Homework 4

Alex Soupir

September 20, 2019

Packages: MASS, GGPlot, Mclust, HSAUR3

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

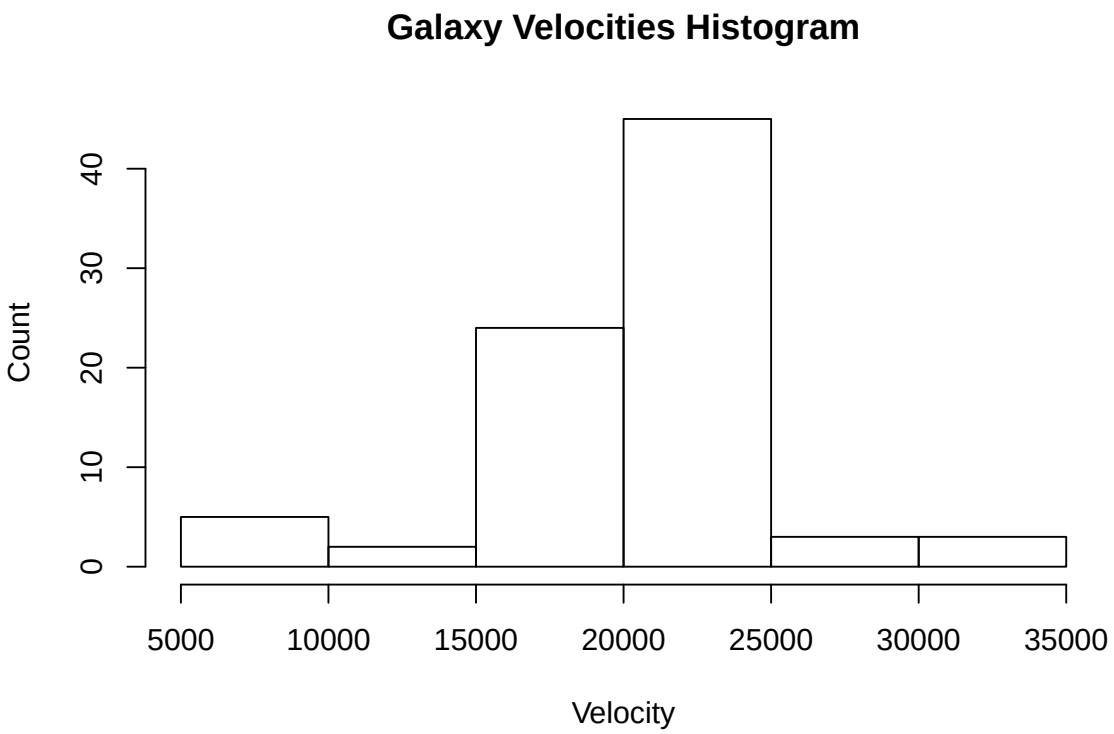
You do not need to include the above statements.

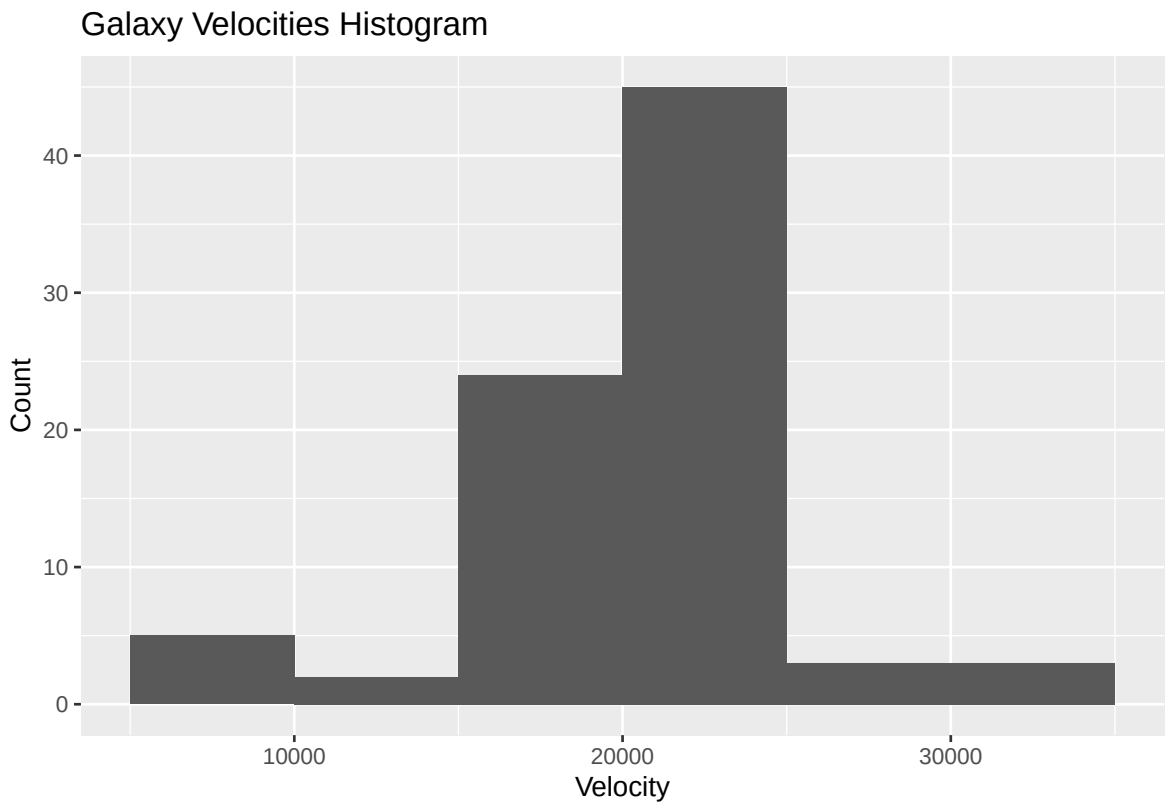
Please do the following problems from the text book R Handbook and stated.

1. 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)

