ASSIGNMENT 2

Harsh Arya (B20043)

Question 1:

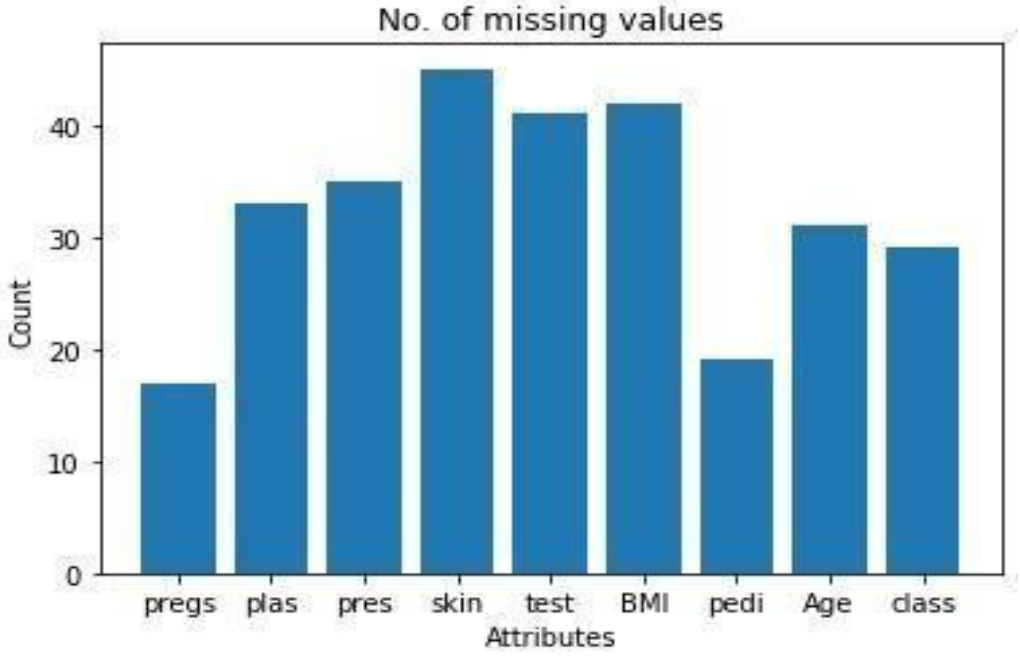

Fig.1

Attribute	pregs	plas	pres	skin	test	BMI	pedi	Age	class
No. of missing values	17	33	35	45	41	42	19	31	29

Table.1

Question 2:

(a) Tuples (rows) having equal to or more than one third of attributes with missing values:

Total no. of tuples deleted: 39

Row no. of deleted tuples:

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

(b) Tuples (rows) having missing value in the target (class) attribute:

Total no. of tuples deleted: 21

Row no. of deleted tuples:

ASSIGNMENT 2

Harsh Arya (B20043)

8 13 28 29 35 62 92 95 107 110 130 131 132 133 149 182 188 218 308 746 748

Question 3:

Attribute	pregs	plas	pres	skin	test	BMI	Pedi	Age	class
No. of missing values after step 2	0	12	9	8	8	12	2	18	0

Table.2

Total no. of missing values: 69

Question 4: Experiments on filling missing values:

a) Replaced the missing values by mean of their respective attribute:

i) Cleaned Data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
Mean	3.885	120.666	69.001	20.348	77.814	32.009	0.476	33.094	0.343
Median	3.000	118.000	72.000	23.000	36.000	32.009	0.382	29.000	0.000
Mode1	1.000	99.000	70.000	0.000	0.000	32.000	0.254	22.000	0.000
Mode2		100.000					0.258		
Standard Deviation	3.373	30.990	19.691	15.946	110.607	7.764	0.333	11.519	0.475

Table.3

Original Data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
Mean	3.845	120.894	69.105	20.536	79.799	31.992	0.471	33.240	0.348
Median	3.000	117.000	72.000	23.000	30.500	32.000	0.372	29.000	0.000
Mode1	1.000	99.000	70.000	0.000	0.000	32.000	0.254	22.000	0.000

ASSIGNMENT 2

Harsh Arya (B20043)

Mode2		100.000					0.258		
Standard Deviation	3.369	31.972	19.355	15.952	115.244	7.884	0.331	11.760	0.476

Table.4

• **Inference:** From *Table.3* and *Table.4* we can see that centre of the new data did not change much from the original data.

