**Survey on details of students**

ARUMILLI MEGAHNA

PES1201701279

DEPARTMENT OF CSE

PES UNIVERSITY

BENGALURU,INDIA

[meghanamaggi2020@gmail.com](mailto:meghanamaggi2020@gmail.com)

Introduction:

Dataset is selected in such a way that there should be something meaningful to clean

This dataset contains 183 row and 10 columns. Hence the dataset is large because n is greater than 30.

It is a survey of students(details of each student).

**Dataset selected**: survey on student details

Dataset gives the description of the following:

Gender

Age

Dob

Height

Weight

Sun sign

CGPA

Rate programming

Programming language

Mean

Median

**Dataset cleaning:**

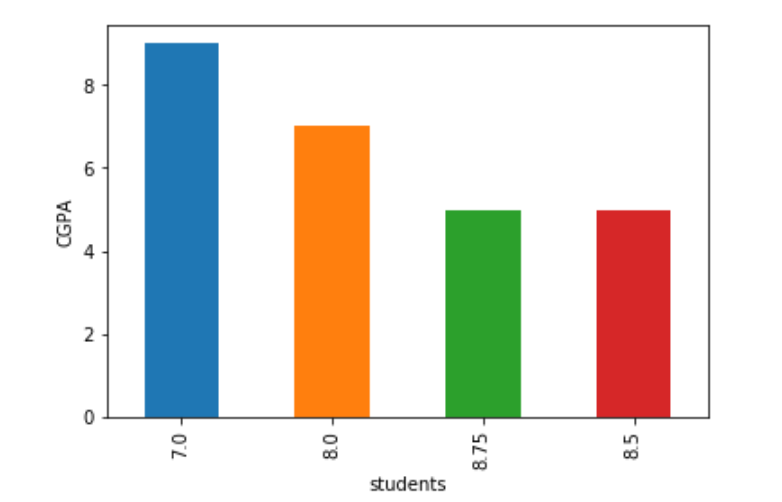
In the above dataset there are totally 21 columns we have removed 11 columns because of unwanted stuff.

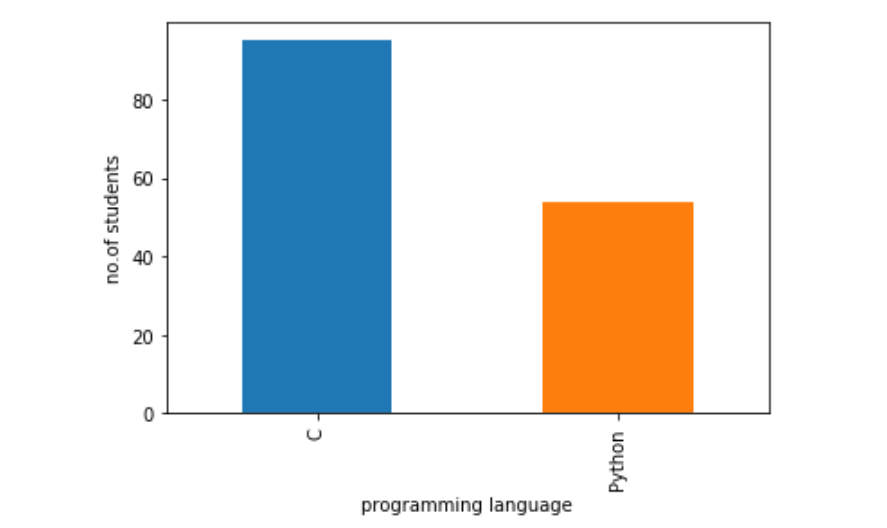
Dropping of unwanted rows and columns by clearly observing the dataset .

When the dataset was analysed it was observed that few of **columns** had insufficient data or no data to go on with the further visualization and comparisons. So we came to the conclusion that not considering those columns would make our visualisation better.

**Visualisation:**

In the above dataset the visualisation is done for the columns CGPA , programming language corresponding to count



  
After observing and analysing the dataset clearly visualisation is done.

