Indian
Institute of
Technology
Mandi

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

Student's Name: Khushi Ladha	**Mobile No: 7665519043**
Roll Number: B20013	**Branch:Bioengineering**

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	no. of times pregnant	3.845052	3	1.0	0	17	3.369578
2	plas	120.894531	117	99	0	199	31.972618
3	pres (in mm Hg)	69.105469	72	70	0	122	19.355807
4	skin (in mm)	20.536458	23	0	0	99	15.952218
5	test (in mu U/mL)	79.799479	30.5	0	0	846	115.244
6	BMI (in kg/m^2)	31.992578	32	32	0	67.1	7.88416
7	pedi	0.471876	0.3725	0.254 , 0.258	0.078	2.42	0.331329
8	Age (in years)	33.240885	29	22	21	81	11.760232

Inferences:

1. If standard deviation is close to zero; mean, median and mode are close to each other. This can be observed from Pedi attribute where SD is 0.34 and hence mean, mode median are close to each other. This is because SD shows spread, and if spread is less, then the central tendencies will lie closer to each other.
2. If mean, mode and median are close to each other, the distribution curve will be less skewed and shift towards being more symmetrically distributed.
3. Higher the range i.e. higher the gap between minimum and maximum value, higher is the spread generally and higher is the SD generally. This can be highlighted in attribute –test, plas, no. of times pregnant etc.

2 a.

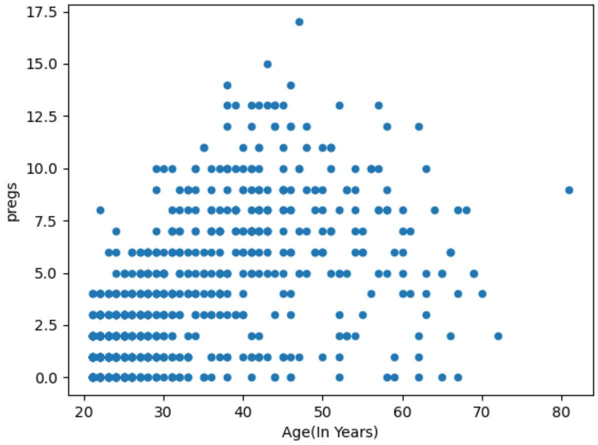


Figure 1 Scatter plot: Age (in years) vs. no. of times pregnant

Inferences:

1. A **weak positive correlation** can be seen between Age and No of times no. of times pregnant .
2. However, There seems to be **no strong association/relationship** between No of times no. of times pregnant and Age as the data points are scattered.
3. There is a slightly higher density of clustering at the lower left corner of plot. This means there is **more density between age group 20 – 35 years** and **No of times pregnant 0 to 6**. So, it can be inferred that in the sample, a large amount of females are from age group 20-35 and these have been no. of times pregnant 0 to 6 times.
4. **2 distinct outliers** can be identified at - 10 no. of times pregnancies till age 80 and highest number of times no. of times pregnant value – 17.

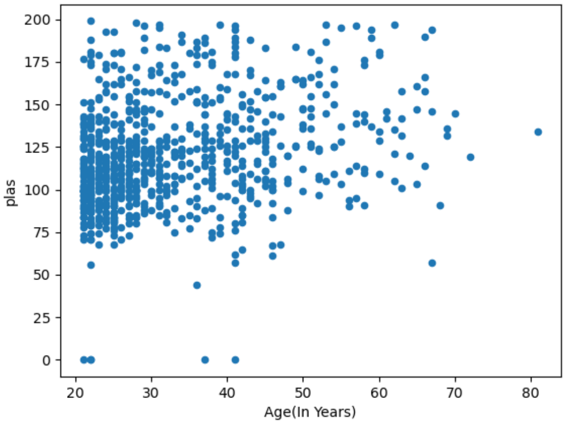


Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. There is a **very weak positive correlation** between Age and plasma glucose levels.
2. There is a higher density and clustering at the middle- upper left section of plot. This means that a good chunk of people in the sample are from **age group 20-35** and have **plasma glucose values between 75 to 150.**
3. Few outliers with nearly 0 plasma glucose values or old age(nearly 80) can be identified.
4. Except the few outliers, **most females have plasma glucose levels values above 60**.
5. Plasma glucose levels here refer to Plasma glucose concentration 2 hours in an oral glucose tolerance test.

