Capstone Project - 3

Retail Sales Prediction

By – Rishanshu Yadav

❑ CONTENT -

1. Problem Statement
2. Data Summary
3. Data Preprocessing
4. Exploratory Data Analysis
5. Feature Engineering
6. Model Implementation
7. Conclusion

❑PROBLEM STATEMENT -

- Rossmann operators over 3,000 drug stores in European countries . Currently , Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance . Store sales are influenced by many factors , including promotions , competition , school and state holidays , seasonality and locality . With thousands of individual managers predicting sales based on their unique circumstances , the accuracy of results can be quite varied.

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "SALES" column for the set . Note that some stores in the dataset were temporarily closed for refurbishment

❑DATA SUMMARY -

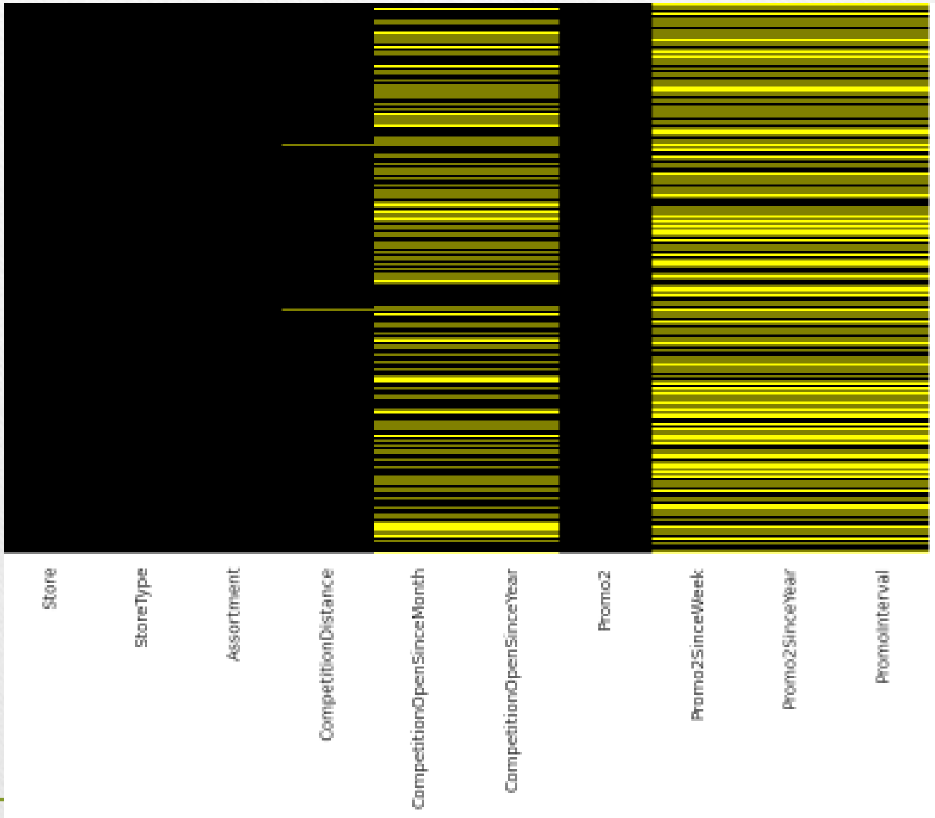
- We have two datasets. Rossman store data is for years 2013, 2014 and 2015 With 10,17,209 observations on 9 variables. Store data with 1115 observations on 10 variables

➢Some important features are -

1. Customer : - The number of customers on a given day in a store.
2. Date :- Showing dates for observations.
3. State Holiday :- Indicating a state holiday.
4. Store Type : Differentiate between 4 different store models (a,b,c,d).
5. Assortment : Describes an assortment level i.e a : basic, b : extra and c : extended.
6. Competition Distance : Distance in meters to the nearest competition store.
7. Promo :- Indicates whether a store is running a promo on that day

