Capstone Project-2

ROSSMANN STORES SALES PREDICTION

RETAIL SALES PREDICTION

PREDICTING SALES OF A MAJOR STORE CHAIN ROSSMANN

Team Members

Palash Pathak

Yogesh Dubey

Let’s Predict The Sales

1. Defining Problem Statement
2. Data Wrangling
3. Exploratory Data Analysis
4. Feature Selection
5. Prepare Dataset For Modeling
6. Applying Model
7. Model Validation & Selection

Overview of the 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.
- Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.
- In order to accomplish this, we organized the whole series into four parts as follows :

1. Data Pre-processing
2. EDA
3. Feature Engineering
4. Model training

Data Summary (DataSet 1)

Stores Daily Sales Data : No of observations & no of features (1017209, 9)
9 Variables (3 numeric, 6 categorical)
Date: 2013/01/01 – 2015/07/31

Column Name	Unique_Count	Nan %	Data_type
Store	1115	0.0	Int64
DayOfWeek	7	0.0	Int64
Date	942	0.0	Object
Sales	21734	0.0	Int64
Customers	4086	0.0	Int64
Open	2	0.0	Int64
Promo	2	0.0	Int64
Stateholiday	5	0.0	Object
Schoolholiday	2	0.0	Int64

DataSet 2

- **Store :** No of observations & no of features (1115, 10)
  10 variables (3 numeric, 7 categorical)

Column Name	Unique_Count	Nan%	Data_type
Store	1115	0.0000	Int64
StoreType	4	0.0000	Object
Assorment	3	0.0000	Object
CompetitionDistance	654	0.27	Float64
CompetitionOpenSince Month	12	31.75	Float64
CompetitionOpenSince Year	23	31.75	Float64
Promo2	2	0.0000	Int64
Promo2SinceWeek	24	48.80	Float64
Prom2SinceYear	7	48.80	Float64
PromoInterval	3	48.80	Object

Data Pipeline

Data Preprocessing :

-Dealing with the missing values: CompetitionDistance , CompetitionAge, CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceYear, Promo2SinceWeek,PromoInterval

-Sanity check: We have dropped entries when sale is zero for a open store

EDA : In this part we've done some exploratory data analysis(EDA) on the various features to see the trend and analyzing.

Feature Engineering : In this process we are selecting, manipulating, and transforming raw data into features that can be used .In order to make machine learning work well

Model Building : Finally, we create models. We show how to start with a simple model, then slowly add complexity for better performance.

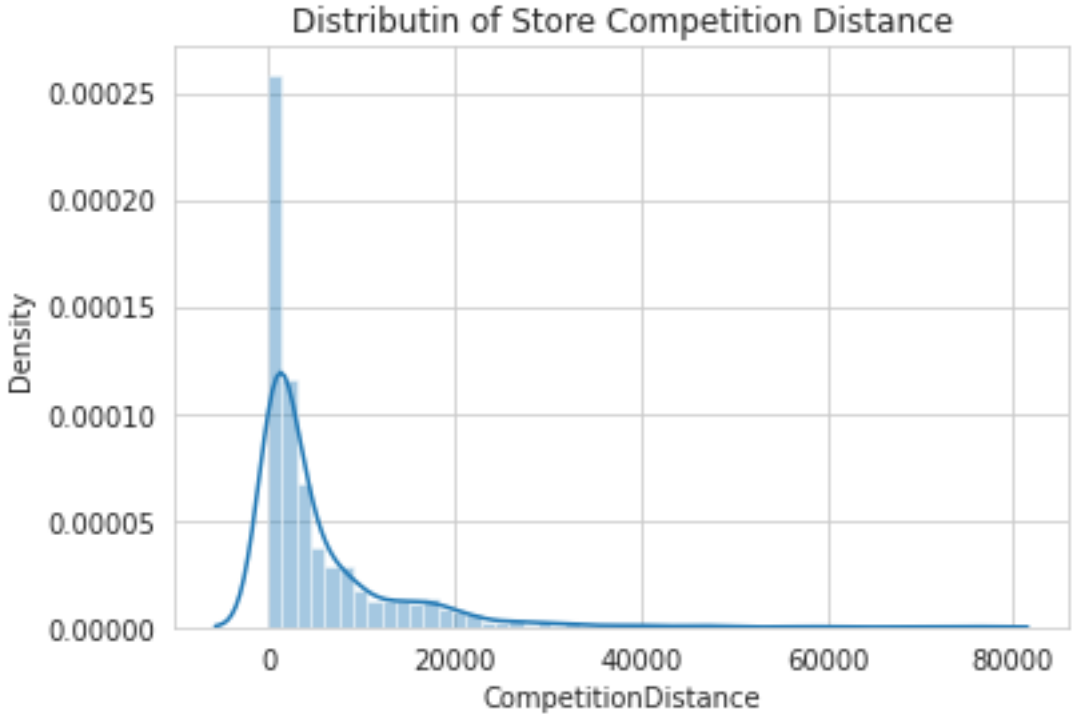
Some of the important fields that we need to understand

- 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 we 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 a school holiday: 0= no holiday, 1= holiday

- 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

Dealing With Missing Values - Competition Distance

CompetitionDistance is distance in meters to the nearest competitor store
Let's first have a look at its distribution



We have 3 null entries for Competition Distance. Replacing them with median would be a good choice

Other features With Apparent Missing Values

Competition Open Since Month/Year :

Both these features give info about opening time of the nearest competition to the particular store , so we added a calculated field stating age of competition in no of months since competition opening and we have replaced all nulls to zero .

Promo 2 Since Week/Year :

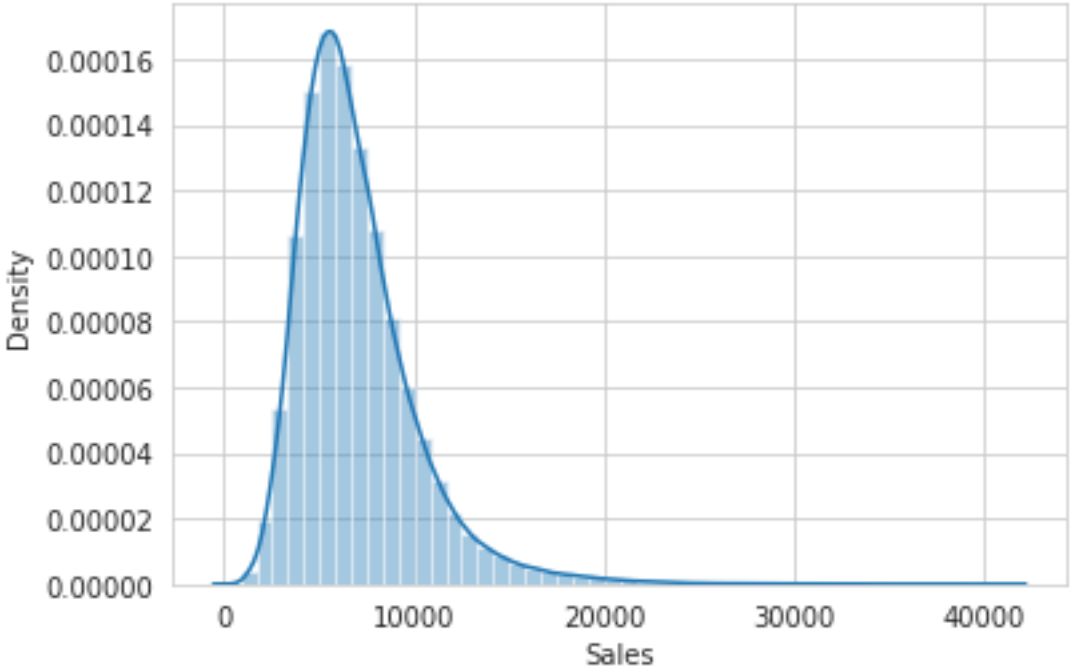
Both these features give info about the time since that store is participating in promo 2. So we have added a binary feature stating if store is currently running promo2 or not. All nulls in both these features refers to the stores that are not participating in promo2

Promo Interval:

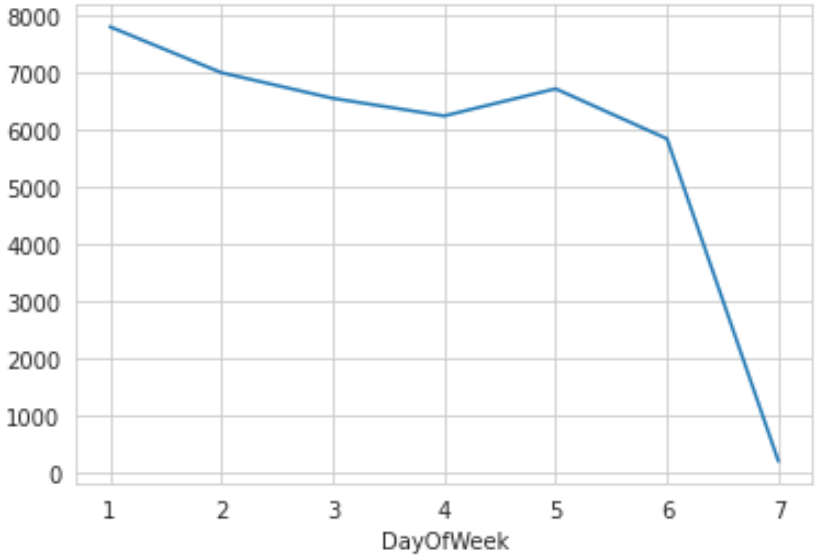
This variable has a list of months in which a new round of promo2 is started for the stores that are participating in it. So, we have again added a binary feature representing if a new round of promo2 is started in the current month or not. Here also the nulls are referring to the non-participating stores.

