Homework #4

Justin Robinette

September 18, 2018

No collaborators for any problem

Problem #1, Part A: The **galaxies** data from **MASS** contains the velocities of 82 galaxies from six well-separated conic sections of space (Postman et al., Roeder, 1990). The data are intended to shed light on whether or not the observable universe contains supperclusters of galaxies surrounded by large voids. The evidence for the existence of superclusters would be the multimodality of the distribution of velocities. Construct a histogram of the data and add a variety of kernel estimates of the density function. What do you conclude about the possible existence of superclusters of galaxies? (8.1 Handbook)

Construct historgrams using the following functions:
- hist() and ggplot() + geom_histogram()
- truehist() and ggplot + geom_histogram() (pay attention to the y-axis!)
- qplot()

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

Results: First, we corrected a typo in the **galaxies** dataset by changing the value of the 78th observation from *26690* to *26960*. Then, Figure 1.1 uses the **hist()** function. We see a relatively normal distribution centered around the 20,000 'velocity' mark. For similarity, I matched the number of bins and the breaks in the corresponding **ggplot**. Here we see a relatively normal distribution centered on 20,000 km/sec.

Figure 1.2 uses the **truehist()** function. We see the same basic plot, with relatively normal distribution, to what we saw in Figure 1.1. The main difference is the labeling of the y-axis ticks. Here we see the y-axis labeled with density values rather than a frequency count as we saw in Figure 1.1. The corresponding 'ggplot' has been created with the same number of bins and a y axis that labels density as opposed to count/frequency. Here we see a relatively normal distribution centered on 20,000 km/sec.

Figure 1.3 looks considerably different than the prior two examples. This figure uses the **qplot** function. First, we notice that this plot contains more bins than the prior two plots. Additionally, the distribution is not as normal as the prior two examples. In this plot, we see a multimodal distribution with an increase around 10,000 km/sec and another larger increase at 20,000 km/sec. Additionally, there are increases between 20,000 and 25,000 km/sec and again between 30,000 and 35,000 km/sec.

Figure 1.1: Histogram of Galaxies - ggplot

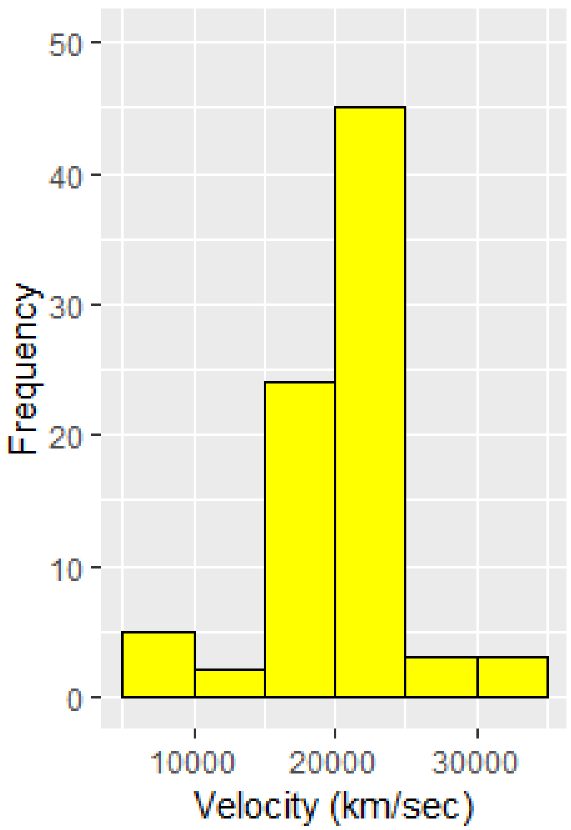
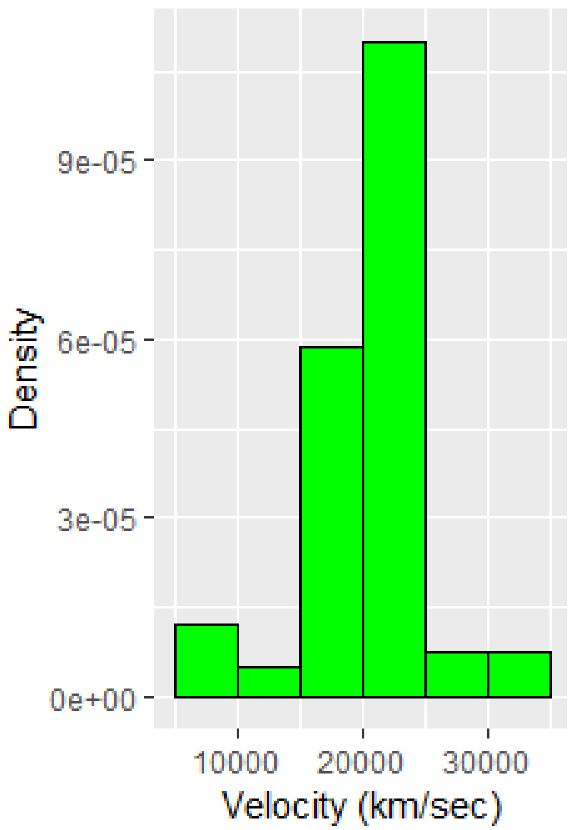
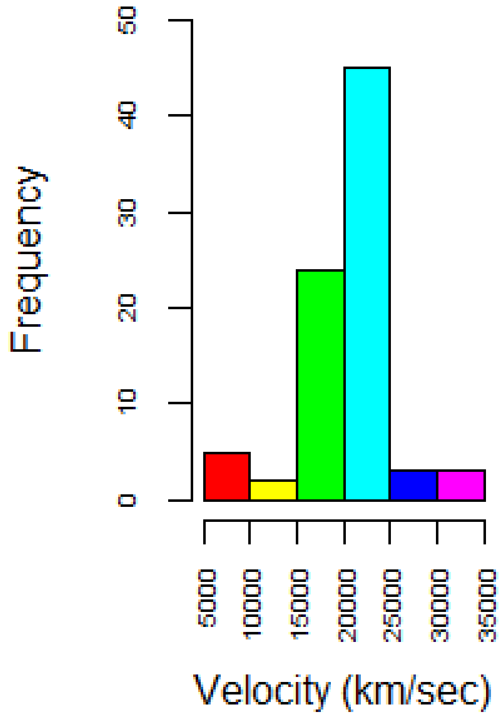


Figure 1.2: Histogram of Galaxies - ggplot



Histogram of Galaxies Base R (Fig 1.1)



Truehist of Galaxies Base R (Fig.1.2)

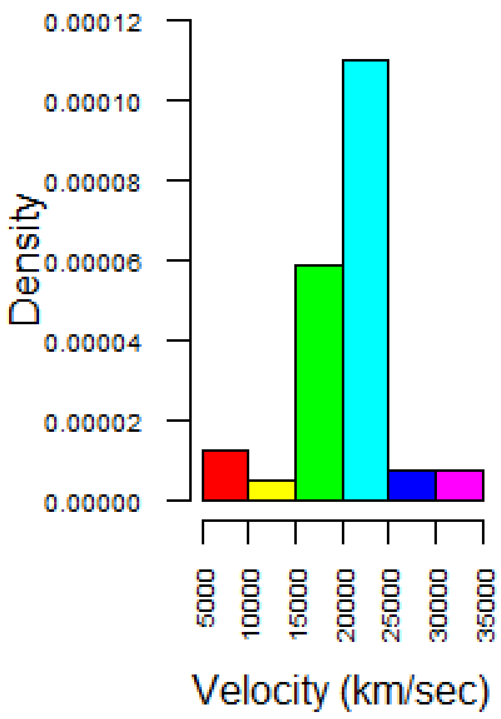
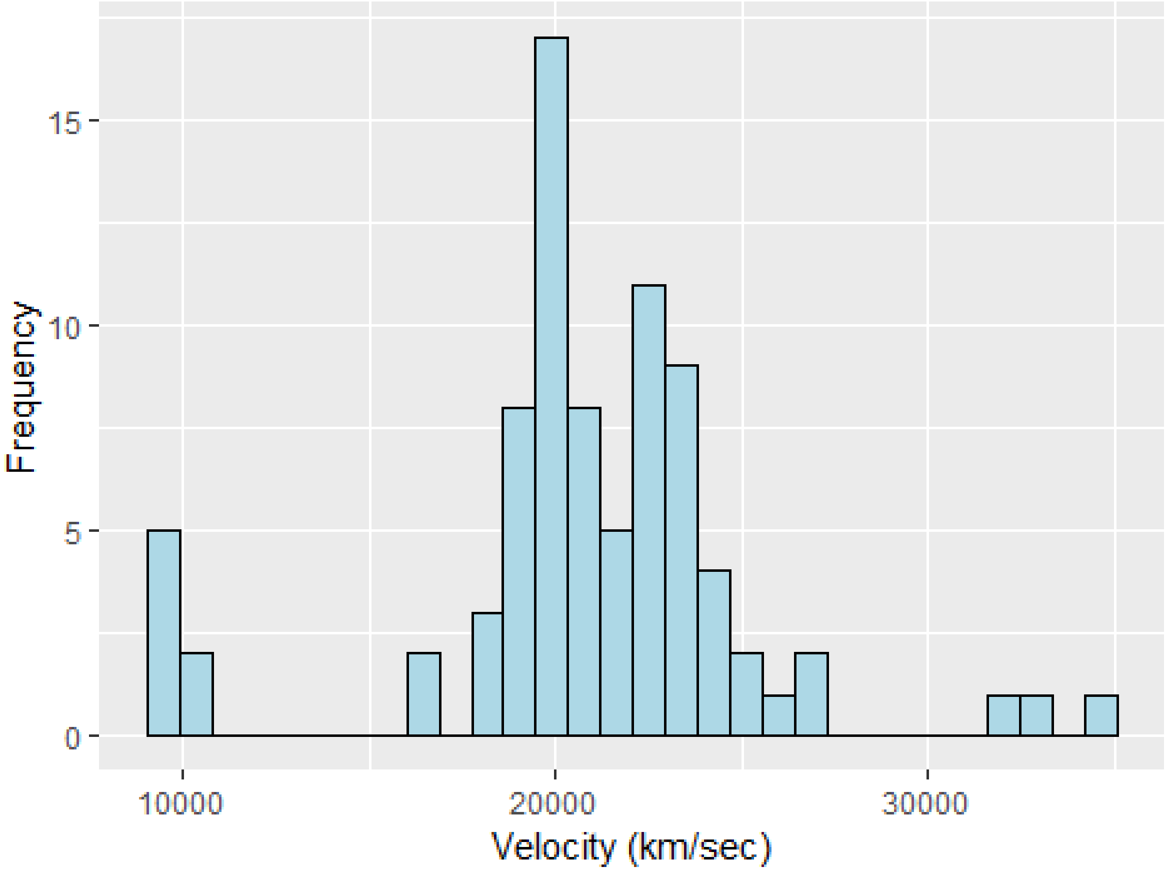


Figure 1.3: Histogram of Galaxies - qplot



Problem #1, Part B: Create a new variable *loggalaxies* = log(galaxies). Construct three histograms using the above functions and comment on the shape.

Results: Figure 1.4 uses the hist() function. We see a left-skewed distribution with 7 bins peaking around the 10 'log(velocity)' axis tick. For similarity, I set the number of bins in the corresponding ggplot to 7. Oddly, when I set it to 7 bins, the ggplot only contains 6 bins. Here we see a similar shape with the 1st and 2nd bins from the hist() plot being combined in the ggplot.

Figure 1.5 uses the **truehist()** function. We see the same basic plot to what we saw in Figure 1.1. The main difference is the labeling of the y-axis ticks. Here we see the y-axis labeled with density values rather than a frequency count as we saw in Figure 1.4. This time, there are more differences in the *truehist* plot and the *ggplot*. Despite setting 'bins = 7', in an attempt to mirror the truehist plot, the ggplot only has 6 bins and a slighty different distribution than that of the truehist plot. It appears in the ggplot that the 1st and 2nd bins, from the truehist plot, were combined into one.

Figure 1.6 looks considerably different than the prior two examples. This figure uses the **qplot** function. First, we notice that this plot contains more bins than the prior two plots. Additionally, the distribution is more similar to the prior 2 examples than it was in Part A of this exercise. In this plot, we see a multimodal distribution with an increase around 8.5 log(km/sec) and another larger increase near 9.8 log(km/sec). Additionally, there are increases at 10 log(km/sec) and again near 10.3 log(km/sec).

Figure 1.4: Histogram of Galaxies - ggplot

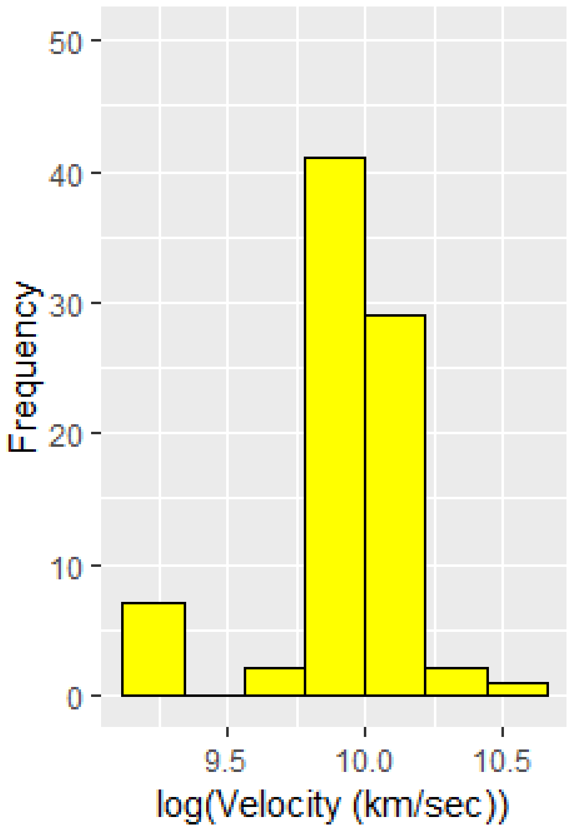
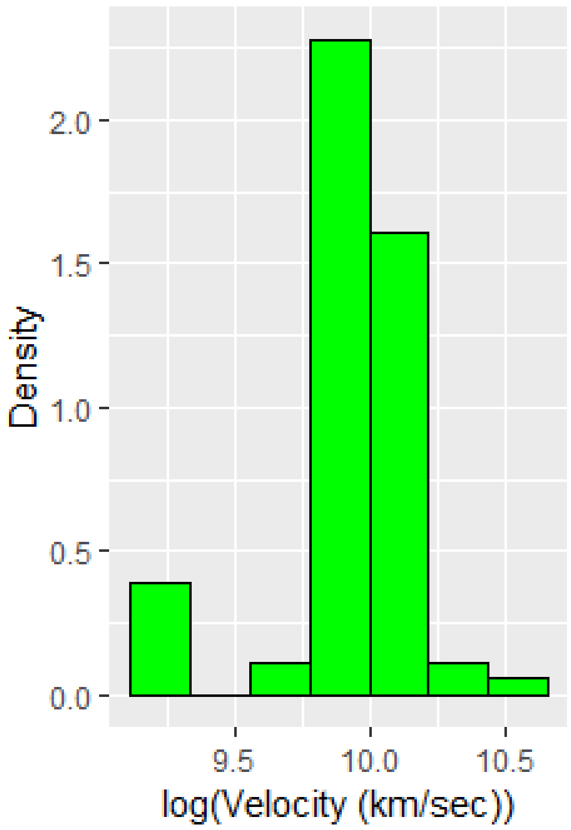
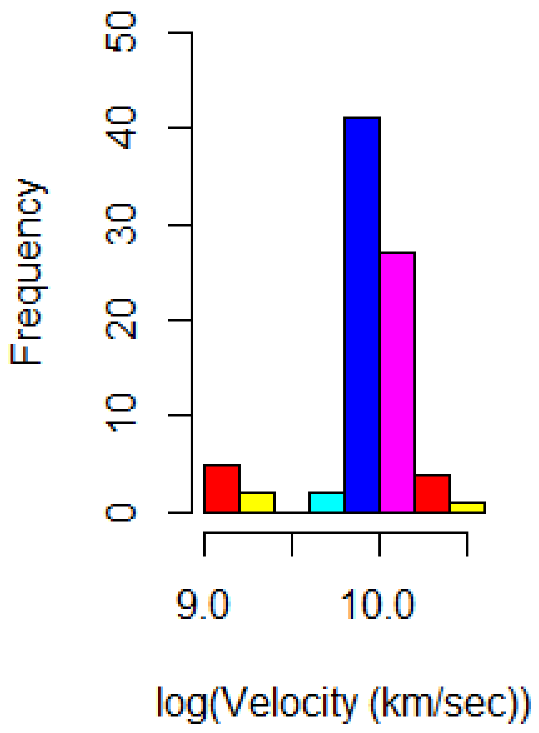


