3080_final_project

Nemmo Ciccone

2025-11-02

R Markdown

Introduction

Gasoline prices are a vital economic indicator and a significant component of consumer budgets. While national average prices are widely reported, they obscure the substantial price differences and volatility experienced by consumers at a regional level. These regional variations are driven by distinct local supply chains, state and local taxes, and environmental regulations mandating specific fuel blends (e.g., reformulated vs. conventional gasoline).

This project investigates the statistical properties of U.S. gasoline prices on a regional basis. The analysis will use weekly price data from the U.S. Energy Information Administration (EIA) from 2003 to the present. The research aims to answer two primary questions:

> Center: Is there a statistically significant difference in the overall mean weekly price of regular gasoline between distinct U.S. regions (e.g., the West Coast vs. the Gulf Coast)?
>
> Spread: Is there a statistically significant difference in the price volatility (i.e., the standard deviation of weekly price changes) between U.S. regions (e.g., the Midwest vs. the East Coast)?

By answering these two statistical questions, this project aims to provide a quantitative foundation for exploring the broader empirical drivers of regional price disparities. Should statistically significant differences in mean price or volatility be found, the analysis will conclude by discussing the potential real-world implications, such as differing supply chain logistics, varying state tax burdens, regional differences in economic welfare, or distinct consumer search behaviors across these markets.

Data Summary

(a) & (b) Data Collection The data represents a sample, not a population. It is sourced from the U.S. Energy Information Administration's (EIA) Form EIA-878, "Motor Gasoline Price Survey." Every Monday, the EIA collects retail price data as of 8:00 a.m. from a representative sample of approximately 800 retail gasoline outlets across the United States.

This sample includes major national retailers, small chains, and independent owners. Data is collected via a multi-mode strategy, including telephone, email, text, fax, and web surveys. The final published figures are volume-weighted average prices for each region.

(c) Data Modifications The raw data will be modified in two ways for this analysis:Data Selection: The analysis will use the "Regular All Areas All Formulations" price series. This series is the most representative of the consumer market and is the most appropriate for making "apples-to-apples" comparisons between regions that have different fuel formulation requirements.Transformation (for Volatility): To analyze price volatility (spread) independent of long-term trends, a second dataset will be created by calculating the first difference ($Price_t - Price_{t-1}$) of the weekly price data. This transforms the dataset from "price" to "week-over-week change."

(d) Potential Issues Non-Stationarity: The raw price data is non-stationary, meaning its mean and variance change over time (e.g., prices from 2003 are not comparable to prices from 2023). This trend could skew a simple calculation of variance, which is why the first-difference transformation is necessary to analyze volatility.

Missing Data: The dataset may contain NA or blank values for certain weeks or regions, which will need to be identified and omitted.

Methodology Changes: The EIA updated its sampling and estimation methodology in May 2018. This could create a structural break in the data, though its impact on comparing parameters between regions (which are sampled concurrently) is likely minimal.

(e) Appropriateness The EIA data is ideal for this research. Its weekly frequency, long historical time-span, and clear regional segmentation (PADDs) provide robust, high-integrity data to calculate and compare the long-term parameters of center (mean) and spread (variance)

