Exploratory Cluster Analysis for Josie Schafer

Michael Flynn, Prior Analytics, LLC.

Cleaning Data

This first code block loads the data and performs any necessary cleaning, rescaling, etc.

First, there don't appear to be any missing values in any rows.

Second, for now I'm focusing primarily on broader demographic and institutional indicators for now, but also some more targeted variables that would likely help to explain disparate economic outcomes. For example, high-research universities granting PhDs or a higher number of community hospital beds. There are other variables that we could include that might be useful for some purposes (e.g. Medicare recipients by region) but I expect that these will be closely tracking other age-related demographic variables.

Third, I'm rescaling the variables by dividing the observed value by the largest value of X as follows:

$$\frac{X_i}{\max X}$$

This puts all observed values on a $0 - 1$ scale.

My understanding of clustering techniques is that when they calculate the distance between units, they will treat the scale of the variables equivalently. The idea here is to scale all of the cluster inputs so they are all on a $0 - 1$ scale, thereby treating all of them equivalently. That way variables with large values and large ranges don't dominate the clustering procedure.

That said, if there's reason to want to weight input variables differently for clustering we can explore that with more time.

```
# Read in data and select relevant variables for clusters.
# Focus is on demographic and anchor institution variables.

# Read in raw data file
data <- readxl::read_xlsx(here("data/anchor regions analysis.xlsx"))
```

```r
# List of variables to include in clustering
varlist <- c("totpop_19",  # Total pop
             "popchange",  # pop change
             "medage",     # Median age
             "labfor",     # percent population in labor force
             "pov",       # percent population living in poverty
             "poc",       # people of color as percent of pop
             "highed",     # Percent population with at least bachelor's
             "forborn",    # Percent population foreign born
             "net_mig",    # Net domestic migration
             "highered_emp_qcew",
             "highered_estab_qcew",
             "hospital_emp_qcew",
             "hospital_estab_qcew",
             "inst_ipeds_enrollment_all",
             "inst_ipeds_doctoralunihighrese",
             "inst_ipeds_pellawards",
             "inst_hosp_ahacommunityhospitals",
             "inst_hosp_ahabeds",
             "inst_hosp_nihresearchfunding")

# Rescale the variables from 0-1
data.clean <- data |>
  mutate(across(all_of(varlist), # Variables to scale
              ~.x/max(.x),                                              # Scale relative to
              .names = "{col}_max")) |>                                 # Add "max" suffix
  dplyr::select(MSA, ends_with("_max")) |>                              # select chosen vari
  column_to_rownames("MSA")
```

Clustering Methods

Here I start with agglomerative/hierarchical clustering methods. The goal as I understand it is to find a happy medium number of groups that illustrates the variability across regions and anchor institutions while still being tractable for analyses.

The priority here is to construct clusters on the basis of 1) anchor institution characteristics, and 2) demographic characteristics of the surrounding region. For now I'll combine these into a single cluster, but we may want to think about constructing two clusters, one on the basis of demographic traits and the other on the basis of anchor institution traits. This would help parse out effects later if the client is interested in using these as predictors in subsequent regression analyses.

I'm going to create a few different clusters and we can compare the characteristics and performance of each, and then choose which one the client likes best.

I chose the "complete" method for the `hclust()` function because it generates a better distribution of clusters than the other methods. For example, others tend to produce either very flat distributions, in which case you may just as well use dummy variables for each MSA or city, or they produce oddly concentrated clusters with 80-90% of observatiosn falling into cluster group 2.

```
distance <- dist(data.clean) # calculate Euclidian distance between obs

hc.tree <- hclust(distance, method = "complete") # Create cluster groupings based on dista


# List of distances to use in generating clusters
cluster.size.list <- list("5" = 5,
                          "10" = 10,
                          "15" = 15,
                          "20" = 20,
                          "25" = 25,
                          "30" = 30,
                          "35" = 35,
                          "40" = 40)

cluster.ids <- map(
  .x = seq_along(cluster.size.list),
  .f = ~ cutree(hc.tree, k = cluster.size.list[[.x]])
                 ) |>
  bind_cols()
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
```

```
names(cluster.ids) <- c("cluster_5", "cluster_10", "cluster_15", "cluster_20", "cluster_25

data.out <- data.clean |>
  bind_cols(cluster.ids)
```