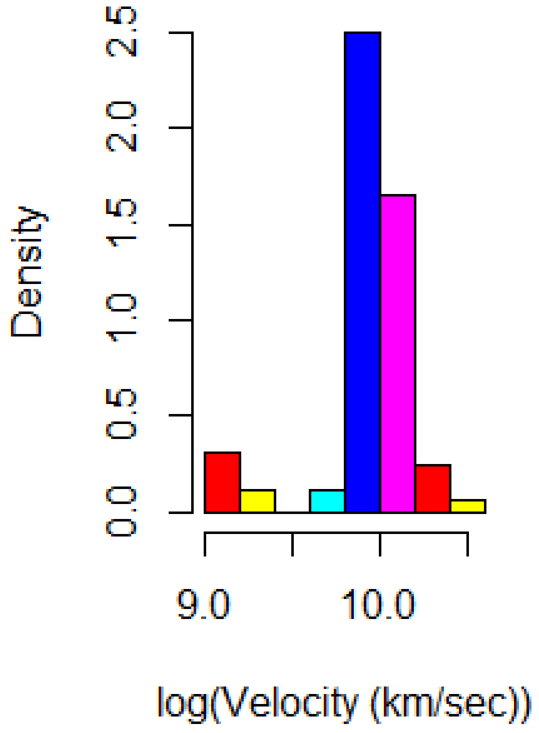
Figure 1.5: Histogram of Galaxies - ggplot



Histogram of Galaxies Base R



Truehist of Galaxies Base R



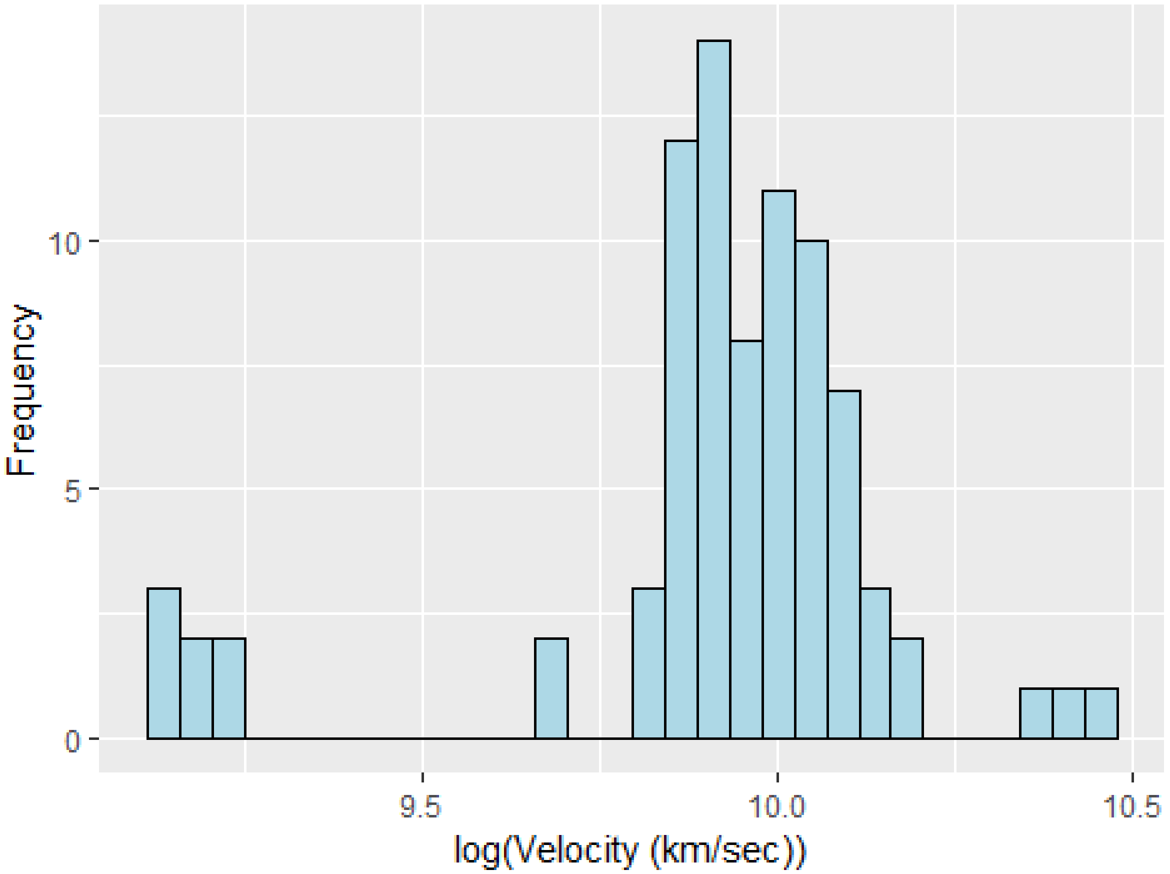


Figure 1.6: Histogram of Galaxies - qplot

Problem #1, Part C: Construct kernel density estimates using 2 different choices of kernel functions and 3 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 6 different kernel density estimates plots. Discuss your results. You can use the log scale or original scale for the variable.

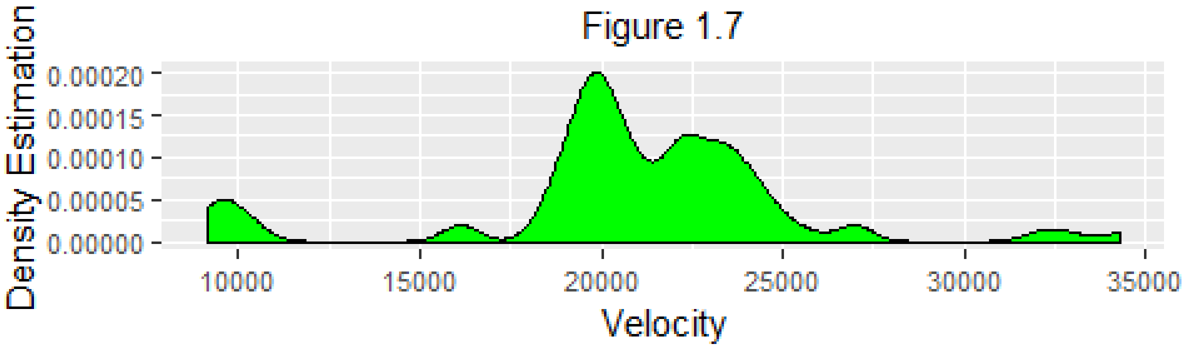
Results: I plotted the three different kernel density estimates. One that undersmooths, one that is appropriate, and one that oversmooths. To do so, I used **Silverman's Rule of Thumb** with **bw.nrd0** since Dr. Saunders mentioned in the lecture that that is his preferred method.

Each of the three estimates were done using the "Gaussian" kernel and the "Rectangular" kernel for a total of 6 plots using **ggplot**.

We can see, the oversmooth plots definitely oversimplify the distribution. With the undersmooth plots, we see a multimodal distribution that may be evidence of the existence of superclusters - as the original question discussed.

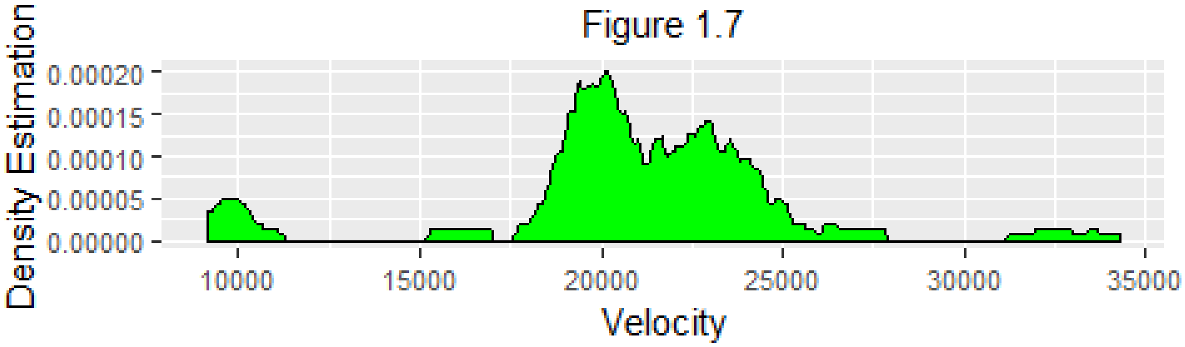
Gaussian Undersmooth
Galaxy Kernel Density

Figure 1.7



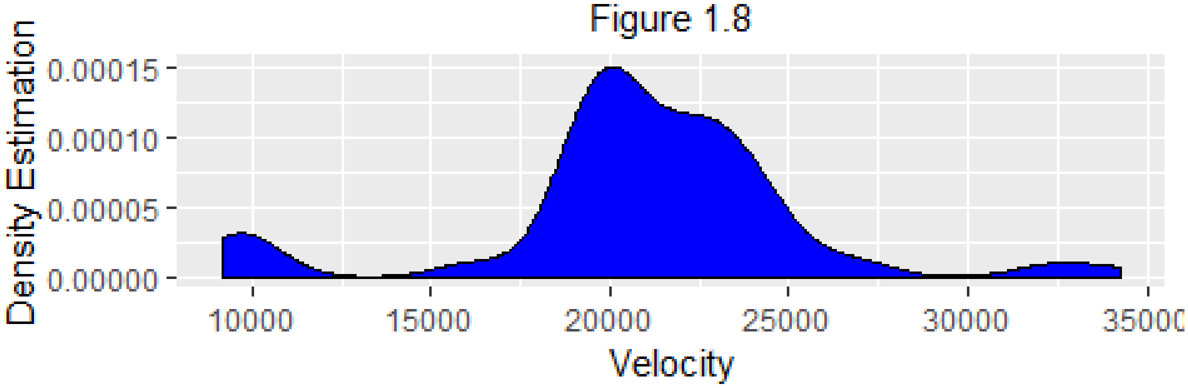
Rectangle Undersmooth
Galaxy Kernel Density

