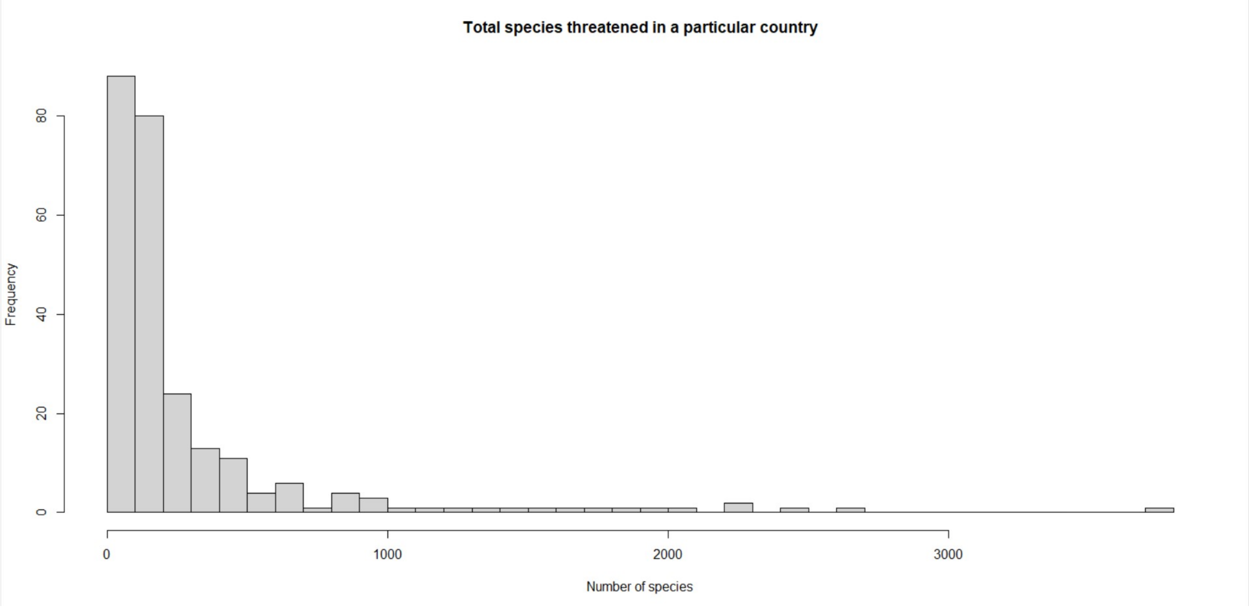


Figure 8: Histogram for given data

Using the packages 'MASS' and 'fitdistrplus' in R software, fitting of a distribution was carried out.(Refer Appendix)

The density curve and the corresponding cumulative distribution for the given data were plotted to get a general idea of the distribution.

