

Drugstore Sales Analysis

Group – 9

BUAN-6356 | Business Analytics with R | 9 December 2018

Contents

[Executive Summary 2](#_Toc532157388)

[Project Motivation 2](#_Toc532157389)

[Data Description 2](#_Toc532157390)

[Exploratory data analysis 3](#_Toc532157391)

[Day Level Patterns 12](#_Toc532157392)

[Models and Analysis 23](#_Toc532157393)

[Findings: 27](#_Toc532157394)

[Conclusion 32](#_Toc532157395)

# Executive Summary

* The Drug Store operates over 3,000 drug stores in 7 European countries. Currently, the 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.
* We are focused on predicting the sales for the 1115 stores that are located across Germany. By forecasting the Sales, it will enable the store managers to create effective staff schedules that increase the productivity and motivation. Our primary aim is to create a prediction model that will help the store manager to stay focused on what is most important to them, their customers and their teams.
* The data set contain information on all the stores that are in Germany. This dataset has totally 5 lakh observations and each observation row correspond to the sales in each one of the stores present across Germany. This Data sent has 14 columns which has details regarding the store type, number of customers visited on a given day, distance to the nearest competitor, state holiday, school holiday etc.

# Project Motivation

In this project we applied the various analysis and data exploration techniques to a real-word problem of analyzing and predicting the store sales. This will be able to help staff create proper and effective schedules that increase productivity and motivation. We have applied the statistical techniques learnt in this course using R.

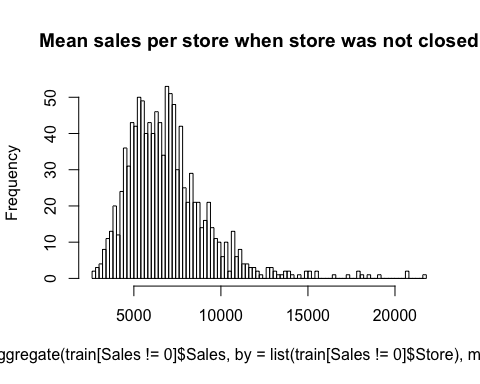
# Data Description

|  |  |  |
| --- | --- | --- |
| ·       Store | a unique Id for each store | |
| ·       Sales | the turnover for any given day (this is what you are predicting) | |
| ·       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 | |
|  | |  | |
|  | |  | |

# Exploratory data analysis

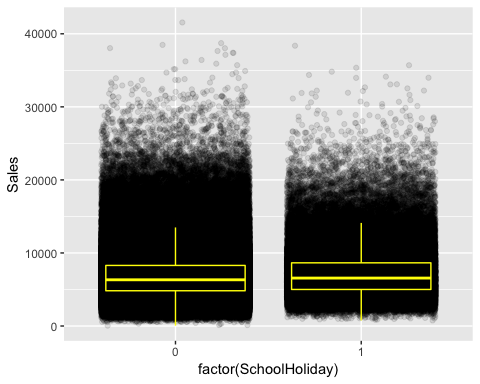
##### Trying to understand the data by plotting graphs

1. Mean sales per store when store was not closed

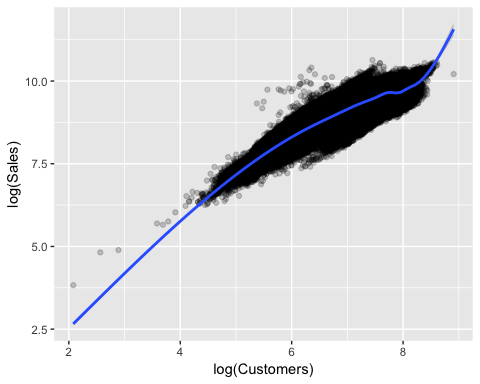


1. Effect of School Holidays on Sales

HA

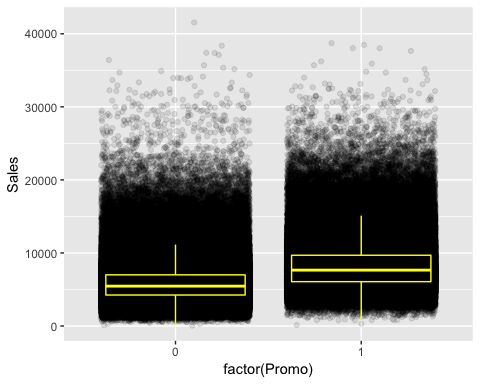


1. Relation between Sales and Customers

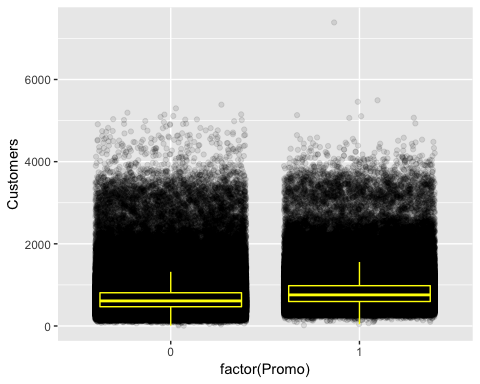


1. Effect of Promotions on Sales

HA



1. Effect of promotions on Customers HA

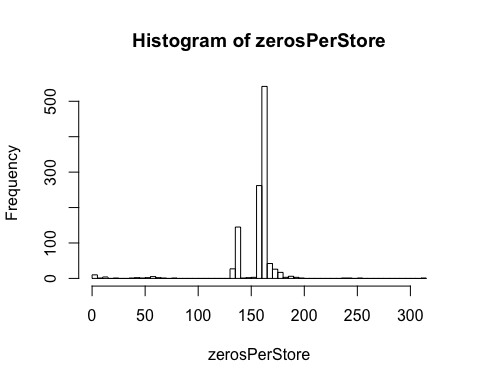


Sales, as expected, is strongly correlated with the number of customers. It looks like Boxplots of customers overlap a little more than the boxplots of sales. This would mean that the promos are not mainly attracting more customers but make customer spend ore. The mean amount spent per customer is about one Euro Higher.

1. Tabulating some more details:

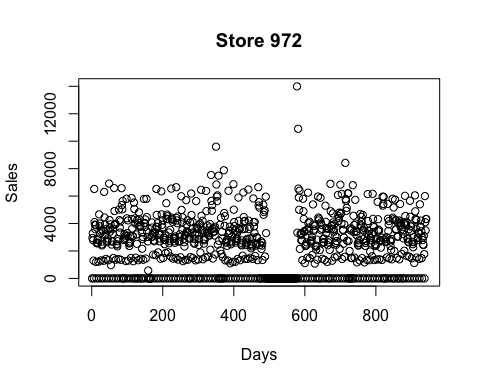
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | ***No Promo*** | ***Promo*** | | ***Sales = 0*** | *161666* | *11205* | | ***Sales>0*** | *467463* | *376875* | |  | |  |  |  | | --- | --- | --- | |  | ***No Promo*** | ***Promo*** | | ***Sales***  ***= 0*** | *161666* | *11205* | | ***Sales>0*** | *467463* | *376875* | |
|  | |  |  |  |  | | --- | --- | --- | --- | |  | ***Customers***  ***= 0*** | ***Customers***  ***>0*** |  | | ***Closed*** | ***172817*** | ***0*** |  | | ***Open*** | ***52*** | ***844340*** |  | |  |
| |  |  |  |  | | --- | --- | --- | --- | |  | ***Customers***  ***= 0*** | ***Customers***  ***>0*** |  | | ***Closed*** | ***172817*** | ***0*** |  | | ***Open*** | ***52*** | ***844340*** |  | |  | |  |  |  |  | | --- | --- | --- | --- | |  | ***Customers***  ***= 0*** | ***Customers***  ***>0*** |  | | ***Sales = 0*** | ***172869*** | ***2*** |  | | ***Sales>0*** | ***0*** | ***844338*** |  | |

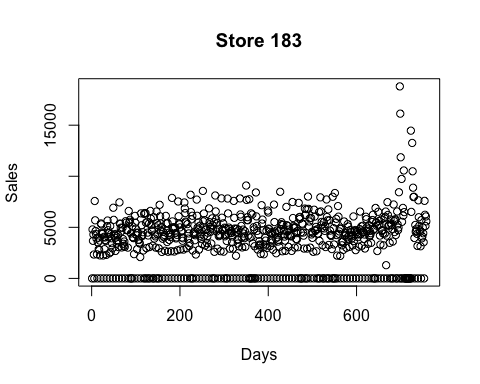
1. Checking zero sales per store HA

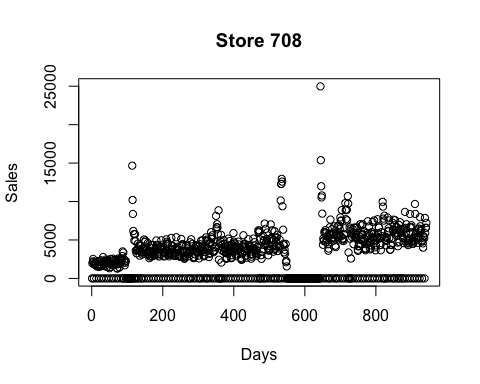


The stores have different amounts of days with zero sales. There ar espikes in e sales before the stores close and after the stores reopen.

