Capstone Project – 2

Machine Learning – Regression

Team **Champions** : Rossmann Retail Sales Prediction

Team Members

Varsha Rani

Vivek Chandrakant Pawar

Rabista Parween

Tushar Gaikwad

POINTS FOR DISCUSSION

- Problem Statement & Objective
- Data Summary
- Importing Libraries & Loading Data
- Exploratory Data Analysis
- Pre – processing & Feature Engineering
- Data Modeling
- Conclusion

Problem Statement & Objective

Rossmann operates over 3,000 drug stores in 7 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.

We are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set.

Note – some stores in the dataset were temporarily closed for refurbishment.

Data Summary

We are given two data sets :

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

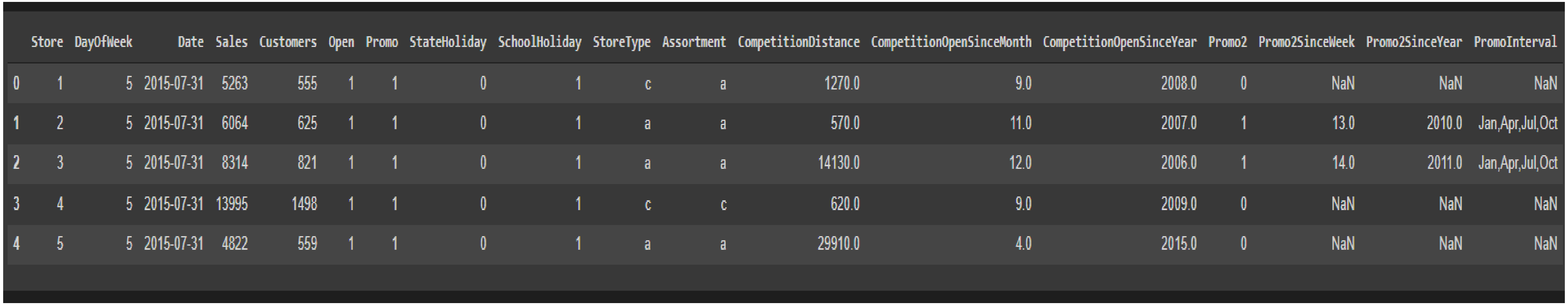
Data fields

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)
- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
- StoreType - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- CompetitionDistance - distance in meters to the nearest competitor store
- CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
- Promo - indicates whether a store is running a promo on that day
- Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
- PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store.

Importing Libraries & Loading Data

- Pandas – Manipulation of tabular data in Dataframes
- Numpy – Mathematical operations on arrays
- Matplotlib – Visualization
- Seaborn – Visualization
- Sklearn – Data Modeling

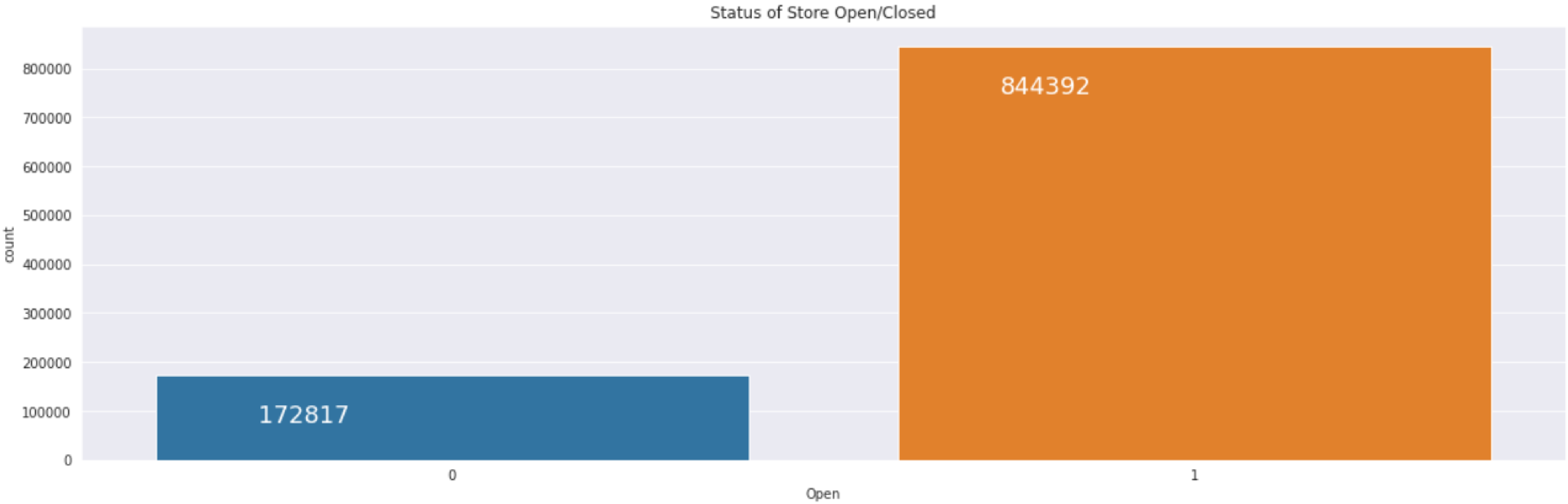


	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	1	5	2015-07-31	5263	555	1	1	0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN	NaN
1	2	5	2015-07-31	6064	625	1	1	0	1	a	a	570.0	11.0	2007.0	1	13.0	2010.0	Jan,Apr,Jul,Oct
2	3	5	2015-07-31	8314	821	1	1	0	1	a	a	14130.0	12.0	2006.0	1	14.0	2011.0	Jan,Apr,Jul,Oct
3	4	5	2015-07-31	13995	1498	1	1	0	1	c	c	620.0	9.0	2009.0	0	NaN	NaN	NaN
4	5	5	2015-07-31	4822	559	1	1	0	1	a	a	29910.0	4.0	2015.0	0	NaN	NaN	NaN

We have merged both the dataset Rossmann Stores Data.csv and store.csv for further implementations.

Exploratory Data Analysis

When stores are closed?



Over those two years, 172817 is the number of times that different stores were closed on given days. From those closed events, 2263 times the stores were closed because there was a school holiday. For closed event 30140 times it occurred because of either a bank holiday or easter or Christmas. Sometimes the stores were closed because they were undergoing refurbishments.

