Capstone Project - 2

Retail Sales Prediction

Content

Problem Statement

- Rossmann runs more than 3000 pharmacies throughout seven nations in Europe.
- Given historical sales information for 1,115 Rossmann pharmacies stores. Numerous factors affect sales, and it is your job to predict Sales for the next six weeks.

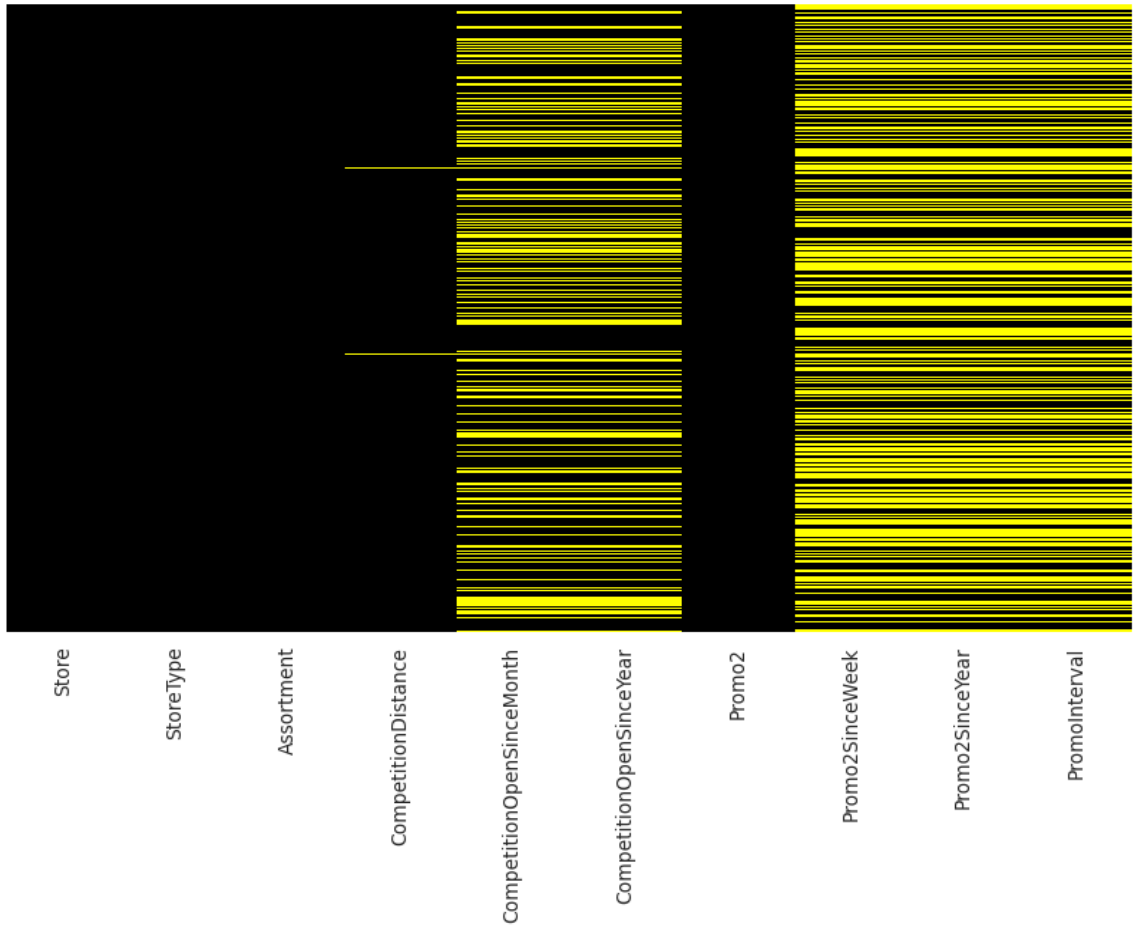
Data Summary

Two datasets exist. Rossman storage data includes 10,17,209 observations over 9 variables for the years 2013, 2014, and 2015. 1115 observations on 10 variables are stored in the data. Some key characteristics include:

1. Customer: the number of customers who visit a store on a particular day.
2. Date : Showing dates for observations.
3. State Holiday : Indicating a state holiday.
4. Store Type: Compare and contrast the four distinct store models (a, b, c, and d).
5. Assortment: describes a level of assortment, such ‘a’ as basic, ‘b’ as extra, and ‘c’ as extended.
6. Competition Distance: Distance in meters to the nearest competition store.
7. Promo: shows if a store is having a promotion that day.

