Forecasting Crude Oil and Gasoline Prices: Analysis of Time-Series Data from 1986-2016

Project Group 6
Arun Srivatsan Swaminathan, Kannan Walter, Mason Zepnick, Prageeshwar Chandran

Problem Statement

The prices of crude oil and gasoline play a crucial role in the global economy, affecting various industries and carrying significant geopolitical implications. The main objective of this analysis is to use historical data to forecast future prices of crude oil and gasoline. To achieve this goal, we will need to identify any patterns, trends, or seasonality present in the data and select appropriate forecasting methods to model these patterns. Furthermore, we will evaluate the accuracy of our forecasts and identify any potential sources of error. Our analysis will also investigate whether any seasonal patterns or trends exist in the data and assess whether we can make accurate predictions about future prices of crude oil and gasoline based on historical data. Throughout this report, we will employ the R programming language to analyze and visualize the trends in crude oil and gasoline prices.

Motivations

In the oil and gas industry, making informed business and investment decisions relies on solving engineering problems that are unique and specific. Companies in this industry are particularly susceptible to fluctuations in the prices of crude oil and gasoline, so it is essential to accurately forecast how these prices will behave. With the right forecasting tools, companies can gain a critical competitive edge that directly affects their profitability. To seize on a potential uptick in oil costs, businesses might opt to invest in new drilling or exploration initiatives when they can effectively anticipate the trend.

Another primary motivation behind this project was to inform public policy decisions related to transportation and energy. For example, if it is forecasted that gasoline prices will increase in the upcoming years, policymakers would opt to implement policies that incentivize alternatives such as transportation or fuel.

Lastly, studying trends in crude oil and gasoline prices over time can provide insight into broader economic and global trends. Major geopolitical events such as wars or changes in trading policies can have a significant impact on oil prices, and analyzing these trends can help policy makers and industry leaders better understand the potential impacts of such events.

Objectives

The main objective of the project is to forecast future crude oil and gasoline prices based on historical data. Also, we emphasize answering the following questions as a part of learning and analyzing the data in depth to understand the price distribution and comparison over a period of time.

- Are there any seasonal patterns or trends in the data?
- Can we accurately predict future prices of crude oil and gasoline based on historical data? What is the level of uncertainty in our predictions?
- How do political events, such as wars or sanctions, affect crude oil and gasoline prices? Can we quantify the impact of these events on the price series?
- Can we identify any outliers or anomalies in the data, and if so, what caused them.

Trend Analysis

We start the analysis by obtaining summary statistics of the crude oil prices using the "summary" function in R. This provides us with key measures such as mean, median, minimum, maximum, and quartiles for each type of crude oil. Next, we visualize the trends in crude oil and gasoline prices using line plots. We use the "ggplot" function in R to create line plots for each type of crude oil and gasoline, with the years on the x-axis and the average prices on the y-axis. The line plots show the trends in crude oil and gasoline prices over the years, providing insights into price movements and patterns. The results reveal that for various crude oil and gasoline prices (Cushing, US, NY), there is an increasing trend observed. Although an increasing trend is observed, there is a dip in the data from 2008 - 2009 and from 2014 to 2016, as the economy had a recession and there was excessive oil production respectively.

