Stats 503 Homework 5

Sam Edds

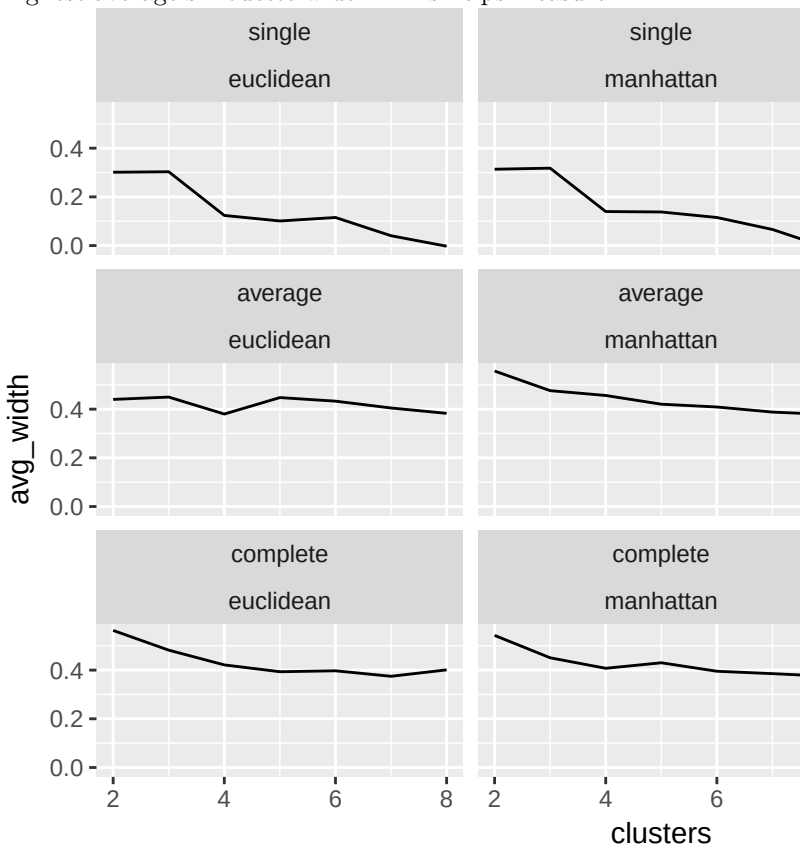
4/17/2018

For this analysis we perform cluster analysis on the crabs dataset, which examines Blue and Orange crab species, by sex, and by traits (frontal lobe size, rear width, carapace length and width, and body depth). We examine how our results classify for both sex and species.

We first convert all our continuous measures into centimeters and compute the distance matrix. We also use MDS to output two variables so we can plot our results in 2 dimensions.

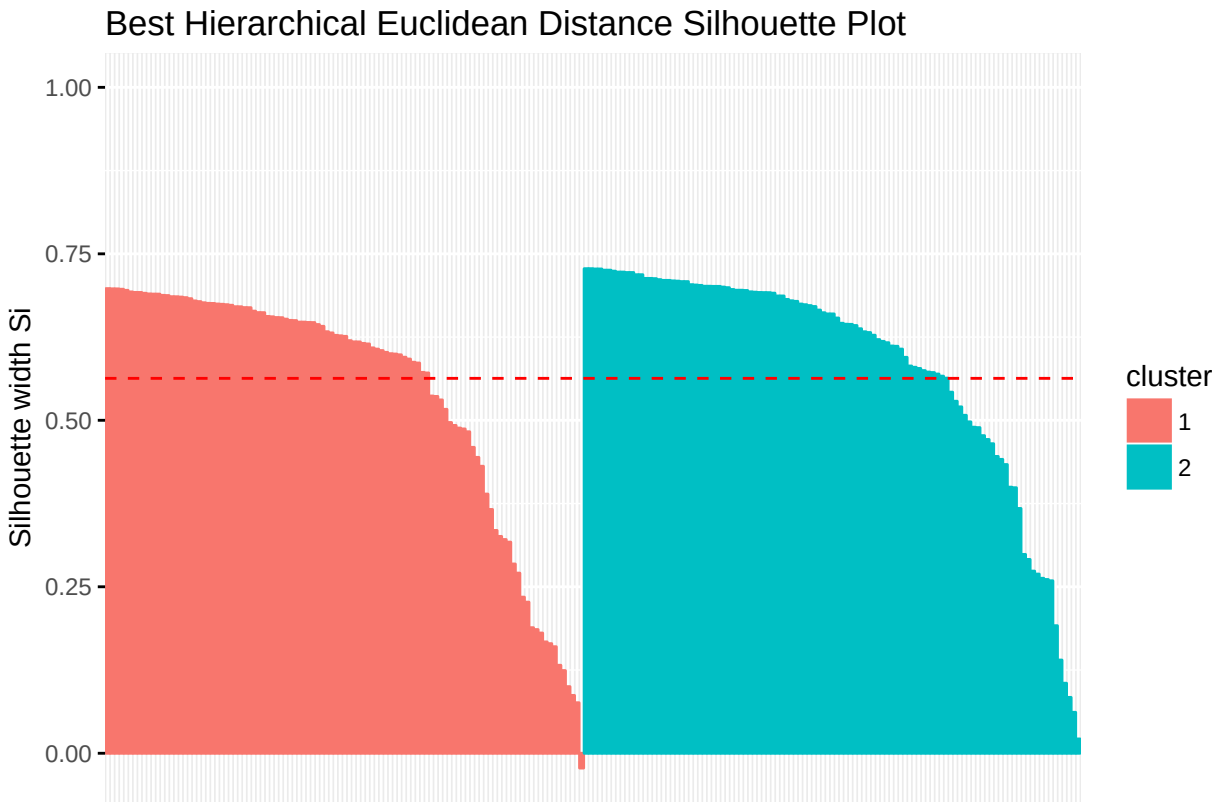
Initially we perform hierarchical clustering using single, average, and complete methods, the distances euclidean, manhattan, and gower, and on 2-8 potential clusters. We examine our best results and notice that it is 2 clusters with Euclidean distance, and complete linkage.We notice our results are pretty consistent within linkage type, and distance seems to matter less than linkage type.

