Capstone Project

Retail Sales Prediction
Rossman Stores

by

Lovejeet Singh

Contents

1. Rossmann Store
2. Problem Statement
3. Data Summary
4. Exploratory Data Analysis
5. Feature Engineering
6. Feature Selection
7. Model Implementation
8. Challenges Faced
9. Conclusion

Rossmann Store

- Rossmann Store is one of the largest drug store chains in Europe with around 56,200 employees and more than 4000 stores. The product range includes up to 21,700 items and can vary depending on the size of the shop and the location.
- In addition to drugstore goods with a focus on skin, hair, body, baby and health, Rossmann also offers promotional items ("World of Ideas"), pet food, a photo service and a wide range of natural foods and wines.

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.

- The objective of the project is to come with a optimal machine learning model to predict sales.

Data Summary

The Dataset provided by the Firm are as follow :-

- **Rossmann Stores Data.csv –** Historical data including Sales
- **store.csv -** Supplemental information about the stores.

Details of Dataset

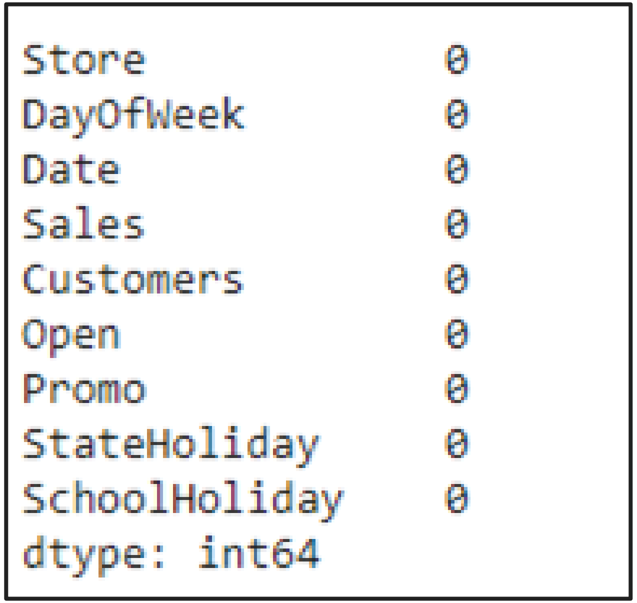
- Rossmann stores Data.csv has 9 feature and 1017209 observations.
- Store.csv has 10 feature and 1115 observations

Main Features

1. **Sales -** The turnover for any given day (this is what we are predicting).
2. **Open -** An indicator for whether the store was open or closed. 0 = closed, 1 = open.
3. **Store type -** Differentiates between 4 different store models (a, b, c & d).
4. **Assortment -** Describes an assortment level: a = basic, b = extra, c = extended.
5. **Promo -** Indicates whether a store is running a promo on that day
6. **Promo2 -** Promo2 is a continuing and consecutive promotion for some stores.
7. **Store -** A unique Id for each store.
8. **Customer -** The number of customers on a given day.
9. **Competition Distance -** Distance in meters to the nearest competitor store.
10. **Promo Interval -** Describes the consecutive intervals Promo2 is started, naming the months the promotion is started a new.
11. **Promo2Since [Year/week] -** Describes the year and calendar week when the store started participating in Promo2.

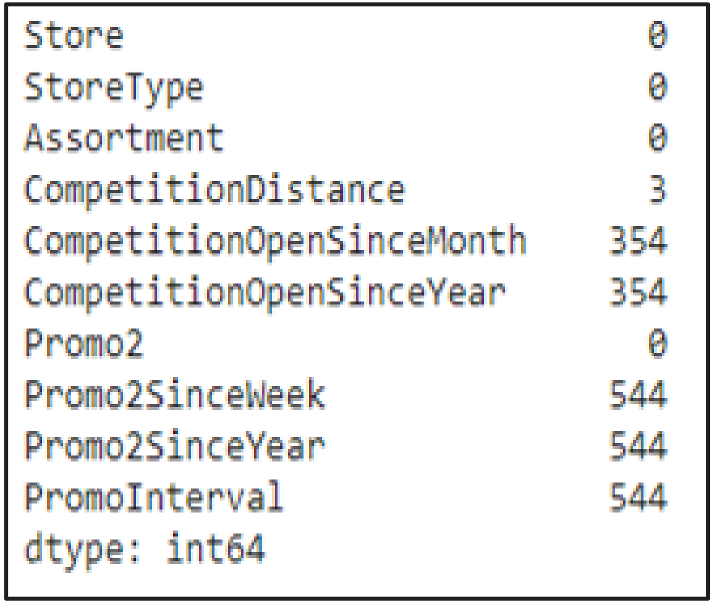
Null Values

Rossmann.csv has zero null values in



```
Store            0
DayOfWeek        0
Date             0
Sales            0
Customers        0
Open             0
Promo            0
StateHoliday     0
SchoolHoliday    0
dtype: int64
```

