IC 272: DATA SCIENCE– III
LAB ASSIGNMENT – I
Data Visualization and Descriptive analytics

Student's Name: Sachin Mahawar **Roll Number:** B20129

Mobile No: 9166843951 **Branch:** CSE

1.

Table 1 : Mean, median, mode, minimum, maximum and standard deviation for all the attributes

S. No.	Attributes	Mean	Median	Mode	Max.	Min.	S.D.
1	pregs	3.8450	3.0	1	17	0	3.3673
2	plas	120.8945	117.0	99,100	199	0	31.9517
3	pres (in mm Hg)	69.1054	72.0	70	122	0	19.3432
4	skin (in mm)	20.5364	23.0	0	99	0	15.9418
5	test (in mu U/mL)	79.7994	30.5	0	846	0	115.1689
6	BMI (in kg/m^2)	31.9925	32.0	32.0	67.1	0	7.8790
7	pedi	0.4718	0.3725	0.254, 0.258	2.42	0.078	0.3311
8	Age (in years)	33.2408	29.0	22	81	21	11.7525

Inferences:

1. When standard deviation of data set is close to zero, it means that distribution is tightly grouped around mean while high value of standard deviation means data set is widely spread. That's why, mean, median and mode of the data set with low std. deviation is close to each other.

2 (a).

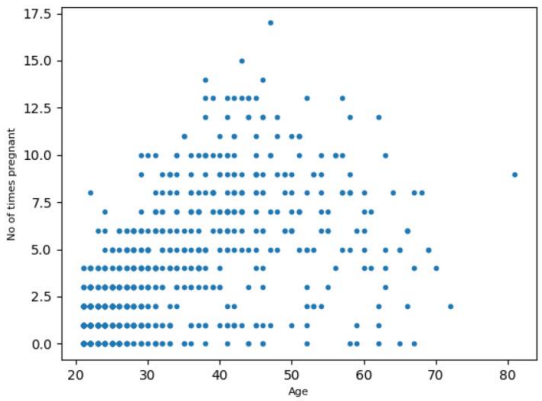


Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. As with increase in age, there is a increase in pregs. but not in whole dataset as for high values of Age, dataset is spreading, that's why both are moderately correlated.
2. Density of plot is high for lower values but decreases with high values of age.

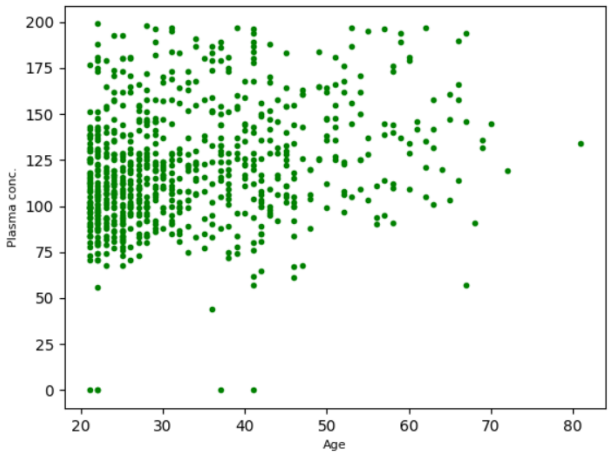


Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. Since the spread of data is so high and random, there is very low correlation between both the attributes or can say negligibly correlated.
2. The density of data points is larger for younger age groups, but it decreases with age. However, we can observe that the values of plasma concentration are also high for older age groups, implying that both are weakly correlated.

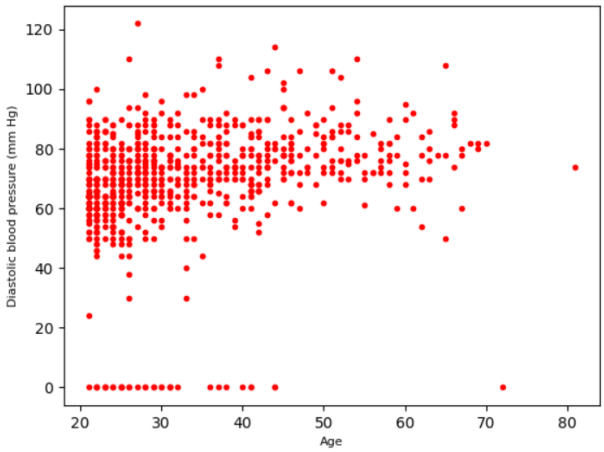


Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. As with the increase in age, there appears to be no influence on diastolic pressure values, which remain concentrated around 70 mm-Hg, implying that both qualities are negligibly correlated or weakly correlated.

