IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – I

Data visualization and statistics from data

Student's Name: Arpit Singh **Mobile No: 6265104315**

Roll Number: B20084 **Branch: CSE**

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	3.8450	3	1	0	17	3.369
2	plas	120.894	117	99 100	0	199	31.972
3	pres (in mm Hg)	69.105	72	70	0	122	19.355
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)	31.992	32	32	0	67.1	7.884
7	pedi	0.471	0.372	0.254 0.258	0.078	2.42	0.331
8	Age (in years)	33.240	29	22	21	81	11.760

Inferences:

1. From the above table in pres, BMI attribute value of standard deviation is relatively small therefore the value of mean, mode and median is relatively closer to each other than other attributes.

2 a.

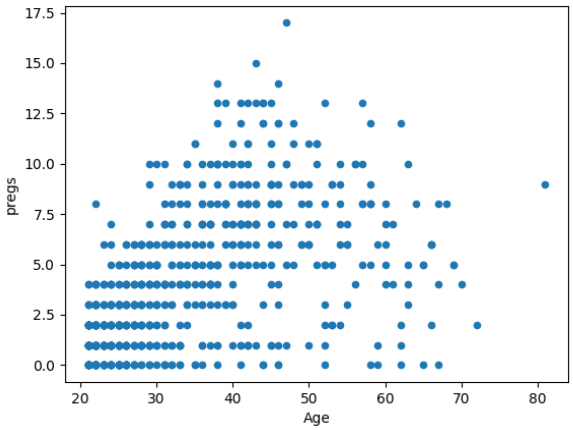


Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. In the scatter plot between age and pregs as the value of age is increasing we can observe an increase in no. of pregs. So, we can say there is positive correlation between the attributes.
2. Density of points in the scatter plot is more towards the origin of the plot i.e., at age less than 40 years and pregs less than 6.

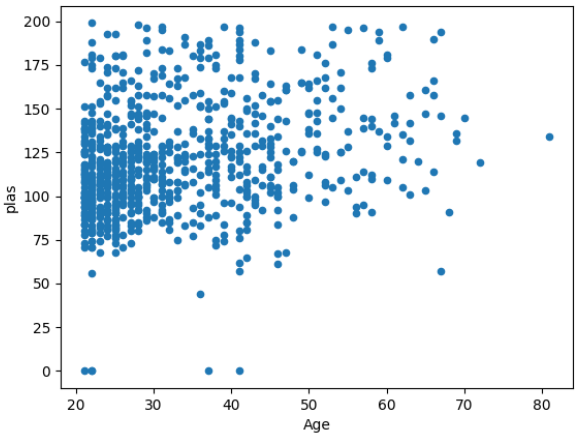


Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. In the scatter plot we can observe that data points are clustered toward the y axis and no increasing and decreasing relation can be observed, which means the plas is not correlated to age attribute.
2. Density of the points is above 75 in plas and less than 50 years in age.

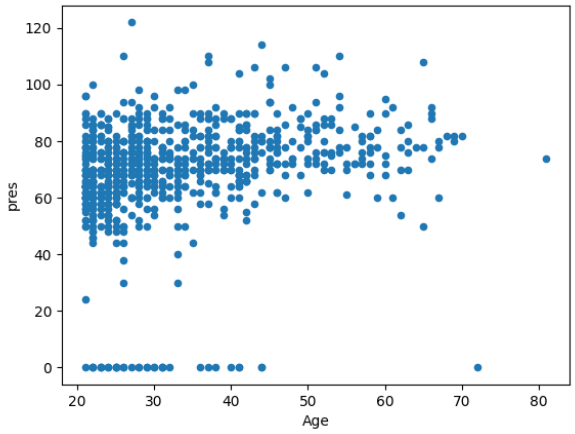


Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. In the scatter plot between age and pres we can observe that data points are clustered toward the y axis, above 40 mm Hg and no increasing and decreasing relation can be observed, which means the pres is not correlated to age attribute.
2. The density of the points are mainly between 50 mm Hg to 90 mm Hg.

