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

Student's Name: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:EE**

1 Mean, median , minimum , maximum , mode and standard deviation.

S. No.	Attributes	Mean	Median	Mode	Min.	Max.	S.D.
1	pregs	3.845	3.000	1.000	0.000	17.000	3.370
2	plas	120.900	117.000	99.000	0.000	199.000	31.980
3	pres (in mm Hg)	69.105	72.000	70.000	0.000	122.000	19.360
4	skin (in mm)	20.540	23.000	0.000	0.000	99.000	15.960
5	test (in mu U/mL)	79.800	30.500	0.000	0.000	846.000	115.245
6	BMI (in kg/m^2)	31.100	32.000	32.000	0.000	67.100	7.890
7	pedi	0.500	0.380	0.254	0.078	2.420	0.331
8	Age (in years)	33.240	29.000	22.000	21.000	81.000	11.760

Inferences:

1. It can be seen that as standard deviation of attribute pedi (Diabetes pedigree function) approaches to 0 , the value of mean , median and mode also approaches to 0.
2. Since mean , median and mode approaches to 0 which means that for most cases the pedi value approaches 0.

2 a.

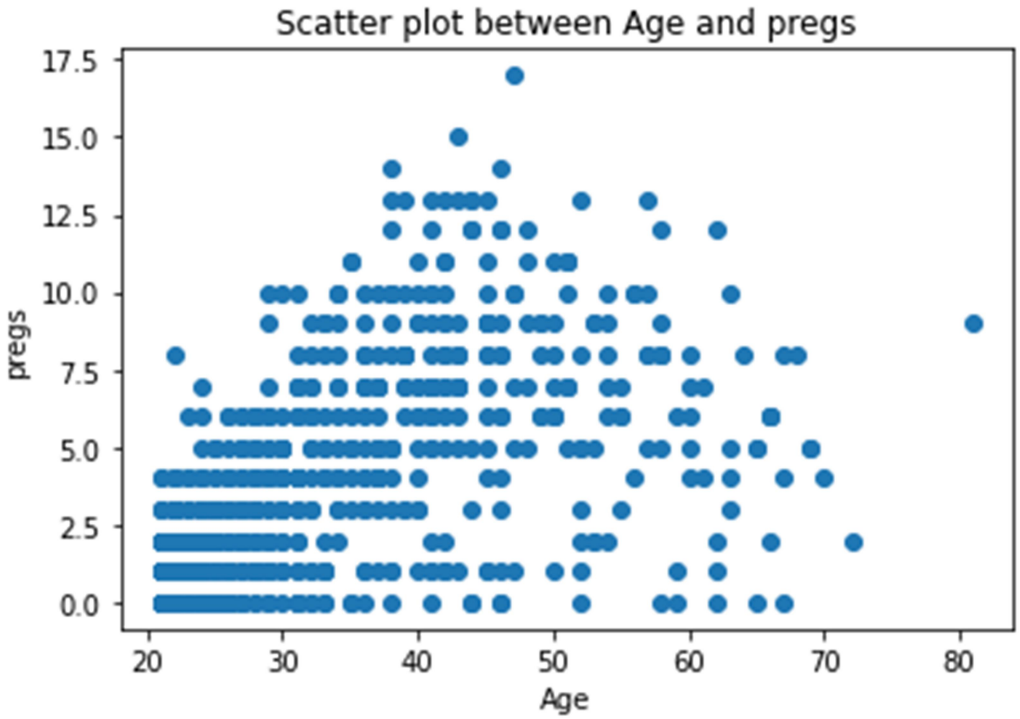

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. As the age of women increases , the number of times she's pregnant decreases.
2. The age group between 20 -30 years are having more pregnancy.
3. The average of number of times of pregnant for a women is in between 0 to 4 .
4. There is most probably one women who got pregnant at age of 80 years.

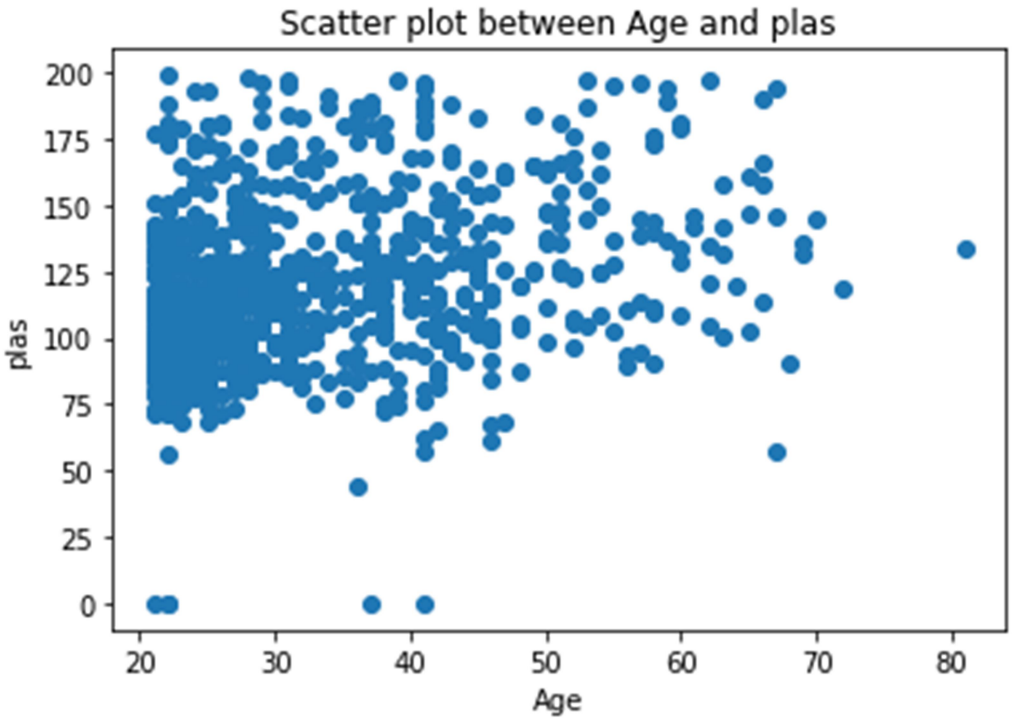

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. The age group between 20 to 30 years have concentration of plas more than any other age groups.
2. The concentration of plas is more in between 75 to 150.
3. The maximum concentration of plas in any women goes till upto 200.
4. The minimum concentration of plas in any women is found to be in betwee 50.

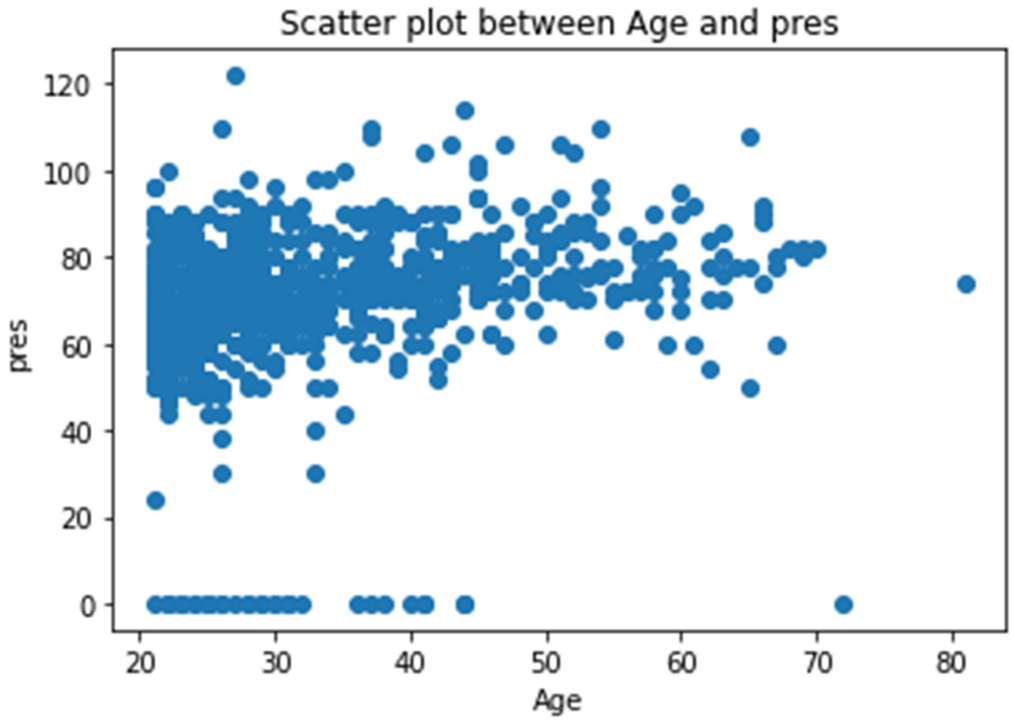

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. The age group 20 to 30 years have diastolic blood pressure level mostly in between 40 and 100 mm hg.
2. The diastolic blood pressure level for age groups between 50 to 60 year is in between 60 and 100 mm hg.
3. The minimum diastolic blood pressure level for any age groups lie in between 20 to 45 mm hg.

