DATA MINING

ASSIGNMENT 1

WQD170041 MUHAMMAD IZZUDDIN BIN AHAMAD SFHAFI

Business Background Study

Rossman is a dataset that was derived from Kaggle website. The dataset was for 1115 store throughout Germany. The dataset is available in the Kaggle website. Rossmann operates over 3,000 drug stores in 7 European countries.. 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. The challenge was intended to predict the sales for all stores. Using tableau, we are going to make visualization to ex plain what happen for each variable in the dataset.

https://www.kaggle.com/c/rossmann-store-sales/data

The variables from acquired from Kaggle website consist of:

- 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
- 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 if the (Store, Date) was affected by the closure of public schools
- 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

The external datasets were derived from other participant that contribute to this challenge consist of weather data, google trend, and store name.

Data Mining Objective

To do Explanatory Data Analysis to find out factors that influence the sales of Rossman Store using tableau visualization tools.

Star Schema

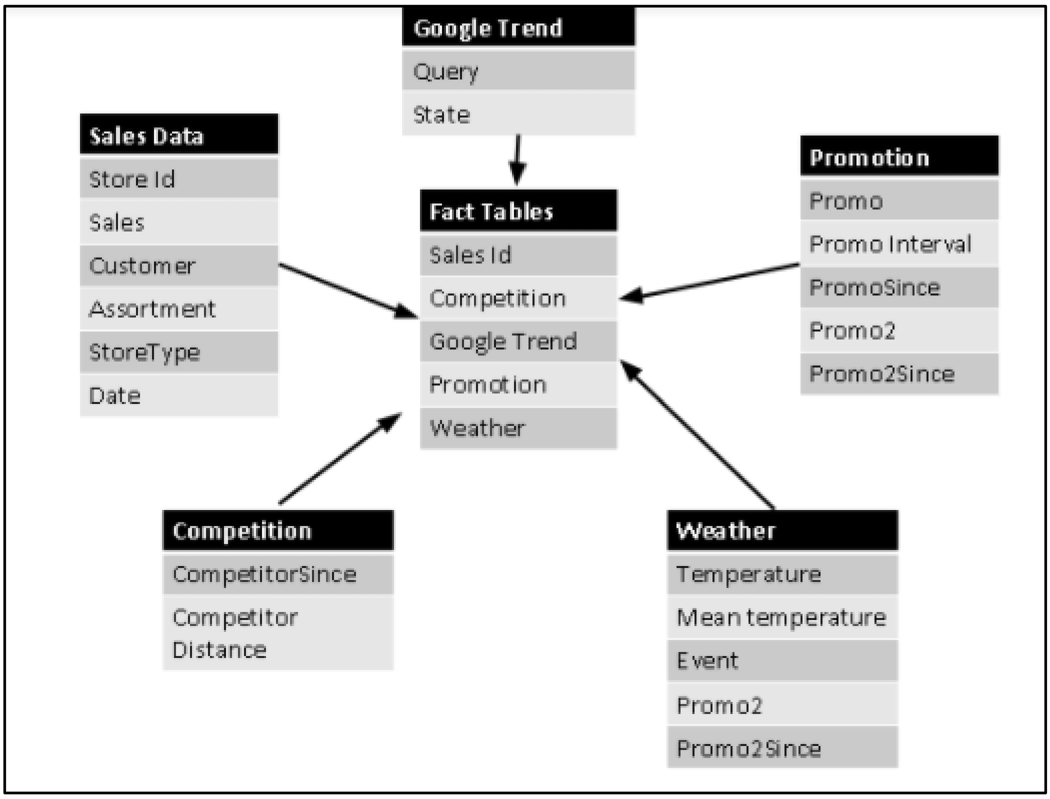
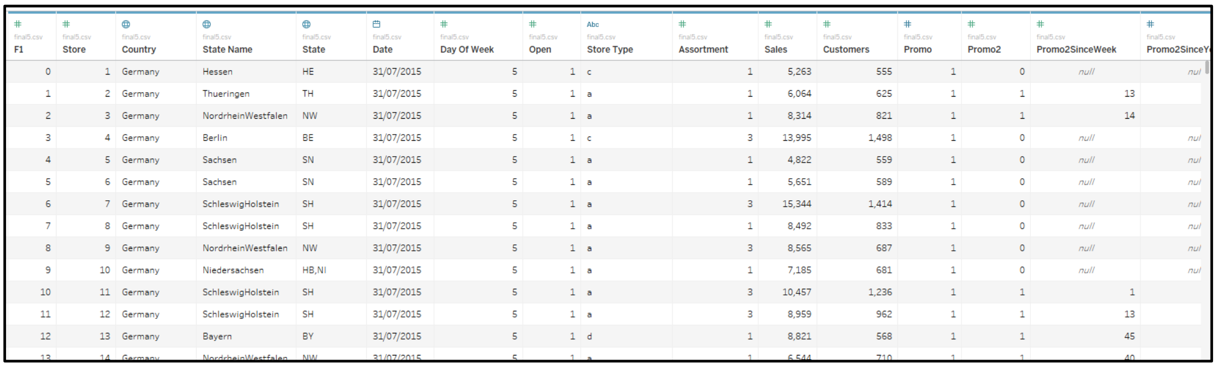


Figure 1: Star Schema

The star schema created inside tableau joined the Sales Data, Google Trend, Promotion, Competition and Weather dataset.



F1	Store	Country	State Name	State	Date	Day Of Week	Open	Store Type	Assortment	Sales	Customers	Promo	Promo2	Promo2SinceWeek	Promo2SinceY
0	1	Germany	Hessen	HE	31/07/2015	5	1	c	1	5,263	555	1	0	null	nul
1	2	Germany	Thueringen	TH	31/07/2015	5	1	a	1	6,064	625	1	1	13	
2	3	Germany	NordrheinWestfalen	NW	31/07/2015	5	1	a	1	8,314	821	1	1	14	
3	4	Germany	Berlin	BE	31/07/2015	5	1	c	3	13,995	1,498	1	0	null	nul
4	5	Germany	Sachsen	SN	31/07/2015	5	1	a	1	4,822	559	1	0	null	nul
5	6	Germany	Sachsen	SN	31/07/2015	5	1	a	1	5,651	589	1	0	null	nul
6	7	Germany	SchleswigHolstein	SH	31/07/2015	5	1	a	3	15,344	1,414	1	0	null	nul
7	8	Germany	SchleswigHolstein	SH	31/07/2015	5	1	a	1	8,492	833	1	0	null	nul
8	9	Germany	NordrheinWestfalen	NW	31/07/2015	5	1	a	3	8,565	687	1	0	null	nul
9	10	Germany	Niedersachsen	HB,NI	31/07/2015	5	1	a	1	7,185	681	1	0	null	nul
10	11	Germany	SchleswigHolstein	SH	31/07/2015	5	1	a	3	10,457	1,236	1	1	1	
11	12	Germany	SchleswigHolstein	SH	31/07/2015	5	1	a	3	8,959	962	1	1	13	
12	13	Germany	Bayern	BY	31/07/2015	5	1	d	1	8,821	568	1	1	45	
13	14	Germany	NordrheinWestfalen	NW	31/07/2015	5	1	a	1	6,544	710	1	1	40	

Figure 2:The dataset

We have in total 59 columns and 1017209 rows of dataset for 1115 stores dated from 1/01/2013 to 31/07/2015

States Name for Rossman Store

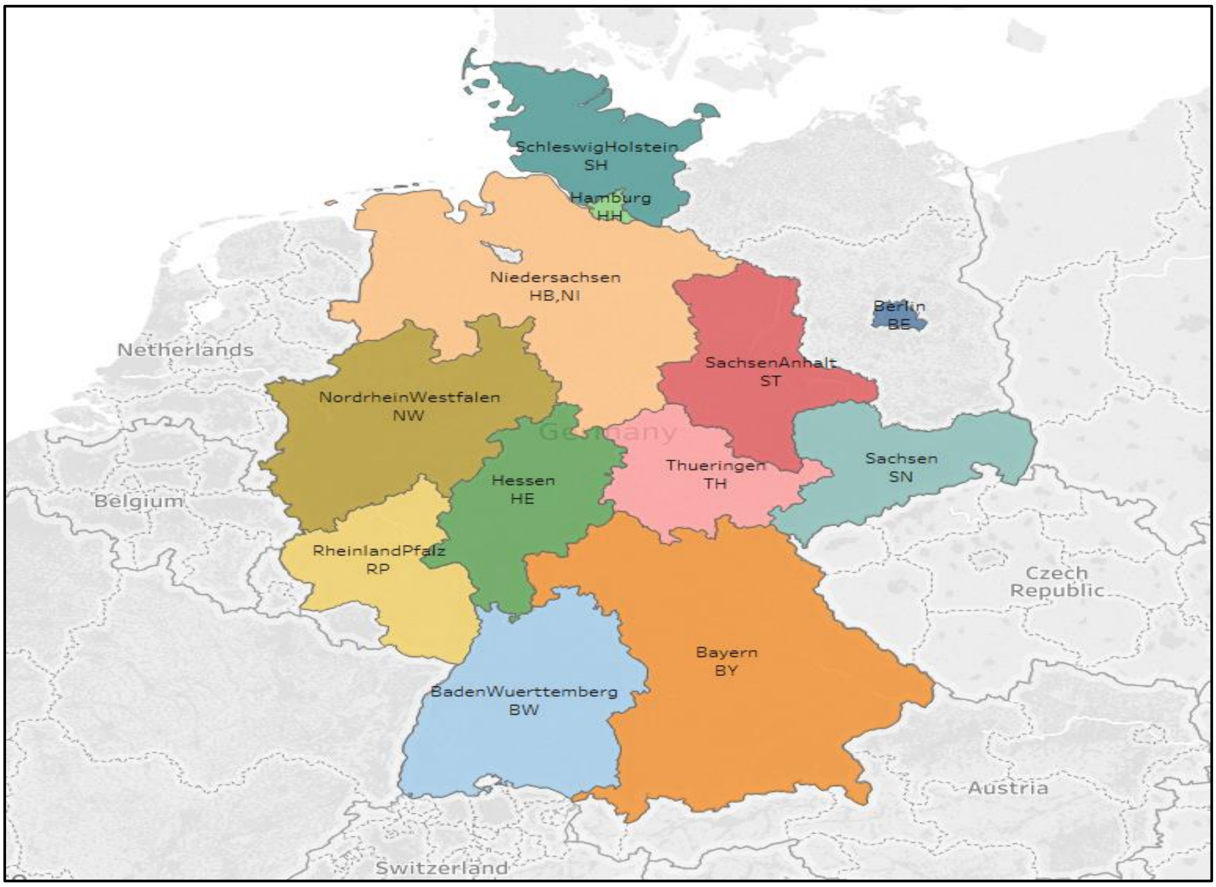


Figure 3:States name for this study

Using map diagram. A map that represent the states name was generated to show the area of study for this analysis. All the states names available were given abbreviation(BE,BW,BY,HE,HH..etc etc) as this abbreviation will be used as key value that connect other dataset (weather and google trend)