**HA**Taking a random stores , for instance,



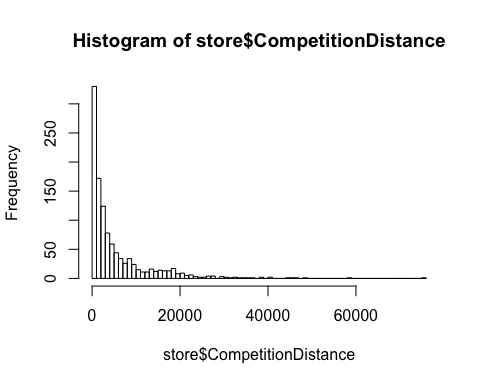




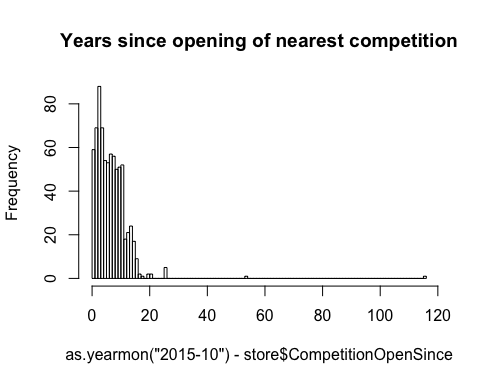
There are stores that have no zeros in their sales. these are exceptions because they are opened also on Sundays/holidays.

1. Taking a look into the competition

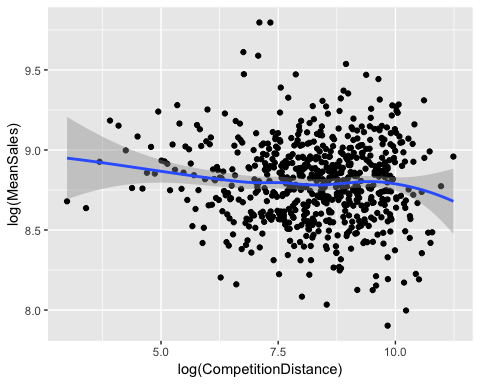
**HA**



A huge number of stores are closely located.



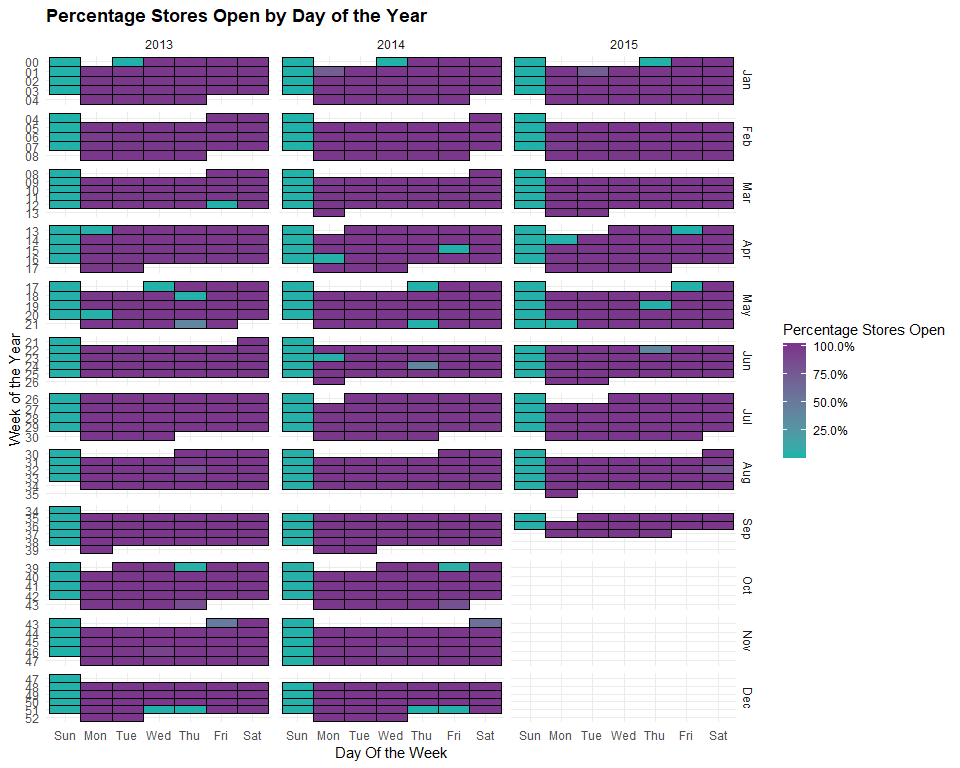
1. Effect of competition Distance



The effect of the distance to the next competitor is a little counterintuitive. Lower distance to the next competitor implies (slightly, possibly not significantly)higher sales. This may occur(my assumption) because stores with a low distance to the next competitor are located in inner cities or crowded regions with higher sales in general. Maybe the effects of being in a good / bad region and having a competitor / not having a competitor cancel out.

