Capstone Project - 2

Retail Sales Prediction

Content

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

Problem Statement

1. Rossmann operates over 3000 drug stores in 7 European countries.
2. Provided with historical sales data for 1,115 Rossmann stores. The sales are influenced by many parameters and task is to forecast the "Sales" for 6 weeks in advance.

Data Summary

We have two datasets. Rossman store data is for years 2013, 2014 and 2015 with 10,17,209 observations on 9 variables. Stores data with 1115 observations on 10 variables. Some important features are:-

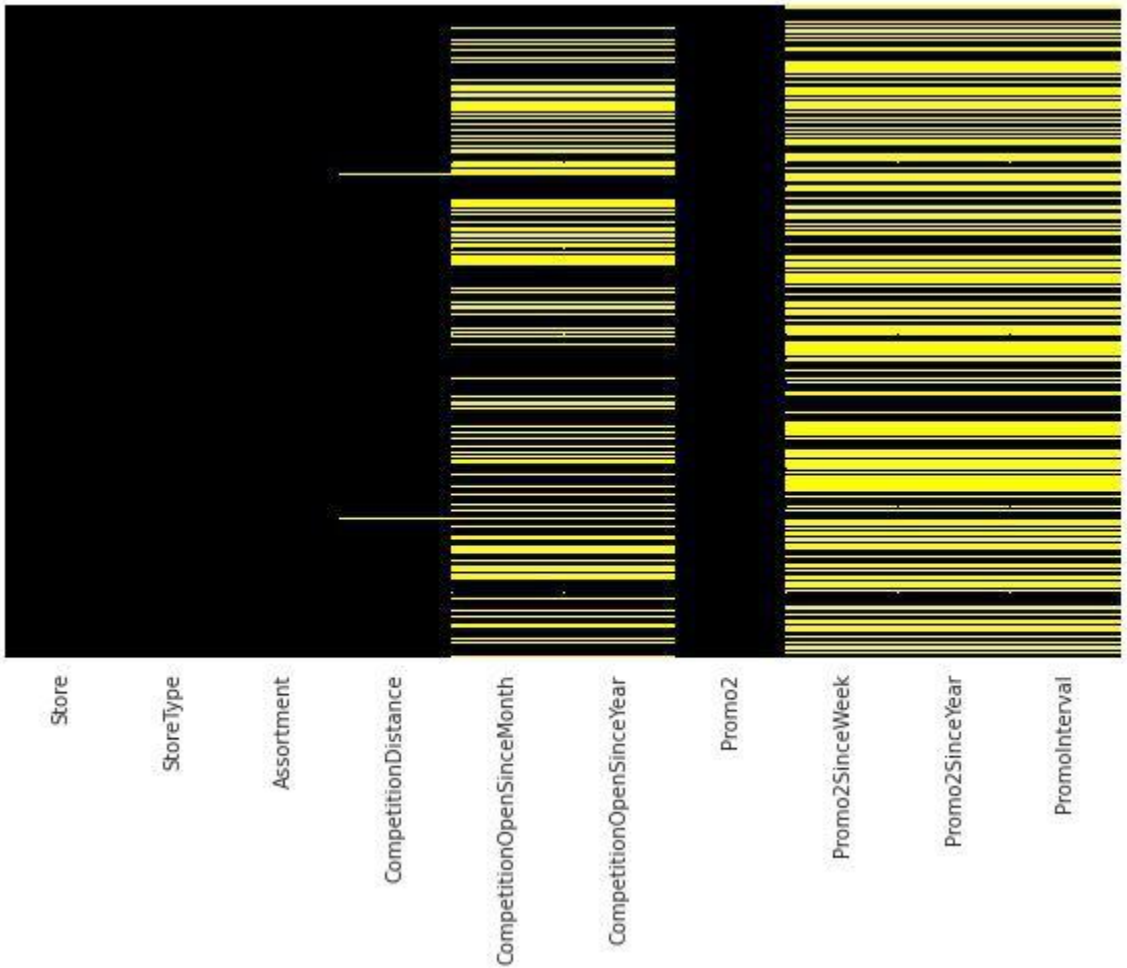
1. Customer : - The number of customers on a given day in a store.
2. Date :- Showing dates for observations.
3. State Holiday :- Indicating a state holiday.
4. Store Type : Differentiate between 4 different store models (a,b,c,d).
5. Assortment : Describes an assortment level i.e a : basic, b : extra 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.

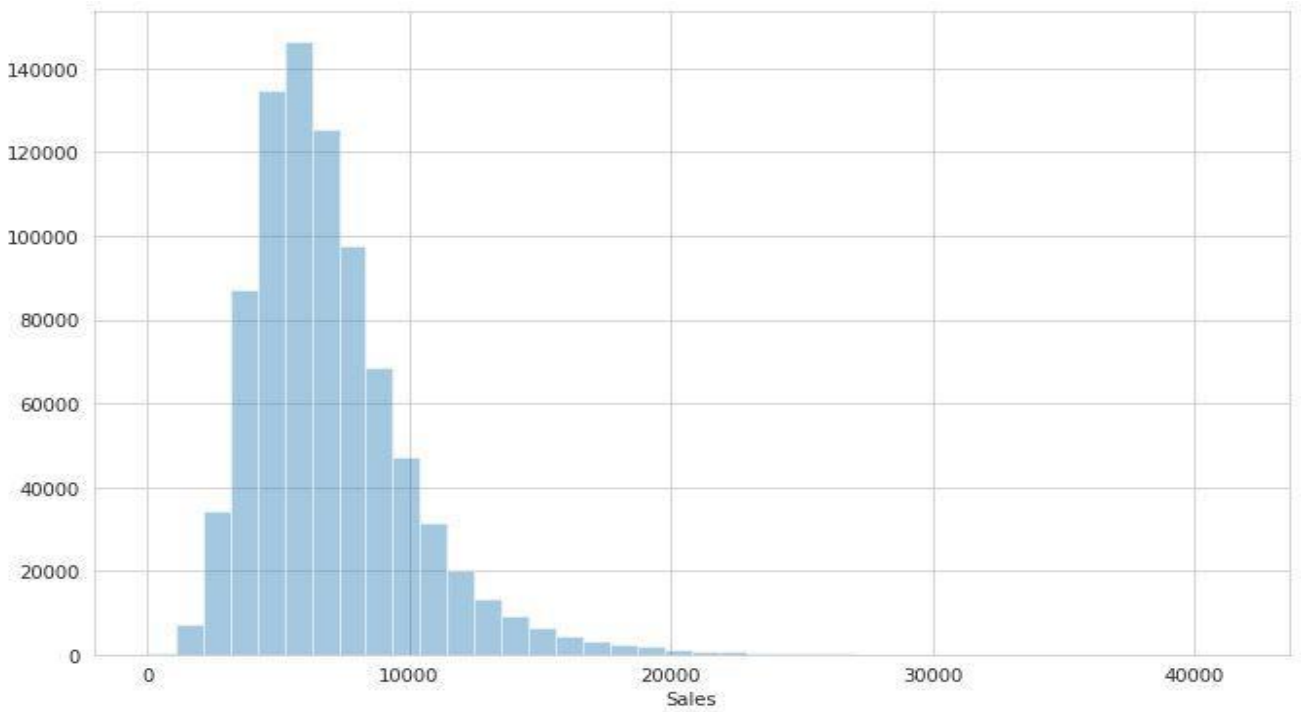
Null values in ‘Competition Distance’ are imputed with median of feature.

