IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – I

Data visualization and statistics from data

Student's Name: Amit Maindola
Roll Number: B20079

Mobile No.: +91 7470985613
Branch: Computer Science & Engineering

1

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

S. No.	Attribute	Mean	Median	Mode	Min.	Max.	S.D.
1	pregs	3.845	3.000	1.000	0.000	17.000	3.370
2	plas	120.895	117.000	99.500	0.000	199.000	31.973
3	pres (in mm Hg)	69.105	72.000	70.000	0.000	120.000	19.356
4	skin (in mm)	20.536	23.000	0.000	0.000	99.000	15.952
5	test (in mu U/mL)	79.799	30.500	0.000	0.000	846.000	115.244
6	BMI (in kg/m^2)	31.993	32.000	32.000	0.000	67.100	7.884
7	pedi	0.472	0.372	0.256	0.078	2.420	0.331
8	Age (in years)	33.241	29.000	22.000	21.000	81.000	11.760

Inferences:

1. In the above table standard deviation is closest to zero in 'pedi' and hence the mean, median, and mode values are closest in this case.
2. Standard deviation have a small value for 'pedi' it means its values are almost equal for all the rows in csv file.

2

a.

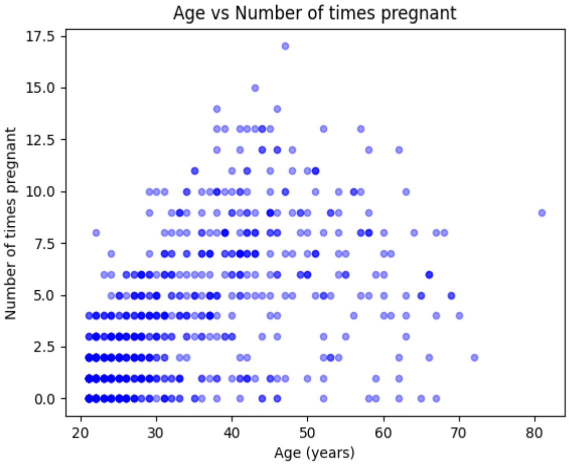

Figure 1: Scatter plot: Age (in years) vs. pregs

Inferences:

1. ’preg’ and ’Age’ are a very strongly correlated,and ’preg’ increases highly with increase in the ’Age’ for most of the points.
2. The scatter plot density is more for lower values of ’Age’ and decreases with increase in ’Age’.

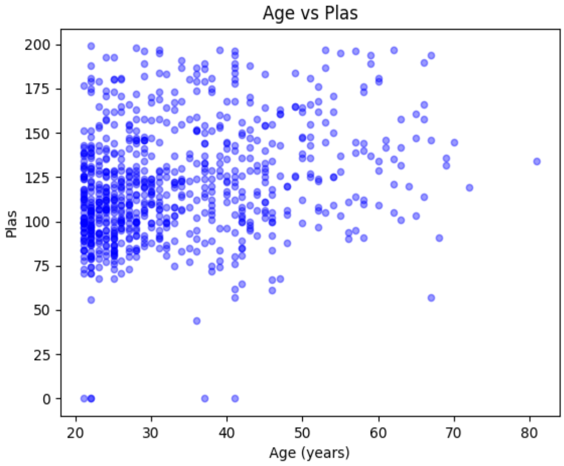

Figure 2: Scatter plot: Age (in years) vs. plas

Inferences:

1. ’plas’ and ’Age’ have a moderate correlation. ’plas’ increases by a remarkable value with a increase in’Age’.
2. The scatter plot density is more for lower values of ’Age’ and decreases with increase in ’Age’.

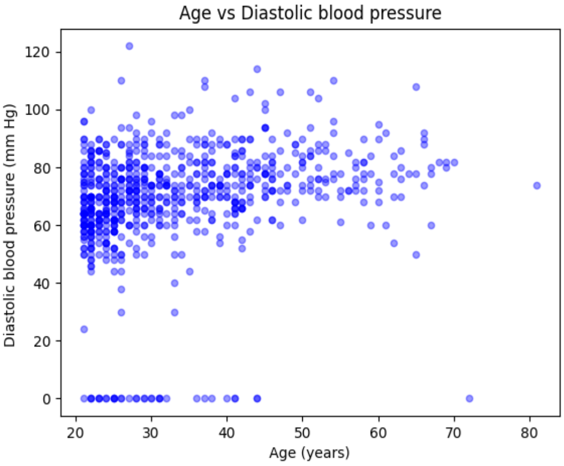

Figure 3: Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. ’pres’ and ’Age’ have a moderate correlation. ’pres’ increases by a remarkable value with a increase in ’Age’.
2. The scatter plot density is higher for lower values of ’Age’.

