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

Student's Name: Mayank Bansal **Mobile No: +919636993445**

Roll Number: B20156 **Branch:CSE**

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	1	0	17	3.367
2	plas	120.895	117	99, 100	0	199	31.952
3	pres (in mm Hg)	69.105	72	70	0	122	19.343
4	skin (in mm)	20.536	72	0	0	99	15.942
5	test (in mu U/mL)	70.799	30.5	0	0	846	115.169
6	BMI (in kg/m^2)	31.993	32	32	0	67.1	7.879
7	pedi	0.472	0.372	.254, .258	0.078	2.42	0.331
8	Age (in years)	33.241	29	22	21	81	11.753

Inferences:

1. BMI as seen have almost same value of mean, median and mode which indicates the graph would be approximately symmetric.
2. Plas and pedi as seen have 2 values of mode.
3. The standard deviation of pedi is very small as compared to others which indicates that its spread is very small and its peak attained should be very narrow.

2 a.

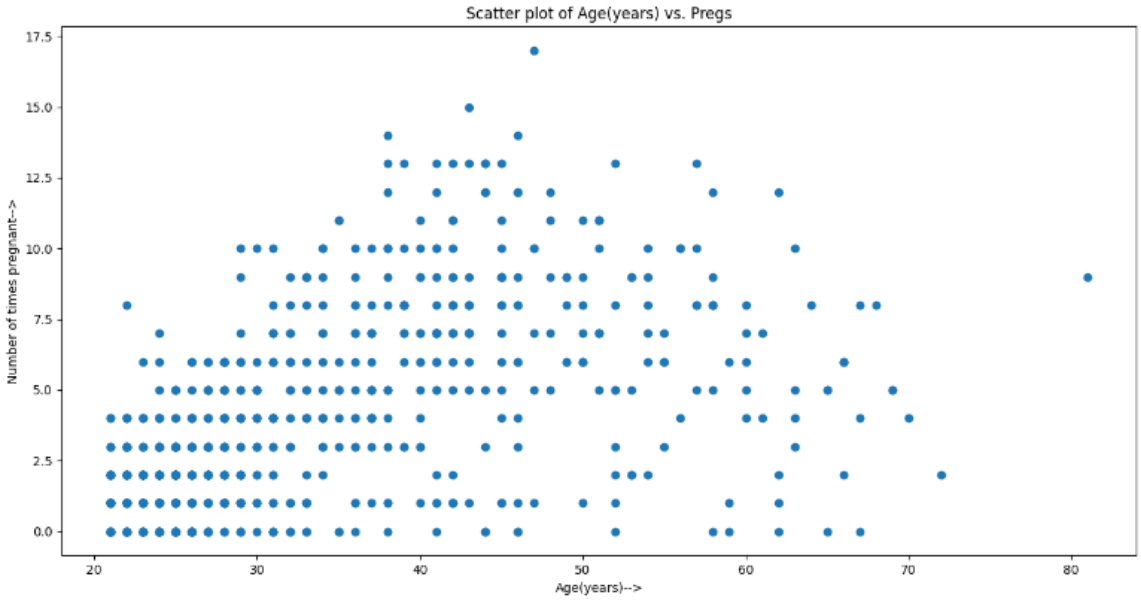

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. Distribution of scatter points indicate that less number of pregnancies are common among people of age 20-40 , whereas higher number of pregnancies are common within age group of 40-60
2. Density of pregs is denser in people of lesser age whereas it is less dense in people of higher age group.
3. Both attributes are positively correlated.

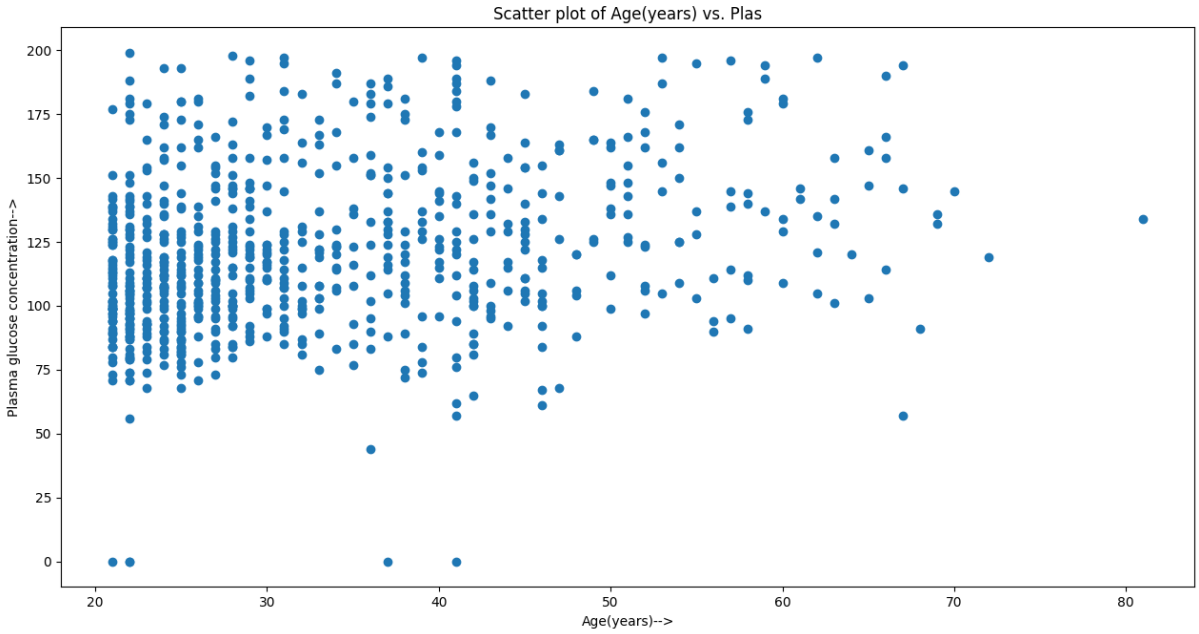

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. Both the attributes are positively correlated.
2. Higher Plasma glucose concentration is denser within age group of(20-35years)
3. People within the higher age group have plas density spread almost evenly within the range of 75-200

