Capstone Project

Retail Sales Prediction

Team Members

Gaurav Kinhikar

Amalkrishna N

OVER VIEW

- Data restoration
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Modelling
- Conclusion

Problem Statement :

- 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.
- You are provided with historical sales data for 1,115 Rossmann stores.
- The task is to forecast the SALES

Data summary :

We have two datasets

* Stores having different features – “Store”
* Stores with sales on a particular day – “Rossmann Stores Data”

- **Columns -** 'Store', ‘Day Of Week', 'Date', 'Sales', 'Customers', 'Open', 'Promo', 'State Holiday', 'School Holiday', 'Year', 'Month', 'Day', 'Week', 'Week Of Year', 'Store Type', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2', 'Promo2SinceWeek', 'Promo2SinceYear', 'Promo Interval'

2. Data Cleaning

- Check whether any duplicate entries
- Dealing With Missing Values
- Count missing values in each dataset
- Replacing missing values with appropriate values
- Joining Tables
- Drop part Of Data Where Might Cause Bias
- Removing zero sale entries

- Column “CompetitionOpenSinceMonth” and “CompetitionOpenSinceYear” had null values – after exploring I got to know that I should replace the null by mode in this case.

- Column “CompetitionDistance” had null values – after exploring I got to know that I should replace the null by median in this case.

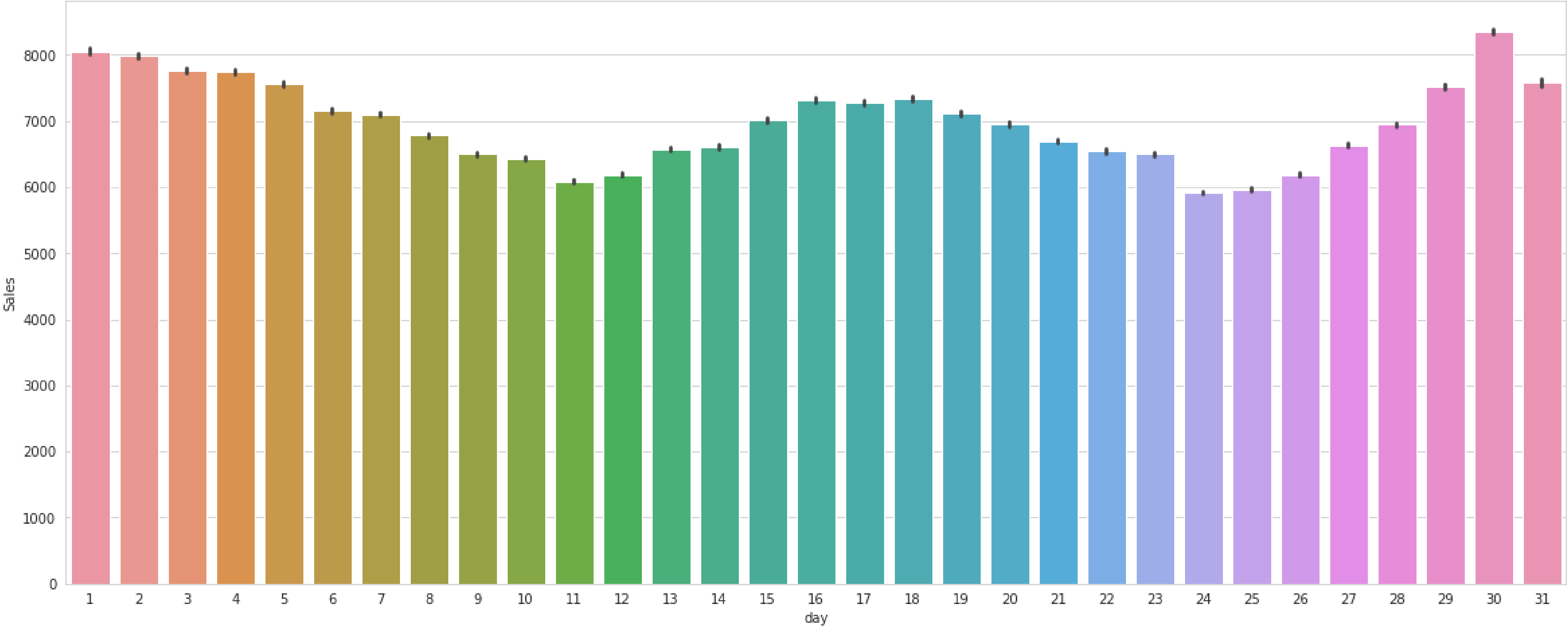
- Column “Promo2SinceWeek”, “Promo2SinceYear”, “PromoInterval” was having a lot of Null values, because those stores have not started any promotion, so they should be zero for our Dataset.

- I have split the “Date” into Month, Year, week of year, day, and Week

- Finally I have merged the two Datasets into one, named “df”

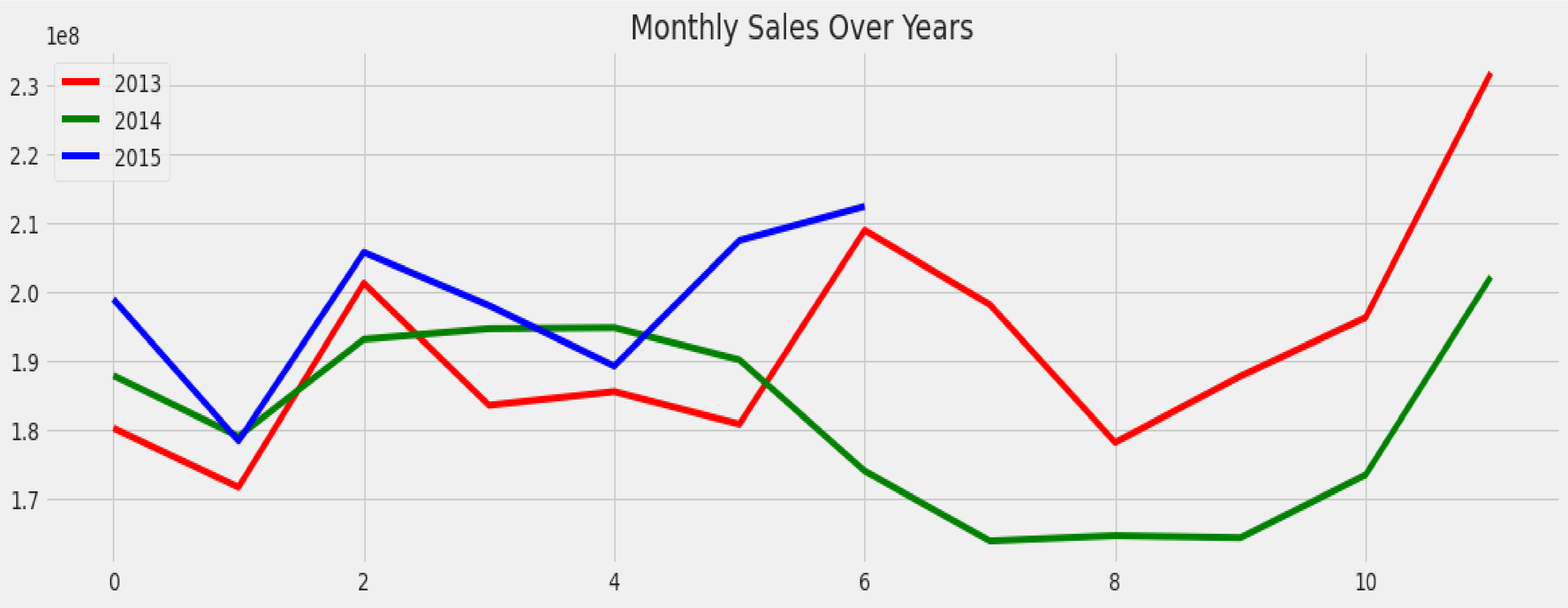
3. Exploratory Data Analysis (EDA)

