Inferential Analysis of United States Gas Prices

Thomas Hollingshead

Golden Gate University

Data 190: Capstone

April, 2024

Table of Contents

Abstract	3
Introduction	4
Literature Review	5
Test design/approach	6
Results of data analysis/Conclusion	15
References	17

Abstract

The domestic price of gasoline in the United States is very important for several reasons. This research paper will have a particular focus on the domestic price of gasoline in the United States for the consumer, as we have recently had the highest gasoline prices ever recorded in June 2022, in the United States. In this research paper, we will analyze the domestic price of gasoline within the United States, and will analyze rises and a falls in the price while attempting to determine whether time itself is a significant independent variable that can predict the rise and fall of the price in gasoline with the intent to grasp a better understanding of the fluctuations in price the domestic price of gasoline has had since the turn of the century.

Introduction

Gasoline is a natural resource that is important to the world economy for several reasons. Gasoline is a fuel that exists for a variety of use-cases, as it is what is used when operating combustion engines. Because of the necessity of gasoline word wide, gasoline has a significant impact on the world economy, as it is being bought and sold by every country on the planet. Due to the high supply-and-demand created by gasoline, gasoline has the ability to effect the GDP of a country where countries with higher gas prices are expected to have a lower GDP per capita, as opposed to a higher GDP per capita when gas prices are lower. (Purcell, Sekar 2015)

Gasoline is without a shadow of a doubt, one of the most important global energy resources, and due to influences such as politics, economic strategies, and the balance of supply and demand globally, gasoline prices of a tendency to fluctuate. Despite the importance of the resource, and the importance of being able to predict the price of gasoline, understanding the volatility of gasoline remaisn to be a challenging task. Government all over the world regulate the gasoline market, meanign there are recurring patterns in the fluctuation in its price in response to international conditions. Regardless, and as we will delve into in this research paper, we have determined that global occurences have a significant impact on the price of gasoline.

Literature Review

In the fall of 2008, during the financial crisis, Americans began to purchase more premium gasoline, in contrast to what was expected – that Americans would cut back on the expenditure of non-essentials. (Sumo, 2013) Research by Professor Jesse M. Shapiro at The University of Chicago sheds light on this, challenging the assumption in the economic model.

While it was expected that as income levels declined during the financial crisis of 2008, which would theoretically result in consumers reducing spending on gasoline, the data says otherwise. In contrast to the financial constraints of the average American Household, people began to purchase more premium and mid-grade gasoline during this time. Researchers suggest this unexpected behavior created volatility in the market, resulting in the raising of the domestic price of gasoline.

The onset of the unprecendeted Covid-19 Global Pandemic of 2020 resulted in a significant drop in the domestic price of gasoline due to the severe lack of demand in the United States. As demand plummented, so did surplus supply – and there was a limited storage capacity. When lockdown regulations were lifted, there was a huge influx of people trying to purchase gasoline, resulting in price hikes for the fuel. The Monthly Labor Review noted that the influx in gasoline purchases caused by the pandemic resulted in extreme fluctuations in volatility in the domestic price of gasoline.

Demand factors hold a heavier weight than supply when it comes to the price of gasoline. In the United States, Gasoline consumption is seasonal, rising substantially

during the winter due to the necessity of heating. Its typical gas prices peak in the winter months in the United States. Its vital to forecast gasoline prices accurately due to thigns like enviornmentla conservation, energy investment, policy writing, and economic planning.

Test Design/Approach

Inferential Analysis, Predictive Analytics, Time Series Analysis

Random sample of the data retrieved from the U.S Energy Infomration Administration (EIA):