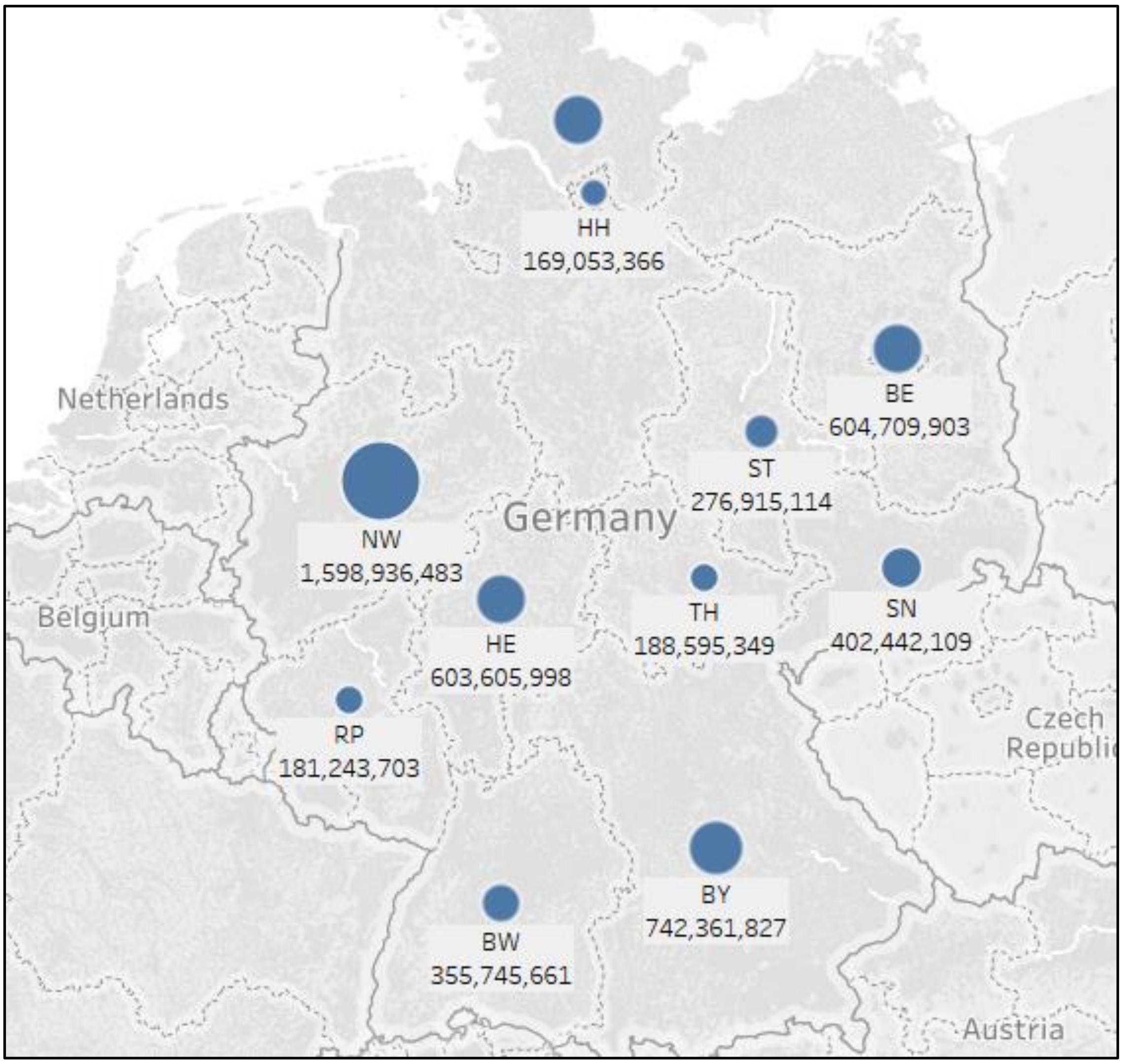


Figure 4: Sum of sales for each state

The diagram above show the Total sales for each states using mark "circle". Using the maps function, a map was generated to determine the sales and customers for each states. The mark "Circle" is used to visualize the total number of sales and customer for each country together with "label" sum of sale and sum of customer.

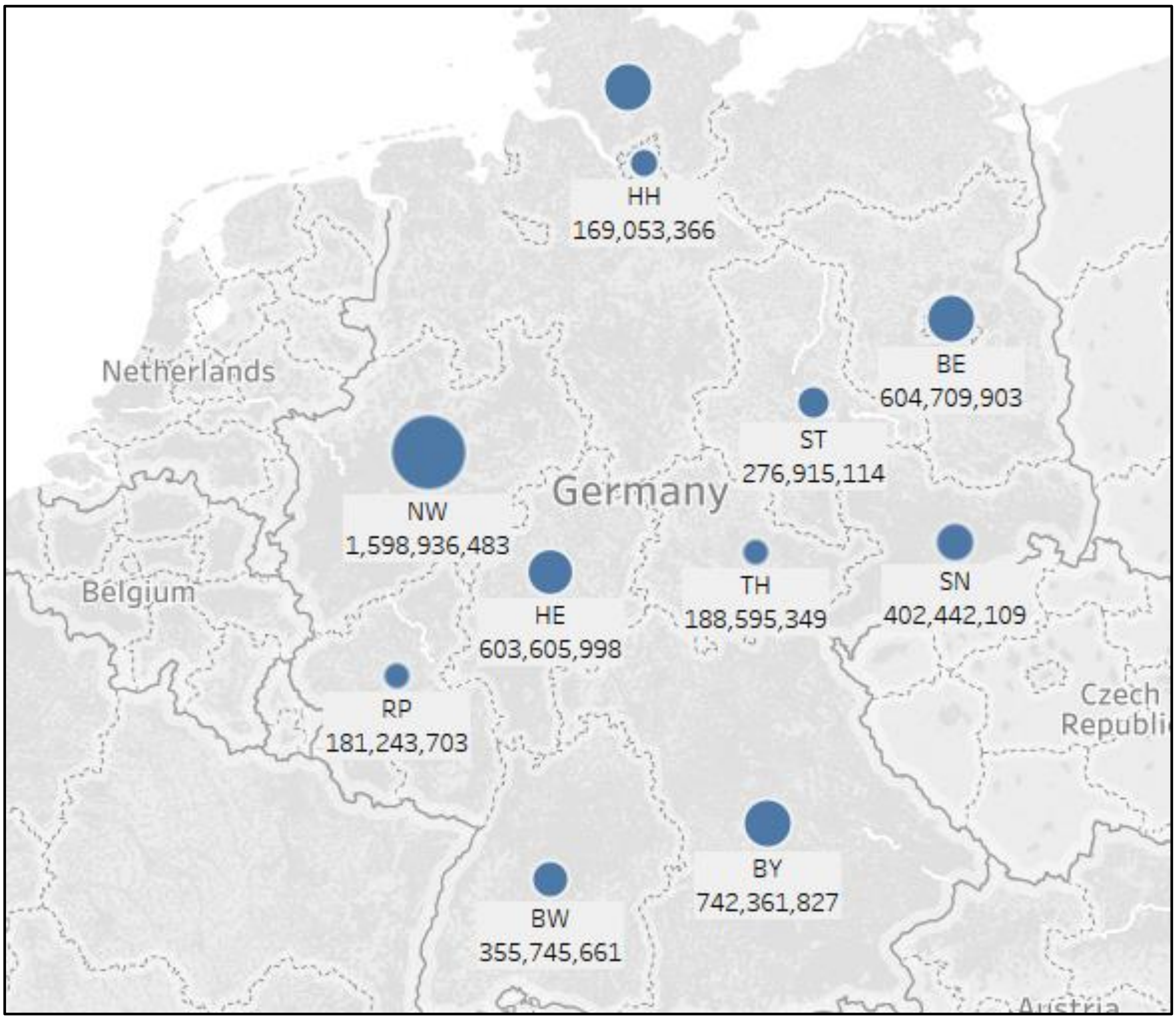


Figure 5: Sum of customers for each state

The diagram above show the total customer that visit the stores throughout the study period using mark "circle" function.

