Homework 4

Amin Baabol

Note: I am getting the hang of remembering to cite my sources, I get engrossed in the coding and the interpretations that forget in the end to include my sources. Collaboration: I did not collaborate with anyone on this assignment.

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.

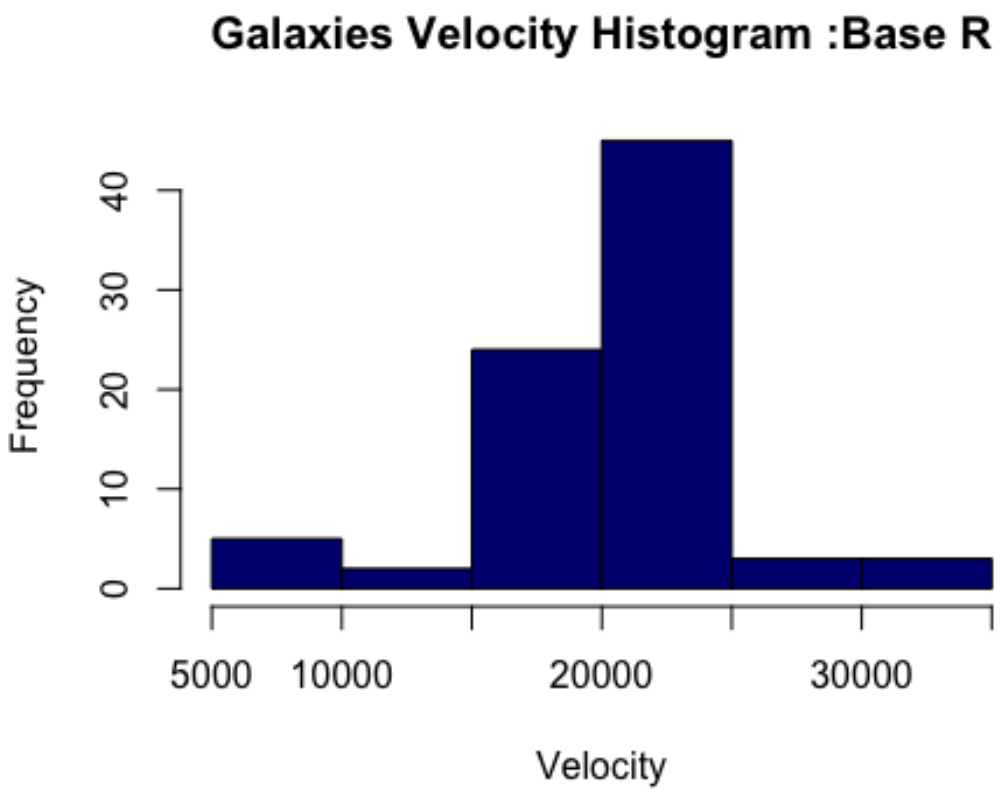
1. (Ex. 8.1 in HSAUR, modified for clarity) The data from contains the velocities of 82 galaxies from six well-separated conic sections of space (Postman et al., 1986, Roeder, 1990). The data are intended to shed light on whether or not the observable universe contains superclusters of galaxies surrounded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

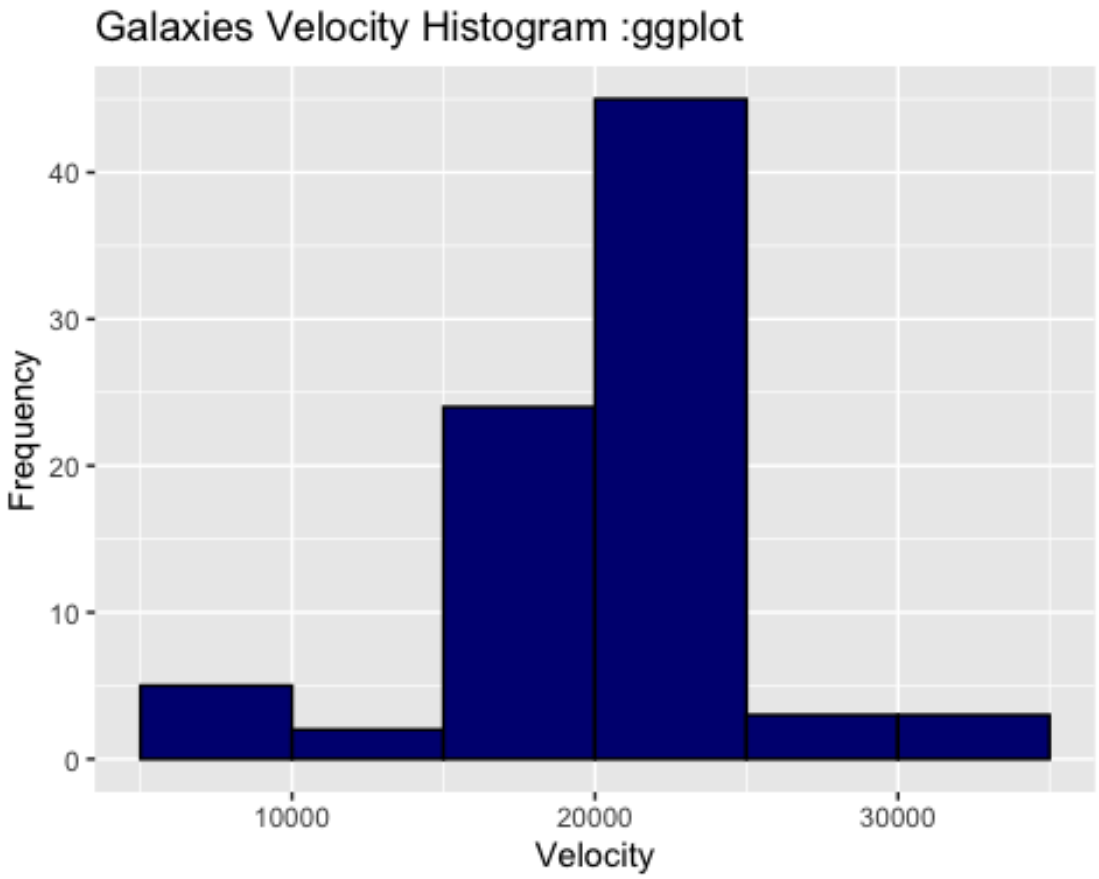
a) Construct histograms using the following functions:

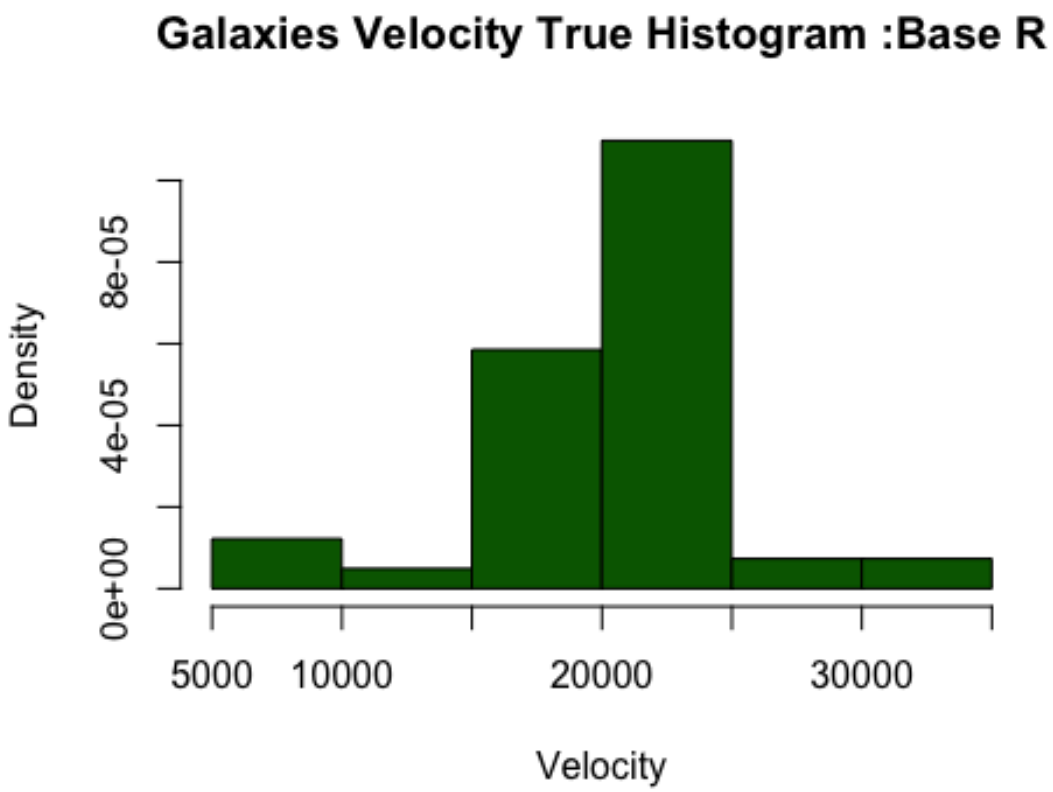
```
 -hist() and ggplot()+geom_histogram()

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

 -qplot()
```

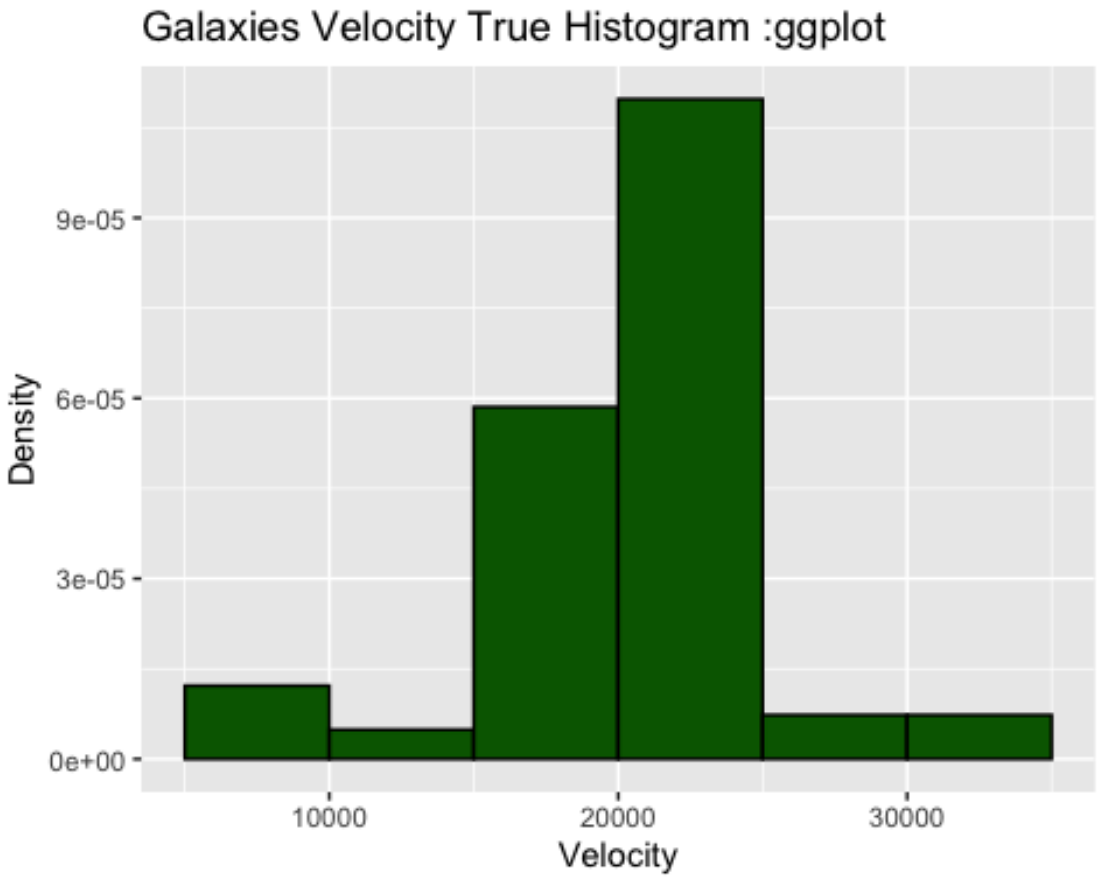
Comment on the shape and properties of the variable based on the five plots. Do you notice any sets of observations clustering? (Hint: You can adjust bin number or bin size as you try to determine the properties of the variable, but use the same bin settings between plots in your final analysis. You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

