IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Data visualization and statistics from data

Student's Name: Akshar Singh **Mobile No: 7428357700**

Roll Number: B20147 **Branch: Computer Science**

1

Table 1 Mean, median, mode, minimum, maximum and standard deviation for all the attributes

S. No.	Attributes	Mean	Median	Mode	Min.	Max.	S.D.
1	pregs	3.845	3	1	0	17	3.370
2	plas	120.895	117	99	0	199	31.973
3	pres (in mm Hg)	69.105	72	70	0	122	19.359
4	skin (in mm)	20.536	23	0	0	99	15.952
5	test (in mu U/mL)	79.799	30.5	0	0	846	115.244
6	BMI (in kg/m^2)	31.993	32	32	0	67.1	7.884
7	pedi	0.472	0.373	0.254	0.078	2.420	0.331
8	Age (in years)	33.241	29	22	21	81	11.760

Inferences:

1. When the standard deviation is close to zero then mean, median and mode are close to each other which is the case for pedi and pregs comparitively.

2 a.

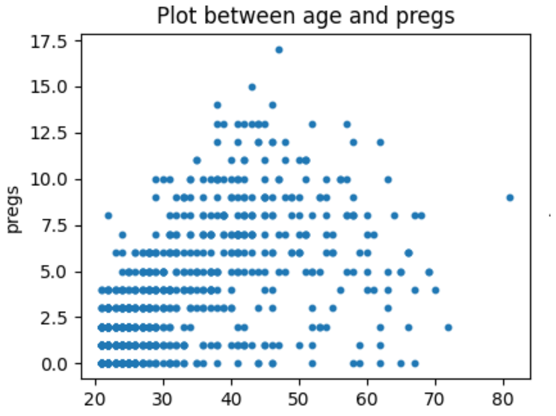

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. The value of no. of times pregnant increases with age and correlation value is 0.544 which justifies it.
2. Density seems to be more at early ages (20-40 years).

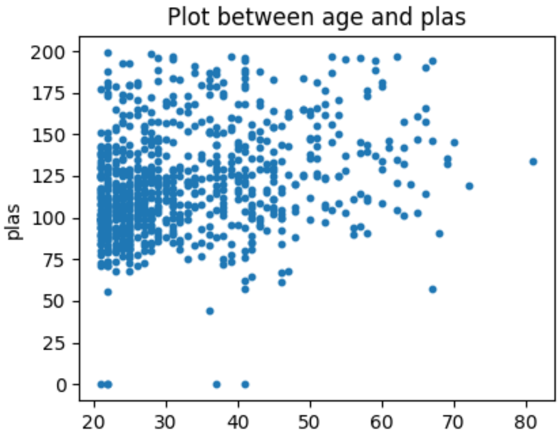

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. We see increase in values of Plasma glucose concentration 2 hours in an oral glucose tolerance test with age and correlation value is 0.264 which justifies it.
2. Density seems to be more at early ages (20-40 years).

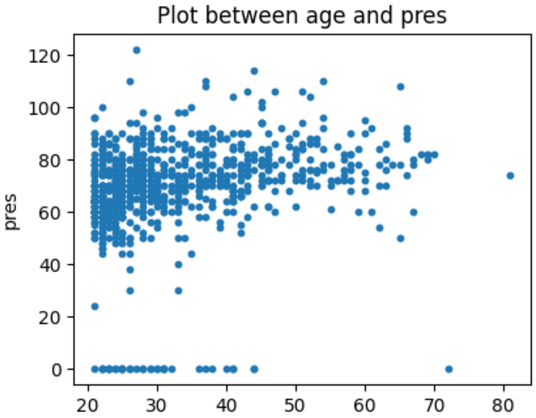

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. We see appreciable amount of increase in diastolic blood pressure with age and correlation value is 0.240 which justifies it.
2. Density seems to be more at early ages and appreciable at middle ages.

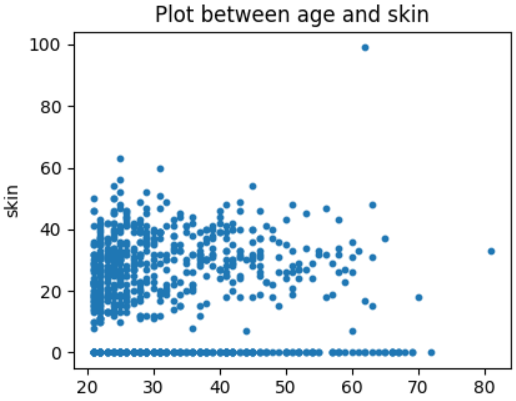

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. We see pretty much decrease in Triceps skin fold thickness with age and correlation value is -0.114 which justifies it.
2. Density seems to be more at early ages (20-30 years).

