

## **\*\*Observation and Processes\*\***

### **Strategic Importance:**

- Accurate forecasting of gas prices is critical for strategic planning, budgeting, and risk management in sectors dependent on energy costs.
- Insights derived from the model can inform procurement strategies, pricing adjustments, and long-term investment decisions.

### **Variables Used:**

- **Natural Gas Prices**
- **What it is:** The cost per unit of natural gas as traded on commodity exchanges, such as the Henry Hub, which is a benchmark for North American natural gas prices.
- **Influencing factors:** Prices are influenced by factors such as weather conditions, production levels, storage volumes, and changes in demand from residential, commercial, and industrial users.
- **Crude Oil Prices**
- **What it is:** The cost of a barrel of crude oil on global markets, with major benchmarks including WTI (West Texas Intermediate) and Brent.
- **Influencing factors:** Prices fluctuate based on geopolitical events, changes in global supply and demand, OPEC policies, and technological developments in extraction and refining.
- **General Gasoline Prices**
- **What it is:** The average price consumers pay for gasoline at the pump, which can vary regionally.
- **Influencing factors:** Influenced by crude oil prices, refinery output, seasonal demand, taxes, and local market conditions.

- **RBOB Gasoline Prices**

- **What it is:** RBOB (Reformulated Blendstock for Oxygenate Blending) is the benchmark gasoline contract traded on the NYMEX.
- **Influencing factors:** Prices are affected by crude oil costs, refining processes, seasonal driving patterns, and regulations regarding fuel formulations.

- **Heating Oil Prices**

- **What it is:** The price of heating oil, used primarily for residential and commercial heating, particularly in the northeastern U.S.
- **Influencing factors:** Influenced by crude oil prices, seasonal demand in colder months, and changes in inventory levels.

- **Diesel Prices**

- **What it is:** The price of diesel fuel, used in various engines and machinery, with prices typically quoted at the pump.
- **Influencing factors:** Diesel prices are affected by crude oil prices, refinery capacity, and demand from transportation and industrial sectors.

- **Kerosene Prices**

- **What it is:** The cost of kerosene, another petroleum product used for heating, lighting, and as jet fuel.
- **Influencing factors:** Prices are influenced by crude oil prices, production levels, and its use in aviation and household sectors in certain regions.

- **Propane Prices**

- **What it is:** The price of propane gas, used for heating, cooking, and as a fuel for certain vehicles.
- **Influencing factors:** Influenced by production rates, especially as a byproduct of natural gas processing and oil refining, as well as seasonal demand.

- **Oil Stock Levels**

- **What it is:** The quantity of oil held in storage, which includes strategic reserves and commercial inventories.
- **Influencing factors:** Stock levels change based on production rates, import levels, consumption rates, and policy decisions regarding strategic reserves.

- **Drilling Activity**
- **What it is:** The number of rigs actively exploring or extracting oil and gas.
- **Influencing factors:** Drilling activity is influenced by oil and gas prices which determine the profitability of new drilling, and regulatory policies.
  
- **Gas Production Volumes**
- **What it is:** The total amount of gas produced, typically measured in cubic feet or cubic meters.
- **Influencing factors:** Production volumes are influenced by market demand, prices, technological advancements, and natural resource availability.
  
- **Gas Consumption Volumes**
- **What it is:** The amount of gas consumed by end users, including residential, commercial, and industrial sectors.
- **Influencing factors:** Consumption is driven by economic activity, efficiency of gas usage, seasonal heating or cooling needs, and population growth.
  
- **Gas Storage Levels**
- **What it is:** The volume of gas held in storage facilities to manage supply and demand over time, ensuring stability in gas supply.
- **Influencing factors:** Storage levels are affected by production rates, consumption patterns, and seasonal variations in demand.
  
- **Gas Import Volumes**
- **What it is:** The total volume of gas imported from other countries, often measured in cubic feet or cubic meters.
- **Influencing factors:** Import volumes are driven by domestic production shortfalls, international market conditions, and cross-border trade agreements.
  
- **Gas Import Prices**
- **What it is:** The price paid for gas imported from other countries, which can vary significantly based on the source and type of gas.
- **Influencing factors:** Prices are influenced by international supply and demand, transportation costs, and geopolitical factors.

