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

Student's Name: Madhur Jajoo **Mobile No: 7597389137**

Roll Number: B20211 **Branch: DSE**

1

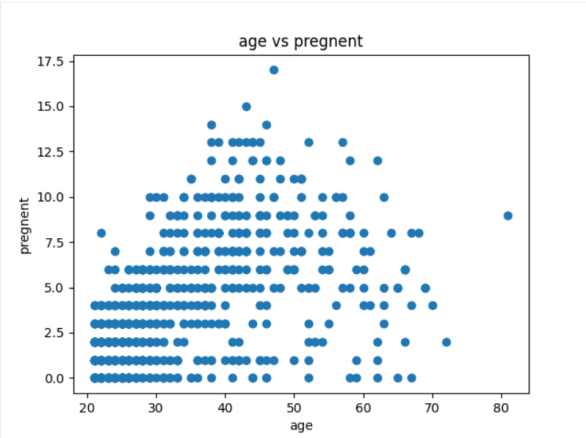
Table 1 Mean, median, mode, minimum, maximum and standard deviation for all the attributes

S. No.	Attributes	Mean	Median	Mode	Min.	Max.	S.D.
1	pregs	3.845	3.0	1	0	17	3.367
2	plas	120.895	117.0	99,100	0	199	31.952
3	pres (in mm Hg)	69.105	72.0	70	0	122	19.343
4	skin (in mm)	20.536	23	0	0	99	15.942
5	test (in mu U/mL)	79.799	30.5	0	0	846	115.169
6	BMI (in kg/m^2)	31.993	32.0	32	0.0	67.1	7.879
7	pedi	0.472	0.372	0.258,0.254	0.078	2.42	0.331
8	Age (in years)	33.241	29.0	22	21	81	11.753

Inferences:

1. As we can see that the value of standard deviation for "pedi" is close to zero therefore its mean mode median are also close to each other.

2 a.



Inferences:

1. Age and pregnant are in positive correlation
2. Women between age 20 to 30 years have been pregnant for 0 to 5 times

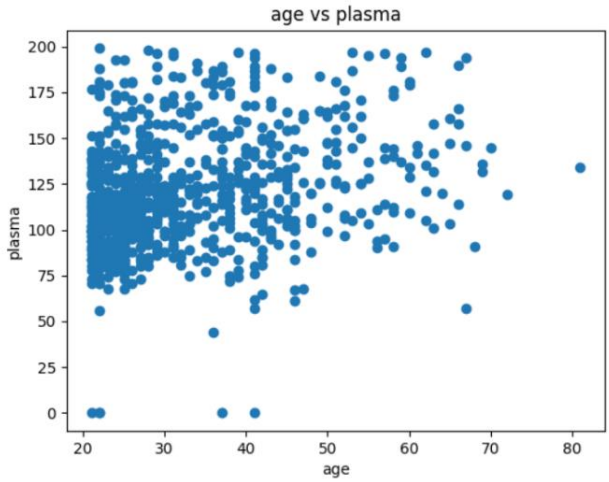


Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. Age and Plasma are in low positive correlation
2. Most of the patients are between age of 20 to 30 years and have plasma between 75 to 150

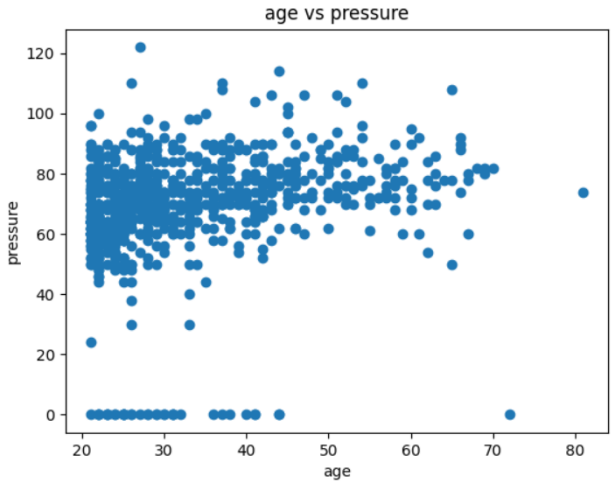


Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. Age and pressure have no correlation
2. Most of the patients of 20 to 30 years of age and have pressure between 60 to 90 mm HG

