Activity 3 - Aryan Khandelwal

**Question 1:** Download Act3-CustomerTrans.dat and put it in your STA 3000 folder. This is the entirety of some customer’s credit card transaction data. Each record has the amount of the transaction, the merchant involved (coded by, M326, for example, UNKNOWN when not available), the type of merchant (industry and super industry), the merchant’s and customer’s zip codes when available (along with the distance between these zips), whether the purchase was made on the internet, and the date of the purchase (measured in # of days since the beginning of data collection).

1. Import the file and name the dataframe CUST. Try this using read.csv. Determine the number of rows and columns in the dataset by running nrow(), ncol(), or dim() (using CUST as the argument to these functions), then show the first four lines of the data using head (see RBasic slides).

CUST <- read.csv("Act3-CustomerTrans.dat")

1. Show row 27 of the data on the screen. What is the amount spent for this transaction?

#Show row 27 on the screen  
CUST[27,]

## Amount DaysSinceStart Industry SuperIndustry Internet CZIP MZIP  
## 27 95 362 Insurance Finance and Insurance No 10583 51401  
## ID Merchant Distance  
## 27 37783 M179 949

*Comment:* The amount is 95 Dollar

1. What is the column name and number that contains the customer’s zip code?

#Hint: try using data.frame(names(CUST))  
data.frame(names(CUST))

## names.CUST.  
## 1 Amount  
## 2 DaysSinceStart  
## 3 Industry  
## 4 SuperIndustry  
## 5 Internet  
## 6 CZIP  
## 7 MZIP  
## 8 ID  
## 9 Merchant  
## 10 Distance

*Comment:* The column name is MZIP and it is the 6th column, that contains customer’s zip codes.

1. Using either summary or table, find how many times this person purchased something on the internet, column Internet.

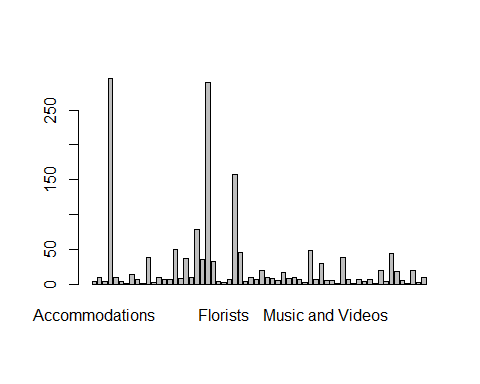
#Using summary or table on the CUST dataframe and look at the result for Internet  
table(CUST$Internet)

##   
## No Yes   
## 1405 146

*Comment:* 146 times someone purchased something from internet, while 1405 people did not purchase something from internet.

1. Make a barchart (see RBasic slides) of the Industry column. Sometimes plots in R are not that useful, this one of them (try resizing the plot to see which is the most common value; you can’t, and would have to use table instead).

#Barchart of Industry column  
barplot(table(CUST$Industry))



1. You see that Distance (how far away the customer is away from the store) is -999 in row 563, a bogus value. Replace this value by the number 2 using the left arrow symbol. Show row 563 on the screen after this is done to verify the change. Note: this is done very much like x <- 2, except instead of x you refer to the particular row/column with CUST[563,"Distance"] or CUST$Distance[563] or CUST[563,10] (since it is the 10th column).

#Code assigning the relevant entry to a 2  
#Showing row 563 on the screen  
CUST$Distance[563]<-2  
CUST[563,]

## Amount DaysSinceStart Industry SuperIndustry Internet  
## 563 21 378 T+E Railroad Transportation and Warehousing No  
## CZIP MZIP ID Merchant Distance  
## 563 10583 12312 37783 M2333 2

1. There are many values of Distance that are -999 (really a stand-in for an unknown distance). Let’s make a subset of the dataset called CUSTD that has the two columns Amount and Distance where the distance are at least 0. See RBasic slides. Once this is done, find the average amount of the Amount column and produce a summary of the Distance column.

#Making CUSTD to be a subset of CUST where Distance>=0, selecting only the Amount and Distance columns  
#mean of Amount column, e.g., mean(CUSTD$Amount)  
#summary of Distance column in the CUSTD dataset  
  
CUSTD<- subset(CUST, Distance>=0, select=c("Amount", "Distance"))  
mean(CUSTD$Amount)

## [1] 48.8022

summary(CUSTD$Distance)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 3.0 106.1 17.0 2245.0

**Question 2:** Make a vector called y that is a sequence that goes from 8 to 103 in increments of 1.8 (note: the sequence will actually end at 101.6, see RBasic slides). Using the length function combined with which, how many of the elements are larger than 90? See RBasic slides.

#create the described vector y using the seq command  
#find how many elements of y are larger than 90 using the length and which commands  
  
y = seq(8,103,by=1.8)  
length(y[which(y>90)])

## [1] 7