Our silhouette plot shows our best result, which has the highest average silhouette width. This helps measure the space between clusters, with more space being better.

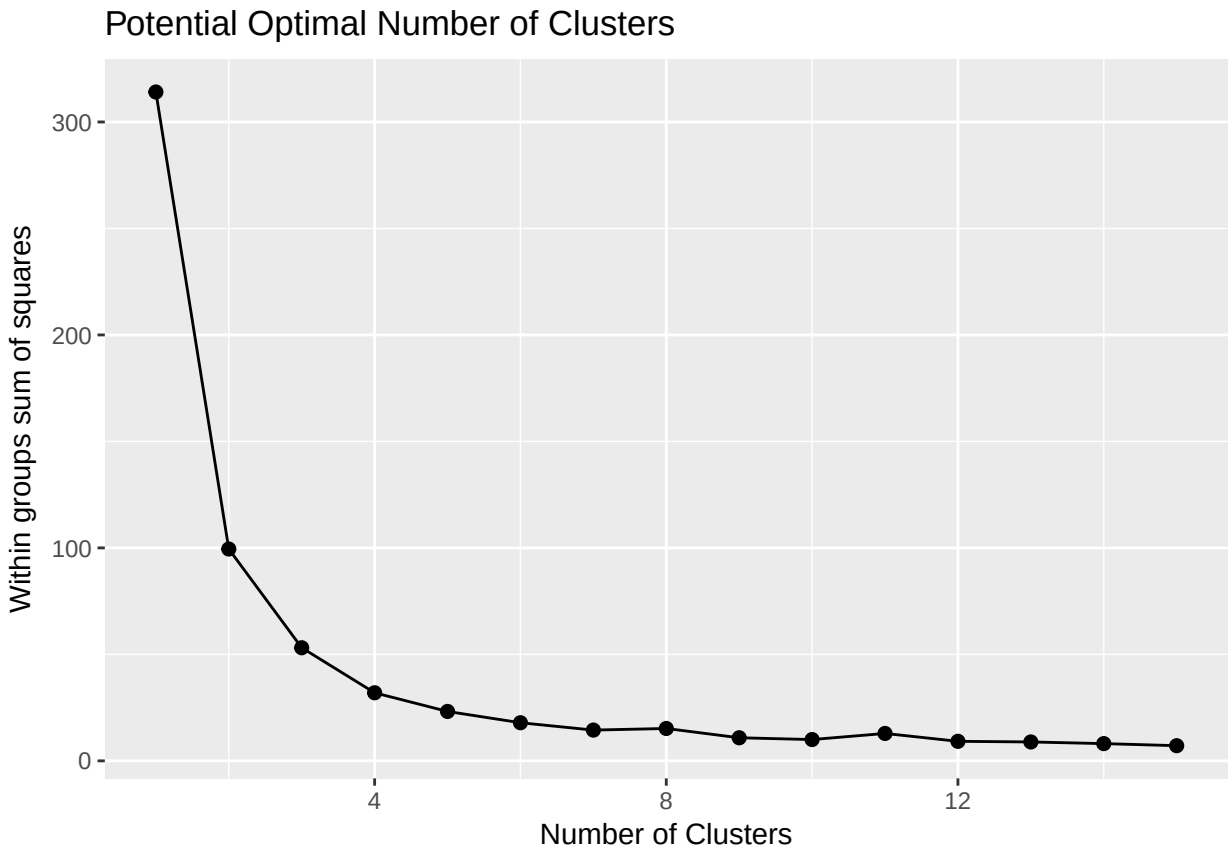



```
## mapping: colour = meth, group = meth
## geom_line: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

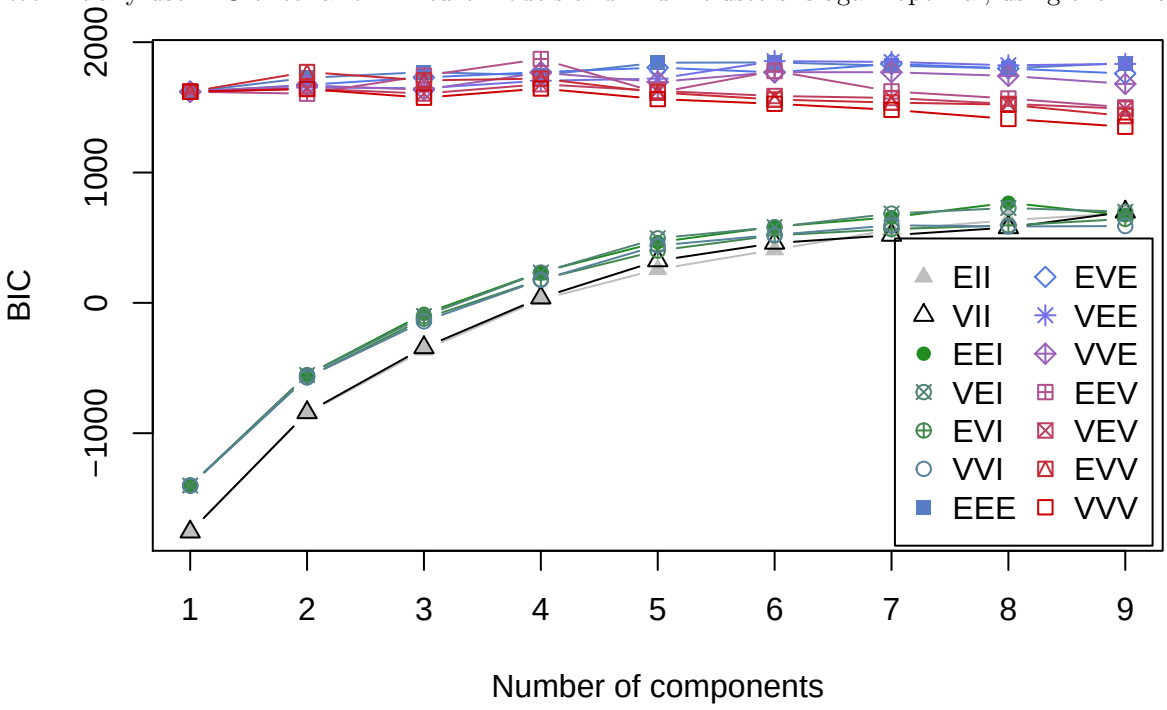
```
##   cluster size ave.sil.width
## 1       1  105          0.54
## 2       2  109          0.58
```



For K-Means we initially compute an elbow plot to determine the number of optimal clusters, which appears to be 4 (K-Means only uses Euclidean distances and we tested k=2 to 15).



