Final Project Deliverable

IST 687

Antwarn Sanders

Ben Heindl

Chris Knight

Eryka Gosselin-Preston

Victoria Haley

For our final deliverable, our Team looked at a Customer Analytics dataset from Kaggle, a website that provides access to data sets (<https://www.kaggle.com/datasets/prachi13/customer-analytics>). This data set consists of 10,999 unique rows and 12 columns. Each row was associated with a customer ID and each column was associated with a different variable that was tracked for purchases. The full list of column names and definitions can be found in the Appendix to this deliverable.

After initially looking through the data set, we developed a list of 5 business questions that we would analyze. These business questions are ones that we felt would be important to know as the company this data came from. We would want to determine if there are patterns or correlations between variables and would want to ensure we analyzed this information to determine an answer or determine if more data might need to be analyzed. The list of questions included:

* Are repeat customers likely to purchase more expensive products? - Chris
* Which method of shipment is most likely to reach customers on time? - Victoria
* Do customers tend to leave better reviews on products that have a discount? - Eryka
* How do customers who spend more than the median cost for products rate their experience? -Antwarn
* Do customers with prior purchases receive more discount offers? - Ben

We divided this list of questions between us, and each took our own approach to determining the answers. The following deliverable lists the questions along with a summarized answer of our findings. This is then followed by the methodology that each team member took in solving their assigned questions and any other pertinent information such as the lines of codes and the graphs or diagrams that were created.

As you will see, there were some questions that we were able to technically answer, but no usable information was provided. This might mean there was no correlation or pattern determined to exist. We kept these findings since this information would be important to note for our operations, but also noted ways to possibly re-analyze this information such as using more data or a larger data set for future attempts.

**Business Question 1**

**Q: Are repeat customers likely to purchase more expensive products?**

**A: There is no evidence to show that repeat customers are likely to purchase more expensive products. However, this is caveated with the fact that the sample data is limited to the most recent purchase only. If the sample data were expanded to include all purchases from possibly the last 3-6 months, the information would be more accurate in showing a correlation or relationship between the two variables.**

*Methodology:*

First, I imported the data set into R Studio and saved to a data frame called “ProjectData”

I then looked to find the min and max for both the “Prior\_pruchases” and “Cost\_of\_the\_Products” variables to get the range of both.

I next created a new data frame that only included the “Prior\_pruchases” and “Cost\_of\_the\_Products” variables since these were the ones important to my analysis, and I named this data frame “PriorMixCost”.

I then ran the str() command to ensure that all the data in these two variables were in integer form.

I then made a histogram of both variables to see the frequencies and see if there was any patter (ie more frequency of higher cost of products and more frequency of prior purchases).

After this I made a basic scatter plot looking at the data with the “Cost\_of\_the\_Product” on the x-axis and “Prior\_purchases” on the y-axis. This showed no discernable pattern or correlation.

I then used a correlation formula to see if R could identify any correlation between the values. For this I installed the “ggpubr” package and ran the “cor()” function on the two variables using the “Pearson” correlation method. This returned a correlation coefficient of .12, which means there is likely no noticeable correlation.

I then ran the cor.test() function and correlation coefficient was the same. This function also shows the p-value which shows the significance level of any correlation, and the level was <2.2e-16 which means there is not a strong significance. This was already assumed due to the low correlation coefficient.

Lastly, just to check again, I ran the variables into a scatterplot that also used the “Pearson” correlation method and added a regression line. This scatter plot, like the previous, came back with no discernable relationship between the two variables.

*Conclusion:*

While it is possible there might be a correlation between customers with more prior purchases buying more expensive products, the data in this data set did not confirm this. It showed no noticeable correlation when graphed or when ran through different formulas. If there was more data on prior purchase costs, it might be able to allow for a more accurate measure of any possible correlations. This data set only shows the product cost for one item (the most recent) for each customer. It might be a better measure to have the cost of all purchases over the past x number of months since that would provide a better average of how much they are spending as they increase the number of prior purchases.