Dependent Variable : Sales

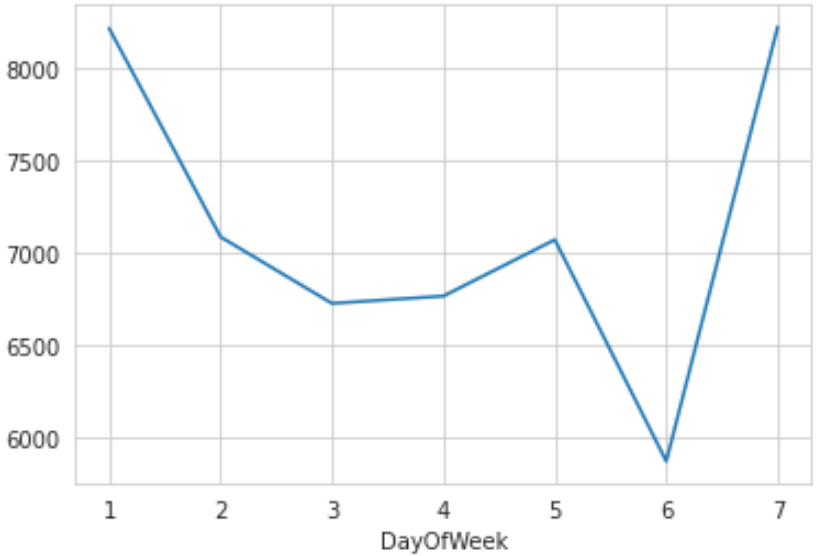
The turnover for any given day (this is what we are predicting)



There is positive skewness, so we should use transformation before training our model



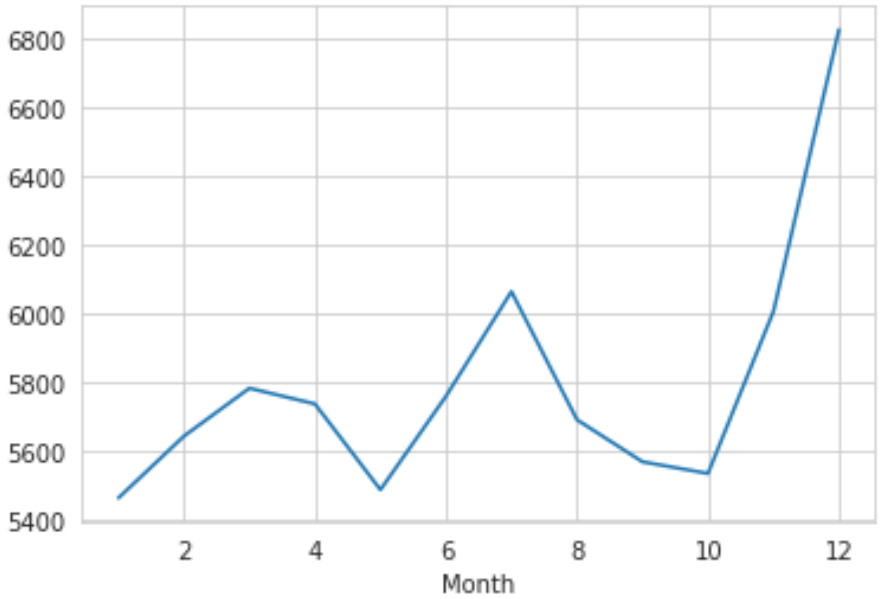
Week-days-wise average sales



Week-days-wise average sales only when store is open

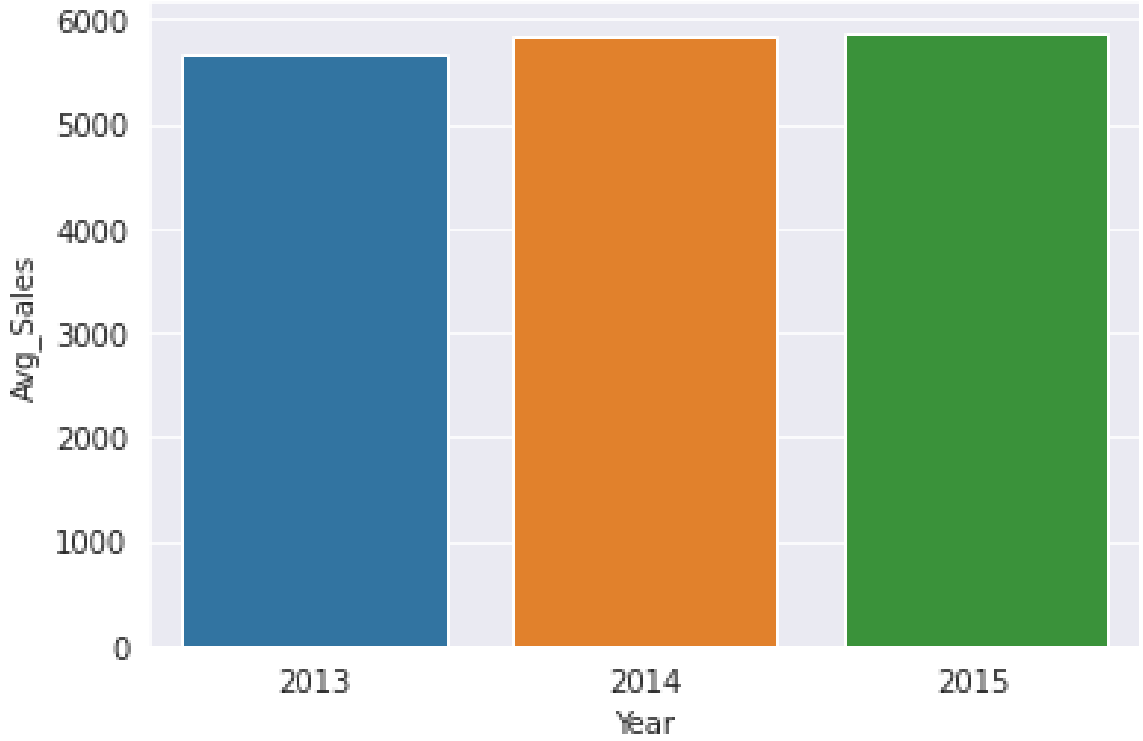
- **As it is evident from the entries day 1 is Monday.**
- **While stores are mostly closed on Sundays ,sales is highest when open.**
- **Apart from Sunday , sale is maximum on Mondays**

Monthly Average Sales
As we don’t have data for 2015 whole year, sum of sales wont be the correct statistic, lets take average.

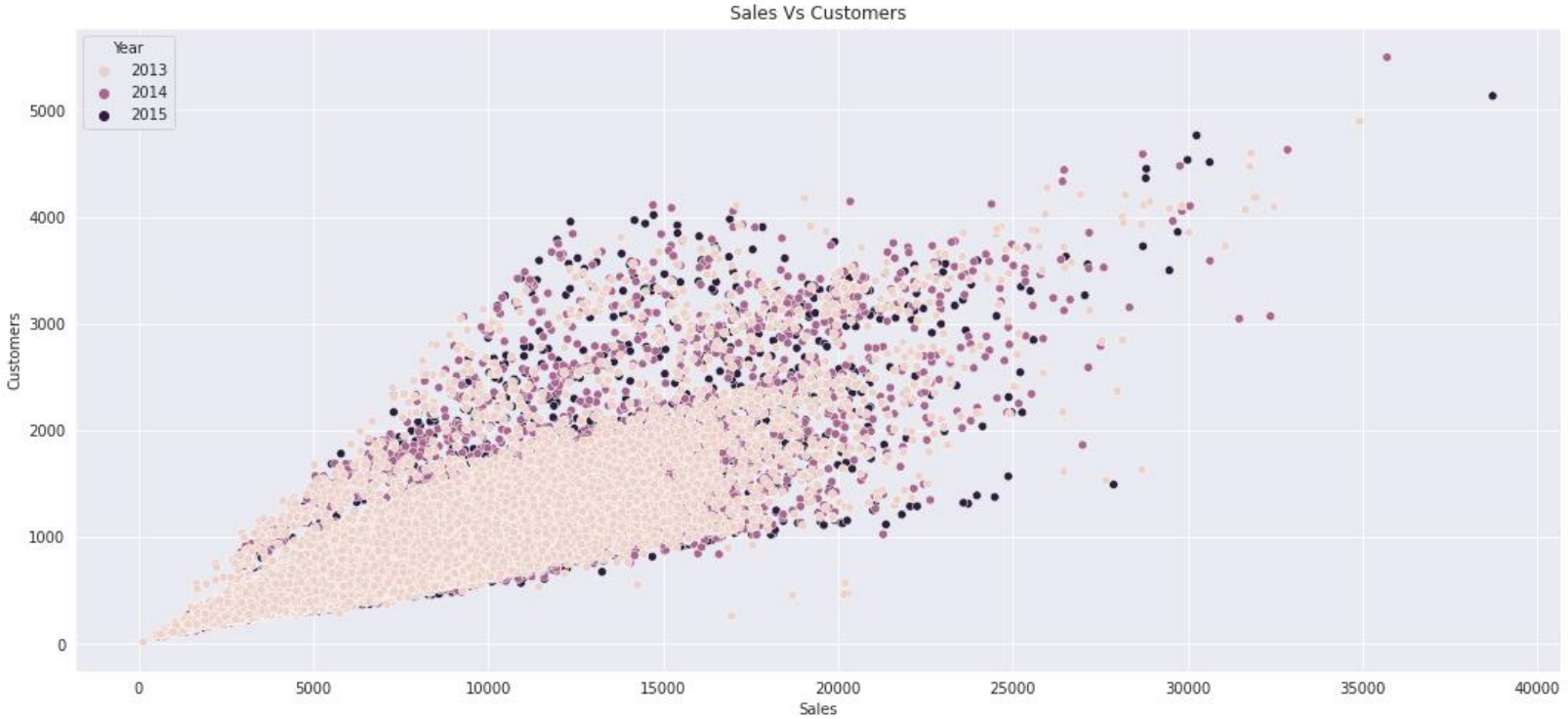


WE CAN CLEARLY SEE THAT
TOWARDS YEAR ENDING SALES ARE BETTER ON AN AVERAGE,

Year vs sales(average)