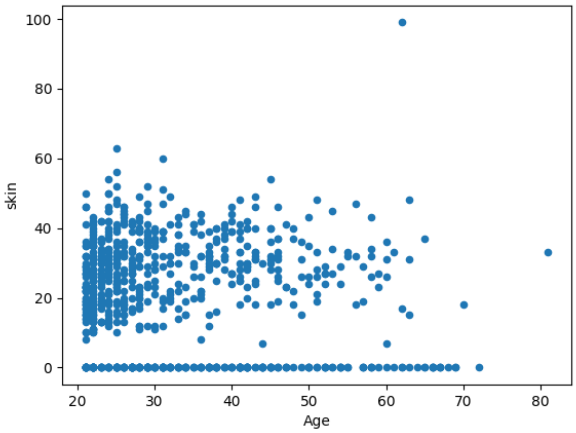


Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. In the scatter plot we can observe that data points are clustered toward the y axis and skin of 0 mm for all ages is observed. As there is no proper trend is there which means plas is not correlated to age.
2. density of points are at 0 mm skin and less than 40 years of age above 10 mm skin.

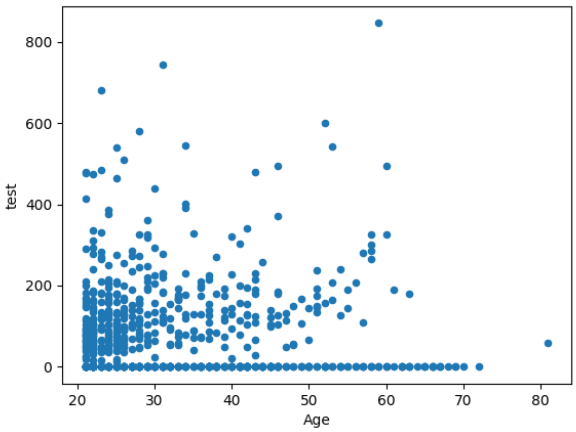


Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. In the scatter plot between age and test we can observe a decreasing trend in the value of test as age increases. So, attributes are negatively correlated.
2. density of the points are below 200 mm U/ml

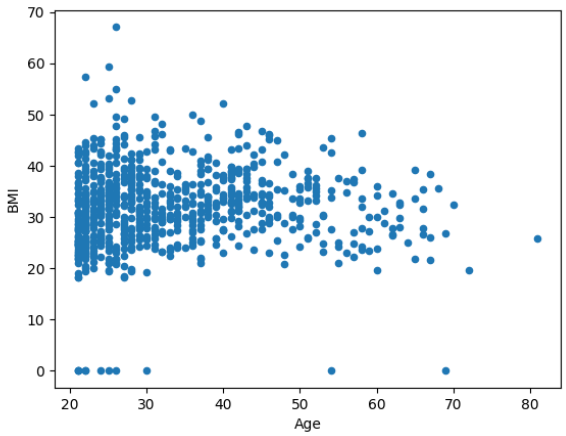


Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. In the scatter plot between age and BMI, the attribute doesn't follow any increasing or decreasing trend. So, they are not correlated with each other.
2. The density of points is above 20 kg/m2.

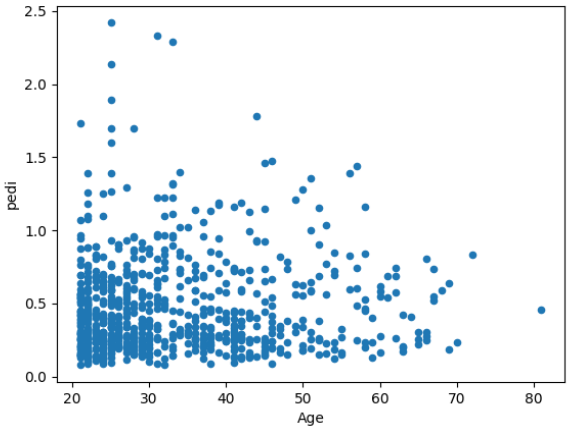


Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. In the scatter plot between age and pres we can observe that data points are clustered toward the origin, below 1 pedi and no increasing and decreasing relation can be observed.
2. density of point is more below 1 pedi and between 20 to 50 years of age.

b.

