Name: Anooshka Bajaj

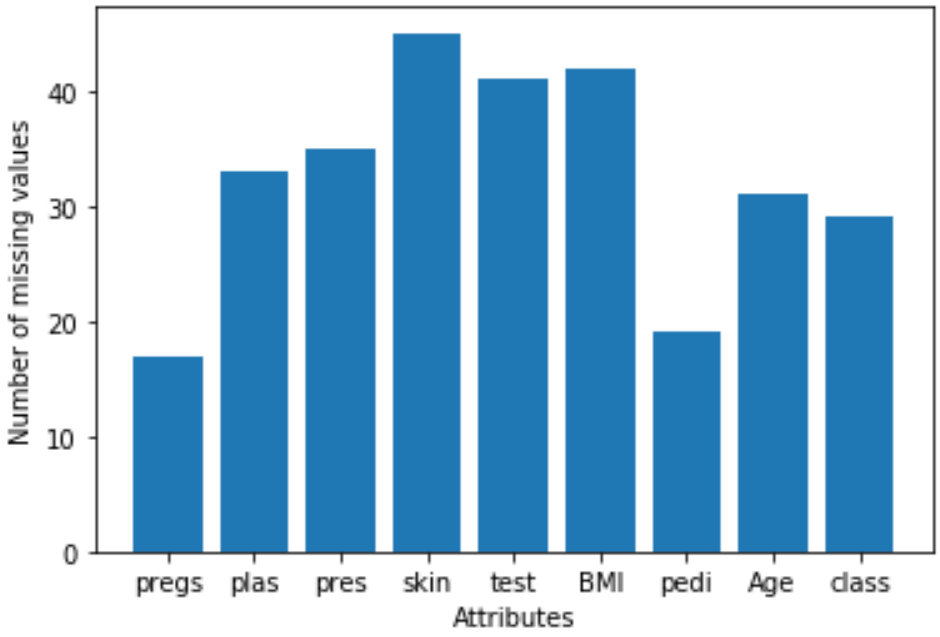
1:

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

Inference:

Attribute skin has the maximum number of missing values (45)

Attribute pregs has the minimum number of missing values (17)



2:

(a) Tuples (rows) having equal to or more than one third of attributes with missing values:
Total number of tuples deleted: 39

Row numbers of the 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 number of tuples deleted: 21

Row numbers of the deleted tuples: [8, 13, 28, 29, 35, 62, 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 308, 746, 748]

3:

The number of missing values in each attribute:

pregs 0

plas 12

pres 9

skin 8

test 8

BMI 12

pedi 2

Age 18

class 0

The total number of missing values in the file: 69

4:

(a) Replacement of missing values by mean of their respective attribute:

(i)

Original data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
mean	3.88	120.63	69.06	20.46	79.61	32.06	0.47	33.10	0.34
median	3.00	117.00	72.00	23.00	29.00	32.00	0.37	29.00	0.00
std deviation	3.37	31.79	19.53	15.96	115.36	7.84	0.33	11.64	0.48
mode 1	1.0	99.0	70.0	0.0	0.0	32.0	0.254	22.0	0.0
mode 2							0.258		

Data filled by mean:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
mean	3.89	120.67	69.00	20.35	77.81	32.01	0.48	33.09	0.34
median	3.00	117.00	72.00	23.00	27.00	32.05	0.38	29.00	0.00
std deviation	3.37	31.26	19.82	16.04	111.24	7.83	0.33	11.67	0.48
mode 1	1.0	99.0	70.0	0.0	0.0	32.0	0.254	22.0	0.0
mode 2		100.0					0.258		

Inference: The centre of data filled by mean does not change much from the centre of original data.