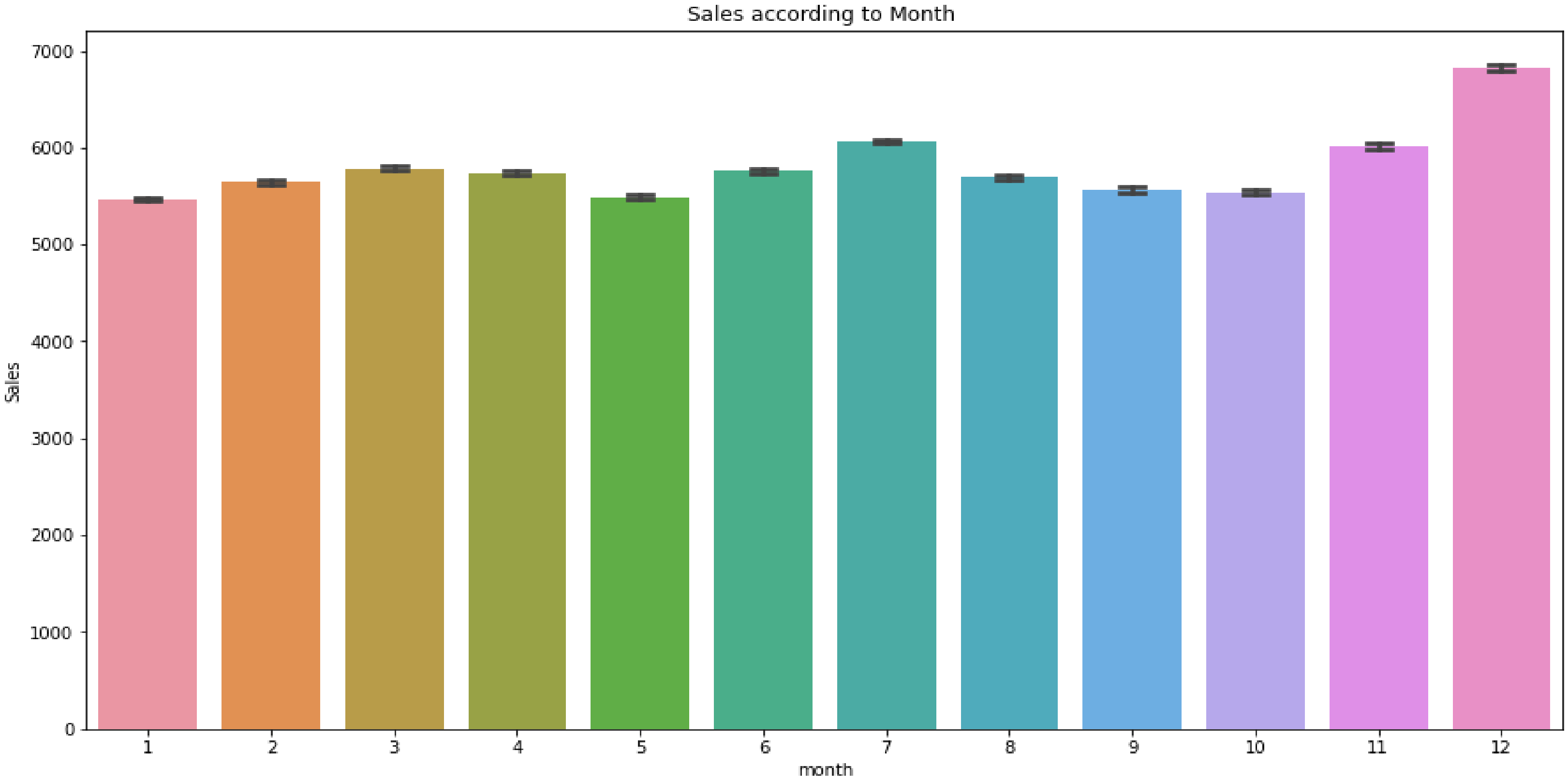
Store.csv has lots of null values



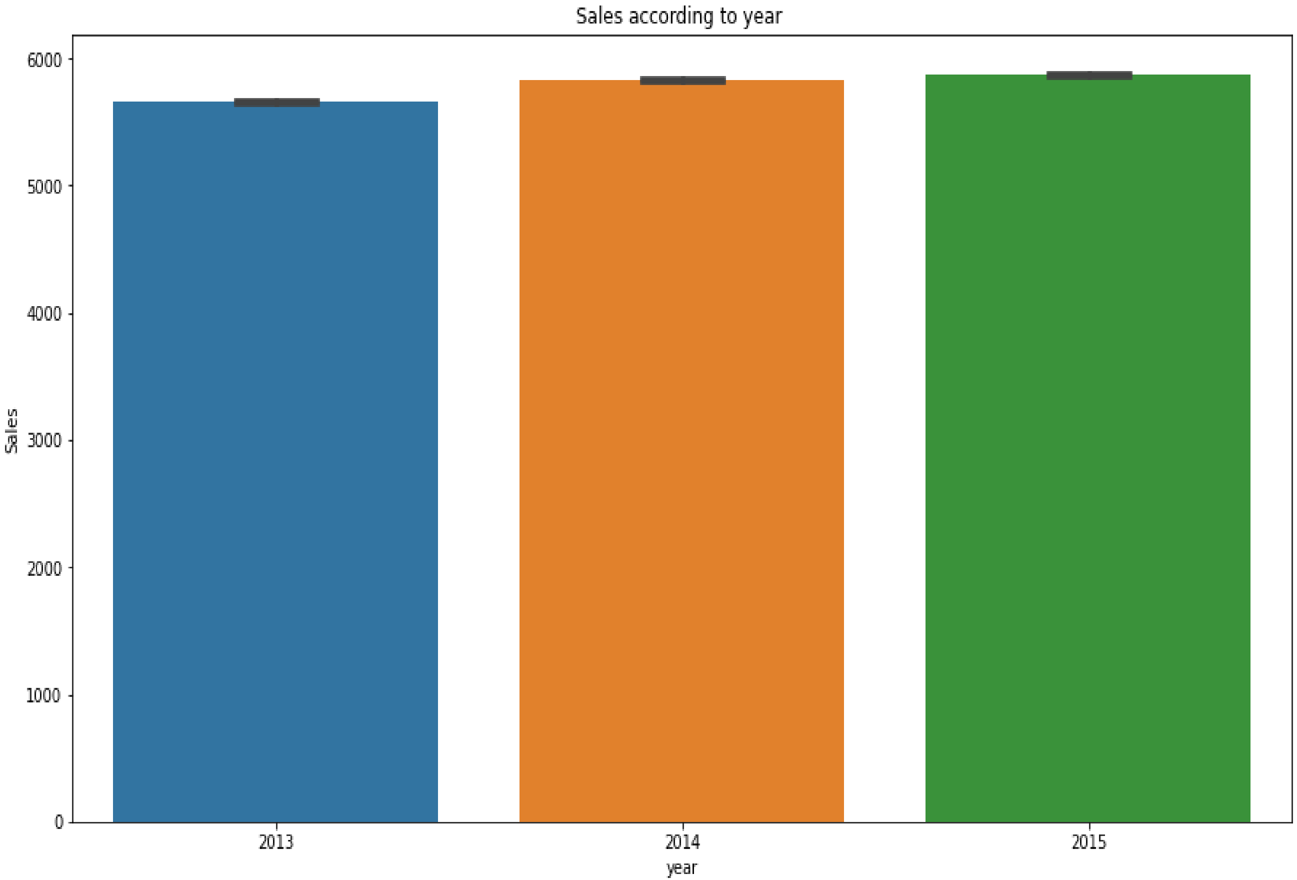
```
Store                          0
StoreType                      0
Assortment                     0
CompetitionDistance            3
CompetitionOpenSinceMonth    354
CompetitionOpenSinceYear     354
Promo2                         0
Promo2SinceWeek              544
Promo2SinceYear              544
PromoInterval                544
dtype: int64
```

