Assignment -2

Data Science – III

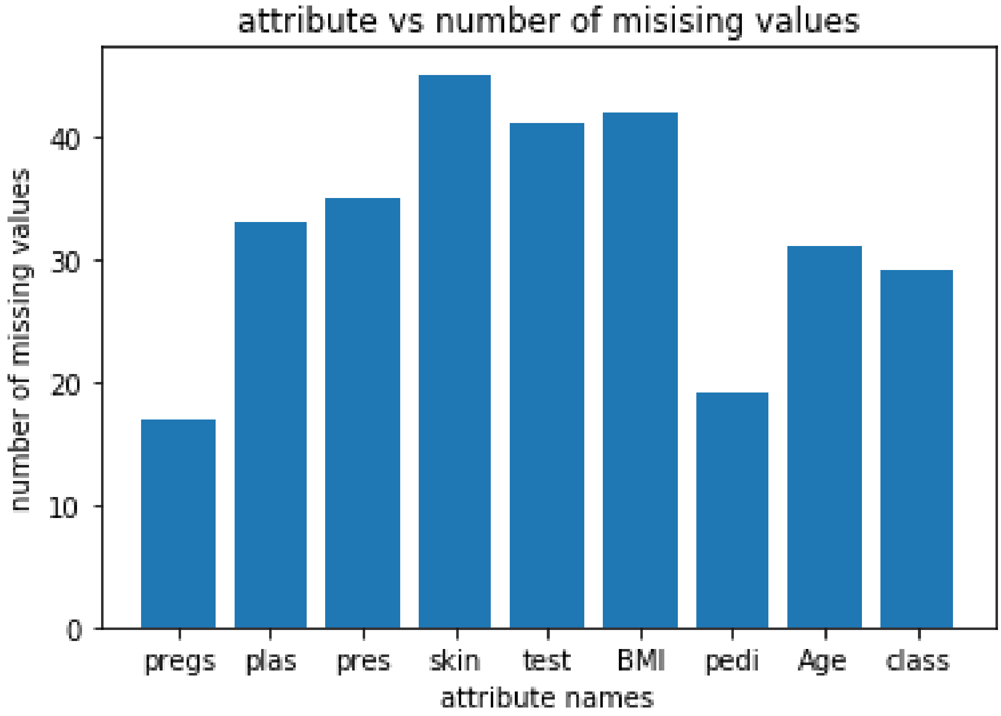
Name – Sumit Kumar

Roll no – B19118

Mobile No – 7549233722

Q1.

Bar graph of the attributes and the number of the missing values is given below:



- The attribute 'skin' contains maximum number of missing values whereas the attribute 'pregs' contains minimum number of missing values.

Q2.(a)

Total number of tuples deleted having equal to or more than one third of the attributes with the missing values is :- **39.**

The row numbers of the deleted tuples are :-

2, 40, 41, 54, 55, 84, 90, 104, 126, 137, 146, 211, 212, 213, 214, 250, 251, 255, 281, 282, 285, 315, 322, 336, 430, 431, 450, 451, 452, 472, 473, 474, 475, 719, 720, 721, 722, 754, 767

Q2.(b)

Total number of tuples deleted having class attribute containing the missing values is :- **21.**

The row numbers of the deleted tuples are :-

9, 14, 29, 30, 36, 63, 93, 96, 108, 111, 131, 132, 133, 134, 150, 183, 189, 219, 309, 747, 749

Q3.

The number of missing values in attribute pregs is :- 0

The number of missing values in attribute plas is :- 12

The number of missing values in attribute pres is :- 9

The number of missing values in attribute skin is :- 8

The number of missing values in attribute test is :- 8

The number of missing values in attribute BMI is :- 12

The number of missing values in attribute pedi is :- 2

The number of missing values in attribute Age is :- 18

The number of missing values in attribute class is :- 0

Total number of missing value remaining in the file is :- **69**

CONCLUSION:-

- There is no any missing value left for the attribute **pregs** and **class** is zero.
- The attribute **Age** contains maximum 18 missing values.
- After performing step 2, the data is still not cleaned as there are still many missing values.

