# Analyzing Oil and Gas Stocks in Election Years

Insights from the 2016 and 2020 American Presidential Elections

**Statistics in Engineering (INE2002) Term Project**

at Bahcesehir University
at the Department of Artificial Intelligence Engineering
**https://github.com/bugraaldal/ine2002**

| | |
|---|---|
| Course Instructor: | Dr. Tankut ATAN |
| Lab Instructor: | Gulfem ER |
| Course Section: | 2 |
| Lab Section: | 903 |


| | |
|---|---|
| Author: | Bugra Aldal<br>bugra.aldal@bahcesehir.edu.tr<br>2101669 |


| | |
|---|---|
| Submission: | 6th June 2023 |

# Abstract

The oil sector is crucial to the world economy, and it is affected by a variety of factors, including political events like presidential elections. The 2016 and 2020 US presidential elections' effects on oil and gas company stock prices are examined in this study. Techniques of statistical analysis are used to examine the connection between election results and stock market performance. This study uses data from the three months before and three months after the elections to identify any possible election-related effects. To investigate the connection between election outcomes and oil stock prices, statistical techniques are applied.

# Contents

# List of Figures

## List of Tables

# 1 Introduction

The oil and gas stock market, which acts as a financial marketplace for the trading of shares in businesses engaged in the exploration, production, refining, and distribution of oil and gas, is crucial to the world economy. This market includes the energy sector, which is essential for supplying the energy needs of the global economy and a variety of sectors. The oil and gas industry, as well as the stock market, are significantly impacted by current events. Events that have the potential to undermine economic stability, change industry dynamics, and influence investor mood can have a significant impact on the stock market. These occurrences might include everything from natural disasters and geopolitical conflicts to political developments and regulatory shifts. Any disruptions in the global oil supply chain or changes in demand patterns can lead to substantial fluctuations in oil prices, which in turn impact the performance of oil and gas stocks.

The stock market is significantly influenced by political changes as well. The profitability and operations of oil and gas firms can be significantly impacted by changes in governmental rules, policies, and trade agreements.

For instance, the American presidential elections might result in policy and regulatory changes that have a direct influence on the oil market both locally and internationally. Energy policy is one area where presidential elections may affect the oil market. Candidates may take different positions on topics relating to energy, including as the production of fossil fuels, financial incentives for renewable energy, and environmental restrictions. The winning candidate's policies might influence sectors like drilling permits, pipeline building, and offshore exploration, impacting the course of the oil industry. The long-term prospects of the oil market can also be impacted by regulations relating to climate change, carbon emissions, and international accords like the Paris Agreement. Additionally, presidential elections can indirectly impact the oil market through broader economic policies. Candidates' proposed tax reforms, infrastructure spending plans, and fiscal policies can influence economic growth, consumer spending, and business investment. The overall health of the economy and its impact on oil demand can have implications for the oil market.

This report aims to investigate the potential impact of the 2016 and 2020 presidential elections on the stock prices of oil and gas companies using statistical methods.

# 2 Data Collection & Sampling

The data for this study was mainly taken from the dataset named “Stock Market Data (NASDAQ, NYSE, S&P500)” from Kaggle. The Kaggle dataset contains JSON and CSV time series files of over 1076 stocks from Forbes2000, NASDAQ, NYSE, and SP500 [1]. Stocks from Yahoo Finance were filtered for the energy sector and oil & gas industry and their information was extracted to a text file [2]. Their names were parsed with the following code:

```
parse_stock_names <- function(file_path){
  text_lines <- readLines(file_path, warn=FALSE)
  # Create an empty vector to store every other line
  selected_lines <- c()

  # Store every other line in the selected_lines vector
  for (i in seq(1, length(text_lines), by = 2)) {
    selected_lines <- c(selected_lines, text_lines[i])
  }
  return (as.list(selected_lines))
}
oil_names <- parse_stock_names("oil_and_gas.txt")
```

Then the extracted names were used to match oil  gas stocks with the Kaggle dataset. Results were filtered by 3 months before and after 2016 & 2020 US presidential elections. 3 months were chosen as a time limit since it included most of the data points and 3 months equal to a quarter in stock market. The reason why 2016 and 2020 elections were chosen is basic. In 2016 elections republicans won and in 2020 elections democrats won. We want to observe if the winner of the elections have an impact on oil & gas industry.

```
check_stocks_elections <- function(csv_paths_list){
  election_dates <- c(
    as.Date("2016-11-08"),
    as.Date("2020-11-02")
  )
  for (cp in csv_paths_list){
    data <- read.csv(cp)
    parsed_string <- strsplit(cp, "/")[[1]][length(strsplit(cp, "/")
       [[1]])]
    parsed_string <- stringr::str_remove(parsed_string, "\\.csv")
    data$Date <- as.Date(data$Date, format="%d-%m-%Y")
    subset_data <- data %>% filter(Date >= (election_dates - 90) & Date
       <= (election_dates + 90))
    if (nrow(subset_data) != 0){
      write.csv(subset_data, file = paste("oil_subset_elections_stock_
         market/elections_stock_market/", parsed_string, ".csv", sep="")
         , row.names = FALSE)
    }}}
```

After these filters, we ended up with data from 84 oil & gas industry companies. We considered their adjusted closing price for our analysis since the closing price refers to the cost of shares at the end of the trading day, but the adjusted closing price takes dividends, stock splits, and new stock offerings into consideration. The adjusted closing price is a more accurate indicator of stock value since it starts where the closing price finishes.

```
division_date_2016 <- as.Date("2016-11-08")
division_date_2020 <- as.Date("2020-11-02")

for (pth in files){
  data <- read.csv(pth)
  stock_name <- parse_stock_name(pth)
  data$Year <- format(as.Date(data$Date), "%Y")
  # Split the data into two based on the date ranges
  data_2016_2017 <- data[data$Year %in% c("2016", "2017"), ]
  data_2020_2021 <- data[data$Year %in% c("2020", "2021"), ]
  if (nrow(data_2016_2017!=0)){
    mean_list_2016 <- add_mean(data_2016_2017, mean_list_2016, stock_
        name)

    before_subset_2016 <- data_2016_2017[data_2016_2017$Date < division_
        date_2016, ]
    after_subset_2016 <- data_2016_2017[data_2016_2017$Date >= division_
        date_2016, ]

    mean_list_2016_after <- add_mean(after_subset_2016, mean_list_2016_
        after, stock_name)

    mean_list_2016_before <- add_mean(before_subset_2016, mean_list_2016
        _before, stock_name)
    }
  if(nrow(data_2020_2021)!=0){
    if(stock_name != "CRC"){ # CRC causes an issiue
    mean_list_2020 <- add_mean(data_2020_2021, mean_list_2020, stock_
        name)

    before_subset_2020 <- data_2020_2021[data_2020_2021$Date < division_
        date_2020, ]
    after_subset_2020 <- data_2020_2021[data_2020_2021$Date >= division_
        date_2020, ]

    mean_list_2020_after <- add_mean(after_subset_2020, mean_list_2020_
        after, stock_name)

    mean_list_2020_before <- add_mean(before_subset_2020, mean_list_2020
        _before, stock_name)
  }}

}

mean_dataframe_2020 <- create_dataframe(mean_list_2020)
mean_dataframe_2016 <- create_dataframe(mean_list_2016)
```

