Typology Work New

Ebba Mark

2022-05-12

Set-up

Libraries

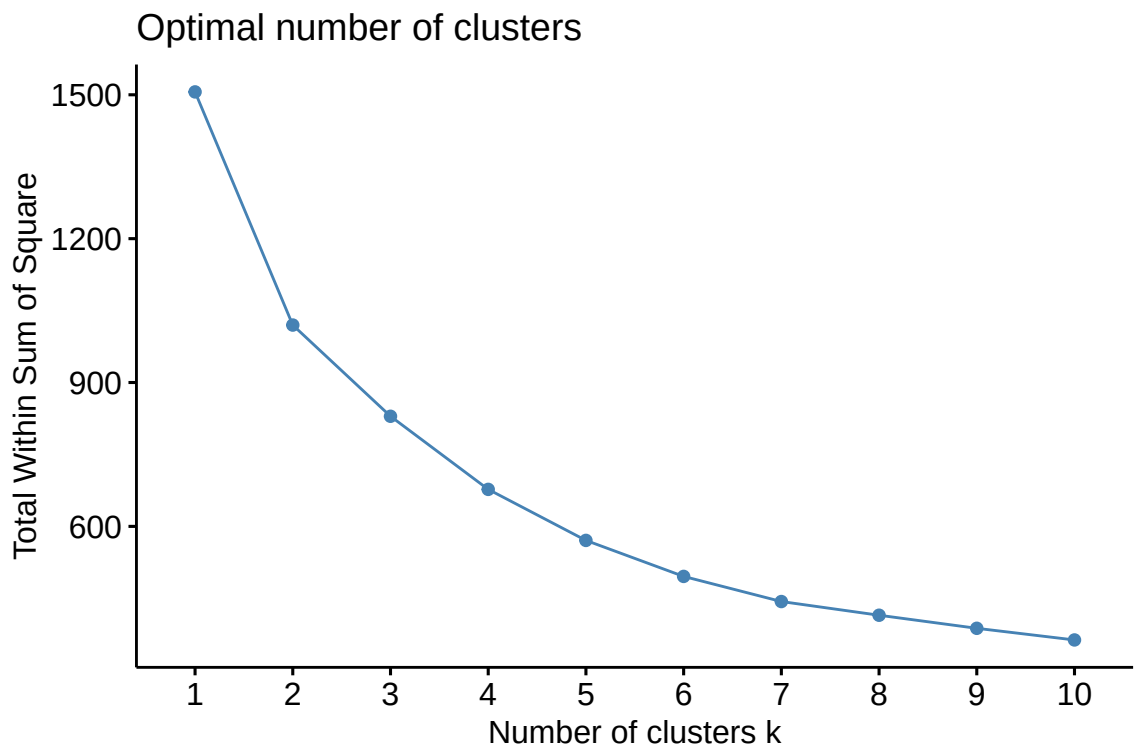
```
# core libraries
library(conflicted)
library(tidyverse)
library(readxl)
library(here)
library(kableExtra)
library(vtable)
library(cluster)
library(factoextra)
library(flexclust)
library(xtable)
library(stargazer)
library(usmap)
library(ggplot2)
library(viridis)

conflict_prefer("filter", "dplyr")
conflict_prefer("select", "dplyr")
source("ref_lists.R")
```

Data

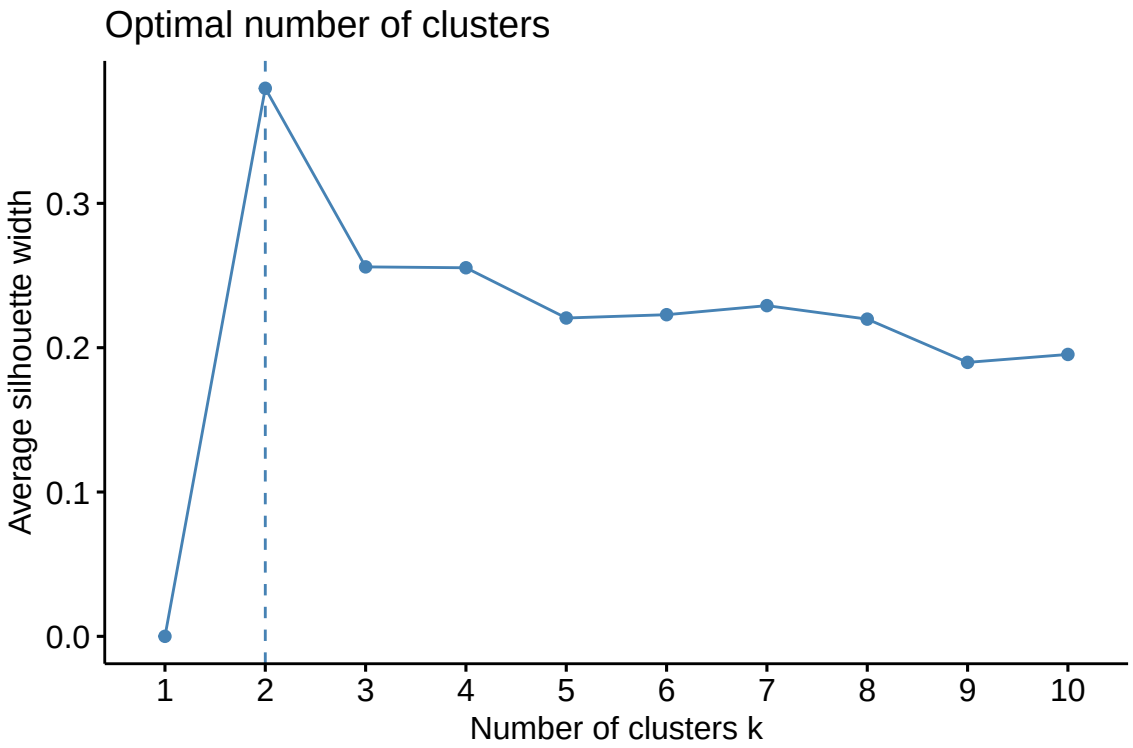
Analysis

```
# Clustering on 251 coal counties claiming analysed in regression section
cluster_coal <- cc[,c("fips","RUC_2013","POPESTIMATE2019",
                          "ed_over_25_bachelor_or_higher", "med_earnings",
                          "lfpr_20_64_female", "diversity_index")]

cluster_coal2 <- data.frame(cluster_coal, row.names = 1)
scaled_coalcluster <- scale(cluster_coal2)
```

