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

Student's Name: Sarthak Saptami Kumar Jha **Mobile No: 8825319259**

Roll Number: B20317 **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	3.84	3.00	1	0	17	3.36
2	plas	120.89	117.00	99, 100	0	199	31.97
3	pres (in mm Hg)	69.10	72.00	70	0	122	19.35
4	skin (in mm)	20.53	23.00	0	0	99	15.95
5	test (in mu U/mL)	79.79	30.50	0	0	846	115.24
6	BMI (in kg/m^2)	31.99	32.00	32	0.078	67.10	7.884
7	pedi	0.47	0.37	0.254,0.258	21	2.42	0.33
8	Age (in years)	33.24	29.00	22	0	81.00	11.76

Inferences:

1. We can clearly see from the data that whenever S.D is close to 0 which is in the case of pedi and pregs, the mean, median and mode are very close to each other and in the case of test, where S.D is 115, the mean, median and mode are well separated.
2. Most of the population has a single pregnancy, with 3.8 being the mean.
3. We can also infer that the population is young in age, with mean, median and mode being around 25.

2 a.

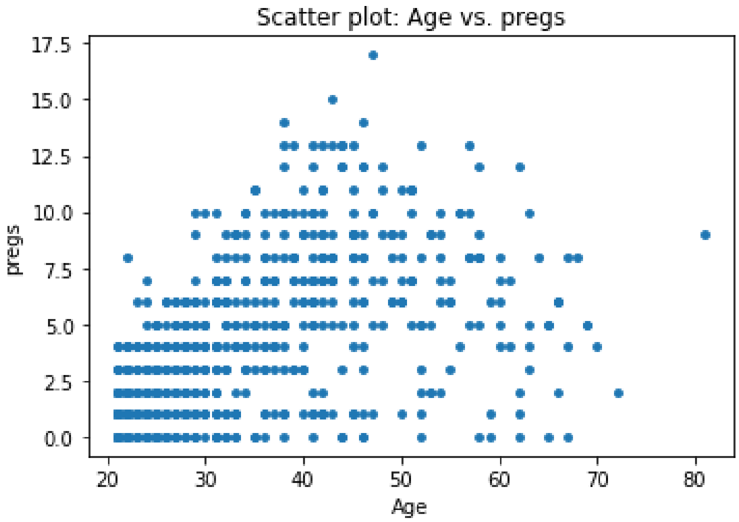

Figure 1 Scatter plot: Age (in years) vs. pregs

Inferences:

1. There is a moderate, positive correlation in linear manner.
2. The points are highly dense at the origin, inferring to the fact that most women in 20-30 age range have 0 – 2 children.
3. There are very few women of higher age ranges who have less than 2 children.
4. Most of the above 10 pregnancies, were in the ages of 40 – 55.
5. The number of women having 0 – 2 children is way less in above 35 age range compared to below 35.

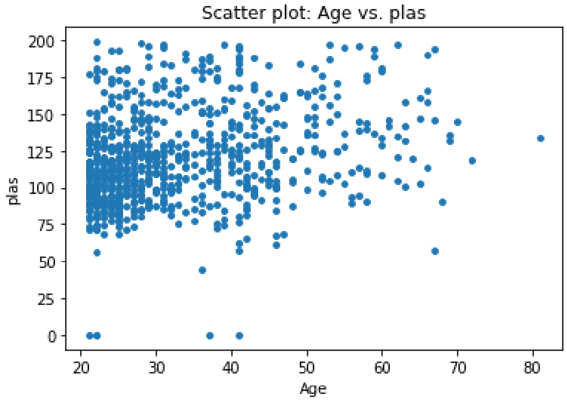

Figure 2 Scatter plot: Age (in years) vs. plas

Inferences:

1. Plasma glucose concentration is positively, but weakly correlated in a linear manner with age.
2. We see that high density of points lies in low age ranges from 20-30 with glucose plasma concentration range around 75 – 150 which can be considered the normal plasma glucose level range.
3. The weak positive correlation corresponds to the fact that with increasing age, plasma glucose concentration increases.
4. There are a few points with 0 data value, and can be attributed as missing data point.

