USF

UNIVERSITY OF
SOUTH FLORIDA

ISM6137.902S18

Rossmann Sales Data Analysis

Assessing Promotion Impact on Sales and Help Business to Formulate Promotion Strategies

Team members:

Sai Seetha Ram Nomula
Venkata Praveen Kumar Dusi
Gauri Vasant Naik
Anurag Kunchamwar
Laxmi Shrestha

Table of content

- **Problem Significance:**

The Multi-level panel data are becoming increasingly important and prevalent in business and industry. Therefore, we felt that it would be interesting to learn more about the differences and nuances of interpreting panel data. To that end, we decided to use the dataset provided in a Kaggle's competition titled "Rossmann Store Sales".

Rossmann is the largest drugstore in Germany. Moreover, it operates over 3,000 drug stores in 7 European countries. Store sales are influenced by many factors, including promotions, store types, assortment, competition, school and state holidays, seasonality, and locality. There could be many objectives for analyzing sales data such as predicting sales, increasing customer base, future demand etc. One which we find interesting is to assess promotion impact on sales which can help team to design promotion strategies to expect high revenue. These strategies are vital for expanding sales and are designed based on customer purchase patterns, day of week, assortment type and many more. Thus, our goal is to analyze effectiveness of promotions based on past data which can help business team to formulate future strategies.

- **Data Source/Preparation:**

Details of Data provided on Kaggle:

Data Set	Variables	No of Variables	No of Observations
Train	store, day of week, date, sales, customers, open, promo, state holiday, school holiday	9	1017210
Store	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10	1115
Test	id, store, dayofweek, date, open, promo, state holiday, school holiday	8	41089

The train.csv contains historical sales data. The store.csv has all supplemental store information. We have merged the data into single file based on store id. All the store attributes are not present in test.csv data. Thus, we are not including test.csv file in our analysis.

Glimpse of dataset, after merging the variables of 'train' and 'store' dataset:

Variables	Measurement Scale	Possible Values
Store	Nominal	1 to 1115
Dayofweek	Nominal	1,2,3,4,5,6,7

Promo	Nominal	0(No Promotion), 1 (Offering Promotion)
Store type	Nominal	a, b, c, d
Assortment Type	Nominal	a: Basic b: Extra c: Extended
State holiday	Nominal	a: Public Holiday b: Easter Holiday c: Christmas Holiday 0: None

Since this is daily sales data from 2013-15, we are not including promo interval variable. Instead to observe impact of putting promotions on day/week of month/month/week of year, we have created new variables from Date column. Also, created Sales per customer variable from Sales and Customer Variable.

When we proceeded toward data cleaning, we were focused not to use convenient methods of data cleaning such as data imputation with mean/median values, deleting columns to name a few.

While exploring the data, we have found out that for 6 months (2014, July 1 to 2014, Dec 31st) data is missing for 200 stores. If we include this data it will be biased data as number of observations for each store will not be equal for each year. Thus, we removed corresponding 200 stores entirely from our analysis. In addition to this, some people may think that it is not required to include the data when a store is closed, as it is obvious that sales will be zero when store is closed. However, if a store is closed for a day, the next day when it will open we can expect possible hike in sales. If we don't include such data points, our model will lack to learn this pattern.

