Data Science III (IC272)

Lab Report

On

Data Cleaning – Handling Missing Values and Outlier Analyses

By

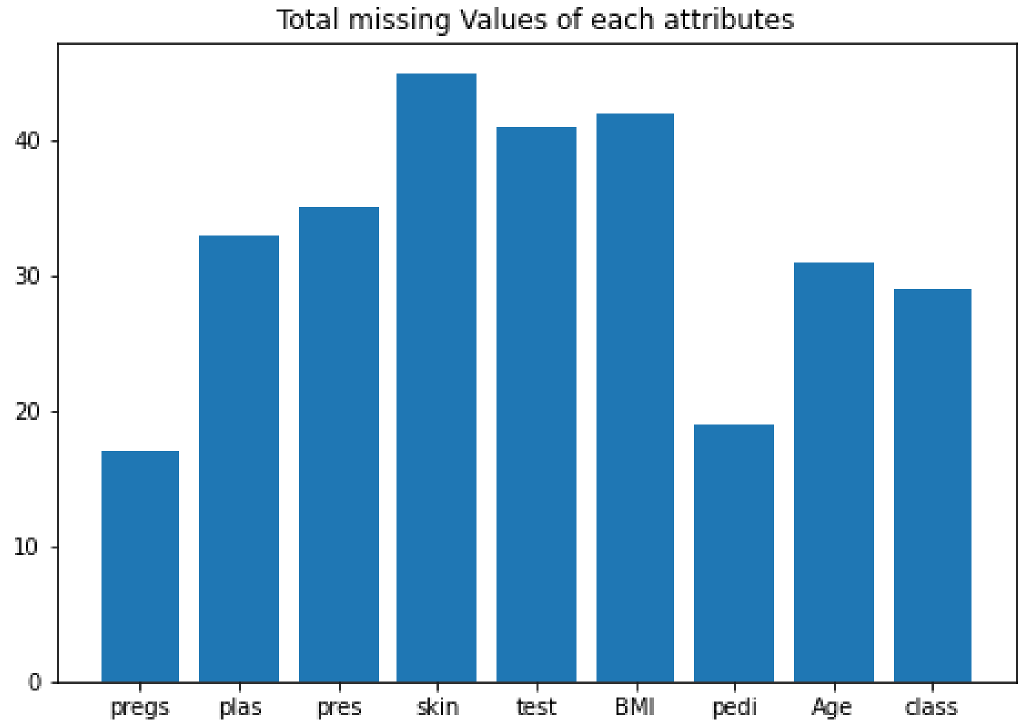
Tarun Singla

B19198

Contact No: - 8872526396

Question 1

The bar plot below indicates the frequency of missing values in each attribute. From this plot, we can infer that most of the missing values are in the attribute ‘skin’, ‘BMI’ and ‘test’ and least of them are in the attribute ‘pregs’



Question 2

2a. In this part, the tuples with missing values in equal to or more than one third of attributes (>= 3) were deleted.

Total number of tuples deleted 39

Row number of 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

2b in this part, the tuples with missing values in target attribute (‘class’) were deleted.

Total number of deleted tuples 21

Row number of deleted rows are 8, 13, 28, 29, 35, 62, 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 308, 746, 748

Question 3

3. Below shows the number of missing values remaining after the deletion of redundant tuples in each attribute with most of them in attribute ‘Age’ and least in attributes ‘pregs’ and ‘class’

Total Missing values are 69.

Missing Values of each Attribute: -

pregs 0

plas 12

pres 9

skin 8

test 8

BMI 12

pedi 2

Age 18

class 0

Question 4

4.

MEAN MEDIAN MODE STANDARAD DEVIATION OF ORIGINAL DATA

…………………………………………….

