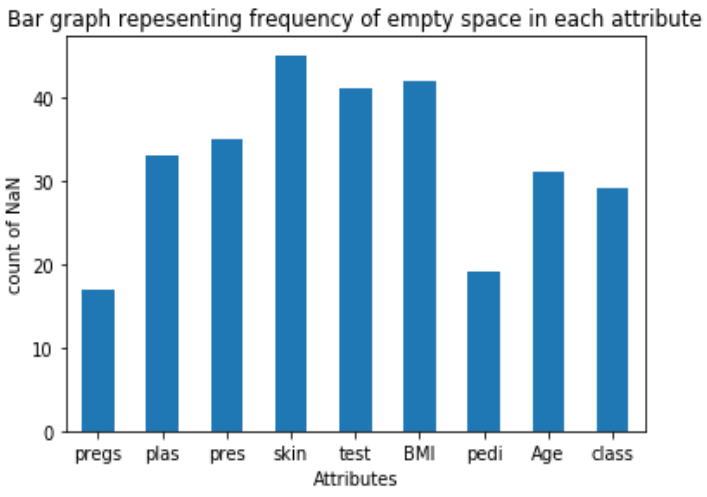
IC272

Data Science III

Lab Report

Topic : Data Cleaning – Handling Missing Values and Outlier Analyses

1. Bar graph of the attributes with their total number of missing values:



2. The row index of all the delete tuple is:
   (a) Tuple having more than or equal to one third of attributes containing missing values, the index with respect to the given csv file is:

[3, 41, 42, 55, 56, 85, 91, 105, 127, 138, 147, 212, 213, 214, 215, 251, 252, 256, 282, 283, 286, 316, 323, 337, 431, 432, 451, 452, 453, 473, 474, 475, 476, 720, 721, 722, 723, 755, 768]

Total number of deleted tuples is: 39

(b) Index of tuple having missing values in class attribute with respect to the given csv file is:

[10, 15, 30, 31, 37, 64, 94, 97, 109, 112, 132, 133, 134, 135, 151, 184, 190, 220, 310, 748, 750]

Total number of deleted tuples having missing values in class attribute is: 21

3. After step 2 the number of missing values in each attribute is given as:

pregs 0
plas 12
pres 9
skin 8
test 8
BMI 12
pedi 2
Age 18
class 0
Total number of missing values in the file (after the deletion of tuples) is: 69

4. (a)(i) After replacing the missing values of every attributes with the mean of it , Comparison of mean, median, mode and standard deviation attributes with the original data

Mean of all the attributes of missing data using mean method is
pregs 3.885593
plas 120.666667
pres 69.001431
skin 20.348571
test 77.814286
BMI 32.009339
pedi 0.476042
Age 33.094203
class 0.343220
Mean of all the attributes of original data is
pregs 3.845052
plas 120.894531
pres 69.105469
skin 20.536458
test 79.799479
BMI 31.992578
pedi 0.471876
Age 33.240885

class 0.348958

Inference: The value of mean is approximately equal with the original mean values of all the attributes as most of the changes happen after decimal.

Median of all the attributes of missing data using mean method is
pregs 3.000000
plas 118.000000
pres 72.000000
skin 23.000000
test 36.000000
BMI 32.009339
pedi 0.382500
Age 29.000000
class 0.000000
Median of all the attributes of original data is
pregs 3.0000
plas 117.0000
pres 72.0000
skin 23.0000
test 30.5000
BMI 32.0000
pedi 0.3725
Age 29.0000
class 0.0000
Inference: The value of median is approximately equal with the original median values of all the attributes and also for some attributes it's totally equal to the original median.

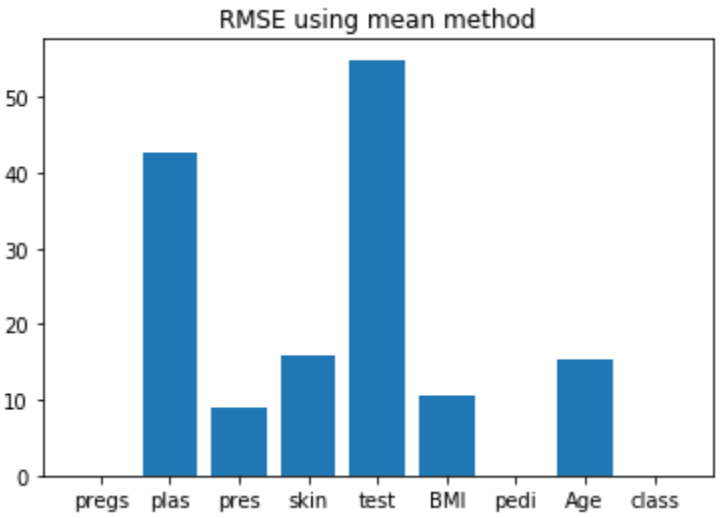
Mode of all the attributes of missing data using mean method is
pregs plas pres skin test BMI pedi Age class
1.0 99.0 70.0 0.0 0.0 32.0 0.254 22.0 0.0
Mode of all the attributes of original data is
pregs plas pres skin test BMI pedi Age class
1.0 99 70.0 0.0 0.0 32.0 0.254 22.0 0.0
Inference: The value of mode is equal to the original mode values of all the attributes that using mean doesn't affect the mode value of all attributes.

