PROJECT-2 REPORT

DATA ANALYTICS - CS40003

TOPIC -4

Roll no: 16CS30033 Name: Shaikh Moin Dastagir
Mob no: 7709400624

- The language used for completing this project is R.
- The software used for completing this project is RStudio.
- For choosing random 50 samples of size 30 each I first made a vector consisting of values from 1 to 768(no of rows in the data set), then i randomly chose 30 values from that vector using the sample function and stored the corresponding rows in my sample , repeated the process 50 times to get 50 samples, stored it all in a list of dataframes.
- For the second part , I used the shapiro normality test check .
  This test gives the p value , if the p value is less than 0.05 one can say surely that the data is not normally distributed , which i found is the case with all the columns of the Data.
  Shapiro values found:: 2.2e-16 1.628e-15 2.2e-16 2.2e-16 2.2e-16 2.2e-16 2.2e-16 2.2e-16 2.2e-16 2.2e-16
  For the columns respectively , all these values are way less than 0.05 .
  When the shapiro test was applied to a sample of the data set for the column X1 gave the value -> 0.008 , which is still less than 0.05.
- The histogram made from the data set also proves that the Data is not normal.
- The qqnorm and qqform also proves that the data is not normal.
- For part 3 of the question
  Samples are taken as in question 1, the mean is computed for each sample, and then the means are used as the data, rather than individual scores being used. The sample is a sampling distribution of the sample means.
- The mean of the sample means will be the mean of the population
  After calculating the mean of the sample means I found the below result for the Data set .

X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
0.7662333	669.9443	318.4183	175.763	5.280333	3.465333	0.2343333	2.775333	22.54521	24.85174