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

Student's Name: Vikas Dangi **Mobile No: 9406661661**

Roll Number: B20238 **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	3.845	3.000	1	0	17.000	3.369
2	plas	120.894	117.000	99 &100	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	0.372	0.254 &0.258	0.078	2.420	0.331
8	Age (in years)	33.240	29	22	21.000	81.000	11.760

Inferences:

1. As the standard deviation in the Diabetes pedigree function is very near to zero, we also observe that Mean, Median and Mode does not hold much difference between them. They are near to each other which shows that the data is less spread and most of the data lies in a particular range only.
2. The standard deviation of number of pregnant women is also less i.e. 3.369 which gets synchronize with the mean and median values being very close to each other.
3. The most spread attribute is serum insulin concentration and the least spread is Diabetes pedigree function.

2 a.

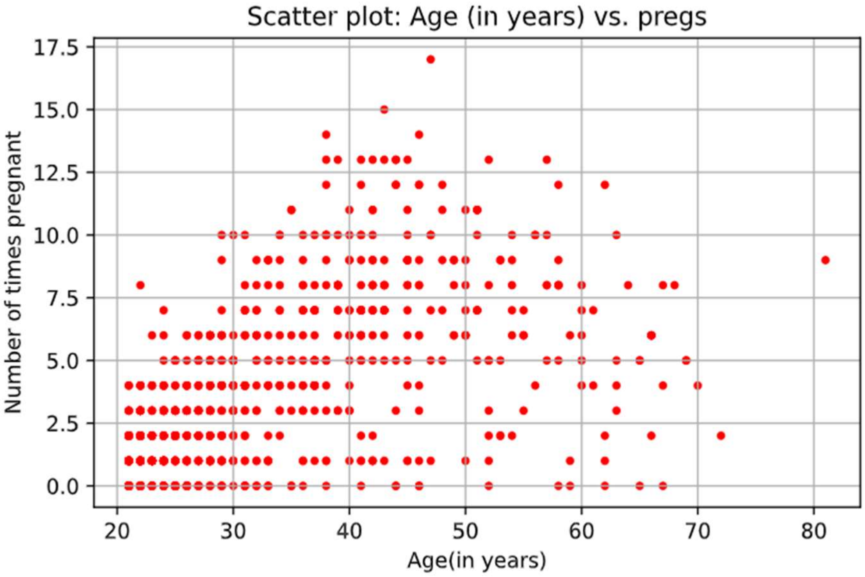

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. Both the age and the number of times pregnancy are strong +vely correlated as we can see that pregs variable increases with the increase in age at least till 50 years of age
2. Most of the dots lies towards the young age and ladies having 0,1 or two children which signifies more population towards young age as a result of population growth of the country
3. It is evident from the plot that for having greater number of children one has to be more in age as it takes at least 1.5(generally 2) years for a new baby to born.

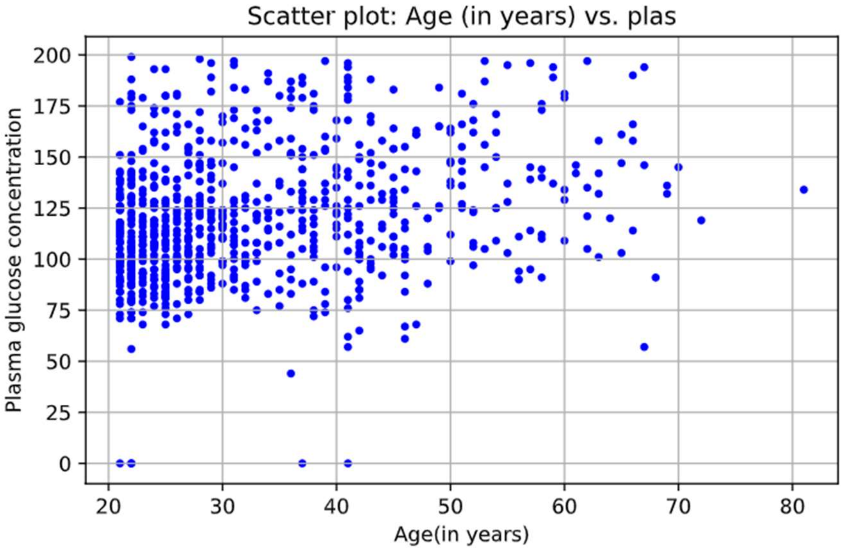

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. The plasma glucose conc. is moderately and +vely correlated with the age as we can see some of increment in the Y axis values (mainly in lower scattered region). This thing gets more insight by the fact that correlation value is 0.26.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Majority of Plasma Glucose conc. has the value in between 75 and 140 and there are certain higher values more dominant in the older people.

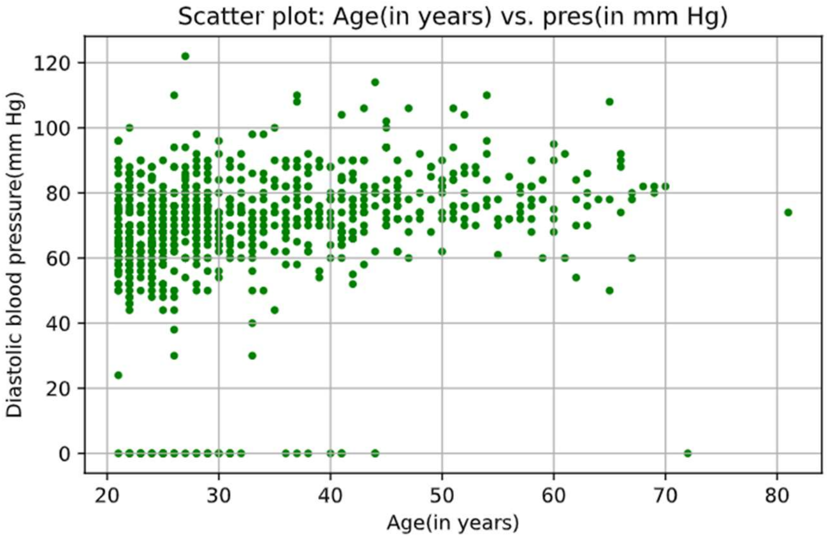

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. It is moderately +ve correlated as there is a slight increase in the values of pressure with age we can observe this in the lower scattered area.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Density of the dots is quite high in the region from 50mm Hg to 90 mm Hg.