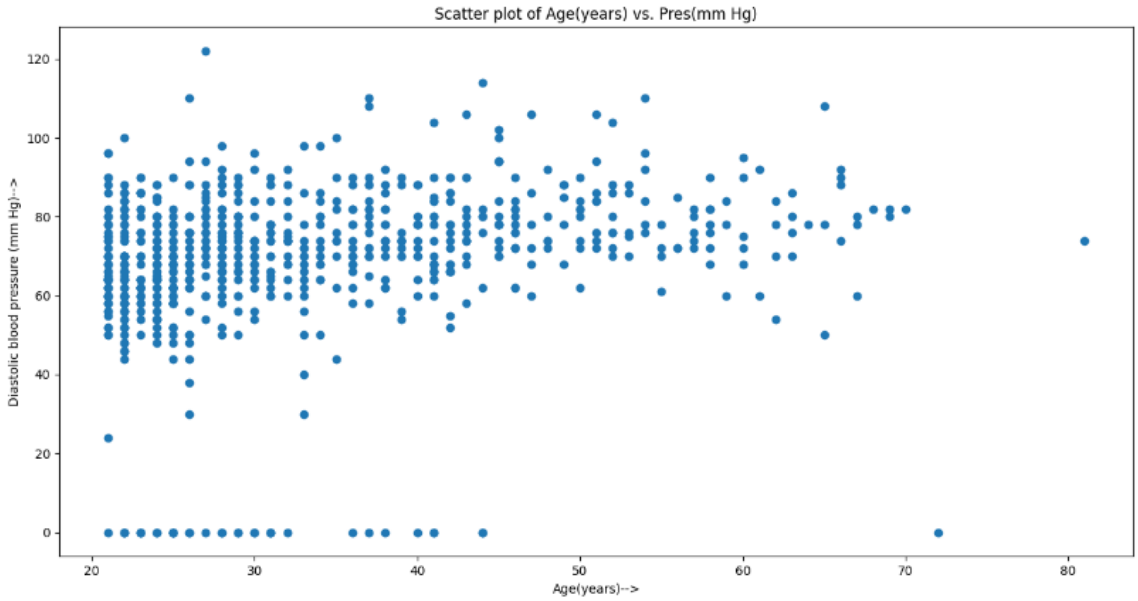

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. Both the attributes are positively correlated.
2. Density of scatter points is more among younger people.
3. Some mistakes in the data given can be seen at blood pressure measured for some people can be seen as 0mm Hg which is not at all possible.

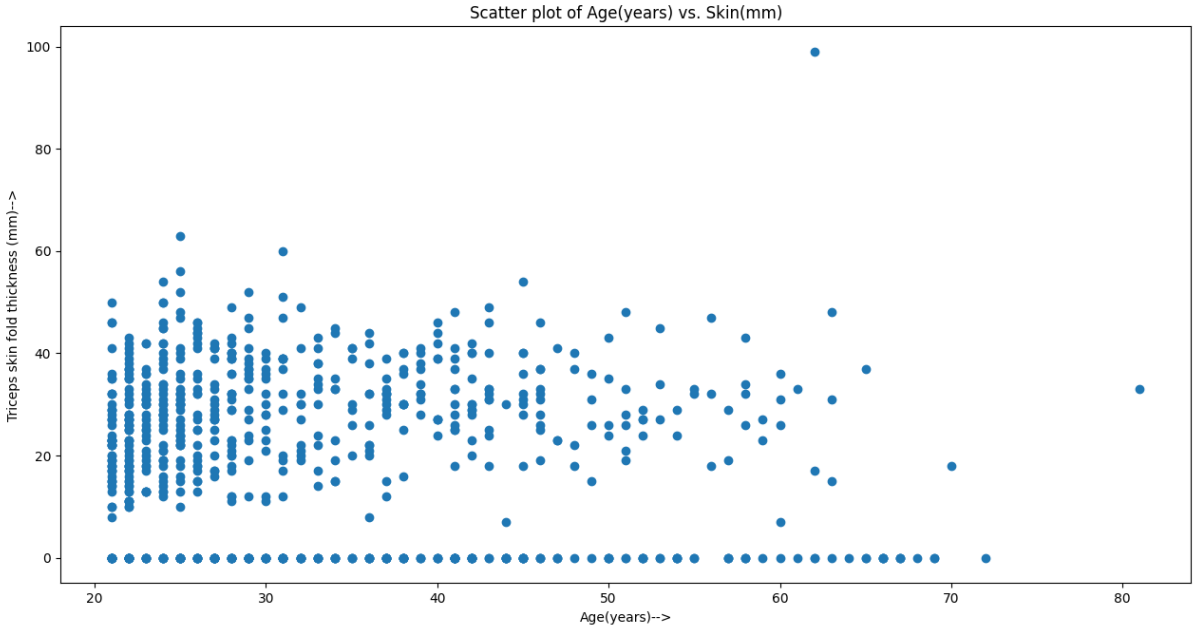

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. Both the given attributes are negatively correlated as seen by the graph.
2. Density of scatter points is higher among people of lower age group(20-40years)

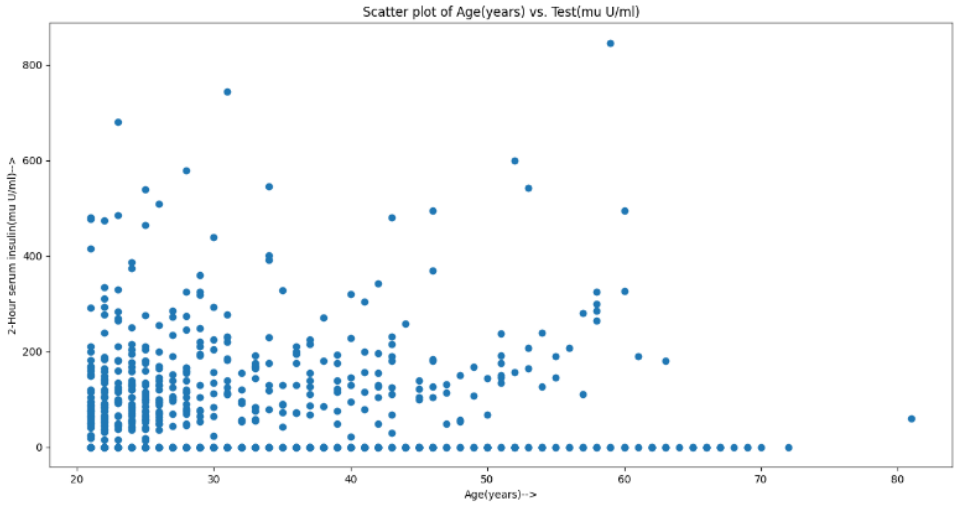

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. Age and test are negatively correlated
2. Density is higher among people of lower age group(20-30years) and w=is mostly less. It is very less dense in people of older age group.

