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

Student's Name: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: Computer Science and Engineering**

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	3.000	1	0	17.000	3.369
2	plas	120.894	117.000	99	0	199.000	31.973
3	pres (in mm Hg)	69.105	72.000	70	0	122.000	19.356
4	skin (in mm)	20.536	23.000	0	0	99.000	15.952
5	test (in mu U/mL)	79.799	30.500	0	0	846.000	115.244
6	BMI (in kg/m^2)	31.992	32.000	32	0	67.100	7.884
7	pedi	0.472	0.373	0.254	0.078	2.420	0.331
8	Age (in years)	33.240	29	22	21.000	81.000	11.760

Inferences:

1. We can see that Mean, Median and Mode has not much difference in their value i.e. they are near to each other showing that all the data is symmetrical about some value.
2. Seeing standard deviation one can also find that concentration of Serum insulin is most spread and that of Diabetes pedigree function is having least.
3. For skin thickness and insulin test, mode and minimum value is matching which infers that minimum valued data is more in number.

2 a.

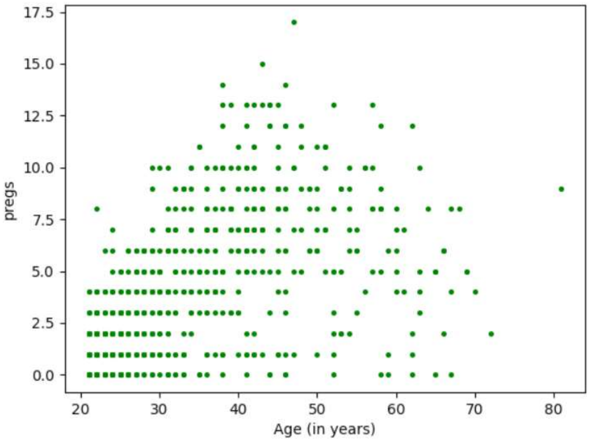

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. Increasing with the age the pregnancy count is increases hence they are positively correlated.
2. As the Age increase number of patients are almost same but the frequency of pregnancy is spread.
3. The dot density can be seen at lower age which infer that most of the patients are of lower age.
4. Peak of data is at middle value hence it infer that pregnancy count for the age in between 30 to 55 is more than that of other age patients.
5. It can also be observed that more patients were pregnant less than 7 time.

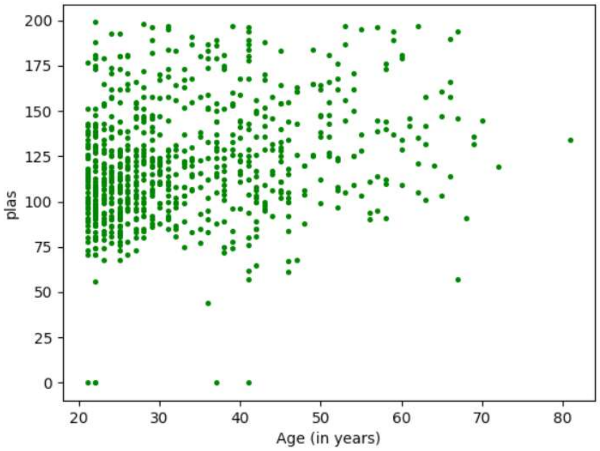

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. For the smaller age most of the patients has less plasma concentration values while as the age increases the patients have high plasma concentration values mostly, hence one can infer that they are slightly positively correlated.
2. With the increase in age plasma glucose concentration spread remains constant.
3. Most of the patients of are of plasma density (specially in between 75 to 150 concentration) and are of lower age(mostly in between 21 to 40 years).
4. Almost all patient's plasma concentration lies in between 70 to 200 unit.

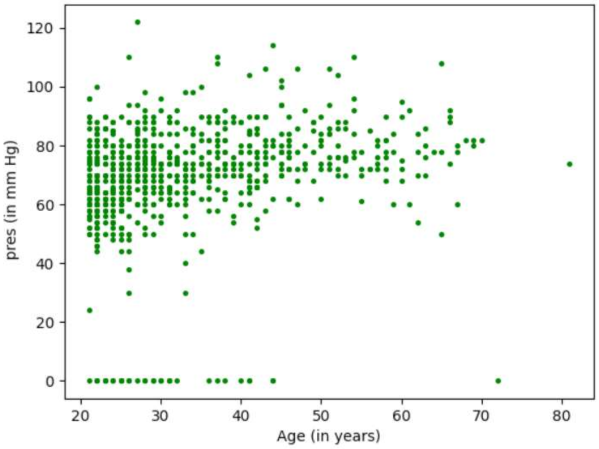

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. In the smaller age most of the patients has less blood pressure values while as the age increases the patients are left with high blood pressure mostly, hence they are slightly positively correlated.
2. As the age increases Diastolic blood pressure spread remains constant.
3. Most of the patients of different plasma density (specially in between 50 to 90 mm Hg) are of lower age(mostly in between 21 to 40 years).
4. For the age of 20 to 45 there is a huge gap in the minimum diastolic blood pressure and second minimum diastolic blood pressure.