We initially use BIC criteria for mixture models and find 4 clusters is again optimal, using the EEV model.

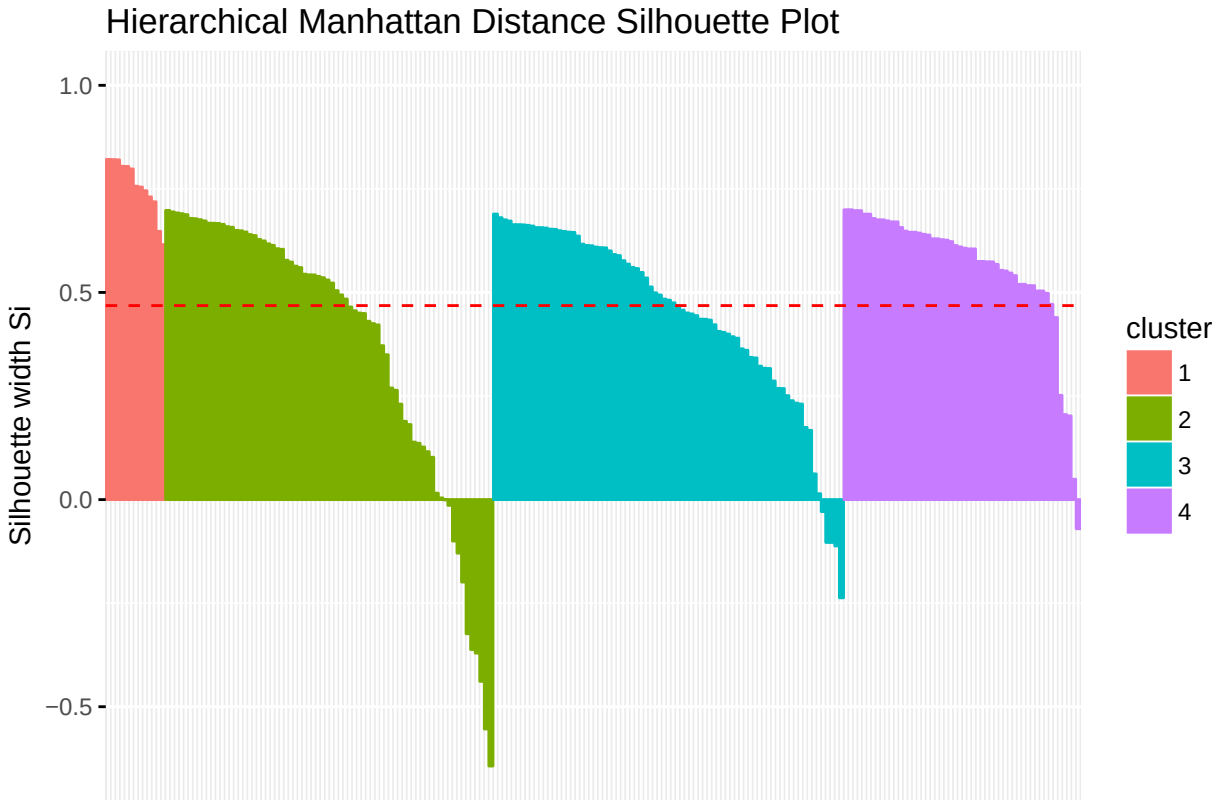



```
## Best BIC values:
##             EEV,4      VEE,6      VEE,7
## BIC      1871.011 1854.58734 1848.48505
```

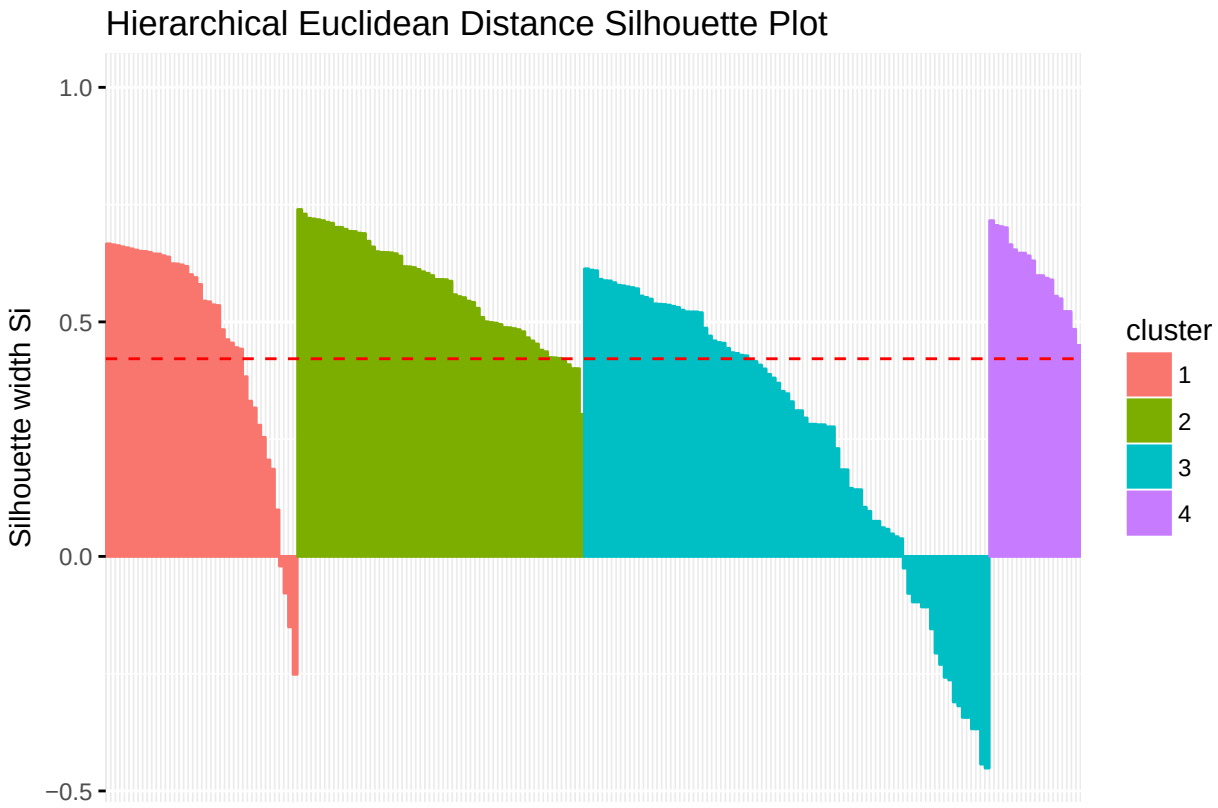
```
## BIC diff    0.000  -16.42338  -22.52567
```

Taking our BIC model from part 1, we now use 4 clusters with each of three methods to compare our results against sex and species. For hierarchical clustering we see our resulting silhouette plots differ among distances. Euclidean distance (average linkage) performs the best, followed by Manhattan (complete linkage), and then Gower (also complete). Again, higher average width being better.