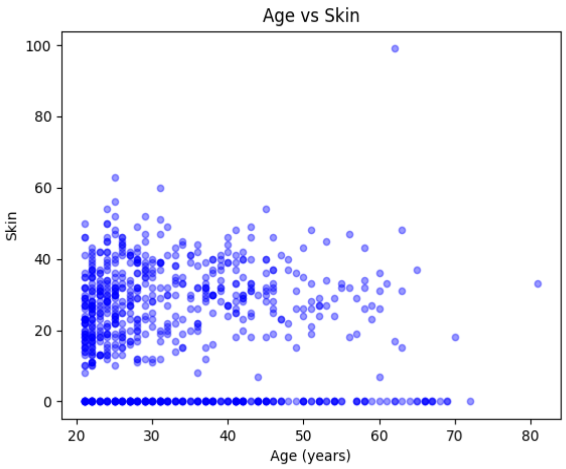

Figure 4: Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. ’skin’ and ’Age’ have a moderate correlation. ’skin’ decreases by a remarkable value with a increase in ’Age’.
2. The scatter plot density is higher for lower values of ’Age’.
3. A huge number of scatters in the plot have value of ’skin’ as zero.

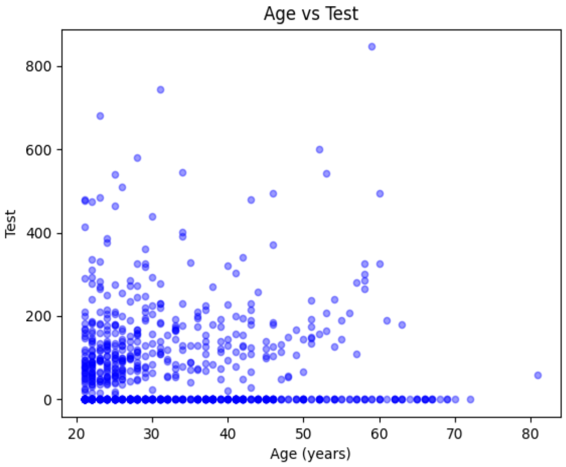

Figure 5: Scatter plot: Age (in years) vs. test (in mu U/mL)

Inferences:

1. ’test’ and ’Age’ have a weak correlation, there is neglegible effect of increasing ’Age’ on ’test’.
2. The scatter plot density is higher for lower values of ’Age’.
3. A huge number of scatters in the plot have value of ’test’ as zero.

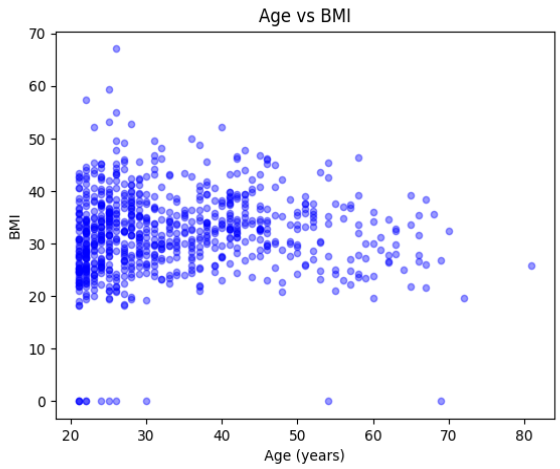

Figure 6: Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. 'BMI' and 'Age' have a weak correlation, there is neglegible effect of increasing 'Age' on 'BMI'.
2. The scatter plot density is higher for lower values of 'Age'.
3. A little number of scatters in the plot have value of 'BMI' as zero.

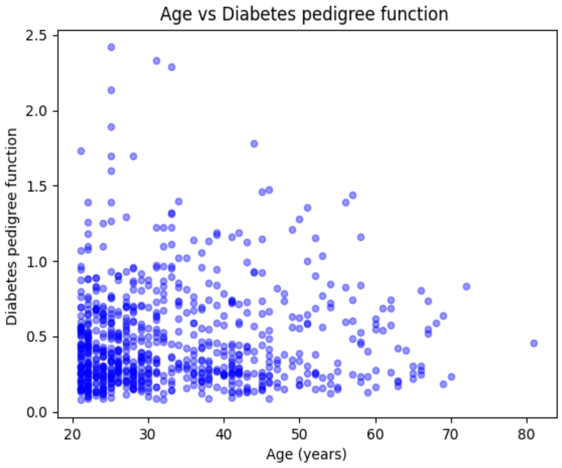

Figure 7: Scatter plot: Age (in years) vs. pedi

Inferences:

1. 'pedi' and 'Age' have a weak correlation, there is neglegible effect of increasing 'Age' on 'pedi'.
2. The scatter plot density is higher for lower values of 'Age' and 'pedi'.

b.

