# **Milestone 1: Data Collection, Exploration, and Preprocessing**

## [**1. Introduction**](#1-introduction)

### 1.1 Business Problem

Retailers face two critical and interconnected challenges: accurately forecasting product demand and setting optimal prices. Poor demand prediction often results in excess inventory or stockouts, both of which impact profitability and customer satisfaction. Additionally, static pricing strategies fail to respond effectively to fluctuating demand, holidays, and customer behavior.

This project addresses both challenges by building a **Demand Prediction Model** and a **Dynamic Pricing Model** tailored to the operations of **Corporacion Favorita**, a major grocery retailer in Ecuador. By forecasting demand at a granular level and dynamically adjusting prices based on key variables, the project aims to help the company make data-driven decisions around inventory, promotions, and revenue optimization.

### 1.2 Project Goal

This project centers on **time series forecasting of unit sales** for **Corporación Favorita** using historical transaction data and external signals such as holidays and oil prices. The ultimate goal is to:

**1-Predict daily item-level sales** for each store to support inventory management, staffing, and supply chain operations.

**2-Design a dynamic pricing framework** that adjusts product prices based on demand patterns, product attributes, seasonal trends, and macroeconomic indicators.

Specifically, the goal of Milestone 1 is to prepare the data for future modeling by:

* Collecting and understanding all relevant datasets.
* Cleaning the data, handling missing values, and correcting data types.
* Exploring the data through descriptive statistics and visualizations.
* Identifying key patterns, trends, and anomalies in sales behavior.
* Testing business hypotheses related to seasonality, promotions, and store/item attributes.
* Merging datasets to create a consistent and clean base for future modeling.

### 1.3 Dataset Overview

The dataset for this project comes from a Kaggle competition hosted by **Corporacion Favorita**, a grocery retailer in Ecuador. It includes multi-year time series data designed to support forecasting of daily product family sales across various stores.

#### **1.4 Key Files**

**train.csv**: Contains historical sales data, including **date, store\_nbr, family,** and **onpromotion** status. The target variable is **sales**. The training dataset contains **3,000,888 entries** and **6 columns**: **id, date, store\_nbr, family, sales,** and **onpromotion**.

**test.csv**: Similar structure to train.csv, but the **test dataset contains 28,512 entries** and **5 columns**: **id, date, store\_nbr, family,** and **onpromotion**. The **sales** column is absent from the test data, as this is the target variable we aim to predict.

**stores.csv**: Metadata about each store, including **city, state, type,** and **cluster**.

**holidays\_events.csv**: A calendar of national, regional, and local holidays and events, with indicators for transferred holidays, bridge days, and special working days.

**transactions.csv**: Contains daily transaction data, which may provide insights into customer buying behavior and sales trends.

**oil.csv**: Daily oil prices, which can influence consumer spending and sales, given Ecuador's dependency on oil.

**sample\_submission.csv**: A template for submitting predictions in the required format.

#### **1.5 Contextual Notes**

Public sector wages are typically paid on the **15th and last day** of each month, which could potentially influence supermarket sales patterns.

A significant **earthquake** struck Ecuador on **April 16, 2016**, leading to an increase in sales of essential products in the aftermath, which should be considered as a potential anomaly in sales behavior for that period.

This rich, multi-source dataset provides the foundation for understanding sales dynamics and supports accurate demand forecasting.

### 1.6 Initial Findings

The **sales data** appears to have reasonable granularity for forecasting, with daily records spanning multiple years. However, there are occasional missing entries, particularly on days with no sales data, which will need to be handled appropriately.

**Store and family-level data** is quite diverse, with hundreds of unique stores and product families, requiring careful treatment of categorical variables during preprocessing.

The **onpromotion** column shows variation over time, indicating promotions are being used dynamically across stores and families. There is potential to leverage this for demand forecasting.

The presence of **oil price data** and **holiday events** could provide important features for modeling, as both are likely to influence sales, especially for specific product families and time periods.

## **2. Exploratory Data Analysis (EDA)**

### 2.1 Understanding the datasets

An in-depth exploration of the datasets is presented to gain insights into the available variables,their distributions and relationships. This step will provide an initial undertanding of the datasets to identify any data quality issues that will inform the cleaning and pre-processing.

### 2.2 Shape of The Datasets

The train dataset contains 3,000,888 rows and 6 columns while the test dataset contains 28,512 rows and 5 columns.

The train dataset is significantly larger than the test dataset in terms of the number of rows. This is expected, as the train dataset is usually larger to provide sufficient data for model training.

The Holiday Events dataset contains 350 rows and 6 columns. This dataset provides information about various holidays and events.

The Oil dataset consists of 1,218 rows and 2 columns. This dataset includes information about the daily price of oil.

The Stores dataset contains 54 rows and 5 columns. This dataset provides details about different stores, such as their locations, types, and clusters.

The Transactions dataset contains 83,488 rows and 3 columns. This dataset contains information about the number of transactions made at each store on specific dates.

