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

Student's Name: Nikhil **Mobile No: 8949463760**

Roll Number: B20219 **Branch: Computer Science & 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	3.37
2	plas	120.895	117.000	99;100	0	199	31.973
3	pres (in mm Hg)	69.105	72.000	70	0	122	19.356
4	skin (in mm)	20.536	23.000	0	0	99	15.952
5	test (in mu U/mL)	79.799	30.500	0	0	846	115.244
6	BMI (in kg/m^2)	31.993	32.000	32	0	67.1	7.884
7	pedi	0.472	0.372	0.254;0.258	0.078	2.42	0.331
8	Age (in years)	33.241	29.000	22	21	91	11.76

Inferences:

1. From the above observations we see that mean, median and mode are not necessarily closer to each other when standard deviation is lower. Hence, we conclude that there is no relation between standard deviation and mean, mode and median.

2 a.

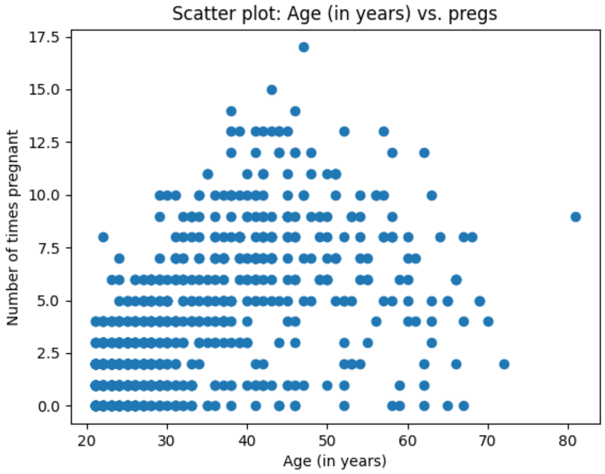

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. As age increases, we see that there is more range to the number of times a woman is pregnant. There is a slight shift of the number of times of pregnancy to the higher side as age increases. Hence, we conclude that correlation between age and pregs is positive.
2. The plot shows that there is more density of points on the lower age, lower number of pregnancy side, concluding that a greater number of women participating in the survey were of smaller age and had a smaller number of pregnancies.

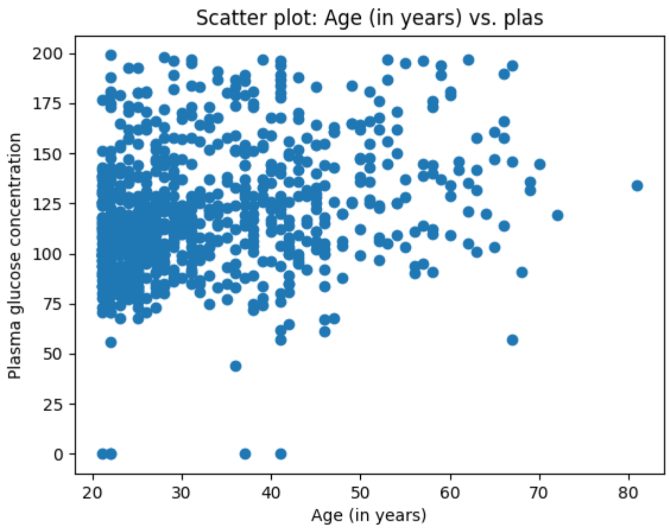

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. As age increases, the concentration of points on the higher plasma glucose concentration increases slightly, indicating that Age has a small positive correlation coefficient with plasma glucose concentration.
2. There is more density of points on the lower age, lower concentration side, indicating that a greater number of women in the survey were young and had lower concentration of plasma glucose.

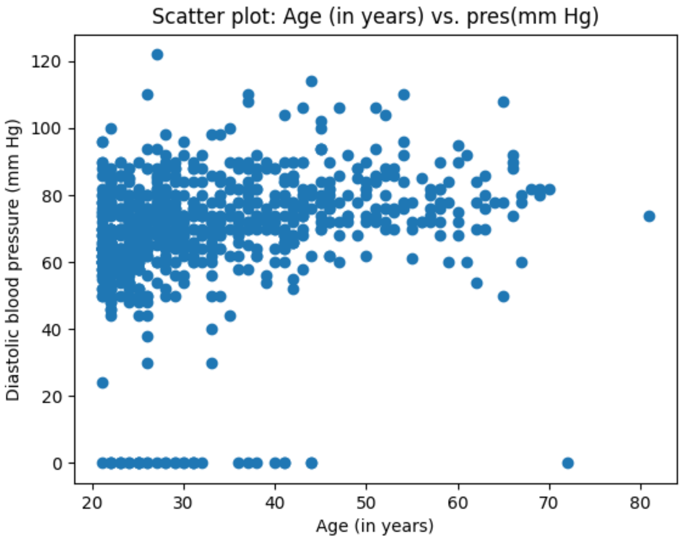

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. The plot shows that there is only a very slight shift of points to the higher-pressure side as age increases. This suggests a very low positive correlation coefficient.
2. The density of points is more on low pressure, lower age side, suggesting that a greater number of participating women were young and had lower blood pressure.

