Capstone Project – 2

Supervised ML - Regression

Rossmann Sales Prediction

By

Team Alma Phoenix

Azhar Ali

Mohd Taufique

Pushpam Raghuvanshi

Content

1. Problem Statement
2. Data Summary
3. Exploratory Data Analysis
4. Feature Engineering
5. Machine Learning Model Building
6. Challenges Faced
7. Conclusions
8. Recommendations

Problem Statement

-Rossmann operates over 3000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for upto six weeks in advance.

-Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

-We are provided with historical sales data for 1,115 Rossmann stores.

-The task is to build Machine Learning model to forecast the "Sales" column for the test set.

DATA SUMMARY

Details of Datasets Provided

S. no.	Dataset	Variables	No. of Variables	No. of observations
1	Rossmann Stores Data - gives historical data including Sales	store, day of week, date, sales,customers, open, promo, state holiday, school holiday	9	1017209
2	store - supplement information about the stores	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10	1115

Important Variables

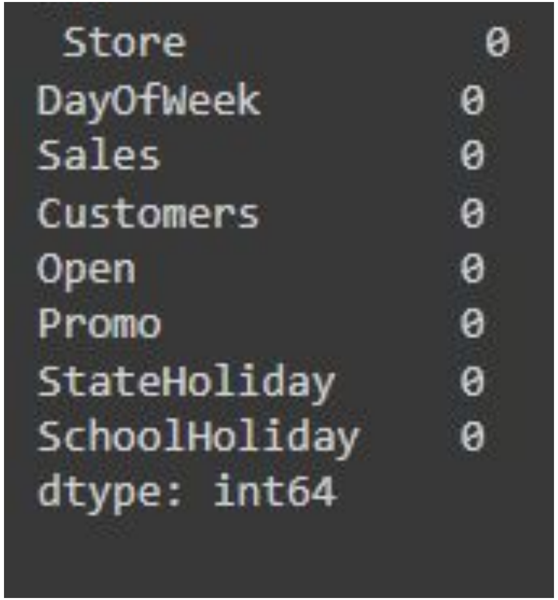
1. **Sales -** The turnover for any given day (this is what we are predicting).
2. **Open -** An indicator for whether the store was open or closed. 0 = closed, 1 = open.
3. **Store type -** Differentiates between 4 different store models (a, b, c & d).
4. **Assortment -** Describes an assortment level: 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.

Important Variables(cont.)

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.

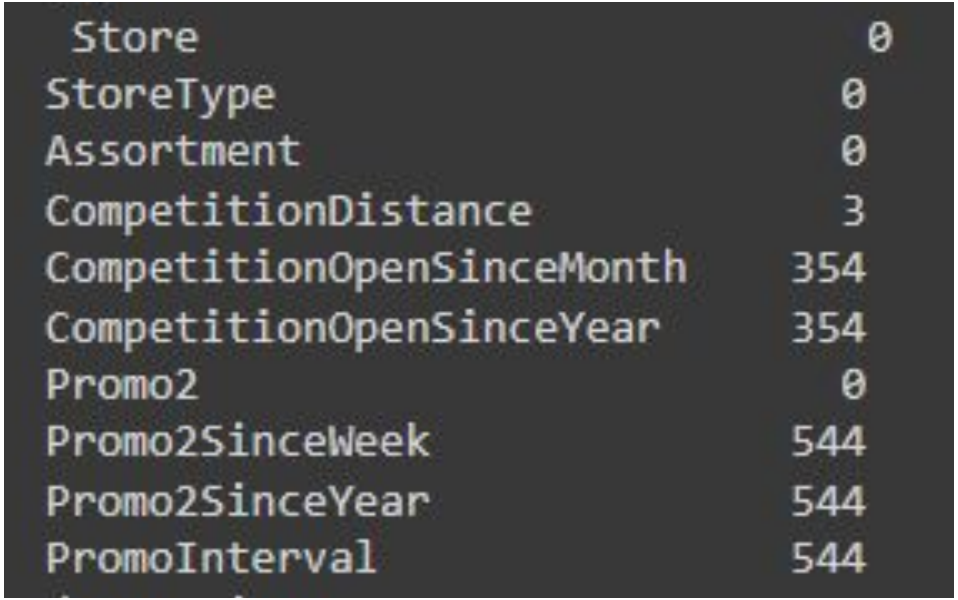
Missing Values

- Number of Missing values present in both the Datasets provided are :



