Homework 4

Snigdha Peddi

Question 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)

References: Ref1- Chapter_8_modified.R, Ref2,Ref3, Ref4, Ref5,Ref6, Ref7, Ref8, Ref9, Ref10

Answer 1: Converted the galaxies data set to a data frame. As per the R documentation there is a typo in the 78th observation. Instead of 26960 the velocity 26690 was included.This observation is corrected in the new data set.

a) Construct histograms using the following functions:

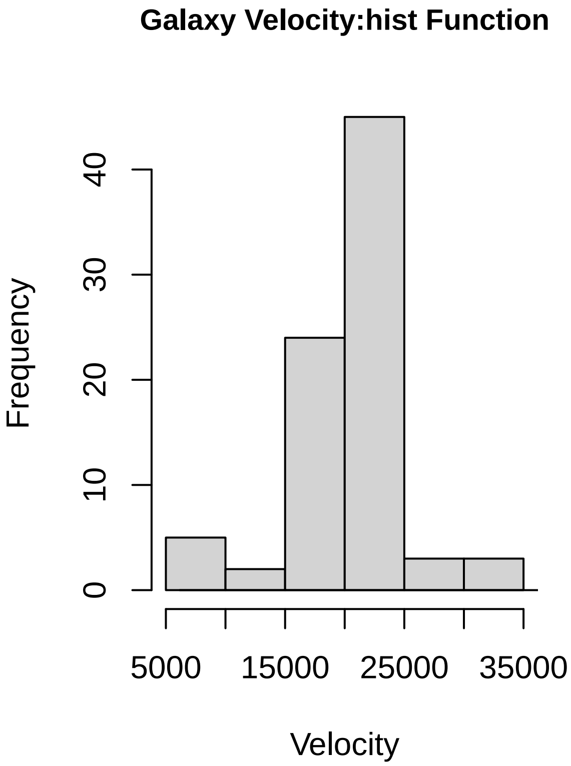
-hist() and ggplot()+geom_histogram()

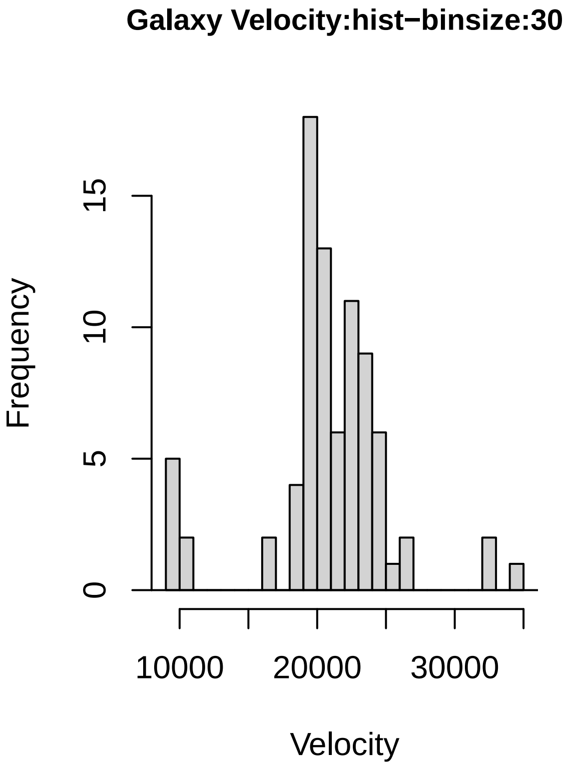
-truehist() and ggplot+geom_histogram() (make sure that the histograms show proportions, not counts.)

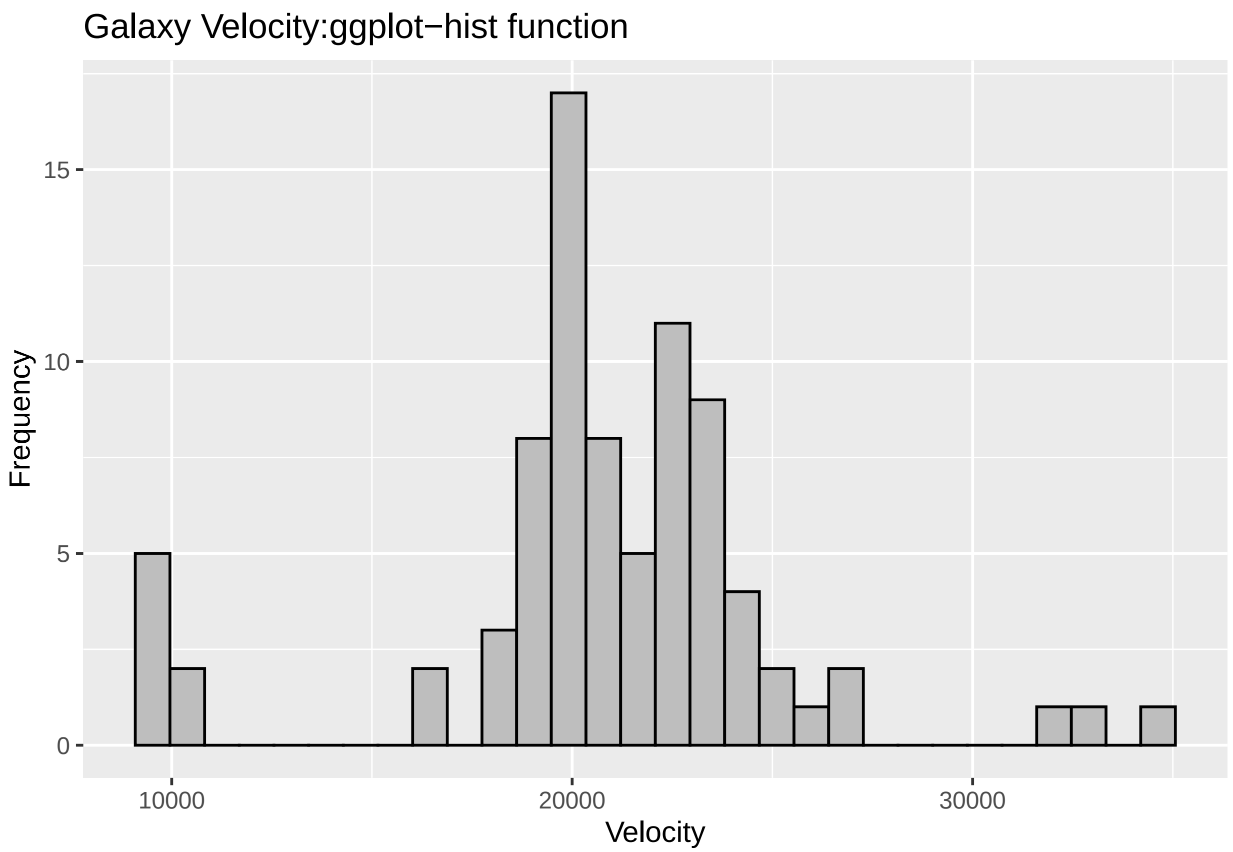
-qplot()

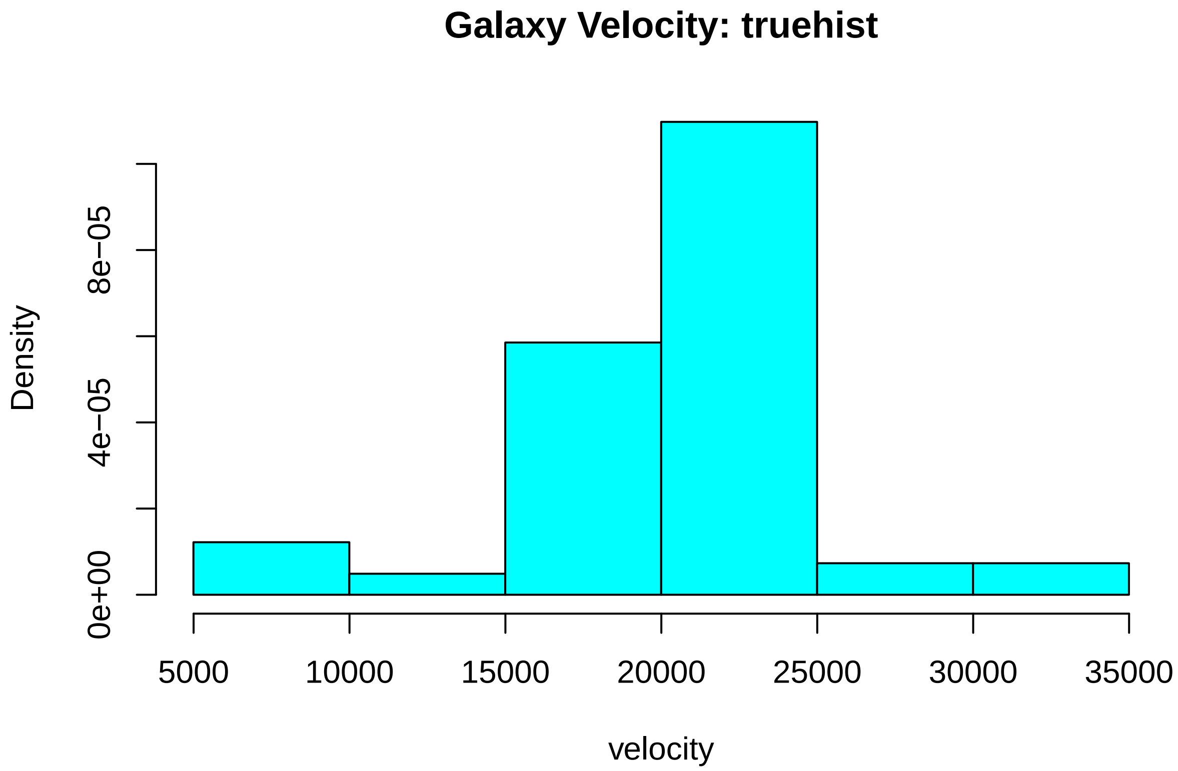
Comment on the shape and properties of the variable based on the five plots. Do you notice any sets of observations clustering? (Hint: You can adjust bin number or bin size as you try to determine the properties of the variable, but use the same bin settings between plots in your final analysis. You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