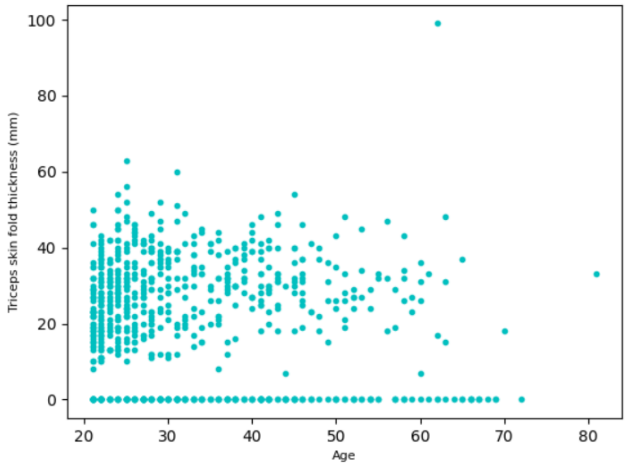


Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. Here, with increase in value of age, there appears to be no effect on skin fond thickness which remain concentrated around 30mm.
2. Density of plot is high for younger age group but decreases with increase in age and for high values of age, concentration of data points in higher around 0 which implies both attributes are weakly negatively correlated.

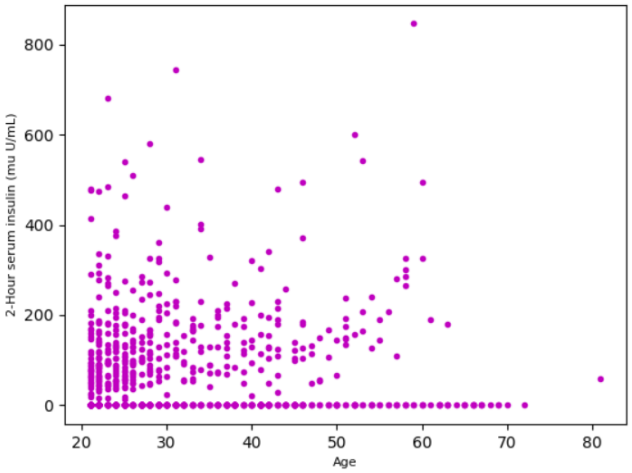


Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. With increase in value of age, there is no specific change (increase or decrease) in values of test (mm U/mL). So, both are not correlated.
2. Density is high for lower age group which means large number of younger women have insulin level less than 200 mm U/ml.

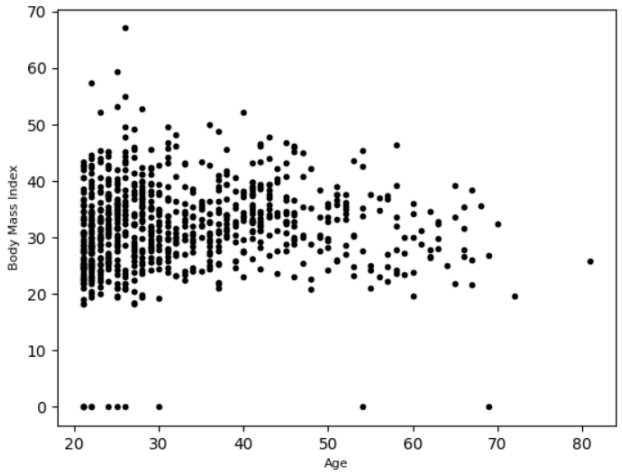


Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. With increase in values of age, there is no specific effect on values of BMI which remains concentrated around 30 kg/m. So, both the attributes are not correlated or have zero value of correlation coefficient.
2. Density decreases with increase in value of Age.

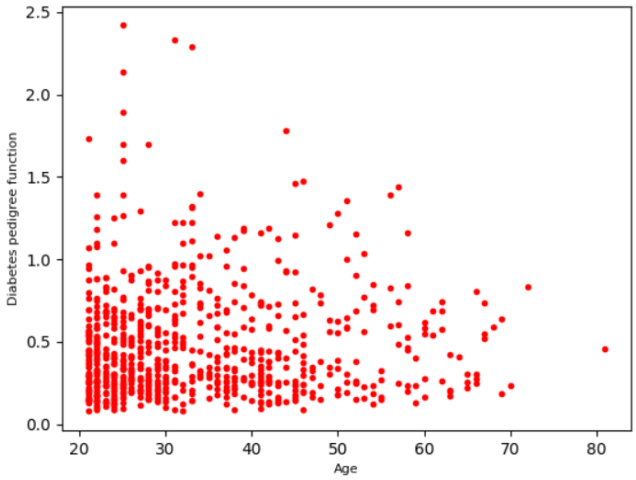


Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. With increase in values of age, there is no specific effect on values of pedi which remain concentrated around 0.3. So, both are not correlated.
2. Density of data point in decreasing with increase in age which implies that data set has large number of young women.

