Determinants of NY State Gas Prices

Group Name: The X-Ponents

Group Members: Aaron Medley (Deadpool), Patrick Sicurello (Wolverine), Jennifer Zhuang (Psylocke)

Professor Z, Theory of Machine Learning 625.742

Fall 2024

Overview

Introduction

Data

Results & Analysis

Conclusions & Further Research

References

Introduction

Goal: Understand the factors that effect gas prices in New York State and use these to predict future prices.

Understanding these drivers has implications for economic analysis, transportation policy, and understanding consumer behavior.

We will analyze data for multiple cities throughout NY State.

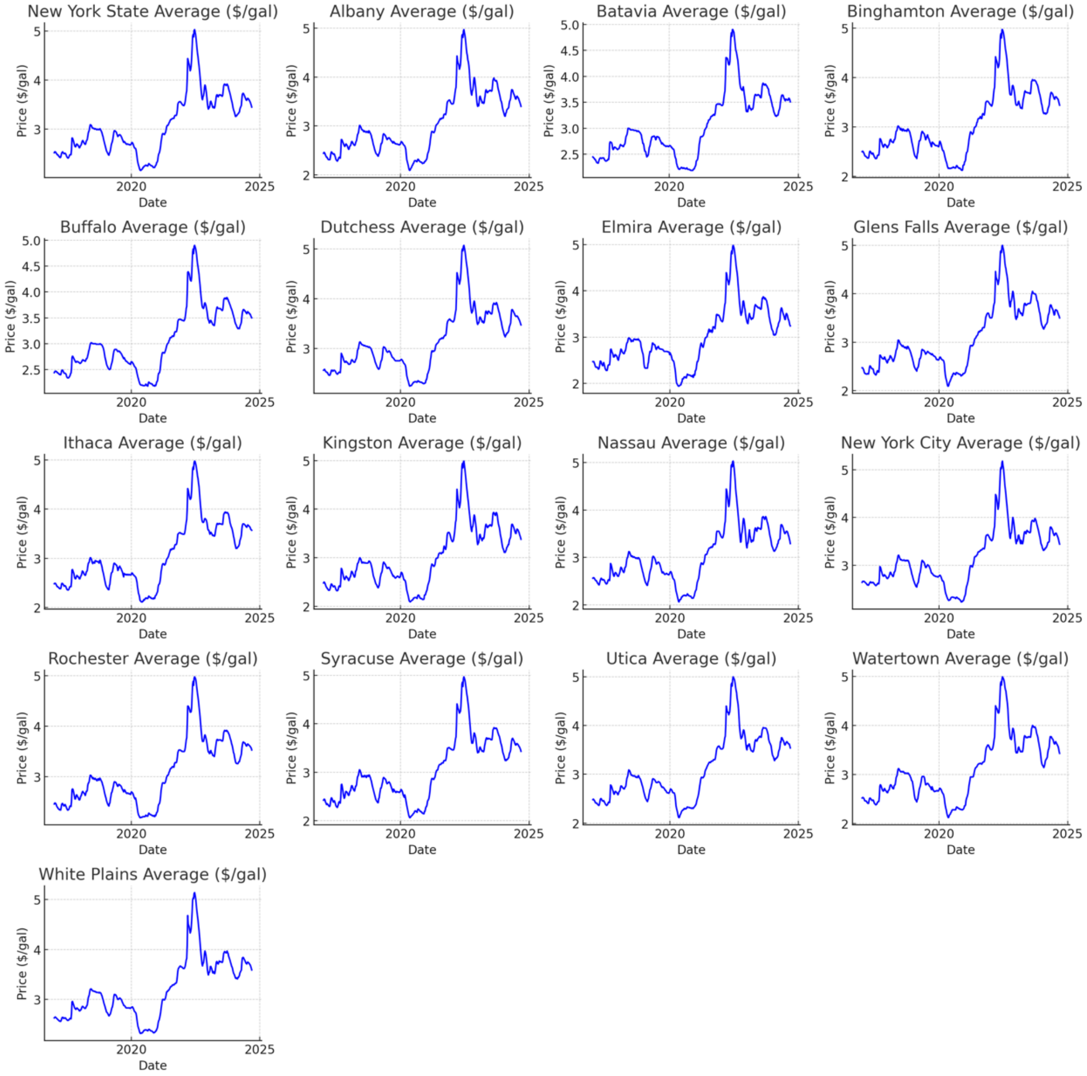
All data from data.ny.gov.

Data - Overview

- NY Government Database (data.ny.gov).
    - Date range: Jan 2017 – September 2024. Data are weekly.
    - Average gasoline prices (dollars per gallon) from Albany, Batavia, Binghamton, Buffalo, Dutchess, Elmira, Glens Falls, Ithaca, Kingston, Nassau, NYC, Rochester, Syracuse, Utica, Watertown, White Plains, and NY State Average.
        - Also have average diesel prices (dollars per gallon) from same regions.
    - Supply Data
        - US
            - Jet fuel and gasoline production (thousands of barrels per day),
            - Crude oil and gasoline stocks (thousand barrels).
        - East Coast:
            - Jet fuel and gasoline production (thousands of barrels per day).
            - Crude oil, ethanol, and gasoline stocks (thousand barrels).
        - Mid Atlantic:
            - Ultra low sulfur diesel and gasoline stocks (thousand barrels).
    - Demand Data
        - US gasoline demand (thousand barrels per day).
    - Spot prices
        - NY conventional gasoline, NY ultra-low sulfur diesel, WTI and Brent crude oil.

