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

Student's Name: Garvit Verma **Mobile No: 8272840777**

Roll Number: B20098 **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	3.845	3	1	0	17	3.369
2	plas	120.894	117	99	0	199	31.973
3	pres (in mm Hg)	69.105	72	70	0	122	19.356
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	67.10	7.884
7	pedi	0.472	0.372	0.254	0.078	2.42	0.331
8	Age (in years)	33.241	29	22	21	81	11.760

Inferences:

1. In the attribute 'pedi', S.D. is close to '0', and also its Mean, Median and Mode are also close to each other.
2. Mean, Median and Mode of attribute 'BMI' are almost equal to each other. So, we can conclude that the data of BMI column is symmetrically distributed.

2 a.

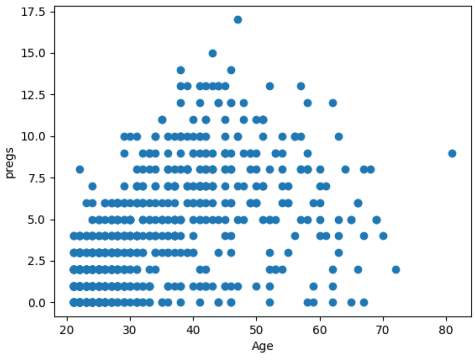

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. Attribute ‘Age’ is very strongly correlated to attribute ‘pregs’.
2. Density of points is high along a line with some positive slope.

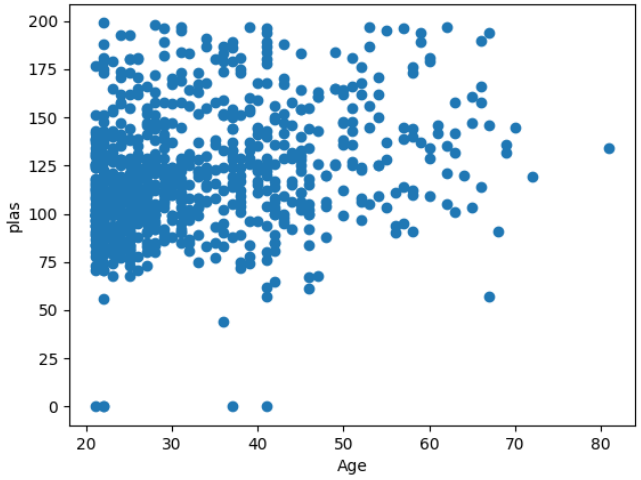

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. Attribute ‘Age’ is moderately correlated to attribute ‘plas’.
2. Density of points is not very high along a line with some positive slope.

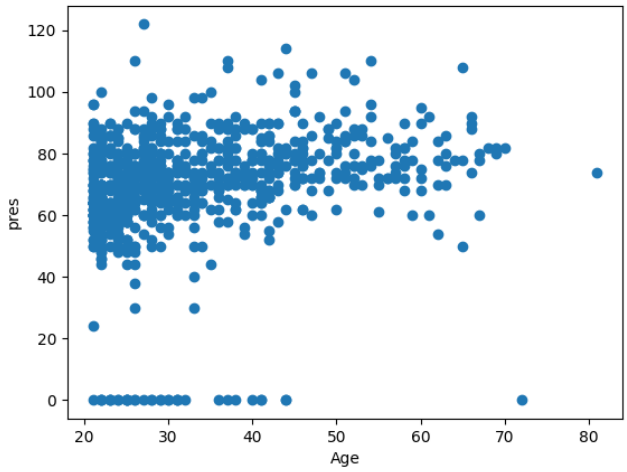

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. Attribute ‘Age’ is moderately correlated to attribute ‘pres’.
2. Density of points is not very high along a line with some positive slope.

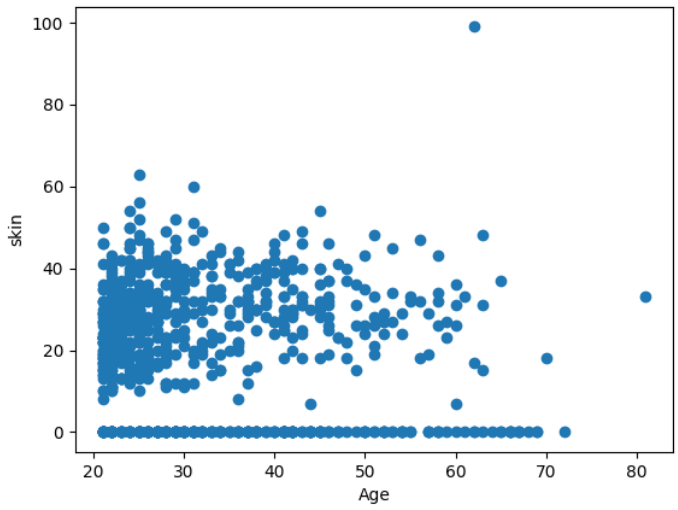

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. Attribute ‘Age’ is moderately correlated to attribute ‘skin’.
2. Density of points is not very high along a line with some negative slope.

