MSIS 2506 R Programming Project 2
Bingbing Pan, Jialin Liang, Shruti Deshpande, Jaskaran Kohli, Tam Anh Pham

Data Preparation

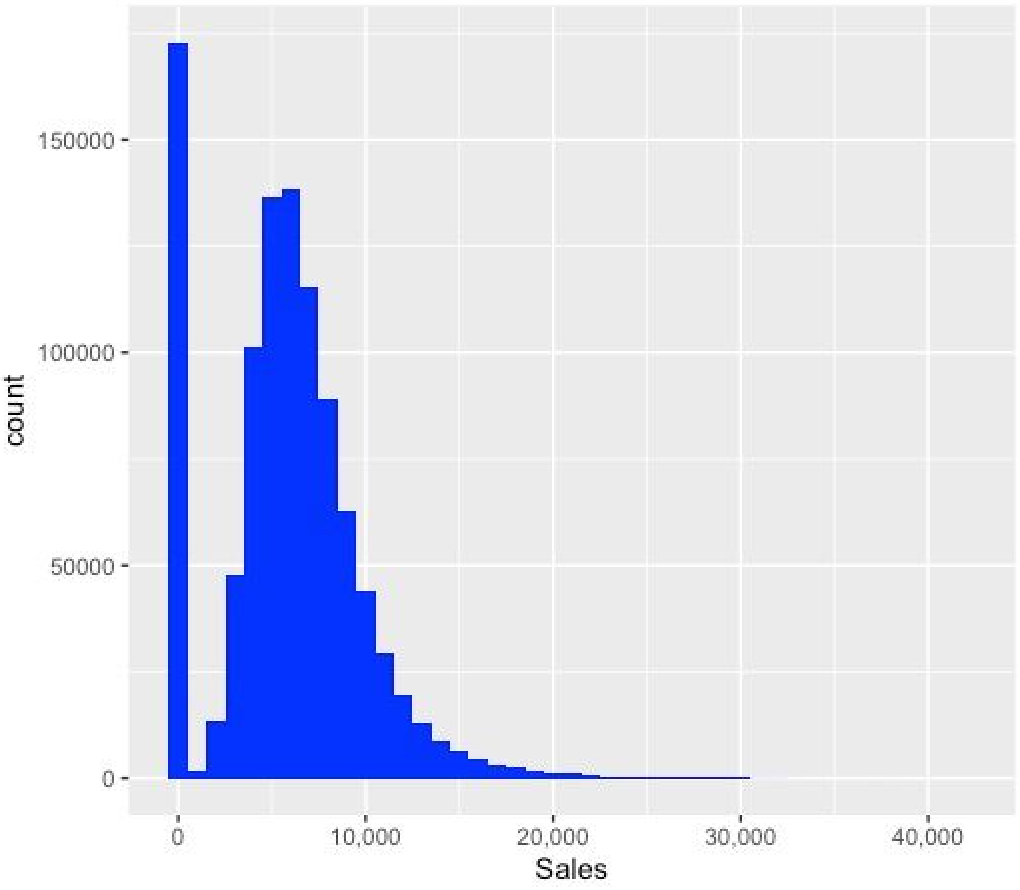
We participated in the Rossmann Store Sales competition at Kaggle. Our goal is to predict Rossman's daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

First of all, we got three datasets from the Kaggle, which are test dataset, train dataset and the dataset with additional information. Since train dataset only contains variables of store number, day of the week, date, sales, customers number, open, promo, state holidays and school holidays, we are not able to obtain enough information for the prediction. We gathered the information from train dataset and the dataset with additional variables by SQL so that we have a combined dataset with 17 variables regarding sales.

