Assignment 5

Due at 11:59pm on November 26.

You may work in pairs or individually for this assignment. Make sure you join a group in Canvas if you are working in pairs. Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it. Include the GitHub link for the repository containing these files.

```
library(censusapi)
library(tidyverse)
library(magrittr)
library(factoextra)
library(dplyr)
library(stringr)
library(ggmap)
library(ggplot2)
library(lubridate)
library(corrplot)
library(RSocrata)
```

Exploring ACS Data

In this notebook, we use the Census API to gather data from the American Community Survey (ACS). This requires an access key, which can be obtained here:

https://api.census.gov/data/key_signup.html

```
cs_key <- read_file("census-key.txt")
```

```
acs_il_c <- getCensus(name = "acs/acs5",
                      vintage = 2016,
                      vars = c("NAME", "B01003_001E", "B19013_001E", "B19301_001E"),
                      region = "county:*",
```

```
                    regionin = "state:17",
                    key = cs_key) %>%
            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
```

Pull map data for Illinois into a data frame.

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

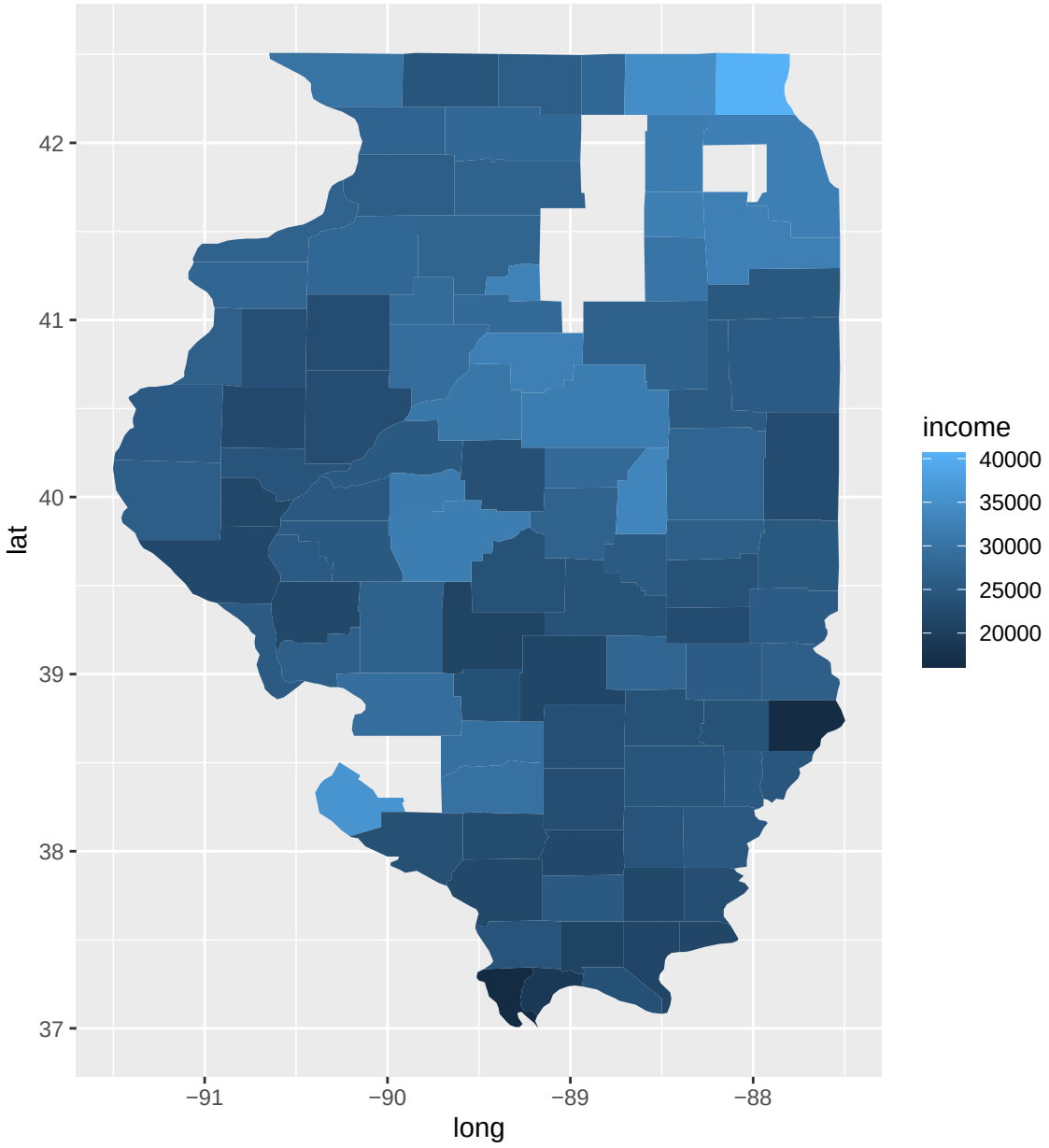
```
       long      lat group order   region subregion
1 -91.49563 40.21018     1     1 illinois     adams
2 -90.91121 40.19299     1     2 illinois     adams
3 -90.91121 40.19299     1     3 illinois     adams
4 -90.91121 40.10704     1     4 illinois     adams
5 -90.91121 39.83775     1     5 illinois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

Join the ACS data with the map data. Note that `il_map` has a column `subregion` which includes county names. We need a corresponding variable in the ACS data to join both data sets. This needs some transformations, among which the function `tolower()` might be useful. Call the joined data `acs_map`.