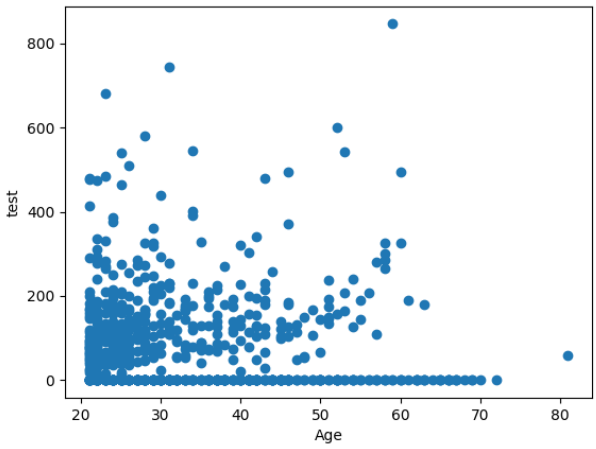

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. Attribute ‘Age’ is weakly correlated to attribute ‘test’.
2. Density of points is very low along a line with some negative slope.

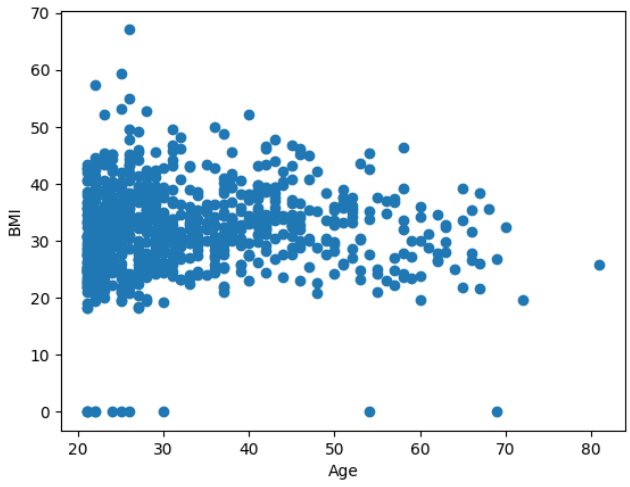

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m²)

Inferences:

1. Attribute ‘Age’ is weakly correlated to attribute ‘BMI’.
2. Density of points is very low along a line with some positive slope.

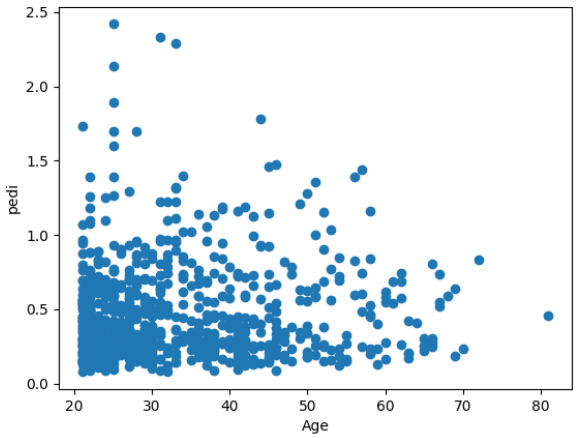

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. Attribute ‘Age’ is weakly correlated to attribute ‘pedi’.
2. Density of points is very low along a line with some positive slope.

b.

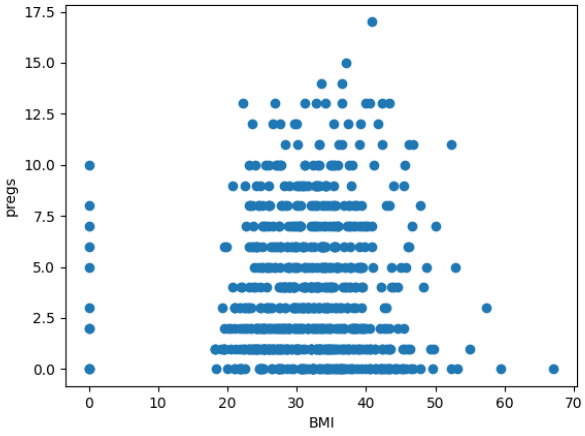

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. Attribute ‘BMI’ is weakly correlated to attribute ‘pregs’.
2. Density of points is very low along a line with some positive slope.

