Homework 4

STAT 601

Instructions

Discuss how your results address each question.

Please use compile to HTML when sharing your results on the message board.

This file can be used as a template.

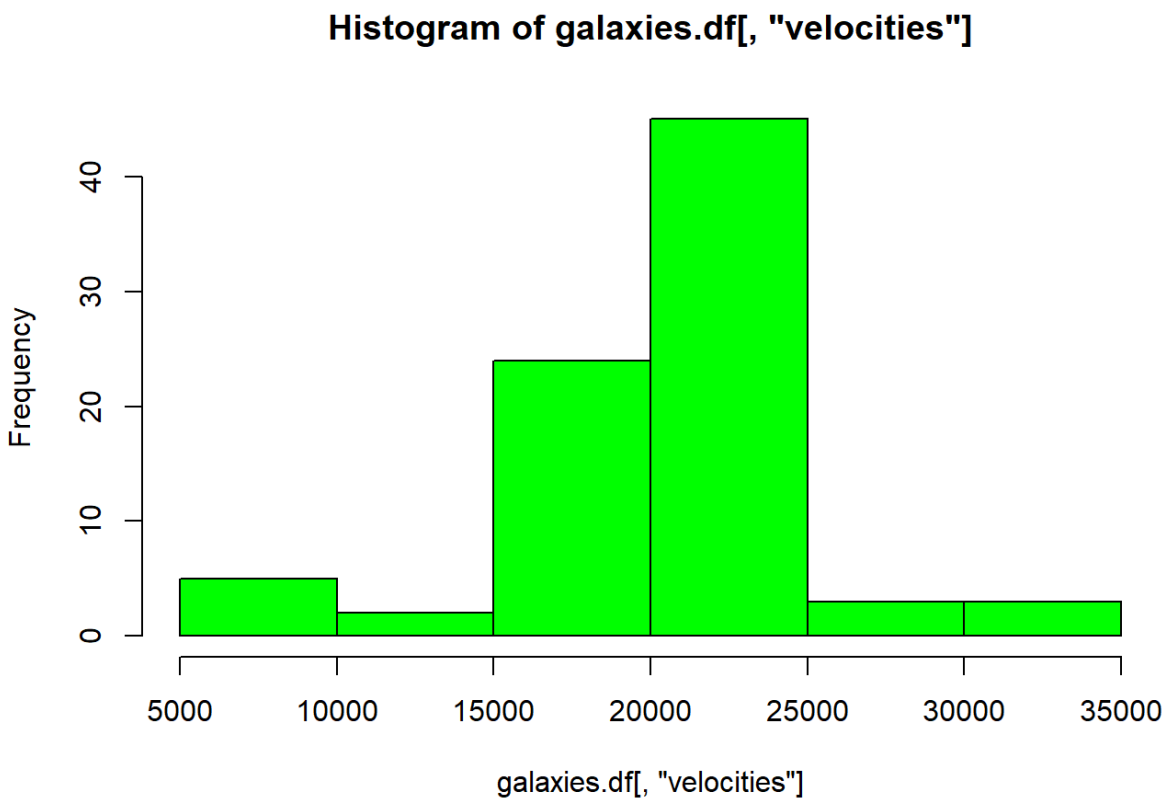
Exercises

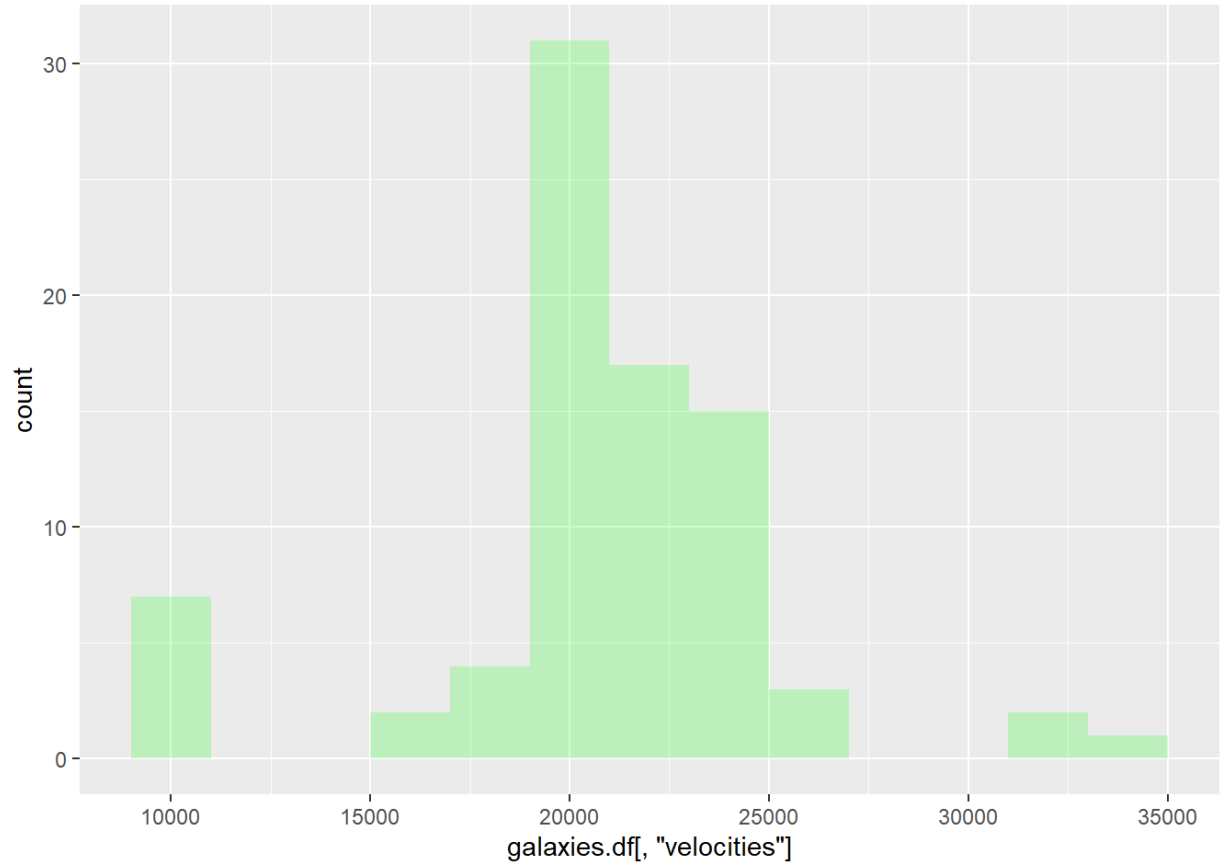
Warning: There are only three questions, however they will require more time coding. You may need to review function calling conventions and whether the optional arguments and their default parameters are appropriate.

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)

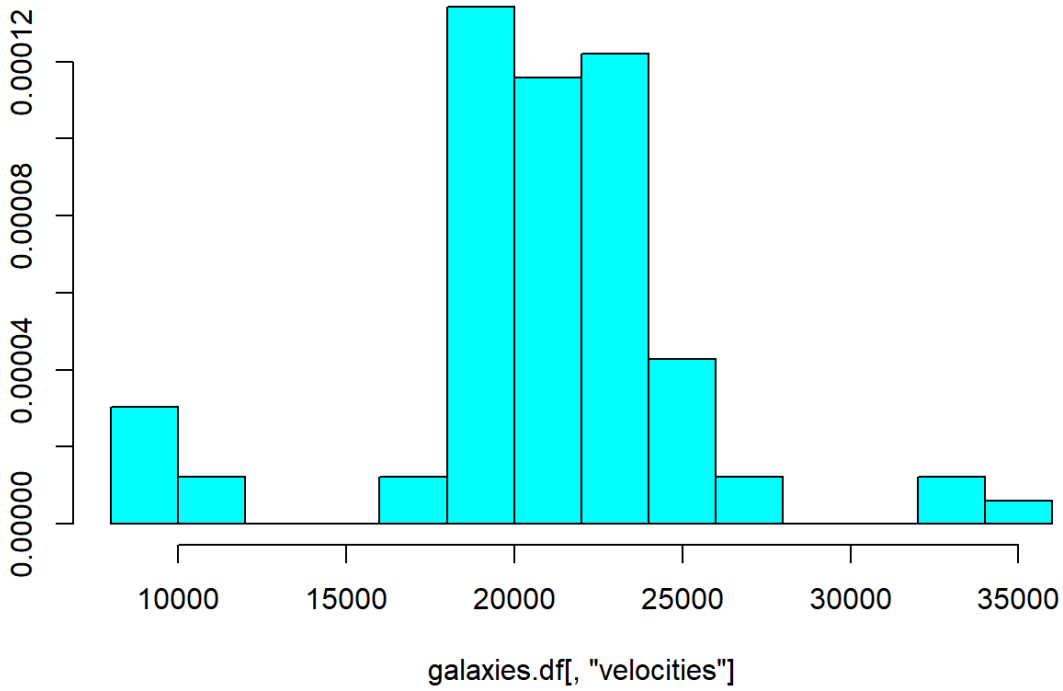
    a. Construct histograms using the following functions:

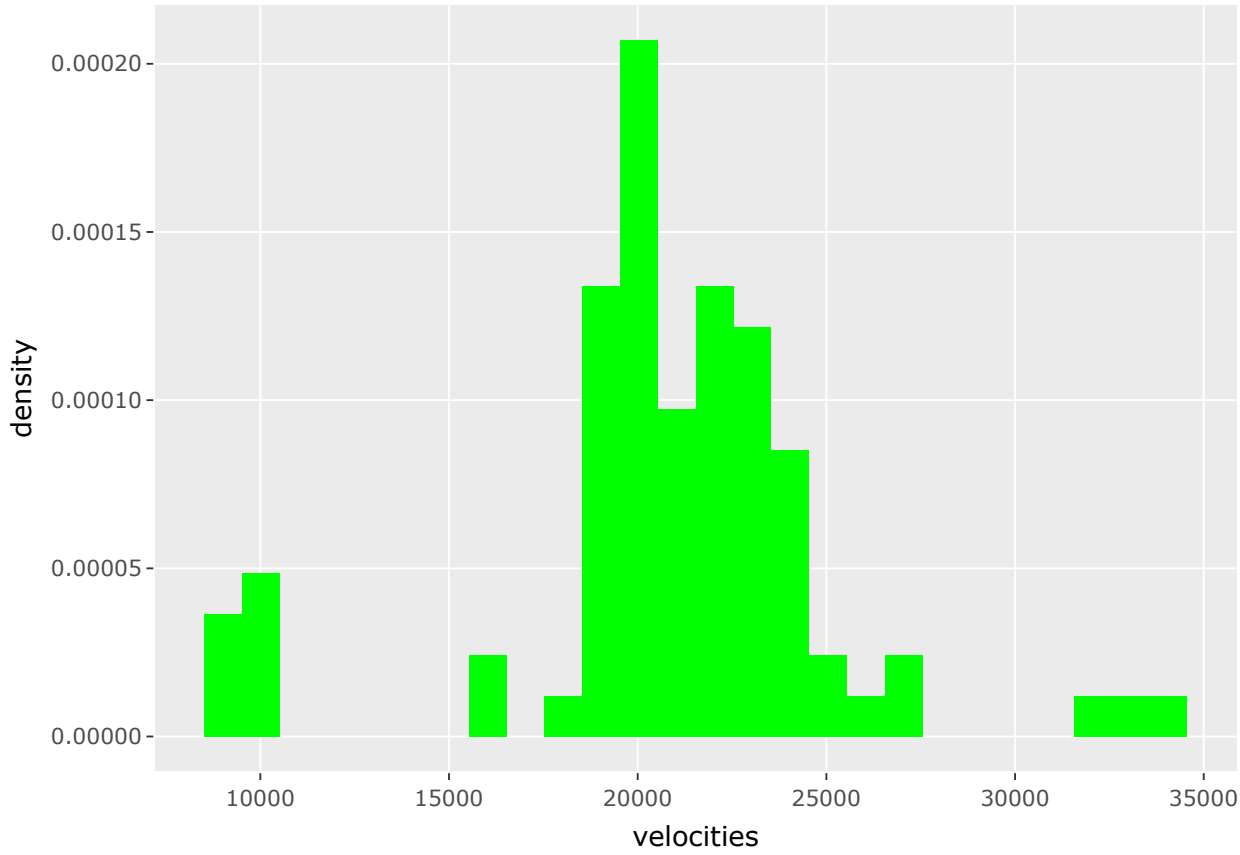
    -hist() and ggplot()+geom_histogram()



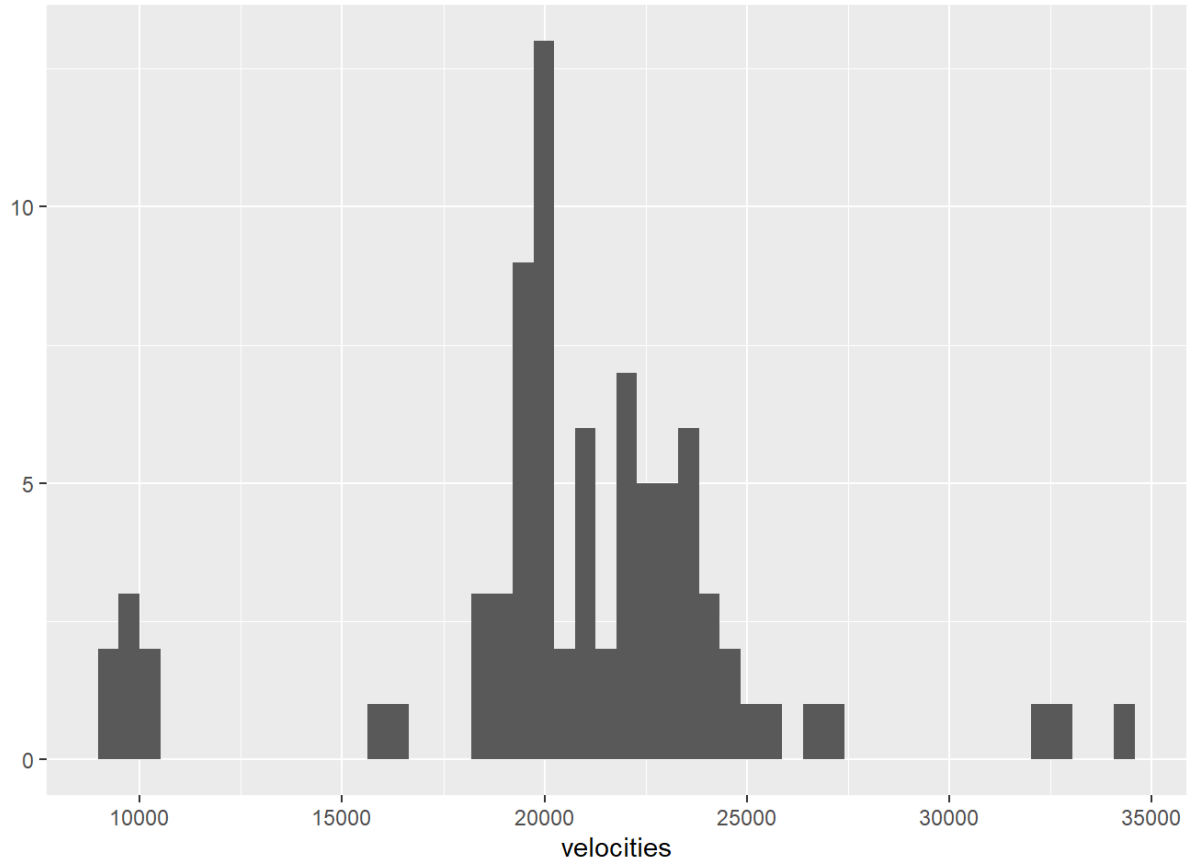


-truehist() and ggplot+geom_histogram() (make sure that the histograms show proportions, not counts.)


