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

Student's Name: Yash Sharma **Mobile No: 8802131138**

Roll Number: B20241 **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	3.0	1.0	0.0	17.0	3.367
2	plas	120.895	117.0	100.0	0.0	199.0	31.952
3	pres (in mm Hg)	69.105	72.0	70.0	0.0	122.0	19.343
4	skin (in mm)	20.536	23.0	0.0	0.0	99.0	15.942
5	test (in mu U/mL)	79.799	30.5	0.0	0.0	846.0	115.169
6	BMI (in kg/m^2)	31.993	32.0	32.0	0.0	67.1	7.879
7	pedi	0.472	0.372	0.254	0.078	2.42	0.331
8	Age (in years)	33.241	29.0	22.0	21.0	81.0	11.753

Inferences:

1. It is observed that for the attribute "pedi" the standard deviation is very close to zero from which we may infer that mean, median and mode will be close to each other. The same can also be concluded from their observed values. We may also say that the data for the same can be approximated to be somewhat symmetrically distributed.
2. It is observed that for the attribute "test" the mean is observed to be very high compared to the median. From here we may infer that the sample data for the same may contain outliers having values that are unusual compared to the rest of the data set by being especially small or large in numerical value. This also shows that mean is not robust to outliers.
3. It is observed for the attribute "BMI", the mean, median and mode are equal. Thus, it may be inferred that the sample data for the same is symmetrically distributed.
4. As mean and median for some of the attributes are not same, we may conclude here that the data set distribution may be skewed right/left and will be asymmetrically distributed. i.e. "pregs" may have left skewed distributed data set etc.

2 a.

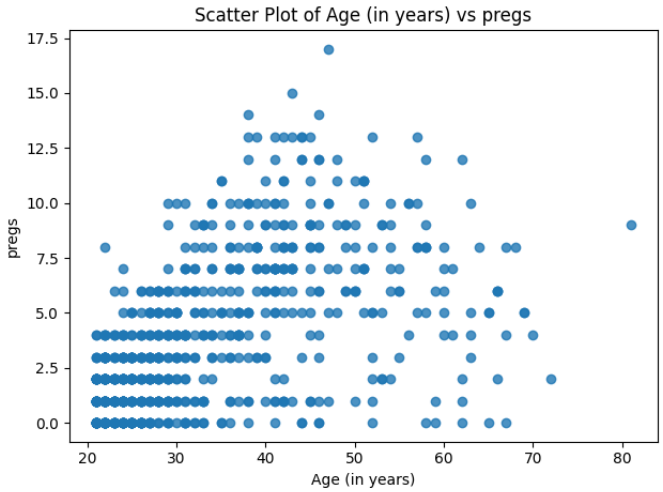

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. It is observed that attribute 1 (“Age”) is moderately positively correlated with attribute 2 (“pregs”) as when “Age” increases “pregs” also shows moderate tendency to increase.
2. The region of “Age” from 20 yrs. to 40 yrs. and “pregs” from 0 to 7.5 seems to be denser as it is heavily populated compared to rest of the plot.
3. The scatter plot is quite scattered everywhere.
4. There are also some outliers present which lie considerably far from the visible clusters.

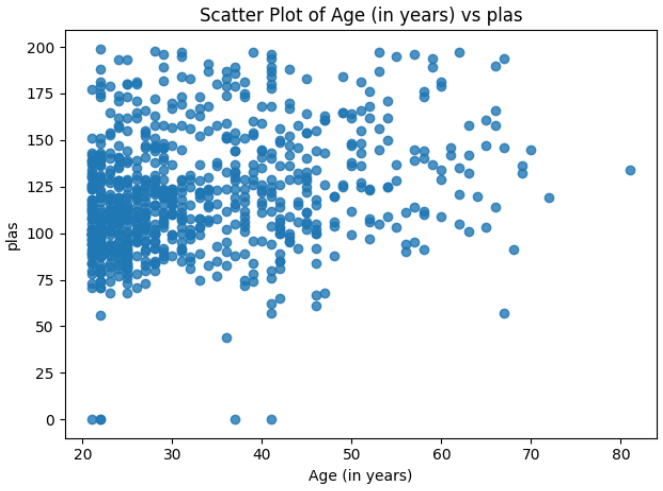

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. It is observed that as "Age" increases, the "plas" remains nearly constant or increases very slightly which is very hard to spot. Thus, there is very weak positive correlation.
2. Region: Age (20-35 years) and plas (60-150) seems to be very dense in the plot.
3. Considerable number of outliers are also visible.

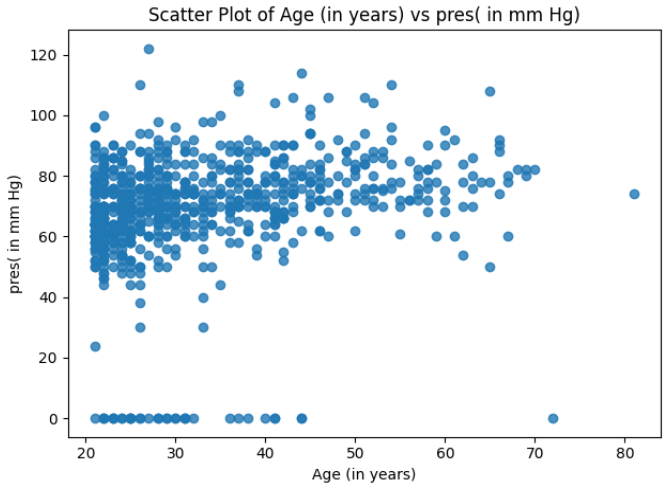

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. It is observed that as "Age" increases, the "pres" remains nearly constant or increases very slightly which is very hard to spot. Thus, there is very weak positive correlation.
2. High density of data points is visible in the region: Age (20-50 years) and pres (41-95 mm-Hg).
3. There are also may outliers present in the plot.