Figure 1.7



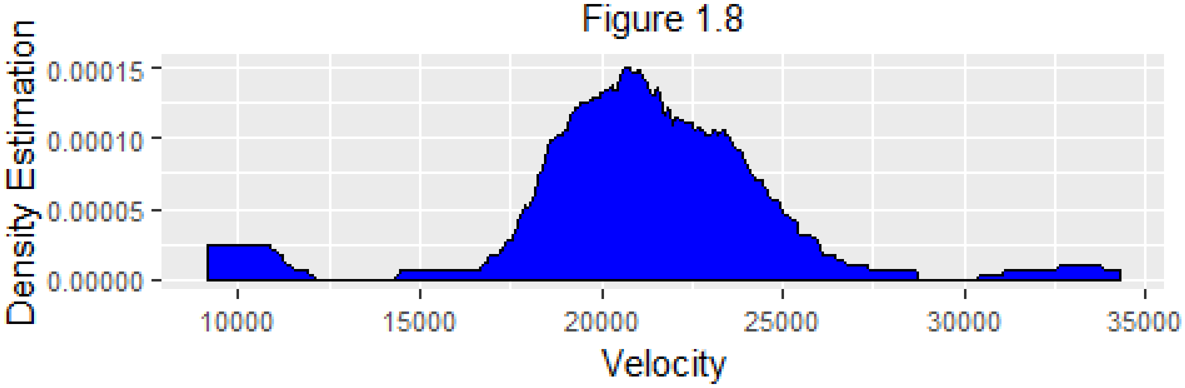
Gaussian Galaxy Kernel Density

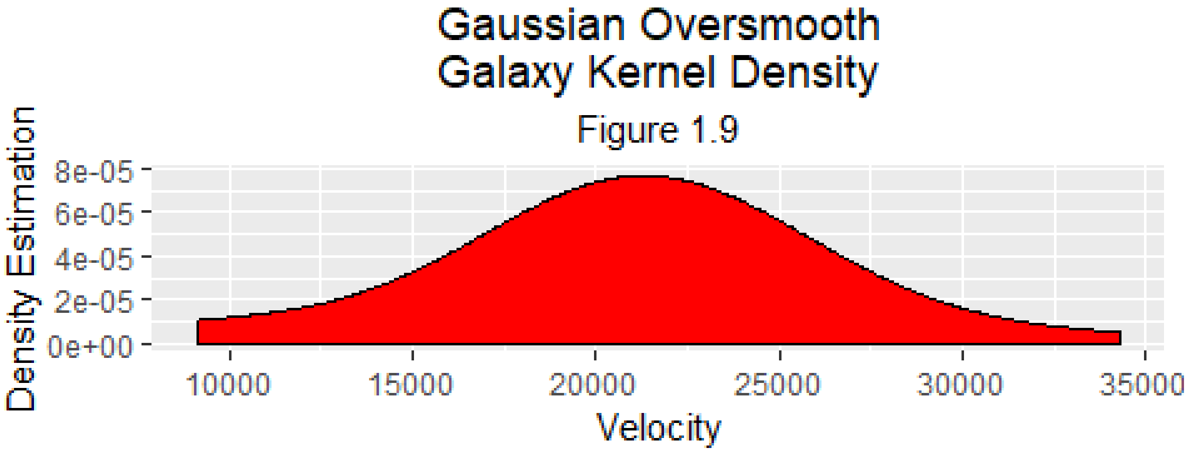
Figure 1.8

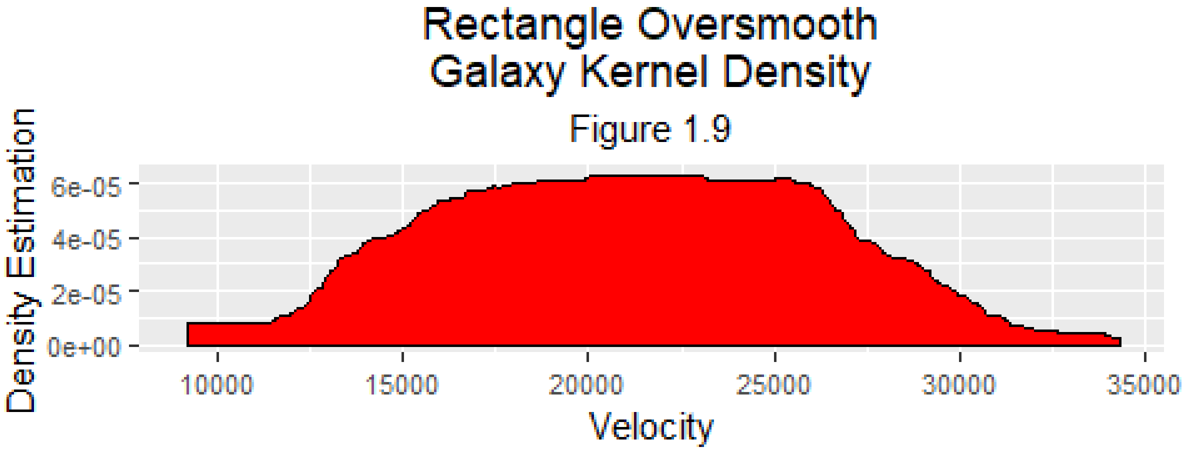


Rectangle Galaxy Kernel Density

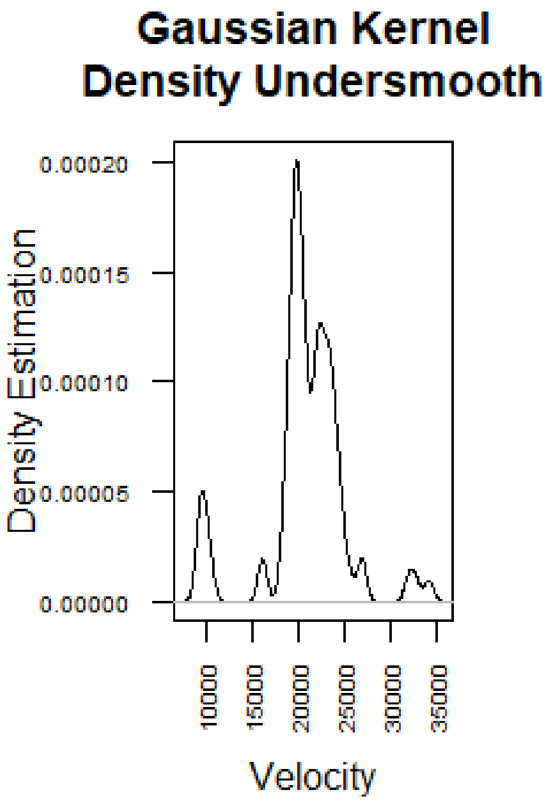
Figure 1.8

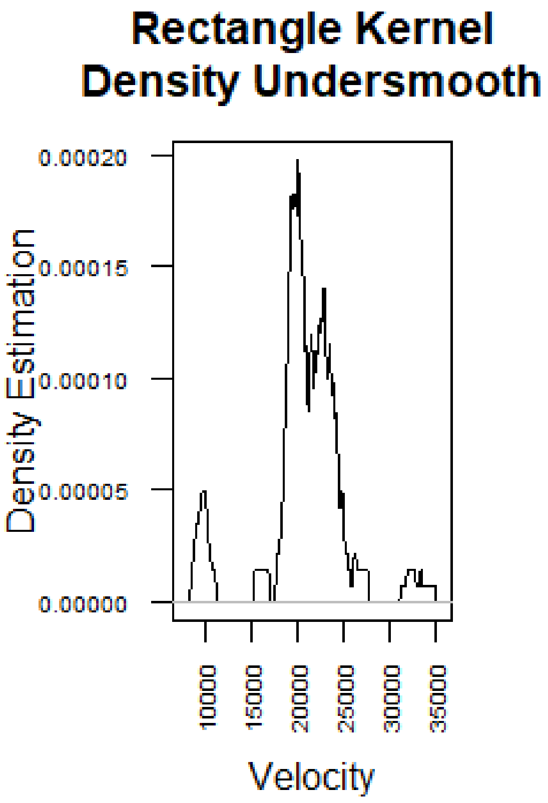


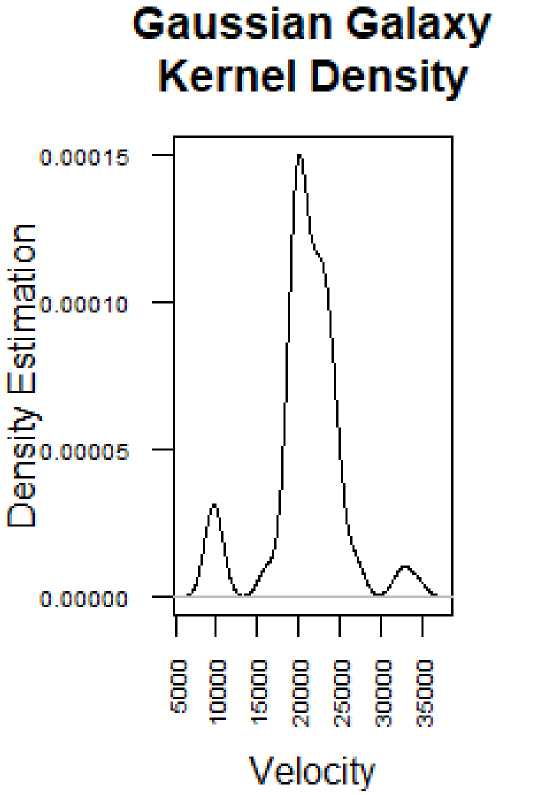


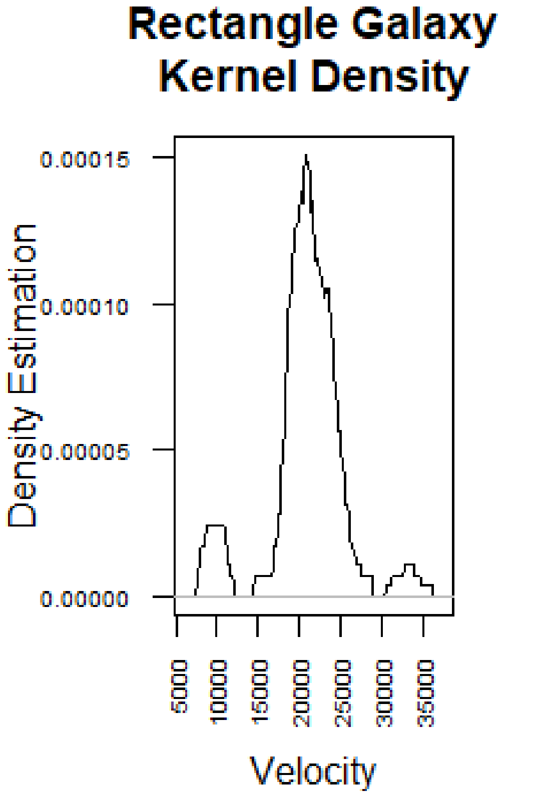


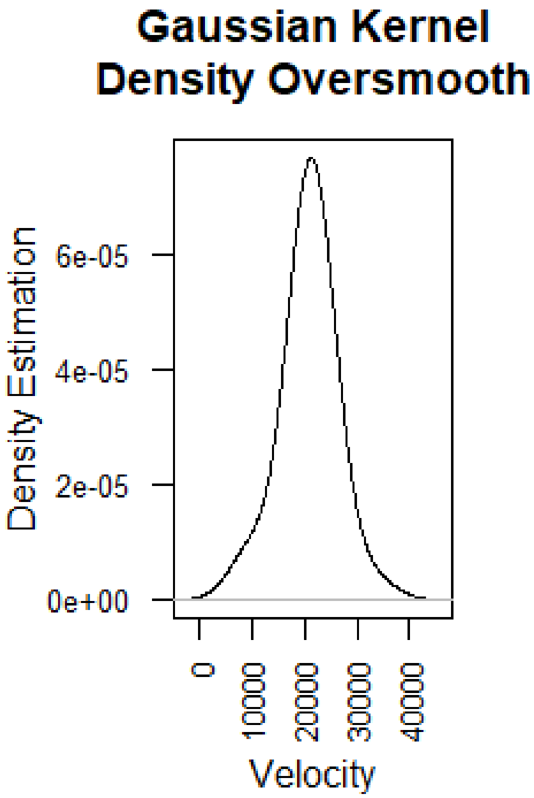
Results, con't: Per our homework instructions, here are 6 comparable base R plots. Again we see the oversimplification from the oversmooth plots and the multimodality of the undersmooth plots.

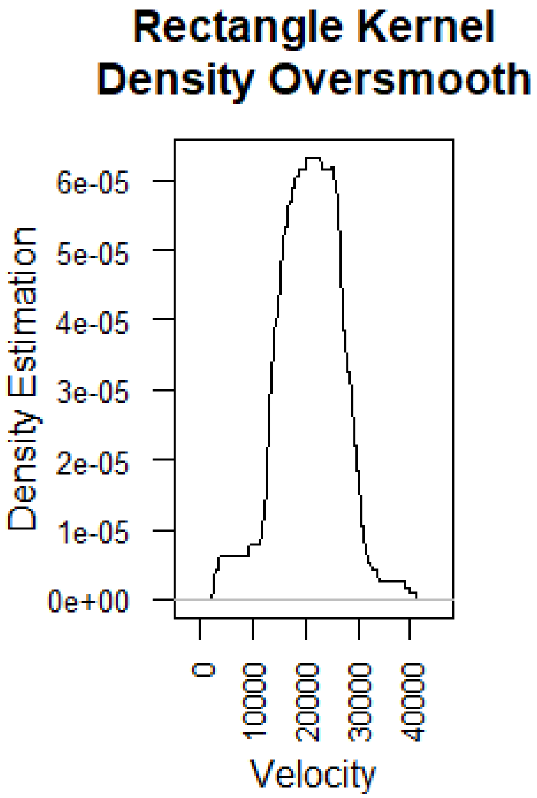





Gaussian Galaxy
Kernel Density
Density Estimation
0.00015
0.00010
0.00005
0.00000
5000
10000
15000
20000
25000
30000
35000
Velocity


Rectangle Galaxy
Kernel Density
Density Estimation
0.00015
0.00010
0.00005
0.00000
5000
10000
15000
20000
25000
30000
35000
Velocity


Gaussian Kernel
Density Oversmooth
Density Estimation
6e-05
4e-05
2e-05
0e+00
0
10000
20000
30000
40000
Velocity


Rectangle Kernel
Density Oversmooth
Density Estimation
6e-05
5e-05
4e-05
3e-05
2e-05
1e-05
0e+00
0
10000
20000
30000
40000
Velocity

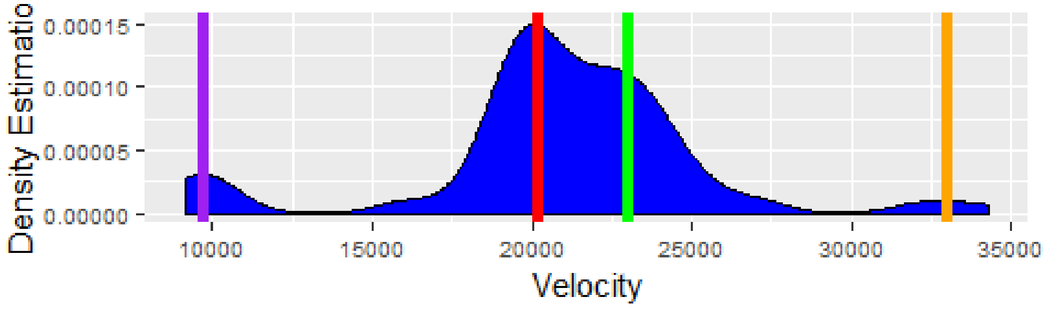
Problem #1, Part D: What is the conclusion about the possible existence of superclusters of galaxies? How many superclusters (1,2,3…)?

Results: I believe Figures 1.10 & 1.11 (shown below) confirm the possible existence of superclustered galaxies. Based on the multimodality of the distribution of the velocities, I would predict 4 superclusters. Below I've added vertical lines showing my interpretation of the possible presence of the superclusters. For comparison, I've included the lines on both the Gaussian and Rectangular plots.

There are not analogous Base R plots as Problem 1, Part D does not request any plots.

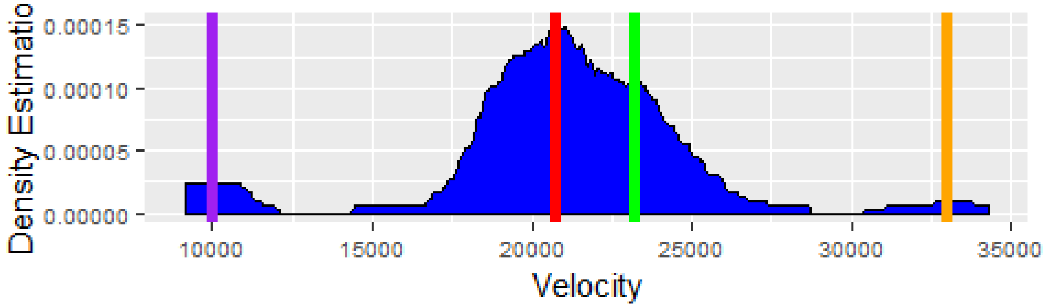
Gaussian Galaxy Kernel Density

Figure 1.10



Rectangle Galaxy Kernel Density

Figure 1.11



Problem #1, Part E: Using Mclust function in R. How many clusters did it find? Did it match with your answer from (d) above? Report parameter estimates and BIC of the best model.