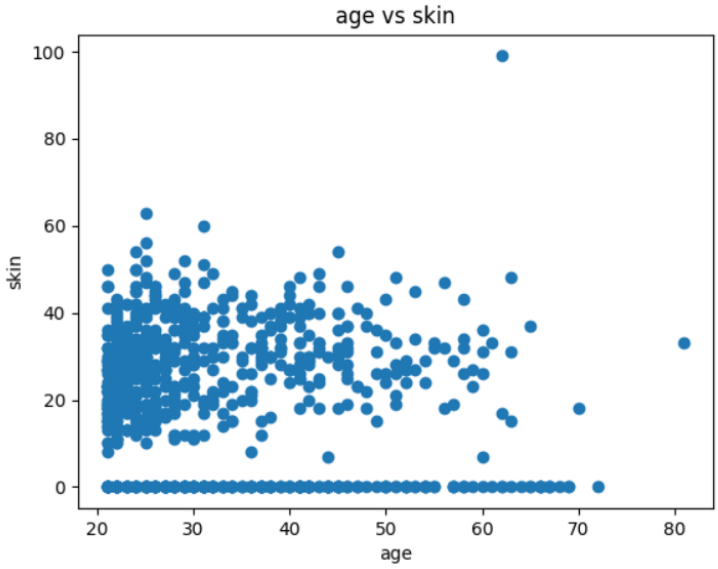


Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. Age and Skin are very low negatively correlated
2. Most of the patients are of age 20 to 30 years and have skinfold thickness of 10 to 40 mm

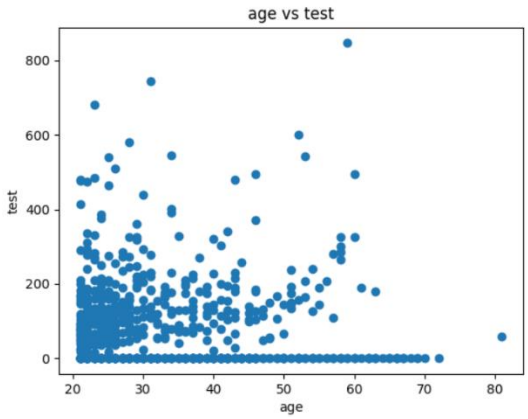


Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. Age and insulin are barely correlated
2. Most of the patients are of age of 20 to 30 years and have insulin level below 200

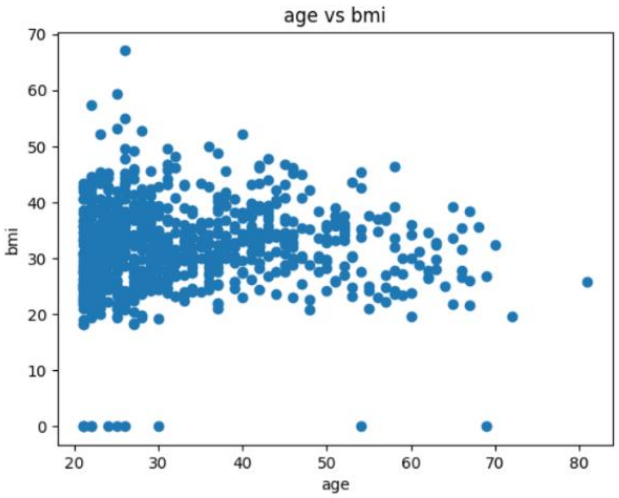


Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. Age and BMI are not correlated
2. Most of the Patients of any age group have BMI between 25 to 40 kg/m^2

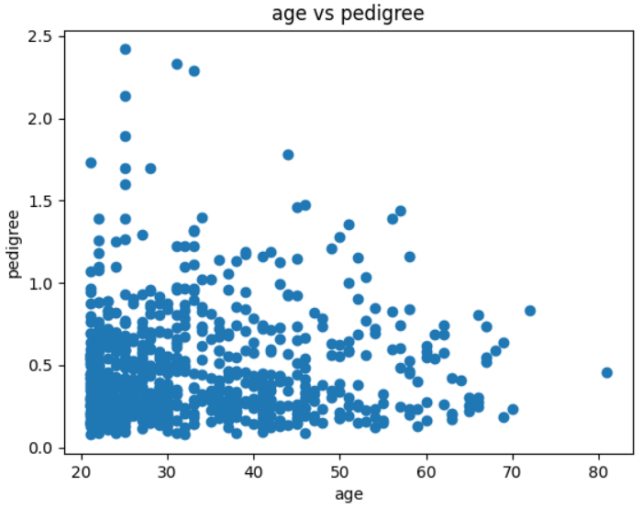


Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. Age and pedigree in are too low positive correlation

2. Most patients are of age group 20 to 30 years and have diabetes pedigree function value between 0.0 and 0.8

b.