*R Code:*

# Importing Data Set from Computer

ProjectData <- read.csv("E:\\School\\Syracuse\\IST687\\Project\\train.csv")

# Checking to see minimum and maximum prior purchases

min(ProjectData$Prior\_purchases)

max(ProjectData$Prior\_purchases)

# No customer has made zero prior purchases. The minimum is 2 prior and maximum is 10 prior

# What are the minmum and maximum costs for current orders?

min(ProjectData$Cost\_of\_the\_Product)

max(ProjectData$Cost\_of\_the\_Product)

# Minimum cost is $96 and maxim um cost is $310

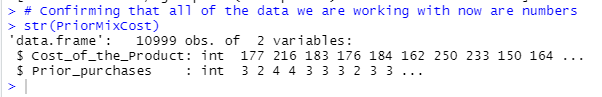
# Limit the Data Set to only show the Cost of the Product and the Prior Purchase values

PriorMixCost <- ProjectData[ , c("Cost\_of\_the\_Product", "Prior\_purchases")]

PriorMixCost

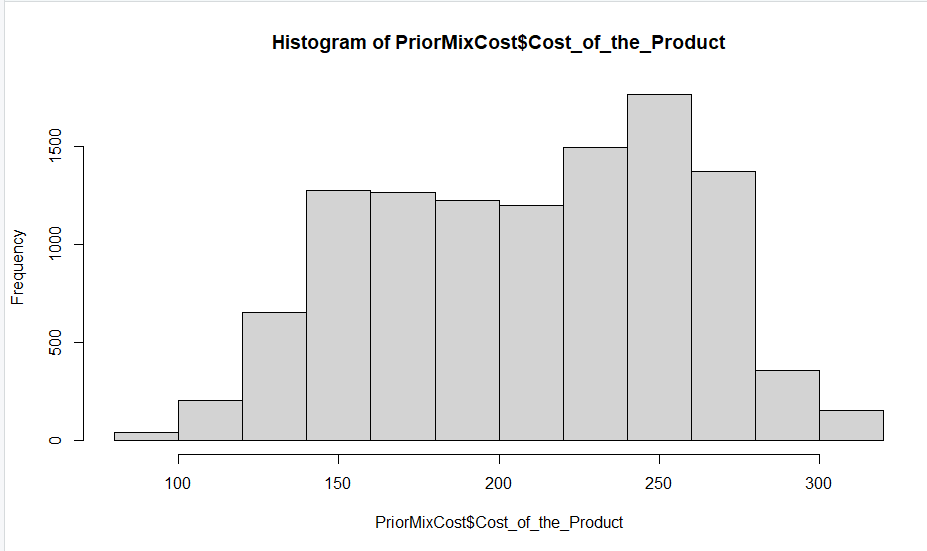
# Confirming that all of the data we are working with now are numbers

str(PriorMixCost)

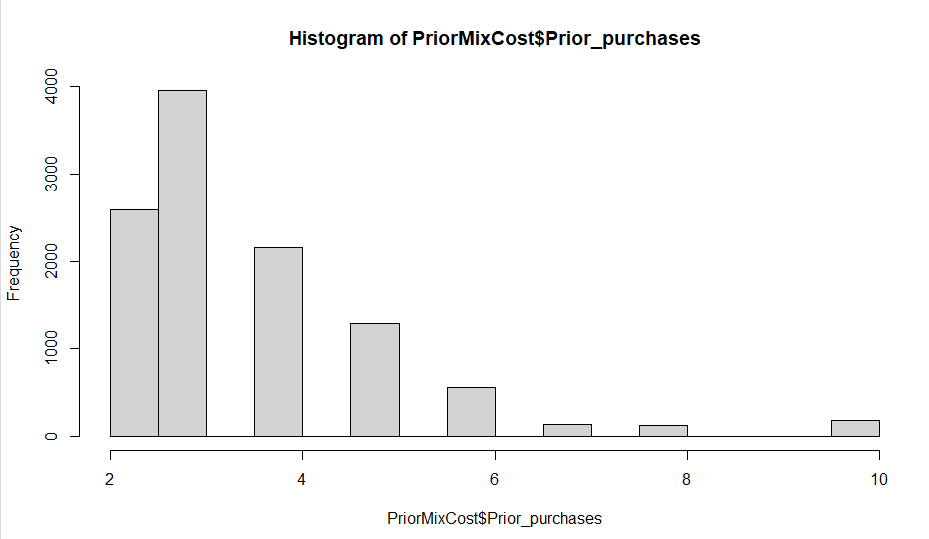


# Making a histogram of both the Cost and Prior purchase values to see how often each number comes up

hist(PriorMixCost$Cost\_of\_the\_Product)



hist(PriorMixCost$Prior\_purchases)