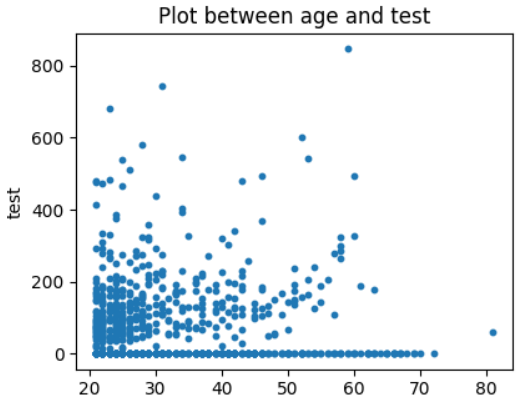

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. We see less values of 2-Hour serum insulin at higher age values and correlation value is -0.042 which justifies it.
2. Density seems to be more at early ages (20-30 years).

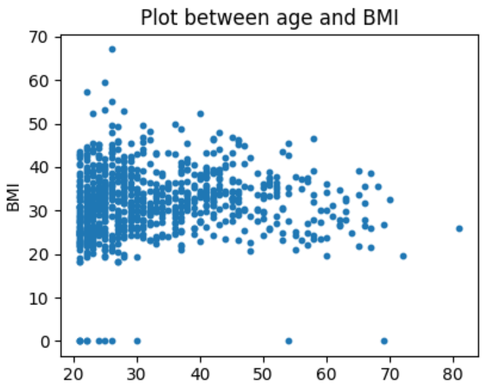

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. We see that BMI is approximately in the same range for all values of ages and correlation is 0.036 which is close to 0 and thus justifies it.
2. Density seems to be more at early ages and middle ages.

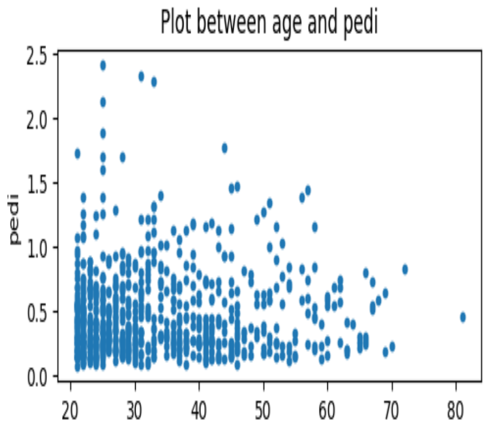

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. We see that Diabetes pedigree function is approximately in the same range for all values of ages and correlation is 0.036 which is close to 0 and thus justifies it.
2. Density seems to be more at early ages(20-30 years).

b.

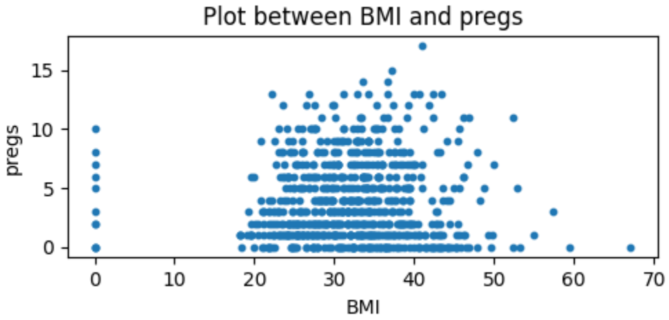

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. We see less increase in BMI with pregs , they seem to be uncorrelated and correlation value is 0.018 which is close to zero.
2. More dense in between 20-40 BMI values.

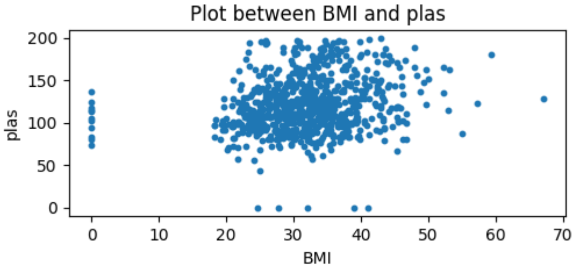

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. High value of Plasma glucose concentration 2 hours in an oral glucose tolerance test with BMI and correlation is 0.221 which justifies it.
2. High density for BMI values in between 20-43.

