CustomersData

Abanather Negusu

ADS500B

7/26/2021

```
#install.packages("xlsx")
library("xlsx")
#install.packages("ggplot2")
library("ggplot2")
#install.packages("tidyverse")
library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

#install.packages("dplyr")
library("dplyr")

#Import your .csv file to your Global Environment
custdata <- read.csv("custdata.csv", header = TRUE)
```

#1 Write a multiplication script using either a “for” loop or a “while” loop.Show your script.(5 points)

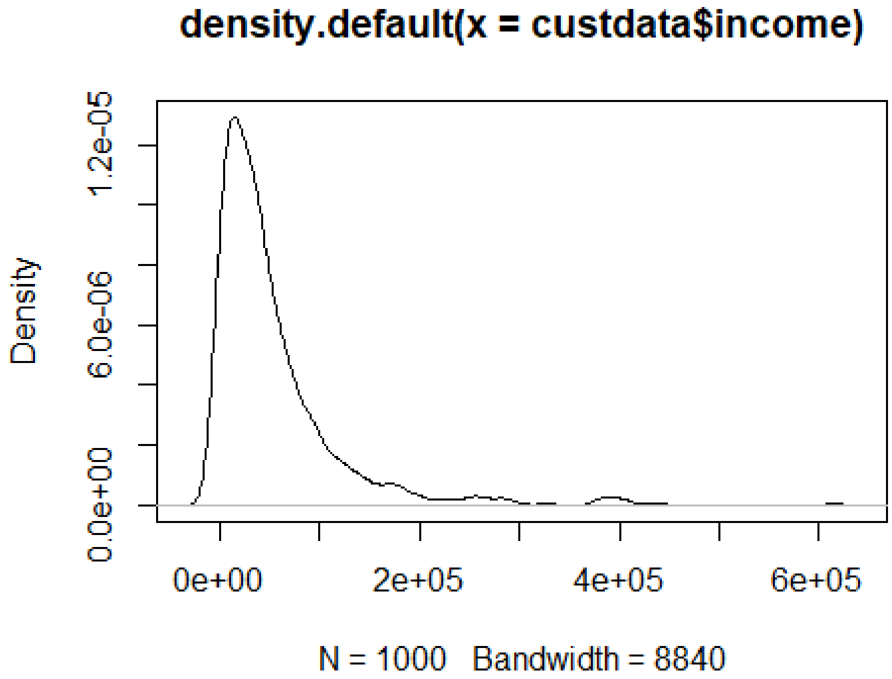
```
x <- 2
while (x < 5)
{
  print(x)
  x <- x * 2
}

## [1] 2
## [1] 4
```

#2.1: Figure out how to plot density of income. (5 points) #2.2: Provide a couple of sentences of description along with the plot. Imagine you #are explaining this to your manager or a senior leader. (5 points)

This density plot below shows a smoothed distribution of points along the numeric axis. According to the plot we can see a high distribution on the left side of our graph.

```
#2.1
plot(density(custdata$income))
```

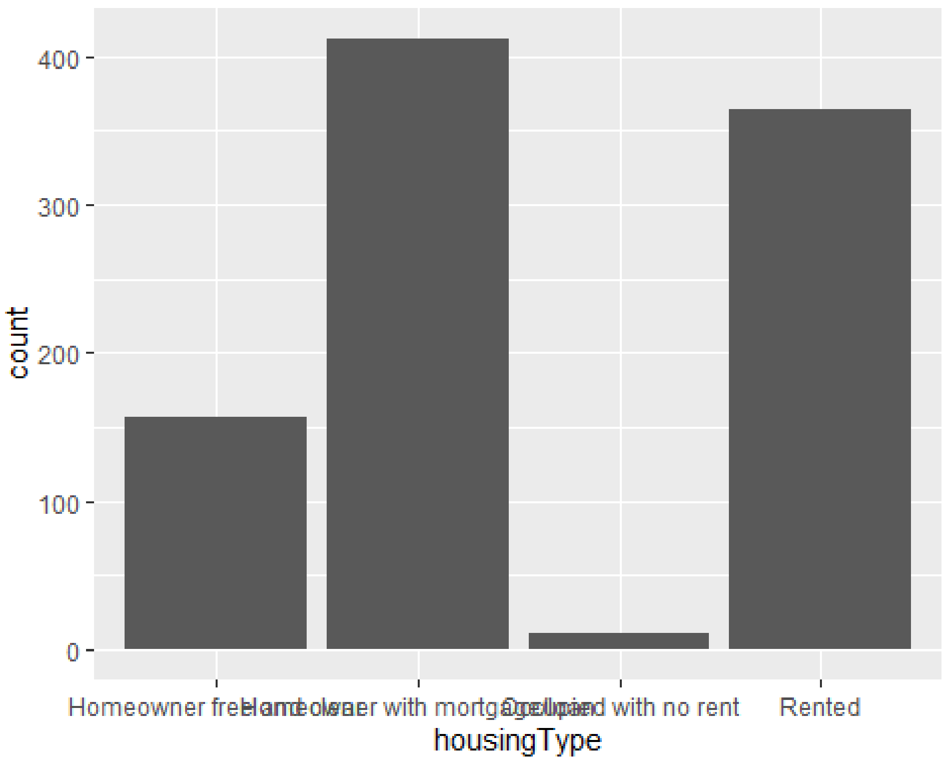


```
Xlab = "income"
ylab = "density"
```