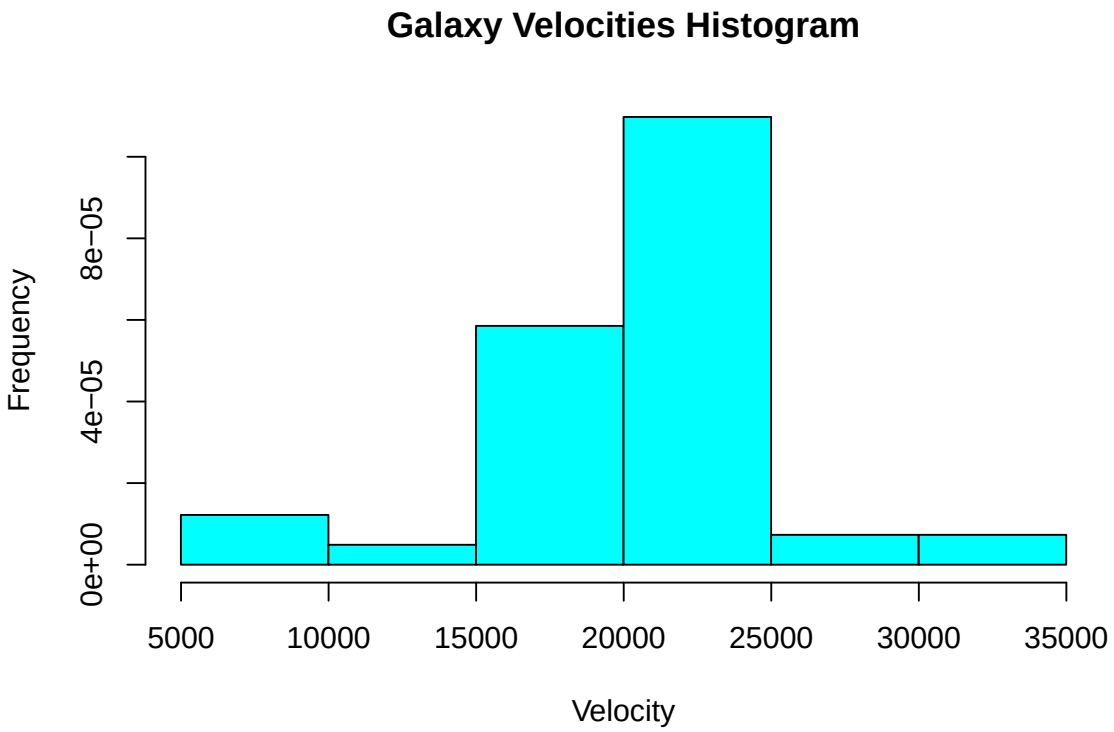
   a.) Construct histograms using the following functions:

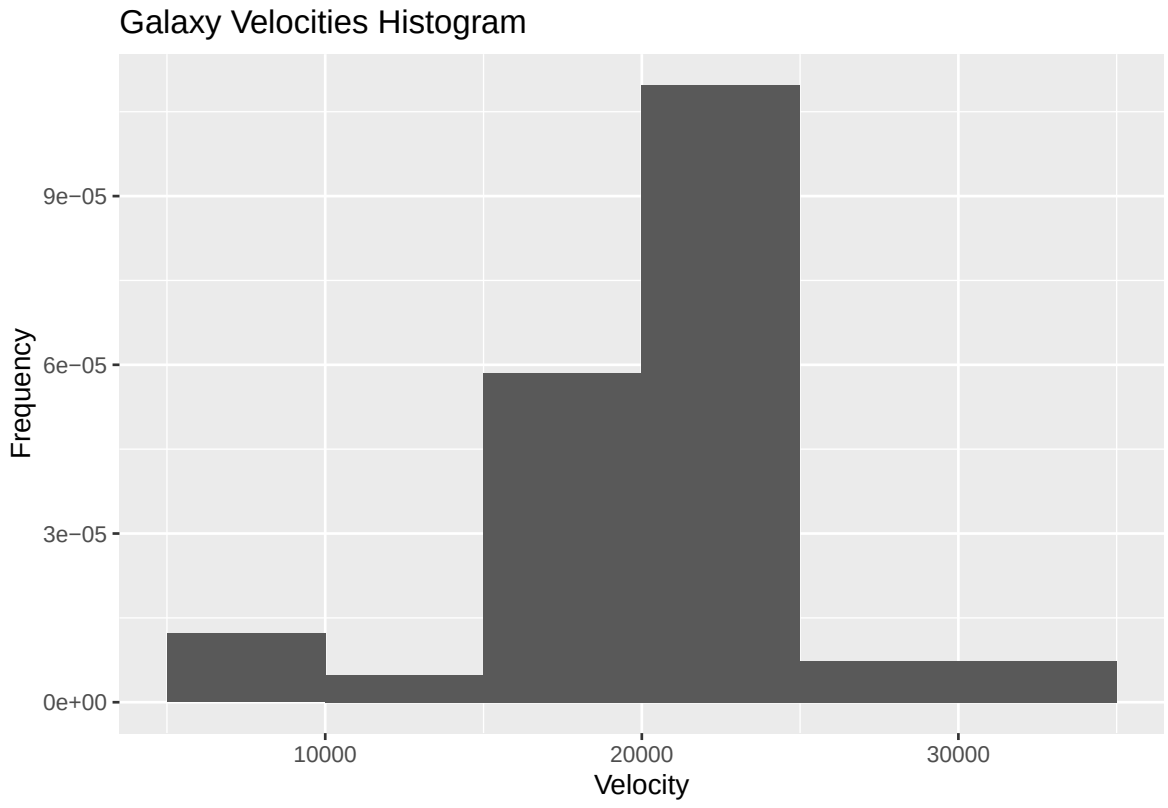
   `-hist() and ggplot()+geom_histogram()`



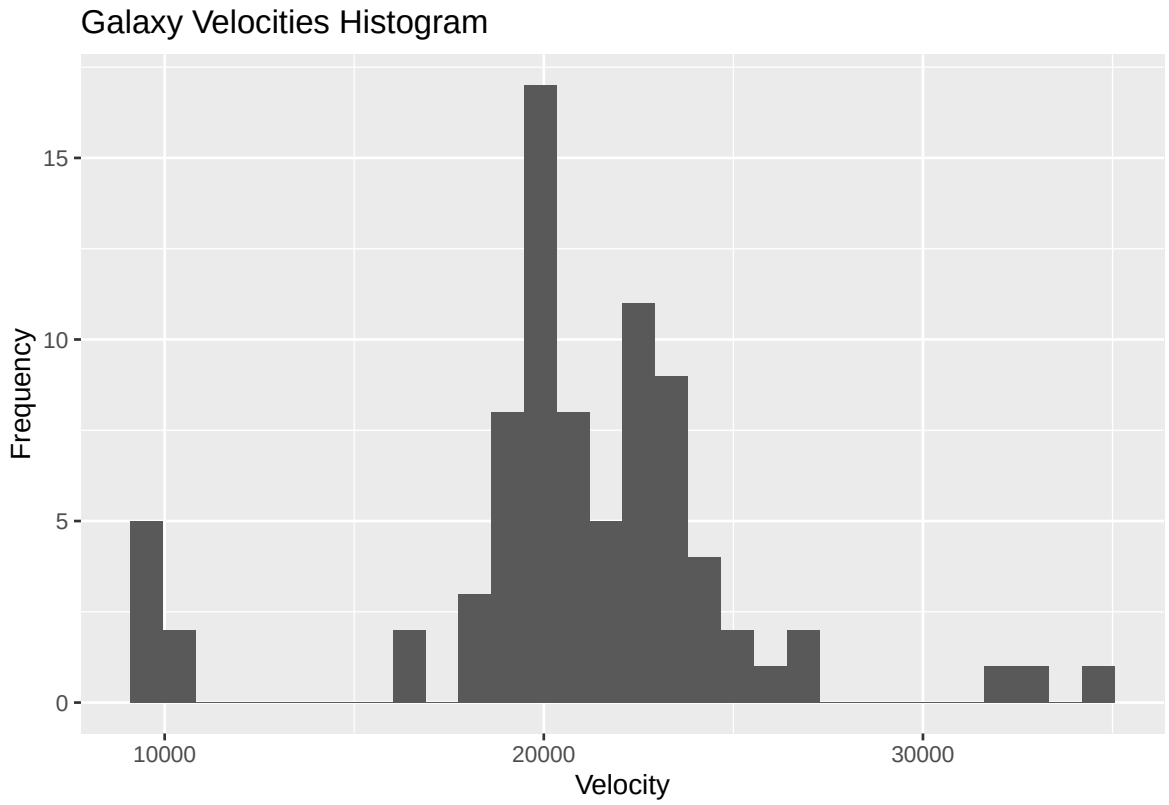


-truehist() and ggplot+geom_histogram() (pay attention to the y-axis!)