Exploring Data

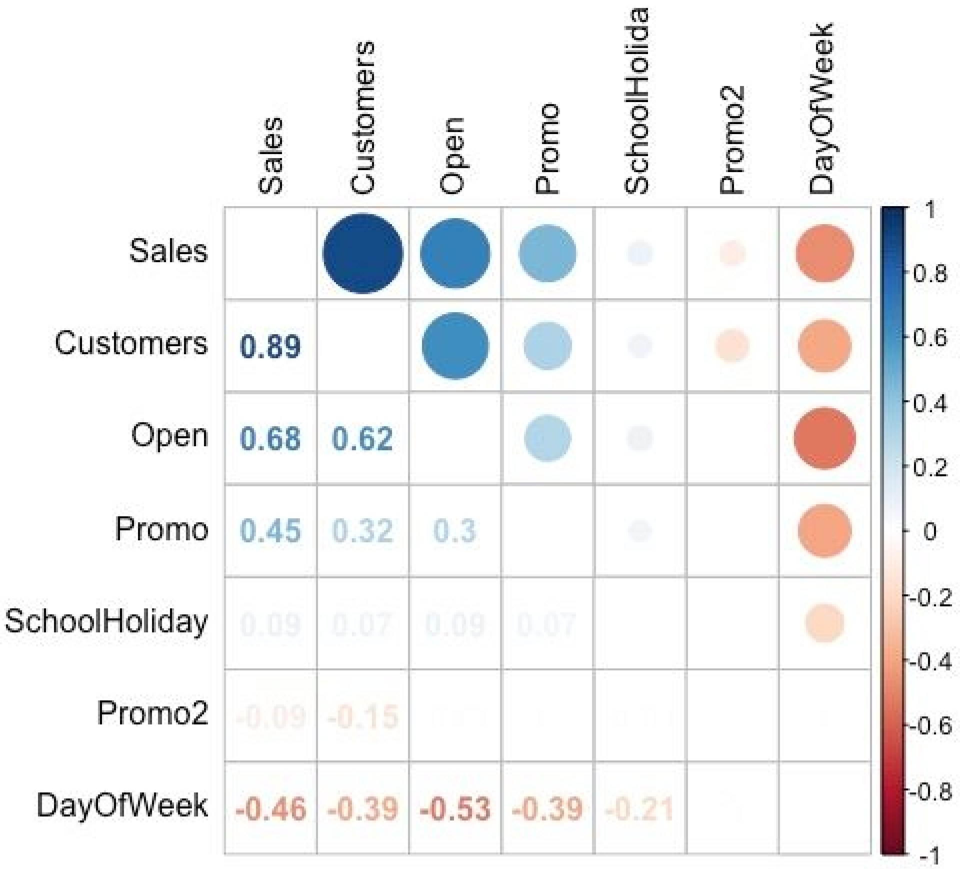
After preparing the combined dataset, we performed exploratory data analysis.

Graph 1 - Exploring sales data



We drew a histogram to visualize the distribution of sales. We found that the sales was skewed distributed and that there are many days with zero sales. We need to further investigate why sales were zero in these days.

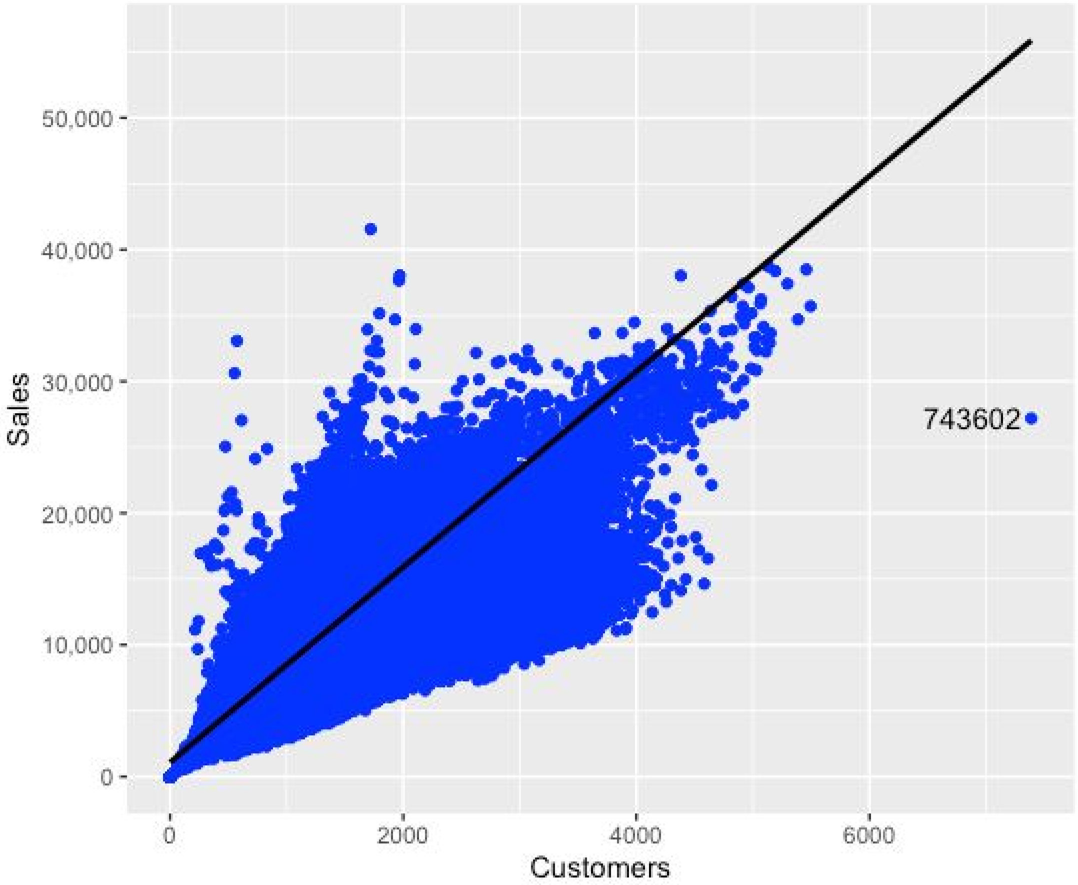
Graph 2 - Finding which numeric variable has the highest correlation with the sales.



