**Introduction**

Retail Store Sales Prediction using Python and Machine learning (Big-Data Analysis).

**Background and Literature Review**

Sales Forecasting plays a substantial role in identifying the sales trends of products for the future era in any organization. These forecasts are also important for determining the profitable retail operations to meet customer demand, maintain storage levels and to identify probable losses.

https://www.researchgate.net/publication/360626956\_Retail\_Sales\_Forecasting\_Using\_Deep\_Learning\_Systematic\_Literature\_Review

**Methodology**

Initially, a literature review is performed to identify machine learning methods suitable for forecasting the sales of truck components and then based on the results obtained, several experiments were conducted to evaluate the performances of the chosen models

Data Summary

We have two datasets. Rossman store data is for years 2013, 2014 and 2015 with 10,17,209 observations on 9 variables. Stores data with 1115 observations on 10 variables. Some important features are:-

1. Customer : - The number of customers on a given day in a store.

2. Date :- Showing dates for observations.

3. State Holiday :- Indicating a state holiday.

4. Store Type : Differentiate between 4 different store models (a,b,c,d).

5. Assortment : Describes an assortment level i.e a : basic, b : extra and c : extended.

6. Competition Distance : Distance in meters to the nearest competition store.

7. Promo :- Indicates whether a store is running a promo on that day.

Data Preprocessing

Columns having >30% null values are dropped.

Null values in ‘Competition Distance’ are imputed with median of feature.

Removing those stores observations that are temporarily closed (~ 17.3K) & stores generating zero sales.

**Data Analysis**

We have done data visualization like

1. Sales are normally distributed with slightly right tail skewed.
2. Impact of Promo on sales
3. Day Wise trends in Sales
4. School and State holidays effect on sales
5. Monthly trends in Sales
6. Yearly Distribution of Sales according to store types
7. Store Types and average sales/customer/spending relation
8. Impact of Competition Distance on Sales and Customers

**Findings**

1. Sales are highly correlated to customers.

2. Stores opened on ‘State Holiday’ makes a good amount of sales.

3. There is no such significant difference in sales on ‘School Holidays’.

4. Even though store type ‘b’ has very less number of stores but these are outperforming other store types in terms of sales and avg customers.

5. Sales are consistent for the second quarter of the year but it starts increasing in the last quarter.

Next Step is to:

1. Extracting week, month, year from Date and adding them in dataset.
2. Merging both dataset.
3. One hot encoding for Storetype, Assortment.
4. Splitting dataset into Training and Test set and applying MinMaxScaler for scaling dataset.

