**PERFORMING DISTRIBUTION -ANALYSIS ON THE DATASET**

Name-Rajrishi Sarkar

Roll-203110061

**Concrete Compressive Strength**

**Data Type: multivariate**

Abstract. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.

**Variable Information:**

Given is the variable name, variable type, the measurement unit and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

**Name – Data Type – Measurement – Description**

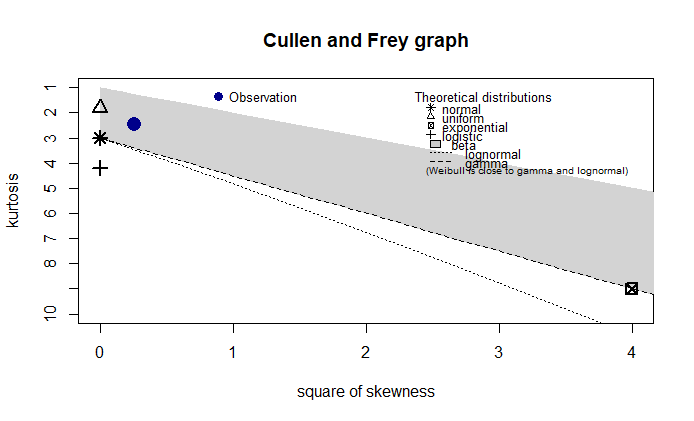
Cement (component 1) – quantitative – kg in a m3 mixture – Input Variable Blast Furnace Slag (component 2) – quantitative – kg in a m3 mixture – Input Variable Fly Ash (component 3) – quantitative – kg in a m3 mixture – Input Variable Water (component 4) – quantitative – kg in a m3 mixture – Input Variable Superplasticizer (component 5) – quantitative – kg in a m3 mixture – Input Variable Coarse Aggregate (component 6) – quantitative – kg in a m3 mixture – Input Variable Fine Aggregate (component 7) – quantitative – kg in a m3 mixture – Input Variable Age – quantitative – Day (1~365) – Input Variable Concrete compressive strength – quantitative – MPa – Output Variable

———————————

**ANALYSING THE FEATURES:**

**1)CEMENT**

Visualizing the skewness kurtosis plot



**summary statistics**

min: 102 max: 540

median: 272.9

mean: 281.1679

estimated sd: 104.5064

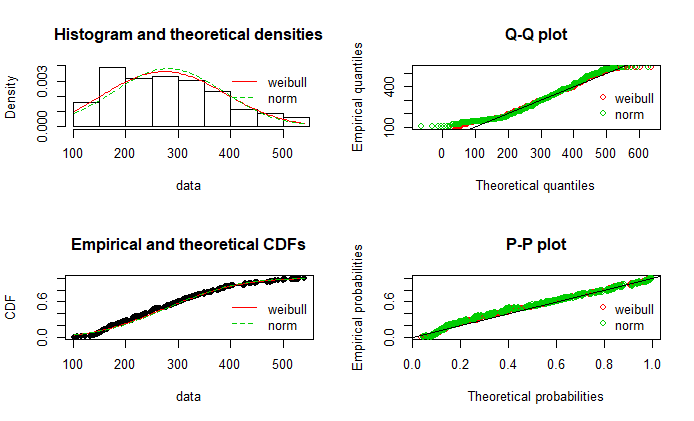
estimated skewness: 0.5094812

estimated kurtosis: 2.479348

observation lies in the beta distribution region

We try to fit the Weibull and Normal distribution and Beta distribution to the data

-Fitting distributions:



Both the distributions seem to fit quite nicely, however judging from the shape error it seems that the Weibull distribution is a better fit for the data.

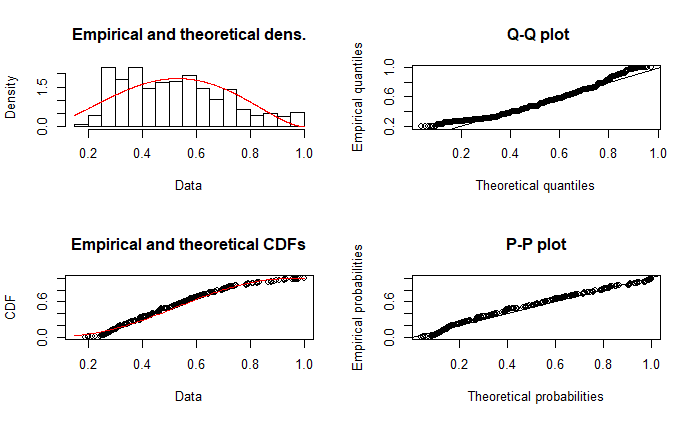
Comparing the AIC values of the three distributions

Weibull distribution : 11466.42

Normal distribution: 9952.473

Normal distribution is better fit of the two because of the lower AIC value. Now we try to fit the data into beta distribution.