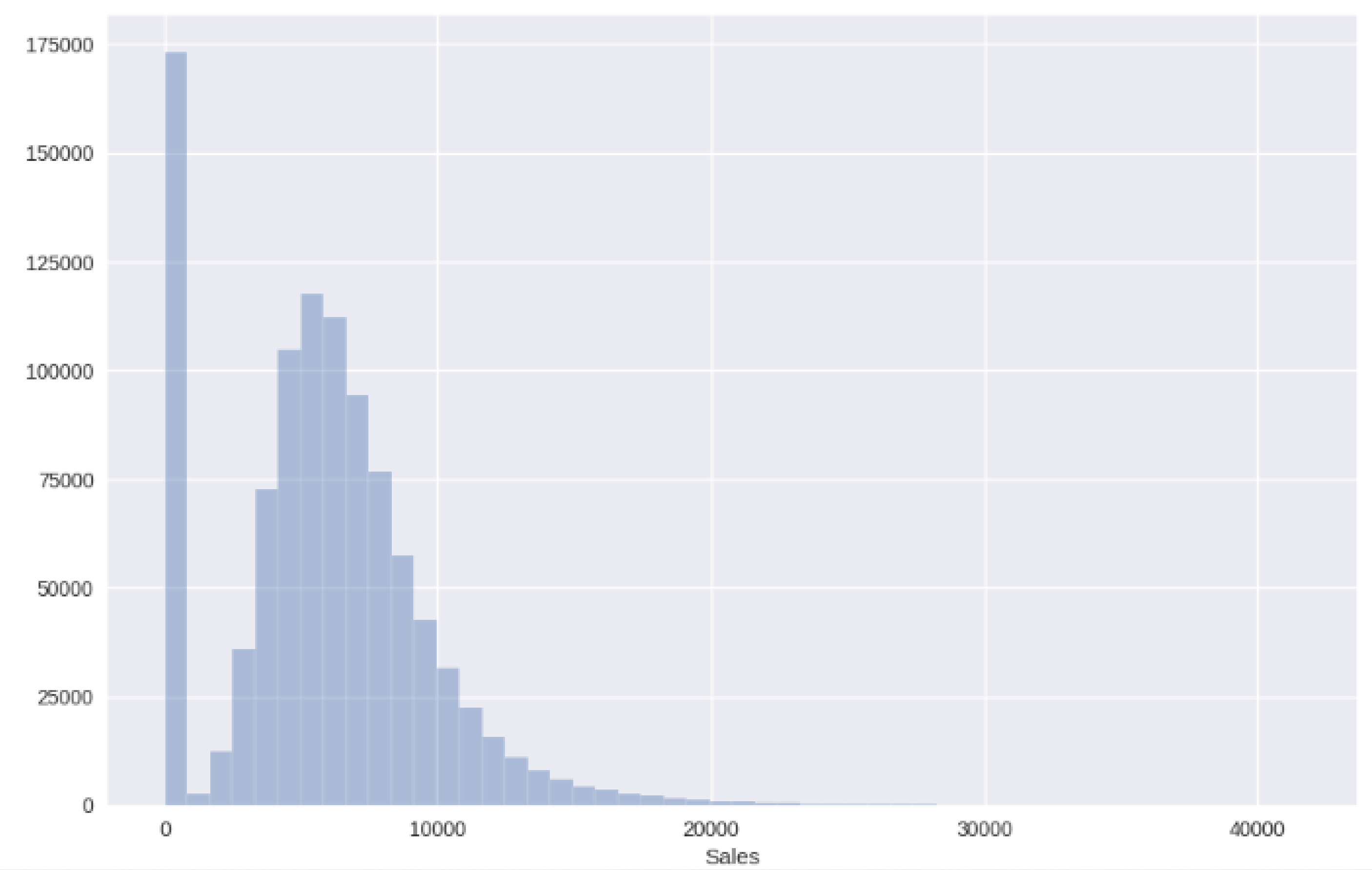
❑ DATA PREPROCESSING -

- Columns having >30% null values are dropped.
- Null values in ‘Competition Distance’ are imputed with median of feature.
- Removing those stores observations that are temporarily closed (~ 17.3K) & stores generating zero sales