Exploratory Data Analysis

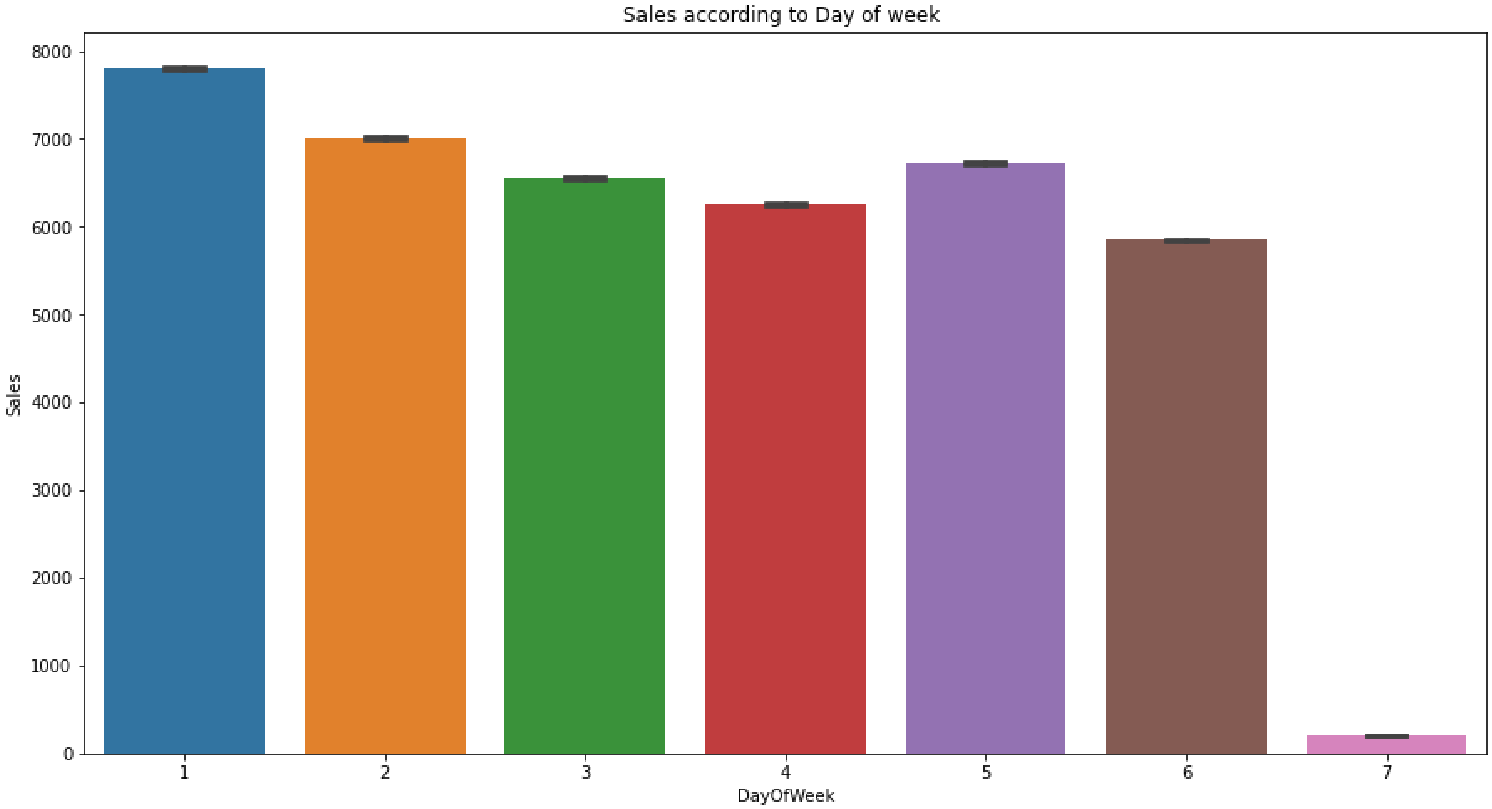
Sales According To Month



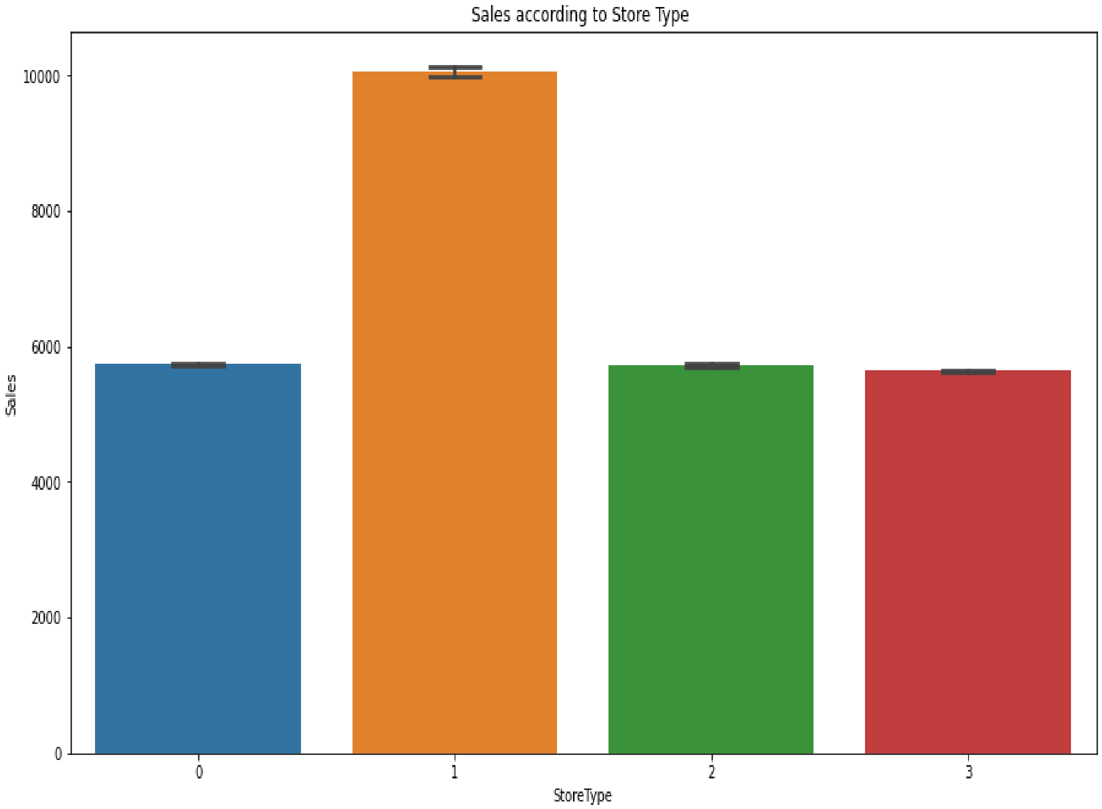
Sales According To Year



Sales According To Day Of Week