Galaxy Velocities Histogram
Frequency
0e+00
4e−05
8e−05
5000
10000
15000
20000
25000
30000
35000
Velocity


Galaxy Velocities Histogram
Frequency
0e+00
3e−05
6e−05
9e−05
10000
20000
30000
Velocity

```
-qplot()
```

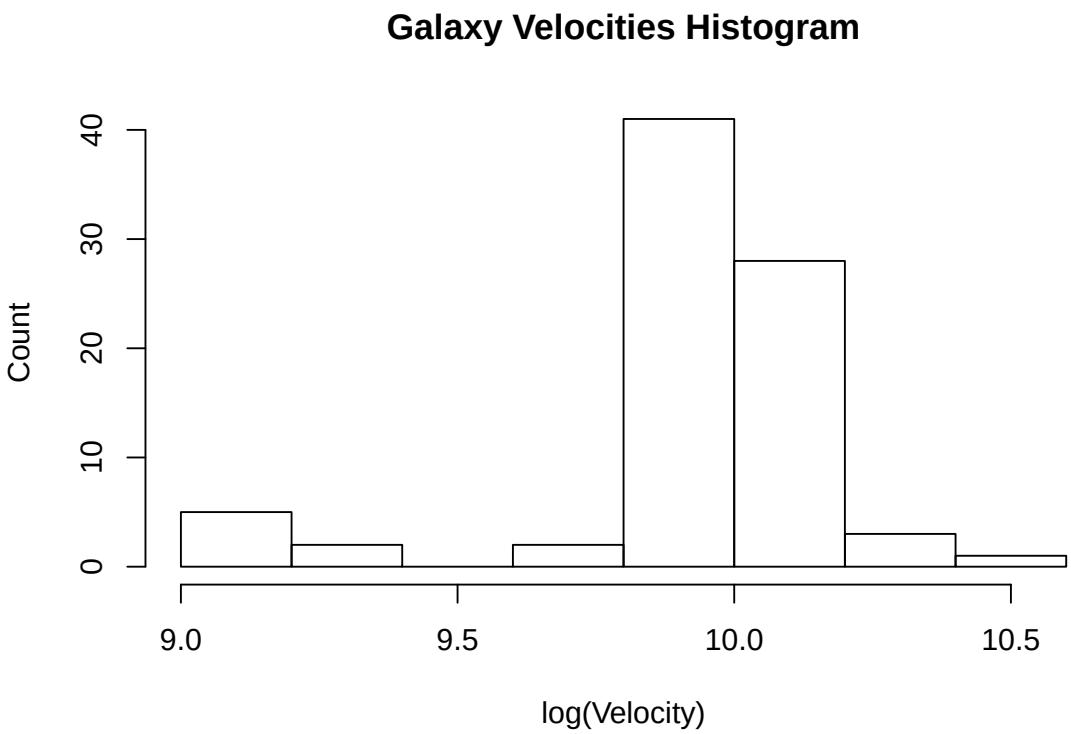


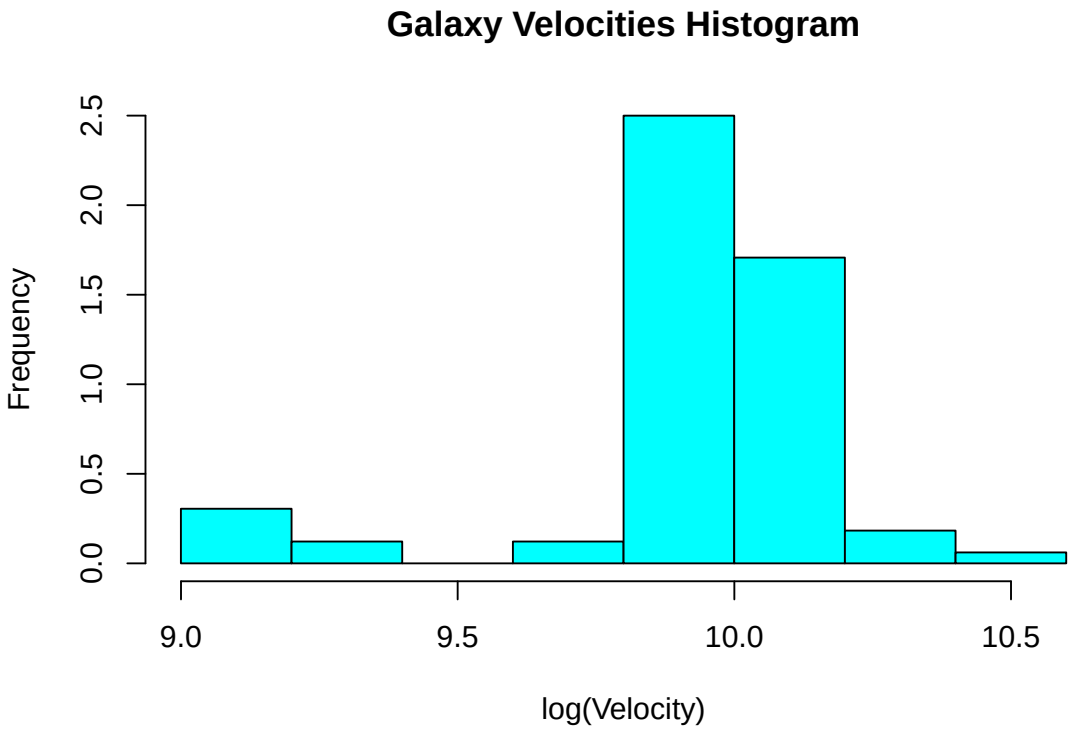
Comment on the shape and distribution of the variable based on the three plots. (Hint: Also play around with binning)

