INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Forecasting natural gas prices using a hybrid deep learning model and news

René Alexandre Porto da Franca Rocha Filho

Master in Data Science

Mentor:
Prof. Diana E. Aldea Mendes, Associate Professor
Department of Quantitative Methods, Iscte - IUL

September, 2023

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Forecasting natural gas prices using a hybrid deep learning model and news

René Alexandre Porto da Franca Rocha Filho

Master in Data Science

Mentor:
Prof. Diana E. Aldea Mendes, Associate Professor
Department of Quantitative Methods, Iscte - IUL

September, 2023

To my beloved grandfather,

You were a beacon of wisdom and love in my life. Your stories of courage and determination have inspired me to pursue knowledge and embrace challenges. Although you are no longer here, your spirit lives on in my heart, and this dissertation is dedicated to you, my dear grandfather, with profound gratitude for the beautiful memories we shared.

May this work honor the love, support, and encouragement of my family, whose unwavering presence has made this achievement possible. Thank you for being my guiding light and for always believing in me.

With love and appreciation,

René Porto

Acknowledgment

I am deeply grateful for Professor Diana Mendes' invaluable guidance, unwavering support, and insightful advice during my academic journey. Her mentorship has been instrumental in shaping my research and navigating the complexities of this dissertation.

Professor Mendes' expertise and profound knowledge in the field have been a constant source of inspiration, motivating me to strive for excellence. Her willingness to share wisdom and provide constructive feedback has been crucial in refining the ideas and methodologies of this study.

I sincerely appreciate the countless hours Professor Mendes dedicated to reviewing and discussing my research progress. Her encouragement during moments of doubt and belief in my capabilities have been pivotal in overcoming challenges.

I also extend my appreciation to the entire academic community at the University Institute of Lisbon for fostering an environment of intellectual growth and inquiry.

Lastly, my heartfelt acknowledgment goes to my wife, for her unwavering encouragement and love throughout this academic journey, providing me with the strength to pursue my passion for research and education.

This dissertation owes its success to Professor Diana Mendes' guidance, support, and the contributions of all those who have played a significant role in shaping my academic and personal growth. I am truly honored and humbled by their involvement in this work.

Thank you all for being an integral part of this rewarding journey.

Sincerely,

René Porto.

Resumo

A transição para fontes de energia mais limpas na União Europeia prioriza o gás natural, no entanto, a Guerra Russo-Ucraniana causou flutuações imprevisíveis nos preços. Nosso estudo visou aprimorar modelos preditivos explorando dados do GDELT, analisando o desempenho pré e pós-guerra, e comparando modelos de "Deep Learning" (RNN, LSTM, GRUNN). A incorporação de dados de petróleo bruto e sentimento médio da nóticia melhorou significativamente as previsões. Fatores geopolíticos exigem mais pesquisas para garantir segurança energética e desenvolvimento econômico. Empregando a metodologia CRISP-DM, estabelecemos uma abordagem sistemática para enfrentar esses desafios. Nosso estudo contribui com insights valiosos para aprimorar as previsões e adaptar modelos aos complexos mercados de energia.

Abstract

The transition to cleaner energy in the European Union prioritizes natural gas, yet the Russo-Ukrainian War caused unpredictable price fluctuations. Our study aimed to enhance predictive models by exploring GDELT data, analyzing pre- and post-war performance, and comparing deep learning models (RNN, LSTM, GRUNN). Incorporating crude oil and average tone data significantly improved predictions. Geopolitical factors necessitate further research to ensure energy security and economic development. Employing CRISP-DM methodology, we established a systematic approach to address these challenges. Our study contributes valuable insights to enhance predictions and adapt models to complex energy markets.

Contents

CHAPTER 1

Introduction

The shift towards cleaner energy sources is a top priority for the European Union, with natural gas being widely embraced by countries to achieve emission reduction goals. This energy resource is predominantly transported through pipelines and offers convenient storage options.
However, the stability of natural gas prices was significantly impacted by the Russo-Ukrainian War. The weaponization of gas sales by Russia during the conflict caused a drastic increase in gas prices, leading to unprecedented price fluctuations. As a result, the predictive models developed in this study were not trained to anticipate such extraordinary events, and their performance was affected.
This unforeseen instability in natural gas prices during the war highlights the need for further research and model adaptation to account for geopolitical factors that can influence energy markets. The ability to forecast such events accurately will be crucial for ensuring energy security and sustainable economic development in the future.

The primary objectives of this study are as follows:

(1) To investigate whether the utilization of data extracted from GDELT (Global Database of Events, Language, and Tone) contributes to an enhancement in model performance.
(2) To assess whether the predictive model trained with data before the Russo-Ukrainian War demonstrates a similar performance to the model that did not anticipate this historical phase.
(3) To compare the performance of different deep learning models, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN), and determine which one yields the best results.

The dataset of TTF natural gas price, obtained from the yfinance Python library[1], spans from October 23, 2017. Our focus in this study was specifically on the time period between January 2, 2018, and December 30, 2022, comprising a total of 1292 data points. To train the models, we utilized only the data from the first year of the war, employing an 80/20 ratio for training and testing data.
In anticipation of the modeling phase, a set of 15 distinctive features was meticulously

[1]https://pypi.org/project/yfinance/

constructed, encompassing a diverse range of attributes. An exhaustive examination of the collective attributes was undertaken, as detailed in Table 4.1 at page 40. However, this process of evaluation led to a deliberate focus on the distinct constituent elements within the feature set. The ensuing selection comprised three pivotal features: natural gas price, crude oil price, and average tone. The strategic amalgamation of these selected features is delineated in Table 4.3 at page 45, guided by a rationale anchored in the incorporation of a natural gas price baseline intertwined with the nuanced interplay of crude oil price and average tone attributes.

Throughout our study, we observed that no prior research had attempted to forecast natural gas prices during periods of war. Additionally, none of the existing studies utilized the Cameo list of events[2] to predict natural gas prices.

The employed methodology was CRISP-DM, and the basic flow can be observed in Figures 3.3, 3.1, 3.5, and 4.1, at pages 23, 21, 27, and 37, respectively:

Step 1: Extraction of natural gas, crude oil, GDELT news, and weather data.
Step 2: Data analysis.
Step 3: Individual transformation and preprocessing of data.
Step 4: Training and fitting of models using Keras Tuner to select the best hyperparameters.
Step 5: Evaluation of models.

The dissertation is structured as follows:

(1) Introduction: This part gives a quick overview of why the research is important and what it aims to achieve. It introduces the main problem, goals, and questions that the rest of the dissertation will explore.
(2) Methodology and Literature Review: In this section, we explain how we did the research, like collecting and studying data. We also talk about what other people have researched on this topic before. This helps set the groundwork for our own research.
(3) Data and Preprocessing: Now we talk about the information we used for the research. We describe where we got it from, what it includes, and how we made sure it was good to use. This step is really important to make sure our analysis is accurate.
(4) Conclusion: This is the last chapter of the dissertation. We summarize what we found out from our research and discuss how it answers the questions we had. We also talk about what our findings mean for the subject and how they could be useful in real life. We mention any limitations in our research and suggest ideas

[2]http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

for future studies. This chapter gives a nice ending to our whole research journey.

Each of these segments assumes a distinctive and pivotal function in molding the framework and substance of the dissertation, contributing to a comprehensive and unified scholarly composition.

CHAPTER 2

Methodology and Literature Review

2.1. Methodology

The methodology adopted for this study was CRISP-DM (Cross-Industry Standard Process for Data Mining), a widely recognized framework that guides data mining and machine learning projects. CRISP-DM offers a structured approach, breaking down complex problems into manageable steps and ensuring a systematic and well-documented process. The detailed flow of the CRISP-DM methodology is presented in Figures 3.3, 3.1, 3.5, and 4.1 at pages 23, 21, 27, and 37 respectively. This methodology encompasses the following key steps:

(1) **Business Understanding:** In the initial phase, we defined the research objectives and formulated research questions to address the challenges of predicting natural gas prices.

(2) **Data Understanding:** The subsequent step involved data collection and exploration. We obtained and thoroughly examined data on natural gas, crude oil, GDELT news, and weather.

(3) **Data Preparation:** After collecting the data, we performed extensive cleaning and preprocessing, handling missing values, and outliers, and ensuring data quality.

(4) **Modeling:** Various machine learning models, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN), were employed for forecasting natural gas prices. The Keras Tuner was utilized to select the most suitable hyperparameters for the models.

(5) **Evaluation:** The performance of each model was evaluated using metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

By adhering to the CRISP-DM methodology, our research followed a systematic and transparent approach, effectively addressing the challenges of predicting natural gas prices.

2.2. Literature Review

2.2.1. Europe Natural Gas Market

Energy is an important part of social progress and economic development (Kaufmann & Connelly, 2020).

Europe has increased the consumption of natural gas over the years and is transitioning toward a renewable-based energy system (Berrisch & Ziel, 2022).

After the European Union liberalization process in 1998, the market grew fast. These directives issued by the European Commission have given access to infrastructure to third-party companies. Another advantage is creating a Virtual Trading Point, and now natural gas can be traded in any location (Hamie et al., 2020).

The network code was also a game changer because it allows efficient transmission that avoids the congestion in EU gas transmission pipelines is reduced (Hamie et al., 2020).

Investment in cross-border capacity contributes to creating a cross-border relationship between Transmission System Operators via the European Network for Gas long-term contracts. The same investment was also aggregated into the new system, old legacy, or long-term contracts into the new system (Hamie et al., 2020).

The main natural gas hubs in the European Union are presented in Figure 2.1, the red dots indicate the mature hubs with the highest trade rate. Blue dots show active hubs, and yellow dots show hubs with lower trade (Heather, 2021). The two most representative hubs are the Dutch TTF and the British NBP, both classified as mature with 46690 TWh and 10060 TWh in 2020 (Heather, 2021).

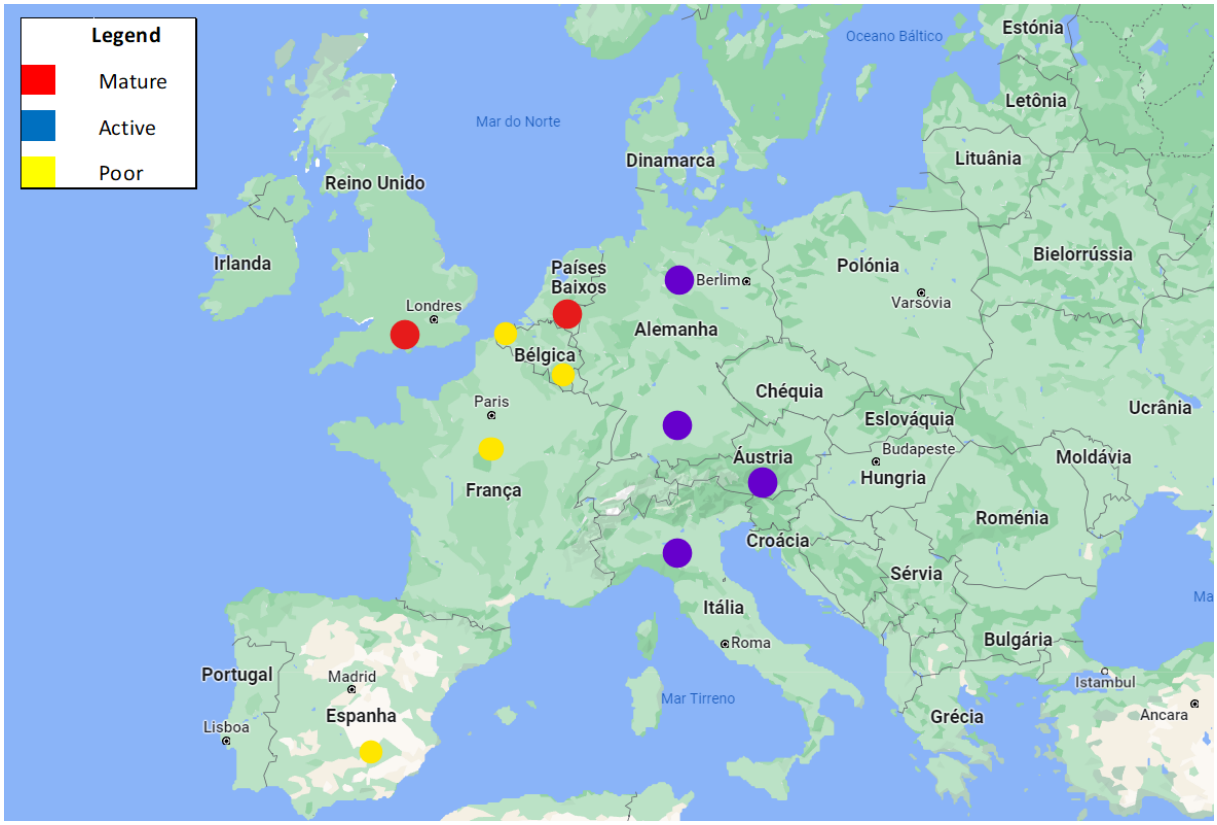

FIGURE 2.1. Map the main natural gas hubs in Europe

The Russian invasion of Ukraine began on Thursday, February 24, 2022. The war has since killed thousands of people, taken the place of millions, and destroyed entire cities (Psaropoulos, 2022).

After the invasion, Russia used the natural gas trade as a weapon, taking advantage of the shortage of European natural gas and the dependence on Russian supply to negotiate

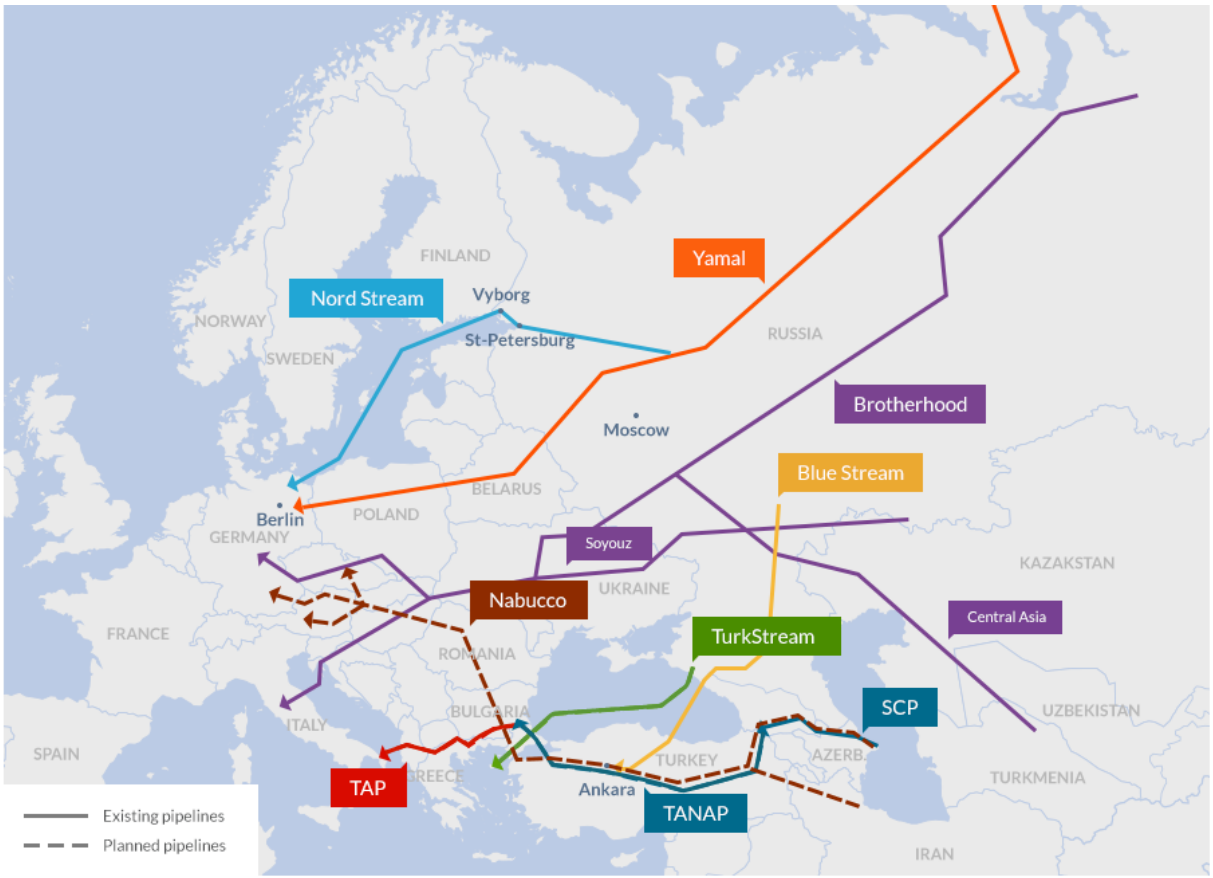

FIGURE 2.2. Map of Gas pipelines between Europe, Russia and Caucasia Sources : Gazprom export ; Gazprom ; Tanap ; Trans Adriatic Pipeline ; BP ; Natural Gaz Europe ; Nord Stream ; South Stream Transport

with European countries not to get involved in the war. In the first 14 days of the invasion, the natural gas price increased by around 180% and Russia started to deliver 60% less gas through the Yamal pipeline (Figure 2.2), this reduction affected imports to France, Austria, Italy, the Czech Republic and Germany. Furthermore, in response to Russia's weaponization of natural gas, the European Union reduced imports through Nord Stream I (Figure 2.2) (*Economic Bulletin Issue 4,* 2022; Halser & Paraschiv, 2022).

The fear of a natural gas shortage in winter made the European Commission propose a new legal obligation to fill underground gas storage to 80% of its capacity by 1 November 2022. Besides, the European Union signed a memorandum for delivering 15 bcm of liquified natural gas with the United States and Qatar (Fabian et al., 2022; *Refilling gas storage for next winter*, 2022).

2.2.2. Forecasting Natural Gas Price

The oldest paper analysed was published in 2010 and the newest was brought out in 2022, in this decade, 2019 was the year with more articles produced, the Figure 2.3 at page 8 presents the number of articles per year.

Table **??** at page **??** shows the most frequent models used to predict natural gas prices found in the literature review. The models are grouped into neural networks, regression,

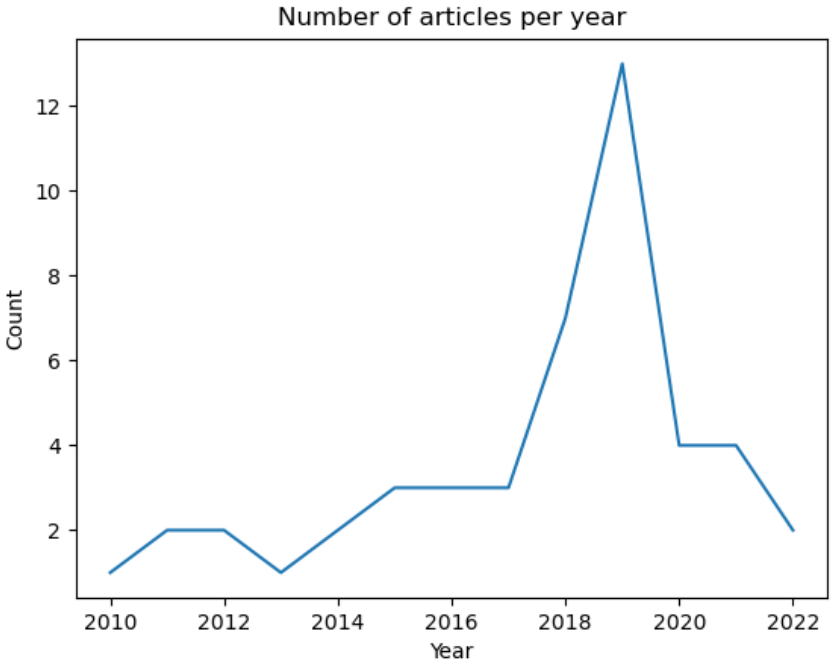

FIGURE 2.3. The number of papers about natural gas prices forecast, per year

TABLE 2.1. Grouped Models in the Literature Review

Grouped Models	Count
Artificial Neural Networks (ANN)	15
Support Vector Regression (SVR)	8
Time Series Models	5
Ensemble Models	3
Regression Models	6
Other Machine Learning Models	8
Other Models	3

auto-regression, decision trees, and other types. Artificial Neural Networks and Support Vector Regression are the most common methods for forecasting natural gas prices as showed in the word cloud in Figure 2.4 at page 9.

The number of hybrid models is slightly more than the traditional method that uses only one model, with 55% and 45% respectively presented on Figure 2.5 at page 9.

The input variables can be a simple natural gas price time series or a list of features (Naderi et al., 2021). After the features related to natural gas, the second most used feature is the input variables linked to Crude oil (Abrishami & Varahrami, 2011; Čeperić et al., 2017; Li et al., 2021; Moting et al., 2019; Naderi et al., 2019; Viacaba et al., 2012). Table 2.2 shows all the features used and counts the occurrences of the exact name. The features are grouped into energy, macroeconomics, weather, and others. The energy group has subgroups such as price, demand, production, consumption, etc.

Carbon, electricity, and natural gas are most affected by meteorological factors (Naderi et al., 2021). The result of Li et al. (2021) presents that the proportion of extremely

FIGURE 2.4. Word cloud of all models used to forecast natural gas price

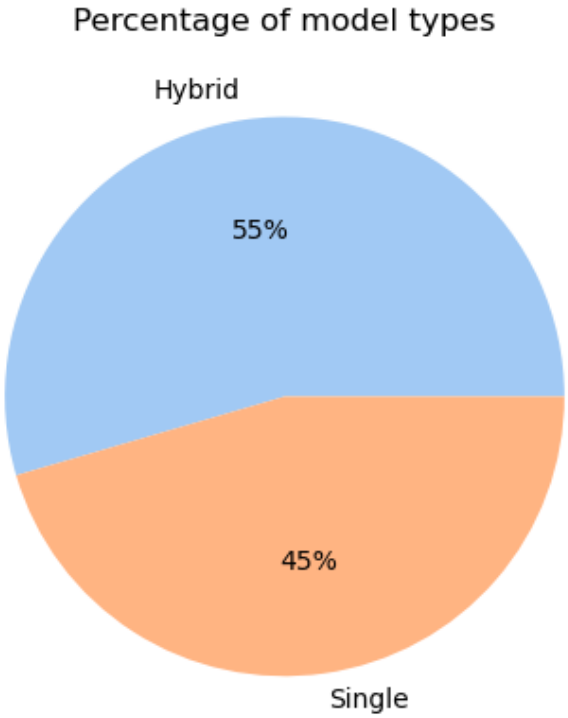

FIGURE 2.5. Percentage of model types

high-temperature weather, the proportion of extremely low-temperature weather, monthly West Texas Intermediate (WTI) crude oil spot prices, natural gas consumption, and natural gas gross withdrawals, all of it come up to predict in different levels of the long-term prices in the Henry Hub natural gas spot, that is the most liquid, but also the most

unpredictable natural gas market in the world (Čeperić et al., 2017). Natural gas prices are insensitive to energy-related and stock-related indexes (Naderi et al., 2021).

Another meaningful connection is between crude oil prices and natural gas prices, the movement of crude oil prices used to influence the natural gas price. Moreover, fluctuations in weather and temperature are used to influence natural gas prices (Moting et al., 2019). The daily total electricity demand in Great Britain has two seasonality effects when it is possible to see less consumption at weekends and a pattern that shows a higher consumption annually during the coldest days of the year (Nguyen & Nabney, 2010).

The analysis of Li et al. (2021) shows that the implementation of natural gas consumption and monthly WTI crude oil prices provides better prediction accuracy for the model that predicts monthly natural gas prices.

News sentiments added as features proved to contain complementary information and can increase the performance of the model by 14.40% compared to a model without the news sentiment (Y. Tang et al., 2019).

Different from the common logic that says "more data is better", Čeperić et al. (2017) finds that for Henry Hub spot prices of natural gas price when it comes to short-term prediction, "less data is better".

The dataset encompassing daily natural gas records presents a substantial variability, spanning from a minimum of 230 data points to a maximum of 5470 data points, as referenced by Naderi et al. (2019) and Siddiqui (2019) respectively. This substantial disparity in the dataset size underscores the diversity in temporal granularity and data availability.

When considering weekly prediction models, Čeperić et al. (2017) employed the dataset with the smallest temporal scope, while Moting et al. (2019a) worked with a significantly larger dataset, containing 886 data points. This wide spectrum in dataset sizes emphasizes the differing preferences and objectives within the field of weekly prediction.

Transitioning to monthly prediction, the dataset sizes exhibit a range of 420 to 2091 records, as reported in studies by Jianwei et al. (2019) and Berrisch and Ziel (2022) respectively. This variation in dataset sizes underscores the distinct temporal resolutions and availability of data within the monthly prediction context.

For the prediction of yearly trends, the sole study conducted by Azadeh et al. (2012) utilized a dataset comprising 40 data points. This notably limited dataset size reflects the challenges inherent in yearly prediction due to the scarcity of available observations.

For a comprehensive overview of these data sizes, refer to Table 5.1, which encapsulates the aforementioned ranges and sizes, encapsulating the diversity and nuances present in the datasets utilized across various prediction timeframes.

For all types of energy prediction studied in the literature review by Naderi et al. (2021) the application of feature engineering led to an average increase of 54.59% in accuracy in the models analyzed (see Table 2.3 at page 12).

TABLE 2.2. List of features used to forecast natural gas price

Input variable(s)	Count
Historical gas price	22
Crude oil price	2
Heating oil price	2
Annual interest rate	1
Proportion of extreme high temperature weather	1
Natural gas price differences	1
Natural gas price rotary rigs	1
Natural gas total consumption	1
Natural gas underground storage volume	1
News	1
OPEC cut production	1
Population	1
Taxes placed on gas price	1
Natural gas imports	1
Temperature	1
Total renewable energy consumption	1
U.S. LNG imports	1
U.S. natural gas gross withdrawals	1
U.S. natural gas marketed production	1
U.S. natural gas pipeline imports	1
U.S. natural gas total consumption	1
WTI crude oil prices	1
WTI crude oil prices differences	1
Natural gas marketed production	1
Natural gas gross withdrawals	1
Natural gas consumption	1
Annual natural gas consumption	1
Coal price	1
Coal price differences	1
Consumer price index	1
Cooling degree-days	1
Economic parameters	1
Electricity price	1
Environmental policy	1
GDP	1
Global demand for crude oil	1
Global demand for gas	1
Heating degree-days	1
Heating oil price difference	1
Historical data of energy demand	1
Internet search	1
Meteorological parameters	1
Monthly WTI crude oil prices	1
Monthly oil price	1
WTI oil spot price	1

Table 2.3. List of feature engineering methods provided on models

feature engineering method(s)	Count
Ensemble Empirical Mode Decomposition (EEMD)	2
Variational Mode Decomposition (VMD)	2
Discrete Wavelet Decomposition (DWD)	1
Feature Selection (FS)	1
Group Method of Data Handling (GMDH)	1
Improved Pattern Sequence Similarity Search (IPSS)	1
Independent Component Analysis (ICA)	1

Table 2.4. List of optimization methods provided on models

Optimizer(s)	Count
Particle Swarm Optimization (PSO)	3
Adaptive Learning Strategy (ALS)	1
Bat Algorithm (BA)	1
Genetic Algorithm (GA)	1

The most common optimization method applied is Particle Swarm Optimization (PSO), which shows an increase in the performance of final models (Čeperić et al., 2017; Li et al., 2021; J. Wang et al., 2021). In our study, we utilized the random search method to find the optimal configuration for our model. We focused on three critical hyperparameters: the number of layers, units, and epochs. The epoch value was consistently set to 20, and we implemented early stopping after 5 epochs without improvement. To explore the impact of the layer count, we conducted trials with a range of one to five layers, increasing by one layer for each attempt. For the units, we varied the values between 32 and 512, with increments of 32 units for each trial.
In the second trial using Keras Tuner, we refined our search based on the best-performing models from the previous round. We restricted the number of layers to either one or two, and the units were limited to the range of 32 to 512, maintaining the same increment value.

It is possible to see in Figure 2.6 that the most common period to forecast is the daily price of natural gas, on which 14 of the articles are working. The second period is the monthly period for 7 papers. The third period is the weekly period with six publications. The last one with only one paper is the yearly period. Researchers do not predict the natural gas price with a horizon of two months, quarters, and semesters.

The single model most used to forecast the price of natural gas are artificial neural networks (ANN), and the second model most used is auto-regressive moving average (ARMA) and support vector regression (SVR).

The idea behind the hybrid model is to combine more than one model to get better performance. Table 2.6 at 15 lists all the models used to develop hybrid models.

The combination varies between two and five models, which can be different or similar, Naderi et al. (2019) worked in a combination of four models of least squares support

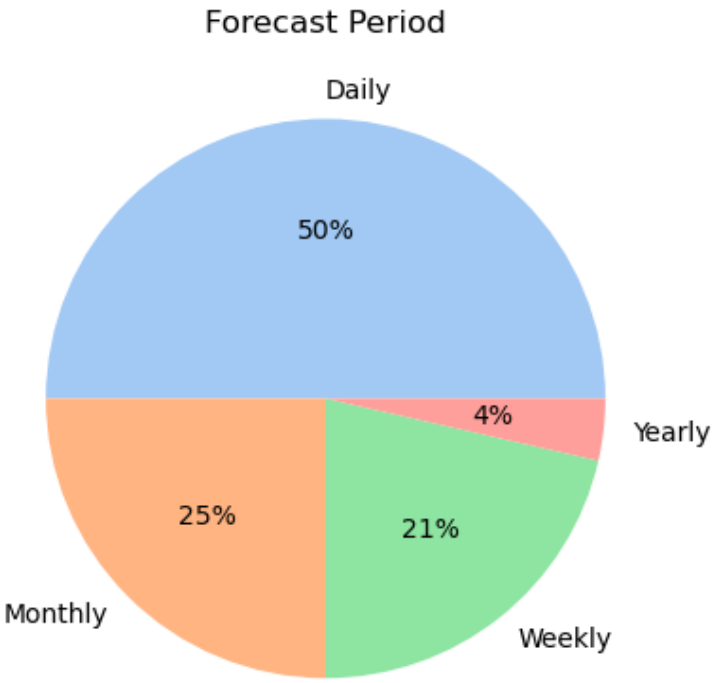

FIGURE 2.6. Forecast period horizon

TABLE 2.5. List of single models

Models	Count
Artificial Neural Networks (ANN)	3
Auto-Regressive Moving Average (ARMA)	2
Support Vector Regression (SVR)	2
Back-Propagation Neural Networks (BPNN)	1
Dynamic Local Linear Regression (DLLR)	1
Extreme Learning Machine (ELM)	1
Gamma test (GT)	1
Gaussian Process Regression (GPR)	1
Autoregressive Neural Networks (ARNN)	1
Group Method of Data Handling (GMDH)	1
Least Square Support Vector Machine (LSSVM)	1
Least squares Regression Boosting (LSBoost)	1
Local Linear Regression (LLR)	1
Random Kitchen Sink (RKS)	1
Random Vector Functional Links (RVFL)	1
Support Vector Machines (SVM)	1
Gradient boosting machines (GBM)	1

vector machine (LSSVM), genetic programming (GP), artificial neural networks (ANN),

and auto-regressive integrated moving average (ARIMA) to predict the price of oil, the annual interest rate, and the daily price of gas.

The study of Li et al. (2021) and J. Wang et al. (2020) compare the hybrid model with a single model and find that the hybrid model performs better. J. Wang et al. (2020) complement with hybrid models when combined with different time-series methods tend to have a better performance. The hybrid model of Li et al. (2021) is a combination of a Deep Belief Network (DBN) with feature engineering of variational mode decomposition (VMD) and particle swarm optimization (PSO).

The study of Čeperić et al. (2017) compared hybrid model applies Support Vector Regression (SVR), Steepwise (SW), and Feature Selection (FS) with another hybrid model that uses Artificial Neural Networks (ANN), Particle Swarm Optimization (PSO), and Feature Selection (FS), and find out that first has a better performance.

The article presented by Jin and Kim (2015) shows that not all hybrid models perform better. The combination of artificial neural networks with wavelet decomposition does not improve the model when compared with Artificial Neural Networks without wavelet decomposition. The other experiment with Auto-Regressive Integrated Moving Average with Wavelet decomposition in the same study demonstrated only a small improvement. But the combination of Auto-Regressive Integrated Moving Average with Wavelet decomposition created the best case for a four-step forecast (Jin & Kim, 2015).

FIGURE 2.7. Word cloud of hybrid models used to forecast natural gas price

2.2.3. GDELT

Global Database of Events, Language, and Tone (GDELT) (https://www.gdeltproject.org/) is a platform that scans news media as printed media, broadcast, and web formats. GDELT creates a database that saves important information in more than 100 languages. The managed data links between every person, organization, location, theme, news source, and event in each corner of the planet. The sentiment extracted from this massive database can be precious in finding the world's feelings (GDELT Project, n.d.).

TABLE 2.6. List of hybrid models

Reference / Abbreviations	Models
Nguyen and Nabney, 2010	Wavelet Transform (WT) Radial Basis Functions Neural Networks (RBFNN) Linear Regression (LR) Group Method of Data Handling (GMDH)
Abrishami and Varahrami, 2011	Group Method of Data Handling (GMDH) Genetic Algorithm (GA) Rule-based Exert System (RES)
Azadeh et al., 2012	Artificial neural networks (ANN) Fuzzy linear regression (FLR) Conventional regression (CR)
Thakur et al., 2015	Moving Average Neural Networks (MANN) Back-Propagation Neural Networks (BPNN)
Jin and Kim, 2015	Discrete Wavelet Decomposition (DWD) Auto-Regressive Integrated Moving Average (ARIMA) Artificial neural networks (ANN)
Dey and Salem, 2017	Gated Recurrent Unit (GRU) Recurrent neural networks (RNN)
Čeperić et al., 2017	Strategic seasonality-adjusted (SSA) Support vector regression machines (SVR) Neural networks (NN) Feature selection (FS) Particle Swarm Optimization (PSO)
Jianwei et al., 2019	Variational Mode Decomposition (VMD) Independent Component Analysis (ICA) Gated Recurrent Unit Neural Networks (GRUNN)
Qin et al., 2019	Ensemble Empirical Mode Decomposition (EEMD) Local Linear Prediction (LLP)
Naderi et al., 2019	Bat Algorithm (BA) Least Square Support Vector Machine (LSSVM) Genetic Programming (GP) Artificial Neural Networks (ANN) Auto-Regressive Integrated Moving Average (ARIMA)
J. Wang et al., 2020	Improved Pattern Sequence Similarity Search (IPSS) Support Vector Regression (SVR) Long-term and Short-term Memory Networks (LSTM)
Li et al., 2021	Variational Mode Decomposition (VMD) Particle Swarm Optimization (PSO) Deep Belief Networks (DBN)
J. Wang et al., 2021	Complete ensemble empirical mode decomposition (CEEMD) Adaptive noise-sample entropy (AN-SE) Particle Swarm Optimization (PSO) Adaptive learning strategy (ALS) Gated Recurrent Unit (GRU)

GDELT is an open platform for research and analysis available for unlimited and unrestricted use for academic, commercial, or government without a fee. The data can be accessed by API from the GDELT website or via the Google Cloud Platform (GCP) (Google Cloud Platform Blog, 2014).

The GDELT project has two versions of event databases, the "GDELT 1.0 Event Database" and the "GDELT 2.0 Event Database". Version 1.0 starts in 1979 through March 31, 2013, and it was updated daily and does not include events reported in the 65 live translated languages. The latest version starts April 1, 2013, has new features, updates every 15 minutes, and includes events reported in articles published in 65 live

translated languages. Furthermore, the project GDELT also has the Global Knowledge Graph data source to analyze images, and other data sets normalized, such as "GDELT 1.0 Event Database Normalization Files" (GDELT Project, n.d.).

The GDELT project has been used in many areas of knowledge. Kwak and An (2014) showed the structure of global news coverage of disasters and their cause. The article finds strong regionalism in the geography news.

To detect events of occupied protests, Qiao et al. (2015) compared the results using a base model using the GDELT database. Models with GDELT features proved to be better with higher accuracy. Hammond and Weidmann (2014) used GDELT to study political violence.

The literature review applied by W. Wang et al. (2016) compared the GDELT project with other similar databases. The study anticipates that there should be a high correlation between GDELT and another database, but the overall correlation does not match the expectations, returning a small correlation. But when the comparison is filtered in each database by country, the experience results in a better correlation.

To analyze public opinion on the energy policy of the Spanish government, Bodas-Sagi and Labeaga (2016) shows a negative feeling about the solar energy policy introduced in 2016.

To predict social unrest events, studies are applying several models such as the hidden Markov model, neural networks, random forest, LSBoost, LSTM, and others (Galla & Burke, 2018; Qiao et al., 2017).

The GDELT project is also successfully employed by Bourgeois et al. (2018) to identify bias in news with success. In the financial field, the GDELT project is used to help predict political crises, oil prices, stock market, and macroeconomic index with considerable improvement (Alamro et al., 2019; Elshendy et al., 2018; Tilly et al., 2020; Zhang et al., 2019).

No paper used the GDELT project to predict the price of natural gas, Y. Tang et al. (2019) used news but from a different source with a single model to predict the price of natural gas.

2.2.4. Results

Table 2.7 presents results of the prediction of the natural gas price by single models. The articles use different performance metrics making the MSE the most common. The best

Table 2.7. List of results of single models to predict daily natural gas price

Reference	Performance
L. Tang et al., 2018 EEMD-based model	MAPE=0.5850
Moting et al., 2019a LSBoost	MAE=0.4493 MSE=0.4376 RMSE=0.6615 R2=0.91
Berrisch and Ziel, 2022 ARMA	MAE=0.3863 CRPS=0.2834 RMSE=1.0843
Salehnia et al., 2013 LLR / DLLR / ANN	LLR (t) MSE=0.29113 DLLR (t) MSE=0.13977 ANN (t) MSE=0.3366
Al-Sharoot and Alramadhan, 2019 ARMA / GMDH	MAE=0.01539 MSE=0.0214
Y. Tang et al., 2019 ANN	MAE = 0.0956; 0.1002; 0.0987; 0.0902 RMSE = 0.1368; 0.137; 0.133; 0.1284
Hu and Trafalis, 2011 SVR	MSE=0.0903 R2=0.9822
Siddiqui, 2019 ARNN	MSE=0.026

Table 2.8. List of results of single models to predict weekly natural gas price

Reference	Performance
Moting et al., 2019a LSBoost	MAE=0.4761 MSE=0.5116 RMSE=0.7153 R2=0.9
Salehnia et al., 2013 LLR / DLLR / ANN	LLR (t) MSE=3.4317 DLLR (t) MSE=0.25566 ANN (t) MSE=0.8268
Viacaba et al., 2012 SVR	RMSE <0.03

result for daily prediction using a single model is MSE equal to 0.0214 by Al-Sharoot and Alramadhan (2019) with an ARMA-GARCH model.

The results of single models that forecast natural gas prices weekly are presented in Table 2.8 with just a few articles, and the best precision comes from the study of Salehnia et al. (2013) with an MSE equal to 0.25566 by applying the dynamic local linear regression model (DLLR).

TABLE 2.9. List of results of single models to predict monthly natural gas price

Reference	Performance
Moting et al., 2019 ANN/SVM/GBM/GPR	ANN R2=0.8904 MAE=0.5115 MSE=0.5363 RMSE=0.7247 MAPE=0.1117 SVM R2=0.8437 MAE=0.5673 MSE=0.7673 RMSE=0.8757 MAPE=0.1202 GBM R2=0.8006 MAE=0.6490 MSE=0.9786 RMSE=0.9888 MAPE=0.1366 GPR R2=0.8374 MAE=0.6026 MSE=0.7980 RMSE=0.8932 MAPE=0.1270
Moting et al., 2019a LSBoost	MAE=0.6859 MSE=1.1166 RMSE=1.0567 R2=0.78
Berrisch and Ziel, 2022 SVR	CRPS=0.2126 MAE=0.3010 RMSE=0.3995
Salehnia et al., 2013 ARMA	Monthly LLR (t) MSE=3.864 DLLR (t) MSE=2.5932 ANN (t) MSE=0.9831

Table 2.9 shows the articles that predict natural gas prices monthly, and the best result is obtained by an Artificial Neural Networks with an MSE = 0.5663 (Moting et al., 2019).

The only article that predicts yearly natural gas prices applies a single model of Artificial neural networks (ANN), fuzzy linear regression (FLR), and conventional regression (CR). These models archive the best result from conventional regression with MAPE = 0.260. Table 2.10 shows all results.

Table 2.10. List of results of hybrid models to predict yearly natural gas price

Reference	Performance
Azadeh et al., 2012 ANN-FLR-CR	MAPE (Average) CR = 0.2260 ANN = 0.2978 FLR = 0.2470

Table 2.11. List of results of hybrid models to predict daily natural gas price

Reference	Performance
Qin et al., 2019 EEMD-LLP	RMSE = 0.035 MAPE = 0.01244 Dstat = 0.908
J. Wang et al., 2020 IPSS-SVR-LSTM	MAPE = 0.0555 MER = 0.0549
Abrishami and Varahrami, 2011 GMDH-GA-RES	Dstat ¿ 0.7 RMSE ¡ 2.942
Thakur et al., 2015 MANN / BPNN	MSE ¡ 0.1
Naderi et al., 2019 BA-LSSVM-GP-ANN-ARIMA	R2 = 0.9611 RMSE = 0.06
Čeperić et al., 2017 SSA-SVR/NN-FS-PSO 5 variables	SVR SW (Steepwise) MAPE = 0.221 RMSE = 0.1401 10 variables SVR SW (Steepwise) MAPE = 0.218 RMSE = 0.1375

The most widely used evaluation metric is the root mean square error (RMSE). Table 2.11 shows all models that forecast daily natural gas prices using different hybrid models. The most accurate prediction is given by a combination of Ensemble Empirical Mode Decomposition (EEMD) and Local Linear Prediction (LLP), resulting in an RMSE = 0.035 developed (Qin et al., 2019).

The list of articles that predict the price of natural gas weekly is in Table 2.12. The best result is proposed by Jin and Kim (2015) using a hybrid model of discrete wavelet decomposition (DWD) and artificial neural networks (ANN) with RMSE = 0.1278 precision.

For a monthly forecast of natural gas prices, Table 2.13 lists three articles, the one with the best results has a MAPE between 0.001691 and 0.00413. The models that meet this precision are the combination of Variational Mode Decomposition (VMD), Independent Component Analysis (ICA), and Gated Recurrent Unit Neural Networks (GRUNN), applied by Jianwei et al. (2019).

Table 2.12. List of results of hybrid models to predict weekly natural gas price

Reference	Performance
Jin and Kim, 2015 DWD-ANN / DWD-ARIMA	Wavelet with ANN MAE = 0.0985 MAPE = 0.033747 RMSE = 0.1278 Wavelet with ARIMA MAE = 0.1112 MAPE = 0.037018 RMSE = 0.1366
J. Wang et al., 2021 CEEMDAN-SE-SO-ALS-GRU	Dstat = 0.519 MAE = 0.114 MSE = 0.025 RMSE = 0.158 R2 = 0.889
Čeperić et al., 2017 SSA-SVR / NN-FS-PSO 5 variables	SVR SW (Steepwise) MAPE = 0.423 RMSE = 0.2904 10 variables SVR SW (Steepwise) MAPE = 0.431 RMSE = 0.2782

Table 2.13. List of results of hybrid models to predict monthly natural gas price

Reference	Performance
Jianwei et al., 2019 VMD–ICA–GRUNN–SVR	Dstat = 0.730159-0.845238 MAD = 0.0201-0.0776 MAPE = 0.001691-0.00413 RMSE = 0.0407-0.1196 R2 = 0.95-0.991
Li et al., 2021 VMD-PSO-DBN	MAE = 0.125 MAPE = 0.0481 RMSE = 0.082 FLR = 0.2470
Nguyen and Nabney, 2010 WT-RBFNN-LR-GARCH	MAE = 0.01699 MAPE = 2.019 MSE = 0.15384

CHAPTER 3

Exploratory Data Analysis and Preprocessing

3.1. Exploratory Data Analysis

We utilized four primary data sources for our analysis, namely natural gas price, news data, weather data, and crude oil price. The integration of these features to the main dataset was done individually, the flow of the ETL process is illustrated in Figure 3.3 at page 23. The news data, extracted from the GDELT project, underwent a separate integration process. A Spark environment was created due the number of lines to be processed, the first step of the flow was a extraction of the raw data from GDELT API, after that we analyzed, cleaned, trandformed, and evaluated the news data, last step we exported the data into a parquet data format to optimize the size of our dataset (Figure 3.1). The final integration step involved combining these two features with weather data and crude oil price at the same level, as shown in Figure 3.5.

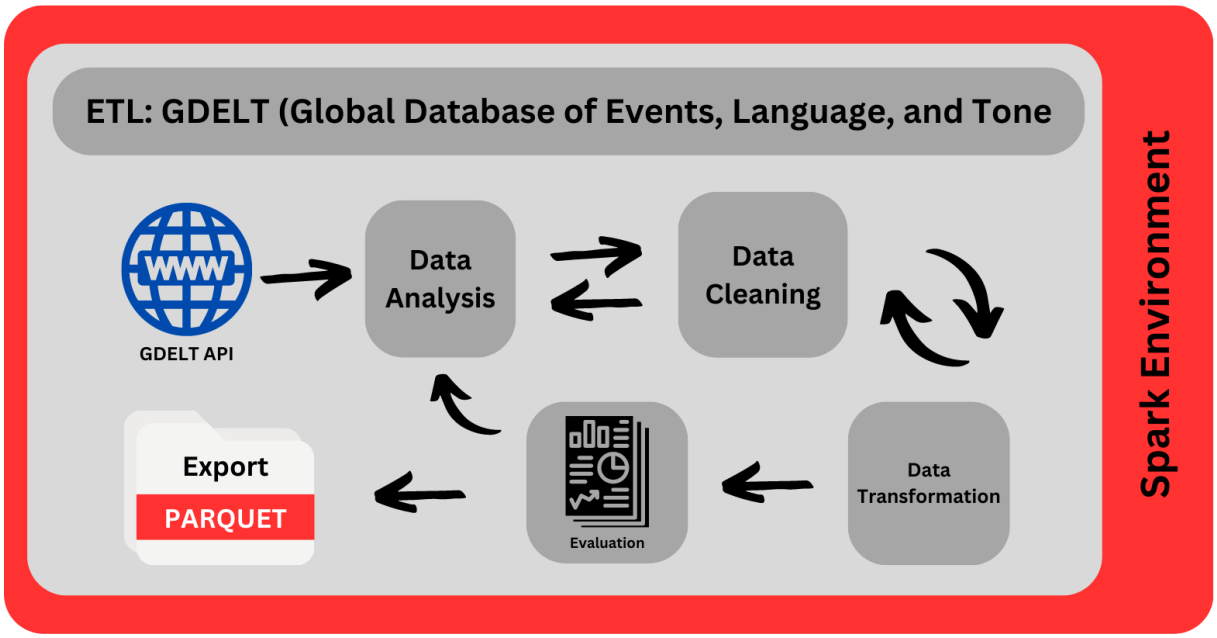

FIGURE 3.1. ETL process for extraction, analysis, cleaning, transformation, and evaluation of news data from GDELT API, with exportation to a parquet file.

Multiple trading hubs exist for natural gas prices, and for our forecasting purposes, we chose to focus on the Dutch Title Transfer Facility (TTF) hub price like Berrisch and Ziel (2022). We obtained the necessary data using the Yahoo Finance API through the yfinance Python library. The description of each column can be found in Table 3.1, and Figure 3.2 shows natural gas prices, and 3.4 presents the natural gas price volume.

The dataset of TTF natural gas prices spans from October 23, 2017 and has been updated daily. For our study, we focused on the period between January 2, 2018, and December 30, 2022, which encompasses a total of 1260 daily natural gas prices. However, it is important to note that there are 32 missing values within this time frame, resulting in a total of 1292 days included in our analysis.

The initial analysis involved performing a statistical summary on Open, High, Low, Volume, Dividends, Stock Splits, and Close columns, as presented in Table **??**. Through this analysis, we observed that both the Dividends and Stock Splits variables exhibited a consistent value of zero across rows.

After conducting the statistical analysis and examining the time-series visualization on natural gas price, during this analysis columns Open, High, Low, and Volume did not match the correlation criteria between -0.70 and 0.70 and do not have strong Granger Causality, and Dividends and Stock Splits all values is equal to zero, is this case we made a decision to retain only the "Close" column as variable and target.

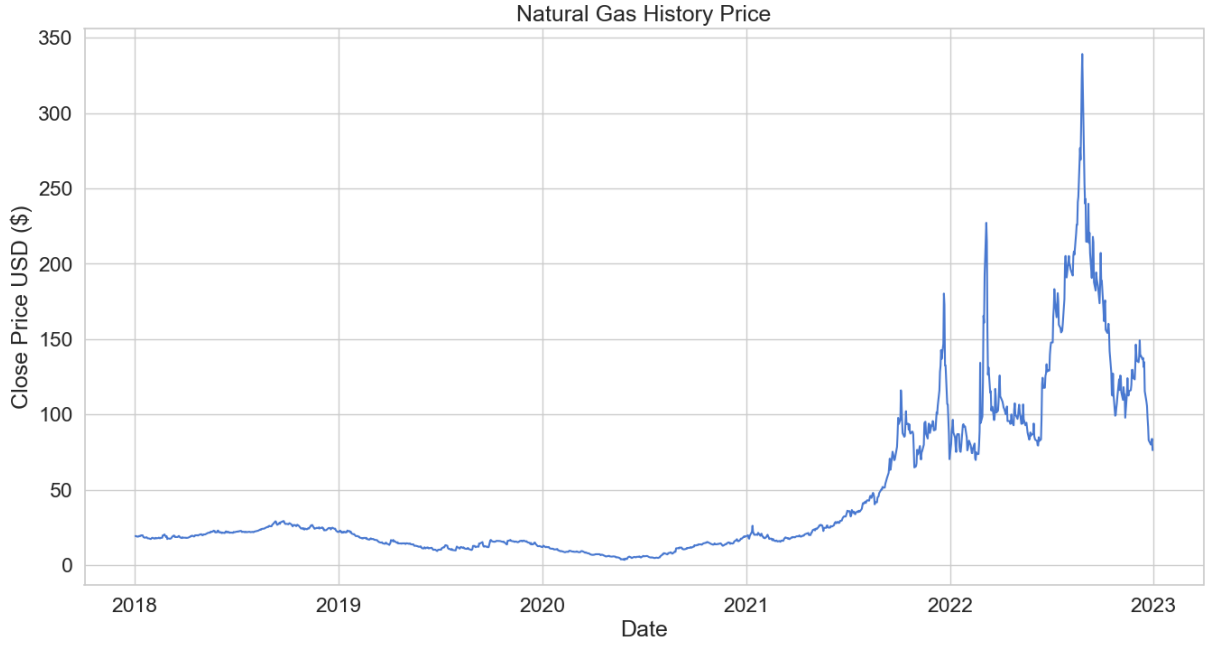

FIGURE 3.2. Natural gas price time-series.

TABLE 3.1. Natural Gas Price TTF Data Columns

Column	Description
Open	The opening price of the natural gas
High	The highest price of the natural gas during the day
Low	The lowest price of natural gas during the day
Volume	The trading volume of the natural gas
Dividends	Any dividends issued for the natural gas
Stock Splits	Any stock splits that occurred for the natural gas
Close	The closing price of the natural gas

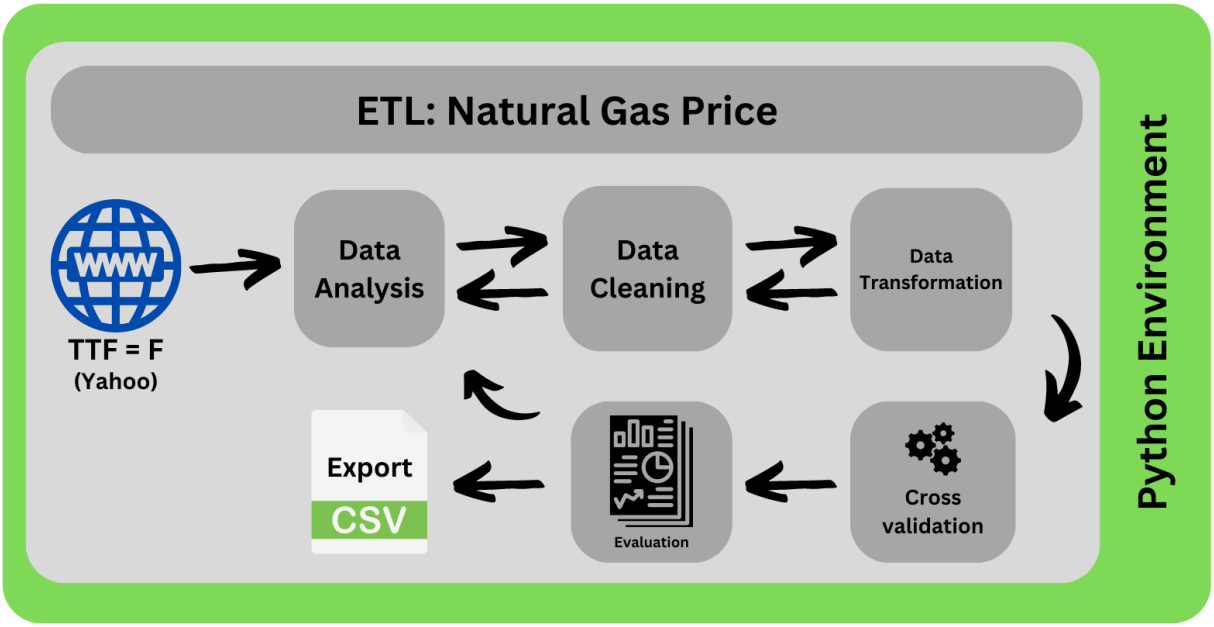

FIGURE 3.3. ETL process for extraction, analysis, cleaning, transformation, cross-validation, and evaluation of TTF natural gas price data from Yahoo API, with Exportation to a CSV file.

3.1.1. GDELT - Global Database of Events, Language, and Tone

In order to handle the large volume of data, we utilized a Spark environment and developed a function to extract news data from the GDELT API. The function includes a filtering mechanism to retrieve only the news articles based on specific Cameo codes from Event Data Project, Department of Political Science, Pennsylvania State University (March 2012), as listed in Table 3.2. Subsequently, the extracted news data was saved in a Parquet format for further processing. The entire process of integrating the news data extracted from the GDELT API is illustrated in Figure 3.1.

The total number of lines obtained from the extraction process amounted to 18,700,299, comprising 58 columns. These columns can be categorized into different types, The main topic of the news is actors, for example, the Actor of news that informs about the deal of European Union with US to deliver natural gas price, the Actor one is European Union and Actor two is US. Actor one and two columns have informations about the name, country, ethnic, religion, and location, shown in Tables 5.2 and 5.3. Geographic information is summarized in Table 5.4. Numeric variables store numbers about the news, as the tone of the news, number articles related to the event, and number of sources reporting the event, all numeric variables of GDELT are described in Table 5.5, and columns slated for deletion can be found in Table 3.3.

3.1.2. Crude Oil Price

The data source of crude oil price was obtained from Investing.com (Accessed on July 19, 2023) with the same period of natural gas price. The structure of the columns is

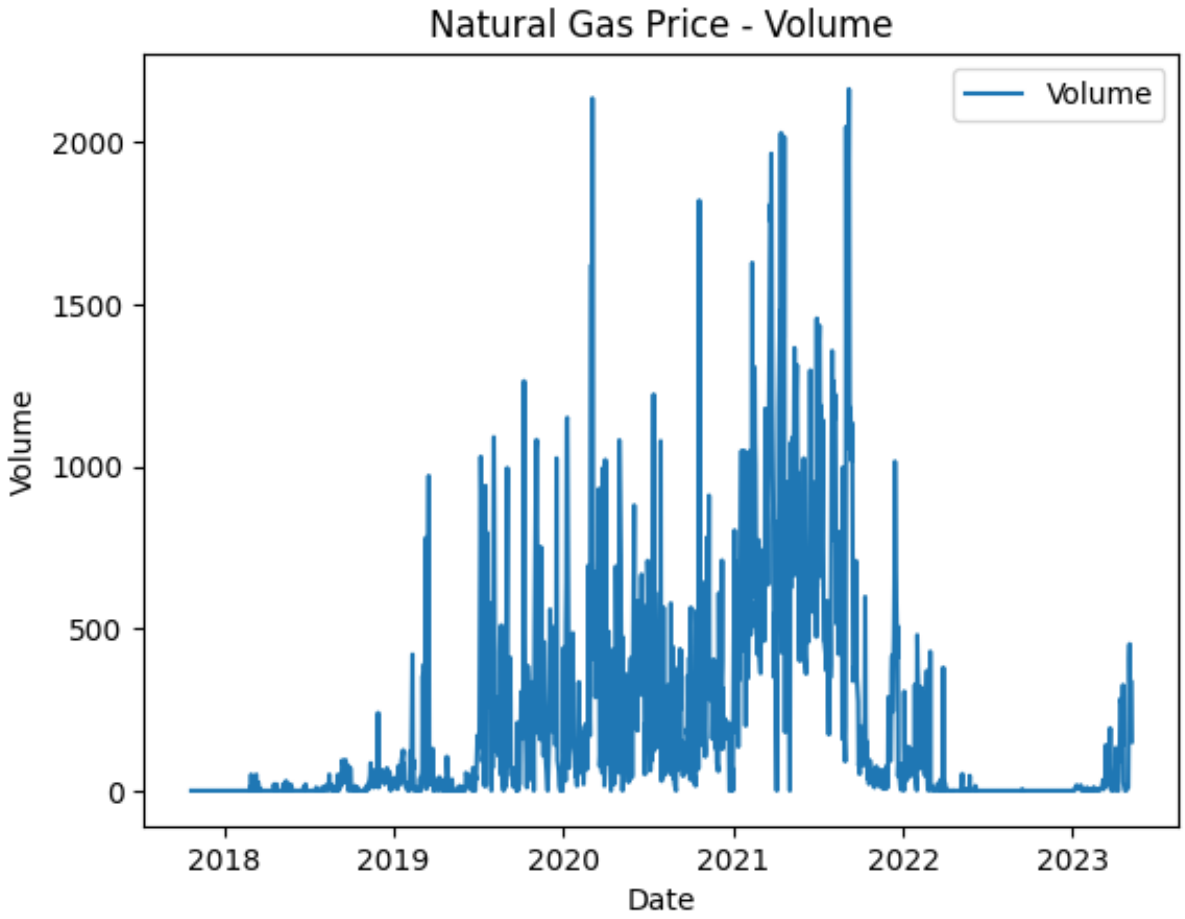

FIGURE 3.4. Visualization of natural gas prices volume.

TABLE 3.2. Cameo event root codes used.

Event Root Code	**Event Root Name**
13	Threaten
14	Protest
15	Exhibit force posture
16	Reduce relations
19	Fight
20	Use unconventional mass violence

very similar to the natural gas price, with Open price, High price, Low price, and Change percentage price compared with previous date, as can be see in Table 3.5 Following the completion of the statistical analysis and a thorough examination of the time-series visualization for crude oil time series, it was observed that the columns labeled Open price, High price, Low price, and Change percentage price did not meet the correlation criteria falling within the range of -0.70 to 0.70. Furthermore, these columns did not exhibit significant Granger Causality and based on this we only used the column Price_oil as feature.