Removing those stores observations that are temporarily closed (~ 17.3K) & stores generating zero sales.



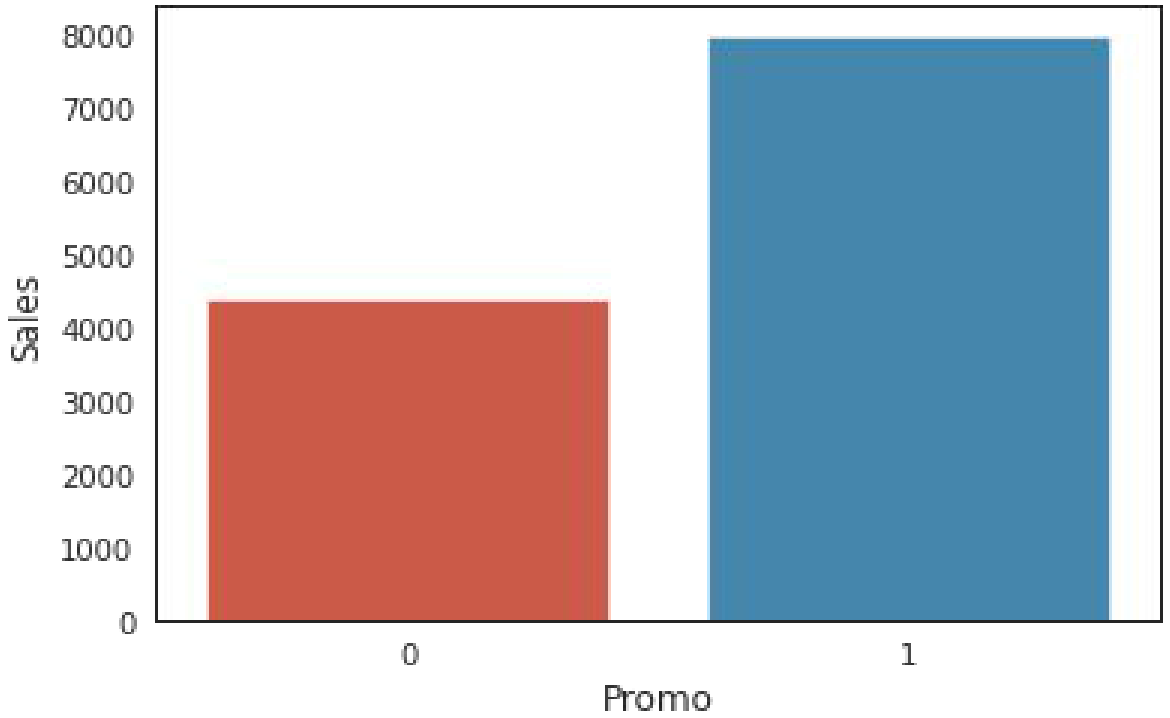
Exploratory Data Analysis

Sales are normally distributed with slightly right tail skewed.



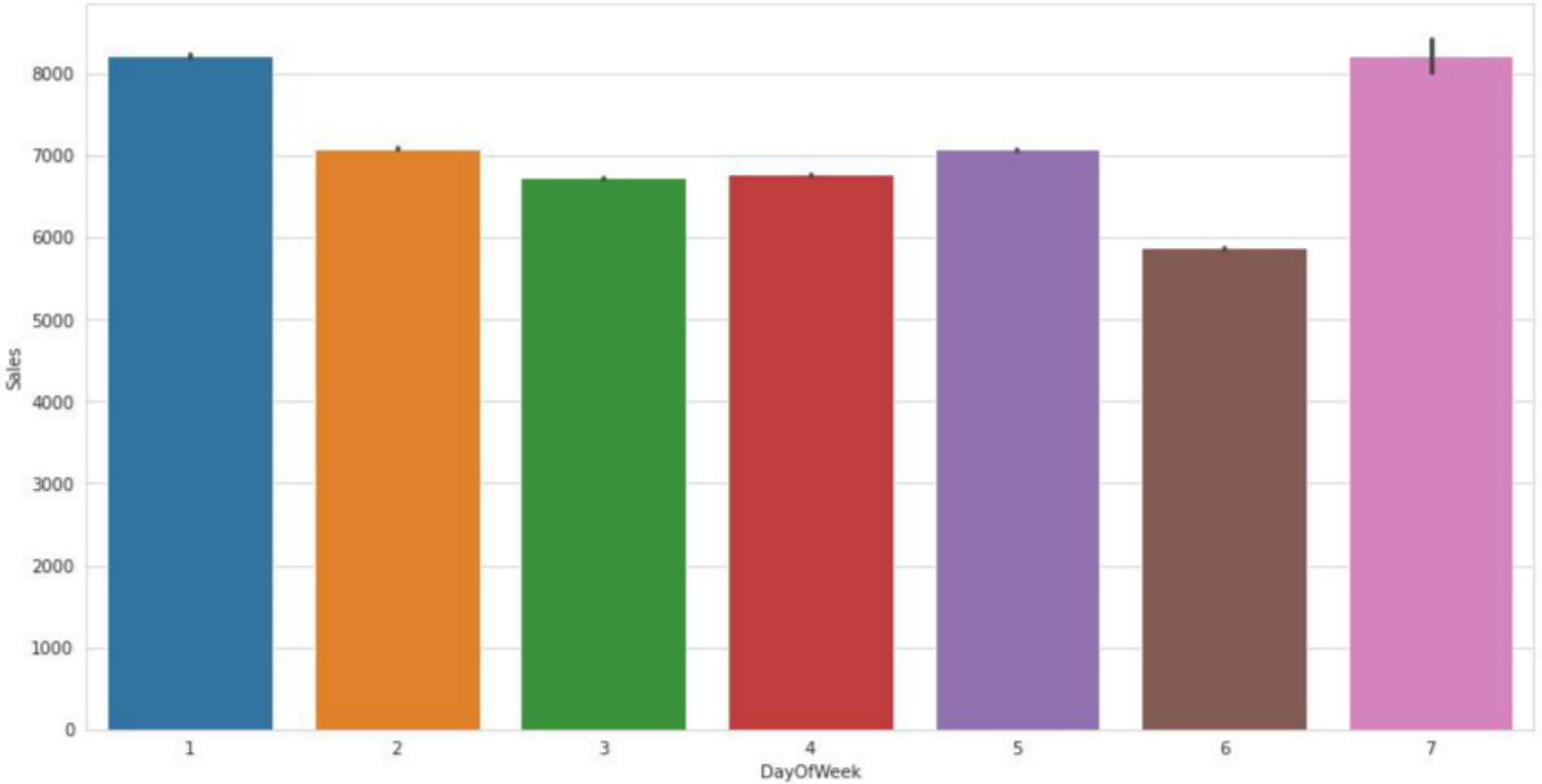
EDA (contd..)

