MSCI 433 Assignment 4

Q1:

Q1.a.i ) The reason it will take a large amount of time to build a hierarchal clustering model is due to the fact that we have a large amount of observations which will be to great for a hierarchal cluster model.

Q1.a.ii) I would recommend a cluster size of 6 because you would like to not overwhelm the reader too many category recommendations.

Chart

Description automatically generated

Q1.a.iii)

Each cluster has the following number of observations

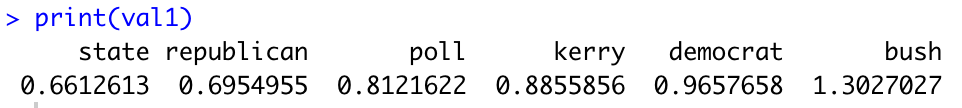
|  |  |
| --- | --- |
| Cluster | # of observations |
| 1 | 2220 |
| 2 | 436 |
| 3 | 324 |
| 4 | 166 |
| 5 | 228 |
| 6 | 56 |

Table

Description automatically generated

Q1.a.iv) Description of each cluster. Yes there is a cluster mostly about the Iraq war which is cluster 5. Yes there is a cluster mostly about the democratic party which is cluster 4.

Cluster 1:



Cluster 2:

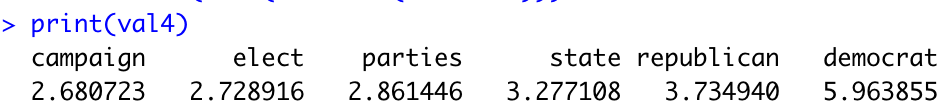
Text, letter

Description automatically generated

Cluster 3:



Cluster 4:

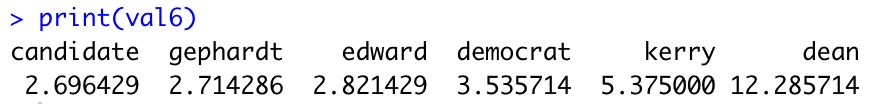


Cluster 5:

Text

Description automatically generated

Cluster6:



Q1.b)

Q1.b.i) The answer is not exactly the same as in hierarchal clustering. This is because hierarchal clustering and k means clustering utilize different methods to produce the clusters, also because the clustering occurs across 2 different instances the cluster sizes may not be exactly the same.

Table

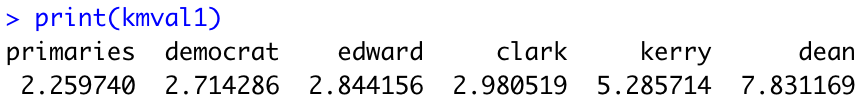
Description automatically generated

Q1.b.ii)

The clusters that are the same across the 2 models are: cluster 1 in the hierarchal and cluster 5 in the kmeans is similar where they have the same top 4 words.

cluster 6 in the hierarchal and cluster 1 in the kmeans is similar where they have the same top 4 words. Yes there are similarities between clusters such as the examples I provided above.

Kmcluster 1:

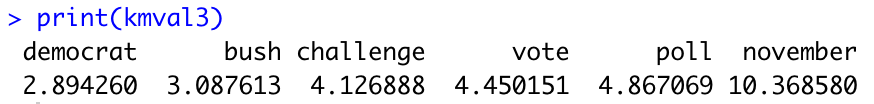


Kmcluster2:

Text

Description automatically generated

Kmcluster3:

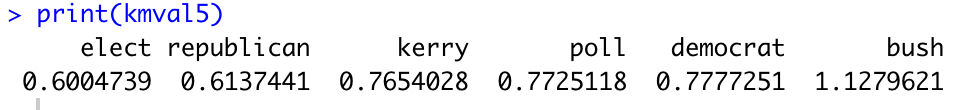


Kmcluster4:

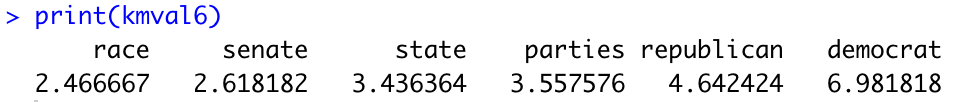
Text

Description automatically generated

Kmcluster5:



Kmcluster6:



Q1.c)

The results are similar between the 2 sections this might be due to the fact that the cluster size difference between the 2 sections is small. I do not have direct preference in cluster size.

