**BANA 200**

**Assignment 1 Data Cleaning**

**Due Wednesday, August 11th on Canvas by 6PM Pacific Standard Time (1AM UTC Time)**

**50 Points**

**You must include your R script with all of your code in order to receive full credit**

Imagine for a moment that you have been hired by Starbuck’s Corporation as a data scientist. Your job over the next several weeks is to conduct some insightful analysis to help senior management understand more about how to improve customer engagement and profitability. In this first assignment, you will prepare the dataset for analysis by cleaning it.

**Q1 (10 Points)**

Import the “starbucks final data.txt” dataset into R. Using R, report the **number of missing values** (NA values) for each one of the 27 variables. How many missing values are there for **each variable** and also for the **entire dataset**? That is, report both the missing values for each variable and the sum of missing values across all variables.

**Answer:-**

The number of missing values for each column is as follows:-

X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13

173 187 207 194 178 215 209 175 190 188 183 199 194

X14 X15 X16 X17 X18 X19 X20 X21 X22 satis100 recommend profits ZipCode

216 189 220 200 212 179 190 209 210 206 206 197 0

Income

0

The total sum of missing values across all variables is **4926**.

**Code:-**

colSums(is.na(data1))

sum(is.na(data1))

**Q2 (10 Points)**

Using R, **strip out** all rows of data where there are ANY missing values (NA values). Once you have removed any and all rows with missing values, report below the **number of rows in the dataset that remain**. How many non-missing rows are there? Does it seem like we are throwing away a lot of data by removing all the rows with missing data?

**Answer:-**

There are **6121 non-missing rows**. Considering that there are 10000 rows in total it definitely seems that we are **throwing away a lot of data** by removing all the rows with missing data.

**Code:-**

omitted <- data1[complete.cases(data1),]

head(omitted)

dim(omitted)

dim(data1)

**Q3 (10 Points)**

The 22 variables X1 – X22 should only contain the values 1,2,3,4, or 5. Report for each one of these 22 variables the number of impossible values. Impossible values are defined as any values that less than 1 or values greater than 5. How many impossible values are there for each variable? Also report the total number of impossible values across all 22 variables (the sum). Use your cleaned dataset from Q2 above for answering this question (perform this analysis on the dataset with NO NA values). For example, the variable X15 has the following values:

table(starbucks.complete$X15)

-1 0 1 2 3 4 5 6 7

2 9 190 1044 2120 1969 686 93 8

Therefore, there are a total of 2 + 9 + 93 + 8 = 112 impossible values. The first row are the values in the variable X15 and the second row is the count of the number of surveys with those values. For example, there were 9 customers who answered a “0”. Repeat this calculation for all 22 variables X1 – X22 and report the total number of impossible values for each variable and the sum of impossible values across all 22 variables below.

**Answer:-**

[1] "For column 1 the number of invalid values is 103"

[1] "For column 2 the number of invalid values is 71"

[1] "For column 3 the number of invalid values is 68"

[1] "For column 4 the number of invalid values is 131"

[1] "For column 5 the number of invalid values is 101"

[1] "For column 6 the number of invalid values is 68"

[1] "For column 7 the number of invalid values is 22"

[1] "For column 8 the number of invalid values is 19"

[1] "For column 9 the number of invalid values is 20"

[1] "For column 10 the number of invalid values is 12"

[1] "For column 11 the number of invalid values is 26"

[1] "For column 12 the number of invalid values is 13"

[1] "For column 13 the number of invalid values is 10"

[1] "For column 14 the number of invalid values is 92"

[1] "For column 15 the number of invalid values is 103"

[1] "For column 16 the number of invalid values is 102"

[1] "For column 17 the number of invalid values is 115"

[1] "For column 18 the number of invalid values is 125"

[1] "For column 19 the number of invalid values is 189"

[1] "For column 20 the number of invalid values is 151"

[1] "For column 21 the number of invalid values is 67"

[1] "For column 22 the number of invalid values is 77"

The sum of impossible values across the 22 columns is **1685.**

**Code:-**

tot\_inv\_val=0

for (i in 1:22)

{

print(paste("For column ",i," the number of invalid values is ",sum((omitted[,i]>5)\*1)+sum((omitted[,i]<0)\*1)))

tot\_inv\_val = tot\_inv\_val + sum((omitted[,i]>5)\*1)+sum((omitted[,i]<0)\*1)

}

print(tot\_inv\_val)

**Q4 (10 Points)**

Management has asked that for variables X1-X22, you **replace** the impossible values with better numbers. Specifically, they have asked that you do the following:

a. For any values less than 1 (< 1), replace these values with 1. For example, replace -1 with 1, replace 0 with 1 etc.

b. For any values greater than 5, replace these values with 5. For example, replace 6 with 5, replace 7 with 5 etc.

c. Once you have replaced all of these values for X1 – X22, do a **frequency count** and report the total numbers of 1s, 2s, 3s, 4s, and 5s across all 22 variables AFTER replacement. If you did this correctly, there should no longer be any impossible values in the dataset for X1-X22.

