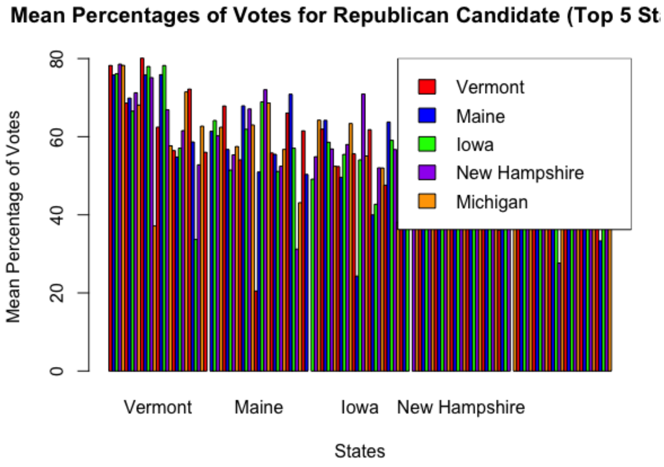
# Clustering States Based on Votes for Republican Candidates in Presidential Elections

## GOAL

To group states together based on their historical voting behavior to find any trends or patterns between clustered states.

## ANALYSIS

To begin this data mining process in R, I installed and loaded the dataset and the following packages: “pastecs”, “cluster”, and “factoextra”. After double-checking that the “votes.repub” dataset was a data frame, I moved on to the next step of data exploration in which I investigated the attributes’ and rows’ names and the head and tail rows of the dataset, and obtained an overall summary of this data, using the summary() function and pastecs::stat.desc() function. From this information, I gathered that there are 50 rows and 31 columns, in which the 50 rows are the 50 states and 31 columns are the 31 presidential elections from 1856 to 1976. I also found the top 5 states with the highest mean percentages for republican candidate votes as well as the bottom 5 states. This can be seen in the histogram I created in Figure 1 and Figure 2. I thought this might be helpful to see if the 5 states for the bottom and the top might have any commonalities on the surface level. For example, California and New York are both part of the bottom states, and based on my knowledge, these states are blue states. Unfortunately, after this data exploration, I discovered many missing values in this dataset.



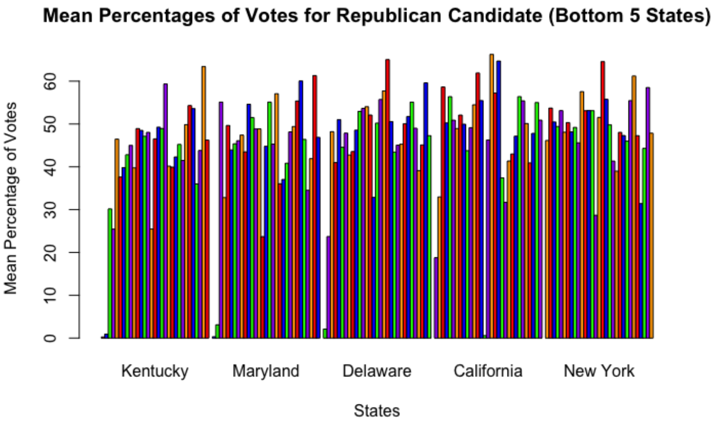
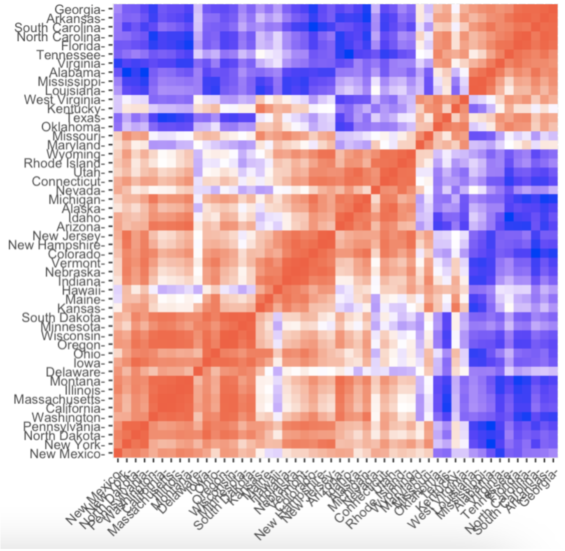