SALES OVER DAYS IN MONTH



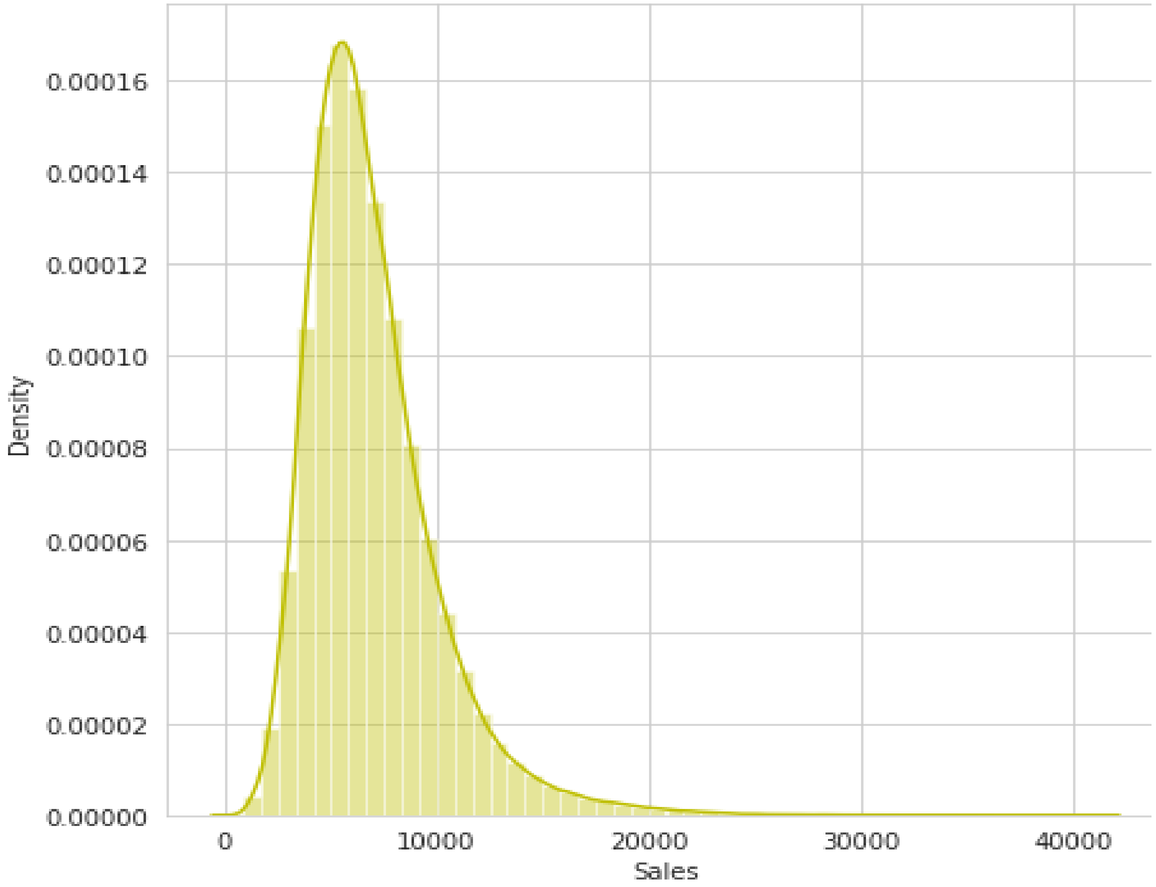
- Sales over month showing that a clear little sinusoidal trend on the graph
- Sales are at its peak in starting days and ending days of month

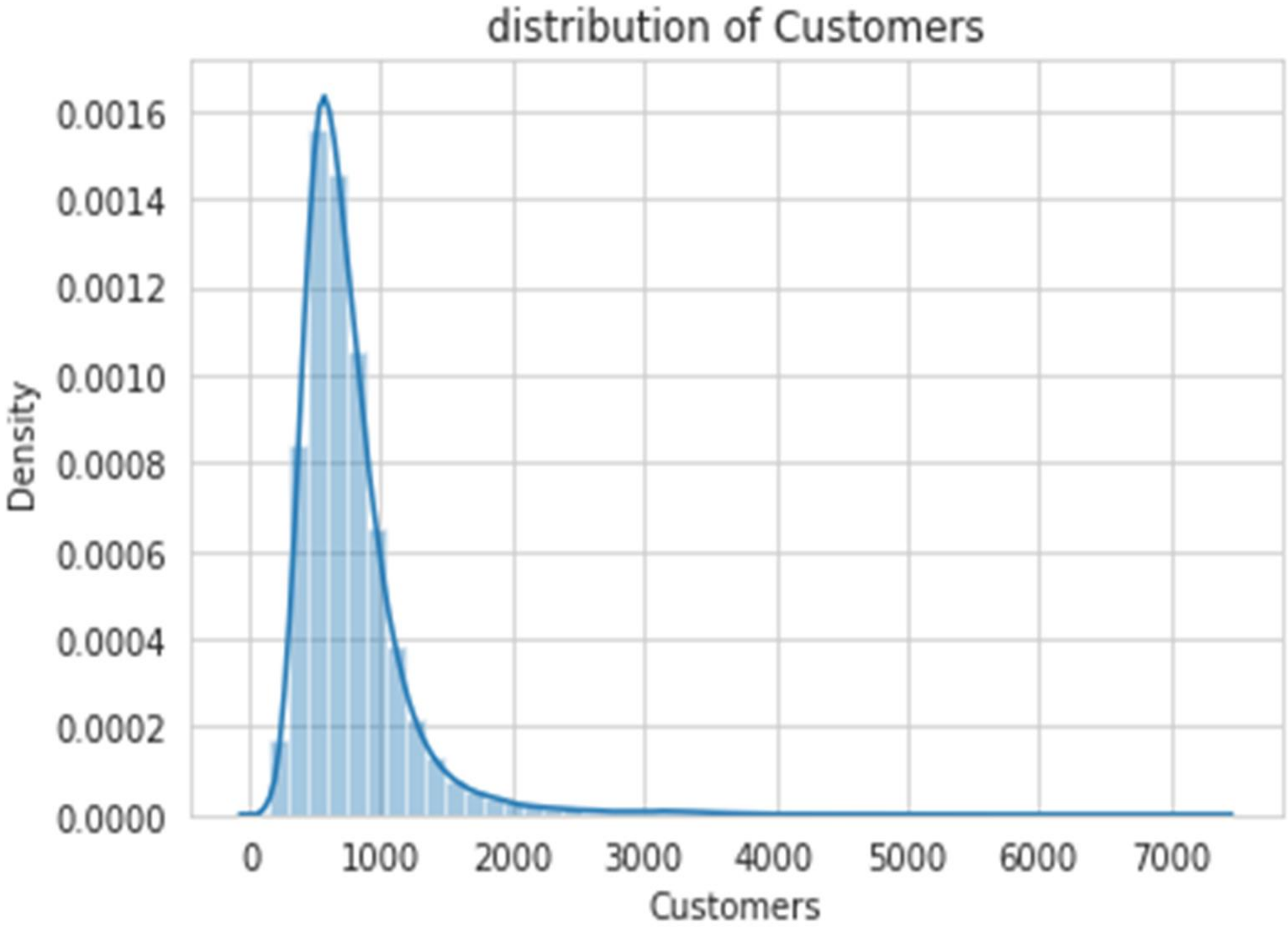
MONTHLY SALES



- Second month of every year will make a dip as compared to the first month.
- August, September and October are the months with low monthly sales.
- After October in every year there is a sharp hike in monthly sales

DISTRIBUTION OF SALES

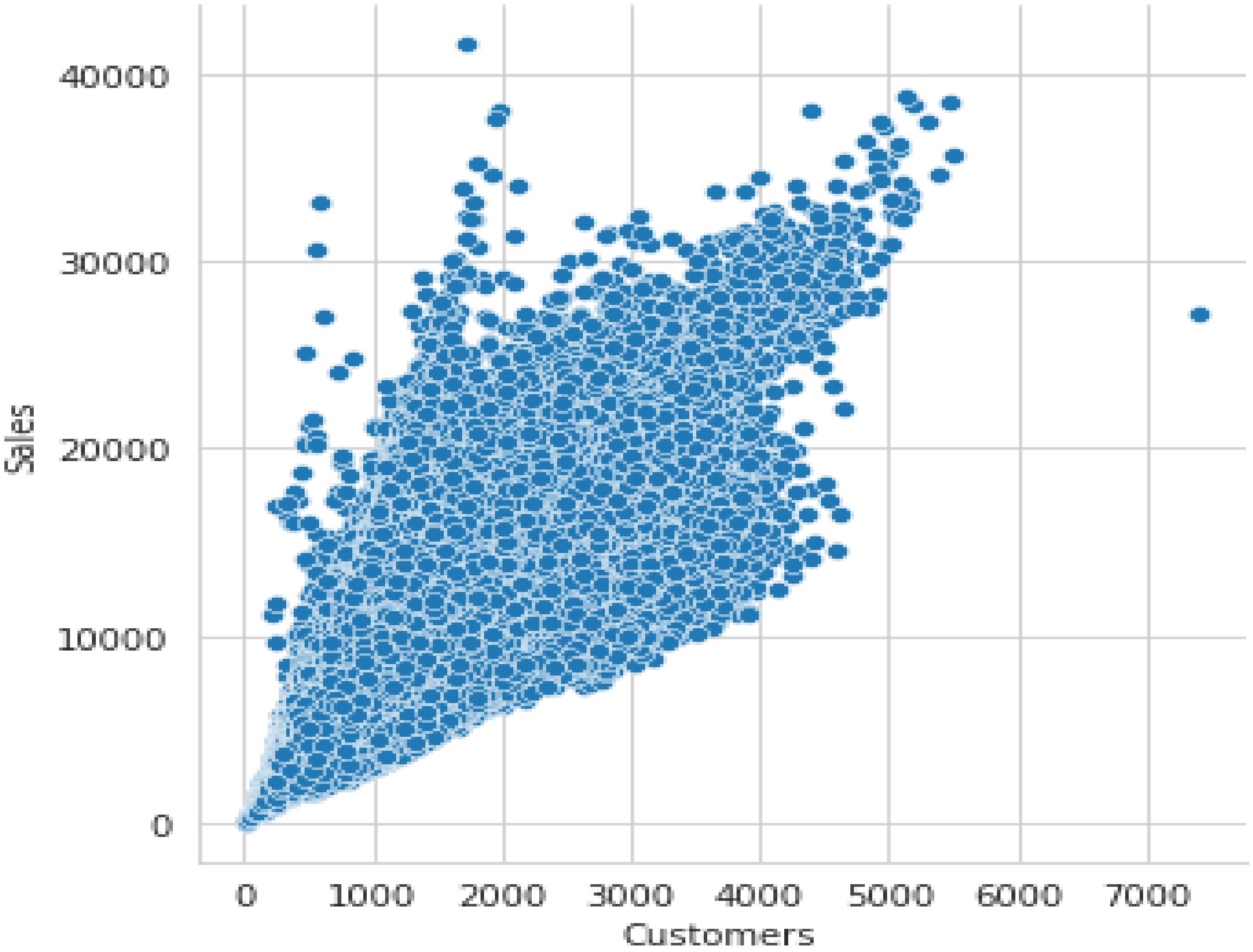




- Conclude that a perfect right skewed distribution of sales in this graph
- Most of the sales are not concentrated on the higher average ranges