Q4.(a).(i)

After filling the missing values with the mean of the respective attributes: ---

For attribute pregs (after replacing the missing values):-

Mean is :- 3.885593220338983
Mode is :- 1.0
Median is :- 3.0
Standard deviation is :- 3.373860130387226

For attribute pregs (from original data) :-
Mean is :- 3.8450520833333335
Mode is :- 1

```
Median is :-  3.0
Standard deviation is :-  3.3695780626988623


For attribute plas (after replacing the missing values):-
Mean is :-  120.66666666666667
Mode is :-  99.0
Median is :-  117.0
Standard deviation is :-  31.25657724548738


For attribute plas (from original data) :-
Mean is :-  120.89453125
Mode is :-  99
Median is :-  117.0
Standard deviation is :-  31.97261819513622


For attribute pres (after replacing the missing values):-
Mean is :-  69.00143061516452
Mode is :-  70.0
Median is :-  72.0
Standard deviation is :-  19.817903058043974


For attribute pres (from original data) :-
Mean is :-  69.10546875
Mode is :-  70
Median is :-  72.0
Standard deviation is :-  19.355807170644777


For attribute skin (after replacing the missing values) :-
Mean is :-  20.34857142857143
Mode is :-  0.0
Median is :-  23.0
Standard deviation is :-  16.037195168032092


For attribute skin (from original data) :-
Mean is :-  20.536458333333332
Mode is :-  0
Median is :-  23.0
Standard deviation is :-  15.952217567727677
```

```
For attribute test (after replacing the missing values):-
Mean is :-  77.81428571428572
Mode is :-  0.0
Median is :-  27.0
Standard deviation is :-  111.23875170759175


For attribute test (from original data) :-
Mean is :-  79.79947916666667
Mode is :-  0
Median is :-  30.5
Standard deviation is :-  115.24400235133837


For attribute BMI (after replacing the missing values):-
Mean is :-  32.00933908045977
Mode is :-  32.0
Median is :-  32.05
Standard deviation is :-  7.831501532477941

For attribute BMI (from original data) :-
Mean is :-  31.992578124999977
Mode is :-  32.0
Median is :-  32.0
Standard deviation is :-  7.8841603203754405


For attribute pedi (after replacing the missing values):-

Mean is :-  0.4760424929178469
Mode is :-  0.254
Median is :-  0.3815
Standard deviation is :-  0.33367122461347043


For attribute pedi (from original data) :-
Mean is :-  0.4718763020833327
Mode is :-  0.254
Median is :-  0.3725
Standard deviation is :-  0.33132859501277484
```

For attribute Age (after replacing the missing values) :-
Mean is :- 33.094202898550726
Mode is :- 22.0
Median is :- 29.0
Standard deviation is :- 11.669174457848875

For attribute Age (from original data) :-
Mean is :- 33.240885416666664
Mode is :- 22
Median is :- 29.0
Standard deviation is :- 11.76023154067868

For attribute class (after replacing the missing values):-
Mean is :- 0.3432203389830508
Mode is :- 0.0
Median is :- 0.0
Standard deviation is :- 0.47511996196348166

For attribute class (from original data) :-
Mean is :- 0.3489583333333333
Mode is :- 0
Median is :- 0.0
Standard deviation is :- 0.4769513772427971

CONCLUSION:-

- After filling the missing values with the mean of the particular attributes, we found that for most of the attributes have same value of mode and median as compared from the original data.
- There is a very little difference between Mean and Standard deviation of the attributes as compared to the original one.
- For the attribute **test** almost all parameters are different as compared to the original one.
- From the above data, we can say that we can clean the missing data by replacing with the mean of their attributes because there is a very slight difference as compared to the original one.