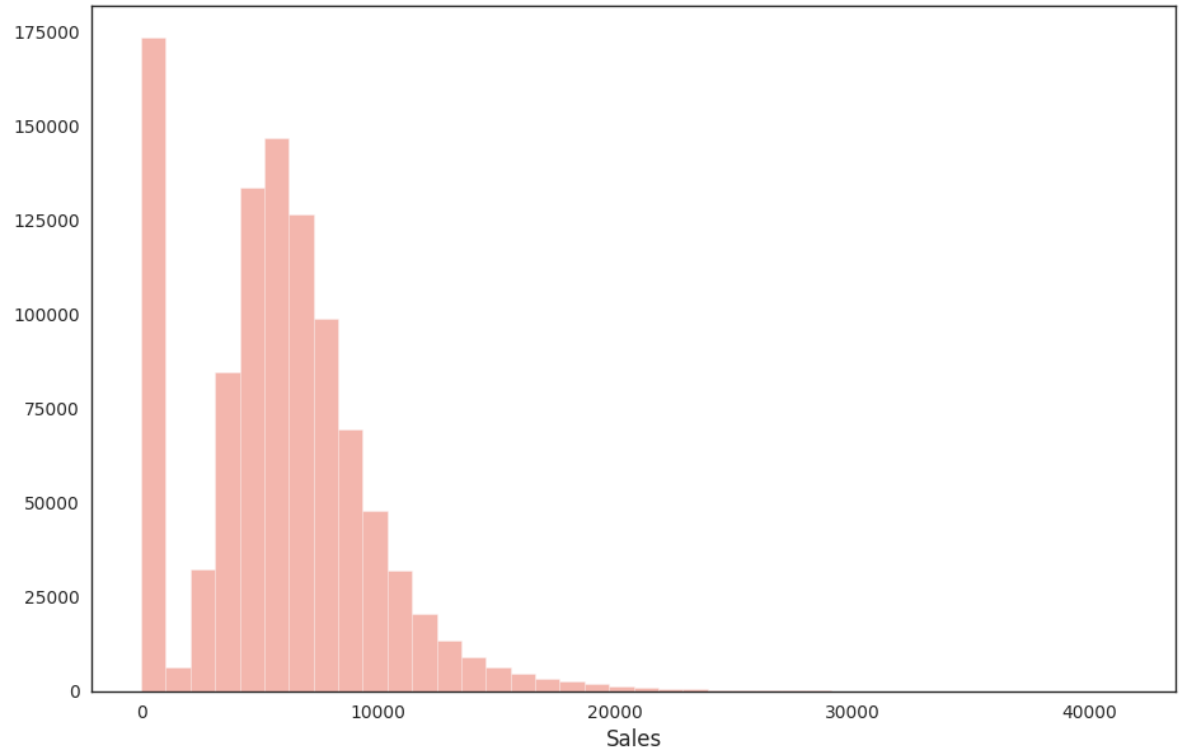
Data Preprocessing

- Columns with more than 30% null values are removed.
- The median of the feature is used to impute null values in the competition distance column.
- removing the 17.3K store observations that are now closed and the stores that had no sales.

