Seif Eldin Omar - Ali Sherief - Mohammed Rabee - Mohammed Walid - Abdelhamid Eisa -
Hana Nabhan

Sales Forecasting and Optimization: Data Exploration Report

1. Introduction

The **Sales Forecasting and Optimization** project aims to predict future sales trends and optimize related business processes. By analyzing historical sales data and identifying key factors, the goal is to enhance decision-making, streamline operations, and improve profitability.

Accurate sales forecasting is essential for informed decisions on inventory management, pricing strategies, and marketing efforts. This project will develop a forecasting model using data such as product pricing, promotions, seasonality, and regional demand variations to provide actionable insights for optimizing sales performance.

The ultimate goal is to use data-driven insights to improve sales efficiency, reduce costs, and better align resources with consumer demand.

2. Data Overview

Data Source: The dataset is publicly available on Kaggle as part of the **Store Sales Time Series Forecasting** competition. It includes historical sales data for stores in Ecuador, along with external factors such as oil prices and holidays that may influence sales.

Dataset Description:

Size: The dataset contains **3,054,348 rows** across multiple files.
Time Period: Data spans from 2013 to 2017, with both training and test data available for forecasting.
Files:

- `train.csv`: Contains the sales data for the training set.
- `stores.csv`: Includes metadata about the stores.
- `oil.csv`: Provides daily oil prices.
- `holidays_events.csv`: Contains data on holidays and special events.

Features/Columns:

1. `train.csv` (Sales Data):

- **store_nbr**: Identifies the store where the sales occurred.
- **family**: Identifies the product family (e.g., dairy, meat, etc.).
- **sales**: The total sales for a product family at a store on a specific date. Fractional values are possible as products can be sold in fractional units.

- **onpromotion**: The total number of items in a product family promoted at a store on a given date.

2. `stores.csv` (Store Metadata):

- **store_nbr**: as the index of the table
- **city**: The city where the store is located.
- **state**: The state where the store is located.
- **type**: The type of store (e.g., supermarket, hypermarket).
- **cluster**: A grouping of similar stores based on certain features.

3. `oil.csv` (Oil Prices Data):

- **date**: The date when the oil price is recorded.
- **dcoilwtico**: Daily oil prices. Given Ecuador's dependence on oil, these prices could have a significant impact on sales.

4. `holidays_events.csv` (Holidays and Events Data):

- **date**: The date of the holiday or event.
- **type**: The type of holiday or event (e.g., public holiday, regional holiday).
- **locale**: The location of the holiday (either 'national', 'regional', or 'local').
- **locale_name**: Name of the region or locality.
- **description**: Description of the holiday or event.

5. `transactions.csv` (Holidays and Events Data):

- **store_nbr**: Identifies the store where the sales occurred.
- **transactions**: Identifies the store where the sales occurred.

Target Variable:

- **sales**: The target variable is the total sales for each product family in a store on a given date, which will be used for forecasting future sales.

3. Data Cleaning and Preprocessing

Missing Values:

- **Holiday-related Columns**: Columns like `holiday_type`, `locale`, `locale_name`, `description`, and `transferred` have missing values. We will analyze and decide whether to impute or ignore these based on their relevance.
- **Oil Prices**: The `dcoilwtico` column has missing values, which will be handled with appropriate imputation methods (e.g., forward-fill or rolling average).
- **Transactions**: Missing values in the `transactions` column will be assessed and imputed if necessary.

Duplicates:

- We have verified that there are no duplicates in the dataset, so no further action is needed in this regard.

Outliers:

- **Sales**: There are 2,788,335 outliers in the `sales` column, indicating extreme sales events. We will review these instances and decide whether to transform or cap the outliers, depending on their relevance.
- **Transactions**: There are 154,308 outliers in the `transactions` column, suggesting abnormal transaction counts. These will be investigated for potential removal or transformation.

Data Types and Formatting:

- We will ensure that columns like `date`, categorical columns, and numerical columns are in the correct formats (e.g., `datetime`, categorical types, and numeric types).

Feature Engineering:

- **Time-based Features**: We will use existing features like `day_of_week`, `month`, and `year`, and create a binary `is_holiday` feature.
- **Lag Features and Days Since Last Promotion**: Additional features capturing trends over time and promotional events will be created to improve forecasting accuracy.

