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

Overview:

As mentioned in class, perhaps one of the most important elements to good data science is making sure your data is correct before it can be analyzed. This first homework assignment will provide you with hands-on experience with some of the typical kinds of data cleaning and data manipulations required of many datasets before they can be analyzed.

The tab-delimited text file "starbucks final data.txt" contains survey data on a random sample of 10,000 Starbucks Coffee customers. The survey was done in Orange County, CA, and contains the following data:

1. **X1**: Overall, how would you rate the beverages served at Starbucks? - Taste
2. **X2**: Overall, how would you rate the beverages served at Starbucks? - Overall quality
3. **X3**: Overall, how would you rate the beverages served at Starbucks? - Temperature
4. **X4**: Overall, how would you rate the beverages served at Starbucks? - Freshness
5. **X5**: Overall, how would you rate the beverages served at Starbucks? - Presentation
6. **X6**: Overall, how would you rate the beverages served at Starbucks? - Variety
7. **X7**: Overall, how would you rate the food served at Starbucks? - Temperature
8. **X8**: Overall, how would you rate the food served at Starbucks? - Variety
9. **X9**: Overall, how would you rate the food served at Starbucks? - Taste
10. **X10**: Overall, how would you rate the food served at Starbucks? - Overall quality
11. **X11**: Overall, how would you rate the food served at Starbucks? - Presentation
12. **X12**: Overall, how would you rate the food served at Starbucks? - Freshness
13. **X13**: How do you rate the value for the money?
14. **X14**: How would you rate the Starbucks staff along the following dimensions? - Well dressed and appear neat
15. **X15**: How would you rate the Starbucks staff along the following dimensions? - Remembering your name
16. **X16**: How would you rate the Starbucks staff along the following dimensions? - Knowledgeable
17. **X17**: How would you rate the Starbucks staff along the following dimensions? - Personal treatment
18. **X18**: How would you rate the Starbucks staff along the following dimensions? - Polite
19. **X19**: How would you rate the Starbucks staff along the following dimensions? - Remembering your order correctly
20. **X20**: How would you rate the Starbucks staff along the following dimensions? - Friendly/attentive
21. **X21**: How would you rate the Starbucks staff along the following dimensions? - Have your best interest at heart

22. **X22**: How would you rate the Starbucks staff along the following dimensions? - Providing prompt service
23. **satis100**: A customer satisfaction variable that ranges from 0 to 100. Customers were asked the following question: "Overall, how satisfied are you with Starbucks? 0 = very dissatisfied; 100 = very satisfied."
24. **recommend**: "How likely are you to recommend Starbucks to others? 0 = definitely WILL NOT recommend; 10 = definitely WILL recommend." This variable ranges from 0 to 10.
25. **profits**: Average monthly profits that Starbucks earns on each customer (in US Dollars). Some profit numbers may be negative (i.e. Starbucks loses money on some customers).
26. **ZipCode**: The five digit zip code associated with the customer's place of residence.
27. **Income**: Estimated annual income of each customer (reported in US Dollars), based on the US Census Bureau Zip Code demographics data.

Variables X1 – X22 are all measured on a 5 point scale (1 = terrible, 2 = poor, 3 = average, 4 = good, 5 = excellent).

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.

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?

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.

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.

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.