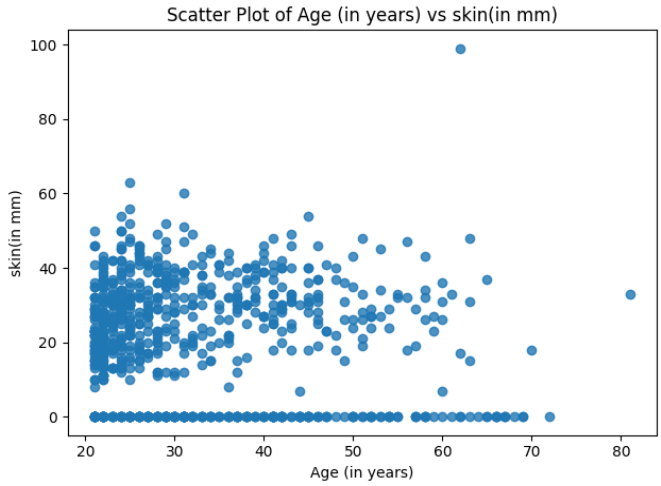

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. It is observed that as "Age" increases, the "skin" remains nearly constant or decreases very slightly which is very hard to spot. Thus, there is very weak negative correlation.
2. The region: Age (20-33 years) and skin (10-50 mm) seems to contain high density of points.
3. There are also some amount of outliers present.

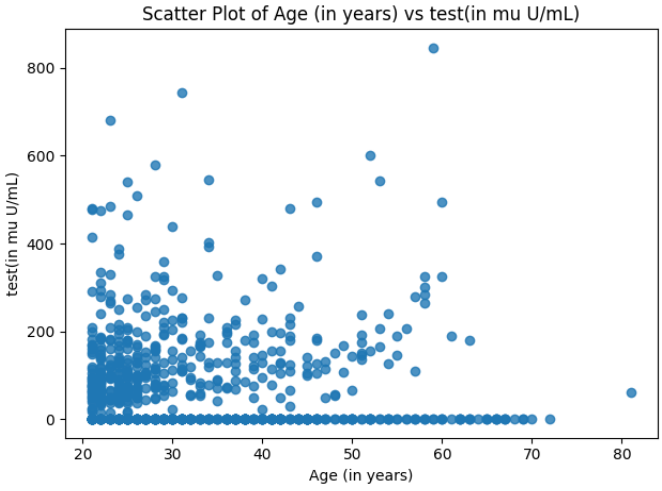

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. The plot between "Age" and "Test" is scattered everywhere, moreover no exact relationship is visible via the scatter plot. Thus, we may say that there is very low correlation.
2. The region: Age (20-33 years) and test (0-200 mu-U/ml) seems to contain high density of points.
3. There are also many outliers present in the plot.

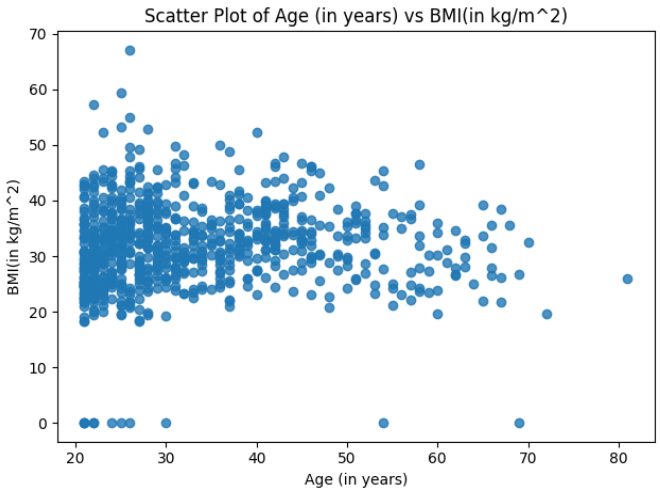

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m²)

Inferences:

1. There seems to no clear relationship between the two attributes. As "Age" increases the "BMI" seems to be following no specific direction, thus it may be concluded that there is very low positive/negative correlation.
2. The region: Age (20 – 45 years) and BMI (20 - 45 kg/m^2) seems to contain high density of points.
3. Low number of Outliers are visible in the plot.

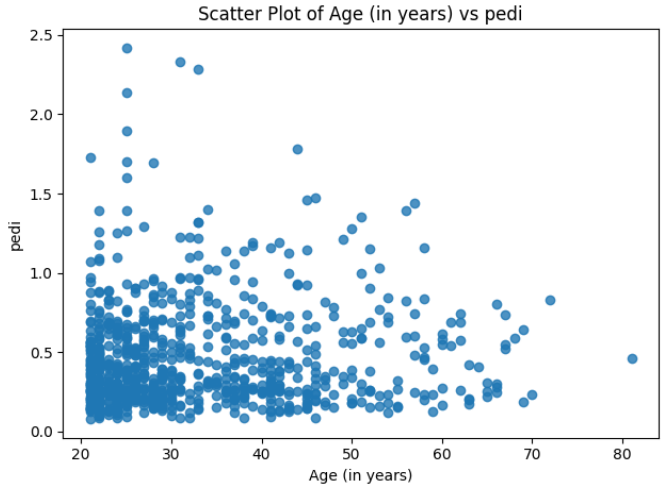

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. There seems to no clear relationship between the two attributes. The plot is scattered. We may conclude that there is very low correlation visible.
2. The region: Age (20-30 years) and pedi (0-1) seems to contain high density of points.
3. There are also considerable number of outliers present in the plot.

b.

