**Part 1: Structured Data & ML Challenge**

**DATA PREPROCESSING**

The data was loaded and checked for any missing values and mismatched data types. The data summary statistics were carried out to have an understanding of the data. The data type error for date was corrected for. More insight could be derived from the date, therefore features such as: ‘day of the week’, ‘month’ and whether the day fell on a ‘weekend’ were derived. For the missing values, there missing categorical values were replaced with “Unknown” where as the missing numerical values were replaced by the median.

The insight from the summary statistics showed that:

* The range of customers range between 51-285 within the stores, with an average of 106 customers per day.
* The range of total sales range between 250.01 - 4202.36 within the stores, with an average of 1310.48 sales per day.
* The range of online sales range between 2.0 - 3371.68 within the stores, with an average of 525.70 sales per day.
* The range of instore sales range between 13.26 - 3178.17 within the stores, with an average of 784.13 sales per day. Performing better than the online store sales.
* The range of average transactions range between 2.62 - 27.86 within the stores, with an average of 12.477143 per day.

Before running Exploratory Data Analysis (EDA), I created histograms and boxplots from the numerical variables and the following insights were observed:

* Customer numbers at an interquartile range of 80-120.
* Total sales at an interquartile range of 1000 - 1500.
* Online sales at an interquartile range of 300 - 700.
* Instore sales at an interquartile range of 500 - 1000.
* Average transactions at an interquartile range of 10 - 15.
* Return rate ranging at an interquartile range of 0.04 - 0.06
* There is a balance between majority of the categorical variables: Categories, Weather, Promotions, Dominant Age groups and Store ID. However the special events vary greatly

I then run a Bivariate Analysis and created bar plots to show the relationships between the categorical variables and continuous variables. The following relationships were looked into:

* ***Weather Vs Total Sales, Weather Vs Online Sales, Weather Vs In Store sales:*** Sales performed better during the sunny weather across all sales. However, they performed better for in store sales as compared to online sales
* ***Promotion Vs Total Sales, Promotion Vs Online Sales, Promotion Vs In Store sales:*** Sales performed better during the BOGO promotions across all sales. However, they performed better for in store sales as compared to online sales
* ***Dominant Age group Vs Total Sales, Dominant Age group Vs Online Sales, Dominant Age group Vs In Store sales:*** No notable change between the age groups. However, they performed better for in store sales as compared to online sales
* ***Category Vs Total Sales, Category Vs Online Sales, Category Vs In Store sales:*** No notable change between the *categories*. However, they performed better for in store sales as compared to online sales
* ***Weekend Vs Total Sales, Weekend Vs Online Sales, Weekend Vs In Store sales, Day of week Vs Total Sales, Day of week Vs Online Sales, Day of week Vs In Store sales:*** Sales on the weekend, Saturday and Sunday, performed better than during the week. They performed better for in store sales as compared to online sales
* ***Month Vs Total Sales, Month Vs Online Sales, Month Vs In Store sales:*** Sales performed best in December and worst in November.

**CUSTOMER SEGMENT INSIGHTS**

It was noted that across all stores, there was a higher number of customers:

* On weekends, where more customers were registered on both Saturday and Sunday
* In December, where the number of customers were notably higher than all other months
* On Sunny days, the highest number of customers were recorded followed by Cloudy, Windy, Rainy and Snowing, as the weather conditioned worsened.
* Where promotions and special events were offered

It was noted that there was no great variance in customers numbers when the data was aggregated by categories, dominant age groups, or any other month excluding December.

**Strategy recommendations**

* Running promotions and special events during the weekdays.
* Running promotions and special events through online sales to encourage more online sale customers
* To attract more customers, there could be weekly promotions cutting across the various categories per day to encourage more awareness of products and thus attract some variance in customers by products.
* Given the challenges of weather, one could optimize sales during sunny weather by running promotions and special events. However, bring in a product that could encourage customers to shop and have their products delivered or kept in store for them to pick up once it’s convenient for them, at no extra cost. This could be successfully done by encouraging online shopping and free delivery and storage.

**SALES FORECASTING**

Upon the data being prepared, the categorical data was first encoded using the One-hot encode and a StandardScaler run across the numerical data. I then split the test data and training data, before running the model.