0.00000
0.00004
0.00008
0.00012
10000
15000
20000
25000
30000
35000
galaxies.df[, "velocities"]


0.00020
0.00015
0.00010
0.00005
0.00000
density
10000
15000
20000
25000
30000
35000
velocities

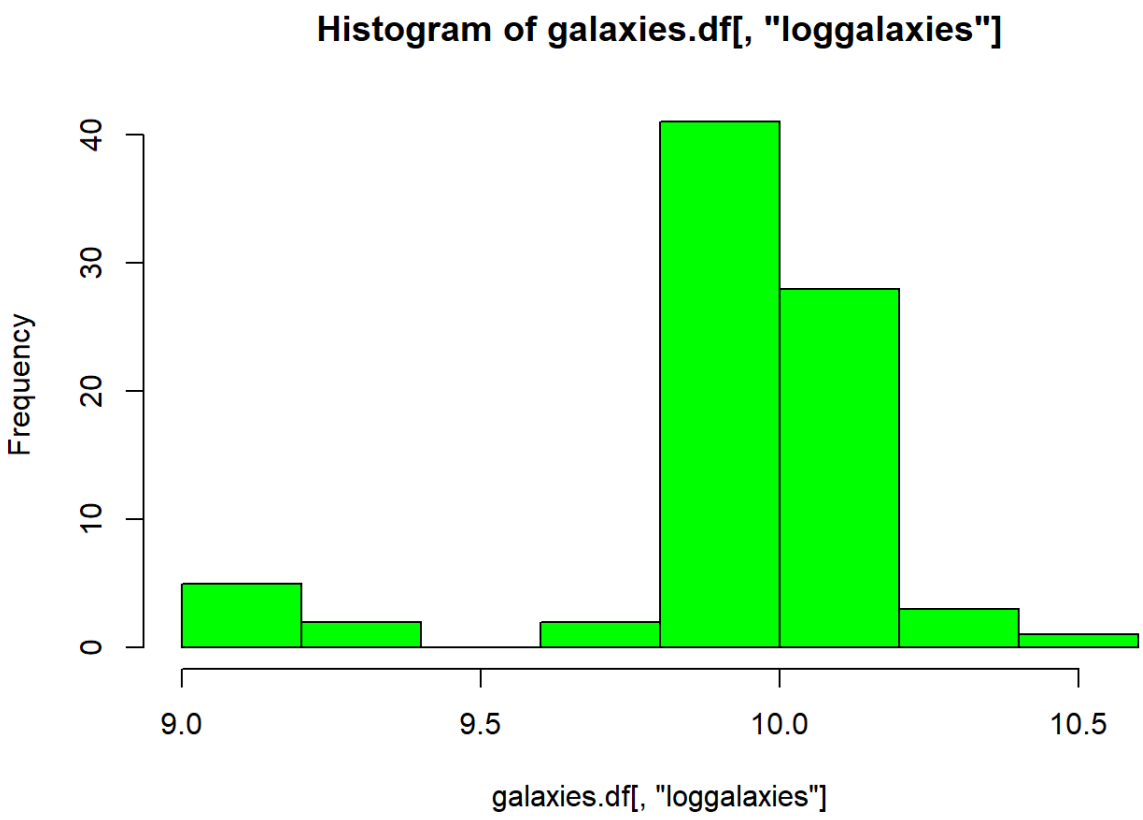
-qplot()

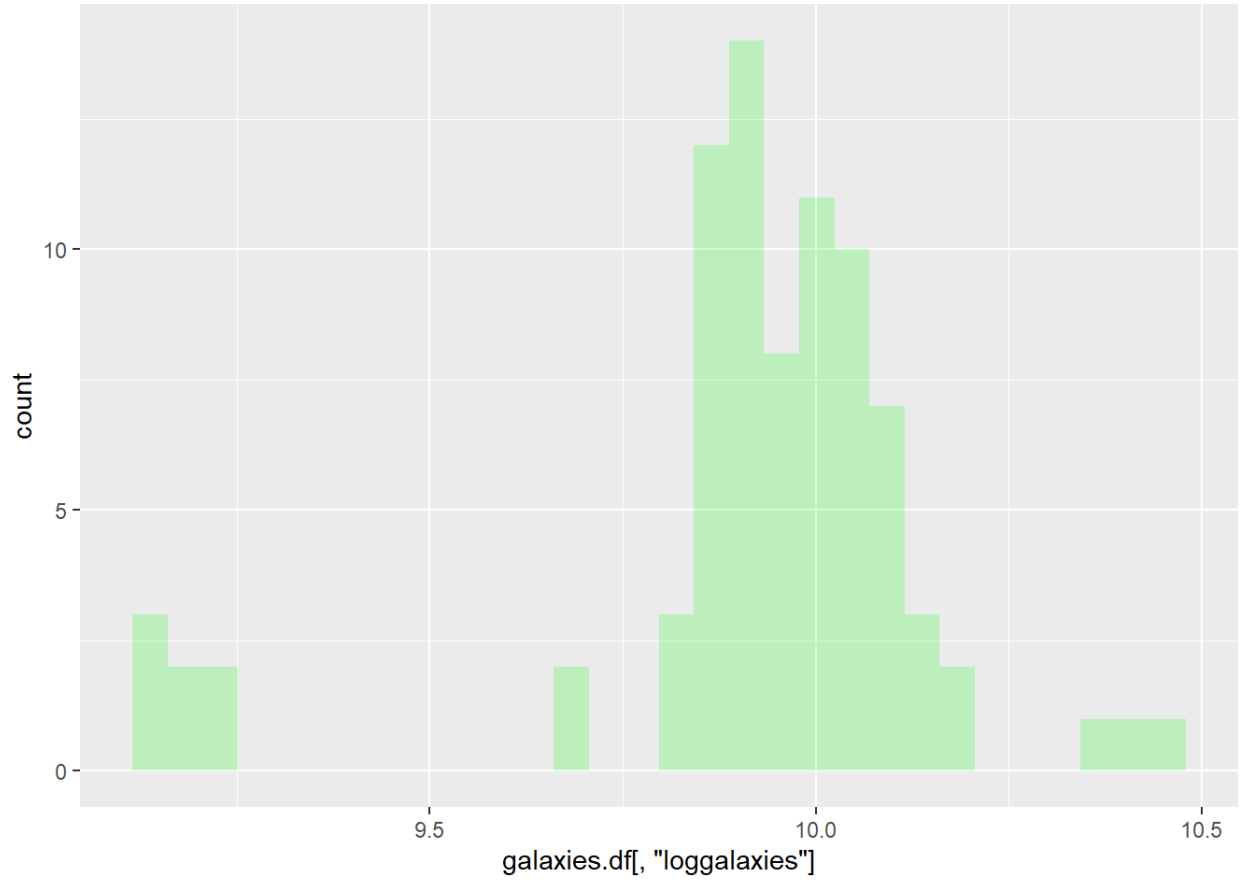


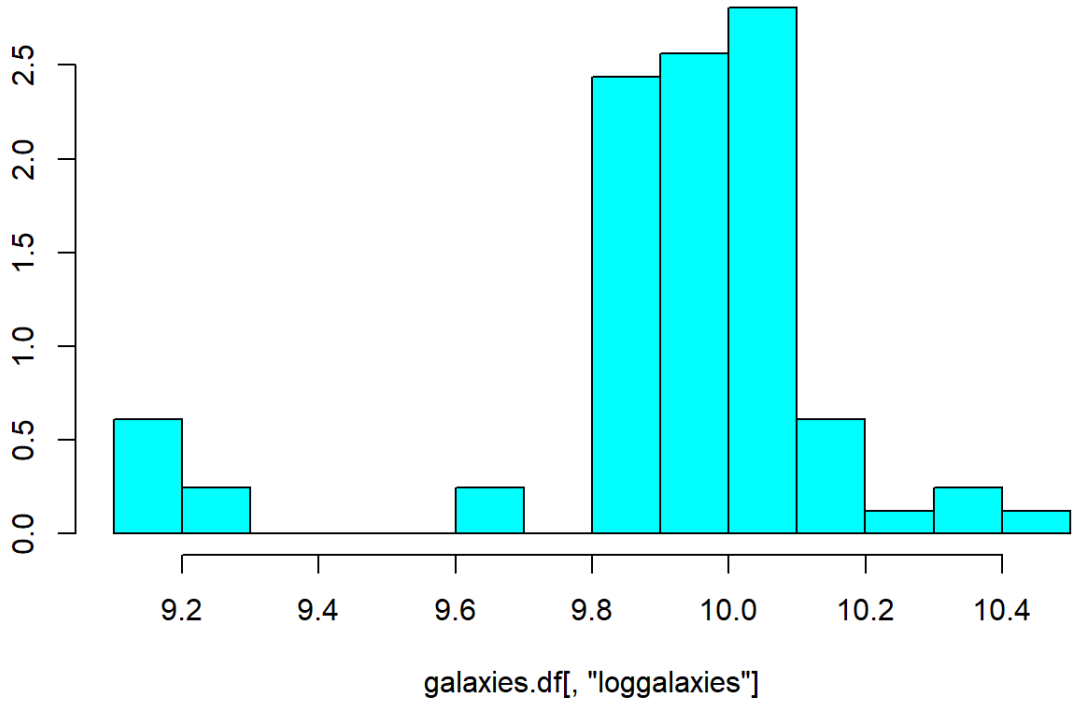
Comment on the shape and properties of the variable based on the five plots. Do you notice any sets of observations clustering? (Hint: You can adjust bin number or bin size as you try to determine the properties of the variable, but use the same bin settings between plots in your final analysis. You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