1. Graph is plotted – student vs CGPA

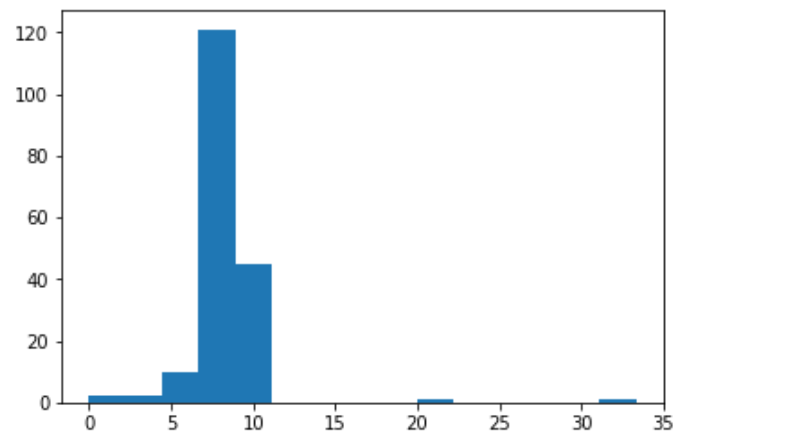
2. Graph is plotted

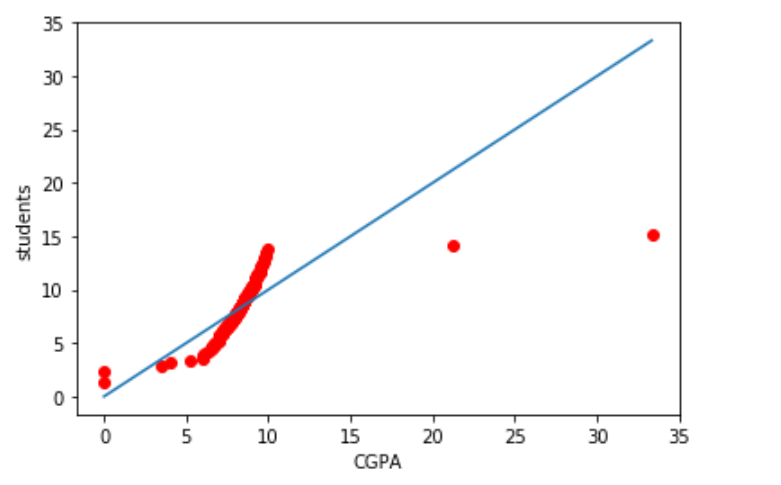
Based on no.of students who opted for C and Python.

Normalization:

We have checked normality test for the above data set.

The normality checking was not a normal plot





**Hypothesis testing:**

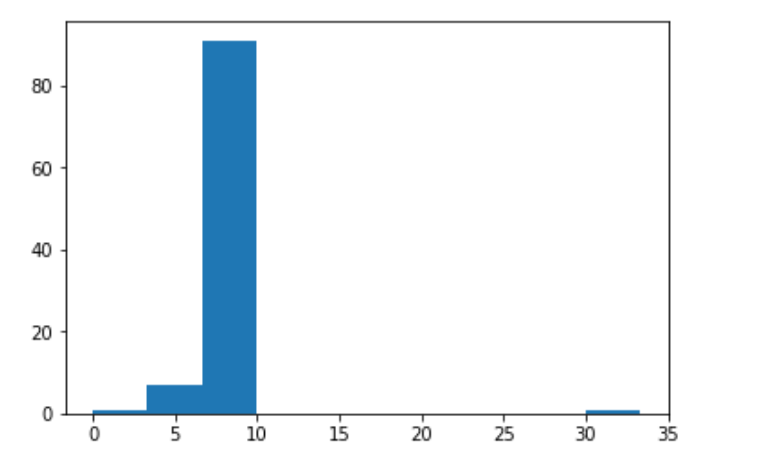
Since our dataset is not normally distributed so we had conducted chi-square test for the column CGPA .

We are taking the sample of CGPA of 100 students.

Alternate and null hypothesis are:

Ha : p!=0 Ho : p=0

Test statistic is 938249922458.019, pvalue is 0.0040



**Confidence interval:**

For a sample size of 100 we got mean as 8.45915 and standard deviation as 3.147260162029825

We are calculating confidence interval for confidence value of 95%

Margin of error , MoE =Z(alpha/2)\*standard deviation/sqrt(n)

MoE=0.6169

Confidence interval =[mean - MoE , mean + MoE]

CI=[7.8422,9.0764]

**Result and conclusion:**

Since p value of hypothesis testing is 0.0040 and confidence value is 95% ,we are rejecting the null hypothesis as the p value is less then 0.005

We are rejecting null hypothesis