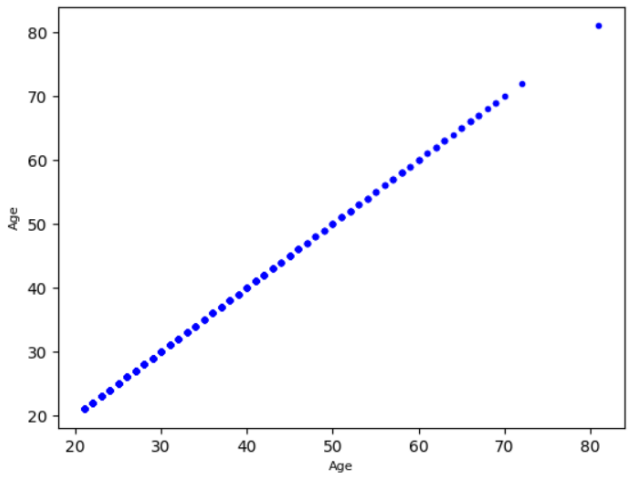


Figure 8 Scatter plot: Age (in years) vs. Age (in years)

Inferences:

1. Since there is same set of data on x-axis and y-axis, plot is highly dense and both attributes (age and age) are strongly correlated with correlation coefficient equals to 1.

2 (b).

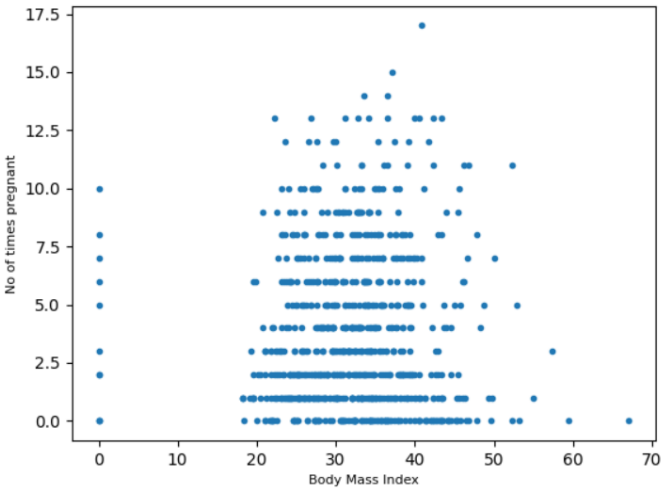


Figure 9 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. With increase in value of pregs., there is no specific change in BMI and the value of BMI seems to be remain between 20 to 40 irrespective of value of pregs. Thus both are not correlated.
2. The density of plot shows that practically all of the women had a BMI of 20 to 40 kg/m^2.

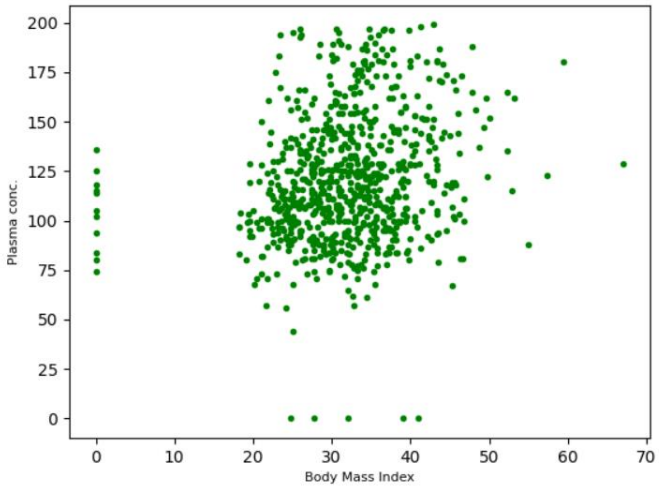


Figure 10 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. Both attributes shows weak correlation because data is highly concentrated but overall we can see that value of plasma conc. is increasing with age.
2. Density of data set is very high as almost all data points are concentrated around BMI value of 30 and plasma concentration value of 110.

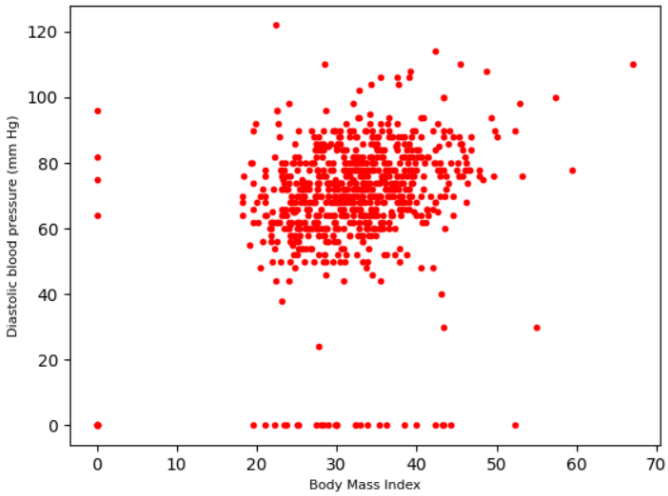


Figure 11 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. Data set has very high density around BMI value of 30 kg/m^2 and 70 mm-Hg diastolic blood pressure.
2. Overall, we can see that the value of diastolic blood pressure is higher for older age group than young age group thus both attributes are weakly correlated.