Models Implemented

1. Linear Regression (Baseline Model)

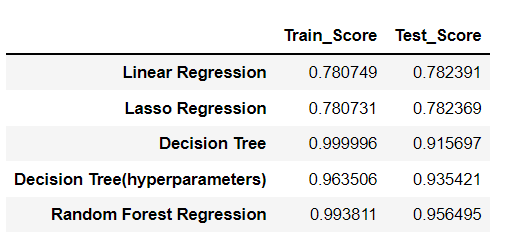
2. Lasso Regression

3. Decision Tree Regression

4. K-Nearest Neighbors Regression

5. Random Forest Regressor

Model Evaluation

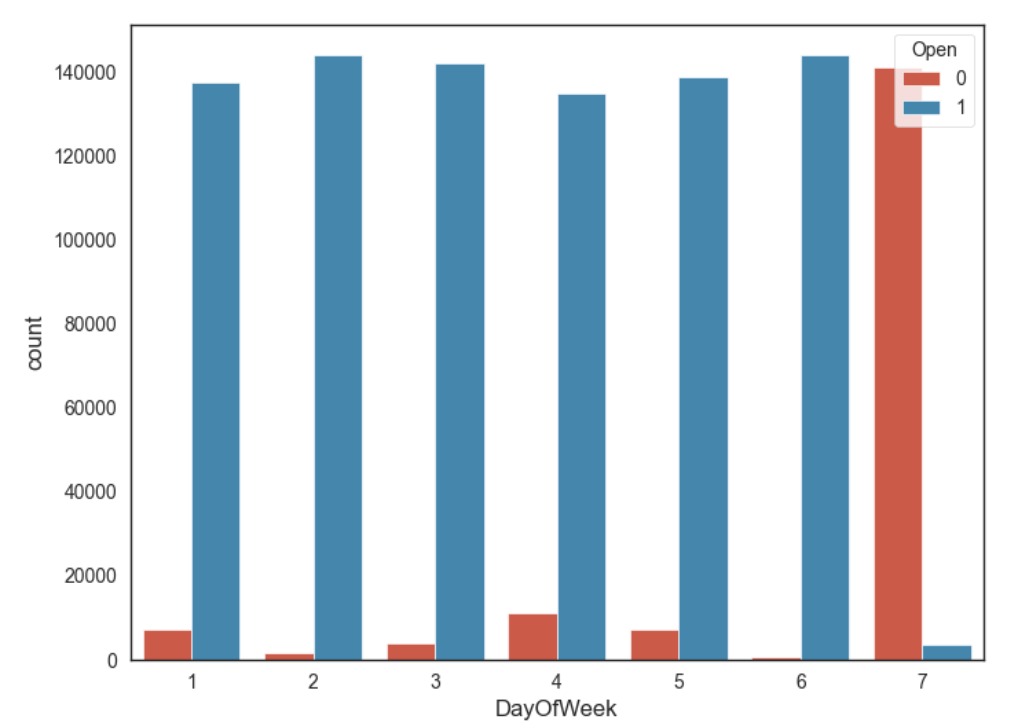
****

**Conclusion**

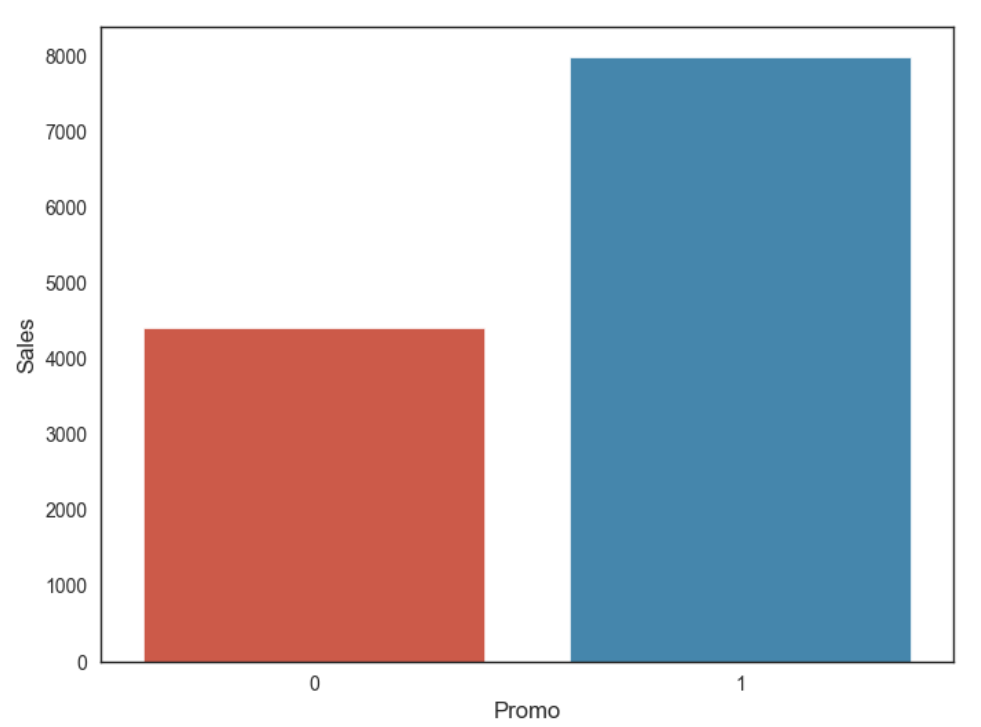
Our model shows that Customers, Competition distance, Store type are some of the most important features in our sales prediction. We need to focus on these aspects to maximize our profits for the next 6 weeks.

**Analysis and Specification**

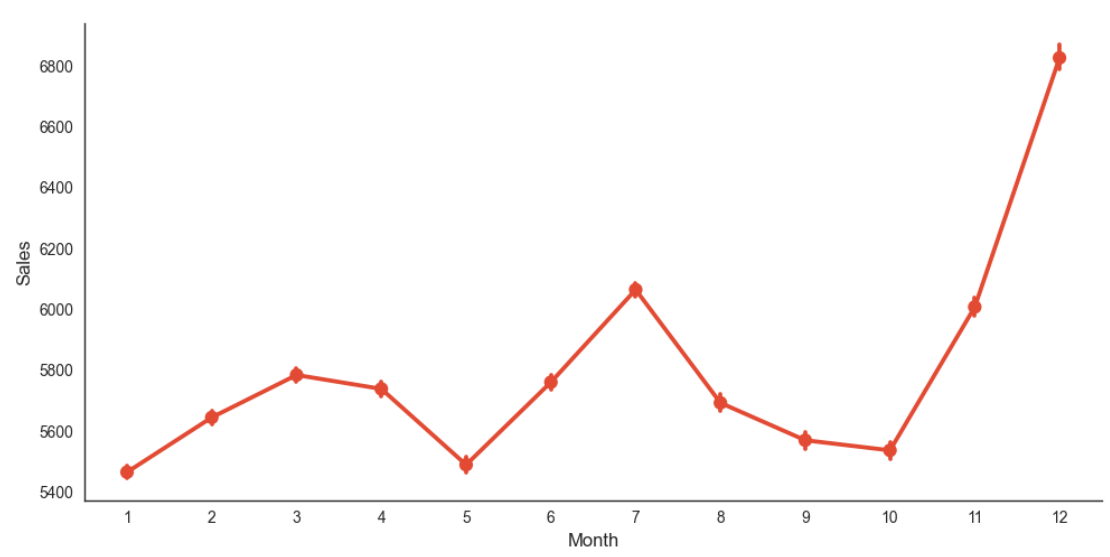
The first step of the analysis is to study the data set, which contains the sales information from the retail store.



