Aaron: Gasoline Prices with Economic Indicators

**Objective:** Build a machine learning model to predict future gasoline prices for New York State or specific cities.

**Steps:**

1. **Data Integration:**
   * Use crude oil prices, Consumer Price Index (CPI), and local/state gasoline tax data as external features.
2. **Feature Engineering:**
   * Include time-related features like seasonality (month, week, holidays).
   * Compute rolling averages and volatility of crude oil prices.
3. **Model Building:**
   * Train regression models such as Random Forest, Gradient Boosting (e.g., XGBoost), or LSTM for time-series forecasting.
   * Evaluate using metrics like RMSE, MAE, and MAPE.
4. **Insights:**
   * Identify key drivers of gasoline prices using feature importance analysis (SHAP or permutation importance).

Jen: Cluster Analysis

Uncover spatial or temporal patterns in fuel prices across regions in NY state.

**Objective**

* Group regions based on similarities in average fuel prices and trends.
* Identify high-cost and low-cost clusters or regions with distinct pricing patterns.

**Steps**

1. **Data Preprocessing**
   * Standardize numerical variables (e.g., average prices, production, stocks) to ensure equal weighting.
   * Aggregate data by region and/or time if needed (e.g., monthly averages).
   * Handle missing values using imputation techniques.
2. **Feature Selection**
   * Use columns like:
     + Gasoline/Diesel prices (by location and statewide).
     + Regional production or stock levels.
     + Spot prices for related commodities (e.g., NY Conventional Gasoline Spot Price, WTI Crude Oil Spot Price).
   * Consider adding external data:
     + **Traffic Data**: Regional congestion levels (e.g., using INRIX or Google Maps API).
     + **Weather Data**: Regional seasonal patterns (e.g., heating degree days from NOAA).
3. **Clustering Techniques**
   * **K-Means Clustering**: Group regions based on fuel price similarities.
     + Example: Cluster regions into 3-5 groups (e.g., high-price, medium-price, low-price).
   * **DBSCAN (Density-Based Spatial Clustering)**: Identify clusters of regions with unusual fuel pricing behavior.
   * **Time Series Clustering**:
     + Use Dynamic Time Warping (DTW) to group regions with similar price trends over time.
4. **Evaluation**
   * Visualize clusters using:
     + Heatmaps (prices by region over time).
     + Scatter plots (cluster assignments on geographical maps).

Pat: **Anomaly Detection**

Anomaly detection focuses on identifying unusual price fluctuations or discrepancies across regions and time.

**Objective**

* Detect and analyze unexpected changes in fuel prices that might be driven by supply chain disruptions, weather events, or economic policies.

**Steps**

1. **Data Preprocessing**
   * Detrend and deseasonalize time series data using moving averages or seasonal decomposition.
   * Aggregate data weekly or monthly for consistency.
2. **Feature Engineering**
   * Calculate price differentials:
     + Between regions (e.g., Albany vs. Buffalo).
     + With spot prices (e.g., NY Conventional Gasoline Spot Price vs. Albany Average).
   * Derive additional features:
     + Rolling averages/volatility (e.g., 7-day or 30-day).
     + External data:
       - **Weather Data**: Look for extreme conditions or storm events (e.g., NOAA).
       - **Traffic Data**: Examine congestion spikes (e.g., INRIX).
3. **Anomaly Detection Models**
   * **Statistical Methods**:
     + Z-scores for price outliers.
     + Seasonal-trend decomposition to detect abrupt deviations from seasonal norms.
   * **Machine Learning Models**:
     + **One-Class SVM**: Identify anomalies by modeling normal price distributions.
     + **Autoencoders**: Train neural networks to reconstruct normal price patterns; deviations signal anomalies.
4. **Visualization**
   * Plot anomalies on time series charts for easy interpretation.
   * Overlay external factors (e.g., storms, demand spikes) to explain detected anomalies.

External Data Sources:  
  
**Traffic Data**

* **Source**: INRIX, Google Maps API.
* **Use**: Understand price surges related to holiday or commuter traffic. Cluster regions based on their sensitivity to traffic patterns.

**Weather Data**

* **Source**: NOAA Climate Data Online.
* **Use**: Analyze weather-related disruptions (e.g., hurricanes causing price spikes) and include these patterns in anomaly detection.

**Commodity Futures**

* **Source**: CME Group.
* **Use**: Incorporate futures market volatility into anomaly detection to flag instances of speculative price behavior.