Standard deviation of all the attributes of missing data using mean method is
pregs 3.373860
plas 30.990181
pres 19.691360
skin 15.946203
test 110.607605
BMI 7.764755
pedi 0.333199
Age 11.519670
class 0.475120
Standard deviation of all the attributes of original data is
pregs 3.369578

plas 31.972618
pres 19.355807
skin 15.952218
test 115.244002
BMI 7.884160
pedi 0.331329
Age 11.760232
class 0.476951

Inference: The value of Standard deviation is approximately equal with the original Standard deviation values of all the attributes.

(ii) Root mean square error (RMSE) between the original attributes and the replace value is given as:



(b) (i) After replacing the missing values of every attributes with the linear interpolated values of it, Comparison of mean, median, mode and standard deviation of the attributes with the original data

Mean of all the attributes of missing data using interpolation method is
pregs 3.885593
plas 120.349576
pres 69.109463
skin 20.392655
test 77.355226
BMI 32.046328
pedi 0.477325
Age 33.216102
class 0.343220
Mean of all the attributes of original data is
pregs 3.845052
plas 120.894531
pres 69.105469
skin 20.536458
test 79.799479

BMI 31.992578
pedi 0.471876
Age 33.240885
class 0.348958

Inference: The value of mean is approximately equal with the original mean values of all the attributes as most of the changes happen after decimal which shows how close the result is with the original value.

Median of all the attributes of missing data using interpolation method is
pregs 3.0000
plas 117.0000
pres 72.0000
skin 23.0000
test 27.0000
BMI 32.2500
pedi 0.3825
Age 29.0000
class 0.0000
Median of all the attributes of original data is
pregs 3.0000
plas 117.0000
pres 72.0000
skin 23.0000
test 30.5000
BMI 32.0000
pedi 0.3725
Age 29.0000
class 0.0000

Inference: The value of median for more than 90% attributes is equal to the original median values of all the attributes which shows using interpolation method median should not affected much.

Mode of all the attributes of missing data using interpolation method is

pregs	plas	pres	skin	test	BMI	pedi	Age	class
1.0	99.0	70.0	0.0	0.0	32.0	0.254	22.0	0.0

Mode of all the attributes of original data is

pregs	plas	pres	skin	test	BMI	pedi	Age	class
1.0	99	70.0	0.0	0.0	32.0	0.254	22.0	0.0

Inference: The value of mode is equal to the original mode values of all the attributes that using mean doesn't affect the mode value of all attributes.

Standard deviation of all the attributes of missing data using interpolation method is

pregs 3.373860
plas 31.274798

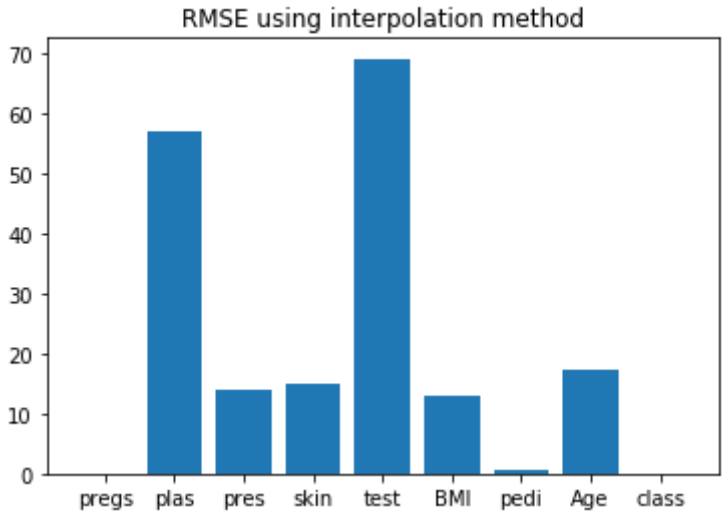
pres 19.735986
skin 15.975849
test 110.755991
BMI 7.792615
pedi 0.334248
Age 11.652648
class 0.475120

Standard deviation of all the attributes of original data is

pregs 3.369578
plas 31.972618
pres 19.355807
skin 15.952218
test 115.244002
BMI 7.884160
pedi 0.331329
Age 11.760232
class 0.476951

Inference : The value of Standard deviation is approximately equal or very close to the original Standard deviation values of all the attributes.

