Capstone Project - 2

RETAIL SALES PREDICTION

Predicting Sales of a major store chain Rossmann

Team Members

Ajith Varpe
Harisha Chennozwala
Niharika Soni
Satya Prakash

AGENDA

1. Rossmann Overview
2. Problem Formulation & Business Objective
3. Data Overview - variable description
4. Data Wrangling
5. Exploratory Data Analysis
6. Machine Learning Data Modeling
7. Experiments & Results
8. Model Selection Conclusions
9. Feature Importance
10. Business Insights & Recommendations
11. Challenges & Conclusion

The **Rossmann** is one of the largest drugstore chains in Europe with around 56,200 employees and more than 4000 stores. In 2019 Rossmann had more than €10 billion turnover in Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain. The company was founded by Dirk Rossmann in Germany. The product range includes up to 21,700. In addition to drugstore goods with a focus on skin, hair, body, baby and health, Rossmann also offers promotional items ("World of Ideas"), pet food, perfume range, a photo service and a wide range of natural foods and wines. Rossmann has 29 private brands with 4600 products (as of 2019).

Problem Formulation & Business Objective

- **Client:** Rossmann is one of the largest drugstore chains in Europe & largest in Germany.

- **Objective**: Sales Forecast - Predict sales for 6 weeks in advance given the data.

- **Challenges:** Provided with historical sales data for 1115 different stores and has been provided from Jan'2013 through July' 2015(2 years 7 months).

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

- To provide useful insights which can help increase in productivity.
- What model is appropriate to predict sales?

Data Overview

S.no	Data Set	Variables	No. of Variables	No. of Observation
1.	Rossmann Stores Data.csv	store, day of week, date, sales, customers, open, promo, state holiday, school holiday	9	1017209
2.	Stores.csv	store, storetype, assortment, competition distance, competition open since month, promo2, promo2 since week, promo2 since year, promo interval	10	1115

Variables Description

S.No	Variables	Measurement Scale	Possible Values
1.	Store	Nominal	1 to 1115
2.	DayofWeek	Nominal	1,2,3,4,5,6,7
3.	Date	Interval	1/1/2013 to 7/31/2015
4.	Sales	Ratio	0 to 41551
5.	Customers	Ratio	0 to 7338

Variables Description

S.No	Variables	Measurement Scale	Possible Values

6.	Open	Nominal	0(Closed)
			1(Open)

7.	Promo	Nominal	0(No Promotion)
			1(Offering Promotion)

8.	State Holiday	Nominal	a.Public Holiday b.Easter Holiday c.Christmas Holiday 0. None

9.	School Holiday	Nominal	0(No)
			1(Yes)

Variables Description (cont..)

S.No	Variables	Measurement Scale	Possible Values
10.	Store Type	Nominal	a, b, c, d(Store Models)
11.	Assortment	Nominal	a. Basic b. Extra c. Extended
12.	Competition Distance	Ratio	20-75860
13.	Competition Open since month	Interval	1(Jan) To 12(Dec)

Variables Description (cont..)

S.No	Variables	Measurement Scale	Possible Values
14.	Competition Open since year	Interval	1900-2015
15.	Promo2(Long Term Promotion)	Nominal	0,1
16.	Promo2 since week	Interval	1-50
17.	Promo2 since year	Interval	2009-2015
18.	Promo Interval	Ordinal	(jan, apr, jul, oct) (feb, may, aug, nov) (mar, jun, sept, dec)

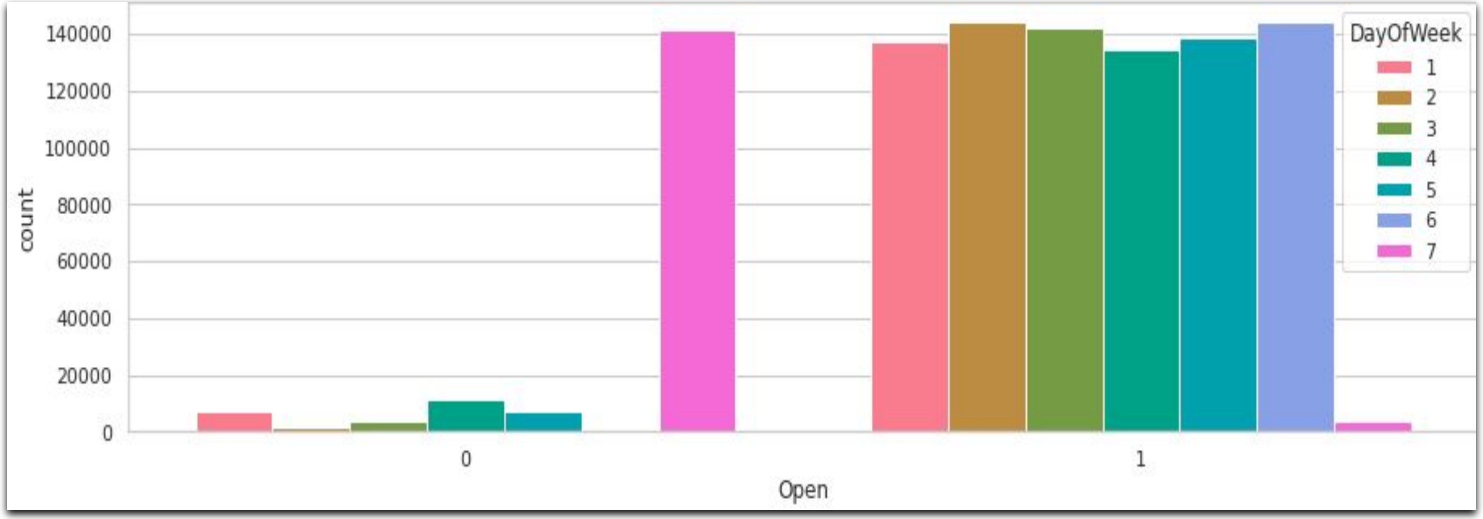
Data Wrangling

- Merged store information and historical sales data. Store type and Assortment is merged into each entry of historical sales data.
- Combined Promo2, Promo2SinceWeek, Promo2SinceYear and Promointerval to a promotion 2 indicator in historical sales data. The indicator indicates on a certain day whether a certain store is on promotion 2.
- Similarly, we combined CompetitionDistance, CompetitionOpenSinceMonth, CompetitionOpenSinceYear to a competitor indicator. The indicator indicates on a certain day whether a certain store has a competitor.
- We created a Month and Year feature based on the Date feature. Month and Year are used as features, since they correlate with sales data.
- The final training dataset used includes the following features.

StoreID	**Open**	**Promo2 indicator**
DayOfWeek	**StateHoliday**	**Store Type**
Month	**SchoolHoliday**	**Assortment**
Year	**Promo**	**CompetitionDistance**