In order to work with continuous data, we took their means. We classified these in 2 main categories.

- Mean Data from 2016 Elections
  - Mean stock prices within 3 months range of 2016 elections
  - Mean stock prices from 3 months before 2016 elections
  - Mean stock prices from 3 months after 2016 elections
- Mean Data from 2020 Elections
  - Mean stock prices within 3 months range of 2020 elections
  - Mean stock prices from 3 months before 2020 elections
  - Mean stock prices from 3 months after 2020 elections

# 3 Data Analysis & Visualization

## 3.1 Scatter Plot & Histogram

Our data immediately shows signs of outliers when we plot scatter plots.

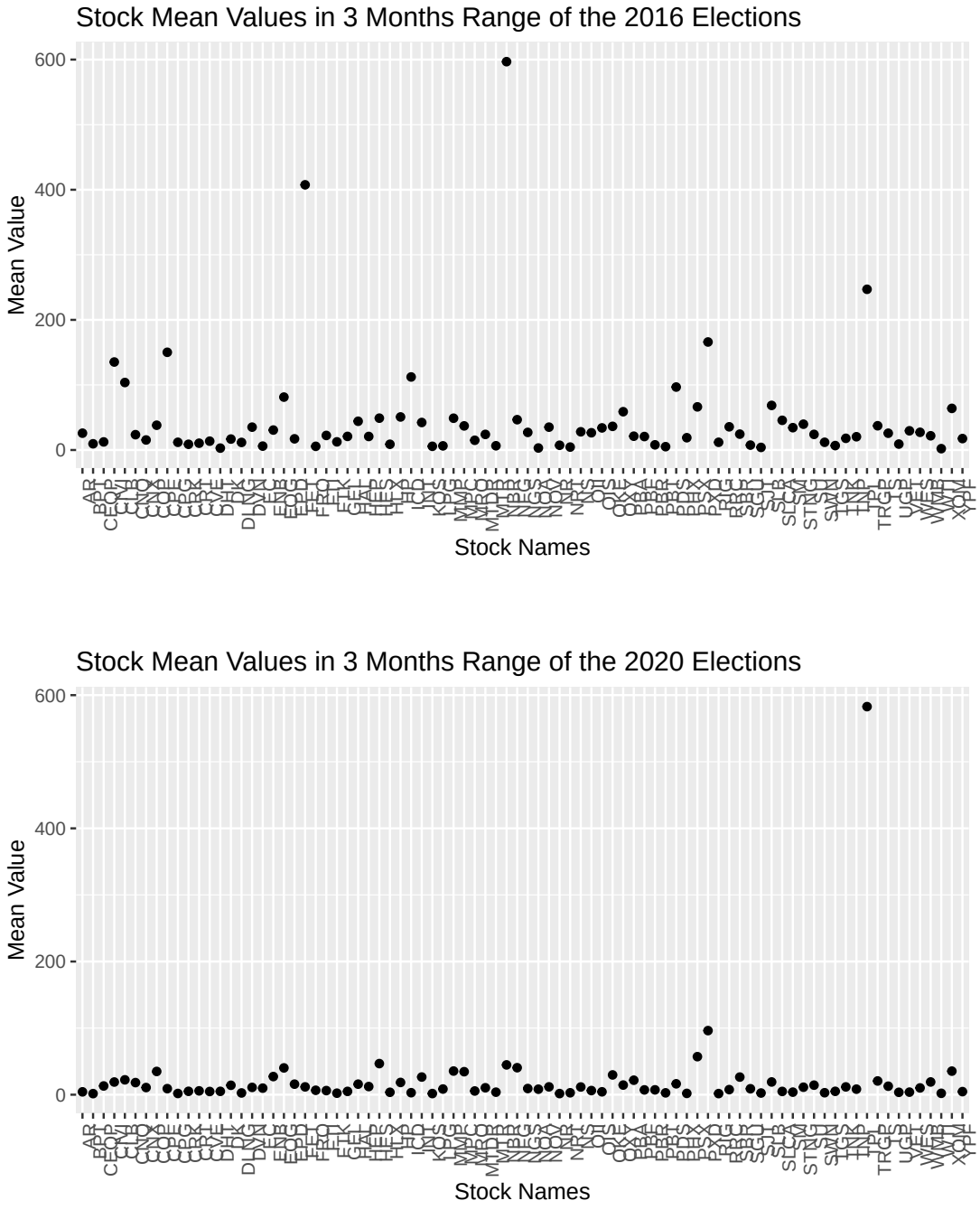



Figure 1 Scatter plots of 2016 & 2020 mean adjusted closing price

```
add_scatter_plot <- function(mean_data, year, before=FALSE, after=FALSE)
    {
  title <- get_title(year, before, after)
  ggplot(mean_data, aes(x = Stock, y = Mean_Value)) +
  geom_point() +
  labs(title = title, x = "Stock Names", y = "Mean Value") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
}
add_scatter_plot(mean_dataframe_2020, 2020)
add_scatter_plot(mean_dataframe_2016, 2016)
```

Plotting the stock by years in the same plot can help us understand more.

