Prashant Sharma

Grading:
20 marks- Complete submission of worksheet8
40 marks- Problem 1 and 40 marks- Problem 3
Solution-1(a):

```
> Scores=scan("https://www.isibang.ac.in/~athreya/Teaching/PaSwR/Scores")
> ##scan the data set Scores from the course website.
> summary(Scores)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   22.00   30.00   29.28   40.00   50.00
```

```
> ##Using summary() function to get summary of Scores and this function gives the minimum value,\\
> ##1st quartile, median, mean, 3rd quartile and maximum value of a data set.
```

Solution-1(b):

We have to calculate the proportion of data that is 1-standard deviation, 2-standard deviation, 3-standard deviation far from the mean.
Now, in general,

$$P(|Scores - mean(Scores)| < k * sd(Scores)); k = 1, 2, 3$$

$$= P(\frac{|Scores - mean(Scores)|}{sd(Scores)} < k)$$

$$= P(|cs| < k); cs = \frac{|Scores - mean(Scores)|}{sd(Scores)}$$

$$= P(cs > -k \& cs < k)$$

```
> mean(Scores)
```

```
[1] 29.27835
```

```
> ##this command gives the mean of Scores
> sd(Scores)
```

```
[1] 12.06816
```

```
> ##this command gives the standard deviation of Scores
> cs=(Scores-mean(Scores))/sd(Scores)
> onesdcs=cs[cs>-1 & cs<1]##Within one sd
> twosdcs=cs[cs>-2 & cs<2]##Within two sd
> threesdcs=cs[cs>-3 & cs<3]##Within three sd
> length(onesdcs)/length(cs)
```

```
[1] 0.628866
```

```
> ##Calculating proportion of data that is 1-Standard deviation far from mean
> length(twosdcs)/length(cs)
```

```
[1] 0.9587629
```

```
> #Calculating proportion of data that is 2-Standard deviation far from mean
> length(threesdcs)/length(cs)
```

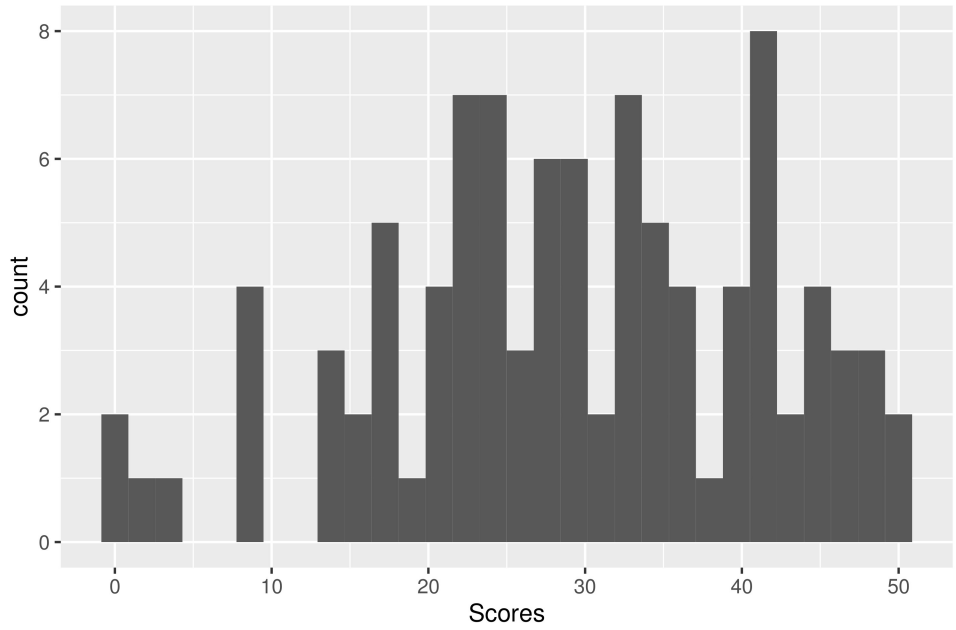
```
[1] 1
```

```
> #Calculating proportion of data that is 3-Standard deviation far from mean
```