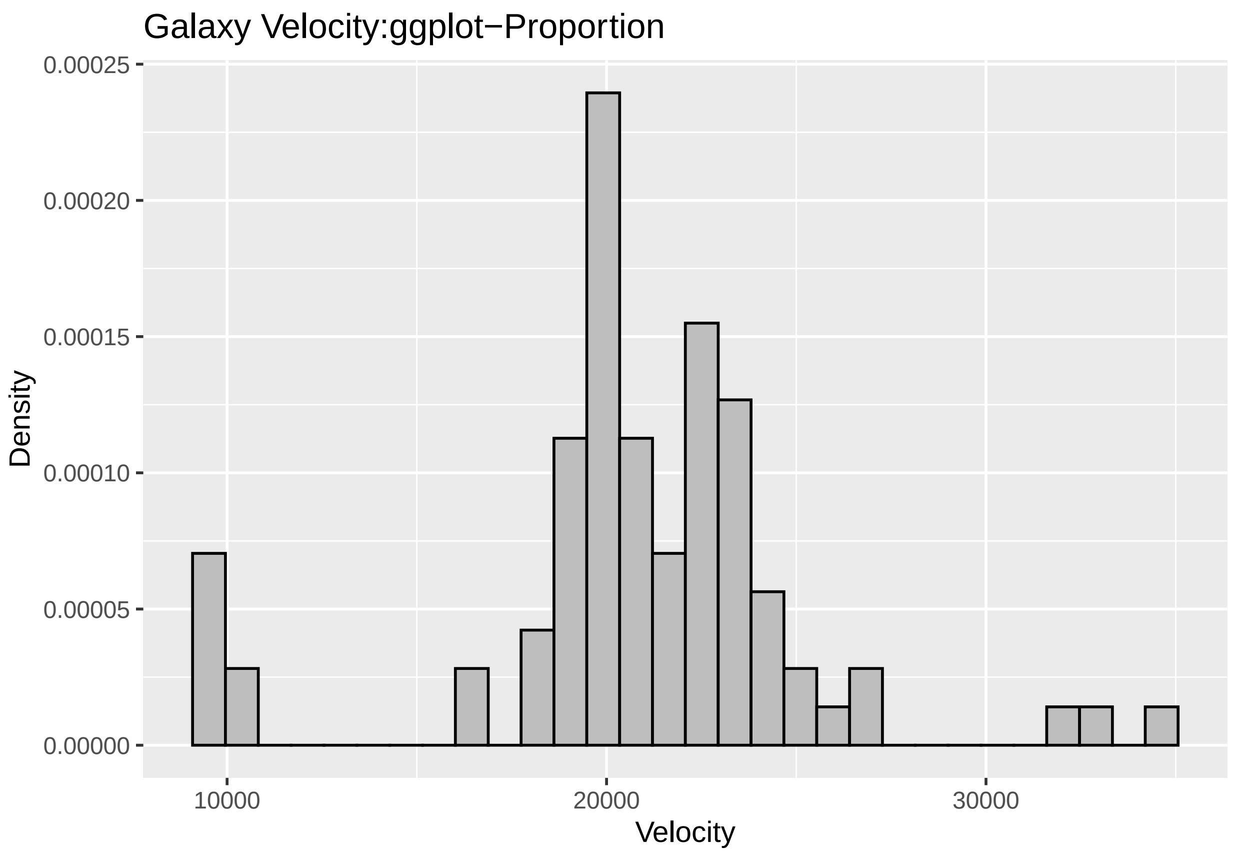
Answer 1.a: Applying hist() function from Graphics package on the velocities show that higher proportion of velocities are in between 15000 and 25000 km/sec. However, clustering of observations is observed at bin size of 30. A ggplot with default bin size of 30 also shows similar trend. The truehist function from Mass Package outputs similar histogram as histogram from graphics package. Only difference is that the bars are filled with color. The ggplot produce similar histogram even when using proportions. The qplot function from ggplot2 package also outputs similar histogram with 5 clusters with majority of observations in the center and 3 clusters with very few observations.The outline of each bar is not visible as they are filled. The 3 histogram functions from different packages yield similar output when the bin sizes are same.The observations show five clusters with majority of velocities in the center ranging between 15000 and 25000 km/sec.

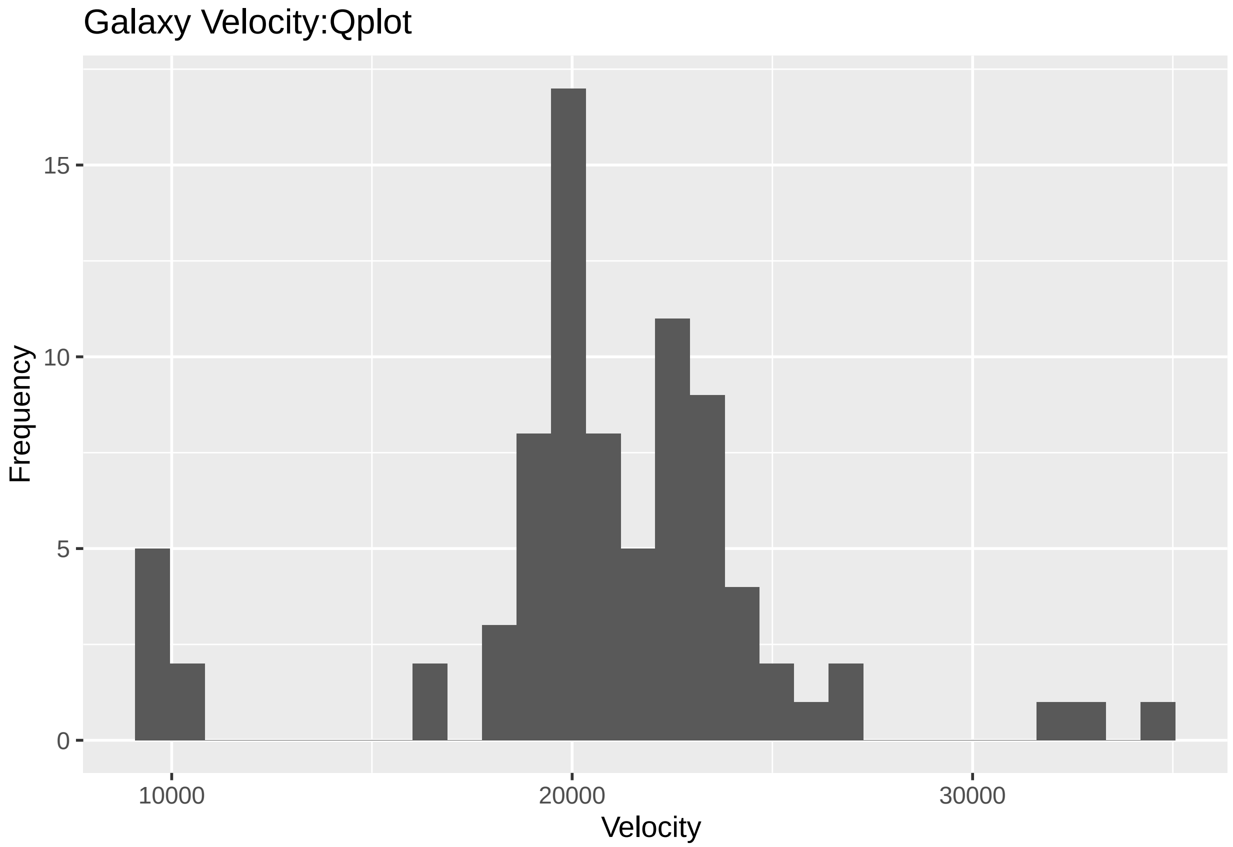

Galaxy Velocity:hist Function
Frequency
0
10
20
30
40
5000
15000
25000
35000
Velocity


Galaxy Velocity:hist−binsize:30
Frequency
0
5
10
15
10000
20000
30000
Velocity


Galaxy Velocity:ggplot−hist function
Frequency
15
10
5
0
10000
20000
30000
Velocity


Galaxy Velocity: truehist
Density
0e+00
4e−05
8e−05
5000
10000
15000
20000
25000
30000
35000
velocity

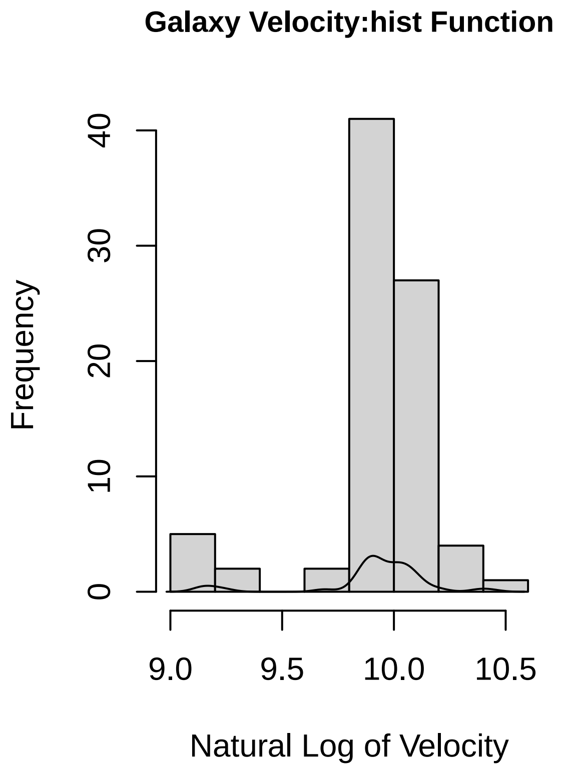

Galaxy Velocity:ggplot−Proportion
Density
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
10000
20000
30000
Velocity

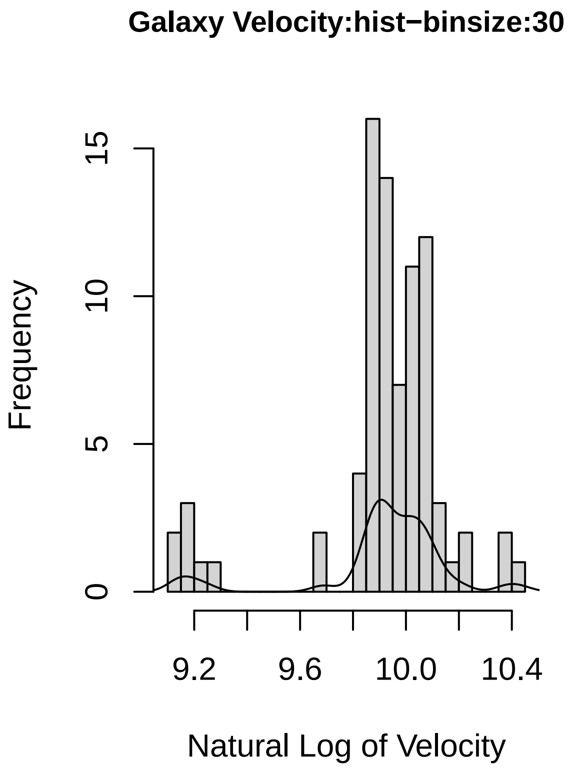


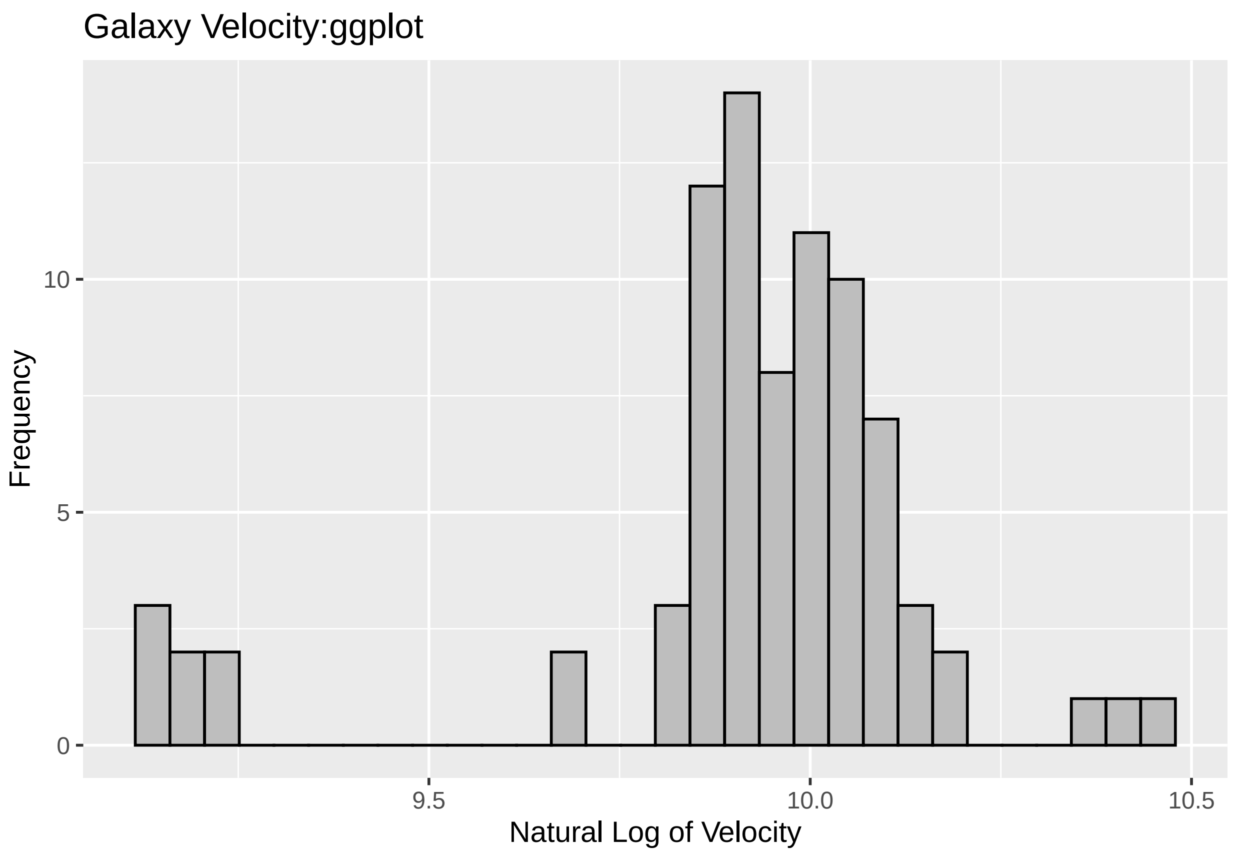
b) Create a new variable *loggalaxies* = log(galaxies). Repeat part a) using the `loggalaxies` variable. Does this affect your interpretation of the graphs?