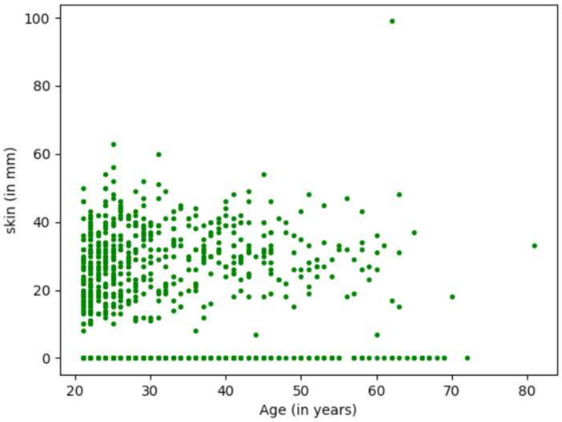

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. As the age increases skin thickness population density decreases hence it infers that they are slightly negatively correlated.
2. Most of the patients are of lower age
3. Mostly the thickness of skin is of 10mm to 45 mm.
4. Here one can see a small gap of 10mm in the minimum and second minimum skin thickness.

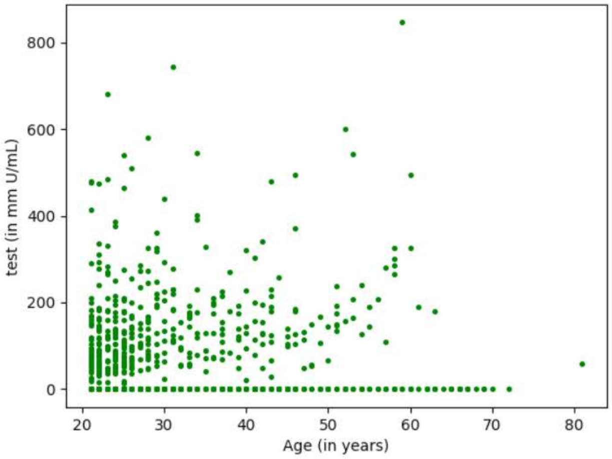

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. With the increasing of the age the patients of higher concentration are decreasing hence it infers that they are negatively correlated.
2. Mostly the patients are of lower age.
3. Spread of the data is consistent but the number of patients decreases with increase in age.
4. For almost every age we have patients of insulin concentration 0 mm U/mL.

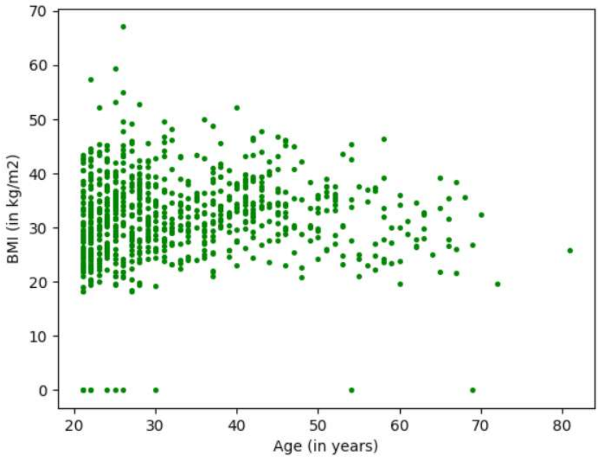

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. Since with the increasing of age patients with low mass index as well high mass index both decreases hence it is hard to say that are correlated to each other.
2. As the age increases the spread of the data slightly decreases.
3. Most of the patients are of lower age and their body mass index is in range of 19 to 45 kg/m^2.
4. Very few patients have body mass index 0 kg/m^2 and that gap can be seen in the graph.