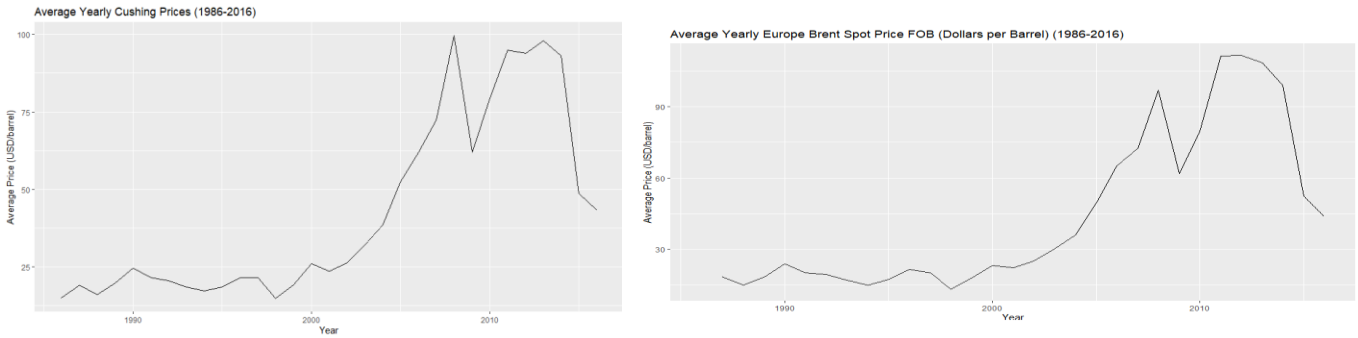


Fig 1 Trend Analysis for Cushing and Europe Brent prices

Decade-wise Analysis

We further analyze the crude oil and gasoline prices by grouping them into decades using the "Decade" variable created in R. We use box plots to compare the average prices of crude oil and gasoline for each decade. The box plots provide information on the median, quartiles, and outliers of the data, allowing us to identify any significant differences in prices between decades. The decade wise analysis gives us a clear insight on how the prices range over the span of 30 years in a range of 4 decades from 1980 to 2010 for Cushing, US and NY crude oil and gasoline prices and it is seen that prices spiked during the recession period (2008) as seen an outlier in the data for the decade 2010.

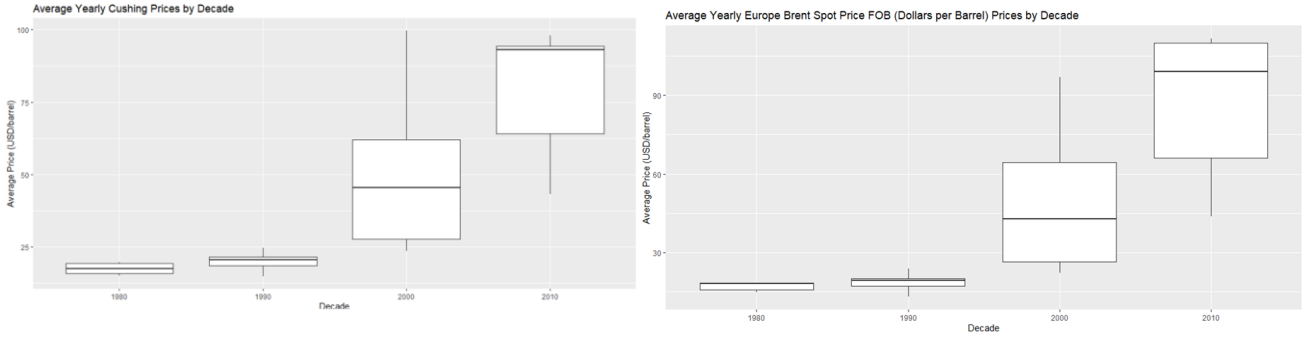


Fig 2 Decade-wise Analysis for Cushing and Europe Brent Prices

Distribution Analysis

We also analyze the distribution of crude oil and gasoline prices using histograms. We create histograms for each type of crude oil and gasoline, showing the frequency of prices within certain price ranges.

The histograms provide insights into the distribution of prices and can help identify any skewness or asymmetry in the data.