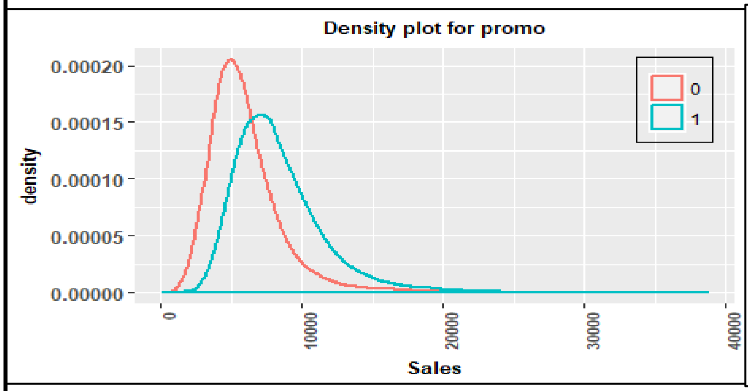
- **Hypotheses:**

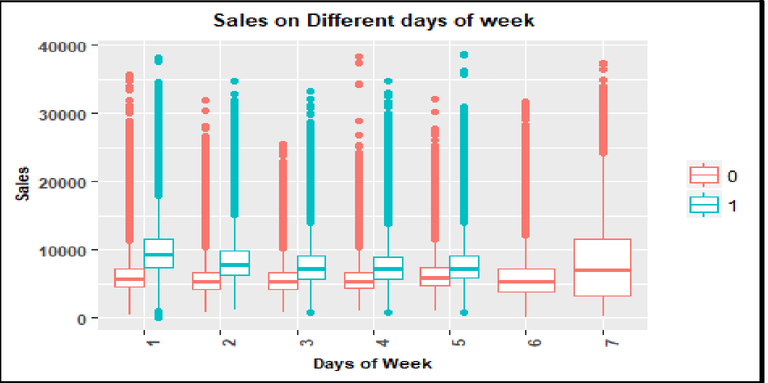
Before drawing hypothesis's, we had one question in mind. How promotions are impacting sales? After careful consideration, we came with following hypothesis which could possibly have impact on sales and assessing them could help team improve promotion strategies.

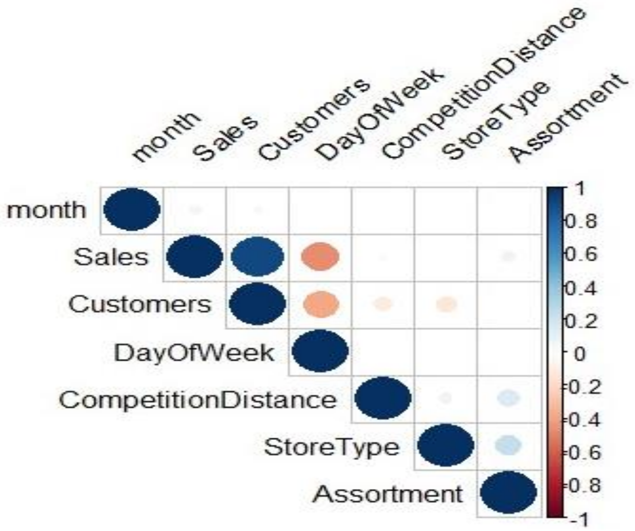
- What is the effect of promotions on different month?
- For what store/assortment type promotion impact is more?
- Does giving promotions boost sales in cases where competitor distance is smaller?

- **Descriptive Analysis:**

We started exploring data to see the impact of promotions on average sales. As expected density plot showed high sales for promotions compared to non-promotion days. We observed that promotions were only kept on weekdays for which they gave comparative high sales to non-promo days. Also, overall sales median is high on Monday and Sunday. In all sales pattern don't vary much from Tuesday to Friday.