The dataset has 13 numeric variables in total. While plotting the correlation graph, we decided to keep the top 7 correlated variables as the goal was to see which numeric variable has the highest correlation with the sales. After the correlation graph was created, we found that Customers and Open are the two variables with the highest correlation with Sales. Therefore, we decided to look into the relationship of Sales between Customers and Open.

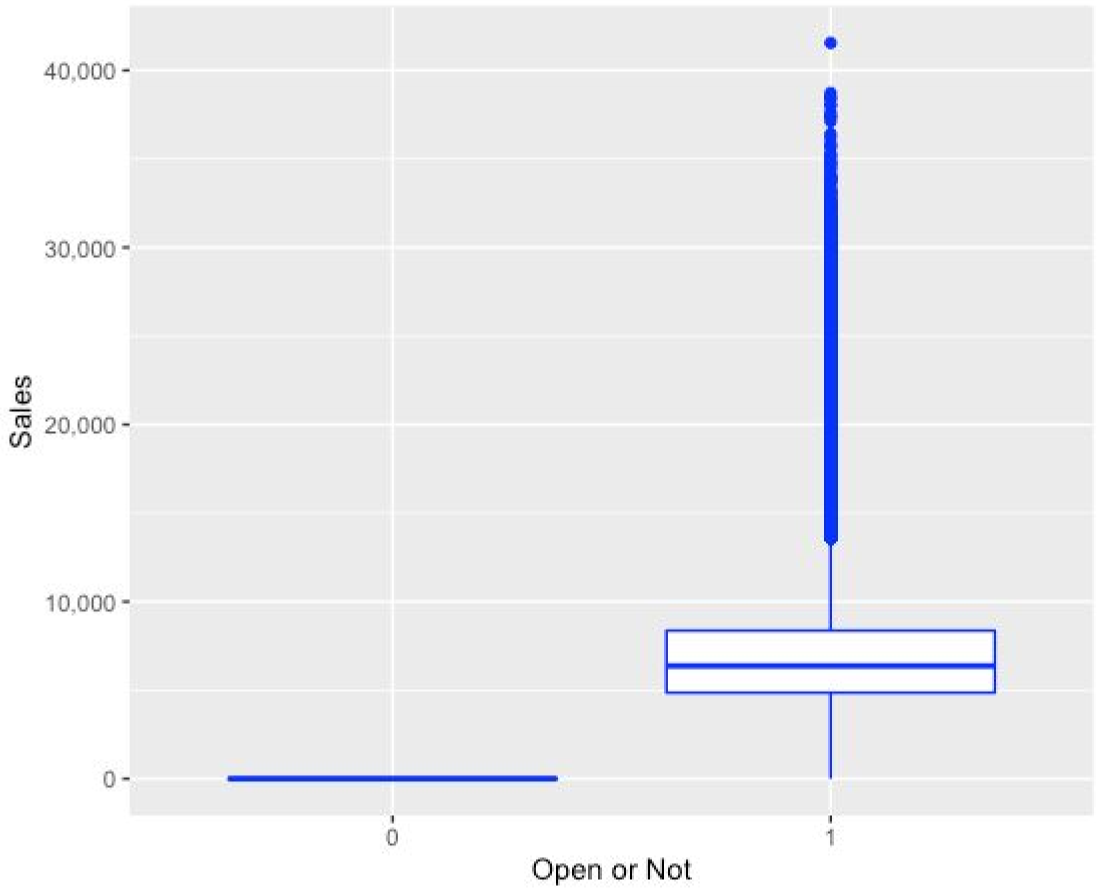
We further investigate the two variables with the highest correlations with sales.