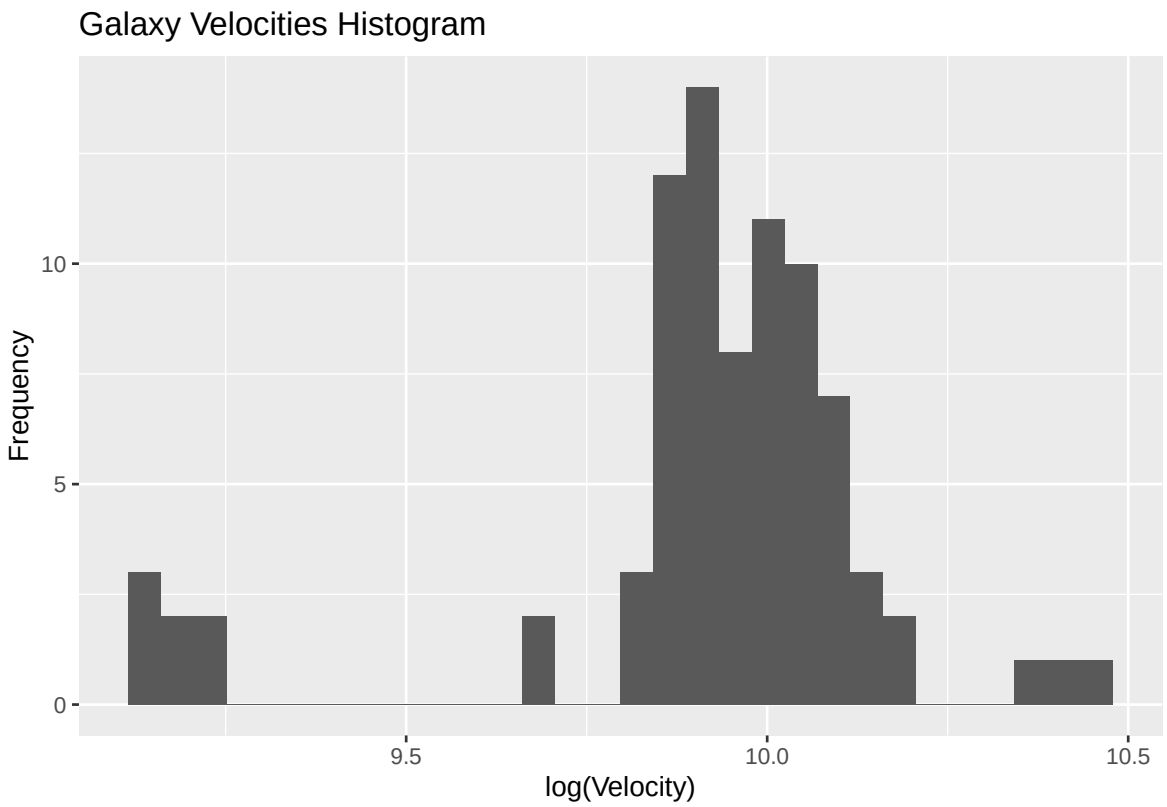
The shape of both the regular histogram and the *truehist()* is the same, but the y-axis of *truehist()* is the frequency, with the total area being 1 while *hist()* provides the number of galaxies in each bin. The *qplot()* function also provides the counts for the number of galaxies in each provided bin. The number of bins (and bin width) for *hist()* and *truehist()* is the same by default while *qplot()* chooses a much smaller bin width which tells a much different story.

GGPlot picks a different way to plot its histogram, picking different starting bin locations. This caused some difficulties since they weren't looking the same as the default outputs for the other first functions. The origin for ggplot had to be set at 5000 in order to make the plots look similar.

b.) Create a new variable *loggalaxies* = log(galaxies). Construct histograms using the functions in part a.) and comment on the shape and differences.

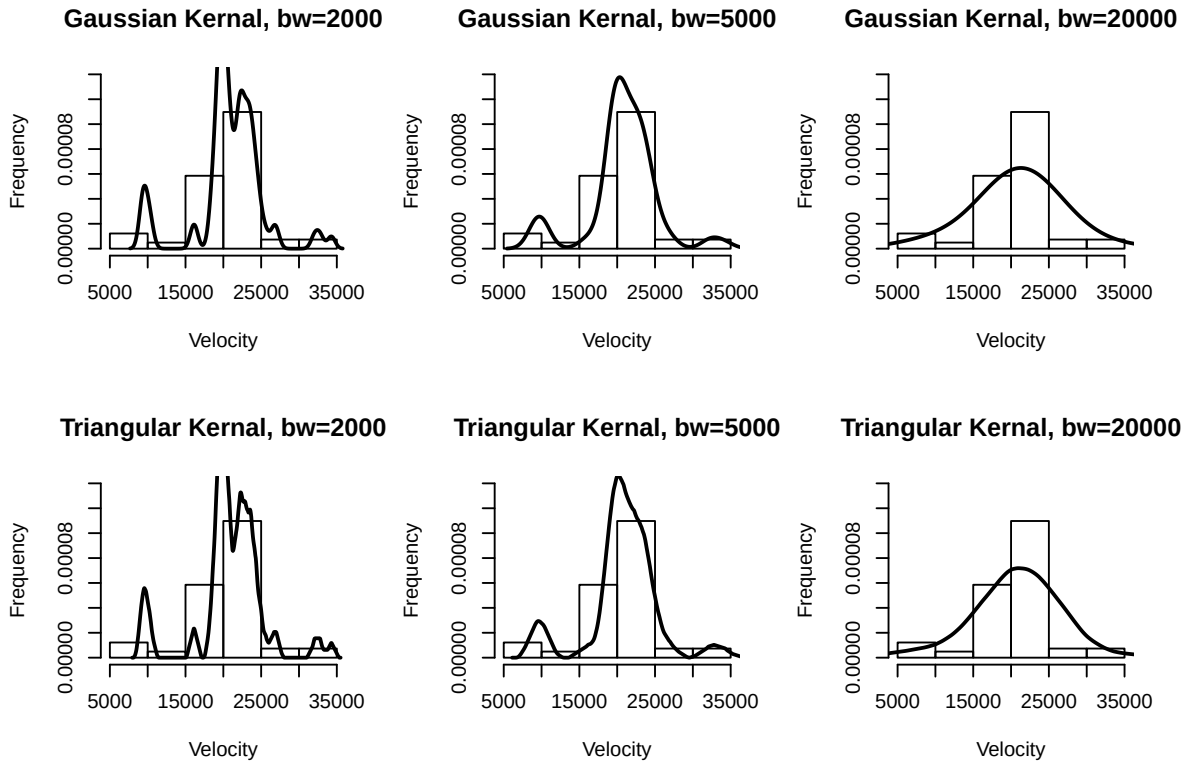

Galaxy Velocities Histogram
Count
0
10
20
30
40
9.0
9.5
10.0
10.5
log(Velocity)


Galaxy Velocities Histogram
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
9.0
9.5
10.0
10.5
log(Velocity)



The *hist()* and *truehist()* look similar when plotting the log of velocity as to when plotting the velocity. However, the *truehist()* plot y-axis now shows a value that doesn't have a total area equal to 1. The bin size of *qplot()* is also different than the other 2 methods.

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. Discuss your results. You can use the log scale or original scale for the variable.



The different bin widths changed the appearence of the line a great deal. With a bin width too low, the line is very sporadic and the bin width being too large causes over-smoothing. Between the gaussian kernal and triangular kernal, there doesn't seem to be much difference besides in the bin width of 2000 where triangular is slightly more jagged.

d.) What is your conclusion about the possible existence of superclusterd of galaxies? How many superclusters (1,2, 3, ...)?