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

Rossmann Stores Data.csv



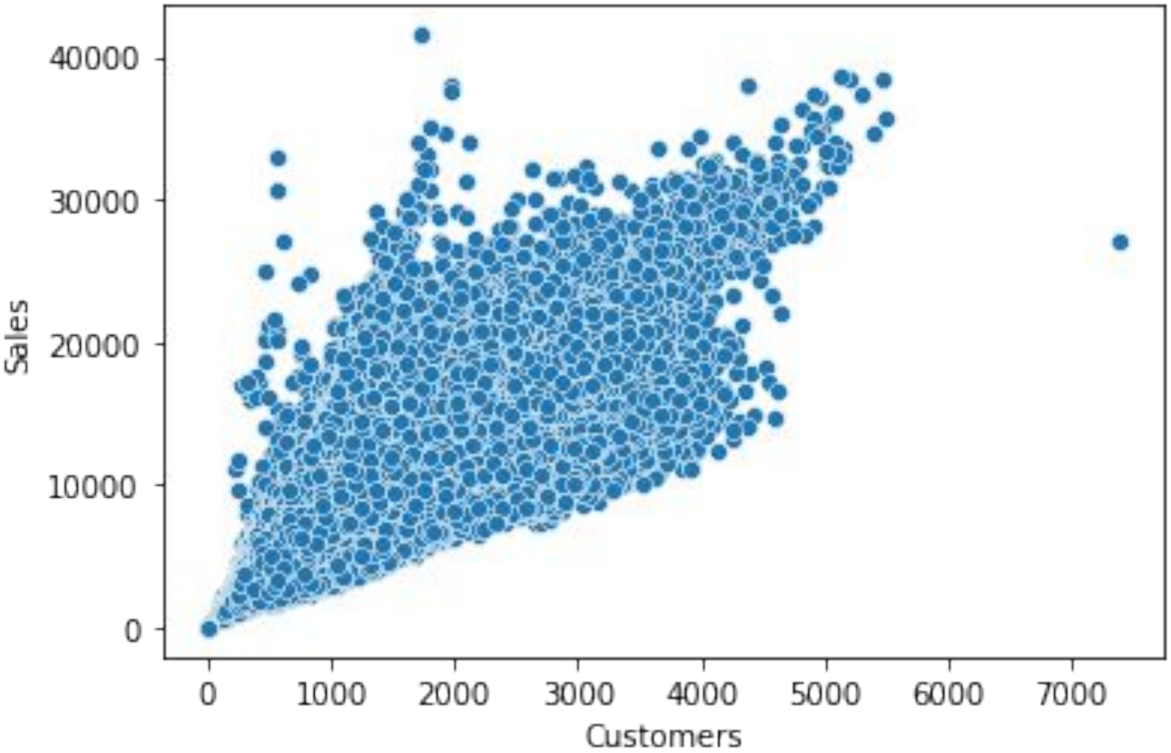
```
 Store                          0
StoreType                      0
Assortment                     0
CompetitionDistance            3
CompetitionOpenSinceMonth    354
CompetitionOpenSinceYear     354
Promo2                         0
Promo2SinceWeek              544
Promo2SinceYear              544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

store.csv

EXPLORATORY DATA ANALYSIS

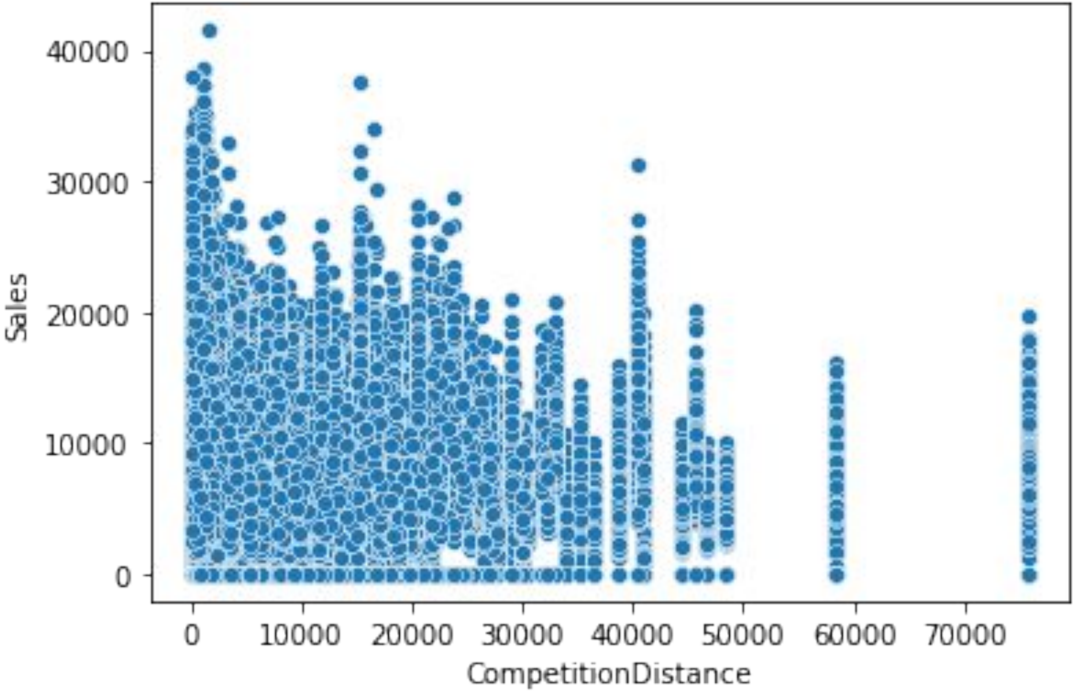
Scatterplot - Sales and Customer

- Scatterplot shows that relation between Sales and customer is sort of linear. Sales is increasing with the number of Customers increasing which is pretty obvious.

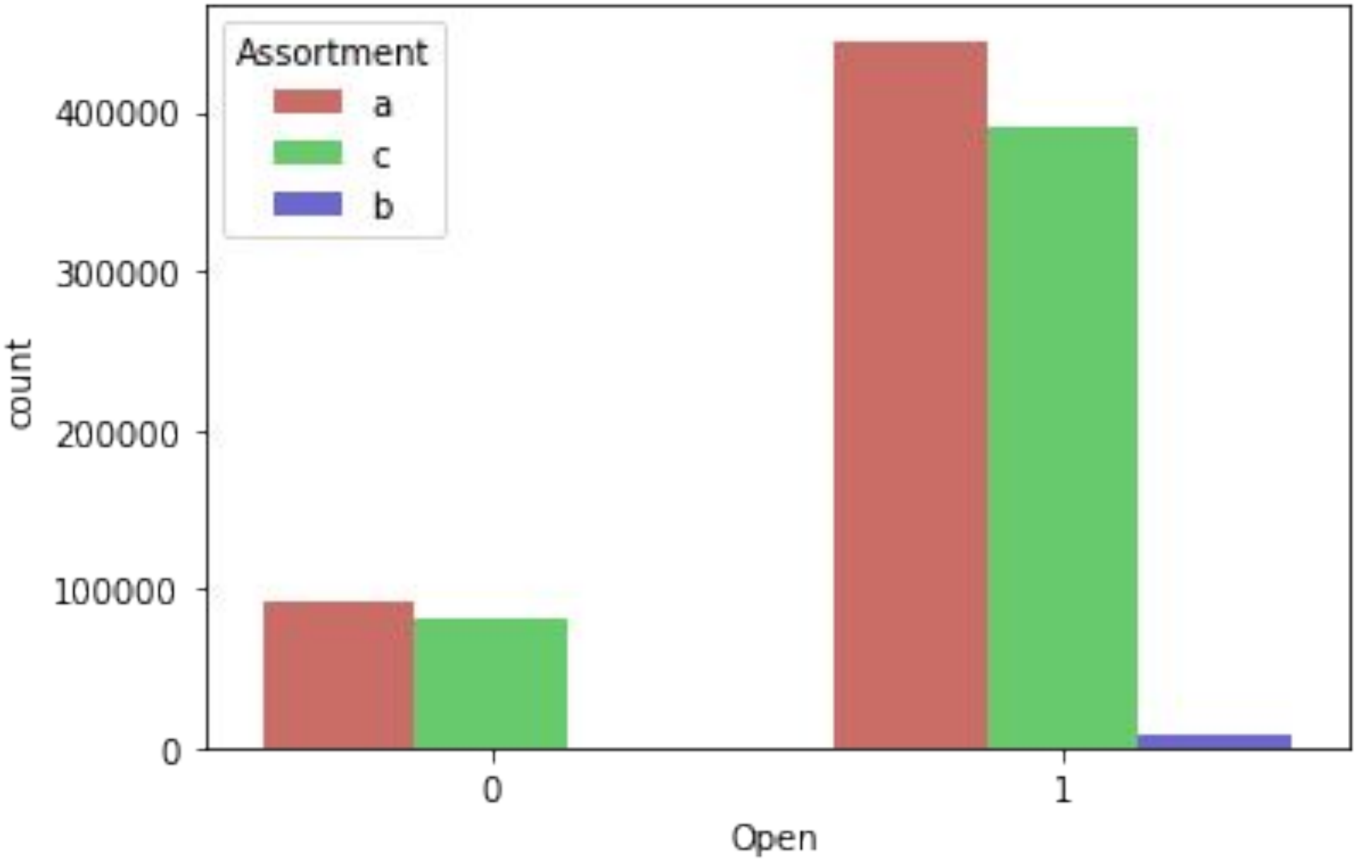


Scatterplot - Sales and Competition Distance

- It can be observed that mostly the competitor stores weren't that far from each other and the stores densely located near each other saw more sales.



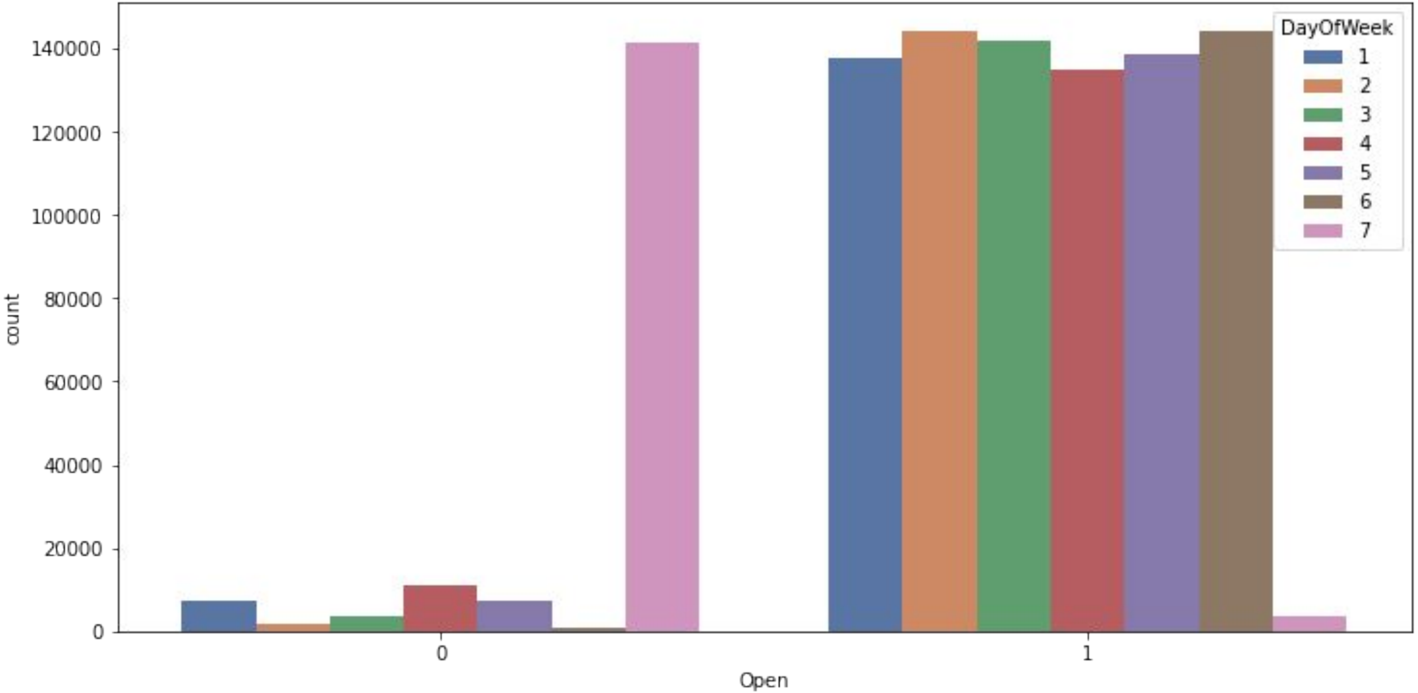
Countplot - Store Open/Closed



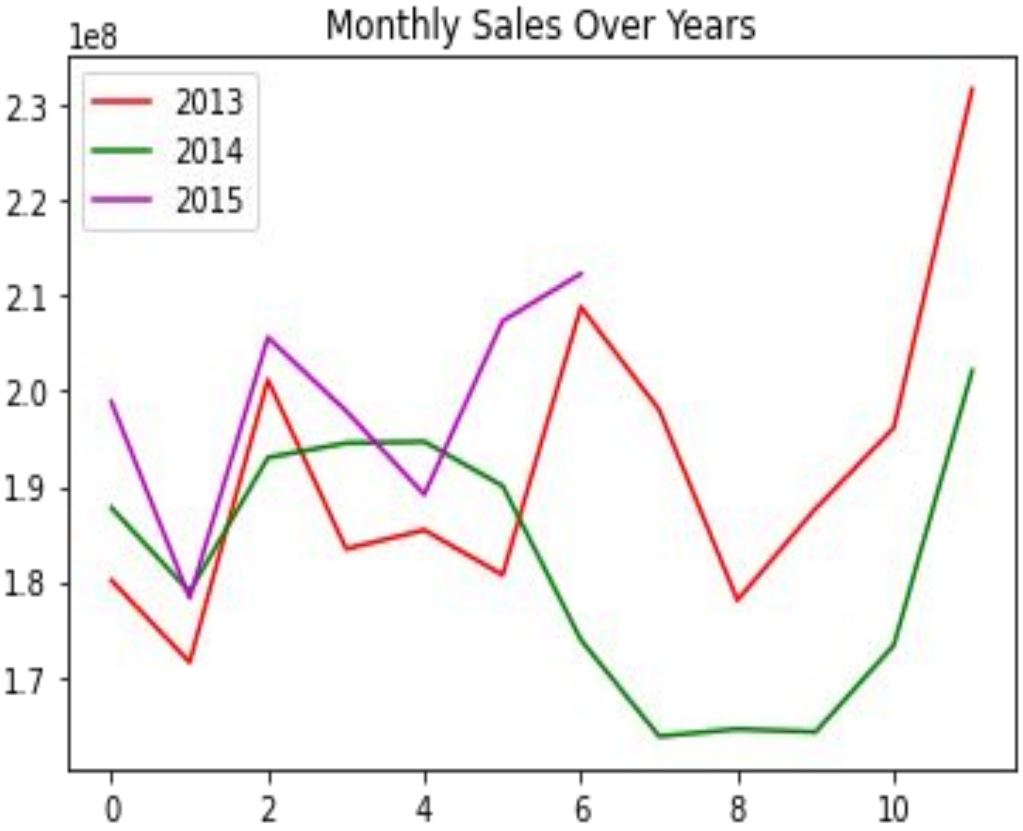
0 - closed
1 - opened

Store Open/Closed over Day of Week

