IC272: Lab 2
Data Cleaning
Handling Missing Values and Outlier Analysis

Anmol Bishnoi
Group 01
IIT Mandi, b19069@students.iitmandi.ac.in
7042845211

QUESTION 1

Plot a graph of the attribute names (x-axis) with the number of missing values in them (y-axis).
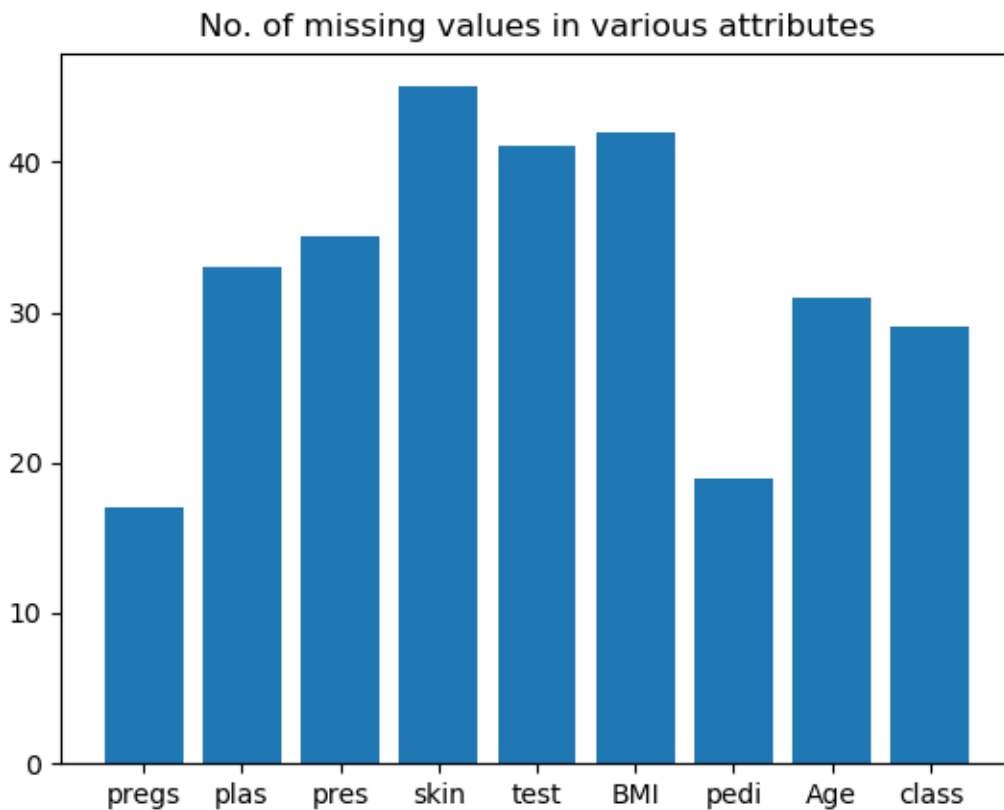

Fig 1

The plot shows the number of missing values in every individual attribute. For a database with around 800 tuples, this isn't that bad.

QUESTION 2A

Delete (drop) the tuples (rows) having equal to or more than one third of attributes with missing values. Print the total number of tuples deleted and also print the row numbers of the deleted tuples.

```
39 tuples were deleted
The row number for the tuples deleted 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]
```

Fig 2

QUESTION 2B

Drop the tuples (rows) having missing value in the target (class) attribute. Print the total number of tuples deleted and also print the row numbers of the deleted tuples

```
21 tuples were deleted
The row number for the tuples deleted are [8, 13, 28, 29, 35, 62,
 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 308, 7
46, 748]
```

Fig 3

QUESTION 3

After step 2, count and print the number of missing values in each attributes. Also find and print the total number of missing values in the file (after the deletion of tuples).

```
The number missing values in the different attributes are:
pregs - 0
plas - 12
pres - 9
skin - 8
BMI - 12
pedi - 2
Age - 18
class - 0
The total number of missing values in the file are 69
```

Fig 4

QUESTION 4A

Replace the missing values by mean of their respective attribute.

