ROSSMANN STORE – DATA ANALYSIS & SALES FORECAST

-- Pradeep Joshi

-- Abhishek Goswami

-- Rahul

University at Buffalo The State University of New York

CONTENT

INTRODUCTION

Predicting sales performance is one of the key challenges every business face. It is important for firms to predict customer demands to offer the right product at the right time and at the right place. It is very important for retail stores to save money on their inventory and increase profit by meeting demands. Thus, it will help a lot on stores' earnings if their sales are predictable.

The importance of this issue is underlined by the fact that figuratively a bazillion consulting firms are on the market trying to offer sales forecasting services to businesses of all sizes. Some of these firms rely on advanced data analytics techniques, the kind of which we also have covered in our analysis and prediction.

The topic was chosen, because the problem is intuitive to understand. We have a well understanding of the problem from our daily life, which makes us more focused on training methodology.

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. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

OBJECTIVE:-

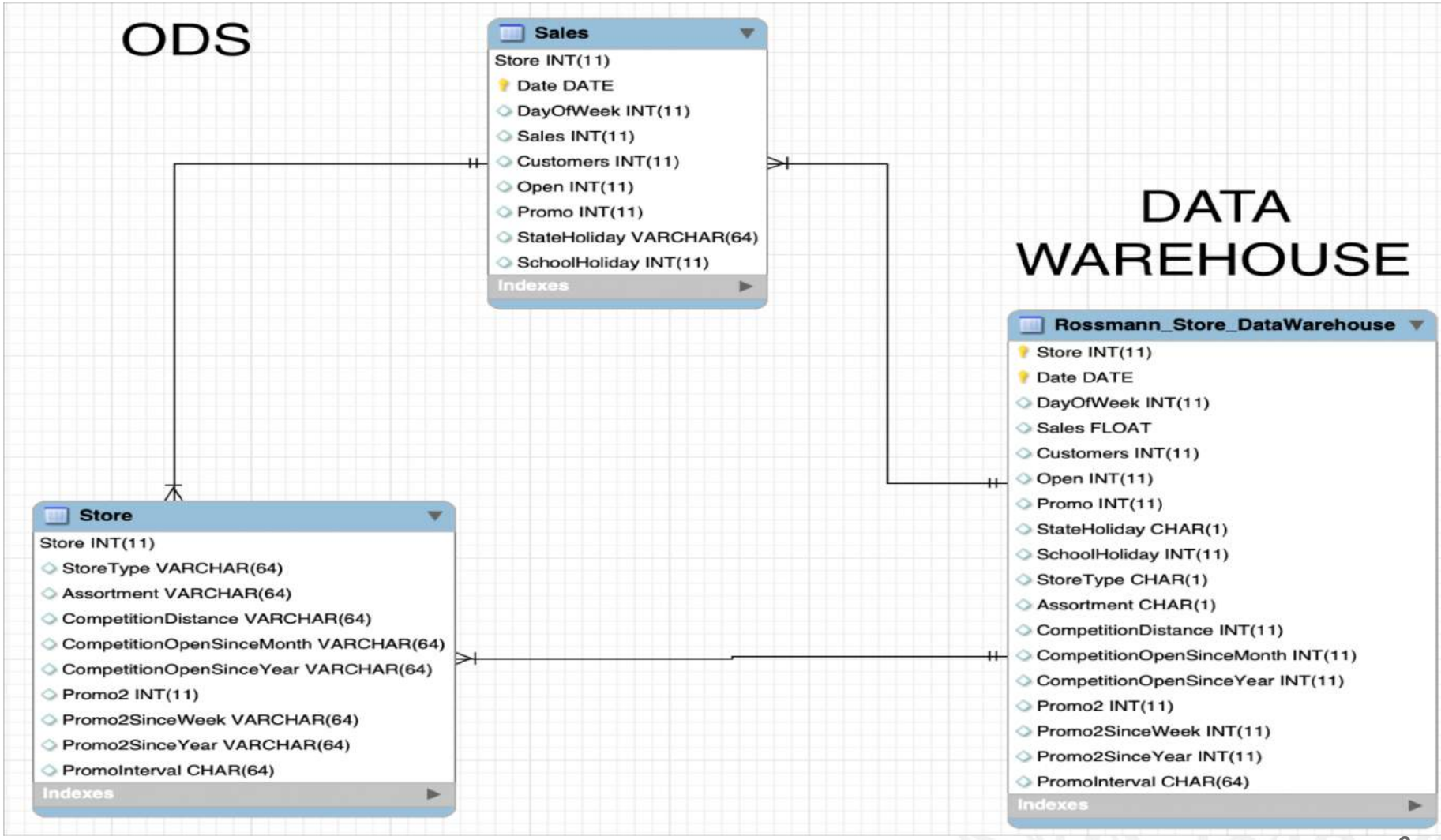
- Help Rossmann create a robust sales prediction model.

DATA SET(Overview)

Historical data up to 2 year 7 month is provided(Jan 2013 to July 2015)

SNo	Data Set	Variables	No of variables	No of Observations
1.	Sales	store, day of week, date, sales, customers, open, promo, state holiday, school holiday	9	1017210
2.	Store	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10	1115

DATA MODEL


ODS
Sales
Store INT(11)
Date DATE
DayOfWeek INT(11)
Sales INT(11)
Customers INT(11)
Open INT(11)
Promo INT(11)
StateHoliday VARCHAR(64)
SchoolHoliday INT(11)
Indexes
DATA WAREHOUSE
Rossmann_Store_DataWarehouse
Store INT(11)
Date DATE
DayOfWeek INT(11)
Sales FLOAT
Customers INT(11)
Open INT(11)
Promo INT(11)
StateHoliday CHAR(1)
SchoolHoliday INT(11)
StoreType CHAR(1)
Assortment CHAR(1)
CompetitionDistance INT(11)
CompetitionOpenSinceMonth INT(11)
CompetitionOpenSinceYear INT(11)
Promo2 INT(11)
Promo2SinceWeek INT(11)
Promo2SinceYear INT(11)
PromoInterval CHAR(64)
Indexes
Store
Store INT(11)
StoreType VARCHAR(64)
Assortment VARCHAR(64)
CompetitionDistance VARCHAR(64)
CompetitionOpenSinceMonth VARCHAR(64)
CompetitionOpenSinceYear VARCHAR(64)
Promo2 INT(11)
Promo2SinceWeek VARCHAR(64)
Promo2SinceYear VARCHAR(64)
PromoInterval CHAR(64)
Indexes

PREVIOUS WORK

We read a number of old posts from Kaggle giving us various approaches on how to improve the scores on the algorithms, and how to fine tune the best performing algorithms.

From our literature review, the highest scoring teams in the Kaggle competition added weather features to improve their performance. The supplementary data sets for this are available on the Kaggle forums, but we decided not to incorporate this data since it would break the generality of our model.

Normally you would not know what the weather was going to be in the future, so it would not be a reliable feature to add to the test set. In the case of the Kaggle competition, the sales that were being predicted were already in the past, and the actual weather during that period was known.

Since the goal in the competition was to super tune the models in order to win the prize, they were not concerned with the generality of their models. But we preferred to build our model to actually work for future predictions when the future weather would not be known.

DATA PRE-PROCESSING

SALES DATA

- There were no null values.
- The data ranged from 2013-01-01 to 2015-07-31.
- Average Sale per customer = *'9.493619'*.
- Sales column had 0 for *'172871'* records, out of which *'172817'* records were having sales 0 because the store was closed.
- Predicting sales for closed shops is trivial, hence removed all data for closed store. Number of records left = *'844392'*.

STORE DATA

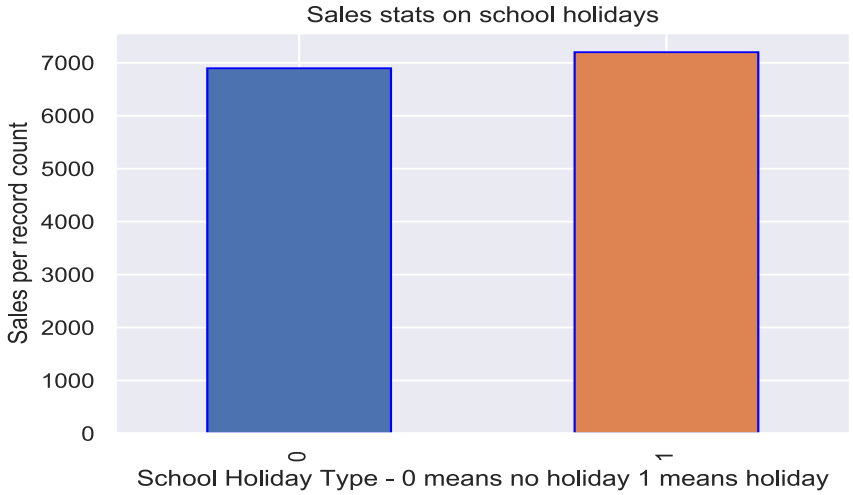
- Null values were present and needed to be fixed.
- Replaced by 0 for Promo2 related columns.
- Replaced using "most occurring data", for competition distance related columns.