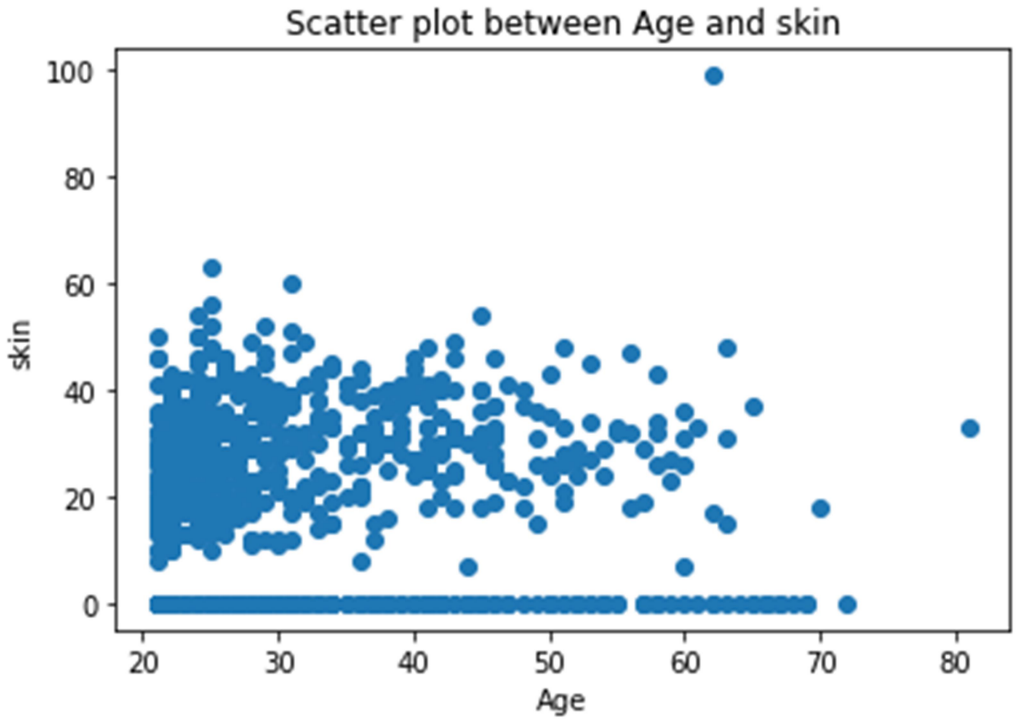

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. The value for triceps of fold thickness is mostly concentrated in between 5 to 45 mm for the age group 20 to 30 years.
2. For the age groups between 45 to 60 years , the value of triceps of fold thickness lies in between 20 to 60 mm.
3. The maximum value of triceps fold thickness goes upto 62mm.

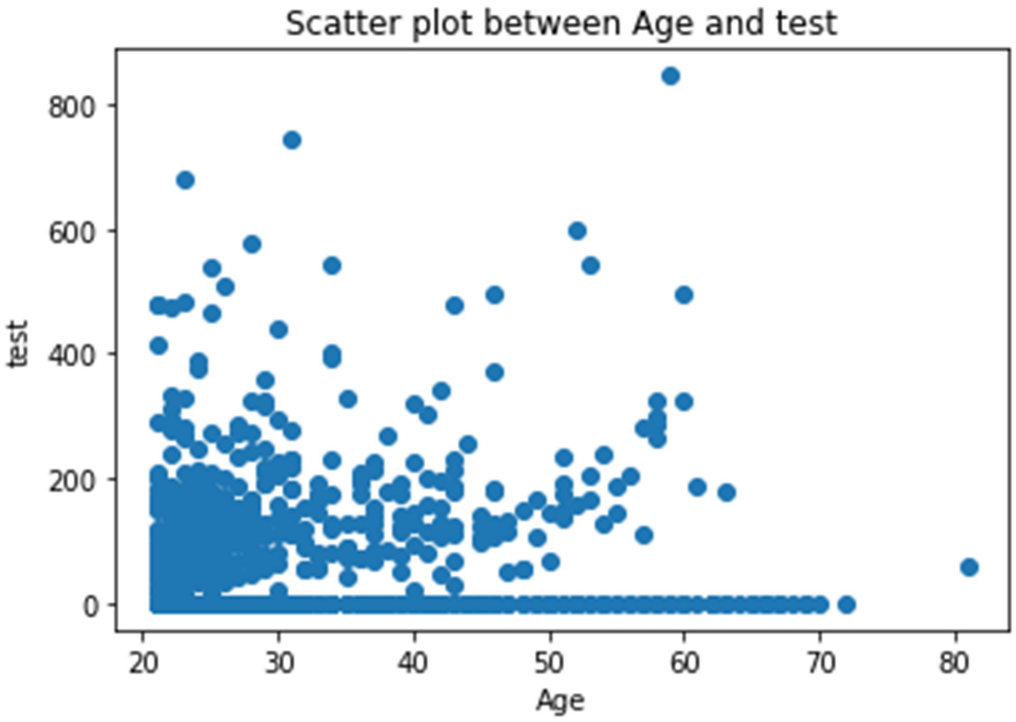

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. The 2-Hour serum insulin for age in between 20 to 30 years is mostly concentrated in between 10 to 200 mu U/mL .
2. The 2 hour serum insulin for age in between 30 to 45 years is mostly concentrated in between 50 to 200 mu U/mL .
3. Very few women shows 2 hour serum insulin in between 200 to 700 mu U/mL for all age groups.

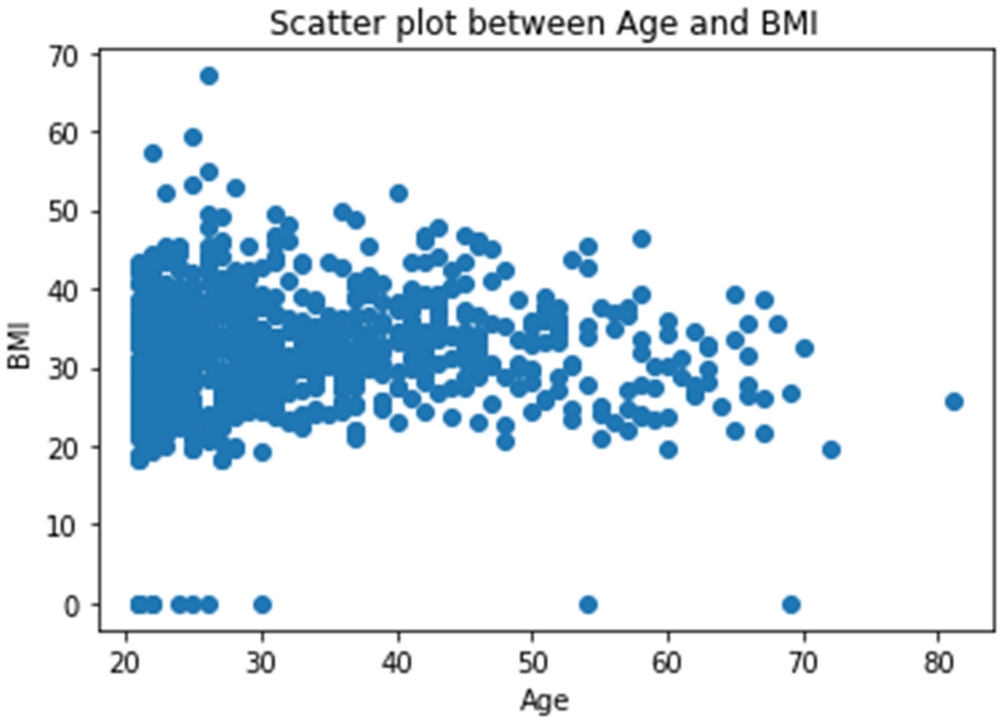

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. The BMI for age in between 20 to 70 years lies in between 20 to 45 kg/m^2.
2. The maximum value of BMI for the age group 20 to 30 years equals 69 in kg/m^2.

