Petroleum (Vehicle Fuel) Sales

2022-11-17

Introduction

The impacts of petroleum consumption for vehicle fuel has long been a debated topic in regards to how to handle the finite resource and it's potential environment impacts. Understanding how fast we are selling and delivering this resource can allow us to mitigate risks of resource consumption in the future while continuing to access affordable transportation fuel methods. The research to be performed is an analysis of potential predictors to determine how petroleum sales has changed over time to provide the knowledge on how to support the industry as well as determine the need to find and implement alternative resources. For the purpose of this report, motor gasoline is the only petroleum product being considered.

Research Questions

The research questions that are considered for this report's analysis are:

- How has petroleum sales changed in the last 10 years?
- What grade of vehicle fuel has the greatest sales?
- What impact does the units of vehicles sold have on the amount of petroleum sales/deliveries for transportation?
- Are there any seasonal changes in petroleum sales?
- What amount of annual and monthly growth is expected in the future for petroleum sales/deliveries based on the above considerations?

Approach

To address the problem of predicting the growth of petroleum sales for vehicle fuel, I will perform an analysis on the current delivery amount as well as calculate the relationship to the national average gas price and vehicle units sold. To provide an accurate prediction based on the current market conditions, only the last 10 years (2012-2021) will be taken into consideration.

This approach will partially solve the problem by creating a model to predict the annual and monthly petroleum sales for vehicle fuel based on potential predictors that best represents the variance of petroleum sales/deliveries. Once the sales/deliveries can be estimated, additional research can be performed to determine a timeline on when need will outgrow the supply chain or when the resource may be limited. From there policies and research into alternative energy methods can be put in place to prevent future impacts to transportation.

Data

Based on the approach stated above, the following datasets were captured for analysis:

U.S. Energy Information Administration: Petroleum & Other Liquids Sales Volumes

Vehicle fuel sales/consumption of petroleum has been recorded monthly since 1983 by the U.S. Energy Information Administration. This data has been collected and analyzed in order to promote understanding of resource consumption and provide a public understanding of energy. The dataset contains 14 variables including date and various usage categories for petroleum measured in thousand barrels per day. (U.S. Energy Information Administration 2022a)

U.S. Energy Information Administration: U.S. Gasoline and Diesel Retail Prices

The U.S. Energy Information Administration collects weekly data regarding retail gasoline and diesel prices by grade of the product. The original purpose of the dataset is to analyze energy information to encourage sound policy making and an efficient market. Data has been collected since August 1990. The dataset contains 16 variables including the date and the various product grades. The monthly averages are an average of the weekly data collected. If there was an incomplete weekly data series, the respective monthly and/or annual averages are not available. (U.S. Energy Information Administration 2022b)

U.S. Bureau of Economic Analysis: Supplemental Estimates, Motor Vehicles

The U.S. Bureau of Economic Analysis collects data from over 360 surveys and other collections are sponsored by other federal agencies. The original purpose of the data collection is to measure U.S. gross domestic product (GDP). A subset of the larger dataset includes the supplemental estimates of motor vehicle sales by month. This information has been collected at a monthly basis since 1976. The dataset for Vehicle Sales includes 2 variables and they are date and total sales (millions of units). Revisions are made to the dataset during quarterly reviews of the estimation methodology. The dataset being used for the analysis was exported by the Federal Reserve Bank of St. Louis noting that the source was a subset of the data originating from the U.S. Bureau of Economic Analysis. (Federal Reserve Bank of St. Louis 2022)

Required Packages

The packages required to complete this analysis are as follows:

- dplyr: tools to enable dataframe manipulation
- ggplot2: provide graphics for data analysis
- gridExtra: arrange multiple grid-based plots
- lubridate: work with date-times and time-spans
- magittr: offers set of operators to make code more readable
- readr: read data from csv and tsv files
- readxl: read excel files
- stringr: implementation of string manipulation

Plots and Table Needs

To display the results of the analysis, the following plots and tables will be created to answer the research questions above.

Key fields are described as petroleum sales, national average gas prices, and car units sold.

Plots

- Histogram of key fields
- Normality plot (q-q plot) of key fields
- Scatterplot to show relationship between petroleum sales and national average gas prices as well as petroleum sales and cars sold
- Scatterplot for month and yearly trends for petroleum sales
- Line chart to show the grade of vehicle fuel with the greatest sales

Tables

- Summary Statistics table for key fields
- Covariance of all key fields
- Correlation Coefficients and Coefficient of Determination

Questions for Future Steps

To perform the analysis above, new knowledge will need to be gained regarding reshaping data in excel files and manipulating dates by month and year. The datasets provided by the U.S. Energy Information

Administration are not in the expected format where each variable is in its own column and each observation is in its own row. Utilizing dates in this dataset by month and year will allow grouping of data to provide predictions of how it changes monthly and annually. By using the lubridate package, I should be able to split this in order to summarize the data appropriately and in a common format.

Data Preparation

The following steps were followed in order to prepare the datasets to be combined for the purpose of this report's analysis:

- The datasets were imported into RStudio as dataframes
  - Due to the export format, the excel documents from U.S. Energy Information Administration required additional parameters to select the appropriate tab and skip extra header rows.
  - The dataset provided by U.S. Bureau of Economic Analysis (summarized by Federal Reserve Bank of St. Louis) was downloaded as a csv.
- The column names of the datasets were renamed to be shorter and specific to the dataset it related to.
- Utilizing the date variable, two new columns were created for year and month for each dataset.
- Dyplr was used to filter and select the appropriate columns and rows of the dataset to obtain data in the last 10 years (2012-2021)
- Scatterplots and Histograms were used to validate that there were no significant outliers and that the data was approximately normal. No significnat issues were found.
- All 3 datasets were combined into one dataframe joined on year and month.

