Unsupervised learning exercise from HSAUR

Yamuna Dhungana

1. (Ex. 8.1 in HSAUR, modified for clarity) The **galaxies** data from **MASS** contains the velocities of 82 galaxies from six well-separated conic sections of space (Postman et al., 1986, Roeder, 1990). The data are intended to shed light on whether or not the observable universe contains superclusters of galaxies surrounded by large voids. The evidence for the existence of superclusters would be the multimodality of the distribution of velocities.(8.1 Handbook)

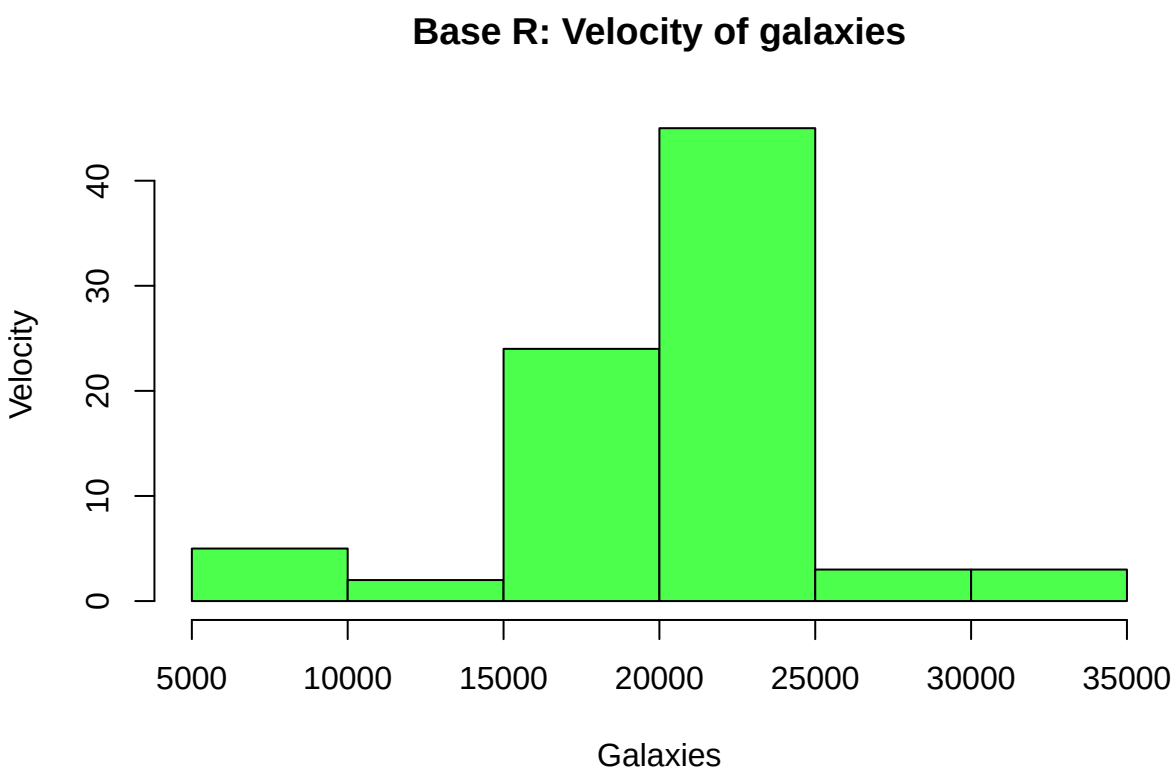
   a) Construct histograms using the following functions:

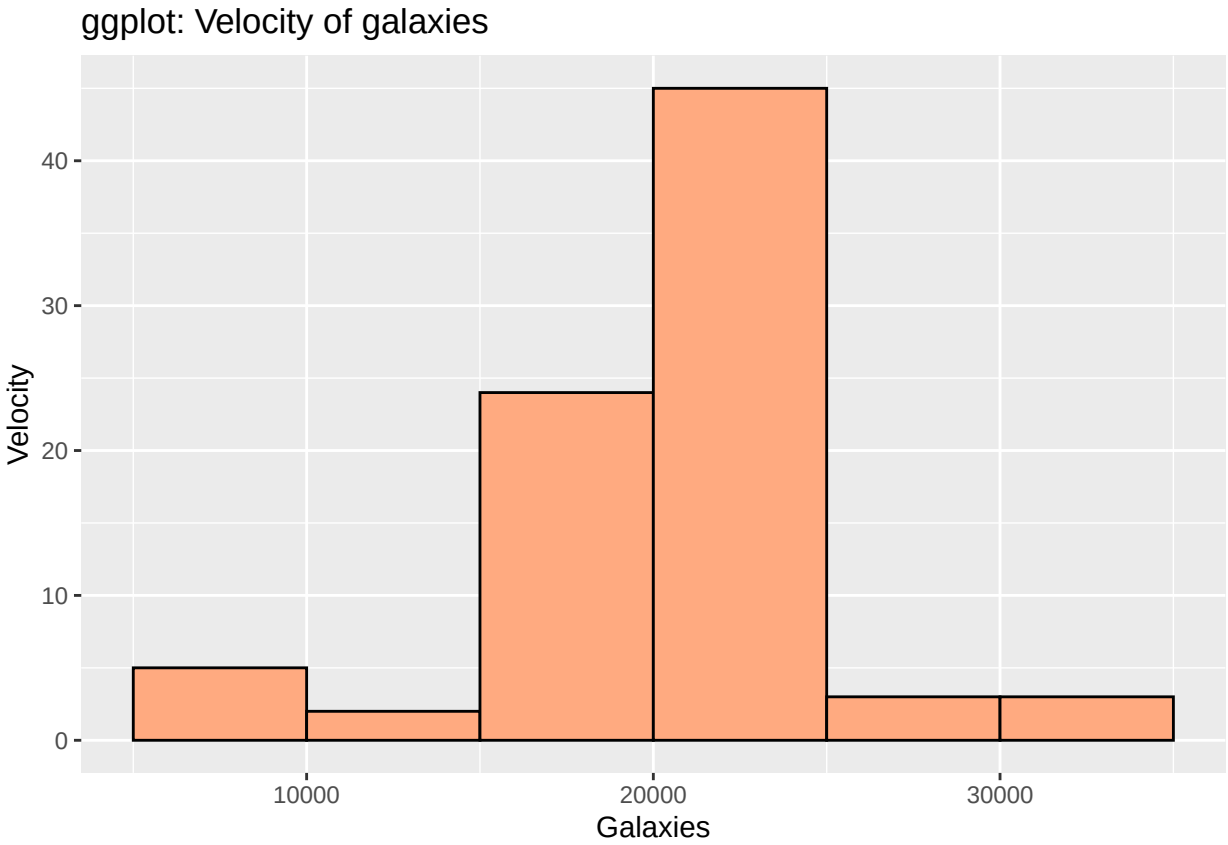
   -hist() and ggplot()+geom_histogram()

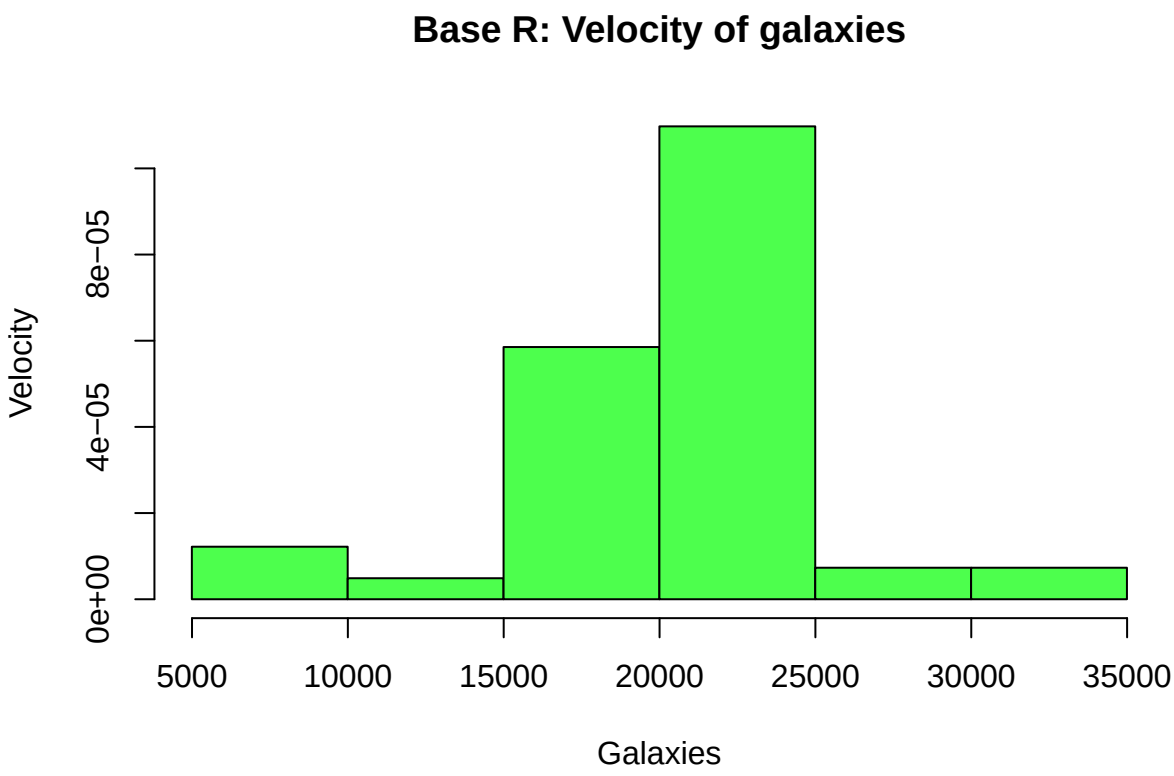
   -truehist() and ggplot+geom_histogram() (make sure that the histograms show proportions, not counts.)

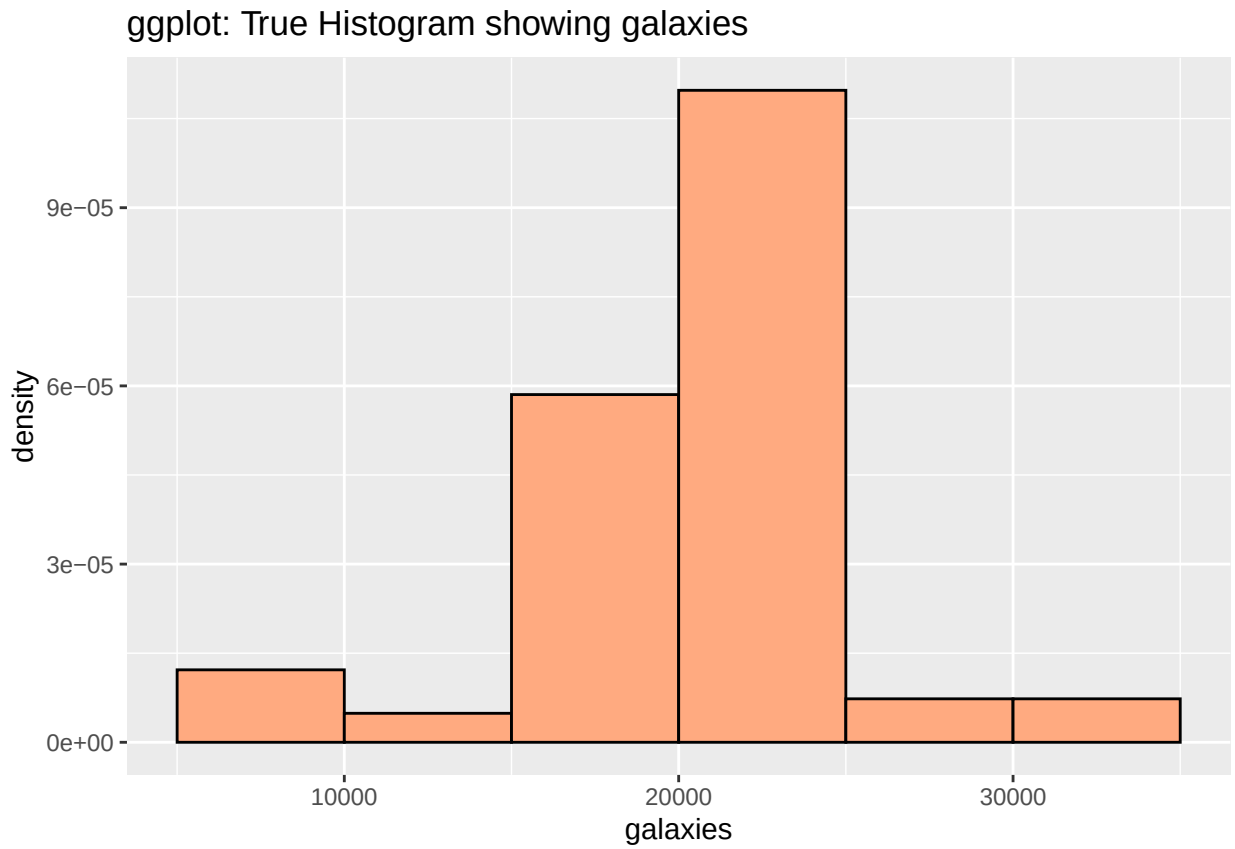
   -qplot()

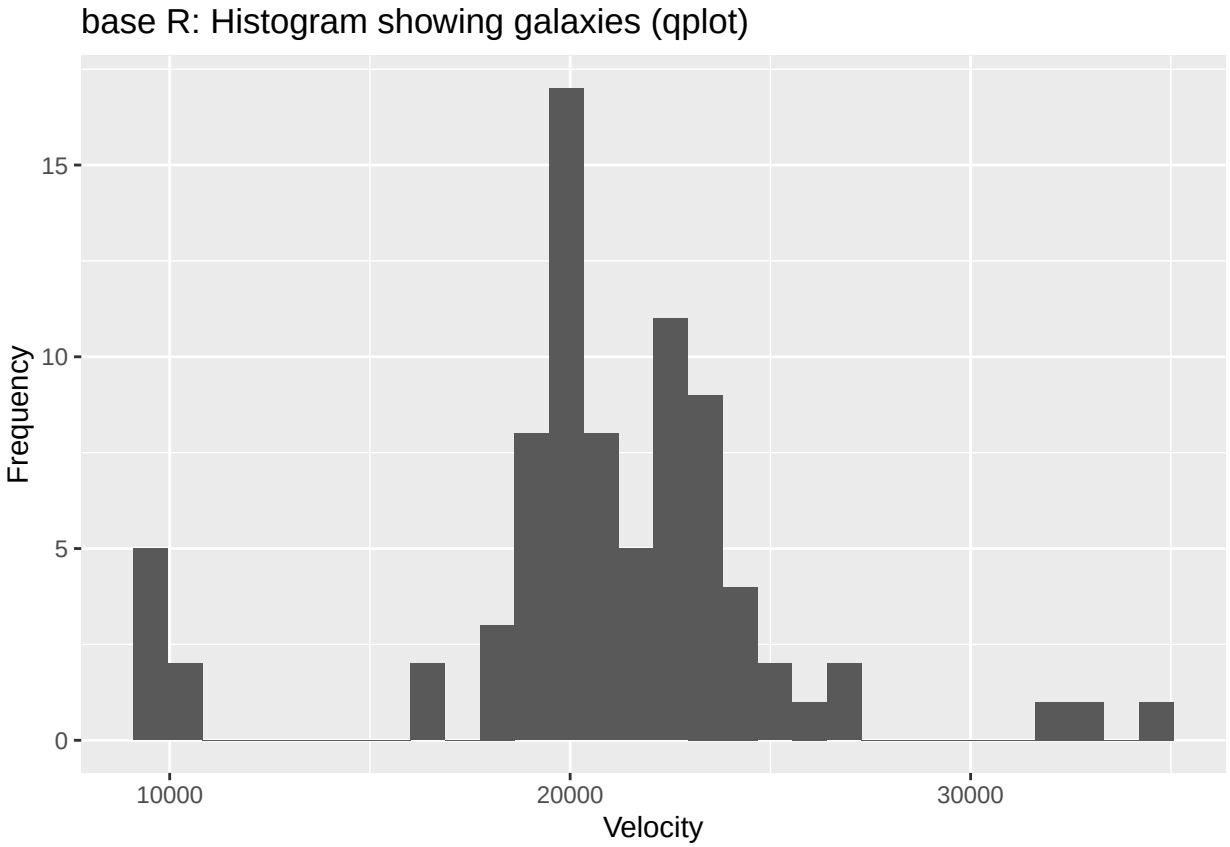
   Comment on the shape and properties of the variable based on the five plots. Do you notice any sets of observations clustering? (Hint: You can adjust bin number or bin size as you try to determine the properties of the variable, but use the same bin settings between plots in your final analysis. You can also overlay the density function or use the rug command.)