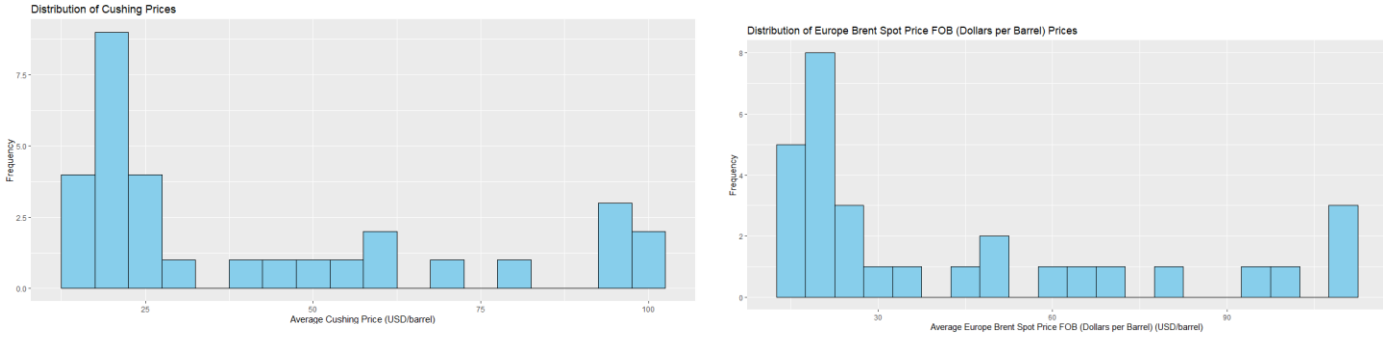


Fig 3 Distribution Analysis of Cushing and Europe Brent Prices

Based on the two histograms created to show the frequency of prices from both Cushing and Europe we notice most values in the data are on the lower end. This can be attributed to the dataset's values dating back to 1986 when the prices were more stable and lower. Observing the rest of the data shows that the prices are around 50 as often as they jump up to 100. While this seems like the data has extreme variance this can be explained by comparing the histograms to the trend analysis and noticing that these large spikes in prices are always followed by an extreme decline. The main two instances of this happening are in 2008 and 2016. In 2008, the financial crisis occurred and caused prices to crash and in 2016 the industry experienced a glut. Meaning that there was an abundance of crude oil and gasoline in the market forcing the price to crash again.

Overall, this shows us that when the prices of crude oil and gasoline peak there is likely to be a global event that causes an extreme drop off in price. This is useful information as it relates to one of our key motivations behind the project which is to provide insight for public policy decisions as well as predicting global events.

Decomposition Analysis

The Decomposition analysis is done first for the original yearly data set. Since we faced some challenges in decomposing the model for the yearly dataset, we decided to go with a monthly dataset to get a better understanding of seasonality, trend, cyclic and random components. The visualization of Cushing and Europe Brent Prices gave us an clear idea that there was an increasing additive trend, there was seasonality in the dataset in cyclic pattern and also randomness in the data.

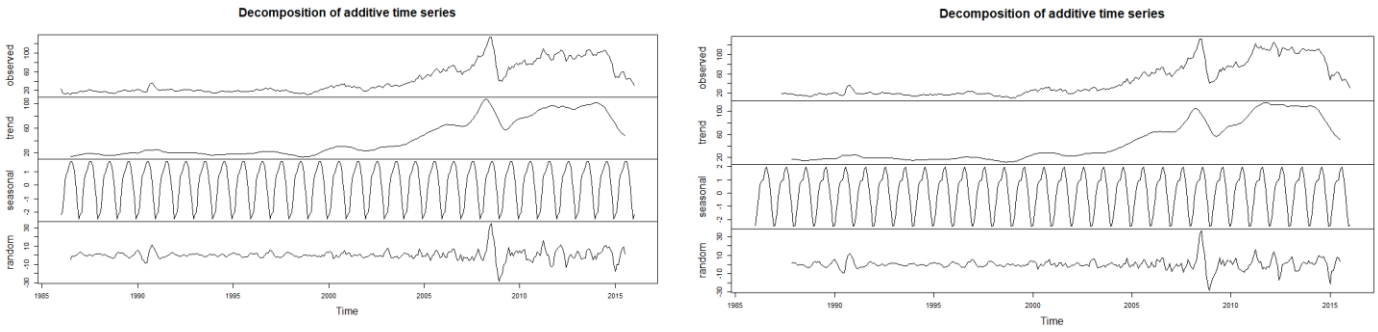


Fig 4 Decomposition of additive time series for Cushing and Europe Brent Prices (monthly dataset)