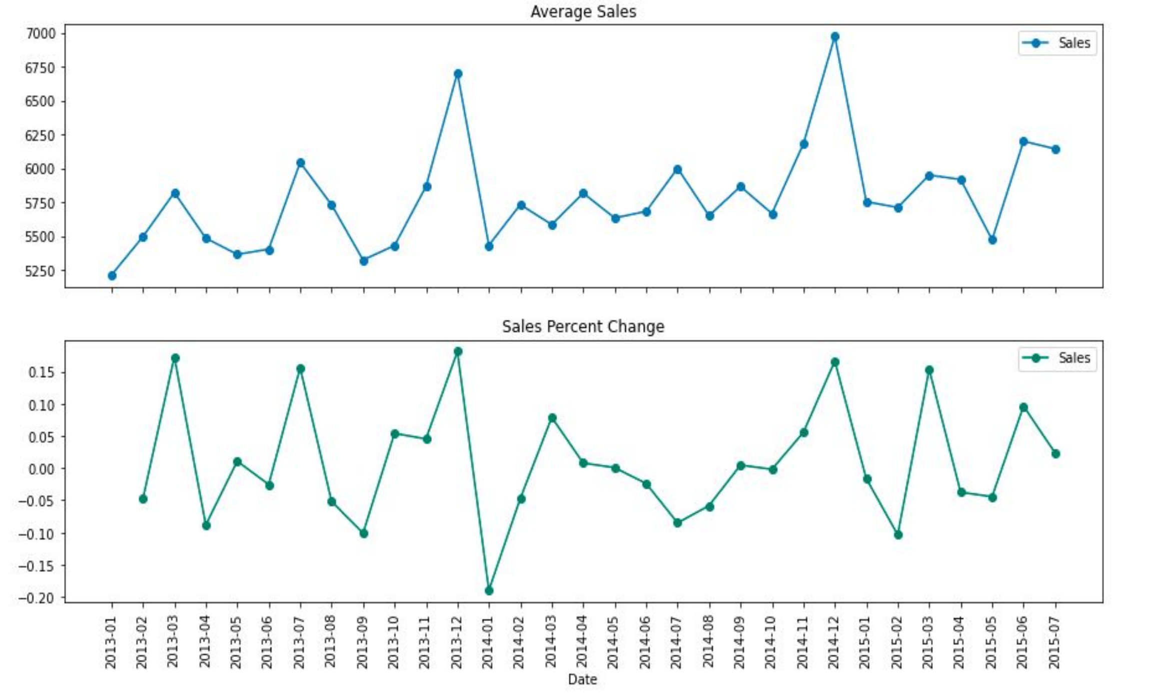
Exploratory Data Analysis

Open



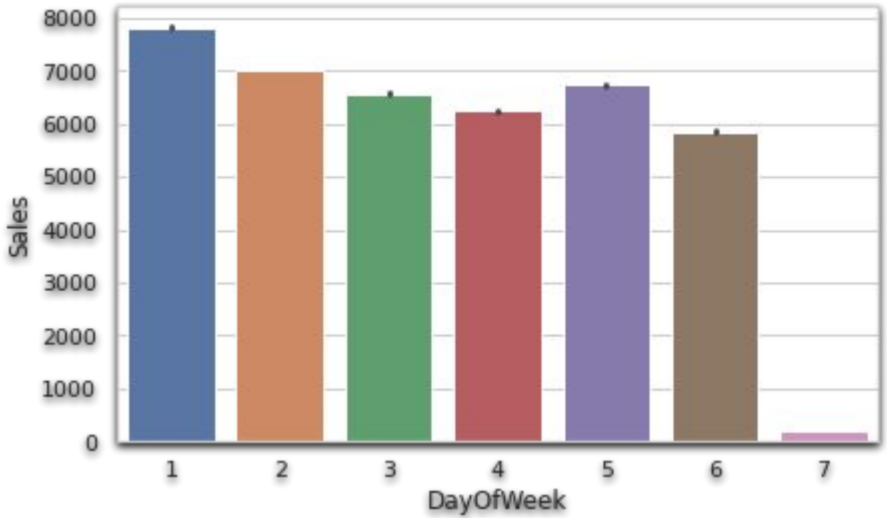
"Open" indicates if this store is open or not on a given specified day. It clearly shows that most of the stores remain closed during Sundays. Some stores were closed on weekdays too, this might be due to State Holidays as stores are generally closed during State Holidays and opened during School Holidays.

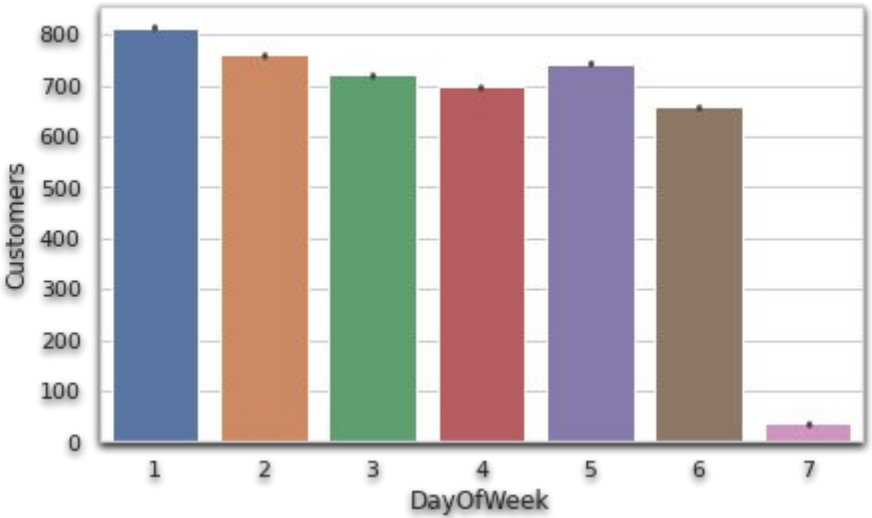
Average Sales & Sales percent change



Christmas and New Year lead to increase in sales. As Rossmann Stores sells health and beauty products, it may be guessed that during Christmas and New Year people buy beauty products as they go out to celebrate and, this might be the cause of sudden increase in sales.

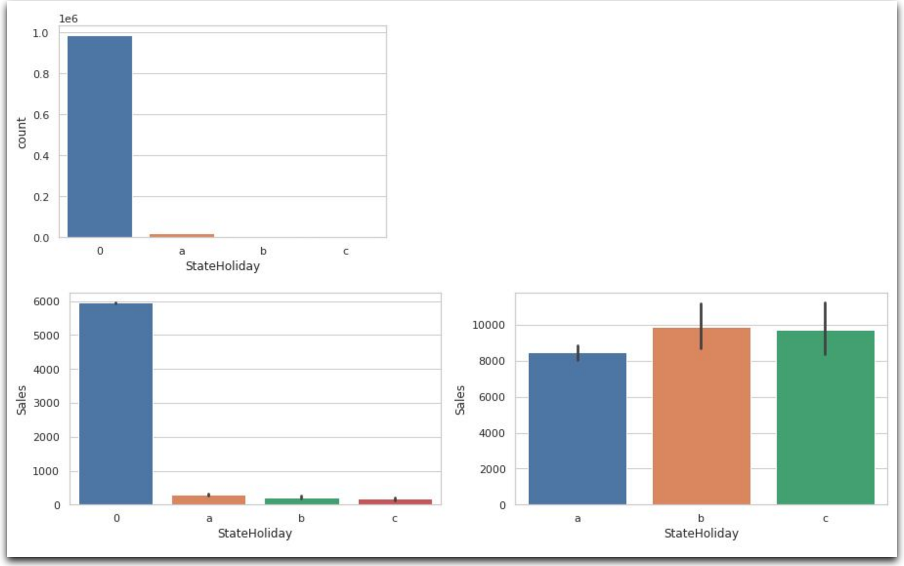
Day of week

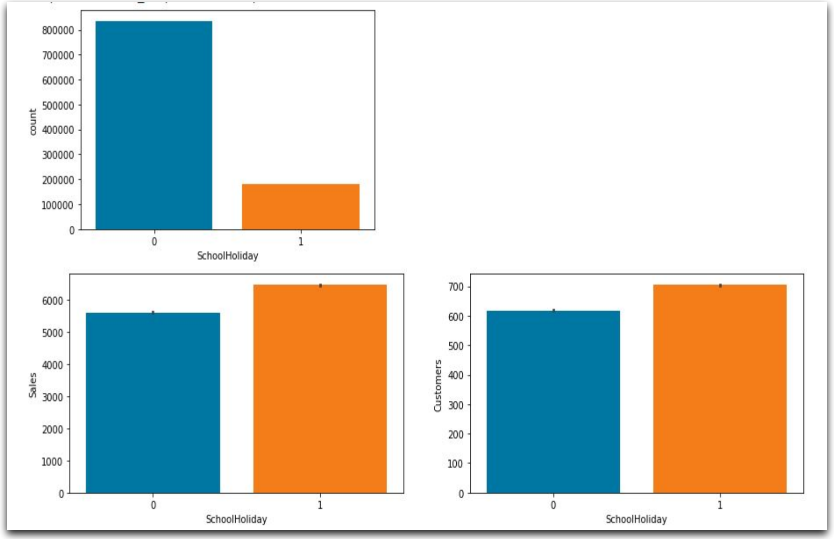




DayOfWeek for both Sales and Customers are very less on Sundays as most of the stores are closed on Sunday. Also, Sales on Monday are the highest in the whole week. This might be due to the fact that stores are closed on Sundays.