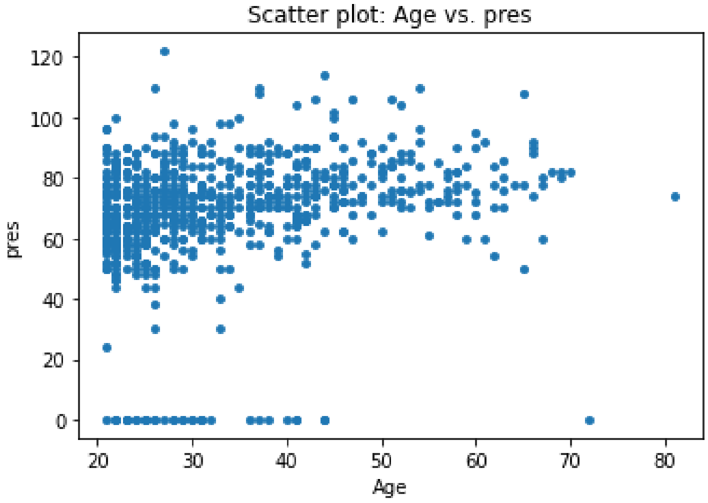

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

Inferences:

1. We can see a weak positive correlation of linear nature.
2. There is a dense concentration between 60 – 80 mmHg of pressure for less than 40 years of age, after which, number of outliers start increasing.
3. Majority of data is within the limits of 55 – 90 mmHg which can be considered normal levels, and there is very minor presence of extreme variations in data.
4. There are a large number of points with 0 data value, and can be attributed to missing data.

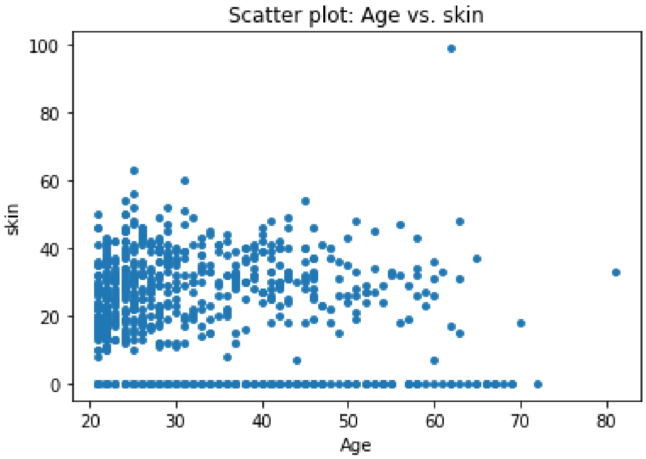

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

Inferences:

1. A very weak negative correlation.
2. The majority of population is lying between around 10 – 40mm.
3. There are very few outliers outside the 0 – 50 mm range.
4. There are a large number of points with 0 data value, and can be attributed to missing data.

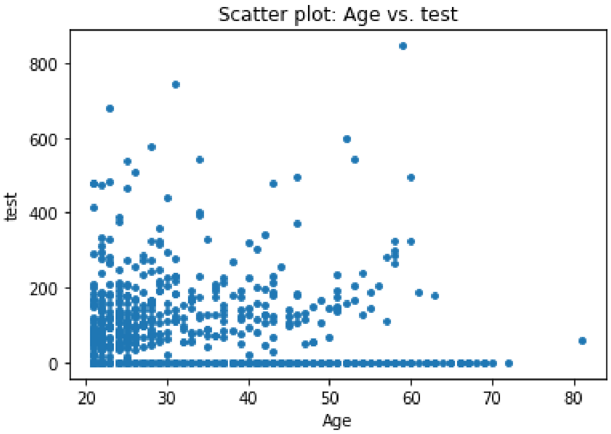

Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

Inferences:

1. There is extremely weak negative correlation between age and test.
2. Most data points lay within 200 mu U/mL
3. The outliers in the data are spread across evenly till the age of 60.
4. There are a large number of points with 0 data value, and can be attributed to missing data.