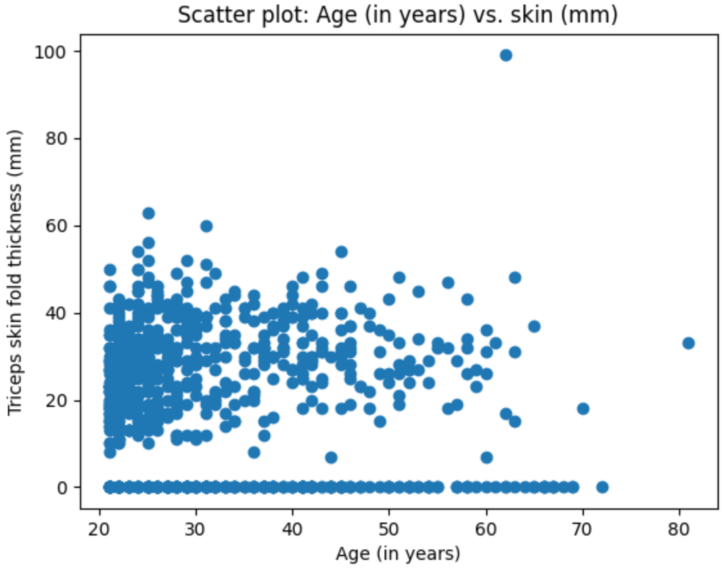

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. The plot shows that there is almost no relation between increasing age and skin thickness. Since we see a very slight shift of thickness to lower side as age increases, the correlation is almost negligible and on the negative side.
2. More density of points on the lower age side shows that more number of women participating were younger.

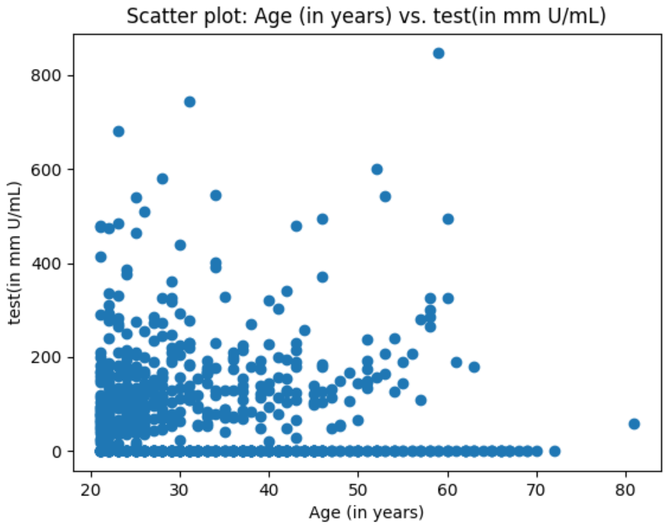

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. Since we see no net directional shift of points on the higher age side, we conclude that increasing age has no relation with the test.
2. We see more spread of points on the higher age side, suggesting that older women are more likely to have a very low as well as very high test. Also, more density of points on the lower age side suggests that more women participating were young.

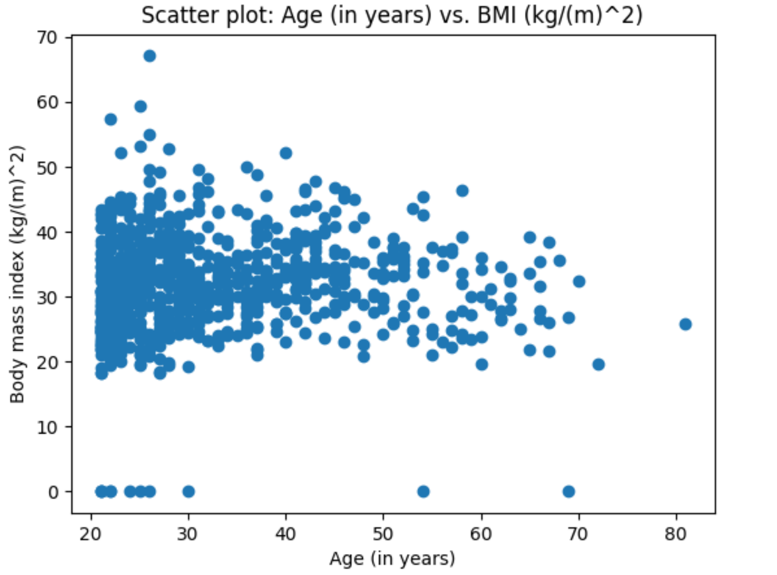

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m²)

Inferences:

1. The plot shows almost no change in distribution(orientation) if the points with increase in age, suggesting that BMI is independent of age.
2. More density of points on the lower age side suggests that a greater number of women who participated were younger. We see wider range of points on the younger side but this does not mean that younger women are more likely to have greater BMI, as the survey had more number of younger women participating than the old women.

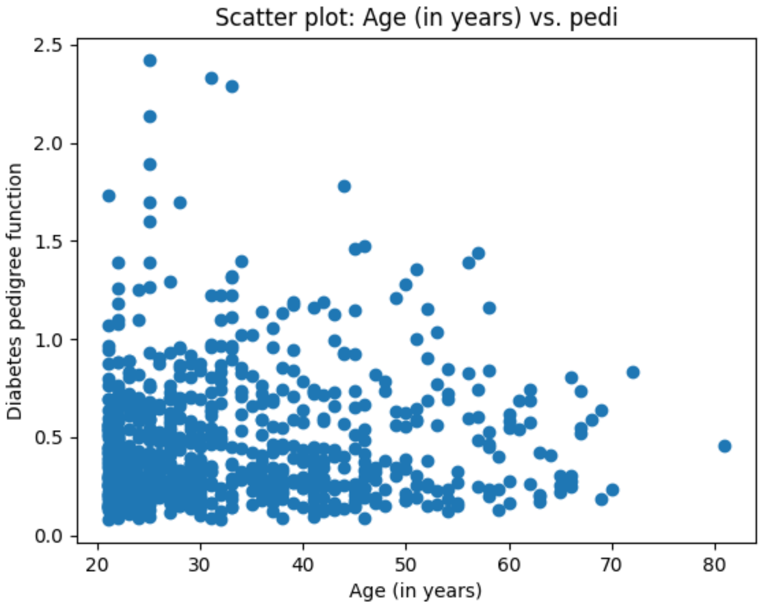

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. We see almost no change in orientation of points on the older age side(though we see less spread, but that is because of lower number of women on older age side), suggesting that DPF is independent of Age.
2. More density of points on the lower age side suggests that a greater number of women who participated were younger.

b.