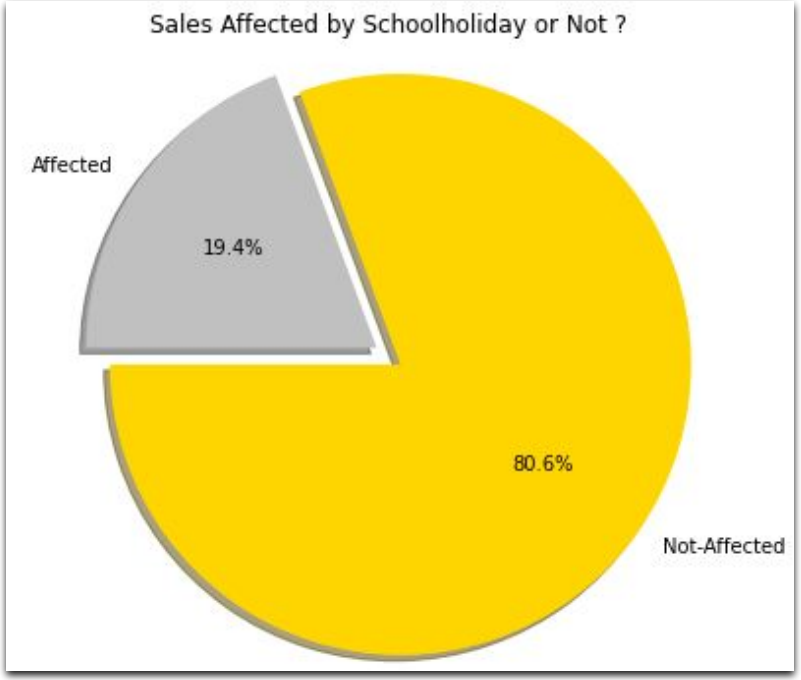
Sales & customers according to State & School Holidays





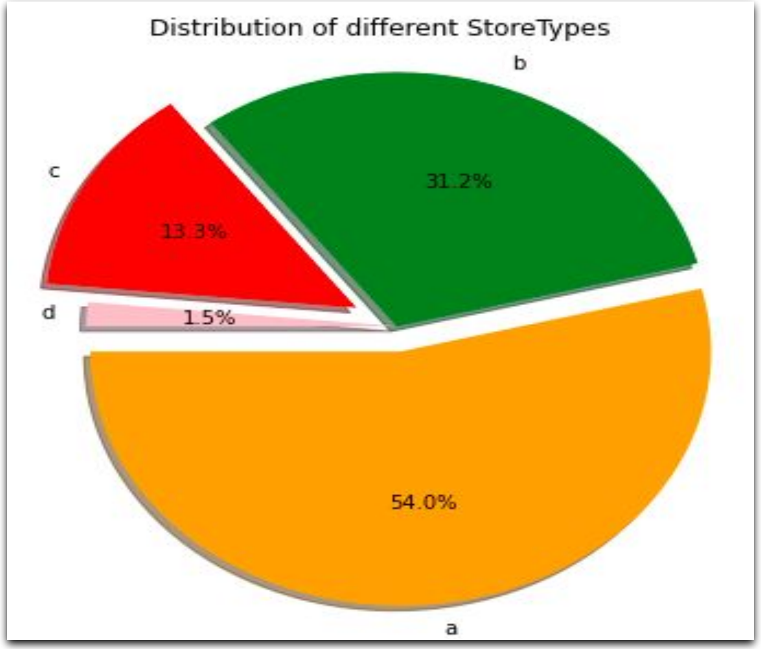
Most of the stores remain closed during State and School Holidays. The number of stores opened during School Holidays were more than those opened during State Holidays. And the stores which were opened during School holidays had more sales than normal.

Sales affected by school holiday or not?



On examining the effect of school holiday on sales, we can see the impact is not so significant.

Sales by Store types

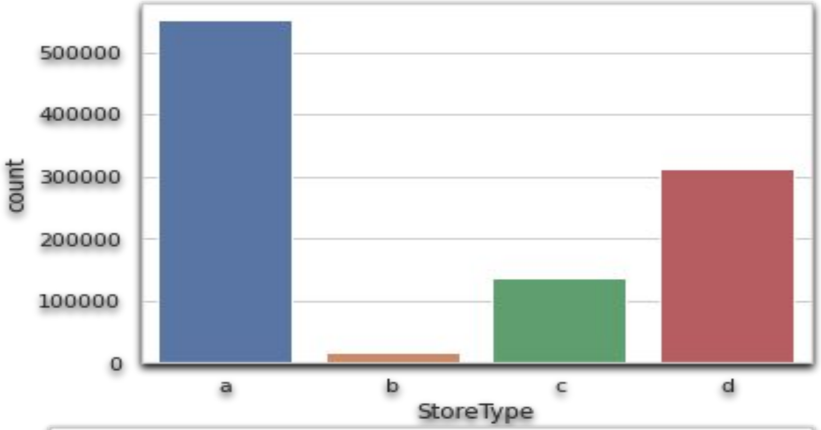


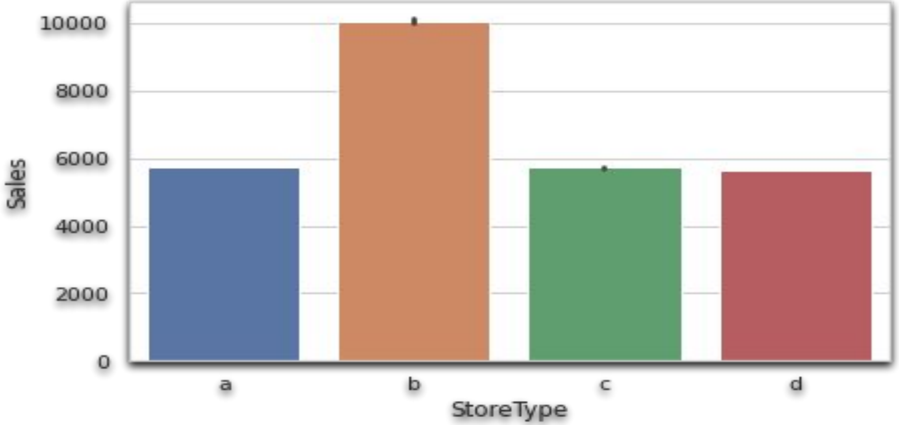
We can see that stores of “type A” have a higher amount of total distribution of storetypes. StoreType D goes on the last place in both Sales and Customer.