Results: Here I've used the Mclust function to fit a finite mixture model. The model found 4 clusters. This matches my observation in *Part D* of this exercise.

The parameter estimates are included in *Figure 1.12*. As we can see, 75.66% of the observations are in clusters #2 and #3. These clusters have mean velocities (km/sec) of 19,807 and 22,880, respectively.

The BIC value is reported in *Figure 1.13* as **-1579.862**. The summary after Figure 1.13 shows the best model, as well as the other two BIC values that were nearest to the BIC value. As we see, the model with 4 clusters and varying variances (V) is the best model. The second best model is with 3 clusters and equal variances. The third best model is with 7 clusters and equal variances. *Figure 1.14* visually summarizes the BIC values for 1-9 clusters by model (E or V). E stands for **equal variance** and V represents **varying variance**

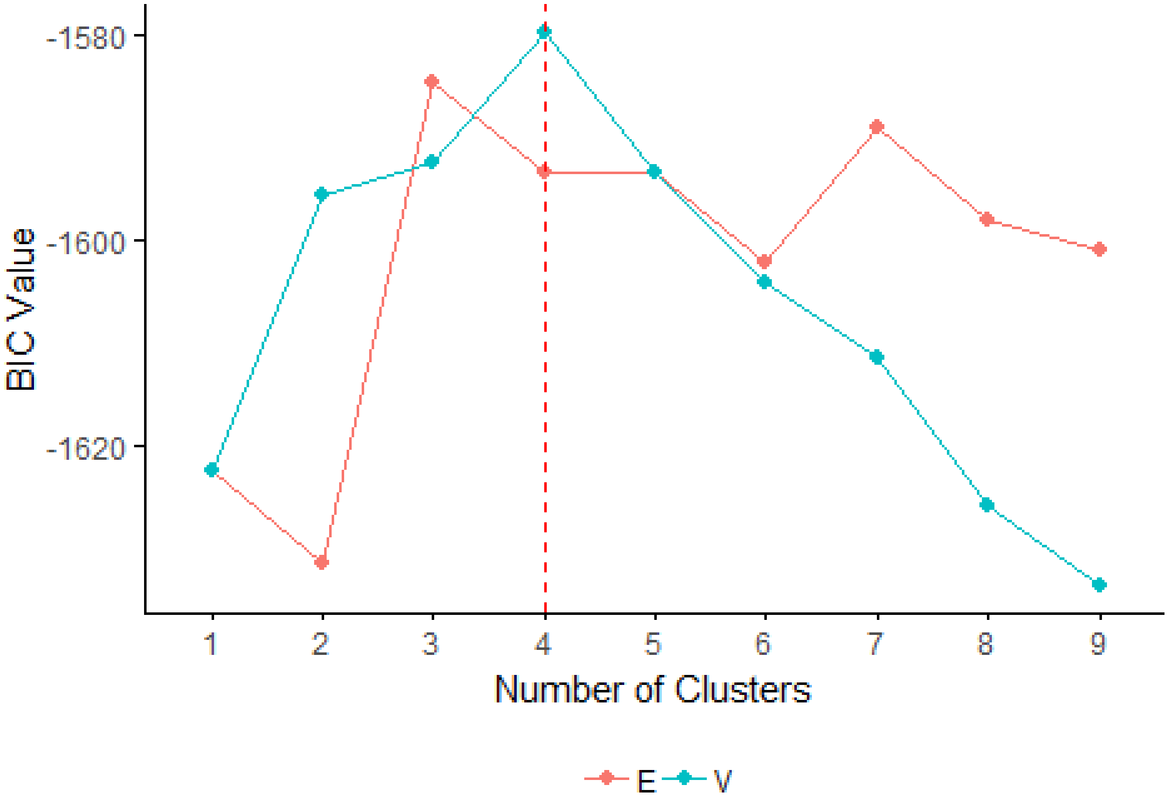
Figure 1.12: Mclust Supercluster Parameter Estimates

Clusters	Mixing Probabilities	Means	Variances
1	**0.0844193**	**9707.522**	**177311.8**
2	**0.3876859**	**19806.592**	**437746.2**
3	**0.3689634**	**22880.348**	**1231115.8**
4	**0.1589315**	**24483.603**	**34305975.7**

Figure 1.13: Mclust Supercluster BIC

BIC
-1579.937

```
## Best BIC values:
##                 V,4          E,3          E,7
## BIC      -1579.937 -1584.672708 -1589.152853
## BIC diff     0.000    -4.735519    -9.215663
```



Model selection

Figure 1.14: Best model: V | Optimal clusters: n=4

Results: If we look at our plots, again, we may be able to visualize why the 3 and 7 cluster models were 2nd and 3rd best choices, according to BIC.

Looking at *Figure 1.15*, although there are 4 distinct modes in the plot, the three that are highlighted below and the most distinct / largest.

If we look at our undersmoothed plot, *Figure 1.16*, we can see 7 distinct clusters.

While the 4 cluster model is the best, these two examples help us understand why the 3 cluster and 7 cluster models were the next best choices using the 'Mclust()' function.

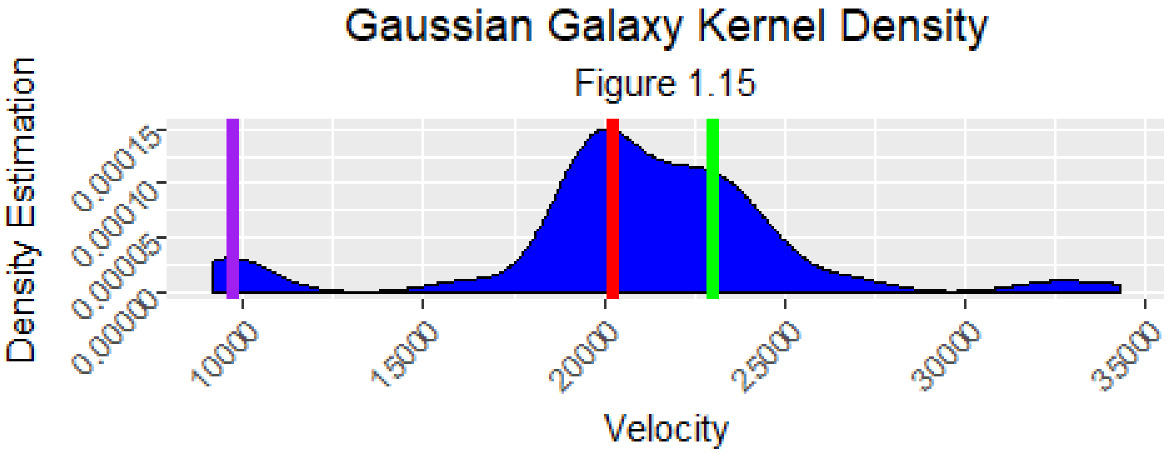

Gaussian Galaxy Kernel Density

Figure 1.15

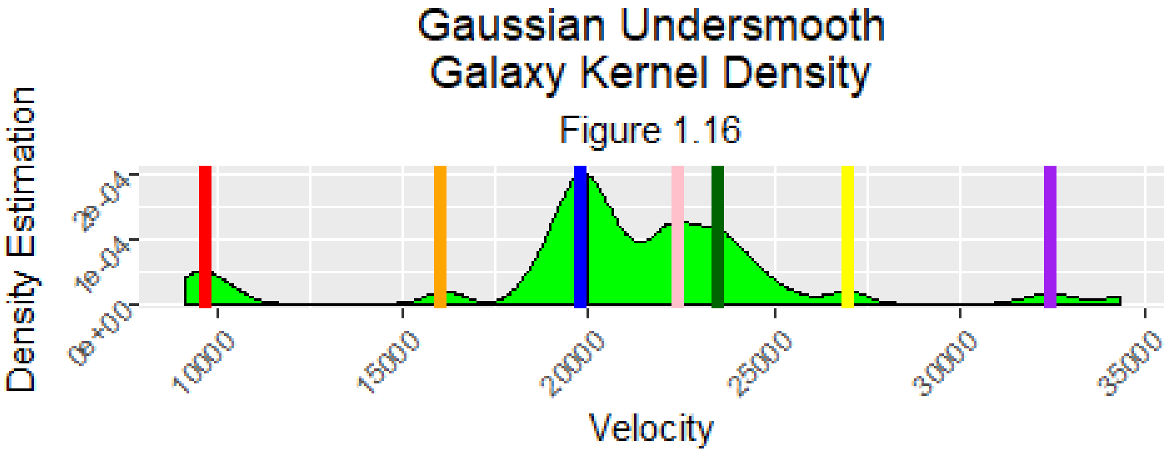
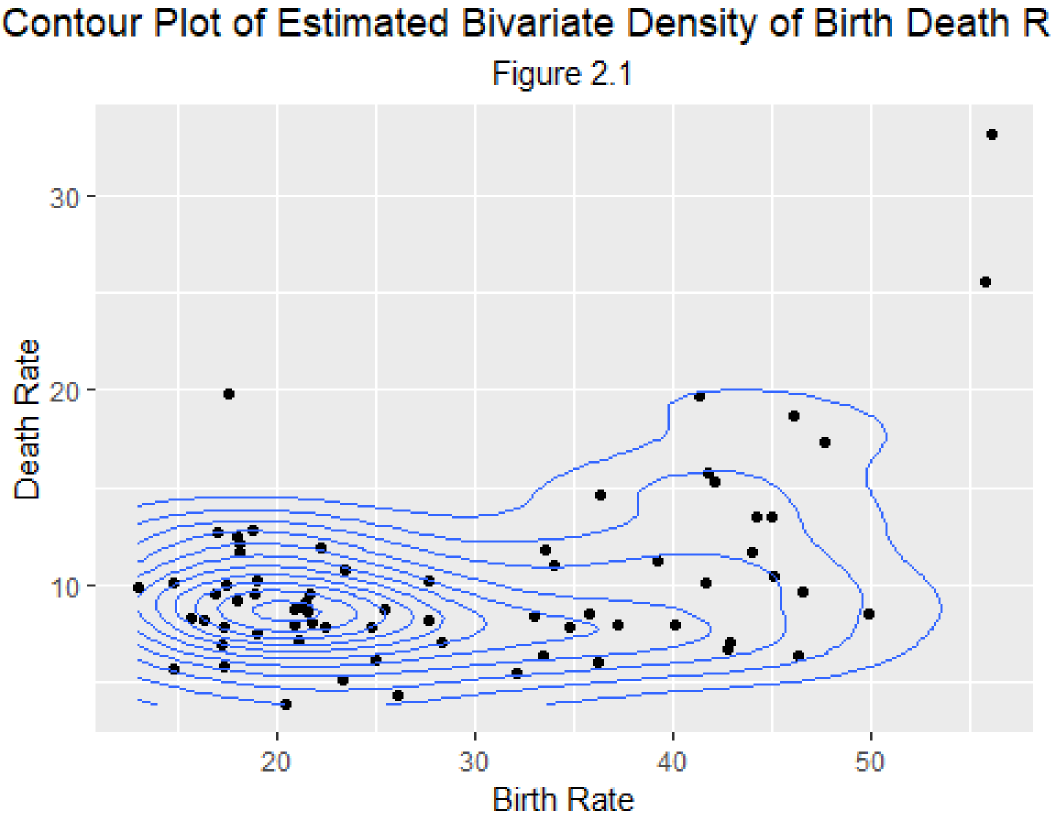

Gaussian Undersmooth
Galaxy Kernel Density

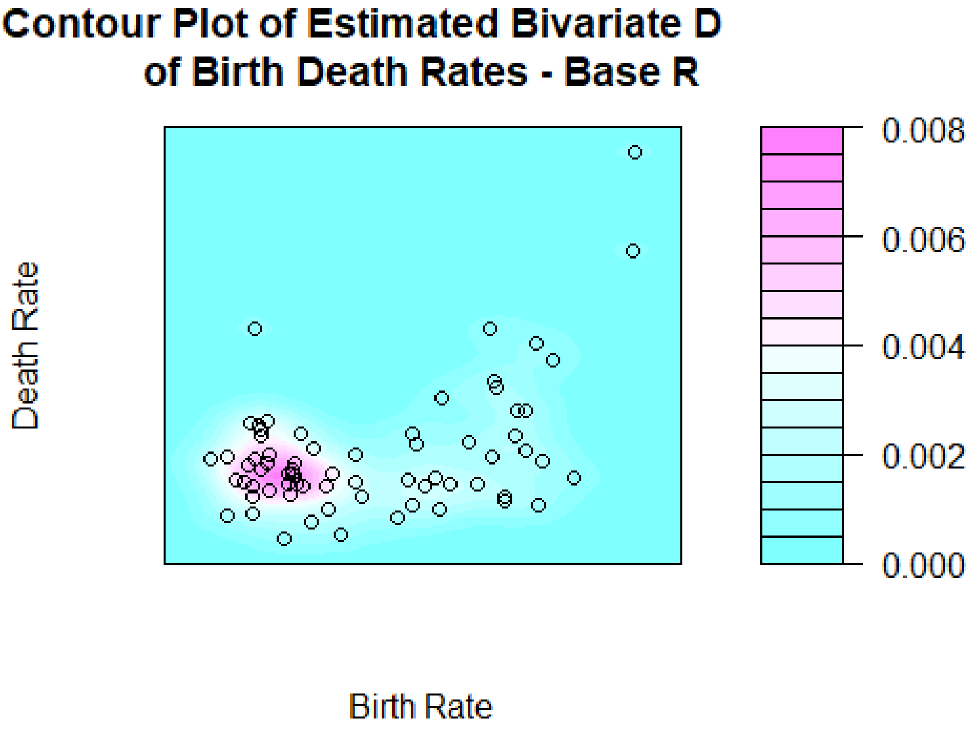
Figure 1.16

Problem #2, Part A: The **birthdeathrates** data from **HSAUR3** gives the birth and death rates for 69 countries (8.2 Handbook).

Produce a scatterplot of the data that shows a contour plot of the estimated bivariate density. Put the original points on the contour plot.

Results: In *Figure 2.1* we see the data are clustered near the point when 'birth rate' = 20 and 'death rate' = 10. We also see a few posts outside the contour lines. The general distribution of the points, and especially the points outside the contour, will be examined in greater detail in our next plot.





Problem #2, Part B: Does the plot give you any interesting insights into the possible structure of the data?

Results: It does. Looking at *Figure 2.2* below, we see that, for most of the data points, birth rate is at least 2 times higher than death rate. In many of the points, birth rate is even higher, proportionate to death rate.