The final set contains 120 records with 12 variables listed below:

- date
- year
- month
- total gas sales
- regular gas sales
- midgrade gas sales
- premium gas sales
- all gas price
- regular gas price
- midgrade gas price
- premium gas price
- car sale volume

A summarized version of the final dataset is displayed below with data grouped by year:

year	Gas Delivered (Thousands)	Average Gas Price	Total Cars Sold (Millions)
2012	4,166,723	3.69	177.37
2013	4,183,412	3.58	190.58
2014	4,159,550	3.44	202.30
2015	4,332,136	2.51	214.28
2016	4,460,596	2.25	214.55
2017	4,500,437	2.53	210.77
2018	4,494,336	2.82	212.54
2019	4,406,278	2.69	209.86
2020	3,870,048	2.26	178.55
2021	4,204,281	3.09	184.91

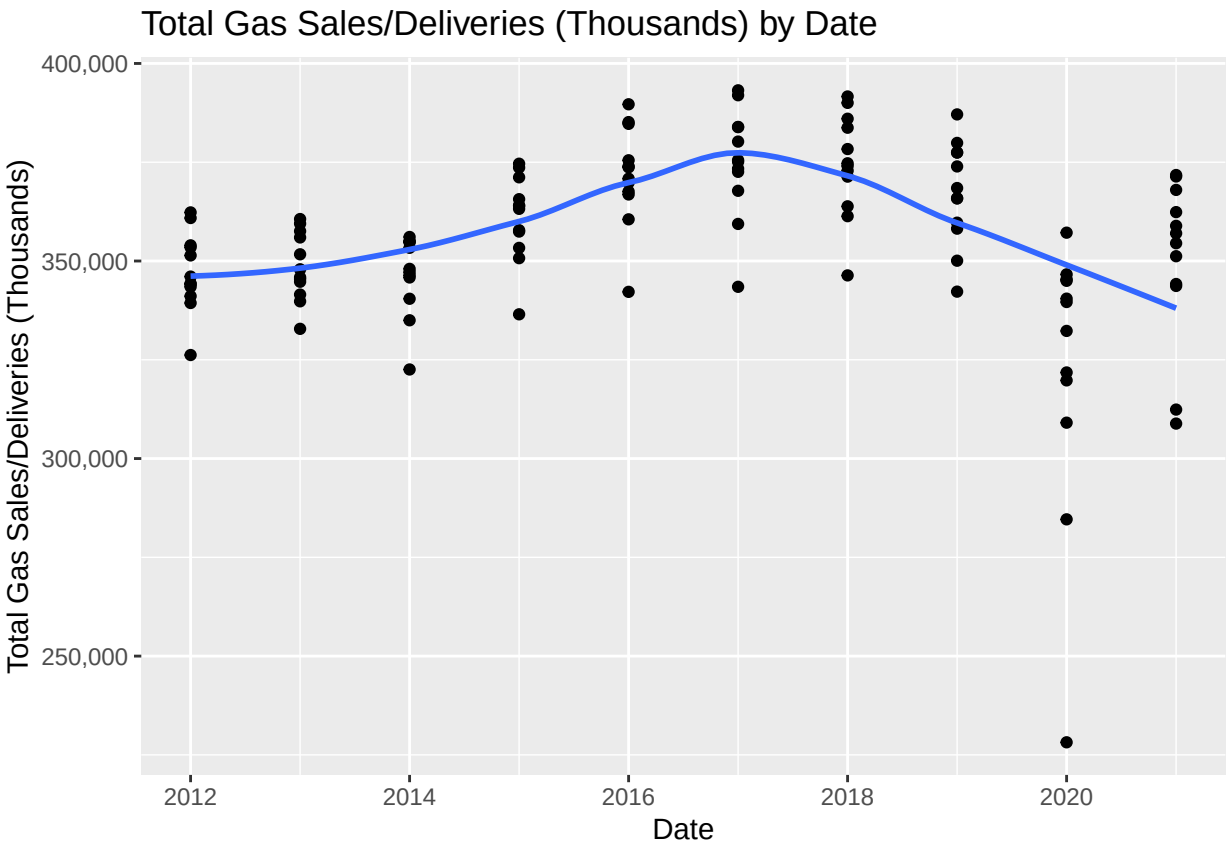
At this time, it appears that the dataset import and cleanup was done to satisfaction for the purpose of this report. In order to uncover new information, I will be looking at the correlation between the variables to determine what effect gas prices and volume of cars sold has on gas being delivered. If significant correlation

is found, the subgroups for gas including regular, midgrade, and premium will be used to determine the impact each grade has.

To answer the questions state above, I will be using the charts described in Plot and Table Needs to determine the relationship between the variables. Beginning analysis has been demonstrated in the charts below.

