# IC272-ASSIGNMENT2

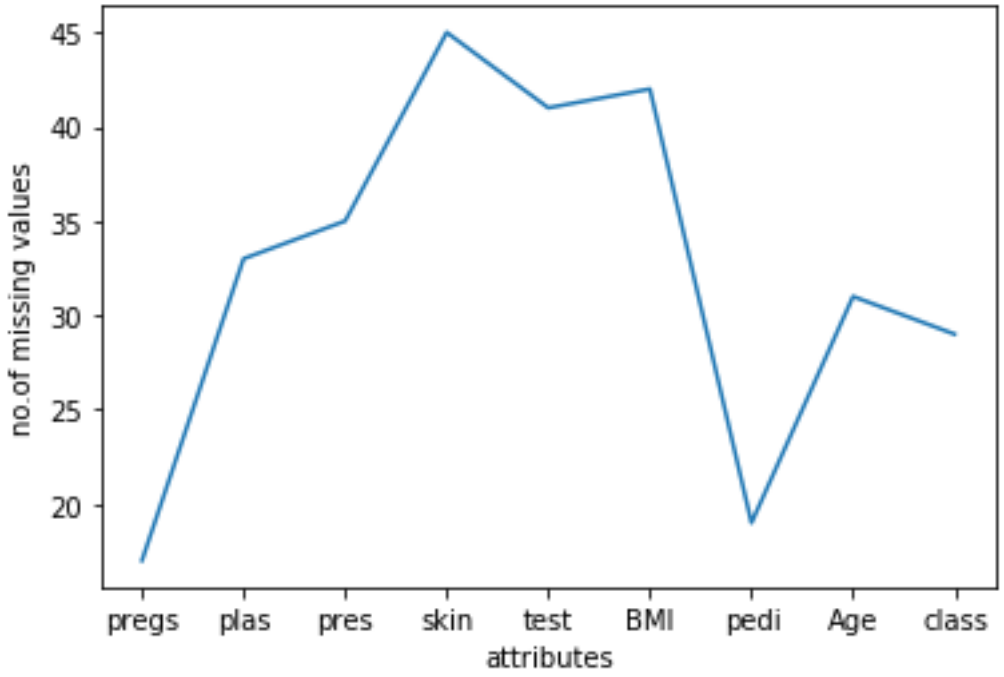
## ROLLNO: B19084

## NAME: SRISHTI GINJALA

## PHONE: 9440000900

## QUESTION1)

The number of missing values in all attributes are shown in this graph.




Hence, we conclude that skin has the highest number of missing values and Number of Pregnancies (pregs) the lowest.

## QUESTION2)

Now we delete the tuples having equal to or more than one third of attributes with missing values.The row numbers of the deleted tuples are:[1, 39, 40, 53, 54, 83, 89, 103, 125, 136, 145, 210, 211, 212, 213, 249, 250, 254, 280, 281, 284, 314, 321, 335, 429, 430, 449, 450, 451, 471, 472, 473, 474, 718, 719, 720, 721, 753, 766]. Number of tuples deleted are 39.

We also delete tuples (rows) having missing value in the target (class) attribute. The row numbers of the tuples having missing value in the target (class) attribute are:

[1, 8, 13, 28, 29, 35, 54, 62, 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 249, 250, 254, 280, 281, 284, 308, 746, 748]

Number of tuples deleted are 29.

## QUESTION3)

The number of missing values in attributes are

pregs 0

plas 12

pres 9

skin 8

test 8

BMI 12

pedi 2

Age 18

class 0

The total number of missing values in the file are 69. (after the deletion of tuples).

## QUESTION4)

a) Replacing the missing values by mean of their respective attribute, we get

| | pregs | plas | pres | skin | test | bmi | pedi | age | class |
|---|---|---|---|---|---|---|---|---|---|
| mean | 3.886 | 120.67 | 69.00 | 20.35 | 77.83 | 32.01 | 0.476 | 33.1 | 0.343 |
| median | 3.000 | 118.0 | 72.00 | 23.00 | 36 | 32.06 | 0.382 | 29 | 0 |
| mode | 1.0 | 99,100 | 70.00 | 0 | 0 | 32.00 | 0.254 | 22 | 0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| std | 3.374 | 30.99 | 19.69 | 15.95 | 115.2 | 7.765 | 0.333 | 11.52 | 0.475 |
| rmse | 0 | 42.65 | 8.944 | 15.89 | 55.23 | 10.46 | 0.04 | 15.36 | 0 |