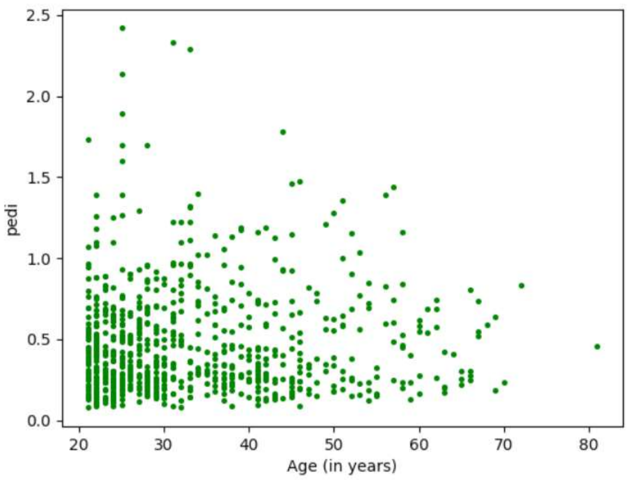

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. With increase in age patients density decreases for smaller pedigree function value which infers that their correlation is slightly positive.
2. As the age increases Diabetes pedigree function population density decreases means patients decreases.
3. Most of the patients are of age 21 to 50 years and have Diabetes pedigree function in between 0 to 1 unit hence most diabetic patients have not the disease inherited.

b.

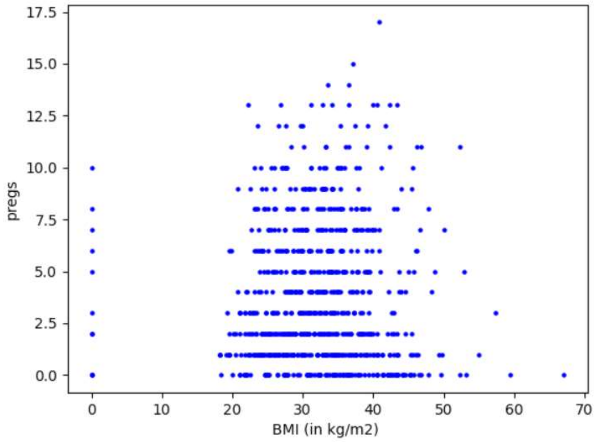

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. With increase in mass index the patients decreases uniformly in for every pregnancy count hence we do not get any inference for correlation.
2. Data is symmetrical.
3. Corresponding to a range of BMI (in 18 to 50 kg/m^2) almost all the patient pregnancy data lies.
4. There is no patients in between BMI of 0 to 17 kg/m^2.

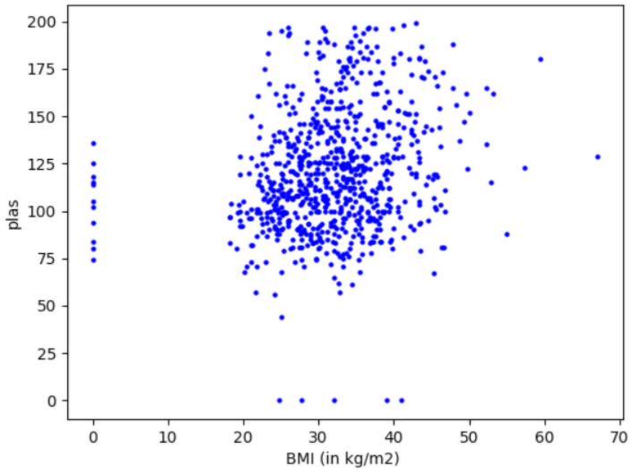

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. With the increasing of body index, patients of lower plasma concentration decreases which infers that BMI is positively correlated with the plasma concentration.
2. All the data is condensed i.e. for particular range of values of BMI and plasma concentration, all the patients lie.
3. There is a large gap in between minimum BMI and second minimum BMI value.
4. There is also large gap in minimum plasma concentration and second minimum plasma concentration.

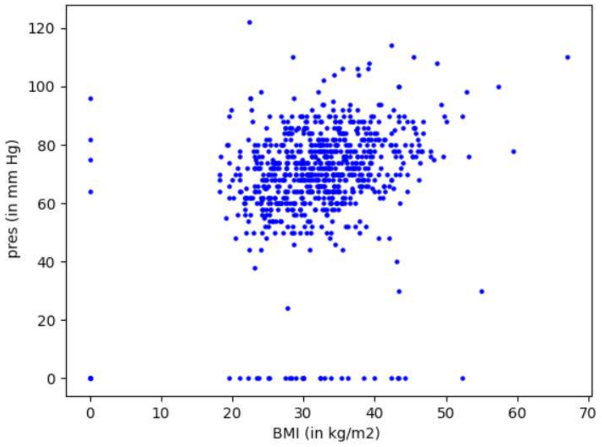

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. As the blood pressure slightly increases with the increase in BMI value hence one can say that both are slightly positively correlated.
2. Data is condensed.
3. There is a large gap between min blood pressure and second minimum blood pressure of the patients and similarly there is large gap in between minimum BMI value and second minimum BMI value.