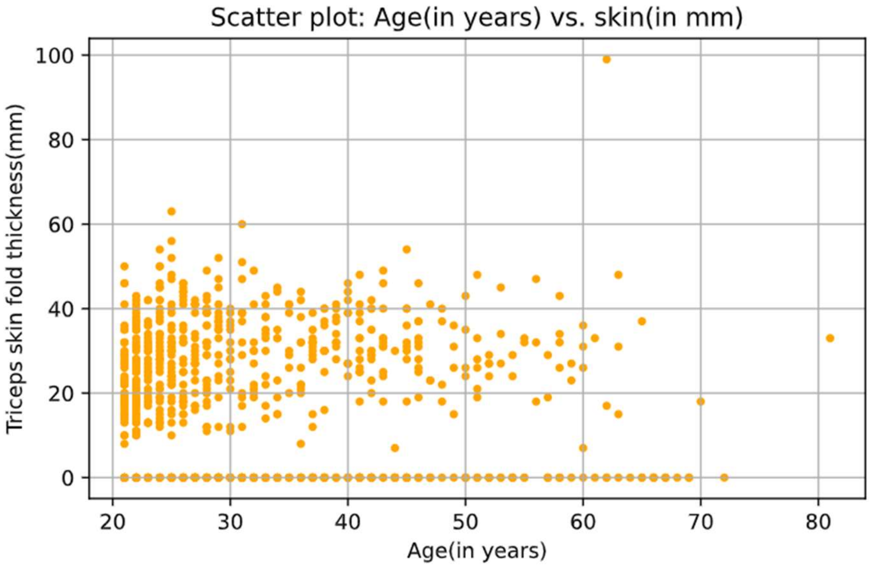

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. The Skin is weakly and slightly negatively correlated as the change of Triceps skin fold with the X-axis is negligible only a slight decrease is observed in the upper outliers.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Be it young or old each one of them have their values lying in certain fix range which shows this attribute does not depend much upon age.

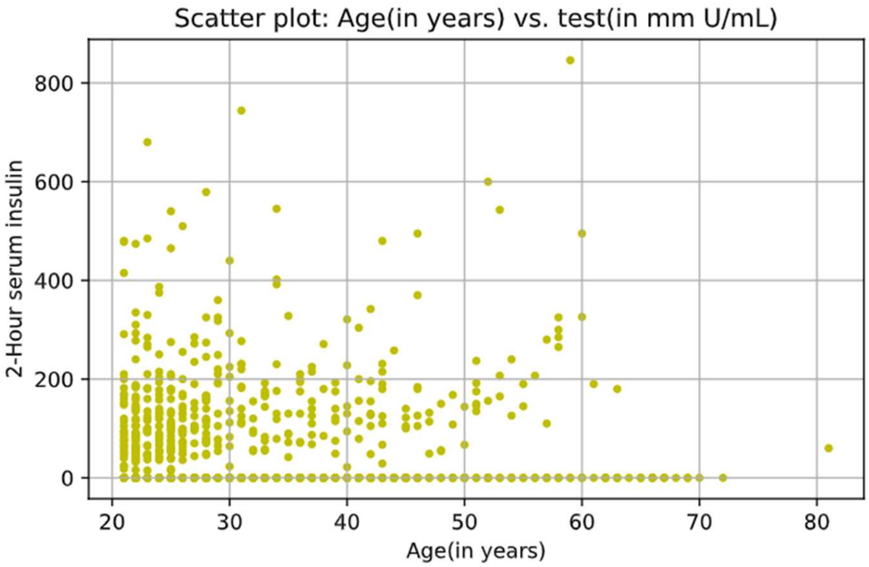

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. It is little or not correlated as there is not some fix increment or decrement in the test with the age.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. There are more outliers in this plot which means we can expect even high values of the Insulin conc. and the outliers increases with the age.

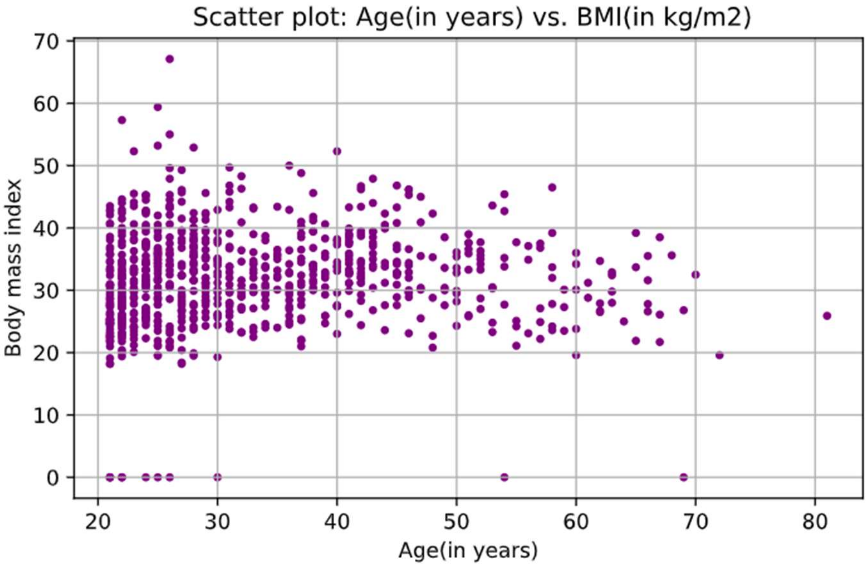

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. It is little or no correlated as there is not some fix increment or decrement in the BMI with the age.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Be it young or old each one of them have their values lying in certain fix range which shows BMI index does not depend much upon age.

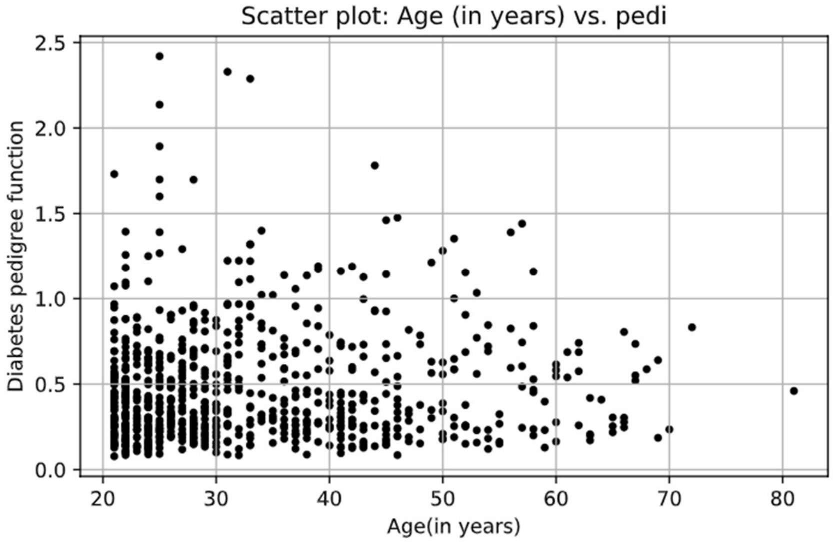

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. It is little or no correlated as there is not some fix increment or decrement in the Pedigree function with the age.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Some of the young people seems to reach high Diabetes pedigree function value which is not seen in the older people. Also we are having more outliers in the higher values than in the lower values which conveys the function's value to deviate to the higher values rather than some lower value.

b.