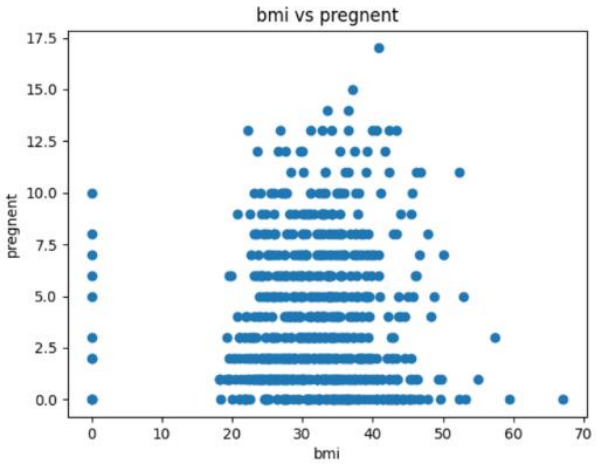


Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. These are not correlated
2. Pregnant women have BMI between 20 to 40 kg/m^2

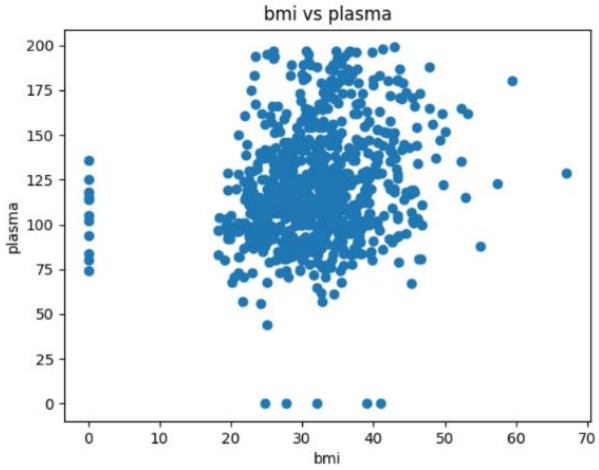


Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. These are not correlated as plot is discrete
2. Patients having BMI between 20 to 40 kg/m^2 have plas between 75 to 175

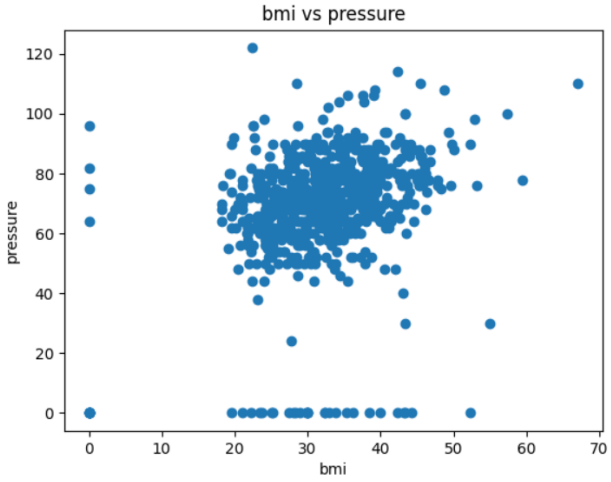


Figure 10 Scatter plot: BMI (in kg/m²) vs. pres (in mm Hg)

Inferences:

1. These are not correlated.
2. Patients having BMI between 20 to 40 kg/m^2 have pressure value between 50 to 90 mm HG

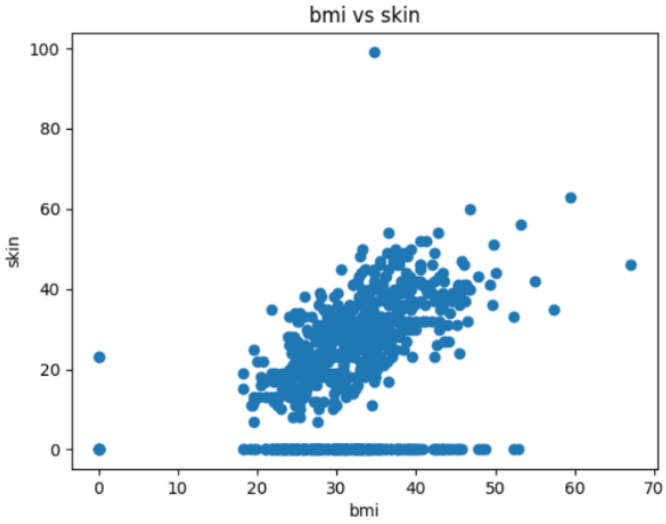


Figure 11 Scatter plot: BMI (in kg/m²) vs. skin (in mm)

Inferences:

1. These are positively correlated
2. Patients having BMI between 20 to 40 kg/m^2 have skinfold between 15 to 45 mm

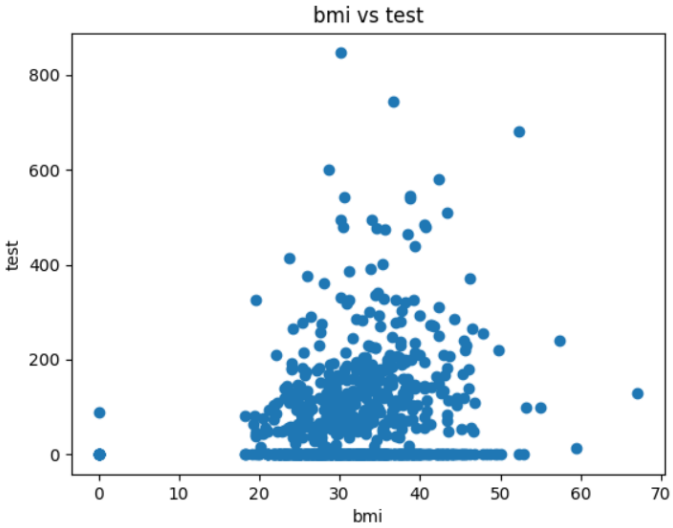


Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. These are not correlated
2. Patients having BMI between 20 to 40 kg/m^2 have pressure below 200mm U/ml