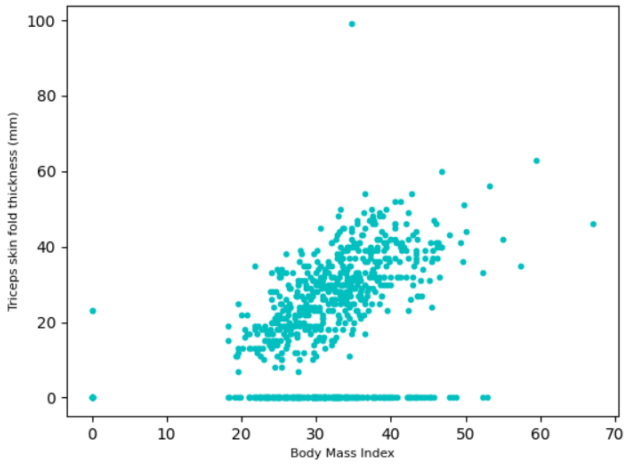


Figure 12 Scatter plot: BMI (in kg/m²) vs. skin (in mm)

Inferences:

1. Here, with increase in value of BMI, there is increase in the value of skin, but many value for skin remains low (around zero), thus both are moderately positively correlated.
2. Density of data points seems to be constant for whole data.

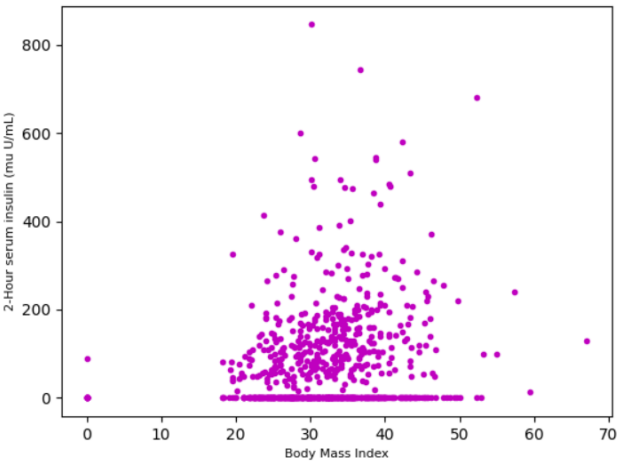


Figure 13 Scatter plot: BMI (in kg/m²) vs. test (in mm U/mL)

Inferences:

1. Both attributes have very low or negligible correlation due to the randomness of the data points.
2. Density of data points is higher for low values of insulin and very low for high value of insulin.

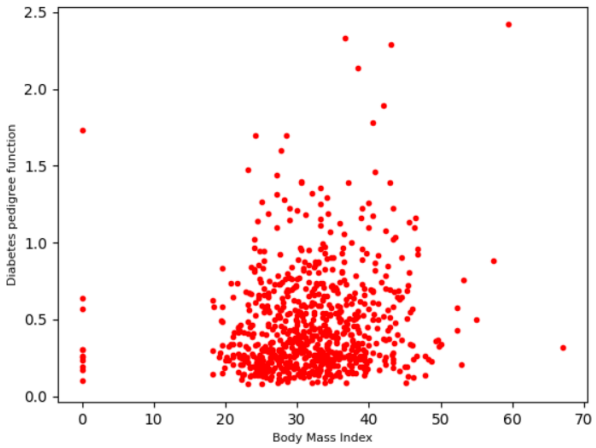


Figure 14 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. Data set is concentrated around BMI value of 30 kg/m^2 and pedi value of 0.2 and shows no specific relation between both attributes so both are weakly correlated.
2. Density of plot seems to be high for value if BMI around 30 and data spreads with increase in value of BMI to 40 kg/m^2.