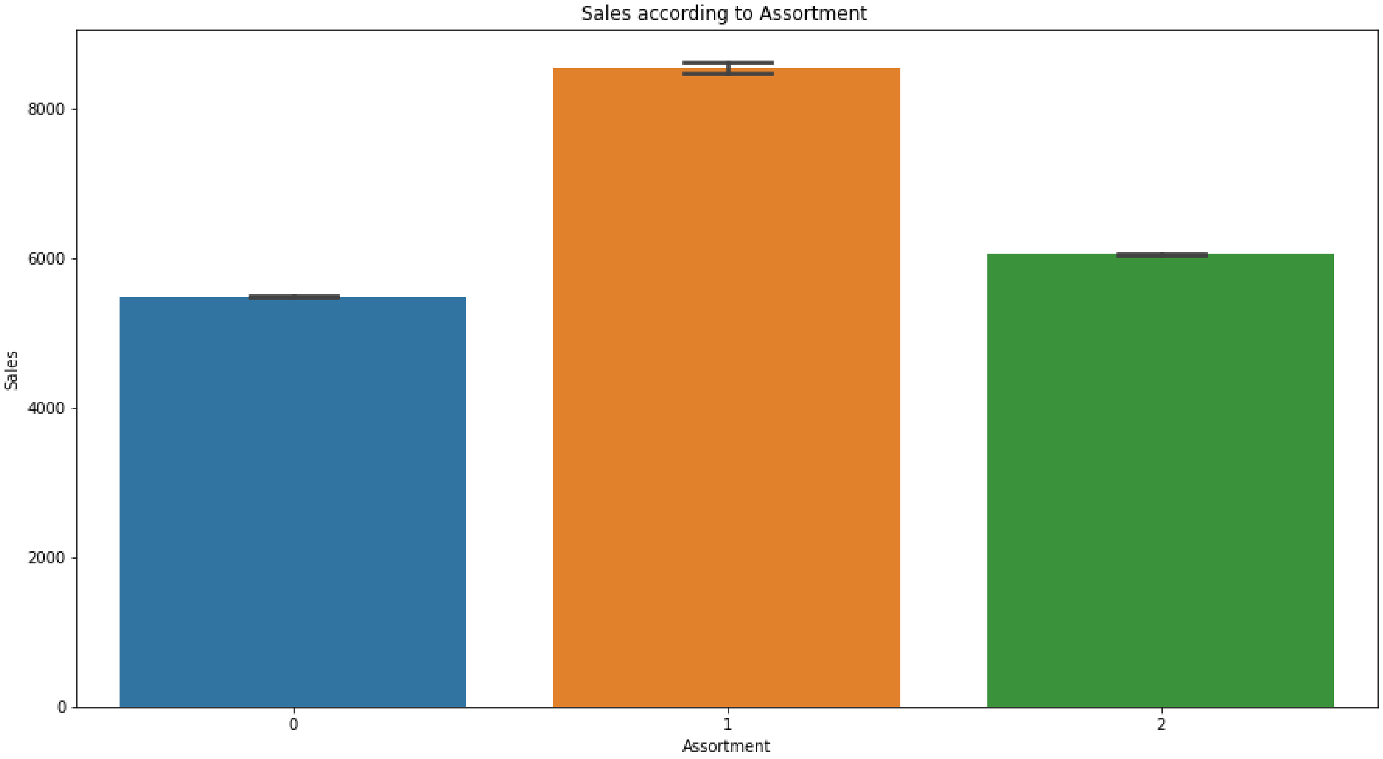


Sales Vs Store Type

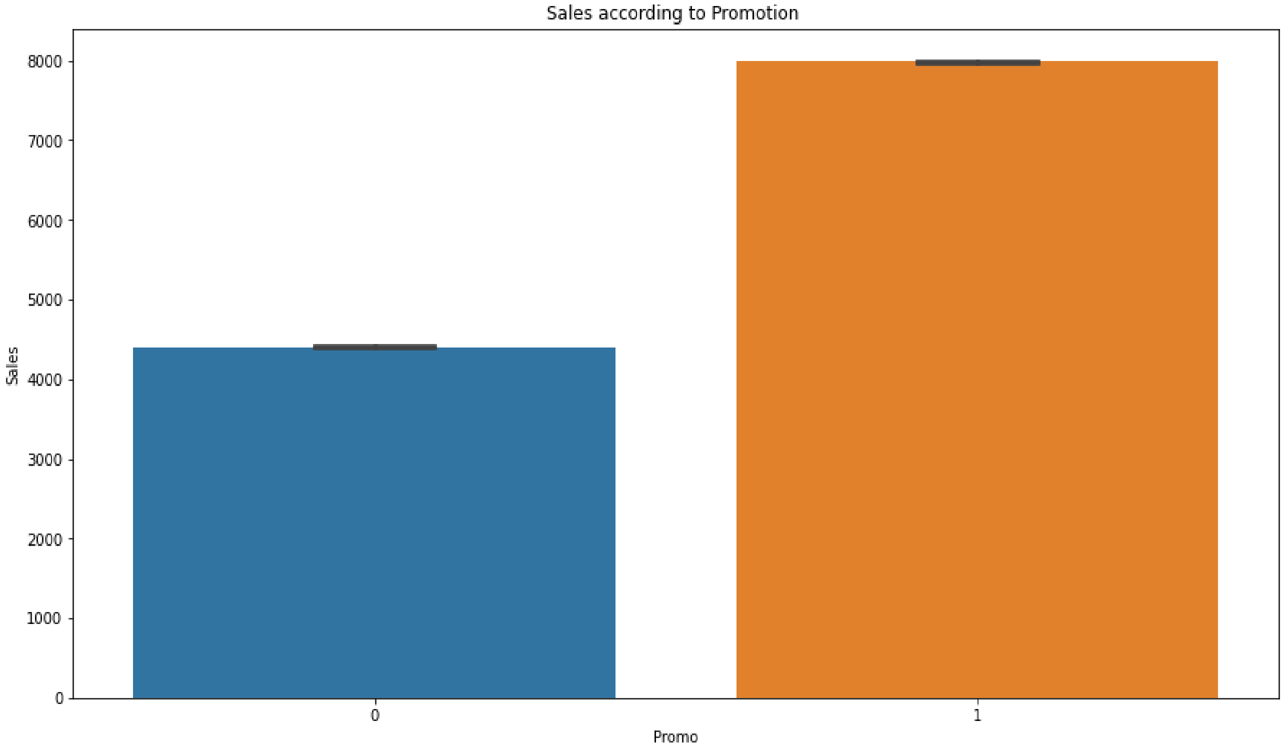


Store Type 0 = Small Size Store
Store Type 1 = Medium Size Store
Store Type 2 = Large Size Store
Store Type 3 = Huge Size Store

