Homework 9

Applied Mutlivariate Analysis

Emorie Beck

November 1, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(gridExtra)
library(knitr)
library(kableExtra)
library(MASS)
library(vegan)
library(smacof)
library(scatterplot3d)
library(ape)
library(ade4)
library(ecodist)
library(cluster)
library(factoextra)
library(ggdendro)
library(lme4)
library(plyr)
library(tidyverse)
```

1.2 data

The file, Set_7_A.csv, contains data for 183 participants in a Pew Center Political Survey conducted in April 2017. Participants were asked their opinions about government spending using the following question stem:

"If you were making up the budget for the federal government this year, would you increase spending, decrease spending or keep spending the same for _________"?

The spending areas considered were:

- Scientific research
- Militarydefense
- Government assistance for the unemployed
- Medicare
- Environmental protection

- Economic assistance to needy people around the world
- Education

Responses were coded 1 = Decrease spending, 2 = Keep spending the same, and 3 = Increase spending.

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework9"

datA <- sprintf("%s/Set_7_A.csv", wd) %>% read.csv(., stringsAsFactors = F)
datB <- sprintf("%s/Set_7_B.csv", wd) %>% read.csv(., stringsAsFactors = F)

head(datA)

##   Spend_Science Spend_Military Spend_Unemployed Spend_Medicare
## 1             1              3                1              3
## 2             1              3                2              1
## 3             1              3                2              2
## 4             1              3                1              2
## 5             1              3                1              3
## 6             1              3                1              1
##   Spend_Environment Spend_World_Needy Spend_Education ID
## 1                 1                 1               3  1
## 2                 1                 1               1  2
## 3                 1                 1               2  3
## 4                 1                 1               2  4
## 5                 1                 1               2  5
## 6                 1                 1               2  6

head(datB)

##   Age   Area_F               Trust_F             Gun_Control_F Sex_F
## 1  59    Rural                 Never Protect Right to Own Guns   Men
## 2  82    Rural Only Some of the Time Protect Right to Own Guns Women
## 3  40 Suburban Only Some of the Time Protect Right to Own Guns   Men
## 4  61    Rural Only Some of the Time Protect Right to Own Guns Women
## 5  70    Rural Only Some of the Time Protect Right to Own Guns   Men
## 6  63 Suburban Only Some of the Time Protect Right to Own Guns Women
##             Education_F Race_F     Income_F     Party_F        Ideology_F
## 1 Completed High School  White   30K to 40K  Republican      Conservative
## 2          Some College  White   50K to 75K  Republican      Conservative
## 3      Four-Year Degree  White   30K to 40K Independent Very Conservative
## 4          Some College  White 100K to 150K  Republican          Moderate
## 5         Post-Graduate  White   50K to 75K  Republican      Conservative
## 6      Four-Year Degree  White 100K to 150K  Republican          Moderate
##   ID
## 1  1
## 2  2
## 3  3
## 4  4
## 5  5
## 6  6
```

2 Question 1

First, use Wards procedure with Euclidean distance to determine how many clusters of respondents best describe this sample. Use the following procedures to arrive at the best choice:

2.1 Part A-D

Run Wards procedure on a file that has been re-ordered using values for the military spending variable. Save the cluster memberships for a two-cluster solution. Repeat this procedure after re-ordering the file using values for the environmental protection variable. Repeat this procedure after re-ordering the file using values for the education variable. When you are done, you will have three new cluster membership variables, each from a Wards method applied to a different ordered version of the file. Create a cross-classification table for each pair of cluster membership variables. Are cases classified consistently regardless of how the data were ordered?

