Capstone Project-2

Retail Sales Prediction

Prepared by:

Suhail Shaikh

Contents

- □ Problem Statement
- □ Data Summary
- □ Data Cleaning
- □ Exploratory Data Analysis
- □ Model Implementation
- □ Model Validation
- □ Model Selection
- □ Hyperparameter Tuning
- □ Conclusion
- □ Challenges
- □ References

Problem Statement

Problem Statement

- Rossmann operates over 3000 drug stores in 7 European countries. Rossmann Managers are tasked with predicting their sales for 6 weeks in advance.

- The sales is influenced by many parameters and the task is to predict the sales based on the parameters.

Data Summary

Data Summary

Data set has 19735 rows and 28 columns. The columns in data set have information as mentioned below :

- **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

Cont…

- **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

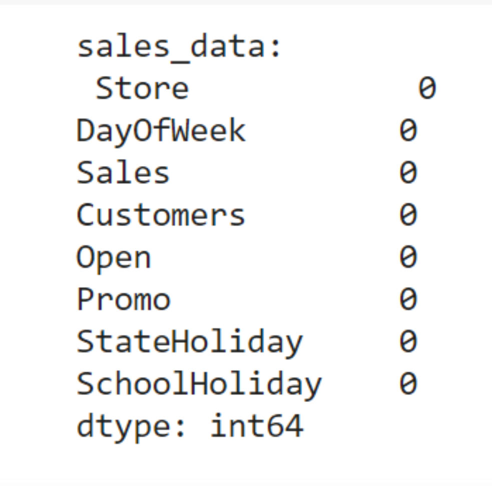
Cont…

- **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

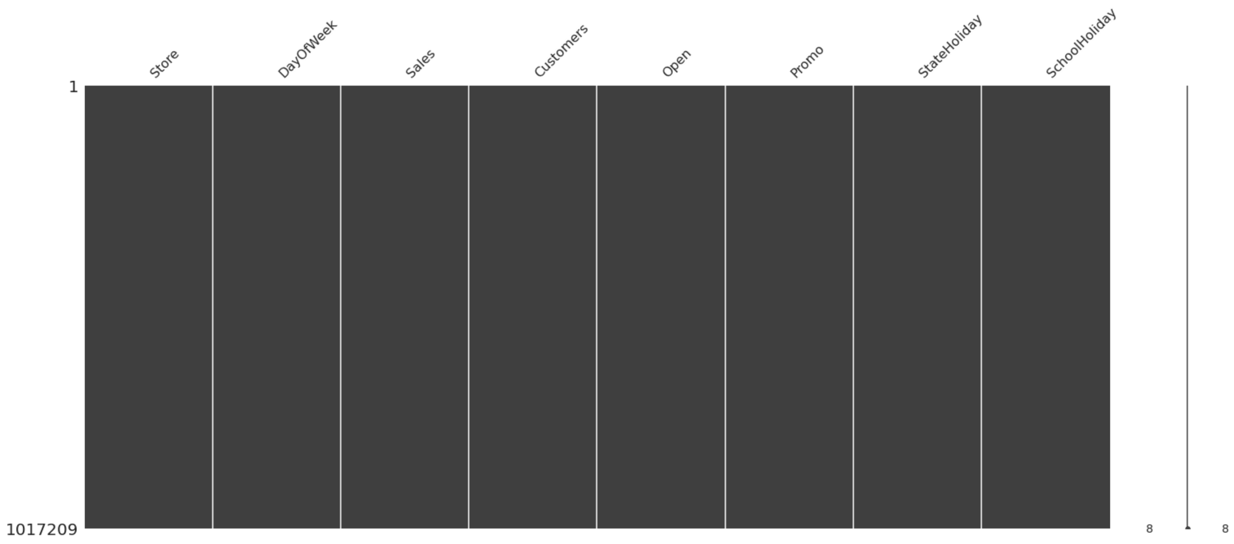
Data Cleaning

Null Values Detection

- **Checking missing values in Sales Data Set**

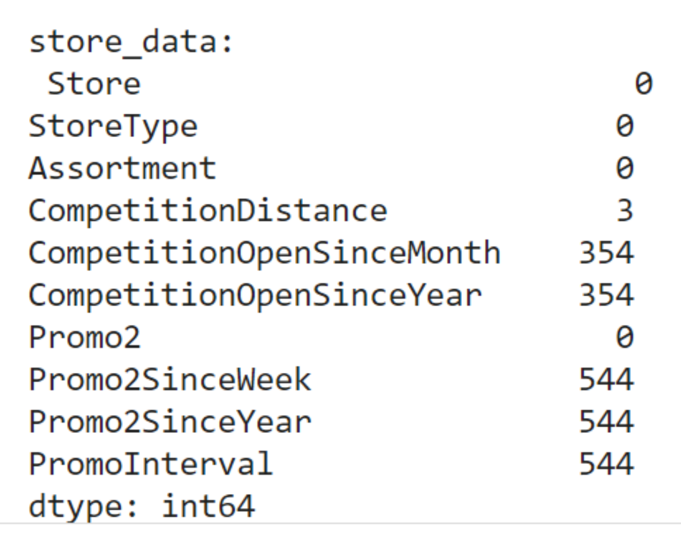


```
sales_data:
 Store           0
DayOfWeek       0
Sales           0
Customers       0
Open            0
Promo           0
StateHoliday    0
SchoolHoliday   0
dtype: int64
```