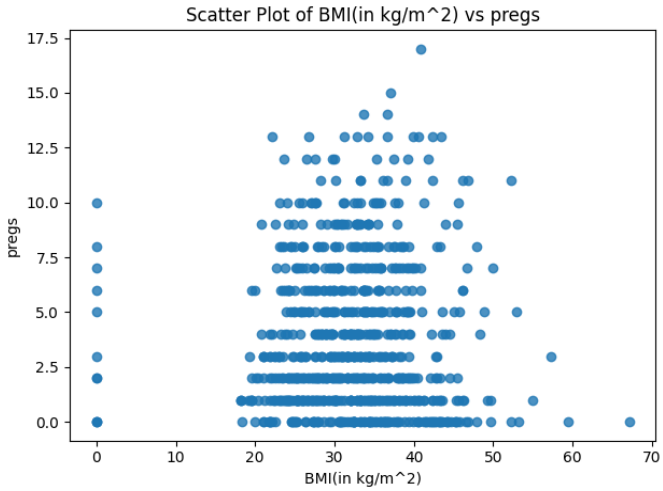

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. There seems to no clear relationship between the two attributes. ”BMI” seems to remain nearly constant or increase very slightly while “pregs ” increases. Thus, there is very negligible positive correlation.
2. A cluster of points in the bottom-middle plot is visible.
3. Many outliers are also visible in the plot.

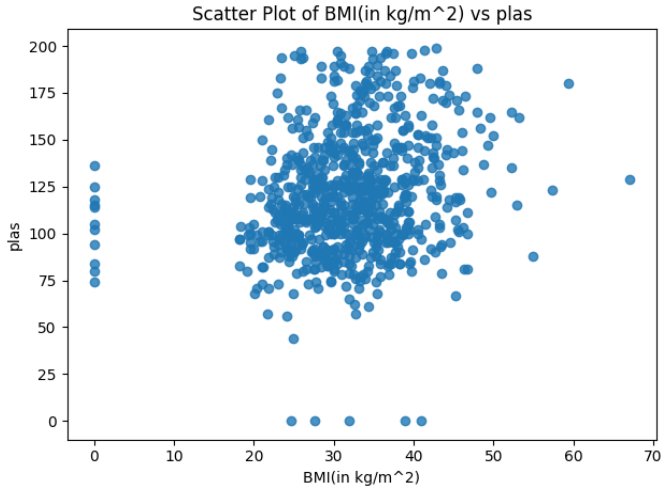

Figure 9 Scatter plot: BMI (in kg/m²) vs. plas

Inferences:

1. It is observed that as “BMI” increases, “plas” also seems to increase weakly. Thus, a low positive correlation is visible.
2. A clear cluster dense points is visible in region: BMI (20-42 kg/m^2) and plas (75-150 mm Hg).
3. Many outliers are also visible in the scatter plot.

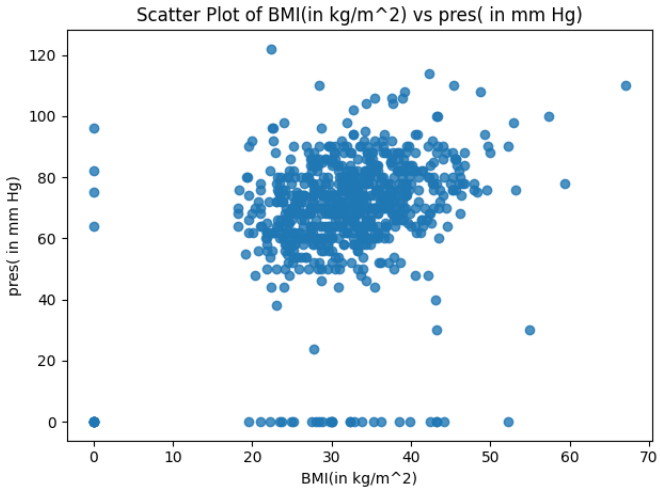

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

4. It is observed that as “BMI” increases, “pres” also have weak tendency to increase. Thus, a low positive correlation is visible.
5. A clear cluster dense points is visible in region: BMI (20-42 kg/m^2) and pres (40-100 mm Hg).
6. Many outliers are also visible in the scatter plot.

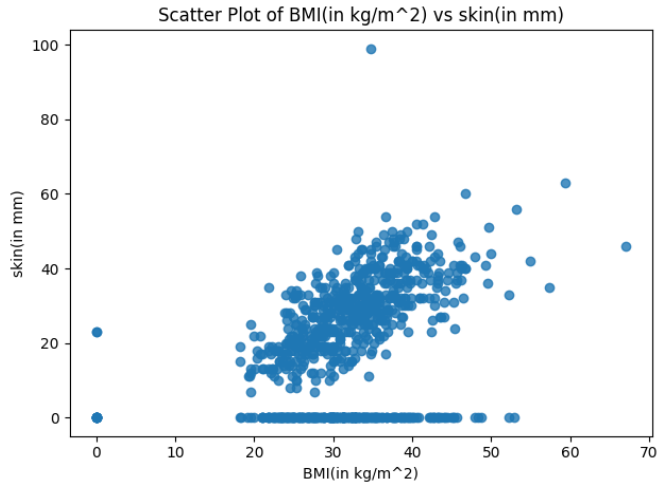

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. It is observed that as “BMI” increases, “skin” also increases. Thus, we may conclude that there is moderate to low positive correlation visible.
2. A clear cluster of points is visible in region: BMI (20-45 kg/m^2) and skin (20-60 mm).
3. A low number of outliers are also visible in the plot.