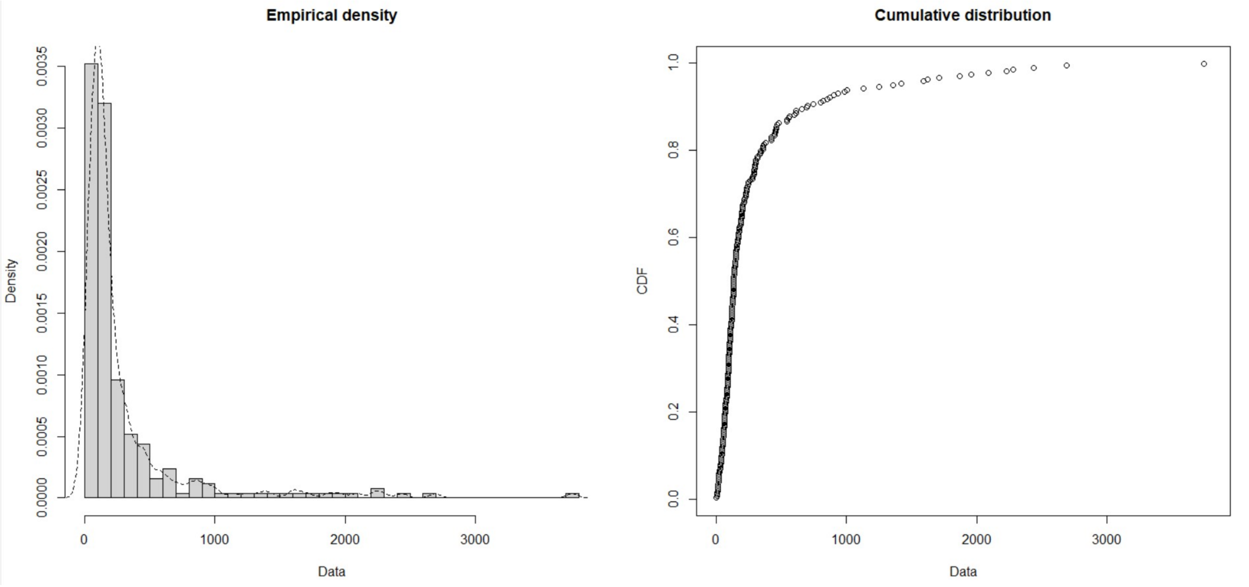
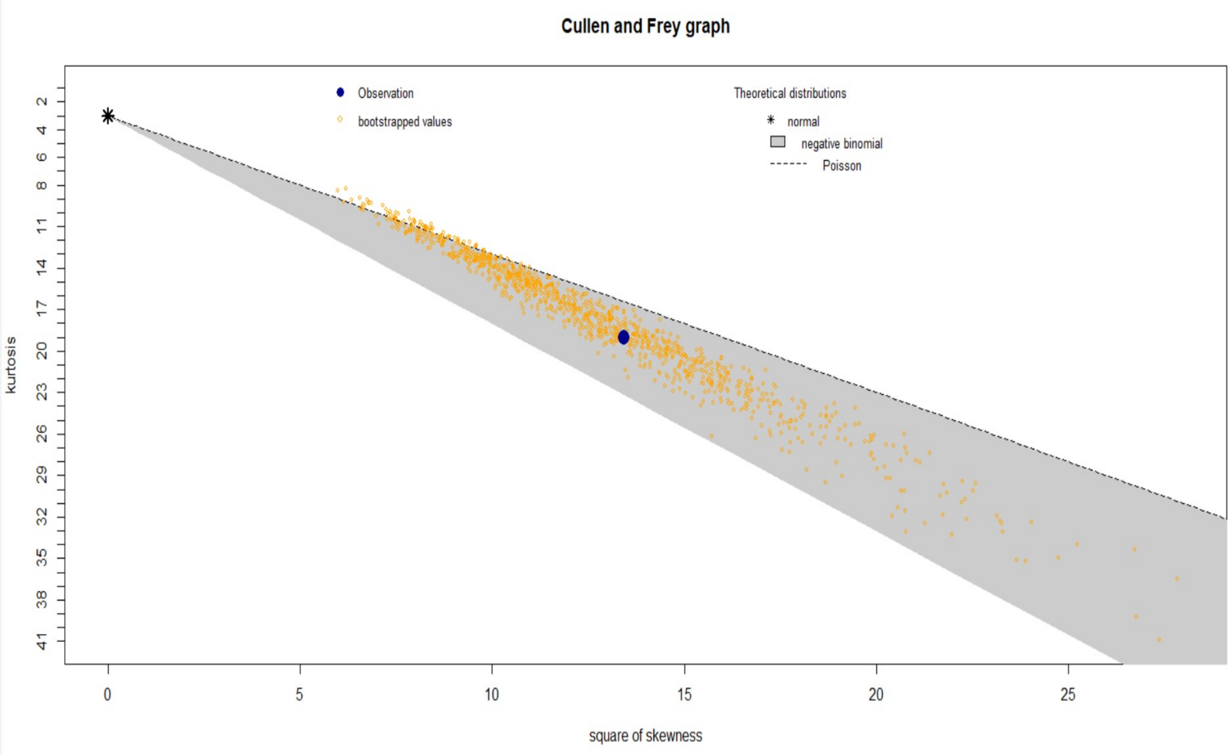
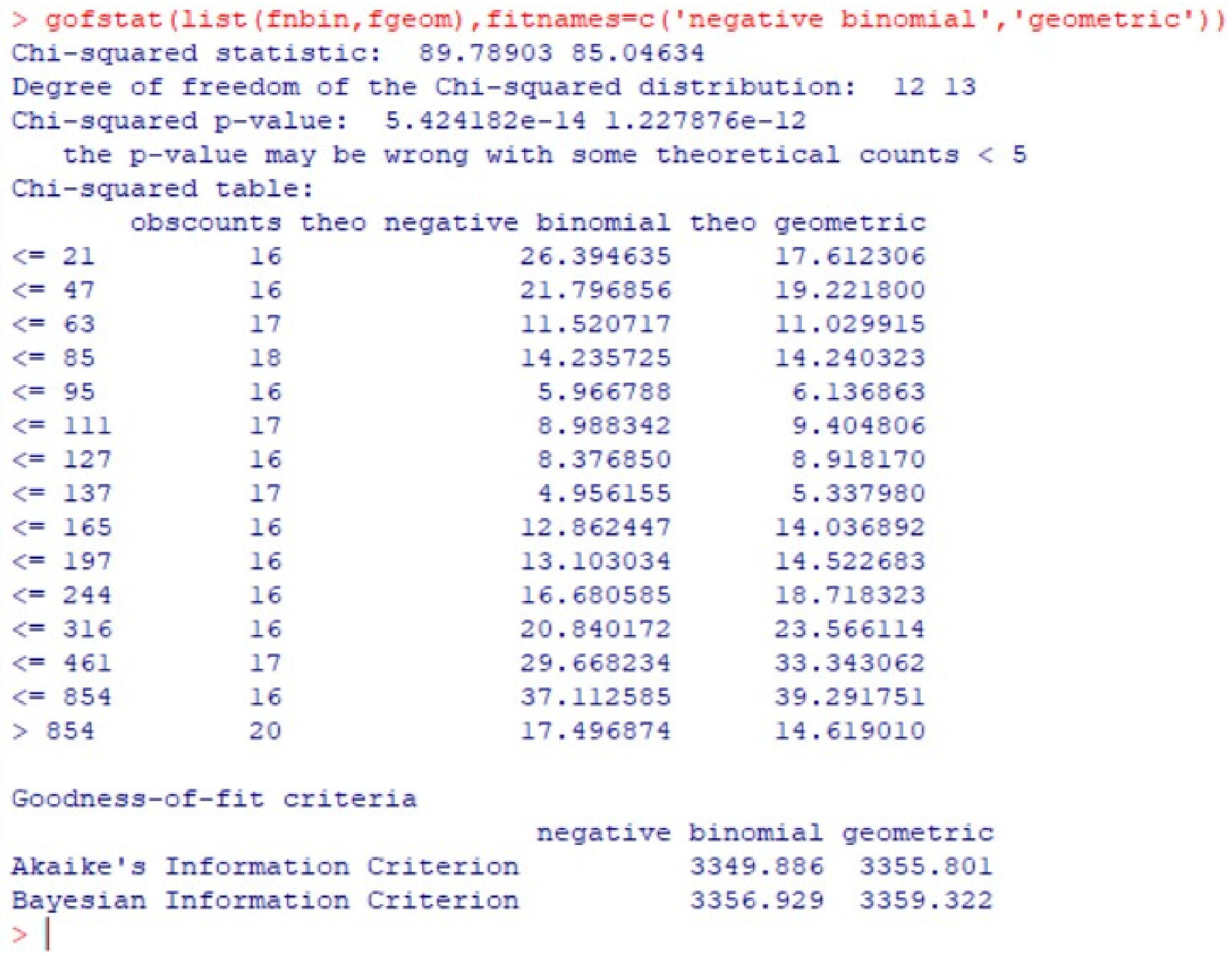