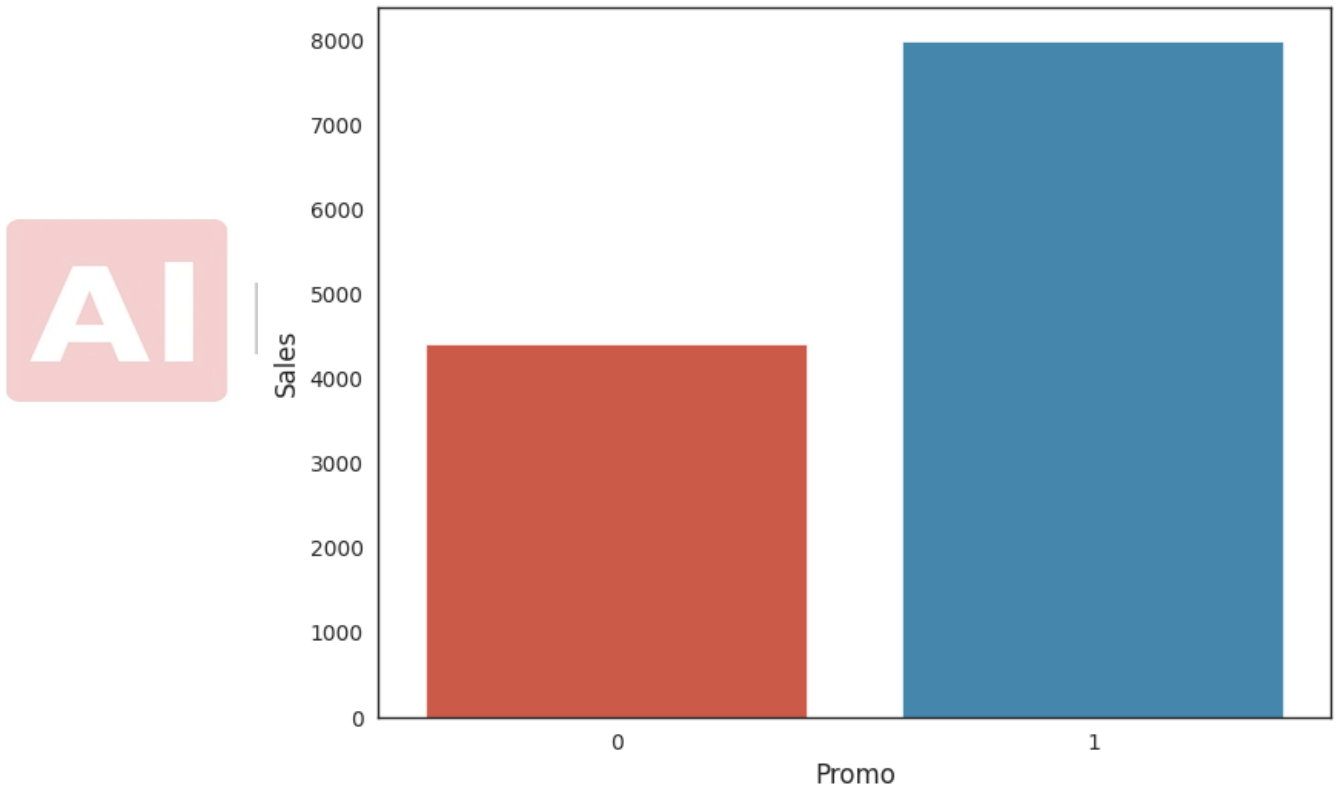


Exploratory Data Analysis

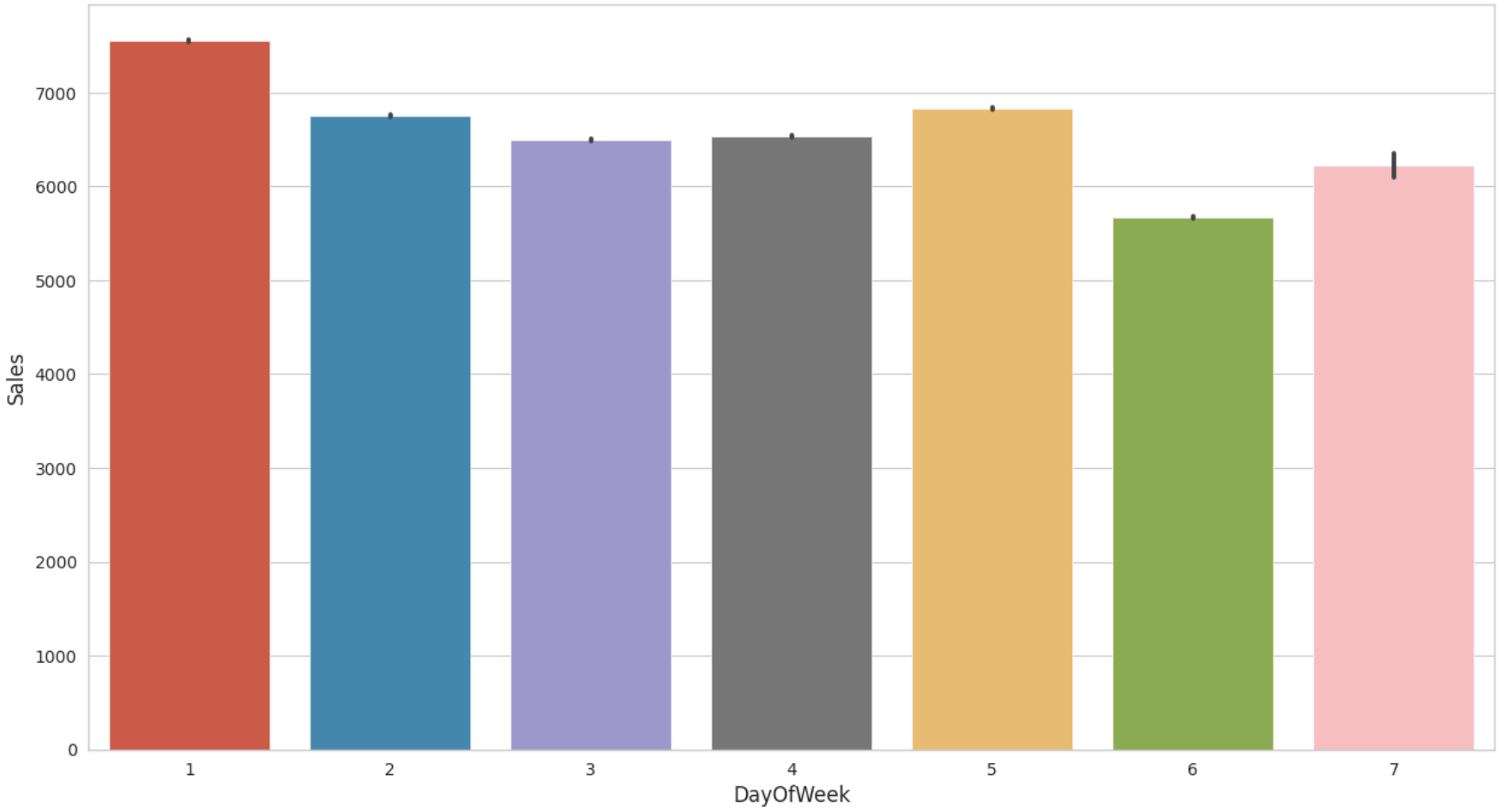
- In this chat we can observer sales are typically skewed significantly to the right in their distribution.