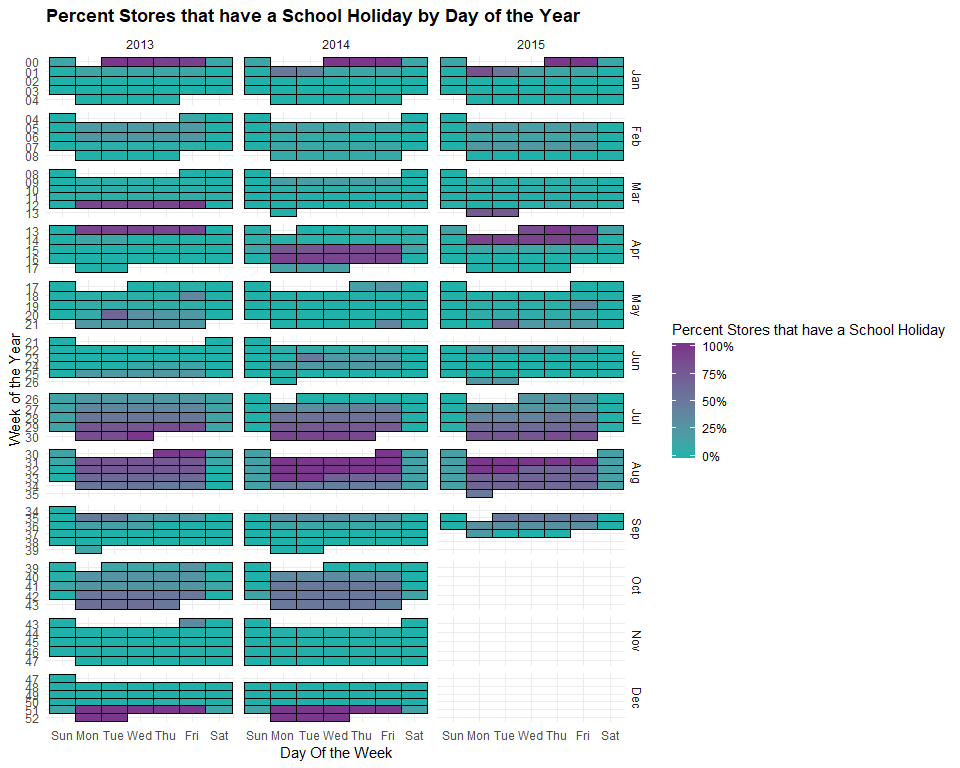
# Day Level Patterns

##### To see for every observation, the percentage of stores that are open:



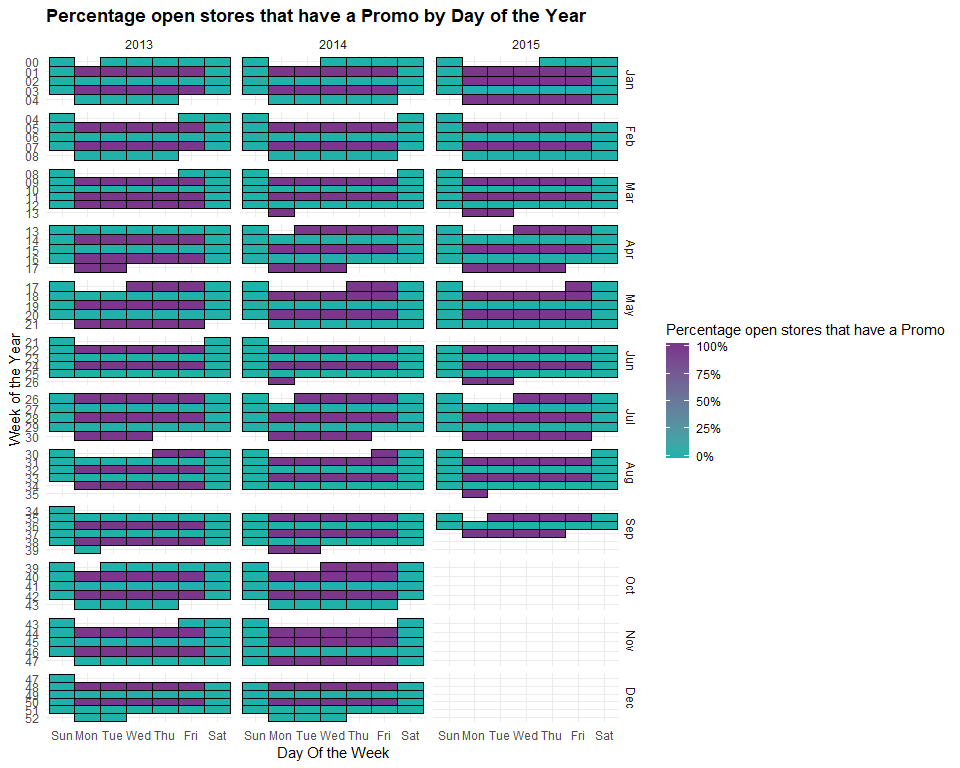
We can see from the plot above, 100% stores are open most days of the year. Mostly the stores are closed on Sundays and state holidays. The stores that are open on State Holidays, there must be a special reason as to why they are open when others are closed.

##### To see the percentage of schools that have a School Holiday:



We can clearly see that there is a seasonal pattern here. Most of the stores recognize school holidays between Late July and Early August, New Year’s week and the week prior. Some of the stores recognoze school holidays within October. We also see that the timing of the school holidays shifts from last week of March and first week of April in 2013 to two weeks in April in 2014 and then back to last week of March and First week of April in 2015.

##### The percentage of stores that have a promotion:



Evident from the plot above, there are only two situations possibles. THe store either has a promotion or doesnt. There was no variation in stores that had the promotion or those that did not. Every store had or did not have the promotion during a day.

Taking promotions into consideration, we can see a pattern. There is a promotion every alternate week. Also, we can see some back-to-back promotions. For example: March 2013, April - May 2013, April - May 2014, July - August 2014, September - October 2014, November 2014, January 2015, January - February 2015, and April - May 2015.

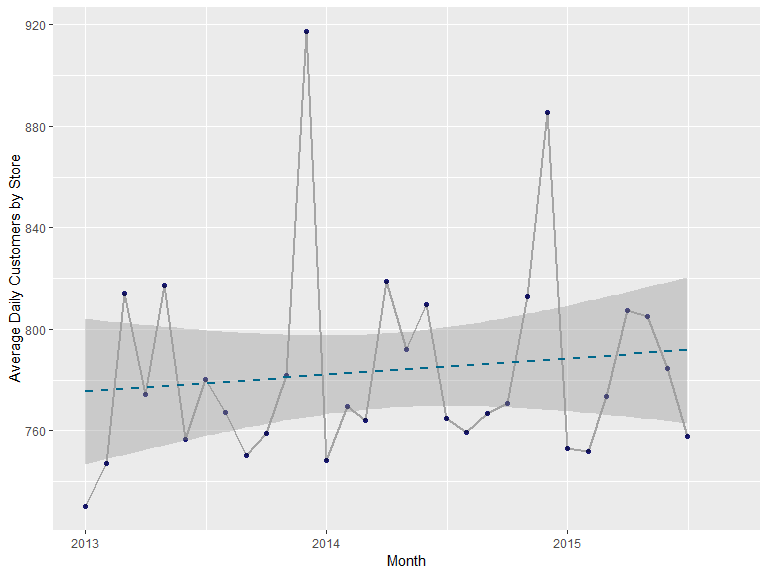
Comparing with the real-life scenarios, it is quite possible that the second week of these back to back sales has less sales because the promotion has been extended too long and has reached a saturation point.

#### Yearly-trends

We shall now attempt to study Yearly trends in the data. If we see some pattern in the trends, they could be due to some economic factors or other macro factors. Understanding them will help us in future predictions.

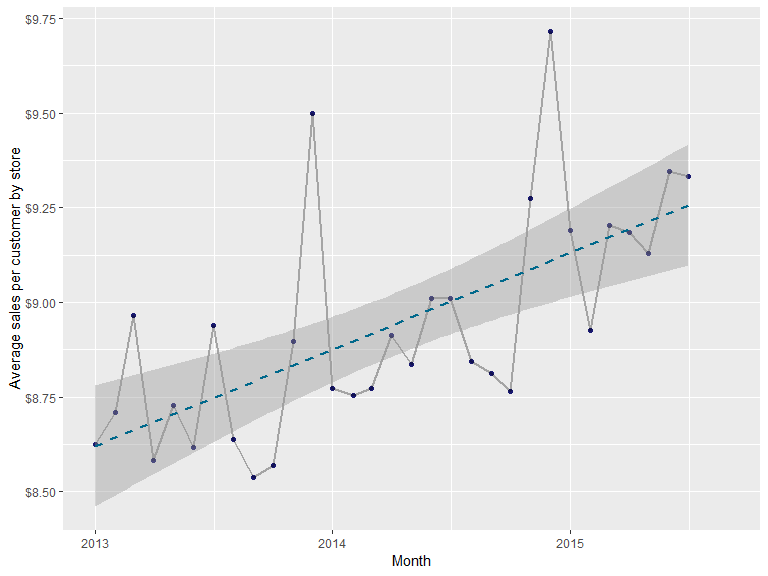
Because we are predicting the sales and how and what factors affect it. Sales will be our response variable or the dependent variable. We can try to find the relation between customers and average sales per customer.

##### Looking at the linear fit of the past sales



We can clearly see a surge in customers during the holiday season, i.e. December. Each year, March, April and May also perform well each year.

##### Plotting Average Sales per Customer in the past

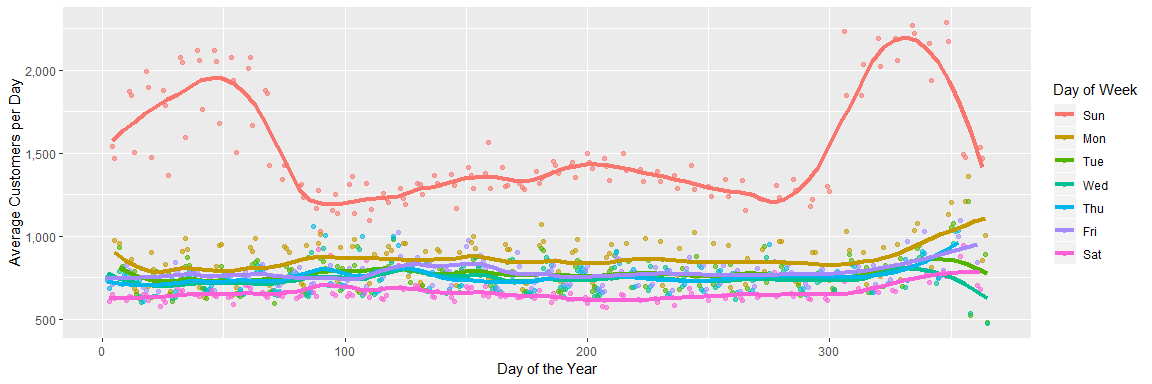


It turns out that December is the month with the highest average sales per customer. But, unlike the previous plot, March, April and May don’t seem to have the same effect on average sales per customer. The trend line seems steeper that the previous plot.

#### Within-Year Trends

In order to continue studying variation, we can study further in to the day patterns across years and taking into consideration the holidays both state and school and promotion patterns.

##### Getting the daily customer trends over the course of a year

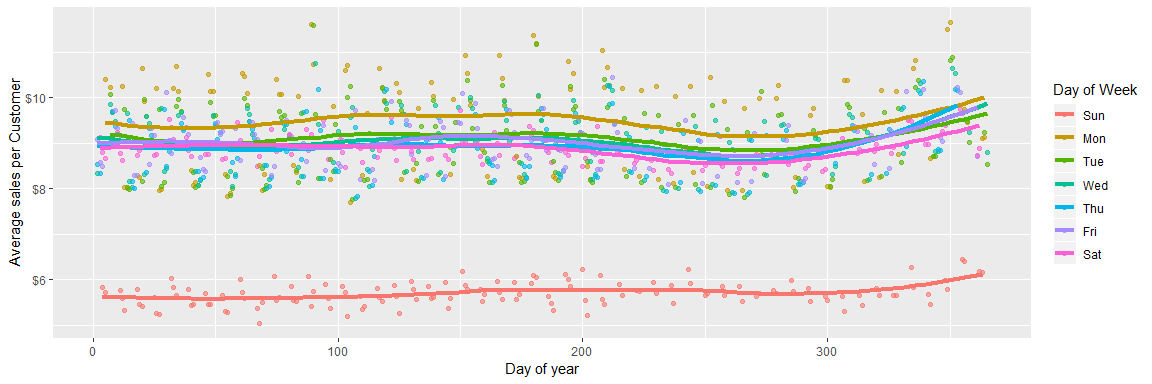


Sundays peak out the most in the number of customers. While Sundays, in general top in the number of customers among other days, the initial Sundays into the year and the Sundays to the end of the year peak in the number

We must remember that on Sundays, most of the stores are closed. Also, from the plot we can see that the relation ship between the days of the year and average customers is not linear, like for the other days. So, we can say that for the very few stores that are open on sundays, have a unique pattern of customers going on

The rest of the days are consistent throughtout the year. Mondays are the most popular among the rest of the days and Saturdays are the least popular. Also to be noted is that towards the end of the year, in December, there is a slight upward trend in the average number of customers except on Tuesdays and Wednesdays.