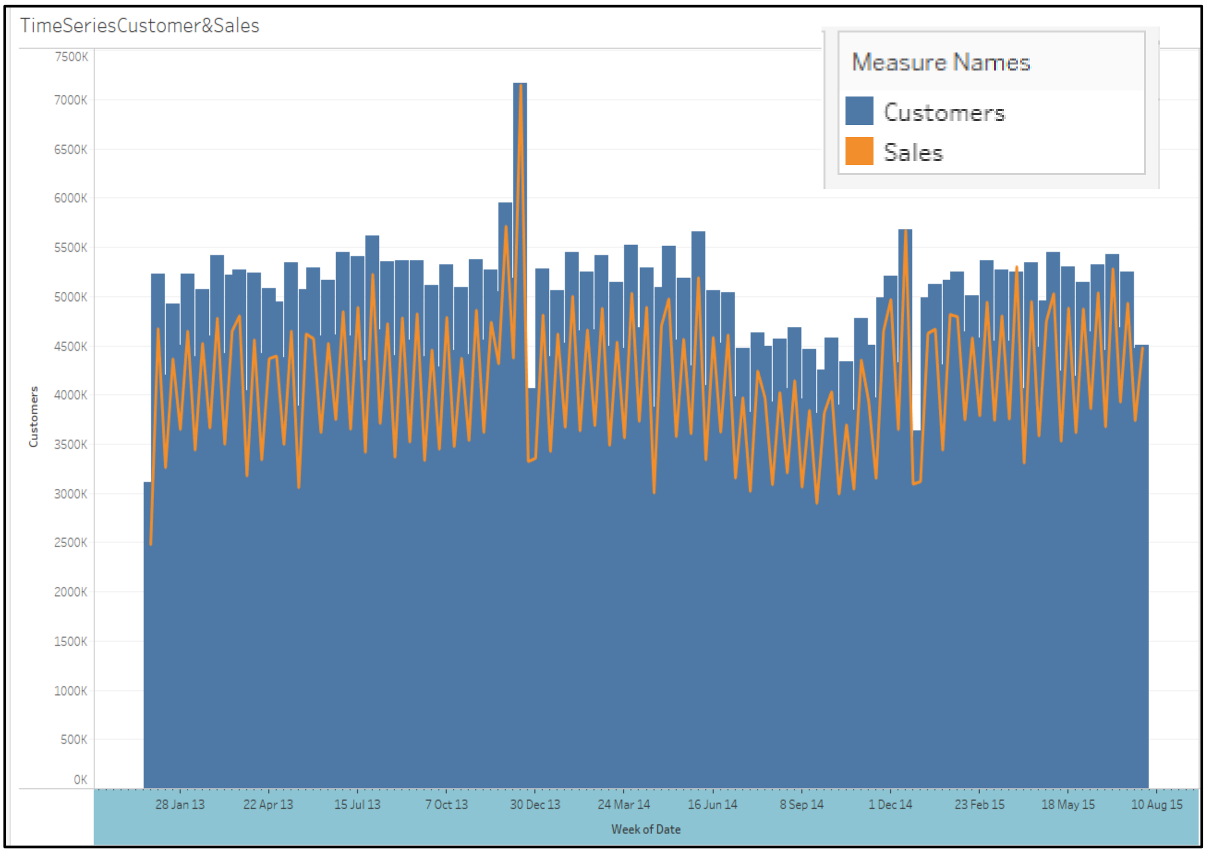


Figure 6: Time Series Customer & Sales

Based on the dual combination below, we can see that total number of sales and total number of customers is highly correlated