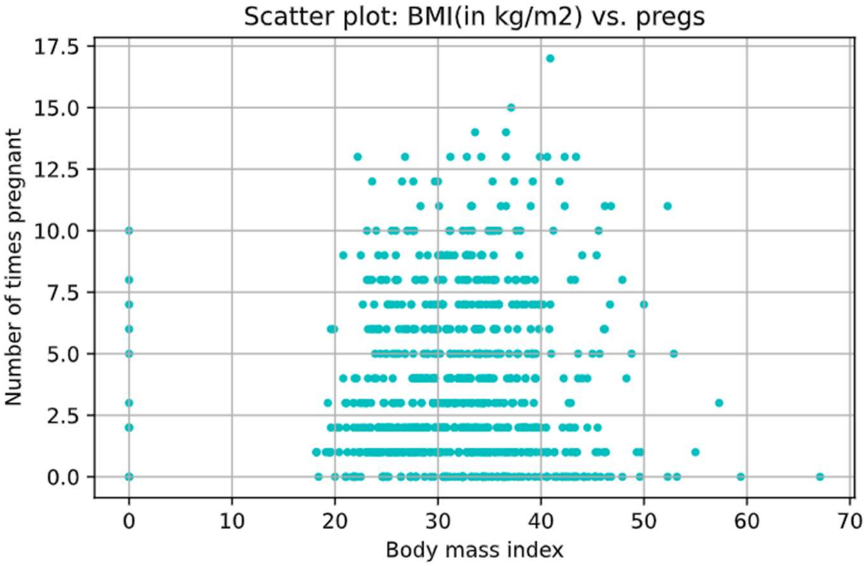

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. It is little or no correlated as there is not some fix increment or decrement in the number of times pregnant with the BMI index.
2. The data seems to be distributed over the y axis which means the number of times women upto getting pregnant is spread over a good range. Yet it is dense when it comes to the number 0,1,2.

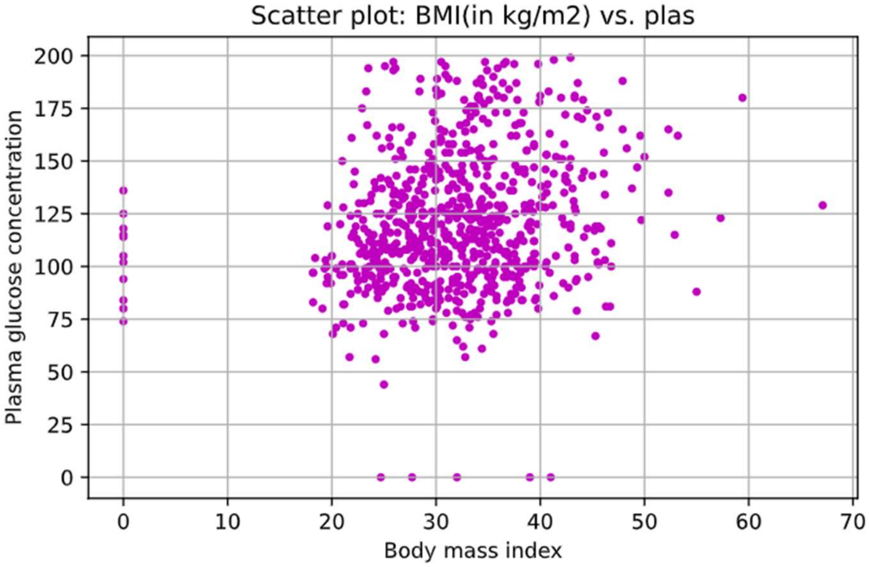

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. The plasma glucose conc. is moderately and +vely correlated with the BMI index as we can see some of increment in the Y axis values with BMI. This thing gets more insight by the fact that correlation value of 0.221. And also the graph seems to be tilted slightly in a way in which it looks like it has got a small positive slop.
2. The plot has outliers in all the region, the major dots lie in the 20-40 range which shows us that majority of the people have their BMI in this range.

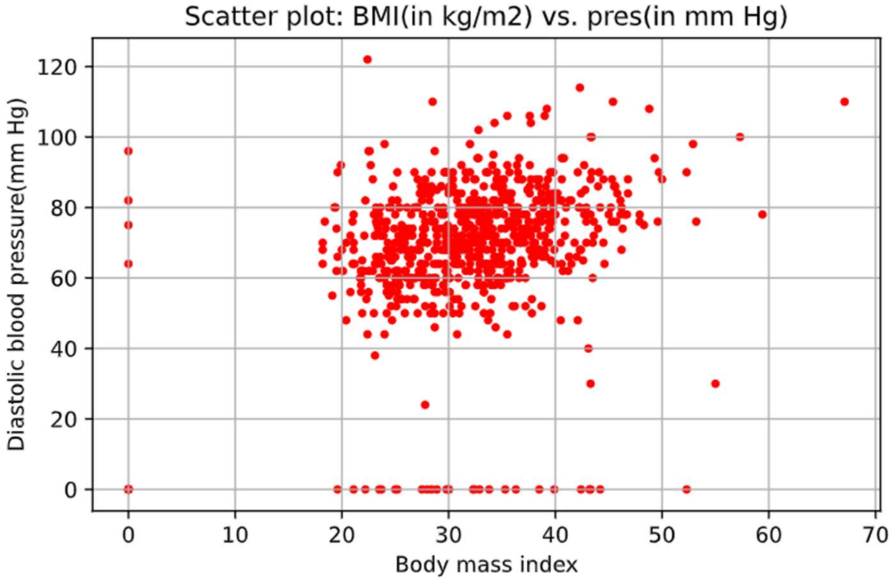

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. Diastolic Blood pressure is moderately and +vely correlated with the age as we can see some of increment in the Y axis values with BMI. And also, the graph seems to be tilted slightly in a way in which it looks like it has got a fairly small positive slop (greater than the last one).
2. The data seems to get distributed over the range 50-90 mm Hg and as the population has BMI index lying in 20-40 kg/m^2 the data is not much spread.

