Project 4: Cluster Analysis

Brooke O'Donnell

SOCI556

Professor Wang

This dataset is part of the "The Economist's" (a magazine) rating of "democracy index" for 2018. The democracy index is calculated for each country based on several dimensions of democratic institution and practices.

The dataset contains the following variables:

NOMINAL

1. [COUNTRY] :Country name

CONTINUOUS: measured on a 0 to 10 scale

2. [PLURALISM] a measure of pluralism
3. [GOVFUNC]: a measure of how the government functions (government efficacy)
4. [POLIPARC]: a measure of political participation
5. [POLICUL]: a measure of political culture
6. [CIVLIB]: a measure of civil liberty

Your task is to cluster (categorize) all 167 countries in the dataset into groups using different clustering techniques and scenarios.

A. Please create a dataset that contains only the the five variables that measure aspects of democracy; and scale all the measures.

Interpretation: I created two dataframes for project 4. DF.DEM only contains the five scaled variables measuring democracy. The second dataframe, DEMwCountry has the five variables measuring democracy, scaled, and the country names.	**Output:** N/A
Given code: mydf1<-na.omit(mydf)str(mydf1)mydf2<-scale(mydf1[, -1]) **Code:** names(DEM)[1]<-"country" DF.DEM<-select(DEM, pluralism, govfunc, poliparc, policul, civlib)	

```
na.omit(DF.DEM)
DF.DEM<-scale(DF.DEM)
str(DF.DEM) DEMwCOUNTRY<-DEM%>% mutate_at(c(2,3,4,5,6), funs(c(scale(.))))
na.omit(DEMwCOUNTRY)
```

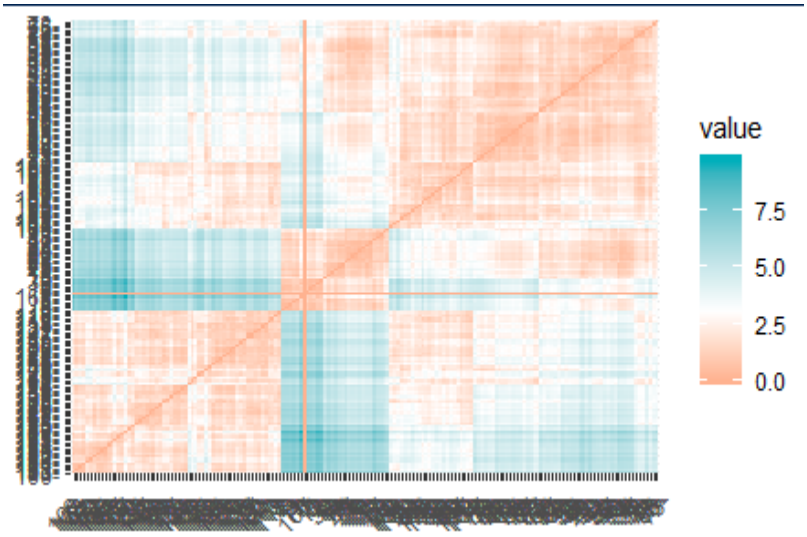
B. Please create a "distance" matrix ("euclidean distance") and visualize it. Do you think the colors in the plot justify some form of cluster analysis? Why or why not?

Interpretation: Yes, the clusters justify some form of cluster analysis. There are many instances of highly concentrated areas on the graph with only blue and only orange cases. It is up to the researcher to decide what # of clusters are going to produce best model.

Output:

```
      1     2     3
1 0.000 0.666 0.991
2 0.666 0.000 0.402
3 0.991 0.402 0.000
```

```
             1         2         3
2   0.6655584
3   0.9911368 0.4024151
4   1.4149861 1.2486864 1.3229929
5   1.0962021 0.5788280 0.4732149
6   1.5494670 1.0790503 0.9229355
            4         5         6
```



Given code:

Code: hclust_dist<- as.dist(clus)
hclust_dist[is.na(hclust_dist)] <- 0
hclust_dist[is.nan(hclust_dist)] <- 0
sum(is.infinite(hclust_dist))
h <- hclust(hclust_dist, "ward.D2")
fviz_dist(hclust_dist, gradient=list (low="#FC4E07", mid="white", high="#00AFBB"))

C. Perform a k-mean cluster analysis with 3 clusters. Include the 3 cluster centers in your output. What do the 3 cluster centers inform us about the three clusters of countries?