Exploratory analysis

```
library(readxl)
library(tidyverse)
```

```
# List all sheets in the Excel file
excel_sheets("eia_gas_data.xls")
```

```
##  [1] "Contents" "Data 1"   "Data 2"   "Data 3"   "Data 4"   "Data 5"
##  [7] "Data 6"   "Data 7"   "Data 8"   "Data 9"   "Data 10"  "Data 11"
## [13] "Data 12"
```

```
# Read a specific sheet, we will use "Data 3", which is Regular All Areas All Formulations
#first 2 rows are headers to skip
reg_gas<-read_excel("eia_gas_data.xls", sheet = "Data 3", skip = 2)
#colnames(reg_gas)
```

```
# Remove initial rows with NA values
reg_gas_regions<-reg_gas_regions[-(1:511), ]
head(reg_gas_regions)
```

```
## # A tibble: 6 x 8
##   Date                ‘New England‘ ‘Central Atlantic‘ ‘Lower Atlantic‘ Midwest
##   <dttm>                      <dbl>              <dbl>            <dbl>   <dbl>
## 1 2000-06-05 00:00:00          1.59               1.56             1.46    1.65
## 2 2000-06-12 00:00:00          1.62               1.59             1.49    1.80
```

```
## 3 2000-06-19 00:00:00          1.64               1.62             1.54    1.87
## 4 2000-06-26 00:00:00          1.66               1.64             1.54    1.79
## 5 2000-07-03 00:00:00          1.67               1.65             1.54    1.68
## 6 2000-07-10 00:00:00          1.65               1.65             1.53    1.59
## # i 3 more variables: `Gulf Coast` <dbl>, `Rocky Mountain` <dbl>,
## #   `West Coast` <dbl>
```

```r
# Check for missing values in the dataset
sapply(reg_gas_regions, function(x) sum(is.na(x)))
```

```
##             Date      New England Central Atlantic   Lower Atlantic
##                0                0                0                0
##          Midwest       Gulf Coast   Rocky Mountain       West Coast
##                0                0                0                0
```

```r
# Transform data to long format for easier analysis
reg_gas_long<- reg_gas_regions %>%
    pivot_longer(cols = -Date, names_to = "Region", values_to = "Price")

# Calculate week-over-week price changes for volatility analysis
reg_gas_logs <- reg_gas_long %>%
  group_by(Region) %>%
    mutate(
    Log_Price = log(Price),
    Log_Change = Log_Price - lag(Log_Price) # This is the % change
   ) %>%
   ungroup()

# Calculate each region's deviation from all regions' weekly average price
regional_diffs<- reg_gas_long %>%
    group_by(Date) %>%
    mutate(Avg_Price_All_Regions = mean(Price),
           Deviation_From_Avg = Price - Avg_Price_All_Regions) %>%
    ungroup()
```

```r
# Numerical Summaries for Prices by Region
summary_prices <-na.omit(reg_gas_long) %>%
    group_by(Region) %>%
    summarize(
        Mean_Price = mean(Price),
        Median_Price = median(Price, na.rm = ),
        SD_Price = sd(Price),
        Min = min(Price),
        Max = max(Price)
    )
print(summary_prices)
```