- This countplot clearly shows that majority of stores are closed on sunday. Some stores were also closed on other days of the week may be due to public holidays & refurbishment, as stores are usually closed on public holidays and are open during school vacations.

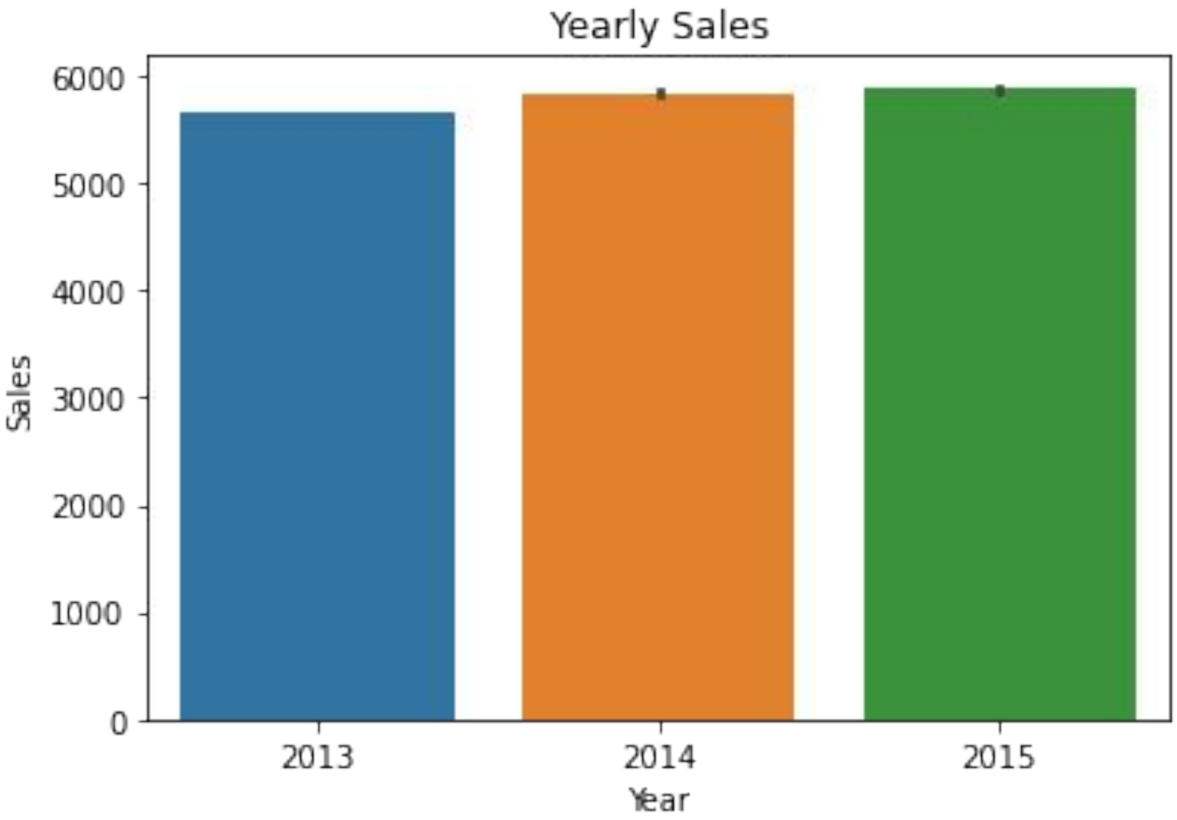


Average Sales Over Year/Month

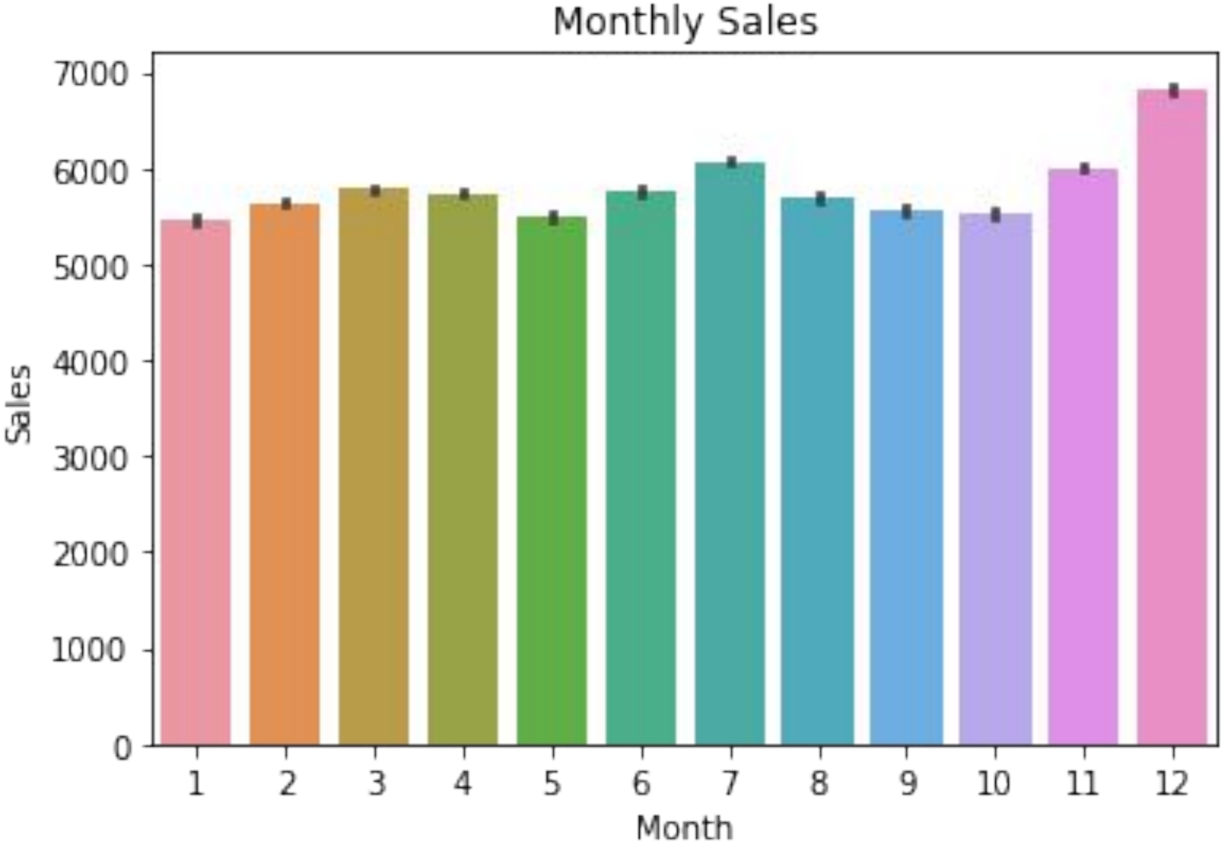


- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.

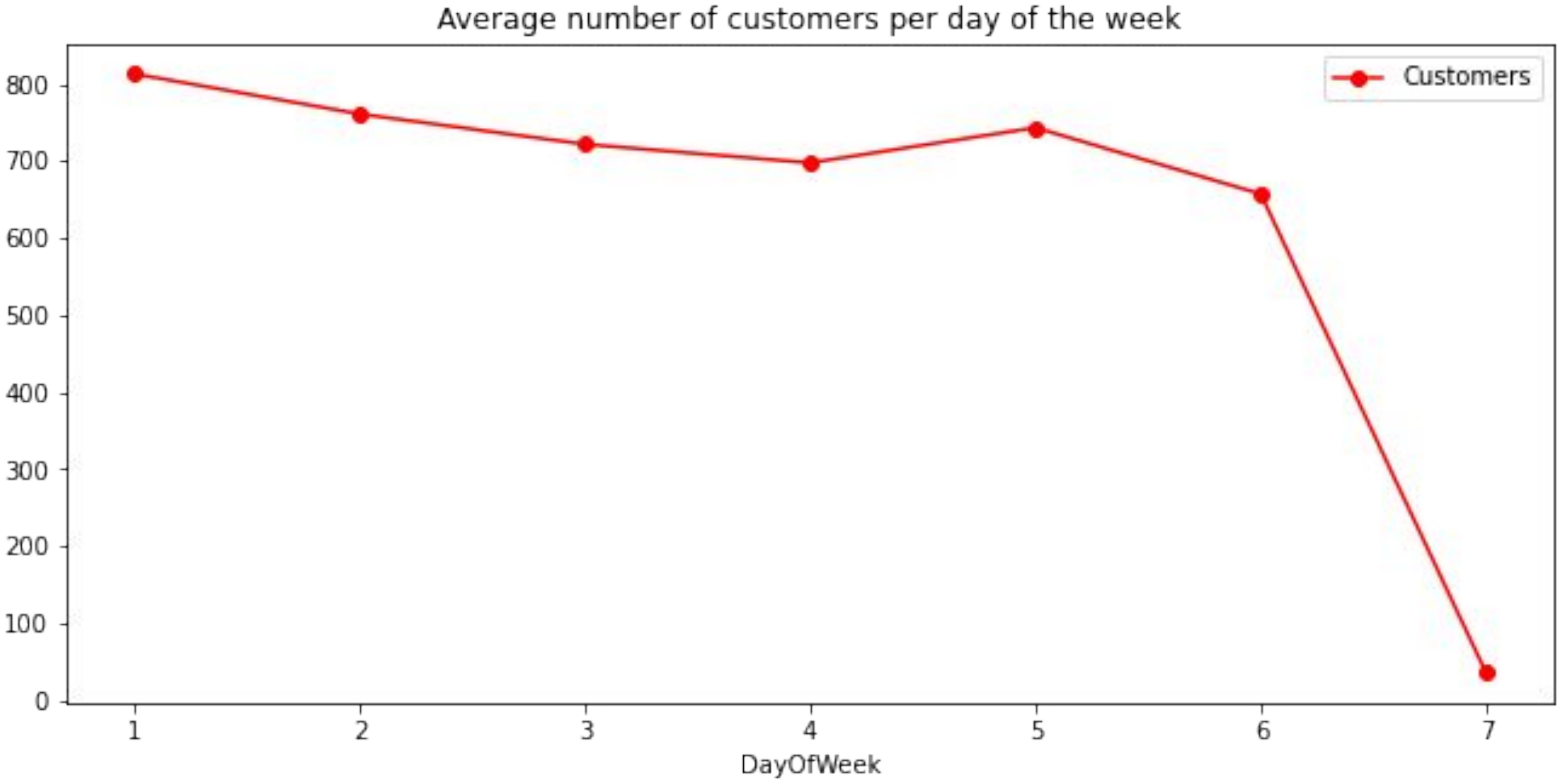
Total Sales in successive Years



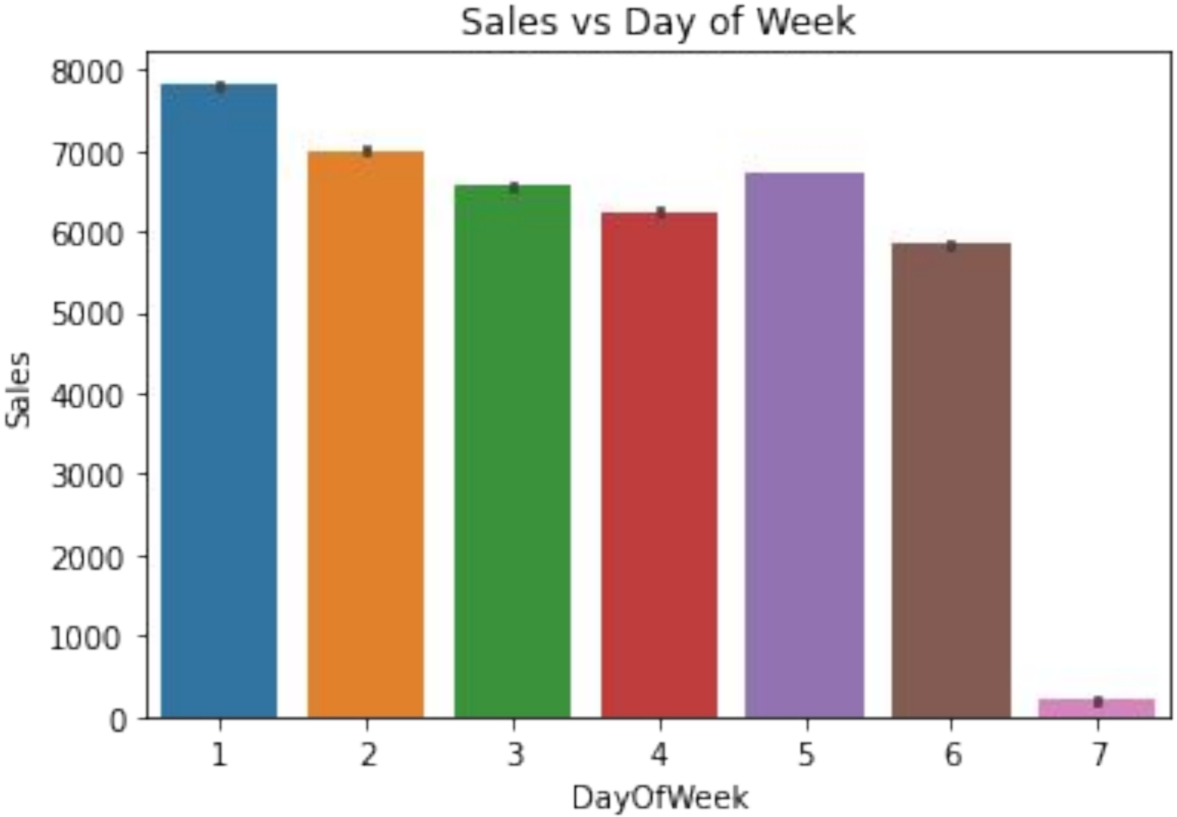
Total Sales in Months



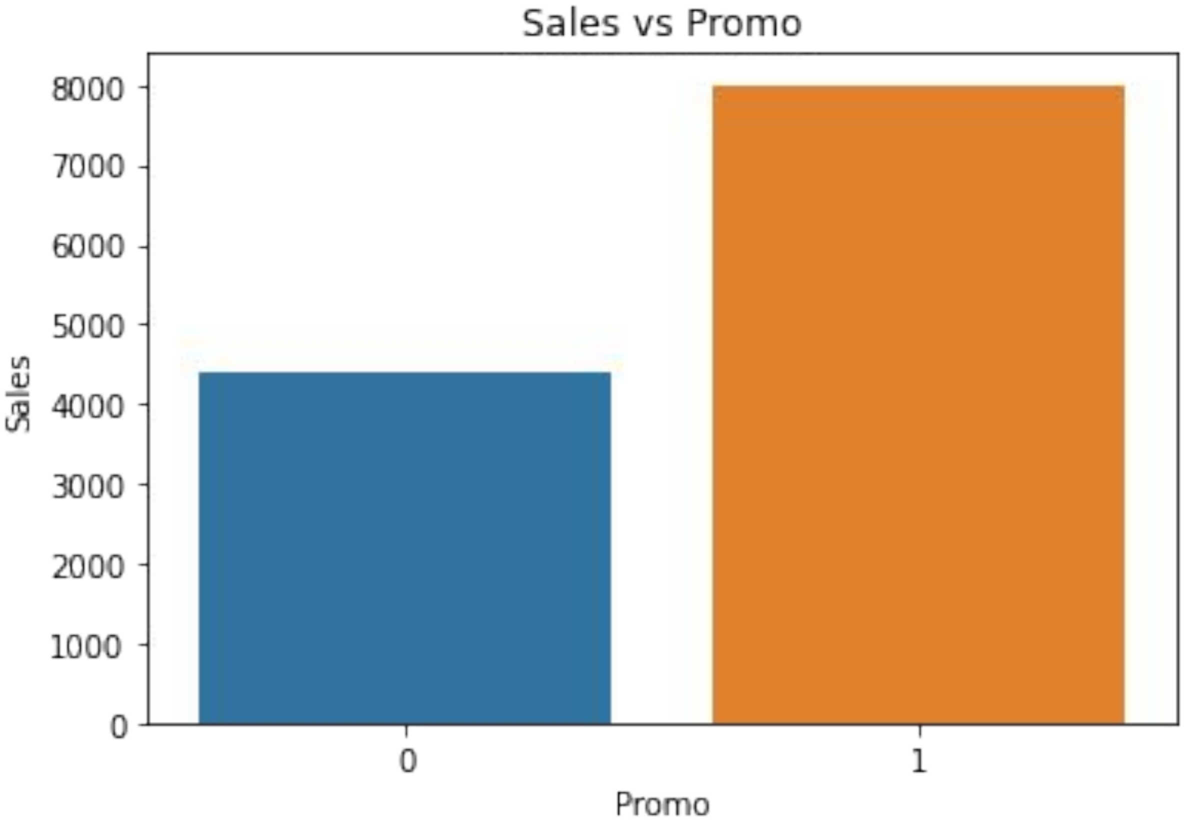
Trend - Average Customer per Day of Week



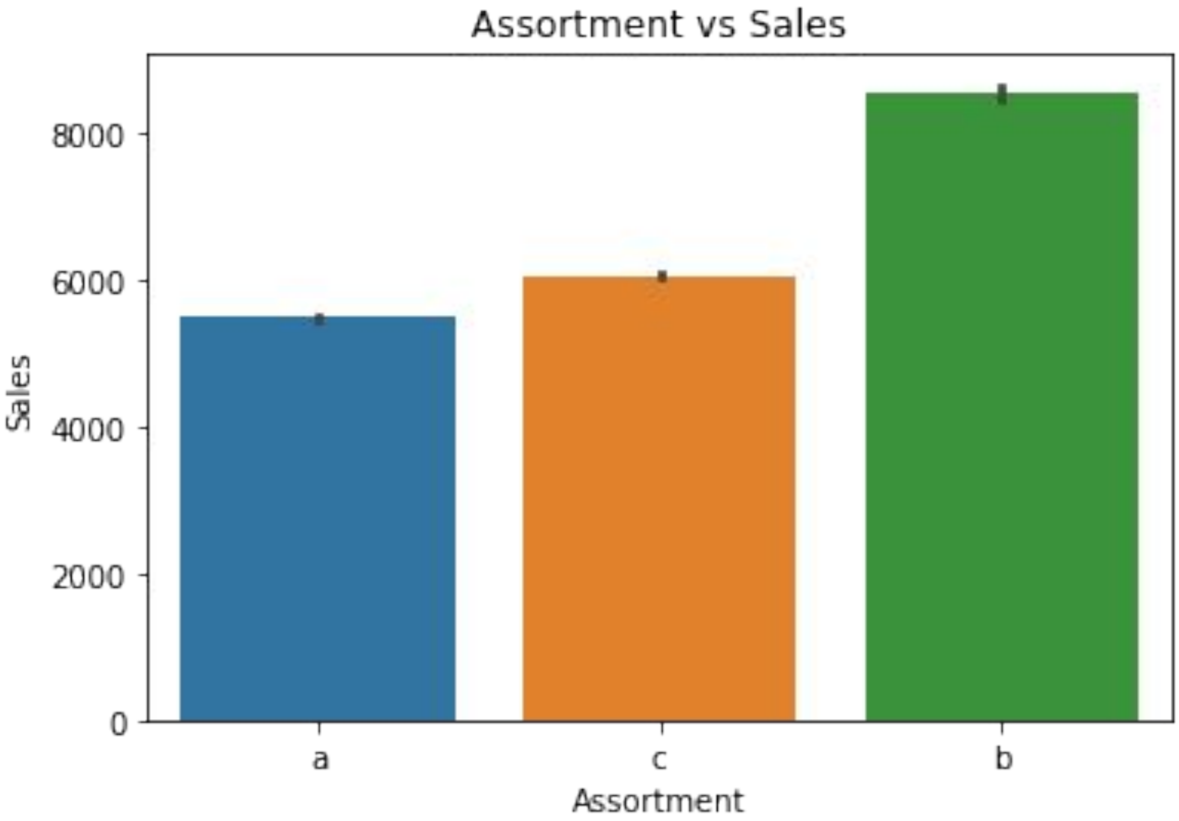
Total Sales on Weekdays



Sales VS Promo

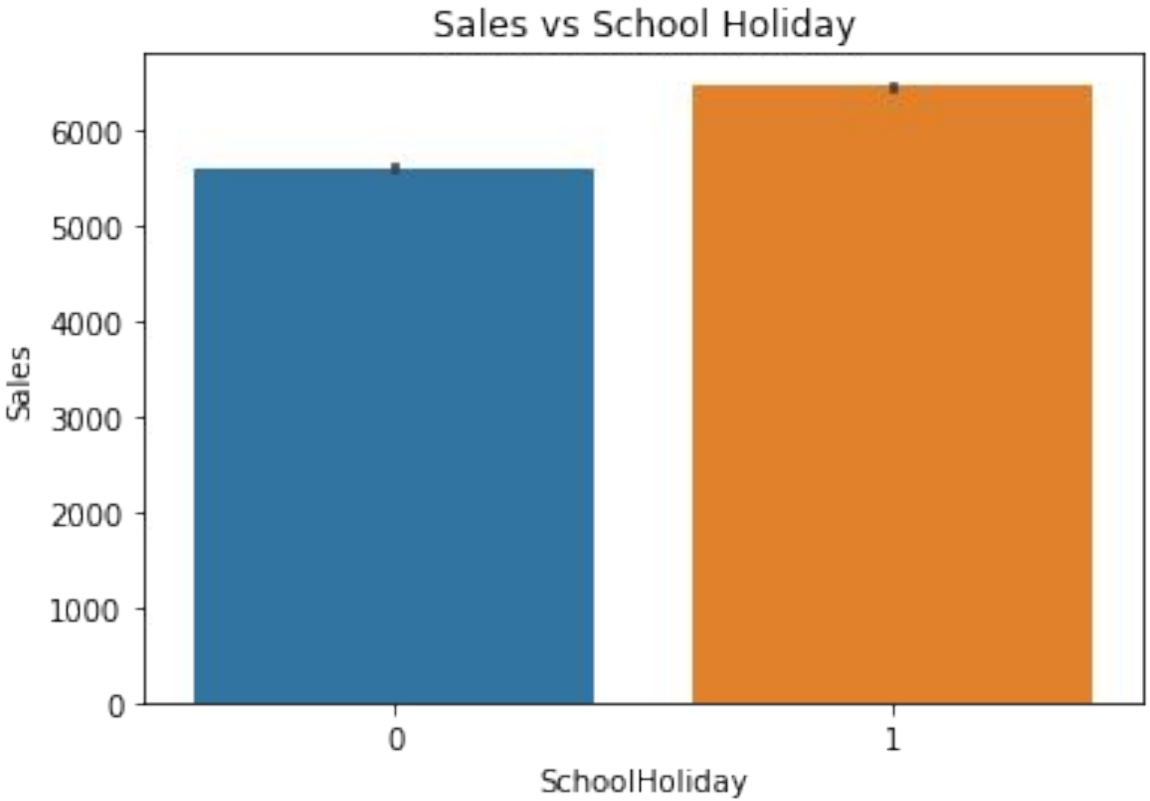


Sales vs Assortment



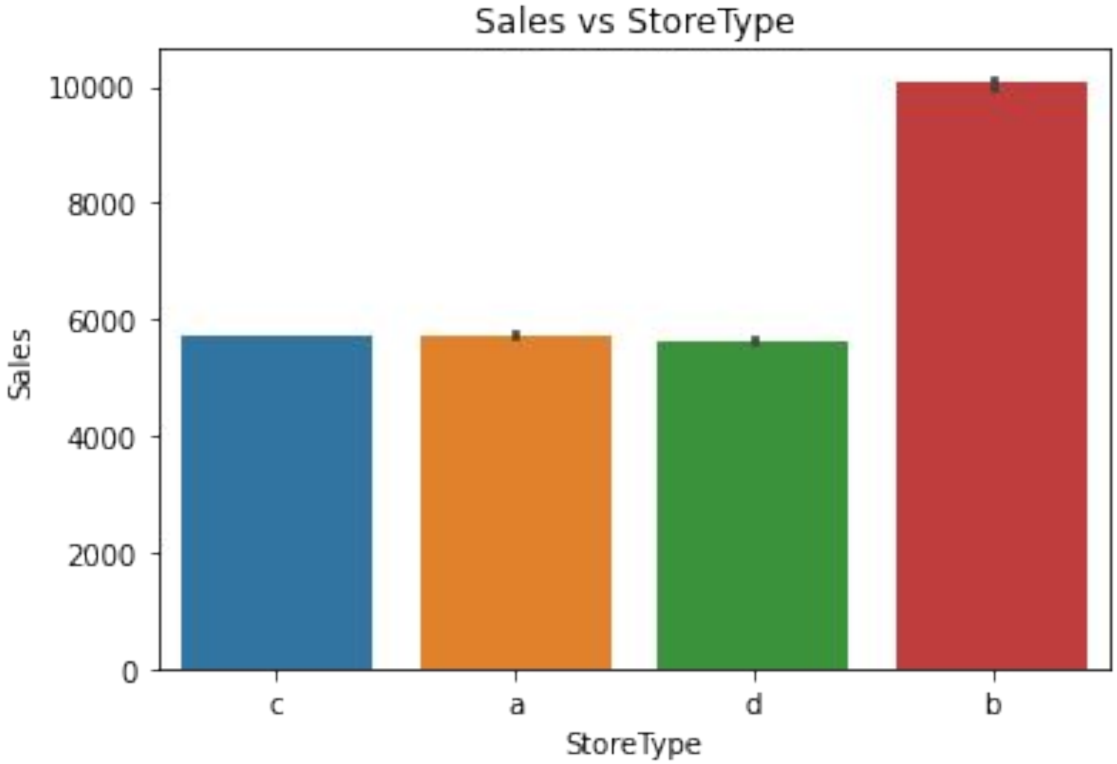
a - basic
b - extra
c - extended

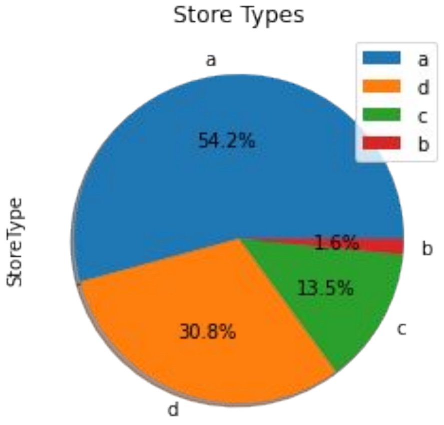
Sales vs School Holiday

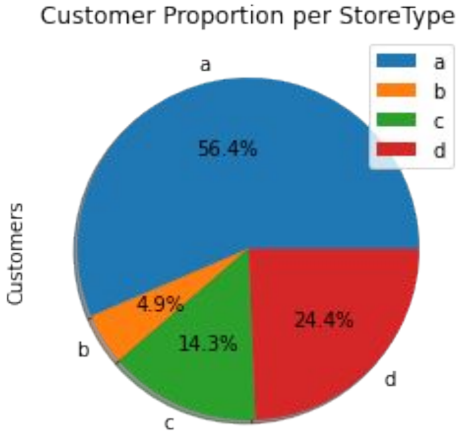


0 - Not Holiday
1 - School Holidays

Proportion of Store Types and Total Sales per Store Type







FEATURE ENGINEERING

Replacing Missing Values