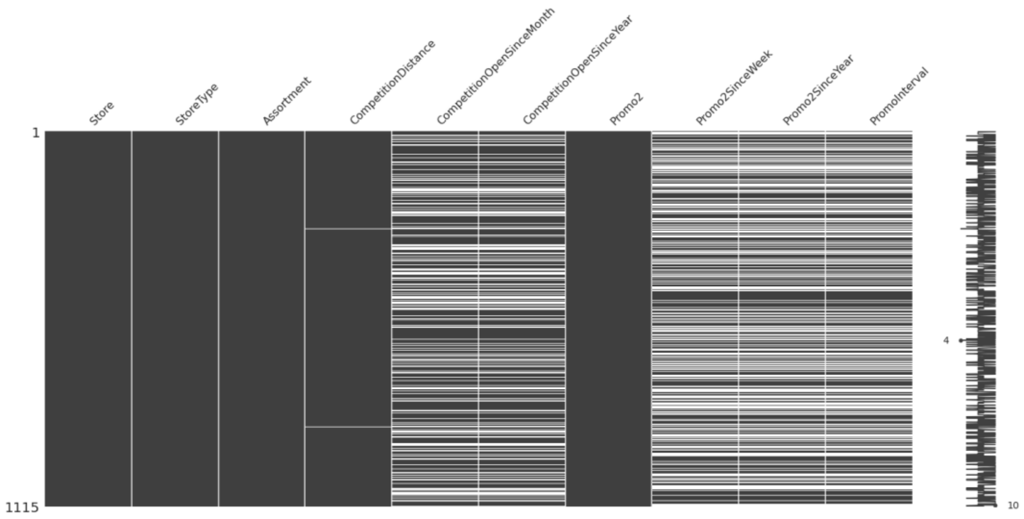


Null Values Detection

- **Checking missing values in Store Data Set**

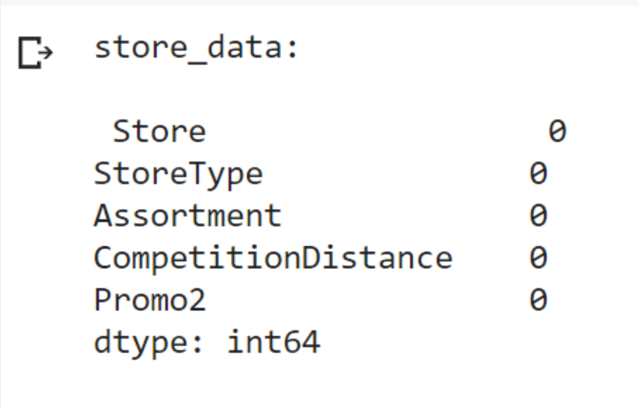


```
store_data:
 Store                         0
StoreType                     0
Assortment                    0
CompetitionDistance           3
CompetitionOpenSinceMonth   354
CompetitionOpenSinceYear    354
Promo2                        0
Promo2SinceWeek             544
Promo2SinceYear            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

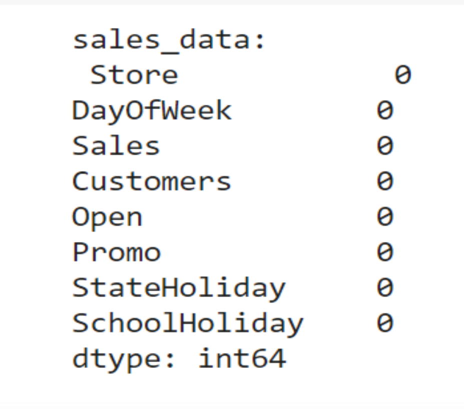


Null Values Removal

- We have dropped the columns with missing values more than 30% and in other columns we have replaced null values by median value.



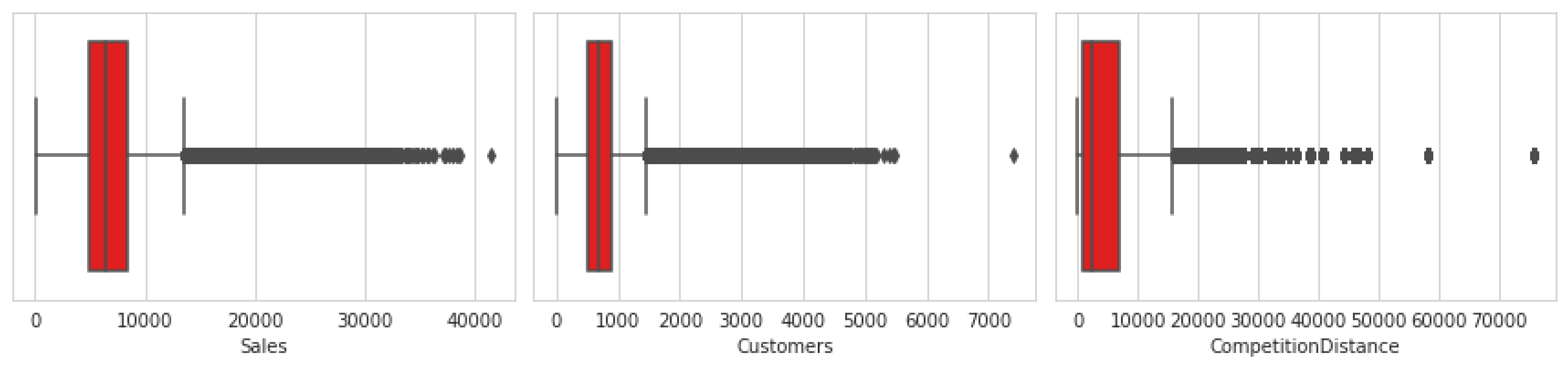
```
store_data:

 Store                  0
StoreType              0
Assortment             0
CompetitionDistance    0
Promo2                 0
dtype: int64
```



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

Outlier Detection

- Box Plot Before Outlier Removal



- **We have plotted boxplot for numerical features to visualize the outliers.**

IQR (Interquartile Range) Method

- We have used IQR Method to remove the outliers from data set. The important parameters of IQR are as follows:

  Q1 = 25% quantile

  Q3 = 75% quantile

  IQR = Q3-Q1

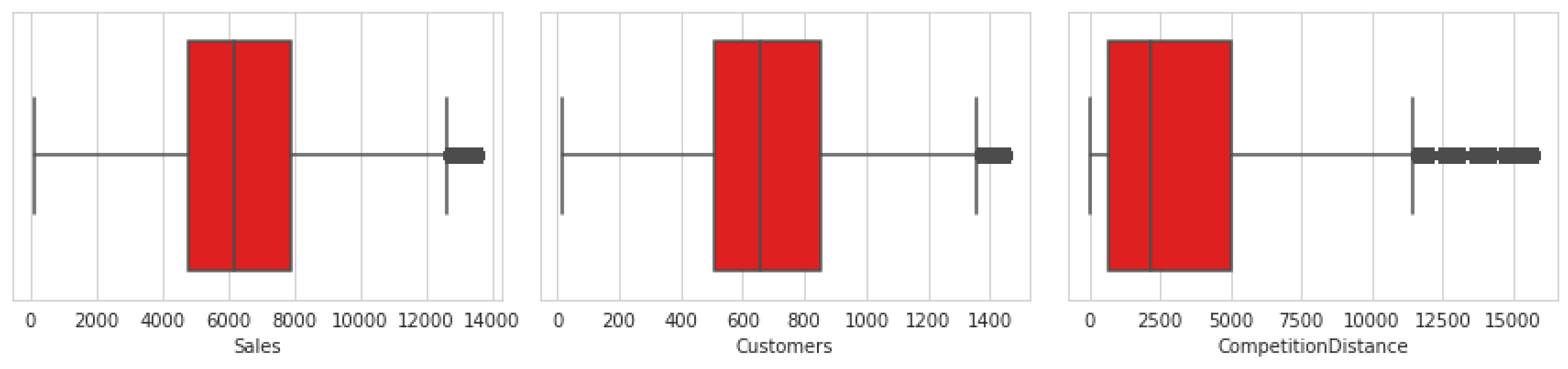
  Lower limit = Q1 - 1.5 * IQR

  Upper limit = Q3 + 1.5 * IQR

- We identified lower and upper limit for our numerical features and set the limits to remove the outliers.

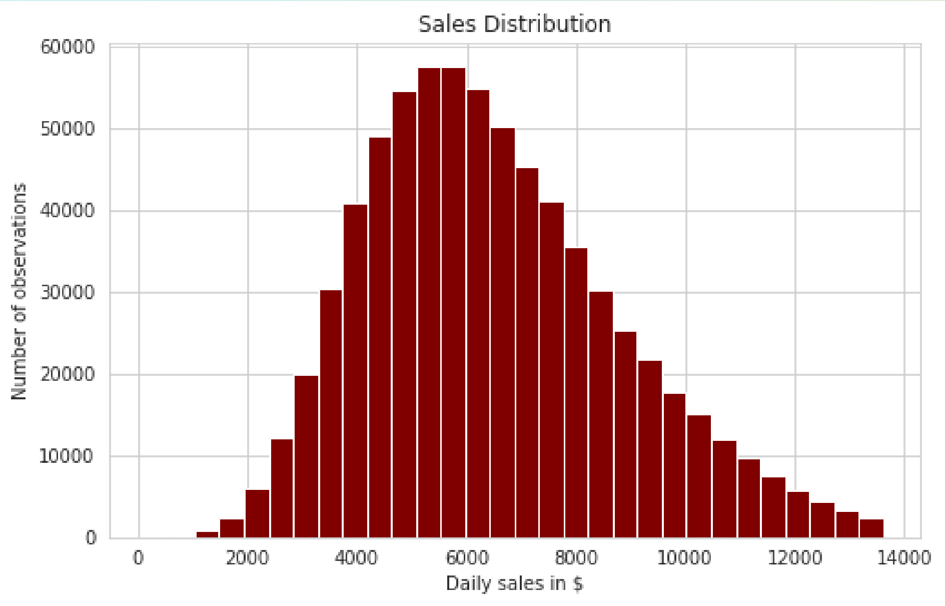
Outliers Removal

- **Box Plot After Outliers Removal**

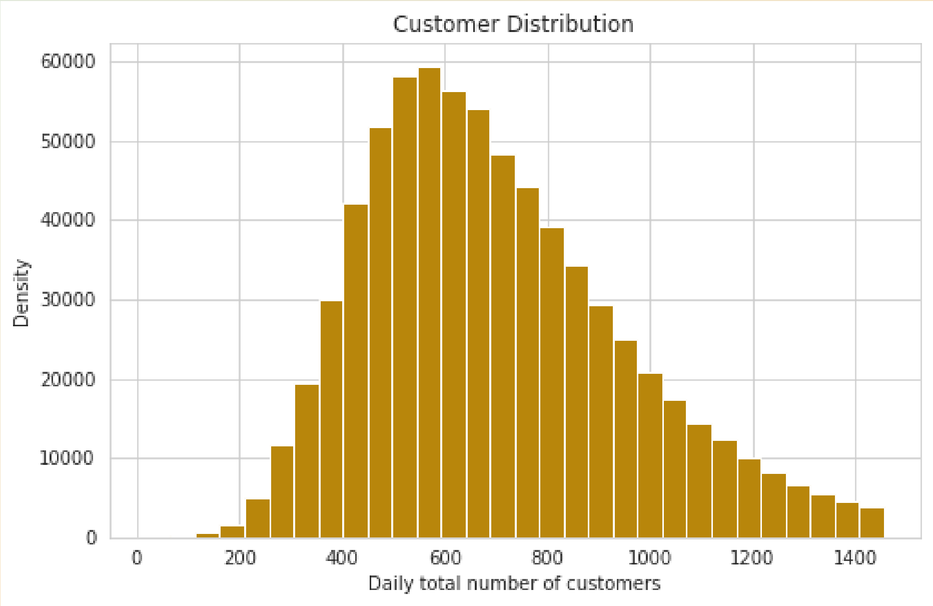


Exploratory Data Analysis

Sales and Customer Distribution



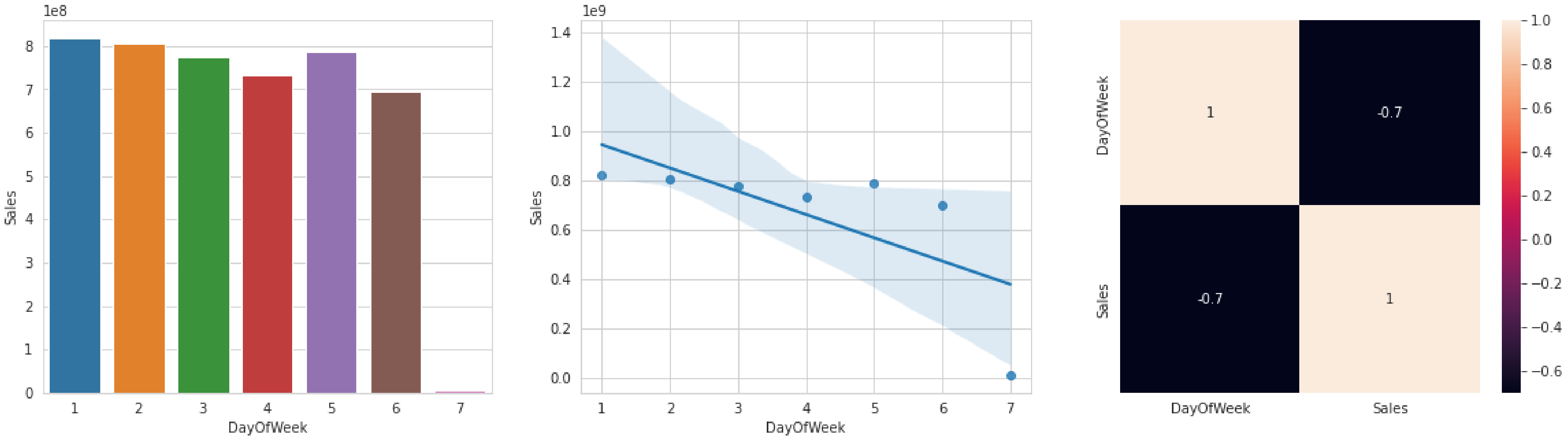
Sales Distribution



Customer Distribution

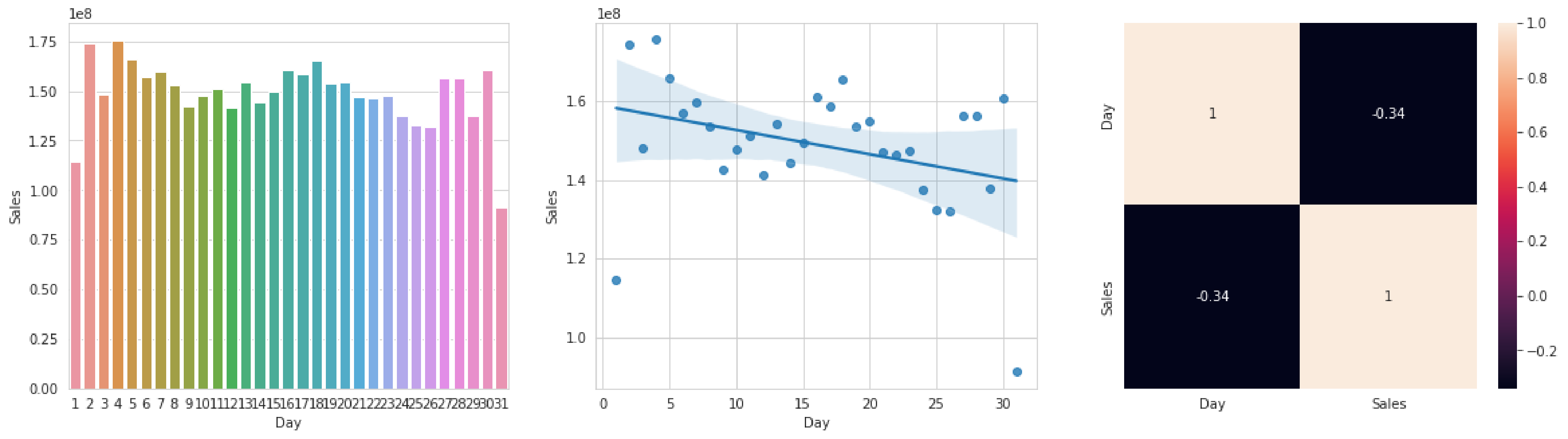