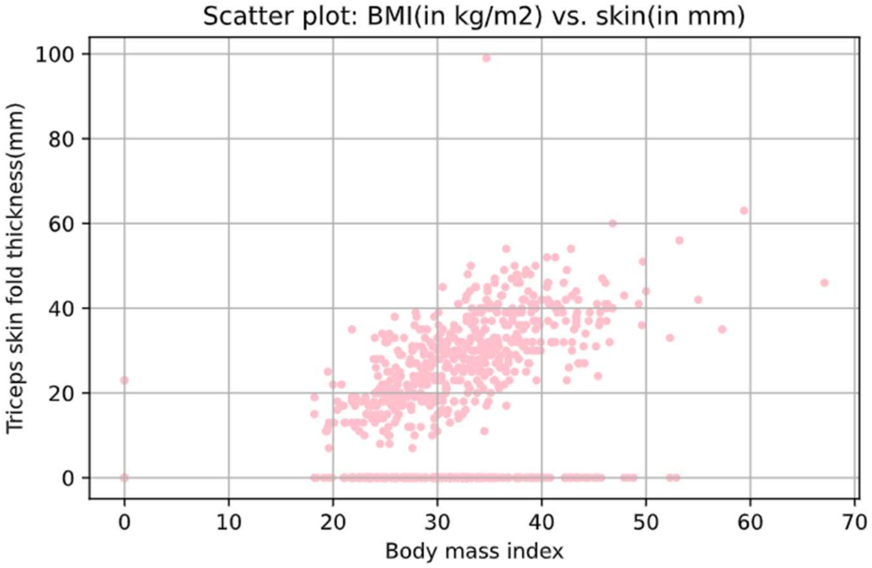

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. The skin parameter is strongly correlated with the BMI as the more obesity invites more skin thickness and it is quite clear from the graph too which shows us the graph positively tilted towards right side giving us a fairly positive strong correlation.
2. The graph doesn't have extreme outliers and the density lies in a well-defined range.

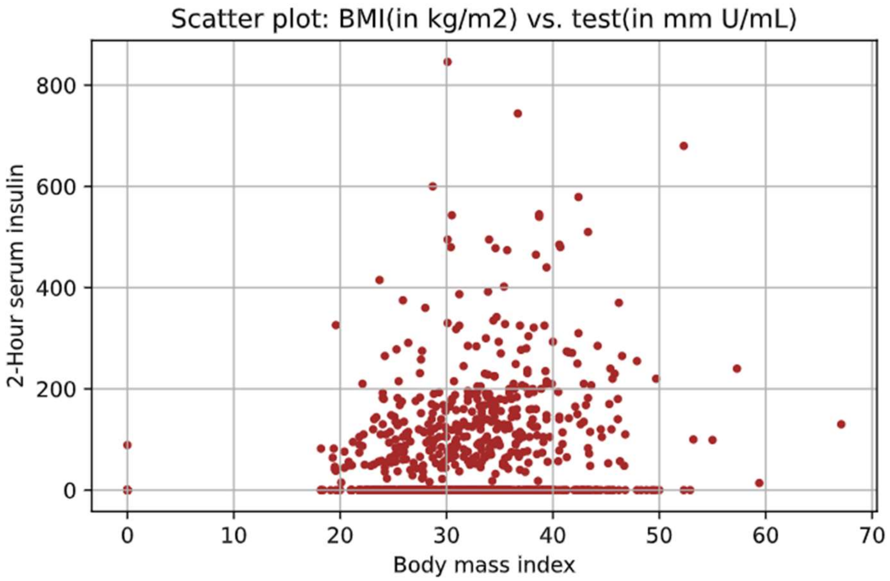

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. It is weak and positively correlated and it can be seen intense in the upper scattered data and slow in the lower region.
2. Major density of points lies in lower range under 200 km/m^2 but the outliers are shooting up real high values.

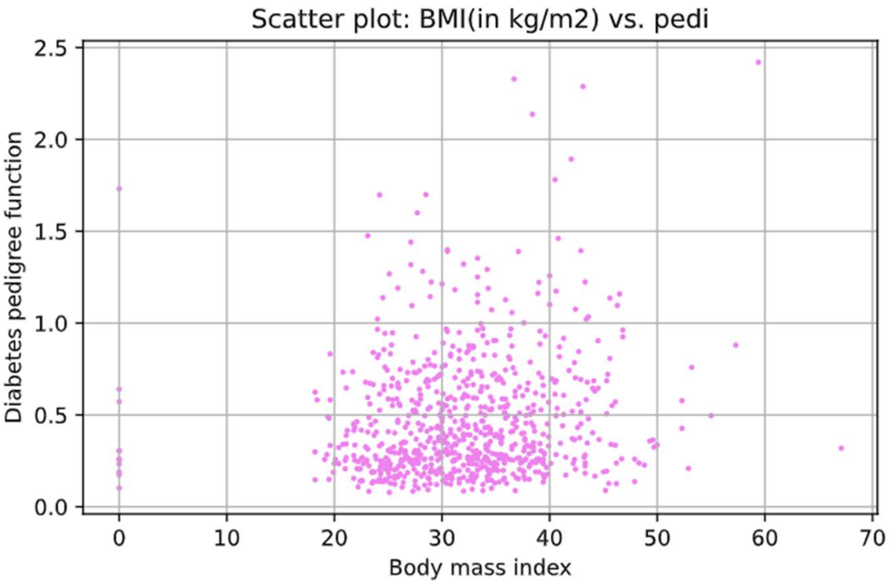

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. The Diabetes Pedigree function seems to be weakly but positively correlated, the diabetes function values above 0.5 seems to get increase with BMI.
2. Majority of the values lies below 0.5 and are condensed but the values above it are not so condensed at a particular region and tends to achieve higher values.