- Merged the two files.

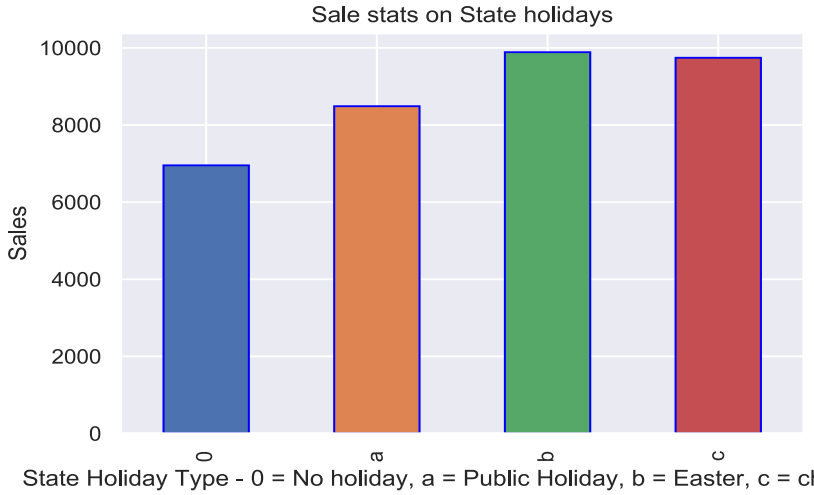
```
Store                        0
StoreType                    0
Assortment                   0
CompetitionDistance          3
CompetitionOpenSinceMonth  354
CompetitionOpenSinceYear   354
Promo2                       0
Promo2SinceWeek            544
Promo2SinceYear            544
PromoInterval    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

Exploratory Data Analysis (EDA)

School Holiday v/s Sales



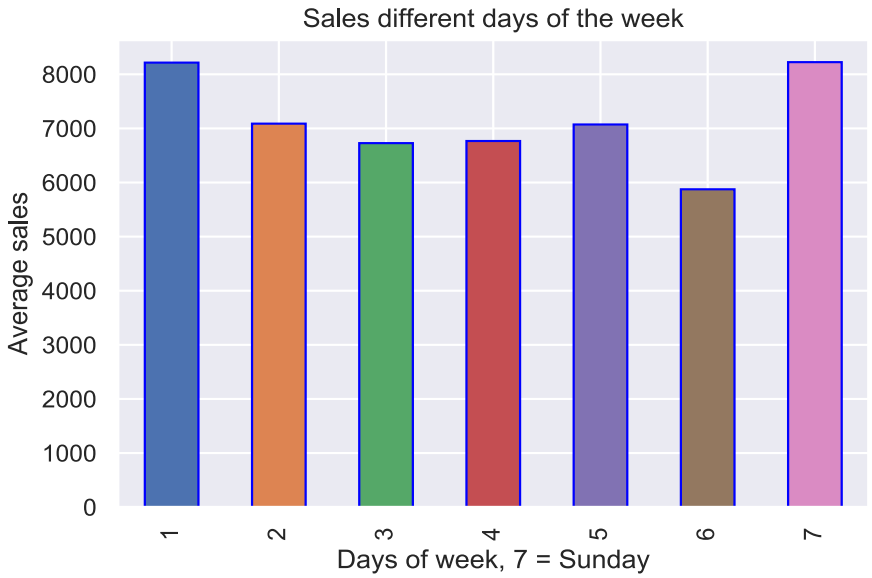
State Holiday v/s Sales



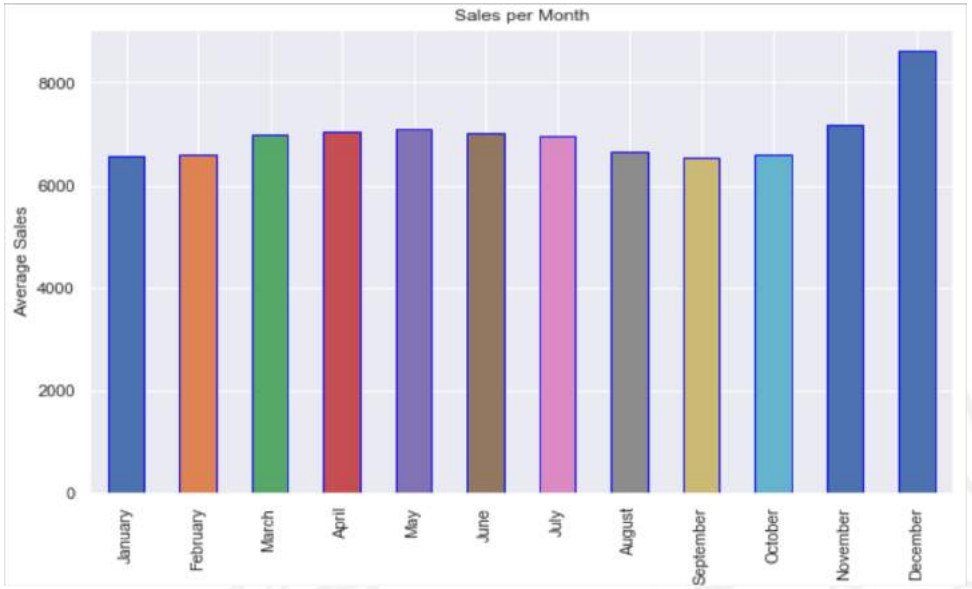
Promo v/s Sales

Inference: Sales are quite high when there is holiday ,especially, when there is Easter or Christmas. Moreover sales are also high when there is promo.

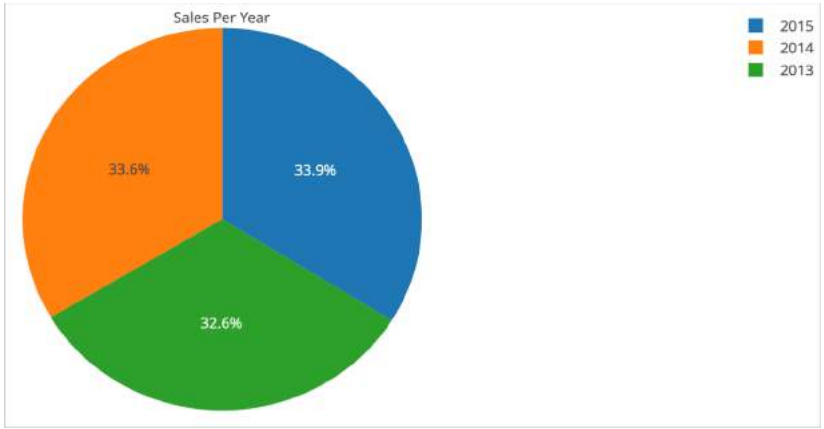
WEEKLY SALE



MONTHLY SALE