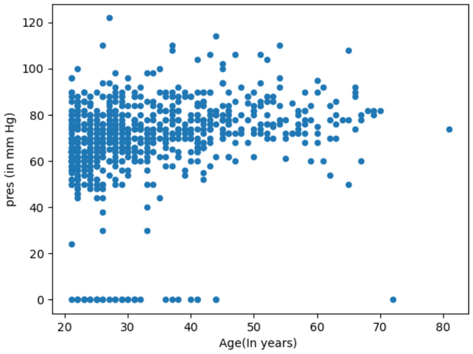


Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. A very weak positive correlation can be seen between Age and Blood Pressure.
2. However, the correlation appears to be so weak, the trace of data points looks nearly constant i.e. no visible change in blood pressure with increasing age.
3. Blood pressure values are concentrated between 50 – 90 mm Hg.
4. Major part of age group appears to be lying between 20-30(in years) ages, followed by 30-40 years of age, based on density of data points.
5. However, few outliers can be seen in all age groups with high blood pressure (hypertension) or blood pressure (hypotension).

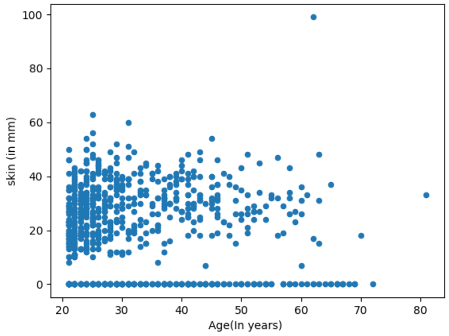


Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. There is nearly no association or a very weak correlation between skin fold thickness and age.
2. Again, there is a higher density of data points in the lower left part of plot. This is because of larger chunk of participants having age group 20-35(in years).
3. In the 20-35 (in years) age groups, skin folds are seen to be between 10-45 mm. Since, skin folds lie in such a large range, there appears to be very little relation between age and skin folds.
4. As age increase, density is decreasing and data points appear to be more scattered.

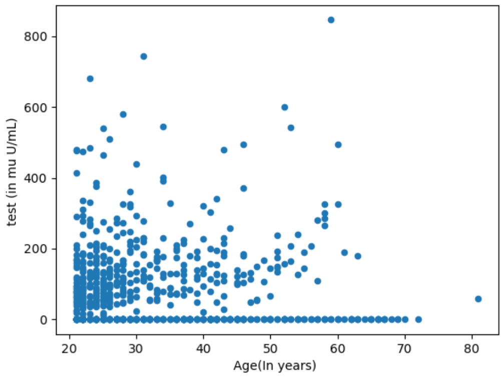


Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. A moderately dense cluster can be identified when age is between 20-30 years and insulin test values are between 0 – 200 mu U/mL. This tells that for females of age 20-30(in years), insulin values are usually between 0 -200 mu U/mL.
2. As age increases, there seems to be a slightly upward non linear trend in insulin values.
3. Quiet a number of outliers are present in all age group, indicating that all age groups have some females which have higher than average insulin levels.

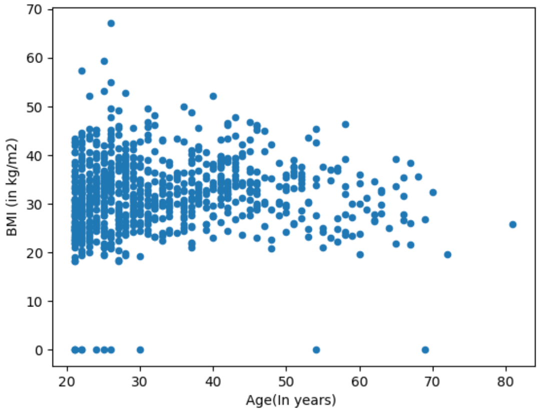


Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. There is no correlation between age and BMI.
2. Density of data points decreases as age increases. This means that most of the females are in age group 20-35(in years).
3. BMI values are in between 20-45(in kg/m^2) for most women.
4. Since most BMI values are greater than 20 (in kg/m^2), so very female in the sample are underweight, however nearly half of the women have BMI above 30(in years), which means half of them are overweight.
5. BMI values are spread evenly for age groups, so **mean BMI** can turn out to be a **nearly constant value.**