Sales is increasing with time



Sales vs customers

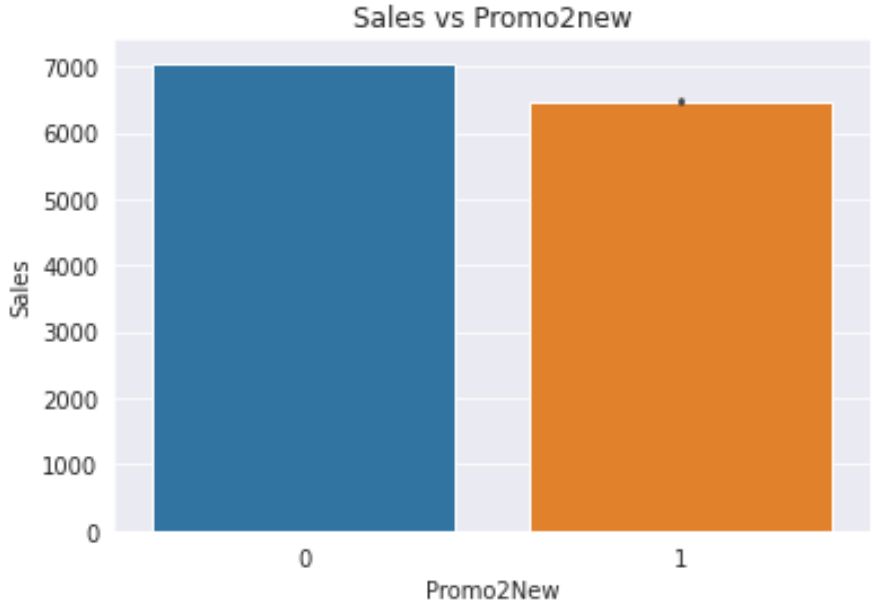
Number of stores running promo

Lets see how promo is impacting sales

WE CAN CLEARLY SEE THAT PROMO IS EFFECTIVE

Number of stores participating in promo2

More than 50% of stores are participating in Promo2

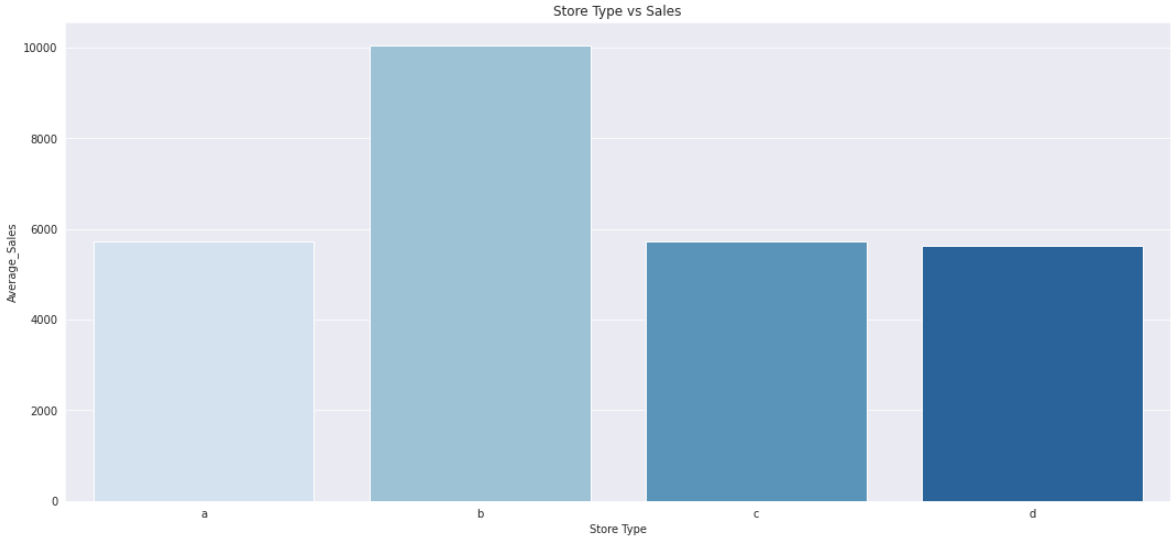


Promo 2 is ineffective

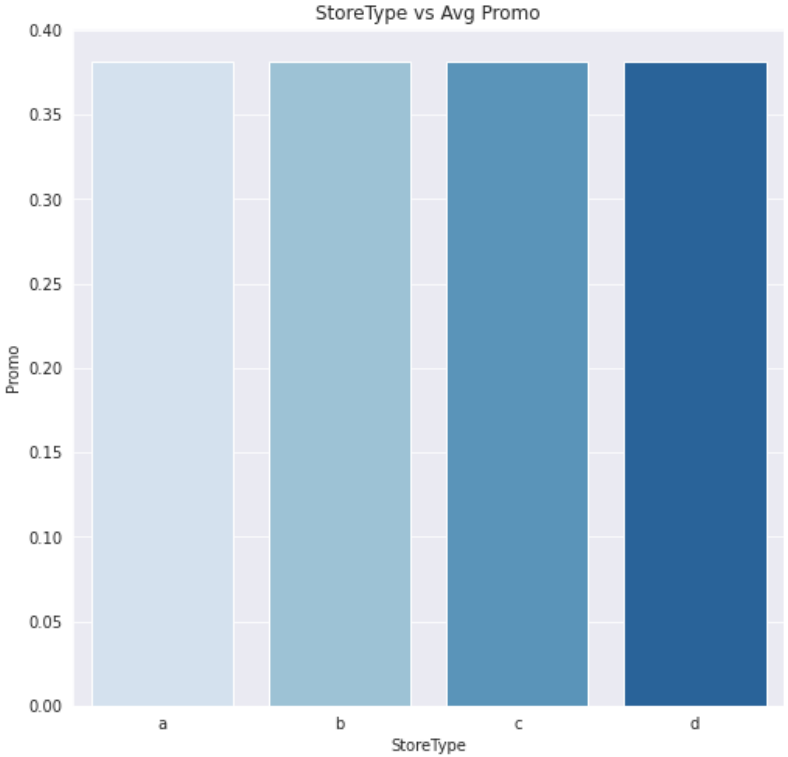
Lets look at the store type :

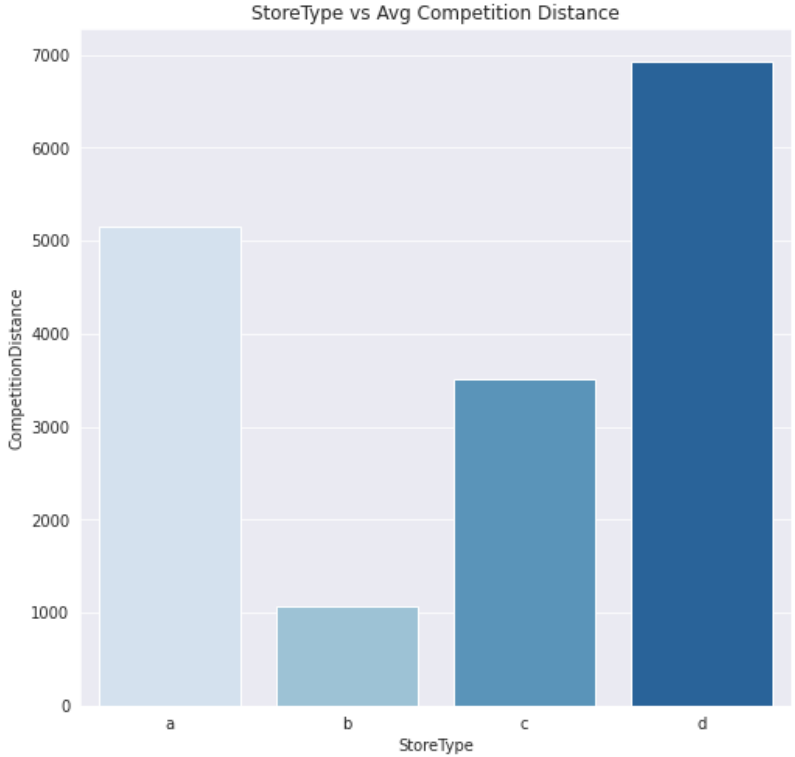
a : 602
d : 348
c : 148
b : 17

So, majority of stores are of 'a' type , followed by d and c while b type are the least



Store b are the least in numbers but on an average outperforms other store types.

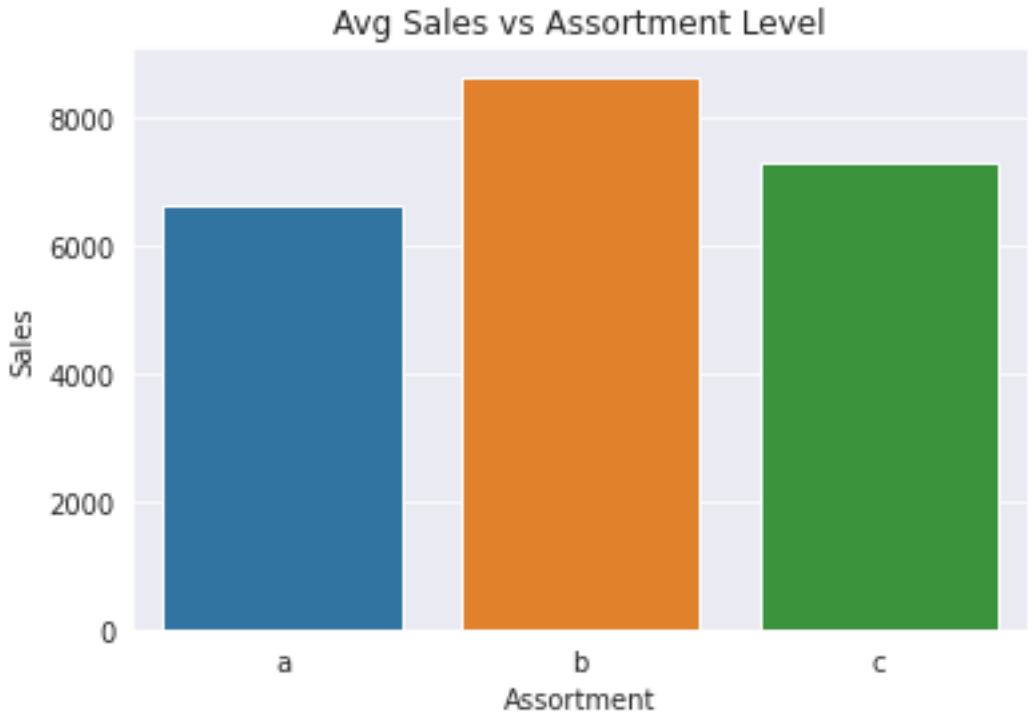




Store type b has the least Competition, that means distance from the nearest competition is the lowest on an average for store type b.