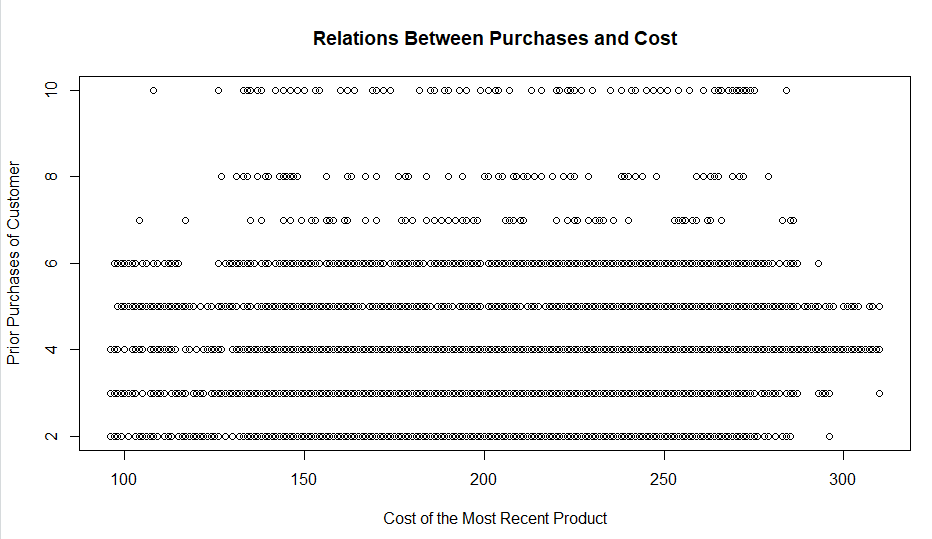


# Making a scatter plot to look for a pattern

attach(PriorMixCost)

plot(Cost\_of\_the\_Product, Prior\_purchases, main="Relations Between Purchases and Cost",

xlab="Cost of the Most Recent Product", ylab="Prior Purchases of Customer")



# No patterns seen in this scatter plot. Going to use correlation functions to look for any correlation between the variables.

# Using Correlation tools

install.packages("ggpubr")

library(ggpubr)

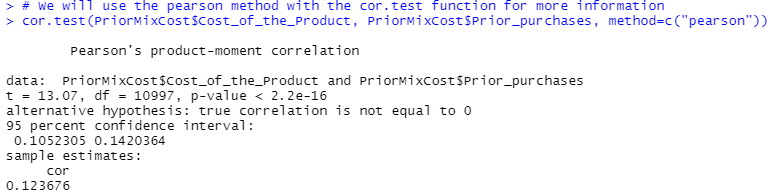
# Using the pearson correlation method to see if there is a correlation between prior purchases and product cost

cor(PriorMixCost$Cost\_of\_the\_Product, PriorMixCost$Prior\_purchases, method=c("pearson"))

# The correlation coefficient is 0.12 This means there is likely no correlation between previous purchases and cost of products in the current purchase.

# We will use the pearson method with the cor.test function for more information

cor.test(PriorMixCost$Cost\_of\_the\_Product, PriorMixCost$Prior\_purchases, method=c("pearson"))



# With this function we see the same correlation coefficient. We also see that the p-value, which shows the significance level, is < 2.2e-16 which shows there is not a strong significance.