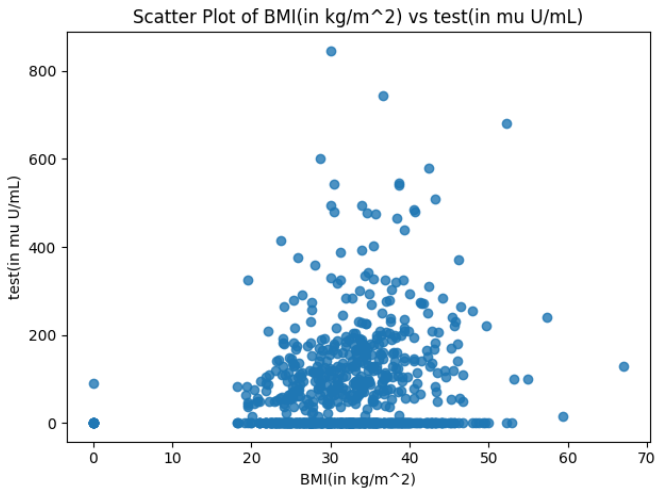

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. The plot is quite scattered and contains a cluster of points. As “BMI” increases very slight increase in “test” is visible. Thus, there is very low positive correlation visible.
2. A dense cluster of points is visible in region: BMI (20-45 kg/m^2) and test (0-230 mu U/mL).
3. Outliers are also visible in the plot.

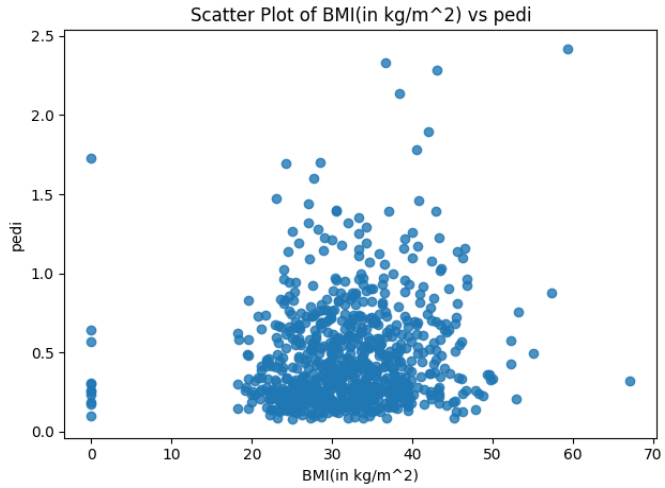

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. There seems to no clear relationship between the two attributes. There seems to be a cluster formed in the plot. No specific dependence is visible. Thus, we may conclude that there is very negligible correlation between the attributes.
2. A clear dense cluster of points is visible in region: BMI (20-45 kg/m^2) and pedi (0-1).
3. There are also some outliers visible in the plot.

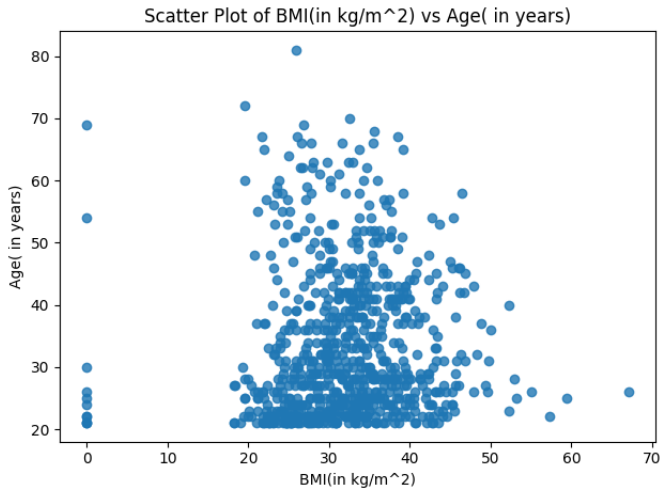

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. There seems to no clear relationship between the two attributes. As “Age” increases the “BMI” seems to be following no specific direction, thus it may be concluded that there is very low correlation.
2. The region: Age (20-45years) and BMI (20-45 kg/m^2) seems to contain high density of points.
3. Low number of Outliers are visible in the plot.

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)	0.240
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

Inferences:

1. It is observed that the attribute “pregs” has high (strong) degree of correlation while all the rest of the attributes (plas, pres (in mm Hg), skin (in mm), test (in mu U/mL) , BMI (in kg/m2), pedi) have low degree of correlation with the attribute “Age”.
2. Attributes: {pregs, plas, pres (in mm Hg), BMI (in kg/m2), pedi } are positively correlated with the attribute “Age” which means if “Age” increases or decreases consequently the other attribute will increase or decrease. While for the attributes {skin (in mm), test (in mu U/mL)}, the negative correlation signifies that if “Age” decreases or increases the other attribute will go in opposite direction i.e. increase or decrease.
3. The inferences corresponding to the respective scatter plots on their correlation coefficient is much closer to that visible from the above correlation table.

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
7	pedi	0.141
8	Age (in years)	0.036

Inferences:

1. It is observed that the attribute “skin (in mm)” has moderate (medium) degree of correlation while all the rest of the attributes (pregs, plas, pres (in mm Hg), test (in mu U/mL) , BMI (in kg/m^2), pedi) have low degree of correlation with the attribute “BMI”.
2. Attributes: { pregs, plas, pres (in mm Hg), test (in mu U/mL) , skin (in mm), BMI (in kg/m^2), pedi } are positively correlated with the attribute “BMI” which means if “BMI” increases or decreases consequently the other attribute will increase or decrease.
3. The inferences corresponding to the respective scatter plots on their correlation coefficient is much closer to that visible from the above correlation table.