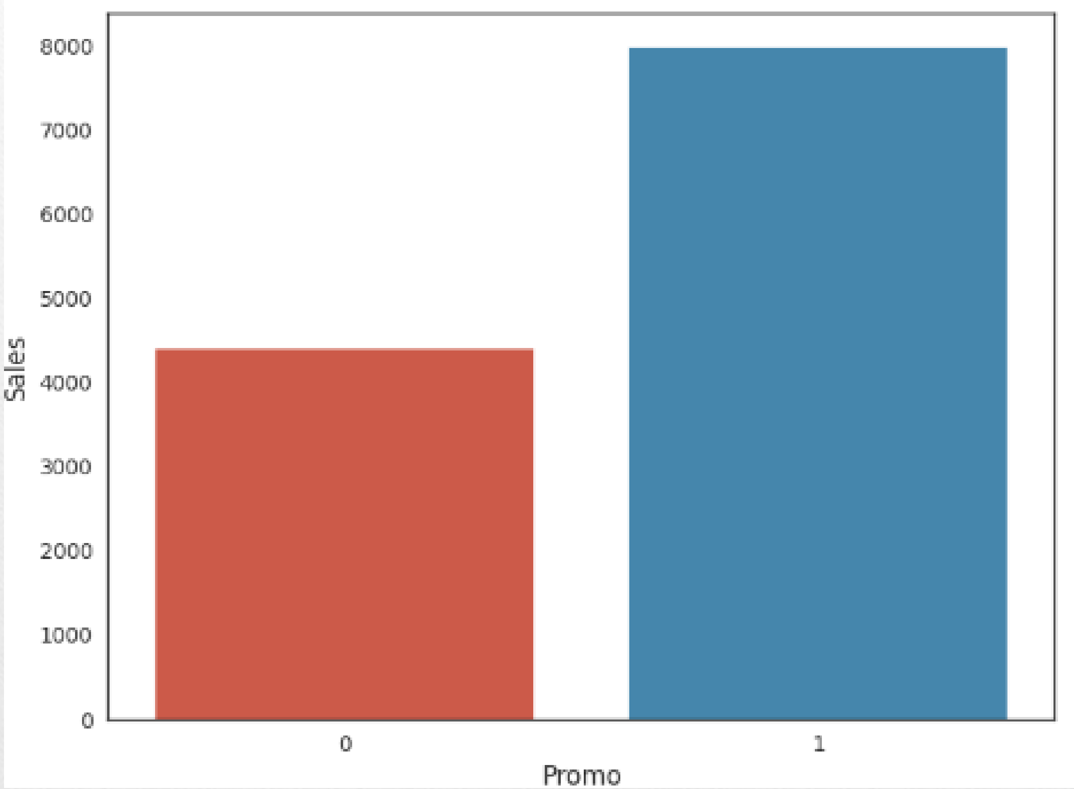
❑ EXPLORATORY DATA ANALYSIS (EDA) -

- ➢ Sales are normally distributed with slightly right tail skewed .



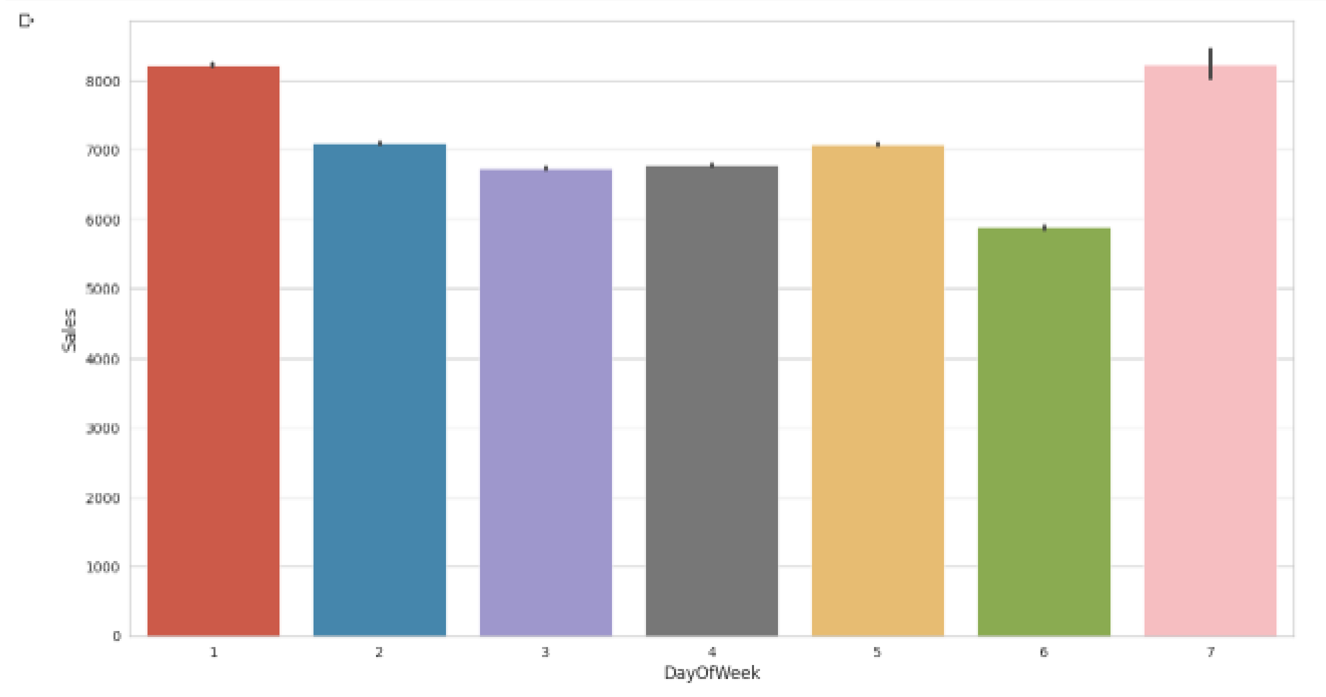
❑EDA (Contd ..)