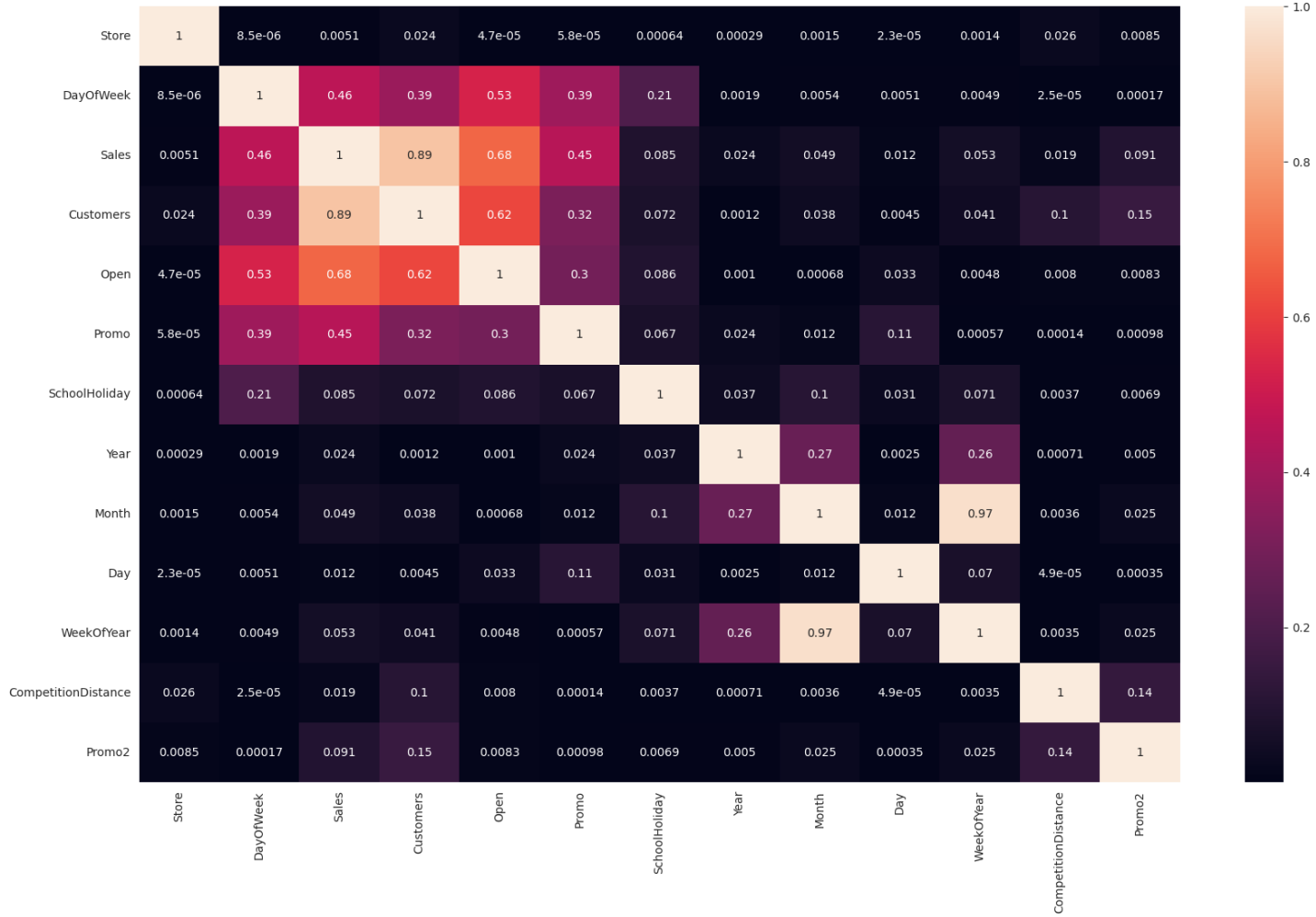
- Impact of Promo on sales



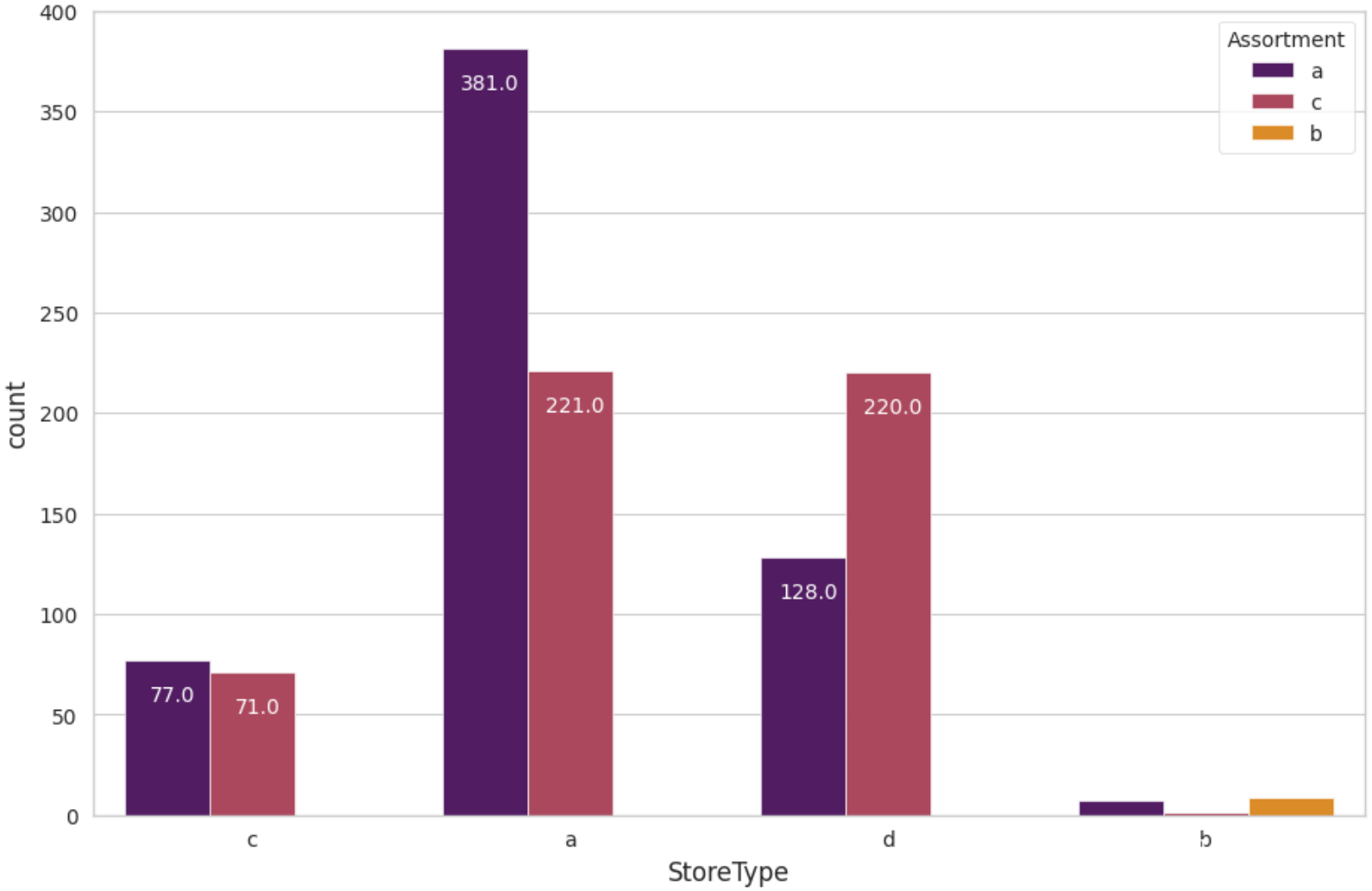
- Day Wise trends in Sales



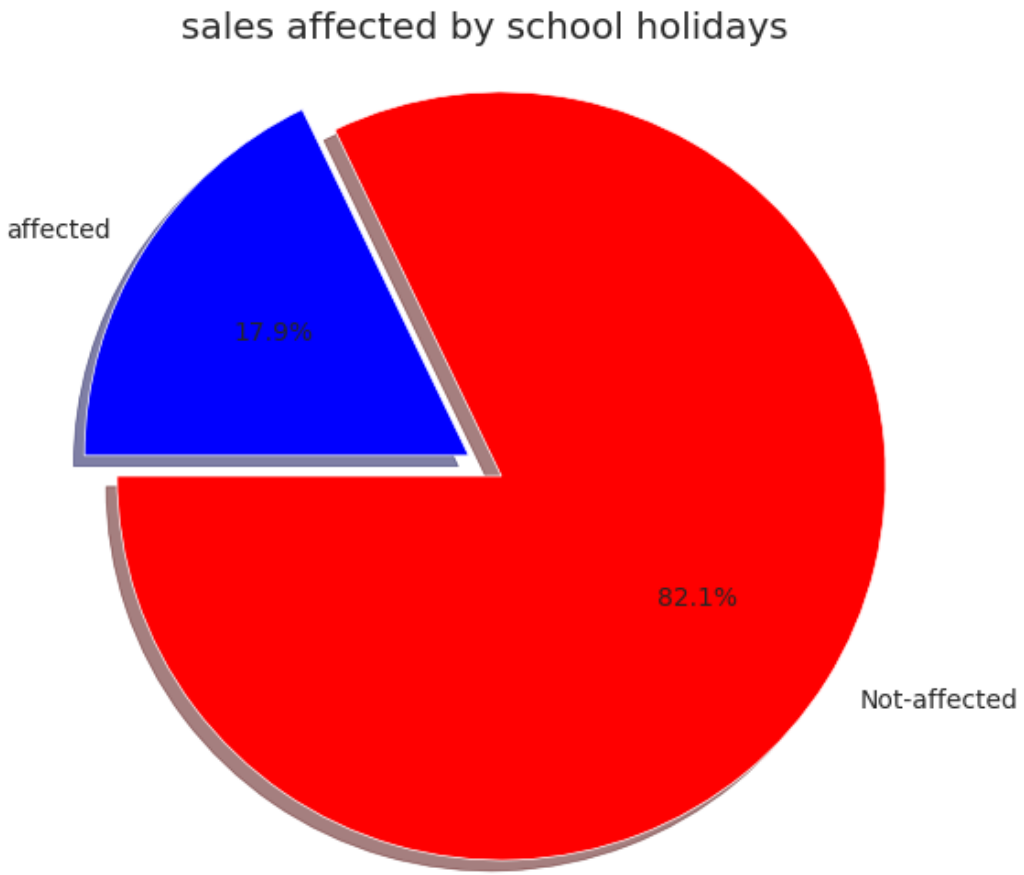
- Heatmap for merged dataset

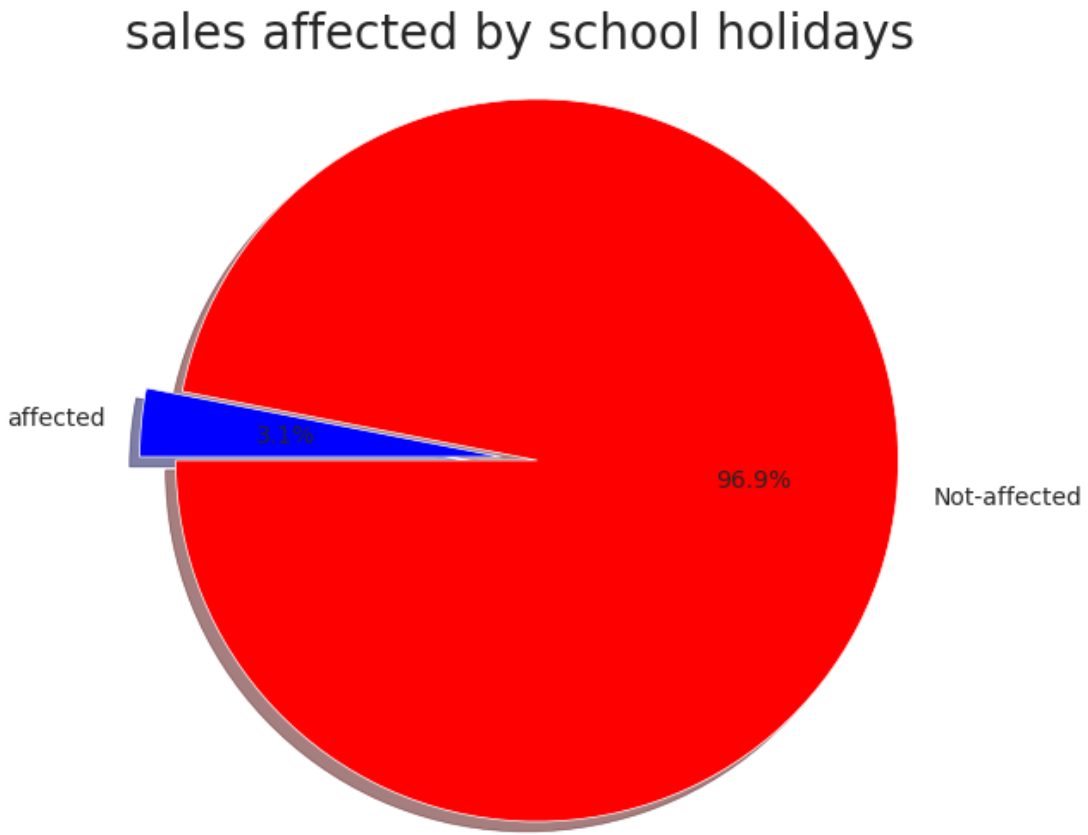


Analysis of Store Types with their respective assortment

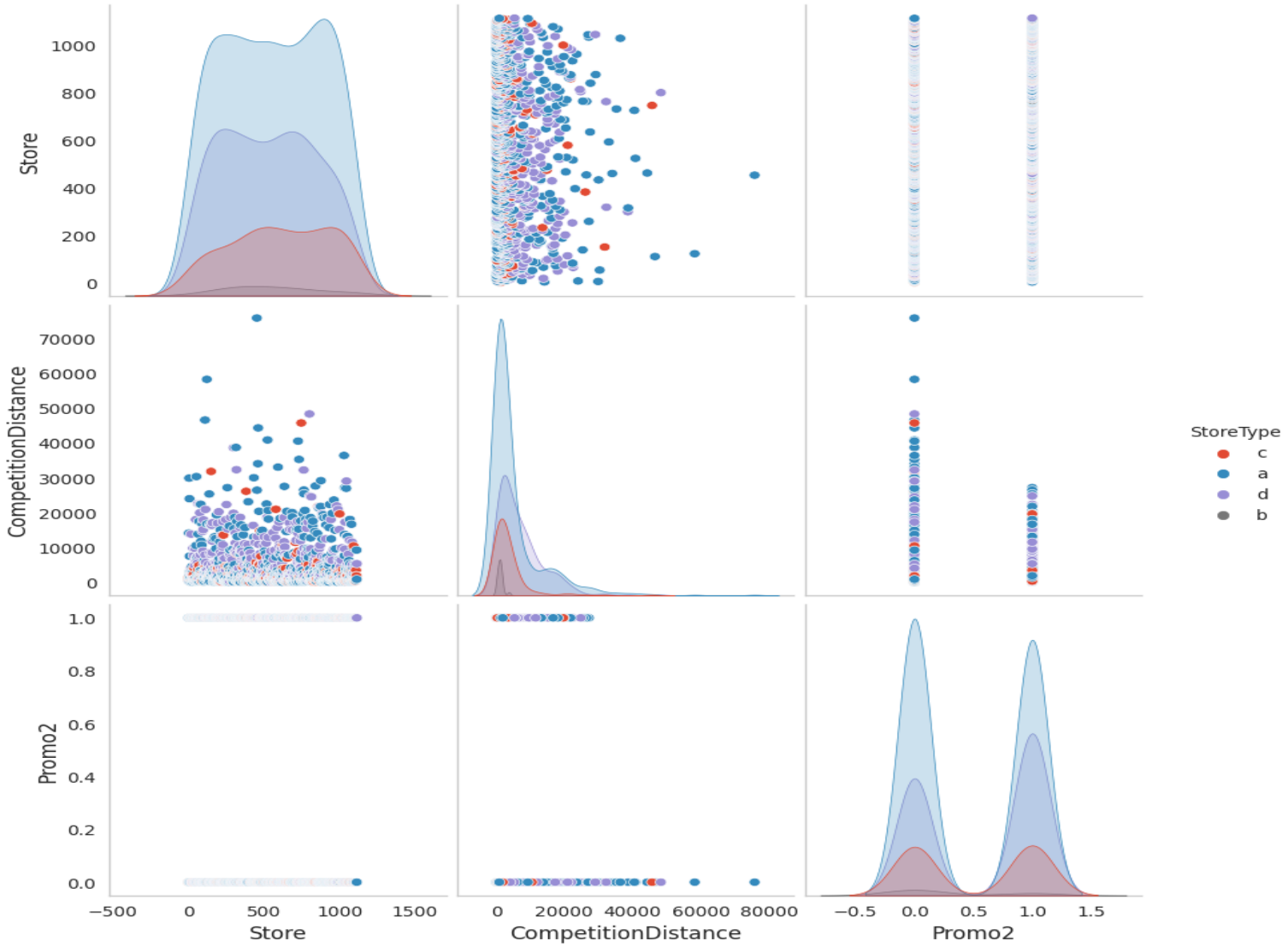


School and State holidays effect on sales