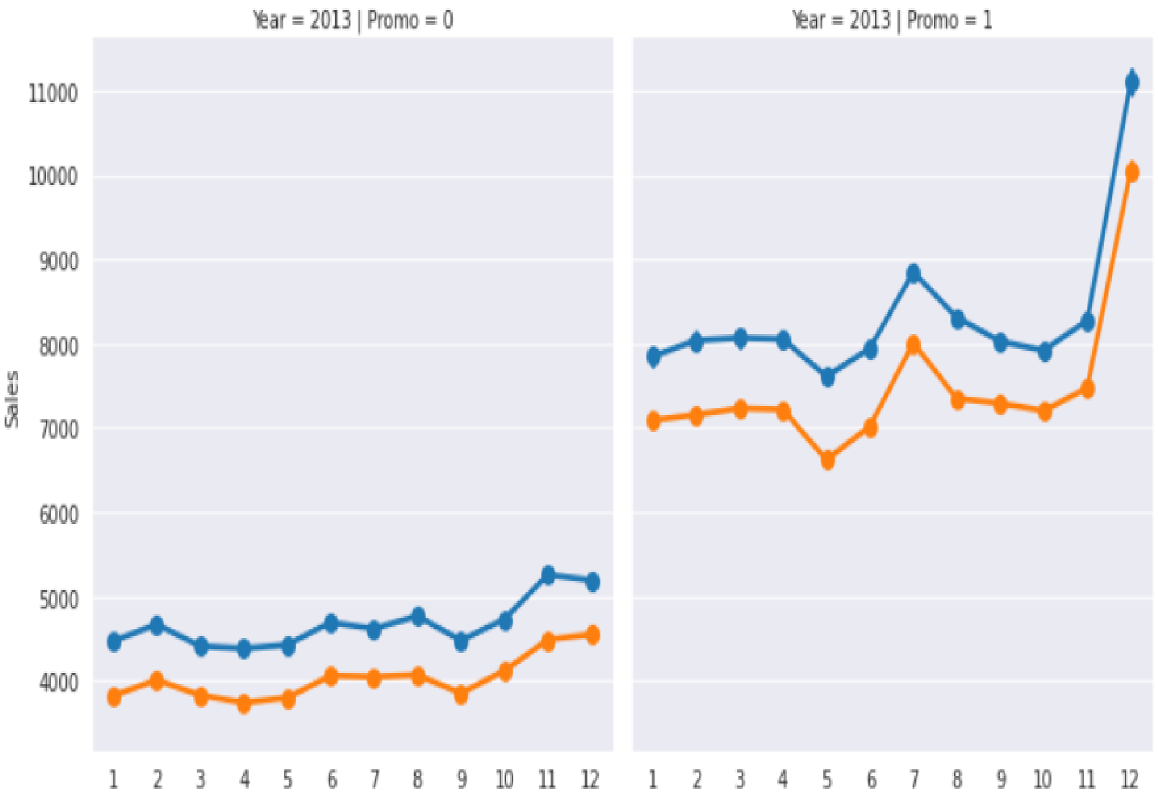
EDA Continued…

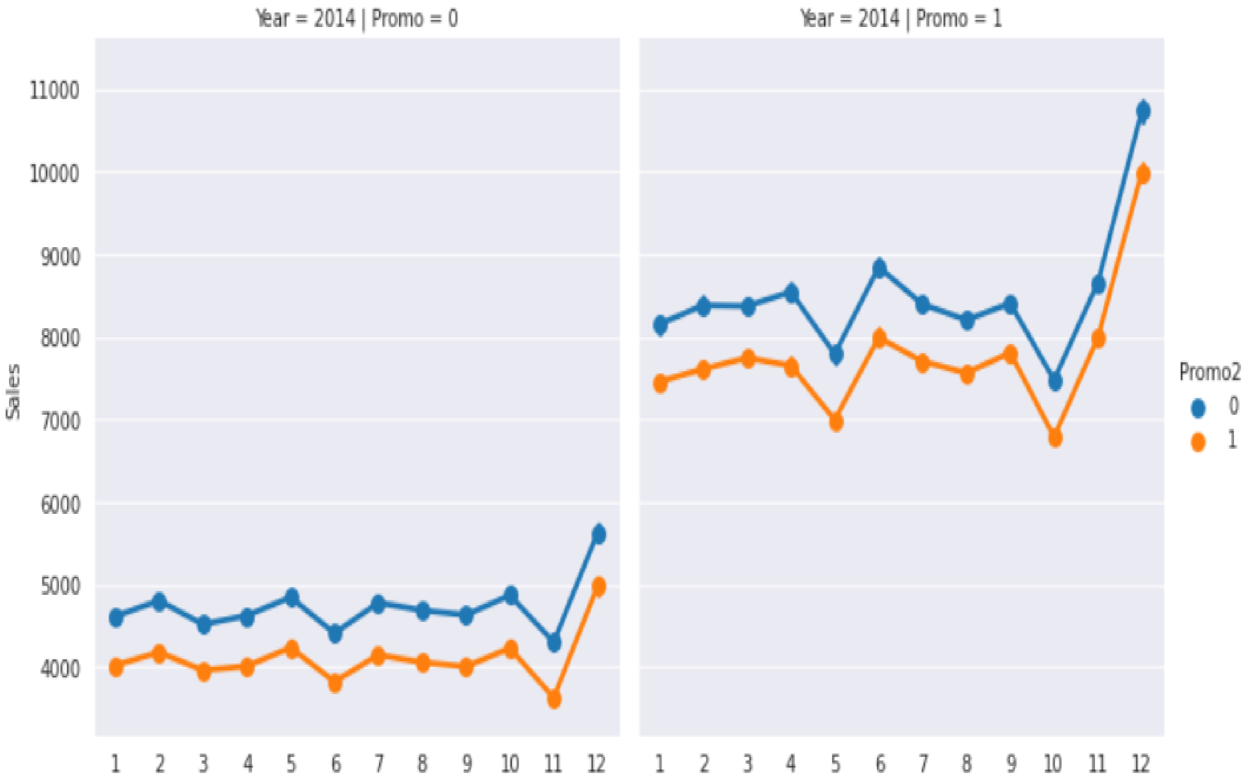
Around 62% stores were not running promotion on that day.

Sales with no promotion = 5928.965569239575

Sales with promotion = 8223.920367241544

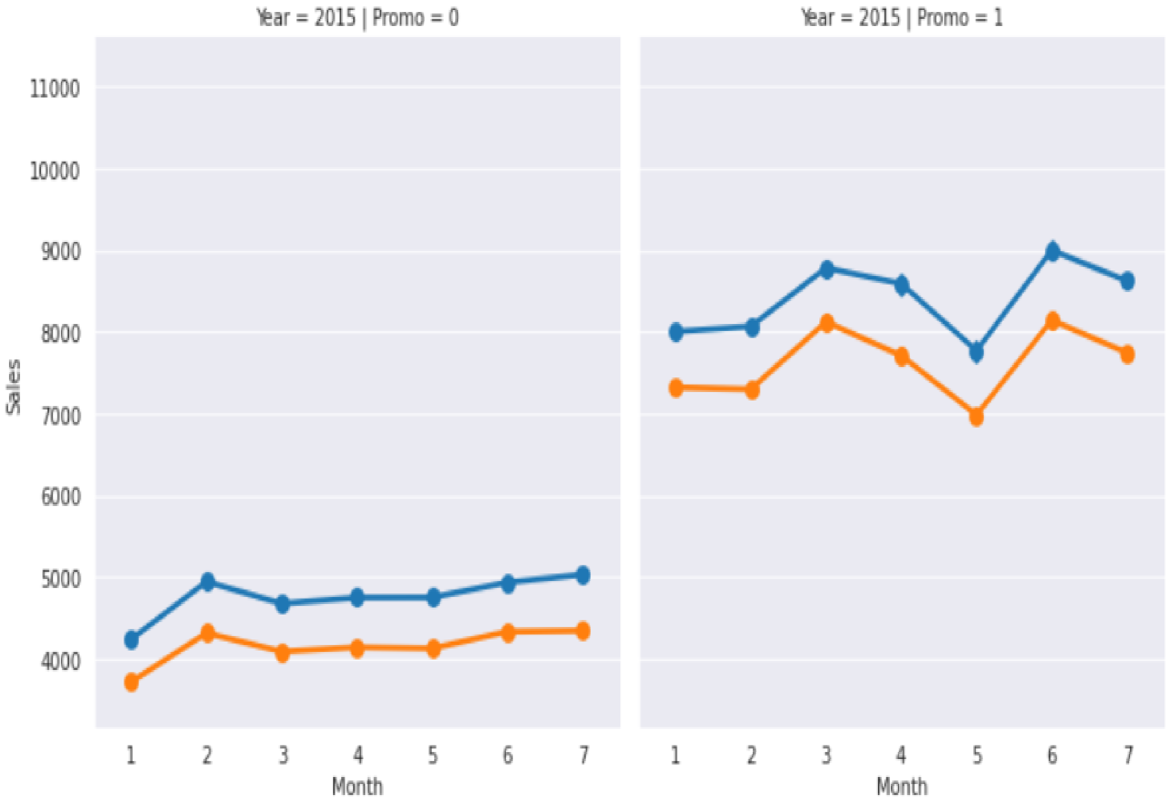
Effect of running promotional ads on sales :





EDA Continued…

Effect of running promotional ads on sales :

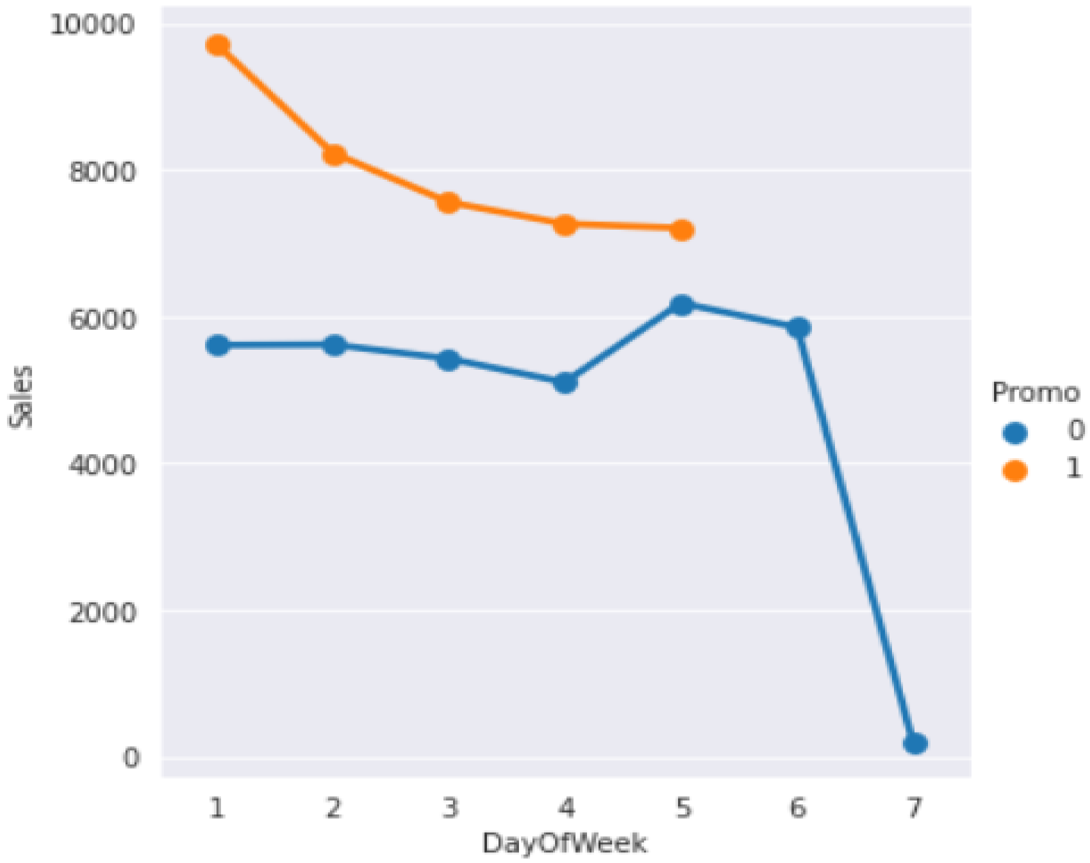


- Analysis clearly states that if we run promotional advertisement then there is jump in average sales, So we can say running promotion is beneficial for stores.