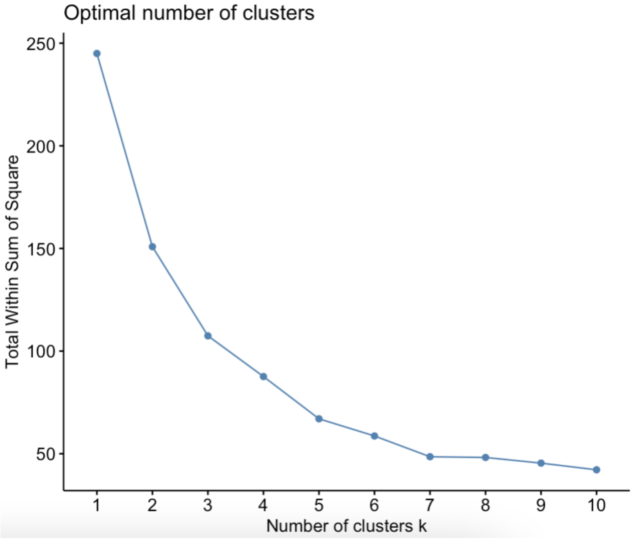
*Figure 1. Mean Percentage of Votes for Republican Candidate (Top 5 States). Figure 2. Mean Percentage of Votes for Republican Candidate (Bottom 5 States).*

After investigating the dataset, the preprocessing step begins with finding these null values. I need to delete these null values since I plan on using k-means clustering which does not work unless the data has no missing values. To solve this issue, I begin sum(is.na(votes.repub)), confirming that there are 217 missing values. From here, I decided to not use a list-wise approach to deleting these NA values and instead delete the attributes (elections) that contain the null values. Deleting by row would eliminate many of the states that would've been used and placed in the clusters. It would also be counterproductive to the goal of this cluster analysis since we want to see the voting behavior of states. Therefore, I found all the names of the attributes containing the highest number of null values and placed them in a subset. There were 26 attributes with null values in the subset. Finally, I created a new dataset called 'data2' that omitted the subset (using the na.omit() function) and no longer had any missing values. This data2 is to be used for k-means clustering but for the hierarchical clustering, I will just use the original votes.repub dataset. Additionally, to create clusters, the data attribute types must be numeric so using str(), I double-checked the attributes in both the original votes.repub and data2. All of the attributes were numeric, so I just needed to scale all numeric values' ranges so that they have the same frame of reference for when they are compared. I did this for both votes.repub and data2 because hierarchical clustering and k-means clustering need data that is scaled and normalized to ensure that all the variables have the same range and importance.

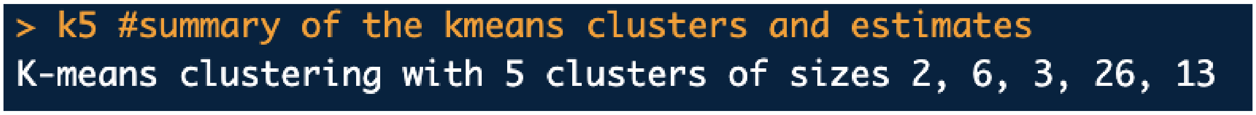
Before moving on to the clustering step, it is important to note why I am using k-means clustering and hierarchical clustering for this analysis. K-means clustering will separate the states into groups based on the centroids which represent the mean voting behavior of the states within each cluster. Meanwhile, hierarchical clustering will tell us how states are grouped based on their similarities or dissimilarities in voting behavior. This structure of voting behavior is then presented as a dendrogram. It is good to have hierarchical clustering to validate the k-means clustering. Additionally, I am not using density-based clustering for two reasons. First, this technique of clustering does not make sense for the goal of clustering states together based on their similar or dissimilar behavior in voting. Density-based clustering is formed from high density regions that are separated from one another by regions with low density. Additionally, density-based clustering is designed for spatial data, and the votes.repub dataset is temporal data.

For the k-means clustering, I calculated the distance matrix using the get.dist() function which calculates the distances between all data objects (being states) and returns those distances in a matrix. Then, I used the Pearson correlation as a tool to find proximity measures because Euclidean distance is not a good tool to use when scaling data since it is not invariant. On the left is the visualization of the matrix, and on the right is the line graph that determines the number of clusters needed for the k-means clustering using the fviz_nbclust() function. Based on the elbow method of choosing clusters for k-means, clusters 5, 6, and 7 are the best candidates for k in the clustering.