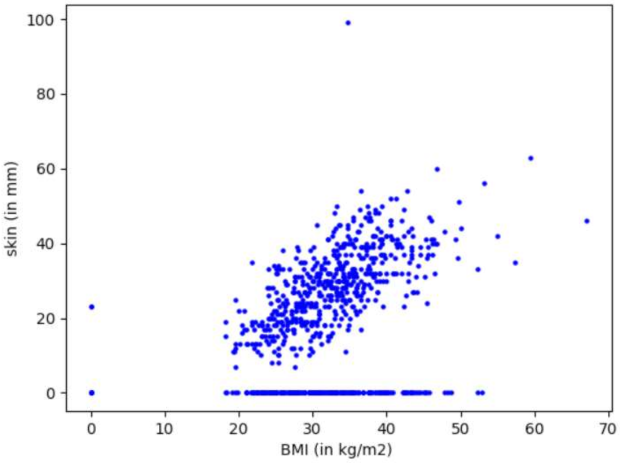

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. Skin thickness increases with increase in BMI value hence one can say that they are positively correlated.
2. Inference based on density of points
3. Inference 3(You may add or delete the number of inferences)

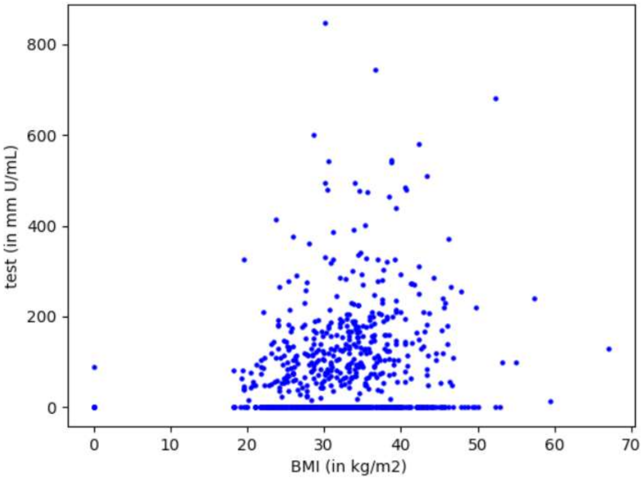

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. Since the insulin test concentration is slightly increasing hence they are slightly positively correlated to each other.
2. Most of the patients lies in some range of values for BMI and concentration of insulin test.
3. For every specified above range age there is some patients having zero concentration of insulin test.

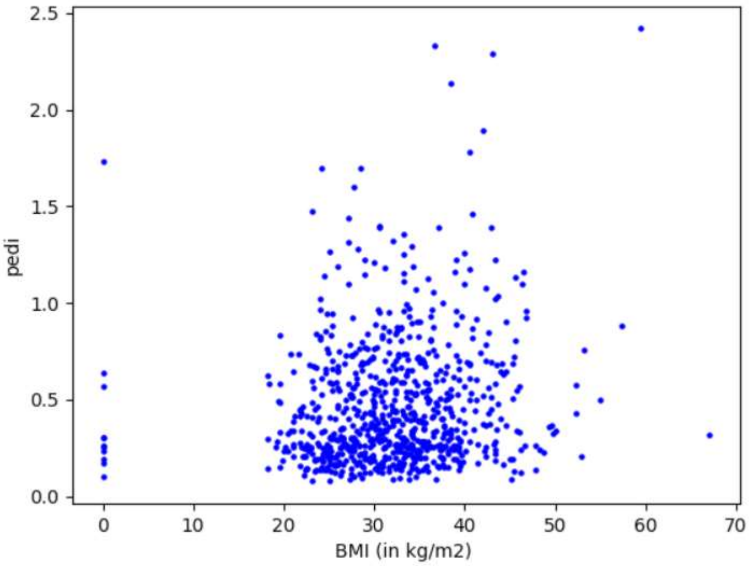

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. One can see that the value of pedigree function increase with the increase in BMI, which infer that the pedigree function and BMI is slightly positively correlated to each other.
2. Data is condensed means most of the patient lie in some specific range values of BMI and pedigree function.
3. There is large gap in between the minimum BMI value and second minimum BMI value.
4. Most of the patients are having less pedigree function value i.e. about 0 to 1.