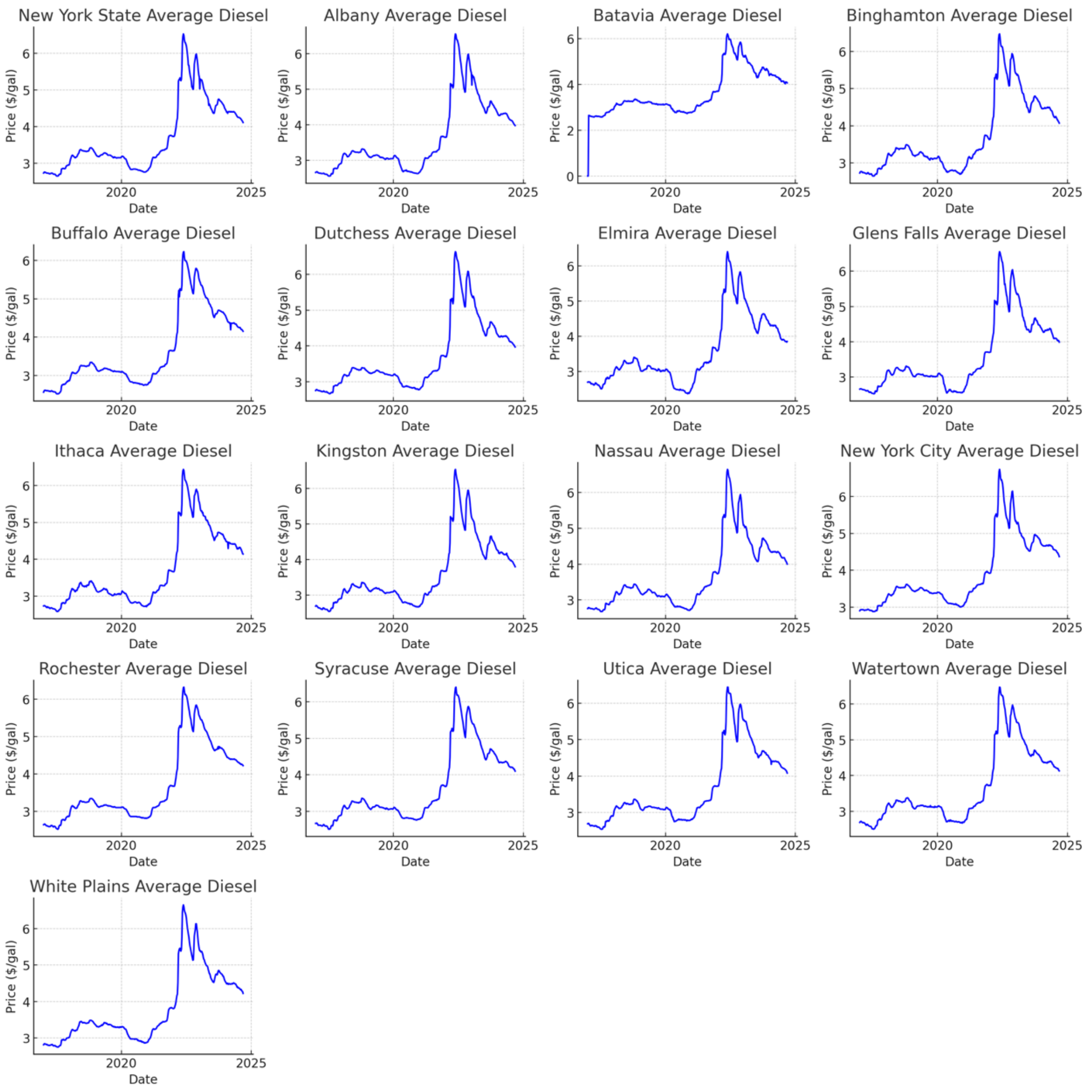
Data:
NY Gas Prices

- Series all appear to be highly correlated.
- Linear regressions should likely avoid using these variables as features, since our response variable is NY State Average.
    - Could possibly lead to issues with multicollinearity.