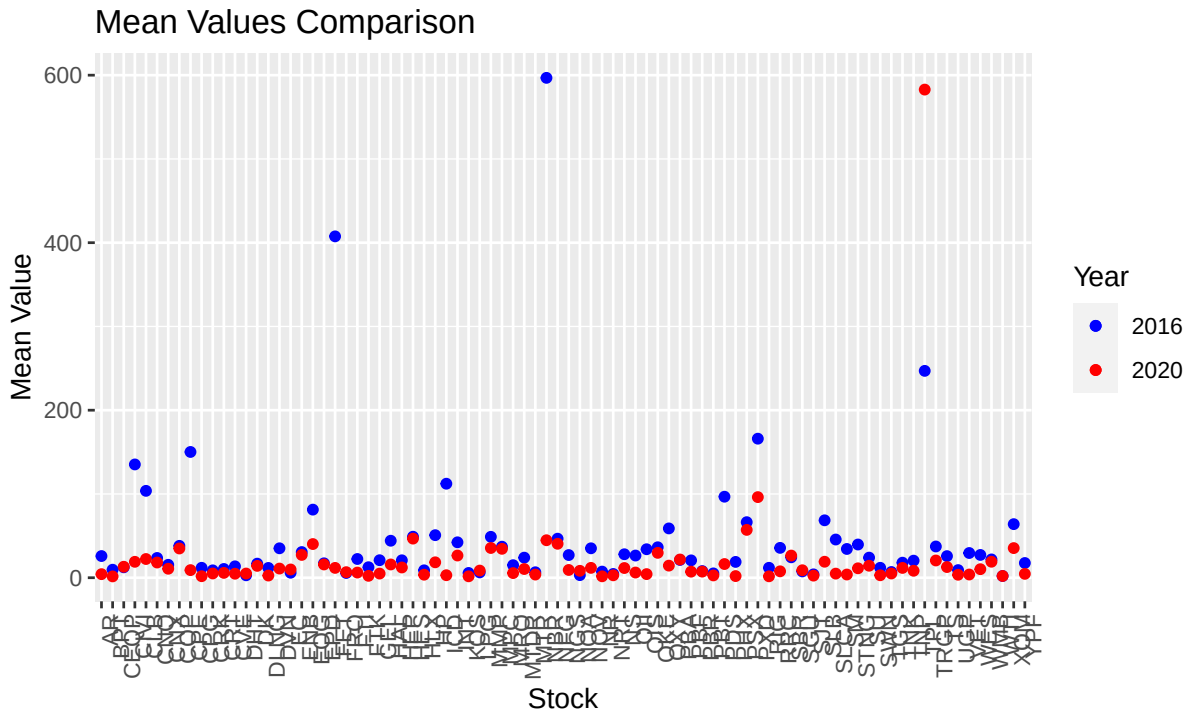



Figure 2 Scatter plot of 2016 & 2020 Elections

```
# Merge the dataframes based on the "Stock" column
merged_dataframe <- merge(mean_dataframe_2016, mean_dataframe_2020, by =
    "Stock", suffixes = c("_2016", "_2020"))

# Plot the scatter plot
ggplot(merged_dataframe) +
  geom_point(aes(x = Stock, y = Mean_Value_2016, color = "2016")) +
  geom_point(aes(x = Stock, y = Mean_Value_2020, color = "2020")) +
  labs(x = "Stock", y = "Mean Value", color = "Year") +
  ggtitle("Mean Values Comparison") +
  scale_color_manual(values = c("2016" = "blue", "2020" = "red")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme_bw()
```

2016 mean values seem to be generally above 2016 values. Some 2016 values seem to be way above their 2016 prices.

Histogram shows us that our data seems to be left skewed.

```
plot_ditribution <- function(mean_data, year, before=FALSE,  after=FALSE
    ){
  title <- get_title(year, before, after)
  ggplot(mean_data, aes(x = Mean_Value)) +
    geom_histogram(binwidth = NUM_BREAKS(mean_data), color="black", fill
        = "steelblue", linewidth=0.5) +
    labs(title = paste("Histogram Plot of", title), x = "Mean Value", y
       = "Count") +
    geom_vline(aes(xintercept=mean(Mean_Value)),
               color="red", linetype="dashed", linewidth=1)
}
plot_ditribution(mean_dataframe_2016, 2016)
plot_ditribution(mean_dataframe_2020, 2020)
```