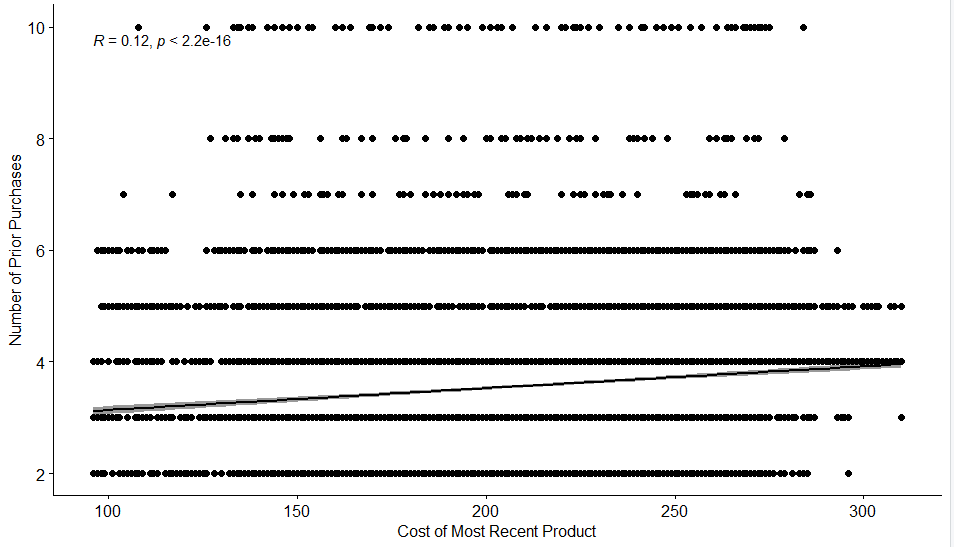
# Lastly, we will see if there is a correlation by using a scatter plot

ggscatter(PriorMixCost, x = "Cost\_of\_the\_Product", y = "Prior\_purchases",

add = "reg.line", conf.int = TRUE,

cor.coef = TRUE, cor.method = "pearson",

xlab = "Cost of Most Recent Product", ylab = "Number of Prior Purchases")



#Like the previous scatter plot and the correlation formulas, there seems to be no correlation shown in this graph between the two variables.

**Data Science Project Q2**

**Q: Which shipping method is most likely to reach customers on time?**

**A: Based on the data, customers are more likely to receive their order on time if the mode of shipment is by Ship.**

* **Ship:** 3003/4436
* Flight: 708/4436
* Road: 725/4436

*Methodology:*

First, I stored the e-commerce data that our group chose to work on into R into a data frame that I called “Shipping” from the Train.csv file that we found on <https://www.kaggle.com/datasets/prachi13/customer-analytics>

Then, after examining the data, I determined that I needed to separate the products that were delivered on time (0) from those that weren’t (1). After that, I was able to simply use the table() function that was mentioned in chapter 11 (pg 243) to count each instance of individual modes of shipment in the Mode\_of\_Shipment column in the OnTime data frame. From there, I was able to see at a glance that Ship had the most occurrences, meaning that it was the shipment method that was most likely to be on time, but I ran additional codes (which.max(), max(), which.min()) to supplement my findings.

To help visualize these findings, I created a new dataframe called chart in order to create a bar chart that makes it even easier to see that Ship is the fastest mode of shipment. This new dataframe was created by making a dataframe out of the results of the table() code. I had to clean up the columns so that they labelled the column for mode of shipment and how many times that mode occurred in the on-time data. From there, I used ggplot2 to help create my chart.

*R Code:*

#To create new data frame with only on time shipments from original Shipping data frame:

OnTime <- Shipping[Shipping$Reached.on.Time\_Y.N==0, ]

#To summarize the mode of shipment:

table(OnTime$Mode\_of\_Shipment)

#to determine the max of the counts from above:

which.max(table(OnTime$Mode\_of\_Shipment))

#New dataframe for bar chart:

chart <- data.frame(table(OnTime$Mode\_of\_Shipment))

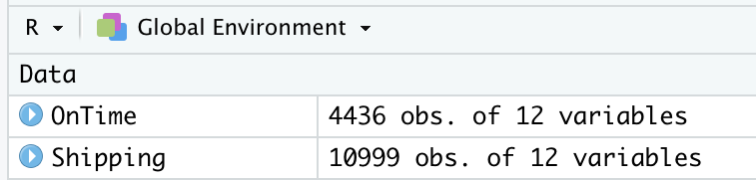
cnames <- c("Mode\_of\_Shipment", "Occurrences")

colnames(chart) <- cnames

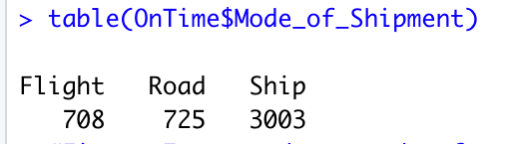
#Bar chart:

ggplot(data=chart, aes(x=Mode\_of\_Shipment, y=Occurrences)) + geom\_bar(stat="identity", aes(fill=Mode\_of\_Shipment))