Figure 9: Density and PMF plot

To estimate best distribution to be fitted, we make use of the Cullen-Frey graph.



Thus it can be observed that, the observations(blue dot) and a majority of the bootstrapped values lie in the grey shaded region, indicating that our data might be following a negative binomial distribution.



```
> gofstat(list(fnbin,fgeom),fitnames=c('negative binomial','geometric'))
Chi-squared statistic:  89.78903 85.04634
Degree of freedom of the Chi-squared distribution:  12 13
Chi-squared p-value:  5.424182e-14 1.227876e-12
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo negative binomial theo geometric
<= 21        16                26.394635      17.612306
<= 47        16                21.796856      19.221800
<= 63        17                11.520717      11.029915
<= 85        18                14.235725      14.240323
<= 95        16                 5.966788       6.136863
<= 111       17                 8.988342       9.404806
<= 127       16                 8.376850       8.918170
<= 137       17                 4.956155       5.337980
<= 165       16                12.862447      14.036892
<= 197       16                13.103034      14.522683
<= 244       16                16.680585      18.718323
<= 316       16                20.840172      23.566114
<= 461       17                29.668234      33.343062
<= 854       16                37.112585      39.291751
> 854        20                17.496874      14.619010

Goodness-of-fit criteria
                               negative binomial geometric
Akaike's Information Criterion          3349.886  3355.801
Bayesian Information Criterion          3356.929  3359.322
> |
```