ACF and PACF

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are critical statistical tools used in time series analysis to measure the correlation patterns in a time series data. ACF measures the similarity between current and lagged values, while PACF accounts for the influence of intermediate lags to determine the direct impact of each lag on the current value. These functions were utilized in R to analyze our dataset's correlation and identify the appropriate ARIMA models. The ACF and PACF plots illustrate correlation coefficients against lags on the x-axis and display positive or negative correlations. The presence of autocorrelation or partial autocorrelation outside the confidence interval indicates significant coefficients, revealing insights into the underlying structure of the time series data for accurate forecasting. The first dataset we wanted to observe was from Cushing.

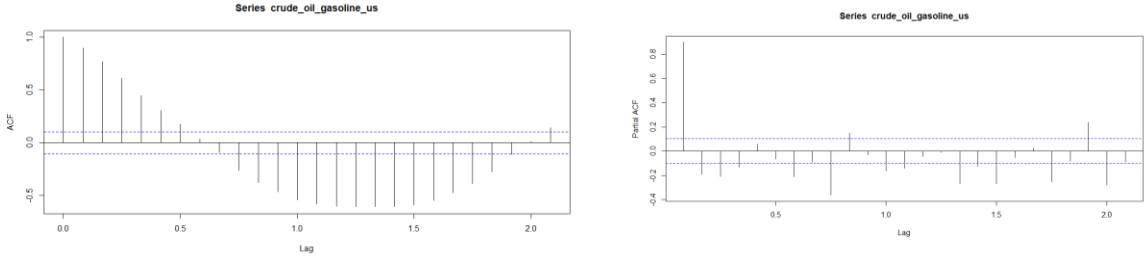


Fig 5 ACF and PACF plot for Cushing Prices

The ACF plot shows a significant negative correlation between the crude oil time series and its lagged values at lags 1 to 1.5, with the highest negative correlation coefficient of -0.6 occurring at lag 1.2. This suggests that the crude oil prices are negatively autocorrelated at these lags, meaning that high prices are likely to be followed by lower prices and vice versa. The PACF plot shows a significant spike at lag 0, which indicates that the current value of the crude oil prices is highly correlated with its immediate past value after controlling for all other lags. The large spike at lag 0 and the relatively small PACF values at other lags suggest that the crude oil prices exhibit a random walk behavior, which means that the current price is not significantly related to any of its past values beyond the immediate past value.

Overall, the ACF and PACF plots suggest that the crude oil prices exhibit a mix of negative autocorrelation and random walk behavior, with high correlation with its immediate past value and little correlation with other past values. However, the negative autocorrelation at lags 1 to 1.5 could be used in conjunction with other relevant information to develop more accurate price forecasts. Further investigating the ACF and PACF plots from the US and NY datasets we find that the results are the same.

