Cluster Analysis on votes.repub dataset

Nathacia Nathacia

2023-02-13

Contents

```
library(cluster)
library(factoextra)
```

Loading necessary packages

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(dbscan)
```

I began the script by selecting the necessary packages for the analyses.

Cluster Analysis

votes.repub dataset

Importing dataset

```
data(package='cluster')
data("votes.repub")
View(votes.repub)
```

I started by importing the votes.repub dataset from the cluster package.

Investigating data

```
str(votes.repub)
```

```
## ’data.frame':    50 obs. of  31 variables:
##  $ X1856: num  NA NA NA NA 18.8 ...
##  $ X1860: num  NA NA NA NA 33 ...
##  $ X1864: num  NA NA NA NA 58.6 ...
##  $ X1868: num  51.4 NA NA 53.7 50.2 ...
##  $ X1872: num  53.2 NA NA 52.2 56.4 ...
##  $ X1876: num  40 NA NA 39.9 50.9 ...
##  $ X1880: num  37 NA NA 39.5 48.9 ...
##  $ X1884: num  38.4 NA NA 40.5 52.1 ...
##  $ X1888: num  32.3 NA NA 38.1 50 ...
##  $ X1892: num  3.95 NA NA 32.01 43.76 ...
##  $ X1896: num  28.1 NA NA 25.1 49.1 ...
##  $ X1900: num  34.7 NA NA 35 54.5 ...
##  $ X1904: num  20.6 NA NA 40.2 61.9 ...
##  $ X1908: num  24.4 NA NA 37.3 55.5 ...
##  $ X1912: num  8.26 NA 12.74 19.73 0.58 ...
##  $ X1916: num  22 NA 35.4 28 46.3 ...
##  $ X1920: num  31 NA 55.4 38.7 66.2 ...
##  $ X1924: num  27 NA 41.3 29.3 57.2 ...
##  $ X1928: num  48.5 NA 57.6 39.3 64.7 ...
##  $ X1932: num  14.2 NA 30.5 12.9 37.4 ...
##  $ X1936: num  12.8 NA 26.9 17.9 31.7 ...
##  $ X1940: num  14.3 NA 36 20.9 41.4 ...
##  $ X1944: num  18.2 NA 40.9 29.8 43 ...
##  $ X1948: num  19 NA 43.8 21 47.1 ...
##  $ X1952: num  35 NA 58.4 43.8 56.4 ...
##  $ X1956: num  39.4 NA 61 45.8 55.4 ...
##  $ X1960: num  41.8 50.9 55.5 43.1 50.1 ...
##  $ X1964: num  69.5 34.1 50.4 43.9 40.9 38.7 32.2 39.1 48.9 54.1 ...
##  $ X1968: num  14 45.3 54.8 30.8 47.8 50.5 44.3 45.1 40.5 30.4 ...
##  $ X1972: num  72.4 58.1 64.7 68.9 55 62.6 58.6 59.6 71.9 75 ...
##  $ X1976: num  43.5 62.9 58.6 35 50.9 ...
```

```
summary(votes.repub)
```

```
##      X1856           X1860           X1864           X1868      
##  Min.   : 0.19   Min.   : 0.93   Min.   :30.17   Min.   :25.45  
##  1st Qu.:26.08   1st Qu.:33.72   1st Qu.:52.56   1st Qu.:50.24  
##  Median :47.31   Median :53.71   Median :55.89   Median :55.02  
##  Mean   :39.47   Mean   :44.59   Mean   :57.88   Mean   :54.14  
##  3rd Qu.:55.71   3rd Qu.:57.03   3rd Qu.:62.20   3rd Qu.:60.80  
##  Max.   :78.23   Max.   :75.79   Max.   :78.61   Max.   :78.57  
##  NA’s   :30      NA’s   :27      NA’s   :25      NA’s   :17     
##      X1872           X1876           X1880           X1884      
##  Min.   :40.71   Min.   :27.94   Min.   :23.95   Min.   :23.72
##  1st Qu.:52.17   1st Qu.:42.47   1st Qu.:40.11   1st Qu.:46.03
##  Median :54.60   Median :50.21   Median :49.17   Median :48.16
##  Mean   :57.15   Mean   :48.49   Mean   :47.91   Mean   :47.97
##  3rd Qu.:62.67   3rd Qu.:52.45   3rd Qu.:51.89   3rd Qu.:52.34
##  Max.   :78.26   Max.   :68.58   Max.   :69.88   Max.   :66.54
##  NA’s   :13      NA’s   :13      NA’s   :12      NA’s   :12
##      X1888           X1892           X1896           X1900
##  Min.   :17.27   Min.   : 2.64   Min.   : 7.27   Min.   : 7.04
##  1st Qu.:45.07   1st Qu.:39.00   1st Qu.:30.09   1st Qu.:43.37
##  Median :49.34   Median :45.23   Median :49.30   Median :52.43
##  Mean   :46.58   Mean   :41.05   Mean   :45.25   Mean   :47.93
##  3rd Qu.:53.22   3rd Qu.:48.25   3rd Qu.:55.89   3rd Qu.:57.12
##  Max.   :71.24   Max.   :68.10   Max.   :80.10   Max.   :75.79
##  NA’s   :12      NA’s   :8       NA’s   :6       NA’s   :6
##      X1904           X1908           X1912           X1916
##  Min.   : 4.63   Min.   : 5.97   Min.   : 0.58   Min.   : 2.43
##  1st Qu.:46.18   1st Qu.:45.55   1st Qu.:18.12   1st Qu.:36.14
##  Median :57.30   Median :52.46   Median :23.11   Median :45.52
##  Mean   :51.90   Mean   :47.82   Mean   :22.42   Mean   :40.59
##  3rd Qu.:63.73   3rd Qu.:57.47   3rd Qu.:30.39   3rd Qu.:49.80
##  Max.   :77.98   Max.   :75.12   Max.   :37.46   Max.   :62.44
##  NA’s   :6       NA’s   :5       NA’s   :3       NA’s   :2
##      X1920           X1924           X1928           X1932
##  Min.   : 3.90   Min.   : 2.21   Min.   : 8.54   Min.   : 1.89
##  1st Qu.:50.99   1st Qu.:41.63   1st Qu.:53.62   1st Qu.:30.42
##  Median :58.90   Median :49.52   Median :58.07   Median :36.59
##  Mean   :55.00   Mean   :47.42   Mean   :55.88   Mean   :34.98
##  3rd Qu.:65.86   3rd Qu.:57.87   3rd Qu.:63.84   3rd Qu.:43.52
##  Max.   :77.79   Max.   :78.22   Max.   :72.02   Max.   :57.66
##  NA’s   :2       NA’s   :2       NA’s   :2       NA’s   :2
##      X1936           X1940           X1944           X1948
##  Min.   : 1.43   Min.   : 4.19   Min.   : 4.46   Min.   : 2.62
##  1st Qu.:27.49   1st Qu.:37.20   1st Qu.:40.53   1st Qu.:40.75
##  Median :36.77   Median :45.18   Median :46.90   Median :46.13
##  Mean   :32.88   Mean   :40.38   Mean   :42.67   Mean   :41.90
##  3rd Qu.:40.22   3rd Qu.:48.11   3rd Qu.:49.43   3rd Qu.:49.56
##  Max.   :56.44   Max.   :57.41   Max.   :60.25   Max.   :61.55
##  NA’s   :2       NA’s   :2       NA’s   :2       NA’s   :2
##      X1952           X1956           X1960           X1964
##  Min.   :30.34   Min.   :24.46   Min.   :24.67   Min.   :19.10
##  1st Qu.:51.53   1st Qu.:54.24   1st Qu.:48.58   1st Qu.:34.88
##  Median :55.45   Median :57.89   Median :50.39   Median :39.80
##  Mean   :55.76   Mean   :56.13   Mean   :49.94   Mean   :41.14
##  3rd Qu.:60.63   3rd Qu.:61.13   3rd Qu.:54.42   3rd Qu.:44.48
##  Max.   :71.46   Max.   :72.18   Max.   :62.07   Max.   :87.10
##  NA’s   :2       NA’s   :2
##      X1968           X1972           X1976
##  Min.   :13.50   Min.   :45.20   Min.   :33.02
##  1st Qu.:40.58   1st Qu.:58.52   1st Qu.:46.90
##  Median :45.10   Median :62.55   Median :50.42
##  Mean   :44.13   Mean   :62.71   Mean   :50.10
##  3rd Qu.:50.58   3rd Qu.:67.67   3rd Qu.:52.87
##  Max.   :59.80   Max.   :78.20   Max.   :64.94
##
```

