e-Commerce Exploratory Data Analysis (EDA)

Dorcas Mbaeri

March 3, 2018

### Structure of data

This is the introduction to my senior project at the University of Houston-Downtown. We will begin by exploring the dataset containing Zappos.com’s customer transactions.

But first let’s load the data.

mydata <- read.csv("C:/Users/Dorcas/OneDrive - University of Houston Downtown/SeniorProject/Senior Project/Analytics\_Challenge\_Data.csv", header = TRUE, row.names = NULL, na = "NA")  
  
my\_data <- data.table(mydata)

Now, let’s take look at the structure of the data and missing values. From the below, we see that there are 8259 blank values in *new\_customer* as well as 2469 blank values for *conversion rate*, *bounce\_rate*, *add\_to\_cart\_rate*. Note that the last three columns are calculated using *orders*, *bounces*, *add\_to\_cart*, and *visits*. Conversion rate is calculted by diving the number of orders by visits; bounce rate is calculated by dividing bounces by visits; and, add to cart rate is dividing add to cart by visits. So, if there is a division by 0 (meaning with 0 visits), this woud be *null*.

summary(mydata)

## day site new\_customer platform   
## 12/19/2013 0:00: 86 Acme :7392 Min. :0.000 iOS :3435   
## 11/29/2013 0:00: 85 Botly : 804 1st Qu.:0.000 Android:3172   
## 12/11/2013 0:00: 85 Pinnacle:5725 Median :0.000 Windows:2399   
## 12/7/2013 0:00 : 85 Sortly :5532 Mean :0.448 MacOSX :2054   
## 12/2/2013 0:00 : 84 Tabular : 804 3rd Qu.:1.000 Linux :2036   
## 12/5/2013 0:00 : 84 Widgetry: 804 Max. :1.000 Unknown:1641   
## (Other) :20552 NA's :8259 (Other):6324   
## visits distinct\_sessions orders gross\_sales   
## Min. : 0 Min. : 0 Min. : 0.00 Min. : 1   
## 1st Qu.: 3 1st Qu.: 2 1st Qu.: 0.00 1st Qu.: 79   
## Median : 24 Median : 19 Median : 0.00 Median : 851   
## Mean : 1935 Mean : 1515 Mean : 62.38 Mean : 16473   
## 3rd Qu.: 360 3rd Qu.: 274 3rd Qu.: 7.00 3rd Qu.: 3145   
## Max. :136057 Max. :107104 Max. :4916.00 Max. :707642   
## NA's :9576   
## bounces add\_to\_cart product\_page\_views search\_page\_views  
## Min. : 0.0 Min. : 0.0 Min. : 0 Min. : 0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 3 1st Qu.: 4   
## Median : 5.0 Median : 4.0 Median : 53 Median : 82   
## Mean : 743.3 Mean : 166.3 Mean : 4358 Mean : 8584   
## 3rd Qu.: 97.0 3rd Qu.: 43.0 3rd Qu.: 708 3rd Qu.: 1229   
## Max. :54512.0 Max. :7924.0 Max. :187601 Max. :506629   
##   
## conversion\_rate bounce\_rate add\_to\_cart\_rate  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.1429 1st Qu.:0.0223   
## Median :0.0000 Median :0.3118 Median :0.1667   
## Mean :0.2201 Mean :0.3396 Mean :0.2935   
## 3rd Qu.:0.3571 3rd Qu.:0.5024 3rd Qu.:0.5000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :2469 NA's :2469 NA's :2469

dim(mydata)

## [1] 21061 15

str(mydata)

## 'data.frame': 21061 obs. of 15 variables:  
## $ day : Factor w/ 268 levels "1/1/2013 0:00",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ site : Factor w/ 6 levels "Acme","Botly",..: 1 1 4 1 2 1 4 4 1 1 ...  
## $ new\_customer : int 1 1 1 1 1 1 1 1 0 0 ...  
## $ platform : Factor w/ 15 levels "","Android","BlackBerry",..: 2 3 6 14 2 9 2 14 8 7 ...  
## $ visits : int 24 0 0 922 11 384 14 1 41 448 ...  
## $ distinct\_sessions : int 16 0 0 520 10 214 10 0 27 368 ...  
## $ orders : int 14 0 0 527 11 213 4 0 6 36 ...  
## $ gross\_sales : int 1287 13 98 60753 1090 28129 432 31 705 4637 ...  
## $ bounces : int 4 0 0 149 0 65 4 0 6 80 ...  
## $ add\_to\_cart : int 16 0 0 610 11 245 7 0 12 79 ...  
## $ product\_page\_views: int 104 1 0 3914 4 1783 33 2 130 722 ...  
## $ search\_page\_views : int 192 0 0 7367 19 3255 52 2 272 1073 ...  
## $ conversion\_rate : num 0.583 NA NA 0.572 1 ...  
## $ bounce\_rate : num 0.167 NA NA 0.162 0 ...  
## $ add\_to\_cart\_rate : num 0.667 NA NA 0.662 1 ...

head(mydata)

## day site new\_customer platform visits distinct\_sessions  
## 1 1/1/2013 0:00 Acme 1 Android 24 16  
## 2 1/1/2013 0:00 Acme 1 BlackBerry 0 0  
## 3 1/1/2013 0:00 Sortly 1 iPad 0 0  
## 4 1/1/2013 0:00 Acme 1 Windows 922 520  
## 5 1/1/2013 0:00 Botly 1 Android 11 10  
## 6 1/1/2013 0:00 Acme 1 Macintosh 384 214  
## orders gross\_sales bounces add\_to\_cart product\_page\_views  
## 1 14 1287 4 16 104  
## 2 0 13 0 0 1  
## 3 0 98 0 0 0  
## 4 527 60753 149 610 3914  
## 5 11 1090 0 11 4  
## 6 213 28129 65 245 1783  
## search\_page\_views conversion\_rate bounce\_rate add\_to\_cart\_rate  
## 1 192 0.5833333 0.1666667 0.6666667  
## 2 0 NA NA NA  
## 3 0 NA NA NA  
## 4 7367 0.5715835 0.1616052 0.6616052  
## 5 19 1.0000000 0.0000000 1.0000000  
## 6 3255 0.5546875 0.1692708 0.6380208

colSums(sapply(mydata, is.na))

## day site new\_customer   
## 0 0 8259   
## platform visits distinct\_sessions   
## 0 0 0   
## orders gross\_sales bounces   
## 0 9576 0   
## add\_to\_cart product\_page\_views search\_page\_views   
## 0 0 0   
## conversion\_rate bounce\_rate add\_to\_cart\_rate   
## 2469 2469 2469

We want to identify the number of missing values in each numeric column.

num\_var <- names(my\_data)[which(sapply(my\_data, is.numeric))]  
  
summary(my\_data[,.SD, .SDcols = num\_var])