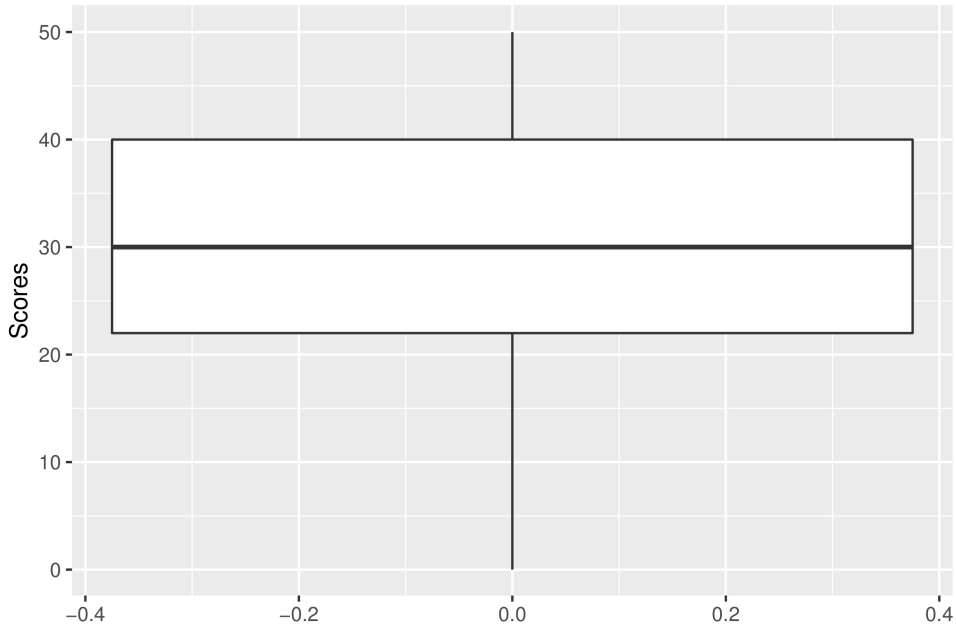
Thus, we get the proportion of data that is 1-Standard deviation, 2-Standard deviation, 3-Standard deviation far from mean are 0.628866, 0.9587629, 1 which is not satisfying the 3σ rule of normal distribution. Hence, we can conclude that the data set Scores is not normally distributed.

Solution-1(c):

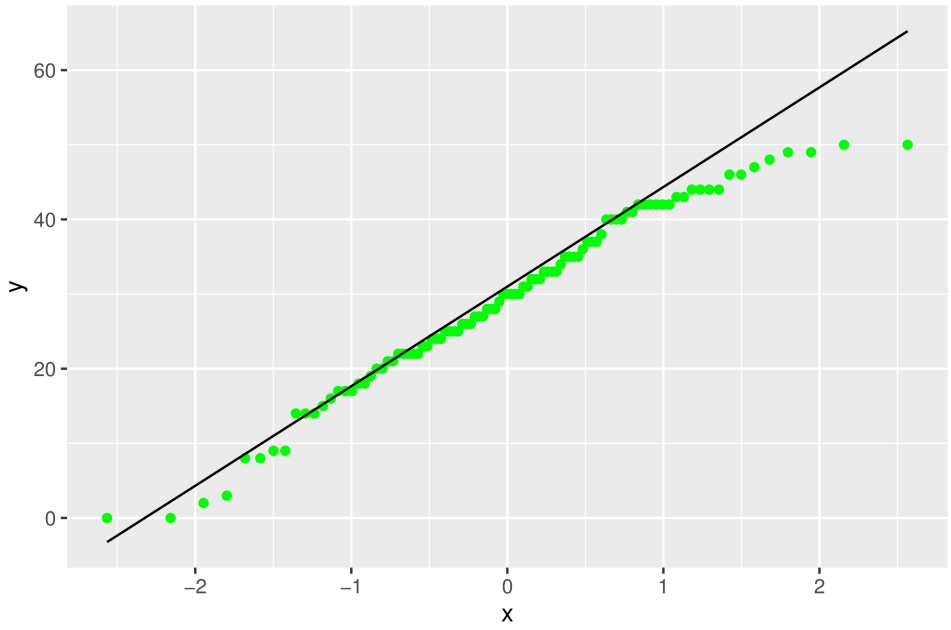
```
> #Plotting histogram of Scores
> library(ggplot2)
> df=data.frame(Scores)
> #hist_plot=ggplot(df, aes(x=Scores)) + geom_histogram(color="black",binwidth=1)
> hist_plot=ggplot(df,aes(x=Scores))+geom_histogram()
> #Hs=hist(Scores)
> #Hs
```




```
> #Plotting boxplot of Scores
> box_plot=ggplot(df,aes(y=Scores))+geom_boxplot()
> #Bs=boxplot(Scores)
> #Bs
```




```
> #Plotting Q-Q Plot of Scores
> QQ_plot=ggplot(df) + stat_qq(aes(sample = Scores), colour = "green")+stat_qq_line(aes(sample = Scores)
```



As we observe from the shape of boxplot and histogram that distribution of Scores is not symmetric. Also, in the Q-Q normal plot ,the plotted points are deviate on straight line. Hence, We can conclude that the distibution of Scores is not Normal.

Solution-1(d):

```
> library(moments)
> skewness(Scores)
```

```
[1] -0.3548957
```

```
> #computing skewness of the dataset Scores
> kurtosis(Scores)
```

```
[1] 2.591014
```

```
> #Computing kurtosis of the dataset Scores
```

Since the skewness of Scores is not 0 and the kurtosis is not 3(As for normal distribution, skewness should be 0 and kurtosis should be 1). Therefore, we can conclude that the distribution of Scores is not Normal distribution.