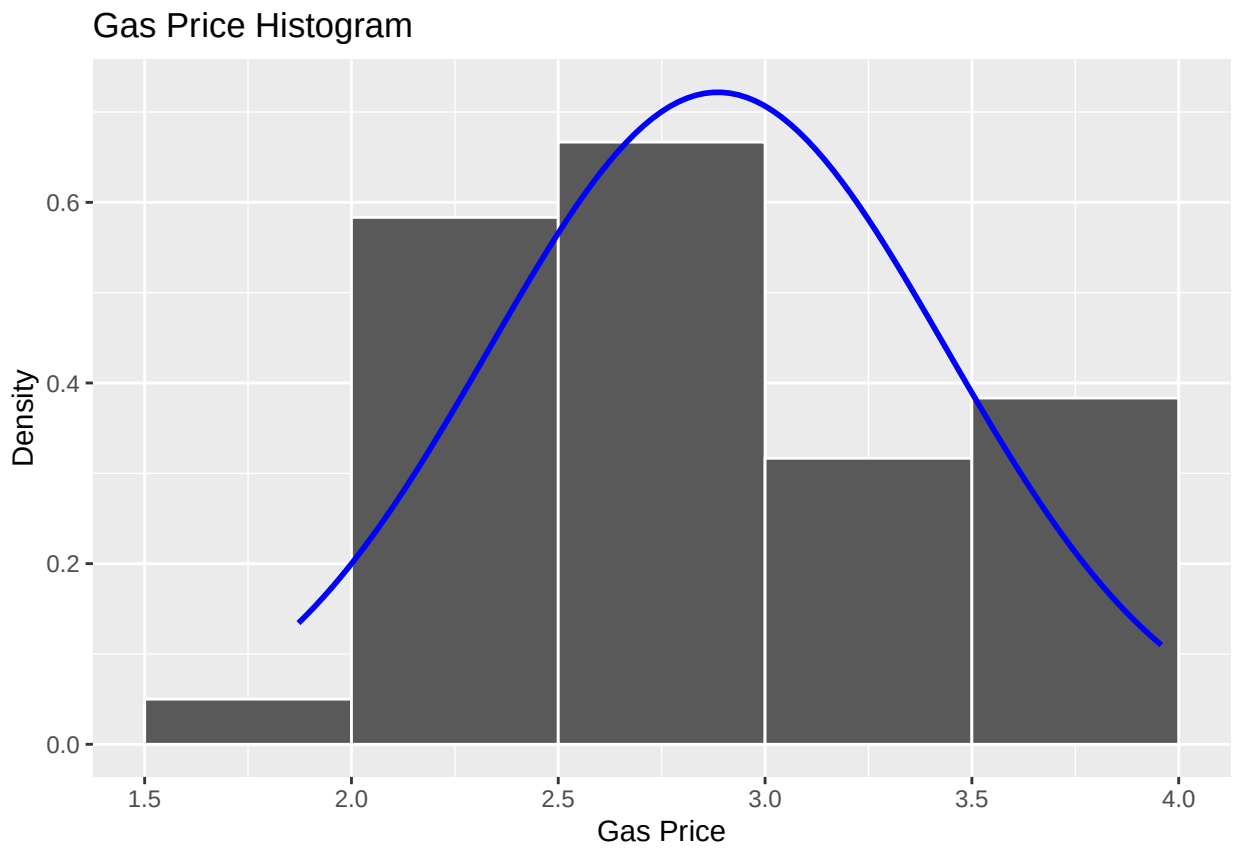
Summary information for the key areas are listed as follows:

```
##  totalGasSales     regGasSales      midGasSales     premGasSales
##  Min.   :228185   Min.   :197611   Min.   : 3357   Min.   :27217
##  1st Qu.:345296   1st Qu.:300469   1st Qu.: 5579   1st Qu.:36129
##  Median :357680   Median :310807   Median : 6771   Median :41149
##  Mean   :356482   Mean   :309441   Mean   : 7348   Mean   :39692
##  3rd Qu.:371968   3rd Qu.:323256   3rd Qu.: 7326   3rd Qu.:42929
##  Max.   :393200   Max.   :341532   Max.   :14596   Max.   :48293
##   allGasPrice    regGasSales.1     midGasPrice     premGasPrice
##  Min.   :1.872   Min.   :197611   Min.   :2.021   Min.   :2.242
##  1st Qu.:2.419   1st Qu.:300469   1st Qu.:2.602   1st Qu.:2.844
##  Median :2.748   Median :310807   Median :3.003   Median :3.256
##  Mean   :2.886   Mean   :309441   Mean   :3.075   Mean   :3.286
##  3rd Qu.:3.396   3rd Qu.:323256   3rd Qu.:3.571   3rd Qu.:3.747
##  Max.   :3.958   Max.   :341532   Max.   :4.028   Max.   :4.162
##     carSales
##  Min.   : 8.923
##  1st Qu.:15.823
##  Median :17.177
##  Mean   :16.631
##  3rd Qu.:17.708
##  Max.   :18.655
```

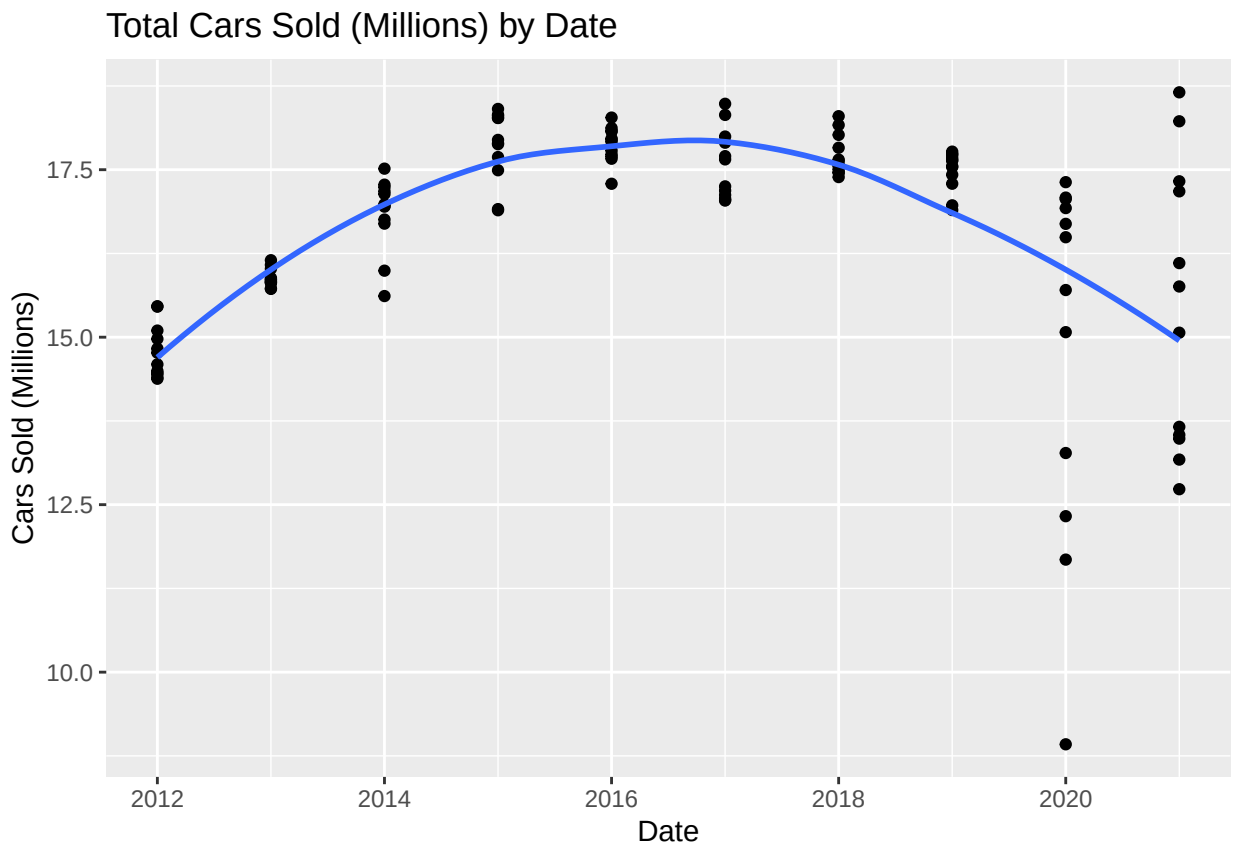


In the chart above, a data point was noticed in 2020 that showed lower sales than previous months. After analyzing the subgroups of gas, it appears that this was a low point across all types of gas. Although this deviates from the rest of the data points, it is still be considered for analysis.