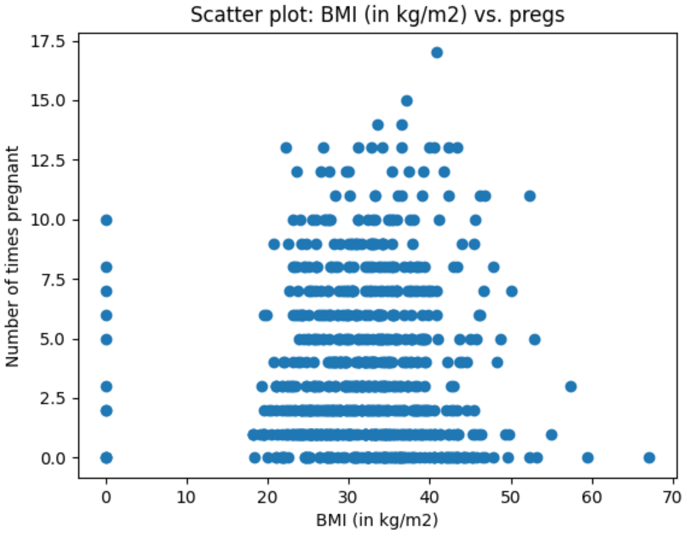

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. Based on the orientation of points from low to high BMI, we can see that there is no significant shift of points in any direction, there is only a change in density. This suggests that they are independent.

2. We see most density on the BMI range 20-40, suggesting that a greater number of women who participated were overweight.

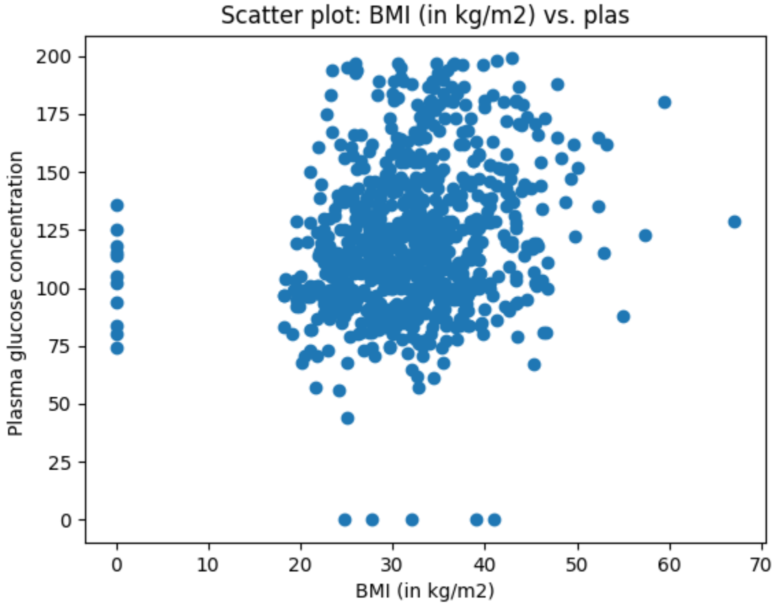

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. In the plot we can see that there is a shift of points on the higher BMI side to more plasma glucose concentration. This suggests that the women who are more overweight are sightly more likely to have more plasma concentration. The correlation is a moderate and positive.
2. We see most density on the BMI range 20-40, suggesting that a greater number of women who participated were overweight and had mediocre plasma concentration.

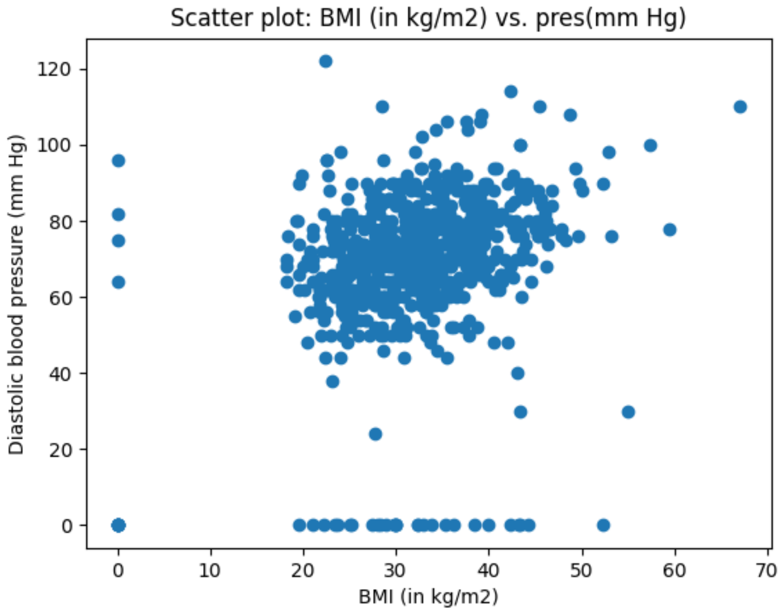

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. We see a shift of points to the higher blood pressure side as BMI increases, suggesting that BMI is moderately positively correlated with blood pressure.
2. We see most density on the BMI range 20-40, suggesting that a greater number of women who participated were overweight.