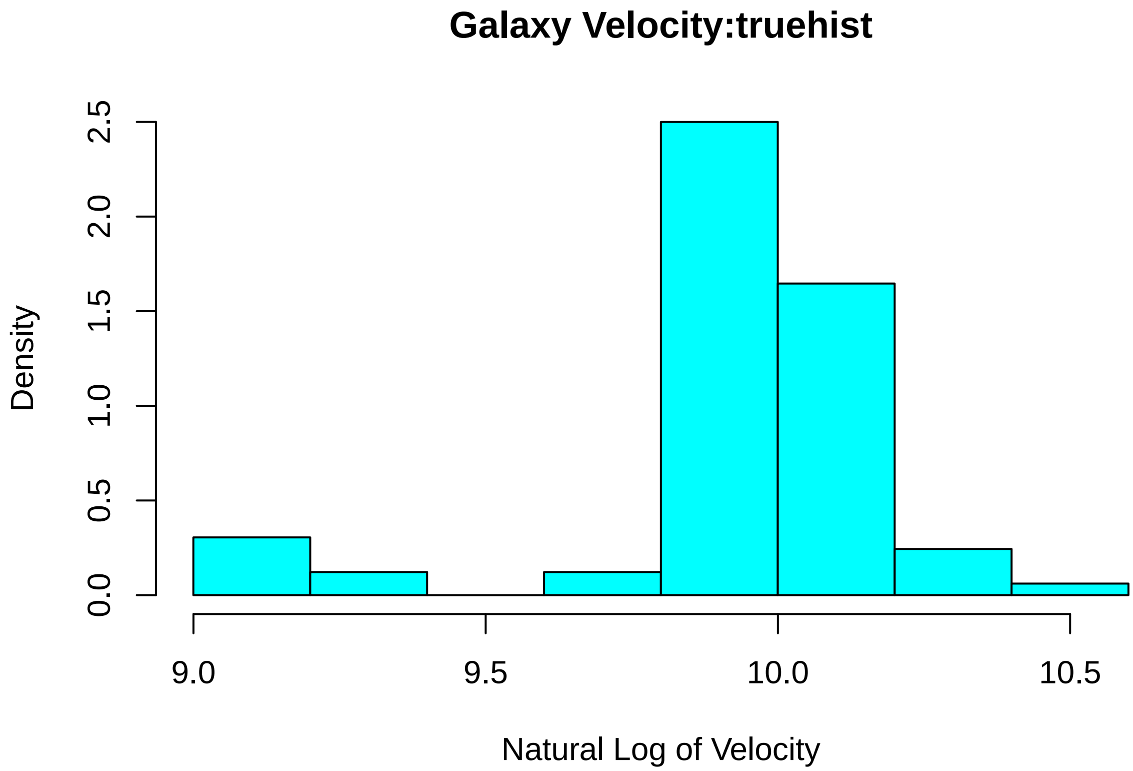
Answer 1.b: Applying hist() function from Graphics package on the log velocities show that the data is skewed towards left with a long tail to the left indicating a median closer to third quartile. It shows two clusters.However, 4 clusters are observed at bin size of 30. A ggplot with default bin size of 30 also shows similar trend. A density curve is added to the plots and it shows that the selected bin size is not good enough to show correct velocity densities and needs more investigation. The truehist function from Mass Package outputs similar histogram as histogram from graphics package. Only difference is that the bars are filled with color. The ggplot produce similar histogram even when using proportions. The qplot function from ggplot2 package also outputs similar histogram.

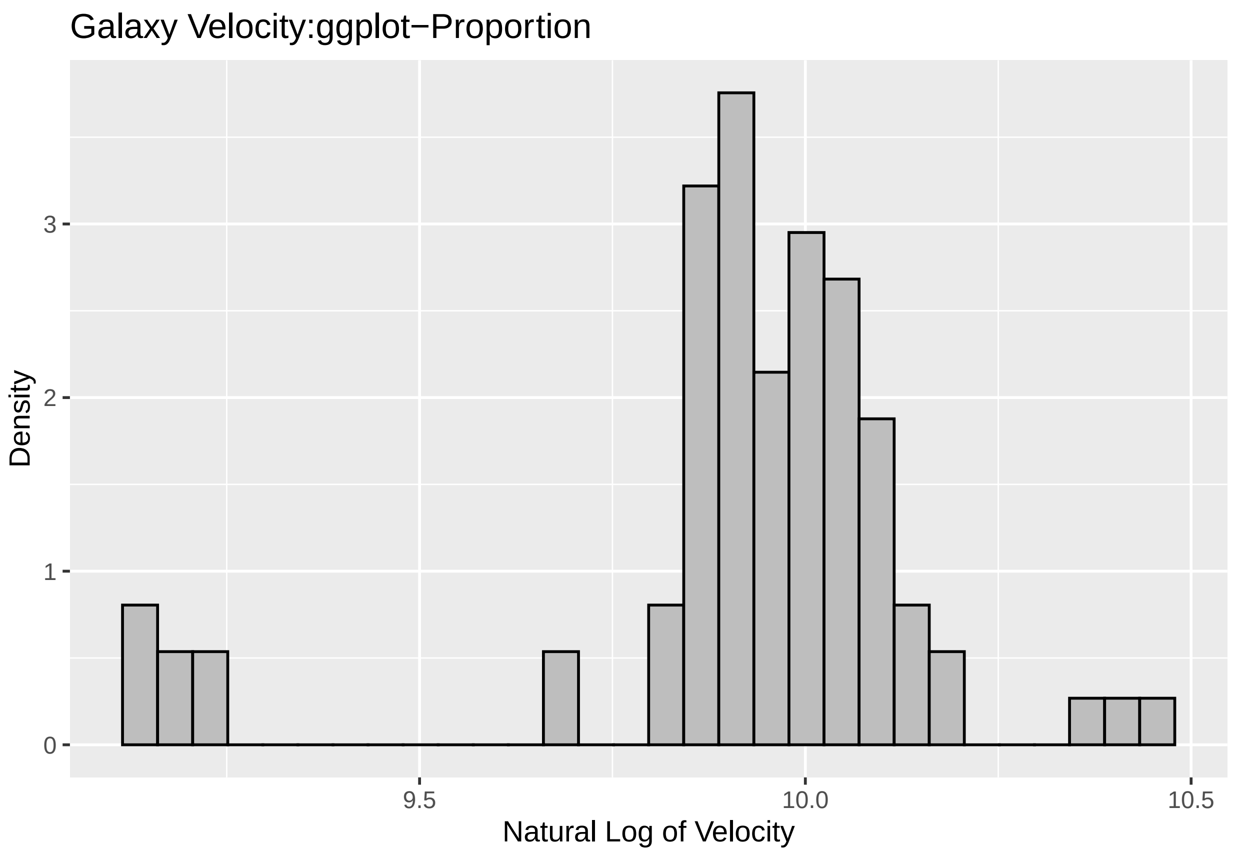
Histogram with velocities showed normal distribution with majority of observations at the center and 5 clusters (3 clusters with very few observations). But, the log velocities show a left handed skewness with 4 clusters,affecting my previous interpretation.I believe that, as the two clusters towards the higher velocities are not too far,taking natural log have grouped them into one cluster.

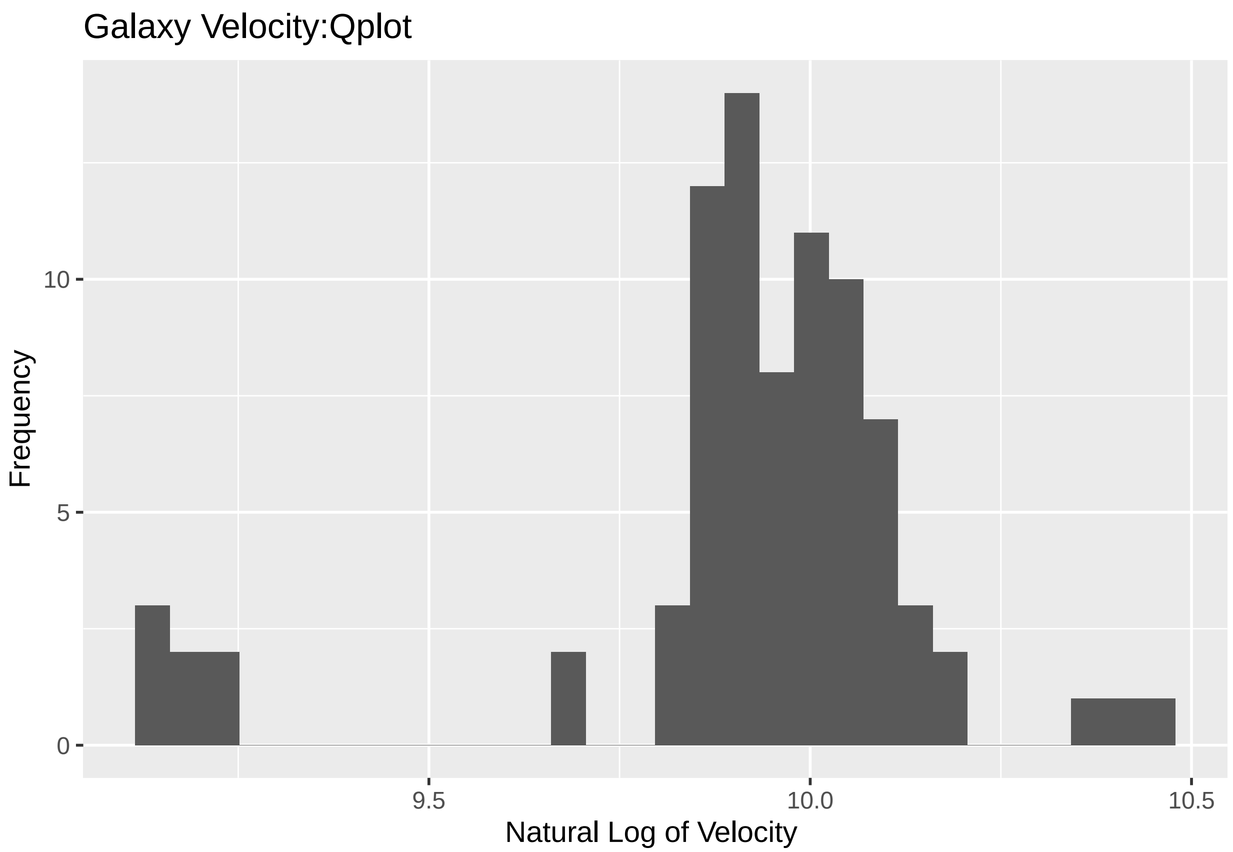

Galaxy Velocity:hist Function
Frequency
0
10
20
30
40
9.0
9.5
10.0
10.5
Natural Log of Velocity


Galaxy Velocity:hist−binsize:30
Frequency
0
5
10
15
9.2
9.6
10.0
10.4
Natural Log of Velocity