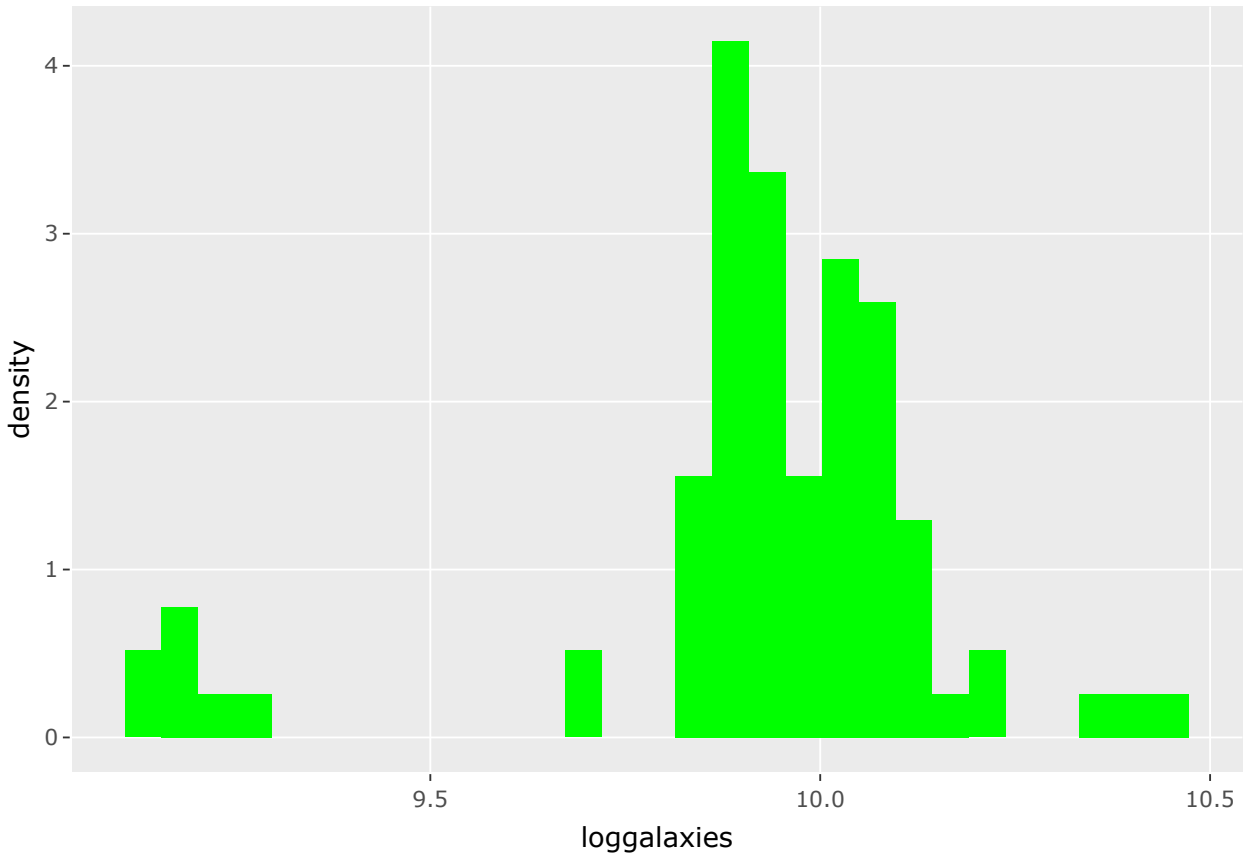
Going by the plots, most of the data seems to be centered around 19000 and 24000 velocity with the zenith being at 20000. There are a couple of other minor clusters at around 9000 to 10000 and 32000 to 33000.

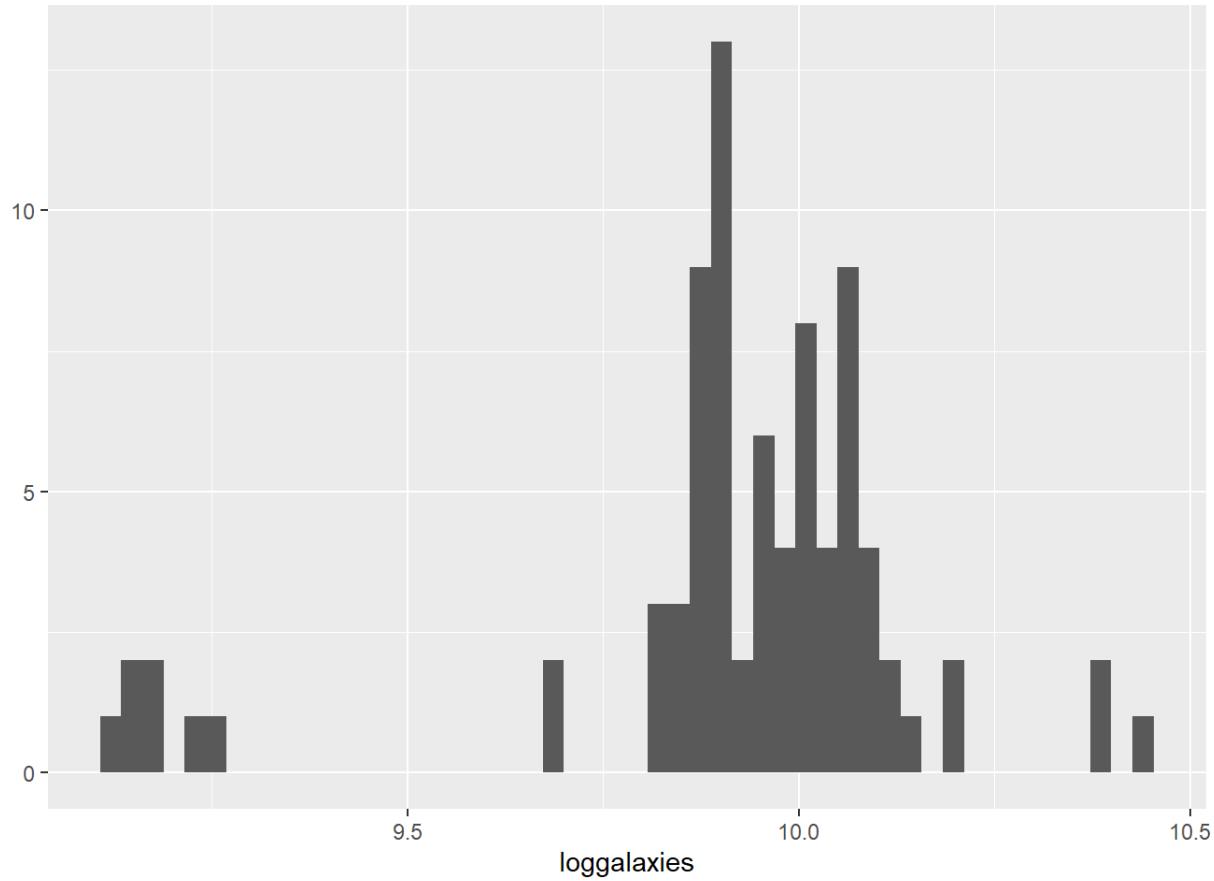
```
b) Create a new variable \textit{loggalaxies} = $\log$(galaxies). Repeat part a) using the `loggalax
ies` variable. Does this affect your interpretation of the graphs?
```