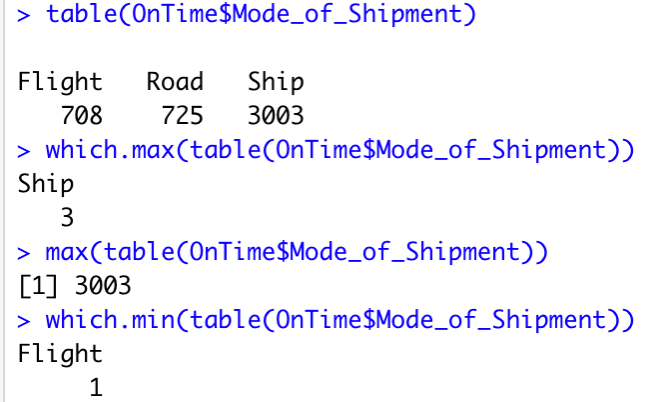
Screenshots:



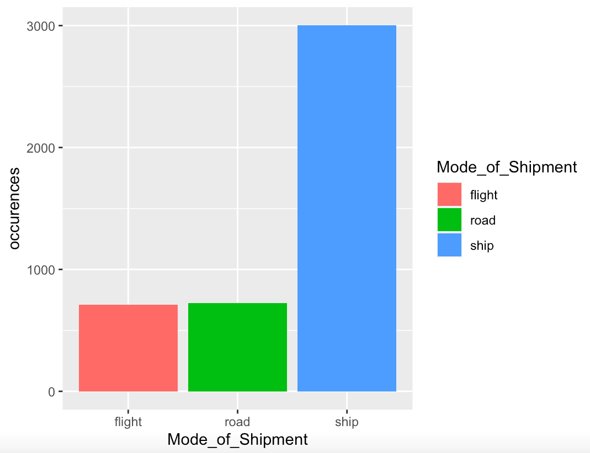
After creating a new data frame for products that were on time, we can see that only 4,436 of the 10,999 customers (columns 1 = ID) had products that were delivered on time



By using the table() code, we can see that 708/4436 of the on time shipments were by Flight, 725/4436 were by Road, and 3003/4436 were by Ship.



Here, we can see that which.max() determined that Ship has the maximum # of occurrences from the table() code. For further testing, I also ran the max() of the table() code, as well as the which.min().



Bar chart of results pictured above.

**Question #3:**

**Q: Do customers that have been offered discounts tend to leave better ratings?**

**A: There is zero correlation between customer ratings and the discounts offered. Through further investigation, the data shows that there are slightly negative correlations between discounts offered when paired against prior purchases (-0.08), customer care calls (-0.13), and the cost of the product (-0.14). As for customer ratings, there was virtually zero correlation amongst those same variables.** **All the other variable pairings have positive correlations that were statistically significant. The highest correlation was found to be between customer care calls and the cost of the product (0.32)**

*Methodology:*

First, I loaded the dataset and saved it to a data frame called shipData. After inspecting the data frame, I decided to focus on the variables outlined in the business question; Customer\_rating and Discount\_offered.

Then, I created a linear between both pairs to see whether the paired variables had a positive or negative relationship. When analyzing the summary, the pairing resulted in no correlation with a coefficient of -0.0002 and p-value of 0.743. I then plotted the model to visually confirm that there was in fact no correlation, which held true.

Since there was no correlation, I wanted to investigate to see if either the ratings or discounts offered correlated with other variables in the dataset. I created a new data frame called shipdf with the previous variables and additional variables; Customer\_care\_calls, Cost\_of\_the\_Product, and Prior\_purchases.

From there, I created a correlation coefficient matrix using the Pearson method followed by another correlation matrix for p-values. To visualize the outputs, I created a correlogram plotting the coefficients of the variables and then created a heatmap to confirm the results.