Table 3.3. Deleted columns from GDELT.

Column Name	Description
MonthYear	Month and year of the event
Year	Year of the event
FractionDate	Fractional date representation of the event
EventCode	Code representing the event category
Actor1Geo_FeatureID	Feature ID of the geographic location for Actor 1
Actor2Geo_FeatureID	Feature ID of the geographic location for Actor 2
ActionGeo_FeatureID	Feature ID of the geographic location for the action
DATE ADDED	Date the record was added
SOURCEURL	URL of the source for the event

Table 3.4. Sum of null values for each GDELT column that has null values.

Column Name	Sum of Null
IsRootEvent	18604188
Actor1Geo_Lat	2274519
Actor1Geo_Long	2271529
Actor2Geo_Lat	5505621
Actor2Geo_Long	5503481
AActionGeo_Lat	563853
ActionGeo_Long	560217

Table 3.5. Description of Columns in Crude Oil Price Data

Column	Description
Dates	The dates corresponding to the crude oil price observations
Price_oil	The closing price of crude oil for the given date
Open_oil	The opening price of crude oil for the given date
High_oil	The highest price of crude oil reached during the date
Low_oil	The lowest price of crude oil reached during the date
Change_%_oil	The percentage change in crude oil price compared to the previous date

3.1.3. Weather

The weather dataset used in this study was obtained from NCEI (Accessed on July 19, 2023), specifically from the Rhein-Main station located in Germany. The dataset includes various weather measurements such as precipitation, snowfall, snow depth, maximum temperature, minimum temperature, and average temperature. However, for our analysis, we focused solely on the average temperature time series as it was the relevant variable for our study. The detailed description of each column can be found in Table 3.6.

3.1.4. Data Integration

The integration of all the features was conducted within a Spark environment due to the voluminous nature of the data being processed. The natural gas price, news data, and other relevant features were combined by joining them together into a Spark data frame.

TABLE 3.6. Description of Columns in Weather Data

Column	Description
PRCP	Precipitation (Rainfall)
SNOW	Snowfall
SNWD	Snow Depth
TMAX	Maximum Temperature
TMIN	Minimum Temperature
TAVG	Average Temperature

TABLE 3.7. Sum of null values for each weather column that has null values.

Column Name	Sum of Null
Prcp_temp	60
Prcp_temp	1824
Snow_temp	60
Wind_temp	60
Max_temp	60
Min_temp	60
Avg_temp	3

TABLE 3.8. Days with average temperature null.

Dates	Prcp_temp	Snow_temp	Wind_temp	Max_temp	Min_temp
20210211	0	null	10	-9	-9.7
20210212	0	null	10	-5	-7.3
20210211	0	null	0	16.5	8

Following this integration, a series of steps were executed, including analysis, data cleaning, transformation, and evaluation. As a result, a comprehensive file containing all the features was generated and saved for further analysis. For a visual representation of this integration process, please refer to Figure 3.5.

3.2. Preprocessing

In order to preprocess the dataset, several steps were undertaken to optimize the data before training the model. The preprocessing began with addressing missing values by employing backward and forward fill, the next method was linear, cubic, and quadratic interpolation, another one was imputation of mean value, and the last one was seasonal decomposition, that we replaced the null values by trend and seasonal. Subsequently, the data was categorized, with textual information transformed into index codes. Correlations and causality between variables were identified to filter and select the correlated features between -0.7 and 0.7. Outliers were also detected and handled, by taking the mean value of before and after data of time series. Lastly, statistical and mathematical operations were performed to aggregate the data on a daily basis. These preprocessing steps aimed

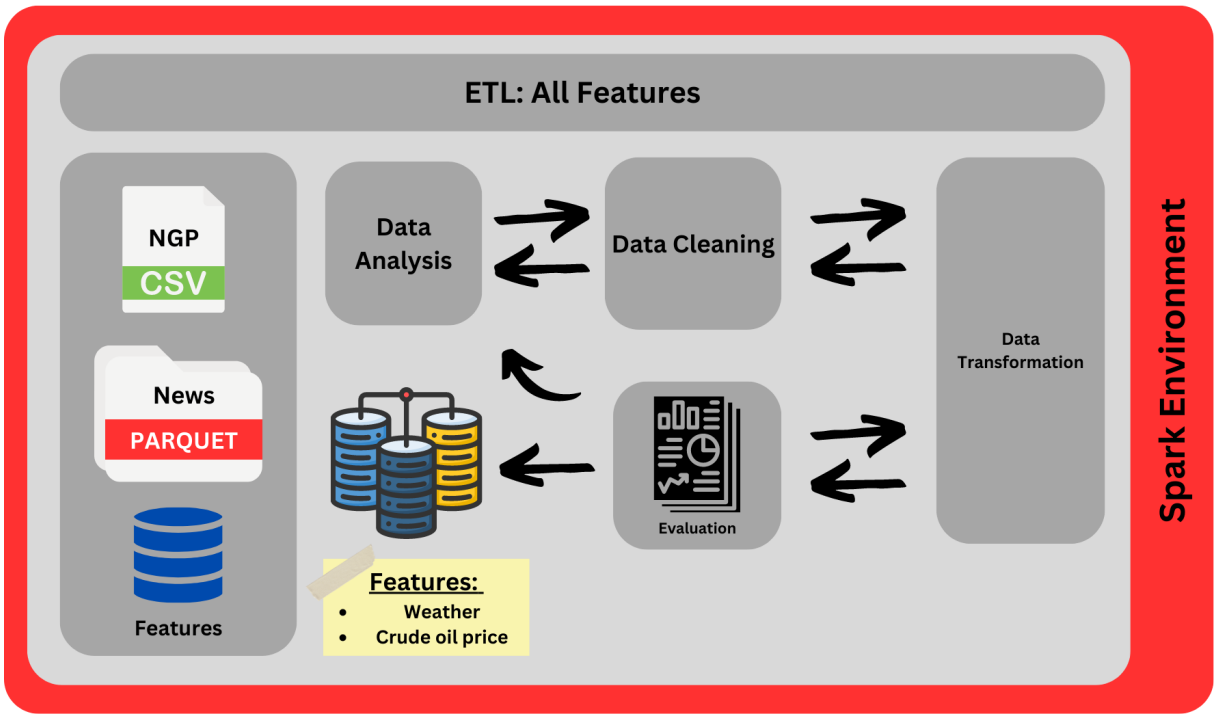

FIGURE 3.5. ETL process for analysis, cleaning, and integration of features into a parquet file.

to ensure the quality and preparedness of the dataset for subsequent modeling tasks.

3.2.1. Replacement of Null Values

For natural gas prices, a unique scenario arises where the identification of missing data points is not solely based on their absence. Instead, an additional step is taken to determine if these null values coincide with weekends or holidays. Through this analysis, it was revealed that there were 32 missing values that required further handling. Several methods were then applied to impute these missing values, aiming to obtain the most accurate representation of the actual values.

The first approach employed for filling the missing values was backward and forward filling techniques. Backward filling involved propagating the last observed value backward in time to fill the missing values, ensuring a smooth and continuous representation of the data. This method is illustrated in Figure 5.11. Similarly, forward filling was applied to propagate the next observed value forward in time to fill any remaining missing values. The application of forward filling can be visualized in Figure 5.12 (van Buuren, 2012).

In addition to backward and forward filling, another method used for imputing missing values was interpolation. Cubic interpolation is shown in Figure 5.13, linear interpolation is shown in Figure 5.14, and quadratic interpolation is shown in Figure 5.15. The visual analysis indicates a satisfactory evolution of the time series after applying these interpolation techniques (van Buuren, 2012).

Another method employed was seasonal decomposition filling. Seasonal decomposition

with the trend is depicted in Figure 5.19, while seasonal values are presented in Figure 5.17. However, it was observed that this method did not provide a natural substitution for the null values. Therefore, a mean value substitution was also utilized, as shown in Figure 5.16 applying the value for time-series (van Buuren, 2012).
In Figure 3.6, it is evident that many of the applied methods did not result in a natural filling of the null values. To gain a better understanding of the filled values, a closer look is provided in Figures 3.7 and 3.8, which depict a specific slice of the data with a null value in the middle. These zoomed-in plots clearly show that the mean and seasonal decomposition methods do not naturally fill the missing values, as the substituted values noticeably differ from the surrounding data points.
To determine the most suitable method for handling missing values in the natural gas price dataset, we conducted a correlation analysis. The results of this analysis, as presented in Table 3.9, indicated that the mean and seasonal decomposition with trend methods performed well. These results suggest that the natural gas price exhibits a tendency towards the mean value and a seasonal pattern. However, instead of choosing one of these methods, we decided to use forward filling. This decision was based on the observation that forward filling effectively fills the null values without introducing significant deviations from the surrounding data points.

TABLE 3.9. R-square results of cross-validation for all methods used to fill null values of natural gas price.

Methods	**R2**
Mean	0.999298
Seasonal Decompose: Trend	0.999135
Forward fill	0.999126
Interpolation: linear	0.999125
Backward fill	0.999105
Interpolation: Cubic	0.999087
Interpolation: Quadratic	0.999081
Seasonal Decompose: Seasonal	0.998753

3.2.2. Categorization

The categorical columns in the dataset were derived from the news data and can be observed in Table 3.10, which provides the distinct count for each category. In order to preprocess these categorical columns, we utilized the Pyspark machine learning function called StringIndexer. This function was selected for its efficiency in handling large datasets, as memory management is crucial for successful transformation. The outcome of this categorization process was a more compact and manageable dataset (Apache Spark, Accessed on July 19, 2023).

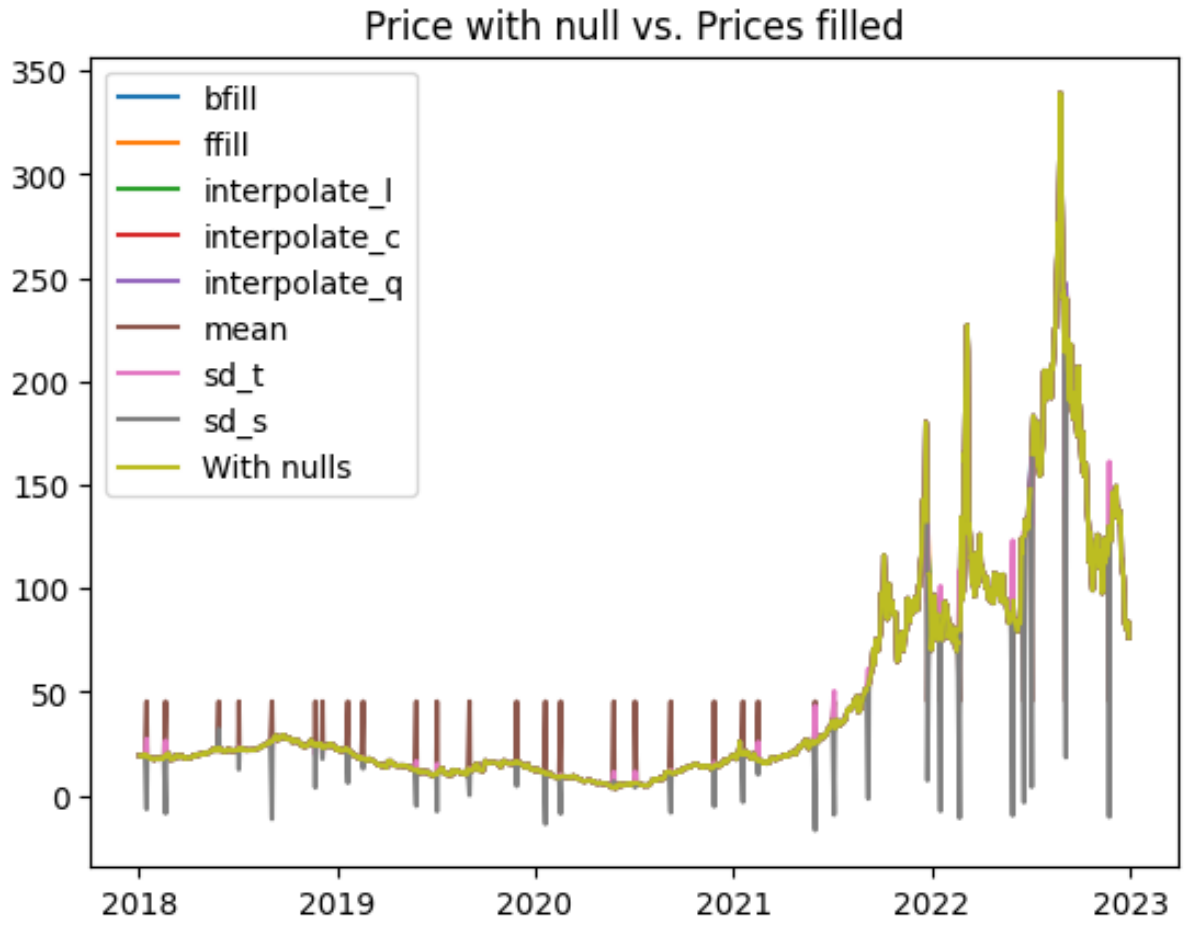

FIGURE 3.6. Compilation of methods to fill null natural gas prices for all time series.

3.2.3. Correlations

The initial step involved generating a Pearson Correlation Matrix to examine the linear relationships between all variables in the dataset. This matrix can be visualized in Figure 3.9, and the corresponding correlation values were presented in Table 5.7 for negative correlations and Table 5.8 for positive correlations. Upon analysis, it was observed that the natural gas price exhibited a positive correlation of 0.722026 with the crude oil price. To further investigate this finding, a zoomed-in plot (Figure 3.10) focusing on these two columns was generated (Downey, Accessed on July 19, 2023).
In order to filter the columns based on their correlation values, a criterium was set to include only those with correlations greater than 0.7 and less than -0.7 (see Table 5.6) (Downey, Accessed on July 19, 2023).

A series of scatter plots was generated to explore potential non-linear correlations. The columns "Actor1Name_Idx" and "Actor1Code_Idx" did not exhibit any significant correlation with the natural gas price, as depicted in Figures 5.2 and 5.1. Similarly, the "Avg_temp" feature displayed a discernible pattern, but it was not conclusive in determining its correlation with the natural gas price, as shown in Plot 5.3.

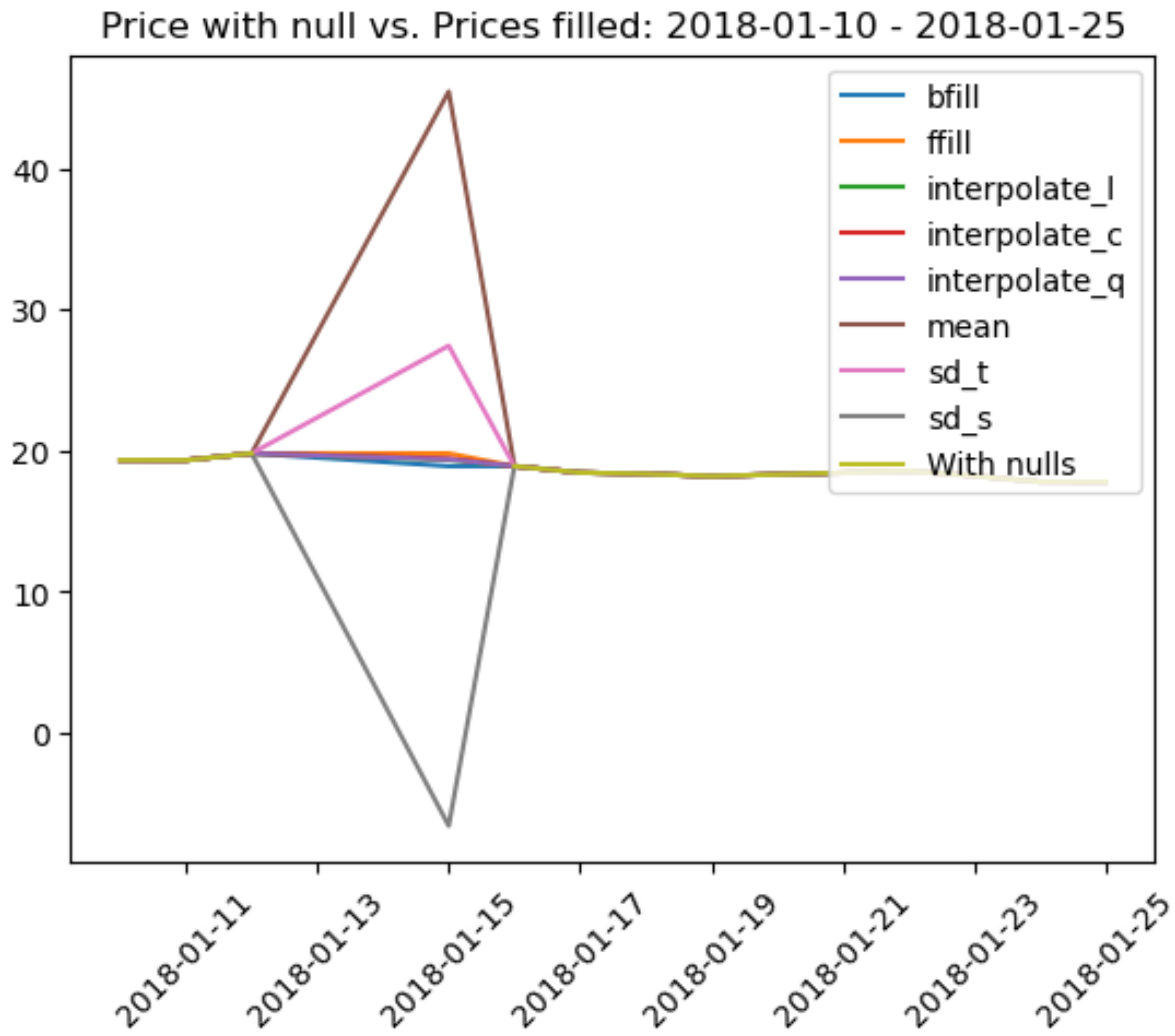

FIGURE 3.7. Compilation of methods to fill null natural gas prices for 2018-01-15.

On the other hand, the "Avg_Tone" feature revealed a wider range of tone values when the price of natural gas was below 50, as depicted in Figure 5.4. This trend was also observed for the "NumArticles," "NumMentions," and "NumSources" columns, indicating a higher volume of articles when the natural gas price was low (respectively in Figures 5.5, 5.6, and 5.7).

Regarding the weather variables, "Prcp_temp" and "Wind_temp" did not display any clear patterns or correlations with the natural gas price, as seen in Figures 5.8 and 5.9.

Overall, the most visually correlated feature with the natural gas price was the crude oil price, as demonstrated in Figure 5.10. This observation is consistent with the results obtained from the Pearson Correlation analysis, indicating a strong correlation between the prices of natural gas and crude oil.

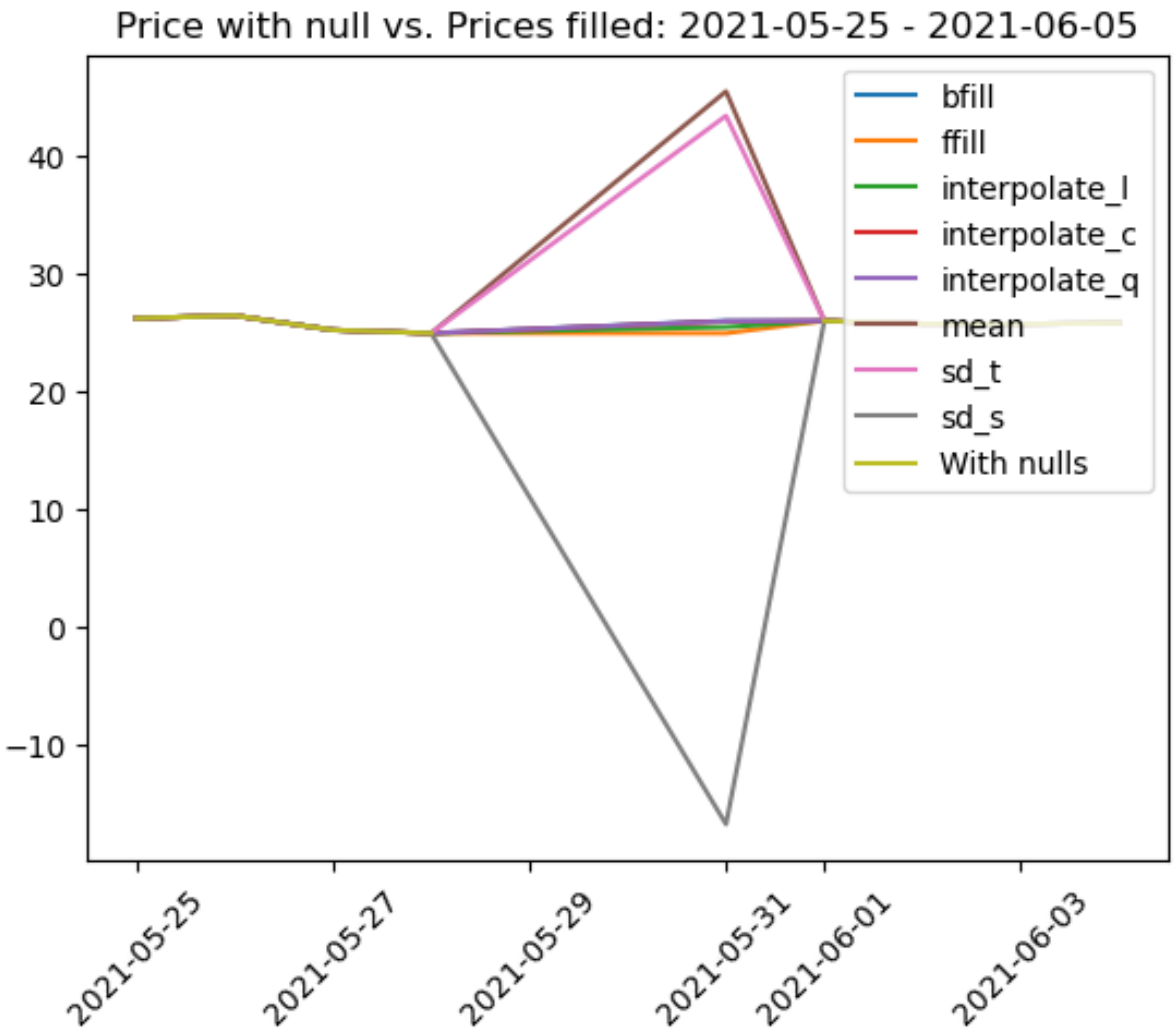

FIGURE 3.8. Compilation of methods to fill null natural gas prices for 2021-05-31.

3.2.4. Granger Causality

Granger causality is a statistical technique employed to evaluate whether a given time series can anticipate or project the behavior of another time series. This entails examining whether past data points from one-time series offer valuable insights into anticipating future data points from another, suggesting a plausible causal connection within time-dependent datasets. It's crucial to emphasize that Granger causality doesn't necessarily establish a direct cause-and-effect relationship; instead, it identifies predictive correlations rooted in statistical trends.

To conduct the Granger Causality analysis, we set the significance level (alpha) to 0.05, and the lag parameter to values of 1, 5, and 10. The analysis was performed for the following features:

When analyzing Granger causality with lag parameters of 1, 5, and 10, consistent findings emerged. Specifically, the variables Wind_temp, Price_oil, and Avg_temp displayed a lack of substantial Granger causality. Tables 3.11, 3.12, 3.13, 3.14, 3.15, 3.16, 3.18, 3.17, and

TABLE 3.10. Sum of distinct values for all category features.

Columns	**Value**
IsRootEvent	1
EventBaseCode	38
EventRootCode	6
Actor1Code_Idx	10499
Actor1Name_Idx	8697
Actor1CountryCode_Idx	220
Actor1KnownGroupCode_Idx	56
Actor1EthnicCode_Idx	357
Actor1Religion1Code_Idx	16
Actor1Religion2Code_Idx	20
Actor1Type1Code_Idx	34
Actor1Type2Code_Idx	27
Actor1Type3Code_Idx	24
Actor2Code_Idx	9754
Actor2Name_Idx	8159
Actor2CountryCode_Idx	221
Actor2KnownGroupCode_Idx	55
Actor2EthnicCode_Idx	344
Actor2Religion1Code_Idx	16
Actor2Religion2Code_Idx	20
Actor2Type1Code_Idx	33
Actor2Type2Code_Idx	28
Actor2Type3Code_Idx	23
QuadClass_Idx	2
Actor1Geo_Type_Idx	6
Actor1Geo_FullName_Idx	189346
Actor1Geo_CountryCode_Idx	255
Actor1Geo_ADM1Code_Idx	4136
Actor2Geo_Type_Idx	6
Actor2Geo_FullName_Idx	163323
Actor2Geo_CountryCode_Idx	253
Actor2Geo_ADM1Code_Idx	4088
ActionGeo_Type_Idx	6
ActionGeo_FullName_Idx	200253
ActionGeo_CountryCode_Idx	255
ActionGeo_ADM1Code_Idx	4160

3.19 presents details of the Granger causality result.

Features GoldsteinScale, NumMentions, NumSources, NumArticles, QuadClass, AvgTone, Threaten, Protest, Exhibit_force_posture, Reduce_relations, Fight, and Use_unconvetional_mass_vi show a strong causality relationship with the natural gas price.

TABLE 3.11. Granger causality results of feature Wind_temp with one lag, without a strong Granger causality.

Wind_temp - Lag = 1				
ssr based F test	F=2.8113	p=0.0938	df_denom=1288	df_num=1
ssr based chi2 test	chi2=2.8179	p=0.0932	df=1	
likelihood ratio test	chi2=2.8148	p=0.0934	df=1	
parameter F test	F=2.8113	p=0.0938	df_denom=1288	df_num=1

TABLE 3.12. Granger causality results of feature Price_oil with one lag, without a strong Granger causality.

Price_oil - Lag = 1				
ssr based F test	F=0.2684	p=0.6045	df_denom=1288	df_num=1
ssr based chi2 test	chi2=0.2690	p=0.6040	df=1	
likelihood ratio test	chi2=0.2690	p=0.6040	df=1	
parameter F test	F=0.2684	p=0.6045	df_denom=1288	df_num=1

TABLE 3.13. Granger causality results of feature Avg_temp with one lag, without a strong Granger causality.

Avg_temp - Lag = 1				
ssr based F test	F=0.4324	p=0.5110	df_denom=1288	df_num=1
ssr based chi2 test	chi2=0.4334	p=0.5103	df=1	
likelihood ratio test	chi2=0.4333	p=0.5104	df=1	
parameter F test	F=0.4324	p=0.5110	df_denom=1288	df_num=1

TABLE 3.14. Granger causality results of feature Wind_temp with five lags, without a strong Granger causality.

Wind_temp - Lag = 5				
ssr based F test	F=0.5456	p=0.7418	df_denom=1288	df_num=5
ssr based chi2 test	hi2=2.7518	p=0.7382	df=5	
likelihood ratio test	hi2=2.7518	p=0.7382	df=5	
parameter F test	F=0.5456	p=0.7418	df_denom=1288	df_num=5

TABLE 3.15. Granger causality results of feature Price_oil with five lags, without a strong Granger causality.

Price_oil - Lag = 5				
ssr based F test	F=1.8175	p=0.1065	df_denom=1288	df_num=5
ssr based chi2 test	chi2=9.1659	p=0.1026	df=5	
likelihood ratio test	chi2=9.1334	p=0.1039	df=5	
parameter F test	F=1.8175	p=0.1065	df_denom=1288	df_num=5

TABLE 3.16. Granger causality results of feature Avg_temp with five lags, without a strong Granger causality.

Avg_temp - Lag = 5				
ssr based F test	F=1.4011	p=0.2210	df_denom=1288	df_num=5
ssr based chi2 test	chi2=7.0660	p=0.2158	df=5	
likelihood ratio test	chi2=7.0467	p=0.2172	df=5	
parameter F test	F=1.4011	p=0.2210	df_denom=1288	df_num=5

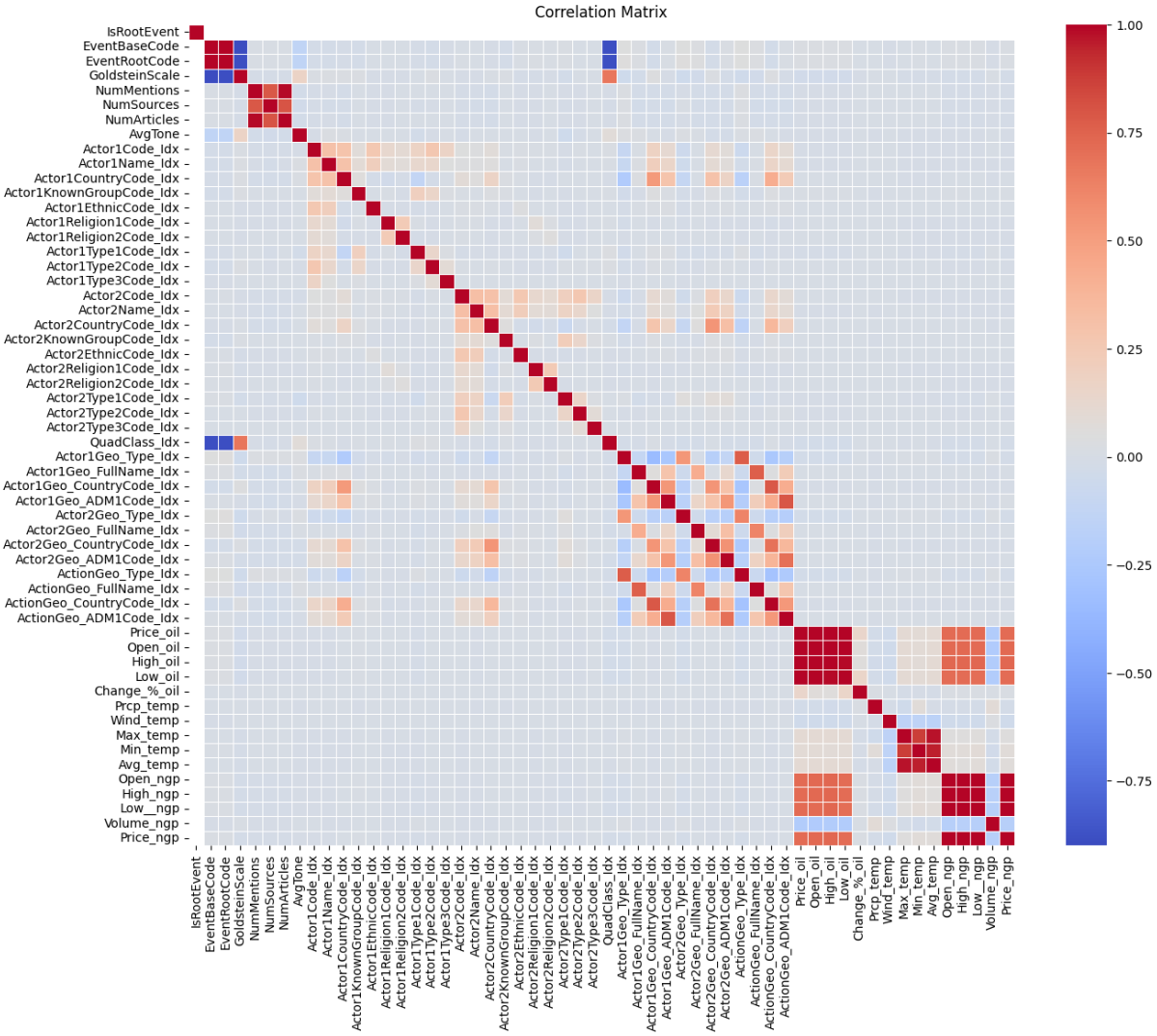

FIGURE 3.9. Pearson correlation heatmap of all features.

TABLE 3.17. Granger causality results of feature Price_oil with ten lags, without a strong Granger causality.

Price_oil - Lag = 10				
ssr based F test	F=1.8175	p=0.1065	df_denom=1288	df_num=10
ssr based chi2 test	chi2=9.1659	p=0.1026	df=10	
likelihood ratio test	chi2=9.1334	p=0.1039	df=10	
parameter F test	F=1.8175	p=0.1065	df_denom=1288	df_num=10

3.2.5. Outliers

The columns Avg_temp, Avg_tone, and Price_oil exhibited outliers, as observed in Figure 5.23, Figure 5.24, and Figure 5.25, respectively. The total count of outliers can be found in Table 3.20. To address these outliers, a mean value was computed using the values before and after the outlier, and the outlier value was replaced with the computed mean. It is important to note that not all outliers were removed, as they provide valuable insights into real-world occurrences.

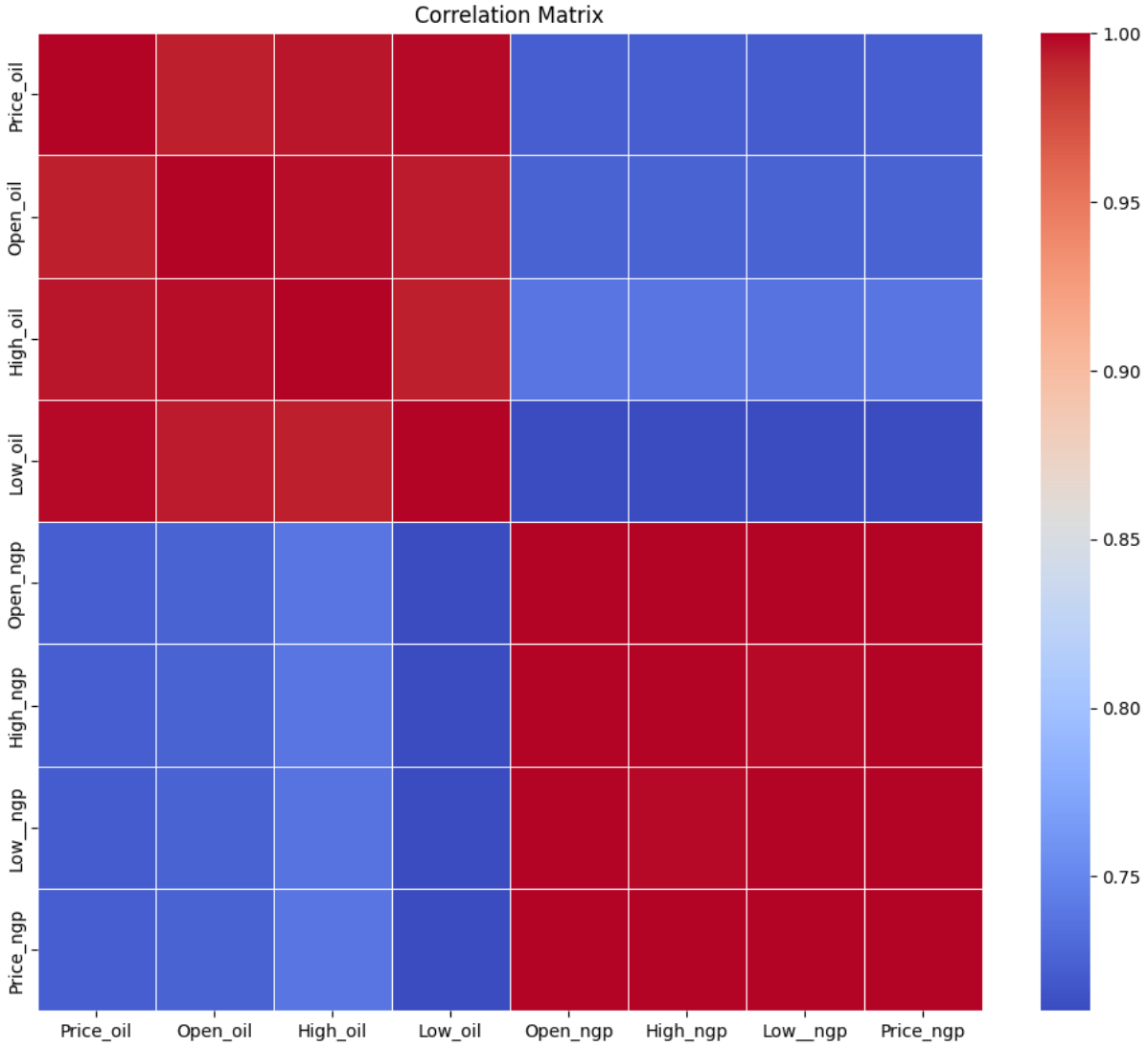

FIGURE 3.10. Pearson correlation heatmap between natural gas features and crude oil features.

TABLE 3.18. Granger causality results of feature Wind_temp with ten lags, without a strong Granger causality.

Wind_temp - Lag = 10				
ssr based F test	F=0.5456	p=0.7418	df_denom=1288	df_num=10
ssr based chi2 test	chi2=2.7518	p=0.7382	df=10	
likelihood ratio test	chi2=2.7488	p=0.7386	df=10	
parameter F test	F=0.5456	p=0.7418	df_denom=1288	df_num=10

3.2.6. Aggregation and Lags

In this preprocessing step, the first action performed was pivoting the variable EventRootCode into separate columns: Threaten, Protest, Exhibit_force_posture, Reduce_relations, Fight, and Use_unconventional_mass_violence. Following this, the news features were aggregated on a daily basis. The columns subjected to summation during the aggregation process were NumMentions, NumSources, NumArticles, QuadClass_Idx, and the columns

TABLE 3.19. Granger causality results of feature Avg_temp with ten lags, without a strong Granger causality.

Avg_temp - Lag = 10				
ssr based F test	F=1.4011	p=0.2210	df_denom=1288	df_num=10
ssr based chi2 test	chi2=7.0660	p=0.2158	df=10	
likelihood ratio test	chi2=7.0467	p=0.2172	df=10	
parameter F test	F=1.4011	p=0.2210	df_denom=1288	df_num=10

TABLE 3.20. Summary of outliers

Column	**q1**	**q3**	**iqr**	**lower_bound**	**upper_bound**	**count_outliers**	**perc_outliers**
AvgTone	-6.404	-2.046	4.358	-12.941	4.491	224133	1.210
Price_oil	52.64	70.73	18.09	25.505	97.865	454154	2.453
Avg_temp	58.0	178.0	120.0	-122.0	358.0	23309	0.126

derived from the EventRootCode column. Additionally, the columns Wind_temp, AvgTone, Price_oil, and Avg_temp were averaged.
To ensure that no calculations were performed on the label column, the minimum value of the natural gas price was selected.
As part of the preprocessing phase, we introduced a lag of 5 and 10 days to all time series. This lagging process entails shifting the values of each column backward in time by the specified number of days. By incorporating these lagged values as additional features, our objective was to capture the temporal relationships in the data. The inclusion of lagged features enables the models to take into account the historical values of each variable when making predictions.

3.2.7. Scaler

In the last preprocessing step, we used a min-max scaler to transform the data. The min-max scaler rescales the values of each feature to a range between zero and one. This normalization technique, implemented using the Scikit-learn library, helps to ensure that all features are on a comparable scale. By applying the min-max scaler, we aimed to facilitate the training process of the models by reducing the impact of varying feature magnitudes (Pedregosa et al., 2011).

CHAPTER 4

Modeling and Performance Evaluation

The deep learning models chosen for this study were the Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN) from Tensorflow library. The selection of these models was complemented with the use of Keras Tuner, a framework for hyperparameter optimization. This allowed us to automatically search for the best hyperparameters for each model.

To assess the performance of the models, several evaluation metrics were employed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE), R-Squared (R2), Explained Variance Score (EVS), and Root Mean Squared Error (RMSE).

The complete modeling workflow is depicted in Figure 4.1, illustrating the sequential steps involved in model development and evaluation.

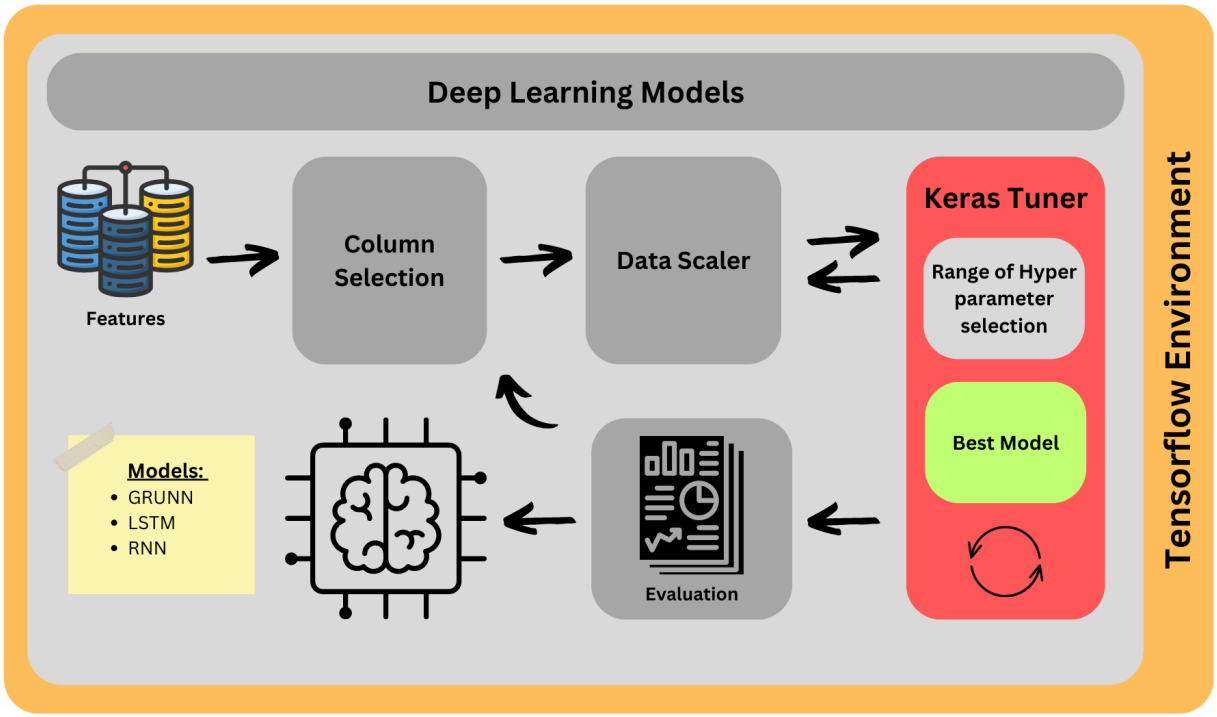

FIGURE 4.1. ETL process for column selection, data scaler, keras tuner, and evaluation of the best model.

4.1. Modeling

4.1.1. Keras Tuner - Hyperparameters

Keras Tuner is a specialized library that optimizes hyperparameters in deep learning models built using Keras. Hyperparameters are external settings that impact the training process, such as the learning rate, number of layers, and units per layer.
Keras Tuner automates the search for the best hyperparameters by utilizing algorithms like random search, grid search, and Bayesian optimization. These strategies efficiently explore the hyperparameter space to identify the combination of settings that yield optimal model performance (O'Malley et al., 2019).
In our study, we specifically used the random search approach to find the best model configuration. We focused on three key hyperparameters: the number of layers, units, and epochs. The epoch value was consistently set to 20, with early stopping applied after 5 epochs without improvement. For the number of layers, we tested a range from one to five, incrementing by one layer per trial. The units were varied between 32 and 512, incrementing by 32 units per trial. Commencing with a modest value like 32 and gradually increasing it by steps of 32 covers a diverse spectrum of options without needing to meticulously test each and every value. This method achieves a harmonious equilibrium between exploring a sufficiently extensive hyperparameter space and preventing excessive consumption of computational resources and time.
In the second Keras Tuner trial, we refined the search based on the best models from the previous attempt. We limited the number of layers to one or two, and the units were constrained to the range of 32 to 512, maintaining the same increment value.

4.1.2. Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are specialized neural networks designed to process sequential data, such as time series or natural language sequences. Unlike traditional neural networks, RNNs have connections that allow them to retain information from previous time steps, enabling them to capture the temporal relationships in sequential data. At the core of an RNN is the hidden state, which acts as a memory of past inputs and is updated at each time step. This hidden state the current input and past information, allowing RNNs to learn and model the patterns and dynamics of sequential data. However, a limitation of RNNs is the vanishing gradient problem, which hampers their ability to capture long-term dependencies. To address this issue, advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) has been developed. These variants incorporate gating mechanisms to alleviate the vanishing gradient problem and improve the RNNs' capacity to capture and remember long-term dependencies (Goodfellow et al., 2016).
Our investigation employed the random search strategy to find the optimal model configuration. We focused on three crucial hyperparameters: the number of layers, units, and

epochs.

4.1.3. Long Short Term Memory (LSTM)

LSTM, a variation of Recurrent Neural Networks (RNNs), addresses the challenge of capturing long-term dependencies in sequential data. By incorporating specialized memory cells, LSTM models can retain information for extended periods, enabling them to effectively learn and represent temporal dependencies. Unlike traditional RNNs, LSTM networks employ a gating mechanism that allows for selective retention and forgetting of information at each time step. This mechanism comprises three gates: the input gate, forget gate, and output gate. The input gate controls the flow of new information, while the forget gate determines which information to discard. The output gate regulates the output of the memory cell. Through dynamic memory updates and control, LSTM networks excel in capturing and preserving long-term dependencies, making them highly suitable for tasks such as speech recognition, language modeling, and machine translation (Goodfellow et al., 2016).

4.1.4. Gated Recurrent Unit Neural Networks (GRUNN)

Gated Recurrent Unit for Neural Networks (GRUNN) is an advanced variation of recurrent neural networks (RNNs) that overcomes limitations found in traditional RNN architectures. GRUNN incorporates gating mechanisms to control the flow of information within the network, allowing it to selectively retain or update information at each time step. This addresses the issue of the vanishing gradient problem, which can hinder training in deep neural networks. By selectively preserving relevant information and discarding irrelevant information, GRUNN models can effectively capture long-term dependencies in sequential data. The gated recurrent units in GRUNN consist of a reset gate and an update gate, which govern the information flow through the network. The reset gate determines what information from previous time steps should be forgotten, while the update gate controls the blending of new input with the existing hidden state. This adaptive gating mechanism empowers GRUNN models to capture intricate temporal patterns and dependencies, making them highly suitable for tasks involving time series forecasting, natural language processing, and speech recognition (Goodfellow et al., 2016).
The GRUNN has its function, which dynamically uses the Keras tuner to find the best layer and unit number. The compile is set with Adam optimizer, loss is MSE, and the metric is RMSE.

4.2. Performance Evaluation

In each model, we developed a baseline with either 5 or 10 lagged values, consisting solely of the natural gas price feature. Subsequently, we conducted model training using all available features listed in Table 4.1.

With a test dataset of 258 data points, we predicted the same size time daily.
Abbreviations for the different feature combinations can be found in Table 4.3. The optimal hyperparameters for each trained model are presented in Table 4.2, while the performance of each model is ranked in Table 4.4.

TABLE 4.1. All features applied to the model.

Column Name	**Descritption**
GoldsteinScale	Numeric measure indicating the level of conflict or cooperation in political events.
NumMentions	Represents the number of mentions of an event in various sources
NumMentions	Number of mentions of the event
NumArticles	Number of articles related to the event
QuadClass_Idx	Index for the QuadClass category
AvgTone	Average tone of the event
Price_oil	Crude oil price
Avg_temp	Average temperature
Threaten	Cameo code to threaten
Protest	Cameo code to protest
Exhibit_force_posture	Cameo code to exhibit force posture
Reduce_relations	Cameo code to reduce relations
Fight	Cameo code to fight
Use_unconvetional_mass_violence	Cameo code to use unconventional mass violence
Price_ngp	Natural gas price

4.2.1. Best model: Recurrent Neural Networks (RNN)

Among all the models trained, the RNN model outperformed the others. This model incorporated the features of natural gas price, crude oil price, and average tone of the extracted news. With ten lagged values, one layer, and 224 units, the RNN model achieved an RMSE of 11.925 euros. The prediction made by the RNN model is illustrated in Figure 4.2 in the form of a time-series plot.

4.2.2. Best model: Long Short Term Memory (LSTM)

The LSTM model, which utilized the same set of features as the best model (natural gas price, crude oil price, and average tone), ranked fourth among the best models. With ten lagged values, one layer, and 320 units, the LSTM model achieved an RMSE of 11.954. The prediction made by the LSTM model is depicted in Figure 4.4 as a time-series plot.

4.2.3. Best model: Gated Recurrent Unit Neural Networks (GRUNN)

The GRUNN model, utilizing natural gas price and crude oil price as features, secured the second position among the best models. With five lagged values, one layer, and 480 units, the GRUNN model obtained an RMSE of 11.935. The prediction generated by the GRUNN model is illustrated in Figure 4.3 in the form of a time-series plot.

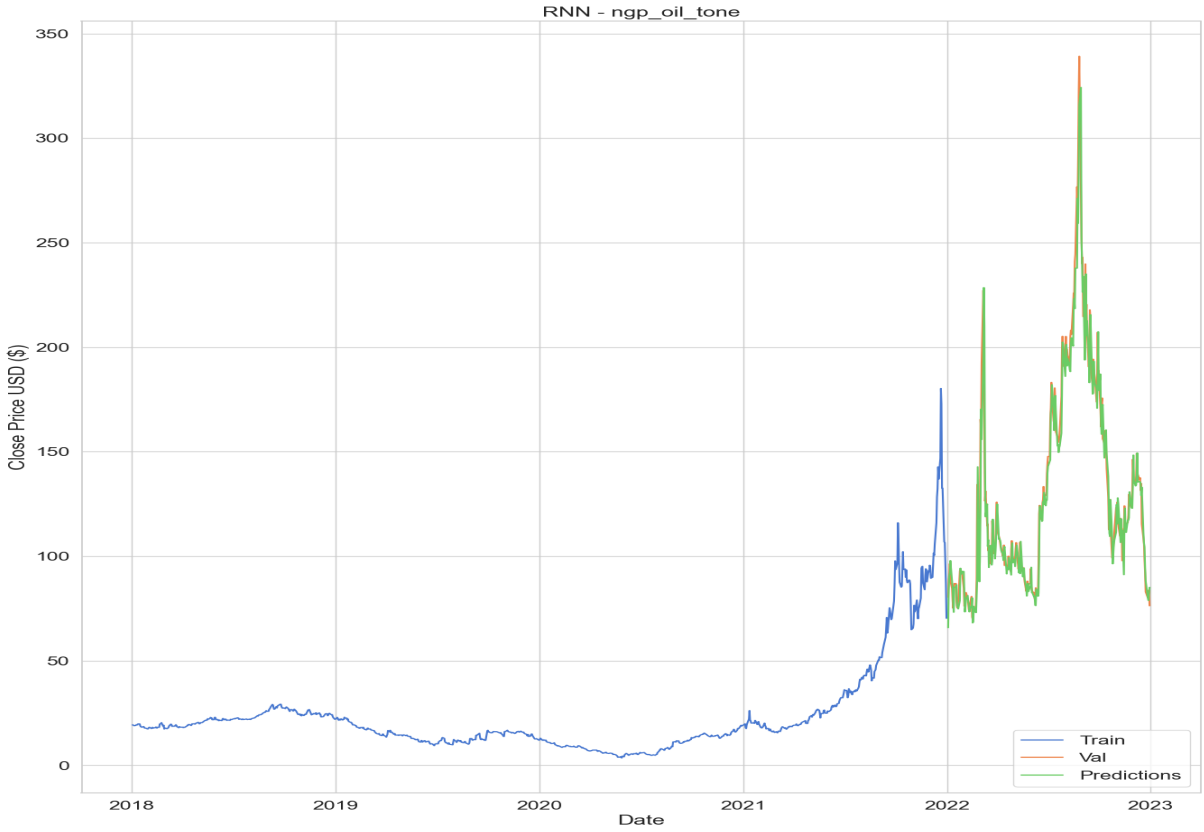

FIGURE 4.2. Prediction plot curve of RNN model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

4.2.4. Comparative Analysis

The literature review covered various prediction horizons, utilizing different models, features, and preprocessing methods. Our comparison focuses on single and hybrid models for predicting daily natural gas prices, similar to other studies.

Among the single models reviewed, the study by Al-Sharoot and Alramadhan (2019) achieved the best performance using Auto-regressive moving average (ARMA) and Group Method of Data Handling (GMDH) models with 527 observations from August 29, 2016, to August 27, 2018, without exogenous variables. Their mean squared error (MSE) was 0.0214.

Another notable single model from Qin et al. (2019) employed Ensemble Empirical Mode Decomposition (EEMD) and Local Linear Prediction (LLP) with 1678 observations from January 4, 2010, to August 15, 2016, also without exogenous variables. Their root mean squared error (RMSE) was 0.035.

Our study's best-performing model was a Recurrent Neural Network (RNN) with 10 lags, incorporating natural gas price, crude oil price, and average tone as exogenous variables. We used 1292 observations from January 2, 2018, to December 30, 2022, and achieved an RMSE of 11.925.

However, it is important to note that our results were not as favorable as the best results in the literature. This disparity is primarily attributed to the complexity of the prediction

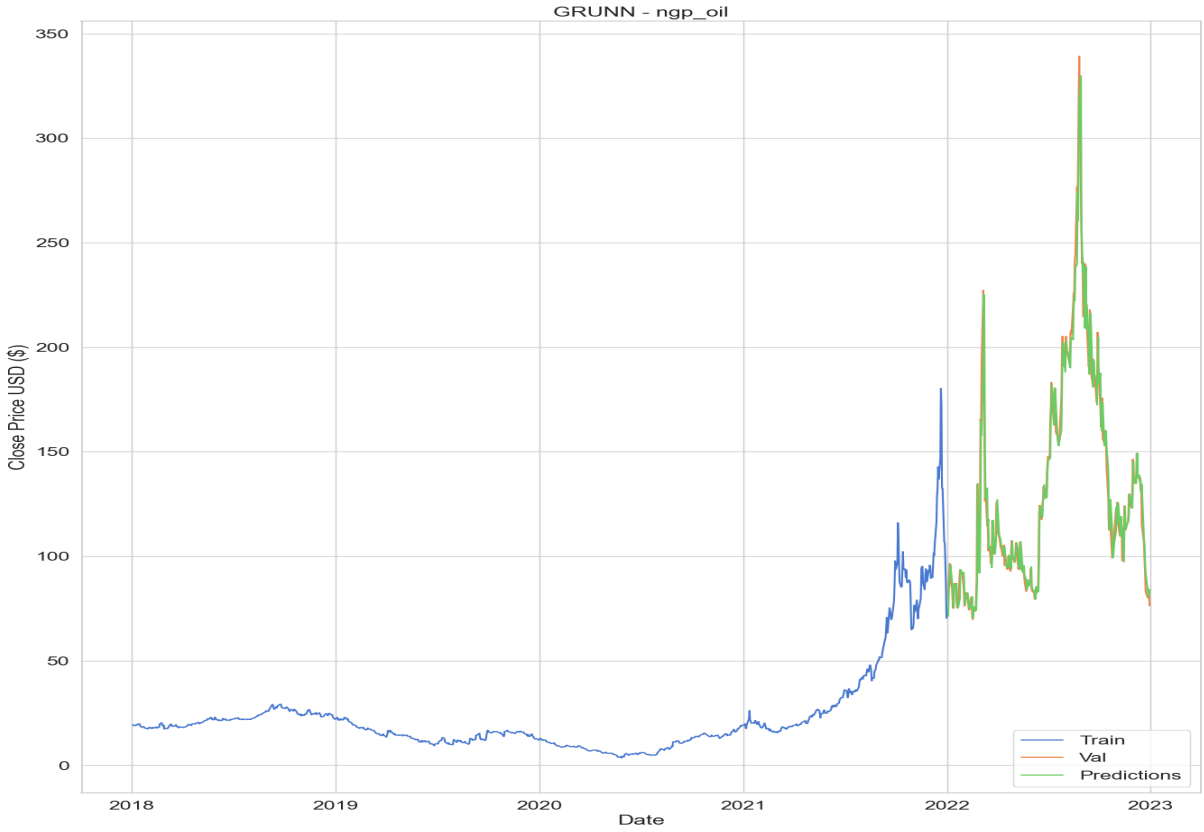

FIGURE 4.3. Prediction plot curve of GRUNN model with natural gas price and crude oil price as features, and lag equal to 5.

period chosen, which coincided with the Russo-Ukrainian War. The geopolitical situation during this period likely contributed to the model's reduced performance.

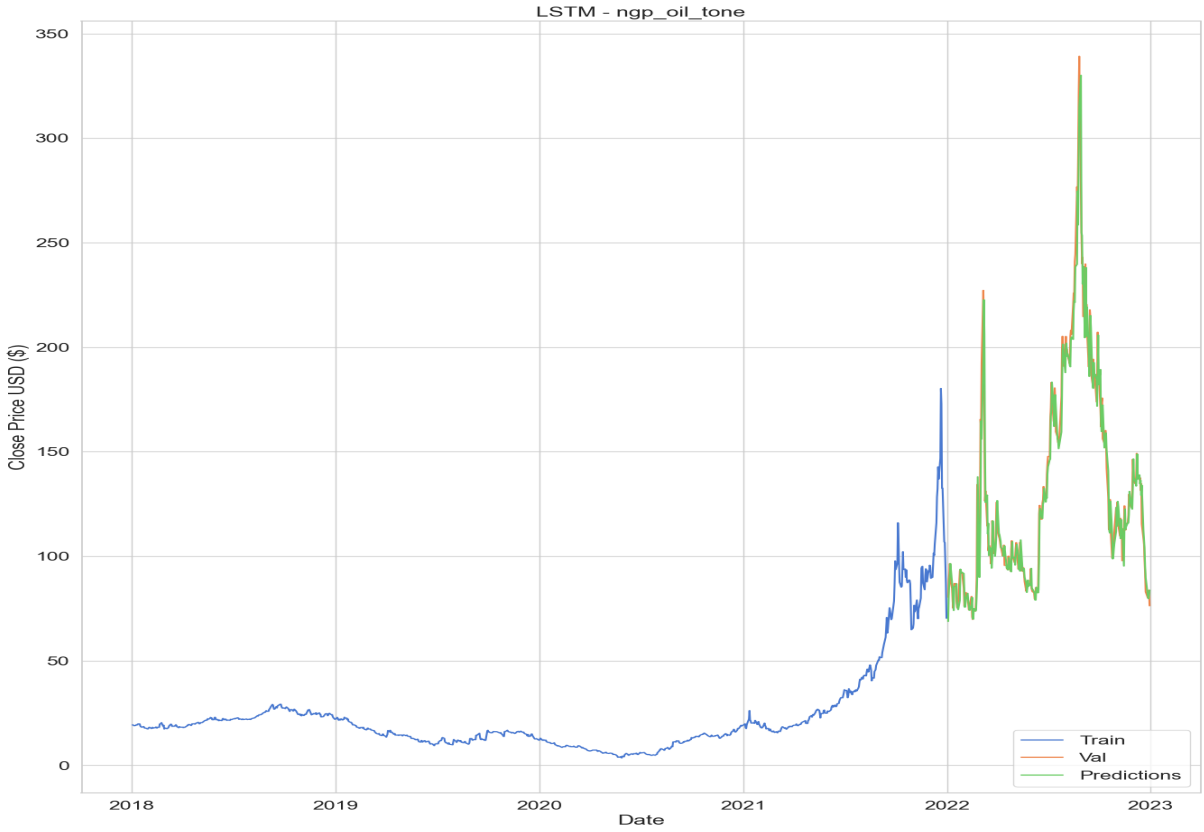

FIGURE 4.4. Prediction plot curve of LSTM model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

TABLE 4.2. All models with the best hyperparameter selected by Keras Tuner.

