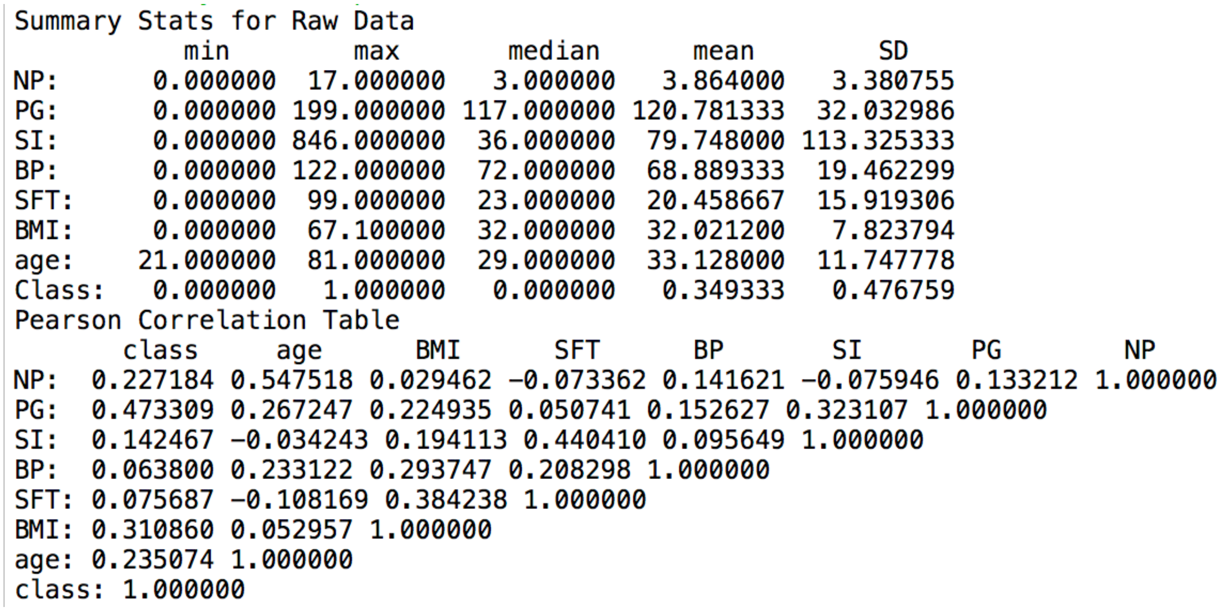
HW04 — STAT/CS 287
NAME: Galen Byrd
DATE: 10/30/18

P1.1

My loadData function reads in one json object at a time, appending each datapoint to a list (lines 22-29) that serves as the "column" for that key. One observation would be found by getting the value from each list at identical indices (lines 33-48). I then create a dictionary where the keys are the column keys and the value is the list of data for that column (lines 49-56).

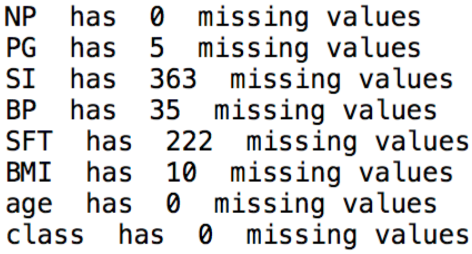
P1.2



```
Summary Stats for Raw Data
              min        max       median       mean        SD
NP:       0.000000  17.000000   3.000000   3.864000   3.380755
PG:       0.000000 199.000000 117.000000 120.781333  32.032986
SI:       0.000000 846.000000  36.000000  79.748000 113.325333
BP:       0.000000 122.000000  72.000000  68.889333  19.462299
SFT:      0.000000  99.000000  23.000000  20.458667  15.919306
BMI:      0.000000  67.100000  32.000000  32.021200   7.823794
age:     21.000000  81.000000  29.000000  33.128000  11.747778
Class:    0.000000   1.000000   0.000000   0.349333   0.476759
Pearson Correlation Table
       class     age      BMI      SFT      BP       SI       PG       NP
NP:   0.227184 0.547518 0.029462 -0.073362 0.141621 -0.075946 0.133212 1.000000
PG:   0.473309 0.267247 0.224935 0.050741 0.152627 0.323107 1.000000
SI:   0.142467 -0.034243 0.194113 0.440410 0.095649 1.000000
BP:   0.063800 0.233122 0.293747 0.208298 1.000000
SFT: 0.075687 -0.108169 0.384238 1.000000
BMI: 0.310860 0.052957 1.000000
age: 0.235074 1.000000
class: 1.000000
```