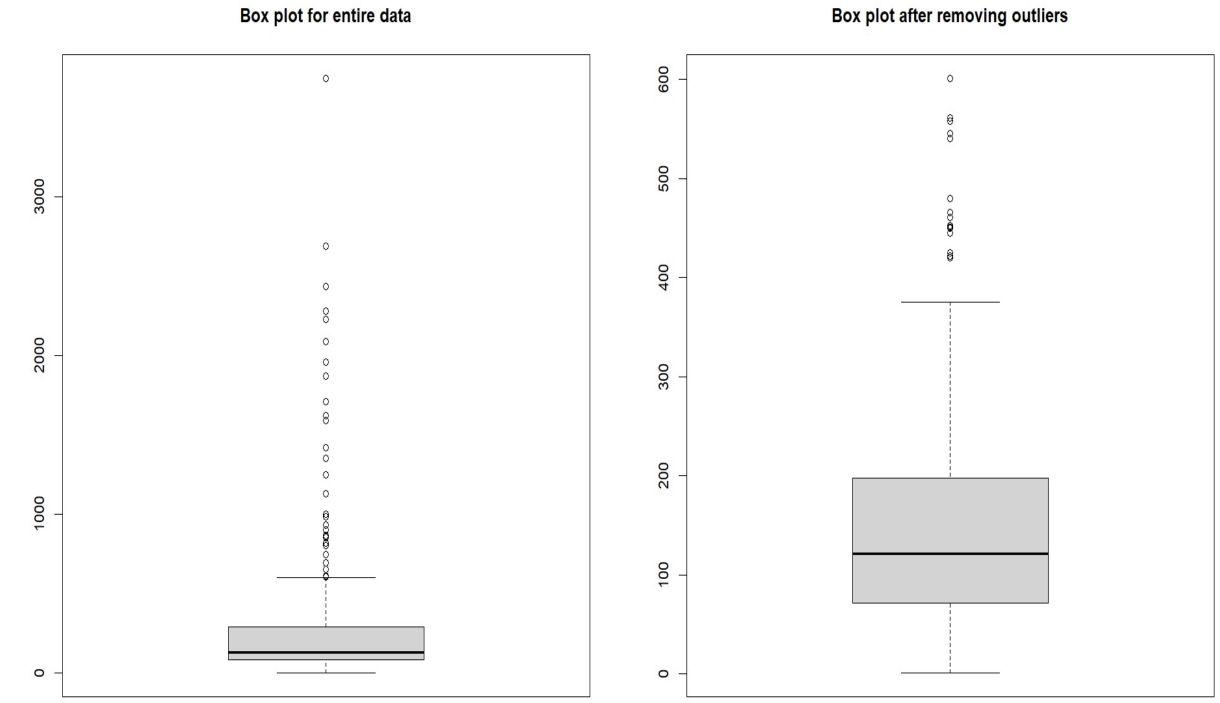
However, here the p-value is around zero, which suggests neither fit is good

Reason:- The main reason for this negligible p-value is the presence of too many outliers. Referring to the histogram for the data, it is evident that there are many values which heavily distort the measures of central tendency and dispersion.