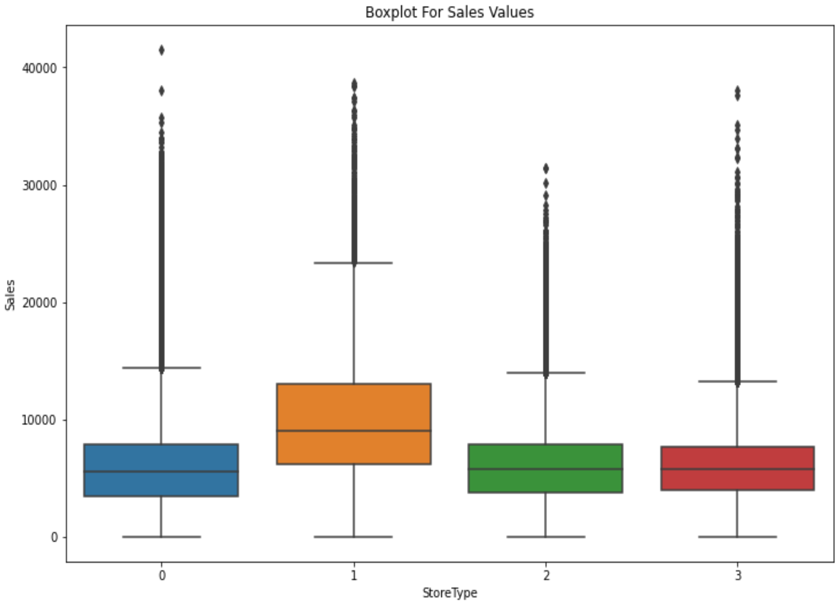
Sales According To Assortment

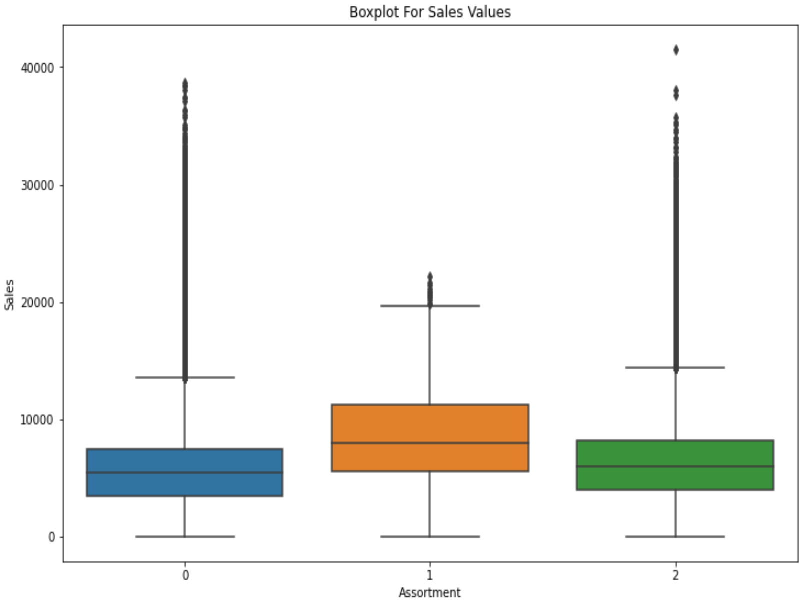


Sale According To Promotion



Box Plot of Sales at different Stores & Assortment



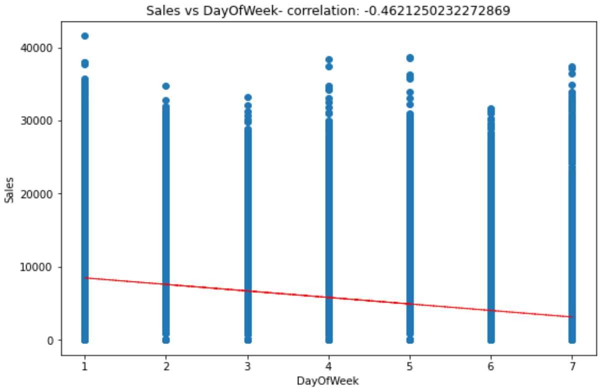


Feature Engineering

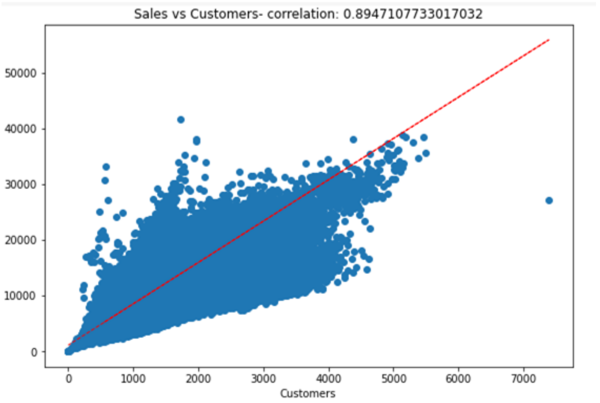
Null Values Treatment

- In Store Dataset there are many Nan values.
- In Competition Distance there are only 3 Nan value, so by Checking its distribution which look like right skewed so we decided to replace with it median.
- For rest other features there are lots of missing values and nothing much information giving about them.
- Some of Feature has more than 40% null vales to we simply decided to drop these feature.

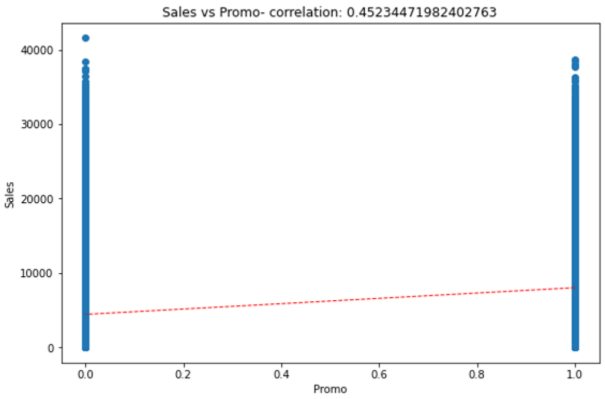
Correlation between Independent Features With Target Variable Sales



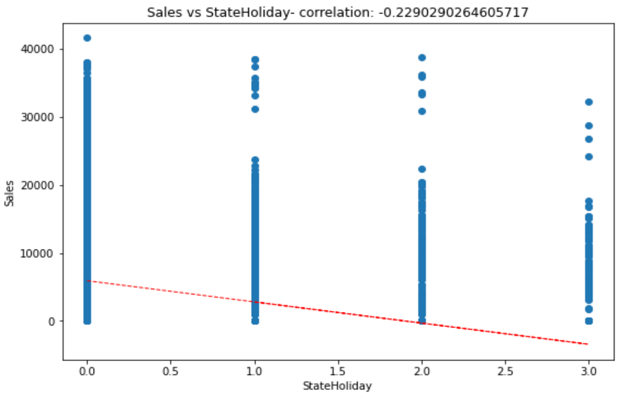
Sales vs DayOfWeek- correlation: -0.4621250232272869



Sales vs Customers- correlation: 0.8947107733017032

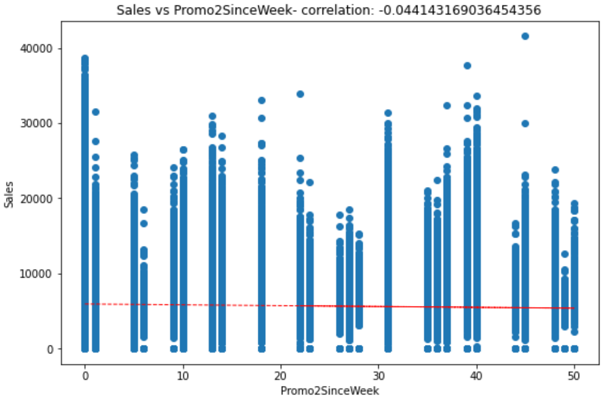


Sales vs Promo- correlation: 0.45234471982402763

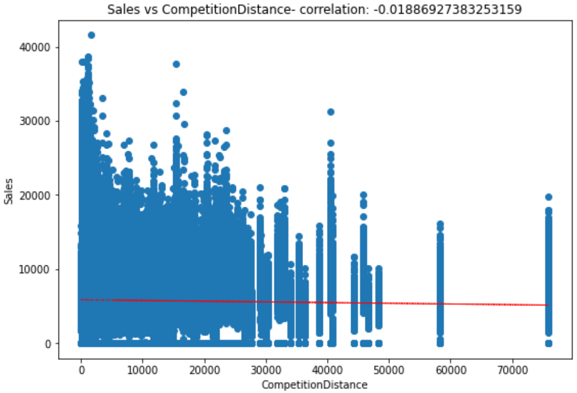


Sales vs StateHoliday- correlation: -0.2290290264605717

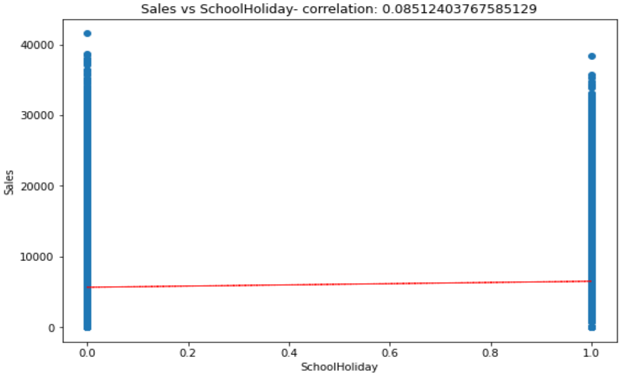
Contd..



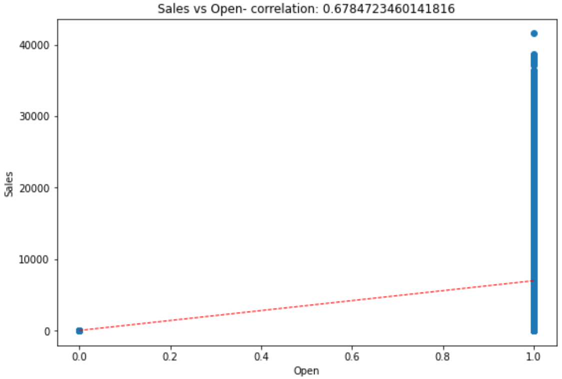
Sales vs Promo2SinceWeek- correlation: -0.044143169036454356



