Data Science – Retail Forecasting

Group Name	RFuMFM (Retail Forecasting Using Combined ML and Deep Learning Multivariate Forecasting Models)
Name	Tanmay Potbhare
Email	tupotbhare@gmail.com
Country	Ireland
College	Dublin City University
Specialization	Data Science

Problem Description

The large company who is into beverages business in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed forecast of each of products at item level every week in weekly buckets.

Data Understanding

1. **Data Collection**

2. **Data Description**
   - The data is all numerical with 12 columns namely 'Product', 'date', 'Sales', 'Price Discount (%)', 'In-Store Promo', 'Catalogue Promo', 'Store End Promo', 'Google Mobility', 'Covid Flag', 'V_DAY', 'EASTER', 'CHRISTMAS'. Product Column has 6 different products with 204 value counts for each product. Further, the data is divided as Sales, Price discount, Promotions, and Holiday.
   - The promotions include 'In-store Promo', 'Catalogue Promo', 'Store End Promo' which are already classified in 0 and 1 which is readable for machine learning and deep learning models. They have no missing values.
   - The Holidays include 'Covid Flag', 'V-day ', 'Easter', 'Christmas' which are also classified in 0 and 1 with no missing values.

3. **Data Quality**

```
In [4]: data.isna().sum()

Out[4]: Product              0
        date                 0
        Sales                0
        Price Discount (%)   0
        In-Store Promo       0
        Catalogue Promo      0
        Store End Promo      0
        Google Mobility      0
        Covid_Flag           0
        V_DAY                0
        EASTER               0
        CHRISTMAS            0
        dtype: int64
```

The data is the data frame I have defined and with getting the sum of null values, it can be said that there are no missing values in the data provided.

There are no missing values, but there are outliers which are found using Z Score with its formula mean divided by standard deviation and also some of the data are skewed not highly skewed but not normal as well. The outliers are considered for the threshold limit set greater than 3 and less than 3. So, using z score or imputing these data points I will be removing or transforming the outliers with mean, median or mode.

4. **What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?**
   I am using Z Score for detecting and removing or imputing the outliers as it is efficient method where you can set a good threshold limit and identify the outliers very well. The data where the outliers are seen are in sales and as the data is already classified, I will be removing the outliers. But if the percentage of outliers are more and if it will affect the data then imputing the data values will be a good fit.

GitHub Repo link - **https://github.com/TanmayPotbhare/Retail-Forecast-Using-Multivariate-Forecasting-Models/blob/main/Week%208_Deliverables.pdf**