Output:

3-cluster model:

```
K-means clustering with 3 clusters of sizes 67, 60, 40

(between_SS / total_SS =  71.0 %)
Cluster means:
Cluster means:
   pluralism    govfunc   poliparc    policul     civlib
1  0.4083755  0.1479025  0.2606102 -0.2228408  0.2871479
2 -1.1406485 -0.9814841 -0.9771140 -0.6306445 -1.1186176
3  1.0269438  1.2244895  1.0291490  1.3192250  1.1969537
```

Interpretation: This output suggests that if the researcher were to accept the 3-cluster model the countries might suggest the following interpretation; 'hybrid-democratic' countries are in group one, 'low-democratic' countries are in group 2, and 'democratic' countries are in group 3.

D. Based on the model in step 3, how much of the total sum of squares can be attributed to between-cluster sum of squares? Does the proportion or percentage justfy a 3-cluster model?

Interpretation:

(between_SS / total_SS = 71.0 %). 71.0% suggests there are good clusters. Normally researchers recommend a percentage between 70-90. Increasing the number of clusters to 5 makes our between SS / total SS percentage rise 6.9%. However, I would suggest to the researcher to accept a 3-cluster model over a 5-cluster model. This is because there is less variation between 67,30, and 40 than compared to the variation between 47, 40, 17, 35 and 28

Output:

3-cluster model:

```
(between_SS / total_SS =
71.0 %)
K-means clustering with
3 clusters of sizes 67,
60, 40
```

5-cluster model:

```
(between_SS / total_SS =
77.9 %)
K-means clustering with
5 clusters of sizes 47,
40, 17, 35, 28
```

Given code: set.seed(33)model.km<-kmeans(mydf2, 3, nstart = 10)model.km
Code: set.seed(97)
model.km<-kmeans(DF.DEM, 5, nstart = 10)
model.km

E. One key component of cluster analysis is to examine the clusters by applying empirical knowledge one has (sometimes common sense). For instance, "Is the U.S. in the same cluster with the U.K. and France etc.?" or "Is China in the same cluster with Russia or Iran?" So, please choose a few of your "favorite" countries to "make sense" of the clusters.

Interpretation:
This assignment made me realize I know nothing about other governments. I approached this question by googling "What type of government does _____ have?" Norway and Japan both list 'Parliamentary system' and they both appear in the 'democratic' countries cluster. Russia, North Korea, Iran, and Irag all appeared in the same county cluster. All these countries are notoriously known for being 'low democratic' countries. Finally, Mexico, Moldova, Mongolia all appear in the 'hybrid-democratic' countries. There were often multiple descriptions to describe the type of government in these countries.

Output:

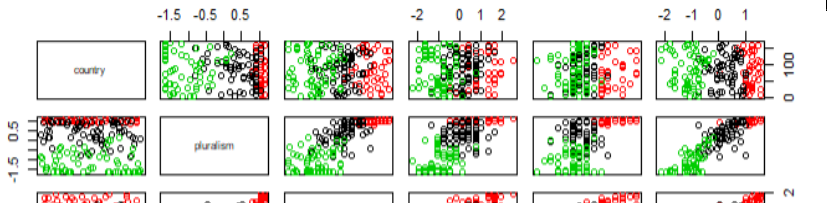
```
Norway                  0 0 1
Japan                   0 0 1

Russia                  0 1 0
North Korea             0 1 0
Iran                    0 1 0
Iraq                    0 1 0

Mexico                  1 0 0
Moldova                 1 0 0
Mongolia                1 0 0
Montenegro              1 0 0
Morocco                 1 0 0
```

Given code: table(mydf1$country, model.km$cluster)
Code: table(DEM$country, model.km$cluster)

F. Please plot the "clusters" from your model against the dataframe you created in step 1 (it should contain all the variables). As this plot provides multiple 2-dimensional scatterplots, please discuss which pairs of variables produce clarity in the clustering; which pairs do not.