Impact of Promo on sales



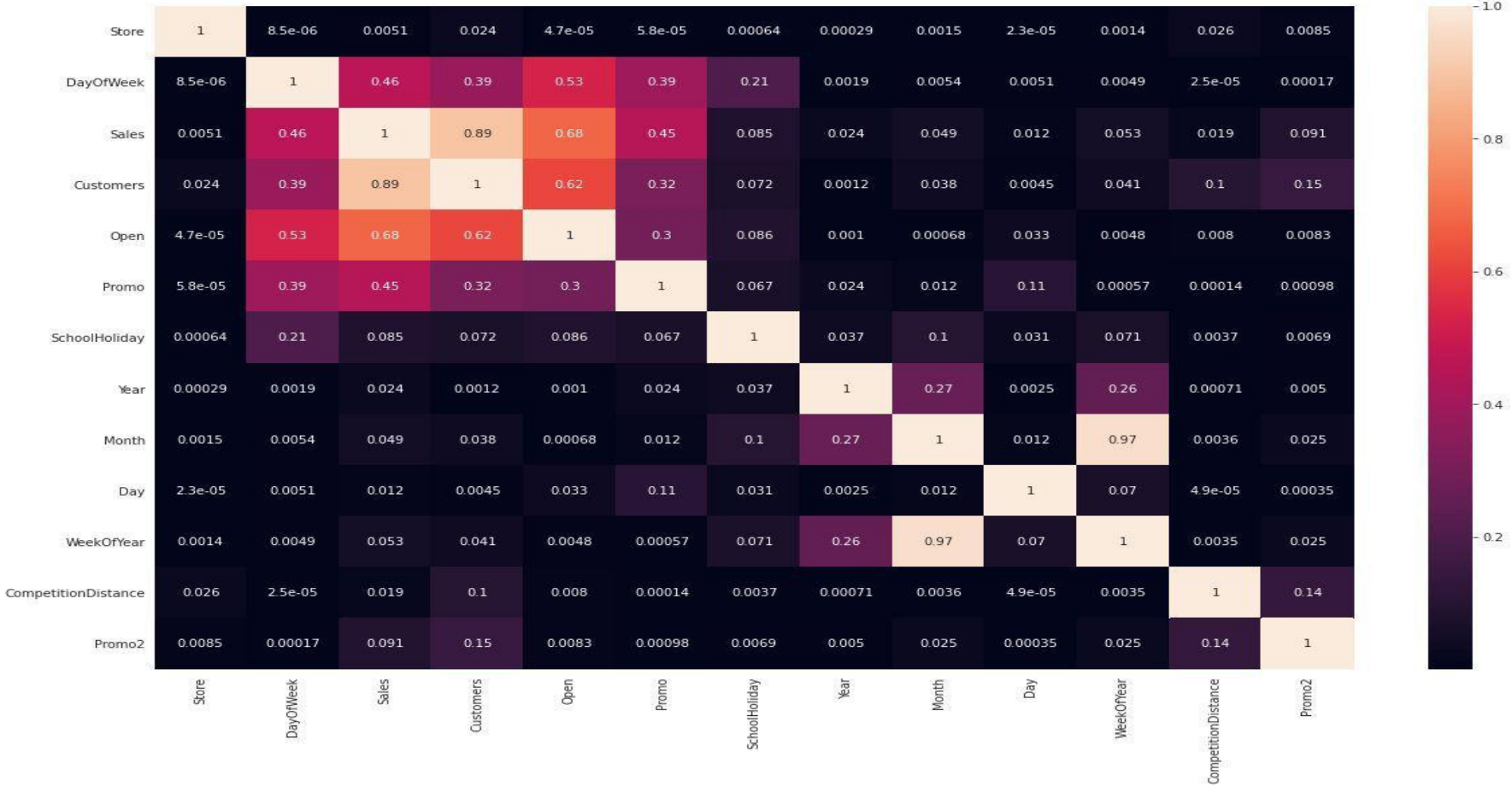
EDA (contd..)