Histogram of galaxies.df[, "loggalaxies"]
Frequency
0
10
20
30
40
9.0
9.5
10.0
10.5
galaxies.df[, "loggalaxies"]


count
0
5
10
9.5
10.0
10.5
galaxies.df[, "loggalaxies"]

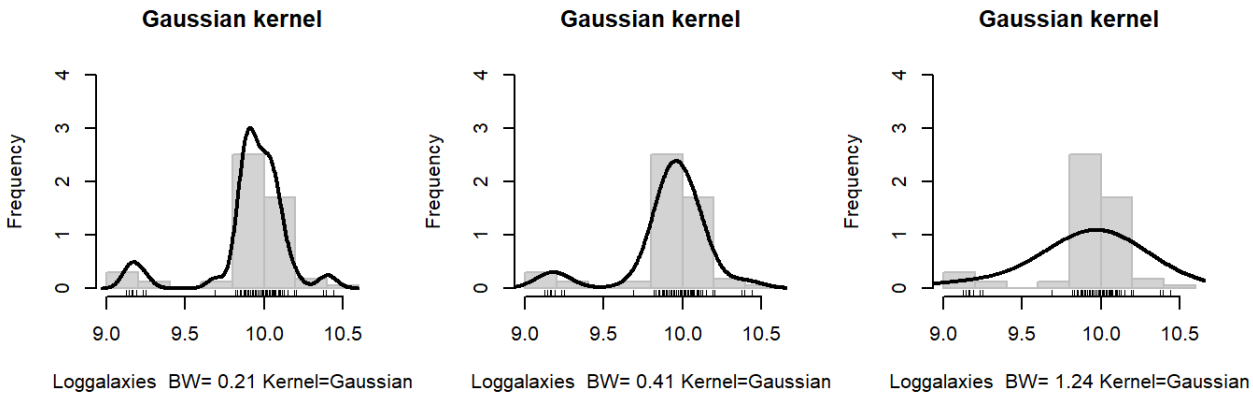

0.0
0.5
1.0
1.5
2.0
2.5
9.2
9.4
9.6
9.8
10.0
10.2
10.4
galaxies.df[, "loggalaxies"]

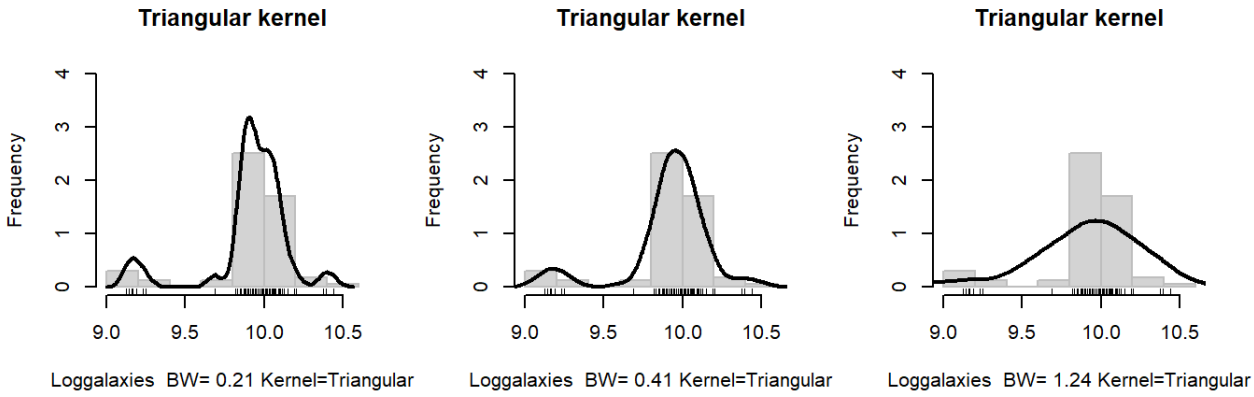

density
0
1
2
3
4
9.5
10.0
10.5
loggalaxies



No, my interpretation is the same as the log transformation seems to have just lowered the scale of the data. The distribution is around the same.

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.





My results

for the kernel density estimates were as expected. I used bandwidth formula to get a baseline bandwidth and then divided that by 2 as well as multiplied it by 3 to get 2 new plots that under- and over-smooth the graph.

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

The results of the galaxies survey seem to allude to the possibility of a supercluster. I came to this conclusion as my plot results reveal 3 clusters, 1 much larger than the other 2. According to this source(https://imagine.gsfc.nasa.gov /features/cosmic/nearest_superclusters_info.html (https://imagine.gsfc.nasa.gov/features/cosmic /nearest_superclusters_info.html)) superclusters contain chains of at least a dozen galaxies, while one of the clusters of galaxies I found in the data contains around 30 galaxies that are around the same velocity.

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 X (univariate normal) model with 1 component:
##
##  log-likelihood  n df       BIC       ICL
##       -806.7738 82  2 -1622.361 -1622.361
##
## Clustering table:
##  1
## 82
```

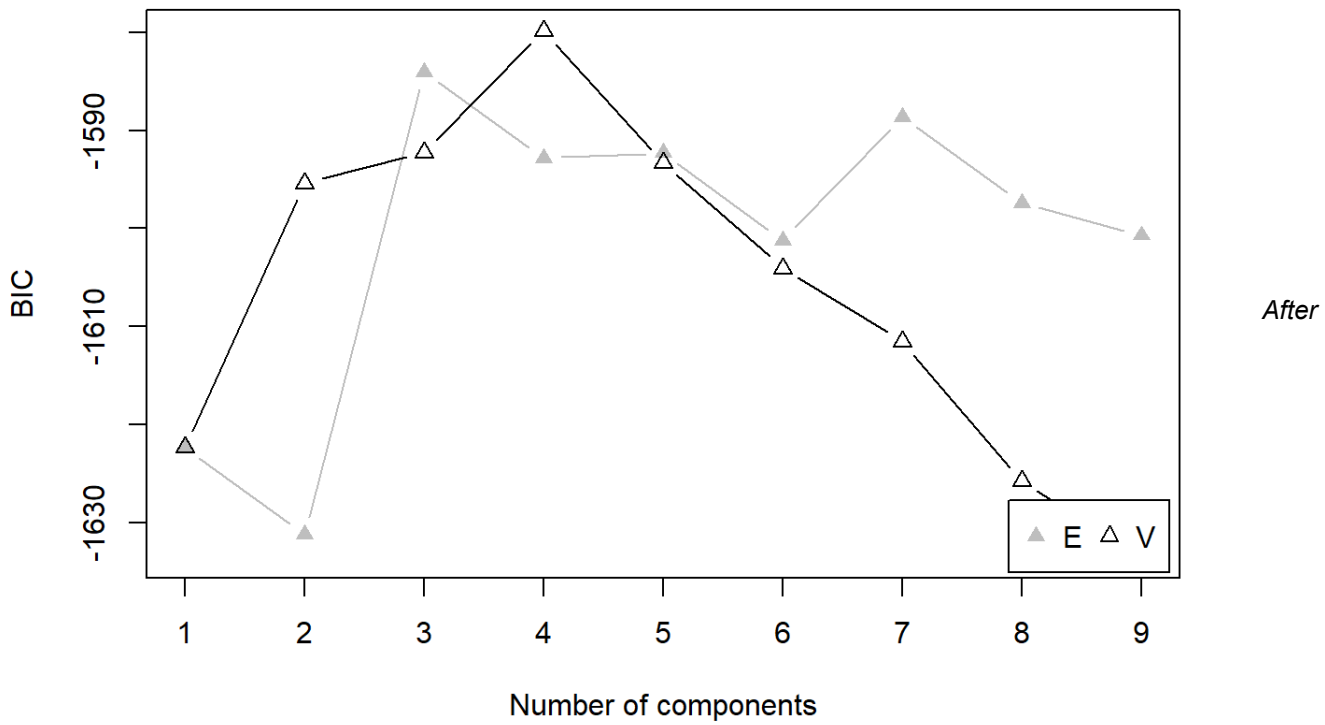
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -786.6847 82  5 -1595.403 -1615.064
##
## Clustering table:
##  1  2
## 14 68
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 3 components:
##
##  log-likelihood  n df       BIC       ICL
##       -778.7878 82  6 -1584.016 -1584.178
##
## Clustering table:
##  1  2  3
##  7 72  3
```