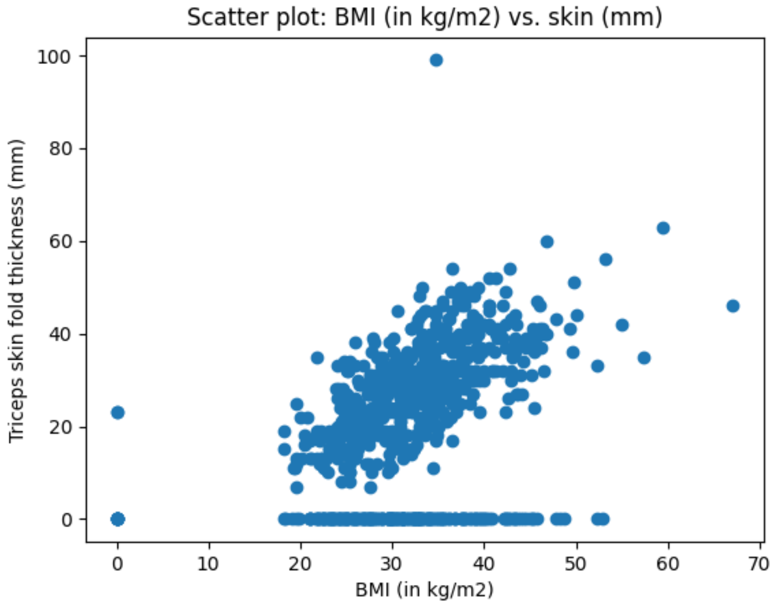

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. We clearly see a shift of points to the higher skin thickness side as BMI increases, indicating that BMI has a moderate to strong positive correlation with skin thickness.
2. We see most density on the BMI range 20-40, suggesting that a greater number of women who participated were overweight.

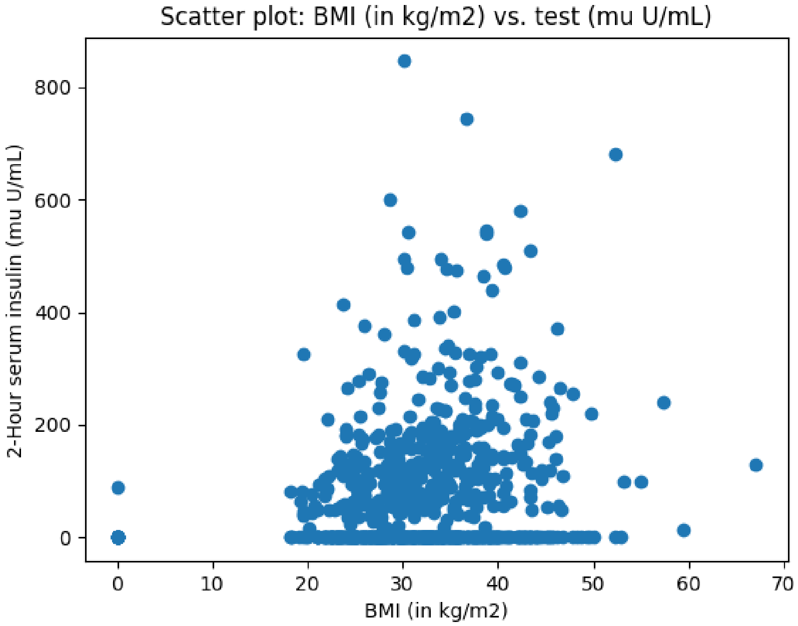

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. We see a very slight shift of points to the higher ‘test’ side as BMI increases, indicating a weak positive correlation between them.
2. We see most density on the BMI range 20-40 and ‘test’ range 0-200, suggesting that a greater number of women who participated were overweight and showed ‘test’ less than 200 U/mL.

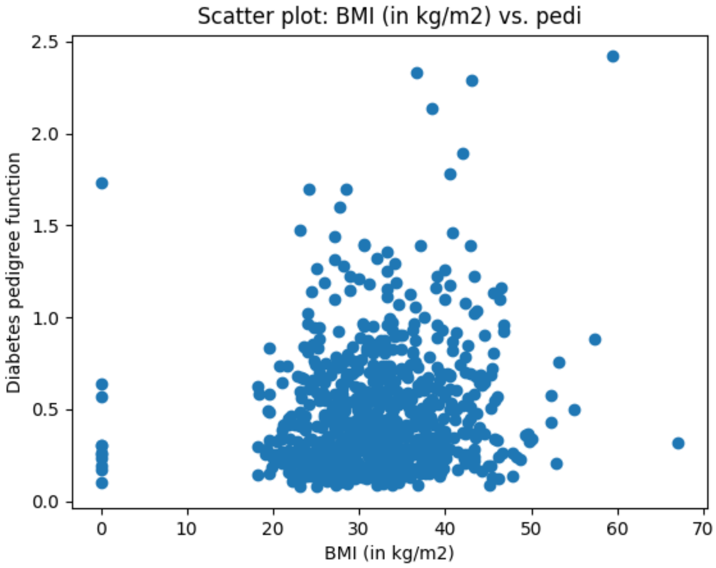

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. Since we see almost no shift of points towards the higher DPF side as BMI increases, the correlation coefficient is a very small quantity.
2. We see most density on the BMI range 20-40 and DPF<1.0, suggesting that a greater number of women who participated were overweight and had DPF<1.0.