```
# creating new variable 'subregion' in acs_il_c to match the 'subregion' variable in il_map
acs_il_c <- acs_il_c %>%
  mutate(subregion = tolower(str_remove(NAME, " County, Illinois")))

# joining the ACS data with the map data
acs_map <- inner_join(acs_il_c, il_map, by = "subregion")
```

After you do this, plot a map of Illinois with Counties colored by per capita income.

```
ggplot(acs_map) +
geom_polygon(aes(x = long, y = lat, group = group, fill = income))
```



Hierarchical Clustering

We want to find clusters of counties that are similar in their population, average household income and per capita income. First, clean the data so that you have the appropriate variables to use for clustering. Next, create the distance matrix of the cleaned data. This distance matrix can be used to cluster counties, e.g. using the ward method.

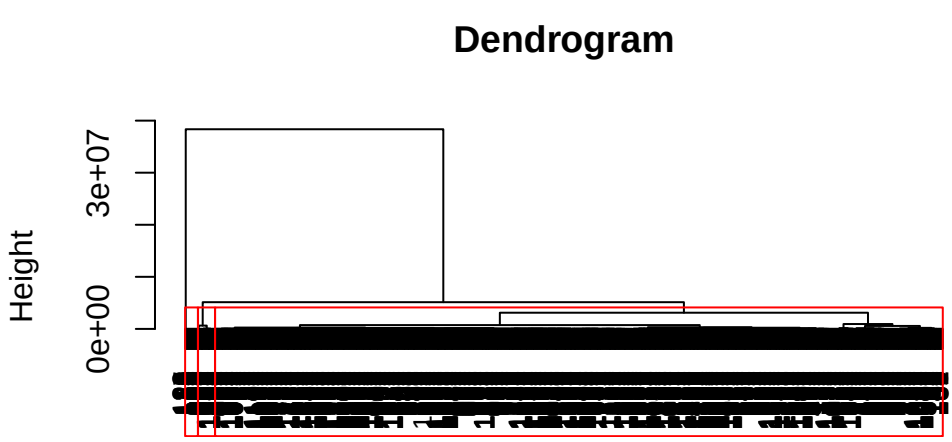
```
# cleaning the data so that I have the appropriate variables to use for clustering
acs_map_clean <-
  acs_map %>%
  select(pop, hh_income,income)
```

```
# scaling the data
scaled_data <- scale(acs_map_clean)