Mean 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


Median 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


Mode of Original data is

pregs     1

plas      100

pres      70

skin      0

test      0

BMI       32.0

pedi      0.254

Age       22

class     0


Standard Deviation 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
```

4a in this, the missing values were replaced by the mean of their respective attribute. Then, the mean, median, mode and standard deviation for each attribute was calculated and compared with that of the original data as shown in the figure below

Mean after filling with mean 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
```

Median after filling with mean 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

Mode After filling with mean is

pregs     1.0

plas     100.0

pres     70.0

skin     0.0

test     0.0

BMI     32.0

pedi     0.254

Age     22.0

class     0.0

Standard Deviation after filling with mean 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
```

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.

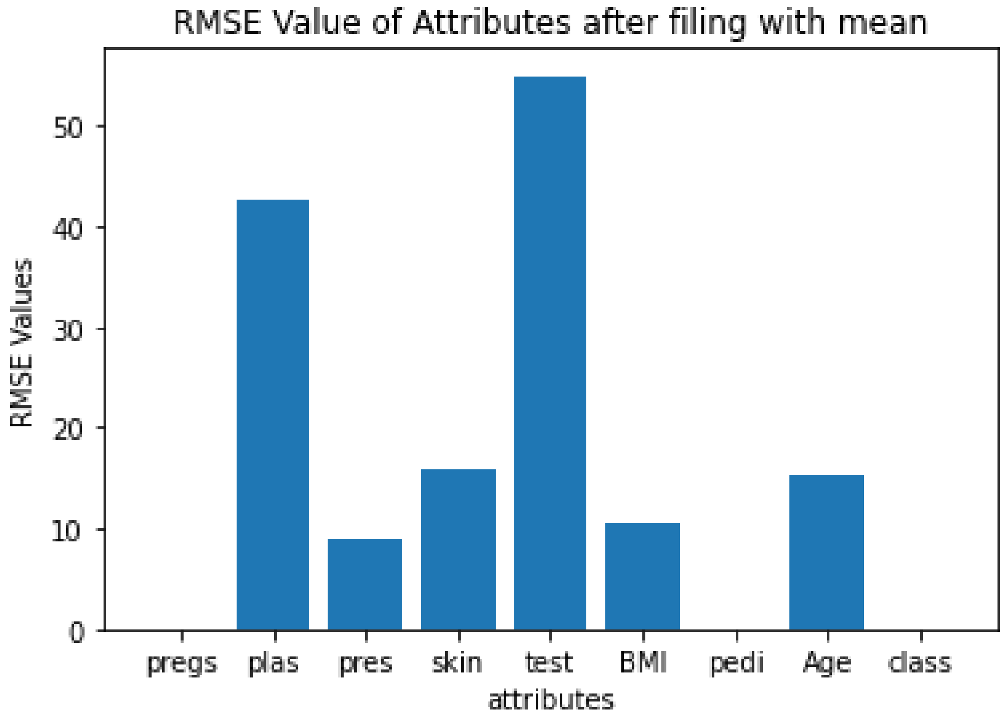
- For the attribute **test** almost, all parameters are different as compared to the original one.
- From the above data, we can say that we can clean the missing data by replacing with the mean of their attributes because there is a very slight difference as compared to the original one.

RMSE Values after Replacing with mean

pregs	**0**
plas	**42.64387412044079**
pres	**8.950321330960236**
skin	**15.839442244354595**
test	**54.969720793193346**
BMI	**10.450965534783302**
pedi	**0.046762740833851374**
Age	**15.365829400182065**
class	**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 **test** attribute is very high which states that we cannot use this method of data cleaning. We cannot fill the missing values by the mean of the attributes because huge error is there.
- The RMSE value of **pregs** and **class** is zero because there is no any missing value in these attribute.
- The RMSE value of **pedi** is very low which shows very low error and this method of cleaning the data is suitable here.



4a in this the missing values in each attribute were replaced using the linear interpolation technique. Then, the mean, median, mode and standard deviation for each attribute was calculated and compared with that of the original data as shown below.

Mean after filling with interpolation 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**

```
Median after filling with interpolation 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

Mode after filling with interpolation is

pregs      1.0

plas      100.0

pres      70.0

skin      0.0

test      0.0

BMI      32.0

pedi      0.254

Age      22.0

class      0.0

Standard Deviation after filling with interpolation 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

CONCLUSION

- After filling the missing values with the interpolation 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.
- But for attribute **test** the values are different from the original one so it's not a good method of data cleaning for attribute **test.**