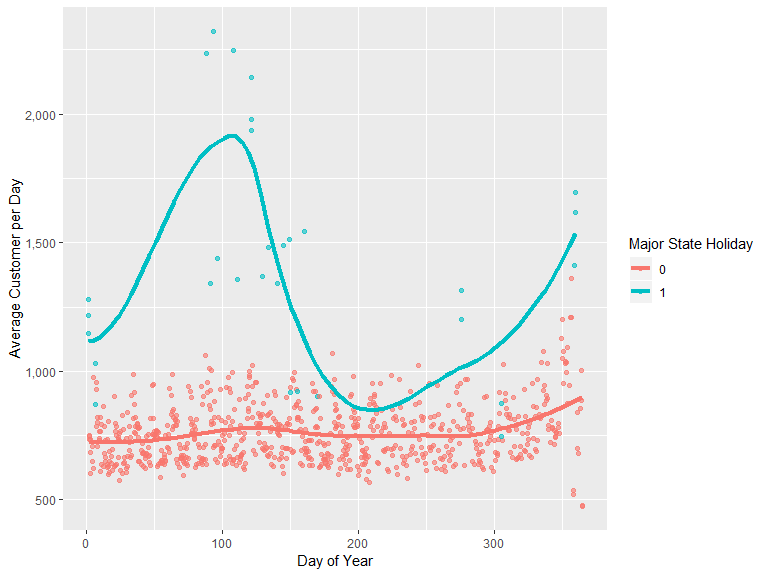
##### Plotting Average sales per customer every day of the week.



Monday has the highest average sales per customer and Saturday has the lowest. We can see a lot of variation in average sales per customer in the days. Similar to the previous plot, we see an upward trend at the end of the year.

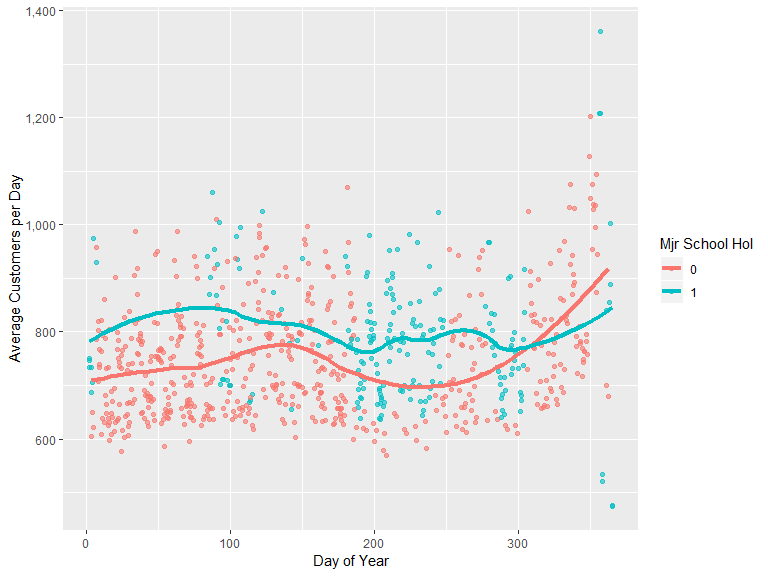
##### Studying State Holidays.

We put a condition previously, where we recognized major state holidays where more that 25% of the stores recognized that state holiday. We can study the customer trends during these major state holidays:



##### Studying School Holidays

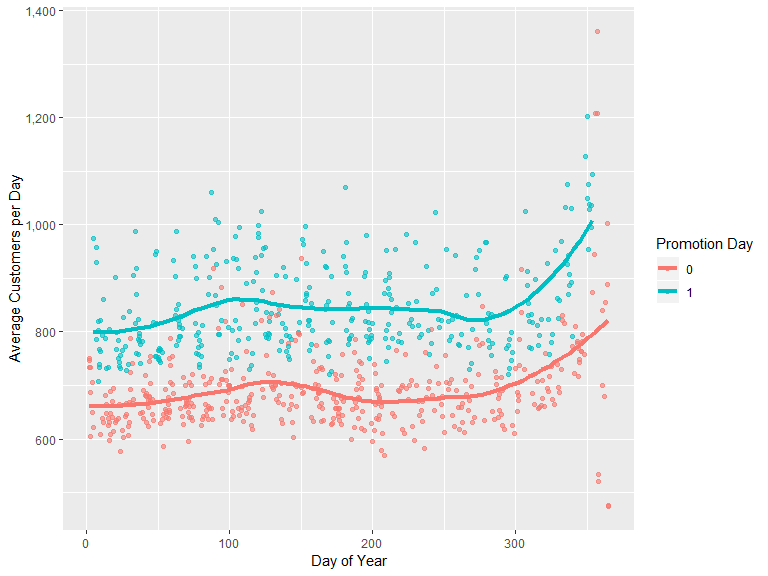
Major School Holidays are the ones where more than 25% stores recognize the school holidays. Studying them as below:



We can see that School holidays have more customers than non-school holidays. Also, the majority school days are around Day 200, which is Late July and early August. We see that similar to the other plots, we can see the similar pattern of heightened sales and number of customers around March, April, December and January.

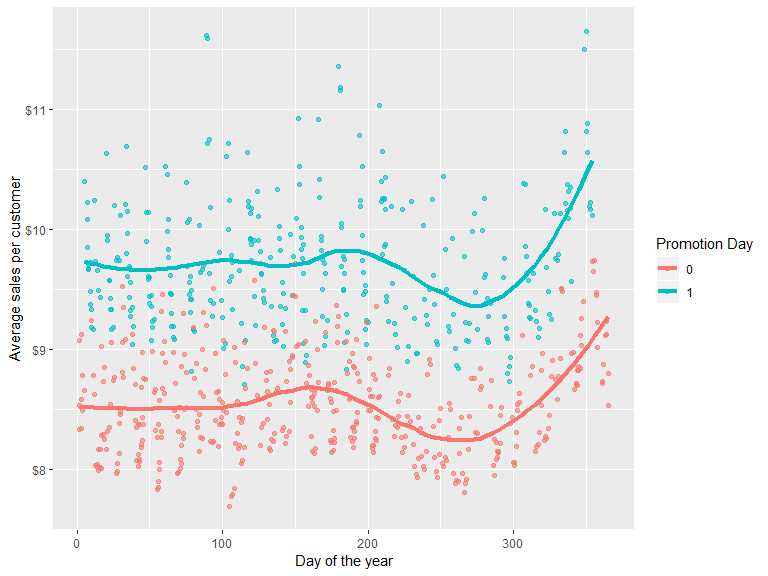
##### Studying the Promotion Pattern

##### Average Customers Per day during a Promotion Period.



Promotion days definitely have a greater number of customers that non-promotion days. We can see that, on average, the promotion days have 100 to 150 additional customers per day.

##### Average Sales Per Customer during a Promotional Period

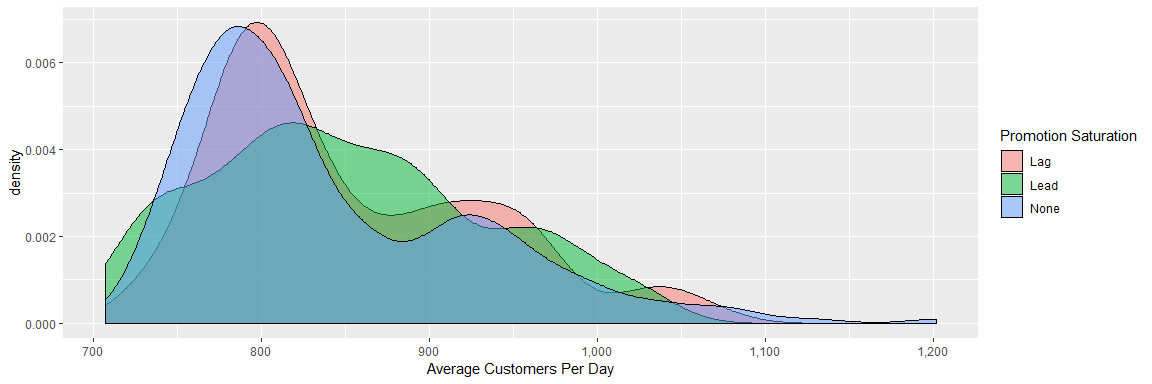


We see that promotions don’t just get more customers but more average sales per customer!. Also, the amount of sales is also high!