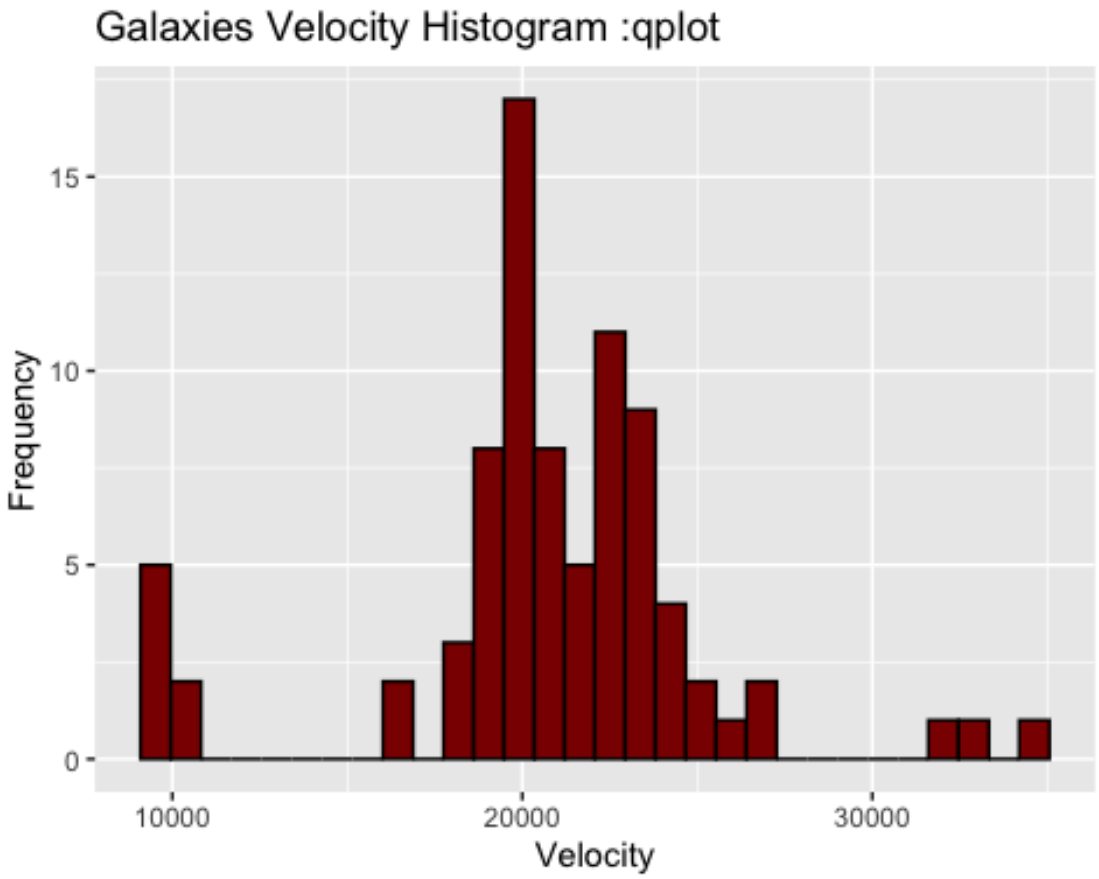
```
## [1] 26690
```


Galaxies Velocity Histogram :Base R
Frequency
0
10
20
30
40
5000
10000
20000
30000
Velocity


Galaxies Velocity Histogram :ggplot
Frequency
40
30
20
10
0
10000
20000
30000
Velocity


Galaxies Velocity True Histogram :Base R
Density
0e+00
4e-05
8e-05
5000
10000
20000
30000
Velocity

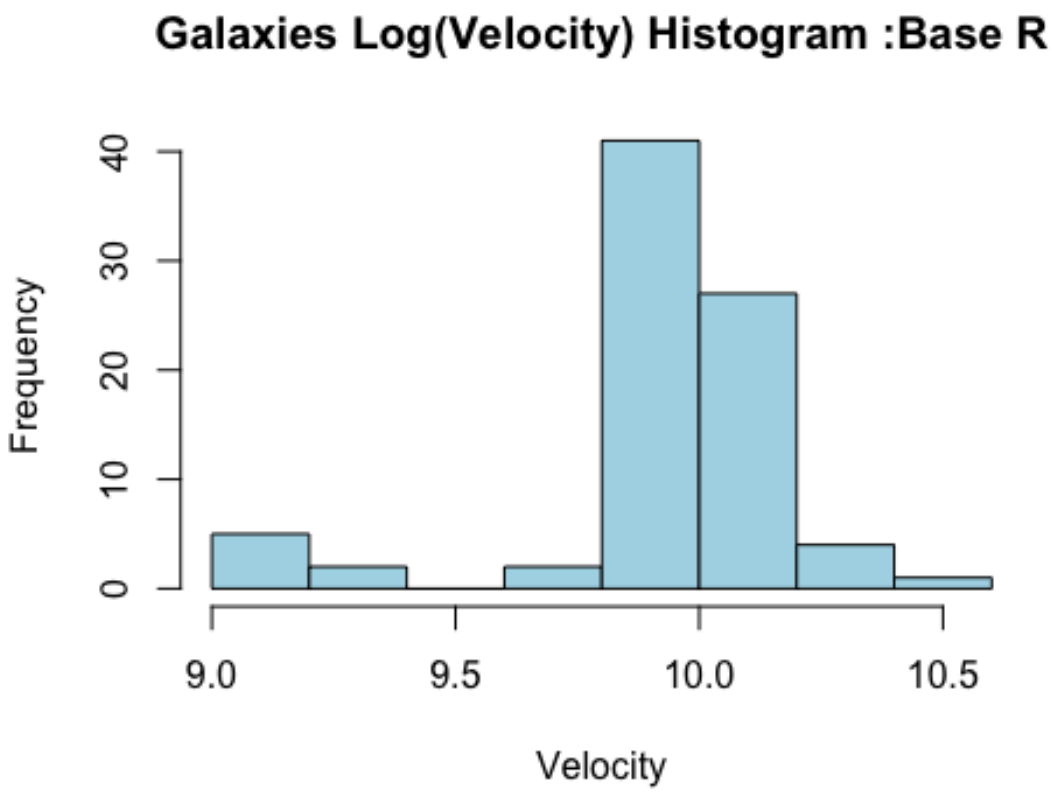

Galaxies Velocity True Histogram :ggplot
Density
9e-05
6e-05
3e-05
0e+00
10000
20000
30000
Velocity