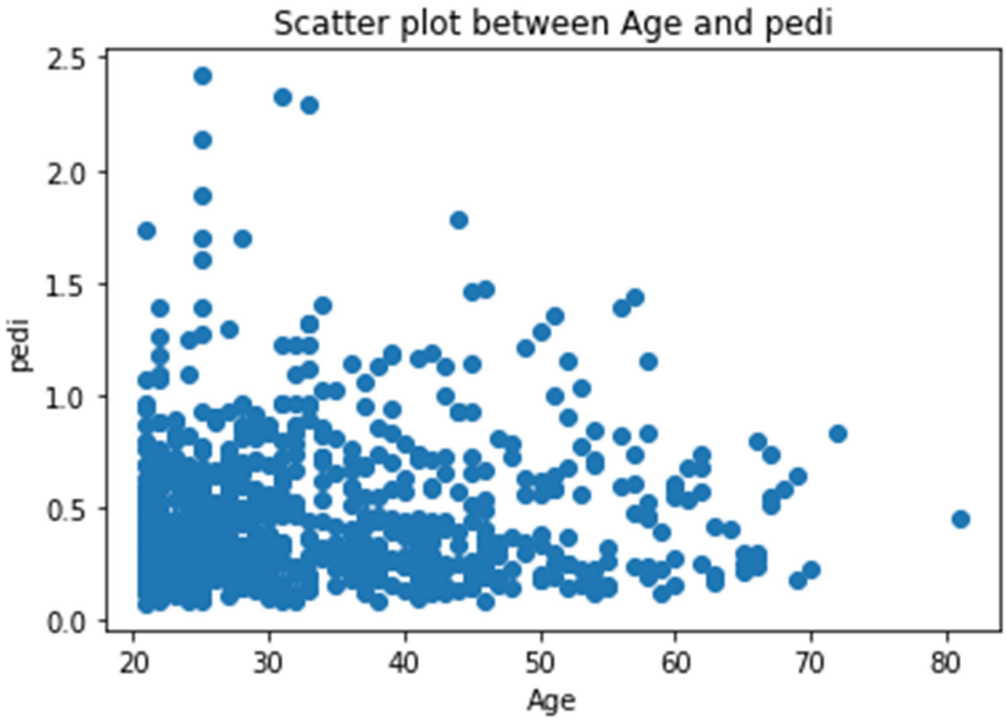

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. The age 20 to 30 years have diabetes pedigree function that lies in between 0.1 to 1.
2. The maximum value of diabetes pedigree function for the age group 20 to 30 years is 2.4 .
3. As the age increases the value of diabetes pedigree function decreases.

b.

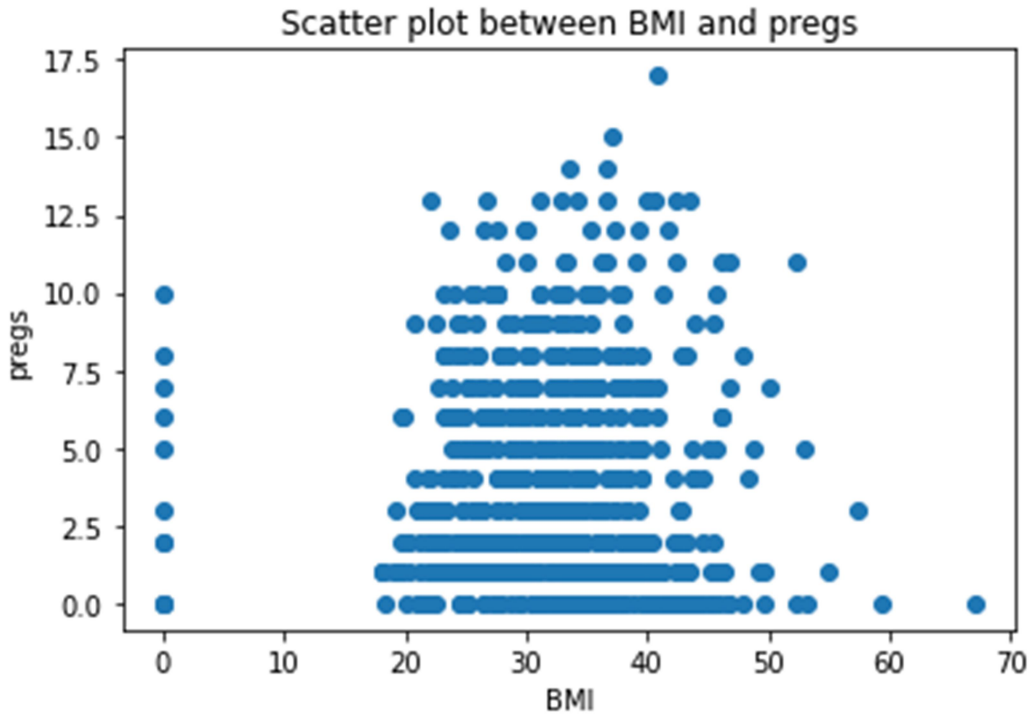

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. As the BMI value increases upto 42 kg/m^2 , the number of times pregnant value also increases upto 17.5 .
2. The number of times pregnant lies in between 0 to 15 when the bmi of women lies in between 20 to 50 kg/m^2.
3. The number of times pregnant mostly concentrated in between 0 to 10 when BMI lies in between 25 to 40 kg/m^2.

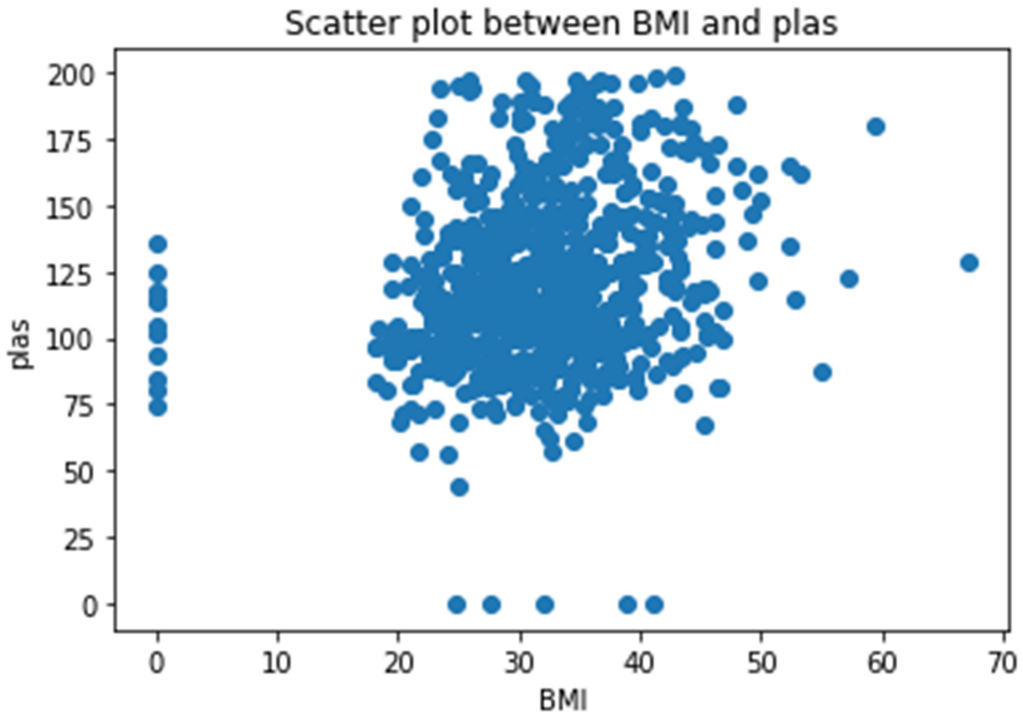

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. The value of plas lie in between 75 to 200 when BMI lies in between 20 to 45 kg/m^2.
2. The value of plas mostly concentrated in between 75 to 150 for the BMI value in between 25 to 36 kg/m^2.
3. The attribute BMI and plas Is uncorrelated.

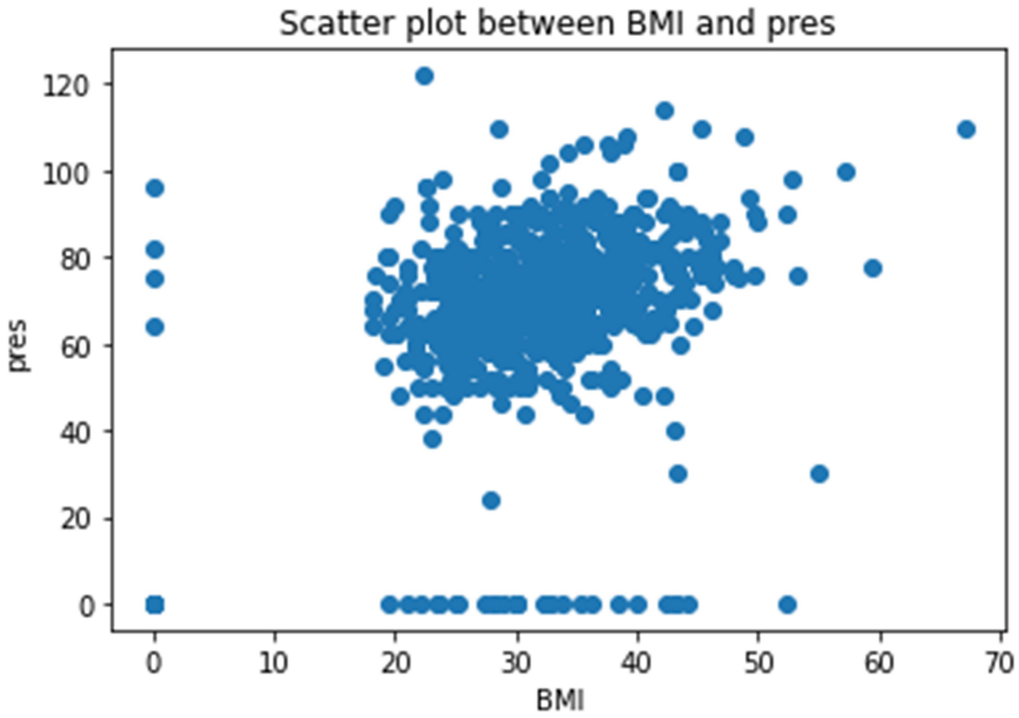

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. The attribute BMI is uncorrelated with attribute pres.
2. When the value of BMI lies in between 20 to 50 years then the value of Diastolic blood pressure lies in between 40 to 100 mm Hg.
3. The maximum value of Diastolic blood pressure equals 120 mm Hg .