4 a.

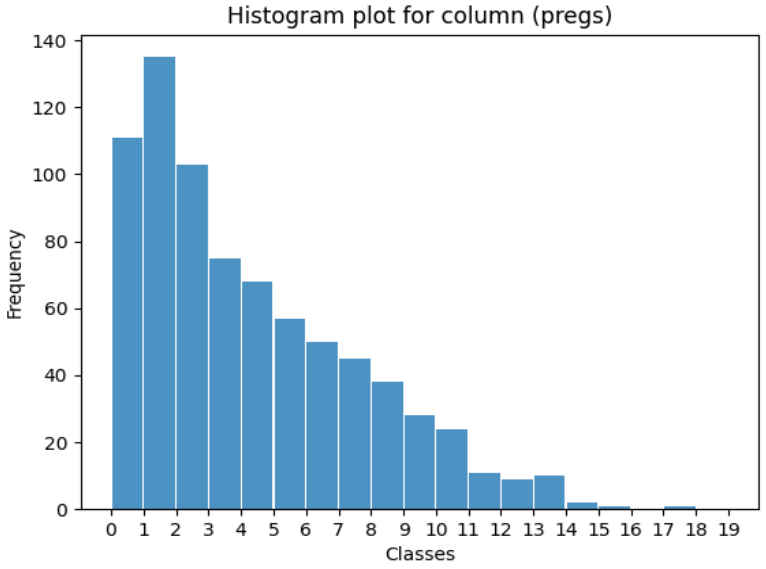

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. It can be observed that the bin height after class (1,2) is decreasing continuously which suggests that the given data may be exponentially distributed. We may conclude that as the value of attribute "preg" increases a gradual decrease in its frequency in given data is observed.
2. It is observed that the class of range 1-2 has maximum frequency or has highest bin height. Thus, the mode (most frequently occurring data point) for the attribute "pregs" lies in bin of class in range 1-2.
3. It is a Unimodal left skewed histogram with approximately exponentially distributed data.

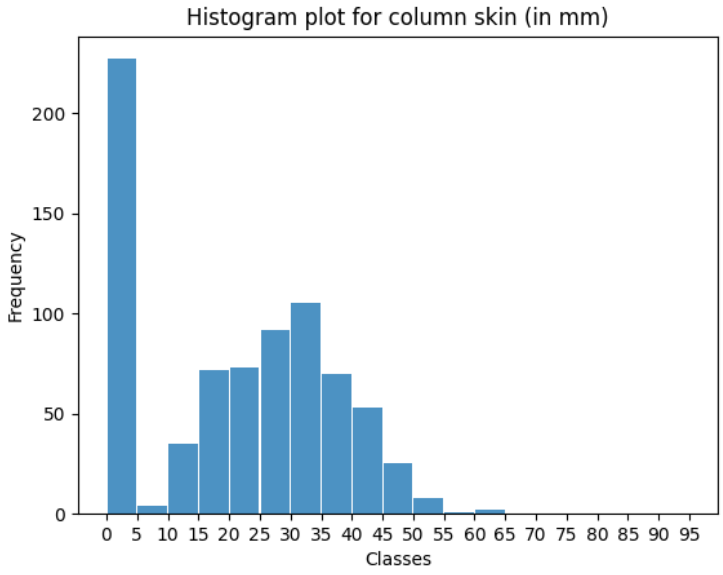

Figure 16 Histogram depiction of attribute skin

Inferences:

1. It is observed that the frequency of each bin (referring to its height) is highest for class in range (0-5) but after that the frequency increases, reaches peak and then decreases uniformly. The latter part of the histogram is much closer to symmetric distribution.
2. The bin of class in range (0-5) has the highest frequency. Thus, we may conclude that the mode for the attribute "skin" lies in that bin only.
3. It is a Bimodal/Multimodal Histogram (as it has 2 peaks).
4. It is a right-skewed Histogram. (A lump of data lies on left side followed by tail of rest of the data towards left.)