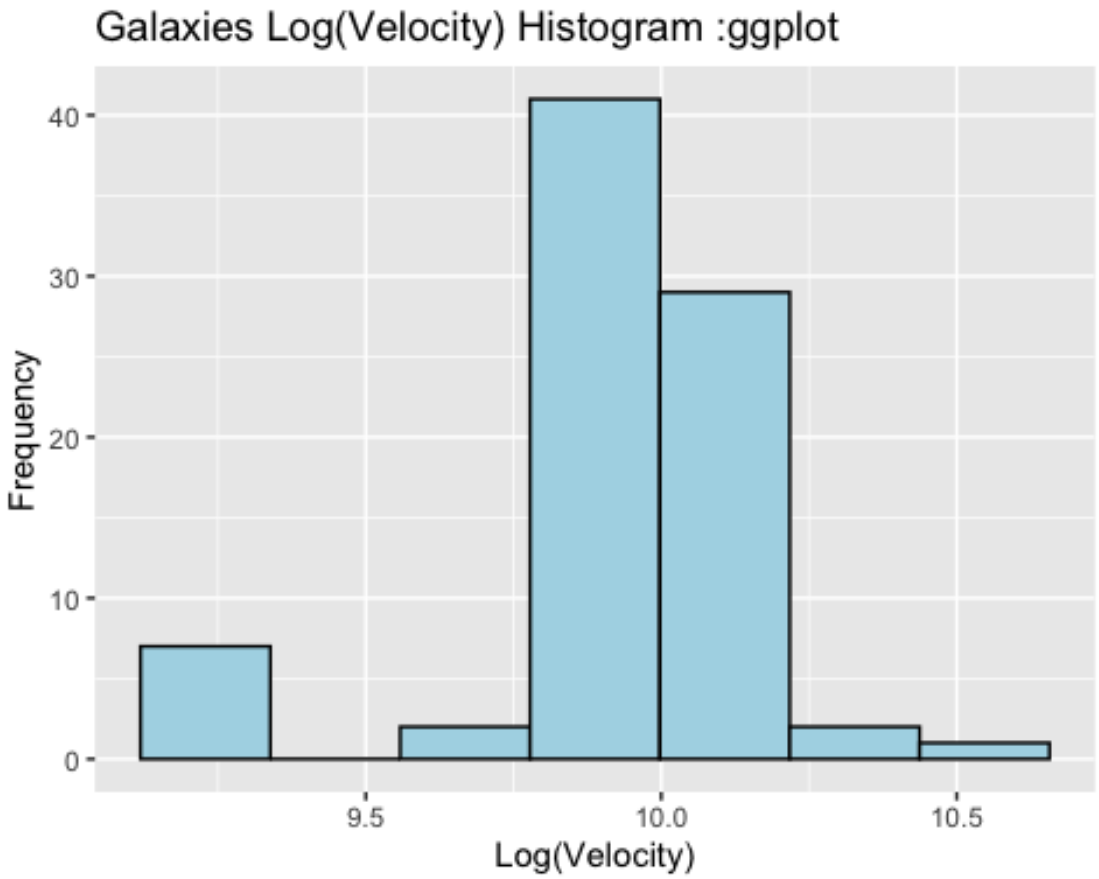


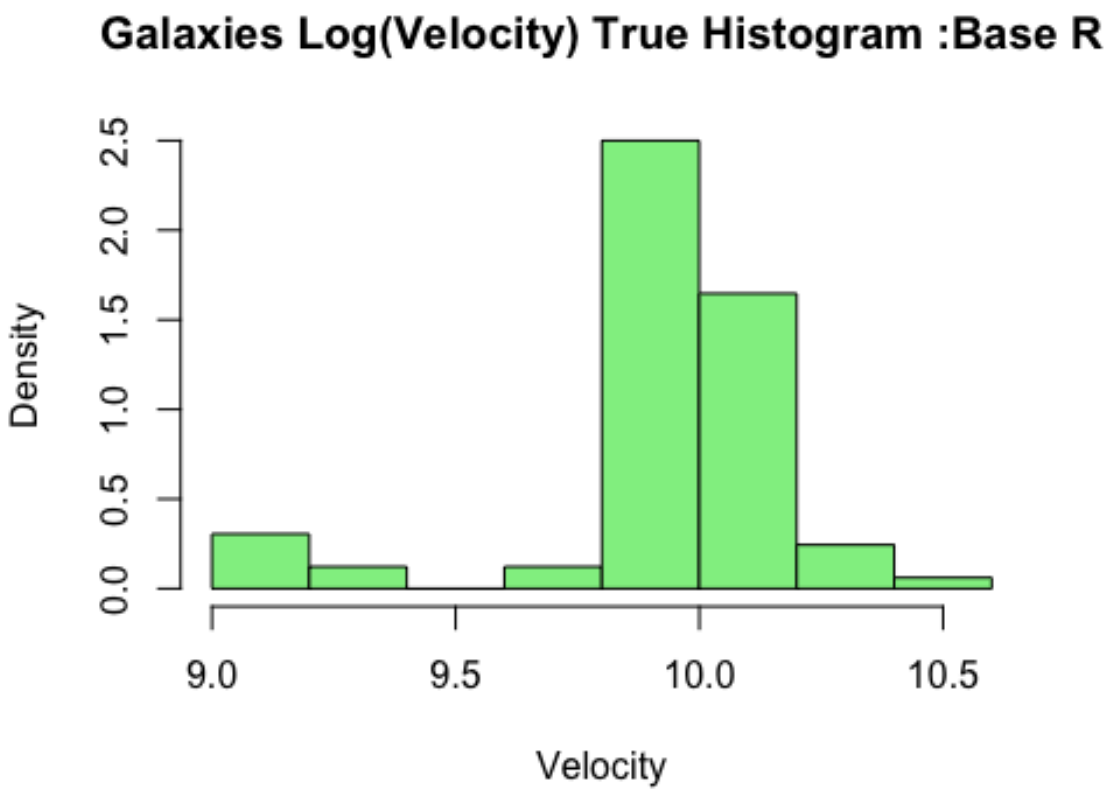
Discussion

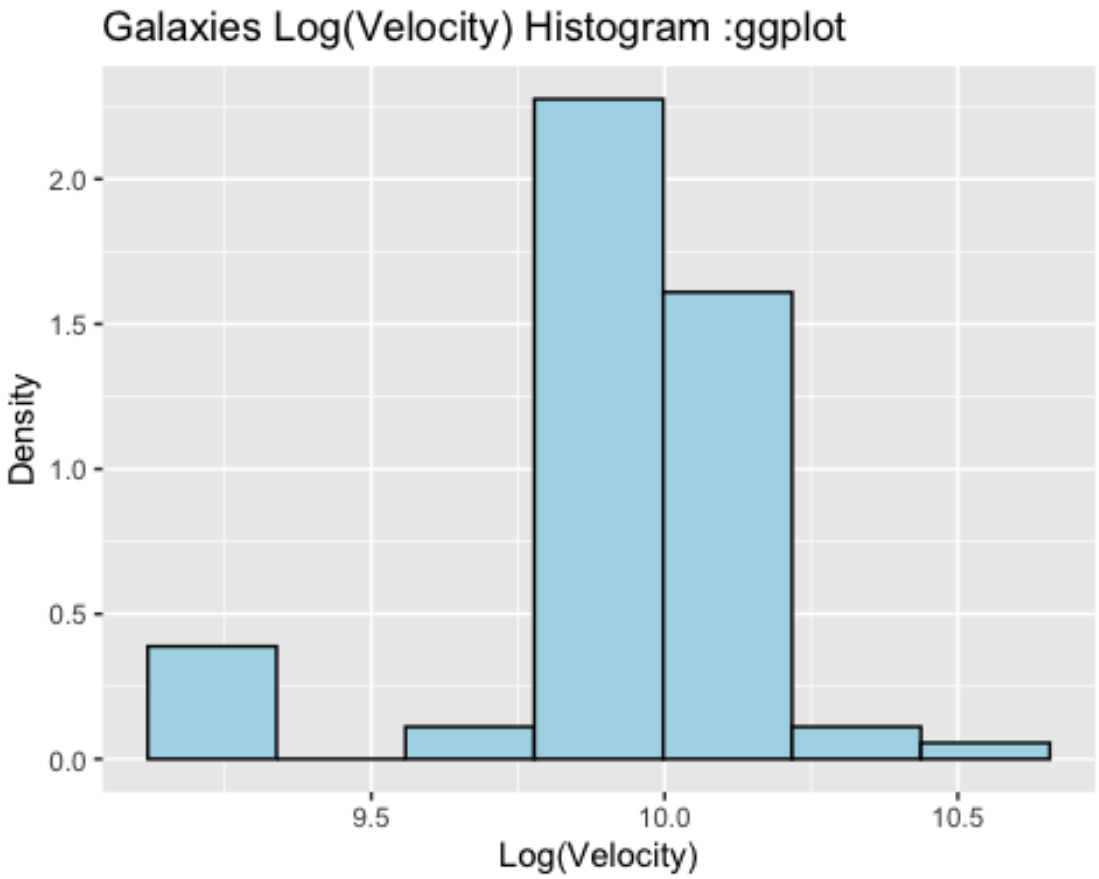
part 1a: There's a typo in the original galaxies dataset in the 78th observation 26690 which should be 26960 according to R, we corrected it. Histograms in figure1a.1 indicate that there is a normal distribution clustered around around 20,000 with frequency count as the y-axis.True histogram in figure 1a.2 reveals the same normal distribution seen in the regular histogram in figure 1a.1. It's however, important to note that the true histrogram function uses density values which is probability as its y-axis.Qplot in figure1a.3 reveals that the distribution isn't exactly normal like we initially assumed, there is one main cluster in and around 20,00 and there are also three smaller clusters which suggests multimodal distribution with four clusters.

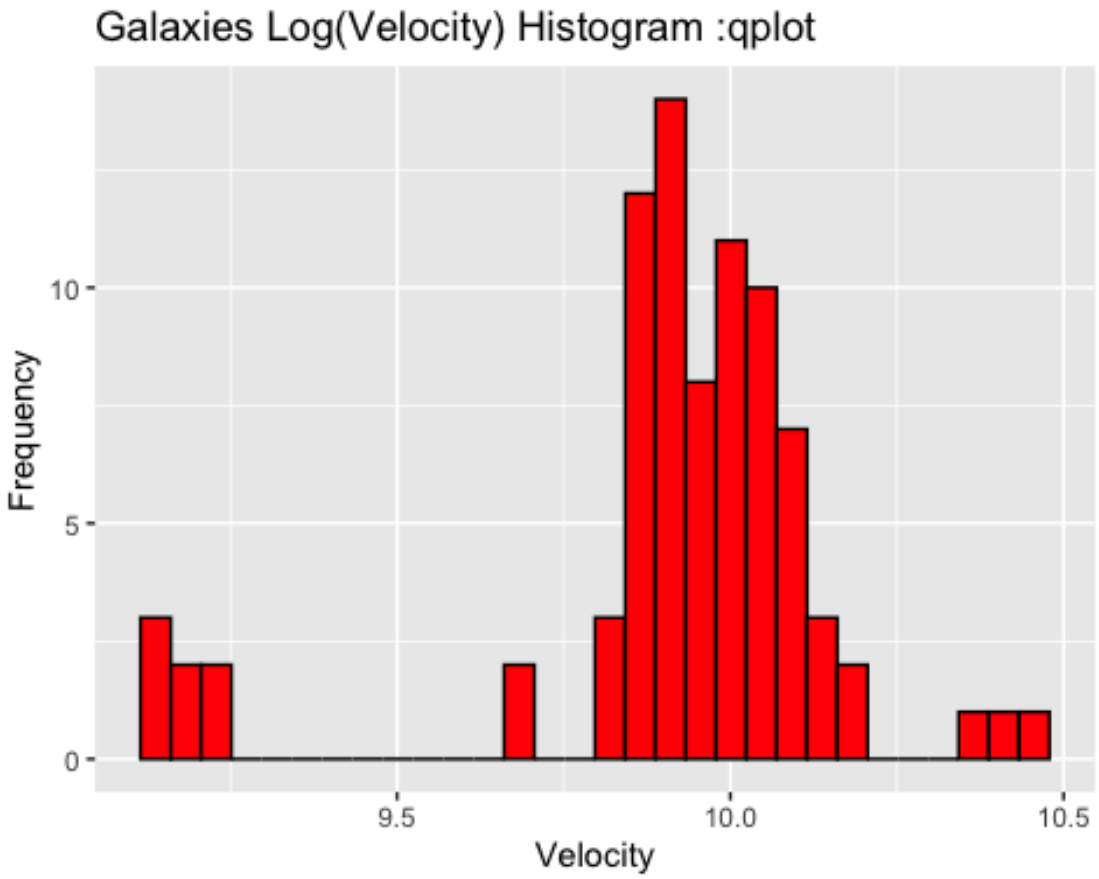
b) Create a new variable = log(galaxies). Repeat part a) using the `loggalaxies` variable. Does this affect your interpretation of the graphs?