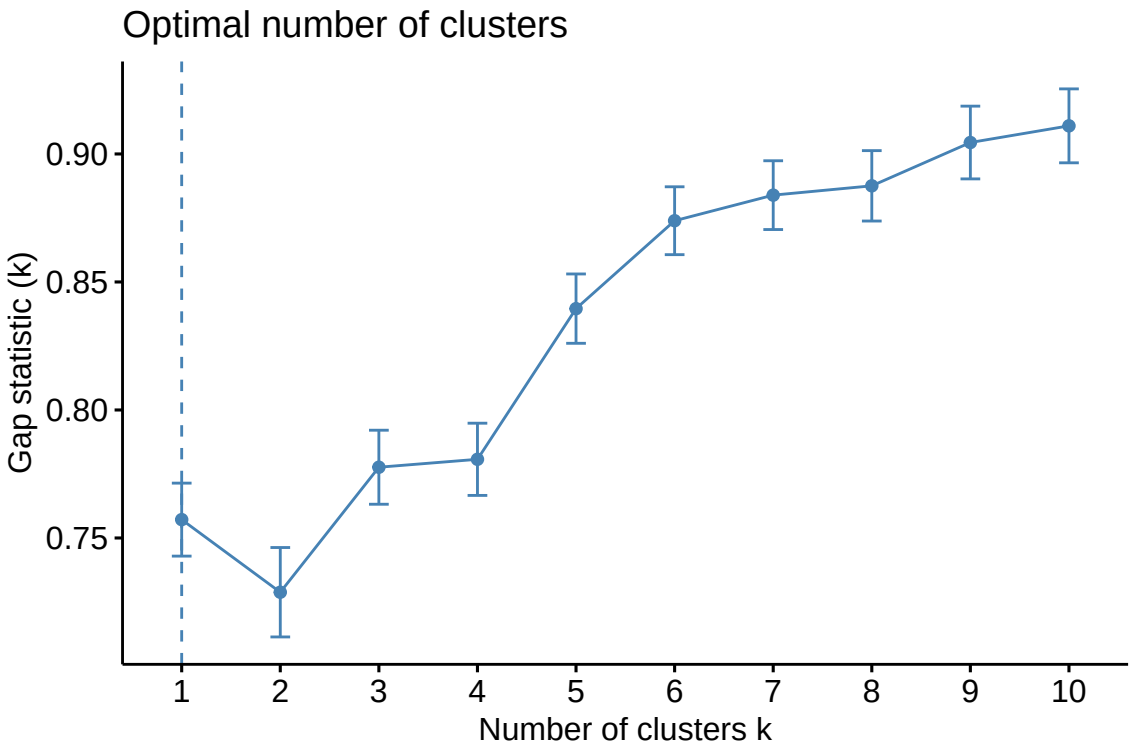
```
# Indicates optimal clusters between 2-4
fviz_nbclust(scaled_coalcluster, hcut, method = "wss")
```




```
# Indicates optimal clusters 2
fviz_nbclust(scaled_coalcluster, hcut, method = "silhouette")
```