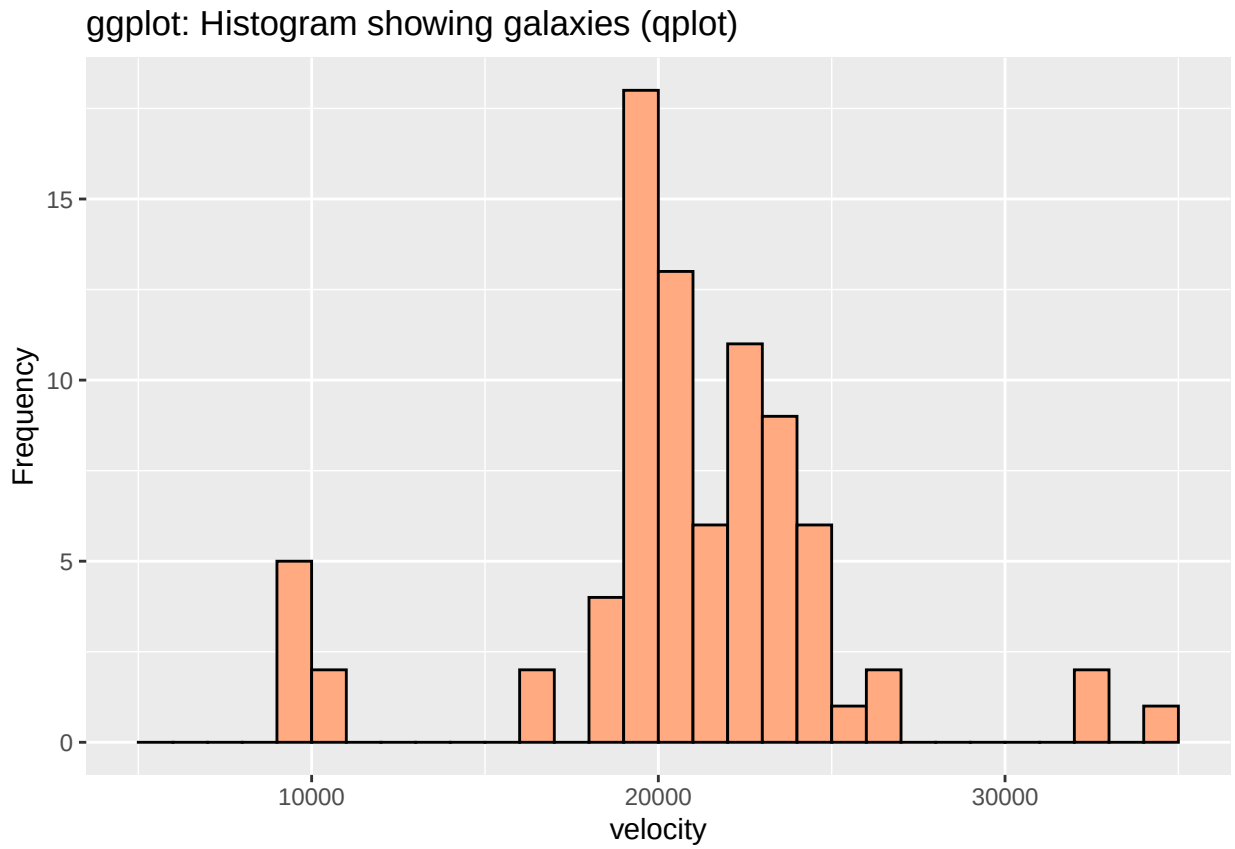



ggplot: Velocity of galaxies
Velocity
40
30
20
10
0
10000
20000
30000
Galaxies


Base R: Velocity of galaxies
Velocity
0e+00
4e−05
8e−05
5000
10000
15000
20000
25000
30000
35000
Galaxies

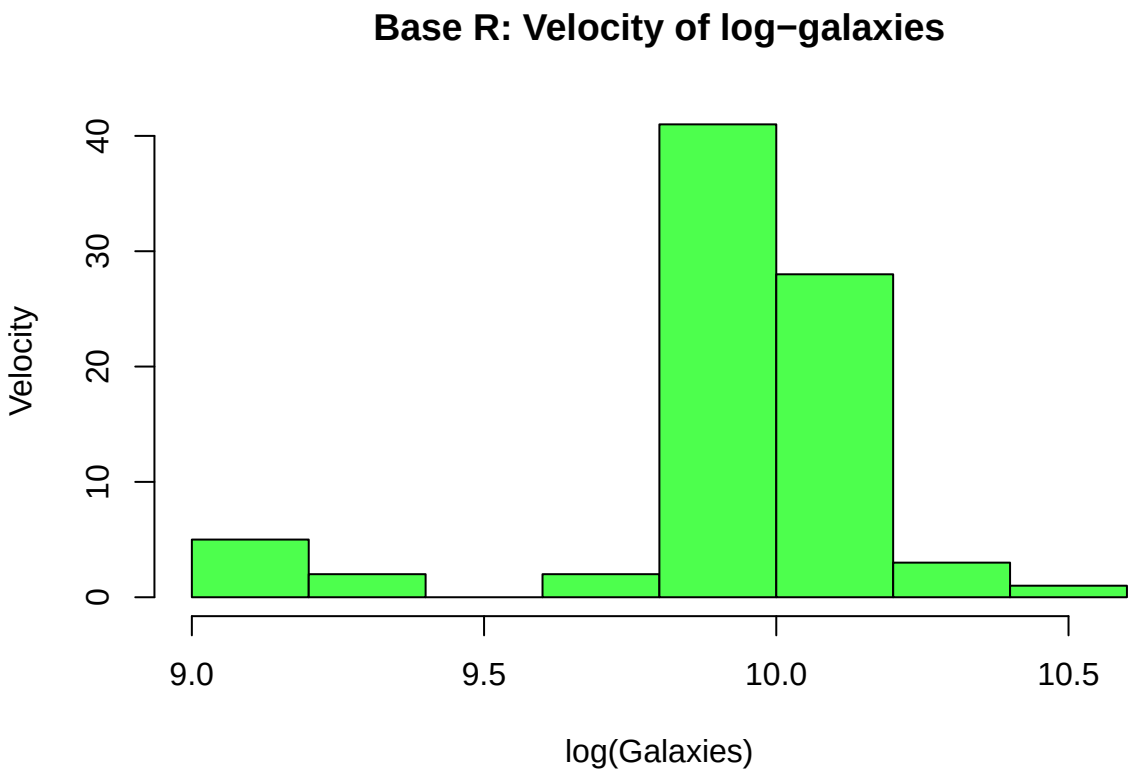

ggplot: True Histogram showing galaxies
density
9e−05
6e−05
3e−05
0e+00
10000
20000
30000
galaxies

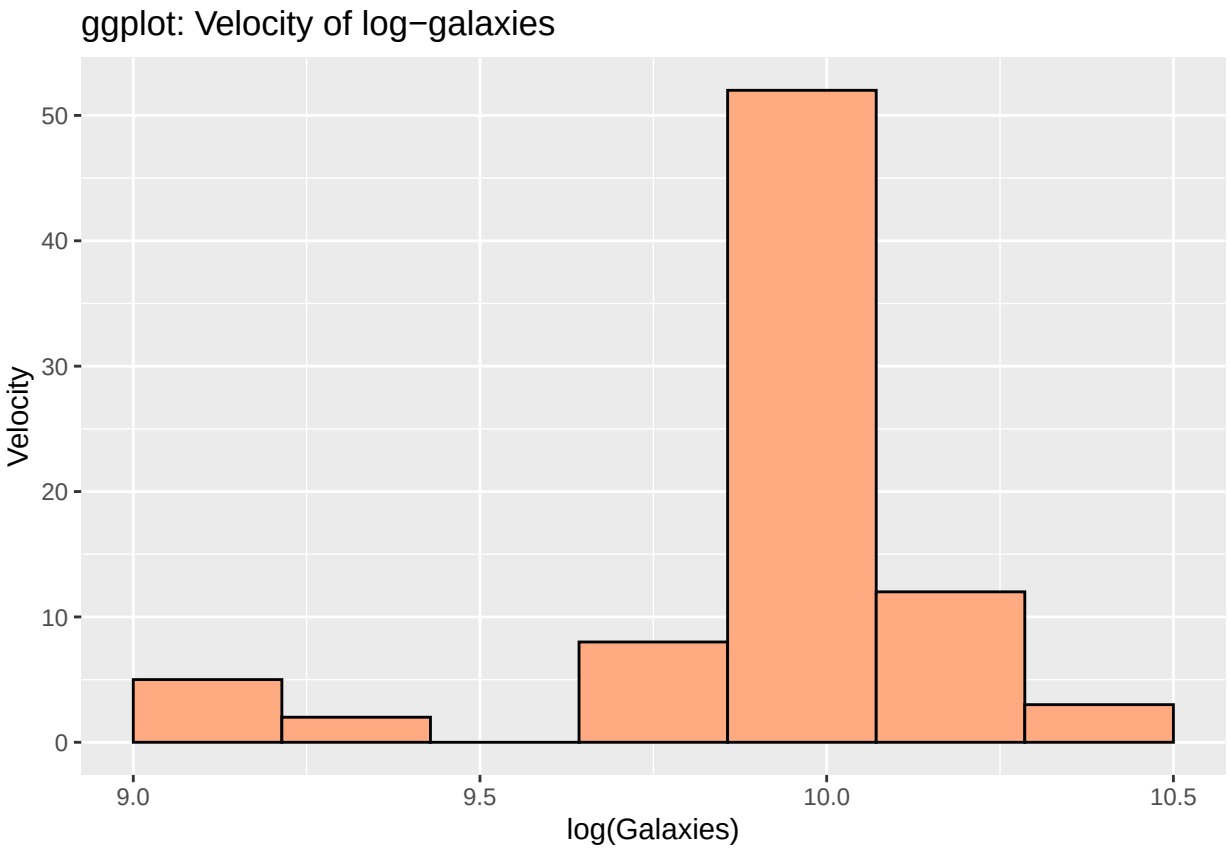

base R: Histogram showing galaxies (qplot)
Frequency
15
10
5
0
10000
20000
30000
Velocity

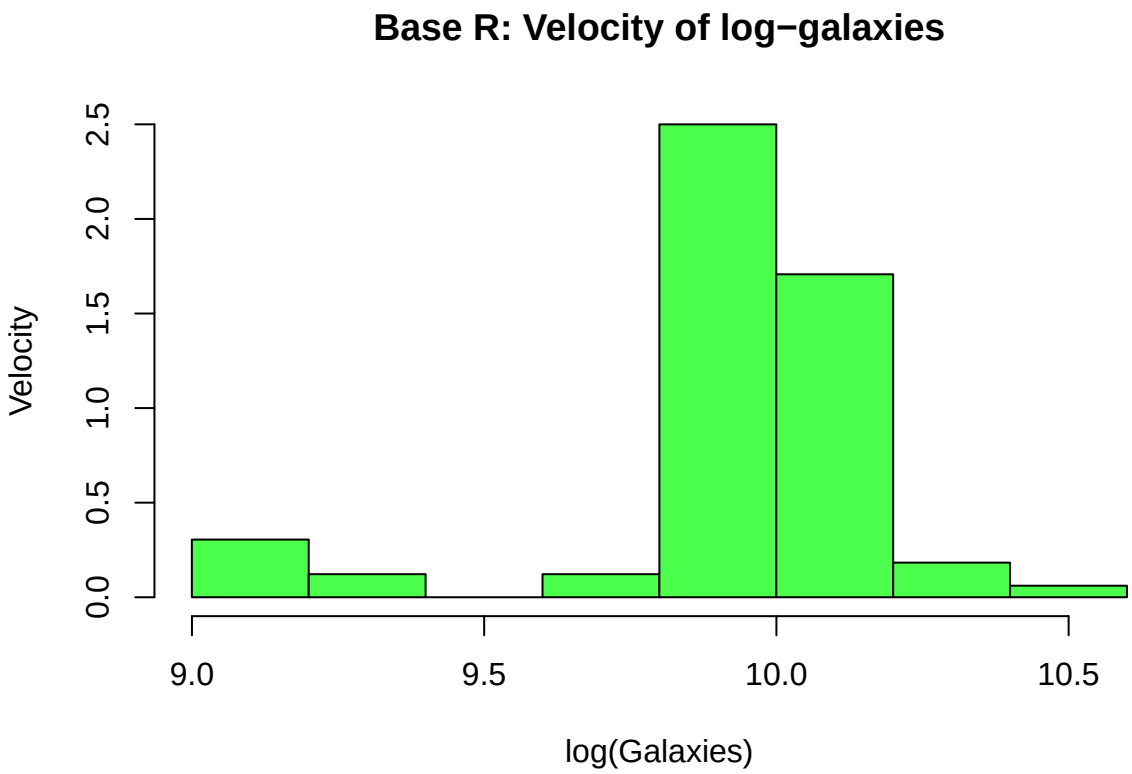


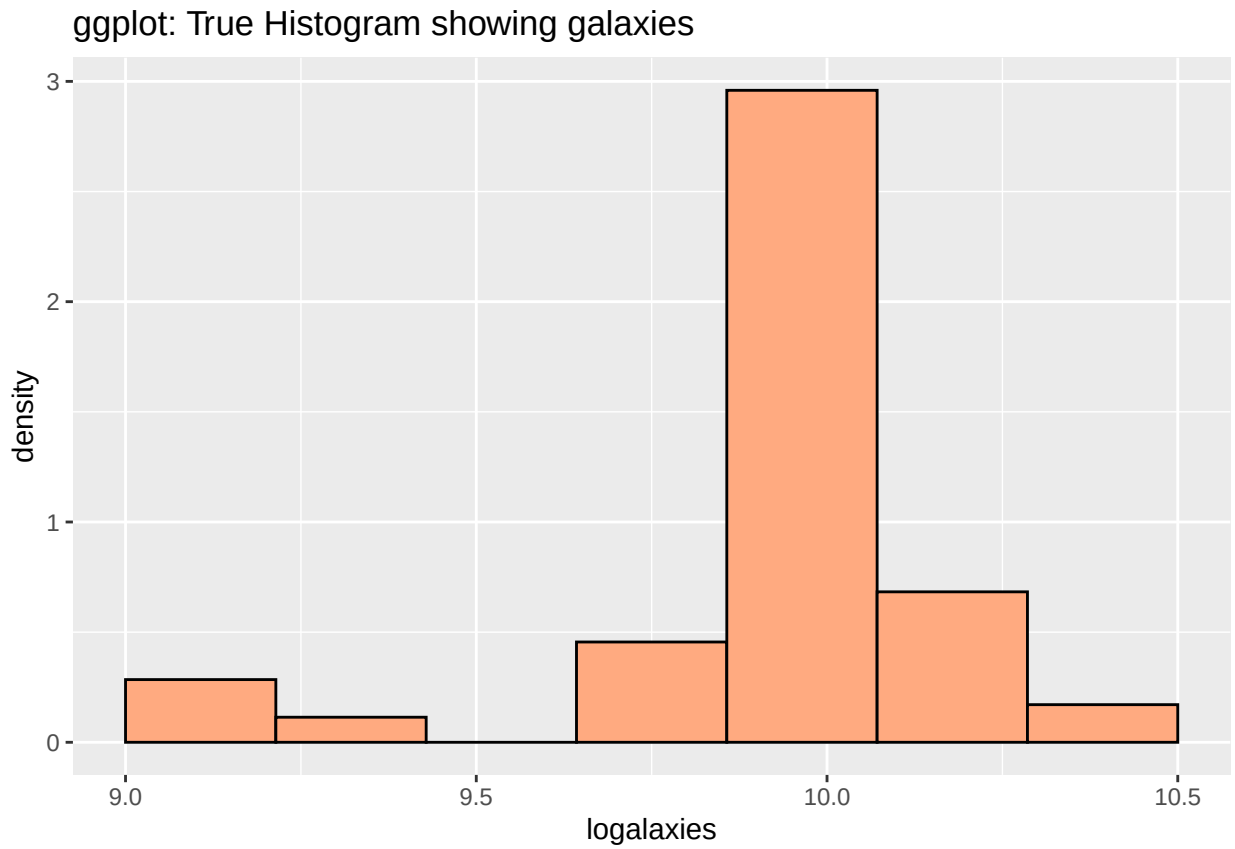
From the graphs, it appears that all the graphs for hist and true-hist appear to be the same. The only difference from the graphs is the value of the y-axis. Histograms give the frequency. Whereas, true-hist gives the probability of the hist. We can see the clustering group clustered in the middle. We also can see some clusters on either side of the cluster. We can say that there are three clusters, but in the plot, there is an extra one cluster in the middle cluster. Therefore we can assume there ate four clusters in the galaxies data.

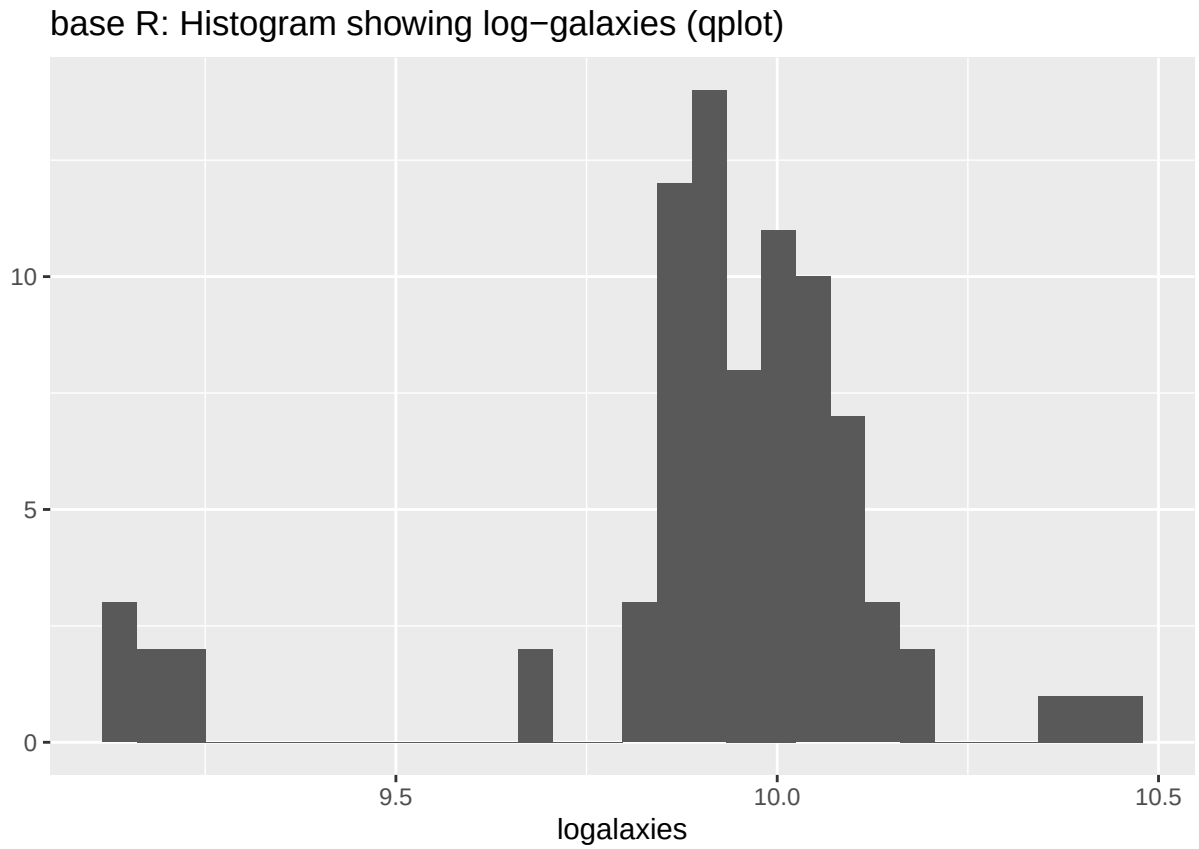
b) Create a new variable *loggalaxies* = log(galaxies). Repeat part a) using the `loggalaxies` variable. Does this affect your interpretation of the graphs?