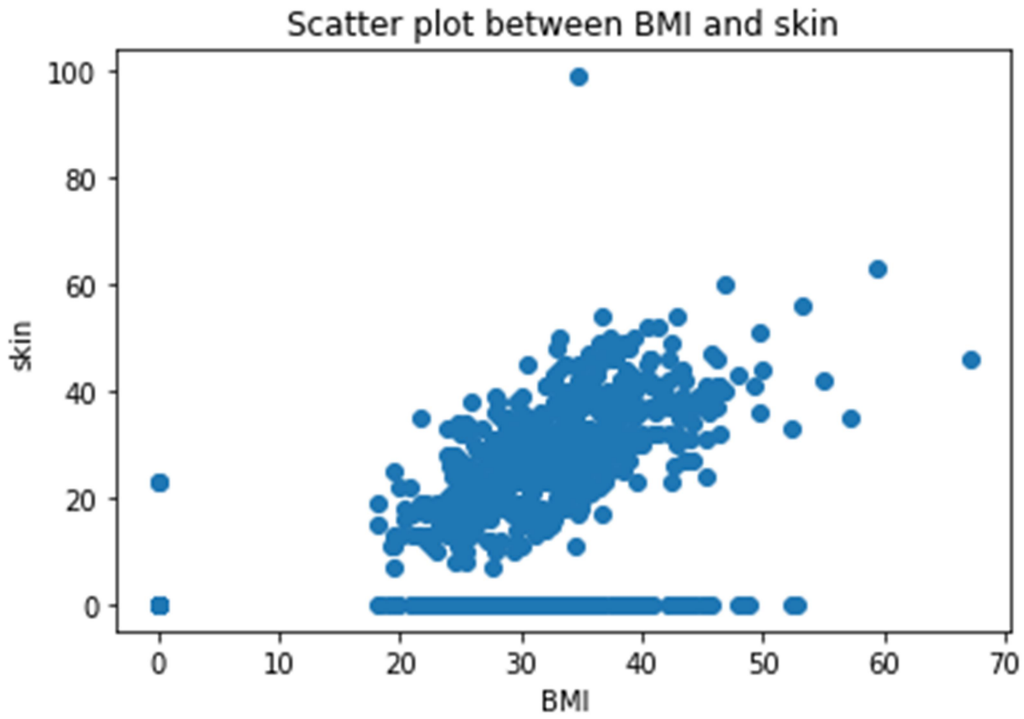

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. The attribute BMI is positively correlated with attribute skin. As value of BMI increases , value of skin increases.
2. The value of skin increases lies in between 20 to 40 mm when BMI lies in between 30 to 40 kg/m^2.
3. The maximum value of skin equals 100 mm when BMI equals 36 kg/m^2.

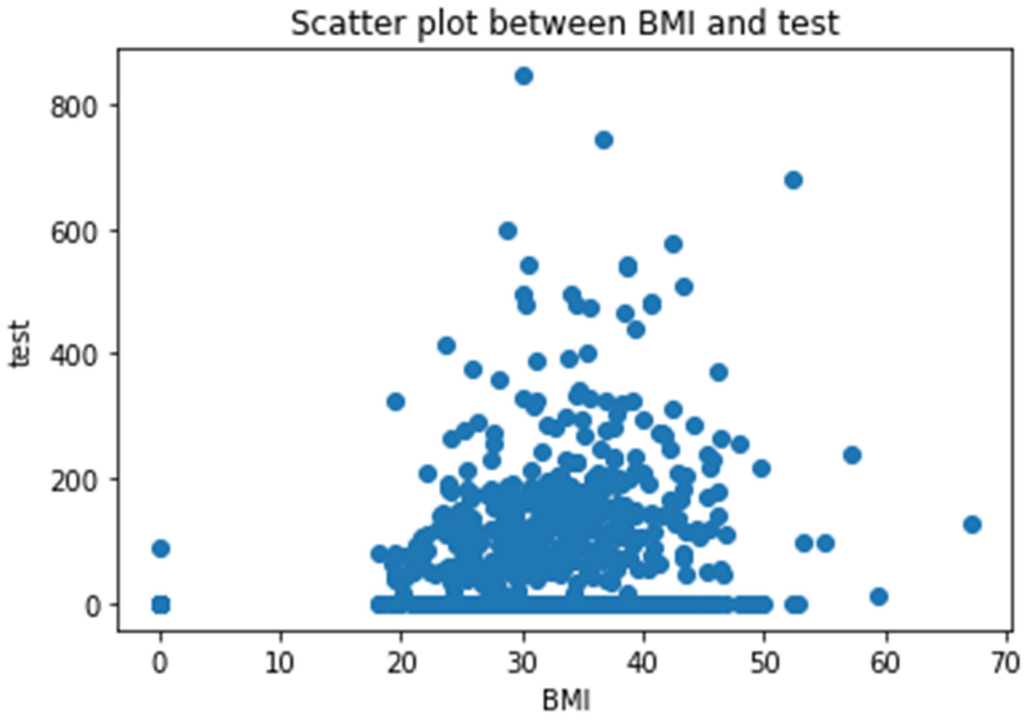

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. The value of test lies in between 10 to 200 U/mL when BMI lies in between 20 to 50 kg/m^2.
2. The maximum value of test lies in between 600 to 800 U/mL when BMI lies in between 20 to 55 kg/m^2.
3. The attribute test is uncorrelated with attribute BMI.

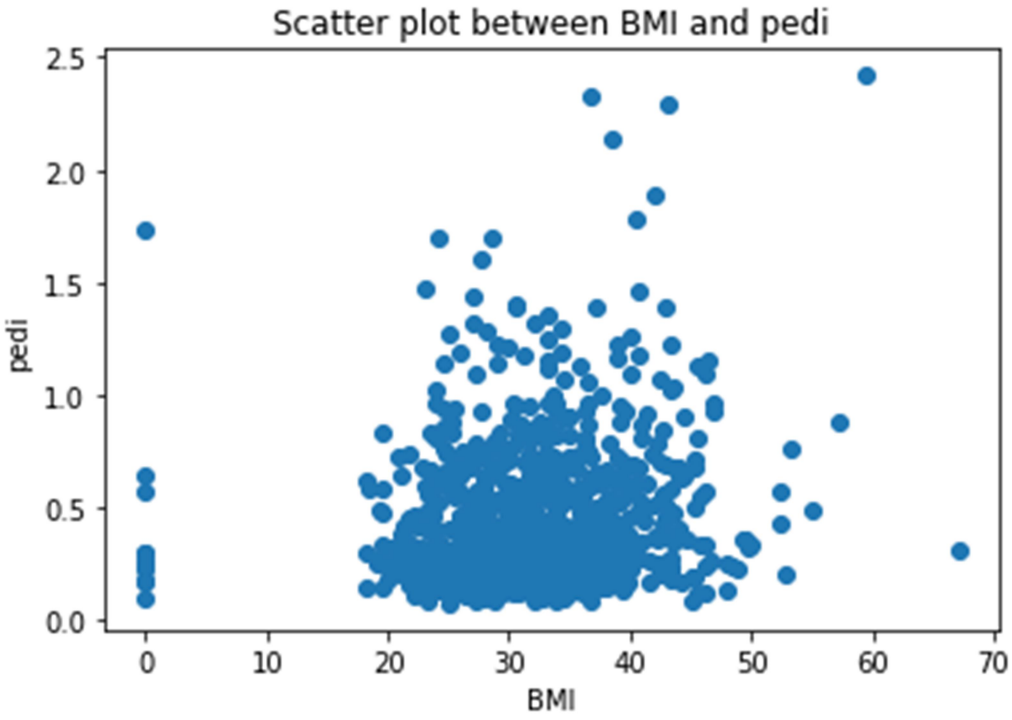

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. The value of pedi lies in between 0.1 to 1.5 when BMI lies in between 20 to 50 kg/m^2 .
2. The maximum value of pedi lies in between 1.8 to 2.4 when BMI lies in between 38 to 45 kg/m^2 .
3. The attribute pedi is uncorrelated with attribute BMI.

