**APPROACH FOR SALES PREDICTION**

* **UPLOADING LIBRARIES**
* Firstly uploaded the libraries like pandas, numpy, matplotlib and seaborn.
* **IMPORTING DATASET**
* Uploaded the test and train data.
* **DATA CONCATENATION**
* The test data doesn’t consist of order and sales columns.
* Added the order and sales columns to test data by keeping constant variable as test.
* Combined both train and test data row wise by using concat statement .
* **DATA CLEANING**
* Made a check on head, tail, shape and datatypes of the combined dataset.
* Extracted year, month, date from date column and assigned them to three variables by using lambda statement.
* Appended the three extracted variables to the combined dataset.
* Dropped ID, Store\_id, Order, Date columns from combined dataset.
* Checked null values by using **isnull( ).sum( )** statement.
* **DATA PREPROCESSING**
* Divided combined dataset into numerical and object data.
* Divided the categorical data from numerical data and dropped the categorical columns from numerical data.
* Transferred sales and order columns from object data to numerical data.
* Dropped sales and order columns from object data.
* Made dummies for both object and categorical data by using **pd.get\_dummies(variable name )** statement.
* Concatenated numerical data, object dummies data, categorical dummies data column wise and named it as **cleandf**.
* Made a check on head, tail, and shape of cleandf.
* Divided cleandf data into train and test data.
* Dropped order and sales columns from the test data.
* Plotted the heatmap for both train and test data.
* Plotted the graphical representations like distplot, piechart, histograms, boxplot, countplot for variables like sales, Holiday, Region\_code and Discount.
* Made **Hypothesis test** for Holiday and sales columns using **groupby statement** and **ttest\_ind statement.**
* Made **chi-square test** for Store\_ type and Location\_type using **chi2\_contingency statement**.
* **MACHINE LEARNING MODELS**
* Divided the new train data into **independent variable(X)** and **dependent variable (y)**.
* The **dependent variable (y)** consists only **sales column**.
* The **independent variable (X)** consists **all the columns** **except sales column**.
* Sent **independent variable (X)** to machine learning models like **Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost** for model building by using **fit(X,y) Statement** .
* R-square (Accuracy) found for every model by using **score(X,y) statement**.
* Predicted the values for test data and converted to csv file.
* Predicted the values for independent variable(X) and named it as**(Xp).**
* Predicted the residual for every model by using **actual value(y) – predicted value(Xp) statement** and named it as **res.**
* Predicted the Root Mean Squared Error(RMSE) by using **np.sqrt(np.mean(res\*\*2))** statement