Sales vs CompetitionDistance- correlation: -0.01886927383253159

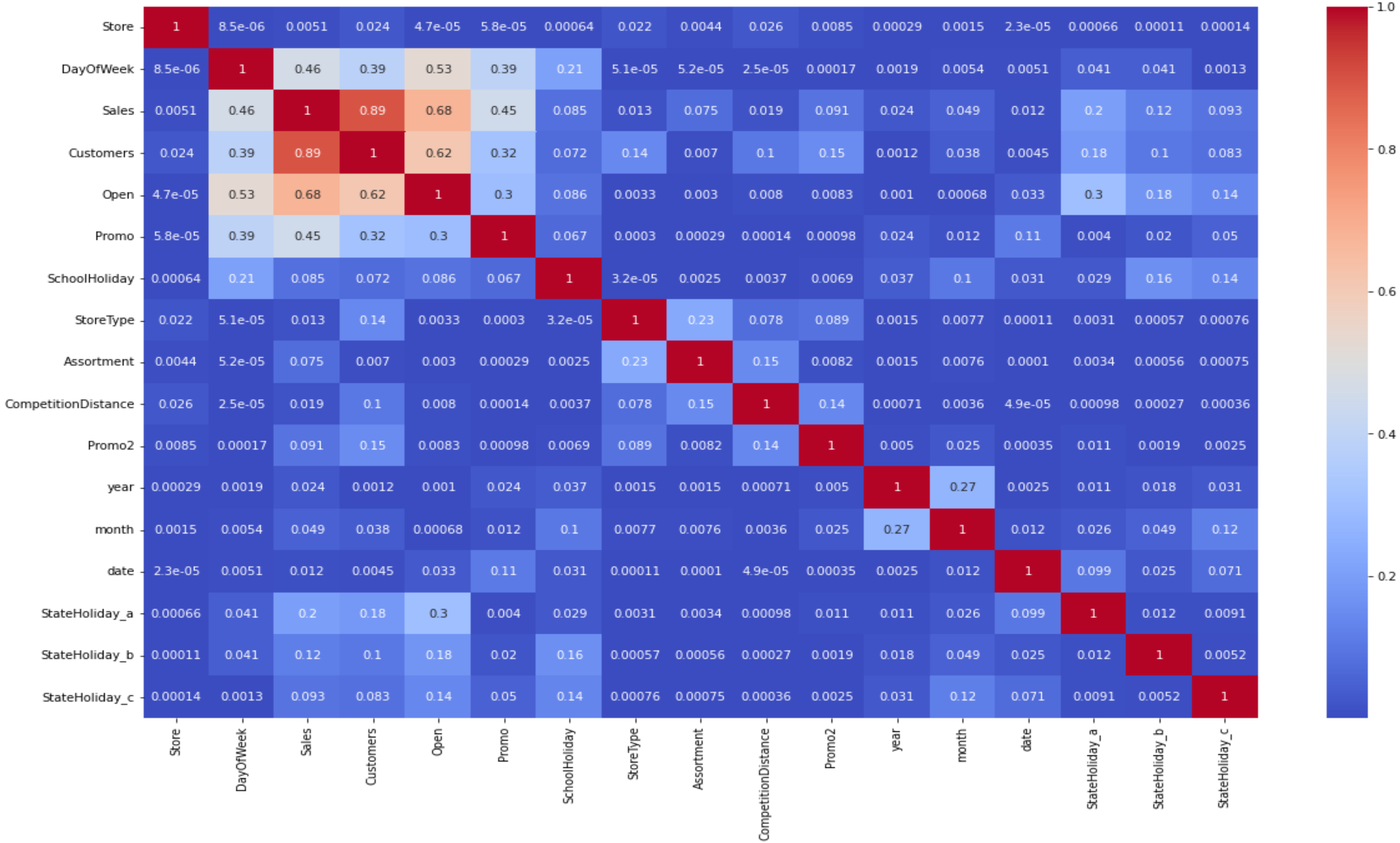


Sales vs SchoolHoliday- correlation: 0.08512403767585129



Sales vs Open- correlation: 0.6784723460141816

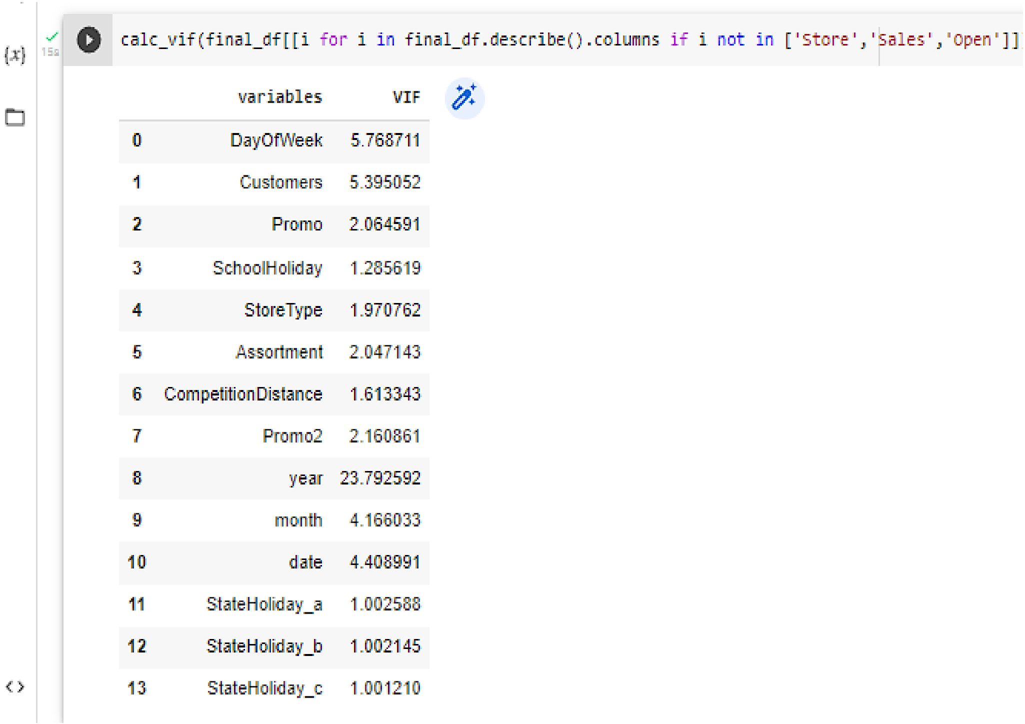
Correlation Heatmap



	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Promo2	year	month	date	StateHoliday_a	StateHoliday_b	StateHoliday_c
Store	1	8.5e-06	0.0051	0.024	4.7e-05	5.8e-05	0.00064	0.022	0.0044	0.026	0.0085	0.00029	0.0015	2.3e-05	0.00066	0.00011	0.00014
DayOfWeek	8.5e-06	1	0.46	0.39	0.53	0.39	0.21	5.1e-05	5.2e-05	2.5e-05	0.00017	0.0019	0.0054	0.0051	0.041	0.041	0.0013
Sales	0.0051	0.46	1	0.89	0.68	0.45	0.085	0.013	0.075	0.019	0.091	0.024	0.049	0.012	0.2	0.12	0.093
Customers	0.024	0.39	0.89	1	0.62	0.32	0.072	0.14	0.007	0.1	0.15	0.0012	0.038	0.0045	0.18	0.1	0.083
Open	4.7e-05	0.53	0.68	0.62	1	0.3	0.086	0.0033	0.003	0.008	0.0083	0.001	0.00068	0.033	0.3	0.18	0.14
Promo	5.8e-05	0.39	0.45	0.32	0.3	1	0.067	0.0003	0.00029	0.00014	0.00098	0.024	0.012	0.11	0.004	0.02	0.05
SchoolHoliday	0.00064	0.21	0.085	0.072	0.086	0.067	1	3.2e-05	0.0025	0.0037	0.0069	0.037	0.1	0.031	0.029	0.16	0.14
StoreType	0.022	5.1e-05	0.013	0.14	0.0033	0.0003	3.2e-05	1	0.23	0.078	0.089	0.0015	0.0077	0.00011	0.0031	0.00057	0.00076
Assortment	0.0044	5.2e-05	0.075	0.007	0.003	0.00029	0.0025	0.23	1	0.15	0.0082	0.0015	0.0076	0.0001	0.0034	0.00056	0.00075
CompetitionDistance	0.026	2.5e-05	0.019	0.1	0.008	0.00014	0.0037	0.078	0.15	1	0.14	0.00071	0.0036	4.9e-05	0.00098	0.00027	0.00036
Promo2	0.0085	0.00017	0.091	0.15	0.0083	0.00098	0.0069	0.089	0.0082	0.14	1	0.005	0.025	0.00035	0.011	0.0019	0.0025
year	0.00029	0.0019	0.024	0.0012	0.001	0.024	0.037	0.0015	0.0015	0.00071	0.005	1	0.27	0.0025	0.011	0.018	0.031
month	0.0015	0.0054	0.049	0.038	0.00068	0.012	0.1	0.0077	0.0076	0.0036	0.025	0.27	1	0.012	0.026	0.049	0.12
date	2.3e-05	0.0051	0.012	0.0045	0.033	0.11	0.031	0.00011	0.0001	4.9e-05	0.00035	0.0025	0.012	1	0.099	0.025	0.071