Graph 3 - The relationship between customer numbers and sales



After we run the linear regression model, we find Customers and Sales are positive correlated, which means as the store getting more customers, the more sales will be generated. We also find an outlier, which is row 743602, this row of data has a relatively higher amount of customers, but normal level of sales. However, we will not take it out yet, as taking outliers can be dangerous.

Graph 4 - The relationship between sales and open



We further graphed the boxplot in order to find the relationship between sales and open. We have two findings in total:

1. When the store is closed, there's no sales.
2. The distribution of sales is skewed.

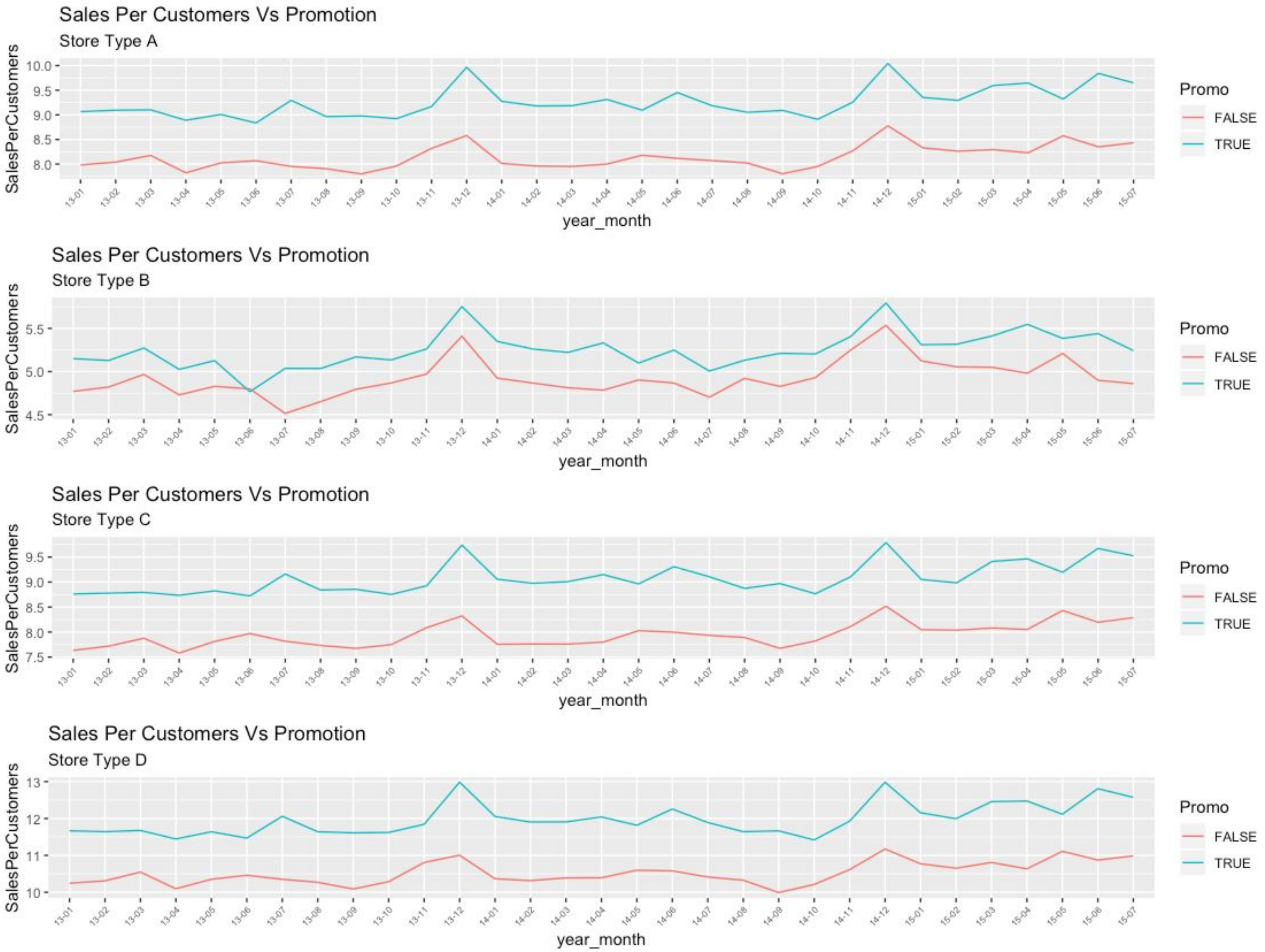
Cleaning Data