Create Dashboard for StateSales and StateCustomers

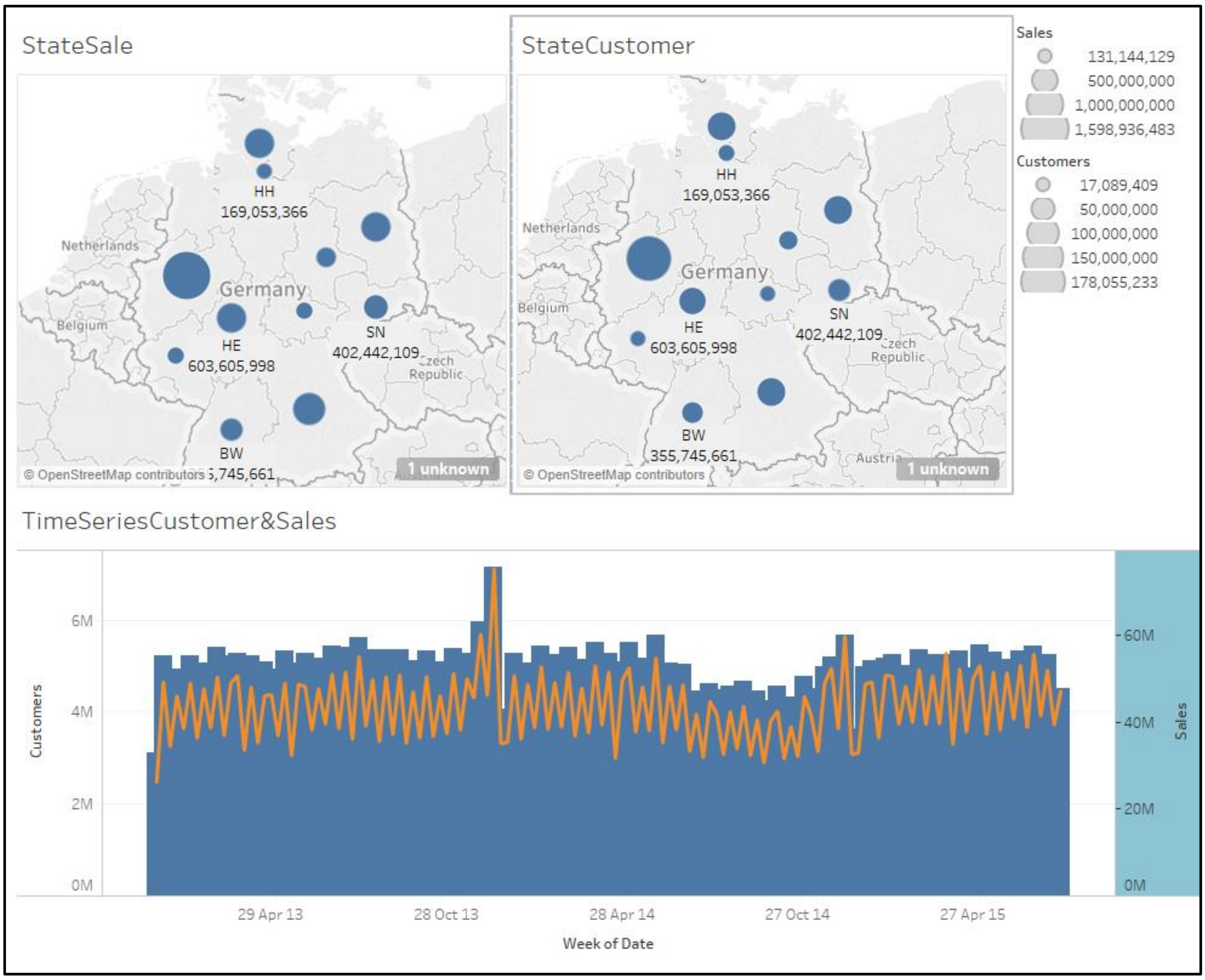


Figure 7: Dashboard for StateSale & StateCustomer

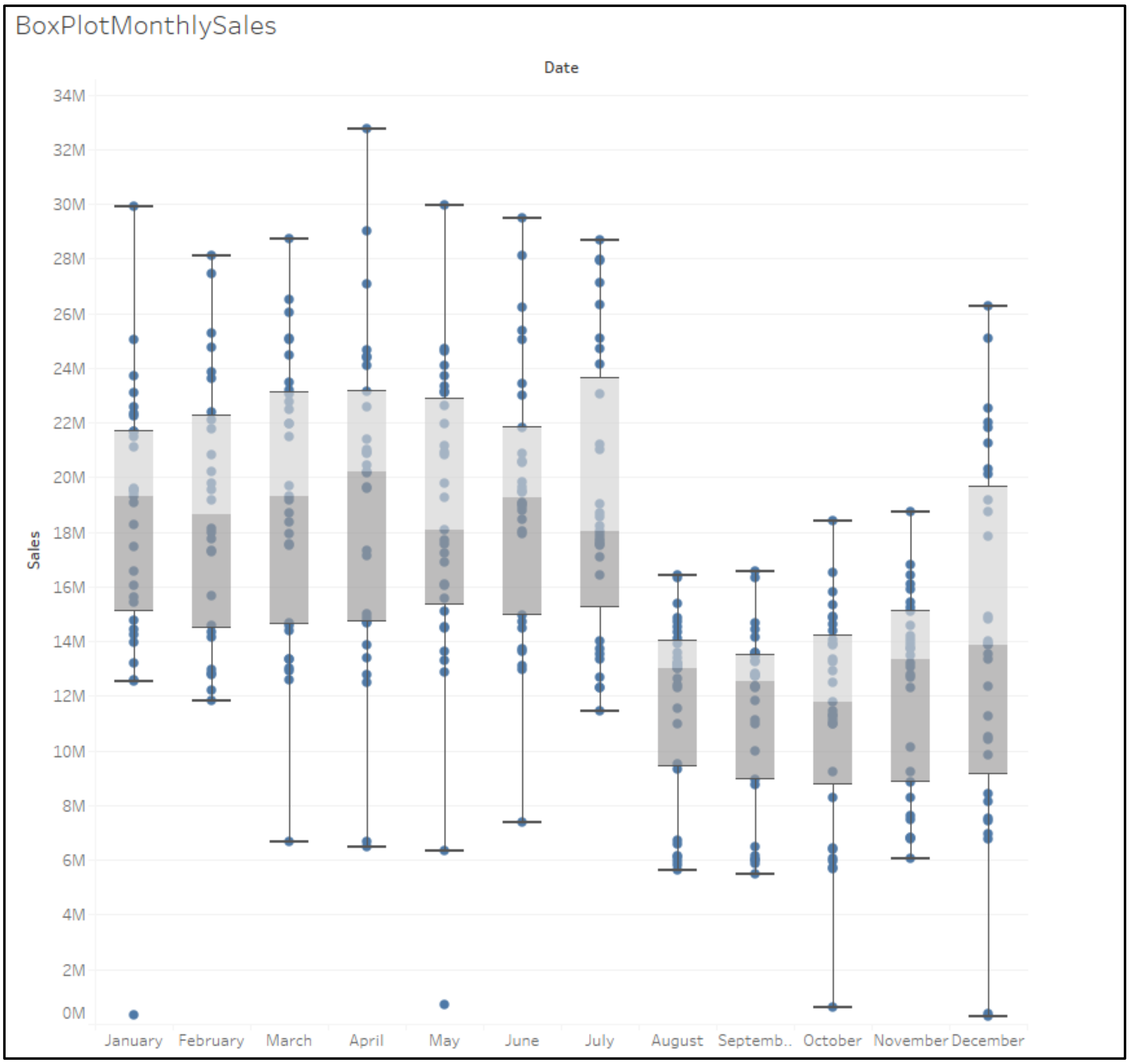


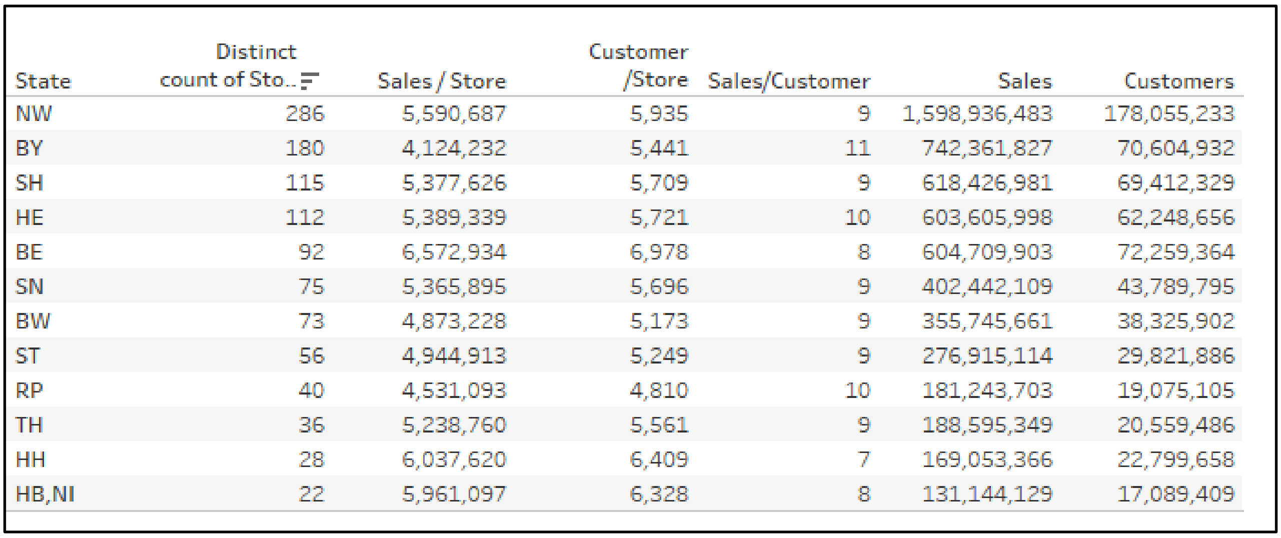
Figure 8: Box Plot Monthly Sale

Boxplot showed first half of the year (Jan-July) contribute to higher revenue frequency compare to the second half of the year (Aug-Dec)

Store Analysis

Store Performance studies based on states

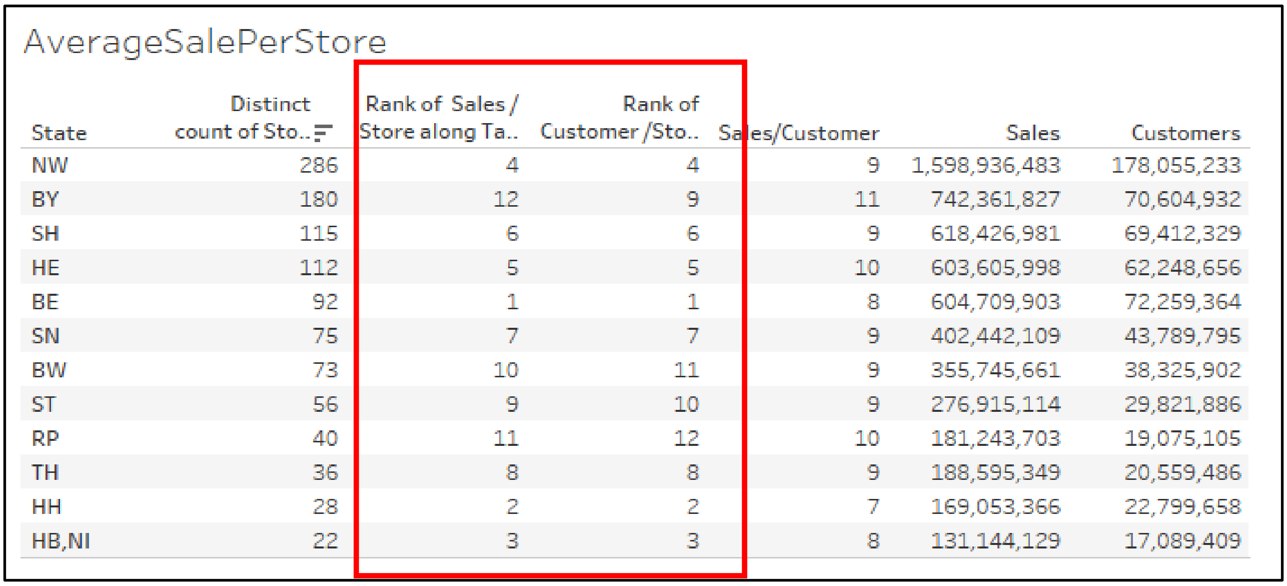
We are going to dissect each store performance for each states using the text tables. Text tables allow us to study multiple columns of data simultaneously. We use custom create field to generate figure for Sales/Store, Customer/Store and Sales/Customer



	Distinct		Customer			
State	count of Sto..	Sales / Store	/Store	Sales/Customer	Sales	Customers
NW	286	5,590,687	5,935	9	1,598,936,483	178,055,233
BY	180	4,124,232	5,441	11	742,361,827	70,604,932
SH	115	5,377,626	5,709	9	618,426,981	69,412,329
HE	112	5,389,339	5,721	10	603,605,998	62,248,656
BE	92	6,572,934	6,978	8	604,709,903	72,259,364
SN	75	5,365,895	5,696	9	402,442,109	43,789,795
BW	73	4,873,228	5,173	9	355,745,661	38,325,902
ST	56	4,944,913	5,249	9	276,915,114	29,821,886
RP	40	4,531,093	4,810	10	181,243,703	19,075,105
TH	36	5,238,760	5,561	9	188,595,349	20,559,486
HH	28	6,037,620	6,409	7	169,053,366	22,799,658
HB,NI	22	5,961,097	6,328	8	131,144,129	17,089,409

Figure 9: Store Performance for each state

We sorted the table above starting with highest count of stored. From the figure above, State "NW" has the highest number of store (286). And contribute to the highest total sales and total number of customers per state. On the other hand, Sales/store, Customers/Store and Sales/Customer based on States indicated that NW did not top the ranking. Due to the limitation of tableau to colour different column with different range of colour, we decide to use ranking number to assess performance for column Sale/Store and Customer/Store



AverageSalePerStore

	Distinct	Rank of Sales /	Rank of			
State	count of Sto..	Store along Ta..	Customer /Sto..	Sales/Customer	Sales	Customers
NW	286	4	4	9	1,598,936,483	178,055,233
BY	180	12	9	11	742,361,827	70,604,932
SH	115	6	6	9	618,426,981	69,412,329
HE	112	5	5	10	603,605,998	62,248,656
BE	92	1	1	8	604,709,903	72,259,364
SN	75	7	7	9	402,442,109	43,789,795
BW	73	10	11	9	355,745,661	38,325,902
ST	56	9	10	9	276,915,114	29,821,886
RP	40	11	12	10	181,243,703	19,075,105
TH	36	8	8	9	188,595,349	20,559,486
HH	28	2	2	7	169,053,366	22,799,658
HB,NI	22	3	3	8	131,144,129	17,089,409

Figure 10: Ranking performance for Sales/Store and Customers/Store

State “BE” Contribute to the highest Sales and Customer per unit store and surprisingly, State “HB,BB” and “HH” which has the lowest and second lowest of total number of store contribute to the third and second highest sales and customer per store correspondingly. Yet, The total average sales per customer per store in state BE was amongst the lowest ($8/customer) compare to other state

Store Type Performance

Each store consist of 4 models (a,b,c,d). We are going to assess the performance of each store type using pie chart

