3080 Project Part 1: Regional Gasoline Price Analysis

Nemmo Ciccone

2025-11-02

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 June 5, 2000 through October 27, 2025. 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

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 comparisons between regions that have different fuel formulation requirements.

Transformation: To analyze price volatility (spread) independent of long-term trends, a second dataset will be created by calculating the log first difference. This data may be referred to as weekly log changes or simply weekly price changes. It is calculated as:

$\text{Log}(Price_t - Price_{t-1})$ of the weekly price data.

(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 and spread across U.S. regions.

Exploratory analysis

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

Data Import and Cleaning

```
# 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 in "Data 3" sheet, which is Regular All Areas All Formulations; remove headers
reg_gas<-read_excel("eia_gas_data.xls", sheet = "Data 3", skip = 2)
```

```r
# 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()
```

Summary Statistics

```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),
        SD_Price = sd(Price),
        Min = min(Price),
        Max = max(Price)
    )
print(summary_prices)
```

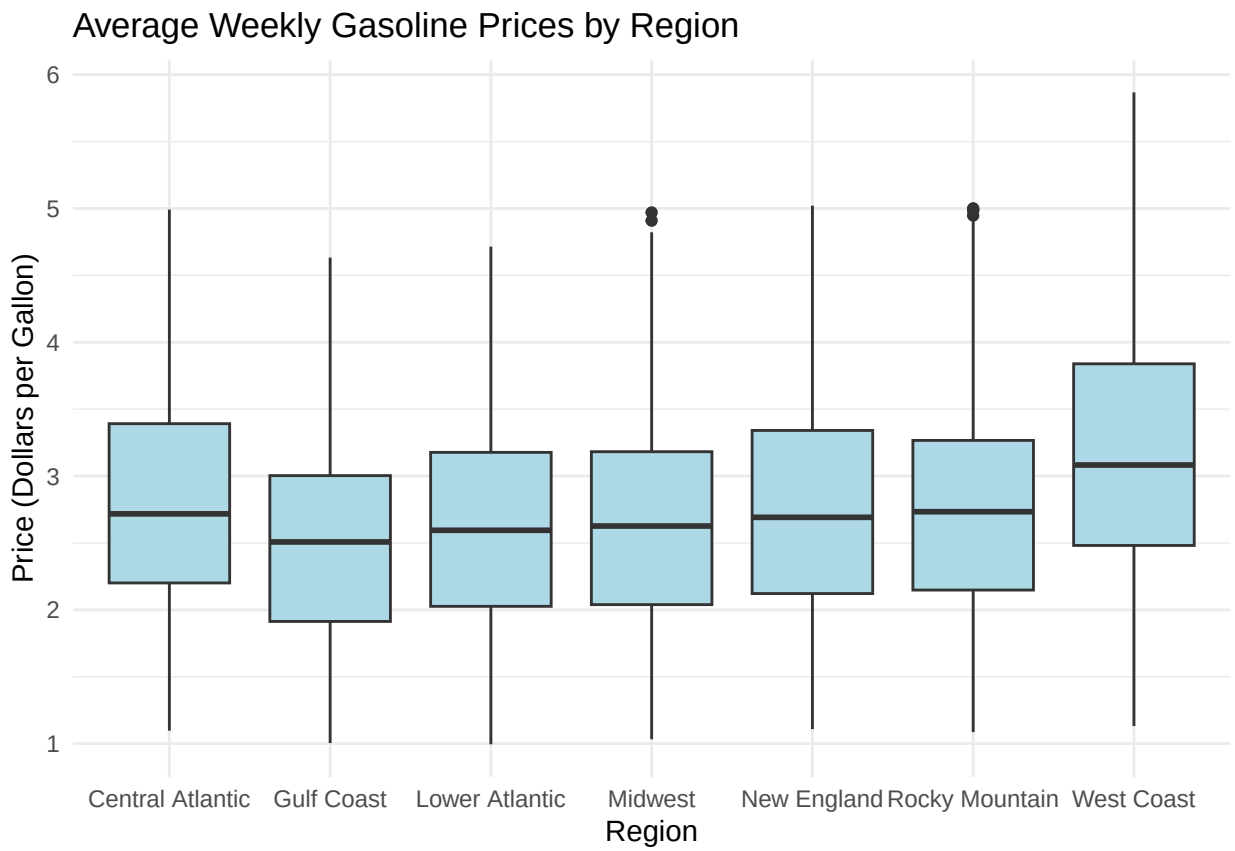
```
## # A tibble: 7 x 6
##   Region           Mean_Price Median_Price SD_Price   Min   Max
##   <chr>                 <dbl>        <dbl>    <dbl> <dbl> <dbl>
## 1 Central Atlantic       2.72         2.72    0.771 1.10   4.99
## 2 Gulf Coast             2.46         2.51    0.714 1.00   4.63
## 3 Lower 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
```