```
class(votes.repub)
```

```
## [1] "data.frame"
```

```
names(votes.repub)
```

```
##  [1] "X1856" "X1860" "X1864" "X1868" "X1872" "X1876" "X1880" "X1884" "X1888"
## [10] "X1892" "X1896" "X1900" "X1904" "X1908" "X1912" "X1916" "X1920" "X1924"
## [19] "X1928" "X1932" "X1936" "X1940" "X1944" "X1948" "X1952" "X1956" "X1960"
## [28] "X1964" "X1968" "X1972" "X1976"
```

I then proceeded to investigate the data to gain some insight on the data using various functions including View(), str(), summary(), class(), etc. The View() and summary() function showed a lot of NA values within the dataset, which would need to be omitted later on in the preprocessing stage. The data type is a data frame and using the pairs() function did not work as the figure margins were too large.

Preprocessing

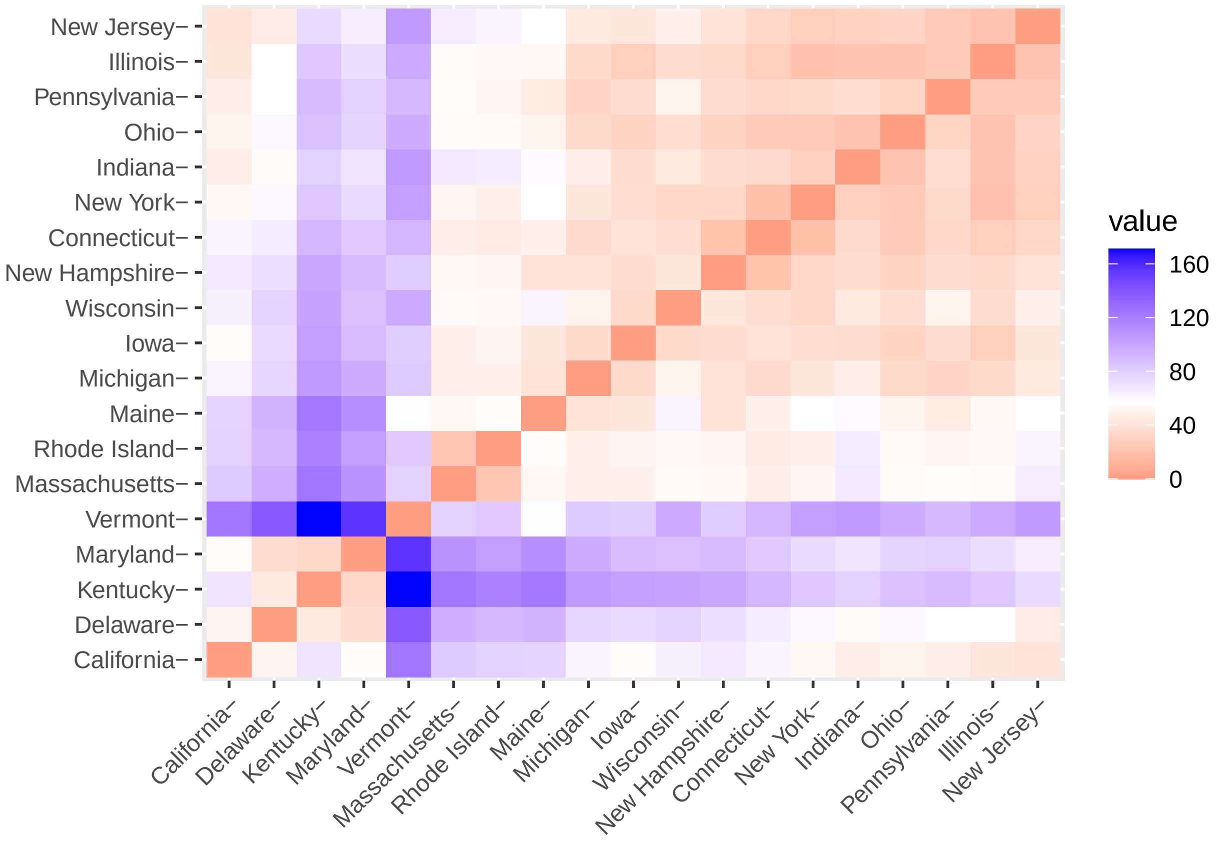
```
votesdset <- na.omit(votes.repub)
```

Following this is the preprocessing step where I used the na.omit() function on the votes.repub dataset and assigned it to a new variable called votesdset. The number of observations dropped to 19 from 50.

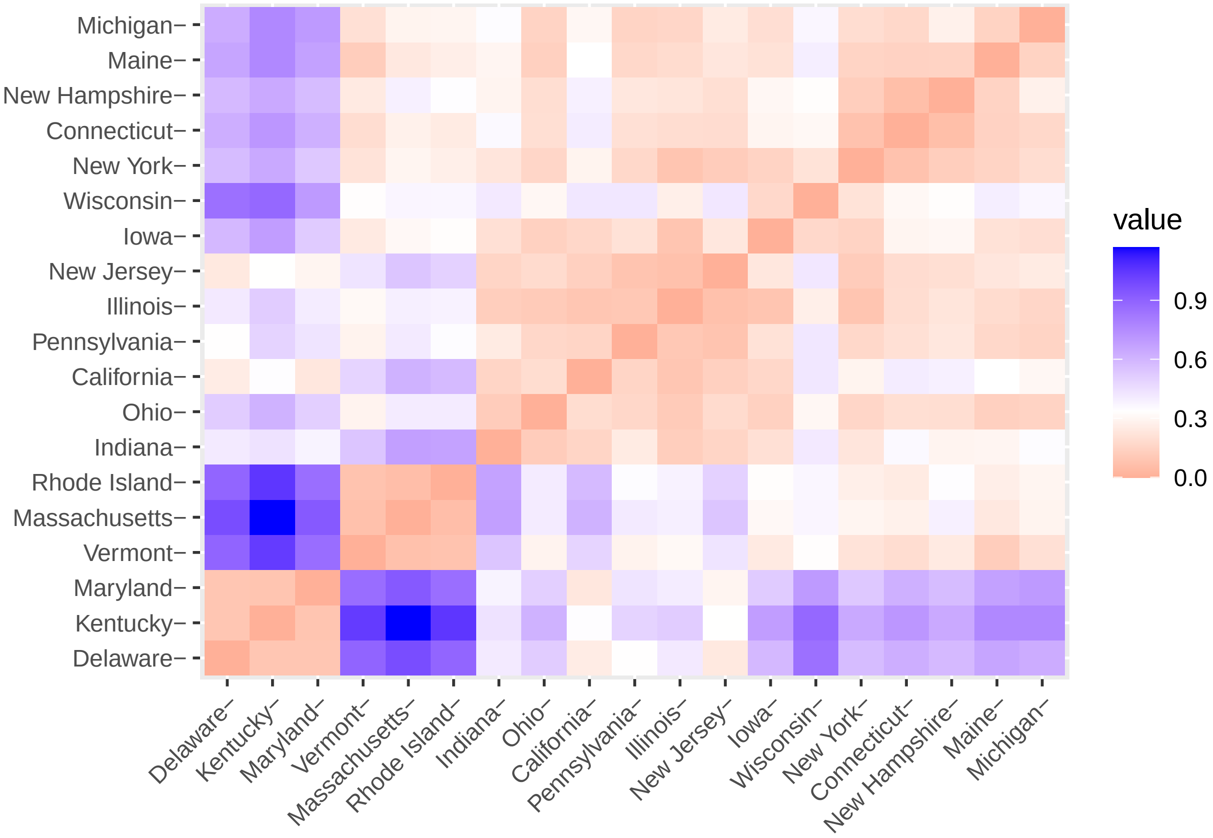
Visualizations

I then continued by visualizing the data using the function get_dist() using both the euclidean and pearson method and assigning it to a new variable called dist and dist1 respectively. By using the function fviz_dist() on the new dist variable, I was able to see this result:

```
dist <- get_dist(votesdset, method = 'euclidean')
fviz_dist(dist)
```



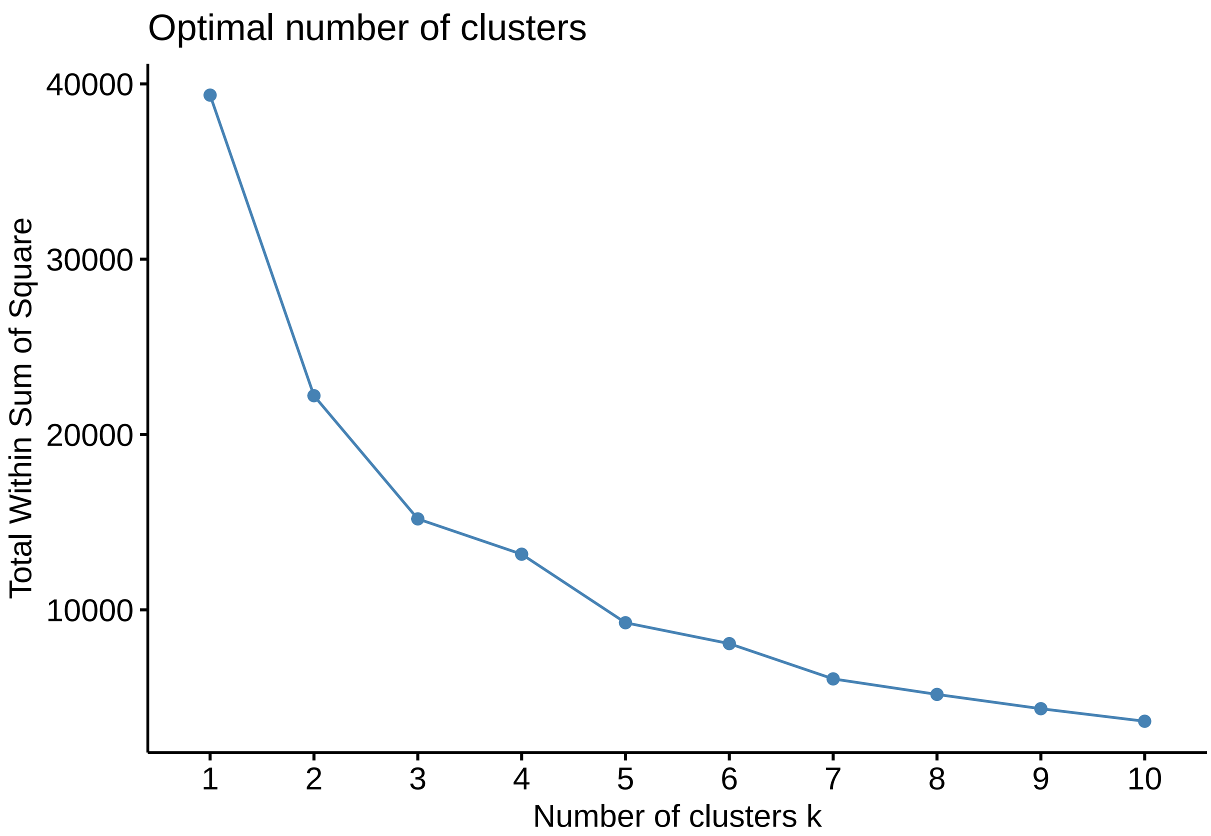
```r
dist1 <- get_dist(votesdset, method = 'pearson')
fviz_dist(dist1)
```



Kmeans clustering analysis

Following this step, I proceeded to do the kmeans portion using the function fviz_nbclust() in order to see the estimated optimal number of centroids from the dataset for k. Here I am using the within sum of squares (wss) method to see the trend of within some of squares between different clusters and data points in the dataset. There is a limitation to kmeans insofar that the initial centroids are randomly placed in the dataset based on which clusters are identified. Hence, we will be using the set.seed(123) function. Based on the results, I use the kmeans() function on the votesdset using 5 as center and 25 as nstart, and assign it to a new variable called k1. The results gave me between_SS / total_SS = 76.5 %, which means that about 76.5% of the variation in the dataset is explained by the clusters. The cluster sizes in the 5 clusters are 3, 2, 11, 1, 2. I visualized it using the code 'fviz_cluster(k1, data=votesdset)' and the cluster plot is as follows:

```r
fviz_nbclust(votesdset, kmeans, method = 'wss')
```

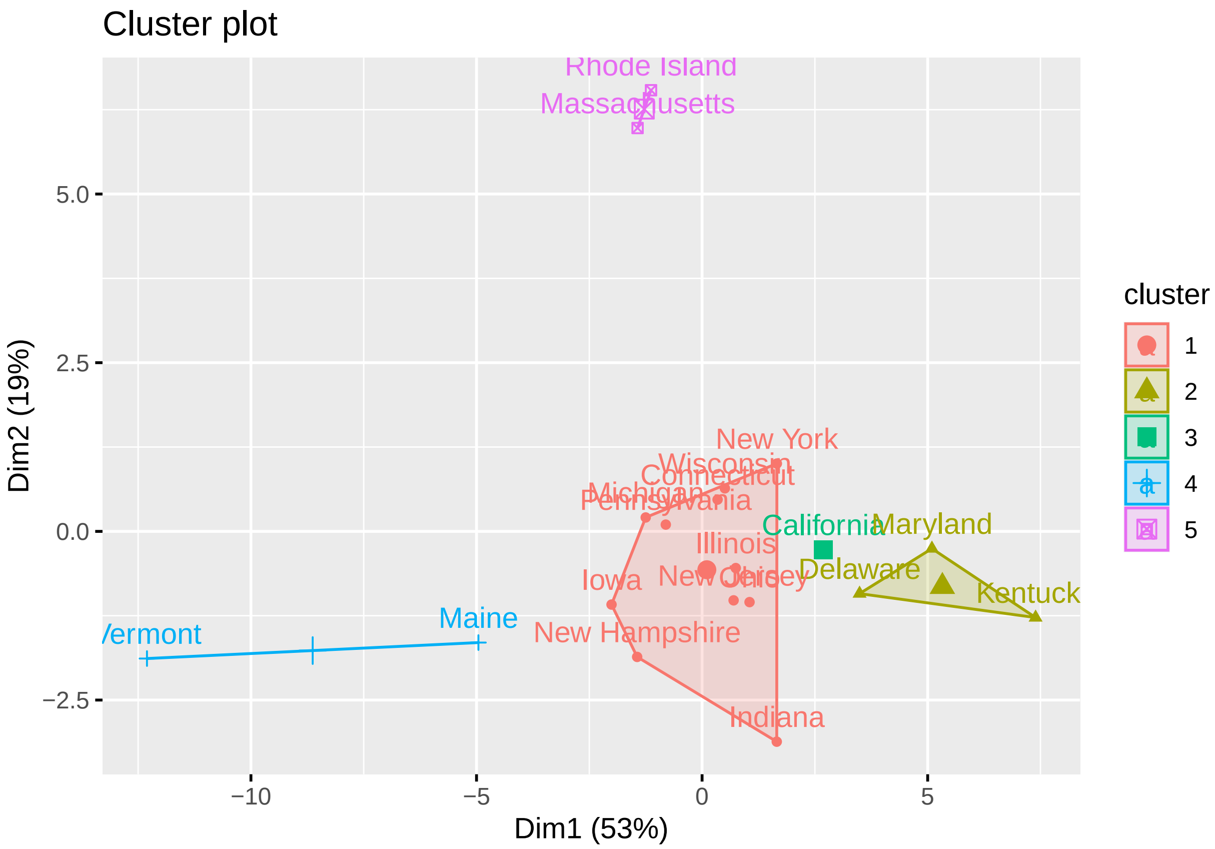