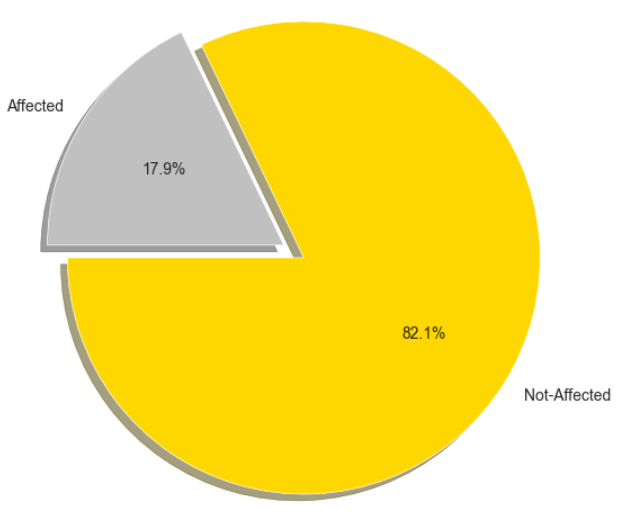
The above bar graph shows days in a week where store is open/close.



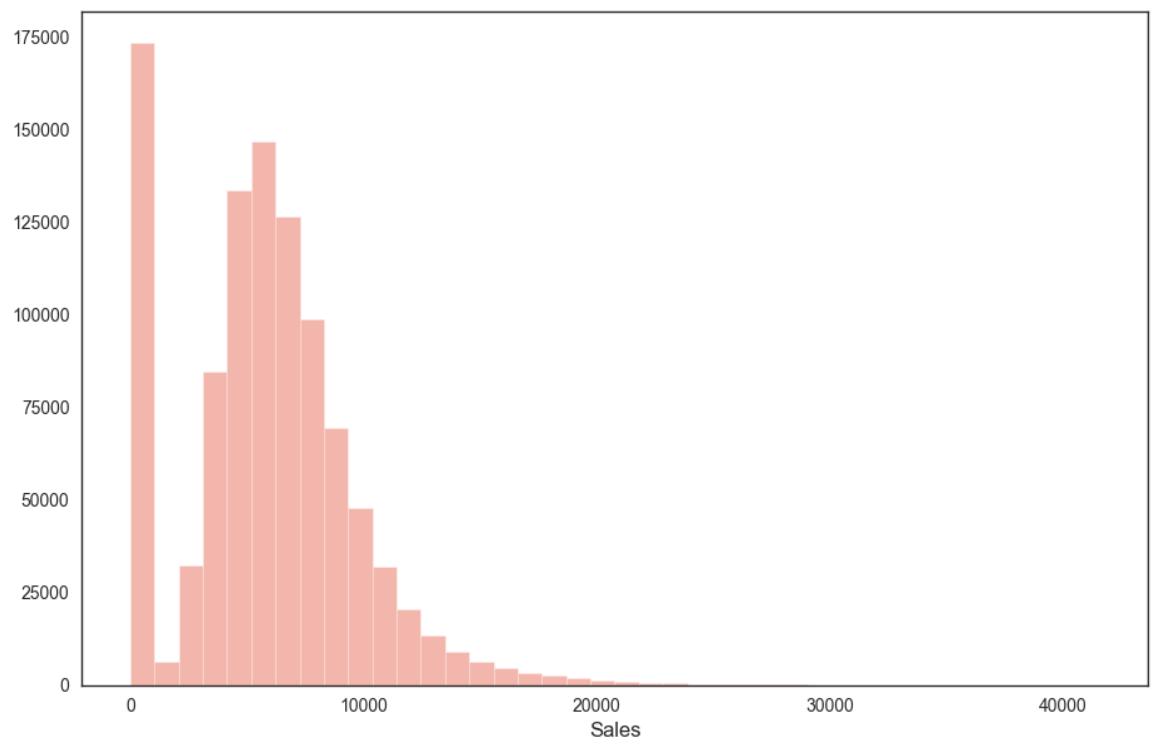
**Above graph shows that sales are nearly doubled High when Promo is running.**

****

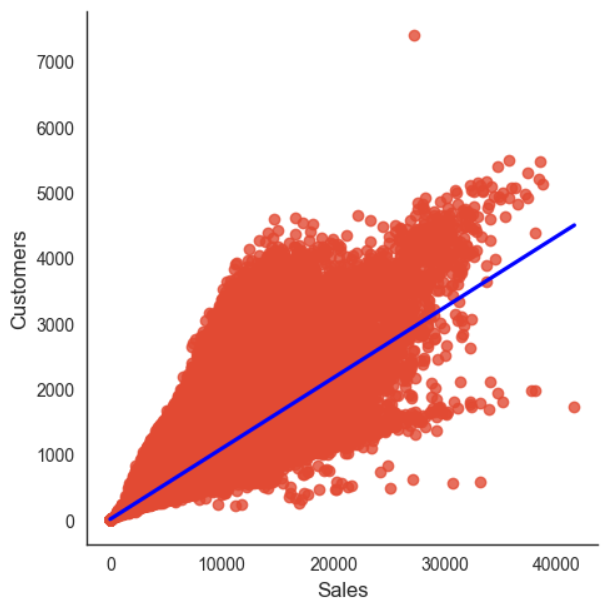
**As We can see that in the month of November and Specially in December Sales is increasing Rapidly every year on the Christmas eve.**

