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

Student's Name: Priyanka Kumari **Mobile No:** 8328354314

Roll Number: B20307 **Branch:** Electrical 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	3.84	3.0	1	0	17	3.367
2	plas	120	117.0	99	0	199	31.95
3	pres (in mm Hg)	69.105	72.0	70	0	122	19.34
4	skin (in mm)	20.53	23.0	0	0	99	15.94
5	test (in mu U/mL)	79.799	34.0	0	0	846	115.16
6	BMI (in kg/m^2)	31.99	32.0	32	0	67.1	7.87
7	pedi	0.47	0.3745	0.254	0.078	2.42	0.33
8	Age (in years)	33.24	29.0	22	21	81	11.75

Inferences:

1. By the above given values of mean median mode, we can say that column pres and column BMI have approximately same mean, median and mode values, therefore the spread of the data is most likely to be symmetrical.
2. The column pres and pedi have values of mean and median close to each other while the value of standard deviation is relatively low as compared to other attributes.
3. The column test has a large difference between mean and median, also there is large difference between maximum and minimum values, hence have a wide range. The standard deviation is very high.

2 a.

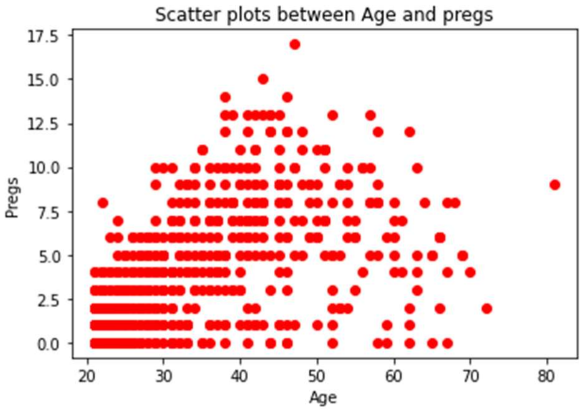

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. The above scatter plot tells us that the correlation between age and pregs is almost zero.
2. From the above graph we can conclude that as the age value increases the spread of the data values of pregs decreases.
3. The density of no. of women getting pregnant is more at the age in range 20 to 40 while the density decreases as the age value increases.
4. The highest value of pregs is 17 at the age in tge range approximately in the range 45 to 50.
5. This scatter plot tells us that women most likely get pregnant during the age of 20 to 40 while they don’t prefer to get pregnant during their late stage in the age 50 to 80 hence the density of number of women getting pregnant in the age 50 to 80 is less.

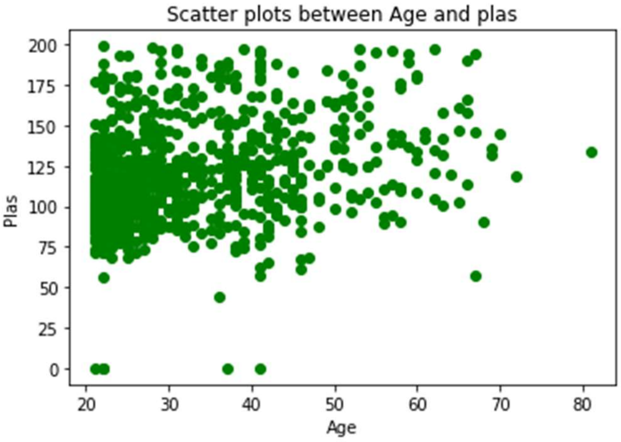

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. The above scatter plots tell us that the correlation between age and plas is almost zero.
2. From the above graph we can conclude that in the age between 20 to 40 the Plasma glucose concentration 2 hours in an oral glucose tolerance test have highest spread of data and while the values decrease as the age increases.
3. The density of plasma glucose concentration is highest in the range of 20 to 40 while its density decreases as the value of age increase.
4. The density of plas is inversely related with the age value.
5. We can see from the above graph that the concentration of plasma glucose is highly dense in the range of 20 to 30 age year while its slowly becomes less dense with the increasing age and is least dense at the age of 80.
6. We can also conclude that women in their late 80s have less plasma glucose concentration than the women in the age of 20 to 30s.

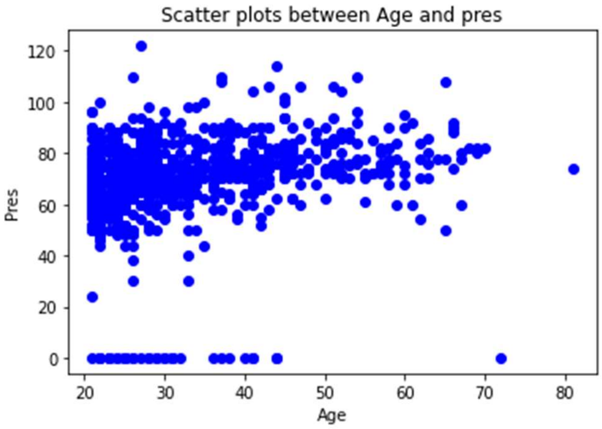

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. The correlation is almost zero between age and pres.
2. The spread of data is more in the range of 20 to 30 while it slowly decreases as the value of age increases.
3. The density of pres is high in the range of 20 to 40 years of age while it becomes less dense in the range of 50 to 60 and least in the range of 70 to 80 years of age.
4. From above graph we also conclude that the diastolic blood pressure in women is more in the age of 20 to 40 while it decreases in the age to 60 to 80.