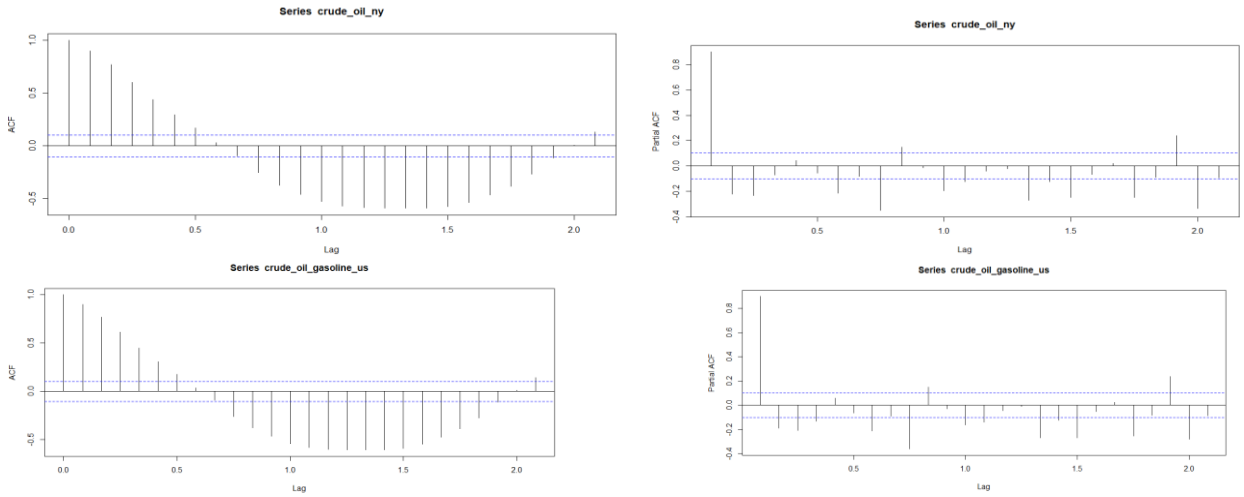


Fig 6 ACF and PACF plot for Gasoline Price - NY and Gulf Harbor

ARIMA Modeling

ARIMA (Autoregressive Integrated Moving Average) modeling is a time series analysis method that has been widely used in various fields for forecasting and understanding time series data. The fundamental basis of ARIMA models is the idea that the data under analysis is stationary, i.e., that its statistical characteristics do not alter over time. The models may be used to time series data to identify linear trends, seasonal fluctuations, and other patterns.

The accuracy of the ARIMA forecasts can be evaluated using statistical metrics such as the mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics provide a quantitative measure of the difference between the forecasted values and the actual values of the time series data.

The following steps are implemented in ARIMA Modeling

- Plot the data. Identify any unusual observations.
- If necessary, transform the data (using a Box-Cox transformation) to stabilize the variance.
- Use auto.arima() function to automatically select a model.
- Check the residuals from your chosen model by plotting the ACF of the residuals and doing a test of the residuals. If they do not look like white noise, try a modified model.
- Once the residuals look like white noise, calculate forecasts.

The results are visualized and is shown below:

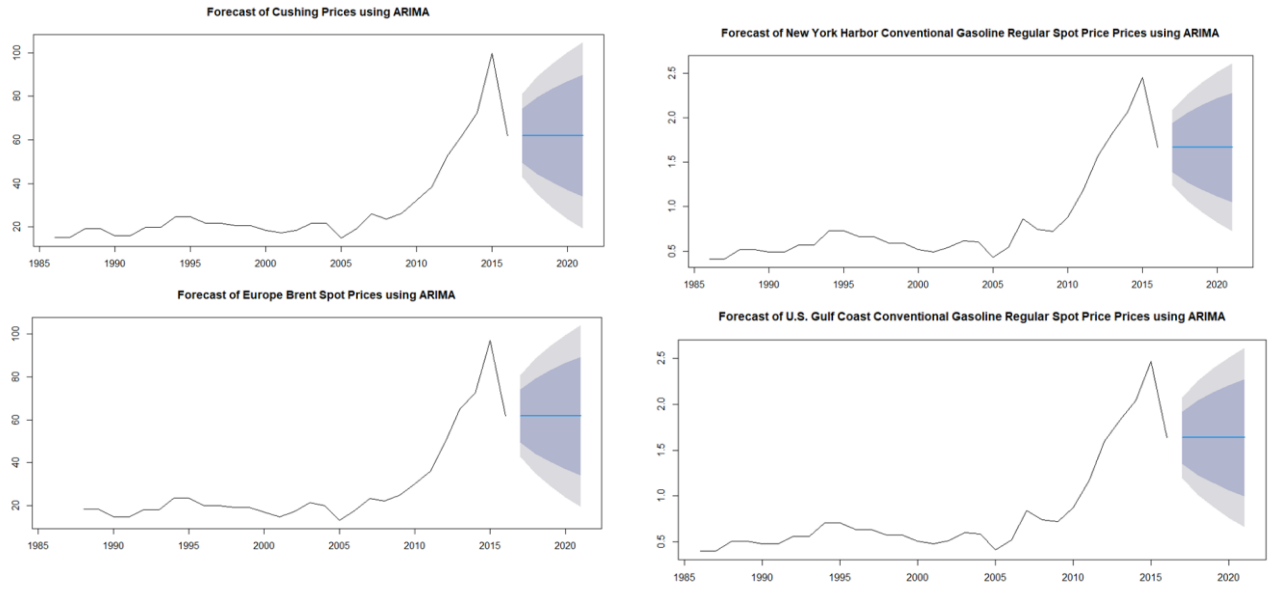


Fig 7 Forecasted Prices for Cushing, Europe, Gasoline Price NY and Gulf Harbor using ARIMA

Exponential Smoothing

Exponential smoothing is a time series forecasting method that is widely used in business and economics to make short-term forecasts. It is a flexible and adaptable method that can be applied to many different types of time series data. Holt Winter method is used as it incorporates the trend and seasonality components in addition to the level component. ETS function is used in R to figure out the best model and Holt's winter method is applied. The method of exponential smoothing used in the code is the "Additive Error, Additive

Trend, and Additive Seasonality" (AAN) model. The results showing the error metrics are tabulated below in the results section and the plots are visualized to get a better understanding.

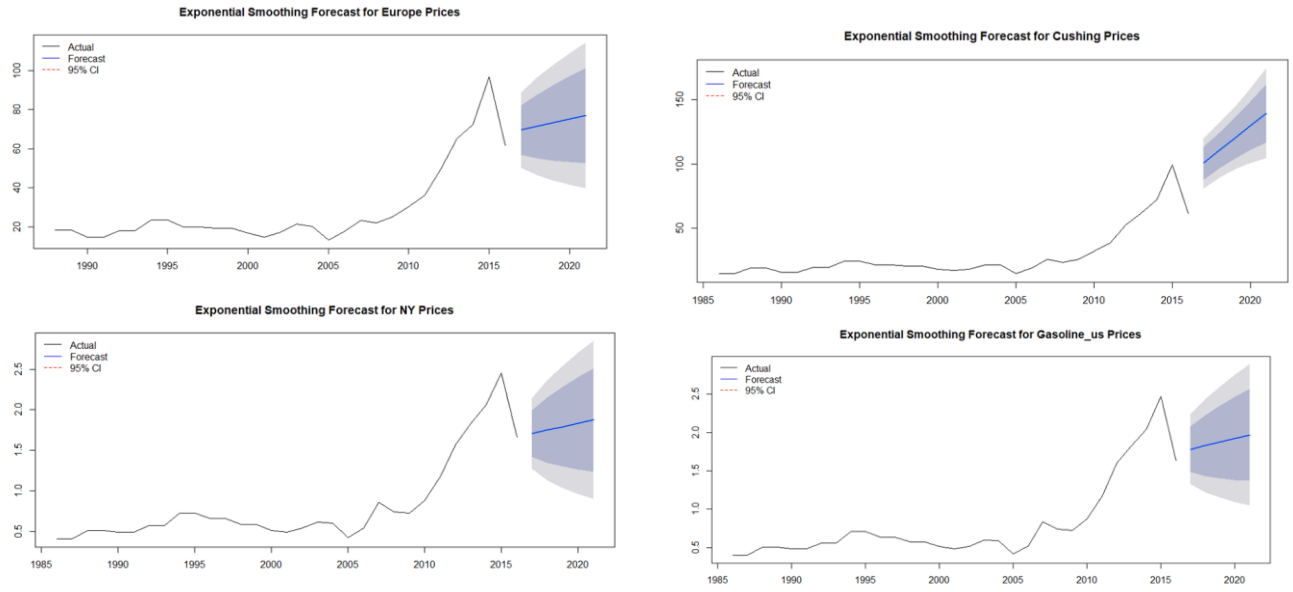
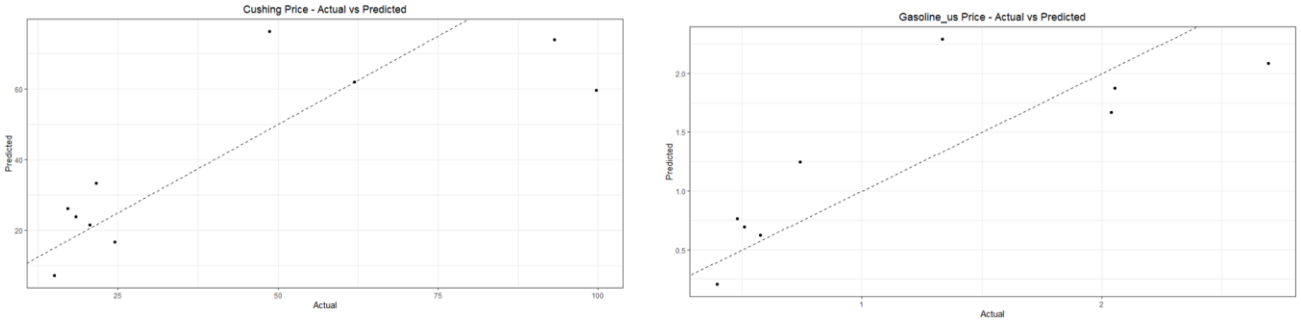