Summary Statistics

```
## # A tibble: 7 x 6
```

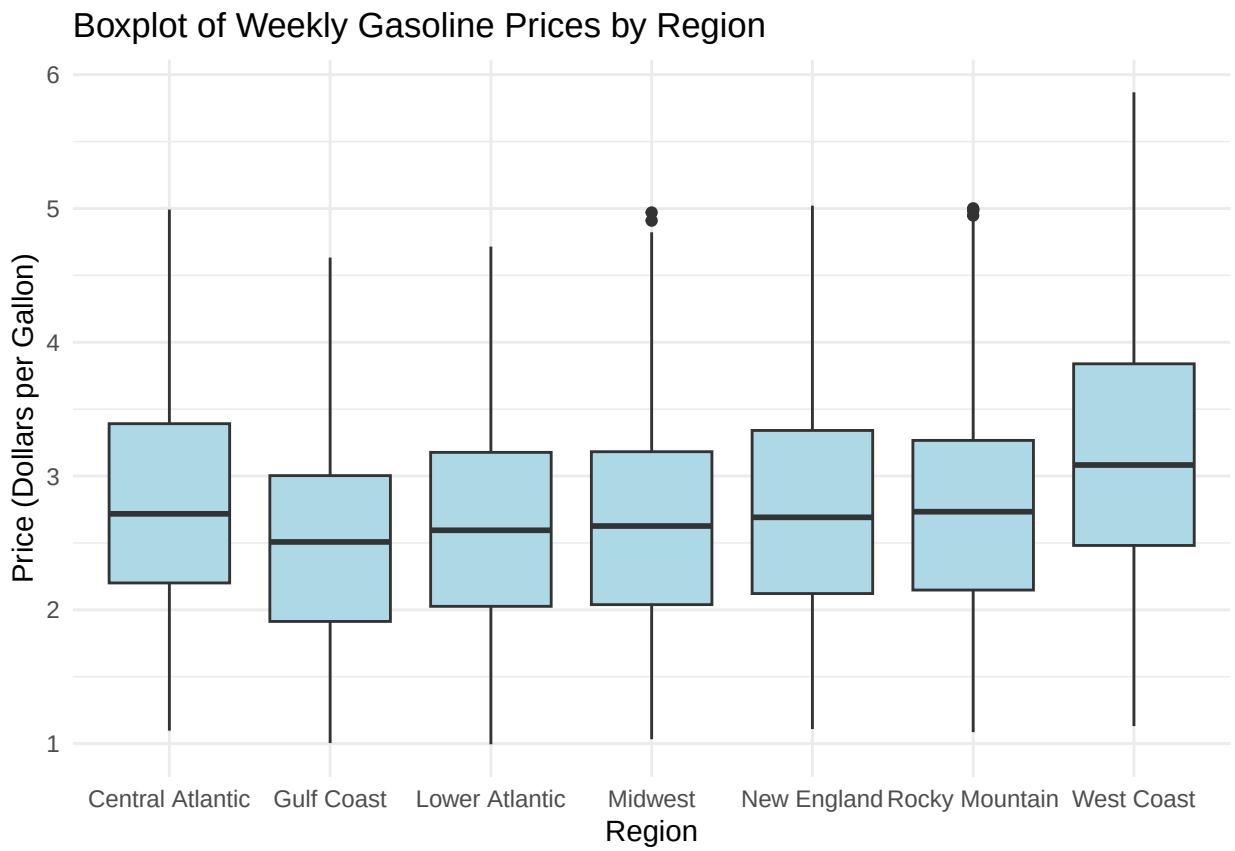
```
##   Region           Mean_Price Median_Price SD_Price   Min   Max
##   <chr>                 <dbl>        <dbl>    <dbl> <dbl> <dbl>
## 1 Central Atlantic       2.72         2.72    0.771 1.10   4.99
## 2 Gulf Coast             2.46         2.51    0.714 1.00   4.63
## 3 Lower Atlantic         2.56         2.59    0.745 0.995  4.72
## 4 Midwest                2.60         2.63    0.757 1.03   4.97
## 5 New England            2.69         2.69    0.770 1.11   5.02
## 6 Rocky Mountain         2.67         2.73    0.767 1.09   5.00
## 7 West Coast             3.11         3.08    0.940 1.13   5.87
```

```r
# Numerical Summaries for Price Changes by Region
summary_price_changes <-reg_gas_logs %>%
    group_by(Region) %>%
    summarize(
        Mean_Change = mean(Log_Change, na.rm = TRUE),
        Median_Change = median(Log_Change, na.rm = TRUE),
        SD_Change = sd(Log_Change, na.rm = TRUE),
        Min_Change = min(Log_Change, na.rm = TRUE),
        Max_Change = max(Log_Change, na.rm = TRUE)
    )
print(summary_price_changes)
```