- ❖ Impact of Promo on sales.



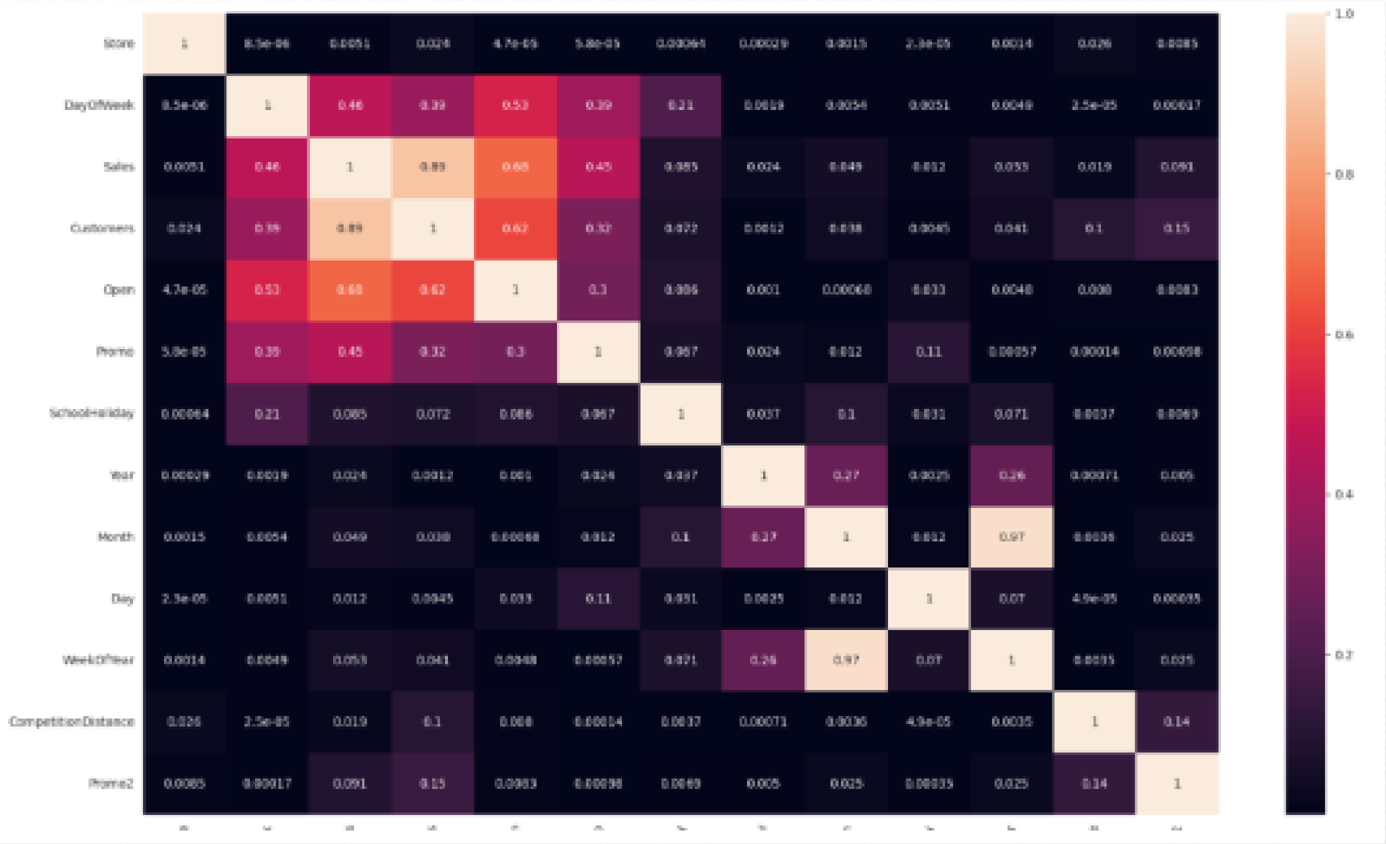
❑EDA (Contd ..)

- Day Wise trends in Sales.



