# 1. Import the data into R and change the first variable, State, into a factor variable.

dt<- read.csv(file.choose(), stringsAsFactors = F)

str(dt)

dt$State <- factor(dt$State)

dt$State

# 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?

summary(dt$X1856)

#min - 0.19

#median - 47.31

#missing values - 30

# 3. Find the mean votes' percentage for the year 1936 using function mean ( ).

# Hint: you need to remove missing values, ie na.rm=TRUE

mean(dt$X1936,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

# a. that takes a column parameter (say x).

# b. 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.

impute <- function(x)

{

avg <- mean(x,na.rm = TRUE)

x <- ifelse(is.na(x),avg,x)

}

# 5. Check if impute function works for X1856.

dt$X1856 <- impute(dt$X1856)

dt$X1856

# 6. Apply impute function on all of the columns except the first column.

dt3 <- dt

new <- data.frame(lapply(dt3[-1],impute))

head(new)

# 7. Scale the variables in the new data frame.

dt3.scale <- data.frame(lapply(new,scale))

# 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.

set.seed(5)

clusters <- kmeans(dt3.scale[-1], 5)

# 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.

dt3$cluster <- clusters$cluster

table(dt3$State,dt3$cluster)

# cluster 3 - 7

# cluster 4 - 9

# 10. List out all states for the cluster 3

#Iowa, Kansas, Maine, Massachusetts, Minnesota,Rhode Island ,Vermont

# 11. Look at the center of clusters. What center values of x1876 do cluster 1, 2 and 3 have?

clusters$cluster

#cluster 1 = 0.1826023 ; cluster 2 - 1.2659163 ; cluster 3 - 1.5535077

# 12. What are the average votes' percentages of year 1900 and 1892 for cluster 5?

aggregate(data = dt3, dt$X1900 ~ cluster, FUN = mean) # 55.93250

aggregate(data = dt3, dt$X1892 ~ cluster, FUN = mean) # 46.93000

# 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")

table(dt3$State,dt3$cluster)

# Alabama :cluster 2

#california : cluster 5

#utah : cluster 1