# creating the distance matrix of the cleaned data
dist_matrix <- dist(acs_map_clean)
```

```
# clustering counties using different methods
hc_ward <- hclust(dist_matrix, method = "ward.D2")
```

Plot the dendrogram to find a reasonable number of clusters. Draw boxes around the clusters of your cluster solution.

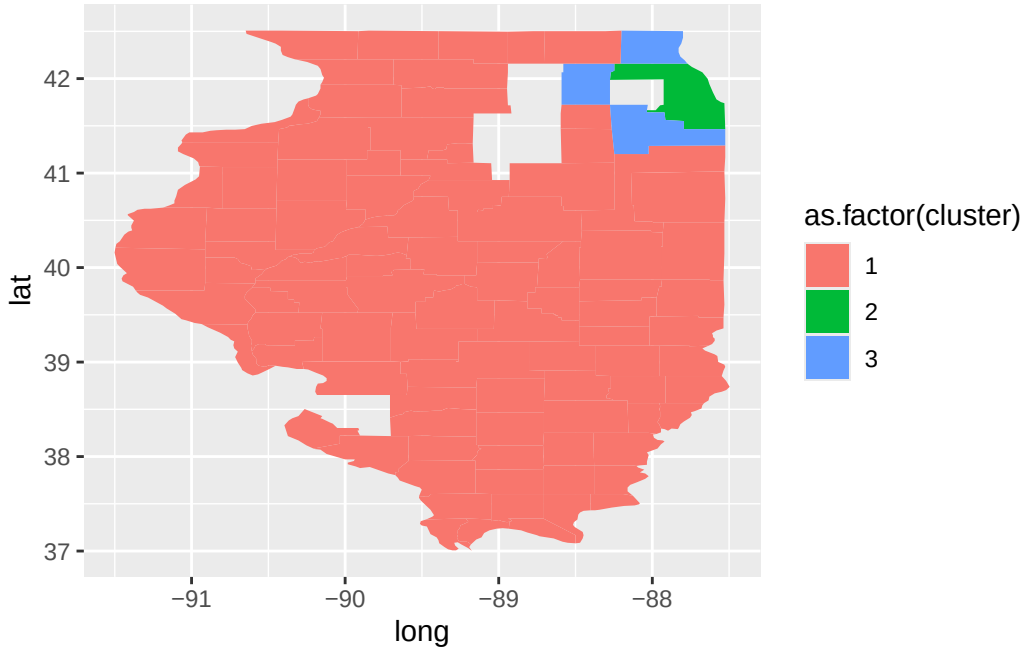
```
# plotted the dendrogram and found 3 clusters (in red)
plot(hc_ward, main = "Dendrogram", xlab = "", sub = "", cex = 0.8)

rect.hclust(hc_ward,
            k = 3,
            border = "red")
```



Visualize the county clusters on a map. For this task, create a new `acs_map` object that now also includes cluster membership as a new column. This column should be called `cluster`.

```
clusters <- cutree(hc_ward, k = 3)

# creating a new acs_map object that now also includes cluster membership as a new column 'cl
acs_map <- acs_map %>%
  mutate(cluster = clusters) %>%
  select(pop, hh_income,income, cluster, long, lat, group)
```

```
# visualizing the county clusters on a map
#| fig.height = 6.5, fig.width = 6

 ggplot(acs_map) +
 geom_polygon(aes(x = long, y = lat, group = group, fill = as.factor(cluster)))
```



Census Tracts

For the next section we need ACS data on a census tract level. We use the same variables as before.

```
acs_il_t <- getCensus(name = "acs/acs5",
                      vintage = 2016,
                      vars = c("NAME", "B01003_001E", "B19013_001E", "B19301_001E"),
                      region = "tract:*",
                      regionin = "state:17",
                      key = cs_key) %>%
             mutate_all(funs(ifelse(.==-666666666, NA, .))) %>%
             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
```

k-Means

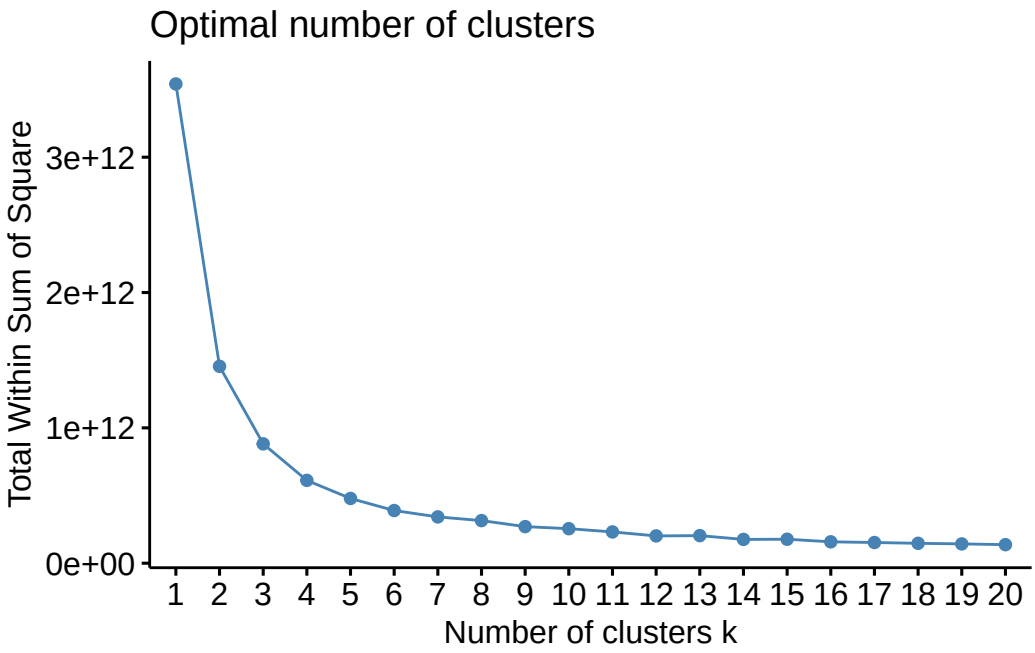
As before, clean our data for clustering census tracts based on population, average household income and per capita income.