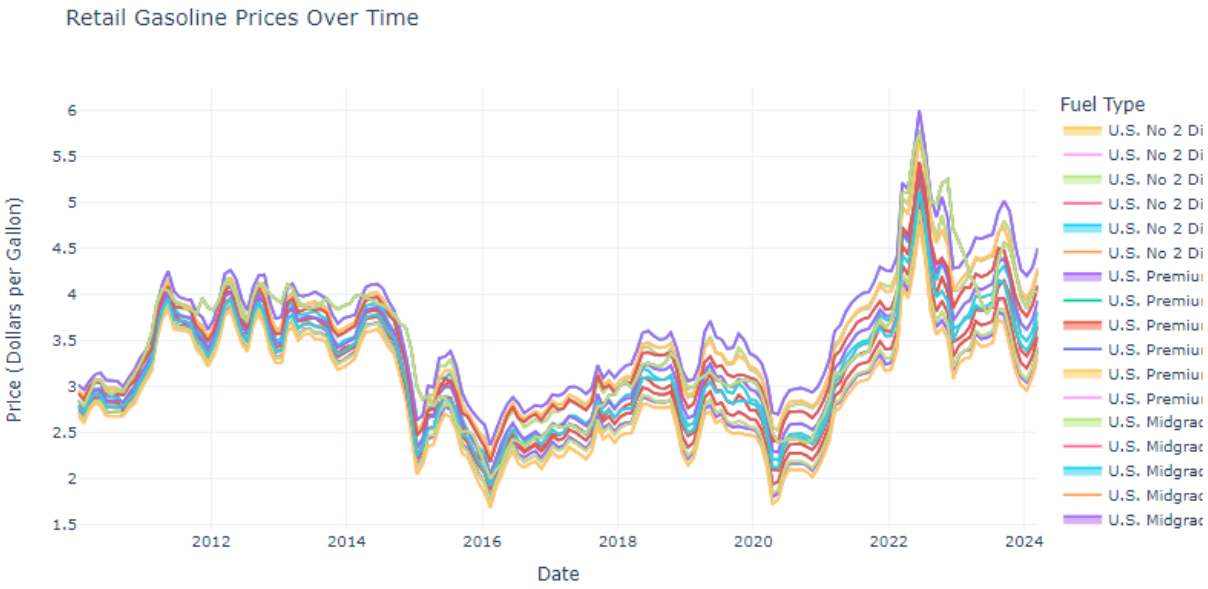
U.S. All Grades All Formulations Retail Gasoline Prices (Dollars per Gallon)	U.S. All Grades Conventional Retail Gasoline Prices (Dollars per Gallon)	U.S. All Grades Reformulated Retail Gasoline Prices (Dollars per Gallon)	U.S. Regular All Formulations Retail Gasoline Prices (Dollars per Gallon)	U.S. Regular Conventional Retail Gasoline Prices (Dollars per Gallon)	U.S. Regular Reformulated Retail Gasoline Prices (Dollars per Gallon)	U.S. Midgrade All Formulations Retail Gasoline Prices (Dollars per Gallon)	U.S. Midgrade Conventional Retail Gasoline Prices (Dollars per Gallon)	U.S. Midgrade Reformulated Retail Gasoline Prices (Dollars per Gallon)	U.S. Premium All Formulations Retail Gasoline Prices (Dollars per Gallon)	U.S. Premium Conventional Retail Gasoline Prices (Dollars per Gallon)	U.S. Premium Reformulated Retail Gasoline Prices (Dollars per Gallon)	U.S. No 2 Diesel Retail Prices (Dollars per Gallon)
1.177	1.159	1.261	1.131	1.119	1.207	1.216	1.209	1.337	1.316	1.303	1.402	1.166
1.132	1.115	1.213	1.086	1.075	1.161	1.172	1.166	1.288	1.273	1.261	1.355	1.12
1.096	1.082	1.157	1.049	1.042	1.105	1.138	1.133	1.229	1.237	1.228	1.299	1.084
1.064	1.055	1.098	1.017	1.014	1.046	1.109	1.107	1.169	1.203	1.199	1.242	1.063
1.077	1.064	1.114	1.03	1.024	1.067	1.12	1.116	1.182	1.216	1.209	1.253	1.067
1.105	1.088	1.165	1.064	1.049	1.114	1.152	1.138	1.232	1.244	1.23	1.298	1.069
1.103	1.086	1.165	1.064	1.048	1.11	1.155	1.133	1.228	1.241	1.224	1.302	1.041
1.094	1.078	1.153	1.055	1.039	1.098	1.146	1.124	1.217	1.234	1.216	1.293	1.029
1.065	1.049	1.128	1.026	1.011	1.071	1.119	1.097	1.194	1.207	1.188	1.273	1.007
1.049	1.033	1.113	1.009	0.994	1.054	1.103	1.081	1.179	1.193	1.174	1.259	1.024
1.059	1.045	1.115	1.019	1.006	1.057	1.113	1.094	1.181	1.204	1.188	1.26	1.039
1.036	1.02	1.105	0.995	0.979	1.045	1.093	1.071	1.173	1.183	1.164	1.252	1.022
0.987	0.964	1.081	0.945	0.923	1.016	1.046	1.016	1.152	1.137	1.11	1.232	0.973
0.98	0.957	1.07	0.939	0.917	1.008	1.038	1.009	1.141	1.128	1.101	1.22	0.967
0.962	0.94	1.047	0.921	0.9	0.984	1.02	0.993	1.119	1.112	1.086	1.199	0.959
1.022	1	1.107	0.982	0.961	1.043	1.081	1.051	1.181	1.166	1.141	1.247	0.997
1.171	1.137	1.305	1.131	1.099	1.228	1.23	1.186	1.392	1.31	1.272	1.434	1.079
1.171	1.143	1.282	1.131	1.103	1.212	1.229	1.194	1.359	1.314	1.282	1.419	1.073