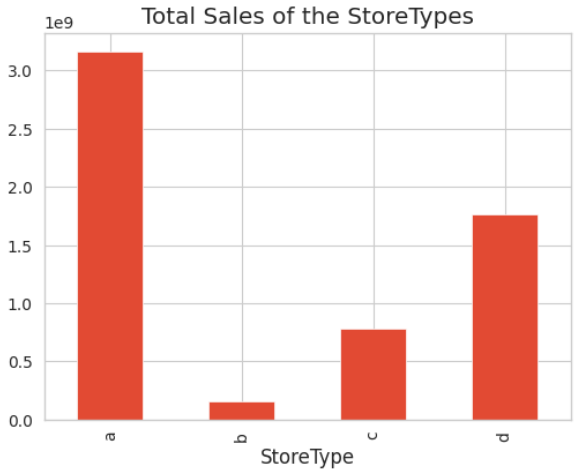


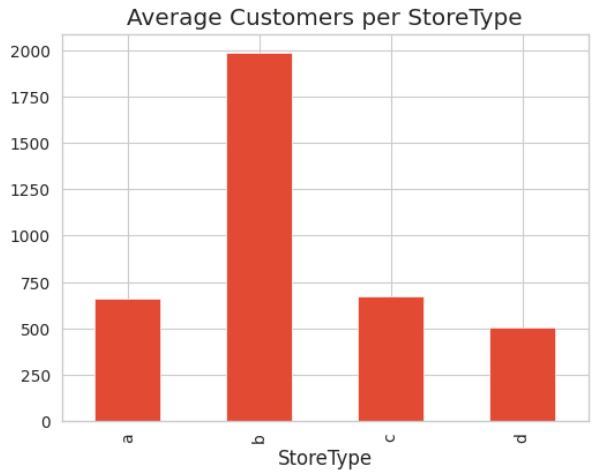


Yearly Distribution of Sales according to store types

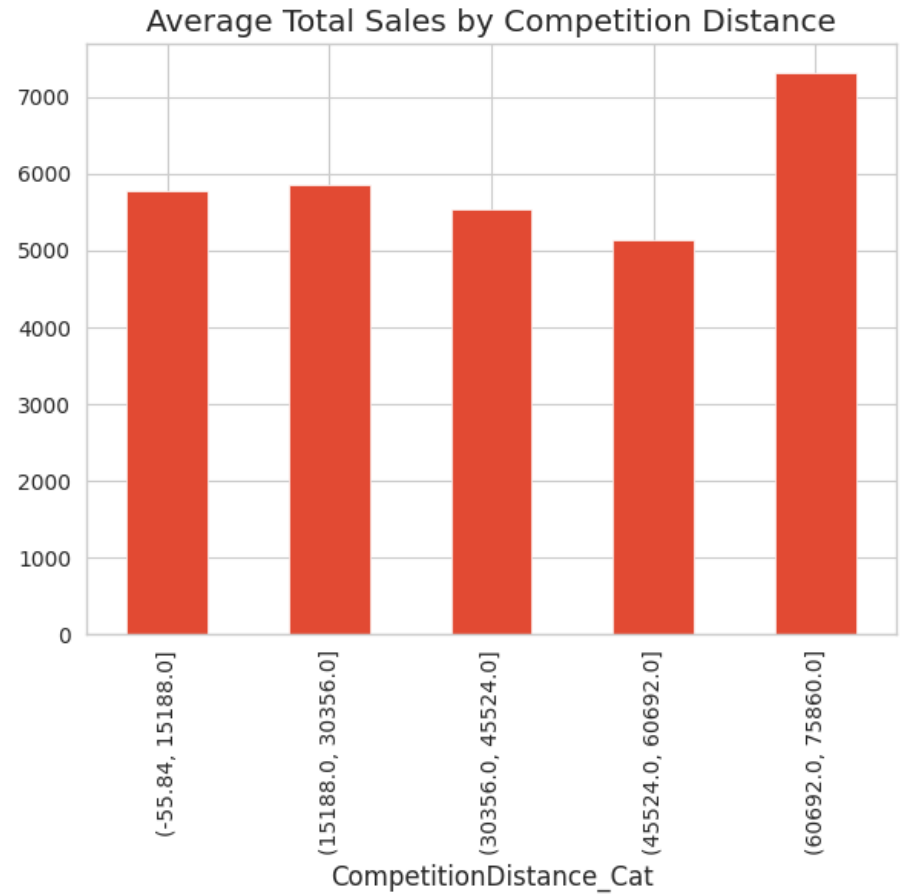


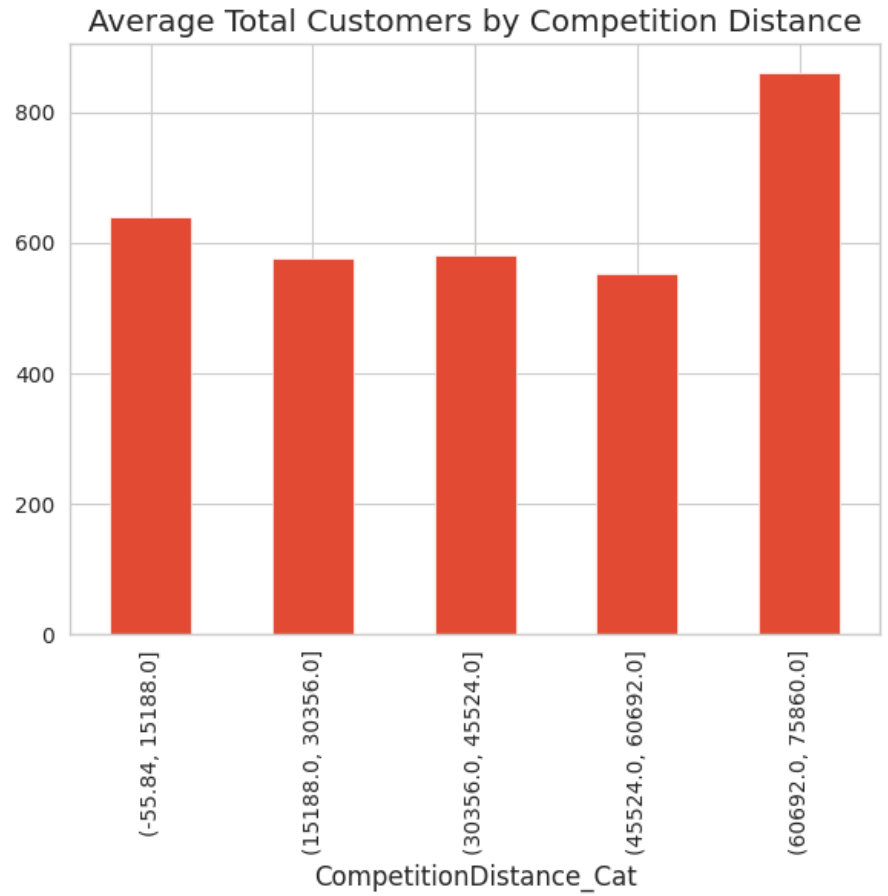
Store Types and average sales/customer/spending relation





Impact of Competition Distance on Sales and Customers





summary