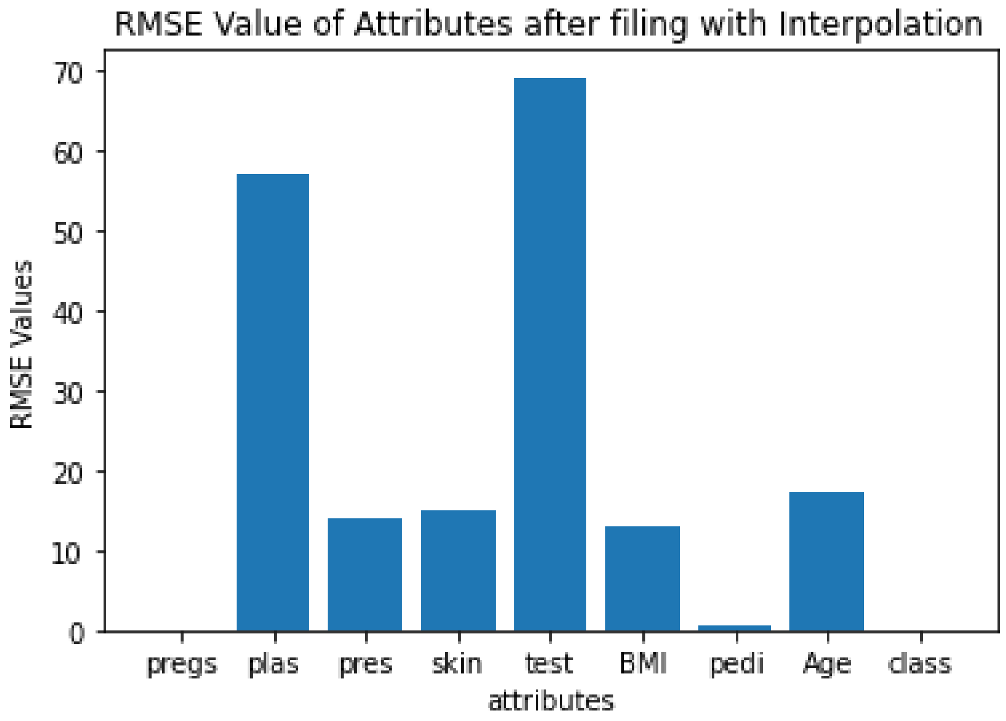
RMSE Values after filling missing values with interpolation

pregs	**0**
plas	**57.055832791709875**
pres	**13.771347065556077**
skin	**14.875828641718678**
test	**68.98482623012107**
BMI	**12.819238291348297**
pedi	**0.5085297434762297**
Age	**17.399712641305314**
class	**0**

CONCLUSION

- The RMSE value of **test** attribute is very high which states that we cannot use this method of data cleaning. We cannot fill the missing values by the mean of the attributes because huge error is there.
- The RMSE value of **pregs** and **class** is zero because there is no any missing value in these attribute.
- The RMSE value of **pedi** is very low which shows very low error and this method of cleaning the data is suitable here.

- 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 using by their mean is the most suitable method here because the root mean square error is least in this case.