- If we look over the years, there is a slight increase Year over Year but we don't see any major change from 2013 to 2015 and we actually see a very similar pattern in the months over the years with major spikes first around Easter period in March and April then in Summer in May, June and July and then finally around the Christmas period in November and December.

EDA Continued…

Sales according to days of week :



The above graph shows sales according to days of week.

Only 33 stores are open on Sunday.

For Sunday since a very few stores opens on Sundays (only 33); if anyone needs anything urgently and don't have the time to get it during the week, he will have to do some distance to get to the open ones even if it's not close to his house. This means that those 33 open stores on Sunday actually accounts for the potential demand if all Rossman Stores were closed on Sundays. This clearly shows us how important it is for stores to be opened on Sundays.

EDA Continued…

Promotional Ads and Customers count:

Percentage increase in customer when promotional ads are running = 9.665641134987318

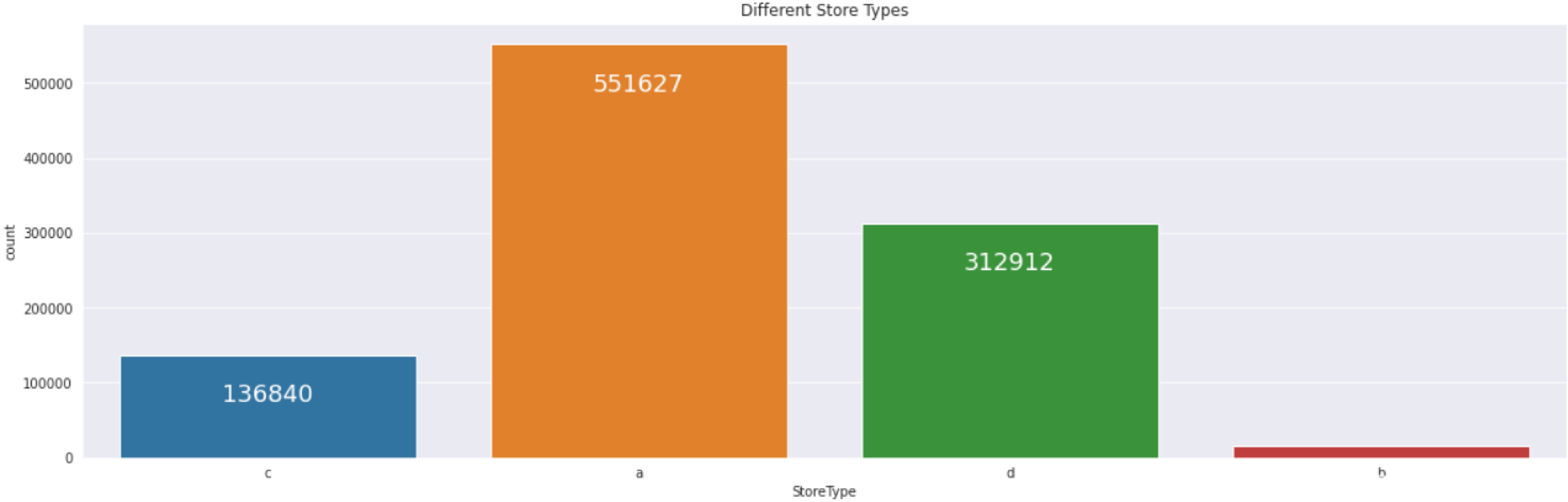
Running promotional ads gives 9% hike in customers.

Percentage Shop Running Continuous Promotions:

The Promo2SinceWeek,Promo2SinceYear and PromoInterval variables has almost 51% fill rate since they are actually NULL values because there are no continuous promotion for those stores.

EDA Continued…

Different Store Types

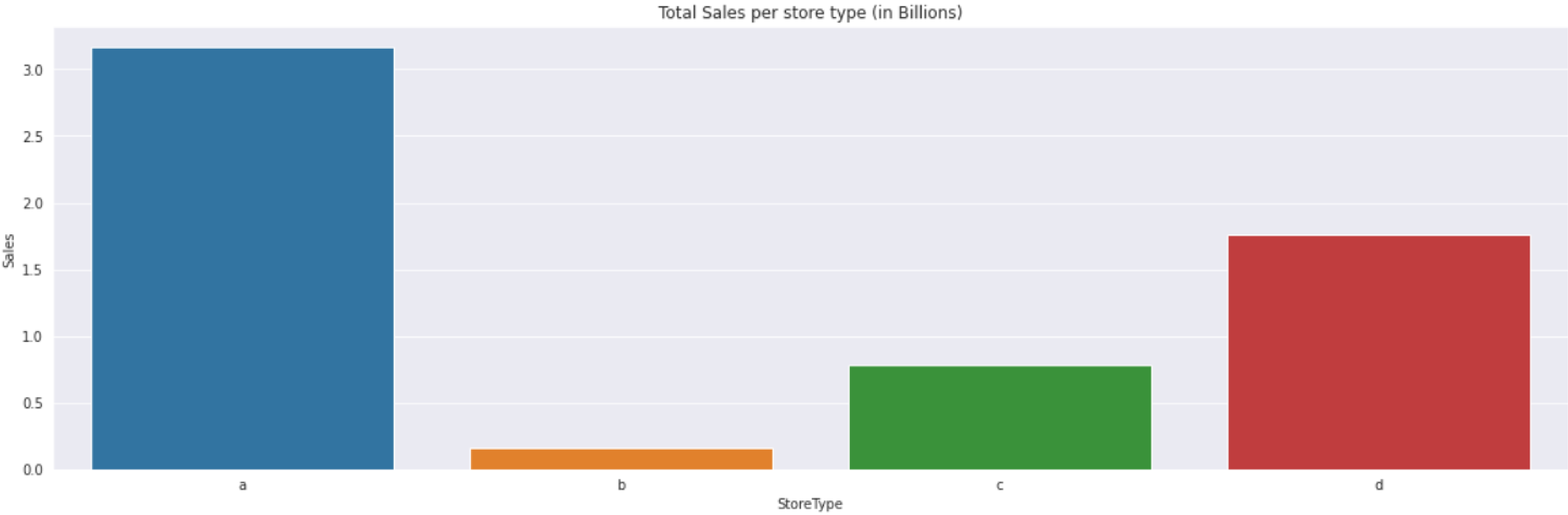


There are four different types of stores – a, b, c and d.

There are 551627 stores in type a, very few number of stores in type b, 136840 stores in type c and 312912 stores in type d.

EDA Continued…

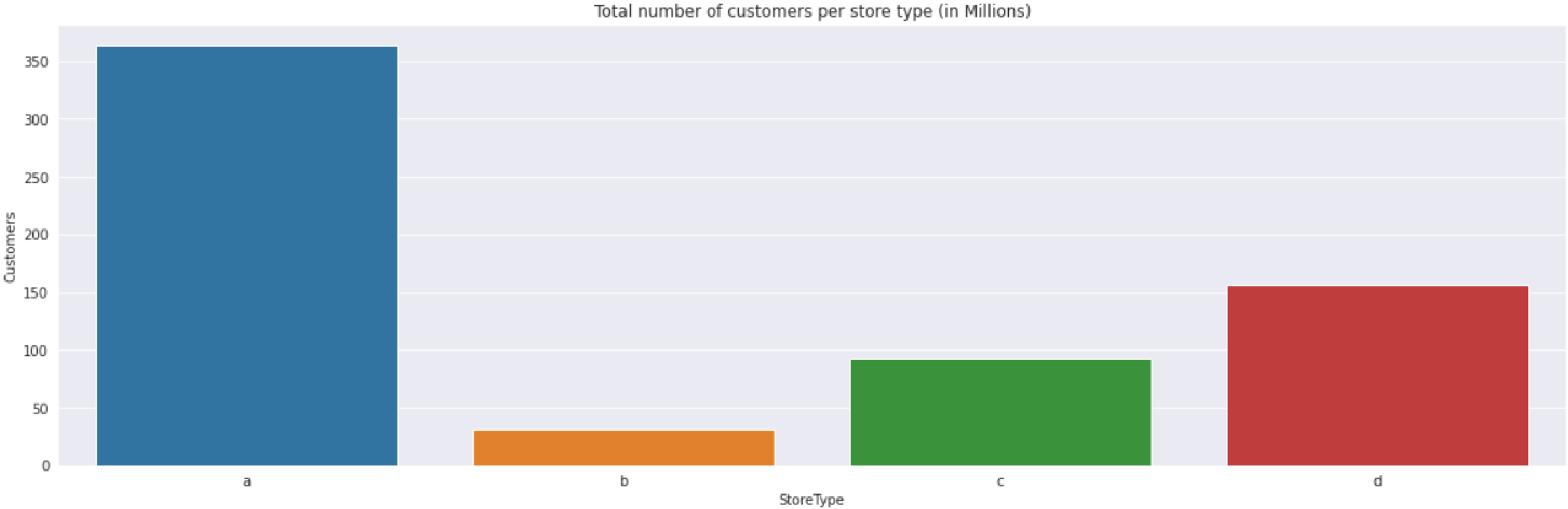
Total Sales per Store Type (in Billions)



Type a store is having more than 3 Billion sales.
Type b store is having less than 0.5 Billion sales.
Type c store is having less than 1 Billion sales.
Type d store is having less than 2 Billion sales.

