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.

```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",  # 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 observations falling into cluster group 2.

```r
distance <- dist(data.clean) # calculate Euclidean 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 |>
  bind_cols(cluster.ids) |>
  dplyr::select(starts_with("cluster"), Name, State, MSA, varlist)
```

```
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(data.out) <- c("Cluster 5",
                     "Cluster 10",
                     "Cluster 15",
                     "Cluster 20",
                     "Cluster 25",
                     "Cluster 30",
                     "Cluster 35",
                     "Cluster 40",
                     "Name",
                     "State",
                     "MSA",
                     "Total Population (2019)",
                     "Population Change",
                     "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")

write_csv(data.out,
          here::here("data/raw-data-with-cluster-ids.csv"))
```

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.

The “ideal” number of clusters will also depend on modeling considerations as I briefly address below. Depending on what the final models look like you may want to use a smaller number of clusters. At some point there’s going to be a tradeoff between the total number of clusters and the value added with respect to model inputs.

```
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)
```

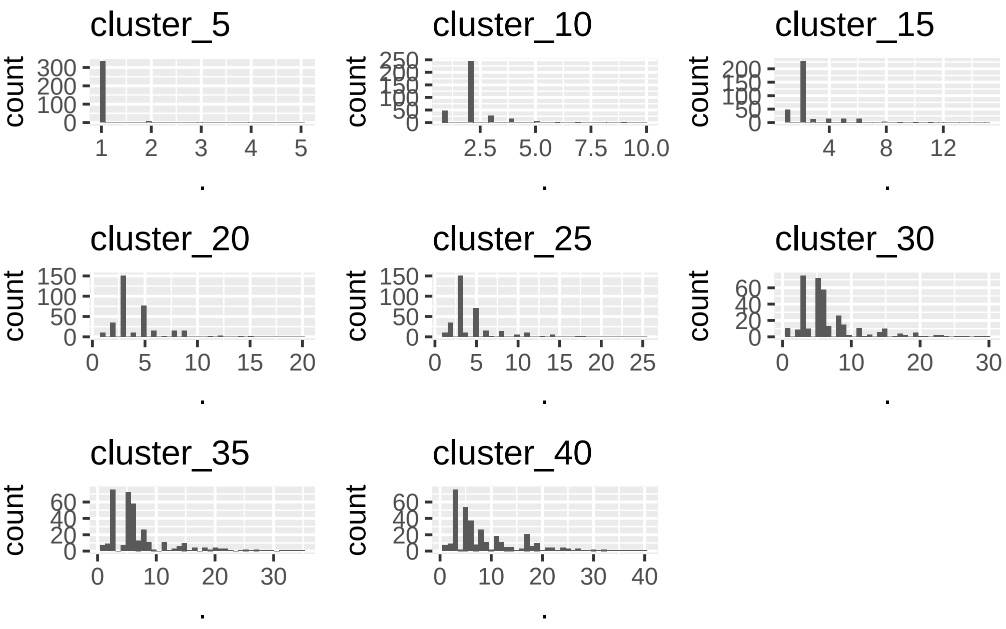


Figure 1: Histograms showing the distribution of clusters depending on the number of clusters chosen.

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

# Save output for table to send to client.
write_csv(table.data,
          here::here("data/table-data-output"))

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

One issue to consider is whether or not the final data will ultimately have more than ~350 observations. 25–35 dummy indicator variables and possibly various other covariates may be a lot relative to the total number of observations.

I use the GDP index and per capita income here because they should facilitate a pretty simple linear model. The residual checks at the end provide some basic support for this. They're not perfectly normally distributed, so in the future some adjustments to the models would be helpful to provide better fit.

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

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

# Per Capita Income models
m5 <- lm(percapita_personal_income_21 ~ factor(cluster_10), data = model.data)
m6 <- lm(percapita_personal_income_21 ~ factor(cluster_20), data = model.data)
m7 <- lm(percapita_personal_income_21 ~ factor(cluster_30), data = model.data)
m8 <- lm(percapita_personal_income_21 ~ factor(cluster_40), data = model.data)


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

modelsummary(mlist,
             fmt = 3,
             stars = TRUE,
             estimate = "estimate",
             statistic = "std.error",
             title = "Linear Regression Models",
             output = "kableExtra") |>
  kable_styling("striped") |>
  add_header_above(c(" " = 1, "2021 GDP Index (1-4)" = 4, "2021 Per Capita Income (5-8)" =
```

```
# GDP index models
mm1 <- lmer(index_real_gdp_21 ~ factor(cluster_10) + (1|State), data = model.data)
mm2 <- lmer(index_real_gdp_21 ~ factor(cluster_20) + (1|State), data = model.data)
mm3 <- lmer(index_real_gdp_21 ~ factor(cluster_30) + (1|State), data = model.data)
mm4 <- lmer(index_real_gdp_21 ~ factor(cluster_40) + (1|State), data = model.data)

# Per Capita Income models
mm5 <- lmer(percapita_personal_income_21 ~ factor(cluster_10) + (1|State), data = model.da
mm6 <- lmer(percapita_personal_income_21 ~ factor(cluster_20) + (1|State), data = model.da
mm7 <- lmer(percapita_personal_income_21 ~ factor(cluster_30) + (1|State), data = model.da
```

```
boundary (singular) fit: see help('isSingular')
```

```
mm8 <- lmer(percapita_personal_income_21 ~ factor(cluster_40) + (1|State), data = model.da
```

```
boundary (singular) fit: see help('isSingular')
```

```
mmlist <- list(mm1, mm2, mm3, mm4, mm5, mm6, mm7, mm8)

modelsummary(mmlist,
             fmt = 3,
             stars = TRUE,
             estimate = "estimate",
             statistic = "std.error",
             title = "Multilevel Regression Models",
             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)" =
```