Solution-2(a):

```
> ?faithful
> ##command to get information and details about data set faithful
```

The data set faithful is a dataframe which describes the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowsatone National Park, Wyoming, USA. faithful is a data frame with 272 observations on 2 variables. And, the variable "eruptions" stores the eruption time in minutes and is a numerical type of data.

```
> ?ToothGrowth
> ##command to get information and details about data set ToothGrowth
```

The data set ToothGrowth is a data frame with 60 observations on 3 variables which includes the observation of the effect of vitamin C on Tooth Growth in Guinea Pigs. Also, the variable "len" stores the data of tooth lengths of numeric type.

Solution-2(b):

```
> eruption=faithful$eruptions
> #command to get summary(i.e, minimum value,1st Quartile, median, mean, 3rd Quartile,maximum value) of
> ##variable eruption
> summary(eruption)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
```

```
> mean(eruption)
```

```
[1] 3.487783
```

```
> sd(eruption)
```

```
[1] 1.141371
```

```
> cs1=(eruption-mean(eruption))/sd(eruption)
> onesdcs1=cs1[cs1>-1 & cs1<1]##Within one sd
> twosdcs1=cs1[cs1>-2 & cs1<2]##Within two sd
> threesdcs1=cs1[cs1>-3 & cs1<3]##Within three sd
> ##Calculating proportion of data that is 1-Standard deviation far from mean
> length(onesdcs1)/length(cs1)
```

```
[1] 0.5514706
```

```
> #Calculating proportion of data that is 2-Standard deviation far from mean
> length(twosdcs1)/length(cs1)
```

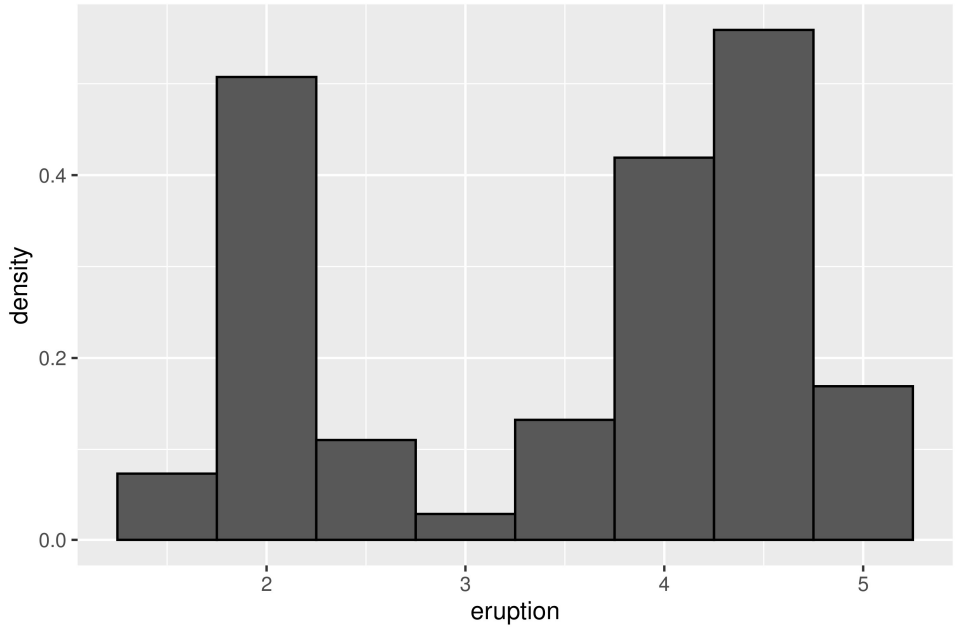
```
[1] 1
```

```
> #Calculating proportion of data that is 3-Standard deviation far from mean
> length(threesdcs1)/length(cs1)
```

```
[1] 1
```

Thus, we get the proportion of data that is 1-Standard deviation, 2-Standard deviation, 3-Standard deviation far from mean are 0.5514706, 1, 1 which is not satisfying the 3σ rule of normal distribution. Hence, we can conclude that the data set eruption is not normally distributed.
Now,

```
> #Plotting histogram of eruption
> df1=data.frame(eruption)
> hist_erup=ggplot(df1)+geom_histogram(mapping=aes(x=eruption,y=..density..), color="black",binwidth=0.5
```



From the plotted histogram, we can conclude that the distibution of eruption is not Normal.

```
> skewness(eruption)
```

```
[1] -0.415841
```

```
> kurtosis(eruption)
```

```
[1] 1.4994
```

Since the skewness of eruption is -0.415841 and the kurtosis is 1.4994(As for normal distribution, skewness should be 0 and kurtosis should be 1). Therefore, we can conclude that the distribution of eruption is not Normal distribution.
Hence, from the above descriptive measures and graphs, it is clear that distribution of eruptions does not follow normal distribution.