EDA Continued…

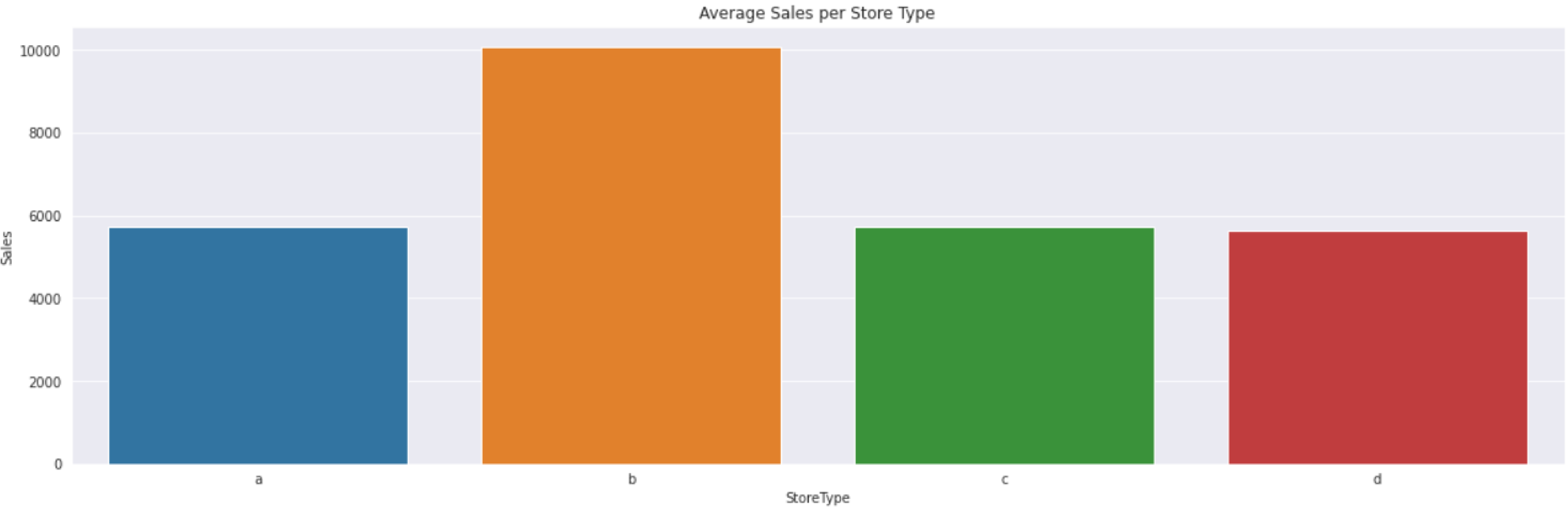
Total number of customers per Store Type (in Millions)



There are more than 350 Million customers for type a store.
There are less than 50 Million customers for type b store.
There are approximately 100 Million customers for type c store.
There are approximately 150 Million customers for type d store.

EDA Continued…

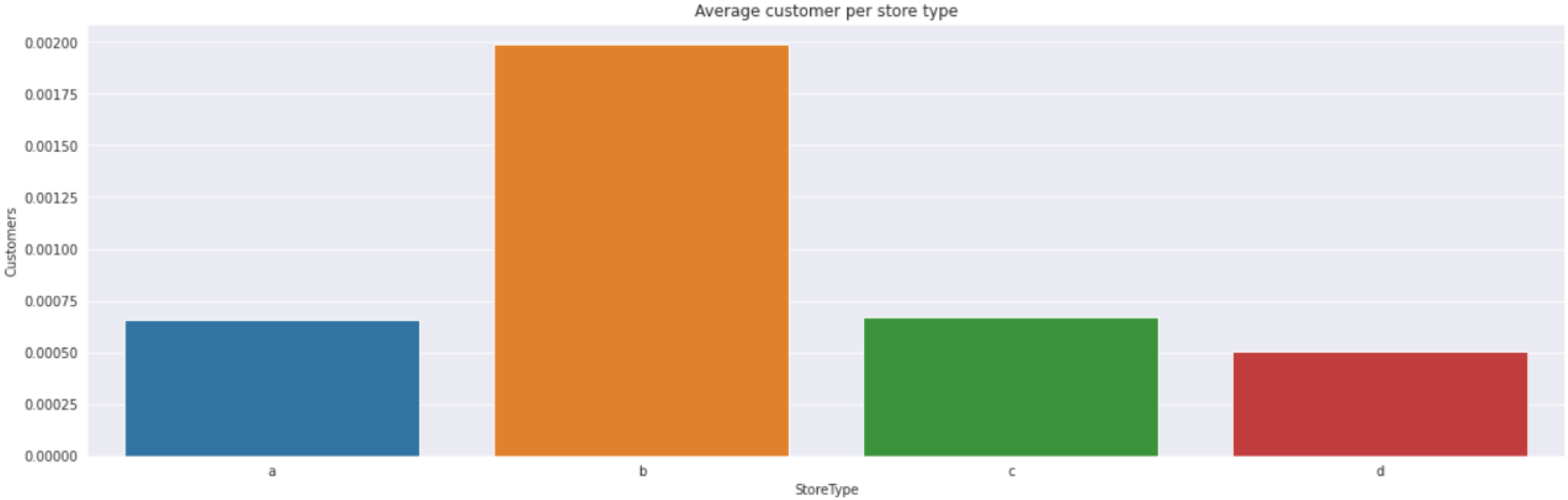
Average Sales per Store Type



The average sale of type b store is higher than all other types of stores because, according to number of stores there are more number of customers for store type b when compared with other types of stores.

EDA Continued…

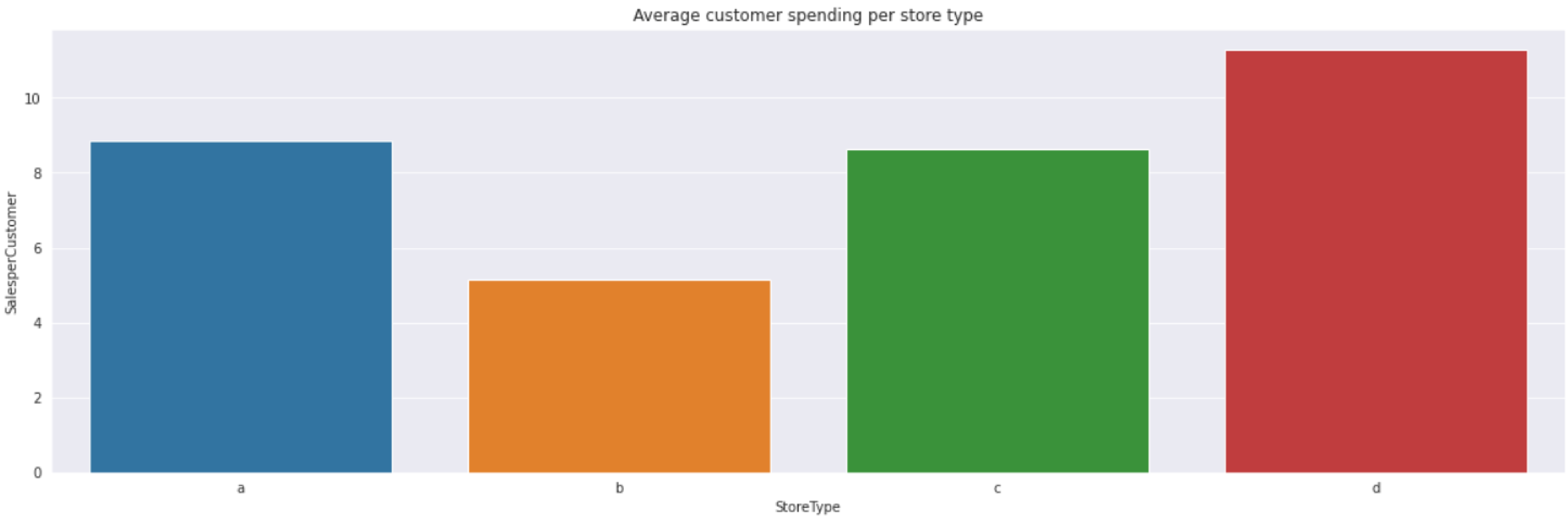
Average customer per Store Type



The average customer of type b store is higher than all other types of stores because, according to number of stores there are more number of customers for store type b when compared with other types of stores.