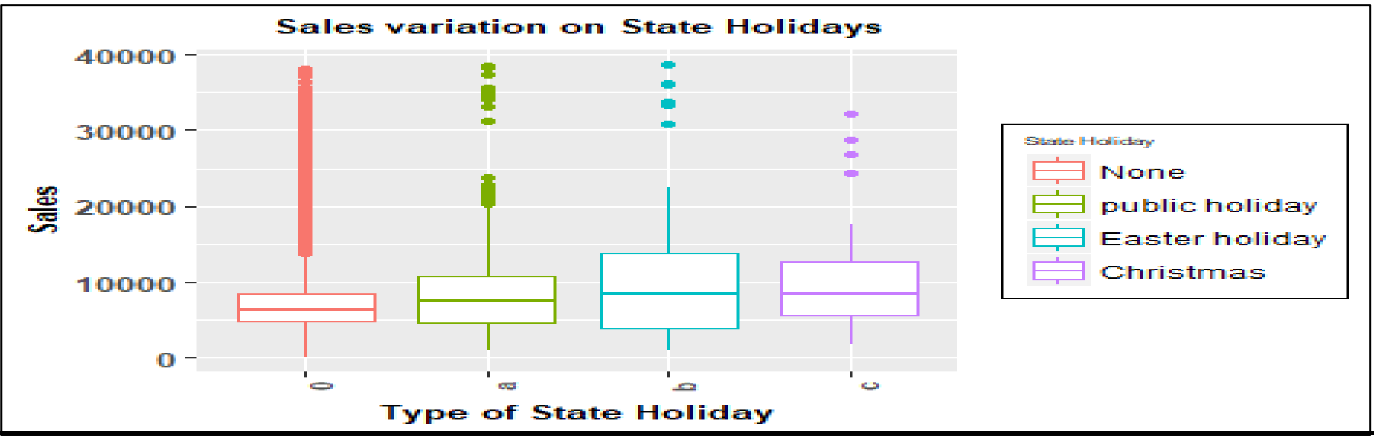






The correlation plot shows that customers and Sales values are highly correlated. This is obvious, when customers are more sales will be more. Thus, while data modeling we have not included Customers attribute. Reason for doing so is, store attributes responsible for higher sales will also be responsible for high customer count. Business decisions are made based on store attributes that are in control of sales team and modifying them will help team assess their impact on sales.

As we expect high sales on holidays, Easter and Christmas holidays showed high sales compared to Public holidays.

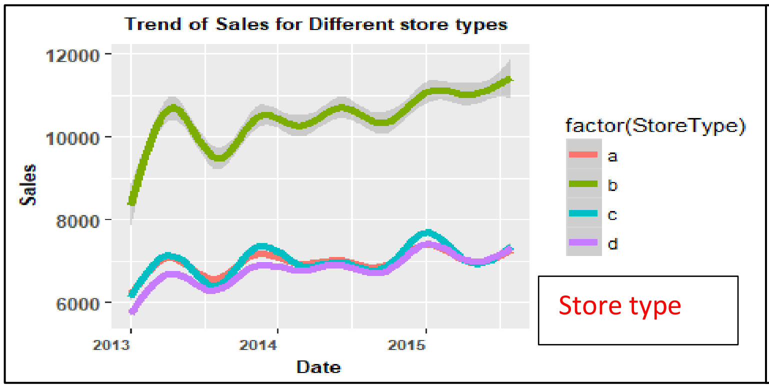


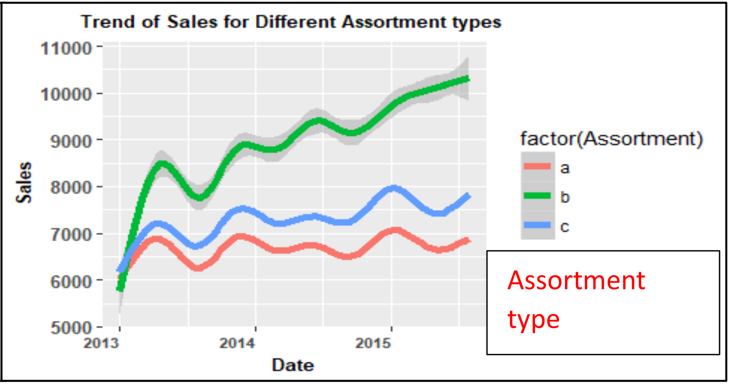
Store level daily sales trend:

For Type B store, sales were higher compared to other store types which can be seen from the below store type graph. Also, store types of a, c, d almost follow the same trend which is quite different from b type store.

Assortment level daily sales trend:

Assortment pertains to depth of the products. Assortment type a, c follows similar trend to that of a, c, d store type. While b assortment type is higher and similar to o b store type. This might be because store types of a, c, d may not have assortment type of b whereas b store type might have its major assortment type as b.

