HW04 — STAT/CS 287  
NAME: Cecily Page   
DATE: October 30, 2018

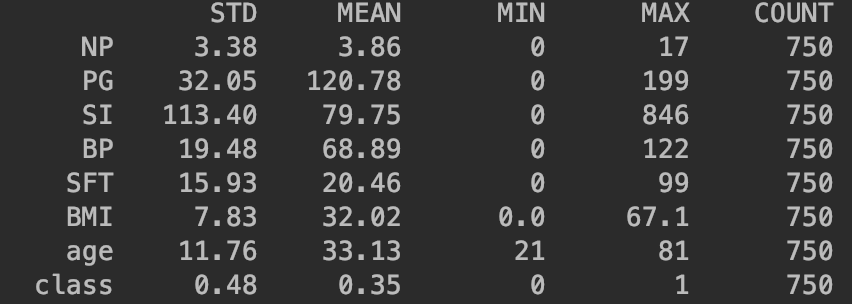
## P1.1

provide a narrative describing how load\_data() works (not how to use it).

Load\_data() works by taking in a file path, opening the file, reading it line for line , reading each line in as a json dictionary, then adding that dictionary to a list. The function then returns the list to be further used.

## P1.2

describing all the columns and what you may or may not understand about them



This is the summary statistics for the raw data. Because the minimum is 0 for blood pressure (meaning the patient would be dead) it implied that 0 is the default value when there is no other value. The same goes for PG, BMI, STF, SI. Because 0 is a value and counts towards the mean and standard deviation, this implied that those two statistics are not reliable because having x amount of 0s in a list of values can seriously sway their statistical meanings.

## P2.1

many observations are missing for each variable?

Total Missing: 635

SI Missing: 363

BMI Missing: 10

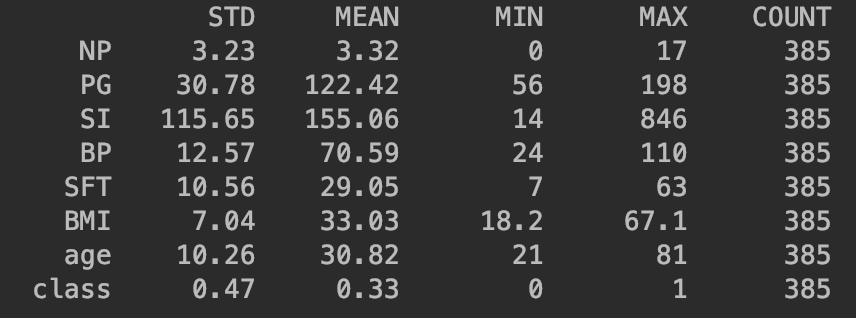
BP Missing: 35

PG Missing: 5

SFT Missing: 222

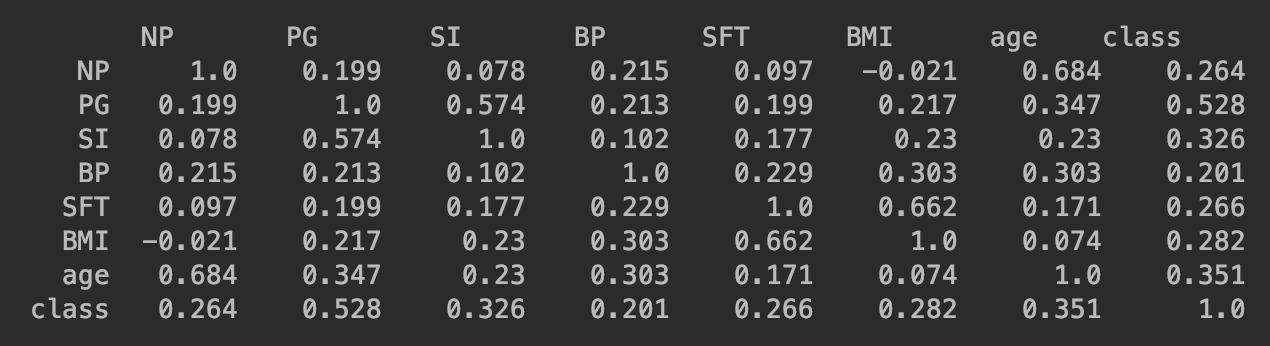
## P2.2

Report the (Pearson) correlation coefficient between each pair of variables on the sanitized dataset



Cleaned Data Set Stats

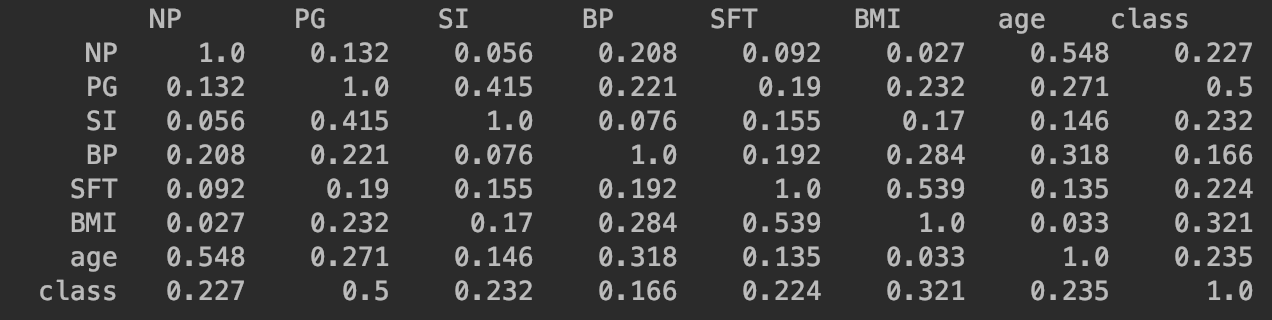
These statistics are more accurate. As you can see there are real minimum values have appeared and the means have all gone up because they’re not accounting for so many 0s. The standard deviations have gotten smaller (for the most part) implying that their data sets were much more spread apart before because of these 0s. there are also almost only half as many counts (number of rows still used) because 365 rows were removed with the listwise\_deletion() method. I feel much more comfortable with these numbers because they are within ranges of viable medical statistics.



Correlation Coefficients:

No values stand out too much to me except SFT and BMI (0.662) and NP to age (0.684). NP to age makes sense, because the older you are the more time you’ve had to have more pregnancies considering the fact that they’re 9 months each and women can’t have children until 13 or so, someone who’s 20 has only had the chance to have 9 pregnancies if they’re having a kid one day and pregnant the next. SFT and BMI probably have something to do with the amount of fat you have, so the higher BMI gets the thicker your skin (Just a guess, I’m not a medical expert)

## P3.1



Correlation coefficients for imputation means.

The correlations are overall weaker than before because we are somewhat “neutralizing” the missing values by making them the mean instead of removing that line completely.

Here are the graphs. The Red are the values that have mean changed to the mean.