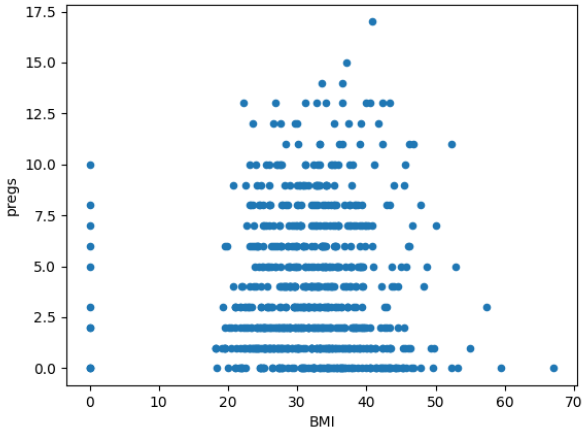


Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. There is not any increasing or decreasing trend between the attributes which means they are not correlated.
2. The density of points is between the BMI of 20 to 50 kg/m2 .

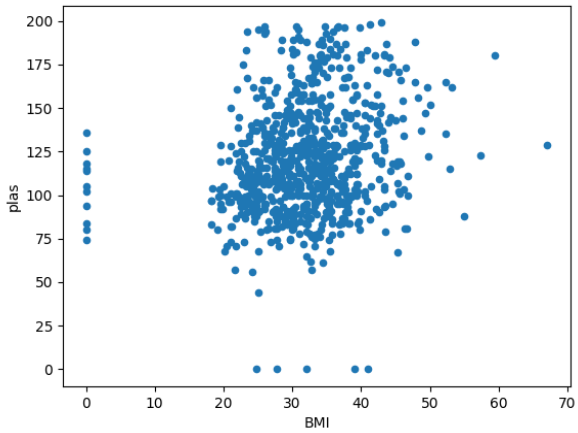


Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. data points are clustered at one region and no increasing or decreasing trend can be observed so, attributes are not correlated to each other.
2. The density of points is above 75 plas and between 20 to 50 kg/m2 BMI.

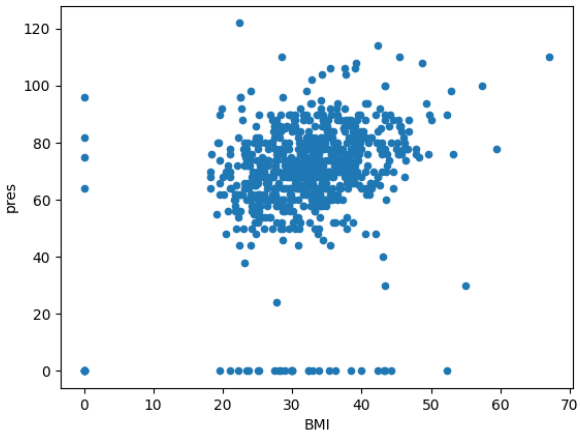


Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. data points are clustered at one region and no increasing or decreasing trend can be observed so, BMI and pres are not correlated to each other.