Overall, the total gas sales/deliveries seem to be in a decline over the last few years.



The histogram of gas price shows a normal distribution with the average around $2.88 per gallon.



Total cars sold shows a decrease from 2017-2022 after a recovering from 2012. 2020 shows a wider spread of points on a monthly basis as well.

Relationships

Relationship between the variables may not come from all three areas of interest, so individual areas will also be explored to see if a prediction could be made on gas deliveries. Depending on those results, new dataframes may be created for new summary information. New variables will be created to understand the mean and sum at a monthly and yearly interval. This will also allow to determine if there are any seasonal impact to gas delivered.

For the the next steps in the project, I will need to do additional research to determine what machine learning can be used in this scenario to train the model to predict gas deliveries. A linear regression model in under consideration due to wanting to predict the volume of gas deliveries over time. Questions that still need to be answered is understanding how much of a relationship is involved with the variables (if any) or if there is any trend to help predict gas deliveries for the future.

Problem Statement Addressed

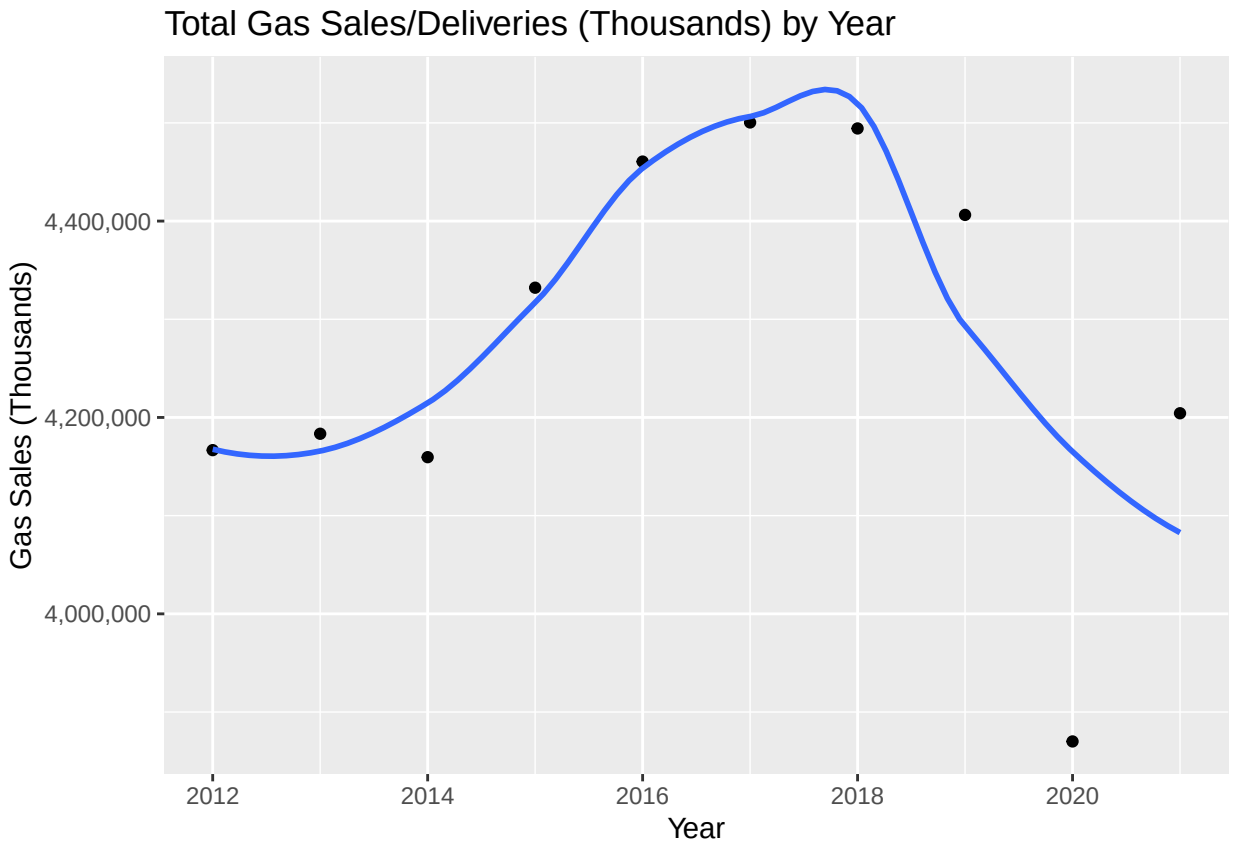
Utilizing the consolidated dataset prepared using the steps listed above, analysis was performed in order to determine the significance of the relationship of the variables to explain and predict petroleum sales/deliveries. To understand the trends of petroleum sales and deliveries the data was analysed at a yearly and monthly level. At a high-level the relationship between petroleum sales, gas price per gallon, and cars sold were also evaluated. Utilizing this information will allow an evaluation on what factors should be considered when predicting future petroleum sales for environmental and fuel industry purposes.