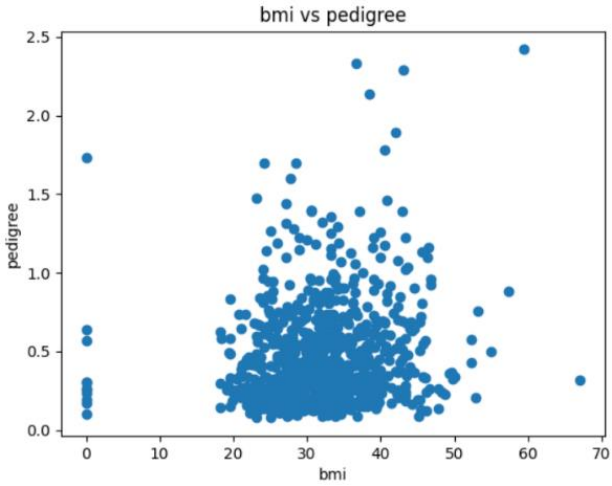


Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. These are not correlated.
2. Patients having BMI between 20 to 40 kg/m^2 have diabetes pedigree function value below 1.0

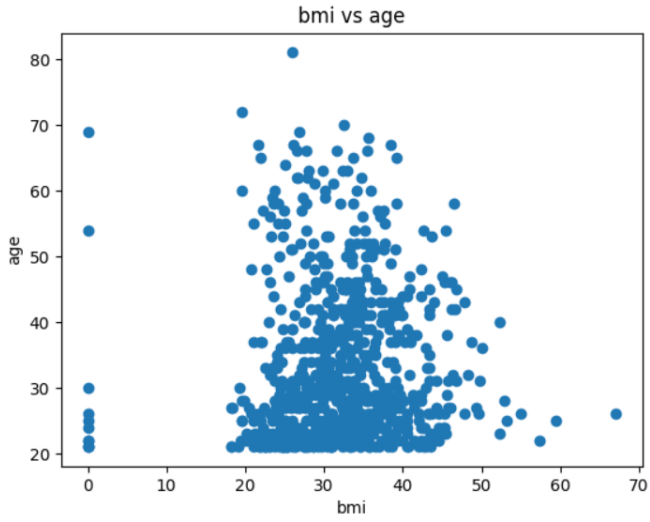


Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. These are not correlated
2. Patients having BMI between 20 to 40 kg/m^2 are below 40 years of age
3. Inference 3(You may add or delete the number of inferences

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)	0.240mm HG
4	skin (in mm)	-0.114 mm
5	test (in mu U/mL)	-0.042 mu u/ml
6	BMI (in kg/m^2)	0.036 kg/m^2
7	pedi	0.034
8	Age (in years)	1.000 years

Inferences:

1. Age is almost not correlated with insulin test, BMI, diabetes pedigree, and triceps skinfold and is in weak correlation with plasma and pressure, while it have a moderate correlation with pregnant.
2. As age is in very less correlation with insulin test, BMI, diabetes pedigree, and triceps skinfold therefore increase in age barely have effect in these. While it is in less correlation with plasma and pressure therefore as age increases these also increases, and as we can see no. of time pregnant and age are highly correlated therefore as age increases it also increases.
3. As from the plots it can be seen that pregnant increases with increase in age therefore they are highly correlated, the value of plasma and pressure increases or decreases at a low pace therefore they are less correlated, and all other attributes changes very gradually therefore they have value of correlation near zero.

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 mm HG
4	skin (in mm)	0.393 mm
5	test (in mu U/mL)	0.198 mu U/ml
6	BMI (in kg/m^2)	1.000 kg/m^2
7	pedi	0.141
8	Age (in years)	0.036 years

Inferences:

1. BMI is almost not correlated with age and pregnant and is in weak correlation with plasma, insulin test, triceps skinfold, pressure and diabetes pedigree.
2. As BMI is in very less correlation with age and pregnant therefore increase in BMI barely have effect in these. While it is in less correlation with plasma, insulin test, triceps skinfold, pressure and diabetes pedigree therefore as BMI increases these also increases.
3. As from the plots it can be seen that the value of plasma, insulin test, triceps skinfold, pressure and diabetes pedigree increases or decreases at a low pace therefore they are less correlated, and all other attributes changes very gradually therefore they have value of correlation near zero.