Based on the above findings, we decided to drop all the rows that has “0” value on Open column, because those data will not help us with predicting sales. We also decided to use median to replace data with NA values, because median will be more accurate than average since the distribution of sales is skewed. We then replaced NA values in column Promo2SinceWeek and column Promo2SinceYear with zero, because zero means the stores did not have promotion during that day. We replaced NA values in column CompetitionOpenSinceMonth and CompetitionOpenSinceYear with 0, because zero means the stores did not have competitor around.

MSIS 2506 R Programming Project 2
Bingbing Pan, Jialin Liang, Shruti Deshpande, Jaskaran Kohli, Tam Anh Pham

Visualizing Findings

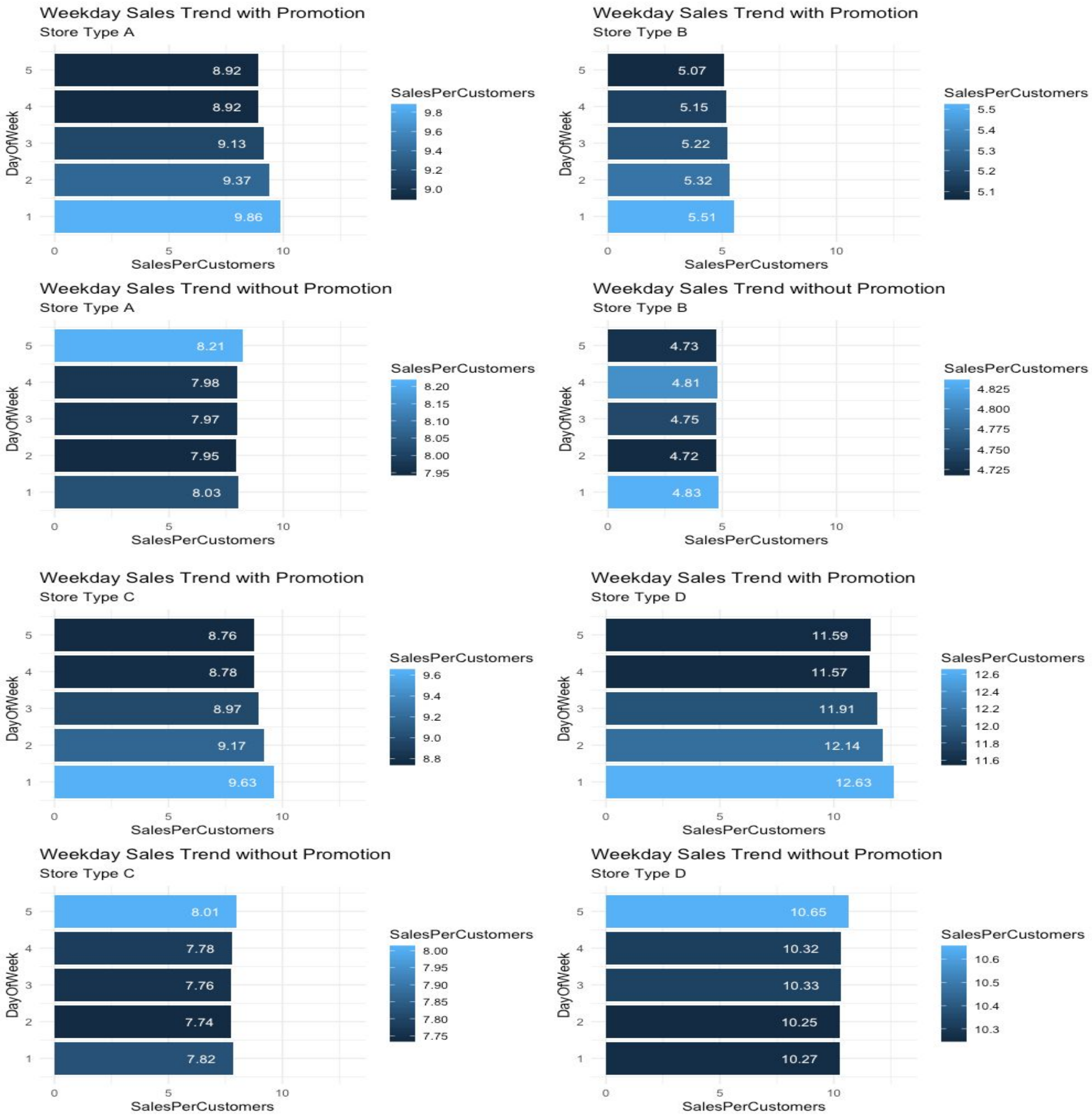
Graph 5 - Find out the relationship among store types and promo and sales.