EDA Continued…

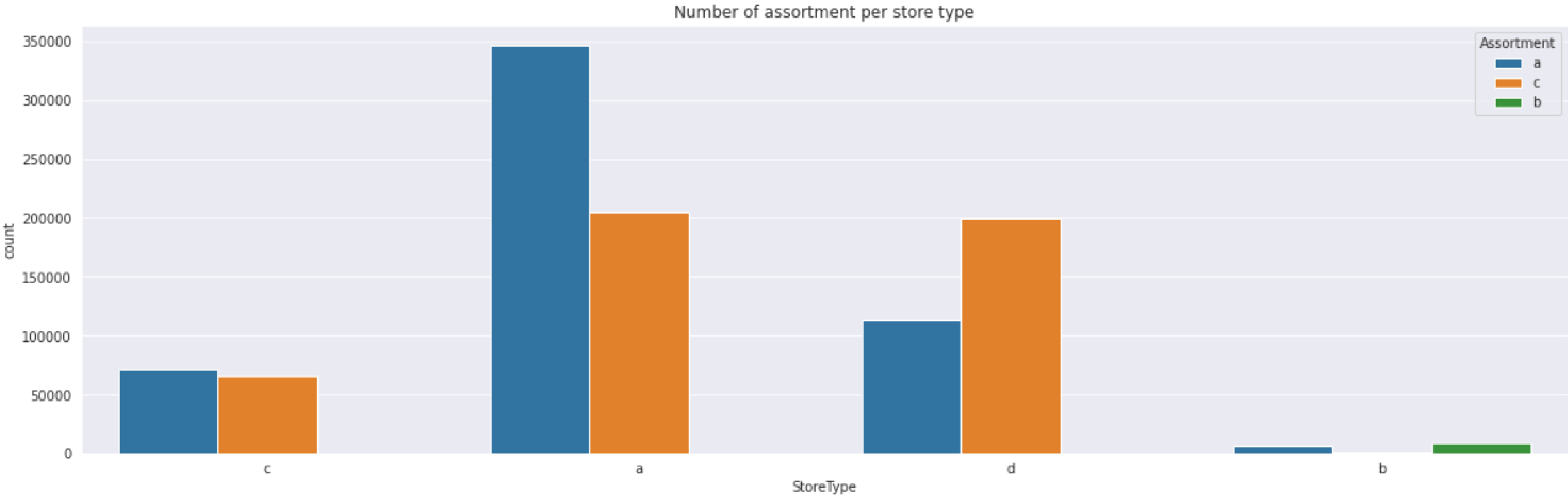
Average customer spending time per Store Type



According to above graph, the customers spend more time in store type d and less time in store type b.

EDA Continued…

Assortments and Store Types

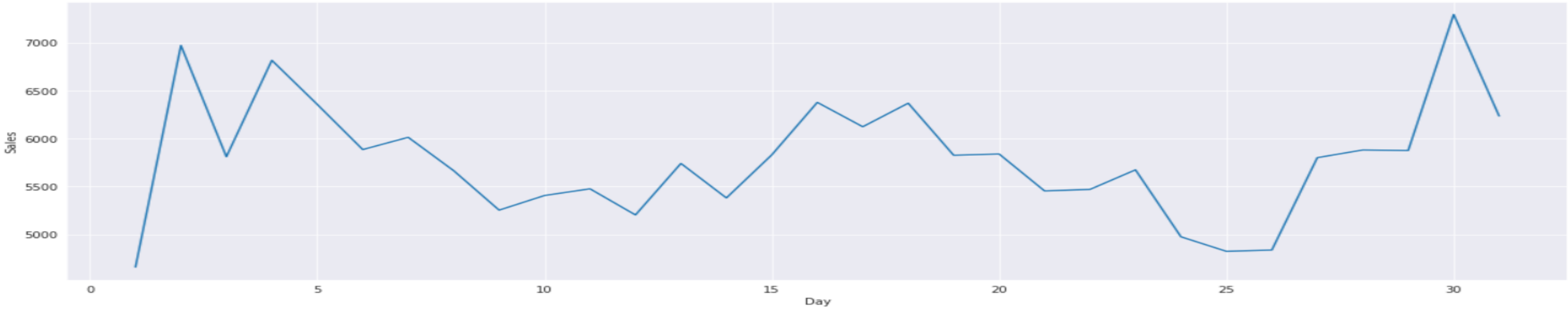


- We can clearly see here that most of the stores have either a assortment type or c assortment type.
- Interestingly enough Store Type d which has the highest Sales per customer average actually has mostly c assortment type, this is most probably the reason for having this high average in Sales per customer. Having variety in stores always increases the customers spending pattern.