    i.   **Compute the mean, median, mode and standard deviation for each attribute and compare the same with that of the original file.**

   ii.   **Calculate the root mean square error (RMSE) between the original and replaced values for each attribute. Plot these RMSE with respect to the attributes.**

```
Original
                    Attributes        Mean     Median          Mode  Standard Deviation
0                  Pregnancies    3.845052     3.0000           [1]            3.367384
1                      Glucose  120.894531   117.0000     [99, 100]           31.951796
2               Blood Pressure   69.105469    72.0000          [70]           19.343202
3               Skin Thickness   20.536458    23.0000           [0]           15.941829
4                      Insulin   79.799479    30.5000           [0]          115.168949
5                          BMI   31.992578    32.0000        [32.0]            7.879026
6   Diabetes Pedigree Function    0.471876     0.3725  [0.254, 0.258]          0.331113
7                          Age   33.240885    29.0000          [22]           11.752573
8                      Outcome    0.348958     0.0000           [0]            0.476641
After substituting with mean
                    Attributes        Mean     Median          Mode  Standard Deviation
0                  Pregnancies    3.885593   3.000000         [1.0]            3.371477
1                      Glucose  120.666667 118.000000  [99.0, 100.0]          30.968288
2               Blood Pressure   69.001431  72.000000        [70.0]           19.677448
3               Skin Thickness   20.348571  23.000000         [0.0]           15.934938
4                      Insulin   77.814286  36.000000         [0.0]          110.529464
5                          BMI   32.009339  32.009339        [32.0]            7.759269
6   Diabetes Pedigree Function    0.476042   0.382500  [0.254, 0.258]          0.332964
7                          Age   33.094203  29.000000        [22.0]           11.511532
8                      Outcome    0.343220   0.000000         [0.0]            0.474784
```

Fig 5

The main advantage of substituting with mean is that the values of the central tendencies does not change after addition. We see a change in the values Mean, Median and Standard Deviation because we were asked to do calculations on the original data, before the deletion of rows.

The value of Mode more or less remains unchanged as not many tuples with the Mode were removed. Had that been the case, the values for Mode would've been found to be starkly different too.
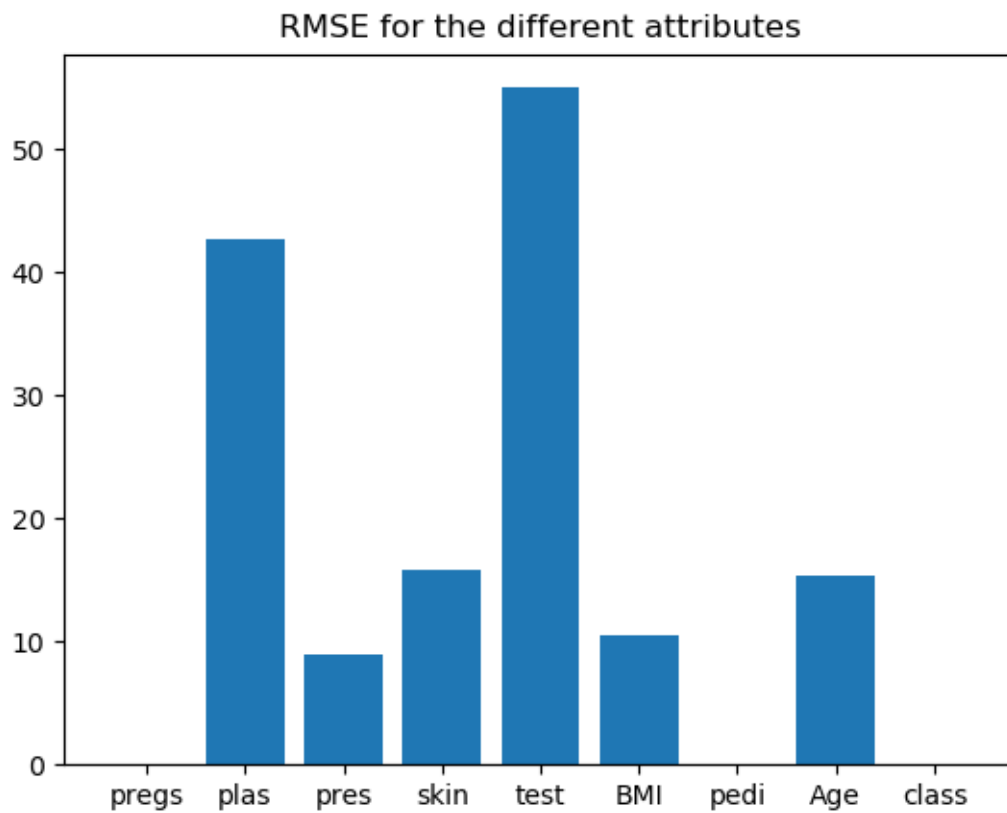
RMSE for the different attributes

Fig 6

We can observe that the values we put in were very different from the actual values atleast for the attributes test and plas. This is one of the downfalls of simply substituting with mean as it is pretty unlikely that mean is also what the actual/predicted value is.