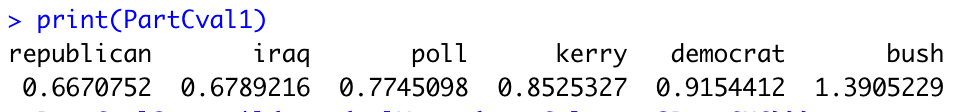
Using 4 clusters in hierarchical clustering

|  |  |
| --- | --- |
| Cluster | # of observations |
| 1 | 2448 |
| 2 | 436 |
| 3 | 324 |
| 4 | 222 |

Table

Description automatically generated

hierarchical cluster 1:



hierarchical cluster 2:

Text

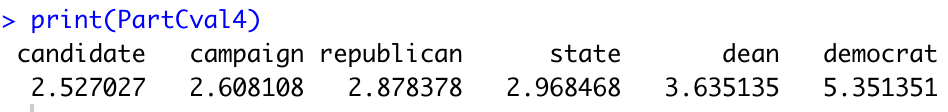
Description automatically generated

hierarchical cluster 3:

Text

Description automatically generated

hierarchical cluster 4:



Using 4 clusters in kmeans clustering

Table

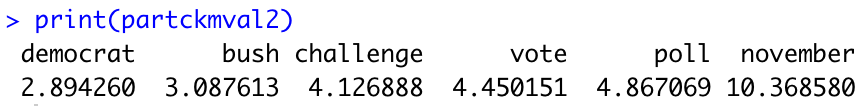
Description automatically generated

Kmcluster1:

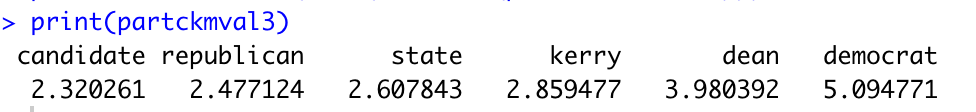
Letter

Description automatically generated with medium confidence

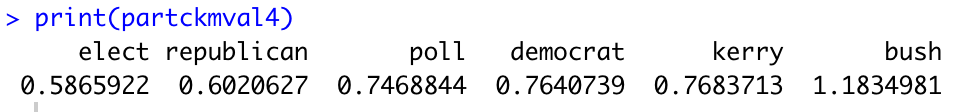
Kmcluster2:



Kmcluster3:



Kmcluster4:



#############################################################################

Q2)

Q2.a.i) The reason it is important to normalize data is to ensure that the dimensions and features during clustering are treated uniformly. Furthermore it is important to normalize data because we need to remove duplicated data the can lead to inaccurate results. What could occur if we do not normalize the data the clusters will be based of off very similar data points and therefore it will have a large amount of bias.

Q2.a.ii)

Table

Description automatically generated

Q2.b.i)

Table

Description automatically generated

Q2.b.ii)

Q2.b.iii)

Q2.b.iv)

Q2.c.i)

Table

Description automatically generated

Q2.c.ii)

Q2.c.iii)

I think the reduced number of clusters has lead to overfitting on some of the data points as some of the data points are added to clusters that may not be the most accurate or relevant and this is due to the reduced cluster size.

#############################################################################

Q3)

Q3.a)

Q3.a.i) The age bracket with the most is 20-24, while the age bracket with the least is 60-65

Chart, histogram

Description automatically generated

Default histogram

Chart, histogram

Description automatically generated

Histogram with binwidht=5

Q3.a.ii) The argument color blue adds a blue outline to the histogram bars

Chart, histogram

Description automatically generated

Histogram with binwidht=5 and colour blue



Q3.b.i) The insights gained from this data is on average there are more males than females in respective age brackets.

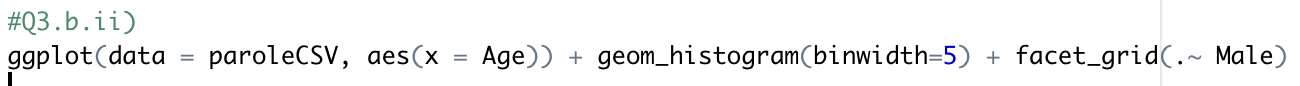
Text

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Q3.b.ii) This argument makes the histograms appear side by side



Chart, histogram

Description automatically generated

Q3.b.iii)



Chart, histogram

Description automatically generated

Q3.c)



Chart, histogram

Description automatically generated

Q3.c.i) The most common time served is from 4 to 5 months

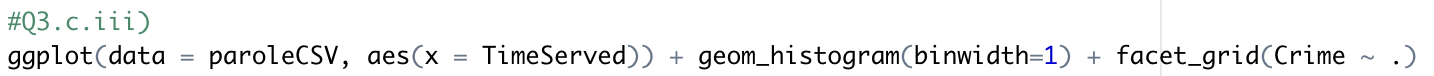
Q3.c.ii) The most common time served is from 3 to 3.10



Chart, histogram

Description automatically generated

Q3.c.iii) The average time severed for other crimes is the highest. The time served for drugs is on average higher than the time served for driving.



Chart, histogram

Description automatically generated

Q3.c.iv) Faced histograms are better than overlay histograms. The reason is it is visually much easier to differentiate and identify the values when each respective histogram is separated in its own chart

Chart, bar chart

Description automatically generated