Model Name	Features Code	RMSE	Best Num Layers	Best Num Units
RNN	ngp_oil_tone	11,925	1	224
GRUNN	ngp_oil	11,935	1	480
RNN	ngp_oil	11,942	1	96
LSTM	ngp_oil_tone	11,954	1	320
LSTM	ngp_oil	11,962	1	96
RNN	base_line	11,964	1	192
GRUNN	base_line	11,979	1	224
GRUNN	ngp_oil	11,996	1	32
GRUNN	all_features	12,001	1	32
RNN	base_line	12,010	1	32
GRUNN	base_line	12,023	1	32
GRUNN	ngp_oil	12,043	1	32
LSTM	all_features	12,045	1	384
RNN	ngp_oil	12,060	1	160
GRUNN	all_features	12,066	1	128
RNN	ngp_tone	12,077	1	32
GRUNN	ngp_oil	12,087	2	32
LSTM	ngp_oil_tone	12,090	1	128
GRUNN	all_features	12,098	1	160
GRUNN	ngp_oil_tone	12,116	2	32
RNN	all_features	12,124	1	320
GRUNN	ngp_oil_tone	12,125	1	96
LSTM	ngp_oil_tone	12,129	1	192
RNN	ngp_oil_tone	12,141	1	96
LSTM	ngp_oil_tone	12,142	1	256
RNN	ngp_oil	12,145	1	96
GRUNN	ngp_tone	12,161	1	128
GRUNN	ngp_tone	12,165	1	160
RNN	ngp_tone	12,178	1	96
RNN	all_features	12,206	1	128
LSTM	all_features	12,244	1	384
GRUNN	ngp_oil_tone	12,250	2	64
LSTM	base_line	12,258	1	32
GRUNN	all_features	12,268	1	192
LSTM	ngp_tone	12,279	1	64
GRUNN	ngp_oil_tone	12,282	2	64
RNN	ngp_tone	12,293	1	480
RNN	ngp_oil_tone	12,302	1	384
LSTM	ngp_oil	12,323	1	96
LSTM	ngp_tone	12,338	1	64
LSTM	base_line	12,339	1	224
RNN	ngp_tone	12,340	2	64
LSTM	ngp_oil	12,341	1	160
RNN	ngp_oil	12,363	1	32
RNN	ngp_oil_tone	12,369	1	32
LSTM	ngp_oil	12,384	1	320
LSTM	ngp_tone	12,483	1	160
RNN	all_features	12,520	1	64
GRUNN	ngp_tone	12,555	1	480
LSTM	all_features	12,606	1	416
RNN	all_features	12,617	1	416
LSTM	ngp_tone	12,717	1	32
GRUNN	ngp_tone	13,286	3	512

TABLE 4.3. Features code description.

Code	**Description**
ngp_oil_tone	Natural gas price, crude oil price, and average tone
ngp_oil	Natural gas price and crude oil price
base_line	Natural gas price
all_features	All features
ngp_tone	Natural gas price and average tone

TABLE 4.4. Result for each model executed with features applied, lags used, MAE, MSE, MAPE, MedAE, R2, EVS, and RMSE performance metrics.

Model Name	Features Code	Lags	MAE	MSE	MAPE	MedAE	R2	EVS	RMSE
RNN	ngp_oil_tone	10	7,846	142,215	5,759	4,759	0,946	0,946	11,925
GRUNN	ngp_oil	5	7,811	142,438	5,712	4,470	0,946	0,946	11,935
RNN	ngp_oil	10	7,955	142,604	5,795	4,925	0,946	0,947	11,942
LSTM	ngp_oil_tone	10	7,902	142,907	5,777	4,609	0,946	0,946	11,954
LSTM	ngp_oil	10	7,956	143,096	5,836	4,791	0,946	0,946	11,962
RNN	base_line	5	7,881	143,142	5,749	4,989	0,946	0,946	11,964
GRUNN	base_line	5	7,903	143,498	5,732	4,771	0,946	0,946	11,979
GRUNN	ngp_oil	5	7,838	143,903	5,705	4,780	0,946	0,946	11,996
GRUNN	all_features	10	7,941	144,028	5,760	4,957	0,946	0,946	12,001
RNN	base_line	10	8,028	144,248	5,860	4,859	0,946	0,946	12,010
GRUNN	base_line	10	7,987	144,543	5,770	5,030	0,946	0,947	12,023
GRUNN	ngp_oil	10	7,890	145,023	5,763	4,802	0,945	0,946	12,043
LSTM	all_features	10	8,005	145,089	5,896	4,964	0,945	0,945	12,045
RNN	ngp_oil	5	7,940	145,433	5,744	4,961	0,945	0,946	12,060
GRUNN	all_features	5	7,851	145,579	5,814	5,044	0,945	0,945	12,066
RNN	ngp_tone	10	8,139	145,842	5,918	5,222	0,945	0,947	12,077
GRUNN	ngp_oil	10	8,071	146,096	5,795	5,006	0,945	0,946	12,087
LSTM	ngp_oil_tone	5	7,968	146,174	5,743	4,711	0,945	0,946	12,090
GRUNN	all_features	10	7,859	146,367	5,814	4,869	0,945	0,945	12,098
GRUNN	ngp_oil_tone	10	8,053	146,788	5,834	4,928	0,945	0,946	12,116
RNN	all_features	5	8,046	146,984	5,857	4,786	0,945	0,945	12,124
GRUNN	ngp_oil_tone	5	7,901	147,023	5,722	4,785	0,945	0,945	12,125
LSTM	ngp_oil_tone	10	8,027	147,106	5,786	5,038	0,945	0,945	12,129
RNN	ngp_oil_tone	5	7,965	147,411	5,826	5,232	0,944	0,945	12,141
LSTM	ngp_oil_tone	5	7,960	147,423	5,756	4,870	0,944	0,945	12,142
RNN	ngp_oil	10	8,019	147,511	5,852	4,585	0,944	0,945	12,145
GRUNN	ngp_tone	10	8,132	147,880	5,857	5,268	0,944	0,946	12,161
GRUNN	ngp_tone	5	8,061	147,980	5,830	5,080	0,944	0,946	12,165
RNN	ngp_tone	5	8,111	148,306	5,917	5,085	0,944	0,945	12,178
RNN	all_features	10	8,196	148,987	5,925	5,143	0,944	0,945	12,206
LSTM	all_features	10	8,143	149,917	6,098	5,279	0,943	0,944	12,244
GRUNN	ngp_oil_tone	5	8,112	150,070	5,782	4,965	0,943	0,945	12,250
LSTM	base_line	5	8,231	150,256	5,883	5,011	0,943	0,945	12,258
GRUNN	all_features	5	8,106	150,505	6,006	5,217	0,943	0,944	12,268
LSTM	ngp_tone	5	8,259	150,777	5,892	5,405	0,943	0,945	12,279
GRUNN	ngp_oil_tone	10	8,319	150,853	5,973	5,515	0,943	0,946	12,282
RNN	ngp_tone	5	8,061	151,108	5,928	5,040	0,943	0,943	12,293
RNN	ngp_oil_tone	5	8,125	151,348	5,947	5,258	0,943	0,944	12,302
LSTM	ngp_oil	5	8,182	151,856	5,836	4,919	0,943	0,945	12,323
LSTM	ngp_tone	10	8,293	152,219	5,897	5,398	0,943	0,945	12,338
LSTM	base_line	10	8,271	152,239	5,943	5,230	0,943	0,945	12,339
RNN	ngp_tone	10	8,343	152,281	6,017	5,541	0,943	0,945	12,340
LSTM	ngp_oil	5	8,189	152,294	5,848	4,932	0,943	0,944	12,341
RNN	ngp_oil	5	8,226	152,839	5,886	5,420	0,942	0,945	12,363
RNN	ngp_oil_tone	10	8,262	152,999	6,021	5,089	0,942	0,945	12,369
LSTM	ngp_oil	10	8,244	153,367	5,887	4,949	0,942	0,944	12,384
LSTM	ngp_tone	10	8,388	155,832	5,953	5,604	0,941	0,945	12,483
RNN	all_features	5	8,414	156,747	6,014	5,639	0,941	0,944	12,520
GRUNN	ngp_tone	5	8,443	157,629	5,970	5,639	0,941	0,945	12,555
LSTM	all_features	5	8,229	158,906	6,101	5,023	0,940	0,940	12,606
RNN	all_features	10	8,557	159,199	6,246	5,302	0,940	0,941	12,617
LSTM	ngp_tone	5	8,573	161,718	6,010	5,560	0,939	0,944	12,717
LSTM	all_features	5	9,072	171,336	6,681	6,078	0,935	0,936	13,090
GRUNN	ngp_tone	10	9,162	176,510	6,437	6,370	0,933	0,942	13,286

CHAPTER 5

Conclusions

Throughout our study, we explored various features for predicting natural gas prices, considering different scenarios such as using only natural gas prices or combinations with crude oil prices and average tone from news sources. Among the top five performing models, incorporating crude oil price as an exogenous variable significantly enhanced the predictive accuracy, consistent with previous findings (Li et al., 2021). Surprisingly, the best model emerged when we included the feature of average tone in our input data, resulting in a 7.82% improvement compared to the same model without it.

By filtering the extensive GDELT Big Data using specific Cameo codes (see Table 3.2), we achieved improved performance during conflict times. Contrary to the notion proposed by Čeperić et al. (2017), who suggested "less data is better" for short-term prediction, we found that using the relevant variables is crucial for accurate predictions.

Our optimization process, employing the Random Search optimizer, demonstrated an efficient selection of hyperparameters and facilitated in-depth analysis of each interaction. Despite these efforts, the best model's RMSE of 11.925 fell short when compared to the literature, mainly due to an abrupt change in values caused by the Russo-Ukrainian War and Europe's high dependence on Russian natural gas prices.

Nonetheless, our study's best-performing model remained the Recurrent Neural Networks (RNN) with 10 lags, incorporating natural gas price, crude oil price, and average tone as exogenous variables. The entire observation has 1292 data points from January 2, 2018, to December 30, 2022, we achieved promising results, although the worst model, a GRUNN with natural gas price and average tone, had an RMSE of 13.286 (see Table 4.4).

In conclusion, this study highlights the negative impact on the performance of natural gas price models during war times and emphasizes the positive influence of specific Cameo codes on model results. Moreover, the findings reinforce the strong correlation and causation between crude oil and natural gas prices, contributing to improved model performance.

To enhance the methodology, implementing version management tools like MLFlow could have been beneficial. Additionally, further exploration of different Cameo codes and exhaustive testing of all possible feature combinations extracted from GDELT could have been conducted.

References

Abrishami, H., & Varahrami, V. (2011). Different methods for gas price forecasting. *Cuadernos de Economia*, *34*, 137–144. https://doi.org/10.1016/S0210-0266(11)70013-9

Alamro, R., McCarren, A., & Al-Rasheed, A. (2019). Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling. *Communications in Computer and Information Science*, *1097 CCIS*, 317–328. https://doi.org/10.1007/978-3-030-36365-9_26

Al-Sharoot, M. H., & Alramadhan, O. M. (2019). Forecasting the gas prices in investing.com's weekly economic data table using linear and non-linear arma-garch models for period 2016-2018. *AIP Conference Proceedings*, *2096*. https://doi.org/10.1063/1.5097818

Apache Spark. (Accessed on July 19, 2023). pyspark.ml.feature.StringIndexer.

Azadeh, A., Sheikhalishahi, M., & Shahmiri, S. (2012). A hybrid neuro-fuzzy simulation approach for improvement of natural gas price forecasting in industrial sectors with vague indicators. *International Journal of Advanced Manufacturing Technology*, *62*, 15–33. https://doi.org/10.1007/s00170-011-3804-6

Berrisch, J., & Ziel, F. (2022). Distributional modeling and forecasting of natural gas prices. *Journal of Forecasting*, *41*, 1065–1086. https://doi.org/10.1002/for.2853

Bodas-Sagi, D., & Labeaga, J. (2016). Using gdelt data to evaluate the con dence on the spanish government energy policy. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*, 38. https://doi.org/10.9781/ijimai.2016.366

Bourgeois, D., Rappaz, J., & Aberer, K. (2018). Selection bias in news coverage: Learning it, fighting it. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 535–543. https://doi.org/10.1145/3184558.3188724

Čeperić, E., Žiković, S., & Čeperić, V. (2017). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. *Energy*, *140*, 893–900. https://doi.org/10.1016/j.energy.2017.09.026

Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (gru) neural networks. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. https://doi.org/10.1109/MWSCAS.2017.8053243

Downey, A. B. (Accessed on July 19, 2023). *Think stats: Exploratory data analysis in python*. https://greenteapress.com/wp/think-stats-2e/

Economic bulletin issue 4, (2022, June). European Central Bank. https://www.ecb.europa.eu/pub/pdf/ecbu/eb202204.en.pdf

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, *44*, 408–421. https://doi.org/10.1177/0165551517698298

Event Data Project, Department of Political Science, Pennsylvania State University. (March 2012). *CAMEO Conflict and Mediation Event Observations Event and Actor Codebook*. http://eventdata.psu.edu/

Fabian, J., Wingrove, J., & Krukowska, E. (2022, May). *U.s., eu reach lng supply deal to cut dependence on russia*. https://www.bloomberg.com/news/articles/2022-03-25/u-s-and-eu-reach-energy-supply-deal-to-cut-dependence-on-russia?leadSource=uverify%5C%20wall

Galla, D., & Burke, J. (2018). Predicting social unrest using gdelt. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10935 LNAI*, 103–116. https://doi.org/10.1007/978-3-319-96133-0_8

GDELT Project. (n.d.). *The gdelt project*. https://www.gdeltproject.org/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Google Cloud Platform Blog. (2014). *World's largest event dataset now publicly available in bigquery*. https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html

Halser, C., & Paraschiv, F. (2022). Pathways to overcoming natural gas dependency on russia—the german case. *Energies*, *15*. https://doi.org/10.3390/en15144939

Hamie, H., Auer, H., & Hoayek, A. (2020). Modeling post-liberalized european gas market concentration-a game theory perspective. https://doi.org/10.3390/forecast

Hammond, J., & Weidmann, N. B. (2014). Using machine-coded event data for the micro-level study of political violence. *Research and Politics*, *1*. https://doi.org/10.1177/2053168014539924

Heather, P. (2021). *European traded gas hubs : German hubs about to merge*.

Hu, Y., & Trafalis, T. B. (2011). *New kernel methods for asset pricing: Application to natural gas price prediction new kernel methods for asset pricing* (2).

Investing.com. (Accessed on July 19, 2023). *Crude Oil Historical Data*. https://www.investing.com/commodities/crude-oil-historical-data

Jianwei, E., Ye, J., He, L., & Jin, H. (2019). Energy price prediction based on independent component analysis and gated recurrent unit neural network. *Energy*, *189*. https://doi.org/10.1016/j.energy.2019.116278

Jin, J., & Kim, J. (2015). Forecasting natural gas prices using wavelets, time series, and artificial neural networks. *PLoS ONE*, *10*. https://doi.org/10.1371/journal.pone.0142064

Kaufmann, R. K., & Connelly, C. (2020). Oil price regimes and their role in price diversions from market fundamentals. *Nature Energy*, *5*, 141–149. https://doi.org/10.1038/s41560-020-0549-1

Kwak, H., & An, J. (2014). A first look at global news coverage of disasters by using the gdelt dataset. *6th International Conference, SocInfo*.

Li, J., Wu, Q., Tian, Y., & Fan, L. (2021). Monthly henry hub natural gas spot prices forecasting using variational mode decomposition and deep belief network. *Energy*, *227*. https://doi.org/10.1016/j.energy.2021.120478

Moting, S., Zongyi, Z., Ye, Z., & Donglan, Z. (2019a). Data-driven natural gas spot price forecasting with least squares regression boosting algorithm. *Energies*, *12*. https://doi.org/10.3390/en12061094

Moting, S., Zongyi, Z., Ye, Z., Donglan, Z., & Wenying, W. (2019). Data driven natural gas spot price prediction models using machine learning methods. *Energies*, *12*. https://doi.org/10.3390/en12091680

Naderi, M., Khamehchi, E., & Karimi, B. (2019). Novel statistical forecasting models for crude oil price, gas price, and interest rate based on meta-heuristic bat algorithm. *Journal of Petroleum Science and Engineering*, *172*, 13–22. https://doi.org/10.1016/j.petrol.2018.09.031

Naderi, M., Khamehchi, E., & Karimi, B. (2021, February). *Energy price prediction using data-driven models: A decade review*. https://doi.org/10.1016/j.cosrev.2020.100356

NCEI. (Accessed on July 19, 2023). *National Centers for Environmental Information (NCEI) - Daily Summaries*. https://www.ncei.noaa.gov/access/search/data-search/daily-summaries?dataTypes=TAVG&bbox=54.304,-3.406,41.919,20.984&startDate=2017-10-23T00:00:00&endDate=2023-02-07T23:59:59&stations=GMW00035032

Nguyen, H. T., & Nabney, I. T. (2010). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. *Energy*, *35*, 3674–3685. https://doi.org/10.1016/j.energy.2010.05.013

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). Kerastuner.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Psaropoulos, J. (2022, August). *Timeline: Six months of russia's war in ukraine*. https://www.aljazeera.com/news/2022/8/24/timeline-six-months-of-russias-war-in-ukraine

Qiao, F., Li, P., Deng, J., Ding, Z., & Wang, H. (2015). Graph-based method for detecting occupy protest events using gdelt dataset. *Proceedings - 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015*, 164–168. https://doi.org/10.1109/CyberC.2015.77

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society, 2017*. https://doi.org/10.1155/2017/8180272

Qin, Q., Xie, K., He, H., Li, L., Chu, X., Wei, Y. M., & Wu, T. (2019). An effective and robust decomposition-ensemble energy price forecasting paradigm with local linear prediction. *Energy Economics*, *83*, 402–414. https://doi.org/10.1016/j.eneco.2019.07.026

Refilling gas storage for next winter. (2022, March). European Comission. https://ec.europa.eu/commission/presscorner/detail/en/fs_22_1938

Salehnia, N., Falahi, M. A., Seifi, A., & Adeli, M. H. M. (2013). Forecasting natural gas spot prices with nonlinear modeling using gamma test analysis. *Journal of Natural Gas Science and Engineering*, *14*, 238–249. https://doi.org/10.1016/j.jngse.2013.07.002

Siddiqui, A. W. (2019). Predicting natural gas spot prices using artificial neural network. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. https://doi.org/10.1109/CAIS.2019.8769586

Tang, L., Wu, Y., & Yu, L. (2018). A randomized-algorithm-based decomposition-ensemble learning methodology for energy price forecasting. *Energy*, *157*, 526–538. https://doi.org/10.1016/j.energy.2018.05.146

Tang, Y., Wang, M., Wang, Q., Wang, Z., & Xu, W. (2019). Natural gas price prediction with big data. *2019 IEEE International Conference on Big Data (Big Data)*.

Thakur, A., Kumar, S., & Tiwari, A. (2015). *Hybrid model of gas price prediction using moving average and neural network*.

Tilly, S., Ebner, M., & Livan, G. (2020). Macroeconomic forecasting through news, emotions and narrative. https://doi.org/10.1016/j.eswa.2021.114760

van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman; Hall/CRC.

Viacaba, A., Poursaeidi, M. H., & Kundakcioglu, O. E. (2012). *Natural gas price forecasting via selective support vector regression*. https://www.researchgate.net/publication/289446769

Wang, J., Lei, C., & Guo, M. (2020). Daily natural gas price forecasting by a weighted hybrid data-driven model. *Journal of Petroleum Science and Engineering*, *192*. https://doi.org/10.1016/j.petrol.2020.107240

Wang, J., Cao, J., Yuan, S., & Cheng, M. (2021). Short-term forecasting of natural gas prices by using a novel hybrid method based on a combination of the ceemdan-se-and the pso-als-optimized gru network. *Energy*, *233*. https://doi.org/10.1016/j.energy.2021.121082

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016, September). *Growing pains for global monitoring of societal events*. https://doi.org/10.1126/science.aaf6758

Zhang, C., Xiao, C., & Liu, H. (2019). Spatial big data analysis of political risks along the belt and road. *Sustainability (Switzerland)*, *11*. https://doi.org/10.3390/su11082216

Appendix

TABLE 5.1. Data size used to forecast natural gas price

References	Daily	Weekly	Monthly	Yearly
Naderi et al., 2019	230			
Al-Sharoot and Alramadhan, 2019	527			
J. Wang et al., 2021	1042			
Thakur et al., 2015	1146			
Hu and Trafalis, 2011	1457			
Qin et al., 2019	1678			
Čeperić et al., 2017	1800	260		
Abrishami and Varahrami, 2011	1825			
Berrisch and Ziel, 2022	2068		2091	
Y. Tang et al., 2019	2372			
Salehnia et al., 2013	3803	792	182	
Moting et al., 2019a	4260	886	204	
L. Tang et al., 2018	4873			
Siddiqui, 2019	5470			
Nguyen and Nabney, 2010			130	
Moting et al., 2019			240	
Azadeh et al., 2012				40
Jianwei et al., 2019			420	
Jin and Kim, 2015		726		

TABLE 5.2. Category columns related to actor number one in the articles.

Column	**Description**
Actor1Code	Code representing the primary actor involved in the event
Actor1Name	Name of the primary actor involved in the event
Actor1CountryCode	Country code associated with the primary actor
Actor1KnownGroupCode	Code representing a known group associated with the primary actor
Actor1EthnicCode	Code representing the ethnic group associated with the primary actor
Actor1Religion1Code	Code representing the primary religion associated with the primary actor
Actor1Religion2Code	Code representing the secondary religion associated with the primary actor
Actor1Type1Code	Code representing the primary type of the primary actor
Actor1Type2Code	Code representing the secondary type 1 of the primary actor
Actor1Type3Code	Code representing the secondary type 2 of the primary actor
QuadClass	Code representing the high-level category of the event
Actor1Geo_Type	Type of the geographic location associated with the primary actor
Actor1Geo_FullName	Full name of the geographic location associated with the primary actor
Actor1Geo_CountryCode	Country code associated with the geographic location of the primary actor
Actor1Geo_ADM1Code	ADM1 code associated with the geographic location of the primary actor

TABLE 5.3. Category columns related to actor number two in the articles.

Column	Description
Actor2Code	Code representing the secondary actor involved in the event
Actor2Name	Name of the secondary actor involved in the event
Actor2CountryCode	Country code associated with the secondary actor
Actor2KnownGroupCode	Code representing a known group associated with the secondary actor
Actor2EthnicCode	Code representing the ethnic group associated with the secondary actor
Actor2Religion1Code	Code representing the primary religion associated with the secondary actor
Actor2Religion2Code	Code representing the secondary religion associated with the secondary actor
Actor2Type1Code	Code representing the primary type of the secondary actor
Actor2Type2Code	Code representing the secondary type 1 of the secondary actor
Actor2Type3Code	Code representing the secondary type 2 of the secondary actor
Actor2Geo_Type	Type of the geographic location associated with the secondary actor
Actor2Geo_FullName	Full name of the geographic location associated with the secondary actor
Actor2Geo_CountryCode	Country code associated with the geographic location of the secondary actor
Actor2Geo_ADM1Code	ADM1 code associated with the geographic location of the secondary actor

TABLE 5.4. Category geographic columns.

Column	Description
ActionGeo_Type	Type of the geographic location associated with the action
ActionGeo_FullName	Full name of the geographic location associated with the action
ActionGeo_CountryCode	Country code associated with the geographic location of the action
ActionGeo_ADM1Code	ADM1 code associated with the geographic location of the action

TABLE 5.5. Integer and decimal columns of GDELT.

Column Name	**Description**
GKGRECORDID	Unique identifier for the GKG record
Date	Date of the event
EventBaseCode	Code representing the base event category
EventRootCode	Code representing the root event category
Actor1Geo_Lat	Latitude of the geographic location for Actor 1
Actor1Geo_Long	Longitude of the geographic location for Actor 1
Actor2Geo_Lat	Latitude of the geographic location for Actor 2
Actor2Geo_Long	Longitude of the geographic location for Actor 2
ActionGeo_Lat	Latitude of the geographic location for the action
ActionGeo_Long	Longitude of the geographic location for the action
NumMentions	Represents the number of mentions of an event in various sources
NumSources	Number of sources reporting the event
NumArticles	Number of articles related to the event
AvgTone	Average tone of the event
GoldsteinScale	Numeric measure indicating a level of conflict or cooperation in political events.

TABLE 5.6. Pearson correlation with values between -0.7 and 0.7 with natural gas price.

Variable	**Correlation**
Price_ngp	1.000000
Price_oil	0.722026
Open_oil	0.725523
High_oil	0.737906
Low_oil	0.710763
Open_ngp	0.999004
High_ngp	0.999524
Low__ngp	0.999375

TABLE 5.7. Pearson correlation with negative values with natural gas price.

Variable	**Correlation**
Volume_ngp	-0.189236
Wind_temp	-0.040241
Actor2Geo_Type_Idx	-0.023385
Actor1Geo_Type_Idx	-0.020253
QuadClass_Idx	-0.019151
ActionGeo_Type_Idx	-0.018957
GoldsteinScale	-0.017096
Prcp_temp	-0.016933
NumSources	-0.015359
NumArticles	-0.013915
NumMentions	-0.013703
Actor2Type1Code_Idx	-0.010697
Actor1Type1Code_Idx	-0.010565
Actor2Name_Idx	-0.008764
Actor1Name_Idx	-0.008247
Actor2Type2Code_Idx	-0.006322
Actor1Religion1Code_Idx	-0.004739
Actor2Code_Idx	-0.004414
AvgTone	-0.004131
ActionGeo_CountryCode_Idx	-0.003934
Actor1Code_Idx	-0.003631
Actor2Religion1Code_Idx	-0.003412
Actor1Type2Code_Idx	-0.003282
Actor2Geo_CountryCode_Idx	-0.003195
Actor1Geo_CountryCode_Idx	-0.002933
Actor1Religion2Code_Idx	-0.002894
Actor2Religion2Code_Idx	-0.002311
Actor1Type3Code_Idx	-0.001329
Actor2Type3Code_Idx	-0.001270
Actor1EthnicCode_Idx	-0.001230
Actor2EthnicCode_Idx	-0.001021
Actor2CountryCode_Idx	-0.000579
Actor1CountryCode_Idx	-0.000467

TABLE 5.8. Pearson correlation with positive values with natural gas price.

Variable	Correlation
ActionGeo_FullName_Idx	0.000369
Actor2Geo_FullName_Idx	0.000370
Actor1Geo_FullName_Idx	0.000572
Actor2KnownGroupCode_Idx	0.003669
ActionGeo_ADM1Code_Idx	0.005056
Actor2Geo_ADM1Code_Idx	0.005138
Actor1Geo_ADM1Code_Idx	0.005856
Actor1KnownGroupCode_Idx	0.005979
EventRootCode	0.013956
EventBaseCode	0.015387
Change_%_oil	0.017322
Max_temp	0.055238
Avg_temp	0.068243
Min_temp	0.078298
Low_oil	0.710763
Price_oil	0.722026
Open_oil	0.725523
High_oil	0.737906
Open_ngp	0.999004
Low__ngp	0.999375
High_ngp	0.999524
Price_ngp	1.000000
IsRootEvent	NaN

TABLE 5.9. Pearson correlation matrix with values between -0.7 and 0.7 with natural gas price.

Variable	Price_oil	Open_oil	High_oil	Low_oil	Open_ngp	High_ngp	Low__ngp	Price_ngp
Price_oil	1.000000	0.992711	0.994759	0.997807	0.721894	0.722315	0.721408	0.722026
Open_oil	0.992711	1.000000	0.997544	0.994174	0.725619	0.725832	0.725099	0.725523
High_oil	0.994759	0.997544	1.000000	0.992507	0.737587	0.738181	0.737114	0.737906
Low_oil	0.997807	0.994174	0.992507	1.000000	0.710740	0.711108	0.710183	0.710763
Open_ngp	0.721894	0.725619	0.737587	0.710740	1.000000	0.999309	0.999440	0.999004
High_ngp	0.722315	0.725832	0.738181	0.711108	0.999309	1.000000	0.998746	0.999524
Low__ngp	0.721408	0.725099	0.737114	0.710183	0.999440	0.998746	1.000000	0.999375
Price_ngp	0.722026	0.725523	0.737906	0.710763	0.999004	0.999524	0.999375	1.000000

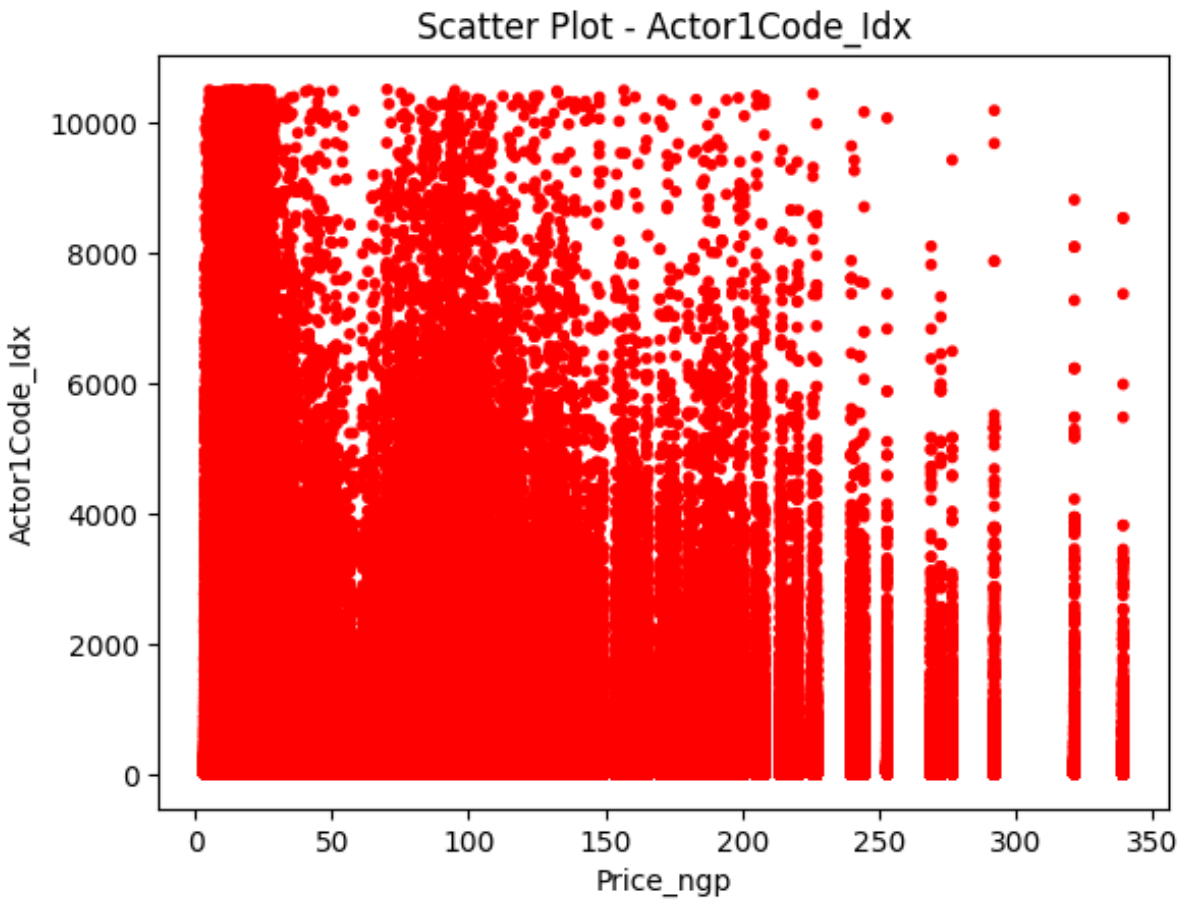

FIGURE 5.1. Scatter plot between natural gas price and Actor1Code_Idx feature.

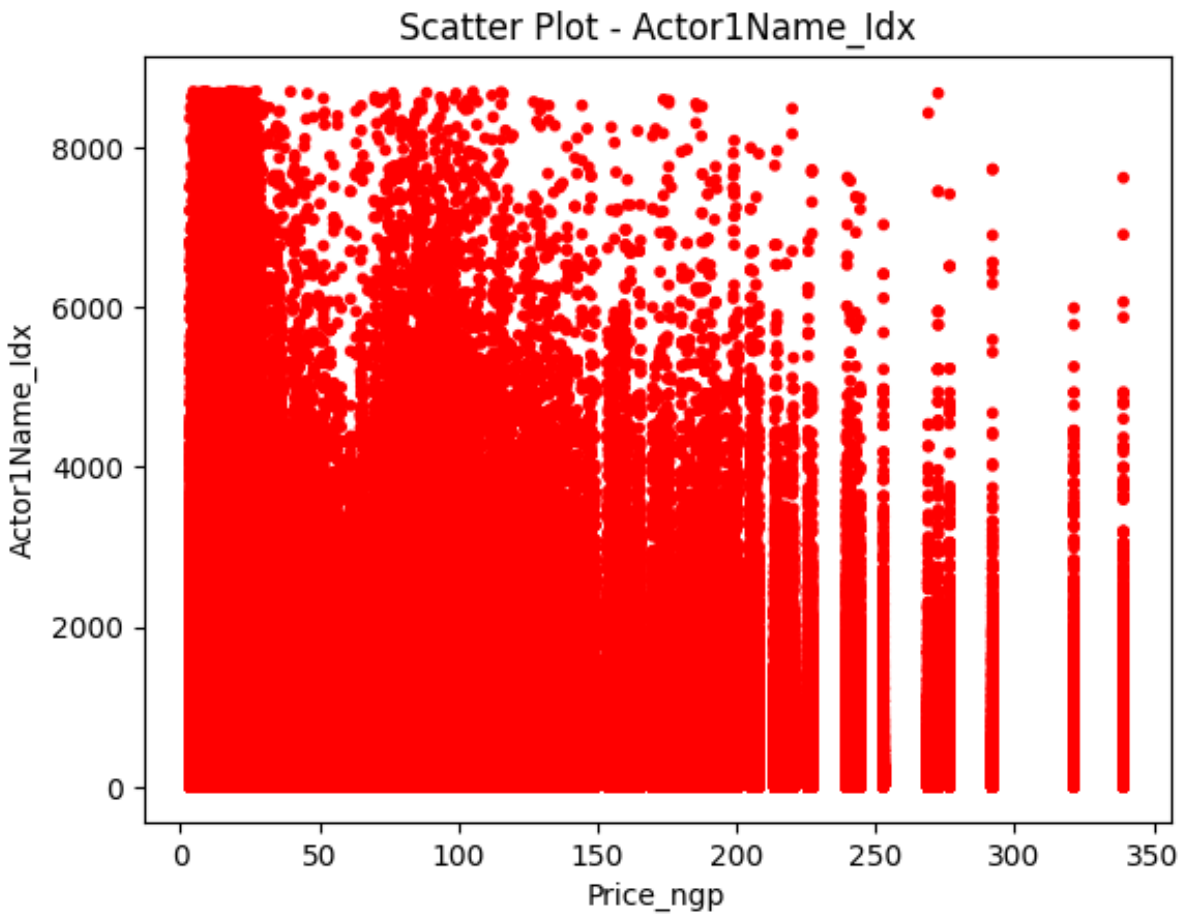

FIGURE 5.2. Scatter plot between natural gas price and Actor1Name_Idx feature.

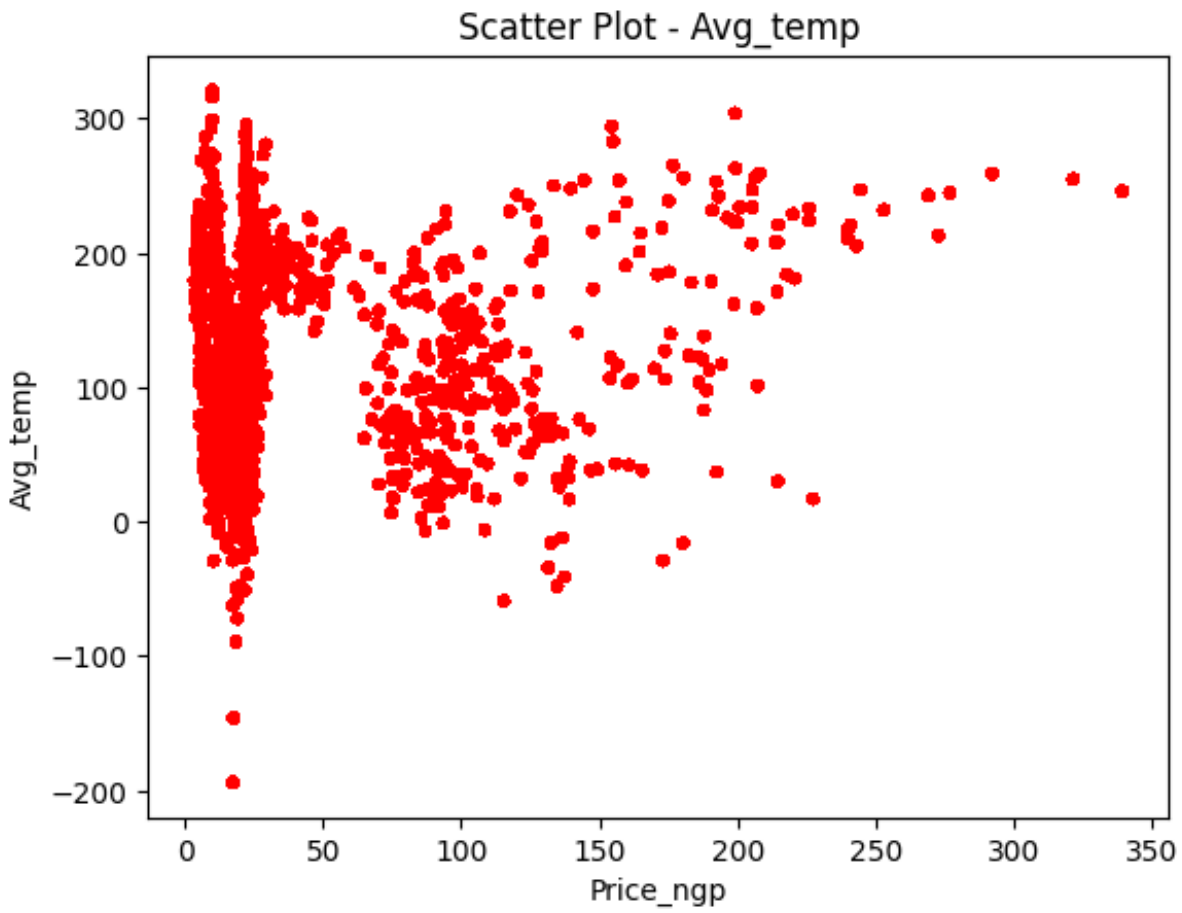

FIGURE 5.3. Scatter plot between natural gas price and average temperature feature.

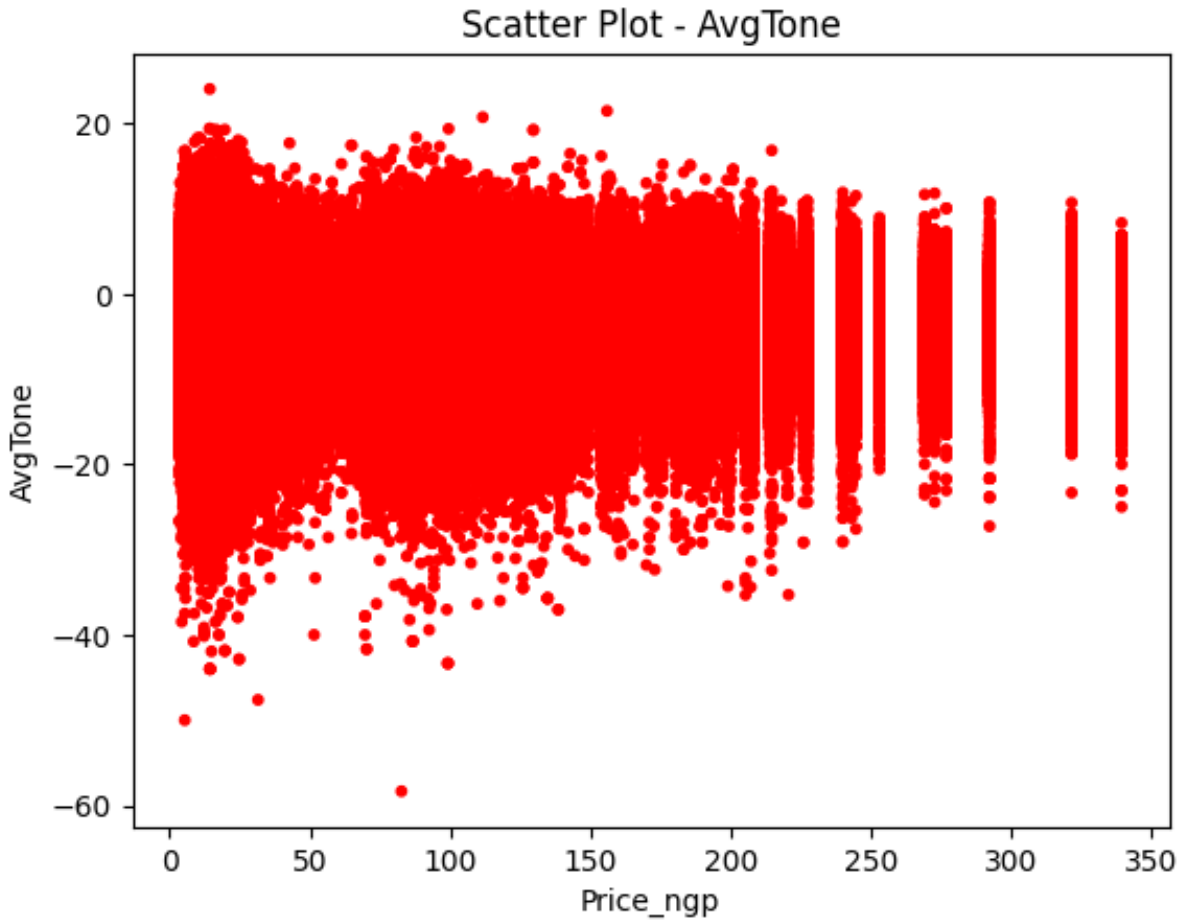

FIGURE 5.4. Scatter plot between natural gas price and average tone feature.

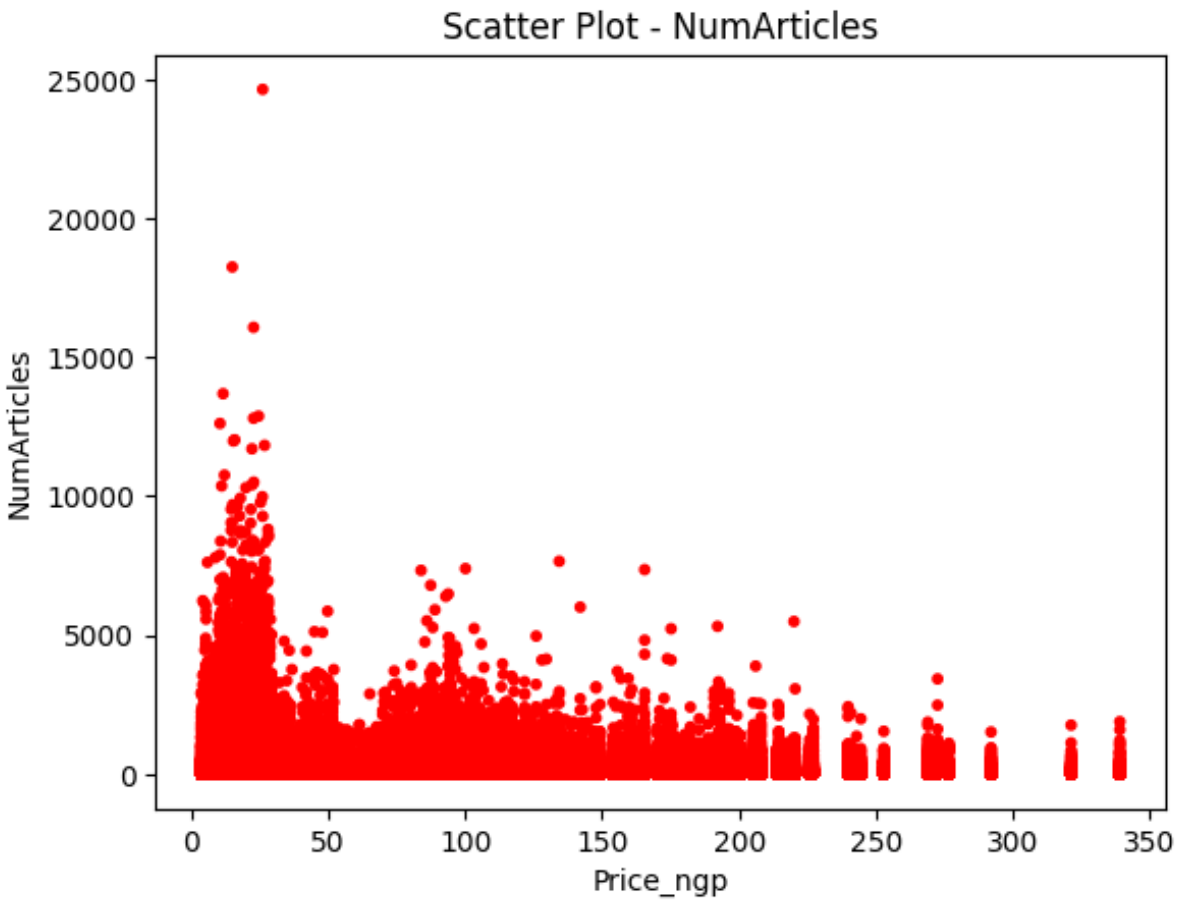

FIGURE 5.5. Scatter plot between natural gas price and the number of articles feature.

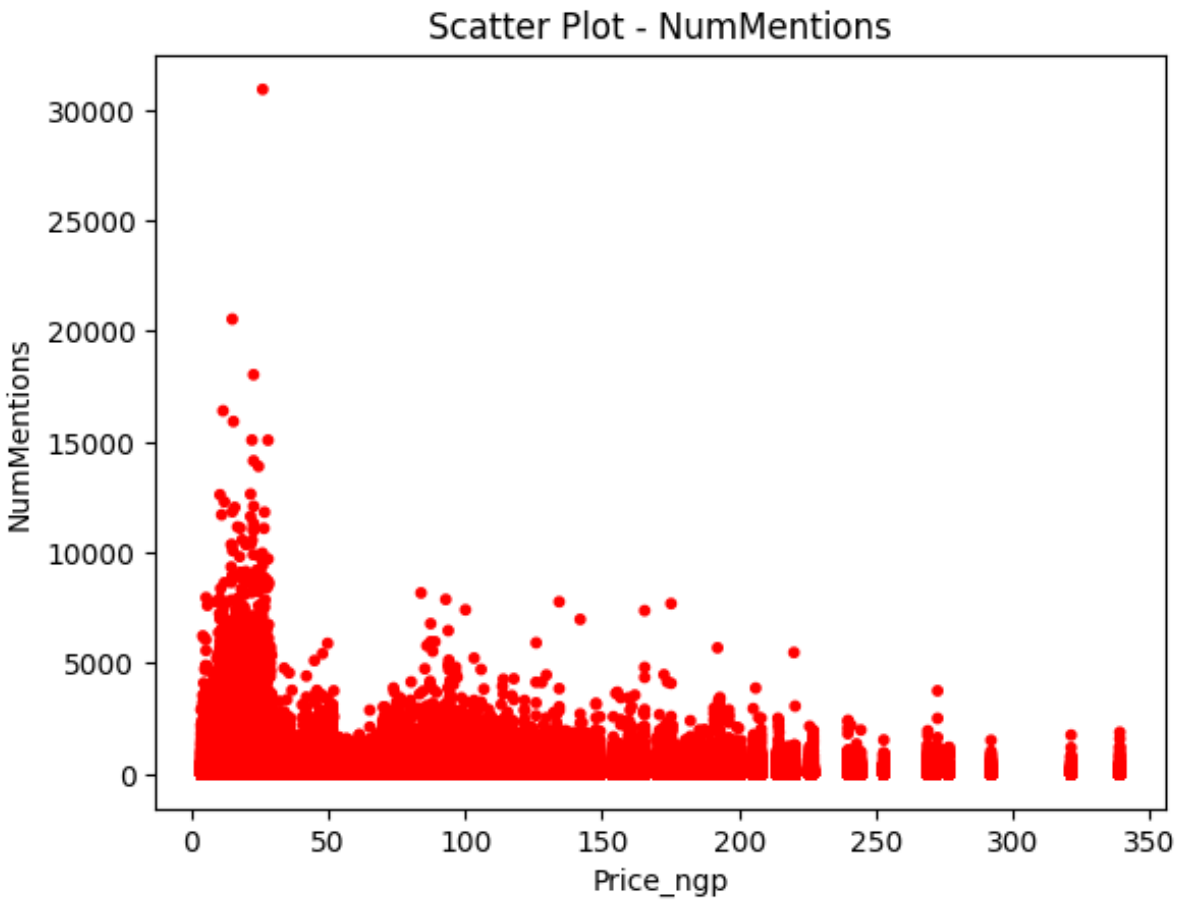

FIGURE 5.6. Scatter plot between natural gas price and the number of mentions feature.

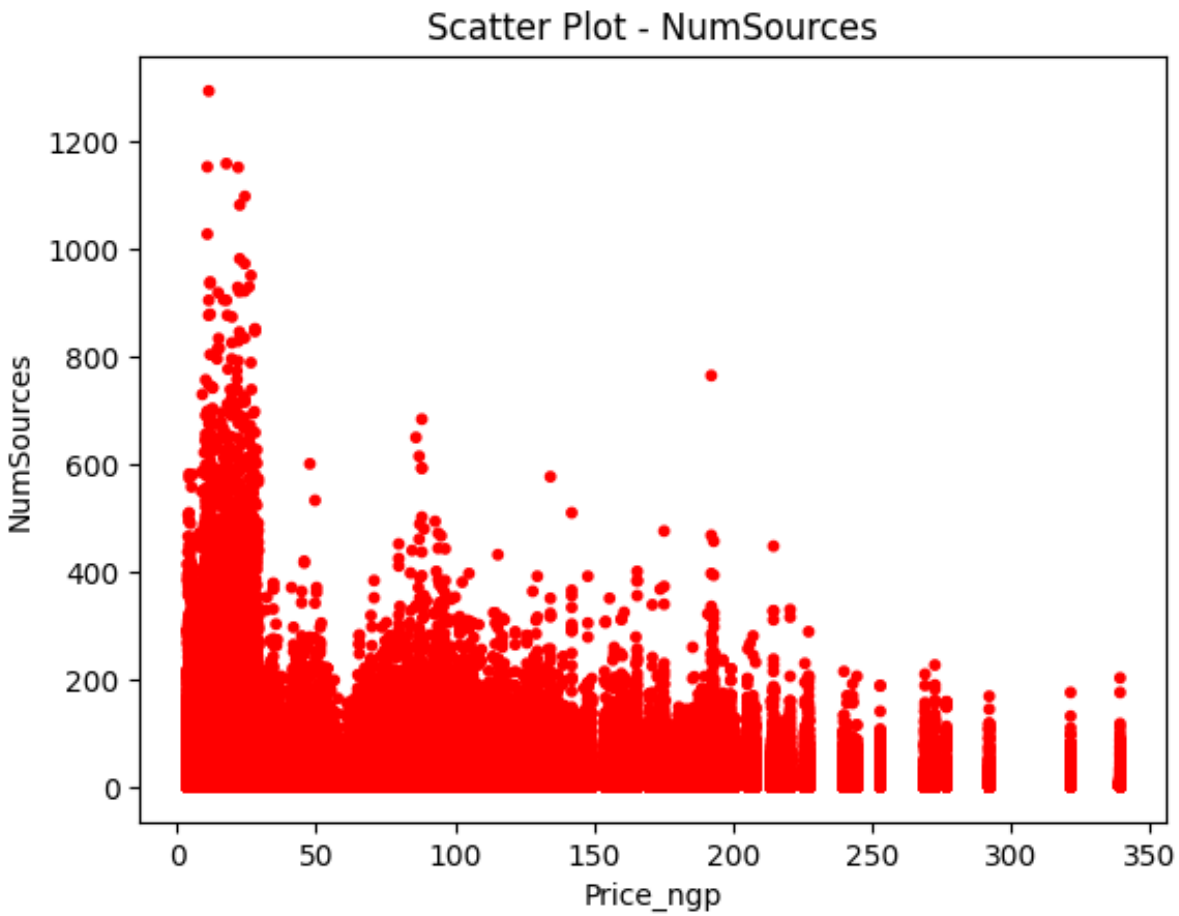

FIGURE 5.7. Scatter plot between natural gas price and the number of sources feature.

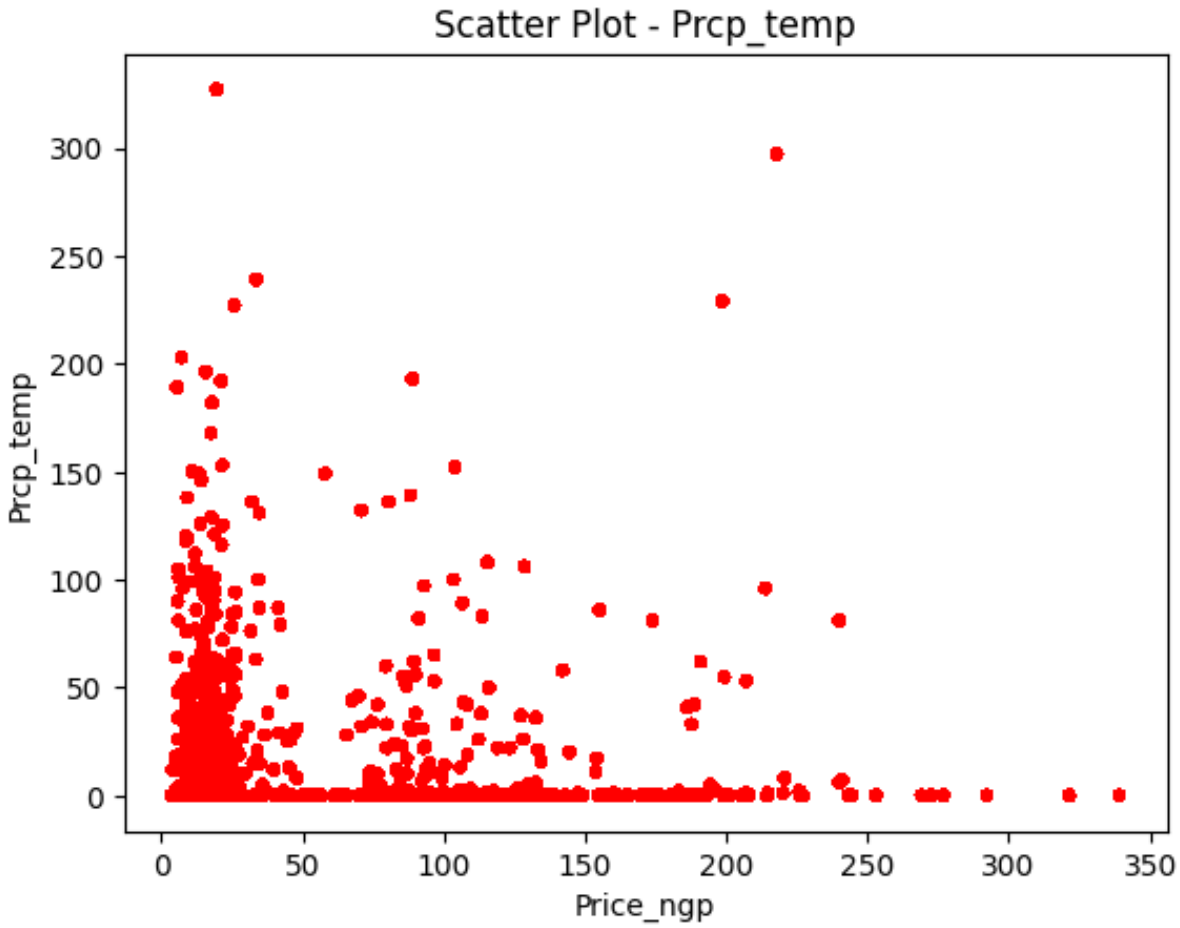

FIGURE 5.8. Scatter plot between natural gas price and precipitation feature.

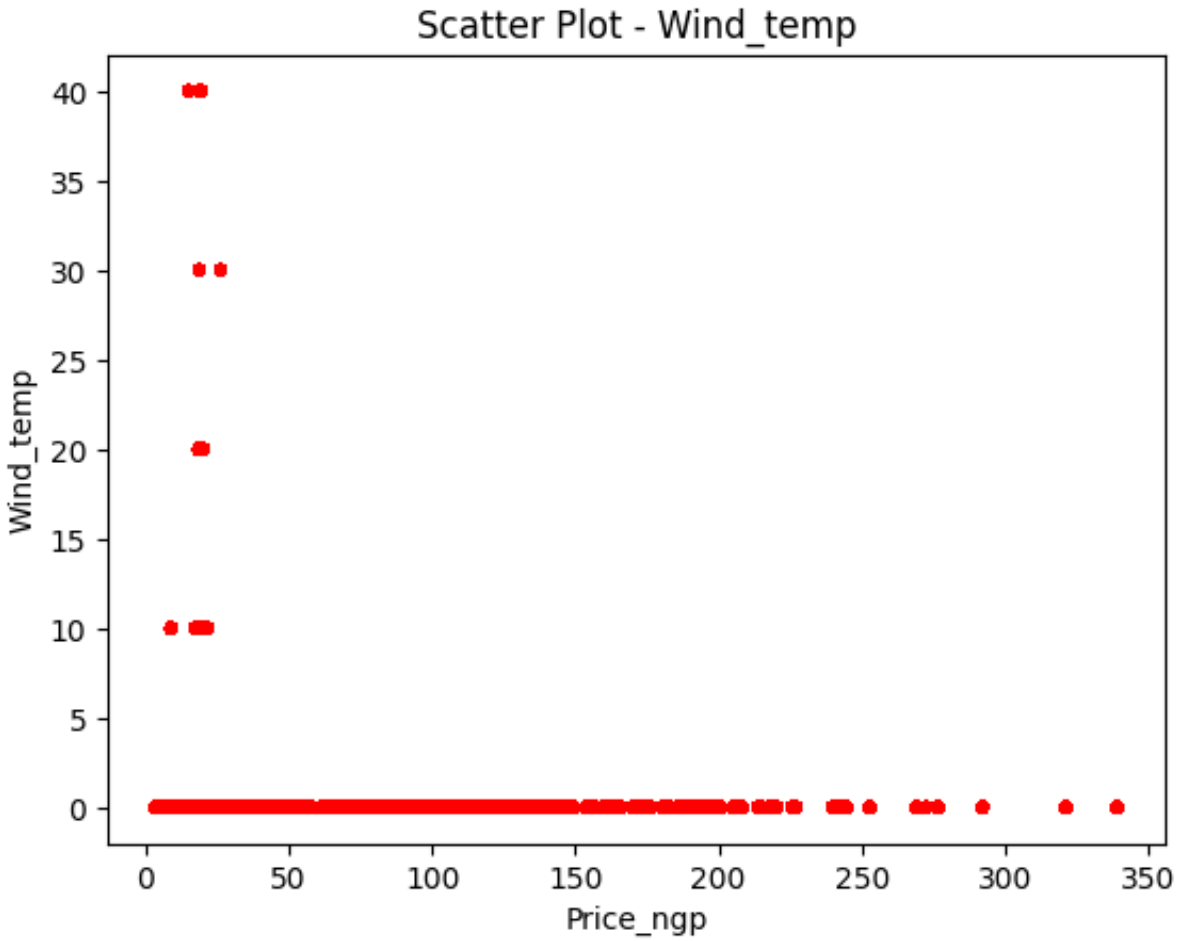

FIGURE 5.9. Scatter plot between natural gas price and wind feature.