Fitting beta distribution:



It is supposed to be the best fit

Fitting of the distribution ' beta ' by matching moments

Parameters:

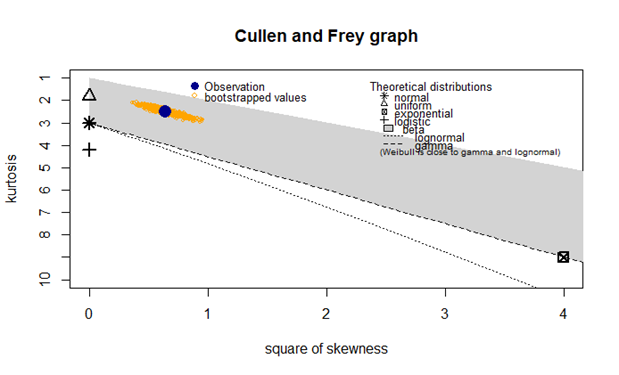
estimate

shape1: 2.952218

shape2: 2.717696

**2)BLAST FURNACE SLAG**

Analyzing the Cullen Frey graph



Summary statistics:

min: 0

max: 359.4

median: 22

mean: 73.89583

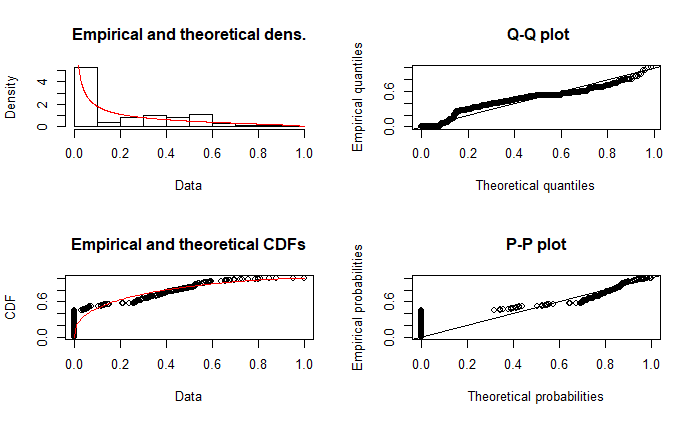
estimated std: 86.27934

estimated skewness: 0.8007169

estimated kurtosis: 2.491825

As visible from the graph the observation lies in the beta distribution region.

Fitting in the beta distribution:



Parameters:

estimate

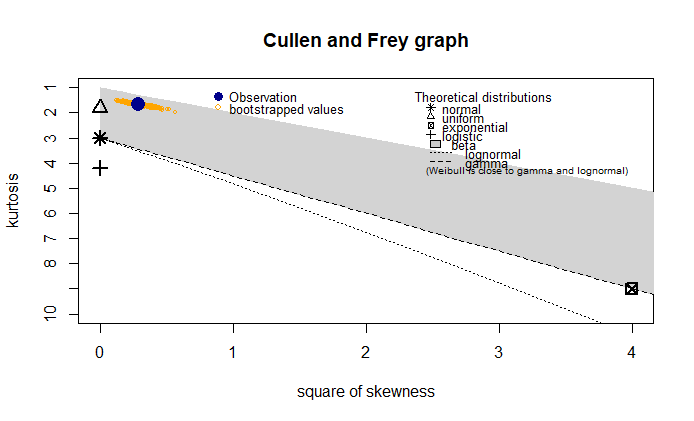
shape1: 0.3776783

shape2: 1.4591991

the beta distribution fits the distribution nicely although there is some irregularity in the pp plot.