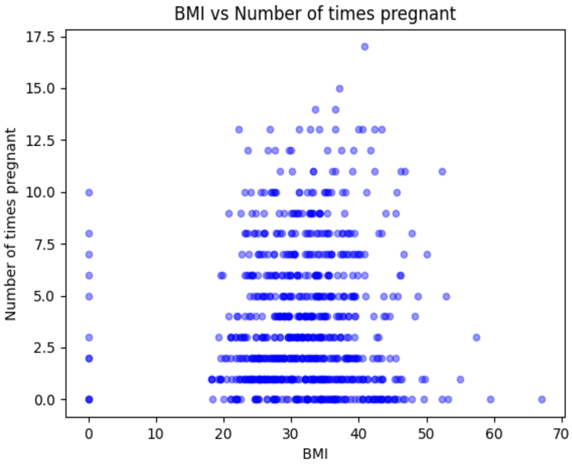

Figure 8: Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. ’pregs’ and ’BMI’ have a weak correlation, ’pregs’ increases with increase in ’BMI’
2. The scatter plot density is slightly higher for lower values of ’pregs’.
3. A remarkable number of dots have ’BMI’ as zero.

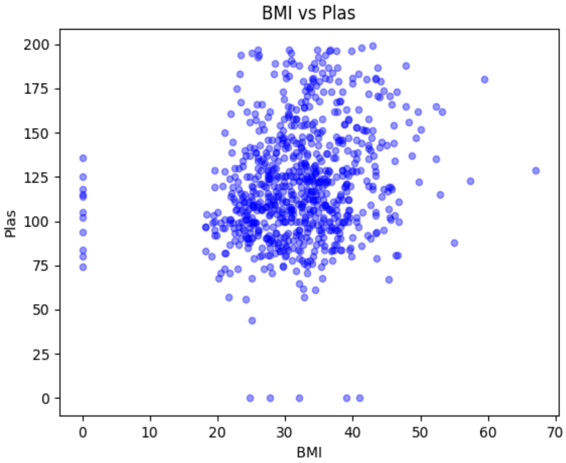

Figure 9: Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. ’plas’ and ’BMI’ have a moderate correlation, ’plas’ increases with increase in ’BMI’.
2. The scatter plot density is almost same.
3. A remarkable number of dots have either ’BMI’ or ’plas’ as zero.

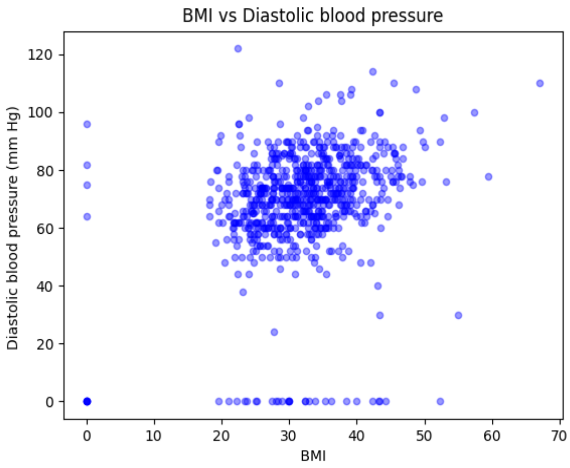

Figure 10: Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. 'pres' and 'BMI' have a moderate relationship, 'pres' increases with increase in 'BMI'
2. The scatter plot density is almost same.
3. A remarkable number of dots have either 'BMI' or 'pres' as zero.

