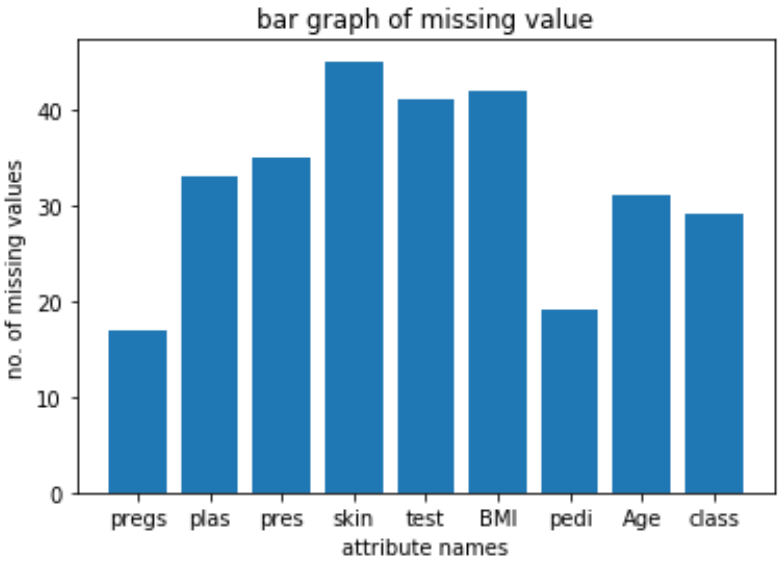
Indian Institute of Technology, Mandi
IC272 – Data Science 3
Lab Assignment 2

Pooja Patidar (B19255)
Mob no.:- 8516921968

Q1.)
Graph of the attribute names with the number of missing values in them



This bar graph shows the no. of missing values in respective attributes.

PREGS	17
PLAS	33
PRES	35
SKIN	45
TEST	41
BMI	42
PEDI	19
AGE	31
CLASS	29

OBSERVATION:

- Many tuples have missing values for several attributes.
- Pandas library is used to detect missing values as “NaN”

Q2.)

In second question we just clean our data by removing tuples having equal or more than 1/3rd missing values.

Total deleted rows: 39

Row no. od 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]

In part (b) we just remove tuples having missing value in target(class) attributes.

OBSERVATION:

- Ignoring the tuples is effective when target variable (here, class) is missing because class attribute is very important attribute here, it tells us about the patient having diabetes or not.

Total deleted rows: 21

Row no. of deleted tuples: 8 13 28 29 35 62 92 95 107 110 130 131 132 133 149 182 188 218 308 746 748

Q3.)

Total no. of missing values after deleting some of the tuples: 69

PREGS	0
PLAS	12
PRES	9
SKIN	8
TEST	8
BMI	12
PEDI	2
AGE	18
CLASS	0

This table shows missing values in each attribute after cleaning data in que 2.

Q4.)

a.)

i.)statistics of missing data after replacing missing values with its mean

	Mean	Median	Mode1	Mode2	Standard Dev
pregs	3.885593	3.000000	1.000	NaN	3.373860
plas	120.6666	118.0000	99.000	100.000	30.990181
pres	69.001431	72.000000	70.000	NaN	19.691360
skin	20.348571	23.000000	0.000	NaN	15.946203
test	77.814286	36.000000	0.000	NaN	110.607605
BMI	32.009339	32.009339	32.000	NaN	7.764755
pedi	0.476042	0.382500	0.254	0.258	0.333199
Age	33.094203	29.000000	22.000	NaN	11.519670

class	0.343220	0.000000	0.000	NaN	0.475120

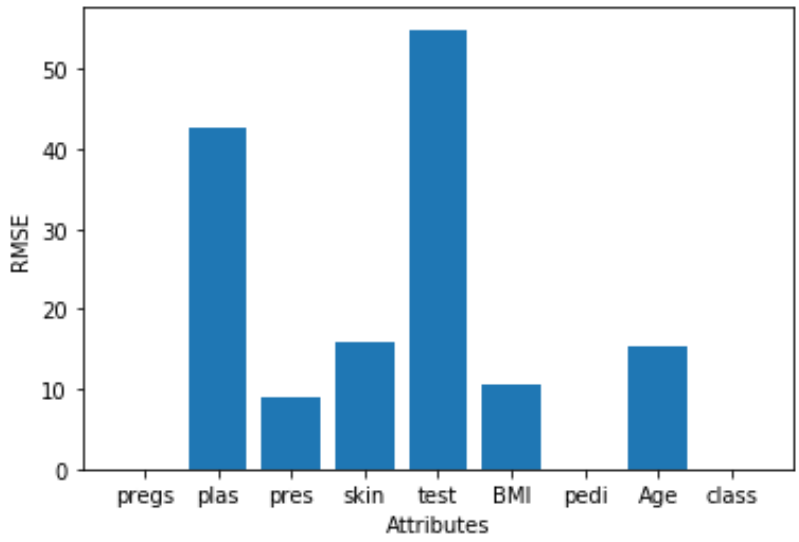
statistics of Original Data:

	Mean	Median	Mode1	Mode2	Standard Dev
pregs	3.845052	3.0000	1.000	NaN	3.369578
plas	120.89453	117.000	99.000	100.000	31.972618
pres	69.105469	72.0000	70.000	NaN	19.355807
skin	20.536458	23.0000	0.000	NaN	15.952218
test	79.799479	30.5000	0.000	NaN	115.244002
BMI	31.992578	32.0000	32.000	NaN	7.884160
pedi	0.471876	0.3725	0.254	0.258	0.331329
Age	33.240885	29.0000	22.000	NaN	11.760232
class	0.348958	0.0000	0.000	NaN	0.476951

OBSERVATION:

- By observing the above two tables we say that after replacing missing value with mean of respective attribute there is negligible change in their central tendency. So, this is an effective way to clean the data or replace missing values.
- However, it does not preserve the relationship with other variables.

ii.)



PREGS	0
PLAS	42.64387
PRES	8.9503
SKIN	15.8394
TEST	54.969
BMI	10.4509
PEDI	0.04676
AGE	15.3658
CLASS	0

OBSERVATION:

- RMSE is computed between replaced value and its corresponding original value.

$$\text{RMSE}=\sqrt{1/\text{Na}\ \ (\sum(\text{yi}-\text{xi})2\ )}$$

Q4 b.)

i.) statistics of missing data after replacing missing values using linear interpolation

	Mean	Median	Mode1	Mode2	Standard Dev
pregs	3.885593	3.0000	1.000	NaN	3.373860
plas	120.349576	117.0000	99.000	100.000	31.274798
pres	69.109463	72.0000	70.000	NaN	19.735986
skin	20.392655	23.0000	0.000	NaN	15.975849
test	77.355226	27.0000	0.000	NaN	110.755991
BMI	32.046328	32.2500	32.000	NaN	7.792615
pedi	0.477325	0.3825	0.254	0.258	0.334248
Age	33.216102	29.0000	22.000	NaN	11.652648
Class	0.343220	0.0000	0.000	NaN	0.475120

statistics of Original Data:

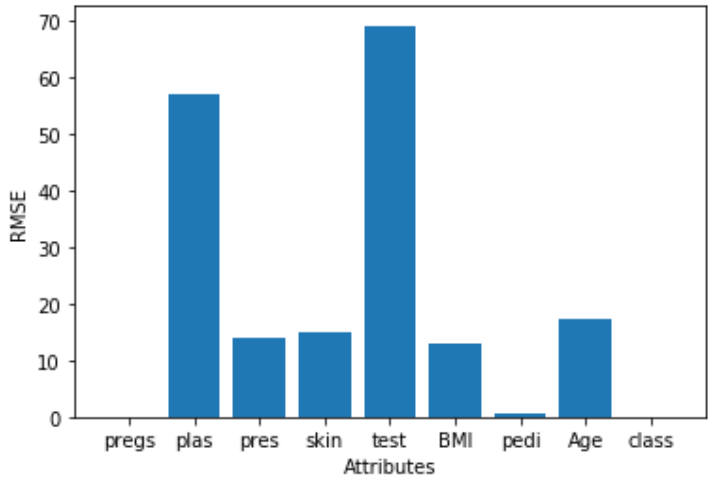
	Mean	Median	Mode1	Mode2	Standard Dev
pregs	3.845052	3.0000	1.000	NaN	3.369578
plas	120.894531	117.0000	99.000	100.000	31.972618
pres	69.105469	72.0000	70.000	NaN	19.355807
skin	20.536458	23.0000	0.000	NaN	15.952218
test	79.799479	30.5000	0.000	NaN	115.244002
BMI	31.992578	32.0000	32.000	NaN	7.884160
pedi	0.471876	0.3725	0.254	0.258	0.331329
Age	33.240885	29.0000	22.000	NaN	11.760232
class	0.348958	0.0000	0.000	NaN	0.476951

OBSERVATION:

- *Linear Interpolation* is achieved by geometrically rendering a straight line between two adjacent points on a graph. It happens column wise.
- By observing the above two tables we say that after replacing missing value using linear interpolation of respective attribute, there is negligible change in their central tendency. So, this is an effective way to clean the data or replace missing values.