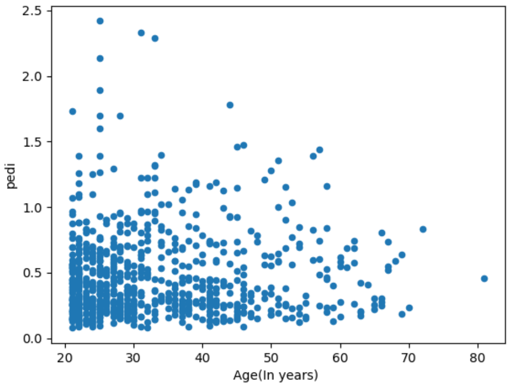


Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. There is no strong correlation between Age and Pedigree function value.
2. Higher density of data points is located for age group between 20-30 years and diabetes pedigree function value between 0 to 0.75.
3. Outliers indicated that some females have a very high pedigree value even in the age group of 20-40 (in years), so these can be warned at earlier times before the onset of disease. (Note: Pedigree function value tells about the risk of developing type 2 diabetes. So, higher the value of function, higher the risk.)

b.

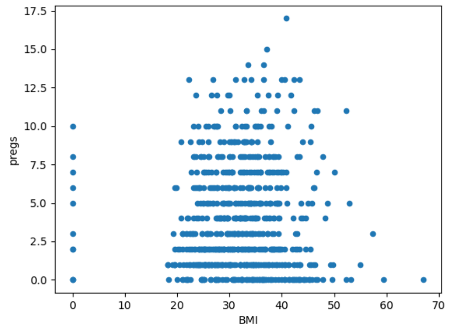


Figure 8 Scatter plot: BMI (in kg/m^2) vs. no. of times pregnant

Inferences:

1. There is no correlation/association between BMI value and no of times women has been no. of times pregnant value.
2. The points are not continuous in the vertical trace. This is because no of time no. of times pregnant value has to take only integer values.
3. Outliers with 0 BMI are present which is due to unclean/ noisy data.
4. The number of times no. of times pregnant value appears to be having most density in between range 1-5.

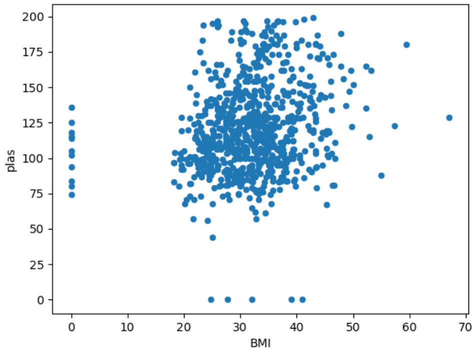


Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. A weak positive correlation can be observed between BMI and plasma glucose concentration.
2. A denser cluster can be observed for BMI 20-35 (in kg/m^2) and plasma glucose concentration values 75-150.

3. The outliers with 0 values of BMI or plasma glucose concentration could be due to noisy data.

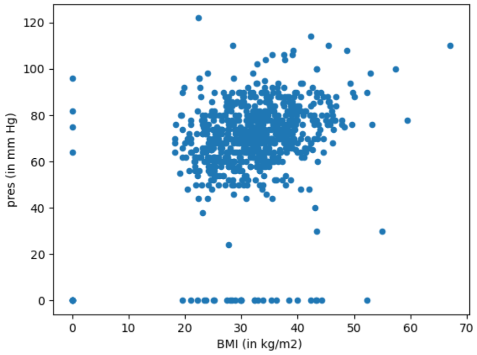


Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. A weak positive correlation can be observed between BMI and blood pressure.
2. The data points are dense around BMI values 20-40 (in kg/m^2) and pressure values 50-90 mm Hg.
3. Few outliers are present in BMI range 20-60 kg/m^2, having both high and low blood pressure.