- There are many Missing values in our Store Dataset.

Feature Name	Number of Missing values	Replaced with	Remark
Promo2SinceWeek	544	0	Not much Information given about the features.
Promo2SinceYear	544		
PromoInterval	544		
CompetitionOpenSinceMonth	354	Mode	
CompetitionOpenSinceYear	354		
CompetitionDistance	3	Median	As they are only 3.

Changing Datatype of Columns

- In Store Dataset - StoreType, Assortment and PromoInterval is of Object type.

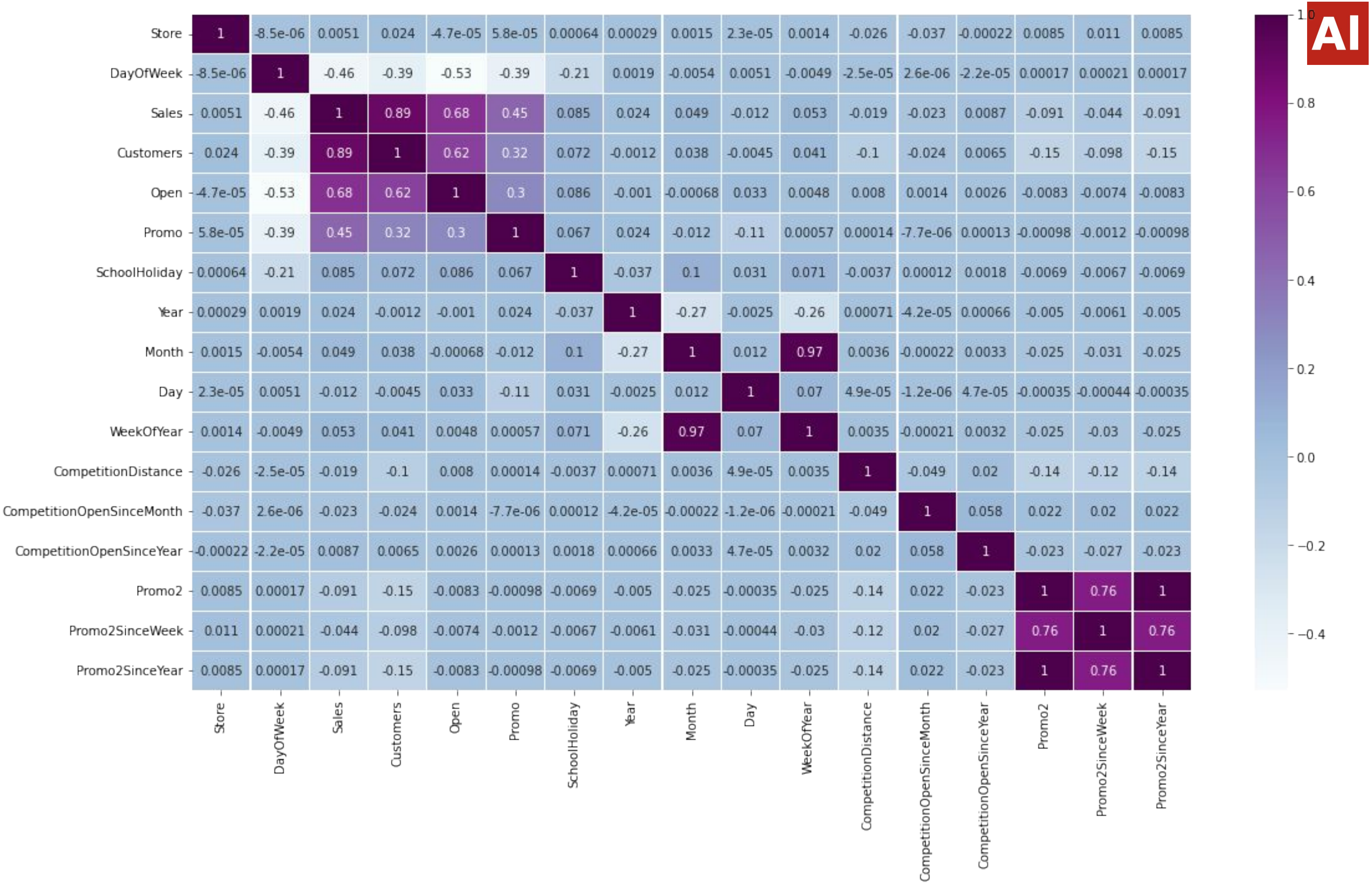
- StoreType and Assortment have values [a,b,c,d] and [a,b,c] respectively.

- Changing these to Numerical values [0,1,2,3] and [0,1,2] respectively.

- PromoInterval has values [0, 'Feb,May,Aug,Nov', 'Jan,Apr,Jul,Oct', 'Mar,Jun,Sept,Dec']

- Changing these to dummy variables.

Correlation Heatmap

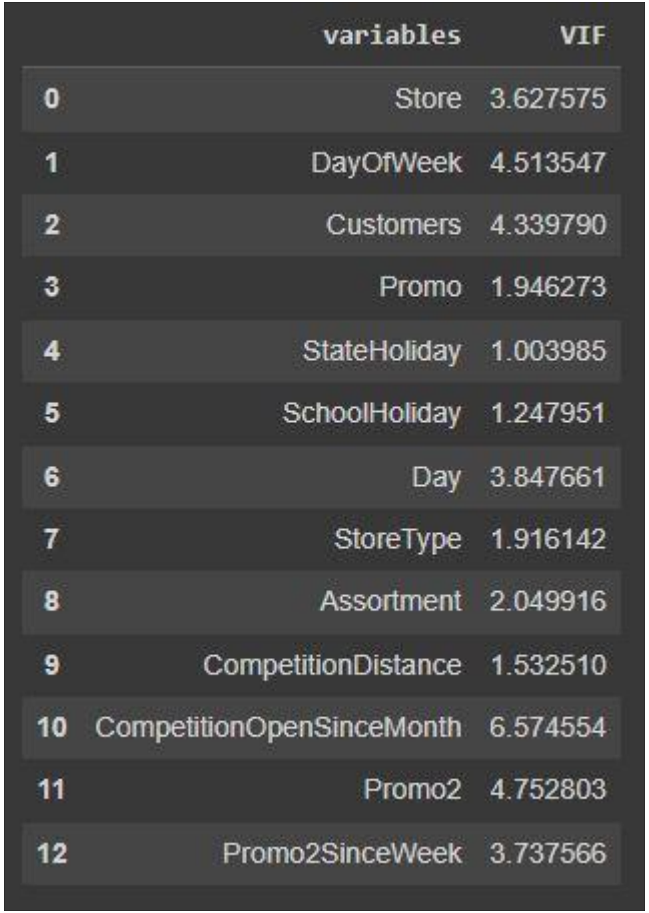


	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	Year	Month	Day	WeekOfYear	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
Store	1	-8.5e-06	0.0051	0.024	-4.7e-05	5.8e-05	0.00064	0.00029	0.0015	2.3e-05	0.0014	-0.026	-0.037	-0.00022	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	-0.0049	-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	