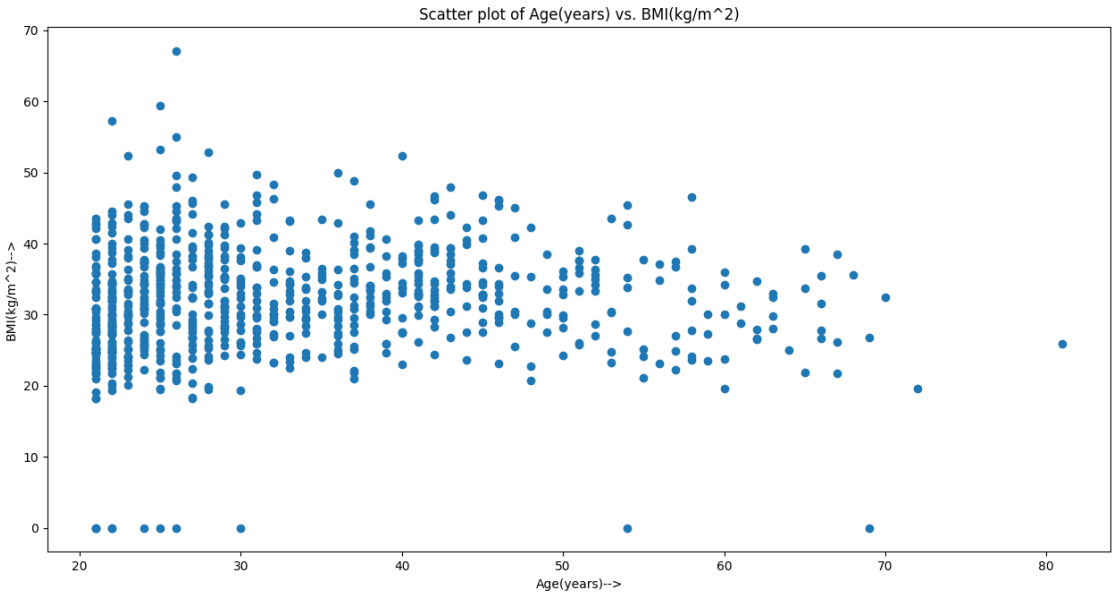

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. Both the attributes are positively correlated. But since their coefficient<0.1 ,the correlation is very weak.
2. Density is more in age group(20-45years).

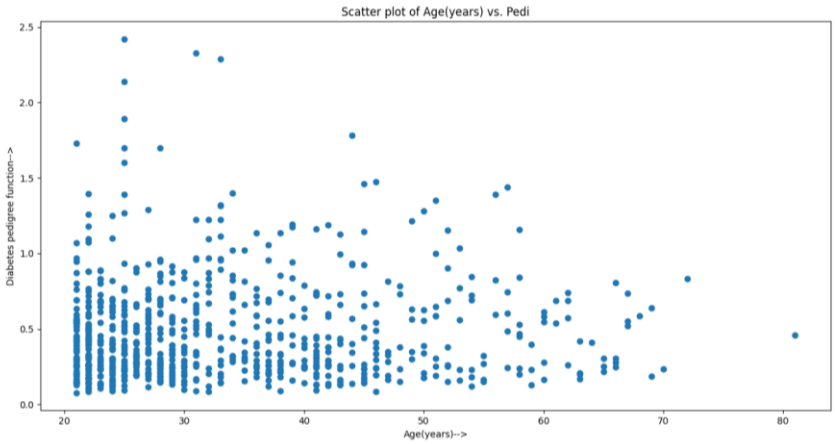

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. Both the attributes are positively correlated. But the correlation is very weak
2. Density is higher among people of age group(20-30years)
3. People of age group 20-30 years have full variation of pedi(0-2.5)

b.

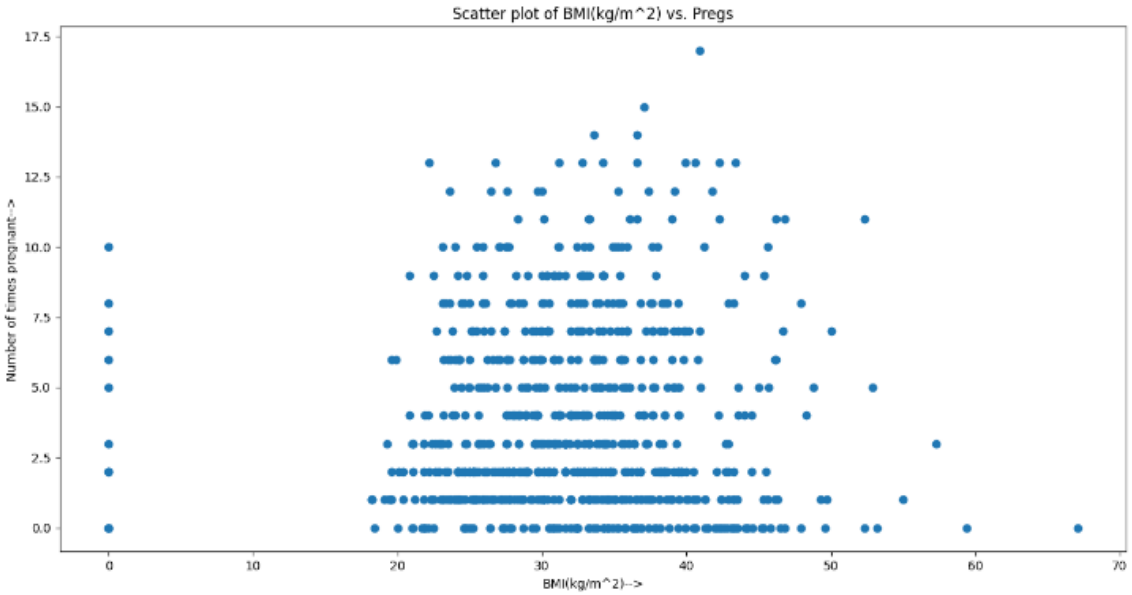

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. Both the attributes are positively correlated but very weakly.
2. Density is high in area of range Age(20-50) and pregs(0-8)
3. Some data show people with BMI 0 having pregnancies which is impossible and it indicates some wrong information the spreadsheet.