## new\_customer visits distinct\_sessions orders   
## Min. :0.000 Min. : 0 Min. : 0 Min. : 0.00   
## 1st Qu.:0.000 1st Qu.: 3 1st Qu.: 2 1st Qu.: 0.00   
## Median :0.000 Median : 24 Median : 19 Median : 0.00   
## Mean :0.448 Mean : 1935 Mean : 1515 Mean : 62.38   
## 3rd Qu.:1.000 3rd Qu.: 360 3rd Qu.: 274 3rd Qu.: 7.00   
## Max. :1.000 Max. :136057 Max. :107104 Max. :4916.00   
## NA's :8259   
## gross\_sales bounces add\_to\_cart product\_page\_views  
## Min. : 1 Min. : 0.0 Min. : 0.0 Min. : 0   
## 1st Qu.: 79 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 3   
## Median : 851 Median : 5.0 Median : 4.0 Median : 53   
## Mean : 16473 Mean : 743.3 Mean : 166.3 Mean : 4358   
## 3rd Qu.: 3145 3rd Qu.: 97.0 3rd Qu.: 43.0 3rd Qu.: 708   
## Max. :707642 Max. :54512.0 Max. :7924.0 Max. :187601   
## NA's :9576   
## search\_page\_views conversion\_rate bounce\_rate add\_to\_cart\_rate  
## Min. : 0 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 4 1st Qu.:0.0000 1st Qu.:0.1429 1st Qu.:0.0223   
## Median : 82 Median :0.0000 Median :0.3118 Median :0.1667   
## Mean : 8584 Mean :0.2201 Mean :0.3396 Mean :0.2935   
## 3rd Qu.: 1229 3rd Qu.:0.3571 3rd Qu.:0.5024 3rd Qu.:0.5000   
## Max. :506629 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :2469 NA's :2469 NA's :2469

colSums(sapply(my\_data[,.SD, .SDcol = num\_var], is.na))

## new\_customer visits distinct\_sessions   
## 8259 0 0   
## orders gross\_sales bounces   
## 0 9576 0   
## add\_to\_cart product\_page\_views search\_page\_views   
## 0 0 0   
## conversion\_rate bounce\_rate add\_to\_cart\_rate   
## 2469 2469 2469

The summary statistics helps to see the distribution of the numerical variables. For example, the mean number of *visits* in the data is **1935**, the median is **24**, and the maximum value is **136057**. In this scenario, we can conclude that the spread of this dimension is ***skewed right*** or positively skewed (with the mean to the right of the median). We will test some of these columns using a boxplot to visualize their spread.

Let’s also do a comparison of missing values in the categorical columns.

char\_var <- names(my\_data)[which(sapply(my\_data, is.factor))]  
  
summary(my\_data[,.SD, .SDcols = char\_var])

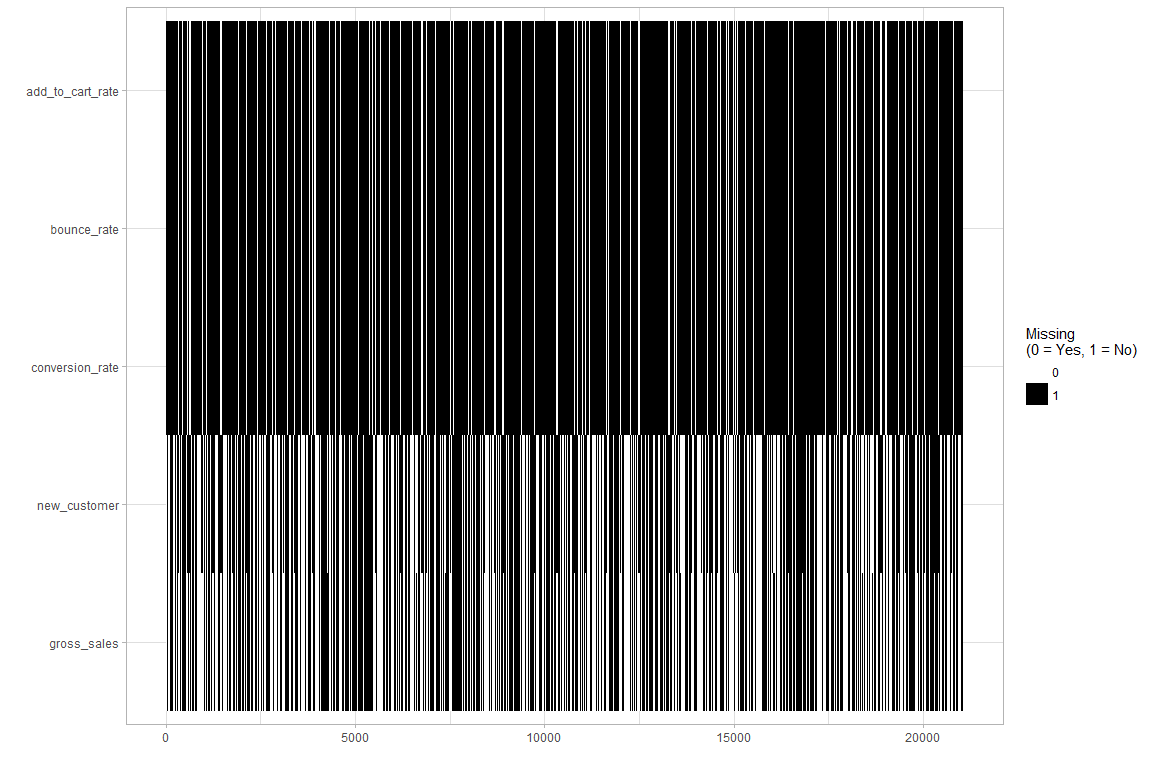
## day site platform   
## 12/19/2013 0:00: 86 Acme :7392 iOS :3435   
## 11/29/2013 0:00: 85 Botly : 804 Android:3172   
## 12/11/2013 0:00: 85 Pinnacle:5725 Windows:2399   
## 12/7/2013 0:00 : 85 Sortly :5532 MacOSX :2054   
## 12/2/2013 0:00 : 84 Tabular : 804 Linux :2036   
## 12/5/2013 0:00 : 84 Widgetry: 804 Unknown:1641   
## (Other) :20552 (Other):6324

colSums(sapply(my\_data[,.SD, .SDcols = char\_var], is.na))

## day site platform   
## 0 0 0

It’s always a lot easier when we can visualize any missing data representation. So we will create a function that will visually display the comparison of missing values to non-missing values in our data set. *(kaggle post by AiO)*

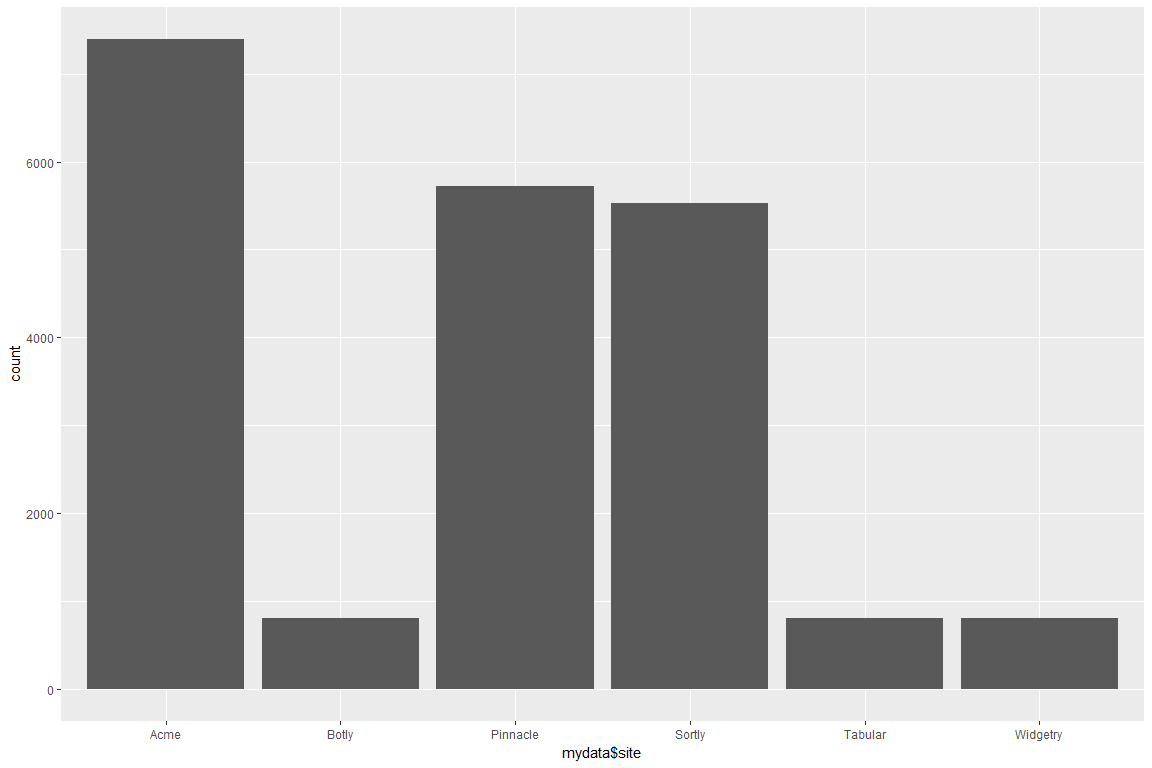
Missing\_Values <- function(input\_data) {  
 temp\_data <- as.data.frame(ifelse(is.na(input\_data), 0, 1))  
 temp\_data <- temp\_data[, order(colSums(temp\_data))]  
 data\_temp <- expand.grid(list(x= 1:nrow(temp\_data), y=colnames(temp\_data)))  
 data\_temp$m <- as.vector(as.matrix(temp\_data))  
 data\_temp <- data.frame(x = unlist(data\_temp$x), y = unlist(data\_temp$y), m = unlist(data\_temp$m))  
   