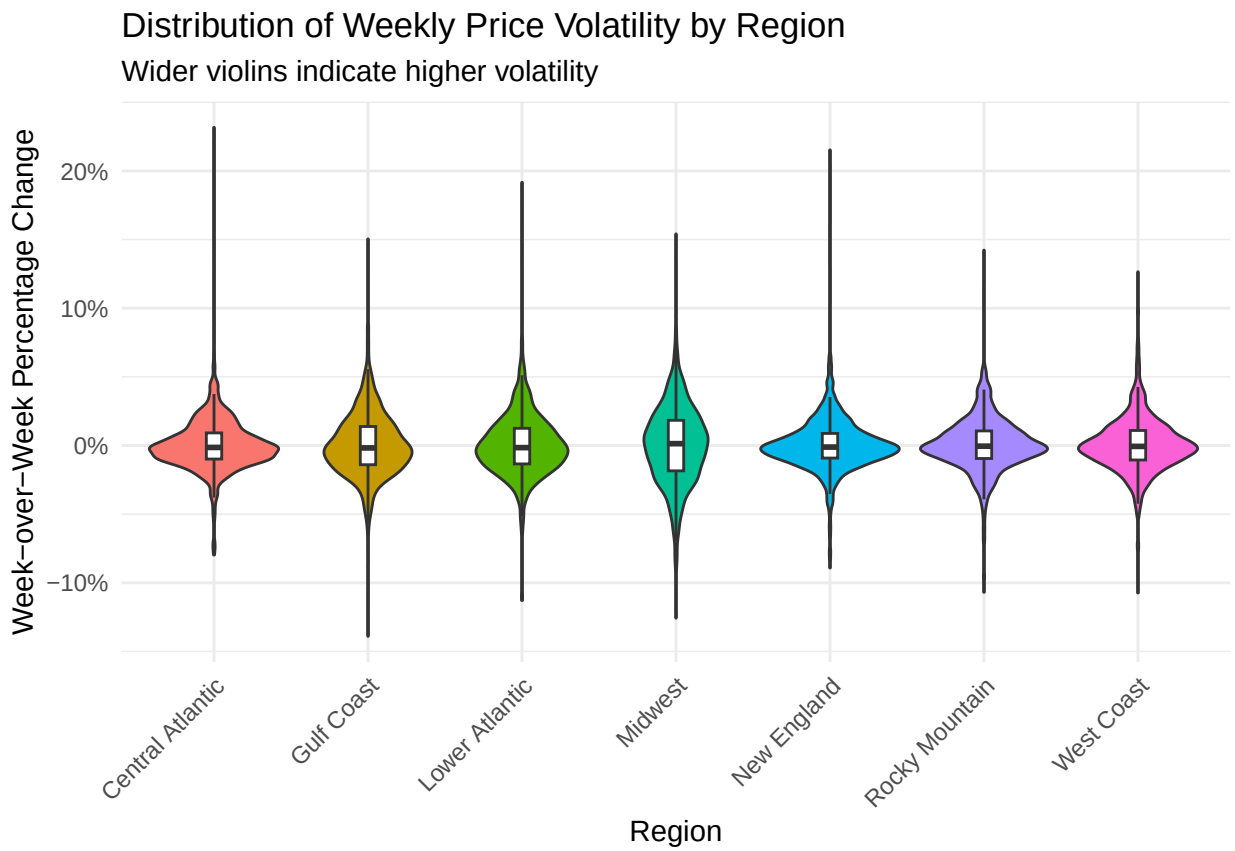
```
## # A tibble: 7 x 6
##   Region           Mean_Change Median_Change SD_Change Min_Change Max_Change
##   <chr>                  <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 Central Atlantic    0.000510     -0.00140     0.0195    -0.0797      0.232
## 2 Gulf Coast          0.000427     -0.00177     0.0240    -0.139       0.150
## 3 Lower Atlantic      0.000493     -0.00155     0.0236    -0.113       0.192
## 4 Midwest             0.000420      0.00134     0.0296    -0.126       0.154
## 5 New England         0.000462     -0.00113     0.0199    -0.0892      0.215
## 6 Rocky Mountain      0.000506     -0.000501    0.0204    -0.107       0.142
## 7 West Coast          0.000718     -0.000654    0.0206    -0.107       0.127
```