Table 1: Linear Regression Models

	2021 GDP Index (1-4)				2021 Per Cap	
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	110.142***	119.942***	119.942***	118.225***	57 330.875***	77 711.818***
	(2.485)	(4.930)	(4.772)	(5.925)	(1512.775)	(2800.600)
factor(cluster_10)2	5.932*				−1706.654	
	(2.719)				(1654.896)	
factor(cluster_10)3	18.192***				7508.051**	
	(4.142)				(2521.291)	
factor(cluster_10)4	9.925+				−11 744.475***	
	(5.093)				(3100.265)	
factor(cluster_10)5	10.609				13 624.925**	
	(8.092)				(4925.237)	
factor(cluster_10)6	24.695*				5186.625	
	(12.426)				(7563.873)	
factor(cluster_10)7	68.487***				72 693.625***	
	(12.426)				(7563.873)	
factor(cluster_10)8	14.310				34 959.125**	
	(17.397)				(10 589.422)	
factor(cluster_10)9	6.689				16 575.625*	
	(12.426)				(7563.873)	
factor(cluster_10)10	3.965				27 805.125**	
	(17.397)				(10 589.422)	
factor(cluster_20)2		−12.614*				−27 008.390***
		(5.652)				(3210.672)
factor(cluster_20)3		−7.327				−24 554.865***
		(5.106)				(2900.815)
factor(cluster_20)4		22.018**				−10 919.718**
		(7.144)				(4058.454)
factor(cluster_20)5		−0.734				−18 137.584***
		(5.270)				(2993.967)
factor(cluster_20)6		0.125				−32 125.418***
		(6.491)				(3687.161)
factor(cluster_20)7		−14.445				−16 495.818*
		(12.569)				(7140.156)
factor(cluster_20)8		14.311*				−17 532.085***
		(6.491)				(3687.161)
factor(cluster_20)9		−0.844				−13 554.552***
		(6.491)				(3687.161)
factor(cluster_20)10		4.614				−22 013.818*
		(17.078)				(9701.562)
factor(cluster_20)11		9.538				−17 526.318*
		(12.569)				(7140.156)
factor(cluster_20)12		7.060				−10 519.152+
		(10.650)				(6049.987)
factor(cluster_20)13		13.703				−19 403.818*
		(17.078)				(9701.562)
factor(cluster_20)14		58.688***				52 312.682***
		(12.569)				(7140.156)
factor(cluster_20)15		−8.566				−1111.318
		(12.569)				(7140.156)
factor(cluster_20)16		16.087				−10 984.818