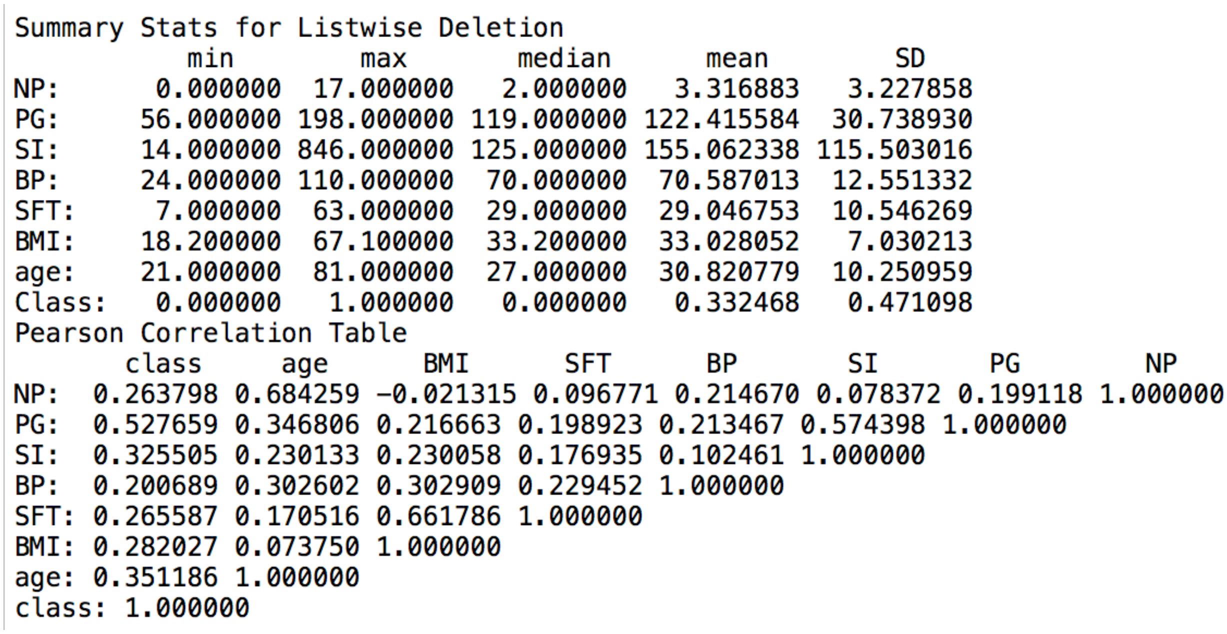
Many of these columns have many zero values when they represent something that can not possibly be zero. Plasma Glucose, Serum Insulin, Blood Pressure, Skin Fold Thickness, Body Mass Index and age all can not be zero. The class and Number of times Pregnant can both be zero.

P2.1



```
NP  has  0  missing values
PG  has  5  missing values
SI  has  363  missing values
BP  has  35  missing values
SFT  has  222  missing values
BMI  has  10  missing values
age  has  0  missing values
class  has  0  missing values
```