Day Wise trends in Sales



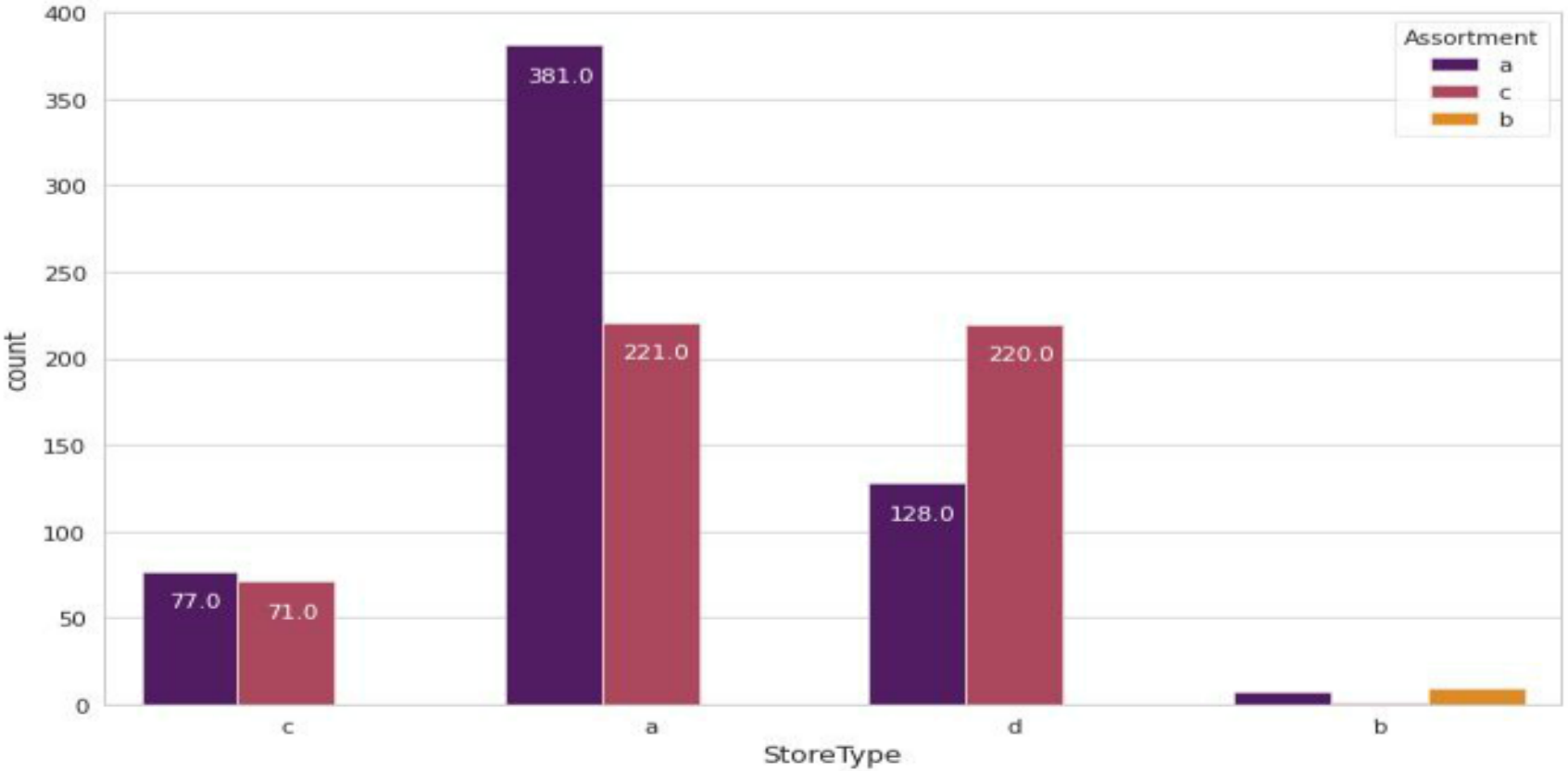
EDA (contd..)

Heatmap for merged dataset



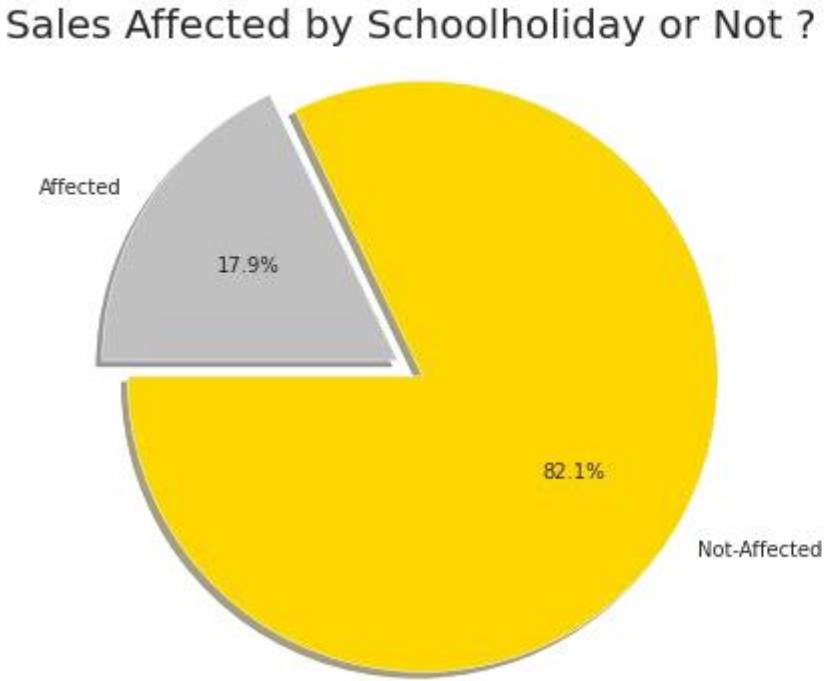
EDA (contd..)

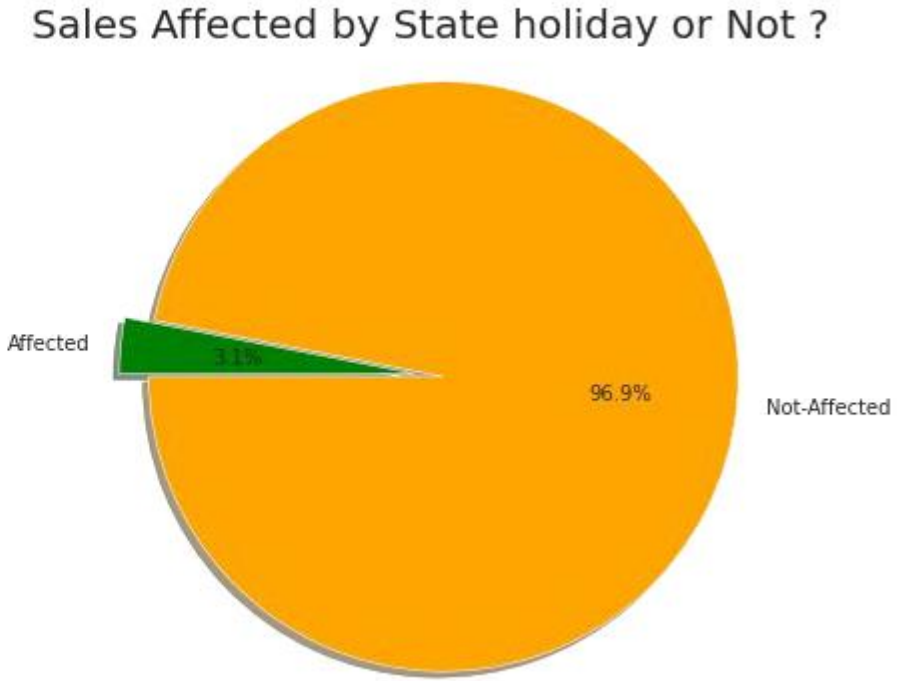
Analysis of Store Types with their respective assortment.