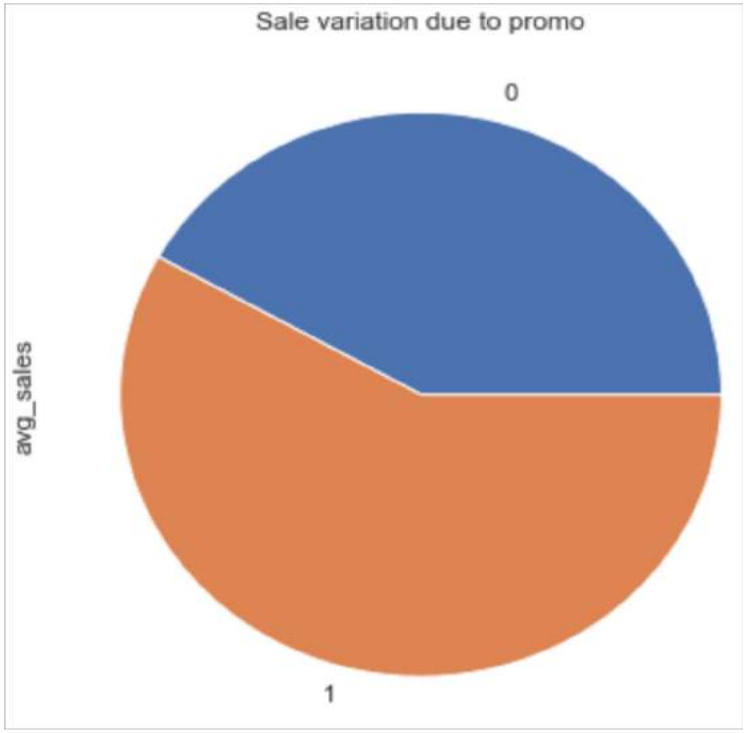


YEARLY SALE

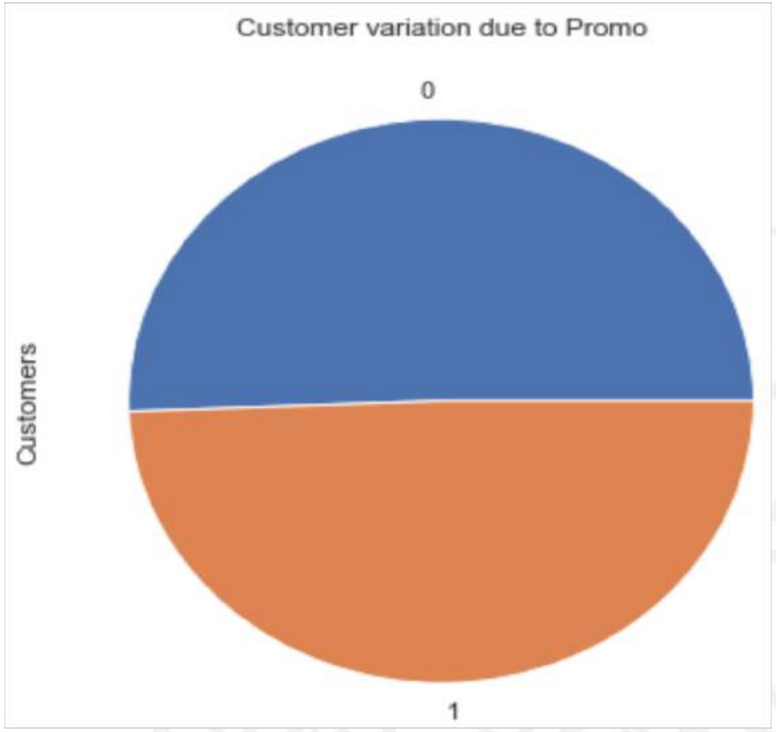


Inference: Sales are quite high in November and December.

SALE VARIATION wrt. PROMO

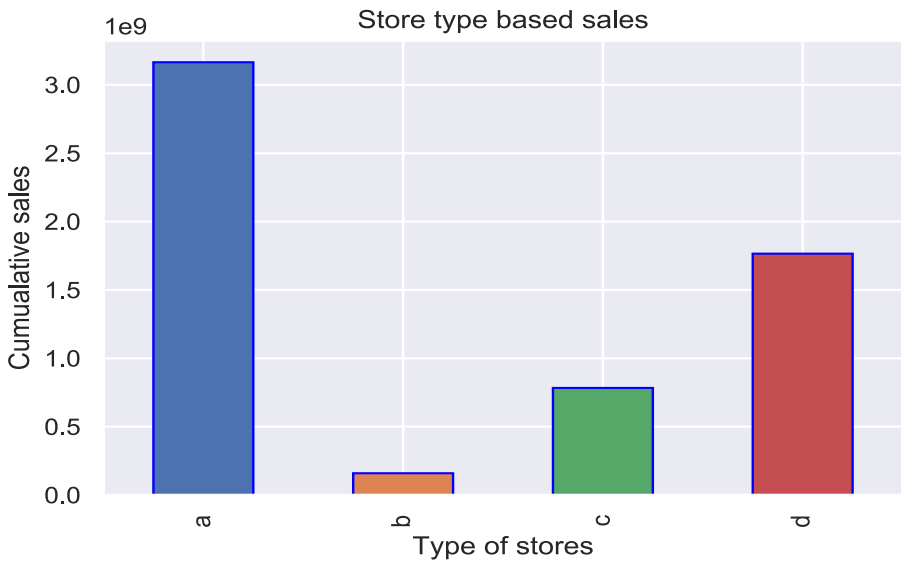


CUSTOMER VARIATION wrt. PROMO

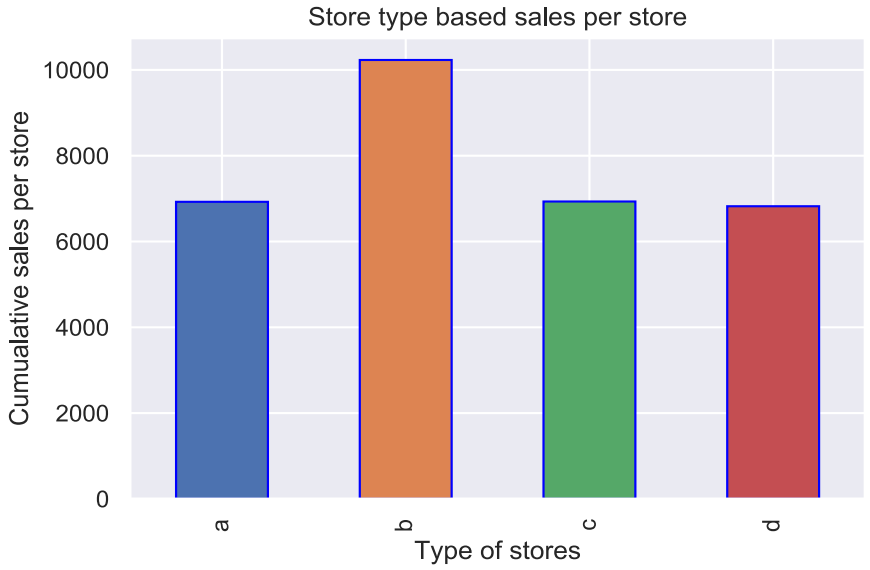


Inference: Although average sales increase during promotion , number of customers stay the same.

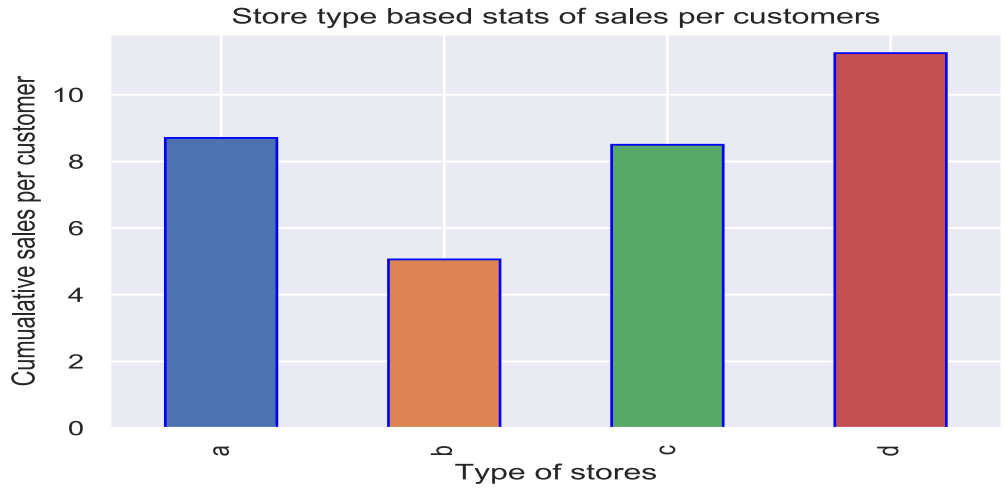
SALES PER STORE TYPE



SALES PER STORE PER STORE TYPE

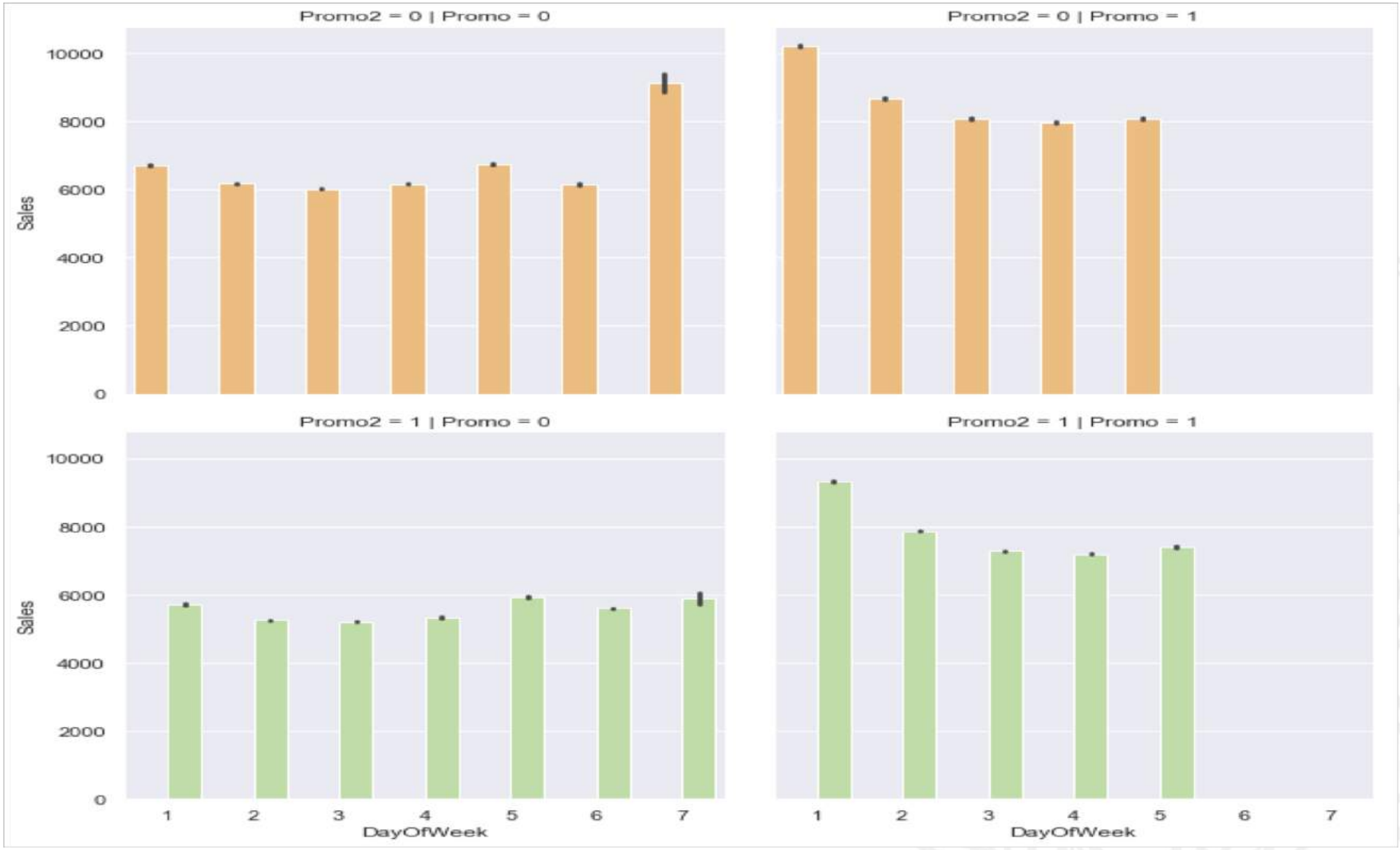


SALES PER CUSTOMER PER STORE TYPE

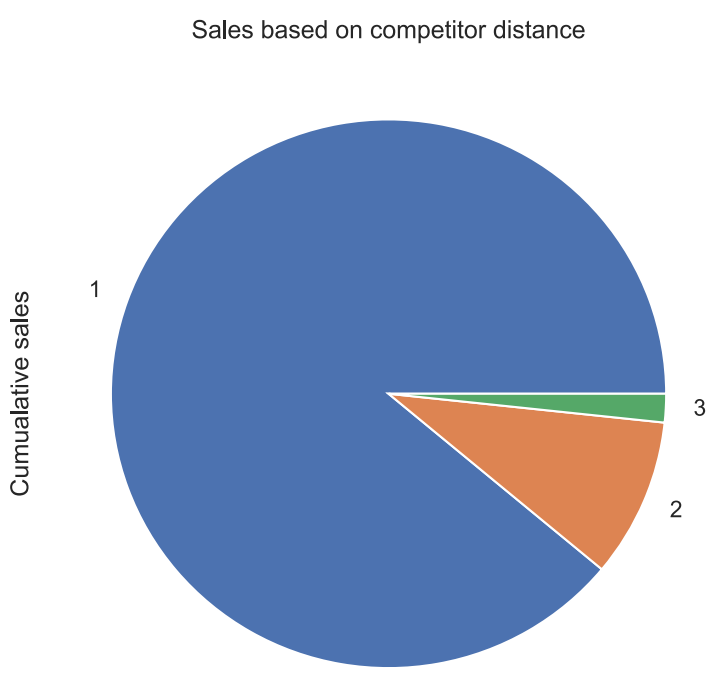


Inference: Cumulative sales are quite high for type-a store and lowest for type-b store but cumulative sales per store is maximum for type-b store and almost same for other store types. And cumulative sales per customer is highest in type-d store.