### 2.3 Column Information of The Datasets

The train dataset contains 3,000,888 entries and 6 columns: id, date, store\_nbr, family, sales, and onpromotion.

The test dataset contains 28,512 entries and 5 columns: id, date, store\_nbr, family, and onpromotion.

As expected, the test dataset does not have the sales column. This column is not needed because sales is the variable we want to predict. The goal is to use the trained model to predict or forecast the sales in the test data based on the other available features.

The Holiday Events Dataset contains 350 entries and 6 columns: date, type, locale, locale\_name, description, and transferred.

The Oil Dataset contains 1,218 entries has 2 columns: date and dcoilwtico.

The dcoilwtico column has 1,175 non-null values, indicating that there are some missing values in this column.

The Stores dataset contains 54 entries and 5 columns: store\_nbr, city, state, type, and cluster.

The Transactions dataset contains 83,488 entries and 3 columns: date, store\_nbr, and transactions.

### 2.4 Transforming the date column to datetime format

Date Column Data Type After Transformation:

Train dataset: datetime64[ns]

Test dataset: datetime64[ns]

Holiday Events dataset: datetime64[ns]

Oil dataset: datetime64[ns]

Transactions dataset: datetime64[ns]

It was converted to datetime to enable further Time-series Analysis

### 2.5 Summary statistics of the datasets

To understand the structure and nature of the data, we examined summary statistics for each key dataset used in this project:

#### **Train Dataset**

**Entries:** 3,000,888

**Columns:** id, date, store\_nbr, family, sales, onpromotion

**Sales** ranged from **0** to **124,717**, with a **mean of 358** and a **median of 11**, indicating a highly skewed distribution.

**Onpromotion** values also show many zeros, with a max of **741 promotions** in a record, and a mean of **2.6**.

#### **Test Dataset**

**Entries:** 28,512

**Similar structure** to train data, but without the sales column (which is the target variable).

**Onpromotion** has a mean of **7**, with some entries having up to **646 promotions**, suggesting variability in marketing strategies.

#### **Holiday Events Dataset**

**Entries:** 350 holiday and event records.

Dates span from **2012-03-02** to **2017-12-26**, with a median around **mid-2015**, covering local, regional, and national events.

#### **Oil Prices Dataset**

**Entries:** 1,218 days of data; however, only **1,175** have valid oil prices.

Prices range from **26.19** to **110.62**, with an average of **67.7**.

The missing values were later handled using a **backfill strategy**.

#### **Stores Metadata**

**Entries:** 54 unique stores.

Each store belongs to a **cluster (1–17)** and is located in different **cities/states**.

The store numbers range from **1 to 54**, distributed relatively evenly.

#### **Transactions Dataset**

**Entries:** 83,488

Records daily transaction volume for each store, ranging from **5 to 8,359 daily transactions**.

The average transaction volume is **1,695**, with noticeable variation across stores and time.

### 2.6 Checking for Missing Values in The Datasets

Of all the columns in the different datasets, it’s only the dcoilwtico (daily crude oil prices) column in the Oil dataset that has missing values. The column has 43 missing values.

### 2.7 Checking for the completeness of the date column in the Train and Test Datasets

The train dataset is incomplete. The following dates are missing:

[2013-12-25, 2014-12-25, 2015-12-25, 2016-12-25]

The test dataset is complete. It includes all the required dates.

### 2.8 Merging the Train Dataset with the Stores, Transactions, Holiday Events, and Oil Dataset

The use of an inner merge in this time series forecasting project for **Corporation Favorita** helps to ensure data consistency, avoid missing values, and focus on the relevant data for accurate predictions.

An inner merge type retains only the rows with matching values in the specified columns. In the context of time series forecasting, it allows us to merge datasets based on a common time index or timestamp. By performing an inner merge, we ensure that only the rows with corresponding timestamps in both datasets are included in the merged result. This is important for time series forecasting because you want to align the data from different sources based on their timestamps to build a consistent and accurate forecasting model.

With an inner merge, you eliminate any non-matching timestamps, which may not be useful for forecasting and could introduce inconsistencies in the data. By focusing on the intersection of the datasets, we can create a merged dataset that contains the necessary information for accurate time series forecasting.

The merged dataset, after merging the train dataset with additional datasets, contains 322,047 rows and 16 columns. Two columns have been renamed as a result of the merging, type\_x and type\_y.

The type\_x column represents the store type.

The type\_y column represents the holiday type.

We then renamed type\_x to store\_type and type\_y to holiday\_type

### 2.9 Checking for Missing Values in the Merged dataset

The merged dataset has no missing values.

### 2.10 Checking for Duplicate Values in The Datasets

There are no duplicates in the merged and test datasets

## [**3. Data Cleaning and Preparation**](#2-data-cleaning-and-preparation)

The first step in preparing the data for modeling is ensuring that it is clean, consistent, and structured in a way that allows for effective analysis. This phase includes handling missing values, optimizing memory usage, and correcting any data type inconsistencies. Below are the key steps involved in the data cleaning and preparation process.