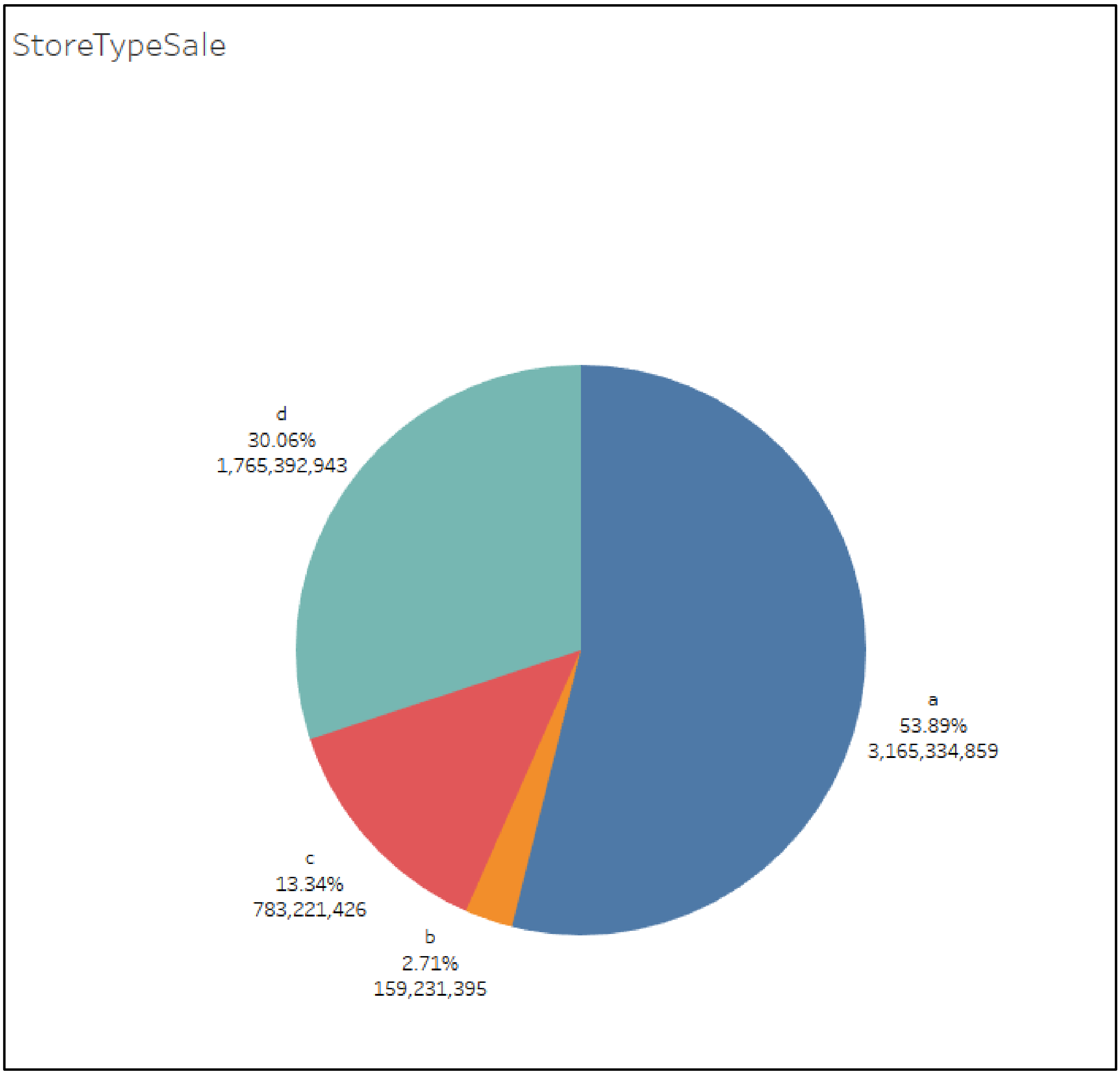
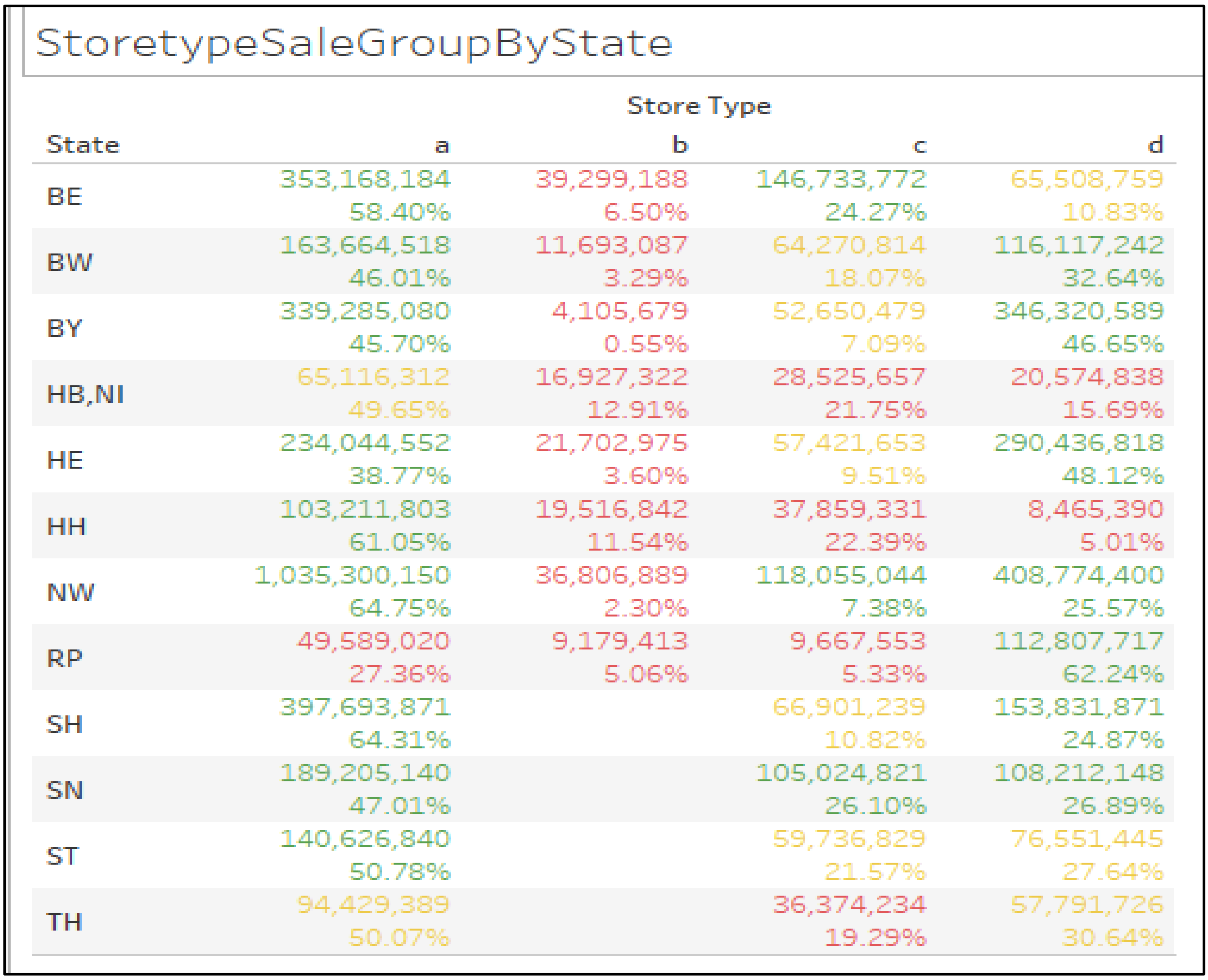


Figure 11: Store Type Sales

Store Type Performance Based on States



StoretypeSaleGroupByState

	Store Type			
State	a	b	c	d
BE	353,168,184 58.40%	39,299,188 6.50%	146,733,772 24.27%	65,508,759 10.83%
BW	163,664,518 46.01%	11,693,087 3.29%	64,270,814 18.07%	116,117,242 32.64%
BY	339,285,080 45.70%	4,105,679 0.55%	52,650,479 7.09%	346,320,589 46.65%
HB,NI	65,116,312 49.65%	16,927,322 12.91%	28,525,657 21.75%	20,574,838 15.69%
HE	234,044,552 38.77%	21,702,975 3.60%	57,421,653 9.51%	290,436,818 48.12%
HH	103,211,803 61.05%	19,516,842 11.54%	37,859,331 22.39%	8,465,390 5.01%
NW	1,035,300,150 64.75%	36,806,889 2.30%	118,055,044 7.38%	408,774,400 25.57%
RP	49,589,020 27.36%	9,179,413 5.06%	9,667,553 5.33%	112,807,717 62.24%
SH	397,693,871 64.31%		66,901,239 10.82%	153,831,871 24.87%
SN	189,205,140 47.01%		105,024,821 26.10%	108,212,148 26.89%
ST	140,626,840 50.78%		59,736,829 21.57%	76,551,445 27.64%
TH	94,429,389 50.07%		36,374,234 19.29%	57,791,726 30.64%

Figure 12:Store Type Performance Based on States

Using custom conditional function, Any sales above 100 million will be green (great performance), above 50 million will be yellow (ok performance) and below 50 million will be red (bad performance). From the rough observation, it can be seen that the store type b perform really bad for all states that it is open in. while store a perform really well in most of the state.

Promotion Analysis

Promotion performance

We are going to investigate how promotion days influence the sales

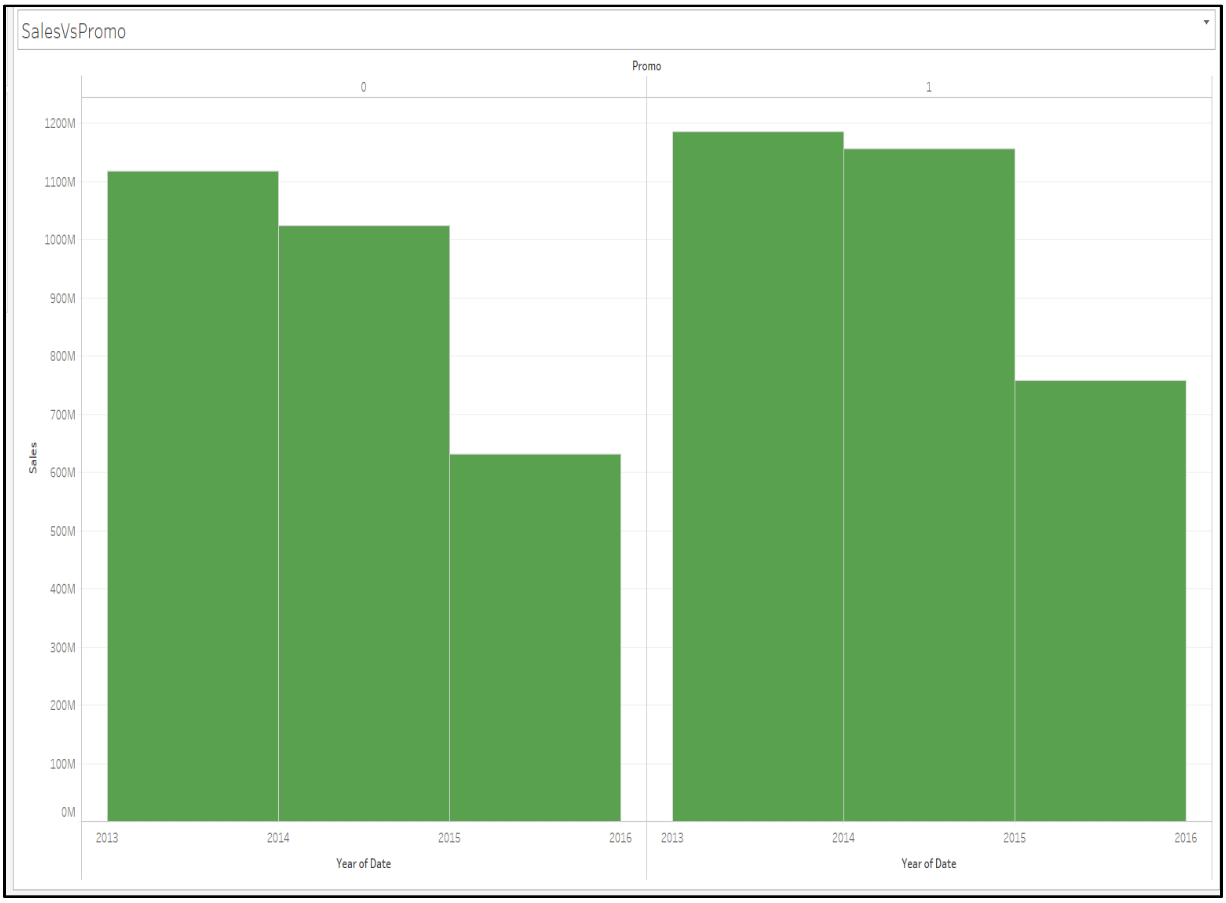


Figure 13: Sales performance based on promotion

1 stand for the day with promotion and 0 stand for the day without promotion. The diagram above showed that not having promotion or not did not have significant influence to the bottom line revenue.