```
set.seed(123)
k1 <- kmeans(votesdset, center=5, nstart=25)
k1
```

```
## K-means clustering with 5 clusters of sizes 11, 3, 1, 2, 2
##
## Cluster means:
##        X1856    X1860    X1864    X1868    X1872    X1876    X1880    X1884
## 1 45.9590909 53.77909 53.96818 53.95273 56.33455 50.57818 51.65818 49.75455
## 2  0.8966667  9.25000 44.49000 33.07667 49.03000 42.03333 44.34333 43.88333
## 3 18.7700000 32.96000 58.63000 50.24000 56.38000 50.88000 48.92000 52.08000
## 4 69.8000000 69.97000 68.16000 70.49500 73.06000 62.65500 60.66500 60.94500
## 5 61.2850000 61.98500 67.21000 68.08000 70.59500 58.51500 60.37500 53.19500
##      X1888    X1892    X1896    X1900    X1904    X1908    X1912    X1916
## 1 49.83636 47.65000 57.97545 56.11364 60.42455 55.77364 27.41818 50.77909
## 2 45.31333 43.92333 52.15333 51.21667 49.99333 49.66000 27.33333 47.16667
## 3 49.95000 43.76000 49.13000 54.48000 61.90000 55.46000  0.58000 46.26000
## 4 64.36500 61.08000 74.00000 68.84000 72.54000 69.06000 28.80500 