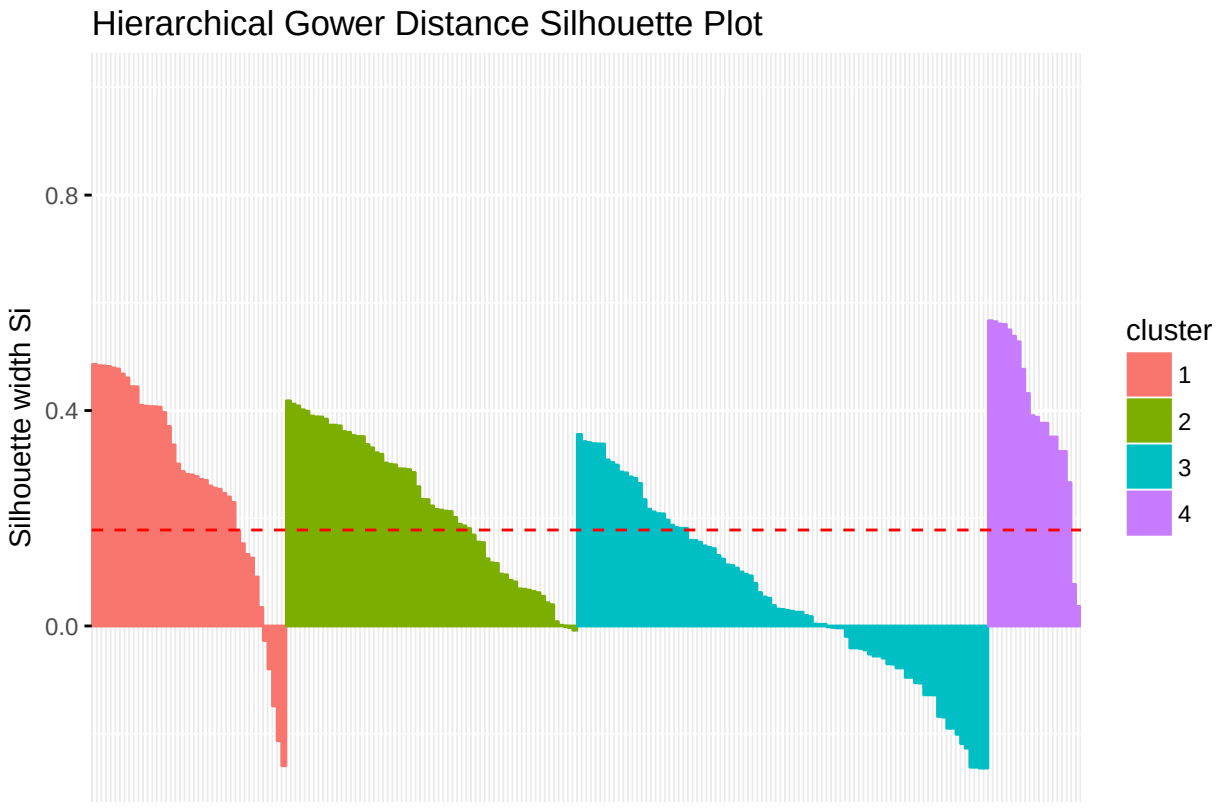
```
##   cluster size ave.sil.width
## 1       1   13          0.76
## 2       2   72          0.38
## 3       3   77          0.44
## 4       4   52          0.56
```



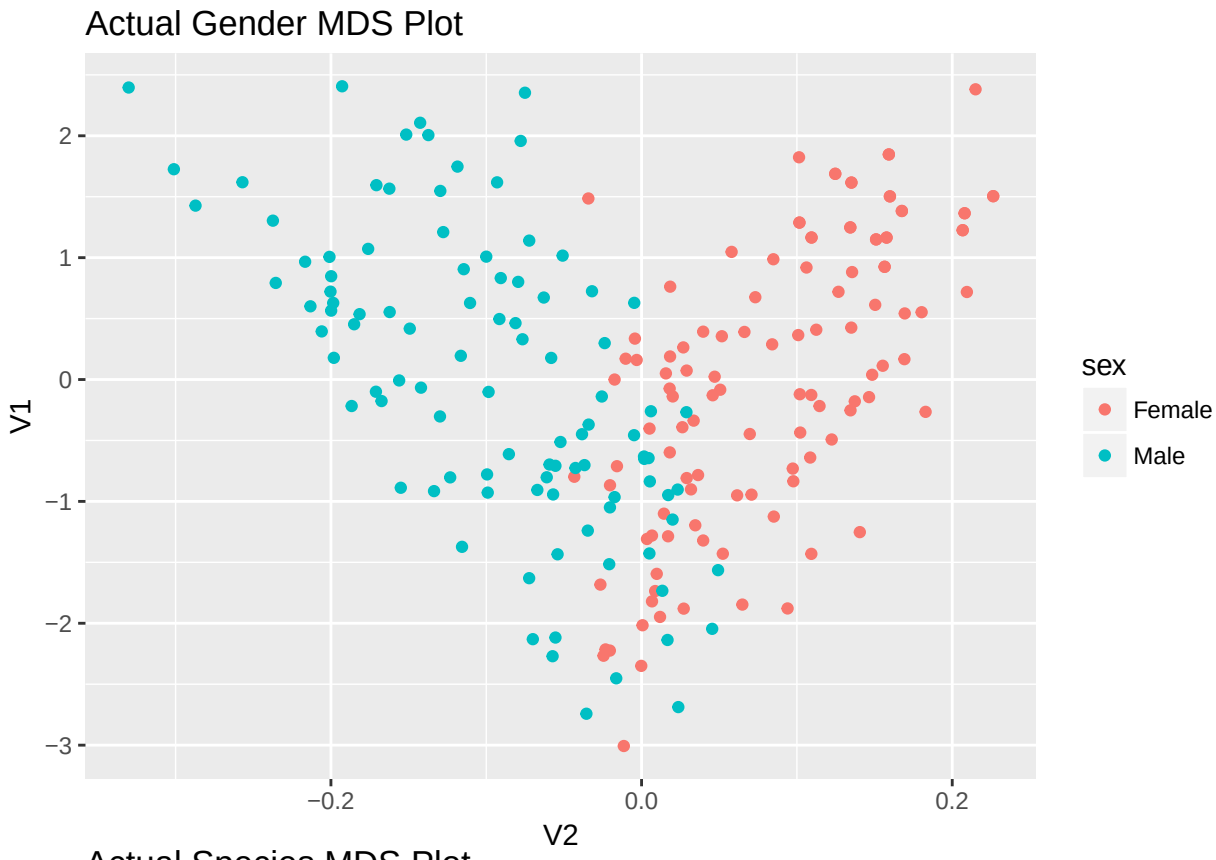

```
##   cluster size ave.sil.width
## 1       1   42          0.46
## 2       2   63          0.57
## 3       3   89          0.25
## 4       4   20          0.61
```

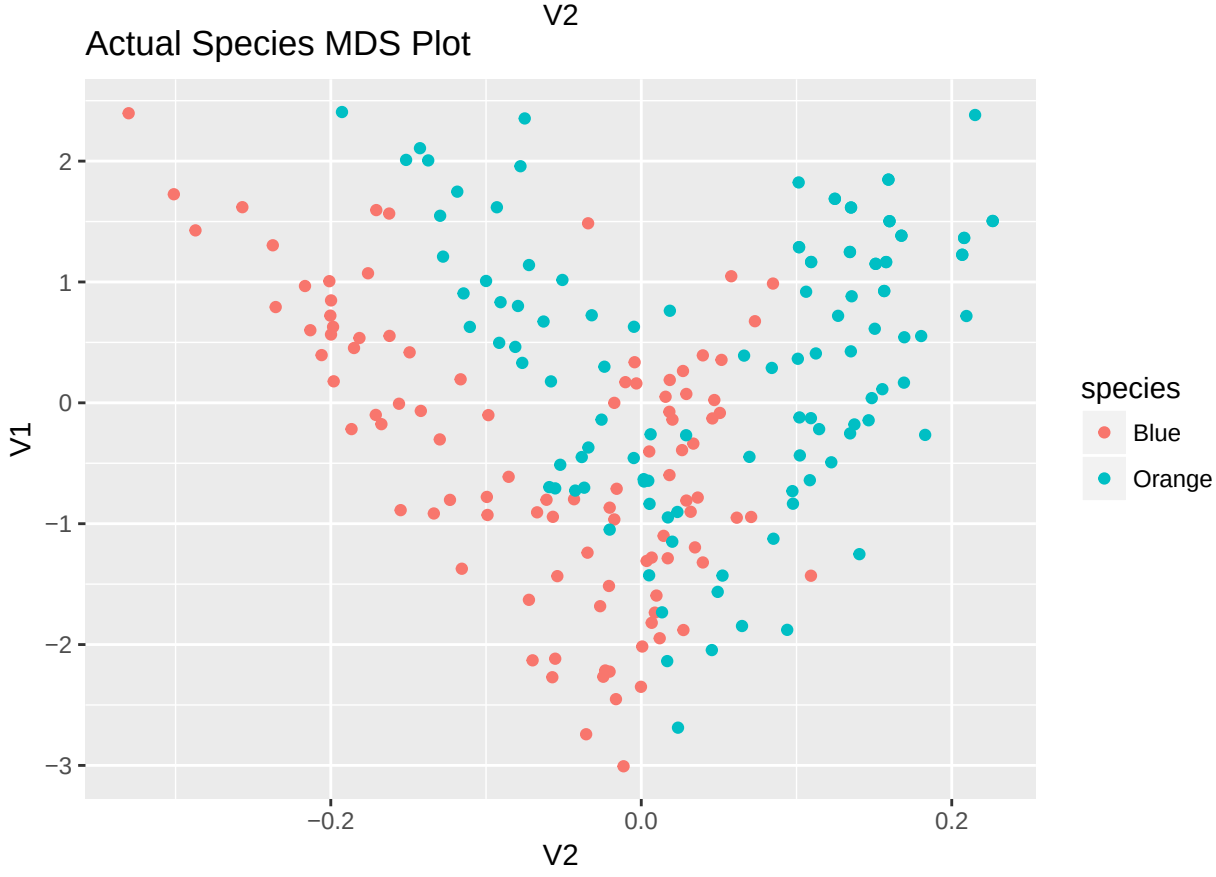


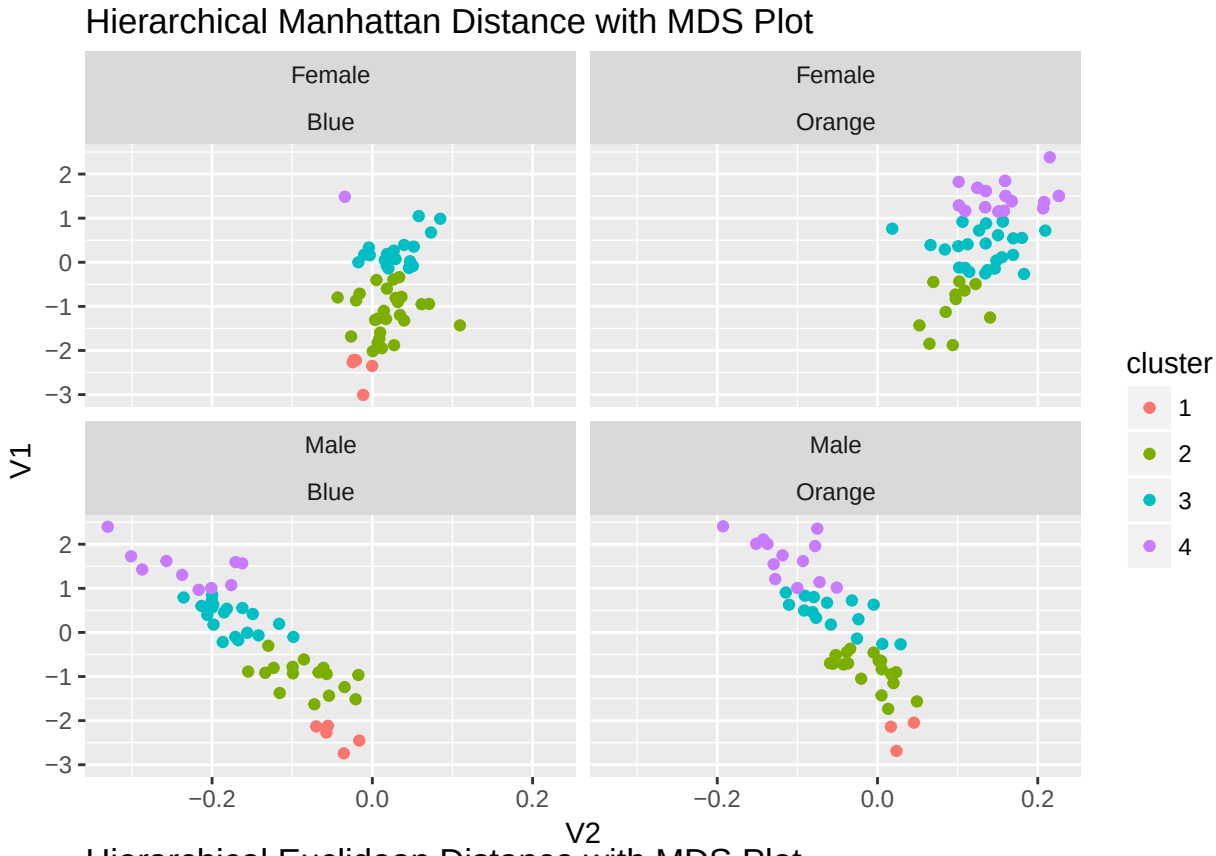

```