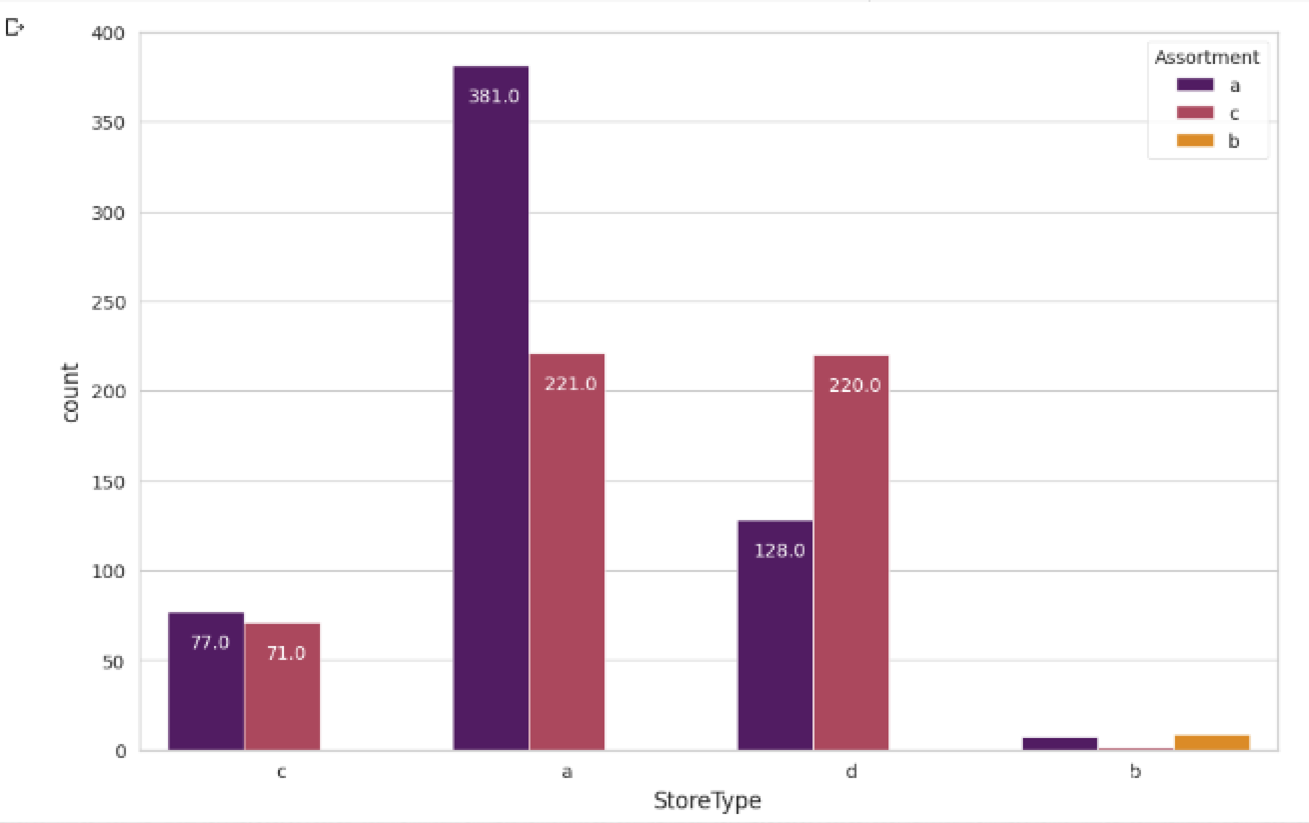
❑EDA (Contd ..)

- ❖ Heatmap for merged dataset



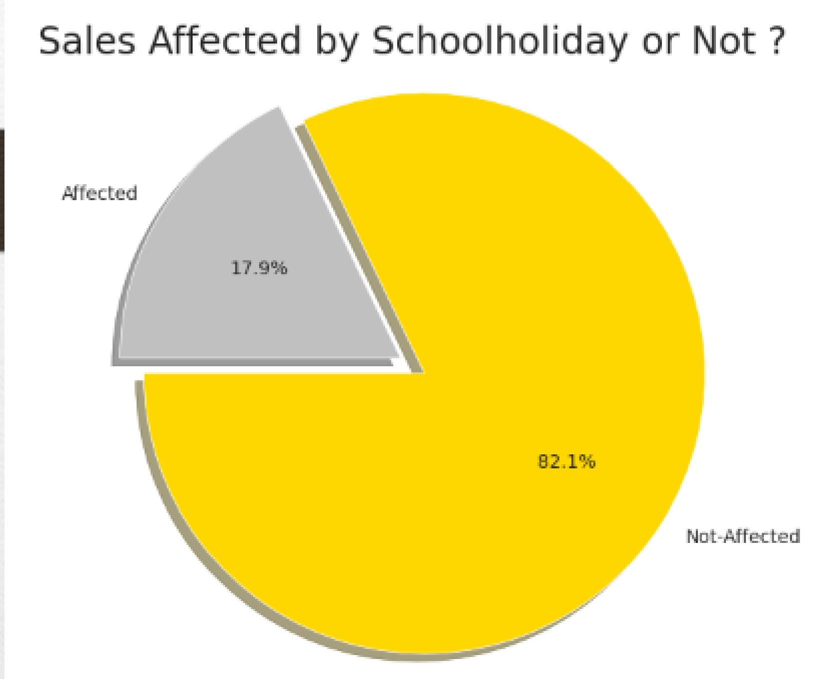
❑EDA (Contd ..)