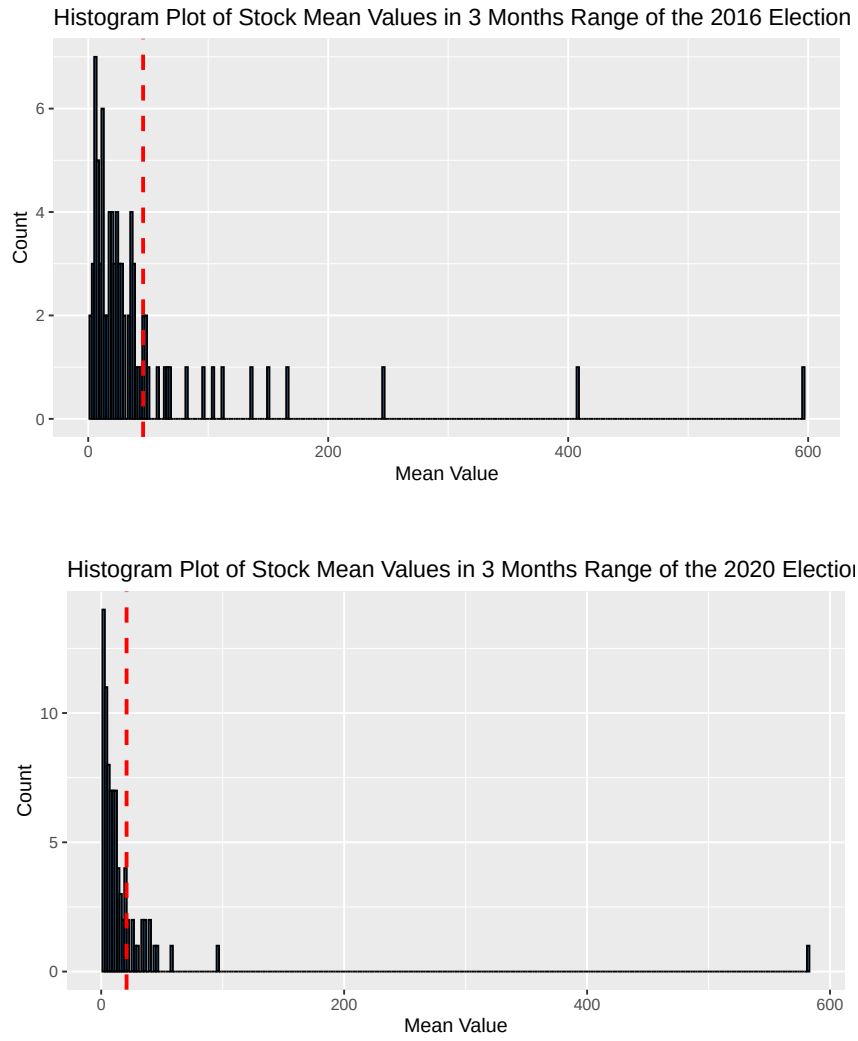



Figure 3 Histogram plots of 2016 & 2020 mean adjusted closing price

## 3.2 Data Description

### 3.2.1 Measures of Central Tendency

```
# Define a function to calculate the mode
get_mode <- function(x) {
  ux <- unique(x)
  return(ux[which.max(tabulate(match(x, ux)))])
}
measures_of_central_tendency <- function(df){
# Calculate the median
median_value <- median(df$Mean_Value)


# Calculate the mode
mode_value <- get_mode(df$Mean_Value)

# Calculate the midrange
min_value <- min(df$Mean_Value)
max_value <- max(df$Mean_Value)
midrange <- (min_value + max_value) / 2
print(paste(median_value, mode_value, midrange))
}
```

| | Median | Mode | Midrange | Mean |
|---|---|---|---|---|
| **2016** | 23.7739 | 25.8719 | 299.4341 | 45.7405 |
| **2016 3 Months After** | 24.2738 | 25.0965 | 353.0682 | 49.4060 |
| **2016 3 Months Before** | 22.5267 | 26.6474 | 245.7999 | 42.1026 |
| **2020** | 9.5586 | 4.2605 | 292.1915 | 20.8814 |
| **2020 3 Months After** | 10.7259 | 5.1713 | 341.2350 | 23.5004 |
| **2020 3 Months Before** | 8.5826 | 3.4066 | 246.1405 | 18.3890 |

Table 1 Measures of Central Tendency

The median value for the 3-month period after 2016 and 2020 is higher compared to the corresponding 3-month period before. This indicates a shift towards higher stock prices during those periods.

### 3.2.2 Measures of Variation

```
calculate_range <- function(data) {
  max_val <- max(data)
  min_val <- min(data)
  range_val <- max_val - min_val
  return(range_val)
}

calculate_variance <- function(data) {
  variance_val <- var(data)
  return(variance_val)
}
```

```
calculate_std_dev <- function(data) {
  std_dev_val <- sd(data)
  return(std_dev_val)
}
```

| | Range | Variance | Standard Deviation |
|---|---|---|---|
| **2016** | 594.7378 | 6939.886 | 83.30598 |
| **2016 3 Months After** | 701.2015 | 9073.137 | 95.25302 |
| **2016 3 Months Before** | 488.274 | 5145.432 | 71.73167 |
| **2020** | 581.1934 | 4076.287 | 63.8458 |
| **2020 3 Months After** | 678.7806 | 5537.874 | 74.41689 |
| **2020 3 Months Before** | 489.8511 | 2912.65 | 53.96897 |

Table 2 Measures of Variation

In comparison to the same 3-month time previously, the range is wider during the 3-month period between 2016 and 2020. This implies that there might have been more fluctuation in stock values during those times. The 3-month period between 2016 and 2020 has a higher variation than the 3-month period before, which is similar to the range. This suggests that there was greater price variation during those times. Consistent with the range and variance, the standard deviation is higher in the 3-month period after 2016 and 2020 compared to the 3-month period before. This suggests greater volatility and fluctuations in stock prices during those periods. Overall, these measures indicate that there may have been higher volatility and variability in stock prices during the 3-month periods after 2016 and 2020 compared to the corresponding 3-month periods before.

### 3.2.3 Exploratory Data Analysis & Outliers

Using the $\pm 1.5 * IQR$ method, we detect outliers.

```
identify_outliers <- function(dataframe, column_name) {
  # Calculate the IQR
  q1 <- quantile(dataframe[[column_name]], 0.25)
  q3 <- quantile(dataframe[[column_name]], 0.75)
  iqr <- q3 - q1

  # Determine the lower and upper bounds for outliers
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr

  # Identify outliers
  outliers <- dataframe[dataframe[[column_name]] < lower_bound |
     dataframe[[column_name]] > upper_bound, ]

  # Return outliers
```

```
  return(outliers)
}
outliers2020 <- identify_outliers(mean_dataframe_2020, "Mean_Value")
outliers2016 <- identify_outliers(mean_dataframe_2016, "Mean_Value")
```

Table 3 Outlier Mean Values for 2016

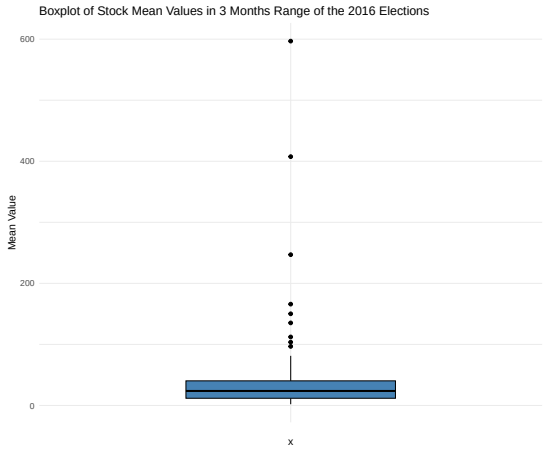
| Stock | Value |
|---|---|
| CIVI | 135.25731 |
| CLB | 103.73921 |
| CPE | 150.16452 |
| FET | 407.55161 |
| ICD | 112.24194 |
| NBR | 596.80294 |
| PDS | 96.74839 |
| PXD | 166.01280 |
| TPL | 246.97938 |

Table 4 Outlier Mean Values for 2020

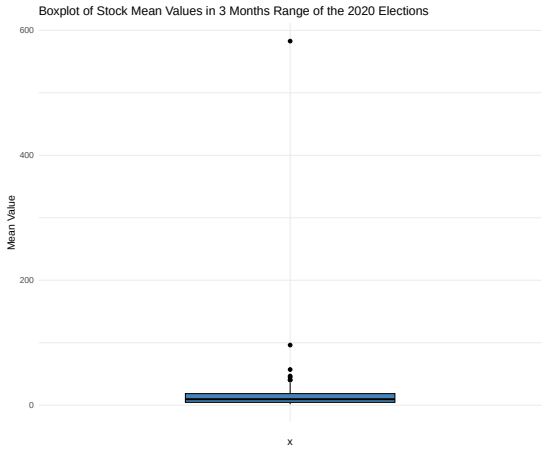
| Stock | Value |
|---|---|
| EOG | 40.21232 |
| HES | 46.69785 |
| NBR | 44.75419 |
| NFG | 40.53040 |
| PSX | 57.12918 |
| PXD | 96.24816 |
| TPL | 582.78825 |