Repeat this process, but for a three-cluster solution.

```
dist_fun <- function(dat){dist(dat %>% select(-ID))}
cut_fun_w <- function(dat, C, k){dat$CW <- cutree(C, k = k); dat}
cut_fun_k <- function(dat, C, k){dat$CK <- C$cluster; dat}

nested.models <- tribble(
  ~k, ~order, ~data,
  2, "Spend_Military", datA %>% arrange(Spend_Military),
  3, "Spend_Military", datA %>% arrange(Spend_Military),
  4, "Spend_Military", datA %>% arrange(Spend_Military),
  5, "Spend_Military", datA %>% arrange(Spend_Military),
  2, "Spend_Environment", datA %>% arrange(Spend_Environment),
  3, "Spend_Environment", datA %>% arrange(Spend_Environment),
  4, "Spend_Environment", datA %>% arrange(Spend_Environment),
  5, "Spend_Environment", datA %>% arrange(Spend_Environment),
  2, "Spend_Education", datA %>% arrange(Spend_Education),
  3, "Spend_Education", datA %>% arrange(Spend_Education),
  4, "Spend_Education", datA %>% arrange(Spend_Education),
  5, "Spend_Education", datA %>% arrange(Spend_Education)
) %>%
  mutate(D = map(data, dist_fun),
         C = map(D, ~hclust(., method = "ward.D2")),
         D = pmap(list(data, C, k), cut_fun_w),
         K = map2(data, k, kmeans),
         D = pmap(list(D, K, k), cut_fun_k))

cross.class <- nested.models %>% unnest(D) %>%
  select(k, order, CW, ID) %>%
  spread(key = order, value = CW) %>%
  select(-ID) %>%
  group_by(k) %>%
  summarize(`Edu-Env` = list(table(Spend_Education, Spend_Environment)),
            `Edu-Mil` = list(table(Spend_Education, Spend_Military)),
            `Env-Mil` = list(table(Spend_Environment, Spend_Military)))

(cross.class %>% filter(k == 2))$`Edu-Env`[[1]]

##                  Spend_Environment
## Spend_Education   1   2
##               1  63   0
##               2   0 120
```

```
(cross.class %>% filter(k == 2))$`Edu-Mil`[[1]]
```

```
##                Spend_Military
## Spend_Education   1   2
##               1   0  63
##               2 120   0
```

```
(cross.class %>% filter(k == 2))$`Env-Mil`[[1]]
```

```
##                  Spend_Military
## Spend_Environment   1   2
##                 1   0  63
##                 2 120   0
```

```
(cross.class %>% filter(k == 3))$`Edu-Env`[[1]]
```

```
##                Spend_Environment
## Spend_Education  1  2  3
##               1 63  0  0
##               2  0  0 46
##               3  0 74  0
```

```
(cross.class %>% filter(k == 3))$`Edu-Mil`[[1]]
```

```
##                Spend_Military
## Spend_Education  1  2  3
##               1  0  0 63
##               2  0 46  0
##               3 74  0  0
```

```
(cross.class %>% filter(k == 3))$`Env-Mil`[[1]]
```

```
##                  Spend_Military
## Spend_Environment  1  2  3
##                 1  0  0 63
##                 2 74  0  0
##                 3  0 46  0
```

```
(cross.class %>% filter(k == 4))$`Edu-Env`[[1]]
```

```
##                Spend_Environment
## Spend_Education  1  2  3  4
##               1 29  0 34  0
##               2  0  0  0 46
##               3  0 32  0  0
##               4  0 42  0  0
```

```
(cross.class %>% filter(k == 4))$`Edu-Mil`[[1]]
```

```
##                Spend_Military
## Spend_Education  1  2  3  4
##               1  0  0 29 34
##               2  0 46  0  0
##               3 32  0  0  0
##               4 42  0  0  0
```

```
(cross.class %>% filter(k == 4))$`Env-Mil`[[1]]

##                  Spend_Military
## Spend_Environment  1  2  3  4
##                 1  0  0 29  0
##                 2 74  0  0  0
##                 3  0  0  0 34
##                 4  0 46  0  0

(cross.class %>% filter(k == 5))$`Edu-Env`[[1]]

##                Spend_Environment
## Spend_Education  1  2  3  4  5
##               1 29  0  0  0  0
##               2  0  0 34  0  0
##               3  0  0  0  0 46
##               4  0 22  0 10  0
##               5  0  0  0 42  0

(cross.class %>% filter(k == 5))$`Edu-Mil`[[1]]

##                Spend_Military
## Spend_Education  1  2  3  4  5
##               1  0  0  0 29  0
##               2  0  0  0  0 34
##               3  0  0 46  0  0
##               4 17 15  0  0  0
##               5  5 37  0  0  0

(cross.class %>% filter(k == 5))$`Env-Mil`[[1]]

##                  Spend_Military
## Spend_Environment  1  2  3  4  5
##                 1  0  0  0 29  0
##                 2  9 13  0  0  0
##                 3  0  0  0  0 34
##                 4 13 39  0  0  0
##                 5  0  0 46  0  0
```

The data are relatively inconsistently classified depending on the ordering. With larger k, the data seem to be classified more consistently.