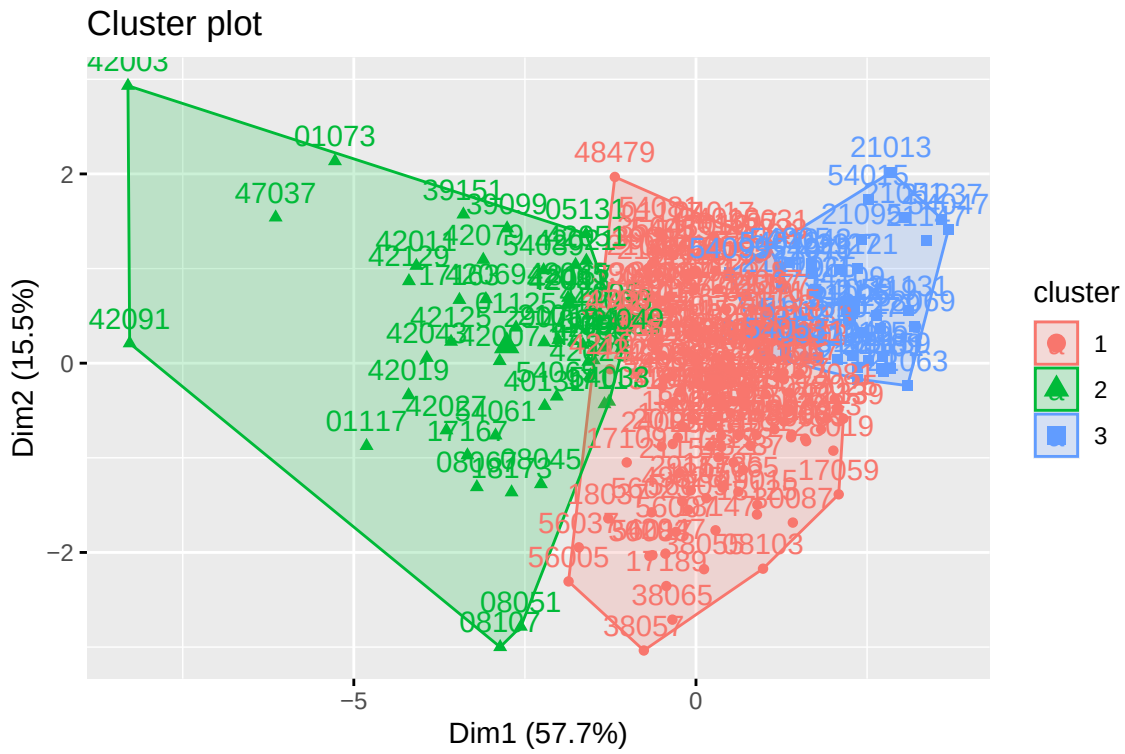

```
# Indicates optimal clusters 1
gap_stat <- clusGap(scaled_coalcluster, hcut, K.max = 10)
fviz_gap_stat(gap_stat)
```



```
res_k3 <- eclust(scaled_coalcluster, k = 3,  "hclust")
res_k3$size
```

```
[1] 160  50  42
```

```
fviz_cluster(res_k3)
```




```
cc <- data.frame(cc, res_k3$cluster)
names(cc)[names(cc) == "res_k3.cluster"] <- "cluster"

ct_voting <- function(clust) {
  n_clust = sum(with(cc, cc$cluster == clust))
  p_16 = sum(with(cc, party_16 == "REPUBLICAN" & cc$cluster == clust))
  p_20 =sum(with(cc, party == "REPUBLICAN" & cc$cluster == clust))
  if(p_16 == p_20) {return(p_16)}else{return(as.list[p_16,p_20])}
}

# Create table of cluster characteristics
cc_output <- as.data.frame(aggregate(cluster_coal2, by=list(cluster=res_k3$cluster), mean))
cc_output <- cbind(cc_output, res_k3$size)
cc_output <- as.data.frame(t(cc_output))
cc_output <- cc_output[, c(2, 1, 3)]
colnames(cc_output) = c("Type 1", "Type 2", "Type 3")
cc_output <- rbind(cc_output, c(ct_voting(2), ct_voting(1), ct_voting(3)))
typ_us$party[which(typ_us$fips == "06077")] = "DEMOCRAT"
us_avg <- t(summarize(typ_us, cluster = "us_total", mean(RUC_2013),mean(POPESTIMATE2019),mean(ed_over_25
cc_output1 <- cbind(cc_output, us_avg)
```

```r
# NOTE THAT FIVE VIRGINIA COUNTIES ARE MISSING FROM THIS GROUP
typ_us$fips[which(is.na(typ_us$diversity_index))]
```

```
[1] "51091" "51093" "51095" "51097" "51099"
```

```r
cc_output1 <- cc_output1[-1,]
print(cc_output1)
```

```
                                    Type 1      Type 2       Type 3    us_avg
RUC_2013                      2.540000e+00     5.28125     7.261905  4.988091
POPESTIMATE2019               1.942618e+05 36100.10625 19143.380952  104726.9
ed_over_25_bachelor_or_higher 2.911000e+01    16.50250    11.423810   21.9422
med_earnings                  3.868624e+04 33947.76250 29845.214286  35826.82
lfpr_20_64_female             7.265600e+01    66.09063    49.207143  69.69936
diversity_index               7.219598e-01     1.01777     1.112773 0.9980209
res_k3$size                   5.000000e+01   160.00000    42.000000      3107
9                             3.600000e+01   156.00000    41.000000      2568
```