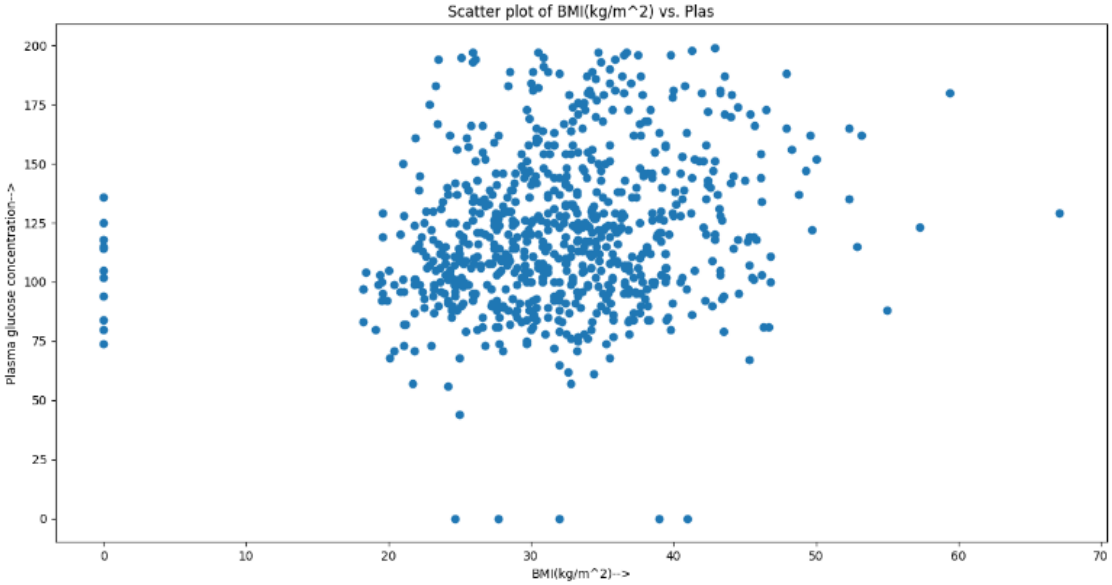

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. Both BMI and plas are positively correlated.
2. Density is high in range age(20-50years) and plas(75-175)
3. Some points show people with BMI 0 which is impossible and it indicates some wrong information the spreadsheet.

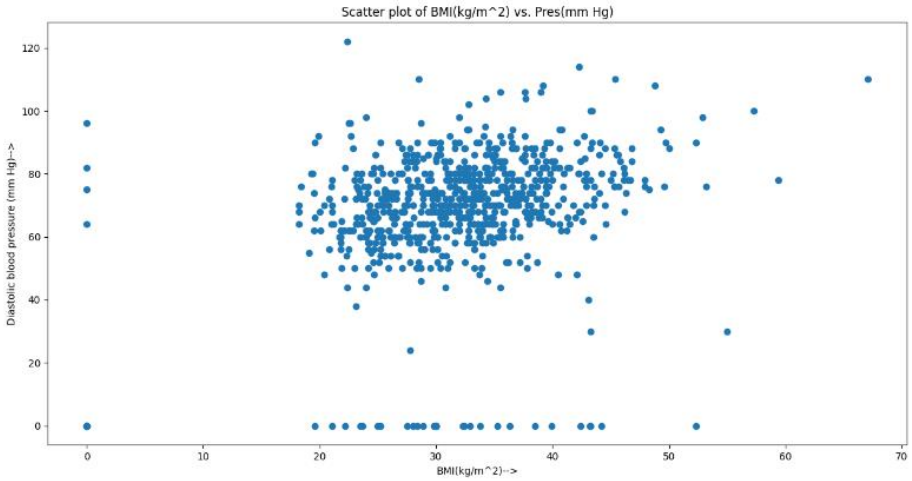

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. Both BMI and pres are positively correlated.
2. Density is higher in range pres(50-90mm Hg).

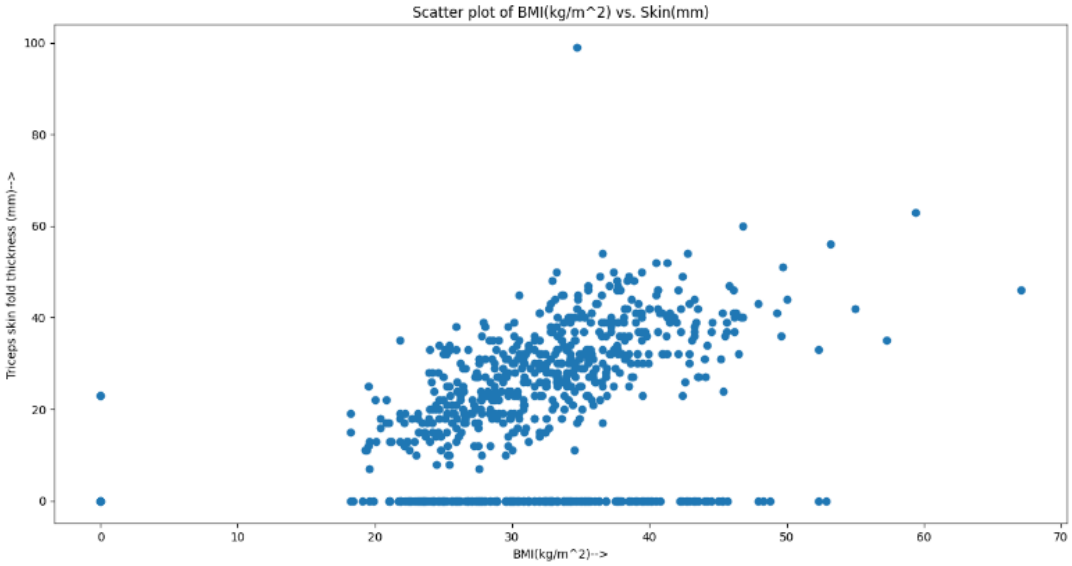

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. BMI and skin both are positively correlated.
2. Density is divided uniformly as the graph proceeds in a tilted line from skin(10-60mm) and age(20-50years)
3. Graph here as seen proceeds in a tilted line