```
# for later on (displaying the most frequent county that can be observed within each cluster)
acs_il_t_clean2 <-
  acs_il_t %>%
  select(pop, hh_income,income, NAME) %>%
  na.omit()
```

```
# cleaning the data so that I have the appropriate variables to use for clustering
acs_il_t_clean <-
  acs_il_t %>%
  select(pop, hh_income,income) %>%
  na.omit()
```

Since we want to use K Means in this section, we start by determining the optimal number of K that results in Clusters with low within but high between variation. Plot within cluster sums of squares for a range of K (e.g. up to 20).

```
fviz_nbclust(acs_il_t_clean, #data set we want to use
             kmeans, #cluster method
             method = "wss", #method used for estimating the optimal number of clusters
             k.max = 20)
```



Run `kmeans()` for the optimal number of clusters based on the plot above.

```
km <- kmeans(acs_il_t_clean, 3, nstart = 20)
print(km$size)
```

```
[1] 1553  337 1219
```

```
print(km$centers)
```

```
       pop hh_income   income
1 3665.401  39382.67 20507.19
2 4437.356 122378.74 62174.49
3 4637.467  72015.19 34810.54
```

Find the mean population, household income and per capita income grouped by clusters. In addition, display the most frequent county that can be observed within each cluster.

```
# adding cluster information to acs_il_t_clean
acs_il_t_clean$cluster <- km$cluster

# finding the mean of population, household income, and per capita income by cluster
```

```r
cluster_means <- acs_il_t_clean %>%
  group_by(cluster) %>%
  summarise(
    mean_pop = mean(pop, na.rm = TRUE),
    mean_hh_income = mean(hh_income, na.rm = TRUE),
    mean__income = mean(income, na.rm = TRUE)
  )
cluster_means
```

```
# A tibble: 3 x 4
  cluster mean_pop mean_hh_income mean__income
    <int>    <dbl>          <dbl>        <dbl>
1       1    3665.         39383.       20507.
2       2    4437.        122379.       62174.
3       3    4637.         72015.       34811.
```

```r
# adding cluster information to acs_il_t_clean2
acs_il_t_clean2$cluster <- km$cluster

# finding the most frequent county within each cluster
most_frequent_county <- acs_il_t_clean2 %>%
  group_by(cluster) %>%
  summarise(most_frequent_county = names(sort(table(NAME), decreasing = TRUE))[1])
most_frequent_county
```

```
# A tibble: 3 x 2
  cluster most_frequent_county
    <int> <chr>
1       1 Census Tract 1, Coles County, Illinois
2       2 Census Tract 12.04, Champaign County, Illinois
3       3 Census Tract 1, Adams County, Illinois
```

As you might have seen earlier, it's not always clear which number of clusters is the optimal choice. To automate K Means clustering, program a function based on `kmeans()` that takes K as an argument. You can fix the other arguments, e.g. such that a specific dataset is always used when calling the function.

```r
automate_km <- function(K, data) {
  km_result <- kmeans(data, centers = K, nstart = 20)
  return(km_result$cluster)
}
```

We want to utilize this function to iterate over multiple Ks (e.g., K = 2, ..., 10) and – each time – add the resulting cluster membership as a new variable to our (cleaned) original data frame (`acs_il_t`). There are multiple solutions for this task, e.g. think about the `apply` family or `for` loops.

```
# removing the 'cluster' column from the previous questions
acs_il_t_clean <- subset(acs_il_t_clean, select = -cluster)
```

```
for (K in 2:10) {
  cluster_numbers <- automate_km(K, acs_il_t_clean)
  acs_il_t_clean[[paste("cluster_K", K, sep = "_")]] <- cluster_numbers
}
```

Finally, display the first rows of the updated data set (with multiple cluster columns).

```
head(acs_il_t_clean)
```

```
   pop hh_income income cluster_K_2 cluster_K_3 cluster_K_4 cluster_K_5
1 7304     56975  23750           2           1           1           3
2 7577     53769  25016           2           1           1           3
3 2684     62750  30154           2           3           1           3
4 2590     53583  20282           2           1           1           3
5 3594     40125  18347           2           1           4           4
6 1521     63250  31403           2           3           1           3
  cluster_K_6 cluster_K_7 cluster_K_8 cluster_K_9 cluster_K_10
1           2           4           3           5            4
2           4           5           3           2            2
3           2           4           3           5            4
4           4           5           3           2            2
5           4           5           7           8           10
6           2           4           3           5            4
```