Now, for ”len” variable

```
> len=ToothGrowth$len
> #command to get summary(i.e, minimum value,1st Quartile, median, mean, 3rd Quartile,maximum value) of
> ##variable len
> summary(len)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20   13.07   19.25   18.81   25.27   33.90
```

```
> mean(len)
```

```
[1] 18.81333
```

```
> sd(len)
```

```
[1] 7.649315
```

```
> cs2=(len-mean(len))/sd(len)
> onesdcs2=cs2[cs2>-1 & cs2<1]##Within one sd
> twosdcs2=cs2[cs2>-2 & cs2<2]##Within two sd
> threesdcs2=cs2[cs2>-3 & cs2<3]##Within three sd
> ##Calculating proportion of data that is 1-Standard deviation far from mean
> length(onesdcs2)/length(cs2)
```

```
[1] 0.6666667
```

```
> #Calculating proportion of data that is 2-Standard deviation far from mean
> length(twosdcs2)/length(cs2)
```

```
[1] 1
```

```
> #Calculating proportion of data that is 3-Standard deviation far from mean
> length(threesdcs2)/length(cs2)
```

```
[1] 1
```

Thus, we get the proportion of data that is 1-Standard deviation, 2-Standard deviation, 3-Standard deviation far from mean are 0.6666667, 1, 1 which is not satisfying the 3σ rule of normal distribution. Hence, we can conclude that the data set len is not normally distributed.
Now,

```
> skewness(len)
```

```
[1] -0.1461768
```

```
> kurtosis(len)
```

```
[1] 2.024403
```

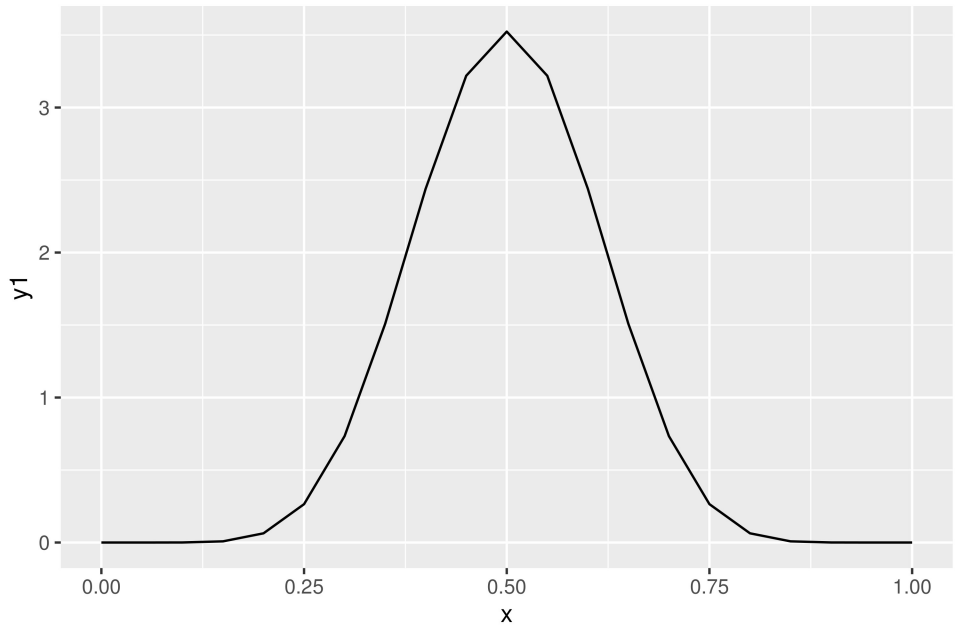
Since the skewness of len is -0.1461768 and the kurtosis is 2.024403(As for normal distribution, skewness should be 0 and kurtosis should be 1). Therefore, we can conclude that the distribution of len is not Normal distribution.
Hence, from the above descriptive measures and graphs, it is clear that distribution of len does not follow normal distribution.

Solution-3(a):

```
> #creating a sequence from 0 to 1 by jump of 0.05
> x = seq(0,1, by=0.05)
> x
```

