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.

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.

```r
# 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"))

# List of variables to include in clustering
varlist <- c("totpop_19",
             "popchange",
             "real_gdp_21",
             "labfor",
             "pov",
             "poc",
             "highed",
             "forborn",
             "net_mig",
             "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(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")
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `across(varlist, ~.x/max(.x), .names = "{col}_max")`.
Caused by 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>.
```

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.

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

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


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

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`
```

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

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

```r
table.out <- kbl(data.out) |>
  kable_styling(font_size = 9)

table.out
```

	totpop_19_max	popchange_max	real_gdp_21_max	labfor_max
Bridgeport-Stamford-NorwalkCT	0.0489227	0.0187062	0.0495386	0.9037433
Virginia Beach-Norfolk-Newport NewsVA-NC	0.0913086	0.0260385	0.0551656	0.8930481
LimaOH	0.0053475	-0.0140466	0.0047577	0.8315508
Orlando-Kissimmee-SanfordFL	0.1300373	0.0896005	0.0865064	0.8582888
Pensacola-Ferry Pass-BrentFL	0.0253053	0.0418151	0.0123642	0.8048128
New Orleans-MetairieLA	0.0657076	0.0646487	0.0435803	0.8355615
SpartanburgSC	0.0159435	0.0464697	0.0093690	0.8221925
BismarckND	0.0066083	0.0916022	0.0042012	0.9491979
BillingsMT	0.0092811	0.0713106	0.0062377	0.8863636
HuntsvilleAL	0.0236860	0.0604353	0.0181842	0.8382353
Athens-Clarke CountyGA	0.0108041	0.0453603	0.0058690	0.8208556
JohnstownPA	0.0068937	-0.0355771	0.0026744	0.7339572
Hagerstown-MartinsburgMD-WV	0.0146752	0.0306810	0.0064328	0.8288770
HattiesburgMS	0.0087164	0.0917819	0.0036654	0.8101604
El PasoTX	0.0435610	0.0387597	0.0192853	0.8288770
ColumbiaSC	0.0427215	0.0472655	0.0247162	0.8582888
Kahului-Wailuku-LahainaHI	0.0086025	0.0000000	0.0052180	0.8930481
EvansvilleIN-KY	0.0163240	-0.0504403	0.0109535	0.8449198
Florence-Muscle ShoalsAL	0.0076358	0.0033593	0.0030579	0.7339572
WilmingtonNC	0.0149442	-0.0768353	0.0092775	0.8235294
East StroudsburgPA	0.0087089	0.0000000	0.0042460	0.8262032
MissoulaMT	0.0060800	0.0409968	0.0034269	0.9371658
WinchesterVA-WV	0.0071328	0.0428451	0.0042788	0.8328877
ReadingPA	0.0216658	0.0115467	0.0115861	0.8729947
ColumbusGA-AL	0.0165543	0.0441611	0.0083706	0.7981283
Lake Havasu City-KingmanAZ	0.0107646	0.0187710	0.0035250	0.6029412
Davenport-Moline-Rock IslandIA-IL	0.0197559	0.0051718	0.0127006	0.8475936
Des Moines-West Des MoinesIA	0.0352664	0.1012588	0.0317817	0.9572193
SpringfieldIL	0.0108409	0.0024838	0.0066837	0.8502674
Durham-Chapel HillNC	0.0324809	0.1241611	0.0317185	0.8636364
Omaha-Council BluffsNE-IA	0.0482931	0.0446070	0.0381353	0.9451872
LongviewWA	0.0055342	0.0239916	0.0030172	0.7526738
Austin-Round Rock-GeorgetownTX	0.1095893	0.1312995	0.1036072	0.9438503
Sioux CityIA-NE-SD	0.0074554	0.0063520	0.0051249	0.9264706
New BernNC	0.0064675	0.0000000	0.0032921	0.7794118
The VillagesFL	0.0064809	0.0000000	0.0027375	0.3008021
KankakeeIL	0.0057342	-0.0057283	0.0036816	0.8155080
North Port-Sarasota-BradentonFL	0.0416554	0.0687870	0.0228092	0.6818182
Providence-WarwickRI-MA	0.0838731	0.0042298	0.0496741	0.8703209
BakersfieldCA	0.0460055	0.0387153	0.0297355	0.7794118
AlbuquerqueNM	0.0472736	0.0254258	0.0257661	0.8155080
York-HanoverPA	0.0230932	0.0178266	0.0120477	0.8770053
San AngeloTX	0.0062629	0.0446797	0.0051926	0.8622995
Rapid CitySD	0.0071732	0.0546489	0.0038305	0.8970588
Sebring-Avon ParkFL	0.0053610	0.0000000	0.0015194	0.5655080
Lansing-East LansingMI	0.0283386	0.0787429	0.0153930	0.8475936
AnchorageAK	0.0206746	0.0363207	0.0144566	0.9264706
ReddingCA	0.0092884	0.0057215	0.0046898	0.7245989
Deltona-Daytona Beach-Ormond BeachFL	0.0334964	0.1329331	0.0129085	0.6951872
AbileneTX	0.0088456	0.0203918	0.0045748	0.8061497
TucsonAZ	0.0532391	0.0285551	0.0260376	0.7780749
BangorME	0.0078663	-0.0033292	0.0039473	0.8074866
Hilton Head Island-BlufftonSC	0.0111304	0.0000000	0.0057635	0.7553476
OwensboroKY	0.0061405	0.0188920	0.0033439	0.8168449
Vineland-BridgetonNJ	0.0078731	-0.0100233	0.0037212	0.7433155
FargoND-MN	0.0124608	0.0847837	0.0090638	1.0000000
Little Rock-North Little Rock-ConwayAR	0.0381987	0.0355376	0.0219965	0.8342246
State CollegePA	0.0083942	0.0305805	0.0049779	0.7660428
SumterSC	0.0072931	0.1404587	0.0028383	0.7606952
PeoriaIL	0.0210883	0.0359933	0.0120265	0.8181818
Raleigh-CaryNC	0.0690523	0.1077590	0.0575318	0.9237968
Mount Vernon-AnacortesWA	0.0065103	0.0395421	0.0042331	0.7981283
Shreveport-Bossier CityLA	0.0207118	0.0069053	0.0130787	0.7901070