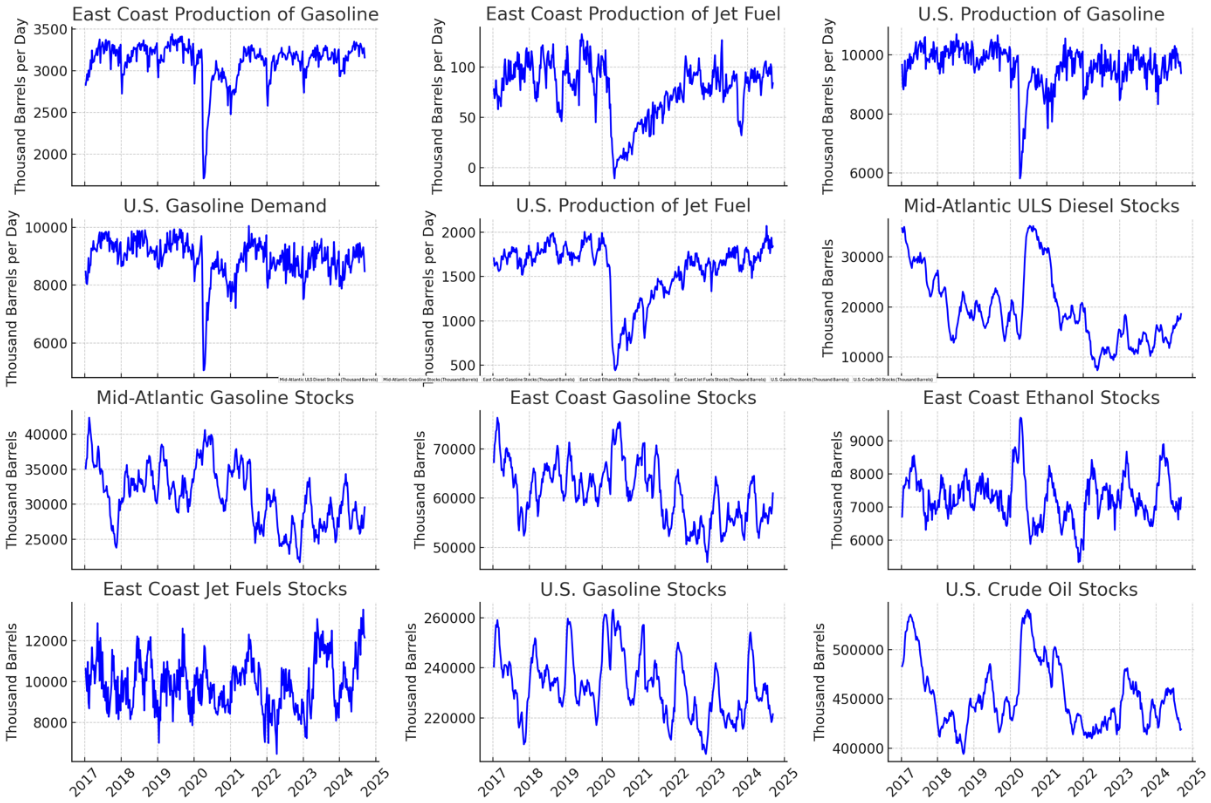
Data:
NY Diesel Prices

- Series all appear to be highly correlated.
- Plots have similar shape as the gasoline price data, indicating possibly high correlation between this data and our response variable.
- Will not be used in linear regression models.