#3.1: Create a bar chart for housing type using the customers data. Make sure to #remove the “NA” type. [Hint: You can use subset function with an appropriate #condition on housing type field.] Provide your commands and the plot. #(5 points)

Below the boxplot can show that there are a high proportion of customers that are home owners with loans or are renting.

```
#3.1
#removing nulls from house column
custdata2 <- custdata %>% filter(custdata$housingType != "")
#plotting
box_plot <- ggplot(custdata2, aes(x = housingType)) + geom_bar()
box_plot
```



#4.1: Extract a subset of customers that are married and have an income more than $50,000.(5 points)

```
#4.1
custdata3<- subset(custdata,income>50000 & marital.stat == "Married")
custdata3

##      custid sex is.employed income marital.stat health.ins
## 12    17134   M        TRUE 220000      Married       TRUE
## 24    30768   M        TRUE  80000      Married       TRUE
## 41    52197   M          NA  65100      Married       TRUE
## 44    52436   F        TRUE 139000      Married       TRUE
## 46    53214   M        TRUE  84010      Married       TRUE
## 48    54177   M       FALSE  51500      Married       TRUE
## 52    62999   M        TRUE  91000      Married       TRUE
## 55    67776   M        TRUE  52000      Married       TRUE
## 57    68221   M        TRUE  78000      Married       TRUE
## 58    69062   M        TRUE 120300      Married       TRUE
## 60    74447   M        TRUE 162000      Married       TRUE
## 63    78476   M        TRUE  76000      Married       TRUE
## 66    80549   M          NA  85200      Married       TRUE
## 67    82503   M        TRUE  70000      Married       TRUE
## 74    90863   M        TRUE 285020      Married       TRUE
## 76    94743   M        TRUE 299000      Married       TRUE
## 77    96964   M        TRUE 266200      Married       TRUE
## 79    98086   M          NA  52100      Married       TRUE
```

#4.2: What percentage of these customers have health insurance? (5 points

```
#Getting information of customers  that are married with health insurance
custdata4<- subset(custdata,income>50000 & marital.stat == "Married")
custdata4 <- custdata4[ , c( "custid", "marital.stat", "health.ins")]

# Using the same data from above to create table that shows How many people
are True and how many are False for having insurance
cust_insurance <- table(custdata4$health.ins) %>% data.frame()
cust_insurance

##    Var1 Freq
## 1 FALSE    8
## 2  TRUE  208

 # Referencing the row of Trues from the table of how many people have
insurance or not
insurance_True <- cust_insurance[2, ]


# Percentage of people who are married and have an income of $50000+ have
insurance
insurance_True$Freq *100/sum(cust_insurance$Freq)

## [1] 96.2963
```

#4.3: How does this percentage differ from that for the whole data set? (5 points)

The percentage of 84.1% differs from the whole data because that is how many customers have health insurance. According to the table below 841 customers have health insurance compared to the 159 who do not.

```
#4.3:
cust_insurance2 <- table(custdata$health.ins) %>% data.frame() # table that
shows How many people are True and how many are False
cust_insurance2

##    Var1 Freq
## 1 FALSE  159
## 2  TRUE  841

insurance_True2 <- cust_insurance2[2, ] # Referencing the row of Trues from
the table of how many people have insurance or not


# Percentage of all the people in the data set who have insurance
insurance_True2$Freq *100/sum(cust_insurance2$Freq)

## [1] 84.1
```

#5.1: In the customers data, do you think there is any correlation between age, #income, and number of vehicles? Explain why or why not. (5 points)

```
#5.1 Here I am cleaning my data to filter and display ages greater/ equal to
18 & less than equal to 93 with an income greater than zero.
Clean_data <- custdata %>% filter(age >= 18 & age <= 93 &  income > 0) %>%
select(num.vehicles, age, income)
#5.1 Here is my correlation
Clean_data$income <- as.numeric(Clean_data$income)
cor(Clean_data %>% select(num.vehicles,income,age))

##              num.vehicles      income         age
## num.vehicles   1.00000000  0.10566245 -0.03425412
## income         0.10566245  1.00000000 -0.02249358
## age           -0.03425412 -0.02249358  1.00000000
```

#5.2: Report your correlation numbers and interpretations. [Hint: Make sure to #remove invalid data points, otherwise you may get incorrect answers!] (10 points)

I believe that my correlation numbers show that most are inversely correlated. Looking at the correlation between income and age I could see some patterns there. Although there might not be causation there can still be a pattern found within those two variables.