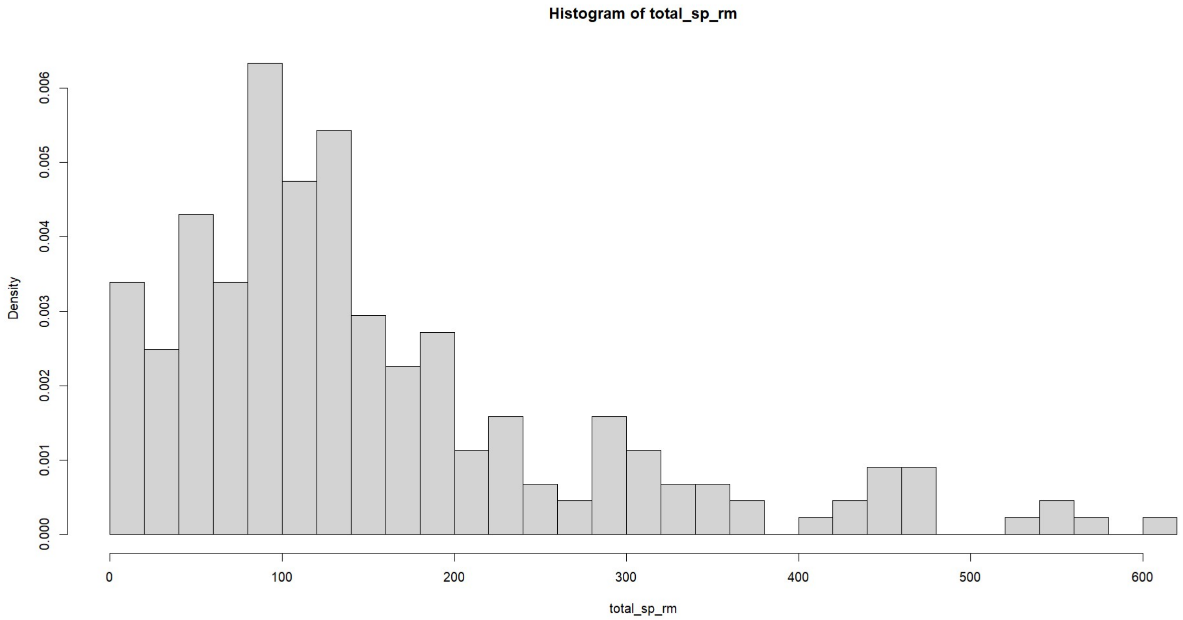
To cater to this problem, we shall fit a model after removing its outliers. The procedure is very similar to that used above.

Detection of outliers:- The easiest way of detecting outliers, is by drawing a box plot of the data. Here we compare the box plots of the data before and after removing the outliers.(Refer Appendix for R program).

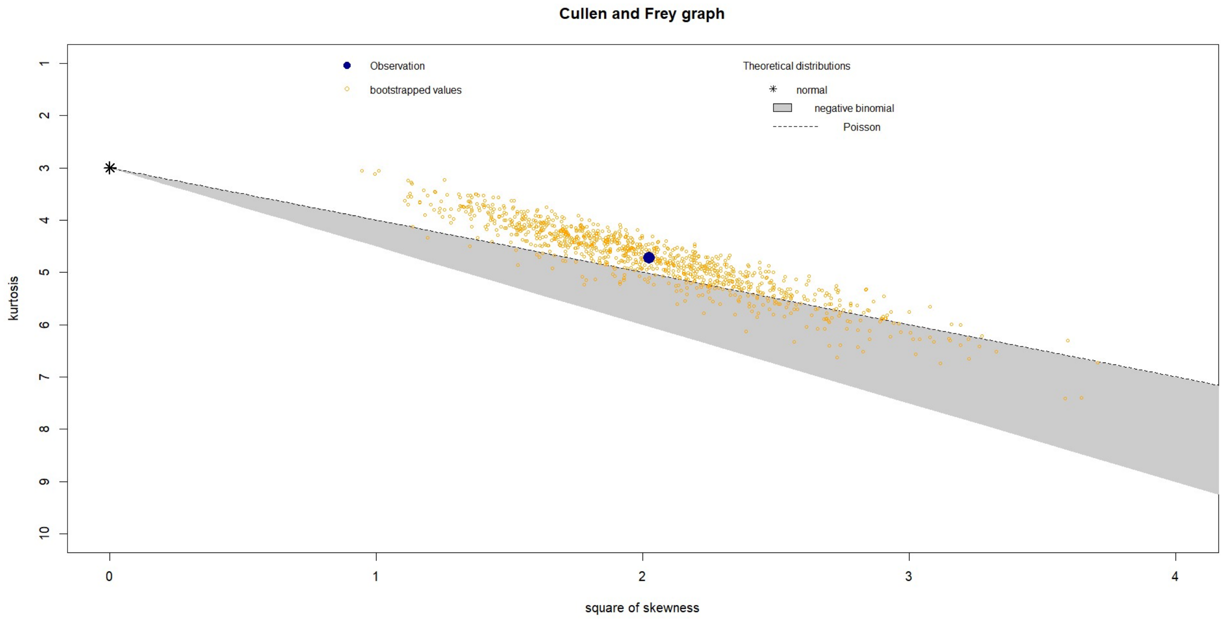
A total of 29 observations have been outliers and they have been removed to increase the goodness/reliability of the model.