## **Derived Features:**

## 1. Price Ratios

- **Ratios of different fuel prices** such as crude oil price to gas prices, or diesel price to heating oil price. These ratios can reveal relative pricing trends and substitutions that might occur depending on price movements.
- **Example: `diesel_to_heating_oil_ratio = diesel price / heating oil price`**
- **Reasoning:** Price ratios can help capture cross-dependencies between different fuel types which might be influenced by similar market forces or seasonal demand changes.

## 2. Historical Volatility

- **30-day or 60-day rolling volatility of prices** for commodities like crude oil, gas, and others.
- **Example: `rolling_volatility_crude_oil = crude oil price.rolling(window=30).std()`**
- **Reasoning:** Volatility measures how much the price of a commodity fluctuates over a period and can be indicative of market uncertainty or potential future price movements.

## 3. Price Differences

- **Differences between related commodity prices**, such as the spread between crude oil prices and various types of gas prices.
- **Example: `crude_vs_gas_price_diff = crude oil price - conventional gas price`**
- **Reasoning:** Price differences or spreads between similar commodities can indicate relative supply and demand dynamics, refining margins, or logistical costs.

## 4. Seasonal Indicators

- **Cyclical features derived from the week number**, such as sine and cosine transformations to capture seasonality effects throughout the year.
- **Example: `sin_week = np.sin(2 * np.pi * week number / 52)`, `cos_week = np.cos(2 * np.pi * week number / 52)`**
- **Reasoning:** Energy consumption patterns are highly seasonal; transforming week numbers into cyclical features allows the model to better learn and predict seasonal impacts on prices and demands.

## 5. Demand-Supply Balance

- **Ratios or differences between production, consumption, storage, and import volumes** to reflect the demand-supply balance in the market.

- **Example:**  $\text{supply\_demand\_ratio} = (\text{gas production} + \text{gas import volume}) / \text{gas consumption}$
- **Reasoning:** A higher ratio or a significant change in the ratio can signal shifts in market equilibrium, potentially impacting prices and availability.

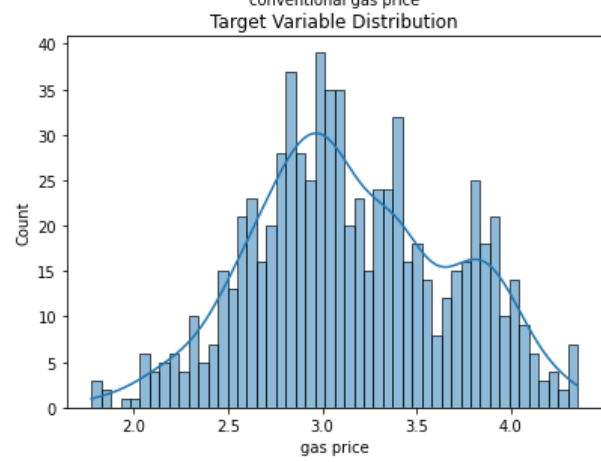
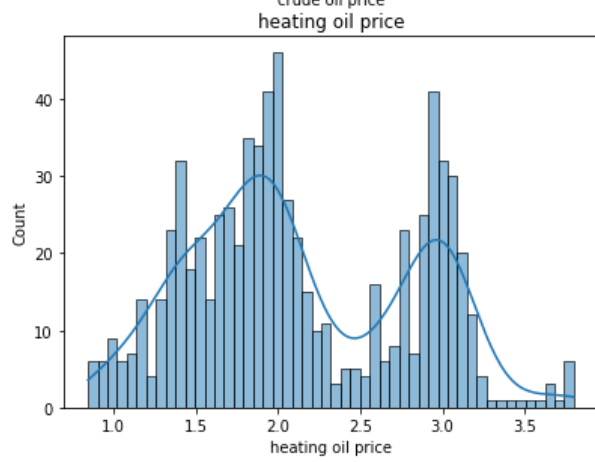
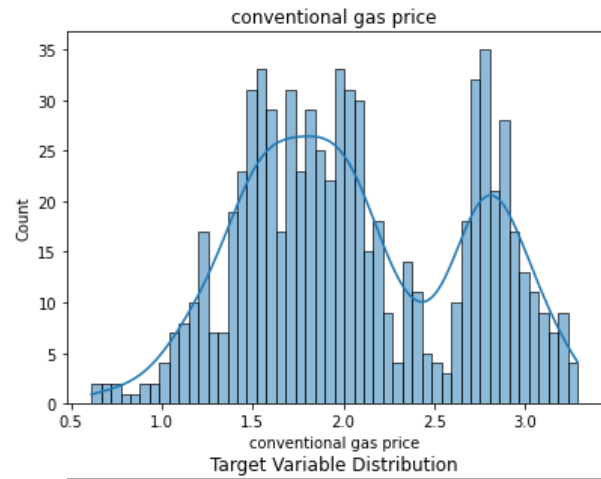
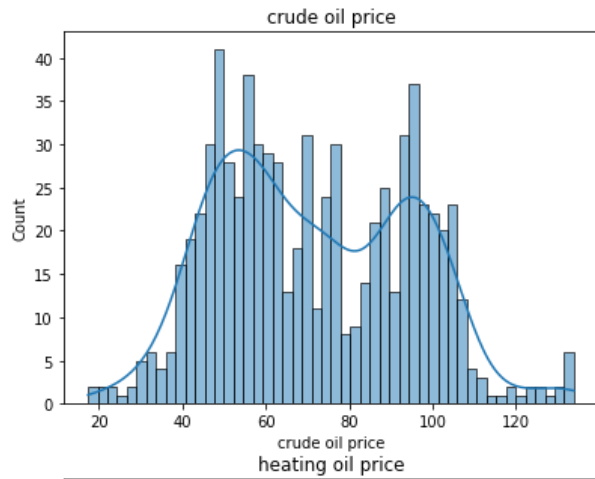
## 6. Import Cost