Graphical Summaries

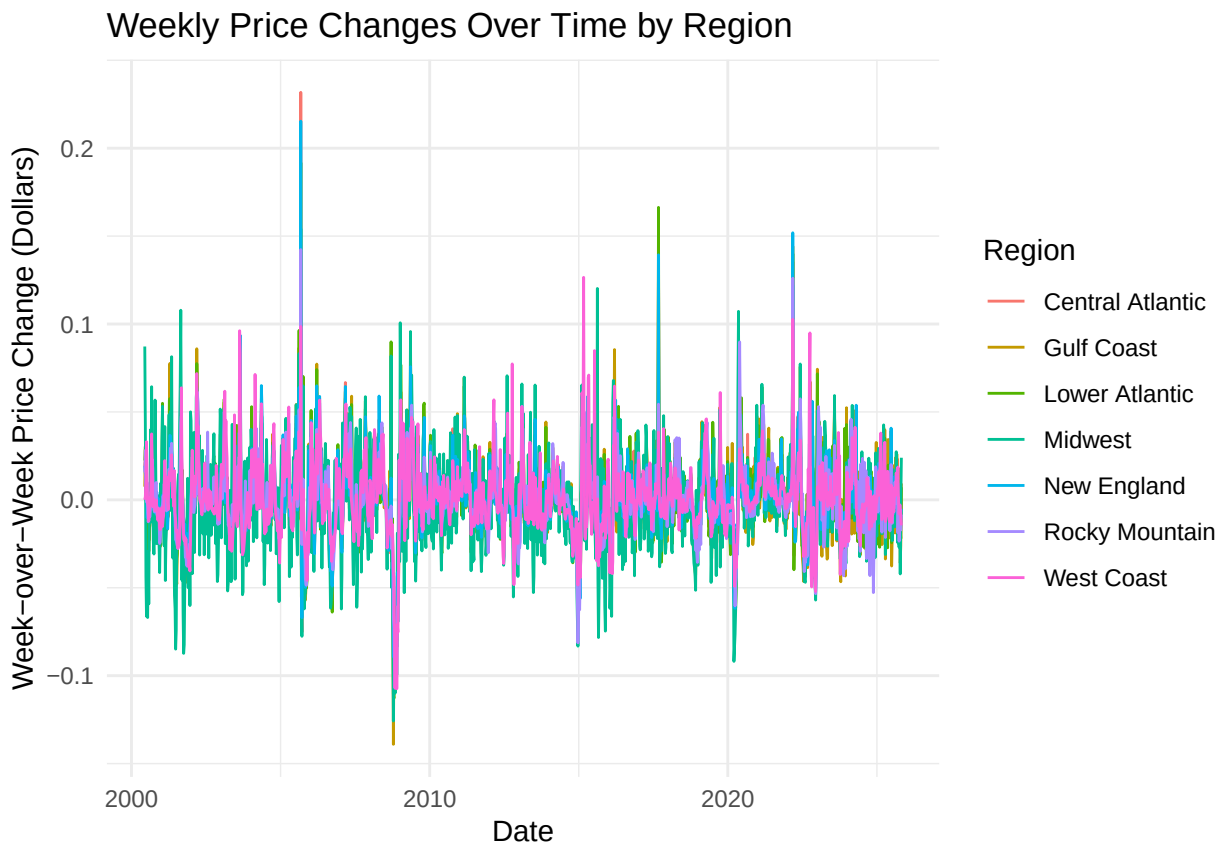
```r
# Boxplot of Prices by Region
ggplot(reg_gas_long, aes(x = Region, y = Price)) +
    geom_boxplot(fill = "lightblue") +
    labs(title = "Boxplot of Weekly Gasoline Prices by Region",
         x = "Region",
         y = "Price (Dollars per Gallon)") +
    theme_minimal()
```