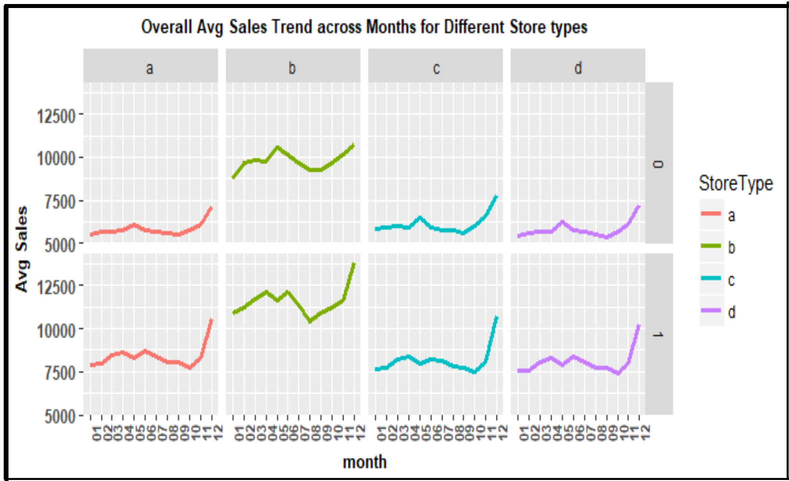


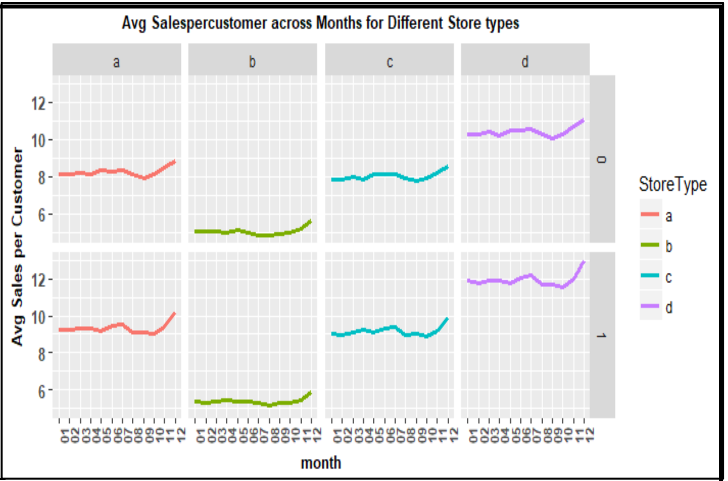


Our understanding got confirmed from the result of Contingency Table for Type of Stores and Assortment Level. Which showed that data is not equally available for all store type and assortment level. Thus, to avoid biased analysis we should consider averaging the sales values to explore promotion impact on different store types.

Type of Stores/Assortment Level	a	b	c
a	313	0	205
b	7	8	1
c	68	0	66
d	90	0	177

This finding, encouraged us to explore overall monthly average sales trend and average sales per customer trend for different store types.





Even though the plots showed Store Type B as the most selling and performant one, further investigation for average sales per customer showed higher for Store Type D.

- **Models and Quality Checks:**

We attempted to study our hypotheses through a few different ways. The first is by visually looking at the data. After looking at the various relationships of the variables we had better grasp on seasonality, trend and effect of promotions on sales.

Model 1: Pooled Model

Our dataset consists of daily sales from year 2013-2015, the sales for one day will not differ much for next day as one-day sales will have impact on next day sales. Thus, each record is not independent and autocorrelation for sales values will be observed. We cannot use simple linear regression as OLS assumptions won't be followed.

This we verified by conducting statistical tests such as Durbin-Watson and PLM test. The Durbin-Watson test on pooled OLS model showed high autocorrelation at lag 1 in model's error terms. Also, PLM test shows panel effect.

Mixed Effect Model: Model 2 & 3

Model 2: When promotions are given which months are giving more sales across year for any store type.