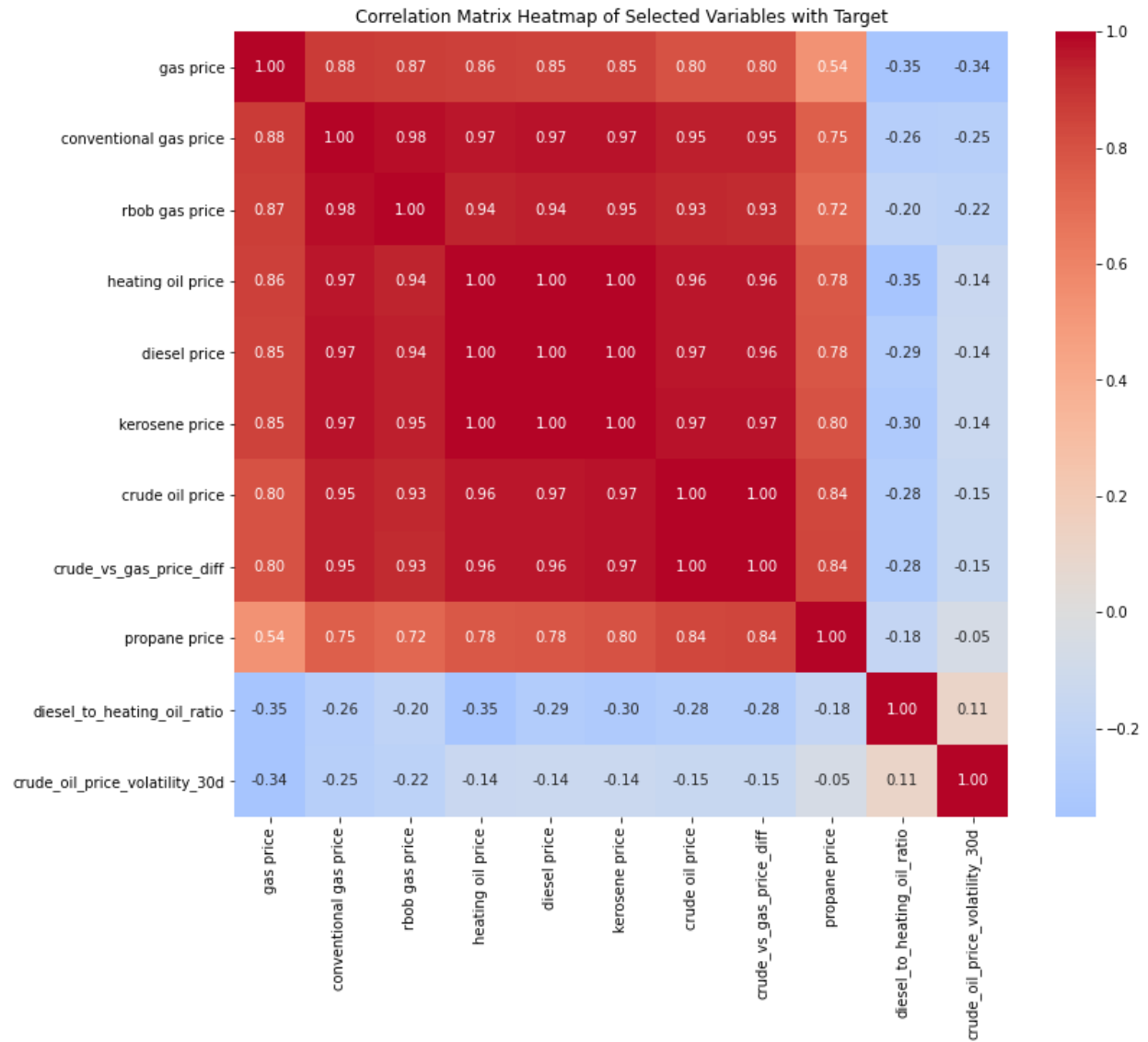
- **Total cost of imports** calculated by multiplying the volume by the price of imports.
- **Example:**  $\text{total\_import\_cost} = \text{gas import volume} * \text{gas import price}$
- **Reasoning:** This feature can indicate the economic burden or benefit derived from importing versus local production, influencing national energy policies and market prices.