Data Visualization Retail Prices of Gasoline (All Types)

Below is figure 1, the Retail Prices of all gasoline types over time. As well as giving us a very specific outlook on the fluctuation of gas price, the graph hints at correlation between gas prices through it's visual representation. By analyzing this data, we are able to to determine that gas prices between different types tend to fluctuate with

each other, meaning the variables have an effect on all types of gasoline, with little to few exceptions (as outliers).

The main focus of this visual representation of retail gasoline prices over time is that we have a clear representation of gasoline prices spiking during the June of 2022, having remained stable prior to this time frame. Another important thing to note is that it *seems* as if gasoline prices hold a significant correlation between one another. Meaning, the forces that heavily influence the price of gasoline whether it is a diesel or regular product, seemingly have an influence on all *different types* of gasoline equally. This is



further explained in the following confusion matrix designed to scientifically define the relationship between different types of gasoline.

Ordinary Least Square Analysis

Regression Equation: Price =-156.1602 +0.0002 * Date_ordinal

OLS Regression Results

Dep. Variable:	Q('U.S. Regular All Formulations Retail Gasoline Prices (Dollars per Gallon)')	R-squared:	0.646
Model:	OLS	Adj. R-squared:	0.645
Method:	Least Squares	F-statistic:	727.3
Date:	Tue, 16 Apr 2024	Prob (F-statistic):	6.05e-92
Time:	20:22:41	Log-Likelihood:	-339.37
No. Observations:	401	AIC:	682.7
Df Residuals:	399	BIC:	690.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-156.1602	5.872	-26.593	0.000	-167.704	-144.616
Date_ordinal	0.0002	8.01e-06	26.969	0.000	0.000	0.000

Omnibus:	24.065	Durbin-Watson:	0.071
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.805
Skew:	0.625	Prob(JB):	1.51e-06
Kurtosis:	3.204	Cond. No.	1.52e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.52e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Regression Equation: Price = -156.1602 + 0.0002 * Date_ordinal

Intercept and slope – b0 (intercept) is -156.1682 and the slope (b1) for the ordinal date is 0.0002.

According to the model, if the ordinal date is increased by 1 unit (one day), the price is expected to increase by $0.0002 on average (assuming a perfectly linear relationship)

Coefficient Significance – slope is < 0.05, indicating that the date is a significant predictor of gasoline price

R-Squared is 0.645 which means that there is about 64.5% variability in gasoline price that can be explained by time

The Durbin Watson is 0.871 which is close enough to one to indicate that there might be positive autocorrelation in the residuals of the model.

Positive autocorrelation means high values to be followed by high values, and low values by low value

The Condition Number was large, suggesting there might be numerical problem or multicollinearity problems. This is common when independent variables (date) have a linear relationship

The warning about multicollinearity and the low Durbin-Watson value suggest that while time is a significant predictor, the model may not be best fit for the data.