Galaxies Log(Velocity) Histogram :Base R
Frequency
0
10
20
30
40
9.0
9.5
10.0
10.5
Velocity


Galaxies Log(Velocity) Histogram :ggplot
Frequency
40
30
20
10
0
9.5
10.0
10.5
Log(Velocity)


Galaxies Log(Velocity) True Histogram :Base R
Density
0.0
0.5
1.0
1.5
2.0
2.5
9.0
9.5
10.0
10.5
Velocity


Galaxies Log(Velocity) Histogram :ggplot
Density
2.0
1.5
1.0
0.5
0.0
9.5
10.0
10.5
Log(Velocity)

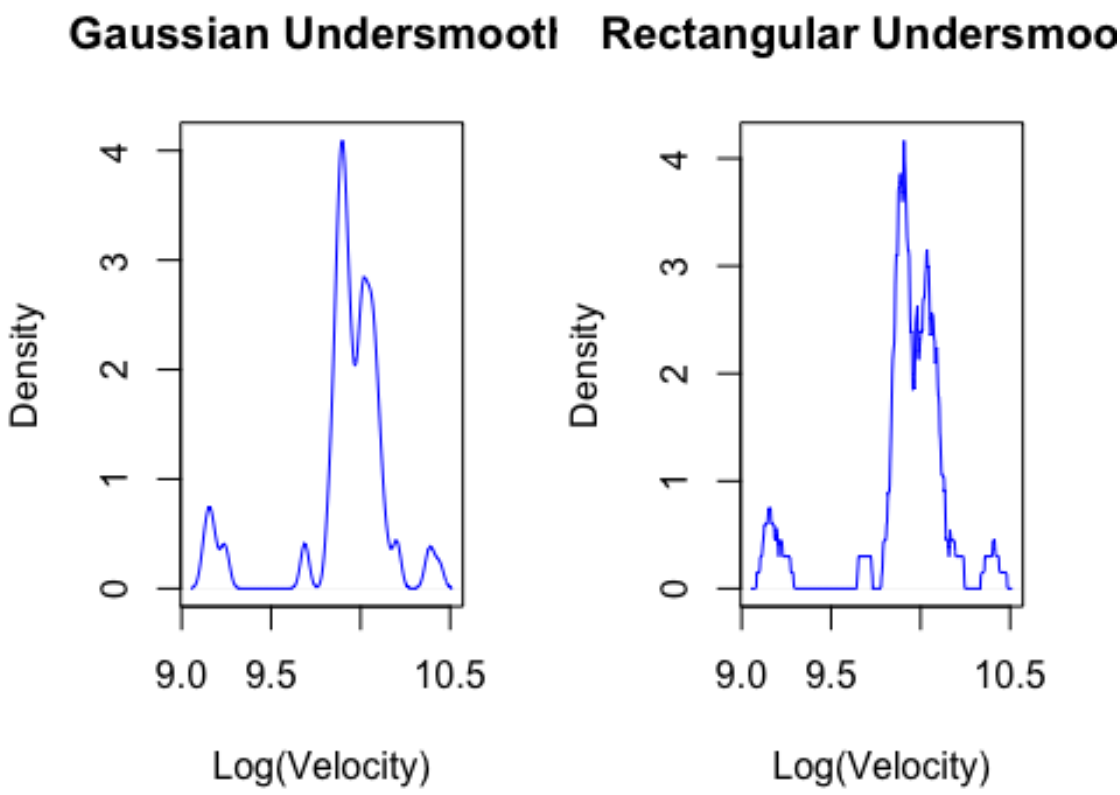


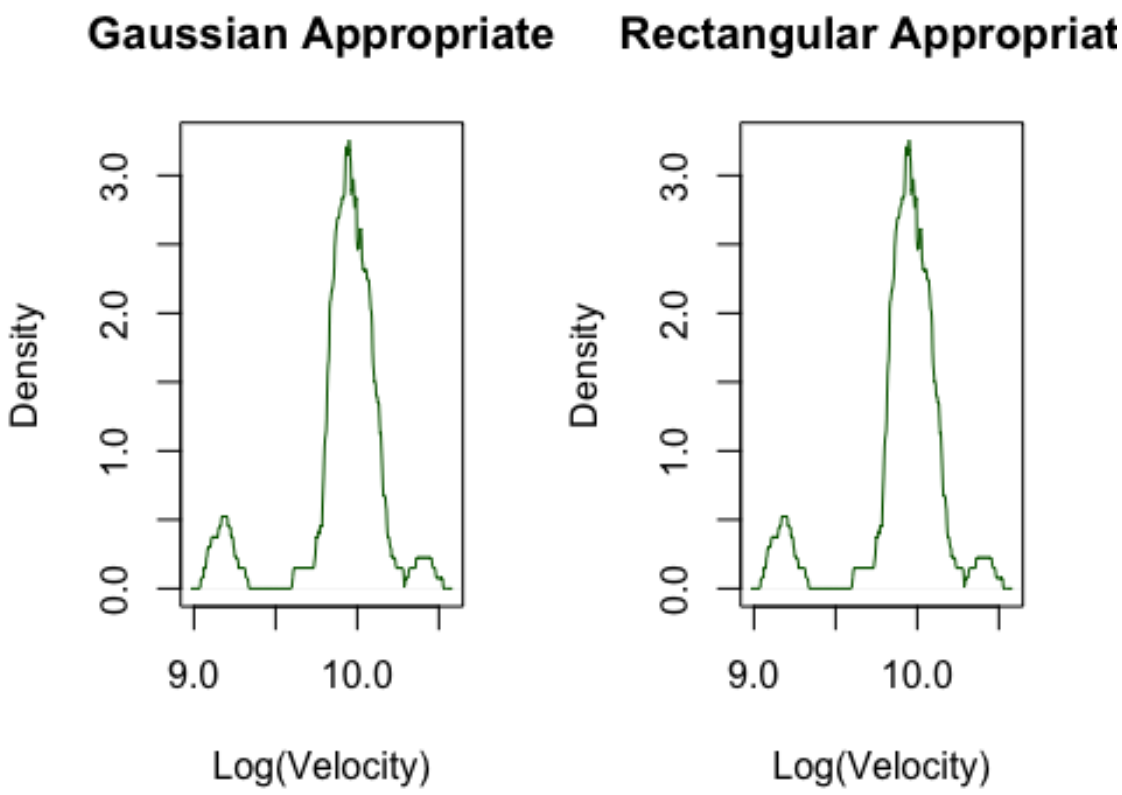
Discussion

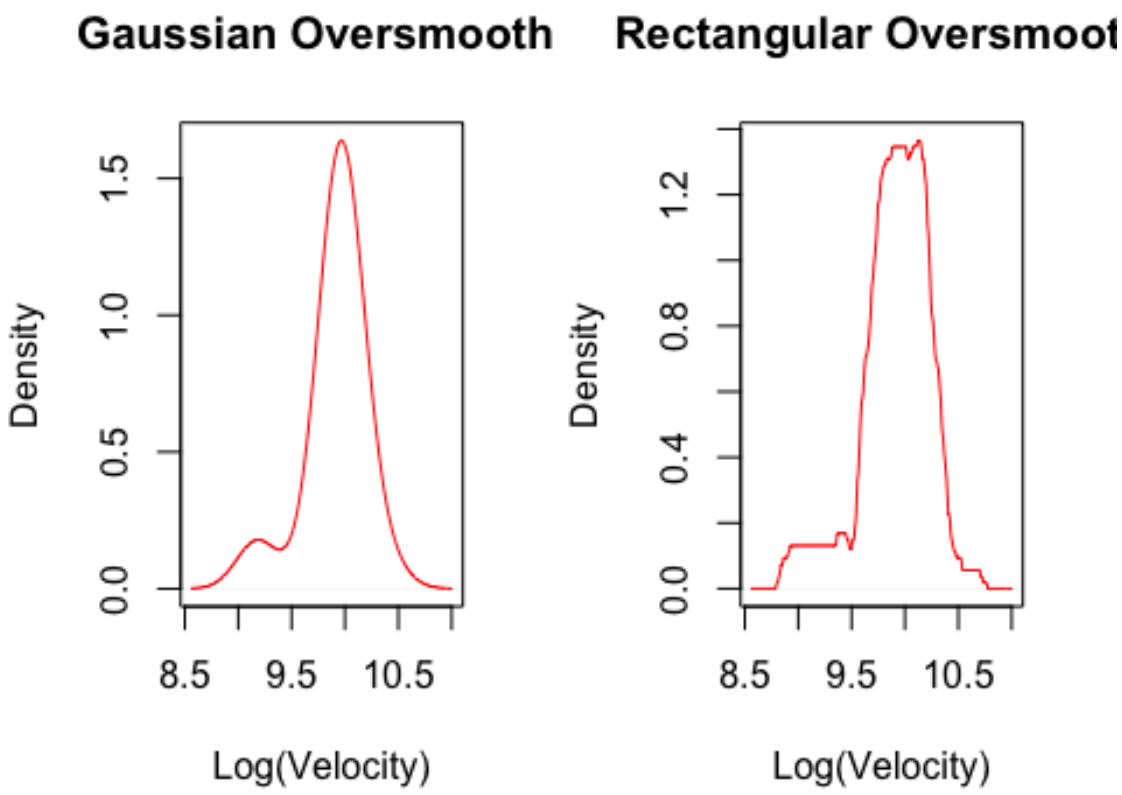
Part 1b:

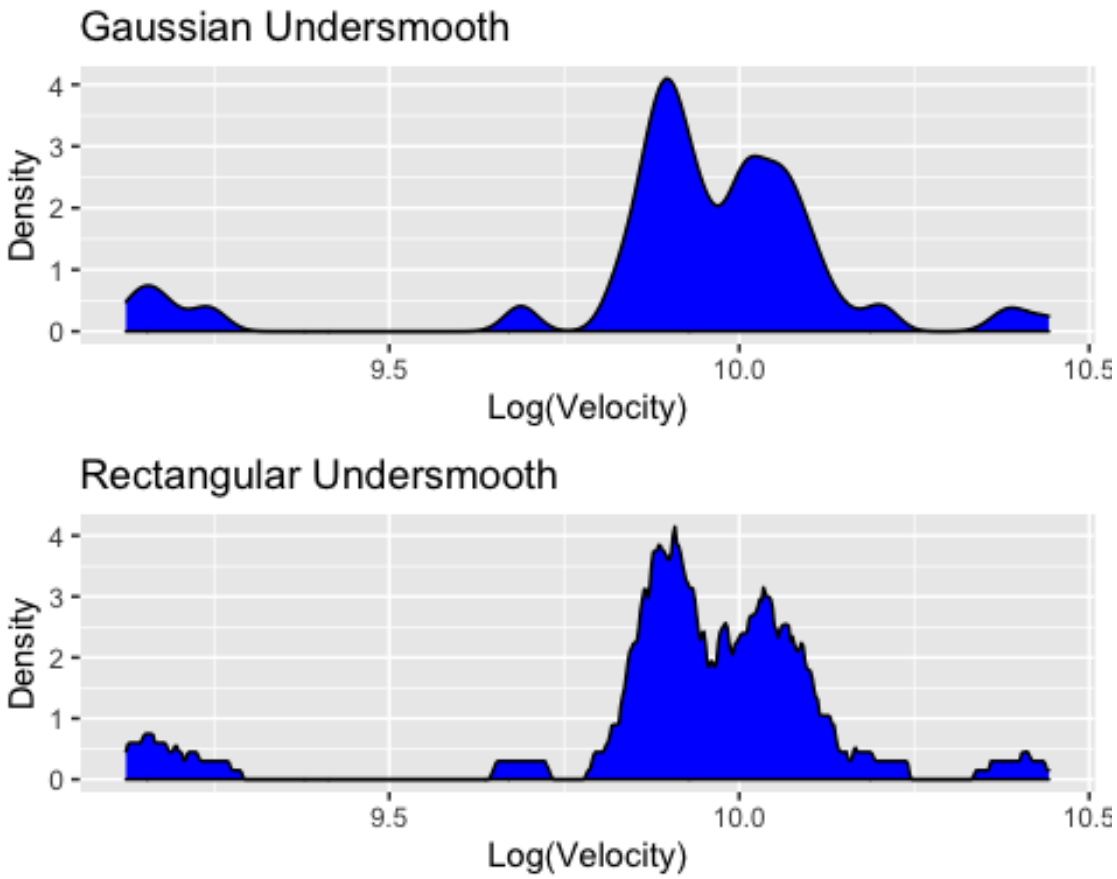
In figure 1b.1 we see normal distribution of the log velocity centered around 10 with negative skewness. Similarly, we also find normal distribution in the true histograms with 7 bins for both the base r and ggplots with true the true histograms still maintaining density y-axis. Figure 1b.3 which utilized the qplot function appears to have 3 clusters which suggests a normal distribution.