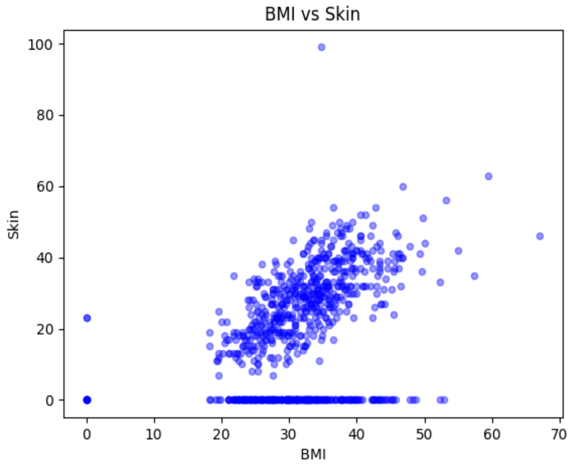

Figure 11: Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. 'skin' and 'BMI' have a strong relationship (one increases with other).
2. The scatter plot density is almost same.
3. A remarkable number of dots have value of 'skin' as zero.

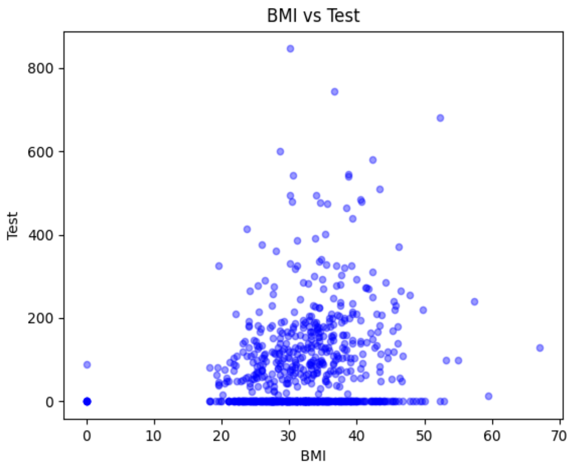

Figure 12: Scatter plot: BMI (in kg/m^2) vs. test (in mu U/mL)

Inferences:

1. 'test' and 'BMI' have a modrate relationship, 'test' increases with increase in 'BMI'
2. The scatter plot density is more for lower positive values of 'test'.
3. A remarkable number of dots have value of 'test' as zero.

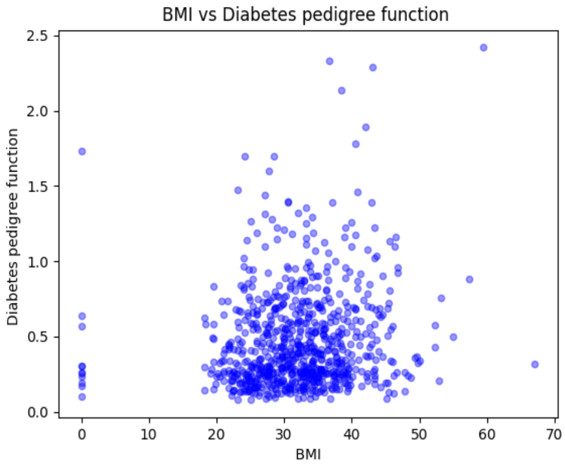

Figure 13: Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. 'pedi' and 'BMI' have a moderate relationship, 'pedi' increases with increase in 'BMI'.
2. The scatter plot density is more for lower values of 'pedi'.
3. A small number of dots have value of 'BMI' as zero.

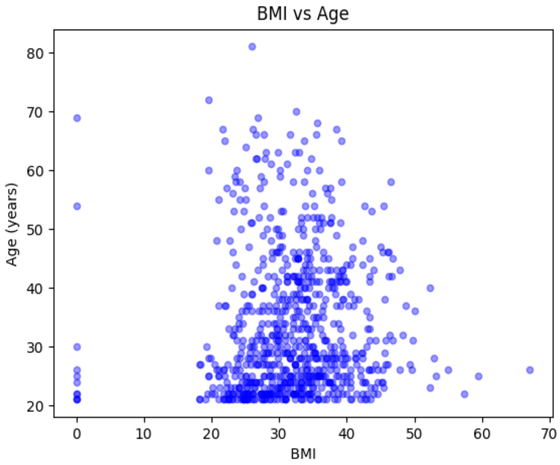

Figure 14: Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. 'Age' and 'BMI' have a weak relationship, 'Age' increases with increase in 'BMI'
2. The scatter plot density is more for lower values of 'Age'.
3. A small number of dots have value of 'BMI' as zero.

3

a.