SALES VARIATION PER DAY wrt PROMO 2 and PROMO

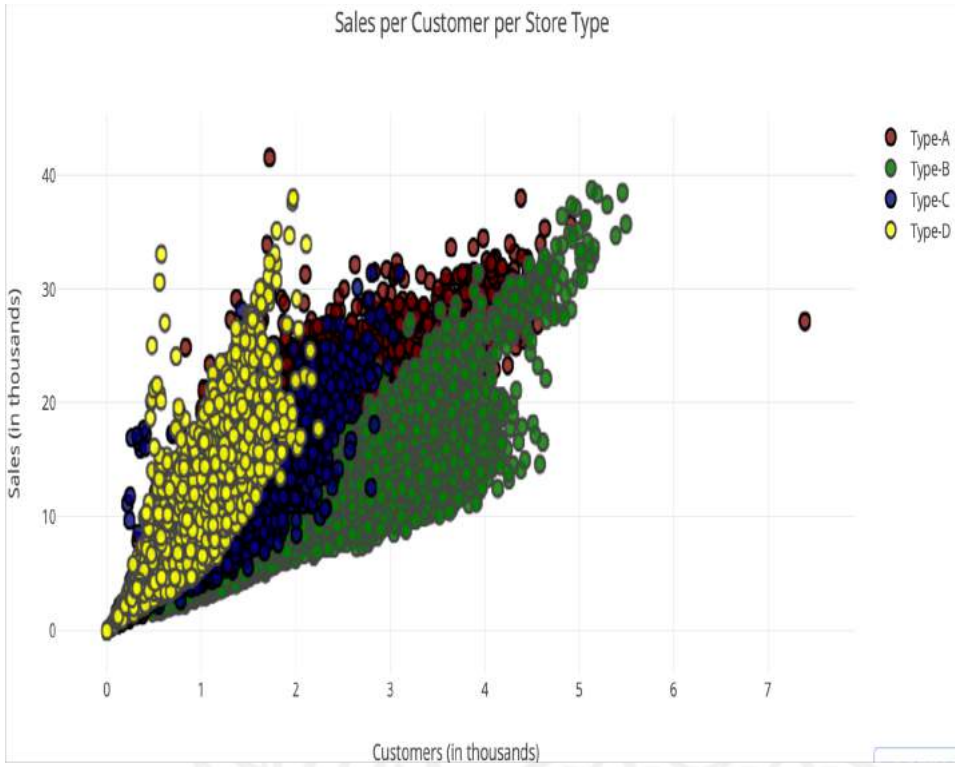


SALES VARIATION wrt COMPETITOR DISTANCE



Distances shown by below 15000 as 1, 15000 to 30000 as 2 and above 30000 as 3

SALES VARIATION PER STORE TYPE wrt CUSTOMERS



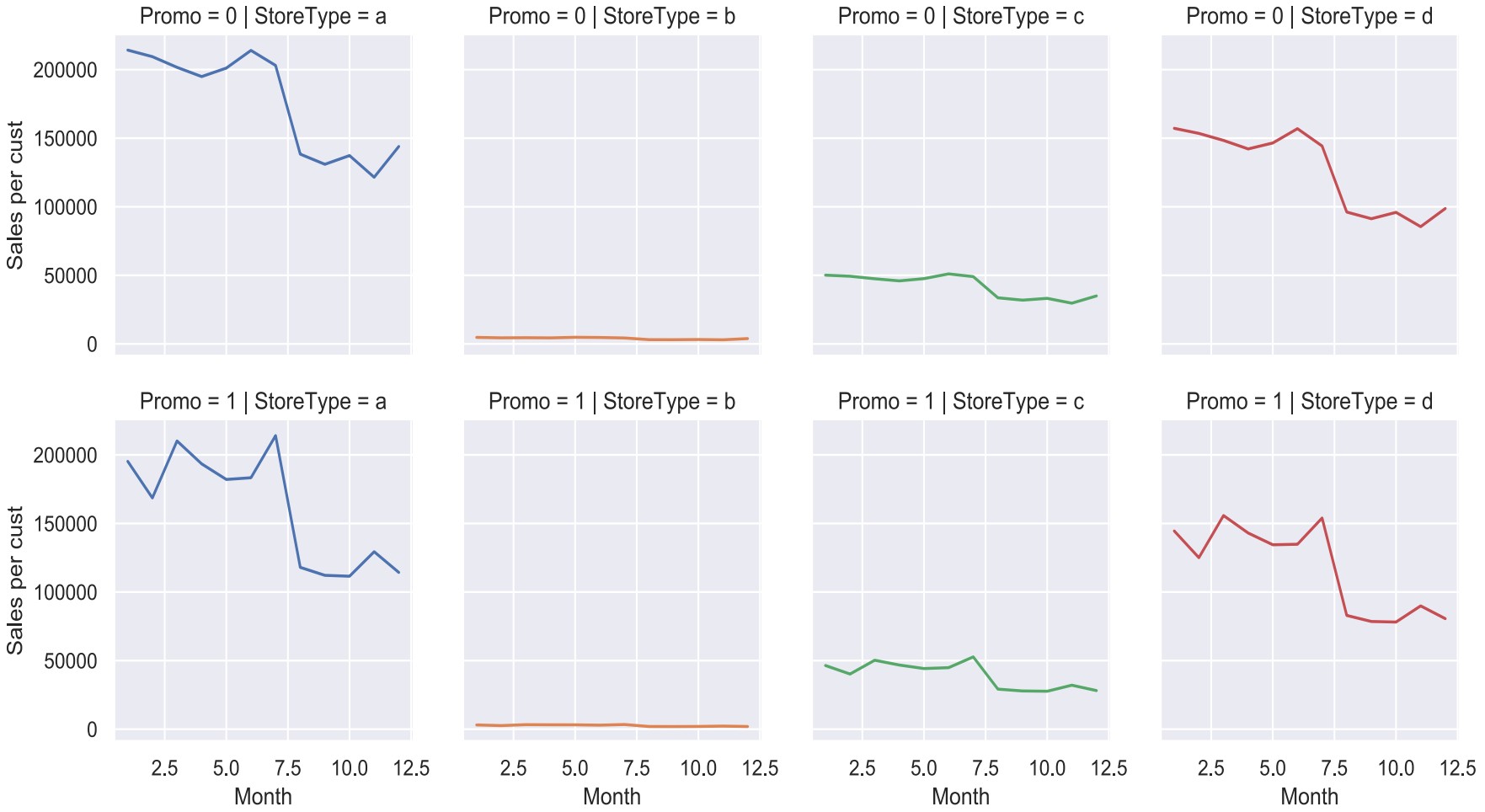
Inference: Cumulative sales are high when competition distance is less.

EDA CONCLUSION

1. Store type A is most crowded and has most sale i.e. 3165334859 and 457077 respectively.
2. Store B has the most numbers when we calculate sale per store in different store type but this is mainly for low data on b 10231.407505.
3. Sale per customer is highest for Store Type D 11.251350.
4. Store type C remains closed on Sundays.
5. Customers tend to by on Sundays if there is no Promotion and on Mondays if there is a promotion.