5

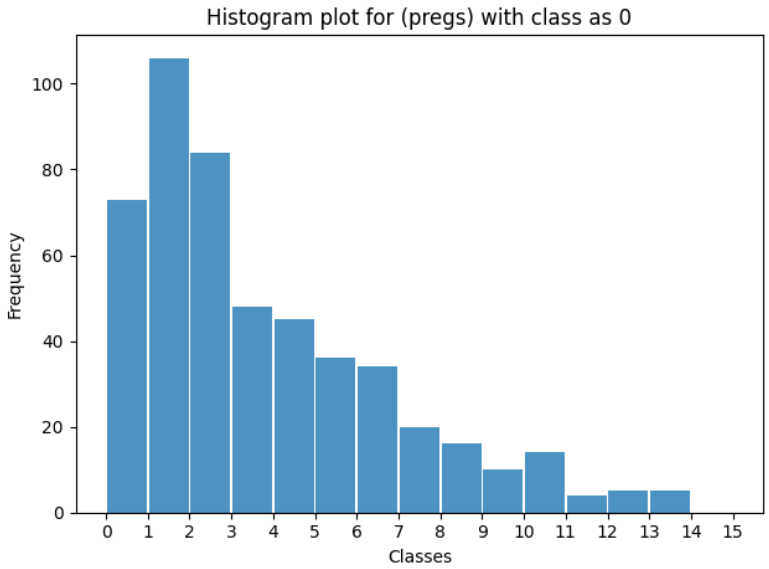

Figure 17 Histogram depiction of attribute pregs for class 0

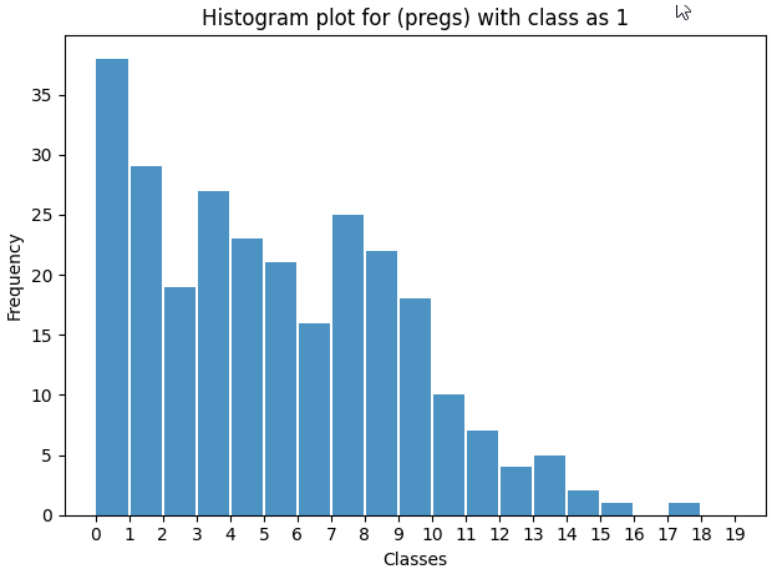

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. From the above two histogram, we may observe that the frequency for bin of modal class 1-2 for attribute “class”-0 is nearly 105 while for bin of modal class 1-2 for attribute “class”-1 is nearly 40. It is clearly visible as the frequency of bin for modal class for “class-0” is nearly twice as that for “class-1”, thus we may conclude that mode of the attribute “pregs” lie in bin with class range 1-2.
2. Both the histograms (excluding some peaks) have uniformly decreasing frequency. The trend is more prominent for the histogram with “class-0”. Thus, in general we may conclude that the data represented in the both the histograms is somewhat exponentially distributed.
3. It may be concluded that both histograms are right -skewed and are Unimodal in nature.

6

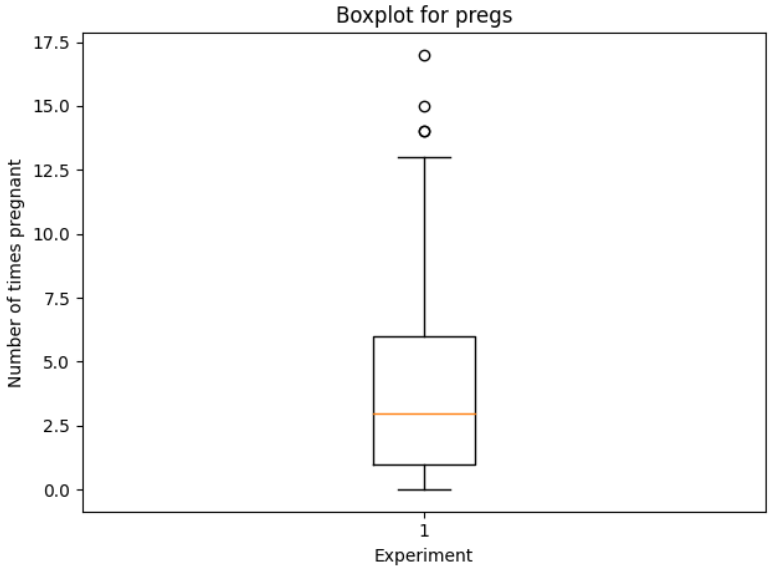

Figure 19 Boxplot for attribute pregs

Inferences:

1. It can be observed that 3 outliers are present above the top whisker with values greater than (QUARTILE 3 +1.5 * Interquartile range or 12.5)
2. The Inter quartile range may be calculated as, IQR = QUARTILE 3-QUARTILE 1 = 6 - 0.99 = 5.01
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=3).
4. It can be observed that the longer part of the box is above the median, thus the data is left-skewed (Negative skewness) .
5. We can clearly observe that the median, max, min from box plot are 3, 17, 0 respectively which exactly matches with the value calculated in the table for Q1.

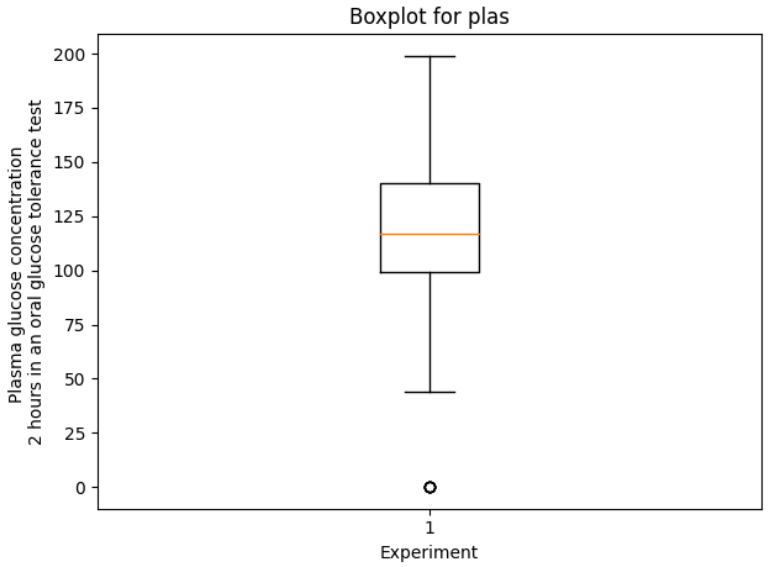

Figure 20 Boxplot for attribute plas

Inferences:

1. It can be observed that 1 outlier is present below the bottom whisker with values lower than QUARTILE 1 -1.5 * Interquartile range or 44.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=140.2-99.2=41
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=117).
4. It can be observed that the longer part of the box is above the median, thus the data is left-skewed (Negative skewness) .
5. We can clearly observe that the median, max, min from box plot are 117, 199, 0 respectively which exactly matches with the value calculated in the table for Q1.