1. Sales and customers have a strong relationship.
2. Stores open on 'State Holiday' generate a considerable amount of revenue.
3. There is not much difference in sales during "School Holidays.“
4. Despite having a very small number of locations, store types 'b' beat other store types in terms of sales and average customers.
5. Sales are consistent for the second quarter of the year but it starts increasing in the last quarter.

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

Feature Engineering

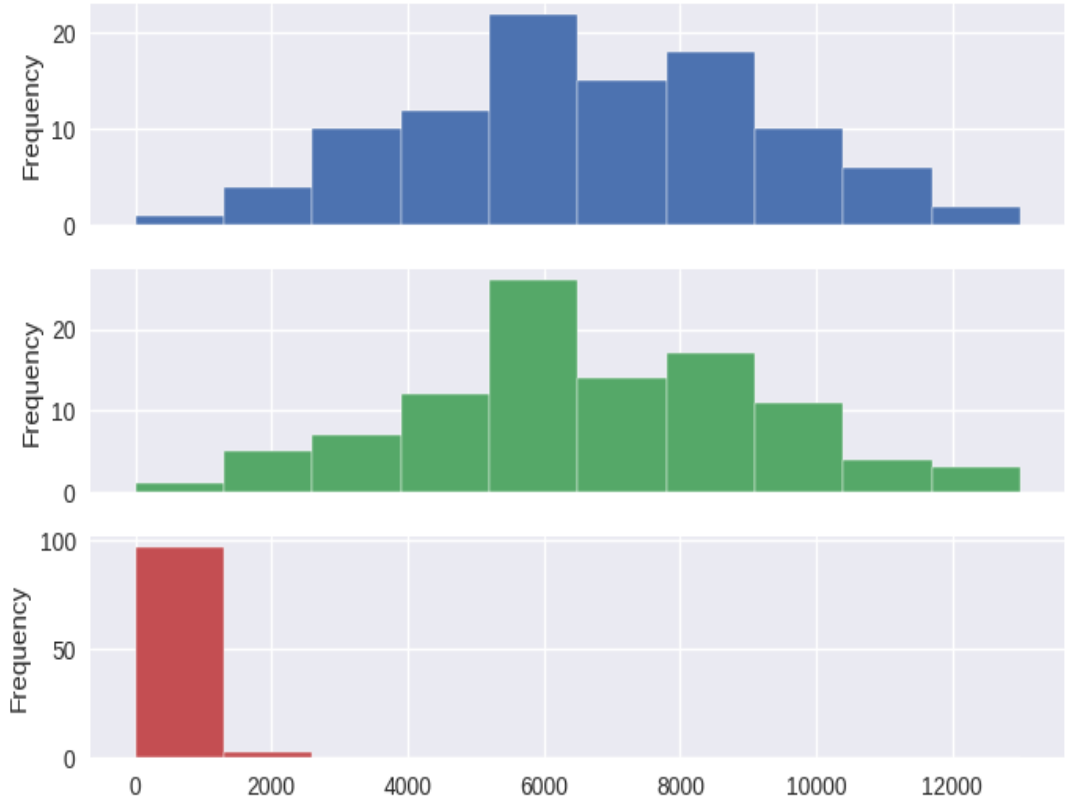
- Week, month, and year are extracted from dates and added to the dataset.
- Merging both dataset.
- One hot encoding for Storetype, Assortment.
- employing MinMaxScaler to scale the dataset after dividing it into the Training and Test sets.