0.053	-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	0.041	-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	0.0048	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	0.00057	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	0.071	-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	-0.26	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	0.97	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	0.07	4.9e-05	-1.2e-06	4.7e-05	-0.00035	-0.00044	-0.00035
WeekOfYear	0.0014	-0.0049	0.053	0.041	0.0048	0.00057	0.071	-0.26	0.97	0.07	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	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	-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	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	-0.025	-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	-0.03	-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	-0.025	-0.14	0.022	-0.023	1	0.76	1

Removing Multicollinearity

- Removing the features which is having VIF>10 because it will affect & interpret the result.

- VIF <= 10 is usually preferred as this can easily explain the variance of 90% i.e, R-square becomes 90%.(VIF=1/1-R^2)

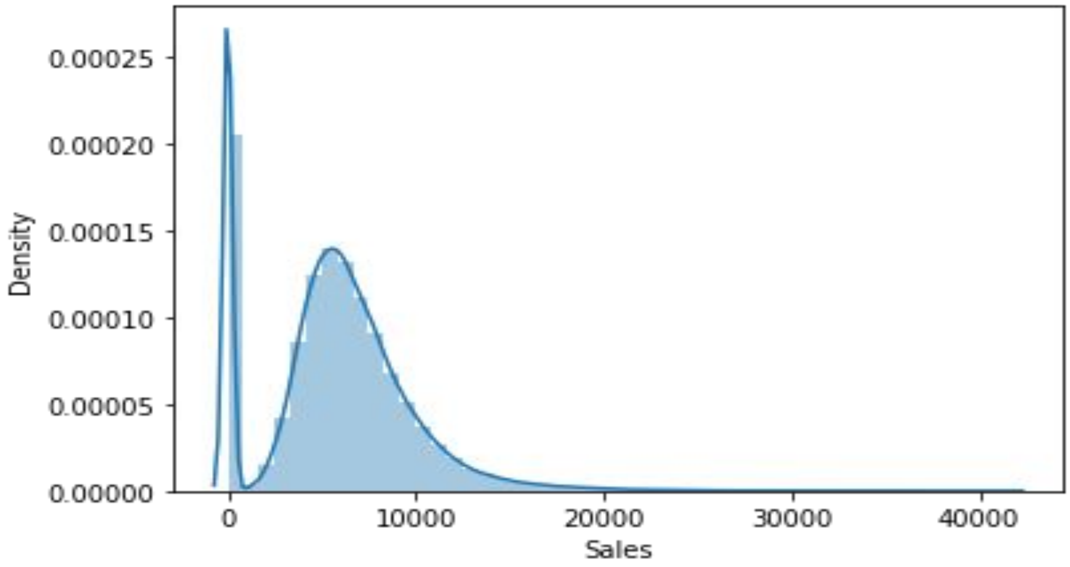


	variables	VIF
0	Store	3.627575
1	DayOfWeek	4.513547
2	Customers	4.339790
3	Promo	1.946273
4	StateHoliday	1.003985
5	SchoolHoliday	1.247951
6	Day	3.847661
7	StoreType	1.916142
8	Assortment	2.049916
9	CompetitionDistance	1.532510
10	CompetitionOpenSinceMonth	6.574554
11	Promo2	4.752803
12	Promo2SinceWeek	3.737566

Filtering Rows(Records)

- Filtering records where stores are closed as they won't generate any Sales.

- Filtering records where stores has Sales equal to 0.



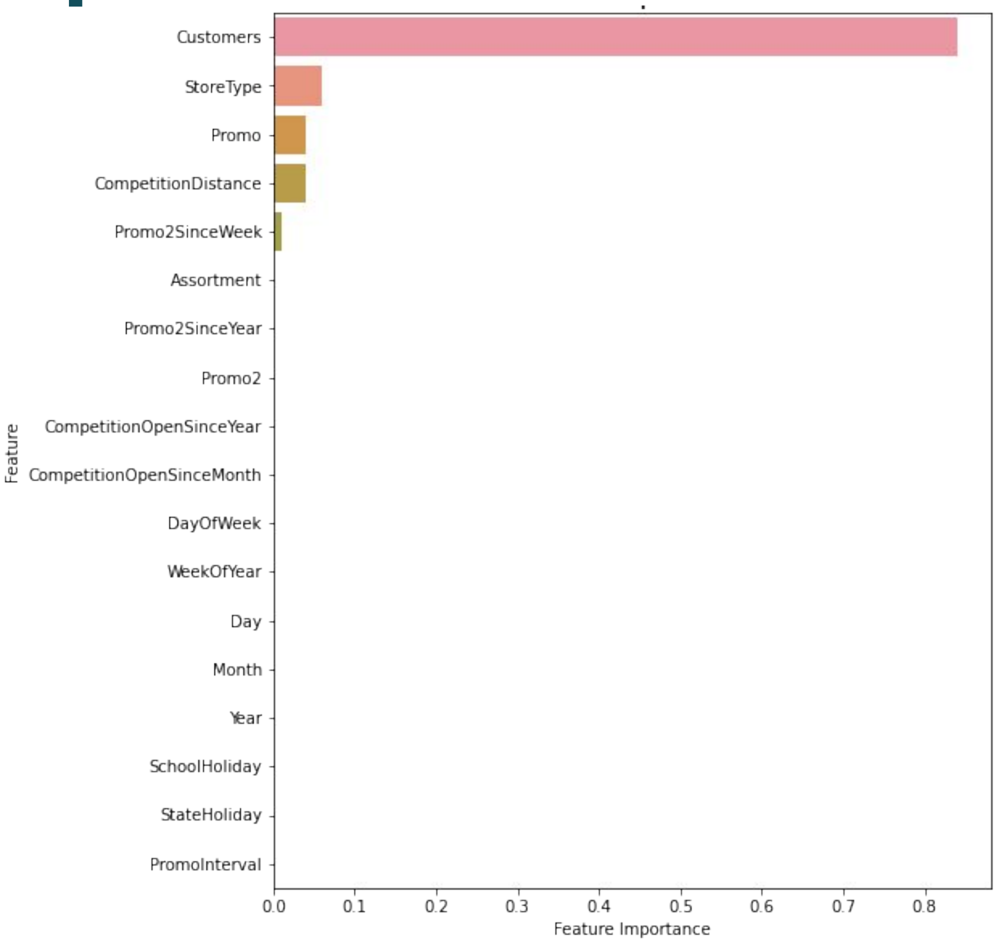
MACHINE LEARNING MODEL BUILDING

Evaluation Metrics

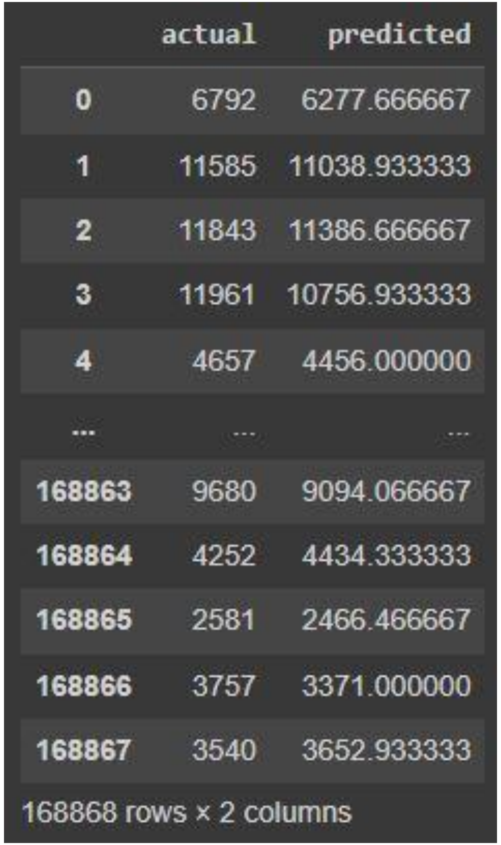
ML Model predict sales for stores which are open and when there is some sales because there is no sales when store is closed.

	Linear Regression (OLS)	Lasso Regression (L1)	Ridge Regression (L2)	Decision Tree	Random Forest	Random Forest with hyperparameter tuning
RMSE	1519.132	1520.620	1519.849	1433.262	1126.660	513.893
MAPE	15.902	15.920	15.908	15.478	12.442	4.993
R2	0.761	0.7607	0.7609	0.7609	0.8686	0.9727

Feature Importance



Actual vs Predicted Values for Random Forest Hyperparameter & Cross-validation Tuned Model