These high outliers makes our boxplot cramped, making them harder to understand.

```
get_boxplot <- function(mean_data, year, before=FALSE, after=FALSE){
  title <- get_title(year, before, after)
  ggplot(mean_data, aes(x = "", y = Mean_Value)) +
  geom_boxplot(fill = "steelblue", color = "black", width = 0.5) +
  labs(title = paste("Boxplot of", title), y = "Mean Value") +
  theme_minimal()
}
get_boxplot(mean_dataframe_2020, 2020)
get_boxplot(mean_dataframe_2016, 2016)
```




(a) Boxplot of 2016

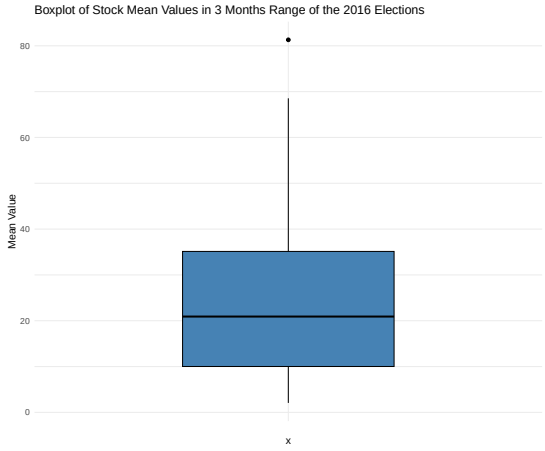



(b) Boxplot of 2020

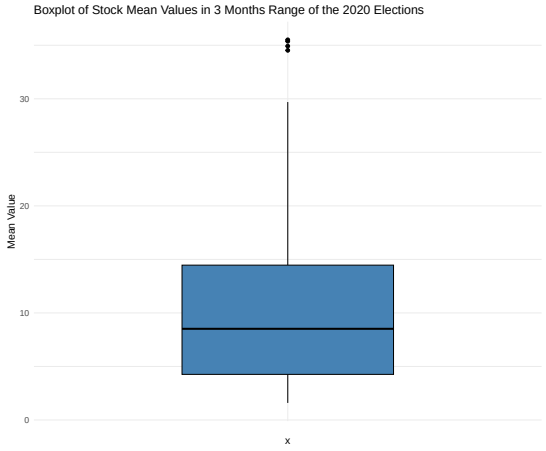
Figure 4 Boxplots of 2016 & 2020 with outliers

Plotting the boxplots without the high outliers shows us clear visualization.

```
mean_dataframe_2016_no_outliers <- mean_dataframe_2016[!(mean_dataframe_
    2016$Mean_Value %in% outliers2016$Mean_Value), ]
mean_dataframe_2020_no_outliers <- mean_dataframe_2020[!(mean_dataframe_
    2020$Mean_Value %in% outliers2020$Mean_Value), ]
get_boxplot(mean_dataframe_2016_no_outliers, 2016)
get_boxplot(mean_dataframe_2020_no_outliers, 2020)
```




(a) Boxplot of 2016 without outliers




(b) Boxplot of 2020 without outliers

Figure 5 Boxplots of 2016 & 2020 without outliers

# 4 Normality

In order the check for normality, we applied Kolgmorov-Smirnov Test, Shapiro-Wilk Test, Pearson Coefficent of Skewness.

## 4.1 Kolgmorov-Smirnov Test

The Kolmogorov-Smirnov test, sometimes referred to as the KS test, is a nonparametric statistical test used to compare two empirical distributions or the empirical distribution of a dataset to a theoretical distribution. It is based on the largest discrepancy between the expected cumulative distribution function (CDF) under the null hypothesis and the cumulative distribution function (CDF) of the actual data. The test determines whether there is a significant departure from the expected distribution in the observed data.

```
# Perform the K-S test
ks_result_2016 <- ks.test(mean_dataframe_2016$Mean_Value, "pnorm", mean(
    mean_dataframe_2016$Mean_Value), sd(mean_dataframe_2016$Mean_Value))

# Print the test statistic and p-value
cat("Test statistic:", ks_result_2016$statistic, "\n")
cat("P-value:", ks_result_2016$p.value, "\n")

# Perform the K-S test
ks_result_2020 <- ks.test(mean_dataframe_2020$Mean_Value, "pnorm", mean(
    mean_dataframe_2020$Mean_Value), sd(mean_dataframe_2020$Mean_Value))

# Print the test statistic and p-value
cat("Test statistic:", ks_result_2020$statistic, "\n")
cat("P-value:", ks_result_2020$p.value, "\n")
```

For the 2016 dataset:

- Test statistic: The test statistic is 0.3091688.
- P-value: The p-value is 1.253637e-07.

For the 2020 dataset:

- Test statistic: The test statistic is 0.381295.
- P-value: The p-value is 1.723577e-11.

Between the empirical cumulative distribution function (ECDF) of the data and the cumulative distribution function (CDF) of the normal distribution, the test statistic for both

datasets shows the largest vertical distance. The null hypothesis that the data follow a normal distribution is strongly refuted by the relatively modest p-values for both datasets. We can therefore conclude from these findings that the data for both the 2016 and 2020 datasets significantly depart from a normal distribution.