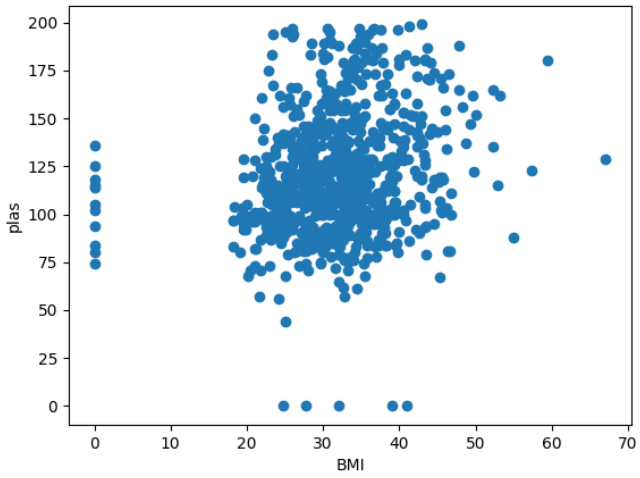

Figure 9 Scatter plot: BMI (in kg/m²) vs. plas

Inferences:

1. Attribute ‘BMI’ is moderately correlated to attribute ‘plas’.
2. Density of points is not very high along a line with some positive slope.

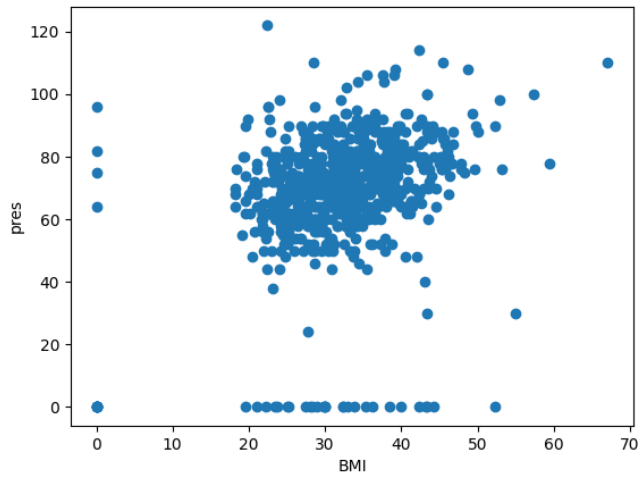

Figure 10 Scatter plot: BMI (in kg/m²) vs. pres (in mm Hg)

Inferences:

1. Attribute 'BMI' is moderately correlated to attribute 'pres'.
2. Density of points is not very high along a line with some positive slope.

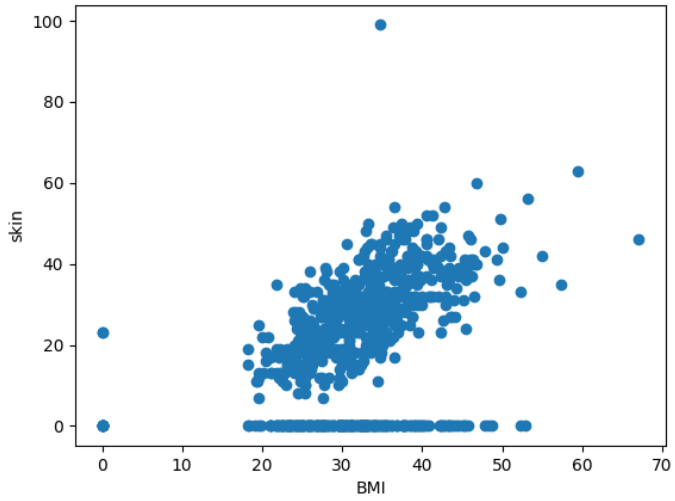

Figure 11 Scatter plot: BMI (in kg/m²) vs. skin (in mm)

Inferences:

1. Attribute 'BMI' is strongly correlated to attribute 'skin'.
2. Density of points is very high along a line with some positive slope.