	actual	predicted
0	6792	6277.666667
1	11585	11038.933333
2	11843	11386.666667
3	11961	10756.933333
4	4657	4456.000000
...	...	...
168863	9680	9094.066667
168864	4252	4434.333333
168865	2581	2466.466667
168866	3757	3371.000000
168867	3540	3652.933333

168868 rows × 2 columns

Model Selection

By Looking at the evaluation metrics obtained on implementing different sort of regression model, we decided to go with the Random Forest Tuned model.The maximum R^2 was seen in tuned Random Forest model with the value 0.97267. It means our best accurate model is able to explain approx/almost 97% of variances in the datasets.

Based on our model; Customer, store Type, Promo & Competition Distance are the most impactful features which are driving the sales more as compared to other features present in the dataset.

Challenges Faced

- Understanding the meaning of some columns.

- Handling Large amount of Sales Data.

- Dealing with Null values, as there are many Null values.

- Understanding the business model of Retail Sales that how they work.

- Also, forming different graphs to show insights from the dataset and to summarize the information and communicate the results and trends to the reader successfully.

- Dealing with Categorical columns to make them numerical for make use in ML model building.

- Selecting appropriate Model to fulfill the purpose.

Conclusions

- From the sales and customer scatterplot, the relationship is sort of linear ie sales is increasing with number of customers increasing which is obvious.
- Stores with Assortment level ‘b’ has the highest sales.
- Approx. 50% stores are of type ‘a’. There are very few stores of type ‘b’.
- Store type ‘b’ has the highest sales and all other store types ‘a’,’c’,’d’ has nearly equal sales.
- December records the highest monthly sales. This may be due to Christmas and New Year.
- Sales is more when promos/offers are running on stores.

Recommendations from our Analysis

- More stores should be encouraged for promos.

- Store type ‘b’ should be increased in number.

- There is seasonality involved. Hence, the stores should be encouraged to promote and take advantages of the holidays.

Thank You

Hope our best model will help the Store managers to predict the daily sales accurately in advance.

CHEERS ✌!!!

Team Alma Phoenix