 ggplot(data\_temp) +  
 geom\_tile(aes(x=x, y=y, fill=factor(m))) +  
 scale\_fill\_manual(values=c("white", "black"), name = "Missing\n(0 = Yes, 1 = No)") +  
 theme\_light() +  
 ylab("") +  
 xlab("")  
}  
  
Missing\_Values(my\_data[, colSums(is.na(my\_data)) > 0, with = FALSE])



Here are visual distributions of the categorical variables of ***mydata*** (excludes missing values):

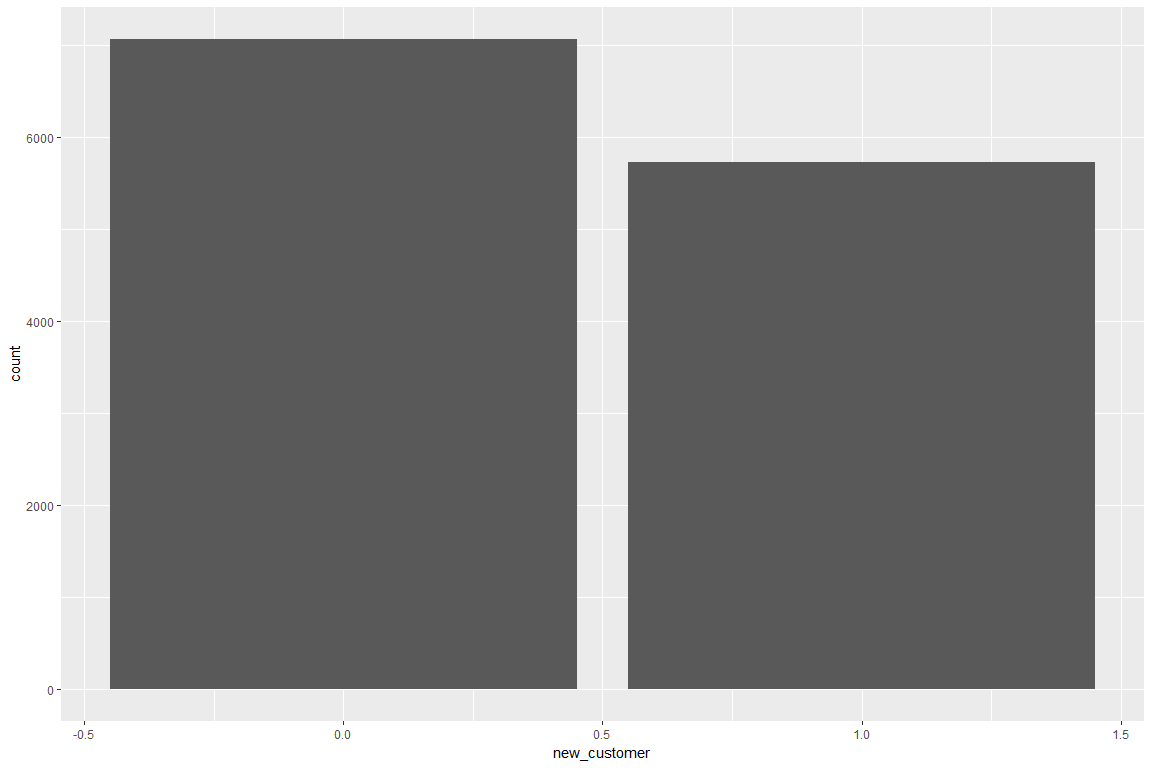
Distribution of *site*

ggplot(data = mydata) +  
 geom\_bar(mapping = aes(x = mydata$site), na.rm = TRUE)



*new\_customer* distribution

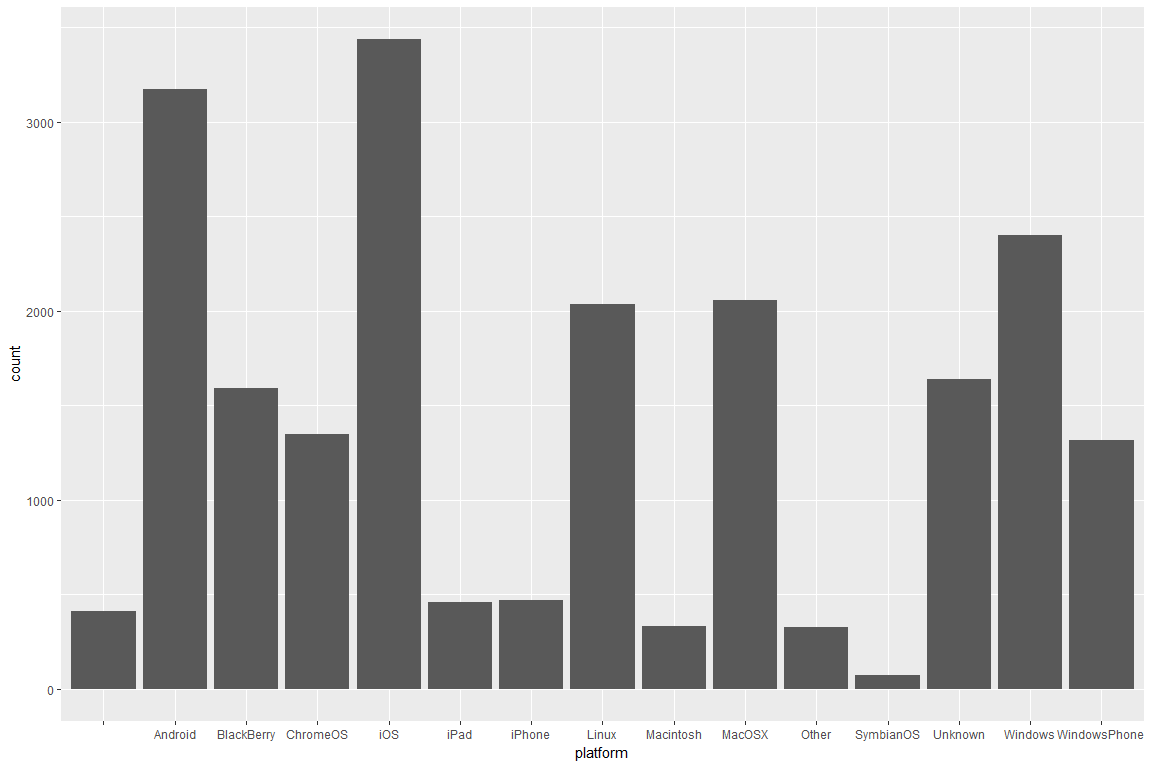
ggplot(data=mydata) +  
 geom\_bar(mapping = aes(x = new\_customer), na.rm = TRUE)



*platform* distribution

ggplot(data = mydata, mapping = aes(x = platform), na.ra = TRUE) +  
 geom\_histogram(stat = "count", position = position\_stack(reverse = TRUE), na.ra = TRUE)

## Warning: Ignoring unknown parameters: binwidth, bins, pad, na.ra



We can take a look at the distribution of *conversion\_rate*, *bounce\_rate* and *add\_to\_cart\_rate* by *new\_customer* as boxplots and frequency plots. From these plots, we see that Acme and Android and iOS are the most most used site and platforms to have items searched for as well as added to cart.