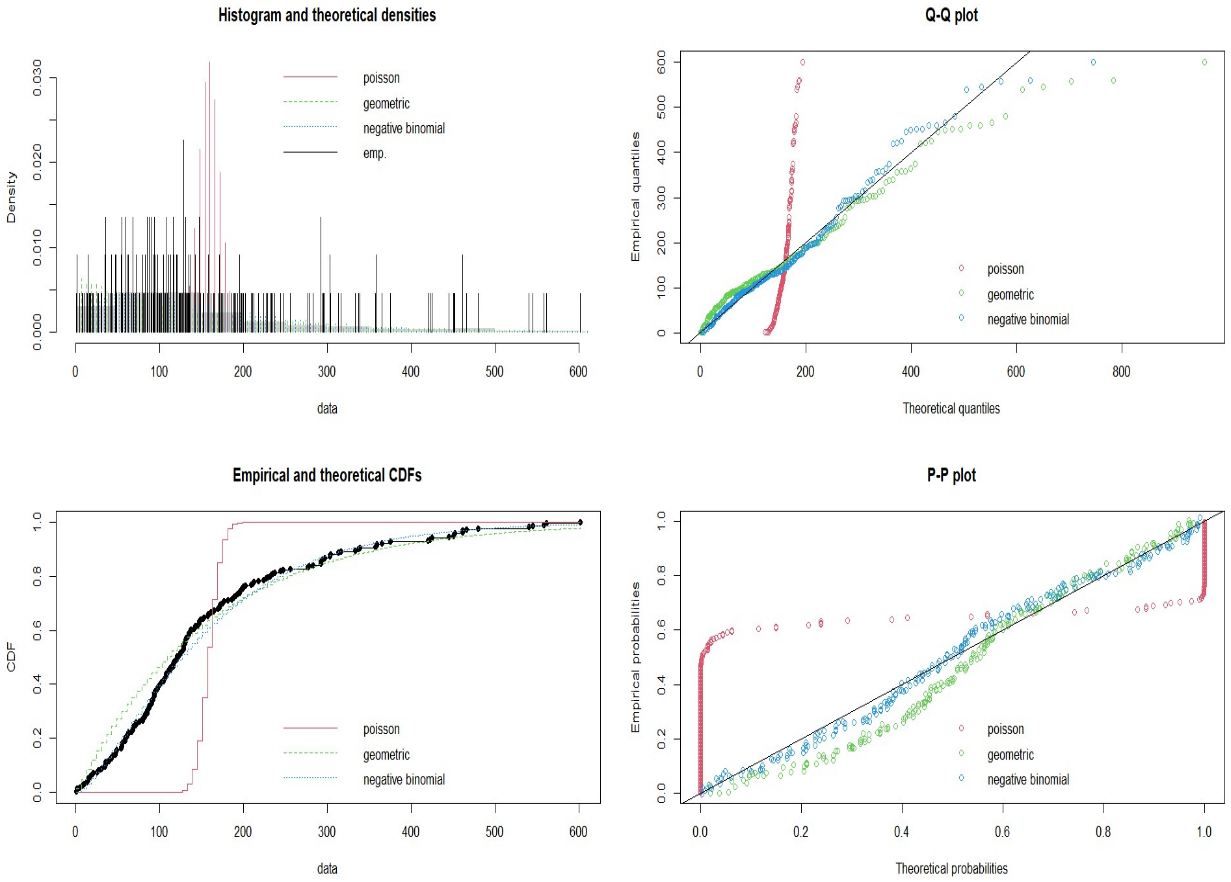
Thus, the corresponding histogram after removal of outliers was obtained.



Thus, the Cullen Frey graph for best fit(given below) suggests negative binomial and poisson distributions to be approximately good.