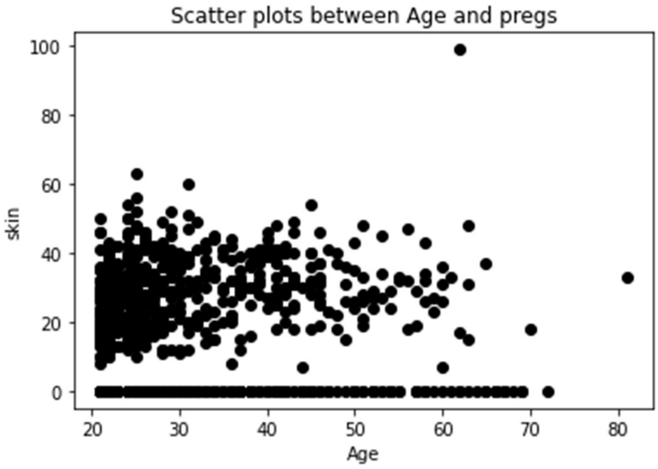

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. The correlation is negative and close to zero between age and skin.
2. From the above graph we can infere that the spread of data is more in the range of 20 to 30 years of age and it decreases as the value of age increases.
3. Triceps skin fold thickness (mm) is highly dense in the range of 20 to 30 years of age while its density decreases as the value of age increases and is least dense in the region of 70 to 80 years of age.
4. The maximum triceps skin fold thickness is 99mm at the age closure to 60 to 65 years range.
5. From above graph we can conclude that the skin fold thickness is more during the early age of 20 to 40 years while it decreases in the range to 50 to 80.

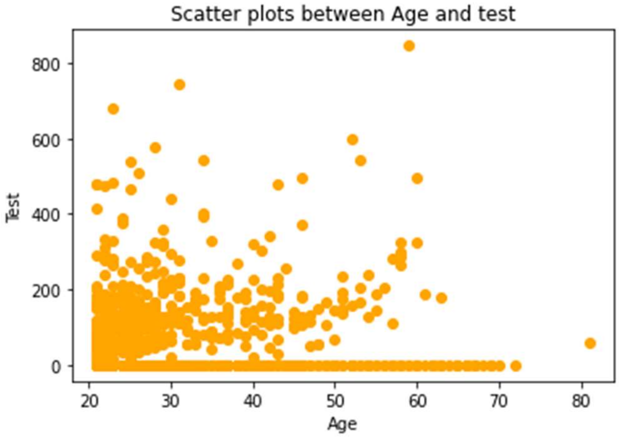

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. The correlation between age and test is moderate and negative.
2. The spread of data is more in the range of 20 to 30 years of age while it decreases with the increasing value of age.
3. The density of 2-Hour serum insulin (mu U/mL) test is more in the region of 20 to 30 years of age while its less dense in the range of 40 to 70 years and is least at 80 years of age.
4. The probability of number of women taking test in at the age of 20 to 30 years of age is more than the probability of women taking test in the range of 40 to 80 years of age.

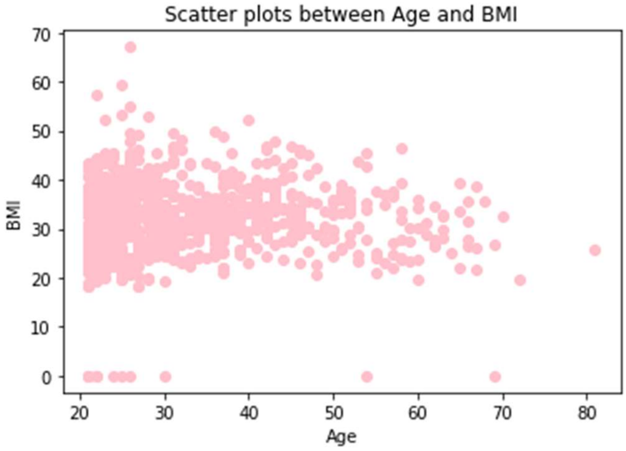

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. The correlation between age and BMI is almost zero.
2. From the above graph we can conclude that the spread of data is more in the range of 20 to 40 while it decreases in the range of 50 to 80.
3. The density of BMI is more in the range of 20 to 30 years of age and it becomes less dense as the age value increases and the density is least in the range of 70 to 80 years of age.
4. The maximum value of BMI is around 68 at the age in the range of 20 to 30 years.
5. From the above graph we can conclude that women in the age of range 20 to 40 have more BMI than compared to the women in the range of 50 to 80 years of age.