2. The density of the plot is between 40 to 100 mm Hg pres and between 20 to 50 kg/m2 BMI.

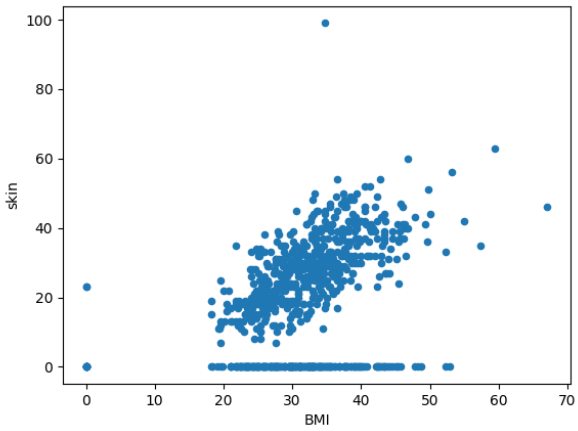


Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. In the plot we can observe an increasing trend in the value of skin as the bmi increases so attributes are positively correlated to each other.
2. density of points is between 20 to 50 kg/m2 BMI.

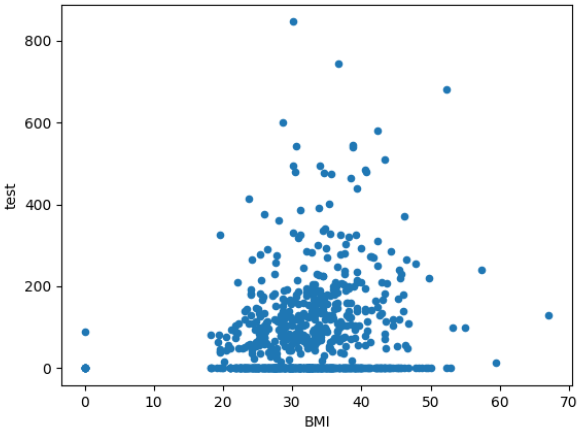


Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. data points are clustered at one region and the test has 0 value for BMI at a region of high density of data points. no increasing or decreasing trend can be observed so, attributes are not correlated to each other.
2. density of points is between 20 to 50 kg/m2 BMI and below 200 mm U/ml test.

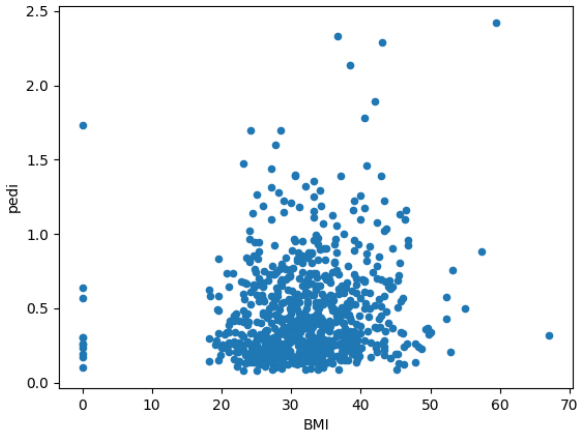


Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. data points are clustered at region between 20 to 50 kg/m2 BMI and no increasing or decreasing trend can be observed so attributes are not correlated to each other.
2. density of points is between 20 to 50 kg/m2 BMI and below 1 pedi.

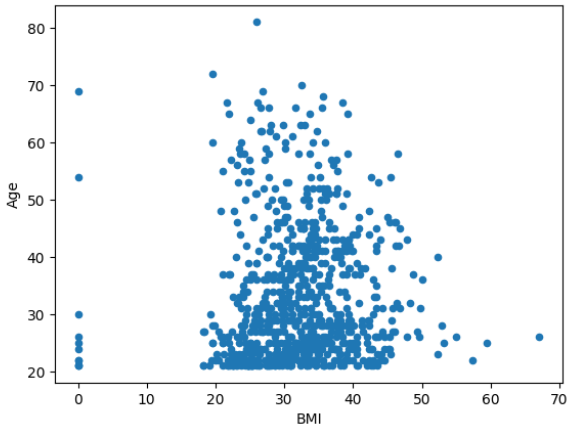


Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. In the scatter plot between age and BMI no trend can be observed as age increases or decreases so they are not correlated.
2. density of points is between 20 to 45 kg/m2 BMI.