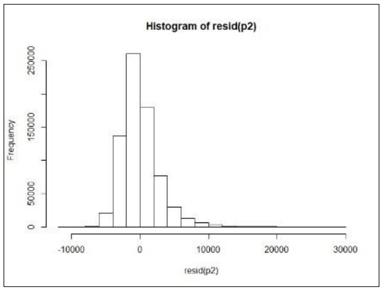
Model 3: It is similar to model 2 with addition of competition distance.

Following are the summary results for respective models:

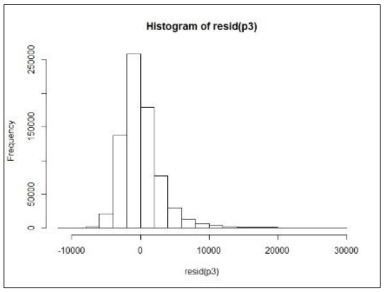
	Model Type	AIC	BIC
Model 1	Pooled	13711272	13711548
Model 2	Random	13676875	13677060
Model 3	Random	13708144	13708317

Residual Histogram:

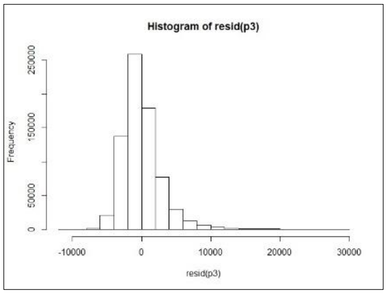
Model 1



Mode 2



Mode 3



Out of all the models we considered model 2 to use it further for our hypothesis analysis because it explained the pattern of sales in a logical manner and competition distance is having practically non-significant effect on sales.

Model 2 Analysis:

Based on the coefficients generated in model 2 we have developed equations for each hypothesis scenario and the trend chart have been plotted in excel.

Insight Question:

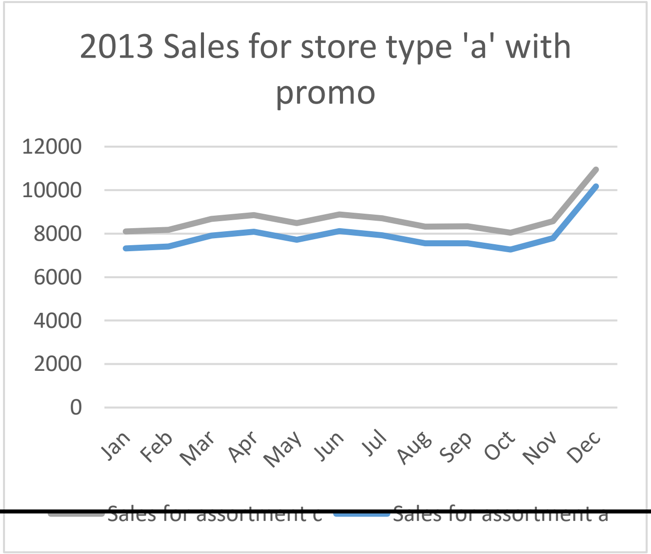
The target audience would like to know the effect of promotion on different months across store types and assortment types over years. The following charts shows the effect of promotions visually.

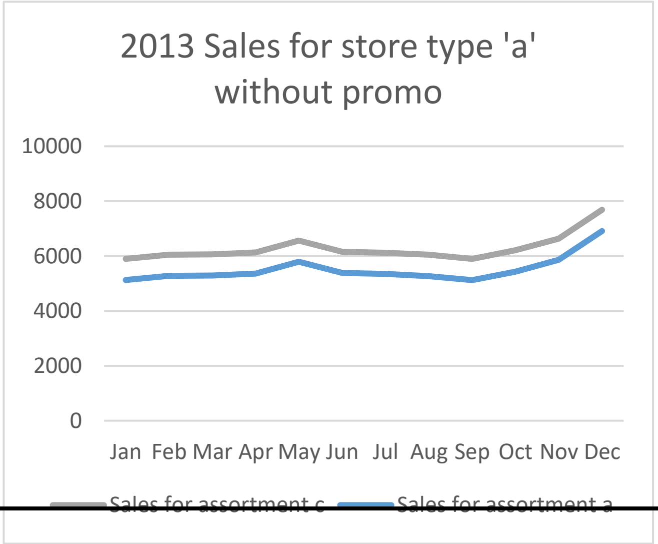
Year 2013, Store Type 'a' during no holidays:

The sales followed the below pattern across all assortment and we can see that the effect of promo amplified the sales during the festive months of December.

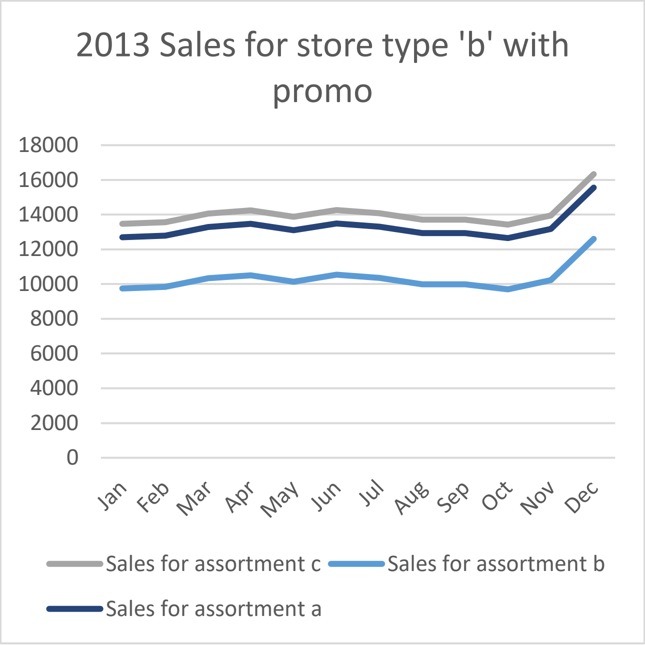
We can see here that the overall amount of sales decreased for Assortment 'b' but the pattern of sales were the same.

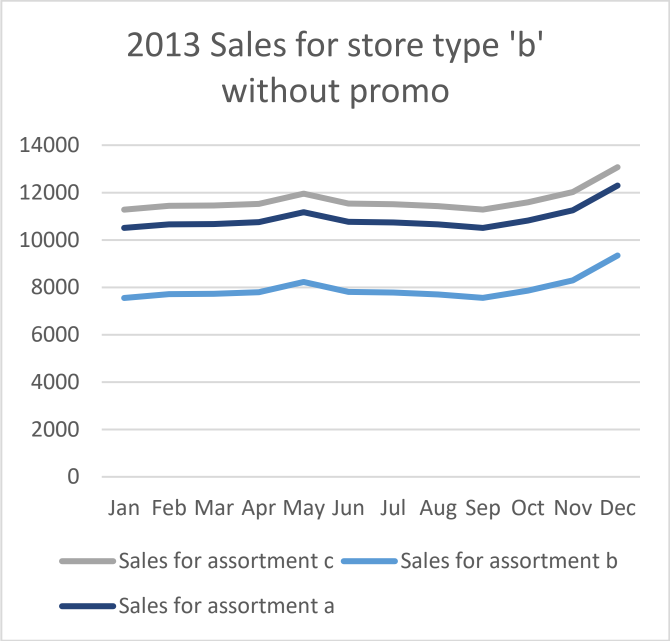
Here we can see that the highest sales for 2013 in store type a were for stores having assortment type c following similar sales pattern as that above.