Table 2: Correlation coefficient value computed between age and all other attributes

S. No.	Attribute	Correlation coefficient value
1	pregs	0.544
2	plas	0.264
3	pres (in mm Hg)	0.240
4	skin (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)	1.000

Inferences:

1. 'pregs' is **very strongly** correlated to 'Age'.
   'plas', 'pres' and 'skin' are **strongly** correlated to 'Age'.
   'test', 'pedi' and 'BMI' are **weakly** correlated to 'Age'.
   'Age' is **perfectly** correlated to 'Age' (as both are same).
2. 'pregs', 'plas', 'pres', 'BMI', 'pedi', and 'Age' **increases** with increase in 'Age' and vica versa.
   'skin' and 'test' **decreases** with increase in 'Age and vica versa'.
3. We can even see the same in correspondig plots where magnitude of correlation coefficient is **high** there the one attribute is showing a **high** increase/decrease on increasing the other attribute. Similarly where magnitude of correlation coefficient is **low** there the one attribute is showing a **low** increase/decrease on increasing the other attribute.
   We can also notice that if sign of correlation coefficient is negetive than one attribute **decreases** with a increase in other. Similarly if sign of correlation coefficient is positive than one attribute **increases** with a decrease in other.

b.

Table 3: Correlation coefficient value computed between BMI and all other attributes

S. No.	Attribute	Correlation coefficient value
1	pregs	0.018
2	plas	0.221
3	pres (in mm Hg)	0.282
4	skin (mm)	0.393
5	test (in mu U/mL)	0.198
6	BMI (in kg/m^2)	1.000
7	pedi	0.141
8	Age (in years)	0.036

Inferences:

1. 'skin' is **very strongly** correlated to 'BMI'.
   'plas', 'pres', 'pedi' and 'test' are **strongly** correlated to 'BMI'.
   'pregs', and 'Age' are **weakly** correlated to 'BMI'.
   'BMI' is **perfectly** correlated to 'BMI' (as both are same).

2. All the attributes **increases** with increase in 'BMI' and vica versa.

3. We can even see the same in correspondig plots where magnitude of correlation coefficient is **high** there the one attribute is showing a **high** increase/decrease on increasing the other attribute. Similarly where magnitude of correlation coefficient is **low** there the one attribute is showing a **low** increase/decrease on increasing the other attribute.
   We can also notice that if sign of correlation coefficient is negetive than one attribute **decreases** with a increase in other. Similarly if sign of correlation coefficient is positive than one attribute **increases** with a decrease in other.

4

a.

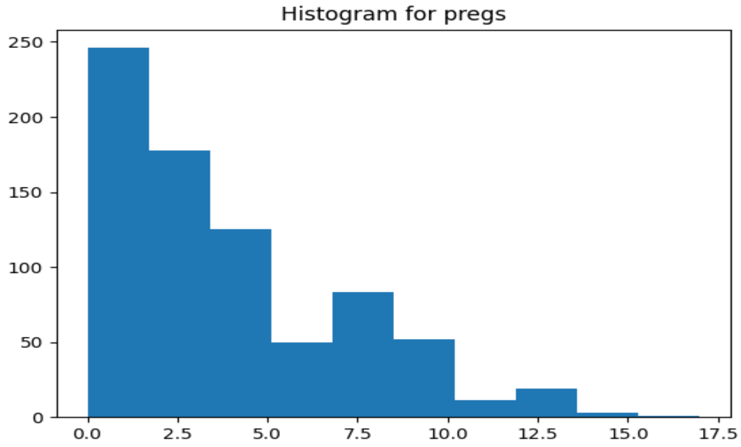

Figure 15: Histogram depiction of attribute pregs

Inferences:

1. There are total **10** bins each one with a width of **1.7**
   Frequency of bins is given below with corresponding range

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. Since mode of 'pregs' attribute is 1, it lies in first bin(0-1.7).

b.

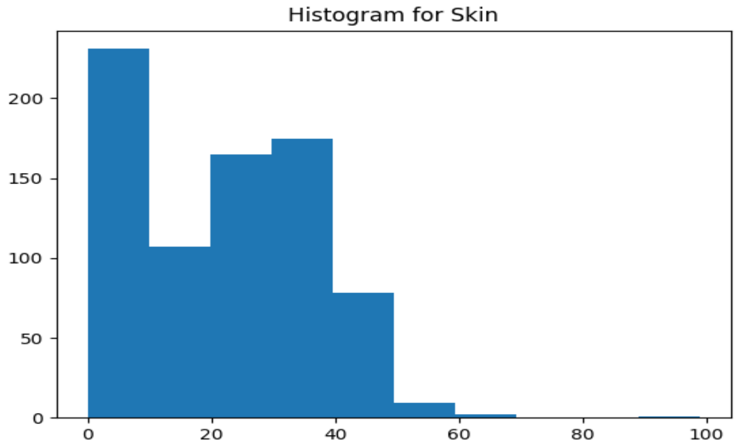

Figure 16: Histogram depiction of attribute skin

Inferences:

1. There are total **10** bins each one with a width of **9.9**
   Frequency of bins is given below with corresponding range
   0-9.9 : 231, 9.9-19.8 : 107, 19.8-29.7 : 175, 29.7-39.6 : 78, 39.6-49.5 : 9, 59.4-69.3 : 2, 69.3-179.2 : 0, 79.2-89.1 : 0, 89.1-99 : 1.