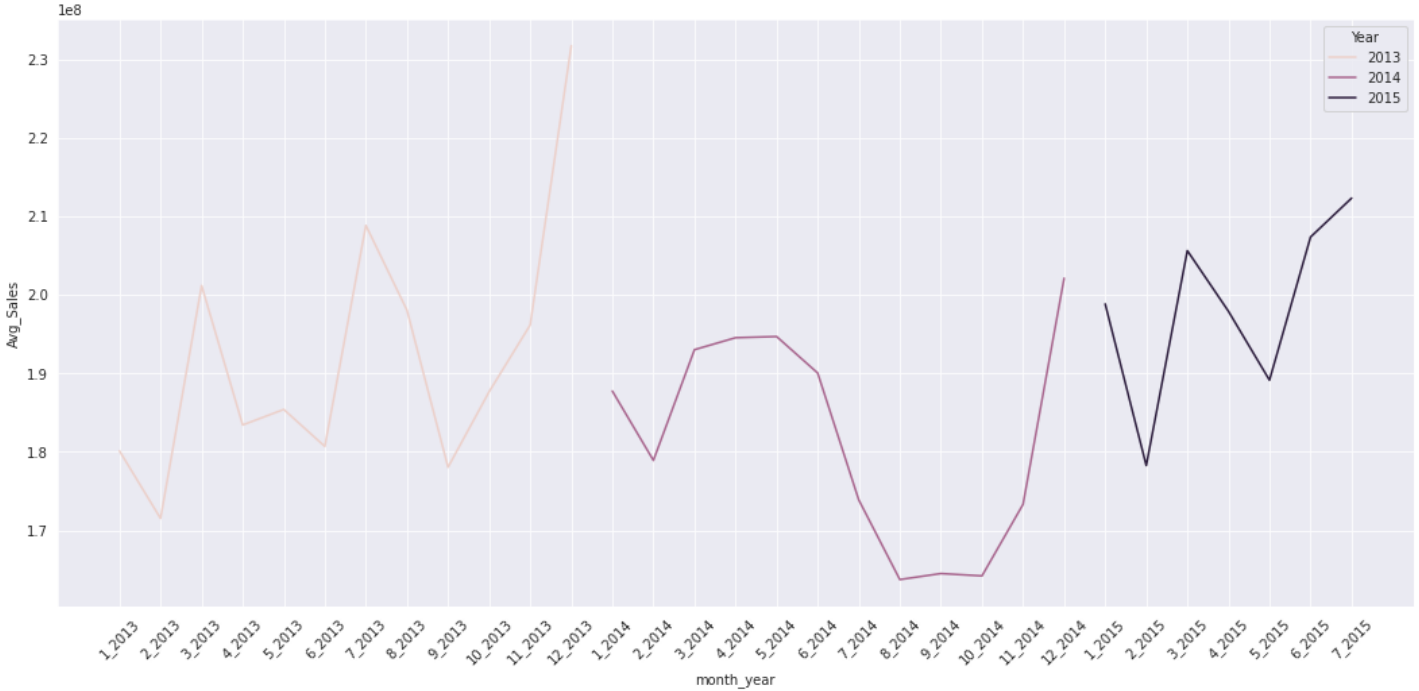
Assortment level of stores :

A : 593
c : 513
b : 9



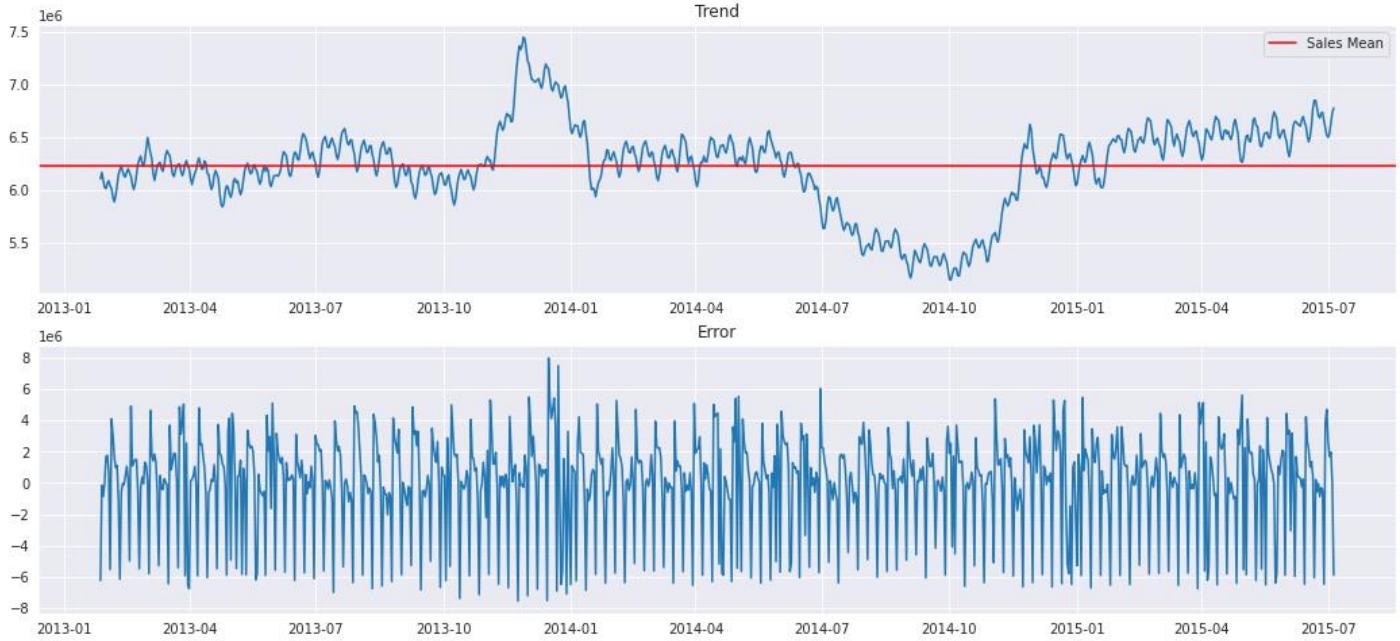
Assortment level indicates a = basic, b = extra, c = extended
So, extra level has highest average sales

Let’s talk about sales trend



We can clearly see that...
There is a increase in sales at year-ending as compared to previous months.
There is a dip in year 2014 mid-way

Let’s look at the trend and seasonal part



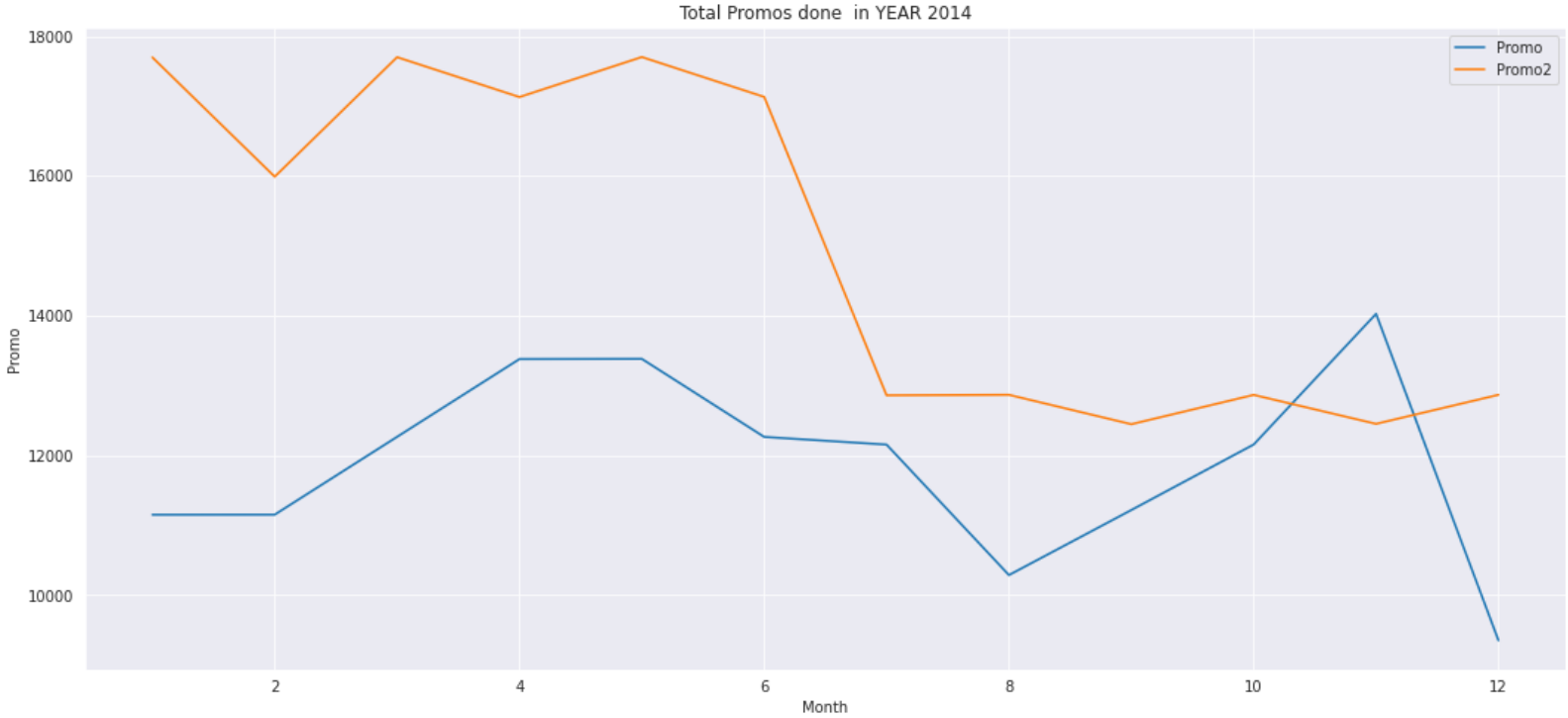
2015 seems to be a good year as the trend line is above the average trend line.
There is a increase in sales at year-ending as compared to previous months.
There is a dip in year 2014 from July-Oct
I wonder what derived that ?

1. For that first let's check the number of stores open :

Many of the stores where closed in 2nd half of 2014 as compared to 2013 & 2015 that contributed to dip in sales