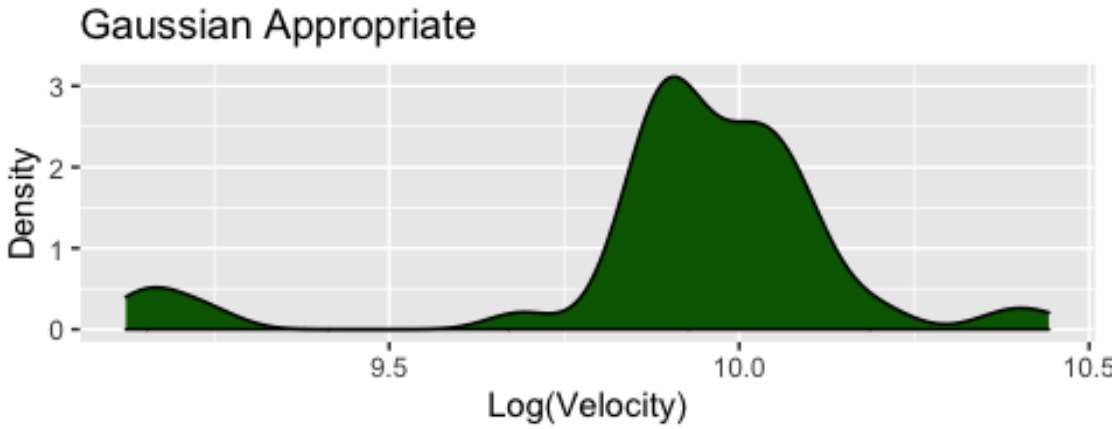
c) Construct kernel density estimates using two different choices of kernel functions and three choices of bandwidth (one that is too large and "oversmooths," one that is too small and "undersmooths," and one that appears appropriate.) Therefore you should have six different kernel density estimates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

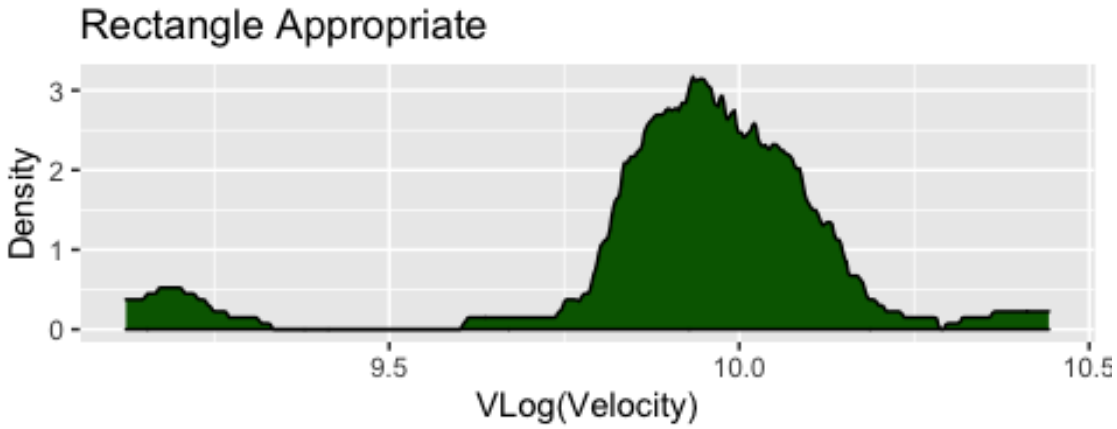

Gaussian Undersmootl
Density
0
1
2
3
4
9.0
9.5
10.5
Log(Velocity)
Rectangular Undersmoo
Density
0
1
2
3
4
9.0
9.5
10.5
Log(Velocity)

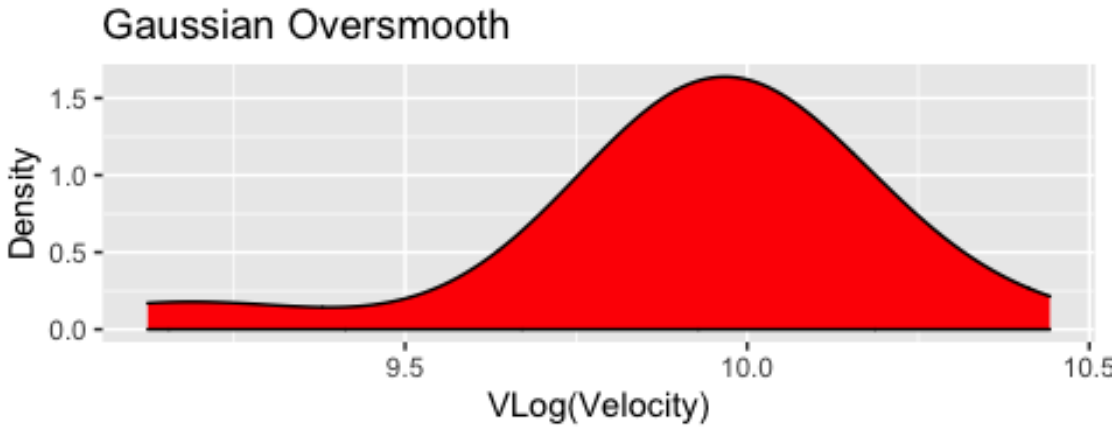

Gaussian Appropriate
Density
0.0
1.0
2.0
3.0
9.0
10.0
Log(Velocity)
Rectangular Appropriat
Density
0.0
1.0
2.0
3.0
9.0
10.0
Log(Velocity)

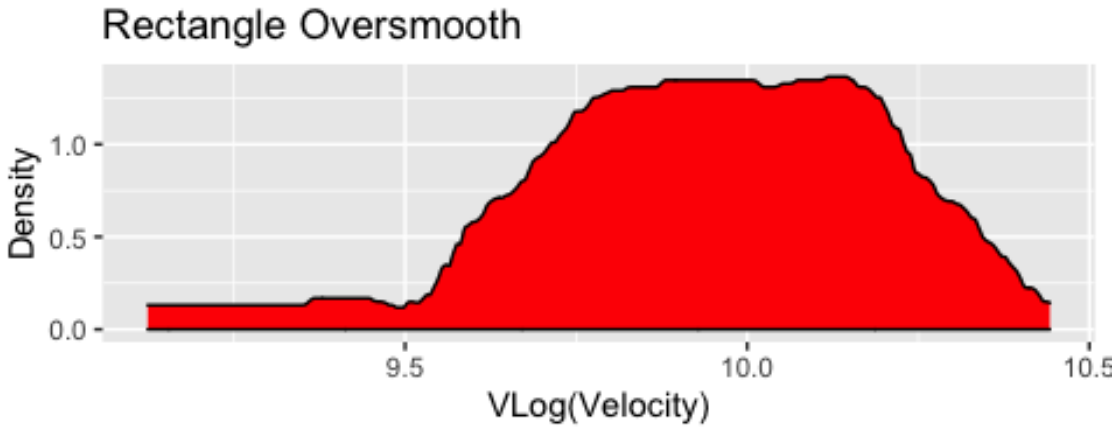

Gaussian Oversmooth
Density
0.0
0.5
1.0
1.5
8.5
9.5
10.5
Log(Velocity)
Rectangular Oversmoot
Density
0.0
0.4
0.8
1.2
8.5
9.5
10.5
Log(Velocity)


Gaussian Undersmooth
Density
0
1
2
3
4
9.5
10.0
10.5
Log(Velocity)
Rectangular Undersmooth
Density
0
1
2
3
4
9.5
10.0
10.5
Log(Velocity)


Gaussian Appropriate
Density
3
2
1
0
9.5
10.0
10.5
Log(Velocity)


Rectangle Appropriate
Density
3
2
1
0
9.5
10.0
10.5
VLog(Velocity)





Discussion part c: Looking at the plots, it's evident that oversmooth plots tend to overgeneralize the density estimation. The caviat of the oversmoothing plots is that they tends to fail at identifying the existence of super clusters, which is where undersmooth plots are very useful in, particularly the undersmooth plots show multimodal distributions which may potentially uncover the existence of superclusters.

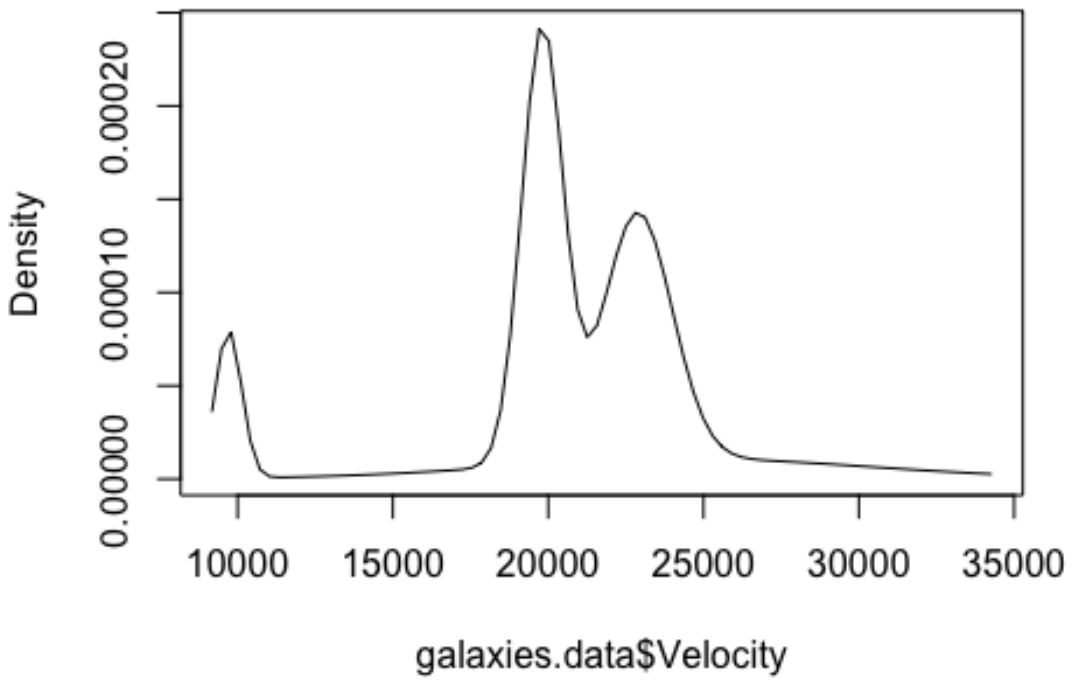
d) What is your conclusion about the possible existence of superclusters of galaxies? How many superclusters (1, 2, 3, …)? (Hint: the existence of clusters implies the existence of empty spaces between galaxies.)