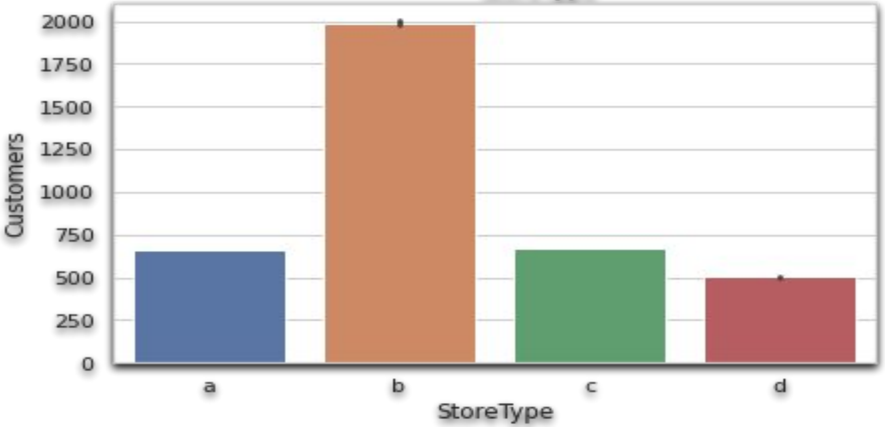
StoreType Vs average sales and customers

In store type B & D has the highest average sales & customers that is likely hyper Rossmann branches.

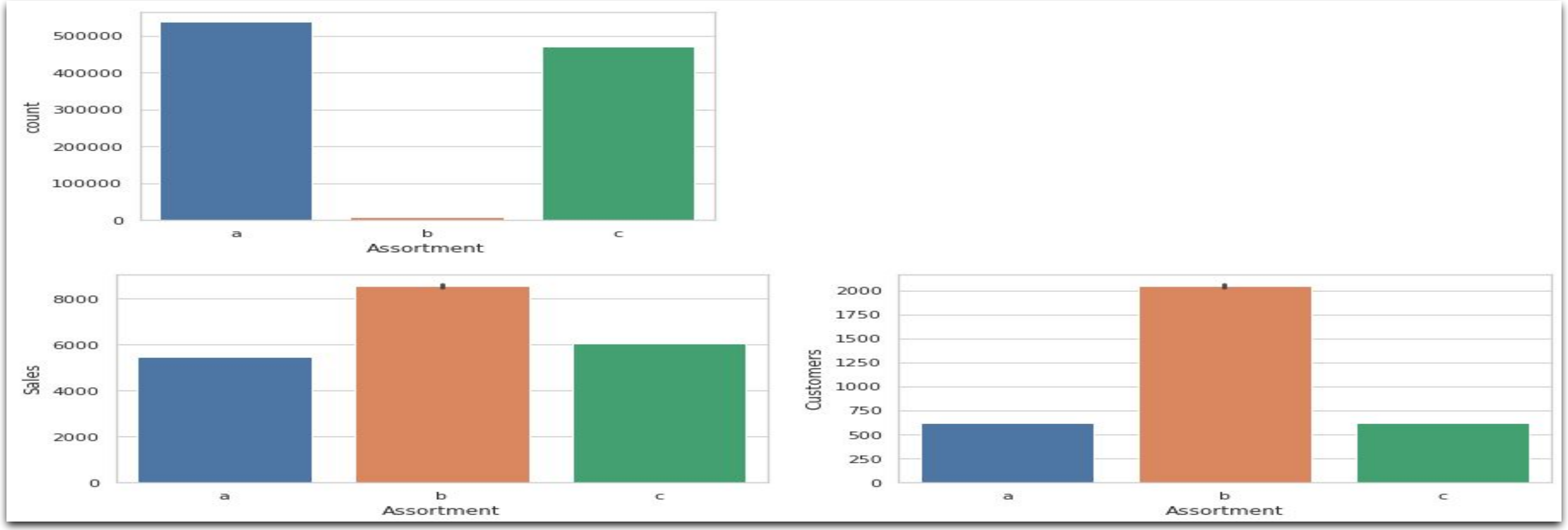
Storetype A would be smaller in size but much more present.