Galaxy Velocity:ggplot
Frequency
10
5
0
9.5
10.0
10.5
Natural Log of Velocity

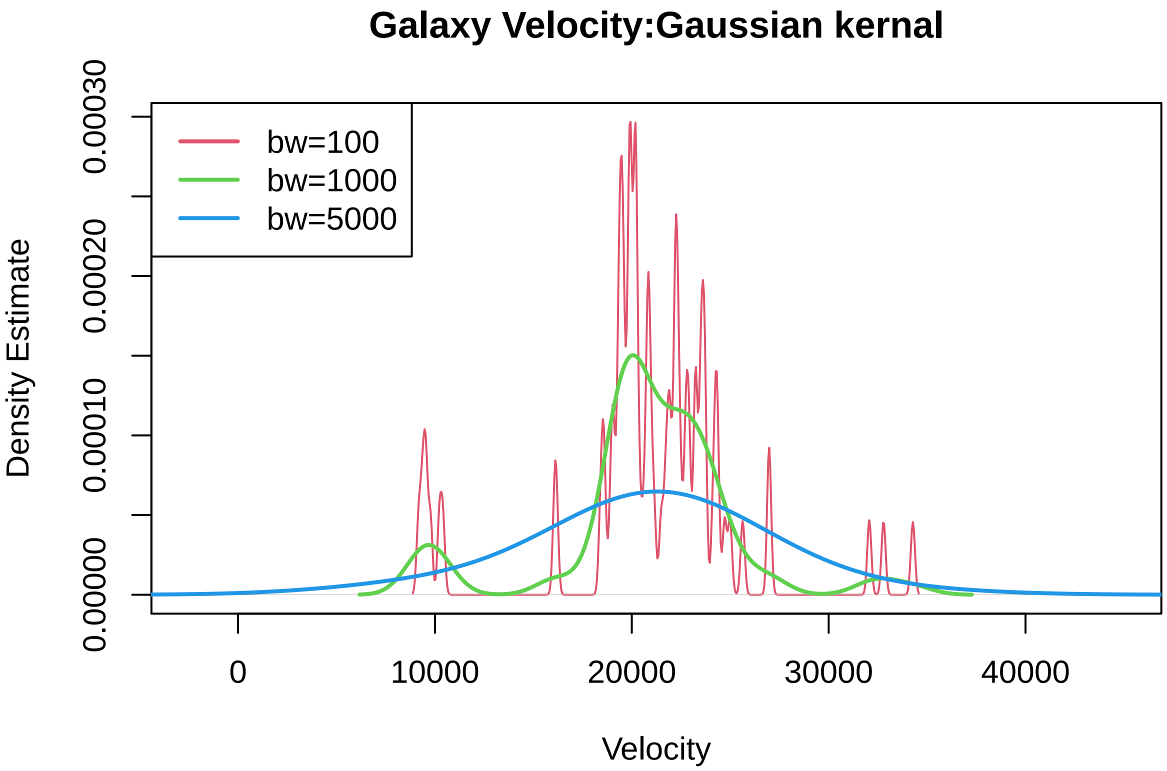

Galaxy Velocity:truehist
Density
0.0
0.5
1.0
1.5
2.0
2.5
9.0
9.5
10.0
10.5
Natural Log of Velocity

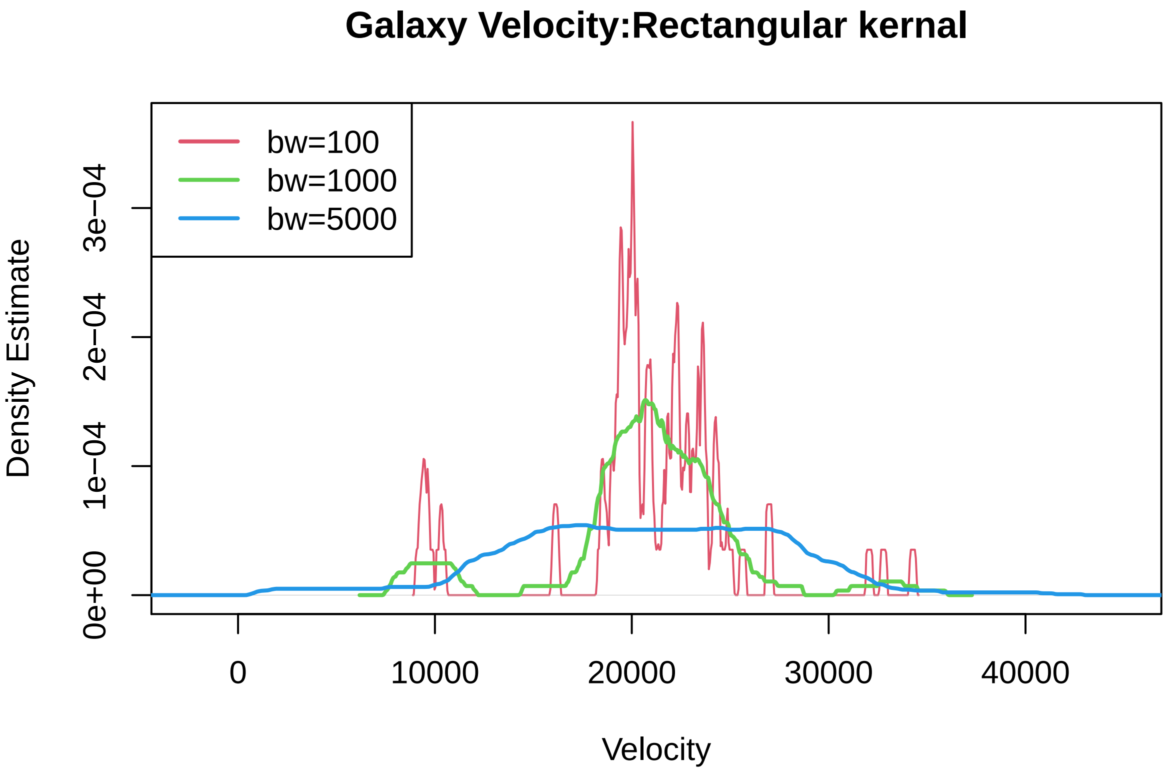

Galaxy Velocity:ggplot−Proportion
Density
0
1
2
3
9.5
10.0
10.5
Natural Log of Velocity



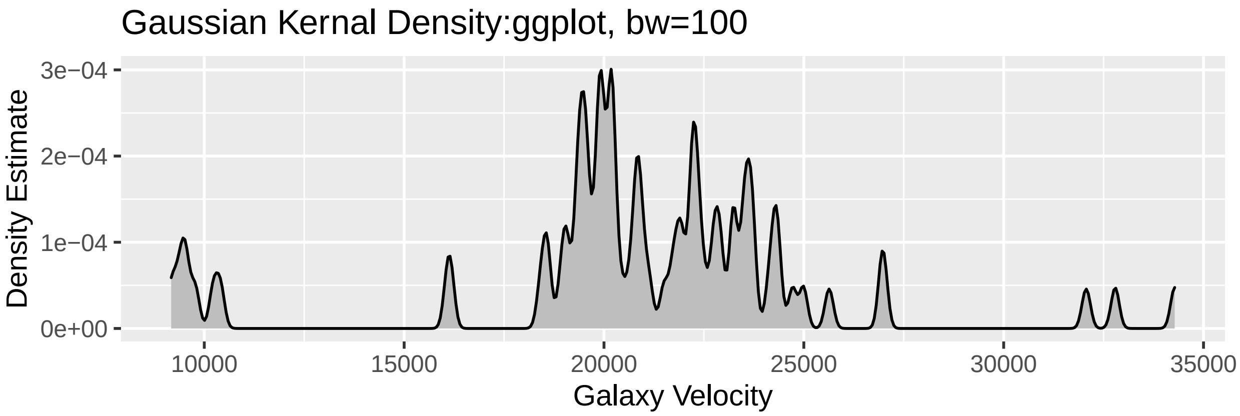
c) Construct kernel density estimates using two different choices of kernel functions and three choices of bandwidth (one that is too large and "oversmooths," one that is too small and "undersmooths," and one that appears appropriate.) Therefore you should have six different kernel density estimates 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.

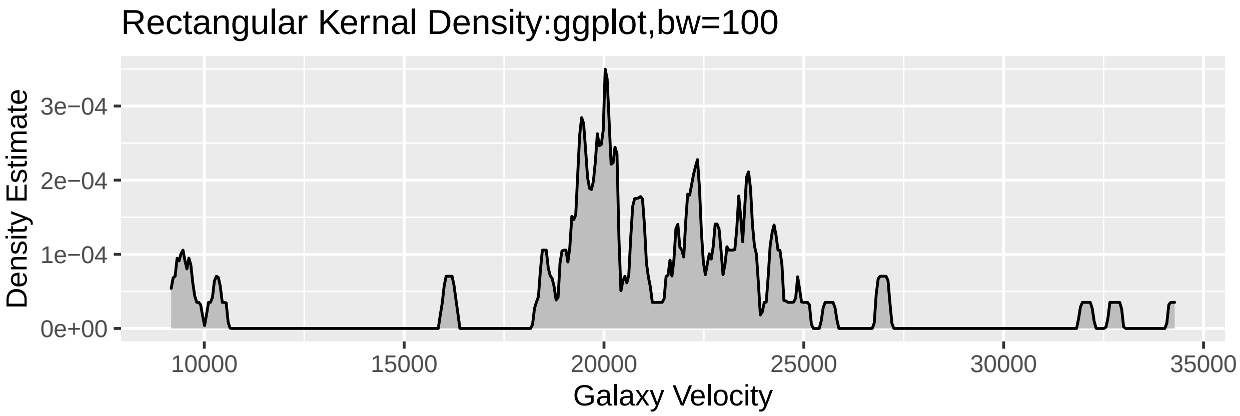
Answer 1.c: I have plotted density estimates of galaxies velocity using Gaussian and Rectangular kernel functions with bandwidths 100(too low and undersmooths),1000(Appropriate) and 5000(too large and oversmooths).For both the kernel functions used, a bandwidth of 100 gave a noisy and overfit estimate, a bandwidth of 1000 gave a smooth estimate distinguishing the clusters of observations and a bandwidth of 5000 is too big and oversmooths the density curve.The plots are separated by the kernel function. Similar trend is observed in plots obtained using ggplot package.The plots are separated by the Bandwidth.


Galaxy Velocity:Gaussian kernal
bw=100
bw=1000
bw=5000
Density Estimate
0.00000
0.00010
0.00020
0.00030
0
10000
20000
30000
40000
Velocity

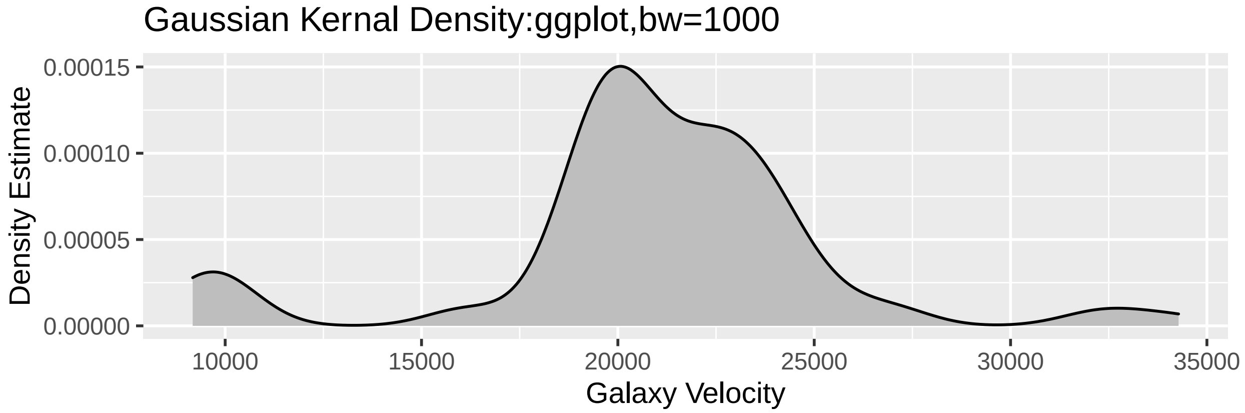

Galaxy Velocity:Rectangular kernal
bw=100
bw=1000
bw=5000
Density Estimate
0e+00
1e−04
2e−04
3e−04
0
10000
20000
30000
40000
Velocity