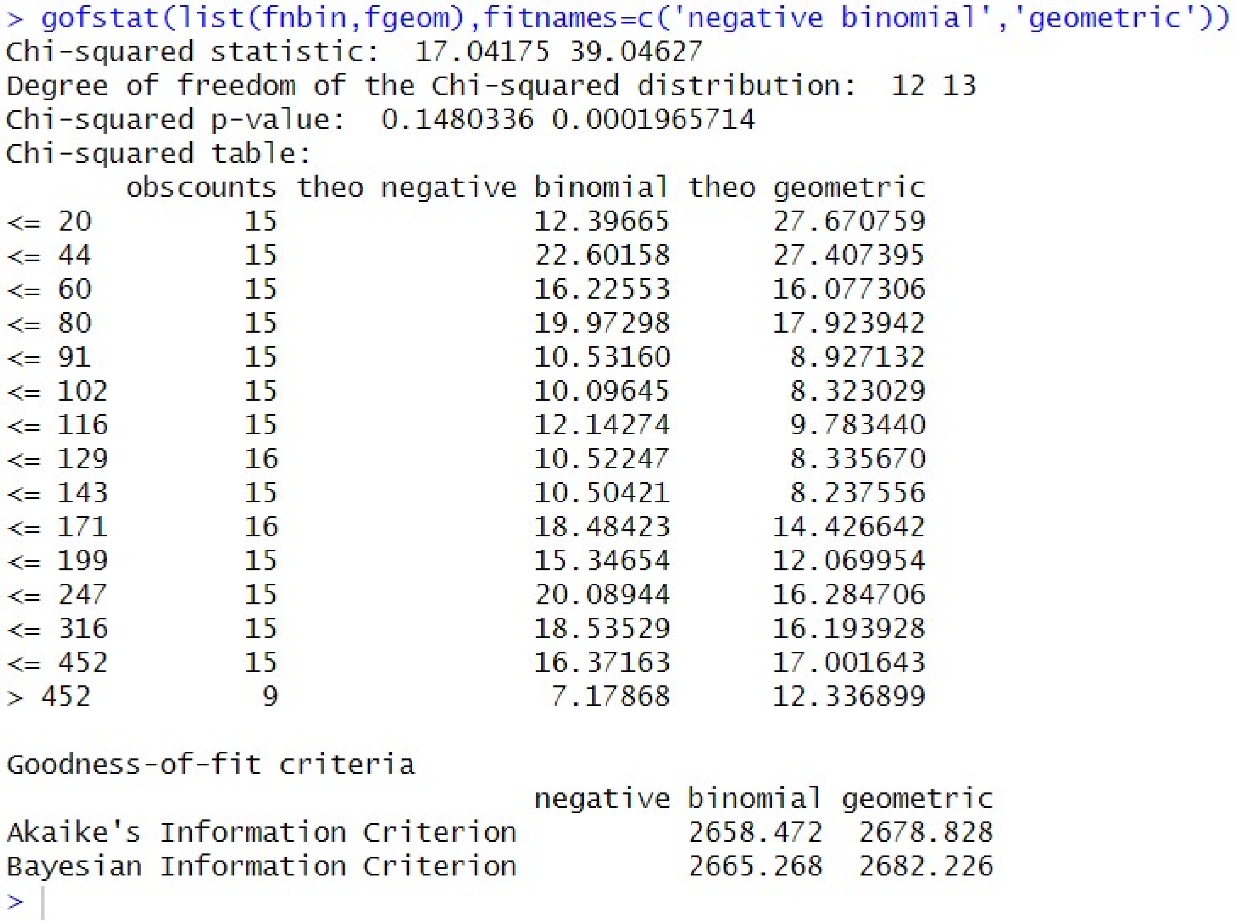


To get a better idea, we make use of the diagnostic plots as done earlier.



Thus it can be observed from the Q-Q,P-P plots and cumulative distribution plot that the negative binomial very accurately describes the data.

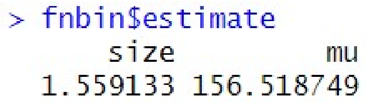
Hence it is important to now check the goodness of fit for fitted models(namely, negative binomial and geometric). The output has been attached for reference.



```
> gofstat(list(fnbin,fgeom),fitnames=c('negative binomial','geometric'))
Chi-squared statistic:  17.04175 39.04627
Degree of freedom of the Chi-squared distribution:  12 13
Chi-squared p-value:  0.1480336 0.0001965714
Chi-squared table:
       obscounts theo negative binomial theo geometric
<= 20         15              12.39665      27.670759
<= 44         15              22.60158      27.407395
<= 60         15              16.22553      16.077306
<= 80         15              19.97298      17.923942
<= 91         15              10.53160       8.927132
<= 102        15              10.09645       8.323029
<= 116        15              12.14274       9.783440
<= 129        16              10.52247       8.335670
<= 143        15              10.50421       8.237556
<= 171        16              18.48423      14.426642
<= 199        15              15.34654      12.069954
<= 247        15              20.08944      16.284706
<= 316        15              18.53529      16.193928
<= 452        15              16.37163      17.001643
> 452          9               7.17868      12.336899

Goodness-of-fit criteria
                               negative binomial geometric
Akaike's Information Criterion          2658.472  2678.828
Bayesian Information Criterion          2665.268  2682.226
> |
```