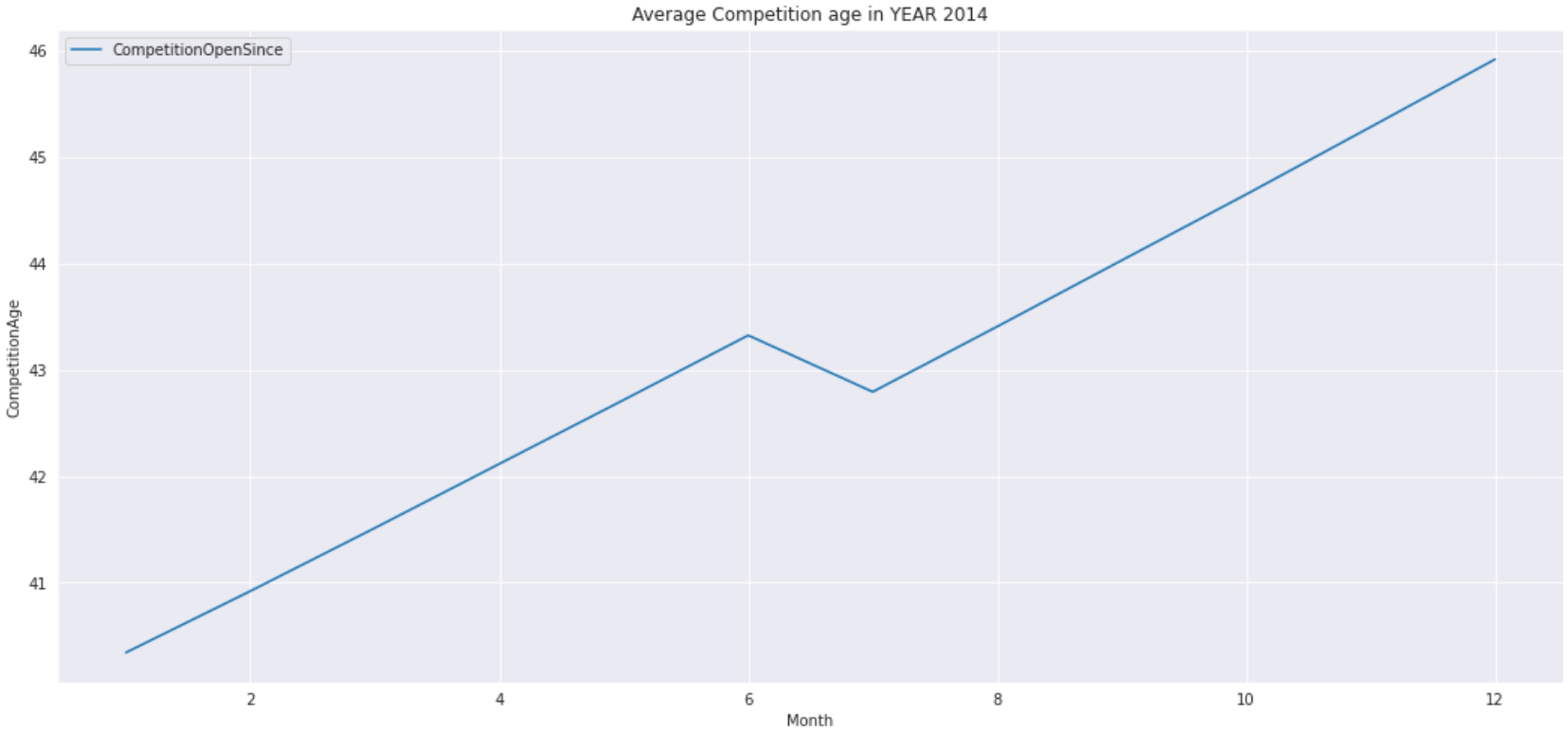
	Month	2013	2014	2015
1		28865	28707	28763.0
2		26682	26791	26766.0
3		27891	29005	29079.0
4		27878	26917	26931.0
5		26199	28021	25879.0
6		27939	26209	28423.0
7		30164	25224	30188.0
8		30023	24388	NaN
9		27980	24341	NaN
10		28990	24301	NaN
11		28412	22989	NaN
12		26901	23492	NaN

2. Promos done in year 2014 :



We can clearly see that stores have decreased there promos after June 2014

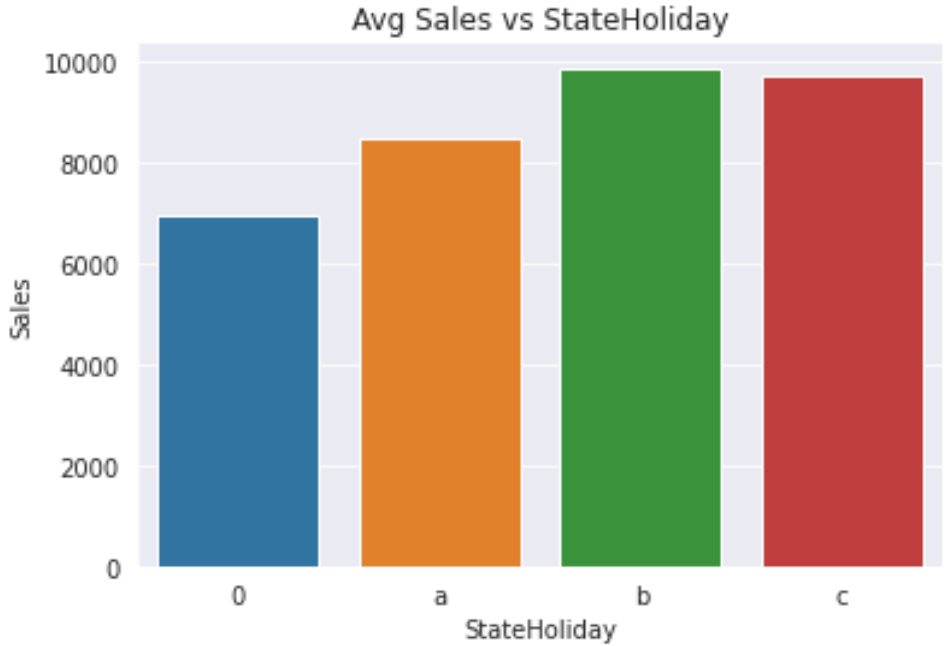
3. Lets check if there is any entry of new competition :



There is new entry in competition which may also have factored in the dip in sales

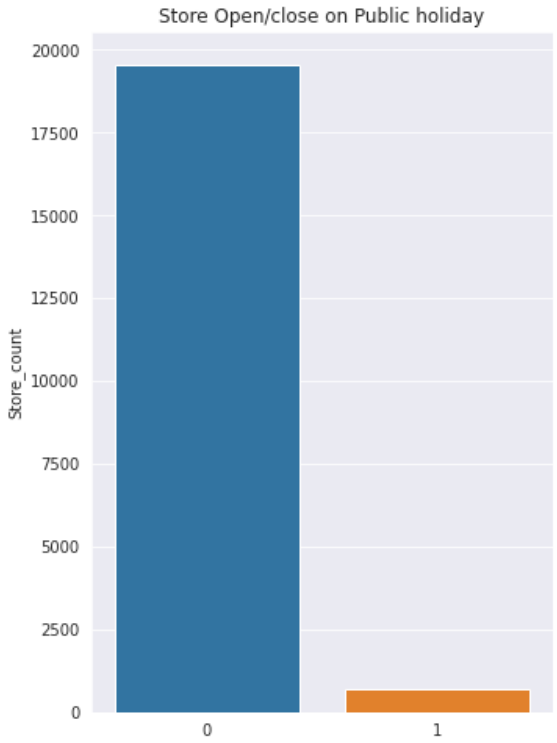
Stateholiday vs Sales

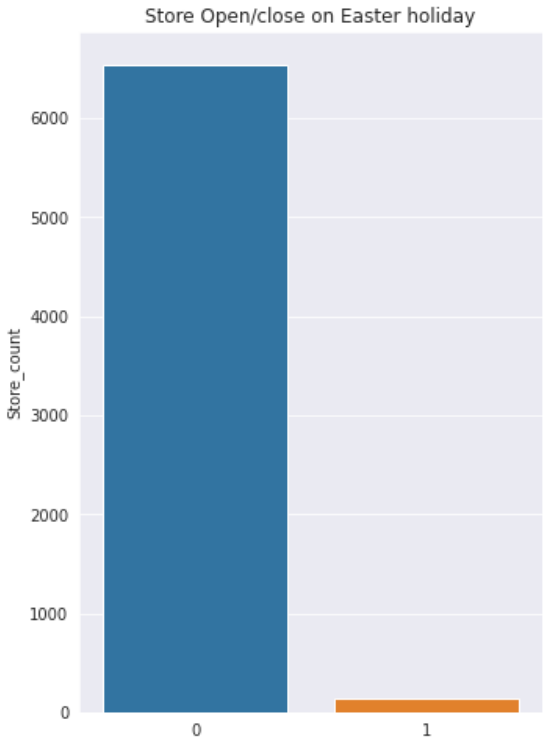
We already know that a = public holiday, b = Easter holiday, c = Christmas, 0 = None

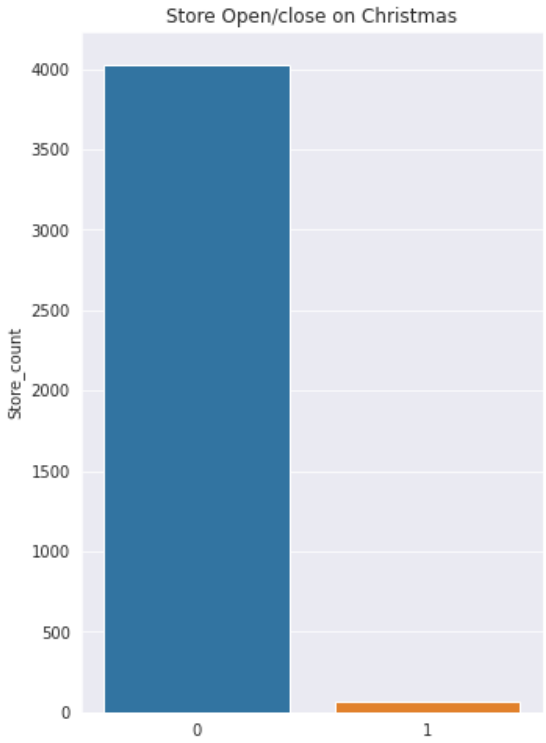


Sales are maximum on Easter holiday on an average followed by Christmas while on a regular day sales are low (on average)

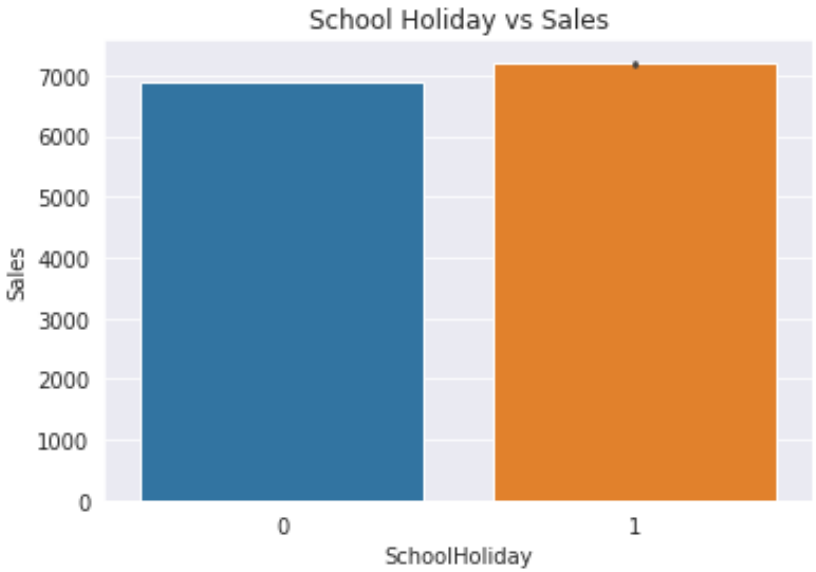
Lets check for all 3 holidays, stores are open/close



