Assignment 5

Due at 11:59pm on December 5th

Akari Oya, Zhouer Wang

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.

Github link: https://github.com/ZuorW/SURV727.git

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

```
acs_il_c <- getCensus(name = "acs/acs5",
                    vintage = 2016,
                    vars = c("NAME", "B01003_001E", "B19013_001E", "B19301_001E"),
                    region = "county:*",
                    regionin = "state:17",
                    key = 'e3ef32e30f690ff26b243cea3315af9cc1ce6ede') %>%
            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. Not 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`.

```
# Manipulating to match datasets for merging
# In ACS data, rename NAME to subregion
acs_il_c <- acs_il_c %>%
  rename(subregion = NAME)

# Take out unnecessary part of subregion variable
acs_il_c$subregion <-
  gsub(' County, Illinois', '', acs_il_c$subregion)

# Take out spaces in subregion variable to match data sets
acs_il_c$subregion <-
  gsub(" ", "", acs_il_c$subregion) %>%
  tolower() #make values lowercase

head(acs_il_c)
```

```
  state county subregion    pop hh_income income
1    17    067   hancock  18633     50077  25647
2    17    063    grundy  50338     67162  30232
```

```
3     17     091  kankakee 111493     54697  25111
4     17     043    dupage 930514     81521  40547
5     17     003 alexander   7051     29071  16067
6     17     129    menard  12576     60420  31323
```

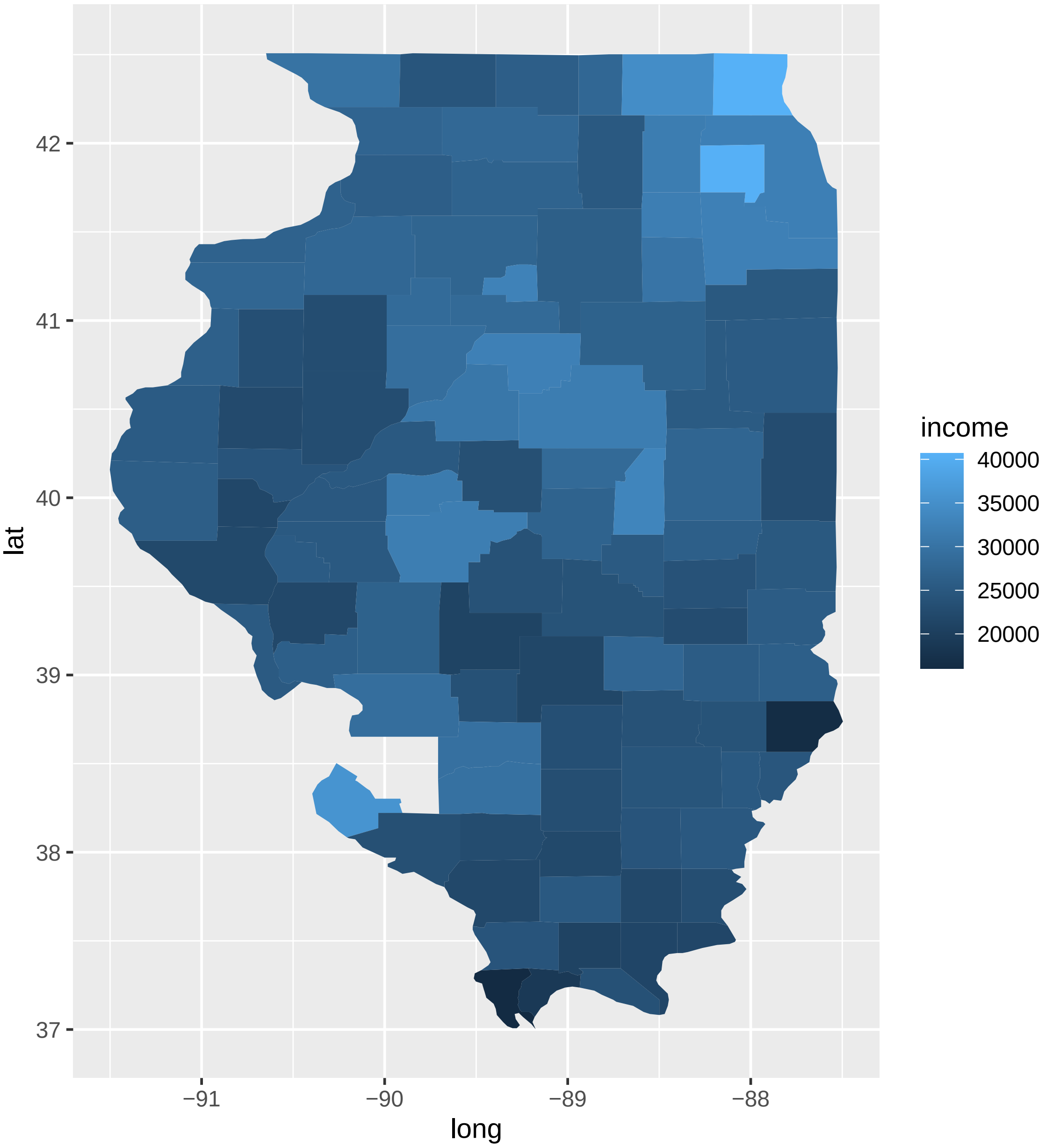
```
# In map data, remove spaces in subregion column
il_map$subregion <- gsub(" ", "", il_map$subregion)
```