6. Promo 2 does not show any affect on sales.
7. Sale is higher for stores that have competitors nearby.

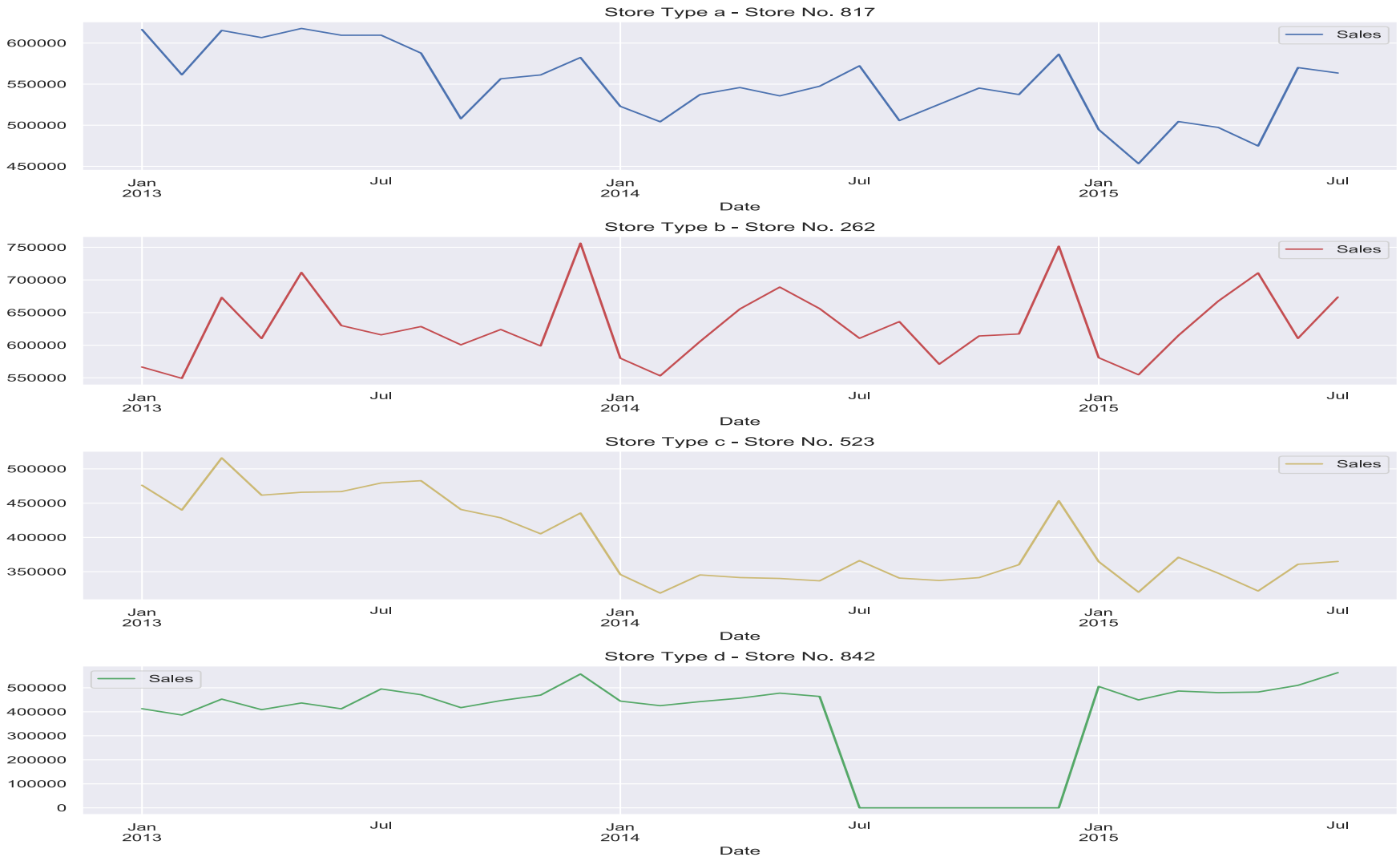
TIME SERIES ANALYSIS

MONTHLY SALES VARIATION wrt PROMO & STORE TYPE

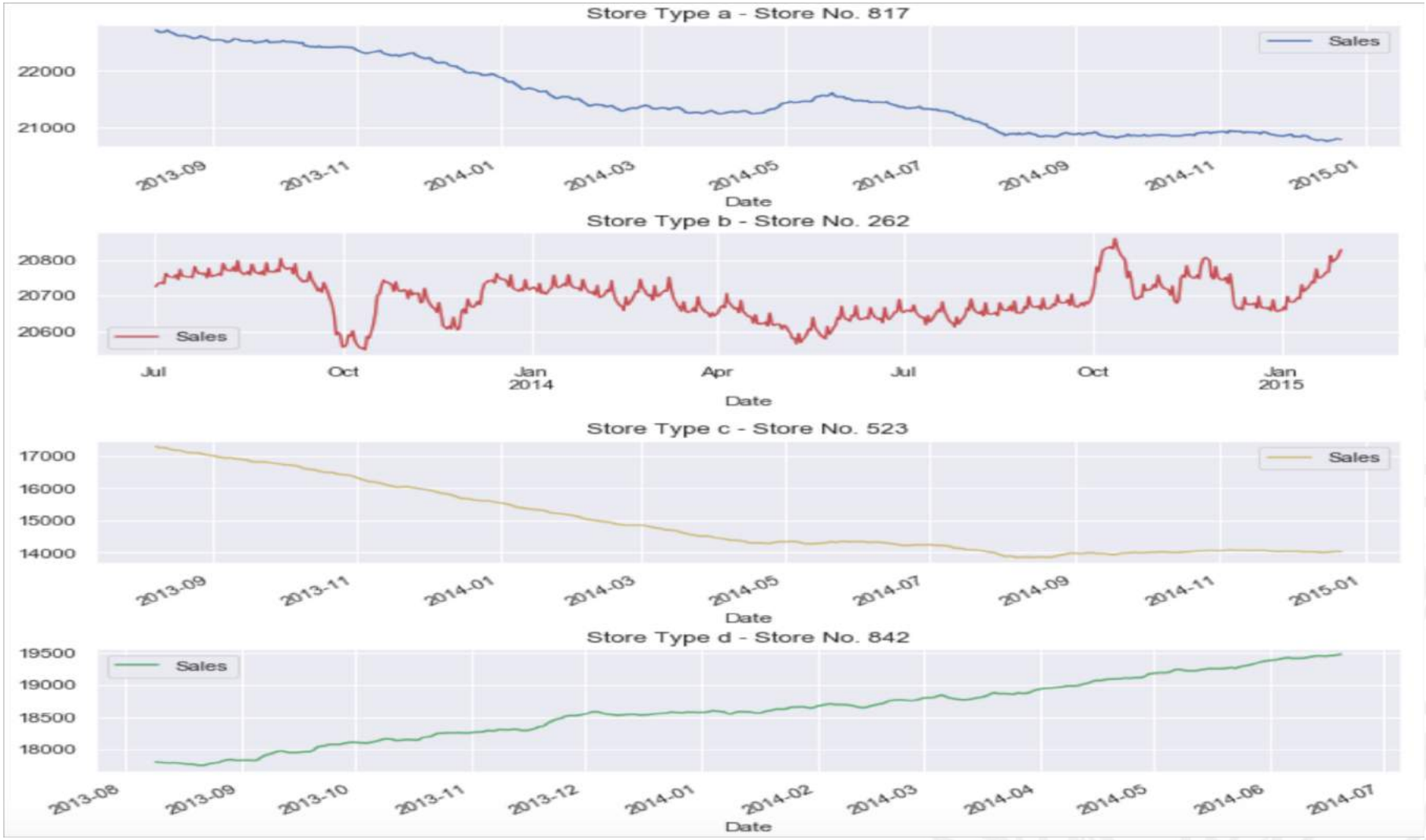


SALES VARIATION PER DAY PER STORE TYPE

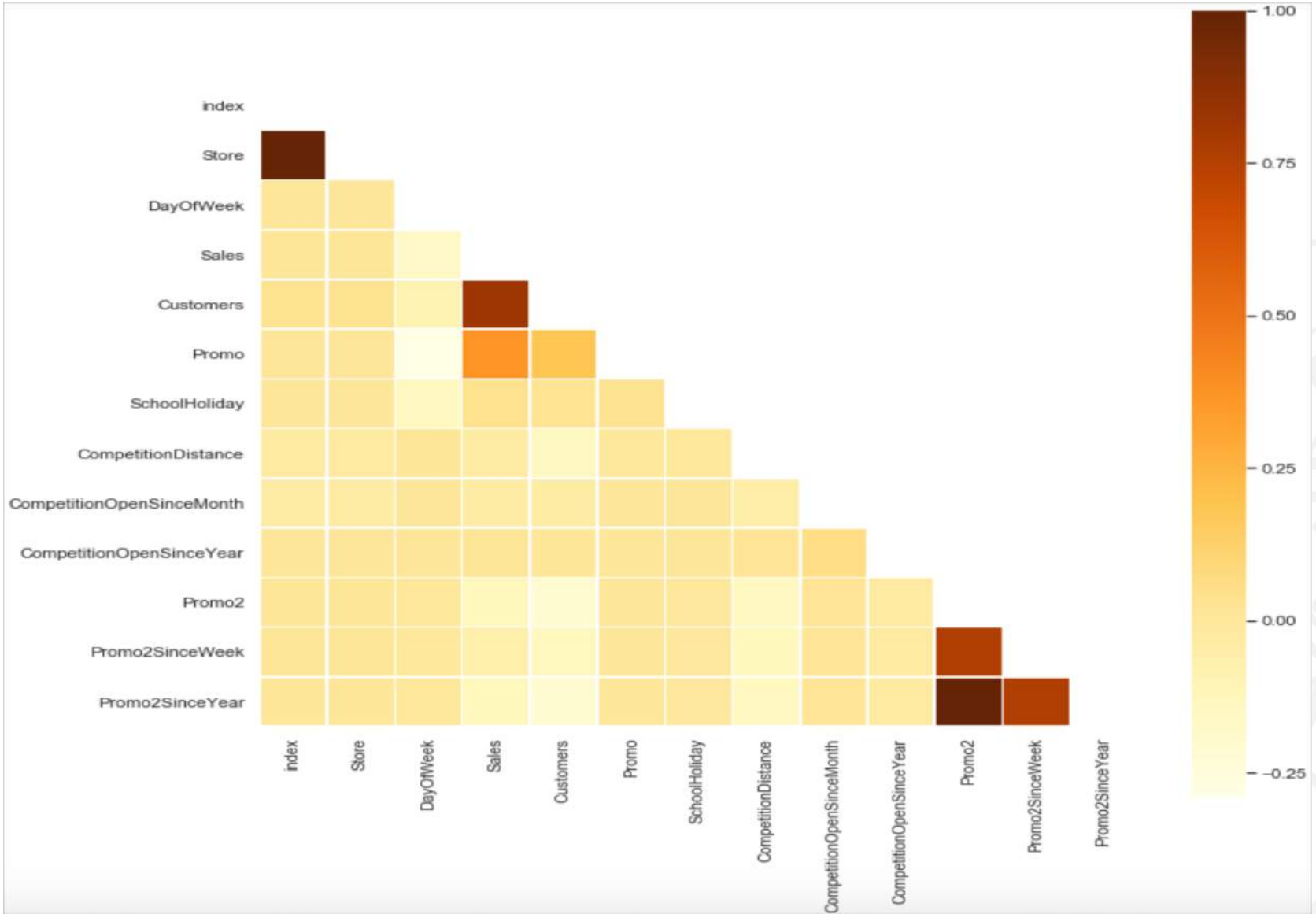
MONTHLY SALES DISTRIBUTION OF STORES WITH HIGHEST SALE



SEASONAL TREND OF STORES WITH HIGHEST SALE



CORRELATION MATRIX



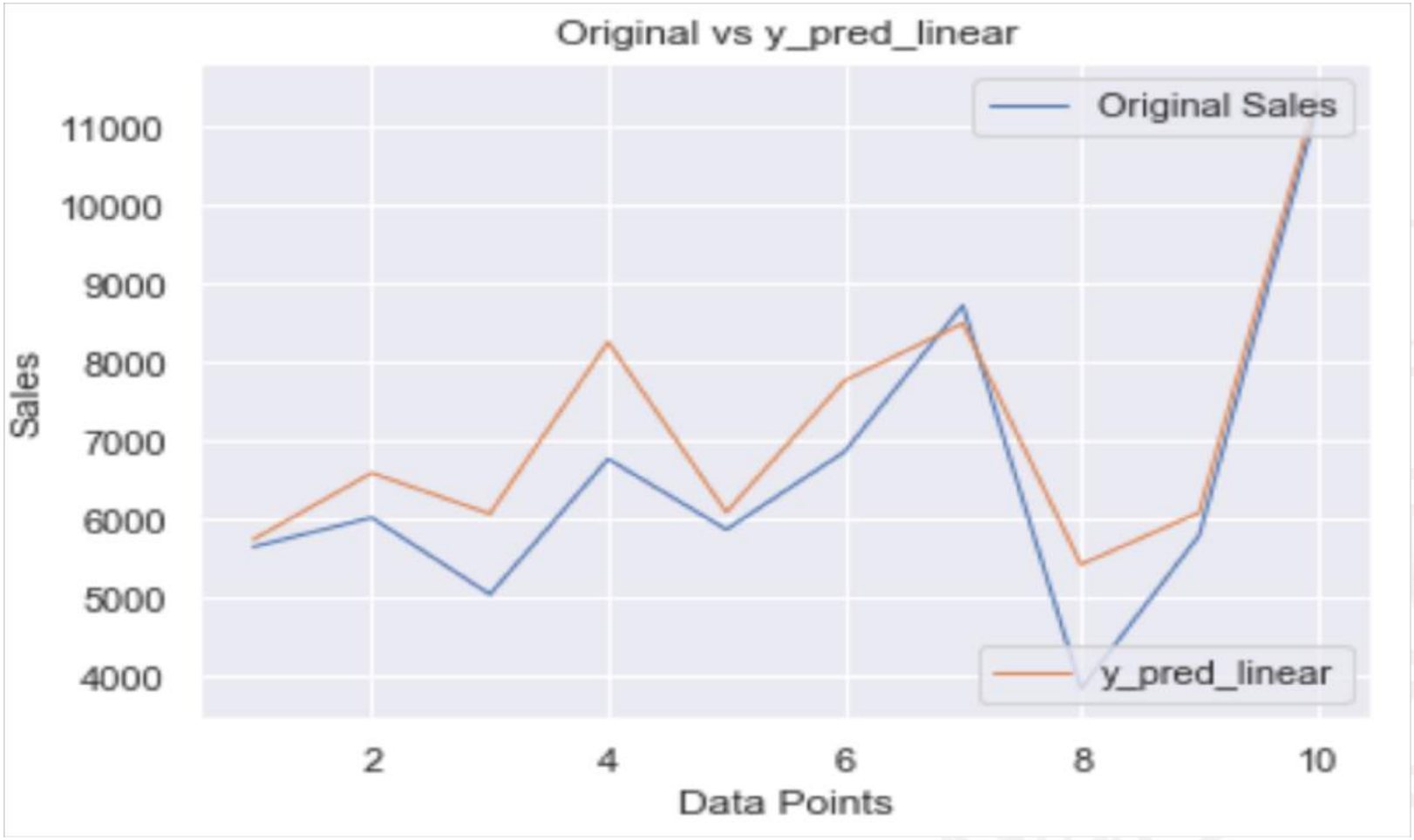
FEATURE ENGINEERING

- Parameters like competitionDistance, Promo and number of customers etc. were more important variables in comparison to other variables.

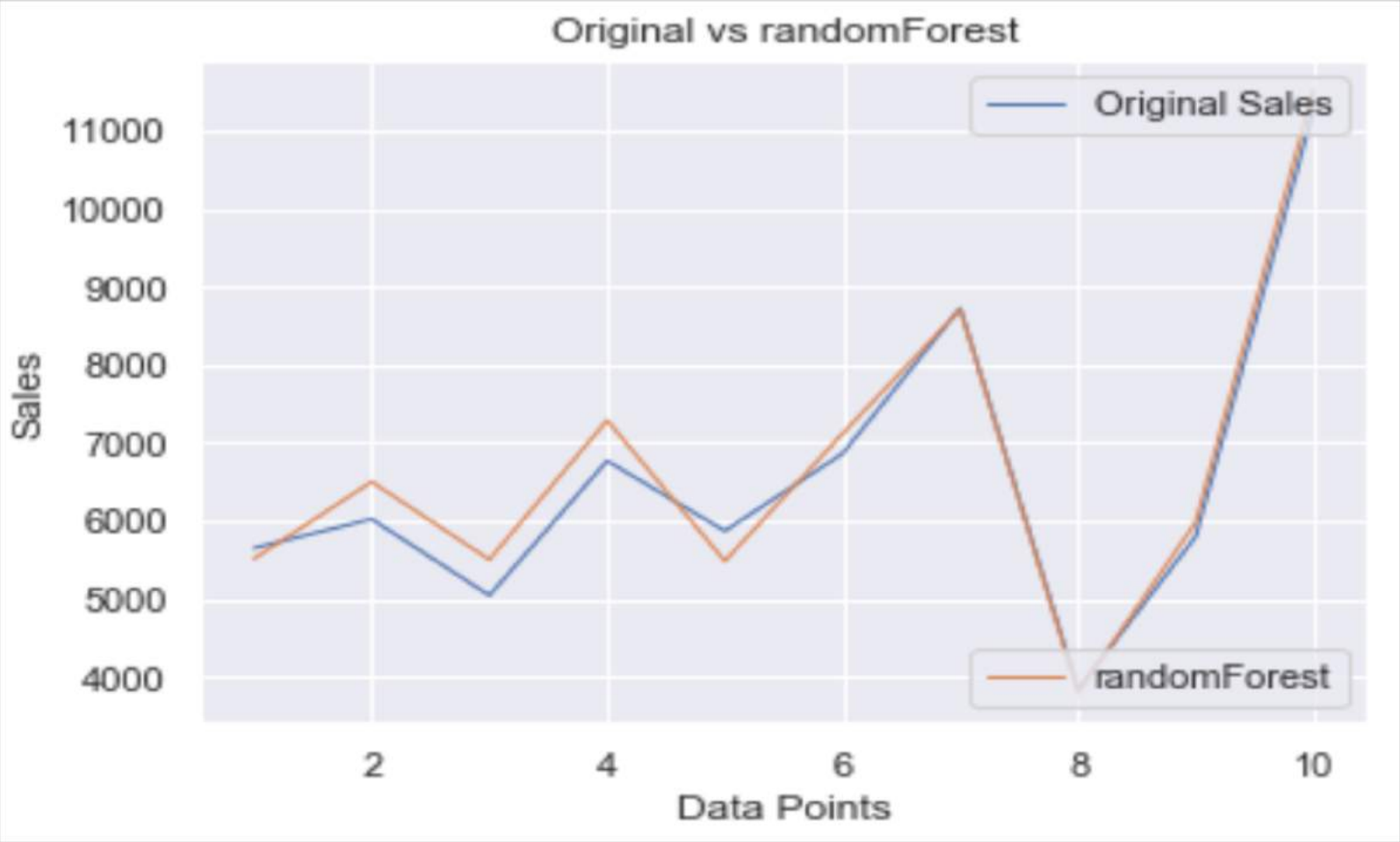