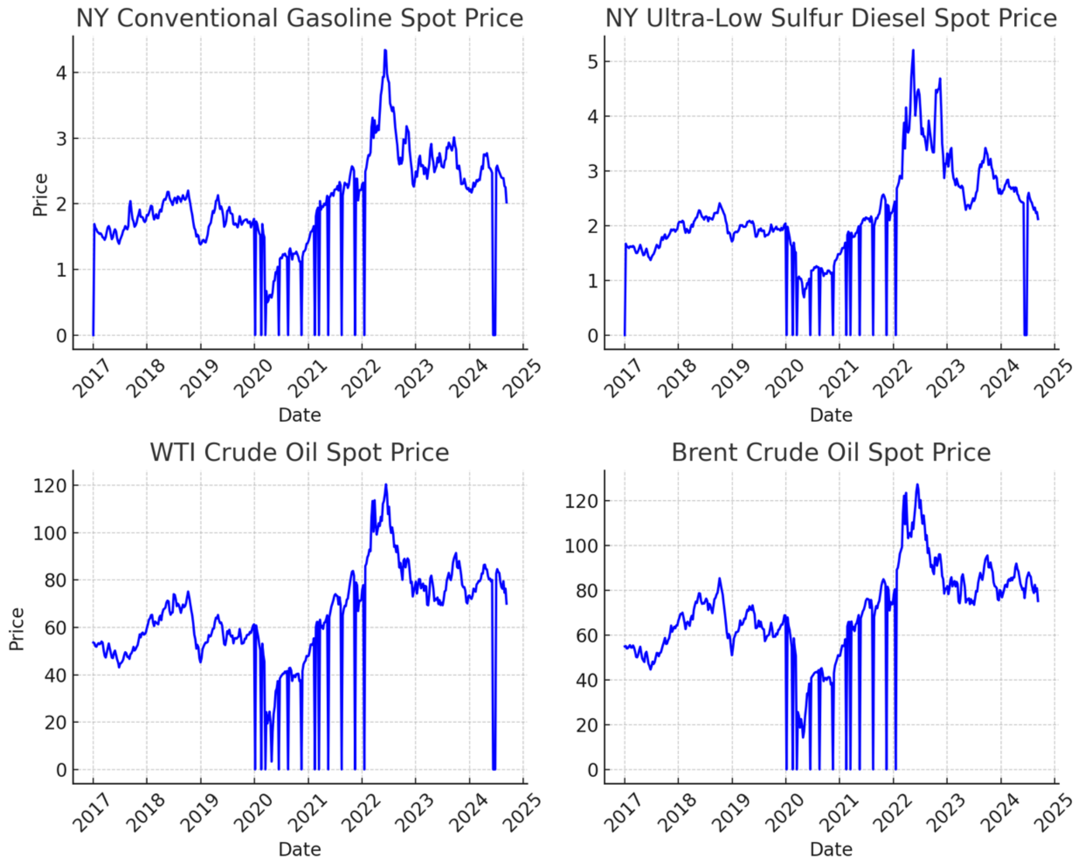
Data:
Supply & Demand

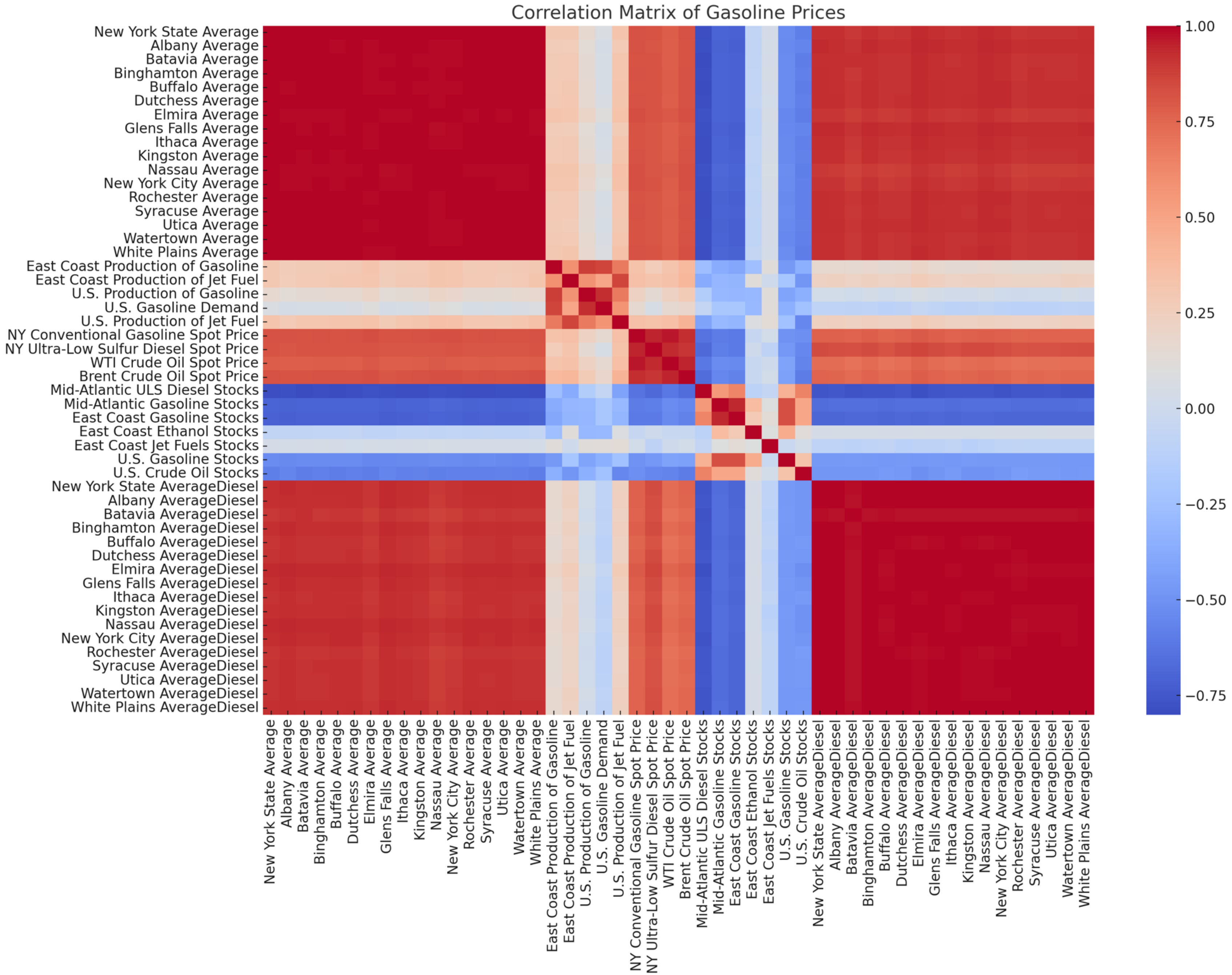
- Data does not appear as correlated with response variable.
- Good candidates for independent variables in linear regression.



Data:
Spot Prices

- Data appear correlated (and similar to the NY gasoline and diesel data).
- Missing data is causing the plots to fall to zero. This will be remedied before any analysis by forward filling prior data.