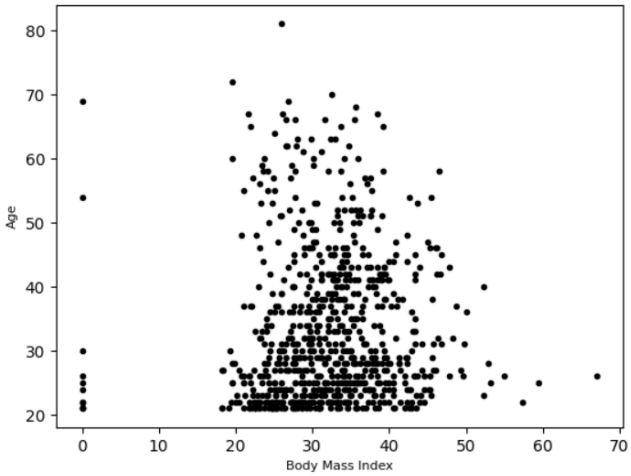


Figure 15 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. Both attributes are not correlated due to the randomness of the plot.

3(a).

Table 3 Correlation coefficient value computed between age and all other attributes

S. No.	Attributes	Correlation Coefficient Value(q)
1	pregs	0.5443
2	plas	0.2635
3	pres (in mm Hg)	0.2395
4	skin (in mm)	-0.1139
5	test (in mu U/mL)	-0.0421
6	BMI (in kg/m^2)	0.0362
7	pedi	0.0335
8	Age (in years)	1.0000

Inferences:

1. With increase in magnitude of correlation coefficient from 0 to 1, degree of correlation between age and other attributes increases.
2. For positive value of correlation coefficient, with increase in age other attributes also increases but for negative value of correlation coefficient, with increase in age other attribute's value decreases.
3. For pregs, $q \approx 0.5$ → moderately correlated
   For plas, $q \approx 0.2$ → weakly correlated
   For pres, $q \approx 0.2$ → weakly correlated
   For skin, $q \approx -0.1$ → negligible negative correlation
   For test, $q \approx -0.04$ → no correlation
   For BMI, $q \approx 0.03$ → no correlation
   For pedi, $q \approx 0.03$ → no correlation
   For Age, $q = 1$ → very strongly correlated

3(b).

Table 4 Correlation coefficient value computed between BMI and all other attributes

S. No.	Attributes	Correlation Coefficient Value
1	pregs	0.0176
2	plas	0.2210
3	pres (in mm Hg)	0.2818
4	skin (in mm)	0.3925
5	test (in mu U/mL)	0.1978
6	BMI (in kg/m^2)	1.0000
7	pedi	0.1406
8	Age (in years)	0.0362

Inferences:

1. With increase in magnitude of correlation coefficient from 0 to 1, degree of correlation between BMI and other attributes increases.
2. For positive value of correlation coefficient, with increase in BMI other attributes also increases but for negative value of correlation coefficient, with increase in BMI other attribute's value decreases.
3. For pregs, q≈0.01 → no correlation
   For plas, q≈0.22 → weakly correlated
   For pres, q≈0.28 → weakly correlated
   For skin, q≈ 0.4 → low correlation
   For test, q≈ 0.2 → weakly correlated
   For BMI, q= 1 → very strongly correlation
   For pedi, q≈ 0.1 → negligibly correlated
   For Age, q=0.03 → no correlation

4.

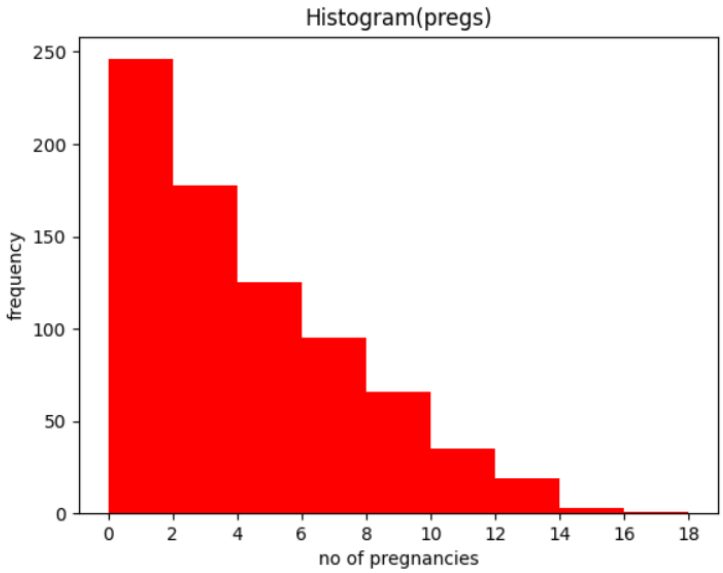


Figure 16 Histogram depiction of attribute pregs

Inferences:

1. Height of each bin decreases with increase in number of pregnancies which means as the number of pregnancies increases, women count decreases. More women had less pregnancies.
2. Modal bin in this plot is bin (0,2) having highest frequency thus mode of pregs lie in (0,2).