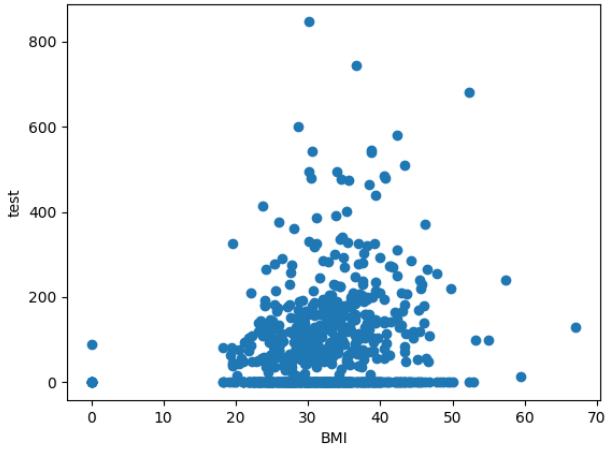

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. Attribute 'BMI' is moderately correlated to attribute 'test'.
2. Density of points is not very high along a line with some positive slope.

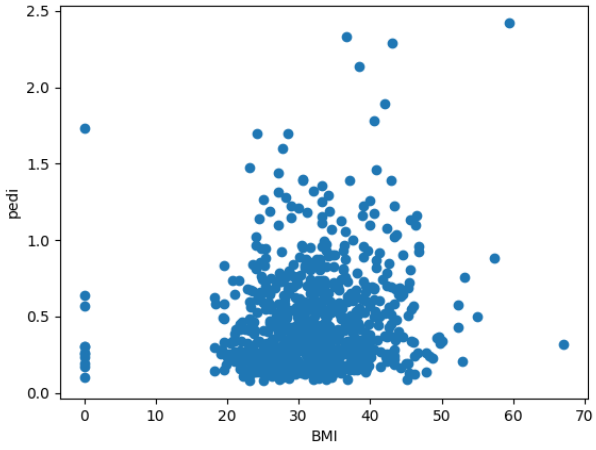

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. Attribute ‘BMI’ is moderately correlated to attribute ‘pedi’.
2. Density of points is not very high along a line with some positive slope.

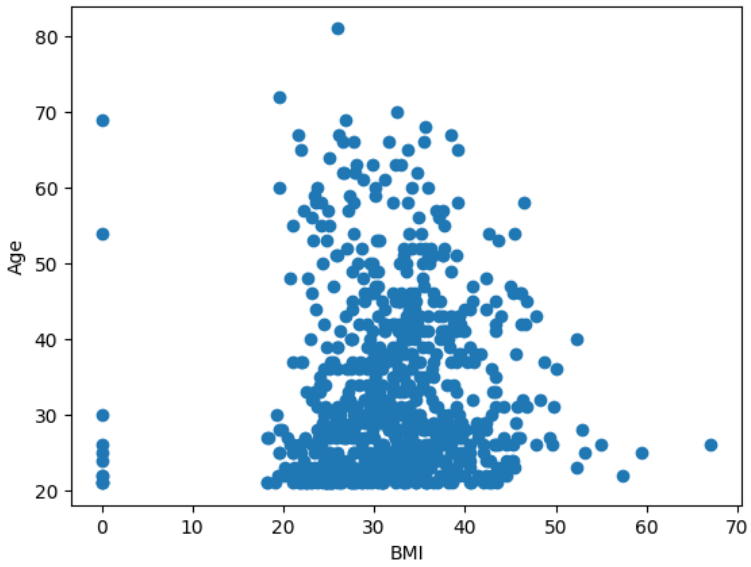

Figure 14 Scatter plot: BMI (in kg/m²) vs. Age (in years)

Inferences:

1. Attribute ‘BMI’ is weakly correlated to attribute ‘Age’.
2. Density of points is very low along a line with some positive slope.

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)	1

Inferences:

1. Age-pregs: **very strong**, Age-plas: **moderate**, Age-pres: **moderate**, Age-skin: **moderate**, Age-test: **weak**, Age-BMI: **weak**, Age-pedi: **weak** correlation.
2. **Age-pregs**: pregs increases as Age increases and vice versa.
   Age-plas: plas increases as Age increases and vice versa.
   Age-pres: pres increases as Age increases and vice versa.
   Age-skin: skin decreases as Age increases and vice versa.
   Age-test: test decreases as Age increases and vice versa.
   Age-BMI: BMI increases as Age increases and vice versa.
   Age-pedi: pedi increases as Age increases and vice versa.
3. **Age-pregs**: There is strong linear relationship between the two attributes.
   Age-plas: The relationship is not very strong between the two attributes.
   Age-pres: The relationship is not very strong between the two attributes.
   Age-skin: The linear relationship is not very strong and as value of 'skin' increases, 'Age' decreases.
   Age-test: The linear relation between two attributes is weak and the line is hard to distinguish and as value of 'skin' increases, 'Age' decreases.
   Age-BMI: The linear relation between two attributes is weak.
   Age-pedi: The linear relation between two attributes is weak and the line is hard to distinguish.