Determine the correlation of Variables

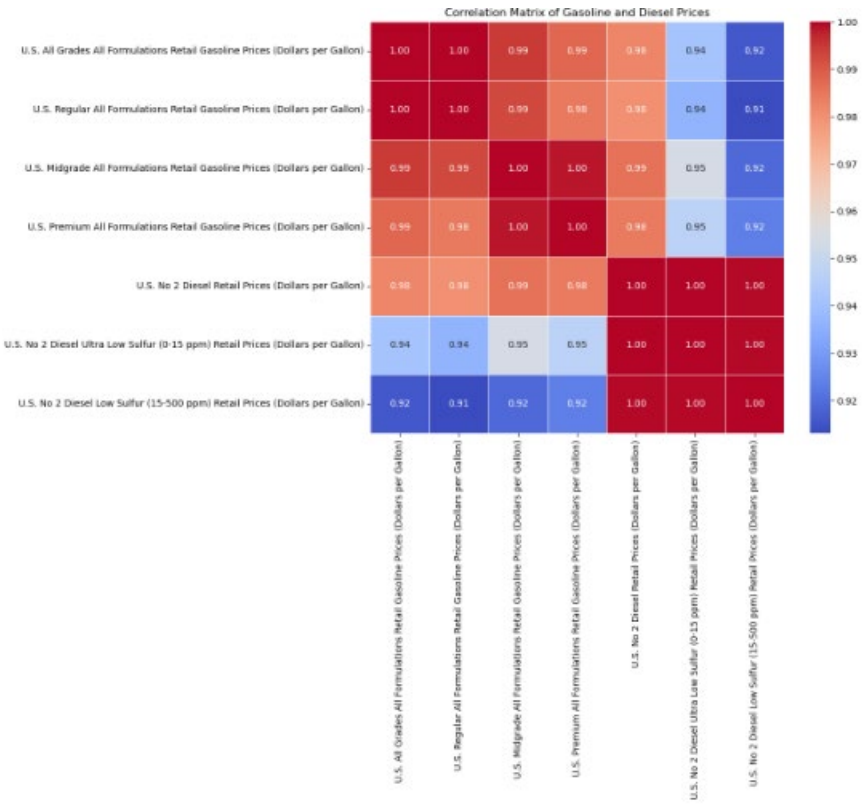
In this case, independent variables are the date and the different gasoline prices. We use these variables to explain the gasoline prices' changes.

This matrix shows the correlation coefficients between the different types of gasoline. The correlation coefficient ranges from -1 to 1, where:

1 - Positive Correlation

-1 – Negative orrelation

0 – No Correlation

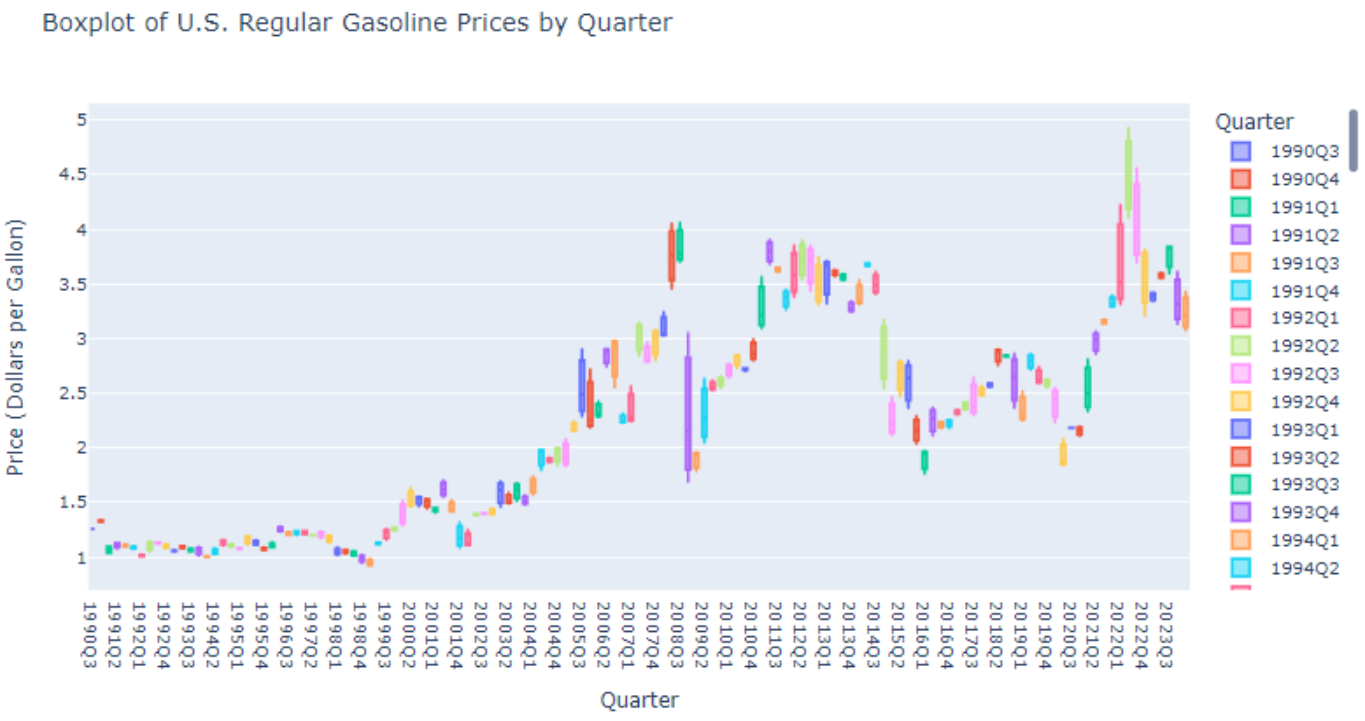


By using the correlation Matrix, we have determined that all of *different types* of gasoline have a significant correlation with one another – meaning no specific type of gasoline's price will change any more significantly than another.

Boxplot

The boxplots in Figure 2 group the Regular All Formulations Retail Gasoline prices quarterly over the years provided by the data. Each boxplot represents a distribution of prices quartlery within the specific year. By analyzing the boxplot we can identify seasonal trends, year-to-year variablity, quarterly fluctuations and outliers.

Within this boxplot we are able to note the significant drop in gasoline price in 2008, which could be speculated to be caused by the 2008 financial crisis when Americans unexpectedly began consuming more gasoline despite it being a non-essential commodity during this period of time. There is also a clear representation of the price



skyrocketing in 2022 following the unprecedented Global Pandemic of 2020, when the demand for gasoline significantly increased.

Calculate the regression equation. Train the model using past data.

The regression line has an equation of the form:

\text{Price} = (b0) + (b1) \times \text{Date_ordinal}

To perform regression, X, or the independent variable, was set as the oridnal date, and y, the dependent variable was set as the gasolince prices. With the linear regression model we are attempting to find a linear relationship between the price of gasoline and the date. The red line is the predicted gasoline prices, and the blue dots are the actual prices. The slope that we produced through this equation indicates that over time, gasoline prices have sustainably been raising.

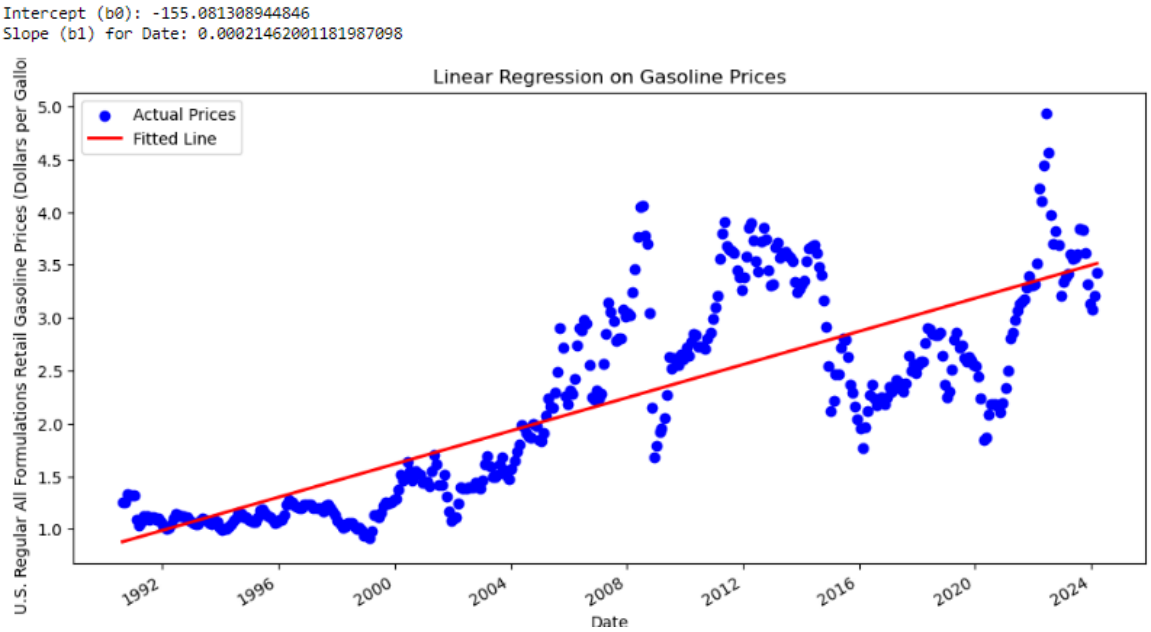
The intercept is the expected value of gasoline prices when the date is set to zero, (January 1, Year 1). This intercept would represent the gasoline priceo n that day. This isn't a meaningful interpretation based on the context of the data, but its how we adjust the regression line to fit the data's future plot points.