2. Since mode of 'skin' attribute is 0, it lies in first bin (0-9.9).

5

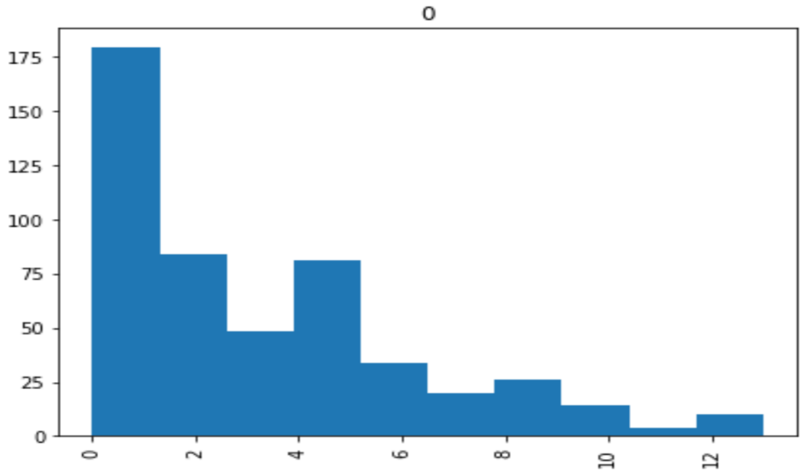

Figure 17: Histogram depiction of attribute pregs for class 0

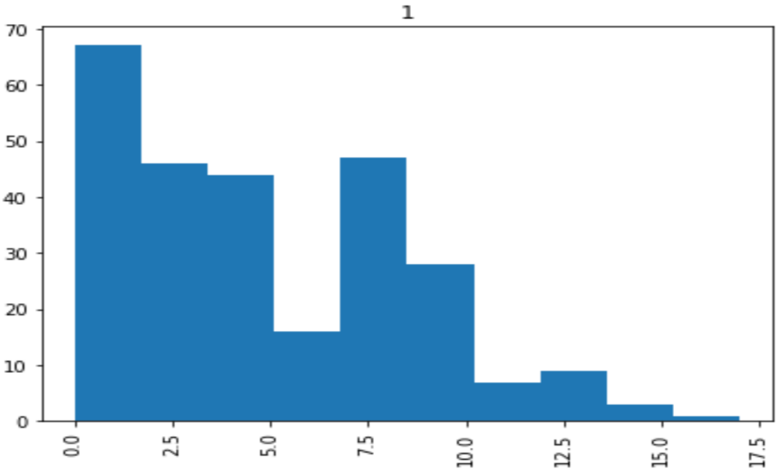

Figure 18: Histogram depiction of attribute pregs for class 1

Inferences:

1. Mode of 'pregs' attribute is **1**.
   For class=0 width of each bin is 1.3, therefore it lies in bin 1(0-1.3)
   For class=1 width of each bin is 1.3, therefore it lies in bin 1(0-1.7)

2. For class = 0, the range and corresponding frequencies are :
   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 = 0, the range and corresponding frequencies are :
   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.
   We can see that frequencies(height) on histogram with class = 0 is more than that on histogram with class = 1.

6

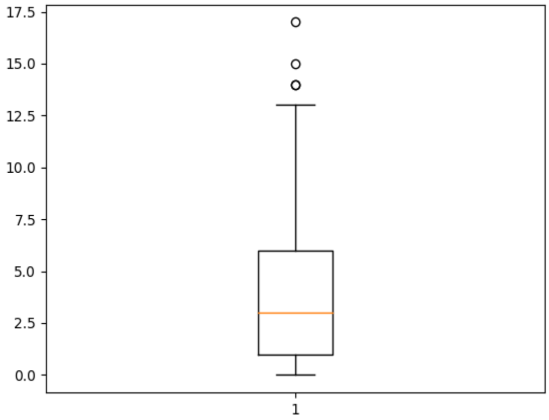

Figure 19: Boxplot for attribute pregs

Inferences:

1. There are three outliers in the above boxplot, with approx. values 14, 15 ans 17.

2. Inter quartile range : **5 (1.25 to 6.25)**.

3. Variability of attribute : Varrying from 0 to nearly 17.

4. skewness : **positive**.

