Assignment 5

James Dill

2025-11-24

- Link to GitHub Repo: https://github.com/jcdill500/SURV_727

```
library(censusapi)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(mapproj)
```

```
## Loading required package: maps
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(maps)
#readRenviron("~/.Renviron")
#usethis::edit_r_environ()
```

```r
acs_il_c<-getCensus(
  name="acs/acs5",
  vintage=2016,
  vars=c("NAME",
         "B01003_001E",
         "B19013_001E",
         "B19301_001E"),
  region="county:*",
  regionin="state:17",
  key = Sys.getenv("CENSUS_API_KEY")
)|>
  dplyr::rename(
    pop=B01003_001E,
    hh_income=B19013_001E,
    income=B19301_001E
  )

head(acs_il_c)
```

```
##   state county                      NAME    pop hh_income income
## 1    17    067   Hancock County, Illinois  18633     50077  25647
## 2    17    063    Grundy County, Illinois  50338     67162  30232
## 3    17    091  Kankakee County, Illinois 111493     54697  25111
## 4    17    043    DuPage County, Illinois 930514     81521  40547
## 5    17    003 Alexander County, Illinois   7051     29071  16067
## 6    17    129    Menard County, Illinois  12576     60420  31323
```

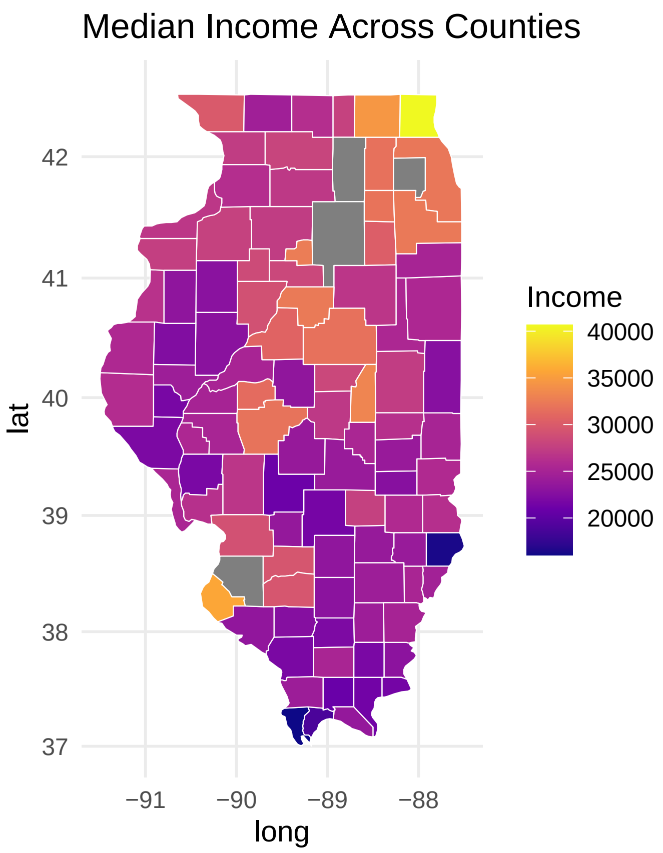
Census Map

```r
il_map<-map_data("county", region = "illinois")
head(il_map)
```

```
##        long      lat group order   region subregion
## 1 -91.49563 40.21018     1     1 illinois     adams
## 2 -90.91121 40.19299     1     2 illinois     adams
## 3 -90.91121 40.19299     1     3 illinois     adams
## 4 -90.91121 40.10704     1     4 illinois     adams
## 5 -90.91121 39.83775     1     5 illinois     adams
## 6 -90.91694 39.75754     1     6 illinois     adams
```

