**All approaches which have been used in Model ::**

**1: Importing the Relevant Libraries**

**2: Data Inspection**

**3: Data Cleaning**

**4: Exploratory Data Analysis**

**5: Feature Engineering**

**6: Building Model**

**7: Hyperparameter Tuning**

**8: Model performance**

**1. Importing the Relevant Libraries:**

Here I have imported some libraries which are numpy, pandas, seaborn, metrics and some ML models etc.

**2. Data Inspection:**

Here I checked number of rows and columns, missing values, outliers, numerical and categorical columns.

As data has no missing values, All columns are categorical therefore data has no outliers except Order column.

**3. Data Cleaning:**

As we know this data has no missing values and outliers, that means the data is already cleaned and we can proceed further.

**4. Exploratory Data Analysis:**

All columns are categorical mostly that’s why we need to check the each categories of every column.

Categorical columns are: Store\_type, Location\_type, Region\_code, Holiday and Discount.

And we got some categories with respect to each column which

was highly repeated and which was least repeated.

**Relationship between categorical columns and Target column (Sales):**

Here we got some amazing results by exporing the relations between each categories of each column with target column Sales.

As we have seen in Code file:

i. Store S4 has maximum sales more than other stores

ii. Location L2 has maximum sales more than other locations

iii. Region R1 has maximum sales more than other regions

iv. Stores are having more sales on working days rather than holidays (obviously there are less hoildays in a week)

v. Stores which are providing discounts are having more sales.

**Relationship between numerical column and target column (Sales):**

As we have seen in Code file:

\* As #Orders are increasing, Sales are also increasing hence it's showing a positive correlation.

\* #Order column is not present in Test dataset that's why this column can be dropped in Train dataset otherwise we will get some unwanted errors while predicting the test dataset.

**5. Feature Engineering:**

This part is very crucial as we have all columns are categorical need to be converted into numerical columns then we will be able to use these columns in model building.

: Discount column is converted into numerical column for Train and Test Dataset.

* **Used Dummy Encoding for some categorical variables:-**

This dummy encoding worked very well and we got each separate column for each category in categorical variables. These all columns contain 0 and 1.

Dummy encoding has been used for both Training and Testing dataset.

* **Feature Importance using RandomForest:-**

Store\_S4, Discount, Location\_L2, Location\_L1, Store\_S2 and Holiday are most important features for predicting Sales.

**6. Model Building:**

Here we have tried many models together to get a generalised model and implemented **Decision Tree Regressor** model for model building, because it has least MSLE errors.

**7. Hyperparameter Tuning:**

It was used for improving model.

**8. Model Performance:**

Finally we predicted the Sales for Test dataset and it’s working well.