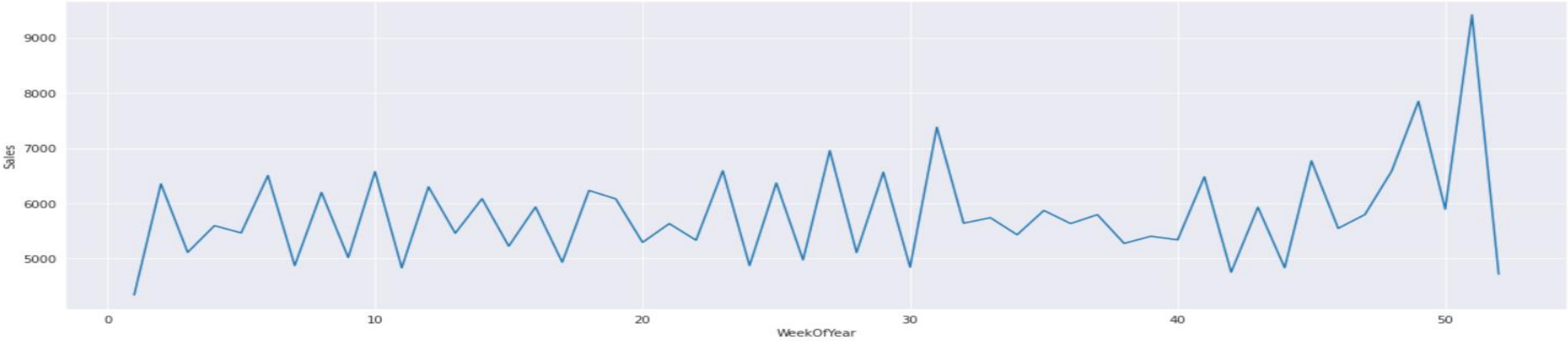
EDA Continued…

Sales Analysis

Sales per day



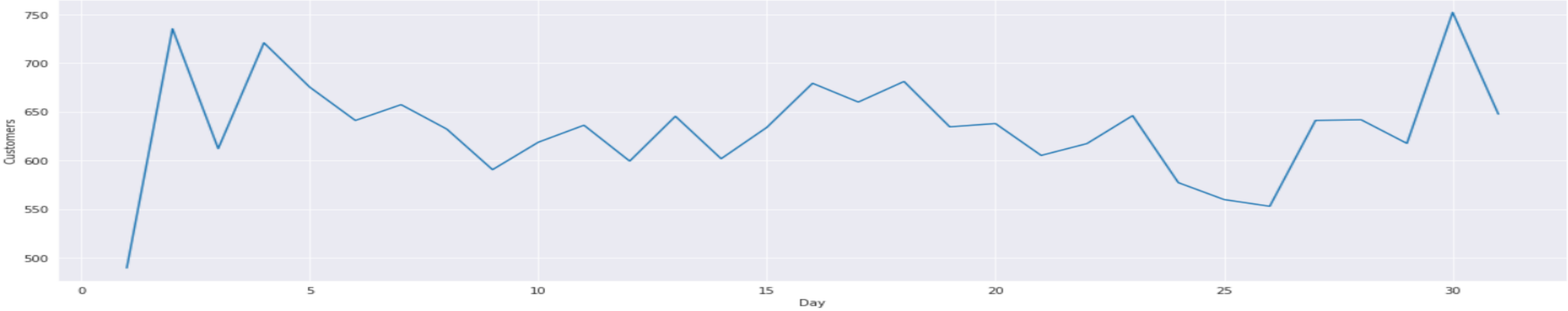
Sales per week



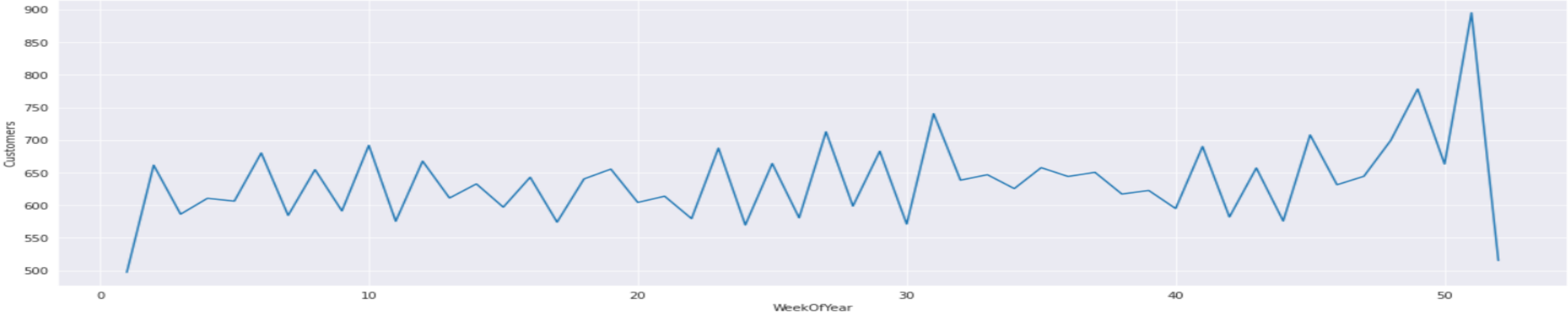
EDA Continued…

Sales Analysis

Customers per day



Customers per week



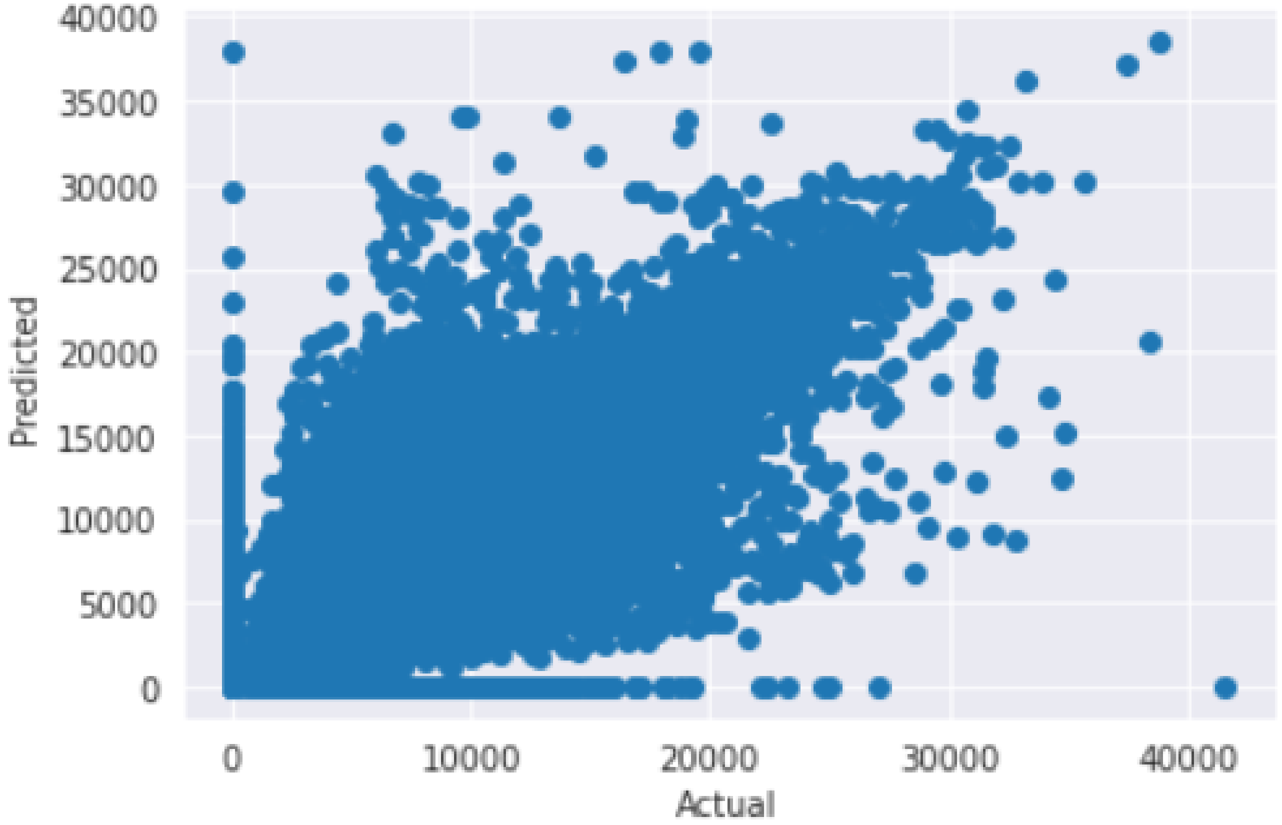
Preprocessing and feature engineering

- **Split the Numerical and Categorical Columns**
- **Scaling**
- **Encoding**
- **Split the Date**