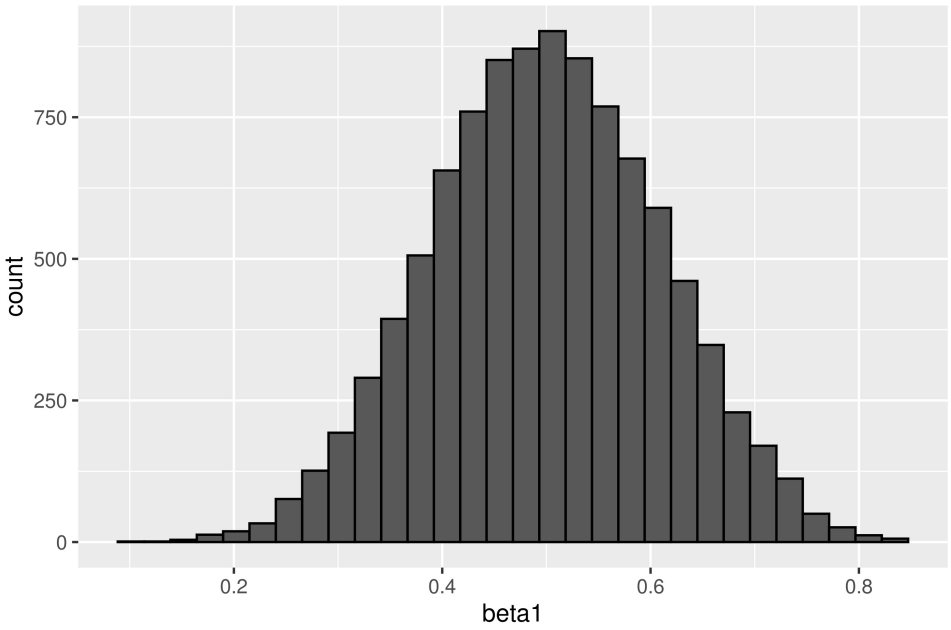
```
 [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70
[16] 0.75 0.80 0.85 0.90 0.95 1.00
```

```
> y1 = dbeta(x, 10,10)
> df_y1=data.frame(y1)
> plot_y1=ggplot(df_y1, aes(x=x, y=y1)) +geom_line()
```

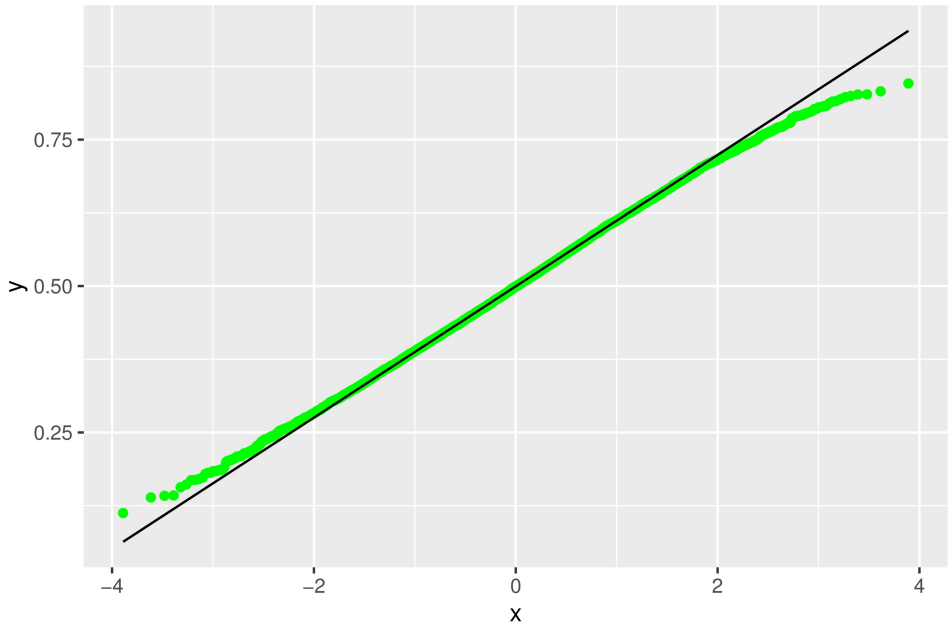


Now,

```
> #generating a random sample of size 10000 from Beta(10,10) and perform the following
> beta1=rbeta(10000,10,10)
> df_beta1=data.frame(beta1)
> beta1_plot=ggplot(df_beta1, aes(x=beta1)) + geom_histogram(color="black")
```




```
> QQ_beta1=ggplot(df_beta1) + stat_qq(aes(sample = beta1), colour = "green")+ stat_qq_line(aes(sample =
```




```
> library(moments)
> skewness(beta1)
```

```
[1] -0.0009771266
```

```
> kurtosis(beta1)
```

```
[1] 2.777612
```