### 3.1 Missing Values

Of all the columns in the different datasets, it is only the dcoilwtico (daily crude oil prices) column in the Oil dataset that has missing values. The column has 43 missing values.

We filled in missing values in the dcoilwtico column using backfill strategy

The **backfill strategy** involves filling missing data with the most recent valid data point available before the missing value. Essentially, if a missing value exists in the dcoilwtico column, it is replaced by the next available valid (non-missing) value further down the time series. This approach ensures that the missing values are filled with the closest possible data points in terms of time.

The date column in the train dataset had some missing rows that we talked about earlier, and we imputed them manually.

### 3.2 Downcasting for Memory Optimization

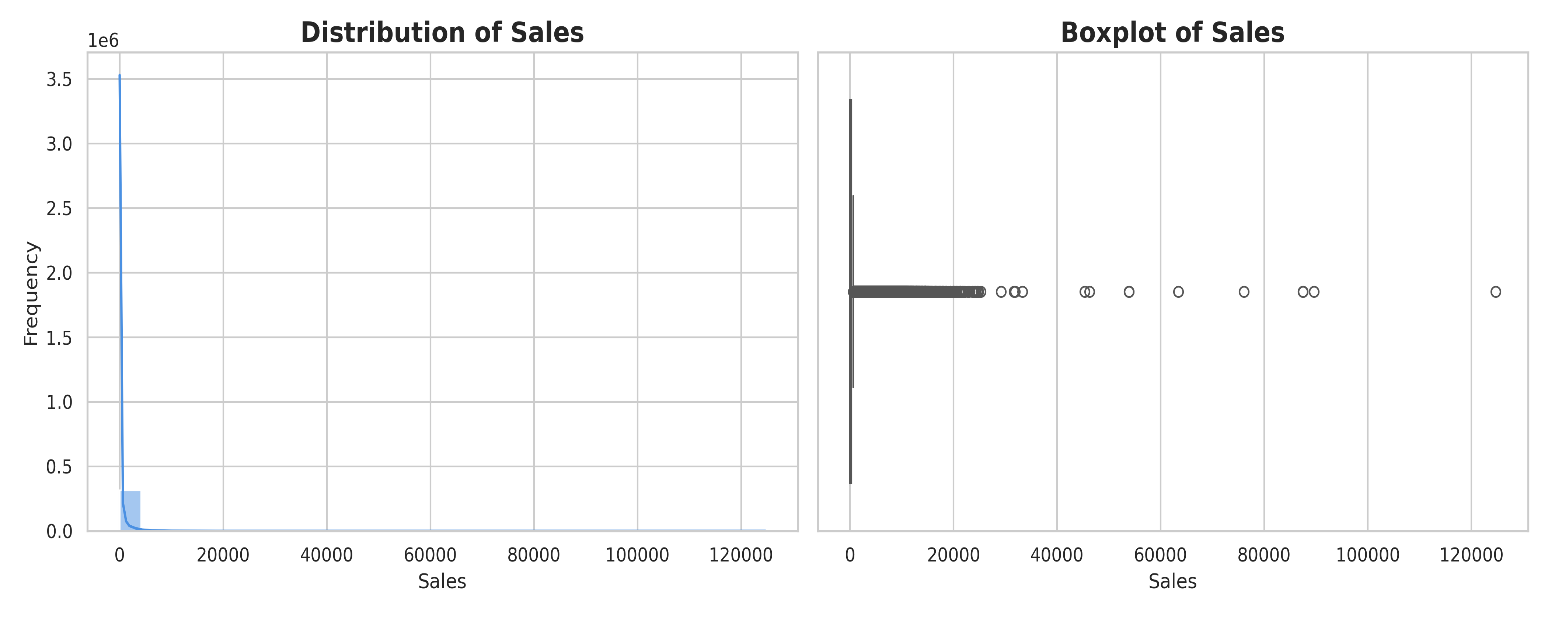
Given the large size of the dataset, memory efficiency was a priority during the cleaning process. To reduce memory usage, we performed **downcasting**, converting numeric columns (such as sales and onpromotion) from float64 to float32 where applicable. Similarly, categorical columns like family were converted from object type to more memory-efficient formats like category.