- Again customer distribution also a good right skewed model
- Customer numbers on stores are also not concentrated to higher vales.
- Stores with higher number of customers are exceptional or lesser in number

RELATIONSHIP B/W SALES & CUSTOMERS

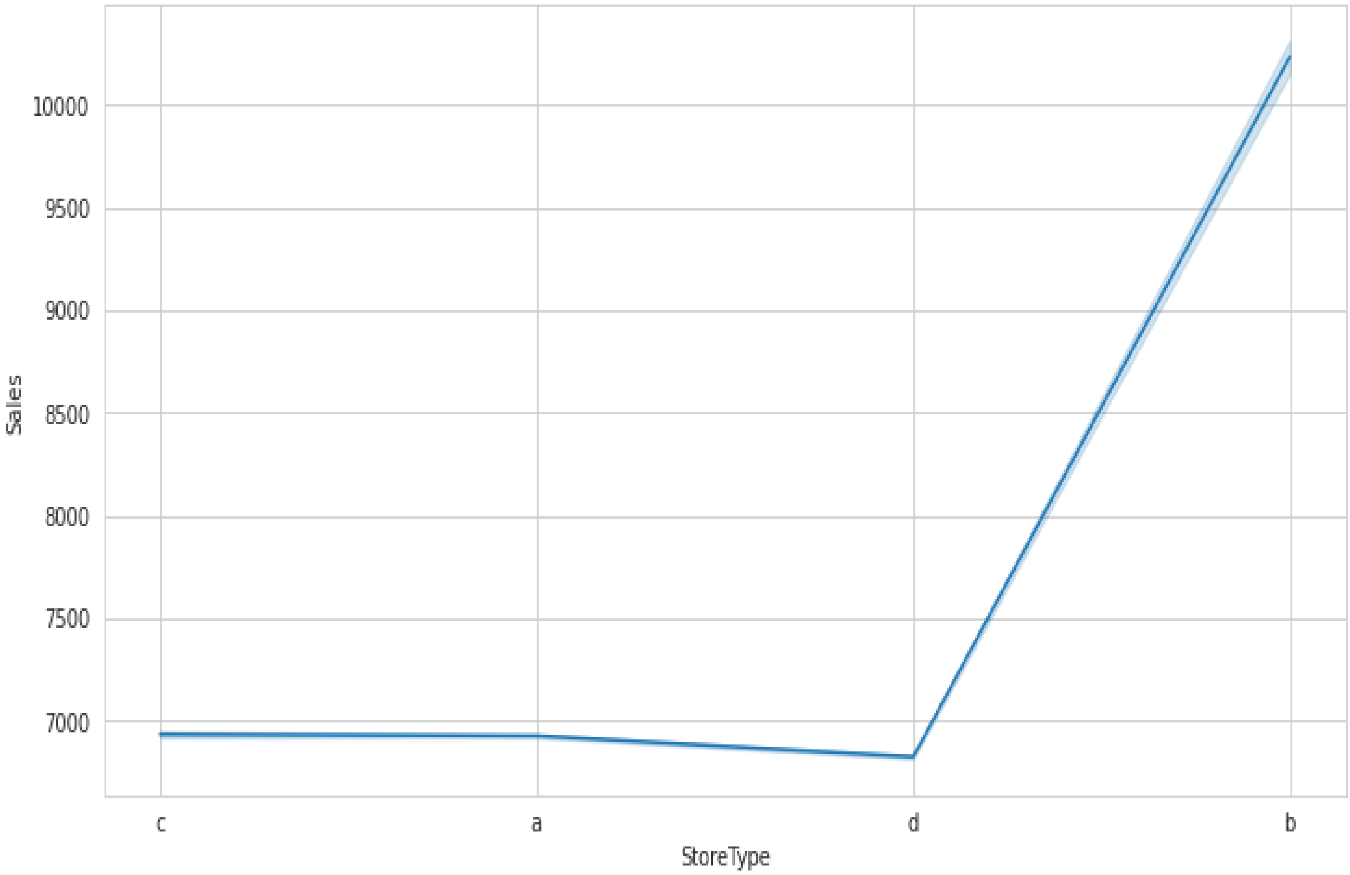


- Scatter plot between sales and customer count plots a clear idea about the relationship between sales and number of customers

- It is clear that the sales and customers are strongly in a linear relationship

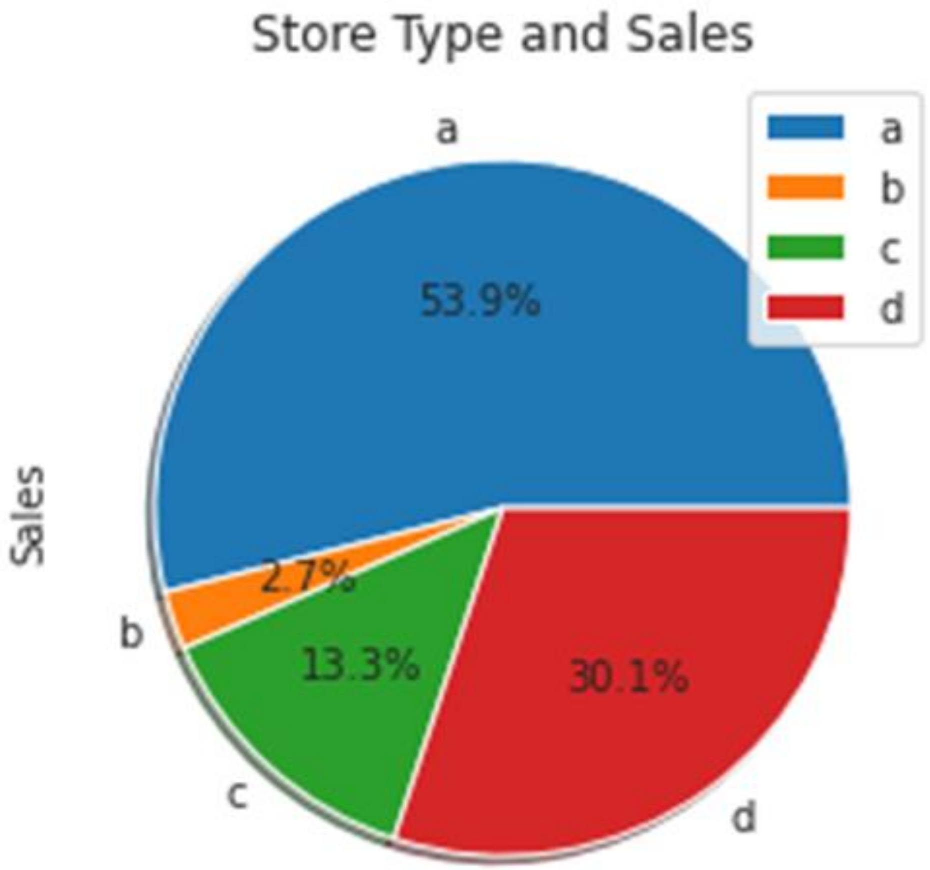
- Customers count will highly affect the sales before any other features