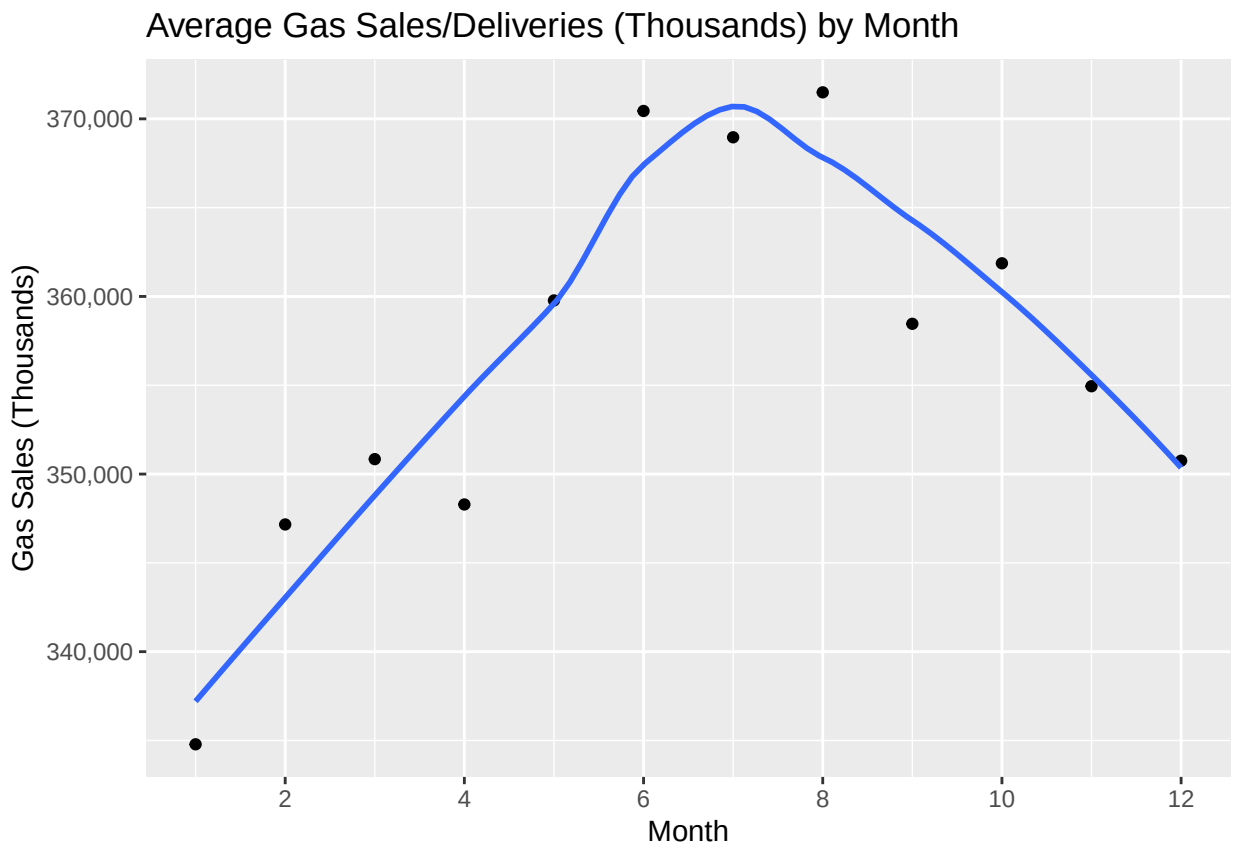
The methodology utilized consistent of a baseline understanding of petroleum sales, a breakdown of gas type sales monthly and yearly, and qualifying the relationship through exploratory analysis, covariance, and

correlation of the high-level variables. With the dataset utilized, a linear regression model could be used to predict future values because it can model the relationship between inputs and a continuous numerical output variable.

Analysis

The month and year trends for petroleum sale and deliveries were taken into consideration for potential patterns. Based on the chart showing the trend of total petroleum sales by year, there appears to be a small decline in the last two years sales with a much larger dip in 2020. This change may be associated with the economic changes from the COVID pandemic.





When evaluating the average sales by month, petroleum sales appear to peak between June and August. A hypothesis can be made that during the warmer months, people are more likely to travel and utilize more gas.

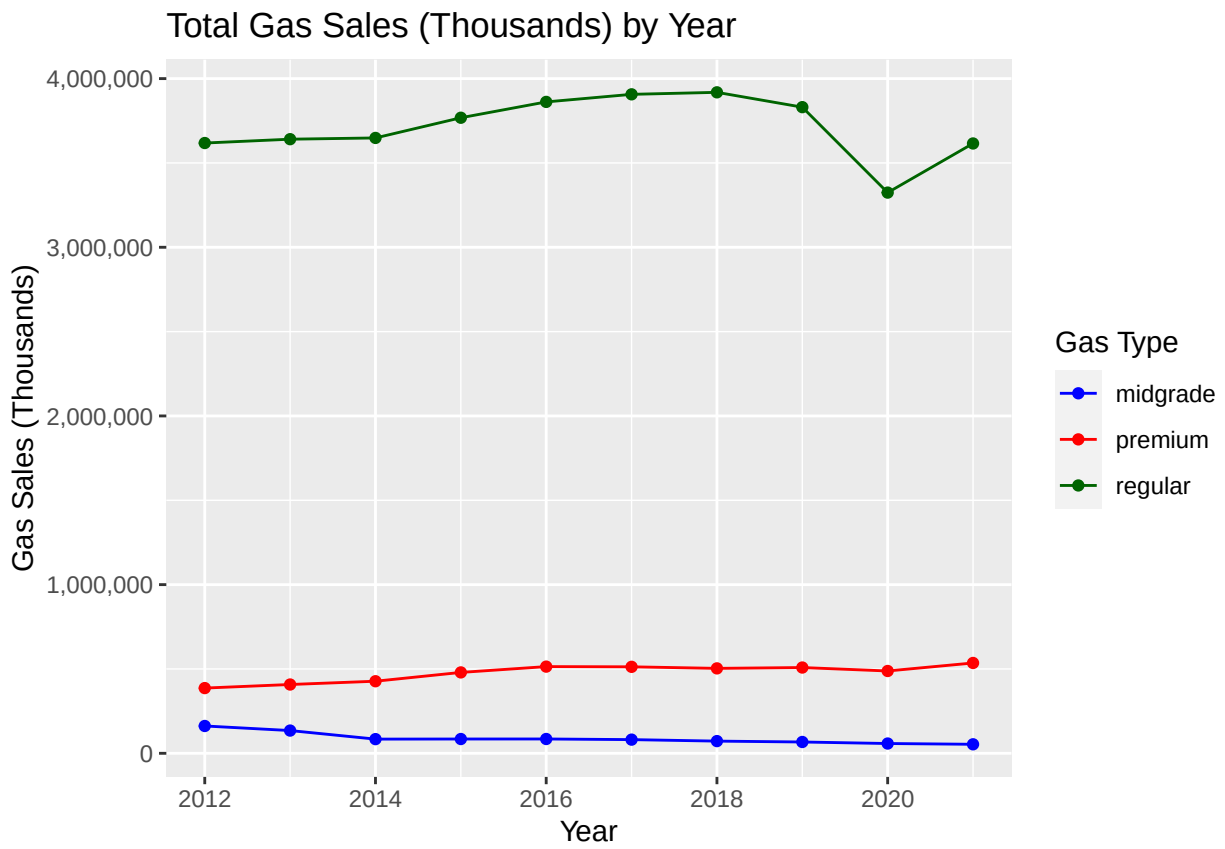
Summary information regarding the sales by year and month are listed below.