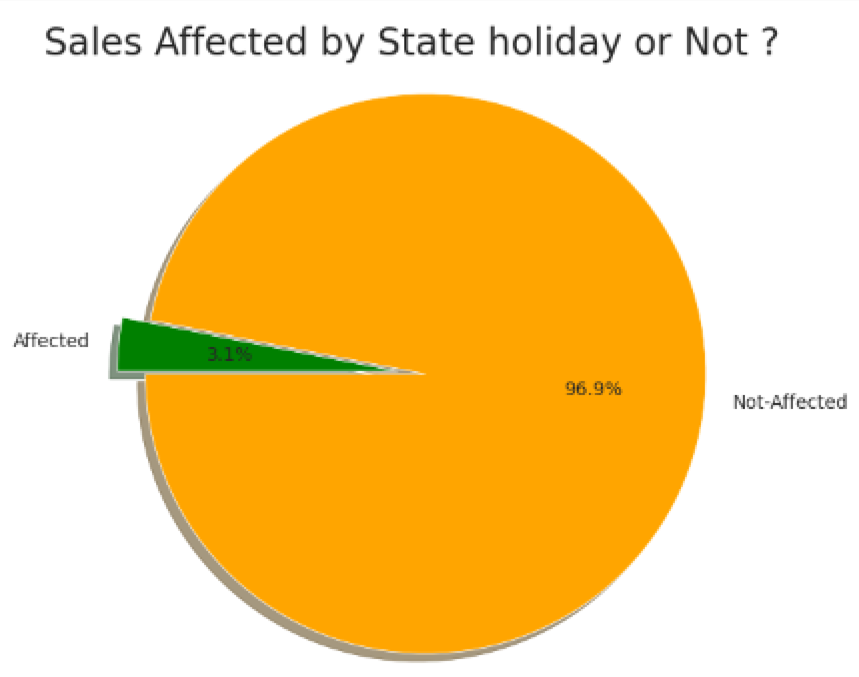
❖ Analysis of Store Types with their respective assortment.



❑EDA (Contd ..)

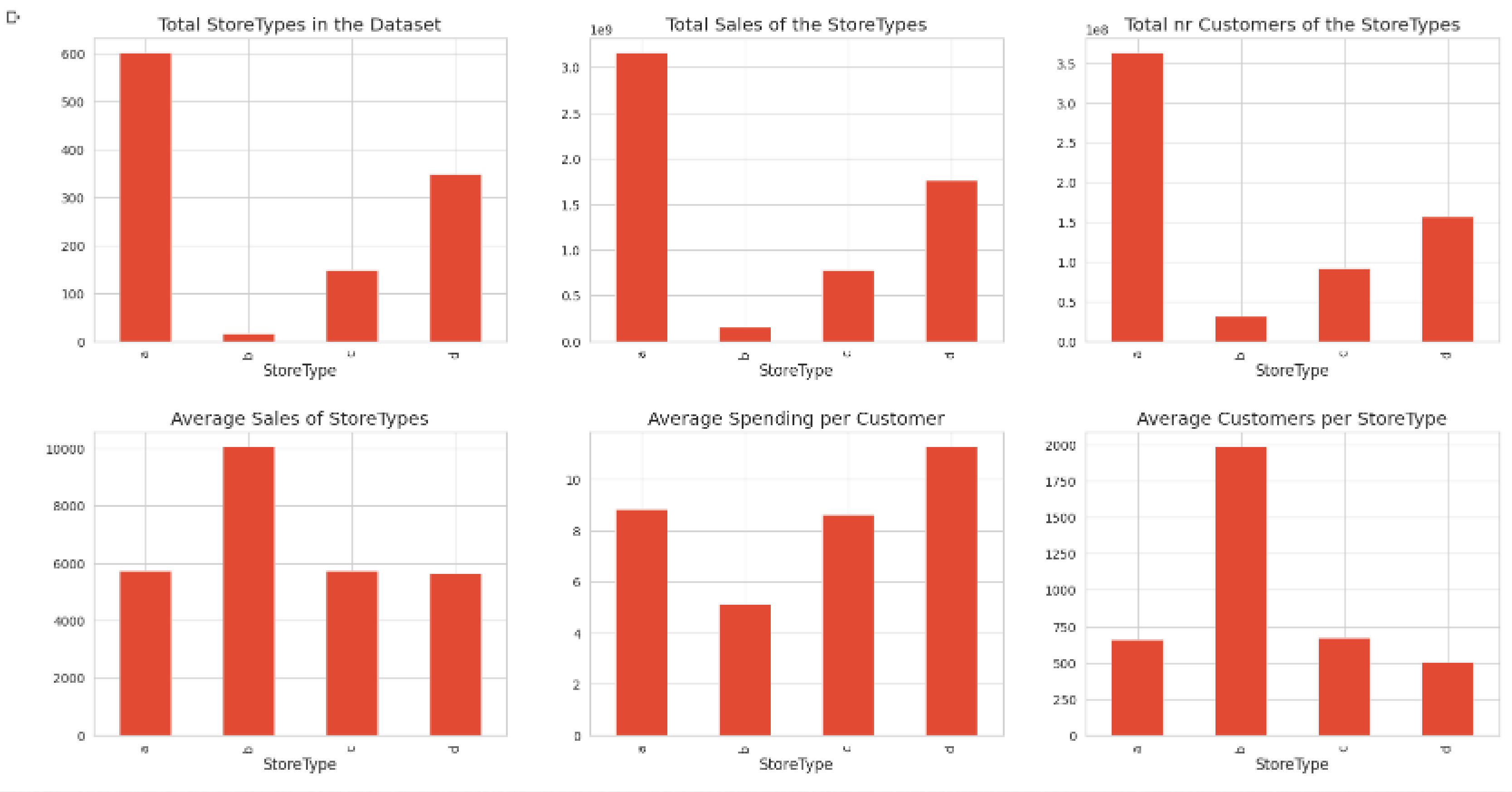
- School and State holidays effect on sales