The slope tells us how much we can expect the gasoline price to change for every unit increase in the (ordinal) date. In this case, the slope describes there being an approximate $0.00021462 change in gasoline price daily.

We can determine the rate of change as about 21.46 cents per 3 years. Although this model assumes a linear relationship, it does not account for factors such as economic events, changes in crude oil price, seasonal variation, policy changes, wars, etc.

$$\text{Price} = \text{Intercept} + \text{Slope} \times \text{Date\_ordinal}$$

Because of how much the “Actual Prices” fluctuate past the “Fitted Line” we got from the regression equation, we can determine that time is not the only independent variable responsible for the fluctuation in the price of gasoline – as speculated by the significant economic events that occurred during the financial crisis of 2008 and the global pandemic of 20202, effecting the supply and demand of the resource during these times.

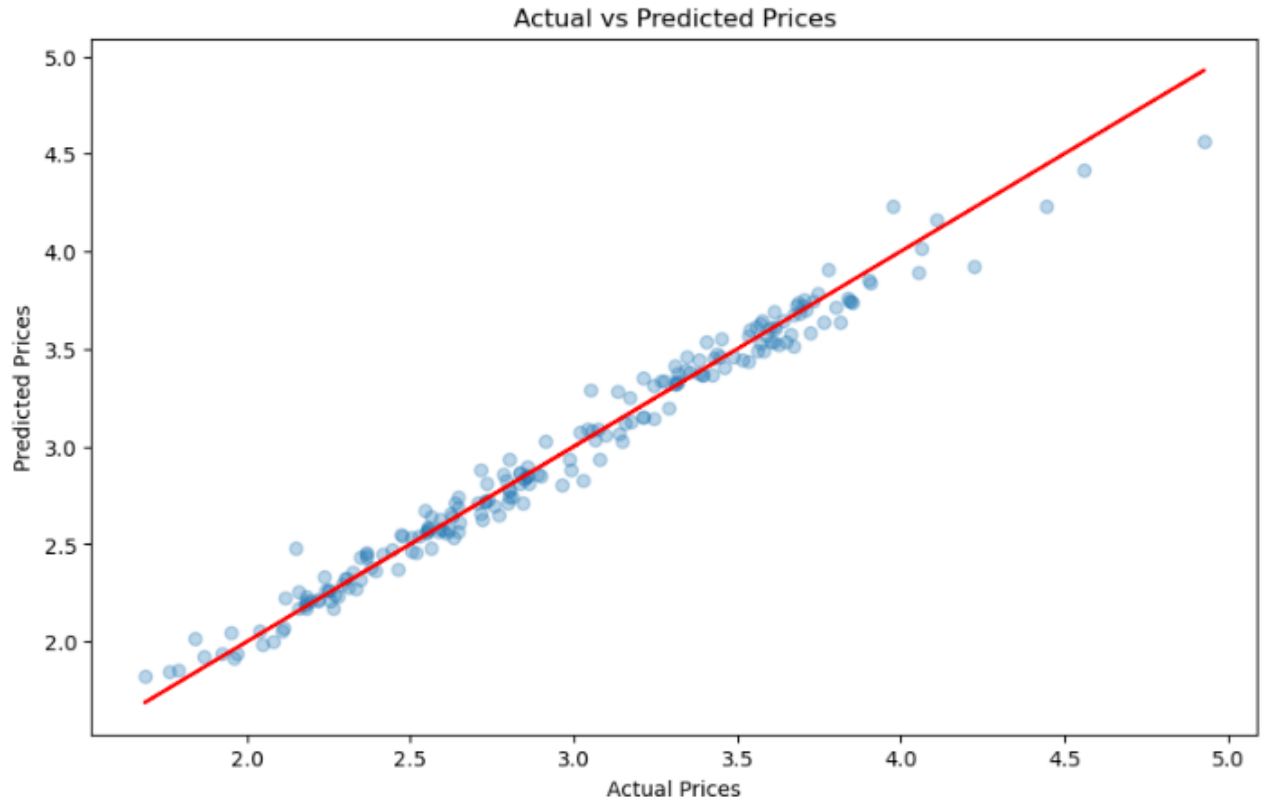


Predict the price of gasoline in future months

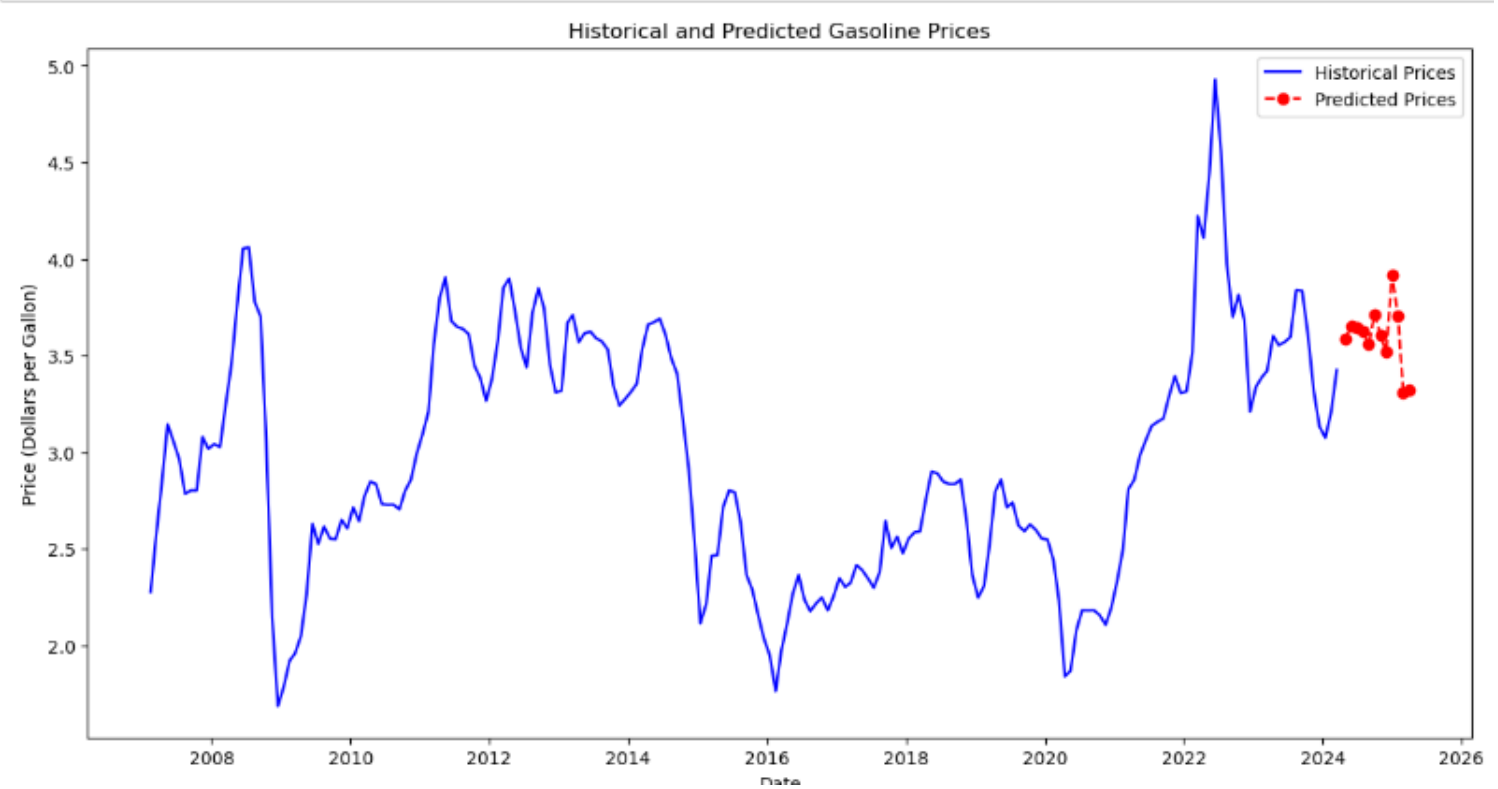
By performing a machine learning analysis on historical gasoline price using a Random Forest Model, we can train the model using it to predict gasoline prices in the future based on the previous day’s prices. Points closer to the line infer better predictions. The Mean Absolute Error between the actual and predicted prices represents a $.06

difference between the two based on the historical data, indicating the model performs at a good rate.

```
Mean Absolute Error on the entire dataset: 0.06432608148404984
```