2.2 Part E

(e) Of these solutions, which one provides the highest number of clusters while being completely immune to the ordering of the variables?

The highest number of clusters that are aimmune to the ordering is 3 clusters.

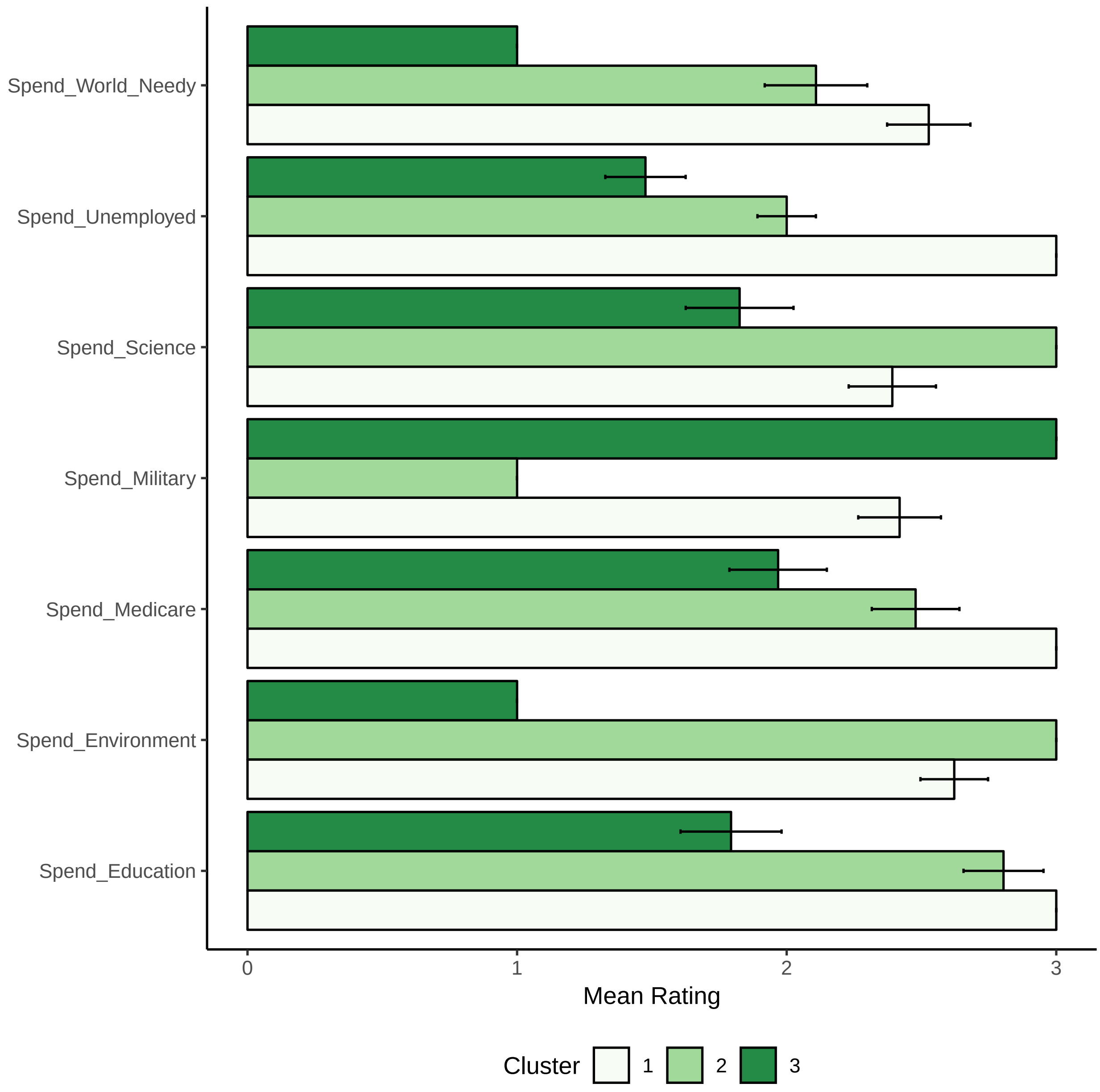
```
(cross.class %>% filter(k == 3))$`Edu-Env`[[1]]

##                Spend_Environment
## Spend_Education  1  2  3
##               1 63  0  0
##               2  0  0 46
##               3  0 74  0
```

```
(cross.class %>% filter(k == 3))$`Edu-Mil`[[1]]

##                 Spend_Military
## Spend_Education  1  2  3
##               1  0  0 63
##               2  0 46  0
##               3 74  0  0

(cross.class %>% filter(k == 3))$`Env-Mil`[[1]]

##                   Spend_Military
## Spend_Environment  1  2  3
##                 1  0  0 63
##                 2 74  0  0
##                 3  0 46  0
```