4 a.

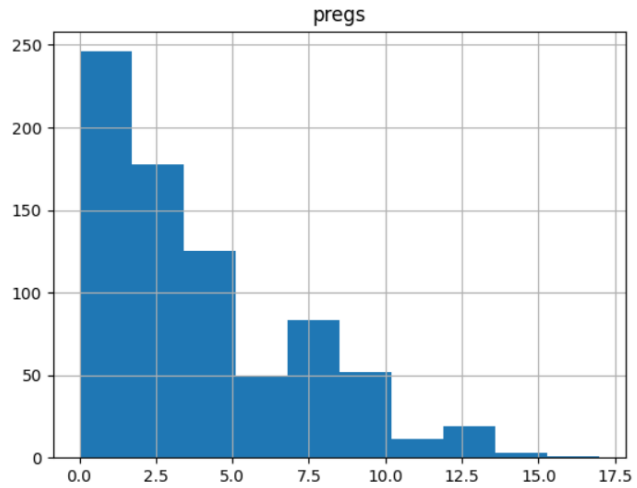


Figure 15 Histogram depiction of attribute pregs

Inferences:

1. As from the histogram it can be seen that the range 0-2 have very high frequency with value almost equal to 250, and the value frequency decreases as the no. of times pregnant increases and it decreases to almost zero in range 16-17.
2. As the frequency of first bin is highest therefore the mode lies in the first bin.

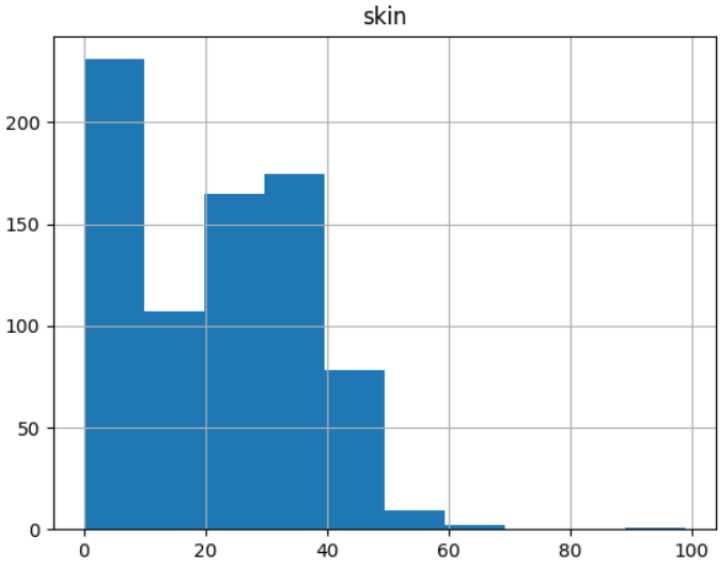


Figure 16 Histogram depiction of attribute skin

Inferences:

1. As from the histogram it can be seen that the range 0-10 have very high frequency with value almost equal to 240, and the value frequency decreases or increases as increases in the skinfold measurement value.
2. As the frequency of first bin is highest therefore the mode lies in the first bin.