Total Gas Sales by Year Summary

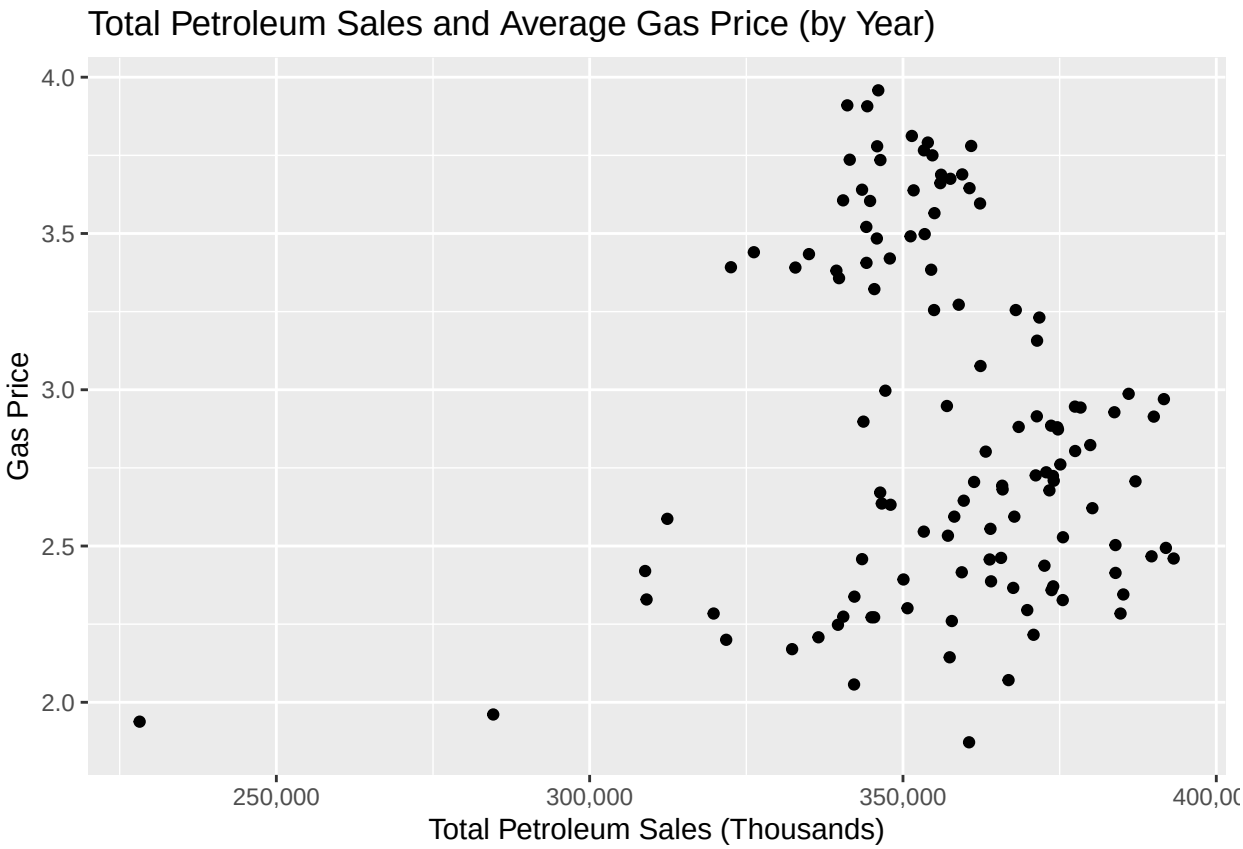
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 3870048 4170895 4268208 4277780 4447017 4500437
```

Average Gas Sales by Month Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  334785  350137  356703  356482  363641  371493
```