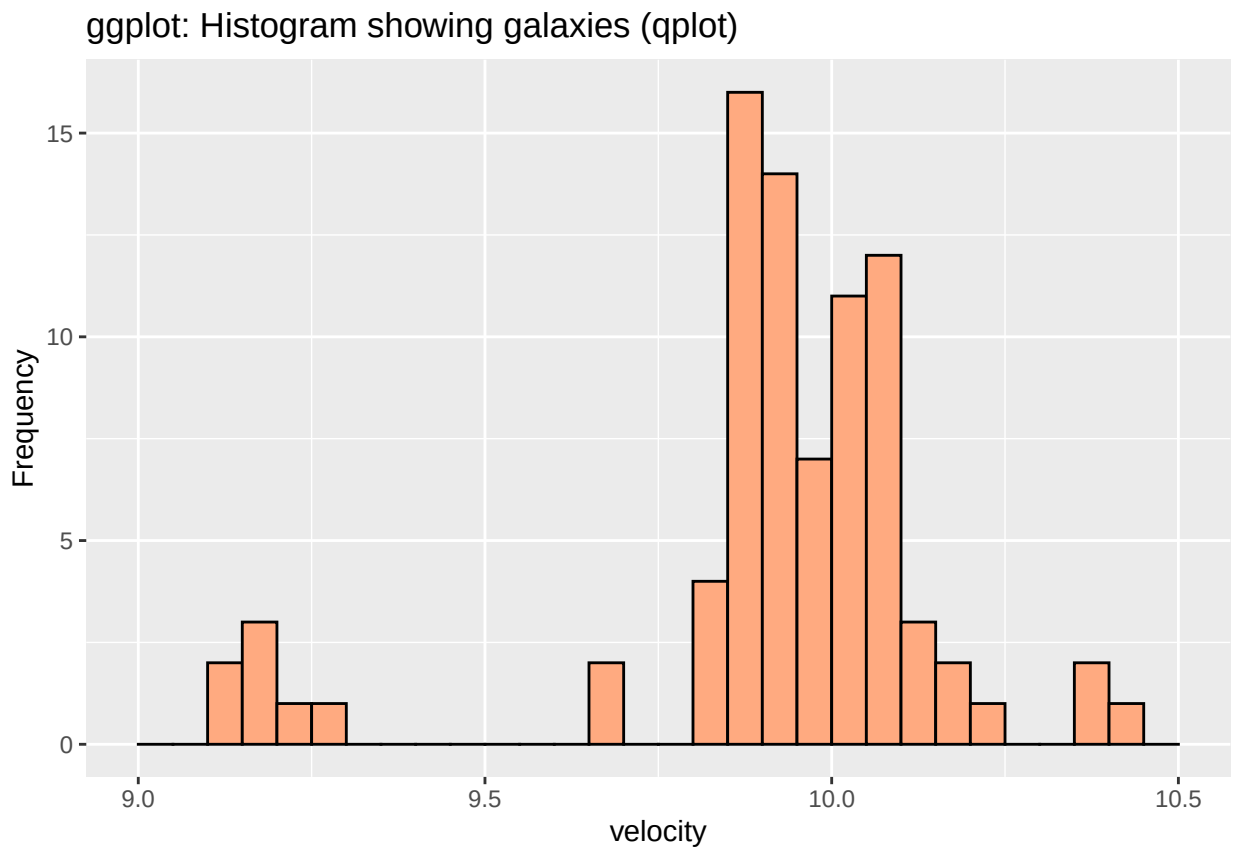

Base R: Velocity of log−galaxies
Velocity
0
10
20
30
40
9.0
9.5
10.0
10.5
log(Galaxies)


ggplot: Velocity of log−galaxies
Velocity
50
40
30
20
10
0
9.0
9.5
10.0
10.5
log(Galaxies)


Base R: Velocity of log−galaxies
Velocity
0.0
0.5
1.0
1.5
2.0
2.5
9.0
9.5
10.0
10.5
log(Galaxies)

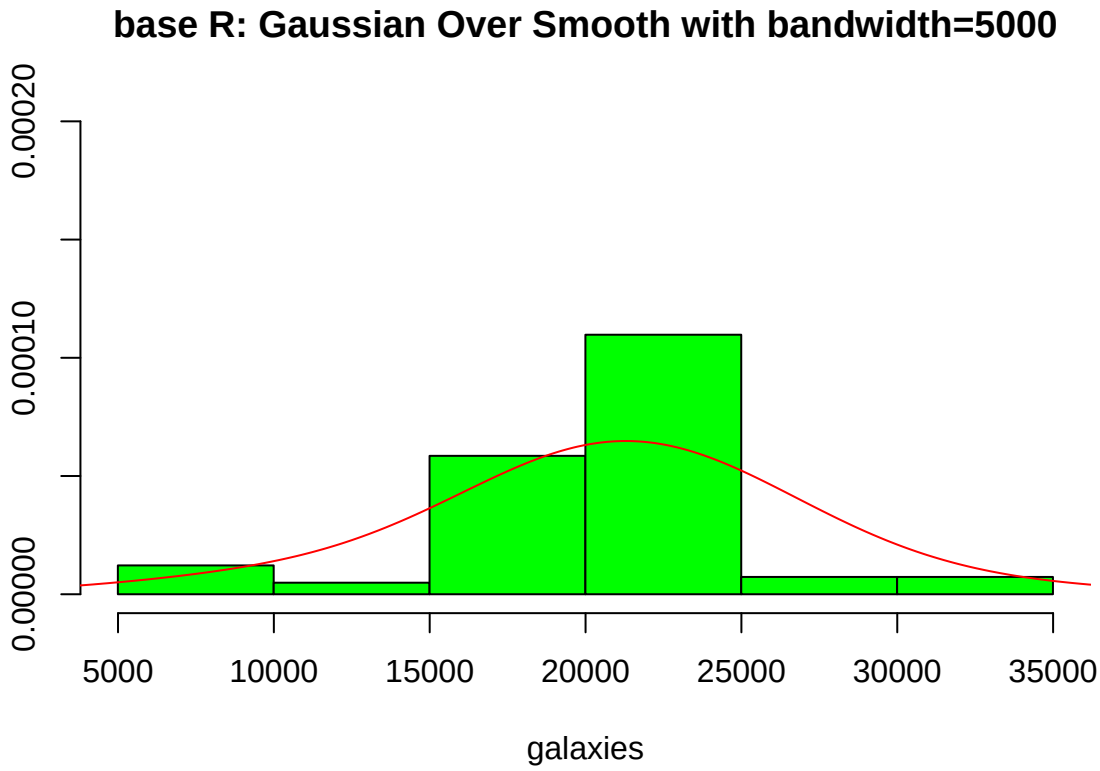

ggplot: True Histogram showing galaxies
density
3
2
1
0
9.0
9.5
10.0
10.5
logalaxies


base R: Histogram showing log−galaxies (qplot)
10
5
0
9.5
10.0
10.5
logalaxies

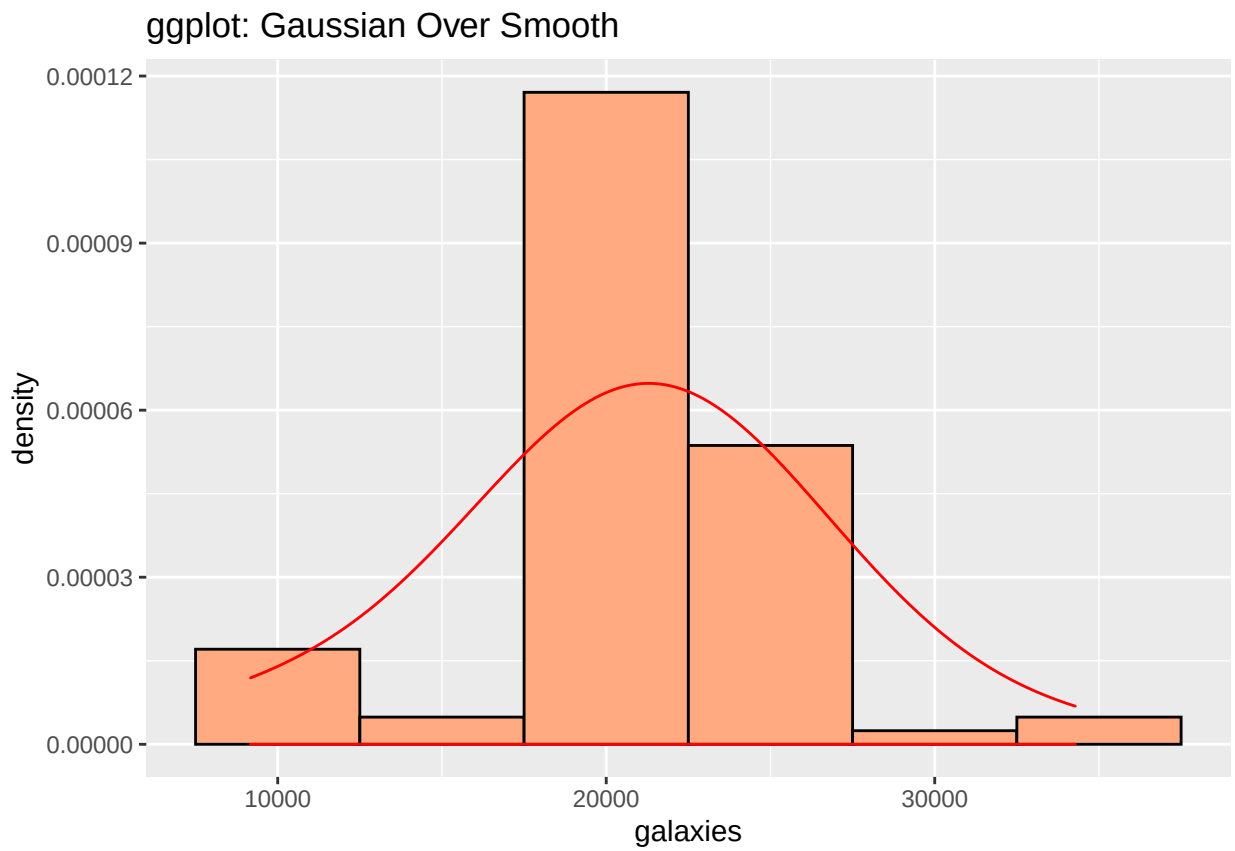


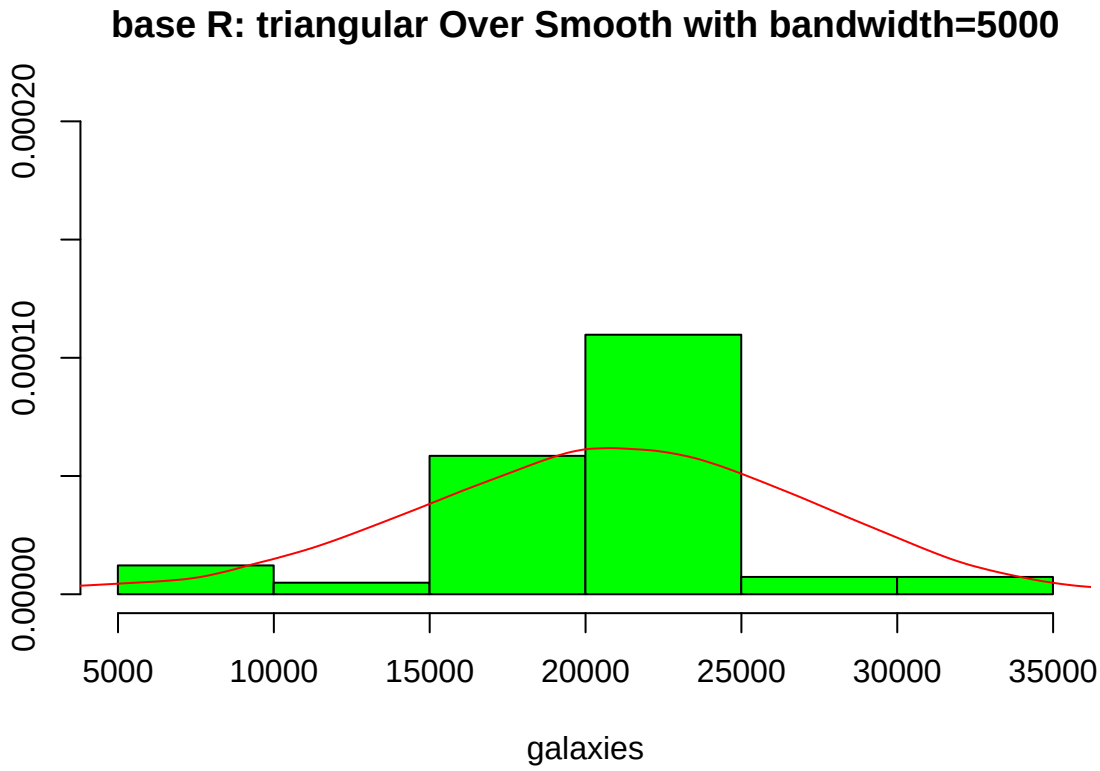
In this question, we constructed the graph, which is the same as the question-a but in the log form. The scale of the values in the log form for the galaxies. A similar description applies to this question as well, and we can say that we can see three main clusters with one small cluster in between the concentrated cluster.

c) Construct kernel density estimates using two different choices of kernel functions and three choices of bandwidth (one that is too large and "oversmooths," one that is too small and "undersmooths," and one that appears appropriate.) Therefore you should have six different kernel density estimates plots (you may combine plots when appropriate to reduce the number of plots made). Discuss your results. You can use the log scale or original scale for the variable, and specify in the plot x-axis which you choose.

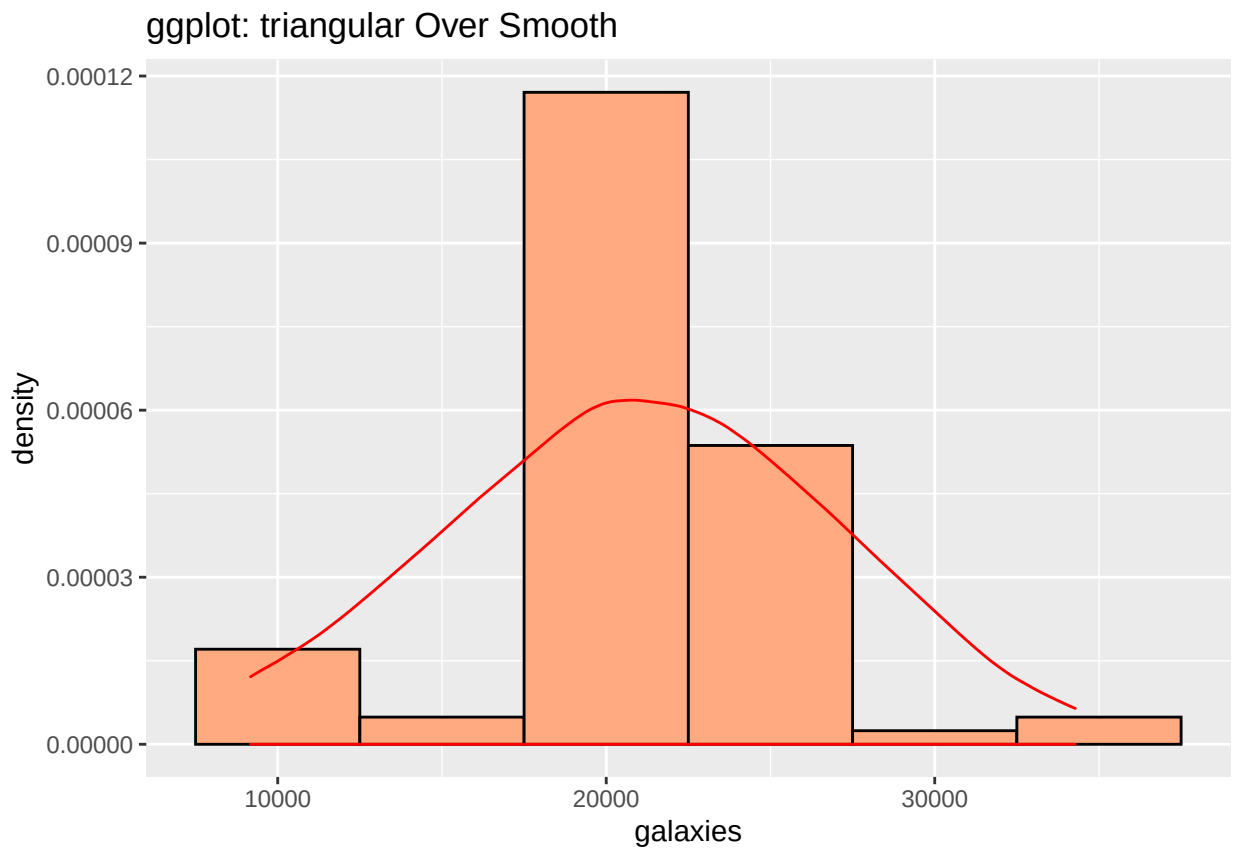

base R: Gaussian Over Smooth with bandwidth=5000
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

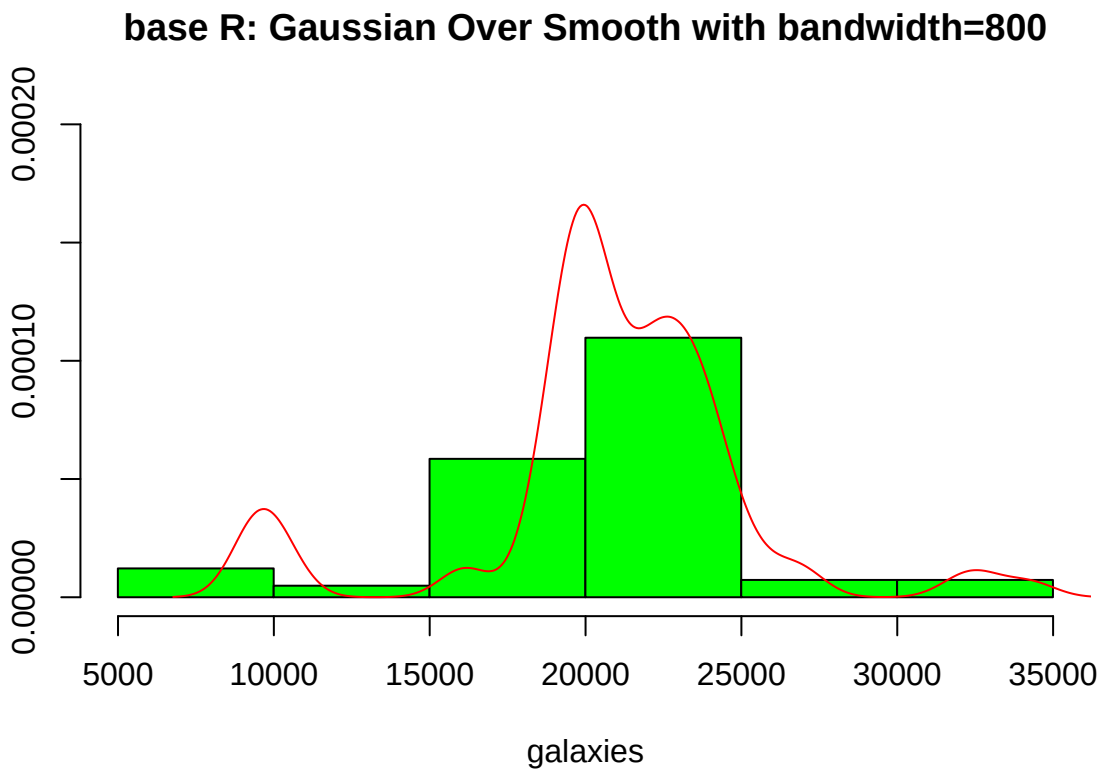
```
## integer(0)
```


ggplot: Gaussian Over Smooth
0.00012
0.00009
0.00006
0.00003
0.00000
density
10000
20000
30000
galaxies