QUESTION 4B

Replace the missing values in each attribute using linear interpolation technique.

    i.   **Compute the mean, median, mode and standard deviation for each attribute and compare the same with that of the original file.**

   ii.   **Calculate the root mean square error (RMSE) between the original and replaced values for each attribute. Plot these RMSE with respect to the attributes.**

```
Original
                        Attributes        Mean    Median          Mode  Standard Deviation
0                      Pregnancies    3.845052    3.0000           [1]            3.367384
1                          Glucose  120.894531  117.0000     [99, 100]           31.951796
2                   Blood Pressure   69.105469   72.0000          [70]           19.343202
3                   Skin Thickness   20.536458   23.0000           [0]           15.941829
4                          Insulin   79.799479   30.5000           [0]          115.168949
5                              BMI   31.992578   32.0000        [32.0]            7.879026
6       Diabetes Pedigree Function    0.471876    0.3725  [0.254, 0.258]           0.331113
7                              Age   33.240885   29.0000          [22]           11.752573
8                          Outcome    0.348958    0.0000           [0]            0.476641
After Interpolation
                        Attributes        Mean    Median          Mode  Standard Deviation
0                      Pregnancies    3.885593    3.0000         [1.0]            3.371477
1                          Glucose  120.349576  117.0000  [99.0, 100.0]          31.252704
2                   Blood Pressure   69.109463   72.0000        [70.0]           19.722043
3                   Skin Thickness   20.392655   23.0000         [0.0]           15.964563
4                          Insulin   77.355226   27.0000         [0.0]          110.677746
5                              BMI   32.046328   32.2500        [32.0]            7.787110
6       Diabetes Pedigree Function    0.477325    0.3825  [0.254, 0.258]           0.334012
7                              Age   33.216102   29.0000        [22.0]           11.644416
8                          Outcome    0.343220    0.0000         [0.0]            0.474784
```

Fig 7

All values change just as we expected them to. The value of Mode does not change as it very unlikely that introducing a few interpolated value will cause the number of instances of some other individual value to increase by a huge margin.
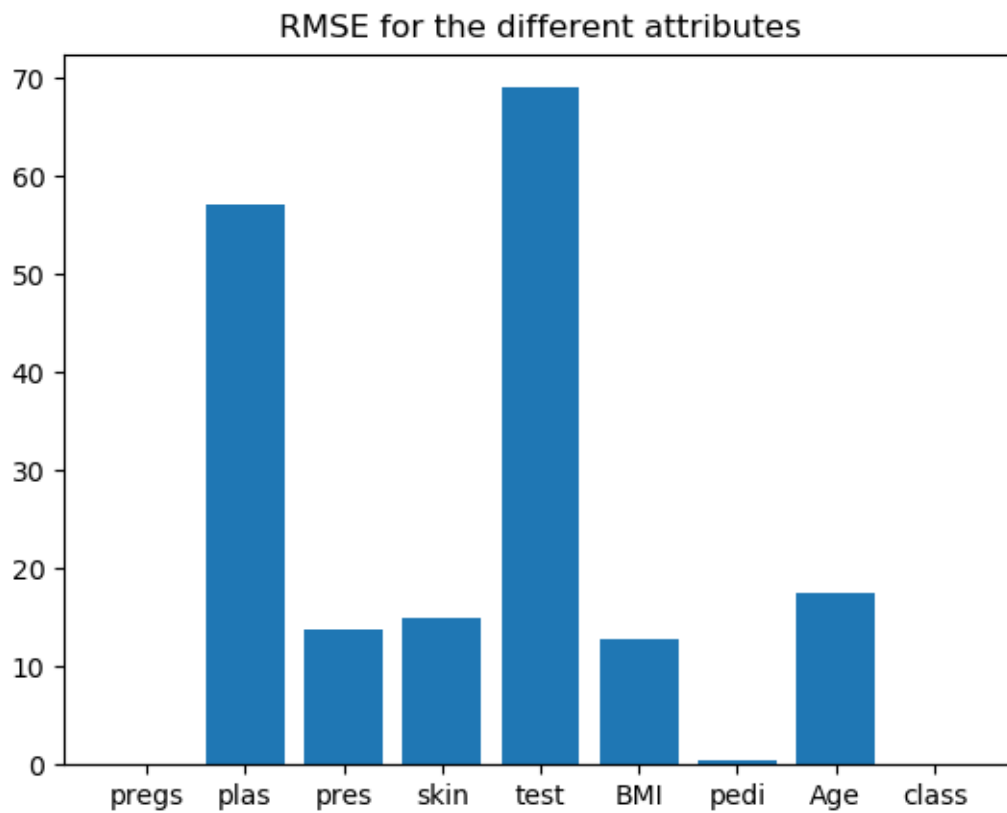
Fig 8