Table 2: Multilevel Regression Models

	2021 GDP Index (1-4)				2021 Per Cap	
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	109.669***	117.942***	116.999***	116.188***	57 999.410***	77 704.001***
	(2.739)	(5.489)	(5.363)	(6.334)	(1658.543)	(2878.854)
factor(cluster_10)2	5.389+				−2322.848	
	(2.745)				(1727.418)	
factor(cluster_10)3	15.282***				5994.450*	
	(3.935)				(2507.244)	
factor(cluster_10)4	5.275				−15 978.639***	
	(4.805)				(3107.068)	
factor(cluster_10)5	7.102				13 210.124**	
	(7.492)				(4787.379)	
factor(cluster_10)6	22.012*				5677.100	
	(11.129)				(7256.584)	
factor(cluster_10)7	61.611***				65 566.506***	
	(11.076)				(7251.389)	
factor(cluster_10)8	14.783				34 290.590**	
	(17.404)				(10 621.931)	
factor(cluster_10)9	3.903				12 580.310+	
	(11.552)				(7341.603)	
factor(cluster_10)10	4.438				27 136.590*	
	(17.404)				(10 621.931)	
factor(cluster_20)2		−10.306+				−26 830.286***
		(6.167)				(3290.978)
factor(cluster_20)3		−6.159				−24 545.327***
		(5.532)				(2976.582)
factor(cluster_20)4		20.102**				−11 020.319**
		(7.218)				(4116.706)
factor(cluster_20)5		2.087				−18 150.019***
		(5.649)				(3065.956)
factor(cluster_20)6		−1.242				−32 146.678***
		(5.912)				(3687.503)
factor(cluster_20)7		−12.445				−16 488.001*
		(14.472)				(7211.100)
factor(cluster_20)8		11.951				−17 568.146***
		(7.535)				(3833.353)
factor(cluster_20)9		−0.229				−13 492.026***
		(6.335)				(3715.735)
factor(cluster_20)10		6.614				−22 006.001*
		(17.600)				(9730.316)
factor(cluster_20)11		8.597				−17 586.455*
		(12.332)				(7173.398)
factor(cluster_20)12		4.295				−10 443.257+
		(10.039)				(6085.553)
factor(cluster_20)13		17.444				−18 992.354+
		(15.931)				(9726.507)
factor(cluster_20)14		55.659***				52 337.663***
		(10.991)				(7109.571)
factor(cluster_20)15		−6.566				−1103.501
		(13.036)				(7175.195)
factor(cluster_20)16		−13.802				−11 116.937