### 3.3 Data Type Corrections:

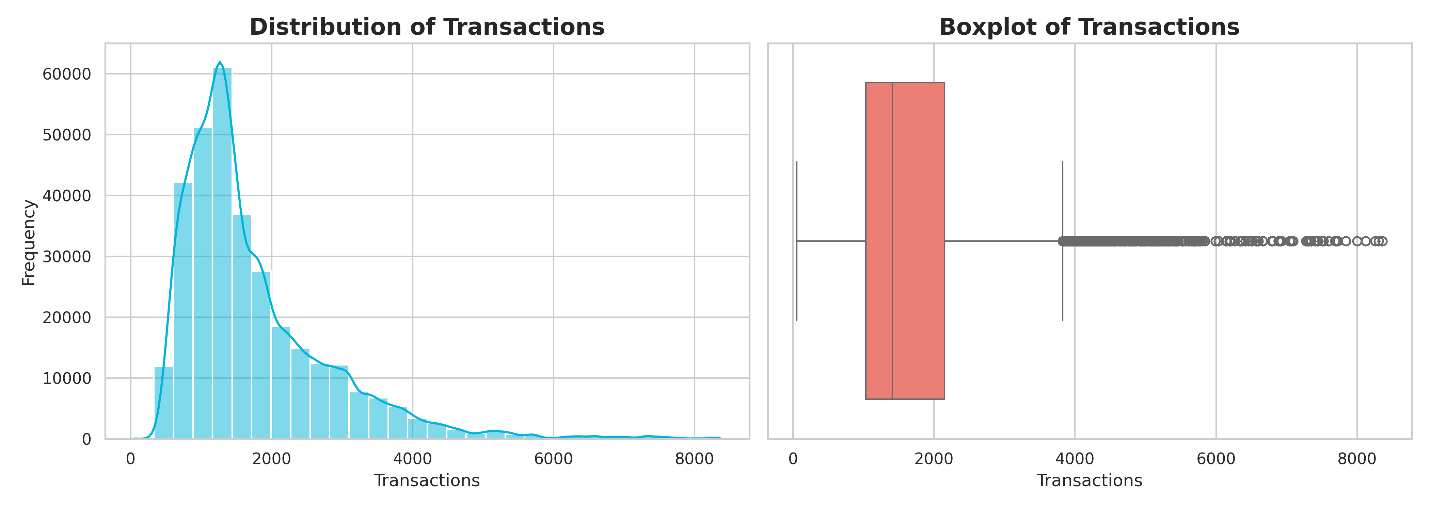
**Date fields** were converted to datetime format to enable time series operations.