P2.2

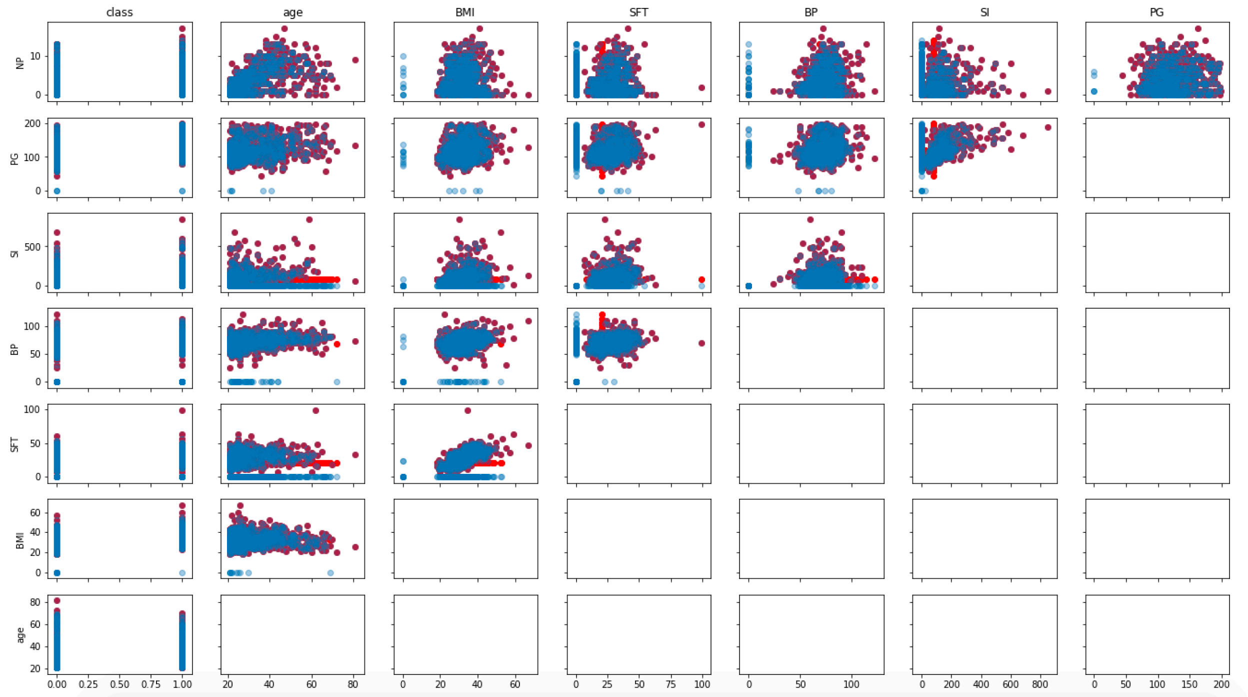


```
Summary Stats for Listwise Deletion
           min        max        median      mean       SD
NP:      0.000000  17.000000   2.000000   3.316883   3.227858
PG:     56.000000 198.000000 119.000000 122.415584  30.738930
SI:     14.000000 846.000000 125.000000 155.062338 115.503016
BP:     24.000000 110.000000  70.000000  70.587013  12.551332
SFT:     7.000000  63.000000  29.000000  29.046753  10.546269
BMI:    18.200000  67.100000  33.200000  33.028052   7.030213
age:    21.000000  81.000000  27.000000  30.820779  10.250959
Class:   0.000000   1.000000   0.000000   0.332468   0.471098
Pearson Correlation Table
       class     age      BMI      SFT      BP       SI       PG        NP
NP:  0.263798 0.684259 -0.021315 0.096771 0.214670 0.078372 0.199118 1.000000
PG:  0.527659 0.346806 0.216663 0.198923 0.213467 0.574398 1.000000
SI:  0.325505 0.230133 0.230058 0.176935 0.102461 1.000000
BP:  0.200689 0.302602 0.302909 0.229452 1.000000
SFT: 0.265587 0.170516 0.661786 1.000000
BMI: 0.282027 0.073750 1.000000
age: 0.351186 1.000000
class: 1.000000
```

Now, we see values we can expect as the minimums for all the columns. We also see there was a huge jump in the means and medians of SI and SFT. This is because of all of the missing values stored as zeros weighing down the original mean.

P3.1

```
Summary Stats for Mean Imputation
           min        max        median       mean         SD
NP:      0.000000  17.000000   3.000000   3.864000   3.380755
PG:     44.000000 199.000000 117.000000 121.586542  30.466533
SI:     14.000000 846.000000  79.748000 118.346032  90.965418
BP:     24.000000 122.000000  72.000000  72.104169  12.123450
SFT:     7.000000  99.000000  23.000000  26.514432   9.636343
BMI:    18.200000  67.100000  32.021200  32.448149   6.881717
age:    21.000000  81.000000  29.000000  33.128000  11.747778
Class:   0.000000   1.000000   0.000000   0.349333   0.476759
Pearson Correlation Table
       class      age      BMI      SFT       BP       SI        PG        NP
NP:  0.227184 0.547518 0.026816 0.022985 0.208884 -0.019458 0.131743 1.000000
PG:  0.500453 0.270801 0.232696 0.155555 0.221848 0.389313 1.000000
SI:  0.195120 0.048137 0.189569 0.238418 0.015734 1.000000
BP:  0.162385 0.319758 0.283647 0.132487 1.000000
SFT: 0.181133 0.034214 0.529711 1.000000
BMI: 0.320969 0.033654 1.000000
age: 0.235074 1.000000
class: 1.000000
```



For my graphs I plotted the mean imputed data in red first, followed by the raw data in blue on top. I set alpha=.4 so you can see purple dots where the points do not change, semi-transparent blue dots that have missing values, and red dots where the mean was inserted for the missing value. In theory, these would be the only meaning of the colors, but since the graphs are small there are many dots piled on top of one another, so a big blue cluster forms where the data is centered. The semi-transparent blue dots (missing values) form a line at x=0 or y=0, and the red dots (imputed values) form a line at x=mean(x) or y=mean(y). Mean imputation gave us a bit smaller correlations overall compared to the correlations in the List-wise Deletion. The only large change in means/SDs was in the SI column, as it was the column with the most missing values.