Q4(a).(ii) :-

The RMSE value of the attribute pregs is :- 0

The RMSE value of the attribute plas is :- 42.64387412044079

The RMSE value of the attribute pres is :- 8.950321330960236

The RMSE value of the attribute skin is :- 15.839442244354593

The RMSE value of the attribute test is :- 54.969720793193346

The RMSE value of the attribute BMI is :- 10.450965534783302

The RMSE value of the attribute pedi is :- 0.046762740833851374

The RMSE value of the attribute Age is :- 15.365829400182067

The RMSE value of the attribute class is :- 0

CONCLUSION:-

- The RMSE value is basically the prediction errors of the particular attribute.
- It denotes how widely the data is dispersed around the regression line.
- The RMSE value of **pregs** and **class** is zero because there is no any missing value in these attribute.

Question 4(b).(i) :-

After filling the missing values using the interpolation method: --

For attribute pregs (after replacing the missing values):-
Mean is :- 3.885593220338983
Mode is :- 1.0
Median is :- 3.0
Standard deviation is :- 3.373860130387226

For attribute pregs (from original data) :-
Mean is :- 3.8450520833333335
Mode is :- 1
Median is :- 3.0
Standard deviation is :- 3.3695780626988623

For attribute plas (after replacing the missing values)) :-
Mean is :- 120.34957627118644
Mode is :- 99.0
Median is :- 117.0
Standard deviation is :- 31.274798286084703

For attribute plas (from original data) :-
Mean is :- 120.89453125
Mode is :- 99

```
Median is :-  117.0
Standard deviation is :-  31.97261819513622


For attribute pres (after replacing the missing values):-
Mean is :-  69.10946327683615
Mode is :-  70.0
Median is :-  72.0
Standard deviation is :-  19.735986079470695


For attribute pres (from original data) :-
Mean is :-  69.10546875
Mode is :-  70
Median is :-  72.0
Standard deviation is :-  19.355807170644777


For attribute skin (after replacing the missing values):-
Mean is :-  20.39265536723164
Mode is :-  0.0
Median is :-  23.0
Standard deviation is :-  15.975849466950478


For attribute skin (from original data) :-
Mean is :-  20.536458333333332
Mode is :-  0
Median is :-  23.0
Standard deviation is :-  15.952217567727677


For attribute test (after replacing the missing values):-
Mean is :-  77.35522598870057
Mode is :-  0.0
Median is :-  27.0
Standard deviation is :-  110.75599102677111


For attribute test (from original data) :-
Mean is :-  79.79947916666667
Mode is :-  0
Median is :-  30.5
Standard deviation is :-  115.24400235133837
```

```
For attribute BMI (after replacing the missing values):-
Mean is :-  32.04632768361581
Mode is :-  32.0
Median is :-  32.25
Standard deviation is :-  7.792615028426985


For attribute BMI (from original data) :-
Mean is :-  31.992578124999977
Mode is :-  32.0
Median is :-  32.0
Standard deviation is :-  7.8841603203754405


For attribute pedi (after replacing the missing values):-
Mean is :-  0.47732485875706193
Mode is :-  0.254
Median is :-  0.38249999999999995
Standard deviation is :-  0.33424800709049673


For attribute pedi (from original data) :-
Mean is :-  0.4718763020833327
Mode is :-  0.254
Median is :-  0.3725
Standard deviation is :-  0.33132859501277484


For attribute Age (after replacing the missing values):-
Mean is :-  33.21610169491525
Mode is :-  22.0
Median is :-  29.0
Standard deviation is :-  11.652648195089654


For attribute Age (from original data) :-
Mean is :-  33.240885416666664
Mode is :-  22
Median is :-  29.0
Standard deviation is :-  11.76023154067868
```

For attribute class (after replacing the missing values):-
Mean is :- 0.3432203389830508
Mode is :- 0.0
Median is :- 0.0
Standard deviation is :- 0.47511996196348166