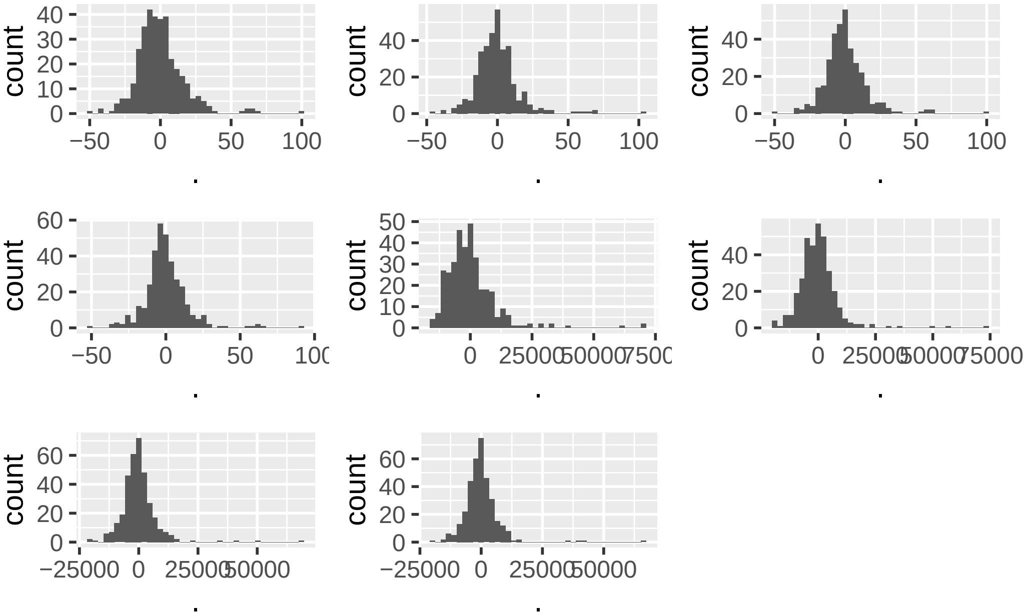
Residual Check

Linear Regression

```
residual.list.lm <- map(
  .x = seq_along(mlist),
  .f = ~ mlist[[.x]]$residuals
)


plot.out <- map(
  .x = residual.list.lm,
  .f = ~ggplot(data = as.data.frame(.x), aes(x = .)) +
      geom_histogram(bins = 40)
)


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



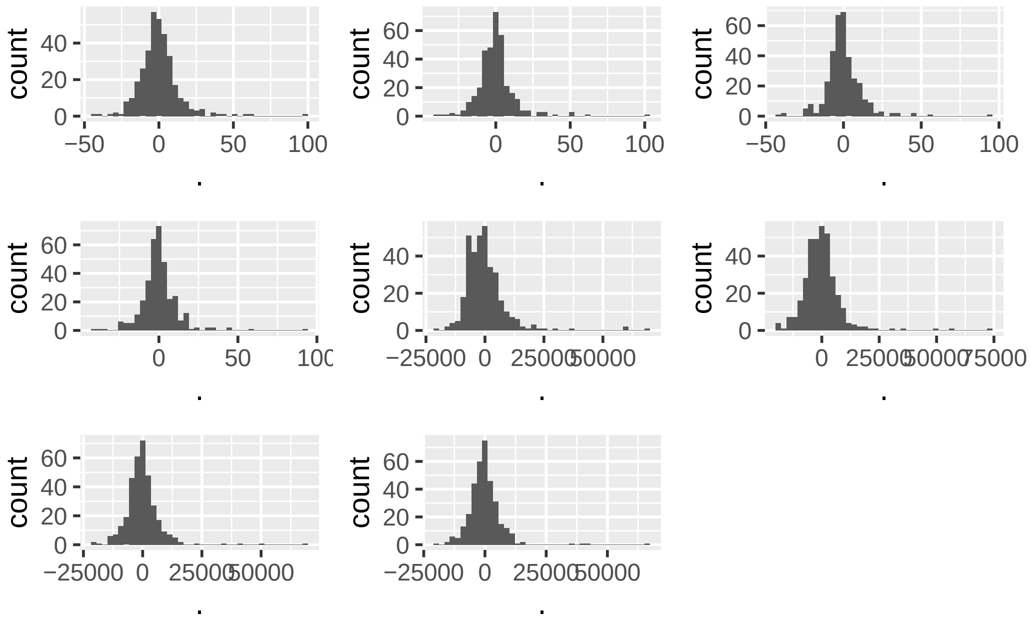
Multilevel Model

The multilevel models look similar but there's generally less dispersion in the residuals as compared to the simple linear model.

```
residual.list.mm <- map(
  .x = seq_along(mmlist),
  .f = ~ residuals(mmlist[[.x]])
)


plot.out <- map(
  .x = residual.list.mm,
  .f = ~ggplot(data = as.data.frame(.x), aes(x = .)) +
      geom_histogram(bins = 40)
)


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


count
count
count
count
10C
count
count
count
count