Now considering the promotions that were back-to-back for 2 weeks. We suspected that the sales might slow down due to promotion saturation. We can confirm our suspicions now!

We have mentioned “Lead” as the first week of back-to-back promotion and “Lag” as the second week. Below, “None” would correspond to the promotions that were there only for a week.

##### Checking the promotion saturation for average customers per day



We can see, from the above plot that Lag week and the None week, have very similar pattern. Whereas, the Lead week manages to get more average customers per day!

##### Checking the promotion saturation for average sales per customer

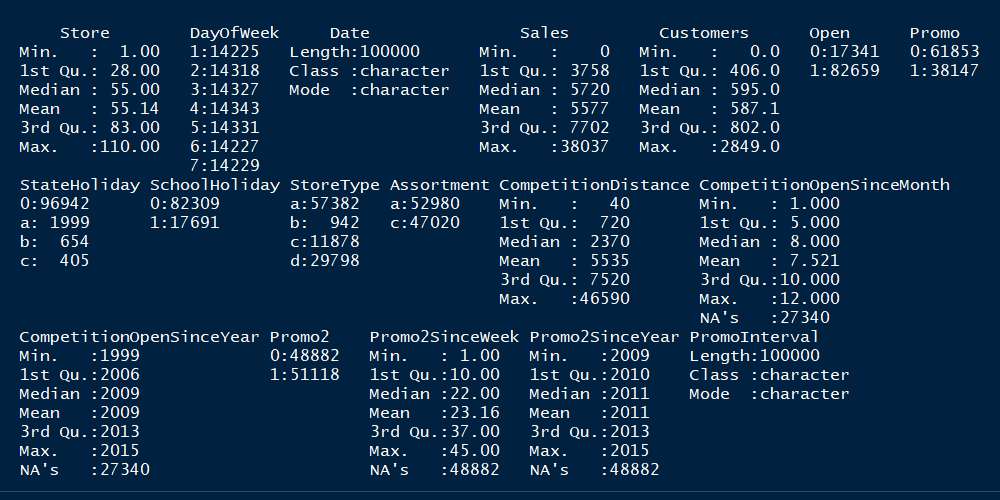
##### 

We can clearly see that average sales are more for Lead and Lag than for None.

We can witness a business strategy here. We see that back-to-back promotions averages more customers and higher sales than non-promotion weeks. It is only logical to hold these promotions frequently after taking into consideration the related costs.

# Models and Analysis

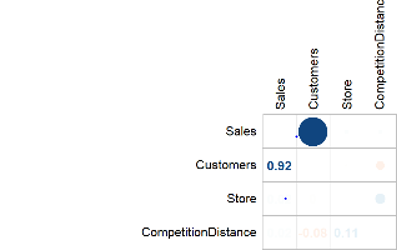
##### General Summary of the data:



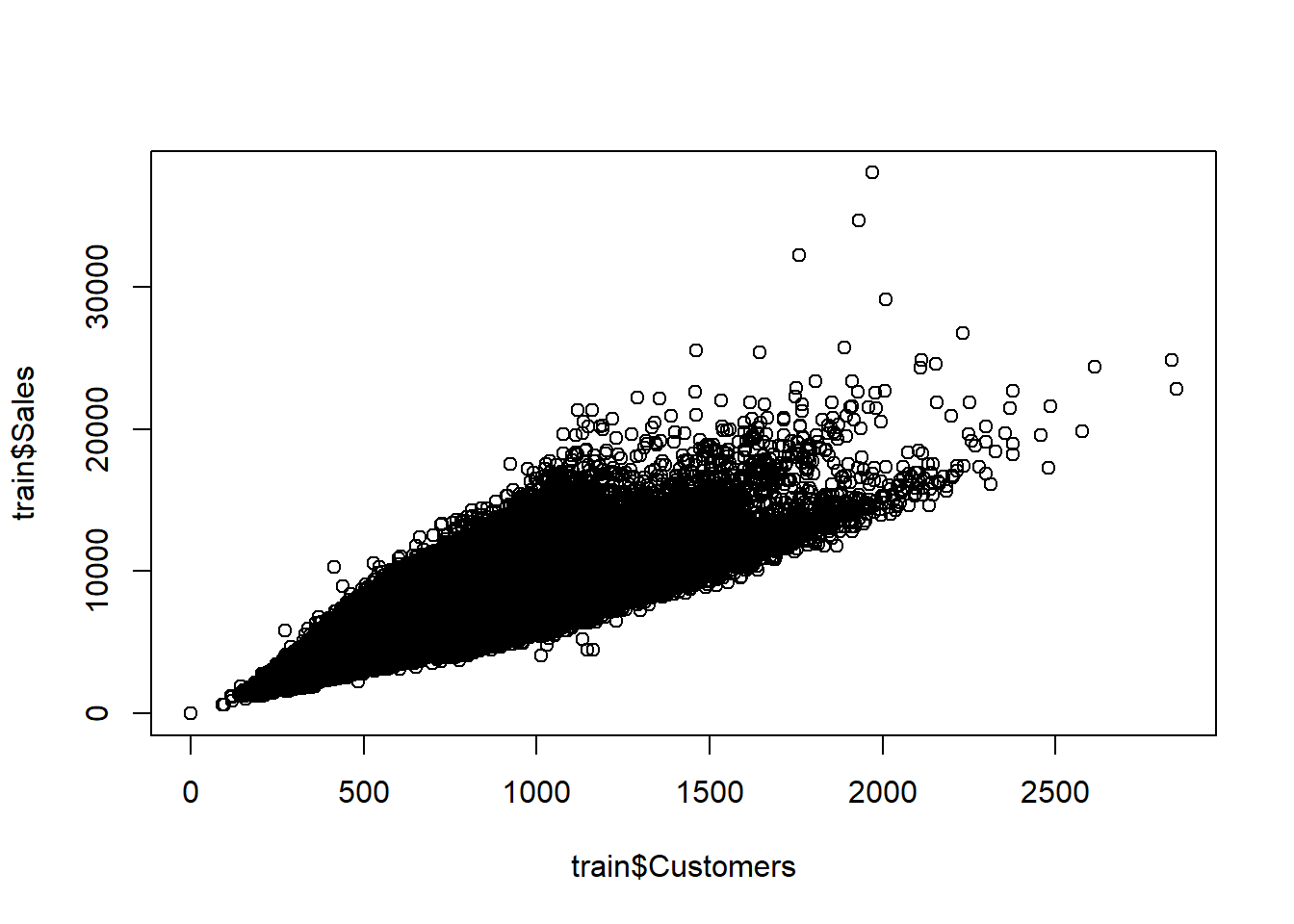
This gives the general summary of all the variables in the data set. It provides us with information such as minimum, maximum, mean, median, etc. for each variable.

##### Predictions

Correlation between the dependent and independent variables:

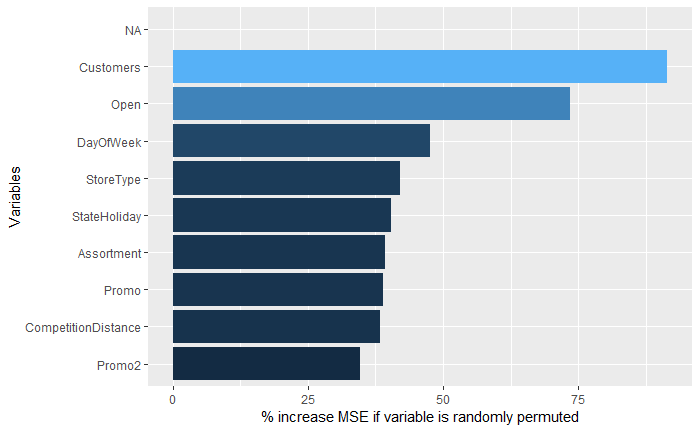


*This graph gives the correlation between the dependent and independent variable. There is seems to be a great positive correlation between the number of customer and Sales. The correlation is 0.92 which is close to 1. This indicates that there is a 92% positive linear relationship between them which says that as the number of customers increase the Sales correspondingly increase..*



Scatter plot displaying the positive linear relationship between Sales and number of customers

##### Finding variable importance with a quick Random Forest



Though correlations give a good view of important numeric variables, we perform Random forest to get an overview of the most important variables including the categorical variables before visualization. This shows that Customers, Open, DayofWeek have greater importance compared to the other variables which are present.

##### Linear Regression:

Linear Regression is a simple approach to supervised learning. Although true linear functions are never true, we still use this method because it is easy to interpret.