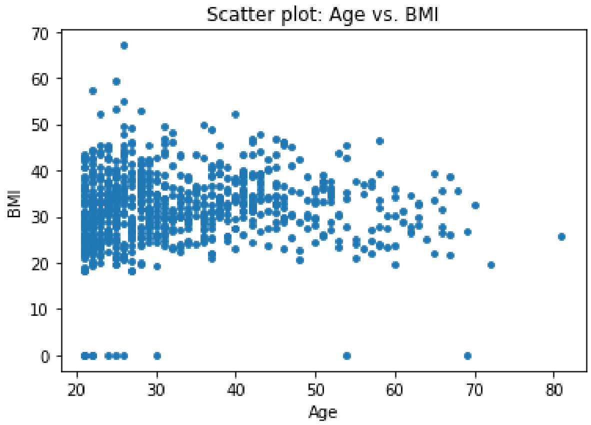

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m^2)

Inferences:

1. There is extremely weak positive correlation between Age and BMI

2. The most density is between 20 – 45 kg/m^2.

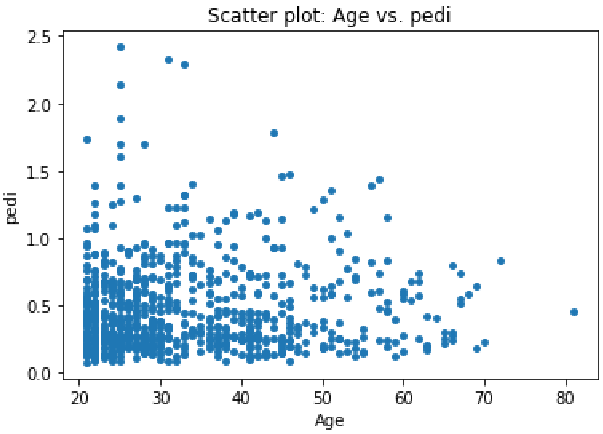

Figure 7 Scatter plot: Age (in years) vs. pedi

Inferences:

1. There is extremely weak positive correlation between Age and pedi
2. The majority data set lies in between 0 - 1.
3. A number of extreme outliers with data values above 1.5 are present in lower age ranges.

b.

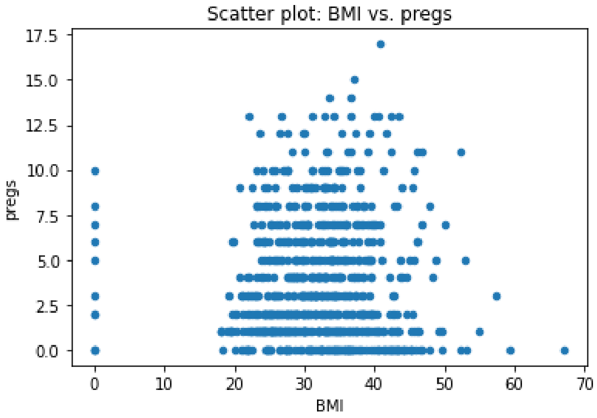

Figure 8 Scatter plot: BMI (in kg/m^2) vs. pregs

Inferences:

1. There is extremely weak positive correlation between BMI and number of times pregnant.
2. We can see a bell curve formation, thus women with average BMI have most pregnancies.
3. It can also be inferred that very low or very high BMI may lead to much fewer number of pregnancies.

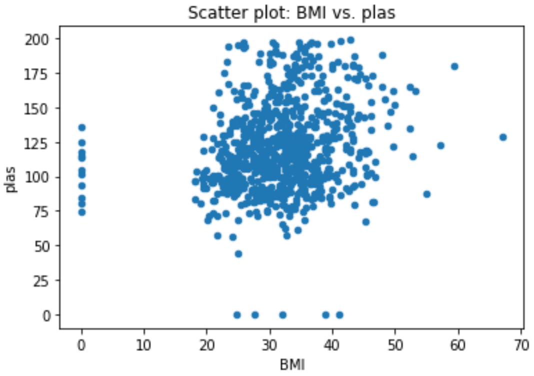

Figure 9 Scatter plot: BMI (in kg/m^2) vs. plas

Inferences:

1. There is a weak positive correlation between BMI and Plasma glucose concentration.
2. We can see that a normal BMI leads to normal Plasma glucose concentration.
3. A low BMI leads to lower Plasma glucose concentration and higher BMI corresponds to high Plasma glucose concentration.