- ***The Sales distribution lived up to the expectation with no irregularities. It seems to be a perfect gaussian distribution with small positive skewness.***

Sales Over the Week



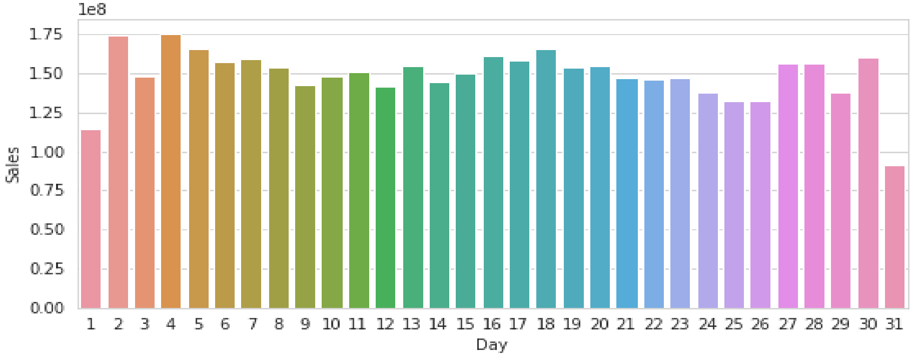
- ***From the graph we can see that sales is very low on Sunday and it is gradually decreasing with the end of week.***

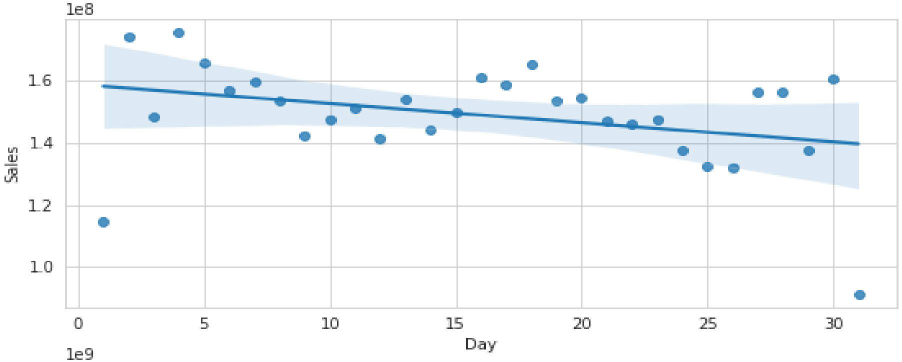
Sales Over the Month

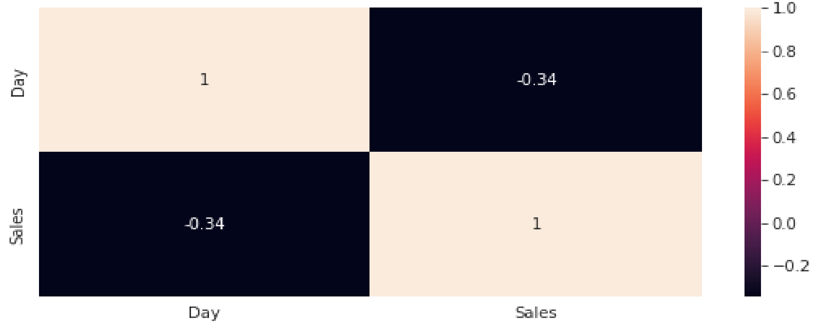


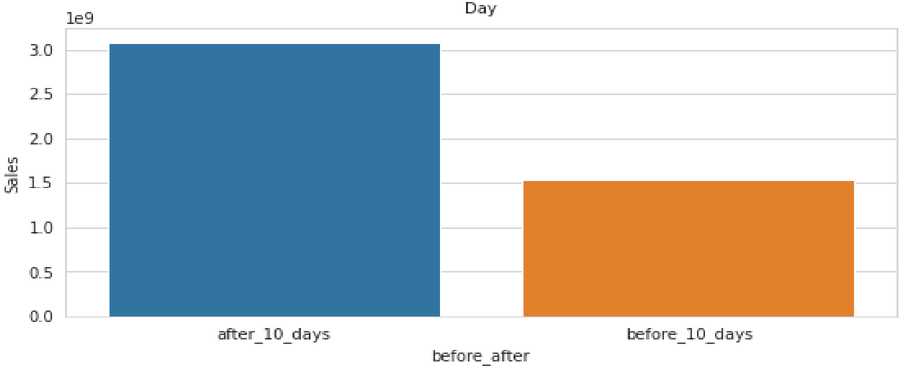
- ***Sales of store is good in first half of the month and it is dropping on second half of the month.***

Sales After 10th in Each Month



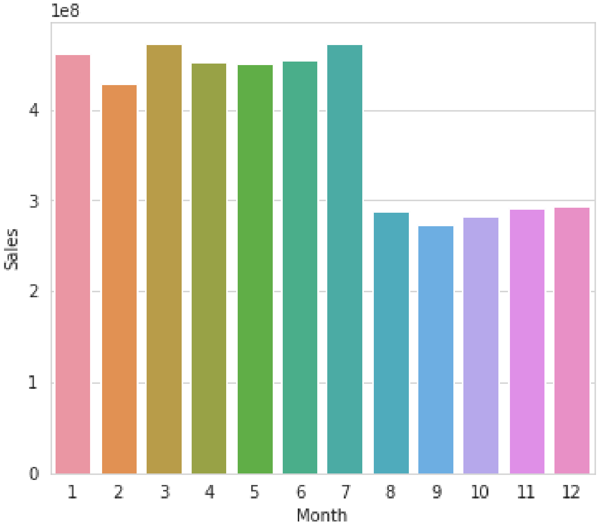


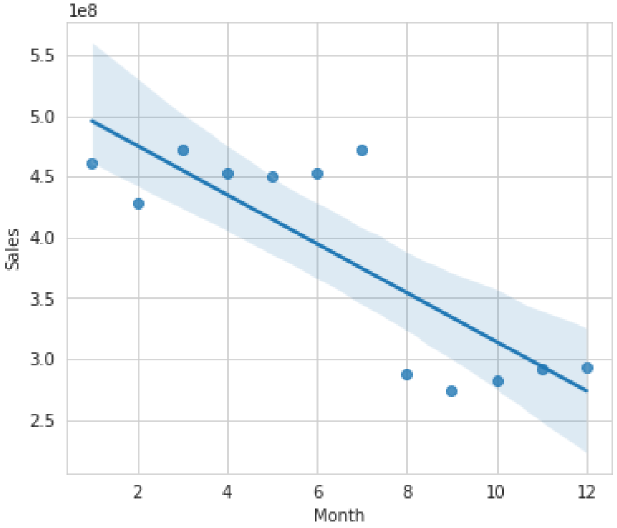


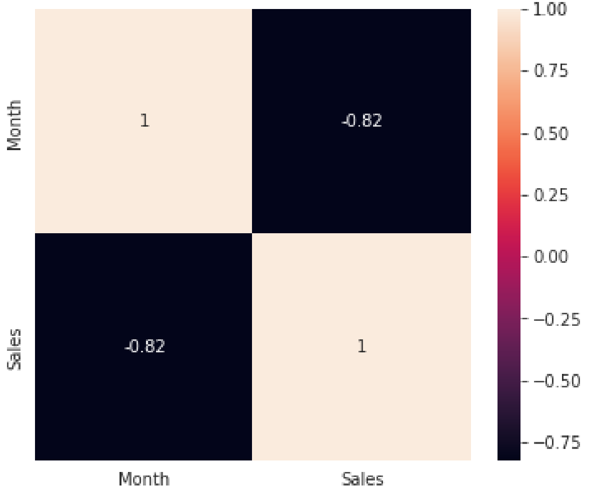


- ***Sales of store is good in first 10 days of month, and it is decreased in second half of month.***

Sales Over the Year

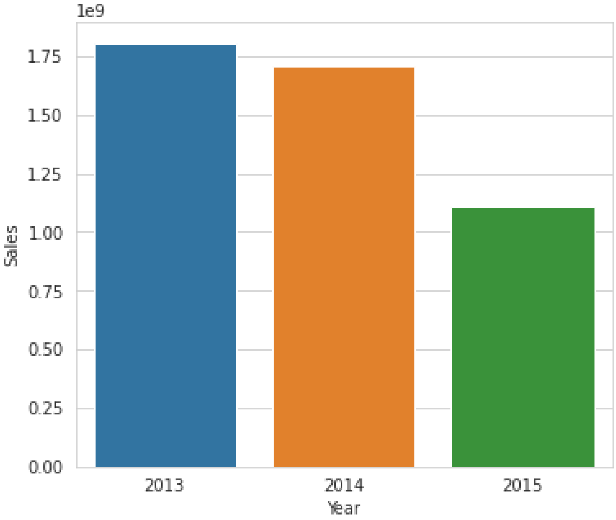


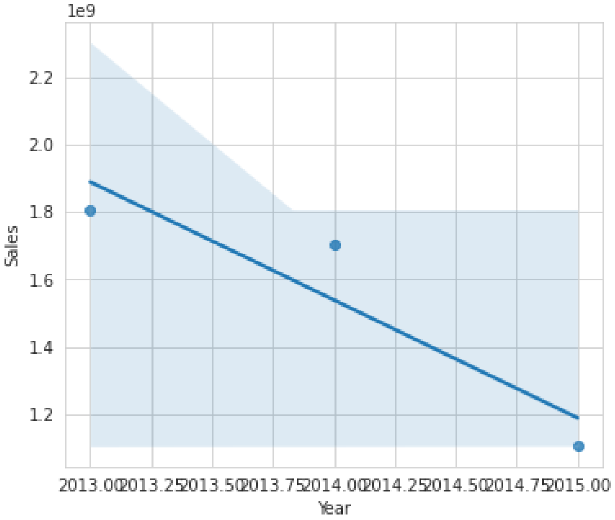


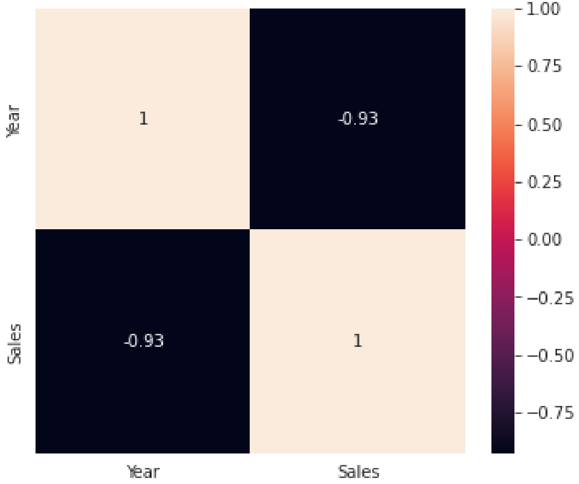