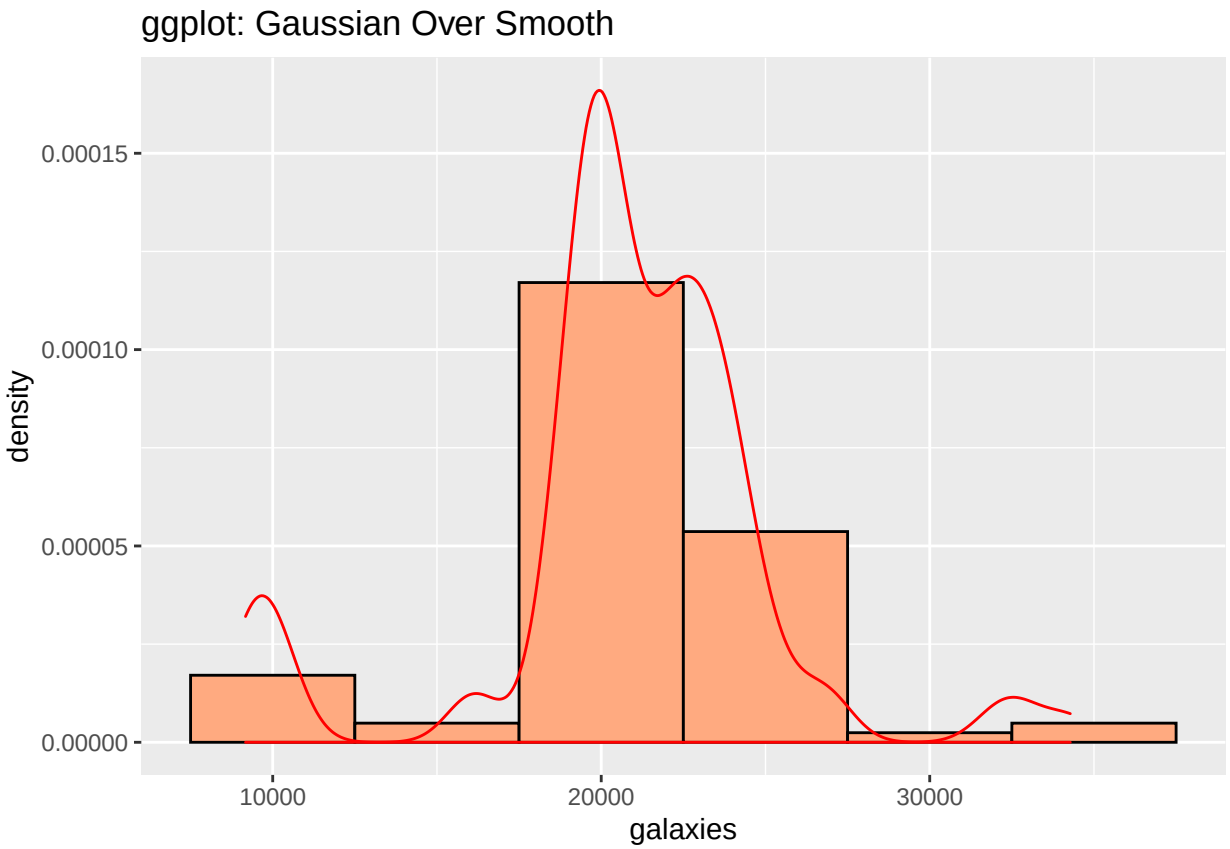

base R: triangular Over Smooth with bandwidth=5000
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

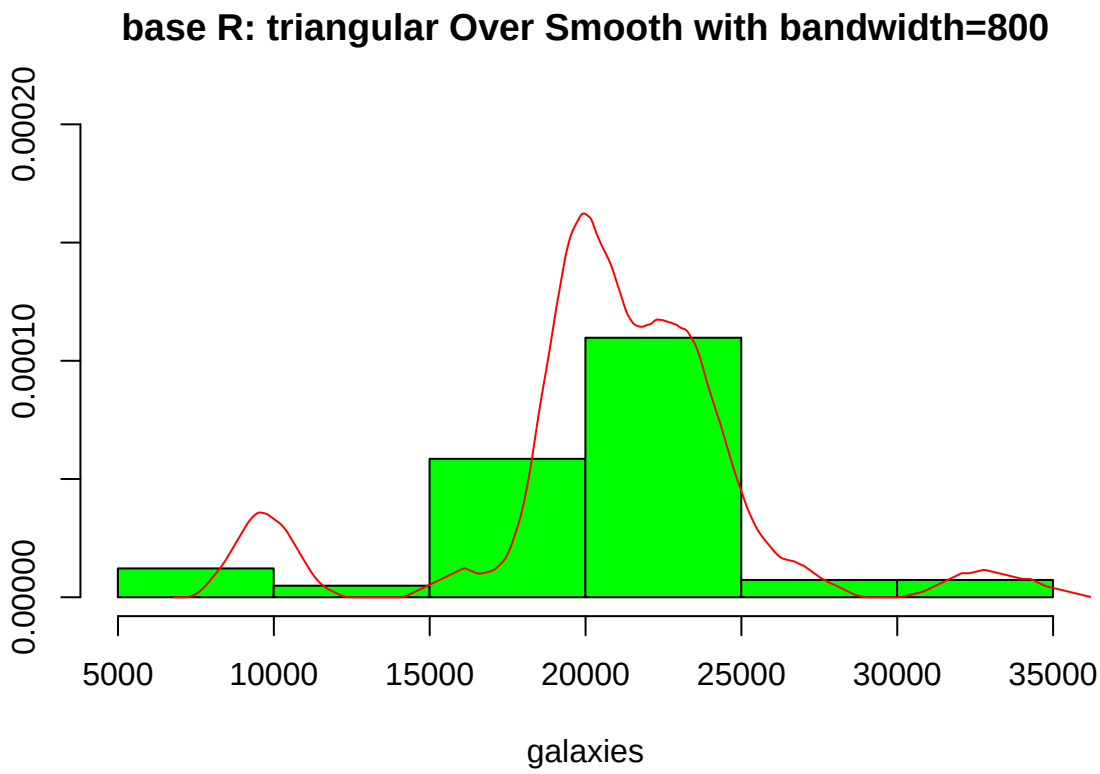
```
## integer(0)
```


ggplot: triangular Over Smooth
0.00012
0.00009
0.00006
0.00003
0.00000
density
10000
20000
30000
galaxies

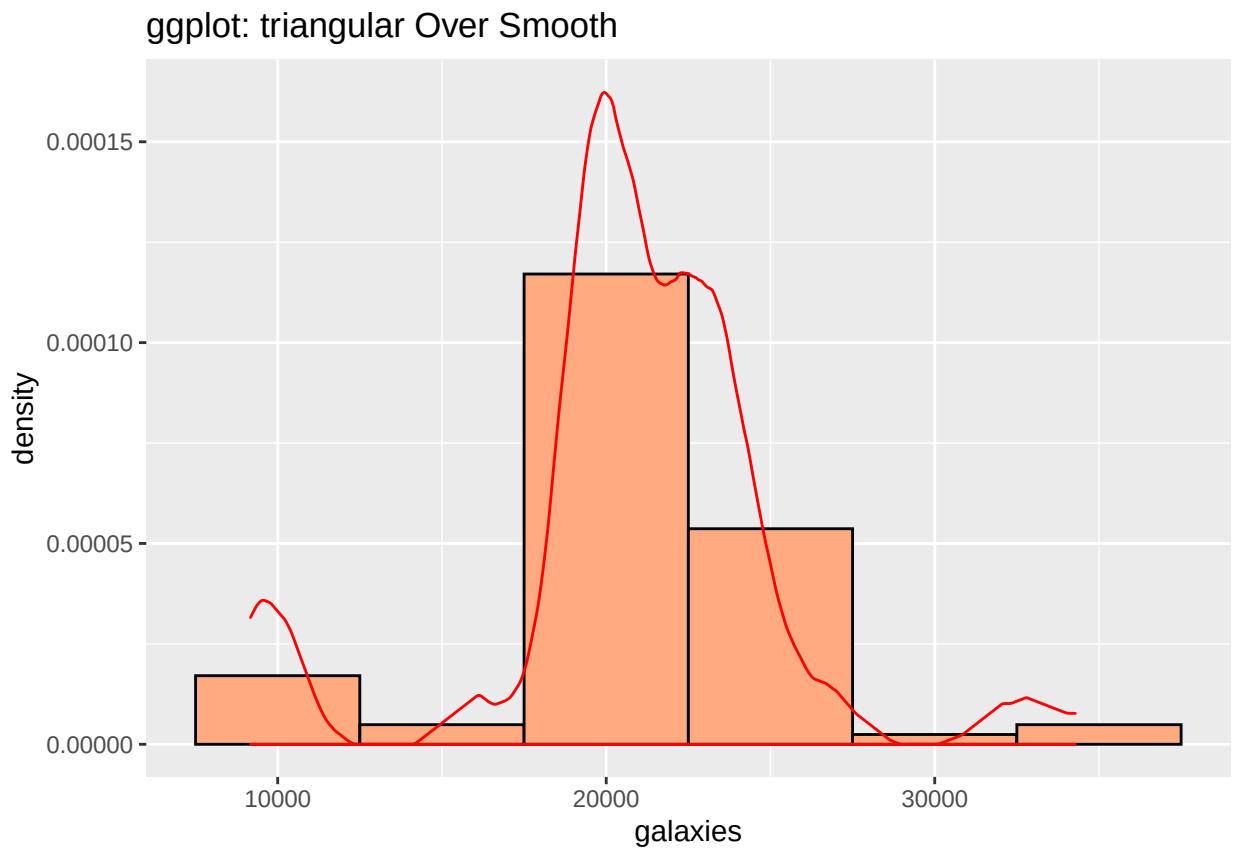

base R: Gaussian Over Smooth with bandwidth=800
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

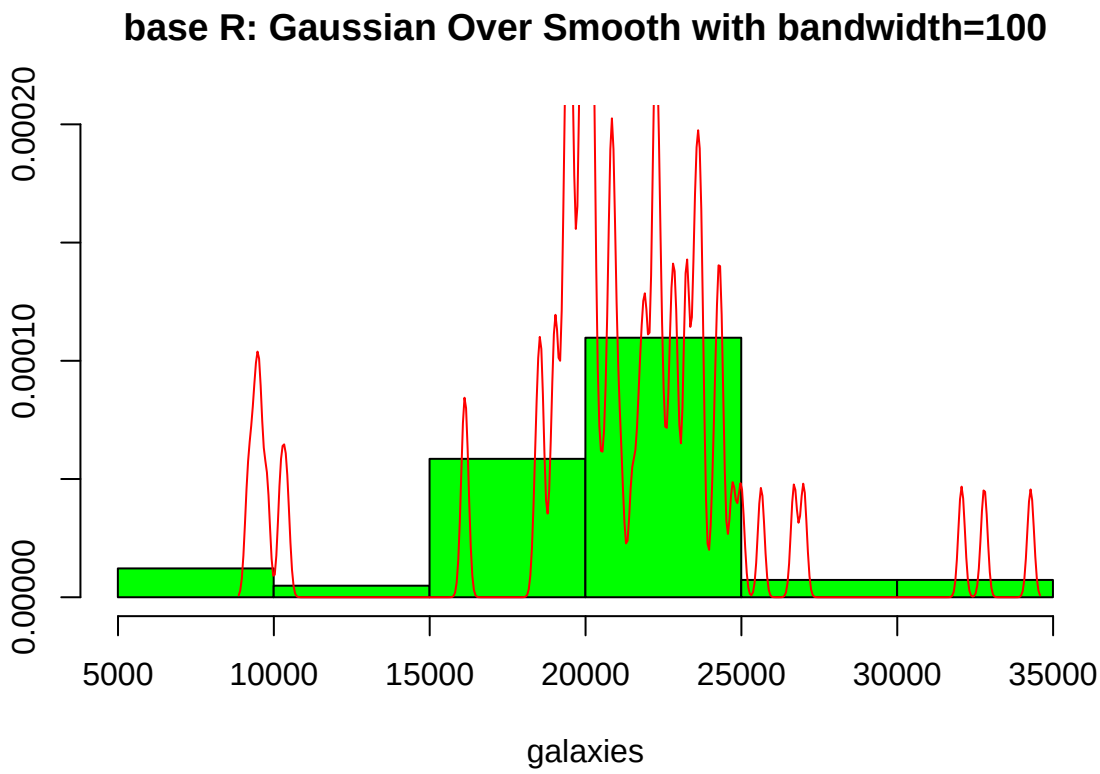
```
## integer(0)
```


ggplot: Gaussian Over Smooth
density
0.00015
0.00010
0.00005
0.00000
10000
20000
30000
galaxies

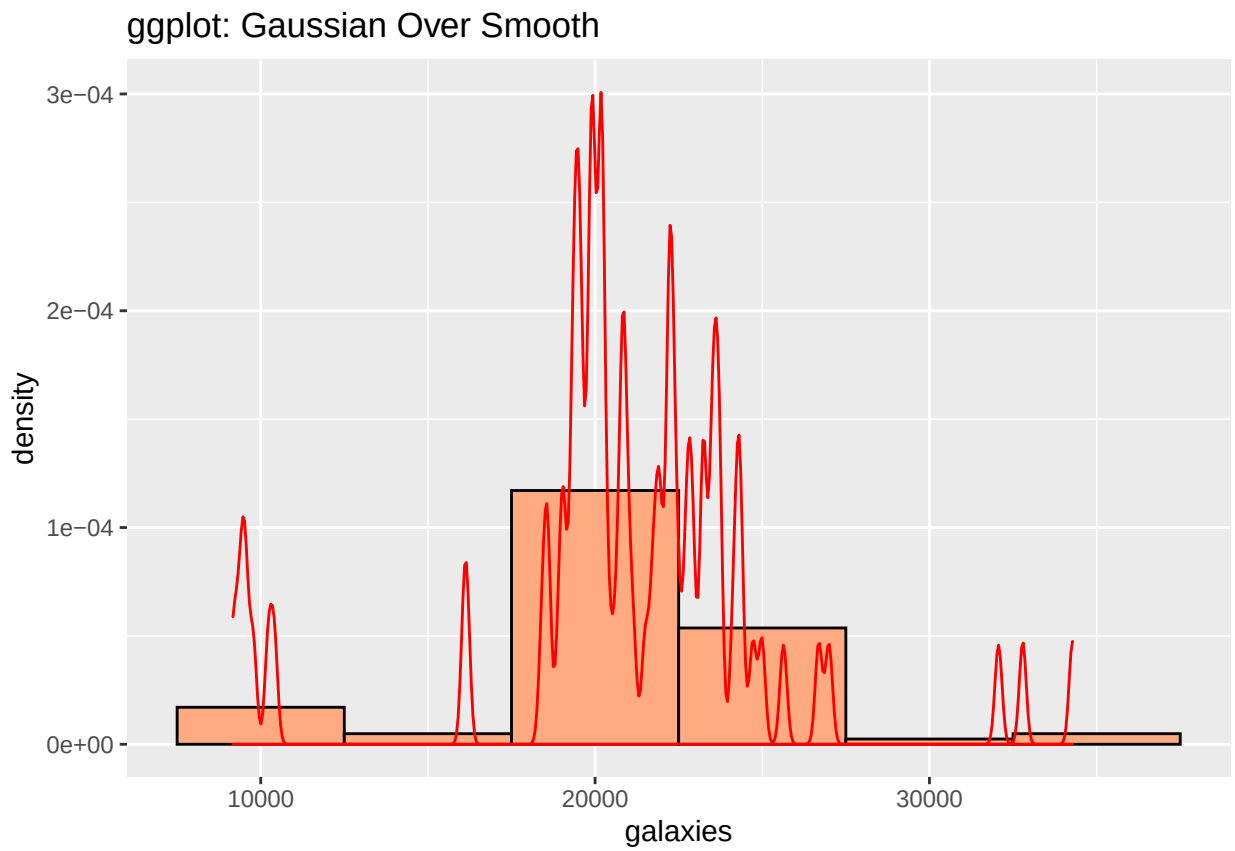

base R: triangular Over Smooth with bandwidth=800
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

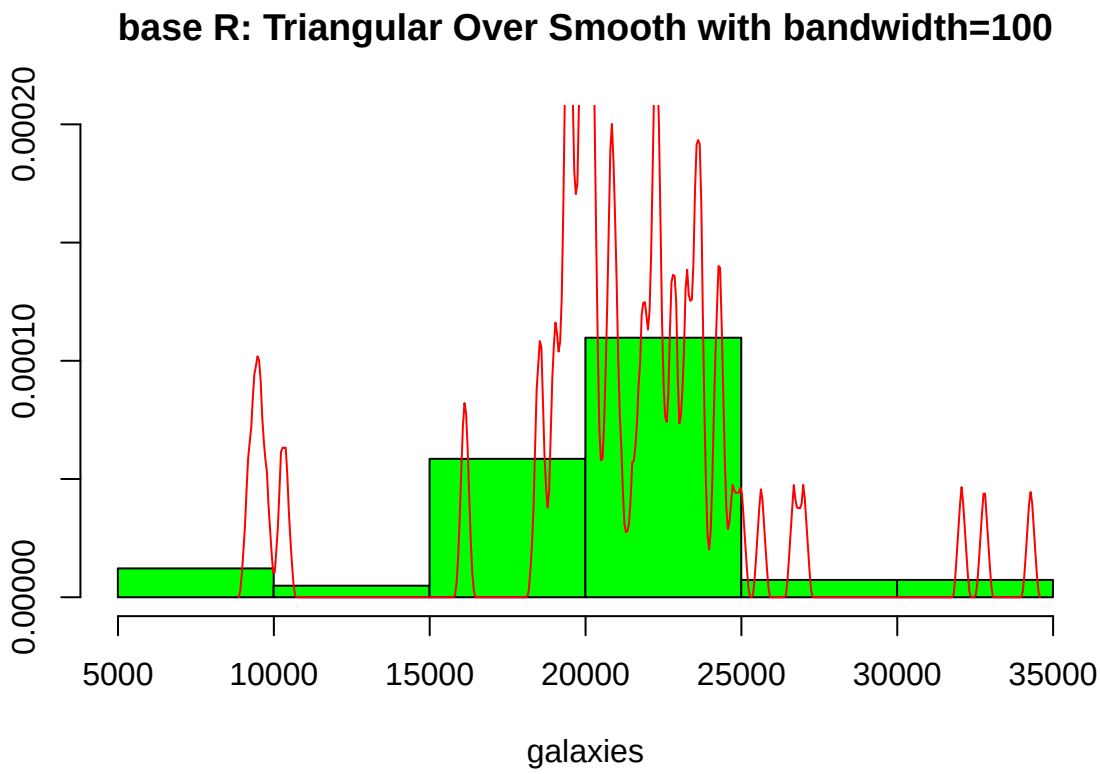
```
## integer(0)
```


ggplot: triangular Over Smooth
density
0.00015
0.00010
0.00005
0.00000
10000
20000
30000
galaxies