## 4.2 Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical test used to assess the normality of a dataset as well. It evaluates whether the observed data significantly deviates from a normal distribution. The test calculates a test statistic and corresponding p-value, and if the p-value is below a specified significance level, it suggests evidence of a departure from normality.

```
# Apply Shapiro-Wilk test
shapiro_test_2020 <- shapiro.test(mean_dataframe_2020$Mean_Value)
shapiro_result_2020 <- ifelse(shapiro_test_2020$p.value < 0.05, "Not
    Normally Distributed", "Normally Distributed")
# Print the results
cat("Shapiro-Wilk Test:")
print(shapiro_result_2020)
# Apply Shapiro-Wilk test
shapiro_test_2016 <- shapiro.test(mean_dataframe_2016$Mean_Value)
shapiro_result_2016 <- ifelse(shapiro_test_2016$p.value < 0.05, "Not
    Normally Distributed", "Normally Distributed")
# Print the results
cat("Shapiro-Wilk Test:")
print(shapiro_result_2016)
```

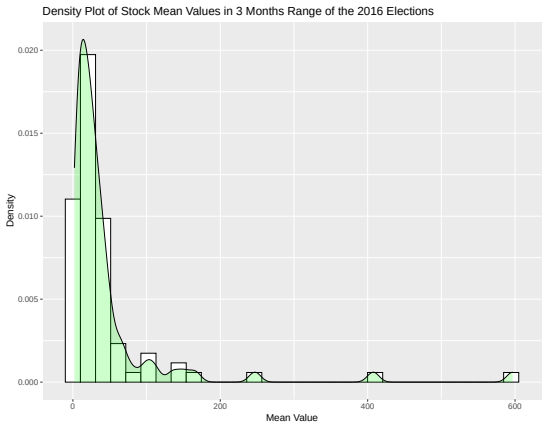
Shapiro-Wilk Test further proves our point that the data is not normally distributed.

## 4.3 Pearson Coefficent of Skewness

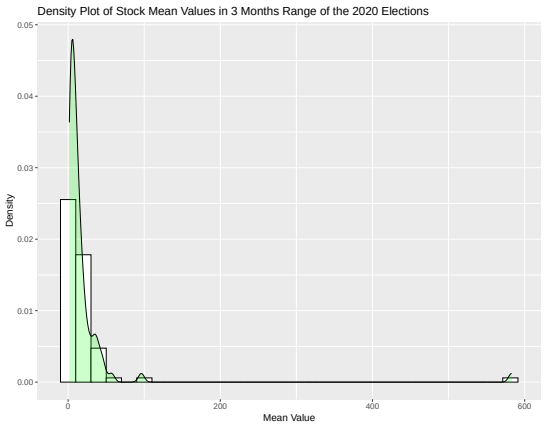
The asymmetry of a probability distribution is measured statistically using the Pearson coefficient of skewness. It reveals details regarding the degree and direction of data skewness. A right-skewed distribution with a longer tail on the right side is indicated by a positive value, whereas a left-skewed distribution with a longer tail on the left is shown by a negative number. A zero value implies a symmetric distribution.

```
> skewness(mean_dataframe_2020$Mean_Value)
[1] 8.27815
> skewness(mean_dataframe_2016$Mean_Value)
[1] 4.723053
```

2016 has a skewness coefficient of 4.723053. This also suggests favorable skewness, but less so than in 2020. However, it is not as strongly right-skewed as 2020. The skewness coefficient for 2020 data is 8.27815. This indicates a strong positive skewness, suggesting that the distribution is highly right-skewed. The majority of the values are likely concentrated on the lower end, with a few extremely large values on the right side. Both of our data is right-skewed.




(a) Density plot of 2016




(b) Density Plot of 2020

Figure 6 Density plots of 2016 & 2020

## 4.4 Margin of Error, Point Estimations, & Confidence Intervals

These results provide estimates of the mean stock values for each election period, along with a measure of uncertainty in the form of confidence intervals. The confidence intervals indicate the range of values within which the true population mean stock value is likely to fall with a 95% level of confidence.

```
> # Calculate Confidence Intervals (assuming 95% confidence level)
> conf_level <- 0.95
> margin_error_2016 <- qnorm(1 - (1 - conf_level) / 2) * sd_2016 / sqrt(
    length(mean_dataframe_2016$Mean_Value))
> conf_interval_2016 <- c(mean_2016 - margin_error_2016, mean_2016 +
    margin_error_2016)
> margin_error_2020 <- qnorm(1 - (1 - conf_level) / 2) * sd_2020 / sqrt(
    length(mean_dataframe_2020$Mean_Value))
> conf_interval_2020 <- c(mean_2020 - margin_error_2020, mean_2020 +
    margin_error_2020)
> margin_error_2016_before <- qnorm(1 - (1 - conf_level) / 2) * sd_2016_
    before / sqrt(length(mean_dataframe_2016_before$Mean_Value))
> conf_interval_2016_before <- c(mean_2016_before - margin_error_2016_
    before, mean_2016_before + margin_error_2016_before)
> margin_error_2016_after <- qnorm(1 - (1 - conf_level) / 2) * sd_2016_
    after / sqrt(length(mean_dataframe_2016_after$Mean_Value))
> conf_interval_2016_after <- c(mean_2016_after - margin_error_2016_
    after, mean_2016_after + margin_error_2016_after)
> margin_error_2020_before <- qnorm(1 - (1 - conf_level) / 2) * sd_2020_
    before / sqrt(length(mean_dataframe_2020_before$Mean_Value))
```

```
> conf_interval_2020_before <- c(mean_2020_before - margin_error_2020_
    before, mean_2020_before + margin_error_2020_before)
> margin_error_2020_after <- qnorm(1 - (1 - conf_level) / 2) * sd_2020_
    after / sqrt(length(mean_dataframe_2020_after$Mean_Value))
> conf_interval_2020_after <- c(mean_2020_after - margin_error_2020_
    after, mean_2020_after + margin_error_2020_after)
> # Print the results
> cat("Point Estimations:\n")
Point Estimations:
> cat("2016 Mean:", mean_2016, "\n")
2016 Mean: 45.74053
> cat("2020 Mean:", mean_2020, "\n")
2020 Mean: 20.88143
> cat("2016 3 Months Before Mean:", mean_2016_before, "\n")
2016 3 Months Before Mean: 42.10259
> cat("2016 3 Months After Mean:", mean_2016_after, "\n")
2016 3 Months After Mean: 49.40598
> cat("2020 3 Months Before Mean:", mean_2020_before, "\n")
2020 3 Months Before Mean: 18.38898
> cat("2020 3 Months After Mean:", mean_2020_after, "\n\n")
2020 3 Months After Mean: 23.50035

> cat("Confidence Intervals (95%):\n")
Confidence Intervals (95%):
> cat("2016 Confidence Interval:", conf_interval_2016, "\n")
2016 Confidence Interval: 27.92558 63.55548
> cat("2020 Confidence Interval:", conf_interval_2020, "\n")
2020 Confidence Interval: 7.228035 34.53483
> cat("2016 3 Months Before Confidence Interval:", conf_interval_2016_
    before, "\n")
2016 3 Months Before Confidence Interval: 26.7628 57.44238
> cat("2016 3 Months After Confidence Interval:", conf_interval_2016_
    after, "\n")
2016 3 Months After Confidence Interval: 29.03616 69.7758
> cat("2020 3 Months Before Confidence Interval:", conf_interval_2020_
    before, "\n")
2020 3 Months Before Confidence Interval: 6.847737 29.93022
> cat("2020 3 Months After Confidence Interval:", conf_interval_2020_
    after, "\n")
2020 3 Months After Confidence Interval: 7.586323 39.41437
```

# 5 Hypothesis Testing & Analysis of Variance

Our study's main aim is to see if there is a significant change before and after each election depending on the winner.