Only two of the countries have a death rate higher than 20 (highlighted in yellow) and one one country has a death rate higher than their birth rate (highlighted in green). I've included the country abbreviation from the dataset for these countries.

Additionally, a red point is used to denote the country if its birth rate is not at least twice as large as its death rate. As we see, my initial comment regarding the ratio of birth to death rates is accurate, with the majority of the countries having at least 2 births per 1 death.

Below *Figure 2.2*, I've included a table showing the top 3 countries for birth:death ratio. We see that the countries 'jor' 'ven' and 'syr' have the largest birth:death ratios, all with ratios in excess of 6:1.

Lastly, I've included a table of the countries that are denoted by red dots (meaning their Birth to Death ratio is less than 2:1). I've sorted the table from lowest to highest. As we see, **dem**, highlighted in green on the plot, is the only country with a B:D ratio less than 1:1.

No base R plot is included since the question does not ask us to plot anything, it only asks us for insights into the previous plots.

Contour Plot of Estimated Bivariate Density
of Birth Death Rates (Descriptive)

Figure 2.2

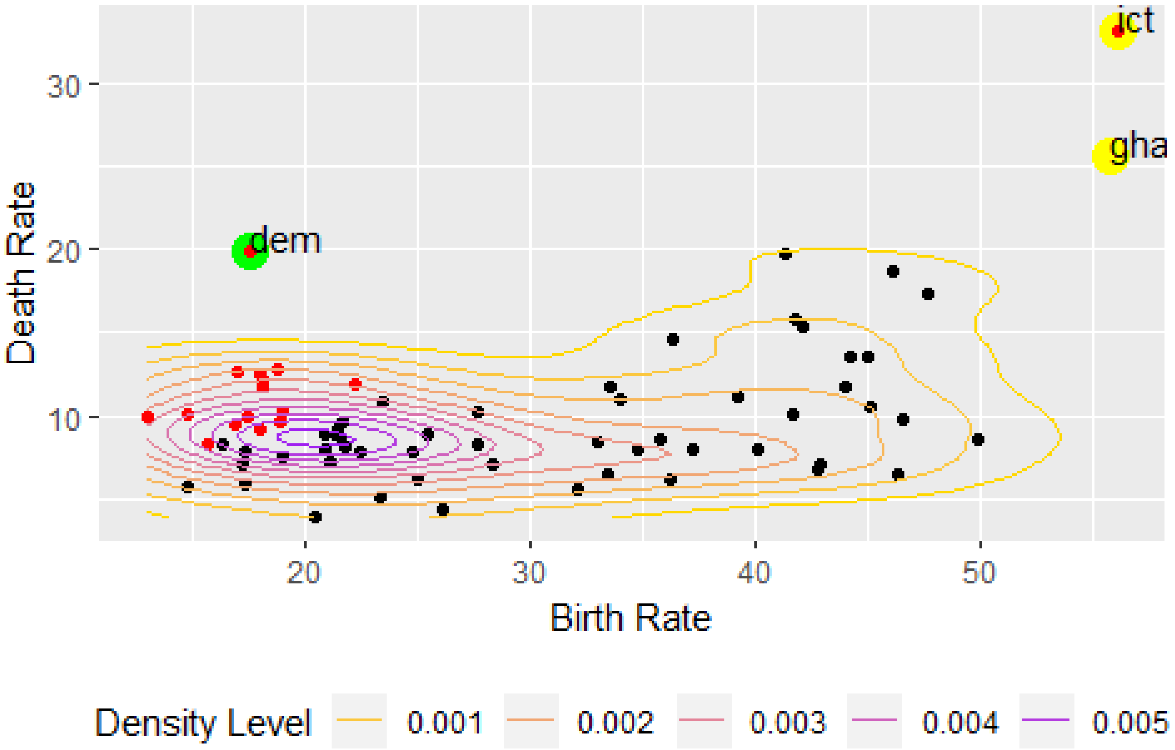


Figure 2.3: Top 3 Birth to Death Ratios

Country	Birth	Death	BirthPerDeath
jor	**46.3**	**6.4**	**7.234375**
ven	**42.8**	**6.7**	**6.388060**
syr	**26.2**	**4.3**	**6.093023**

Figure 2.4: Birth to Death Ratios < 2:1

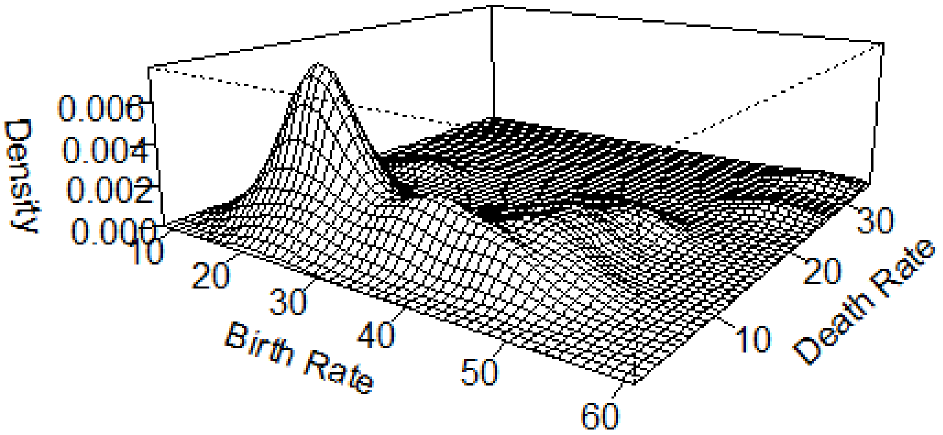
Country	Birth	Death	BirthPerDeath
dem	17.6	19.8	0.8888889
hun	13.1	9.9	1.3232323
bel	17.1	12.7	1.3464567
gmy	18.0	12.5	1.4400000
swe	14.8	10.1	1.4653465
aus	18.8	12.8	1.4687500
brt	18.2	12.2	1.4918033
fra	18.2	11.7	1.5555556
ict	56.1	33.1	1.6948640
now	17.5	10.0	1.7500000
cze	16.9	9.5	1.7789474
ity	19.0	10.2	1.8627451
irl	22.3	11.9	1.8739496
rom	15.7	8.3	1.8915663
fin	18.1	9.2	1.9673913
swz	18.9	9.6	1.9687500

Problem #2, Part C: Construct the perspective plot (persp() in R, GGplot is not required for this question).

Results: With the perspective plot, *Figure 2.5*, we can see a 3 dimensional depiction of *Figure 2.2*. Again, the majority of the data observations have a death rate that is proportionally smaller than the birth rate. Similar to with Figure 2.2, we see a couple of outliers with higher death rates.

No ggplot version included per homework instructions.

Figure 2.5: Birth Death Rate Perspective Plot

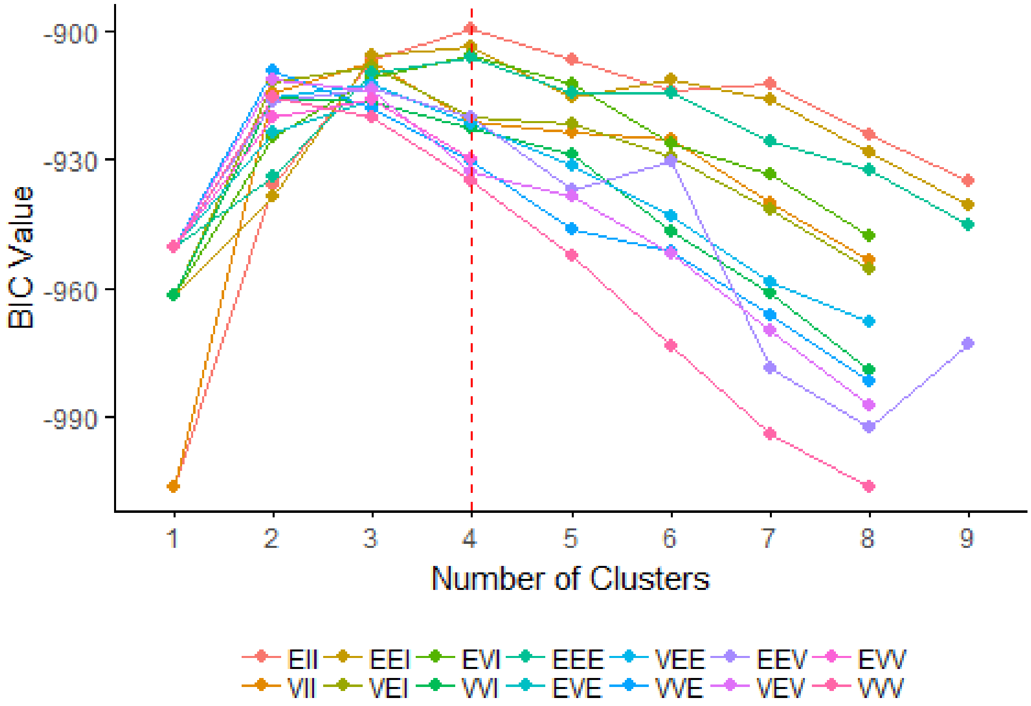


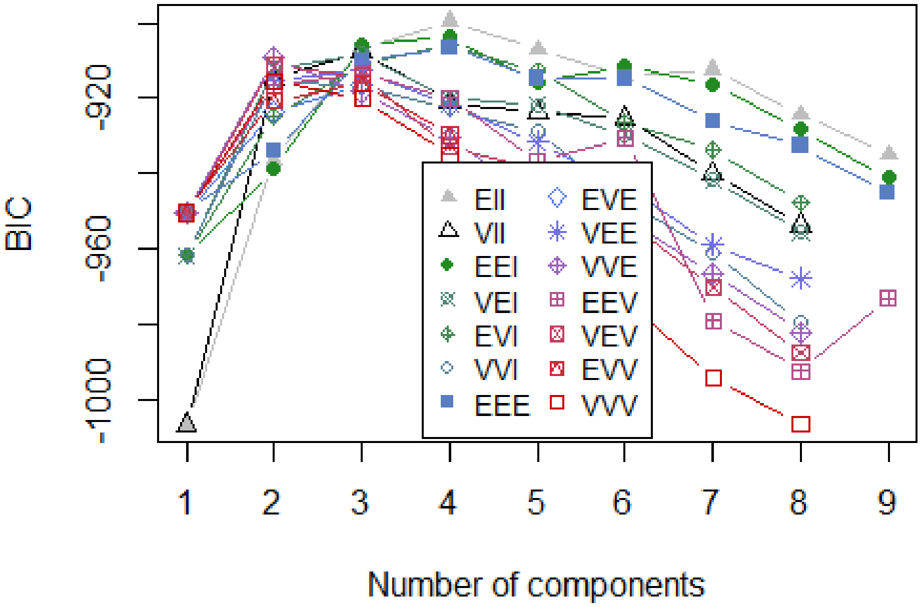
Problem #2, Part D: Use/perform Model-based clustering (Mclust, library mclust). Provide plot of the summary of your fit (BIC, classification, uncertainty, and density).

Results - BIC: *Figure 2.6* shows the BIC values, by cluster, for each multivariate mixture. As we can see, a cluster of 4 with the EII multivariate mixture has the largest BIC indicating the best model. EII tells us that the multivariate model mixture has **spherical, equal model**. The 2nd largest BIC, from Figure 2.6, is a cluster of 4 with EEI multivariate mixture. EEI indicates that the model is diagonal with **equal volume and shape**.

Model selection

Figure 2.6: Best model: EII | Optimal clusters: n=4



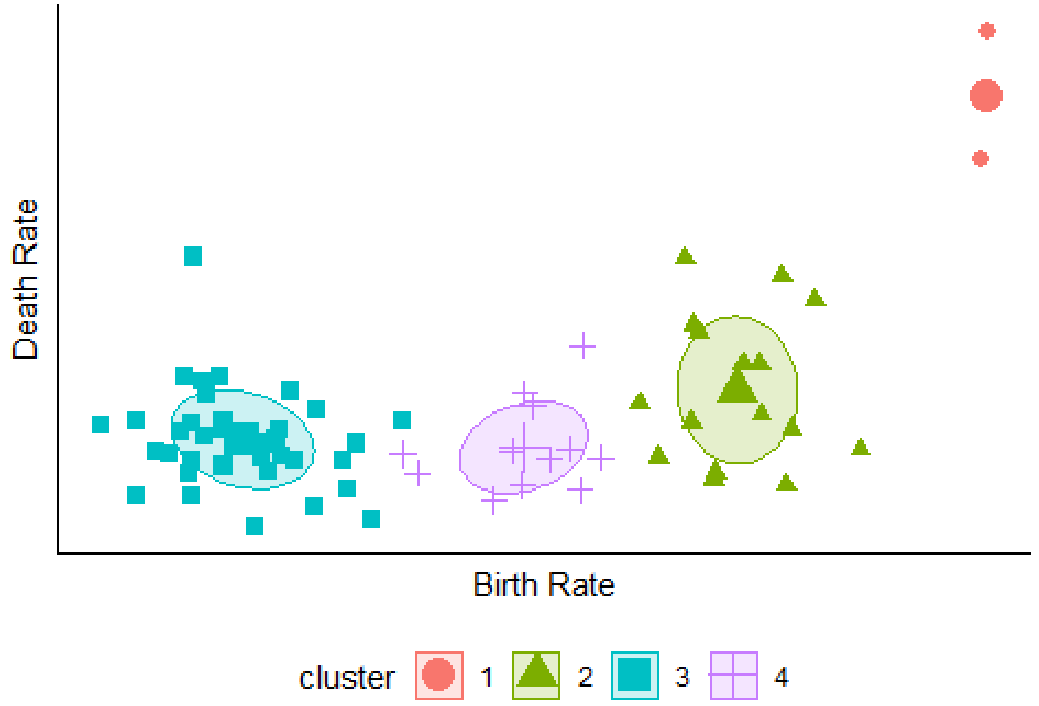


Results - Classification: *Figure 2.7* and the corresponding base R plot shows the four distinct clusters on a scatter plot with an x axis of ‘Birth Rate’ and y axis of ‘Death Rate’. As we see, ‘cluster #1’ contains only 2 points. The remaining 3 clusters contain numerous observations that are pretty distinct with few points outside the cluster ellipse. The larger symbols, that correspond to each plot in Figure 2.7, denote the center of the cluster.