From the first set of histograms, the *qplot()* shows that there may be 3 different superclusters. This is further supprted by the bin width of 5000 gaussian and triangular curves. The kernals with low bin widths show that there may be 2 superclusters located in the center of the velocities for a total of 4, but this may be from just the bin widths now showing appropriate information.

e.) How many clusters did it find? Did it match with your answer from (d) above? 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
```

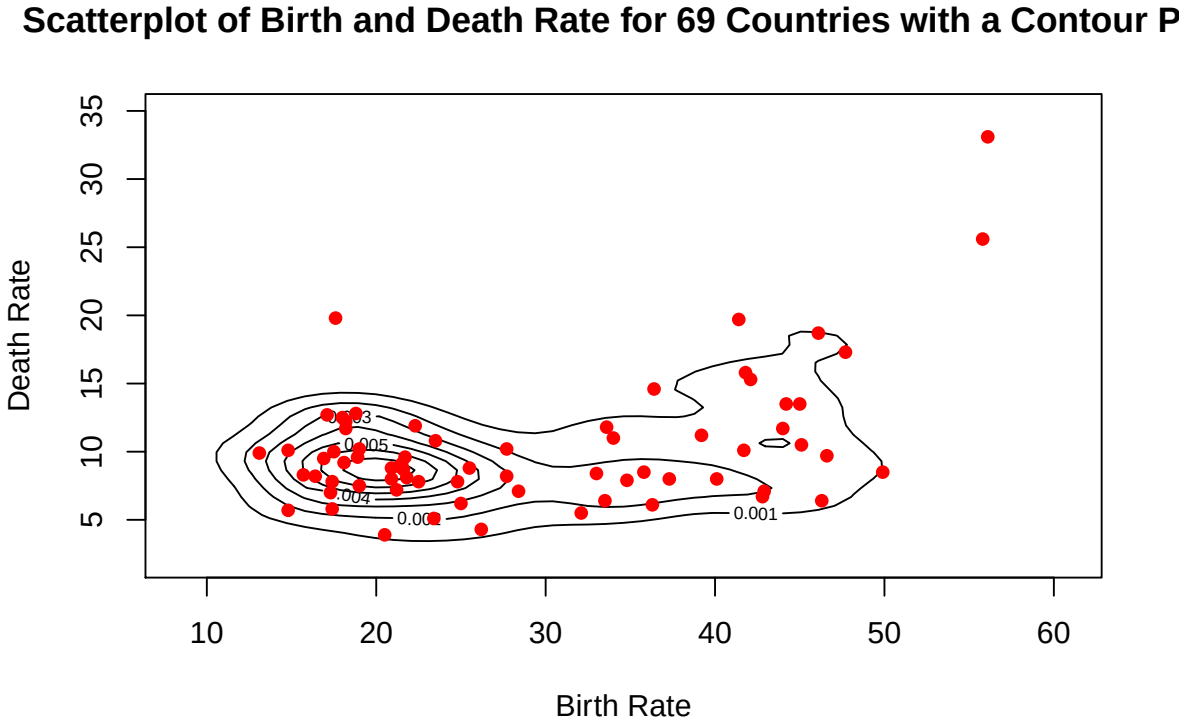
```
## Bayesian Information Criterion (BIC):
##           E         V
## 1 -1622.361 -1622.361
## 2 -1631.243 -1595.403
## 3 -1584.016 -1592.299
## 4 -1592.828 -1579.862
## 5 -1592.299 -1593.277
## 6 -1601.228 -1604.069
## 7 -1588.610 -1611.538
## 8 -1597.427 -1625.804
## 9 -1600.709 -1633.494
##
##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

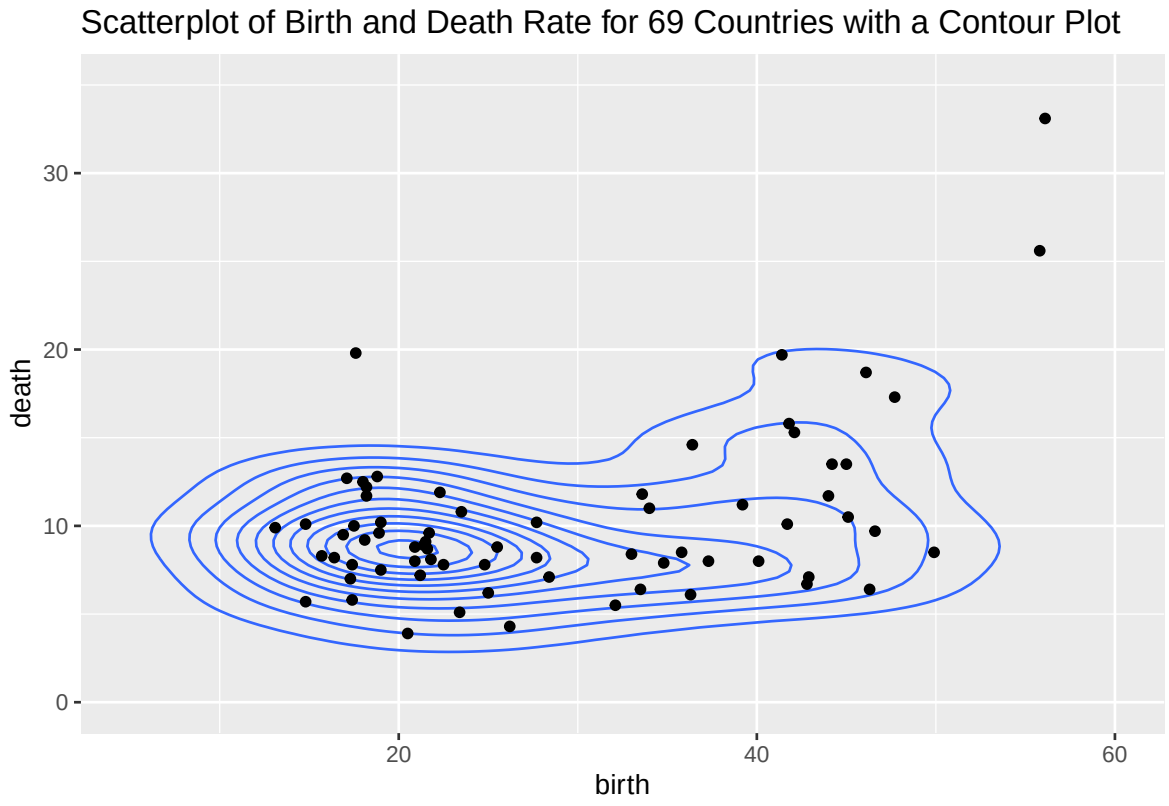
Lecture mentiones that *Mclust()* uses the EM algorithm to find the optimal number of clusters. The number of clusters that were found was 4. This doesn't match up with the initial estimate of there being 3 clusters. It does match the number of clusters estimated after thinking that there may be 2 clusters hidden within the middle peak. The cluster means are 9707.5, 19804.3, 22879.5, and 24459.5.

Looking at the best models using the BIC criterion, 4 superclusters produced the best model with the next best model being 3 superclusters. This agrees with the previous statement.

2. The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (from Hartigan, 1975). (8.2 Handbook)

   a.) Produce a scatterplot of the data and overlay a contour plot of the estimated bivariate density.

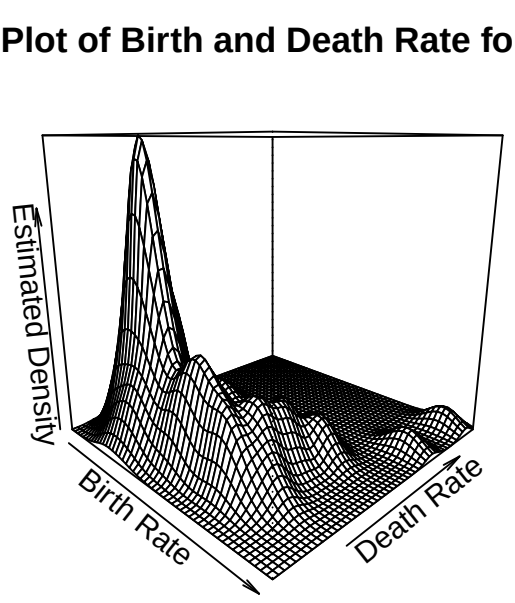




b.) Does the plot give you any interesting insights into the possible structure of the data?