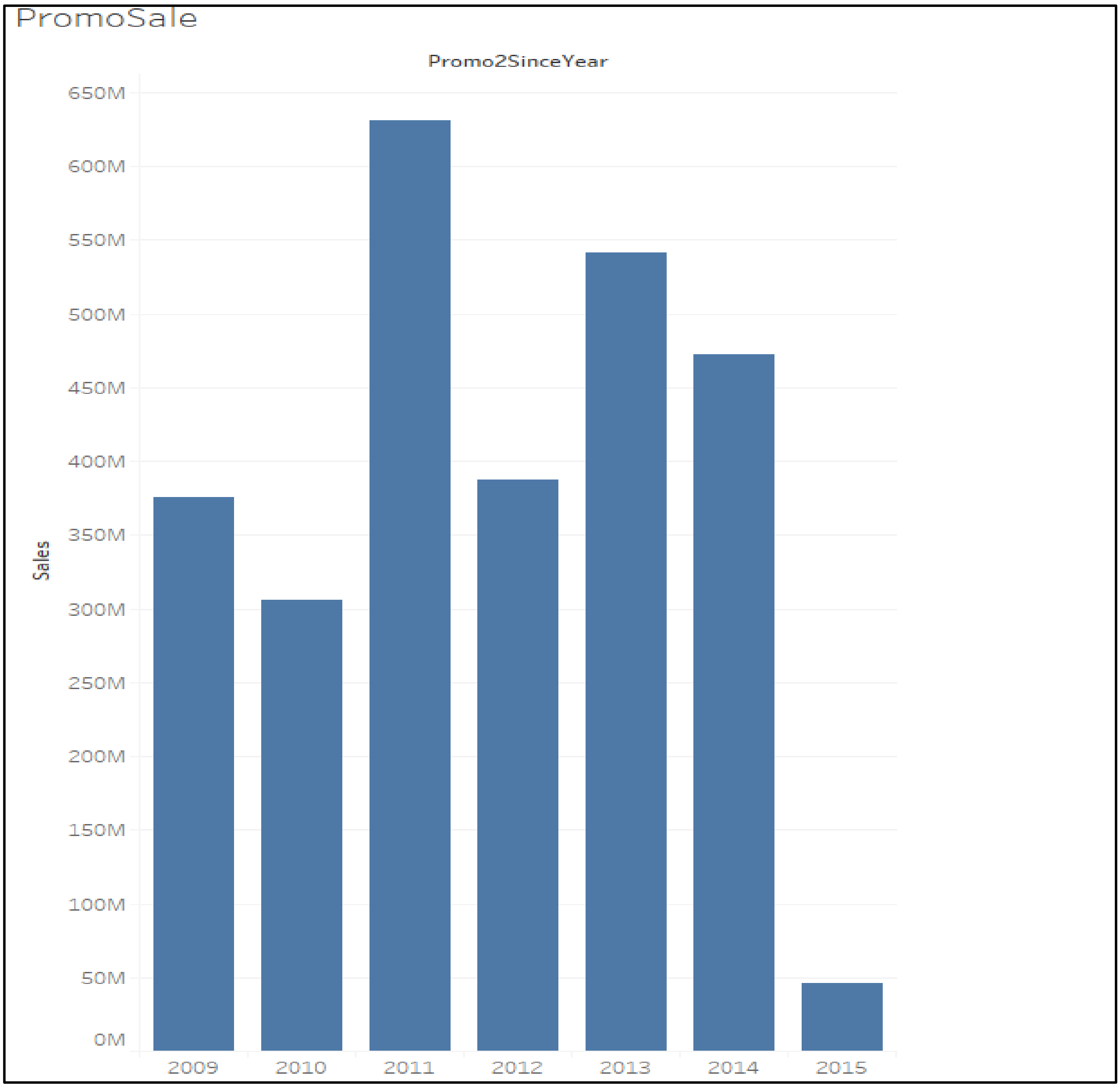
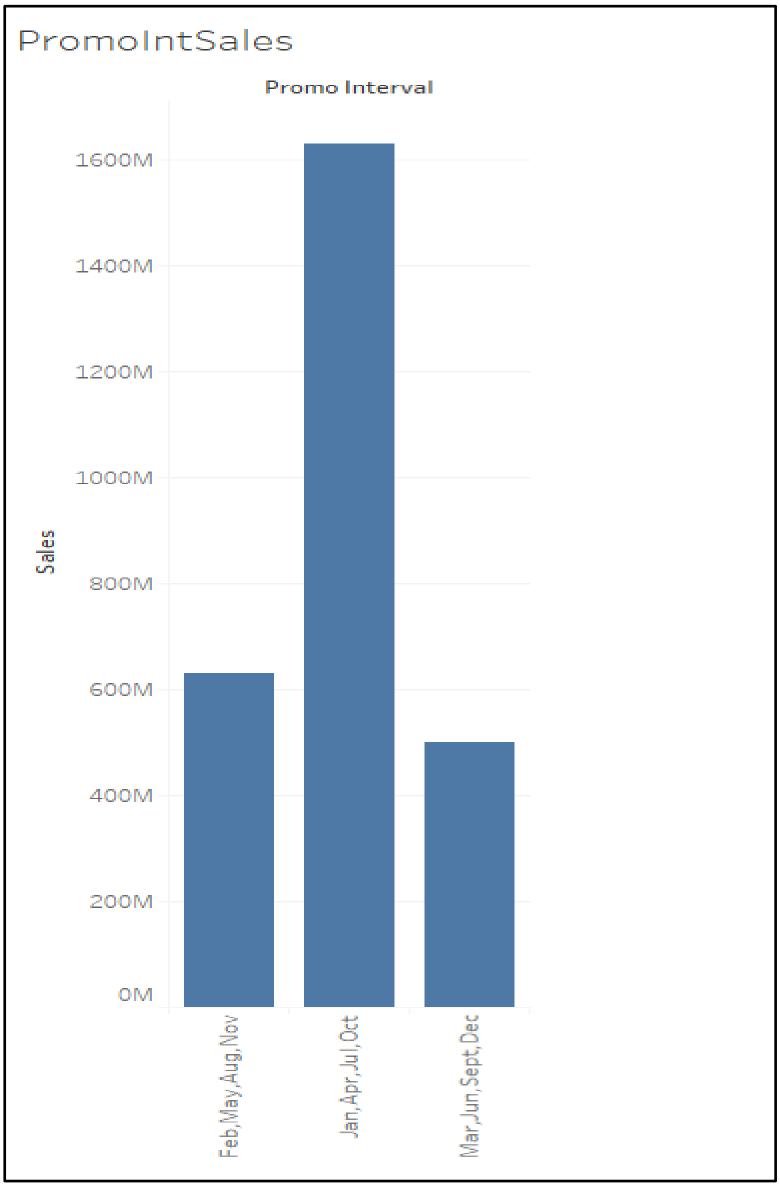


Figure 14:Sales Vs PromotionSinceYear

From above, It can be observed that promotion that start at 2011 contribute to the highest sales

Promo Interval



Based on promotion interval toward sale. We can see that the interval Jan,Apr,Jul,Oct contribute to the highest revenue

Weather Analysis

Temperature

Using dataset from weather, We are going to assess Customers attendance based on event of weather

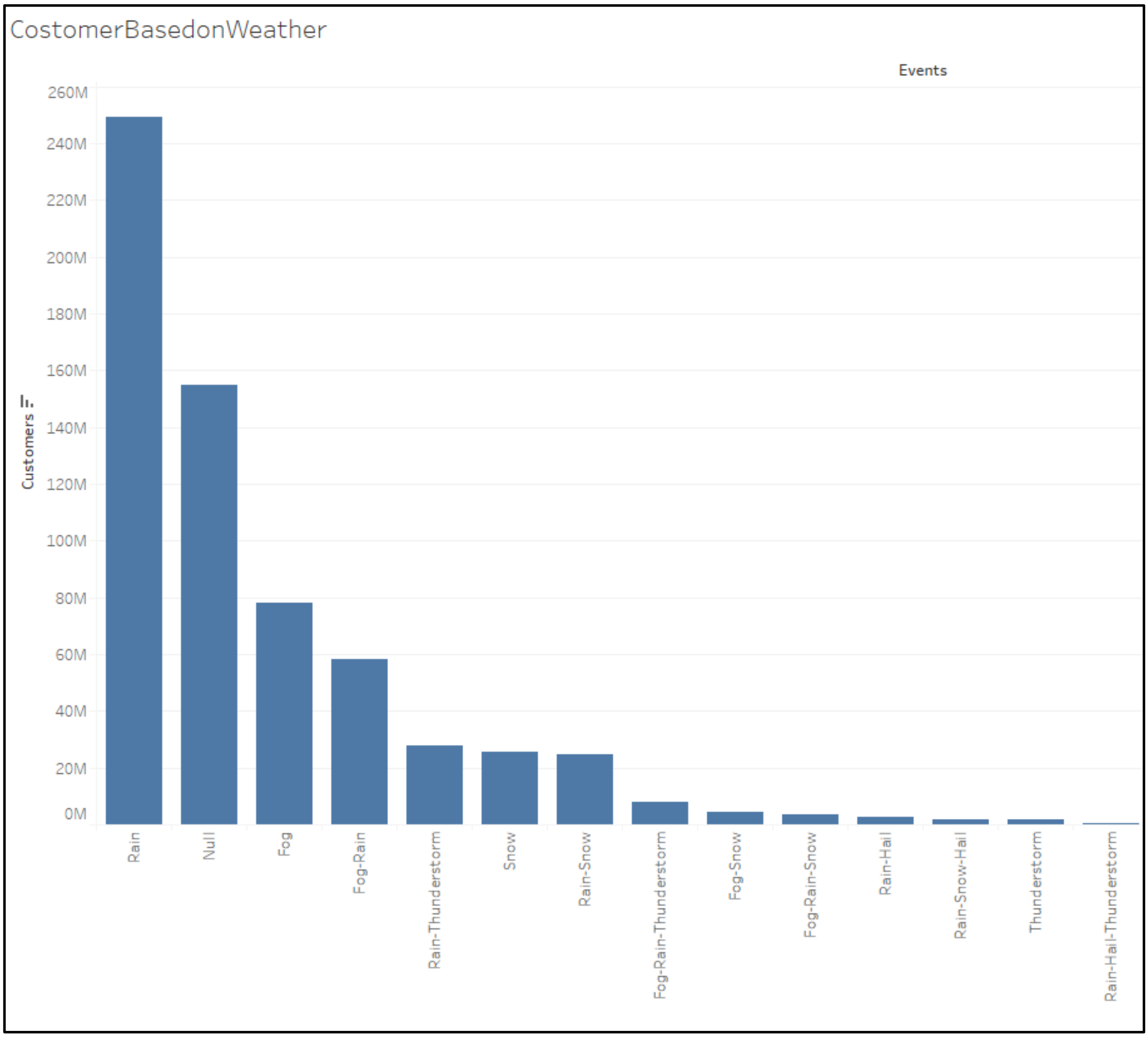


Figure 15: Customers Attendance Based on Weather

We use sort to arrange the bar from highest of customer count(from the left) to the lowest(to the right) can be observed that customer attendance was the highest even during the rainy season. The Customer attendance was sorted from Highest attendance (left) to lowest (right)