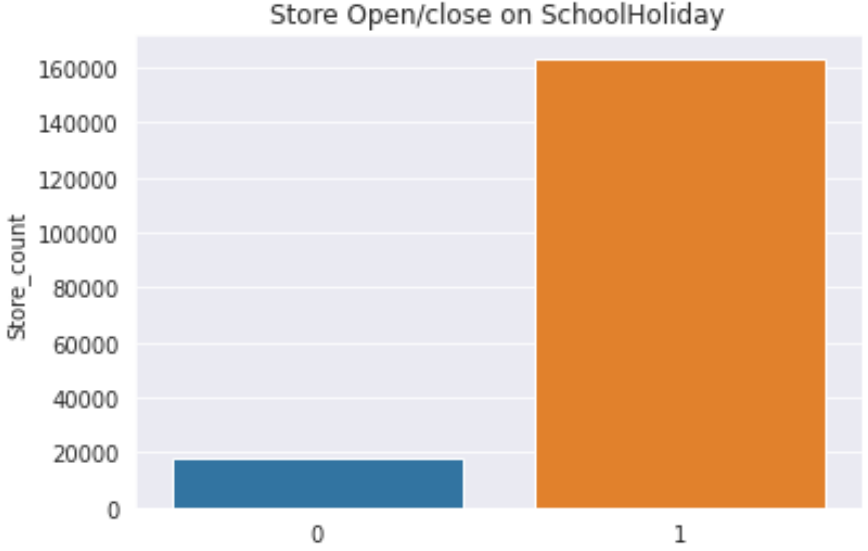




We can see that stores are mostly closed on all state holidays

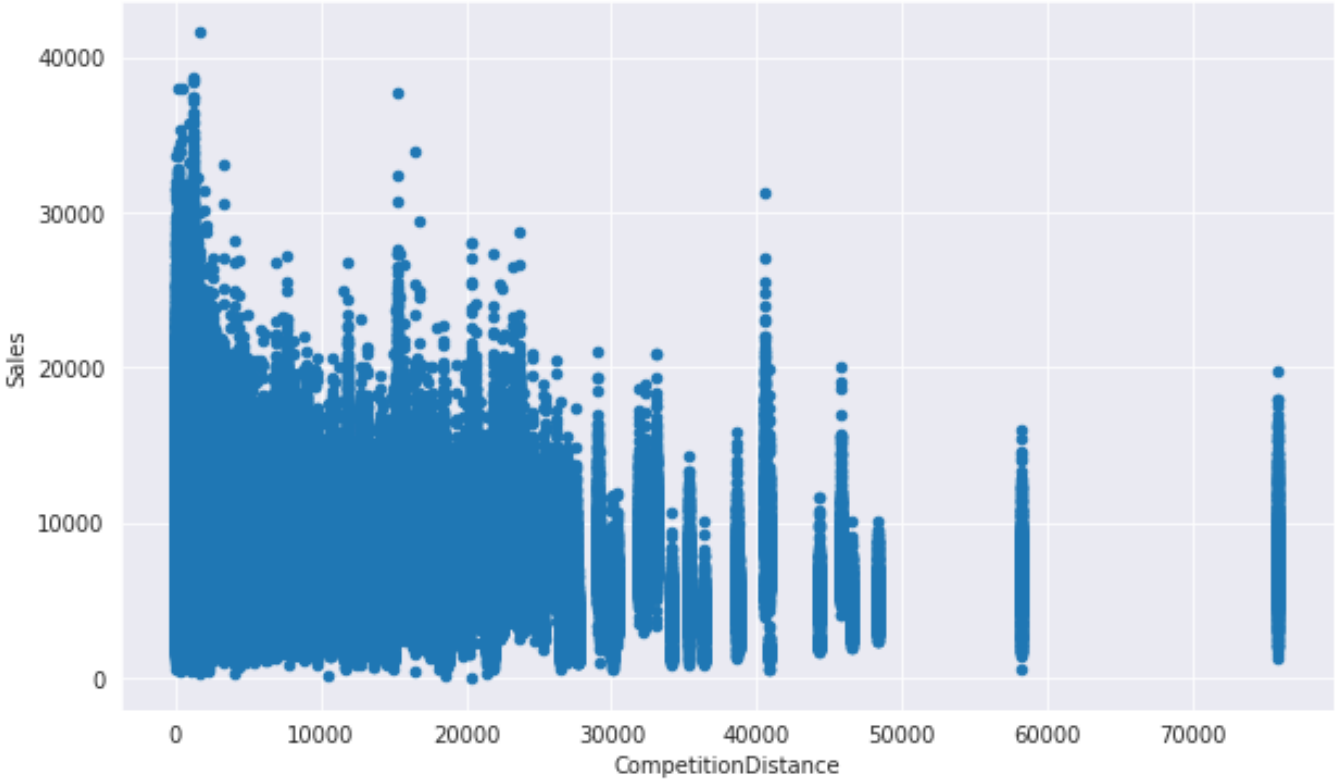


Sales are higher on school holidays



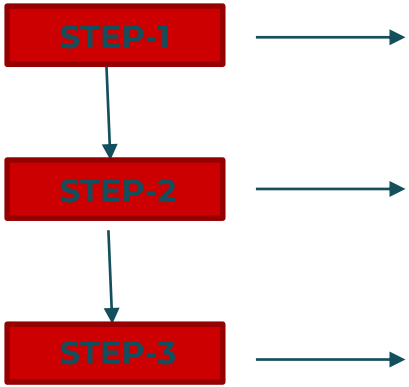
We can say that stores mostly remain open on school holidays.

Competition Distance



Rossman stores are performing well even when competition distance is low.....

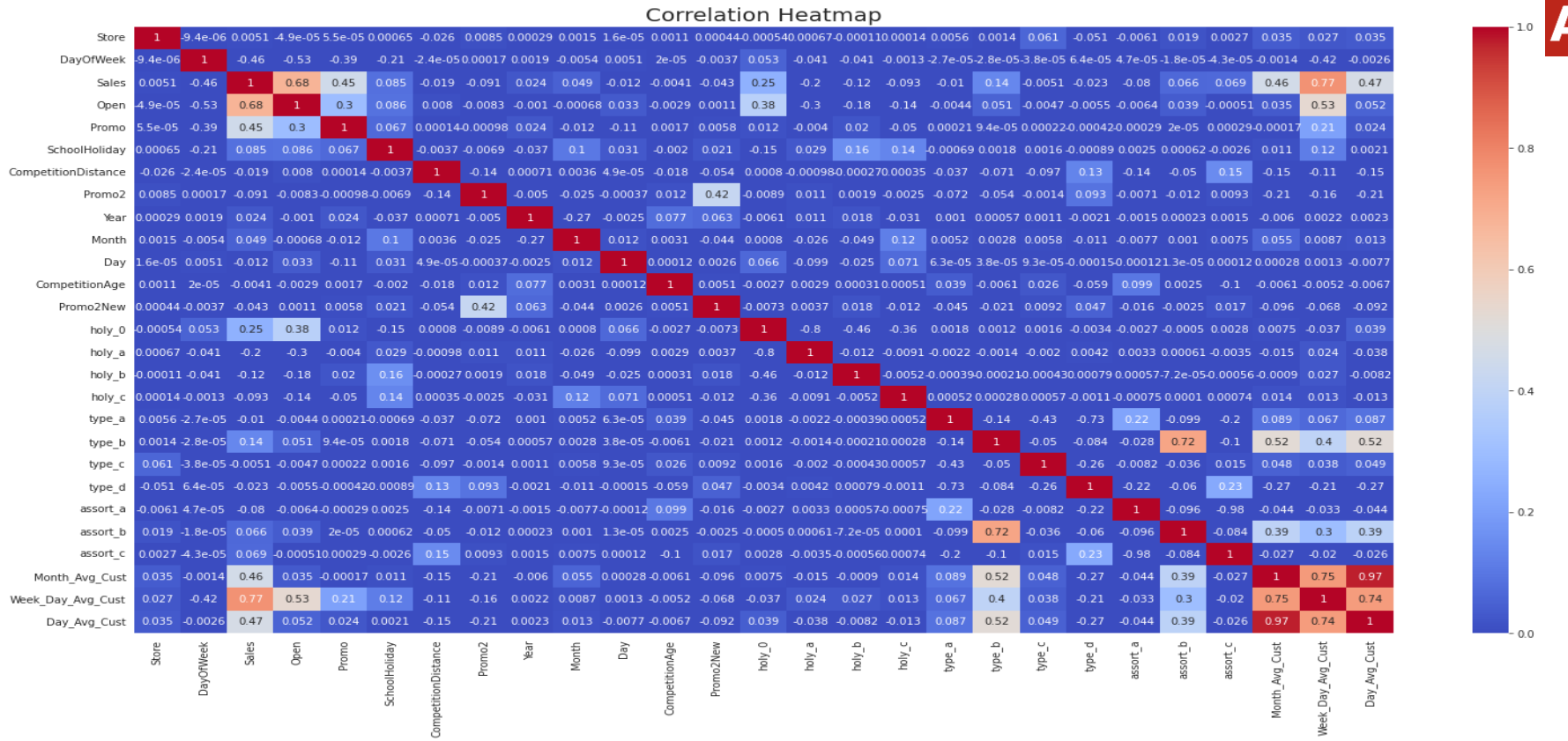
Feature Engineering :



One hot encoding for categorical features

Using average of customers in the past as a feature for future predictions.
(Number of Customers will not be available for predictions, so we cannot use it directly in our model)

Removing all features that are not useful



From the heatmap, we can briefly comment that important features apart from 'Open' are WeekDayAvgCust, DayAvgCust, MonthAvgCust, Promo, holy_0(no holiday), type_b(store_type)

Preparing Dataset for Modeling

Features engineering :
Calculated Features –
- Year, Month , Day from Date
- Competition Age Logt from Competition Open Since Year/Month , log transformed
- Promo2
- Promo2 new from promo interval
- Log transformed Competition distance
- Avg customers for each store monthly, weekly, daywise from store-wise daily customers data

Assumption :
We make an assumption that the knowledge of store being open/close is available beforehand or it is planned for all future days for which we are predicting Sales

Train-Test Split:
We will split data on time basis and keep last 6 weeks of data in the test set .
Last Date in our data : 2015-07-31
Date Before Six Weeks : 2015-06-19