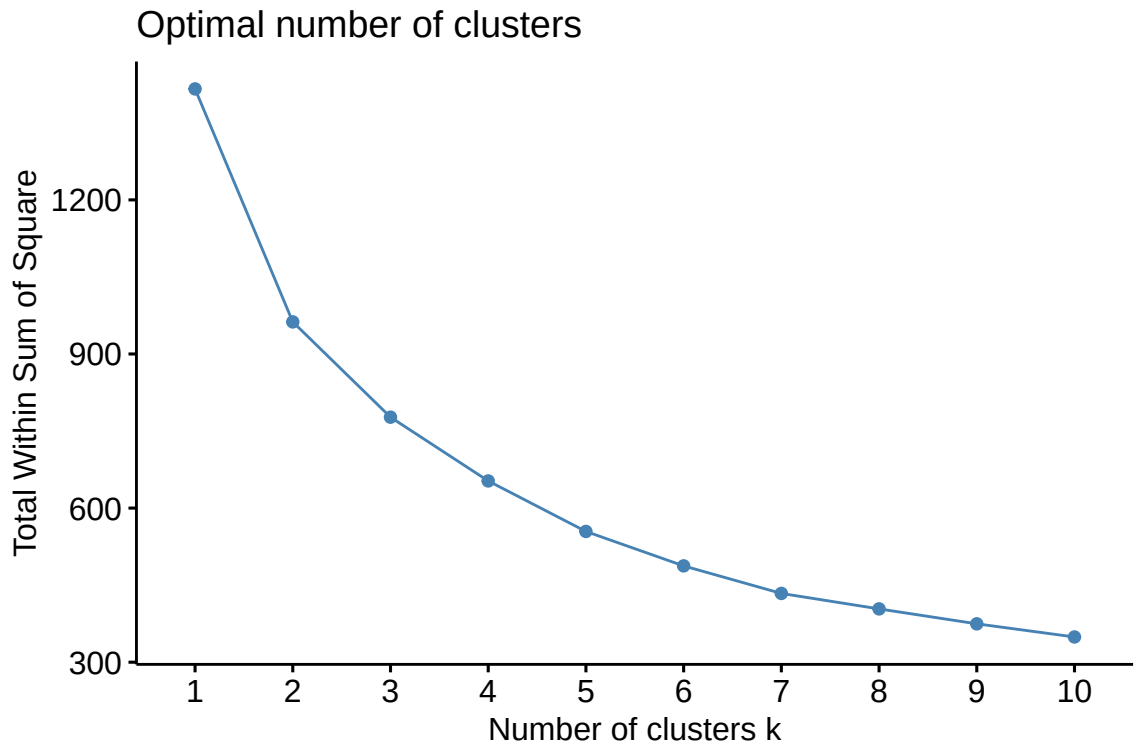
```r
cluster_coal <- cc[,c("fips","RUC_2013","POPESTIMATE2019",
                           "ed_over_25_bachelor_or_higher", "med_earnings",
                           "lfpr_20_64_female", "diversity_index", "party", "party_16")]

cluster_coal$party_20_bin <- ifelse(cluster_coal$party == "REPUBLICAN", 1, 0)
cluster_coal$party_16_bin <- ifelse(cluster_coal$party_16 == "REPUBLICAN", 1, 0)

stargazer(as.data.frame(cluster_coal[,c("RUC_2013","POPESTIMATE2019",
                           "ed_over_25_bachelor_or_higher", "med_earnings",
                           "lfpr_20_64_female", "diversity_index", "party_20_bin", "party_16_bin")]), c
          covariate.labels=c("2013 Rural-Urban Code","Population Size",
                             "Educational Attainment: Bachelor's Degree or Higher %; aged 25-64",
                             "Median Earnings USD","Female Labour Force Participation %; aged 25-64",
                             "Chmura Diversity Index", "Voted for the Republican Party in the 2020 Elec
```

```
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard
% Date and time: Thu, May 12, 2022 - 13:34:38
\begin{table}[!htbp] \centering
  \caption{}
  \label{}
\begin{tabular}{@{\extracolsep{5pt}}lccccccc}
\\[-1.8ex]\hline
\hline \\[-1.8ex]
Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multicolu
\hline \\[-1.8ex]
2013 Rural-Urban Code & 252 & 5.1 & 2.4 & 1 & 3 & 7 & 9 \\
Population Size & 252 & 64,655.3 & 121,468.2 & 1,959 & 16,684.5 & 64,320.2 & 1,216,045 \\
Educational Attainment: Bachelor's Degree or Higher %; aged 25-64 & 252 & 18.2 & 7.7 & 5.4 & 13.6 & 20.
Median Earnings USD & 252 & 34,204.2 & 4,935.0 & 20,268 & 31,357 & 36,705.2 & 54,754 \\
Female Labour Force Participation %; aged 25-64 & 252 & 64.6 & 8.9 & 37.9 & 60.3 & 71.1 & 81.7 \\
Chmura Diversity Index & 252 & 1.0 & 0.2 & 0.5 & 0.9 & 1.1 & 1.5 \\
Voted for the Republican Party in the 2020 Election & 252 & 0.9 & 0.3 & 0 & 1 & 1 & 1 \\
Voted for the Republican Party in the 2020 Election & 252 & 0.9 & 0.3 & 0 & 1 & 1 & 1 \\
```