## 7. Inventory Turnover

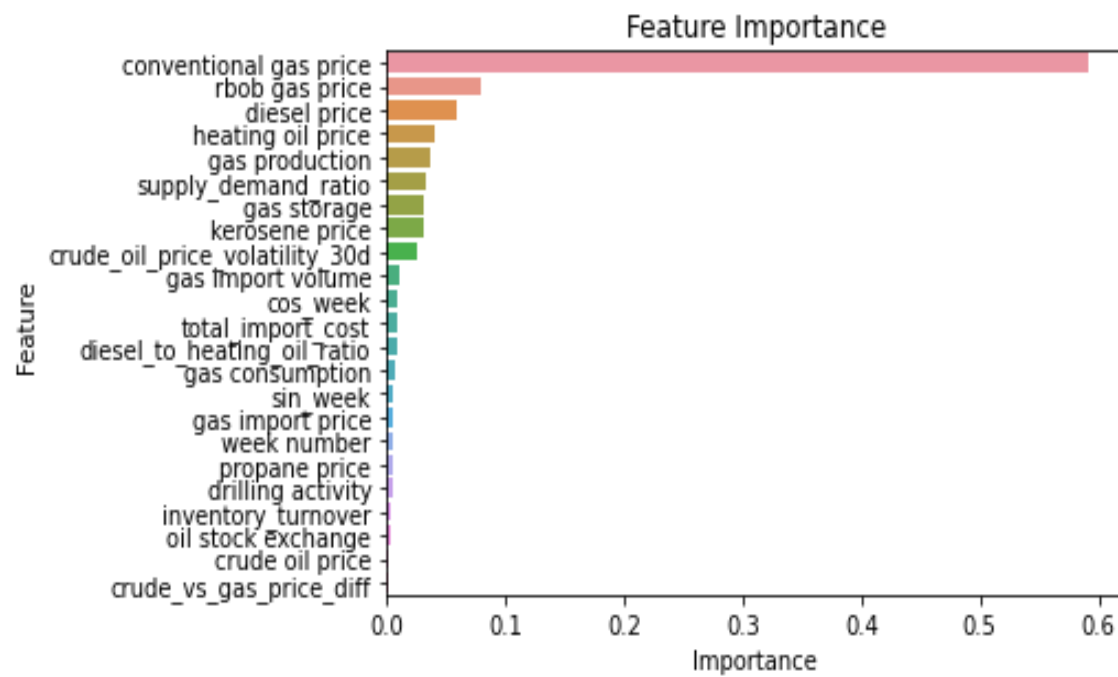
- **Ratio of consumption to storage** to gauge how quickly the stored gas is being used.
- **Example:**  $\text{inventory\_turnover} = \text{gas consumption} / \text{gas storage}$
- **Reasoning:** A lower ratio might indicate oversupply or decreased demand, while a higher ratio could signal tight market conditions or increased demand.