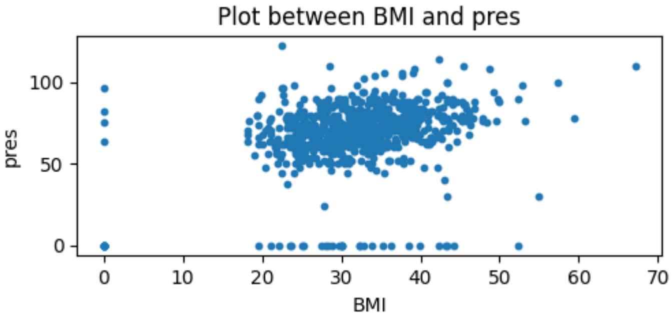

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. High value of Diastolic blood pressure with BMI and correlation is 0.282 which justifies it.
2. High density for BMI values in between 20-43.

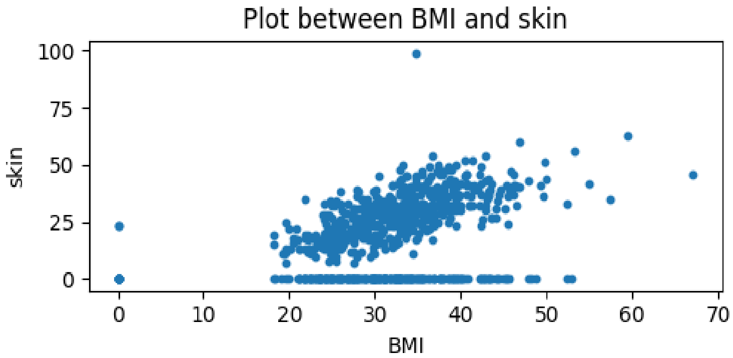

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. We see considerable increase in Triceps skin fold thickness with BMI and correlation value is 0.393 which justifies it.
2. High density for BMI values in between 20-45.

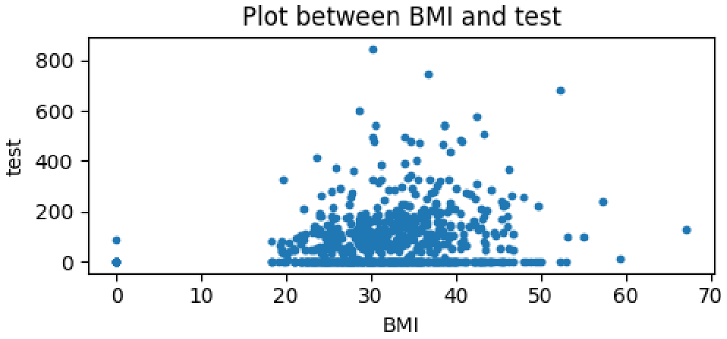

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. High value of 2-Hour serum insulin with BMI and correlation is 0.198 which justifies it.
2. High density for BMI values in between 20-45.

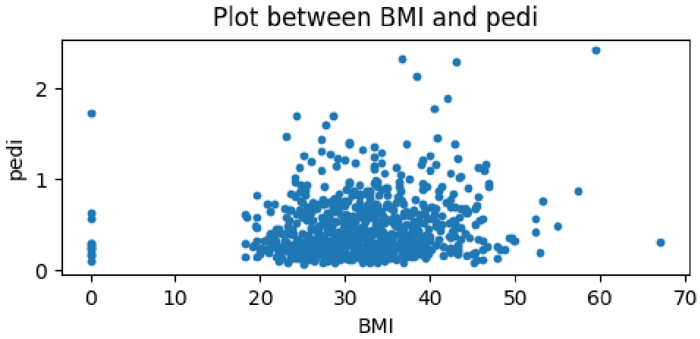

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. High value of Diabetes pedigree function with BMI and correlation is 0.141 which justifies it.
2. High density for BMI values in between 20-45.