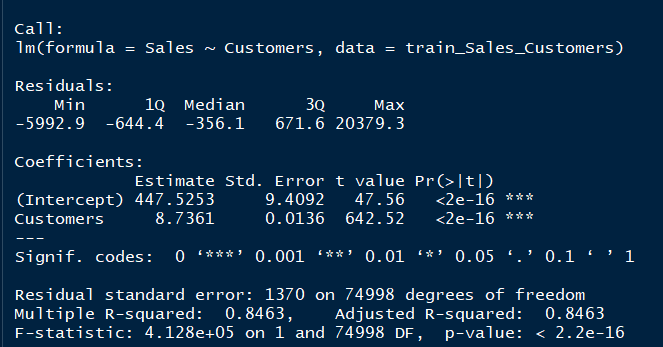
The population regression line is:

Y = Β0 + Β1X

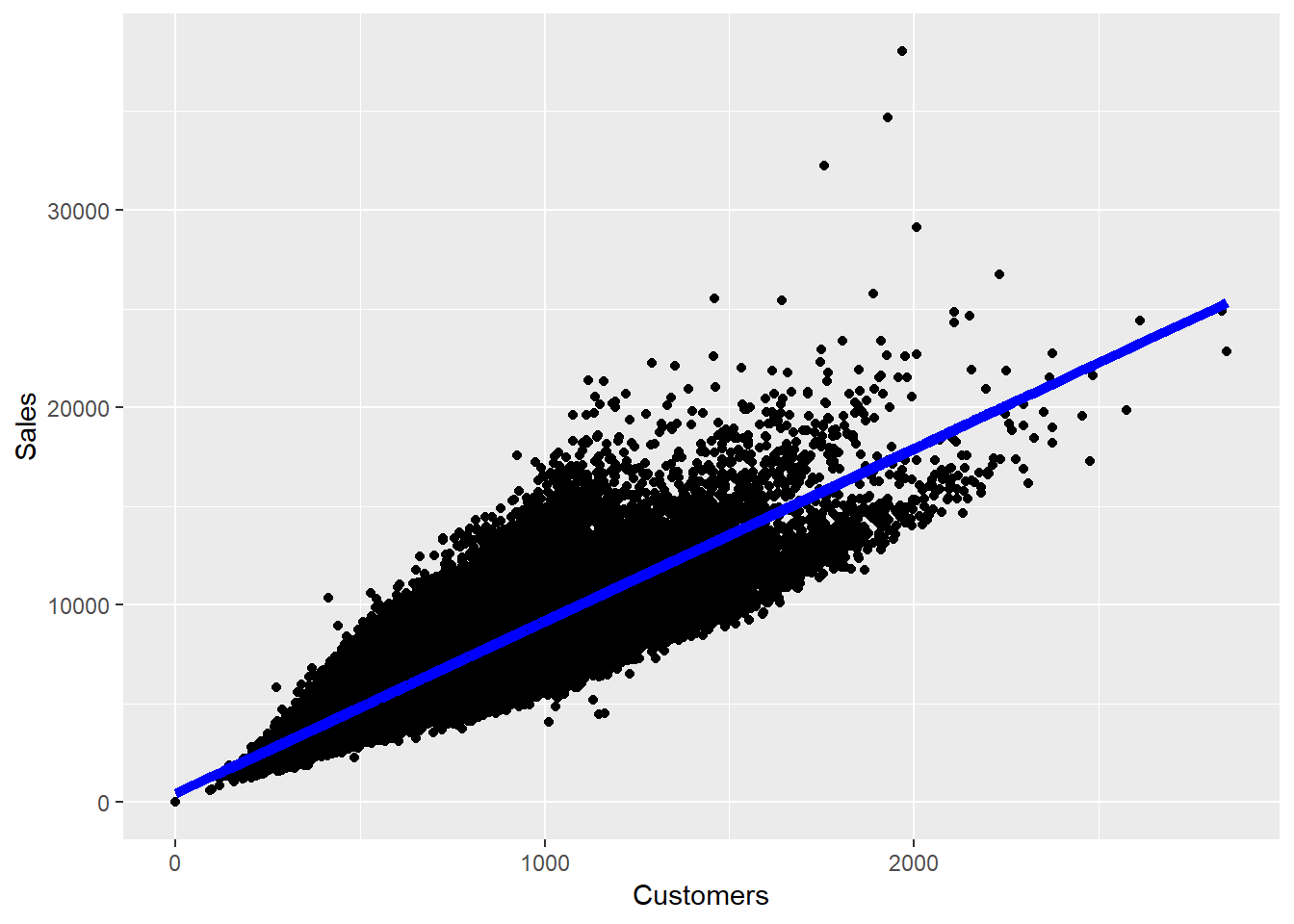
Given a random sample of observations, the population regression line is estimated by:

ŷ = b0 + b1x

where b0 is a constant, b1 is the regression coefficient, x is the value of the independent variable, and ŷ is the predicted value of the dependent variable.



Running the linear regression model for Sales and Customers. The p-value is found to be extremely low and thereby we reject the null hypothesis in support of alternate hypothesis. The R squared value of the model is 84.63% which shows that 84% of the variation in the Sales is explained by the customers and the remaining 16% of the variation is unexplained.



The predictions are plotted, and the blue line indicates the prediction line.

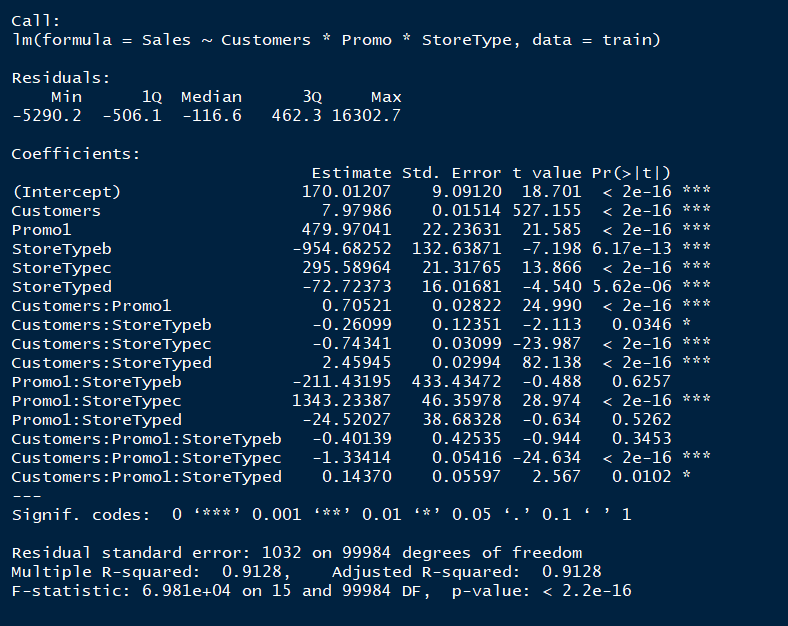
##### Multiple linear regression models:

Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the relationship between the explanatory and response variables.

The model for MLR, given n observations, is:

yi = B0 + B1xi1 + B2xi2 + ... + Bpxip + E where i = 1,2, ..., n

Multiple regression between Sales and Promo, Customers and Store Type.

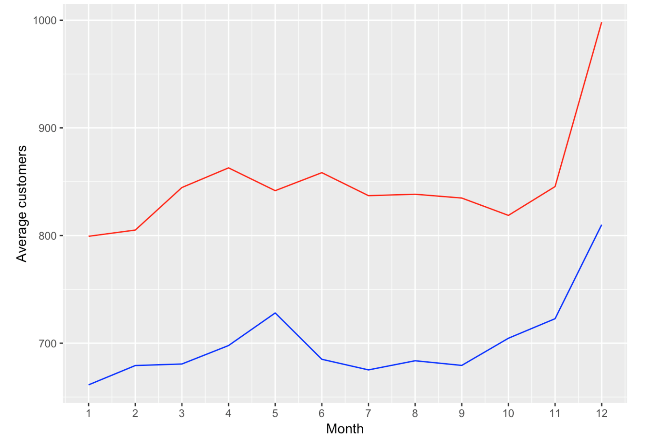
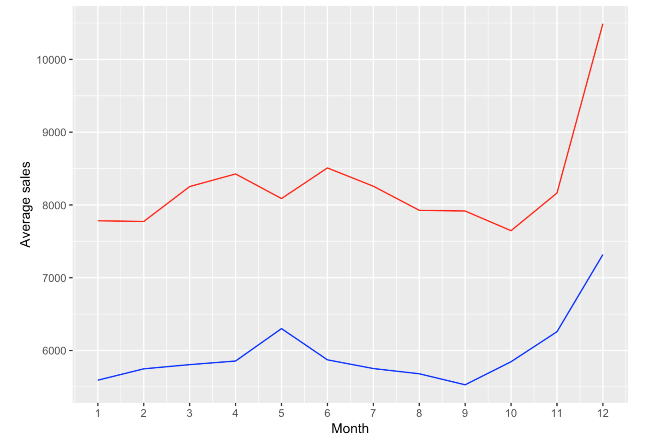


We can see that the R square value is around 91.28% and also the coefficients for each of the independent variables indicate that most of the variables plays a significant role in determining the sales. However the p value of customers: storetypeB, promo1:storeypeB ,promo1:storetypeD and customer:promo1:storetypeB is pretty high which indicates that it supports the null hypothesis and thereby it does not affect the sales. The overall p-value of the model is less and thereby we can say that this is a significant model.