5. Minimum and maximum value from Q1 also show that it varries from 0 to 17, and on the box plot we can relate the same.

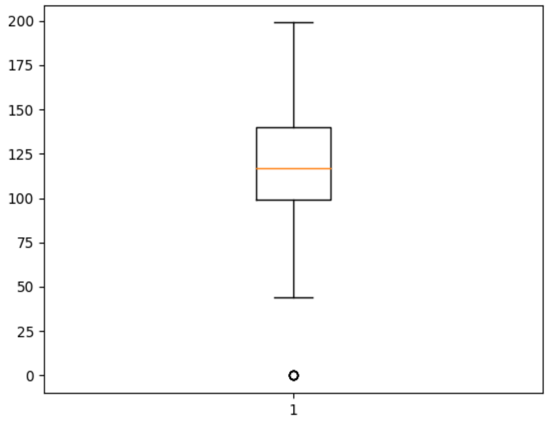

Figure 20: Boxplot for attribute plas

Inferences:

1. There only one outlier is present, with approx. value 0.

2. Inter quartile range : **37.5 (100 to 137.5)**.

3. Variability of attribute : Varrying from approx. 40 to 200.

4. skewness : **positive**.

5. Minimum and maximum value from Q1 also show that it varries from 0 to 199 (approx. 200), and on the box plot we can relate the same.

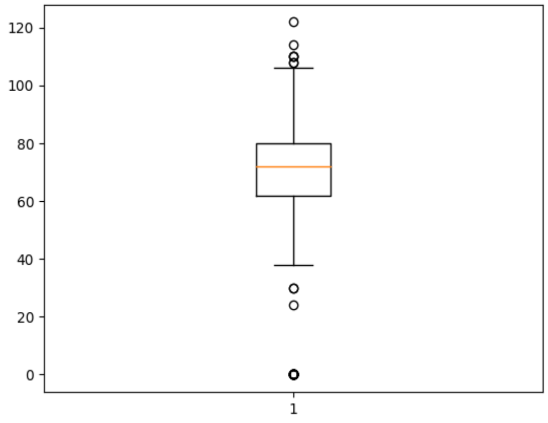

Figure 21: Boxplot for attribute pres (in mm Hg)

Inferences:

1. About four outliers are present, one with value 0, 2 with values in range of 20 to 40 and other four having values greater than 100.

2. Inter quartile range : **approx 18 (62 to 80)**.
3. Variability of attribute : Varrying from approx. 0 to 105.
4. skewness : **negative**.
5. Minimum and maximum value from Q1 also show that it varries from 0 to 120, and on the box plot we can relate the same.

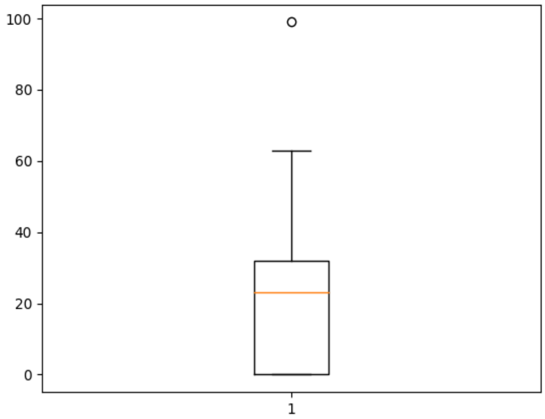

Figure 22: Boxplot for attribute skin (in mm)

Inferences:

1. There only one outlier is present, with approx. value 100.
2. Inter quartile range : **30 (0 to 30)**.
3. Variability of attribute : Varrying from approx. 0 to 65.
4. skewness : **negative**.
5. Minimum and maximum value from Q1 also show that it varries from 0 to 99 (approx. 100) and on the box plot we can relate the same.

