Density Estimation

Mohamed Ahmed

Exercises

Warning: There are only three questions, however they will require more time coding. You may need to review function calling conventions and whether the optional arguments and their default parameters are appropriate.

1. (Ex. 8.1 in HSAUR, modified for clarity) The data from contains the velocities of 8 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)

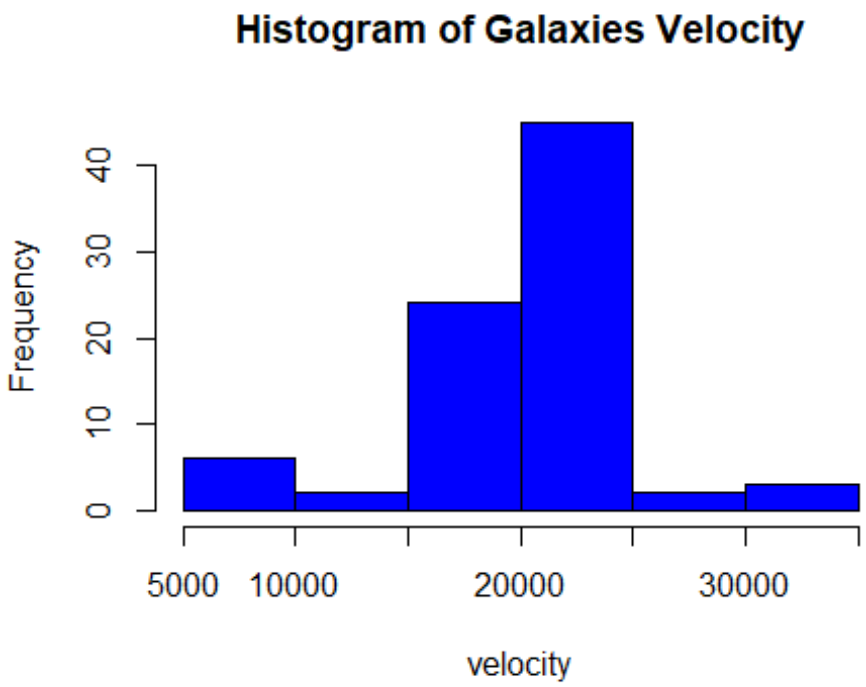
    a) Construct histograms using the following functions:

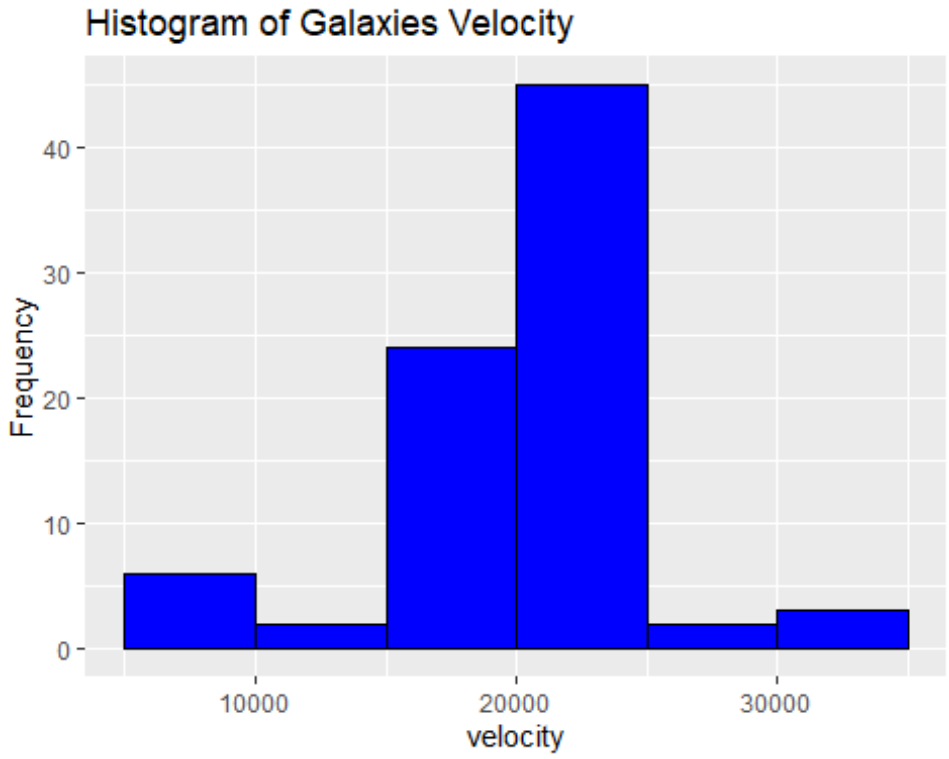
    -hist() and ggplot()+geom_histogram()

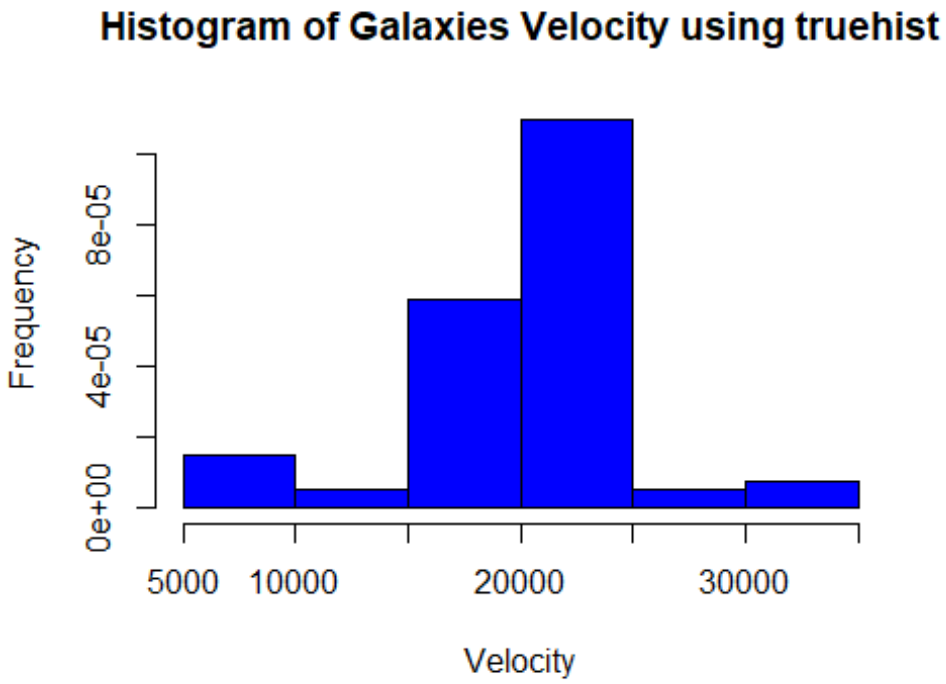
    -truehist() and ggplot+geom_histogram() (make sure that the histograms show proportions, not counts.)

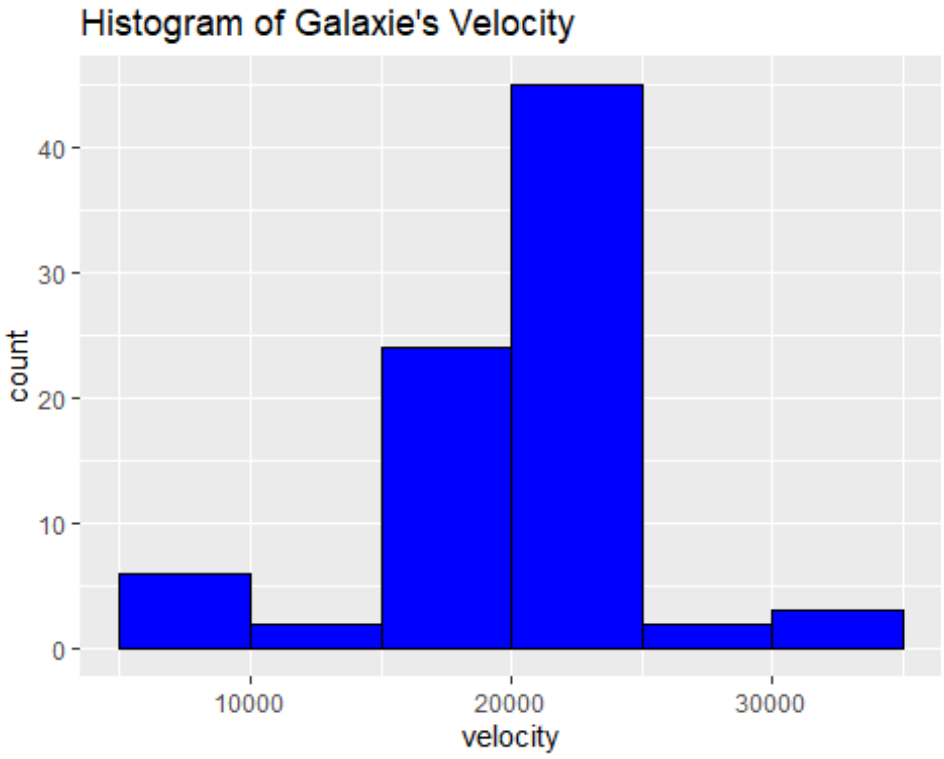
    -qplot()