Choosing the optimal number of clusters

Figure 1 shows the distribution of the observations depending on the number of clusters chosen. In general, 25–35 clusters seems like a nice balance between parsimony and too much detail. Smaller numbers of clusters, like 5 or 10, group too many areas together (see the spike at group #1). In general we see there are regularly spikes like these, but we start to get more variability as we move towards the 25–30 range.

```
cluster.list <- list("cluster_5", "cluster_10", "cluster_15", "cluster_20", "cluster_25",



plot.out <- data.out |>
  dplyr::select(starts_with("cluster")) |>
  map2(
    .y = cluster.list,
    .f = ~{ggplot(data.out, aes(x = .)) +
    geom_histogram(bins = 40) +
        labs(title = .y)}
)


patchwork::wrap_plots(plot.out)
```

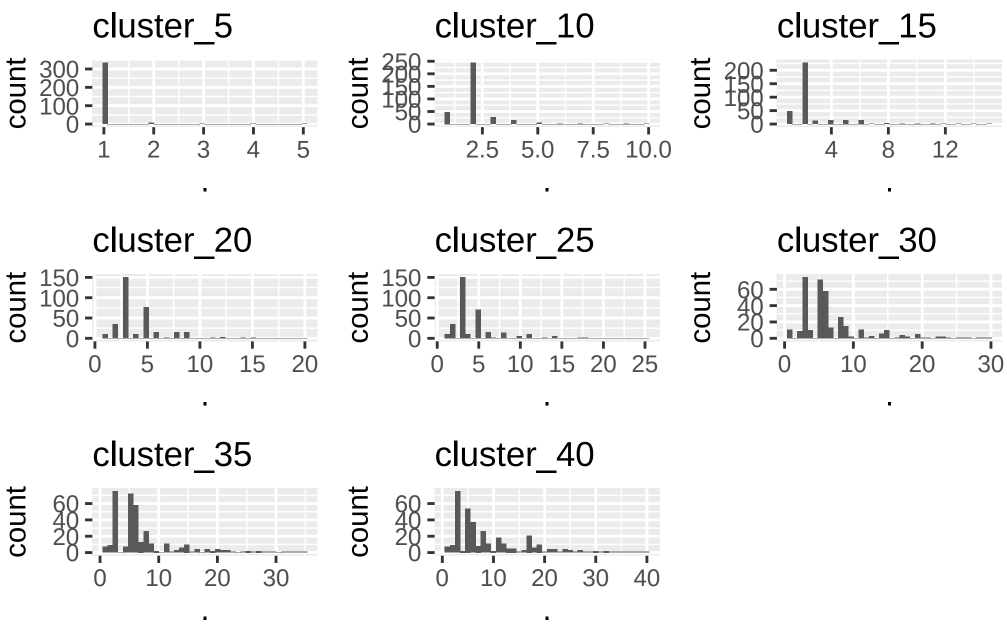

Figure 1: Histograms showing the distribution of clusters depending on the number of clusters chosen.

```
table.data <- data |>
  bind_cols(cluster.ids) |>
  dplyr::select(varlist, cluster_30) |>
  group_by(cluster_30) |>
  dplyr::summarise(across(everything(),
                          mean))
```

```
Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
i Please use `all_of()` or `any_of()` instead.
  # Was:
  data %>% select(varlist)

  # Now:
  data %>% select(all_of(varlist))

See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```
names(table.data) <- c("Cluster",
                        "Total Population (2019)",
                       "Population Change",
```

```r
                        "Median Age",
                        "% Population in Labor Force",
                        "% Population in Poverty",
                        "% Population People of Color",
                        "$% Population with Bachelor's Degree",
                        "% Population Foreign Born",
                        "Net Domestic Migration",
                        "Higher Education Employment",
                        "Higher Education Establishments",
                        "Hospital Employment",
                        "Hospital Establishments",
                        "Higher Education Enrollment",
                        "High Research Doctoral Degree Institutions",
                        "Total Pell Grant Amounts Awarded",
                        "Hospitals/Community Hospitals",
                        "Hospital Beds",
                        "NIH Research Funding")



table.out <- table.data |>
  kbl(longtable = TRUE) |>
  kable_styling(font_size = 8) |>
  scroll_box(height = "600px", width = "800px")

table.out
```

