Data Glacier Internship – Project

Batch: LISUM41 (30 December, 2024 – 30 March, 2025)

Team Member Details:

Name: Mahnoor Farhat

Email: mahnoor.farhat@gmail.com

Country: United Kingdom

College: University of Hertfordshire (Sep 2023 – Oct 2024)

Specialization: Data Science

Problem Description:

The time series data showed a range of patterns, some with trends, some seasonal, and some with neither. At the time, they were using their own software, written in-house, but it often produced forecasts that did not seem sensible. The beverage company wanted to explore power of AI/ML based forecasting to replace their in-house local solution.

What type of data you have got for analysis?

Column Name	Description	Data Type	Example
Product	Product identifier (SKU)	Categorical (String)	SKU1
Date	The date of sales record	Date (String, needs conversion to DateTime)	02/05/2017
Sales	Total sales for the given date	Numerical (Integer)	27,750
Price Discount (%)	Discount applied to the product	Numerical (Percentage, needs conversion to float)	17%
In-Store Promo	Whether the product had an in-store promotion	Categorical (Binary: 0 = No, 1 = Yes)	1
Catalogue Promo	Whether the product was in a catalogue promotion	Categorical (Binary: 0 = No, 1 = Yes)	0
Store End Promo	Whether the product was on a store-end promotion	Categorical (Binary: 0 = No, 1 = Yes)	1

Google_Mobility	External factor indicating mobility trends	Numerical (Integer)	0
Covid_Flag	Indicates if the sales period was affected by COVID	Categorical (Binary: 0 = No, 1 = Yes)	0
V_DAY	Whether the date falls on Valentine's Day	Categorical (Binary: 0 = No, 1 = Yes)	1
EASTER	Whether the date falls on Easter	Categorical (Binary: 0 = No, 1 = Yes)	0
CHRISTMAS	Whether the date falls on Christmas	Categorical (Binary: 0 = No, 1 = Yes)	0

What are the problems in the data?

- Missing Values (NA): Need to check if any missing values exist in Sales, Discount, and Google_Mobility, as they are crucial for demand forecasting.
- Outliers:
    - Sales: Sudden spikes might indicate an outlier.
    - Price Discount (%): Unusual discount values should be validated.
- Skewness:
    - Sales Distribution: Likely right-skewed due to occasional high-demand weeks.
    - Discount: May be skewed depending on the frequency of high-discount periods.
- Inconsistent Date Format:
    - The date column contains different formats e.g., "2/19/2017" vs. "02/05/2017".
    - This can cause errors in time series modeling.

What approaches you are trying to apply on your data set to overcome problems and why?

To ensure accurate analysis and forecasting, I will apply the following data preprocessing approaches to address potential issues:

- **Handling Inconsistent Date Formats:** The date column contains inconsistent formats (e.g., 02/05/2017 vs. 2/19/2017). To fix this, I will convert the date column to a standardized datetime format (YYYY-MM-DD) using pandas.to_datetime(). This ensures uniformity for time-series analysis and prevents errors in sorting, filtering, and grouping by date.
- **Converting Percentage Values to Numeric Format:** The Price Discount (%) column contains percentage values as strings (e.g., 17%, 44%). I will remove the % symbol and convert these values to a float (17% → 17.0). This conversion allows proper numerical operations such as correlation analysis and predictive modeling without text-related processing errors.

- **Handling Missing Values (NA Values):** Some records may have missing values, particularly in columns like Google_Mobility. For numerical columns, I will use mean or median imputation, while for categorical columns, I will replace missing values with the mode or a placeholder such as "Unknown". This approach prevents model bias due to missing data while preserving the dataset's overall integrity.
- **Identifying and Treating Outliers:** I will use box plots and interquartile range (IQR) analysis to detect and evaluate these anomalies. If the outliers are errors, they will be removed or corrected, but if they are valid due to factors like promotions, they will be retained for further insights. This approach ensures reliable analysis by preventing skewed interpretations while maintaining real business trends.
- **Standardizing Promotion Indicators:** The dataset includes multiple binary columns for promotions (In-Store Promo, Catalogue Promo, Store End Promo). I will either keep them separate or aggregate them into a single feature representing total promotional impact. This helps simplify feature selection while ensuring all promotion types are considered in predictive modeling.
- **Checking for Skewness in Sales Data:** Sales data might be skewed due to seasonal fluctuations. If skewness is detected, I will apply log transformation or scaling techniques such as Min-Max Scaling or Standardization. Normalizing the data improves the performance of regression models and forecasting algorithms.
- **Time-Series Feature Engineering:** The dataset lacks explicit time-based insights. I will extract new features like weekday/weekend indicators, month and year extracts, and rolling sales averages (e.g., 7-day and 30-day moving averages). These features will help identify seasonality and trends, ultimately improving forecasting accuracy.