3 Question 2

For the chosen number of clusters from Question 1, create a bar graph of the cluster means for each of the spending variables (include 95% confidence intervals). Based on these graphs, provide a description of each clusters spending attitude profile.

```
nested.models %>% filter(k == 3 & order == "Spend_Military") %>%
  unnest(D) %>%
  gather(key = cat, value = value, contains("Spend")) %>%
  Rmisc::summarySE(., measurevar = "value", groupvars = c("CW", "cat")) %>%
  ggplot(aes(x = cat, y = value, fill = factor(CW))) +
    scale_fill_manual(values = RColorBrewer::brewer.pal(9, "Greens")[c(1,4,7)]) +
    geom_bar(stat = "identity", position = "dodge", color = "black") +
    geom_errorbar(aes(ymin = value - ci, ymax = value + ci),
                  position = position_dodge(.9), width = .1) +
    labs(x = NULL, y = "Mean Rating", fill = "Cluster") +
    coord_flip() +
    theme_classic() +
    theme(legend.position = "bottom")
```

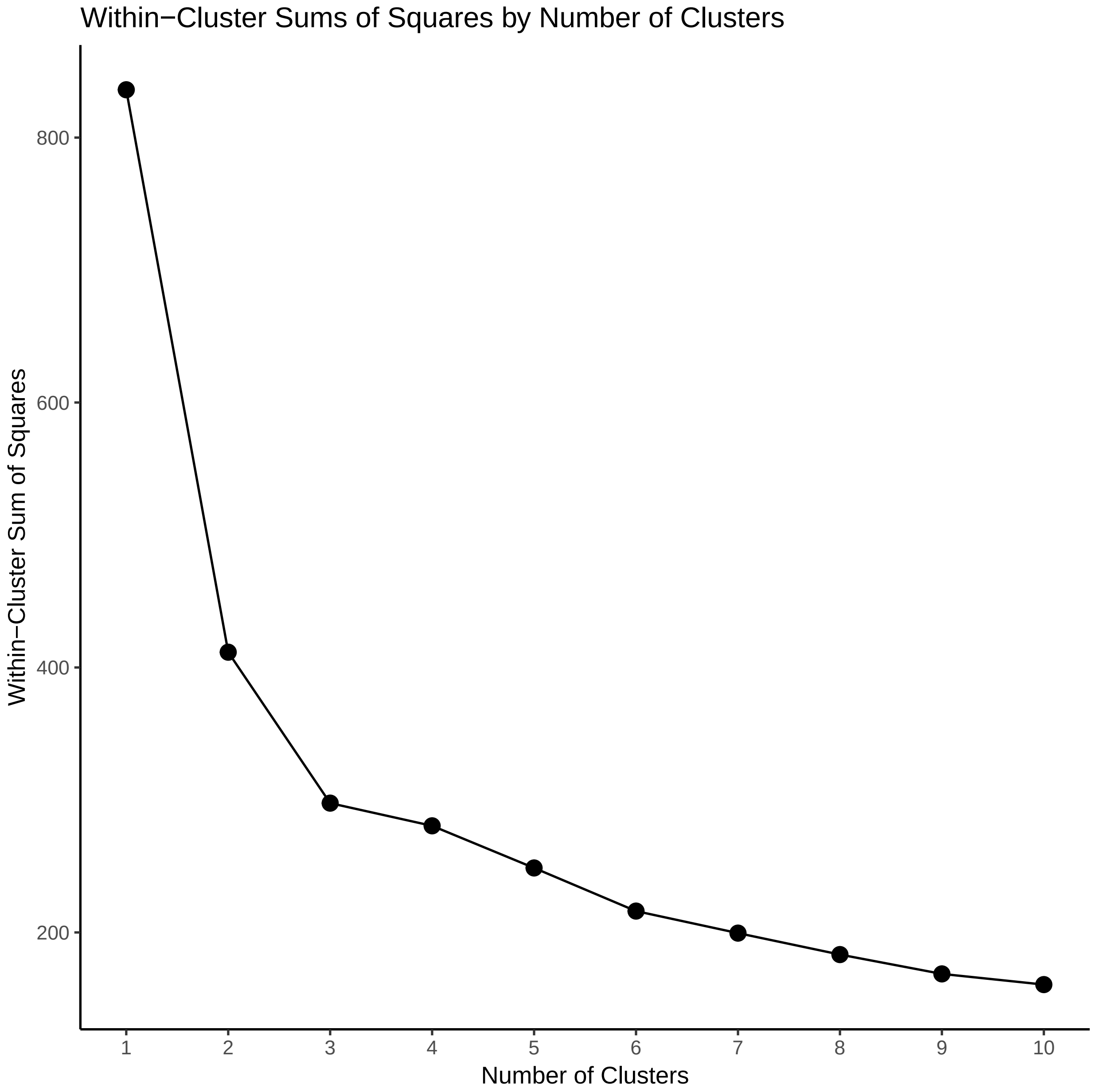


Cluster 3 are clearly conservatives who don't want to spend on anything but the military. Cluster 2 are more radical (and often high SES) liberals who oppose military spending but most other areas of spending. Cluster 1 are more moderate, supporting military spending as well as virtually every other spending category.

4 Question 3

Using K-means clustering, create a plot of the within-cluster sums of squares for cluster solutions up to 10. Does the number of clusters from this plot suggest the same number of clusters identified by Wards procedure?

```
wssplot <- function(data,nc=15,seed=1234) {