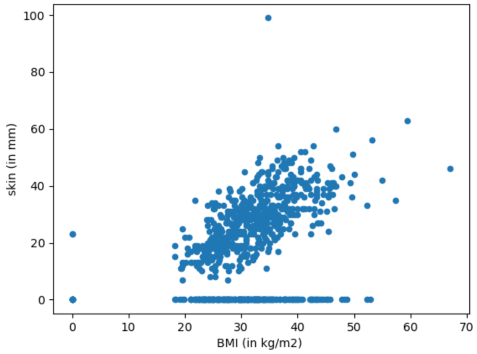


Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. BMI is moderately positively correlated to skin. This means as the BMI increases, body fat percentage increases, so the skin fold thickness value increases .
2. The trace of curve looks increasing and linear.

3. An outlier cluster of horizontally arranged data points at the bottom can be due to noisy data/incorrent value entered because skin fold value can't be 0.
4. Skin fold values range from 10-50 mm and all values are nearly equally distributed. A histogram of skin fold values may have bars of nearly same height.

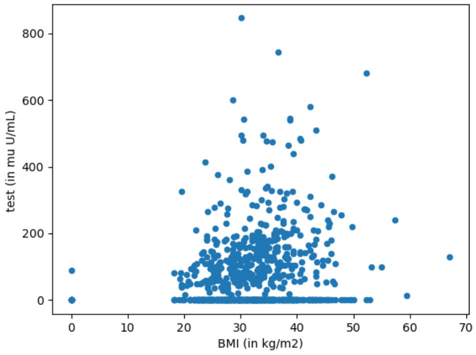


Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. The 2-hour insulin test values have very weak positive or nearly no correlation with BMI values.
2. Most of the test values are concentrated between 0 – 200 mm U/mL .
3. Outliers are present with moderate to high test values for all BMI ranges.

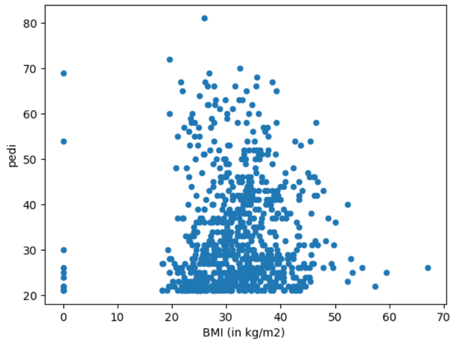


Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. There is nearly no correlation between BMI and pedi. This means that BMI cant be used as an attribute to tell whether a person is more likely to get diabetes based on family history.
2. Data points are concentrated between BMI values 20 – 40 kg/m^2 and pedigree function value 20 – 40.
3. However, outliers are present at even higher pedigree and BMI values.

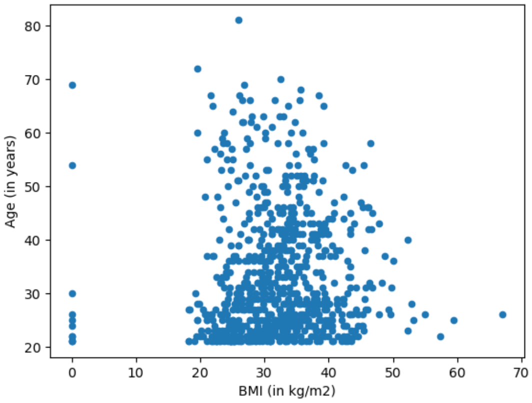


Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. There is no correlation between BMI and age.
2. Denser cluster is present between BMI values 20-40 kg/m^2 and age 20-40.
3. Few outliers with high age or young age but higher BMI are present.
4. Density of data points decreases as age increases. This means that most of the females are in age group 20-35.

3 a.

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

S. No.	Attributes	Correlation Coefficient Value

1	no. of times pregnant	0.544341
2	plas	0.263514
3	pres (in mm Hg)	0.239528
4	skin (in mm)	-0.11397
5	test (in mu U/mL)	-0.04216
6	BMI (in kg/m^2)	0.036242
7	pedi	0.033561
8	Age (in years)	1

Inferences:

1. Correlation between Age and

1.	no. of times pregnant	Moderate	Positive
2.	plas	weak	Positive
3.	pres (in mm Hg)	weak	Positive
4.	skin (in mm)	Very Weak	Negative
5.	test (in mu U/mL)	Very Weak	Negative
6.	BMI (in kg/m^2)	Very Weak	Positive
7.	pedi	Very weak	Positive
8.	Age (in years)	High	Positive

2. In attributes having positive correlation values, with increase or decrease in age each of the attributes will increase or decrease. While in attributes having negative correlation values, with increase in age all other attributes would decrease and with decrease of age, other attributes would increase.
3. The correlation coefficient value for most of the attributes is either weak or very weak. Even on the scatter plots, similar observation was mode that most attributes had feeble association with other attributes. However, for no. of times pregnant attribute the correlation value is moderate but in scatter plot a weak correlation seemed to exist.
4. Test and skin are the only attributes having negative correlation value.

b.