**3)FLY ASH**

Analyzing the Cullen Frey graph:



Summary statistics:

min: 0

max: 200.1

median: 0

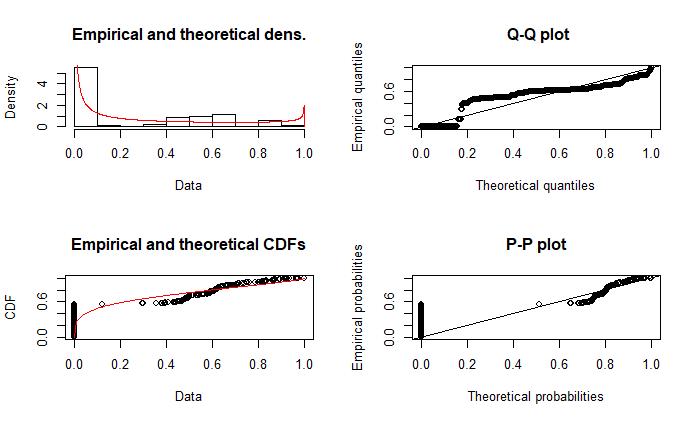
mean: 54.18835

estimated std: 63.997

estimated skewness: 0.5373539

estimated kurtosis: 1.671254

**S**ince the observation lies in the beta region we will try to fit the beta distribution



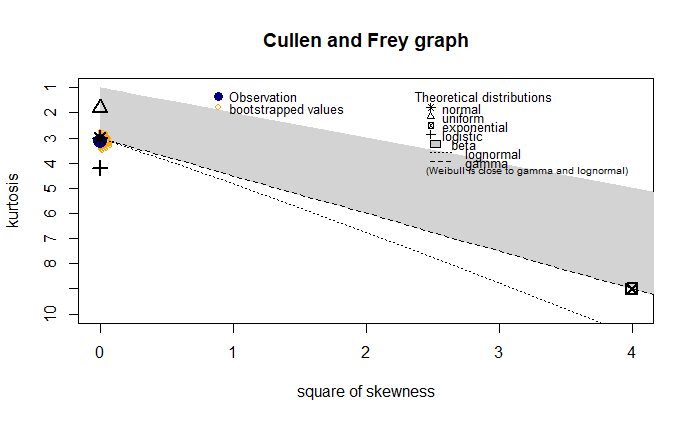
Parameters:

Estimate

shape1: 0.2525016

shape2: 0.6799048

**4)WATER**

Analyzing the Cullen Frey graph

Summary statistics:

min: 121.8

max: 247

median: 185

mean: 181.5673

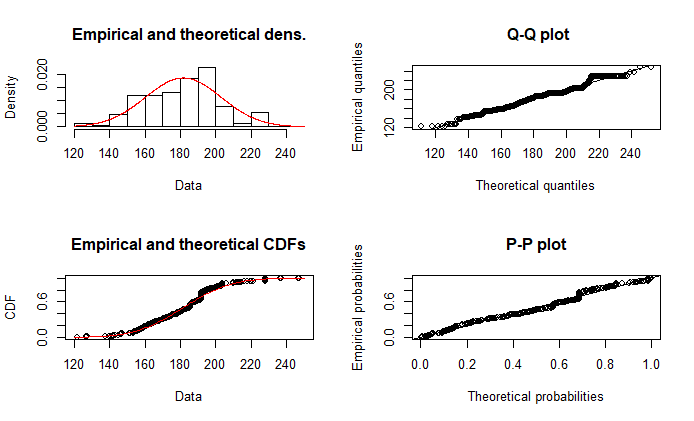
estimated sd: 21.35422

estimated skewness: 0.07462838

estimated kurtosis: 3.122082

The observation lies perfectly in the normal region. we will try fitting the data into the same