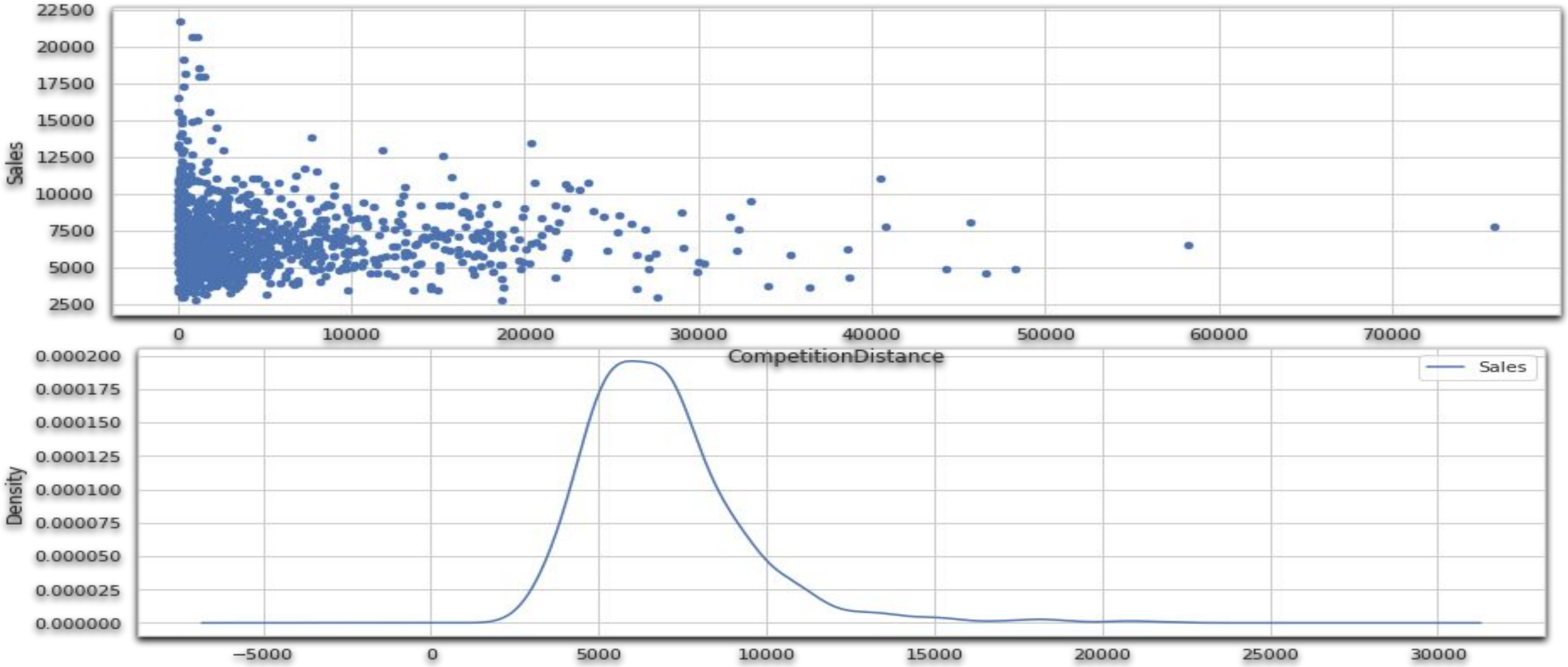


Assortment & Assortment Vs average sales and customers



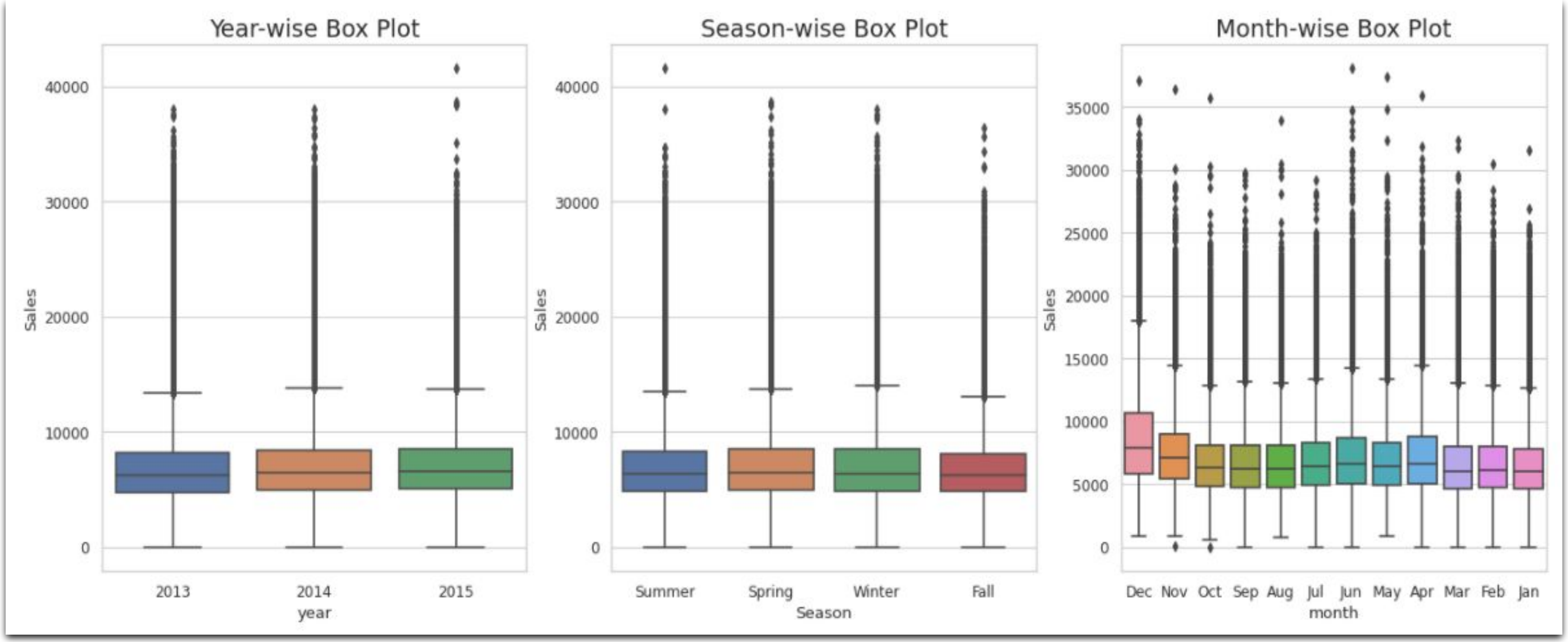
Most of the stores have either an assortment type or c assortment type. Interestingly enough, assortment type B has maximum sales and customers.

Competition Distance



The stores that are the furthest have the highest average sales and number of customers. Drop in Sales observed as the competition opens. We can clearly observe that most of the stores have their competition within 5km range.

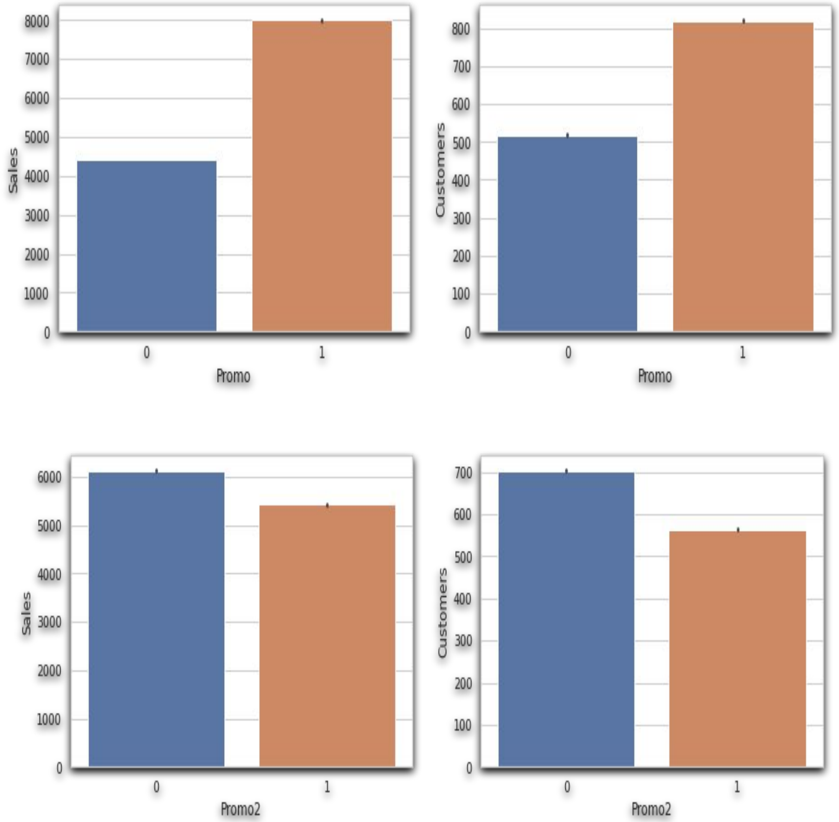
Year, Month and Season wise Sales



We observe that Santa Claus has a special blessing on Rossmann Stores' which means in the month of December sales increases.

Promo, Promo 2 effect, Average sales & Customers

We can see that both Sales and Customers increase by a significant amount during Promotions. This shows that Promotion has a positive effect for a store.