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

S. No.	Attributes	Correlation Coefficient Value
1	no. of times pregnant	0.017683
2	plas	0.221071
3	pres (in mm Hg)	0.281805
4	skin (in mm)	0.392573
5	test (in mu U/mL)	0.197859
6	BMI (in kg/m2)	1
7	pedi	0.140647
8	Age (in years)	0.036242

Inferences:

1. Correlation between Age and

1.	no. of times pregnant	Very weak	Positive
2.	plas	weak	Positive
3.	pres (in mm Hg)	weak	Positive
4.	skin (in mm)	Moderate	Positive
5.	test (in mu U/mL)	Weak	Negative
6.	BMI (in kg/m^2)	HIgh	Positive
7.	pedi	Weak	Positive
8.	Age (in years)	Very Weak	Positive

2. In attributes having positive correlation values, with increase or decrease in age each of the attributes will increase or decrease. While in attributes having negative correlation values, with increase in age all other attributes would decrease and with decrease of age, other attributes would increase.
3. The correlation coefficient value for most of the attributes is either weak or very weak. Even on the scatter plots, similar observation was mode that most attributes had feeble association with other attributes.
4. Test is the only attribute having negative correlation value.

4 a.

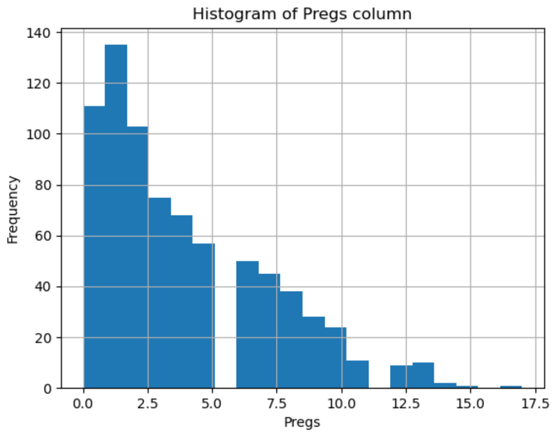


Figure 15 Histogram depiction of attribute no. of times pregnant

Inferences:

Frequency	**Bin size**
111	0 - 0.875
135	0.875 - 1.75
103	1.75 - 2.625
75	2.625 - 3.5
68	3.5 - 4.375
57	4.375 - 5.25
0	5.25 - 6.125
50	6.125 - 7
45	7 - 7.875
38	7.875 - 8.75
28	8.75 - 9.625
24	9.625 - 10.5
11	10.5 - 11.375
0	11.375 - 12.25
9	12.25 - 13.125
10	13.125 - 14
2	14 - 14.875
1	14.875 - 15.75

1. The bars show a decreasing trend with highest no. of times pregnant value in bin 0-1 being 250, lowest no. of times pregnant value frequency in bin 15-17 being 2.
2. Mode of attribute No. of times pregnant lies in bin 0 – 1, with frequency being nearly 250.
3. The histogram looks positively skewed.

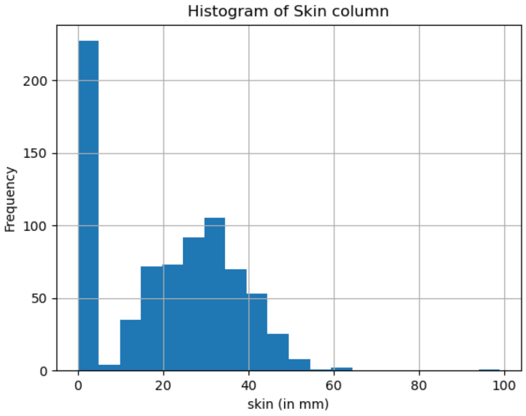


Figure 16 Histogram depiction of attribute skin

Inferences:

Frequency	Bin size
227	0 - 5
4	5 - 10
35	10 - 15
72	15 - 20
73	20 - 25
92	25 - 30
105	30 - 35
70	35 - 40
53	40 - 45
25	45 - 50
8	50 - 55
1	55 - 60
2	60 - 65

0	65	-	70
0	70	-	75
0	75	-	80
0	80	-	85
0	85	-	90

1. The bars show a <u>decreasing, increasing then again decreasing trend.</u>
2. The highest frequency value is in 0-10 bin, followed by 30-40 bins and is nearly 0 in skin values above 60.
3. Mode of attribute Skin lies in bin 0-10, with frequency value being nearly 225.

5

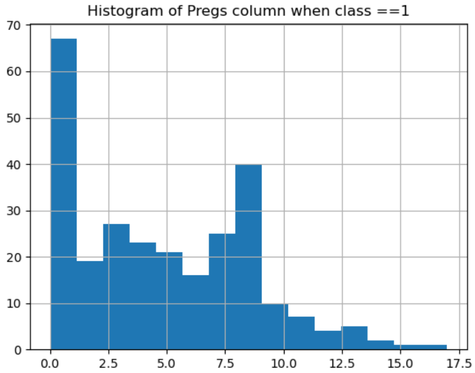
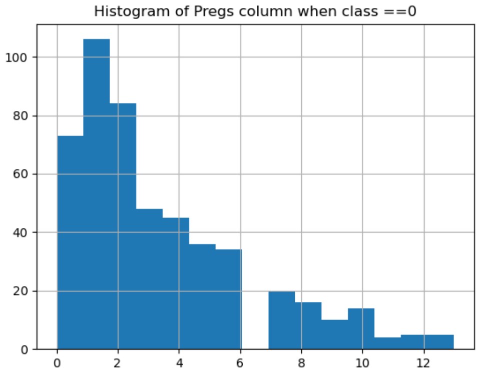


Figure 17 Histogram depiction of attribute no. of times pregnant for class 0

Figure 18 Histogram depiction of attribute no. of times pregnant for class 1



Class = 1				Class = 0			
Frequency	**Bin size**			**Frequency**	**Bin size**		
67	0	-	1.166	73	0	-	1.166
19	1.166	-	2.332	106	1.166	-	2.332
27	2.332	-	3.498	84	2.332	-	3.498
23	3.498	-	4.664	48	3.498	-	4.664
21	4.664	-	5.83	45	4.664	-	5.83
16	5.83	-	6.996	36	5.83	-	6.996
25	6.996	-	8.162	34	6.996	-	8.162
40	8.162	-	9.328	0	8.162	-	9.328
10	9.328	-	10.494	20	9.328	-	10.494
7	10.494	-	11.66	16	10.49	-	11.66
4	11.66	-	12.826	10	11.66	-	12.826
5	12.826	-	13.992	14	12.83	-	13.992
2	13.992	-	15.158	4	13.99	-	15.158
1	15.158	-	16.324	5	15.16	-	16.324
0	16.324	-	17.49	0	16.32	-	17.49

Inferences:

1. For both class 0 and class 1 , mode lies in bin range 0 to 1. The frequency for class 0 is 176 nearly and 68 for class 1.
2. Nearly all bins in class 0 have higher frequency values/height compared to class 1.
3. Overall, it can be seen that both graphs show similar trends in terms of density line trace. Gradually as no of times no. of times pregnant values increases, the frequency decreases for both.
4. Highest frequency in class 1 is 68 in bin range 1-1.3 and lowest is in 15-17.5 , frequency being 2. For class 0, Highest frequency is 176, in bin range 0 - 1 and lowest is in 10-12 , frequency being 5.
5. Both distributions <u>look positively skewed.</u>

6

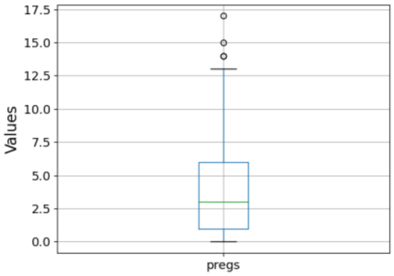


Figure 19 Boxplot for attribute no. of times pregnant

Inferences:

1. An outlier is an observation that is numerically distant from the rest of the data. 3 outliers are present, all above the top whisker. This means that 3 high values of no of time no. of times pregnant and is present in the dataset. No outlier below bottom whisker is present because it is obvious that no. of times pregnant can't take a value below 0.
2. IQR = Q3 – Q1 = 6 -1 = 5.
3. Range = 17, IQR = 5 , 3 outliers , this implies data has moderate variability.
4. The data is positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the outliers, the much higher value of top whisker as compared to lower whisker which didn't even reach 0.5IQR. Also, the fact that median lies closer to Q1 shows positive skewdness
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

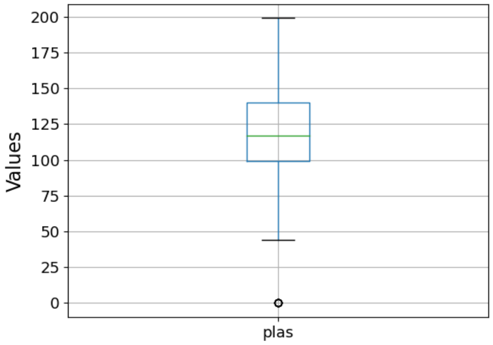


Figure 20 Boxplot for attribute plas

Inferences:

1. Just one outlier is present at the bottom, with plas value = 0 . This might be due to incorrect/noisy data present because otherwise plasma concentration can go to 0.
2. IQR = Q3 – Q1 = 140-100 = 40
3. Range =155, IQR = 40, 1 outlier, this implies data has moderate variability.
4. The data is slightly positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the outliers, higher value of top whisker as compared to lower whisker and the fact that median lies closer to Q1.
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

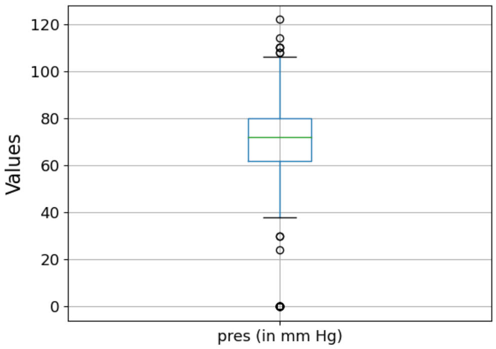


Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. Outliers are present beyond both top and lower whisker. At the top , the values are :- 110, 112, 118 and 122 nearly. At bottom , 31, 25 and 0.
2. IQR = Q3 – Q1 = 82 – 60 = 22
3. Range = Maximum – Minimum = 150 , IQR = 22 , 7 outliers , this implies data has high variability.
4. The data is slightly negatively skewed, i.e. mean is lesser than median. This negative skewedness can be observed by the outliers, the larger spread of top whisker as compared to lower whisker and the fact that median lies closer to Q3.
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

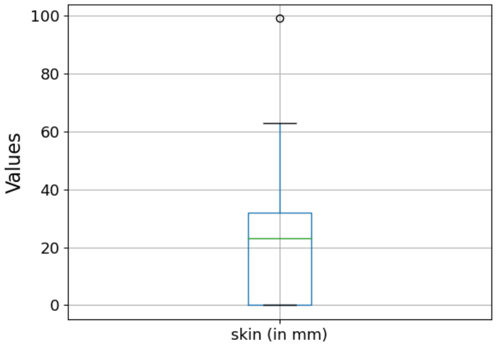


Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. One outlier is present with skin thickness value = 100 mm.
2. IQR = Q3 – Q1 = 32 – 0 = 32
3. Range = 63 , IQR = 32 , 1 outlier , this implies data has moderate variability.
4. The data is highly positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the outlier, the much higher value and spread of top whisker as compared to lower whisker and the fact that median lies closer to Q1..
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