56.69500
## 5 53.63500 51.28000 68.84500 58.70500 59.26000 59.48500 33.73000 50.81000
##      X1920    X1924    X1928    X1932    X1936    X1940    X1944    X1948
## 1 65.11818 58.59545 59.47909 44.21455 39.85909 48.39818 49.23364 49.12364
## 2 53.35667 50.64333 60.47333 42.24667 40.13000 42.72667 46.21333 46.97333
## 3 66.24000 57.21000 64.70000 37.40000 31.70000 41.35000 42.99000 47.14000
## 4 72.39500 75.12500 67.75500 56.74500 55.96500 52.94500 54.75000 59.14500
## 5 66.26000 60.94500 49.35000 44.97500 40.97000 44.80000 44.13000 42.30500
##      X1952    X1956    X1960    X1964    X1968    X1972    X1976
## 1 57.40727 60.81909 50.93545 36.31818 46.89091 59.42727 51.16818
## 2 52.31667 56.47667 49.66000 36.53333 43.60000 61.43333 46.79333
## 3 56.39000 55.40000 50.10000 40.90000 47.80000 55.00000 50.89000
## 4 68.75500 71.52500 57.85000 32.45000 47.95000 62.10000 53.17500
## 5 52.55500 58.82000 37.96000 21.45000 32.35000 49.10000 43.08500
```

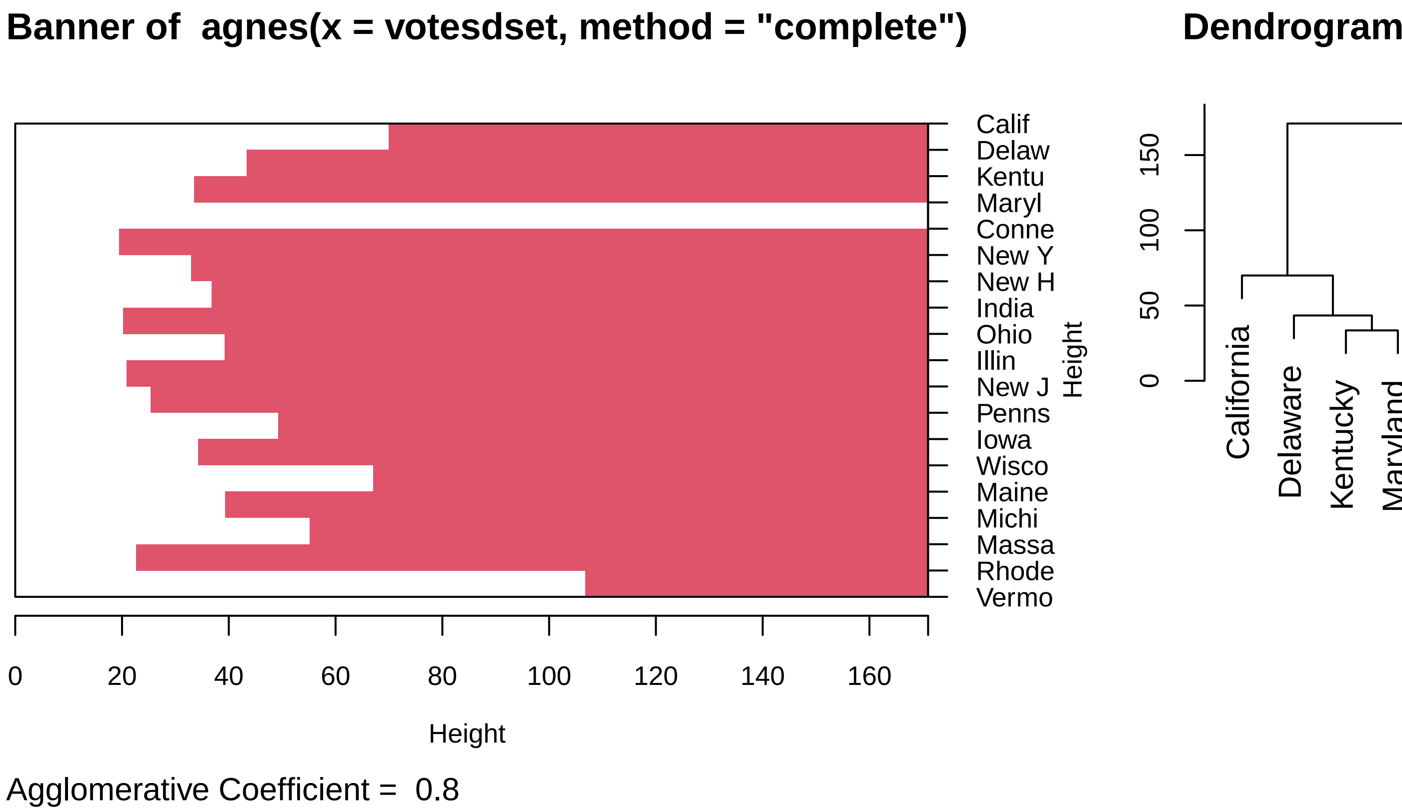
```