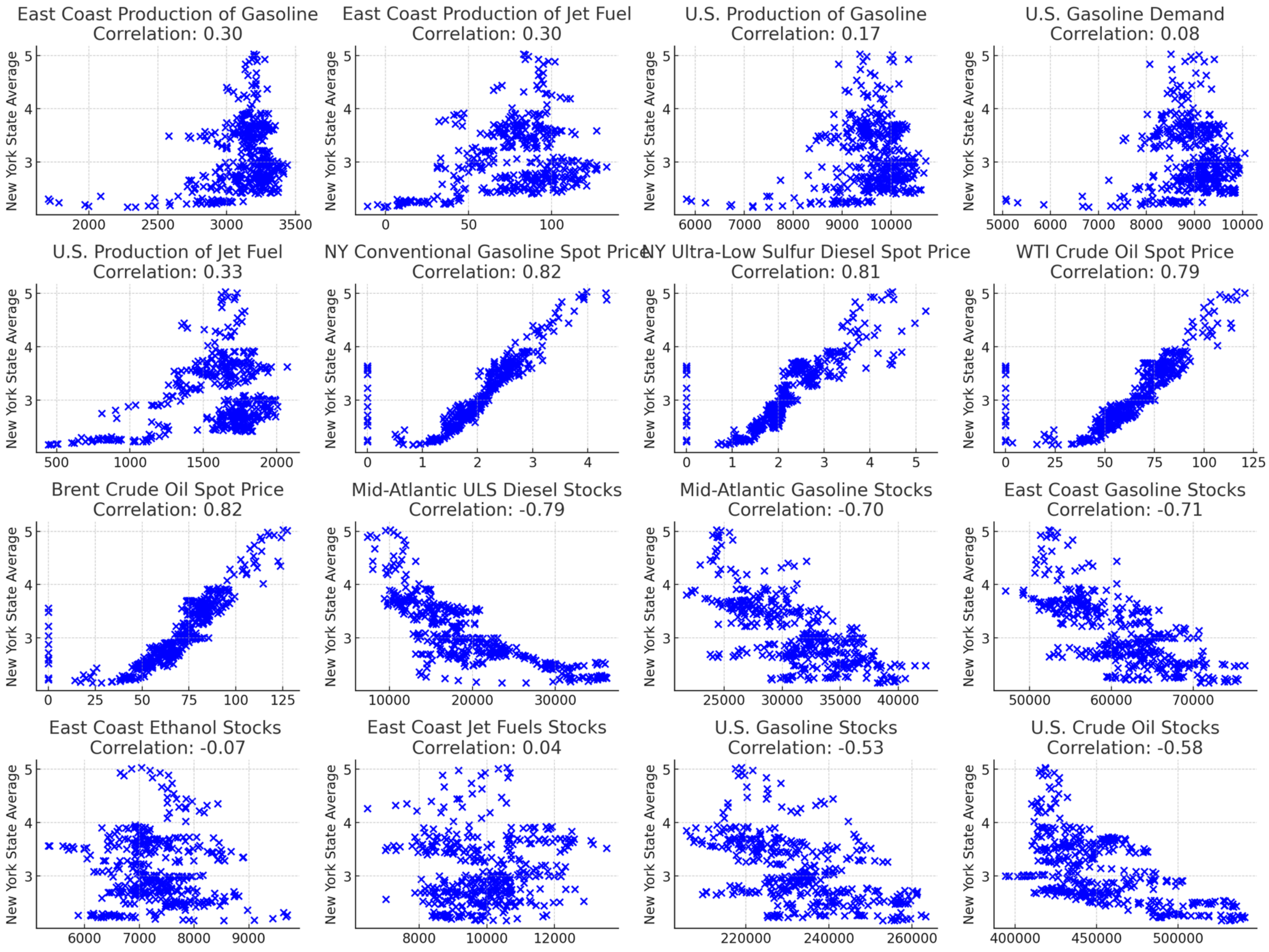



Correlation Matrix of Gasoline Prices

Data: Correlation Matrix

- As expected, the gasoline and diesel data are highly correlated.
- Negative correlation between Stocks (supply) and price of gasoline and diesel.
- Weak correlation between US gasoline demand and gasoline prices.

Data: Scatter Plots



East Coast Production of Gasoline
Correlation: 0.30

East Coast Production of Jet Fuel
Correlation: 0.30

U.S. Production of Gasoline
Correlation: 0.17

U.S. Gasoline Demand
Correlation: 0.08

U.S. Production of Jet Fuel
Correlation: 0.33

NY Conventional Gasoline Spot Price
Correlation: 0.82

NY Ultra-Low Sulfur Diesel Spot Price
Correlation: 0.81

WTI Crude Oil Spot Price
Correlation: 0.79

Brent Crude Oil Spot Price
Correlation: 0.82

Mid-Atlantic ULS Diesel Stocks
Correlation: -0.79

Mid-Atlantic Gasoline Stocks
Correlation: -0.70

East Coast Gasoline Stocks
Correlation: -0.71

East Coast Ethanol Stocks
Correlation: -0.07

East Coast Jet Fuels Stocks
Correlation: 0.04

U.S. Gasoline Stocks
Correlation: -0.53

U.S. Crude Oil Stocks
Correlation: -0.58

- Plot of response variable (New York State Average) versus various regressor variables with correlations.
- Did not include other NY Average gasoline or diesel data due to high correlation.

Results & Analysis:

Linear Regression
R^2 = 0.83
MSE = 0.0741
F-statistic = 117.54 (P-value = 0.000)