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
7	pedi	0.141
8	Age (in years)	0.036

Inferences:

1. BMI-pregs: **weak**, BMI-plas: **moderate**, BMI-pres: **moderate**, BMI-skin: **strong**, BMI-test: **moderate**, BMI-pedi: **moderate**, BMI-Age: **weak** correlation
2. **BMI-pregs**: pregs increases as BMI increases and vice versa.
   BMI-plas: plas increases as BMI increases and vice versa.
   BMI-pres: pres increases as BMI increases and vice versa.
   BMI-skin: skin increases as BMI increases and vice versa.
   BMI-test: test increases as BMI increases and vice versa.
   BMI-Age: Age increases as BMI increases and vice versa.
   BMI-pedi: pedi increases as BMI increases and vice versa.
3. **BMI-pregs**: There is very weak linear relationship between the two attributes and the line is hard to distinguish.
   BMI-plas: The relationship is not very strong between the two attributes.
   BMI-pres: The relationship is not very strong between the two attributes.
   BMI-skin: The linear relationship is strong and as value of 'skin' increases, 'BMI' increases.
   BMI-test: The linear relationship is not very strong.
   BMI-Age: The linear relation between two attributes is weak and the line is hard to distinguish.
   BMI-pedi: The linear relation between two attributes is not very strong.

4 a.

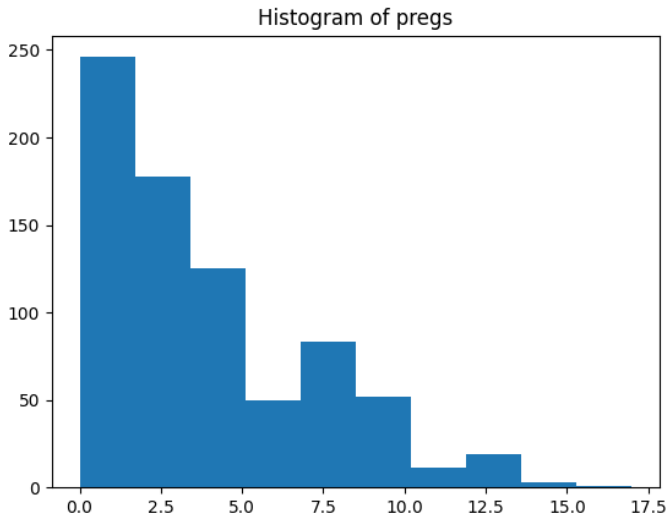

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. Number of bins: **10**, Width of each bin: 1.7
   Range: Frequency of bin
   0-1.7: 246, **1.7-3.4**: 178, **3.4-5.1**: 125, **5.1-6.8**: 50, **6.8-8.5**: 83, **8.5-10.2**: 52, **10.2-11.9**: 11, **11.9-13.6**: 19, **13.6-15.3**: 3, **15.3-17**: 1
2. Mode of the attribute ‘pregs’ is 1. So, it lies in the first bin of histogram which ranges from 0-1.7

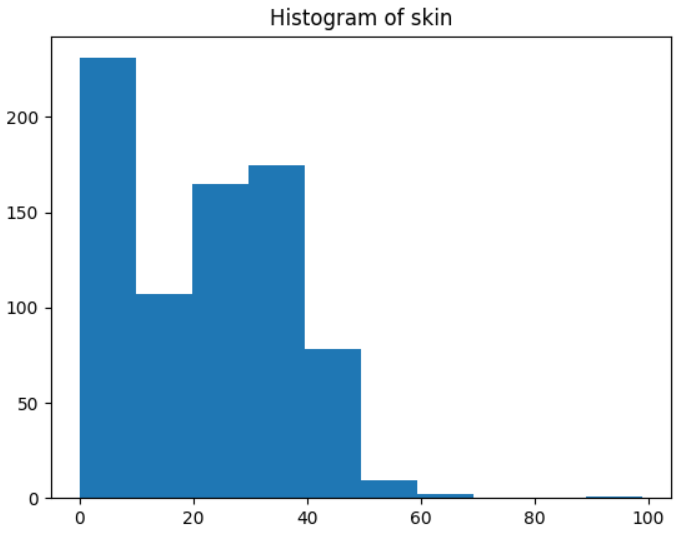

Figure 16 Histogram depiction of attribute skin

Inferences:

1. Number of bins: 10, Width of each bin: 9.9
   Range: Frequency of bin
   0-9.9: 231, **9.9-19.8**: 107, **19.8-29.7**: 165, **29.7-39.6**: 175, **39.6-49.5**: 78, **49.5-59.4**: 9, **59.4-69.3**: 2, **69.3-79.2**: 0, **79.2-89.1**: 0, **89.1-99**: 1
2. Mode of the attribute 'skin' is 0. So, it lies in the first bin of histogram which ranges from 0-1.7

5

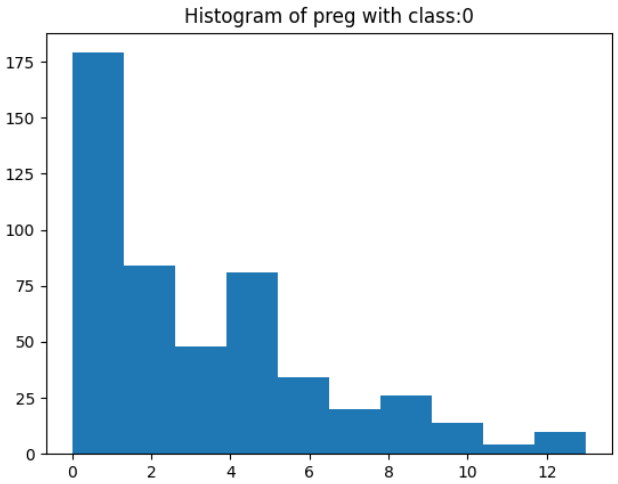

Figure 17 Histogram depiction of attribute pregs for class 0

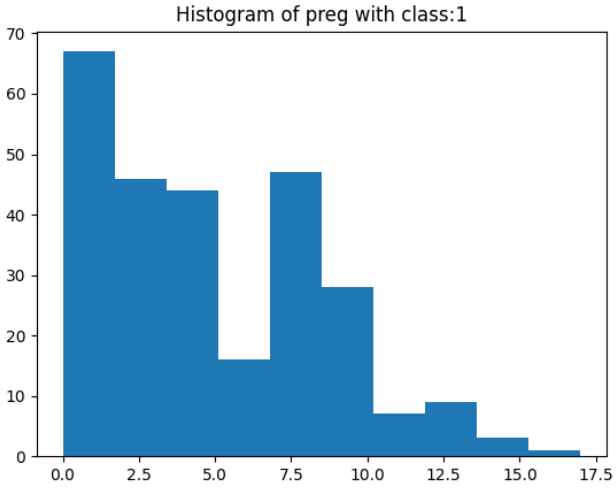

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. Mode of attribute ‘pregs’ is 1. For class 0, mode of ‘pregs’ lies in first bin which ranges from 0-1.3. For class 1, mode of ‘pregs’ lies in first bin which ranges from 0-1.7.
2. For class 0, **Range**: Frequency of bin

0-1.3: 179, **1.3-2.6**: 84, **2.6-3.9**: 48, **3.9-5.2**: 81, **5.2-6.5**: 34, **6.5-7.8**: 20, **7.8-9.1**: 26, **9.1-10.4**: 14, **10.4-11.7**: 4, **11.7-13**: 10.

For class 1, **Range**: Frequency of bin

0-1.7: 67, **1.7-3.4**: 46, **3.4-5.1**: 44, **5.1-6.8**: 16, **6.8-8.5**: 47, **8.5-10.2**: 28, **10.2-11.9**: 7, **11.9-13.6**: 9, **13.6-15.3**: 3, **15.3-17**: 1.

6

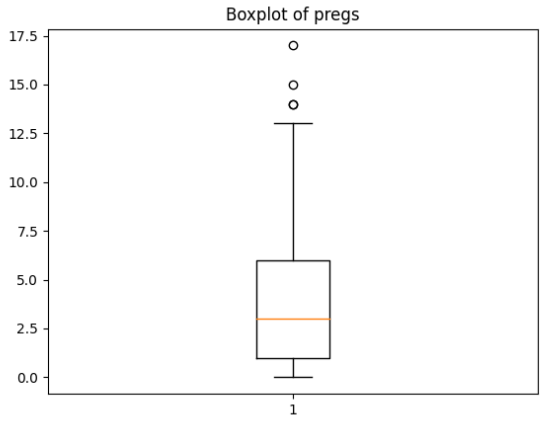

Figure 19 Boxplot for attribute pregs

Inferences:

1. Outliers are the values which lies outside of upper and lower bound. The values of outliers here are 14,15 and 17.
2. Inter quartile range is 5.
3. Approx. all values lie in the range 0-13.
4. Positively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 3, 17 and 0 respectively.