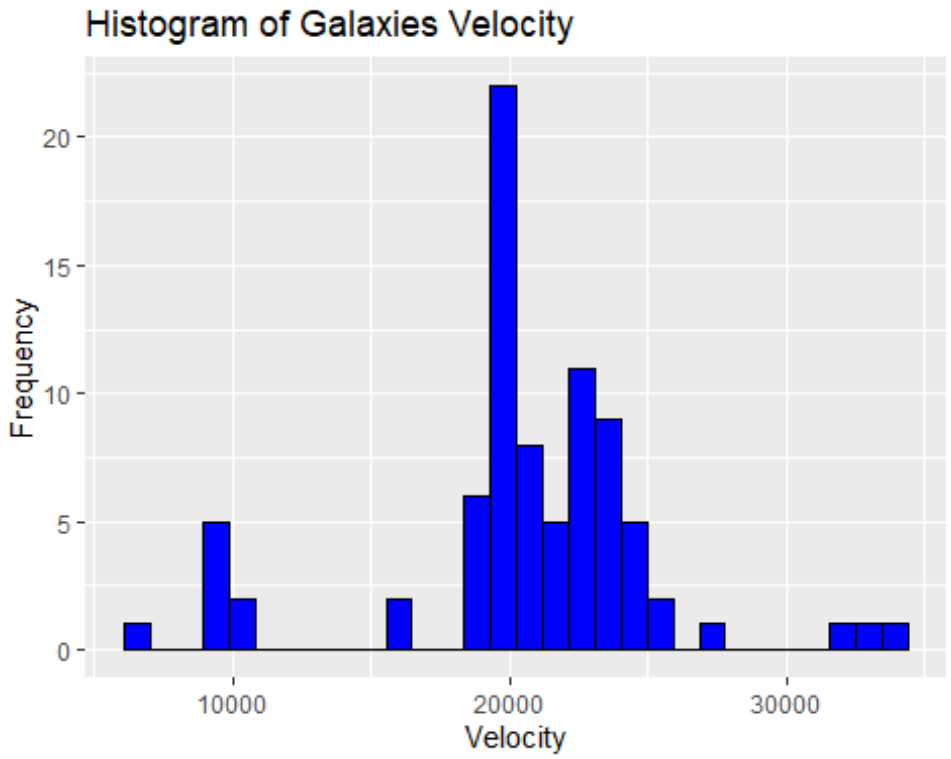
    Comment on the shape and properties of the variable based on the five plots. Do you notice any sets of observations clustering? (Hint: You can adjust bin number or bin size as you try to determine the properties of the variable, but use the same bin settings between plots in your final analysis. You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


Histogram of Galaxies Velocity
Frequency
0
10
20
30
40
5000
10000
20000
30000
velocity


Histogram of Galaxies Velocity
Frequency
0
10
20
30
40
10000
20000
30000
velocity

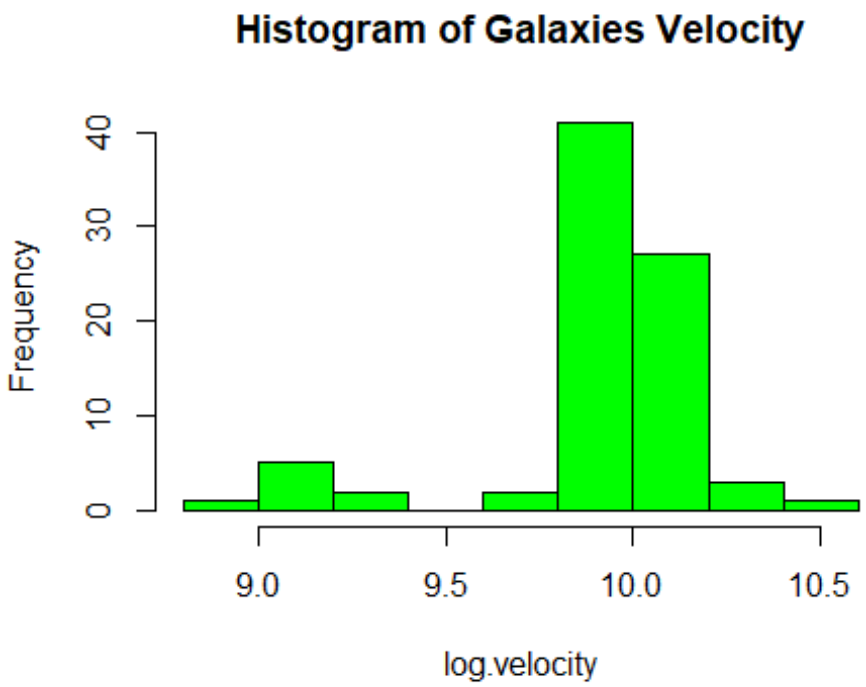

Histogram of Galaxies Velocity using truehist
Frequency
0e+00
4e-05
8e-05
5000
10000
20000
30000
Velocity

