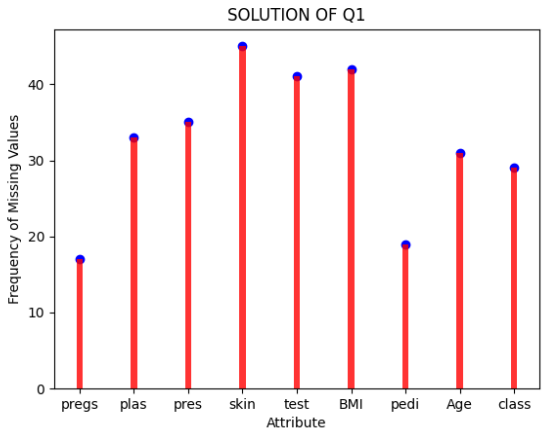
IC272-Data Science III

Report of Lab 2:

Data Cleaning – Handling Missing Values and Outlier Analyses

Name: Prashant Kumar **Roll No.:** B19101 **Mobile No.:** 8700350173

Q1



.

Attribute	Count
Pregs	17
Plas	33
Pres	35
Skin	45
Test	41
BMI	42
Pedi	19
Age	31
Class	29

Q2

a. **Delete (drop) the tuples (rows) having equal to or more than one third of attributes with missing values**
   - **Total number of tuples deleted:** 39
   - **Row number 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. **Drop the tuples (rows) having missing value in the target (class) attribute**
   - **Total number of rows deleted:** 21
   - **Row number 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 Missing Value in updated data: 69

Number of missing values in each attribute:

Attribute	Pregs	Plas	Pres	Skin	Test	BMI	Pedi	Age	Class
Total null values	0	12	9	8	8	12	2	18	0

Q4

Filling Missing Values Using Different Experiment

a). Replacing missing values with mean of each attribute

Missing data filled with mean

	Mean	Median	Mode	Standard Deviation
Pregs	3.885593	3	1	3.373860
Plas	120.672316	118	121	30.990211
Pres	69.001412	72	70	19.69136
Skin	20.344633	23	0	15.946246
Test	77.816384	36	0	110.60760
BMI	32.009181	32	32	7.764755
Pedi	0.474698	0	0.254	0.334157
Age	33.091808	29	22	11.519680
Class	0.343220	0	0	0.475120

Original Data

	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

- **Inference:** Comparing the above two tables, we observe that there is not much deference in the central tendency of the data on filling the missing values with the mean of each attributes.

b). Replacing missing values using interpolation

Missing data filled with interpolation

	Mean	Median	Mode	Standard Deviation
Pregs	3.885593	3	1	3.373860
Plas	120.348870	117	99	31.274096
Pres	69.108757	72	70	19.735687
Skin	20.391243	23	0	15.975610
Test	77.354520	27	0	110.755858
BMI	32.045904	32.25	32	7.792990
Pedi	0.477523	0.3825	0.254	0.334359
Age	33.211864	29	22	11.650511
Class	0.343220	0	0	0.475120

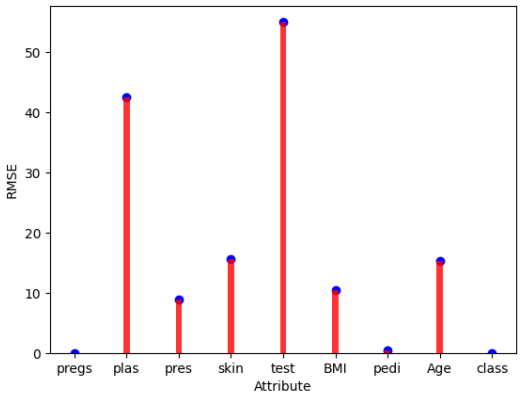
Original Data

	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

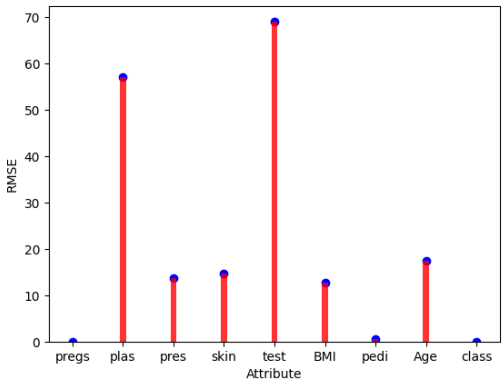
- **Inference:** Comparing the above two tables, we observe that there is not much deference in the central tendency of the data on filling the missing values using interpolation.