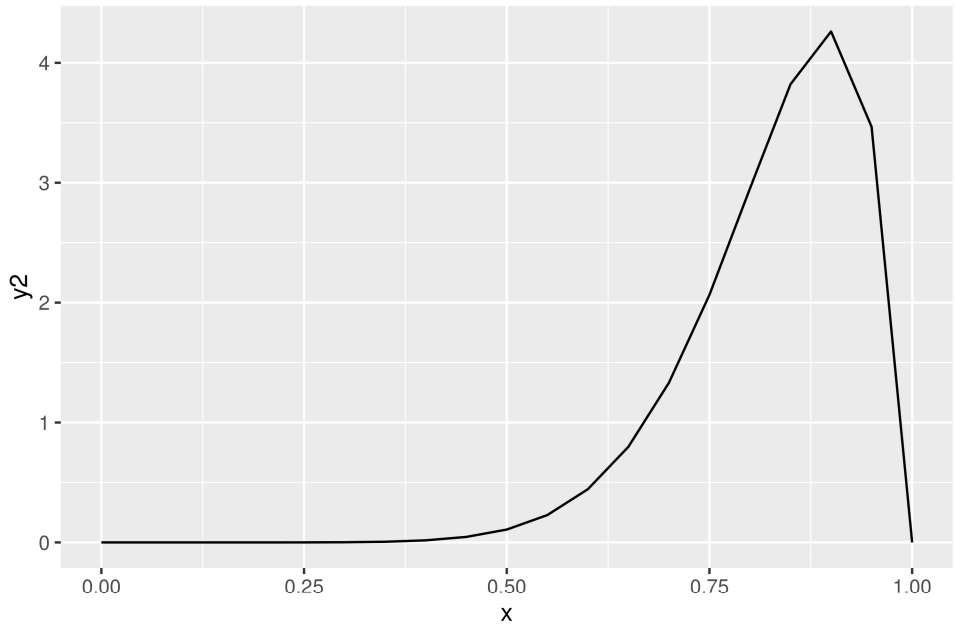
since value of skewness is close to 0 and value of kurtosis is close to 3, therefore we can conclude that Beta(10,10) is approximately normally distributed.
Also, by observing the plotted histogram and Q-Q plot, we can make conclusion for Beta(10,10) in favor of normal distribution.

Solution-3(b):

```
> #creating a sequence from 0 to 1 by jump of 0.05
> x = seq(0,1, by=0.05)
> x
```

```
 [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70
[16] 0.75 0.80 0.85 0.90 0.95 1.00
```

```
> y2 = dbeta(x, 10,2)
> df_y2=data.frame(y2)
> plot_y2=ggplot(df_y2, aes(x=x, y=y2)) +geom_line()
```