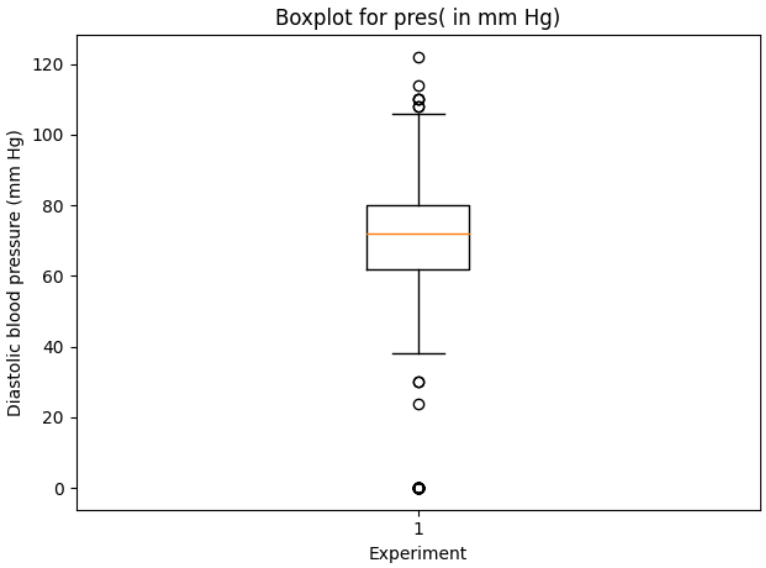

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. It can be observed that the outliers are present on both sides namely above the top whisker with values greater than QUARTILE 3+1.5*IQR or 106 mm Hg and below the bottom whisker with values less than QUARTILE 1-1.5*IQR or 38.1 mm Hg.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=79.7-61.9=17.8 mm Hg.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=72 mm Hg).
4. It can be observed that the longer part of the box is below the median, thus the data is right-skewed (Positive skewness).
5. We can clearly observe that the median, max, min from box plot are 72 mm Hg, 122 mm Hg, 0 mm Hg respectively which exactly matches with the value calculated in the table for Q1.

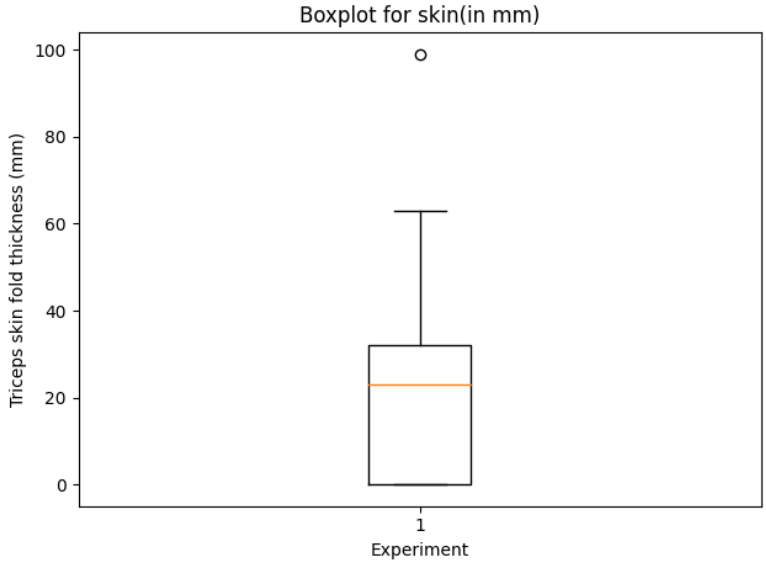

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. It can be observed that 1 outlier is present above the top whisker with values greater than QUARTILE 3+1.5 * IQR or 62.7 mm.
2. The Inter quartile range may be calculated as, IQR = QUARTILE 3 - QUARTILE 1 = 32.1 mm -0.4 mm = 31.7 mm.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=23 mm).
4. It can be observed that the longer part of the box is below the median, thus the data is right-skewed (Positive skewness).
5. We can clearly observe that the median, max, min from box plot are 3 mm, 17 mm, 0 mm respectively which exactly matches with the value calculated in the table for Q1.