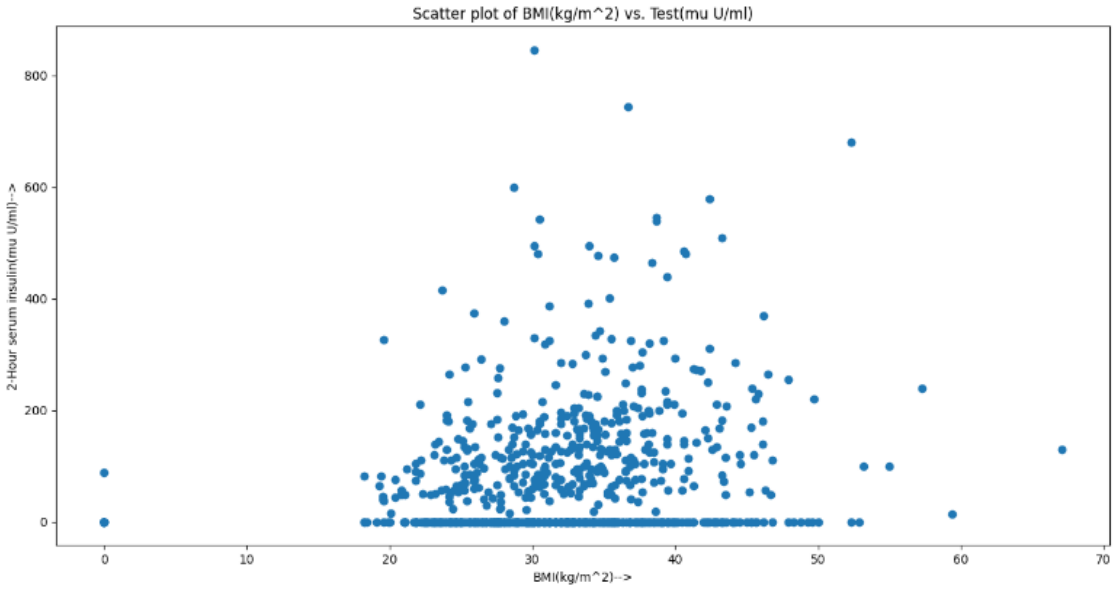

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. Both the attributes are positively correlated
2. Density decreases continuously while going in positive y direction i.e., test from (0 – 600 mu U/ml)

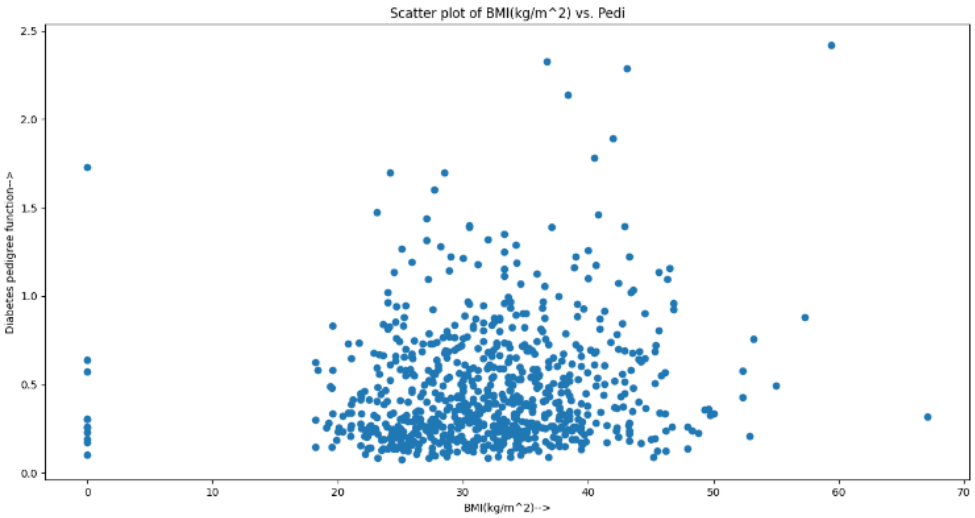

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. Attributes are positively correlated.
2. Density is much higher in lower region.
3. Some points show people with BMI 0 which is impossible and it indicates some wrong information the spreadsheet.

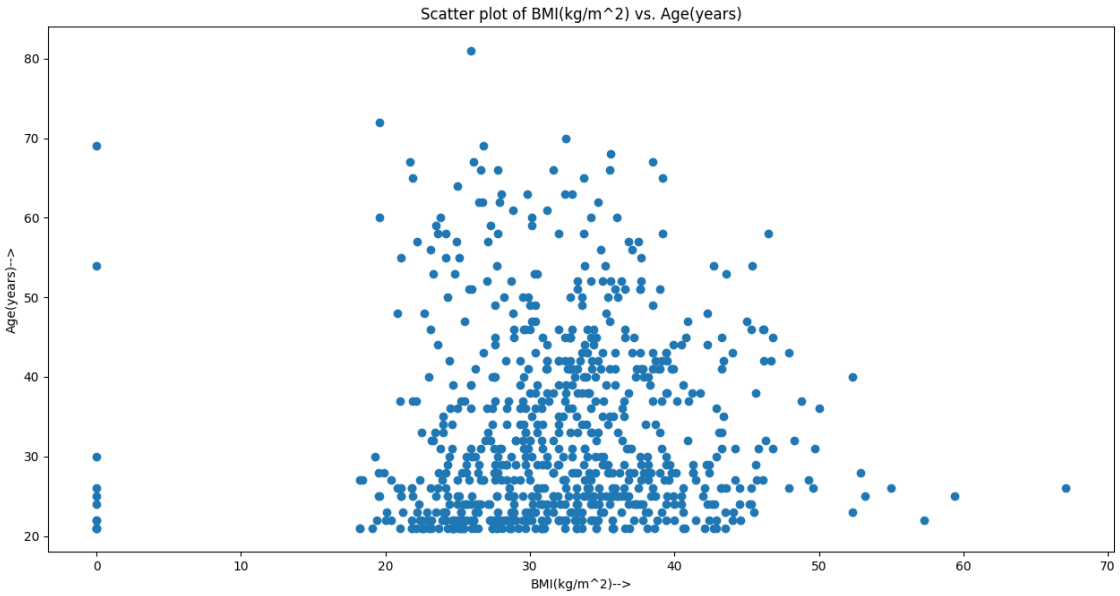

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. Both the attributes are positively correlated. But since their coefficient<0.1 ,the correlation is very weak.
2. Density is more in age group(20-45years).
3. Some points show people with BMI 0 which is impossible and it indicates some wrong information the spreadsheet.

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
8	Age (in years)	1.0

Inferences:

1. Except skin and test all the attributes have positive coefficient of correlation. In case of BMI and pedi correlation is very weak.
2. BMI and pedi have very low coefficient of correlation which explains that while the value of age will increase value of other attributes will also increase but very slowly.
3. Skin and test have negative value of correlation coefficient which means that age and other attributes will increase and decrease in opposite ways. While age will increase other attributes will decrease.
4. Trends of other attributes vary similarly with minor differences.

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)	1.0
7	pedi	0.141
8	Age (in years)	0.036

Inferences:

1. All the attributes have positive coefficient of correlation. In case of Age and pregs correlation is very weak.
2. Age and pregs have very low coefficient of correlation which explains that while the value of BMI will increase value of other attributes will also increase but very slowly.
3. Other properties have similar trends but with very minor differences.