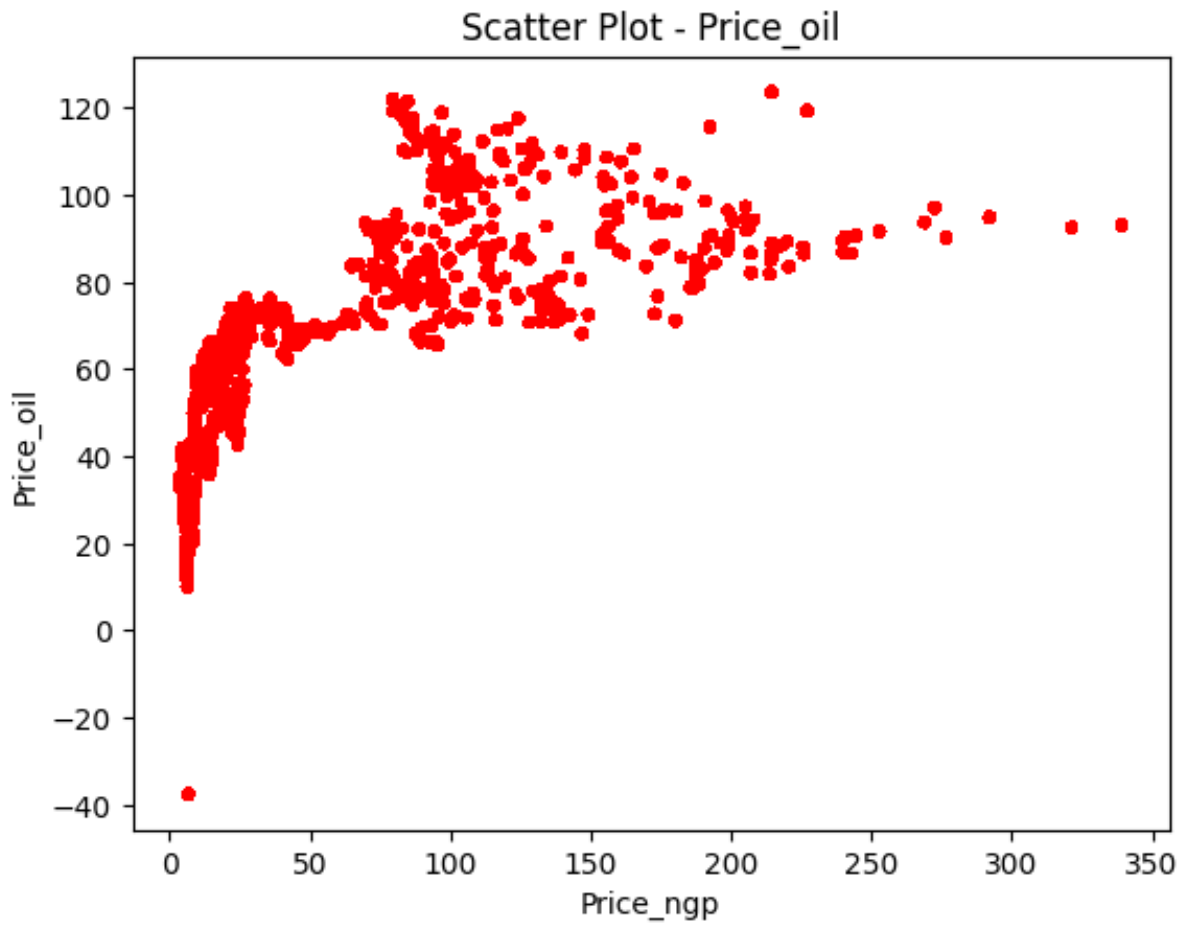

FIGURE 5.10. Scatter plot between natural gas price and crude oil price feature.

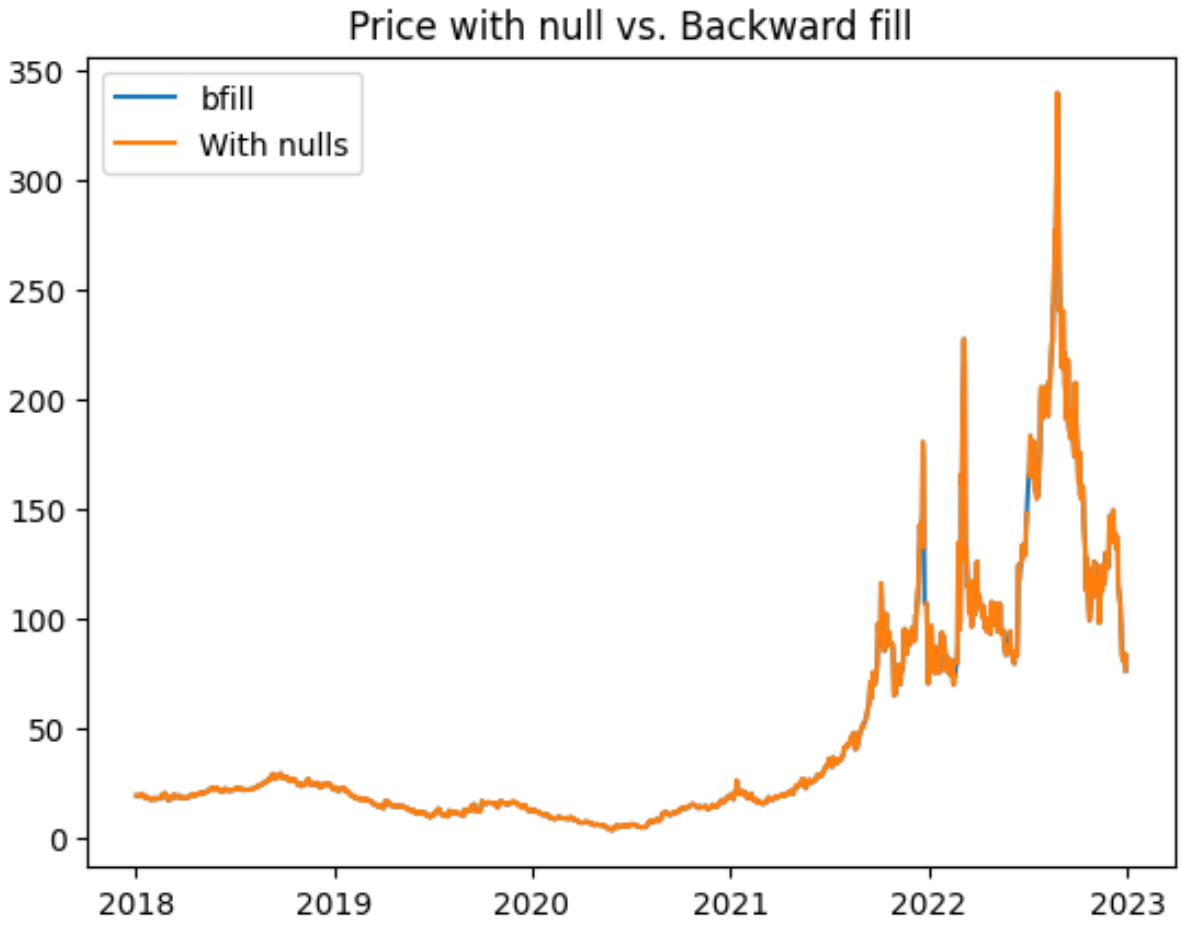

FIGURE 5.11. Visualization of natural gas prices with null values filled with the backward values method.

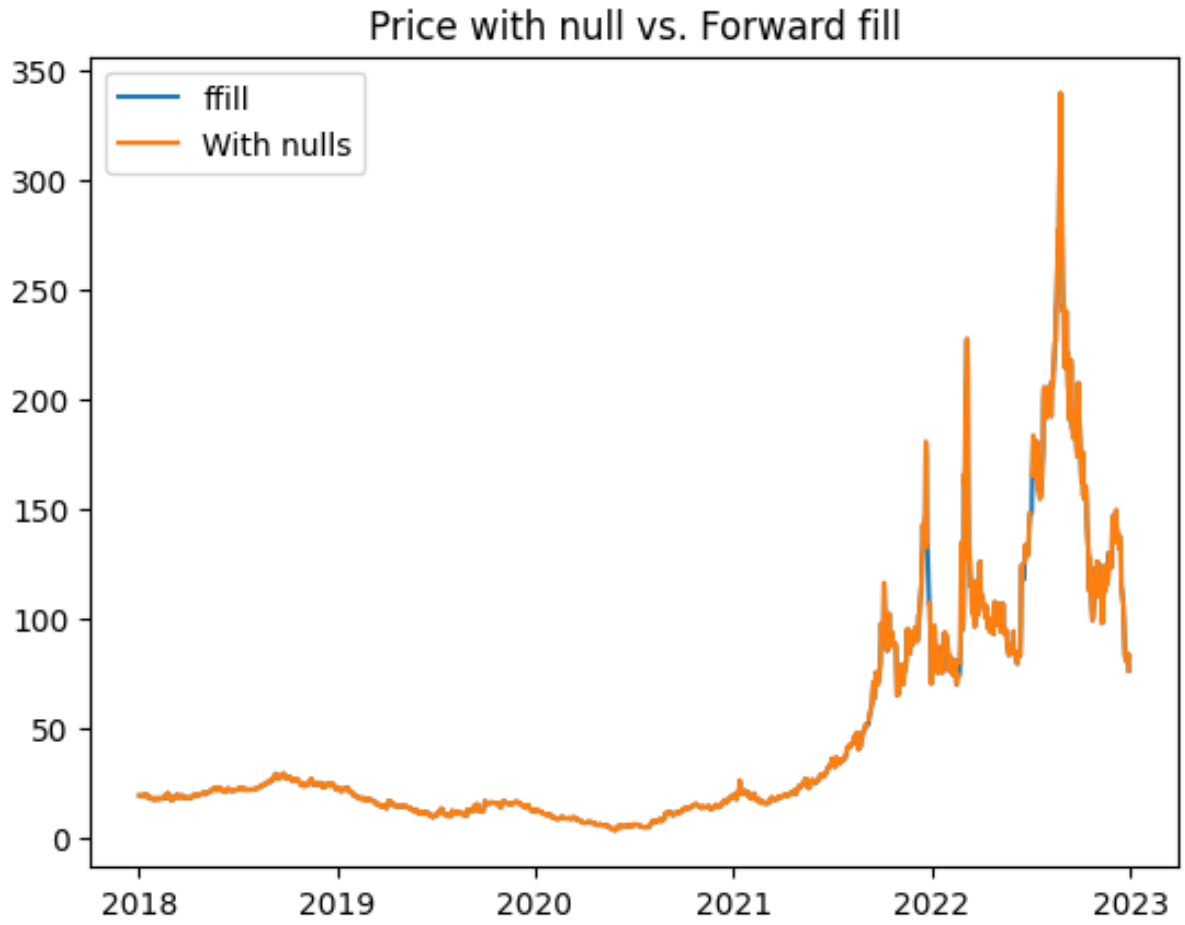

FIGURE 5.12. Visualization of natural gas prices with null values filled with the forward values method.

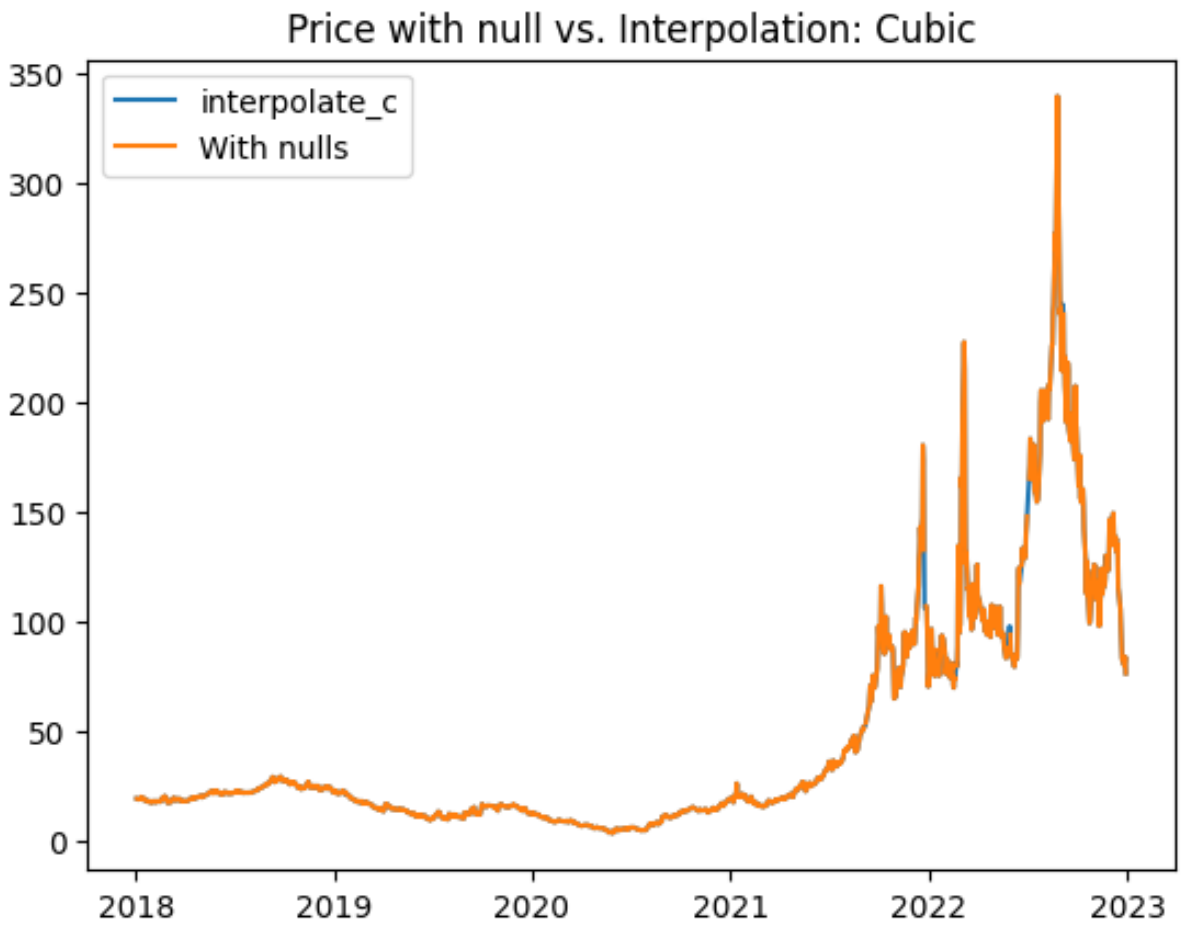

FIGURE 5.13. Visualization of natural gas prices with null values filled with the interpolation cubic values method.

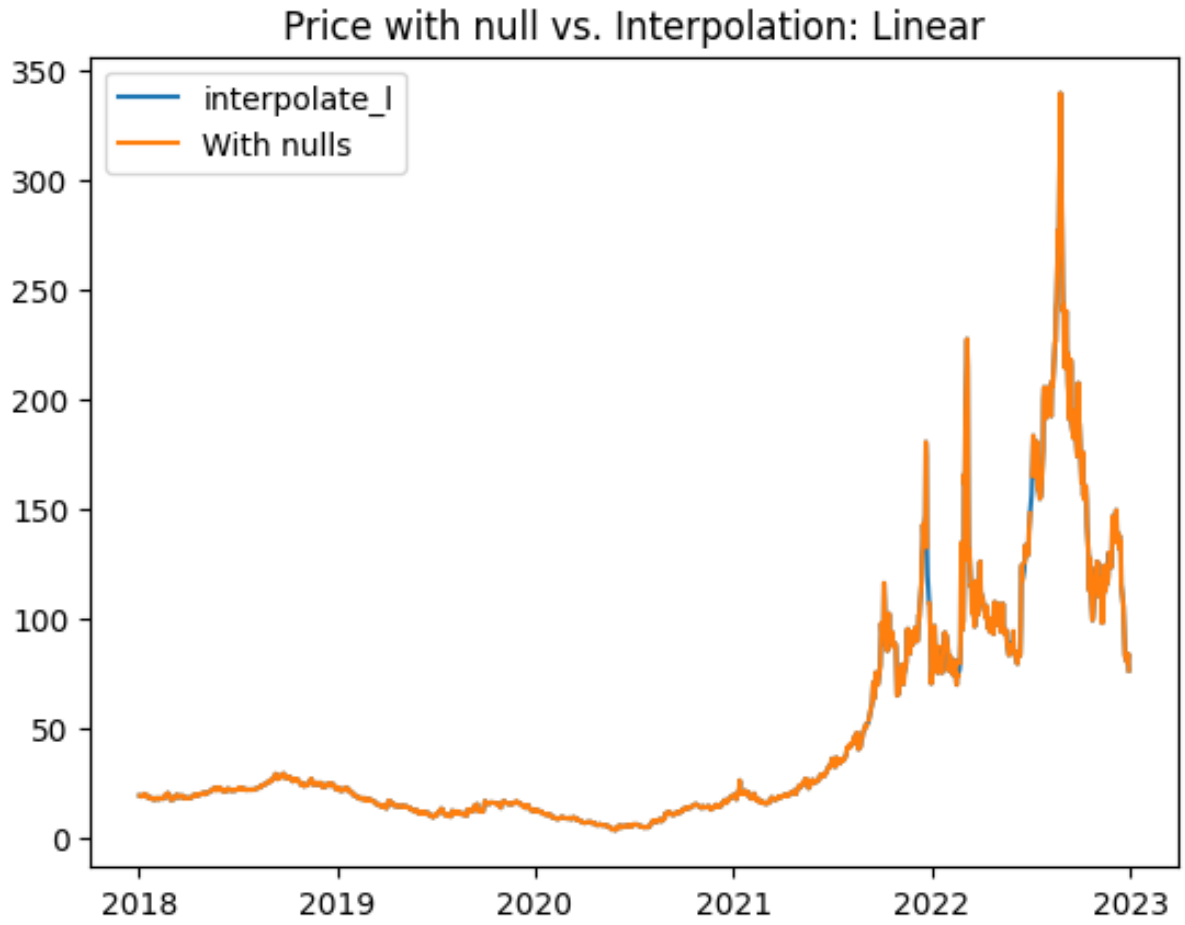

FIGURE 5.14. Visualization of natural gas prices with null values filled with the interpolation linear values method.

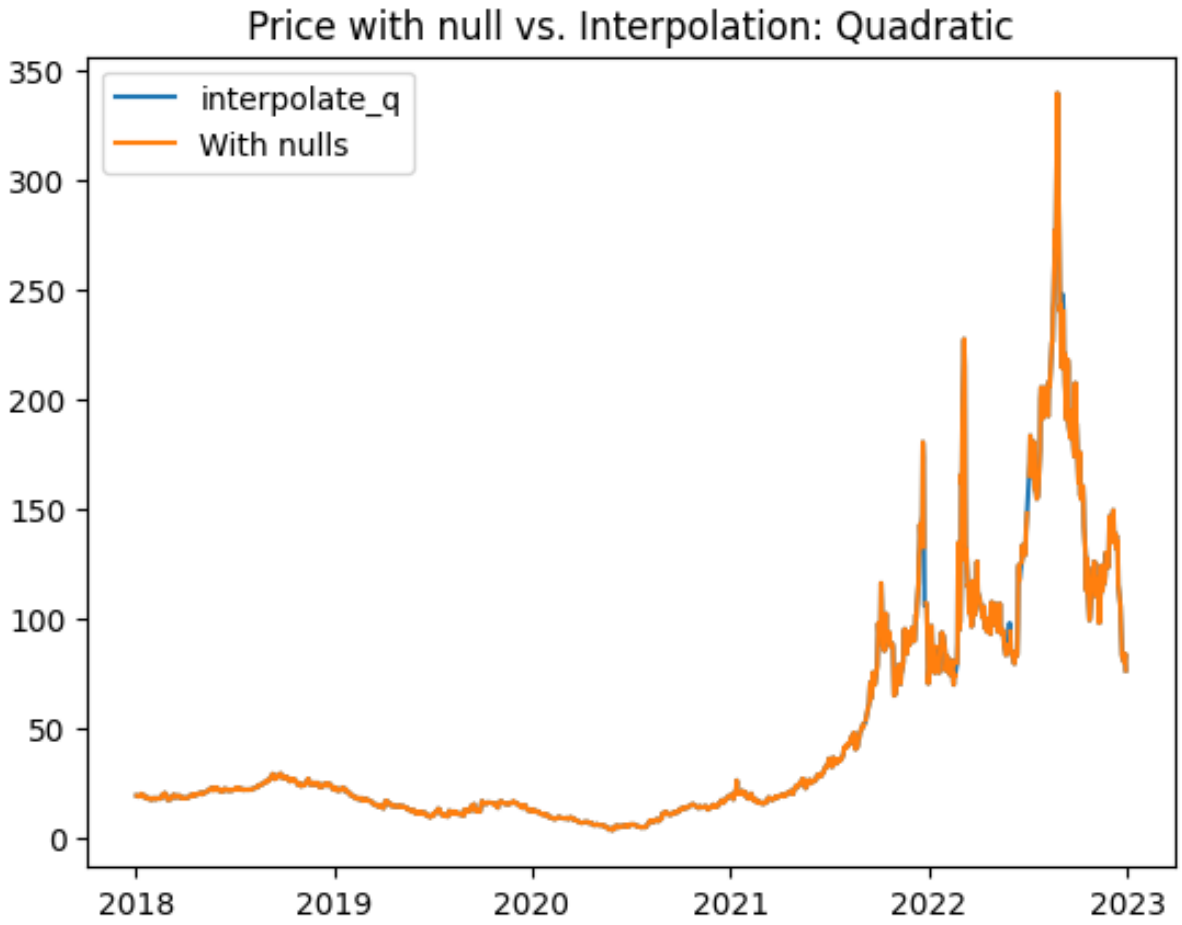

FIGURE 5.15. Visualization of natural gas prices with null values filled with the interpolation quadratic values method.

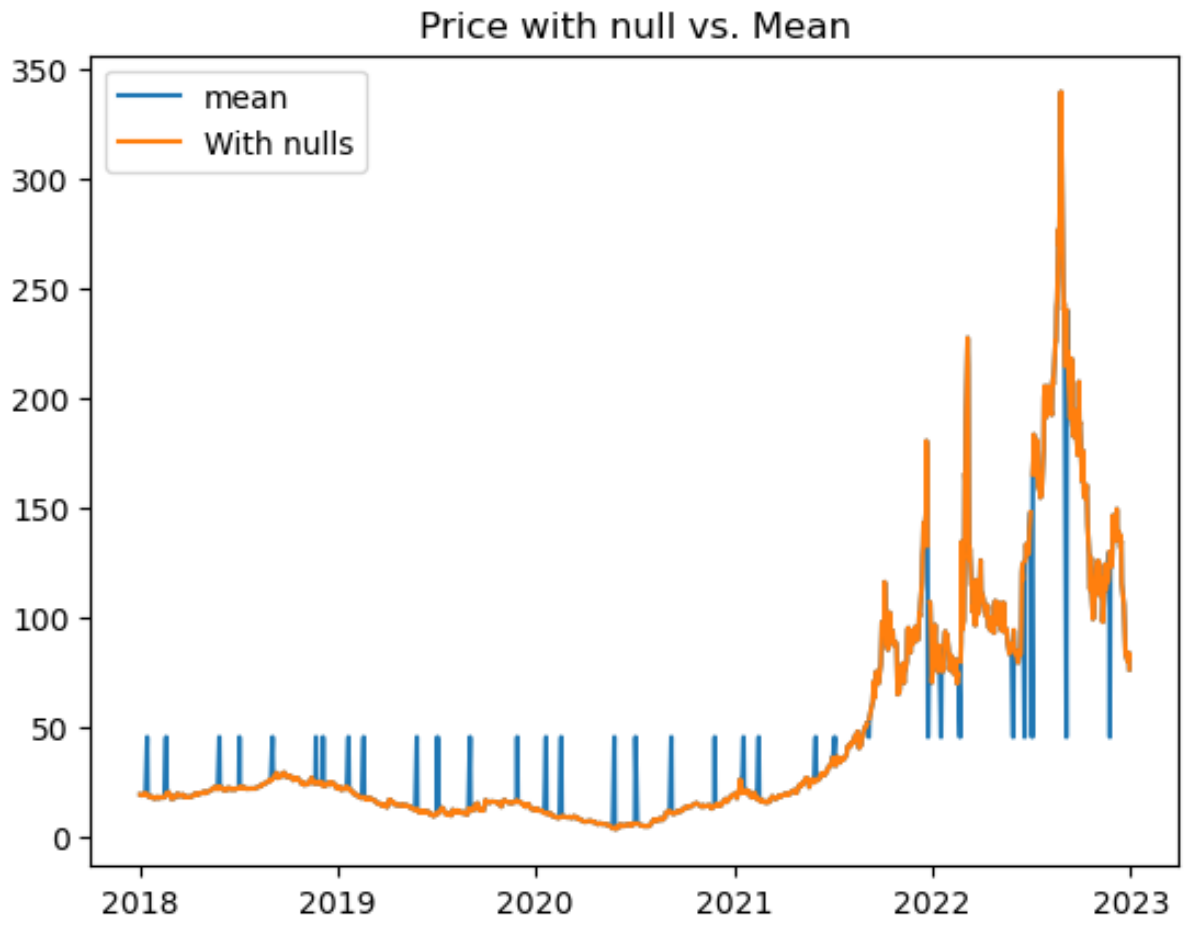

FIGURE 5.16. Visualization of natural gas prices with null values filled with mean value.

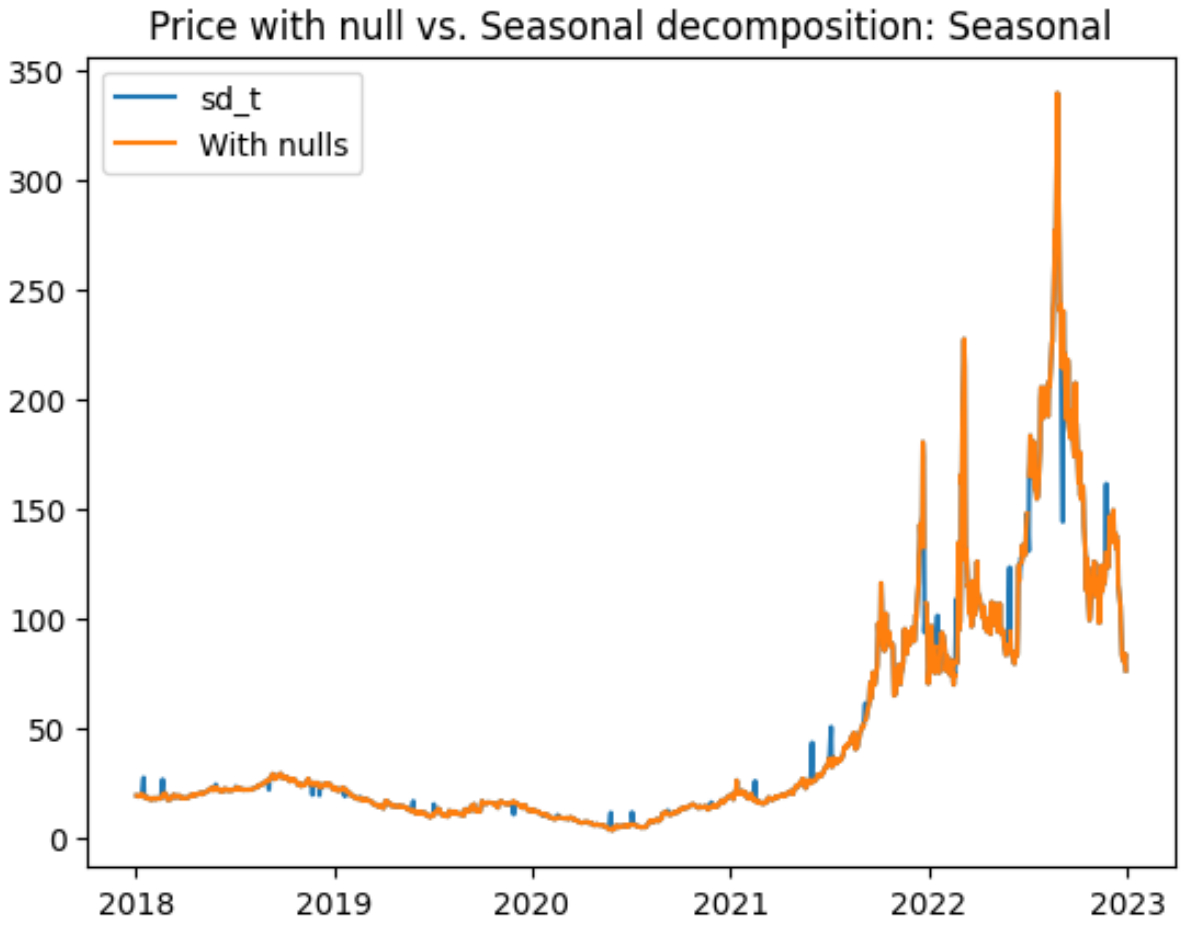

FIGURE 5.17. Visualization of natural gas prices with null values filled with the seasonal values from the seasonal decomposition method.

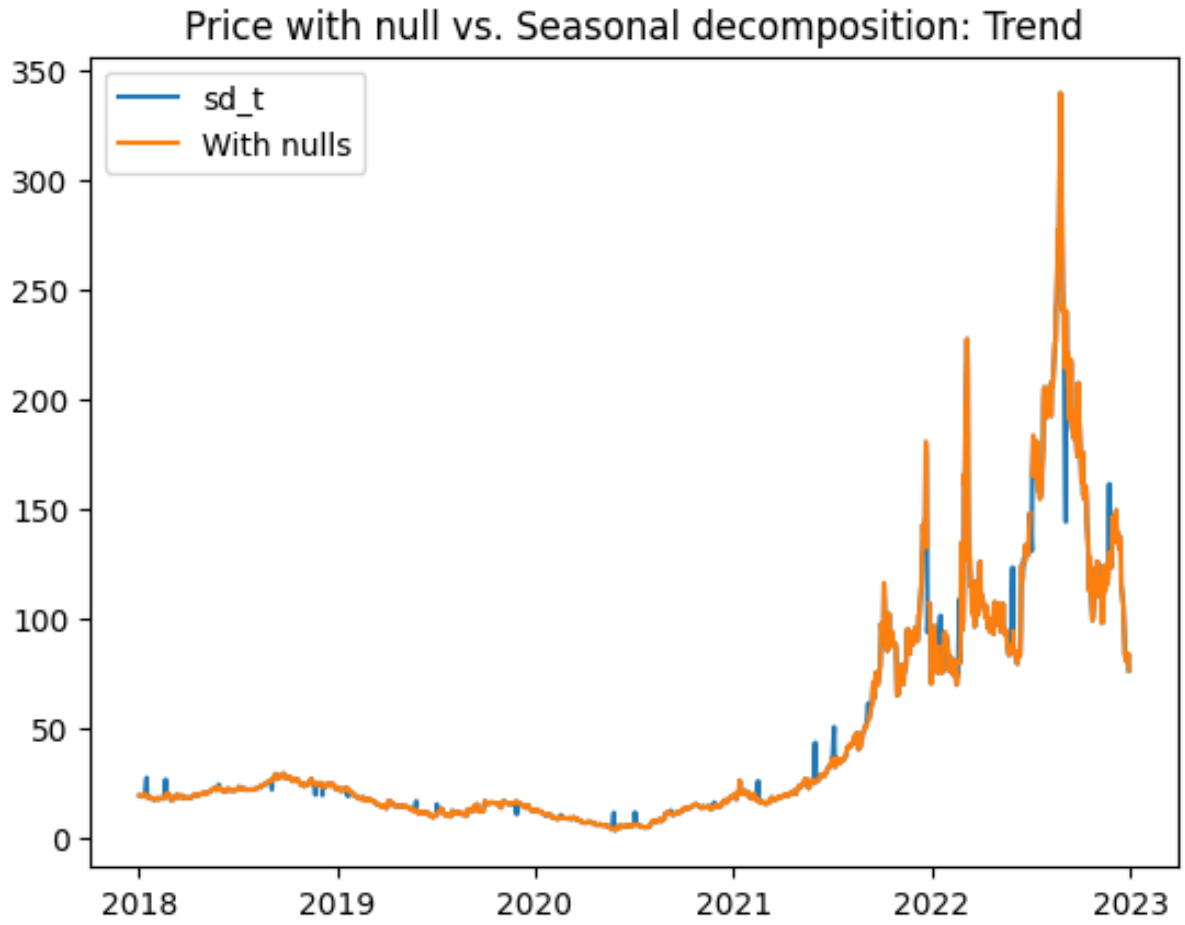

FIGURE 5.18. Visualization of natural gas prices with null values filled with the trend values from the seasonal decomposition method.

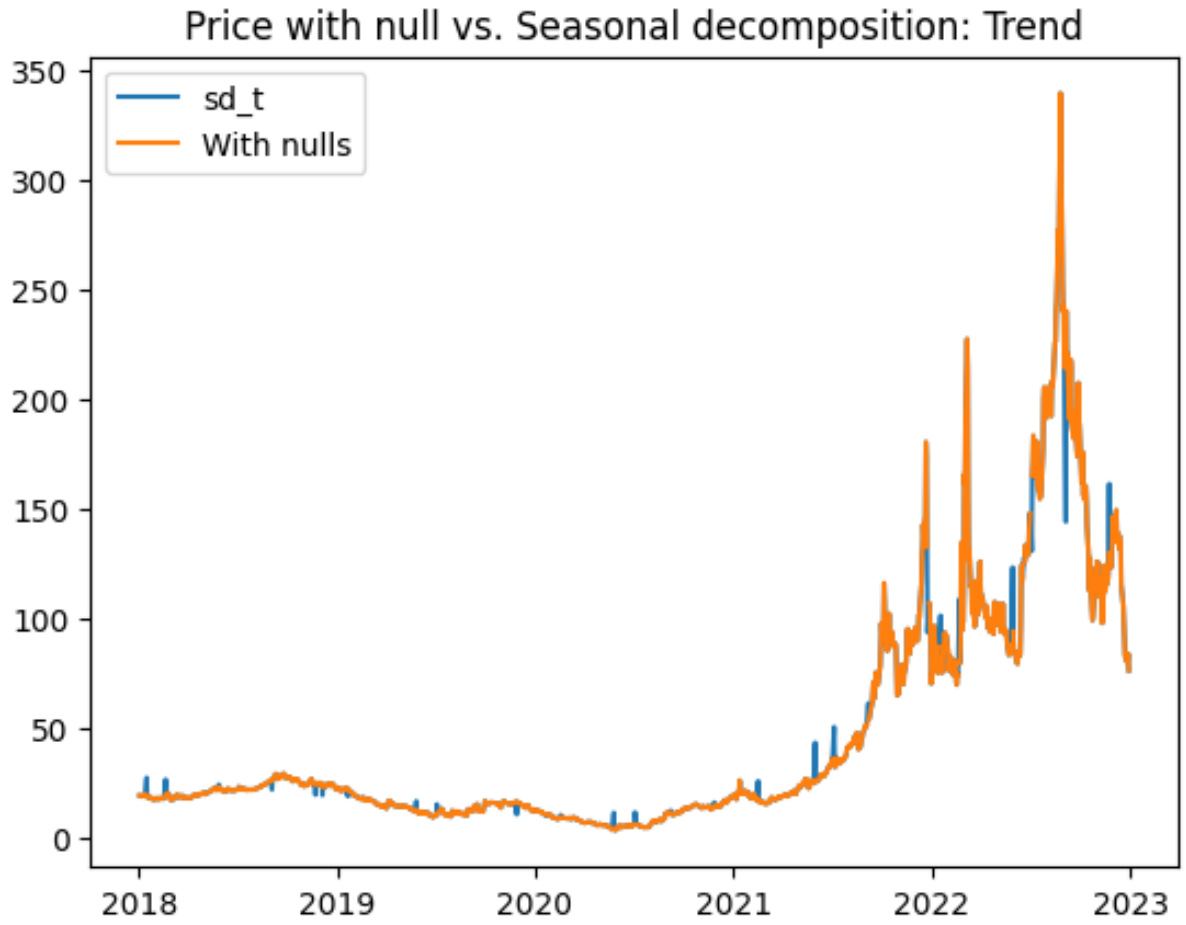

FIGURE 5.19. Visualization of natural gas prices with null values filled with the trend values from the seasonal decomposition method.

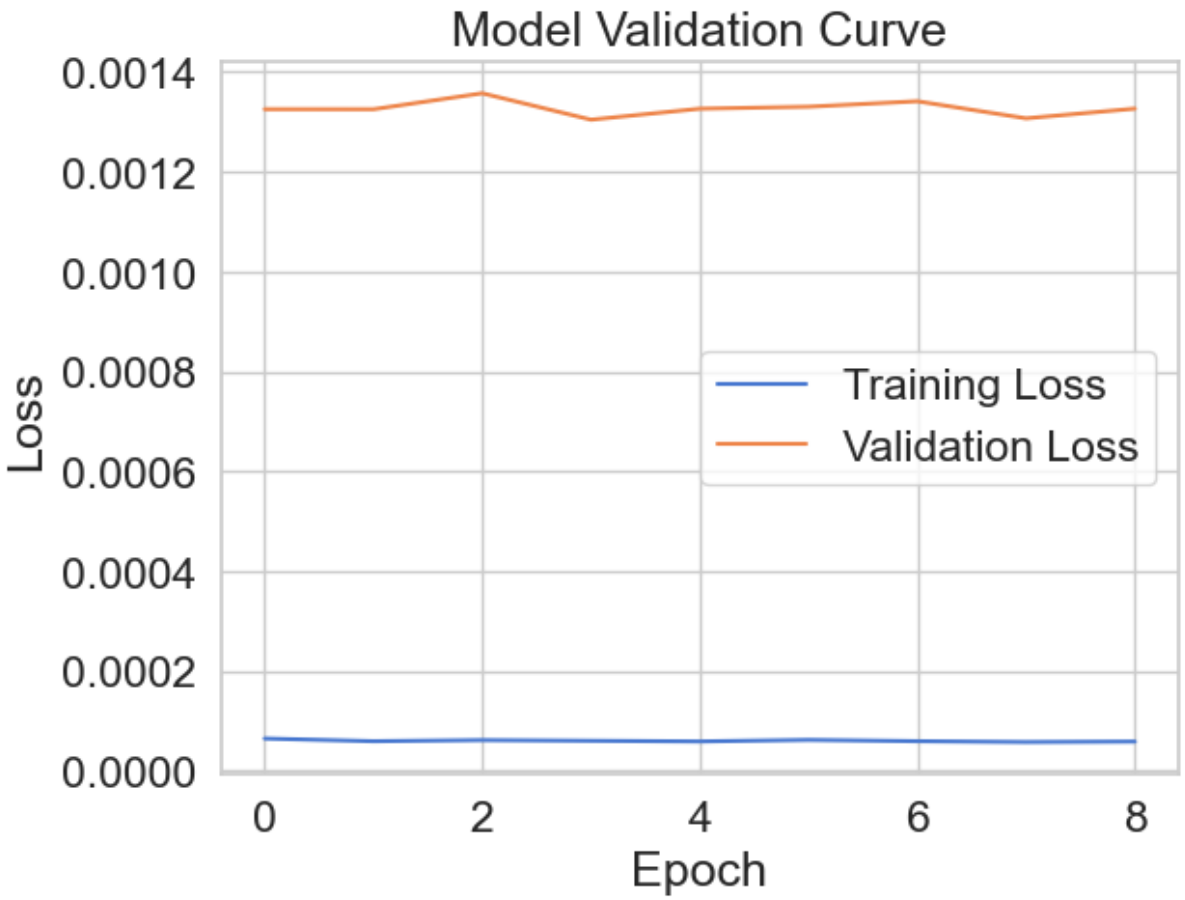

FIGURE 5.20. Validation curve of RNN model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

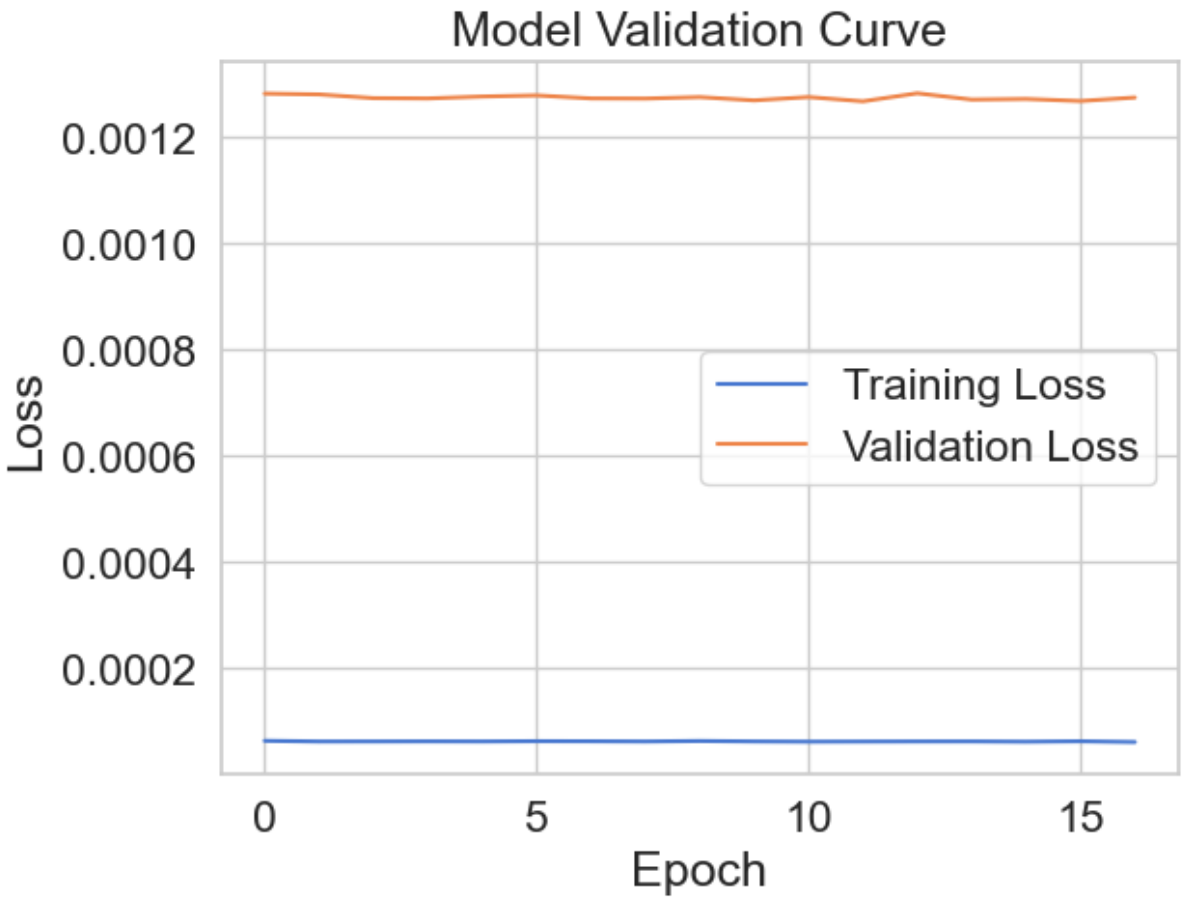

FIGURE 5.21. Validation curve of GRUNN model with natural gas price and crude oil price as features, and lag equal to 5.

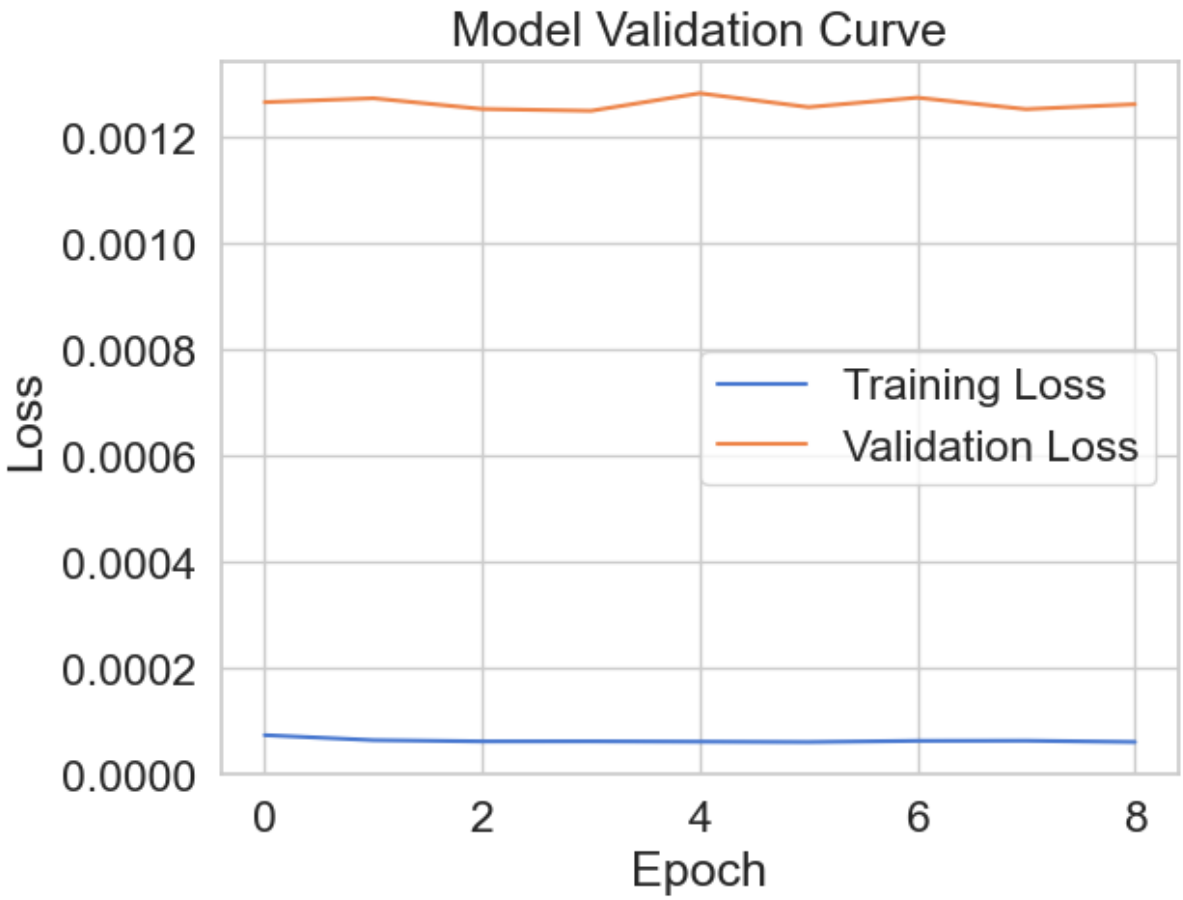

FIGURE 5.22. Prediction plot of RNN model with natural gas price and crude oil price as features, and lag equal to 10.

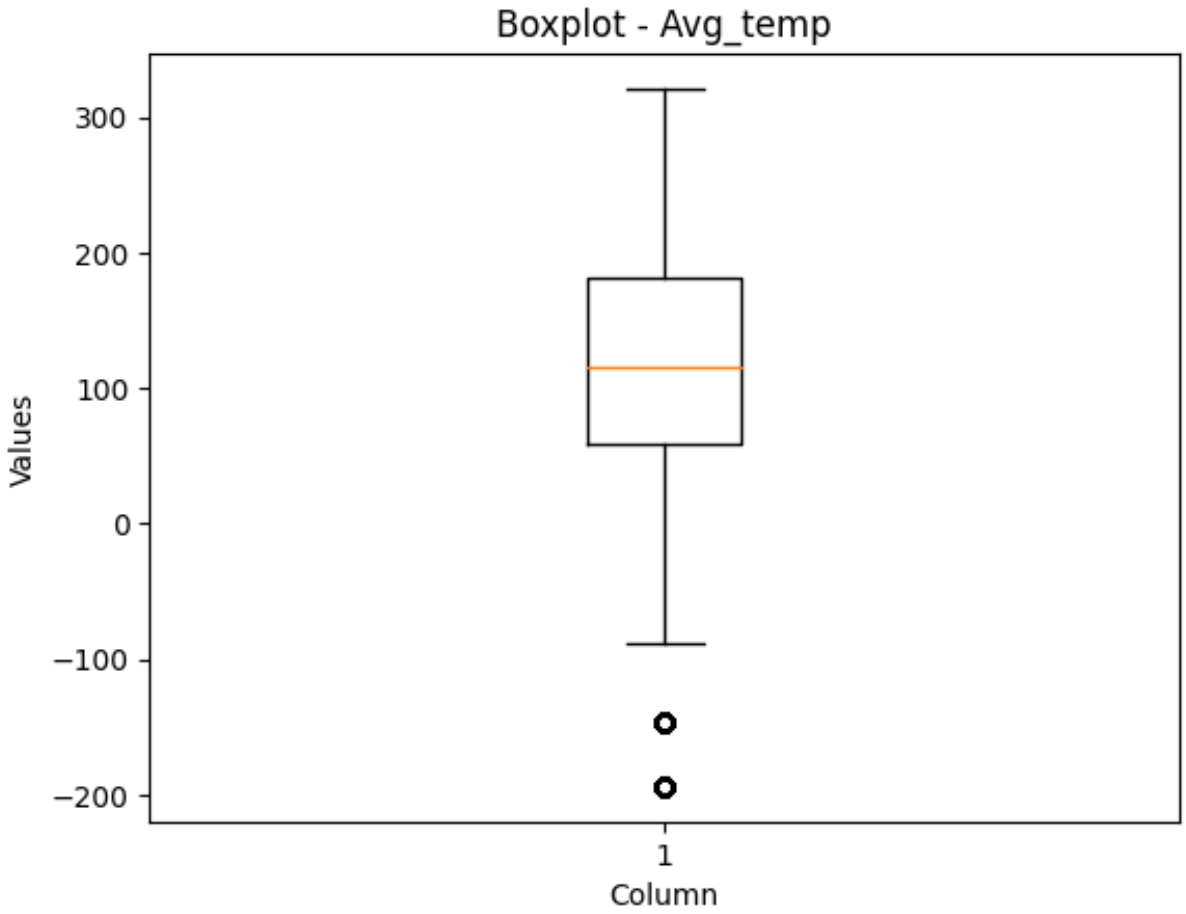

FIGURE 5.23. Boxplot of average temperature.

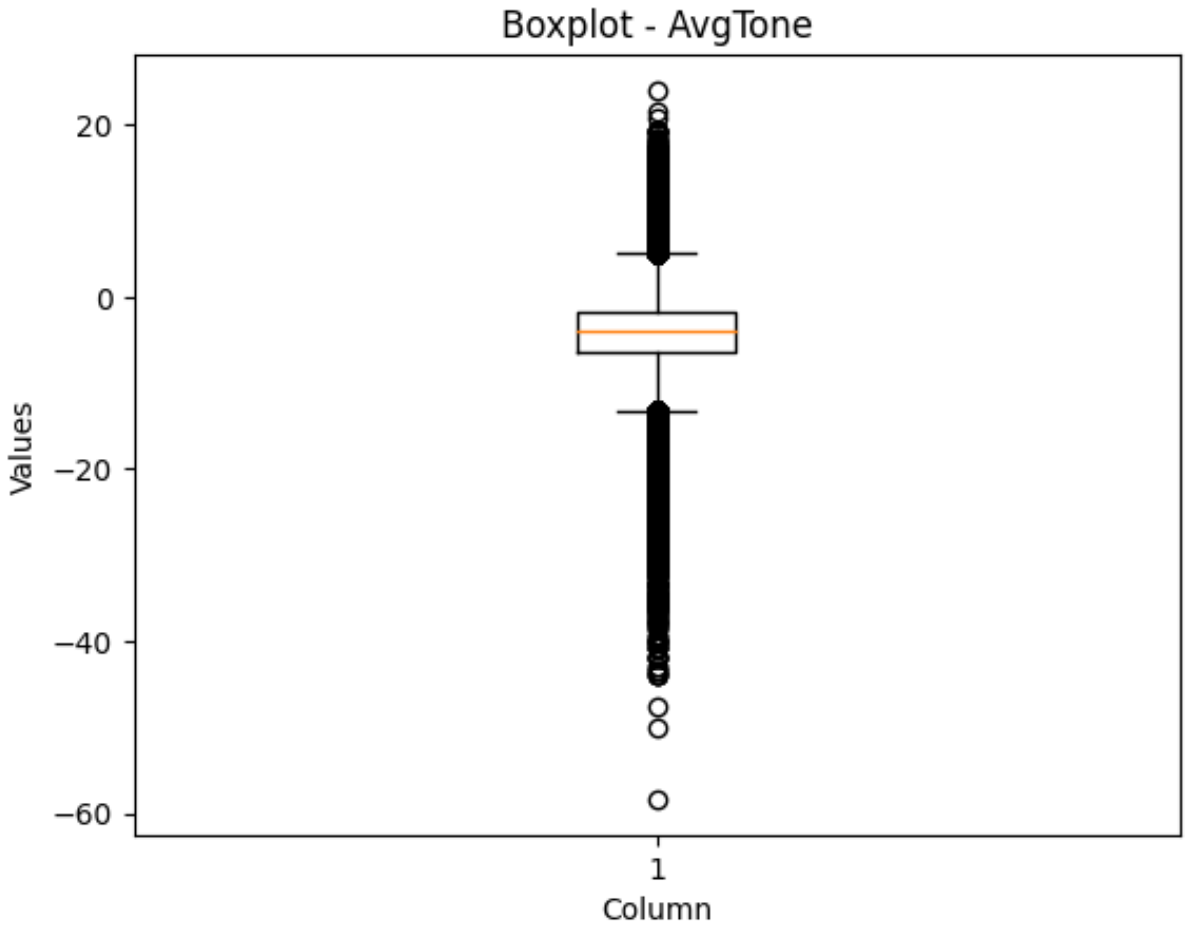

FIGURE 5.24. Boxplot of average tone.

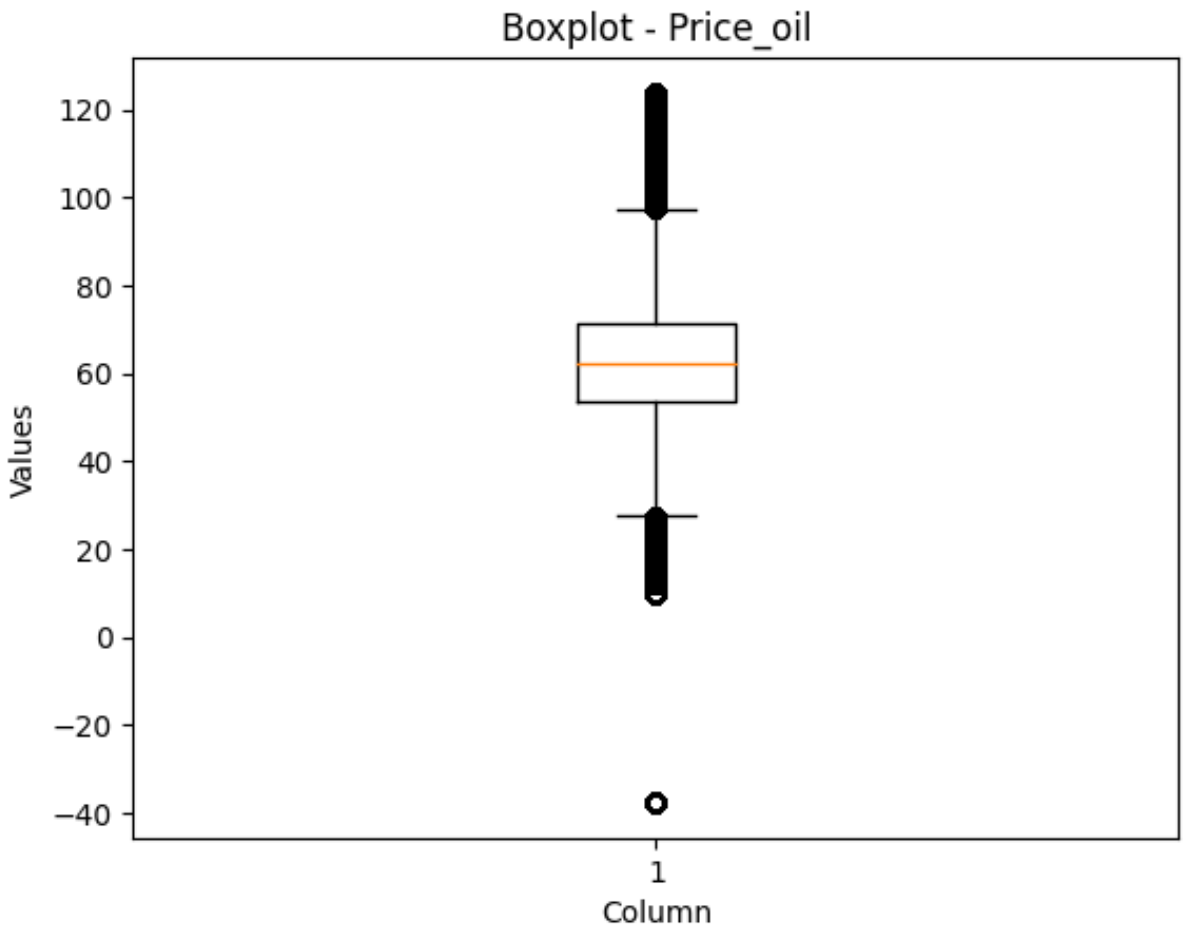

FIGURE 5.25. Boxplot of crude oil price.

Acknowledgment

I am deeply grateful for Professor Diana Mendes' invaluable guidance, unwavering support, and insightful advice during my academic journey. Her mentorship has been instrumental in shaping my research and navigating the complexities of this dissertation.

Professor Mendes' expertise and profound knowledge in the field have been a constant source of inspiration, motivating me to strive for excellence. Her willingness to share wisdom and provide constructive feedback has been crucial in refining the ideas and methodologies of this study.

I sincerely appreciate the countless hours Professor Mendes dedicated to reviewing and discussing my research progress. Her encouragement during moments of doubt and belief in my capabilities have been pivotal in overcoming challenges.

I also extend my appreciation to the entire academic community at the University Institute of Lisbon for fostering an environment of intellectual growth and inquiry.

Lastly, my heartfelt acknowledgment goes to my wife, for her unwavering encouragement and love throughout this academic journey, providing me with the strength to pursue my passion for research and education.

This dissertation owes its success to Professor Diana Mendes' guidance, support, and the contributions of all those who have played a significant role in shaping my academic and personal growth. I am truly honored and humbled by their involvement in this work.

Thank you all for being an integral part of this rewarding journey.

Sincerely,

René Porto.