### 5.0.1 Analysis of Variance (ANOVA)

Our data is a better fit for ANOVA since its essentially means of different stocks.

```
# Combine the data frames for each election year
combined_data_2016 <- rbind(mean_dataframe_2016_before,
   mean_dataframe_2016_after)
combined_data_2020 <- rbind(mean_dataframe_2020_before,
   mean_dataframe_2020_after)
combined_data_full <- rbind(mean_dataframe_2020, mean_dataframe_2016)

# Perform ANOVA for the 2016 election
anova_2016 <- aov(Mean_Value ~ Stock, data = combined_data_2016)

# Perform ANOVA for the 2020 election
anova_2020 <- aov(Mean_Value ~ Stock, data = combined_data_2020)

# Perform ANOVA for the in between 2016 and 2020 elections
anova_full <- aov(Mean_Value ~ Stock, data = combined_data_full)

> summary(anova_2016)
            Df  Sum Sq Mean Sq F value Pr(>F)
Stock       83 1152029   13880   38.41 <2e-16 ***
Residuals   84   30353     361
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1

> summary(anova_2020)
            Df Sum Sq Mean Sq F value Pr(>F)
Stock       83 683596    8236   36.62 <2e-16 ***
Residuals   84  18895     225
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1

> summary(anova_full)
            Df Sum Sq Mean Sq F value  Pr(>F)
Stock       83 610971    7361   1.878 0.00219 **
Residuals   84 329327    3921
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1
```

Both ANOVA results for before and after elections of 2016 and 2020 show us that there are significant differences in the mean values of the stocks.
The ANOVA results in between 2016 and2020 indicate that there are significant dif-

ferences in the mean values of the stocks, suggesting that the stocks had a varying performance between these election periods.

### 5.0.2 Hypothesis Testing

Before starting this out main goal was this.
**Hypothesis:** The mean stock value during the 3 months before the 2016 election is significantly different from the mean stock value during the 3 months after the election.

- **Null Hypothesis (H0):** The mean stock value before the 2016 election is equal to the mean stock value after the election.

- **Alternative Hypothesis (HA):** The mean stock value before the 2016 election is different from the mean stock value after the election.

```
# H0: mean(stock_values_before_2016) = mean(stock_values_after_2016)
# H1: mean(stock_values_before_2016) != mean(stock_values_after_2016)

> # Perform a paired t-test to compare the mean stock values before and
    after the election
> ttest_result_2016 <- t.test(stock_values_before_2016, stock_values_
    after_2016, paired = TRUE)
> # Print the test statistic, p-value, and confidence interval
> cat("Paired t-test (3 Months Before vs. 3 Months After 2016 Election)
    :\n")
Paired t-test (3 Months Before vs. 3 Months After 2016 Election):
> cat("Test statistic:", ttest_result_2016$statistic, "\n")
Test statistic: -2.571801
> cat("P-value:", ttest_result_2016$p.value, "\n")
P-value: 0.01189898
> cat("Confidence Interval:", ttest_result_2016$conf.int, "\n")
Confidence Interval: -12.95164 -1.65515
```

The p-value suggests that there is evidence to reject the null hypothesis at a significance level of 0.05 (or lower), indicating a significant difference between the two periods. Based on these results, we can infer that there is a significant difference between the mean stock values 3 months before and 3 months after the 2016 election.
There is enough evidence to support the claim.
The stock values tend to be lower after the election compared to before the election, with a difference estimated to be between -12.95164 and -1.65515.

**Hypothesis:** The mean stock value during the 3 months before the 2020 election is significantly different from the mean stock value during the 3 months after the election.

- **Null Hypothesis (H0):** The mean stock value before the 2020 election is equal to the mean stock value after the election.

- **Alternative Hypothesis (HA):** The mean stock value before the 2020 election is different from the mean stock value after the election.

```
# H0: mean(stock_values_before_2020) = mean(stock_values_after_2020)
# H1: mean(stock_values_before_2020) != mean(stock_values_after_2020)
> # Collect the stock values for the 3 months before and after the 2020
    election
> stock_values_before_2020 <- mean_dataframe_2020_before$Mean_Value
> stock_values_after_2020 <- mean_dataframe_2020_after$Mean_Value
> # Perform a paired t-test to compare the mean stock values before and
    after the election
> ttest_result_2020 <- t.test(stock_values_before_2020, stock_values_
    after_2020, paired = TRUE)
> # Print the test statistic, p-value, and confidence interval
> cat("Paired t-test (3 Months Before vs. 3 Months After 2020 Election)
    :\n")
Paired t-test (3 Months Before vs. 3 Months After 2020 Election):
> cat("Test statistic:", ttest_result_2020$statistic, "\n")
Test statistic: -2.262165
> cat("P-value:", ttest_result_2020$p.value, "\n")
P-value: 0.02630042
> cat("Confidence Interval:", ttest_result_2020$conf.int, "\n")
Confidence Interval: -9.605423 -0.6173086
```

The p-value suggests that there is evidence to reject the null hypothesis at a significance level of 0.05 (or lower), indicating a significant difference between the two periods. Based on these results, we can infer that there is a significant difference between the mean stock values 3 months before and 3 months after the 2020 election.
There is enough evidence to support the claim.
The stock values tend to be lower after the election compared to before the election, with a difference estimated to be between -9.605423 and -0.6173086. We fail to reject the null hypothesis.
**Hypothesis:** The mean stock value of 2016 and 2020 overall are significantly different from each other.

- **Null Hypothesis (H0):** The mean stock value of 2020 election is equal to the mean stock value of 2016 the election.