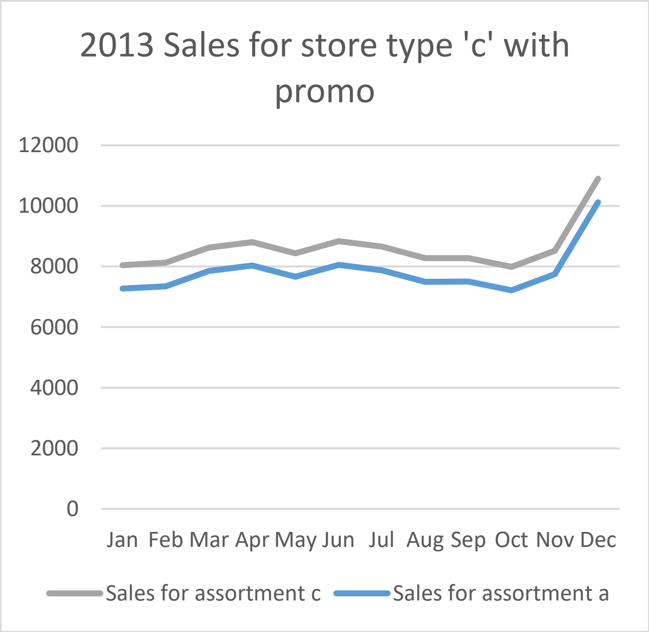


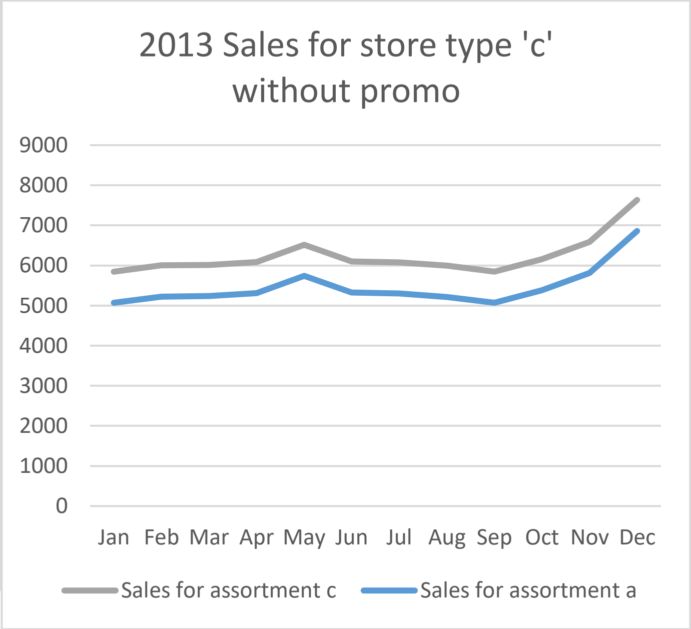
We can see there is overall growth in sales of store type b compared to sales of store type a.