5

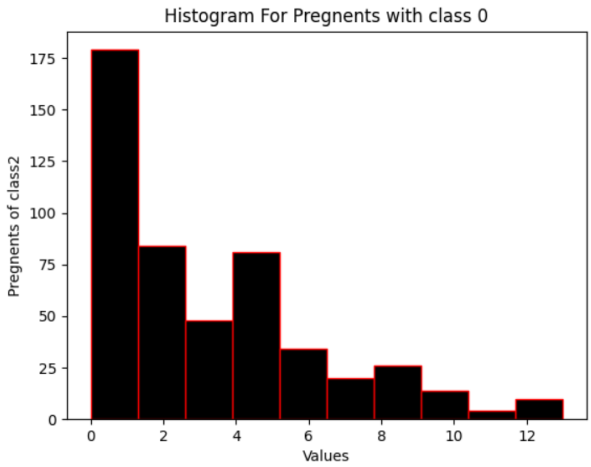


Figure 17 Histogram depiction of attribute pregs for class 0

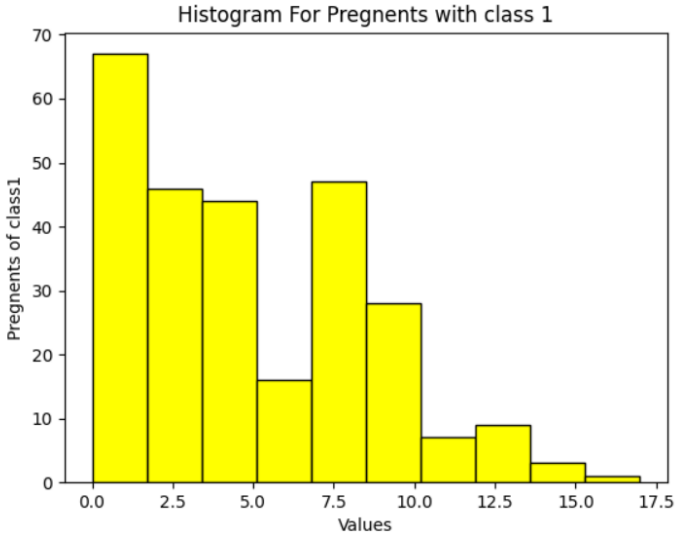


Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. As the frequency of first bin is highest of both classes (class =0 and class=1) therefore the mode lies in the first bin for both classes.
2. The frequency of bin of class 0 is higher than the frequency of bin of class 1 for range 0-5, in range 5-10 bins of class 1 are higher than bins of class 0 and in range 10-15 bins of class1 are higher than class 0.

6

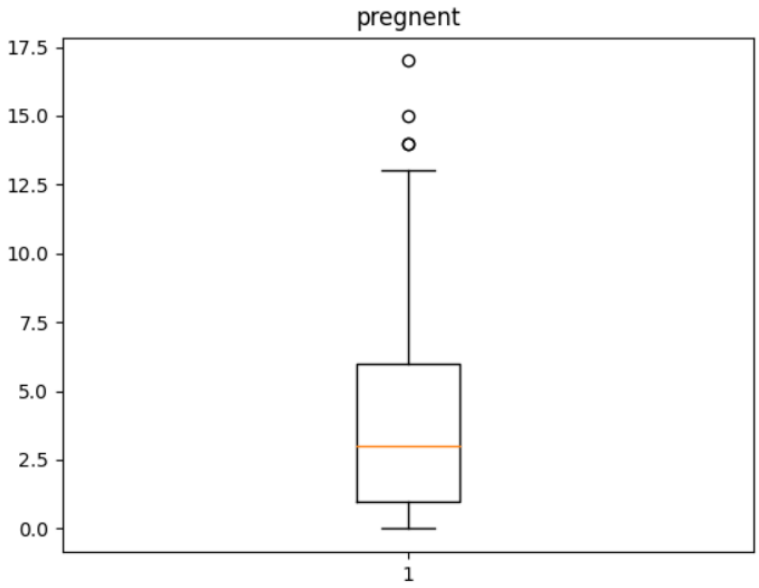


Figure 19 Boxplot for attribute pregs

Inferences:

1. Outliers lies between 14-17 which is larger than 13.5 (Q3 +(1.5*IQR))
2. Inter quartile range is 5 (Q3-Q1)
3. The variable value of no. of time pregnant varies in 0-13 and some values are greater than 13
4. The plot shows that data is skewed right
5. From the plot it can be observed that median is around 3(in Q1 it is 3.000), as per Q1 minimum and maximum are 0 and 17 resp. and the same can be seen in the plot.

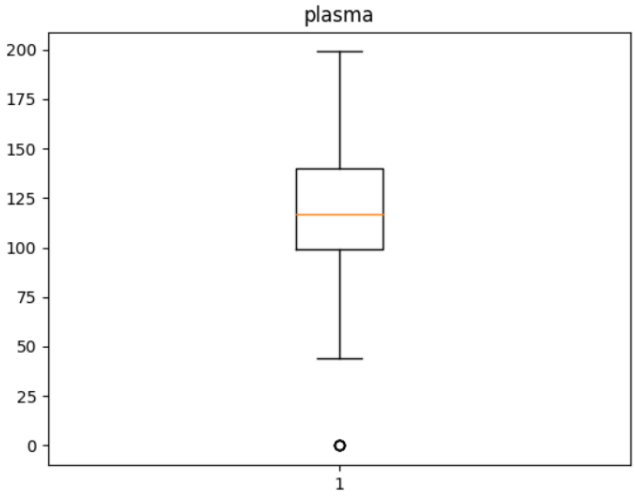


Figure 20 Boxplot for attribute plas

Inferences:

1. Outliers lies around 0 which is smaller than 40 (Q1 -(1.5*IQR))
2. Inter quartile range is 40(Q3-Q1)
3. The variable value of plasma varies between 40-200 and have positive spread
4. The plot shows that data is skewed right
5. From the plot it can be observed that median is around 120(in Q1 it is 117), as per Q1 minimum and maximum are 0 and 199 resp. and the approx. same can be seen in the plot.