The RMSE is again very high for test and plas. Linear Interpolation works and gives more accurate values when the data is sequential but since here the data did not really have any particular sequence, using Linear Interpolation was as good as any other method and as observed even worse than the others.

QUESTION 5A

After replacing the missing values by interpolation method, find the outliers in the attributes "Age" and "BMI". Outliers are the values that does not satisfy the condition (Q1 – (1.5 * IQR)) < X < (Q3 + (1.5 * IQR)), where X is the value of the attribute, IQR is the inter quartile range, Q1 and Q3 are the first and third quartiles. Obtain the boxplot for these attributes.

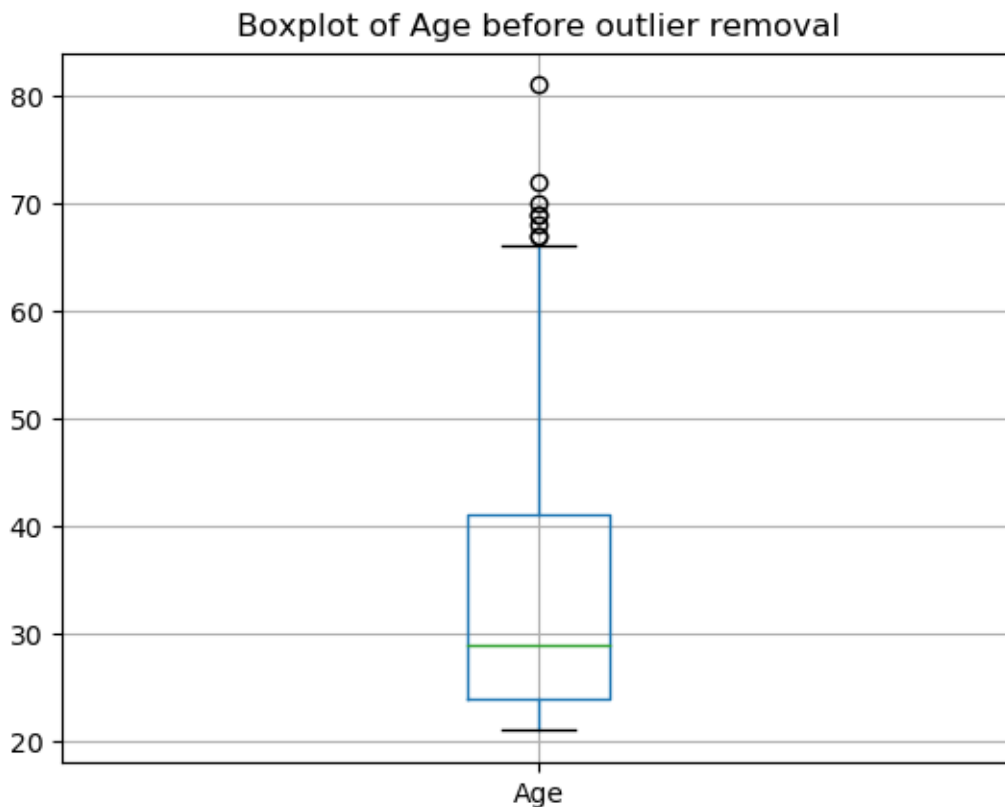The outliers for age are [69., 67., 72., 81., 67., 70., 68., 69.].