- Store type, dayOfWeek, StateHoliday, PromoInterval and assortmentType were categorical in nature so instead of converting them into random numbers, dummy variables were introduced.

Assortment_c	StoreType_a	StoreType_b	StoreType_c	StoreType_d	StateHoliday_0	StateHoliday_a	StateHoliday_b	StateHoliday_c	PromoInterval_0	PromoInterval_Feb,May,Aug,Nov
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1	0

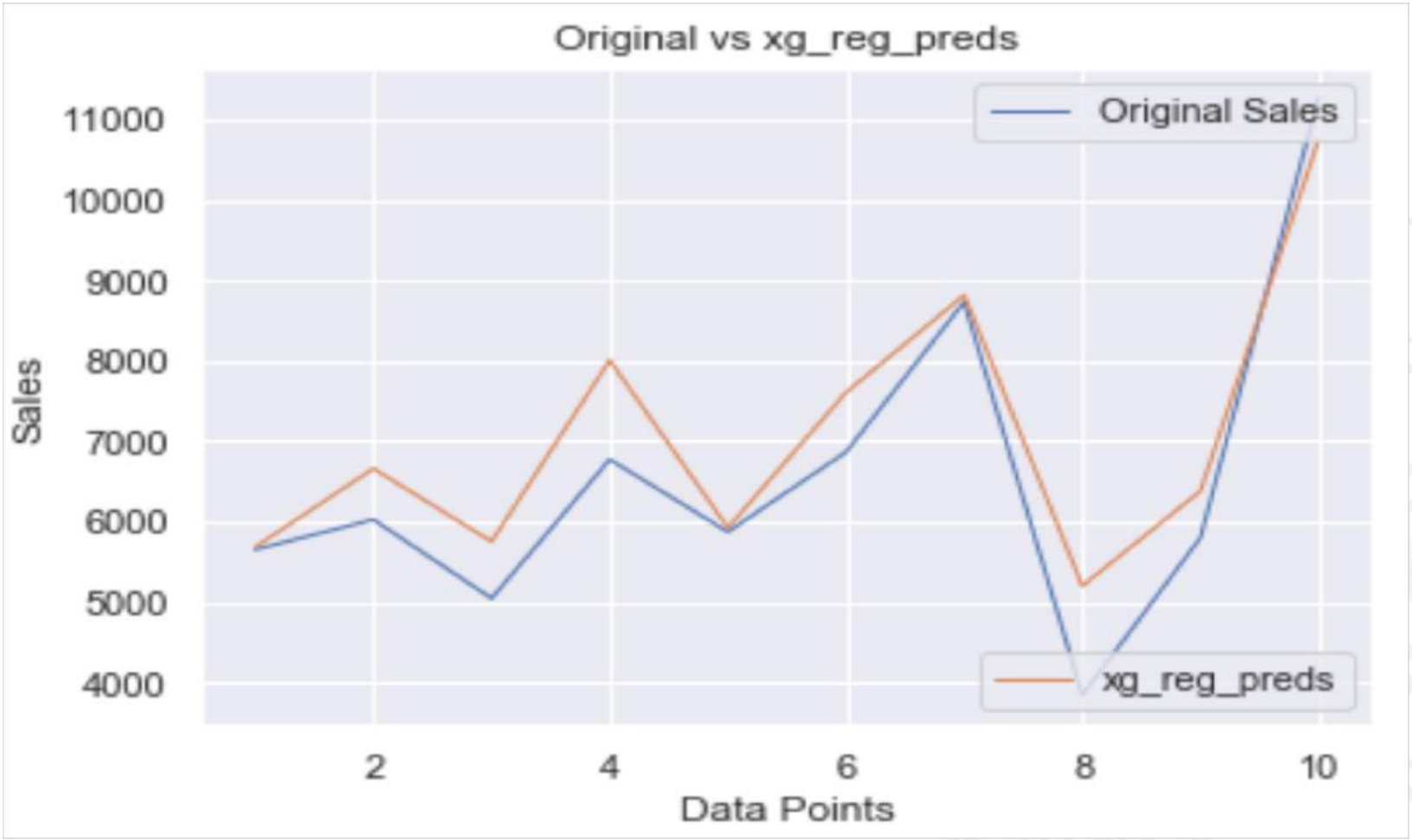
LINEAR REGRESSION



RANDOM FORESTS

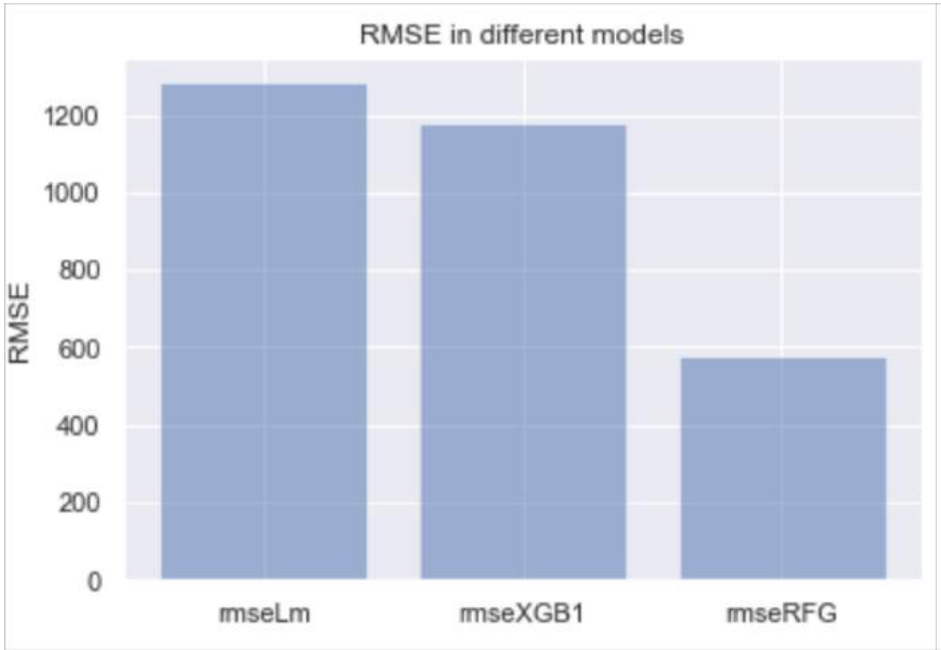


eXtreme Gradient Boosting(XGBoost)

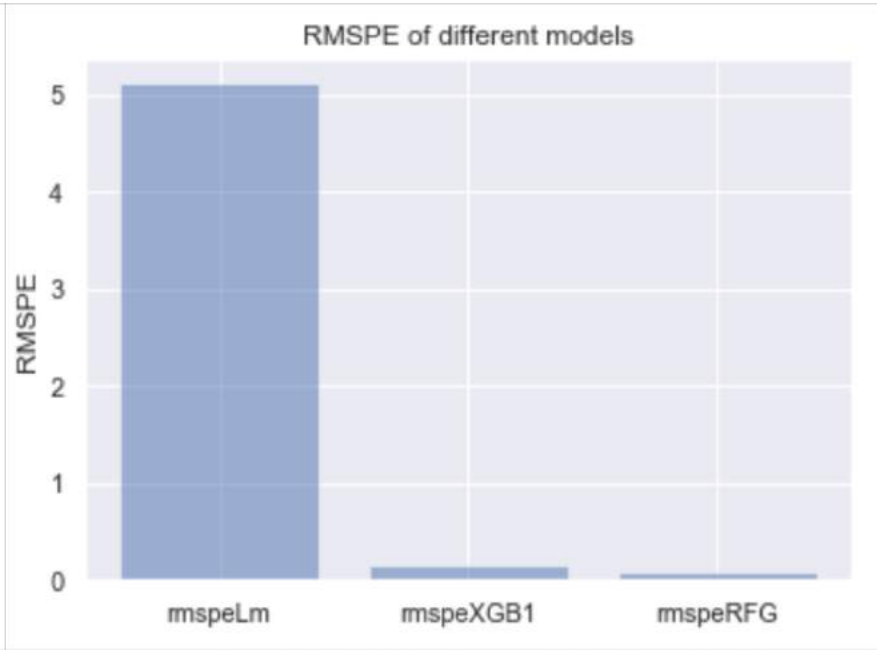


RESULTS

RMSE

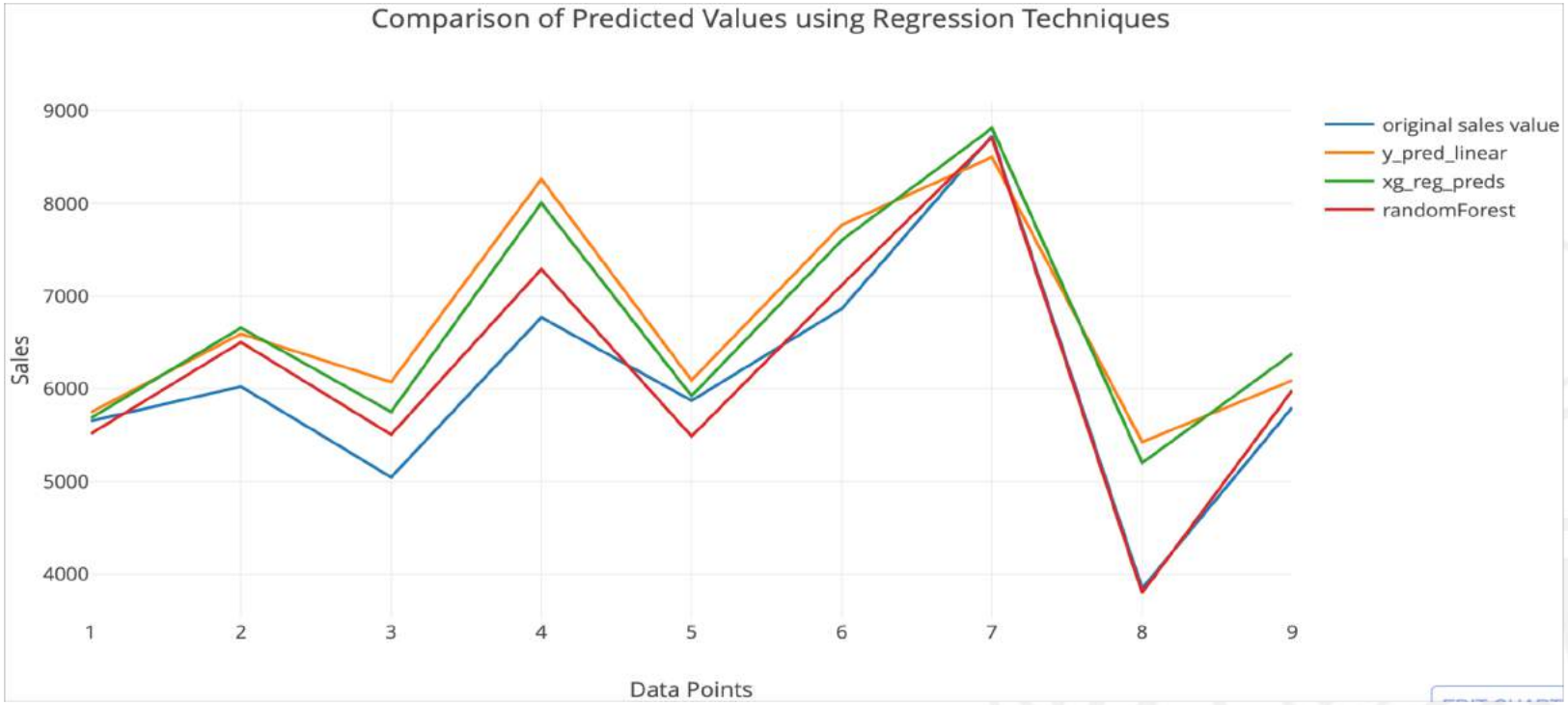


RMSPE



Inference: RMSE seems to be quite high. So used RMSPE and it was quite high for linear and lower for XGBoost regression while lowest for random forest regression.

COMPARISON