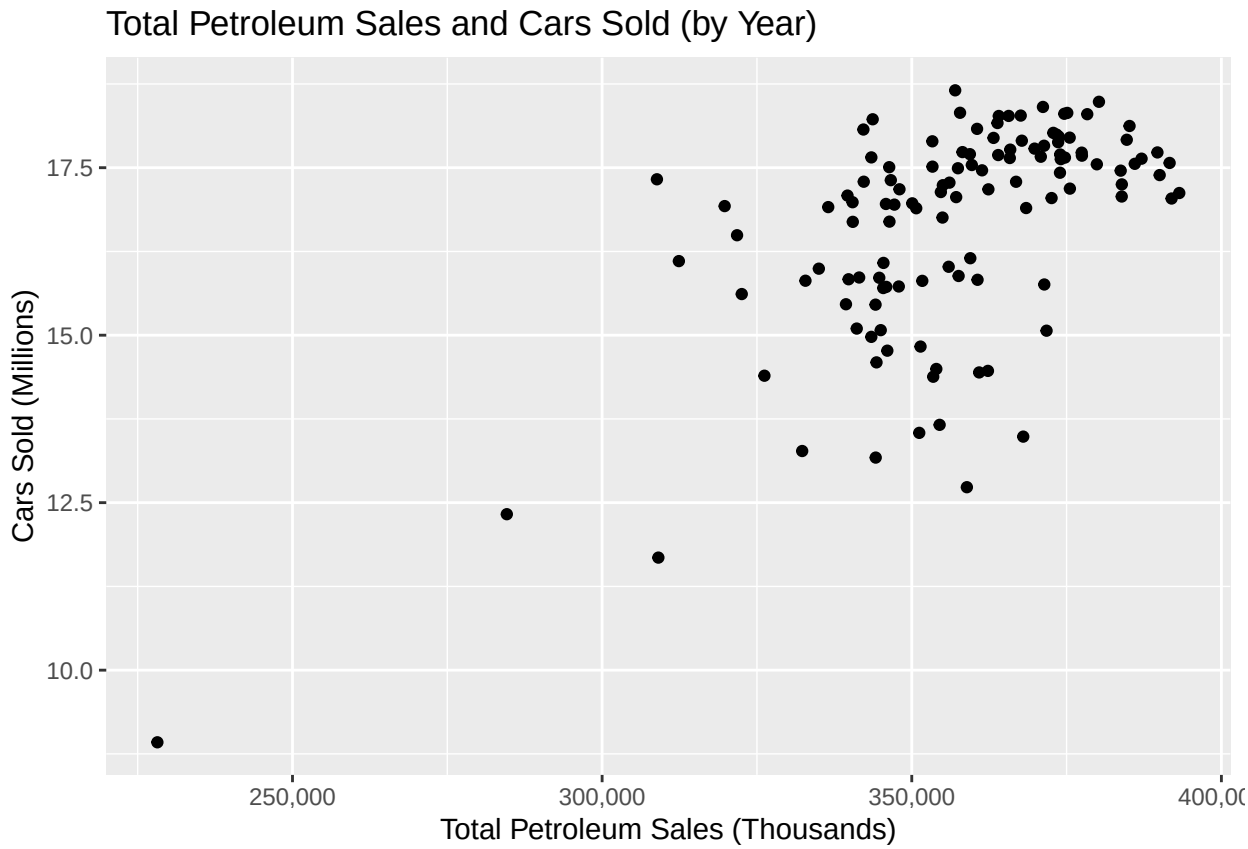


When breaking down petroleum sales by gas type, regular gas sells significantly more than premium and midgrade gas combined. Premium gas is the second most popular gas with midgrade gas at the lowest. We can hypothesize that any changes to regular gas would have a larger impact to the sales than any other gas type.



When analyzing the relationship between the datasets, scatter plots were created in order to visualize any potential correlation.

The relationship between total petroleum sales and average gas price shows a lot of variance without any clear groupings.



In contrast, the relationship between total petroleum sales and cars sold shows some closely related points between 16 and 19 million cars sold showing a potential significant relationship.

After further analysis, it was found that the following relationships exist:

- Total petroleum sales and average gas prices have a negative covariance
- Total petroleum sales and cars sold have a positive covariance
- Average gas prices and car sales have a negative covariance

```
##                totalGasSales  allGasPrice      carSales
## totalGasSales 506913938.8949 -191.0942124 23052.6160019
## allGasPrice        -191.0942    0.3054280    -0.2883094
## carSales          23052.6160   -0.2883094     2.6610017
```

Evaluating the correlation between the variables showed that average gas prices accounted for .02% of the variance of total petroleum sales while car sales accounts for 39.3%.

```
##               totalGasSales  allGasPrice  carSales
## totalGasSales  1.0000000000 0.0002358588 0.3939681
## allGasPrice    0.0002358588 1.0000000000 0.1022736
## carSales       0.3939680948 0.1022735999 1.0000000
```

Overall, car sales appears to be the best predictor with the data available for this analysis; however, it would not be recommended to create the linear regression model with just these predictors. Other factors should be considered.

Implications

The implications involved with the research conducted are that using this research will lead to a better understanding of how petroleum has been used in order to predict the future direction of sales and deliveries in the future. Utilizing the ability to measure cars sold can be one potential predictor to understand how much to expect to sell and separately understanding the seasonality involved with the petroleum sales. With additional analysis and other external factors being considered, it would lead to better understandings of the current rate of petroleum usage to measure the environmental impacts of consuming the resource.

Limitations

The limitations of this analysis are:

- only average gas prices and number of cars sold were considered as potential predictors for a linear regression model
- potential inherited bias using cars sold due to the assumption that petroleum is the primary resource used to power the vehicle
- data collected was conducted by surveys of third party resources and other government agencies as the main source for the datasets utilized
- data was only considered for the last 10 years as a representative sample according to current economic conditions.

Considering other environmental factors would help improve a model that could predict the future petroleum sales.

Concluding Remarks

In conclusion, petroleum sales/deliveries have declined in the last two years potentially due to the pandemic, but see an increase in between June and August on average. Regular gas makes up most of the petroleum sales overall. The average gas price has a negative covariance and accounts for .02% of the variance of petroleum sales. Total car sales accounts for 39.3% of the variance with a positive covariance. A limitation of the dataset used did not specify if petroleum gas was the primary method of fuel for the vehicle. Additional factors should be considered when building a linear regression model to determine the future sales of petroleum. This will have an impact on the fuel industry and environmental policies.

Code Appendix

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
## Set the working directory
setwd("C:/Users/jcamp/Documents/DSC520/Assignments/Final_Project")

library(dplyr)
library(ggplot2)
library(lubridate)
library(readr)
library(readxl)
library(stringr)
library(tidyverse)
library(scales)


## Load total gas sales
gasSales_df <- read_excel("PET_CONS_PRIM_DCU_NUS_M.xlsx", sheet = "Data 1", skip = 2)

## Load avg gas prices
```

```r
gasPrice_df <- read_excel("PET_PRI_GND_DCUS_NUS_M.xlsx", sheet = "Data 1", skip = 2)

## Load the car sales
carSales_df <- read_csv("TOTALSA.csv")

gasSales_df <-
  rename(gasSales_df, totalGasSales = 'U.S. Total Gasoline All Sales/Deliveries by Prime Supplier (Thous
                      regGasSales = 'U.S. Regular Gasoline All Sales/Deliveries by Prime Supplier (Thous
                      midGasSales = 'U.S. Gasoline Midgrade All Sales/Deliveries by Prime Supplier (Thou
                      premGasSales = 'U.S. Premium Gasoline All Sales/Deliveries by Prime Supplier (Thou
                      )

gasPrice_df <-
  rename(gasPrice_df, allGasPrice = 'U.S. All Grades All Formulations Retail Gasoline Prices (Dollars pe
                      regGasPrice = 'U.S. Regular All Formulations Retail Gasoline Prices (Dollars per
                      midGasPrice = 'U.S. Midgrade All Formulations Retail Gasoline Prices (Dollars per
                      premGasPrice = 'U.S. Premium All Formulations Retail Gasoline Prices (Dollars per
  )