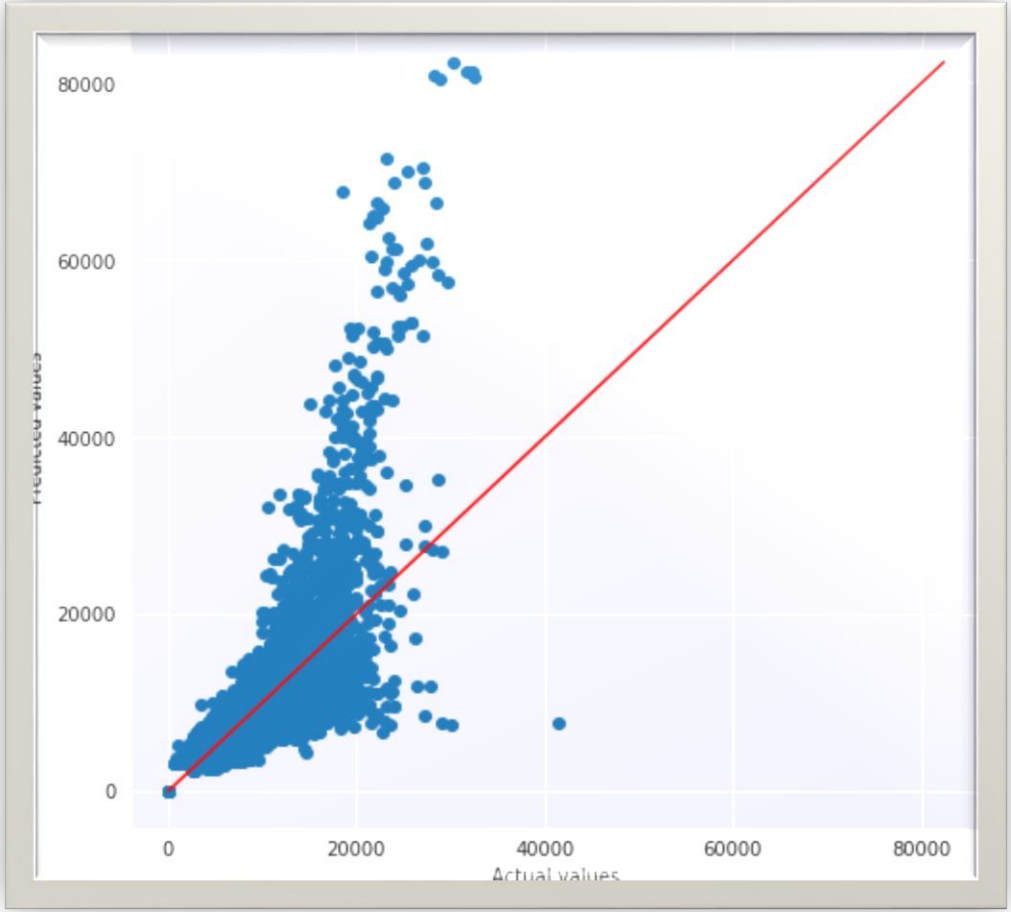
Train df : (970325, 31) from date 01/01/2013 to 19/06/2015
Test df : (46830, 31) from date 20/06/2015 to 31/07/2015

Applying Model

1.Linear Regression:

Linear Intercept - 3.15
Training RMSE : 2270.91
Testing RMSE : 2255.84
Training R2_Score : 0.6530
Testing R2_Score : 0.6334

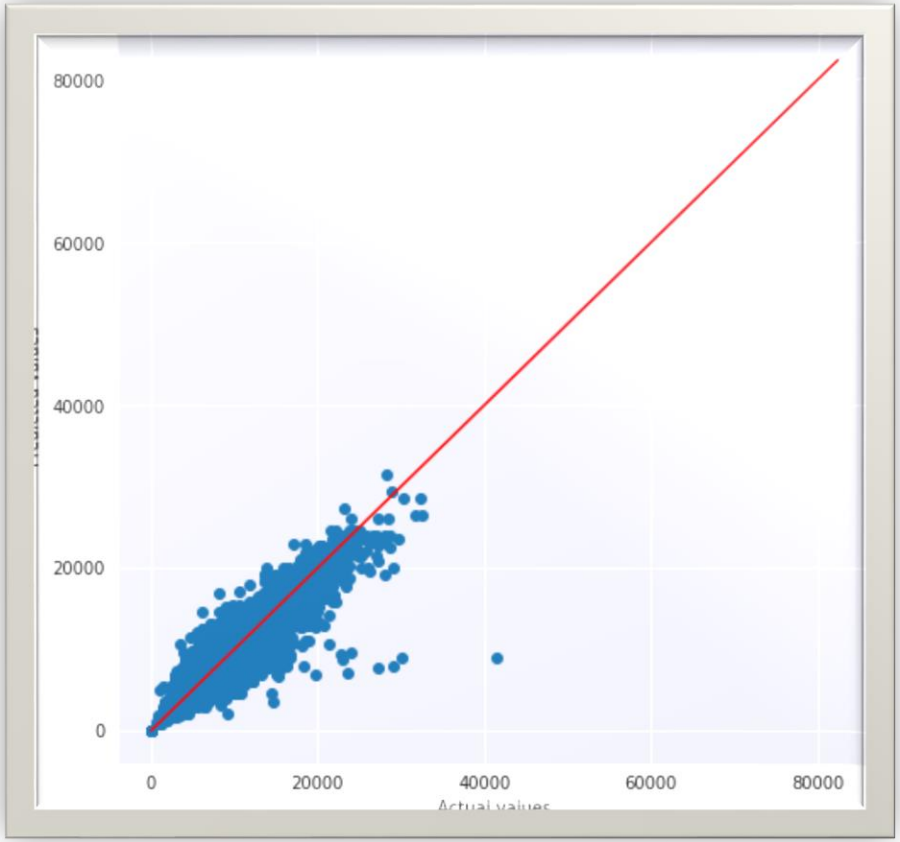
	Test	Pred
0	5263.0	6219.41
1	6064.0	6157.93
2	8314.0	8072.33
3	13995.0	11975.99
4	4822.0	6691.88



2.Decision Tree:

Training RMSE : 1362.15
Testing RMSE : 1386.06
Training R2_Score : 0.8752
Testing R2_Score : 0.8616

Test	Pred_lr	Pred_tree
5263.0	6219.41	5578.54
6064.0	6157.93	5825.74
8314.0	8072.33	8201.15
13995.0	11975.99	10457.56
4822.0	6691.88	5478.54

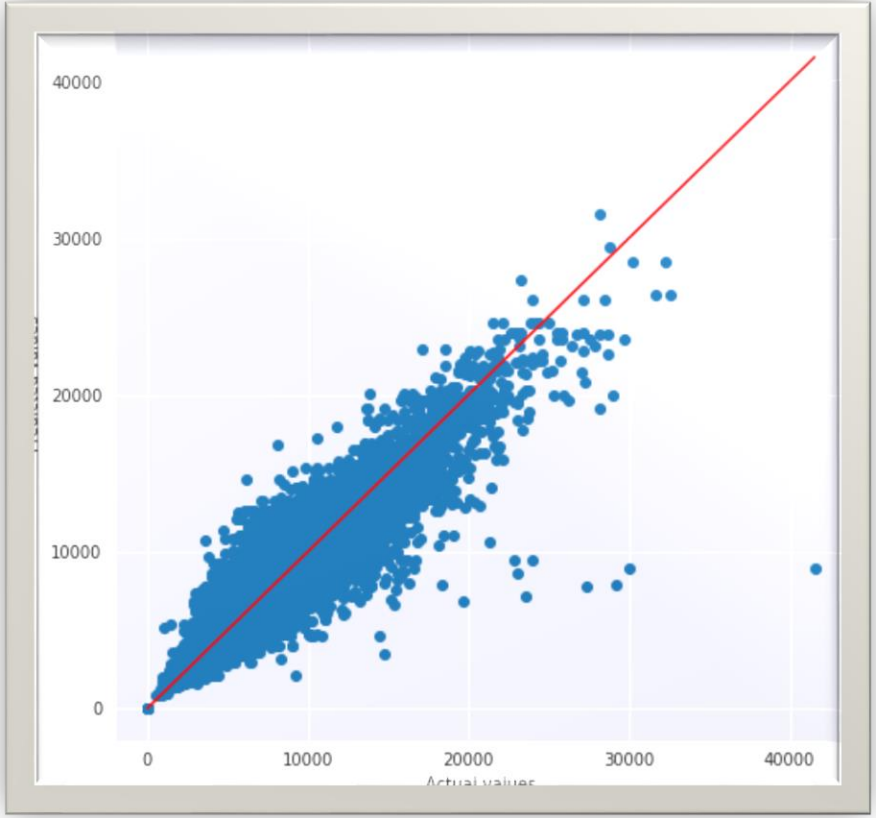


Using Cross-validation for Decision Trees:

The best fit h-par values are found to be :
max_depth: 30,
min_samples_leaf: 15,
squared_error is: -0.0045

Training RMSE : 788.81
Testing RMSE : 1089.20
Training R2_Score : 0.9581
Testing R2_Score : 0.9145

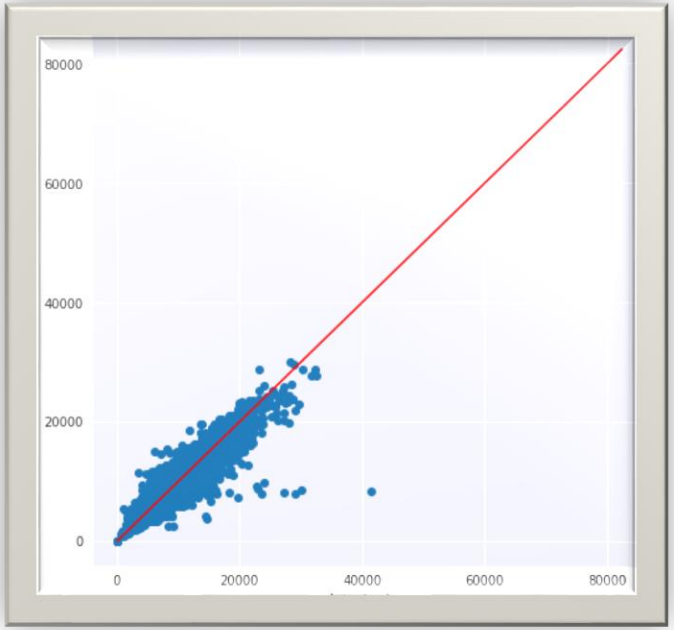
	Test	Pred_lr	Pred_tree	Pred_tree_bestfit
0	5263.0	6219.416931	5478.541440	5548.929442
1	6064.0	6157.927925	5825.714746	6083.961992
2	8314.0	8072.331547	8201.152917	9069.143320
3	13995.0	11975.999452	10457.561812	12918.809436
4	4822.0	6691.881090	5478.541440	6468.042416