RMSE calculation for part (a) and (b).

Replacement by mean



Replacement using intepolation



RMSE

	Replacement by mean	**Replacement using interpolation**
Pregs	0	0
Plas	42.557804	57.028502
Pres	8.950481	13.747727
Skin	15.672428	14.739403
Test	54.994318	68.990941
BMI	10.448884	12.809339
Pedi	0.451146	0.551845
Age	15.394804	17.440375
Class	0	0

- **Inference:** From the above data the RMSE value calculated for each attribute, we observe that the RMSE value obtained by replacing mean is less than the RMSE value obtained using interpolation for most of the attribute.

Q5

Outliner Detection:

i). Without replacing outliners

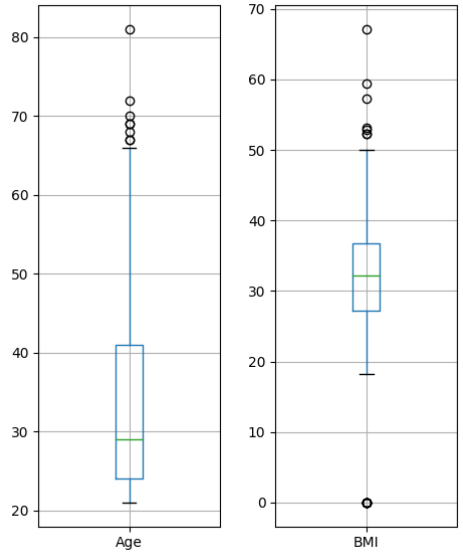
- **Outliners in Age:** 69, 67, 72, 81, 67, 70, 68, 69
- **Outliners in BMI:** 0, 0, 0, 0, 0, 0, 0, 0, 0, 53.2, 67.1, 52.3, 52.3, 52.9, 59.4, 57.3

ii). Replacing Outliner using median

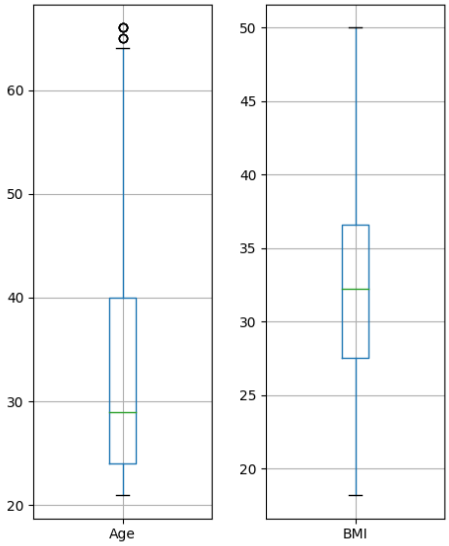
- **Outliner in Age:** 65, 66, 65, 65, 66, 66, 66
- **Outliner in BMI:** No outliner

Box Plot of "Age" and "BMI"

Without replacing outliner



After replacing outliner



	Without replacing outliners		After replacing outliner with median	
	Age	**BMI**	**Age**	**BMI**
Q1-1.5*IQR	-1.5	12.9875	0	13.849
Q3+1.5*IQR	66.5	51.0875	64	50.25
Total Outliners	8	16	7	0

- **Inference:** When we replace the outliners of Age and BMI attribute with the median of the respective attribute, we observe that the whisker get shift. Most of the outliners get reduced but still there are some outliners in the Age attribute but there is no attribute in the BMI attribute. Reason of this is shift of whisker that move toward the median of the data.