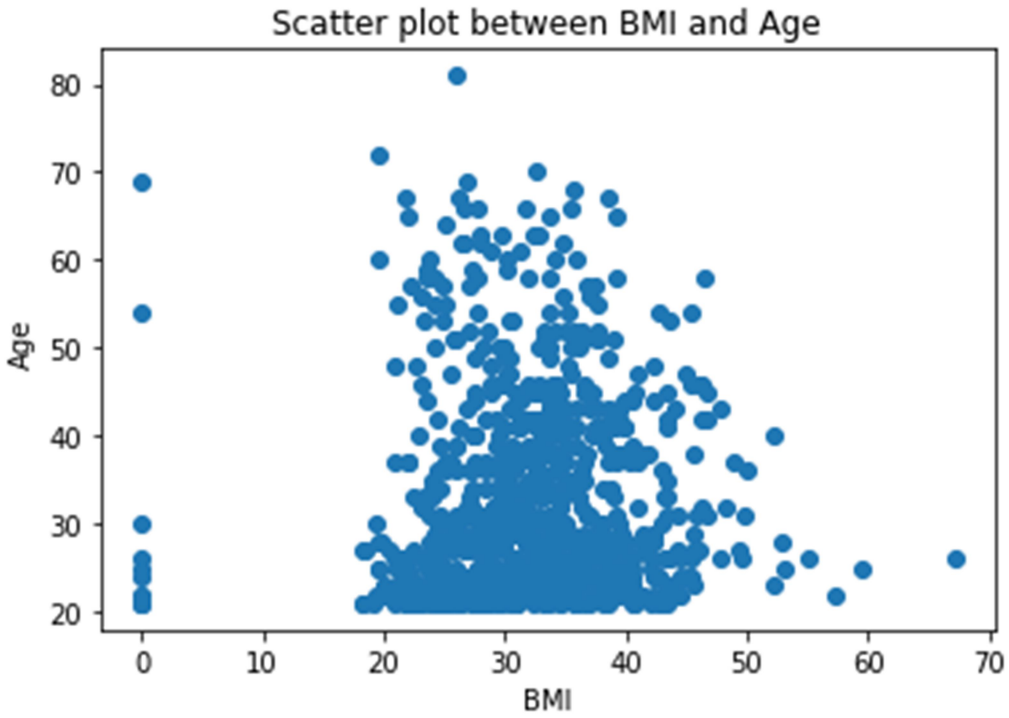

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. The value of age lies in between 20 to 70 years when BMI lies in between 20 to 45 kg/m^2.
2. The maximum age is 80 years when BMI lies in between 20 to 30 kg/m^2.
3. The attribute age is uncorrelated with attribute 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)	0.240
4	skin (in mm)	-0.113
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036
7	pedi	0.340

8	Age (in years)	1.000

Inferences:

1. Age and pregs are strongly correlated , Age and plas are weakely correlated , Age and pres are weakely correlated , Age and skin are negatively correlated , Age and test are negatively correlated , Age and BMI are moderately correlated , Age and pedi are moderately correlated , Age and Age are strongly correlated.
2. With increase in age – pregs increases readily , plas increases slowly , pres increases slowly , skin decreases , test decreases, BMI increases , pedi increases , Age increases .
3. When a line is drawn along the density of scatter plot and
   a. if that line has positive slope then it corresponds to the strong correlation values (like in Age vs Age , pregs vs age).
   b. If the line slope approaches to 0 , then it corresponds to uncorrelation (like age vs test , age vs BMI)
   c. If the line has negative slope than it corresponds to the negative correlation (like age vs skin and age vs test)

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)	0.281
4	skin (in mm)	0.400
5	test (in mu U/mL)	0.200
6	BMI (in kg/m^2)	1.000
7	pedi	0.140
8	Age (in years)	0.036

Inferences:

1. BMI and pregs are weakely correlated , BMI and plas are weakely correlated , BMI and pres are weakely correlated , BMI and skin are moderately correlated , BMI and test are weakely correlated , BMI and BMI are strongly correlated ,BMI and pedi are weakely correlated , BMI and Age are weakely correlated.
2. With increase in BMI – pregs increases slowly , plas increases slowly , pres increases slowly , skin increases moderately , test increases slowly, BMI increases highly , pedi increases slowly , Age increases slowly.
3. When a line is drawn along the density of scatter plot and
   a. if that line has positive slope then it corresponds to the strong correlation values (like in BMI vs plas , BMI vs pres , BMI vs skin , BMI vs BMI).
   b. If the line slope approaches to 0 , then it corresponds to uncorrelation (like age vs BMI , pregs vs BMI).

4 a.

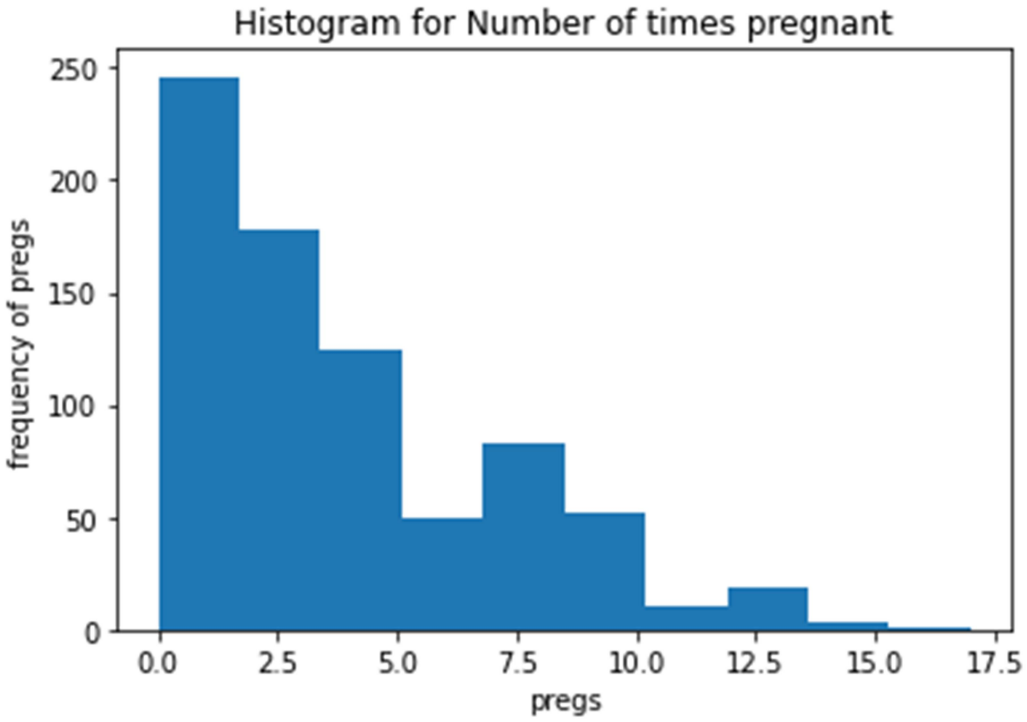

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. When no. of pregnant lies in between 0 to 2.5 the frequency of those corresponds to that no. of pregnant lies in between 0 to 250.
2. As no. of pregnant increases , the frequency for those decreases.
3. Mode for the number of pregnant is lie in between 0 to 2.5.
4. Maximum number of women have number of pregnancy in between 0 to 2.
5. Since the histogram is left skewed then average value of pregs is greater than the median.

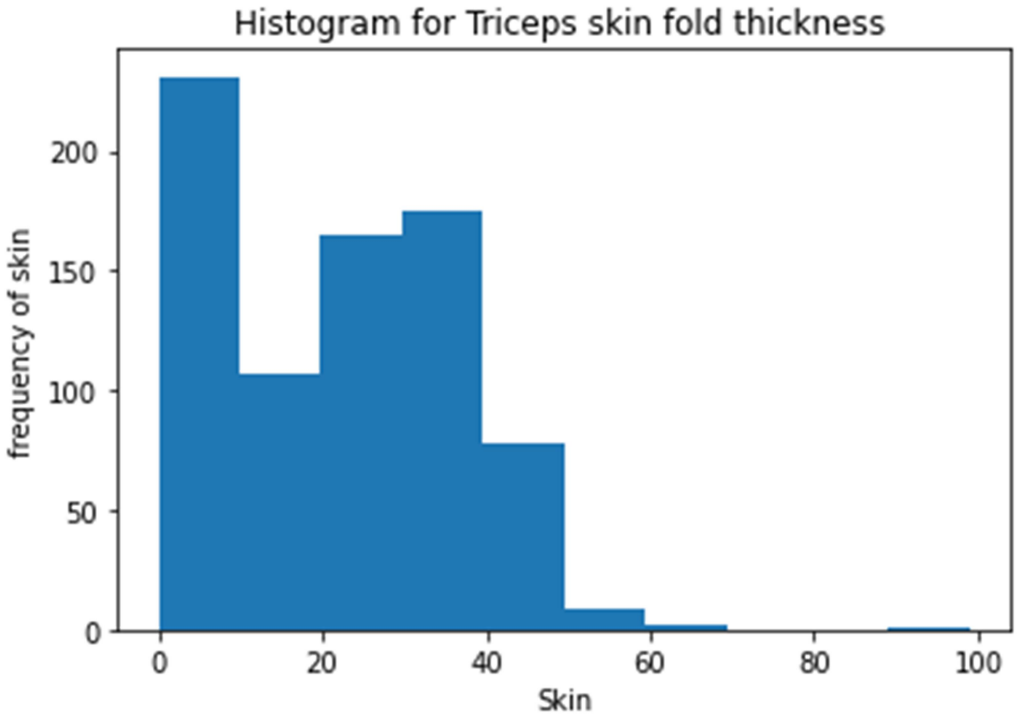

Figure 16 Histogram depiction of attribute skin

Inferences:

1. The mode of attribute skin is 240.
2. As the Triceps skin fold thickness increases , the number of woman for that corresponding thickness decreases.
3. Very few women have triceps skin fold thickness in beyond 60 mm.
4. Since the histogram is left skewed then average value of pregs is greater than the median.