Model Evaluation

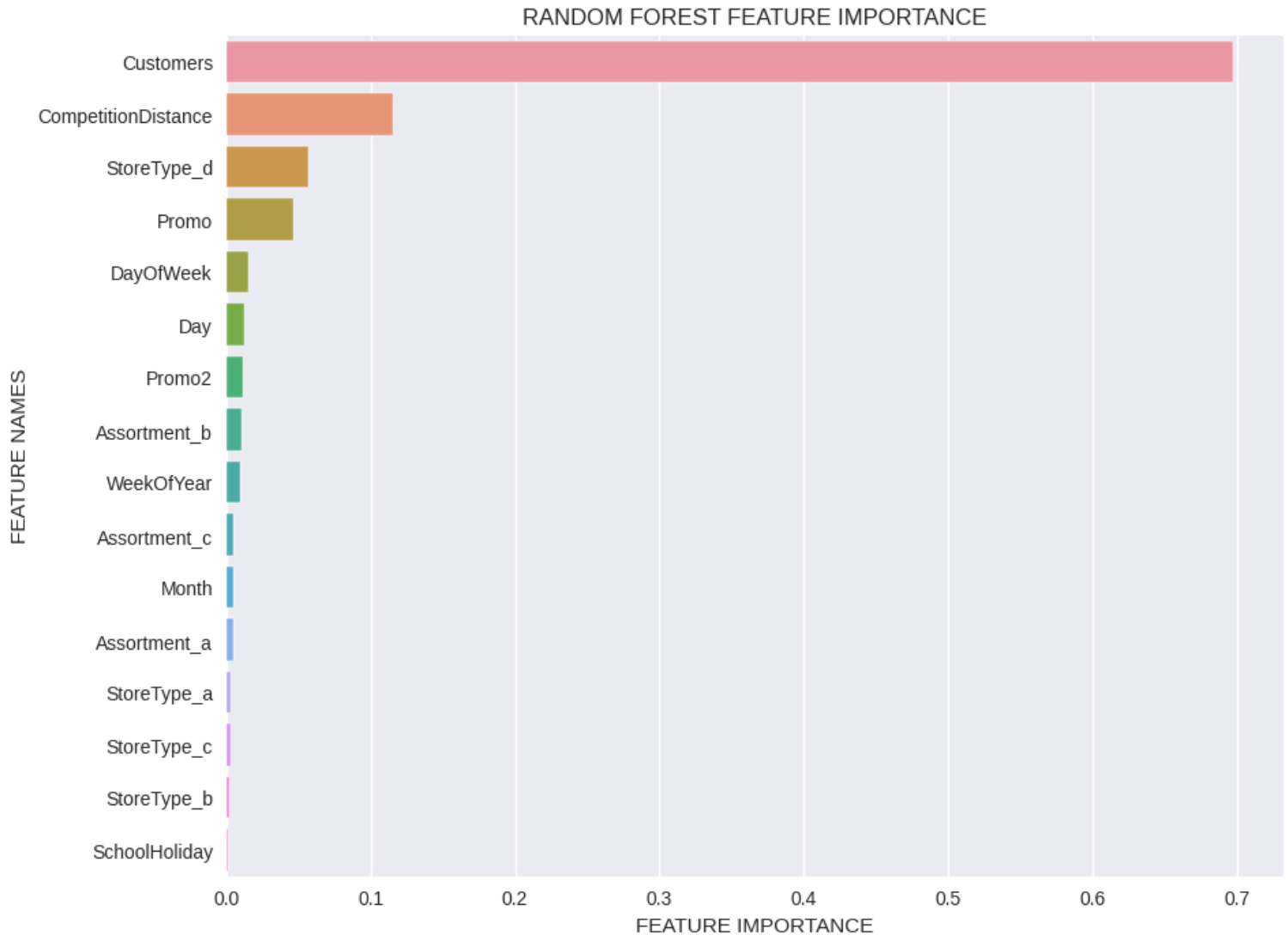
	Train_Score	Test_Score
Linear Regression	0.780750	0.782392
Lasso Regression	0.780731	0.782369
Decision Tree	0.999996	0.915607
Decision Tree(hyperparameters)	0.963506	0.935416
Random Forest Regression	0.993782	0.956380

Insights from Random Forest Regressor

- As a result of our high score, predictions from the random forest model closely match the actual values in our X dataset.The graphic displays the actual values, the anticipated values, and the difference between them.
- MAE is a useful indicator because this is a sales prediction.The MAPE is 5.65% and the mean absolute error is $ 380.



Feature Importance



Conclusion

- Our model demonstrates that some of the most crucial factors in our sales prediction are customers, competition distance, and store type. To increase our profits over the next six weeks, we must concentrate on these factors.

Thank you