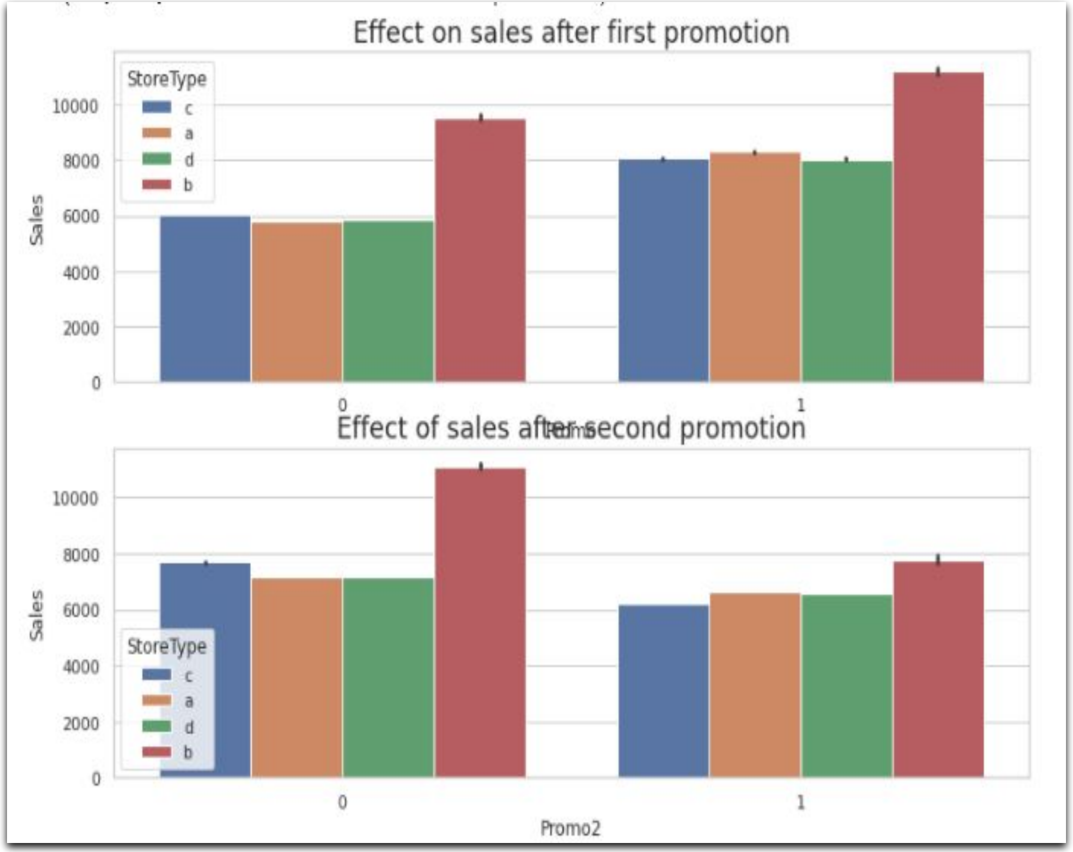


Effect on sales after first and second promotion

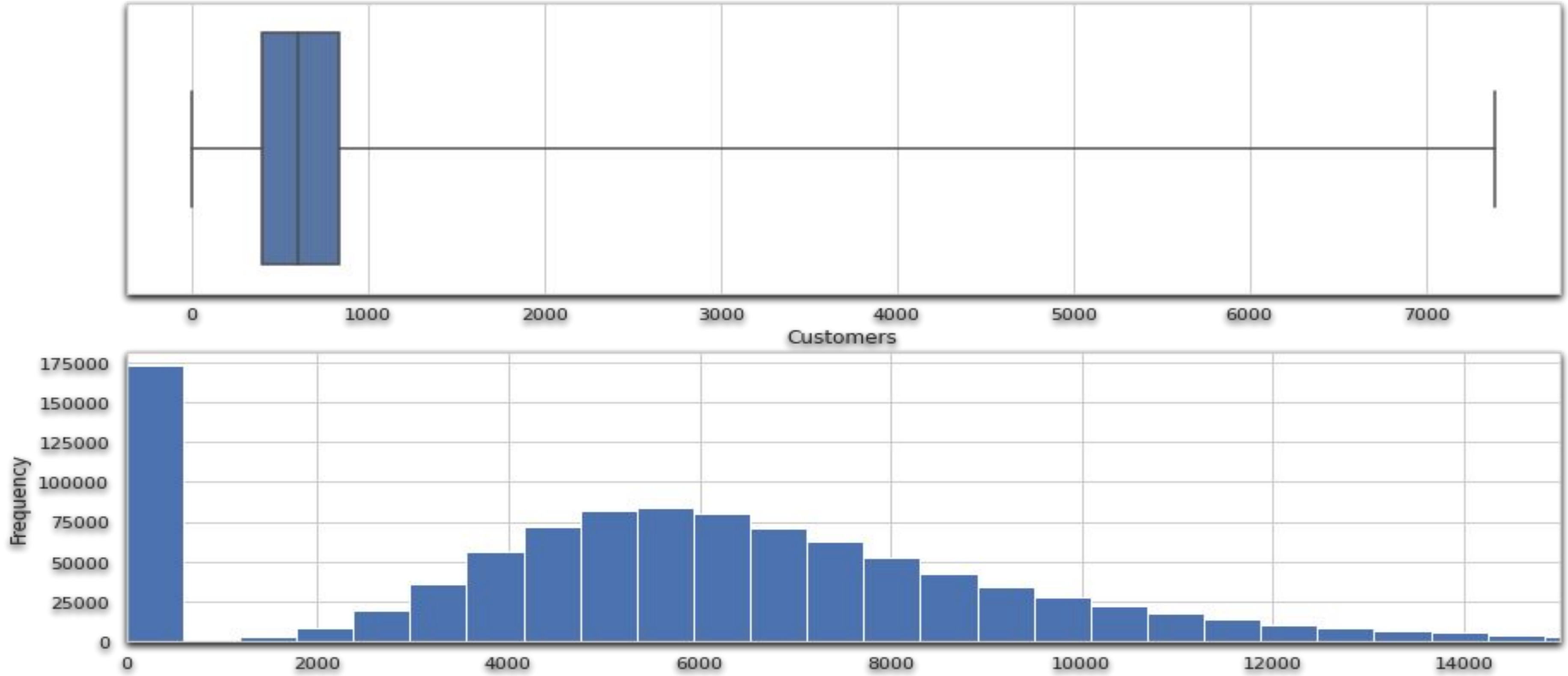
The graph follow the sales trend where sales dropped midyear and increase at the end of the year.

Sales nearly doubled when there was a promo on that day. This is another trend that should be taken into consideration.

First promotion has a positive effect on sales, but the second promotion has a negative impact on sales.

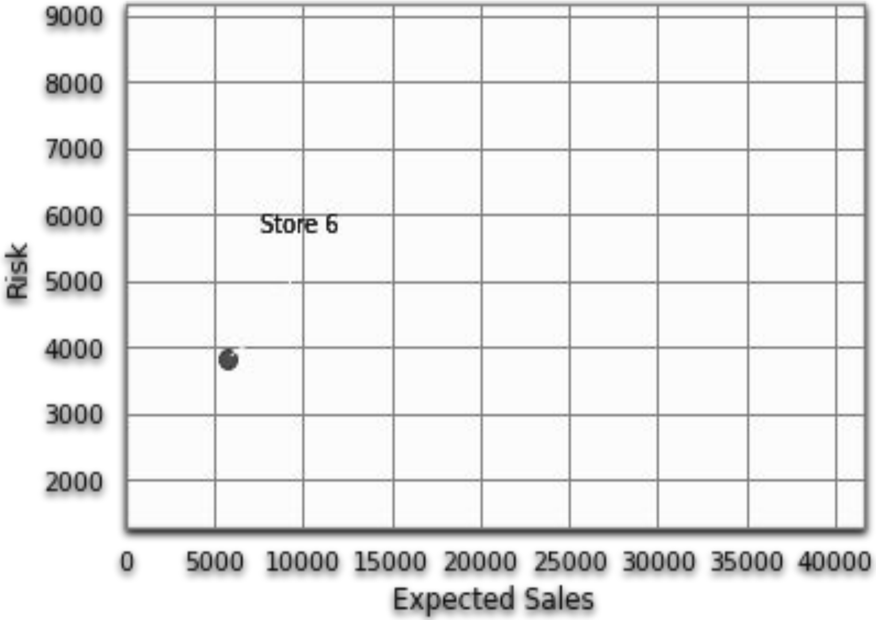


Sales & Customers Distribution



Sales that values with 0 is mostly because the store was closed. Sales is highly correlated to the number of Customers.

Risk Analysis



Forecast not only more probable values of sales but also their distribution. Especially we need it in the risk analysis for assessing different risks related to sales dynamics.

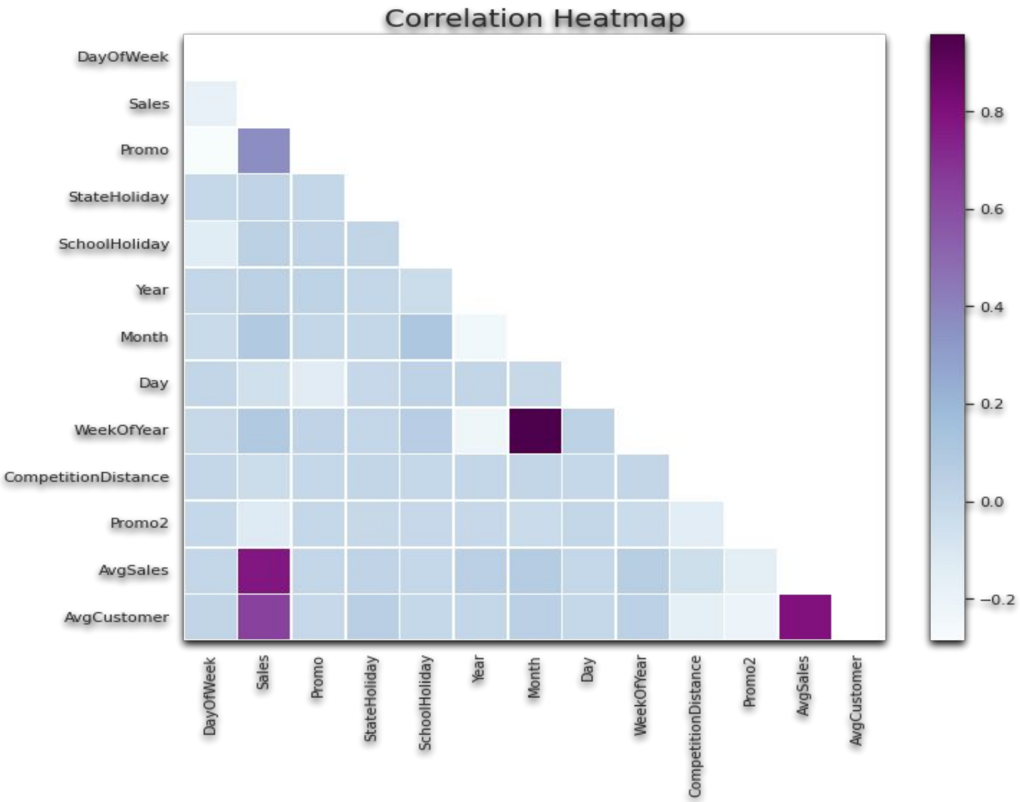
Correlation Heatmap

- Average Customers and Average sales are positively correlated between 0.8

- Sales and Promo (more than 0.2) actually correlate positively

- Sales correlates with Competition Distance(more than 0.1), in a positive manner