4

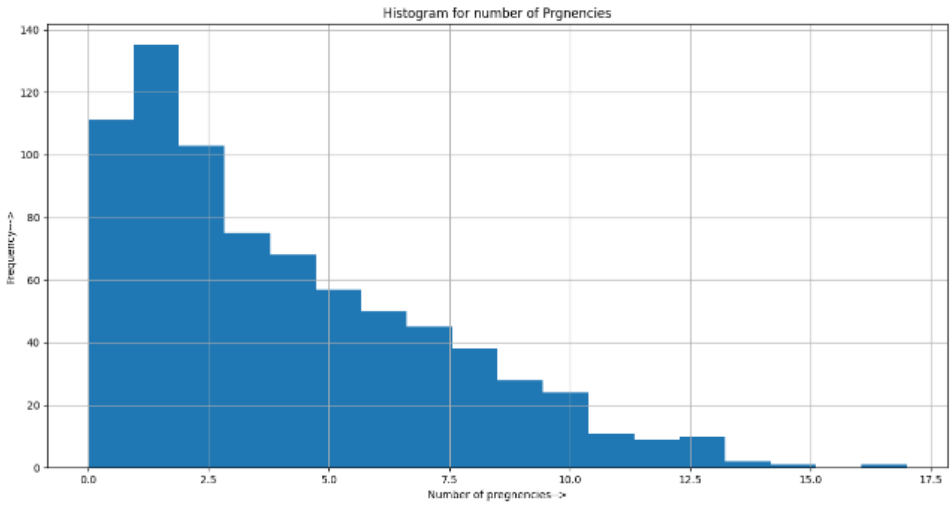

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. Frequency of each bin referring to its height.
   0-1 110
   1-2 135
   2-3 105
   3-4 75
   4-5 70
   5-6 55
   6-7 50
   After this the value of bars continuously decrease on the same rate till, they reach 0.
2. Mode of pregs lies in range 1-2 which can be verified by the function of mode in question 1

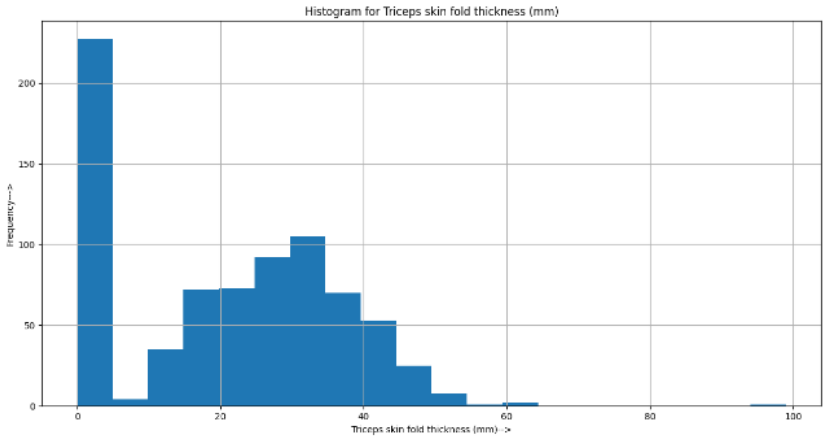

Figure 16 Histogram depiction of attribute skin

Inferences:

1. Frequency of each bin referring to its height. (approx.)

0-5	240	40-45	55
5-10	5	45-50	25
10-15	35	50-55	10
15-20	70		
20-25	70		
25-30	85		
30-35	110		
35-40	70		

   After this range the values of all the other bars almost reach 0.
2. Mode of skin lies in range 0-5 which can be verified by the function of mode in question 1

5

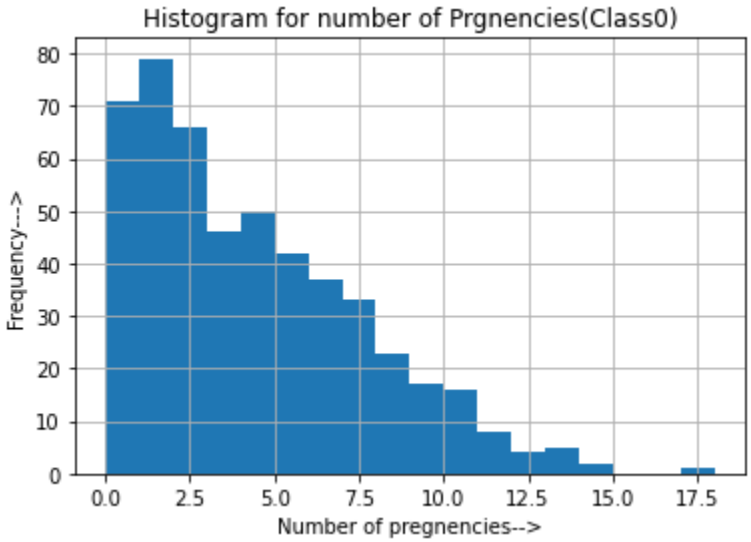

Figure 17 Histogram depiction of attribute pregs for class 0

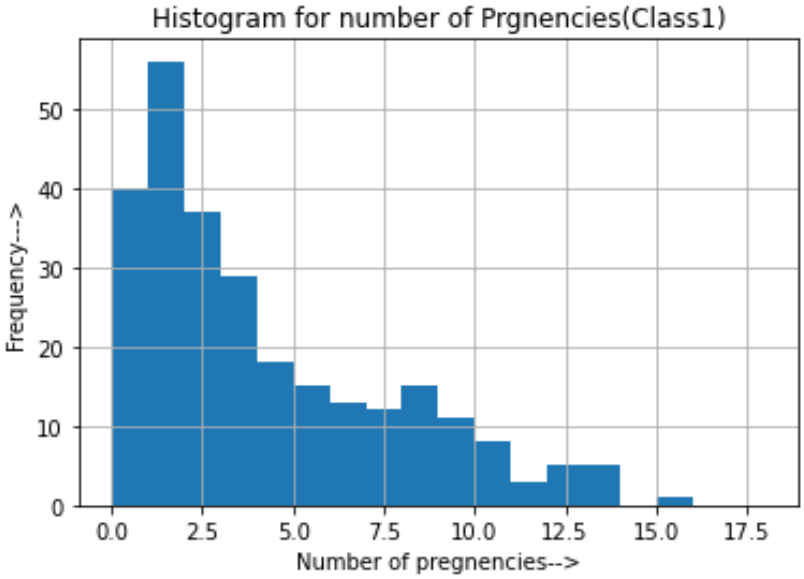

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. In case of class0 mode lies in range while for class1 mode lies in range 1-2.
2. Both the classes possess a same general trend as seen from the plots.