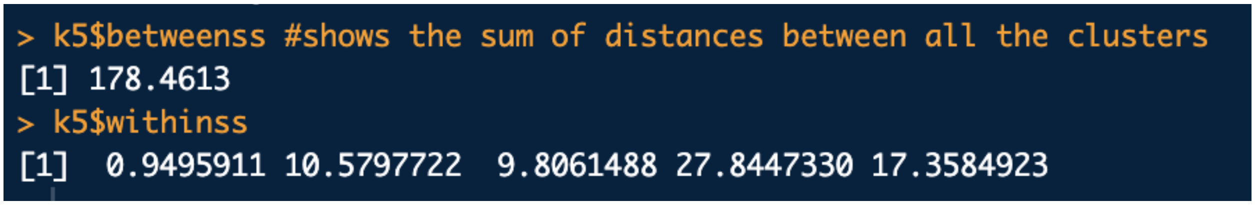




After running a K-means clustering with 5 clusters, it outputted the following information:

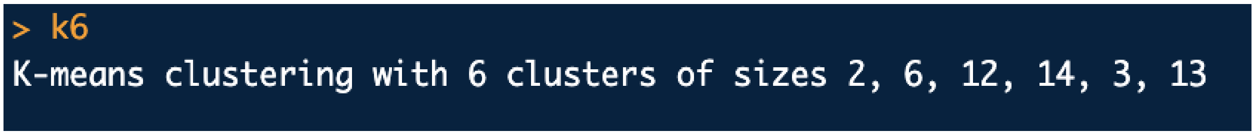


```
> k5 #summary of the kmeans clusters and estimates
K-means clustering with 5 clusters of sizes 2, 6, 3, 26, 13
```

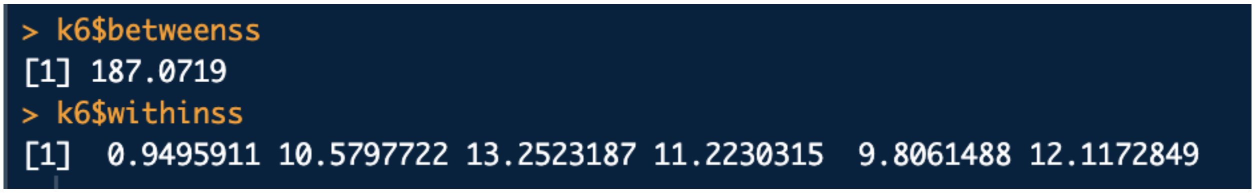


```
> k5$betweenss #shows the sum of distances between all the clusters
[1] 178.4613
> k5$withinss
[1]  0.9495911 10.5797722  9.8061488 27.8447330 17.3584923
```

After running a K-means clustering with 6 clusters, it outputted the following information:

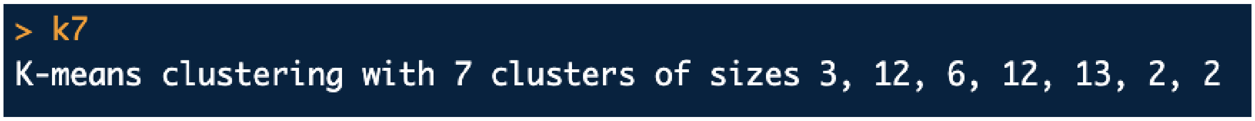


```
> k6
K-means clustering with 6 clusters of sizes 2, 6, 12, 14, 3, 13
```

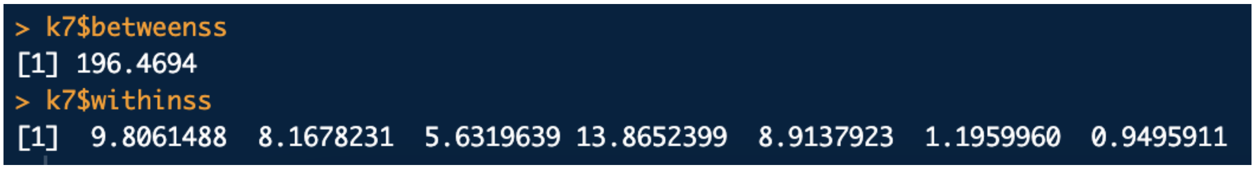


```
> k6$betweenss
[1] 187.0719
> k6$withinss
[1]  0.9495911 10.5797722 13.2523187 11.2230315  9.8061488 12.1172849
```

After running a K-means clustering with 7 clusters, it outputted the following information:



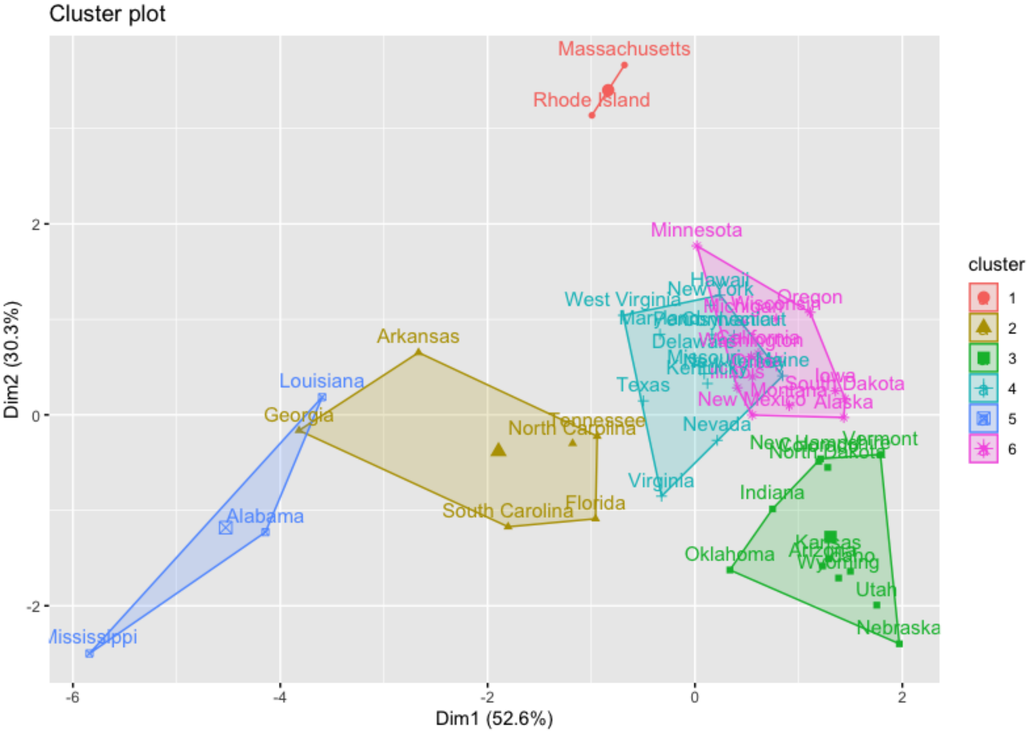
```
> k7
K-means clustering with 7 clusters of sizes 3, 12, 6, 12, 13, 2, 2
```

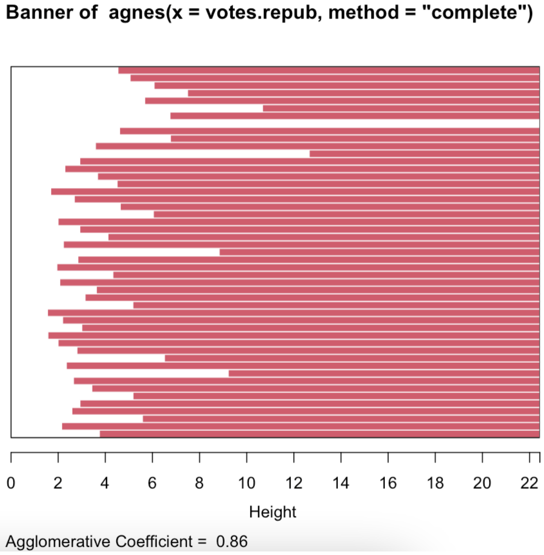


```
> k7$betweenss
[1] 196.4694
> k7$withinss
[1]  9.8061488  8.1678231  5.6319639 13.8652399  8.9137923  1.1959960  0.9495911
```