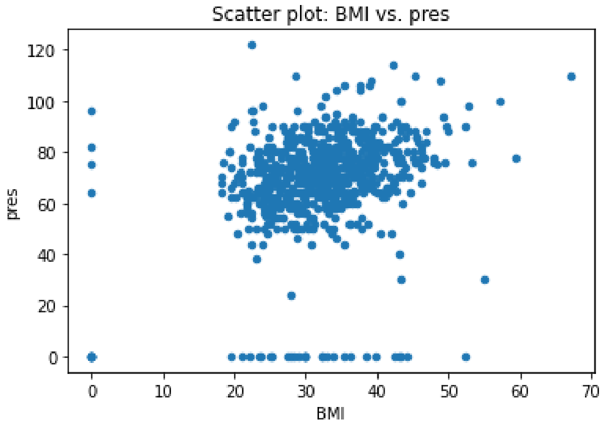

Figure 10 Scatter plot: BMI (in kg/m^2) vs. pres (in mm Hg)

Inferences:

1. There is a moderate positive correlation between BMI and Blood pressure
2. We can see that a normal BMI leads to normal Blood pressure values
3. A low BMI leads to low blood pressure and higher BMI corresponds to definite higher pressure.

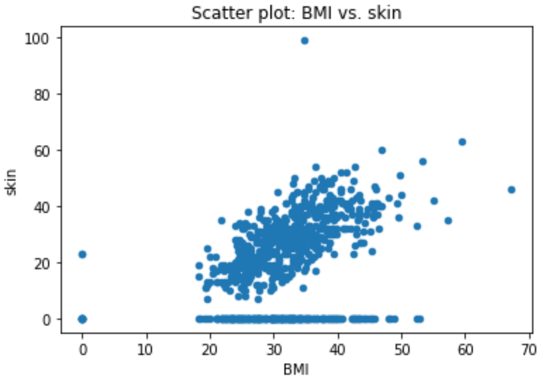

Figure 11 Scatter plot: BMI (in kg/m^2) vs. skin (in mm)

Inferences:

1. There is a strong positive correlation between BMI and skin thickness.
2. A low BMI leads to lower skin thickness and higher BMI leads to higher skin thickness.

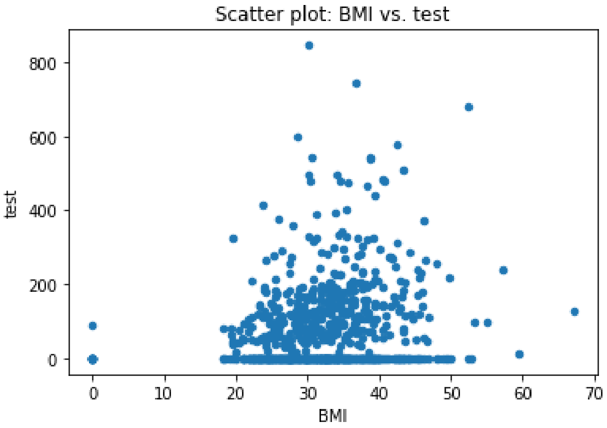

Figure 12 Scatter plot: BMI (in kg/m^2) vs. test (in mm U/mL)

Inferences:

1. There is a very weak positive correlation between BMI and 2-Hour serum insulin levels.
2. Most data points are in the normal range, while the outliers are almost evenly spread, irrespective of BMI

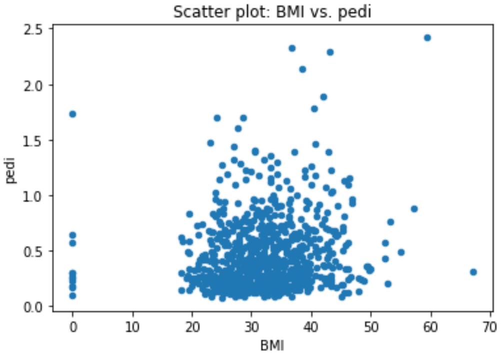

Figure 13 Scatter plot: BMI (in kg/m^2) vs. pedi

Inferences:

1. There is a very weak positive correlation between BMI and Diabetes pedigree function.
2. A normal BMI leads to high probability of having low Diabetes pedigree function value.
3. A high BMI may lead to large Diabetes pedigree function value.

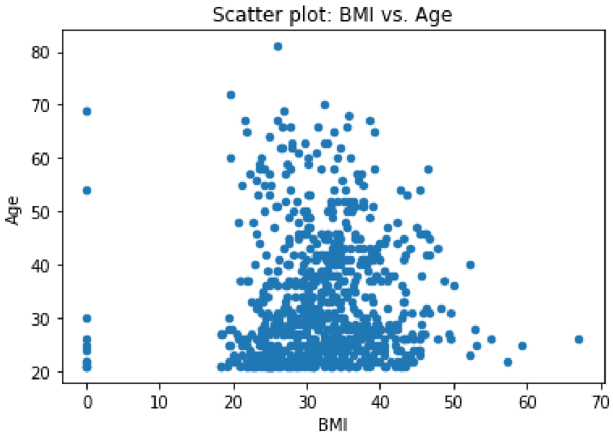

Figure 14 Scatter plot: BMI (in kg/m^2) vs. Age (in years)

Inferences:

1. There is an extremely weak positive correlation between BMI and Age.
2. Most BMI variations occur at younger ages around 20, at older ages, such variation is lesser.
3. In general, there is very minimal correlation between BMI and Age which is evident from the scatter 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	0.263
3	pres (in mm Hg)	0.239
4	skin (in mm)	-0.113
5	test (in mu U/mL)	-0.042
6	BMI (in kg/m^2)	0.036
7	pedi	0.033
8	Age (in years)	1.000

Inferences:

1. **pregs** – There is a very strong correlation between age and number of pregnancies, which is expected as with increase in age, more pregnancies are expected.

   Plas and pres – There is a moderate correlation between both plasma glucose correlation and blood pressure with age, with increase in age there is a higher probability of increased plasma glucose concentration and blood pressure.

   Skin and test – There are very weak negative correlation between both skin thickness and serum insulin levels with age. The scatter plots reflect this data.

   BMI – There is an extremely weak positive correlation between age and BMI, leading to the conclusion that BMI in general remains independent of age.

   Pedi – There is an extremely weak positive correlation between age and Diabetes pedigree function, which is also apparent from the scatter plot.

2. From this data we can infer that, Plasma glucose concentration, Blood pressure and Number of pregnancies are the most correlated to age while other parameters are minimally affected by increase in age.

b.

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

S. No.	Attributes	Correlation Coefficient Value
1	pregs	0.017
2	plas	0.221
3	pres (in mm Hg)	0.281
4	skin (in mm)	0.392
5	test (in mu U/mL)	0.197
6	BMI (in kg/m^2)	1.000
7	pedi	0.140
8	Age (in years)	0.036

Inferences:

1. **pregs** – There is a very weak correlation between BMI and number of pregnancies.

   Plas, pres, test and skin – There is a moderate to strong correlation between plasma glucose correlation, blood pressure, serum insulin levels and skin thickness with BMI, which leads to the fact that, with higher BMI, there is a probability of higher blood glucose level, higher blood pressure, higher serum insulin level and more skin thickness.

   Pedi – There is a weak positive correlation between BMI and Diabetes pedigree function, which is also apparent from the scatter plot.

   age – There is a very weak positive correlation between BMI and age, which means that, there is low probability of age being the cause of extremes in BMI.

2. From this data we can infer that, Plasma glucose concentration, Blood pressure, skin thickness and serum insulin level are the most correlated to BMI.

4. a.

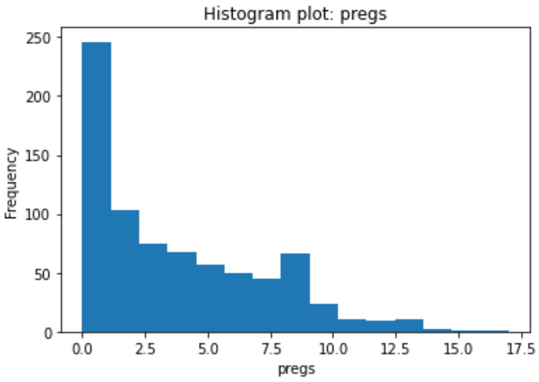

Figure 15 Histogram depiction of attribute pregs

Inferences:

1. From the histogram we can interpret that the highest frequency is for 0 – 3 pregnancies. The frequency of pregnancies drops quite rapidly after 3 pregnancies.
2. The mode of the histogram lies in between 0 – 2.5.

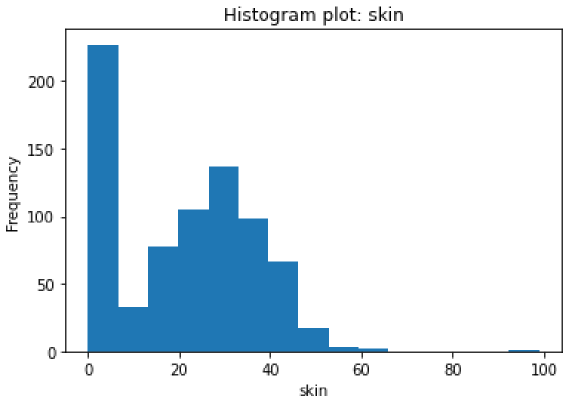

Figure 16 Histogram depiction of attribute skin

Inferences:

1. We can see from the histogram that there are two main peaks in the histogram, the first one is at 0 and the other one being at 20 – 40.
2. The mode lies in the first bin 0 – 20, with 0 being the most measured value for data set.

4

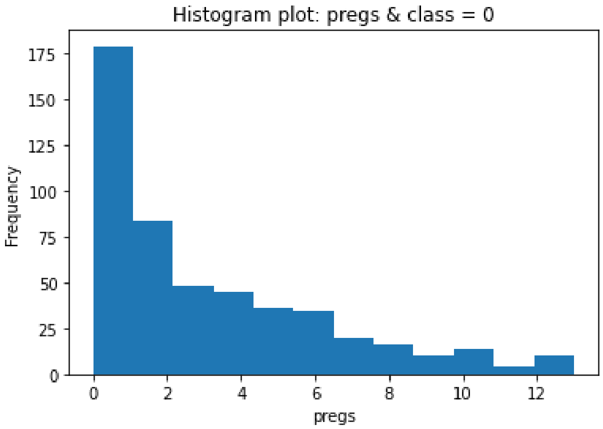

Figure 17 Histogram depiction of attribute pregs for class 0

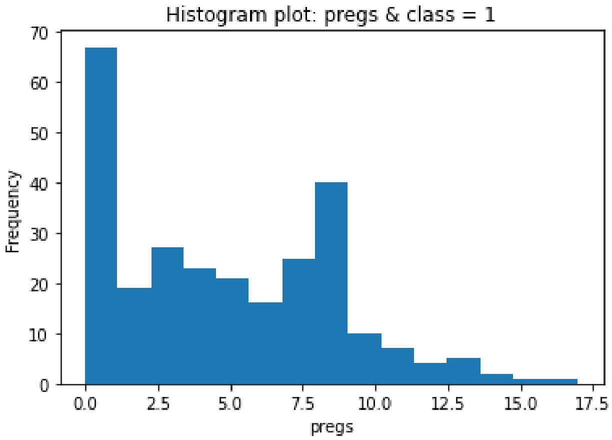

Figure 18 Histogram depiction of attribute pregs for class 1

Inferences:

1. Mode of the number of pregnancies is the same in both class 0 and class 1, it being 0 – 2.5.
2. The regular frequency trend is almost same for both the classes, with each having the highest frequency in the 0 – 2.5 bin, and then a general decline is seen, except for a small variation in class 1 where there is a sudden peak in the 7.5 – 10 bins.

5

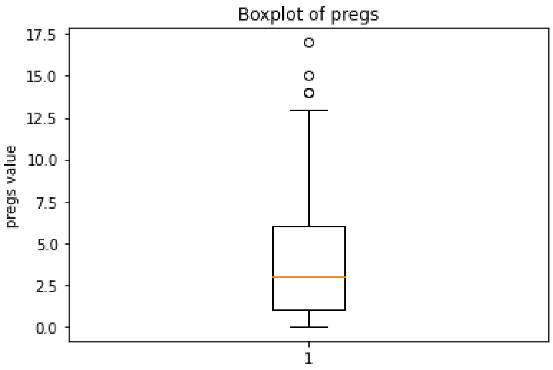

Figure 19 Boxplot for attribute pregs

Inferences:

1. The outliers are having 15 – 17 pregnancies.
2. The interquartile range is from 1 – 6 pregnancies.
3. The variability of the data is 5 pregnancies.
4. The data is skewed towards lower number of pregnancies

5. We can see from the boxplot that median is 3 and the majority of data is between 1 – 6 pregnancies, which can be inferred from the standard deviation.

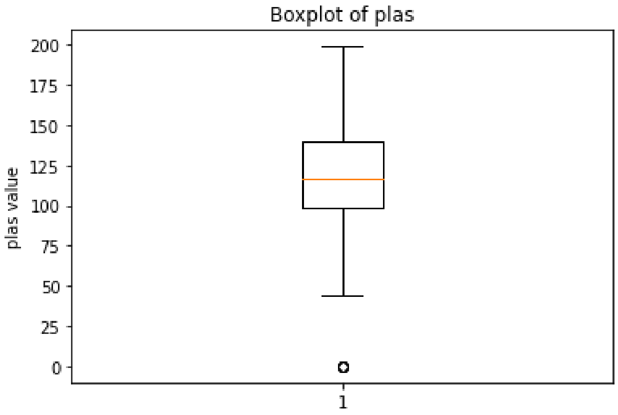

Figure 20 Boxplot for attribute plas

Inferences:

1. The outliers are having 0 plasma glucose concentration, which could infer missing data.
2. The interquartile range is from 100 – 140.
3. The variability of the data is 40.
4. The data is unskewed.
5. We can see from the boxplot that median is 117 and the data is unskewed with average being around 120 and max min being 199 and 0.

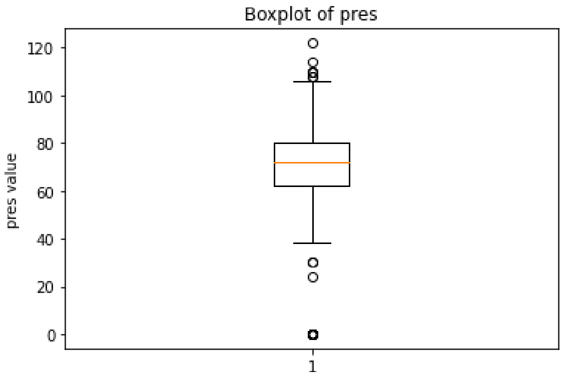

Figure 21 Boxplot for attribute pres(in mm Hg)

Inferences:

1. There are multiple outliers on both ends of the interquartile range. Some of the data can be attributed to missing data.
2. The interquartile range is from 60 - 80
3. The variability of the data is 20.
4. The data is unskewed.
5. We can see from the boxplot that median is close to 70 and the data is unskewed with average also being around 70.

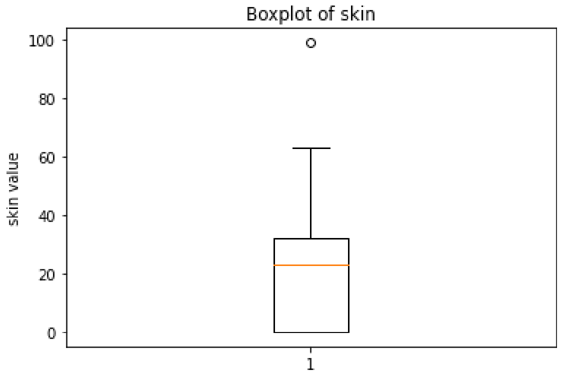

Figure 22 Boxplot for attribute skin(in mm)

Inferences:

1. There is only one outlier at 100.
2. The interquartile range is from 0 – 30.
3. The variability of the data is 30.
4. The data is heavily skewed to lower values of skin thickness.
5. We can see from the boxplot that median is close to 23 and the data is heavily skewed towards lower values, which infers that mode of the data is low in value.

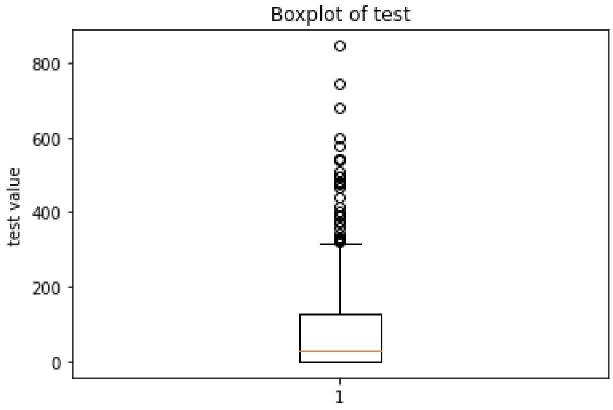

Figure 23 Boxplot for attribute test (mu U/mL)

Inferences:

1. There are a number of outliers, ranging from close to 300 to above 800.
2. The interquartile range is from 0 to approximately 125.
3. The variability of the data is close to 125.
4. The data is heavily skewed to lower values of serum insulin level.
5. We can see from the boxplot that median is close to 30 and the data is heavily skewed towards lower values, which infers that mode of the data is low in value.
6. From the boxplot we can see that while the majority of data is skewed low, there are a wide range of outliers in the data, which extend far above the interquartile range.

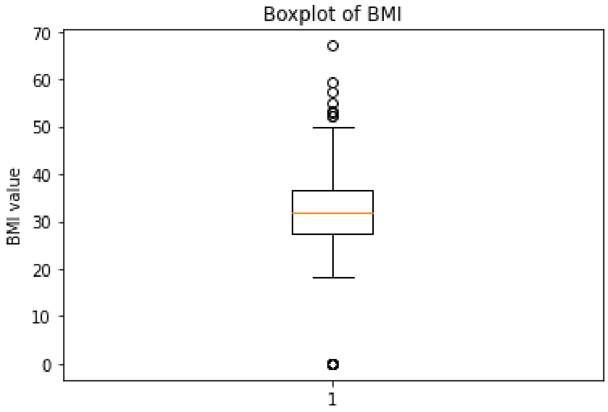

Figure 24 Boxplot for attribute BMI (in kg/m^2)

Inferences:

1. There are a few outliers from 50 to below 70 and some outliers infer to missing data.
2. The interquartile range is approximately from 25 – 35.
3. The variability of the data is approximately 10.
4. The data is unskewed.
5. We can see from the boxplot that median is close to 30 and the data is unskewed so the mean of the data will also be close to the median.

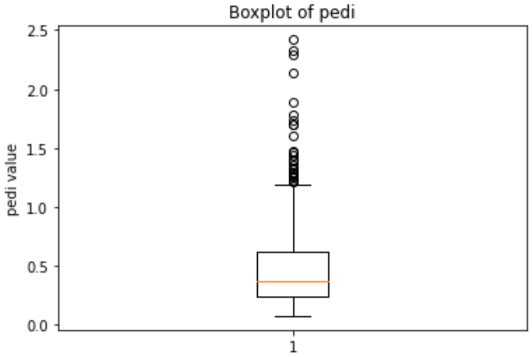

Figure 25 Boxplot for attribute pedi

Inferences:

1. There are large number of outliers, ranging from close to 1.25 to below 2.5.
2. The interquartile range is from 0.25 to approximately 0.6
3. The variability of the data is close to 0.35.
4. The data is heavily skewed to lower values of Diabetes pedigree function.
5. We can see from the boxplot that median is close to 0.3 and the data is heavily skewed towards lower values, which infers that mode of the data is low in value.
6. From the boxplot we can see that while the majority of data is skewed low, there are a wide range of outliers in the data, which extend far above the interquartile range.

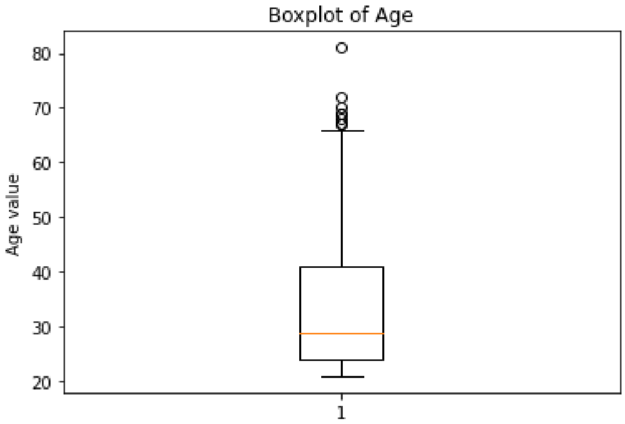

Figure 26 Boxplot for attribute Age (in years)

Inferences:

1. There are large number of outliers above 65 till around 80.
2. The interquartile range is from 10 - 40
3. The variability of the data is close to 30.
4. The data is moderately skewed to lower values of Age.
5. We can see from the boxplot that median is close to 29 and the data is heavily skewed towards lower values, which infers that mode of the data is low in value.