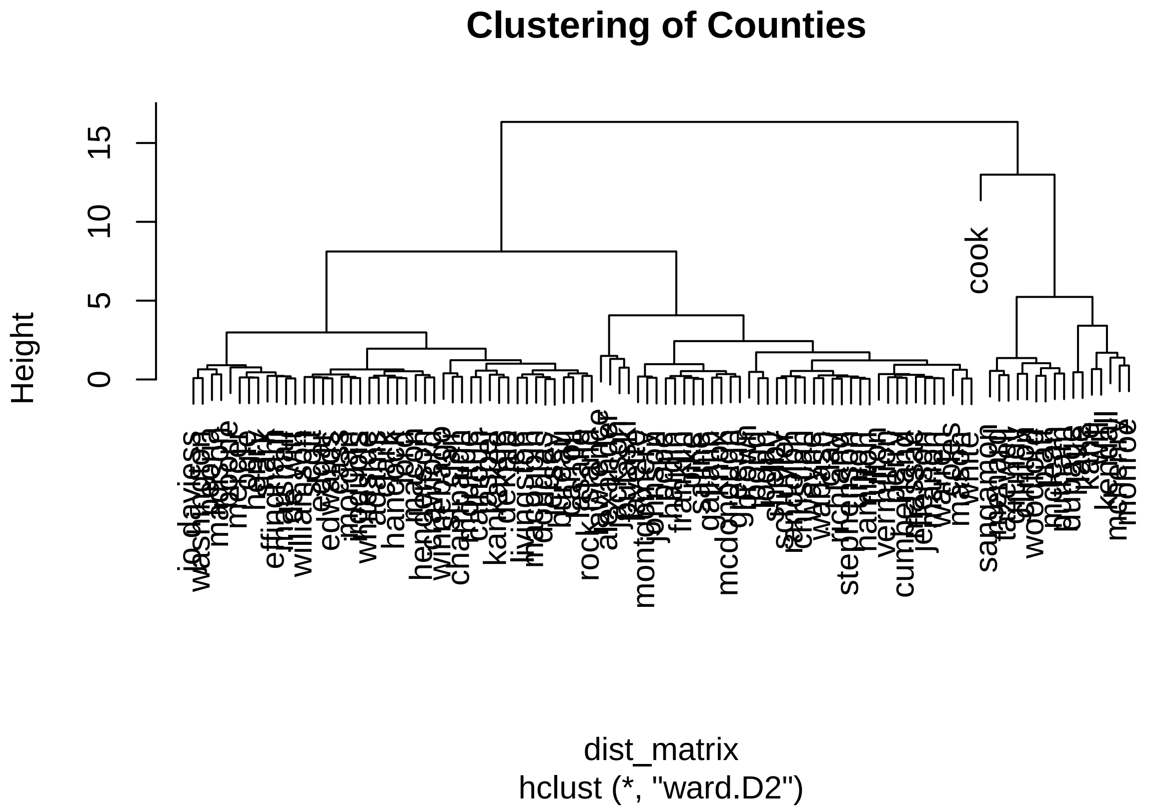
```r
#join
acs_il_c <- acs_il_c|>
  mutate(
    county_name=sub(" County,.*", "", NAME),
    county_name=tolower(county_name)
  )

acs_map <- il_map|>
  left_join(acs_il_c, by=c("subregion"="county_name"))

ggplot(acs_map)+
  geom_polygon(aes(
    x=long,
    y=lat,
    group = group,
    fill = income
  ), color = "white", size=0.2)+
  coord_map()+
  scale_fill_viridis_c(option="plasma")+
  labs(
    title="Median Income Across Counties",
    fill="Income"
  )+
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

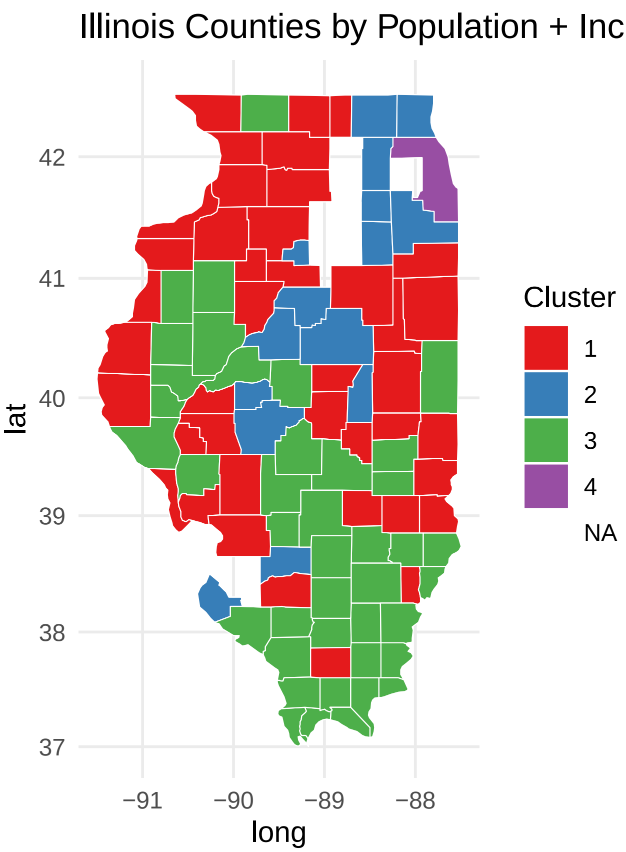


Hierarchical Clustering