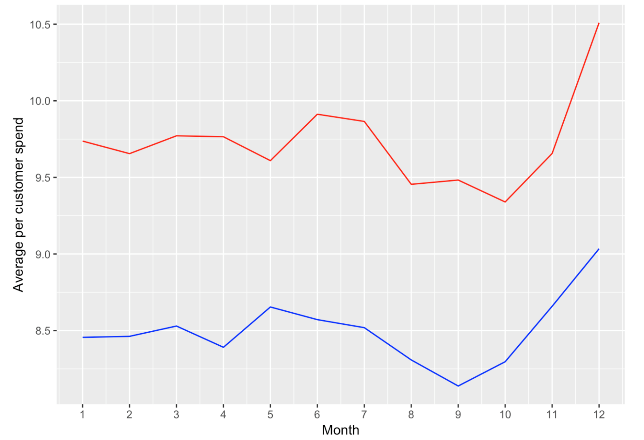
# Findings:

##### Promotion vs. sales & customer:

1. Average sales and customers with promotion is more higher than no promotion, especially in December.

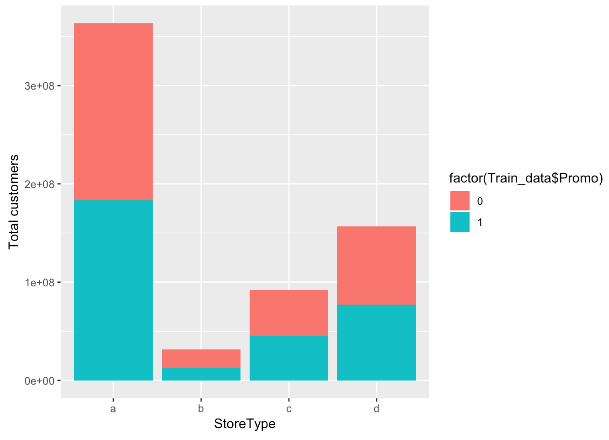
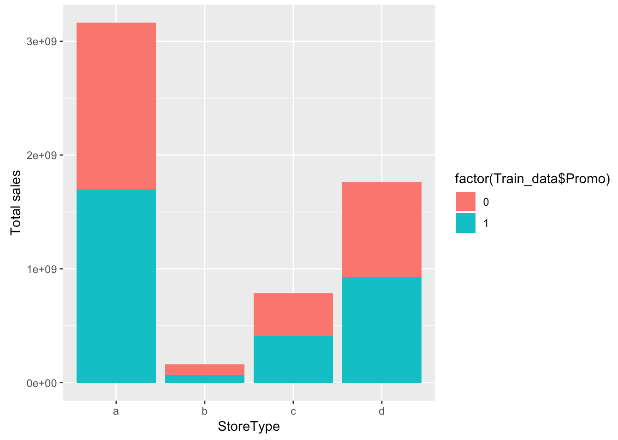


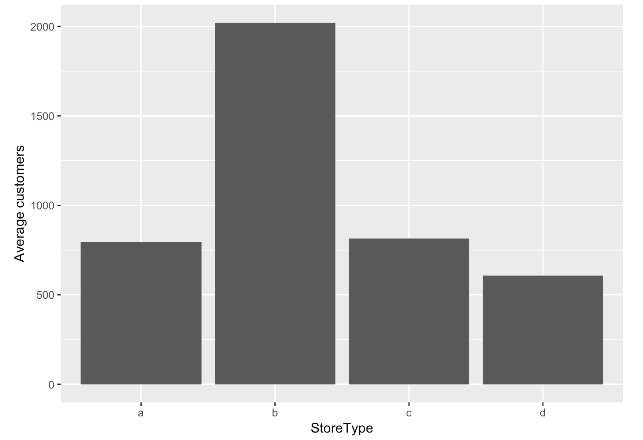
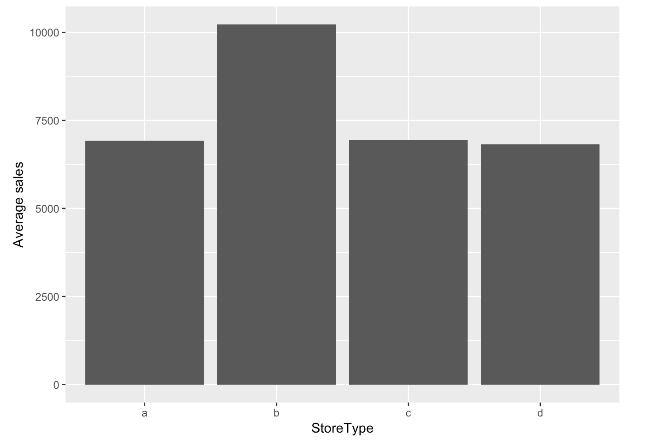
1. Sales per customer spend with promotion is slightly higher than without promotion.
2. Sales per customer spend among months are not much significantly difference, but still can observe that having promotion in Dec. can increase more sales and customer.



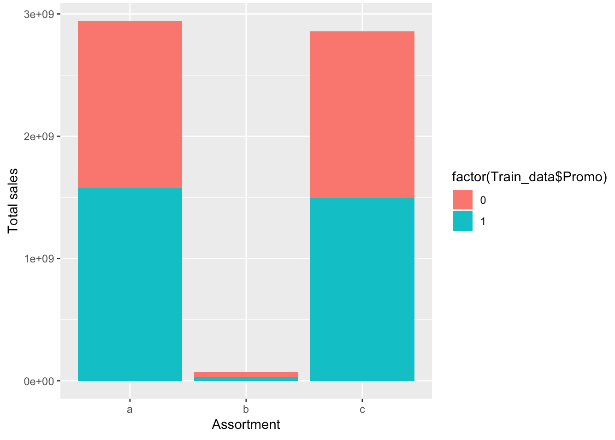
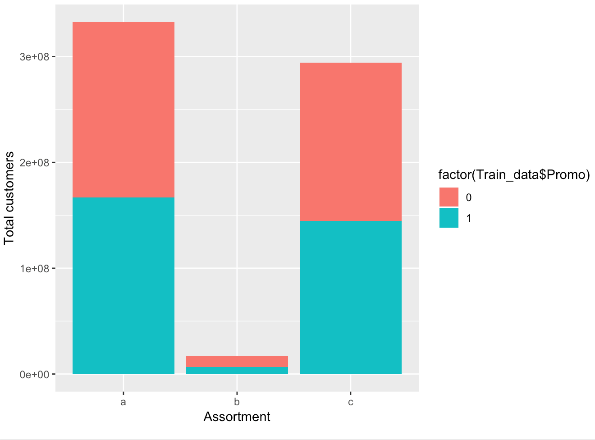
##### Store types & Assortment vs. sales and customer:

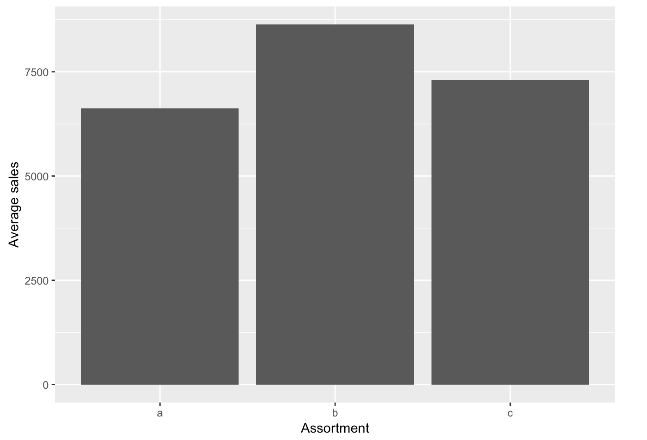
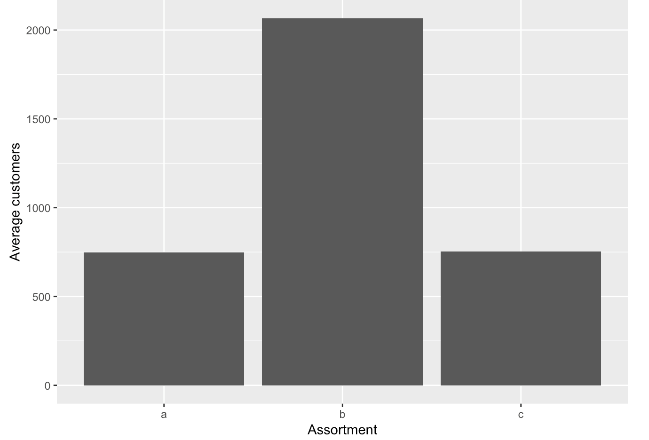
1. Total sales, customers and sales per customer spend of store type B are lower than other store types, but the average sales and customers are higher than other store types, it may because the number of type B store is very few.