AVG. SALES V/S STORE TYPE



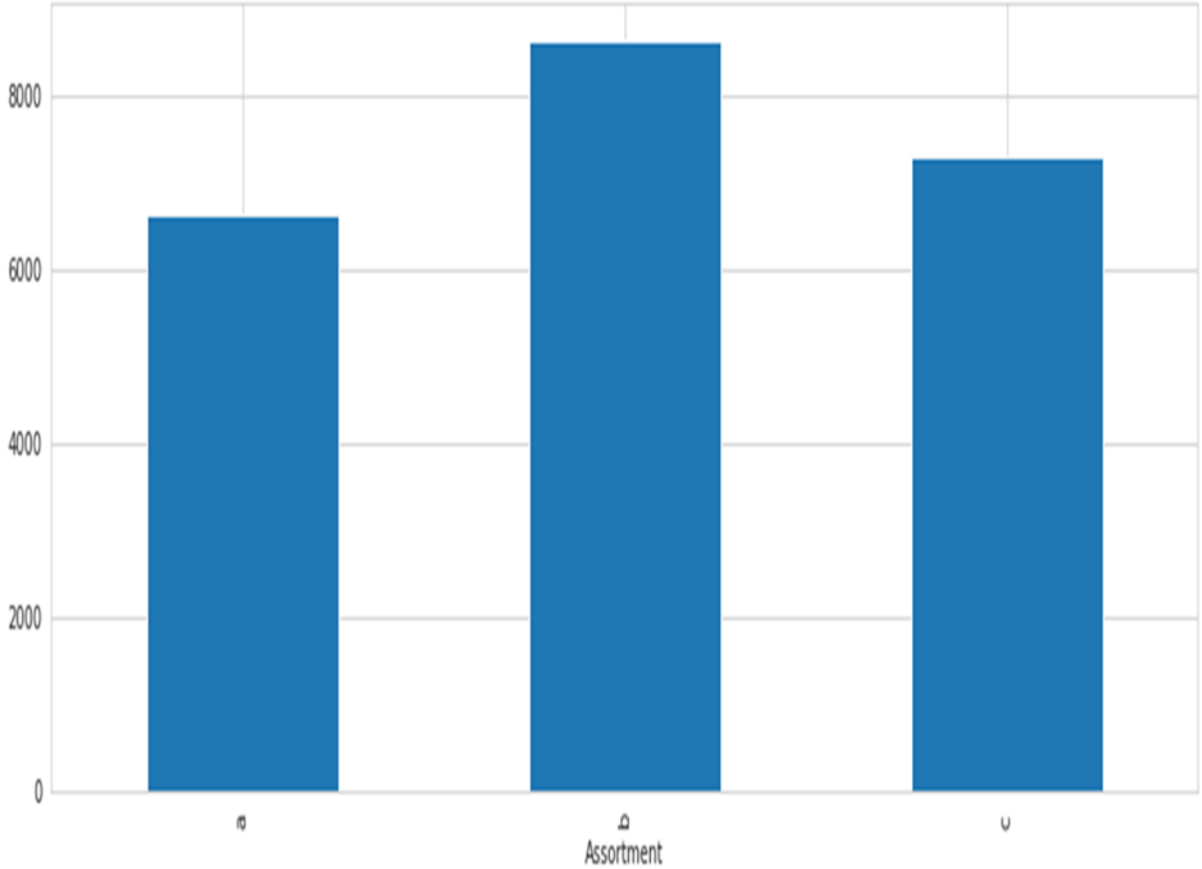
- stores a, c, d making almost same range of mean sales
- store type ‘b’ making a high mean sales value that is around 1.5 times more sales than a, c, d stores

TOTAL SALES

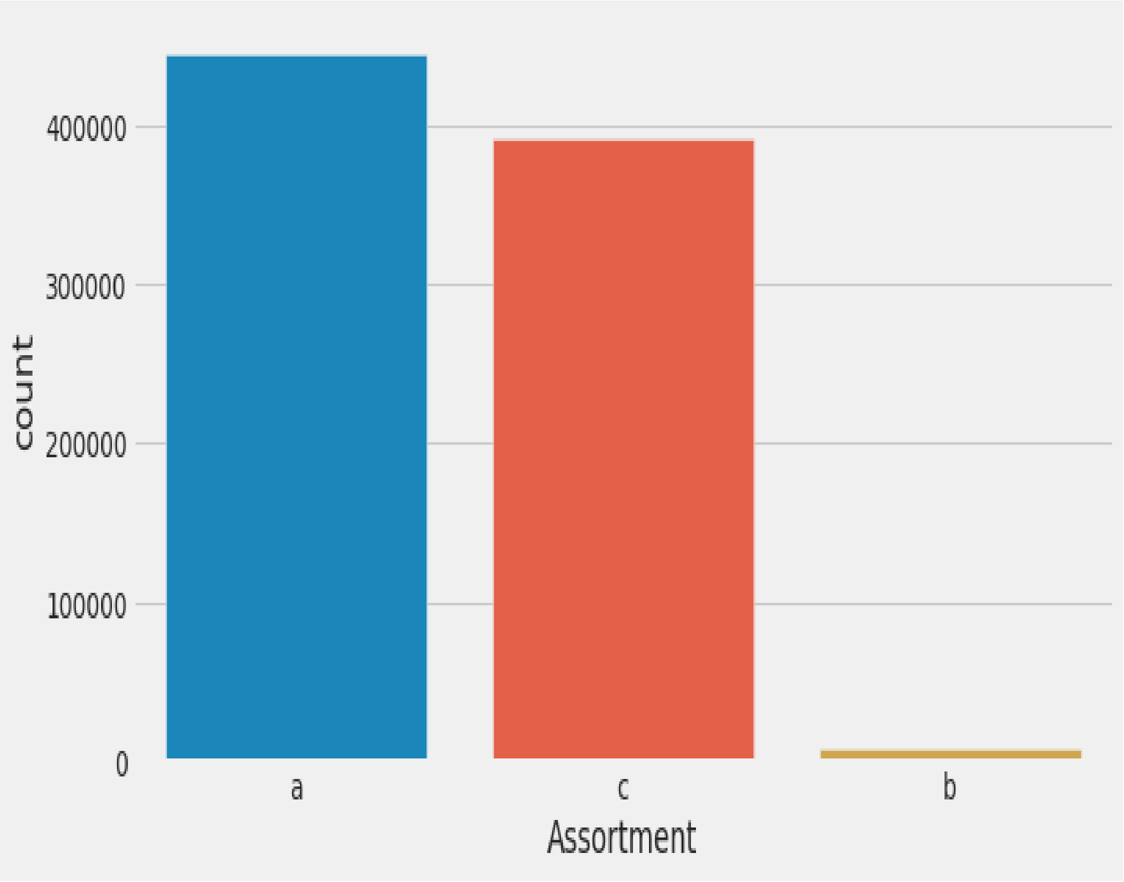


- Store type ‘a’ have a major part on total sales 53% of the total sales are come from store type ‘a’
- followed by store type ‘d’ with 30% and type ‘c’ 13%, type ‘b’ 3% so on