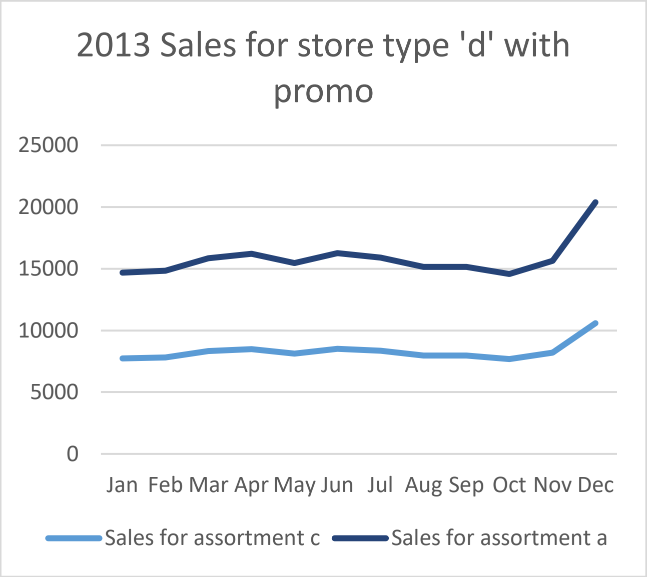


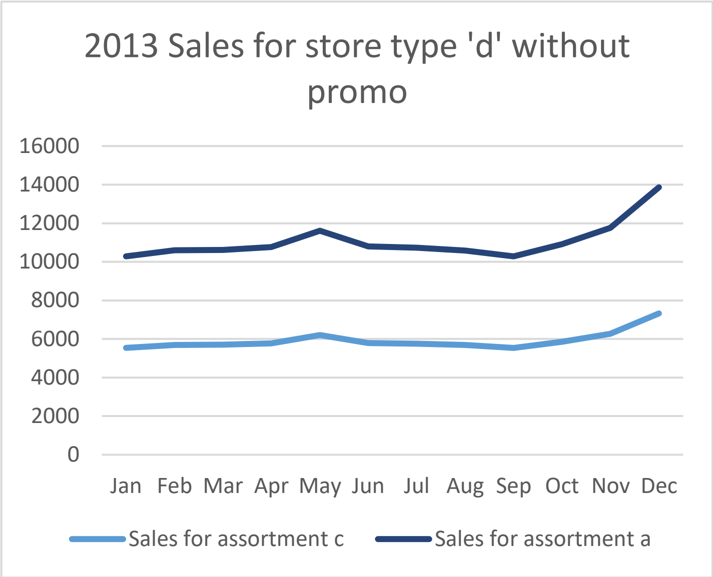


Similarly, we found that the overall sales of store type ‘c’ is less than store type ‘a’ and store type ‘c’.









<u>Comparison of Sales across Years:</u>

We found that overall sales estimate increased from 2013 to 2015 with highest being in 2015 following similar trends across months with or without promo.

The above model analysis answers the initially hypothesis in the following way:

- What is the effect of promotions on different months?
  The promotions have positive impact on sales across all months and all store types overall years.
- For what store/assortment type promotion impact is more?
  Store Type a–Extended assortment (type c)
  Store Type b- Extended assortment (type c)
  Store Type c- Extended assortment (type c)
  Store Type d- Basic assortment (type a)

- Does giving promotions boost sales in cases where competitor distance is smaller?

  We noticed that the inclusion of competitor distance in our model did not have any impact on the sales. Since this effect was not practically significant we cannot accept this hypothesis.

Recommendations:

When promotions are given, irrespective of store type Sales are higher compared to when promotions are not given. Thus, we can say that the promotion strategy has overall positive impact on sales. The effect of promotions amplified the sales for December month for all store and assortment types.

Our year wise analysis showed that the store type d has shown positive impact on sales compared to other store types when promotions are given. For store type d with basic assortment shows more positive impact on sales compared to other assortment types. Thus, we can say that if promotions are to be given on new store type, manager should choose store type d with basic assortment level and similarly if the expansion is for store type a, b, c then the extended assortment type is recommended.

Across all years and all store types we found out that there is a dip in sales in the month of May, October and November even though the promotions were given. We recommend that the new pricing/promotional strategies should be built keeping in view of the events in May, October and November that contribute to this observation.

Appendix:

```
p1<-lm(Sales ~as.factor(StateHoliday)+as.factor(Assortment)+

        as.factor(StoreType)+as.factor(year)+as.factor(month)+as.factor(Promo), data =z)

p2<-lmer(Sales ~as.factor(StateHoliday)+as.factor(Assortment)+

        as.factor(StoreType)+as.factor(year)+(1+Promo|month), data=z)

p3<-lmer(Sales ~as.factor(StateHoliday)+CompetitionDistance+as.factor(Assortment)+

        as.factor(StoreType)+as.factor(year)+(1+Promo|month), data=z)
```

References:

https://www.princeton.edu/~otorres/Panel101.pdf

http://forscenter.ch/wp-content/uploads/2017/03/Slides_2016_all.pdf

http://www.cantab.net/users/bf100/pdf/pd_slides_fingleton.pdf

http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataNotes.htm