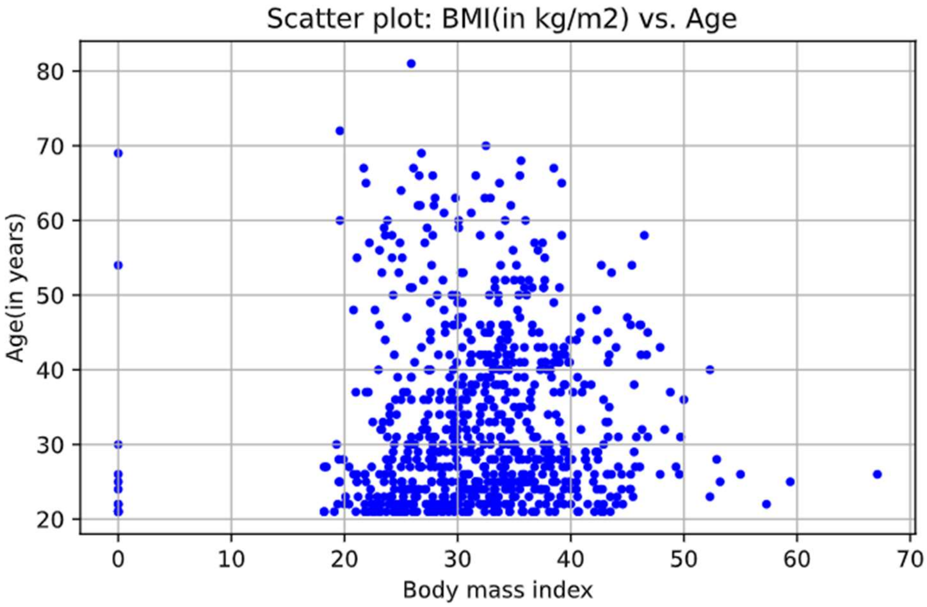

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. It is little or no correlated as there is not some fix increment or decrement in the BMI with the age.
2. We have a high density of dots in the 20 to 30 years of age region which shows us that major of our population lies in that region.
3. Be it young or old each one of them have their values lying in certain fix range which shows BMI index does not depend much upon 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	0.264
3	pres (in mm Hg)	0.239
4	skin (in mm)	-0.114
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036
7	pedi	0.033
8	Age (in years)	1

Inferences:

1) The age is strongly correlated with pregs and moderately correlated with plas and pres. It is less moderately correlated with skin and almost not correlated with test, BMI and pedi. Also, it is totally correlated with itself.

2) Except for skin and test all are positively correlated which means other than these the values of attributes increase to some extend with the increase in age.

3) It was hard to see the correlation of the ones in which it was very less correlated and it took some time for it other than it results seem to match with the inferences we get from the plots.

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)	0.282
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) The BMI index is strongly correlated with skin and moderately correlated with plas, pres and test. It is less moderately with pedi and almost not correlated with pregs and Age. Also, it is totally correlated with itself.

2) All are positively correlated which means the values of these attributes increase to some or more extend with the increase in age. It was hard to see the correlation of the ones in which it was very less correlated and it took some time for it other than this, results seem to match with the inferences we get from the plots

4 a.

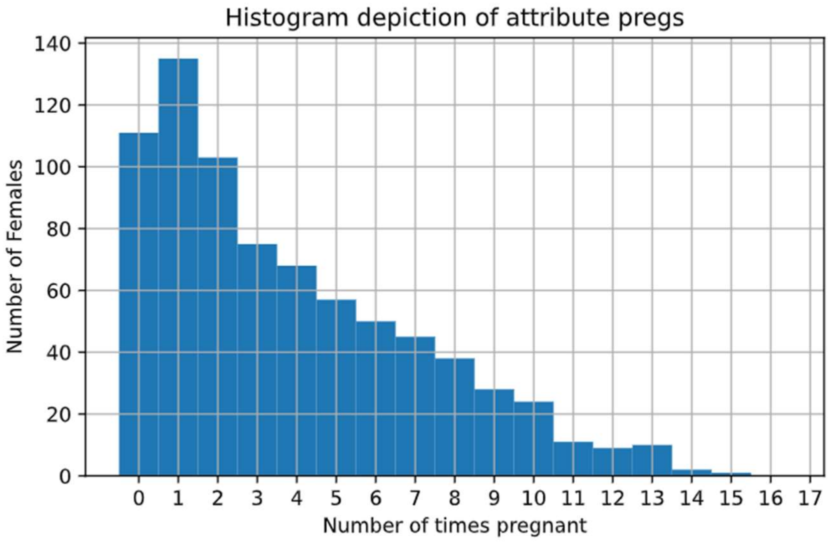

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. The frequency of each bin is getting decreased with the number of times being pregnant due to being having more population in the young age which haven’t got pregnant for a lot of times due to their less age and maybe due to more literacy rate in new people.
2. The mode lies in the second bin having value of 1.
3. The graph seems to get decrease exponentially after 1 which seems quite justifiable as per the human birth capabilities.

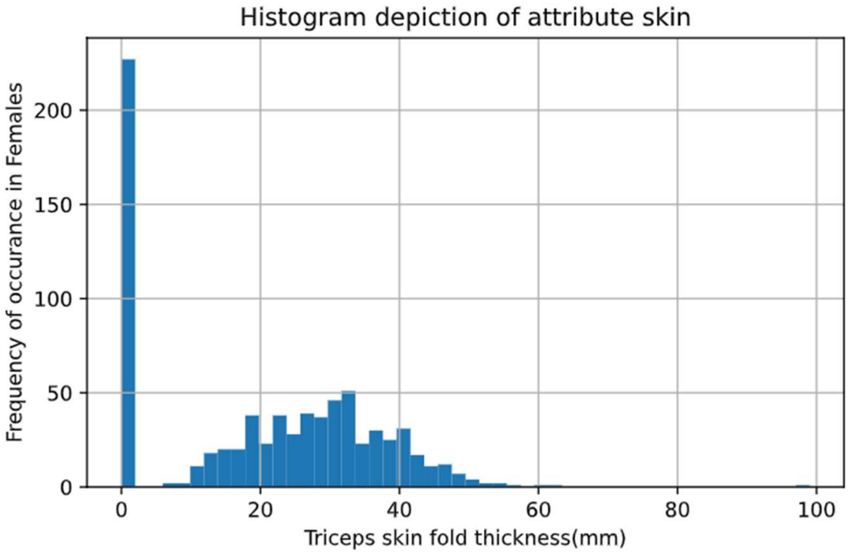

Figure 16 Histogram depiction of attribute skin

Inferences:

1. Except for the value of 1st bin rest of them seems to follow normal distribution having the peak in the middle.
2. Bin 1 with value 0 contains the mode of the attribute skin.
3. Due to the fact that most of the people are having 0 thickness and the people having some thickness its median get shifted more towards left side despite of having normal distribution after the first bin.