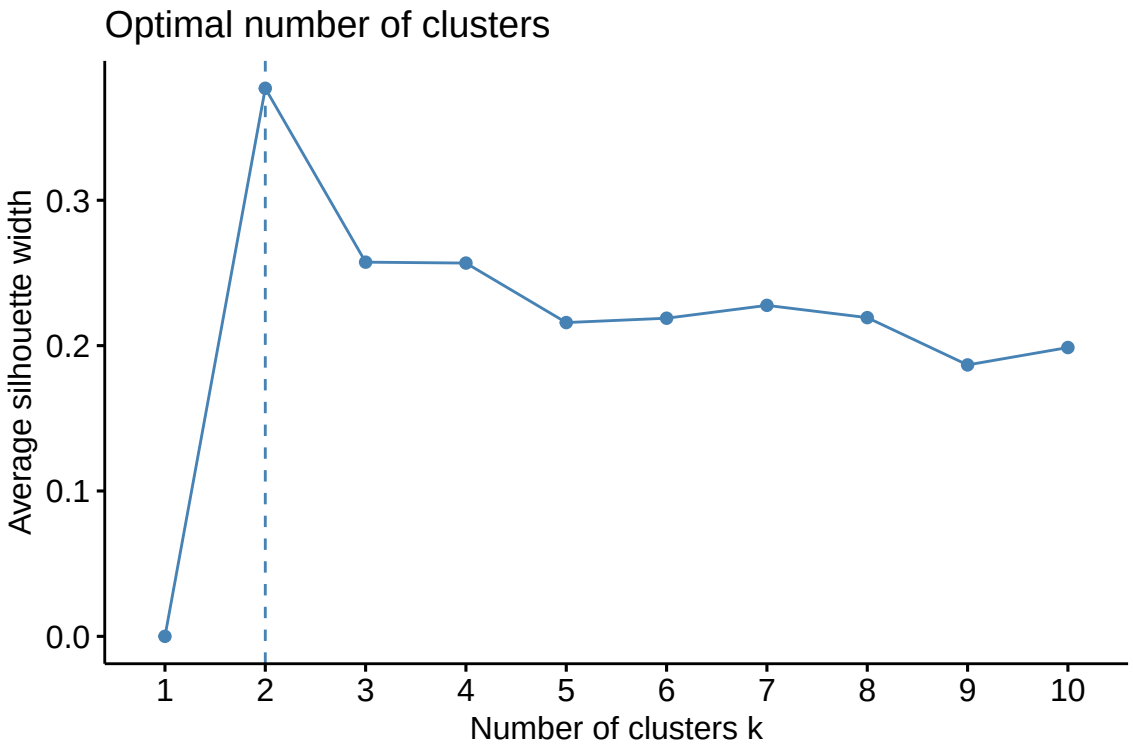
```
\hline \\[-1.8ex]
\multicolumn{8}{l}{* The number of coal counties included in the typology (252) differs from the number
\end{tabular}
\end{table}
```