For attribute class (from original data) :-
Mean is :- 0.3489583333333333
Mode is :- 0
Median is :- 0.0
Standard deviation is :- 0.4769513772427971

CONCLUSION:-

- After filling the missing values using interpolation method, we found that for most of the attributes have same value of mode and median as compared from the original data.
- There is a very little difference between Mean and Standard deviation of the attributes as compared to the original one.
- For the attribute **test** almost all parameters are different as compared to the original one
- From the above data, we can say that we can clean the missing data by replacing with the mean of their attributes because there is a very slight difference as compared to the original one.
- But for attribute **test** the values are different from the original one but if we compare with the mean method then we can see that in this case data is more close to the original one. So interpolation is good way for data cleaning here.

Q4(b).(ii).

The RMSE value of the attribute pregs is :- 0

The RMSE value of the attribute plas is :- 57.055832791709875

The RMSE value of the attribute pres is :- 13.771347065556077

The RMSE value of the attribute skin is :- 14.875828641718678

The RMSE value of the attribute test is :- 68.98482623012107

The RMSE value of the attribute BMI is :- 12.819238291348297

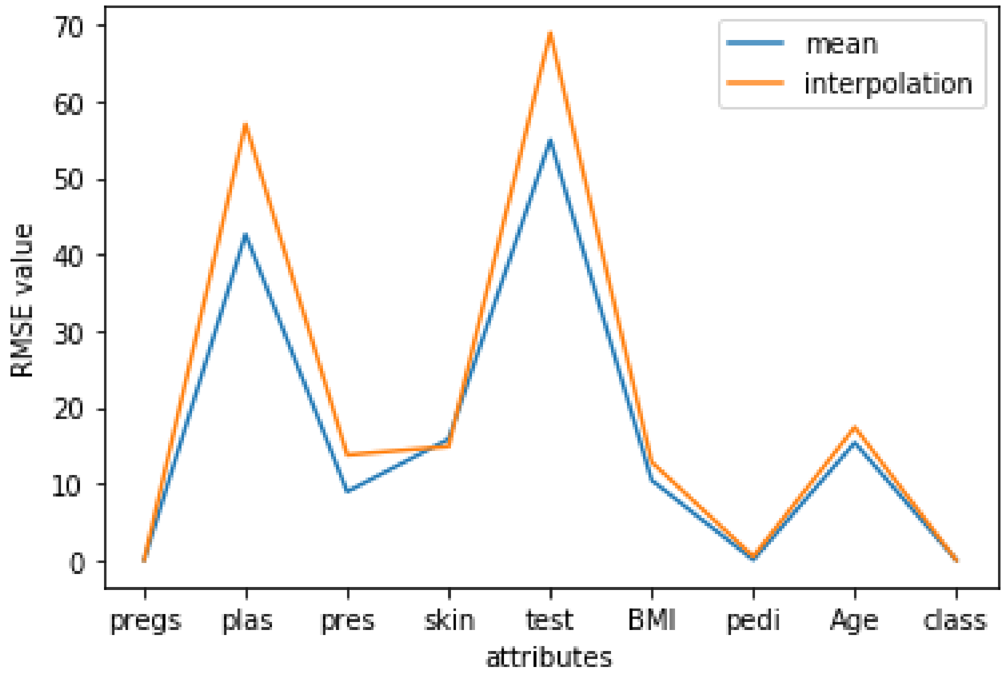
The RMSE value of the attribute pedi is :- 0.5085297434762297

The RMSE value of the attribute Age is :- 17.399712641305314

The RMSE value of the attribute class is :- 0

CONCLUSION:-

- The RMSE value of **pregs** and **class** is zero because there is no any missing value in these attribute.



- For all the attributes we found that the RMSE value as found by replacing the missing values by the mean of the attributes is low as compared to that using the interpolation method.
- So, We can conclude that replacing the missing values by their mean is the most suitable method here because the root mean square error is least in this case.
- The graph of the interpolation method is lying above the graph of mean, therefore errors in interpolation is high.