## [**4. Data Visualization**](#4-data-visualization)

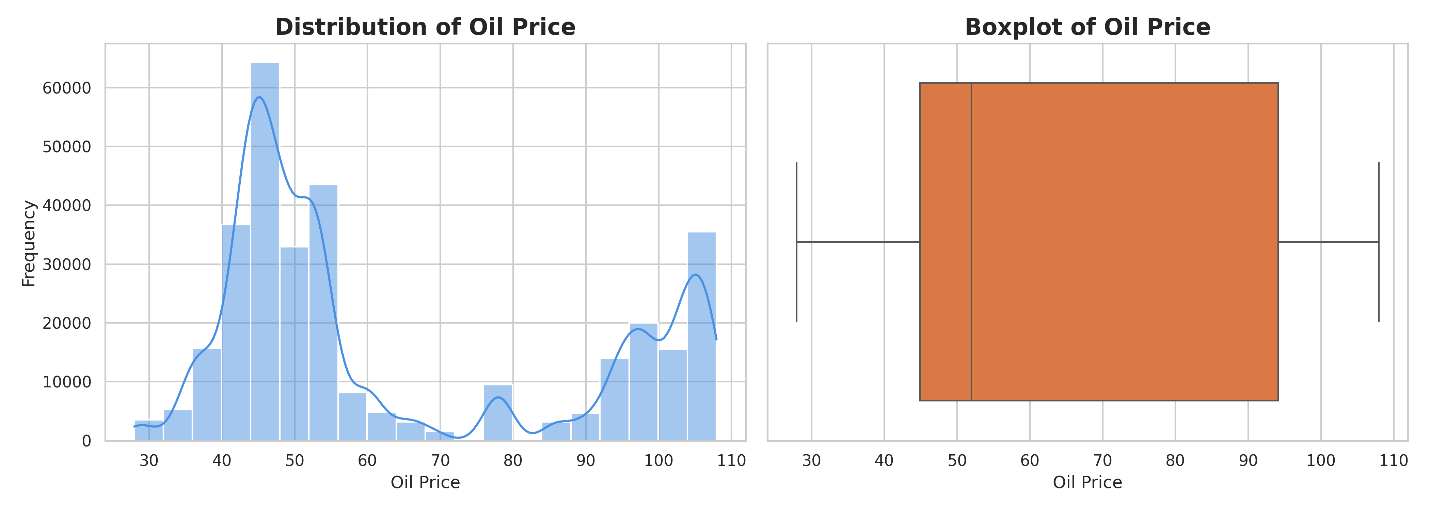
### 4.1 Univariate Analysis



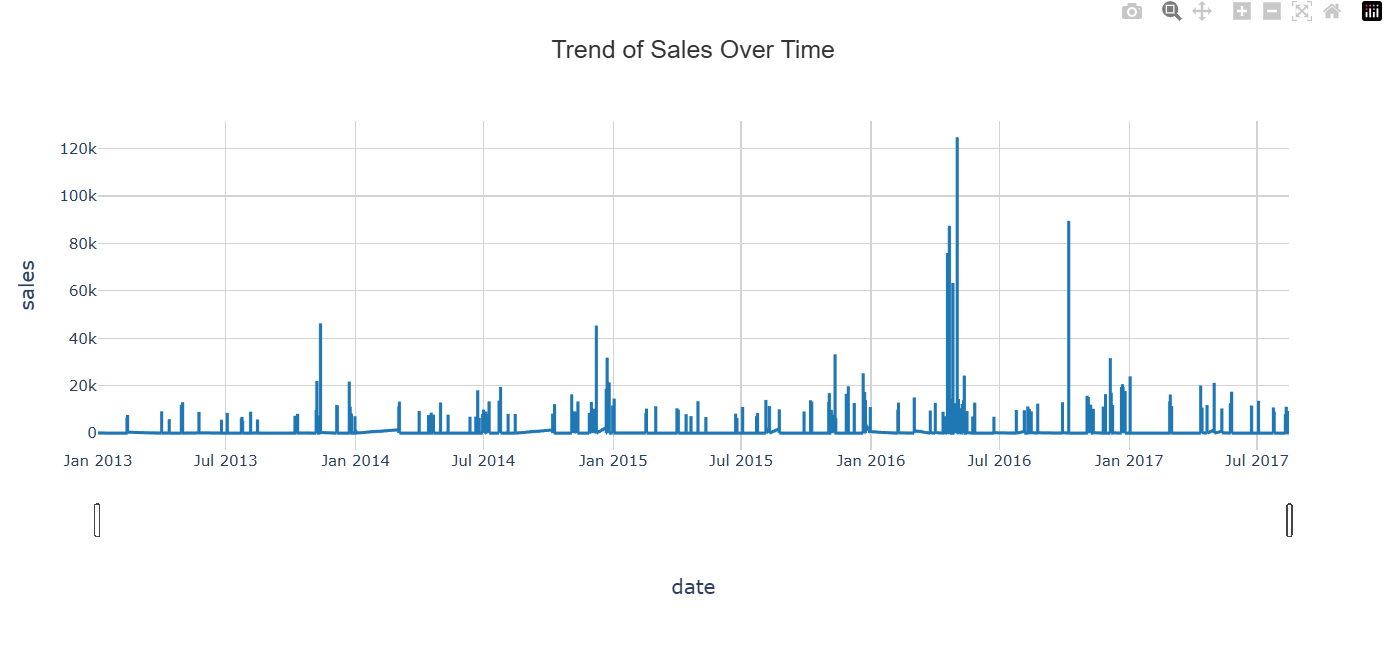
The histogram and boxplot of the sales variable provide insights into its distribution. The histogram shows the frequency distribution of sales values. It reveals that the majority of sales fall within a specific range, represented by the peak in the histogram. However, there are also instances of higher sales values, leading to a right-skewed distribution. This skewness suggests that there are relatively fewer occurrences of very high sales, while the majority of sales are concentrated around lower values. The boxplot further confirms the presence of outliers in the data, as indicated by the points beyond the whiskers.