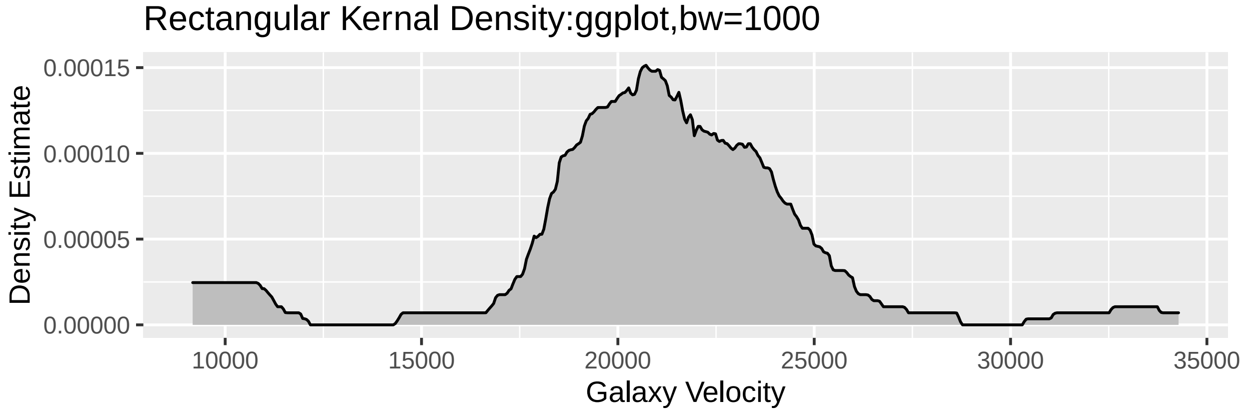
Undersmoothing Bandwidth



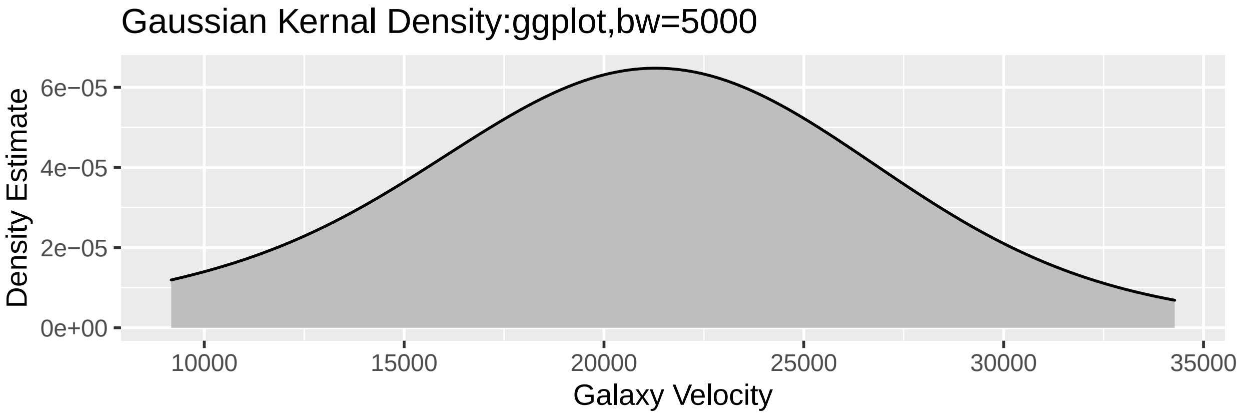


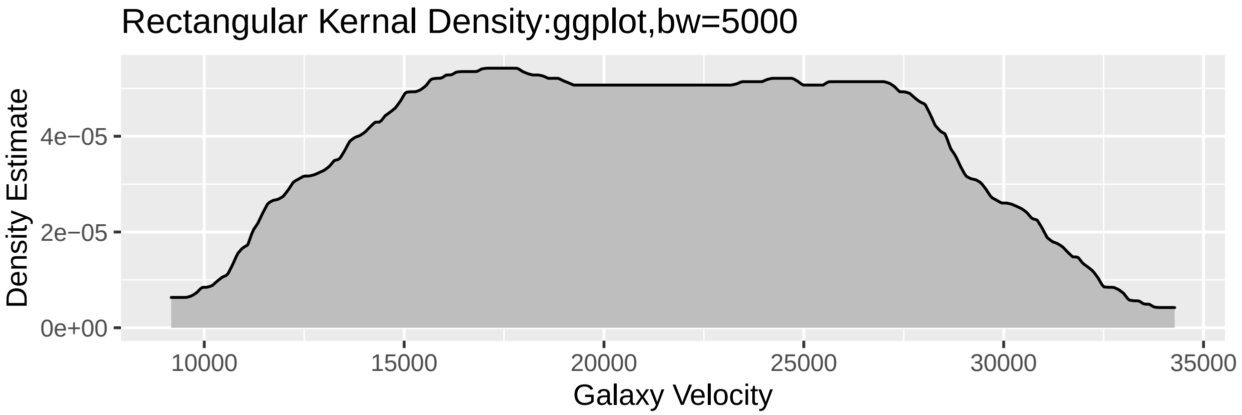
Appropriate Bandwidth





Oversmoothing Bandwidth





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.)

Answer 1.d: Both the histogram and the density plot with appropriate bandwidth shows that the observations are clustered forming an upper and lower tail. I would conclude that 4 superclusters of galaxies exist with large voids in between them like the density curves suggests.

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.

Answer 1.e: The Summary of Finite mixture model shows 4 clusters and confirms the results from graphical inspection.Majority of the observations are in clusters 2 and 3 ranging from average velocities of 19807 to 22880 km/sec.

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -765.7316 82 11 -1579.937 -1598.809
##
## Clustering table:
##  1  2  3  4
##  7 35 32  8
##
## Mixing probabilities:
##          1          2          3          4
## 0.08441927 0.38768587 0.36896338 0.15893147
##
## Means:
##         1         2         3         4
##  9707.522 19806.592 22880.348 24483.603
##
## Variances:
##          1          2          3          4
##   177311.8   437746.2  1231115.8 34305975.7
```

The parameter estimates of the finite mixture models shows the proportions of the observations in each cluster (8%,39%,37% and 16% in clusters, 1,2,3,4 respectively). They also provide mean/center of each cluster, variance type(unequal variance) and the variance of each cluster.