  wss <- (nrow(data-1))*sum(apply(data,2,var))
  for (i in 2:nc) {
    set.seed(seed)
    wss[i] <- sum(kmeans(data,centers=i)$withinss)
  }
  plot_data <- cbind(wss,seq(1,nc,1))
  plot_data <- as.data.frame(plot_data)
  names(plot_data) <- c("wss","nc")
  ggplot(plot_data, aes(x=nc,y=wss)) +
    geom_point(shape=19,size=3) +
    geom_line() +
    scale_x_continuous(breaks=c(seq(1,nc,1))) +
    xlab("Number of Clusters") +
    ylab("Within-Cluster Sum of Squares") +
    theme_classic() +
  ggtitle("Within-Cluster Sums of Squares by Number of Clusters")
}
wssplot(datA[,1:7],nc=10)
```



Yes, K-means clustering also seems to suggest that 3 clusters is optimal.

5 Question 4

For the chosen number of clusters from Question 1, conduct a K-means clustering of the data. Provide a cross-classification table for the Wards and K-means procedures. Do the two procedures identify the same clusters?

```
K <- kmeans(datA %>% select(-ID), centers = 3, iter.max = 1000, nstart = 10)
Clusters_K <- as.data.frame(K$cluster)
names(Clusters_K) <- c("Cluster_K")
Clusters_K$ID <- datA$ID

data_dist <- dist(datA %>% select(-ID), method = "euclidean", diag = TRUE)
```

```
Wards <- hclust(data_dist, method = "ward.D2")
Wards_clusters <- as.data.frame(cutree(Wards, k = 3))
names(Wards_clusters) <- c("Cluster_Wards")
Wards_clusters$ID <- datA$ID

New <- full_join(Clusters_K, Wards_clusters)
Cross_T <- table(New$Cluster_K, New$Cluster_Wards)
Cross_T

##
##      1  2  3
##   1 63  0  0
##   2  0  5 46
##   3  0 69  0
```

The two procedures appear to identify slightly different clusters.

