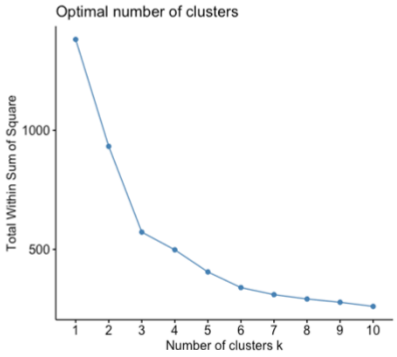
Introduction

The purpose of this report is to perform a cluster analysis on the "Byar" dataset. First, I will test the dataset to find the optimal number of clusters, and then I will apply a number of different clustering methods to the dataset in order to try and separate the patients listed into groups with similar characteristics. The methods I will be employing are k-medoids, hierarchical clustering and clustMD. I will then compare the three different methods and the clusters that each produces.

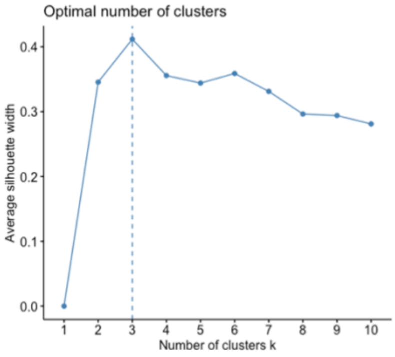
Methods

Testing for Optimal Number of Clusters

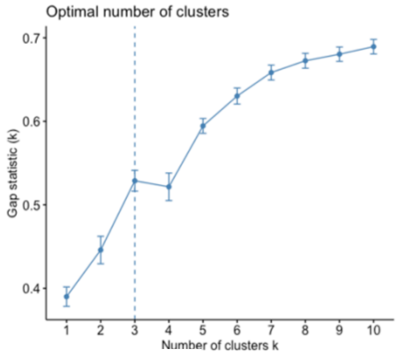
To test for the optimal number of clusters to choose for this dataset, I employed the use of three different methods; the Elbow Method, Silhouette Method and Gap Statistic. The three methods each yielded the same results, pointing at three groups as the optimal number of clusters. This can be seen from the graphs below, the Elbow method shows a kink at 3 clusters and the Silhouette and Gap Statistic also pointed to three clusters.