****

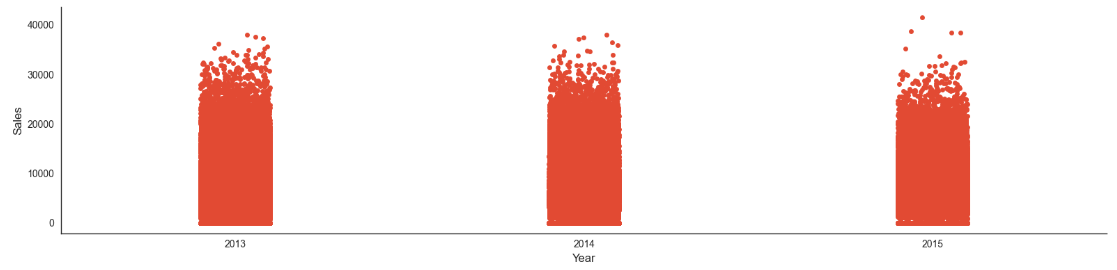
Sales are affected when school holidays are going on.



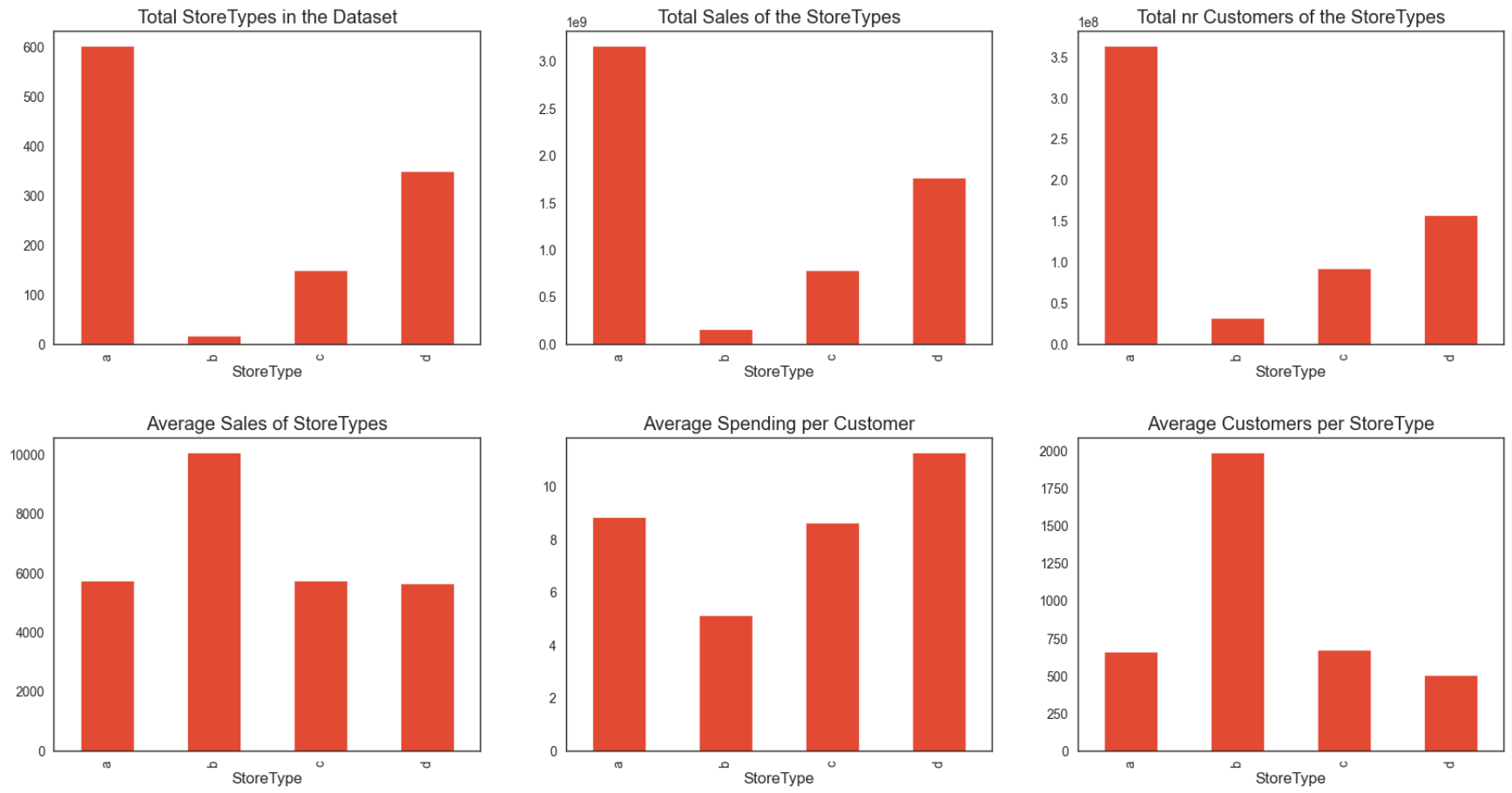
Above is Histogram Representation of Sales. Here 0 is showing because most of the time store was closed.