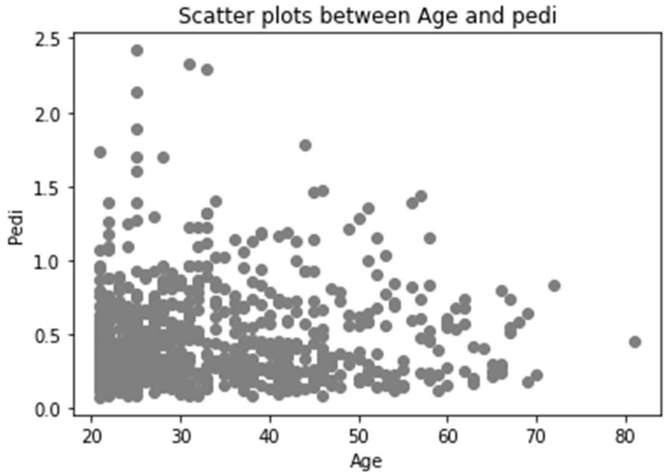

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. The correlation between age and pedi is almost zero.
2. The spread of data is more in the range 20 to 50 years of age and it decreases as the age increases.
3. The density of diabetes pedigree function is more at the age of 20 to 40 years of range and pedi range from 0 to 1 while its density decreases in the range 50 to 80 years of age and 1 to 2.5 range of pedi.
4. The probability of diabetes pedigree function is maximum in the range of 20 to 30 years.
5. From the above graph we can conclude that most women have diabetes at the age of 20 to 40 years and least at the age of 60 to 80 years.

b.

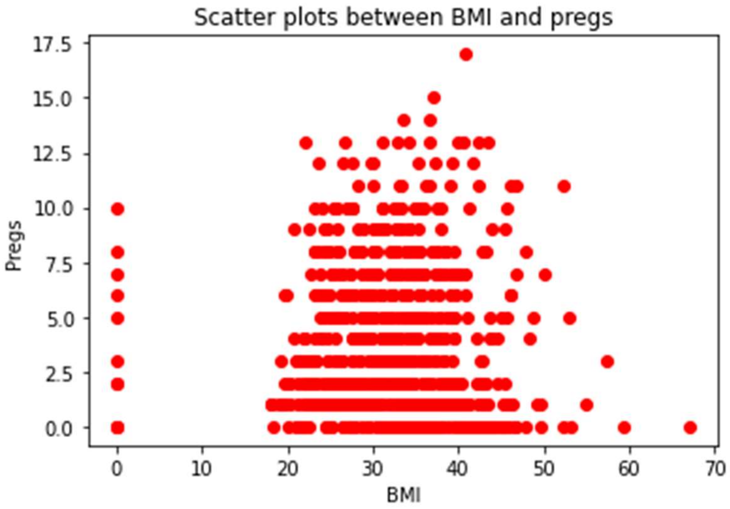

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. The correlation between BMI and pregs is close to zero.
2. The spread of data is more in the range of 20 to 50 BMI while almost zero in the range of 0 to 20 and less in the range of 50 to 70 kg/m^2 BMI.
3. The density of pregs is more in the range of 20 to 50 while it decreases in the range of 50 to 70 range.
4. From the above graph we can conclude that the probability of women being pregnant is more if they have BMI in the range of 20 to 50 kg/m^2.

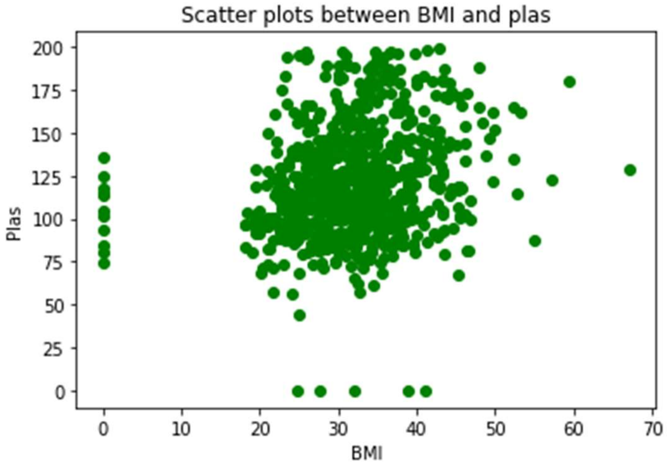

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. The correlation between BMI and plas is weak and positive.
2. The spread of data is more in the range of 30 to 40 kg/m^2 while it is almost zero in the range of 0 to 20 and the spread of data is less in the range of 50 to 70kg/m^2 of BMI.
3. The plasma concentration is highly dense in the region of 20 to 40kg/m^2 BMI and its density decreases in the range of 50 to 70 kg/m^2 BMI.
4. The scatter plot tells us that the women having more plasma level are overweight and obsess i.e., have more BMI value.

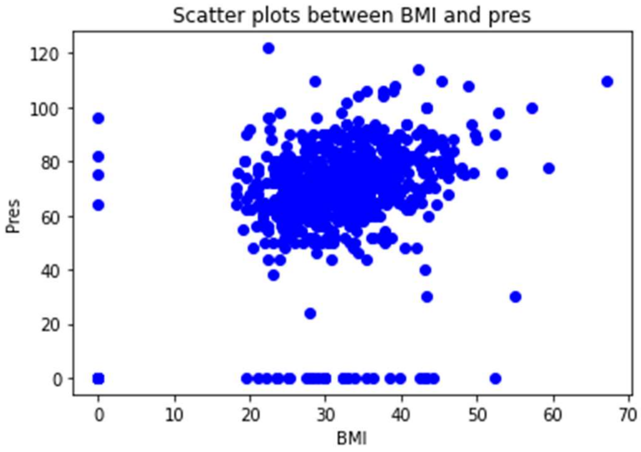

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. The correlation between BMI and pres is positive and weak.
2. The scatter plot has high density in the range of 20 to 50 kg/m^2 BMI and it decreases in the range of 50 to 70 kg/m^2 and is almost zero in the range of 0 to 20 kg/m^2 BMI.
3. From the above graph we can conclude that the women have high probability of getting diabetes if their BMI is in the range of 20 to 50 kg/m^2.