StateHoliday_a	0.00066	0.041	0.2	0.18	0.3	0.004	0.029	0.0031	0.0034	0.00098	0.011	0.011	0.026	0.099	1	0.012	0.0091
StateHoliday_b	0.00011	0.041	0.12	0.1	0.18	0.02	0.16	0.00057	0.00056	0.00027	0.0019	0.018	0.049	0.025	0.012	1	0.0052
StateHoliday_c	0.00014	0.0013	0.093	0.083	0.14	0.05	0.14	0.00076	0.00075	0.00036	0.0025	0.031	0.12	0.071	0.0091	0.0052	1

Feature Selection

Multicollinearity

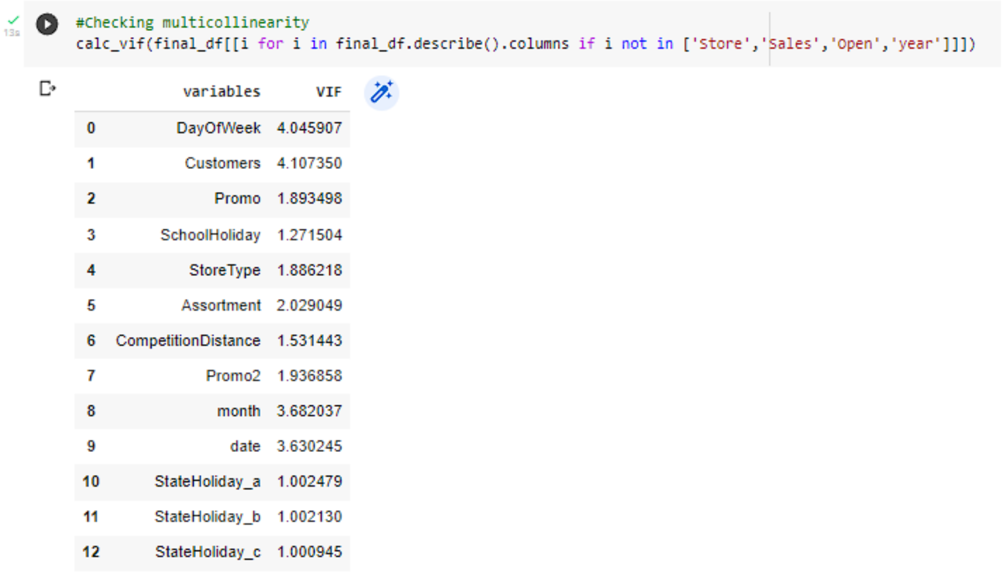


```
calc_vif(final_df[[i for i in final_df.describe().columns if i not in ['Store','Sales','Open']]]
```

	variables	VIF
0	DayOfWeek	5.768711
1	Customers	5.395052
2	Promo	2.064591
3	SchoolHoliday	1.285619
4	StoreType	1.970762
5	Assortment	2.047143
6	CompetitionDistance	1.613343
7	Promo2	2.160861
8	year	23.792592
9	month	4.166033
10	date	4.408991
11	StateHoliday_a	1.002588
12	StateHoliday_b	1.002145
13	StateHoliday_c	1.001210

- Collinearity of year is high. So we have to drop that column only for liner regression algorithm.
- Rest of all algorithm like decision tree, xgboost we are going to use all features.

Contd...



```
#Checking multicollinearity
calc_vif(final_df[[i for i in final_df.describe().columns if i not in ['Store','Sales','Open','year']]])
```

	variables	VIF
0	DayOfWeek	4.045907
1	Customers	4.107350
2	Promo	1.893498
3	SchoolHoliday	1.271504
4	StoreType	1.886218
5	Assortment	2.029049
6	CompetitionDistance	1.531443
7	Promo2	1.936858
8	month	3.682037
9	date	3.630245
10	StateHoliday_a	1.002479
11	StateHoliday_b	1.002130
12	StateHoliday_c	1.000945

- After removing year, Now these are our final feature for Linear regression algorithm.

Model Implementation

Algorithm used

Following are the regression algorithm used.

1. Linear Regression
2. Linear Regression with Regularization (L1,L2 and Elastic Net).
3. Decision Tree
4. Random Forest
5. XGBoost
6. Gradient Boosting Regressor.
7. Gradient Boosting Regressor With GridSearchCV