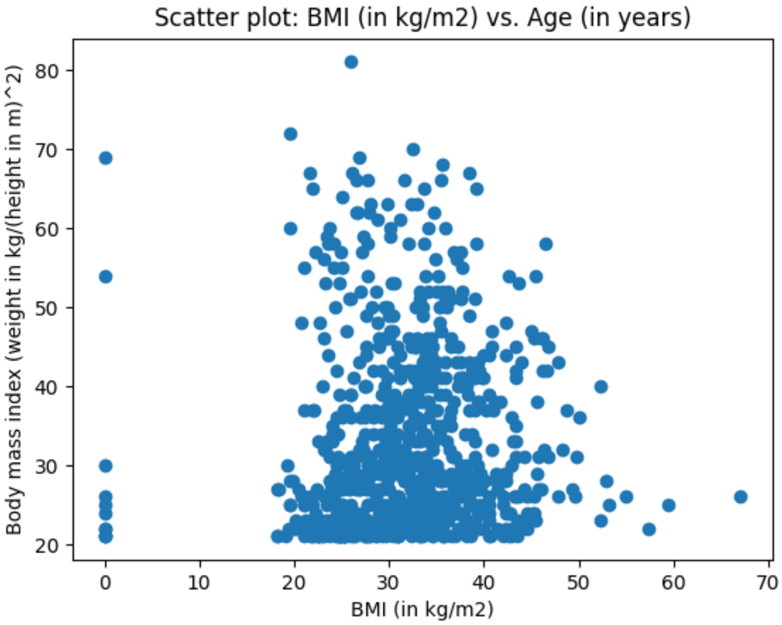

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. Since we see almost no shift of points towards the higher DPF side as BMI increases, the correlation coefficient is a very small quantity.
2. We see most density on the BMI range 20-40, suggesting that a greater number of women who participated were overweight.

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)	0.24
4	skin (in mm)	-0.114
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036

7	pedi	0.034

Inferences:

1. Strength of correlation coefficient of age with all the attributes and comment on whether with increase in age each of the attributes will increase or decrease:

Attribute	Strength	Behavior on increase in Age
pregs	Moderate	Increase
plas	Weak	Increase
pres	Weak	Increase
skin	Very Weak	Decrease
test	Very Weak	Decrease
BMI	Very Weak	Increase
pedi	Very Weak	Increase

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	pedi	0.141
7	Age (in years)	0.036

Inferences:

1. Strength of correlation coefficient of BMI with all the attributes and comment on whether with increase in BMI each of the attributes will increase or decrease:

Atribute	Strength	Behavior on increase in BMI
Pregs	Very weak	Increase
Plas	Weak	Increase
Pres	Weak	Increase
Skin	Weak	Increase
Test	Weak	Increase
Pedi	Very weak	Increase
age	Very weak	Increase

4 a.

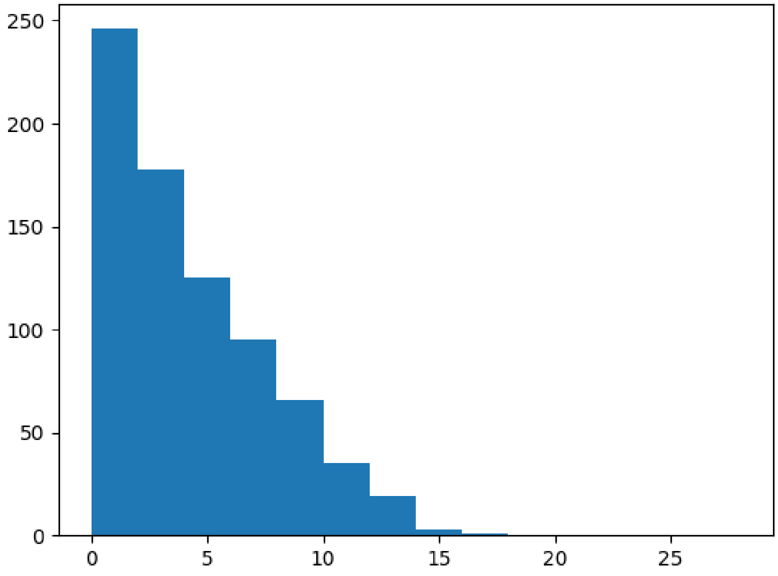

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. We can see that the frequency of women with 0-2 times pregnant is very high and almost reaches 250. Frequency in each bin after this bin monotonically decreases, with reaching almost 0 in the bin 16-18, showing that a woman is lesser likely to be more number of times pregnant.
2. Since the bin 0-2 has most frequency, the mode must lie in this bin.

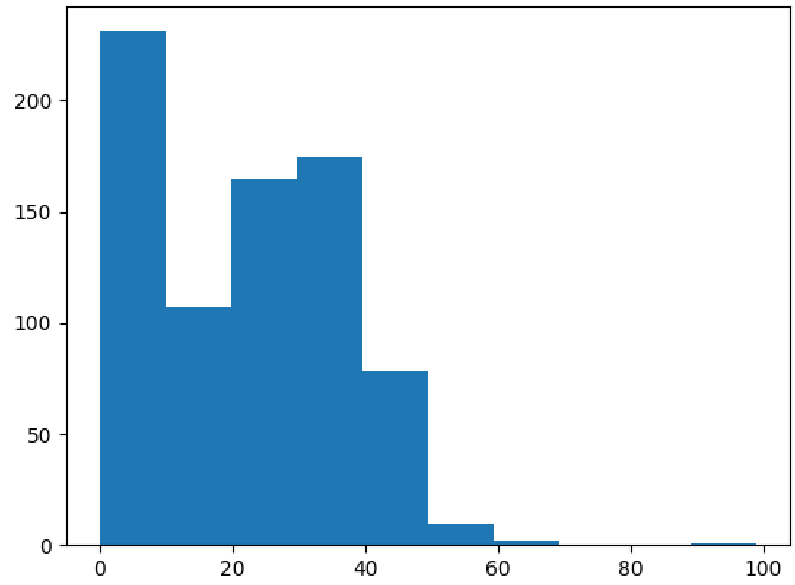

Figure 16 Histogram depiction of attribute skin

Inferences:

1. From the plot it can be seen that the bin 0-10 is the highest frequency, followed by the bin 30-40, and the bin 20-30. Bins 40-50 and 50-60 have the lowest frequency. This indicates that a woman is more likely to have a very thin or moderately thick skin.
2. The bin 0-10 has the highest frequency so the mode thickness of the skin has to lie in this bin.