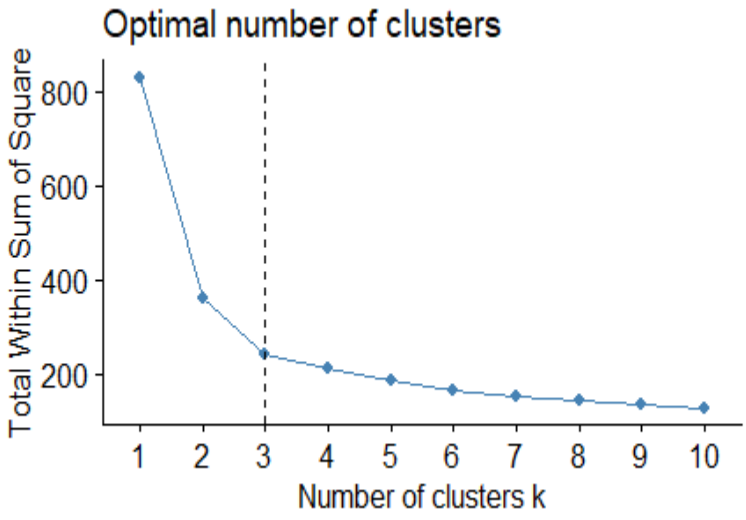
Interpretation:

Output:

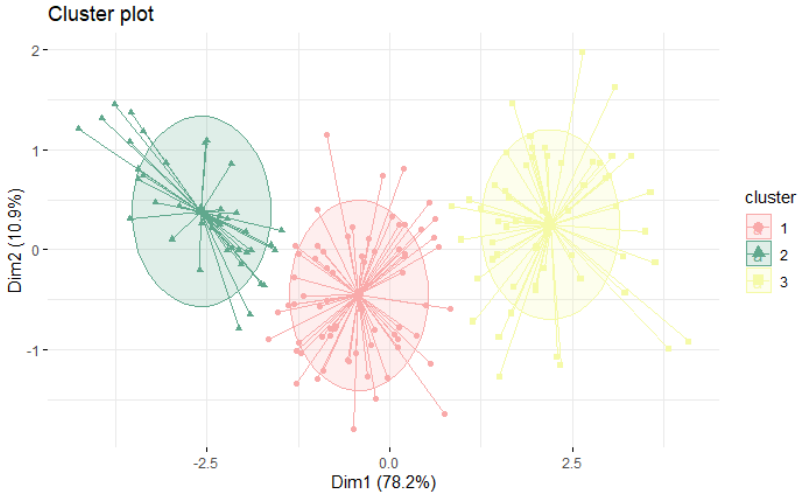


Variables that produce clarity: Nothing looks good against country. All the other variables work well with one another. The variables that contributes the least is **pluralism** against **policul.** I believe the variables that contribute most together is **pluralism** against **civil lib.**	
Given code: plot(mydf1, col=model.km$cluster) **Code**: plot(DEMwCOUNTRY, col=model.km$cluster)	

G. Use the "wss" (within sum of sqaure) reduction method to plot your scaled data to decide on the best number of clusters.

Interpretation: Does increasing the number of clusters to make our findings better? I would answer no. The optimal number of clusters will be 3 because it is in the 'elbow' when plotted.	**Output:** 
Given code: fviz_nbclust(mydf2, kmeans, method = "wss") +geom_vline(xintercept = 3, linetype = 2) **Code:** fviz_nbclust(DF.DEM, kmeans, method = "wss") +geom_vline(xintercept = 3, linetype = 2)	

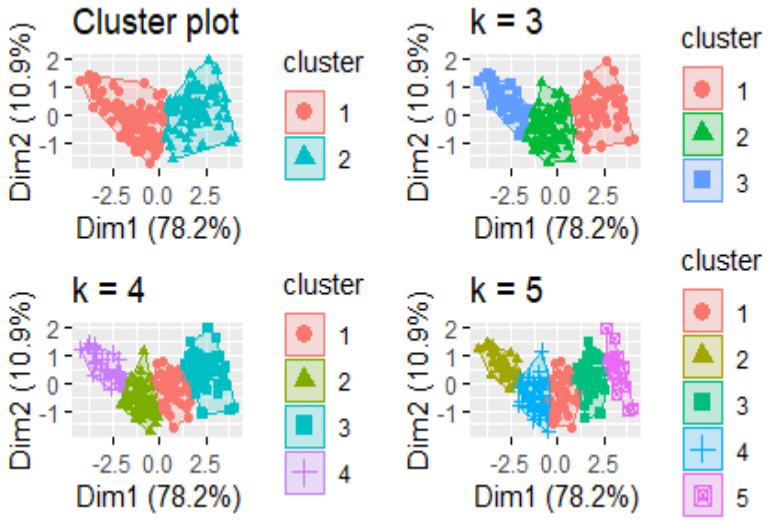
H. Visualize the 3-cluster model using "ellipse.type" and "star.plot" in function "fviz.cluster."