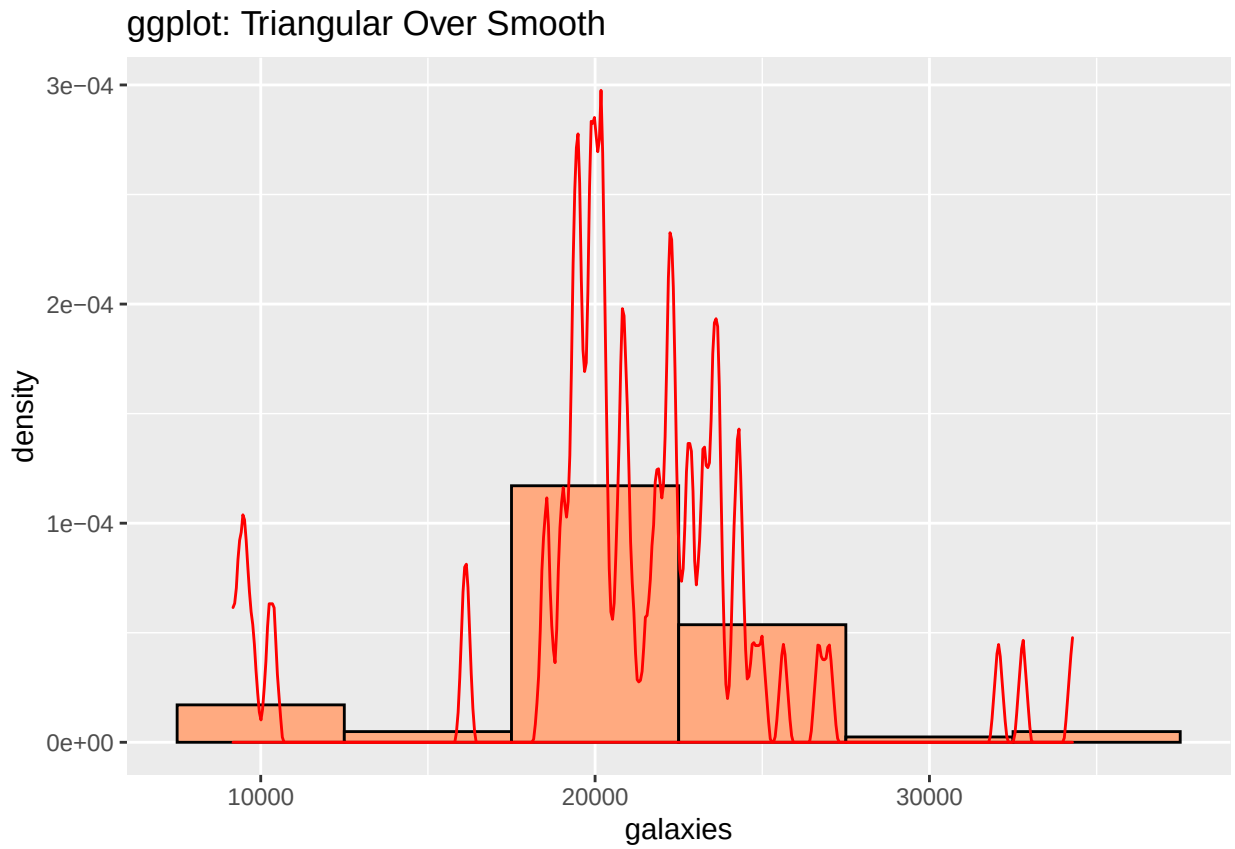

base R: Gaussian Over Smooth with bandwidth=100
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

```
## integer(0)
```

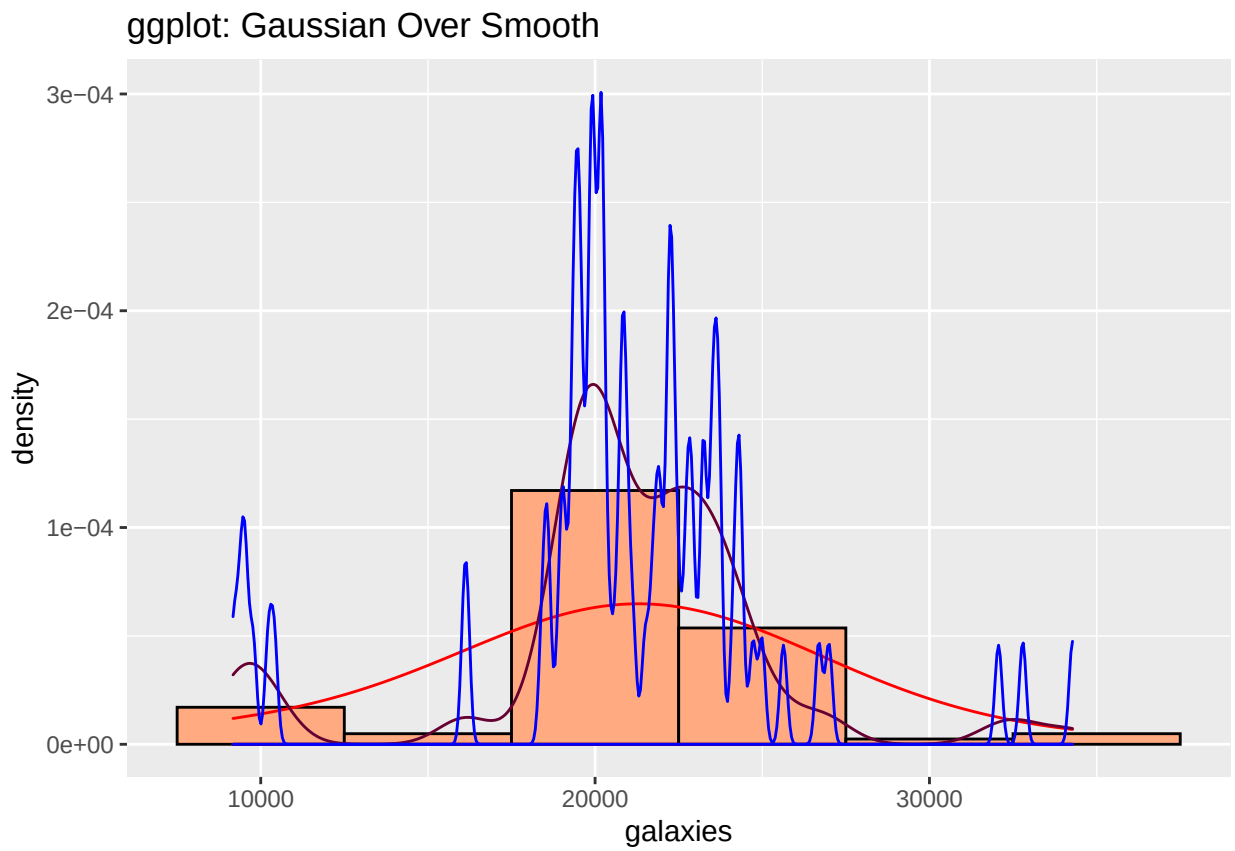

ggplot: Gaussian Over Smooth
3e−04
2e−04
1e−04
0e+00
density
10000
20000
30000
galaxies

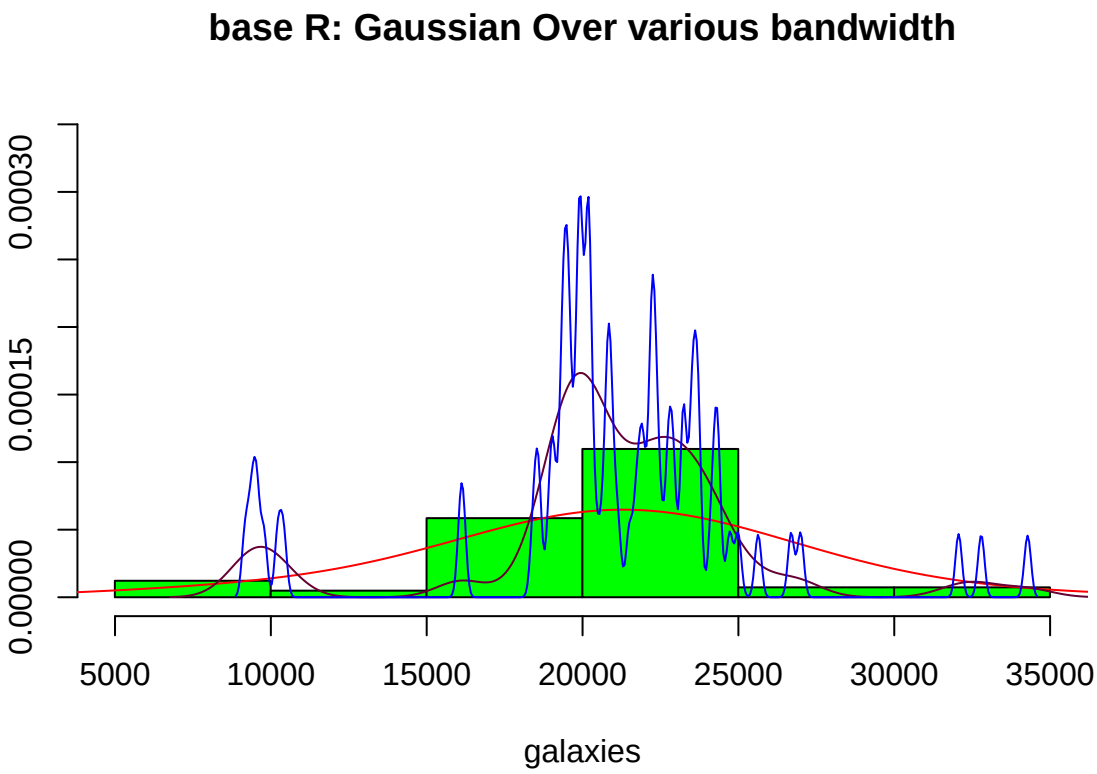

base R: Triangular Over Smooth with bandwidth=100
0.00020
0.00010
0.00000
5000
10000
15000
20000
25000
30000
35000
galaxies

```
## integer(0)
```

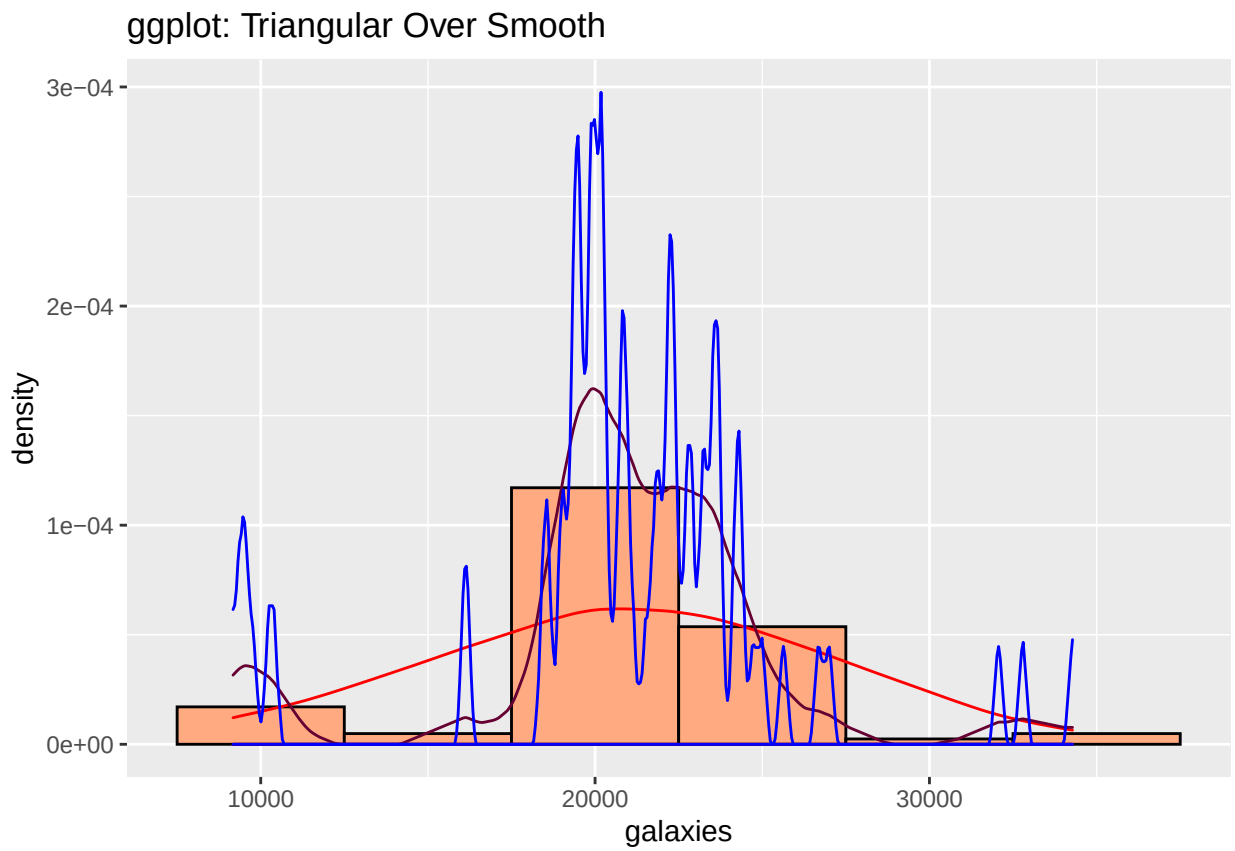


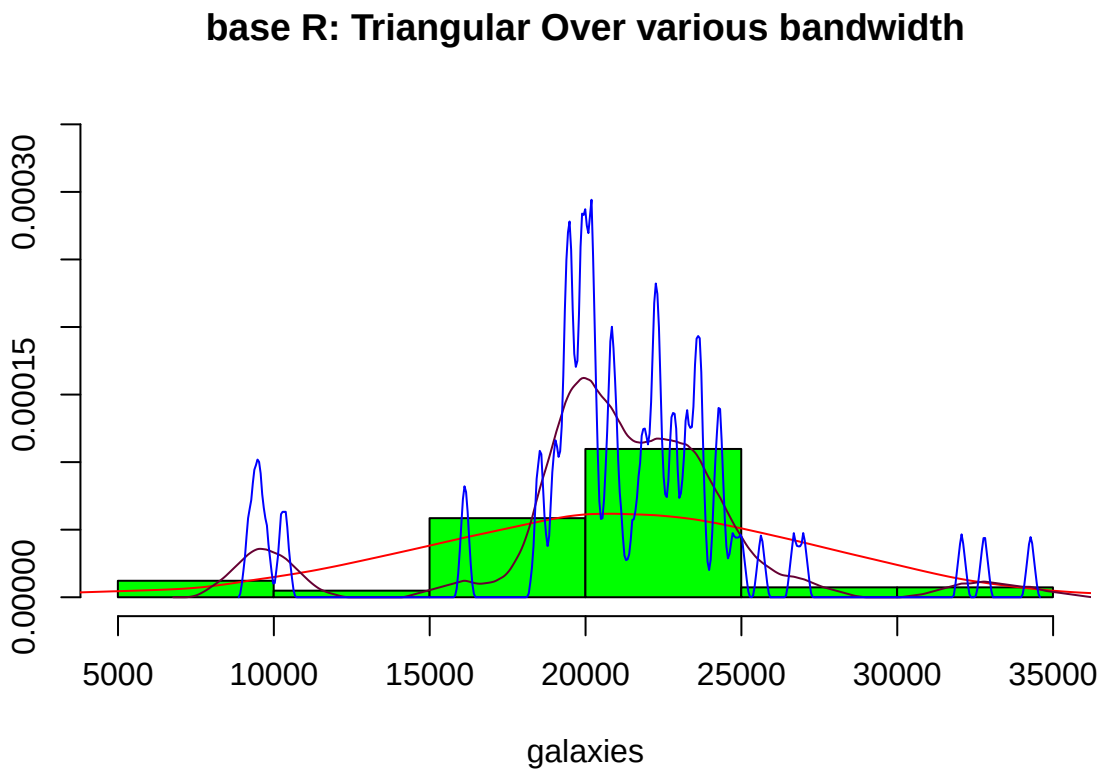
I also wanted to make plot it in the same graph


ggplot: Gaussian Over Smooth
3e−04
2e−04
1e−04
0e+00
density
10000
20000
30000
galaxies


base R: Gaussian Over various bandwidth
0.00000
0.00015
0.00030
5000
10000
15000
20000
25000
30000
35000
galaxies

```
## integer(0)
```


ggplot: Triangular Over Smooth
3e−04
2e−04
1e−04
0e+00
density
10000
20000
30000
galaxies




```
## integer(0)
```

Here, I have used kernel density of Gaussian function and triangular function and have used three different bandwidth that is 5000, 800, and 100. With the highest bandwidth, we couldn't see what is happening in the graph. I couldn't see the imprints of clusters. Likewise, when the bandwidth is small, there were too many bumps. We couldn't see what is happening in the graph. For the proper fit of the kernel density, I repeatedly replaced the value of bandwidth and obtained the graph.