5

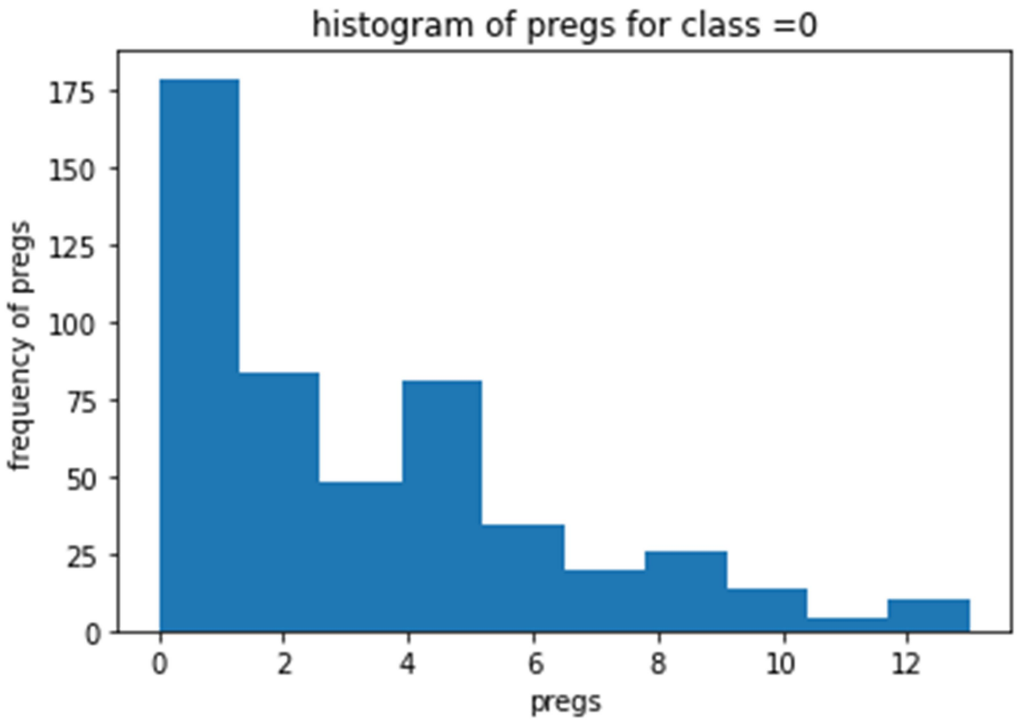

Figure 17 Histogram depiction of attribute pregs for class 0

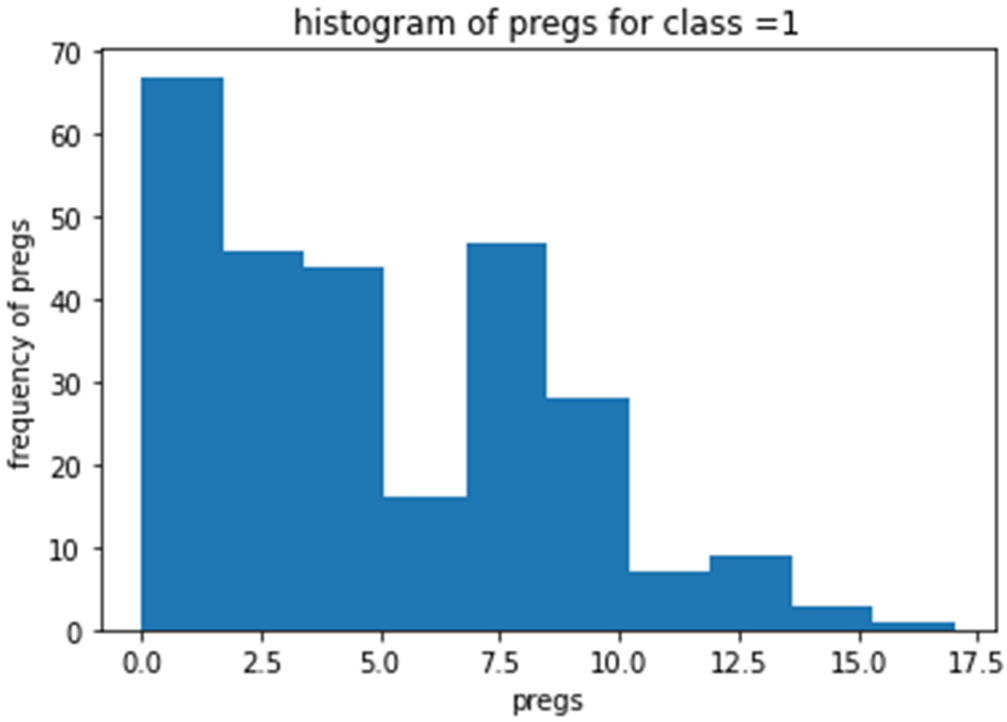

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. For class 0 the mode for pregs lies in the bin 0-2 and for the class 1 the mode for pregs lies in the bin 0-2.5.
2. For
   a. The frequency of pregs for class 0 (bin 0-2) is more than class 1 (bin 0-2.5).
   b. The frequency of pregs for class 0 (bin 2-4) is more than class 1(bin 2.5-5.0).
   c. The frequency of pregs for class 0 (bin 4-6) is more than class 1(bin 5-7.5)
   d. The frequency of pregs for class 0 (bin 6-8) is less than class 1(bin 7.5-10.0)
   e. The frequency of pregs for class 0 (bin 8-10) is more than class 1(bin 10-12.5)
   f. The frequency of pregs for class 0 (bin 10-12) is more than class 1(bin 12.5-15.0)

6

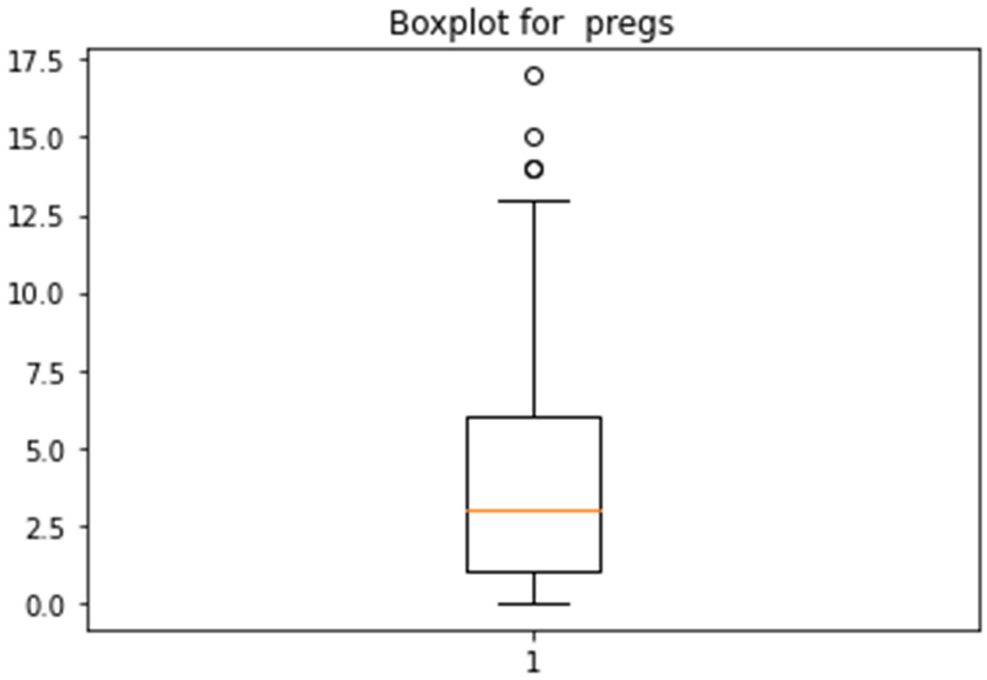

Figure 19 Boxplot for attribute pregs

Inferences:

1. There are very few number of pregnant which lies in between 13 to 17.5 (as they are outliers in the boxplot).
2. 50 percent of the women comes under the group of 1 to 6 number of pregnants (as they occurs in the interquartile).
3. 25 percent of the pregs are 1 (first quartile) , 50 percent of the pregs are 2.5(median) and 75 percent of the pregs are 5.5(third quartile).
4. The data is left skewed.
5. The Q1(first quartile = around 1) are the 25 percent of pregs in the data.