4. Exploratory Data Analysis (EDA)

Univariate Analysis:

Numerical Variables Summary Statistics:

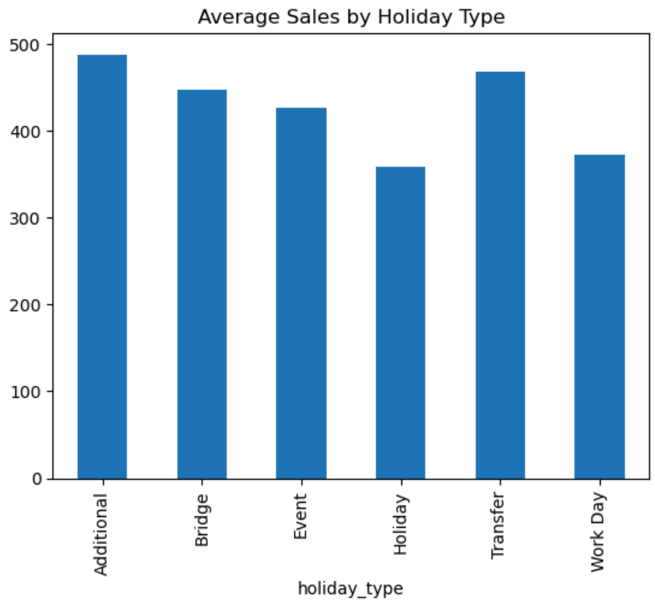
Here are the key summary statistics for the numerical features:

- **Sales**:
  - Mean: 359.02
  - Min: 0.0
  - Max: 124,717.0
  - Standard Deviation: 1,107.29
  - 25th Percentile: 0.0
  - 50th Percentile (Median): 11.0
  - 75th Percentile: 196.01
- **Onpromotion**:
  - Mean: 2.62
  - Min: 0.0
  - Max: 742.0
  - Standard Deviation: 12.25
  - 25th Percentile: 0.0
  - 50th Percentile (Median): 11.0
  - 75th Percentile: 196.01

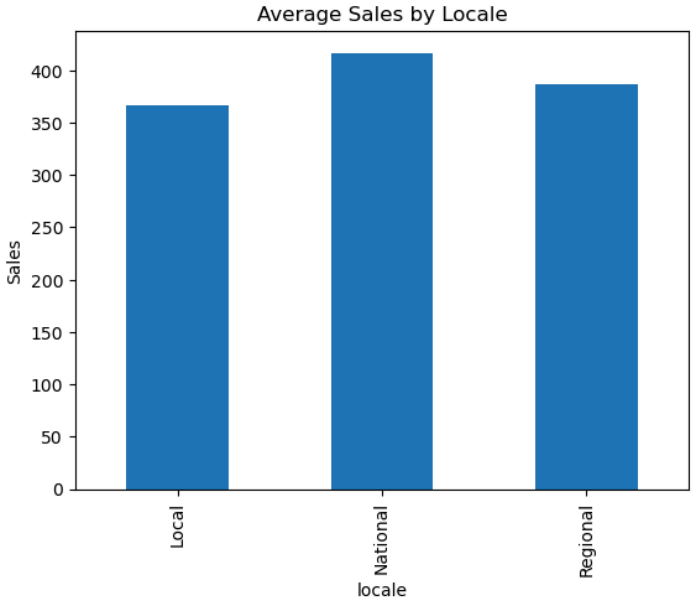
- **Transactions**:
  - Mean: 8,481
  - Min: 5
  - Max: 8,359,000
  - Standard Deviation: 9,668
  - 25th Percentile: 1,046
  - 50th Percentile (Median): 1,395
  - 75th Percentile: 2,081
- **Oil Prices (`dcoilwtico`)**:
  - Mean: 61.39
  - Min: 26.19
  - Max: 110.62
  - Standard Deviation: 25.69

Frequency distributions for categorical variables:

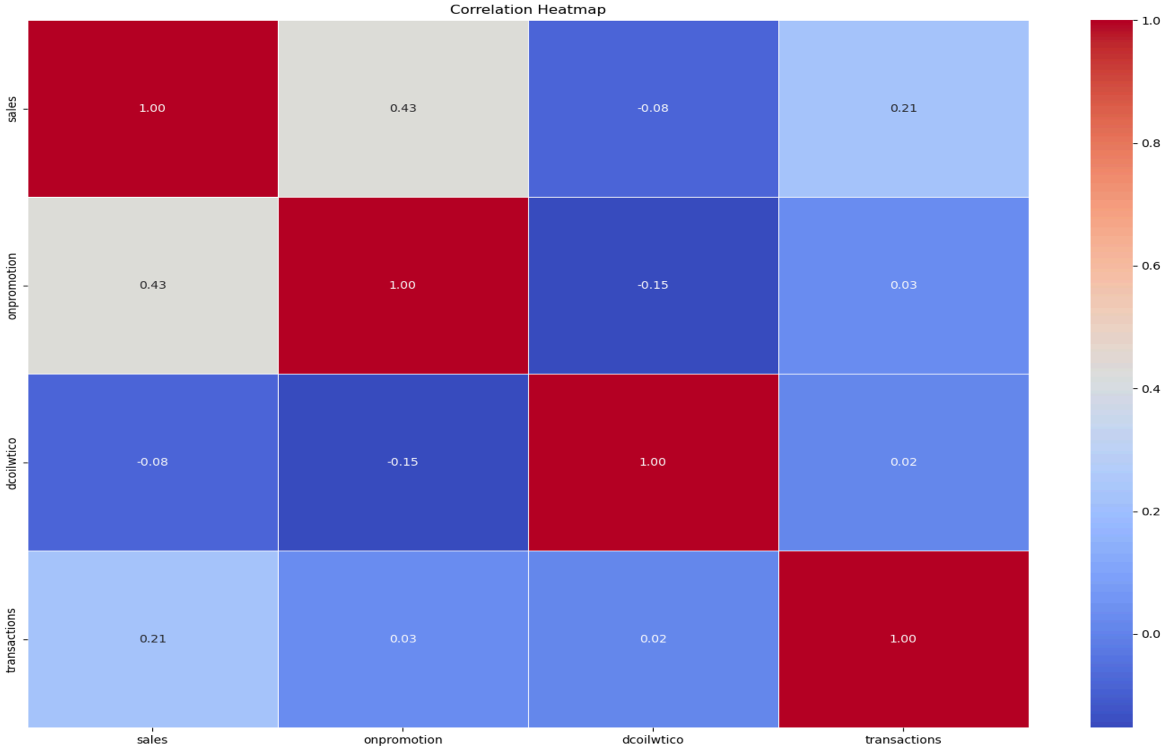
- **Histogram for the distribution of the holiday types**



- Histogram for the distribution of the locale of the holiday



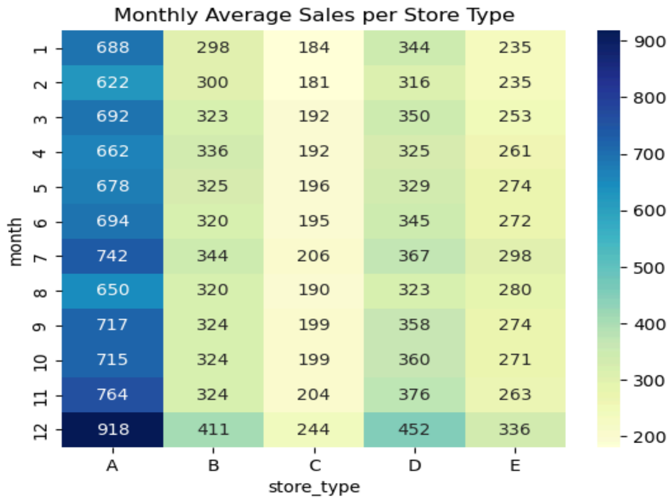
Bivariate Analysis:



- **Correlation Insights**:
  - **Oil Price** Correlation with sales is -0.08, indicating a very weak negative relationship. This suggests that as oil prices increase, sales tend to decrease slightly — but the effect is minimal. Despite the low linear correlation, oil

    prices may still contribute to model performance in more complex, non-linear models or during specific economic events.
    - **Transactions**: Correlation with sales is 0.21, showing a weak positive relationship. This is expected, as more transactions generally mean more sales. Although not strongly correlated linearly, this feature should be kept, especially since it directly reflects customer activity.
    - Both features will be retained for modeling, and further feature engineering (e.g., lags, rolling statistics) may enhance their usefulness.
-

Multivariate Analysis:



- This visualization demonstrates the monthly average sales

5. Trend Analysis

Impact of Holidays, Promotions, or External Factors

- **Holidays**: Analyze how sales are impacted by holidays and whether there are noticeable peaks in sales during specific events or holiday periods.
- **Promotions**: The analysis of **onpromotion** can help determine whether specific promotions (discounts, sales events) significantly affect sales.
- **External Factors**:
    - **Wages**: In Ecuador, public sector wages are paid every two weeks on the 15th and on the last day of the month. Supermarket sales could be affected by these pay periods, as consumers may have more disposable income and therefore increase their spending on food and other products around these dates.