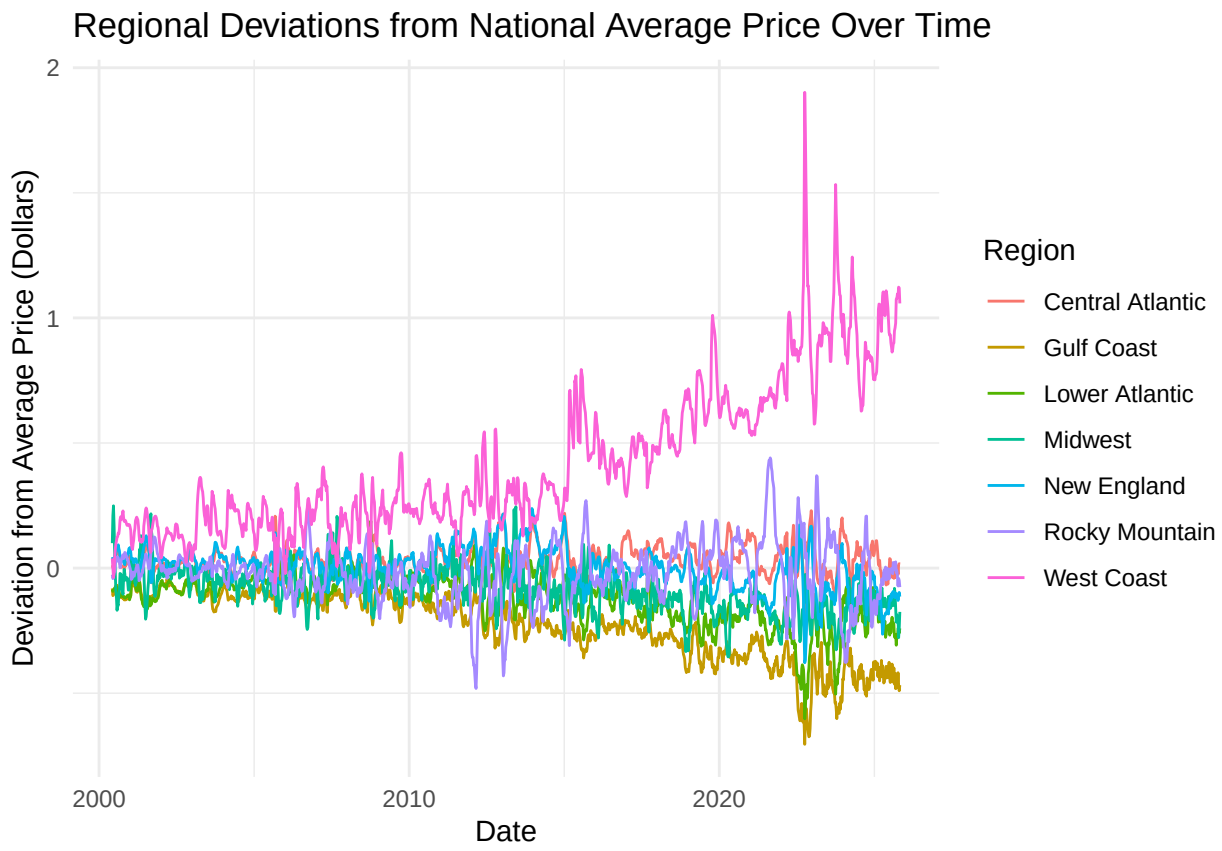

```r
# Create a Violin Plot
ggplot(na.omit(reg_gas_logs), aes(x = Region, y = Log_Change, fill = Region)) +
  geom_violin() +

  # Add a small boxplot inside for more detail (optional but nice)
  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) +

  scale_y_continuous(labels = scales::percent) + # Format Y-axis as %
  theme_minimal() +
  labs(
    title = "Distribution of Weekly Price Volatility by Region",
    subtitle = "Wider violins indicate higher volatility",
    x = "Region",
    y = "Week-over-Week Percentage Change"
  ) +
  theme(legend.position = "none", # Remove legend (redundant)
        axis.text.x = element_text(angle = 45, hjust = 1)) # Angle labels
```




```
ggplot(na.omit(reg_gas_logs), aes(x = Date, y = Log_Change, color = Region)) +
  geom_line() +
  labs(
    title = "Weekly Price Changes Over Time by Region",
    x = "Date",
    y = "Week-over-Week Price Change (Dollars)",
    color = "Region"
  ) +
  theme_minimal()
```




```r
ggplot(regional_diffs, aes(x = Date, y = Deviation_From_Avg, color = Region)) +
  geom_line() +
  labs(
    title = "Regional Deviations from National Average Price Over Time",
    x = "Date",
    y = "Deviation from Average Price (Dollars)",
    color = "Region"
  ) +
  theme_minimal()
```



Conclusions

References

- Baghestani, Hamid, and Bley, Jorg (2020). *Do Directional Predictions of US Gasoline Prices Reveal Asymmetries?* Journal of Economics and Finance, 44, 348–360.

- Tappata, Mariano. (2009). *Rockets and Feathers: Understanding Asymmetric Pricing.* RAND Journal of Economics, 40(4), 673–687.

- EIA Gasoline Data
- EIA Methods
- AAA Gas Prices
- EIA Today in Energy