- ***Sales performance is good in first half of each year but it is decreased in second half of the year.***

Sales in Last 3 Years

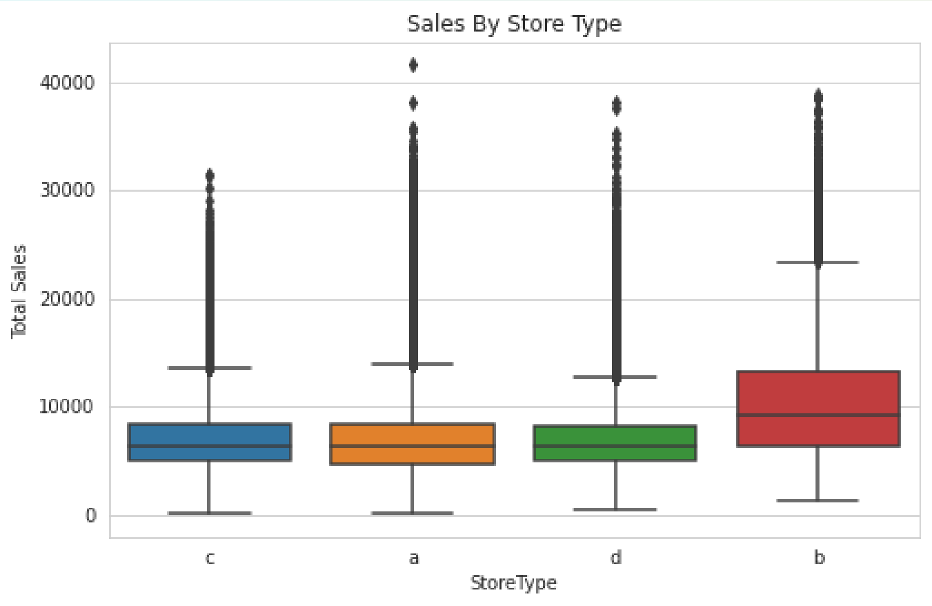


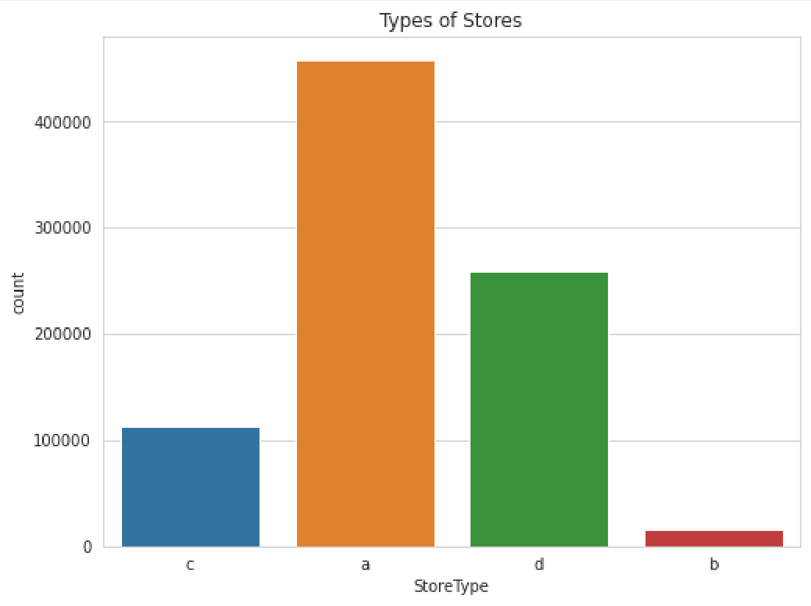




- ***Sales performance was good in year 2013 but it is decreasing each year significantly.***

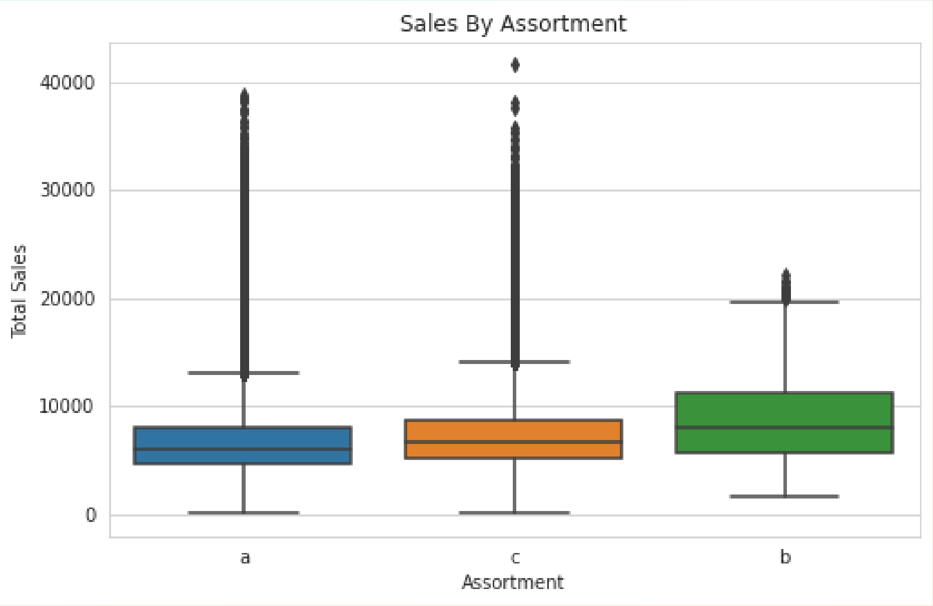
Sales by Store Type

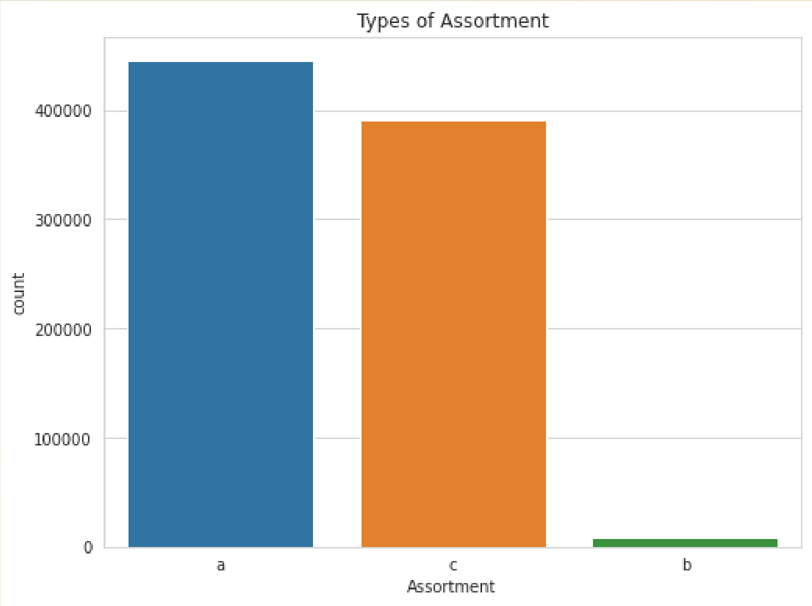




- ***Store type B has higher sales as compared to other types of store.***

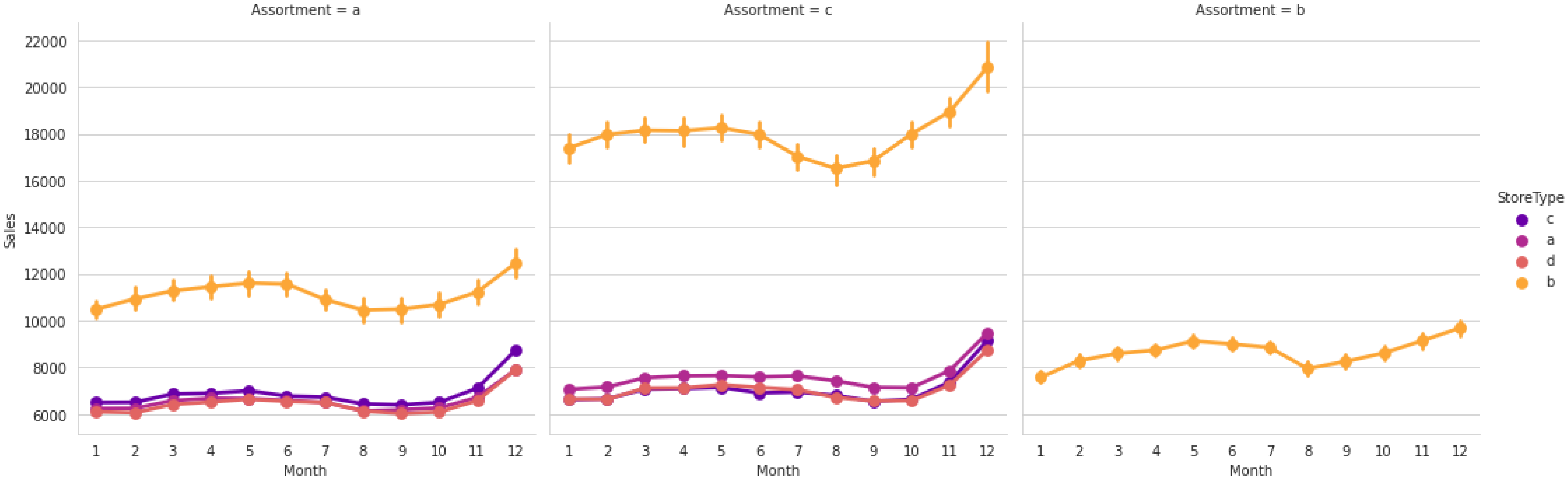
Sales by Assortment Type





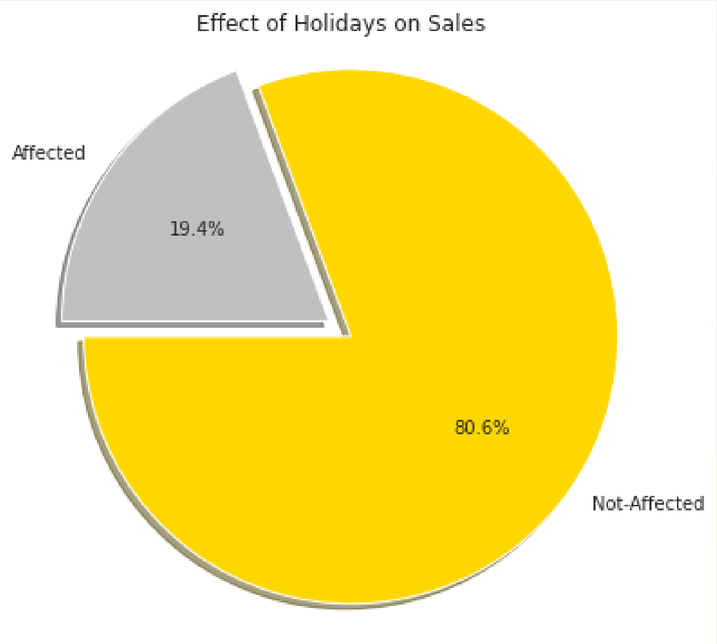
- ***Assortment type B has higher sales as compared to other types.***

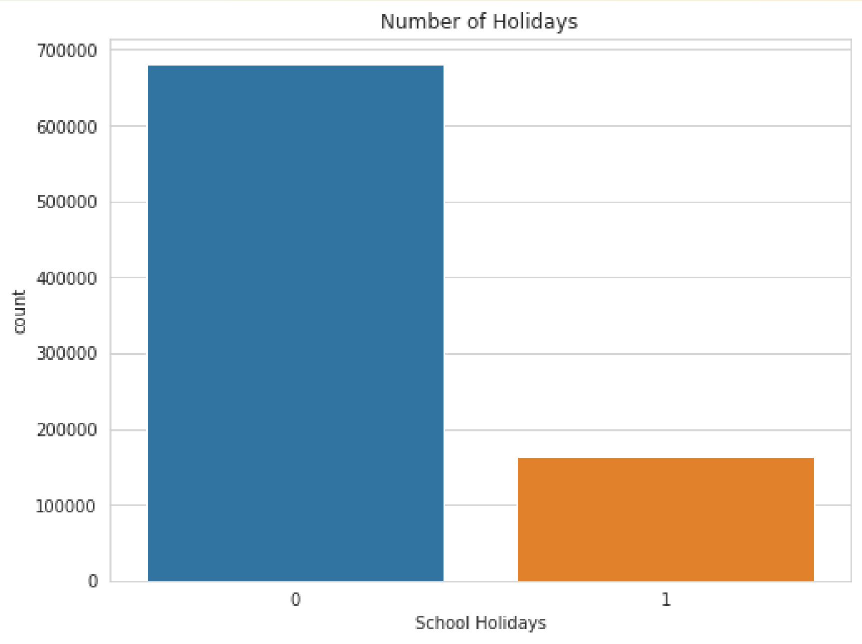
Sales Over Month



- ***Store type B has higher sales over the months in year. In all types of Assortment it has performed well.***

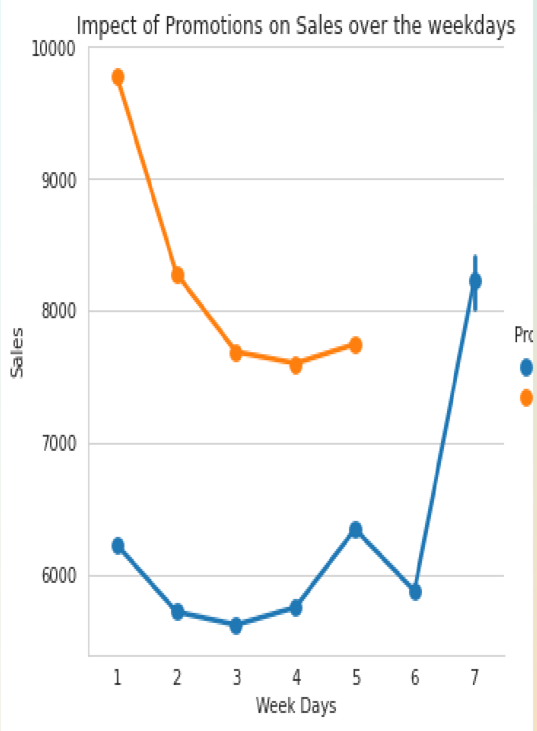
Effect of School Holidays on Sales



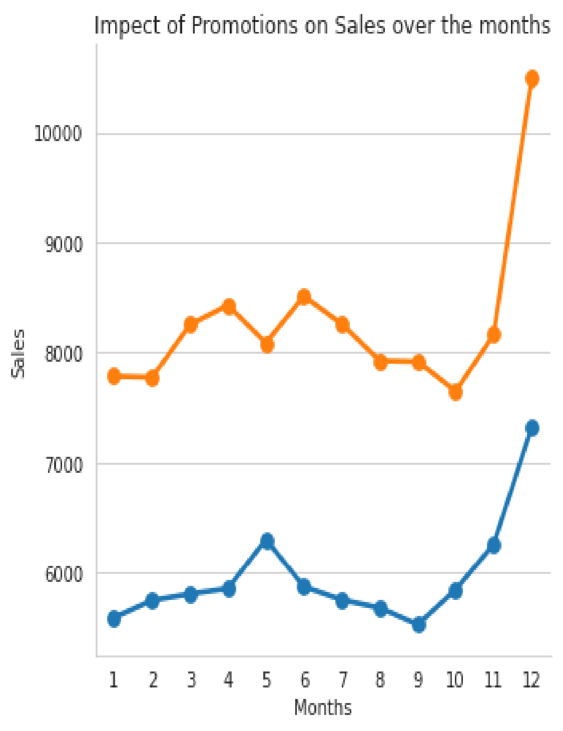


- ***Sales was comparatively lower on school holidays.***

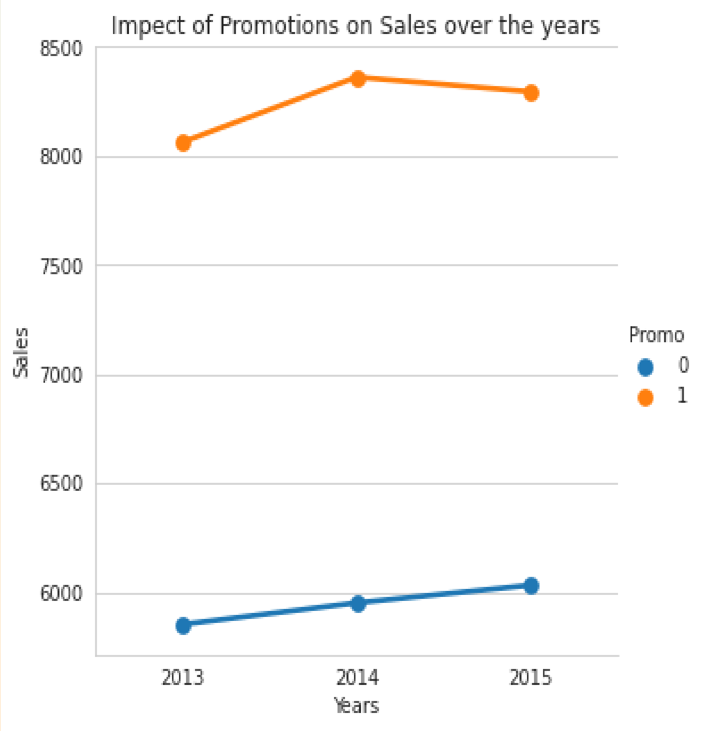
Impact of Promotions on Sales