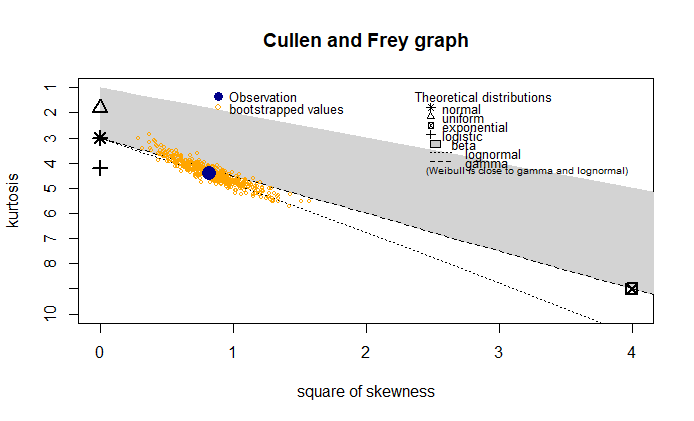
Fitting in normal distribution



The data is normally distributed

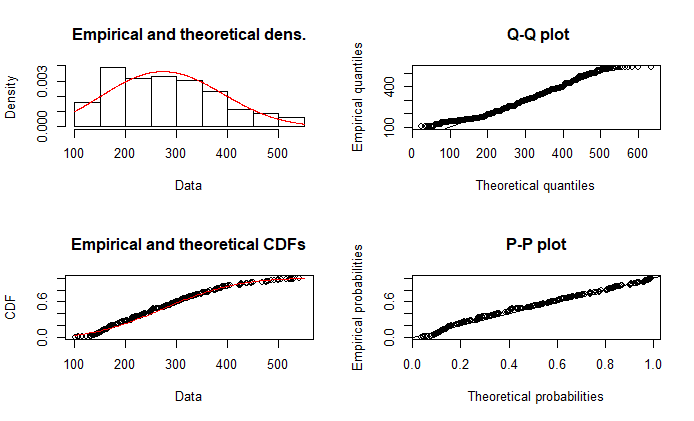
**5)SUPERPLASTICIZER**

Analysis of the Cullen Frey plot:



We will be trying the lognormal and Weibull distribution as the observation lies in between the lognormal and gamma curve.

Fitting to Weibull distribution:

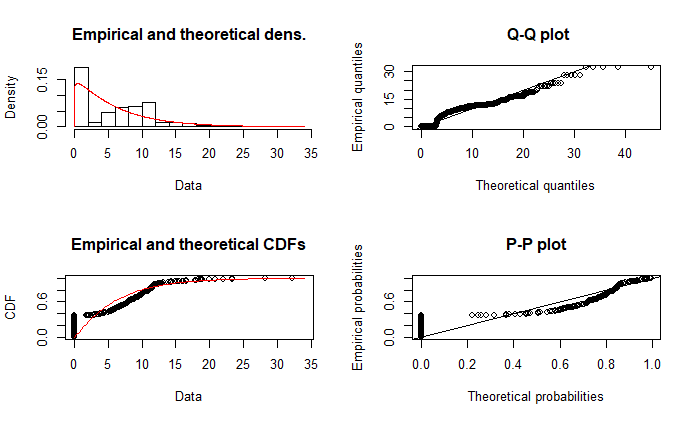


Parameters:

estimate Std. Error

shape 2.910632 0.0692685

scale 315.985410 3.5806575

Fitting the gamma distribution:

Parameters:

estimate

shape 1.0798178

rate 0.1740333

**6)COARSE AGGREGATE**

Analyzing the Cullen Frey graph:

**Summary statistics**

min: 801

max: 1145

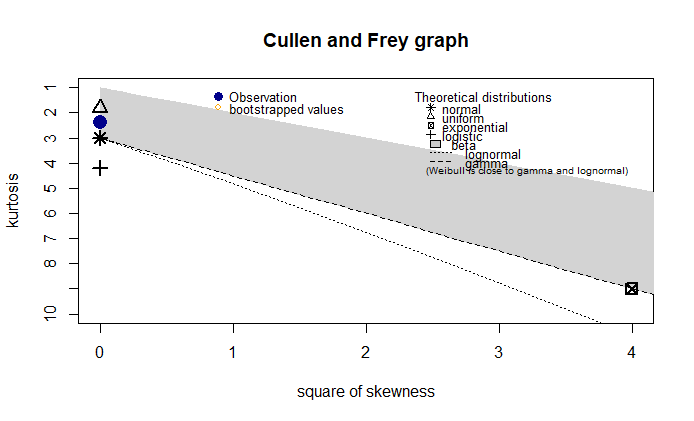
median: 968

mean: 972.9189

estimated sd: 77.75395

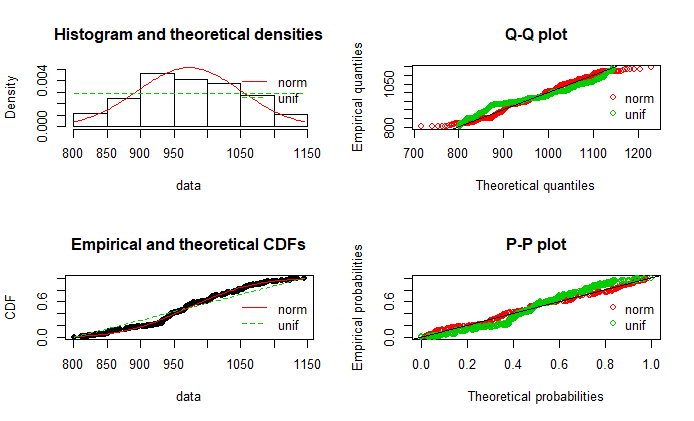
estimated skewness: -0.04021974

estimated kurtosis: 2.400984



from the graph its visible that the observation lies in the beta region. We will try to fit in it the normal and beta distribution along the uniform distribution.