(ii) Root mean square error (RMSE) between the original attributes and the replace value is given as:



Final Conclusion: The root mean square error(RMSE) for interpolation method is slightly higher then the mean method and in both the case rmse for 'test' and 'plas' are higher than all the other attributes. Also using interpolation method we get to know that the mode and the median of this didn't differ much with the original data on the other hand for mode is totally equal for both the values.

5. (i) After replacing missing values of each attribute with the interpolation method all the outliers are given as:

All outliers present in Age attribute

Total no of outliers in Age attribute is: 8

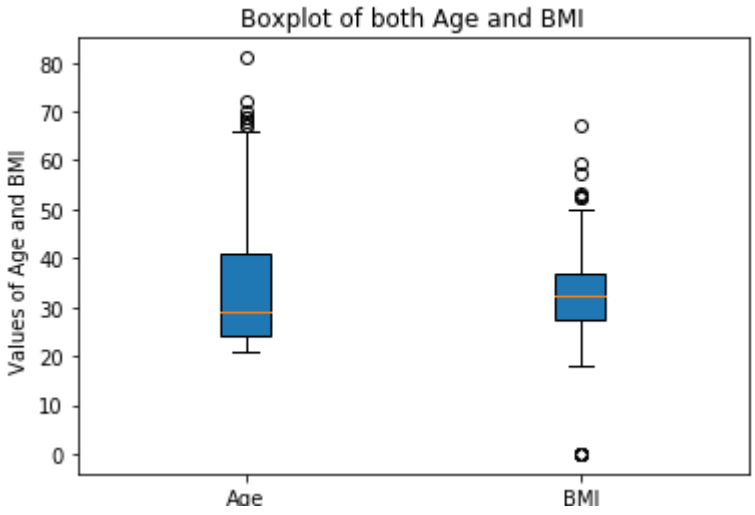
list containing all the outliers in Age attribute: [69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0]

All outliers present in BMI attribute

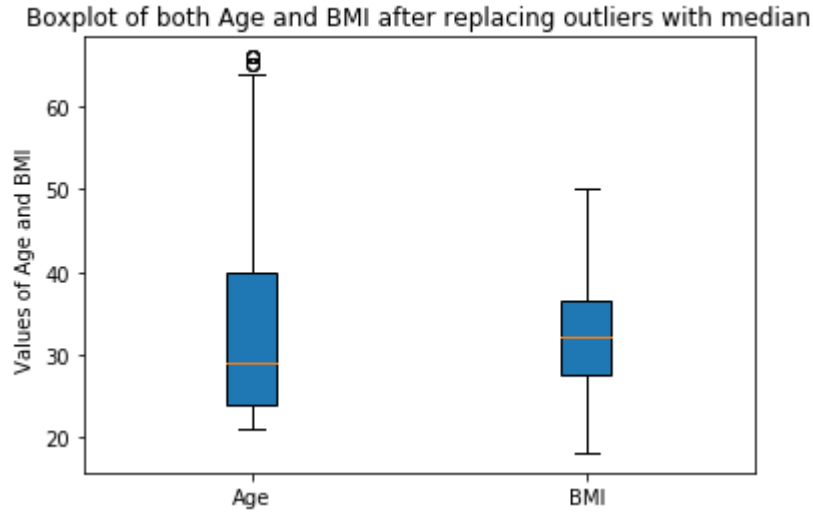
Total no of outliers BMI attribute is: 16

list containing all the outliers in BMI attribute: [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]

Boxplot of “Age” and “BMI” is given as:



(ii)Replacing all the outliers with the median, box plot is given as:



Conclusion: After replacing all the outliers value with the median in the “BMI” attribute all the outliers disappears as the value of range of the outliers are replaced with median so the values of lower as well as upper whisker get balanced and it shows that the new median matches with the previous median(when outliers are not removed) and because of this in BMI attribute no outliers are present. On the other hand in the “Age” attribute when outliers are replaced with the median as the values of upper and lower whisker doesn’t get balanced which implies that new median is different from the previous one that why outliers are still present in this case.