- However, it does not preserve the relationship with other variables.

ii.)



pregs	0
plas	57.055
pres	13.771
skin	14.875
test	68.984
BMI	12.819
Pedi	0.508
Age	17.399
class	0

OBSERVATION:

- RMSE is computed between replaced value and its corresponding original value.

$$\text{RMSE}=\sqrt{1/\text{Na}\ \ (\sum(\text{yi}-\text{xi})2\ )}$$

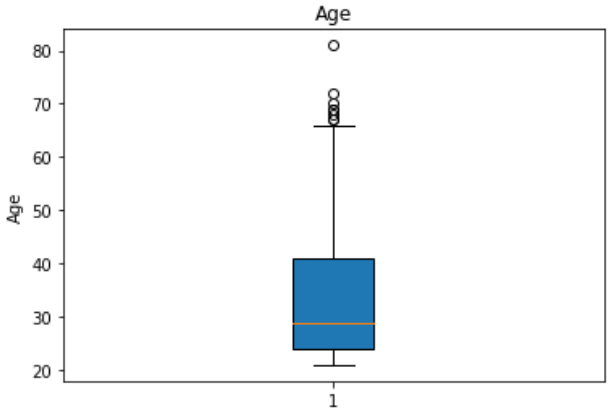
Q5.)

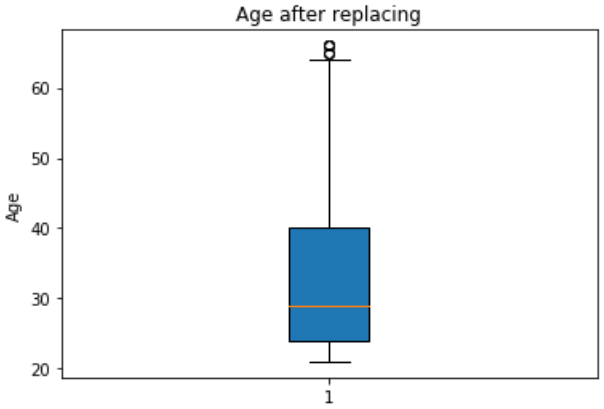
a.)

i.) outliers in Age: [69.0, 67.0, 72.0, 81.0, 67.0, 70.0, 68.0, 69.0] =8

ii.) 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] = 16

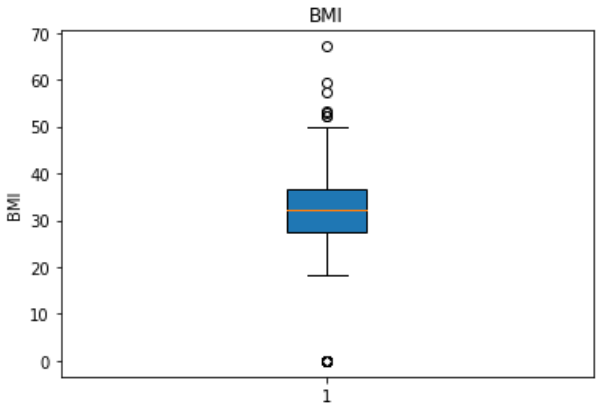
b.)

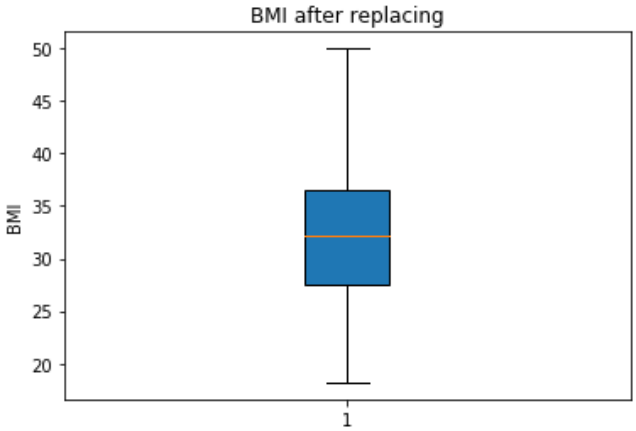




OBSERVATION:

- After replacing outliers with median there is slightly change in value of Q1 and Q3 that's why outliers is appearing in the graph even after replacing.





OBSERVATION:

- There are no outliers in the graph.