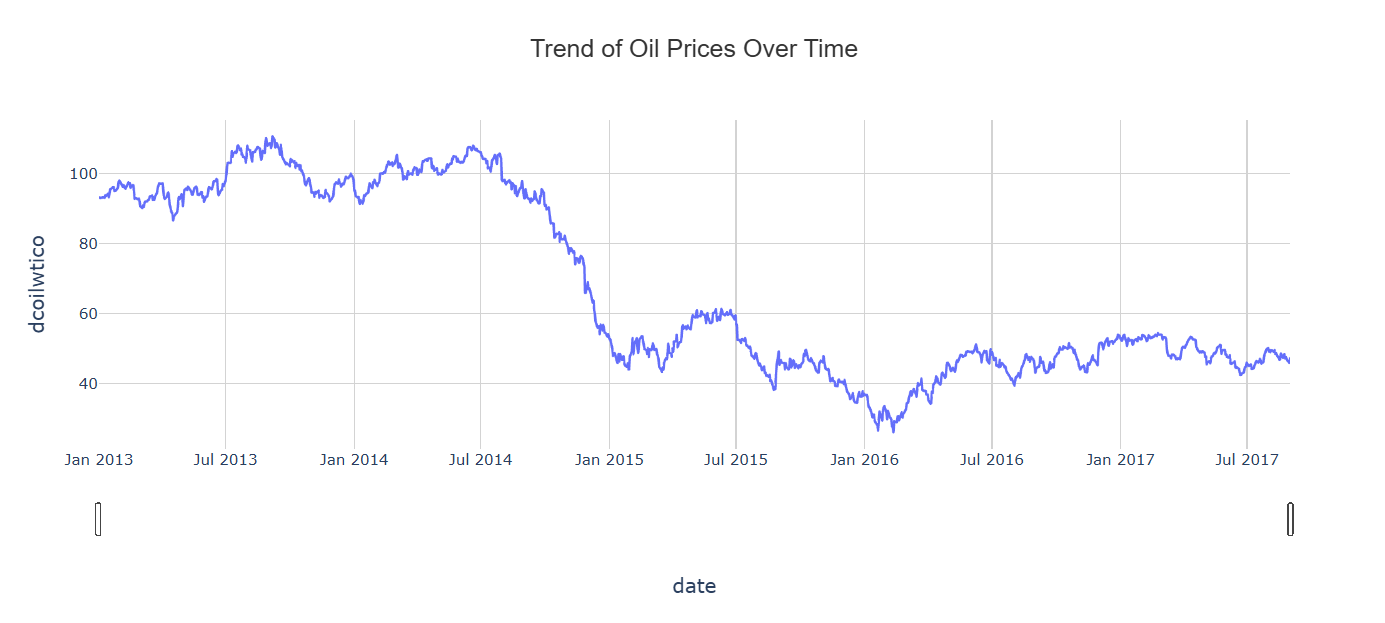


This histogram provides insights into the distribution of transactions in the dataset. The shape of the histogram indicates that the majority of transactions fall into a specific range, which is evident from the high frequency observed on the left side of the histogram. As the transactions increase, the frequency gradually decreases, forming a right-skewed distribution. This suggests that there are relatively fewer instances of high transaction volumes. Overall, the histogram highlights the presence of a cluster of transactions with a lower frequency, indicating a pattern in the data.

The left plot shows a bimodal distribution of oil prices with major peaks around $45 and $100, suggesting two distinct economic periods. The KDE line highlights this dual concentration, possibly due to macroeconomic events like oil market shocks.  
The right-side boxplot illustrates a wide interquartile range (IQR), indicating high variability in oil prices. While there are no extreme outliers, the long whiskers on both sides confirm significant price dispersion, relevant for demand forecasting models that may depend on economic conditions.

### 4.2 Bivariate Analysis

The plot shows the trend of sales over time. From the graph, we can observe that the sales exhibit some variations and fluctuations over time. There are periods of both high and low sales, indicating potential seasonality or other factors affecting sales patterns.



We can see that there is a trend in the in the oil prices over time. We see that oil prices suffered a collapse towards the end of 2014 and have not recovered. In fact despite some volatility, oil prices are at the same level as they were in the beginning of 2015. As a result of this we may see a significant shift in store sales around late 2014. Looking at the unit sales data, this is not readily apparent. Although sales do appear to drop off in the early part of 2015, in late 2014 they are rising.

#### 

What it shows: The number of records (data points) for each store type.

Interpretation: This tells you how many entries exist for each store type in the dataset.

The analysis of the total count of sales by store type provides insights into the sales performance and customer demand across different store types. Store Type D stands out with the highest count of sales, suggesting a strong customer base and popularity of products offered. Store Type C follows with a relatively lower count of sales, indicating a significant customer base as well. On the other hand, Store Types A, B, and E have lower counts, suggesting potential areas for improvement or the need to address competition.

#### 

The total sales amount varies across different store types. Store Type D has the highest total sales, indicating its significant contribution to the overall sales. Store Type A follows closely behind, demonstrating its substantial sales performance. Store Type C ranks third in terms of total sales, while Store Type B and Store Type E have lower sales amounts. Understanding the variations in sales by store type helps identify the key drivers of revenue and highlights the importance of certain store types in driving overall sales.

#### 

The chart reveals notable disparities in average sales across cities. Quito and Cayambe lead significantly, suggesting they are key markets with higher consumer demand or economic activity. In contrast, Playas, Puyo, and El Carmen show the lowest average sales, indicating either smaller markets or reduced commercial activity. The mid-range cities like Guayaquil, Machala, and Esmeraldas represent moderate sales potential and might benefit from targeted strategies to increase market share. This distribution highlights potential opportunities for resource reallocation or market-specific planning.

#### 

The chart highlights clear variations in average sales across Ecuadorian states. Pichincha stands out as the top-performing state by a significant margin, indicating its dominant role in sales performance—likely driven by urban centers like Quito. Tungurahua and Azuay also show strong sales, suggesting robust regional demand. On the lower end, states like Pastaza, Cotopaxi, and Chimborazo exhibit relatively weaker sales performance, which may reflect limited commercial activity or population density. These insights can guide strategic decisions such as where to focus marketing efforts or optimize distribution networks.