EDA (contd..)

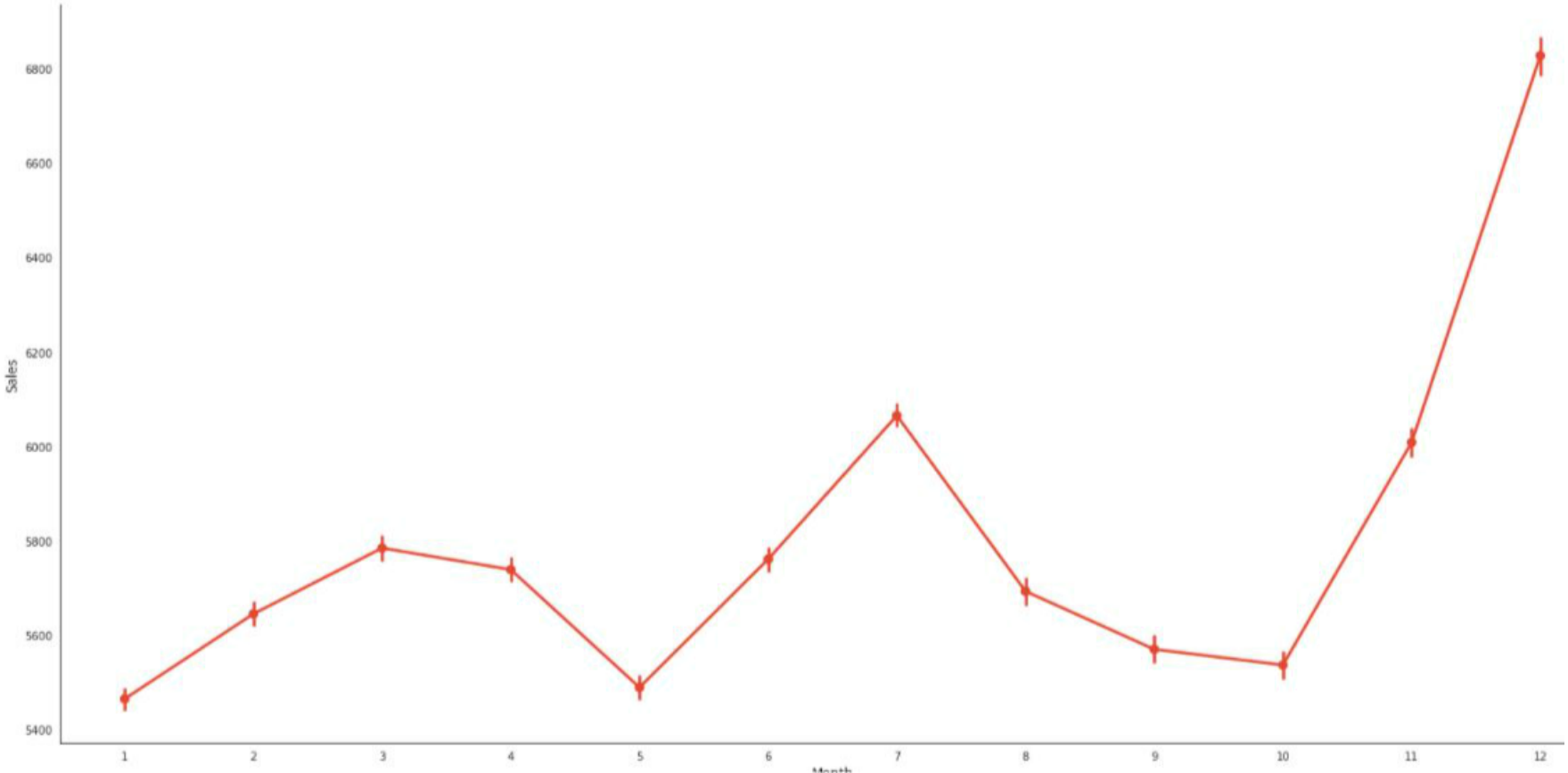
School and State holidays effect on sales