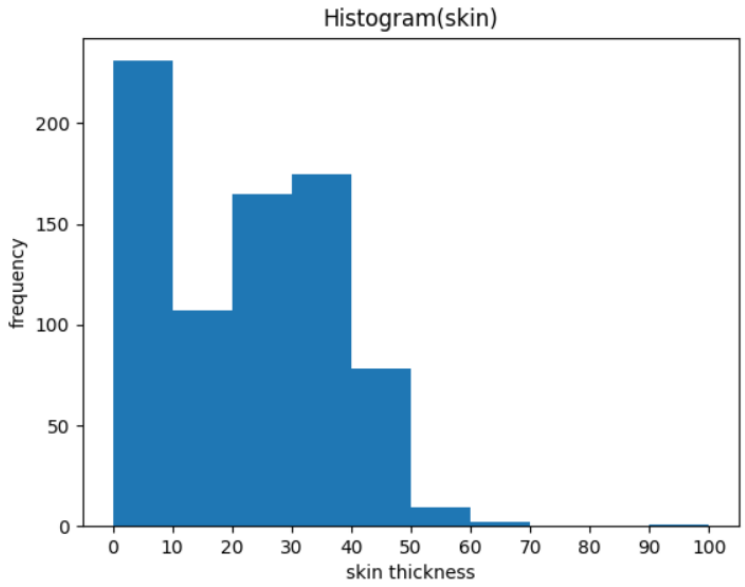


Figure 17 Histogram depiction of attribute skin

Inferences:

1. For most of the women in data set, triceps skin fold thickness is less than 40mm.
2. Modal bin in this plot is bin (0,10) having highest frequency thus mode of attribute skin lie in (0,10).

5.

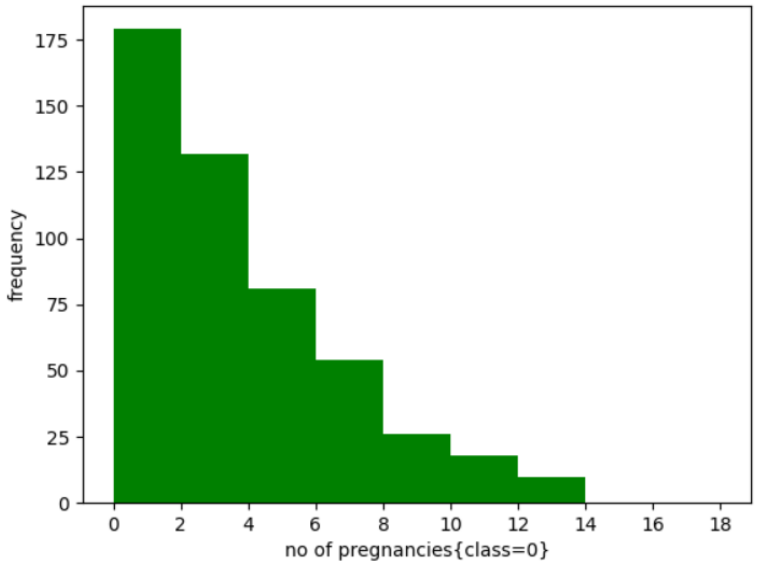


Figure 17 Histogram depiction of attribute pregs for class 0

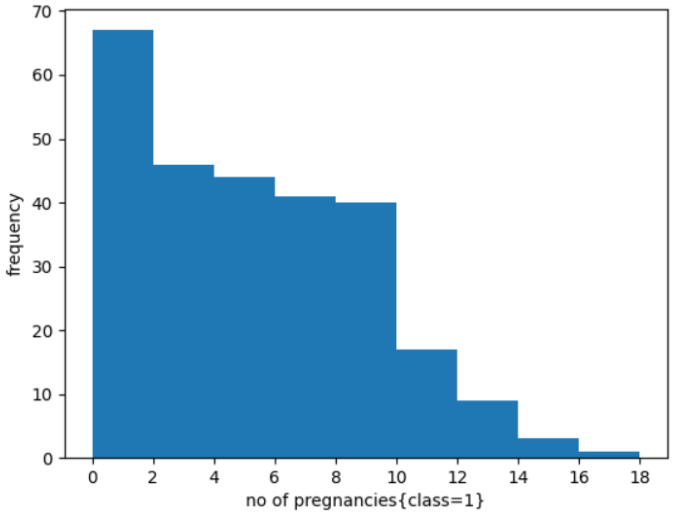


Figure 17 Histogram depiction of attribute pregs for class 1

Inferences:

1. For both class 0 and class 1, mode of attribute pregs lie in (0,2) bin.
2. Height of each bin is count of women, which decreases with increase in the number of pregnancies for both class 0 and 1.
3. For class 0, we can conclude that very few women have pregnancy more than 8.
4. For class 1, very few women have pregnancy more than 10.