6 Question 5

5. The file, Set_7_B.csv, contains additional information about these respondents. Use this file to better define the nature of the clusters identified in Question 4. The variables included in Set_7_B are age, home area (rural, urban, suburban), level of trust in the government, gun control attitude, sex, highest level of education, race, annual income, political party affiliation, and political ideology. Using these variables, what are the defining features of each cluster? Focus on those variables that do the best job of distinguishing the groups.

```
(nested.models %>% filter(k == 3))$D[[1]] %>% full_join(datB) %>%
  select(-(Spend_Science:Spend_Education)) %>%
  group_by(CW) %>% mutate(N = n()) %>%
  gather(key = category, value = value, -ID, -CW, -CK, -Age, -N) %>%
  group_by(CW, category, value) %>%
  summarize(n = n()/N[1]) %>%
  spread(key = CW, value = n)

## # A tibble: 42 x 5
## # Groups:   category [9]
##    category    value                          `1`     `2`    `3`
##    <chr>       <chr>                        <dbl>   <dbl>  <dbl>
##  1 Area_F      Rural                       0.176   0.174  0.270
##  2 Area_F      Suburban                    0.473   0.283  0.460
##  3 Area_F      Urban                       0.351   0.543  0.270
##  4 Education_F Completed High School       0.297   0.0870 0.206
##  5 Education_F Did Not Complete High School 0.0270 NA      0.0476
##  6 Education_F Four-Year Degree            0.216   0.348  0.238
##  7 Education_F Post-Graduate               0.135   0.239  0.206
##  8 Education_F Some College                0.189   0.174  0.159
##  9 Education_F Some Post-Graduate          0.0270 NA      0.0317
## 10 Education_F Two-Year Degree             0.108   0.152  0.111
## # ... with 32 more rows
```

Cluster 1 appears to be moderate to liberal Independents and Democrats who are more suburban, support gun control, and contain a higher proportion of women and African Americans.
Cluster 2 appears to be moderate to liberal Independents and Democrats who are more ubran and support gun control.

Cluster 3 appears to be conservative Republicans who opposed gun control and are disproportionately white and male.

```
(nested.models %>% filter(k == 3))$D[[1]] %>% full_join(datB) %>%
  select(-(Spend_Science:Spend_Education)) %>%
  group_by(CK) %>% mutate(N = n()) %>%
  gather(key = category, value = value, -ID, -CW, -CK, -Age, -N) %>%
  group_by(CK, category, value) %>%
  summarize(n = n()/N[1]) %>%
  spread(key = CK, value = n)
```

```
## # A tibble: 42 x 5
## # Groups:   category [9]
##    category    value                           `1`    `2`     `3`
##    <chr>       <chr>                         <dbl>  <dbl>   <dbl>
##  1 Area_F      Rural                        0.246  0.197   0.180
##  2 Area_F      Suburban                     0.410  0.459   0.393
##  3 Area_F      Urban                        0.344  0.344   0.426
##  4 Education_F Completed High School        0.262  0.164   0.213
##  5 Education_F Did Not Complete High School 0.0164 0.0656 NA
##  6 Education_F Four-Year Degree             0.197  0.279   0.295
##  7 Education_F Post-Graduate                0.164  0.197   0.197
##  8 Education_F Some College                 0.213  0.148   0.164
##  9 Education_F Some Post-Graduate           0.0164 0.0328  0.0164
## 10 Education_F Two-Year Degree              0.131  0.115   0.115
## # ... with 32 more rows
```

Cluster 1 appears to be white and African American conservative to moderate Republicans and Independents who are split on gun control and slightly more rural than other clusters.
Cluster 2 appears to be predominantly white (and disproportionately male) conservative to moderate Republicans and Independents who oppose gun control
Cluster 3 appears to be white urban and suburban Democrats and Independents who support gun control.