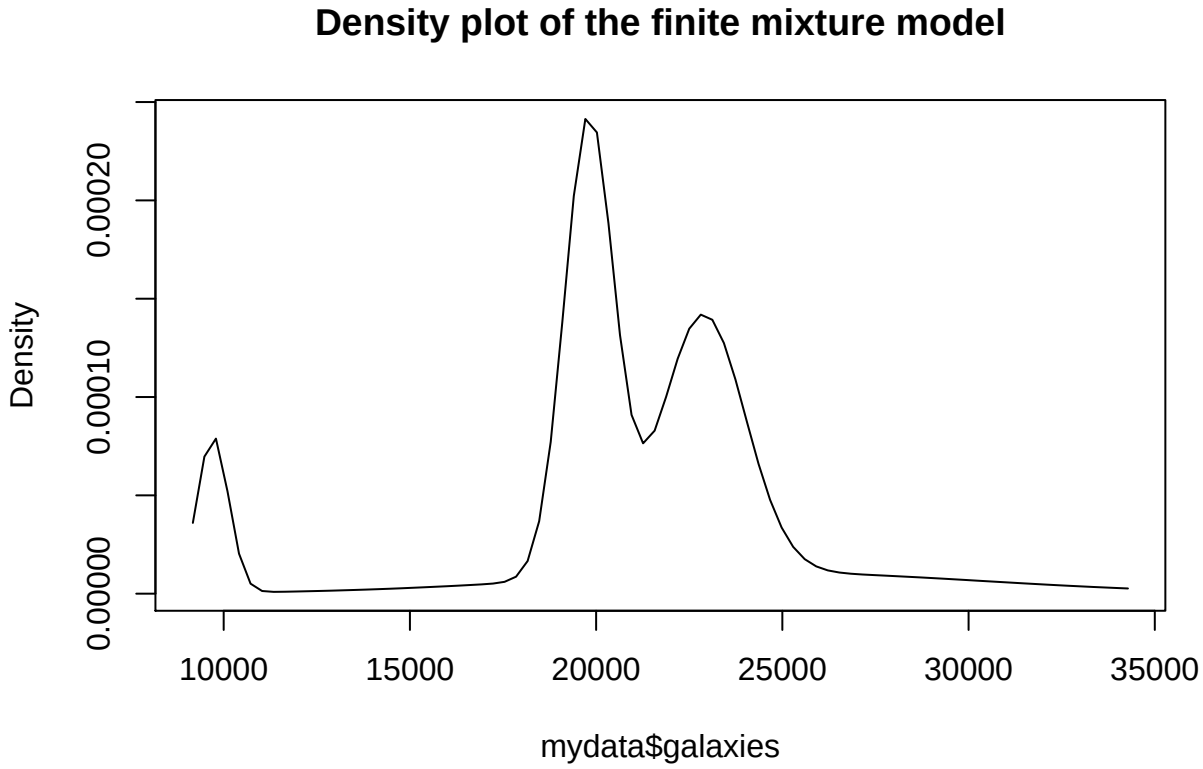
d) What is your conclusion about the possible existence of superclusters of galaxies? How many superclusters (1, 2, 3, ...)? (Hint: the existence of clusters implies the existence of empty spaces between galaxies.)

From the graphs above, we can find the three clusters. However, there can be one more cluster in between the data in the middle. Therefore, I can say that there maybe 3 to 4 clusters.

e) Fit a finite mixture model using the Mclust() function in R (from the mclust library). How many clusters did it find? Did it find the same number of clusters as your graphical inspection? Report parameter estimates and BIC of the best model.

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##        -765.694 82 11 -1579.862 -1598.907
##
## Clustering table:
```

```
##  1  2  3  4
##  7 35 32  8
##
## Mixing probabilities:
##          1          2          3          4
## 0.08440635 0.38660329 0.37116156 0.15782880
##
## Means:
##         1         2         3         4
##  9707.492 19804.259 22879.486 24459.536
##
## Variances:
##          1          2          3          4
##   177296.7   436160.9  1261611.3 34437115.3
```




```
## integer(0)
```

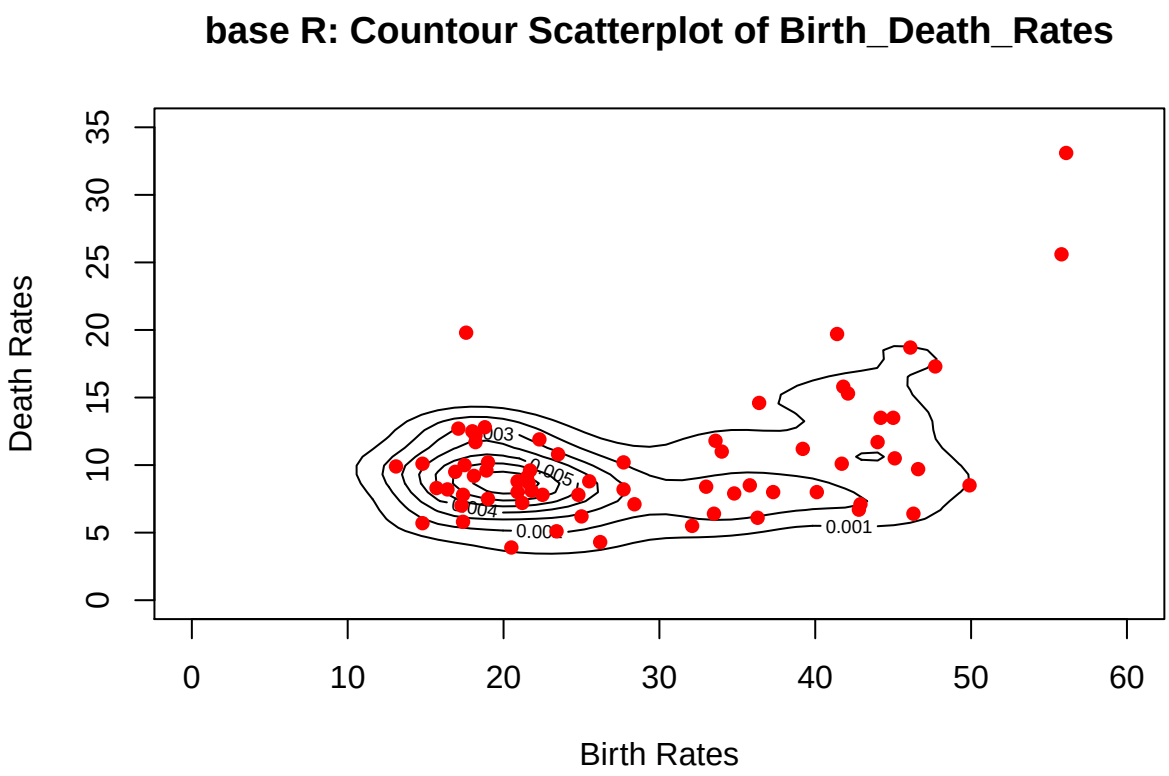
```
## Bayesian Information Criterion (BIC):
##           E         V
## 1 -1622.361 -1622.361
## 2 -1631.243 -1595.403
## 3 -1584.016 -1592.299
## 4 -1592.828 -1579.862
## 5 -1592.299 -1593.277
## 6 -1601.228 -1604.069
## 7 -1588.610 -1611.538
## 8 -1597.427 -1625.804
```

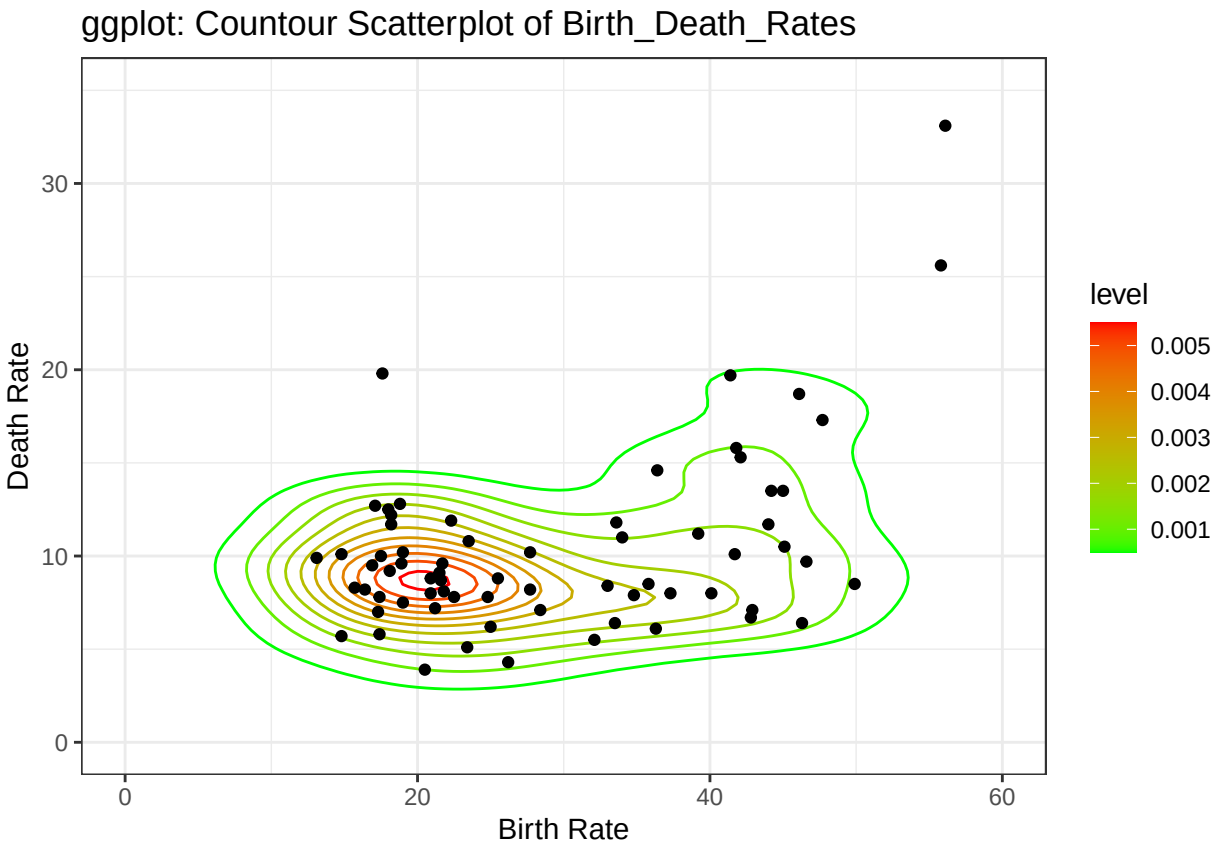
```
## 9 -1600.709 -1633.494
##
## Top 3 models based on the BIC criterion:
##       V,4       E,3       E,7
## -1579.862 -1584.016 -1588.610
```

From the Mclust, we find out that there are 4 clusters in the data. Whereas form the density plot, we found there are 3 clusters. Therefore we can say that there are 3 to 4 clusters in the Galaxies data.

2. (Ex. 8.2 in HSAUR, modified for clarity) The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (from Hartigan, 1975).

    a) Produce a scatterplot of the data. Estimate the bivariate density and overlay the corresponding contour plot on the scatterplot.

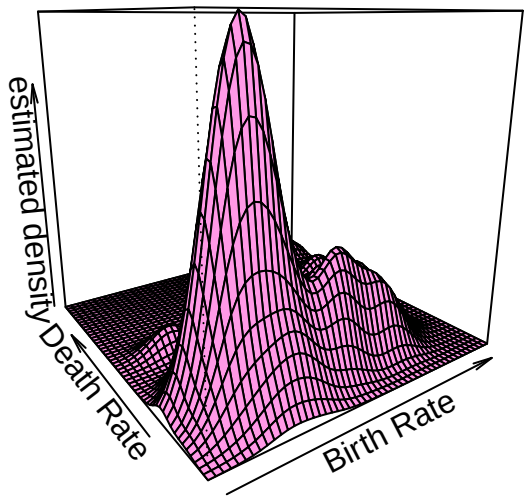




b) What does the contour plot tell you about the structure of the data?

Comparing data for birth rates from 10 to about 50 with the death rates, we can tell that the death rate is relatively slow. Also, twice as many people are being born than they are dying (i.e., 2:1 birth to death ratio).

c) Produce a perspective plot (persp() in R, ggplot is not required for this question).



d) Fit a finite mixture model using the Mclust() function in R (from the mclust library).Summarize this model using BIC, classification, uncertainty, and/or density plots.

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194 69 12 -899.6481 -906.4841
##
## Clustering table:
##  1  2  3  4
##  2 17 38 12
##
## Mixing probabilities:
##          1          2          3          4
## 0.02898652 0.24555002 0.55023375 0.17522972
##
## Means:
##           [,1]     [,2]      [,3]      [,4]
## birth 55.94967 43.80396 19.922913 33.730672
## death 29.34960 12.09411  9.081348  8.535812
##
## Variances:
## [,,1]
```

```
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,2]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,3]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,4]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
```

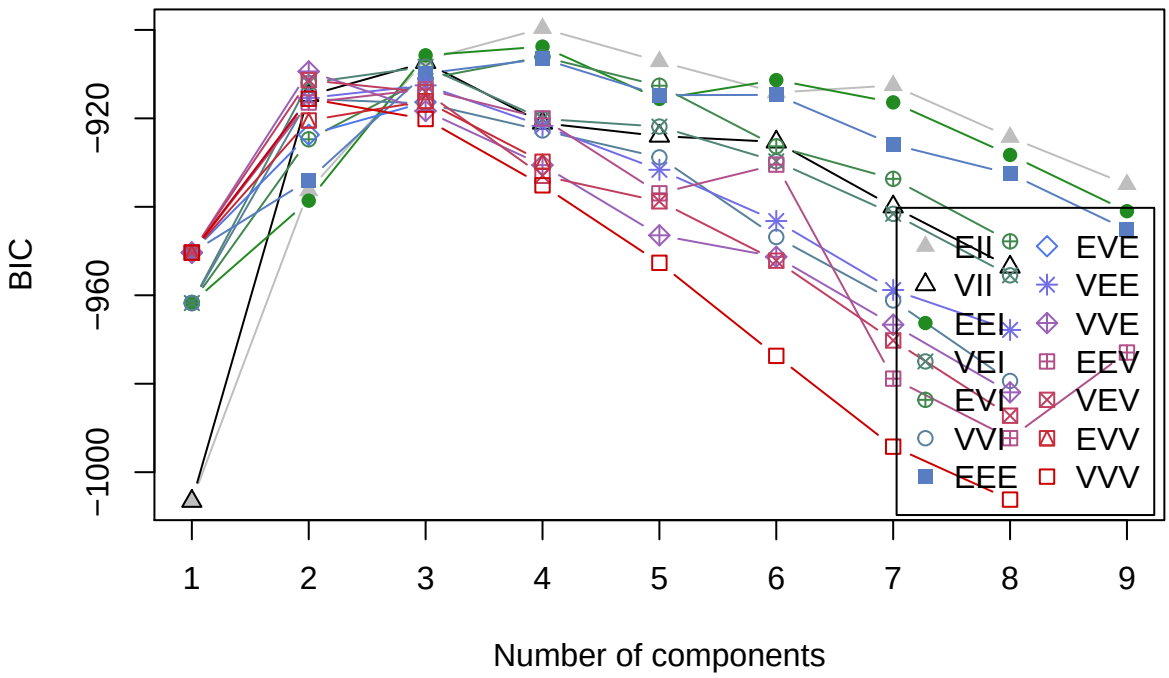
```
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "loglik"
##  [9] "df"             "bic"            "icl"            "hypvol"
## [13] "parameters"     "z"              "classification" "uncertainty"
```

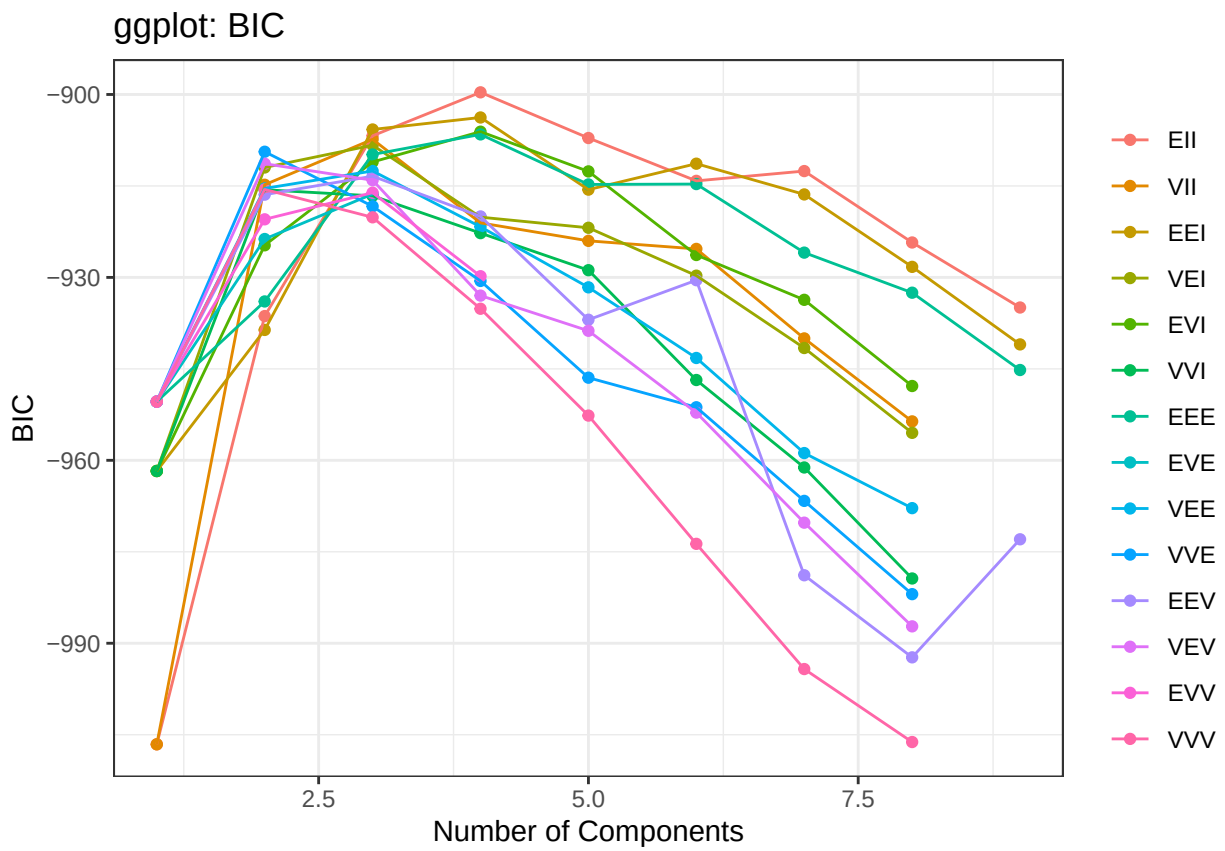
```
## $pro
## [1] 0.02898652 0.24555002 0.55023375 0.17522972
##
## $mean
##           [,1]     [,2]      [,3]      [,4]
## birth 55.94967 43.80396 19.922913 33.730672
## death 29.34960 12.09411  9.081348  8.535812
##
## $variance
## $variance$modelName
## [1] "EII"
##
## $variance$d
## [1] 2
##
## $variance$G
## [1] 4
##
## $variance$sigma
## , , 1
##
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
##
## , , 2
##
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
##
## , , 3
##
##         birth   death
```

```
## birth 10.2108  0.0000
## death  0.0000 10.2108
##
## , , 4
##
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
##
##
## $variance$Sigma
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
##
## $variance$sigmasq
## [1] 10.2108
##
## $variance$scale
## [1] 10.2108
```