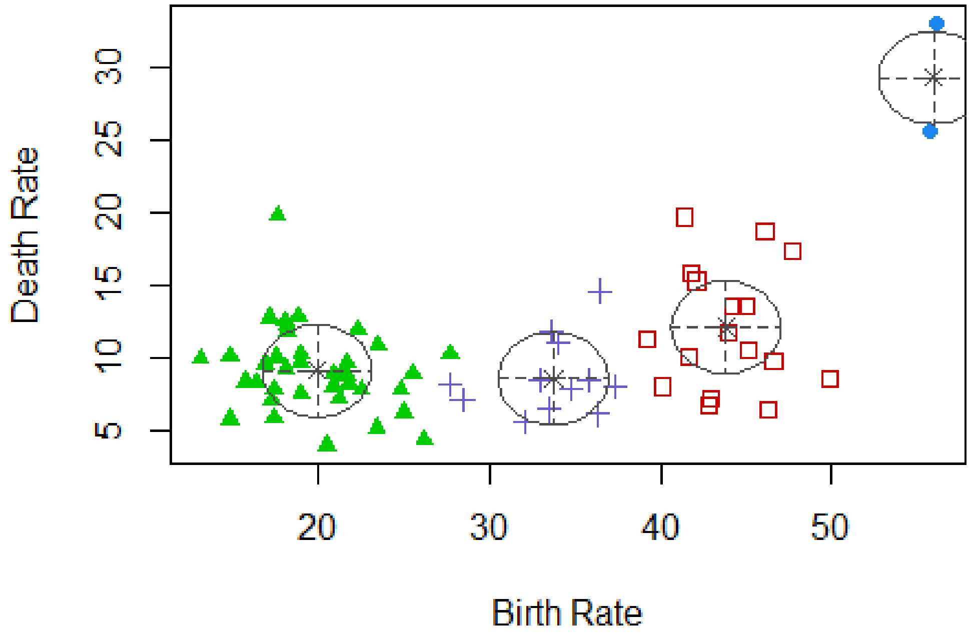
As a note, we see that Figure 2.7 does not produce an ellipse for ‘cluster #1’. This is a product of the visualization tool used, but does not negate the presence of a cluster.

Cluster plot

Figure 2.7: Classification



Classification

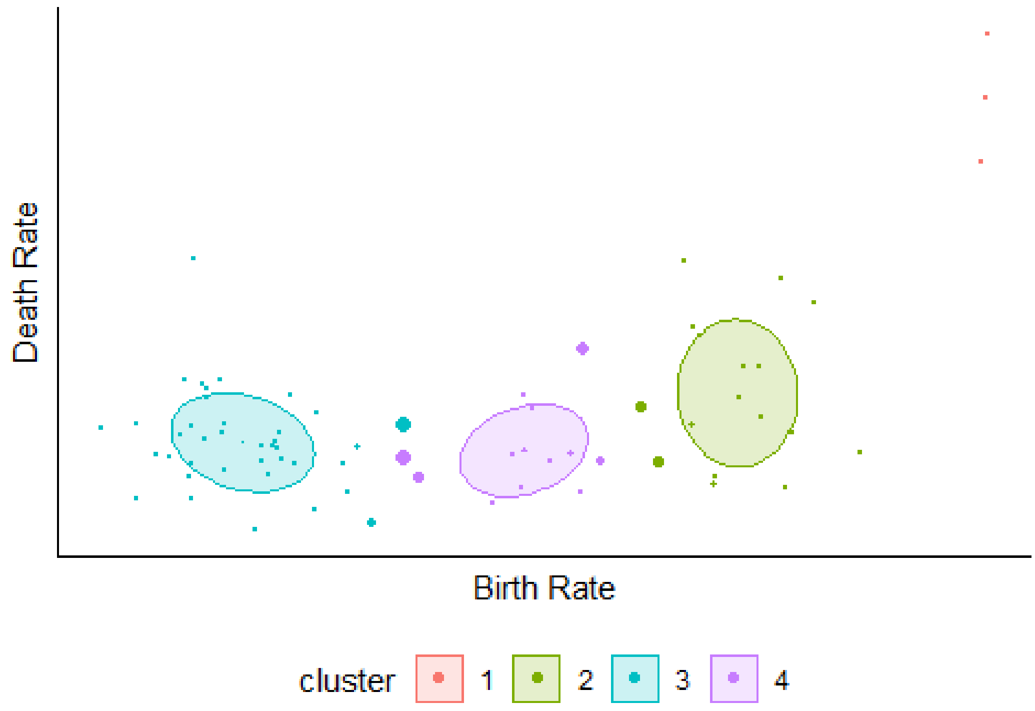


Results - Uncertainty: *Figure 2.8* is similar to *Figure 2.7* but, in this instance, the plot also shows the points' distance from its corresponding cluster. These are highlighted by the size of the points, scaled to correspond with its distance from the cluster center. As we can see, cluster #3 has the most points lying outside the cluster while cluster #4 has the outliers that lie furthest from the cluster.

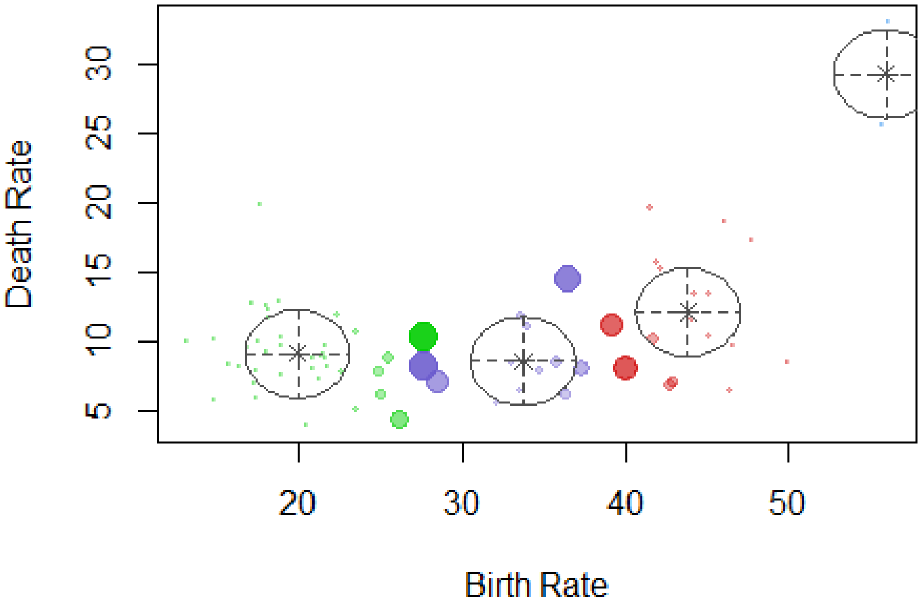
As with *Figure 2.7*, we see that *Figure 2.8* does not produce an ellipse for 'cluster #1'. This is a limitation of the visualization tool used, but does not negate the presence of a cluster.

Cluster plot

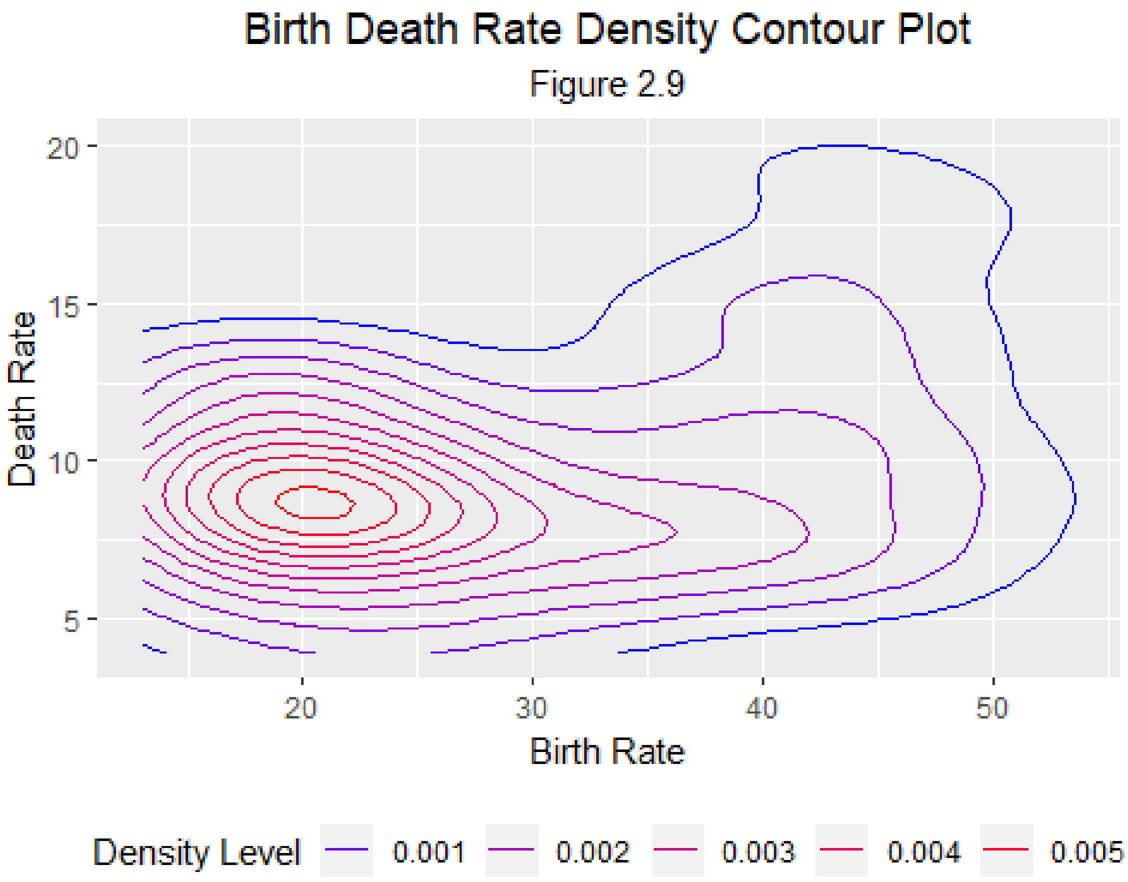
Figure 2.8: Uncertainty

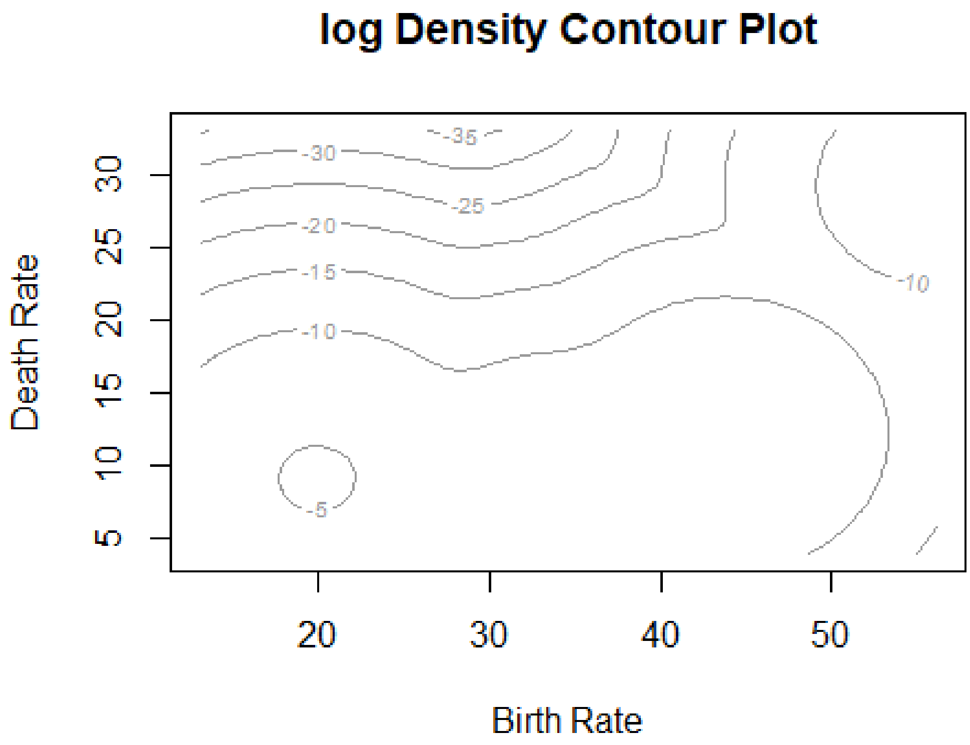


Uncertainty



Results - Density: The density contour, *Figure 2.9* plot is centered around the approximate spot where birth rate is 2 times that of death rate. We also can see that most of the dispersion favors a higher birth to death ratio that 2:1.





Problem #2, Part E: Discuss the results (structure of data, outliers, ...). Write a discussion in the context of the problem.

Results: The plots above provide evidence that the data can be presented in 4 clusters. *Figure 2.6* shows the BIC values, by cluster, for each multivariate mixture. As we can see, a cluster of 4 with the EII multivariate mixture has the largest BIC indicating the best model. EII tells us that the multivariate model mixture has **spherical, equal model**.

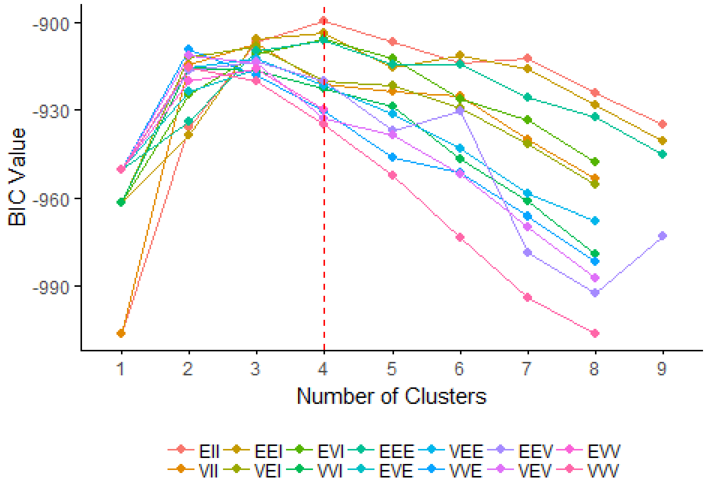
Figure 2.7 and the corresponding base R plot shows the four distinct plots on a scatter plot with an x axis of 'Birth Rate' and y axis of 'Death Rate'. As we see, 'cluster #1' contains only 2 points.

Figure 2.8 is similar to *Figure 2.7* but, in this instance, the plot also shows the points' distance from its corresponding cluster. These are highlighted by the size of the points, scaled to correspond with its distance from the cluster center.

Figure 2.9 has a contour that is centered around the approximate spot where birth rate is 2 times that of death rate. We also can see that most of the dispersion favors a higher birth to death ratio that 2:1.