- Promo also effects sales positively



Machine Learning Data Modeling (for our Prediction)

1. Linear Regression(OLS)
2. Bayesian Ridge Regression
3. LARS Lasso Regression
4. Decision Tree Regression
5. Random Forest Regression
6. K-Nearest Neighbors Regression
7. Facebook Prophet Model

Evaluation metrics - (Equation is shown below)

MAPE(Mean Absolute Percentage Error) is a measure of how accurate a forecast system is.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|A_i - F_i|}{A_i}$$

Ai = actual value
Fi = forecast value
n = total number of observations

RMSE(Root mean square error) method is used to evaluate the prediction quality.

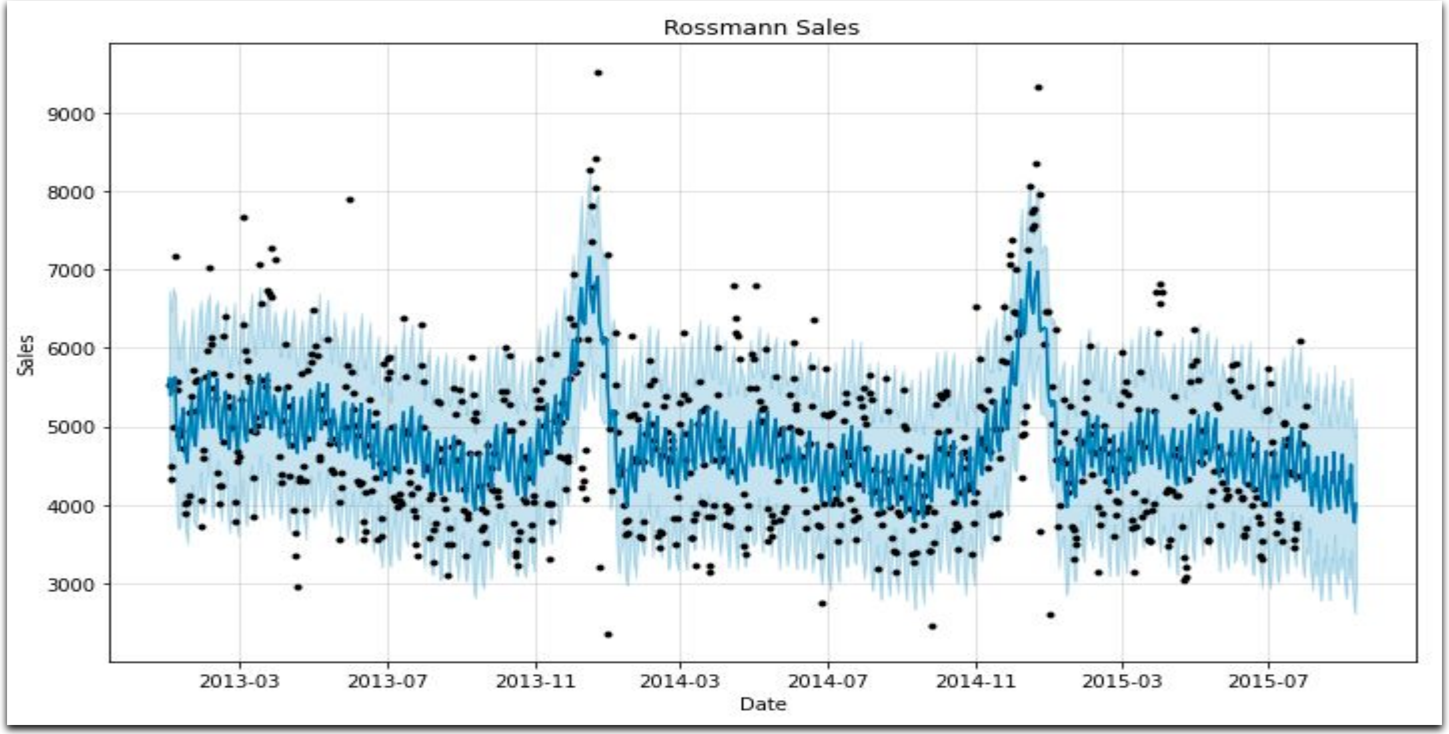
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

N = number of data points
yi = i-th measurement - Predicted sales.
ŷ (i) = the corresponding prediction - Sales.

Experiments & Metrics

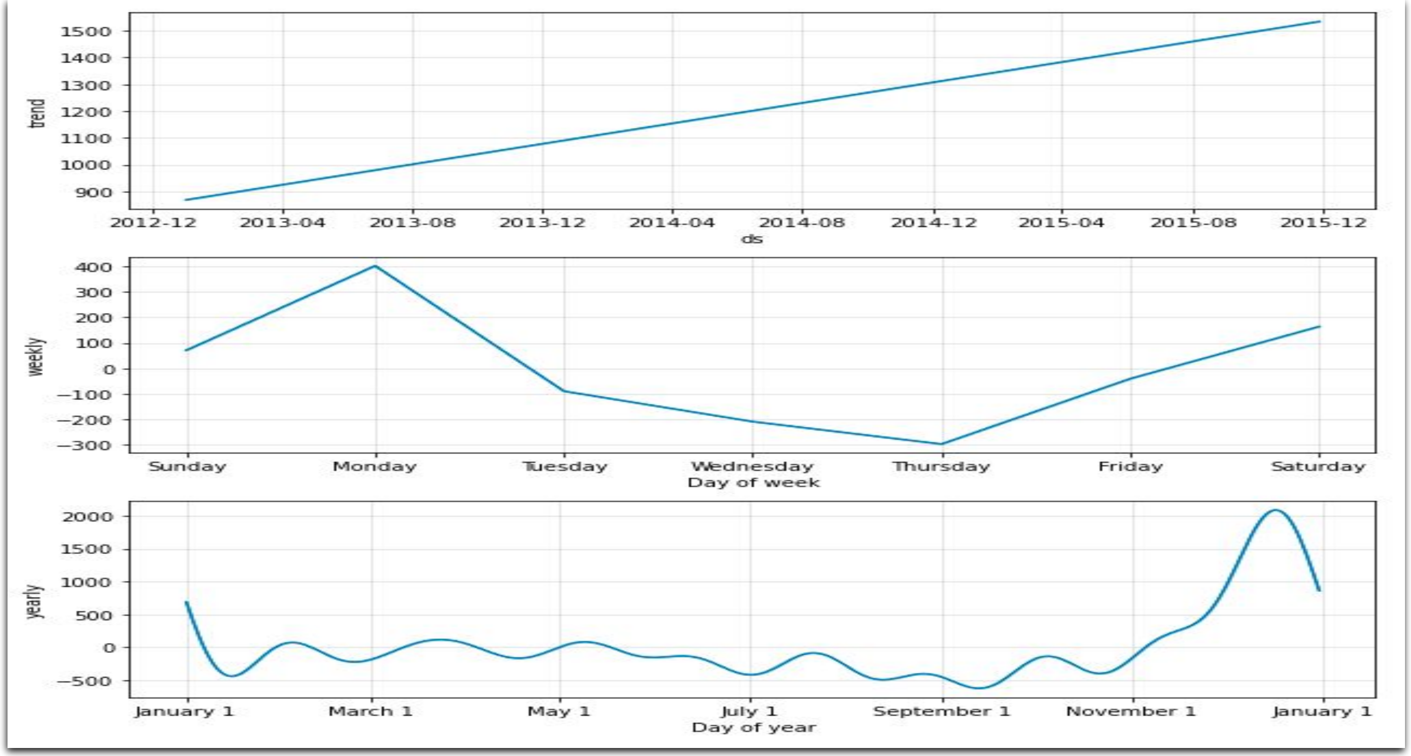
Model	Training MAPE	Testing MAPE	Training RMSE	Testing RMSE	Model Score
Linear Regression	16.95	17.230	1546.46	1563.75	0.7534
Bayesian Ridge Regression	16.95	17.214	1547.10	1562.62	0.7534
LARS Lasso Regression	16.953	17.214	1547.25	1562.73	0.7534
Decision Tree Regression	12.456	14.690	1195.66	1402.53	0.8544
Random Forest Regression	12.456	14.690	1195.66	1402.53	0.9779
K-Nearest Neighbors Regression	22.92	23.86	1915.63	1994.32	0.6275