ii) RMSE:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
RMSE	0	42.643	8.950	15.839	54.969	10.450	0.046	15.36	0

Table.5

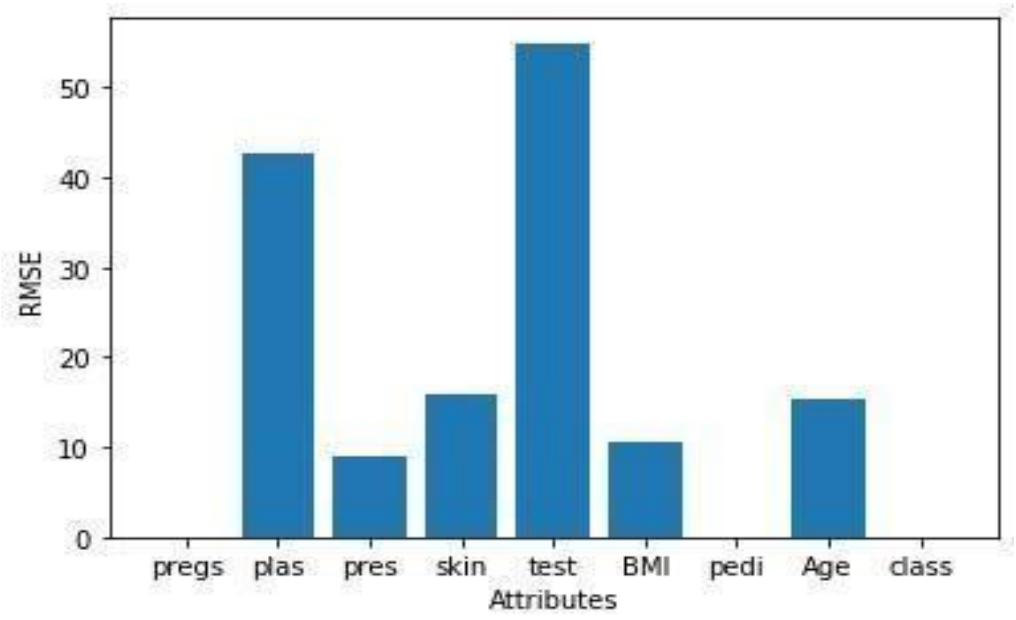

Fig.2

b) Replaced the missing values in each attribute using linear interpolation technique: i)

Cleaned Data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
Mean	3.885	120.349	69.109	20.392	77.355	32.046	0.477	33.216	0.343

ASSIGNMENT 2

Harsh Arya (B20043)

Median	3.000	117.000	72.000	23.000	27.000	32.250	0.382	29.000	0.000
Mode1	1.000	99.000	70.000	0.000	0.000	32.000	0.254	22.000	0.000
Mode2		100.000					0.258		
Standard Deviation	3.373	31.274	19.735	15.975	110.755	7.792	0.334	11.652	0.475

Table.6

Original Data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
Mean	3.845	120.894	69.105	20.536	79.799	31.992	0.471	33.240	0.348
Median	3.000	117.000	72.000	23.000	30.500	32.000	0.372	29.000	0.000
Mode1	1.000	99.000	70.000	0.000	0.000	32.000	0.254	22.000	0.000
Mode2		100.000					0.258		
Standard Deviation	3.369	31.972	19.355	15.952	115.244	7.884	0.331	11.760	0.476

Table.7

- **Inference:** From *Table.6* and *Table.7* we can see that centre of the new data did not change much from the original data.

ii) RMSE:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
RMSE	0	57.055	13.771	14.875	68.984	12.8192	0.508	17.399	0

ASSIGNMENT 2

Harsh Arya (B20043)