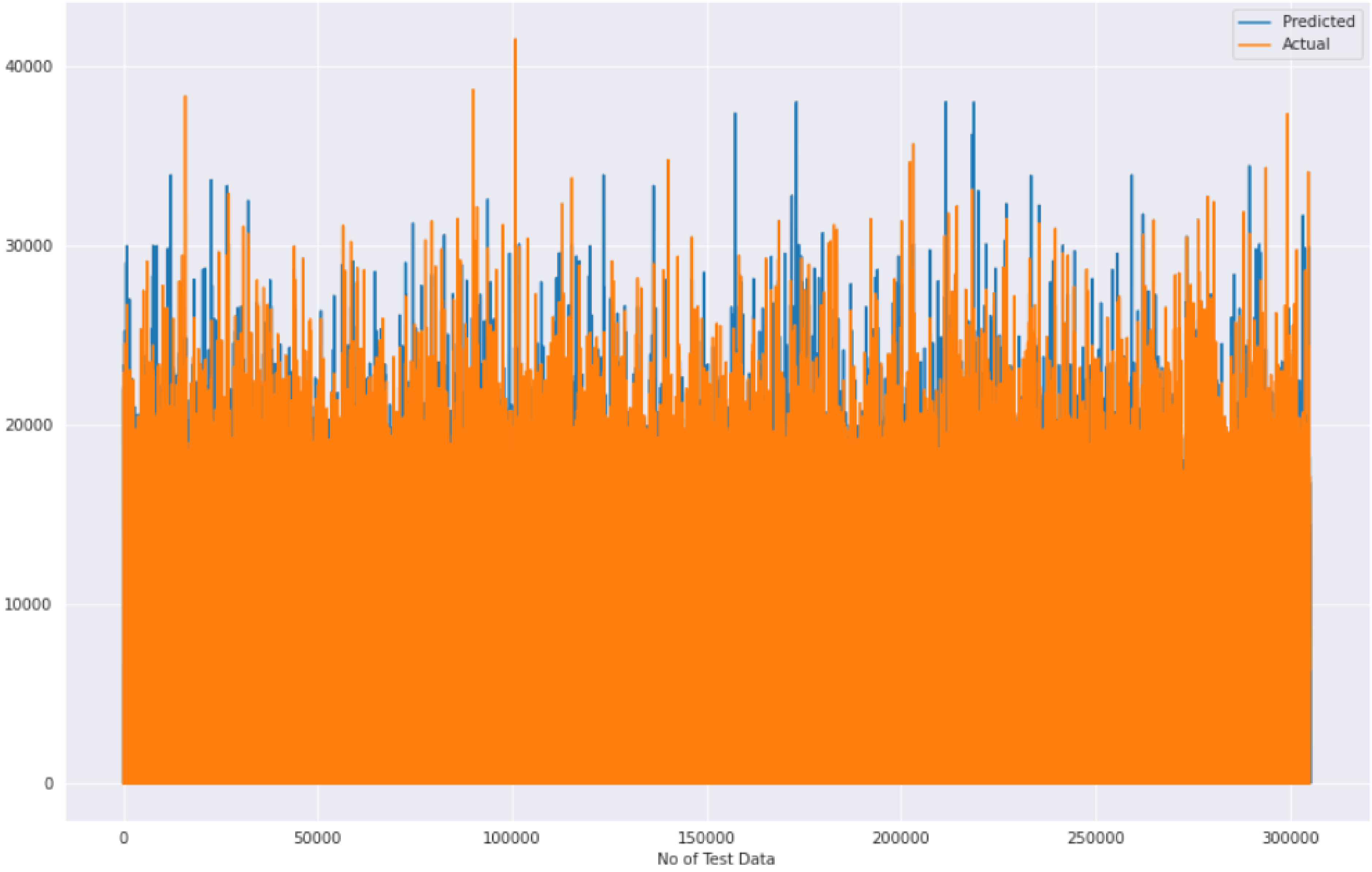
Data Modeling

Decision Tree Regression

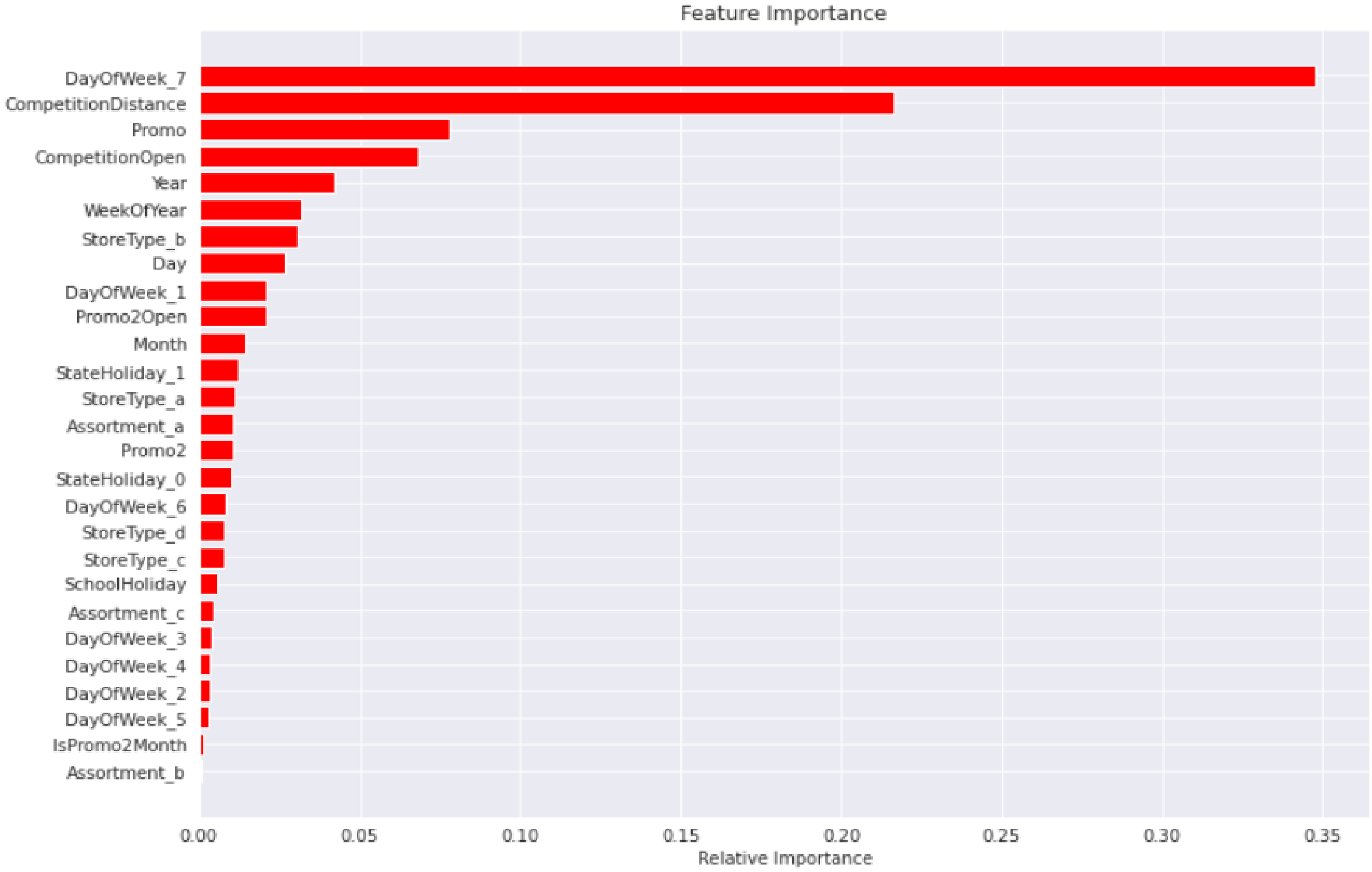
Scatter Plot



Actual vs Predicted values

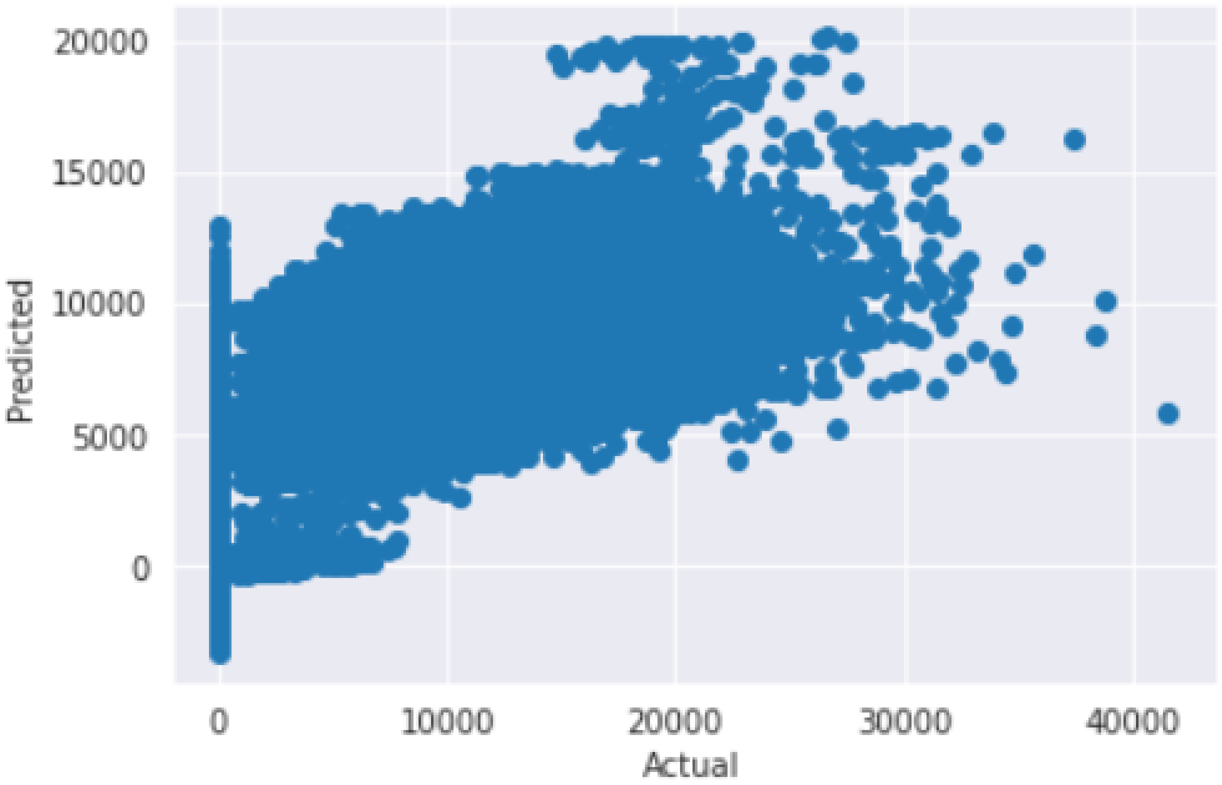


Feature Importance



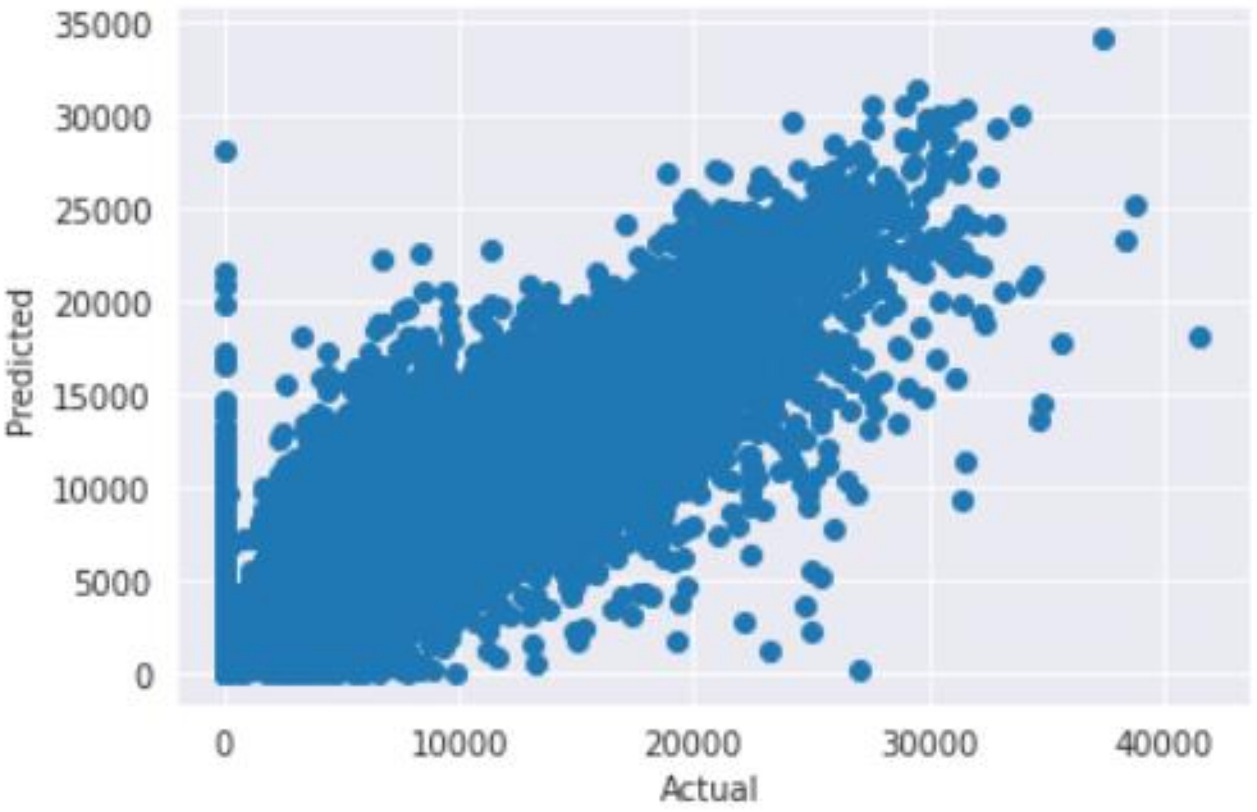
Gradient Boosting

Scatter Plot



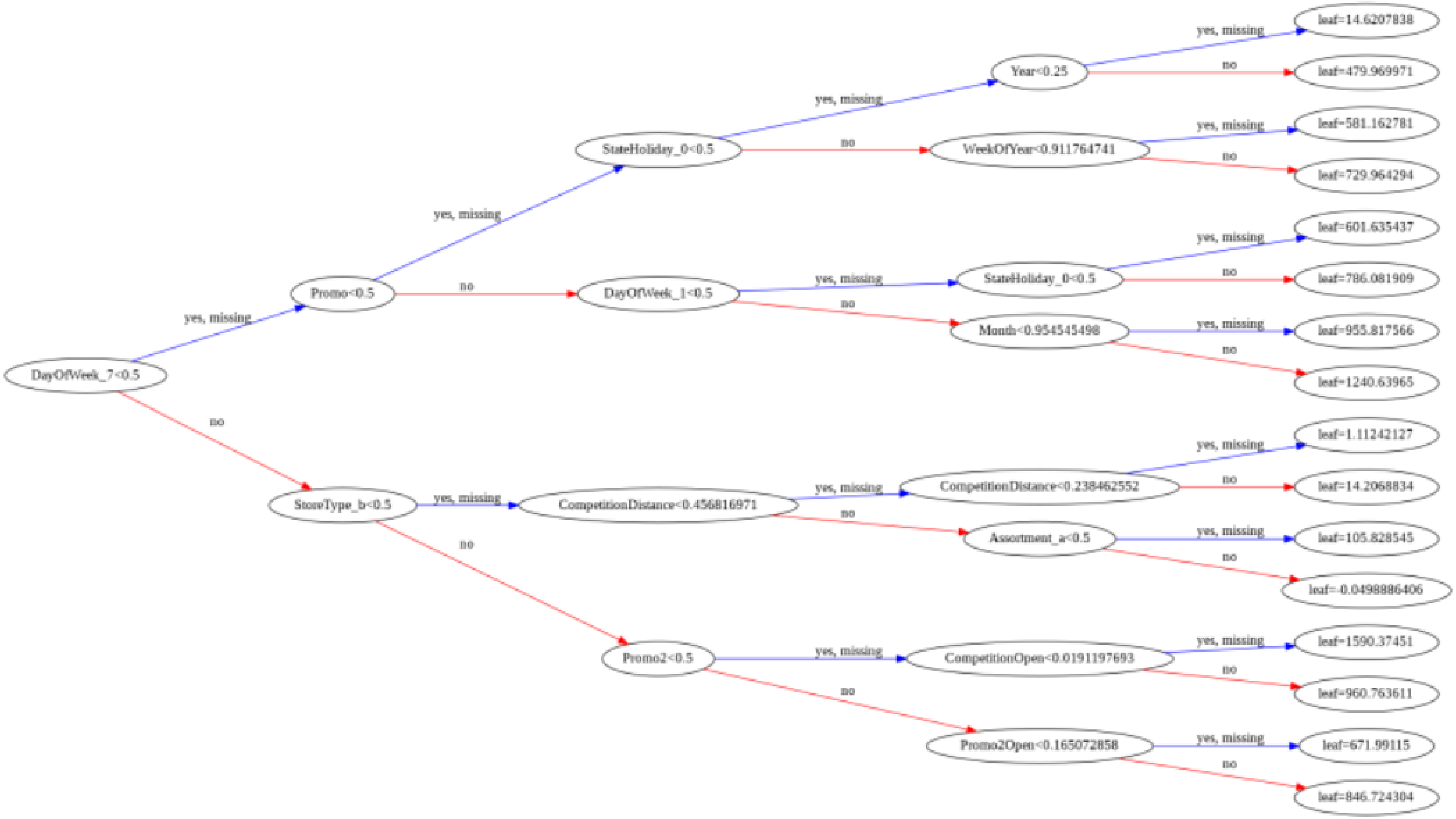
Random Forest

Scatter Plot

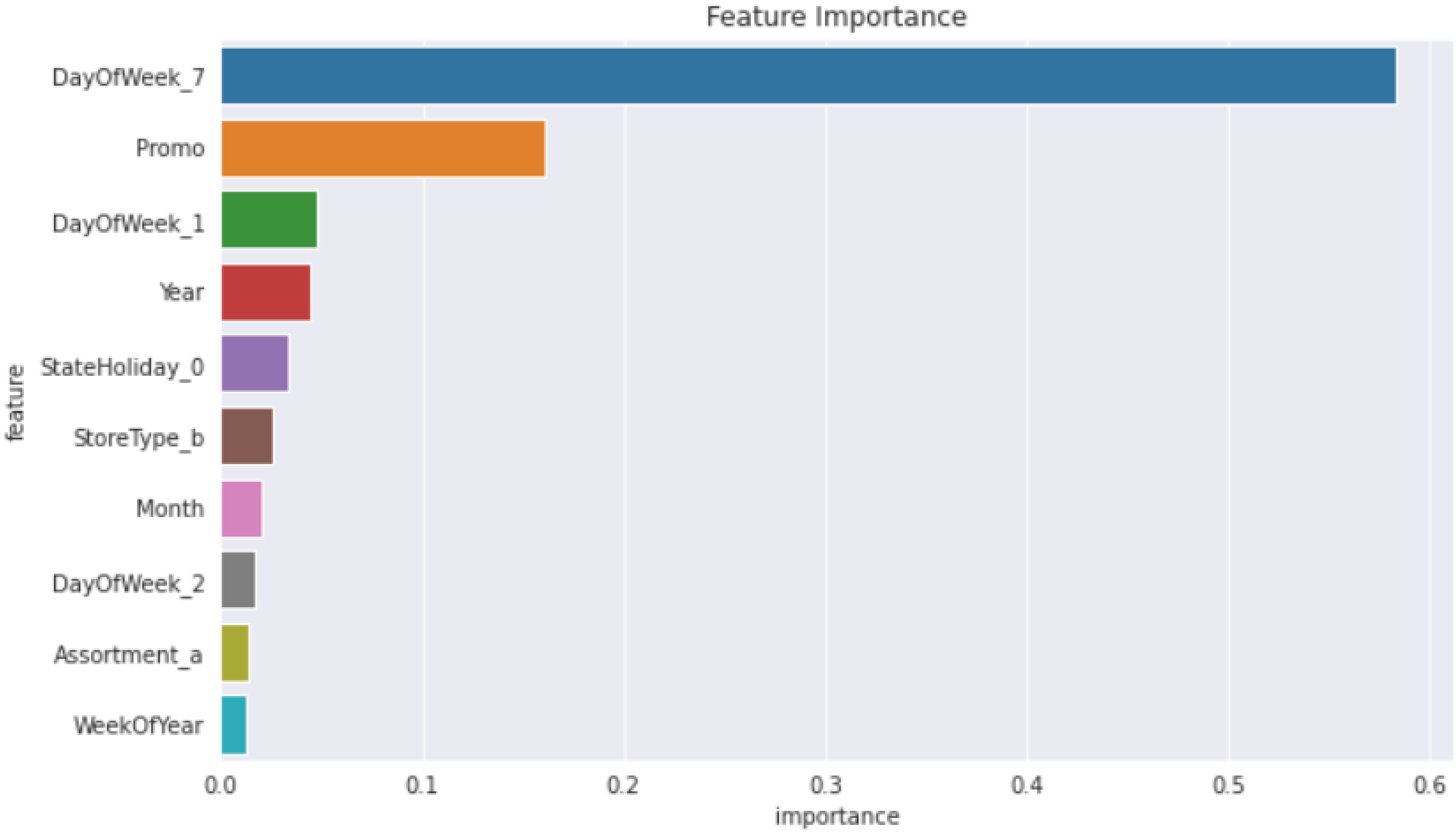


XG Boost

Tree



Feature Importance



Results

Algorithm Used	Adjusted R2 Score
Decision Tree Regression	85.98
Gradient Boosting	61.24
Random Forest	92.5

Conclusion

Conclusion on EDA:

- Over those two years, 172817 is the number of times that different stores closed on given days.
- From those closed events, 2263 times occurred because there was a school holiday.
- For Closed Event 30140 times it occurred because of either a bank holiday or easter or Christmas.
- After reading the description of the this task, Rossman clearly stated that they were undergoing refurbishments sometimes and had to close. Most probably those were the times this event was happening.
- The best solution here is to get rid of closed stores and prevent the models to train on them and get false guidance
- Analysis clearly states that if we run promotional advertisement then there is jump in average sales, So we can say running promotion is beneficial for stores.
- If we look over the years, there is a slight increase Year over Year but we don't see any major change from 2013 to 2015 and we actually see a very similar pattern in the months over the years with major spikes first around Easter period in March and April then in Summer in May, June and July and then finally around the Christmas period in November and December.
- For Sunday since a very few stores opens on Sundays (only 33);if anyone needs anything urgently and don't have the time to get it during the week, he will have to do some distance to get to the open ones even if it's not close to his house. This means that those 33 open stores on Sunday actually accounts for the potential demand if all Rossman Stores were closed on Sundays. This clearly shows us how important it is for stores to be opened on Sundays.