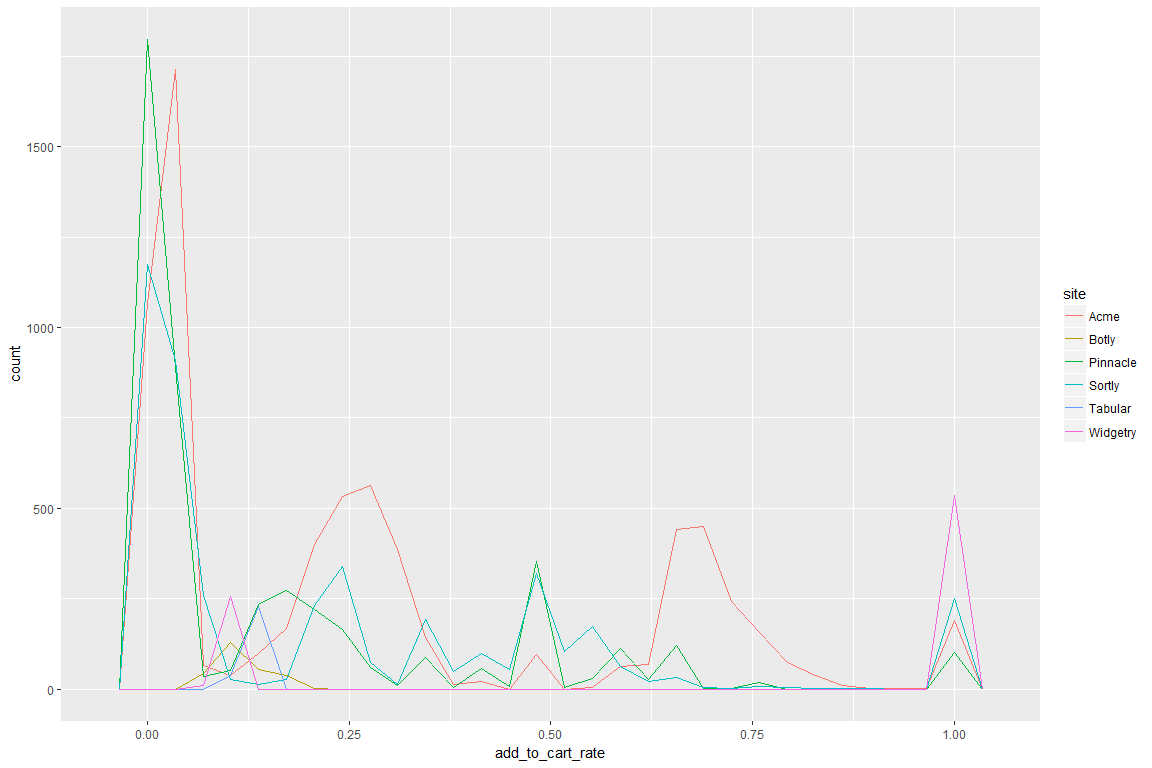
boxplot(mydata$add\_to\_cart\_rate ~ mydata$new\_customer, outline=FALSE, xlab = 'Rate of Items Added to Cart per Visit', ylab='Type of Customer', horizontal=TRUE)



ggplot(data=mydata, mapping = aes(x = add\_to\_cart\_rate)) +  
 geom\_freqpoly(mapping = aes(color = site))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

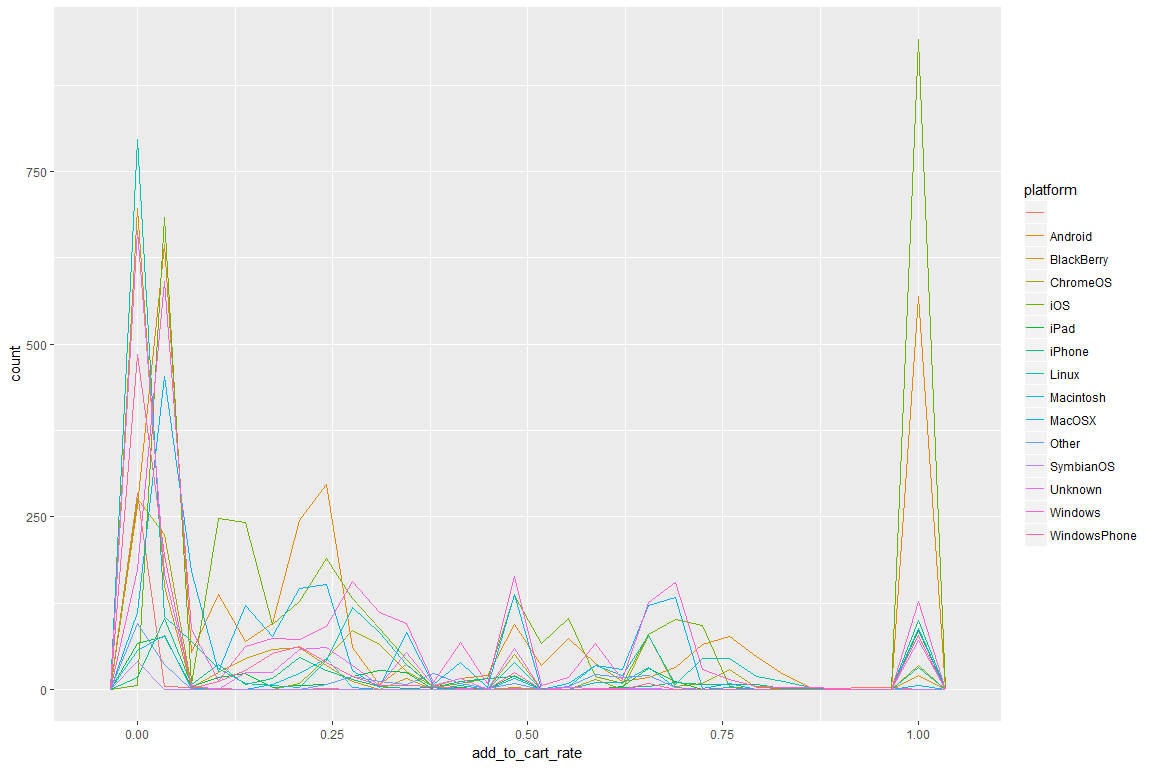
## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



ggplot(data=mydata, mapping = aes(x = add\_to\_cart\_rate)) +  
 geom\_freqpoly(mapping = aes(color = platform))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