When looking at the appropriate kernel density estimation function with the right bandwidth it has two tails and therefore, given the multimodal nature of the velocity distributions I suspect their might be upto 4 superclusters.

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
##       -765.7316 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
```




```
## 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
```

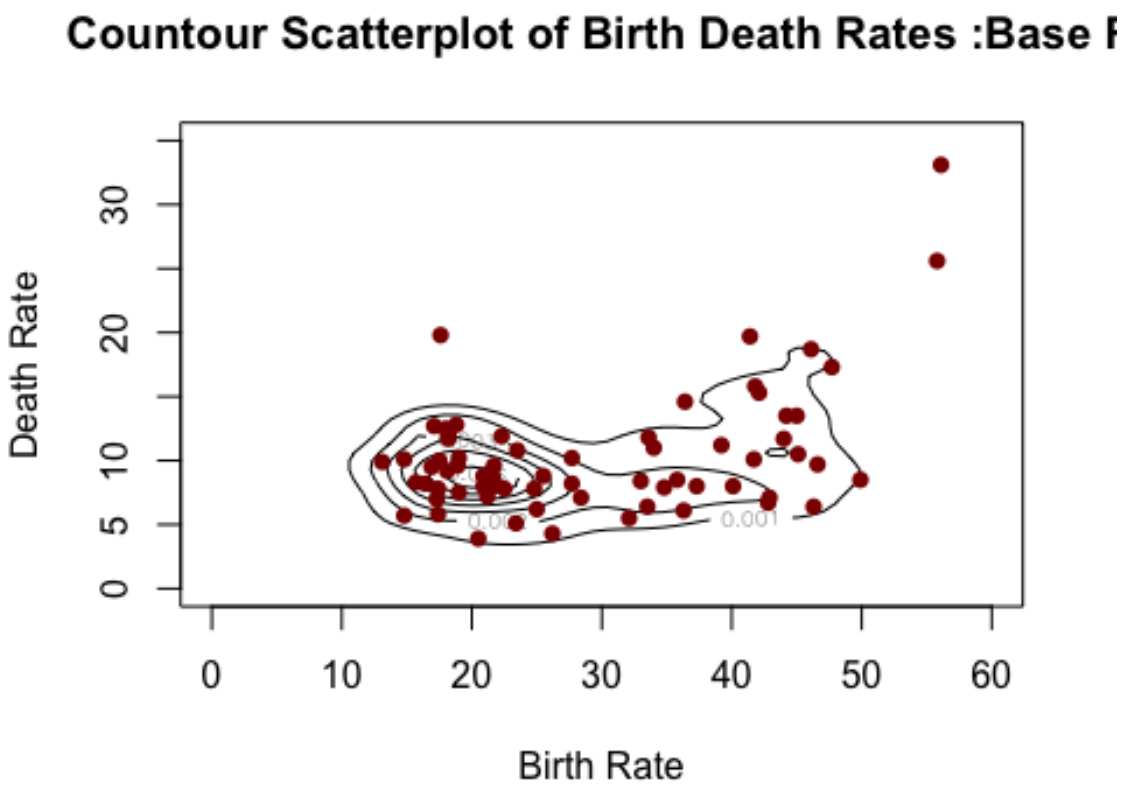
Discussion

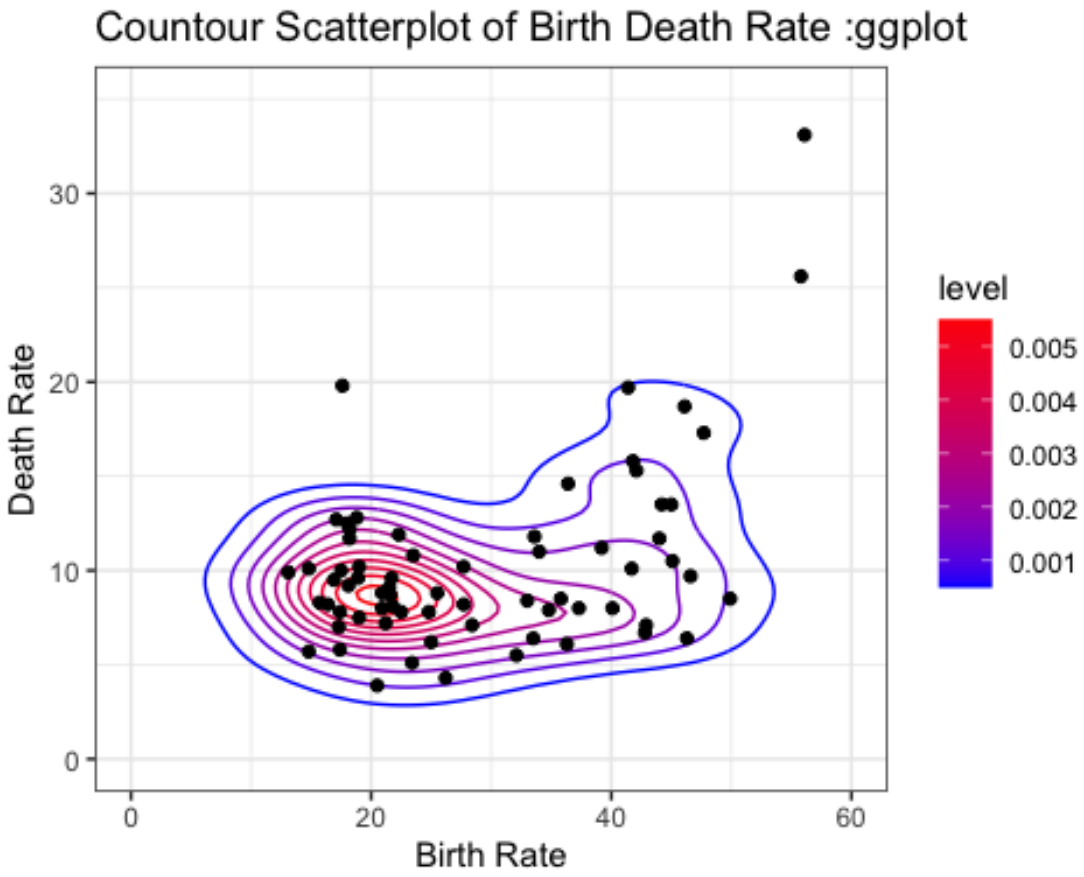
part 1e: The density plot of the model shows three distinct superclusters with the far right tail not being as distinct.After examining the best BIC of the models it is evident that the best possible clustering is at four superclusters, with a three cluster estimate being very close in BIC; most likely due to the low number of observations at the far right tail (high velocity).

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.

```
##     birth death
## alg  36.4  14.6
## con  37.3   8.0
## egy  42.1  15.3
## gha  55.8  25.6
## ict  56.1  33.1
## mag  41.8  15.8