Cluster	Total Population (2019)	Population Change	Median Age	% Population in Labor Force	% Population in Poverty	% Po
1	491022.2	4.987475	38.27273	64.10909	11.21818	
2	994230.7	7.819001	37.26667	63.83333	14.40000	
3	214126.1	4.060407	39.85067	58.44000	15.88533	
4	2319054.6	19.071504	36.69000	67.13000	11.48000	
5	246575.0	7.490234	38.01528	64.73333	11.86806	
6	577068.9	8.442607	37.72586	65.01379	12.75517	
7	253319.8	10.775605	32.23077	61.05385	19.19231	
8	299882.5	9.290796	35.95000	59.55769	19.82692	
9	447905.1	8.002275	32.08667	58.36000	22.06000	
10	575400.0	2.552712	39.55000	66.70000	8.80000	
11	342088.6	10.687640	50.82727	47.78182	14.18182	
12	125044.0	0.000000	67.40000	22.50000	8.20000	
13	637543.3	18.178695	50.60000	52.03333	11.53333	
14	328124.3	9.772405	39.70000	66.56667	10.50000	
15	2197690.2	6.524644	37.91000	67.07000	11.60000	
16	404417.0	227.829477	45.50000	57.40000	12.60000	
17	386114.5	72.536067	40.30000	58.55000	17.30000	
18	2821709.5	31.846792	39.80000	63.45000	12.60000	

19	3461420.8	6.809083	37.22000	63.52000	12.88000
20	6090660.0	11.166374	41.00000	63.00000	14.60000
21	4761603.0	16.685736	36.70000	62.80000	13.70000
22	3344589.0	10.782462	38.05000	67.55000	8.25000
23	6373281.0	17.481792	35.35000	66.95000	12.90000
24	6196585.0	14.397978	37.00000	71.60000	7.80000
25	7320663.0	18.952678	34.80000	68.80000	11.70000
26	4832346.0	7.642613	38.70000	69.20000	9.30000
27	6079130.0	2.833259	38.80000	65.30000	12.40000
28	9508605.0	1.320708	37.50000	66.80000	11.80000
29	13249614.0	4.132679	36.80000	64.90000	13.90000
30	19294236.0	3.173788	38.60000	64.70000	12.80000

Modeling Exploration

Here I'm just running a few models that look at how the clusters perform in predicting outcomes of interest. Again, this is something we can revisit given more time and some discussion to inject more domain knowledge into things.

```r
model.data <- data |>
  bind_cols(cluster.ids)

# GDP index models
m1 <- lm(index_real_gdp_21 ~ cluster_10, data = model.data)
m2 <- lm(index_real_gdp_21 ~ cluster_20, data = model.data)
m3 <- lm(index_real_gdp_21 ~ cluster_30, data = model.data)
m4 <- lm(index_real_gdp_21 ~ cluster_40, data = model.data)

# GDP index models
m5 <- lm(percapita_personal_income_21 ~ cluster_10, data = model.data)
m6 <- lm(percapita_personal_income_21 ~ cluster_20, data = model.data)
m7 <- lm(percapita_personal_income_21 ~ cluster_30, data = model.data)
m8 <- lm(percapita_personal_income_21 ~ cluster_40, data = model.data)


mlist <- list(m1, m2, m3, m4, m5, m6, m7, m8)

modelsummary(mlist,
             fmt = 3,
             estimate = "estimate",
             statistic = "std.error",
             output = "kableExtra") |>
  kable_styling("striped") |>
```