6

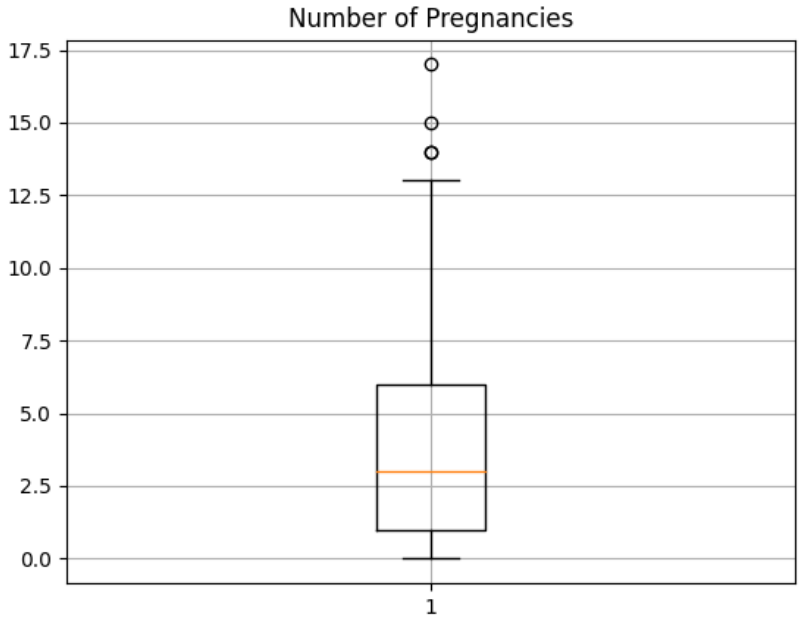

Figure 19 Boxplot for attribute pregs

Inferences:

1. There are 3 outliers for pregs with significantly high value compared to pregs.
2. Inter quartile range=>6-1=5
3. More range of pregs indicates that it has more variability.
4. Box plot shows that pregs is positively skewed.
5. Median is around 3 ,Minimum 0 as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)

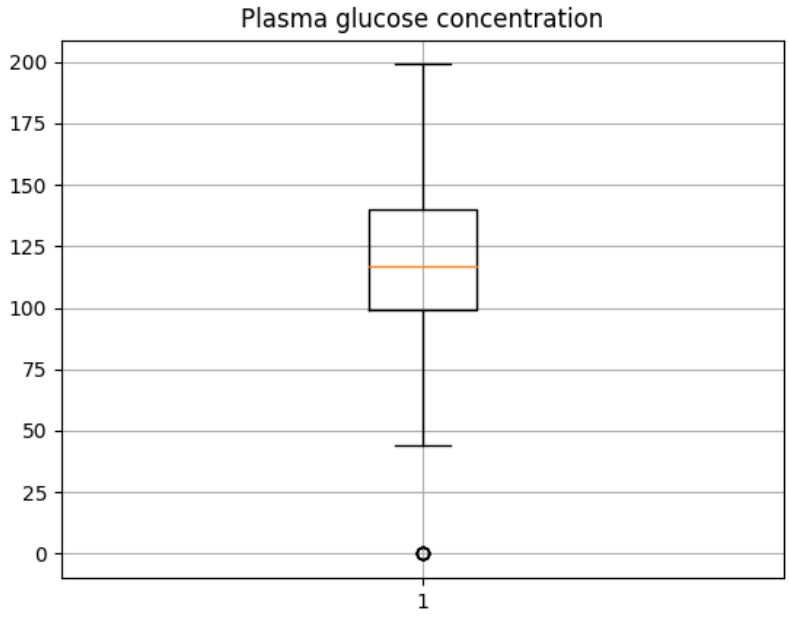

Figure 20 Boxplot for attribute plas

Inferences:

1. There are only 1 outliers for plas with significantly very low value compared to plas.
2. Inter quartile range=>140-100=40
3. More range of plas indicates that it has more variability.
4. Box plot shows that plas is positively skewed.
5. Median is around 120 ,Maximum 199 as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)

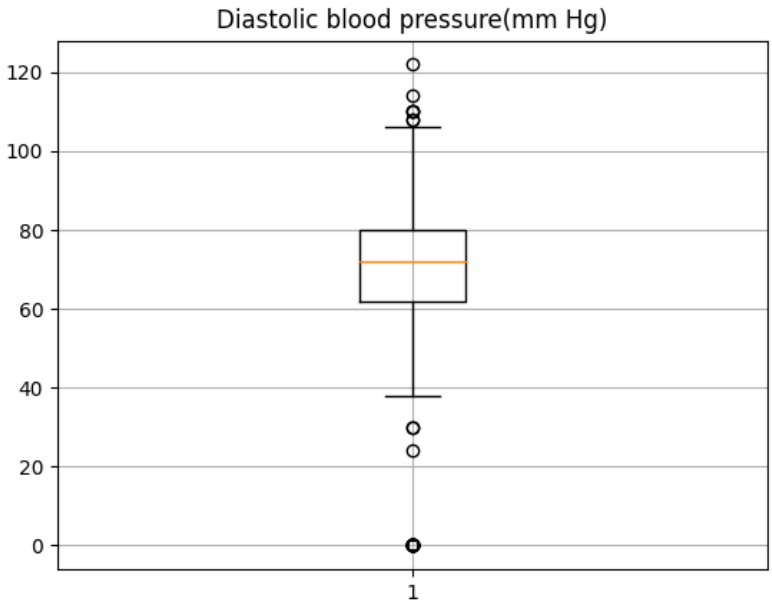

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are only 8-9 outliers for pres with both low and high values compared to boxplot.
2. Inter quartile range=>80-62=18
3. Less range of pres indicates that it has lesser variability.
4. Box plot shows that pres is negatively skewed.
5. Median is around 70mm Hg ,maximum 121mm Hg and minimum 0mm Hg as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)