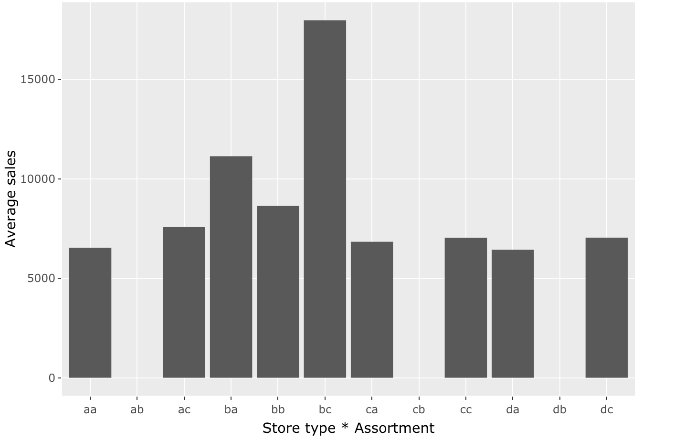
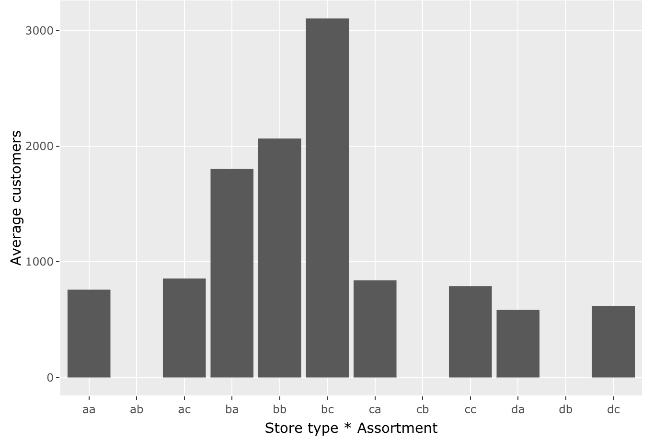


1. Total sales, customers and sales per customer spend of assortment B are lower than the other assortments, but the average sales and customers are higher than other assortments, it may because the number of assortments B is very few.

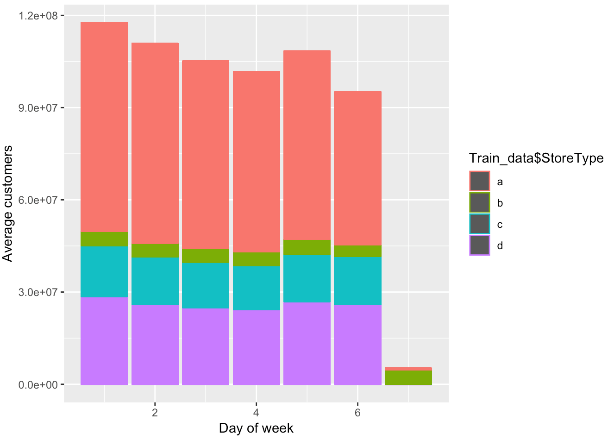
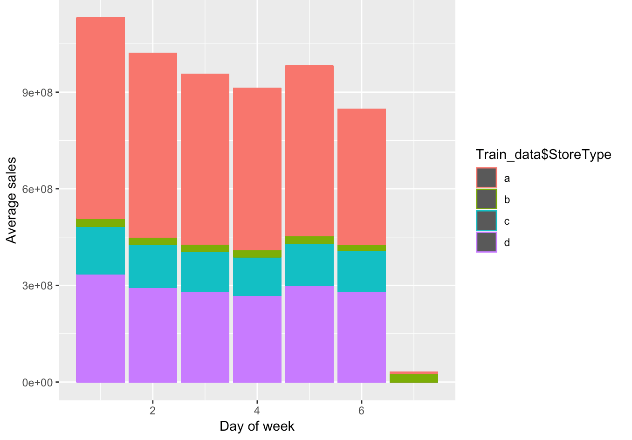
 

3. The store type B with assortment C has the highest average sales and customers.

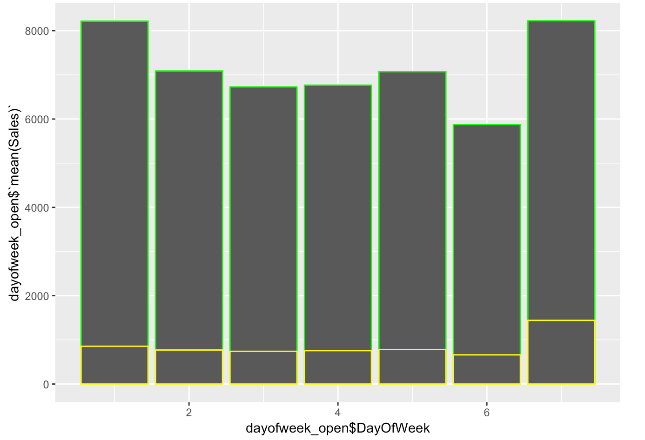
 

##### Day of week vs. Sales& Customers:

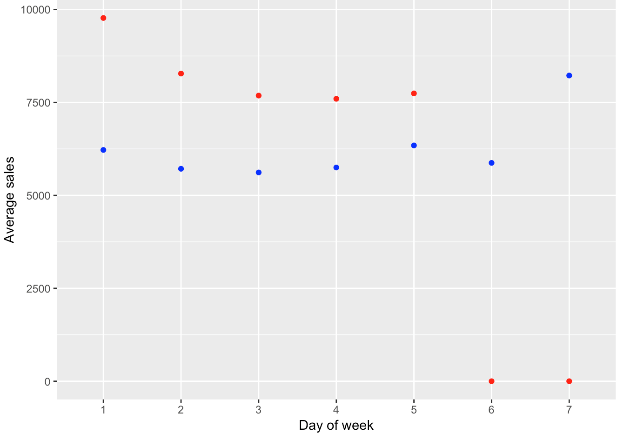
1. The sales and customers on Sunday are very low because few stores are opened and store type C not open on Sunday.



1. While Sunday has the lowest sales and customer, its average sales and customer are the highest. It may because there’s only few stores opened. Its sales per customer spend is also really high even there’s no promotion on Sunday.



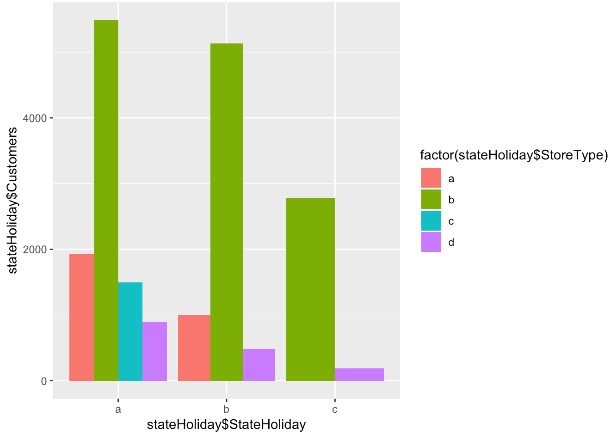
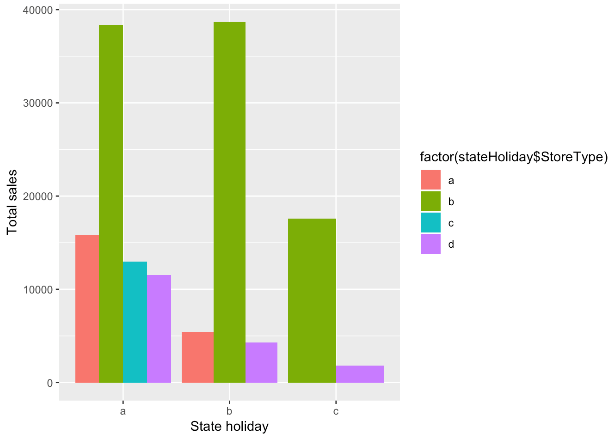
1. It seems like Monday has the highest total sales and customer, and its average sales and customer are also very high. Sales per customer spend is the highest on Monday when stores have promotion.
2. Promotion definitely affect sales and customers, surprisingly Saturday and Sunday don’t have promotion.



##### State Holiday:

1. No promotion running on any stores on Christmas (c), only store type B and D opened on Christmas and store type C not opened on Easter day.

2. Store type B has the significant total sales on each holiday.



# Conclusion

As we can see, the number of customers is the most important factor that affect the sale, so the goal of this company should try to increase the daily customers of each store. The data shows that running promotion is a way to increase promotion. In addition, because average sales and customer of store type B are higher than the other types of store, and store type B with the assortment C makes the most average sales and customers. Besides, store type B can make the most sales and customer on state holiday. However, store type B seems very few now, open more store type B may be helpful.

# 