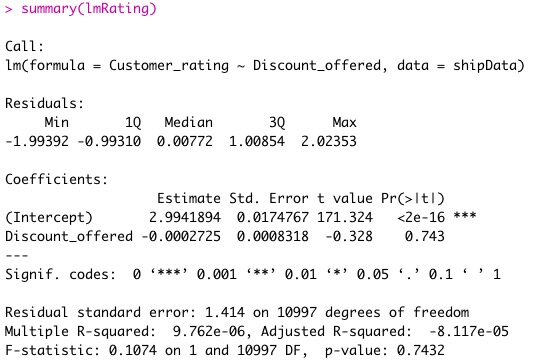
*R Code:*

## Question: Do customers tend to leave better ratings on products that have a discount?

# Generate a linear model to check for correlation

lmRating = lm(Customer\_rating~Discount\_offered, data = shipData)

summary(lmRating)

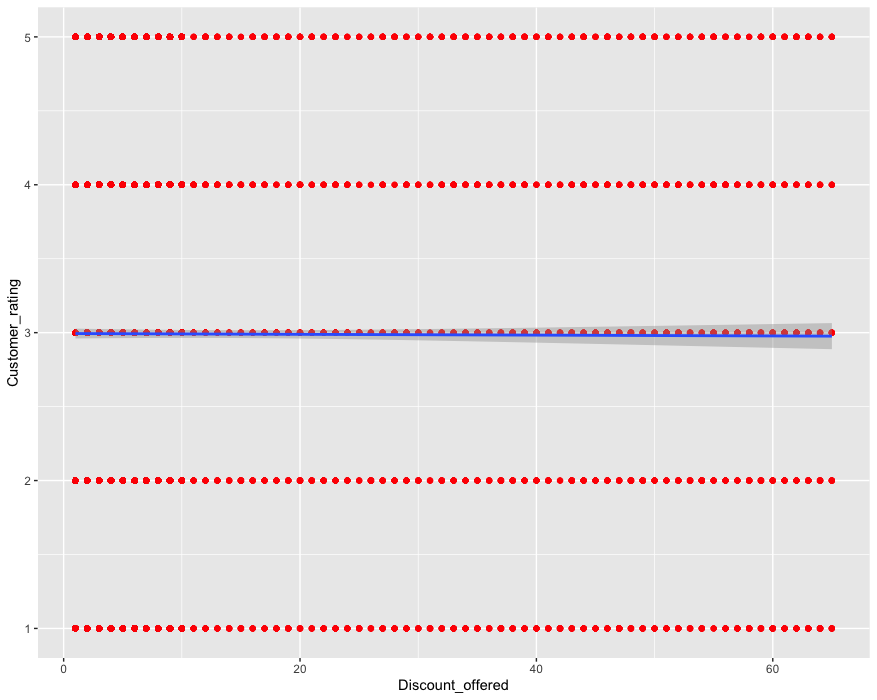


#Plot model

ggplot(shipData, aes(x = Discount\_offered, y = Customer\_rating)) +

geom\_point(colour = "blue") +

stat\_smooth(method = "lm")

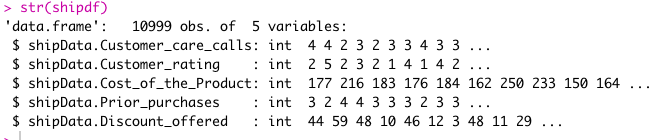


##Further investigation to see if customer ratings and discounts offerred correlates with other variables

# Create Data frame with additional variables

shipdf <- data.frame(shipData$Customer\_care\_calls, shipData$Customer\_rating, shipData$Cost\_of\_the\_Product, shipData$Prior\_purchases, shipData$Discount\_offered)

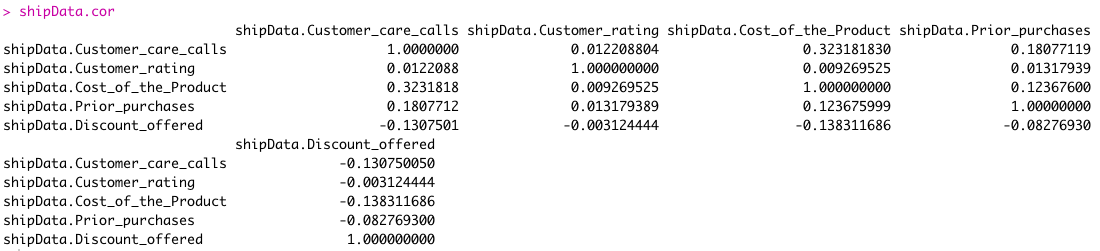
str(shipdf)