## [1] 2

## [1] 69
```


Countour Scatterplot of Birth Death Rates :Base F
Death Rate
0
5
10
20
30
0
10
20
30
40
50
60
Birth Rate
0.001

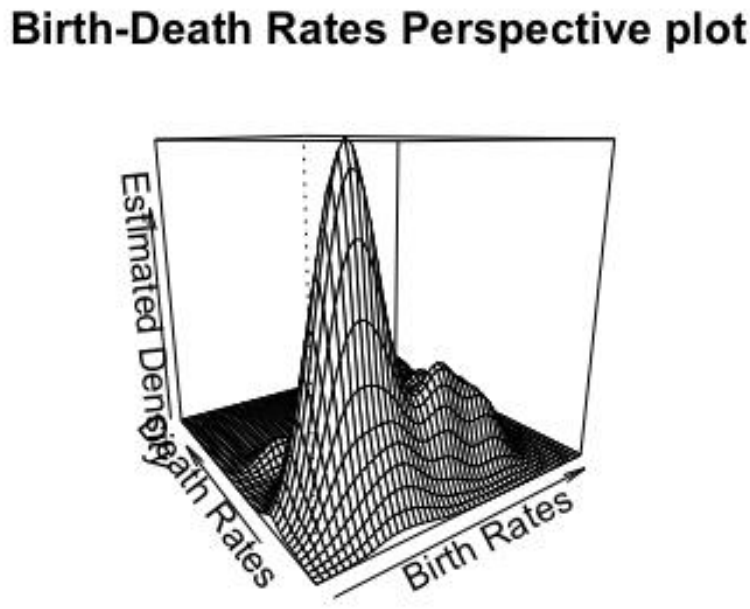


Discussion part 2a: we see the data are clustered near the point when 'birth rate' = 20 and 'death rate' = 10.

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

It shows that most of the data points, birth rate is at least 2 times higher than death rate on a ratio of 2:1 in most countries. There are countries, and clusters of countries, whose birth rate far exceeds that ratio, but only one country has a higher death rate than it does birth rate.

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

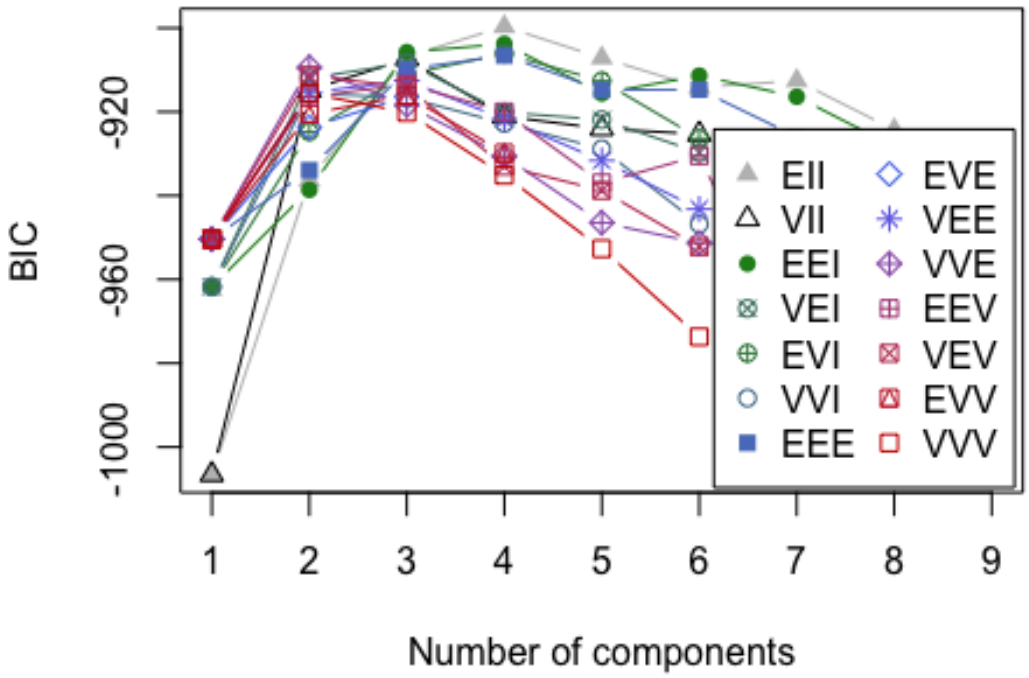


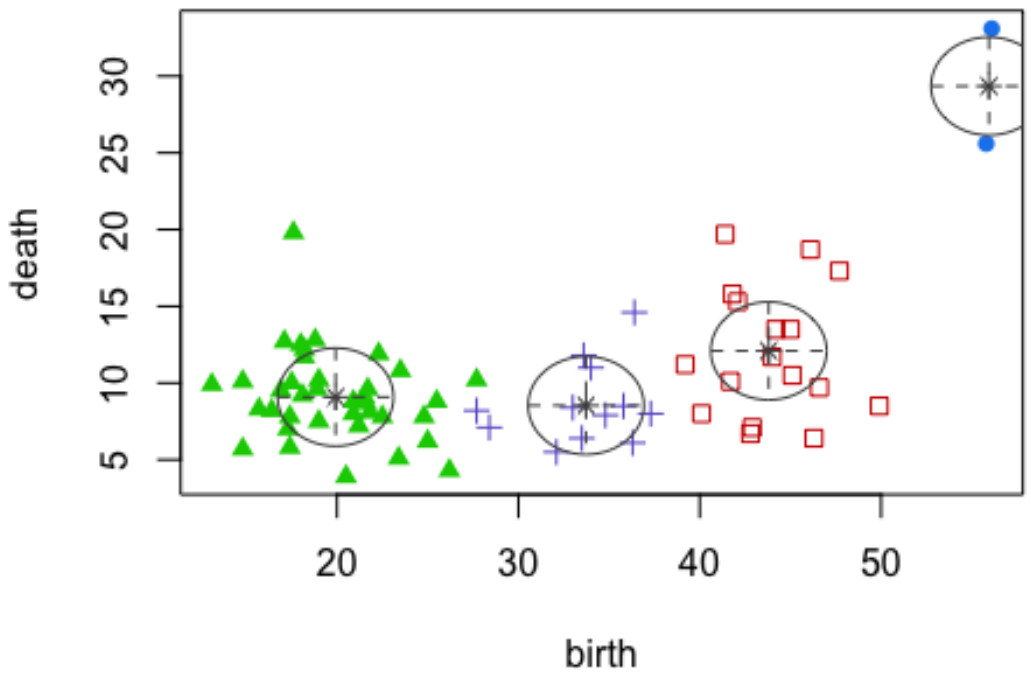
Discussion part 2c: The perspective plot validates what we established in 2b which that birth rate is twice as high as death rate and the majority of the data observations have a death rate that is proportionally smaller than the birth rate.

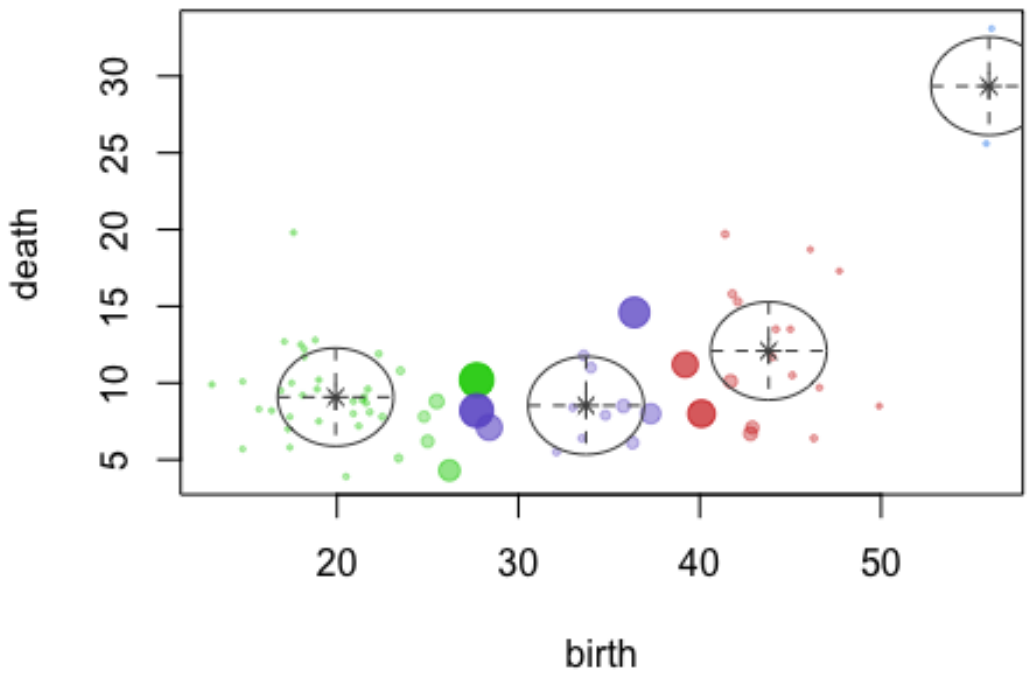
d) Fit a finite mixture model using the Mclust() function in R (from the mclust library). Summarize this model using BIC, classification, uncertainty, and/or density plots.

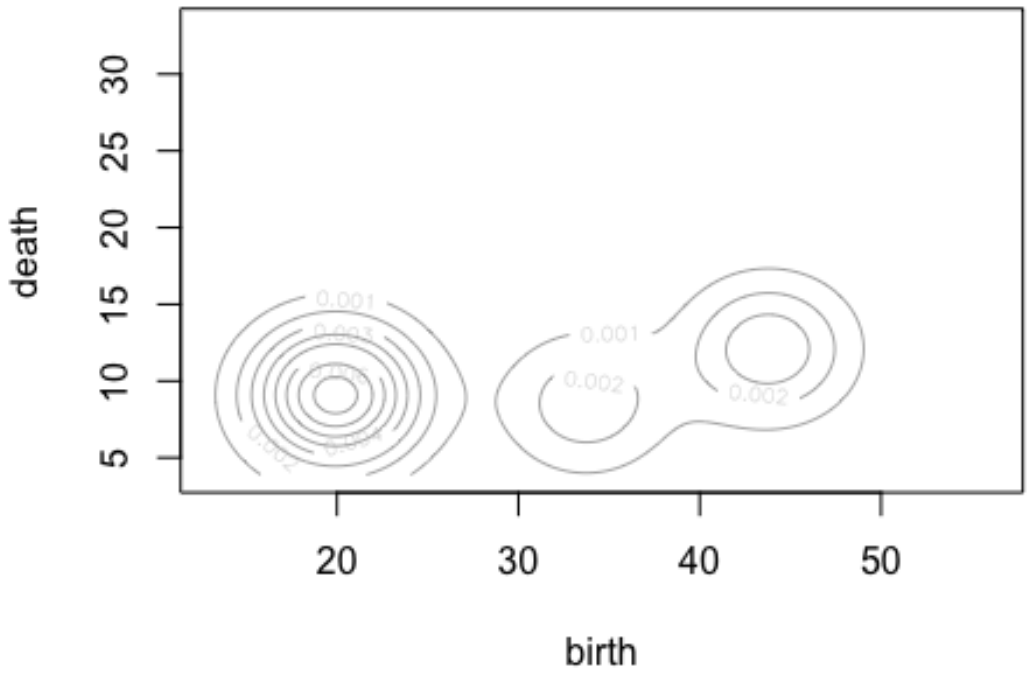
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194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


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


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



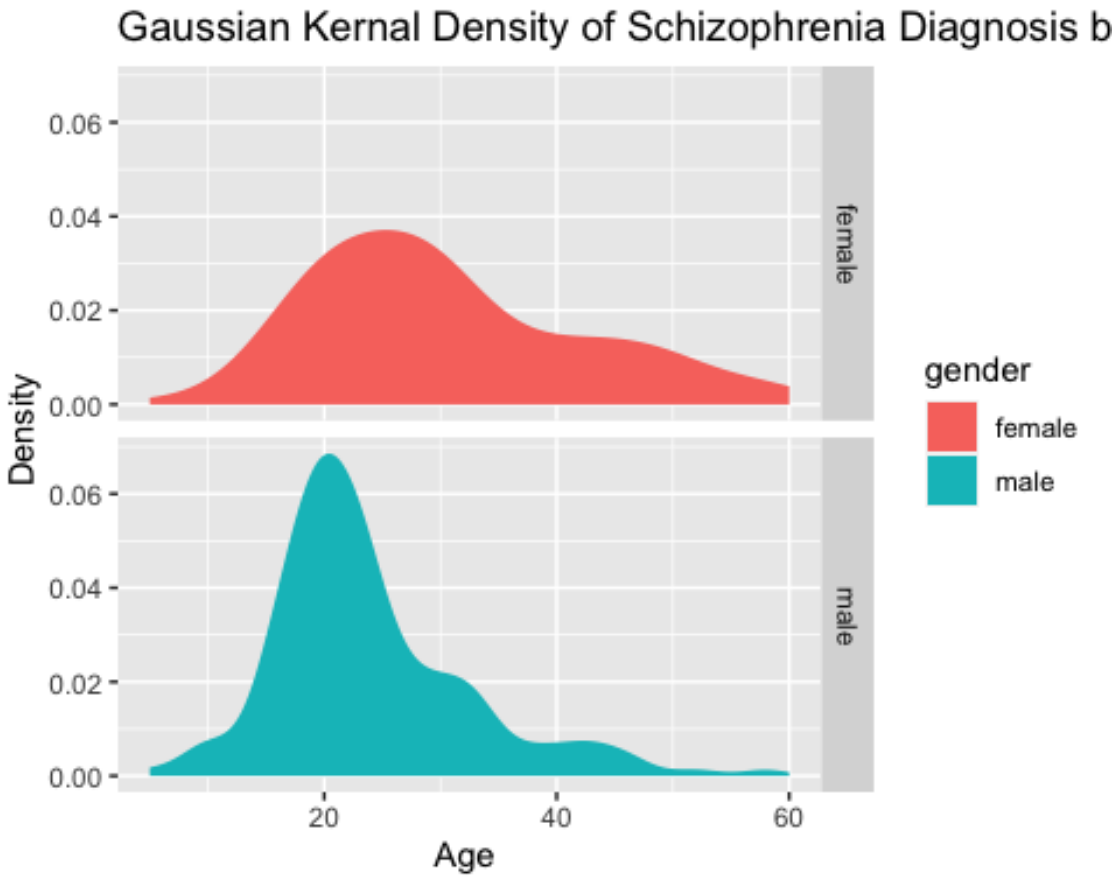

