K-Means Clustering Example: State Election Example

This data frame includes the percentages of votes given to the republican candidate in presidential elections from 1856 to 1976. Rows represent the 50 states, and columns are the 31 elections.

1. Import the data into R and change the first variable, State, into a factor variable.
2. Check summary of votes’ percentage for the year 1856. What is the minimum percentage of vote? What is the median percentage of vote? How many missing values are in year 1856?
3. Find the mean votes’ percentage for the year 1936 using function mean ( ).   
   Hint: you need to remove missing values, ie na.rm=TRUE
4. Create an imputing function to replace missing values in a given column with the mean value for that column after removing missing values.   
   Hint: You need to define a function, say impute
   1. that takes a column parameter (say x).
   2. uses ifelse on this column parameter. Condition is if a record is na then replace it with mean of this column after removing missing values; If record is not na then keep the record as it is.
5. Check if impute function works for X1856.
6. Apply impute function on all of the columns except the first column.
7. Scale the variables in the new data frame.
8. Set seed (e.g. set.seed(5)) for fixed clustering results. Train a k-means clustering models on the scaled election data frame with k =5.   
   Note: Remember not to include the 1st State column while running kmeans.
9. How many States are in cluster 3 and 4?   
   Hint: Add model$cluster to data frame. Make sure that column State is also there in the data frame.
10. List out all states for the cluster 3
11. Look at the center of clusters. What center values of x1876 do cluster 1, 2 and 3 have?
12. What are the average votes’ percentages of year 1900 and 1892 for cluster 5?
13. Which clusters do Alabama, California and Utah belong to?   
    Hint:you need to subset data for which you need to show cluster number, ie dt2$State==c("California", "Utah")