5

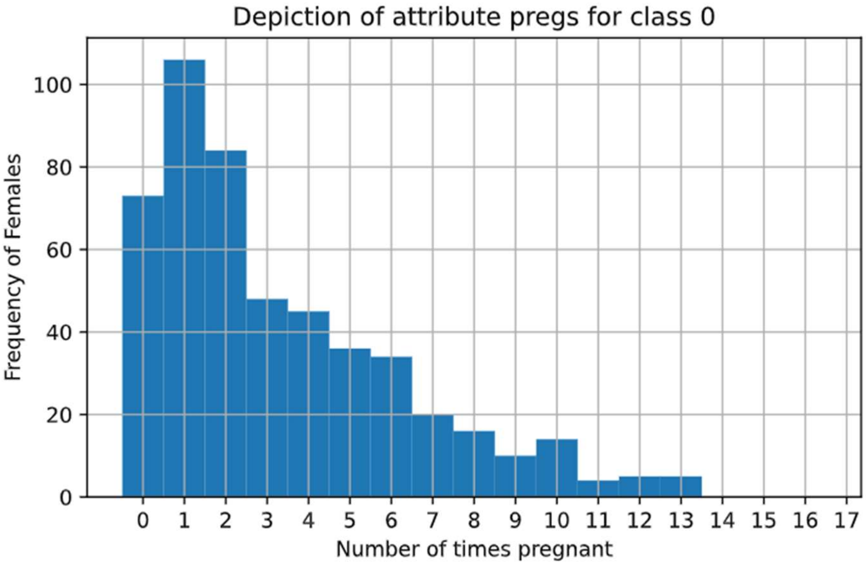

Figure 17 Histogram depiction of attribute pregs for class 0

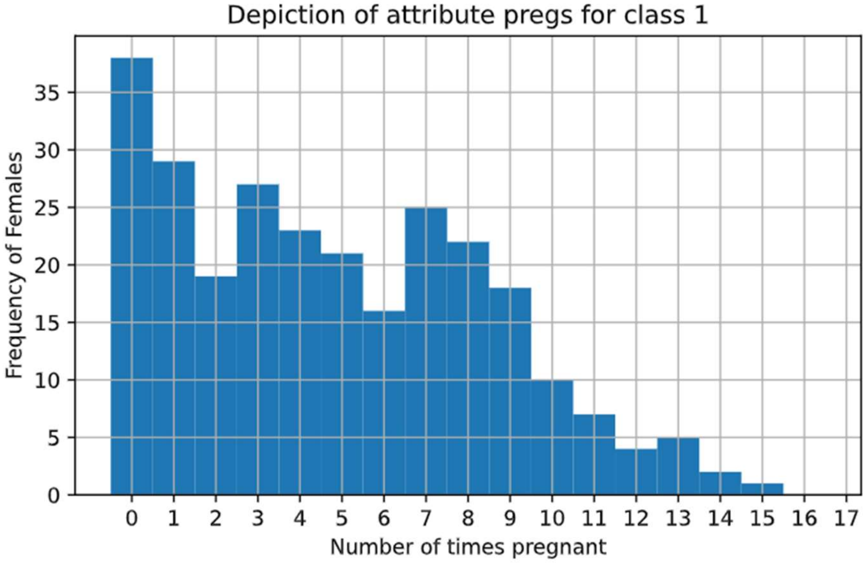

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. For class 0 the mode lies in the bin having value 1 and for class 1 the mode lies in the bin having value 0.
2. The class 0 has its peak at 1 and then the frequency of substituent values decreases but in class 1 the peak is at 0 and then overall we have a decrement later on but is not regular and we are getting ups in the frequency in between
3. Rest it is clearly noticeable that both of them holds strong relation with the plot without classes separately the class 0’s mode becomes dominant and adds 1 as the mode overall.

6

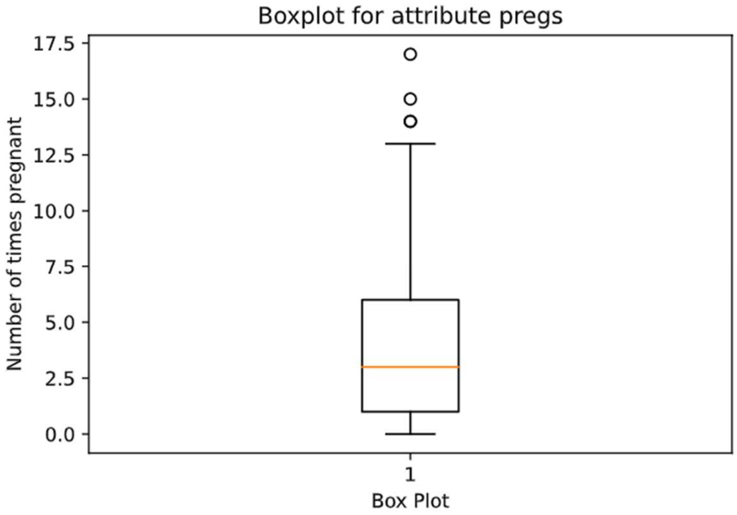

Figure 19 Boxplot for attribute pregs

Inferences:

1. Outliers have higher values like 14,15 and 17. This conveys us that number of times pregnancy can't be such high in normal cases.
2. The Inter quartile range here is 5 with the first quartile on 1 and third on 6. 50 % data lies in the range of 1-5.
3. The interquartile range is 5 which is comparable than the range of the data i.e., 17 which conveys that the data has moderate to low variability.
4. The data is skewed right. We can see that the median does not divide the box into two equal half and the data above the median seems to be more spread.
5. The median is plotted at 3 which synchronize with the value in question number 1 also we have our mode as 1 so that the region containing that value here holds more condensed data.