6.

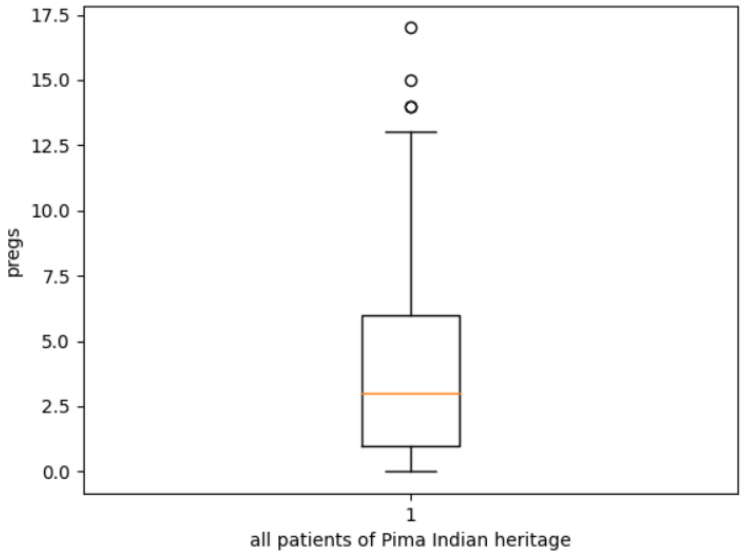


Figure 19 Boxplot for attribute pregs

Inferences:

1. Outliers are present only above the upper whisker [upper Quartile + (1.5*IQR)] that is 6+1.5*5= 13.5.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 6 -1 =5.
3. The attribute pregs varies from 0 to 13.5 with some values higher than 13.5.
4. From plot, we can observe that data set is positively skewed.
5. From plot we can see the value of median, maximum and minimum is 3, 17, 0 respectively which is equal to values in Q1.

Inferences:

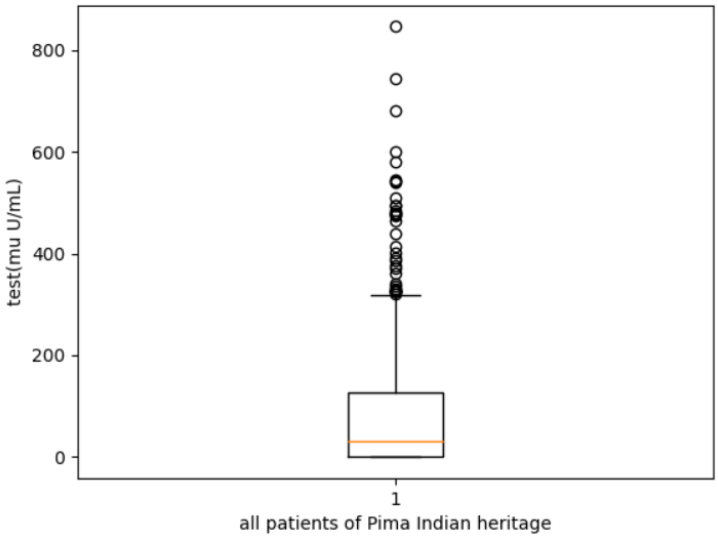
1. Outliers are present only below the lower whisker [lower Quartile - (1.5*IQR)] that is 99.0 - 1.5*41.25 = 78.375.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile– lower quartile. Here, IQR = 140.45 -99 = 41.25.
3. The attribute plas varies from 78.375(lower whisker) to 199(maximum) with some values less than lower whisker.
4. From plot, we can observe that data set is not skewed.
5. From plot we can see the value of median, maximum and minimum is 117, 199, 0 respectively which is equal to values in Q1.

Inferences:

1. Outliers are present below the lower whisker [lower Quartile - (1.5*IQR)] that is 62 - 1.5*18 = 35.0 and also above the upper whisker [upper Quartile + (1.5*IQR)] that is 80 + 1.5*18 = 107.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 80 - 62 = 18.
3. The attribute pres varies from 35(lower whisker) to 107(upper whisker) with some values less than lower whisker and more than upper whisker.
4. From plot, we can observe that data set is not skewed.
5. From plot we can see the value of median, maximum and minimum is 72, 122, 0 respectively which is equal to values in Q1.

Inferences:

1. Outliers are present only above the upper whisker [upper Quartile + (1.5*IQR)] that is 32.0 + 1.5*32 = 80.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 32 -0 = 32.
3. The attribute skin varies from 0(lower whisker) to 99(maximum) with some values greater than upper whisker.
4. From plot, we can observe that data set is negatively skewed.
5. From plot we can see the value of median, maximum and minimum is 23, 99, 0 respectively which is equal to values in Q1.