Clustering on 237 coal counties suggested from PDMIF work

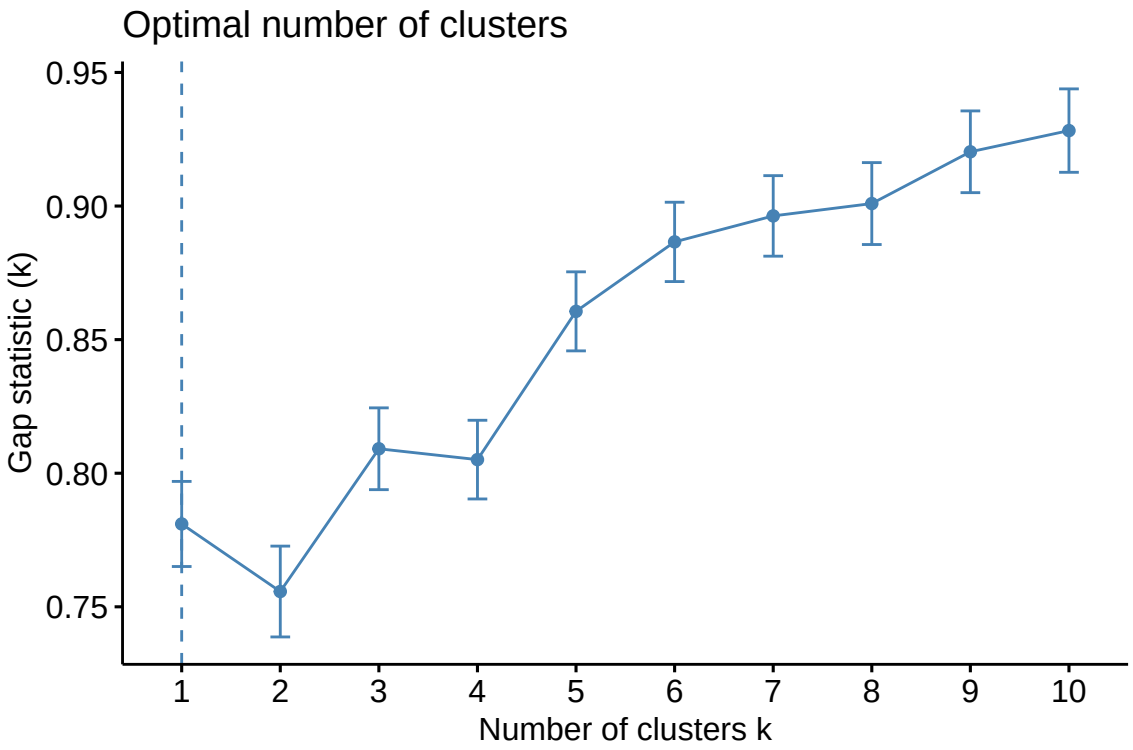
```
cluster_coal_pdmif <- cc_pdmif[,c("fips","RUC_2013","POPESTIMATE2019",
                            "ed_over_25_bachelor_or_higher", "med_earnings",
                            "lfpr_20_64_female", "diversity_index")]

cluster_coal3 <- data.frame(cluster_coal_pdmif, row.names = 1)
pdmifscaled_coalcluster <- scale(cluster_coal3)

# Indicates optimal clusters between 2-4
fviz_nbclust(pdmifscaled_coalcluster, hcut, method = "wss")
```




```
# Indicates optimal clusters 2
fviz_nbclust(pdmifscaled_coalcluster, hcut, method = "silhouette")
```



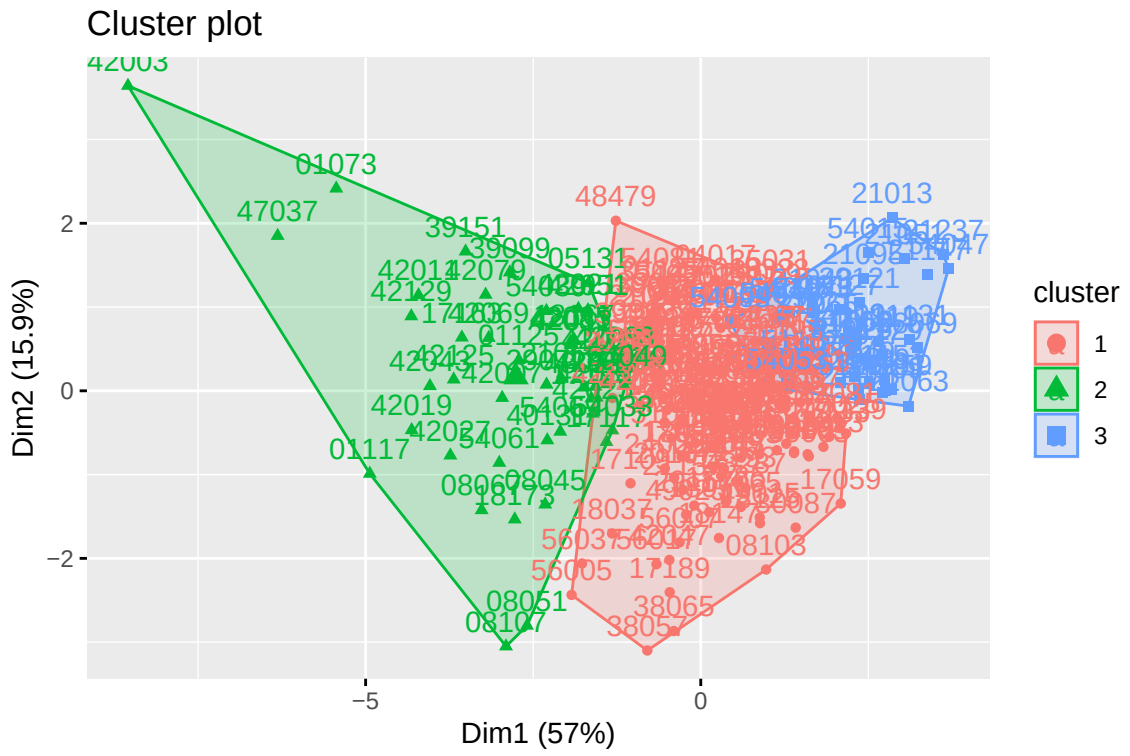

```
# Indicates optimal clusters 1
gap_stat <- clusGap(pdmifscaled_coalcluster, hcut, K.max = 10)
fviz_gap_stat(gap_stat)
```



```
res2_k3 <- eclust(pdmifscaled_coalcluster, k = 3,  "hclust")
res2_k3$size
```

```
[1] 149  47  41
```

```
fviz_cluster(res2_k3)
```




```
cc_pdmif <- data.frame(cc_pdmif, res2_k3$cluster)
names(cc_pdmif)[names(cc_pdmif) == "res2_k3.cluster"] <- "cluster"

ct_voting <- function(clust) {
  n_clust = sum(with(cc, cc$cluster == clust))
  p_16 = sum(with(cc, party_16 == "REPUBLICAN" & cc$cluster == clust))
  p_20 =sum(with(cc, party == "REPUBLICAN" & cc$cluster == clust))
  if(p_16 == p_20) {return(p_16)}else{return(as.list[p_16,p_20])}
}

# Create table of cluster characteristics
cc_pdmif_output <- as.data.frame(aggregate(cluster_coal3, by=list(cluster=res2_k3$cluster), mean))
cc_pdmif_output <- cbind(cc_pdmif_output, res2_k3$size)
cc_pdmif_output <- as.data.frame(t(cc_pdmif_output))
cc_pdmif_output <- cc_pdmif_output[, c(2, 1, 3)]
colnames(cc_pdmif_output) = c("Type 1", "Type 2", "Type 3")
cc_pdmif_output <- rbind(cc_pdmif_output, c(ct_voting(2), ct_voting(1), ct_voting(3)))
typ_us$party[which(typ_us$fips == "06077")] = "DEMOCRAT"
us_avg <- t(summarize(typ_us, cluster = "us_total", mean(RUC_2013),mean(POPESTIMATE2019),mean(ed_over_25
cc_output2 <- cbind(cc_pdmif_output, us_avg)
```

```
# NOTE THAT FIVE VIRGINIA COUNTIES ARE MISSING FROM THIS GROUP
typ_us$fips[which(is.na(typ_us$diversity_index))]
```