Fig 9

The outliers for BMI are [0. , 0. , 0. , 0. , 0. , 0. , 0. , 0. , 0. , 53.2, 67.1 , 52.3, 52.3, 52.9, 59.4, 57.3]
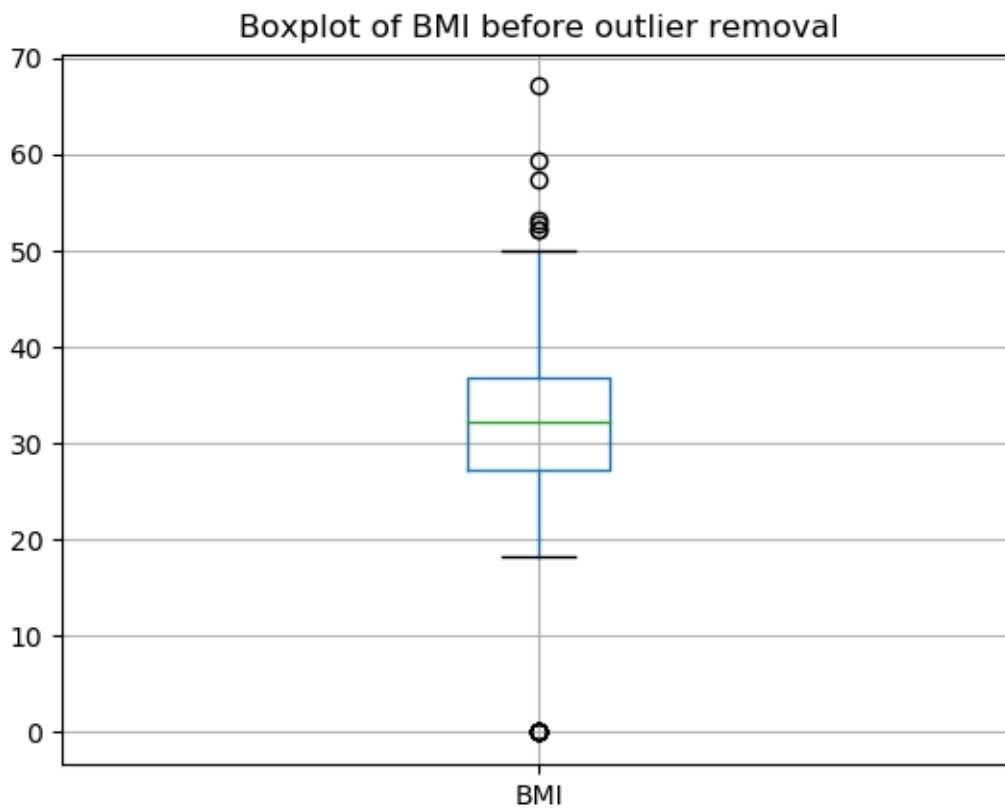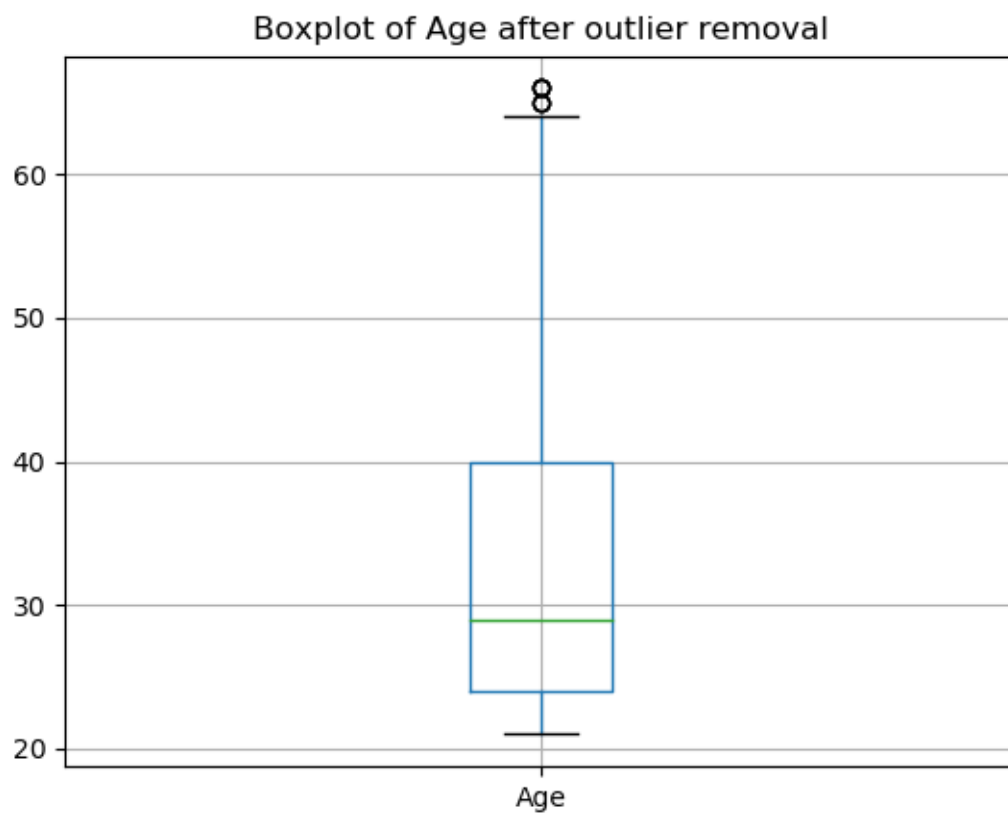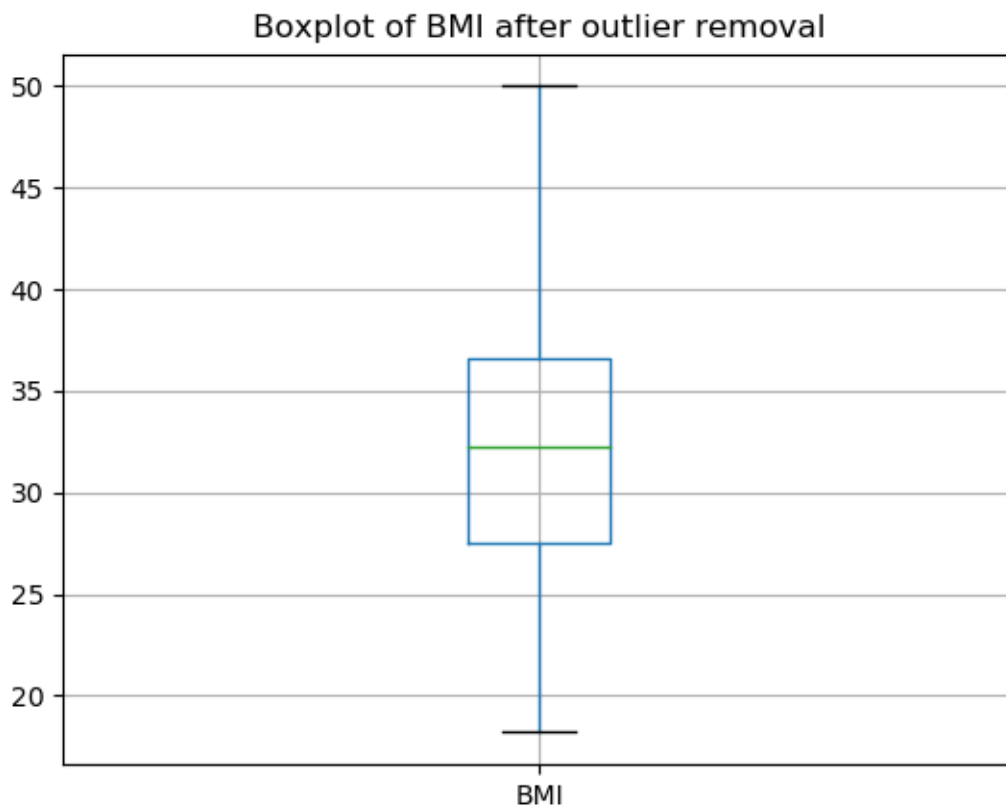
Fig 10

QUESTION 5B

Replace these outliers by the median of the attribute. Plot the boxplot again and observe the difference with that of the boxplot in (5i). Do you still get outliers? Why?Plot the histogram of attribute 'rain' for each of the 10 stations

Boxplot of Age after outlier removal

Boxplot of BMI after outlier removal

We still do get outliers as the overall data has now completely changed and hence the Median and the IQR is different. So, is the boundary for upper whisker and lower whisker and this causes newer extreme values to show up as our outliers.