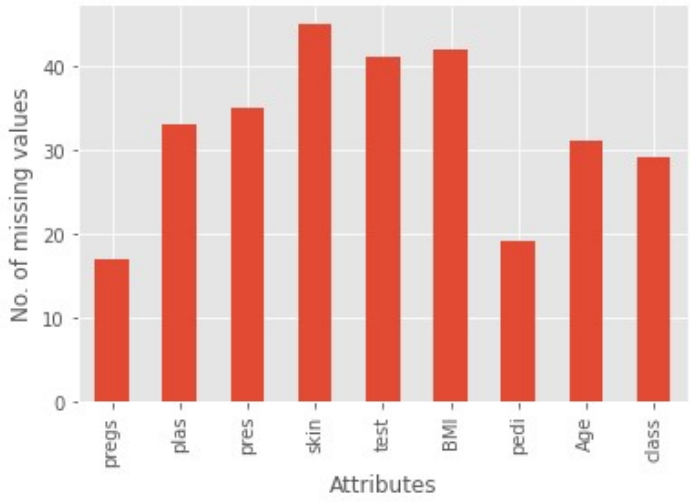
REPORT (LAB 2)

Rashmi, B19218
Ph no. 7015331137

1.



2.(a)

Total no. of tuples deleted = 39

Row numbers of deleted tuples:

[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]

2.(b)

Total no. of tuples deleted = 21

Row numbers of deleted tuples:

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

3. Number of missing values in each attributes:

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) = 69

4. (a) Mean, Median, Mode and Standard Deviation for each attributes of used file when missing values are replaced by mean of the respective attribute:

	Mean	**Median**	**Mode**	**Standard Deviation**
pregs	3.885593	3	1	3.373860
plas	120.666667	118	99	30.990181
pres	69.001431	72	70	19.691360
skin	20.348571	23	0	15.946203
test	77.814286	36	0	110.607605
BMI	32.009339	32.009339	32	7.764755
pedi	0.476042	0.3825	0.254	0.333199
Age	33.094203	29	22	11.519670
class	0.343220	0	0	0.475120

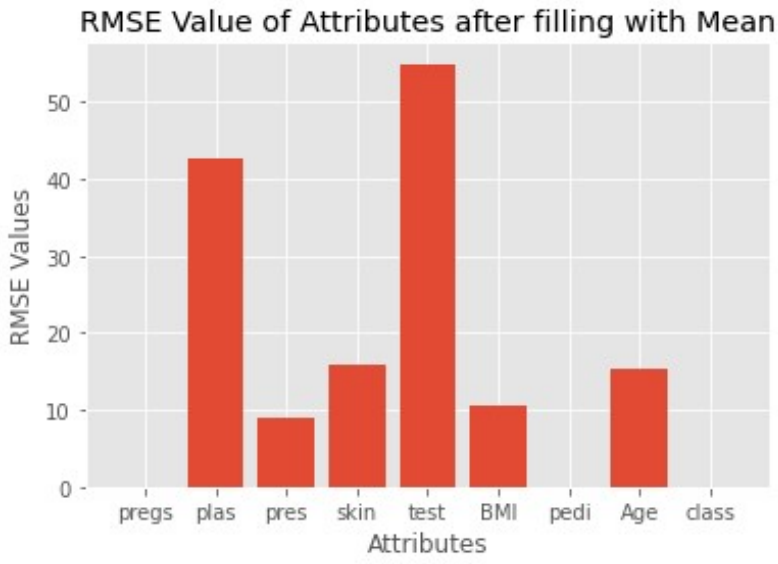
Results from **Original** file:

	Mean	**Median**	**Mode**	**Standard Deviation**
pregs	3.845052	3	1	3.369578
plas	120.894531	117	99	31.972618
pres	69.105469	72	70	19.355807
skin	20.536458	23	0	15.952218
test	79.799479	30.5	0	115.244002
BMI	31.992578	32	32	7.884160
pedi	0.471876	0.3725	0.254	0.331329
Age	33.240885	29	22	11.760232
class	0.348958	0	0	0.476951

The values of the mean, median and mode remain almost same on replacing the missing values with the mean when compared with original data.

Change in standard deviation is more as compared to others.

(ii)



A low RMSE value indicates that the difference in the values in the original dataset and the modified dataset is very less i.e. the values are almost same.

A large value indicates that the difference is large.

Here, highest RMSE is of “test”.

4.(b)) Mean, Median, Mode and Standard Deviation for each attributes of used file when missing values are replaced by linear interpolation technique:

	Mean	Median	Mode	Standard Deviation
pregs	3.885593	3	1	3.373860
plas	120.349576	118	99	31.274798
pres	69.109463	72	70	19.735986
skin	20.392655	23	0	15.975849
test	77.355226	36	0	110.755991
BMI	32.046328	32.009339	32	7.792615
pedi	0.477325	0.3825	0.254	0.334248
Age	33.216102	29	22	11.652648
class	0.343220	0	0	0.475120