Q5(i).

After replacing the missing values using the interpolation method:-

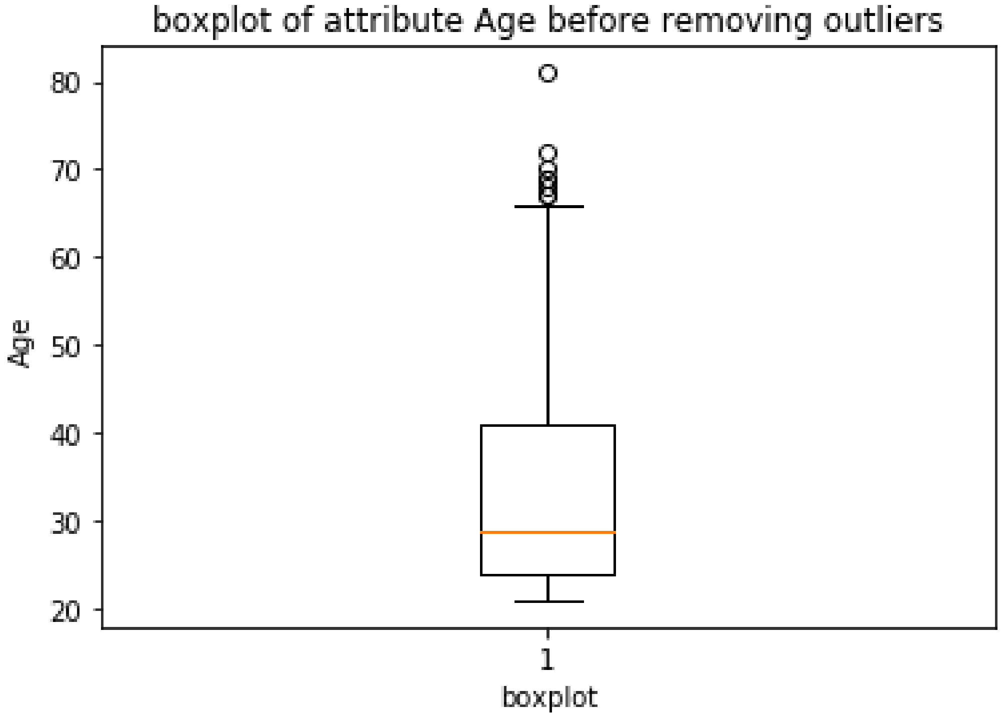
The quartile Q1 of the attribute **Age** is :- **24.0**

The quartile Q3 of the attribute **Age** is :- **41.0**

The interquartile range Q3-Q1 of the attribute **Age** is :- **17.0**

The Outliers of **Age** are :-

69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0



CONCLUSION:-

- The outliers are those values which do not satisfy the condition **(Q1-1.5*IQR) < X < (Q3+1.5*IQR).**
- There are eight outliers in the attribute **Age.** These are the values which differs significantly from the other values.
- The red line is representing the value of median which is close to the first quartile.
- The outliers are mainly in the range 65-80.