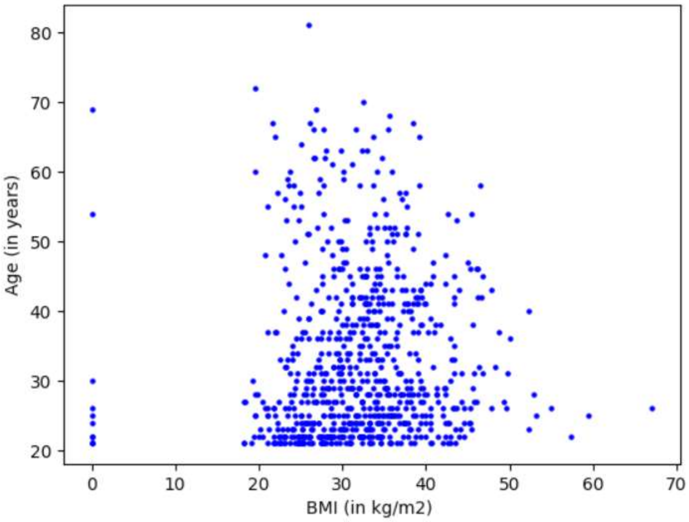

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. Corresponding to every BMI value every age of patients can be found hence there is no correlation one can find.
2. Data is slightly dense for age in between 21 to 50 years corresponding to 20 to 45 kg/m^2 BMI value.
3. Most of the patients age is less than 50 years.

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)	-0.114
5	test (in mu U/mL)	-0.422
6	BMI (in kg/m^2)	0.036
7	pedi	0.336
8	Age (in years)	1.00

Inferences:

1. As the correlation coefficient between age and BMI is about to zero resulting no correlation between them. There is small positive correlation of age with pressure, plasma concentration, pedigree function. There is good positive correlation between age and pregnancy data.
2. Also age have small negative correlation with skin thickness but good negative correlation with insulin concentration.
3. Since in the last two points we have seen the correlation value with sign. So we can say that, as the age increases there is no effect on BMI, while pressure, plasma concentration and pedigree fuction values slightly increases and count of pregnancy increases in well frequency, on the other hand skin thickness slightly decreases while insulin concentration decreases in good amount.
4. From the graph we have found that the correlation of age with pressure, plasma concentration, pedigree function were slightly positive while there is well positive correlation with pregnancy data and slightly negatively related to skin thickness and well negative correlated with the insulin test data

   and we can see that all data corresponding to the table provided above of correlation with age to other attributes, hence the graphs were showing the correlation result in good manner.

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
4	skin (in 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. Here the correlation coefficient between BMI with age and pregnancy data is about to zero resulting no correlation between them. There is small positive correlation of BMI with insulin data, pedigree function. There is good positive correlation with skin thickness.
2. So we can say from inference(1) that as BMI increases age and pregnancy can't be depicted, while small increase in insulin data, pedigree function and good increment in skin thickness.
3. From our plot depiction we get that the correlation of BMI with insulin data, pedigree function were slightly positive while there is well positive correlation with skin thickness and no correlation with age and pregnancy data and we can match that all data corresponding to the table provided above of correlation with BMI to other attributes, and we get that the graphs were showing the correct correlation.

4 a.

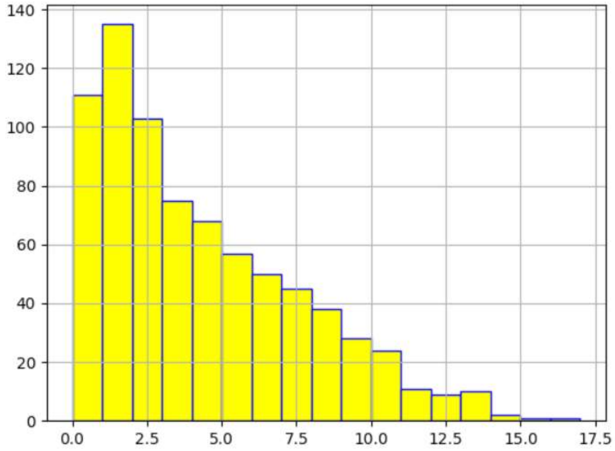

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. From the distribution we can see that as the pregnancy count increase we have very less patients.
2. From plot it can be clearly seen that it lies in second bin i.e. from 1 to 2 pregnancy.
3. One can infer that the data is positively skewed.

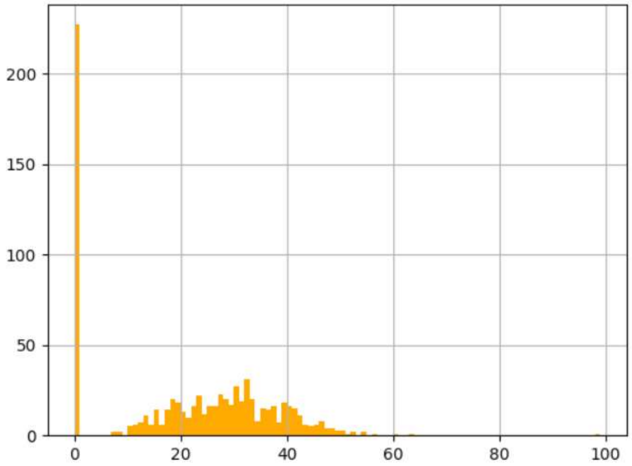

Figure 16 Histogram depiction of attribute skin

Inferences:

1. The frequency of very small thickness is high while there is normal distribution of thickness of the 10 to 55 years.
2. Mode lies in first bin.