We drew a graph of sales' yearly trends with time (year and month) on x-axis and sales per customer on y-axis. We have four sub-graphs for each type of store. The line charts above indicate the following findings:

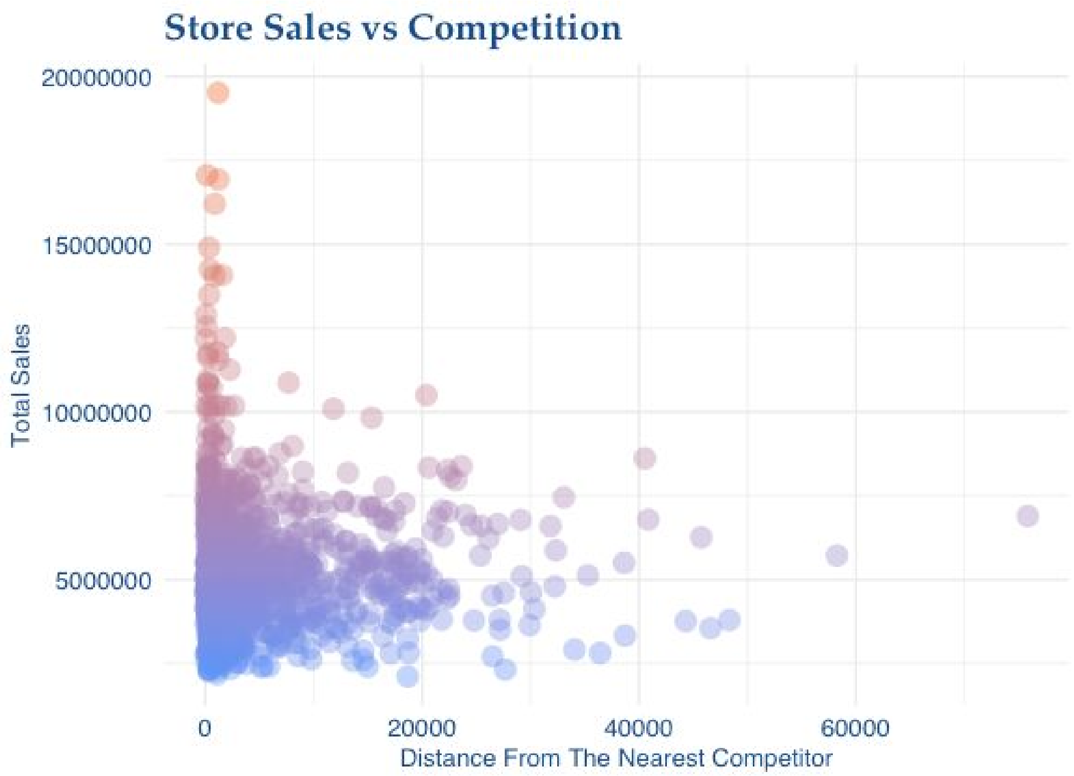
1. Sales per customer is always higher when there's promotion available for all types of stores. Promotion might be one of the most important variables for sales.
2. Sales per customer is relatively higher in store type D compared with that in other stores.
3. Sales per customer always started to grow in October and reached a peak in December. As the holiday season is coming in the last quarter of the year, customers purchase a lot accordingly.