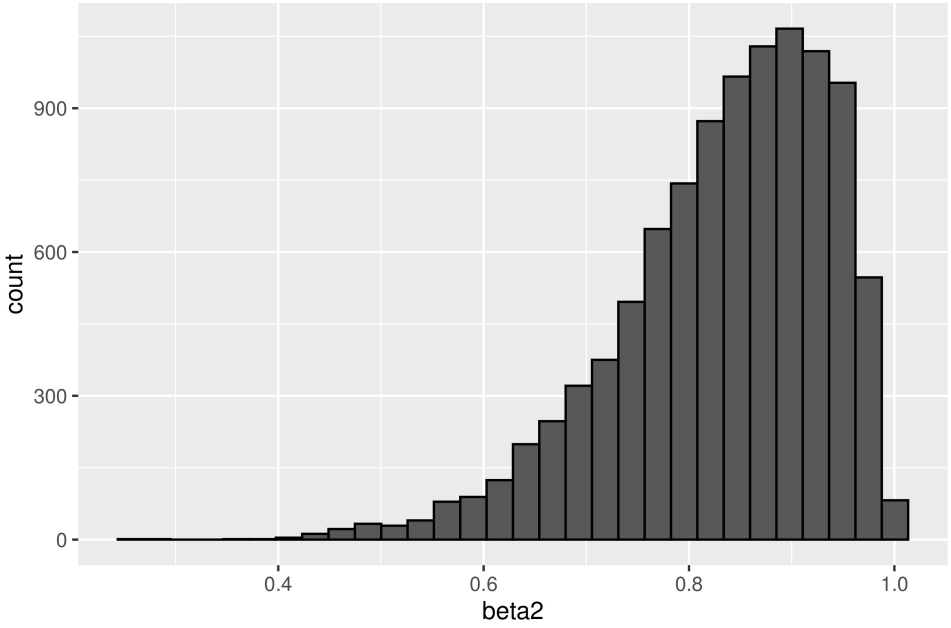


Now,

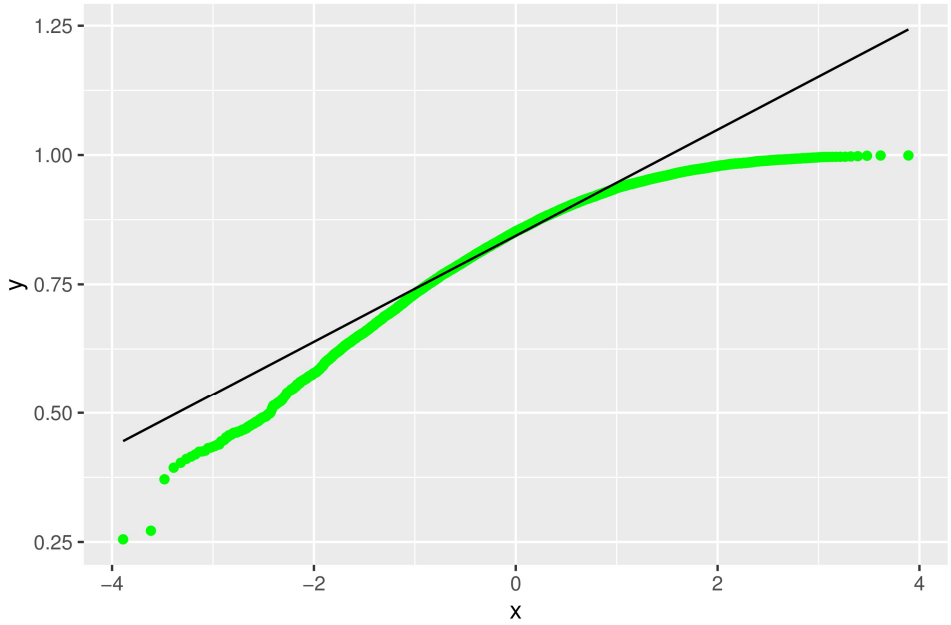
```
> #generating a random sample of size 10000 from Beta(10,2) and perform the following
> beta2=rbeta(10000,10,2)
> summary(beta2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2552  0.7748  0.8522  0.8337  0.9131  0.9992
```

```
> df_beta2=data.frame(beta2)
> beta2_plot=ggplot(df_beta2, aes(x=beta2)) + geom_histogram(color="black")
```




```
> QQ_beta2=ggplot(df_beta2) + stat_qq(aes(sample = beta2), colour = "green")+stat_qq_line(aes(sample = b
```



```
> library(moments)
> skewness(beta2)
```

```
[1] -0.934105
```

```
> kurtosis(beta2)
```

```
[1] 3.862452
```

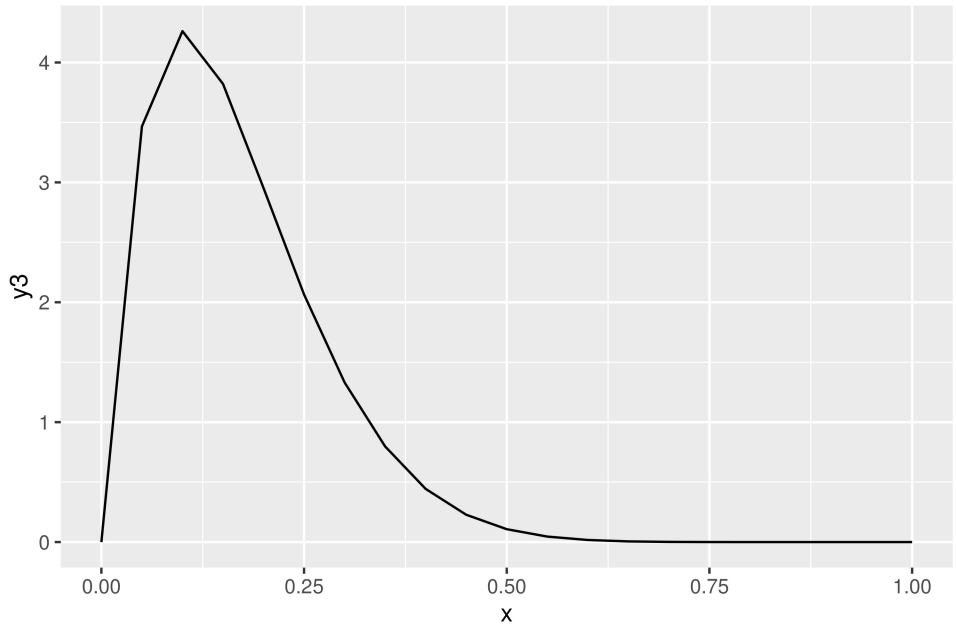
since value of skewness is less than 0 and value of kurtosis is greater than 3, therefore we can conclude that Beta(10,2) is not normally distributed and curve of Beta(10,2) is leptokurtic.
Also, by observing the plotted histogram and Q-Q plot, we can make conclusion for Beta(10,2) is not follow normal distribution and is left skewed curve.