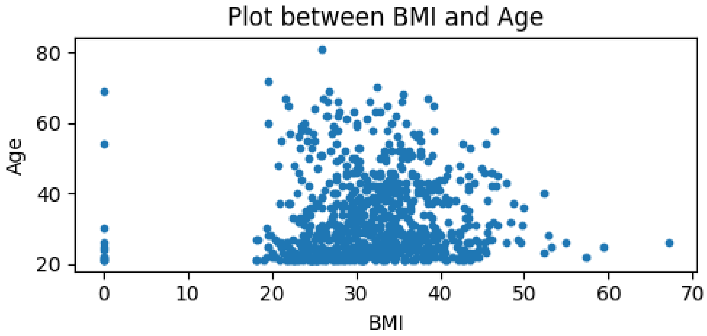

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. There is not much variation of age and BMI and they seem to be weekly correlated and correlation is 0.036 which justifies it.
2. High density for BMI values in between 20-45.

3 a.

Table 3 Correlation coefficient value computed between age and all other attributes

S. No.	Attributes	Correlation Coefficient Value
1	pregs	0.544
2	plas	0.264
3	pres (in mm Hg)	0.240
4	skin (in mm)	-0.114
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036
7	pedi	0.034
8	Age (in years)	1

Inferences:

1. Pregs, plags, pres are positively correlated with age , BMI and pedi are weakly correlated with age whereas skin and test are negatively correlated with age.
2. Pregs, plas, pres, BMI and pedi all increases with age whereas skin and test decreases with age.
3. Pregs is more correlated to age than any other parameter.

b.

Table 4 Correlation coefficient value computed between BMI and all other attributes

S. No.	Attributes	Correlation Coefficient Value
1	pregs	0.018
2	plas	0.221
3	pres (in mm Hg)	0.282
4	skin (in mm)	0.393
5	test (in mu U/mL)	0.198
6	BMI (in kg/m^2)	1
7	pedi	0.141
8	Age (in years)	0.036

Inferences:

1. All attributes are positively correlated with BMI whereas pregs and age are weekly correlated
2. All attributes increases with the increase of BMI.
3. Skin is more correlated to BMI than others

4 a.

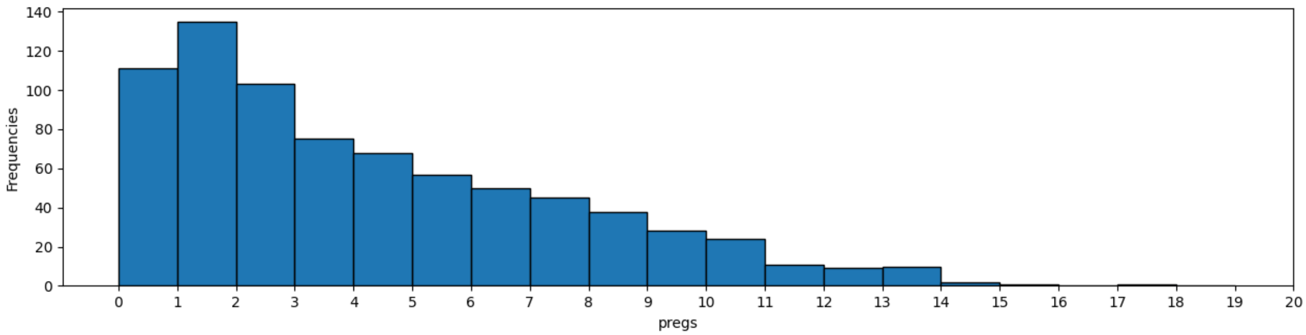

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. By observing the frequency values we can infer that most of the data values lies in 0-4 range.
2. Mode of the attribute pregs lies in the bin 1-2.

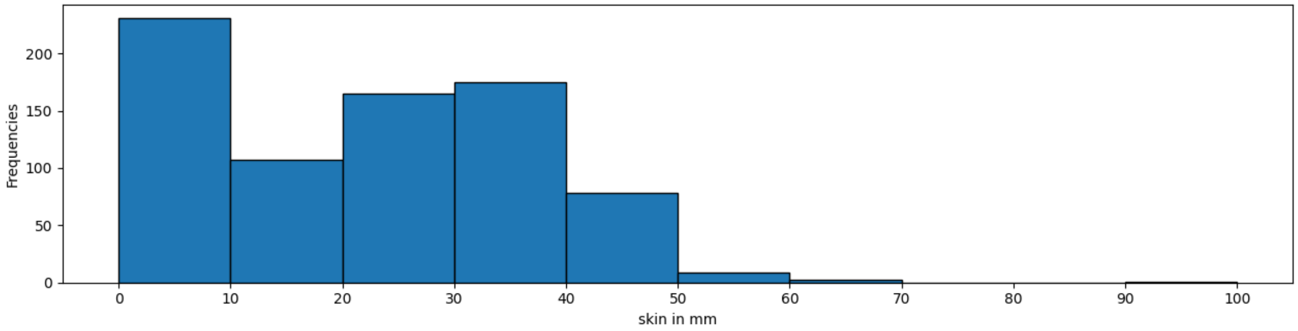

Figure 16 Histogram depiction of attribute skin

Inferences:

1. By observing the frequency values we can infer that most of the data values lies in 0-10 and 20-40 range.

2. Mode of the attribute skin lies in the bin 0-10.

5

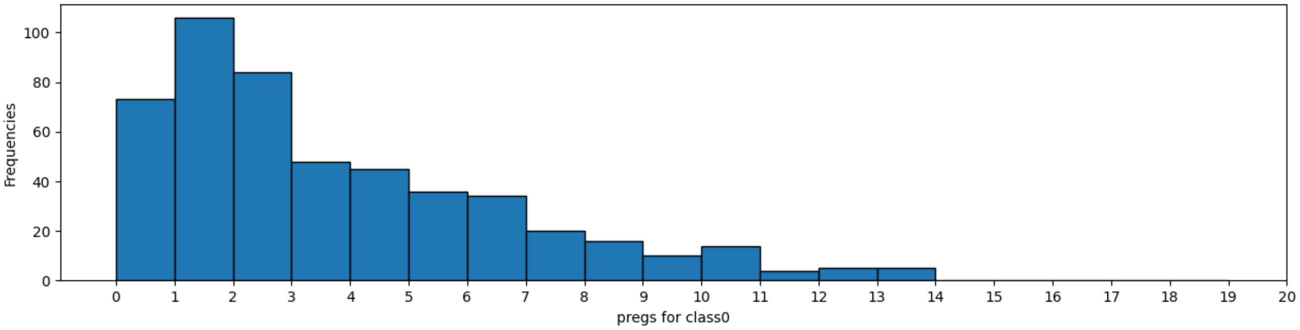

Figure 17 Histogram depiction of attribute pregs for class 0

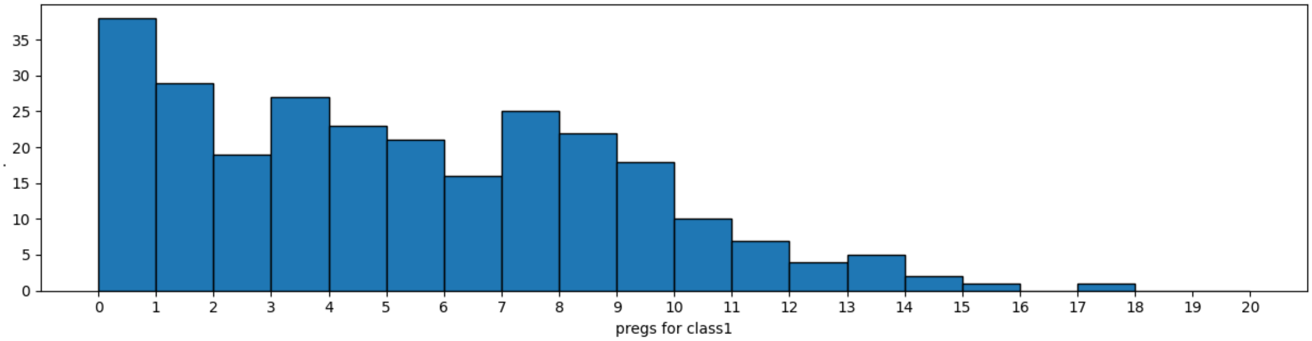

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. Mode of pregs lie in 1-2 bin for class 0 and in 0-1 bin for class 1.
2. Frequencies for class 0 is greater than the frequencies for class 1 and class 1 seems to be more evenly distributed than class 0.

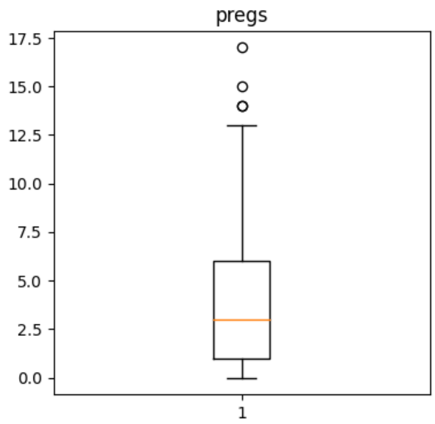

Figure 19 Boxplot for attribute pregs

Inferences:

1. There are 3 outliers above Q3 and their values are greater than 13.5 with highest value 17.
2. IQR = Q3-Q1=6-1=5
3. Variance (11.254) seems to be high for these values.
4. The data seems to be positively skewed (skewed right) as median is 3.0 but the mean of Q1 and Q3 is 3.5.
5. Q1 = 1.0 and the lower whisker is very short hence the values above Q1 have higher variation.