The contour plot does provide someinteresting insight into the data, mainly there being a large cluster around 20 birth - 10 deaths. It shows that there are a few countries that tend to have much greater ratios between birth and death rates, some close to 50 births - 10 deaths. There are some points that are outside the contour plot which may be outliers in the data, but without further emperical tests is hard to say whether or not they actually are outliers. The base R plot and the ggplot show differing bounds of the contour plot, as well. I'm thinking this is probably due to the function used to calculate the contours which is defined as *dpik* for the base R plot, but not defined for the ggplot so it is using some internal calculation.

c.) Construct the perspective plot (persp() in R, GGplot is not required for this question).

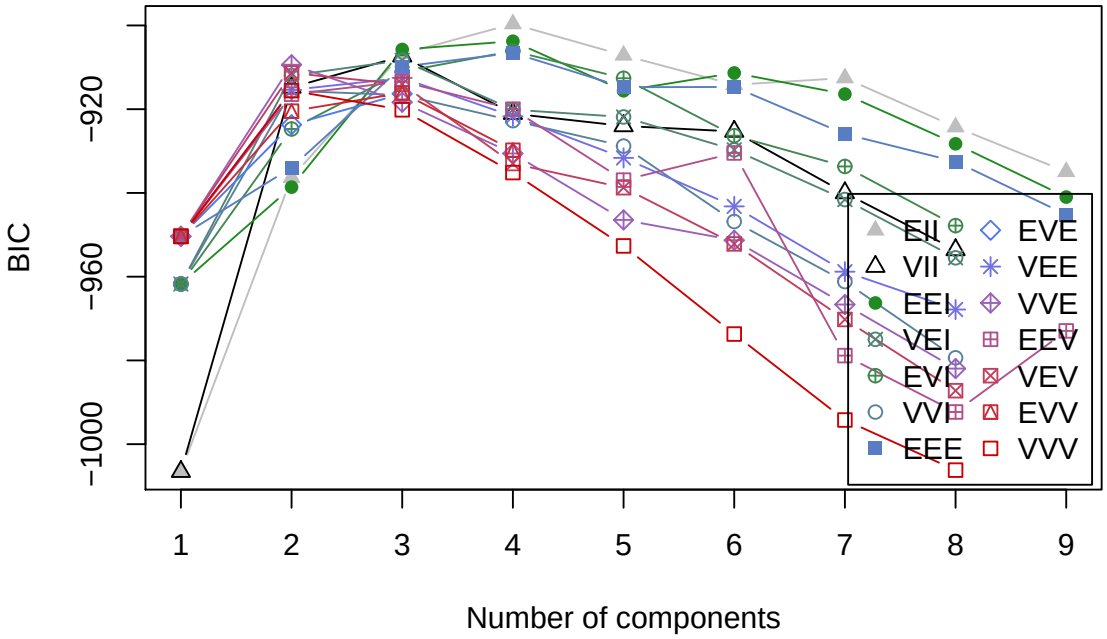


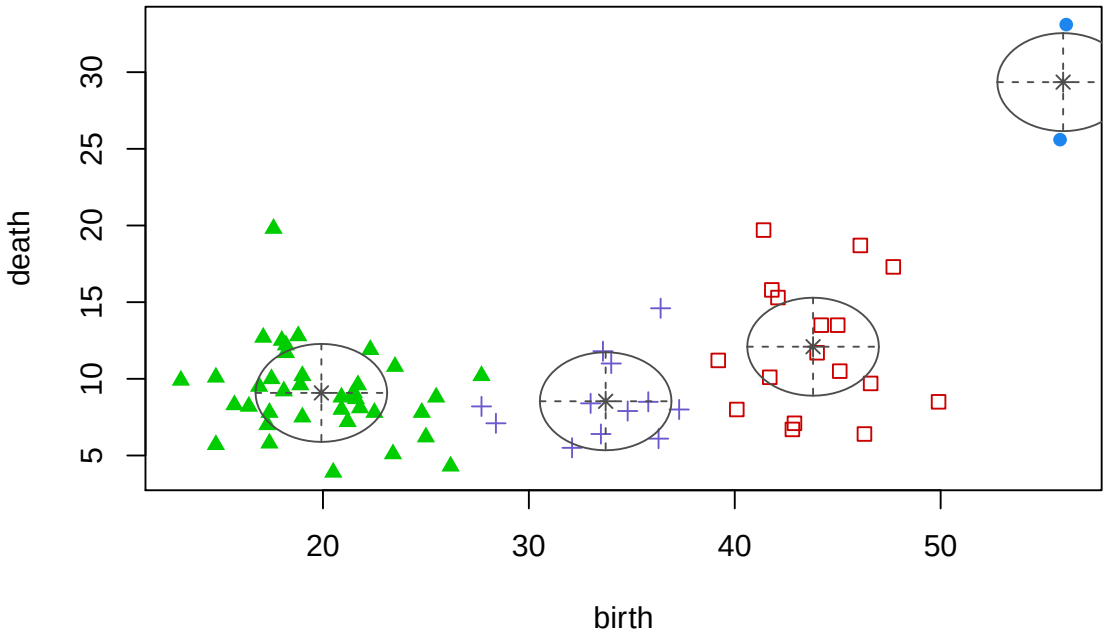
d.) Model-based clustering (Mclust). Provide plot of the summary of your fit (BIC, classification, uncertainty, and density).