#### 

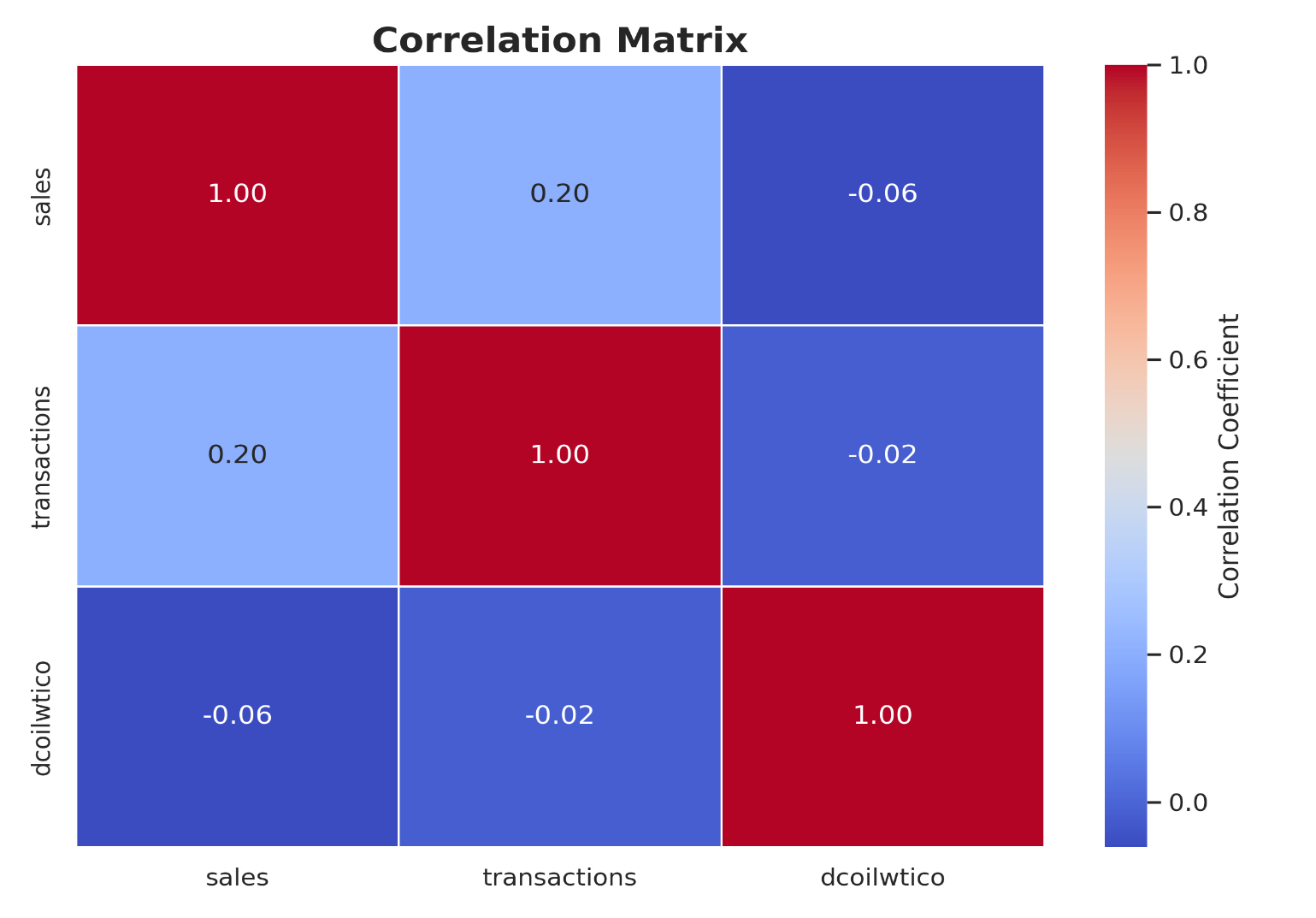
The scatter plot visualizes the relationship between sales and transactions in the dataset. Each data point represents a specific instance with corresponding sales and transaction values. Here are the key insights drawn from the scatter plot:

Clustered Data Points: The majority of data points cluster in the lower sales region, forming a specific concentration. This clustering suggests that there are certain transaction volumes that are consistently associated with particular sales levels. This concentration may indicate a common sales pattern or trend that occurs at specific transaction levels.

Outliers: Several data points deviate from the main cluster and are located at higher sales levels for relatively lower transaction volumes or vice versa. These outliers represent exceptional instances where sales are significantly different from what is typically observed for a given number of transactions. Identifying and understanding these outliers can provide valuable insights into unusual sales scenarios or exceptional business activities.

In summary, the scatter plot provides valuable insights into the relationship between sales and transactions. The clustering of data points around specific sales and transaction levels indicates the presence of common patterns. Additionally, outliers represent exceptional cases that warrant further investigation to understand the factors influencing sales and transactions in unique instances. This analysis can help businesses make informed decisions and devise effective strategies to improve sales performance.

