Data Wrangling QBS 181

Midterm

Instructions:

* You need to work independently and collaboration is NOT allowed
* Please save the files in this format:
* Name of the file should be username(database)\_Midterm\_QBS181
* Please use both SQL and R. You can use R markdown to submit the code written in R. For SQL, please follow the same instructions as in the assignments

Visit the website <https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DIQ_I.htm#Data_Processing_and_Editing>

1. The DIQ\_I.xpt(will be uploaded on canvas) file has some problems with its data (e.g., missing values, numeric columns stored as chars, etc.) and need to be cleaned before further use.
2. List the data-related issues you see in this data set
3. How will you address each data-related issue?
4. Give justification for why you chose a particular way to address each issue. For example, if you decide to address missing values by removing rows or filling empty data cells, justify your decision or if you want to create a PHI field like year of Birth
5. Clean the data by addressing each point listed in 1.

Verify that whether the counts of each code or value for various variables are correct as mentioned in the website

Data frame issues and how I fixed them.

1. In this data set, I first double check the values are same as the website provided. There are five groups of people. They mixed up in a large data frame at first, which are very informative. I first split the large data frame into five small data frames (1,2,3,7,9) based on “Doctor told you have diabetes”. Most of the questions and surveys are based on people who have already been diagnosed diabetes. So, it didn’t make sense if we asked the normal people how many and how often insulin did they intake per day. So, I am mainly focus on and doing the data manipulation in group 1, which only contained people who are diagnosed diabetes. After imputation, we get the complete results for people who have diabetes. We could use these results to do further analysis and prediction.
2. There are many missing values in the data frame. If the column only contains missing values, I will leave them as NA. If the column has the numerical meaning, for example, the length of time taking insulin, how long saw a diabetes specialist, how many times see the doctors… I will impute the mean value to the missing value. For the factor data type column, for example, there is only 1(Yes), 2(No), and NA. I will impute NA as 0. At this time, I could get the informative survey results for the diabetes patients.

For the numerical columns, the way I imputed the mean value for missing value is that I first make a scatter plot. Then, I visualize the value of outliers. I defined a threshold to make the outliers to NA value. Then, I will impute the mean to all of the missing values. This mean is more accurate than we impute the mean first, because there are many outliers first. For example: People cannot live more than 200 years. So, we could set x>200 to make the values which are bigger than 200 become the missing values.

If in the numerical columns, missing values are much more than normal values, I will set the missing values to 0. The mean imputation doesn’t make sense in this scenario.

1. Finally, I merge these five data frames back to the large data frame.

Chart, scatter chart

Description automatically generated

Chart

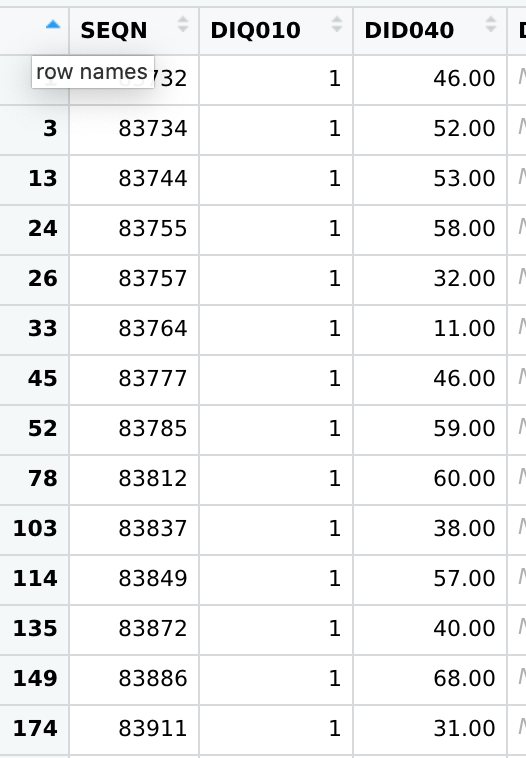
Description automatically generated

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generated



Table

Description automatically generated

A close up of a telephone

Description automatically generated