Facebook Prophet Model



Prophet plots the observed values of our time series (the black dots), the forecasted values (blue line) and the uncertainty intervals of our forecasts (the blue shaded regions). Looks like Prophet has captured the negative trend and the seasonality from this store sales quite well.

Facebook Prophet Model - Weekly & monthly seasonalities trend



The obvious downward trend seems like this store needs some attention from Rossmann. We can clearly see the drop in sales in the fall months and the spike during the Christmas. It is nice to see that this model doesn’t put too much emphasis on the single drop in sales in June 2014. We also get to see the dip in Sales on Thursdays.

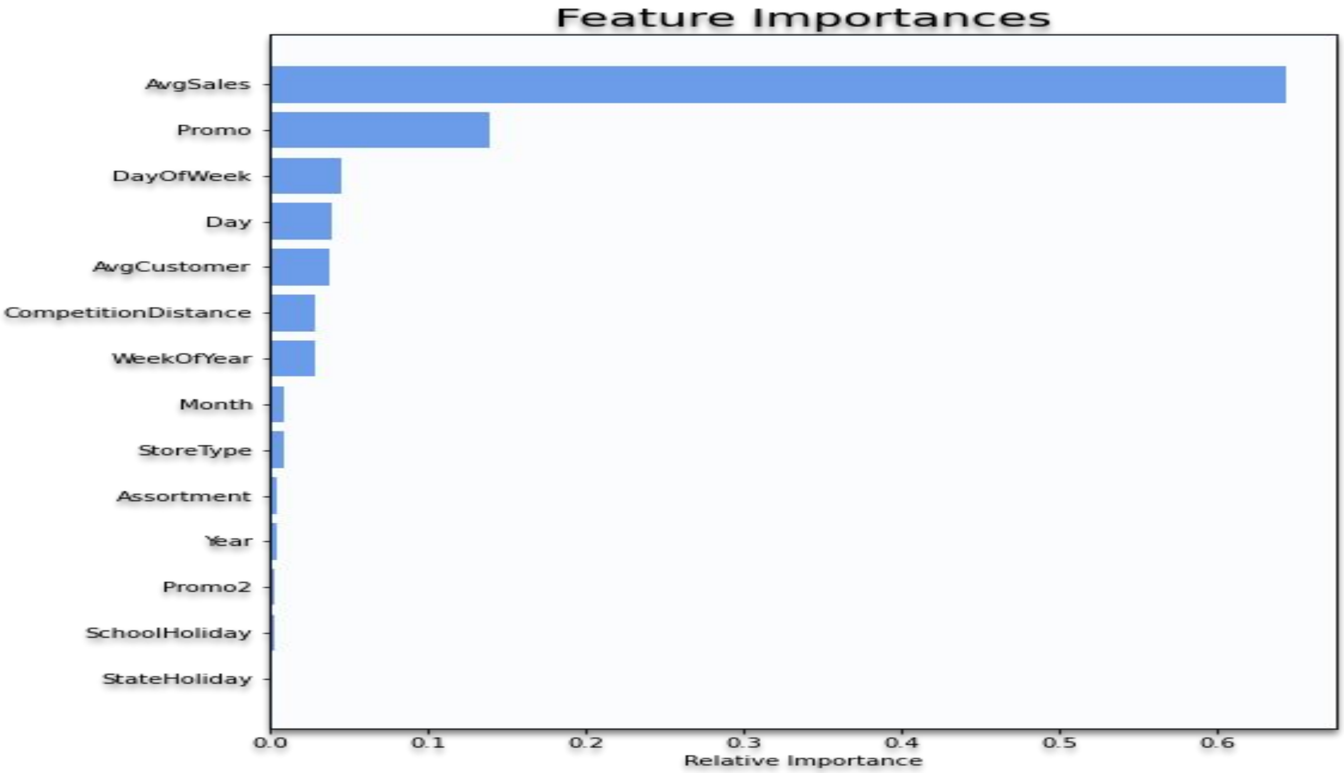
Model Selection Conclusions

- As is shown in the result, among all models, **Random Forest** works the best with the higher model score and least RMSE we have, and provides a reliable prediction of the sales.

- Linear regression, Bayesian Ridge Regression, LARS Lasso Regression, Decision Tree Regression, K-Nearest Neighbors Regression all have their own strengths and limitations.

- However, We have listed out the most significant changepoints in our data. This is representing when the time series growth rate significantly changes, while **Facebook prophet model** calculates the best solid result.

Feature Importance



The important features in all of the Rossmann stores, organized from most important to least. They are equally weighted across all stores to generate all feature importances in the Rossmann dataset.

Business Insights & Recommendations

- Rossmann should focus on increasing the promotional offers per quarter for a,c,d and can minimize for b.
- The most selling and crowded store type is B
- Sales is highly correlated to the number of Customers.
- For all stores, Promotion leads to increase in Sales and Customers both.
- The stores which are opened during the School Holiday have more sales than normal days.
- More stores are opened during School holidays than State holidays.
- Rossman should try to focus on reducing the Promo offers for store type b during StateHolidays as there is no substantial increase in Sales.
- Sales are increased during Christmas week, this might be due to the fact that people buy more beauty products during a Christmas celebration.
- Rossmann can divert some of the Promos from being offered on SchoolHolidays to No SchoolHolidays to maximise the Sales revenue.
- Absence of values in features CompetitionOpenSinceYear/Month doesn't indicate the absence of competition as CompetitionDistance values are not null where the other two values are null.
- After analysing sales using Fourier decomposition, found that there's a little seasonality component in the Sales data.

Conclusion & Challenges

- We have generated sales predictions for the Rossmann store chain 6 weeks in advance using regression models. While conducting EDA for modeling, we found many interesting insights. Although many stores were closed on Sunday, those that were opened saw a great amount of sales revenue. Assortment b had the highest sales, but it was only available in tybe-b stores among 4 types of store in total. We hope that these findings will be helpful for business decision-makers to optimize their profits.
- 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 Promo2C 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 analysed and promotion program or other offers should thereupon be designed.

Challenges:

★ Handling large amount of sales data (10,17,210 observations on 13 variables).

★ Prediction of Sales for individual store (out of 1115)and most of stores have different pattern of sales. A single model cannot fit to all stores.

THANK YOU