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

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 5 components:
##
##  log-likelihood  n df       BIC       ICL
##       -774.1157 82 10 -1592.299 -1655.533
##
## Clustering table:
##  1  2  3  4  5
##  7 24 16 32  3
```

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



After testing with multiple models, the best results seem to be at 4 clusters. This is different than what I predicted previously from the plots, as it seems to have split the large cluster I mentioned earlier into 2. log-likelihood: -765.694; BIC: -1579.862; ICL: -1598.907 Paramater estimate: cluster 1 / mean = 9707.5 Cluster 2 / mean = 19804.3 Cluster 3 / mean

= 22879.5 Cluster 4 / mean = 24459.5

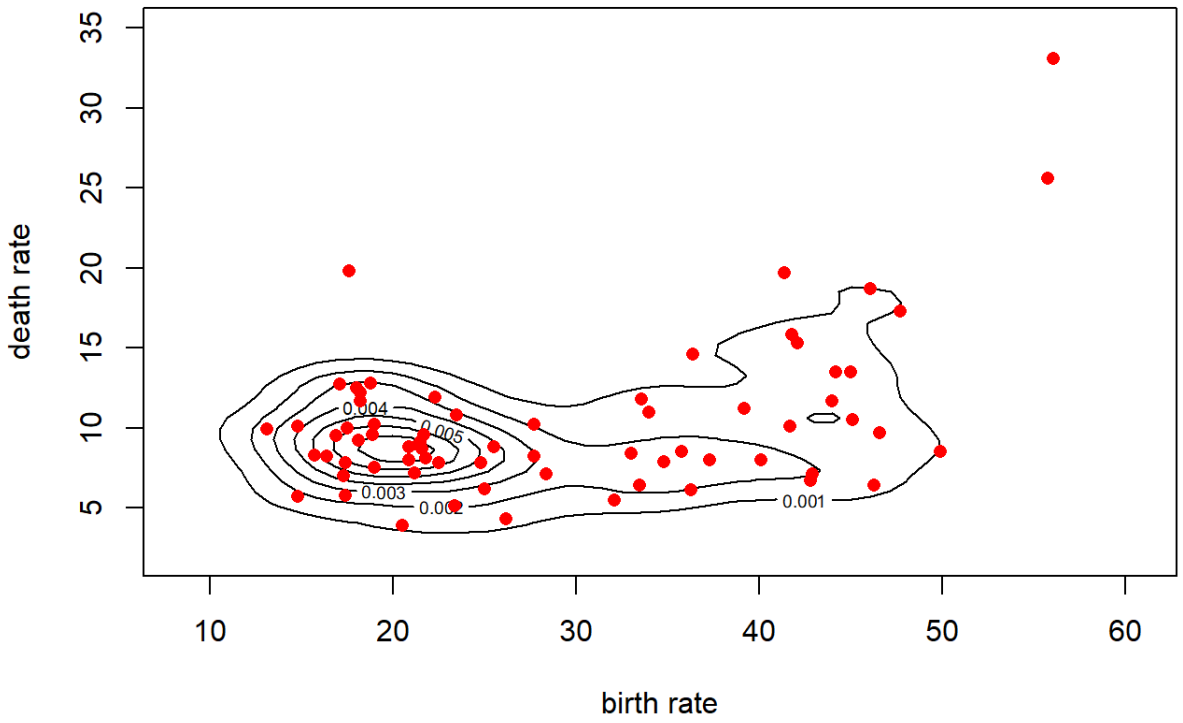
2. (Ex. 8.2 in HSAUR, modified for clarity) The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (from Hartigan, 1975).

	birth <dbl>	**death** <dbl>
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

6 rows

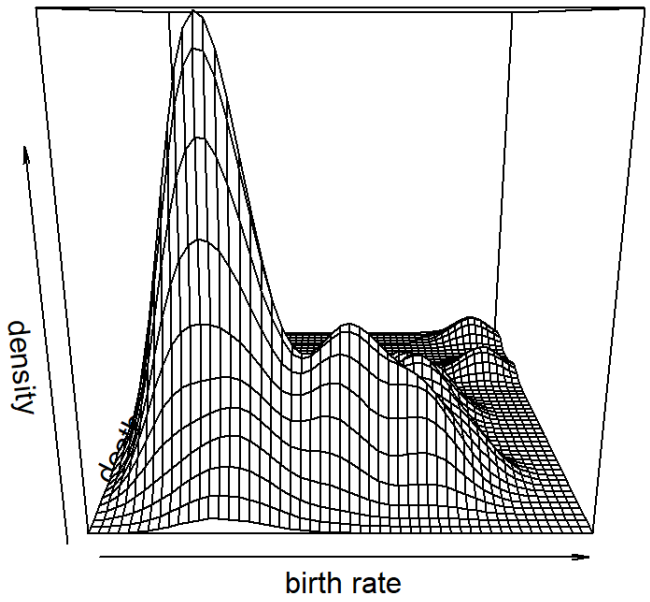
```
##      birth           death       
##  Min.   :13.10   Min.   : 3.90  
##  1st Qu.:18.90   1st Qu.: 7.80  
##  Median :25.00   Median : 9.10  
##  Mean   :29.25   Mean   :10.31  
##  3rd Qu.:40.10   3rd Qu.:11.70  
##  Max.   :56.10   Max.   :33.10
```

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



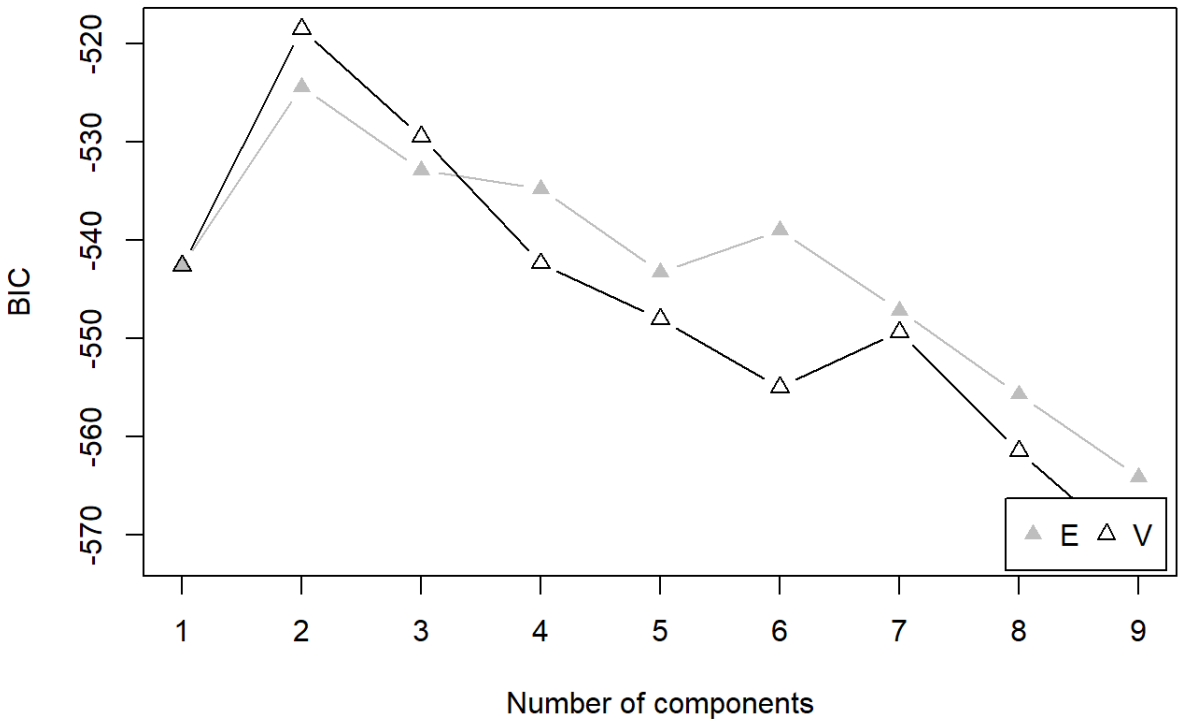
b) What does the contour plot tell you about the structure of the data?
The contour plot shows that the data is clustered at a birth rate of around 20 and a death rate of around 9. Furthermore, it shows that the bulk of the birth rates in these 69 countries have been between 13 and 50. The bulk of the death rates have been between 4 and 16.

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

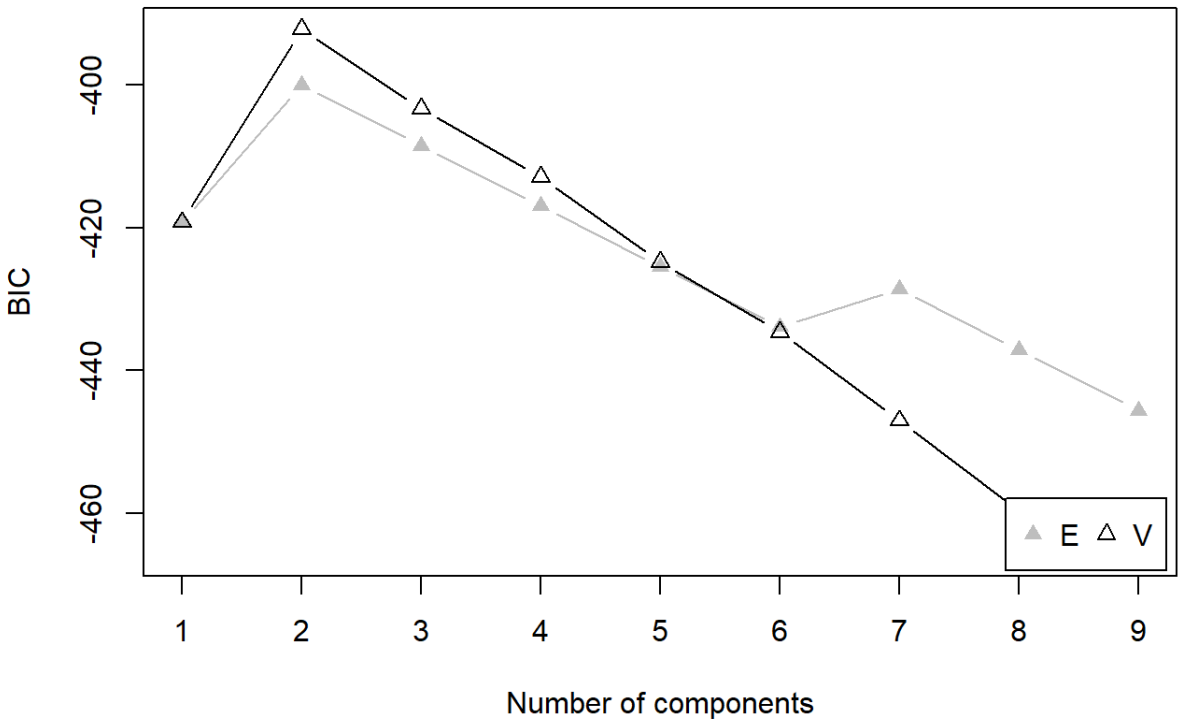


d) Fit a finite mixture model using the Mclust() function in R (from the mclust library). Summarize this model using BIC, classification, uncertainty, and/or density plots.

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -248.6773 69  5 -518.5251 -524.2174
##
## Clustering table:
##  1  2
## 37 32
##
## Mixing probabilities:
##        1        2
## 0.518881 0.481119
##
## Means:
##        1        2
## 19.64140 39.61427
##
## Variances:
##        1        2
## 10.25746 61.76977
```