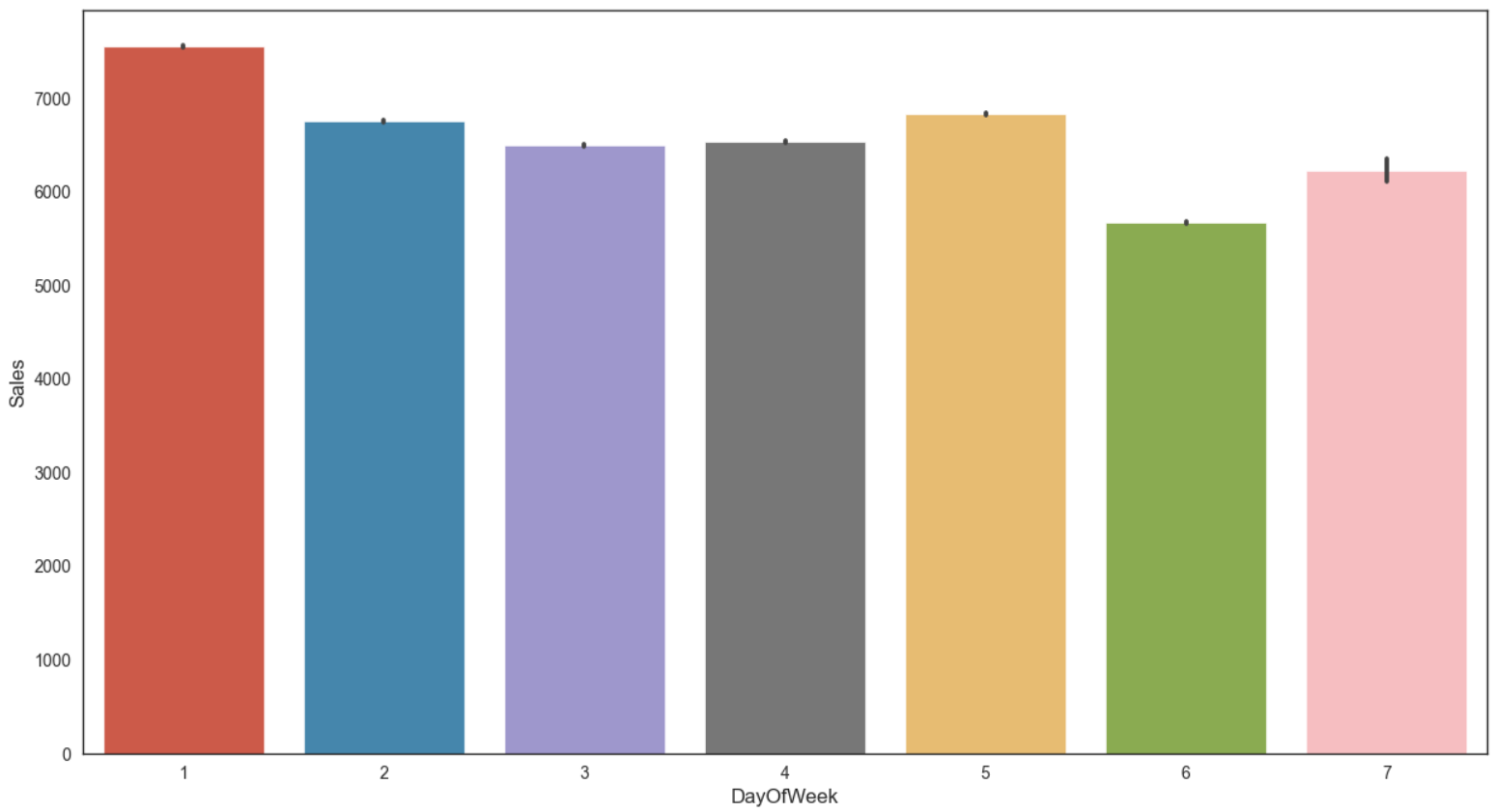
Above is Line graph for showing linear relation between sales and customers.



We can see that there are not such significant differences in these 3 years in terms of sales.