Resumo

A transição para fontes de energia mais limpas na União Europeia prioriza o gás natural, no entanto, a Guerra Russo-Ucraniana causou flutuações imprevisíveis nos preços. Nosso estudo visou aprimorar modelos preditivos explorando dados do GDELT, analisando o desempenho pré e pós-guerra, e comparando modelos de "Deep Learning" (RNN, LSTM, GRUNN). A incorporação de dados de petróleo bruto e sentimento médio da nóticia melhorou significativamente as previsões. Fatores geopolíticos exigem mais pesquisas para garantir segurança energética e desenvolvimento econômico. Empregando a metodologia CRISP-DM, estabelecemos uma abordagem sistemática para enfrentar esses desafios. Nosso estudo contribui com insights valiosos para aprimorar as previsões e adaptar modelos aos complexos mercados de energia.

Abstract

The transition to cleaner energy in the European Union prioritizes natural gas, yet the Russo-Ukrainian War caused unpredictable price fluctuations. Our study aimed to enhance predictive models by exploring GDELT data, analyzing pre- and post-war performance, and comparing deep learning models (RNN, LSTM, GRUNN). Incorporating crude oil and average tone data significantly improved predictions. Geopolitical factors necessitate further research to ensure energy security and economic development. Employing CRISP-DM methodology, we established a systematic approach to address these challenges. Our study contributes valuable insights to enhance predictions and adapt models to complex energy markets.

Contents

CHAPTER 1

Introduction

The shift towards cleaner energy sources is a top priority for the European Union, with natural gas being widely embraced by countries to achieve emission reduction goals. This energy resource is predominantly transported through pipelines and offers convenient storage options.
However, the stability of natural gas prices was significantly impacted by the Russo-Ukrainian War. The weaponization of gas sales by Russia during the conflict caused a drastic increase in gas prices, leading to unprecedented price fluctuations. As a result, the predictive models developed in this study were not trained to anticipate such extraordinary events, and their performance was affected.
This unforeseen instability in natural gas prices during the war highlights the need for further research and model adaptation to account for geopolitical factors that can influence energy markets. The ability to forecast such events accurately will be crucial for ensuring energy security and sustainable economic development in the future.

The primary objectives of this study are as follows:

(1) To investigate whether the utilization of data extracted from GDELT (Global Database of Events, Language, and Tone) contributes to an enhancement in model performance.
(2) To assess whether the predictive model trained with data before the Russo-Ukrainian War demonstrates a similar performance to the model that did not anticipate this historical phase.
(3) To compare the performance of different deep learning models, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN), and determine which one yields the best results.

The dataset of TTF natural gas price, obtained from the yfinance Python library[1], spans from October 23, 2017. Our focus in this study was specifically on the time period between January 2, 2018, and December 30, 2022, comprising a total of 1292 data points. To train the models, we utilized only the data from the first year of the war, employing an 80/20 ratio for training and testing data.
In anticipation of the modeling phase, a set of 15 distinctive features was meticulously

[1]https://pypi.org/project/yfinance/

constructed, encompassing a diverse range of attributes. An exhaustive examination of the collective attributes was undertaken, as detailed in Table 4.1 at page 42. However, this process of evaluation led to a deliberate focus on the distinct constituent elements within the feature set. The ensuing selection comprised three pivotal features: natural gas price, crude oil price, and average tone. The strategic amalgamation of these selected features is delineated in Table 4.3 at page 47, guided by a rationale anchored in the incorporation of a natural gas price baseline intertwined with the nuanced interplay of crude oil price and average tone attributes.

Throughout our study, we observed that no prior research had attempted to forecast natural gas prices during periods of war. Additionally, none of the existing studies utilized the Cameo list of events[2] to predict natural gas prices.

The employed methodology was CRISP-DM, and the basic flow can be observed in Figures 3.3, 3.1, 3.5, and 4.1, at pages 25, 23, 29, and 39, respectively:

Step 1: Extraction of natural gas, crude oil, GDELT news, and weather data.
Step 2: Data analysis.
Step 3: Individual transformation and preprocessing of data.
Step 4: Training and fitting of models using Keras Tuner to select the best hyperparameters.
Step 5: Evaluation of models.

The dissertation is structured as follows:

(1) Introduction: This part gives a quick overview of why the research is important and what it aims to achieve. It introduces the main problem, goals, and questions that the rest of the dissertation will explore.
(2) Methodology and Literature Review: In this section, we explain how we did the research, like collecting and studying data. We also talk about what other people have researched on this topic before. This helps set the groundwork for our own research.
(3) Data and Preprocessing: Now we talk about the information we used for the research. We describe where we got it from, what it includes, and how we made sure it was good to use. This step is really important to make sure our analysis is accurate.
(4) Conclusion: This is the last chapter of the dissertation. We summarize what we found out from our research and discuss how it answers the questions we had. We also talk about what our findings mean for the subject and how they could be useful in real life. We mention any limitations in our research and suggest ideas

[2]http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

for future studies. This chapter gives a nice ending to our whole research journey.

Each of these segments assumes a distinctive and pivotal function in molding the framework and substance of the dissertation, contributing to a comprehensive and unified scholarly composition.

CHAPTER 2

Methodology and Literature Review

2.1. Methodology

The methodology adopted for this study was CRISP-DM (Cross-Industry Standard Process for Data Mining), a widely recognized framework that guides data mining and machine learning projects. CRISP-DM offers a structured approach, breaking down complex problems into manageable steps and ensuring a systematic and well-documented process. The detailed flow of the CRISP-DM methodology is presented in Figures 3.3, 3.1, 3.5, and 4.1 at pages 25, 23, 29, and 39 respectively. This methodology encompasses the following key steps:

(1) **Business Understanding:** In the initial phase, we defined the research objectives and formulated research questions to address the challenges of predicting natural gas prices.

(2) **Data Understanding:** The subsequent step involved data collection and exploration. We obtained and thoroughly examined data on natural gas, crude oil, GDELT news, and weather.

(3) **Data Preparation:** After collecting the data, we performed extensive cleaning and preprocessing, handling missing values, and outliers, and ensuring data quality.

(4) **Modeling:** Various machine learning models, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN), were employed for forecasting natural gas prices. The Keras Tuner was utilized to select the most suitable hyperparameters for the models.

(5) **Evaluation:** The performance of each model was evaluated using metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

By adhering to the CRISP-DM methodology, our research followed a systematic and transparent approach, effectively addressing the challenges of predicting natural gas prices.

2.2. Literature Review

2.2.1. Europe Natural Gas Market

Energy is an important part of social progress and economic development (Kaufmann & Connelly, 2020).

Europe has increased the consumption of natural gas over the years and is transitioning toward a renewable-based energy system (Berrisch & Ziel, 2022).

After the European Union liberalization process in 1998, the market grew fast. These directives issued by the European Commission have given access to infrastructure to third-party companies. Another advantage is creating a Virtual Trading Point, and now natural gas can be traded in any location (Hamie et al., 2020).

The network code was also a game changer because it allows efficient transmission that avoids the congestion in EU gas transmission pipelines is reduced (Hamie et al., 2020).

Investment in cross-border capacity contributes to creating a cross-border relationship between Transmission System Operators via the European Network for Gas long-term contracts. The same investment was also aggregated into the new system, old legacy, or long-term contracts into the new system (Hamie et al., 2020).

The main natural gas hubs in the European Union are presented in Figure 2.1, the red dots indicate the mature hubs with the highest trade rate. Blue dots show active hubs, and yellow dots show hubs with lower trade (Heather, 2021). The two most representative hubs are the Dutch TTF and the British NBP, both classified as mature with 46690 TWh and 10060 TWh in 2020 (Heather, 2021).

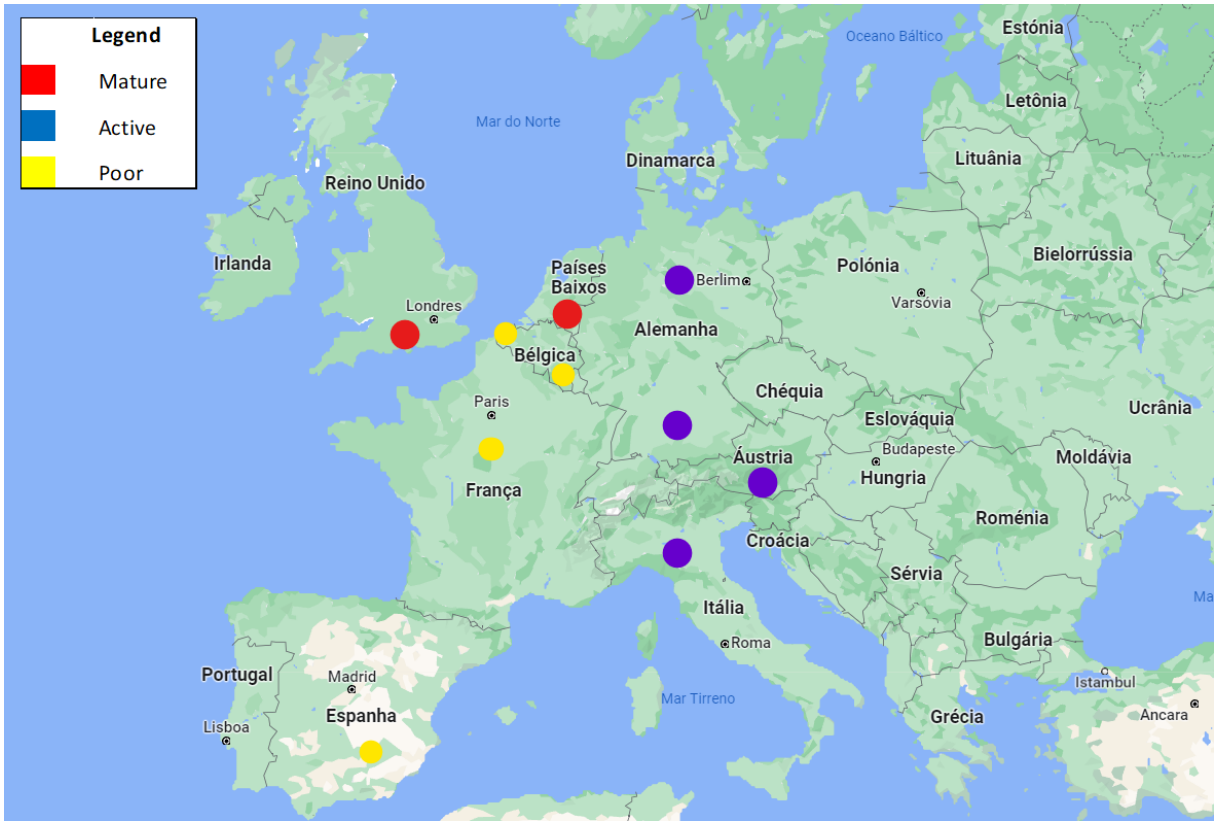

FIGURE 2.1. Map the main natural gas hubs in Europe

The Russian invasion of Ukraine began on Thursday, February 24, 2022. The war has since killed thousands of people, taken the place of millions, and destroyed entire cities (Psaropoulos, 2022).

After the invasion, Russia used the natural gas trade as a weapon, taking advantage of the shortage of European natural gas and the dependence on Russian supply to negotiate

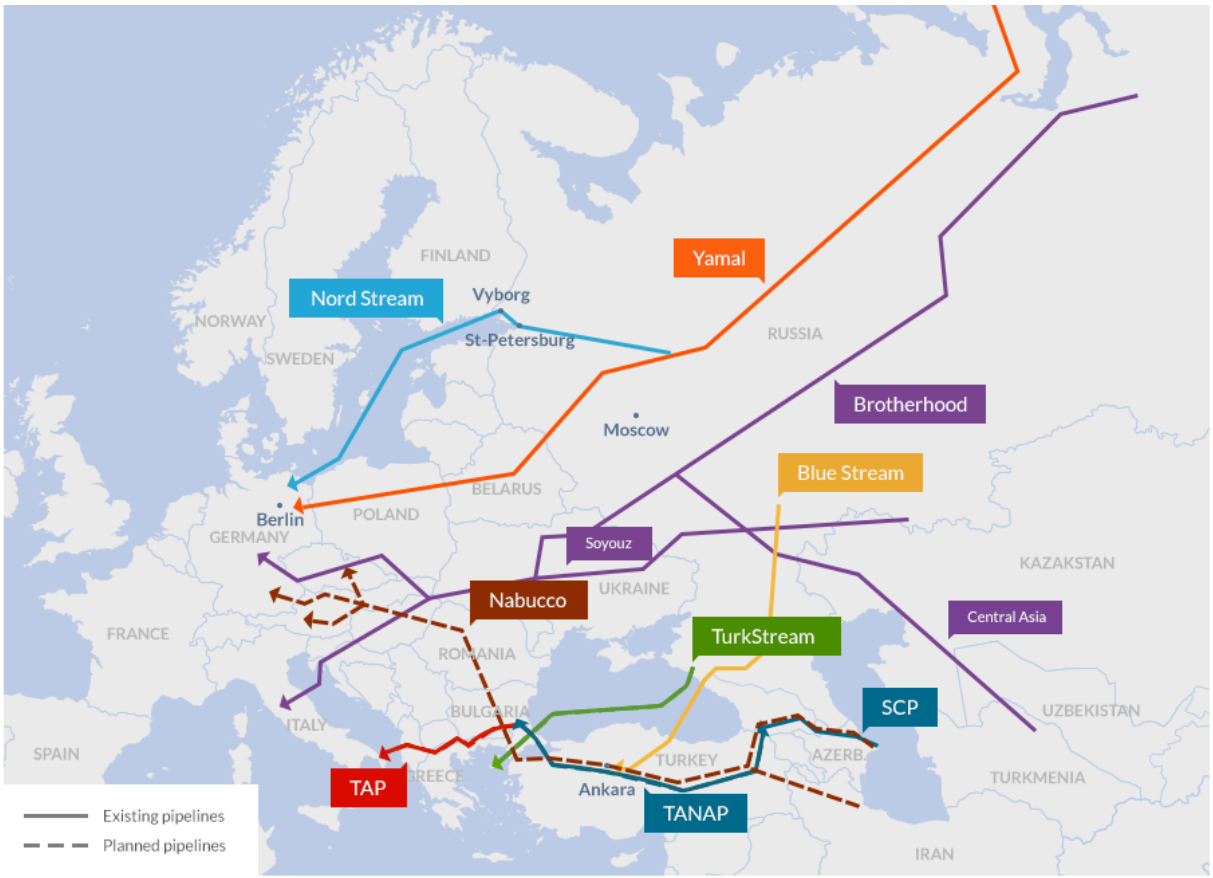

FIGURE 2.2. Map of Gas pipelines between Europe, Russia and Caucasia Sources : Gazprom export ; Gazprom ; Tanap ; Trans Adriatic Pipeline ; BP ; Natural Gaz Europe ; Nord Stream ; South Stream Transport

with European countries not to get involved in the war. In the first 14 days of the invasion, the natural gas price increased by around 180% and Russia started to deliver 60% less gas through the Yamal pipeline (Figure 2.2), this reduction affected imports to France, Austria, Italy, the Czech Republic and Germany. Furthermore, in response to Russia's weaponization of natural gas, the European Union reduced imports through Nord Stream I (Figure 2.2) (*Economic Bulletin Issue 4,* 2022; Halser & Paraschiv, 2022).

The fear of a natural gas shortage in winter made the European Commission propose a new legal obligation to fill underground gas storage to 80% of its capacity by 1 November 2022. Besides, the European Union signed a memorandum for delivering 15 bcm of liquified natural gas with the United States and Qatar (Fabian et al., 2022; *Refilling gas storage for next winter*, 2022).

2.2.2. Forecasting Natural Gas Price

The oldest paper analysed was published in 2010 and the newest was brought out in 2022, in this decade, 2019 was the year with more articles produced, the Figure 2.3 at page 8 presents the number of articles per year.

Table **??** at page **??** shows the most frequent models used to predict natural gas prices found in the literature review. The models are grouped into neural networks, regression,

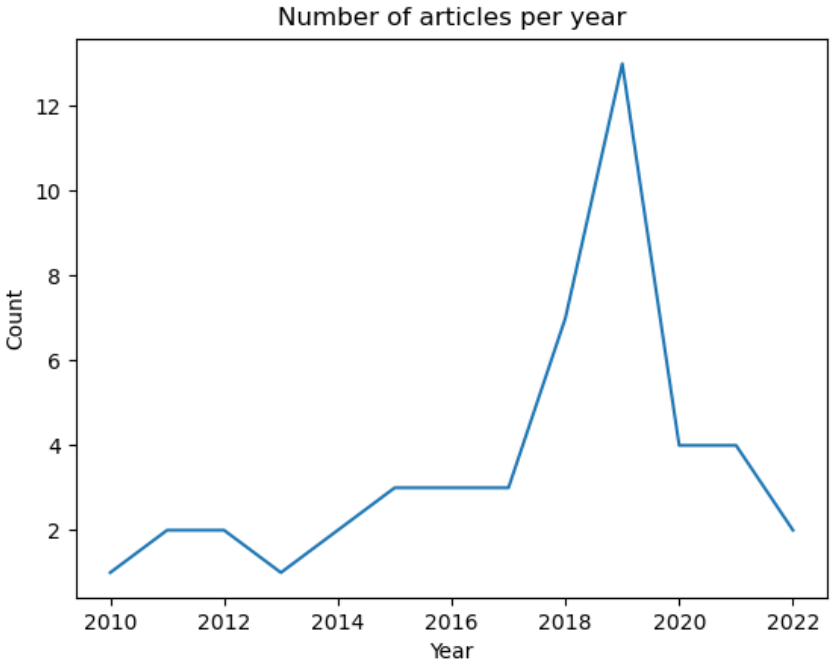

FIGURE 2.3. The number of papers about natural gas prices forecast, per year

TABLE 2.1. Grouped Models in the Literature Review

Grouped Models	Count
Artificial Neural Networks (ANN)	15
Support Vector Regression (SVR)	8
Time Series Models	5
Ensemble Models	3
Regression Models	6
Other Machine Learning Models	8
Other Models	3

auto-regression, decision trees, and other types. Artificial Neural Networks and Support Vector Regression are the most common methods for forecasting natural gas prices as showed in the word cloud in Figure 2.4 at page 9.

The number of hybrid models is slightly more than the traditional method that uses only one model, with 55% and 45% respectively presented on Figure 2.5 at page 9.

The input variables can be a simple natural gas price time series or a list of features (Naderi et al., 2021). After the features related to natural gas, the second most used feature is the input variables linked to Crude oil (Abrishami & Varahrami, 2011; Čeperić et al., 2017; Li et al., 2021; Moting et al., 2019; Naderi et al., 2019; Viacaba et al., 2012). Table 2.2 shows all the features used and counts the occurrences of the exact name. The features are grouped into energy, macroeconomics, weather, and others. The energy group has subgroups such as price, demand, production, consumption, etc.

Carbon, electricity, and natural gas are most affected by meteorological factors (Naderi et al., 2021). The result of Li et al. (2021) presents that the proportion of extremely

FIGURE 2.4. Word cloud of all models used to forecast natural gas price

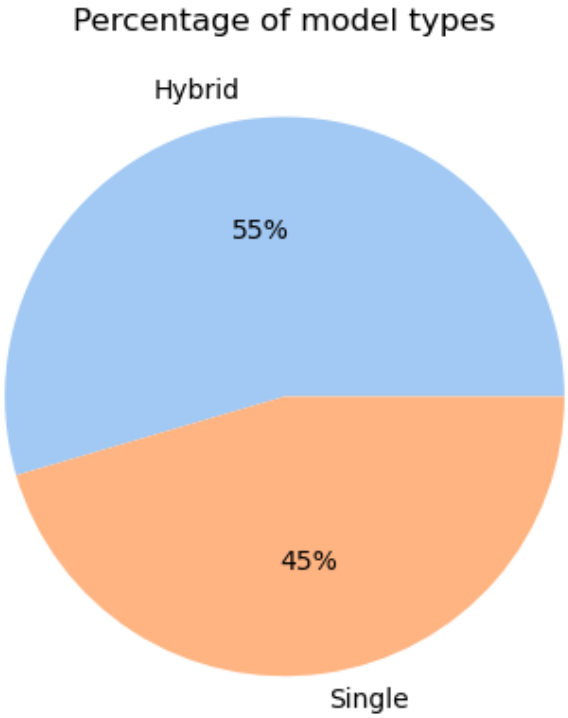

FIGURE 2.5. Percentage of model types

high-temperature weather, the proportion of extremely low-temperature weather, monthly West Texas Intermediate (WTI) crude oil spot prices, natural gas consumption, and natural gas gross withdrawals, all of it come up to predict in different levels of the long-term prices in the Henry Hub natural gas spot, that is the most liquid, but also the most

unpredictable natural gas market in the world (Čeperić et al., 2017). Natural gas prices are insensitive to energy-related and stock-related indexes (Naderi et al., 2021).

Another meaningful connection is between crude oil prices and natural gas prices, the movement of crude oil prices used to influence the natural gas price. Moreover, fluctuations in weather and temperature are used to influence natural gas prices (Moting et al., 2019). The daily total electricity demand in Great Britain has two seasonality effects when it is possible to see less consumption at weekends and a pattern that shows a higher consumption annually during the coldest days of the year (Nguyen & Nabney, 2010).

The analysis of Li et al. (2021) shows that the implementation of natural gas consumption and monthly WTI crude oil prices provides better prediction accuracy for the model that predicts monthly natural gas prices.

News sentiments added as features proved to contain complementary information and can increase the performance of the model by 14.40% compared to a model without the news sentiment (Y. Tang et al., 2019).

Different from the common logic that says "more data is better", Čeperić et al. (2017) finds that for Henry Hub spot prices of natural gas price when it comes to short-term prediction, "less data is better".

The dataset encompassing daily natural gas records presents a substantial variability, spanning from a minimum of 230 data points to a maximum of 5470 data points, as referenced by Naderi et al. (2019) and Siddiqui (2019) respectively. This substantial disparity in the dataset size underscores the diversity in temporal granularity and data availability.

When considering weekly prediction models, Čeperić et al. (2017) employed the dataset with the smallest temporal scope, while Moting et al. (2019a) worked with a significantly larger dataset, containing 886 data points. This wide spectrum in dataset sizes emphasizes the differing preferences and objectives within the field of weekly prediction.

Transitioning to monthly prediction, the dataset sizes exhibit a range of 420 to 2091 records, as reported in studies by Jianwei et al. (2019) and Berrisch and Ziel (2022) respectively. This variation in dataset sizes underscores the distinct temporal resolutions and availability of data within the monthly prediction context.

For the prediction of yearly trends, the sole study conducted by Azadeh et al. (2012) utilized a dataset comprising 40 data points. This notably limited dataset size reflects the challenges inherent in yearly prediction due to the scarcity of available observations.

For a comprehensive overview of these data sizes, refer to Table 5.1, which encapsulates the aforementioned ranges and sizes, encapsulating the diversity and nuances present in the datasets utilized across various prediction timeframes.

For all types of energy prediction studied in the literature review by Naderi et al. (2021) the application of feature engineering led to an average increase of 54.59% in accuracy in the models analyzed (see Table 2.3 at page 12).

TABLE 2.2. List of features used to forecast natural gas price

Input variable(s)	Count
Historical gas price	22
Crude oil price	2
Heating oil price	2
Annual interest rate	1
Proportion of extreme high temperature weather	1
Natural gas price differences	1
Natural gas price rotary rigs	1
Natural gas total consumption	1
Natural gas underground storage volume	1
News	1
OPEC cut production	1
Population	1
Taxes placed on gas price	1
Natural gas imports	1
Temperature	1
Total renewable energy consumption	1
U.S. LNG imports	1
U.S. natural gas gross withdrawals	1
U.S. natural gas marketed production	1
U.S. natural gas pipeline imports	1
U.S. natural gas total consumption	1
WTI crude oil prices	1
WTI crude oil prices differences	1
Natural gas marketed production	1
Natural gas gross withdrawals	1
Natural gas consumption	1
Annual natural gas consumption	1
Coal price	1
Coal price differences	1
Consumer price index	1
Cooling degree-days	1
Economic parameters	1
Electricity price	1
Environmental policy	1
GDP	1
Global demand for crude oil	1
Global demand for gas	1
Heating degree-days	1
Heating oil price difference	1
Historical data of energy demand	1
Internet search	1
Meteorological parameters	1
Monthly WTI crude oil prices	1
Monthly oil price	1
WTI oil spot price	1

Table 2.3. List of feature engineering methods provided on models

feature engineering method(s)	Count
Ensemble Empirical Mode Decomposition (EEMD)	2
Variational Mode Decomposition (VMD)	2
Discrete Wavelet Decomposition (DWD)	1
Feature Selection (FS)	1
Group Method of Data Handling (GMDH)	1
Improved Pattern Sequence Similarity Search (IPSS)	1
Independent Component Analysis (ICA)	1

Table 2.4. List of optimization methods provided on models

Optimizer(s)	Count
Particle Swarm Optimization (PSO)	3
Adaptive Learning Strategy (ALS)	1
Bat Algorithm (BA)	1
Genetic Algorithm (GA)	1

The most common optimization method applied is Particle Swarm Optimization (PSO), which shows an increase in the performance of final models (Čeperić et al., 2017; Li et al., 2021; J. Wang et al., 2021). In our study, we utilized the random search method to find the optimal configuration for our model. We focused on three critical hyperparameters: the number of layers, units, and epochs. The epoch value was consistently set to 20, and we implemented early stopping after 5 epochs without improvement. To explore the impact of the layer count, we conducted trials with a range of one to five layers, increasing by one layer for each attempt. For the units, we varied the values between 32 and 512, with increments of 32 units for each trial.
In the second trial using Keras Tuner, we refined our search based on the best-performing models from the previous round. We restricted the number of layers to either one or two, and the units were limited to the range of 32 to 512, maintaining the same increment value.

It is possible to see in Figure 2.6 that the most common period to forecast is the daily price of natural gas, on which 14 of the articles are working. The second period is the monthly period for 7 papers. The third period is the weekly period with six publications. The last one with only one paper is the yearly period. Researchers do not predict the natural gas price with a horizon of two months, quarters, and semesters.

The single model most used to forecast the price of natural gas are artificial neural networks (ANN), and the second model most used is auto-regressive moving average (ARMA) and support vector regression (SVR).

The idea behind the hybrid model is to combine more than one model to get better performance. Table 2.6 at 16 lists all the models used to develop hybrid models.

The combination varies between two and five models, which can be different or similar, Naderi et al. (2019) worked in a combination of four models of least squares support

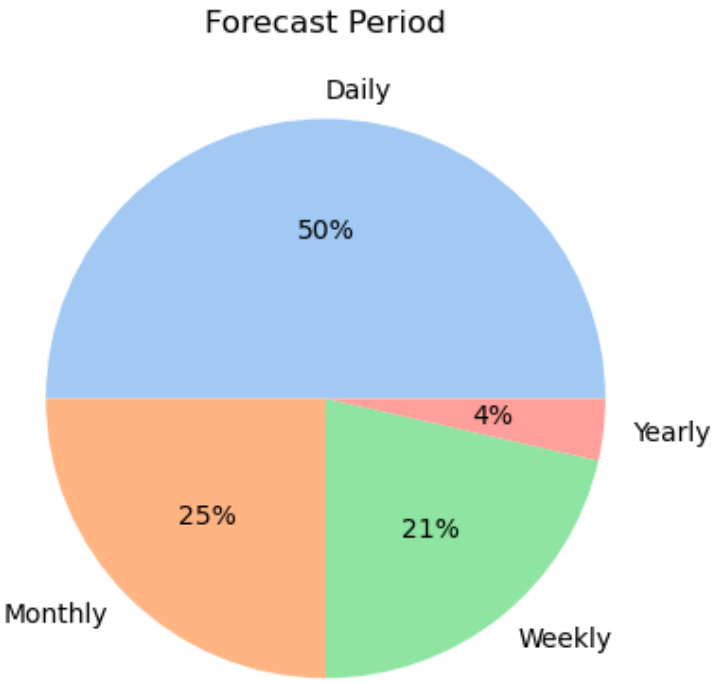

FIGURE 2.6. Forecast period horizon

FIGURE 2.7. Word cloud of single models used to forecast natural gas price

vector machine (LSSVM), genetic programming (GP), artificial neural networks (ANN), and auto-regressive integrated moving average (ARIMA) to predict the price of oil, the annual interest rate, and the daily price of gas.

The study of Li et al. (2021) and J. Wang et al. (2020) compare the hybrid model with a single model and find that the hybrid model performs better. J. Wang et al. (2020)

Table 2.5. List of single models

Models	Count
Artificial Neural Networks (ANN)	3
Auto-Regressive Moving Average (ARMA)	2
Support Vector Regression (SVR)	2
Back-Propagation Neural Networks (BPNN)	1
Dynamic Local Linear Regression (DLLR)	1
Extreme Learning Machine (ELM)	1
Gamma test (GT)	1
Gaussian Process Regression (GPR)	1
Autoregressive Neural Networks (ARNN)	1
Group Method of Data Handling (GMDH)	1
Least Square Support Vector Machine (LSSVM)	1
Least squares Regression Boosting (LSBoost)	1
Local Linear Regression (LLR)	1
Random Kitchen Sink (RKS)	1
Random Vector Functional Links (RVFL)	1
Support Vector Machines (SVM)	1
Gradient boosting machines (GBM)	1

complement with hybrid models when combined with different time-series methods tend to have a better performance. The hybrid model of Li et al. (2021) is a combination of a Deep Belief Network (DBN) with feature engineering of variational mode decomposition (VMD) and particle swarm optimization (PSO).

The study of Čeperić et al. (2017) compared hybrid model applies Support Vector Regression (SVR), Steepwise (SW), and Feature Selection (FS) with another hybrid model that uses Artificial Neural Networks (ANN), Particle Swarm Optimization (PSO), and Feature Selection (FS), and find out that first has a better performance.

The article presented by Jin and Kim (2015) shows that not all hybrid models perform better. The combination of artificial neural networks with wavelet decomposition does not improve the model when compared with Artificial Neural Networks without wavelet decomposition. The other experiment with Auto-Regressive Integrated Moving Average with Wavelet decomposition in the same study demonstrated only a small improvement. But the combination of Auto-Regressive Integrated Moving Average with Wavelet decomposition created the best case for a four-step forecast (Jin & Kim, 2015).

2.2.3. GDELT

Global Database of Events, Language, and Tone (GDELT) (https://www.gdeltproject.org/) is a platform that scans news media as printed media, broadcast, and web formats. GDELT creates a database that saves important information in more than 100 languages. The managed data links between every person, organization, location, theme, news source, and event in each corner of the planet. The sentiment extracted from this massive database can be precious in finding the world's feelings (GDELT Project, n.d.).

FIGURE 2.8. Word cloud of hybrid models used to forecast natural gas price

GDELT is an open platform for research and analysis available for unlimited and unrestricted use for academic, commercial, or government without a fee. The data can be accessed by API from the GDELT website or via the Google Cloud Platform (GCP) (Google Cloud Platform Blog, 2014).

The GDELT project has two versions of event databases, the "GDELT 1.0 Event Database" and the "GDELT 2.0 Event Database". Version 1.0 starts in 1979 through March 31, 2013, and it was updated daily and does not include events reported in the 65 live translated languages. The latest version starts April 1, 2013, has new features, updates every 15 minutes, and includes events reported in articles published in 65 live translated languages. Furthermore, the project GDELT also has the Global Knowledge Graph data source to analyze images, and other data sets normalized, such as "GDELT 1.0 Event Database Normalization Files" (GDELT Project, n.d.).

The GDELT project has been used in many areas of knowledge. Kwak and An (2014) showed the structure of global news coverage of disasters and their cause. The article finds strong regionalism in the geography news.

To detect events of occupied protests, Qiao et al. (2015) compared the results using a base model using the GDELT database. Models with GDELT features proved to be better with higher accuracy. Hammond and Weidmann (2014) used GDELT to study political violence.

The literature review applied by W. Wang et al. (2016) compared the GDELT project with other similar databases. The study anticipates that there should be a high correlation between GDELT and another database, but the overall correlation does not match the expectations, returning a small correlation. But when the comparison is filtered in

Table 2.6. List of hybrid models

Reference / Abbreviations	Models
Nguyen and Nabney, 2010	Wavelet Transform (WT) Radial Basis Functions Neural Networks (RBFNN) Linear Regression (LR) Group Method of Data Handling (GMDH)
Abrishami and Varahrami, 2011	Group Method of Data Handling (GMDH) Genetic Algorithm (GA) Rule-based Exert System (RES)
Azadeh et al., 2012	Artificial neural networks (ANN) Fuzzy linear regression (FLR) Conventional regression (CR)
Thakur et al., 2015	Moving Average Neural Networks (MANN) Back-Propagation Neural Networks (BPNN)
Jin and Kim, 2015	Discrete Wavelet Decomposition (DWD) Auto-Regressive Integrated Moving Average (ARIMA) Artificial neural networks (ANN)
Dey and Salem, 2017	Gated Recurrent Unit (GRU) Recurrent neural networks (RNN)
Čeperić et al., 2017	Strategic seasonality-adjusted (SSA) Support vector regression machines (SVR) Neural networks (NN) Feature selection (FS) Particle Swarm Optimization (PSO)
Jianwei et al., 2019	Variational Mode Decomposition (VMD) Independent Component Analysis (ICA) Gated Recurrent Unit Neural Networks (GRUNN)
Qin et al., 2019	Ensemble Empirical Mode Decomposition (EEMD) Local Linear Prediction (LLP)
Naderi et al., 2019	Bat Algorithm (BA) Least Square Support Vector Machine (LSSVM) Genetic Programming (GP) Artificial Neural Networks (ANN) Auto-Regressive Integrated Moving Average (ARIMA)
J. Wang et al., 2020	Improved Pattern Sequence Similarity Search (IPSS) Support Vector Regression (SVR) Long-term and Short-term Memory Networks (LSTM)
Li et al., 2021	Variational Mode Decomposition (VMD) Particle Swarm Optimization (PSO) Deep Belief Networks (DBN)
J. Wang et al., 2021	Complete ensemble empirical mode decomposition (CEEMD) Adaptive noise-sample entropy (AN-SE) Particle Swarm Optimization (PSO) Adaptive learning strategy (ALS) Gated Recurrent Unit (GRU)

each database by country, the experience results in a better correlation.

To analyze public opinion on the energy policy of the Spanish government, Bodas-Sagi and Labeaga (2016) shows a negative feeling about the solar energy policy introduced in 2016.

To predict social unrest events, studies are applying several models such as the hidden Markov model, neural networks, random forest, LSBoost, LSTM, and others (Galla & Burke, 2018; Qiao et al., 2017).

TABLE 2.7. List of results of single models to predict daily natural gas price

Reference	Performance
L. Tang et al., 2018 EEMD-based model	MAPE=0.5850
Moting et al., 2019a LSBoost	MAE=0.4493 MSE=0.4376 RMSE=0.6615 R2=0.91
Berrisch and Ziel, 2022 ARMA	MAE=0.3863 CRPS=0.2834 RMSE=1.0843
Salehnia et al., 2013 LLR / DLLR / ANN	LLR (t) MSE=0.29113 DLLR (t) MSE=0.13977 ANN (t) MSE=0.3366
Al-Sharoot and Alramadhan, 2019 ARMA / GMDH	MAE=0.01539 MSE=0.0214
Y. Tang et al., 2019 ANN	MAE = 0.0956; 0.1002; 0.0987; 0.0902 RMSE = 0.1368; 0.137; 0.133; 0.1284
Hu and Trafalis, 2011 SVR	MSE=0.0903 R2=0.9822
Siddiqui, 2019 ARNN	MSE=0.026

The GDELT project is also successfully employed by Bourgeois et al. (2018) to identify bias in news with success. In the financial field, the GDELT project is used to help predict political crises, oil prices, stock market, and macroeconomic index with considerable improvement (Alamro et al., 2019; Elshendy et al., 2018; Tilly et al., 2020; Zhang et al., 2019).

No paper used the GDELT project to predict the price of natural gas, Y. Tang et al. (2019) used news but from a different source with a single model to predict the price of natural gas.

2.2.4. Results

Table 2.7 presents results of the prediction of the natural gas price by single models. The articles use different performance metrics making the MSE the most common. The best result for daily prediction using a single model is MSE equal to 0.0214 by Al-Sharoot and Alramadhan (2019) with an ARMA-GARCH model.

The results of single models that forecast natural gas prices weekly are presented in Table 2.8 with just a few articles, and the best precision comes from the study of Salehnia et al. (2013) with an MSE equal to 0.25566 by applying the dynamic local linear regression model (DLLR).

Table 2.8. List of results of single models to predict weekly natural gas price

Reference	Performance
Moting et al., 2019a LSBoost	MAE=0.4761 MSE=0.5116 RMSE=0.7153 R2=0.9
Salehnia et al., 2013 LLR / DLLR / ANN	LLR (t) MSE=3.4317 DLLR (t) MSE=0.25566 ANN (t) MSE=0.8268
Viacaba et al., 2012 SVR	RMSE <0.03

Table 2.9 shows the articles that predict natural gas prices monthly, and the best result is obtained by an Artificial Neural Networks with an MSE = 0.5663 (Moting et al., 2019).

The only article that predicts yearly natural gas prices applies a single model of Artificial neural networks (ANN), fuzzy linear regression (FLR), and conventional regression (CR). These models archive the best result from conventional regression with MAPE = 0.260. Table 2.10 shows all results.

The most widely used evaluation metric is the root mean square error (RMSE). Table 2.11 shows all models that forecast daily natural gas prices using different hybrid models. The most accurate prediction is given by a combination of Ensemble Empirical Mode Decomposition (EEMD) and Local Linear Prediction (LLP), resulting in an RMSE = 0.035 developed (Qin et al., 2019).

The list of articles that predict the price of natural gas weekly is in Table 2.12. The best result is proposed by Jin and Kim (2015) using a hybrid model of discrete wavelet decomposition (DWD) and artificial neural networks (ANN) with RMSE = 0.1278 precision.

For a monthly forecast of natural gas prices, Table 2.13 lists three articles, the one with the best results has a MAPE between 0.001691 and 0.00413. The models that meet this precision are the combination of Variational Mode Decomposition (VMD), Independent Component Analysis (ICA), and Gated Recurrent Unit Neural Networks (GRUNN), applied by Jianwei et al. (2019).

TABLE 2.9. List of results of single models to predict monthly natural gas price

Reference	Performance
Moting et al., 2019 ANN/SVM/GBM/GPR	ANN R2=0.8904 MAE=0.5115 MSE=0.5363 RMSE=0.7247 MAPE=0.1117 SVM R2=0.8437 MAE=0.5673 MSE=0.7673 RMSE=0.8757 MAPE=0.1202 GBM R2=0.8006 MAE=0.6490 MSE=0.9786 RMSE=0.9888 MAPE=0.1366 GPR R2=0.8374 MAE=0.6026 MSE=0.7980 RMSE=0.8932 MAPE=0.1270
Moting et al., 2019a LSBoost	MAE=0.6859 MSE=1.1166 RMSE=1.0567 R2=0.78
Berrisch and Ziel, 2022 SVR	CRPS=0.2126 MAE=0.3010 RMSE=0.3995
Salehnia et al., 2013 ARMA	Monthly LLR (t) MSE=3.864 DLLR (t) MSE=2.5932 ANN (t) MSE=0.9831

TABLE 2.10. List of results of hybrid models to predict yearly natural gas price

Reference	Performance
Azadeh et al., 2012 ANN-FLR-CR	MAPE (Average) CR = 0.2260 ANN = 0.2978 FLR = 0.2470

TABLE 2.11. List of results of hybrid models to predict daily natural gas price

Reference	Performance
Qin et al., 2019 EEMD-LLP	RMSE = 0.035 MAPE = 0.01244 Dstat = 0.908
J. Wang et al., 2020 IPSS-SVR-LSTM	MAPE = 0.0555 MER = 0.0549
Abrishami and Varahrami, 2011 GMDH-GA-RES	Dstat ¿ 0.7 RMSE ¡ 2.942
Thakur et al., 2015 MANN / BPNN	MSE ¡ 0.1
Naderi et al., 2019 BA-LSSVM-GP-ANN-ARIMA	R2 = 0.9611 RMSE = 0.06
Čeperić et al., 2017 SSA-SVR/NN-FS-PSO 5 variables	SVR SW (Steepwise) MAPE = 0.221 RMSE = 0.1401 10 variables SVR SW (Steepwise) MAPE = 0.218 RMSE = 0.1375

TABLE 2.12. List of results of hybrid models to predict weekly natural gas price

Reference	Performance
Jin and Kim, 2015 DWD-ANN / DWD-ARIMA	Wavelet with ANN MAE = 0.0985 MAPE = 0.033747 RMSE = 0.1278 Wavelet with ARIMA MAE = 0.1112 MAPE = 0.037018 RMSE = 0.1366
J. Wang et al., 2021 CEEMDAN-SE-SO-ALS-GRU	Dstat = 0.519 MAE = 0.114 MSE = 0.025 RMSE = 0.158 R2 = 0.889
Čeperić et al., 2017 SSA-SVR / NN-FS-PSO 5 variables	SVR SW (Steepwise) MAPE = 0.423 RMSE = 0.2904 10 variables SVR SW (Steepwise) MAPE = 0.431 RMSE = 0.2782

TABLE 2.13. List of results of hybrid models to predict monthly natural gas price

Reference	Performance
Jianwei et al., 2019 VMD–ICA–GRUNN–SVR	Dstat = 0.730159-0.845238 MAD = 0.0201-0.0776 MAPE = 0.001691-0.00413 RMSE = 0.0407-0.1196 R2 = 0.95-0.991
Li et al., 2021 VMD-PSO-DBN	MAE = 0.125 MAPE = 0.0481 RMSE = 0.082 FLR = 0.2470
Nguyen and Nabney, 2010 WT-RBFNN-LR-GARCH	MAE = 0.01699 MAPE = 2.019 MSE = 0.15384

CHAPTER 3

Exploratory Data Analysis and Preprocessing

3.1. Exploratory Data Analysis

We utilized four primary data sources for our analysis, namely natural gas price, news data, weather data, and crude oil price. The integration of these features to the main dataset was done individually, the flow of the ETL process is illustrated in Figure 3.3 at page 25. The news data, extracted from the GDELT project, underwent a separate integration process. A Spark environment was created due the number of lines to be processed, the first step of the flow was a extraction of the raw data from GDELT API, after that we analyzed, cleaned, trandformed, and evaluated the news data, last step we exported the data into a parquet data format to optimize the size of our dataset (Figure 3.1). The final integration step involved combining these two features with weather data and crude oil price at the same level, as shown in Figure 3.5.

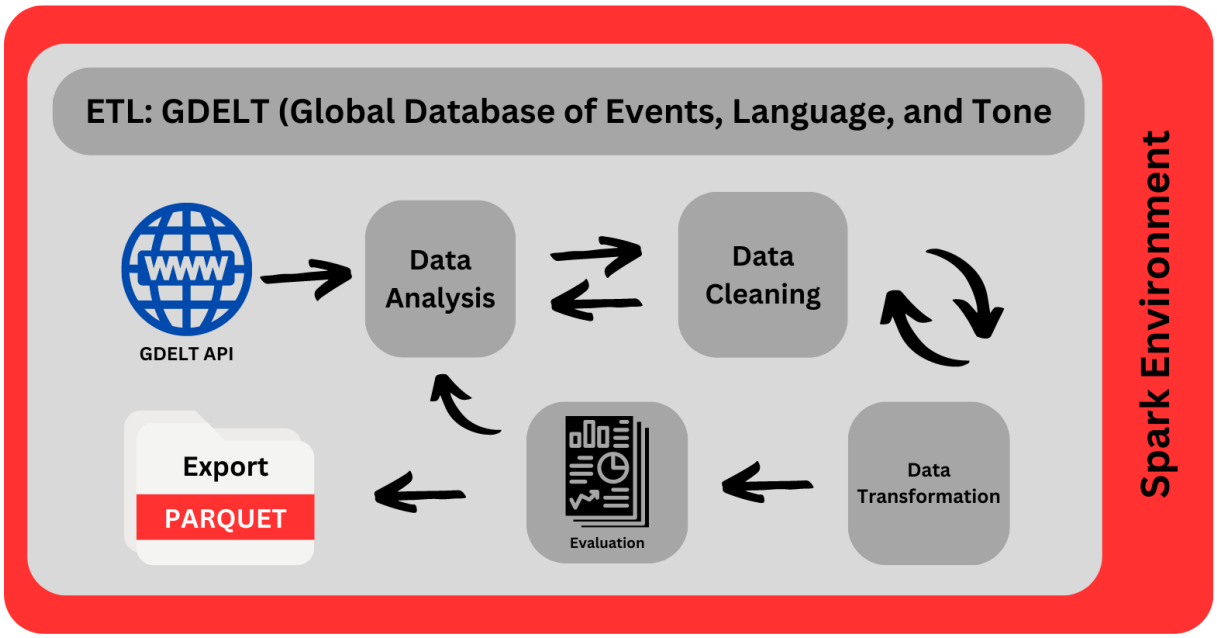

FIGURE 3.1. ETL process for extraction, analysis, cleaning, transformation, and evaluation of news data from GDELT API, with exportation to a parquet file.

Multiple trading hubs exist for natural gas prices, and for our forecasting purposes, we chose to focus on the Dutch Title Transfer Facility (TTF) hub price like Berrisch and Ziel (2022). We obtained the necessary data using the Yahoo Finance API through the yfinance Python library. The description of each column can be found in Table 3.1, and Figure 3.2 shows natural gas prices, and 3.4 presents the natural gas price volume.

The dataset of TTF natural gas prices spans from October 23, 2017 and has been updated daily. For our study, we focused on the period between January 2, 2018, and December 30, 2022, which encompasses a total of 1260 daily natural gas prices. However, it is important to note that there are 32 missing values within this time frame, resulting in a total of 1292 days included in our analysis.

The initial analysis involved performing a statistical summary on Open, High, Low, Volume, Dividends, Stock Splits, and Close columns, as presented in Table **??**. Through this analysis, we observed that both the Dividends and Stock Splits variables exhibited a consistent value of zero across rows.

After conducting the statistical analysis and examining the time-series visualization on natural gas price, during this analysis columns Open, High, Low, and Volume did not match the correlation criteria between -0.70 and 0.70 and do not have strong Granger Causality, and Dividends and Stock Splits all values is equal to zero, is this case we made a decision to retain only the "Close" column as variable and target.

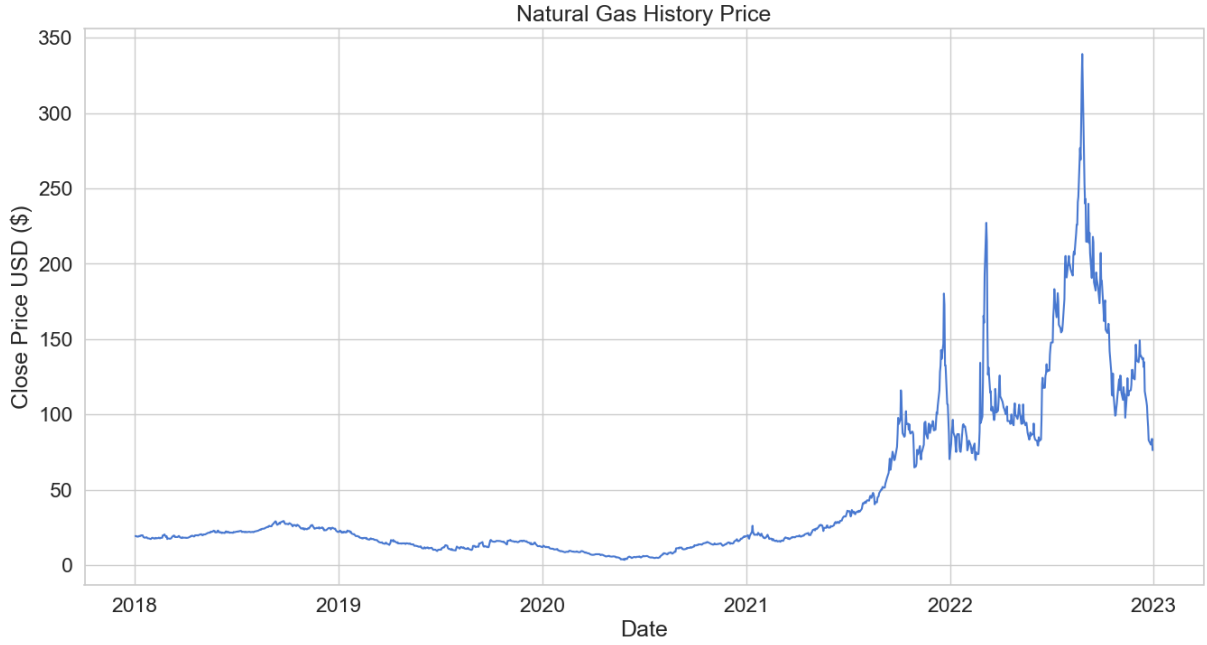

FIGURE 3.2. Natural gas price time-series.

TABLE 3.1. Natural Gas Price TTF Data Columns

Column	Description
Open	The opening price of the natural gas
High	The highest price of the natural gas during the day
Low	The lowest price of natural gas during the day
Volume	The trading volume of the natural gas
Dividends	Any dividends issued for the natural gas
Stock Splits	Any stock splits that occurred for the natural gas
Close	The closing price of the natural gas

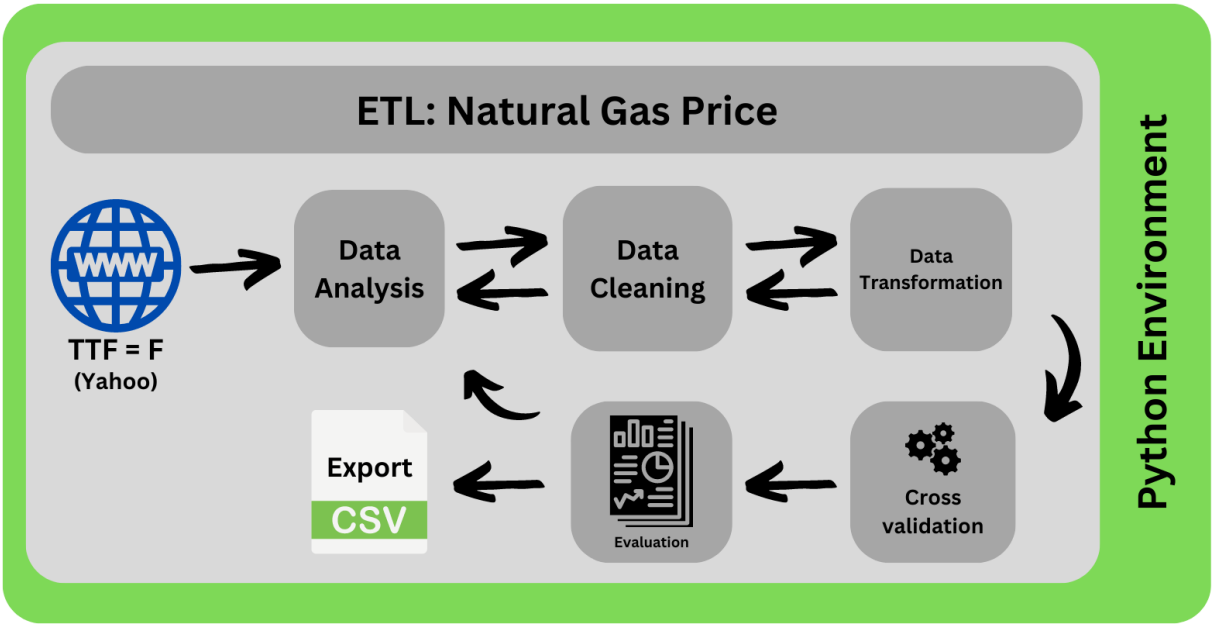

FIGURE 3.3. ETL process for extraction, analysis, cleaning, transformation, cross-validation, and evaluation of TTF natural gas price data from Yahoo API, with Exportation to a CSV file.

3.1.1. GDELT - Global Database of Events, Language, and Tone

In order to handle the large volume of data, we utilized a Spark environment and developed a function to extract news data from the GDELT API. The function includes a filtering mechanism to retrieve only the news articles based on specific Cameo codes from Event Data Project, Department of Political Science, Pennsylvania State University (March 2012), as listed in Table 3.2. Subsequently, the extracted news data was saved in a Parquet format for further processing. The entire process of integrating the news data extracted from the GDELT API is illustrated in Figure 3.1.

The total number of lines obtained from the extraction process amounted to 18,700,299, comprising 58 columns. These columns can be categorized into different types, The main topic of the news is actors, for example, the Actor of news that informs about the deal of European Union with US to deliver natural gas price, the Actor one is European Union and Actor two is US. Actor one and two columns have informations about the name, country, ethnic, religion, and location, shown in Tables 5.2 and 5.3. Geographic information is summarized in Table 5.4. Numeric variables store numbers about the news, as the tone of the news, number articles related to the event, and number of sources reporting the event, all numeric variables of GDELT are described in Table 5.5, and columns slated for deletion can be found in Table 3.3.

3.1.2. Crude Oil Price

The data source of crude oil price was obtained from Investing.com (Accessed on July 19, 2023) with the same period of natural gas price. The structure of the columns is

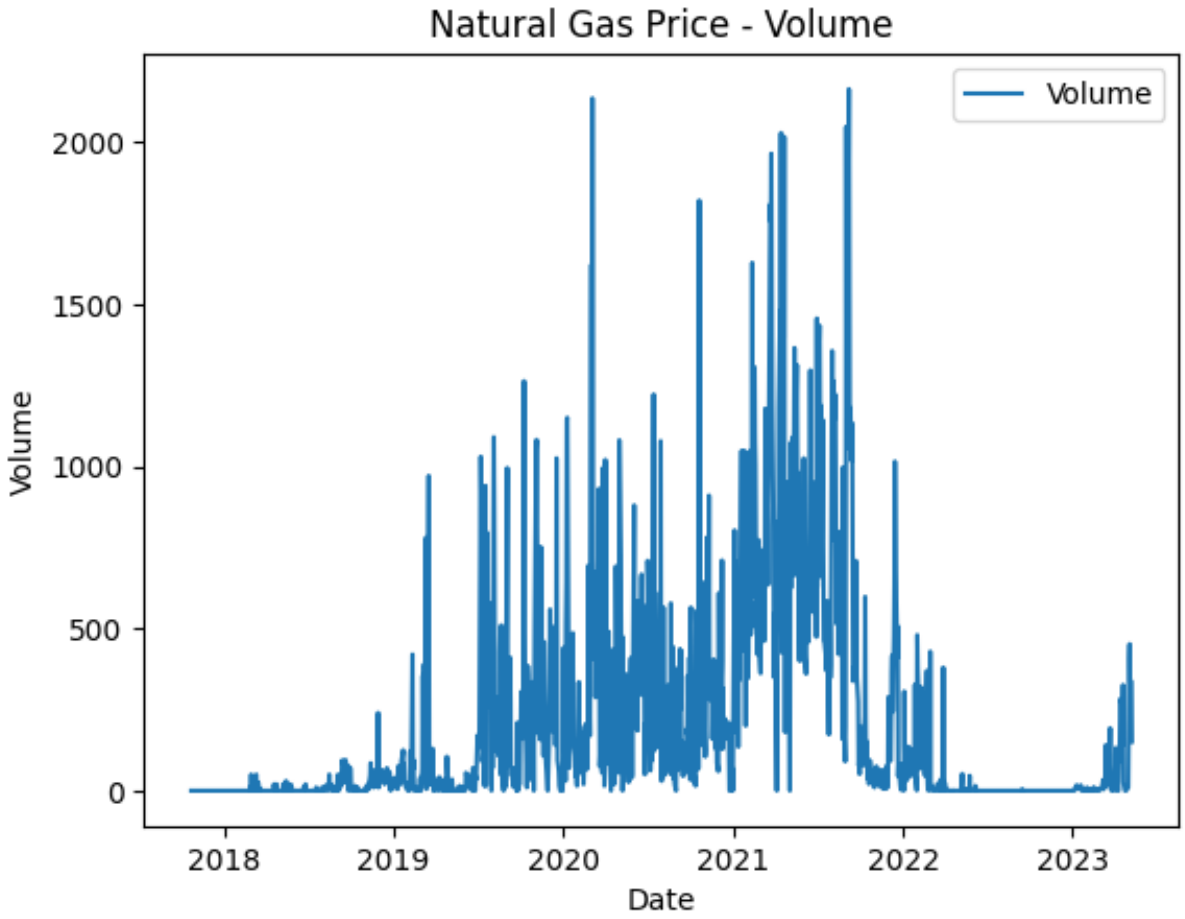

FIGURE 3.4. Visualization of natural gas prices volume.

TABLE 3.2. Cameo event root codes used.

Event Root Code	**Event Root Name**
13	Threaten
14	Protest
15	Exhibit force posture
16	Reduce relations
19	Fight
20	Use unconventional mass violence

very similar to the natural gas price, with Open price, High price, Low price, and Change percentage price compared with previous date, as can be see in Table 3.5 Following the completion of the statistical analysis and a thorough examination of the time-series visualization for crude oil time series, it was observed that the columns labeled Open price, High price, Low price, and Change percentage price did not meet the correlation criteria falling within the range of -0.70 to 0.70. Furthermore, these columns did not exhibit significant Granger Causality and based on this we only used the column Price_oil as feature.

Table 3.3. Deleted columns from GDELT.

Column Name	Description
MonthYear	Month and year of the event
Year	Year of the event
FractionDate	Fractional date representation of the event
EventCode	Code representing the event category
Actor1Geo_FeatureID	Feature ID of the geographic location for Actor 1
Actor2Geo_FeatureID	Feature ID of the geographic location for Actor 2
ActionGeo_FeatureID	Feature ID of the geographic location for the action
DATE ADDED	Date the record was added
SOURCEURL	URL of the source for the event

Table 3.4. Sum of null values for each GDELT column that has null values.

Column Name	Sum of Null
IsRootEvent	18604188
Actor1Geo_Lat	2274519
Actor1Geo_Long	2271529
Actor2Geo_Lat	5505621
Actor2Geo_Long	5503481
AActionGeo_Lat	563853
ActionGeo_Long	560217

Table 3.5. Description of Columns in Crude Oil Price Data

Column	Description
Dates	The dates corresponding to the crude oil price observations
Price_oil	The closing price of crude oil for the given date
Open_oil	The opening price of crude oil for the given date
High_oil	The highest price of crude oil reached during the date
Low_oil	The lowest price of crude oil reached during the date
Change_%_oil	The percentage change in crude oil price compared to the previous date

3.1.3. Weather

The weather dataset used in this study was obtained from NCEI (Accessed on July 19, 2023), specifically from the Rhein-Main station located in Germany. The dataset includes various weather measurements such as precipitation, snowfall, snow depth, maximum temperature, minimum temperature, and average temperature. However, for our analysis, we focused solely on the average temperature time series as it was the relevant variable for our study. The detailed description of each column can be found in Table 3.6.

3.1.4. Data Integration

The integration of all the features was conducted within a Spark environment due to the voluminous nature of the data being processed. The natural gas price, news data, and other relevant features were combined by joining them together into a Spark data frame.

TABLE 3.6. Description of Columns in Weather Data

Column	Description
PRCP	Precipitation (Rainfall)
SNOW	Snowfall
SNWD	Snow Depth
TMAX	Maximum Temperature
TMIN	Minimum Temperature
TAVG	Average Temperature

TABLE 3.7. Sum of null values for each weather column that has null values.