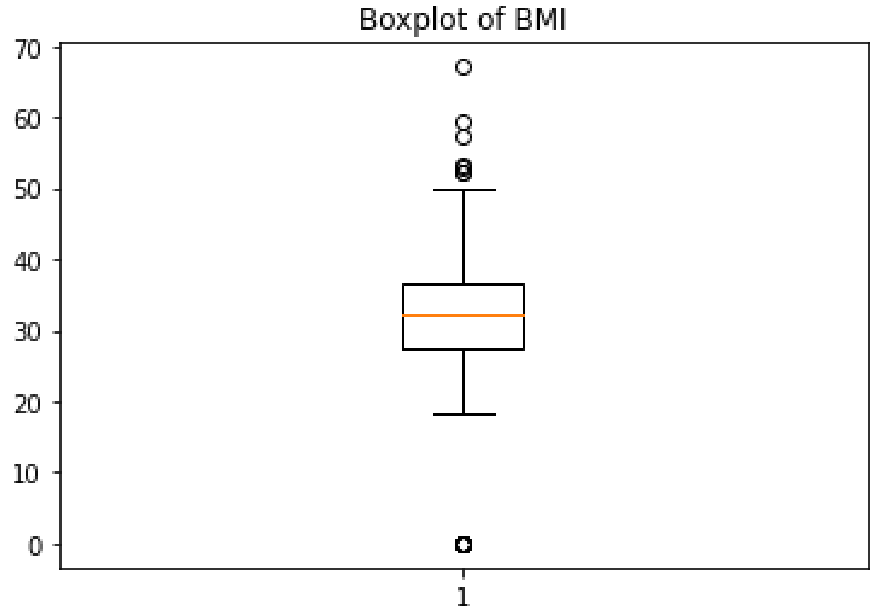
Question 5

5a After replacing the missing values by interpolation method, the outliers in the attributes ‘Age’ and ‘BMI’ were identified as follows:

outliers for BMI 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

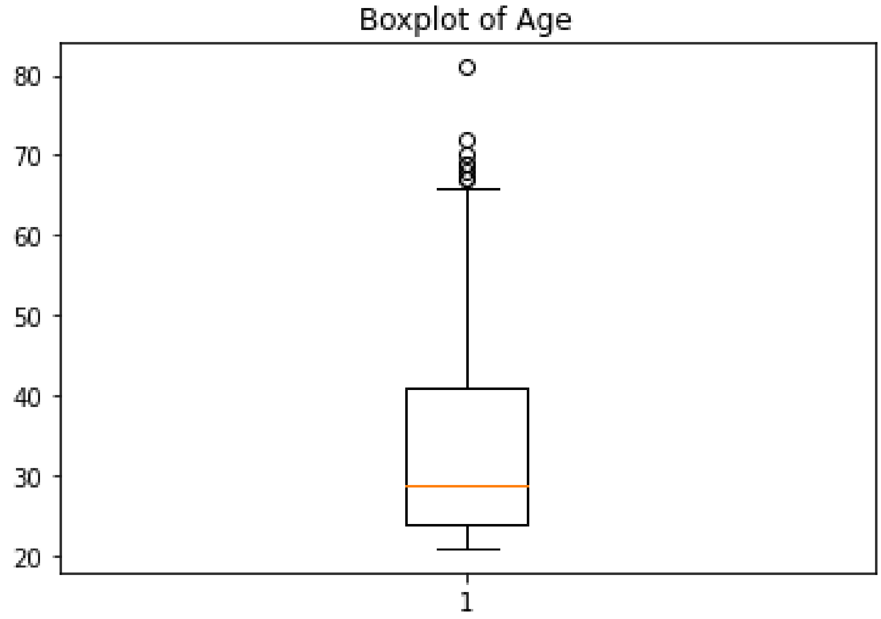
outliers for Age is 0.0, 0.0, 0.0, 53.2, 67.1, 52.3, 52.3, 52.9, 0.0, 0.0, 59.4, 0.0, 0.0, 57.3, 0.0, 0.0, 69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0

Also, the boxplot was plotted for both of these attributes as shown below.