Graph 6 - Find out how promotion affects weekday's sales



1. Monday sales is always the highest when there's promotion. After we looked into the store dataset, we found each Monday will start a new round of sales on different categories of products, which will attract customers to make some purchases.
2. Except for store type A, all the other stores has the highest sales on Friday when there's no promotion available. According to Business Insider (https://www.businessinsider.com/why-we-think-shopping-makes-us-happy-2014-4), shopping can actually make people happier. Since Friday is the last workday of the week, people tend to go shopping as a relaxing activity after 5 days of intense work.

Graph 7 - The relationship between store sales and competitor distance



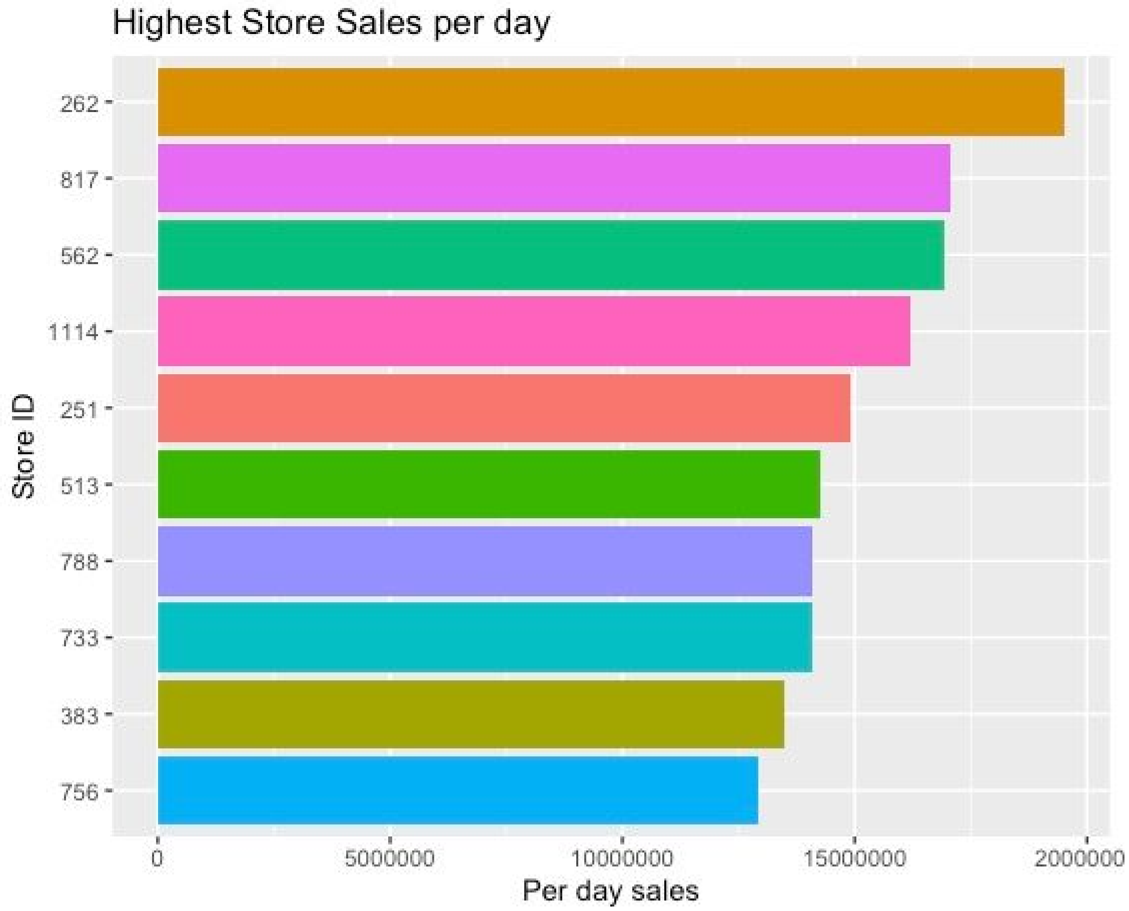
We created a scattergram to explore the relationship between store sales and competition, which is the distance from the nearest competitor. Based on the diagram and data summary, the majority of stores are located within 6875 meters (approximately 4.27 miles) from the closest competitor store. Interestingly, stores with the highest total sales recorded are located relatively close to their competitors, indicating that competitor distance is unlikely to be a factor that adversely affects sales.