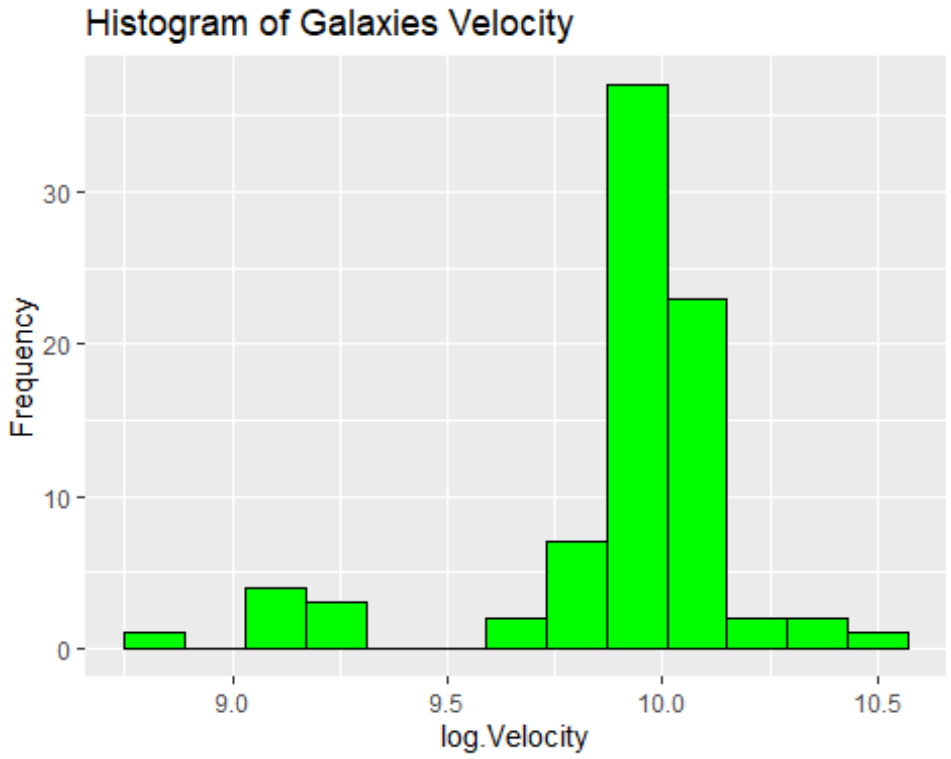

Histogram of Galaxie's Velocity
count
0
10
20
30
40
10000
20000
30000
velocity

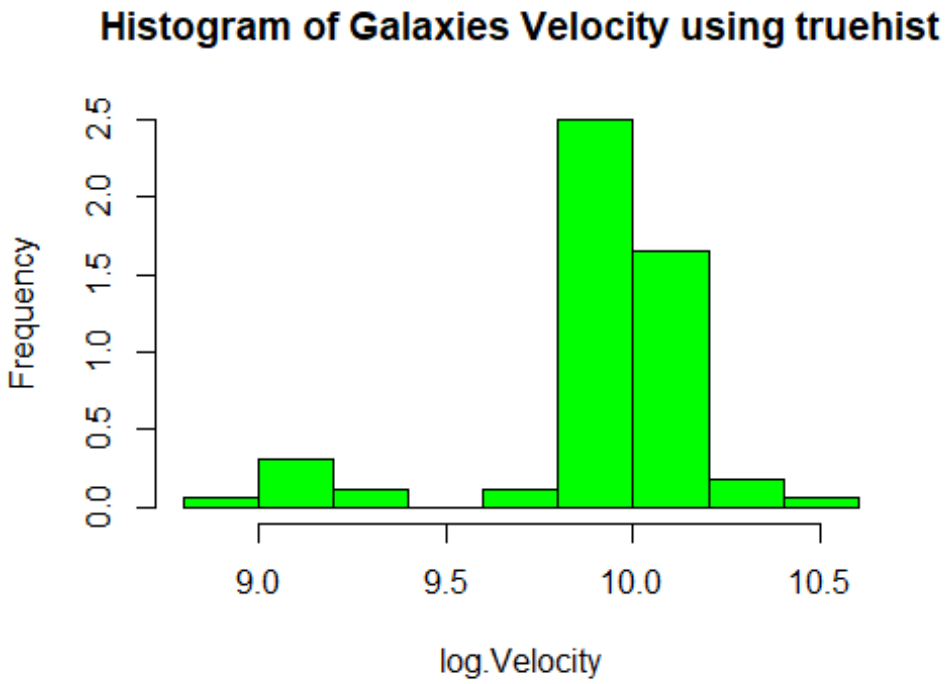


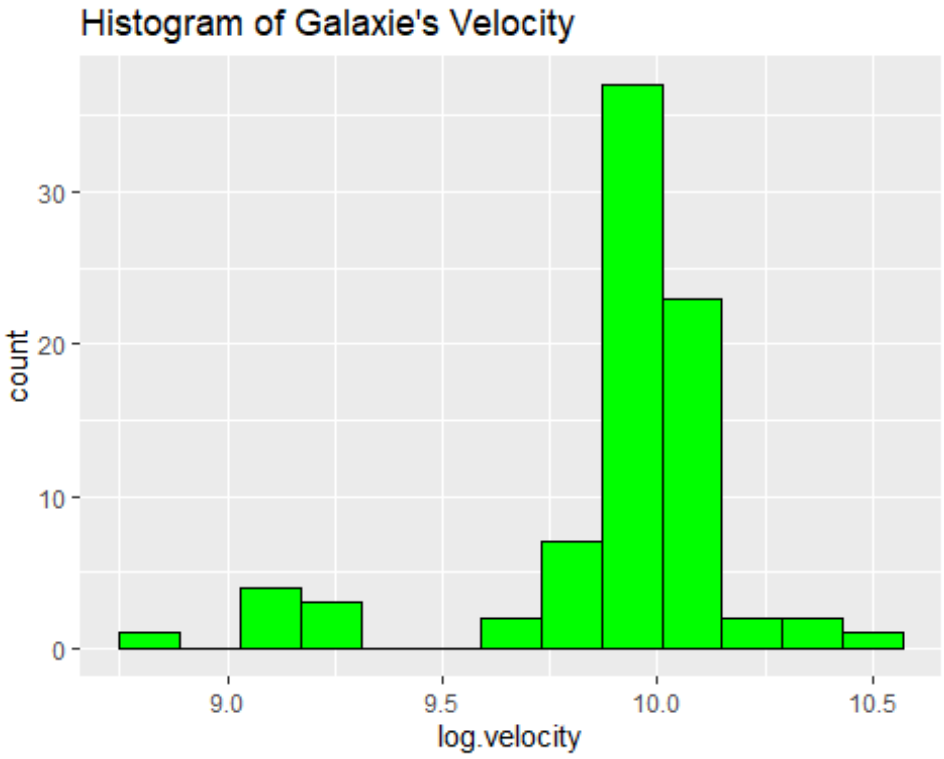
Answer: The qplot histogram shows a biomodal distribution. We can make visually inspect the data to estimate the distribution. The data is concentrated around the middle. My estimation is there is about three clusters. The largest cluster of data is around 20,000. The second cluster is around 10,000. The smallest cluster is around 35,000.

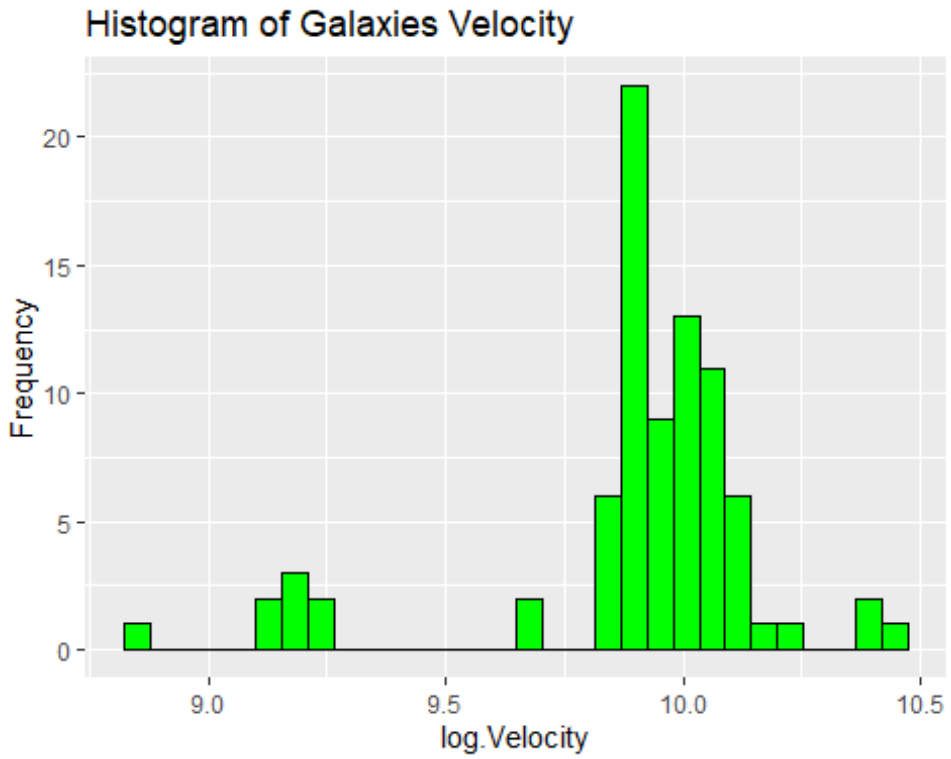
```
b) Create a new variable \textit{loggalaxies} = $\log$(galaxies). Repeat part a) using the `loggalaxies` variable. Does this affect your interpretation of the graphs?
```


Histogram of Galaxies Velocity
Frequency
0
10
20
30
40
9.0
9.5
10.0
10.5
log.velocity


Histogram of Galaxies Velocity
Frequency
0
10
20
30
9.0
9.5
10.0
10.5
log.Velocity


Histogram of Galaxies Velocity using truehist
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
9.0
9.5
10.0
10.5
log.Velocity

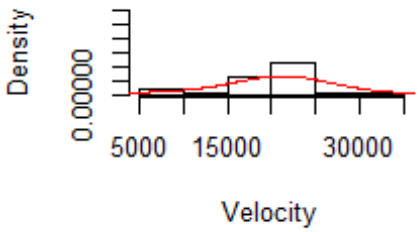

Histogram of Galaxie's Velocity
count
0
10
20
30
9.0
9.5
10.0
10.5
log.velocity



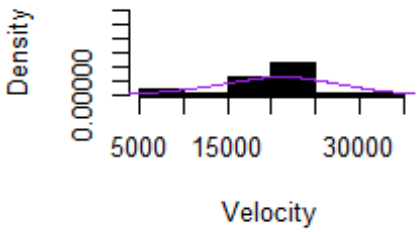
Answer: The main difference between the part a graphs and log scaled graphs is that the log scaled graph shows left skewed distribution. However the graphs still display multiple cluster of data. The log scaled shows mainly two clusters, but the qplot still show three main clusters.Part b findings align with my comments from part a.

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.