CONCLUSION

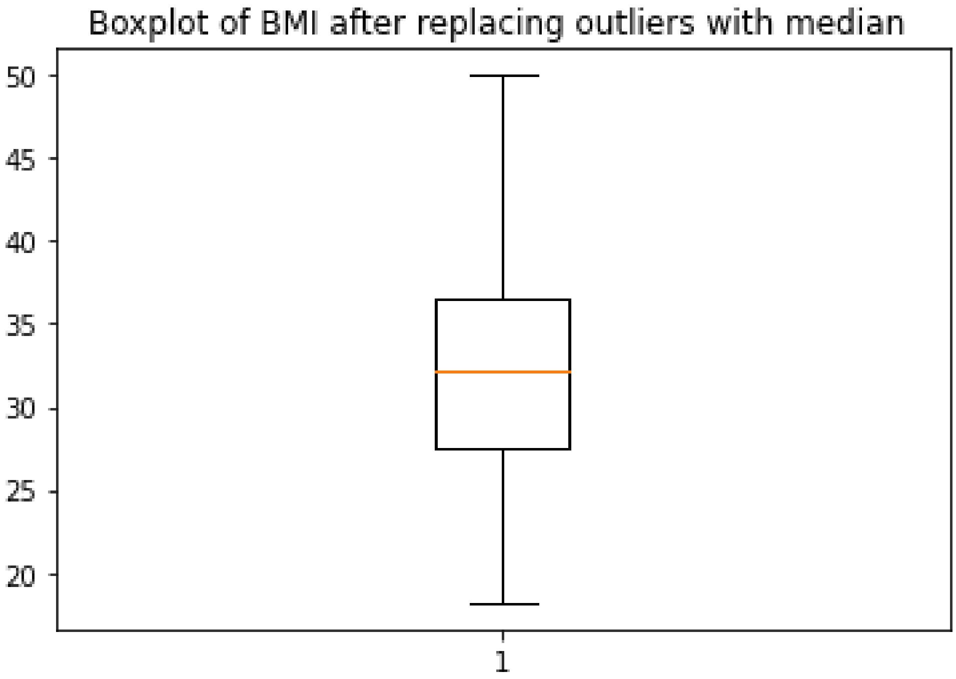
- There are 16 outliers in the attribute BMI.
- The first quartile, median and third quartile are uniformly distributed in the boxplot.
- The outliers are mainly in the range 0 and 50-60.



CONCLUSION

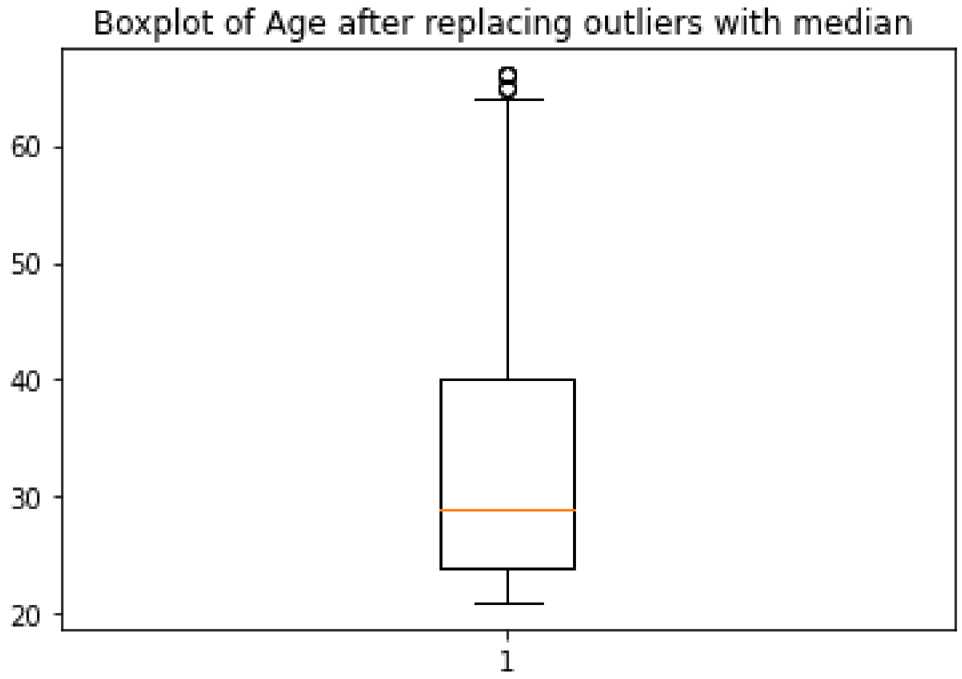
- The outliers are those values which do not satisfy the condition **(Q1-1.5*IQR) < X < (Q3+1.5*IQR).**
- There are eight outliers in the attribute **Age.** These are the values which differs significantly from the other values.
- The red line is representing the value of median which is close to the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

After Replacing 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.



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.