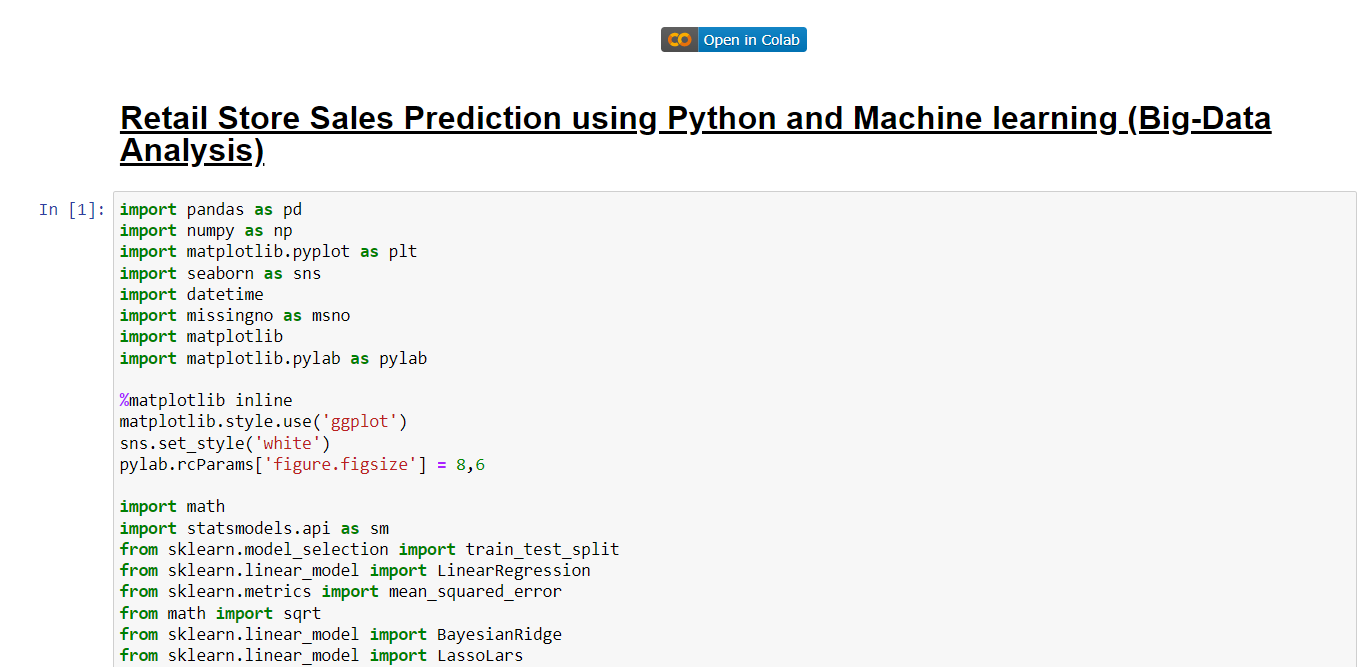


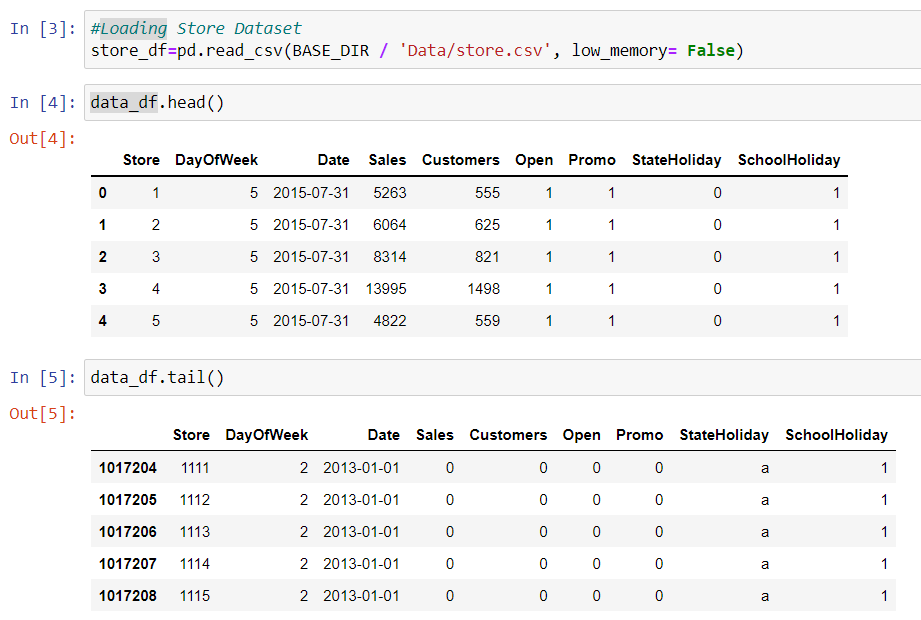
Above Charts shows different store comparisons.

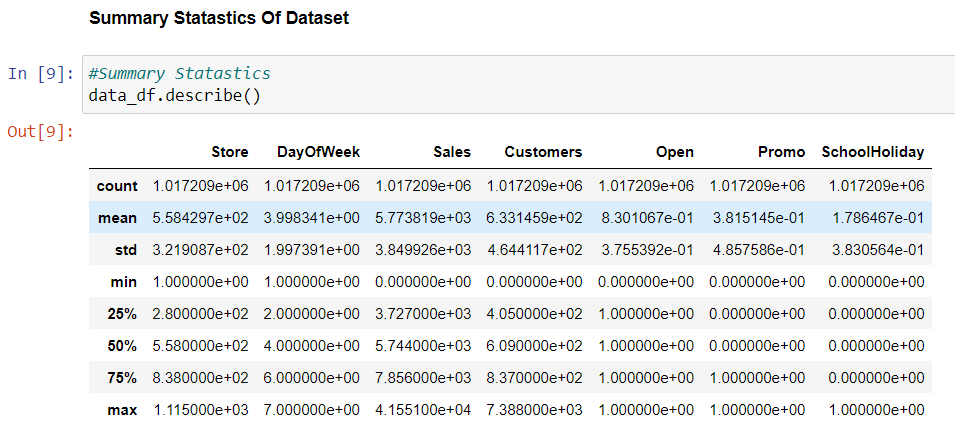


Plot for sales in terms of days of the week

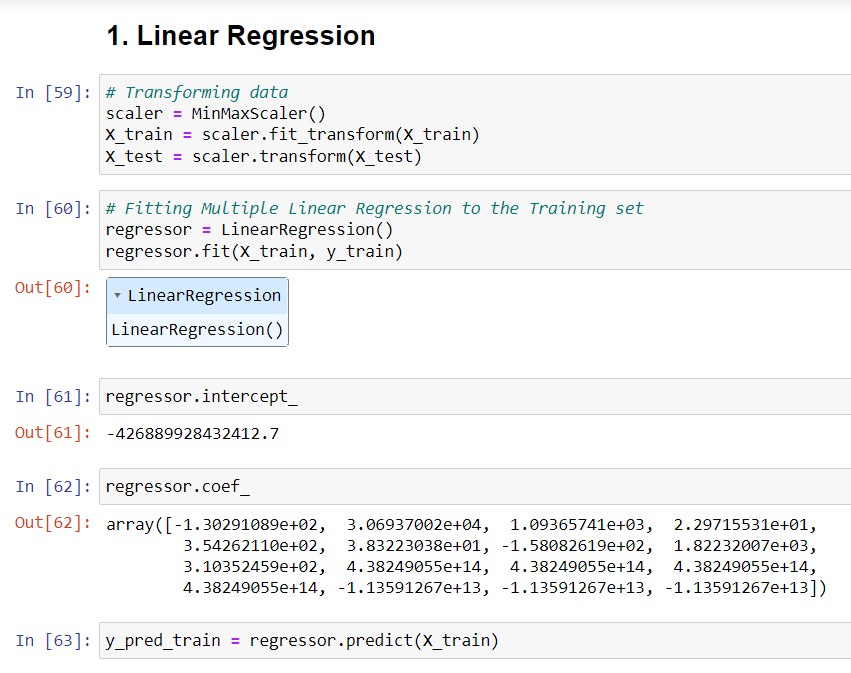
**Implementation**

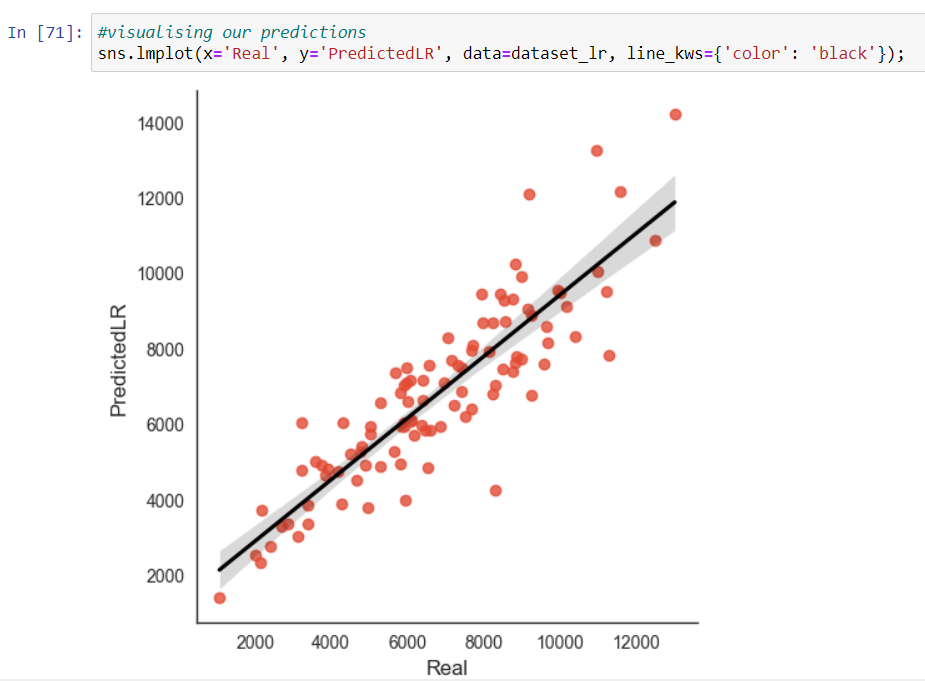


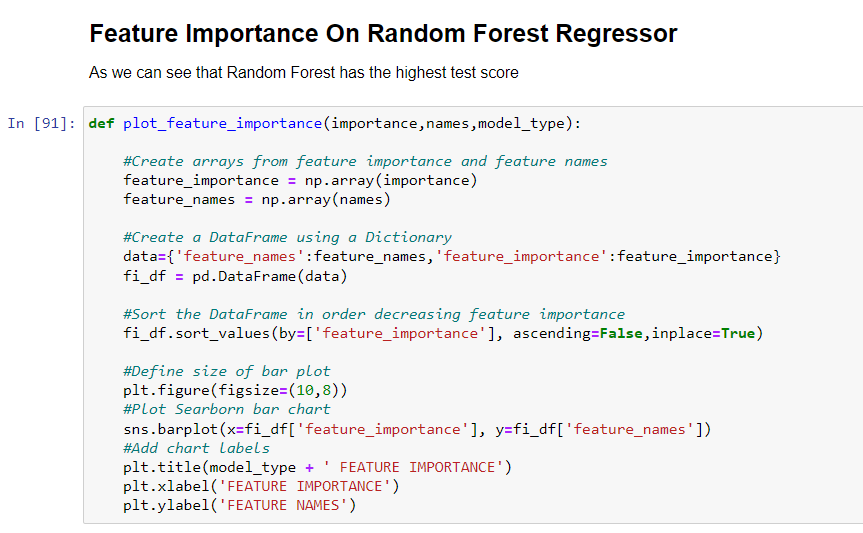


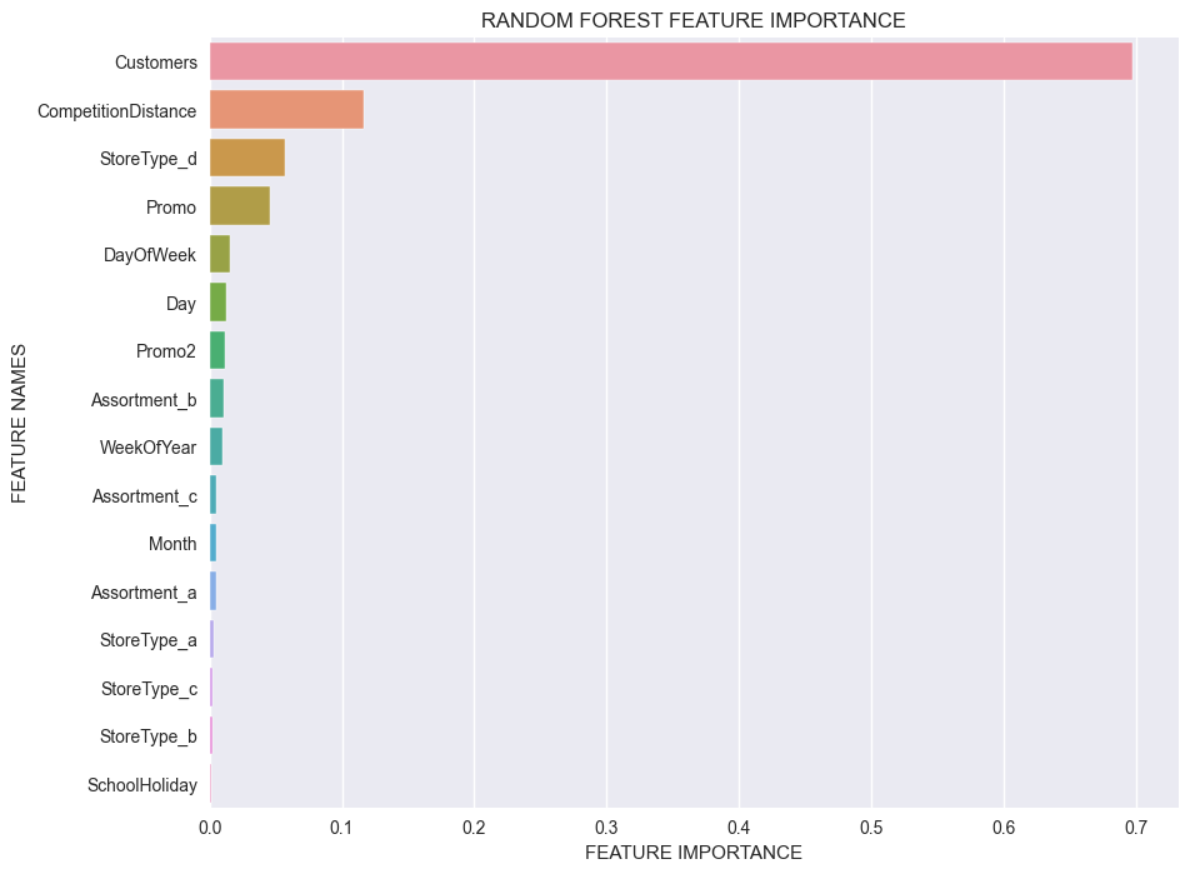


**Testing**









**Outcomes**

Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions and various other growth plans are affected by the revenue the company is going to make in the coming months and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good.

Some important conclusions drawn from the analysis are as follows:

* The positive effect of promotion on Customers and Sales is observable.
* there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.
* The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.
* Random Forest regressor achieved lowest MAPE as 5.65% showing that it is a highly accurate model. MAE is the average magnitude of error produced by your model, the MAPE is how far the model’s predictions are off from their corresponding outputs on average