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	0.263
3	pres (in mm Hg)	0.239
4	skin (in mm)	-0.113
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036
7	pedi	0.033
8	Age (in years)	0.999

Inferences:

1. Age is moderately correlated with pregs and weakly correlated with plas and pres and with all other attributes it is not correlated.
2. From the sign of correlation coefficient value, we can infer that if the sign of correlation is positive with increase or decrease in age each of the attributes will increase or decrease respectively and if the sign is negative with increase or decrease in age each of the attributes will decrease or increase respectively.
3. From the plot we can observe that pregs is positively correlated with Age and plot with other attributes, no significant trend can be observed as its correlation coefficient is very small.

b.

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

S. No.	Attributes	Correlation Coefficient Value
1	pregs	0.017
2	plas	0.221
3	pres (in mm Hg)	0.281
4	skin (in mm)	0.392
5	test (in mu U/mL)	0.197
6	BMI (in kg/m^2)	1.0
7	pedi	0.140
8	Age (in years)	0.036

Inferences:

1. BMI is moderately correlated with skin and weakly correlated with plas, pres, test and pedi and with pregs and age it is not correlated.
2. From the sign of correlation coefficient value, we can infer that if the sign of correlation is positive with increase or decrease in BMI each of the attributes will increase or decrease respectively and if the sign is negative with increase or decrease in BMI each of the attributes will decrease or increase respectively.
3. From the plot we can observe that skin is positively correlated with BMi and plot with other attributes no significant trend can be observed as its correlation coefficient is very small.

4 a.

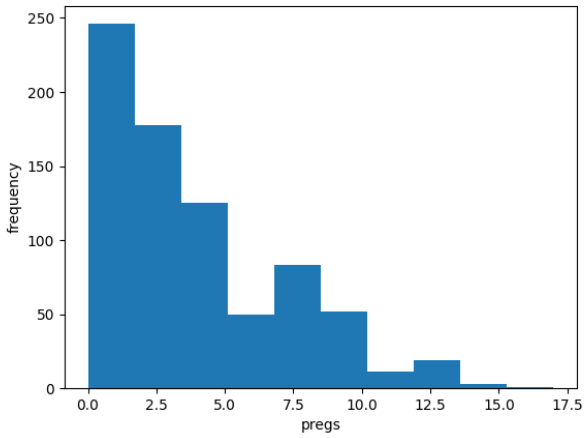


Figure 15 Histogram depiction of attribute pregs

Inferences:

1. In the range of 0-2 pregs frequency of bin is really high and as values of pregs increases frequency of the bins decreases its increases in bin of range 6-8 pregs but again starts decreasing and become almost 0 at bin of range 16-18 pregs.
2. mode of pregs will lie in the bin of value 0-2 pregs as it has the highest frequency.

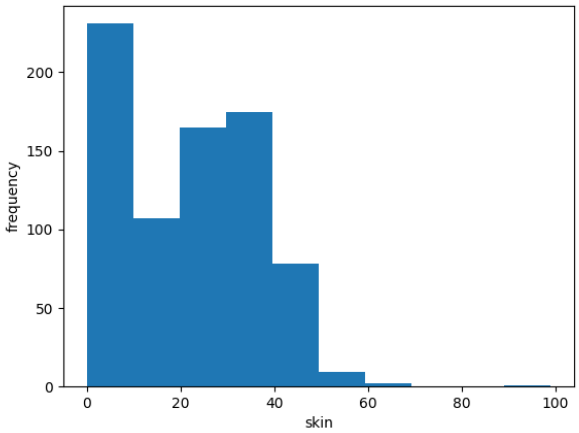


Figure 16 Histogram depiction of attribute skin

Inferences:

1. In the bin range of 0-10 the value of frequency is really high. It decreases abruptly in the next bin but again becomes constant at frequency of 160 for range 20-40 mm skin and decreases rapidly as value of skin increases.
2. mode of pregs will lie in the bin of value 0-10 mm of skin as it has the highest frequency.