NY State Average = 1.9143***

+ 0.000355 * East Coast Production of Gasoline**

- 0.002672 * East Coast Production of Jet Fuel***

- 0.000072 * U.S. Production of Gasoline

- 0.000059 * U.S. Gasoline Demand

+ 0.000014 * U.S. Production of Jet Fuel

+ 0.696472 * NY Conventional Gasoline Spot Price***

- 0.006747 * NY Ultra-Low Sulfur Diesel Spot Price

- 0.021464 * WTI Crude Oil Spot Price***

+ 0.011432 * Brent Crude Oil Spot Price***

- 0.000030 * Mid-Atlantic ULS Diesel Stocks***

- 0.000049 * Mid-Atlantic Gasoline Stocks***

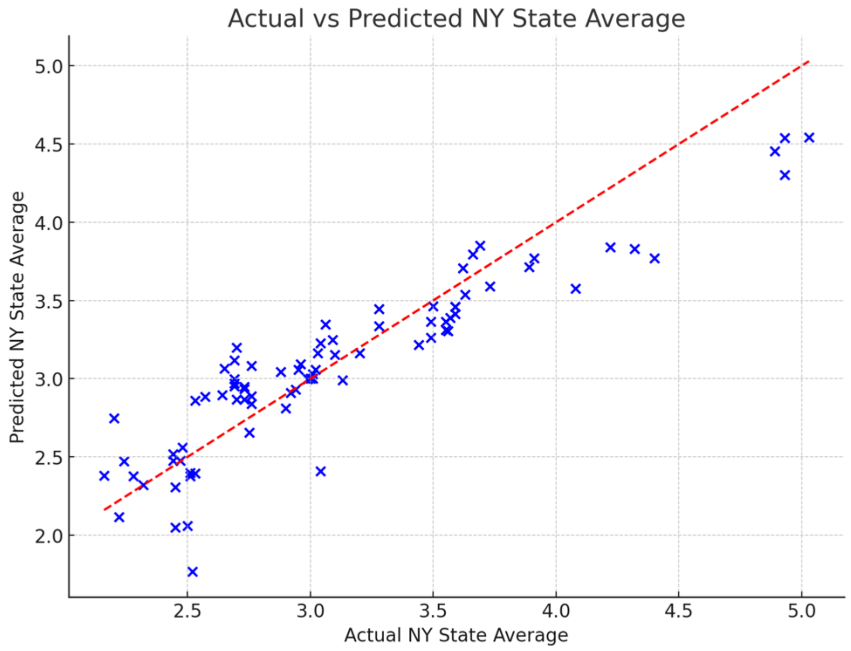
+ 0.000008 * East Coast Gasoline Stocks

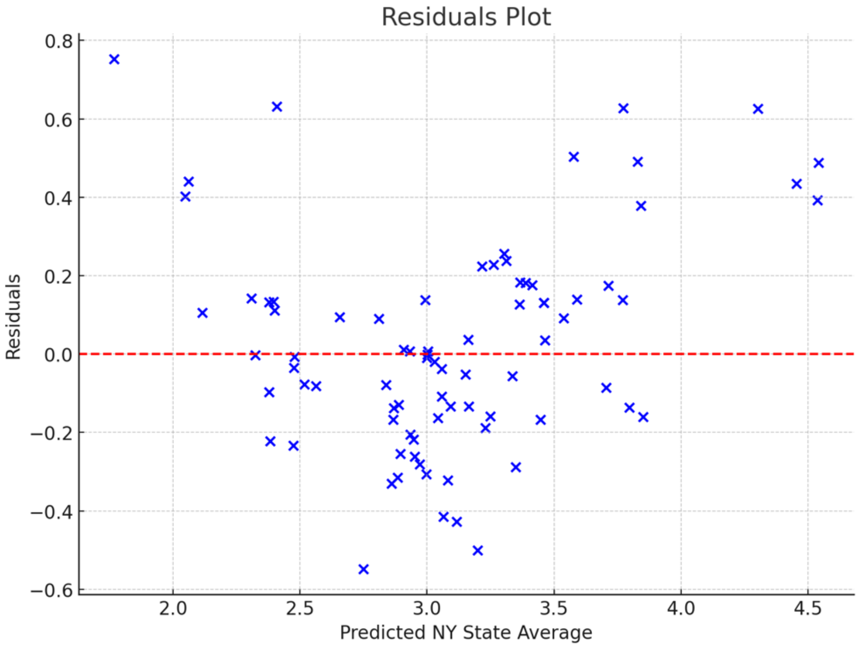
- 0.000011 * East Coast Ethanol Stocks

+ 0.000013 * East Coast Jet Fuels Stocks

+ 0.000006 * U.S. Gasoline Stocks***

- 0.000002 * U.S. Crude Oil Stocks**





Results & Analysis: Variance Inflation Factor (VIF) Test

Value greater than 10 indicates high multicollinearity

East Coast Production of Gasoline: 5.94

East Coast Production of Jet Fuel: 4.50

U.S. Production of Gasoline: 11.51

U.S. Gasoline Demand: 9.36

U.S. Production of Jet Fuel: 6.52

NY Conventional Gasoline Spot Price: 45.57

NY Ultra-Low Sulfur Diesel Spot Price: 15.43

WTI Crude Oil Spot Price: 51.15

Brent Crude Oil Spot Price: 15.37

Mid-Atlantic ULS Diesel Stocks: 3.93

Mid-Atlantic Gasoline Stocks: 11.14

East Coast Gasoline Stocks: 13.11

East Coast Ethanol Stocks: 2.20

East Coast Jet Fuels Stocks: 1.64

U.S. Gasoline Stocks: 5.20

U.S. Crude Oil Stocks: 3.05

Results & Analysis:

Linear Regression

$R^2 = 0.64$
MSE = 0.158
F-statistic = 82.51 (P-value = 0.000)