Avg. sales in assortment & total of assortments

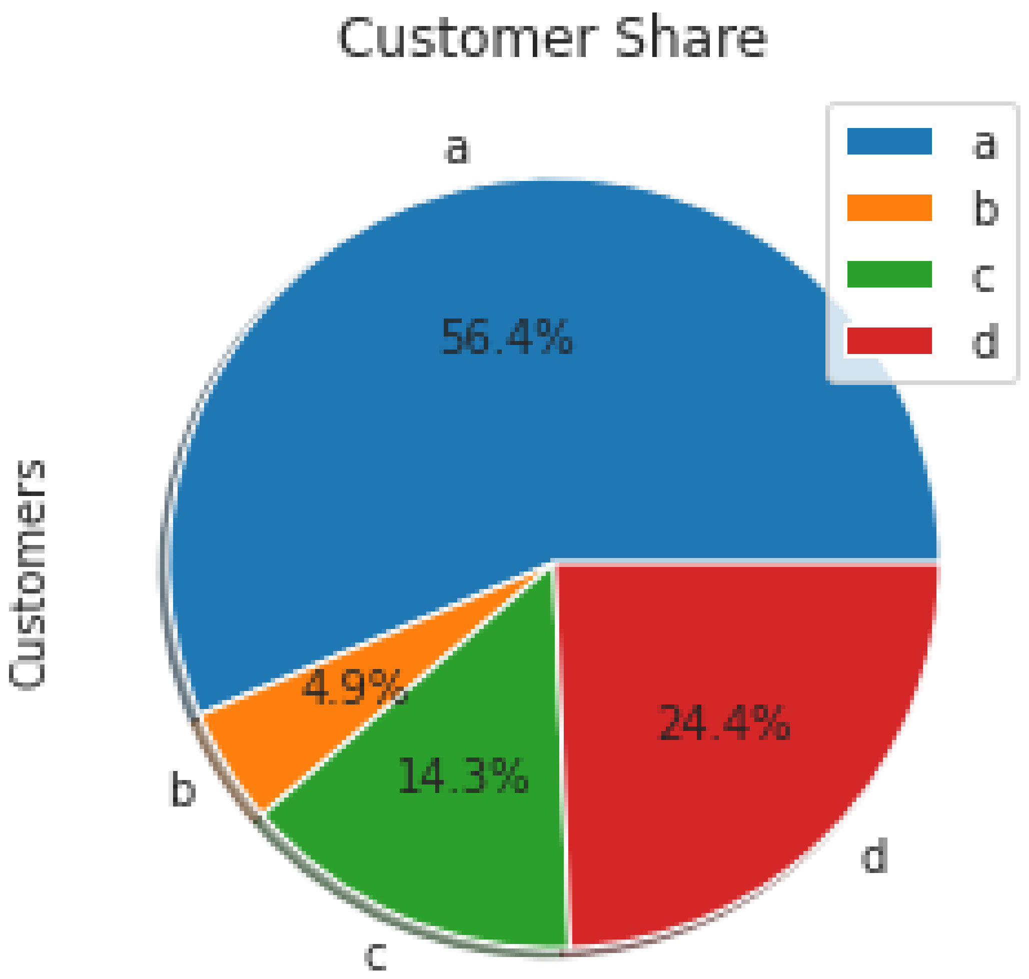


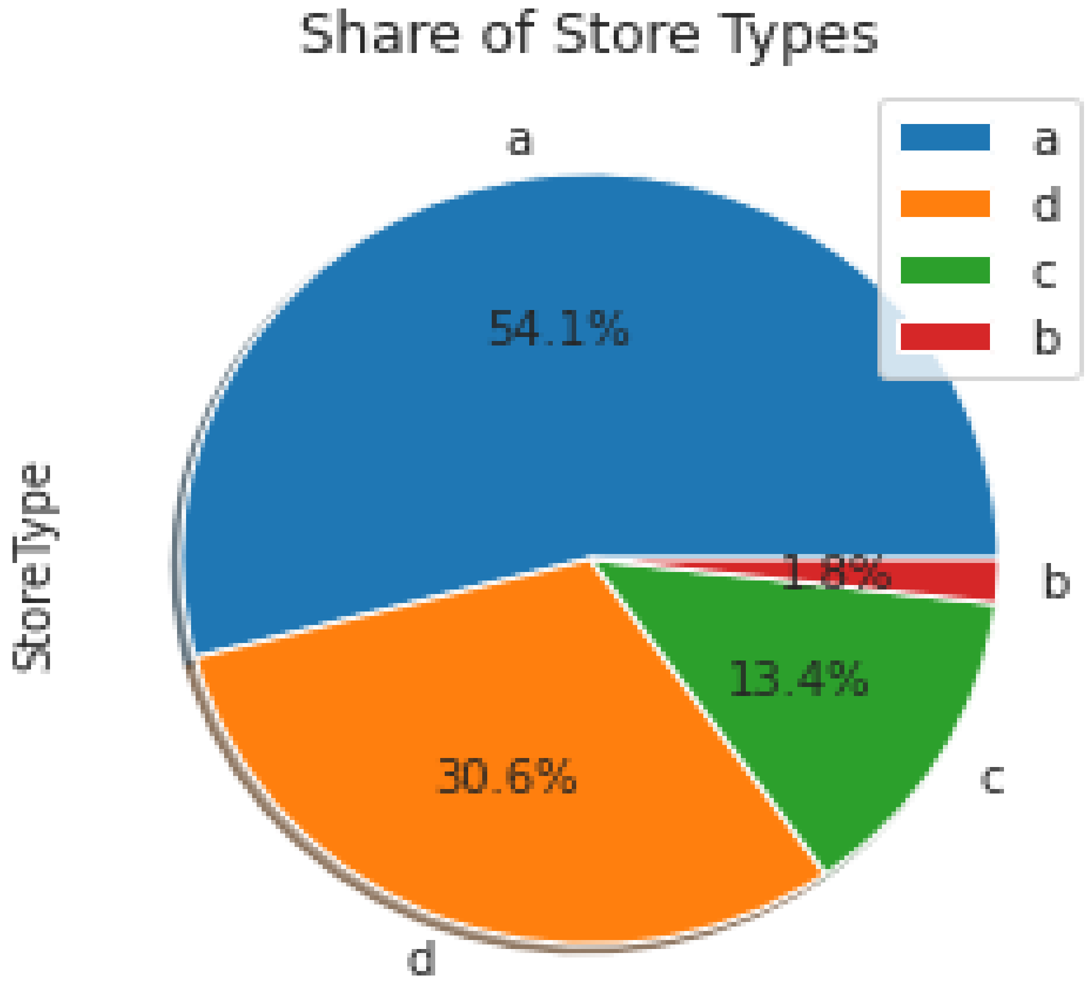
Mean sales over each assortment type of store



- Assortment ‘b’ have more mean value sales then ‘c’ and at last assortment ‘a’
- Around 850$ average sales happening in assortment ‘b’ stores.
- Assortment ’c’ and ‘a’ have sales in between 800$ and 600$

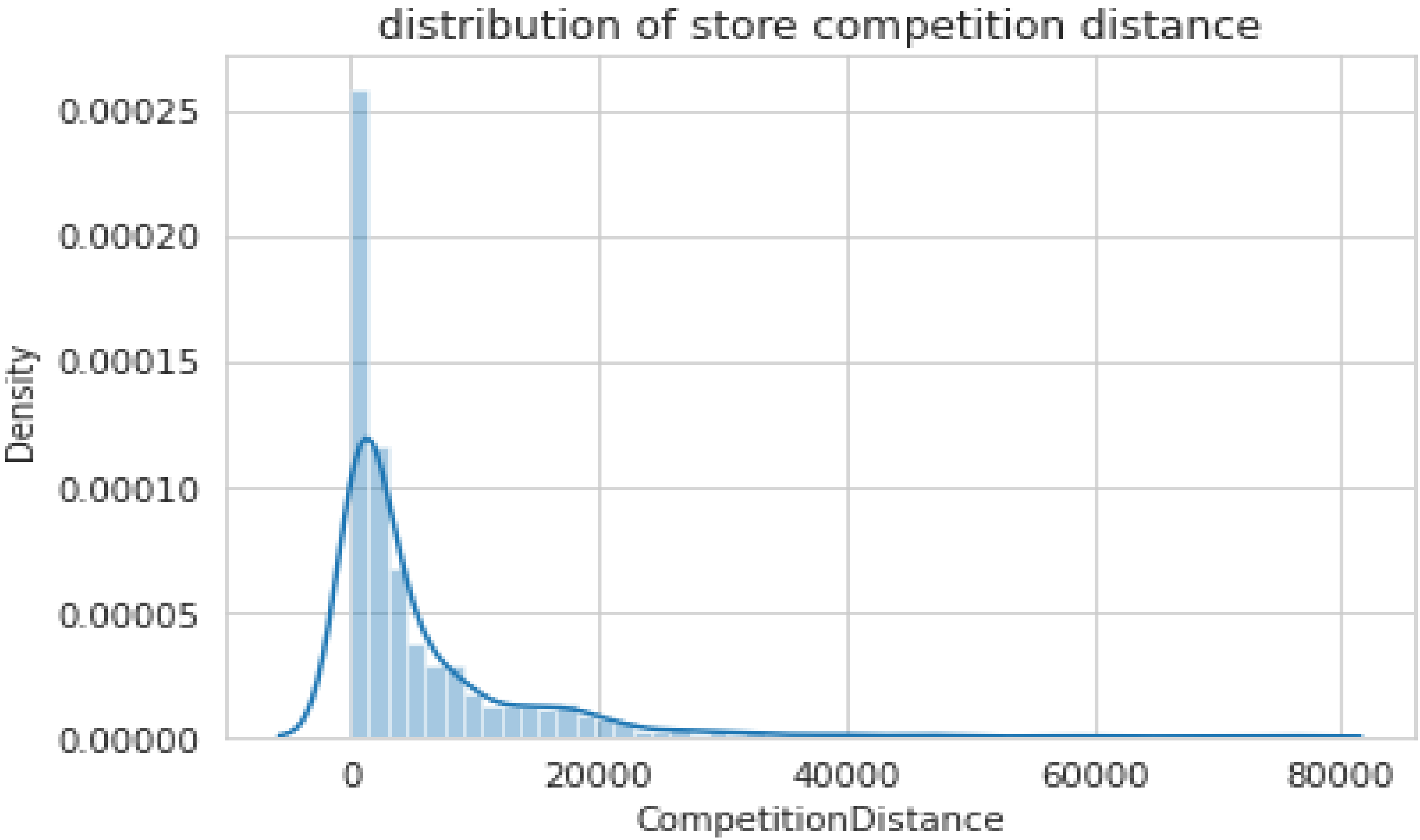
- In the case of assortment of stores most of the shops are came under assortment type a & c
- Assortment type ‘a’ have more number of stores
- ‘b’ type assortment have very less number of stores as compared to ‘a’ and ‘c’