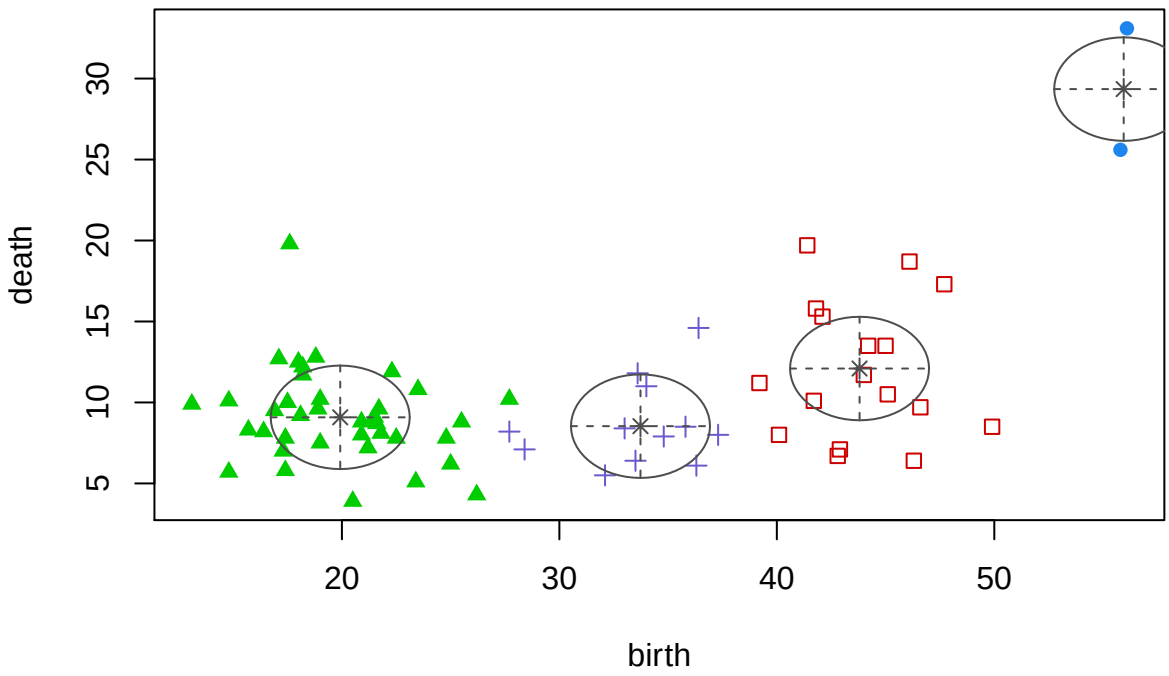
```
## Best BIC values:
##              EII,4      EEI,4       EEI,3
## BIC      -899.6481 -903.77041 -905.740323
## BIC diff    0.0000   -4.12227   -6.092186
```

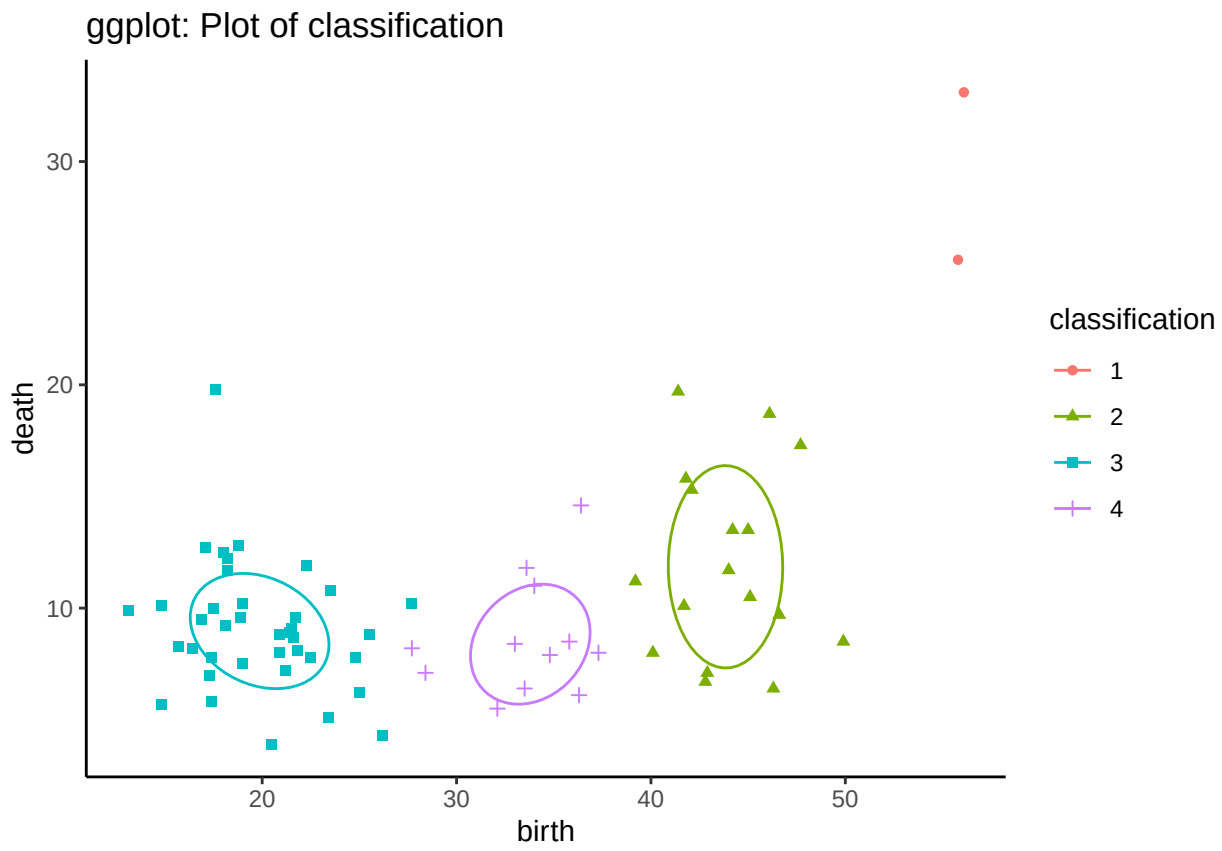
```
## Best BIC values:
##              EII,4      EEI,4       EEI,3
## BIC      -899.6481 -903.77041 -905.740323
## BIC diff    0.0000   -4.12227   -6.092186
```


BIC
−920
−960
−1000
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV

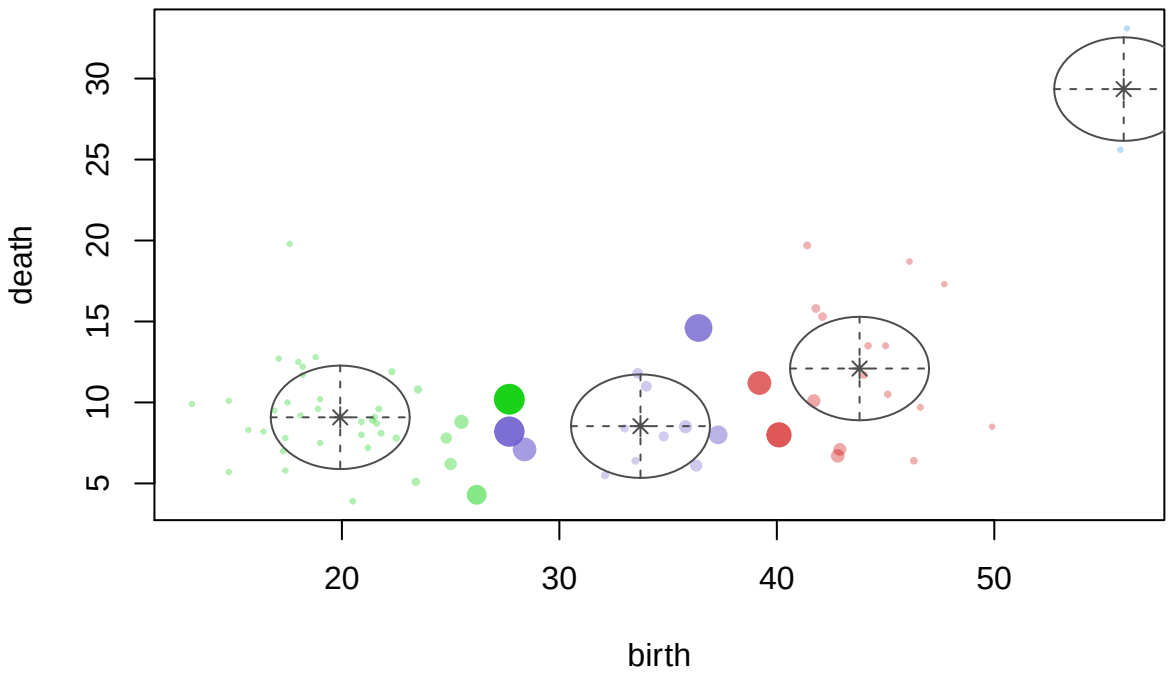


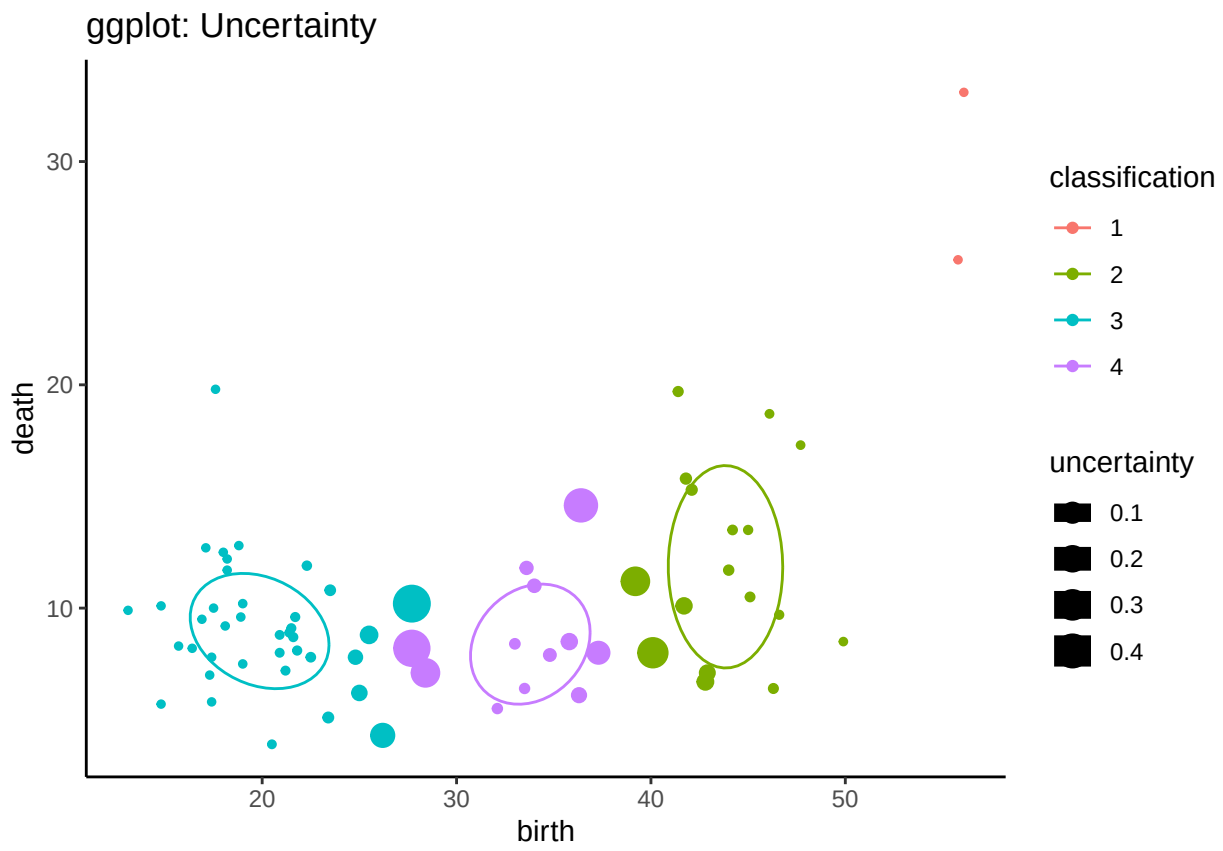

```
## ------------------------------------------------------------------
## Dimension reduction for model-based clustering and classification
## ------------------------------------------------------------------
##
## Mixture model type: Mclust (EII, 4)
##
## Clusters  n
##        1  2
##        2 17
##        3 38
##        4 12
##
## Estimated basis vectors:
##           Dir1     Dir2
## birth -0.94053 -0.22349
## death -0.33971  0.97471
##
##                Dir1      Dir2
## Eigenvalues  0.86401   0.15845
## Cum. %      84.50317 100.00000
```

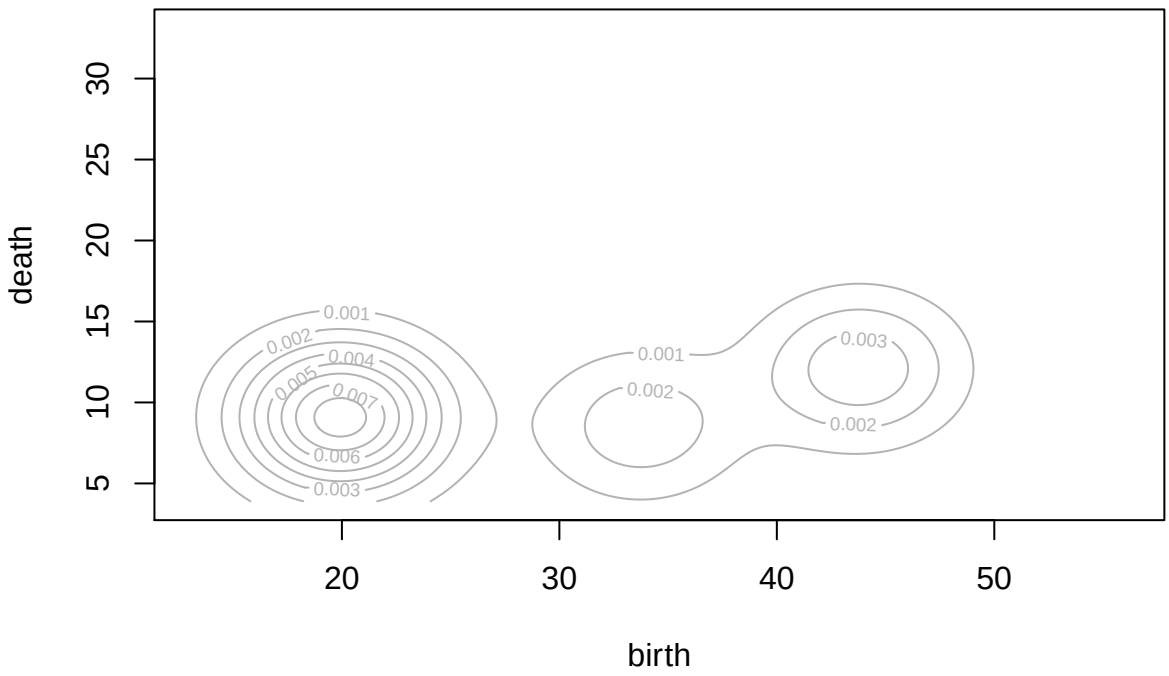

death
5
10
15
20
25
30
20
30
40
50
birth

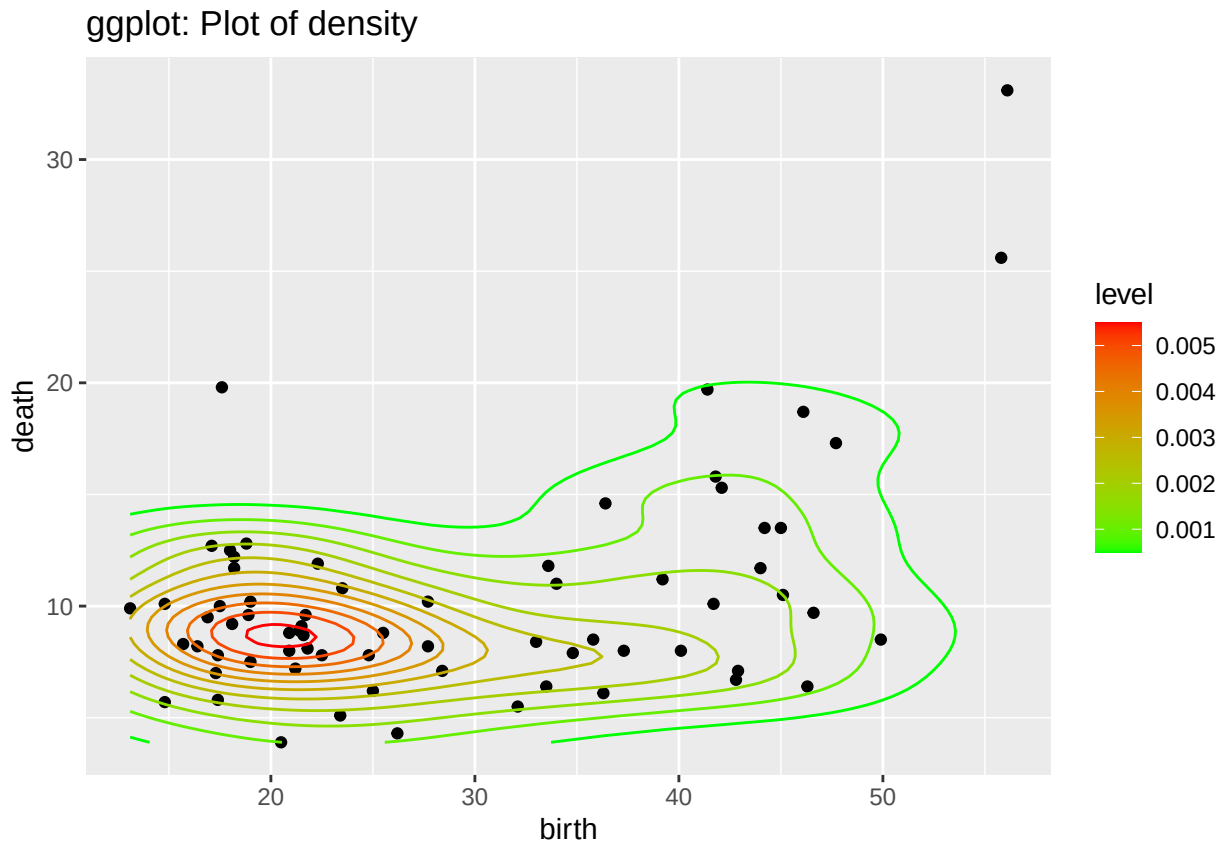



```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000024 0.0003831 0.0405931 0.0146759 0.4965738
```


death
5
10
15
20
25
30
20
30
40
50
birth


ggplot: Uncertainty
death
birth
30
20
10
20
30
40
50
classification
1
2
3
4
uncertainty
0.1
0.2
0.3
0.4


death
birth
5
10
15
20
25
30
20
30
40
50
0.001
0.002
0.004
0.005
0.007
0.006
0.003
0.001
0.002
0.003
0.002