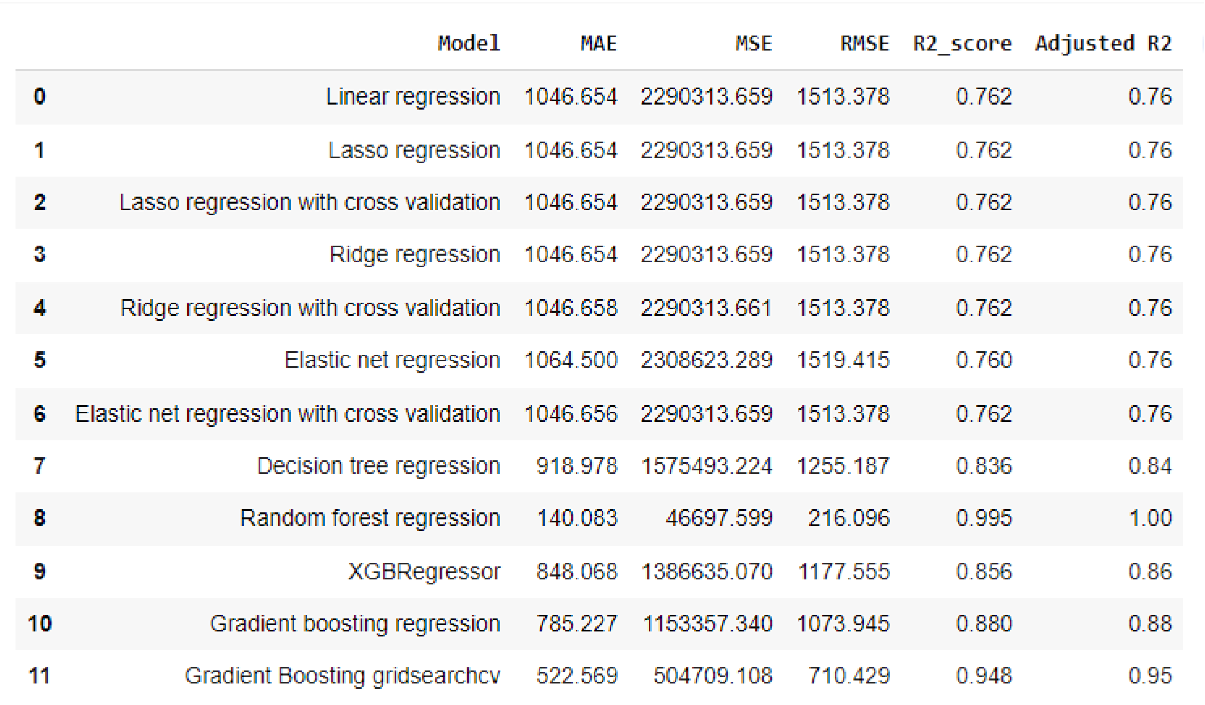
Evaluation Metrics For Regression

Following are the evaluation metrics for regression.

1. Mean Absolute Error(MAE)
2. Mean Squared Error(MSE)
3. Root Mean Squared Error(RMSE)
4. R Squared (R2)
5. Adjusted R Squared

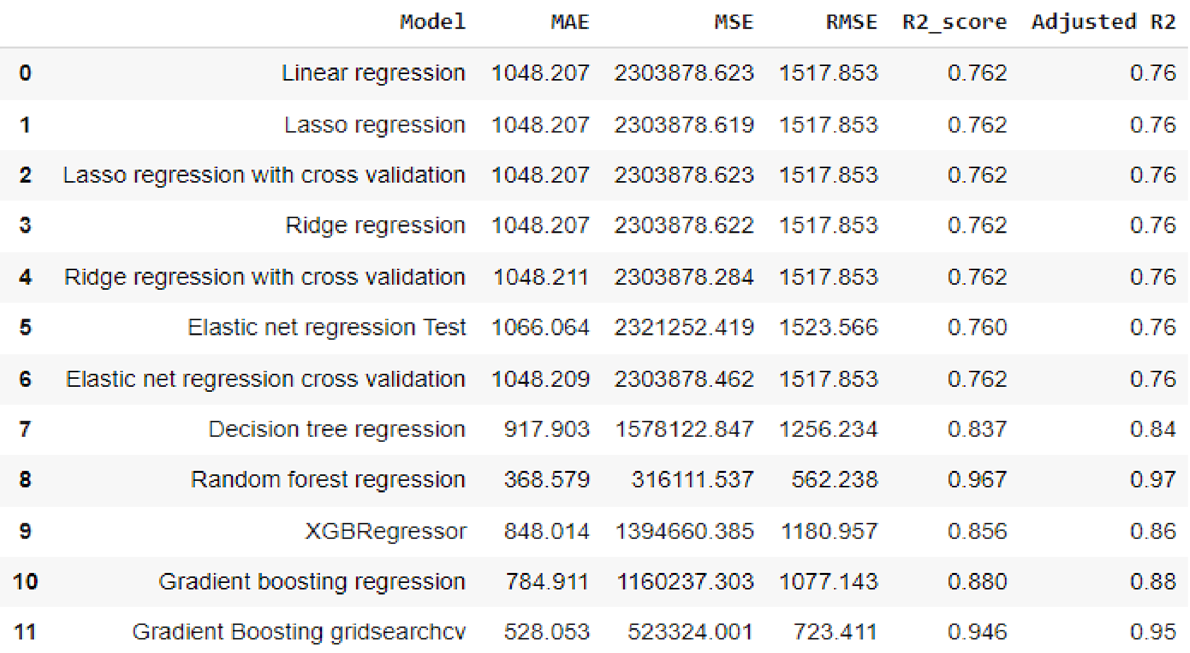
- On the basis of R Squared we evaluate our model performance on both train or test set.

Evaluation Metrics On Train Set



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	1046.654	2290313.659	1513.378	0.762	0.76
1	Lasso regression	1046.654	2290313.659	1513.378	0.762	0.76
2	Lasso regression with cross validation	1046.654	2290313.659	1513.378	0.762	0.76
3	Ridge regression	1046.654	2290313.659	1513.378	0.762	0.76
4	Ridge regression with cross validation	1046.658	2290313.661	1513.378	0.762	0.76
5	Elastic net regression	1064.500	2308623.289	1519.415	0.760	0.76
6	Elastic net regression with cross validation	1046.656	2290313.659	1513.378	0.762	0.76
7	Decision tree regression	918.978	1575493.224	1255.187	0.836	0.84
8	Random forest regression	140.083	46697.599	216.096	0.995	1.00
9	XGBRegressor	848.068	1386635.070	1177.555	0.856	0.86
10	Gradient boosting regression	785.227	1153357.340	1073.945	0.880	0.88
11	Gradient Boosting gridsearchcv	522.569	504709.108	710.429	0.948	0.95

Evaluation Metrics On Test Set



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	1048.207	2303878.623	1517.853	0.762	0.76
1	Lasso regression	1048.207	2303878.619	1517.853	0.762	0.76
2	Lasso regression with cross validation	1048.207	2303878.623	1517.853	0.762	0.76
3	Ridge regression	1048.207	2303878.622	1517.853	0.762	0.76
4	Ridge regression with cross validation	1048.211	2303878.284	1517.853	0.762	0.76
5	Elastic net regression Test	1066.064	2321252.419	1523.566	0.760	0.76
6	Elastic net regression cross validation	1048.209	2303878.462	1517.853	0.762	0.76
7	Decision tree regression	917.903	1578122.847	1256.234	0.837	0.84
8	Random forest regression	368.579	316111.537	562.238	0.967	0.97
9	XGBRegressor	848.014	1394660.385	1180.957	0.856	0.86
10	Gradient boosting regression	784.911	1160237.303	1077.143	0.880	0.88
11	Gradient Boosting gridsearchcv	528.053	523324.001	723.411	0.946	0.95

Model Selection

- By looking evaluation metric on both train and test set. We decided to go with Random Forest Regressor Model in which we got 0.99 R Squared score on train and 0.967 R Squared score on test set.
- The Gradient Boosting Regressor with GridSearchCV model also perform very well and we got 0.94 R Squared Score on both train and test set.
- In all the algorithm there was no overfitting seen.

Challenges

- Handling large amount of sales data (10,17,210 observations on 13 variables)
- Some stores were closed. So we have to drop those observation in which store was closed and sales also 0.
- Due to very large dataset when we fit our model by using optimal algorithm search tool like Gridsearchcv. We face too much system failure.
- When fitting the model using GridSearchCv many times google colab crashes due to exceeding the RAM uses.

Conclusions

- The sales in the month of December is the highest sales among others.
- The Sales is highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales almost near to Zero.
- Those Stores who takes participate in Promotion got their Sales increased.
- Type of Store plays an important role in opening pattern of stores. All Type ‘b’ stores never closed except for refurbishment or other reason.
- We can observe that most of the stores remain closed during State Holidays. But it is interesting to note that the number of stores opened during School Holidays were more than that were opened during State Holidays.
- We can say that random forest regressor model is our optimal model and can be deploy.

THANK YOU