Inference: Since the data set was quite huge , plotted different predictions using various regression techniques against small subset of stores. Random forest was fitting closely.

CHALLENGES

- Handling large amount of sales data (10,17,210 observations on 52 variables)
- Prediction of sales for individual stores(out of 1115) and most of stores have different pattern of sales. A single model cannot fit to all stores.
- Handling large number of categorical variables.
- Time Series Analysis.

LEARNINGS

- Exploring large datasets using visualization tools.
- Concept of Dummy Variables
- Learnt the application of Time Series, and XG Boost.

CONCLUSION

- Rossmann stores should focus on opening and operating stores with the aim of increasing customer count. The associated factors could be visibility, population density etc.
- Competition's nearness adversely affects sales and thus the location of store should accordingly be chosen.
- The Promo2 program needs a revision, as it is not yielding expected results. Although it is a weak variable influencing sales, the stores that signed up for it are not performing as expected. There may be other factors working against them such as staff training, customer satisfaction etc. Such data needs to be analyzed and promotion program or other offers should thereupon be designed.

FUTURE SCOPE

- Usage of neural networks might help in the reduction of error rate.
- Explore Time Series further.
- Explore other algorithms.

REFERENCES

- https://www.kaggle.com/c/rossmann-store-sales/data
- Kaggle competition Forum
- https://xgboost.readthedocs.io/en/latest/
- https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/
- Google, of course.


THANKS!