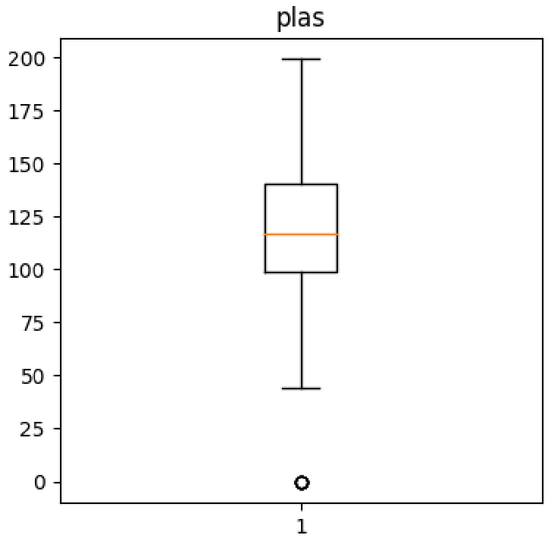

Figure 20 Boxplot for attribute plas

Inferences:

1. There is one outlier below Q1 whose value is 0.
2. IQR = Q3-Q1=140.25-99=41.25
3. Variance is 1022.249 which is high for these values.
4. Data seems to be quite unskewed with median=117 approximately equal to the mean of Q1 and Q3 which is = 119.625.
5. Q1 = 99 and there is one outlier below it , we can say that data is approximately normal distributed with uniform spreading.

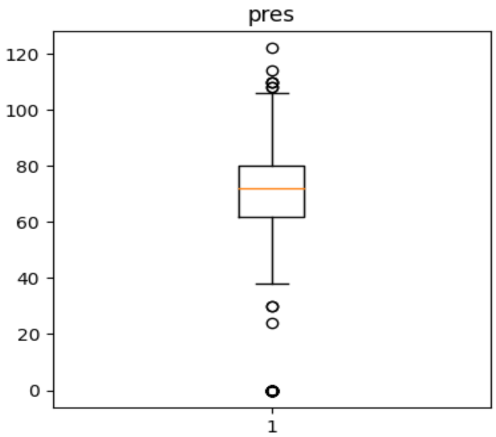

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are many outliers which lie on both sides of the plot with min =0 and max =122.
2. IQR=Q3-Q1=80-62=18.
3. Variance is 374.16 which is quite appreciable for these values.
4. Data seems to be symmetrical since median =72 which is equal to the mean of Q1 and Q3 which is 71.
5. Q1=62, values seems to be equally distributed above and below it.

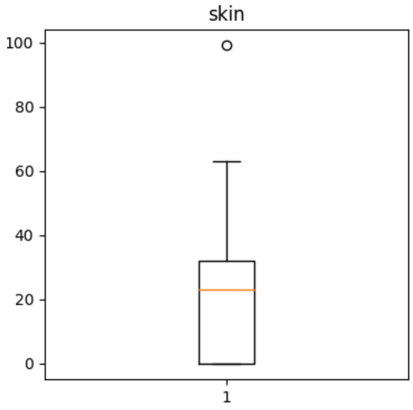

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. There is one outlier above Q3 whose value is 99.
2. IQR=Q3-Q1=32-0=32.
3. Variance = 254.141 which is high for these values.
4. Data seems to be negatively skewed(left skewed) since median = 23 is higher than the mean(16) of Q1 and Q3.
5. Q1 = 0.0, it doesn’t have any value below it.