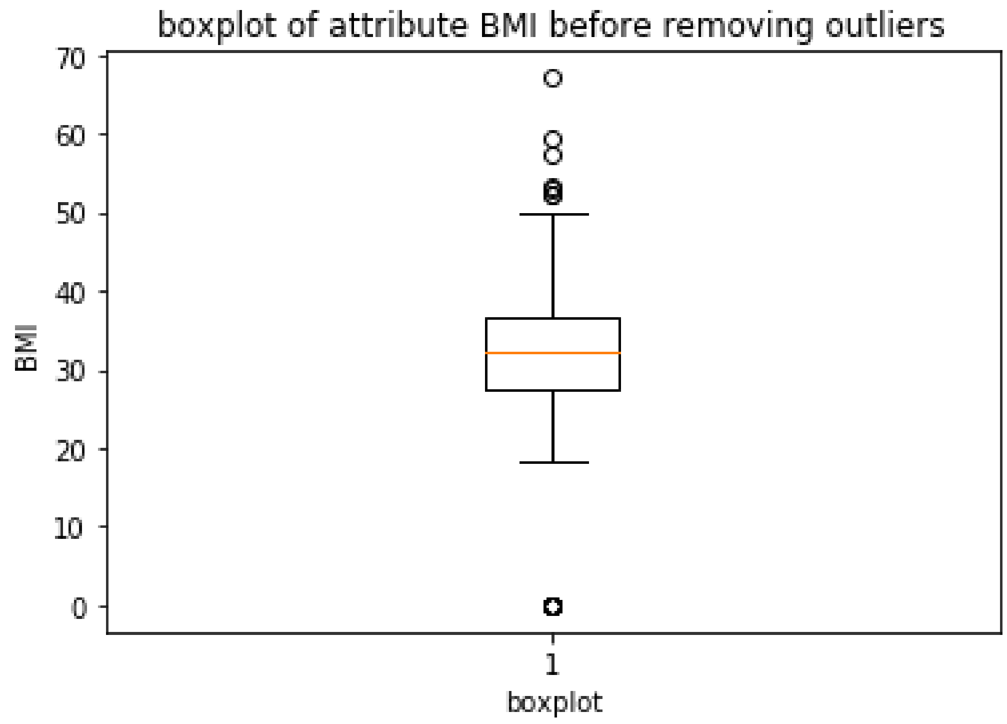
The quartile Q1 of the attribute **BMI** is :- **27.3**

The quartile Q3 of the attribute **BMI** is :- **36.8**

The interquartile range Q3-Q1 of the attribute **BMI** is :- **9.499999999999996**

The Outliers of **BMI** are :-

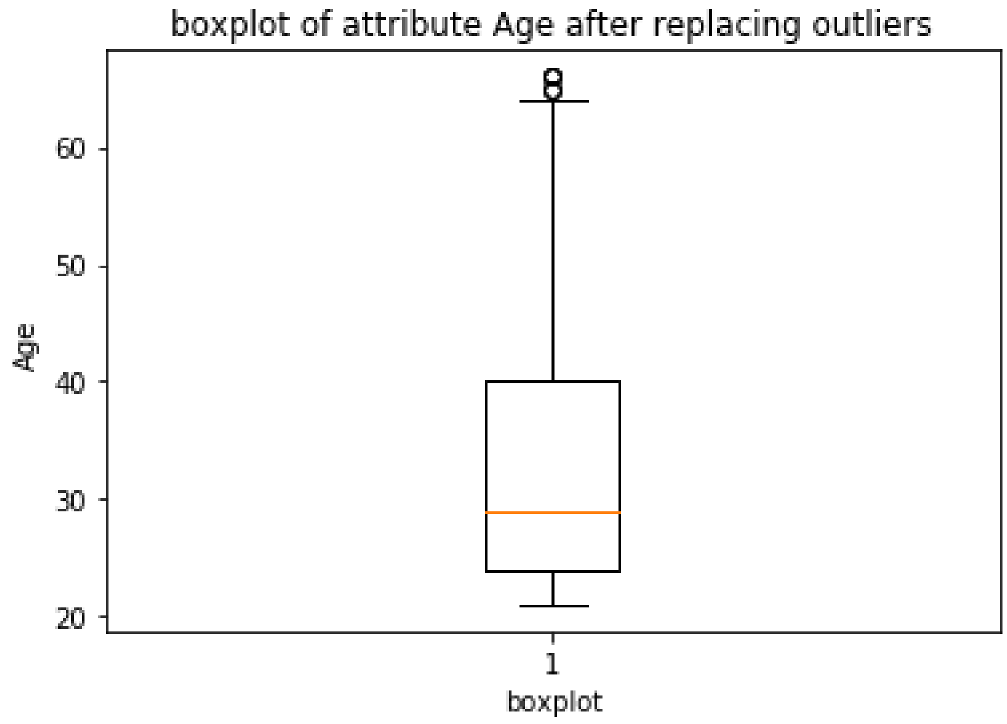
0.0, 0.0, 0.0, 53.2, 67.1, 52.3, 52.3, 52.9, 0.0, 0.0, 59.4, 0.0, 0.0, 57.3, 0.0, 0.0