Impact of Promotions on Sales over the weekdays



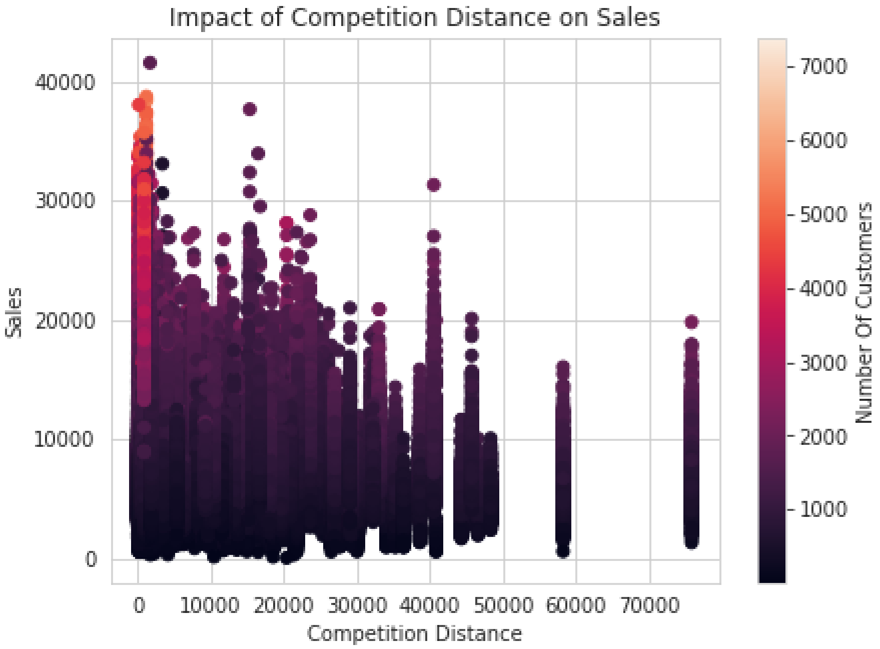
Impact of Promotions on Sales over the months

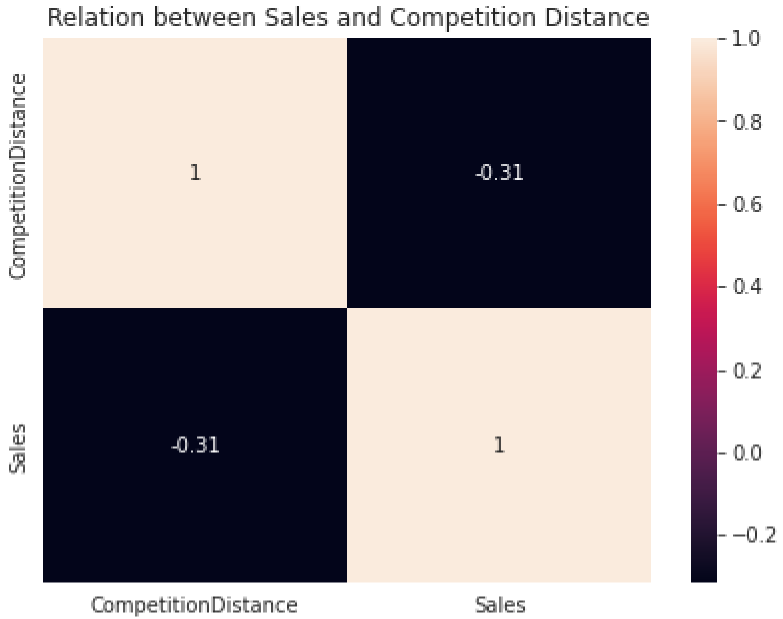


Impact of Promotions on Sales over the years

- ***Approximately 40% sales is increased annually by Promotions.***

Impact of Competition Distance on Sales





- ***At lower competition distance sales is higher because most of the competitors are located where customer demand is high.***

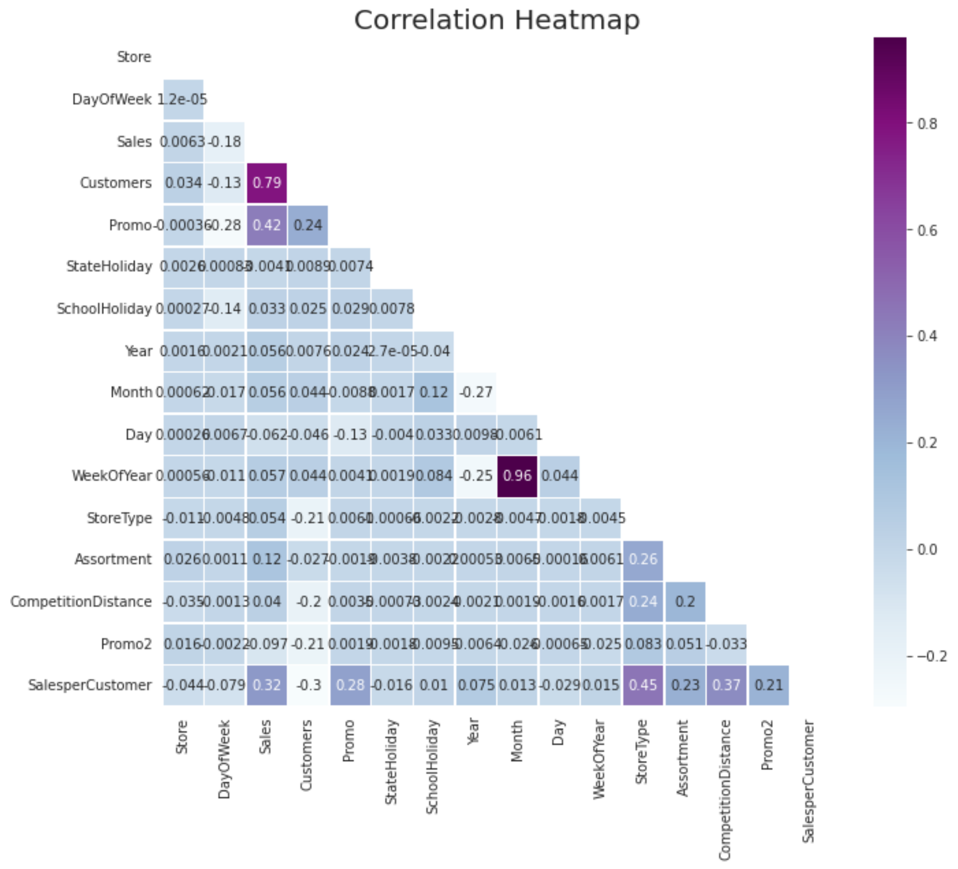
Identifying Correlation

- **Dependent Features/Variables:**

Sales

- **Independent Features/Variables:**

Store, DayOfWeek, Customers, Promo, StateHoliday, SchoolHoliday, Year, Day, WeekOfYear, StoreType, Assortment, CompetitionDistance, Promo2, Month, SalesperCustomer.

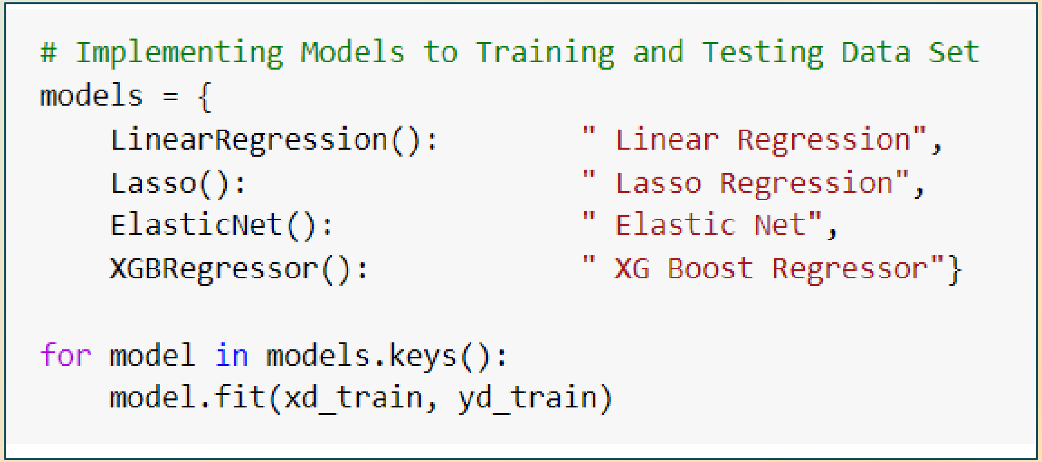


Model Implementation

Model Implementation

We have implemented below mentioned models to our data set.

- **Linear Regression**
- **Lasso Regression**
- **Elastic Net**
- **XGBoost Regressor**

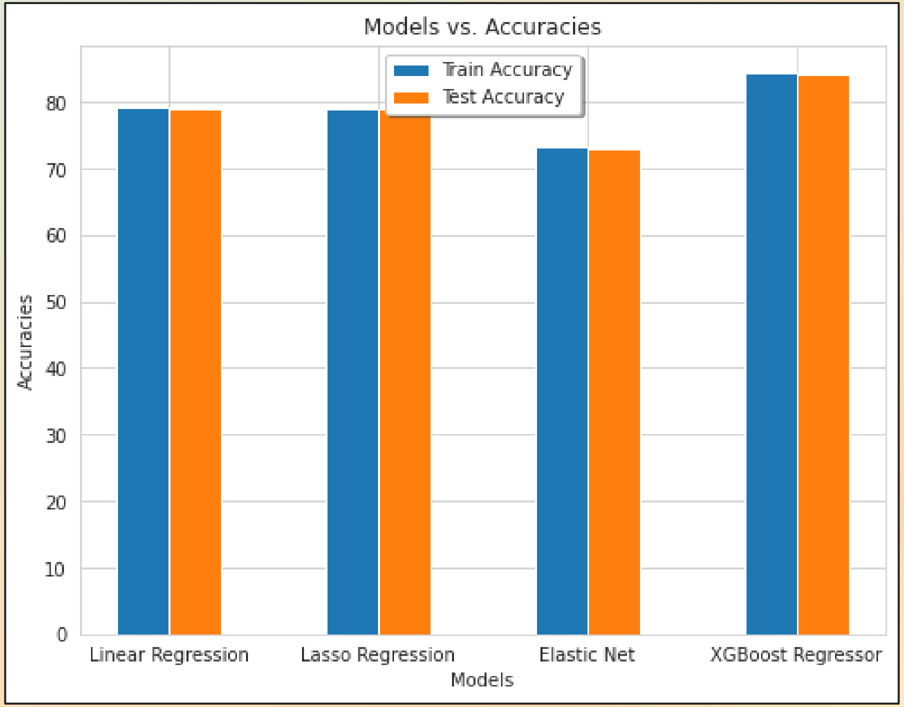