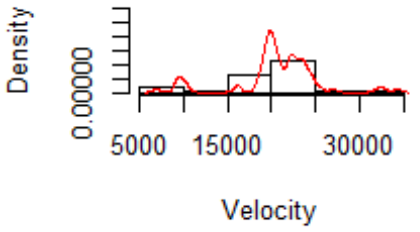
Over Smoothed Gaussian Dens



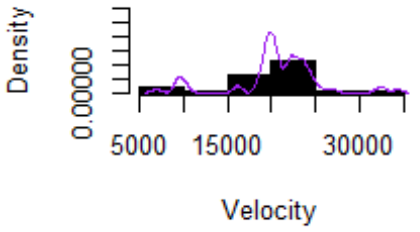
Over Smoothed triangular Den

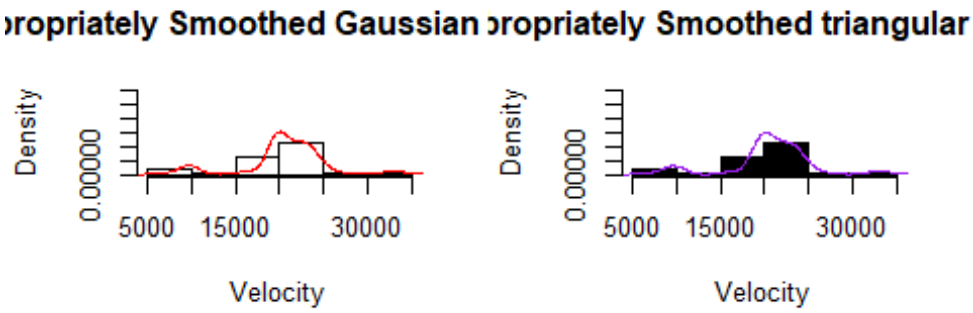


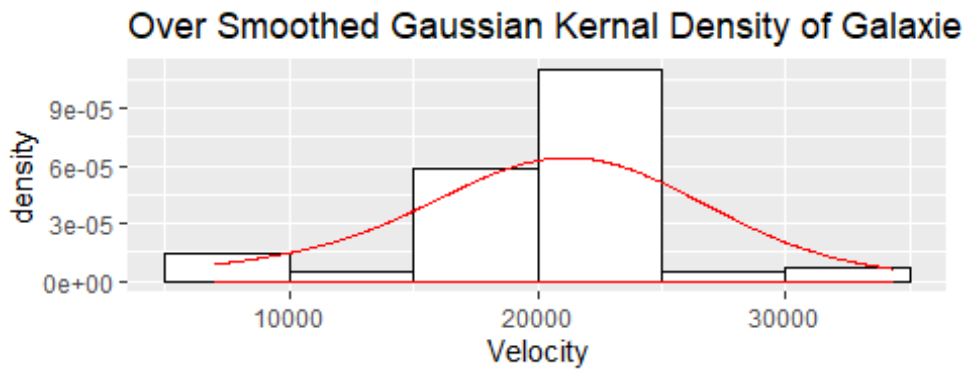
Under Smoothed Gaussian Der

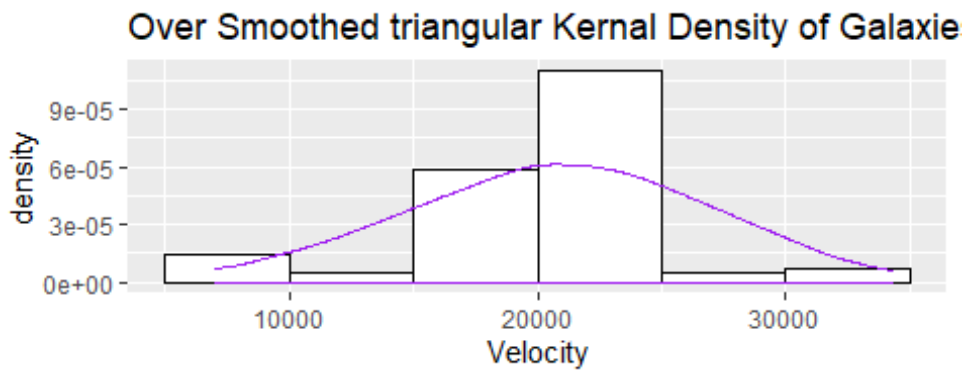


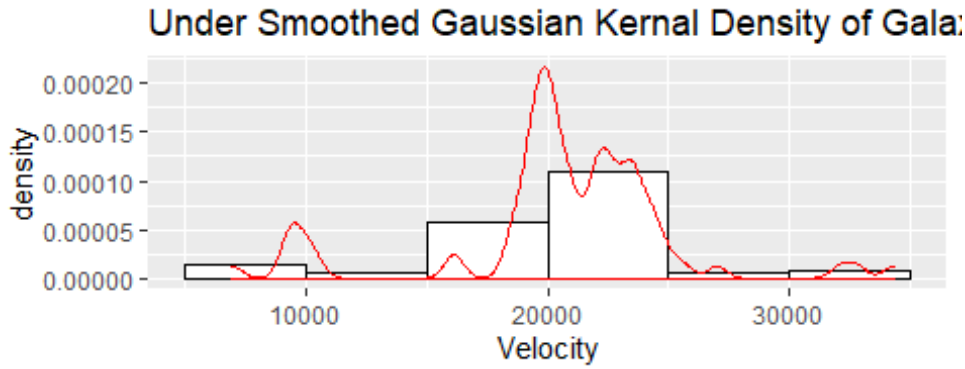
Under Smoothed triangular Der

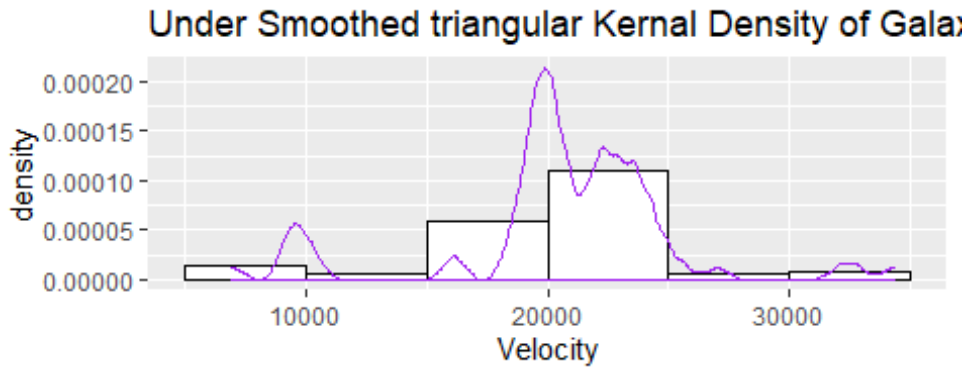


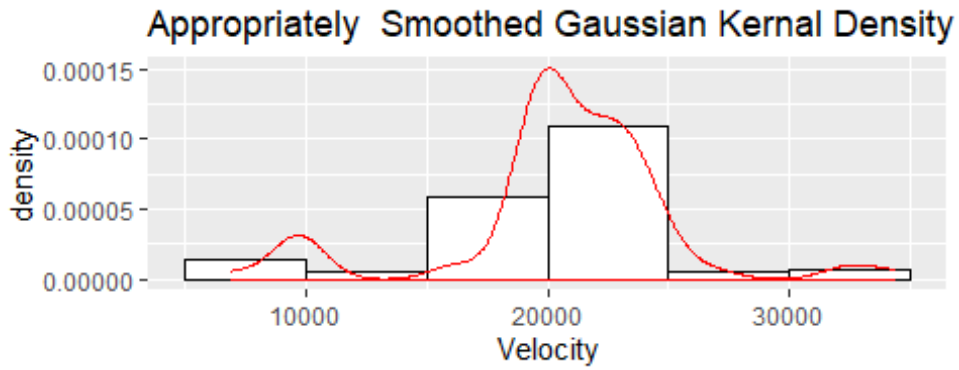

ropriately Smoothed Gaussian
ropriately Smoothed triangular
Density
0.00000
5000
15000
30000
Velocity
Density
0.00000
5000
15000
30000
Velocity

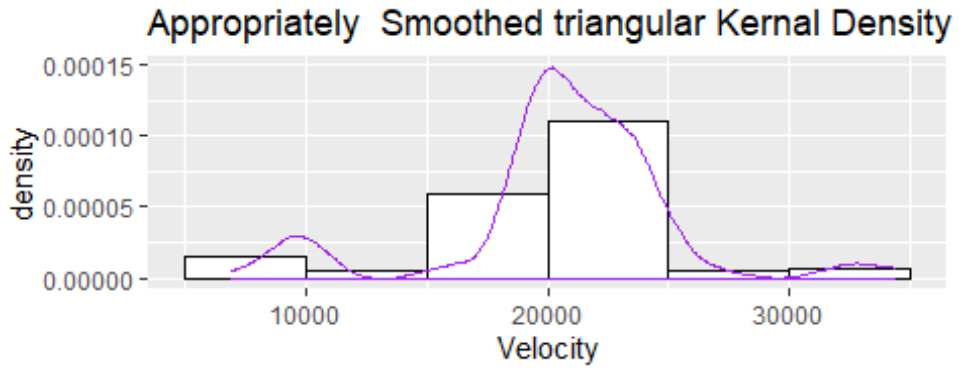

Over Smoothed Gaussian Kernal Density of Galaxie
density
9e-05
6e-05
3e-05
0e+00
10000
20000
30000
Velocity


Over Smoothed triangular Kernal Density of Galaxie
density
9e-05
6e-05
3e-05
0e+00
10000
20000
30000
Velocity


Under Smoothed Gaussian Kernal Density of Gala
density
0.00020
0.00015
0.00010
0.00005
0.00000
10000
20000
30000
Velocity


Under Smoothed triangular Kernal Density of Gala
density
0.00020
0.00015
0.00010
0.00005
0.00000
10000
20000
30000
Velocity





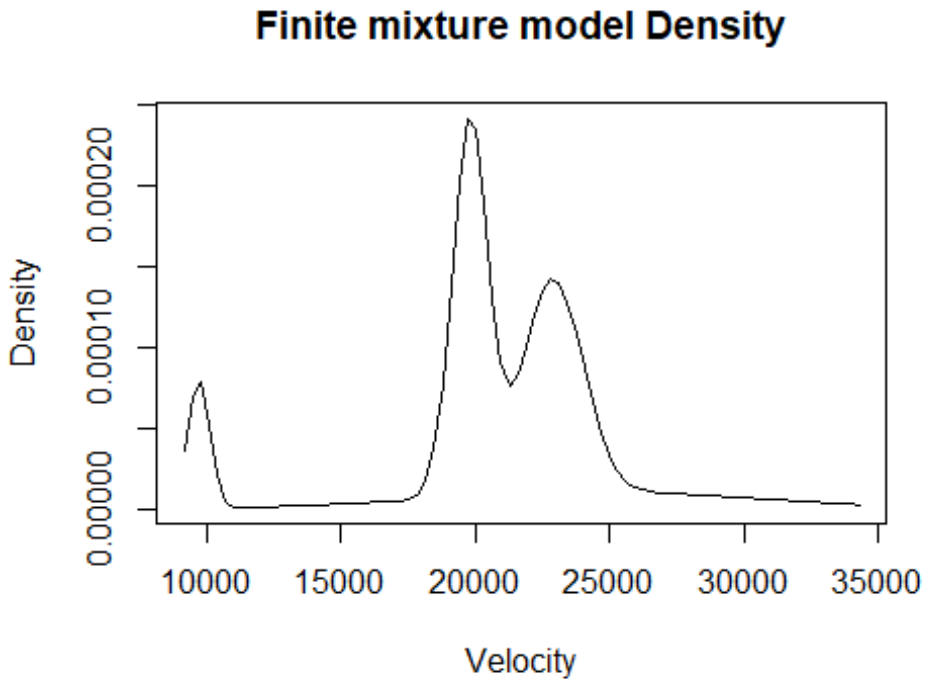
Answer: Choosing an appropriate bin width was a challenging task. It seems that under smoothed density shows about four clusters. two large clusters in the center and two small clusters in the right and the left tail. The over smoothed density shows one smooth curve in the middle. This density plot indicates that there is one large cluster of data. The appropriately smoothed density curve indicates that there three clusters of data or possibly four clusters.

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

Answer: by analyzing the histograms and the density plots, we can conclude that there is about three to four super 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.

