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

GitHub: https://github.com/mahnoor-farhat/data-glacier-project

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.

Data Cleansing Summary

1. Data Import and Initial Inspection:

- Loaded the dataset using pandas.read_csv().
- Inspected data using df.head(), df.info(), and df.describe().
- Identified column types and missing values.

2. Data Type Conversion:

- Converted the date column to datetime format for better handling of time-series analysis.

3. Handling Missing Values:

- Used **mean imputation** for missing values in the Sales column.
- Removed the percentage sign from Price Discount (%) and converted it to a float.
- Applied **KNN Imputation** to fill missing values in Price Discount (%).

4. Handling Outliers:

- Used **Z-score method** to remove data points with z-scores greater than 3.
- Used **Interquartile Range (IQR) method** to remove outliers in Sales.

5. Data Cleaning for Categorical Columns:

- Applied **text normalization** to the Product column by removing special characters and converting to lowercase.
- Used **Label Encoding** to transform categorical Product values into numerical values.