## EDA and Correlational Analysis:

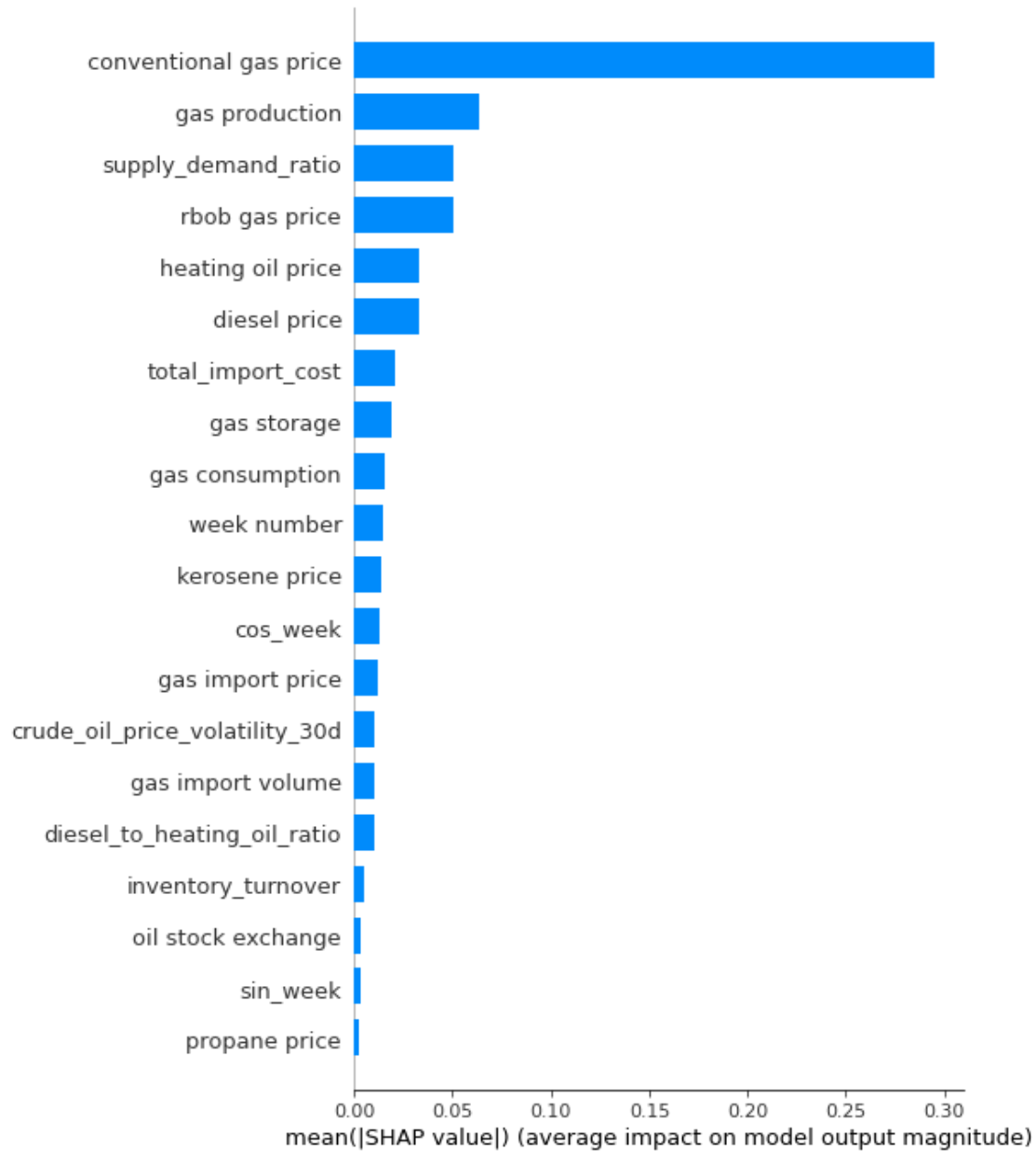




## Feature Importance using RF and SHAP Analysis:







### Outlier Detection:

**The Isolation Forest** algorithm detects outliers by isolating observations. It randomly selects a feature and splits the dataset between the maximum and minimum values of the selected feature; this process repeats recursively, and the logic is that anomalies are

easier to isolate and will have shorter paths in the tree structure, thus identifying them as outliers more efficiently than normal points.