Time Series of relationship between Average Customer Attendance and Median Temperature

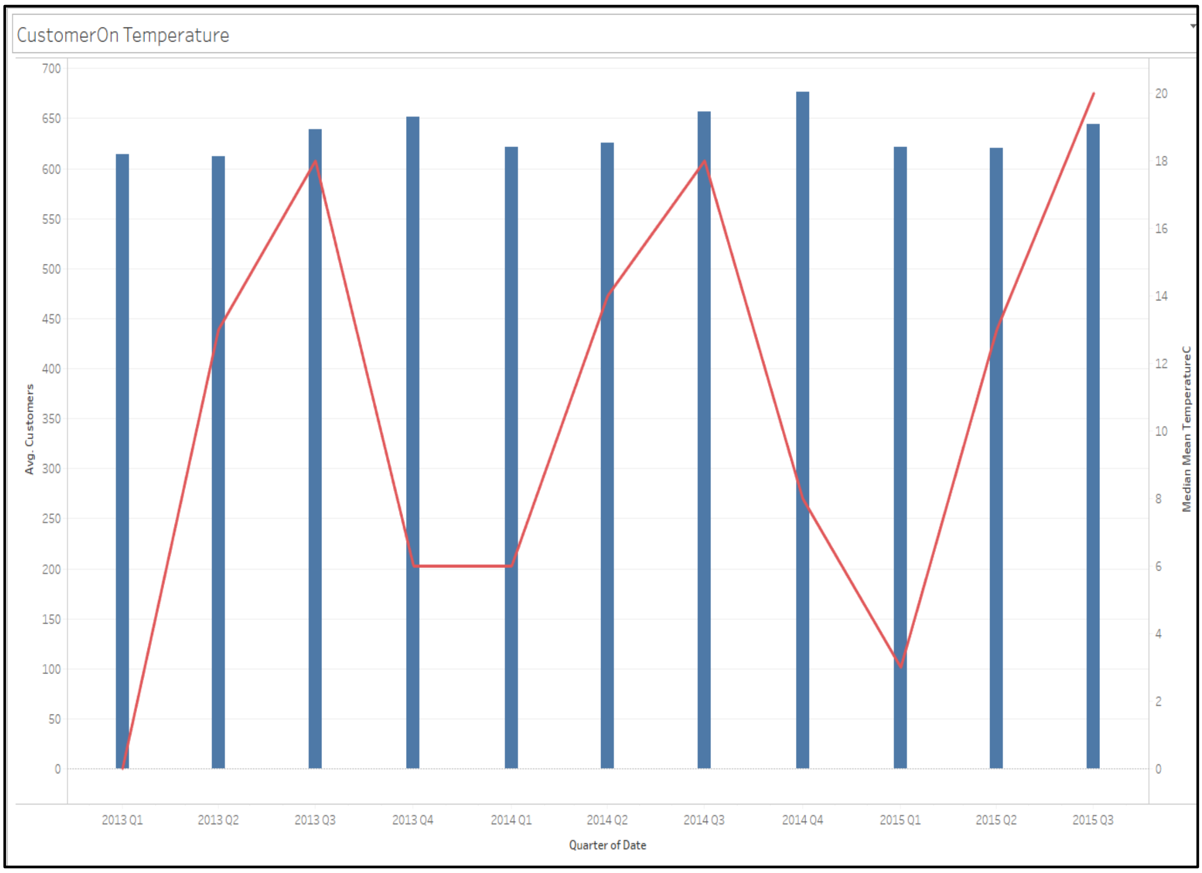


Figure 16:Average Customer Attendance and Median Temperature

From the observation above, the temperature did not posed any correlation with average customer coming to the store. Surprisingly, the attendance of the customer did not drop during the lowest temperature period of the year.

Competitor Analysis

We are using scatter plot to find the relationship between distance of competitor toward the sales.

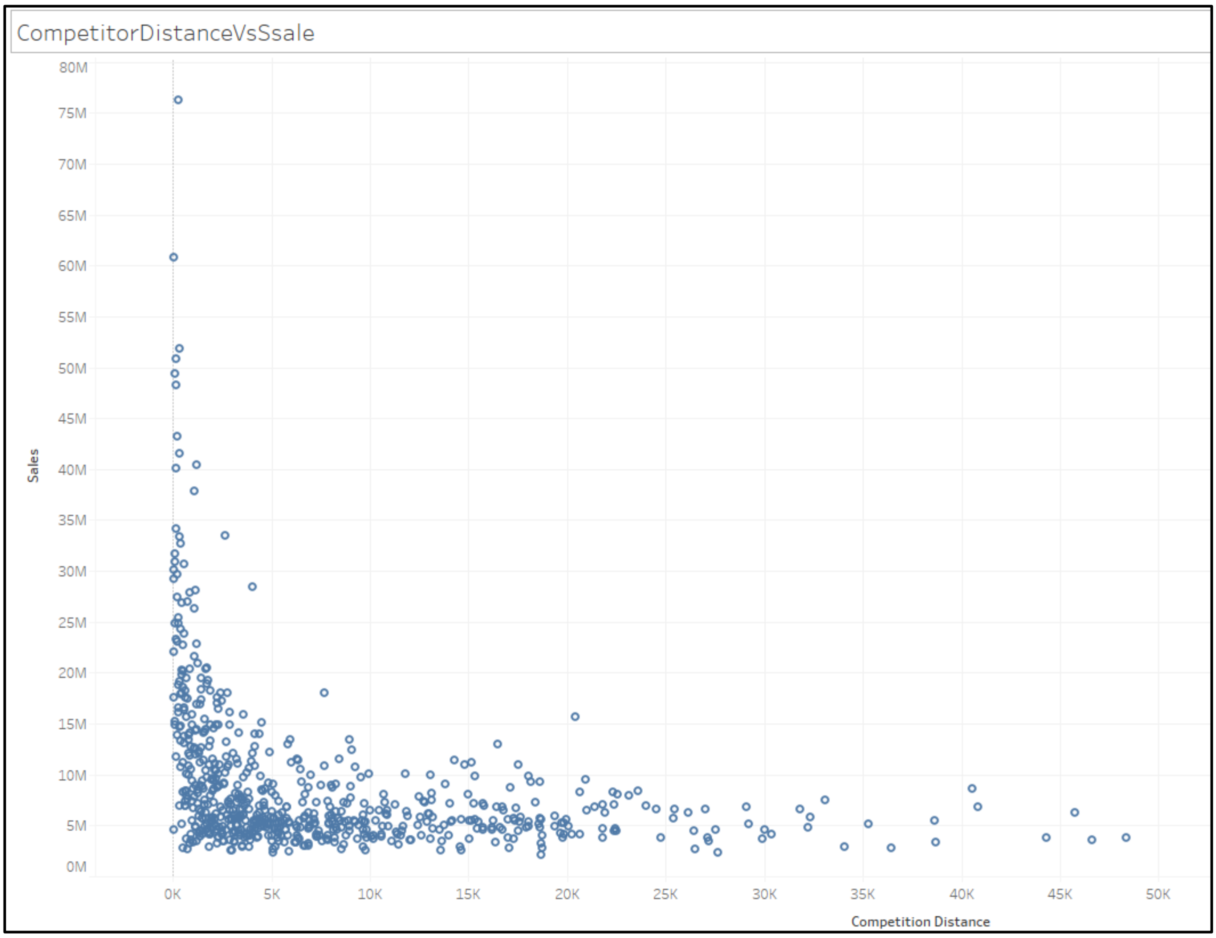


Figure 17:Scatter plot Sales Vs Competitor Distance

The Scatterplot showed negative correlation between Sales and Competitor distance with high magnitude.

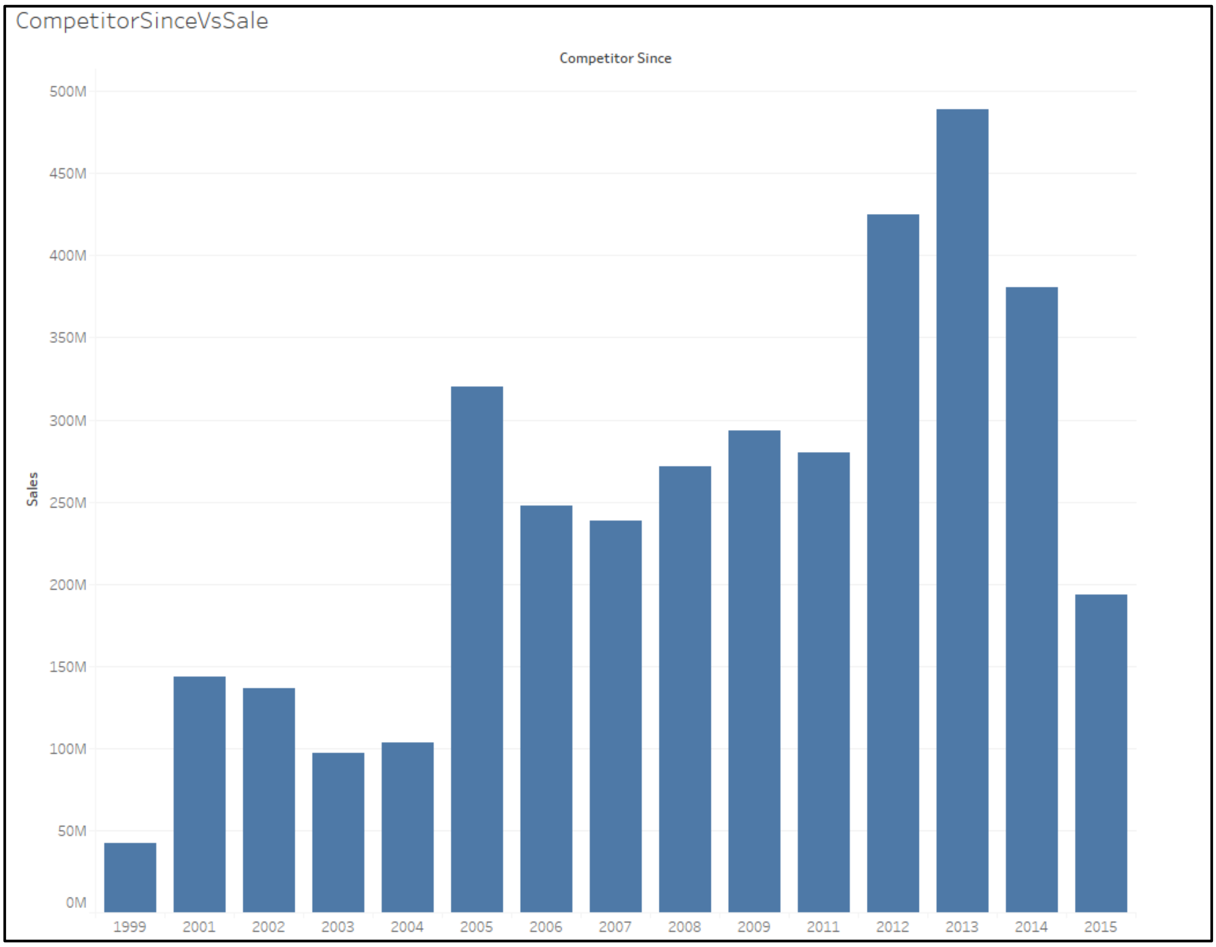


Figure 18: Sales VS Competitor Since

We use **filters** function to filter any Competition data below 1999 as the data removed were not relevant to the study (1961,1981, and null). Based on the finding, it can be observed that any recent establishment of the competitor outlet influence the increase of sales for the Rossman store.

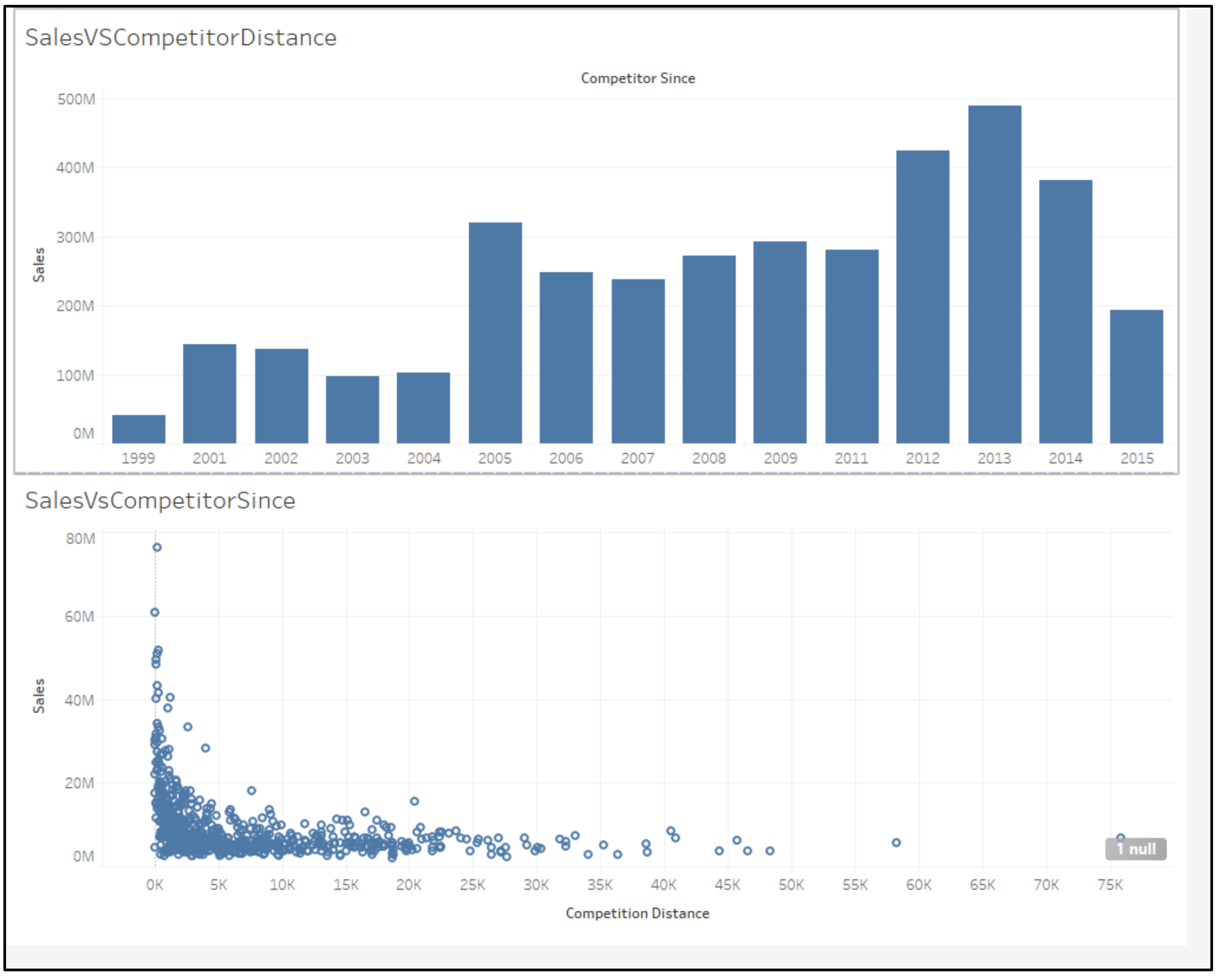


Figure 19:Dashboard for Sales Vs Competitor Factors

We can conclude that the closer the Rossman stores outlet to the competitor's distance and the recent the competitor open their outlet, the higher the total sales of the Rossman Store will be.

Google Trend Analysis

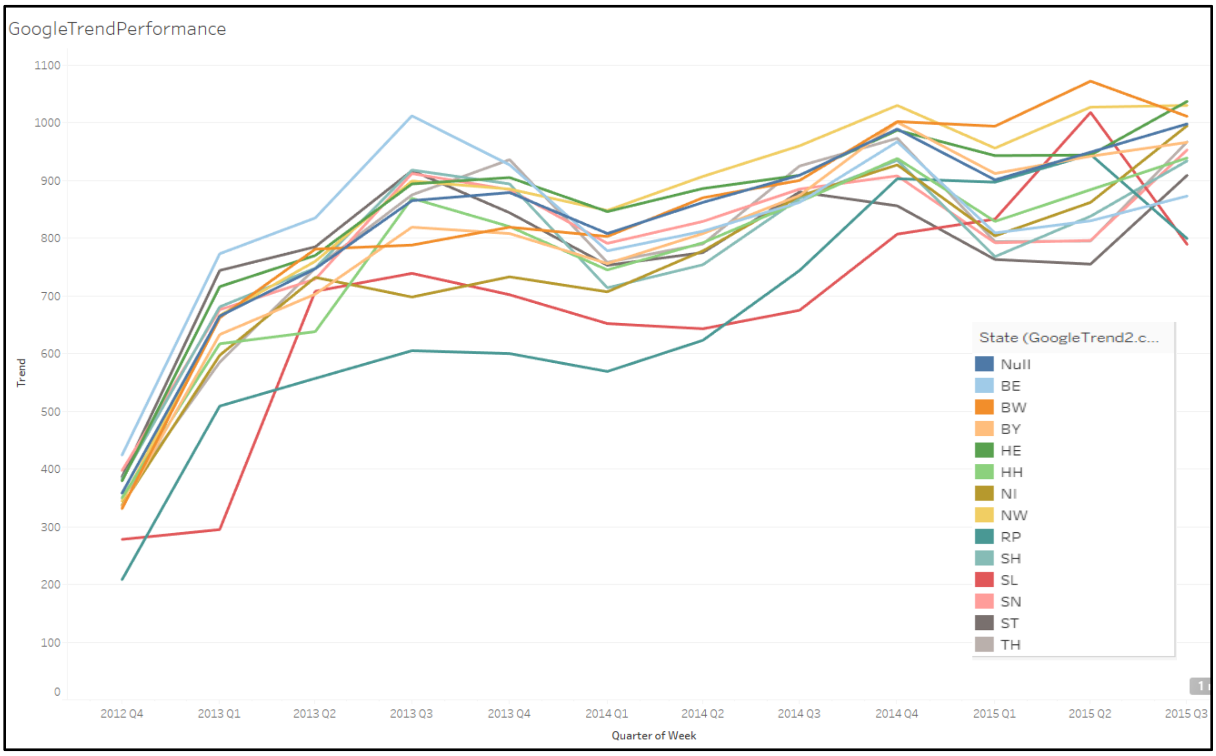


Figure 20:Google trend performance

To assess the performance of google trend over time. We use line continuous diagram. Based on the performance of the line chart, it can be observed that the google trend showed positive correlation over time.

Conclusion

- Total sales has positive correlation with total number of customer with high magnitude
- Store from state NW show the highest Total Sales ($ 1.5 billion) and highest total number of customer (178 mil). While Store HB,NI showed the lowest Sales ($131 mil) and lowest customer (17 mil)
- But in term of per store sales and customer, Store BE showed the highest revenue per store ($6.572 mil) and highest customer per store (6978) with each customer spend 8$ on average. While the NW store was ranked in number 4 for in term of revenue per store ($5.59 mil) and total number of customer per store (5935) with each customer spend $9 on average.
- Based on the box-plot, the sales performance for lower period of the year (Jan-June) perform better than upper half of the year.
- In term of store type, Store type a bring the highest revenue (54%) followed by d(30%), c(13%) and d (2.7%). Bad performed of store type based on state was shown by store b whereby all the performance of store b for all states were considered bad ($ 50 mil and below)
- In term of promotion, the promotion that start from 2011 showed the highest revenue followed by 2013 and 2014
- In term of day of promotion, the were not much difference in term of sales shown for day with promo(10) and day without promo (1)
- In term of interval of promo, Promo Interval of "Jan, Apr, Jul and Oct" contribute to the highest revenue (59%)
- Based Weather Study, It is shown that even in rainy season, the highest customer count come from the event of weather (240 mil customer) compare to day without rain (150 mil)
- Based on the study on effect of temperature toward customer attendance, it can be observed that there were no correlation. The attendance of the customers did not shown any significant changes during low temperature season.
- Based on google trend study, All the states showed positive correlation of search querry
- Based on the Competitor study, the recent the competitor open their establishment, the higher the sales
- the study of the competitor distance toward the sales showed negative correlation, which mean the closer the competitor establishment toward the Rossman store, the higher the sales that the Rossman generate
- Recent establishment of the competitor outlet increase the total sales.
- Based on Google trend study, we can see that the google trend query is increasing as the time progress.