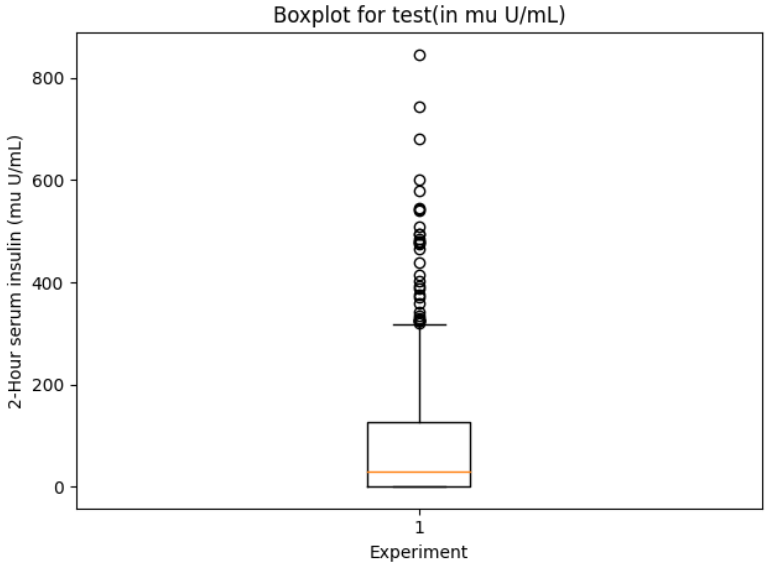

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. It can be observed that many outliers are present above the top whisker with values greater than QUARTILE 3 +1.5 * Interquartile range or 315 mu U/mL.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=126-3=123 mu U/mL.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=30 mu U/mL).
4. It can be observed that the longer part of the box is above the median, thus the data is left-skewed (Negative skewness).
5. We can clearly observe that the median, max, min from box plot are 30 mu U/mL, 845 mu U/mL, 0 mu U/mL respectively which matches with the value calculated in the table for Q1.

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

Inferences:

1. It can be observed that the outliers are present on both sides namely above the top whisker with values greater than QUARTILE 3+1.5*IQR or 50 kg/m^2 and below the bottom whisker with value less than QUARTILE 1-1.5*IQR or 18.3 kg/m^2.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=36.5 kg/m^2 -27.3 kg/m^2 =9.2 kg/m^2.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=31.99 kg/m^2).
4. It can be observed that the median divides both the box in nearly two equal halves and the whiskers can also be assumed of nearly same length. Thus, we may conclude here that the given data may be symmetrically distributed about all the three central tendencies (mean, mode and median are equal here).
5. We can clearly observe that the median from the box plot, QUARTILE 2=31.99 kg/m^2 exactly matches with the value calculated in the table for Q1. Also, the calculated values of mean, mode and median from table are very close to each other which also further concludes the symmetrically distributed nature of data.

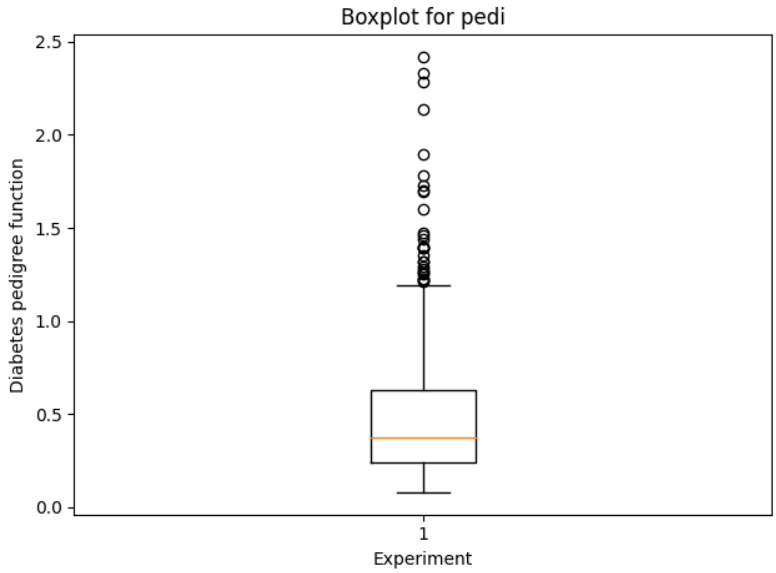

Figure 25 Boxplot for attribute pedi

Inferences:

1. It can be observed that many outliers are present above the top whisker with values greater than QUARTILE 3 +1.5 * Interquartile range or 1.192 .
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=0.623-0.241=0.382.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=0.368).
4. It can be observed that the longer part of the box is above the median and lower whisker is shorter, thus the data is left-skewed (Negative skewness).
5. We can clearly observe that the median, max, min from box plot are 0.368, 2.420, 0.078 respectively which exactly matches with the value calculated in the table for Q1.

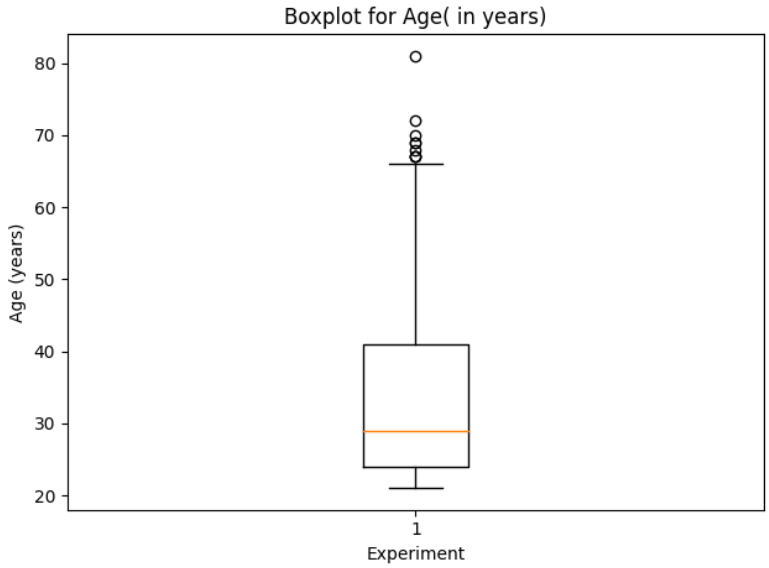

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. It can be observed that many outliers are present above the top whisker with values greater than QUARTILE 3 +1.5 * IQR or 65.9 years.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=41-24=17 years.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=28.9 years).
4. It can be observed that the longer part of the box is above the median and lower whisker is shorter, thus the data is left-skewed (Negative skewness).
5. We can clearly observe that the median, max, min from box plot are 28.9 years, 81 years, 21 years respectively which exactly matches with the value calculated in the table for Q1.