##
## Clustering vector:
##    California   Connecticut      Delaware      Illinois       Indiana
##             3             1             2             1             1
##          Iowa      Kentucky         Maine      Maryland Massachusetts
##             1             2             4             2             5
##      Michigan New Hampshire    New Jersey      New York          Ohio
##             1             1             1             1             1
##  Pennsylvania  Rhode Island       Vermont     Wisconsin
##             1             5             4             1
##
## Within cluster sum of squares by cluster:
## [1] 5878.0460 1445.9572    0.0000 1684.2286  256.1349
##  (between_SS / total_SS =  76.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(k1, data=votesdset)
```

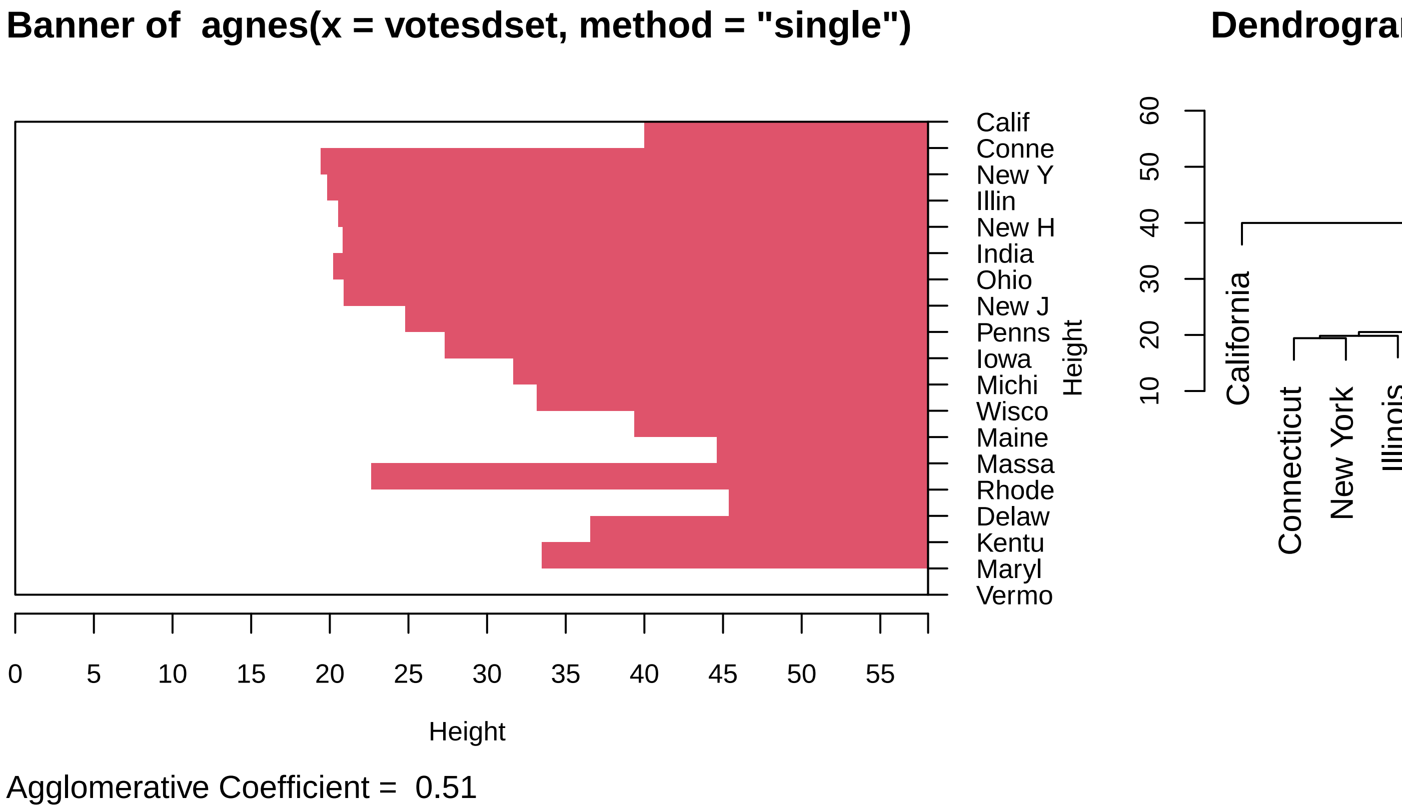


Hierarchical clustering analysis

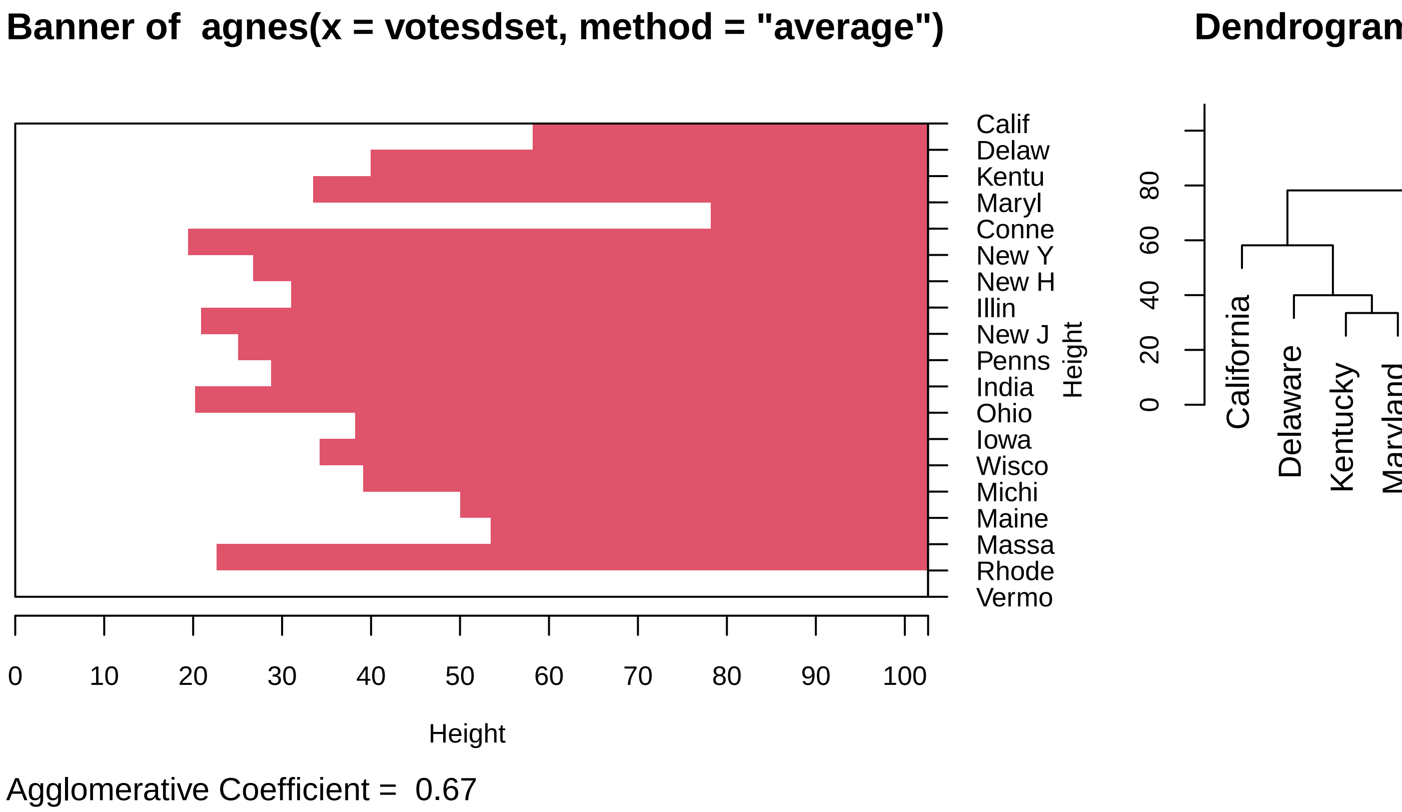
```
hc1 <- agnes(votesdset, method = 'complete')
plot(hc1, cex.axis=0.8, cex.lab=0.8)
```