5

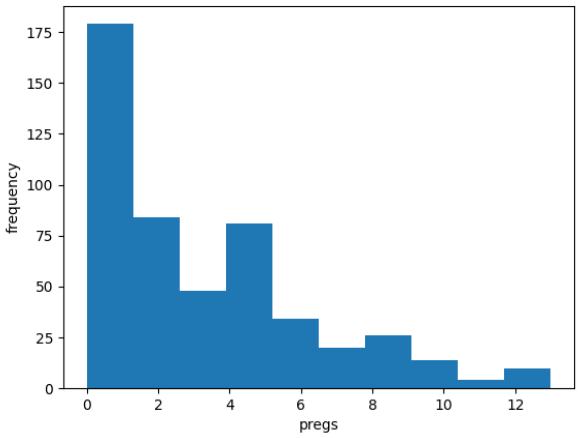


Figure 17 Histogram depiction of attribute pregs for class 0

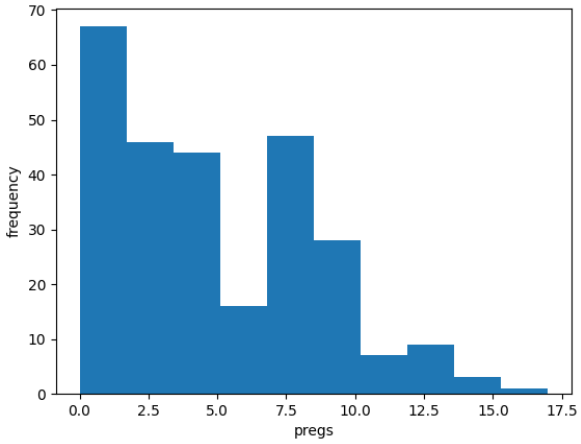


Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. mode of pregs will lie in the bin of value 0-1 pregs for class of 0 and 0-2 pregs for class of 1 as it has the highest frequency.
2. frequency of class 1 is higher than class of 0 in range of 0-5 and in range of 5-6 both have low frequency. In the range above 6 pregs class 1 have rapid decrease in frequency whereas class 0 have almost constant frequency.

6

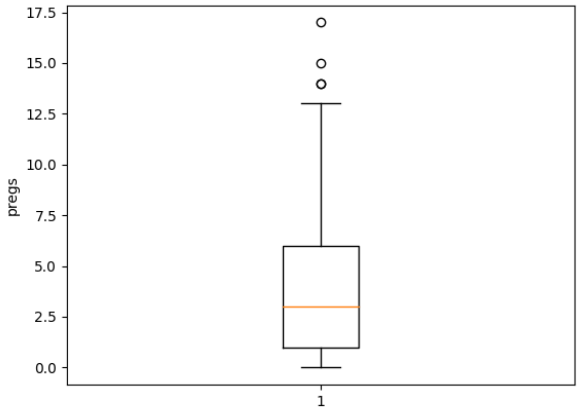


Figure 19 Boxplot for attribute pregs

Inferences:

1. There are 3 outliers and their values lie between 14 to 17 pregs.
2. Inter quartile range is from 1 to 6 pregs.
3. The Value of attribute pregs varies from 0 to 13 pregs with 3 outliers.
4. The data is right skewed.
5. values obtained in the box plot matches with the values obtained in the q1.