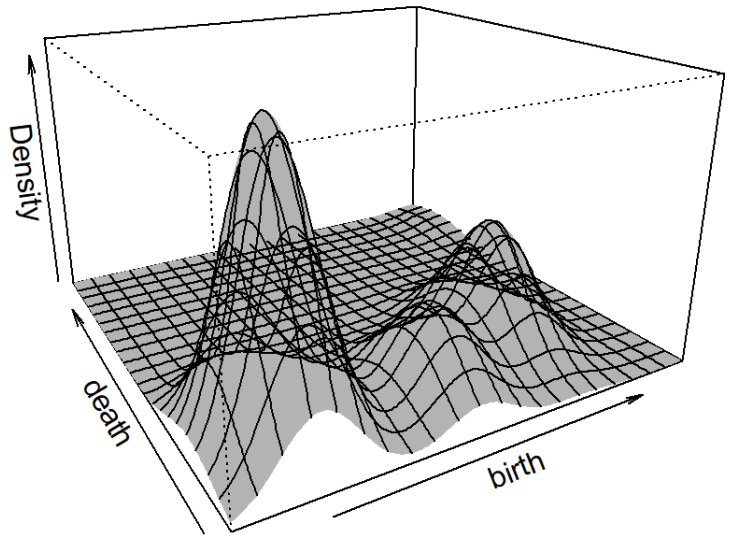


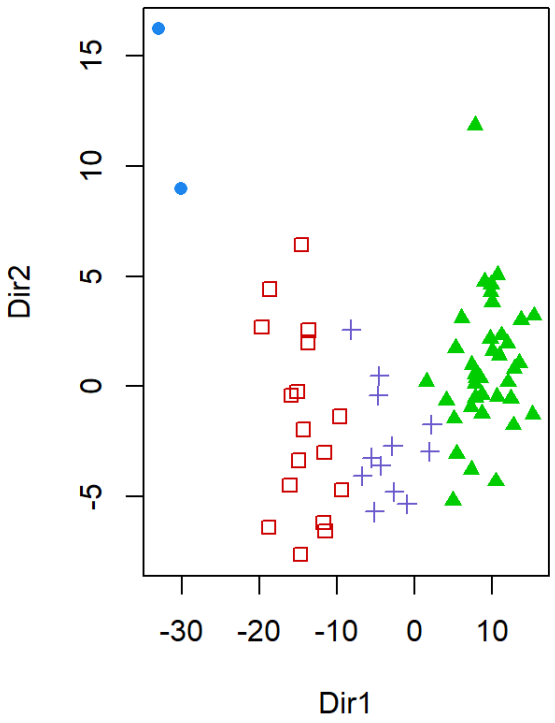

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -185.4814 69  5 -392.1333 -402.2367
##
## Clustering table:
##  1  2
## 60  9
##
## Mixing probabilities:
##       1       2
## 0.82251 0.17749
##
## Means:
##         1         2
##  8.836846 17.153921
##
## Variances:
##         1         2
##  5.145727 46.174184
```

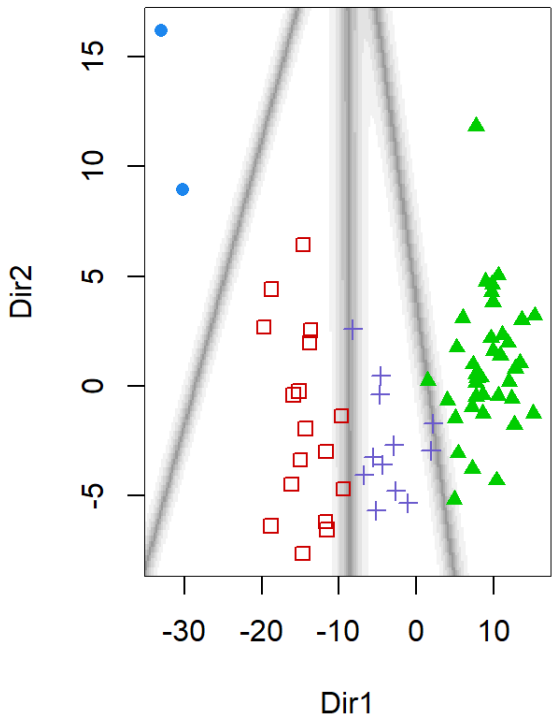



