Oil sales Forecasting

1. **Approach and Assumptions**

This project aims to analyze historical retail sales data and build a predictive model to forecast sales value. The dataset contains product-level and store-level information including city, store name, manufacturer, brand, product class, size, SKU, price bracket, year, month, total sales value, sales volume, and average price.

Problem Definition & Approach

- Objective: Predict continuous sales values → framed as a regression task.
- Target Variable: value_sales.
- Models Applied: Random Forest and Linear Regression, used for comparative performance analysis.

Data Preparation

1. Feature Engineering
    - Removed sku (redundant composite of brand, class, size, and variant).
    - Extracted store_code from store_name..
    - Encoded categorical variables: city, brand,class ,size, and price_bracket.
2. Data Quality Checks
    - Confirmed no missing values.
    - No duplicate records identified.
    - Outliers retained, as they represent legitimate business extremes rather than errors.

Validation Design

- Training Set: Data up to 2023.
- Test Set: Data from 2024 only.
- Rationale: This split simulates real-world forecasting and avoids lookahead bias by ensuring the model is evaluated on unseen future data.

Modeling Strategy

1. Feature Selection: Product attributes, pricing variables, and temporal features.
2. Assumptions:
    - Price–sales relationships remain stable over time.
    - Demand patterns within price brackets are consistent.
3. Evaluation Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 to provide a comprehensive view of predictive accuracy.

The final model achieved strong performance with very high R^2, indicating excellent predictive accuracy.

2. **Key Findings from the Analysis**

- The random forest model achieved high accuracy, demonstrating that sales value can be reliably predicted using product attributes and time-based features.

- Strong relationships were observed between price bracket, volume sales, and total sales value.

---

3. **Business-Relevant Insights**

Insight 1: Sales vs City

- AL BAHA leads with 143,558.12 in sales, significantly higher than other cities.
- This suggests strong demand or effective distribution channels in that region.
- MAKKAH, DAMMAM, TABUK, YANBU, JEDDAH all show sales above 100,000, indicating consistent performance across multiple urban centers.
- These cities represent stable, high-value markets.

Insight 2: Sales vs Years

- 2022: Highest sales at 431,748.10 with 6,974 units sold.
- 2023: Noticeable decline to 379,364.55 (-12% vs 2022) and volume dropped to 6,321 units.
- 2024: Recovery to 418,566.06 (+10% vs 2023), with volume rising to 6,648 units, though still slightly below 2022 levels.

Insight 3: Sales vs Class

- Highest sales value (271,588.36) and largest volume (4,232.5 units).
- Indicates strong demand and market leadership in this category.

Insight 4: Sales vs Brand

- HILAL(13,803.34) and GULF GOLD (13,261.79) command the highest average prices.
- These brands are positioned at the top end of the market, likely reflecting premium quality or strong brand equity.