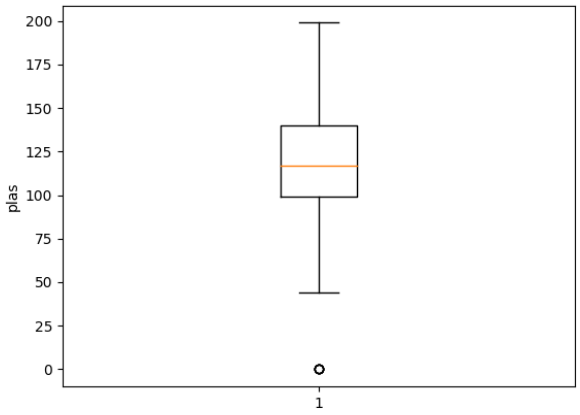


Figure 20 Boxplot for attribute plas

Inferences:

1. There are 1 outlier and its value is 0.
2. Inter quartile range is from 100 to 140 plas.
3. The Value of attribute pregs varies from 45 to 200 plas with 1 outlier at 0.
4. The data is right skewed.
5. values obtained in the box plot matches with the values obtained in the q1.

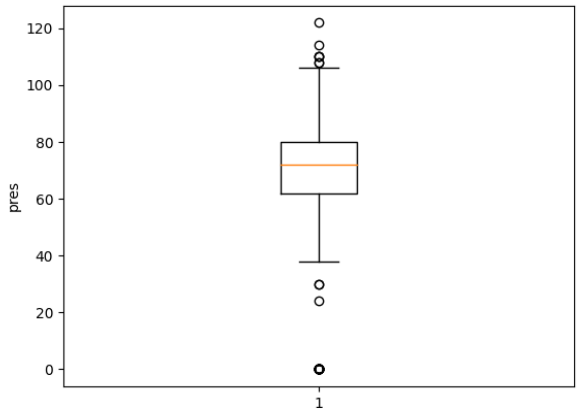


Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are 7 outliers 4 above and 3 below. their values lie between 110 to 125 mm hg pres and 0 to30 mm hg pres.
2. Inter quartile range is from 61 to 80 mm Hg pres.
3. The Value of attribute pregs varies from 38 to 106 mm hg pres with 7 outliers.
4. The data is symmetric.
5. values obtained in the box plot matches with the values obtained in the q1.

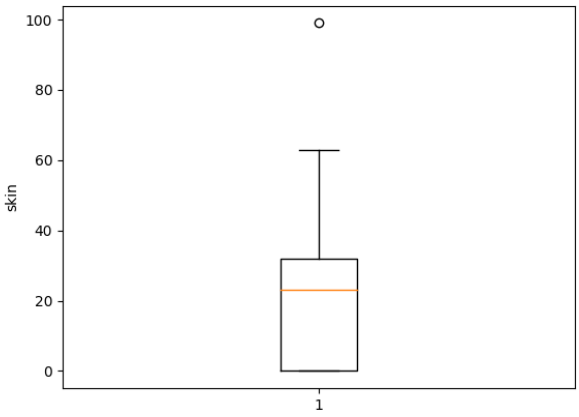
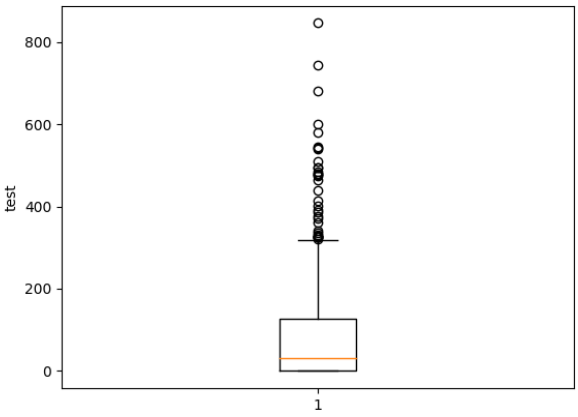


Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. there is 1 outlier and its value is 100mm skin.
2. Inter quartile range is from 0 to 31 mm skin.
3. The Value of attribute pregs varies from 0 to 62.5 mm of skin with 1 outlier at 100 mm of skin.
4. The data is left skewed.
5. values obtained in the box plot matches with the values obtained in the q1.



6.