❑EDA (Contd ..)

❖ Store Types and average sales/customer/spending relation.



❑ EDA (Summary) -

1. Sales are highly correlated to customers.
2. Stores opened on ‘State Holiday’ makes a good amount of sales.
3. There is no such significant difference in sales on ‘School Holidays’.
4. Even though store type ‘b’ has very less number of stores but these are outperforming other store types in terms of sales and avg customers.
5. Sales are consistent for the second quarter of the year but it starts increasing in the last quarter.

❑ FEATURE ENGINEERING -

1. Extracting week, month, year from Date and adding them in dataset.
2. Merging both dataset.
3. One hot encoding for Store type, Assortment.
4. Splitting dataset into Training and Test set and applying Min, Max Scaler for scaling dataset.

❑ MODELS IMPLEMENTED -

1. Linear Regression (Baseline Model)
2. Lasso Regression
3. Decision Tree Regression
4. Decision Tree Regression (with hyperparameters)
5. K-Nearest Neighbors Regression
6. Random Forest Regressor

❑ MODAL EVALUATION -

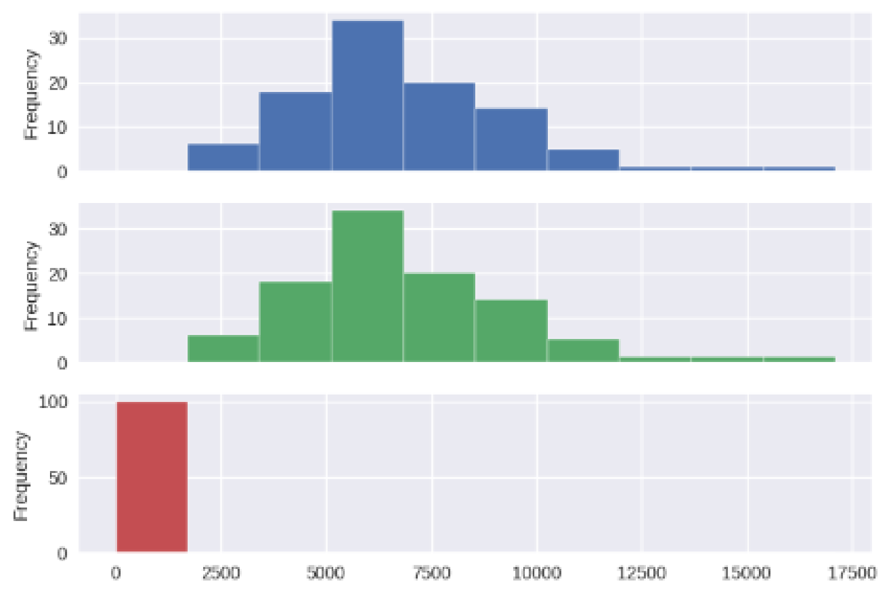
MODEL	TRAINING SCORE	TESTING SCORE
Linear Regression	0.780750	0.782392
Lasso Regression	0.780731	0.780769
Decision Tree Regression	0.99996	0.915942
Decision Tree Regression (with hyperparameters)	0.963506	0.935417
K-Nearest Neighbors Regression	0.73722	0.71665
Random Forest Regressor	0.993783	0.956520