5

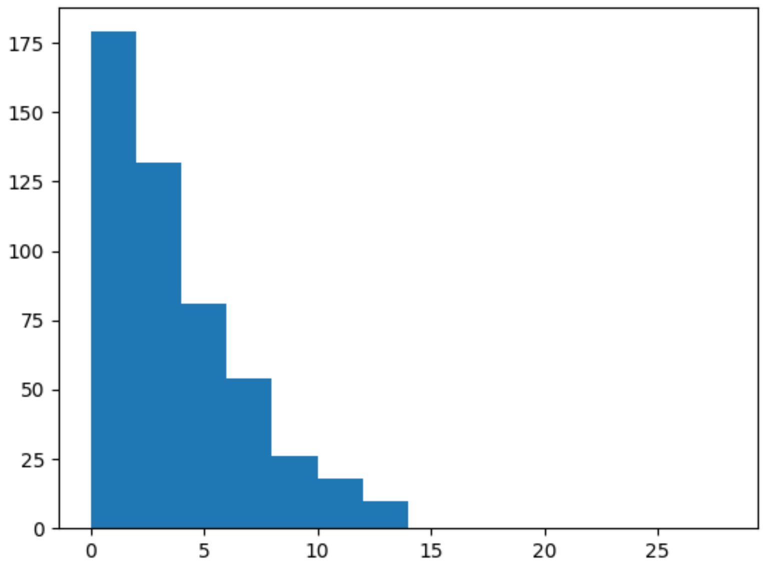

Figure 17 Histogram depiction of attribute pregs for class 0

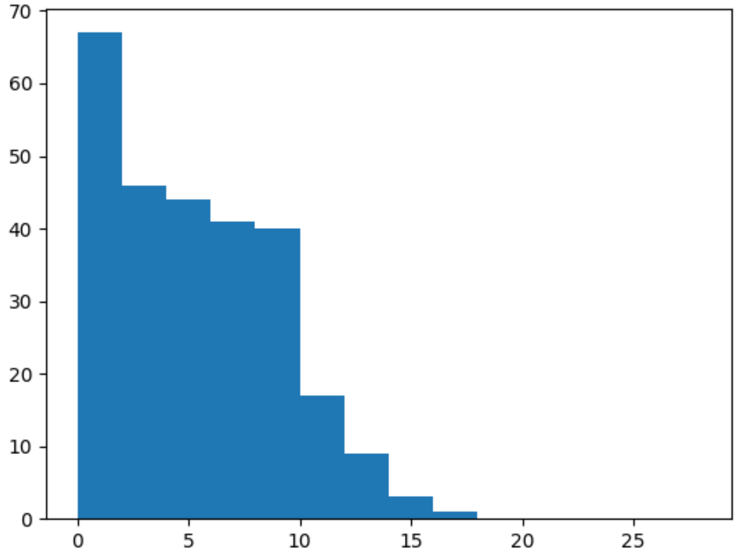

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. From the plot is can be clearly seen that the bin 0-2 has the highest frequency for both 0 and 1 class women. Hence, mode number of pregnancies for each class also lies in this bin.
2. We can see that the frequency of women with 0-2 times pregnant is very high in both classes, as compared to the other bins. Frequency in each bin after this bin monotonically decreases, with reaching almost 0 in the bin 14-16 for class 0 and bin 16-18 for class 1, showing that a woman is lesser likely to be more number of times pregnant, no matter in which class she lies.

.

6

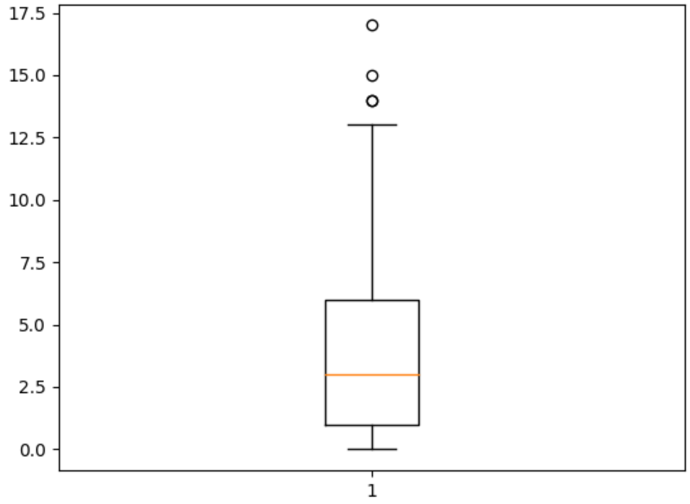

Figure 19 Boxplot for attribute pregs

Inferences:

1. There are 3 outliers on the maximum side with values 14, 15 and 17
2. Inter quartile range is Q3-Q1 = 6-1 = 5
3. The distribution is spread such that it has max value 17, and upper whisker at 13, min value 0 and lower whisker also at 0. It is spread more on the right side than the left with a standard deviation = 3.370
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 3 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 17 and 0 which is exactly what we see in the plot.