**Isolation Forest Setup:** Configures an Isolation Forest with 100 trees, using the entire training dataset for each tree and expecting about 5% of the data to be outliers, with a fixed random state for consistency.

**Model Training:** Fits the Isolation Forest model on `X_train_New` to learn and identify outliers.

**Outlier Detection and Marking:** After fitting, assigns an outlier score and marks outliers directly in the dataset using the `is_outlier` column.

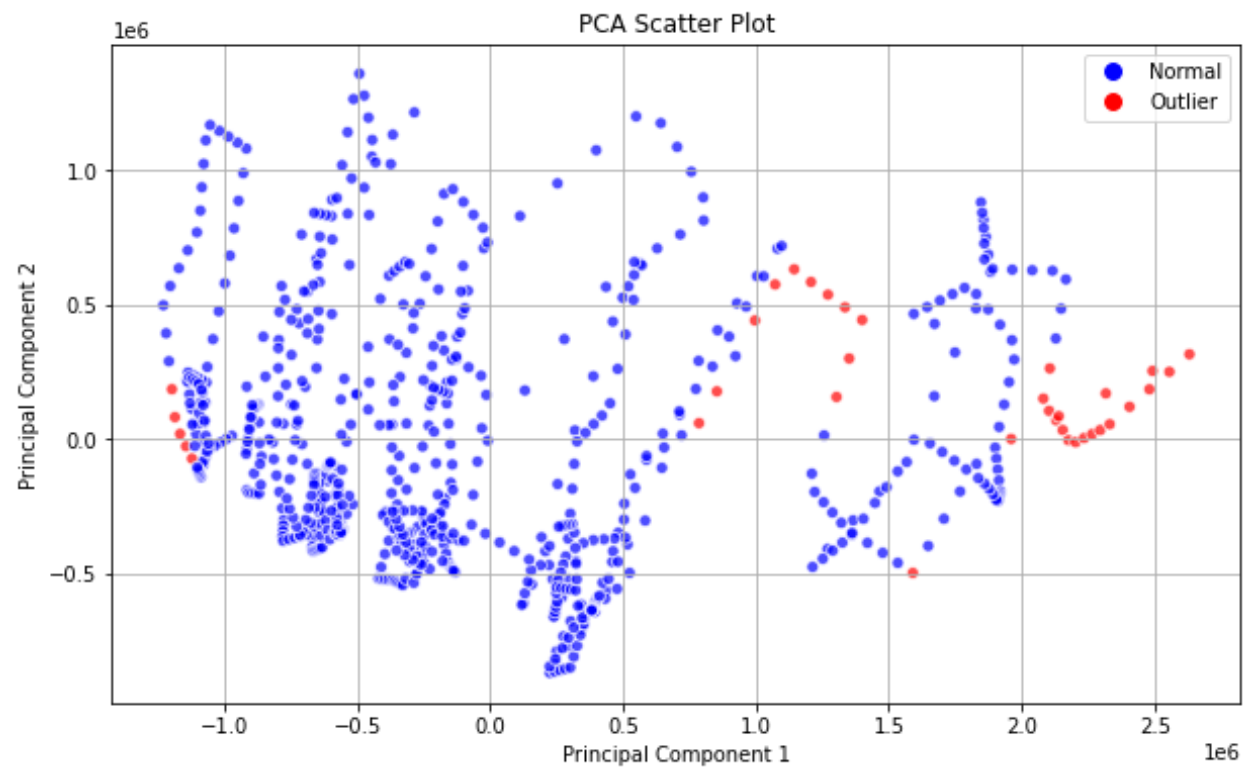
**Outlier Removal:** Filters out data points marked as outliers, retaining the rest in `clean_data_train`.

**Results Display:** Outputs the count of detected outliers and the shape of the dataset after outlier removal.

### **PCA for Outlier Visualization:**

Principal Component Analysis (PCA) is a statistical technique that can be used for outlier detection among other applications like dimensionality reduction. In the context of outlier detection, PCA helps identify anomalies in a dataset by transforming the data into a set of linearly uncorrelated components (principal components) and analyzing the variance captured by these components.

- PCA transforms the original variables to new uncorrelated variables (principal components), ordered so that the first few retain most of the variation present in all of the original variables.
- The transformation to a lower-dimensional space (using the first few principal components) makes it easier to visualize and detect outliers.



Final Prediction:

