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

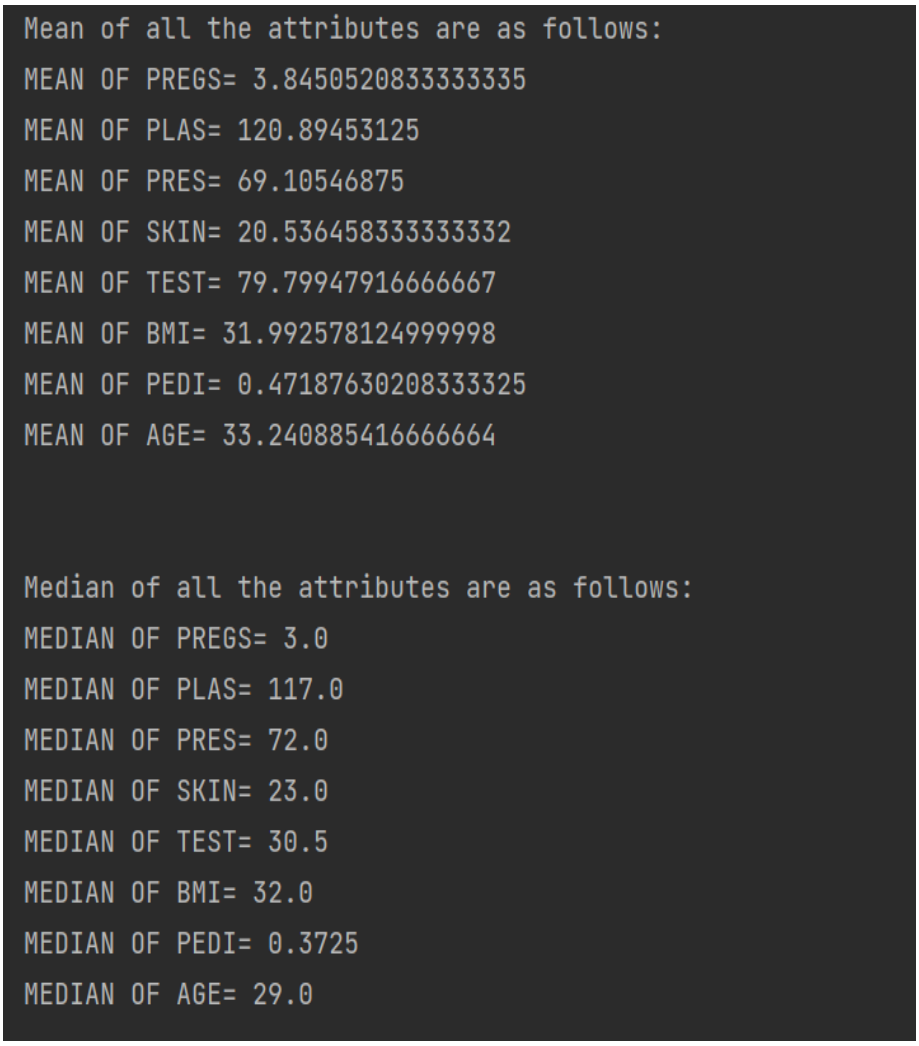
NAME: Sweety Agarwal

ROLL NO: B20234

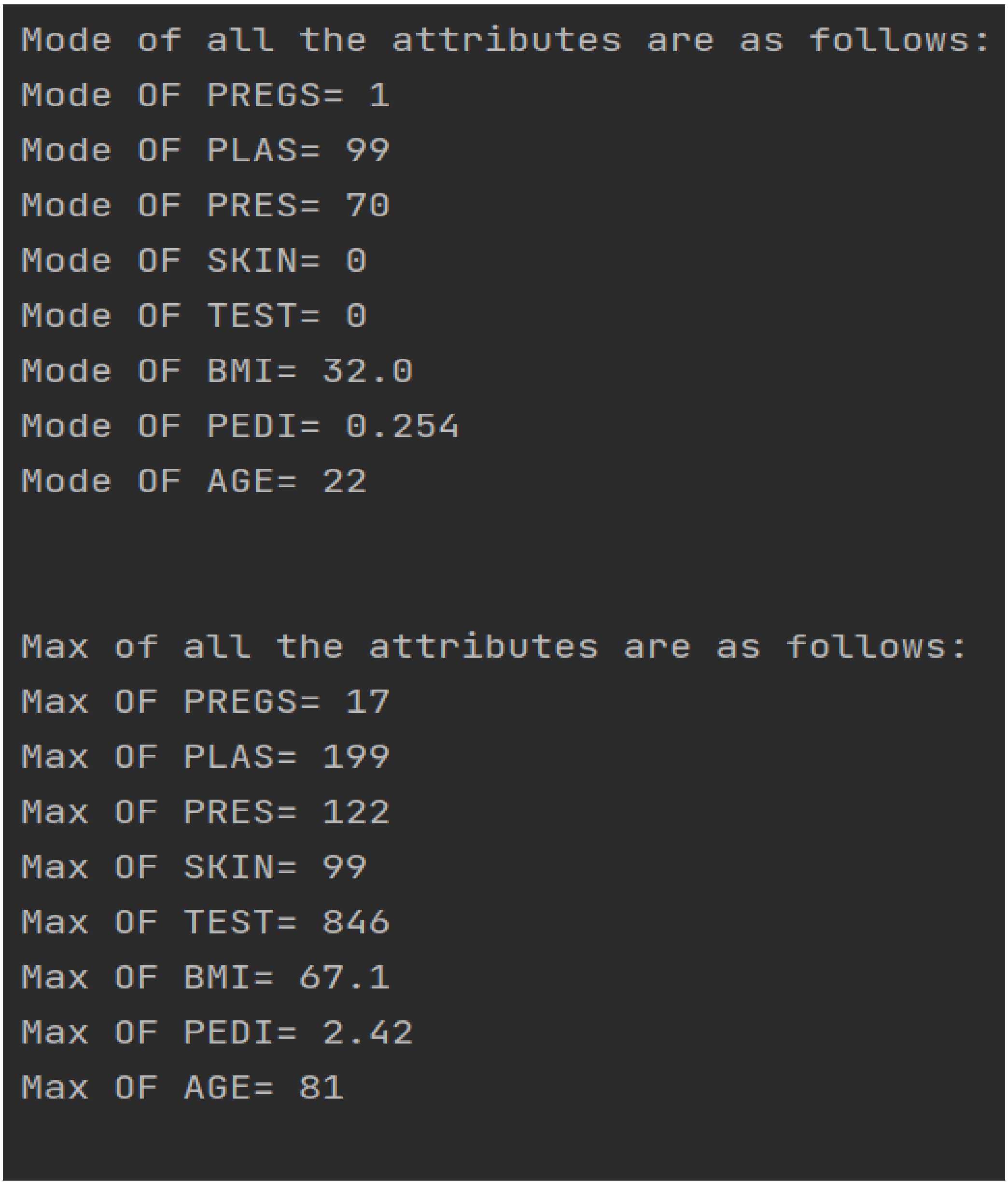
MOBILE NO: 8107876050

BRANCH: DATA SCIENCE & ENGINEERING

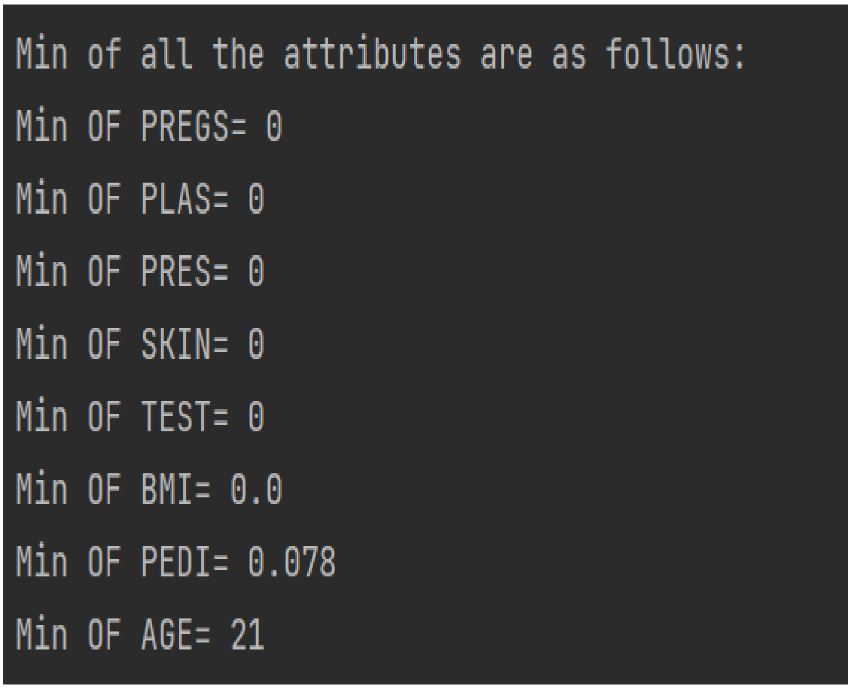
1. Mean, median, mode, minimum, maximum and standard deviation for all the attributes excluding the attribute‘class’.



```
Mean of all the attributes are as follows:
MEAN OF PREGS= 3.8450520833333335
MEAN OF PLAS= 120.89453125
MEAN OF PRES= 69.10546875
MEAN OF SKIN= 20.536458333333332
MEAN OF TEST= 79.79947916666667
MEAN OF BMI= 31.992578124999998
MEAN OF PEDI= 0.47187630208333325
MEAN OF AGE= 33.240885416666664


Median of all the attributes are as follows:
MEDIAN OF PREGS= 3.0
MEDIAN OF PLAS= 117.0
MEDIAN OF PRES= 72.0
MEDIAN OF SKIN= 23.0
MEDIAN OF TEST= 30.5
MEDIAN OF BMI= 32.0
MEDIAN OF PEDI= 0.3725
MEDIAN OF AGE= 29.0
```



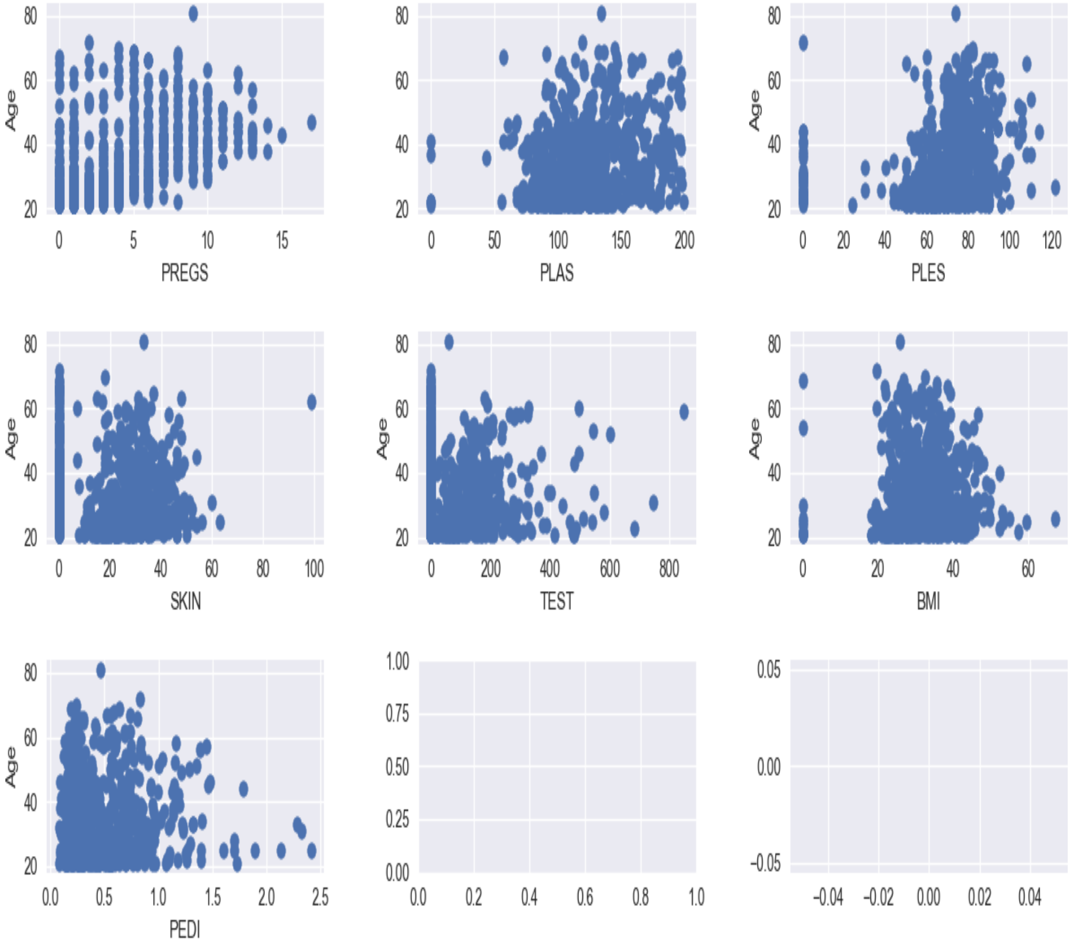
```
Mode of all the attributes are as follows:
Mode OF PREGS= 1
Mode OF PLAS= 99
Mode OF PRES= 70
Mode OF SKIN= 0
Mode OF TEST= 0
Mode OF BMI= 32.0
Mode OF PEDI= 0.254
Mode OF AGE= 22


Max of all the attributes are as follows:
Max OF PREGS= 17
Max OF PLAS= 199
Max OF PRES= 122
Max OF SKIN= 99
Max OF TEST= 846
Max OF BMI= 67.1
Max OF PEDI= 2.42
Max OF AGE= 81
```



```
Min of all the attributes are as follows:
Min OF PREGS= 0
Min OF PLAS= 0
Min OF PRES= 0
Min OF SKIN= 0
Min OF TEST= 0
Min OF BMI= 0.0
Min OF PEDI= 0.078
Min OF AGE= 21
```

The **Mean** or Average of this Data is a central tendency of the data i.e. a number around which a whole data is spread out.
Median is the value that divides the data into 2 equal parts i.e. number of terms on the right side of it is the same as a number of terms on the left side of it when data is arranged in either **ascending or descending order**.
Mode is the term appearing maximum time in data set i.e. term that has the highest frequency.
Standard deviation is the measurement of the average distance between each quantity and mean. That is, how data is spread out from the mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.
From the above code, we can get idea about the **center of our data.**
The **Max and Min** values obtained from the data can be uded to get approximate range of a particular attribute.

2. Scatter plot between a. 'Age' and each of the other attributes, excluding 'class'



Scatter Plots allow us to identify and determine if there is a relationship (correlation) between two variables and the strength of that relationship.

- **Correlation between age and number of times pregnant:**
  From the plot obtained above, we can say that in general, as Age increases, number of time a person is pregnant is also increases. However the correlation is very weak so the one value depends on other value very slightly.
- **Correlation between age and Plasma glucose concentration:**
  From the plot obtained above, we can say that, There is hardly any correlation between Age and plasma glucose concentration. There is a very very weak positive correlation between them.

- **Correlation between age and Diastolic blood pressure:**

  From the plot obtained above, we can say that, There is hardly any correlation between Age and diagonistic blood pressure. There is a very very weak positive correlation between them.

- **Correlation between age and Triceps skin fold thickness:**
  From the plot obtained above, we can say that, in general, as age increases, Triceps skin fold thickness decreases. However, there is a very weal negative correlation between them which is almost negligible.

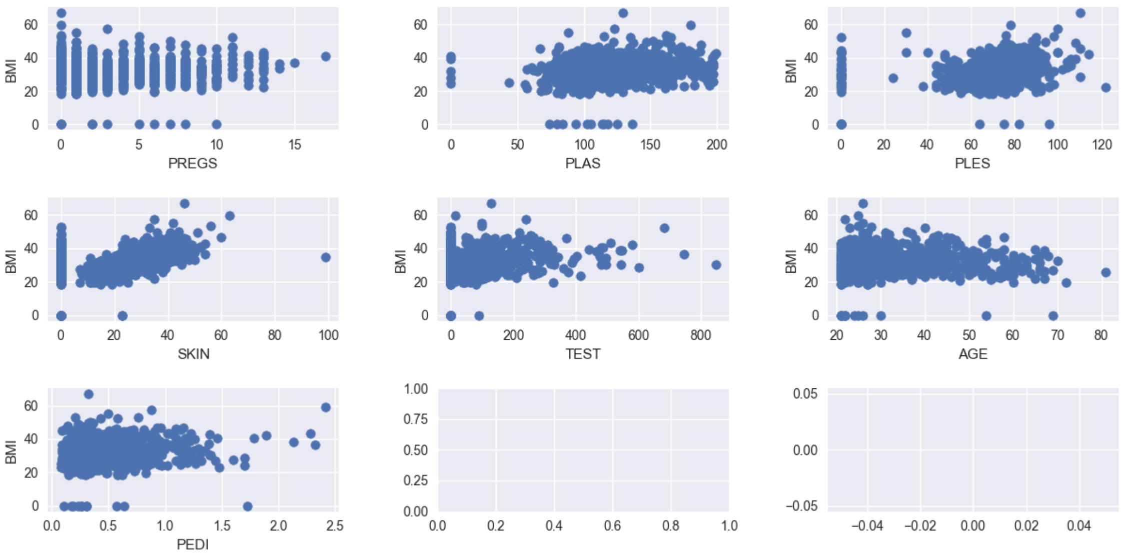
- **Correlation between age and 2-Hour serum insulin:**
- From the plot obtained above, we can say that, There is almost no correlation between the two. However, there is a very weak negative correlation between them which is almost negligible

- **Correlation between age and Body mass index:**
  From the plot obtained above, we can say that, in general, as Age increases, BMI also increases. However the correlation is very weak so the one value depends on other value very slightly.

- **Correlation between age and Diabetes pedigree function:**

  From the plot obtained above, we can say that, in general, as Age increases, Diabetes pedigree function also increases. However the correlation is very weak so the one value depends on other value very slightly.

2. Scatter plot between b. ‘BMI’ and each of the other attributes, excluding ‘class’



- **Correlation between BMI and number of times pregnant:**
  From the plot obtained above, we can say that, There is hardly any correlation between BMI and No of times a person is pregnant. There is a very very weak positive correlation between them.

- **Correlation between BMII and Plasma glucose concentration:**
  From the plot obtained above, we can say that, in general, as BMI increases, plasma glucose concentration also increases. However the correlation is very weak so the one value depends on other value very slightly.

- **Correlation between BMI and Diastolic blood pressure:**
  From the plot obtained above, we can say that, in general, as BMI increases, Diastolic blood pressure also increases. However the correlation is very weak so the one value depends on other value very slightly.

- **Correlation between BMI and Triceps skin fold thickness:**

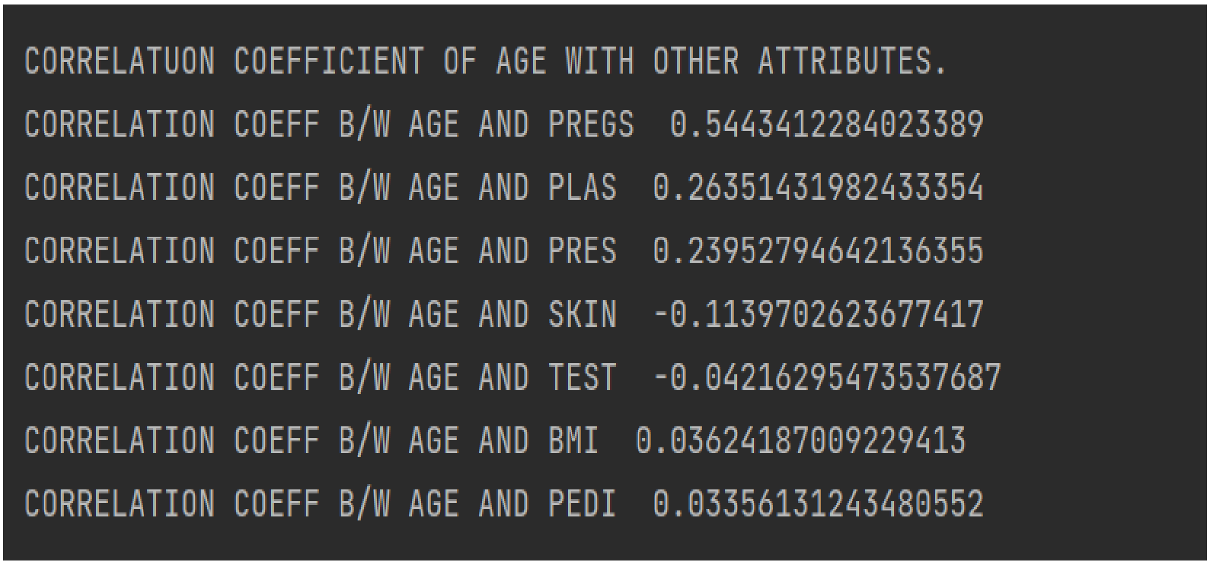
  From the plot obtained above, we can say that in general, as BMI increases, triceps skin fold thickness also increases. However the correlation is very weak so the one value depends on other value very slightly.

- **Correlation between BMI and 2-Hour serum insulin:**
  From the plot obtained above, we can say that, There is hardly any correlation between BMI and 2-Hour serum insulin. There is a very weak positive correlation between them.

- **Correlation between BMI and Age:**
  From the plot obtained above, we can say that, there is a very weak positive correlation between them. However the correlation is very weak and almost negligible.

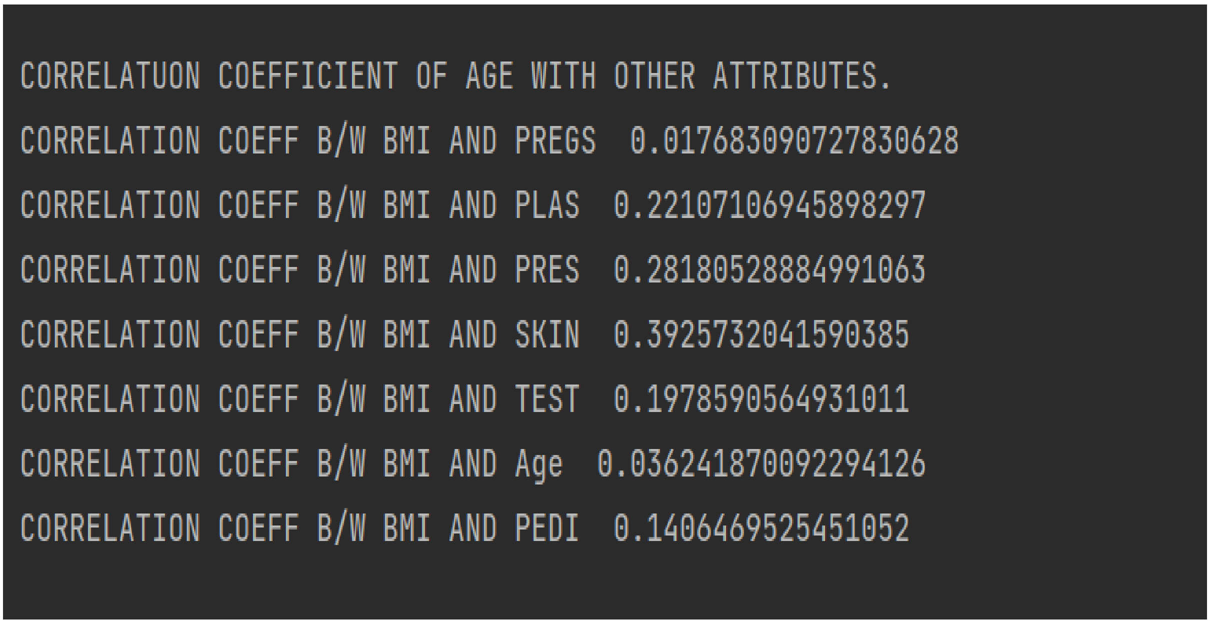
- **Correlation between BMI and**
  From the plot obtained above, we can say that, There is hardly any correlation between BMI and Diabetes pedigree function. There is a very weak positive correlation between them.

3)Correlation coefficient in the following cases: a. 'Age' with all other attributes (excluding 'class').



```
CORRELATUON COEFFICIENT OF AGE WITH OTHER ATTRIBUTES.
CORRELATION COEFF B/W AGE AND PREGS  0.5443412284023389
CORRELATION COEFF B/W AGE AND PLAS  0.26351431982433354
CORRELATION COEFF B/W AGE AND PRES  0.23952794642136355
CORRELATION COEFF B/W AGE AND SKIN  -0.1139702623677417
CORRELATION COEFF B/W AGE AND TEST  -0.04216295473537687
CORRELATION COEFF B/W AGE AND BMI  0.03624187009229413
CORRELATION COEFF B/W AGE AND PEDI  0.03356131243480552
```

3)Correlation coefficient in the following cases: a. ‘BMI’ with all other attributes (excluding ‘class’).



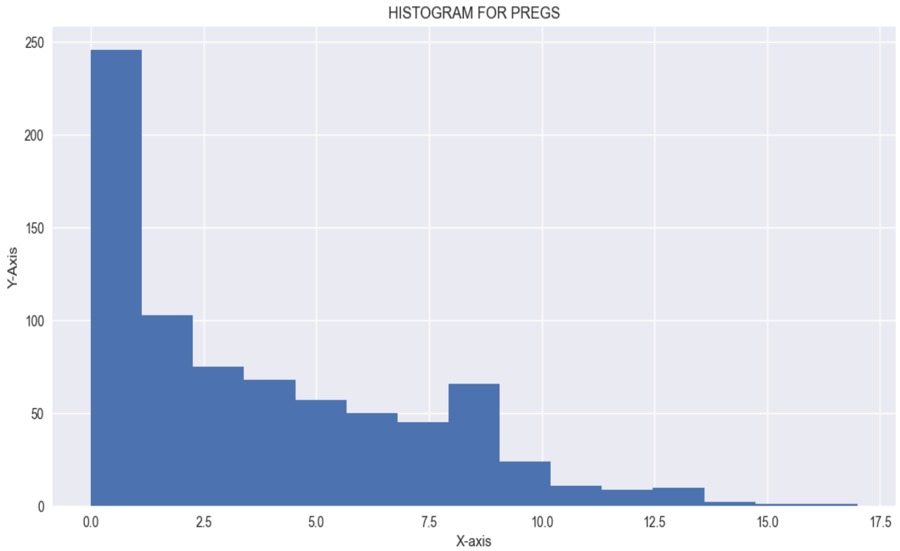
```
CORRELATUON COEFFICIENT OF AGE WITH OTHER ATTRIBUTES.
CORRELATION COEFF B/W BMI AND PREGS  0.017683090727830628
CORRELATION COEFF B/W BMI AND PLAS  0.22107106945898297
CORRELATION COEFF B/W BMI AND PRES  0.28180528884991063
CORRELATION COEFF B/W BMI AND SKIN  0.3925732041590385
CORRELATION COEFF B/W BMI AND TEST  0.1978590564931011
CORRELATION COEFF B/W BMI AND Age  0.036241870092294126
CORRELATION COEFF B/W BMI AND PEDI  0.1406469525451052
```

Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. The possible range of values for the correlation coefficient is -1.0 to 1.0.

- **Positive** correlation coefficient values indicate a positive correlation, where the values of both variables tend to increase together

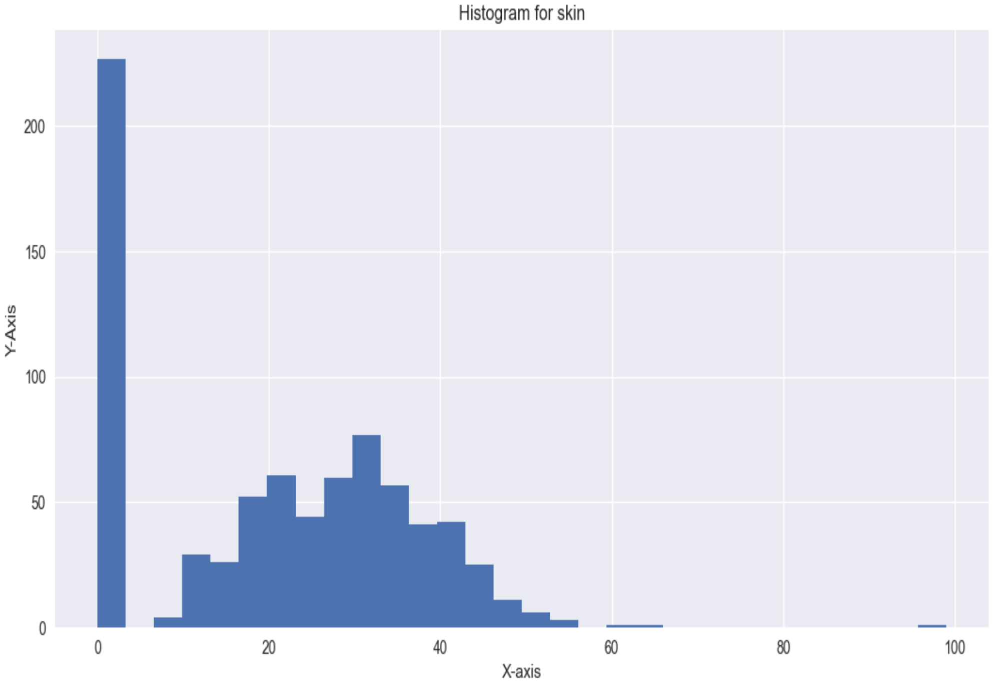
- **Negative** correlation coefficients values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.
- The closer correlation coefficient is to zero, the weaker the linear relationship.

4)Histogram for the attributes ‘preg’



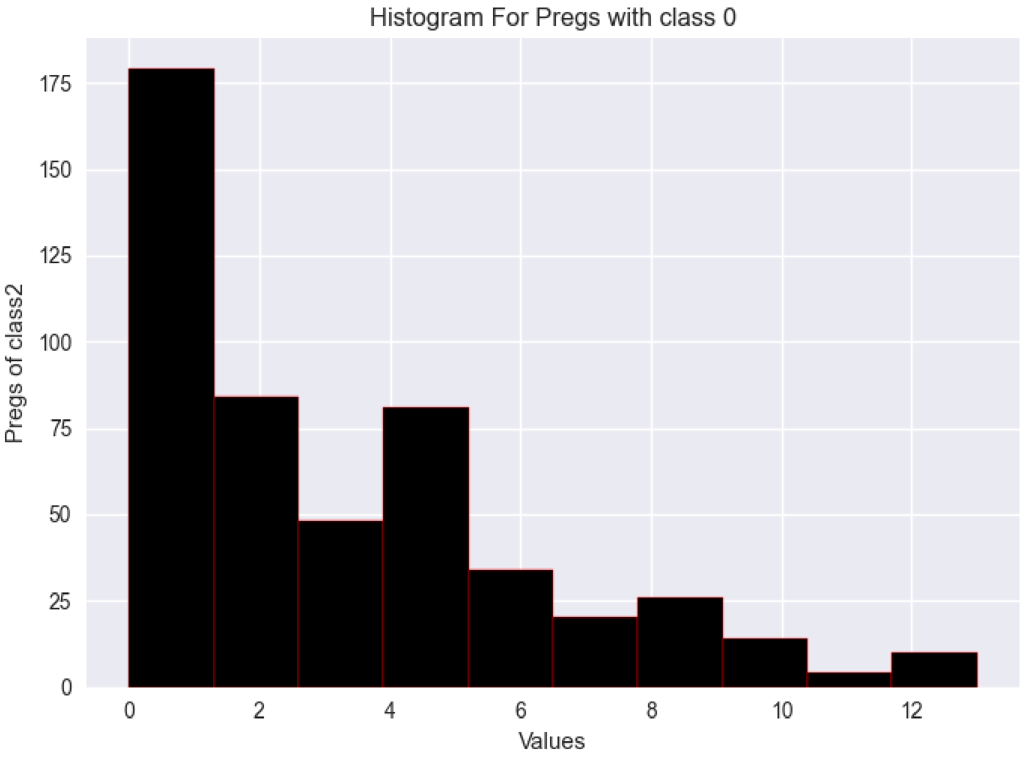
The bin size I have taken for this histogram of number of Times a person is pregnant is 15. From this histogram, we can say that the number of pregnancies of this particular data ranges from 0-17. However it The median of data lies somewhere around 3. Whereas the mode of of the data is approximately 1. We can conclude that This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side).

Histogram for the attributes 'SKIN'

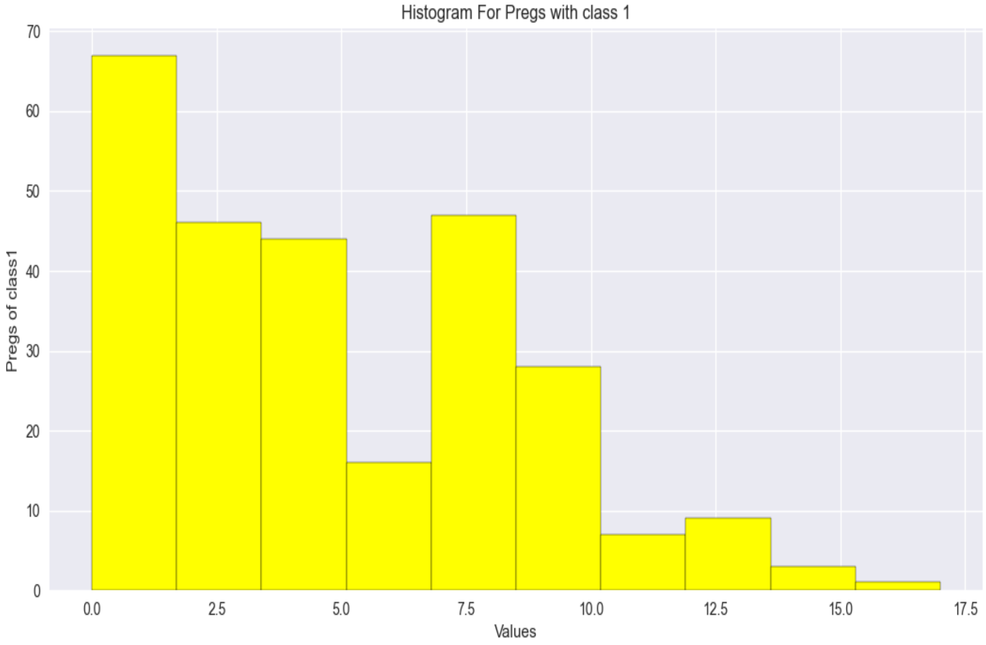


The bin size I have taken for this histogram of number of Times a person is pregnant is 15. From this histogram, we can say that Triceps skin fold thickness ranges from 0-60. In some rare cases, it is near 100. However The median of data lies somewhere around 25. Whereas the mode of of the data is approximately 0. We can conclude that This kind of distribution has a large number of occurrences in the very right side or middle .

5)Histogram of attribute 'preg' for each of the 2 classes individually (Using "groupby" function to group the tuples according to their "class")

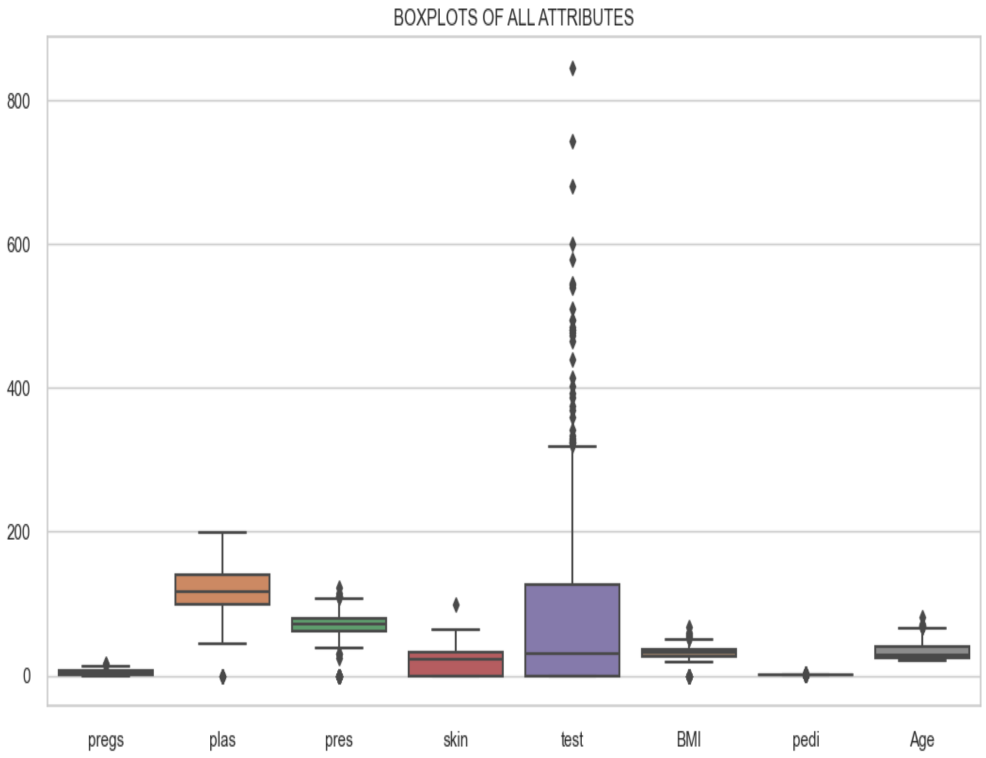


From the histogram, infer in which of the bins mode of the attribute pregs lies for class 0 .

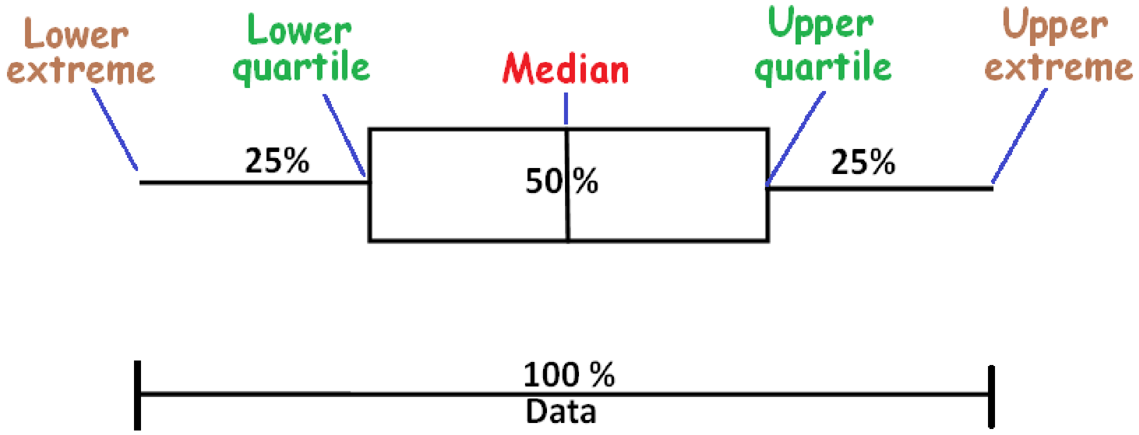


From the histogram, infer in which of the bins mode of the attribute pregs lies for class 1.

6) The boxplot for all the attribute excluding 'class' (Use "boxplot" function)



INFERENCE OF BOX PLOTS:


Lower extreme
Lower quartile
Median
Upper quartile
Upper extreme
25%
50 %
25%
100 %
Data