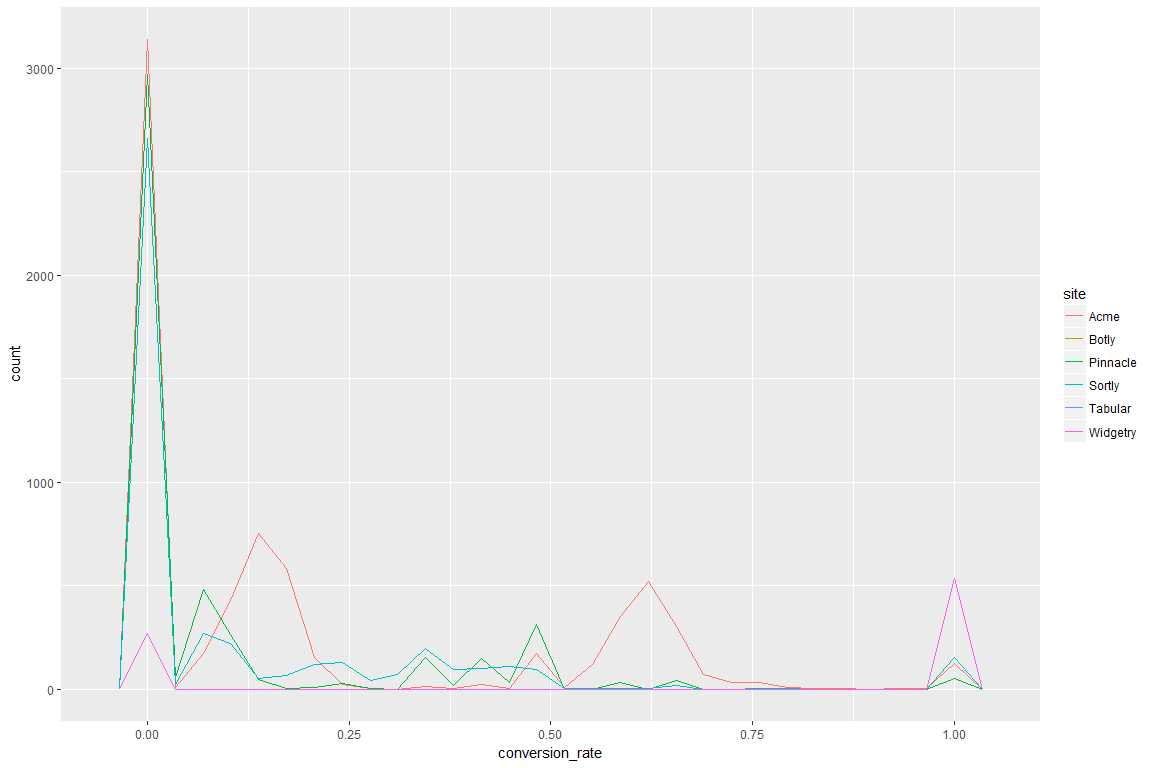
boxplot(mydata$conversion\_rate ~ mydata$new\_customer, outline=FALSE, xlab = 'Rate of Orders Per Visit', ylab='Type of Customer', horizontal=TRUE)



ggplot(data=mydata, mapping = aes(x = conversion\_rate)) +  
 geom\_freqpoly(mapping = aes(color = site))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

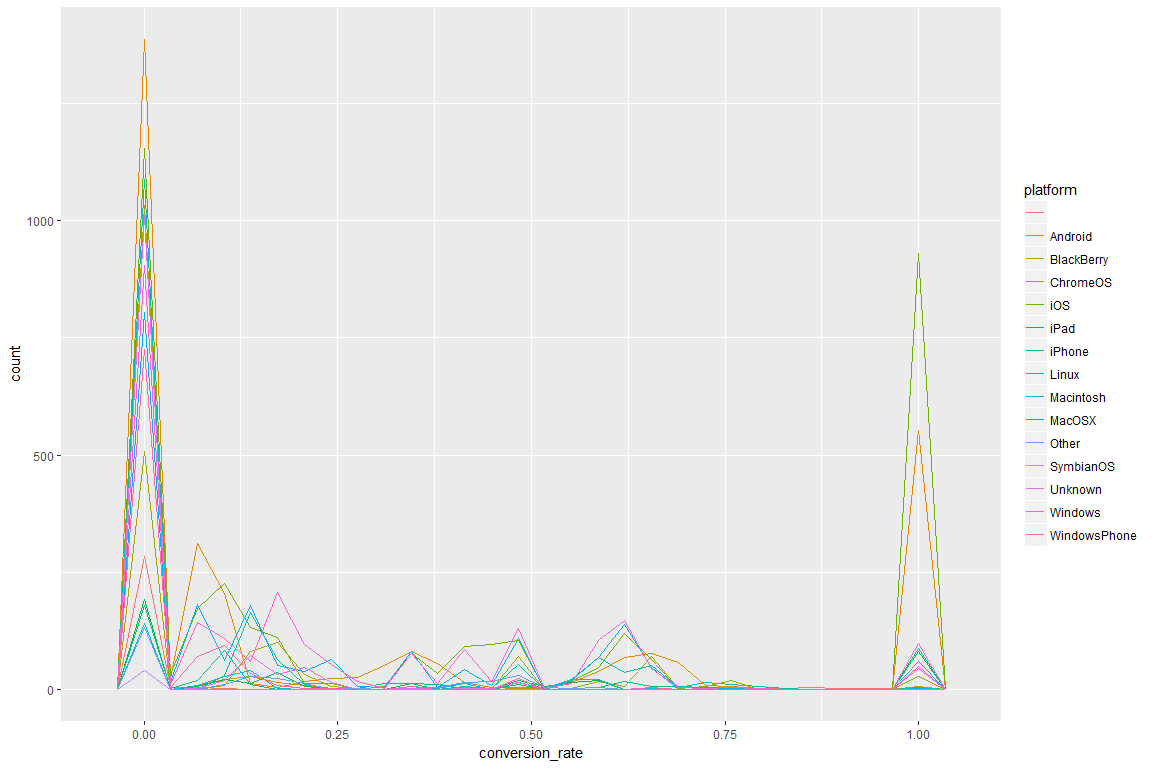
## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



ggplot(data=mydata, mapping = aes(x = conversion\_rate)) +  
 geom\_freqpoly(mapping = aes(color = platform))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



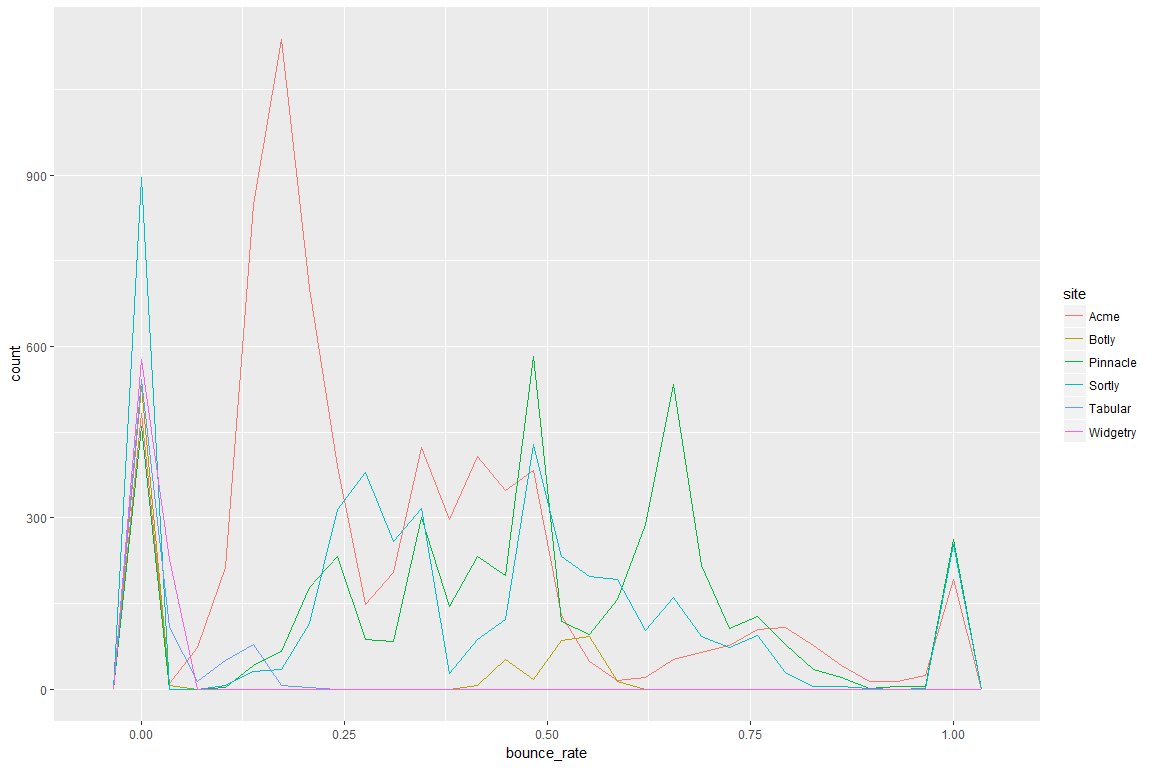
boxplot(mydata$bounce\_rate ~ mydata$new\_customer, outline=FALSE, xlab = 'Rate of Bounces Per Visit', ylab='Type of User', horizontal = TRUE)