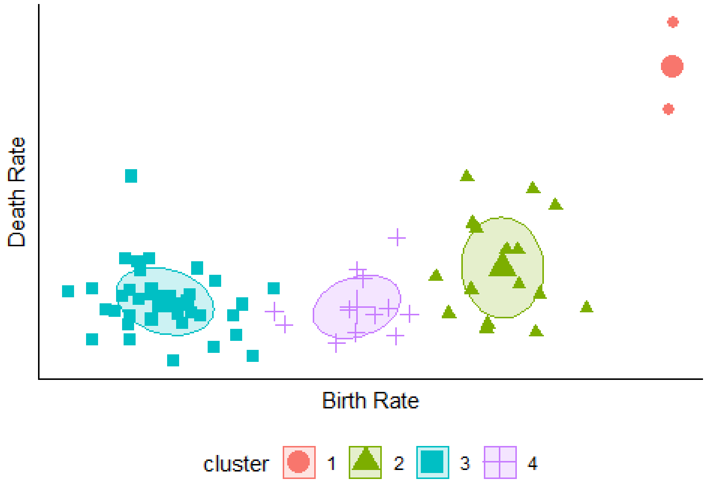
Model selection

Figure 2.6: Best model: EII | Optimal clusters: n=4



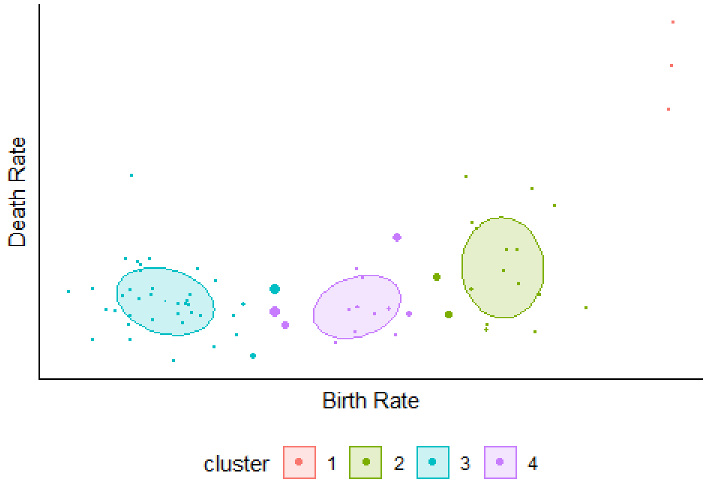
Cluster plot

Figure 2.7: Classification



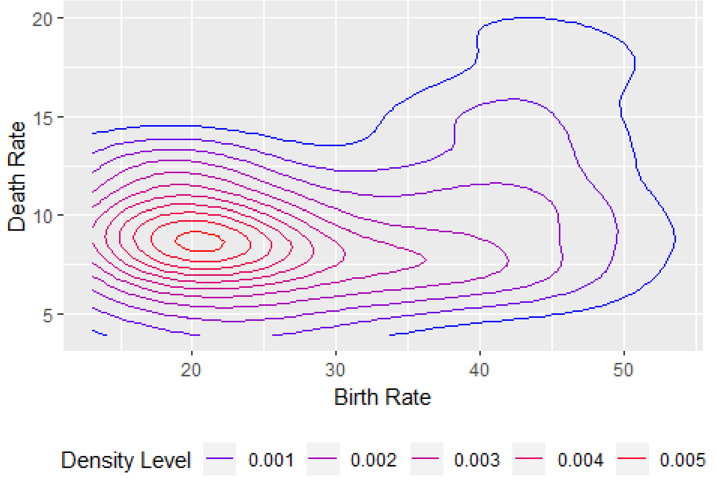
Cluster plot

Figure 2.8: Uncertainty



Birth Death Rate Density Contour Plot

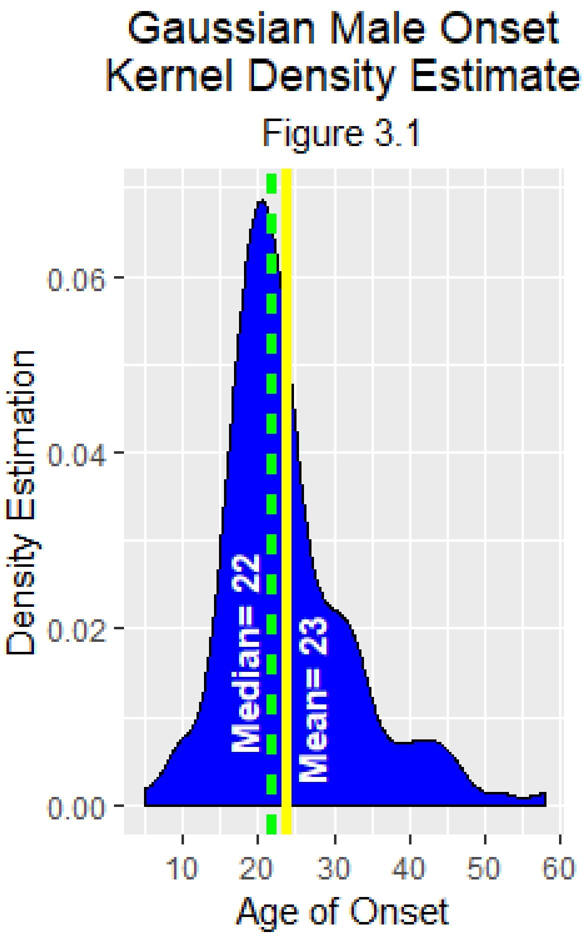
Figure 2.9

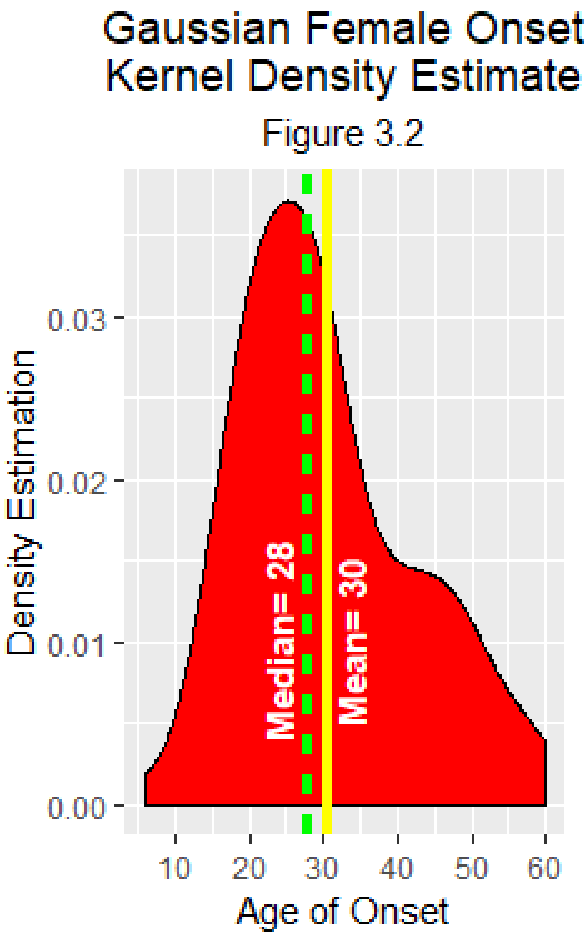


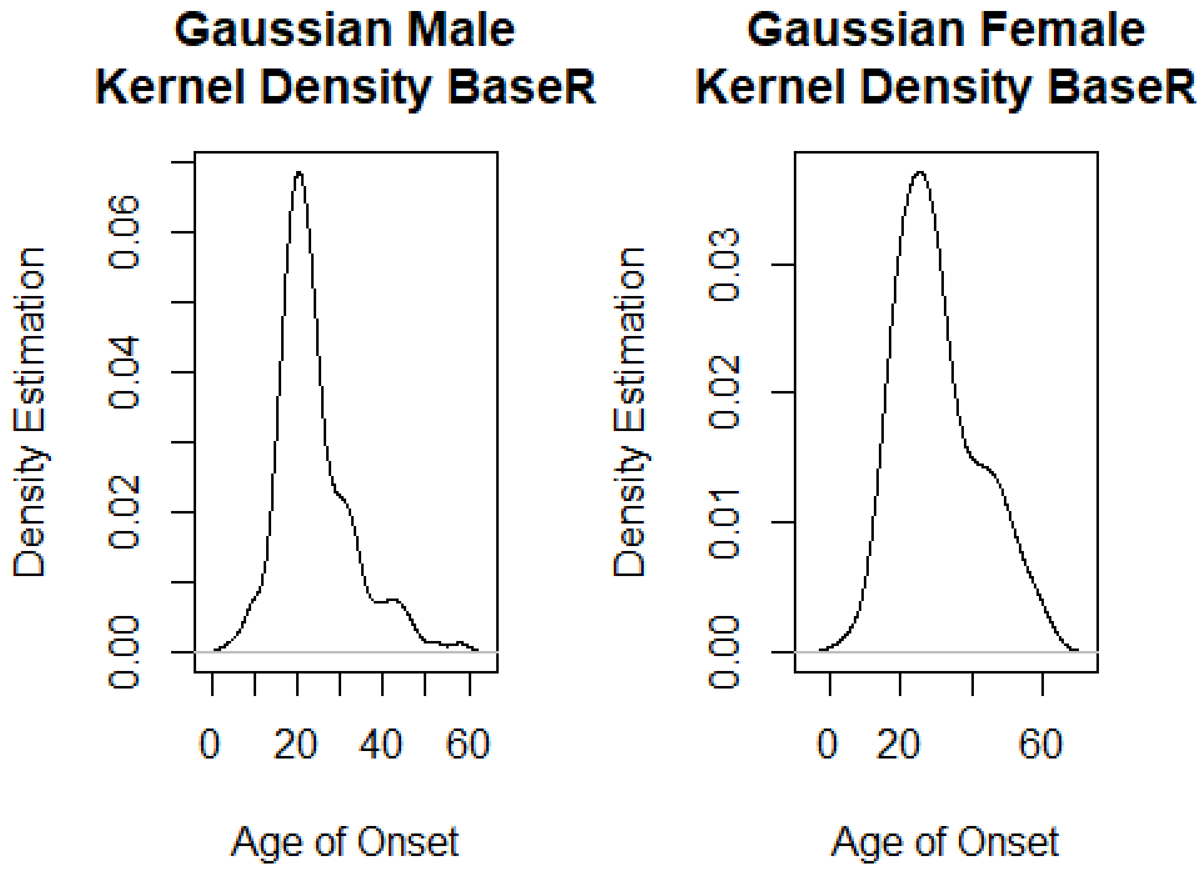
Problem #3: A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequent epidemiological studies of the disorder have consistently shown an earlier onset in men than in women. One model that has been suggested to explain this observed difference is known as the subtype 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 mixtures of normal densities separately to the onset data for men and women given in **schizophrenia** data from **HSAUR3**. See if you can produce some evidence for or against the subtype model. (8.3 Handbook)

Results: First, we plot Gaussian Kernel Density Estimate plots, **Figure 3.1** and **Figure 3.2**. These plots indicate that there is some evidence to support the 'subtype' model's claim which, in part, states that early onset schizophrenia is largely a disorder affecting men and late onset schizophrenia is more associative with women. Both figures show the mean and median ages of onset for each gender.

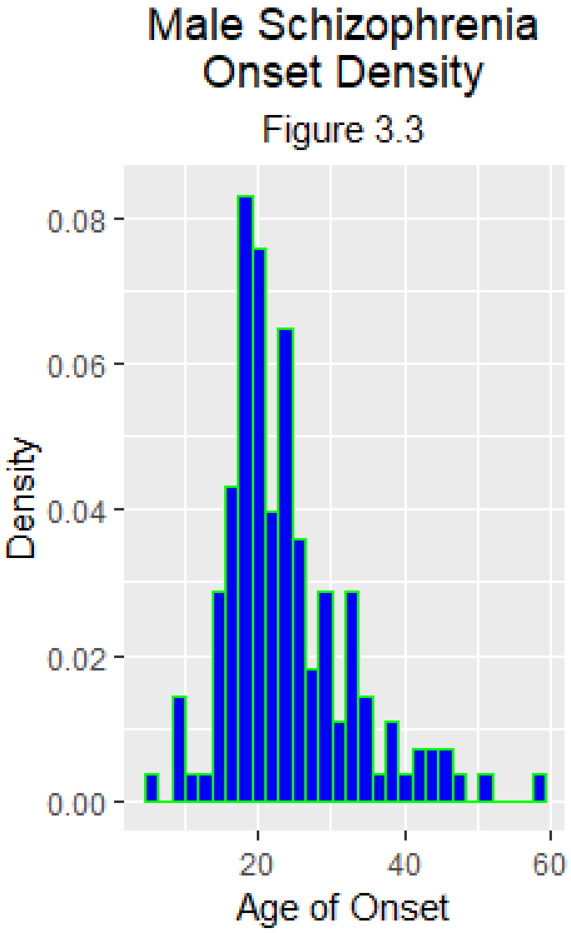
A base R plot is included, as a comparison, per the homework guidelines.