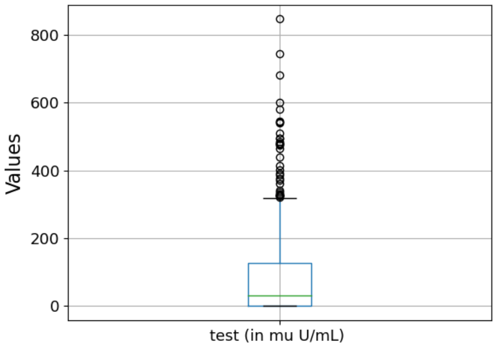


Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. Many outliers are present above the top whisker with maximum of the having values 343 to 542 , highest value being 840.
2. IQR = Q3 – Q1 = 127 – 0 = 127
3. Range = 363 , IQR = 127 many outliers, this implies data has high variability.
4. The data is highly positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the large number of outliers, the much higher value and spread of top whisker as compared to lower whisker and the fact that median lies closer to Q1.
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

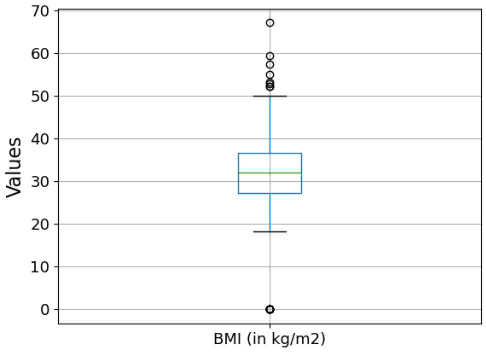


Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. Larger number of outliers is present above the top whisker than the lower one.
2. IQR = Q3 – Q1 = 36.6 – 27.3 = 9.3
3. Range = 23 , IQR = 9.3 , this implies data has moderate variability.
4. The data is slightly negatively skewed,, i.e. mean is lesser than median. This negative skewedness can be observed by the outliers, the larger spread of top whisker as compared to lower whisker and the fact that median lies closer to Q3.
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

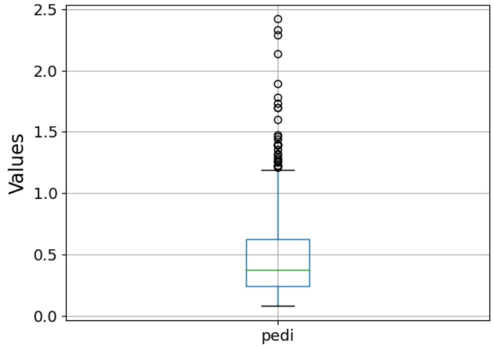


Figure 25 Boxplot for attribute pedi

Inferences:

1. Many Outliers are present above the top whisker with highest value being 2.5 and a denser cluster of outlier lying between 1.2 to 1.5
2. IQR = Q3 – Q1 = 0.62 – 0.24 = 0.38
3. Range =1.2 – 0.8 = 0.4, IQR = 0.38, large number of outliers, this implies data has weak - moderate variability.
4. The data is positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the outliers, larger spread of top whisker as compared to lower whisker and the fact that median lies closer to Q1.
5. The median, minimum and maximum value can be observed to be same as that obtained in question 1.

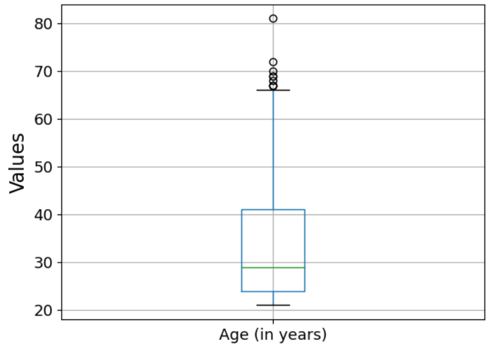


Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. Many Outliers are present above the top whisker with highest value being 82 and most other outliers lying between 66 – 72 .
2. IQR = Q3 – Q1 = 41 – 24 = 17.
3. Range =66 – 24 = 42, IQR = 17 , 7-8 outliers , this implies data has moderate variability.
4. The data is positively skewed, i.e. mean is greater than median. This positive skewedness can be observed by the outliers, larger spread of top whisker as compared to lower whisker and the fact that median lies closer to Q1.
5. The median, minimum and maximum value can be observed to be same as that obtained in question.

Other Observations:-

Some attributes like BMI, skin fold in mm, blood pressure etc had value = 0. However, in reality this is not possible. Some extreme high or low values or 0 values could have come due to incorrect data entered/ noisy data.