ggplot(data=mydata, mapping = aes(x = bounce\_rate)) +  
 geom\_freqpoly(mapping = aes(color = site))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

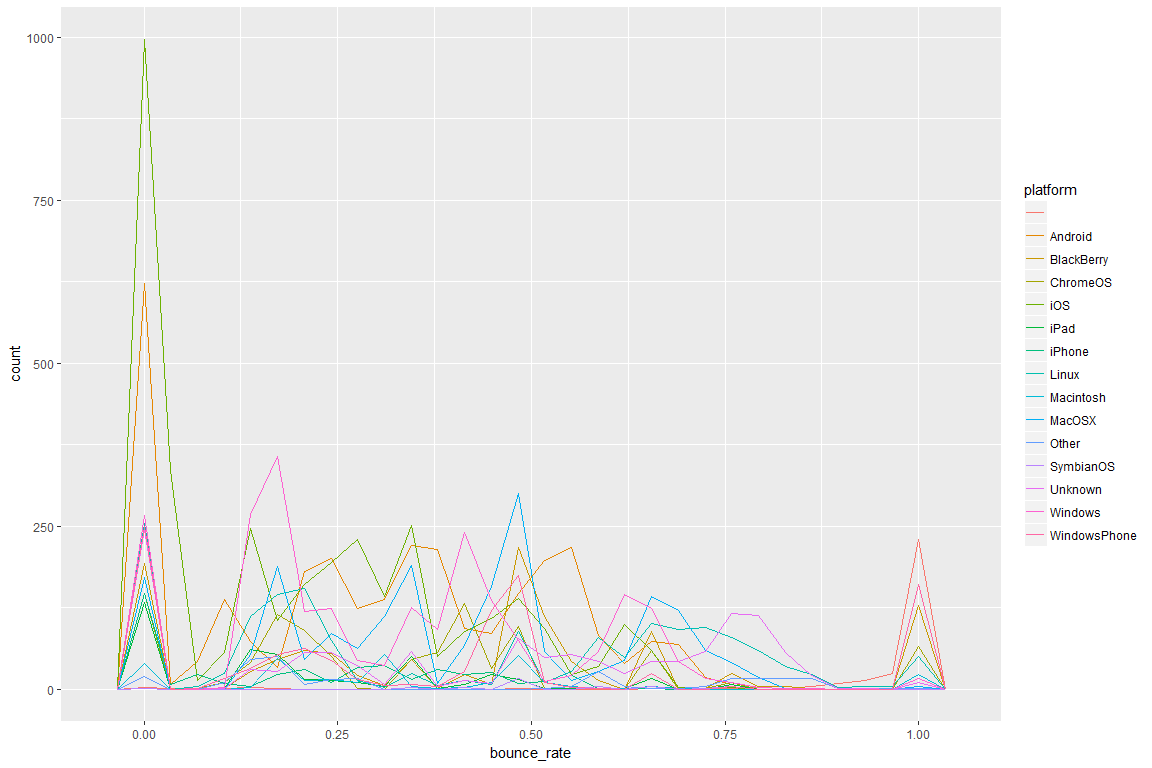
## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



ggplot(data=mydata, mapping = aes(x = bounce\_rate)) +  
 geom\_freqpoly(mapping = aes(color = platform))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2469 rows containing non-finite values (stat\_bin).



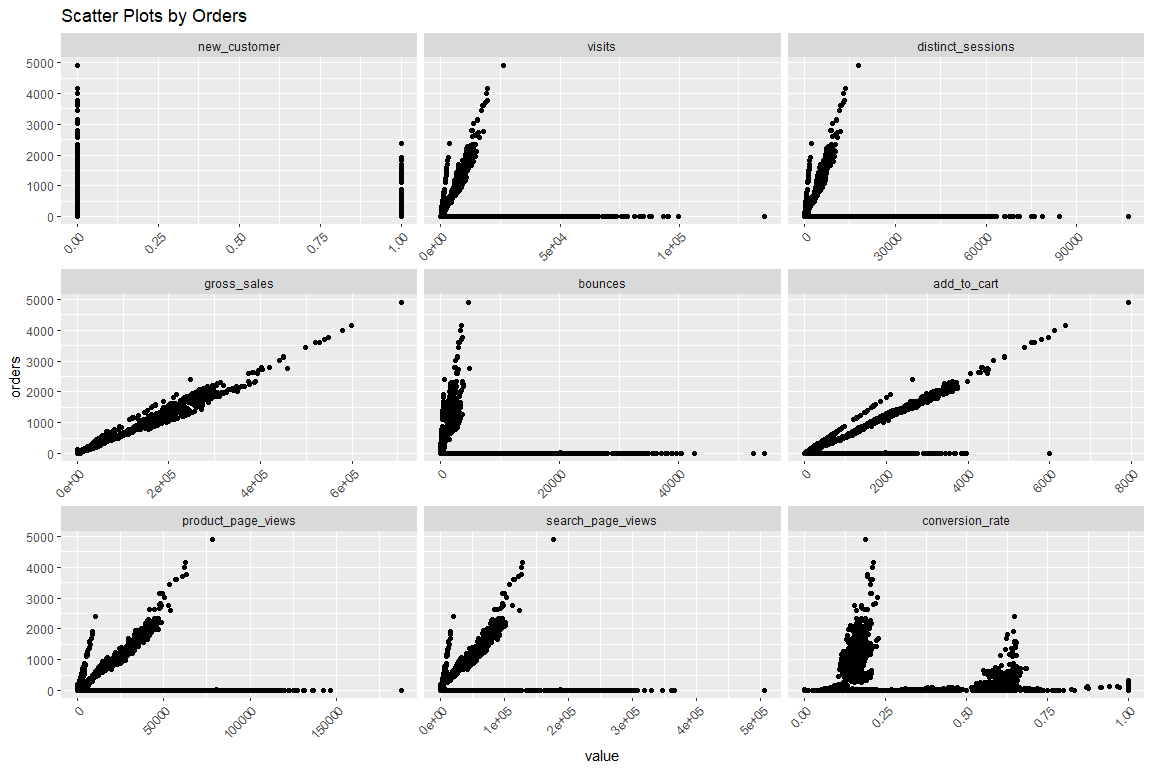
Distribution of the number of orders per type of user/customer as well as a scatterplot of the orders. The scatterplot can show us the relationships or correlation between each column and the orders.

boxplot(mydata$orders ~ mydata$new\_customer, xlab = 'Number of Orders', ylab='Type of User', horizontal = TRUE)