## 'Mclust' model object: (V,4)
##
## Available components:
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "loglik"
##  [9] "df"             "bic"            "icl"            "hypvol"
## [13] "parameters"     "z"              "classification" "uncertainty"
```

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




```
## Bayesian Information Criterion (BIC):
##           E         V
```

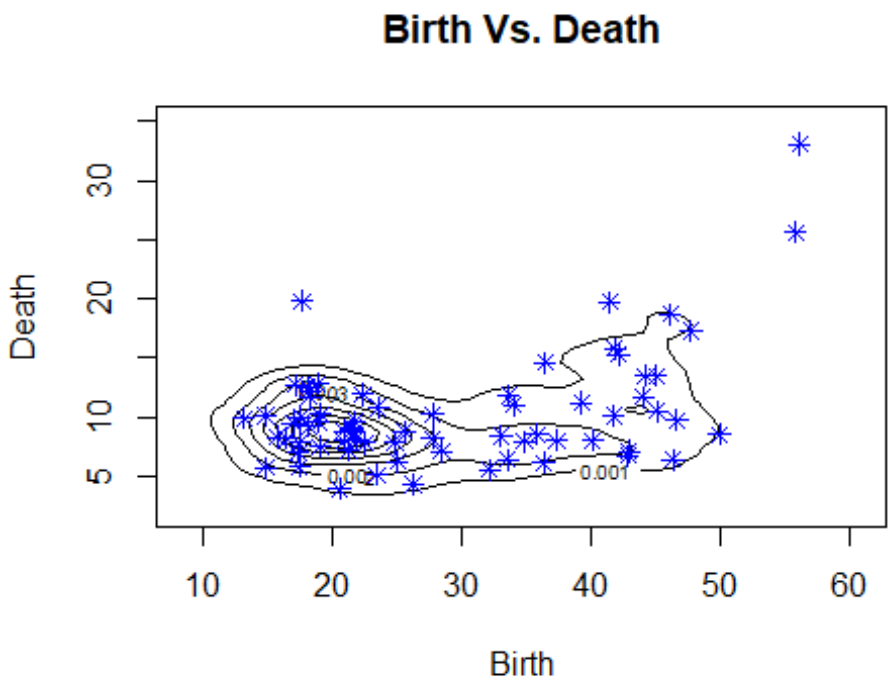
```
## 1 -1622.361 -1622.361
## 2 -1631.243 -1595.403
## 3 -1584.016 -1592.299
## 4 -1592.828 -1579.862
## 5 -1592.299 -1593.277
## 6 -1601.228 -1604.069
## 7 -1588.610 -1611.538
## 8 -1597.427 -1625.804
## 9 -1600.709 -1633.494
##
## Top 3 models based on the BIC criterion:
##       V,4       E,3       E,7
## -1579.862 -1584.016 -1588.610
```

Answer: The model and the density plot found that there is four clusters. We estimated three to four models for parts a,b, and c which is close to what the model found. The BIC for the best model is -1579.862 (V,4). Parameter estimates are reported below