##   cluster size ave.sil.width
## 1       1   42          0.27
## 2       2   63          0.22
## 3       3   89          0.05
## 4       4   20          0.40
```

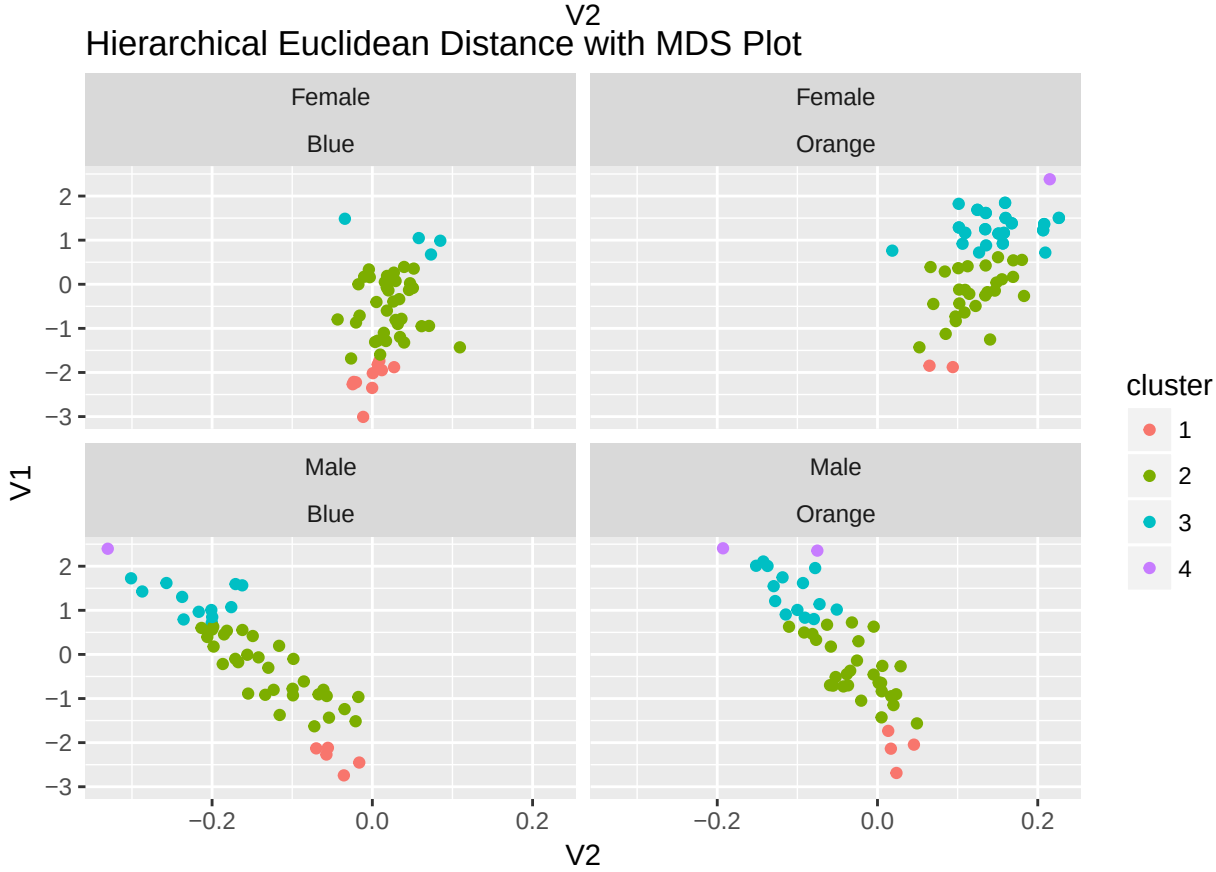


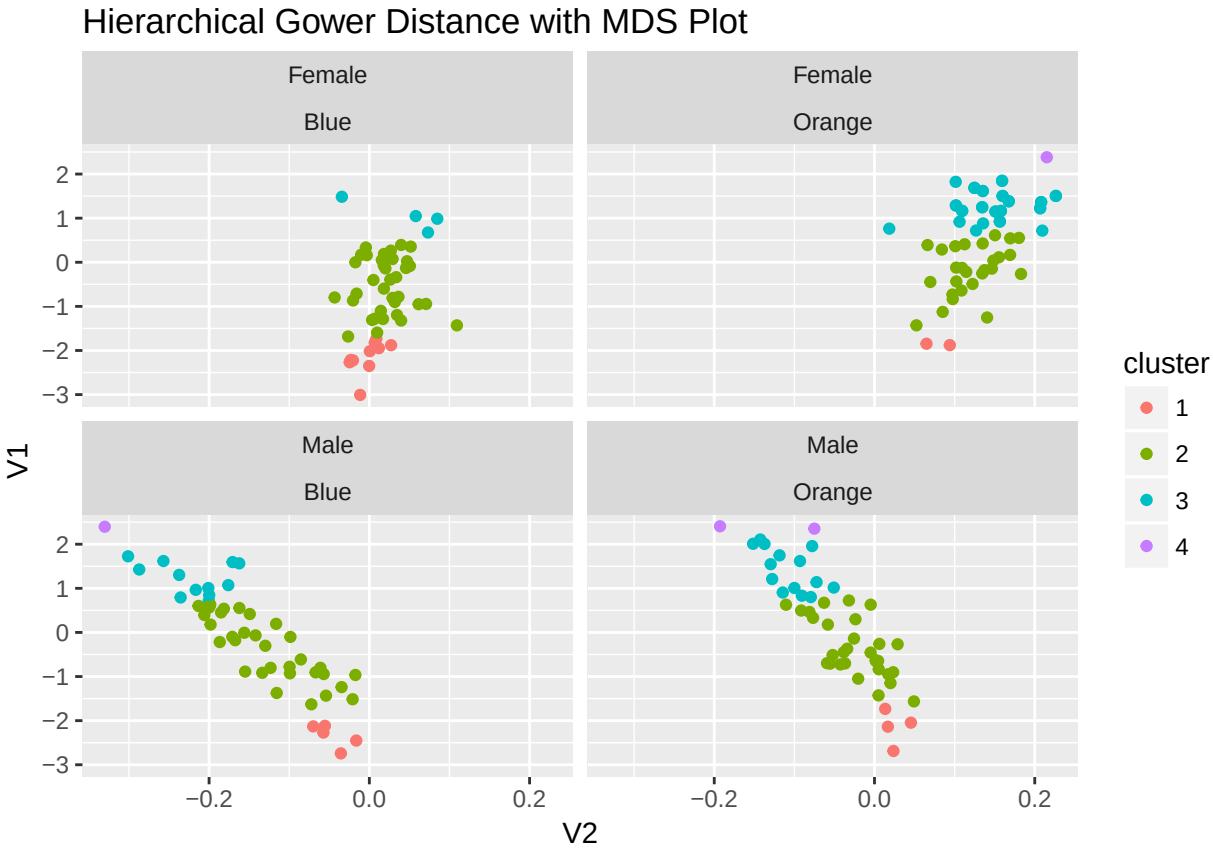
We then plot the actual species and sex variables so we can see how well our classification performs. We purposely break out our graphs by species and sex to allow us to more clearly examine our results. As we can see our results contain observations classified in 3-4 clusters, so it is clear our clustering performs poorly across all of our optimal