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

Inferences:

1. There are multiple no. of outliers and their values are in the range of 250 to 900 mu U/ml.
2. Inter quartile range is from 0 to 125 mu U/ml test.
3. The Value of attribute pregs varies from 0 to 320 mu U/ml test with multiple no. of outliers.
4. The data is right skewed.
5. values obtained in the box plot matches with the values obtained in the q1.

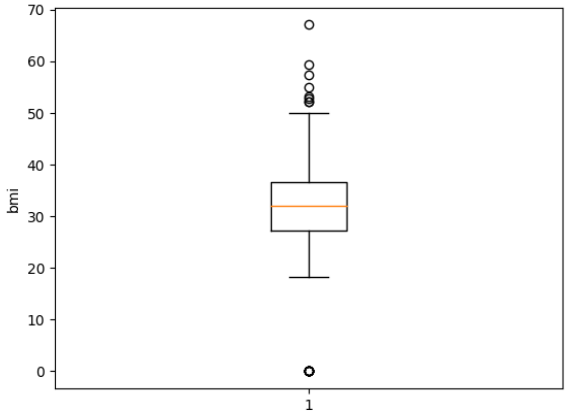


Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. There are multiple no. of outliers 1 below and rest of all above the limit and their values are in the range of 50 to 70 kg/m2 and 0 for outlier below the limit.
2. The Inter quartile range is from 27 to 36 kg/m2.
3. The Value of attribute pregs varies from 18 to 50 kg/m2 BMI with multiple no. of outliers.
4. The data is symmetric.
5. values obtained in the box plot matches with the values obtained in the q1.

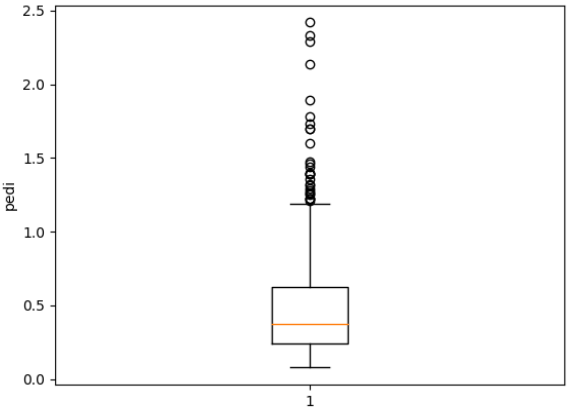


Figure 25 Boxplot for attribute pedi

Inferences:

1. There are multiple no. of outliers and their values are in the range of 1.2 to 2.4 pedi.
2. The Inter quartile range is from 0.25 to 0.6 pedi.
3. The Value of attribute pregs varies from 0.8 to 1.2 pedi with multiple outliers.
4. The data is right skewed.
5. values obtained in the box plot matches with the values obtained in the q1.

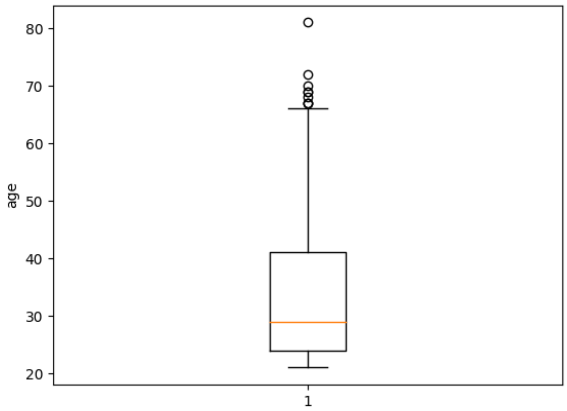


Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. box plots of age have 6 outliers and their values lie between 65 to 80 years of age .
2. The Inter quartile range is from 24 to 40 years of age.
3. The Value of attribute pregs varies from 20 to 65 years of age with 6 outliers.
4. The data is right skewed.
5. values obtained in the box plot matches with the values obtained in the q1.