Column Name	Sum of Null
Prcp_temp	60
Prcp_temp	1824
Snow_temp	60
Wind_temp	60
Max_temp	60
Min_temp	60
Avg_temp	3

TABLE 3.8. Days with average temperature null.

Dates	Prcp_temp	Snow_temp	Wind_temp	Max_temp	Min_temp
20210211	0	null	10	-9	-9.7
20210212	0	null	10	-5	-7.3
20210211	0	null	0	16.5	8

Following this integration, a series of steps were executed, including analysis, data cleaning, transformation, and evaluation. As a result, a comprehensive file containing all the features was generated and saved for further analysis. For a visual representation of this integration process, please refer to Figure 3.5.

3.2. Preprocessing

In order to preprocess the dataset, several steps were undertaken to optimize the data before training the model. The preprocessing began with addressing missing values by employing backward and forward fill, the next method was linear, cubic, and quadratic interpolation, another one was imputation of mean value, and the last one was seasonal decomposition, that we replaced the null values by trend and seasonal. Subsequently, the data was categorized, with textual information transformed into index codes. Correlations and causality between variables were identified to filter and select the correlated features between -0.7 and 0.7. Outliers were also detected and handled, by taking the mean value of before and after data of time series. Lastly, statistical and mathematical operations were performed to aggregate the data on a daily basis. These preprocessing steps aimed

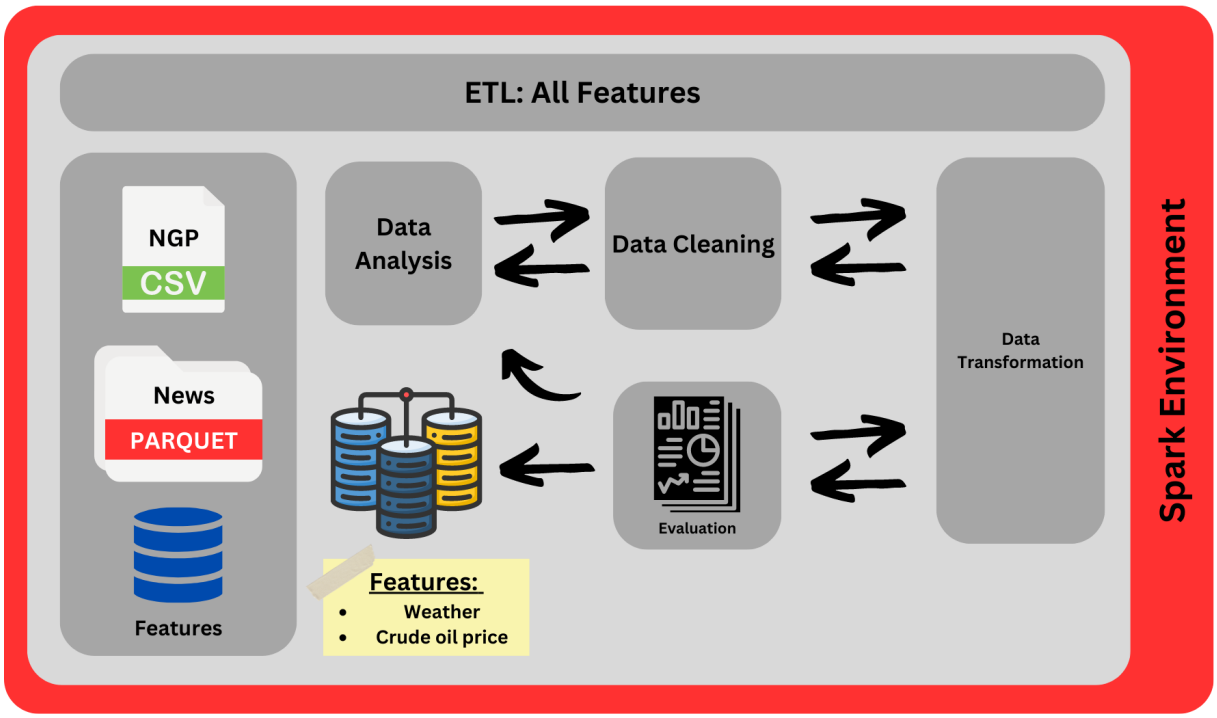

FIGURE 3.5. ETL process for analysis, cleaning, and integration of features into a parquet file.

to ensure the quality and preparedness of the dataset for subsequent modeling tasks.

3.2.1. Replacement of Null Values

For natural gas prices, a unique scenario arises where the identification of missing data points is not solely based on their absence. Instead, an additional step is taken to determine if these null values coincide with weekends or holidays. Through this analysis, it was revealed that there were 32 missing values that required further handling. Several methods were then applied to impute these missing values, aiming to obtain the most accurate representation of the actual values.

The first approach employed for filling the missing values was backward and forward filling techniques. Backward filling involved propagating the last observed value backward in time to fill the missing values, ensuring a smooth and continuous representation of the data. This method is illustrated in Figure 5.11. Similarly, forward filling was applied to propagate the next observed value forward in time to fill any remaining missing values. The application of forward filling can be visualized in Figure 5.12 (van Buuren, 2012).

In addition to backward and forward filling, another method used for imputing missing values was interpolation. Cubic interpolation is shown in Figure 5.13, linear interpolation is shown in Figure 5.14, and quadratic interpolation is shown in Figure 5.15. The visual analysis indicates a satisfactory evolution of the time series after applying these interpolation techniques (van Buuren, 2012).

Another method employed was seasonal decomposition filling. Seasonal decomposition

with the trend is depicted in Figure 5.19, while seasonal values are presented in Figure 5.17. However, it was observed that this method did not provide a natural substitution for the null values. Therefore, a mean value substitution was also utilized, as shown in Figure 5.16 applying the value for time-series (van Buuren, 2012).
In Figure 3.6, it is evident that many of the applied methods did not result in a natural filling of the null values. To gain a better understanding of the filled values, a closer look is provided in Figures 3.7 and 3.8, which depict a specific slice of the data with a null value in the middle. These zoomed-in plots clearly show that the mean and seasonal decomposition methods do not naturally fill the missing values, as the substituted values noticeably differ from the surrounding data points.
To determine the most suitable method for handling missing values in the natural gas price dataset, we conducted a correlation analysis. The results of this analysis, as presented in Table 3.9, indicated that the mean and seasonal decomposition with trend methods performed well. These results suggest that the natural gas price exhibits a tendency towards the mean value and a seasonal pattern. However, instead of choosing one of these methods, we decided to use forward filling. This decision was based on the observation that forward filling effectively fills the null values without introducing significant deviations from the surrounding data points.

TABLE 3.9. R-square results of cross-validation for all methods used to fill null values of natural gas price.

Methods	**R2**
Mean	0.999298
Seasonal Decompose: Trend	0.999135
Forward fill	0.999126
Interpolation: linear	0.999125
Backward fill	0.999105
Interpolation: Cubic	0.999087
Interpolation: Quadratic	0.999081
Seasonal Decompose: Seasonal	0.998753

3.2.2. Categorization

The categorical columns in the dataset were derived from the news data and can be observed in Table 3.10, which provides the distinct count for each category. In order to preprocess these categorical columns, we utilized the Pyspark machine learning function called StringIndexer. This function was selected for its efficiency in handling large datasets, as memory management is crucial for successful transformation. The outcome of this categorization process was a more compact and manageable dataset (Apache Spark, Accessed on July 19, 2023).

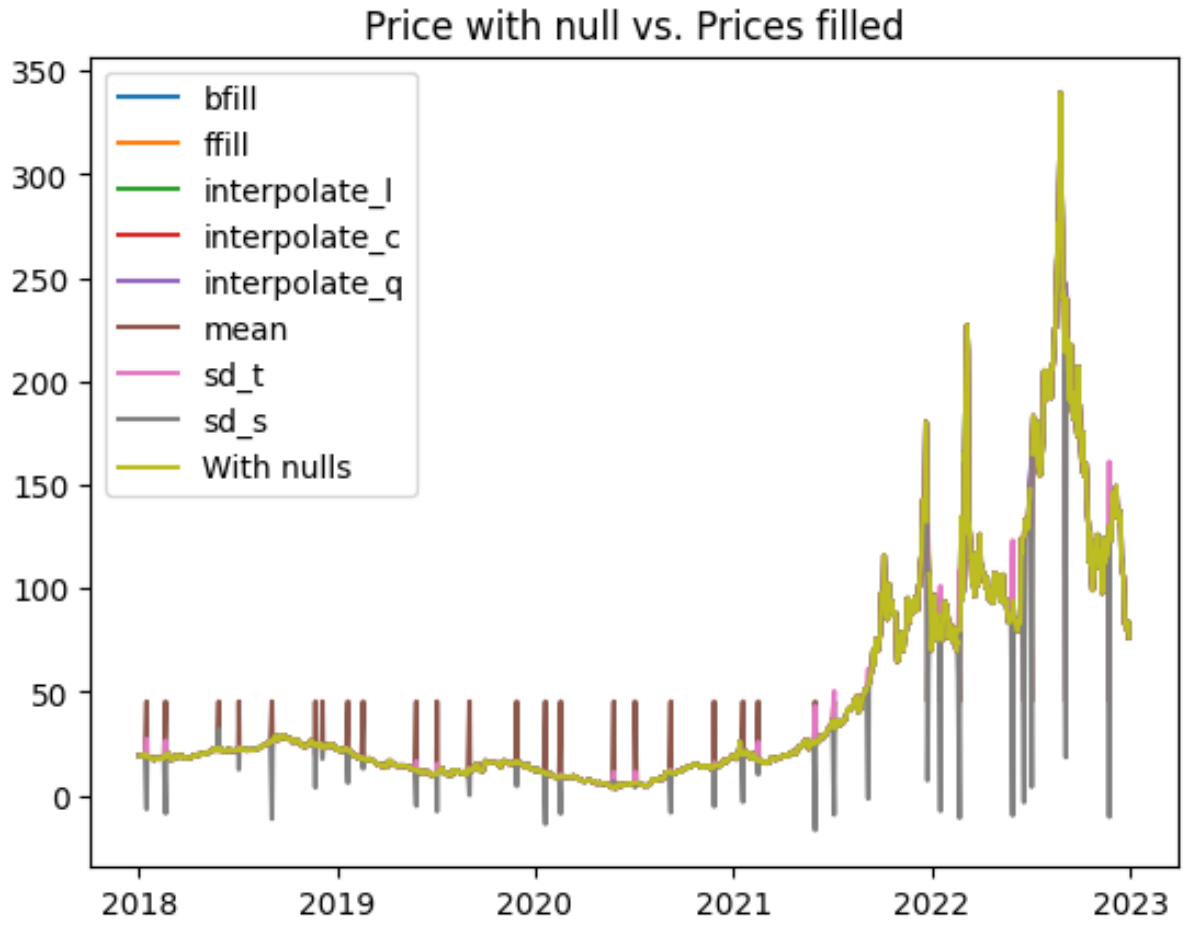

FIGURE 3.6. Compilation of methods to fill null natural gas prices for all time series.

3.2.3. Correlations

The initial step involved generating a Pearson Correlation Matrix to examine the linear relationships between all variables in the dataset. This matrix can be visualized in Figure 3.9, and the corresponding correlation values were presented in Table 5.7 for negative correlations and Table 5.8 for positive correlations. Upon analysis, it was observed that the natural gas price exhibited a positive correlation of 0.722026 with the crude oil price. To further investigate this finding, a zoomed-in plot (Figure 3.10) focusing on these two columns was generated (Downey, Accessed on July 19, 2023).
In order to filter the columns based on their correlation values, a criterium was set to include only those with correlations greater than 0.7 and less than -0.7 (see Table 5.6) (Downey, Accessed on July 19, 2023).

A series of scatter plots was generated to explore potential non-linear correlations. The columns "Actor1Name_Idx" and "Actor1Code_Idx" did not exhibit any significant correlation with the natural gas price, as depicted in Figures 5.2 and 5.1. Similarly, the "Avg_temp" feature displayed a discernible pattern, but it was not conclusive in determining its correlation with the natural gas price, as shown in Plot 5.3.

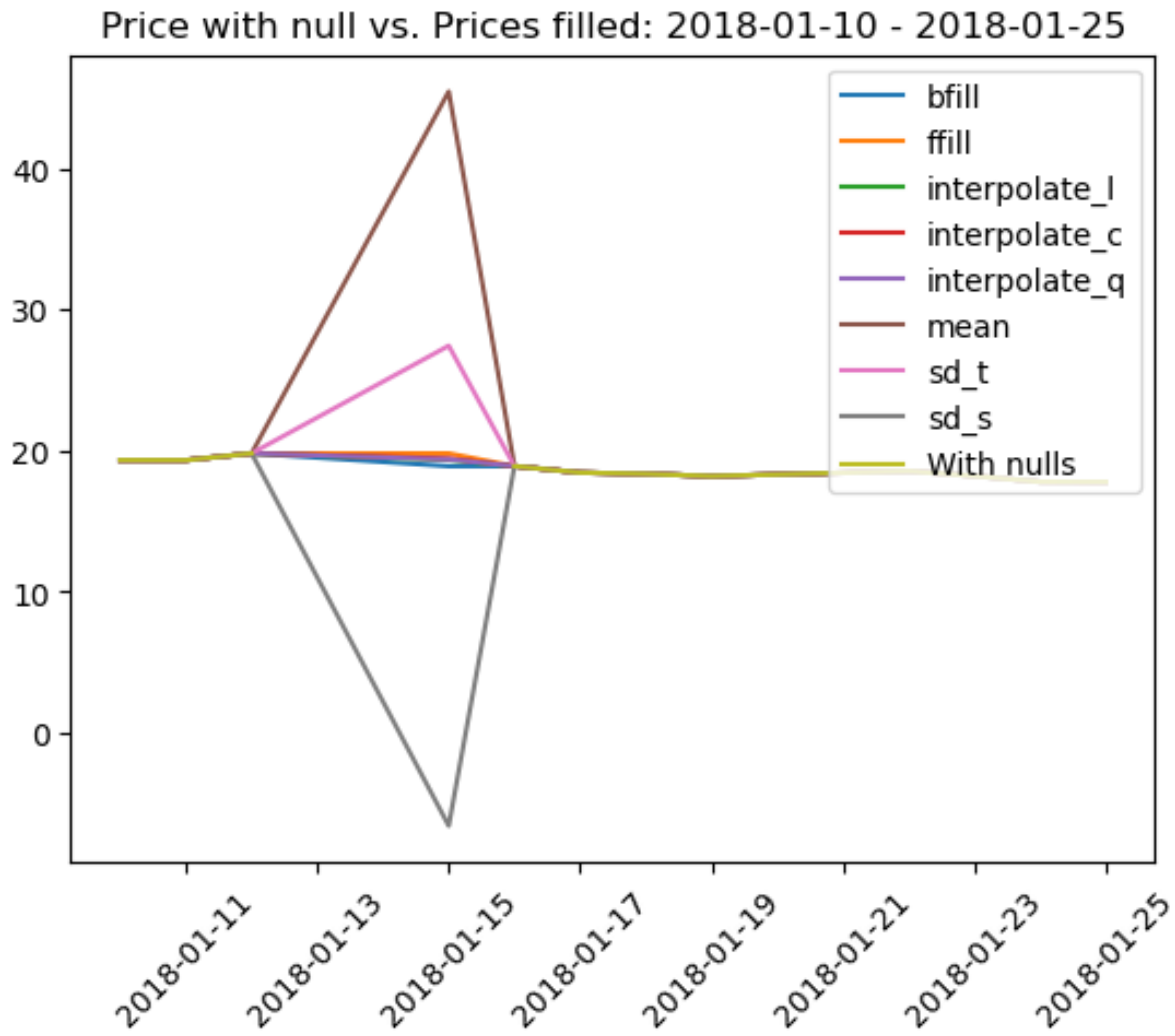

FIGURE 3.7. Compilation of methods to fill null natural gas prices for 2018-01-15.

On the other hand, the "Avg_Tone" feature revealed a wider range of tone values when the price of natural gas was below 50, as depicted in Figure 5.4. This trend was also observed for the "NumArticles," "NumMentions," and "NumSources" columns, indicating a higher volume of articles when the natural gas price was low (respectively in Figures 5.5, 5.6, and 5.7).

Regarding the weather variables, "Prcp_temp" and "Wind_temp" did not display any clear patterns or correlations with the natural gas price, as seen in Figures 5.8 and 5.9.

Overall, the most visually correlated feature with the natural gas price was the crude oil price, as demonstrated in Figure 5.10. This observation is consistent with the results obtained from the Pearson Correlation analysis, indicating a strong correlation between the prices of natural gas and crude oil.

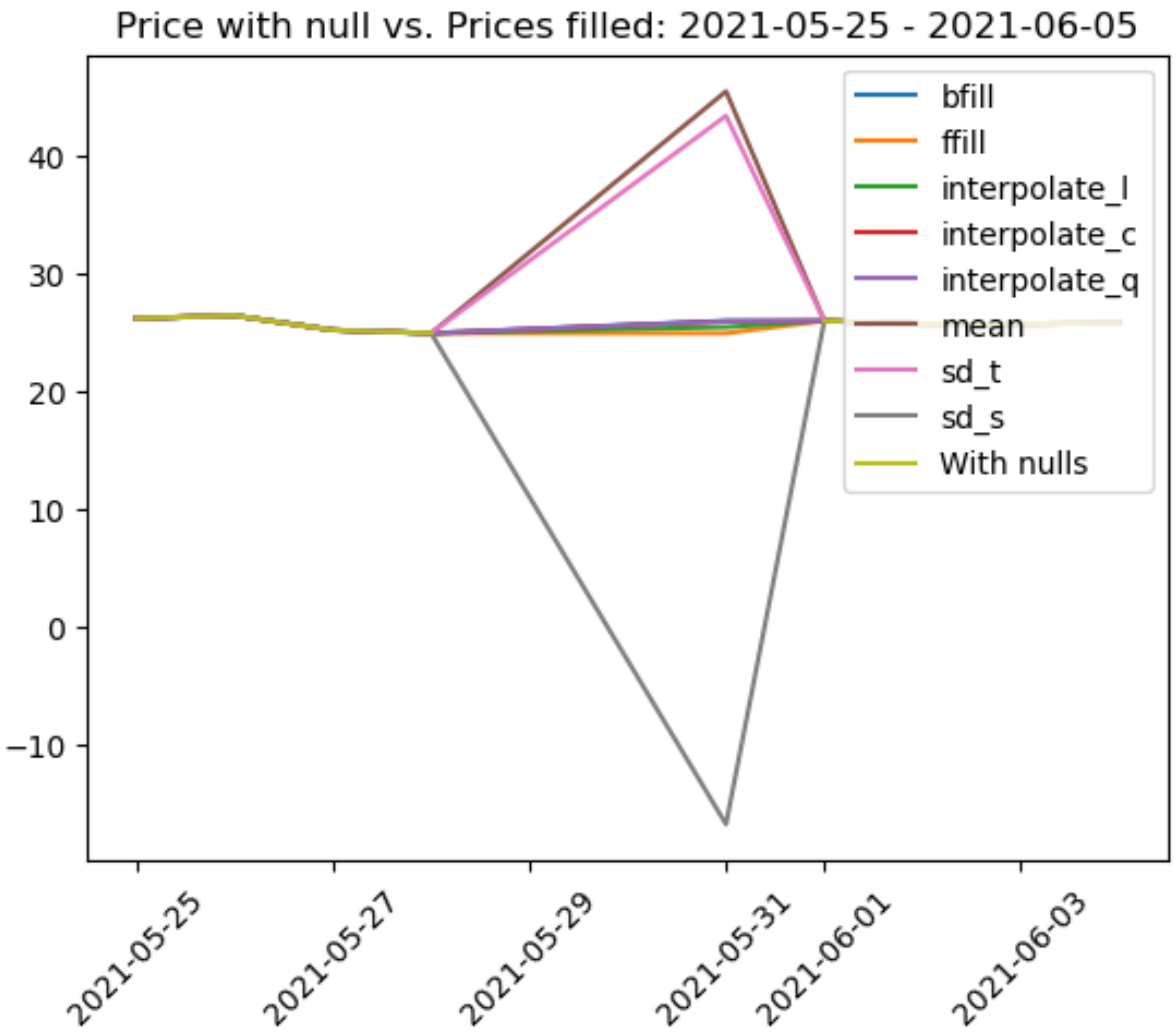

FIGURE 3.8. Compilation of methods to fill null natural gas prices for 2021-05-31.

3.2.4. Granger Causality

Granger causality is a statistical technique employed to evaluate whether a given time series can anticipate or project the behavior of another time series. This entails examining whether past data points from one-time series offer valuable insights into anticipating future data points from another, suggesting a plausible causal connection within time-dependent datasets. It's crucial to emphasize that Granger causality doesn't necessarily establish a direct cause-and-effect relationship; instead, it identifies predictive correlations rooted in statistical trends.

To conduct the Granger Causality analysis, we set the significance level (alpha) to 0.05, and the lag parameter to values of 1, 5, and 10. The analysis was performed for the following features:

When analyzing Granger causality with lag parameters of 1, 5, and 10, consistent findings emerged. Specifically, the variables Wind_temp, Price_oil, and Avg_temp displayed a lack of substantial Granger causality. Tables 3.11, 3.12, 3.13, 3.14, 3.15, 3.16, 3.18, 3.17, and

TABLE 3.10. Sum of distinct values for all category features.

Columns	**Value**
IsRootEvent	1
EventBaseCode	38
EventRootCode	6
Actor1Code_Idx	10499
Actor1Name_Idx	8697
Actor1CountryCode_Idx	220
Actor1KnownGroupCode_Idx	56
Actor1EthnicCode_Idx	357
Actor1Religion1Code_Idx	16
Actor1Religion2Code_Idx	20
Actor1Type1Code_Idx	34
Actor1Type2Code_Idx	27
Actor1Type3Code_Idx	24
Actor2Code_Idx	9754
Actor2Name_Idx	8159
Actor2CountryCode_Idx	221
Actor2KnownGroupCode_Idx	55
Actor2EthnicCode_Idx	344
Actor2Religion1Code_Idx	16
Actor2Religion2Code_Idx	20
Actor2Type1Code_Idx	33
Actor2Type2Code_Idx	28
Actor2Type3Code_Idx	23
QuadClass_Idx	2
Actor1Geo_Type_Idx	6
Actor1Geo_FullName_Idx	189346
Actor1Geo_CountryCode_Idx	255
Actor1Geo_ADM1Code_Idx	4136
Actor2Geo_Type_Idx	6
Actor2Geo_FullName_Idx	163323
Actor2Geo_CountryCode_Idx	253
Actor2Geo_ADM1Code_Idx	4088
ActionGeo_Type_Idx	6
ActionGeo_FullName_Idx	200253
ActionGeo_CountryCode_Idx	255
ActionGeo_ADM1Code_Idx	4160

3.19 presents details of the Granger causality result.

Features GoldsteinScale, NumMentions, NumSources, NumArticles, QuadClass, AvgTone, Threaten, Protest, Exhibit_force_posture, Reduce_relations, Fight, and Use_unconvetional_mass_vi show a strong causality relationship with the natural gas price.

TABLE 3.11. Granger causality results of feature Wind_temp with one lag, without a strong Granger causality.

Wind_temp - Lag = 1				
ssr based F test	F=2.8113	p=0.0938	df_denom=1288	df_num=1
ssr based chi2 test	chi2=2.8179	p=0.0932	df=1	
likelihood ratio test	chi2=2.8148	p=0.0934	df=1	
parameter F test	F=2.8113	p=0.0938	df_denom=1288	df_num=1

TABLE 3.12. Granger causality results of feature Price_oil with one lag, without a strong Granger causality.

Price_oil - Lag = 1				
ssr based F test	F=0.2684	p=0.6045	df_denom=1288	df_num=1
ssr based chi2 test	chi2=0.2690	p=0.6040	df=1	
likelihood ratio test	chi2=0.2690	p=0.6040	df=1	
parameter F test	F=0.2684	p=0.6045	df_denom=1288	df_num=1

TABLE 3.13. Granger causality results of feature Avg_temp with one lag, without a strong Granger causality.

Avg_temp - Lag = 1				
ssr based F test	F=0.4324	p=0.5110	df_denom=1288	df_num=1
ssr based chi2 test	chi2=0.4334	p=0.5103	df=1	
likelihood ratio test	chi2=0.4333	p=0.5104	df=1	
parameter F test	F=0.4324	p=0.5110	df_denom=1288	df_num=1

TABLE 3.14. Granger causality results of feature Wind_temp with five lags, without a strong Granger causality.

Wind_temp - Lag = 5				
ssr based F test	F=0.5456	p=0.7418	df_denom=1288	df_num=5
ssr based chi2 test	hi2=2.7518	p=0.7382	df=5	
likelihood ratio test	hi2=2.7518	p=0.7382	df=5	
parameter F test	F=0.5456	p=0.7418	df_denom=1288	df_num=5

TABLE 3.15. Granger causality results of feature Price_oil with five lags, without a strong Granger causality.

Price_oil - Lag = 5				
ssr based F test	F=1.8175	p=0.1065	df_denom=1288	df_num=5
ssr based chi2 test	chi2=9.1659	p=0.1026	df=5	
likelihood ratio test	chi2=9.1334	p=0.1039	df=5	
parameter F test	F=1.8175	p=0.1065	df_denom=1288	df_num=5

TABLE 3.16. Granger causality results of feature Avg_temp with five lags, without a strong Granger causality.

Avg_temp - Lag = 5				
ssr based F test	F=1.4011	p=0.2210	df_denom=1288	df_num=5
ssr based chi2 test	chi2=7.0660	p=0.2158	df=5	
likelihood ratio test	chi2=7.0467	p=0.2172	df=5	
parameter F test	F=1.4011	p=0.2210	df_denom=1288	df_num=5

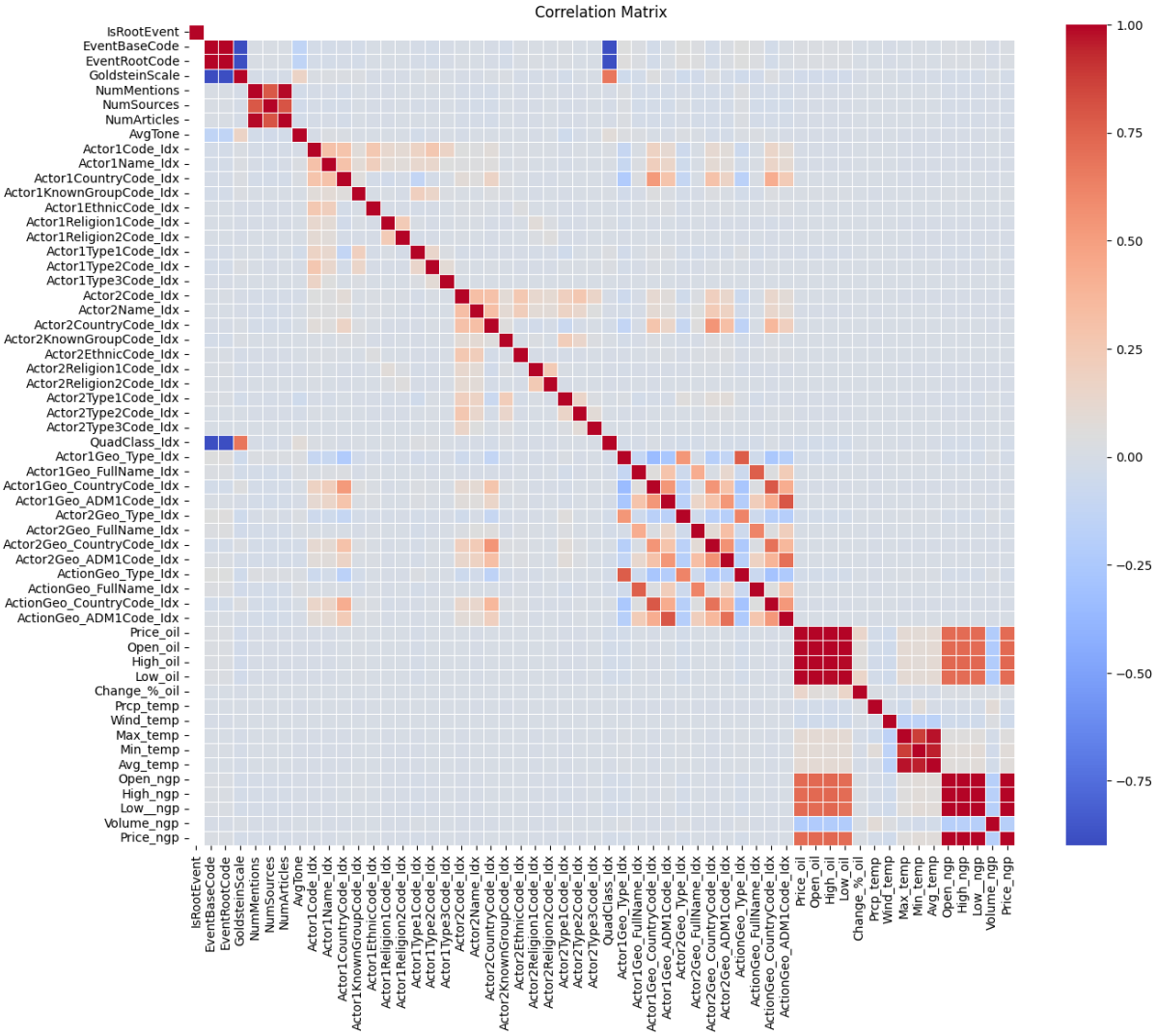

FIGURE 3.9. Pearson correlation heatmap of all features.

TABLE 3.17. Granger causality results of feature Price_oil with ten lags, without a strong Granger causality.

Price_oil - Lag = 10				
ssr based F test	F=1.8175	p=0.1065	df_denom=1288	df_num=10
ssr based chi2 test	chi2=9.1659	p=0.1026	df=10	
likelihood ratio test	chi2=9.1334	p=0.1039	df=10	
parameter F test	F=1.8175	p=0.1065	df_denom=1288	df_num=10

3.2.5. Outliers

The columns Avg_temp, Avg_tone, and Price_oil exhibited outliers, as observed in Figure 5.23, Figure 5.24, and Figure 5.25, respectively. The total count of outliers can be found in Table 3.20. To address these outliers, a mean value was computed using the values before and after the outlier, and the outlier value was replaced with the computed mean. It is important to note that not all outliers were removed, as they provide valuable insights into real-world occurrences.

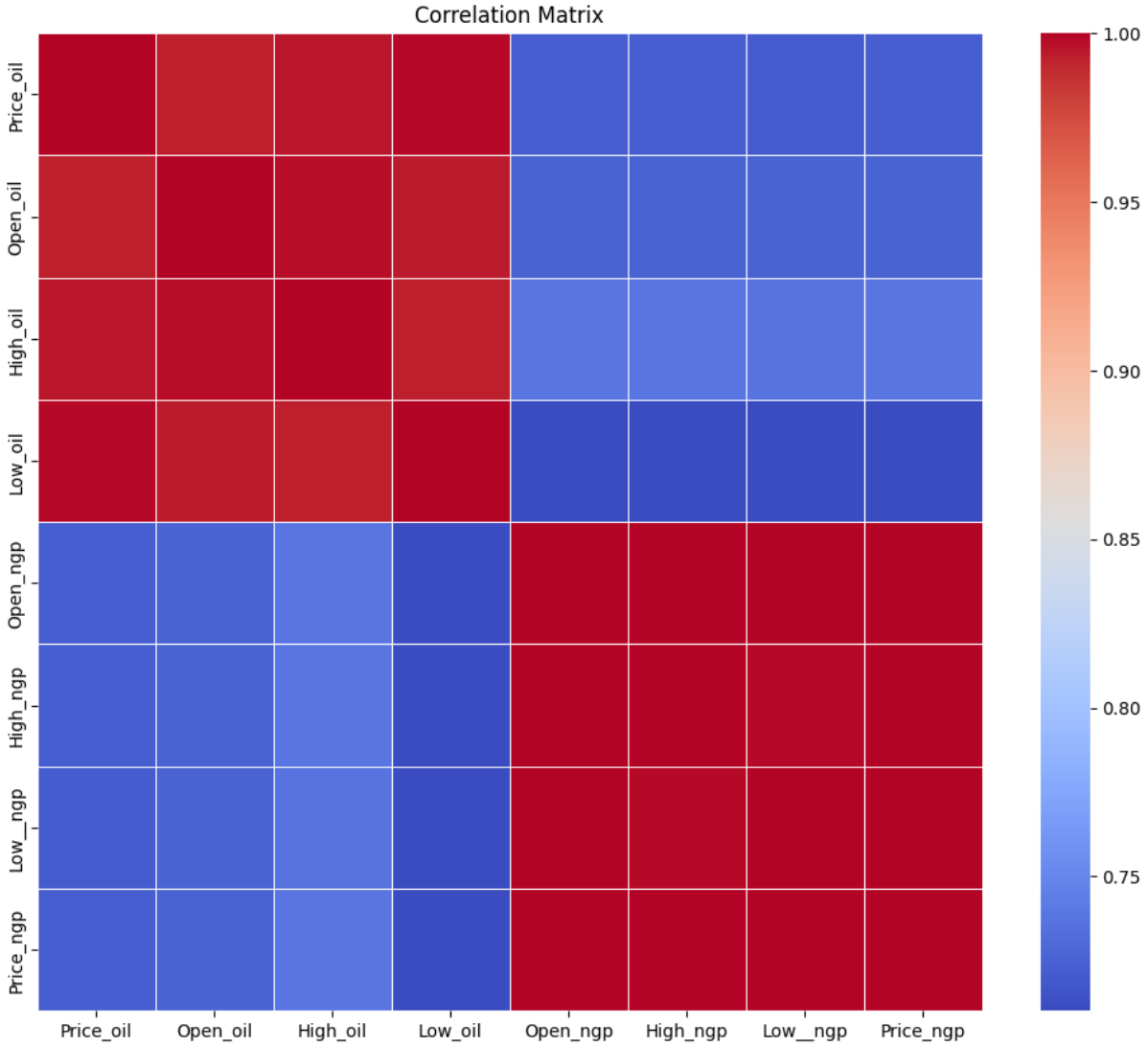

FIGURE 3.10. Pearson correlation heatmap between natural gas features and crude oil features.

TABLE 3.18. Granger causality results of feature Wind_temp with ten lags, without a strong Granger causality.

Wind_temp - Lag = 10				
ssr based F test	F=0.5456	p=0.7418	df_denom=1288	df_num=10
ssr based chi2 test	chi2=2.7518	p=0.7382	df=10	
likelihood ratio test	chi2=2.7488	p=0.7386	df=10	
parameter F test	F=0.5456	p=0.7418	df_denom=1288	df_num=10

3.2.6. Aggregation and Lags

In this preprocessing step, the first action performed was pivoting the variable EventRootCode into separate columns: Threaten, Protest, Exhibit_force_posture, Reduce_relations, Fight, and Use_unconventional_mass_violence. Following this, the news features were aggregated on a daily basis. The columns subjected to summation during the aggregation process were NumMentions, NumSources, NumArticles, QuadClass_Idx, and the columns

TABLE 3.19. Granger causality results of feature Avg_temp with ten lags, without a strong Granger causality.

Avg_temp - Lag = 10				
ssr based F test	F=1.4011	p=0.2210	df_denom=1288	df_num=10
ssr based chi2 test	chi2=7.0660	p=0.2158	df=10	
likelihood ratio test	chi2=7.0467	p=0.2172	df=10	
parameter F test	F=1.4011	p=0.2210	df_denom=1288	df_num=10

TABLE 3.20. Summary of outliers

Column	**q1**	**q3**	**iqr**	**lower_bound**	**upper_bound**	**count_outliers**	**perc_outliers**
AvgTone	-6.404	-2.046	4.358	-12.941	4.491	224133	1.210
Price_oil	52.64	70.73	18.09	25.505	97.865	454154	2.453
Avg_temp	58.0	178.0	120.0	-122.0	358.0	23309	0.126

derived from the EventRootCode column. Additionally, the columns Wind_temp, AvgTone, Price_oil, and Avg_temp were averaged.
To ensure that no calculations were performed on the label column, the minimum value of the natural gas price was selected.
As part of the preprocessing phase, we introduced a lag of 5 and 10 days to all time series. This lagging process entails shifting the values of each column backward in time by the specified number of days. By incorporating these lagged values as additional features, our objective was to capture the temporal relationships in the data. The inclusion of lagged features enables the models to take into account the historical values of each variable when making predictions.

3.2.7. Scaler

In the last preprocessing step, we used a min-max scaler to transform the data. The min-max scaler rescales the values of each feature to a range between zero and one. This normalization technique, implemented using the Scikit-learn library, helps to ensure that all features are on a comparable scale. By applying the min-max scaler, we aimed to facilitate the training process of the models by reducing the impact of varying feature magnitudes (Pedregosa et al., 2011).

CHAPTER 4

Modeling and Performance Evaluation

The deep learning models chosen for this study were the Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Neural Networks (GRUNN) from Tensorflow library. The selection of these models was complemented with the use of Keras Tuner, a framework for hyperparameter optimization. This allowed us to automatically search for the best hyperparameters for each model.

To assess the performance of the models, several evaluation metrics were employed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE), R-Squared (R2), Explained Variance Score (EVS), and Root Mean Squared Error (RMSE).

The complete modeling workflow is depicted in Figure 4.1, illustrating the sequential steps involved in model development and evaluation.

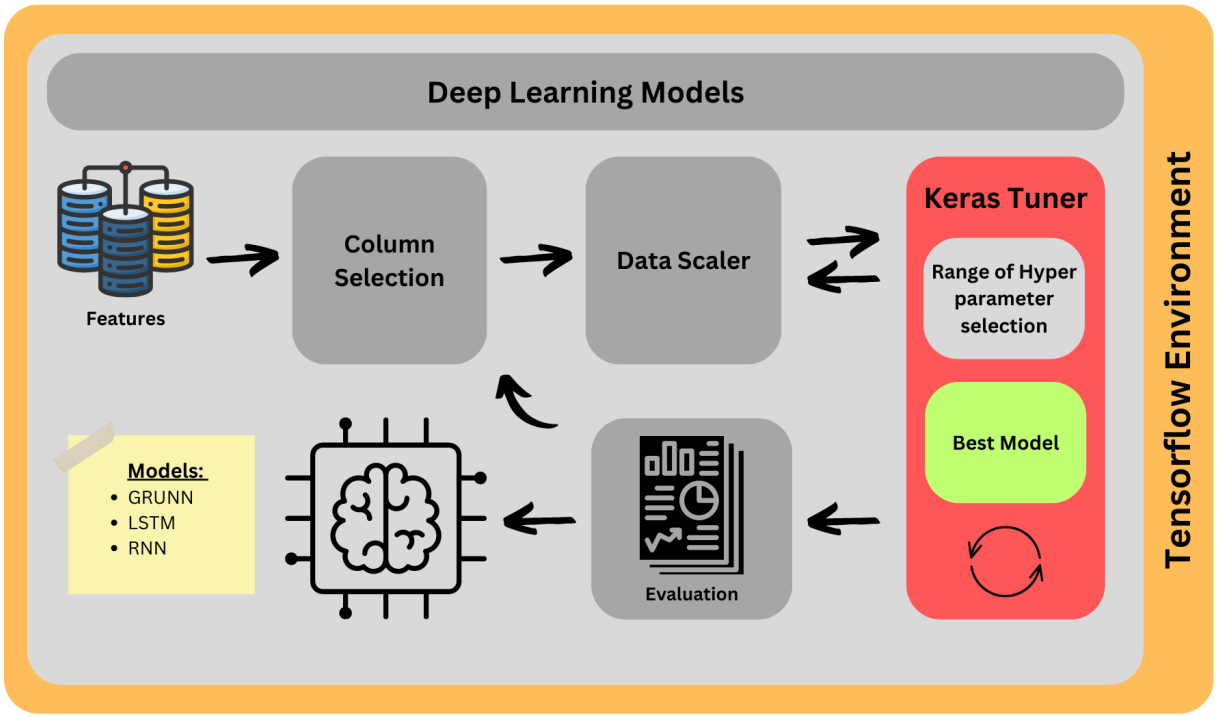

FIGURE 4.1. ETL process for column selection, data scaler, keras tuner, and evaluation of the best model.

4.1. Modeling

4.1.1. Keras Tuner - Hyperparameters

Keras Tuner is a specialized library that optimizes hyperparameters in deep learning models built using Keras. Hyperparameters are external settings that impact the training process, such as the learning rate, number of layers, and units per layer.
Keras Tuner automates the search for the best hyperparameters by utilizing algorithms like random search, grid search, and Bayesian optimization. These strategies efficiently explore the hyperparameter space to identify the combination of settings that yield optimal model performance (O'Malley et al., 2019).
In our study, we specifically used the random search approach to find the best model configuration. We focused on three key hyperparameters: the number of layers, units, and epochs. The epoch value was consistently set to 20, with early stopping applied after 5 epochs without improvement. For the number of layers, we tested a range from one to five, incrementing by one layer per trial. The units were varied between 32 and 512, incrementing by 32 units per trial. Commencing with a modest value like 32 and gradually increasing it by steps of 32 covers a diverse spectrum of options without needing to meticulously test each and every value. This method achieves a harmonious equilibrium between exploring a sufficiently extensive hyperparameter space and preventing excessive consumption of computational resources and time.
In the second Keras Tuner trial, we refined the search based on the best models from the previous attempt. We limited the number of layers to one or two, and the units were constrained to the range of 32 to 512, maintaining the same increment value.

4.1.2. Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are specialized neural networks designed to process sequential data, such as time series or natural language sequences. Unlike traditional neural networks, RNNs have connections that allow them to retain information from previous time steps, enabling them to capture the temporal relationships in sequential data. At the core of an RNN is the hidden state, which acts as a memory of past inputs and is updated at each time step. This hidden state the current input and past information, allowing RNNs to learn and model the patterns and dynamics of sequential data. However, a limitation of RNNs is the vanishing gradient problem, which hampers their ability to capture long-term dependencies. To address this issue, advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) has been developed. These variants incorporate gating mechanisms to alleviate the vanishing gradient problem and improve the RNNs' capacity to capture and remember long-term dependencies (Goodfellow et al., 2016).
Our investigation employed the random search strategy to find the optimal model configuration. We focused on three crucial hyperparameters: the number of layers, units, and

epochs.

4.1.3. Long Short Term Memory (LSTM)

LSTM, a variation of Recurrent Neural Networks (RNNs), addresses the challenge of capturing long-term dependencies in sequential data. By incorporating specialized memory cells, LSTM models can retain information for extended periods, enabling them to effectively learn and represent temporal dependencies. Unlike traditional RNNs, LSTM networks employ a gating mechanism that allows for selective retention and forgetting of information at each time step. This mechanism comprises three gates: the input gate, forget gate, and output gate. The input gate controls the flow of new information, while the forget gate determines which information to discard. The output gate regulates the output of the memory cell. Through dynamic memory updates and control, LSTM networks excel in capturing and preserving long-term dependencies, making them highly suitable for tasks such as speech recognition, language modeling, and machine translation (Goodfellow et al., 2016).

4.1.4. Gated Recurrent Unit Neural Networks (GRUNN)

Gated Recurrent Unit for Neural Networks (GRUNN) is an advanced variation of recurrent neural networks (RNNs) that overcomes limitations found in traditional RNN architectures. GRUNN incorporates gating mechanisms to control the flow of information within the network, allowing it to selectively retain or update information at each time step. This addresses the issue of the vanishing gradient problem, which can hinder training in deep neural networks. By selectively preserving relevant information and discarding irrelevant information, GRUNN models can effectively capture long-term dependencies in sequential data. The gated recurrent units in GRUNN consist of a reset gate and an update gate, which govern the information flow through the network. The reset gate determines what information from previous time steps should be forgotten, while the update gate controls the blending of new input with the existing hidden state. This adaptive gating mechanism empowers GRUNN models to capture intricate temporal patterns and dependencies, making them highly suitable for tasks involving time series forecasting, natural language processing, and speech recognition (Goodfellow et al., 2016).
The GRUNN has its function, which dynamically uses the Keras tuner to find the best layer and unit number. The compile is set with Adam optimizer, loss is MSE, and the metric is RMSE.

4.2. Performance Evaluation

In each model, we developed a baseline with either 5 or 10 lagged values, consisting solely of the natural gas price feature. Subsequently, we conducted model training using all available features listed in Table 4.1.

With a test dataset of 258 data points, we predicted the same size time daily.
Abbreviations for the different feature combinations can be found in Table 4.3. The optimal hyperparameters for each trained model are presented in Table 4.2, while the performance of each model is ranked in Table 4.4.

TABLE 4.1. All features applied to the model.

Column Name	**Descritption**
GoldsteinScale	Numeric measure indicating the level of conflict or cooperation in political events.
NumMentions	Represents the number of mentions of an event in various sources
NumMentions	Number of mentions of the event
NumArticles	Number of articles related to the event
QuadClass_Idx	Index for the QuadClass category
AvgTone	Average tone of the event
Price_oil	Crude oil price
Avg_temp	Average temperature
Threaten	Cameo code to threaten
Protest	Cameo code to protest
Exhibit_force_posture	Cameo code to exhibit force posture
Reduce_relations	Cameo code to reduce relations
Fight	Cameo code to fight
Use_unconvetional_mass_violence	Cameo code to use unconventional mass violence
Price_ngp	Natural gas price

4.2.1. Best model: Recurrent Neural Networks (RNN)

Among all the models trained, the RNN model outperformed the others. This model incorporated the features of natural gas price, crude oil price, and average tone of the extracted news. With ten lagged values, one layer, and 224 units, the RNN model achieved an RMSE of 11.925 euros. The prediction made by the RNN model is illustrated in Figure 4.2 in the form of a time-series plot.

4.2.2. Best model: Long Short Term Memory (LSTM)

The LSTM model, which utilized the same set of features as the best model (natural gas price, crude oil price, and average tone), ranked fourth among the best models. With ten lagged values, one layer, and 320 units, the LSTM model achieved an RMSE of 11.954. The prediction made by the LSTM model is depicted in Figure 4.5 as a time-series plot.

4.2.3. Best model: Gated Recurrent Unit Neural Networks (GRUNN)

The GRUNN model, utilizing natural gas price and crude oil price as features, secured the second position among the best models. With five lagged values, one layer, and 480 units, the GRUNN model obtained an RMSE of 11.935. The prediction generated by the GRUNN model is illustrated in Figure 4.3 in the form of a time-series plot.

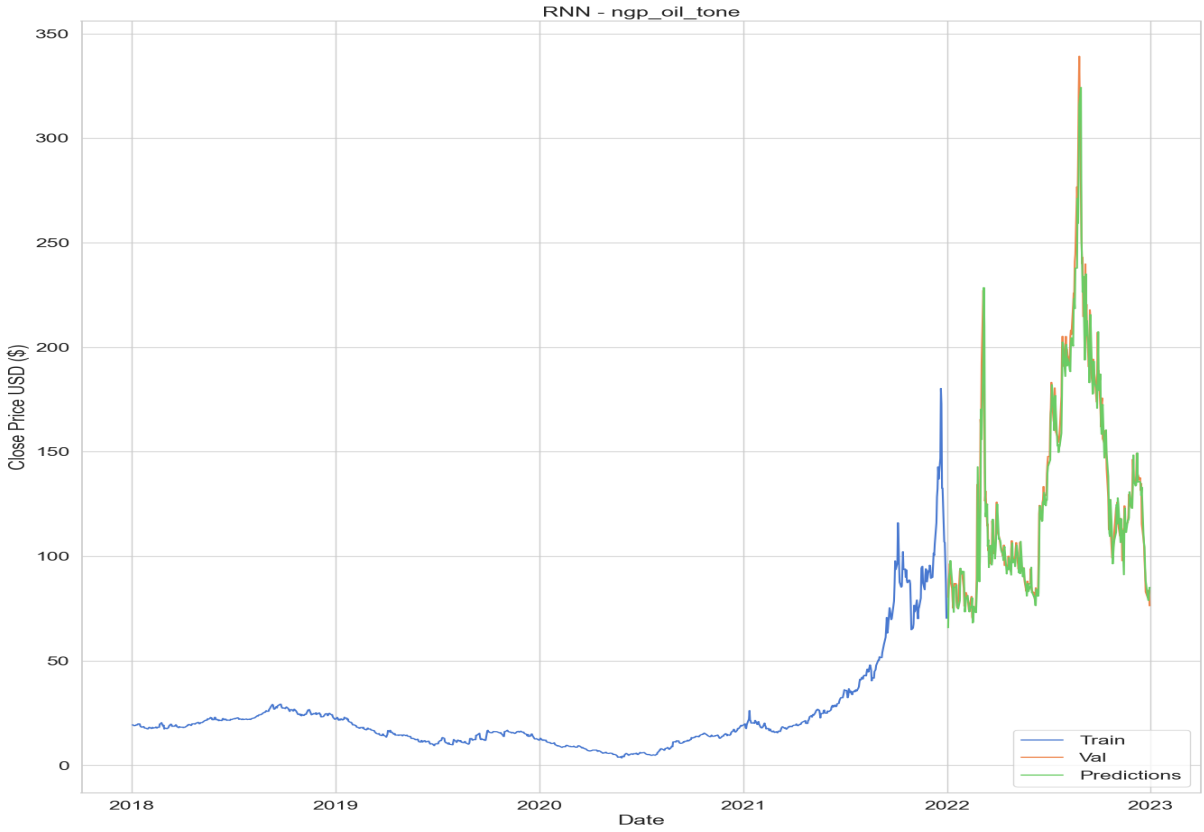

FIGURE 4.2. Prediction plot curve of RNN model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

4.2.4. Comparative Analysis

The literature review covered various prediction horizons, utilizing different models, features, and preprocessing methods. Our comparison focuses on single and hybrid models for predicting daily natural gas prices, similar to other studies.

Among the single models reviewed, the study by Al-Sharoot and Alramadhan (2019) achieved the best performance using Auto-regressive moving average (ARMA) and Group Method of Data Handling (GMDH) models with 527 observations from August 29, 2016, to August 27, 2018, without exogenous variables. Their mean squared error (MSE) was 0.0214.

Another notable single model from Qin et al. (2019) employed Ensemble Empirical Mode Decomposition (EEMD) and Local Linear Prediction (LLP) with 1678 observations from January 4, 2010, to August 15, 2016, also without exogenous variables. Their root mean squared error (RMSE) was 0.035.

Our study's best-performing model was a Recurrent Neural Network (RNN) with 10 lags, incorporating natural gas price, crude oil price, and average tone as exogenous variables. We used 1292 observations from January 2, 2018, to December 30, 2022, and achieved an RMSE of 11.925.

However, it is important to note that our results were not as favorable as the best results in the literature. This disparity is primarily attributed to the complexity of the prediction

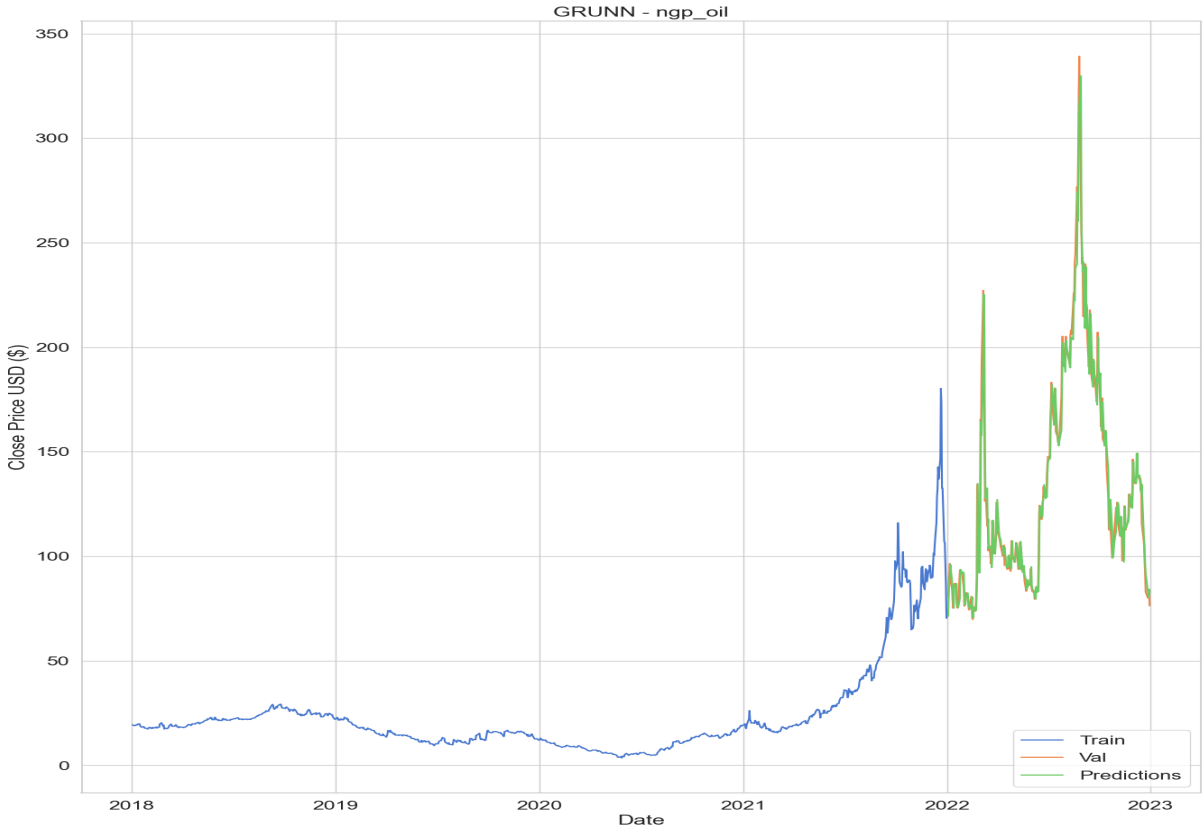

FIGURE 4.3. Prediction plot curve of GRUNN model with natural gas price and crude oil price as features, and lag equal to 5.

period chosen, which coincided with the Russo-Ukrainian War. The geopolitical situation during this period likely contributed to the model's reduced performance.

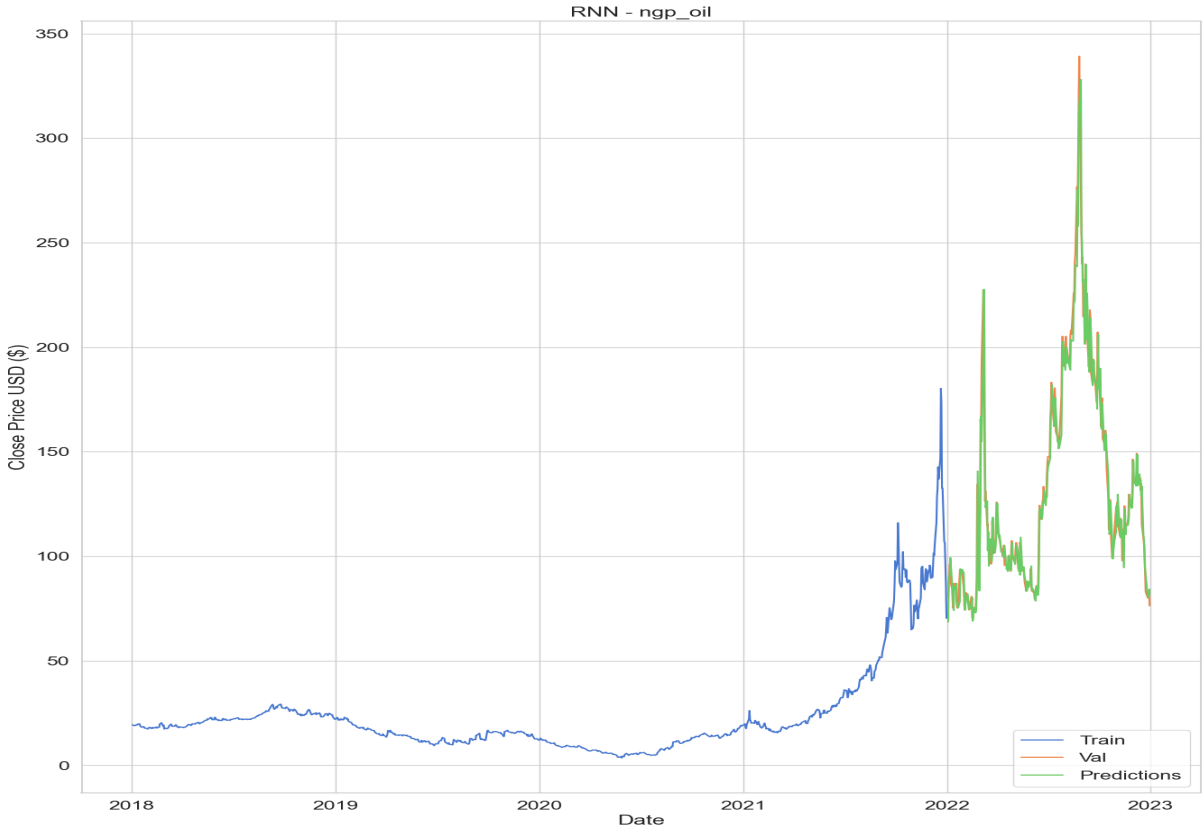

FIGURE 4.4. Validation curve of RNN model with natural gas price and crude oil price as features, and lag equal to 10.

TABLE 4.2. All models with the best hyperparameter selected by Keras Tuner.

Model Name	Features Code	RMSE	Best Num Layers	Best Num Units
RNN	ngp_oil_tone	11,925	1	224
GRUNN	ngp_oil	11,935	1	480
RNN	ngp_oil	11,942	1	96
LSTM	ngp_oil_tone	11,954	1	320
LSTM	ngp_oil	11,962	1	96
RNN	base_line	11,964	1	192
GRUNN	base_line	11,979	1	224
GRUNN	ngp_oil	11,996	1	32
GRUNN	all_features	12,001	1	32
RNN	base_line	12,010	1	32
GRUNN	base_line	12,023	1	32
GRUNN	ngp_oil	12,043	1	32
LSTM	all_features	12,045	1	384
RNN	ngp_oil	12,060	1	160
GRUNN	all_features	12,066	1	128
RNN	ngp_tone	12,077	1	32
GRUNN	ngp_oil	12,087	2	32
LSTM	ngp_oil_tone	12,090	1	128
GRUNN	all_features	12,098	1	160
GRUNN	ngp_oil_tone	12,116	2	32
RNN	all_features	12,124	1	320
GRUNN	ngp_oil_tone	12,125	1	96
LSTM	ngp_oil_tone	12,129	1	192
RNN	ngp_oil_tone	12,141	1	96
LSTM	ngp_oil_tone	12,142	1	256
RNN	ngp_oil	12,145	1	96
GRUNN	ngp_tone	12,161	1	128
GRUNN	ngp_tone	12,165	1	160
RNN	ngp_tone	12,178	1	96
RNN	all_features	12,206	1	128
LSTM	all_features	12,244	1	384
GRUNN	ngp_oil_tone	12,250	2	64
LSTM	base_line	12,258	1	32
GRUNN	all_features	12,268	1	192
LSTM	ngp_tone	12,279	1	64
GRUNN	ngp_oil_tone	12,282	2	64
RNN	ngp_tone	12,293	1	480
RNN	ngp_oil_tone	12,302	1	384
LSTM	ngp_oil	12,323	1	96
LSTM	ngp_tone	12,338	1	64
LSTM	base_line	12,339	1	224
RNN	ngp_tone	12,340	2	64
LSTM	ngp_oil	12,341	1	160
RNN	ngp_oil	12,363	1	32
RNN	ngp_oil_tone	12,369	1	32
LSTM	ngp_oil	12,384	1	320
LSTM	ngp_tone	12,483	1	160
RNN	all_features	12,520	1	64
GRUNN	ngp_tone	12,555	1	480
LSTM	all_features	12,606	1	416
RNN	all_features	12,617	1	416
LSTM	ngp_tone	12,717	1	32
GRUNN	ngp_tone	13,286	3	512

TABLE 4.3. Features code description.