    - **Earthquake in 2016**: A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts, donating water and other essential products, which greatly affected supermarket sales for several weeks after the earthquake. This event might have caused unusual spikes in demand or shifts in purchasing behavior, especially in the weeks immediately following the earthquake.

Stationarity Analysis

- **Definition**: Stationarity refers to a time series where the statistical properties (mean, variance, autocorrelation) do not change over time. Non-stationary data can produce misleading results in modeling, as trends, seasonal patterns, and other variations can lead to incorrect predictions. Thus, stationarity tests help in determining whether data needs to be transformed (e.g., by differencing) before further analysis.
- **KPSS Test**: To test the stationarity of the sales data, we used the **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test**. This test's null hypothesis assumes that the series is stationary. A low p-value indicates that the series is non-stationary.
    - **KPSS Test Results**:
        - **KPSS Statistic**: 147.9957
        - **p-value**: 0.01
        - The low p-value (below 0.05) suggests that the series is non-stationary, meaning the data exhibits trends or seasonality.
    - **Interpretation**: Since the p-value is below the threshold of 0.05, we reject the null hypothesis and conclude that the sales data is **non-stationary**. This implies that trends or seasonality are present in the data.

Next Steps:

- **Differencing**: To transform the series into a stationary one, we can apply **differencing** (either first or second-order differencing) to eliminate the trend component.
- **Transformation**: We could also explore log transformation or other techniques to stabilize variance if needed.

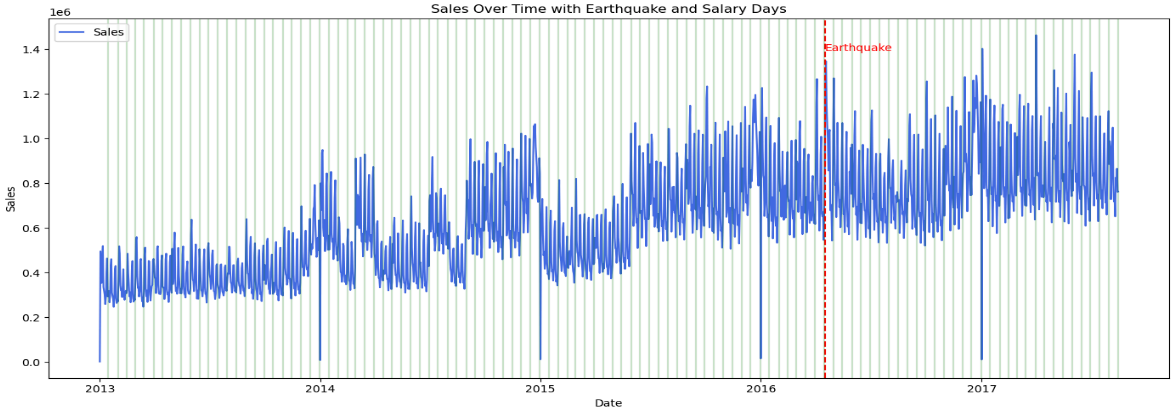
Seasonal Decomposition

We applied **additive seasonal decomposition** on the `sales` time series to better understand its underlying structure in terms of **trend**, **seasonality**, and **residuals**.