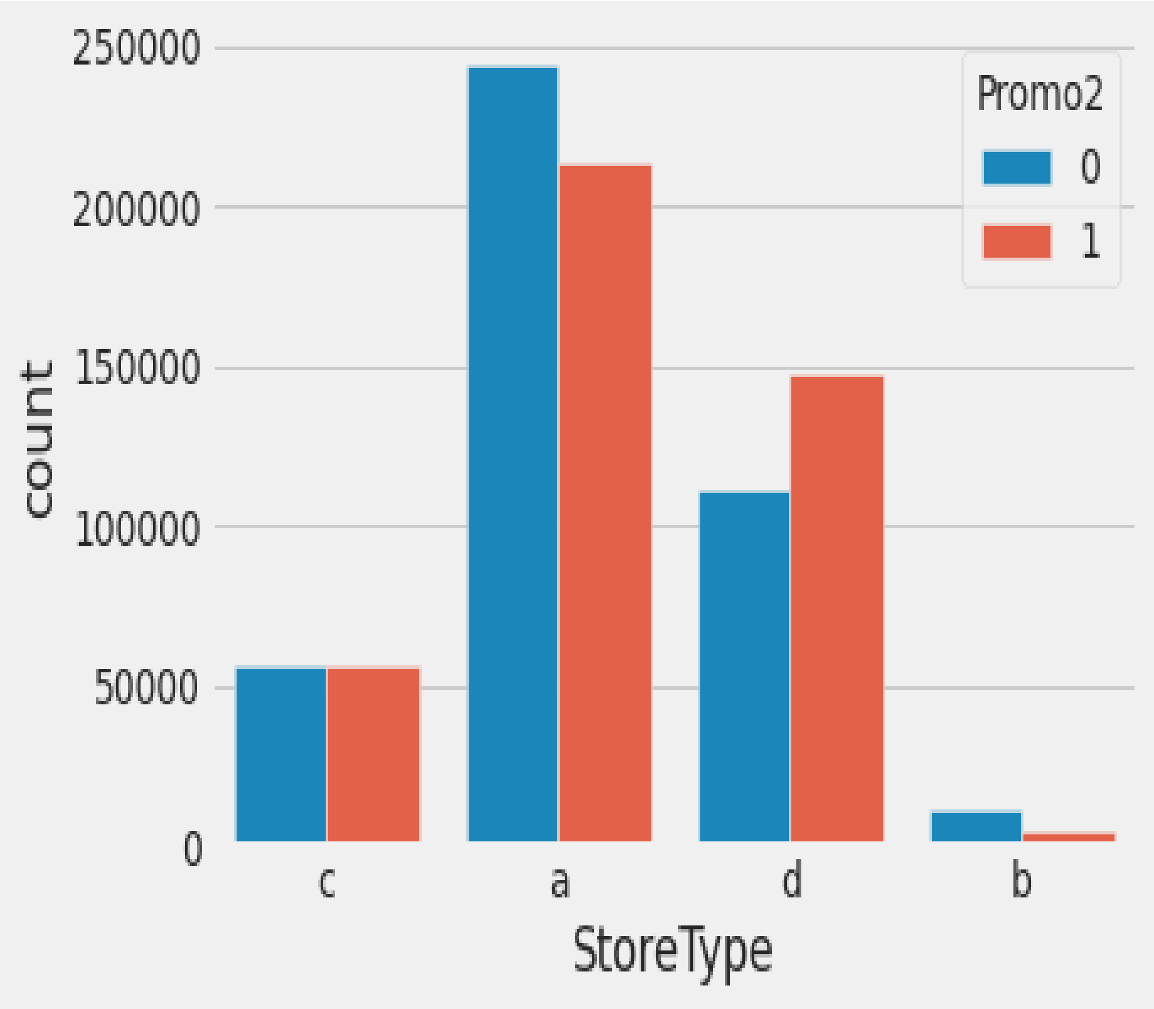
- Number of customers major share is contributed by store type ‘a’ with 54% followed by type ‘d’, ‘c’, ‘b’ as 24 ,14 and 5 percentages respectively

- 54% of stores are in type ‘a’ category 2% of the stores is on ‘b’ category 13% of the stores are in ‘c’ type and 30% of stores are in ‘d’ type



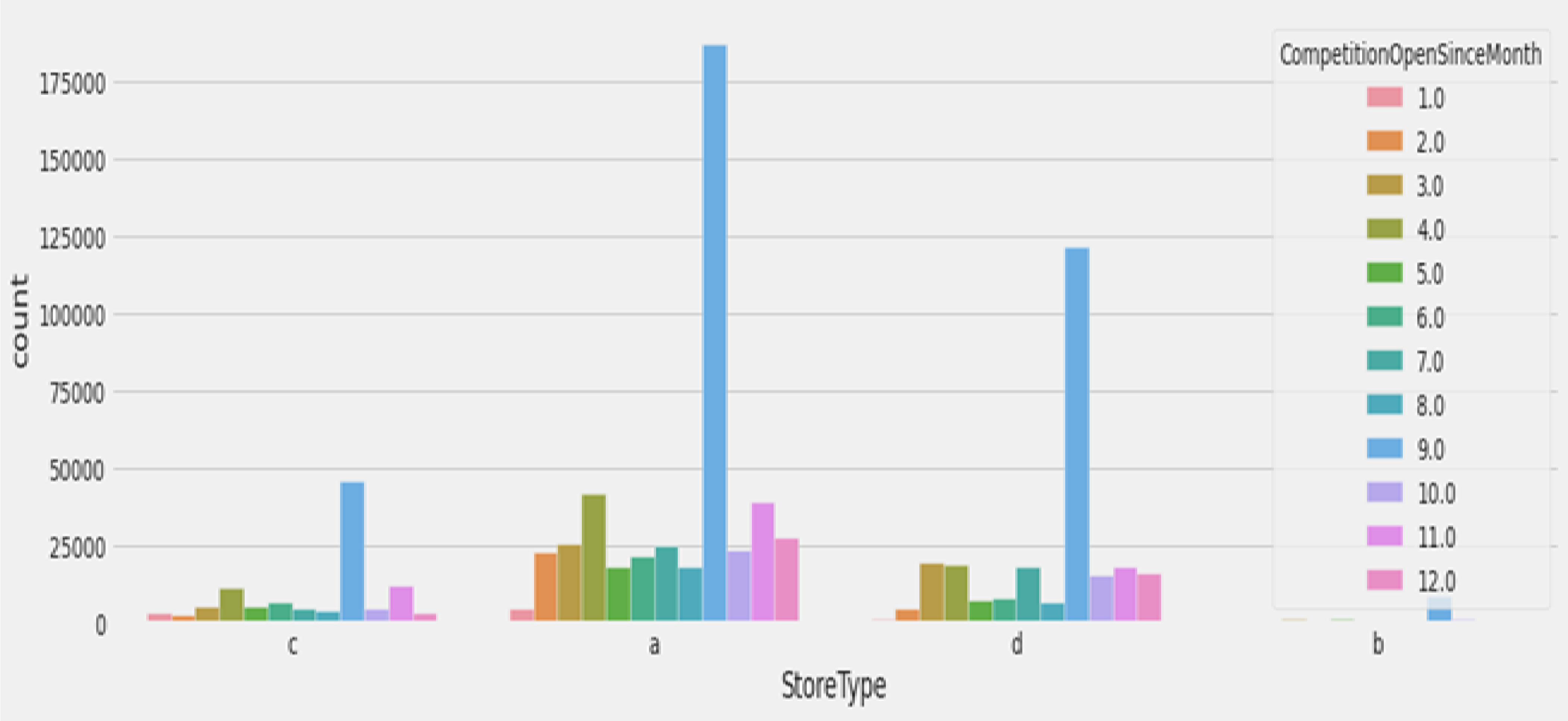
- Distribution of competition distances for drug stores are right skewed
- Which means the competition distance is not much more if there any competition, or there is no competition with much long distance

Promo2 Participation Over Store Type



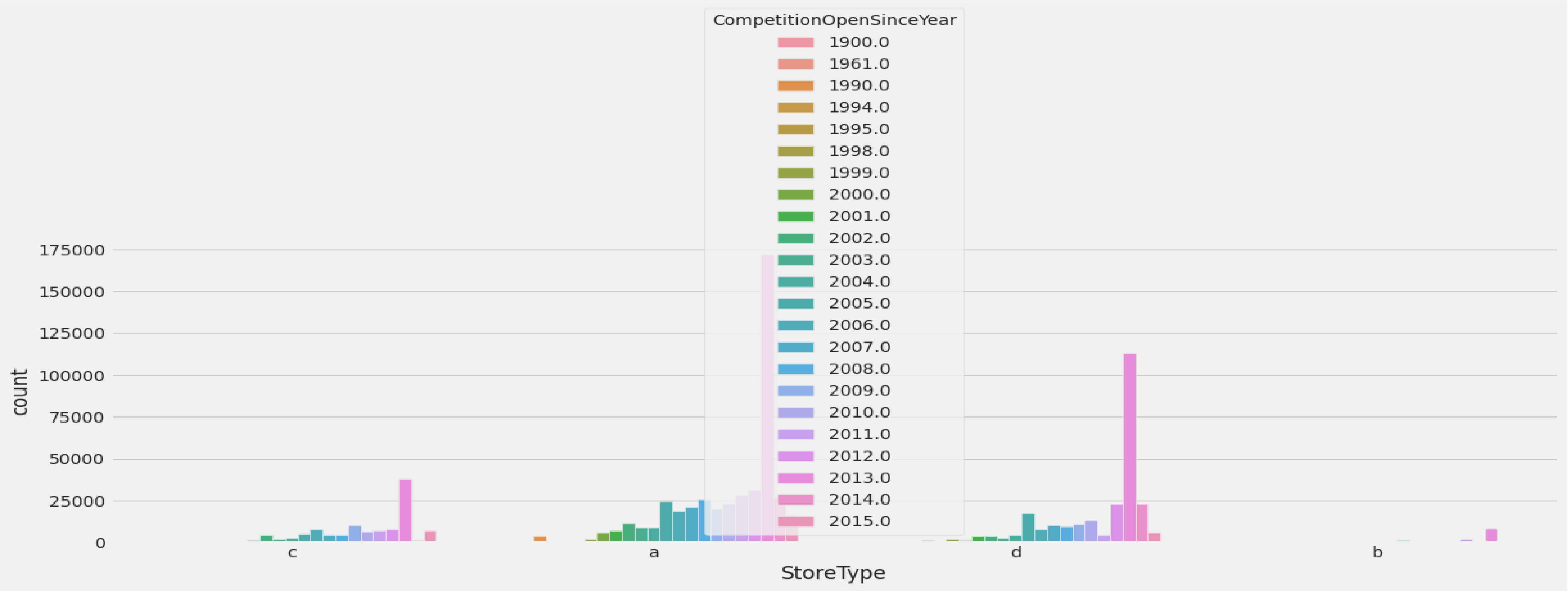
- Store ‘a’ and ‘d’ types are more participating in promotion2 program.
- It is a reflection of shares of store type
- But most of the stores under type ‘a’ are also not participating in promo2 and there is a considerable amount of type ‘d’ stores are also not participating in promo2

