Oil Sales Analysis

1) Approach & assumptions

Approach

- Conducted exploratory data analysis (EDA) to summarize the dataset and spot distributions, seasonality, and anomalies.
- Performed data quality checks (missing values, duplicates, negative numbers, internal price consistency checks) to validate the dataset.
- Detected outliers with the Interquartile Range (IQR) method and visualized them for interpretation (no automatic removal was performed).
- Encoded categorical variables (label encoding) and created modeling features that include product attributes, location, time, and price information.
- Built baseline predictive models using Random Forest (regression to predict sale value and classification to label high/low sales based on the median).

Assumptions

- Each row in the dataset represents a valid, individual sales transaction.
- `volume_sales`, `value_sales`, and `average_price` are numerical; any non-numeric rows were converted/validated before modeling.
- Large transactions flagged as outliers may be legitimate (e.g., wholesale/bulk orders), so the notebook does not remove them automatically.
- Label encoding is appropriate for tree-based models; if different models are used later (e.g., linear models), encoding strategy should be revisited.

2) Key findings from the analysis

1. Revenue concentration in a small set of SKUs

- Top-performing SKUs account for a large share of revenue (top 10 SKUs contribute a meaningful percentage). This shows portfolio concentration and dependence on a few products.

2. Clear seasonality across months

- Monthly revenue analysis shows peaks and troughs indicating predictable seasonality. These trends can be used to optimize inventory and promotion timing.

3. City/region-level differences

- Revenue distribution across cities is uneven. A small number of cities generate most of the revenue opportunities for localized decisions and investment.

4. High-value transactions (outliers) exist and may be business relevant

- Large value transactions are present which might be bulk orders or B2B transactions. They should be treated separately when required for pricing / forecasting decisions.

3) Possible next steps (short and medium term)

Short-term (analysis & modeling)

- Run per-SKU and regional models where data is concentrated to achieve higher precision for top products and top cities.
- Separate high-value/bulk transactions from regular retail sales to build specialized models and prevent skew in forecasting.
- Carry out price elasticity testing (A/B experiments) on groups of SKUs to inform pricing strategy.

Medium-term (data & operations)

- Enrich the dataset with customer-level and channel-level information (customer type, online vs in-store, promotional flags) to enable segmentation and better personalization.

- Build an automated pipeline for feature extraction, model training, and evaluation (CI/CD for data science) to keep models updated with new data.