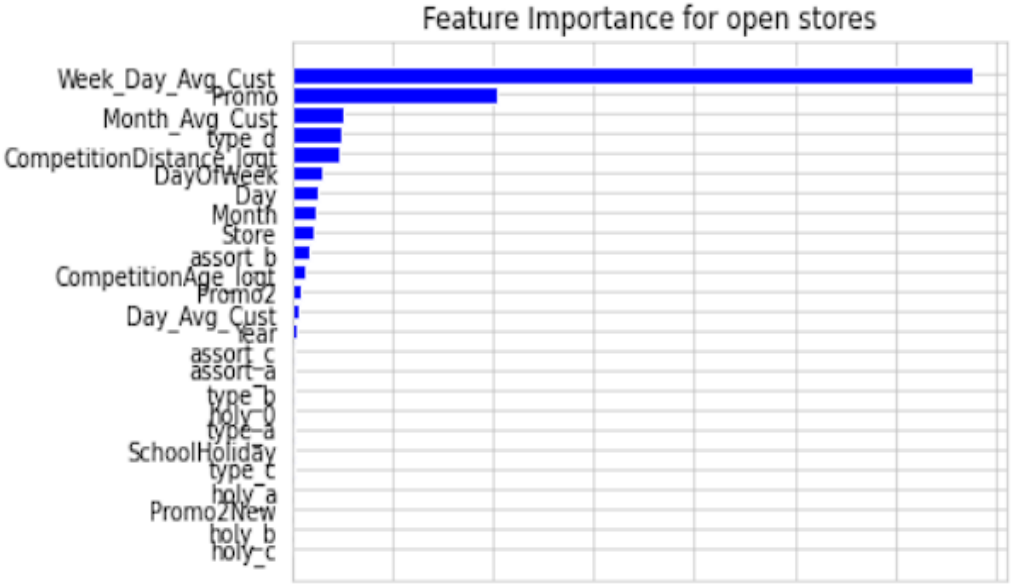


Random Forest :

Training RMSE : 950.85
Testing RMSE : 1090.26
Training R2_Score : 0.9392
Testing R2_Score : 0.9144

	Test	Pred_lr	Pred_tree	Pred_tree_bestfit	Pred_rf
0	5263.0	6219.416931	5478.541440	5548.929442	5716.287226
1	6064.0	6157.927925	5825.714746	6083.961992	6357.033939
2	8314.0	8072.331547	8201.152917	9069.143320	9374.178094
3	13995.0	11975.999452	10457.561812	12918.809436	12363.325058
4	4822.0	6691.881090	5478.541440	6468.042416	5678.621706



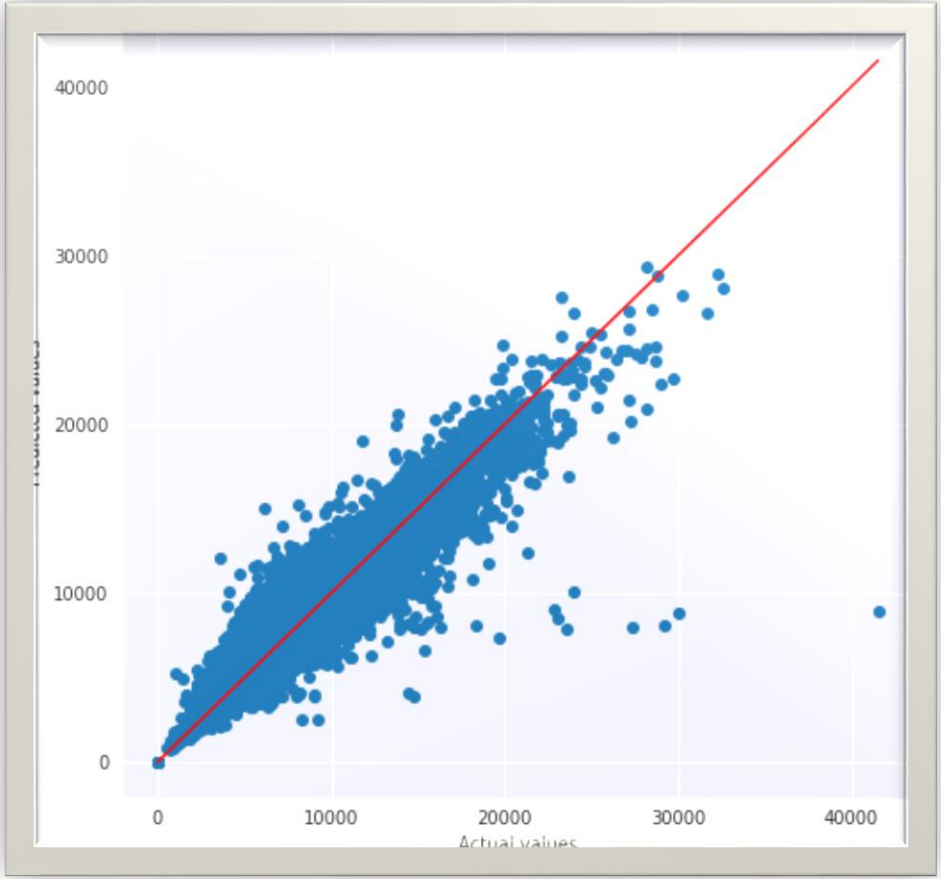


Using Cross Validation for RF:

The best fit h-par values are found to be :
max_depth: 40,
min_samples_leaf: 5,
n_estimators: 40
neg_mean_squared_error is: -0.00337

Training RMSE : 549.42
Testing RMSE : 935.72
Training R2_Score : 0.9833
Testing R2_Score : 0.9369

	Test	Pred_lr	Pred_tree	Pred_tree_bestfit	Pred_rf	Pred_rf_bestfit
0	5263.0	6219.416931	5478.541440	5548.929442	5716.287226	5497.442047
1	6064.0	6157.927925	5825.714746	6083.961992	6357.033939	6407.902044
2	8314.0	8072.331547	8201.152917	9069.143320	9374.178094	9672.962703
3	13995.0	11975.999452	10457.561812	12918.809436	12363.325058	12554.203832
4	4822.0	6691.881090	5478.541440	6468.042416	5678.621706	6120.861395



Model Selection :

	Test	Pred_lr	Pred_tree	Pred_tree_bestfit	Pred_rf	Pred_rf_bestfit
0	5263.0	6219.416931	5478.541440	5548.929442	5716.287226	5497.442047
1	6064.0	6157.927925	5825.714746	6083.961992	6357.033939	6407.902044
2	8314.0	8072.331547	8201.152917	9069.143320	9374.178094	9672.962703
3	13995.0	11975.999452	10457.561812	12918.809436	12363.325058	12554.203832
4	4822.0	6691.881090	5478.541440	6468.042416	5678.621706	6120.861395
...	...	...	...	...	...	...
46825	2738.0	3717.194441	2554.927834	2736.834303	2886.550091	2811.304912
46826	8528.0	5703.875450	6463.323434	7779.561468	7755.107326	8029.321985
46827	5431.0	5314.086421	6010.067634	5195.600987	5832.029980	5350.125382
46828	22183.0	64962.519104	21604.434291	21441.859247	21829.031139	21190.070165
46829	7824.0	5260.642266	5345.574956	7440.580931	5855.624851	7459.042390

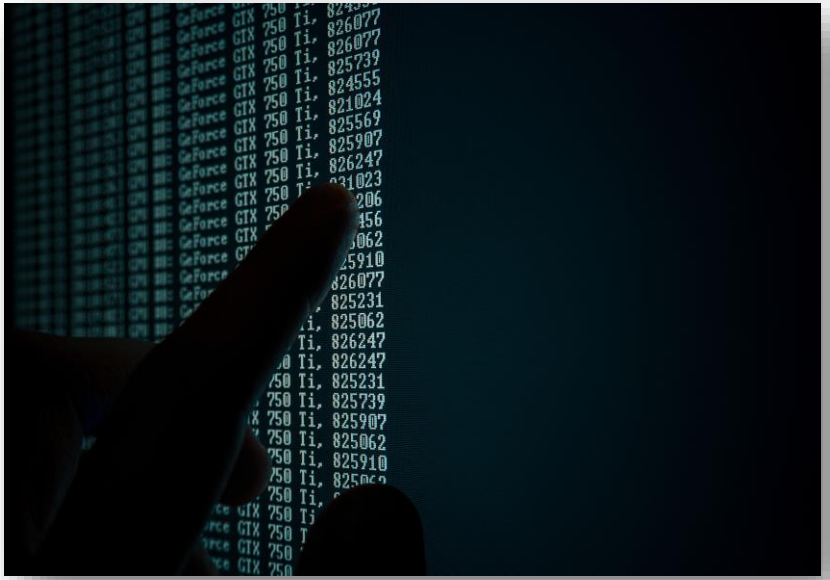
46830 rows × 6 columns

Comparing error and R2 score :

	Train-RMSE	Test-RMSE	Train-R2Score	Test-R2Score
LR	2270.905562	2255.843353	0.653042	0.633471
DTree	788.814612	1089.204086	0.958137	0.914551
RF	594.423068	935.727603	0.976228	0.936935

Zero sales is predicted correctly :

	Test	Pred_lr	Pred_tree	Pred_tree_bestfit	Pred_rf	Pred_rf_bestfit
291	0.0	1.288485	0.0	0.0	0.0	0.0
875	0.0	2.127500	0.0	0.0	0.0	0.0
1406	0.0	1.347775	0.0	0.0	0.0	0.0
1990	0.0	2.835486	0.0	0.0	0.0	0.0
2521	0.0	1.330680	0.0	0.0	0.0	0.0



Observations 1- From the above table we can see that linear regression is not fitting well on our data and it is not able to seperate the zero sales values i.e when store is closed.

Observation 2- Decision tree is working quite well and after tuning we got a score of 91.4% on test set which is quite good.

Observation 3- Random Forest is working exceptionally well, after tuning , we tried to keep a check on overfitting and optimised the test set score around 93.6%

Observation 4- From the above observations we have come to conclusion that we would choose Random Forest as our model.

<u>Assumptions:</u>
Here, we have assumed that store Open/Close information is available/planned in advance for the days on which we are predicting Sales.

Conclusion

- Apart from 'Open' , important feature are WeekDayAvgCust, Promo, MonthAvgCust, type_d (store_type),competition distance.
- Apart from Sunday when max stores are closed , sale is maximum on Mondays
- Towards year ending sales are better on an average
- Promo is Effective whereas Promo2 is ineffective
- There is a dip in sales in year 2014 from July-Oct because many of the stores were closed in 2nd half of 2014 as compared to 2013 & 2015 , stores have decreased their promo participation after June 2014 and there are new entry in competition which may also have factored in the dip in sales
- Stores are generally closed on state holidays and open on school holidays.
- Sales are maximum on Easter holiday compared to other state holidays.
- Rossman stores are performing well even when competition distance is low
- Random Forest is working exceptionally well, after tuning, we tried to keep a check on overfitting and optimised the train/test set score around 97/93%

Thank You