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

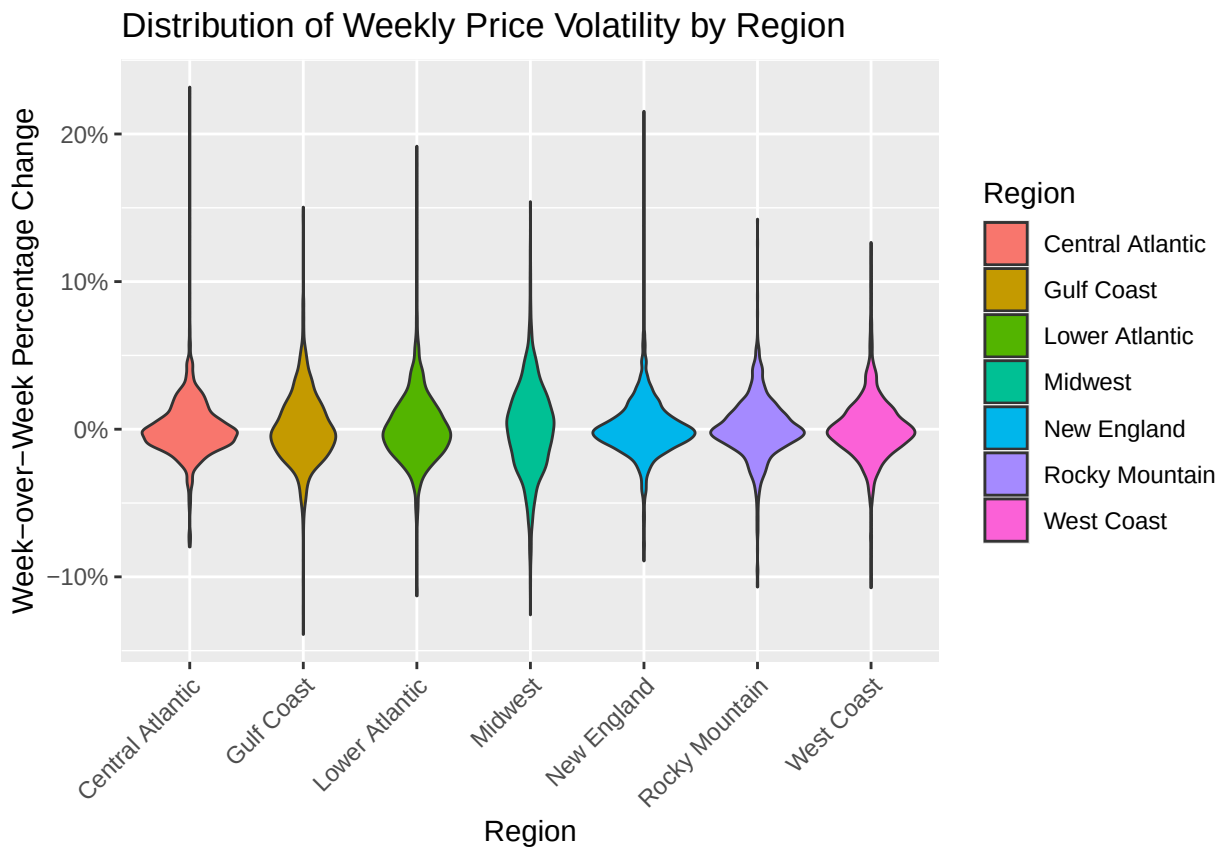
```
## # A tibble: 7 x 6
##   Region           Mean_Change Median_Change SD_Change Min_Change Max_Change
##   <chr>                  <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 Central Atlantic    0.000510     -0.00140     0.0195    -0.0797      0.232
## 2 Gulf Coast          0.000427     -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

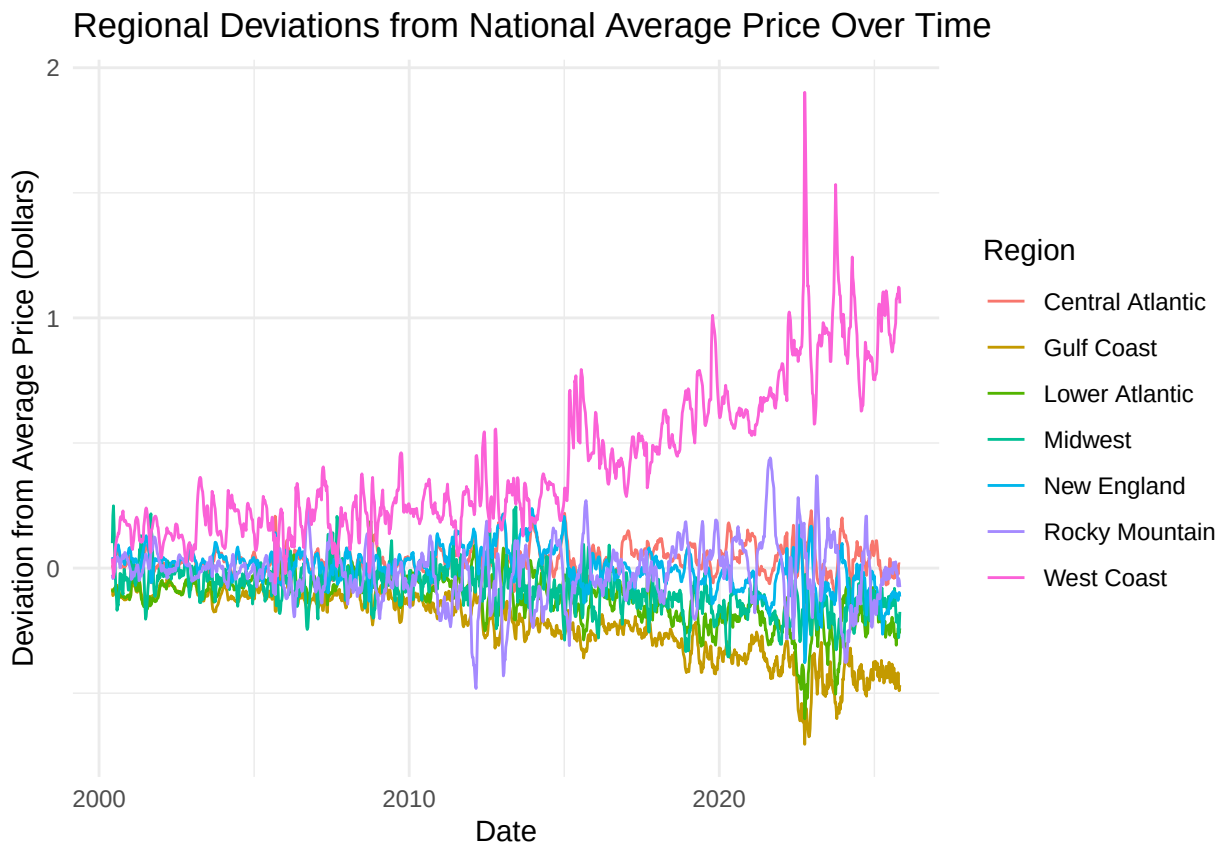
```
# Boxplot of Prices by Region
ggplot(reg_gas_long, aes(x = Region, y = Price)) +
    geom_boxplot(fill = "lightblue") +
    labs(title = "Average Weekly Gasoline Prices by Region",
         x = "Region",
         y = "Price (Dollars per Gallon)") +
    theme_minimal()
```