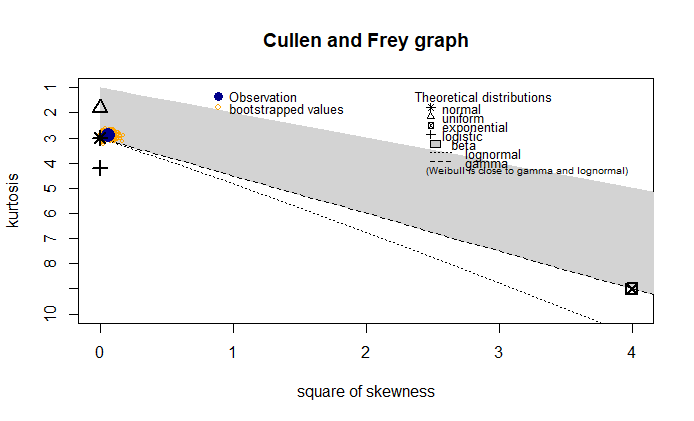
Fitting the normal and uniform distribution:



It can be seen that the normal distribution is a better fit than the uniform distribution

**7)FINE AGGREGATE**

Cullen Frey plot:



Summary statistics:

min: 594 max: 992.6

median: 779.5

mean: 773.5805

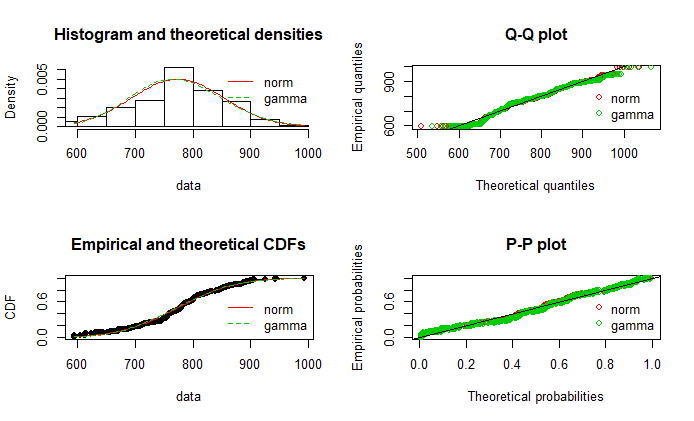
estimated sd: 80.17598

estimated skewness: -0.2530096

estimated kurtosis: 2.897823

The distribution is supposed to follow the normal distribution as the observation from the above plot suggests

Fitting the distributions to the data:



Both the gamma distributions and the normal distributions seem too fit well in the data. However comparing the aic values for the distributions

Normal distribution(AIC)= 11957.51

Gamma distribution(AIC)= 11982.96

Judging by the lower AIC value of normal distribution we fit the data with normal distribution.

**8)AGE:**

Analyzing the Cullen Frey plot. Since Age is discrete variable as observed from the data hence the graph is a bit different.

Summary statistics:

min: 1

max: 365

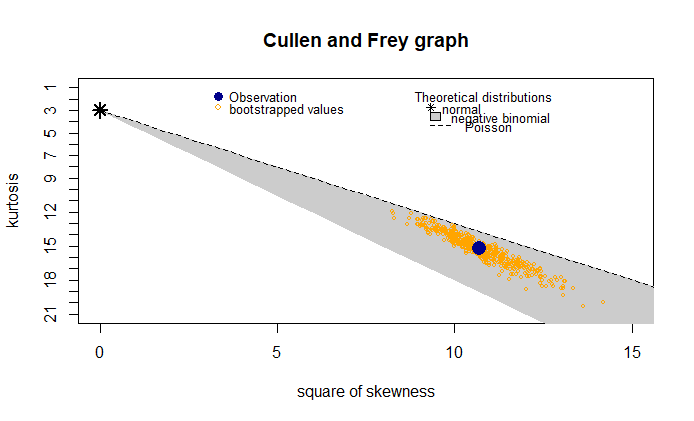
median: 28

mean: 45.66214

estimated sd: 63.16991

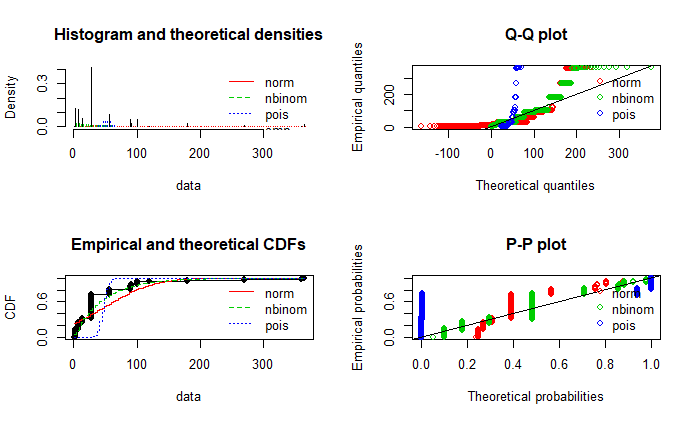
estimated skewness: 3.269177

estimated kurtosis: 15.16899



it is a discrete distribution.

Trying to fit the n-binomial, normal and poisson distribution in the data:



Comparing the AIC values of the three distributions

Normal distribution 11466.42

Negative Binomial distribution 9952.473

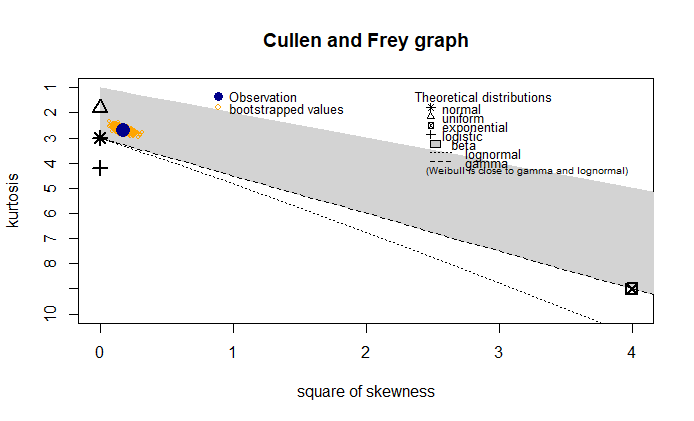
Poisson distribution 61728.4

Judging the lowest AIC value of n-binomial distribution, it’s supposed to be the best fit for “age” feature.

**9)CONCRETE COMPRESSIVE STRENGTH**

This is the **target feature** of the data, the one which we want to predict and analyze w.r.t to all the features.

Analyzing the Cullen Frey plot:



Summary statistics

min: 2.33

max: 82.6

median: 34.445

mean: 35.81796

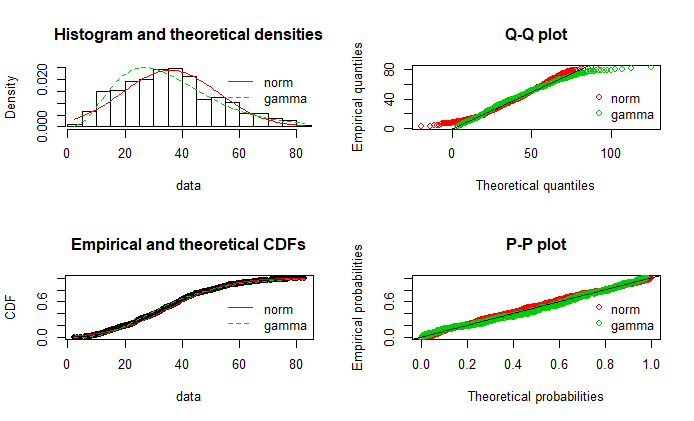
estimated std: 16.70574

estimated skewness: 0.4169773

estimated kurtosis: 2.686275

The observation point lies close to the normal distribution point while falling into the beta distribution region.

Fitting distributions into the target feature:



Comparing the AIC values:

Normal distribution 8726.463

Gamma distribution 8690.263

Judging by the distributions it is visible that the Gamma distribution is a better fit