```
## probabilities

## [1] 0.08440635 0.38660329 0.37116156 0.15782880

## Means

##         1         2         3         4
##  9707.492 19804.259 22879.486 24459.536

## Variances

## $modelName
## [1] "V"
##
## $d
## [1] 1
##
## $G
## [1] 4
##
## $sigmasq
## [1]   177296.7   436160.9  1261611.3 34437115.3
##
## $scale
## [1]   177296.7   436160.9  1261611.3 34437115.3
```

. (Ex. 8. in HSAUR, modified for clarity) The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (from Hartigan, 1975).

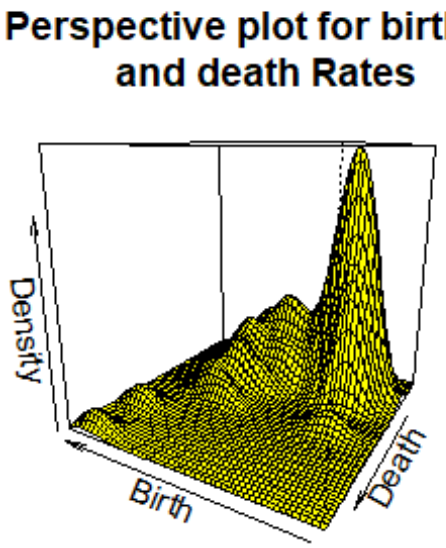
```
a) Produce a scatterplot of the data. Estimate the bivariate density and
overlay the corresponding contour plot on the scatterplot.
```



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

Answer: The data is concentrated around a birth rate of 20 and death rate of 10 for most countries. The biggest cluster is aound those rates. We have more clusters but the data is not concentrated as much in them. We have two to three countries which have high birth and death rate. Most countries have a birth rate around 20 and death rate around 10. Majority of the countries have birth rate that is higher than death rates.

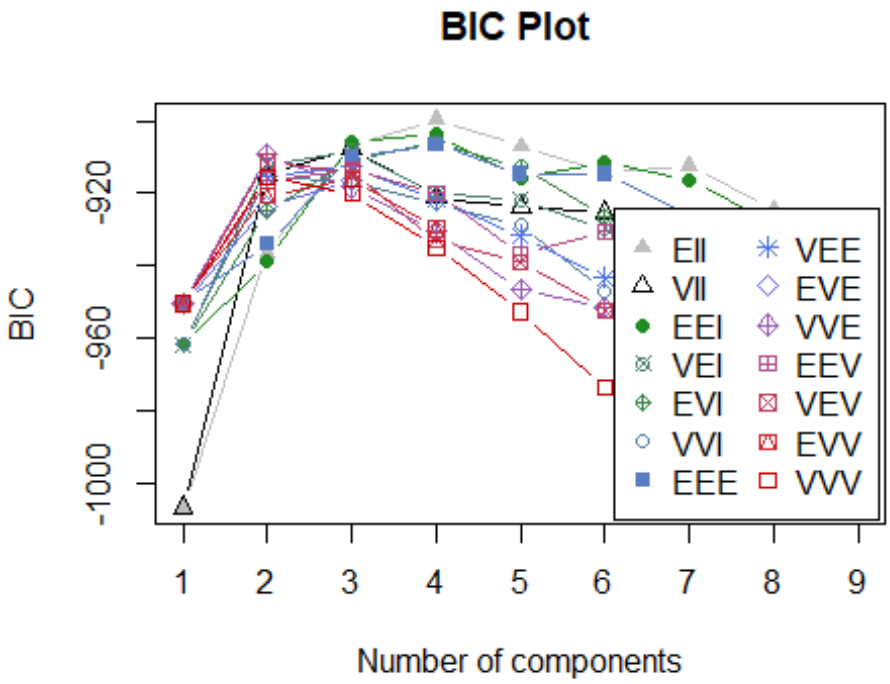
c) Produce a perspective plot (persp() in R, ggplot is not required for this question).

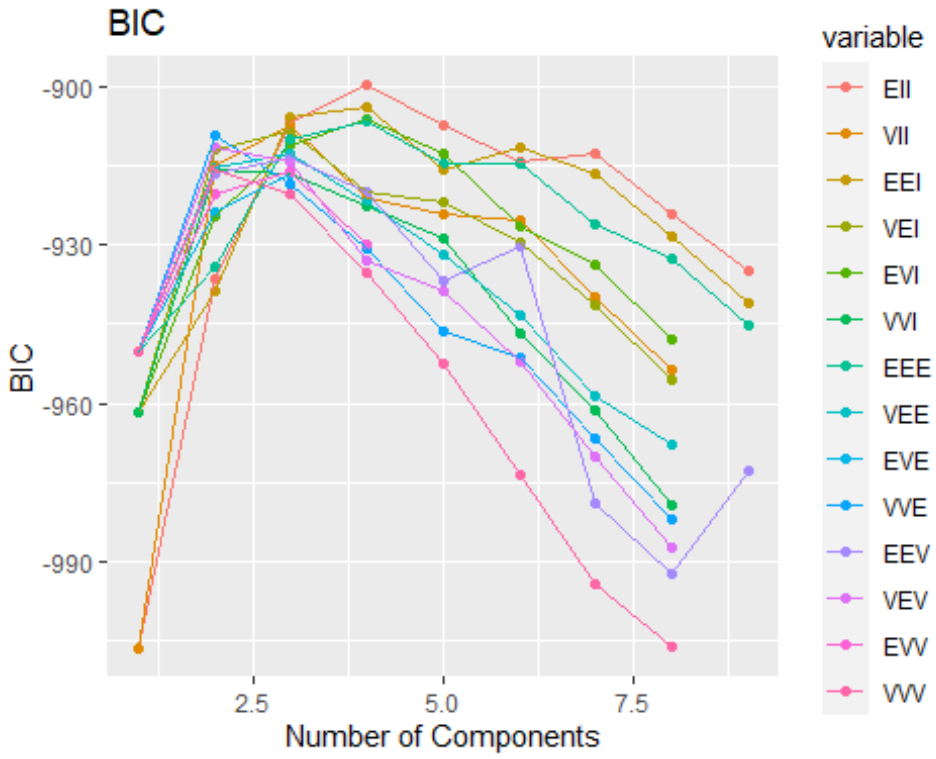


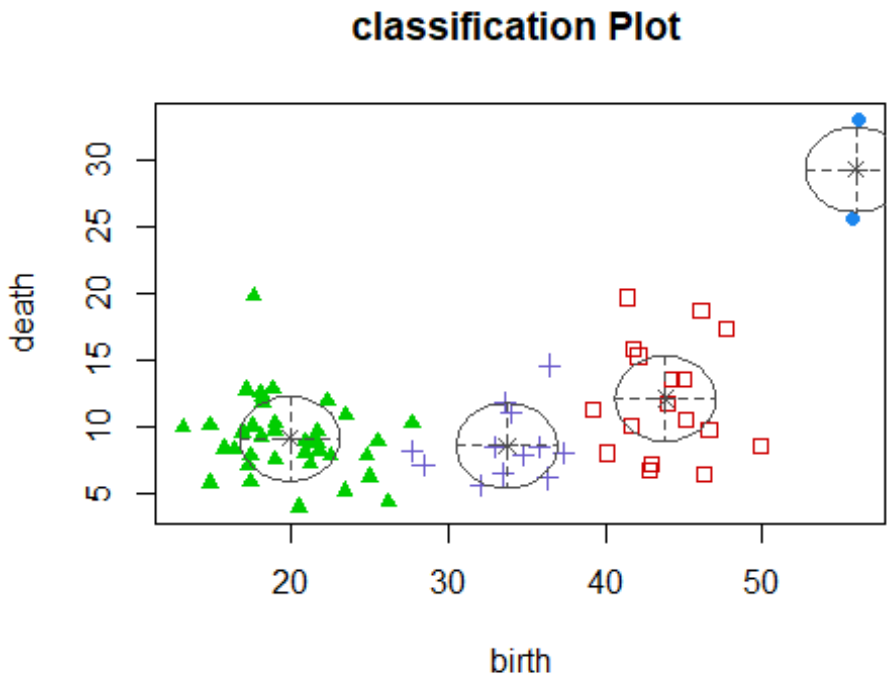
d) Fit a finite mixture model using the Mclust() function in R (from the mclust library). Summarize this model using BIC, classification, uncertainty, and/or density plots.

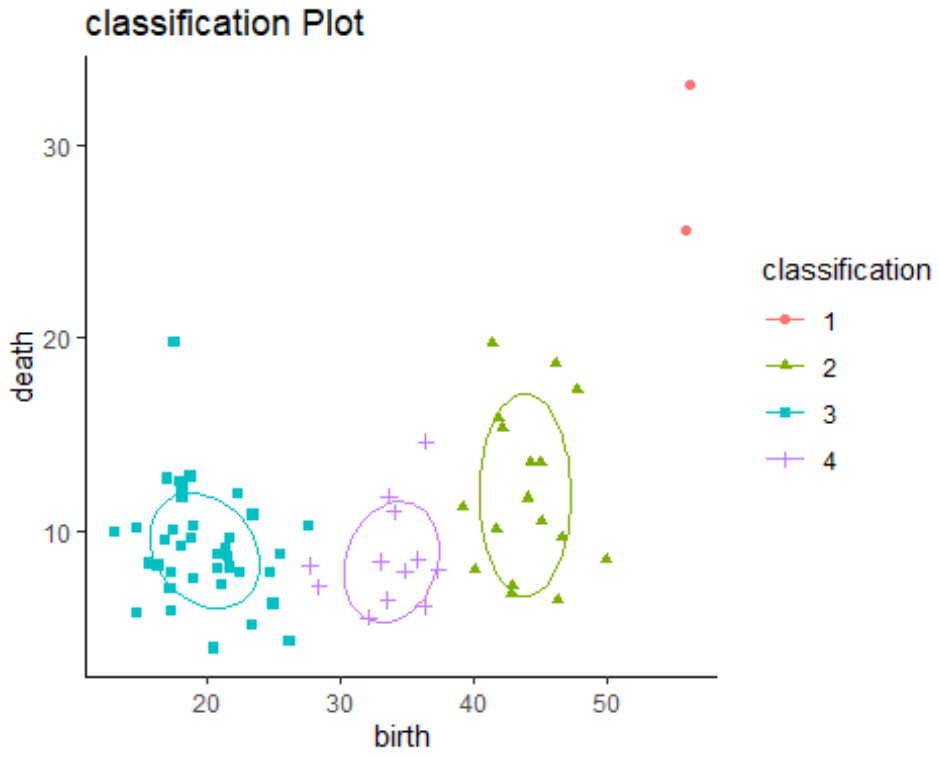
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194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

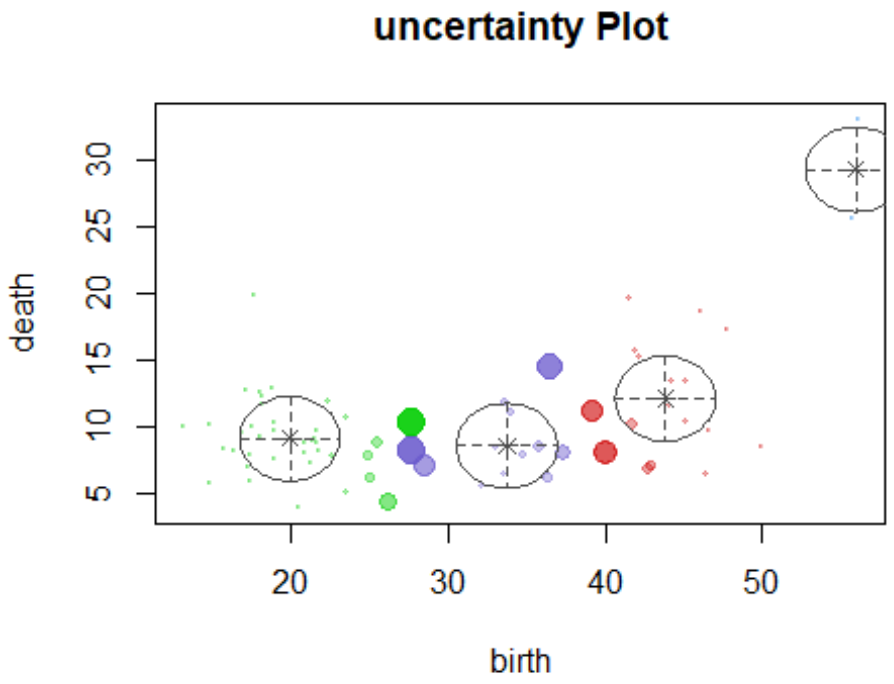
```
## [,,1]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,2]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,3]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
## [,,4]
##         birth   death
## birth 10.2108  0.0000
## death  0.0000 10.2108
```

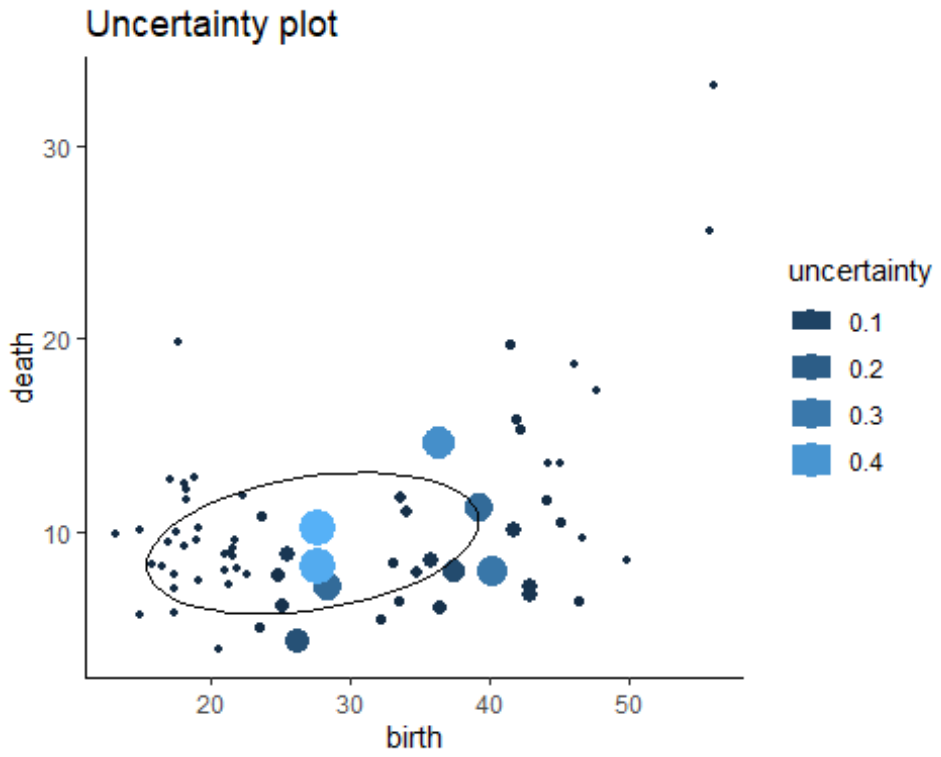

BIC Plot
BIC
-920
-960
-1000
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV

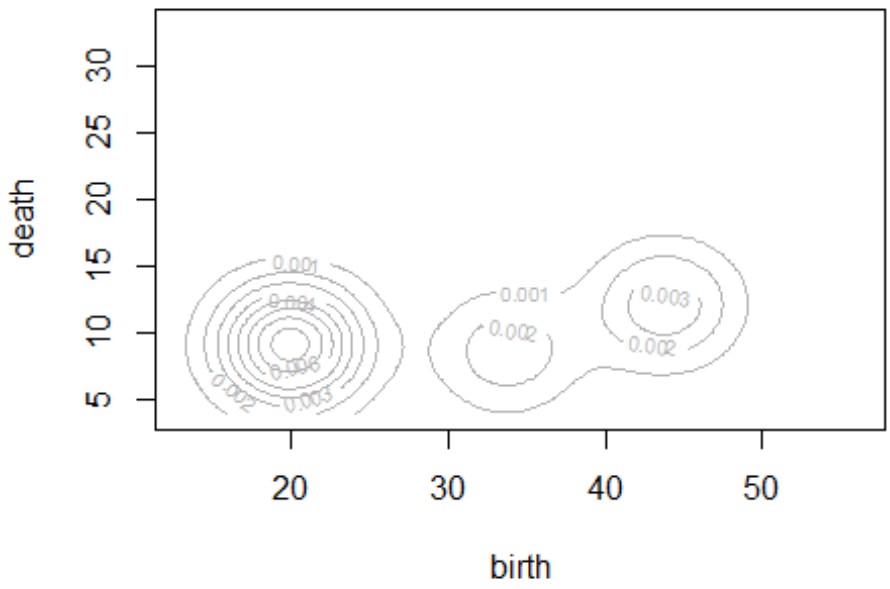

BIC
-900
-930
-960
-990
2.5
5.0
7.5