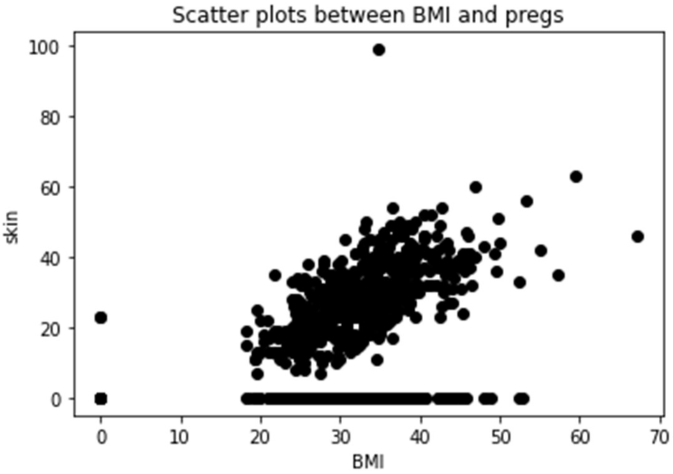

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. The correlation between BMI and skin is positive and moderate.
2. The density of the scatter plot is more in the range of 20 to 30 kg/m^2 BMI and it decreases from 50 t0 70 kg/m^2 BMI.
3. From the above scatter plot, we can conclude that the probability of triceps skin fold thickness (mm) is more if the BMI lies in the range of 20 to 50 kg/m^2.

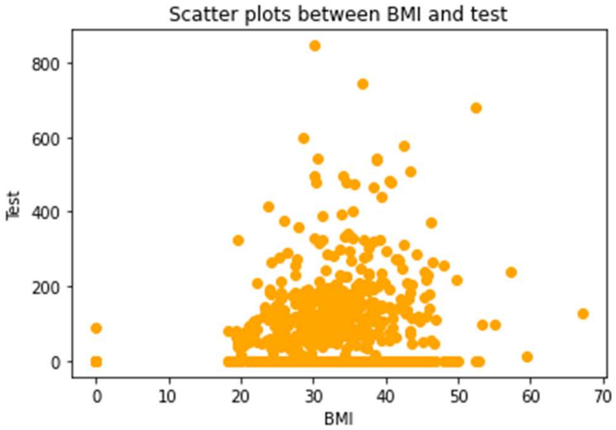

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. The correlation between BMI and test attribute is approximately equal to zero.
2. From the above scatter plot, we can infere that the density is high in the range of 20 to 50 kg/m^2 BMI and it decreases from 50 to 70 kg/m^2, the density is almost zero in the range of 0 to 20kg/m^2 BMI.
3. From the above graph we can conclude that the probability of test (2-Hour serum insulin (mu U/mL) is high with the BMI range in 20 to 50 kg/m^2.

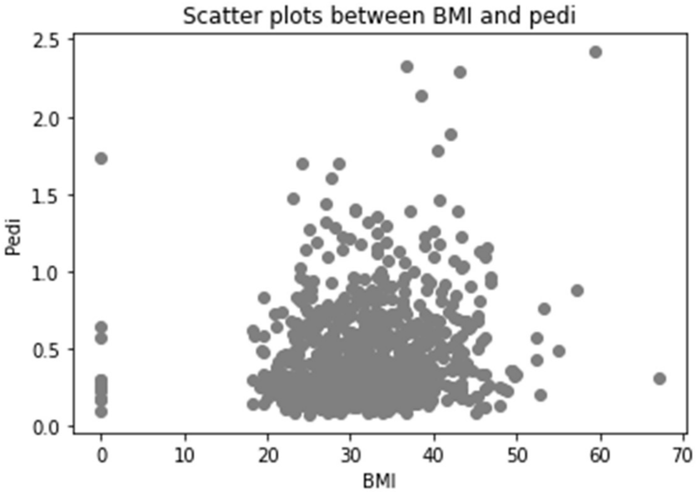

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. The correlation between BMI and pedi attribute is approximately equal to zero.
2. From the above scatter plot, we can infere that the density is high in the range of 20 to 50 kg/m^2 BMI and it decreases from 50 to 70 kg/m^2, the density is almost zero in the range of 0 to 20kg/m^2 BMI.
3. From the above plot we can conclude that the probability of having diabetes is more when the BMI lies in the range of 20 to 40 kg/m^2.