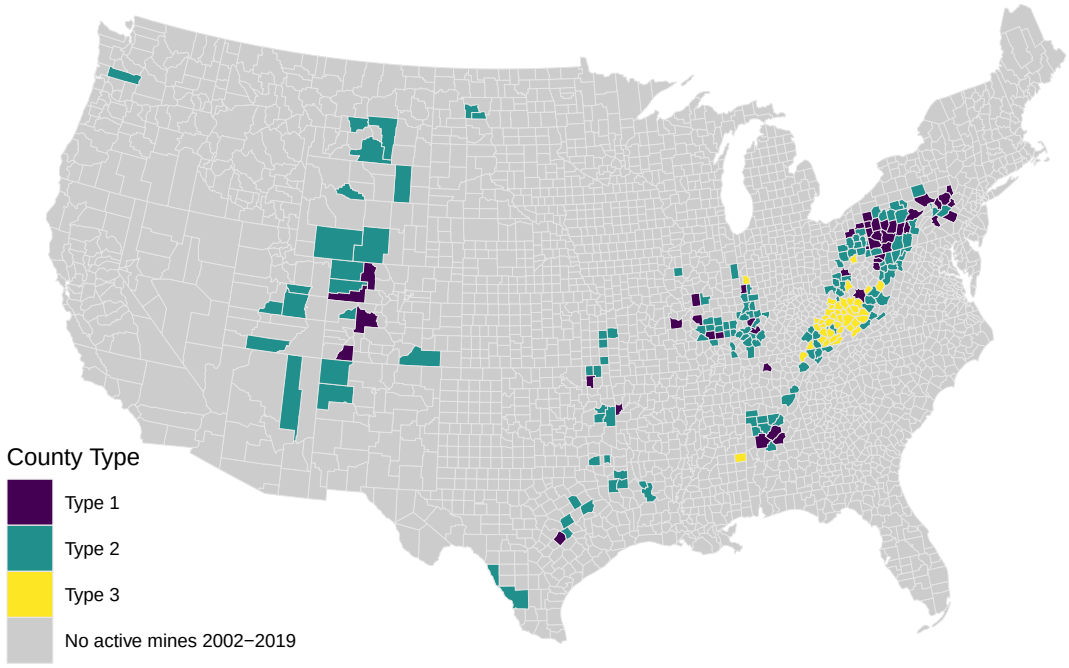
```
[1] "51091" "51093" "51095" "51097" "51099"
```

```
cc_output2 <- cc_output2[-1,]
print(cc_output2)
```