```
# Implementing Models to Training and Testing Data Set
models = {
    LinearRegression():        " Linear Regression",
    Lasso():                   " Lasso Regression",
    ElasticNet():              " Elastic Net",
    XGBRegressor():            " XG Boost Regressor"}

for model in models.keys():
    model.fit(xd_train, yd_train)
```

Model Evaluation

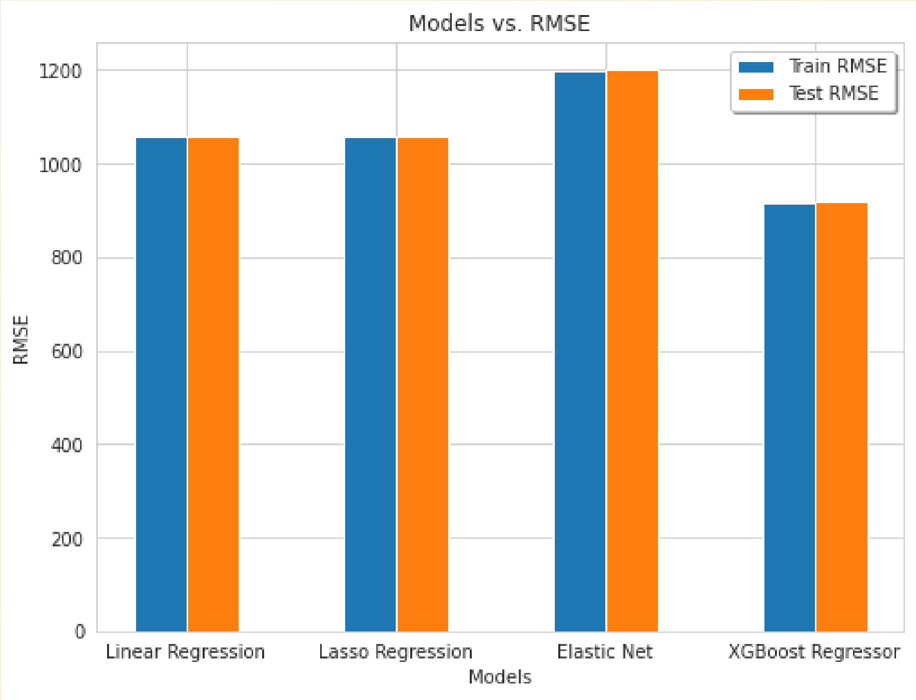
Model Accuracy

Model	Accuracy	
	Train	Test
Linear Regression	79.11%	78.98%
Lasso Regression	79.10%	78.97%
Elastic Net	73.19%	72.99%
XGBoost Regressor	84.35%	84.15%



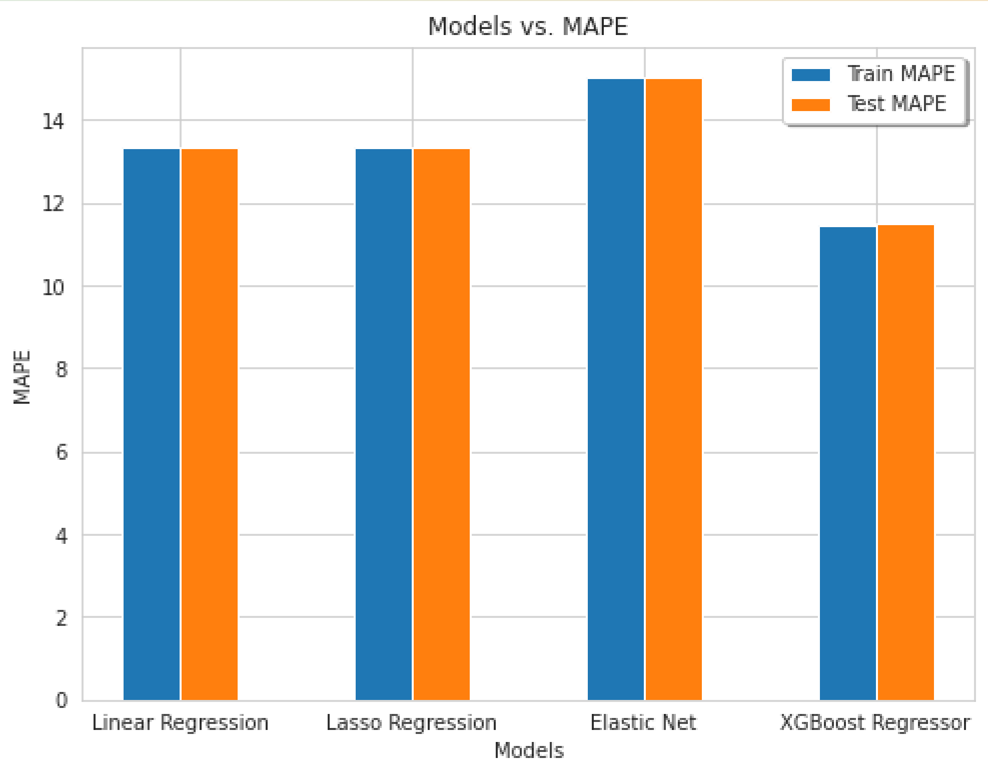
Root Mean Square Error

Model	RMSE	
	Train	Test
Linear Regression	1056.63	1058.99
Lasso Regression	1056.81	1059.23
Elastic Net	1196.95	1200.41
XGBoost Regressor	914.42	919.61



Mean Absolute Percentage Error

	RMSE	
Model	Train	Test
Linear Regression	13.35	13.36
Lasso Regression	13.35	13.36
Elastic Net	15.02	15.03
XGBoost Regressor	11.46	11.53



Model Selection

Model Selection

- We have implemented 4 models, Linear, Lasso, Elastic Net and XGBoost Regressor.
- Performance of Linear and Lasso Regression model is almost similar, both model fitted well to data set but have comparatively less accuracies.
- Elastic Net Regressor has lowest accuracy among all the models.
- XGBoost Regressor has highest accuracy i.e. (84.35%).It has lowest 'Root Mean Square Error' and 'Mean Absolute Percentage Error’.
- So we will select XGBoost Regressor for Data Modelling.

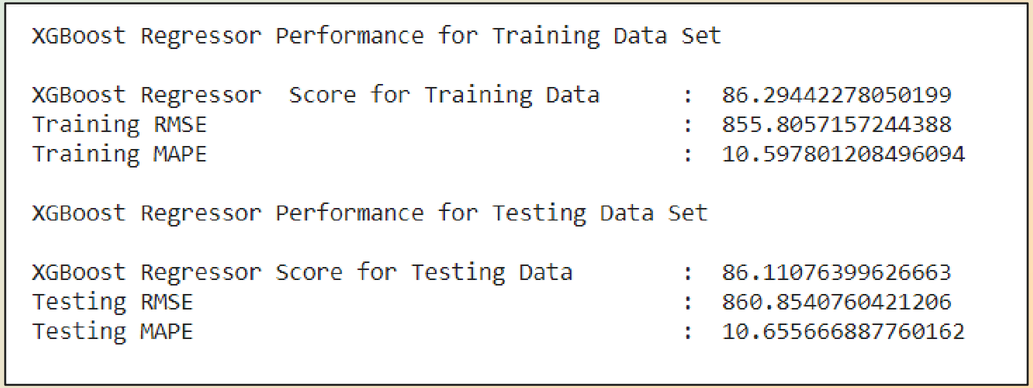
Hyperparameters Tuning

Hyperparameters Tuning

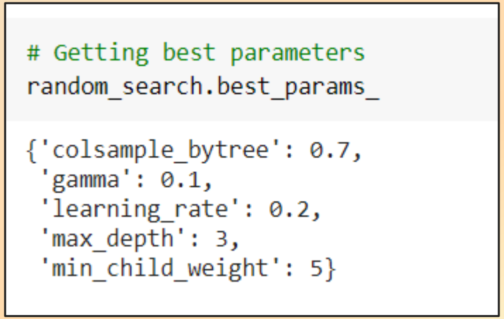
- We have used RandomizedSearchCV to search optimum parameters.

- We used these best parameter to build a optimum model.

- The efficiency of model is increased by 2% with best parameters.