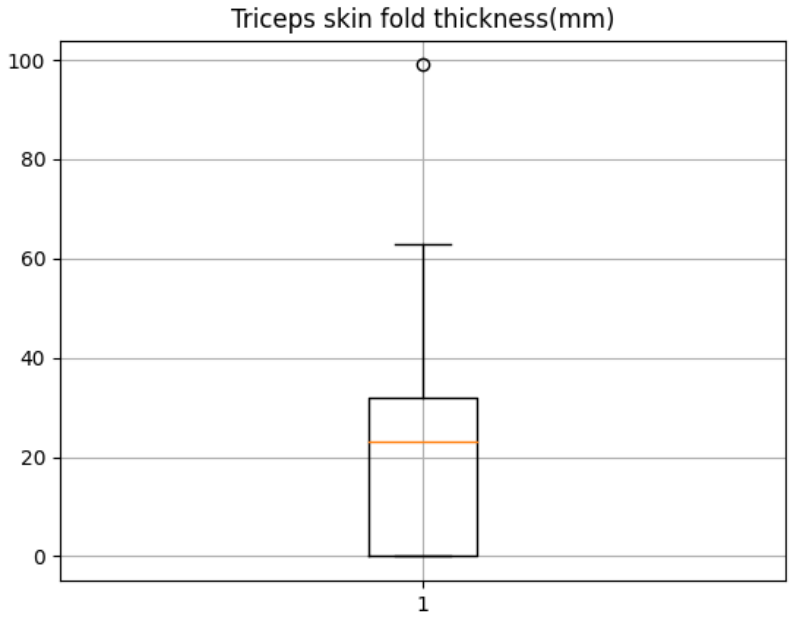

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. There is only 1 outlier for skin with significantly high value compared to boxplot.
2. Inter quartile range=>30-0=20
3. More range of skin indicates that it has more variability.
4. Box plot shows that skin is negatively skewed.
5. Median is around 24mm and minimum 0mm as seen in the boxplot and can be verified using soln1. (All values mentioned in point 5 are approximated)

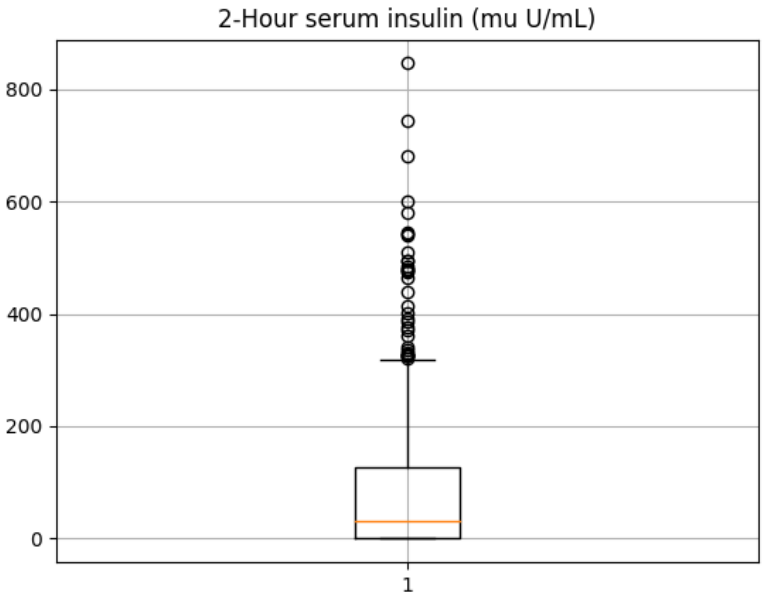

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. There are many outliers for test varying significantly above the boxplot.
2. Inter quartile range=>140-0=140(approx.)
3. More range of skin indicates that it has more variability.
4. Box plot shows that skin is positively skewed.
5. Median is around 35mu U/ml and minimum 0mu U/ml as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)

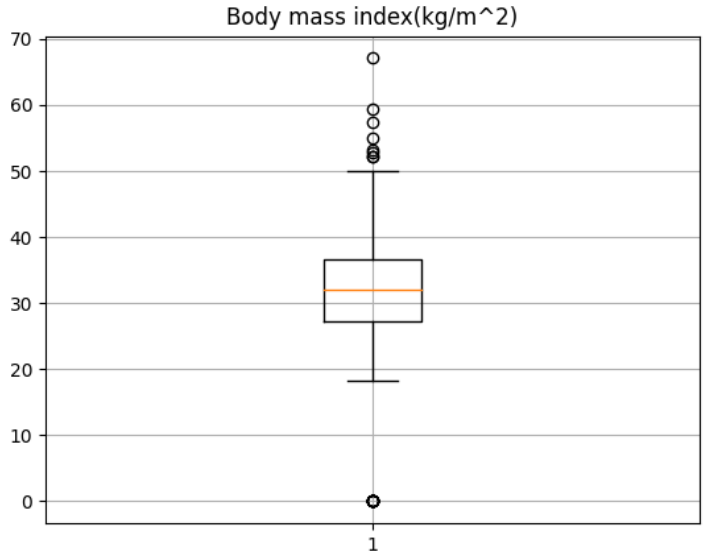

Figure 24 Boxplot for attribute BMI (in kg/m²)

Inferences:

1. There are many outliers for BMI with high values compared to the boxplot.
2. Inter quartile range=>37-27=10(approx.)
3. Less range of BMI indicates that it has lesser variability.
4. Box plot shows that BMI is approximately normally disrtibuted.
5. Median is around 33kg/m^2 ,maximum 70kg/m^2 and minimum 0kg/m^2 as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)

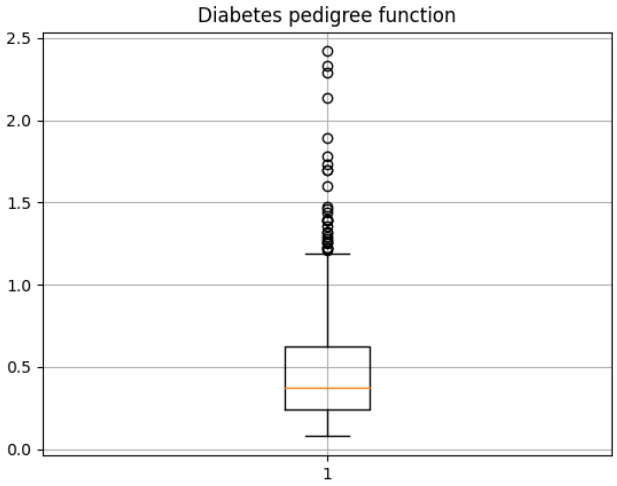

Figure 25 Boxplot for attribute pedi

Inferences:

1. There are many outliers for pedi varying significantly above the boxplot.
2. Inter quartile range=>0.6-0.25=0.35(approx.)
3. Less range of BMI indicates that it has lesser variability.
4. Box plot shows that pedi is positively skewed.
5. Median is 0.35, maximum 2.4 and minimum 0.1as seen in the boxplot and can be verified using soln1. (All values mentioned in point 5 are approximated)

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. There are many outliers for BMI with high values compared to the boxplot.
2. Inter quartile range=>41-25=16(approx.)
3. More range of BMI indicates that it has more variability.
4. Box plot shows that BMI is positively skewed.
5. Median is around 27years, maximum 81years and minimum 19years as seen in the boxplot and can be verified using soln1.
   (All values mentioned in point 5 are approximated)