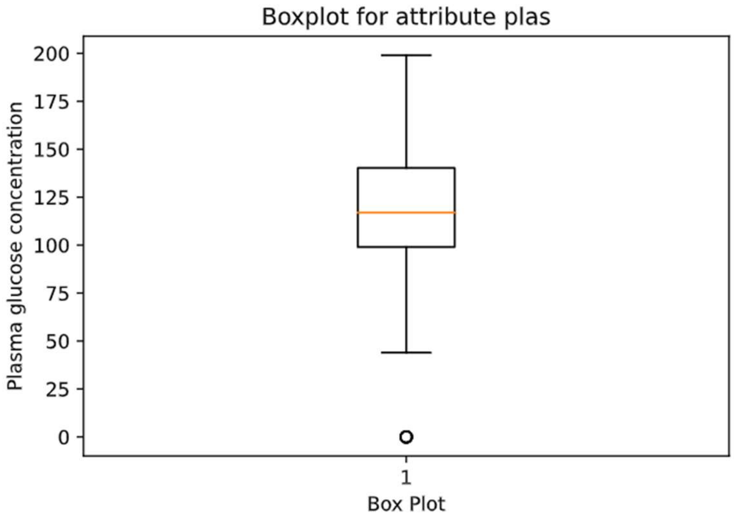

Figure 20 Boxplot for attribute plas

Inferences:

1. There are no such outliers in the plot except one at 0.
2. The interquartile range is around 40 conc. unit in this case and also 50 % data lies in the interquartile range 99-140 conc unit.
3. Almost all the data except the outlier lies in the range of 45-200 concentration unit which shows it has fair variability.
4. The data is less or not skewed as the orange line (representing median) is dividing the box into almost two equal and symmetric halves.
5. The median was 117 which is getting conveyed by the box plot and it seems to lie around the middle of the range of spread data which conveys that the data has very less skewness.

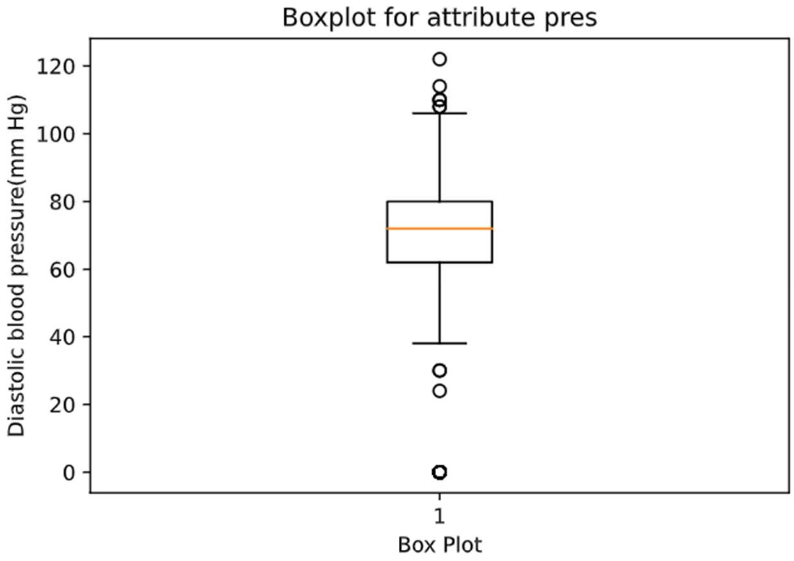

Figure 21 Boxplot for attribute pres (in mm Hg)

Inferences:

1. We have outliers lying on both the extremes 114,122,108 and 110 at the upper side and 0,24 and 30 in the lower outliers.
2. The Inter quartile range is 18 which is not large as compared to the range in the data is taking up the values.
3. The data seems to take values on a fairly large range i.e., 122 although it has a low inter quartile range which gives us information that the data is condensed in this region apart from having values in the larger region.
4. The middle orange line which represents the median divide the box in the almost two equal halves although it is not exactly half, we can say that it is very less skewed or very not skewed. It is skewed very less towards left.
5. The median was 72 which is getting conveyed by the box plot and it seems to lie around the middle of the range of spread data which conveys that the data has very less skewness.

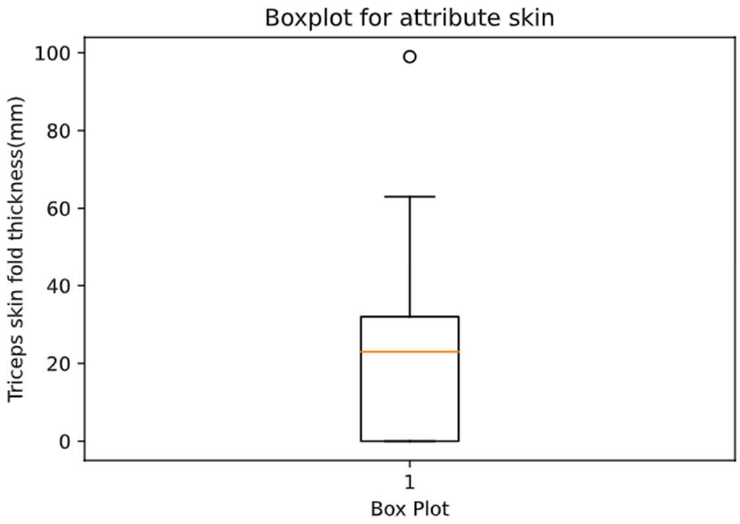

Figure 22 Boxplot for attribute skin (in mm)

Inferences:

1. The plot contains only one outlier having value 99 mm and it lies very far from the rest of the values.
2. The inter quartile range of the data set is 32mm.
3. The variability of the plot is on a moderate level it has a std dev value of 15.95mm the spread would have been more if a large number of data would not have fall on 0 value which makes the data condense at near to it.
4. The data is highly skewed as the orange median line does not divide the box into two equal parts.
5. The median was 23 which is getting conveyed by the box plot and it is clear that it deviates fairly large from the middle of the range of spread data which conveys that the data has very high skewness.

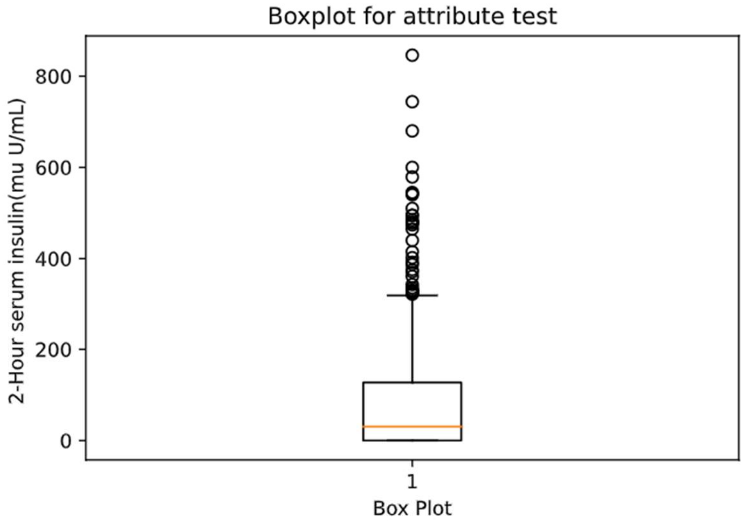

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. The graph has large number of outliers ranging from 320 to 846 it conveys that it is likely that we can get the highly deviated values in the upper range while dealing with this parameter.
2. The data has interquartile range of 127.25 mu U/mL. And it means that the 50% data in the middle lies in this region.
3. The data is spread largely and we get to know this due to its occupancy in the plot and the standard deviation seems to confirm this, it has a standard deviation of 115.244 mu U/mL which is quite large.
4. The data is positively skewed in the right. It is highly skewed due to middle line deviating a lot from dividing it from half.
5. The median was 30.500 mu U/mL which is getting conveyed by the box plot.