Code	Description
ngp_oil_tone	Natural gas price, crude oil price, and average tone
ngp_oil	Natural gas price and crude oil price
base_line	Natural gas price
all_features	All features
ngp_tone	Natural gas price and average tone

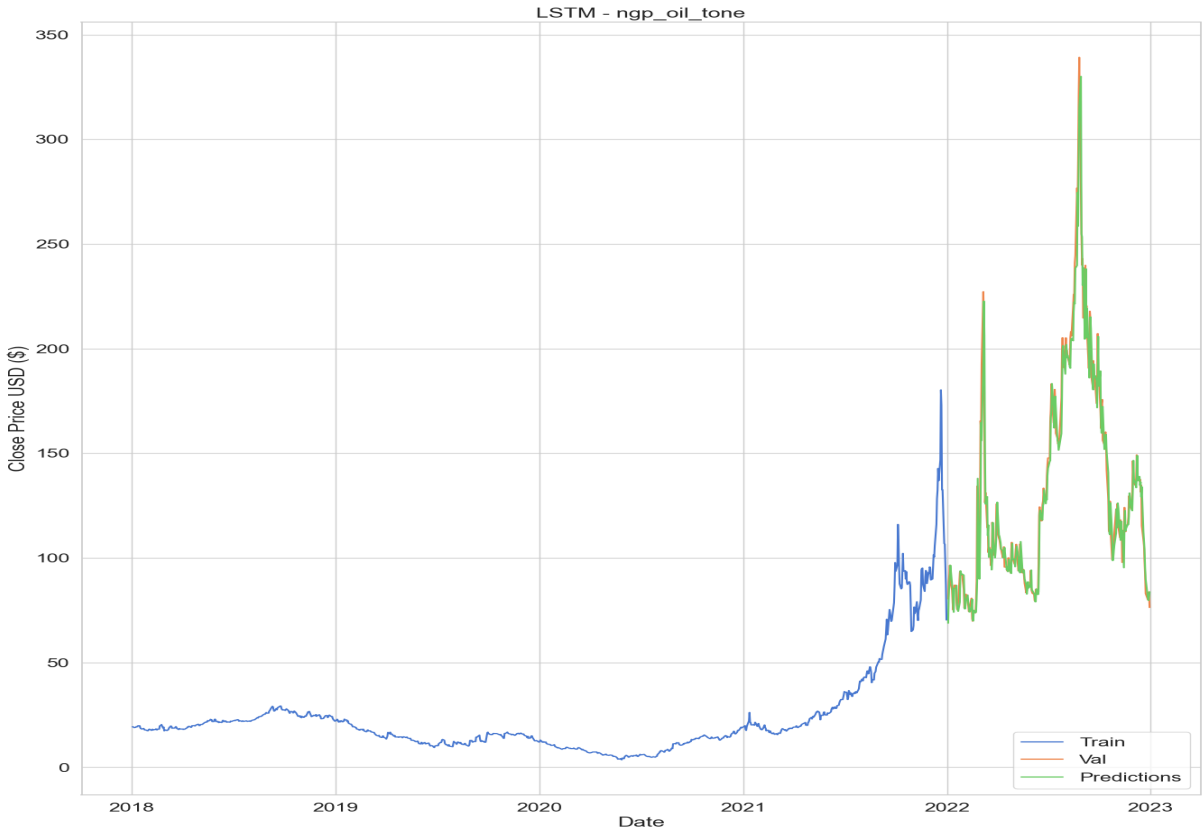

FIGURE 4.5. Prediction plot curve of LSTM model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

TABLE 4.4. Result for each model executed with features applied, lags used, MAE, MSE, MAPE, MedAE, R2, EVS, and RMSE performance metrics.

Model Name	Features Code	Lags	MAE	MSE	MAPE	MedAE	R2	EVS	RMSE
RNN	ngp_oil_tone	10	7,846	142,215	5,759	4,759	0,946	0,946	11,925
GRUNN	ngp_oil	5	7,811	142,438	5,712	4,470	0,946	0,946	11,935
RNN	ngp_oil	10	7,955	142,604	5,795	4,925	0,946	0,947	11,942
LSTM	ngp_oil_tone	10	7,902	142,907	5,777	4,609	0,946	0,946	11,954
LSTM	ngp_oil	10	7,956	143,096	5,836	4,791	0,946	0,946	11,962
RNN	base_line	5	7,881	143,142	5,749	4,989	0,946	0,946	11,964
GRUNN	base_line	5	7,903	143,498	5,732	4,771	0,946	0,946	11,979
GRUNN	ngp_oil	5	7,838	143,903	5,705	4,780	0,946	0,946	11,996
GRUNN	all_features	10	7,941	144,028	5,760	4,957	0,946	0,946	12,001
RNN	base_line	10	8,028	144,248	5,860	4,859	0,946	0,946	12,010
GRUNN	base_line	10	7,987	144,543	5,770	5,030	0,946	0,947	12,023
GRUNN	ngp_oil	10	7,890	145,023	5,763	4,802	0,945	0,946	12,043
LSTM	all_features	10	8,005	145,089	5,896	4,964	0,945	0,945	12,045
RNN	ngp_oil	5	7,940	145,433	5,744	4,961	0,945	0,946	12,060
GRUNN	all_features	5	7,851	145,579	5,814	5,044	0,945	0,945	12,066
RNN	ngp_tone	10	8,139	145,842	5,918	5,222	0,945	0,947	12,077
GRUNN	ngp_oil	10	8,071	146,096	5,795	5,006	0,945	0,946	12,087
LSTM	ngp_oil_tone	5	7,968	146,174	5,743	4,711	0,945	0,946	12,090
GRUNN	all_features	10	7,859	146,367	5,814	4,869	0,945	0,945	12,098
GRUNN	ngp_oil_tone	10	8,053	146,788	5,834	4,928	0,945	0,946	12,116
RNN	all_features	5	8,046	146,984	5,857	4,786	0,945	0,945	12,124
GRUNN	ngp_oil_tone	5	7,901	147,023	5,722	4,785	0,945	0,945	12,125
LSTM	ngp_oil_tone	10	8,027	147,106	5,786	5,038	0,945	0,945	12,129
RNN	ngp_oil_tone	5	7,965	147,411	5,826	5,232	0,944	0,945	12,141
LSTM	ngp_oil_tone	5	7,960	147,423	5,756	4,870	0,944	0,945	12,142
RNN	ngp_oil	10	8,019	147,511	5,852	4,585	0,944	0,945	12,145
GRUNN	ngp_tone	10	8,132	147,880	5,857	5,268	0,944	0,946	12,161
GRUNN	ngp_tone	5	8,061	147,980	5,830	5,080	0,944	0,946	12,165
RNN	ngp_tone	5	8,111	148,306	5,917	5,085	0,944	0,945	12,178
RNN	all_features	10	8,196	148,987	5,925	5,143	0,944	0,945	12,206
LSTM	all_features	10	8,143	149,917	6,098	5,279	0,943	0,944	12,244
GRUNN	ngp_oil_tone	5	8,112	150,070	5,782	4,965	0,943	0,945	12,250
LSTM	base_line	5	8,231	150,256	5,883	5,011	0,943	0,945	12,258
GRUNN	all_features	5	8,106	150,505	6,006	5,217	0,943	0,944	12,268
LSTM	ngp_tone	5	8,259	150,777	5,892	5,405	0,943	0,945	12,279
GRUNN	ngp_oil_tone	10	8,319	150,853	5,973	5,515	0,943	0,946	12,282
RNN	ngp_tone	5	8,061	151,108	5,928	5,040	0,943	0,943	12,293
RNN	ngp_oil_tone	5	8,125	151,348	5,947	5,258	0,943	0,944	12,302
LSTM	ngp_oil	5	8,182	151,856	5,836	4,919	0,943	0,945	12,323
LSTM	ngp_tone	10	8,293	152,219	5,897	5,398	0,943	0,945	12,338
LSTM	base_line	10	8,271	152,239	5,943	5,230	0,943	0,945	12,339
RNN	ngp_tone	10	8,343	152,281	6,017	5,541	0,943	0,945	12,340
LSTM	ngp_oil	5	8,189	152,294	5,848	4,932	0,943	0,944	12,341
RNN	ngp_oil	5	8,226	152,839	5,886	5,420	0,942	0,945	12,363
RNN	ngp_oil_tone	10	8,262	152,999	6,021	5,089	0,942	0,945	12,369
LSTM	ngp_oil	10	8,244	153,367	5,887	4,949	0,942	0,944	12,384
LSTM	ngp_tone	10	8,388	155,832	5,953	5,604	0,941	0,945	12,483
RNN	all_features	5	8,414	156,747	6,014	5,639	0,941	0,944	12,520
GRUNN	ngp_tone	5	8,443	157,629	5,970	5,639	0,941	0,945	12,555
LSTM	all_features	5	8,229	158,906	6,101	5,023	0,940	0,940	12,606
RNN	all_features	10	8,557	159,199	6,246	5,302	0,940	0,941	12,617
LSTM	ngp_tone	5	8,573	161,718	6,010	5,560	0,939	0,944	12,717
LSTM	all_features	5	9,072	171,336	6,681	6,078	0,935	0,936	13,090
GRUNN	ngp_tone	10	9,162	176,510	6,437	6,370	0,933	0,942	13,286

CHAPTER 5

Conclusions

Throughout our study, we explored various features for predicting natural gas prices, considering different scenarios such as using only natural gas prices or combinations with crude oil prices and average tone from news sources. Among the top five performing models, incorporating crude oil price as an exogenous variable significantly enhanced the predictive accuracy, consistent with previous findings (Li et al., 2021). Surprisingly, the best model emerged when we included the feature of average tone in our input data, resulting in a 7.82% improvement compared to the same model without it.

By filtering the extensive GDELT Big Data using specific Cameo codes (see Table 3.2), we achieved improved performance during conflict times. Contrary to the notion proposed by Čeperić et al. (2017), who suggested "less data is better" for short-term prediction, we found that using the relevant variables is crucial for accurate predictions.

Our optimization process, employing the Random Search optimizer, demonstrated an efficient selection of hyperparameters and facilitated in-depth analysis of each interaction. Despite these efforts, the best model's RMSE of 11.925 fell short when compared to the literature, mainly due to an abrupt change in values caused by the Russo-Ukrainian War and Europe's high dependence on Russian natural gas prices.

Nonetheless, our study's best-performing model remained the Recurrent Neural Networks (RNN) with 10 lags, incorporating natural gas price, crude oil price, and average tone as exogenous variables. The entire observation has 1292 data points from January 2, 2018, to December 30, 2022, we achieved promising results, although the worst model, a GRUNN with natural gas price and average tone, had an RMSE of 13.286 (see Table 4.4).

In conclusion, this study highlights the negative impact on the performance of natural gas price models during war times and emphasizes the positive influence of specific Cameo codes on model results. Moreover, the findings reinforce the strong correlation and causation between crude oil and natural gas prices, contributing to improved model performance.

To enhance the methodology, implementing version management tools like MLFlow could have been beneficial. Additionally, further exploration of different Cameo codes and exhaustive testing of all possible feature combinations extracted from GDELT could have been conducted.

References

Abrishami, H., & Varahrami, V. (2011). Different methods for gas price forecasting. *Cuadernos de Economia*, *34*, 137–144. https://doi.org/10.1016/S0210-0266(11)70013-9

Alamro, R., McCarren, A., & Al-Rasheed, A. (2019). Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling. *Communications in Computer and Information Science*, *1097 CCIS*, 317–328. https://doi.org/10.1007/978-3-030-36365-9_26

Al-Sharoot, M. H., & Alramadhan, O. M. (2019). Forecasting the gas prices in investing.com's weekly economic data table using linear and non-linear arma-garch models for period 2016-2018. *AIP Conference Proceedings*, *2096*. https://doi.org/10.1063/1.5097818

Apache Spark. (Accessed on July 19, 2023). pyspark.ml.feature.StringIndexer.

Azadeh, A., Sheikhalishahi, M., & Shahmiri, S. (2012). A hybrid neuro-fuzzy simulation approach for improvement of natural gas price forecasting in industrial sectors with vague indicators. *International Journal of Advanced Manufacturing Technology*, *62*, 15–33. https://doi.org/10.1007/s00170-011-3804-6

Berrisch, J., & Ziel, F. (2022). Distributional modeling and forecasting of natural gas prices. *Journal of Forecasting*, *41*, 1065–1086. https://doi.org/10.1002/for.2853

Bodas-Sagi, D., & Labeaga, J. (2016). Using gdelt data to evaluate the con dence on the spanish government energy policy. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*, 38. https://doi.org/10.9781/ijimai.2016.366

Bourgeois, D., Rappaz, J., & Aberer, K. (2018). Selection bias in news coverage: Learning it, fighting it. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 535–543. https://doi.org/10.1145/3184558.3188724

Čeperić, E., Žiković, S., & Čeperić, V. (2017). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. *Energy*, *140*, 893–900. https://doi.org/10.1016/j.energy.2017.09.026

Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (gru) neural networks. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. https://doi.org/10.1109/MWSCAS.2017.8053243

Downey, A. B. (Accessed on July 19, 2023). *Think stats: Exploratory data analysis in python*. https://greenteapress.com/wp/think-stats-2e/

Economic bulletin issue 4, (2022, June). European Central Bank. https://www.ecb.europa.eu/pub/pdf/ecbu/eb202204.en.pdf

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, *44*, 408–421. https://doi.org/10.1177/0165551517698298

Event Data Project, Department of Political Science, Pennsylvania State University. (March 2012). *CAMEO Conflict and Mediation Event Observations Event and Actor Codebook*. http://eventdata.psu.edu/

Fabian, J., Wingrove, J., & Krukowska, E. (2022, May). *U.s., eu reach lng supply deal to cut dependence on russia*. https://www.bloomberg.com/news/articles/2022-03-25/u-s-and-eu-reach-energy-supply-deal-to-cut-dependence-on-russia?leadSource=uverify%5C%20wall

Galla, D., & Burke, J. (2018). Predicting social unrest using gdelt. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10935 LNAI*, 103–116. https://doi.org/10.1007/978-3-319-96133-0_8

GDELT Project. (n.d.). *The gdelt project*. https://www.gdeltproject.org/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Google Cloud Platform Blog. (2014). *World's largest event dataset now publicly available in bigquery*. https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html

Halser, C., & Paraschiv, F. (2022). Pathways to overcoming natural gas dependency on russia—the german case. *Energies*, *15*. https://doi.org/10.3390/en15144939

Hamie, H., Auer, H., & Hoayek, A. (2020). Modeling post-liberalized european gas market concentration-a game theory perspective. https://doi.org/10.3390/forecast

Hammond, J., & Weidmann, N. B. (2014). Using machine-coded event data for the micro-level study of political violence. *Research and Politics*, *1*. https://doi.org/10.1177/2053168014539924

Heather, P. (2021). *European traded gas hubs : German hubs about to merge*.

Hu, Y., & Trafalis, T. B. (2011). *New kernel methods for asset pricing: Application to natural gas price prediction new kernel methods for asset pricing* (2).

Investing.com. (Accessed on July 19, 2023). *Crude Oil Historical Data*. https://www.investing.com/commodities/crude-oil-historical-data

Jianwei, E., Ye, J., He, L., & Jin, H. (2019). Energy price prediction based on independent component analysis and gated recurrent unit neural network. *Energy*, *189*. https://doi.org/10.1016/j.energy.2019.116278

Jin, J., & Kim, J. (2015). Forecasting natural gas prices using wavelets, time series, and artificial neural networks. *PLoS ONE*, *10*. https://doi.org/10.1371/journal.pone.0142064

Kaufmann, R. K., & Connelly, C. (2020). Oil price regimes and their role in price diversions from market fundamentals. *Nature Energy*, *5*, 141–149. https://doi.org/10.1038/s41560-020-0549-1

Kwak, H., & An, J. (2014). A first look at global news coverage of disasters by using the gdelt dataset. *6th International Conference, SocInfo*.

Li, J., Wu, Q., Tian, Y., & Fan, L. (2021). Monthly henry hub natural gas spot prices forecasting using variational mode decomposition and deep belief network. *Energy*, *227*. https://doi.org/10.1016/j.energy.2021.120478

Moting, S., Zongyi, Z., Ye, Z., & Donglan, Z. (2019a). Data-driven natural gas spot price forecasting with least squares regression boosting algorithm. *Energies*, *12*. https://doi.org/10.3390/en12061094

Moting, S., Zongyi, Z., Ye, Z., Donglan, Z., & Wenying, W. (2019). Data driven natural gas spot price prediction models using machine learning methods. *Energies*, *12*. https://doi.org/10.3390/en12091680

Naderi, M., Khamehchi, E., & Karimi, B. (2019). Novel statistical forecasting models for crude oil price, gas price, and interest rate based on meta-heuristic bat algorithm. *Journal of Petroleum Science and Engineering*, *172*, 13–22. https://doi.org/10.1016/j.petrol.2018.09.031

Naderi, M., Khamehchi, E., & Karimi, B. (2021, February). *Energy price prediction using data-driven models: A decade review*. https://doi.org/10.1016/j.cosrev.2020.100356

NCEI. (Accessed on July 19, 2023). *National Centers for Environmental Information (NCEI) - Daily Summaries*. https://www.ncei.noaa.gov/access/search/data-search/daily-summaries?dataTypes=TAVG&bbox=54.304,-3.406,41.919,20.984&startDate=2017-10-23T00:00:00&endDate=2023-02-07T23:59:59&stations=GMW00035032

Nguyen, H. T., & Nabney, I. T. (2010). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. *Energy*, *35*, 3674–3685. https://doi.org/10.1016/j.energy.2010.05.013

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). Kerastuner.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Psaropoulos, J. (2022, August). *Timeline: Six months of russia's war in ukraine*. https://www.aljazeera.com/news/2022/8/24/timeline-six-months-of-russias-war-in-ukraine

Qiao, F., Li, P., Deng, J., Ding, Z., & Wang, H. (2015). Graph-based method for detecting occupy protest events using gdelt dataset. *Proceedings - 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015*, 164–168. https://doi.org/10.1109/CyberC.2015.77

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society, 2017*. https://doi.org/10.1155/2017/8180272

Qin, Q., Xie, K., He, H., Li, L., Chu, X., Wei, Y. M., & Wu, T. (2019). An effective and robust decomposition-ensemble energy price forecasting paradigm with local linear prediction. *Energy Economics*, *83*, 402–414. https://doi.org/10.1016/j.eneco.2019.07.026

Refilling gas storage for next winter. (2022, March). European Comission. https://ec.europa.eu/commission/presscorner/detail/en/fs_22_1938

Salehnia, N., Falahi, M. A., Seifi, A., & Adeli, M. H. M. (2013). Forecasting natural gas spot prices with nonlinear modeling using gamma test analysis. *Journal of Natural Gas Science and Engineering*, *14*, 238–249. https://doi.org/10.1016/j.jngse.2013.07.002

Siddiqui, A. W. (2019). Predicting natural gas spot prices using artificial neural network. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. https://doi.org/10.1109/CAIS.2019.8769586

Tang, L., Wu, Y., & Yu, L. (2018). A randomized-algorithm-based decomposition-ensemble learning methodology for energy price forecasting. *Energy*, *157*, 526–538. https://doi.org/10.1016/j.energy.2018.05.146

Tang, Y., Wang, M., Wang, Q., Wang, Z., & Xu, W. (2019). Natural gas price prediction with big data. *2019 IEEE International Conference on Big Data (Big Data)*.

Thakur, A., Kumar, S., & Tiwari, A. (2015). *Hybrid model of gas price prediction using moving average and neural network*.

Tilly, S., Ebner, M., & Livan, G. (2020). Macroeconomic forecasting through news, emotions and narrative. https://doi.org/10.1016/j.eswa.2021.114760

van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman; Hall/CRC.

Viacaba, A., Poursaeidi, M. H., & Kundakcioglu, O. E. (2012). *Natural gas price forecasting via selective support vector regression*. https://www.researchgate.net/publication/289446769

Wang, J., Lei, C., & Guo, M. (2020). Daily natural gas price forecasting by a weighted hybrid data-driven model. *Journal of Petroleum Science and Engineering*, *192*. https://doi.org/10.1016/j.petrol.2020.107240

Wang, J., Cao, J., Yuan, S., & Cheng, M. (2021). Short-term forecasting of natural gas prices by using a novel hybrid method based on a combination of the ceemdan-se-and the pso-als-optimized gru network. *Energy*, *233*. https://doi.org/10.1016/j.energy.2021.121082

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016, September). *Growing pains for global monitoring of societal events*. https://doi.org/10.1126/science.aaf6758

Zhang, C., Xiao, C., & Liu, H. (2019). Spatial big data analysis of political risks along the belt and road. *Sustainability (Switzerland)*, *11*. https://doi.org/10.3390/su11082216

Appendix

TABLE 5.1. Data size used to forecast natural gas price

References	Daily	Weekly	Monthly	Yearly
Naderi et al., 2019	230			
Al-Sharoot and Alramadhan, 2019	527			
J. Wang et al., 2021	1042			
Thakur et al., 2015	1146			
Hu and Trafalis, 2011	1457			
Qin et al., 2019	1678			
Čeperić et al., 2017	1800	260		
Abrishami and Varahrami, 2011	1825			
Berrisch and Ziel, 2022	2068		2091	
Y. Tang et al., 2019	2372			
Salehnia et al., 2013	3803	792	182	
Moting et al., 2019a	4260	886	204	
L. Tang et al., 2018	4873			
Siddiqui, 2019	5470			
Nguyen and Nabney, 2010			130	
Moting et al., 2019			240	
Azadeh et al., 2012				40
Jianwei et al., 2019			420	
Jin and Kim, 2015		726		

TABLE 5.2. Category columns related to actor number one in the articles.

Column	**Description**
Actor1Code	Code representing the primary actor involved in the event
Actor1Name	Name of the primary actor involved in the event
Actor1CountryCode	Country code associated with the primary actor
Actor1KnownGroupCode	Code representing a known group associated with the primary actor
Actor1EthnicCode	Code representing the ethnic group associated with the primary actor
Actor1Religion1Code	Code representing the primary religion associated with the primary actor
Actor1Religion2Code	Code representing the secondary religion associated with the primary actor
Actor1Type1Code	Code representing the primary type of the primary actor
Actor1Type2Code	Code representing the secondary type 1 of the primary actor
Actor1Type3Code	Code representing the secondary type 2 of the primary actor
QuadClass	Code representing the high-level category of the event
Actor1Geo_Type	Type of the geographic location associated with the primary actor
Actor1Geo_FullName	Full name of the geographic location associated with the primary actor
Actor1Geo_CountryCode	Country code associated with the geographic location of the primary actor
Actor1Geo_ADM1Code	ADM1 code associated with the geographic location of the primary actor

TABLE 5.3. Category columns related to actor number two in the articles.

Column	Description
Actor2Code	Code representing the secondary actor involved in the event
Actor2Name	Name of the secondary actor involved in the event
Actor2CountryCode	Country code associated with the secondary actor
Actor2KnownGroupCode	Code representing a known group associated with the secondary actor
Actor2EthnicCode	Code representing the ethnic group associated with the secondary actor
Actor2Religion1Code	Code representing the primary religion associated with the secondary actor
Actor2Religion2Code	Code representing the secondary religion associated with the secondary actor
Actor2Type1Code	Code representing the primary type of the secondary actor
Actor2Type2Code	Code representing the secondary type 1 of the secondary actor
Actor2Type3Code	Code representing the secondary type 2 of the secondary actor
Actor2Geo_Type	Type of the geographic location associated with the secondary actor
Actor2Geo_FullName	Full name of the geographic location associated with the secondary actor
Actor2Geo_CountryCode	Country code associated with the geographic location of the secondary actor
Actor2Geo_ADM1Code	ADM1 code associated with the geographic location of the secondary actor

TABLE 5.4. Category geographic columns.

Column	Description
ActionGeo_Type	Type of the geographic location associated with the action
ActionGeo_FullName	Full name of the geographic location associated with the action
ActionGeo_CountryCode	Country code associated with the geographic location of the action
ActionGeo_ADM1Code	ADM1 code associated with the geographic location of the action

Table 5.5. Integer and decimal columns of GDELT.

Column Name	**Description**
GKGRECORDID	Unique identifier for the GKG record
Date	Date of the event
EventBaseCode	Code representing the base event category
EventRootCode	Code representing the root event category
Actor1Geo_Lat	Latitude of the geographic location for Actor 1
Actor1Geo_Long	Longitude of the geographic location for Actor 1
Actor2Geo_Lat	Latitude of the geographic location for Actor 2
Actor2Geo_Long	Longitude of the geographic location for Actor 2
ActionGeo_Lat	Latitude of the geographic location for the action
ActionGeo_Long	Longitude of the geographic location for the action
NumMentions	Represents the number of mentions of an event in various sources
NumSources	Number of sources reporting the event
NumArticles	Number of articles related to the event
AvgTone	Average tone of the event
GoldsteinScale	Numeric measure indicating a level of conflict or cooperation in political events.

Table 5.6. Pearson correlation with values between -0.7 and 0.7 with natural gas price.

Variable	**Correlation**
Price_ngp	1.000000
Price_oil	0.722026
Open_oil	0.725523
High_oil	0.737906
Low_oil	0.710763
Open_ngp	0.999004
High_ngp	0.999524
Low__ngp	0.999375

TABLE 5.7. Pearson correlation with negative values with natural gas price.

Variable	**Correlation**
Volume_ngp	-0.189236
Wind_temp	-0.040241
Actor2Geo_Type_Idx	-0.023385
Actor1Geo_Type_Idx	-0.020253
QuadClass_Idx	-0.019151
ActionGeo_Type_Idx	-0.018957
GoldsteinScale	-0.017096
Prcp_temp	-0.016933
NumSources	-0.015359
NumArticles	-0.013915
NumMentions	-0.013703
Actor2Type1Code_Idx	-0.010697
Actor1Type1Code_Idx	-0.010565
Actor2Name_Idx	-0.008764
Actor1Name_Idx	-0.008247
Actor2Type2Code_Idx	-0.006322
Actor1Religion1Code_Idx	-0.004739
Actor2Code_Idx	-0.004414
AvgTone	-0.004131
ActionGeo_CountryCode_Idx	-0.003934
Actor1Code_Idx	-0.003631
Actor2Religion1Code_Idx	-0.003412
Actor1Type2Code_Idx	-0.003282
Actor2Geo_CountryCode_Idx	-0.003195
Actor1Geo_CountryCode_Idx	-0.002933
Actor1Religion2Code_Idx	-0.002894
Actor2Religion2Code_Idx	-0.002311
Actor1Type3Code_Idx	-0.001329
Actor2Type3Code_Idx	-0.001270
Actor1EthnicCode_Idx	-0.001230
Actor2EthnicCode_Idx	-0.001021
Actor2CountryCode_Idx	-0.000579
Actor1CountryCode_Idx	-0.000467

TABLE 5.8. Pearson correlation with positive values with natural gas price.

Variable	Correlation
ActionGeo_FullName_Idx	0.000369
Actor2Geo_FullName_Idx	0.000370
Actor1Geo_FullName_Idx	0.000572
Actor2KnownGroupCode_Idx	0.003669
ActionGeo_ADM1Code_Idx	0.005056
Actor2Geo_ADM1Code_Idx	0.005138
Actor1Geo_ADM1Code_Idx	0.005856
Actor1KnownGroupCode_Idx	0.005979
EventRootCode	0.013956
EventBaseCode	0.015387
Change_%_oil	0.017322
Max_temp	0.055238
Avg_temp	0.068243
Min_temp	0.078298
Low_oil	0.710763
Price_oil	0.722026
Open_oil	0.725523
High_oil	0.737906
Open_ngp	0.999004
Low__ngp	0.999375
High_ngp	0.999524
Price_ngp	1.000000
IsRootEvent	NaN

TABLE 5.9. Pearson correlation matrix with values between -0.7 and 0.7 with natural gas price.

Variable	Price_oil	Open_oil	High_oil	Low_oil	Open_ngp	High_ngp	Low__ngp	Price_ngp
Price_oil	1.000000	0.992711	0.994759	0.997807	0.721894	0.722315	0.721408	0.722026
Open_oil	0.992711	1.000000	0.997544	0.994174	0.725619	0.725832	0.725099	0.725523
High_oil	0.994759	0.997544	1.000000	0.992507	0.737587	0.738181	0.737114	0.737906
Low_oil	0.997807	0.994174	0.992507	1.000000	0.710740	0.711108	0.710183	0.710763
Open_ngp	0.721894	0.725619	0.737587	0.710740	1.000000	0.999309	0.999440	0.999004
High_ngp	0.722315	0.725832	0.738181	0.711108	0.999309	1.000000	0.998746	0.999524
Low__ngp	0.721408	0.725099	0.737114	0.710183	0.999440	0.998746	1.000000	0.999375
Price_ngp	0.722026	0.725523	0.737906	0.710763	0.999004	0.999524	0.999375	1.000000

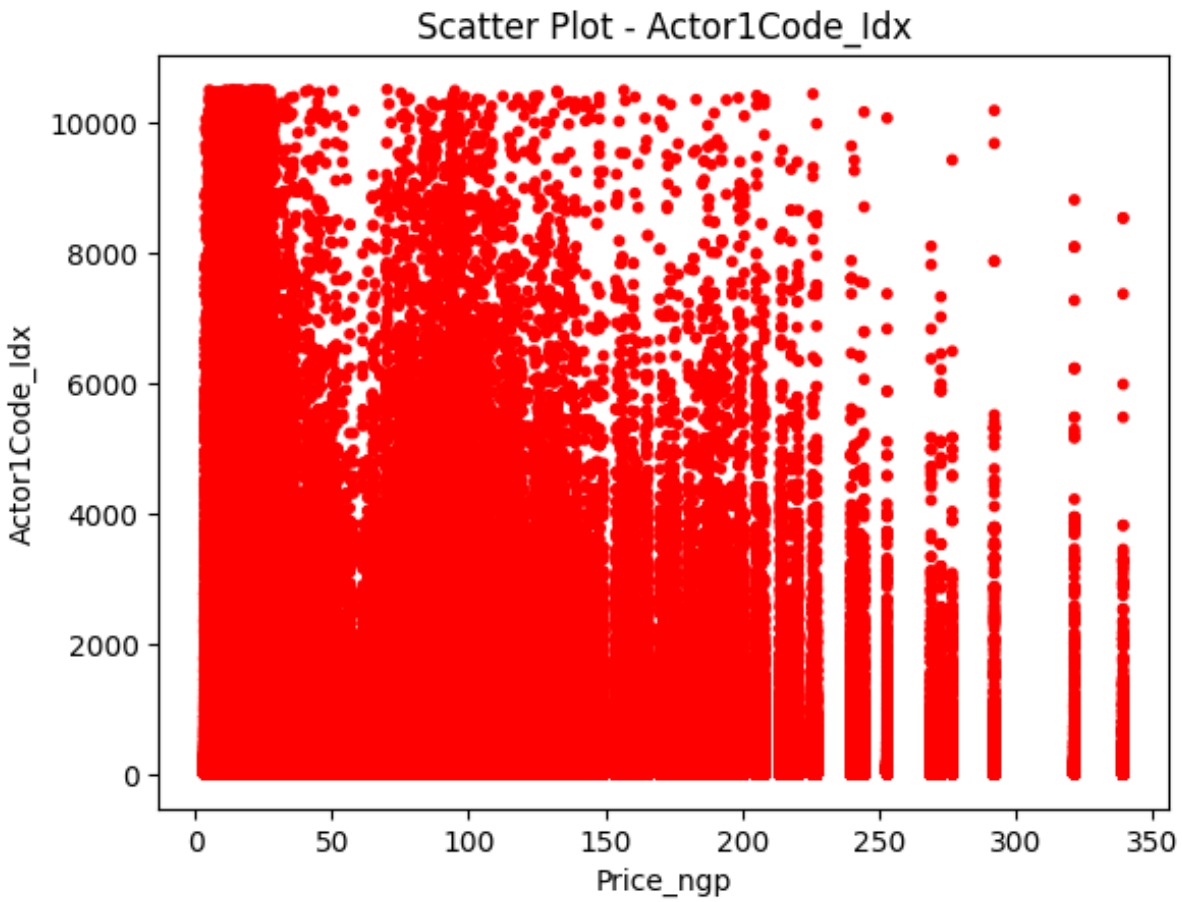

FIGURE 5.1. Scatter plot between natural gas price and Actor1Code_Idx feature.

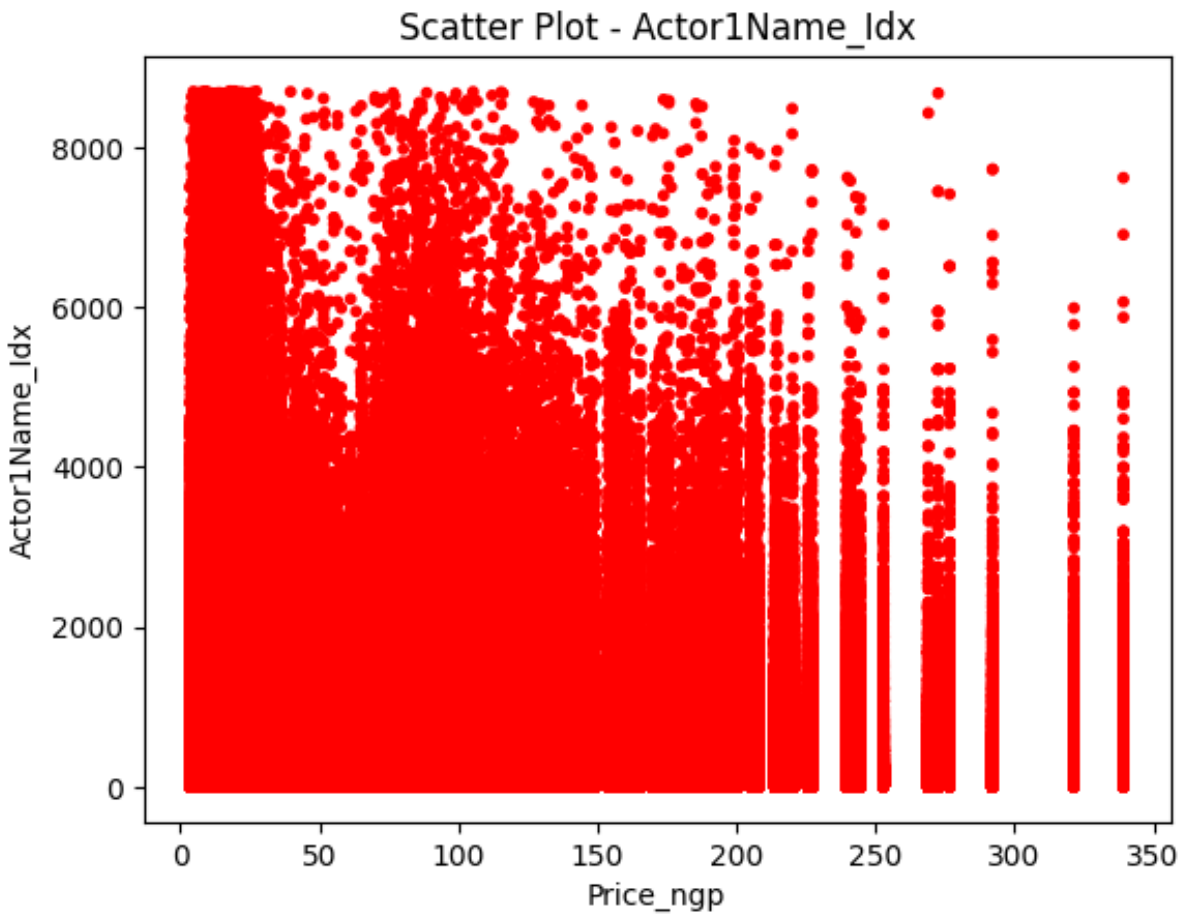

FIGURE 5.2. Scatter plot between natural gas price and Actor1Name_Idx feature.

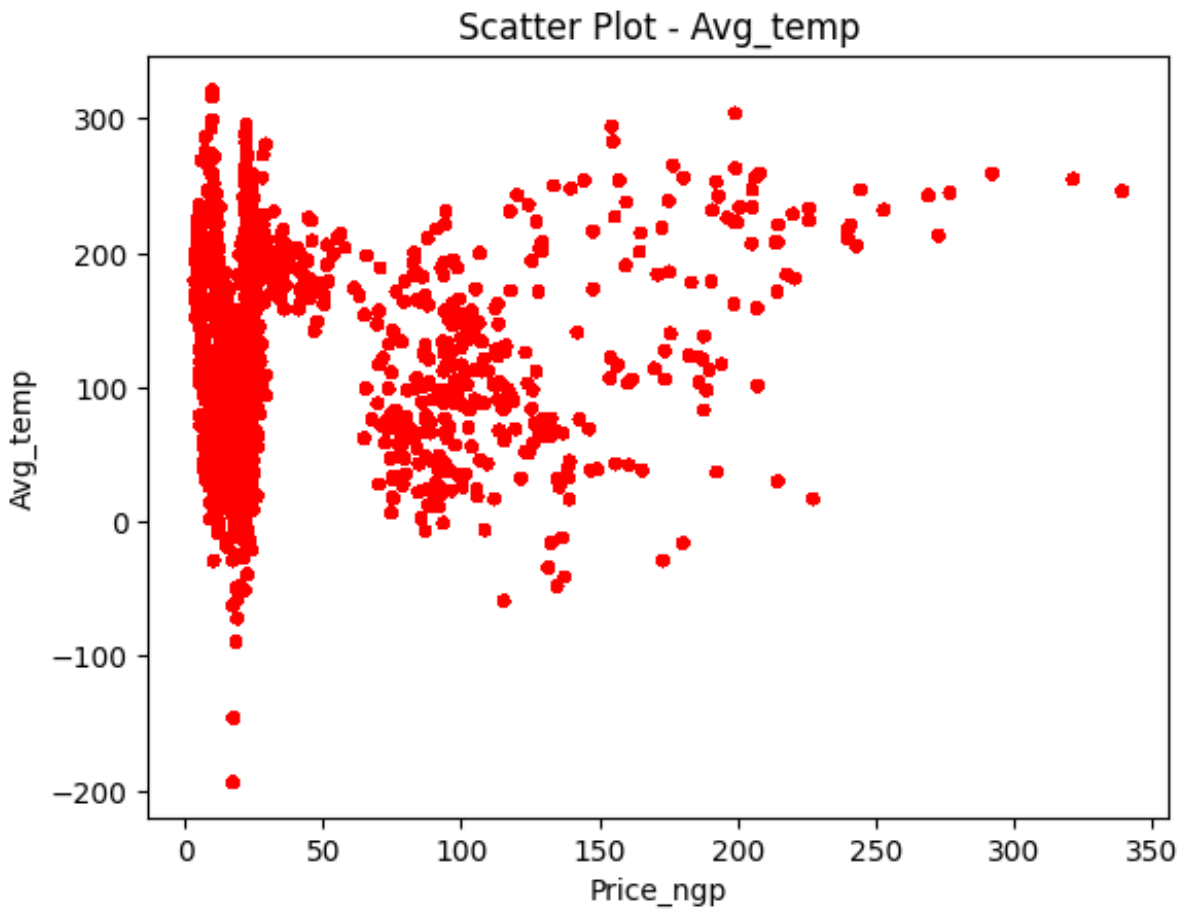

FIGURE 5.3. Scatter plot between natural gas price and average temperature feature.

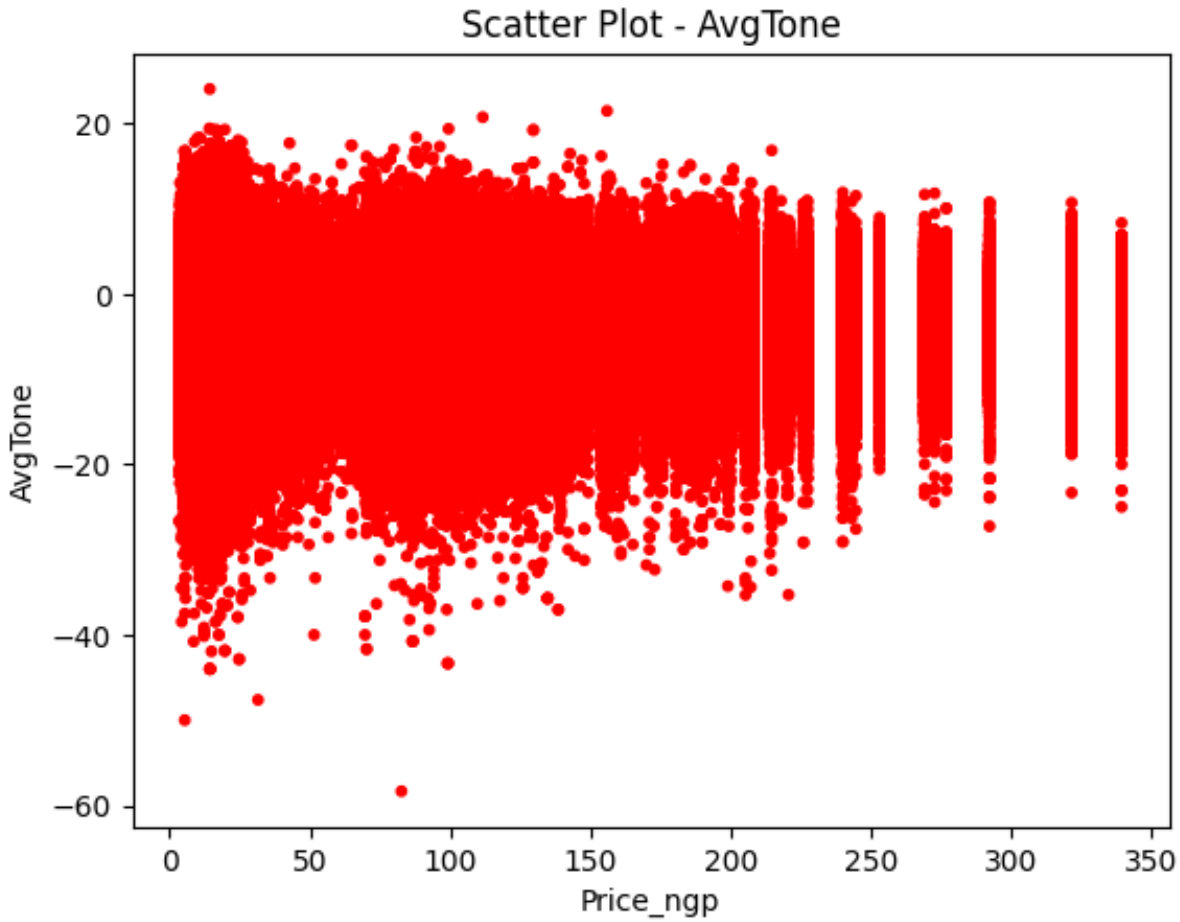

FIGURE 5.4. Scatter plot between natural gas price and average tone feature.

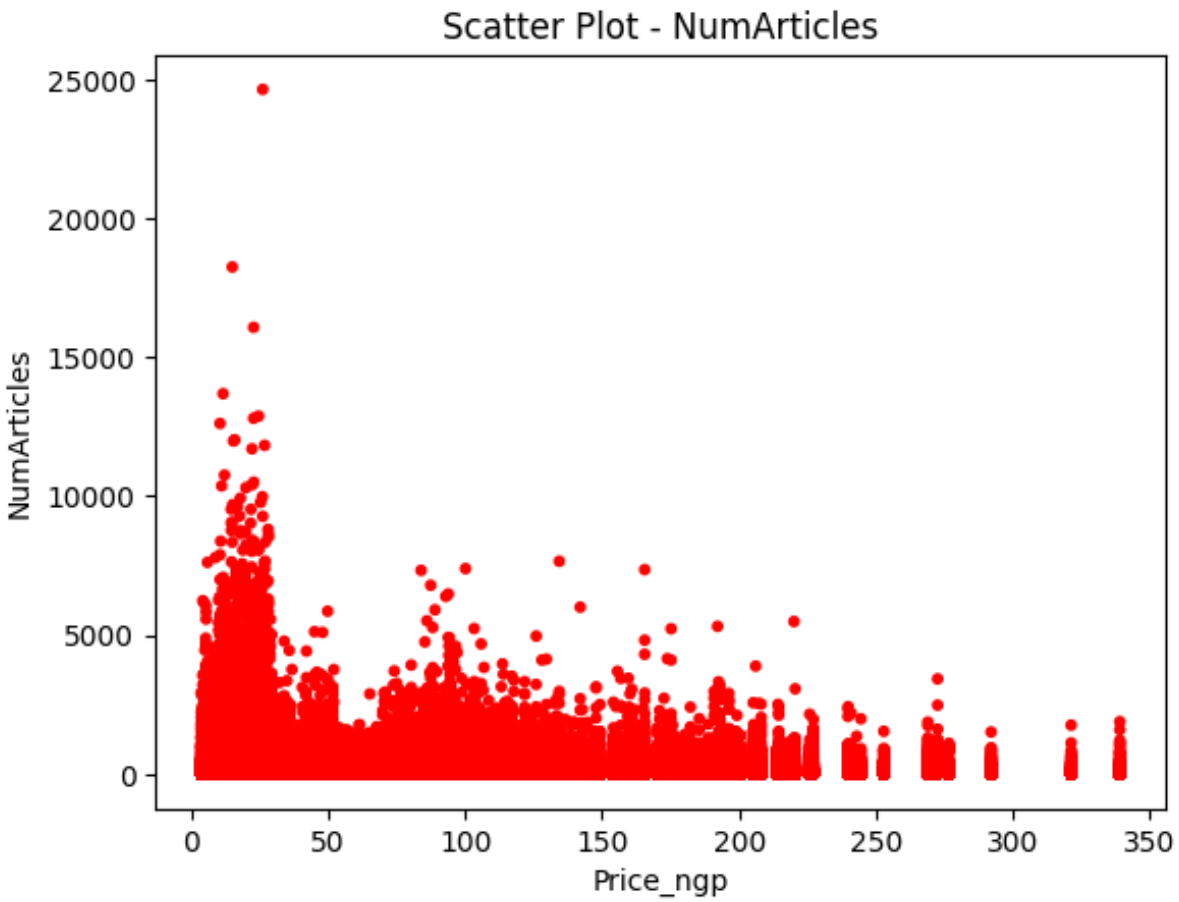

FIGURE 5.5. Scatter plot between natural gas price and the number of articles feature.

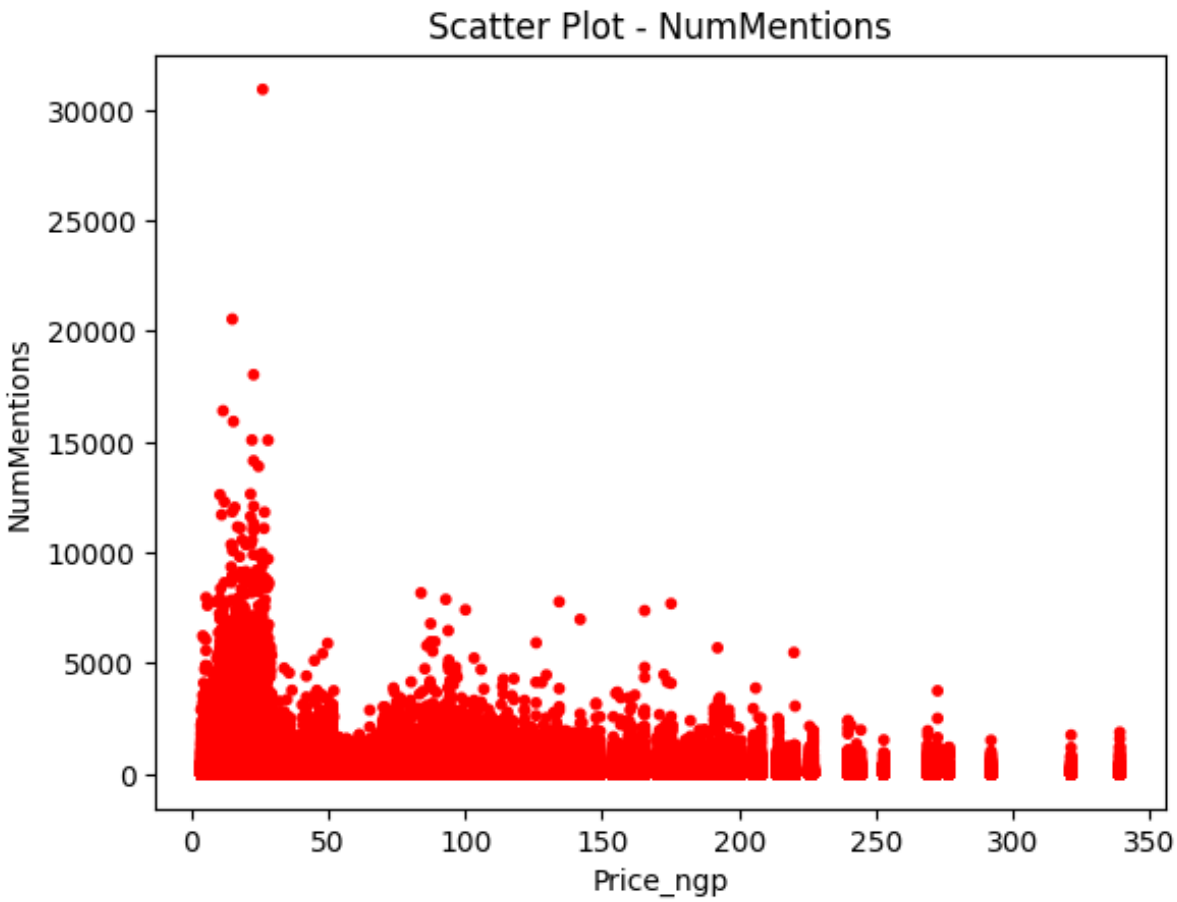

FIGURE 5.6. Scatter plot between natural gas price and the number of mentions feature.

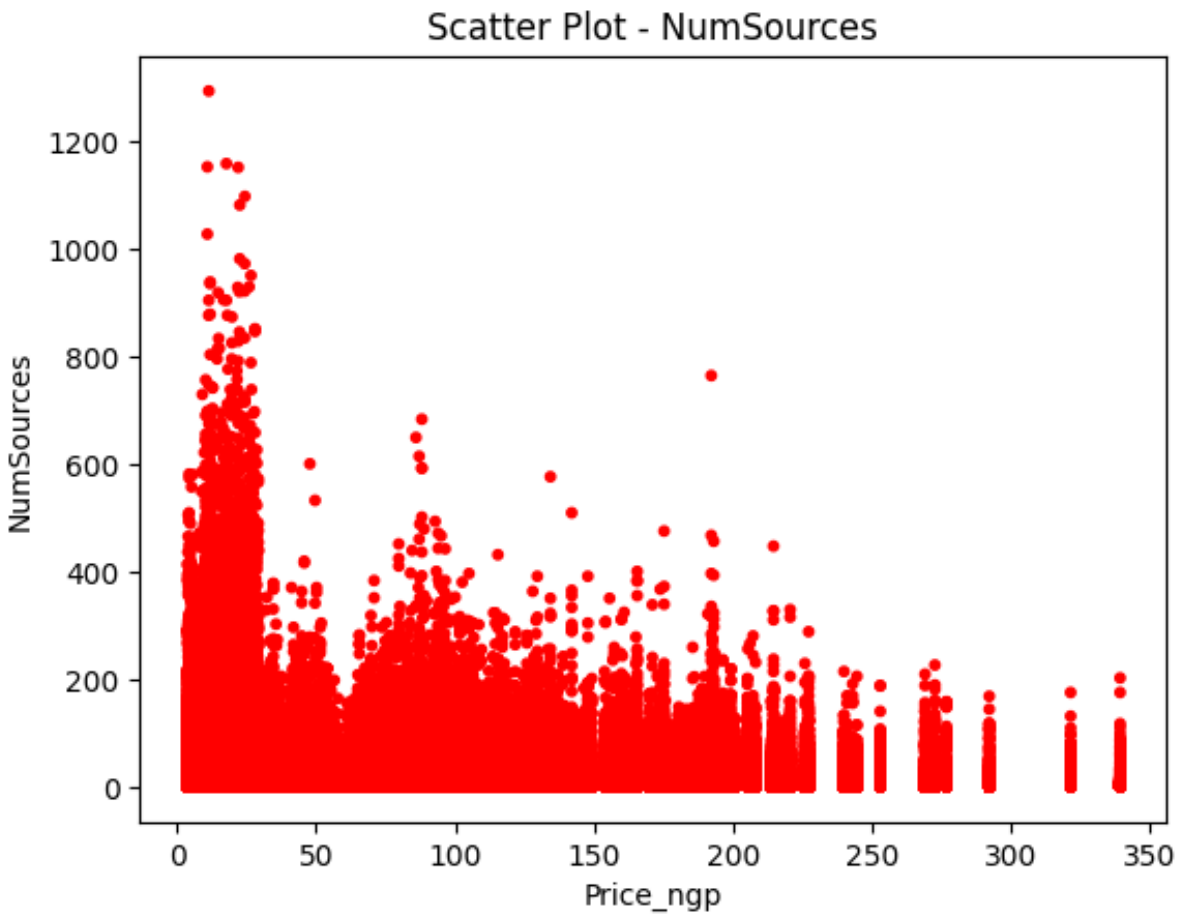

FIGURE 5.7. Scatter plot between natural gas price and the number of sources feature.

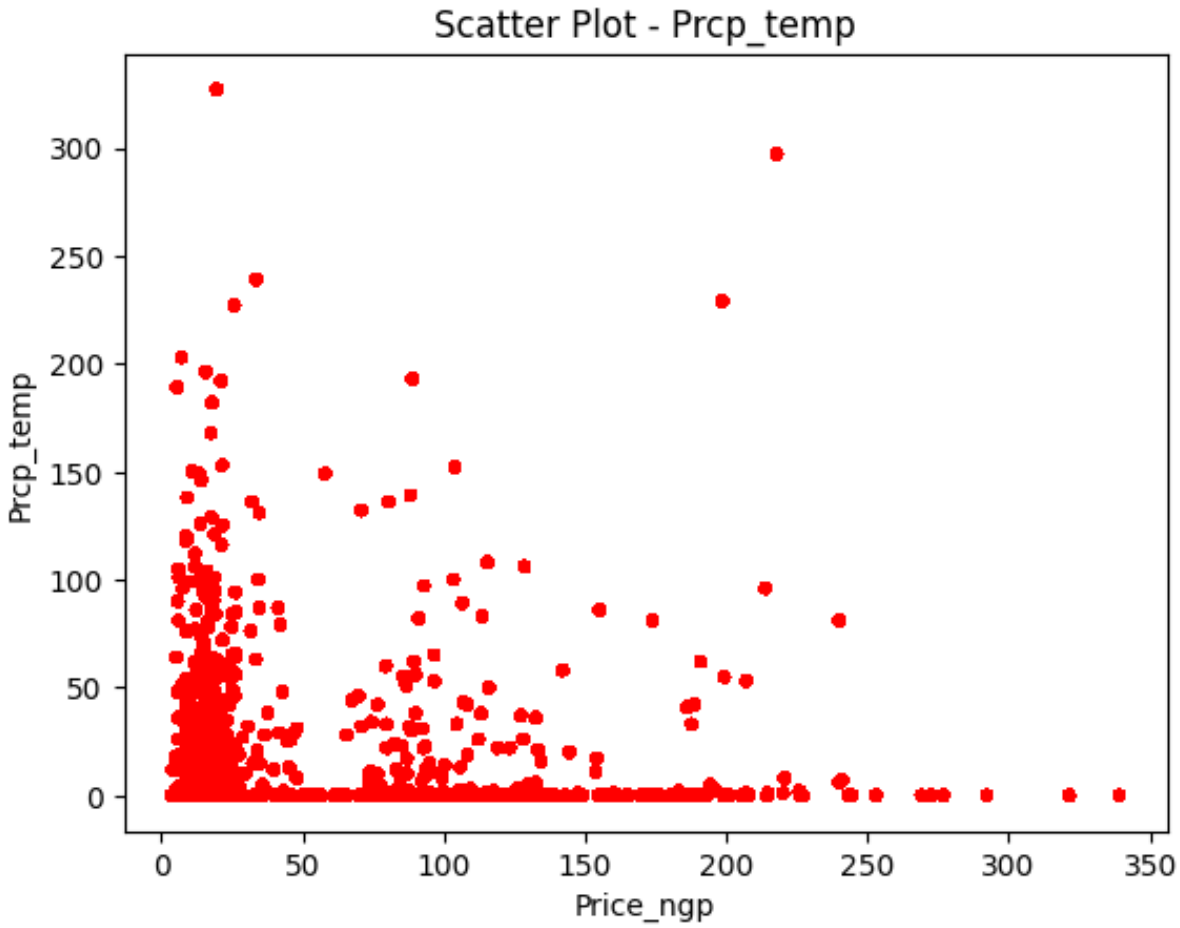

FIGURE 5.8. Scatter plot between natural gas price and precipitation feature.