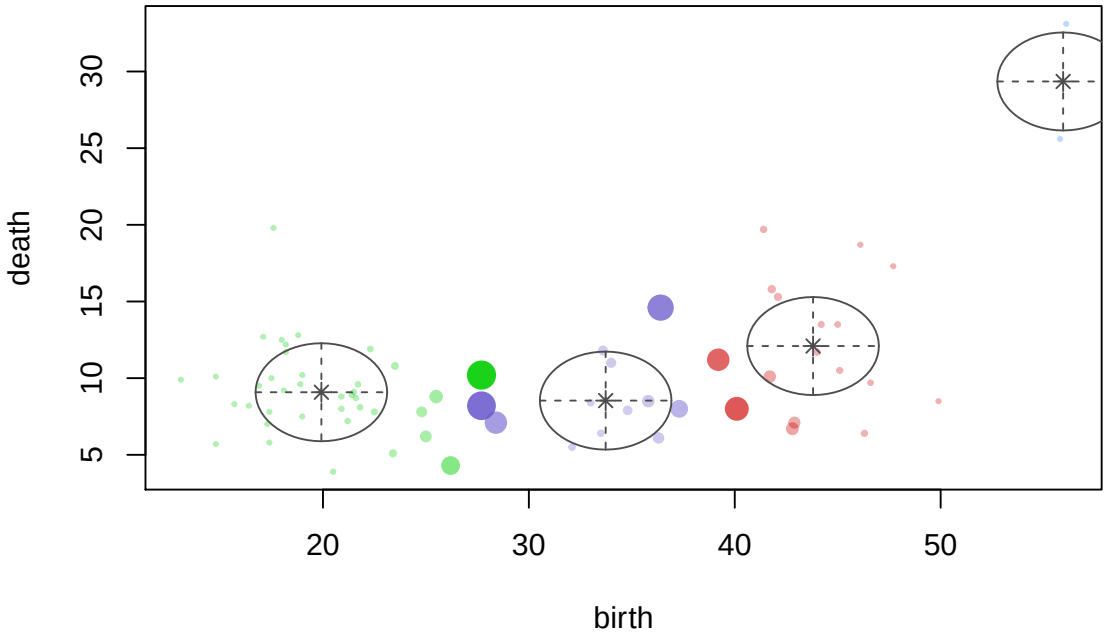
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194 69 12 -899.6481 -906.4841
##
## Clustering table:
##  1  2  3  4
##  2 17 38 12
##
## Mixing probabilities:
```

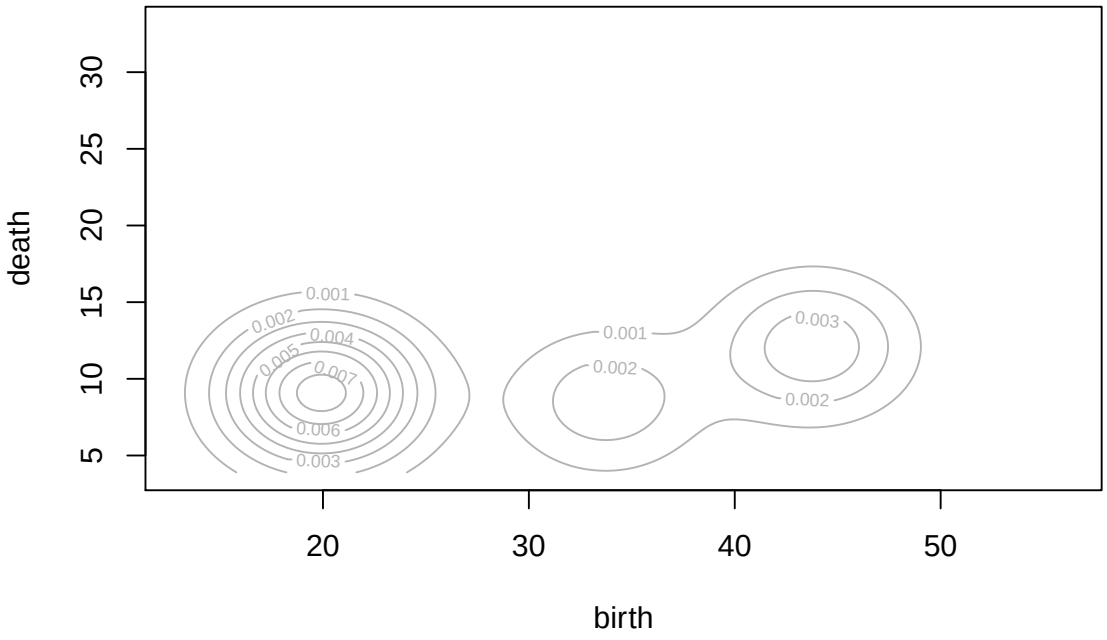
```
##          1          2          3          4
## 0.02898652 0.24555002 0.55023375 0.17522972
##
## Means:
##           [,1]     [,2]      [,3]      [,4]
## birth 55.94967 43.80396 19.922913 33.730672
## death 29.34960 12.09411  9.081348  8.535812
##
## Variances:
## [,,1]
##         birth   death
## 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
```

```
## 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
−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

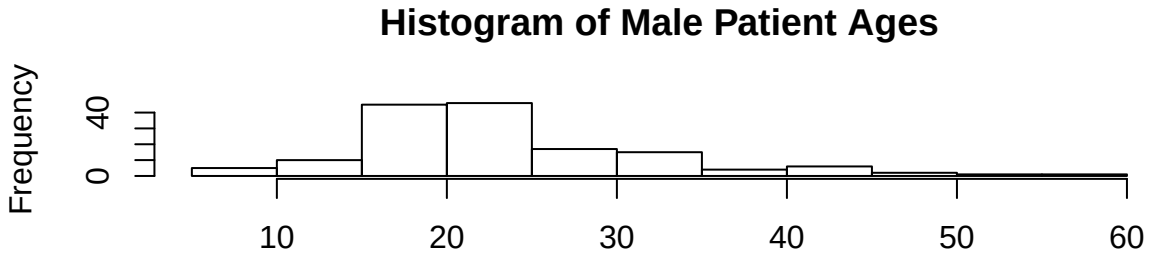


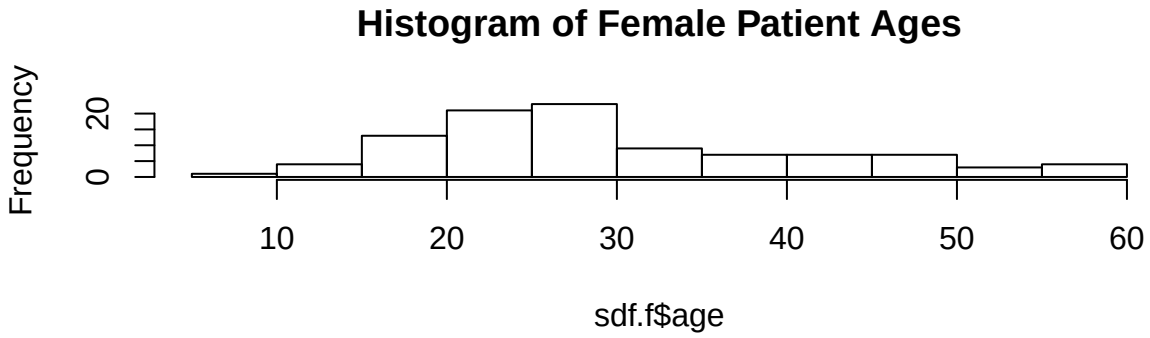


e.) Discuss the results (structure of data, outliers, etc.). Write a discussion in the context of the problem.

The *Mclust* results show that there are 4 clusters located within the data around birth:death means of [55.9:29.3, 43.8:12.1, 19.9:9.1, and 33.7:8.5]. This is confirmed with *mclustBIC* which shows that the top 2 models have 4 unique clusters, while the 3rd best has 3 unique clusters. Observing the density plot we see that a great deal of the data points fall in that 20 births to 10 deaths location, as previously mentioned. This density plot, however, does not show a 4th cluster that the classification plot shows up in the far corner, the high number of births and deaths. There is a country at about 15 births and 20 deaths, which indicates that its population is most likely decreasing, barring emmigration.

3. 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. Fit finite mixutres of normal densities separately to the onset data for men and women given in the **schizophrenia** data from **HSAUR3**. See if you can produce some evidence for or against the subtype model. (8.3 Handbook)





The histograms broke up by the gender of the patients shows that there may be a difference in the ages with males peaking between 15-25 with cases and females peaking between 25-30. There is also a difference in the number of samples per gender (males has 152 observations and females have 99 observations) that could make it difficult to determine whether the difference seen here is due to not having enough samples or true differences.