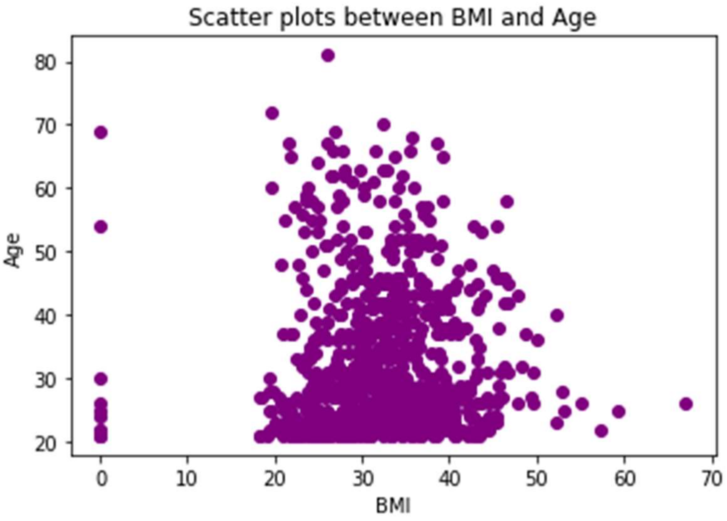

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. The correlation between BMI and age attribute is approximately equal to zero.
2. From the above scatter plot, we can infere that the density is high in the range of 20 to 50 kg/m^2 BMI and it decreases from 50 to 70 kg/m^2, the density is almost zero in the range of 0 to 20kg/m^2 BMI.
3. From the above plot we can conclude that women having high BMI value in the range 20 to 50 are most likely be old in age.

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)	-0.0421
6	BMI (in kg/m^2)	0.03624
7	pedi	0.033

Inferences:

1. From the above magnitude we can conclude that the attribute pregs is strongly correlated with the attribute age. While attribute pedi is weakly correlated with the attribute age.
2. From the above table we can infere that the attribute skin and test is negatively correlated therefore it will decrease with increase in age, while other attribute are positively correlated therefore it will increase with increase in age and the correlation between BMI and pedi attribute is almost zero therefore it will remain constant with increase in age.
3. From the above scatter plot and the table, we can infere that the attributes with strong correlation like pregs have scatter plot less spread. While the attribute having weaker correlation have more spread of data and the attribute with correlation approximately equal to zero have a linear spread of data.
4. From the above table we can conclude that the almost all the attributes are weakly correlated with the attribute age.

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	pedi	0.140
7	Age (in years)	0.036

Inferences:

1. The attribute plas, pres, and skin are moderately correlated while the attribute test and pedi are weakly correlated with BMI and the attribute pregs and age have almost no correlation with BMI.
2. From the above table we can conclude that all the attributes are positively correlated with the attribute BMI and its value will increase with increase in BMI value while the attribute pregs and age remails almost const with increase in BMI value as their magnitude is almost zero.
3. The attribute pres and skin have moderately spread of data while the other attribute has more spread of data as the correlation is weak for other attributes.
4. From the above table we can conclude that almost all the attributes have a positive and weak correlation with the attribute BMI.

4 a.

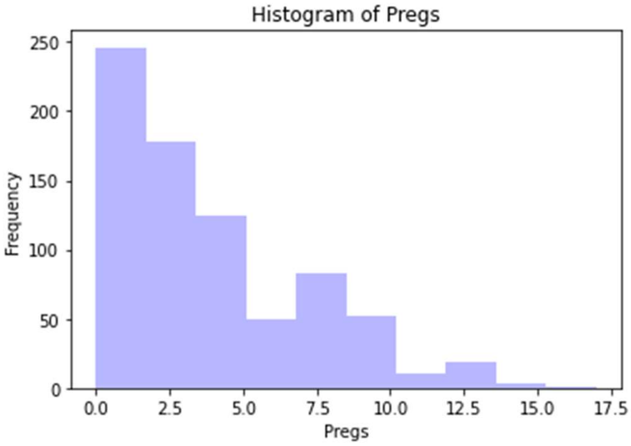

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. The probability of women being pregnant 0 to 2 times is higher as the frequency of that bin is higher.
2. The mode of pregs attribute lies in the bin of 0 to 0.25 as the frequency is high for the bin 0 to 0.25.
3. From the above graph we can conclude that the data is positively skewed or right skewed.

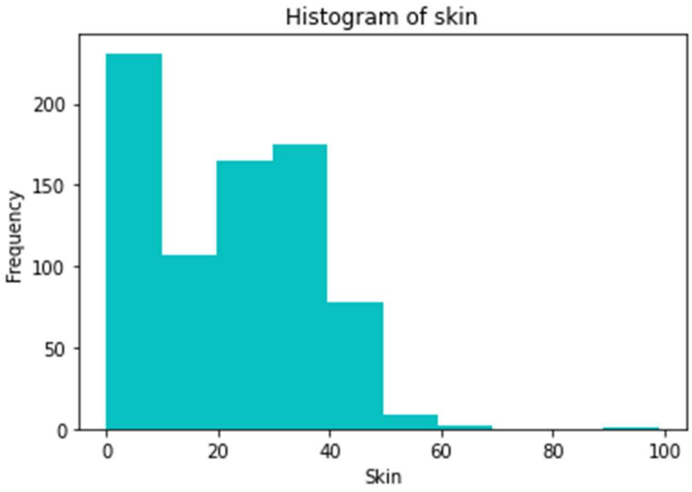

Figure 16 Histogram depiction of attribute skin

Inferences:

1. Most women have skin thickness in the range of 0 to 20 mm as the frequency of that bin is highest
2. The mode of the attribute skin lies in the bins 0 to 20 as the frequency of that bin is high.
3. Data is positively skewed or right skewed.

5.

