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

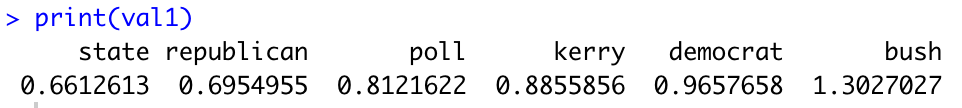
|  |  |
| --- | --- |
| Cluster | # of observations |
| 1 | 2220 |
| 2 | 436 |
| 3 | 324 |
| 4 | 166 |
| 5 | 228 |
| 6 | 56 |

Table

Description automatically generated

Q1.a.iv) Description of each cluster

Cluster 1:



Cluster 2:

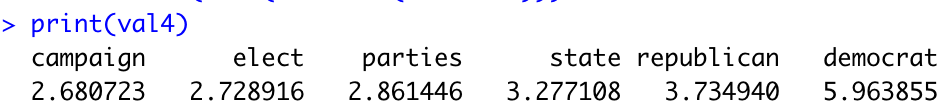
Text, letter

Description automatically generated

Cluster 3:



Cluster 4:

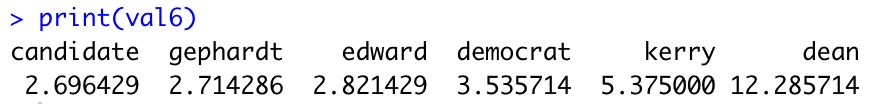


Cluster 5:

Text

Description automatically generated

Cluster6:



Q1.b)

Q1.b.i) The answer is not exactly the same as in hierarchal clustering.

Table

Description automatically generated

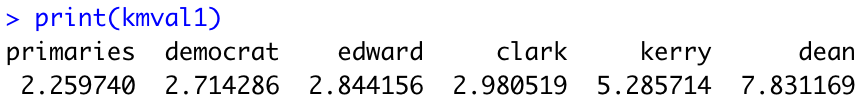
Q1.b.ii)

The clusters that are the same across the 2 models are:

Kmcluster 1:

Text, letter

Description automatically generated

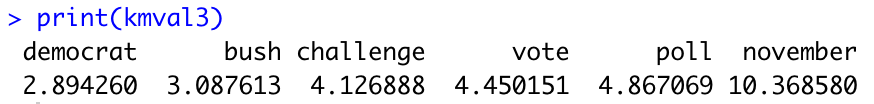


Kmcluster2:

Text

Description automatically generated

Kmcluster3:

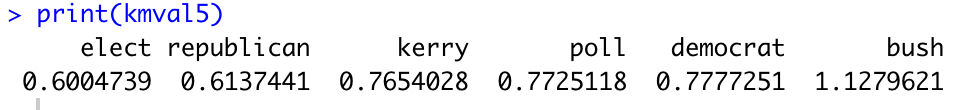


Kmcluster4:

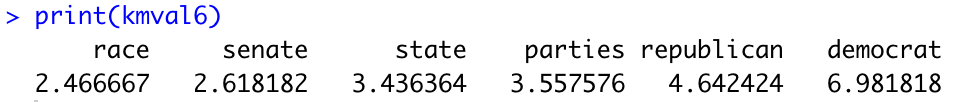
Text

Description automatically generated

Kmcluster5:



Kmcluster6:



Q1.c)

The results

Using 4 clusters in hierarchical clustering

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

Table

Description automatically generated

Using 4 clusters in kmeans clustering

Table

Description automatically generated

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

Q2)

Q2.a.i) The reason it is important to normalize data in

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)

Q2.c.ii)

Q2.c.iii)

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

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.

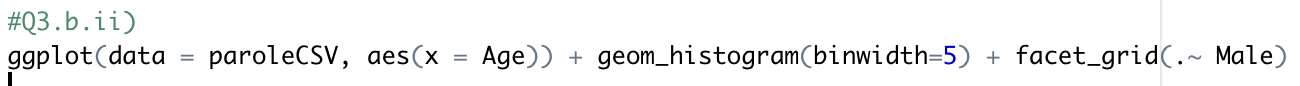
Text

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

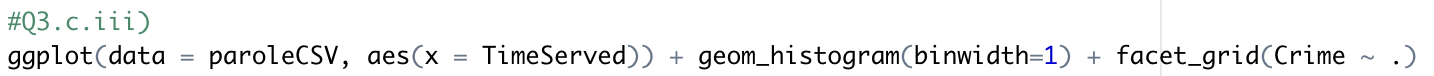
Q3.c.ii) The most common time served is from 3 to 3.10



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, histogram

Description automatically generated