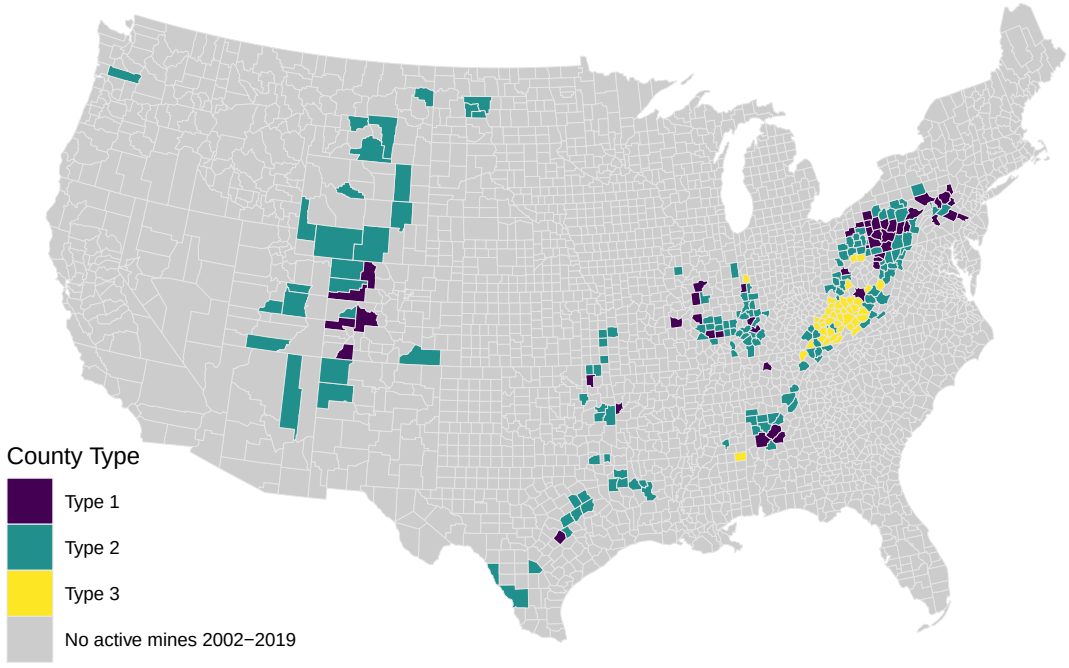
```
                                  Type 1       Type 2       Type 3    us_avg
RUC_2013                      2.531915e+00     5.281879     7.243902  4.988091
POPESTIMATE2019               1.839307e+05 36835.684564 19277.268293  104726.9
ed_over_25_bachelor_or_higher 2.864894e+01    16.462416    11.397561   21.9422
med_earnings                  3.835700e+04 33820.939597 29801.268293  35826.82
lfpr_20_64_female             7.249362e+01    66.044966    49.053659  69.69936
diversity_index               7.227777e-01     1.013947     1.113904 0.9980209
res2_k3$size                  4.700000e+01   149.000000    41.000000      3107
9                             3.600000e+01   156.000000    41.000000      2568
```

```
cluster_coal_pdmif <- cc_pdmif[,c("fips","RUC_2013","POPESTIMATE2019",
                           "ed_over_25_bachelor_or_higher", "med_earnings",
                           "lfpr_20_64_female", "diversity_index", "party", "party_16")]

cluster_coal_pdmif$party_20_bin <- ifelse(cluster_coal_pdmif$party == "REPUBLICAN", 1, 0)
cluster_coal_pdmif$party_16_bin <- ifelse(cluster_coal_pdmif$party_16 == "REPUBLICAN", 1, 0)
```

```
cluster_map <- read_excel(here("data/cc_clusters_251.xlsx"))
cluster_map$type = 0
cluster_map$type[which(cluster_map$cluster == 1)] = 2
cluster_map$type[which(cluster_map$cluster == 2)] = 1
cluster_map$type[which(cluster_map$cluster == 3)] = 3

pdmif_map <- subset(cluster_map, fips %in% cc_pdmif$fips)
cluster_map$type = as.factor(cluster_map$type)
pdmif_map$type = as.factor(pdmif_map$type)


plot_usmap(data = pdmif_map, values = "type", regions = "counties", col = "gray90", size = 0.04, exclud
  scale_fill_viridis_d(name = "County Type", labels = c("Type 1", "Type 2", "Type 3", "No active mines
    theme(panel.background = element_rect(color = "white", fill = "white"),
          plot.title = element_text(face = "bold"), legend.background=element_blank())
```




```
plot_usmap(data = cluster_map, values = "type", regions = "counties", col = "gray90", size = 0.04, exclu
  scale_fill_viridis_d(name = "County Type", labels = c("Type 1", "Type 2", "Type 3", "No active mines 
    theme(panel.background = element_rect(color = "white", fill = "white"),
          plot.title = element_text(face = "bold"), legend.background=element_blank())
```



```
#ggsave("County_types_252.jpg", units="in", width=9, height=6, dpi=300)
```

Create df with clusters for regressions

```
cc_cluster <- cc[c("fips", "cluster")]
cc_cluster$type[which(cc_cluster$cluster == 1)] = 2
cc_cluster$type[which(cc_cluster$cluster == 2)] = 1
cc_cluster$type[which(cc_cluster$cluster == 3)] = 3

cc_cluster$fips[which(cc_cluster$fips == "51195" | cc_cluster$fips == "51720")] <- "51955"
cc_cluster <- cc_cluster[!duplicated(cc_cluster$fips),]

# writexl::write_xlsx(cc_cluster, here("Final Data Products/cc_clusters_251.xlsx"))
```

Testing on Oil and Gas

Testing on entire country