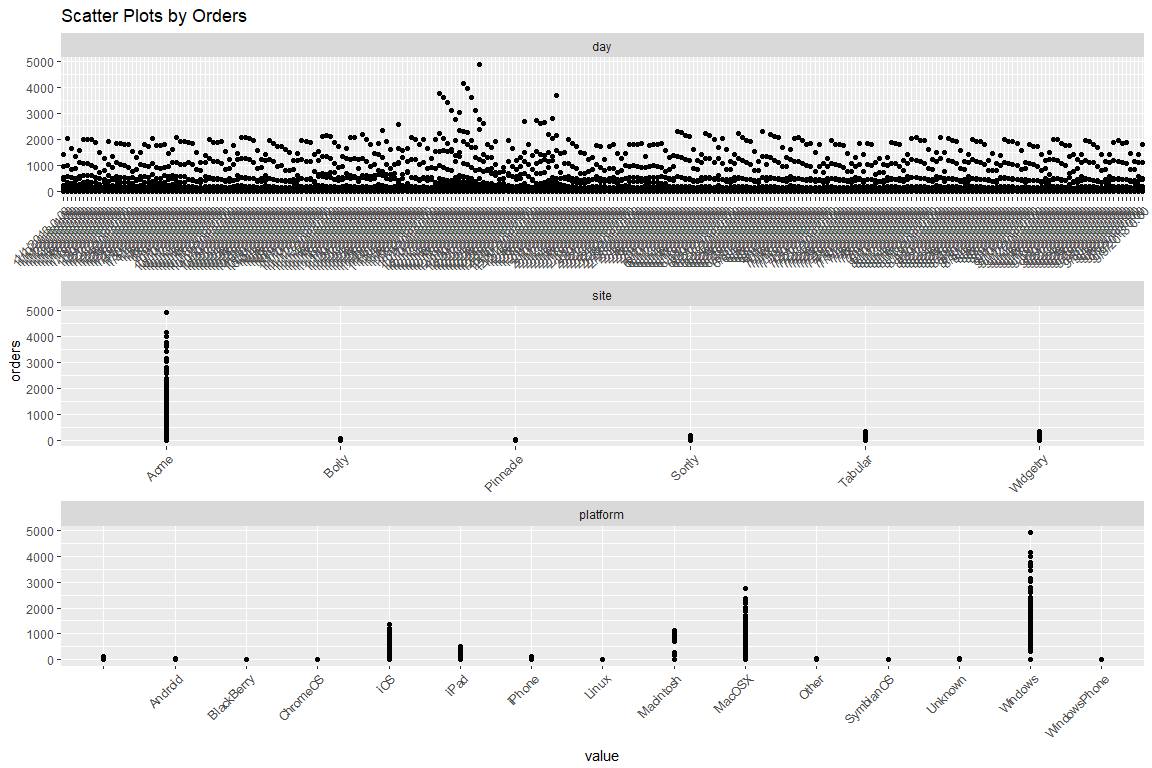
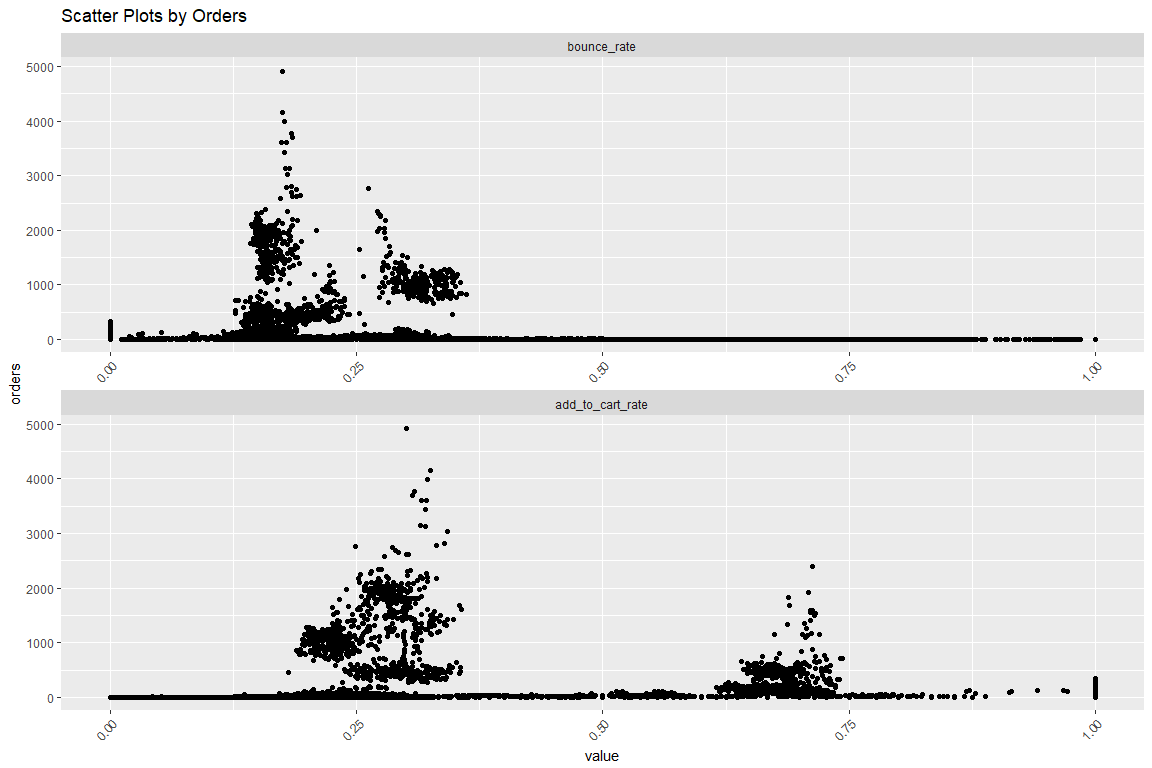


plot\_scatterplot(mydata, "orders", title = "Scatter Plots by Orders")