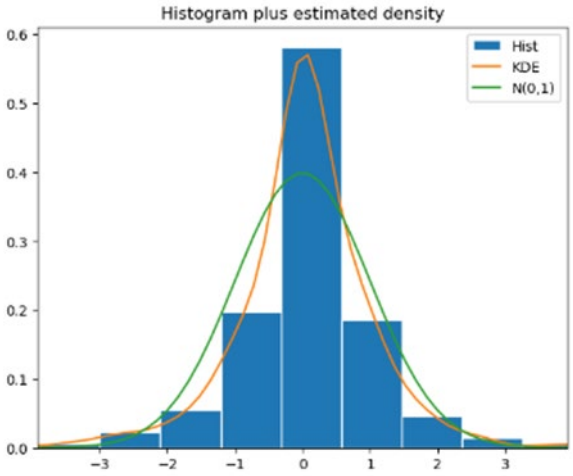
We can create a chart that visualizes both the historical and predicted prices of gasoline by using the aformentioned trained random forest model. By identifying the last known price of gasoline from the dataset, we can use that price as a reference point for predicting the price in the future. We use the most recent predicted price as an input for the next prediction we make.



Residual analysis

By doing residual analysis, we can evaluate the performance of of the predictive movdel. Residuals differentiate between actual data and the predictiosn of the model, which give us insight on potential biases that exist in the model. Because the residuals are disperesed around the zero-line reandomly, this analysis implies the model's errors are random, which means it is performing without bias.

By using the Sarima Model, which is an autoregressive integrated moving average, we have a model that uses the time series data to better understand the data set. The residuals associated with this model are the irregular components, which exist after "trend" and "seasonal" data are cleaned from the original dataset. By scattering the residuals around zero, we are looking to see if there is no apparent pattern.



The histogram residuals followed the bell-shape pattern of normal distribution. This means that the data has performed and is being displayed well. This is because the goal of the histogram is for the KDE to closeley resemble the N(0,1) which suggests normal distribution.

Results of Data Analysis/Testing, Conclusions

By the Correlation Matrix and initial chart, we can determine that all gasoline types hold a significant correlation – meaning their prices depend on another. With regards to the Time Series Analysis, we determined that due to the low Durbin Watson score, multicollinearity and visual representation of the Linear Regression line, we can determine that although time is a significant predictor of gasoline for seasonal and trending cases, the OLS model can't account for absolutely everything. This suggests that outside events have a significant impact on the price of gasoline, as they dramatically effect th esupply and demand of the market. The most signifiacnt volatility we observed in the datasets were during 2008, which was a market low since the turn of the century ,and 2022, which was an all time high. It would be reasonable to speculate that the 2008 financial crisis and the unprecedented global pandemic had a significant impact on the supply and demand of the domestic retail price of gasoline in the united states during these times respectively.

References

Data:
https://www.eia.gov/dnav/pet/pet_pri_gnd_dcus_nus_m.htm

Sources:

O'Donnell, G., & Ferré, I. (2024, March 23). Gas prices: Why the national average could keep going up. Yahoo Finance. https://finance.yahoo.com/news/gas-prices-why-the-national-average-could-keep-going-up-194220210.html

Purcell, L., & Sekar, A. (2015, April 17). Effects of Gasoline Price Levels on GDP Per Capita: A Cross-Country Analysis. [Abstract]. Retrieved from https://repository.gatech.edu/server/api/core/bitstreams/797b4a25-0102-4428-9945-11de7184ee13/content#:~:text=Recent%20and%20past%20studies%20have,have%20lower%20GDP%20per%20capita.

Sumo, V. (2013, September 19). Why Did Drivers Switch to Premium during the Recession? Chicago Booth Review. https://www.chicagobooth.edu/review/why-drivers-switch-premium-during-recession#:~:text=The%20evidence%20points%20to%20the,to%20spend%20more%20on%20gasoline.

U.S. Energy Information Administration. (2023, April 19). Gasoline explained: Why do gasoline prices fluctuate? [Web page]. https://www.eia.gov/energyexplained/gasoline/prices-and-outlook.php

Folger, J. (2021, September 01). How gas prices affect the economy. Investopedia. https://www.investopedia.com/articles/markets/090115/how-gas-prices-affect-economy.asp