```r
#Violin Plot
ggplot(na.omit(reg_gas_logs), aes(x = Region, y = Log_Change, fill = Region)) +
  geom_violin() +
  scale_y_continuous(labels = scales::percent) + # Format Y-axis as %
  labs(
    title = "Distribution of Weekly Price Volatility by Region",
    x = "Region",
    y = "Week-over-Week Percentage Change"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Angle labels
```




```r
# Line Plot of Regional Deviations from National Average Over Time
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

Looking first at the summary statistics for mean regular gas prices from June 2000 to October 2025, there are clear differences between regions. The West Coast has the highest average, while the Gulf Coast has the lowest. This aligns with known factors such as higher state taxes and stricter environmental regulations on the West Coast. This difference in mean prices is visually supported by the boxplot, which shows the West Coast's prices are consistently higher than other regions which cluster lower.

Addressing the volatility of prices, the summary statistics for week-over-week log changes indicate that the Midwest has the highest standard deviation (0.0296), suggesting it experiences the most price volatility. The violin plot further illustrates this, showing a wider distribution of price changes in the Midwest compared to other regions. The regions with the lowest standard deviation in price changes are Central Atlantic (0.0195) and New England (0.0199), indicating more stable prices.

The final graph showing regional deviations from the national average over time highlights persistent patterns. The West Coast consistently deviates positively from the average, and this deviation appears to have increased over time as the distance between the West Coast and other regions has widened. Another takeaway from this graph is the apparent spikes in observed gas prices particular only to certain regions (such as Rocky Mountain spike in 2022, California in 2023). These spikes may be attributable to localized supply disruptions, refinery issues, or regional policy changes affecting fuel prices.

In conclusion, the analysis implies differences in both the mean and volatility of weekly regular gasoline prices across U.S. regions. The West Coast's higher average prices and the Midwest's greater volatility suggest that regional factors such as supply chain logistics, tax policies, and consumer behavior play crucial roles in gasoline market dynamics. Further research could explore these underlying causes in more detail, potentially informing policy decisions aimed at stabilizing prices and reducing regional disparities.

References

AAA Gas Prices. American Automobile Association. https://gasprices.aaa.com/

Baghestani, Hamid, and Jörg Bley. "Do Directional Predictions of US Gasoline Prices Reveal Asymmetries?" *Journal of Economics and Finance*, vol. 44, 2020, pp. 348–360.

"Gasoline and Diesel Fuel Update: Processing Methods." *U.S. Energy Information Administration (EIA).* https://www.eia.gov/petroleum/gasdiesel/gas_proc-methods.php

"Gasoline and Diesel Fuel Update." *U.S. Energy Information Administration (EIA).* https://www.eia.gov/petroleum/gasdiesel/

Tappata, Mariano. "Rockets and Feathers: Understanding Asymmetric Pricing." *RAND Journal of Economics*, vol. 40, no. 4, 2009, pp. 673–687.

"Today in Energy." *U.S. Energy Information Administration (EIA).* https://www.eia.gov/todayinenergy/detail.php?id=65184