## Warning: Removed 20304 rows containing missing values (geom\_point).

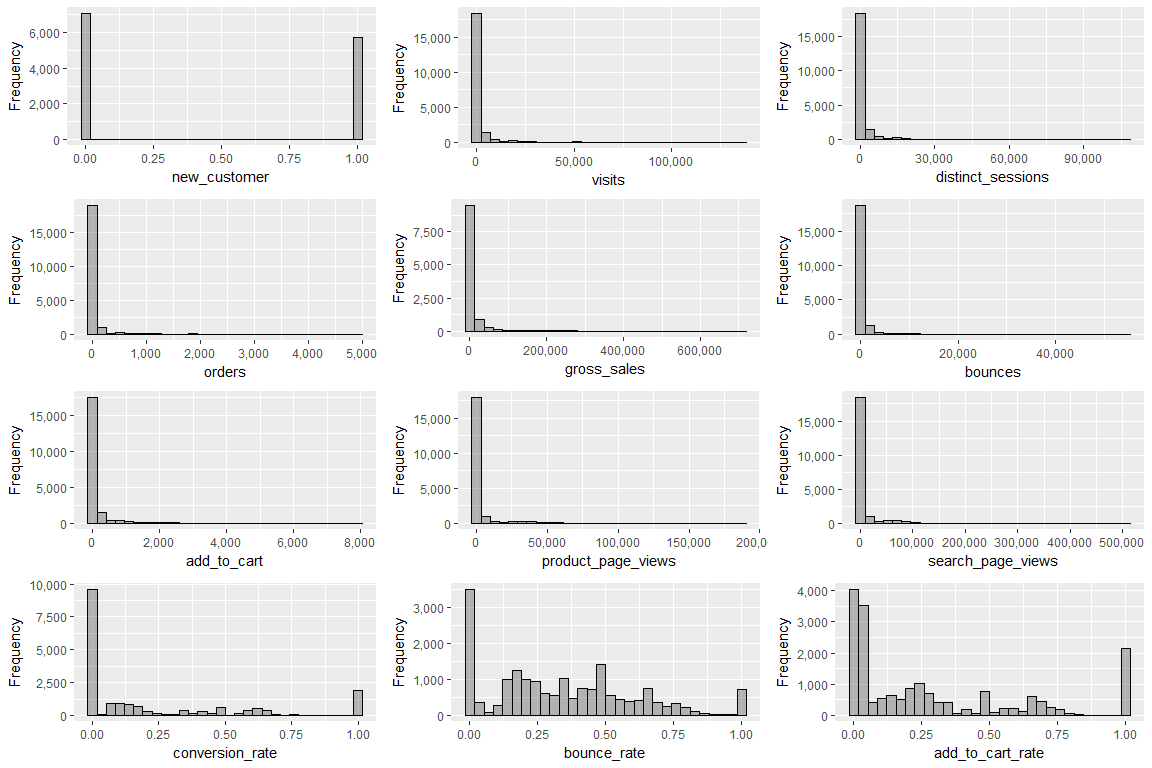


## Warning: Removed 4938 rows containing missing values (geom\_point).



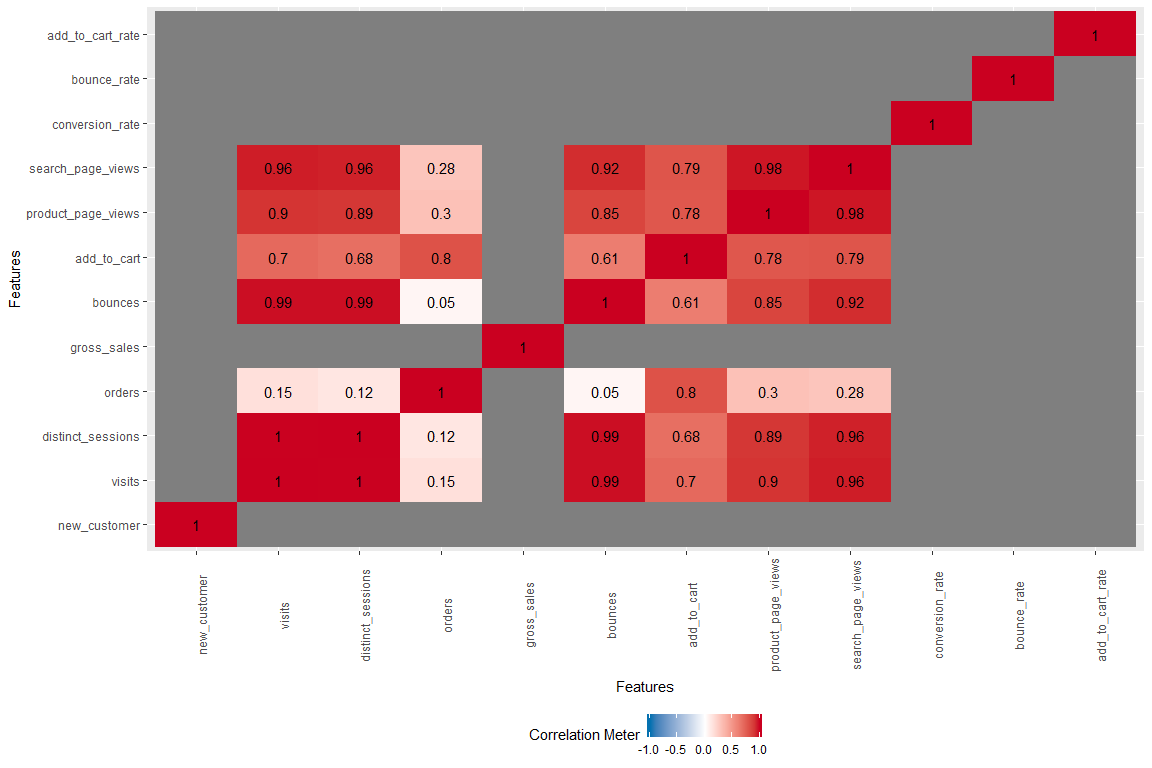
Here is a complete histogram of all continuous variables in the data set:

plot\_histogram(my\_data)

 It seems that most of the continuous data are right skewed with many missing values. Let’s take it a step further and look at the bivariate correlation of some variables with respect to the others. We will begin with the continuous variables.

plot\_correlation(mydata, type = "c")

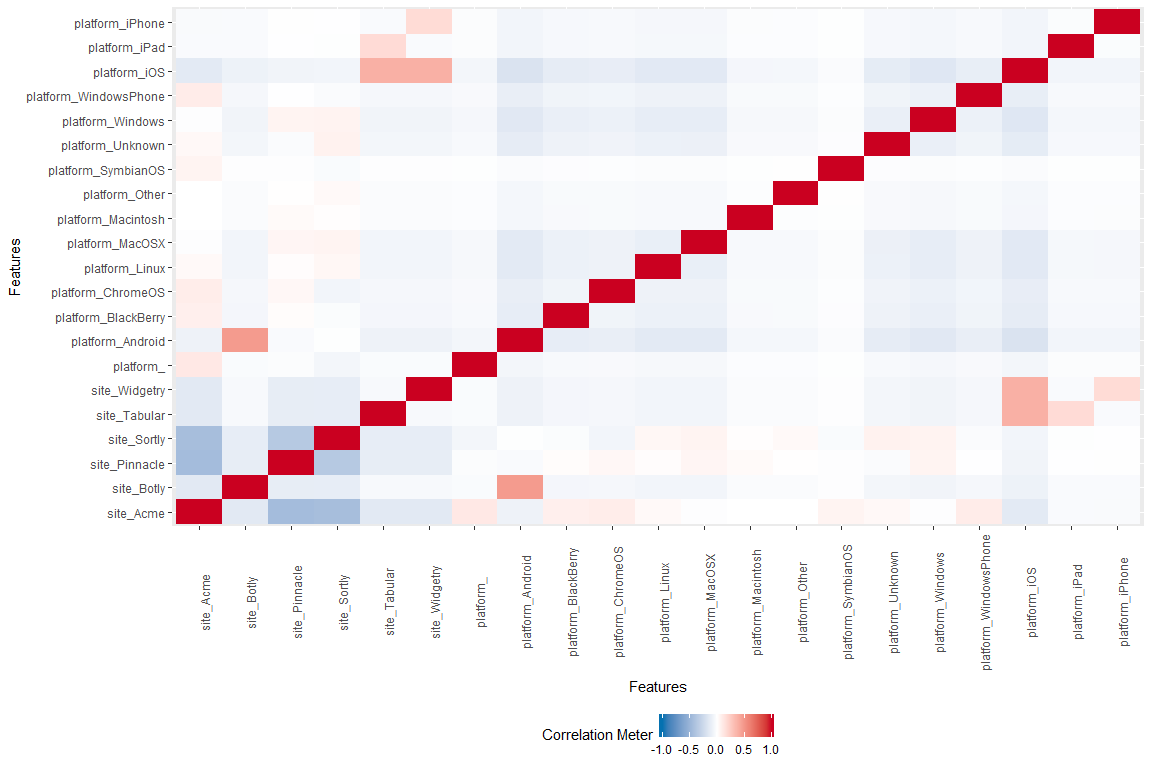
## Warning: Removed 90 rows containing missing values (geom\_text).



Now let’s look at the discrete variables.

plot\_correlation(mydata, type = "d")

## 1 features with more than 20 categories ignored!  
## day: 268 categories

 Using the plot above, Acme shows to have negative correlation to Sortly and Pinnacle, while there is a strong correlation between the use of Android platforms paired with the Botly site. We can also see that iOS users display a stronger correlation with the Widgetry and Tabular sites than with any other sites.