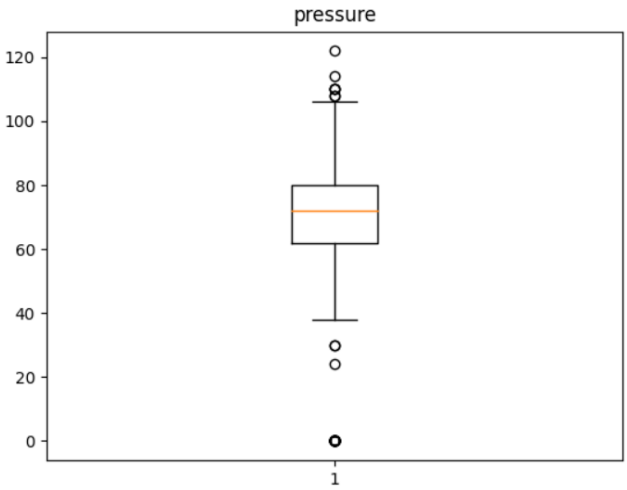


Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. Outliers lies between 0 and 30 mm HG which is smaller than 32.5mm HG (Q1 -(1.5*IQR)) and 109 and 120 mm HG which is larger than 108.5 mm HG (Q3 +(1.5*IQR))
2. Inter quartile range is 19 (Q3-Q1)
3. The variable value of no. of time pregnant varies in 40-105 and some values are greater than these and have positive spread
4. The plot shows that data is skewed left
5. From the plot it can be observed that median is around 73 mm HG (in Q1 it is 72 mm HG), as per Q1 minimum and maximum are 0 and 122 mm HG resp. and the approx. same can be seen in the plot.

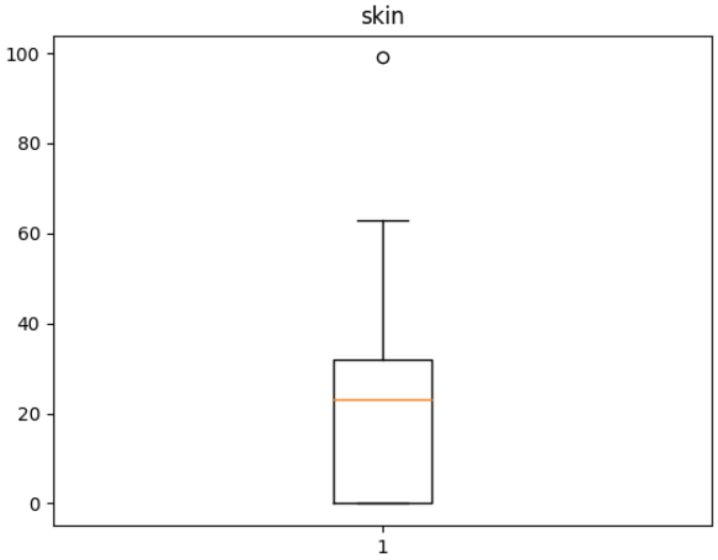


Figure 22 Boxplot for attribute skin (in mm)

Inferences:

1. Outliers lies around 100mm which is larger than 80 mm(Q3 +(1.5*IQR))
2. Inter quartile range is 32mm (Q3-Q1)
3. The variable value of no. of time pregnant varies in 0-62.5 mm.
4. The plot shows that data is skewed left
5. Values are approx. same

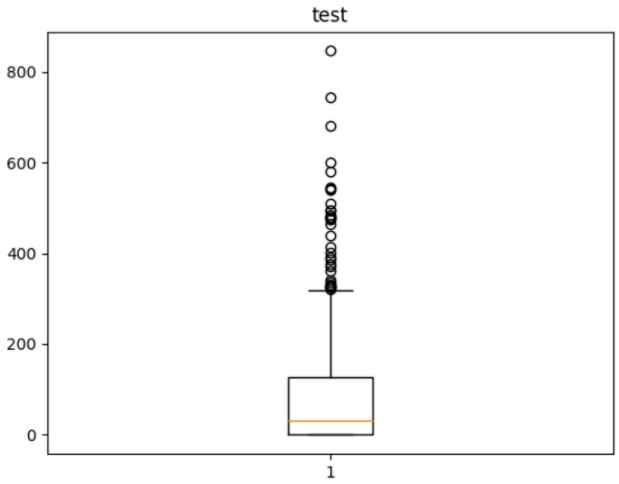


Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. Outliers lies between 380-850 mu u/ml which is larger than 375 mu u/ml (Q3 +(1.5*IQR))
2. Inter quartile range is 150mm (Q3-Q1)
3. The variable value of no. of time pregnant varies in 380-850 mu u/ml.
4. The plot shows that data is skewed right
5. Values are approx. same