- **Alternative Hypothesis (HA):** The mean stock value before the 2020 election is different from the mean stock value after the election.

```
# H0: mean(mean_dataframe_2016) = mean(mean_dataframe_2020)
# H1: mean(mean_dataframe_2016) != mean(mean_dataframe_2020)
> # Perform the paired t-test
> ttest_result_ <- t.test(mean_dataframe_2016$Mean_Value,
    mean_dataframe_2020$Mean_Value, paired = TRUE)
> # Print the test results
> cat("Paired t-test (2016 vs. 2020 Elections):\n")
Paired t-test (2016 vs. 2020 Elections):
> cat("Test statistic:", ttest_result_$statistic, "\n")
Test statistic: 2.66478
> cat("P-value:", ttest_result_$p.value, "\n")
P-value: 0.009255055
> cat("Confidence Interval:", ttest_result_$conf.int, "\n")
Confidence Interval: 6.304568 43.41362
```

Based on the positive test result and the p-value being below the significance level, it seems likely that the mean stock values in 2020 will be higher than they were in 2016. A range for assessing the size of the difference in mean stock values between the two election years is also provided by the confidence interval.

It seems that oil & gas stock prices go down regardless of who wins, whether Republicans or Democrats. Although, both lower and higher bounds of our confidence interval is lower in 2016 elections. This indicates that there is a high level of confidence that the value of price is close to its original value which means the stocks went down less in the year 2016 where Republicans won.

# 6 Nonparametic Tests

### 6.0.1 Wilcoxon Signed-Rank Test

```
> cat("Wilcoxon Signed-Rank Test (2016 Before vs. After Elections):\n")
Wilcoxon Signed-Rank Test (2016 Before vs. After Elections):
> cat("Test statistic:", wilcox_result$statistic, "\n")
Test statistic: 484
> cat("P-value:", wilcox_result$p.value, "\n")
P-value: 6.633855e-09
```

The stock values exhibited a significant change during the 3 months surrounding the 2016 elections, and the observed difference is unlikely to occur by chance alone.

### 6.0.2 Mann-Whitney U test

Unlike Wilcoxon Signed-Rank test, Mann-Whitney U test suggests that there is no significant difference between the values.

```
> mannwhitney_result <- wilcox.test(mean_dataframe_2016_before$Mean_
    Value, mean_dataframe_2016_after$Mean_Value)
> cat("Mann-Whitney U Test (2016 Before vs. After Elections):\n")
Mann-Whitney U Test (2016 Before vs. After Elections):
> cat("Test statistic:", mannwhitney_result$statistic, "\n")
Test statistic: 3351
> cat("P-value:", mannwhitney_result$p.value, "\n")
P-value: 0.5755457
```

# 7 Linear Regression

Using the individual stocks and means for linear regression isn't a good practice. Therefore we picked a random stock and used its data for our model. We excluded the `Date` column since in this case its more of an identity variable. We also excluded the `Close` column because it is highly positively correlated, exactly the same at times, with `Adjusted.Close`.
We split our data into training and testing sets with proportions 80%, 20% respectively.

```
stock_data <- read.csv("oil_subset_stock_market/VET.csv")
# Remove Date
stock_data <- stock_data[, -1]
stock_data
# Set the seed for reproducibility
set.seed(123)

# Calculate the number of rows for training and testing
n_rows <- nrow(stock_data)
train_size <- round(0.8 * n_rows)

# Randomly sample indices for training
train_indices <- sample(seq_len(n_rows), size = train_size, replace =
    FALSE)

# Create the training and testing sets
train_data <- stock_data[train_indices, ]
test_data <- stock_data[-train_indices, ]
```

We used applied linear regression model and calculated its Mean Squared Error error.

```
# Apply linear regression on the training data
lm_model <- lm(Adjusted.Close ~ Low + Open + Volume + High, data = train
    _data)

# Make predictions on the test data
predictions <- predict(lm_model, newdata = test_data)

mse <- mean((test_data$Adjusted.Close - predictions)^2)
[1] 0.08184498

Call:
lm(formula = Adjusted.Close ~ Low + Open + Volume + High, data =
    train_data)

Coefficients:
(Intercept)          Low          Open       Volume          High
  3.599e-01    8.828e-01   -7.177e-01    6.517e-08    8.141e-01
```

# 8 Results & Further Study

There was no discernible relationship between the results of the elections and the performance of the oil and gas sector, according to the analysis of the oil and stock market values based on those of the elections. Although it was anticipated that election results may directly affect the stock market, our statistical research failed to find a meaningful correlation between the two variables.

While our analysis did not find a correlation between election results and the oil and stock market values, further research could explore additional factors that may influence the market, such as economic policies, geopolitical events, or global market trends. Additionally, studying the market reaction to specific policy announcements or regulatory changes implemented by the elected officials could provide valuable insights into the relationship between political events and the oil and stock market.

Furthermore, expanding the study to include a larger dataset spanning multiple election cycles and incorporating data from different countries could provide a broader perspective on the potential impact of elections on the oil and stock market. Additionally, considering other sectors and industries within the stock market could reveal varying patterns of correlation or non-correlation with election outcomes.

It is important to note that the stock market is influenced by numerous complex factors, and a single variable, such as election results, may not be sufficient to explain market movements. Therefore, conducting further studies and employing more sophisticated statistical methods may provide a more comprehensive understanding of the relationship between elections and the oil and stock market values.

## Bibliography

[1] Cunado, J., Gil-Alana, L. A., & Perez de Gracia, F. (2016). *The Impact of Oil Price Volatility on Stock Markets: Evidence from G7 Countries*.

[2] Xu, Y., & Vahid, F. (2021). *Stock Market Reaction to Oil Price Shocks: A Comparative Analysis of the US and Canada*.

[3] Matta, M. P., & Wehrhahn, C. (2019). *Determinants of Stock Prices in the Oil and Gas Industry: A Panel Data Approach*.

[4] Pew Research Center. (2014). *The 2016 U.S. Presidential Election: Political Polarization and Media Habits*.

[5] Federal Election Commission. (2020). *The 2020 Election: The Official Results*.

[6] Rice, J. A. (1986). *Goodness-of-fit Techniques*.

[7] Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric Statistical Methods*.

[8] Shapiro, S. S., & Wilk, M. B. (1989). *A Goodness-of-Fit Test for the Exponential Distribution*.

[9] Neave, H. R. (2012). *Testing for Normality: A Guide for Practitioners*.

[10] Donald, S. G., & Hill, R. C. (2009). *Measures of Skewness and Kurtosis and their Applications in Econometric Models*.

[11] Ott, R. L., & Longnecker, M. T. (2015). *An Introduction to Statistical Methods and Data Analysis*.

[12] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*.