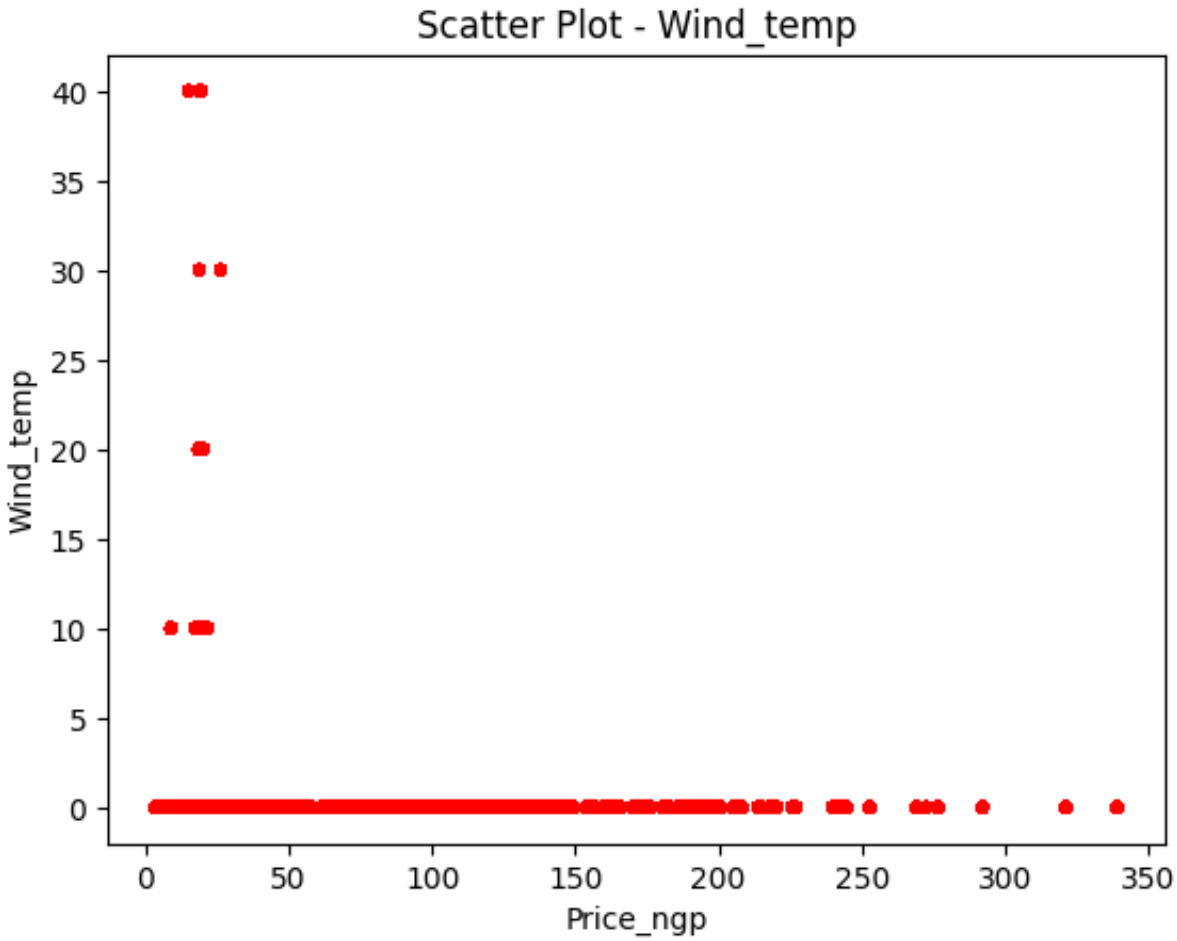

FIGURE 5.9. Scatter plot between natural gas price and wind feature.

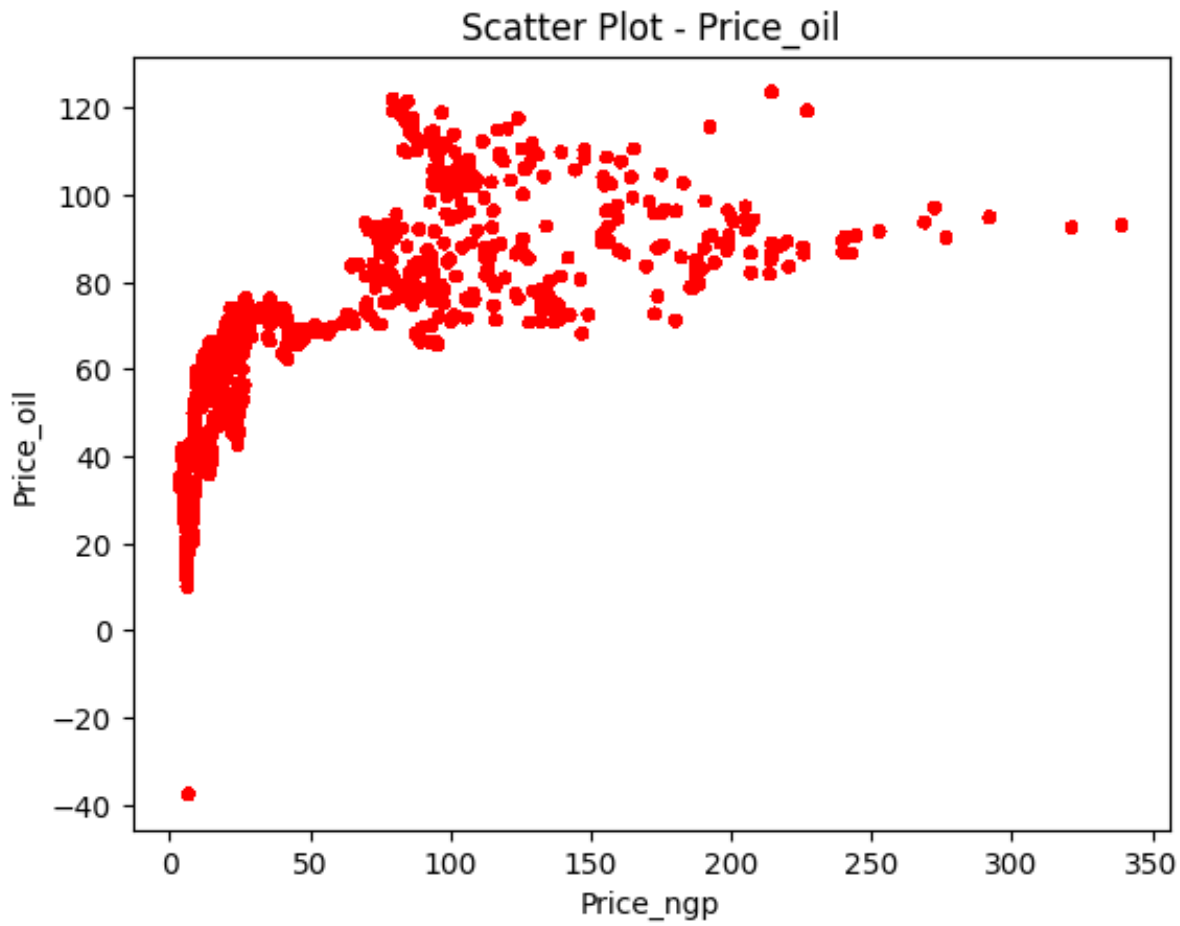

FIGURE 5.10. Scatter plot between natural gas price and crude oil price feature.

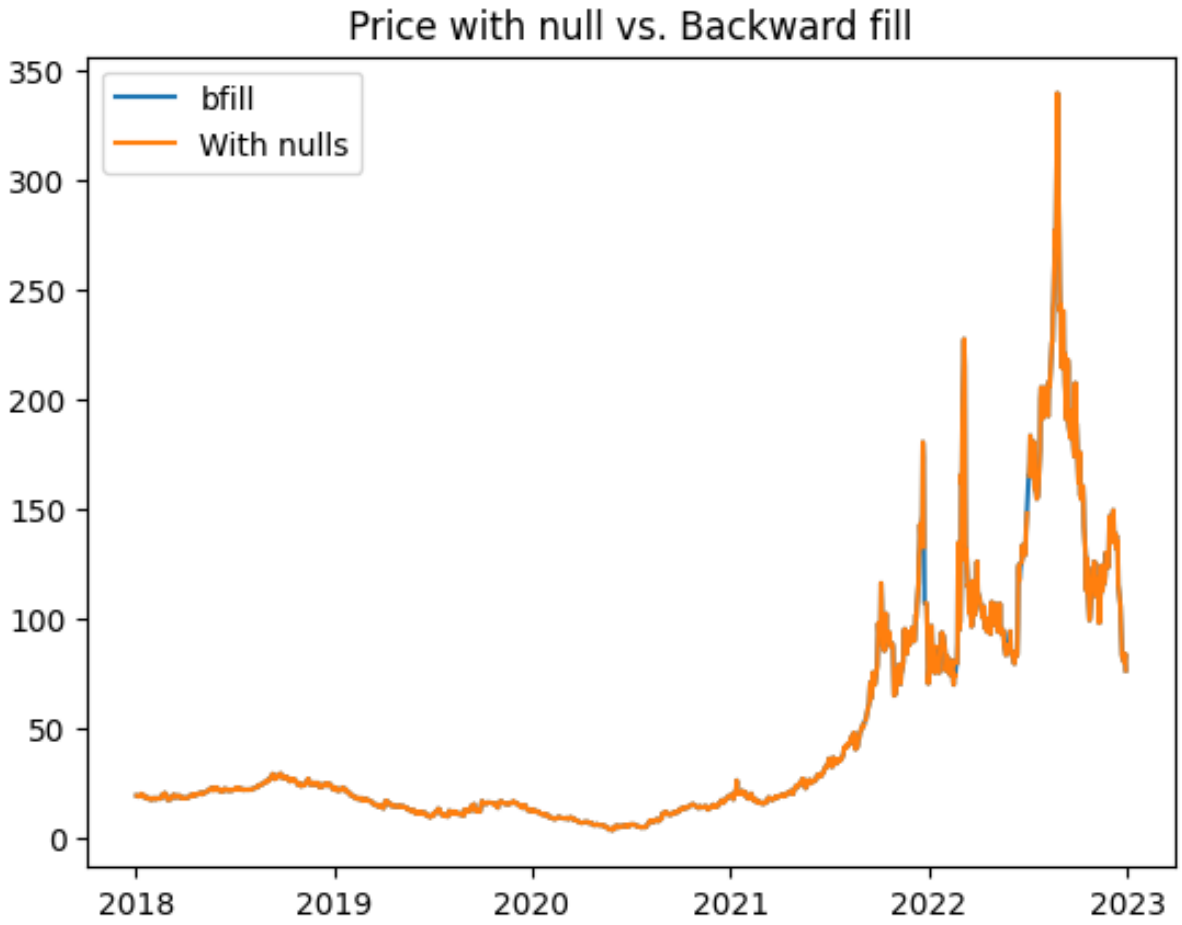

FIGURE 5.11. Visualization of natural gas prices with null values filled with the backward values method.

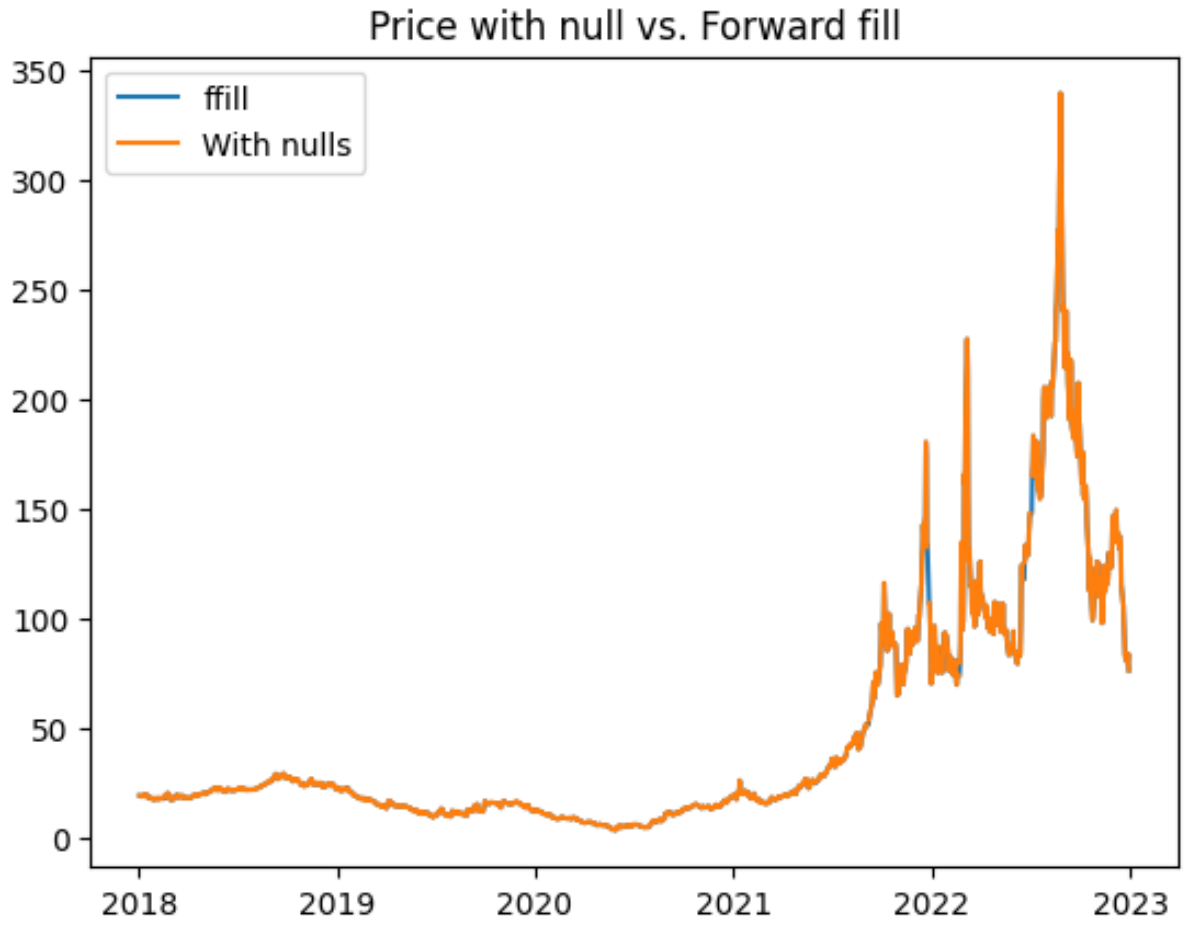

FIGURE 5.12. Visualization of natural gas prices with null values filled with the forward values method.

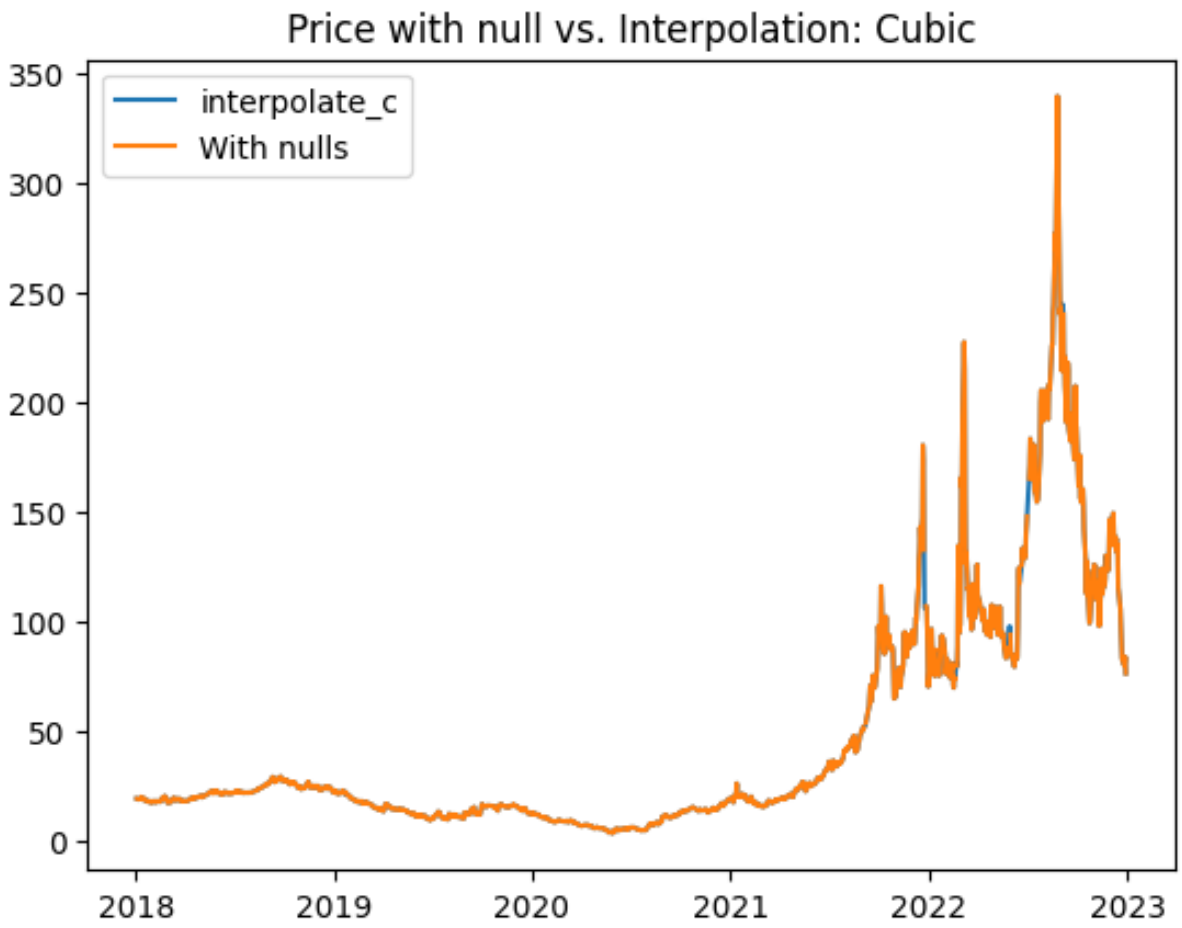

FIGURE 5.13. Visualization of natural gas prices with null values filled with the interpolation cubic values method.

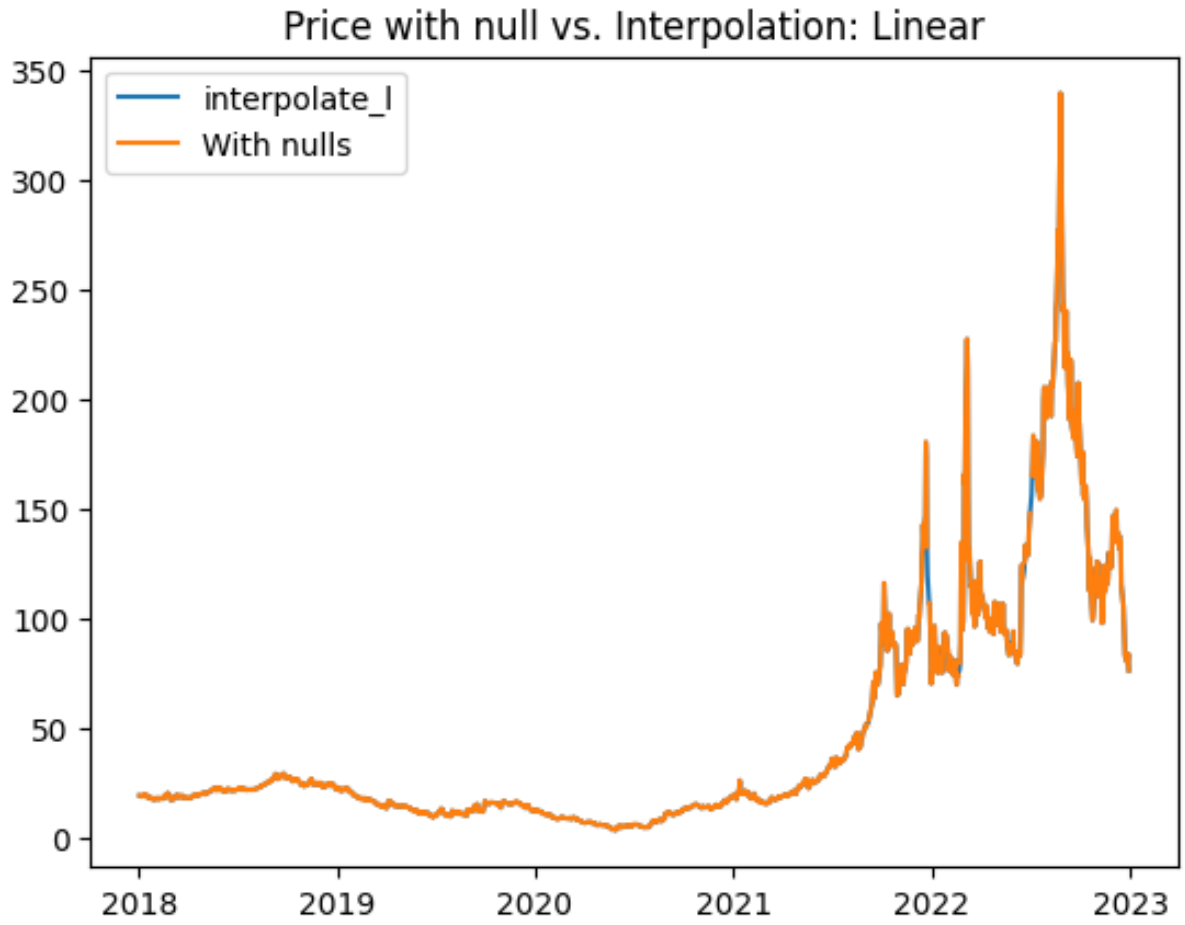

FIGURE 5.14. Visualization of natural gas prices with null values filled with the interpolation linear values method.

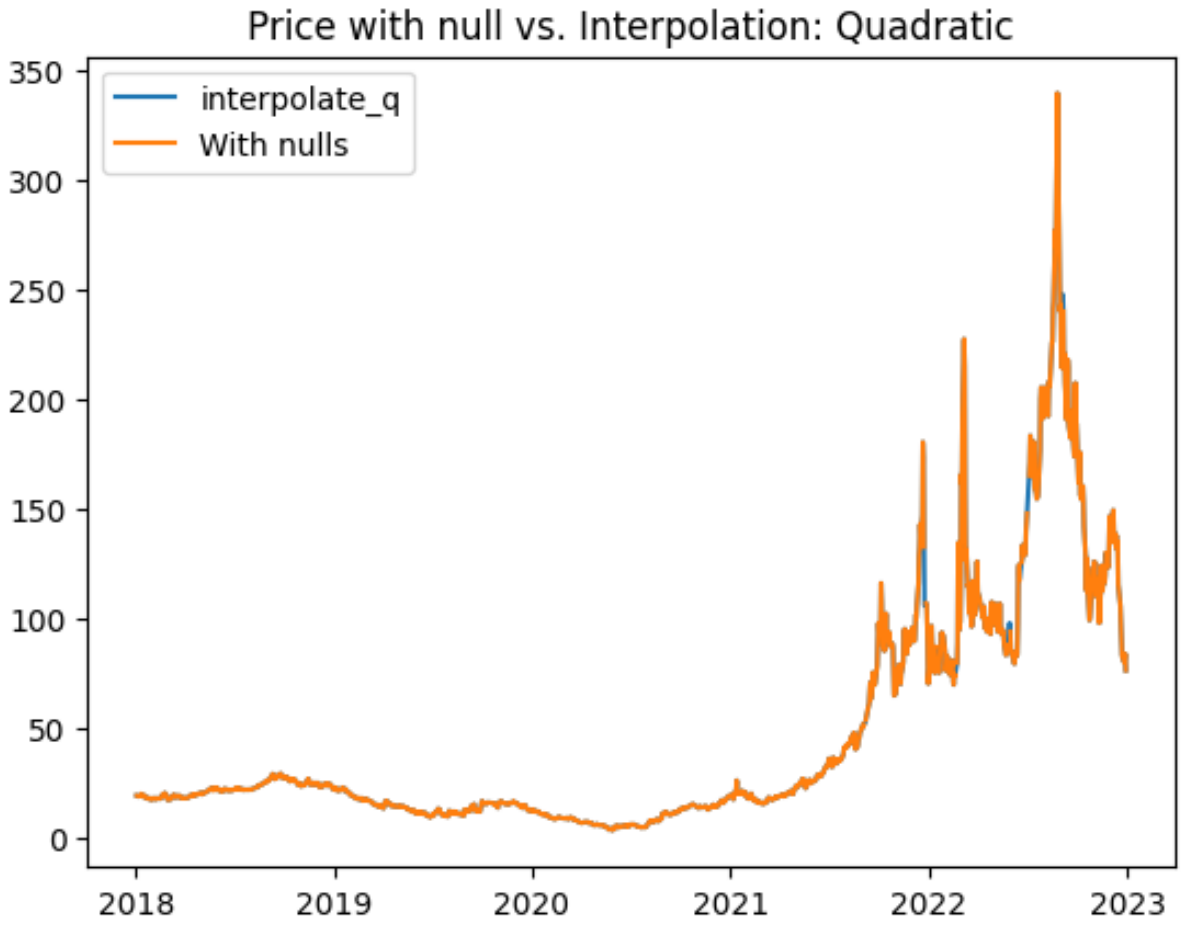

FIGURE 5.15. Visualization of natural gas prices with null values filled with the interpolation quadratic values method.

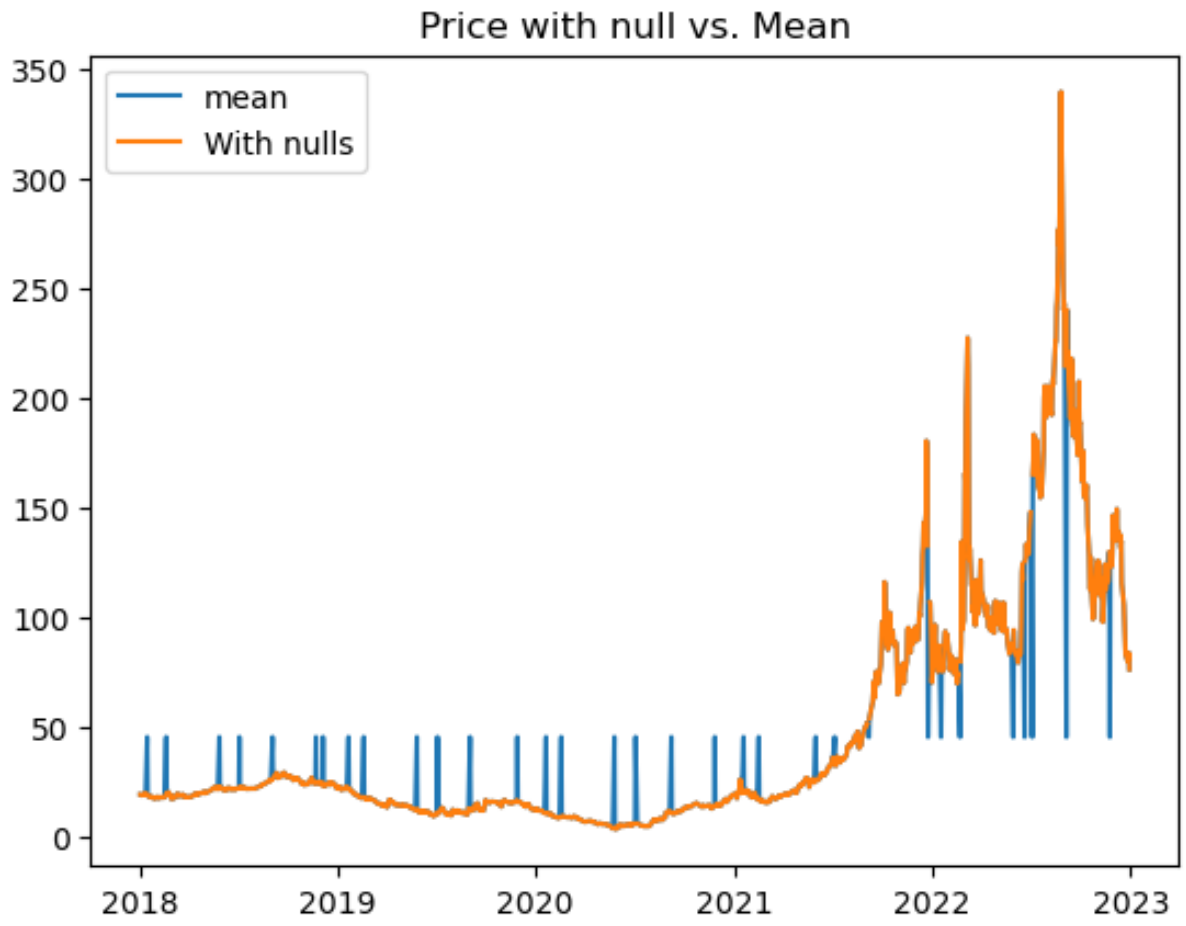

FIGURE 5.16. Visualization of natural gas prices with null values filled with mean value.

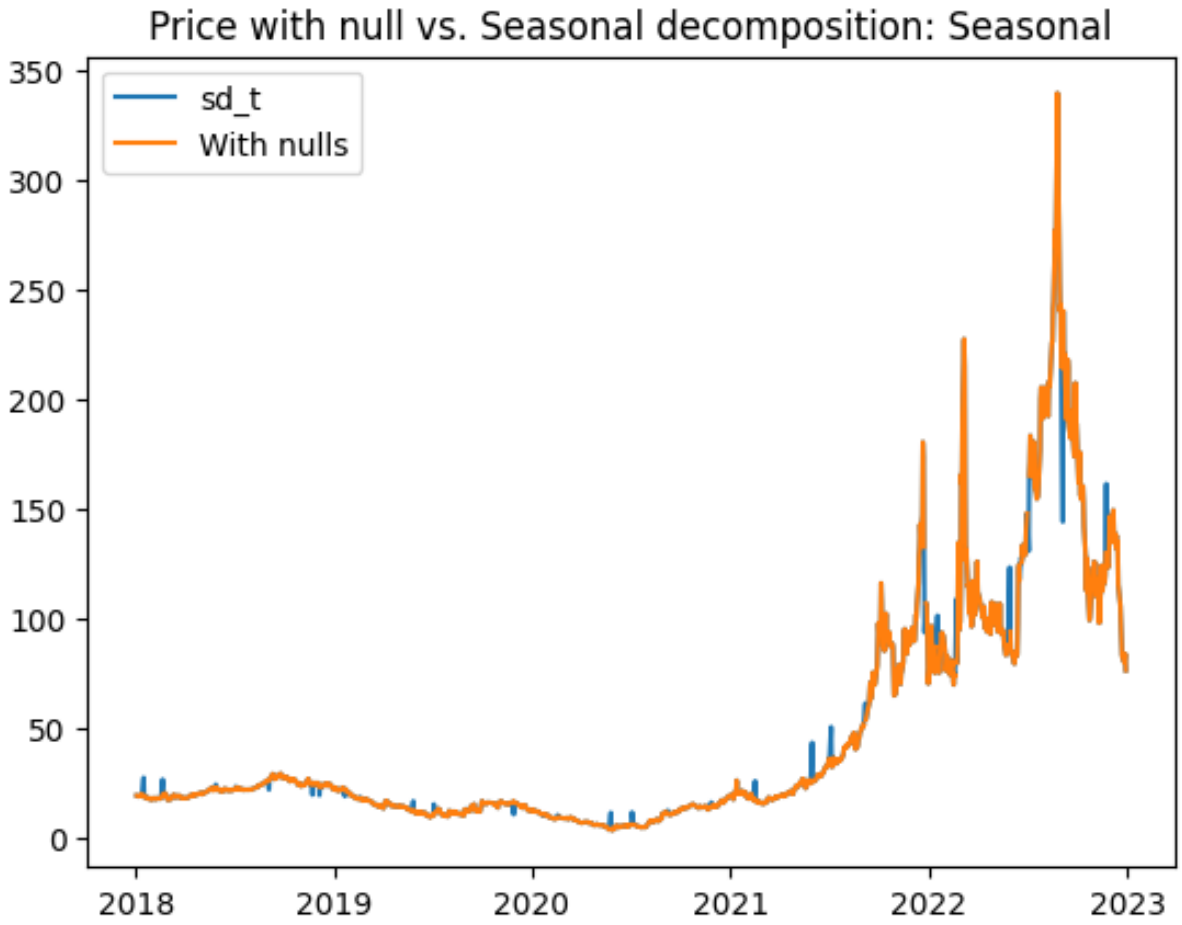

FIGURE 5.17. Visualization of natural gas prices with null values filled with the seasonal values from the seasonal decomposition method.

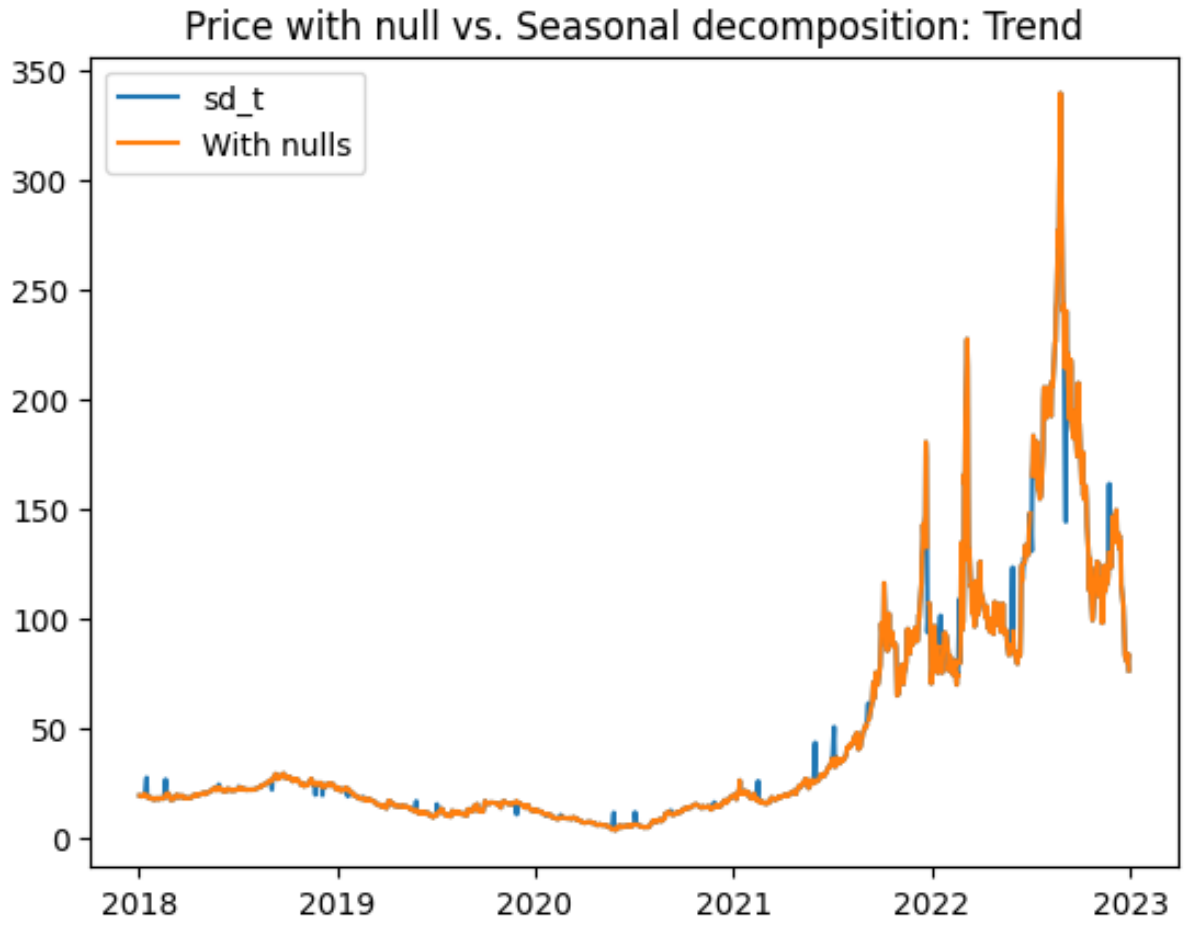

FIGURE 5.18. Visualization of natural gas prices with null values filled with the trend values from the seasonal decomposition method.

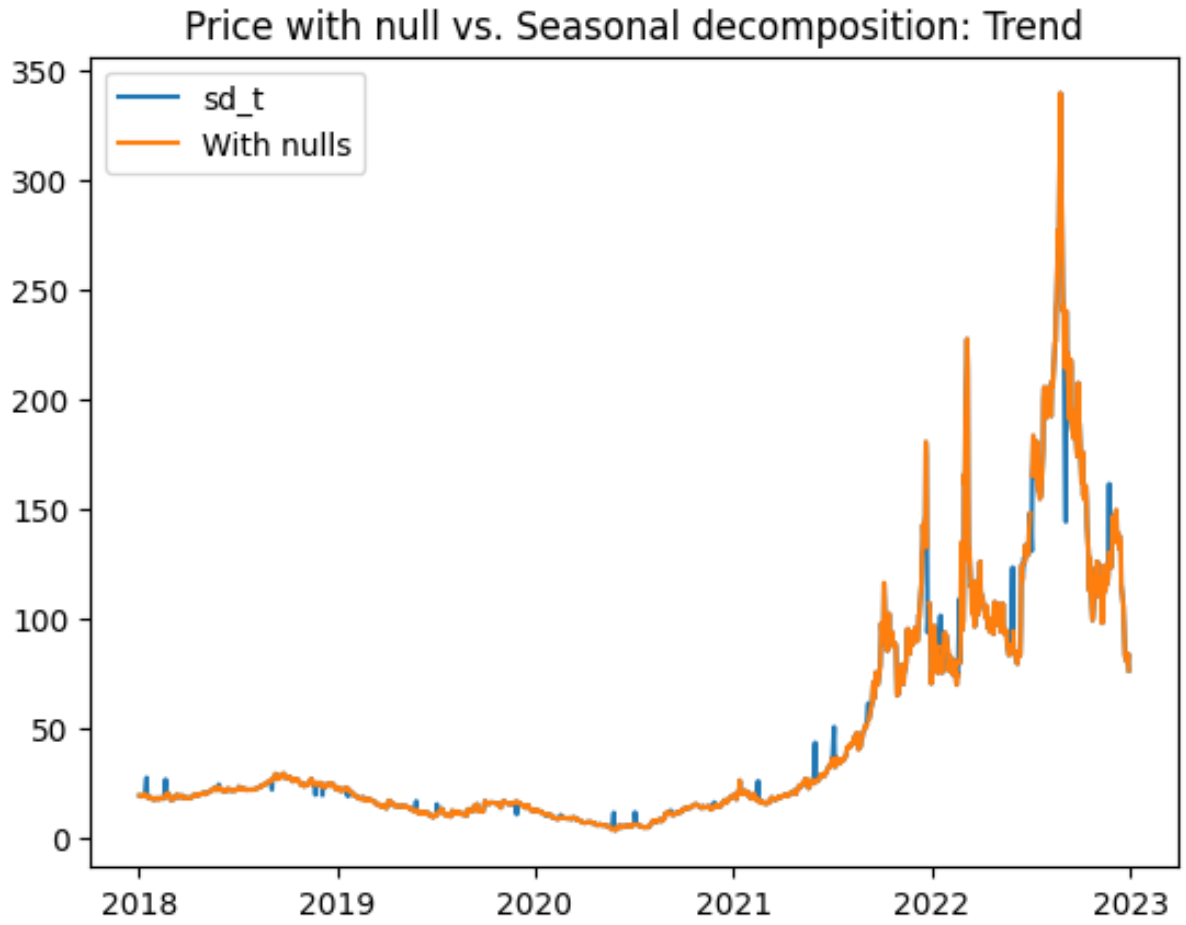

FIGURE 5.19. Visualization of natural gas prices with null values filled with the trend values from the seasonal decomposition method.

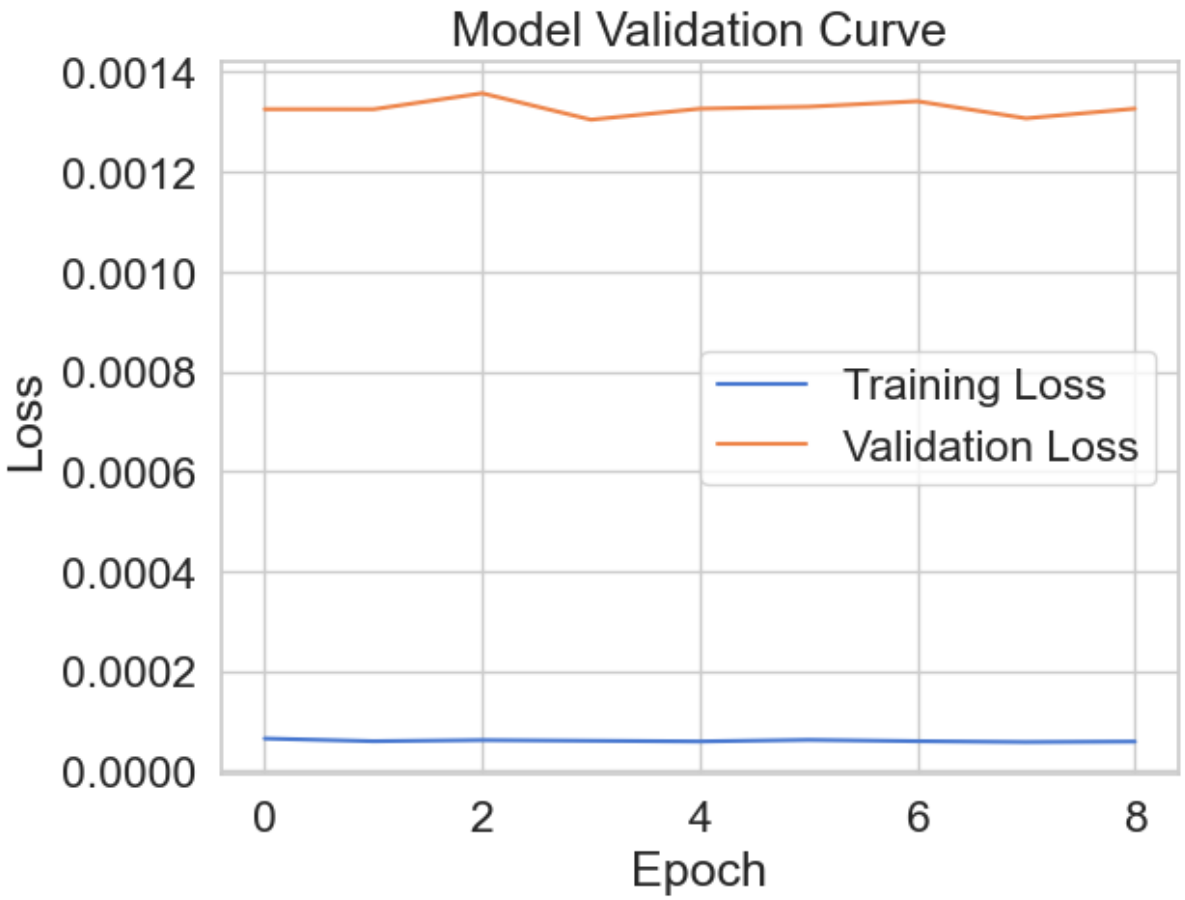

FIGURE 5.20. Validation curve of RNN model with natural gas price, crude oil price, and average tone as features, and lag equal to 10.

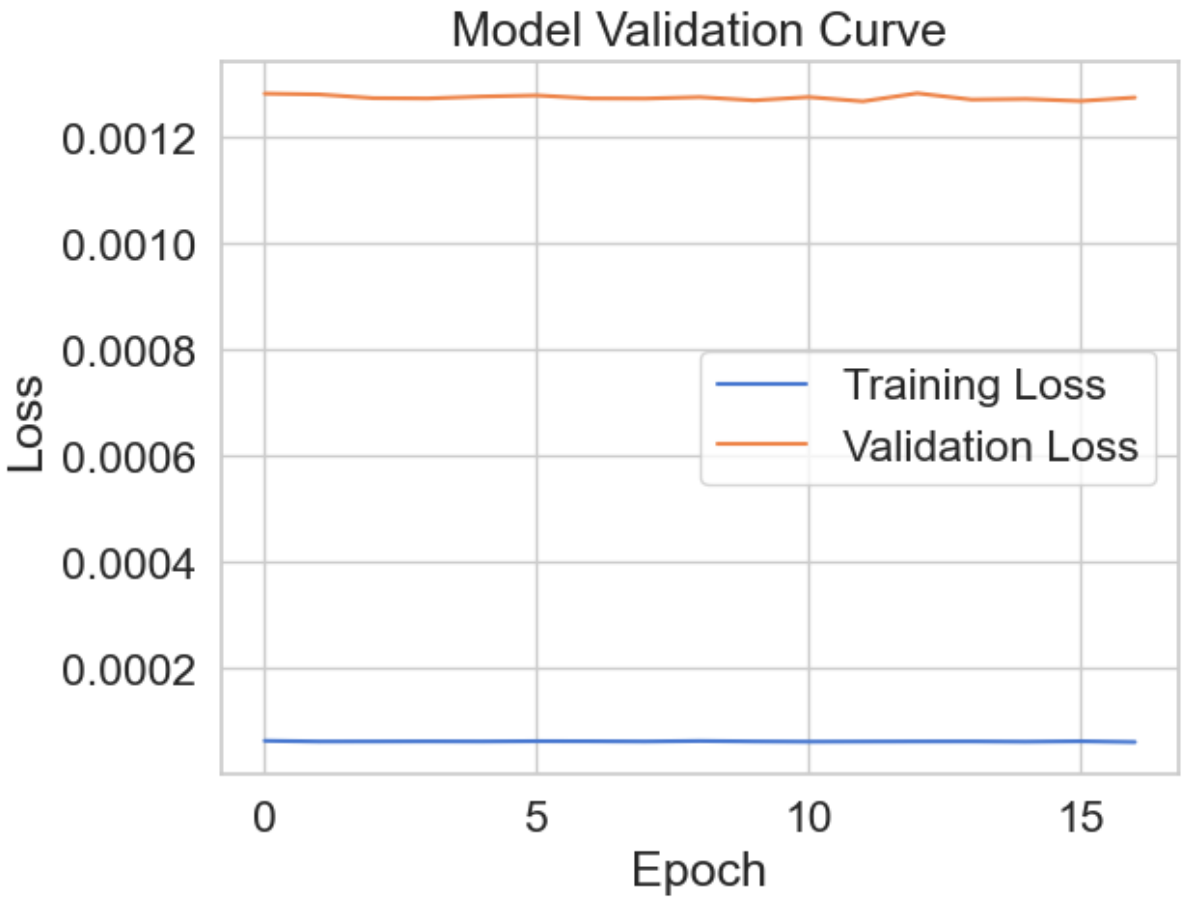

FIGURE 5.21. Validation curve of GRUNN model with natural gas price and crude oil price as features, and lag equal to 5.

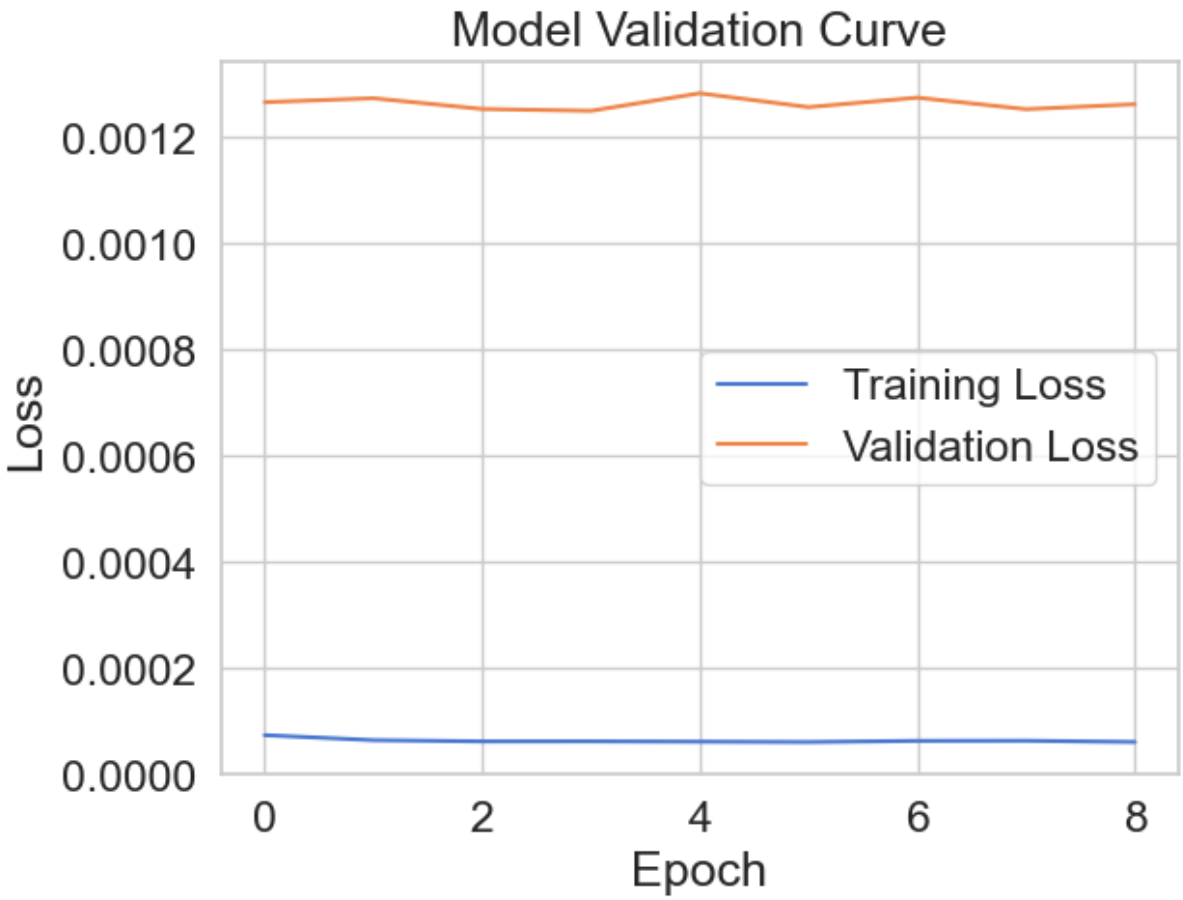

FIGURE 5.22. Prediction plot of RNN model with natural gas price and crude oil price as features, and lag equal to 10.

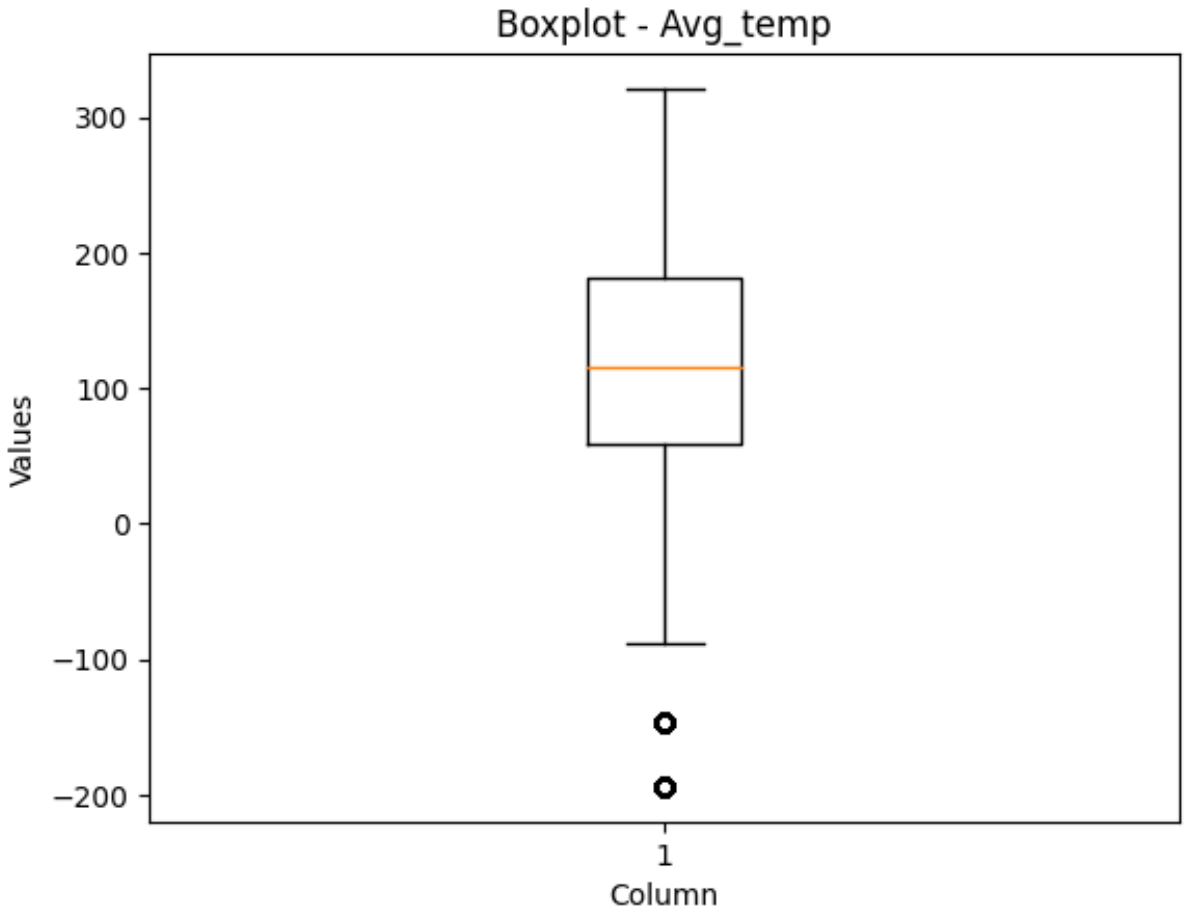

FIGURE 5.23. Boxplot of average temperature.

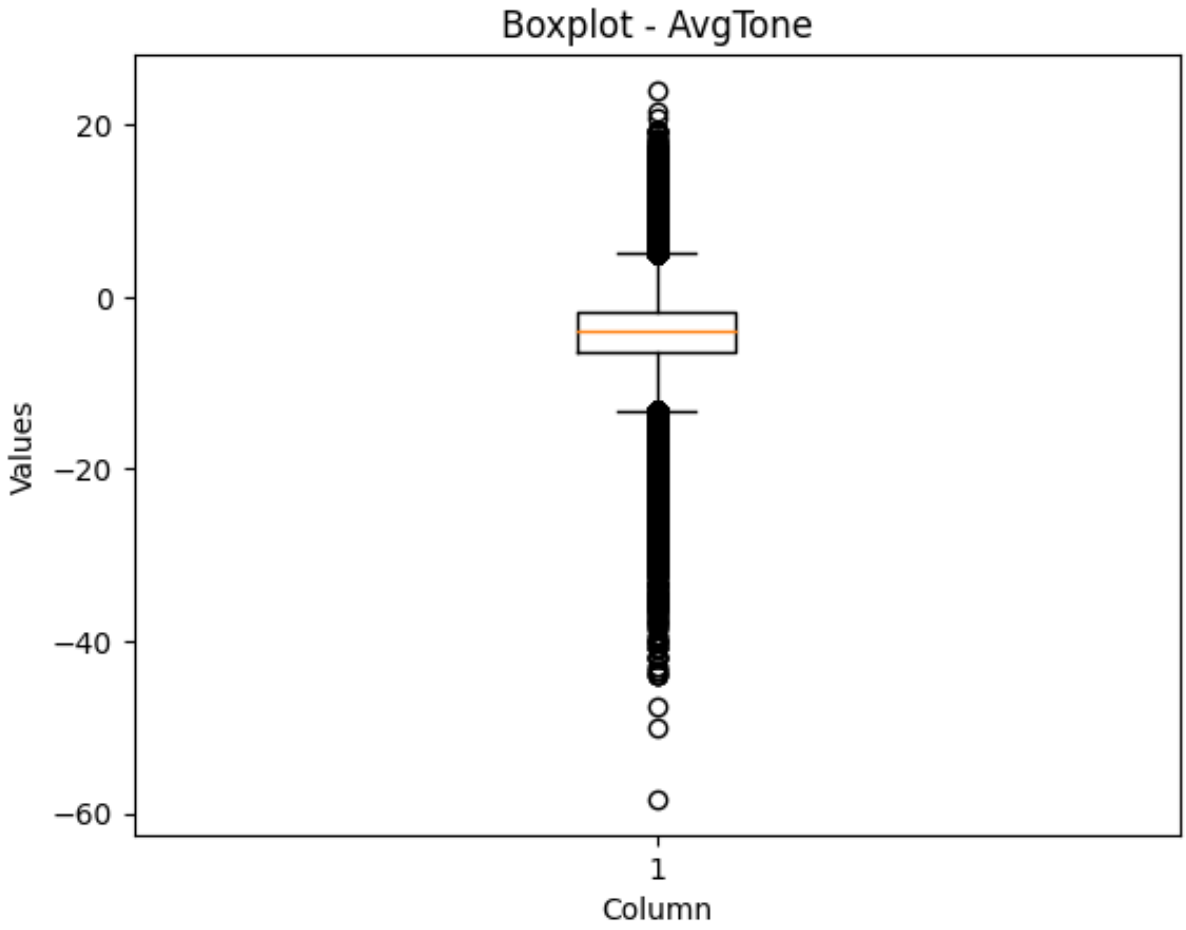

FIGURE 5.24. Boxplot of average tone.

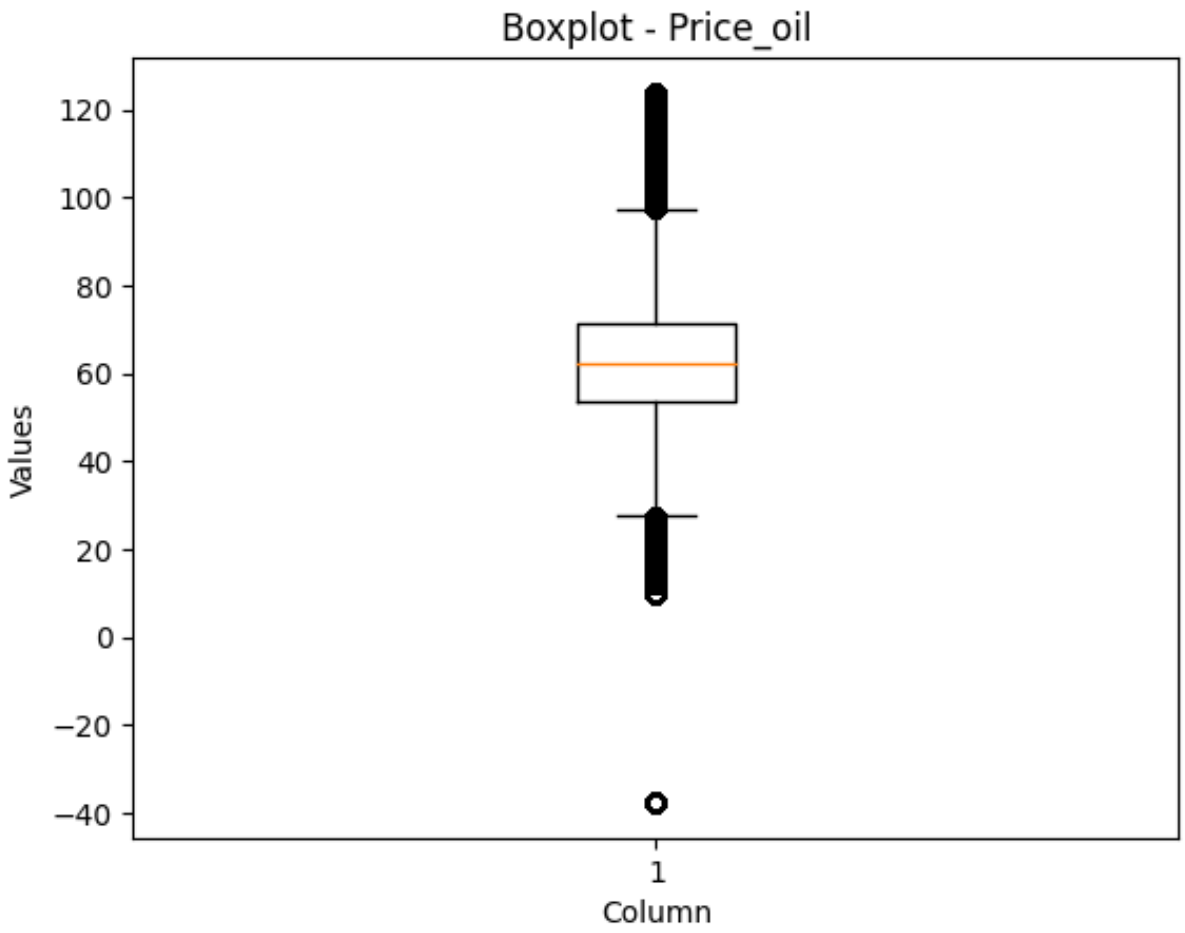

FIGURE 5.25. Boxplot of crude oil price.