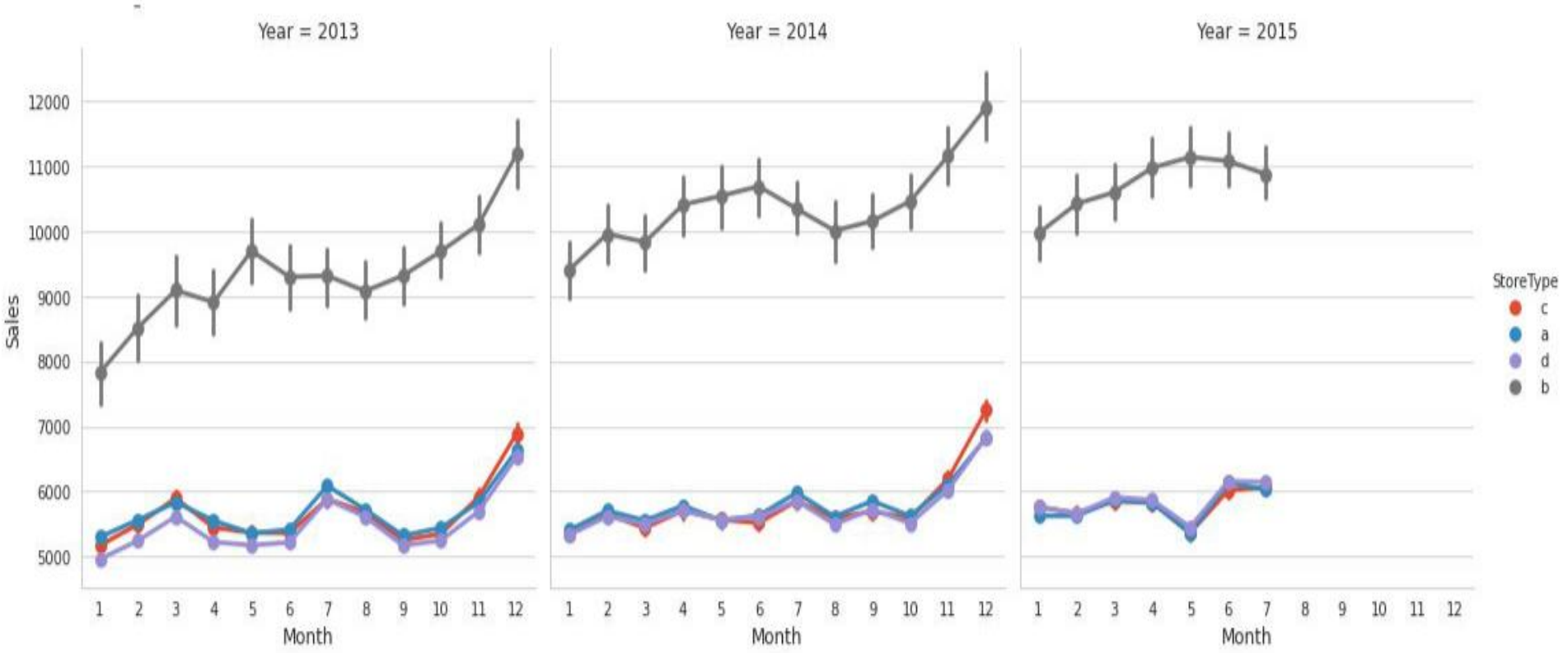
EDA (contd..)

Monthly trends in Sales



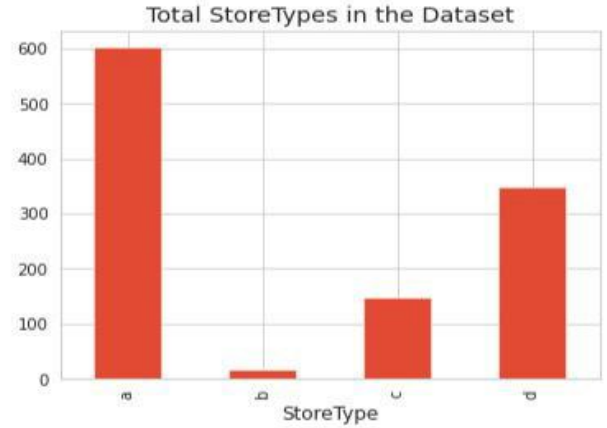
EDA (contd..)

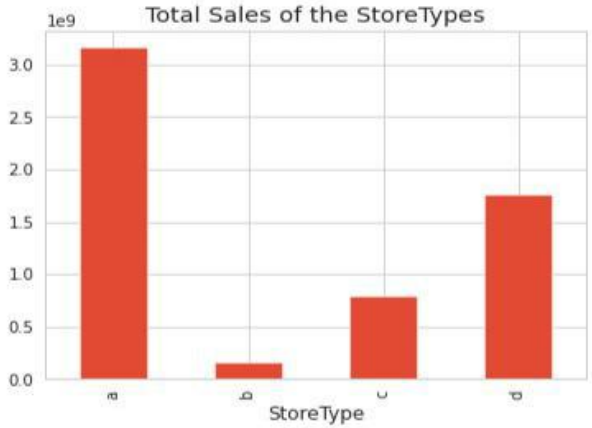
Yearly Distribution of Sales according to store types

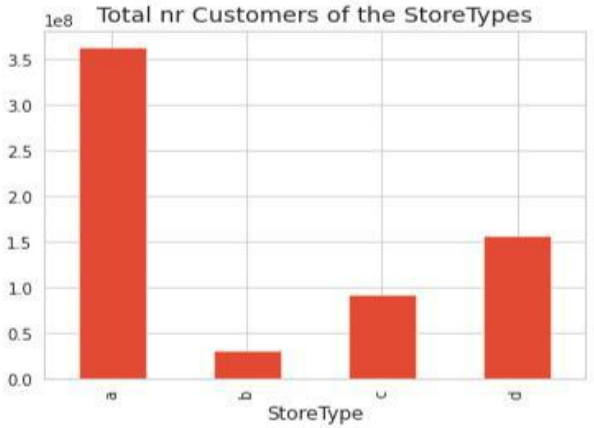


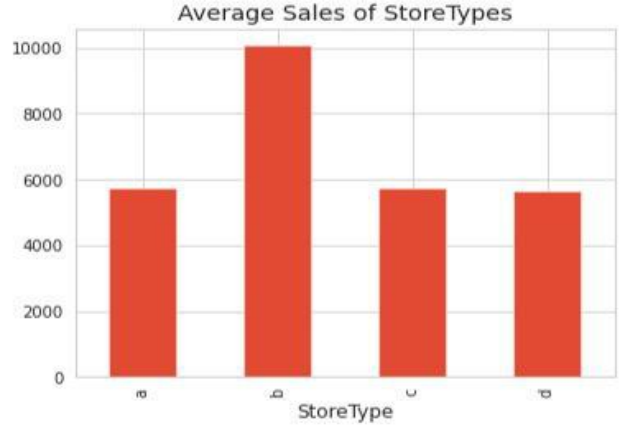
EDA (contd..)