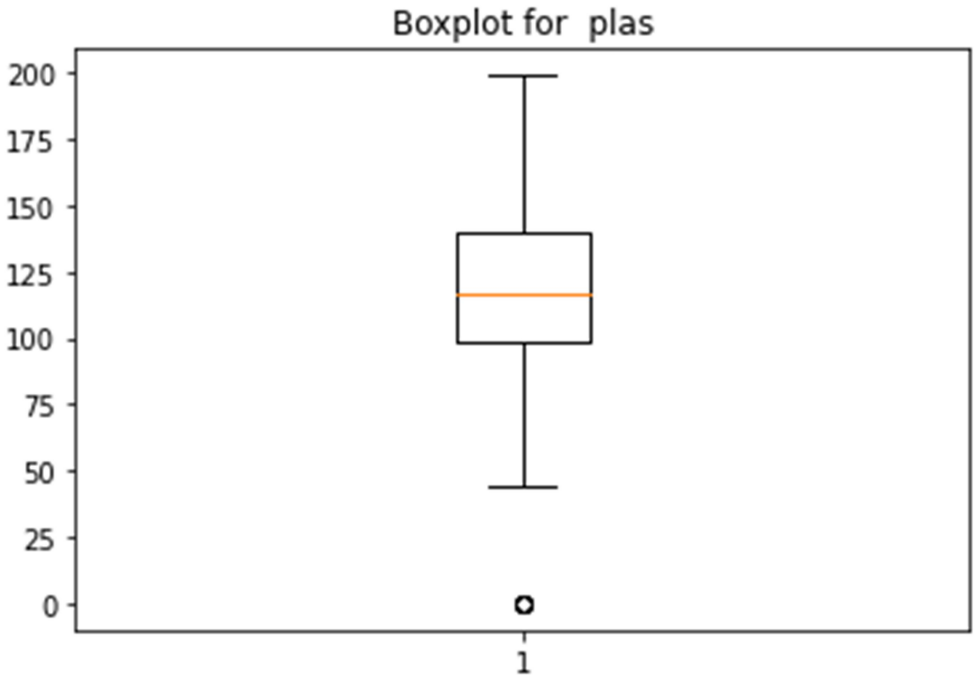

Figure 20 Boxplot for attribute plas

Inferences:

1. There are very few values of plas which equals to 0 (as they are outliers in the boxplot).
2. 50 percent of plas comes under the group of 100 to 140 (as they occurs in the interquartile).
3. 25 percent of the plas are 100 (first quartile) , 50 percent of the plas are 125(median) and 75 percent of the plas are 140(third quartile).
4. The boxplot is symmetric.
5. The Q1(first quartile = around 100) are the 25 percent of plas in the data.

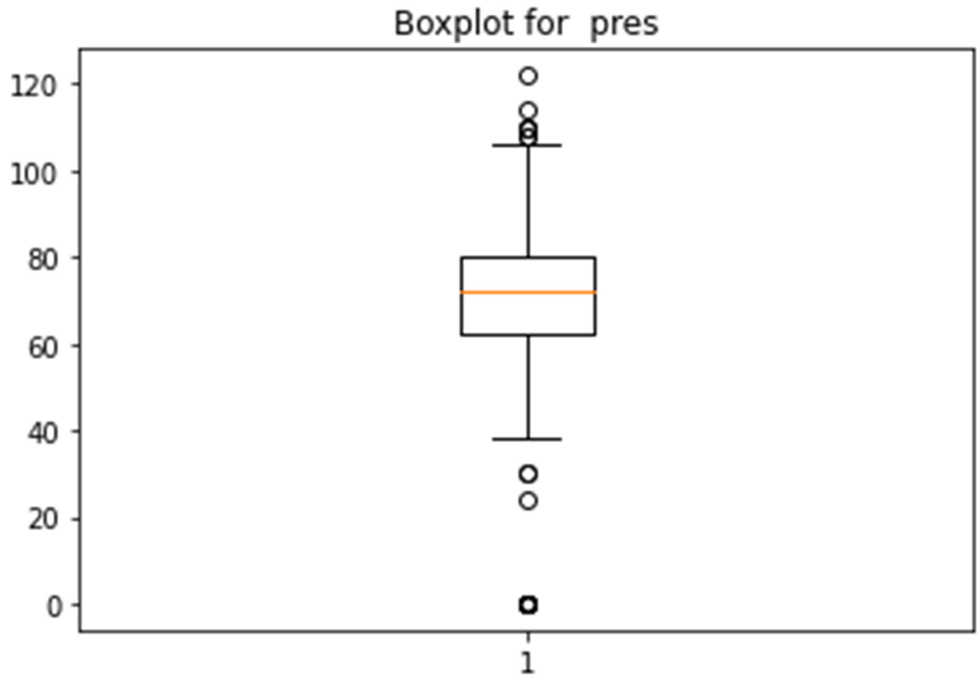

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are very few values of pres which lies in between (0-40 and 105-120 mm hg).
2. 50 percent of pres comes under the group of 60 to 80 mmhg .
3. 25 percent of the pres are 60 mm hg (first quartile) , 50 percent of the pres are 70 mm hg(median) and 75 percent of the pres are 80 mm hg(third quartile).
4. The boxplot is symmetric.
5. The Q1(first quartile = around 60 mm hg) are the 25 percent of pres in the data.

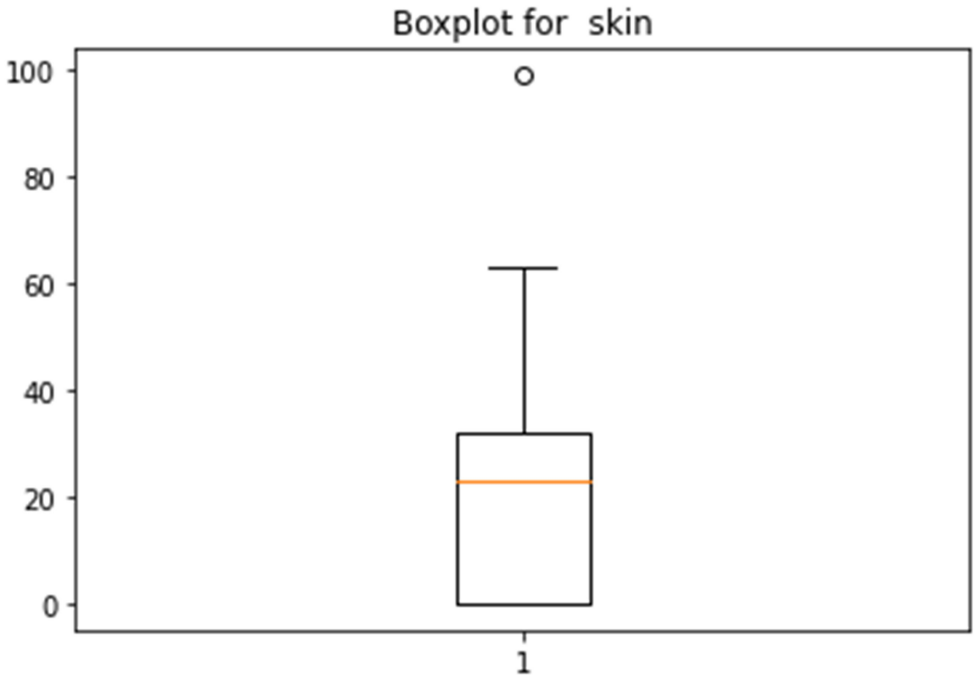

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. There are very few values of skin which equals 100 mm.
2. 50 percent of skin comes under the group of 0 to 30 mm.
3. 25 percent of the skin are 1mm (first quartile) , 50 percent of the pres are 30 mm(median) and 75 percent of the pres are 35 mm(third quartile).
4. The boxplot is Right skewed.
5. The Q1(first quartile = around 1 mm) are the 25 percent of pres in the data.

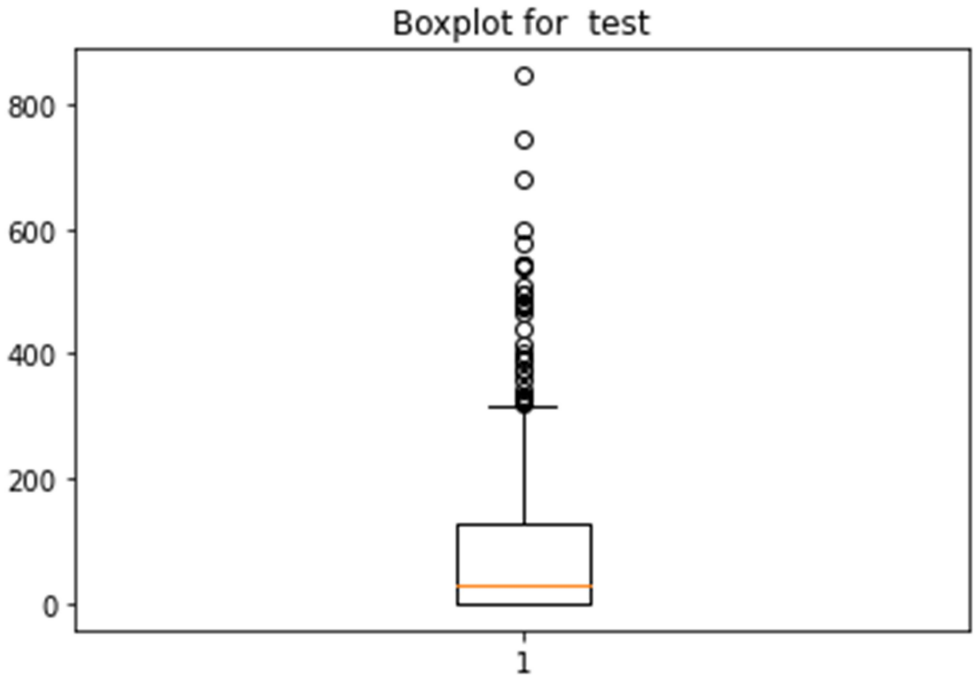

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. There are very few values of test which lies in between 300 to 800 U/mL.
2. 50 percent of test comes under the group of 10 to 180 U/mL.
3. 25 percent of the test are 10 U/mL (first quartile) , 50 percent of the test are 30 U/mL(median) and 75 percent of the test are 180 U/mL(third quartile).
4. The boxplot is left skewed.
5. The Q1(first quartile = around 10 U/mL) are the 25 percent of test in the data.