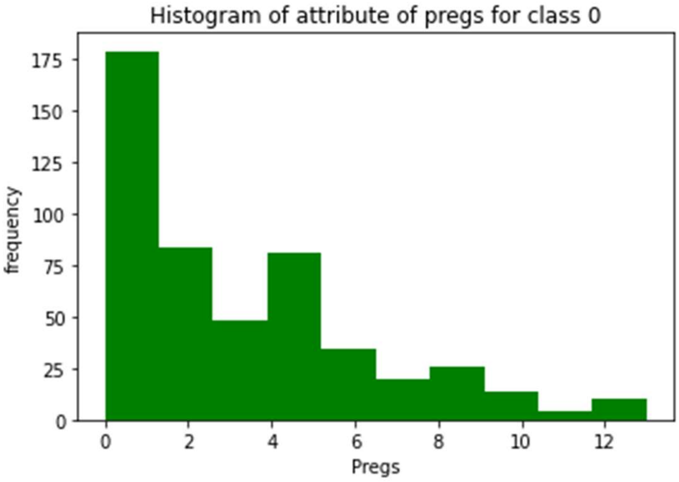

Figure 17 Histogram depiction of attribute pregs for class 0

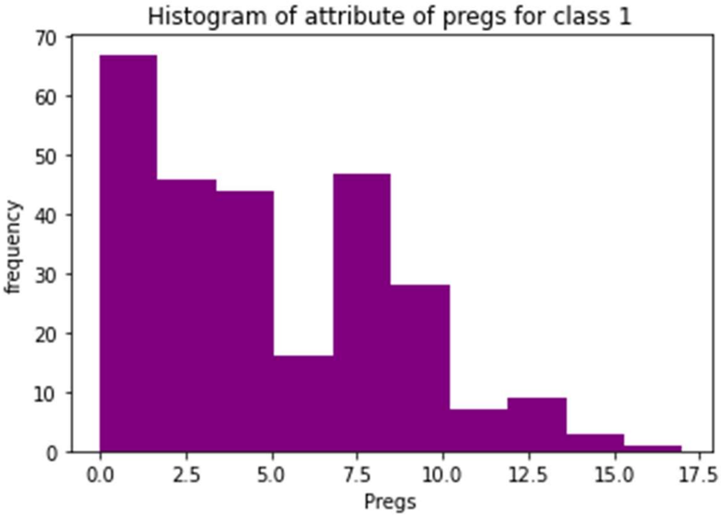

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. From the above histogram for the class 0 frequency of the bin 0 to 2 is highest hence mode lie in 0 to 2 bins while the frequency of 0 to 0.25 bins is highest for class 1 hence the mode lies in 0 to 2.5 bin.
2. From the above histogram we can observe that the frequency of the pregs for class 0 decreases significantly while for class 1 it decreases for 0 to 5 and then increases from 7.5 to 10 with frequency value of 50 and then again decreases until 17.5.
3. The frequency of class 0 with the bin 0 to 2 is greater than the frequency of class 1 with bin value 0 to 2.5 with a frequency of 68.

6.

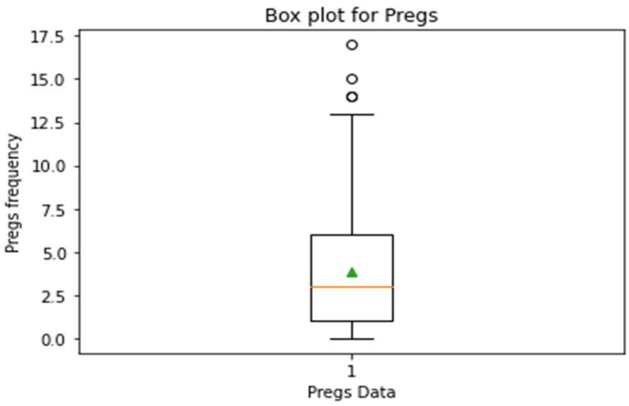

Figure 19 Boxplot for attribute pregs

Inferences:

1. From the above plots we can see that there are three outliers with the values approximately equal to 13.6, 14.9 and 17.4 which is larger than the upper quartile range + (1.5*IQR) = 6+(1.5*6) = 13.5
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 6 – 1 = 5(approx.)
3. The attribute preg has a positive spread of 5 and varies from 0 to 13 (approx.) with some values(outliers) greater than 13.5.
4. From the above boxplot we can infer that the data is positively skewed or right skewed.
5. The median value is approximately equal to 3 and from Q1 we have median value calculated as 3.0, the maximum and minimum values are 0 and 17 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 17.
6. The green triangle in the plot represents the mean value which indicates the mean value is greater than the median value.

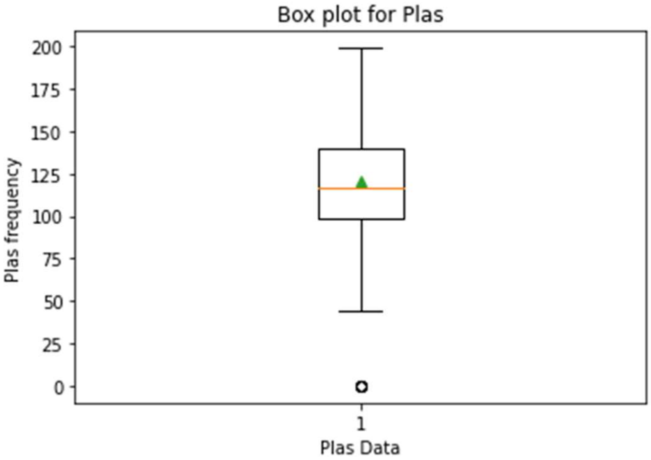

Figure 20 Boxplot for attribute plas

Inferences:

1. From the above plots we can see that there is one outlier with the values approximately equal to 0 which is less than the lower quartile range –(1.5*IQR) = 120 – (1.5*90)= 15
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 140 – 50 = 90(approx.)
3. The attribute plas has a positive spread of 90 and varies from 50 to 200 (approx.) with one value(outliers) less than 15.
4. From the above given plots, we can infer that the plot approximately symmetrical in nature.
5. The median value is approximately equal to 118 and from Q1 we have median value calculated as 117.0, the maximum and minimum values are 0 and 119 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 200.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately equal to the median value.