NY State Average = 7.607***

+ 0.000660* East Coast Production of Gasoline***

+ 0.000863 * East Coast Production of Jet Fuel

- 0.000224 * U.S. Gasoline Demand***

- 0.000221 * U.S. Production of Jet Fuel

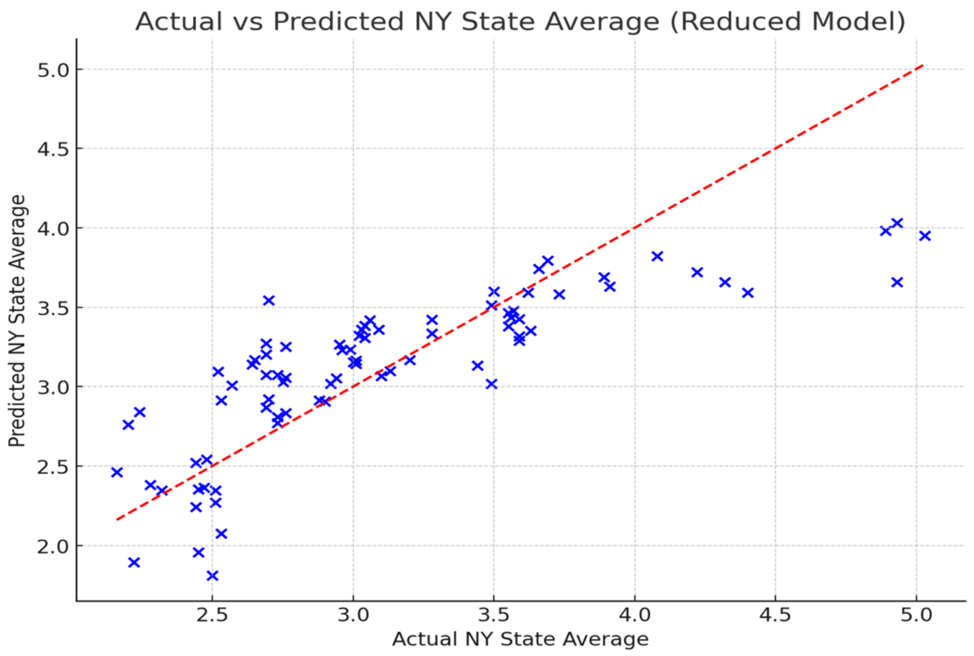
- 0.000051 * Mid-Atlantic ULS Diesel Stocks***

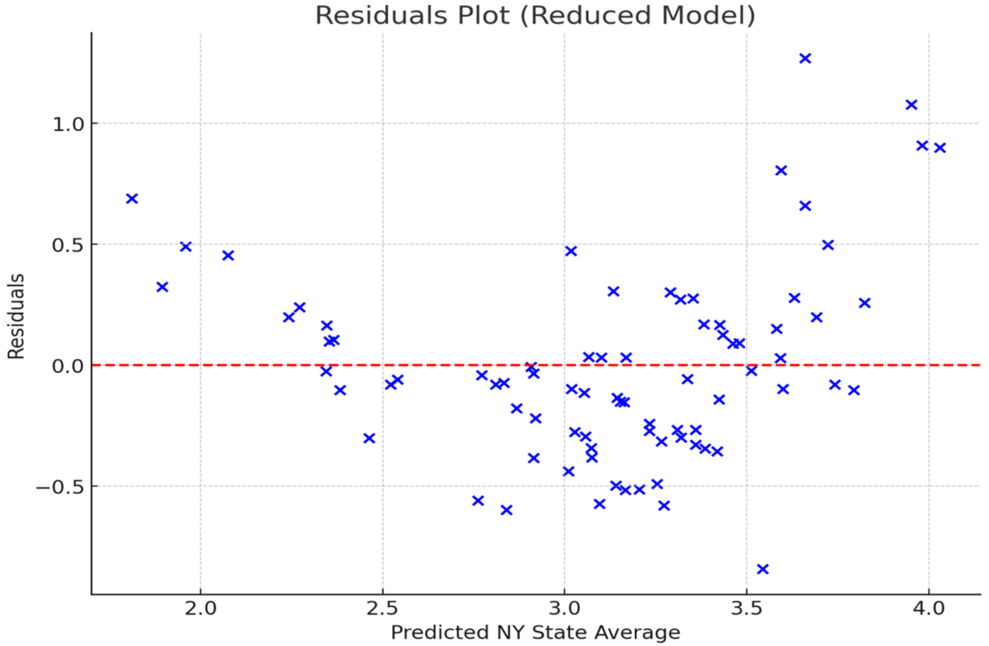
- 0.000025 * East Coast Ethanol Stocks

+ 0.000020 * East Coast Jet Fuels Stocks

- 0.000009 * U.S. Gasoline Stocks***

- 0.000003 * U.S. Crude Oil Stocks**





Results & Analysis:

Stepwise Regression

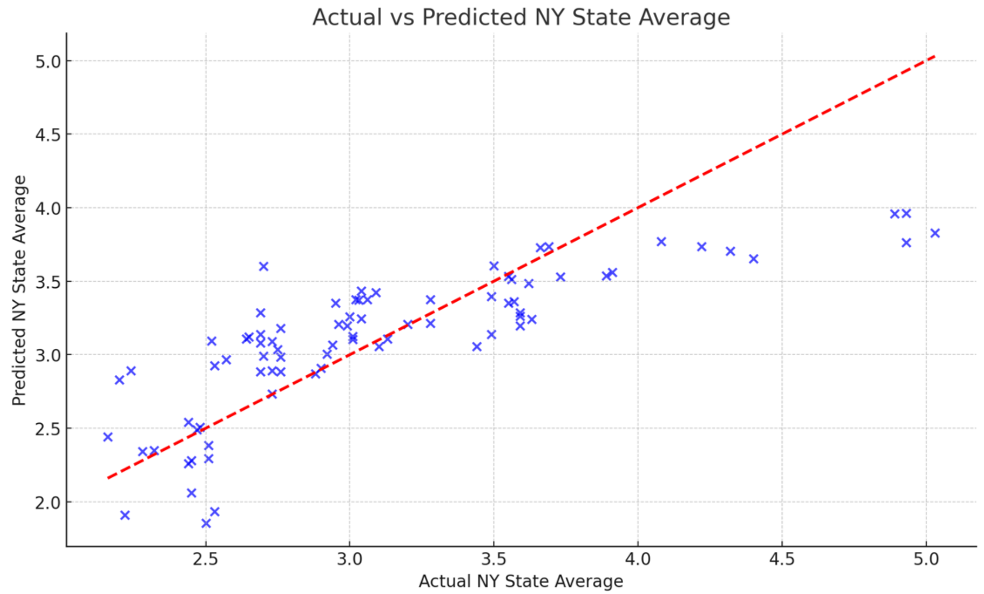
R^2 = 0.689

MSE = 0.165

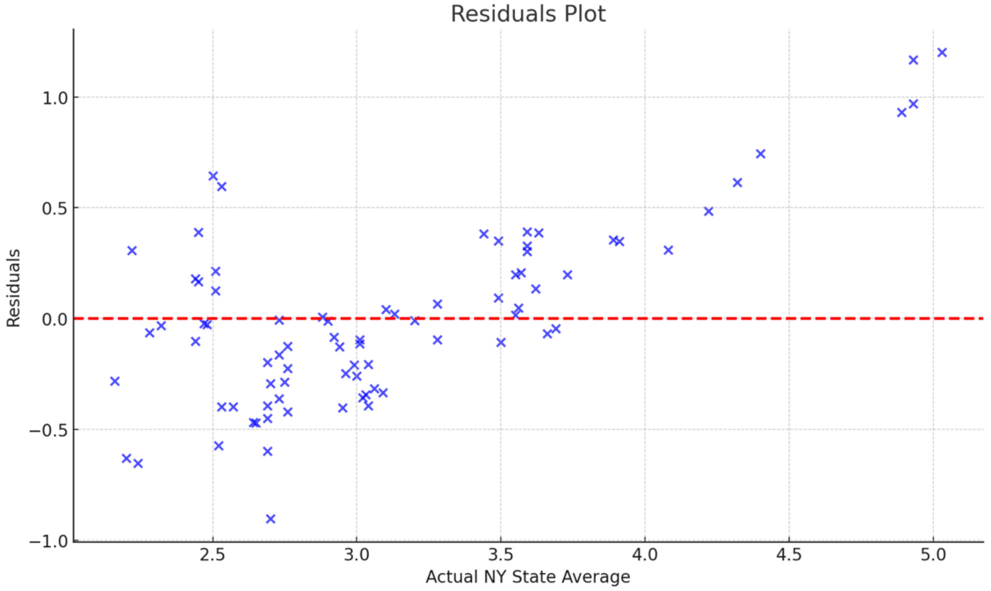
F-statistic = 175.2 (P-value = 0.000)

NY State Average = 7.8924***

- 0.0001 * U.S. Production of Jet Fuel*
- 5.48x10^{-5} * Mid-Atlantic ULS Diesel Stocks***
- 1.06x10^{-5} * U.S. Gasoline Stocks***
- 2.34x10^{-6} * U.S. Crude Oil Stocks***



Actual vs Predicted NY State Average



Residuals Plot

Conclusions

- **Small Direct Effects**:
  - The low magnitude of the coefficients indicates that while statistically significant, the explanatory variables (e.g., stocks and production levels) have small direct impacts on the New York State Average gasoline price.
- **Potential Multicollinearity**:
  - Despite removing high-VIF variables, there could still be some residual multicollinearity, which could limit the effectiveness of the remaining features.
- **Possible Nonlinear Dynamics**:
  - Gasoline prices may be influenced by nonlinear relationships, market dynamics, or interactions between the explanatory variables, which are not fully captured by the linear regression model.
- **Need for Additional Variables**:
  - Factors beyond fuel stocks and production, such as macroeconomic indicators or external events, likely contribute to the price fluctuations and could enhance model accuracy if included.

Further Research

- **Nonlinear Models:**
  - Apply Random Forests or Gradient Boosting to capture nonlinear relationships between fuel stocks, production, and prices.
  - Use Support Vector Machines (SVMs) with nonlinear kernels to explore complex interactions between variables.
- **Feature Engineering:**
  - Incorporate interaction terms to capture potential relationships between stocks and production levels.
  - Introduce lagged variables to account for the delayed impact of stocks and production on gasoline prices.
- **Regularization Methods:**
  - Implement Lasso or Ridge Regression to address multicollinearity and reduce the influence of less important features.
- **Ensemble Learning:**
  - Combine multiple models (e.g., linear regression, decision trees) using Stacking or Blending to improve prediction accuracy.
- **Data Augmentation and Cross-Validation:**
  - Integrate macroeconomic variables (e.g., inflation, unemployment) and weather data to enhance feature richness.
  - Apply k-fold cross-validation to ensure the robustness and generalization of the model across different subsets of the data.