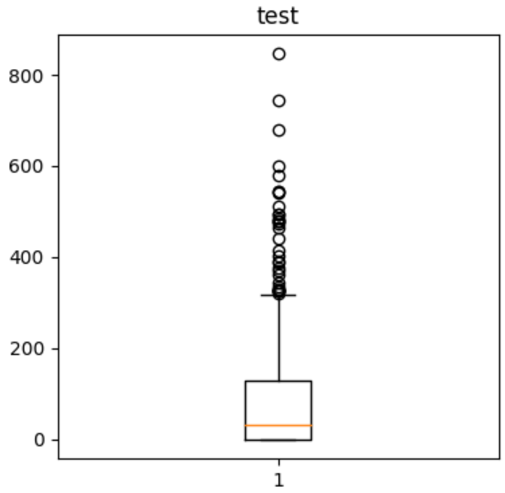

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. There are multiple outliers above Q3 with highest value = 846.
2. IQR = Q3-Q1 = 127.25-0=127.25
3. Variance is 13263.887 which shows that our data is very much spread.
4. Median is 30.5 which is lower than the mean of Q1 and Q3(63.125) hence data is positively skewed (right skewed).
5. Q1=0, and there is no value below it.

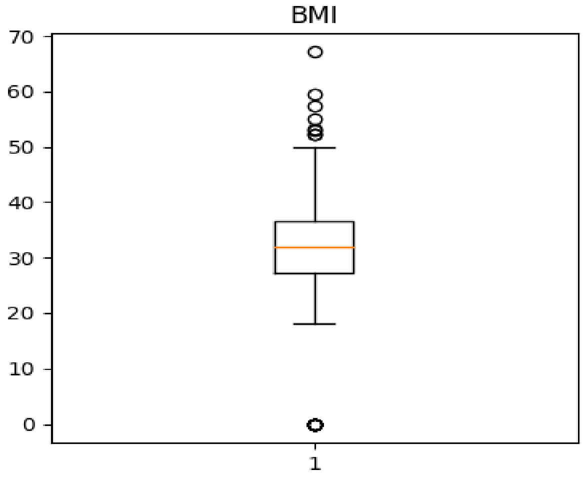

Figure 24 Boxplot for attribute BMI (in kg/m²)

Inferences:

1. Multiple number of outliers are above Q3 with highest value = 67.1 and some are below Q1 with lowest value = 0.
2. IQR=Q3-Q1=36.6-27.3=9.3.
3. Variance is 62.08 which is quite low.
4. Median = 32 which is equal to the mean(31.95) of Q3 and Q1 hence data is symmetrical.
5. Q1=27.3 and data seems to be evenly distributed.

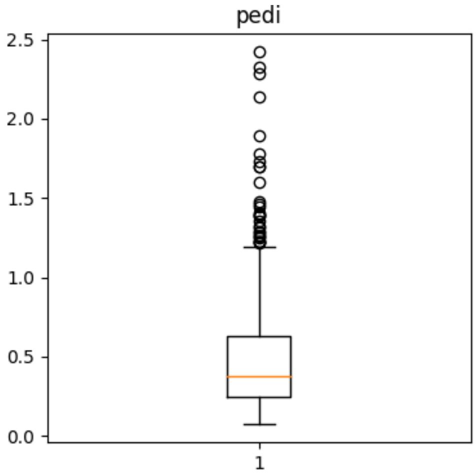

Figure 25 Boxplot for attribute pedi

Inferences:

1. There are multiple number of outliers above Q3 with the highest value = 2.42
2. IQR = Q3-Q1=0.626 – 0.244 = 0.382.
3. Variance is 0.110 which is quite low for these values
4. Median = 0.375 which is less than the mean (0.435) hence the data is positively skewed(right skewed).
5. Q1=0.244 and the whisker is short hence less data below it.

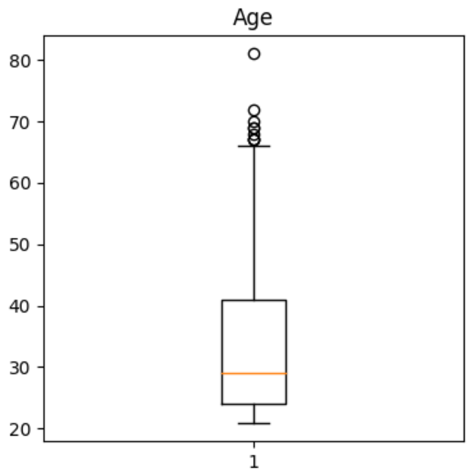

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. There are many outliers above Q3 with highest value = 81
2. IQR = Q3-Q1=41-24=17
3. Variance is 138.123 is high for these values
4. Median = 29 which is lower than the mean(32.5) of Q1 and Q3 so the data is positively skewed(right skewed).