❑Insights from Random Forest Regressor -

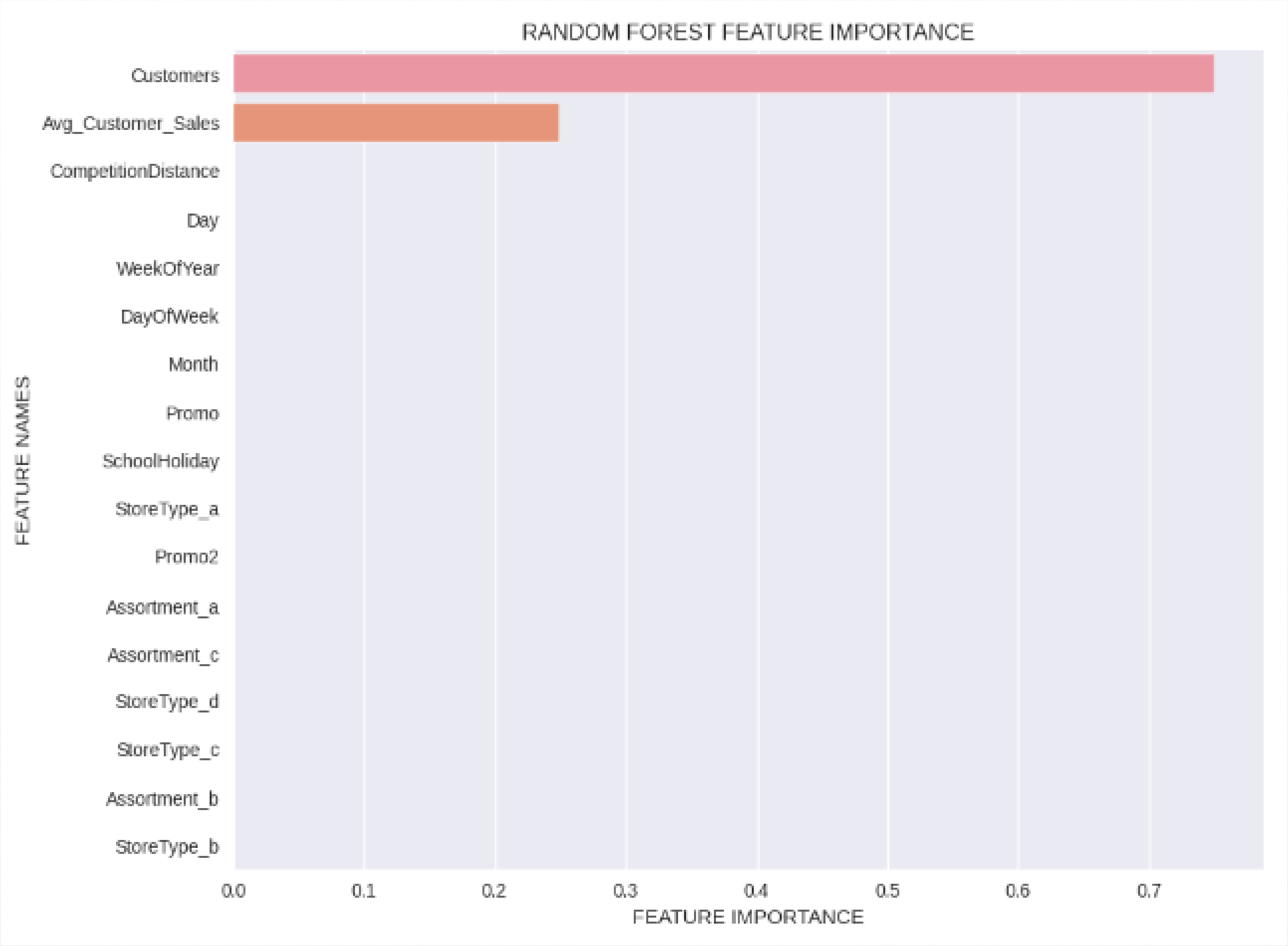
- Predictions from random forest model are very close to actual values in our X dataset as we have good score.

- Predictions from random forest model are very close to actual values in our X dataset as we have good score.

- Predictions from random forest model are very close to actual values in our X dataset as we have good score.
- The figure shows actual values, predicted & the difference between them respectively
- Since this is Sales prediction MAE is a good metric.
- We're getting Mean Absolute Error ~ $ 38
- And MAPE of 5.65%



❑FEATURE IMPORTANCE -



❑ CONCLUSION -

- ❖ Our model shows that Customers, Competition distance, Store type are some of the most important features in our sales prediction. We need to focus on these aspects to maximize our profits for the next 6 weeks.

THANK
you