```
#clean
acs_clean<-acs_il_c|>
  dplyr::mutate(
    county_name=sub(" County,.*", "", NAME),
    county_name=tolower(county_name)
  )|>
  dplyr::select(county_name, pop, hh_income, income)

acs_scaled <- acs_clean|>
  dplyr::select(pop, hh_income, income)|>
  scale()|>
  as.data.frame()

dist_matrix<-dist(acs_scaled, method="euclidean")

#wards method
hc <- hclust(dist_matrix, method="ward.D2")

plot(hc, labels=acs_clean$county_name, main= " Clustering of Counties")
```



```
acs_clean$cluster<-cutree(hc, k=4)

il_map<-map_data("county", region="illinois")
acs_map<-il_map|>
  dplyr::left_join(acs_clean, by= c("subregion"="county_name"))

ggplot(acs_map) +
  geom_polygon(
    aes(x=long, y=lat, group=group, fill= factor(cluster)),
    color="white", size=0.2
  )+
  coord_map()+
  scale_fill_brewer(palette="Set1", name="Cluster") +
  labs(
    title="Illinois Counties by Population + Income") +theme_minimal()
```

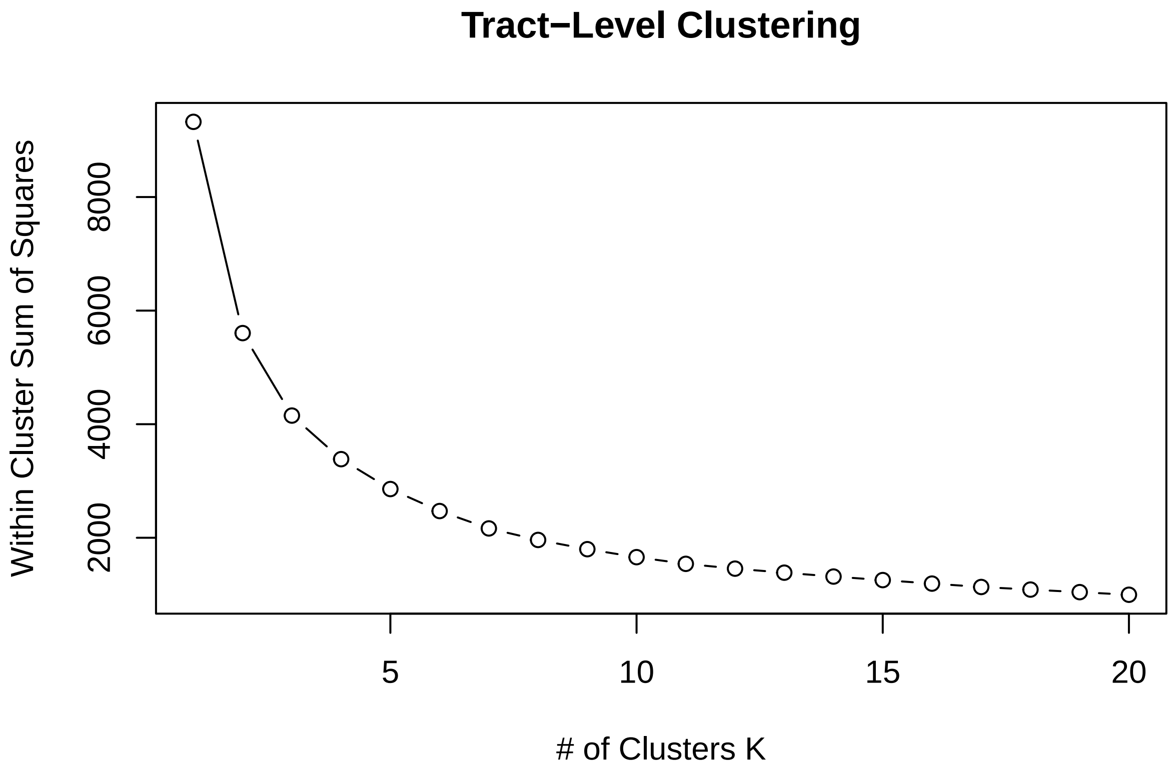


k-means