**Answer:-**

[1] "The frequency count for column 1 is "

1 2 3 4 5

19 382 1955 2665 1100

[1] "The frequency count for column 2 is "

1 2 3 4 5

10 300 1971 2820 1020

[1] "The frequency count for column 3 is "

1 2 3 4 5

2 136 1678 3124 1181

[1] "The frequency count for column 4 is "

0 1 2 3 4 5

1 21 412 1868 2660 1159

[1] "The frequency count for column 5 is "

1 2 3 4 5

7 200 1617 3021 1276

[1] "The frequency count for column 6 is "

1 2 3 4 5

4 178 1568 3066 1305

[1] "The frequency count for column 7 is "

1 2 3 4 5

17 531 2452 2551 570

[1] "The frequency count for column 8 is "

0 1 2 3 4 5

1 68 903 2731 2005 413

[1] "The frequency count for column 9 is "

1 2 3 4 5

28 607 2575 2415 496

[1] "The frequency count for column 10 is "

1 2 3 4 5

28 625 2696 2393 379

[1] "The frequency count for column 11 is "

0 1 2 3 4 5

1 34 632 2545 2355 554

[1] "The frequency count for column 12 is "

1 2 3 4 5

67 871 2631 2145 407

[1] "The frequency count for column 13 is "

0 1 2 3 4 5

4 145 1322 2871 1588 191

[1] "The frequency count for column 14 is "

2 3 4 5

96 1389 3168 1468

[1] "The frequency count for column 15 is "

0 1 2 3 4 5

11 190 1044 2120 1969 787

[1] "The frequency count for column 16 is "

1 2 3 4 5

7 189 1677 2940 1308

[1] "The frequency count for column 17 is "

1 2 3 4 5

8 255 1668 2964 1226

[1] "The frequency count for column 18 is "

1 2 3 4 5

2 69 1200 3153 1697

[1] "The frequency count for column 19 is "

2 3 4 5

146 1280 3017 1678

[1] "The frequency count for column 20 is "

2 3 4 5

97 1227 3166 1631

[1] "The frequency count for column 21 is "

1 2 3 4 5

20 423 2214 2632 832

[1] "The frequency count for column 22 is "

2 3 4 5

101 1420 3308 1292

**Code:-**

for (i in 1:22)

{

omitted[,i][omitted[,i] > 5] <- 5

omitted[,i][omitted[,i] < 0] <- 0

print(paste("The frequency count for column ",i," is "))

print(table(omitted[,i]))

}

**Q5 (10 Points)**

Last but not least, the variables “satis100” and “recommend” also have impossible values. The range of satis100 should lie between 0 and 100, and the range of recommend should be between 0 to 10. For **satis100**, replace any values that are less than 0 with 0, and **replace** any values greater than 100 with 100. For **recommend**, replace any values that are less than 0 with 0, and replace any values that are greater than 10 with 10. Finally, for “**recommend**” only, report below the **counts of the number of unique values** for “recommend” after you’ve replaced the impossible values (i.e. how many 0’s, 1’s…10’s etc. are there in the final dataset for variable “recommend”). Also report the **average values (means) for all variables** in this final cleaned dataset below. That is, report the averages for X1-X22, satis100, recommend, income, profits, and zipcode. Note that the mean for zipcode is meaningless (as it is a postal code) but report it anyway.

**Answer:-**

The counts of the number of unique values for the recommend column are as follows:-

0 1 2 3 4 5 6 7 8 9 10

67 106 237 384 594 819 932 956 761 542 723

The average values for all variables in the final cleaned dataset is as follows:-

[1] "The average of column 1 is 3.72618853128574"

[1] "The average of column 2 is 3.74170887109949"

[1] "The average of column 3 is 3.87338670151936"

[1] "The average of column 4 is 3.73860480313674"

[1] "The average of column 5 is 3.87551053749387"

[1] "The average of column 6 is 3.89691226923705"

[1] "The average of column 7 is 3.51070086587159"

[1] "The average of column 8 is 3.29227250449273"

[1] "The average of column 9 is 3.44829276262049"

[1] "The average of column 10 is 3.40352883515765"

[1] "The average of column 11 is 3.45090671458912"

[1] "The average of column 12 is 3.31922888416925"

[1] "The average of column 13 is 3.05652671132168"

[1] "The average of column 14 is 3.98153896422153"

[1] "The average of column 15 is 3.34079398791047"

[1] "The average of column 16 is 3.87453030550564"

[1] "The average of column 17 is 3.84054892991341"

[1] "The average of column 18 is 4.05767031530796"

[1] "The average of column 19 is 4.01731743179219"

[1] "The average of column 20 is 4.0343081195883"

[1] "The average of column 21 is 3.62620486848554"

[1] "The average of column 22 is 3.94608724064695"

[1] "The average of column 23 is 54.8328704460056"

[1] "The average of column 24 is 6.31971900016337"

[1] "The average of column 25 is 100.992958666884"

[1] "The average of column 26 is 92524.2261068453"

[1] "The average of column 27 is 98287.1563470021"

**Code:-**

omitted$satis100[omitted$satis100 > 100] <- 100

omitted$satis100[omitted$satis100 < 0] <- 0

omitted$recommend[omitted$recommend > 10] <- 10

omitted$recommend[omitted$recommend < 0] <- 0

table(omitted$recommend)

for (i in 1:length(omitted)){

print(paste("The average of column ",i," is ",mean(omitted[,i])))

}