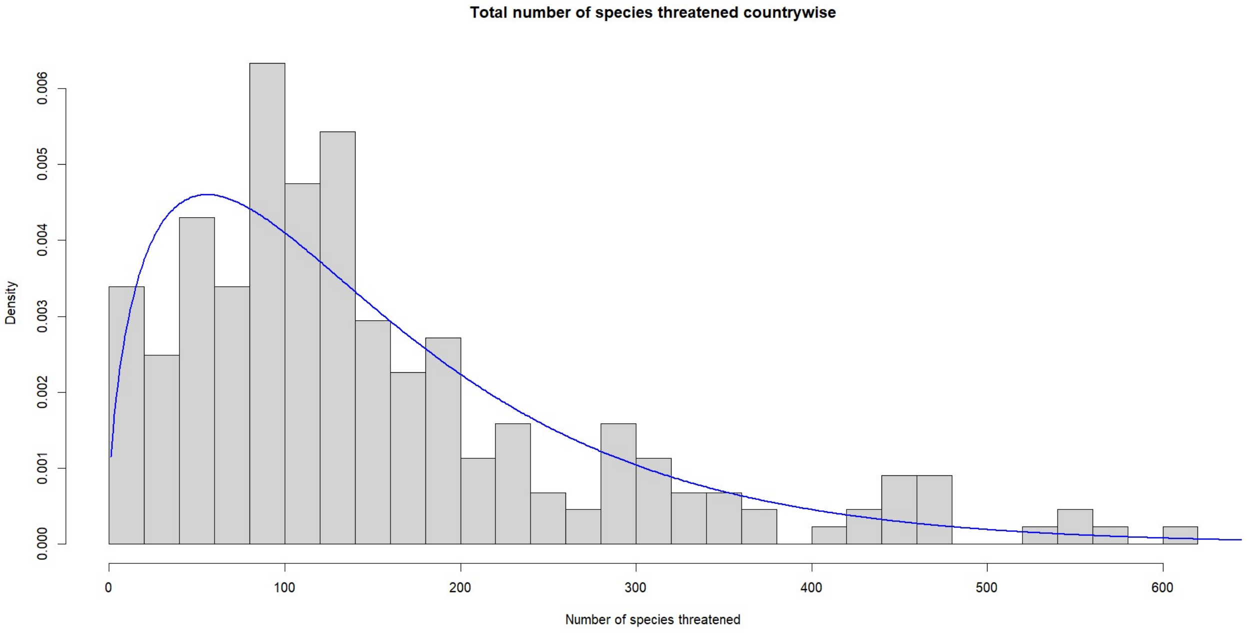
The observed p-value for the model of negative binomial distribution is 0.1480336, while that for geometric distribution is 0.0001965714, clearly suggesting that the negative binomial distribution is a good fit for the given data, as per the chi-square goodness of fit test.
The estimated parameters for the distribution are:-



```
> fnbin$estimate
      size         mu
  1.559133 156.518749
```

That is, the parameters of the negative binomial distribution are:- k= $1.559133 \approx 2$ and p=0.0098631 (or mean = 156.518749), that is

$$X \sim NB(k = 2, p = 0.0098631)$$



Why a negative binomial distribution?

A major indication that the distribution could be following a negative binomial distribution (and not a poisson) was the relation between mean and variance. Here, variance was larger than the mean clearly indicating that the distribution was either geometric or negative binomial(poisson distribution was fitted to the data for demonstration purpose).

Additionally, negative binomial distribution has a wide range of applications in the fields of biology and ecology, further making this a model that can be implemented in our case.

This model can help identifying particular countries that have a higher count of threatened species and those which need targeted conservation efforts further assisting in policy making and decision implementing for conserving a particular species.