## Run correlation coefficient matrix between pairs of variables - Pearson method

shipData.cor = cor(shipdf)

ShipData.cor



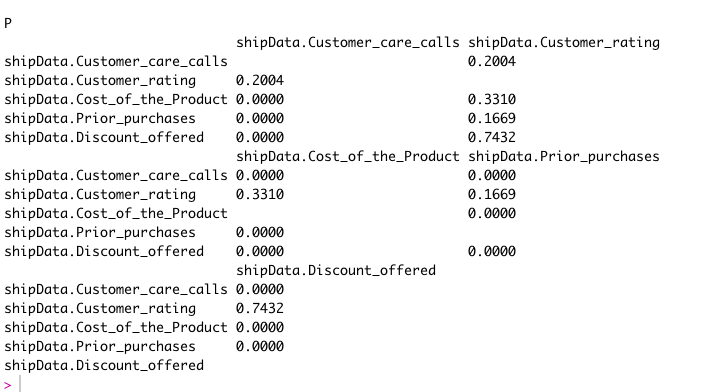
# Run correlation matrix of p-values (significance levels)

install.packages("Hmisc")

library("Hmisc")

shipData.rcorr = rcorr(as.matrix(shipdf))

ShipData.rcorr



# Store correlation coefficients and p-values into objects of class type matrix

shipdf.coeff = shipData.rcorr$r

shipdf.p = shipData.rcorr$P

# Visualize correlation matrix

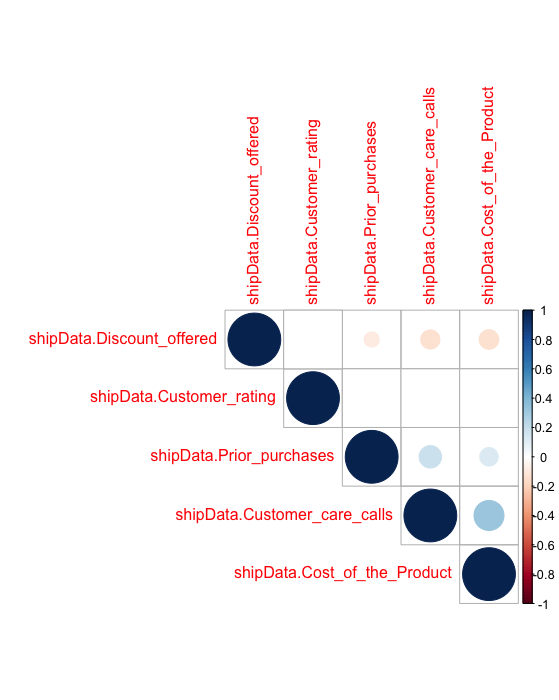
install.packages("corrplot")

library(corrplot)

# Create a correlogram with insignificant correlations leaved blank

corrplot(shipData.rcorr$r, type="upper", order="hclust",

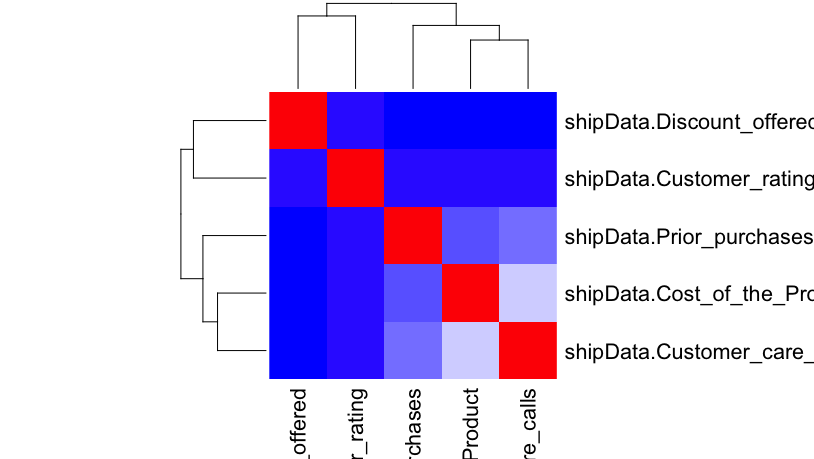
p.mat = shipData.rcorr$P, sig.level = 0.01, insig = "blank")