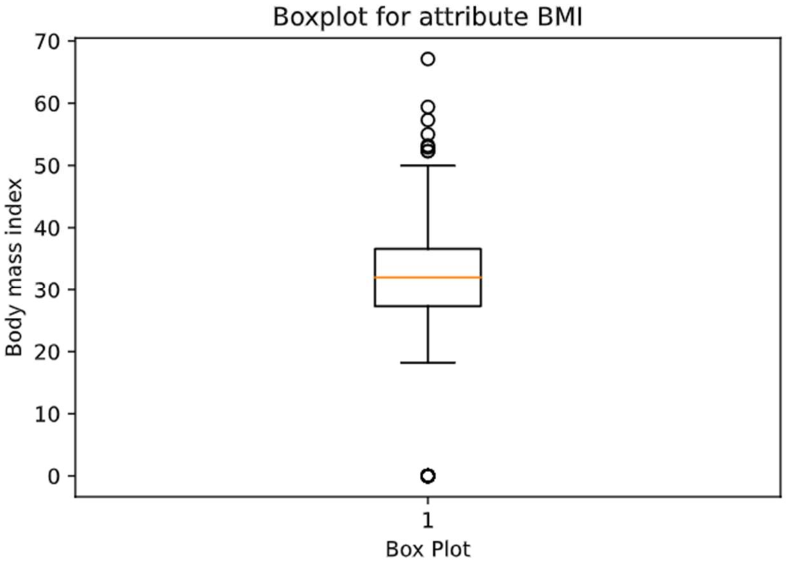

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. The plot has some outliers (57 bmi to 67 bmi) on the top and only one outlier at the bottom(0).
2. The Inter quartile range of the data is 9.3 BMI and our 50% data is spread only in this range.
3. The data is less spread compared to others it has only 9.3 BMI of IQR and has a std dev of 7.884 BMI.
4. The data doesn't seem to be skewed as the median line divides the box into two equal halves.
5. The median was 32 BMI which is getting conveyed by the box plot.

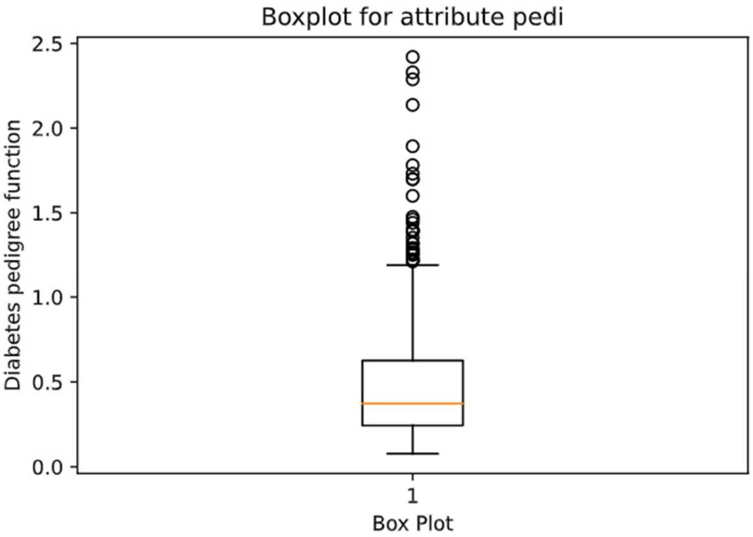

Figure 25 Boxplot for attribute pedi

Inferences:

1. The graph has large number of outliers ranging from 1.2 to 2.4 it conveys that it is likely that we can get the highly deviated values in the upper range while dealing with this parameter.
2. The inter quartile range is 0.382
3. The value of the pedigree function is limited to a small range in which it is spread well. It has a very low std deviation of 0.331.
4. The data is positive skewed as the median line does not divide the box in the two equal half.
5. The median was 0.372 which is getting conveyed by the box plot

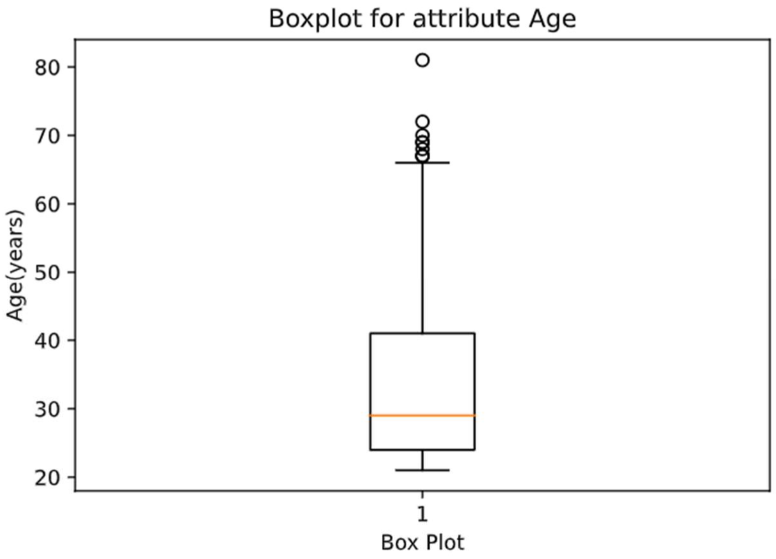

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. The plot has six outliers and their values lies in the range of 67 and 81.
2. The Inter quartile range is 12, 50% women lies in the range between 24 and 41 years of age.
3. The data is spread from the people with higher age for lower age they are condensed below the median. Overall, the spread is less which is also seen by the low std dev of the data.
4. There is a fair skewness in the data, the median does not divide the box in equal halves. It has positive skewness.

Thank You