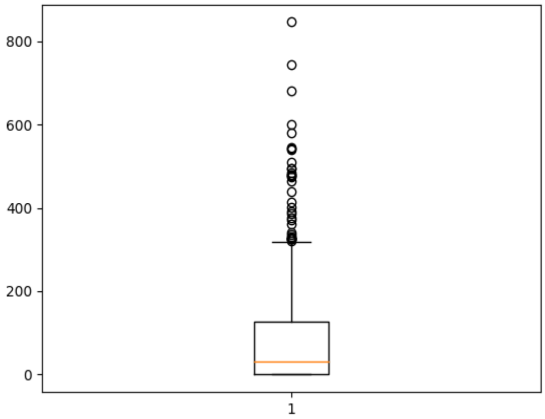

Figure 23: Boxplot for attribute test (in mu U/mL)

Inferences:

1. All outliers have values in a range of approx. 300 to 900.
2. Inter quartile range : **nearly 100 (0 to 100)**.
3. Variability of attribute : Varrying from approx. 0 to 300.
4. skewness : **positive**.
5. Minimum and maximum value from Q1 also show that it varries from 0 to 846 (approx. 850) and on the box plot we can relate the same.

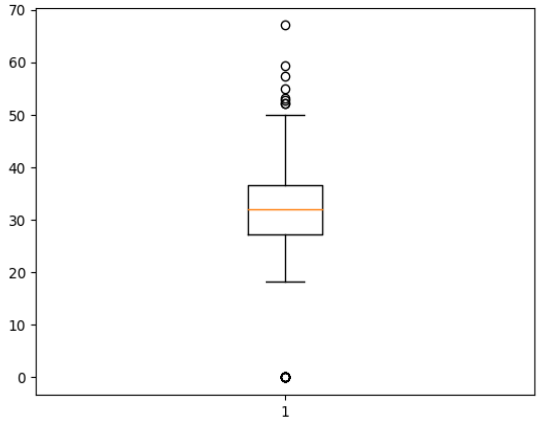

Figure 24: Boxplot for attribute BMI(kg/m^2)

Inferences:

1. One outlier have value 0, and all other outliers have values in a range of approx. 50 to 70.
2. Inter quartile range : **nearly 10 (27 to 37)**.
3. Variability of attribute : Varrying from approx. 18 to 50.
4. skewness : **symmetric data**.
5. Minimum and maximum value from Q1 also show that it varries from 0 to 67.1 (approx. 67) and on the box plot we can relate the same.

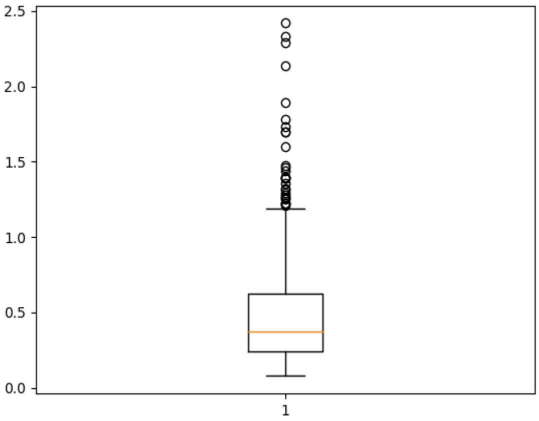

Figure 25: Boxplot for attribute pedi

Inferences:

1. All outliers have values in a range of approx. 1.2 to 2.4.
2. Inter quartile range : **nearly 0.5 (0.2 to 0.7)**.
3. Variability of attribute : Varrying from approx. 0.1 to 1.2.
4. skewness : **positive**.
5. Minimum and maximum value from Q1 also show that it varries from approx 0.078 to 2.42 and on the box plot we can relate the same.

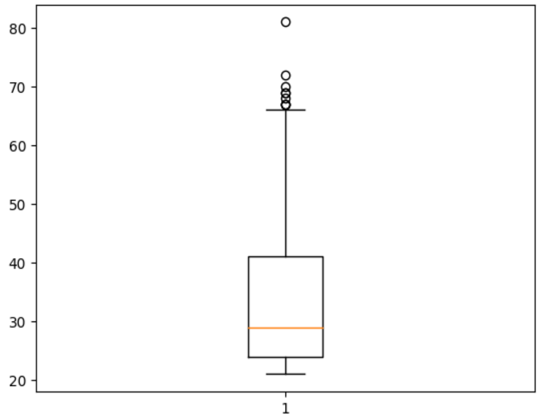

Figure 26: Boxplot for attribute Age (in years)

Inferences:

1. All outliers have values in a range of approx. 65 to 82.
2. Inter quartile range : **nearly 8 (24 to 42)**.
3. Variability of attribute : Varrying from approx. 18 to 50.
4. skewness : **positive**.
5. Minimum and maximum value from Q1 also show that it varries from 21 to 81 and on the box plot we can relate the same.