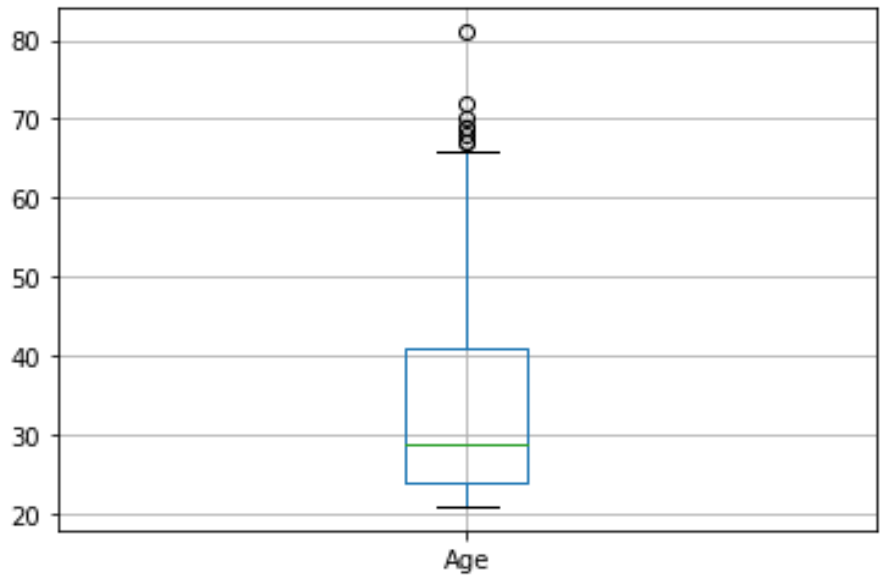
b) Replacing the missing values by interpolation, we get

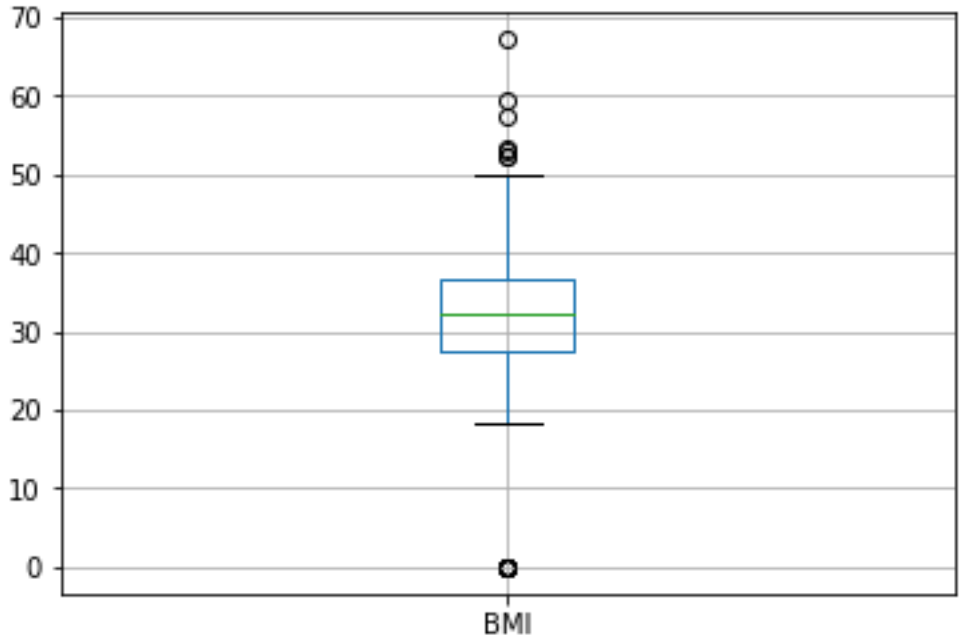
| | pregs | plas | pres | skin | test | bmi | pedi | age | class |
|---|---|---|---|---|---|---|---|---|---|
| mean | 3.886 | 120.35 | 69.10 | 20.39 | 77.35 | 32.05 | 0.477 | 33.21 | 0.343 |
| median | 3.000 | 117.0 | 72.00 | 23.00 | 27 | 32.25 | 0.382 | 29 | 0 |
| mode | 1.0 | 99,100 | 70.00 | 0 | 0 | 32.00 | 0.258 | 22 | 0 |
| std | 3.374 | 31.27 | 19.73 | 15.97 | 110.75 | 7.793 | 0.334 | 11.65 | 0.475 |
| rmse | 0 | 57.05 | 13.77 | 14.87 | 68.98 | 12.82 | 0.5 | 17.39 | 0 |