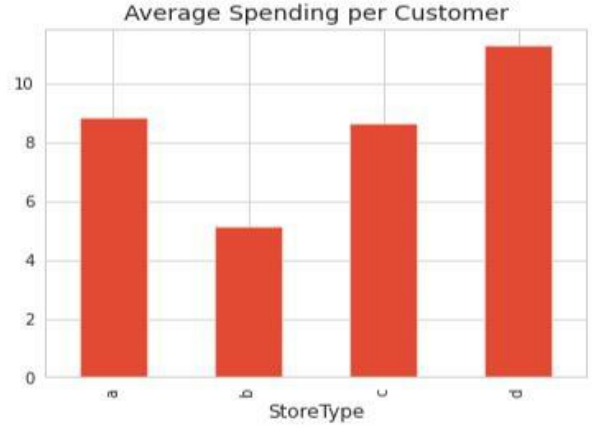
Store Types and average sales/customer/spending relation

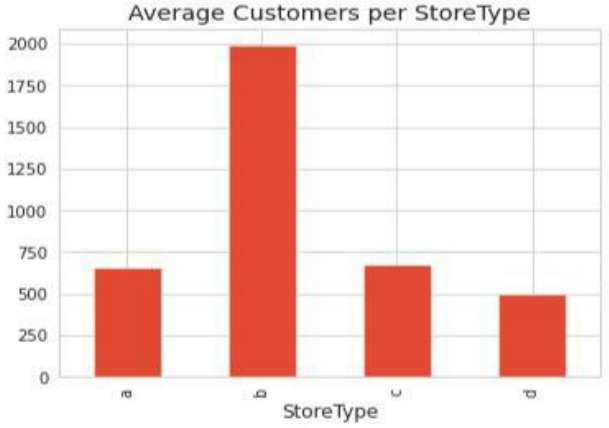






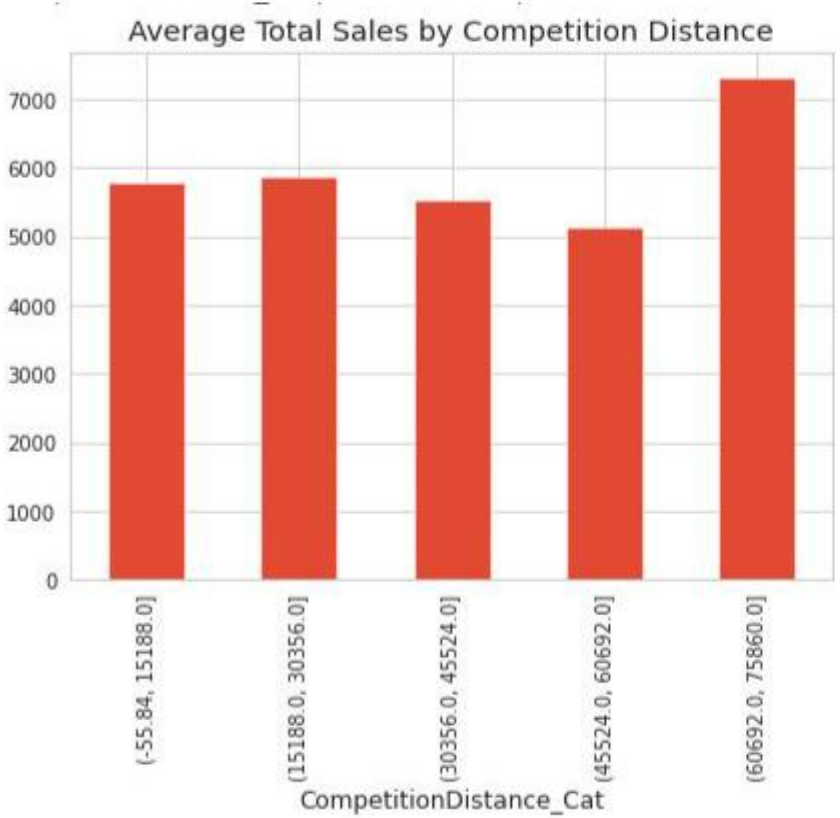


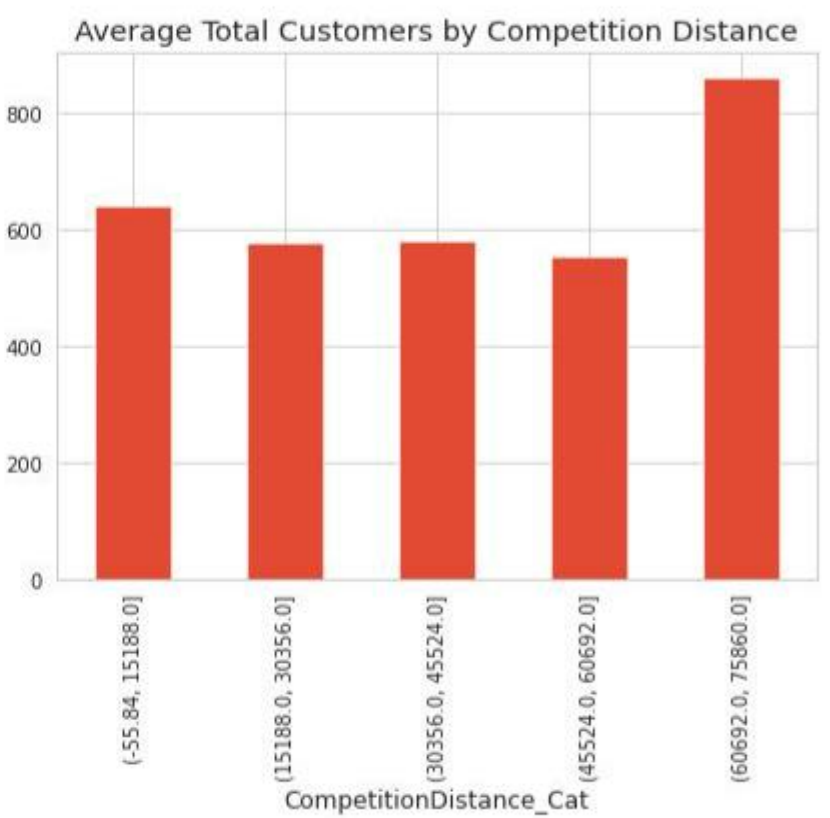




EDA (contd..)

Impact of Competition Distance on Sales and Customers





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 Storetype, Assortment.