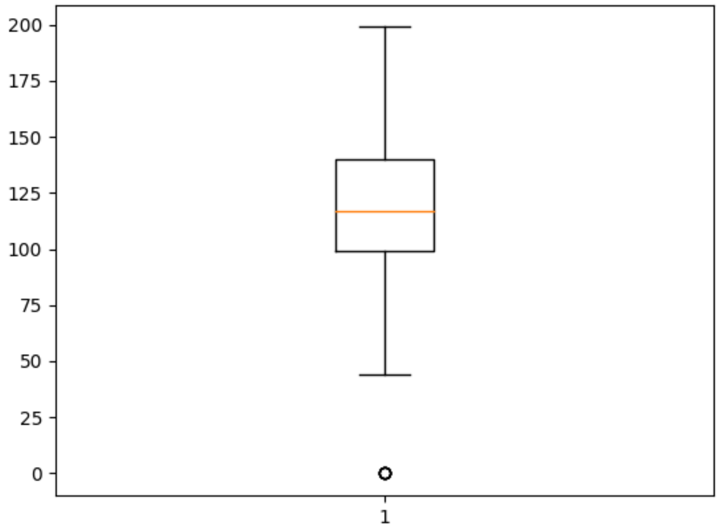

Figure 20 Boxplot for attribute plas

Inferences:

1. There is 1 outlier on the minimum side with value 0.
2. Inter quartile range is Q3-Q1 = 140-100 = 40
3. The distribution is spread such that it has max value around 200, and upper whisker at 200, min value 0 and lower whisker at 45. It is spread more on the right side than the left with a standard deviation = 31.973
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 117 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 199 and 0 which is exactly what we see in the plot.

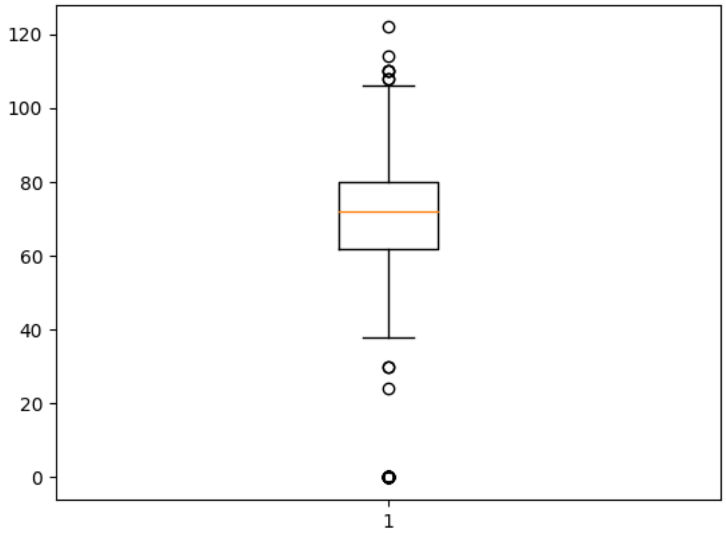

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are 4 outliers on the maximum side with values 122, 114, 110, 108 and 3 on the minimum side with values 30, 25 and 0.
2. Inter quartile range is Q3-Q1 = 80-60 = 20
3. The distribution is spread such that it has max value 122, and upper whisker at 106, min value 0 and lower whisker at 38. It is almost equally spread on both sides with a standard deviation = 19.356
4. The data is almost symmetric.
5. In question 1 we calculated median to be 72 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 122 and 0 which is exactly what we see in the plot.

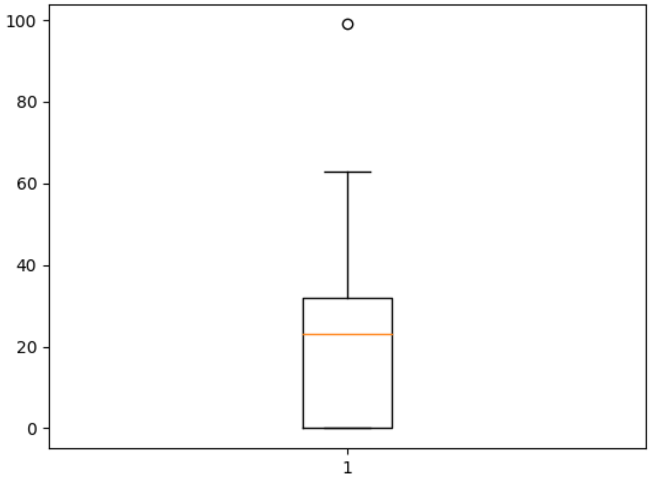

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. There is 1 outlier on the maximum side with value 99.
2. Inter quartile range is Q3-Q1 = 32-0 = 32.
3. The distribution is spread such that it has max value 99, and upper whisker at 63, min value 0 and lower whisker also at 0. It is spread more on the right side than the left with a standard deviation = 15.952
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 23 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 99 and 0 which is exactly what we see in the plot.

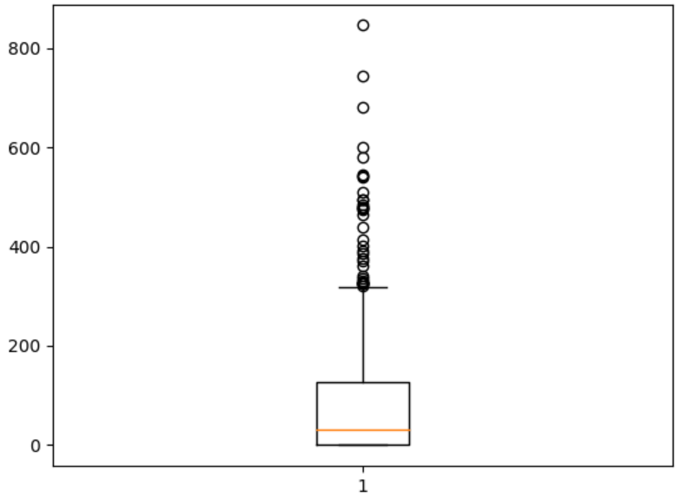

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. There are a large number of outliers on the maximum side with values ranging from 318 to 846.
2. Inter quartile range is Q3-Q1 = 130-0 = 130
3. The distribution is spread such that it has max value 846, and upper whisker at 317, min value 0 and lower whisker also at 0. It is spread more on the right side than the left with a standard deviation = 115.244.
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 30.5 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 846 and 0 which is exactly what we see in the plot.