Number of Components
variable
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


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


classification Plot
death
birth
10
20
30
20
30
40
50
classification
1
2
3
4


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





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

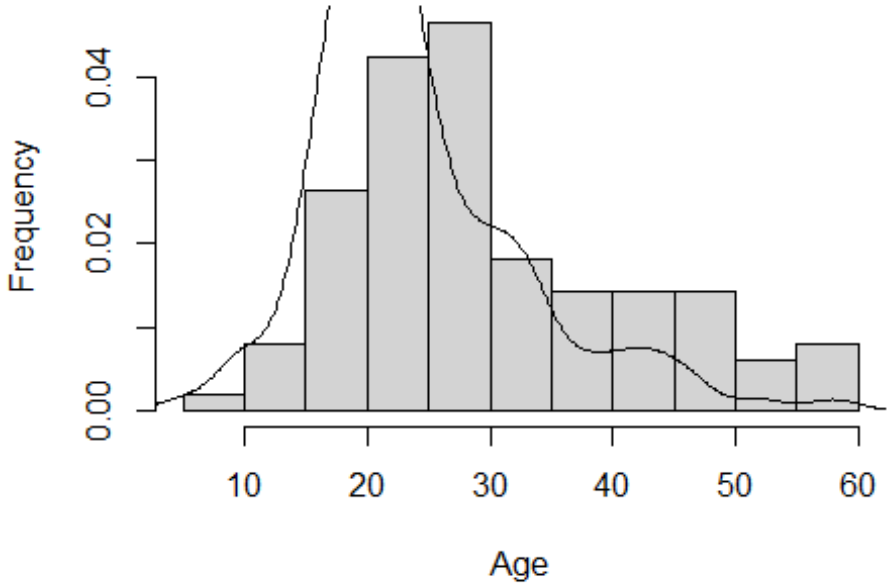
Answer:

The model found that there is four clusters. We can conclude that, generally, the birth rate is twice the death rate in most countries. For one group of countries as birth rate increases from 2 to 27, death rate stays decreases. For countries where their birth is increasing from 28 to 38, death stays at a range of 6 to 15. For countries that have high birth rate that ranges from 39 to 50 , death rate increases slightly to reach 20. There is a group of two countries that have very high birth rates and death rates. Generally, majority of countries are having more babies than people dying except for for countries where where more babies are being born and more people are dying.

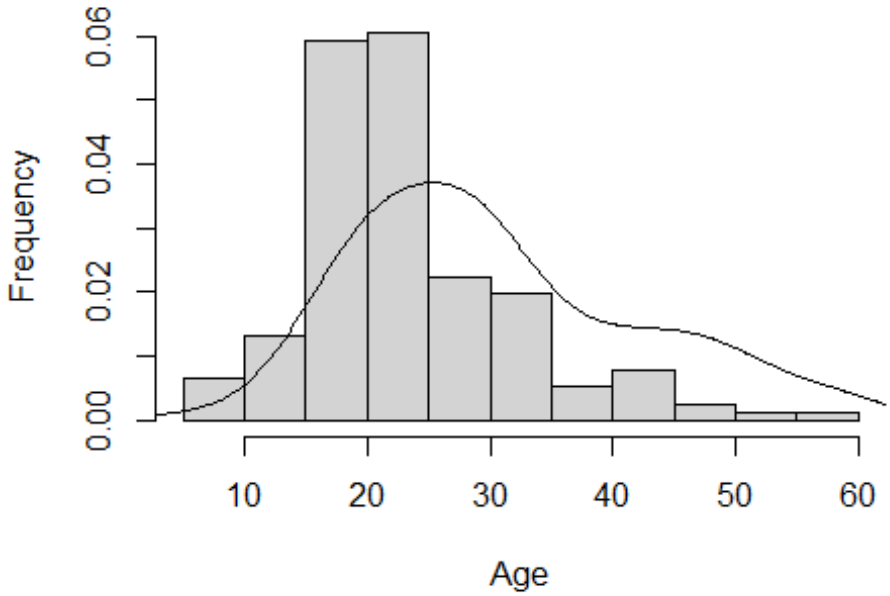
3. (Ex. 8.3 in HSAUR, modified for clarity) Fit finite mixtures of normal densities individually for each gender in the **schizophrenia** data set from **HSAUR3**. Do your models support the *sub-type model* described in the R Documentation?

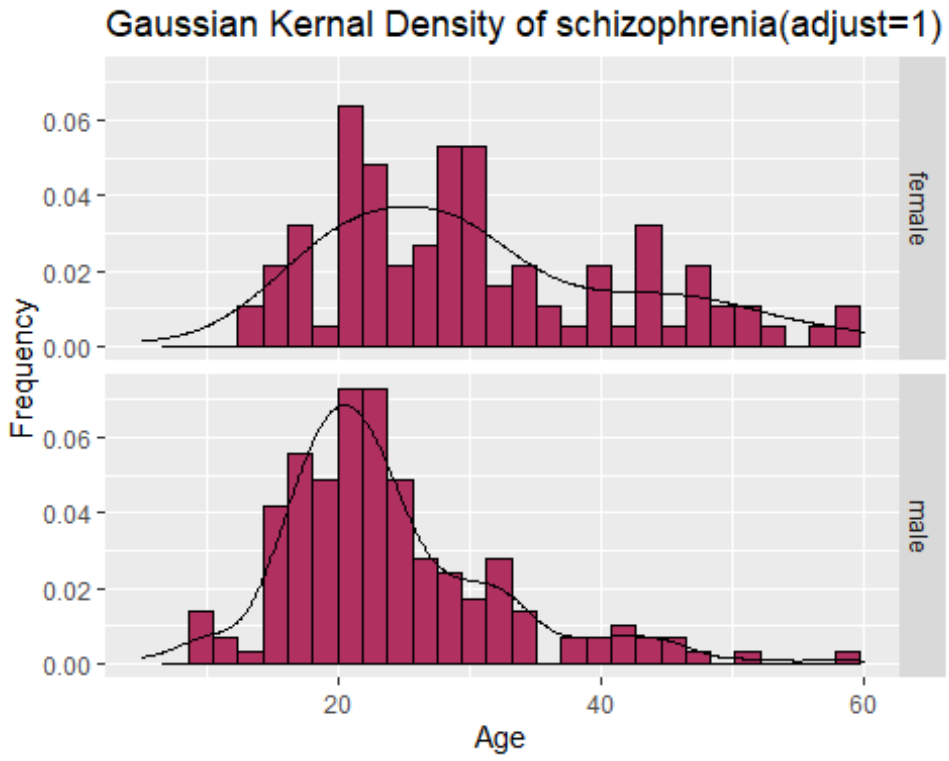
   Quote from the R Documentation: *A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequent epidemiological studies of the disorder have consistently shown an earlier onset in men than in women. One model that has been suggested to explain this observed difference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

Gaussian Kernal for Males with schizophrenia(Adjus



Gaussian Kernal for Males with schizophrenia(Adjus





Answer:

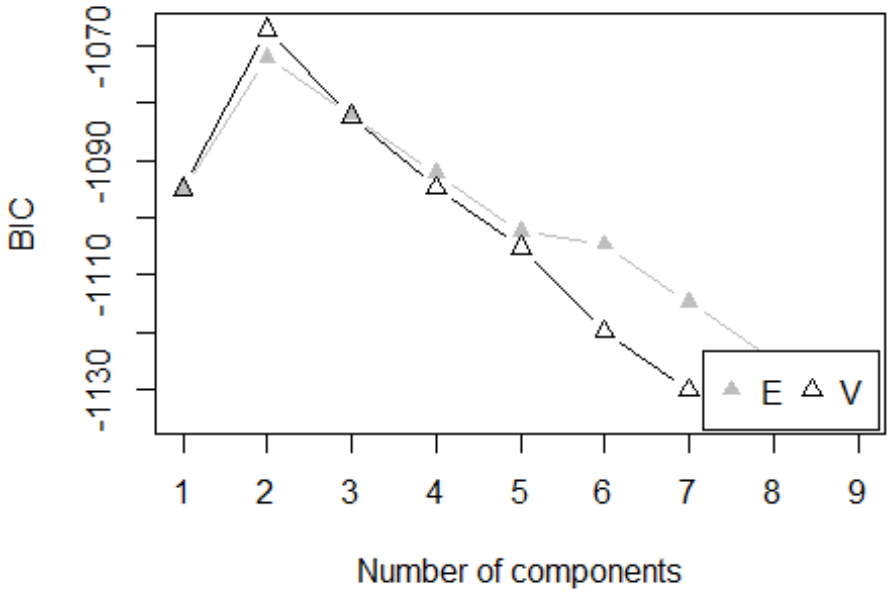
The histogram for males shows that most males have schizophrenia around the age of 20 and 40 where females have schizophrenia through years of life. For males, we estimate that there is about three clusters of data. For females, we estimate tat there about two clusters.

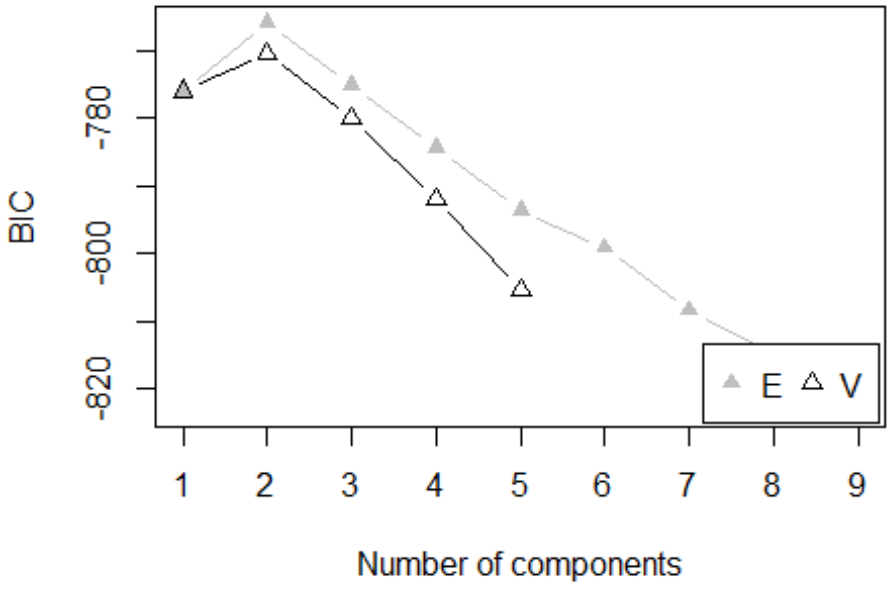
```
## Model Summary For Males