types. Ideally there would be one color per cluster, but this is not at all the case. Overall, hierarchical clustering does poorly.


Actual Gender MDS Plot
V1
V2
2
1
0
−1
−2
−3
−0.2
0.0
0.2
sex
Female
Male


Actual Species MDS Plot
V1
V2
2
1
0
−1
−2
−3
−0.2
0.0
0.2
species
Blue
Orange

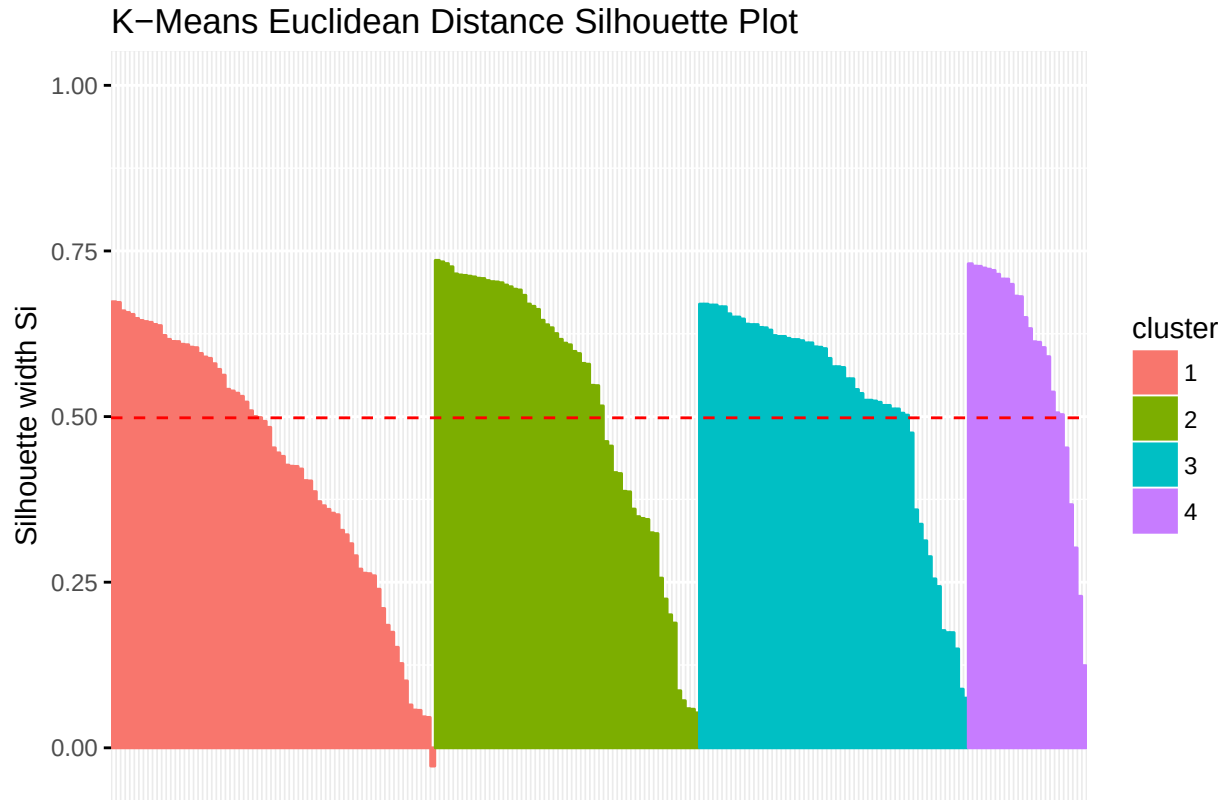

Hierarchical Manhattan Distance with MDS Plot
Female
Blue
Female
Orange
Male
Blue
Male
Orange
V1
V2
2
1
0
−1
−2
−3
−0.2
0.0
0.2
cluster
1
2
3
4

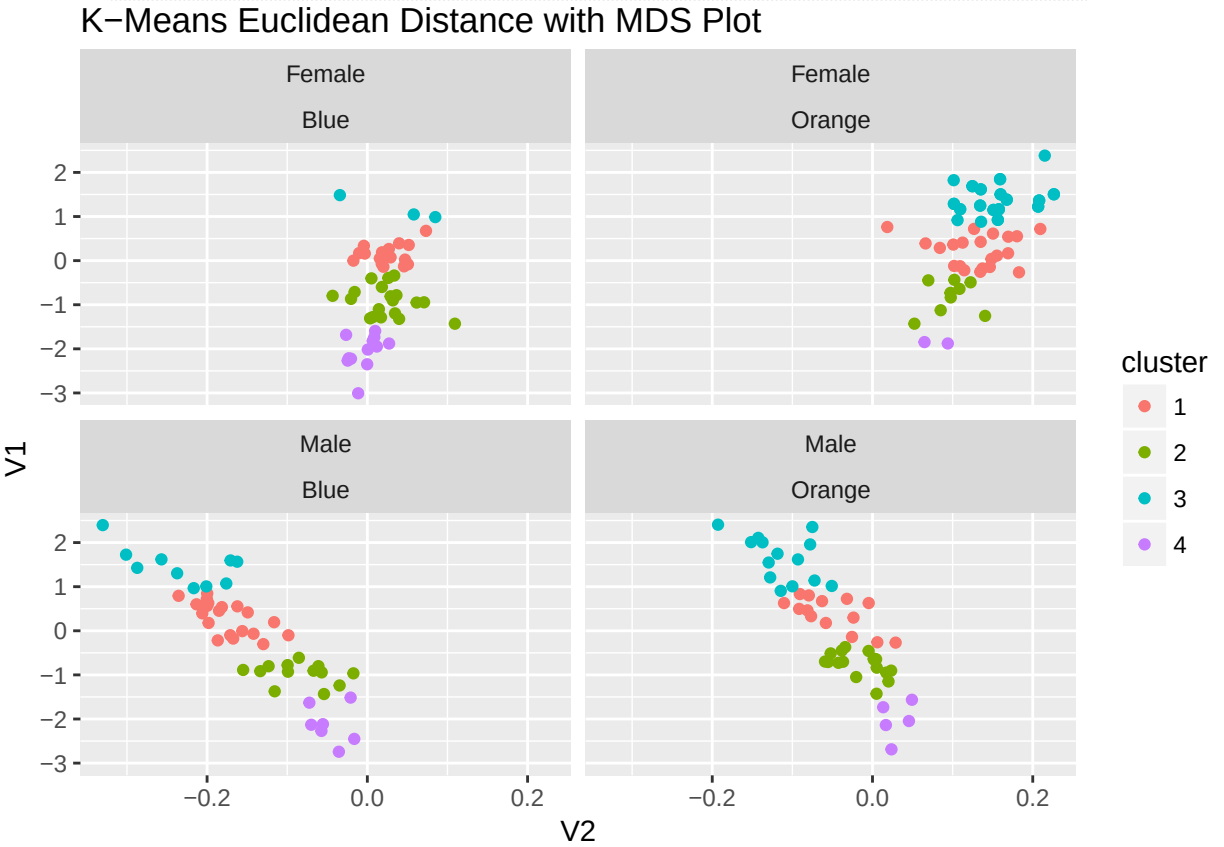

Hierarchical Euclidean Distance with MDS Plot
Female
Blue
Female
Orange
Male
Blue
Male
Orange
V1
V2
2
1
0
−1
−2
−3
−0.2
0.0
0.2
cluster
1
2
3
4



We move on to see how well our K-Means classifies by sex and species. Again we can see there is almost a perfect spread of all 4 cluster types when we expect one to be in each species, sex box. Our K-means results are just as bad as our hierarchical clustering results.

```
##   cluster size ave.sil.width
## 1       1   71          0.43
## 2       2   58          0.52
## 3       3   59          0.52
## 4       4   26          0.59
```





Finally, with mixture models using BIC, we do a decent job classifying by sex and species. We correctly classify all, but one orange female (incorrectly classifying a few orange males only), and we do slightly worse

among the blue species. For the blue species, we classify a number of females as males, but not vice versa. These results look dramatically better than the other types of clustering. This may be a result of the EM algorithm iterations which does a better job than the random assignment then iteration of K-Means or hierarchical clustering.