5

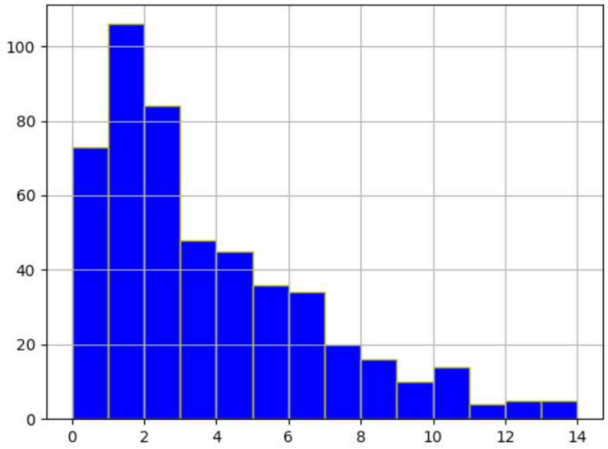

Figure 17 Histogram depiction of attribute pregs for class 0

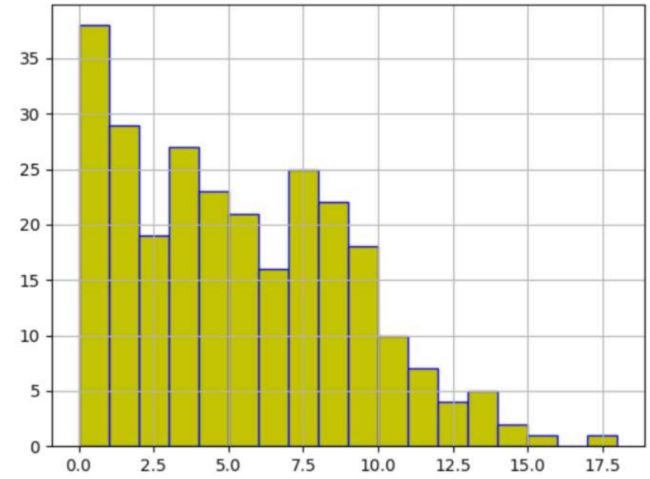

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. Seeing the graph mode inference we get that for class 0 it lies in 2nd bin (means 1 and 2 pregnancy case) while for class 1 it lies in first bin(means 0 and 1 pregnancy case).
2. Both are most like positively skewed data but the frequency of class 1 for larger pregnancy count is greater than the class 0.
3. Class 1 has patients with more pregnancy counts wrt to class 0.

6

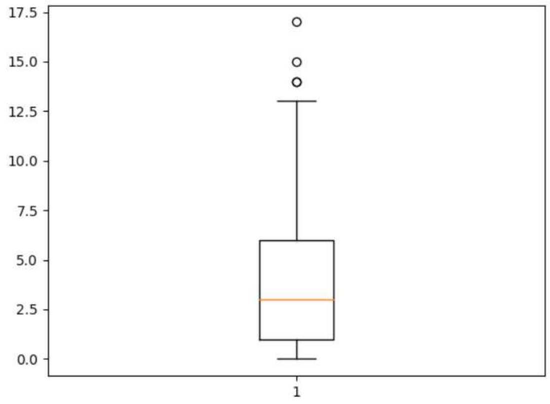

Figure 19 Boxplot for attribute pregs

]Inferences:

1. Outliers are having higher values(13, 15, 17 times pregnant case).
2. Here the interquartile range is of 5 unit having first quartile at 1 and third quartile at 6.
3. As the interquartile range is of 5 unit and our data is spread from 0 to 17.5 unit hence pregnancy count has a moderate variability.
4. As the median line is below the middle of the interquartile box hence data is positively skewed.
5. From que 1 we get that median is 3 which is correctly shown by the orange horizontal line and mode is 1 which lies in box, hence data of que 1 is matching with the boxplot data.

