Step 1: Importing the data sets:

|  |
| --- |
| install.packages("downloader")  library(downloader)  transData <- read\_csv("F:/GitHub/datasciencetest/transData.csv")  View(transData)  testSamples <- read\_csv("F:/GitHub/datasciencetest/testSamples.csv")  View(testSamples) |

Step 2: Merging two data sets for removing duplicates:

|  |
| --- |
| total <- merge(transData,testSamples,by="sample\_id") |

1-What is the approximate probability distribution between the test group and the control group.

|  |
| --- |
| count0 <- length(which(testSamples$test\_group == 0))  count1 <- length(which(testSamples$test\_group == 1))  prop.test(c(44886, 14835), c(59721, 59721)) |

The probability of assigning to group 1 is equal to 0.752 and the probability of assigning to group 0 is equal to 0.248. the approximate probability distribution between the two group is binomial distribution:

1-Is a user that must call-in to cancel more likely to generate at least 1 addition REBILL?

We should compare the means of two groups (0= control group, 1=test group) in terms of the number of transaction by each sample\_id with aid of t-test. In this regard with should divide the data set into two parts, grouping by each group. The hypotheses test would be as follows:

=

If we could reject the H0  then we can conclude that the average of generated REBILL in user that must call-in to cancel significantly is higher than another group.

|  |
| --- |
| total\_1\_Rev <- data.frame(tapply(data1$transaction\_id, data1$sample\_id, FUN=length))  colnames(total\_1\_Rev) <- c("Num.Transaction")  total\_0\_Rev <- data.frame(tapply(data0$transaction\_id, data0$sample\_id, FUN=length))  colnames(total\_0\_Rev) <- c("Num.Transaction")  t.test(total\_1\_Rev$Num.Transaction,total\_0\_Rev$Num.Transaction)  ggplot(total\_1\_Rev, aes("sample\_id\_group1" , Num.Transaction)) + geom\_boxplot()  ggplot(total\_0\_Rev, aes("sample\_id\_group0" , Num.Transaction)) + geom\_boxplot() |

Welch Two Sample t-test:

data: total\_1\_Rev$Num.Transaction and total\_0\_Rev$Num.Transaction

t = -16.247, df = 1142.9, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

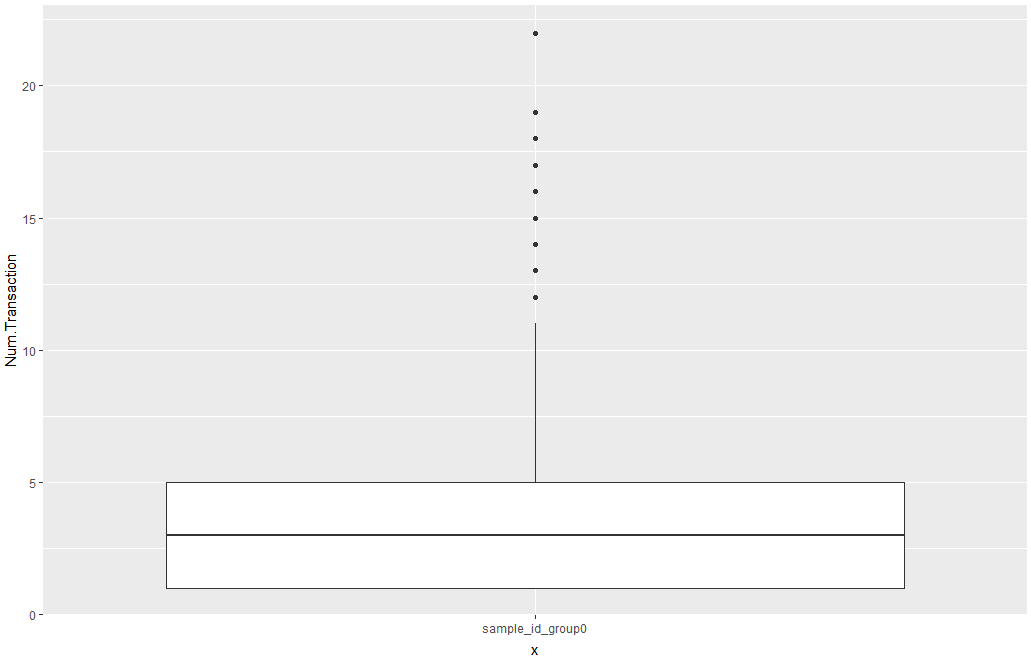
95 percent confidence interval:

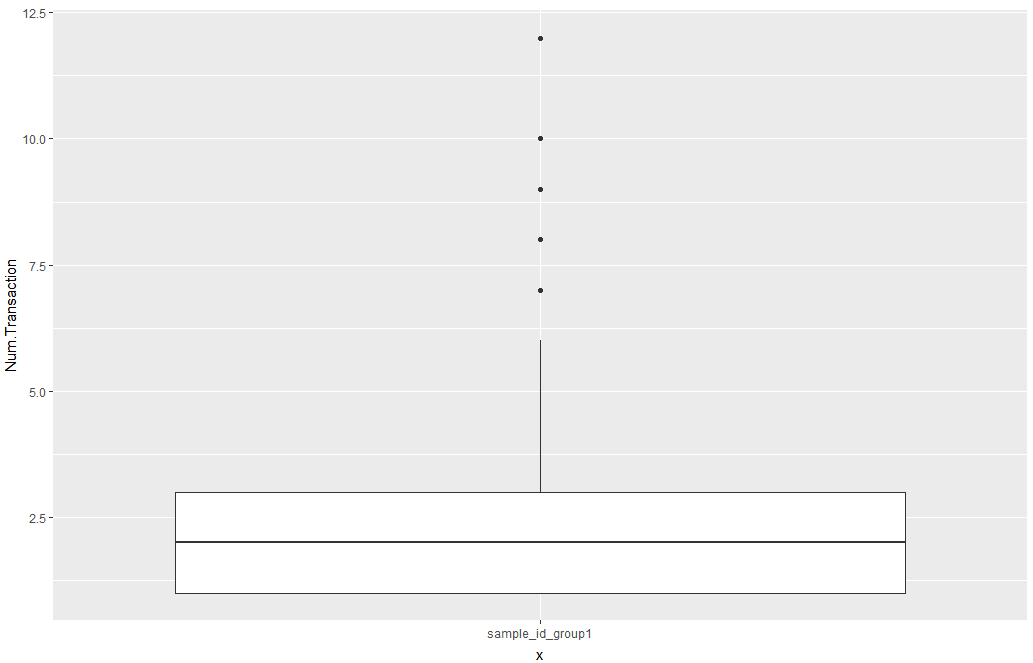
-2.165012 -1.698447

sample estimates:

mean of x mean of y

2.059769 3.991498





Considering the output of t-test we can see that the alternative hypothesis would be accepted so we can confirm that user that must call-in to cancel more likely to generate at least 1 addition REBILL.

3-Is a user that must call-in to cancel more likely to generate more revenues?

We should compare the means of two groups (0= control group, 1=test group) in terms of the amount of transaction by each sample\_id with aid of t-test. In this regard with should divide the data set into two parts, grouping by each group and sum the total amount of transaction for each sample\_id. The hypotheses test would be as follows:

=

If we could reject the H0  then we can conclude that the average amount of transaction in user that must call-in to cancel significantly is more than another group.

|  |
| --- |
| data1\_rev <- filter(total, test\_group == 1)  data0\_rev <- filter(total, test\_group == 0)  data11\_rev <- data.frame(tapply(data1\_rev$transaction\_amount, data1\_rev$sample\_id, FUN=sum))  colnames(data11\_rev) <- c("transaction\_amount")  data00\_rev <- data.frame(tapply(data0\_rev$transaction\_amount, data0\_rev$sample\_id, FUN=sum))  colnames(data00\_rev) <- c("transaction\_amount")  t.test(data11\_rev$transaction\_amount,data00\_rev$transaction\_amount)  ggplot(data11\_rev, aes("sample\_id\_group1" , transaction\_amount)) + geom\_boxplot() |

Welch Two Sample t-test:

data: data11\_rev$transaction\_amount and data00\_rev$transaction\_amount

t = -7.2863, df = 1478.2, p-value = 5.166e-13

alternative hypothesis: true difference in means is not equal to 0

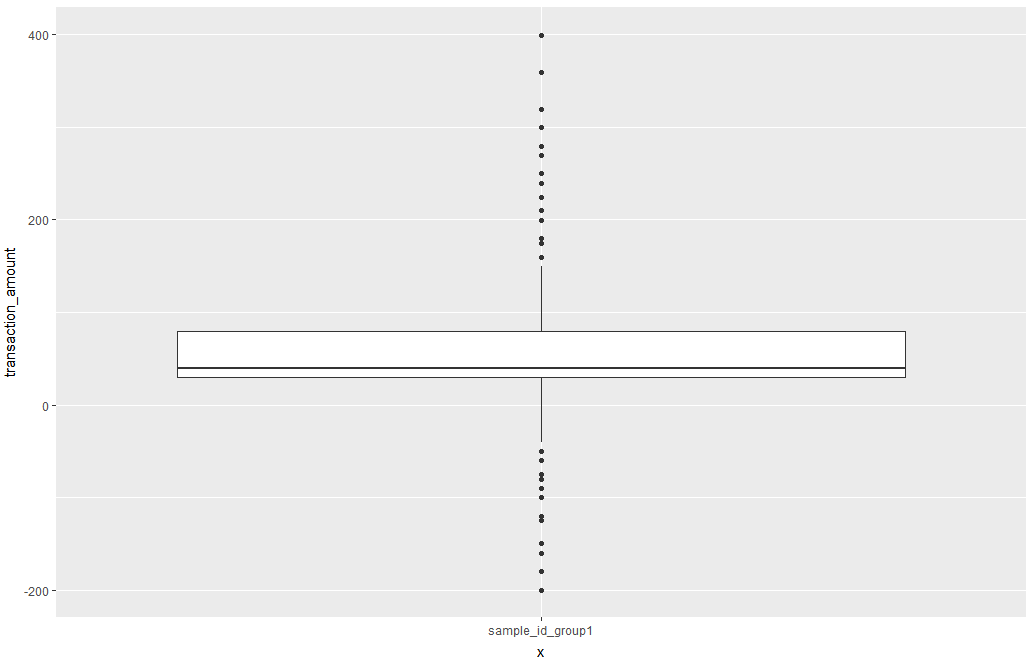
95 percent confidence interval:

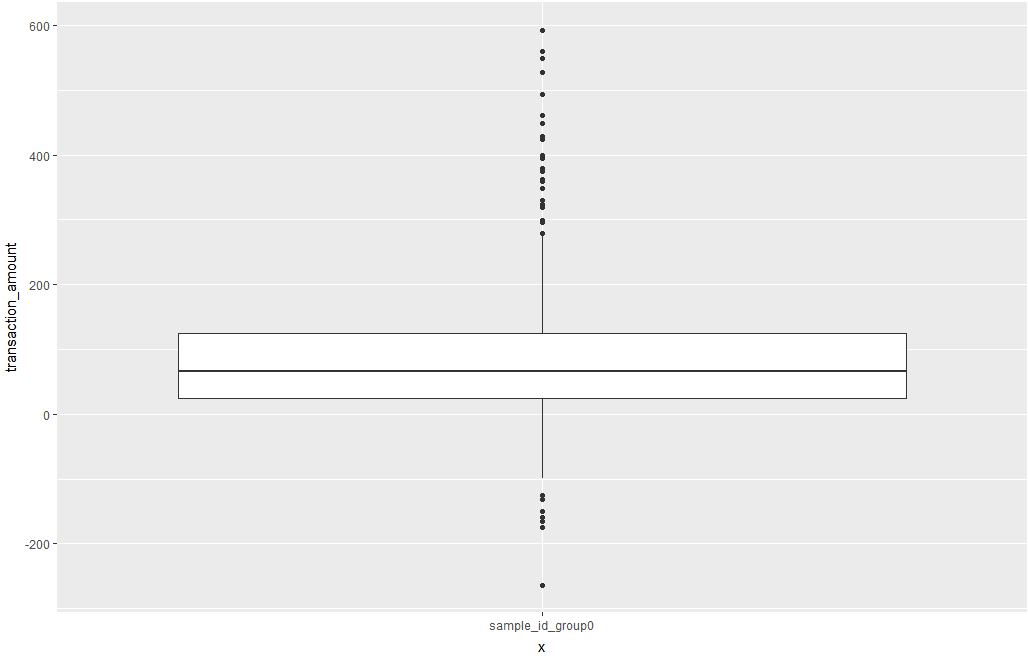
-31.59349 -18.19080

sample estimates:

mean of x mean of y

58.36911 83.26126





Considering the output of t-test we can see that the alternative hypothesis would be accepted. So, we can confirm that user that must call-in to cancel more likely to generate more revenues.

4-Is a user that must call-in more likely to produce a higher chargeback rate(CHARGEBACKs/REBILLs)?

Since we have many cases that users just have CHARGEBACK without any REBILL, for calculating the ratio of " CHARGEBACKs/REBILLs" for those users we will face difficulties. So we have to do two test. At the first test, we compare the proportion of users that only have CHARGEBACK in the two groups of 0 and 1 by t-test and also compare their average of CHARGEBACK amounts for both groups. Then we should compare the proportion and means of users that have CHARGEBACK and REBILLS together by t-test. By considering the result of tests we can reach to a conclusion about them.