```
hc2 <- agnes(votesdset, method = 'single')
plot(hc2, cex.axis=0.8, cex.lab=0.8)
```



```
hc3 <- agnes(votesdset, method = 'average')
plot(hc3, cex.axis=0.8, cex.lab=0.8)
```



```
hc1$ac
```

```
## [1] 0.7972274
```

```
hc2$ac
```

```
## [1] 0.5070565
```

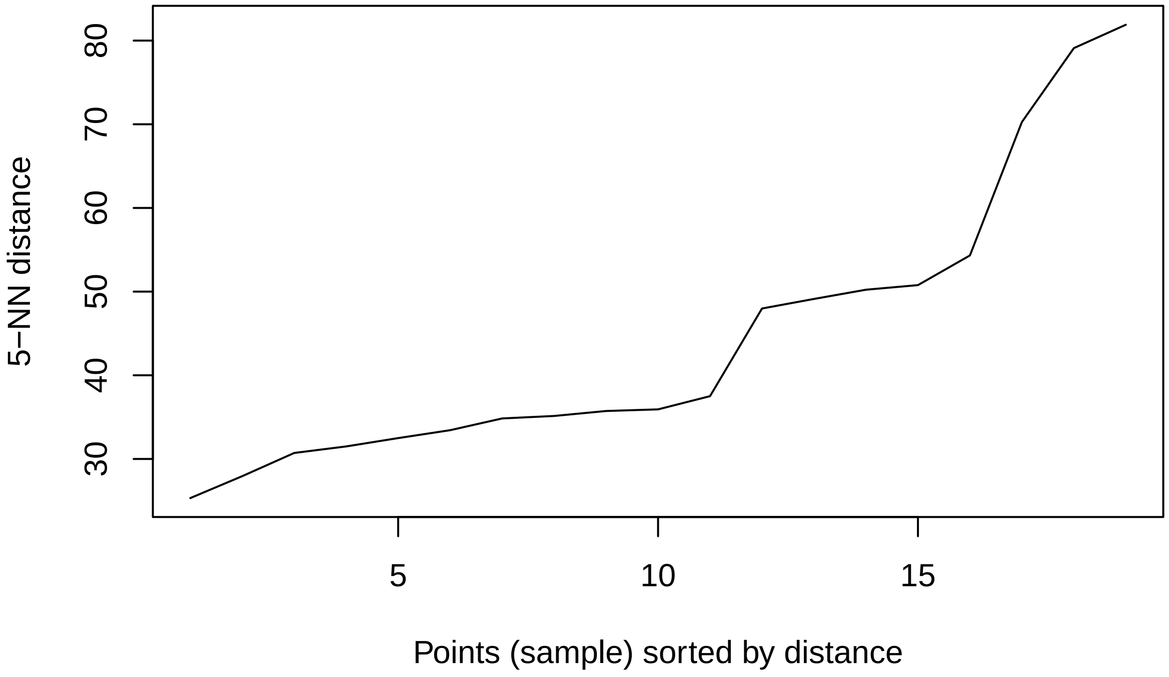
```
hc3$ac
```