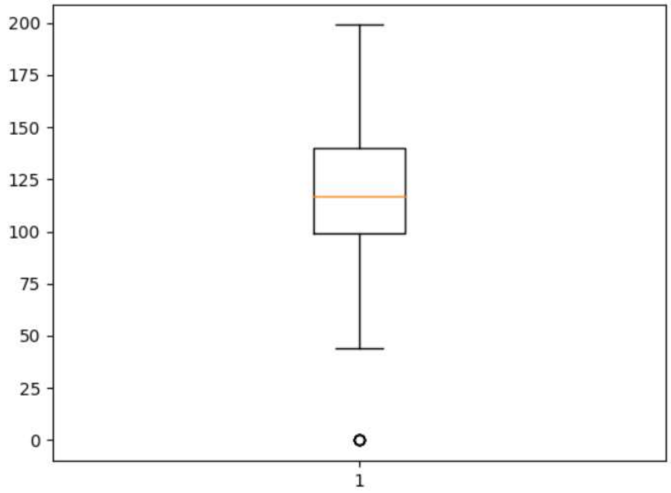

Figure 20 Boxplot for attribute plas

Inferences:

1. Here the outlier is having lower value of 1 unit.
2. The first quartile value is at about 100 unit and third quartile value is at about 140 unit hence our interquartile range is of about 40 units.
3. As there is only one outlier while all data is in boxplot range hence we can say that data has less variability.
4. As the median line is at about middle of the interquartile range hence data is symmetrical.
5. From que 1 median is 117 and mode is 99 which is well depicted by the graph.

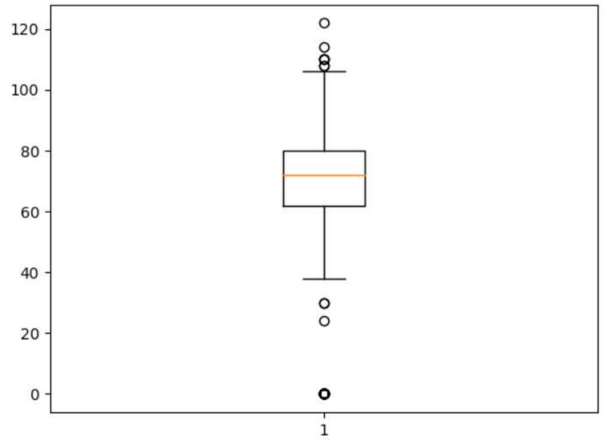

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. Here the outliers are of high values as well low values(i.e values like 106, 108, 112, 121 as well 0, 28, 34).
2. The first quartile is at 62 while third quartile is at 78, hence the interquartile range is of 16 units.
3. Since the interquartile range is of 16 units while the range is of 121 units, hence data has highly variability.
4. Median line is at about middle of the graph, hence data is symmetrical.
5. Median is 72 while mode is 70, hence orange line well depicting the median value while mode is in box hence values are matching.

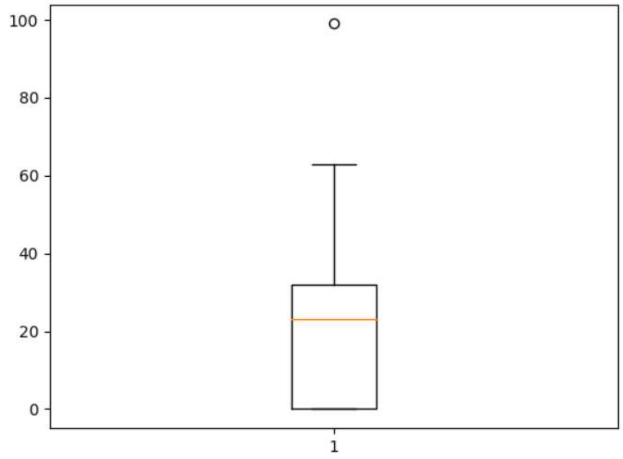

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. Their is one outlier of very high value which is 100.
2. The first quartile is at 0 while third quartile is at 36, hence the interquartile range is of 36 units.
3. Since interquartile range is of 36 units and range is of about 61 units (excluding the outlier), hence data is less variable.
4. Since median line is above the middle of the interquartile range, hence data is negatively skewed.
5. From que 1 we get, Median value is 23 which is depicted by the orange line while mode is 0 which is in box hence values are well matching.

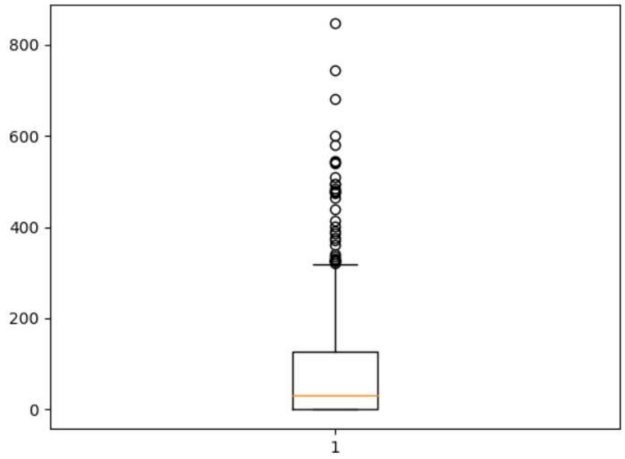

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. Here we have a lot many outliers, all having high values.
2. The first quartile is at 0 while third quartile is at 130, hence the interquartile range is of 130 units.
3. Since data has so many outliers hence data has high variability.
4. Since median line is below the middle of interquartile range, hence data is positively skewed.
5. From que 1 we get, median is 30.5 which is well represented by median line of the boxplot graph while mode is 0 which is in the boxplot hence the values are well coordinating.