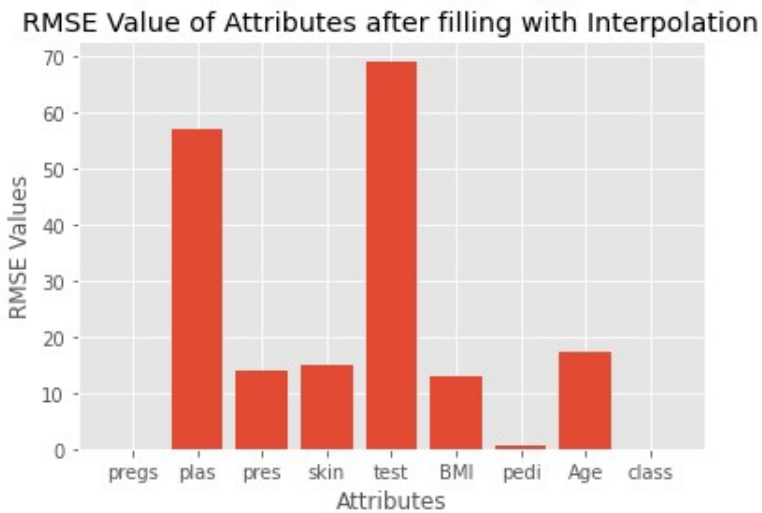
Results from **Original** file:

	Mean	Median	Mode	Standard Deviation
pregs	3.845052	3	1	3.369578
plas	120.894531	117	99	31.972618
pres	69.105469	72	70	19.355807
skin	20.536458	23	0	15.952218
test	79.799479	30.5	0	115.244002
BMI	31.992578	32	32	7.884160
pedi	0.471876	0.3725	0.254	0.331329
Age	33.240885	29	22	11.760232
class	0.348958	0	0	0.476951

Here also, the values of the mean, median and mode remain almost same when compared with original data. The relation between different attributes is also not retained.

Change in standard deviation is more as compared to others.

(ii)

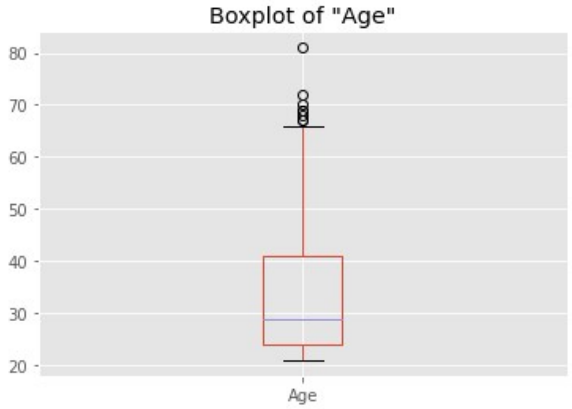


The attributes “pregs” and “class” has no missing values in the modified dataset, so their RMSE values are zero.

In case of “pedi”, there are only 2 missing values therefore low value. Here, highest RMSE is of “test”.

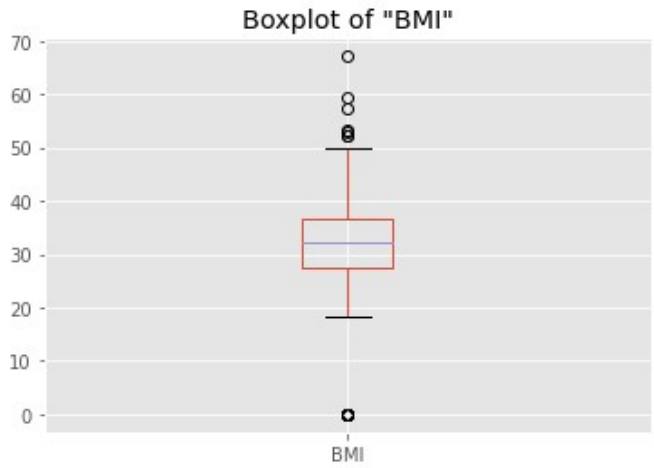
5. Outliers in attribute **"Age"**:

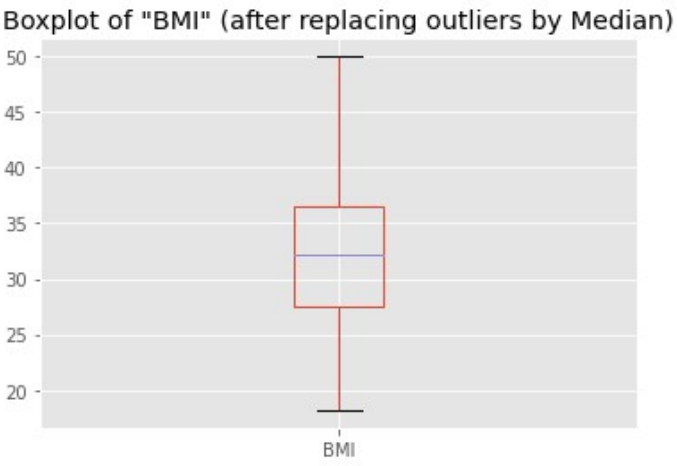
[69, 67, 72, 81, 67, 70, 68, 69]



Outliers in attribute **"BMI"**:

[0, 0, 0, 53.2, 67.1, 52.3, 52.3, 52.9, 0, 0, 59.4, 0, 0, 57.3, 0, 0]





In case of “BMI”, there are no outliers after replacing by Median. But in case of “Age”, there are still outliers even after replacing outliers by Median. This shows that it is not necessary that when we replace outliers by Median, there will be no outliers in the resulting data.

On replacing the values of outliers of two given attributes with their respective medians, their number reduced significantly. Also, the above plots clearly tell us the median doesn’t get affected by the presence of outliers in the data, which is why it’s a better to replace outliers with their median (not mean).

****end****