```
## [1] 0.6700361
```

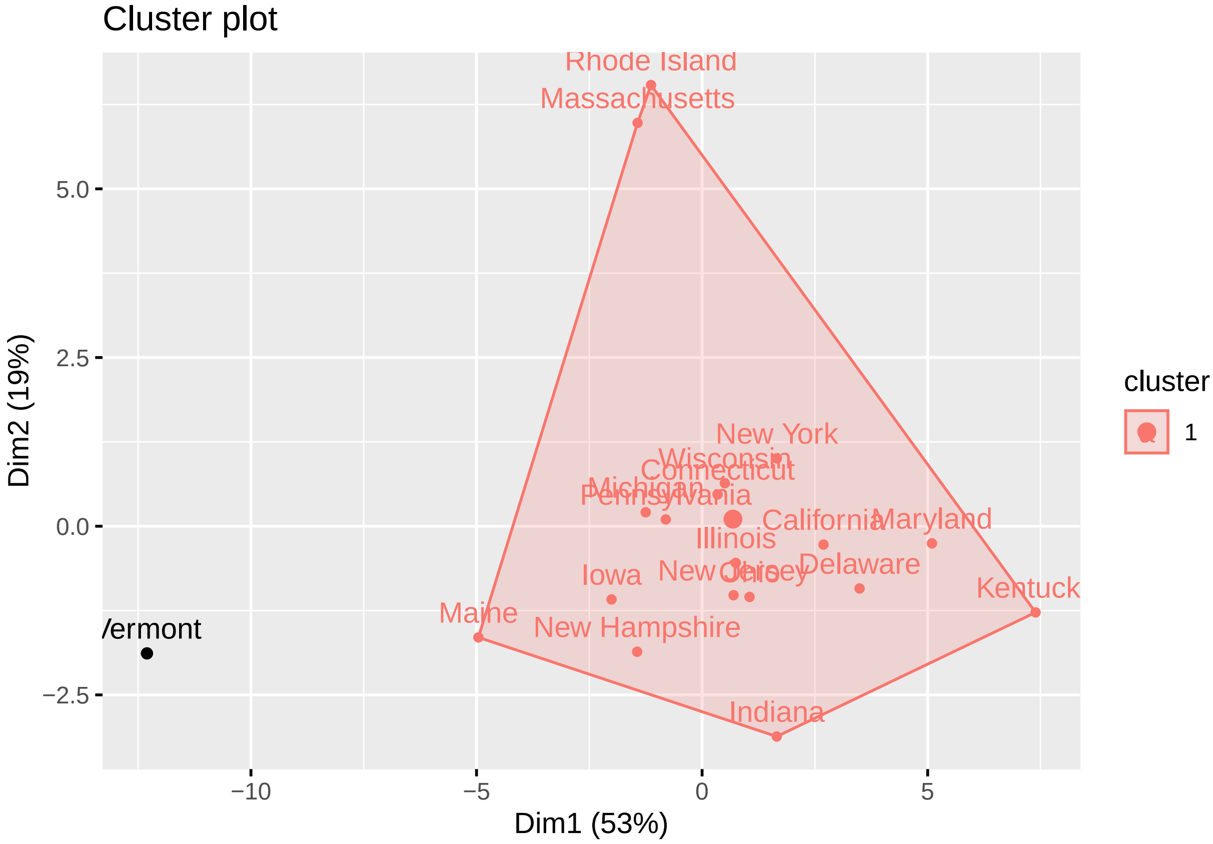
Next, I continued to do the hierarchical clustering using the agnes() function in order to analyze the dataset and identify the nested cluster or hierarchy. Using the complete method, I was able to obtain this result with an agglomerative coefficient of 0.8, which is rather close to 1. Values this high and close to 1 indicate that strong clustering structures are being created. I used the same code with the single and average methods as well. The agglomerative coefficients for the methods complete, single, and average are 0.8, 0.51, and 0.67 respectively. From the dendrograms we can observe that Kentucky and Maryland, for instance, have similar percentages of voters that voted for the republican candidate. The complete method is the best cluster proximity measure as it has the greatest agglomerative coefficient compared to the other methods (closest to 1).

Density based clustering

```
kNNdistplot(votesdset, k=5)
```



```r
db <- dbscan(votesdset, eps = 50, minPts = 5)
fviz_cluster(db, data = votesdset)
```



After this is density based clustering using the kNNdistplot() function using the votesdset data and k as 5. From the graph, I will select 50 as the radius of the core points. To create the dense based cluster, I create a new variable called db and use the dbscan() function using the votesdset dataset, 50 as the eps (radius), and 5 as the minPts. From this cluster plot, we can see that the single outlier is Vermont. This means that the state of Vermont has a much higher percentage of voters who voted for the republican candidate compared to the other states.