```
XGBoost Regressor Performance for Training Data Set

XGBoost Regressor  Score for Training Data      :  86.29442278050199
Training RMSE                                   :  855.8057157244388
Training MAPE                                   :  10.597801208496094

XGBoost Regressor Performance for Testing Data Set

XGBoost Regressor Score for Testing Data        :  86.11076399626663
Testing RMSE                                    :  860.8540760421206
Testing MAPE                                    :  10.655666887760162
```



```
# Getting best parameters
random_search.best_params_
```

```
{'colsample_bytree': 0.7,
 'gamma': 0.1,
 'learning_rate': 0.2,
 'max_depth': 3,
 'min_child_weight': 5}
```

Conclusion

Conclusion

- We have implemented 4 models, Linear, Lasso, Elastic Net and XGBoostcRegressor.
- Performance of Linear and Lasso Regression model is almost similar, both model fitted well to data set but have comparatively less accuracies (78.98% & 78.97%).
- Elastic Net Regressor has lowest accuracy (72.99%) among all the models.
- XGBoost Regressor has highest accuracy i.e (84.35%).It has lowest 'Root Mean Square Error' and 'Mean Absolute Percentage Error’. Hence we selected XGBoost Regressor.
- After tuning hyperparameters, efficiency of model is increased by 2%.

Suggestions to Improve Sales

- ❑ In order to increase the sales and to attract new customer, promotions should be done.
- ❑ Stores should be located where competitors are less and consumer demand is significant.
- ❑ In order to maintain the sales in second half year, discount, coupons and vouchers should be offered to customers.
- ❑ Sales is stores is less on weekend and holidays, so Mega Sale or Discount Sale should be arranged on these days.

Challenges

Challenges

- **The data set was huge, so computational time involved was high.**

- **Hyperparameter tuning was time consuming process to get the optimum parameters for model.**

- **Due to huge data set, time required for figures plotting was high.**

References

References

□ Kaggle

□ Youtube

□ Github

□ Towards data science

□ Analytics Vidya

□ Code basics

□ Stack over flow

Thank You