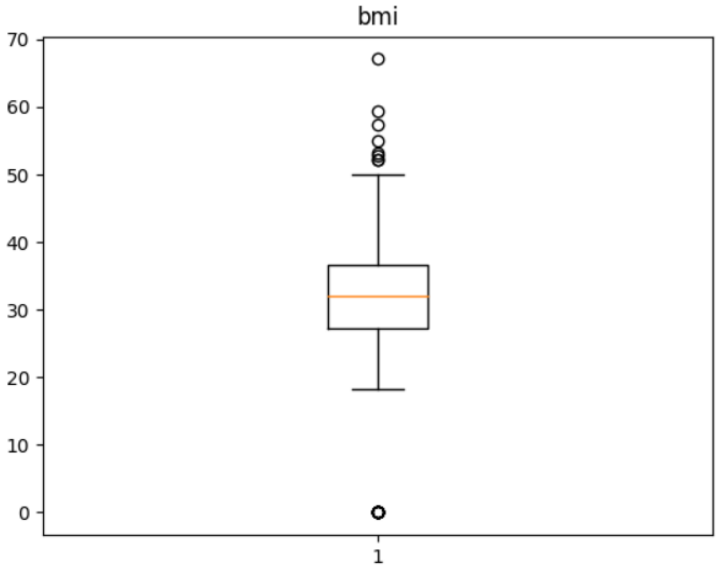


Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. Outliers lies between 50-70 kg/m^2.
2. Inter quartile range is 27-36 kg/m^2 (Q3-Q1)
3. The variable value of no. of time pregnant varies in 18-50 kg/m^2
4. The plot shows that data is symmetric
5. Values are approx. same

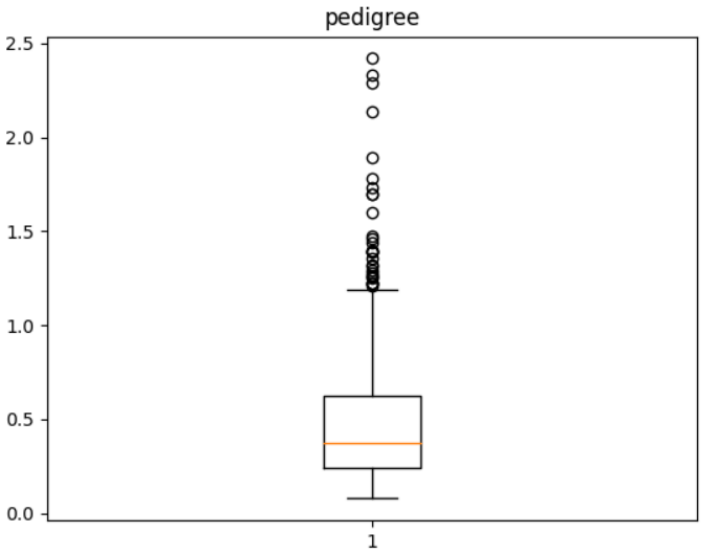


Figure 25 Boxplot for attribute pedi

Inferences:

1. Outliers lies between 1.2-2.4
2. Inter quartile range is 0.25-0.6 (Q3-Q1)
3. The variable value of no. of time pregnant varies in 0.8-1.2
4. The plot shows that data is symmetric
5. Values are approx. same

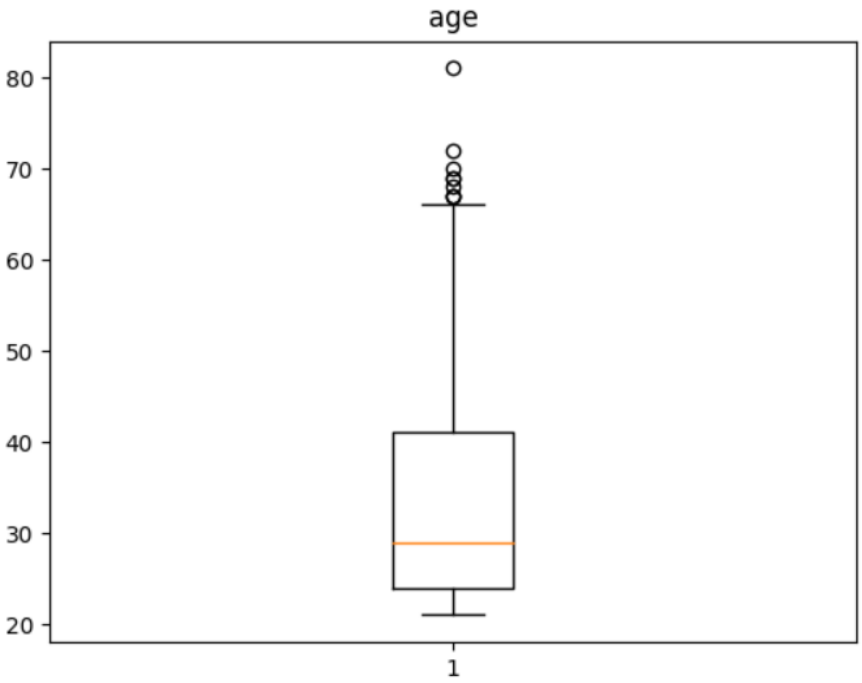


Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. Outliers lies between 65-80 years
2. Inter quartile range is 24-40 years (Q3-Q1)
3. The variable value of no. of time pregnant varies in 20-65
4. The plot shows that data is skewed right
5. Values are approx. same