(ii) RMSE:

(b) Replacement of missing values in each attribute using linear interpolation technique:

(i)

Original data:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
mean	3.88	120.63	69.06	20.46	79.61	32.06	0.47	33.10	0.34
median	3.00	117.00	72.00	23.00	29.00	32.00	0.37	29.00	0.00
std deviation	3.37	31.79	19.53	15.96	115.36	7.84	0.33	11.64	0.48
mode 1	1.0	99.0	70.0	0.0	0.0	32.0	0.254	22.0	0.0
mode 2							0.258		

Data filled by interpolation:

	pregs	plas	pres	skin	test	BMI	pedi	Age	class
mean	3.89	120.35	69.11	20.39	77.36	32.05	0.48	33.22	0.34
median	3.00	117.00	72.00	23.00	27.00	32.25	0.38	29.00	0.00
std deviation	3.37	31.27	19.74	15.98	110.76	7.79	0.33	11.65	0.48
mode 1	1.0	99.0	70.0	0.0	0.0	32.0	0.254	22.0	0.0
mode 2	100.0						0.258		

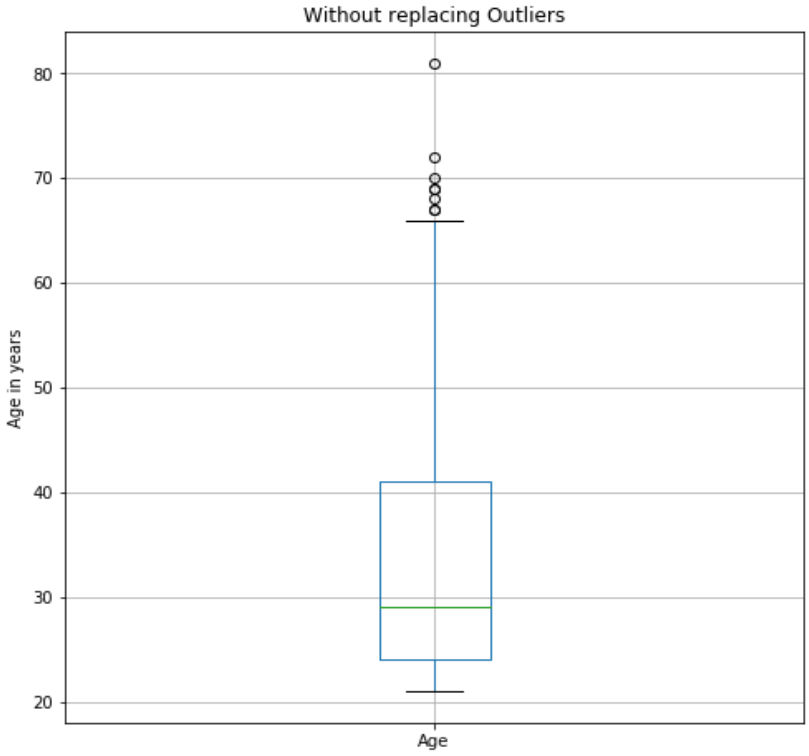
Inference: The centre of data filled by interpolation does not change much from the centre of original data.

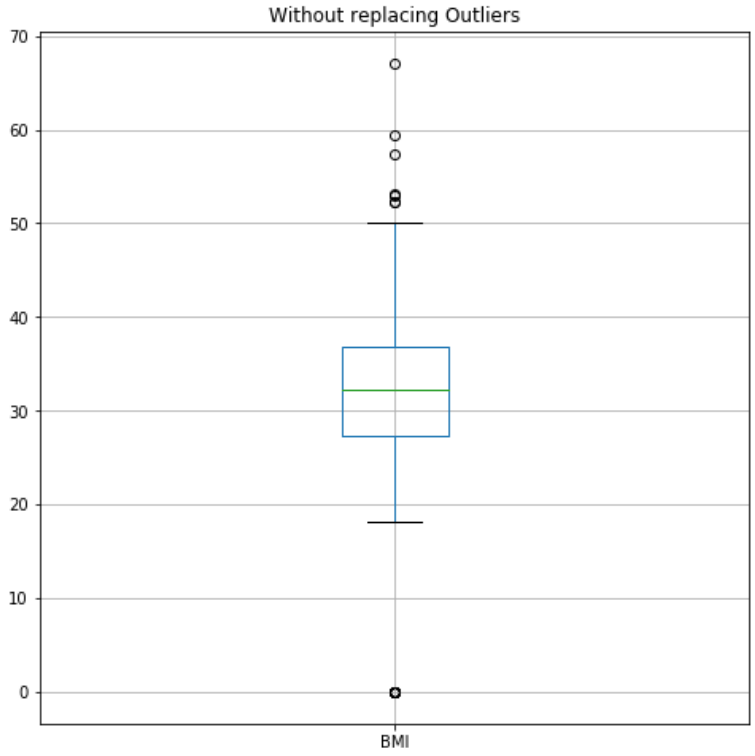
(ii) RMSE:

5:

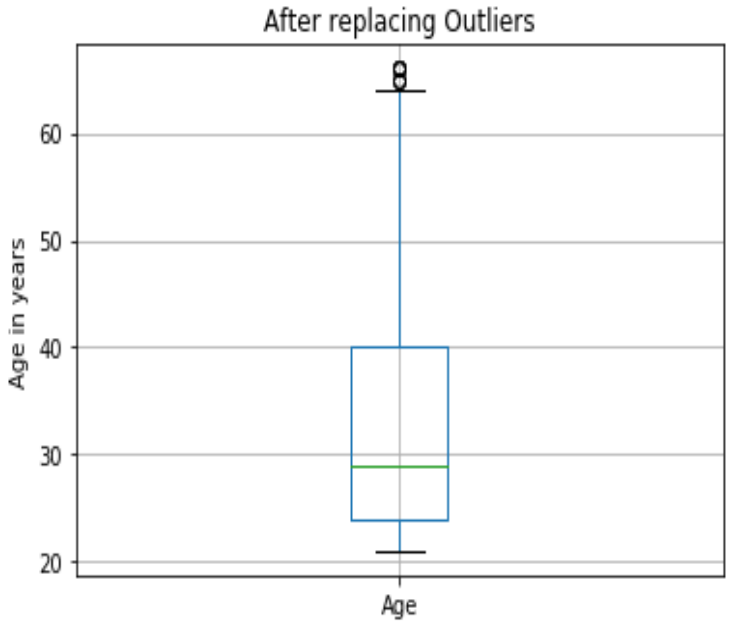
(i) Outliers in Age: 69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0

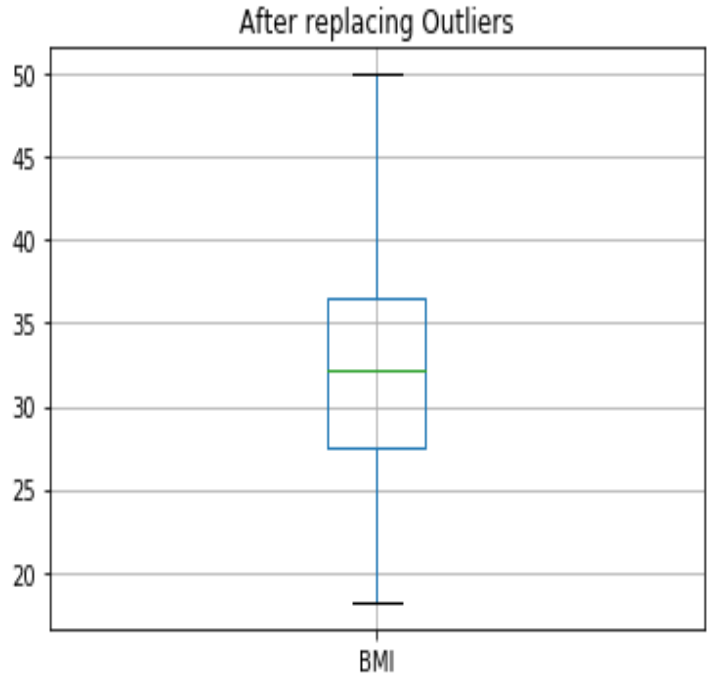
Outliers in BMI: 0.0, 0.0, 0.0, 53.2, 67.1, 52.3, 52.3, 52.9, 0.0, 0.0, 59.4, 0.0, 0.0, 57.3, 0.0, 0.0





(ii) Outliers replaced by the median of the attribute:





Inference: After replacing the outliers with the median value, the value of whiskers shift. Number of outliers has reduced but few outliers are still present in the Age attribute, not in BMI attribute. This is due to the change in the values of whiskers, which slightly moves toward the median value.