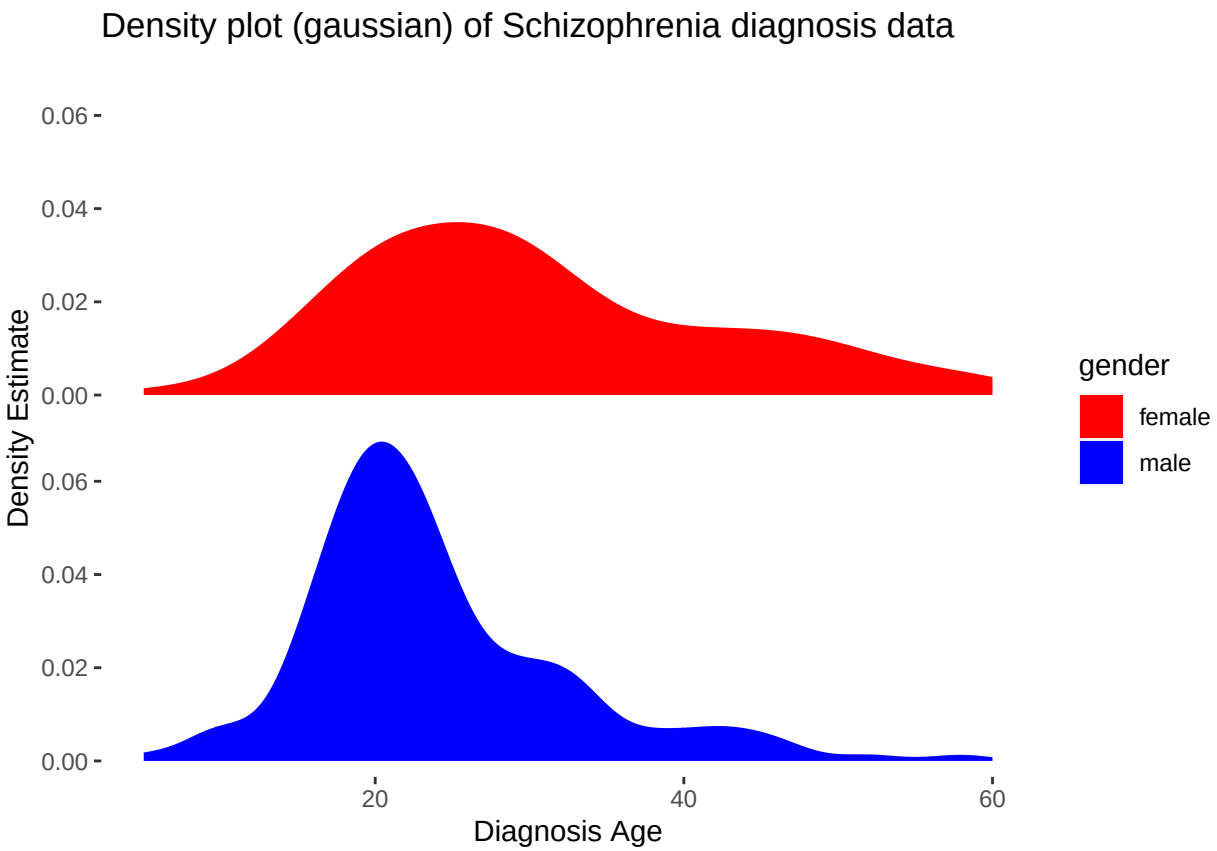
e) Discuss the results in the context of Birth and Death Rates.

The results and the graphs shows that there are four clusters, The mean birth of the these clusters are nearly equal to 20,34,44,54. Likewise, the mean death of the clusters are nearly equal to 8.5, 9, 12, and 29.

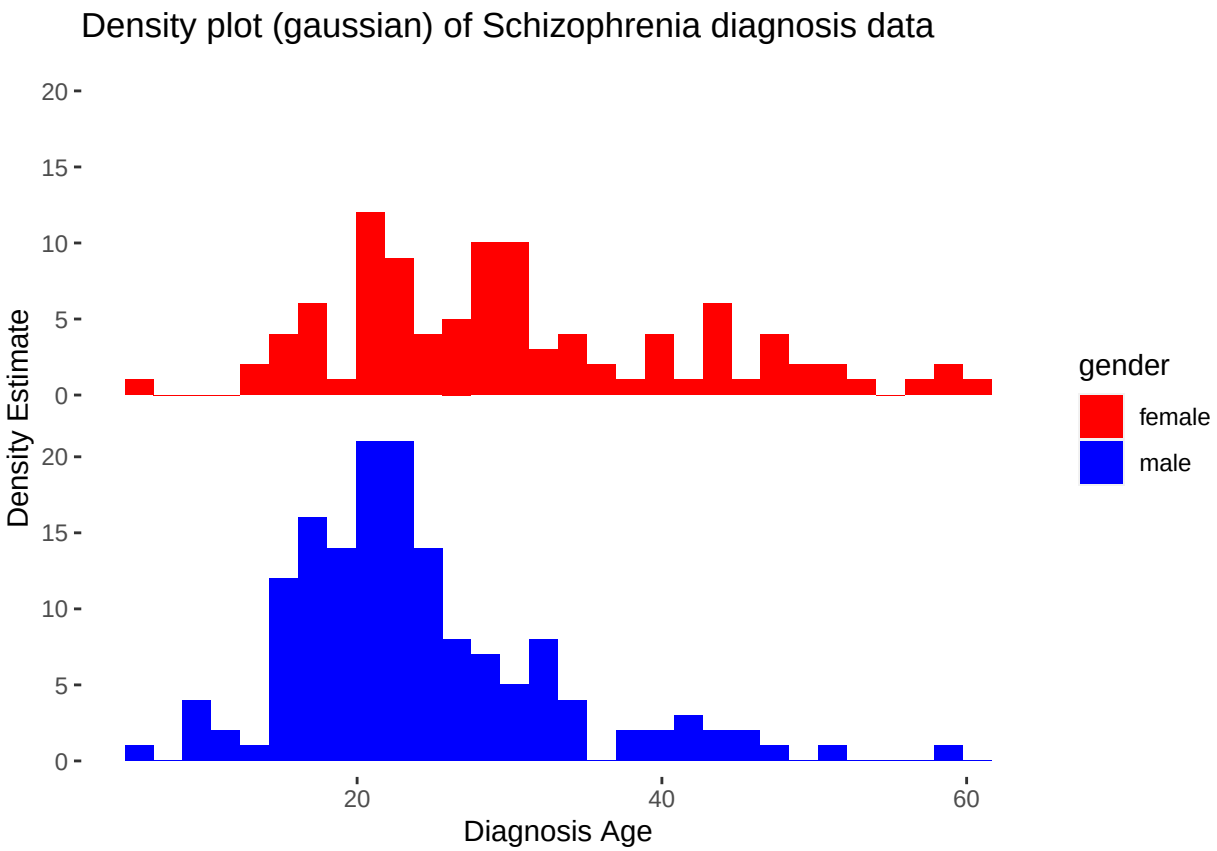
3. (Ex. 8.3 in HSAUR, modified for clarity) Fit finite mixtures of normal densities individually for each gender in the **schizophrenia** data set from **HSAUR3**. Do your models support the *sub-type model* described in the R Documentation?

   Quote from the R Documentation: *A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequent epidemiological studies of the disorder have consistently shown an earlier onset in men than in women. One model that has been suggested to explain this observed difference is known as the subtype model which postulates two types of schizophrenia, one characterized by early onset, typical symptoms and poor premorbid competence; and the other by late onset, atypical symptoms and good premorbid competence. The early onset type is assumed to be largely a disorder of men and the late onset largely a disorder of women.* (See ?schizophrenia)

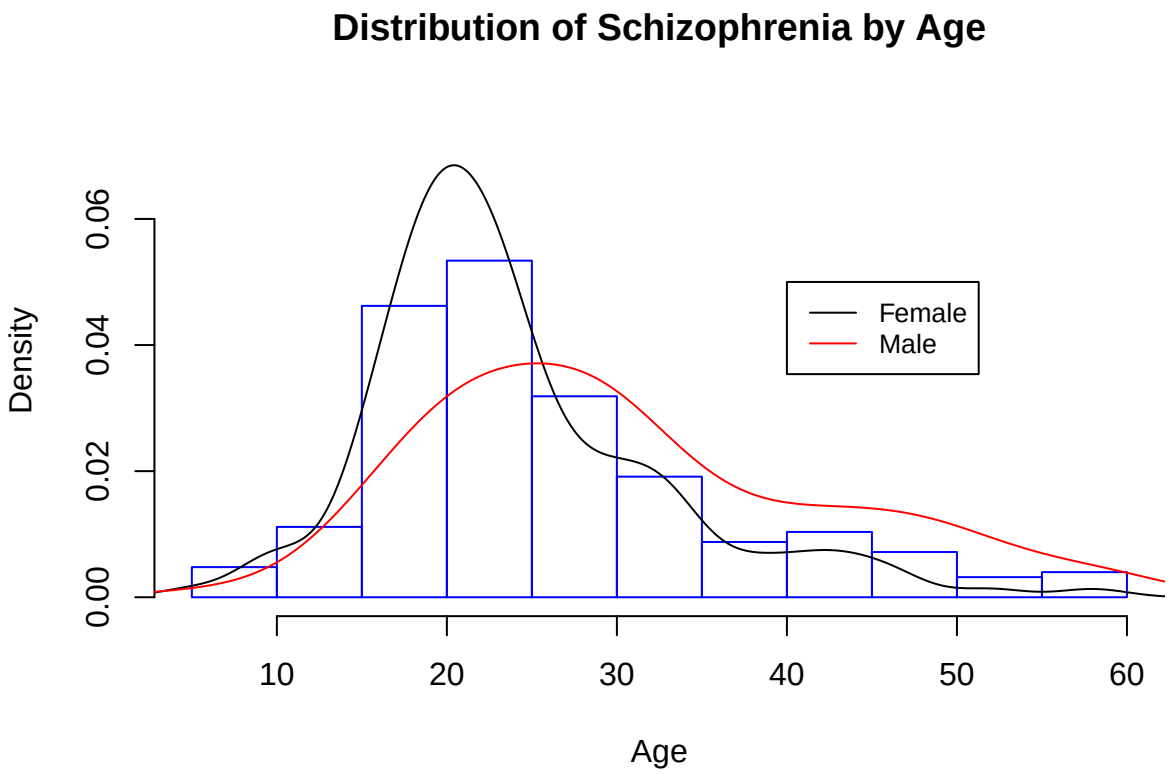
```
##   age gender
## 1  20 female
## 2  30 female
## 3  21 female
## 4  23 female
## 5  30 female
## 6  25 female
```



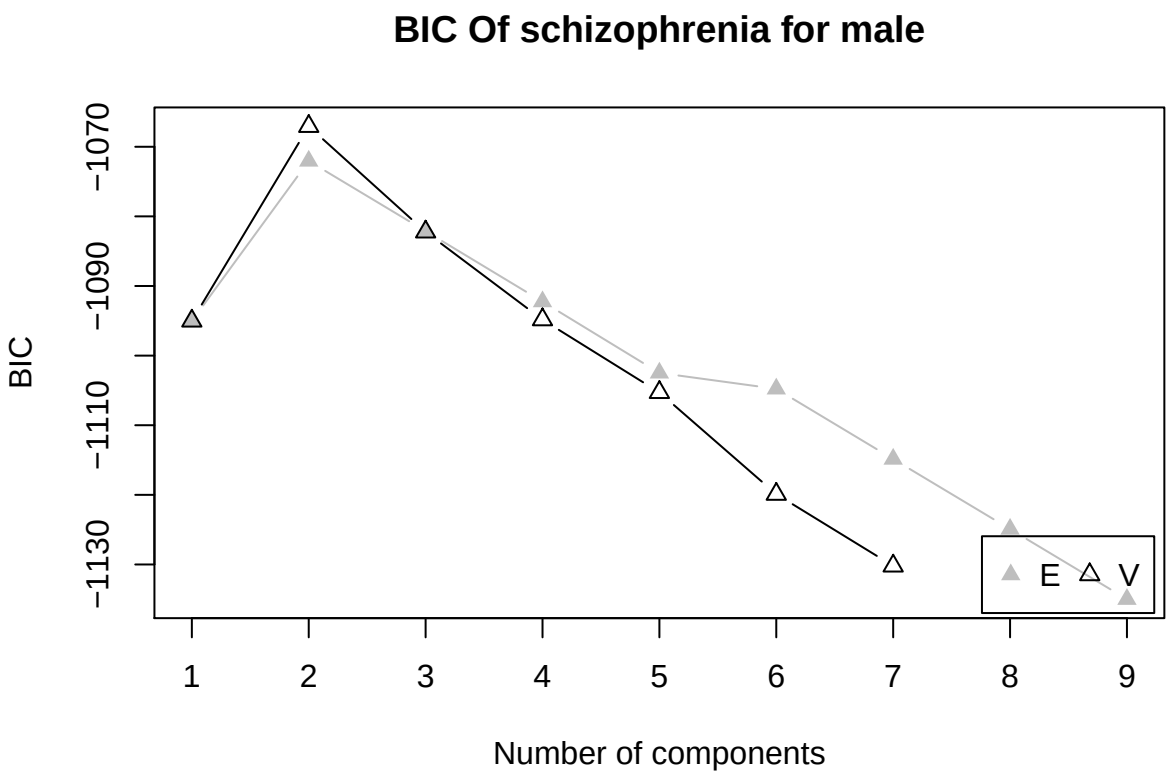
Based on the density plot, we can say that the mean of the age at which the schizophenia starts in male is about 20, Likewise, in female it is throughout the life.



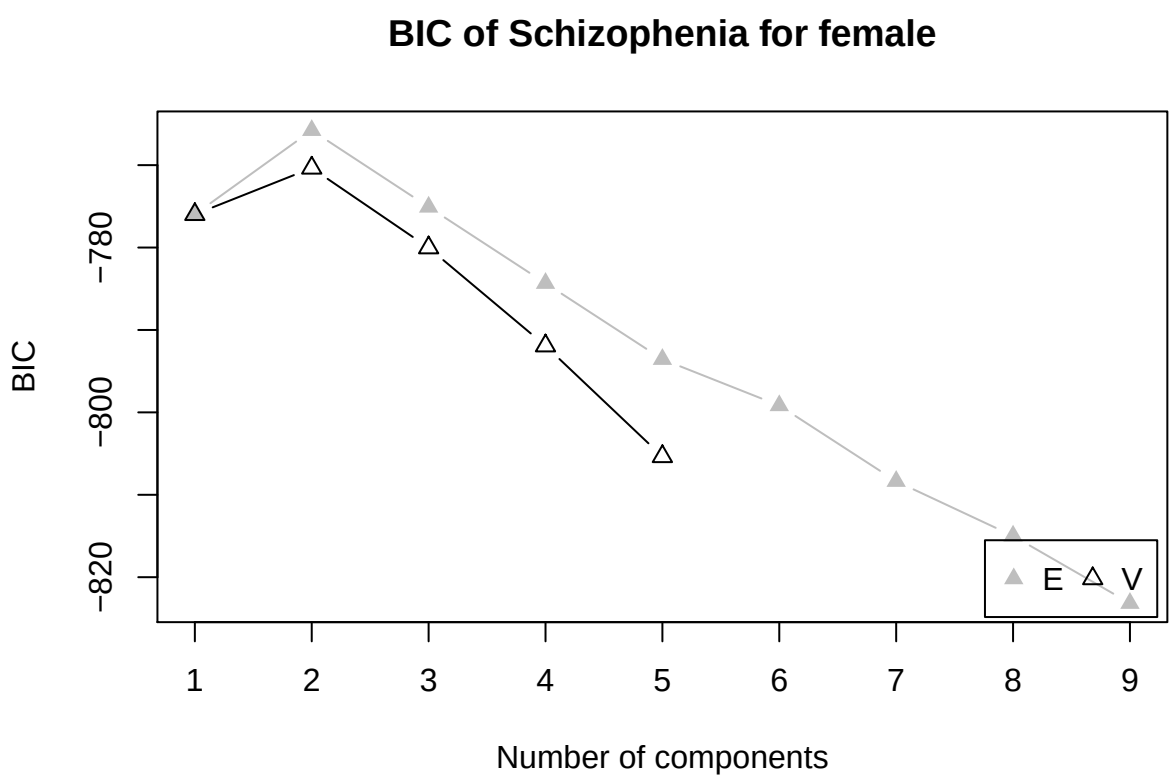
From the histograms, we see that the maximum numbr of men suffered form the disease around the age of the 20-35 and less at the age of 60. Whereas, in female the disease suffered from the teens and it continuous to the lifetime. we can visualize both together and see the same results discussed above by making this plot below:



We can subset the schizophrenia data by male and female for fit of model analysis by gender:


BIC Of schizophrenia for male
BIC
−1070
−1090
−1110
−1130
1
2
3
4
5
6
7
8
9
Number of components
E
V

```
## integer(0)
```




```
## integer(0)
```

```
## Summary for male
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##       -520.9747 152  5 -1067.069 -1134.392
##
## Clustering table:
##  1  2
## 99 53
##
## Mixing probabilities:
##         1         2
## 0.5104189 0.4895811
##
## Means:
##        1        2
## 20.23922 27.74615
##
## Variances:
##          1          2
```

```
##   9.395305 111.997525
```

```
## Summary for female
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -373.6992 99  4 -765.7788 -774.8935
##
## Clustering table:
##  1  2
## 74 25
##
## Mixing probabilities:
##         1         2
## 0.7472883 0.2527117
##
## Means:
##        1        2
## 24.93517 46.85570
##
## Variances:
##        1        2
## 44.55641 44.55641
```

From the model summary above, we can see that the female model showing data points centered at about 25 and age 47 of age marks, whereas for males it was at around 20 and 27 years of age (i.e., within 20s). The BIC plot shows that the optimal number of cluster for both males and females is 2.