Fig 8 Exponential Smoothing models for Cushing, Europe, Gasoline price NY and Gulf Harbor

Linear Regression

Here, Linear Regression model is used to observe the relationship of dependent variable (prices) with respect to the independent variable (year). Linear regression models were created for various price models namely Cushing, Gasoline price - New York and Gasoline price - Gulf Harbor and the error metrics results are tabulated in the results section. It is found that the regression model is the best among various models generated with the least error metrics obtained for various price models.



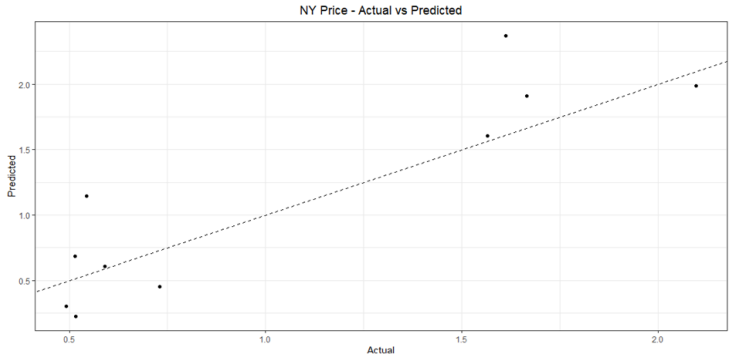


Fig 9 Linear Regression models for Cushing, Gasoline price NY and Gulf Harbor

Neural Network

Among data mining algorithms for predicting cross-sectional data, neural nets are popular for forecasting time series. The problem that we are trying to solve here is to predict the future prices of crude oil and gasoline and visualize the actual vs forecasted crude oil and gasoline prices. Crushing crude oil prices are taken for constructing the model. The model architecture is described as follows: There is one input layer, which is the time series year data from 1986 to 2016, which is split into training data (1986 - 2010, 80%) and validation data (2011-2016, 20%). There are 3 hidden layers used in this model and a linear function is used to visualize the model generated. Neural net function is used to create the model for the above-mentioned inputs and the model is passed as argument to find the forecasted prices. The final output is the forecasted price along with error metrics which are visualized against the actual prices. The results revealed that the error metrics were poor for the model generated and thus the prices for other crude oil and gasoline models and thus one model is alone generated, and results are displayed below.