```
acs_il_t <- getCensus(
  name = "acs/acs5",
  vintage = 2016,
  vars = c("NAME",
           "B01003_001E",
           "B19013_001E",
           "B19301_001E"),
  region = "tract:*",
  regionin = "state:17",
  key = Sys.getenv("CENSUS_API_KEY")
) |>
  dplyr::mutate(across(everything(), ~ ifelse(. == -666666666, NA, .))) |>
  dplyr::rename(
    pop = B01003_001E,
    hh_income = B19013_001E,
    income = B19301_001E
  )

head(acs_il_t)
```

```
##   state county  tract                                        NAME  pop
## 1    17    031 806002 Census Tract 8060.02, Cook County, Illinois 7304
## 2    17    031 806003 Census Tract 8060.03, Cook County, Illinois 7577
## 3    17    031 806400    Census Tract 8064, Cook County, Illinois 2684
## 4    17    031 806501 Census Tract 8065.01, Cook County, Illinois 2590
## 5    17    031 750600    Census Tract 7506, Cook County, Illinois 3594
## 6    17    031 310200    Census Tract 3102, Cook County, Illinois 1521
##   hh_income income
## 1     56975  23750
## 2     53769  25016
## 3     62750  30154
## 4     53583  20282
## 5     40125  18347
## 6     63250  31403
```

```
#clean
acs_t_clean <- acs_il_t |>
  dplyr::mutate(
    GEOID = paste0(state, county, tract),
    county = sub(".*, ", "", NAME),
    county = sub(" County.*", "", county),
    pop = as.numeric(pop),
    hh_income = as.numeric(hh_income),
    income = as.numeric(income)
  ) |>
  dplyr::select(GEOID, county, pop, hh_income, income)

acs_t_scaled_clean <- acs_t_clean |>
  dplyr::select(pop, hh_income, income) |>
  na.omit() |>
  scale() |>
  as.data.frame()

acs_t_clean_used <- acs_t_clean[complete.cases(acs_t_clean[, c("pop", "hh_income", "income")]), ]

#elbow method
wss <- numeric(20)
for (k in 1:20) {
  set.seed(123)
  wss[k] <- kmeans(acs_t_scaled_clean, centers = k, nstart = 25, iter.max = 100)$tot.withinss
}

plot(1:20, wss, type="b",
     xlab="# of Clusters K",
     ylab="Within Cluster Sum of Squares",
     main="Tract-Level Clustering")
```



```
#k means

set.seed(123)
km5 <- kmeans(acs_t_scaled_clean, centers = 5, nstart = 50, iter.max = 100)
acs_t_clean_used$cluster_k5 <- km5$cluster

#clusters summary

#mean values
cluster_summary<-acs_t_clean_used|>
  dplyr::group_by(cluster_k5)|>
  dplyr::summarise(
    mean_pop=mean(pop, na.rm = T),
    mean_hh_income=mean(hh_income, na.rm = T),
    mean_income=mean(income, na.rm = T)
  )
cluster_summary
```

```
## # A tibble: 5 x 4
##   cluster_k5 mean_pop mean_hh_income mean_income
##        <int>    <dbl>          <dbl>       <dbl>
## 1          1    3896.        122368.      67665.
## 2          2    2686.         37123.      19778.
## 3          3    7838.         86010.      38154.
## 4          4    5381.         49260.      23275.
## 5          5    3610.         73195.      35913.
```

```r
#most frequent per cluster
cluster_county <- acs_t_clean_used|>
  dplyr::group_by(cluster_k5, county)|>
  dplyr::summarise(n = n(), .groups = "drop")|>
  dplyr::slice_max(n, n = 1)
cluster_county
```

```
## # A tibble: 1 x 3
##   cluster_k5 county       n
##        <int> <chr>    <int>
## 1          2 Illinois  1016
```

```r
#automated k means

run_kmeans <- function(K) {
  set.seed(123)
  km <- kmeans(acs_t_scaled_clean, centers = K, nstart = 25, iter.max = 100)
  return(km$cluster)
}

for (k in 2:10) {
  colname <- paste0("cluster_k", k)
  acs_t_clean_used[[colname]] <- run_kmeans(k)
}


head(acs_t_clean_used)
```

```
##         GEOID   county  pop hh_income income cluster_k5 cluster_k2 cluster_k3
## 1 17031806002 Illinois 7304     56975  23750          4          2          3
## 2 17031806003 Illinois 7577     53769  25016          4          2          3
## 3 17031806400 Illinois 2684     62750  30154          5          2          2
## 4 17031806501 Illinois 2590     53583  20282          2          2          2
## 5 17031750600 Illinois 3594     40125  18347          2          2          2
## 6 17031310200 Illinois 1521     63250  31403          5          2          2
##   cluster_k4 cluster_k6 cluster_k7 cluster_k8 cluster_k9 cluster_k10
## 1          1          3          6          8          7           8
## 2          1          3          6          8          7           8
## 3          3          5          3          3          6           7
## 4          2          5          2          3          6           7
## 5          2          4          2          1          4          10
## 6          3          5          3          3          6           7
```