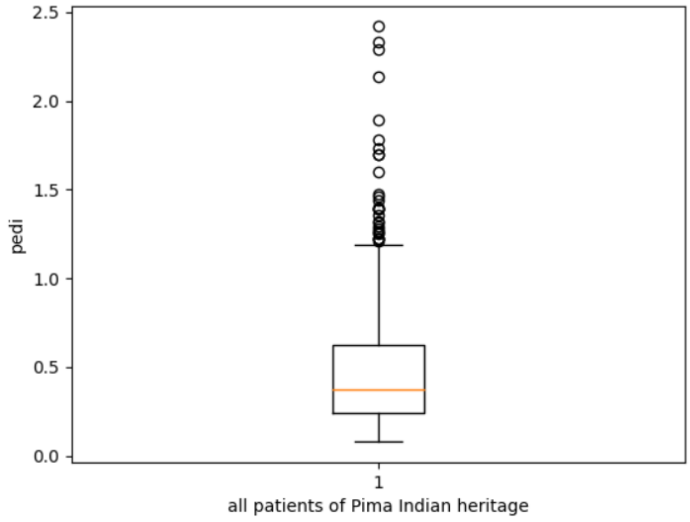


Inferences:

1. Outliers are present only above the upper whisker [upper Quartile + (1.5*IQR)] that is 127.25 + 1.5*127.25 = 318.125.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 127.25 - 0 = 127.25.
3. The attribute test varies from 0(minimum) to 318.125(upper whisker) with some values larger than upper whisker.
4. From plot, we can observe that data set is positively skewed.
5. From plot we can see the value of median, maximum and minimum is 30.5, 846, 0 respectively which is equal to values in Q1.

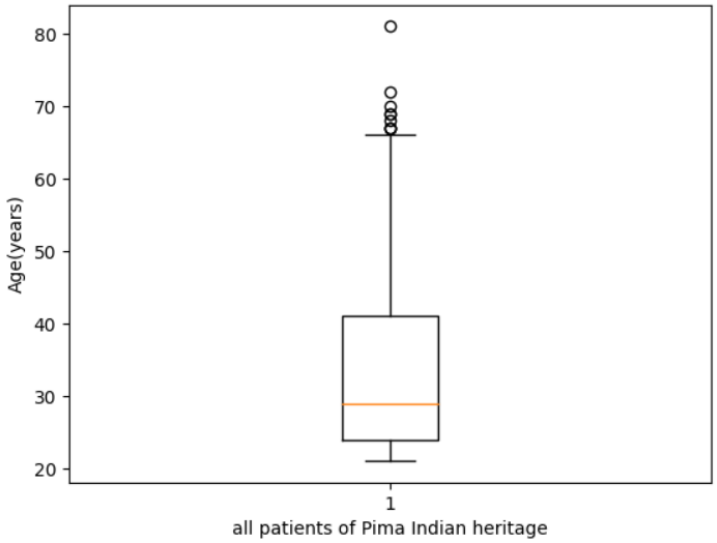
Inferences:

1. Outliers are present below the lower whisker [lower Quartile - (1.5*IQR)] that is 27.3 - 1.5*9.3 = 13.35 and also above the upper whisker [upper Quartile + (1.5*IQR)] that is 36.6 + 1.5*9.3 = 50.55.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 36.6 – 27.3 = 9.3.
3. The attribute BMI varies from 0(lower whisker) to 50.55(upper whisker) with some values less than lower whisker and more than upper whisker.
4. From plot, we can observe that data set is not skewed.
5. From plot we can see the value of median, maximum and minimum is 32, 67.1, 0 respectively which is equal to values in Q1.



Inferences:

1. Outliers are present only above the upper whisker [upper Quartile + (1.5*IQR)] that is 0.62625 +1.5*0.3825 = 1.2.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 0.62625 -0.24375=0.3825.
3. The attribute pedi varies from 0(minimum) to 1.2(upper whisker) with some values larger than upper whisker.
4. From plot, we can observe that data set is positively skewed.
5. From plot we can see the value of median, maximum and minimum is 0.3725, 2.42, 0.078 respectively which is equal to values in Q1.



Inferences:

1. Outliers are present only above the upper whisker [upper Quartile + (1.5*IQR)] that is 41+1.5*17 = 66.5.
2. Inter quartile range represent the middle 50% values of the data set. IQR = upper quartile – lower quartile. Here, IQR = 41 - 24= 17.
3. The attribute Age varies from 0(minimum) to 81(maximum) with some values larger than upper whisker.
4. From plot, we can observe that data set is positively skewed.
5. From plot we can see the value of median, maximum and minimum is 29, 81, 21 respectively which is equal to values in Q1.