```
## Bayesian Information Criterion (BIC):
##          EII        VII       EEI       VEI       EVI       VVI       EEE
## 1 -1006.5723 -1006.5723 -961.7502 -961.7502 -961.7502 -961.7502 -950.3669
## 2  -936.3442  -914.8037 -938.6127 -911.9710 -924.7310 -915.6217 -933.9448
## 3  -906.7729  -907.3547 -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:
```

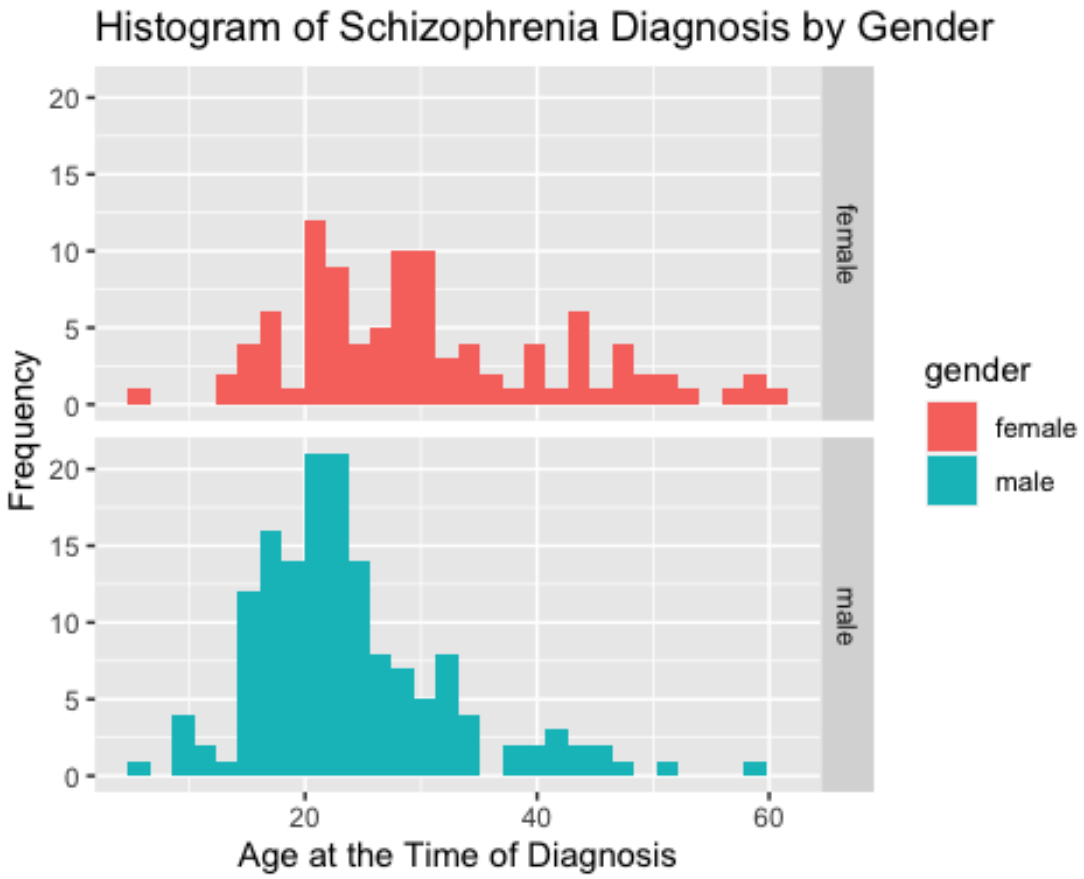
```
##     EII,4     EEI,4     EEI,3
## -899.6481 -903.7704 -905.7403
```

e) Discuss the results in the context of Birth and Death Rates. The plots above provide evidence that the data has 4 unique clusters with birth rate which is the larger cluster having a ratio of 2:1 against death rate, this is shown by the countour and the perspective plots most countries having birth rate of 20 and a death rate of 10.

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)

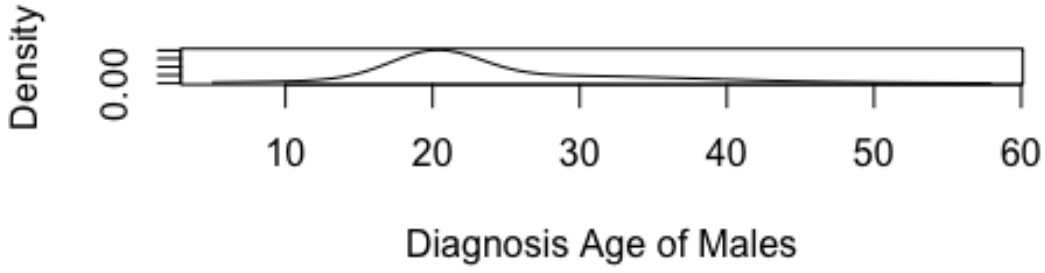

Gaussian Kernal Density of Schizophrenia Diagnosis b
Density
0.06
0.04
0.02
0.00
female
male
20
40
60
Age
gender
female
male

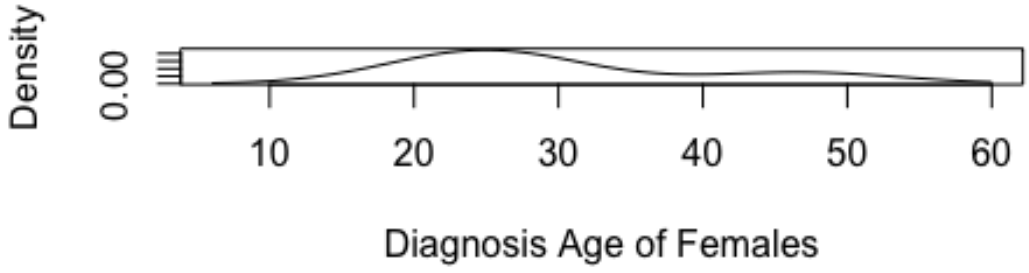


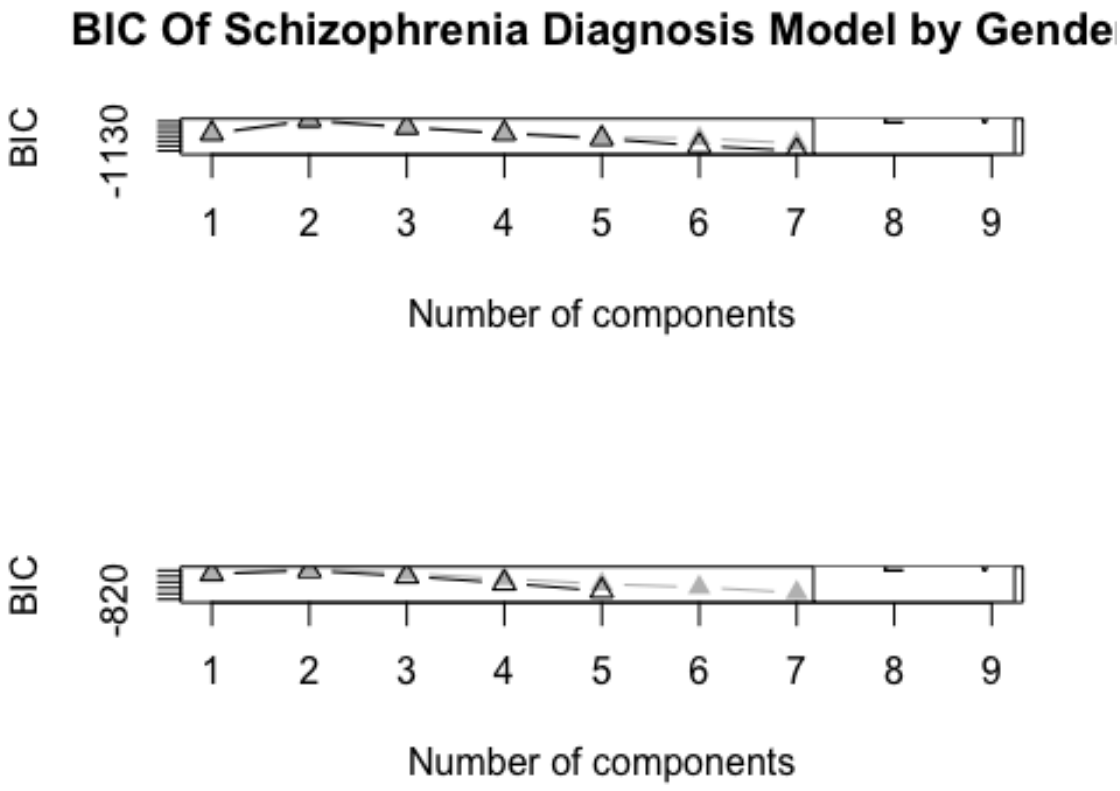

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##       -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

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

ıeled Density Estimate of Schizophrenia Diagnois by







Discussion

Part 3:From our density plots and summary it's reasonably clear that females have lower rates of schizophrenia than their counterparts, particularly in their twenties. Inversely, males suffer from higher rates of schizophrenia in the early to late twenties. Both groups have clustering at different ages, females have clustering at 25 and 46 years whereas males have clustering at 20 and 27 years. Therefore, the clustering and density estimation does supports the theory for early onset schizophrenia in men and late onset schizophrenia in women.

Works Cited

1.Michael and Saunders, Density Estimation:Chapter 8 on HB Non-Parameteric Density Estimation:Kernel Density Estimation 2.Michael and Saunders, Density Estimation:Chapter 8 on HB Parameteric Density Estimation:Finite Mixture Models 3.Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010. 4.Neupane, Achal."Density Estimation." Achal Neupane,1 Sept.2019,achalneupane.github.io/achalneupane.github.io/post/density_estimation/ 5.Kuipers, Kevin."RPubs-STAT 601 Homework 4." RPubs,4 Sept. 2018, rpubs.com/kkuipers/STAT602HW2