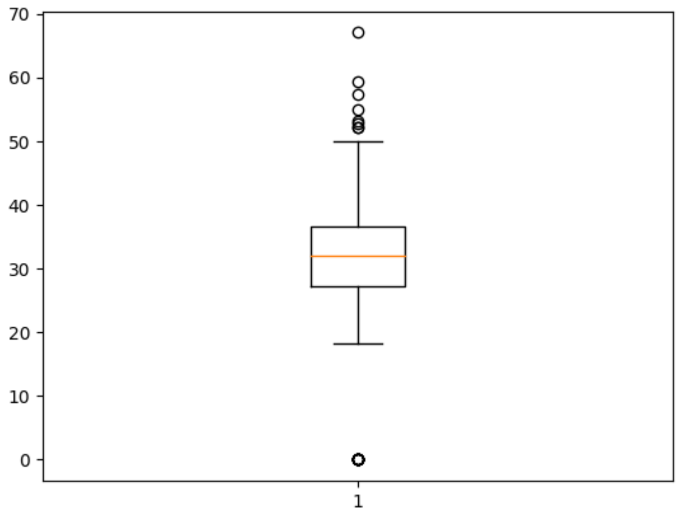

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. There are a large number of outliers on the maximum side with values ranging from 52 to 67 and one outlier on the minimum side with value 0.
2. Inter quartile range is Q3-Q1 = 36-27 = 9.
3. The distribution is spread such that it has max value 67, and upper whisker at 50, min value 0 and lower whisker at 18 It is spread more on the right side than the left with a standard deviation = 7.884.
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 32 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 67.1 and 0 which is exactly what we see in the plot.

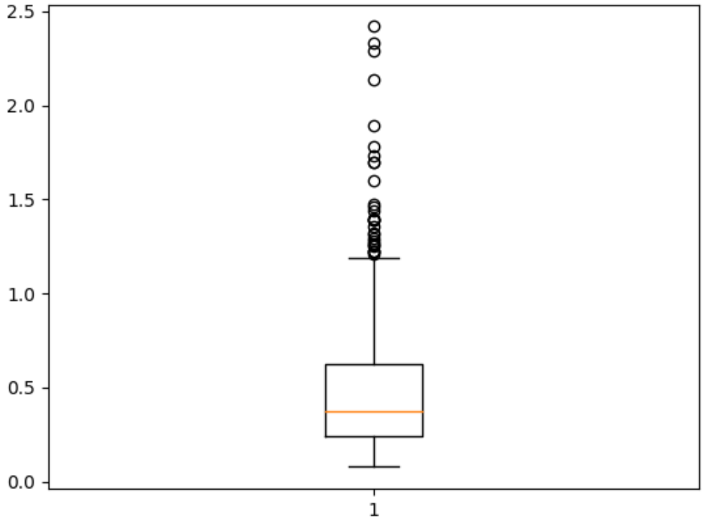

Figure 25 Boxplot for attribute pedi

Inferences:

1. There are a large number of outliers on the maximum side with values ranging from 1.23 to 2.42.
2. Inter quartile range is Q3-Q1 = 0.631-0.248 = 0.383.
3. The distribution is spread such that it has max value 2.42, and upper whisker at 1.2, min value 0.088 and lower whisker also at 0.088. It is spread more on the right side than the left with a standard deviation = 0.331
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 0.372 which is almost what we see in the boxplot. Also, we calculated the maximum and minimum to be 2.42 and 0.078 which is almost what we see in the plot.

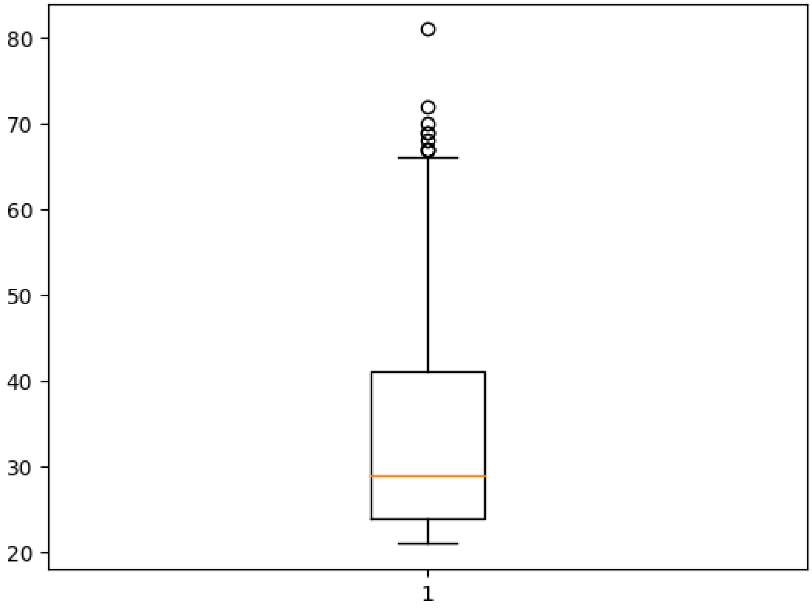

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. There are 6 outliers on the maximum side with values around 67, 68, 69, 72 and 81.
2. Inter quartile range is Q3-Q1 = 41-24 = 17.
3. The distribution is spread such that it has max value 81, and upper whisker at 66, min value 21 and lower whisker also at 21. It is spread more on the right side than the left with a standard deviation = 11.76.
4. The data is positively skewed, as it is more spread after the median than before.
5. In question 1 we calculated median to be 29 which is exactly what we see in the boxplot. Also, we calculated the maximum and minimum to be 81 and 21 respectively, which is exactly what we see in the plot.