# Create Heatmap\*

palette = colorRampPalette(c("blue", "white", "red"))(20)

heatmap(x = shipData.cor, col = palette, symm = TRUE)



*\*Unable to get full image*

**Business Question 4**

**Q: How do customers who spend more than the median cost for products rate their experience?**

**A: There was no clear differentiation between customers who spend more than the median cost and their customer ratings.**

*Methodology:*

**Step One:** Downloaded the data set from “Train.Csv” file and uploaded it into R Studio. Then I stored the data set in data frame called “CAdata”

**Step Two:** Found median for “Cost\_of\_the\_Product” column in data frame “CAdata” using median function in R Studio

**Step Three:** Created new data frame that shows customers who spent more than the median cost and named it “CmoreM”

**Step Four:** Created a histogram and boxplot to show the customer ratings for the new data frame “CmoreM’

**Step Five:** Analyzed the data of the customers who spend more than the median cost and the experience ratings of those customers.

I first took the average of experience ratings from the customers who spend more than the median cost and compared it to the average of experience ratings from all customers.

*Conclusion:*

In the histogram, we can see that the customer ratings from the scale of 1 to 5 are nearly identical despite the amount the customers spent on products.

In the boxplot, we see that the box is symmetrical on both sides meaning that the median is equal difference from the maximum and minimum value. In terms of the data this means that the median cost of products and their customer ratings is a normal distribution. This also proves that the data collected is not skewed negatively nor positively.

In addition, I calculated the average/median of customer ratings in data frame “CAdata” using median function in R and compared it to the average/median customer ratings of the new data frame created which included customers who spend more than the median cost. Surprisingly, the results were the same which meant both of the average customer ratings were 3.

These findings help conclude that customers who spend more than the median cost have little difference in their customer ratings than those who spend less than the median cost.

*R Codes:*

**Step One:**

Install tidyverse

Library(tidyverse)

CAdata <- read\_csv( “Train.csv”)

head(CAdata)

**Step Two:**

Median(CAdata$Cost\_of\_the\_Product)

**Step Three:**

CmoreM <- CAdata$Cost\_of\_the\_Product>214

Step Four:

ggplot(CmoreM)+geom\_boxplot(aes(x=Customer\_rating))

ggplot(CmoreM)+geom\_histogram(aes(x=Customer\_rating))

Step Five:

median(CAdata$Customer\_rating)

median(CmoreM$Customer\_rating)

**Question #5:**

**Q: Do customers with prior purchases receive more discount offers?**

*Methodology:*

I started by storing the dataset in a dataset called ShippingData and used the readr library to read the csv file. After using the view function to examine the data, I determined that I needed to use both the Prior\_purchases and Discount\_offered variables. I used the str() command to be sure the data was in the correct format.

I made a histogram of discounts offered and prior purchases to determine if any noticeable patterns were presented. There was a pattern on each histogram. We can see from the discounts offered histogram that an overwhelming majority of number of discounts offered and 10 or less. In the prior purchased histogram, the number of prior purchases decreases as we move into the higher purchase numbers, and we see that most prior purchases are represented by the 2-4 range.

The next step was to run a correlation analysis using cor() to determine if there was a relationship between the number of discounts offered and prior purchases. The result from the correlation analysis was a coefficient of -0.0828. With the coefficient being so close to zero, this suggests there is a very weak relationship between discounts offered and prior purchases. I wanted to test for statistical significance and learned that the cor.test() function can provide that data. I ran the cor.test() function and the p-value was <2.2e-16 which suggests there is not a strong statistical significance.

I wanted to create another scatterplot that used the “Pearson” correlation method but added a regression line for visual assistance. I expected similar results to the first scatterplot and that is what I got.

*Conclusion:*

I found no evidence to support a relationship between prior purchases and discounts offered. The correlation analysis resulted in a coefficient very close to zero which means there is a very weak relationship between discounts offered and prior purchases.

*R Code:*