carSales_df <-
  rename(carSales_df, carSales = 'TOTALSA')

year <- year(gasSales_df$Date)
month <- month(gasSales_df$Date)
gasSales_df <- cbind(gasSales_df, year, month)

year <- year(gasPrice_df$Date)
month <- month(gasPrice_df$Date)
gasPrice_df <- cbind(gasPrice_df, year, month)

year <- year(carSales_df$DATE)
month <- month(carSales_df$DATE)
carSales_df <- cbind(carSales_df, year, month)

gasSales_df <- gasSales_df %>%
                select(Date, year, month, totalGasSales, regGasSales, midGasSales, premGasSales) %>%
                filter(Date >= '2012-01-01' & Date < '2022-01-01')

gasPrice_df <- gasPrice_df %>%
                select(Date, year, month, allGasPrice, regGasPrice, midGasPrice, premGasPrice) %>%
                filter(Date >= '2012-01-01' & Date < '2022-01-01')

carSales_df <- carSales_df %>%
                filter(DATE >= '2012-01-01' & DATE < '2022-01-01')

carSales_df <- carSales_df %>%
                select(year, month, carSales)

df_list <- list(gasSales_df, gasPrice_df, carSales_df)

gas_df <- df_list %>% reduce(full_join, by.x=c('year', 'month'))

options(scipen = 999)
```

```
gas_summary <- gas_df %>%
  group_by(year) %>%
  summarize('Gas Delivered (Thousands)'=sum(totalGasSales),'Average Gas Price'=mean(allGasPrice), 'Total
  head(21)

gas_summary %>%
  mutate(across('Gas Delivered (Thousands)', ~ format(.x, big.mark = ","))) %>%
  knitr::kable(digits=2, align = "c")
summary(gas_df[c("totalGasSales","regGasSales","midGasSales","premGasSales","allGasPrice","regGasSales"
ggplot(gas_df, aes(year,totalGasSales)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "loess", se = FALSE) +
  scale_y_continuous(label=comma) +
  labs(title = 'Total Gas Sales/Deliveries (Thousands) by Date', x= 'Date', y = 'Total Gas Sales/Delive
h <- hist(gas_df$allGasPrice, breaks = "FD", plot = FALSE)

ggplot(gas_df, aes(x=allGasPrice)) +
  geom_histogram(breaks = h$breaks, col = "white", aes(y = ..density..)) +
  labs(title = 'Gas Price Histogram', x='Gas Price', y = 'Density')+
  stat_function(fun = dnorm,
    args = list(mean = mean(gas_df$allGasPrice),
    sd = sd(gas_df$allGasPrice)),
    col = "blue",
    size = 1)
ggplot(gas_df, aes(year,carSales)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "loess", se = FALSE) +
  labs(title = 'Total Cars Sold (Millions) by Date', x= 'Date', y = 'Cars Sold (Millions)')
# Line chart to show the trend of petroleum sales by year
totalGasSales_df <- gas_df %>%
  group_by(year) %>%
  summarize(totalGasSales = sum(totalGasSales))

ggplot(totalGasSales_df, aes(year, totalGasSales)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "loess", se = FALSE) +
  scale_y_continuous(label=comma) +
  labs(title = 'Total Gas Sales/Deliveries (Thousands) by Year', x= 'Year', y = 'Gas Sales (Thousands)'

# Line chart to show the trend of petroleum sales by month
totalGasSalesM_df <- gas_df %>%
  group_by(month) %>%
  summarize(avgGasSales = mean(totalGasSales))

ggplot(totalGasSalesM_df, aes(month,avgGasSales)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "loess", se = FALSE) +
  scale_y_continuous(label=comma) +
  scale_x_continuous(breaks=c(2, 4, 6, 8, 10, 12)) +
  labs(title = 'Average Gas Sales/Deliveries (Thousands) by Month', x= 'Month', y = 'Gas Sales (Thousand
summary(totalGasSales_df$totalGasSales)
summary(totalGasSalesM_df$avgGasSales)
gas_sales_df <- gather(gas_df, `totalGasSales`, `regGasSales`, 'midGasSales', 'premGasSales',
```

```
       key = "gas_type", value = "gas_sales")

gas_sale_f_df <- gas_sales_df %>%
  filter(gas_type != "totalGasSales") %>%
  group_by(year, gas_type)%>%
  summarize(sumGasSales = sum(gas_sales))


ggplot(gas_sale_f_df, aes(colour = gas_type, x = year, y = sumGasSales)) +
  geom_point() +
  geom_line() +
  scale_y_continuous(label=comma) +
  scale_color_manual(labels = c("midgrade", "premium", "regular"),
                     values = c("blue", "red", "darkgreen")) +
  guides(color = guide_legend(title = "Gas Type")) +
  labs(title = 'Total Gas Sales (Thousands) by Year', x= 'Year', y = 'Gas Sales (Thousands)')
#Scatterplot of petroleum sales and national average gas prices
ggplot(gas_df, aes(totalGasSales,allGasPrice)) +
  geom_point() +
  scale_x_continuous(label=comma) +
  labs(title = 'Total Petroleum Sales and Average Gas Price (by Year)', x= 'Total Petroleum Sales (Thous
#Scatterplot of petroleum sales and cars sold
ggplot(gas_df, aes(totalGasSales,carSales)) +
  geom_point() +
  scale_x_continuous(label=comma) +
  labs(title = 'Total Petroleum Sales and Cars Sold (by Year)', x= 'Total Petroleum Sales (Thousands)',
#Covariance and Correlation
corGas_df <- gas_df %>%
  select(totalGasSales, allGasPrice, carSales)

cov(corGas_df)
cor(corGas_df)^2
```

References

Federal Reserve Bank of St. Louis. 2022. "Total Vehicle Sales." https://fred.stlouisfed.org/series/TOTALSA.

U.S. Energy Information Administration. 2022a. "Petroleum & Other Liquids: Prime Supplier Sales Volumes." https://www.eia.gov/dnav/pet/pet_cons_prim_dcu_nus_m.htm.

———. 2022b. "U.s. Gasoline and Diesel Retail Prices." https://www.eia.gov/dnav/pet/pet_pri_gnd_dcus_nus_m.htm.