Interpretation:	**Output:**
I think that this model serves to be useful model. However, not the best. I would argue that a 4-cluster model will enable an interpretation that matches how society secularizes democracies.	

Given code: fviz_cluster(model.km, data = mydf2,palette = c("#FC4E07", "#00AFBB", "#E7B800", "#fc4e07"),ellipse.type = "euclid", star.plot = TRUE, repel = TRUE, ggtheme = theme_minimal())

Code: fviz_cluster(model.km, data=DF.DEM, palette= c("#F9AAAA", "#60A88D", "#F5FAA7", "#A7C3FA"),ellipse.type="euclid", star.plot=TRUE, repel=TRUE, ggtheme=theme_minimal())

I. Repeat step 3 and try a 4-cluster model. Use all the "clues" or criteria such as common knowledge, wss, or the cluster centers to argue for or against your 4-cluster model.

Interpretation:	**Output:**
Although three has been shown to be in the elbow in the optimal number of cluster plot I would support the 4-cluster model. I believe that the Economist Intelligence team made this decision because the 4-cluster model holds more meaning. Although the 3-cluster model is meaningful separating the clusters into 'hybrid-democratic', 'low-democratic', and high-democratic' countries it won't hold as much meaning when compared to the 4-cluster. By this simple change the researcher can now interpret	

her model by saying there are four types of clusters 'hybrid-democratic', low-democratic', 'high-democratic', and authoritarian' countries. I think this will enable the reader of your models to be more interested in when comparing the clusters of countries with one another.	

Given code: set.seed(33)model.km<-kmeans(mydf2, 4, nstart = 10)model.km

Code::

```
set.seed(72)
kclus5<-kmeans(DEMwCOUNTRY[,2:6], 5, nstart= 10)
DEMwCOUNTRY$kclus5<-as.factor(kclus5$cluster)
kclus5
str(kclus5)

plot(DEMwCOUNTRY, col=kclus5$cluster)
table(DEMwCOUNTRY$country, kclus5$cluster)
plot(DEMwCOUNTRY, col=kclus5$cluster)
DF.DEM<-DF.DEM[-c(168),]
kc2<- kmeans(DF.DEM, centers=2, nstart=10)
kc3 <- kmeans(DF.DEM, centers = 3, nstart = 10)
kc4 <- kmeans(DF.DEM, centers = 4, nstart = 10)
kc5 <- kmeans(DF.DEM, centers = 5, nstart = 10)
        p1<- fviz_cluster(kc2, geom ="point", data =DF.DEM)
        p2 <- fviz_cluster(kc3, geom = "point",  data = DF.DEM) + ggtitle("k = 3")
        p3 <- fviz_cluster(kc4, geom = "point",  data = DF.DEM) + ggtitle("k = 4")
        p4 <- fviz_cluster(kc5, geom = "point",  data = DF.DEM) + ggtitle("k = 5")
library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)

fviz_cluster(kclus, data=DF.DEM, palette = c("#B52A2A", "#2A33B5", "#2AB558"),
ellipse.type = "euclid", star.plot=TRUE, repel=TRUE, ggtheme=theme_minimal())

fviz_cluster(kclus5, data=DF.DEM, palette = c("#B55D2A", "#2AB541", "#2A66B5",
"#F6DC59", "#F659DC"), ellipse.type = "euclid", star.plot=TRUE, repel=TRUE,
ggtheme=theme_minimal())

        library(factoextra)
        fviz_nbclust(DF.DEM, kmeans, method= "wss")+
```

```
geom_vline(xintercept=3, linetype=2)
```