## QUESTION5)

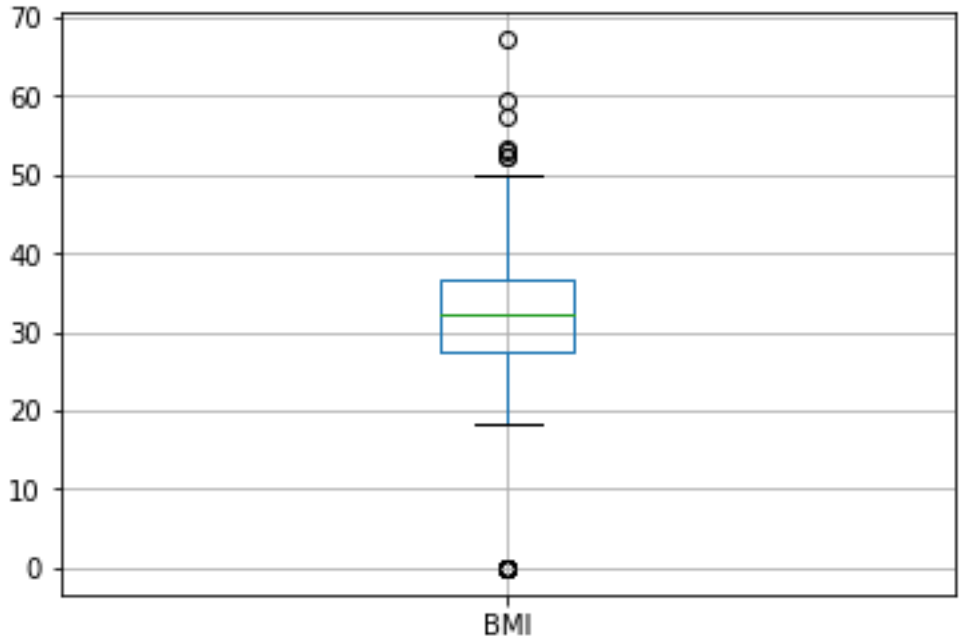
The outliers in Age are [69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0]

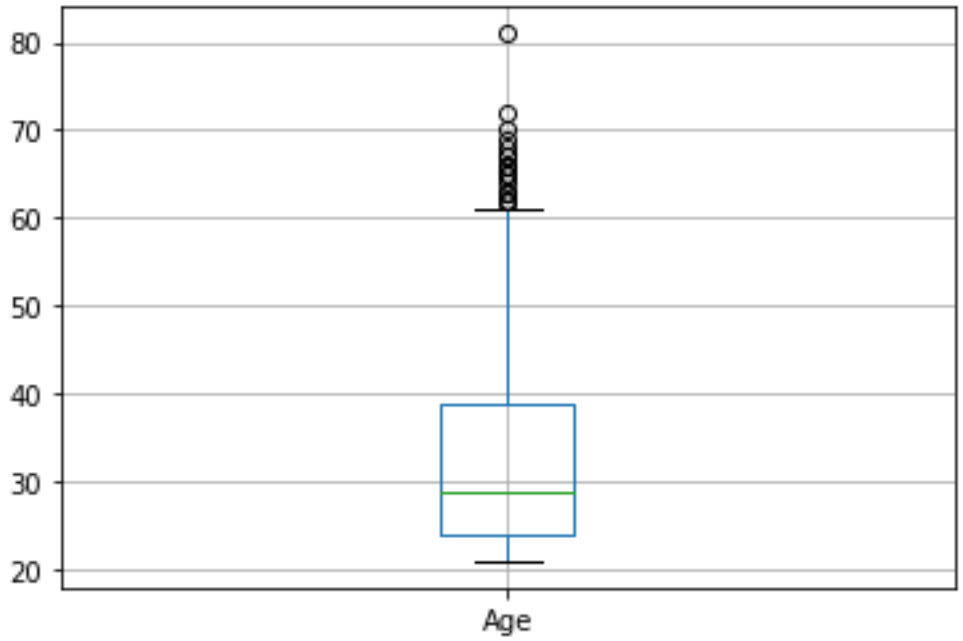
The outliers in BMI are [0.0, 0.0, 0.0, 53.2, 67.1, 52.3, 52.3, 52.9, 0.0, 0.0, 59.4, 0.0, 0.0, 57.3, 0.0, 0.0]






The above figures are the boxplots for Age and BMI. We observe that there are many outliers. After replacing the outliers with median, we get the following boxplots.






We get outliers even after replacing with median because the quartiles change as the numbers get rearranged and there are again chances of outliers.

# Thank you