4. Splitting dataset into Training and Test set and applying MinMaxScaler 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

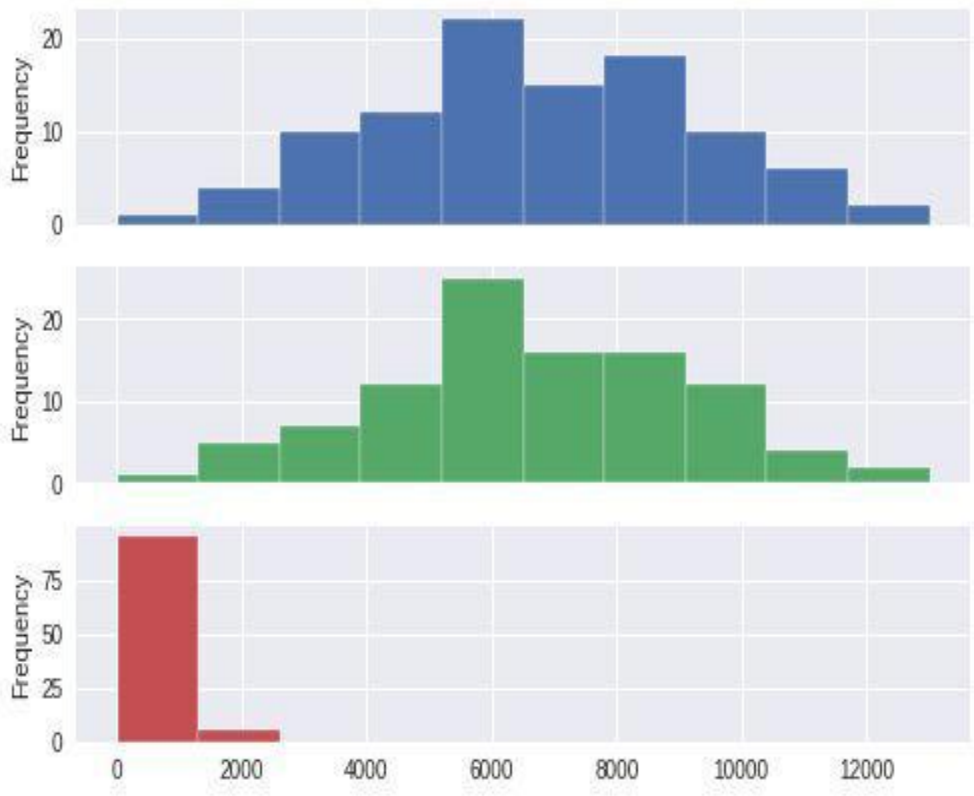
Model Evaluation

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

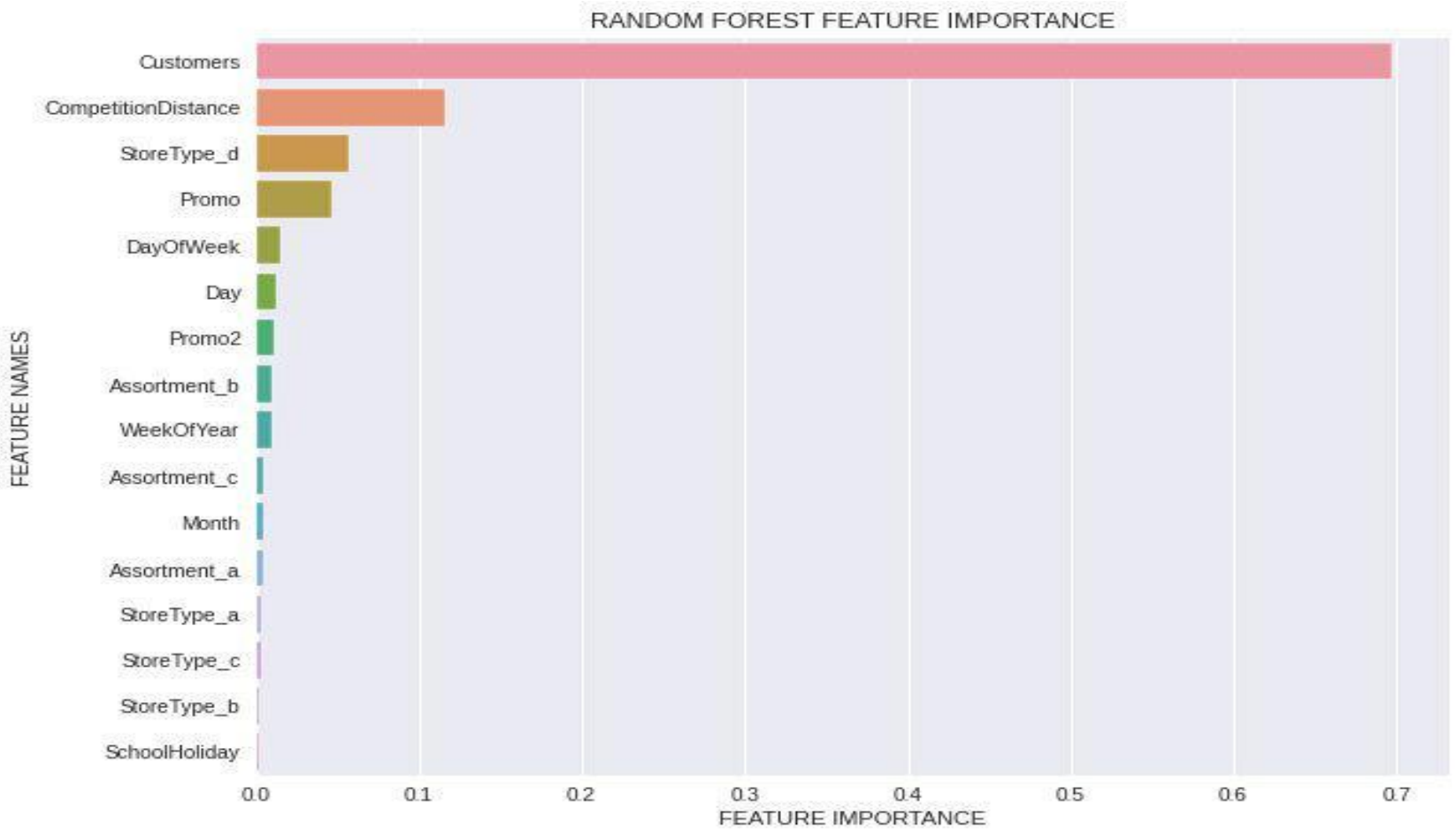
Insights from Random Forest Regressor

Predictions from random forest model are very close to actual values in our X dataset as we have good score. 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 ~ $ 380 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

Presented By:

Sampreet Chakraborty
Data Science Trainee
AlmaBetter