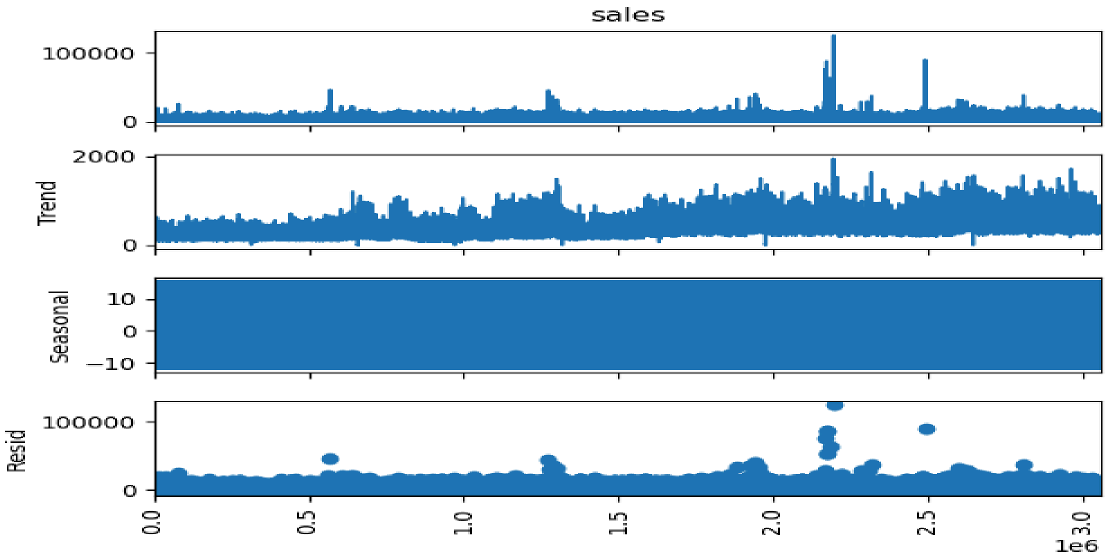
- **Trend Component**:
  The trend line shows a **gradual upward trajectory**, reflecting an overall **increase in sales over time**, with several spikes likely linked to external events (e.g., promotions, holidays, or emergencies like the April 2016 earthquake). This indicates growth in demand or store expansion across the period.
- **Seasonality Component**:
  The **seasonal component appears flat and non-cyclic**, suggesting that **strong periodic seasonality is not present** at the yearly level (period = 365). This might be due to:
  - The aggregation level (daily sales across many stores and products).
  - Non-uniform seasonality patterns that differ by region/store/product.
  - External events dominating regular seasonal patterns.
- To better capture seasonality, **shorter periods (e.g., 7 for weekly or 30 for monthly)** could be explored. The flat line here implies that either the seasonal effect is minimal or masked by more dominant trends and outliers.
- **Residual Component**:
  The residuals reveal **sharp spikes**, especially around certain periods. This suggests **significant irregular variations**, likely caused by **external shocks**, such as:
  - Earthquake (April 2016).
  - Bi-weekly salary cycles (15th and end of each month).
  - Major promotions or holidays.
- These residuals are important for understanding the **unexpected fluctuations** in sales that the trend and seasonality don't explain.

Visualizations for Trend Analysis

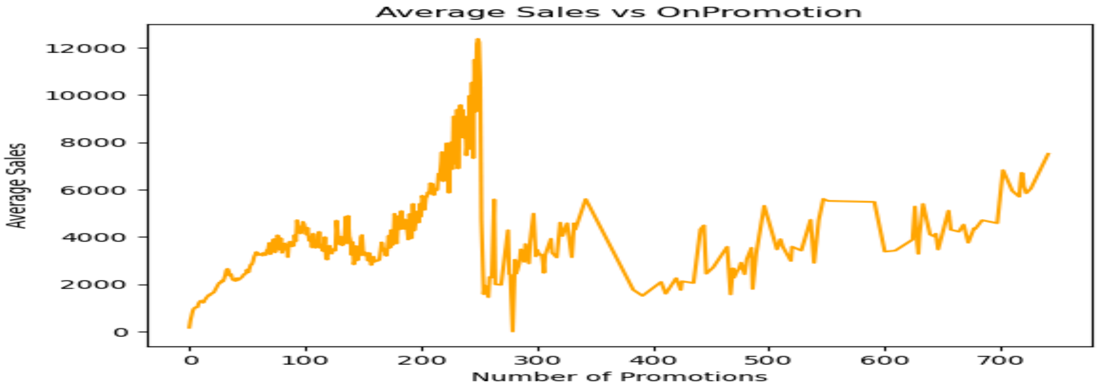
- **Sales over Time**:



- **Seasonality Patterns**:



- **Relation between number of promotion and average sales:**


Average Sales vs OnPromotion
Average Sales
Number of Promotions