```
## ------------------------------------------------------------
## Resampling standard errors
## ------------------------------------------------------------
## Model                       = V
## Num. of mixture components = 2
## Replications                = 999
## Type                        = nonparametric bootstrap
##
## Mixing probabilities:
##          1          2
## 0.07688238 0.07688238
##
## Means:
##         1         2
## 0.8423068 1.9385160
##
## Variances:
##         1         2
##  3.787397 20.784062
```

```
## ------------------------------------------------------
## Density estimation via Gaussian finite mixture modeling
## ------------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood  n df       BIC       ICL
##       -424.4194 69 12 -899.6481 -906.4841
```








```
e) Discuss the results in the context of Birth and Death Rates.
*For the plots above, I used BIC, Line, Density, and Scatter plots, mainly to test out the demonstra
tions from the MCluster documentaiton. Regarding clustered data,the models and BIC chart show 2 clus
ters for the birth rates (means being at 19.64 and 39.61) and the death rates (means being at 8.84 a
nd 17.15). The density plot and the scatter plot show 4 clusters for the countries' birth and death
rates together. This could mean that conditions (health, living, polution, etc.) are similar in thes
e countries where the birth, death, and birth-death combination rates are clustered.*
```

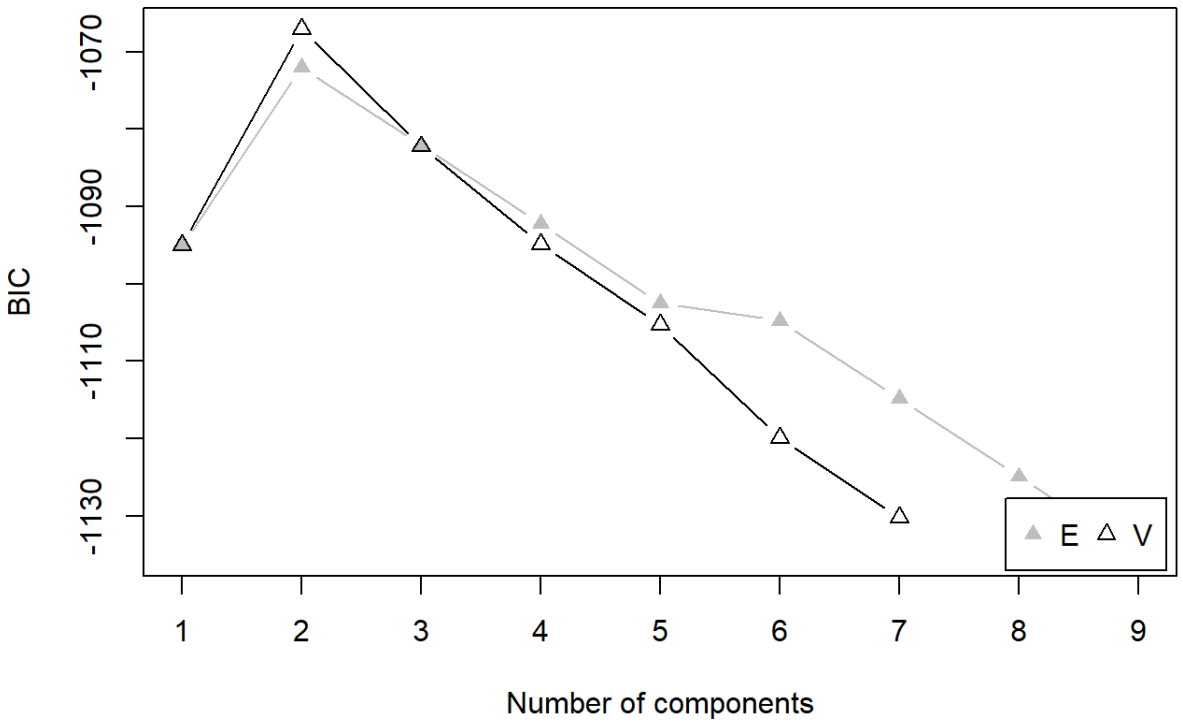
3. (Ex. 8.3 in HSAUR, modified for clarity) Fit finite mixtures of normal densities individually for each gender in the **schizophrenia** data set from **HSAUR3**. Do your models support the *sub-type model* described in the R Documentation?

	age	gender
	<dbl>	<fct>
1	20	female
2	30	female
3	21	female
4	23	female
5	30	female
6	25	female
7	13	female
8	19	female

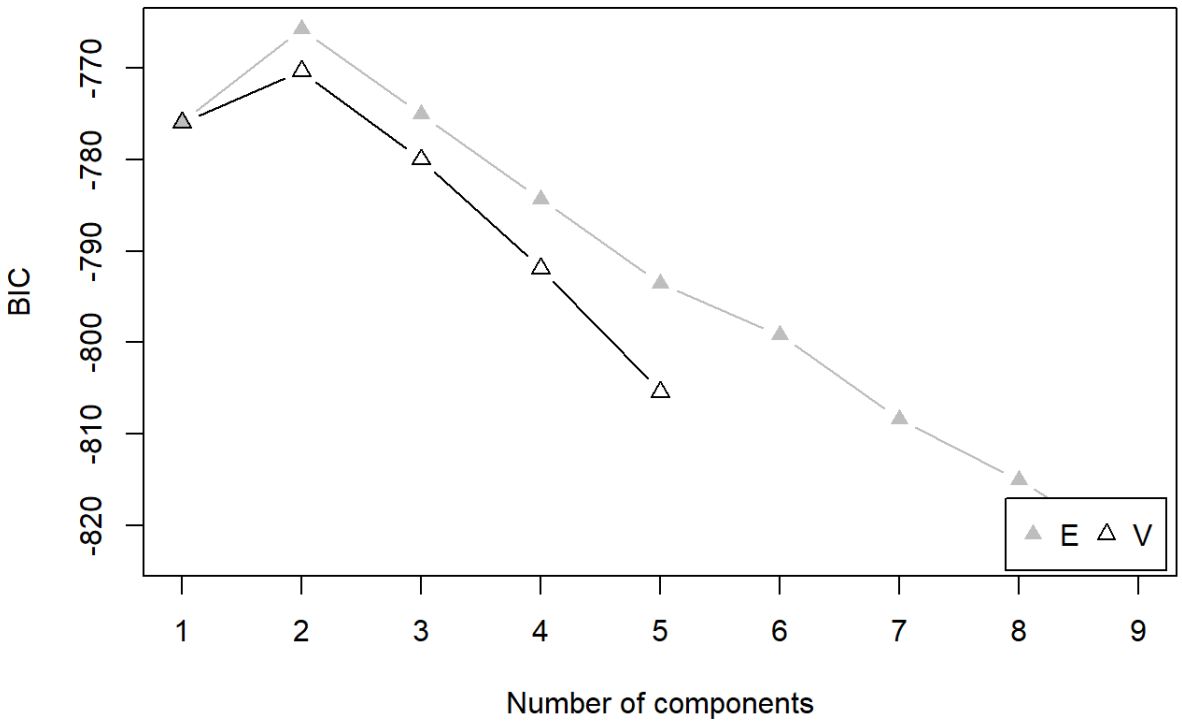
	age	gender
	<dbl>	<fct>
9	16	female
10	25	female

1-10 of 251 rows Previous **1** 2 3 4 5 6 ... 26 Next

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




```
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 earl
ier onset in men than in women. One model that has been suggested to explain this observed differenc
e is known as the subtype model which postulates two types of schizophrenia, one characterized by ea
rly onset, typical symptoms and poor premorbid competence; and the other by late onset, atypical sym
ptoms 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)
```

My results from the use of mclust for finite normal mixure modeling on the schizophrenia data were that there were 2 clusters for male and female schizophrenia patients. For males, the means of the 2 clusters were 20 and 28 years old, while for females, the means of the 2 clusters were at 25 and 47 years old. This outcome suggests that the sub-type model mentioned in the R-Documentation is correct and that males have more of a tendency towards early-onset schizophrenia than females.