Competition open since month over store types



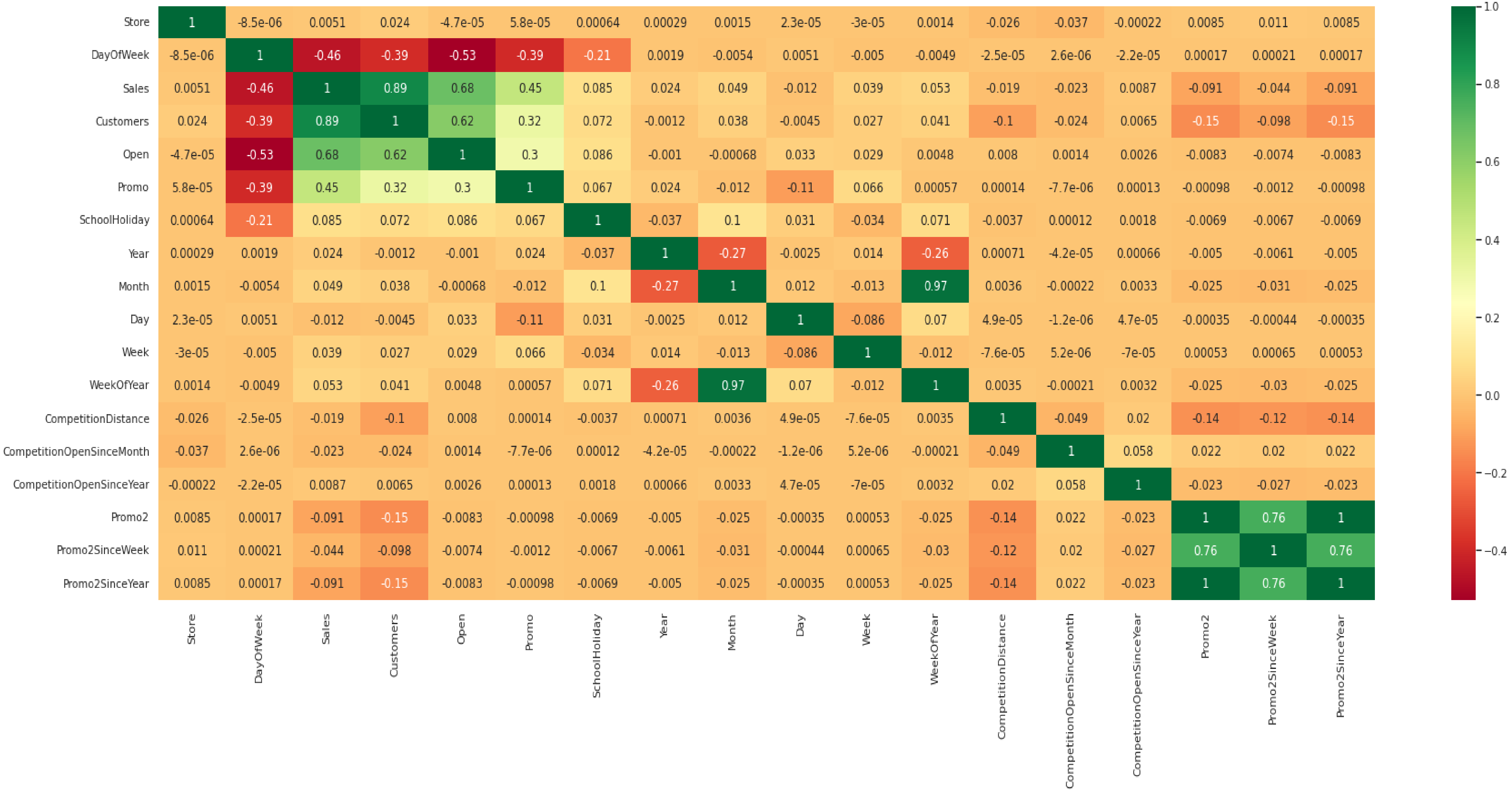
- In case of all store types in November month most of the competitors start to came in action and generally in the year of 2013 it mostly happened

Competition open since year over store type



- Here it is clear that in the year of 2013 most of the competitors are started in field

CORRELATION HEAT MAP



	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	Year	Month	Day	Week	WeekOfYear	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
Store	1	-8.5e-06	0.0051	0.024	-4.7e-05	5.8e-05	0.00064	0.00029	0.0015	2.3e-05	-3e-05	0.0014	-0.026	-0.037	-0.00022	0.0085	0.011	0.0085
DayOfWeek	-8.5e-06	1	-0.46	-0.39	-0.53	-0.39	-0.21	0.0019	-0.0054	0.0051	-0.005	-0.0049	-2.5e-05	2.6e-06	-2.2e-05	0.00017	0.00021	0.00017
Sales	0.0051	-0.46	1	0.89	0.68	0.45	0.085	0.024	0.049	-0.012	0.039	0.053	-0.019	-0.023	0.0087	-0.091	-0.044	-0.091
Customers	0.024	-0.39	0.89	1	0.62	0.32	0.072	-0.0012	0.038	-0.0045	0.027	0.041	-0.1	-0.024	0.0065	-0.15	-0.098	-0.15
Open	-4.7e-05	-0.53	0.68	0.62	1	0.3	0.086	-0.001	-0.00068	0.033	0.029	0.0048	0.008	0.0014	0.0026	-0.0083	-0.0074	-0.0083
Promo	5.8e-05	-0.39	0.45	0.32	0.3	1	0.067	0.024	-0.012	-0.11	0.066	0.00057	0.00014	-7.7e-06	0.00013	-0.00098	-0.0012	-0.00098
SchoolHoliday	0.00064	-0.21	0.085	0.072	0.086	0.067	1	-0.037	0.1	0.031	-0.034	0.071	-0.0037	0.00012	0.0018	-0.0069	-0.0067	-0.0069
Year	0.00029	0.0019	0.024	-0.0012	-0.001	0.024	-0.037	1	-0.27	-0.0025	0.014	-0.26	0.00071	-4.2e-05	0.00066	-0.005	-0.0061	-0.005
Month	0.0015	-0.0054	0.049	0.038	-0.00068	-0.012	0.1	-0.27	1	0.012	-0.013	0.97	0.0036	-0.00022	0.0033	-0.025	-0.031	-0.025
Day	2.3e-05	0.0051	-0.012	-0.0045	0.033	-0.11	0.031	-0.0025	0.012	1	-0.086	0.07	4.9e-05	-1.2e-06	4.7e-05	-0.00035	-0.00044	-0.00035
Week	-3e-05	-0.005	0.039	0.027	0.029	0.066	-0.034	0.014	-0.013	-0.086	1	-0.012	-7.6e-05	5.2e-06	-7e-05	0.00053	0.00065	0.00053
WeekOfYear	0.0014	-0.0049	0.053	0.041	0.0048	0.00057	0.071	-0.26	0.97	0.07	-0.012	1	0.0035	-0.00021	0.0032	-0.025	-0.03	-0.025
CompetitionDistance	-0.026	-2.5e-05	-0.019	-0.1	0.008	0.00014	-0.0037	0.00071	0.0036	4.9e-05	-7.6e-05	0.0035	1	-0.049	0.02	-0.14	-0.12	-0.14
CompetitionOpenSinceMonth	-0.037	2.6e-06	-0.023	-0.024	0.0014	-7.7e-06	0.00012	-4.2e-05	-0.00022	-1.2e-06	5.2e-06	-0.00021	-0.049	1	0.058	0.022	0.02	0.022
CompetitionOpenSinceYear	-0.00022	-2.2e-05	0.0087	0.0065	0.0026	0.00013	0.0018	0.00066	0.0033	4.7e-05	-7e-05	0.0032	0.02	0.058	1	-0.023	-0.027	-0.023
Promo2	0.0085	0.00017	-0.091	-0.15	-0.0083	-0.00098	-0.0069	-0.005	-0.025	-0.00035	0.00053	-0.025	-0.14	0.022	-0.023	1	0.76	1
Promo2SinceWeek	0.011	0.00021	-0.044	-0.098	-0.0074	-0.0012	-0.0067	-0.0061	-0.031	-0.00044	0.00065	-0.03	-0.12	0.02	-0.027	0.76	1	0.76
Promo2SinceYear	0.0085	0.00017	-0.091	-0.15	-0.0083	-0.00098	-0.0069	-0.005	-0.025	-0.00035	0.00053	-0.025	-0.14	0.022	-0.023	1	0.76	1