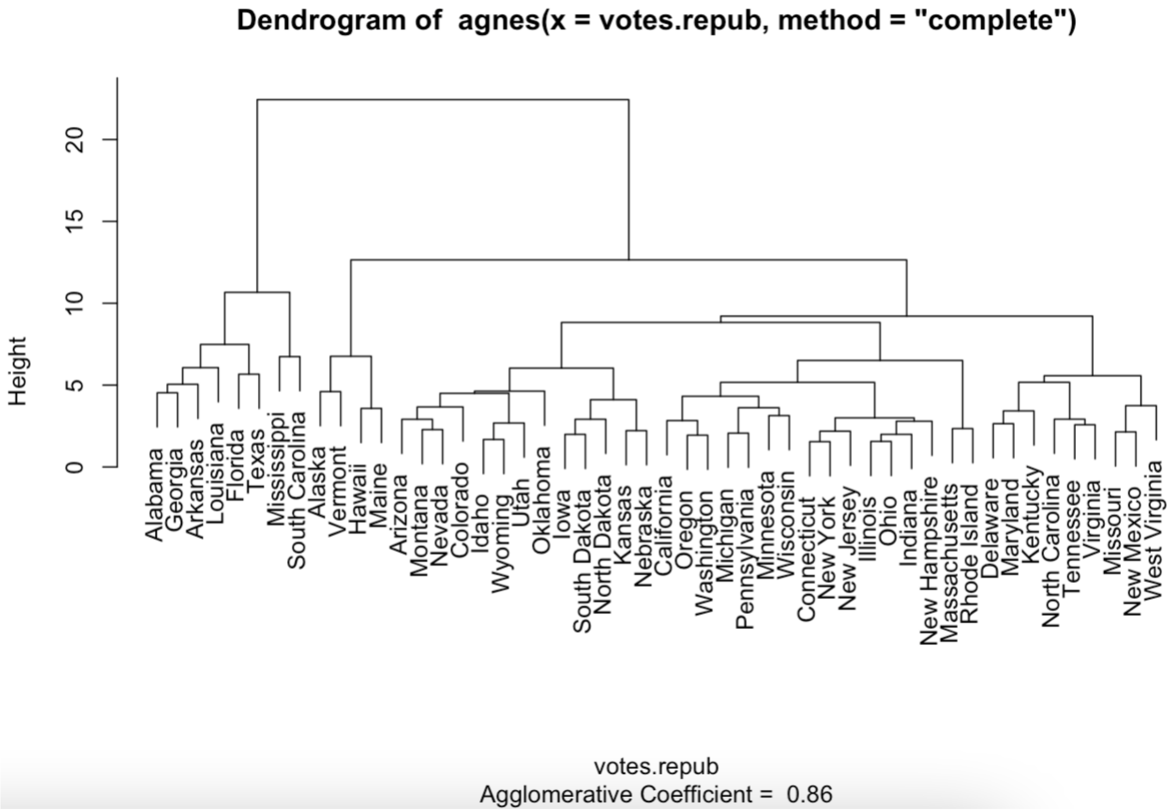
After finding the betweenSS and withinSS for each cluster, it is important to find the cluster with the greatest betweenSS and smallest or lowest withinSS. Even though cluster 7 has the greatest betweenSS, it has the largest withinSS of 9.8 between the other two clusters. Cluster 5 and 6 both have a withinSS of 0.949. Therefore, I chose cluster 6 because it has the next largest betweenSS of 187.07. Finally, I visualized the 6 clusters (seen in the image below) which show the different states and which cluster they belong to based on their voting behavior.

An interesting takeaway from this visualization is that the blue cluster 5 and yellow cluster 2 are all southern states (on the next page). This indicates that they have similar voting behaviors which might be caused by similar cultural, social, and political values in their region. With more detailed attributes that can categorize and describe each state better, one could draw stronger conclusions about the commonalities of the voting behavior between states in the same cluster.





For the hierarchical clustering, I used the 3 methods (single, complete, and average) to build the hierarchical cluster. In the single method, the agglomerative coefficient (ac) was 0.5847012. The complete link method outputted the agglomerative coefficient of 0.8623685, and the average method outputted 0.7490398. For creating the dendrogram, I used the complete link (max) method because it has the greatest agglomerative coefficient. We know higher values of ac mean more coherent clusters that capture the similarities between the states in the dendrogram. After plotting with the complete link method's ac, we see a high height which indicates high dissimilarity between the states.



Meanwhile, in the dendrogram, we see that Alabama and Georgia have very similar voting behaviors since they are connected to the same branch. The further we move up the dendrogram, we see that Arkansas merges with the Alabama-Georgia branch. The height difference means that Arkansas' voting behavior is less similar to Alabama's and Georgia's voting but more similar to the voting behavior of South Carolina which is higher up on the dendrogram structure.

**TOOLS & RESOURCES**

Language: R

Dataset information: "votes.repub" dataset that is pre-installed in R Studio.