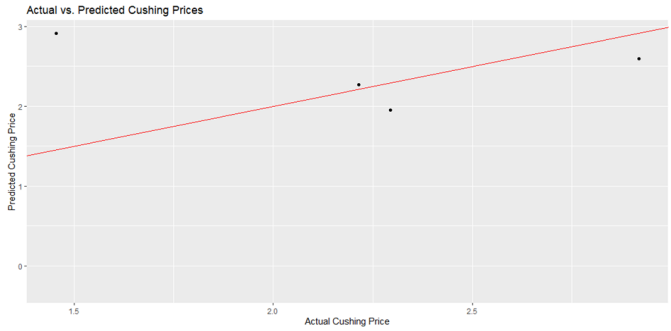


Fig 10 Neural Network models for Cushing, Europe, Gasoline price NY and Gulf Harbor

Results

Error Metrics - Price model / Fitted model	**Cushing**		**Europe Brent Price**		**Gasoline Price - Gulf**		**Gasoline Price - NY**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ARIMA Model	27.65	25.64	35.98	32.27	0.82	0.69	0.88	0.74
Exponential Smoothing	55.94	44.75	32.95	32.18	0.74	0.7	0.85	0.76
Regression	17.65	12.94	28.79	28.15	0.35	0.27	0.43	0.34

Augmented Dickey Fuller Tests were carried out to find the non-stationarity in the data and the results revealed that for Cushing is -1.15 with p value of 0.89, for gasoline price - New York the results were -2.57 with p value of 0.353 and for gasoline price - Gulf harbor the results were -2.66 with p value of 0.3178. The results conclude that the p- value is greater than the significant value and thus we cannot reject the null hypothesis, which states that the data is stationary. Since there were single input and output, we found that the regression model was more accurate (with respect to the accuracy and error metrics) and simple. Also, the predicted prices for different zones using ARIMA model for the prediction period 2017 to 2021 is given as follows:

For Cushing - next 5-year prices were (in dollars) 61.95, 61.96, 61.95, 61.99, 61.95.
For Europe Brent Spot price - next 5-year prices were (in dollars) 61.74,61.76, 61.74, 61.77, 61.74.
For Gasoline price - US Gulf harbor- next 5-year prices were (in dollars) 1.64,1.64, 1.66, 1.67, 1.67.
For Gasoline price - New York harbor - next 5-year prices were (in dollars) 1.67, 1.69, 1.67, 1.67, 1.67.

Possible Improvements for Methods and Data Collection

From the following analysis, we found that there were many missing data for Europe Brent Spot Price and Gasoline prices of Los Angeles for which the data should be referred with respect to daily price changes. Also, the linear regression model can be extended to second order, cubic and polynomial regression models to understand the data deeper and to provide insights on various models. Also, the neural network models can be done by using various functions apart from the linear function used. From the data collection, we understood that there were many anomalies in the dataset provided, which upon doing EDA revealed that there were missing data, which also had some limitations in doing some analysis.

Conclusion

In conclusion, this analysis provides insights into the trends, decade-wise changes, and distribution of crude oil and gasoline prices for different types. The visualizations created using R help us understand the historical movements and patterns in crude oil prices, while the ARIMA modeling allows us to forecast future prices. The Neural Network model provides an insight on how the actual and forecasted prices are related by developing a neural net model. These findings can be useful for decision-making in industries and policymaking at a macroeconomic level.

Lessons Learned

The analysis shows that macroeconomic factors such as GDP growth, inflation, and exchange rates can have a significant impact on crude oil and gasoline prices. For example, during periods of high economic growth, demand for energy increases, leading to an increase in price, whereas the sanctions on Iran from 2014 - 2016 caused an alternate spike and drop in gasoline prices. Also, pandemic played a crucial role, as the demand for oil spiked at rigorous levels, which showed that the forecasted values doesn't match the actual prices that were present during the period 2021-2022. Further analysis, including advanced time series modeling and incorporating external factors, can be conducted to refine the forecasts and gain deeper insights into crude oil and gasoline price dynamics and their implications. Furthermore, efforts can be made to improve the accuracy of the model. This may involve adjusting model parameters, exploring different algorithms, or incorporating additional data features.

References

https://www.eia.gov/dnav/pet/pet_pri_spt_s1_d.htm - Dataset chosen - Gasoline and Crude oil prices.