MODELING

- Linear Regression
- XG Booster
- Decision Tree Regression
- Random Forest Regressor

Evaluation matrix:

Linear Regression

Regression Model Score : 0.8274

Out of Sample Test Score : 0.8278

Training RMSE : 1290.147

Testing RMSE : 1284.715

Training MAPE : 14.258

Testing MAPE : 14.237

R2 score :0.7925

Adjusted R2 : 0.8278

Decision Tree Regression

Regression Model Score : 0.9651

Out of Sample Test Score : 0.9554

Training RMSE : 579.720

Testing RMSE : 653.406

Training MAPE : 5.4758

Testing MAPE : 6.1870

R2 score **:** 0.9538

Adjusted R2 : 0.9554

XG Booster

Regression Model Score : 0.89090

Out of Sample Test Score : 0.89098

Training RMSE : 1025.772

Testing RMSE : 1022.387

Training MAPE : 11.533

Testing MAPE : 11.506

R2 score : 0.86765

Adjusted R2 : 0.8909

Random Forest Regressor

Regression Model Score : 0.9950

Out of Sample Test Score : 0.9735

Training RMSE : 218.87

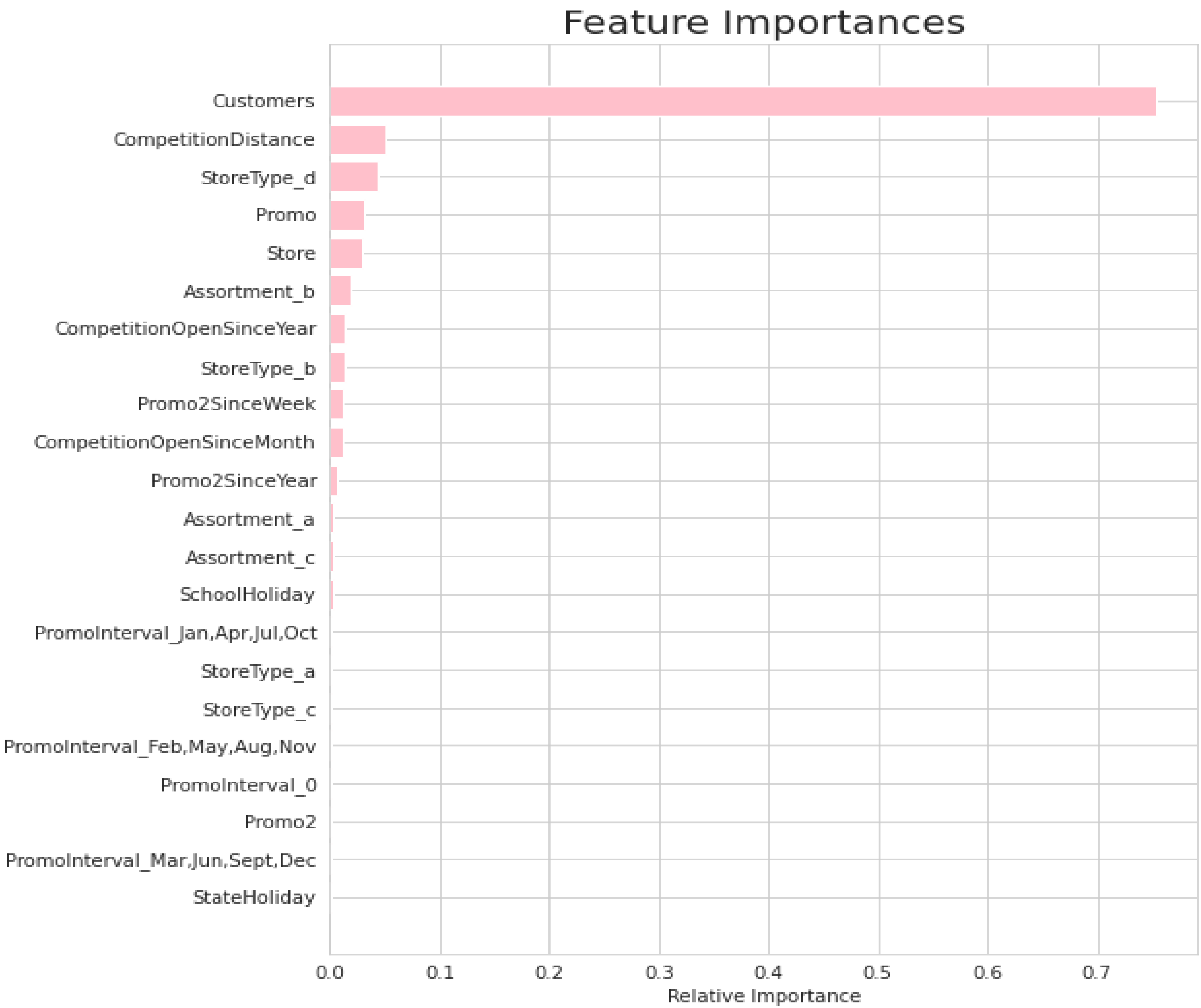
Testing RMSE : 503.49

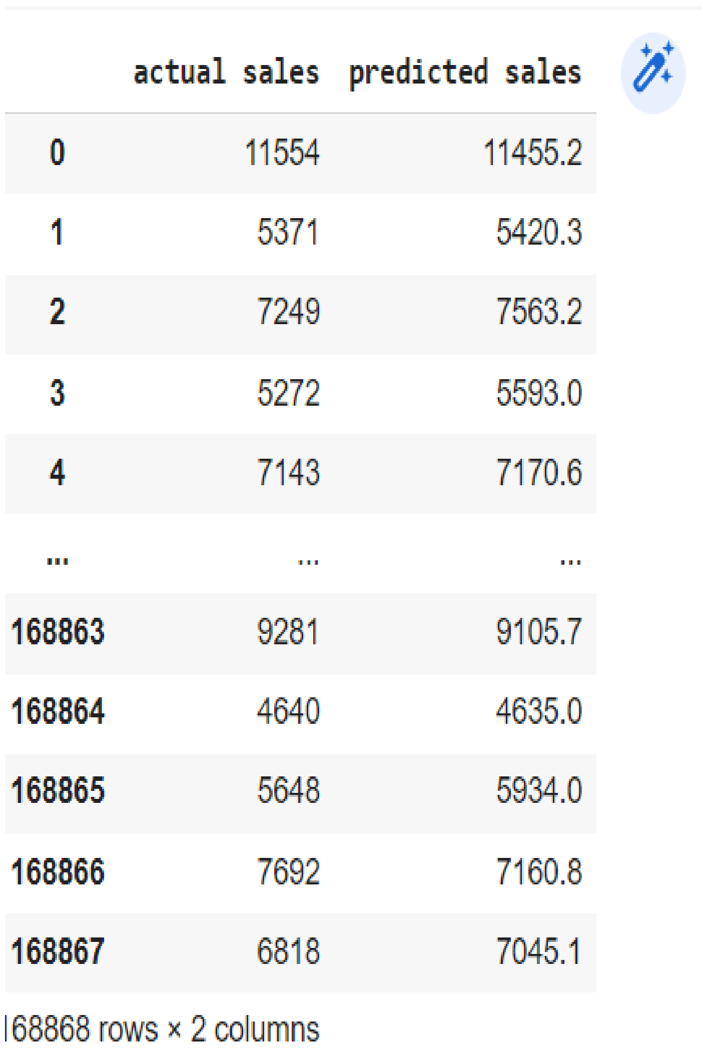
Training MAPE : 2.009

Testing MAPE : 4.972

R2 score : 0.9726

Adjusted R2 : 0.9735





	actual sales	predicted sales
0	11554	11455.2
1	5371	5420.3
2	7249	7563.2
3	5272	5593.0
4	7143	7170.6
...	...	...
168863	9281	9105.7
168864	4640	4635.0
168865	5648	5934.0
168866	7692	7160.8
168867	6818	7045.1

168868 rows × 2 columns

CONCLUSION

Conclusions of model:

By this analysis Random Forest Tuned Model gave the best results and ,which indicates that all the trends and patterns that could be captured by these models without overfitting were done. It built a good model to predict sales with a 97% test regression score. It is a good level of accuracy and also it showing a good range of prediction without much variations

Conclusions of EDA:

- The most selling and crowded store type is A.
- Sales is highly correlated to the number of Customers.
- Absence of values in features CompetitionOpenSinceYear/Month doesn’t indicate the absence of competition as CompetitionDistance values are not null where the other two values are null.
- August, September and October are the months with low monthly sales. After October in every year there is a sharp hike in monthly sales
- Store types ‘a’ and ‘d’ are more participating in promotion2 program
- In a month time frame more sales in the starting days and ending days of month
- Mainly customers and promo are the 2 correlated features with sales as per heatmap matrix
- There are 4 different type of stores among which 54% stores are of type – a which is maximum, and the least is type – b
- we can say maximum sales by store type “b”(by mean sale valuation), but also the number of store with type “b” is minimum so we should consider type “a”

THANK YOU