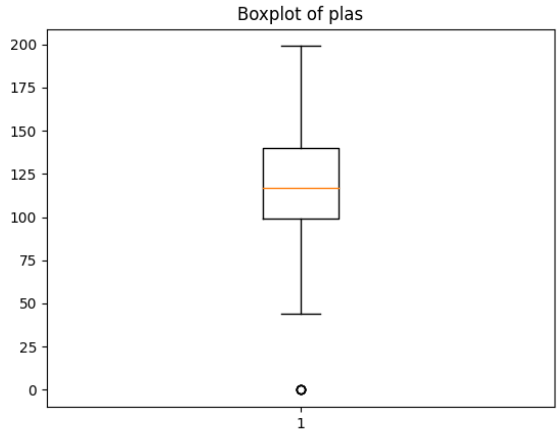

Figure 20 Boxplot for attribute plas

Inferences:

1. Here the outlier is '0' which lies below the lower bound i.e., 99-(1.5*41).
2. Inter quartile range is 140-99= 41.
3. Approx. all values lie in the range 50-199.
4. Positively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 117, 199 and 0 respectively.

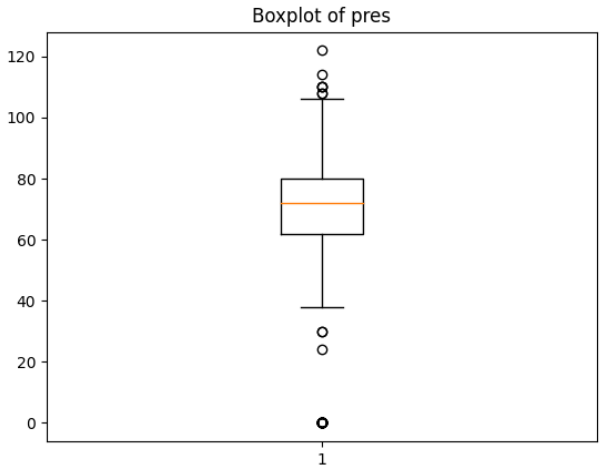

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. Values of outliers are 0,24,30 (lies below lower bound) and 108,110,114,122 (lies above upper bound).
2. Inter quartile range is 80-62= 18.
3. Approx. all values lie in the range 40-110.
4. Data is negatively skewed.
5. The values of median, maximum and minimum are matching from q1, i.e., 72, 122 and 0 respectively.

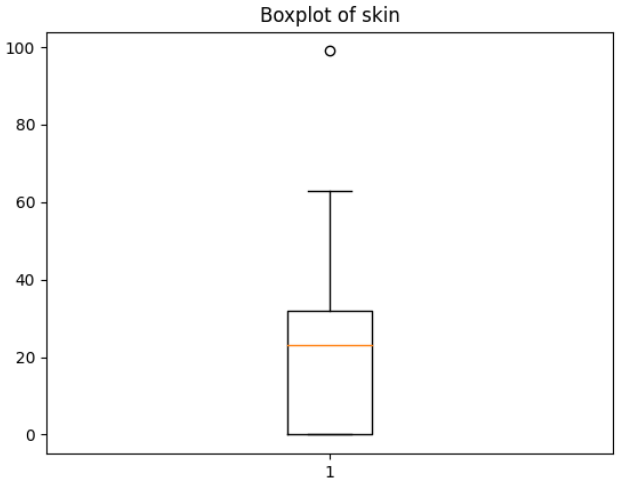

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. Value of outlier here is 99 (lies above upper bound).
2. Inter quartile range is 32-0= 32.
3. Approx. all values lie in the range 0-62.
4. Negatively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 23, 99 and 0 respectively.

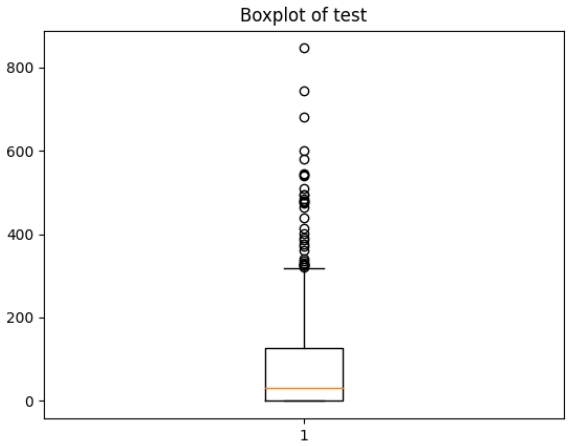

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. Values approximately in the range [317-846] are outliers.
2. Inter quartile range is 127.
3. Approx. all vales lie in the range 0-300.
4. Positively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 30.5, 846 and 0 respectively.