- Store type a has the highest number of stores, sales and customers from the 4 different store types. But this doesn't mean it's the best performing Store type.

- When looking at the average sales and number of customers, actually it is Store type b who was the highest average Sales and highest average Number of Customers. One assumption could be that if b has only 17 stores but such a high amount of average sales and customers, whereas a would be smaller in size but much more present.
- Surprisingly it is Store Type d who has the highest average spending per Customer, this is probably explained by an average competition distance higher than the rest which means each customer will buy more since he knows there isn't a lot of similar shops around.
- We can clearly see here that most of the stores have either a assortment type or c assortment type.
- Interestingly enough Store Type d which has the highest Sales per customer average actually has mostly c assortment type, this is most probably the reason for having this high average in Sales per customer. Having variety in stores always increases the customers spending pattern.

Conclusion on Data Modeling:

- We can understand from this project the flexibility and robustness of a decision tree based model like RandomForest which helped us predict the Store Sales of Rossman based on attributes that defines each store and its surroundings.
- As we can see, it always delivers a good predicition score while not having a lot of modifications and difficulties capturing the patterns hidden in the data. Fortunately we had a train set that was large enough for it to converge but in general RandomForest performs not so bad on small sets since its resampling method (bagging) and its random production of trees allow the bias to remain not so high and in this case always performs good on unseen data where as XGboost has tendency to overfit if not gently and smartly tuned.
- I believe using hyperparameter optimization techniques like Gridsearch and RandomizedSearch is crucial to any Machine Learning problem since it allows the algorithm to not just limit itself on its defaulted parameters but to discover new opportunities of combining those parameters to reach a local optima while training on the data.

Thank You!!

☺