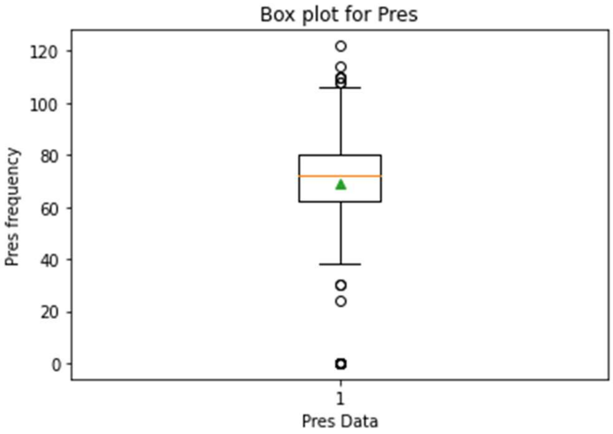

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. From the above plots we can see that there are 7 outliers with the values four of which are larger than the upper quartile range + (1.5*IQR) = 80+(1.5*15) = 102.5 and 3 of which are less than the lower quartile range – (1.5*15) = 60-(15*1.5) = 37.5
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 80 – 65 = 15(approx.)
3. The attribute pres has a positive spread of 15 and varies from 40 to 80 (approx.) with some values(outliers) greater than 102.5(approx.) and some values(outliers) less than 37.5
4. From the above plot we can infer that the data is negatively skewed or left skewed.
5. The median value is approximately equal to 70 and from Q1 we have median value calculated as 69.105, the maximum and minimum values are 0 and 122 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 120.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately less than the median value.

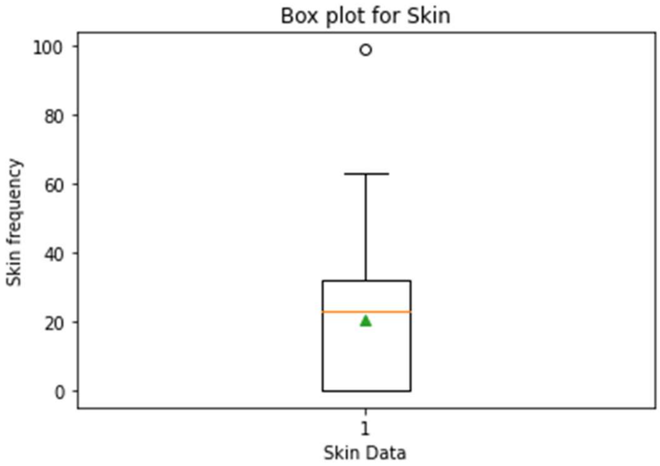

Figure 22 Boxplot for attribute skin (in mm)

Inferences:

1. From the above plots we can see that there is only one outlier with the value approximately equal to 100 which is larger than the upper quartile range + (1.5*IQR) = 30+(1.5*30) = 75.
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 30 – 0 = 30(approx.)
3. The attribute skin has a positive spread of 30 and varies from 0 to 30 (approx.) with one value(outlier) greater than 75.
4. From the above plot we can infer that the data is negatively skewed or left skewed.
5. The median value is approximately equal to 20 and from Q1 we have median value calculated as 23.0, the maximum and minimum values are 0 and 99 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 100.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately less than the median value.

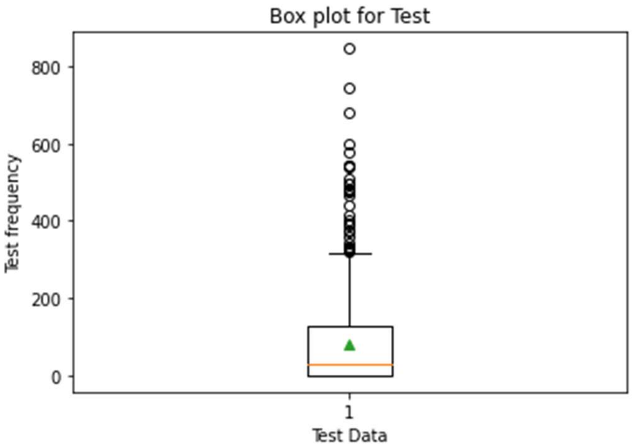

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. From the above plots we can see that there are many outliers with the value approximately equal to larger than the upper quartile range + (1.5*IQR) = 150+(1.5*150) = 375.
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 150 – 0 = 150(approx.)
3. The attribute test has a positive spread of 150 and varies from 0 to 150 (approx.) with many values(outlier) greater than 375.
4. From the above plot we can infer that the data is positively skewed or left skewed.
5. The median value is approximately equal to 30 and from Q1 we have median value calculated as 34.0, the maximum and minimum values are 0 and 846 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 850.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately greater than the median value.