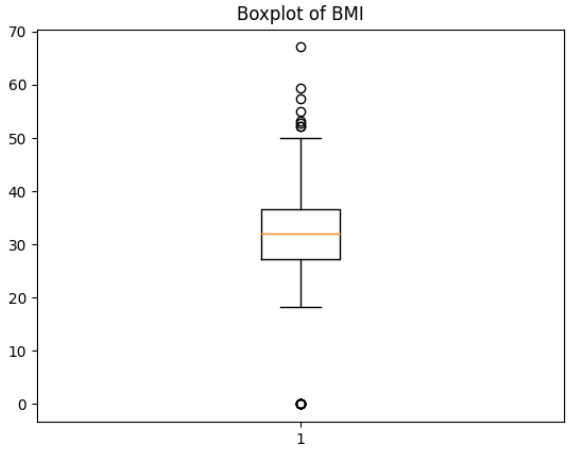

Figure 24 Boxplot for attribute BMI (in kg/m²)

Inferences:

1. Values of outliers are 67.1,55,53.2,52.3,52.9,59.4,57.3 (above upper bound) and 0 (below lower bound).
2. Inter quartile range is 9.
3. Approx. all values lie in the range 20-50.
4. Negatively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 32, 67.1 and 0 respectively

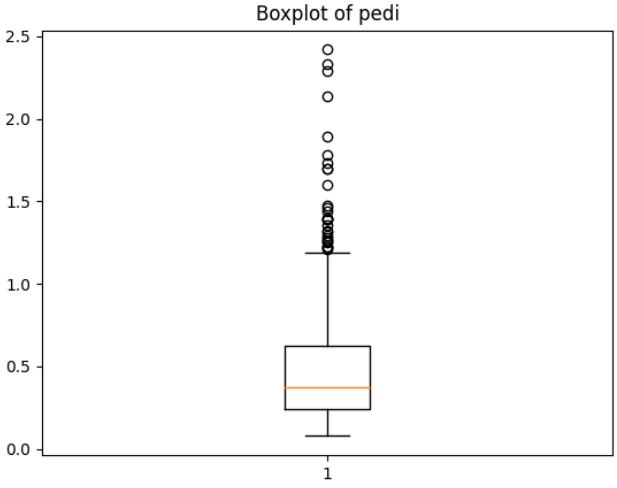

Figure 25 Boxplot for attribute pedi

Inferences:

1. Values above 0.6+(1.5*0.4) are outliers which lies in range [1.001, 2.42].
2. Inter quartile range is 0.4.
3. Approx. all values lie in the range 0-1.2.
4. Positively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 0.372, 2.42 and 0.078 respectively.

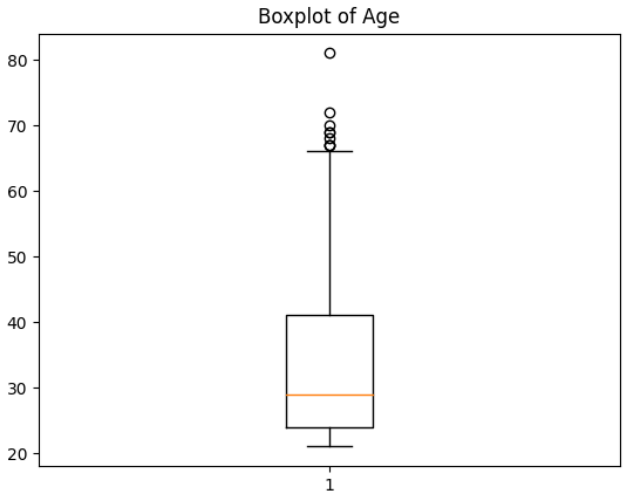

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. Values of outliers are 81, 72, 70, 69, 68, 67 (lies above upper bound).
2. Inter quartile range is 41-24= 17.
3. Approx. all values lie in the range 21-66.
4. Positively skewed data.
5. The values of median, maximum and minimum are matching from q1, i.e., 29, 81 and 21 respectively.