```
## $pro
## [1] 0.08441927 0.38768587 0.36896338 0.15893147
##
## $mean
##         1         2         3         4
##  9707.522 19806.592 22880.348 24483.603
##
## $variance
## $variance$modelName
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 4
##
## $variance$sigmasq
## [1]   177311.8   437746.2  1231115.8 34305975.7
##
## $variance$scale
## [1]   177311.8   437746.2  1231115.8 34305975.7
```

The BIC of the velocities indicates that unequal variance with 4 models (with BIC of -1579.937) is the best model (4 super clusters) with high penalty and flexibility.

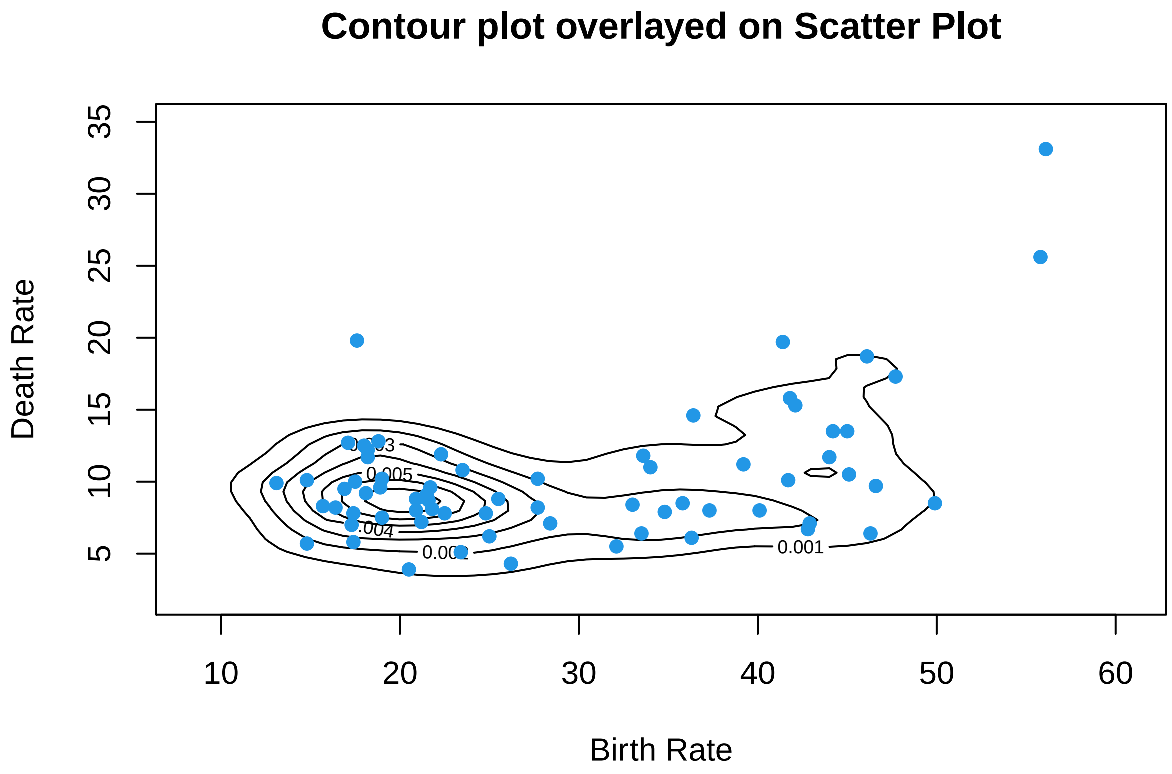
```
## Bayesian Information Criterion (BIC):
##           E         V
## 1 -1622.518 -1622.518
## 2 -1631.401 -1595.633
## 3 -1584.673 -1592.408
## 4 -1593.485 -1579.937
## 5 -1593.361 -1593.345
## 6 -1602.266 -1604.112
## 7 -1589.153 -1611.579
## 8 -1597.984 -1625.847
## 9 -1601.089 -1633.533
##
## Top 3 models based on the BIC criterion:
##       V,4       E,3       E,7
## -1579.937 -1584.673 -1589.153
```

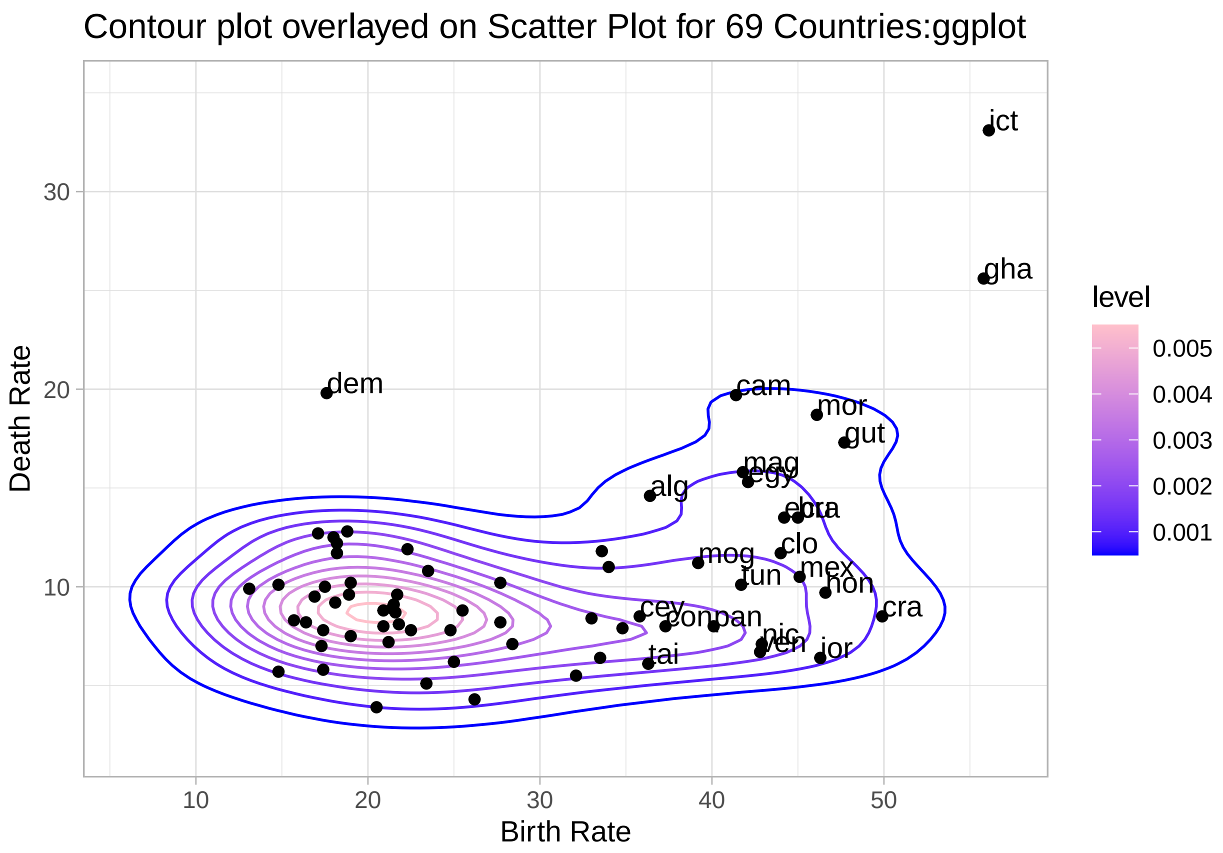
Question 2. (Ex. 8.2 in HSAUR, modified for clarity) The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (from Hartigan, 1975).

References: Ref1- Chapter_8_modified.R,Ref9, Ref11, Ref12, Ref13

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

Answer 2.a: A bivariate density is estimated using bkde2D function, a contour plot is plotted and overlayed on the scatter plot of the data.



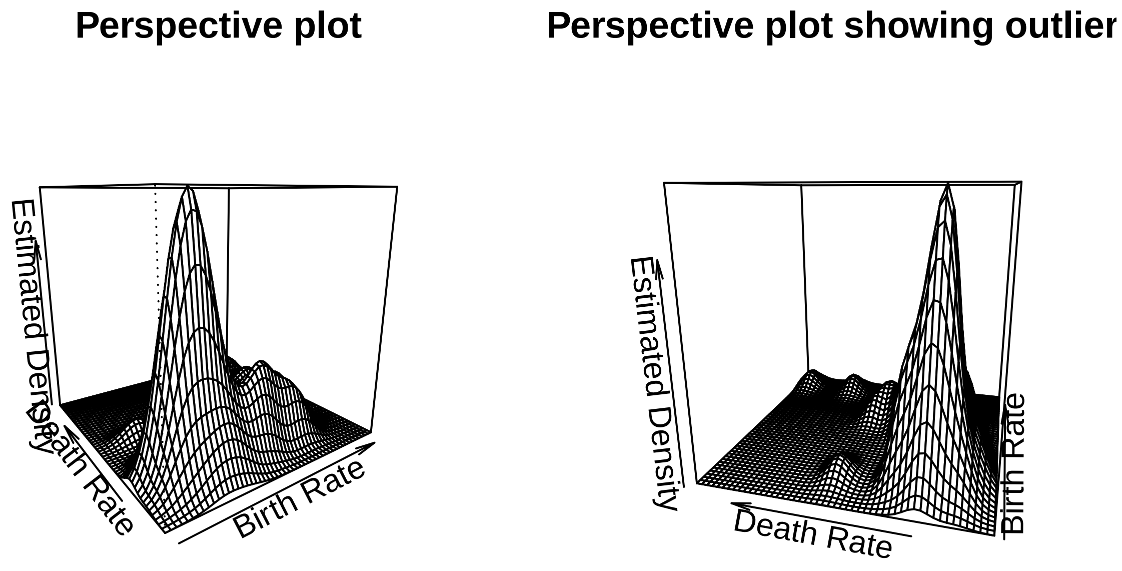


b) What does the contour plot tell you about the structure of the data?

Answer 2.b: The Contour plot shows that for most of the countries the birth rates are concentrated around 13-28 and death rates concentrates around 3-14 with approximately 2:1 ratio. For other countries birth rate is between 32-50 and death rate between 5 and 20. Though the observations are not separated as clusters one can see that the observations are concentrated at 2 different locations on the plot. The plot also shows that there are few observations away from the centroids of the assumed clusters causing a spike.

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

Answer 2.c: Two Perspective plots were produced. One shows high density of observations between 12-50 birthrates and 7-14 death rates. The second plot shows the 3 outliers.



Perspective plot

Perspective plot showing outlier

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.

Answer2.d: The Summary of the Finite Mixture Model gives 4 clusters. Cluster 3 has higher number of observations with a birth rate to death rate ratio of about 2:1 (confirms the observation from scatter plot).It also provides the means and variances of both variables for all 4 clusters.

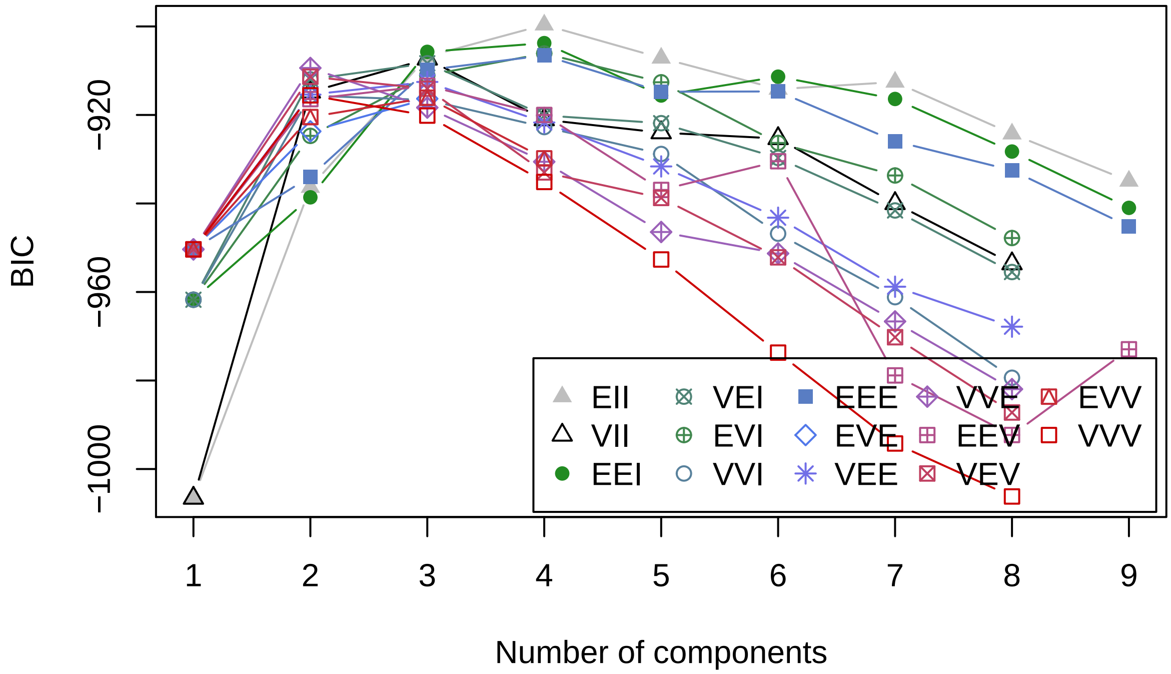
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194 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
```

The BIC of the model shows that 2 of the top 3 models has 4 components/clusters and one has 3 clusters.

```
## Bayesian Information Criterion (BIC):
##          EII        VII       EEI       VEI       EVI       VVI       EEE
## 1 -1006.5723 -1006.5723 -961.7502 -961.7502 -961.7502 -961.7502 -950.3669
## 2  -936.3442  -914.8037 -938.6127 -911.9710 -924.7310 -915.6217 -933.9448
## 3  -906.7729  -907.3547 -905.7403 -908.3174 -911.0701 -916.6248 -909.8428
## 4  -899.6481  -921.0631 -903.7704 -920.1226 -906.1018 -922.7386 -906.5496
## 5  -907.1378  -924.0068 -915.6050 -921.8611 -912.6162 -928.8162 -914.7571
## 6  -914.1679  -925.3259 -911.3484 -929.7137 -926.3244 -946.8290 -914.6918
## 7  -912.5610  -940.0067 -916.3920 -941.5804 -933.6770 -961.1733 -925.9343
## 8  -924.2724  -953.6153 -928.2698 -955.4928 -947.8093 -979.3765 -932.5095
## 9  -934.9379         NA -940.9908        NA        NA        NA -945.1889
##         EVE       VEE       VVE       EEV       VEV       EVV        VVV
## 1 -950.3669 -950.3669 -950.3669 -950.3669 -950.3669 -950.3669  -950.3669
## 2 -923.7050 -915.4055 -909.3891 -916.4290 -911.3583 -920.4713  -915.5710
## 3 -916.3323 -912.5420 -918.3377 -913.3972 -914.0597 -916.1073  -920.1468
## 4        NA -921.7029 -930.5803 -920.0012 -932.9836 -929.8081  -935.1407
## 5        NA -931.6311 -946.4479 -936.9447 -938.7558        NA  -952.6602
## 6        NA -943.2135 -951.2986 -930.4589 -952.1768        NA  -973.6995
## 7        NA -958.8094 -966.6536 -978.8477 -970.2239        NA  -994.2301
## 8        NA -967.8431 -981.9471 -992.3116 -987.2295        NA -1006.1989
## 9        NA        NA        NA -972.9489        NA        NA         NA
##
## Top 3 models based on the BIC criterion:
##     EII,4     EEI,4     EEI,3
## -899.6481 -903.7704 -905.7403
```

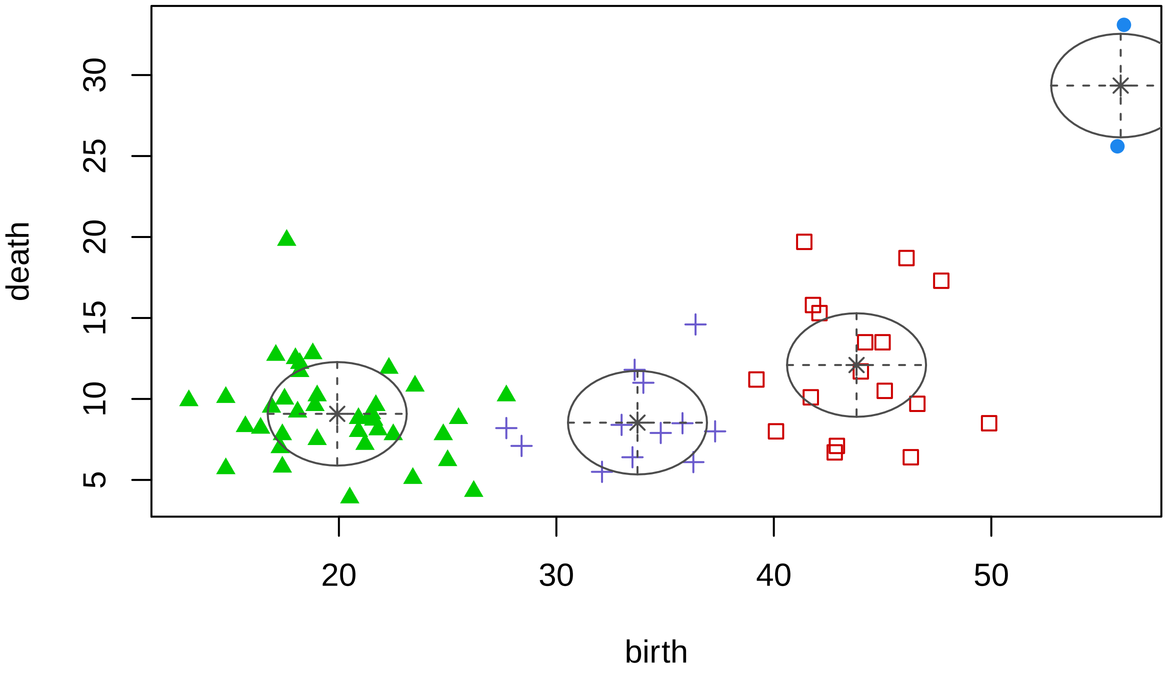
BIC PLOT

The BIC plot shows that the model EII with Spherical distribution and equal volume is best finite mixture model with highest BIC at 4 clusters/components (BIC: -899.6481).