For the purpose of examining the linear association between the two variables, we also constructed a regression model and conducted the significant test on RStudio, which did not yield statistically significant results.

	Estimate	Std. Error	T- value	Pr (>|t|)
(Intercept)	5.302e+06	7.152e+04	74.136	<2e-16
Slope	-6.439e+00	7.639e+00	-0.843	0.399

Based on the findings, we were unable to conclude on the existence of a correlation between store sales and competitor distance. ***Competitor distance, therefore, might not be a factor influencing the total sales of a store.***

Graph 8 - Top 10 Highest Sales Stores

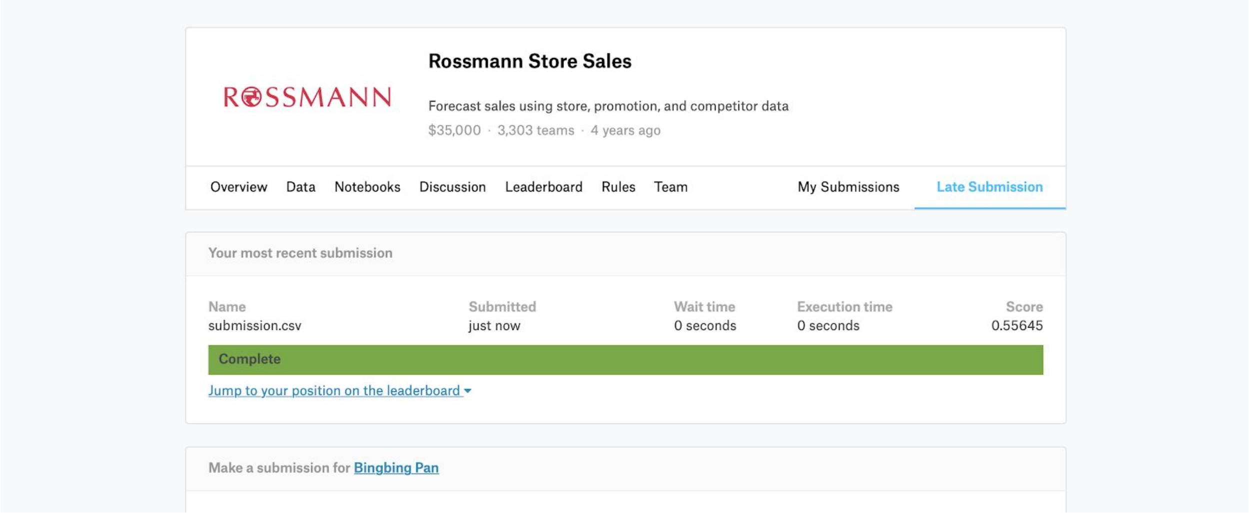


We created a bar plot to find the top 10 stores with the highest sales. One thing we found they are in common is that all 10 stores are continuously having promotion every other day, and most of the 10 stores open seven days a week. We think it’s a really valuable information for other stores which want to improve their sales.

MSIS 2506 R Programming Project 2
Bingbing Pan, Jialin Liang, Shruti Deshpande, Jaskaran Kohli, Tam Anh Pham

Training Data and Make Prediction

We used a simple linear regression model to train the data. We split the dataset into three equal portions and use ⅔ as training data ⅓ as test data. The number of customers is used as the independent variable and sales is used as the dependent variable. We utilized test set to generate sales prediction for the future 41088 days. We then submitted the prediction on Kaggle and got a score of 0.55645.



Rossmann Store Sales

Forecast sales using store, promotion, and competitor data
$35,000 · 3,303 teams · 4 years ago

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	just now	0 seconds	0 seconds	0.55645

Complete

Jump to your position on the leaderboard ▾

Make a submission for Bingbing Pan