CONCLUSION:-

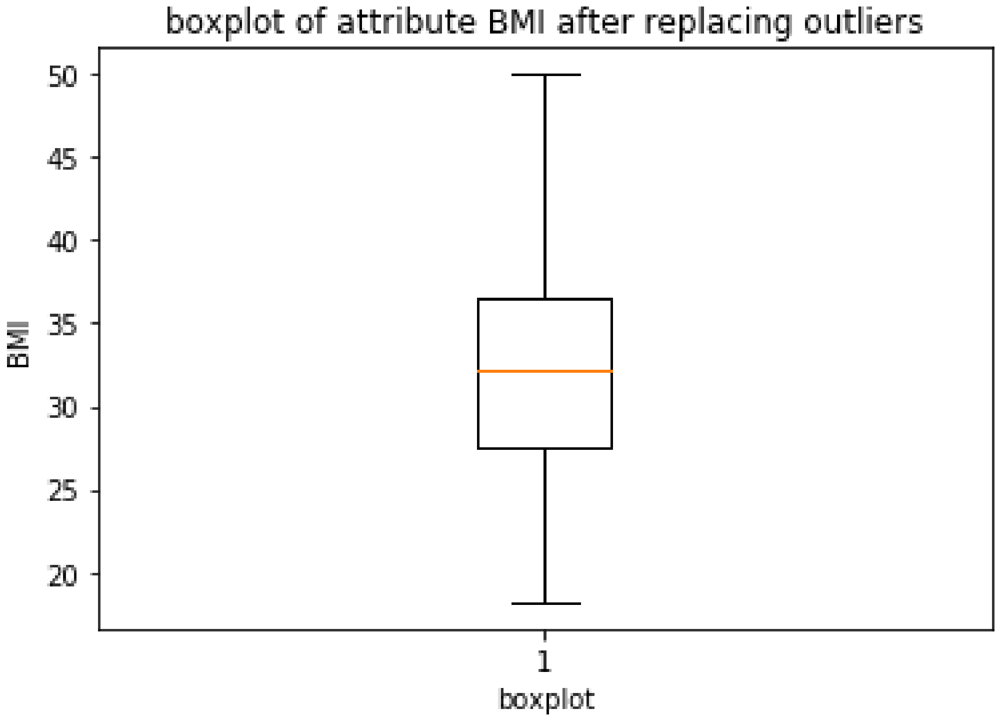
- There are 16 outliers in the attribute BMI.
- The first quartile, median and third quartile are uniformly distributed in the boxplot.
- The outliers are mainly in the range 0 and 50-60.

Q5(ii).



CONCLUSION:-

- The outliers are replaced by the median of the attributes but there are still 6-7 outliers present in the boxplot.
- After replacing the value of outliers the value of Q1, Q3 and IQR also changes so there are still many data points in the attribute Age which satisfy the condition of outliers. Therefore we are still getting outliers.
- There are more values of Age which is around 65-80 because the outliers are in this range and after replacing the outliers the new outliers are also in the range of 65-80, which states that there are good number of values of age in this range.
- The median is very less affected after replacing the outliers.



CONCLUSION:-

- In this case after replacing the outliers, we find that there are no outliers left.
- Earlier the outliers are mainly 0 and in the range 50-60.
- We conclude that after replacing the outliers most of the values lie around the median and no value is satisfying the condition of outliers.