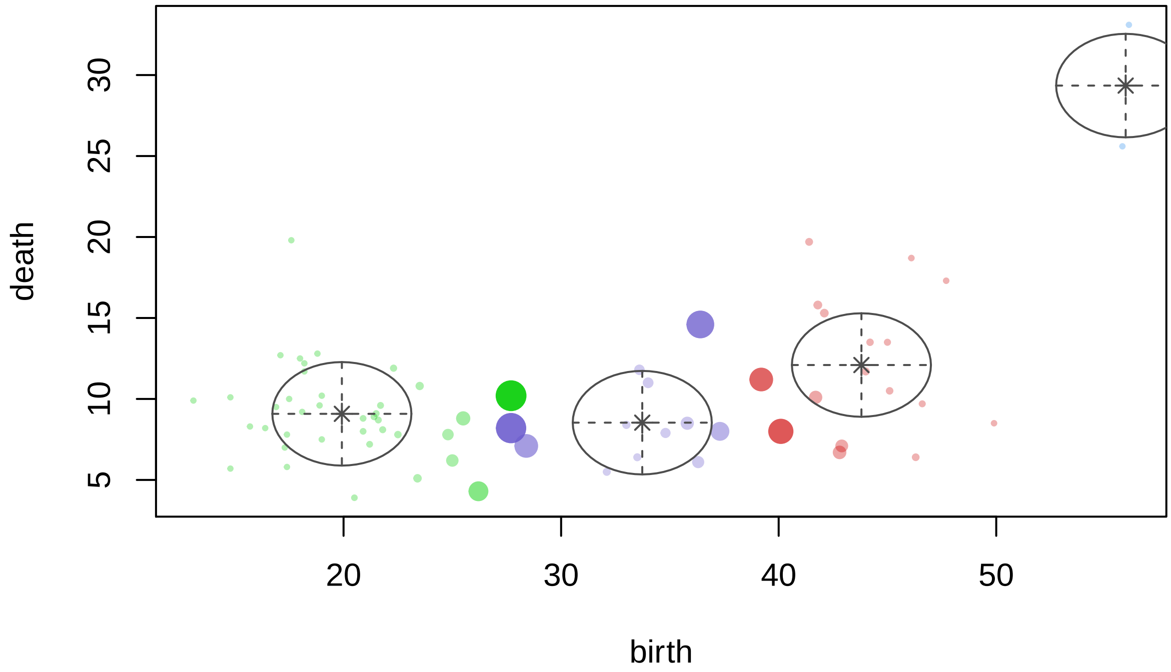
CLASSIFICATION PLOT

This plot shows that the observations are concentrated around 4 points forming 4 cluster which is inline with the summary of the model used.



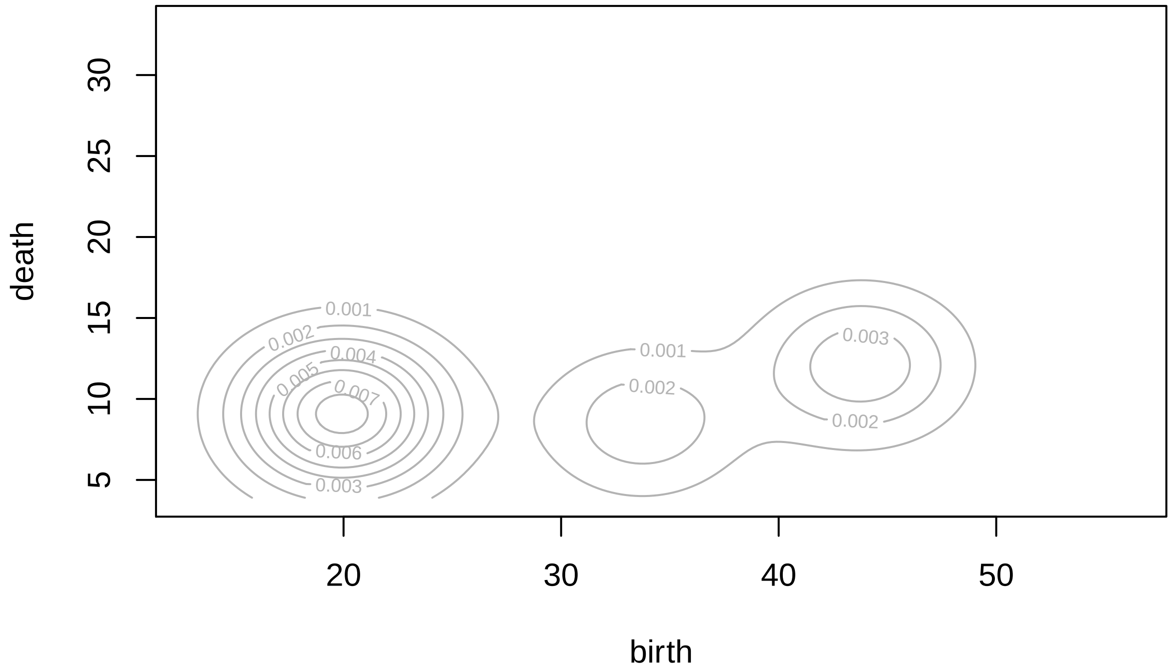
UNCERTAINITY PLOT

The Uncertainty plot indicates that there is an uncertainty in the classification of the observations in the model.Considering the proportions of observations in each cluster there is a large uncertainty in the second cluster.



DENSITY PLOT

This plot shows the density/ spread of the observations.



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

Answer 2.e: All the density plots and the finite mixture models suggest that the birth rate/death rate of the countries can be separated as 4 clusters. The cluster with more number of countries has a birth rate to death rate ratio of 2:1.There are 2 countries(ict,gha) with highest birth rate than death rate and one country (dem) has an opposite situation.The density plots can be used to understand the distribution of the data but does not give a lot of information.

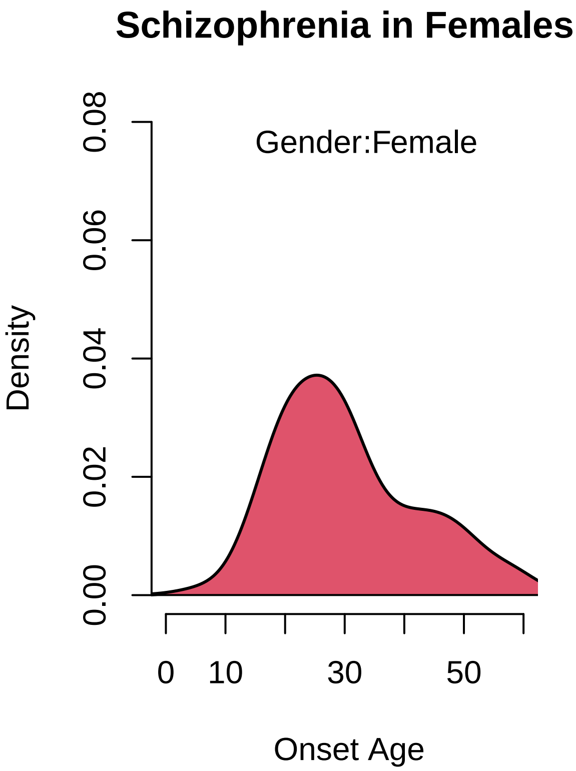
Question 3. (Ex. 8.3 in HSAUR, modified for clarity) Fit finite mixtures of normal densities individually for each gender in the **schizophrenia** data set from **HSAUR3**. Do your models support the *sub-type model* described in the R Documentation?

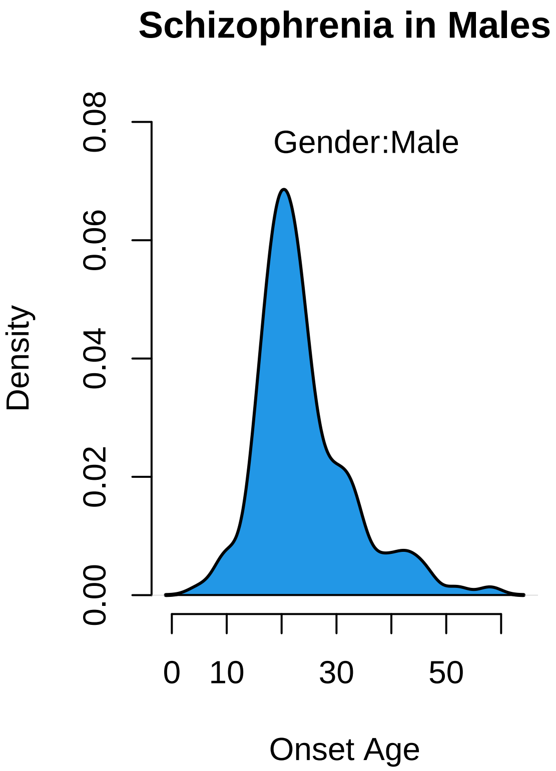
Quote from the R Documentation: *A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequent epidemiological studies of the disorder have consistently shown an earlier onset in men than in women. One model that has been suggested to explain this observed difference 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)

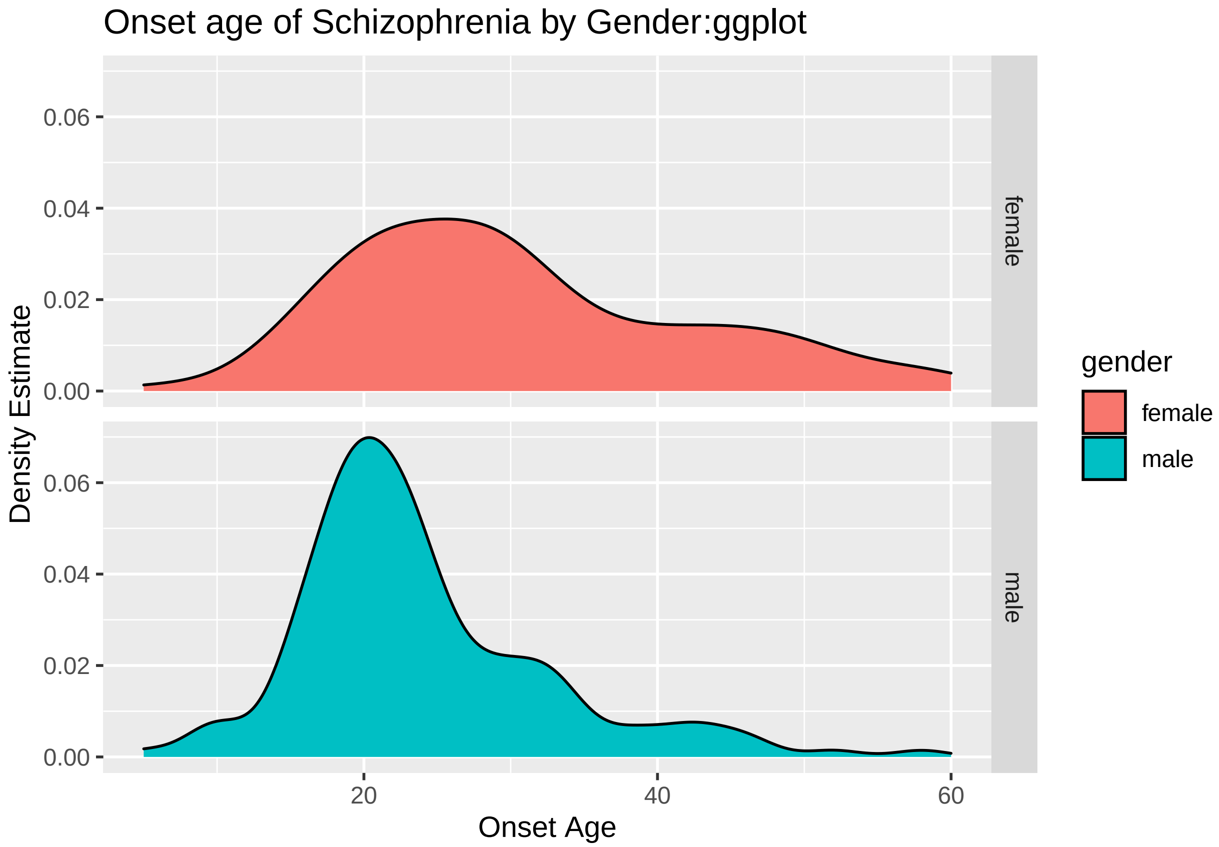
References: Ref1- Chapter_8_modified.R,Ref11, Ref9