Elbow Method



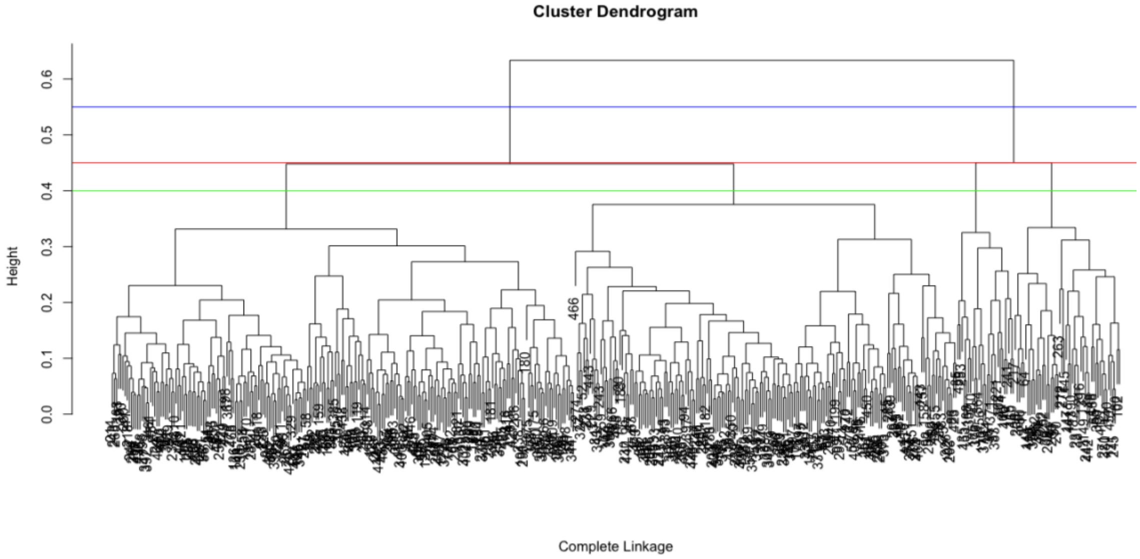
Silhouette Method



Gap Statistic

Hierarchical Clustering

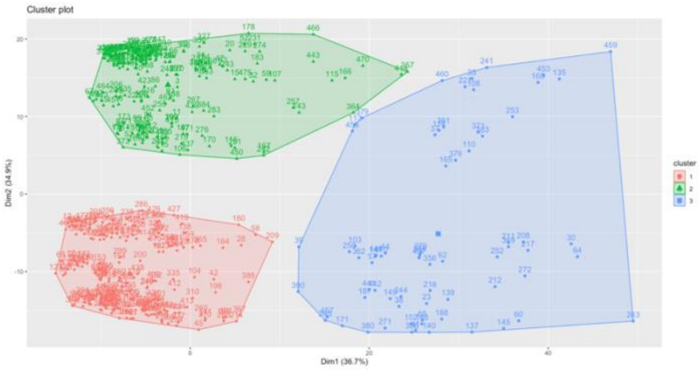
For Hierarchical Clustering, I used Gower's distance method to allow for the fact that this is a mixed variable dataset i.e. not all variables are continuous. I also used complete linkage as it is more robust to outliers, and when compared with average and single linkage it performed much better. The



cluster dendrogram for complete linkage can be seen above. Average and single linkage both produced dendrograms with clusters that were very high up the tree and contained only one observation. Complete linkage however was a much cleaner graph, as shown above. The blue, red and green lines show the possible cuts that could made to the tree to separate into 2, 3 and 4 clusters respectively.

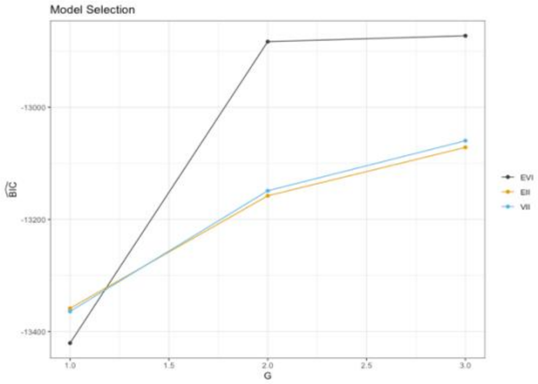
K-Medoids

For K-Medoids, I used the “pam” function in R, and set the number of clusters to three based on the testing I completed earlier. The result from this clustering can be seen below, where the red green and blue sections of the graph represent the three different clusters. As we can see, the red cluster, number 1, has very little spread and all observations appear quite close together, whereas the blue cluster, number 3, has a much larger spread.



ClustMD

To fit the ClustMD model, I first used the clustMDparallel function which ran each of the different possible types of model. As can be seen from the adjacent graph, the model with the largest BIC value is an “EVI” model with a G-score of 3. Hence, I ran this model individually using the clustMD function. This resulted in 3 clusters being produced which I then compared to the original groupings.

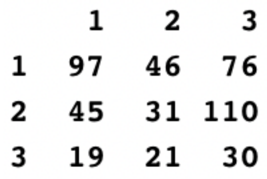


Results

The results of these three analyses showed that clustering analyses would not separate this dataset into groups similar to those created by their EKG values. Firstly, the K-medoids and ClustMD methods produced very similar sized clusters, and the Hierarchical method was much different to these. For the Hierarchical method, I cut the tree at two, three and four clusters as each of these was possible, however no matter which of these I used the resulting clusters did not resemble the EKG groups 1,2 and 3.

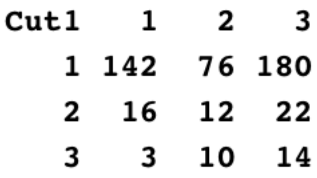
Comparisons can be seen below of the EKG groups and the clusters created by each method. As we can see, there is almost no similarity in each case. From this we can gather that the variables in this dataset may not bear much relevance on the EKG values, as when we use traditional clustering methods based on these variables, very different groups are created.

K-Medoids



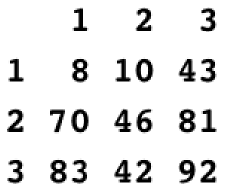
```
     1    2    3
1   97   46   76
2   45   31  110
3   19   21   30
```

Hierarchical



```
Cut1    1    2    3
   1  142   76  180
   2   16   12   22
   3    3   10   14
```

ClustMD



```
    1   2   3
1   8  10  43
2  70  46  81
3  83  42  92
```