### 4.3 Multivariate Analysis



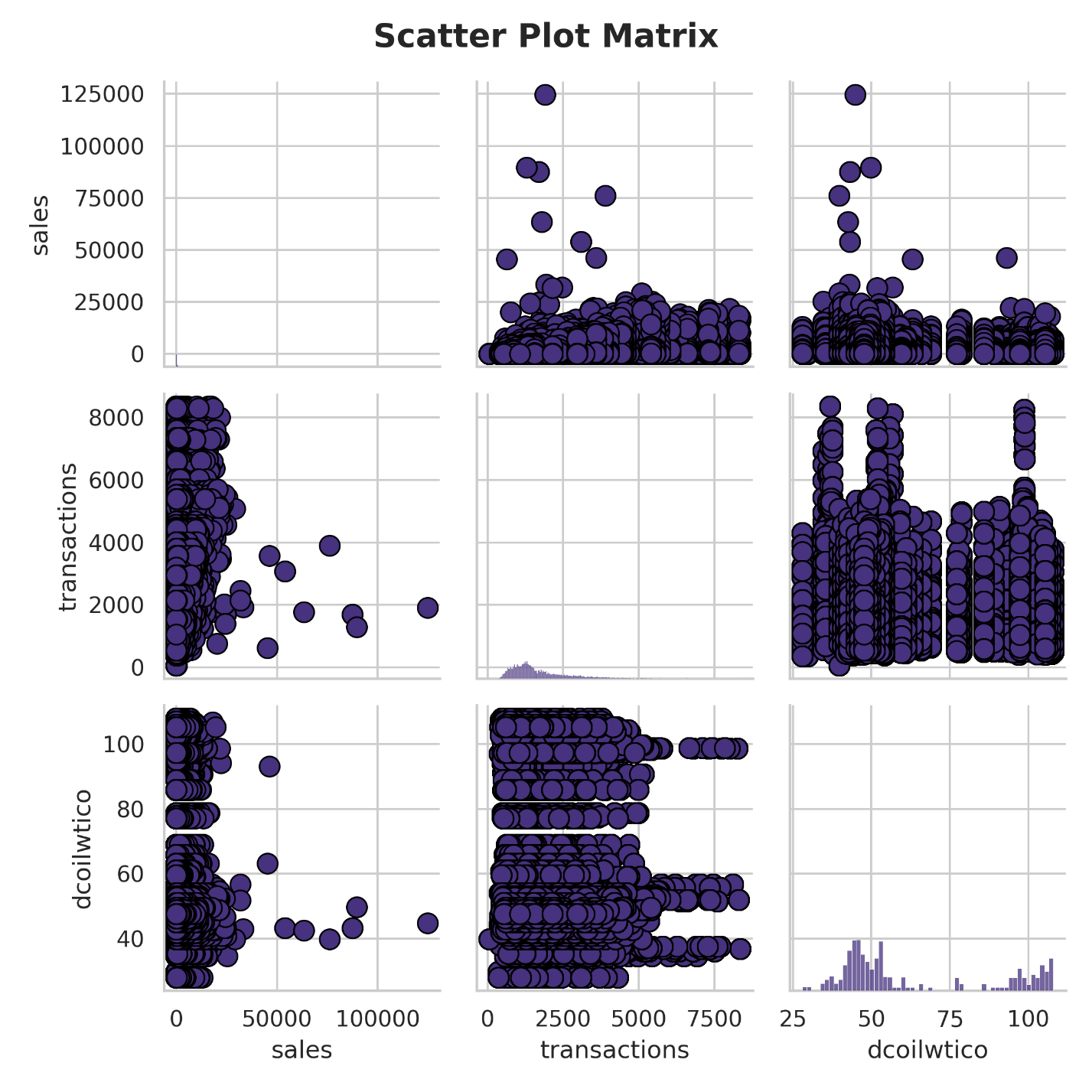
Correlation values range from -1 to 1, where -1 represents a perfect negative correlation, 1 represents a perfect positive correlation, and 0 represents no correlation. The table helps us understand how the variables are related to each other, providing valuable insights into their interactions. In this correlation matrix, we can see the correlations between different variables:

Sales and Transactions: There is a weak positive correlation of approximately 0.200 between Sales and Transactions. This suggests that there is a slight positive relationship between the number of transactions and the sales. It implies that when the number of transactions increases, there is a tendency for sales to increase as well, although the correlation is not very strong.

Sales and Dcoilwito (Oil Prices): There is a weak negative correlation of approximately -0.06 between Sales and Dcoilwito (Oil Prices). This indicates a slight negative relationship between sales and oil prices. It suggests that as oil prices increase, there is a tendency for sales to decrease slightly, though the correlation is not significant.

Transactions and Dcoilwito (Oil Prices): There is a very weak negative correlation of approximately -0.02 between Transactions and Dcoilwito (Oil Prices). This suggests that there is almost no relationship between the number of transactions and oil prices. It indicates that fluctuations in oil prices do not have a significant impact on the number of transactions.

Overall, the correlation values are relatively low, indicating that the relationships between these variables are not very strong. Other factors not considered in this correlation matrix may also influence sales, transactions, and oil prices. It's essential to explore additional factors to gain a more comprehensive understanding of their impact on sales and transactions.

The observations of the scatter plot matrix corroborate the observations from the correlation matrix

## **5. Conclusion:**

This milestone successfully established a clean, well-understood, and integrated dataset suitable for training forecasting models. Key accomplishments include:

1-Consolidation of six disparate datasets into one coherent structure.

2-Handling of all missing values and memory inefficiencies.

3-Initial discovery of seasonal, promotional, and economic effects on sales.

4-Confirmation of key business hypotheses that will guide modeling logic.

This foundational work enables us to confidently proceed to Milestone 2, where we will engineer predictive features and apply statistical analysis to further enrich the dataset.