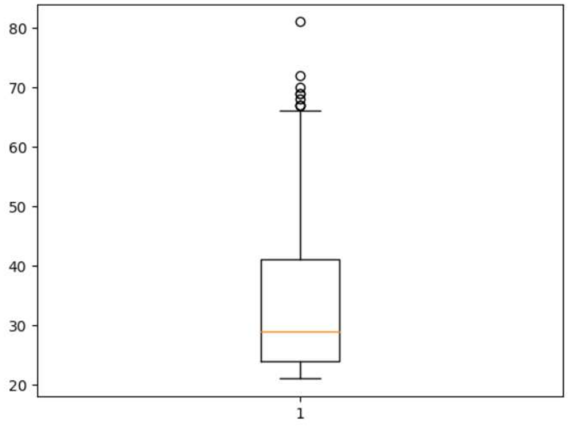

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. Since there are so many outliers with so high values hence data has high variability.
2. The first quartile is at 25 while third quartile is at 40, hence the interquartile range is of 15 units.
3. Since interquartile range of the data is about 15 units and range is about 60 units(with outliers) and 45 units without outliers, hence data has moderate variability.
4. Since median line is below the middle of interquartile range, hence data is slightly positively skewed data.
5. From que 1 we get that median is 32 while mode is also 32 which is correctly coordinating with the graph.

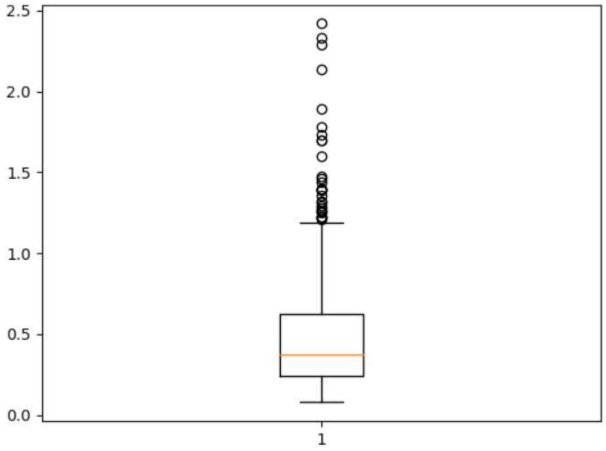

Figure 25 Boxplot for attribute pedi

Inferences:

1. Here we have outliers of high values.
2. The first quartile is at 0.25 while third quartile is at 0.65 hence quartile range is of about 0.4 units.
3. Since interquartile range of the data is about 0.4 units and range is about 2.5 units(with outliers) and 1.3 units without outliers, hence data has high variability.
4. Since the median is below the middle of interquartile range, hence data is positively skewed.
5. From que 1 we get that median is 0.37 which is depicted by orange line in the graph and mode is 0.25 which is also coordinating with the box.

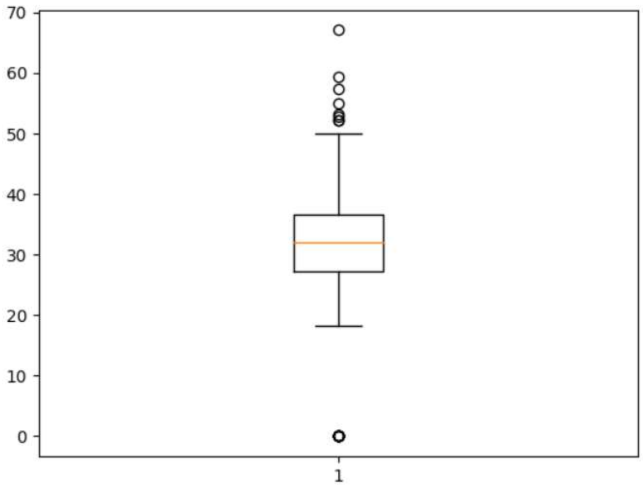

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. Here we have both high values outliers as well lower values outliers.
2. The first quartile is at 28 and third quartile is at 36, hence the interquartile range is of 8 years.
3. Since the interquartile range is of about 8 years and range is about 68 years(with outliers) and 30 years (without outliers) hence data is moderately variable.
4. Since median line is middle of the interquartile range, hence data is symmetrical.
5. Referring to que 1 we get that median is 29 years and mode is 22 years which is coordinating with out boxplot.