Solution-3(c):

```
> #creating a sequence from 0 to 1 by jump of 0.05
> x = seq(0,1, by=0.05)
> x
```

```
 [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70
[16] 0.75 0.80 0.85 0.90 0.95 1.00
```

```
> y3 = dbeta(x, 2,10)
> df_y3=data.frame(y3)
> plot_y3=ggplot(df_y3, aes(x=x, y=y3)) +geom_line()
```

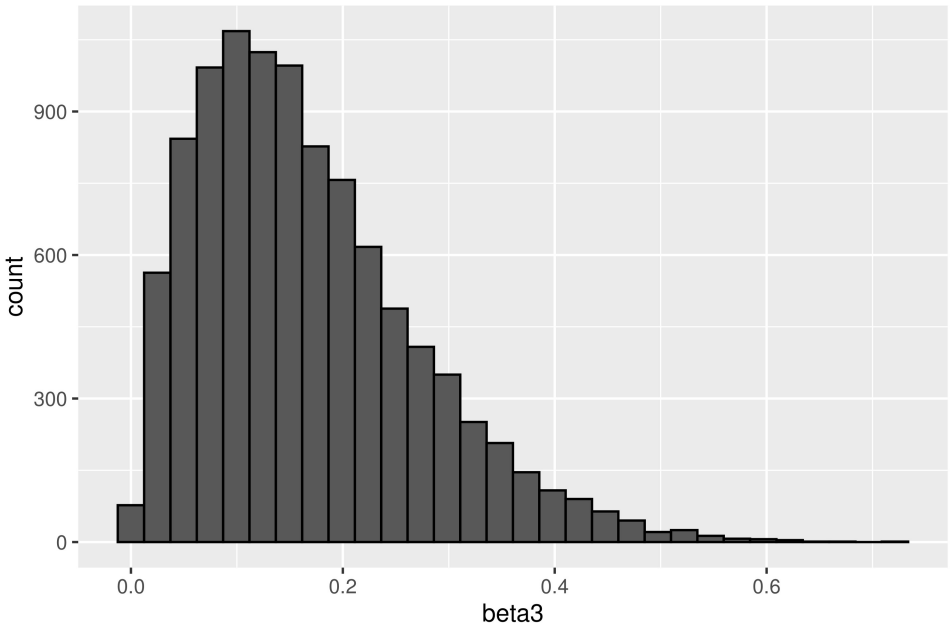


Now,

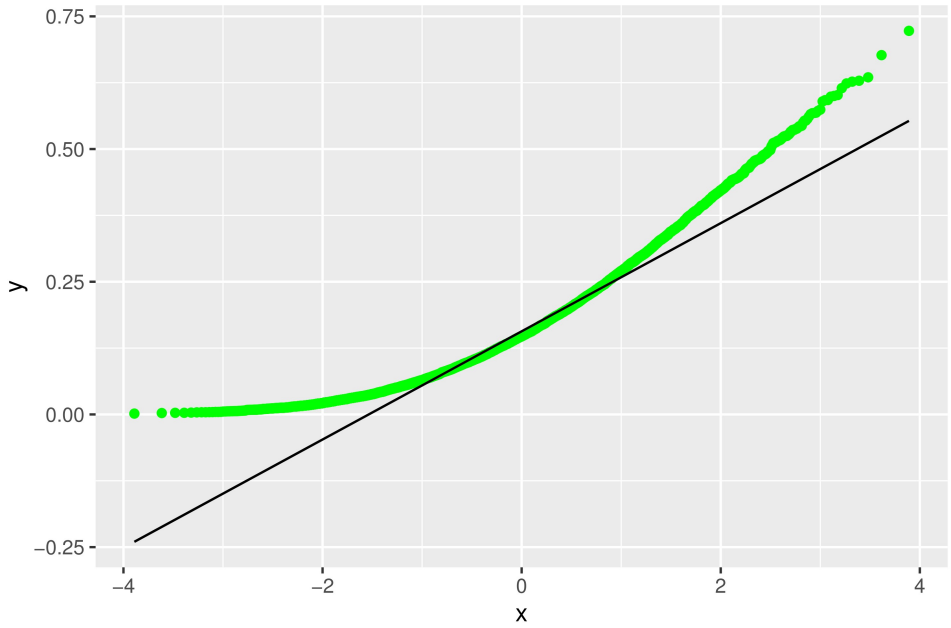
```
> #generating a random sample of size 10000 from Beta(2,10) and perform the following
> beta3=rbeta(10000,2,10)
> summary(beta3)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001572 0.087873 0.147316 0.166433 0.225356 0.722727
```

```
> df_beta3=data.frame(beta3)
> beta3_plot=ggplot(df_beta3, aes(x=beta3)) + geom_histogram(color="black")
```




```
> QQ_beta3=ggplot(df_beta3) + stat_qq(aes(sample = beta3), colour = "green")+stat_qq_line(aes(sample = b
```




```
> library(moments)
> skewness(beta3)

[1] 0.9505731

> kurtosis(beta3)
```

```
[1] 3.879299
```

since value of skewness is greater than 0 and value of kurtosis is greater than 3, therefore we can conclude that Beta(2,10) is not normally distributed and curve of Beta(2,10) is leptokurtic type.
Also, by observing the plotted histogram and Q-Q plot, we can make conclusion for Beta(2,10) is not follow normal distribution and is right skewed curve.