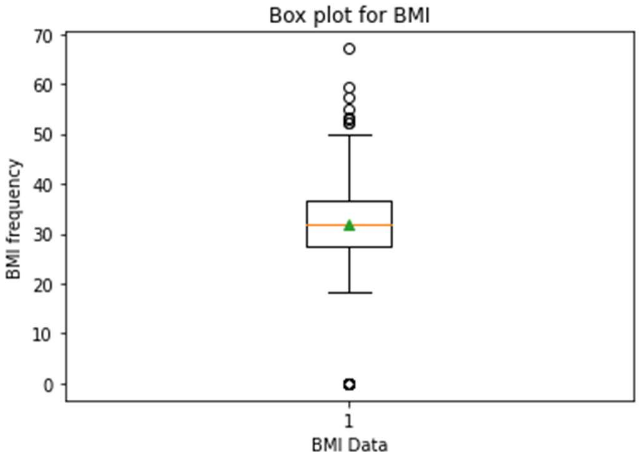

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. From the above plots we can see that there are many outliers with the value approximately equal to larger than the upper quartile range + (1.5*IQR) = 40+(1.5*10) = 55.
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 40 – 30 = 10(approx.)
3. The attribute test has a positive spread of 149 and varies from 1 to 150 (approx.) with many values(outlier) greater than 373.5.
4. From the above plot we can infer that the data is positively skewed or left skewed.
5. The median value is approximately equal to 33 and from Q1 we have median value calculated as 32.0, the maximum and minimum values are 0 and 846 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0 and 70.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately equal the median value.

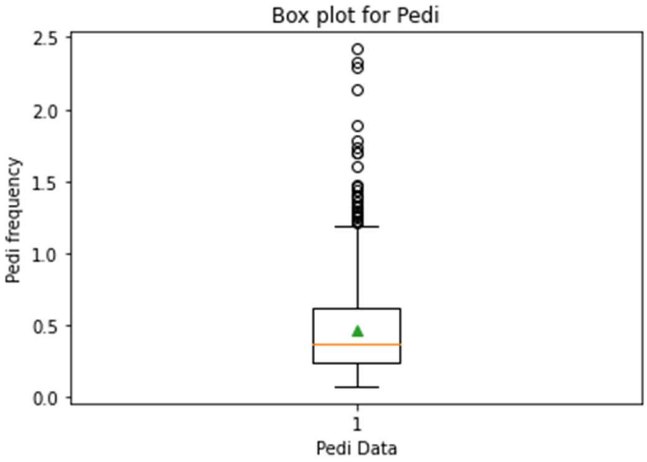

Figure 25 Boxplot for attribute pedi

Inferences:

1. From the above plots we can see that there are many outliers with the value approximately equal to larger than the upper quartile range + (1.5*IQR) = 0.5+(1.5*0.4) = 1.1.
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 0.5 – 0.1 = 0.4(approx.)
3. The attribute test has a positive spread of 0.4 and varies from 0.1 to 0.5 (approx.) with many values(outlier) greater than 1.1.
4. From the above plot we can infer that the data is positively skewed or left skewed.
5. The median value is approximately equal to 0.40 and from Q1 we have median value calculated as 0.374, the maximum and minimum values are 0.078 and 2.42 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 0.06 and 2.40.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately greater than the median value.

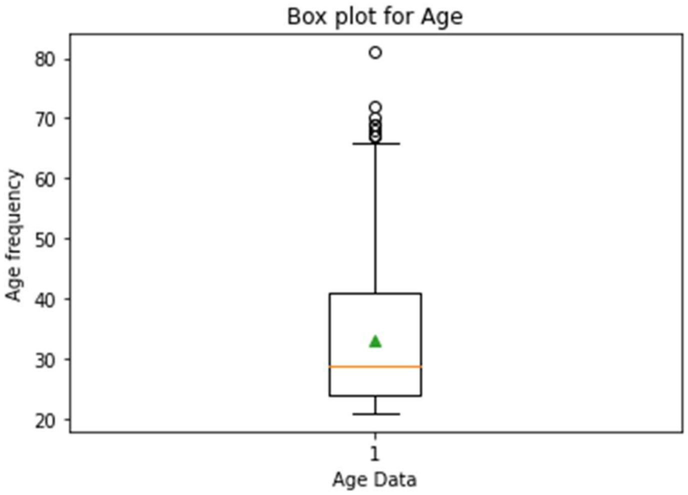

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. From the above plots we can see that there are many outliers with the value approximately equal to larger than the upper quartile range + (1.5*IQR) = 40+(1.5*15) = 62.5
2. The Inter quartile Range is Q3(upper quartile range)-Q1(lower quartile range) = 40 – 25 = 15(approx.)
3. The attribute test has a positive spread of 15 and varies from 25 to 40 (approx.) with many values(outlier) greater than 62.5.
4. From the above plot we can infer that the data is positively skewed or left skewed.
5. The median value is approximately equal to 30 and from Q1 we have median value calculated as 29, the maximum and minimum values are 21 and 81 respectively as calculated in Q1 and from the graph we can see that the max and min value is approximately equal to 20 and 80.
6. The green triangle in the plot represents the mean value which indicates the mean value is approximately greater than the median value.