Answer 3:

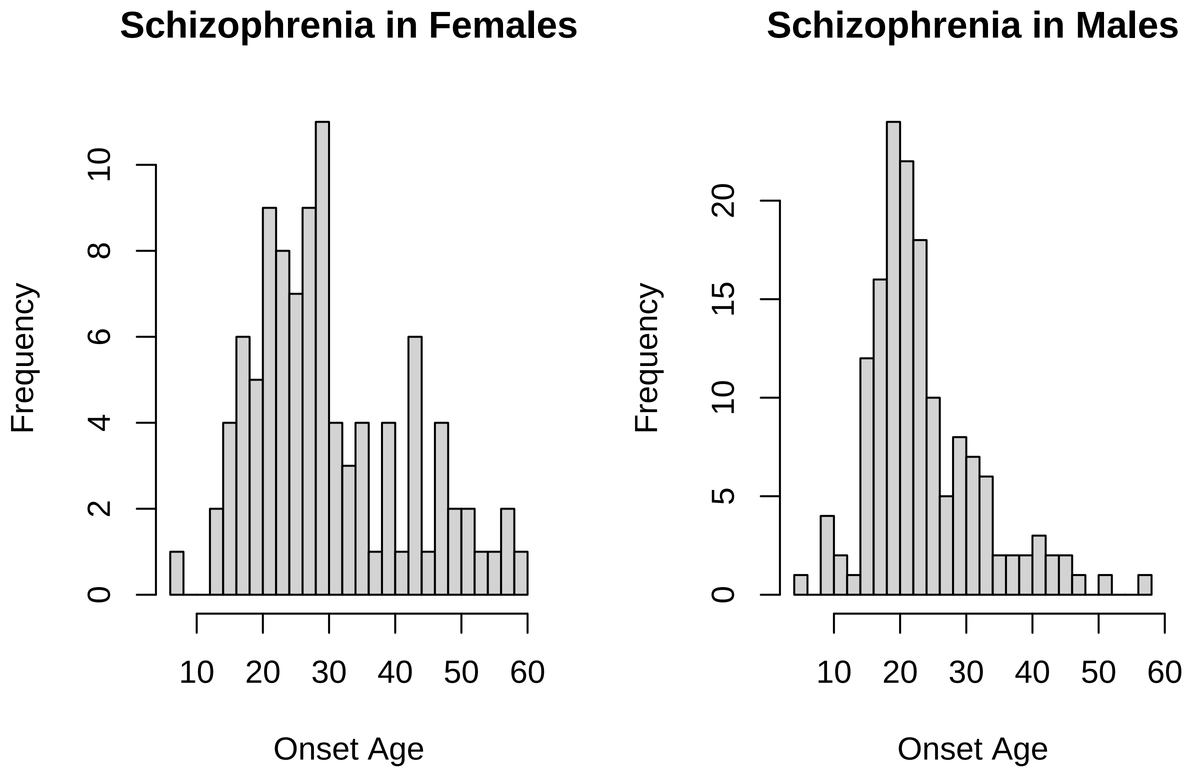
The Density plots of the onset ages of males and females show that the onset age in men is around 15-30 and in females it is spread throughout all the age groups.

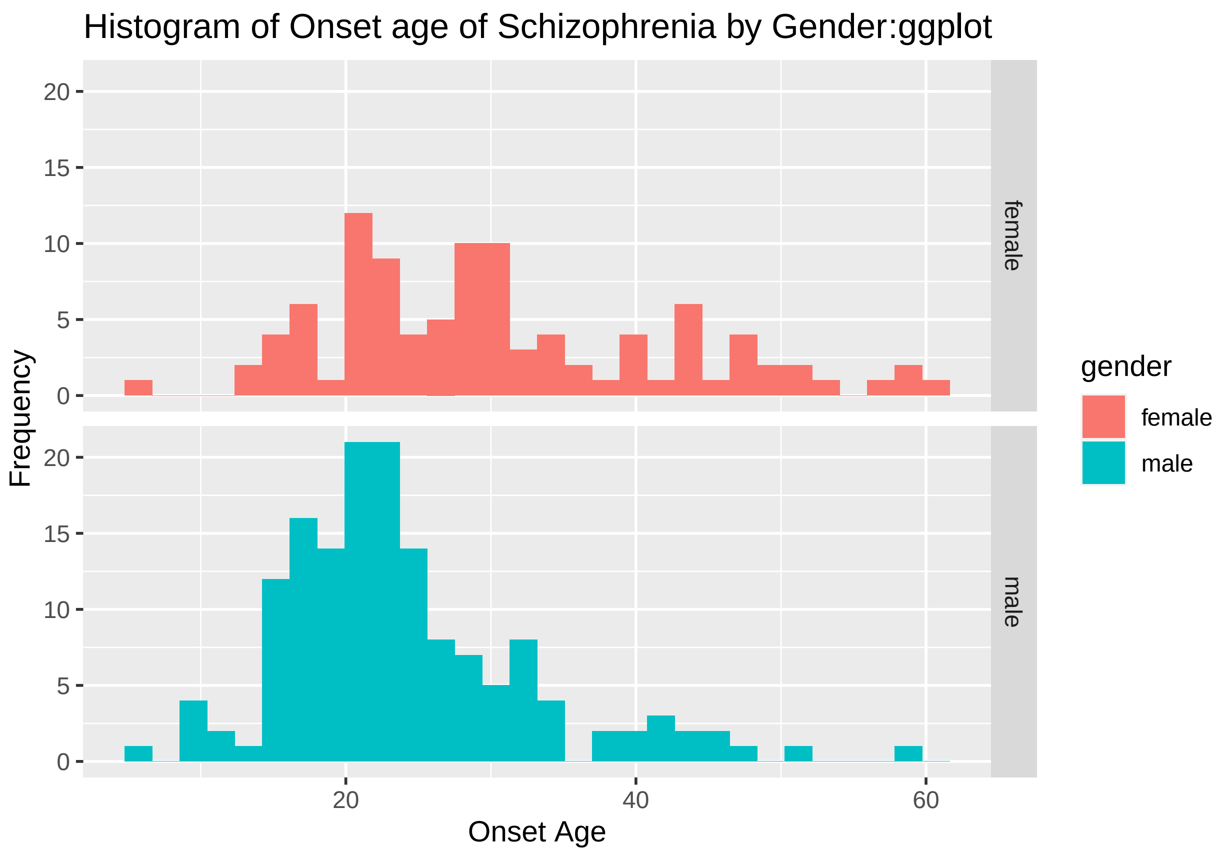

Schizophrenia in Females
Gender:Female
Density
0.00
0.02
0.04
0.06
0.08
0
10
30
50
Onset Age


Schizophrenia in Males
Gender:Male
Density
0.00
0.02
0.04
0.06
0.08
0
10
30
50
Onset Age



The Histograms of the onset ages of males and females show similar pattern as the Density plots.


Schizophrenia in Females
Frequency
0
2
4
6
8
10
10
20
30
40
50
60
Onset Age
Schizophrenia in Males
Frequency
0
5
10
15
20
10
20
30
40
50
60
Onset Age



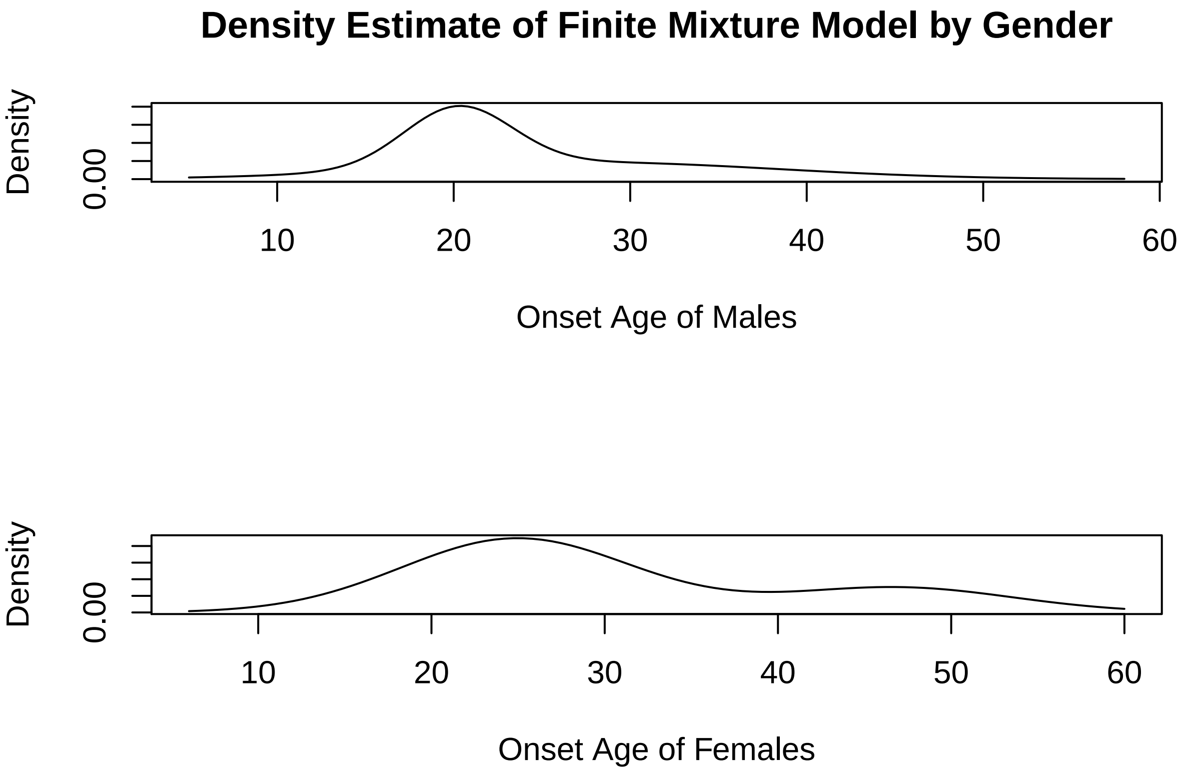
The Mclust model of male onset ages forms 2 univariate, unequal variance clusters with a mean onset ages of 20 and 28.

```
## ----------------------------------------------------
## 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
```

The Mclust model of female onset ages forms 2 univariate,equal variance clusters with a mean onset ages of 25 and 47.Though 75% of the observations have an average onset age of 25,a average late onset age of 47 is only found in female schizophrenia patients.

```
## ----------------------------------------------------
## 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
```

Density Estimate of finite mixture models clearly show an early onset in males and both early and late onset in females.



Final Answer: The summary of Finite Mixture model for female onset ages supports the hypothesis of the subtype model that “the late onset largely is a disorder of women”.However, an early onset is observed in both males (100% of the Schizophrenia data) and females (75% of the schizophrenia data) between the average ages of 20-27.This can be clearly observed in both histogram and the density plots of onset ages separated by gender.Though 100% of males have early onset as per the assumptions of the subtype model that says “the early onset type is assumed to be largely a disorder of men”, my model suggest that early onset is also observed in majority of females.My models supports the subtype model to certain extent but not 100%.