Table.8

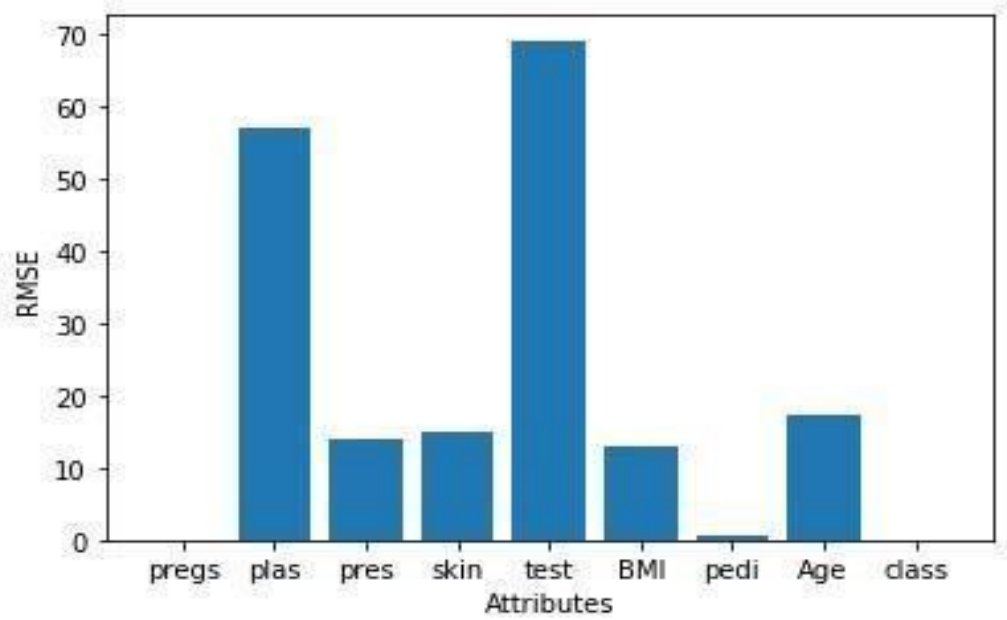

Fig.3

- **Inference:** From *Table.5* and *Table.8*, RMSE value obtained after filling with interpolation method >= RMSE value obtained after filling with mean, for all attributes except skin.

Question5.

i) Outliers in Age:

69.0 67.0 72.0 81.0 67.0 70.0 68.0 69.0

Outliers in BMI:

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

	Q1-1.5*IQR	Q3+1.5*IQR
Age	-1.5	66.5
BMI	13.05	51.05

Table.9

ASSIGNMENT 2

Harsh Arya (B20043)