	2021 GDP Index (1-4)				2021 Per Capita Income (5-8)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8
(Intercept)	109.207	109.875	112.168	111.634	49 714.224	50 077.038	51 695.772	51 401
	(2.083)	(1.732)	(1.604)	(1.481)	(1416.708)	(1161.138)	(1078.270)	(984.
cluster_10	3.512				3367.677			
	(0.841)				(572.083)			
cluster_20		1.585				1581.777		
		(0.326)				(218.236)		
cluster_30			0.691				786.528	
			(0.186)				(125.154)	
cluster_40				0.556				599.
				(0.119)				(79.1
Num.Obs.	347	347	347	347	347	347	347	34
R2	0.048	0.064	0.038	0.059	0.091	0.132	0.103	0.1
R2 Adj.	0.045	0.062	0.036	0.057	0.089	0.130	0.100	0.1
AIC	2985.5	2979.5	2989.0	2981.4	7511.9	7495.9	7507.5	749
BIC	2997.0	2991.1	3000.5	2992.9	7523.4	7507.4	7519.0	750
Log.Lik.	−1489.730	−1486.761	−1491.490	−1487.678	−3752.929	−3744.944	−3750.730	−374
F	17.429	23.685	13.772	21.742	34.653	52.534	39.495	57.3
RMSE	17.71	17.56	17.80	17.61	12 045.16	11 771.15	11 969.08	11 69

```r
  add_header_above(c(" " = 1, "2021 GDP Index (1-4)" = 4, "2021 Per Capita Income (5-8)" =
```

```r
# GDP index models
m1 <- lmer(index_real_gdp_21 ~ cluster_10 + (1|State), data = model.data)
m2 <- lmer(index_real_gdp_21 ~ cluster_20 + (1|State), data = model.data)
m3 <- lmer(index_real_gdp_21 ~ cluster_30 + (1|State), data = model.data)
m4 <- lmer(index_real_gdp_21 ~ cluster_40 + (1|State), data = model.data)

# GDP inerdex models
m5 <- lmer(percapita_personal_income_21 ~ cluster_10 + (1|State), data = model.data)
m6 <- lmer(percapita_personal_income_21 ~ cluster_20 + (1|State), data = model.data)
m7 <- lmer(percapita_personal_income_21 ~ cluster_30 + (1|State), data = model.data)
m8 <- lmer(percapita_personal_income_21 ~ cluster_40 + (1|State), data = model.data)


mlist <- list(m1, m2, m3, m4, m5, m6, m7, m8)

modelsummary(mlist,
```

	2021 GDP Index (1-4)				2021 Per Capita Income (5-8)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	108.983	109.456	111.362	110.357	50 226.250	50 298.044	51 869.232	51 621.1
	(2.304)	(2.035)	(1.912)	(1.815)	(1474.526)	(1233.525)	(1166.839)	(1057.43
cluster_10	2.903				3122.560			
	(0.812)				(579.139)			
cluster_20		1.306				1522.832		
		(0.316)				(220.859)		
cluster_30			0.569				752.824	
			(0.175)				(124.762)	
cluster_40				0.513				571.35
				(0.111)				(79.019
SD (Intercept State)	8.960	9.035	9.078	9.292	2768.106	2829.001	3234.947	2790.27
SD (Observations)	15.310	15.192	15.339	15.038	11 745.817	11 440.996	11 536.033	11 376.8
Num.Obs.	347	347	347	347	347	347	347	347
R2 Marg.	0.033	0.044	0.026	0.051	0.079	0.124	0.095	0.131
R2 Cond.	0.280	0.294	0.279	0.313	0.128	0.174	0.161	0.180
AIC	2940.9	2938.7	2946.1	2936.7	7480.4	7465.4	7476.1	7463.4
BIC	2956.3	2954.0	2961.5	2952.1	7495.8	7480.8	7491.5	7478.8
ICC	0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.1
RMSE	14.40	14.28	14.42	14.10	11 491.47	11 177.61	11 224.92	11 117.3

State used as grouping term.

```r
              fmt = 3,
              estimate = "estimate",
              statistic = "std.error",
              notes = c("State used as grouping term."),
              output = "kableExtra") |>
  kable_styling("striped") |>
  add_header_above(c(" " = 1, "2021 GDP Index (1-4)" = 4, "2021 Per Capita Income (5-8)" =
```