**RETAIL DATA ANALYSIS**

**Main Project Update**

**DATE: 11-06-2022**

The proposed system is used nowadays by shops in order to better predict the number of products, that might get sold and therefore to better estimate how much product should be produced. By the development of the system the organization can easily predict the weekly sales for a store and visualize the data for better insight. This should then help to optimize the manufacturing process and thereby help to increase income while lowering costs.

The dataset is from Kaggle. There are three different dataset that’s been used for this project. So far I have explored the dataset and visualised the given data for a better understanding. I combined the dataset into one which makes it into 16 features and 421570 samples. Checked for the null values. Converted few columns of Boolean and character values into integers. Expanded the “date” column and created many other columns for better understanding and now which makes a total of 70 columns in the dataset.

Split the dataset into testing, training and validation. Used iterative imputer for removing the missing data’s in the markdown 1 to 5. Splitting of x and y .y is weekly sales and x is the rest of the dataset. Trained the model using 4 different algorithm that is linear regression, decision tree, random forest and XGBoost. The platform used is sagemaker from AWS .