## ----------------------------------------------------
## 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

## Model Summary For Females

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

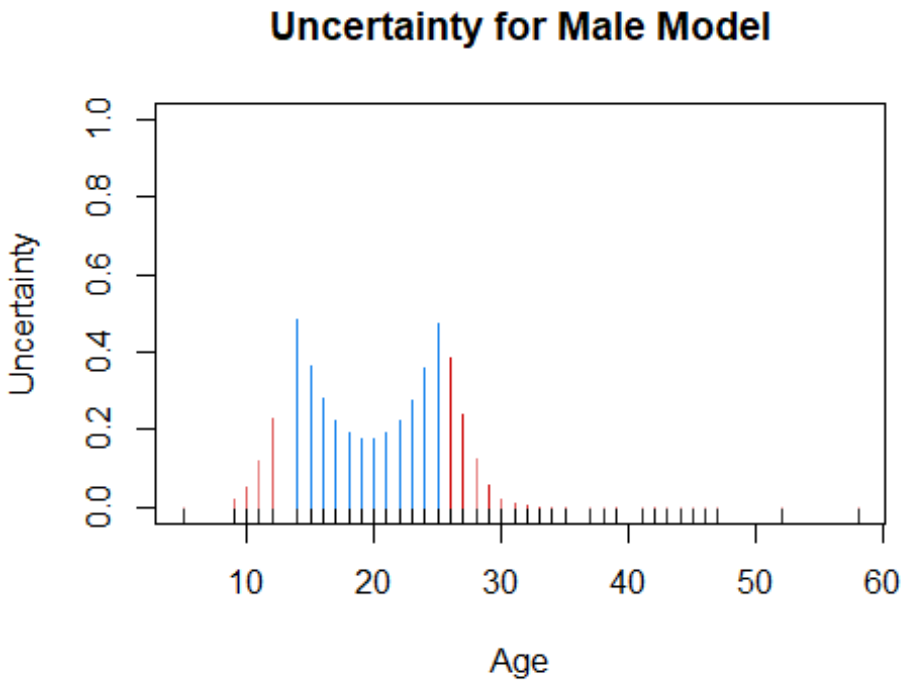
Discussion:

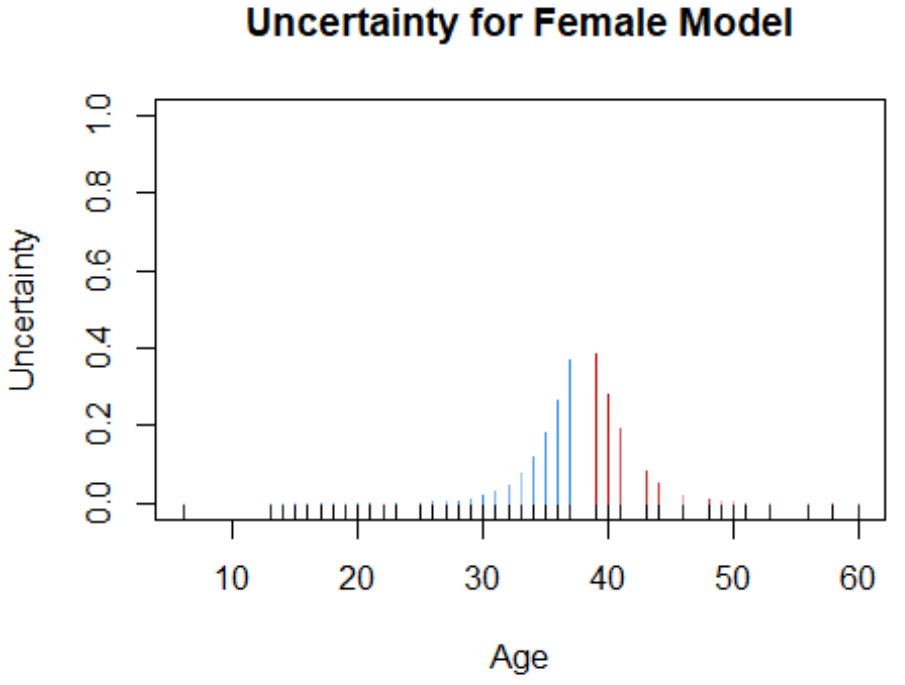
For females, the model found that there is two clusters of data around the age of 25 and 47 which indicates that women are prone to early and late start of schizophrenia. The model found that there two cluster for males with an average of 20 and 28 for each cluster respectively. Females have a larger age mean which indicates that males get the disorder earlier than females.

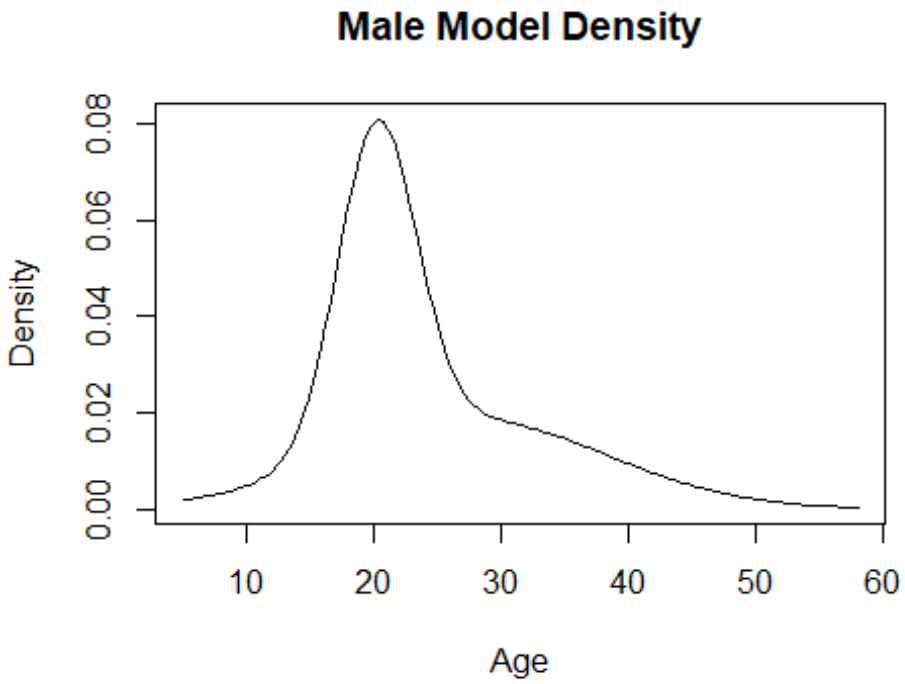


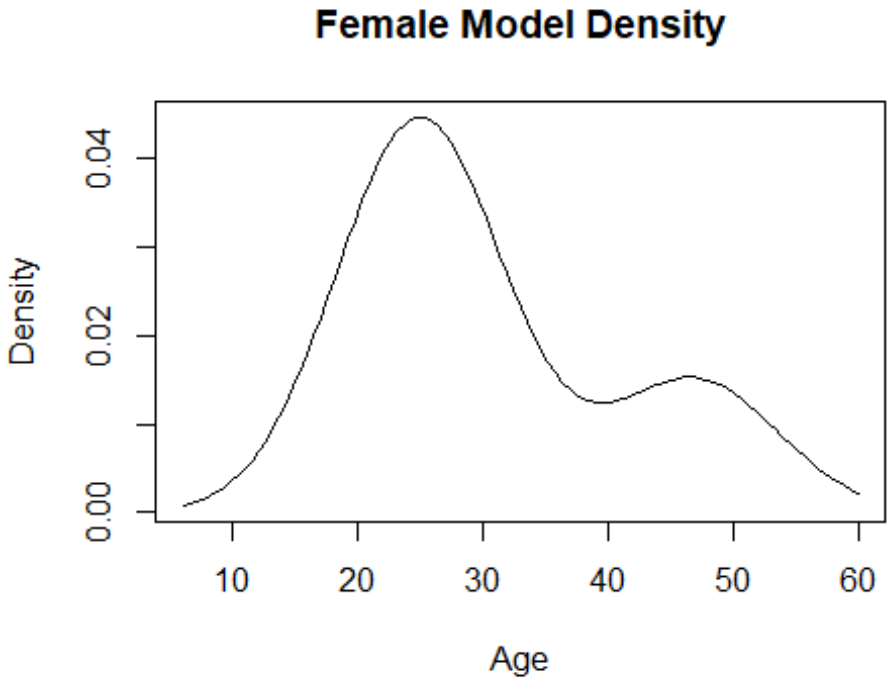


BIC plots The BIC plots show that both males and females have two clusters.


Uncertainty for Male Model
Uncertainty
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
Age


Uncertainty for Female Model
Uncertainty
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
Age





Density plots: The Density plots show that females have early and late of schizophrenia with two clusters around 28 and 48 where most males have an early start of schizophrenia.

Citations:

Kuipers, K. (2018, September 18). Homework 4 - STAT 601. Retrieved October 13, 2020, from https://rpubs.com/kkuipers/529695

Hothorn, T., & Everitt, B. S. (2017). Chapter 8. In A handbook of statistical analyses using R. Boca Raton: CRC Press.