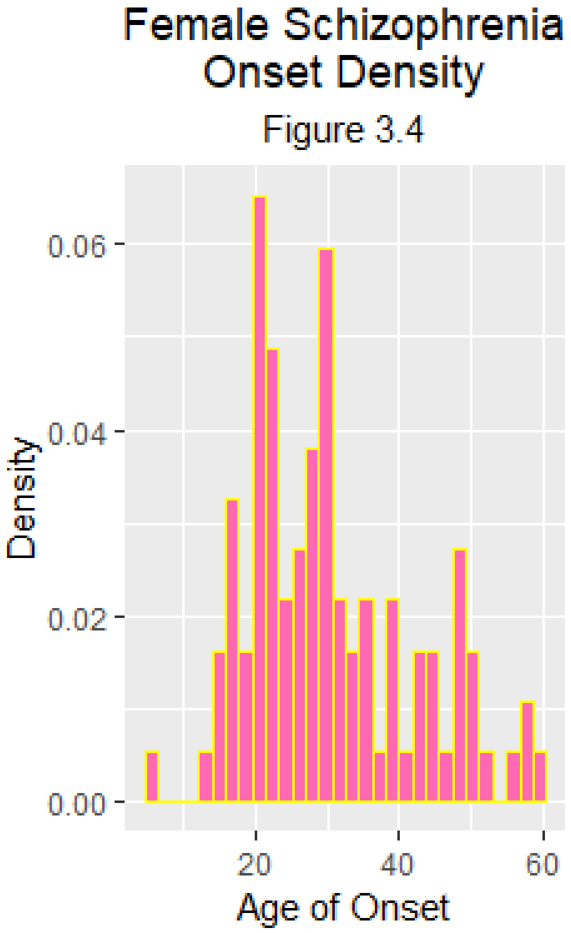


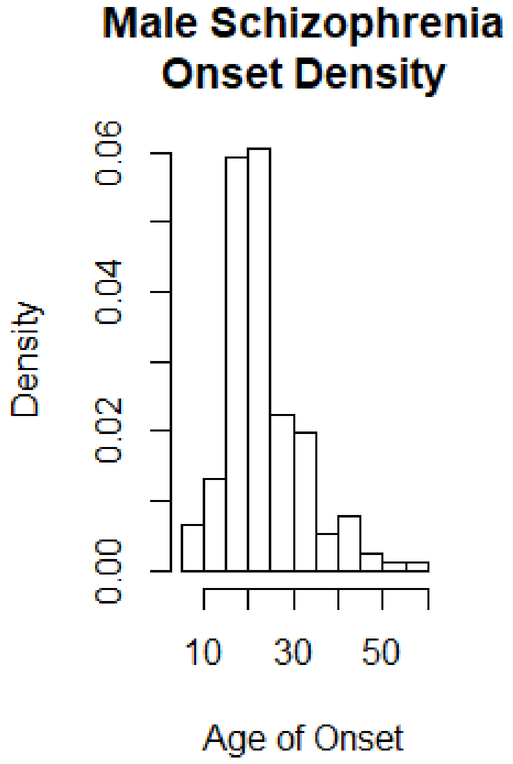


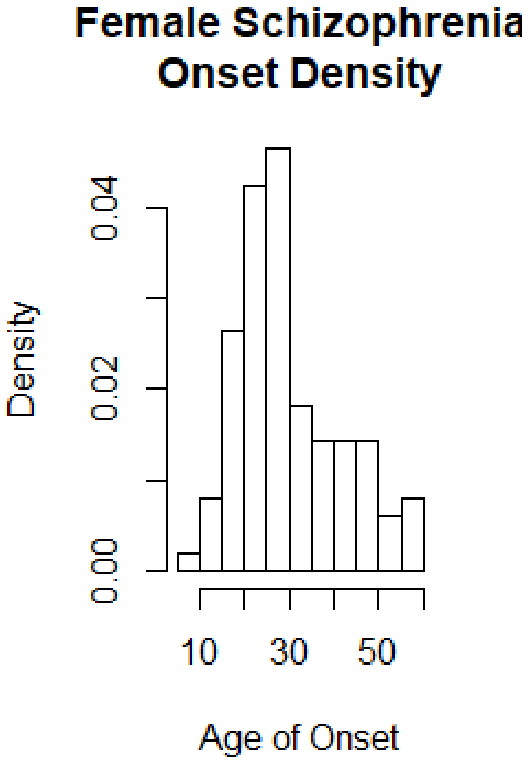

Gaussian Male
Kernel Density BaseR
Density Estimation
0.00
0.02
0.04
0.06
0
20
40
60
Age of Onset
Gaussian Female
Kernel Density BaseR
Density Estimation
0.00
0.01
0.02
0.03
0
20
60
Age of Onset

Results, cont'd: Looking at a histogram comparison, we again see an apparent difference in the age of onset for schizophrenia for men and women. **Figure 3.3** shows us that the peak age for men appears to be around the age of 20. **Figure 3.4** shows us that the age of onset for women appears to be more spread out. An analogous base R plot is included to meet assignment requirements.









Results, cont'd: Below we have two figures futhur explaining the data and differences between men and women onset ages.

Figure 3.5 shows that the 'male' data can be separated into 2 clusters. The first cluster, with a mean of nearly 20, contains over 51% of the male observations.

Figure 3.6 shows that the 'female' data can also be separated into 2 clusters. The first cluster, with a mean age of nearly 25, contains just shy of 75% of the female observations.

One thing I find interesting is that over 25% of the other female observations have an age of onset mean of nearly 47. The 2nd male cluster, which contains just short of 49% of the male observations, has a mean age of onset of less than 28 years old.

These charts show, once again, that the average age of onset for women is slightly higher than that of men. These figures shine a brighter light on the difference in onset for the second clusters from both male and female observations.

Figure 3.5: Mclust Age of Onset Parameter Estimates - Males

Cluster	Mixing Probabilities	Means	Variances
1	0.5104189	20.23922	9.395305
2	0.4895811	27.74615	111.997525

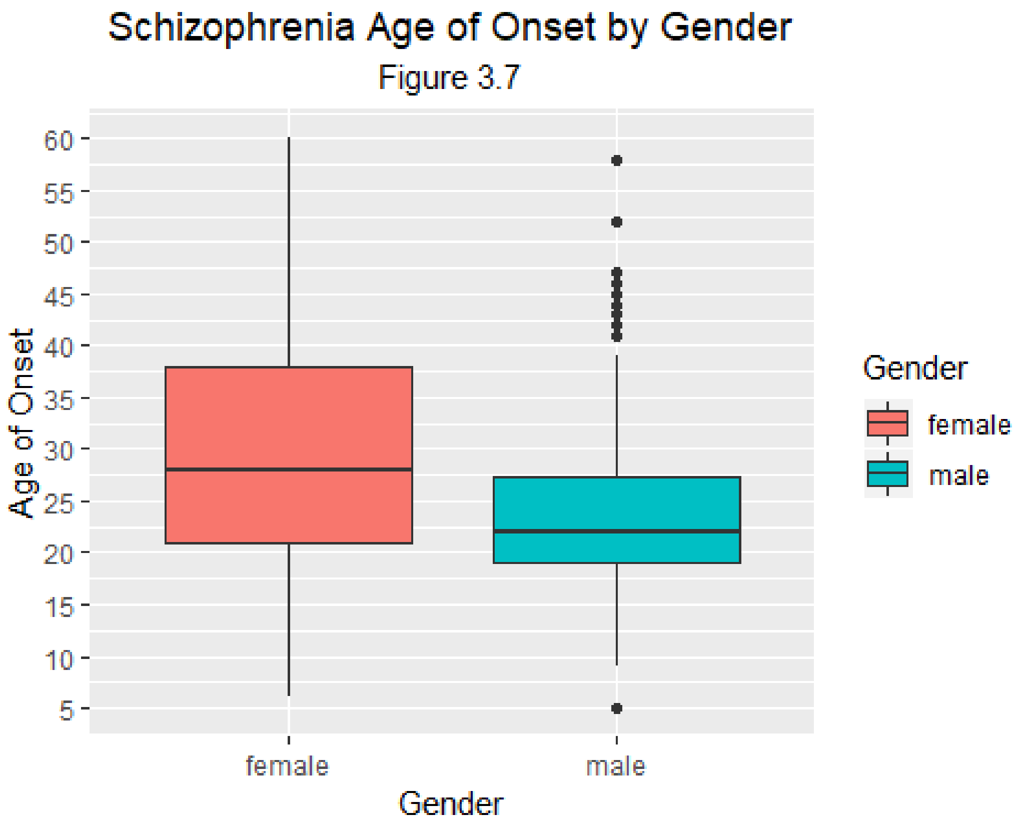
Figure 3.6: Mclust Age of Onset Parameter Estimates - Females

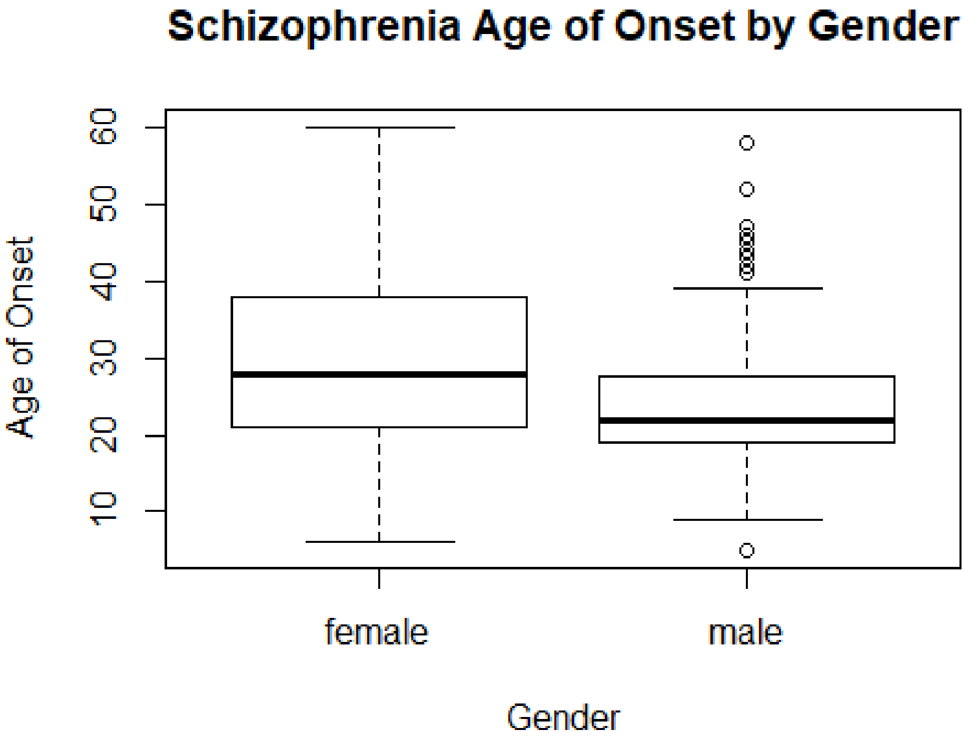
Cluster	Mixing Probabilities	Means	Variances
1	0.7472883	24.93517	44.55641
2	0.2527117	46.85570	44.55641

Results, cont'd: *Figure 3.7* contains a simple boxplot that shows the distribution, among genders, of the age of onset of schizophrenia. As we can see, over 75% of women do not experience schizophrenia until after age 20 where 75% of men have an onset before the age of 28.

We can also see that 50% of men experience the onset of schizophrenia prior to the age of 23, compared to women where 50% don't experience the onset of schizophrenia until after age 27.

An analogous base R plot is included to meet assignment requirements.



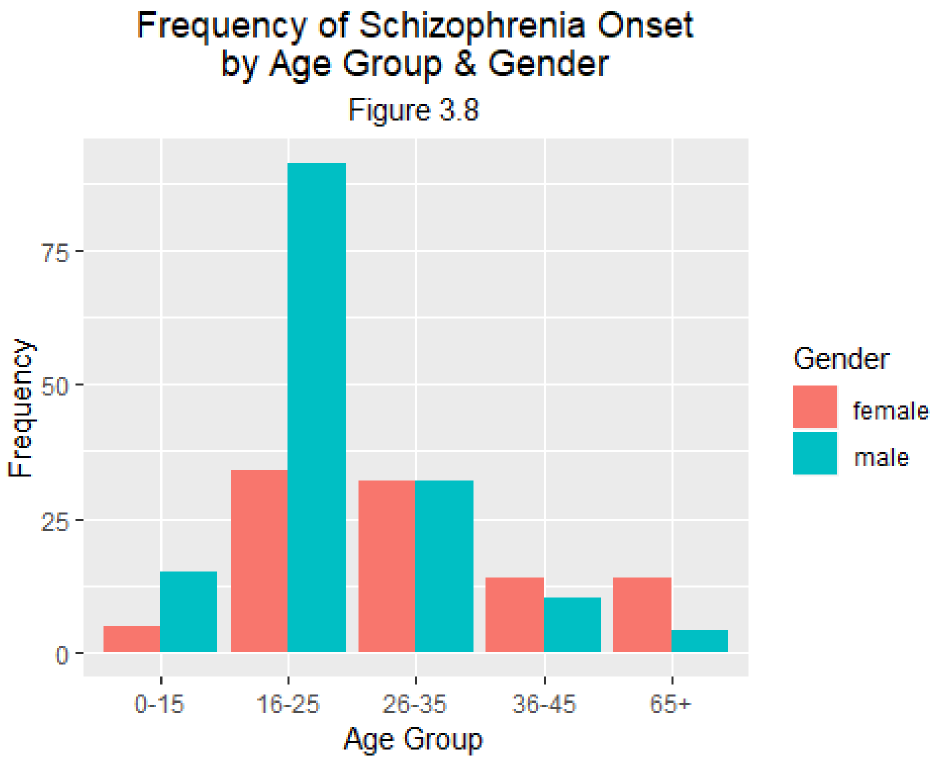


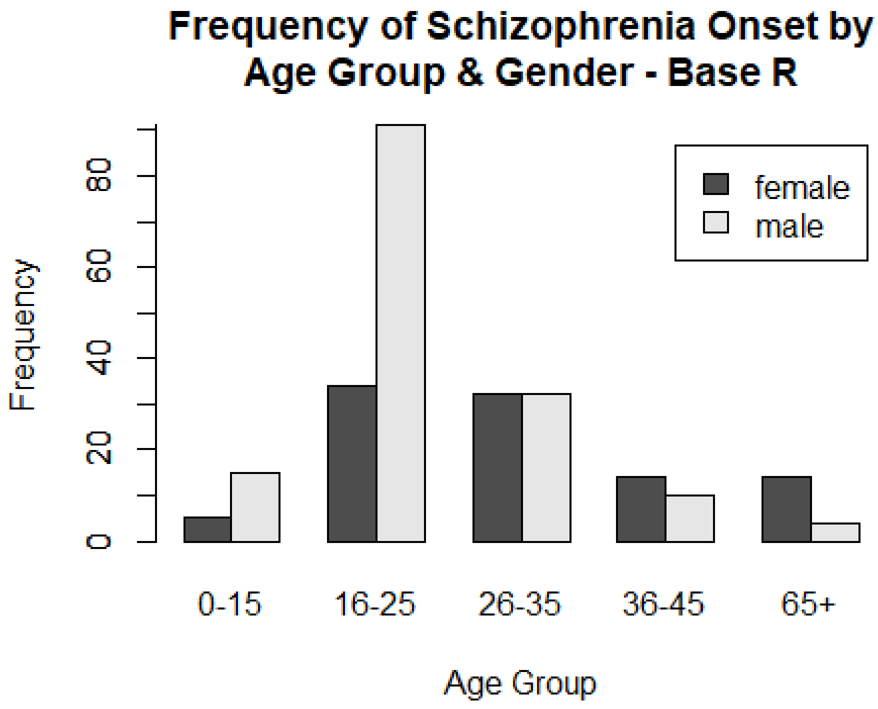
Results, cont'd: So far it appears that the subtype model has validity. Here we'll look at one last plot showing the frequency of onset by age group for each gender.

Figure 3.8 shows the distribution of onset age by gender. As we can see, the bulk of males experience onset between the ages of 16 and 25. This age range is also when women see the bulk of their onsets. On the contrary, women have more instances of onset from 35 - 60. This provides evidence for the subtype model's postulation.

Based on this, and the previous 7 graphs, I think we have quite a bit of evidence in support of the subtype model. Furthermore, we've yet to find any evidence that would disprove the theory that women are more likely to experience onset later in life that men.

A comparable base R plot is included to meet assignment requirements.


Frequency of Schizophrenia Onset
by Age Group & Gender
Figure 3.8
Frequency
Age Group
Gender
female
male


Frequency of Schizophrenia Onset by
Age Group & Gender - Base R
Frequency
female
male
Age Group