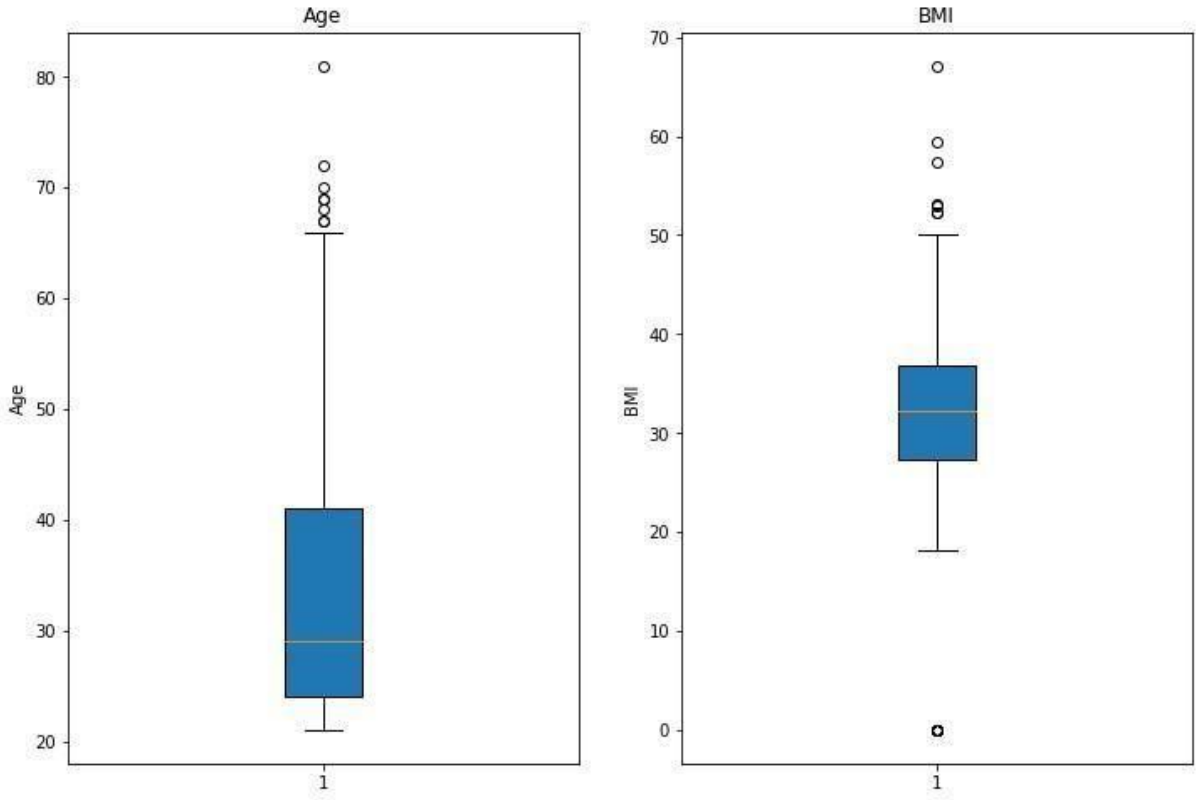

Fig.4

ii) Replaced these outliers by the median of the attribute:

	Q1-1.5*IQR	Q3+1.5*IQR
Age	0.0	64.0
BMI	13.85	50.25

Table.10

ASSIGNMENT 2

Harsh Arya (B20043)

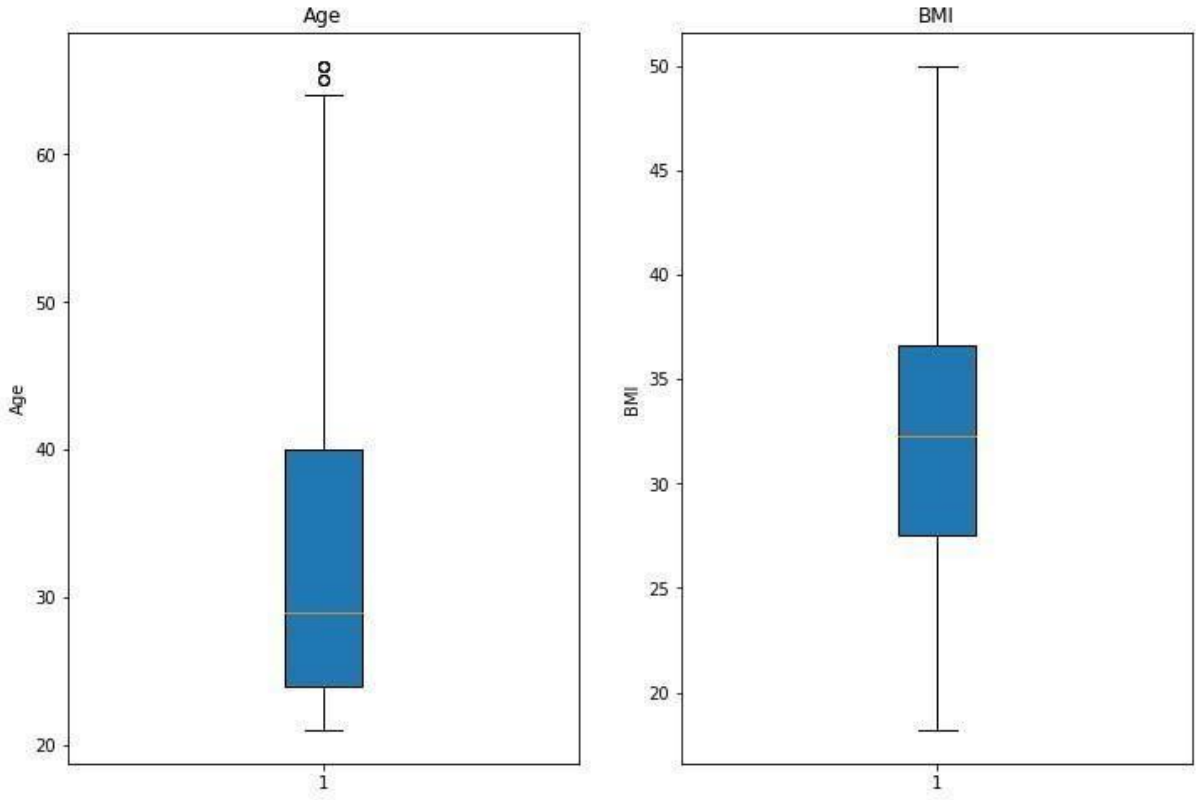

Fig.5

- **Inference:** After replacing the outliers with the median value, the value of whiskers shift. Even though the number of outliers had reduced significantly, few outliers are still present in the Age attribute but not in BMI attribute. This is due to the change in value of whiskers (whiskers slightly shift towards median).