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

```
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##       -912.8923 251  4 -1847.886 -1864.194
##
## Clustering table:
##   1   2
## 213  38
##
## Mixing probabilities:
##         1         2
## 0.8390524 0.1609476
##
## Means:
##        1        2
## 22.90064 45.27658
##
## Variances:
##        1        2
## 40.66788 40.66788

## Bayesian Information Criterion (BIC):
##           E         V
## 1 -1899.227 -1899.227
## 2 -1847.886 -1850.526
## 3 -1859.118 -1861.375
## 4 -1870.313 -1871.969
## 5 -1878.766 -1887.347
## 6 -1889.716        NA
## 7 -1886.809        NA
## 8 -1897.890        NA
## 9 -1908.955        NA
##
## Top 3 models based on the BIC criterion:
##       E,2       V,2       E,3
## -1847.886 -1850.526 -1859.118
```

With the pooled data, there appears to be 2 clusters located around 22.9 and 45.3 age. This doesn't match up with the the histograms though because one of the clusters is on the tails of *BOTH* of the plots. Next, I'll try *Mclust()* on the subset data to see if the means for male and female differ when separated.

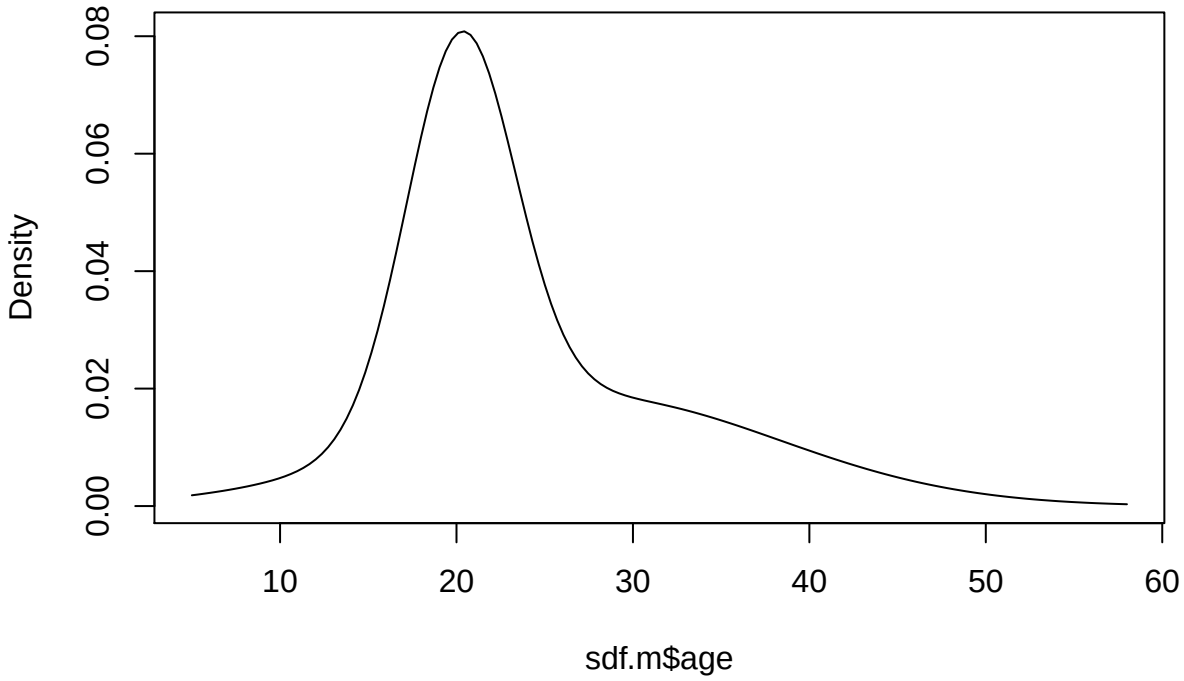
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##      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

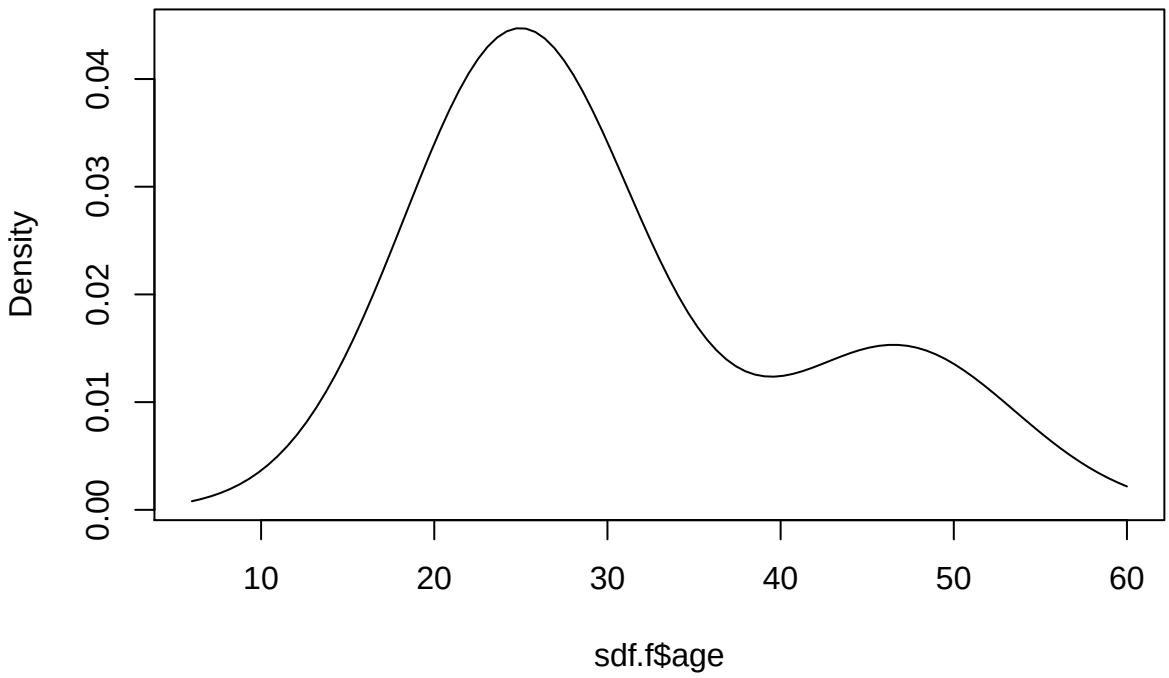
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -373.6992 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
```

With the subset data by gender, we now start to see differences in the means of the formed clusters. Both of the male clusters are below the age of 30 (20.3 and 27.7) while the females have clusters almost twice their ages (24.9 and 46.9).

```
## [1] "Male Density"
```




```
## [1] "Female Density"
```



The density plots of the *Mclust()* model start to tell more of the story in differences between the genders' ages. The male plot has a strong peak around the age of 20. The female plot also has a peak around 25 or so but its more broad, and then there is another peak around 45-50 that is smaller but definitive.

From all of the above, it is fairly evident that both males and females have an optimal cluster number of 2. These cluster are extremely different in the distances from each other within males and females, which relates to there being differences in ages between the genders. Males have 2 clusters between 20 and 30 while females have clusters at 25 and 47. The clusters within the male overvations are more difficult to tell the difference between the 2 as shown in the male density plot (there is a minor bump around 30 that makes the plot appear unnatural {difficult to describe}) while the female plot shows 2 different peaks. Because of these results, I would say that there is evidence that there are potentially different types of schizophrenia that each gender is more prone too, but not found only in an individual gender.

Resources:

- StackExchange
- ggplot2.tidyverse.org (contour plot)