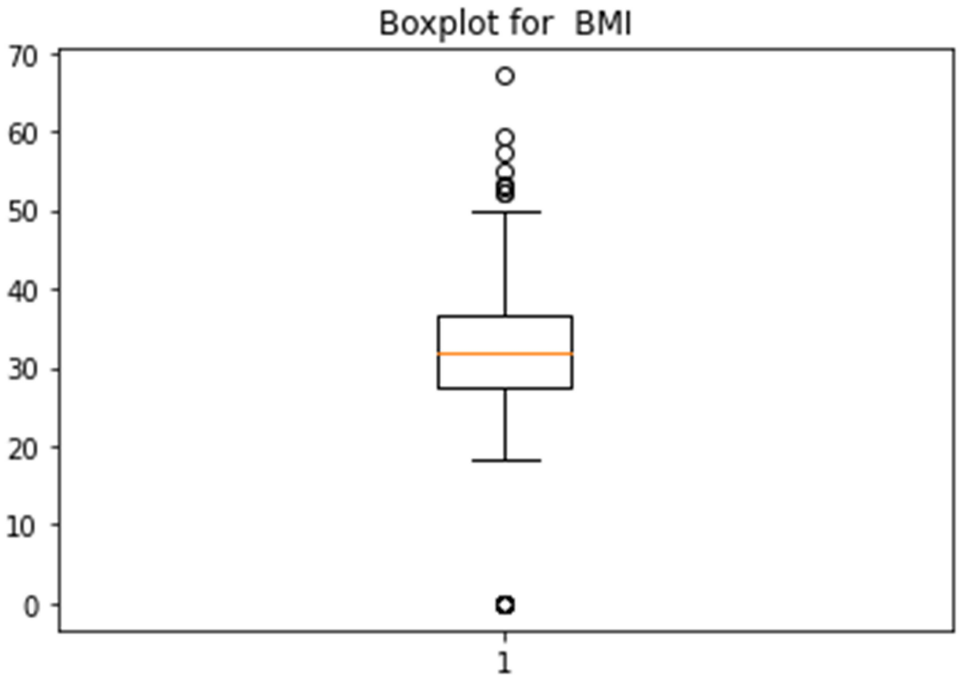

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. There are very few values of BMI which lies in between 50 to 60 kg/m^2.
2. 50 percent of test comes under the group of 30 to 40 kg/m^2.
3. 25 percent of the test are 28 kg/m^2 (first quartile) , 50 percent of the test are 32 kg/m^2 (median) and 75 percent of the test are 38 kg/m^2 (third quartile).
4. The boxplot is symmetric.
5. The Q1(first quartile = around 28 kg/m^2) are the 25 percent of test in the data.

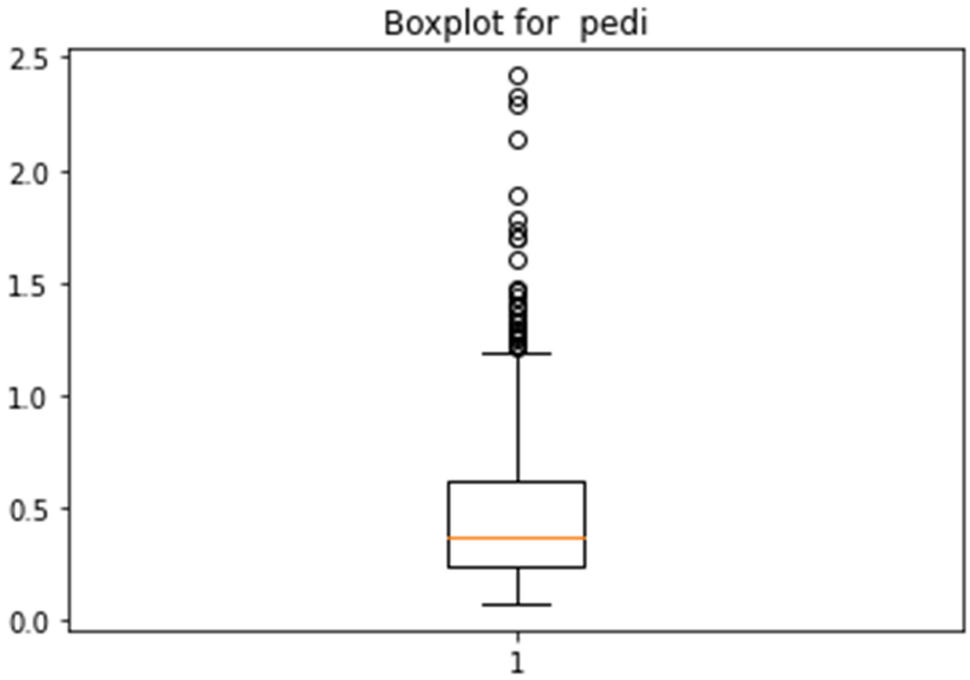

Figure 25 Boxplot for attribute pedi

Inferences:

1. There are very few values of BMI which lies in between 1.3 to 2.5.
2. 50 percent of BMI comes under the group of 0.3 to 0.6 .
3. 25 percent of the BMI are 0.3 (first quartile) , 50 percent of the BMI are 0.4 (median) and 75 percent of the BMI are 0.6 (third quartile).
4. The boxplot is left skewed.
5. The Q1(first quartile = around 0.3) are the 25 percent of BMI in the data.

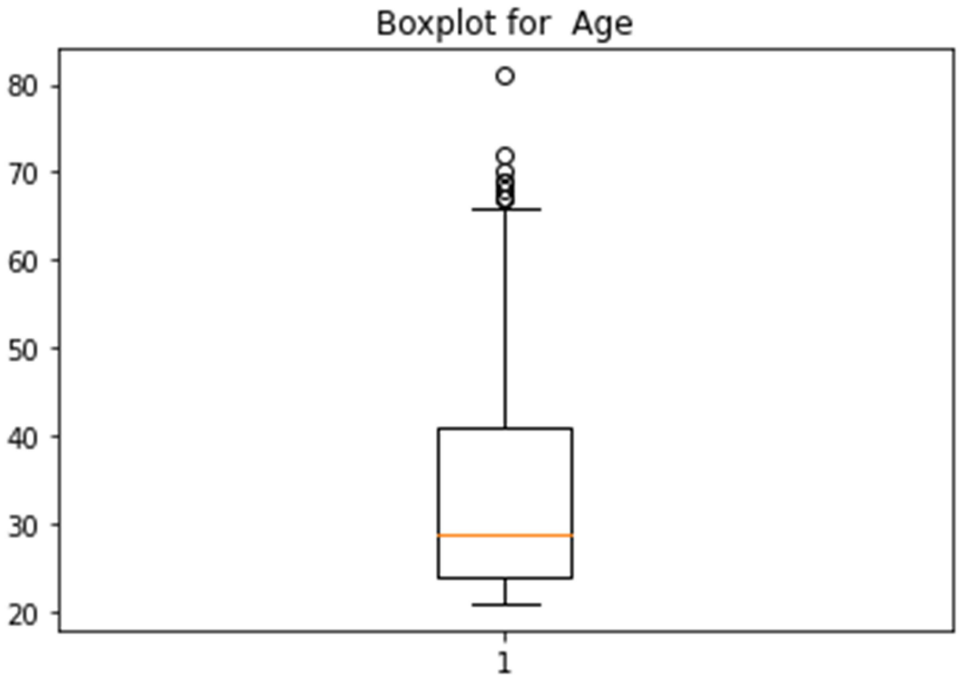

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. There are very few values of Age which lies in between 65 to 80 years.
2. 50 percent of Age comes under the group of 22 to 42 years.
3. 25 percent of the Age are 22 years (first quartile) , 50 percent of the Age are 30 years (median) and 75 percent of the BMI are 42 years (third quartile).
4. The boxplot is left skewed.
5. The Q1(first quartile = around 22) are the 25 percent of Age in the data