```
# Join ACS and map data
acs_map <-
  acs_il_c %>%
  inner_join(il_map, by = "subregion")
head(acs_map)
```

```
  state county subregion   pop hh_income income      long      lat group order
1    17    067   hancock 18633     50077  25647 -91.18623 40.63417    34   573
2    17    067   hancock 18633     50077  25647 -90.89976 40.63417    34   574
3    17    067   hancock 18633     50077  25647 -90.91121 40.27893    34   575
4    17    067   hancock 18633     50077  25647 -90.91121 40.19299    34   576
5    17    067   hancock 18633     50077  25647 -91.49563 40.21018    34   577
6    17    067   hancock 18633     50077  25647 -91.48990 40.25029    34   578
    region
1 illinois
2 illinois
3 illinois
4 illinois
5 illinois
6 illinois
```

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, fill = income, group = group))
```



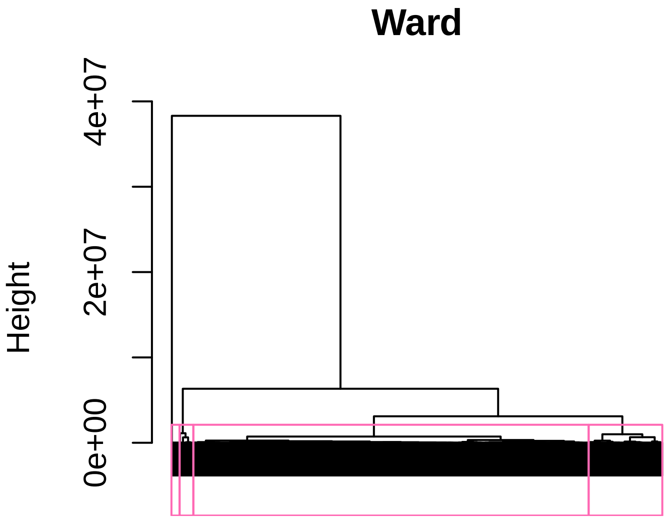
Hierarchical Clustering

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.

```
acs_map$pop <- as.numeric(acs_map$pop)
acs_map$hh_income <- as.numeric(acs_map$hh_income)
acs_map$income <- as.numeric(acs_map$income)

acs_map <- na.omit(acs_map)

hclust_data <- acs_map[, c("pop", "hh_income", "income")]

hclust_d <- dist(hclust_data)
hc_ward <- hclust(hclust_d, method = "ward.D2")
```

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

```
plot(hc_ward, main = "Ward", xlab = "", sub = "", labels = FALSE)
rect.hclust(hc_ward,
            k = 4,
            border = "hotpink")
```

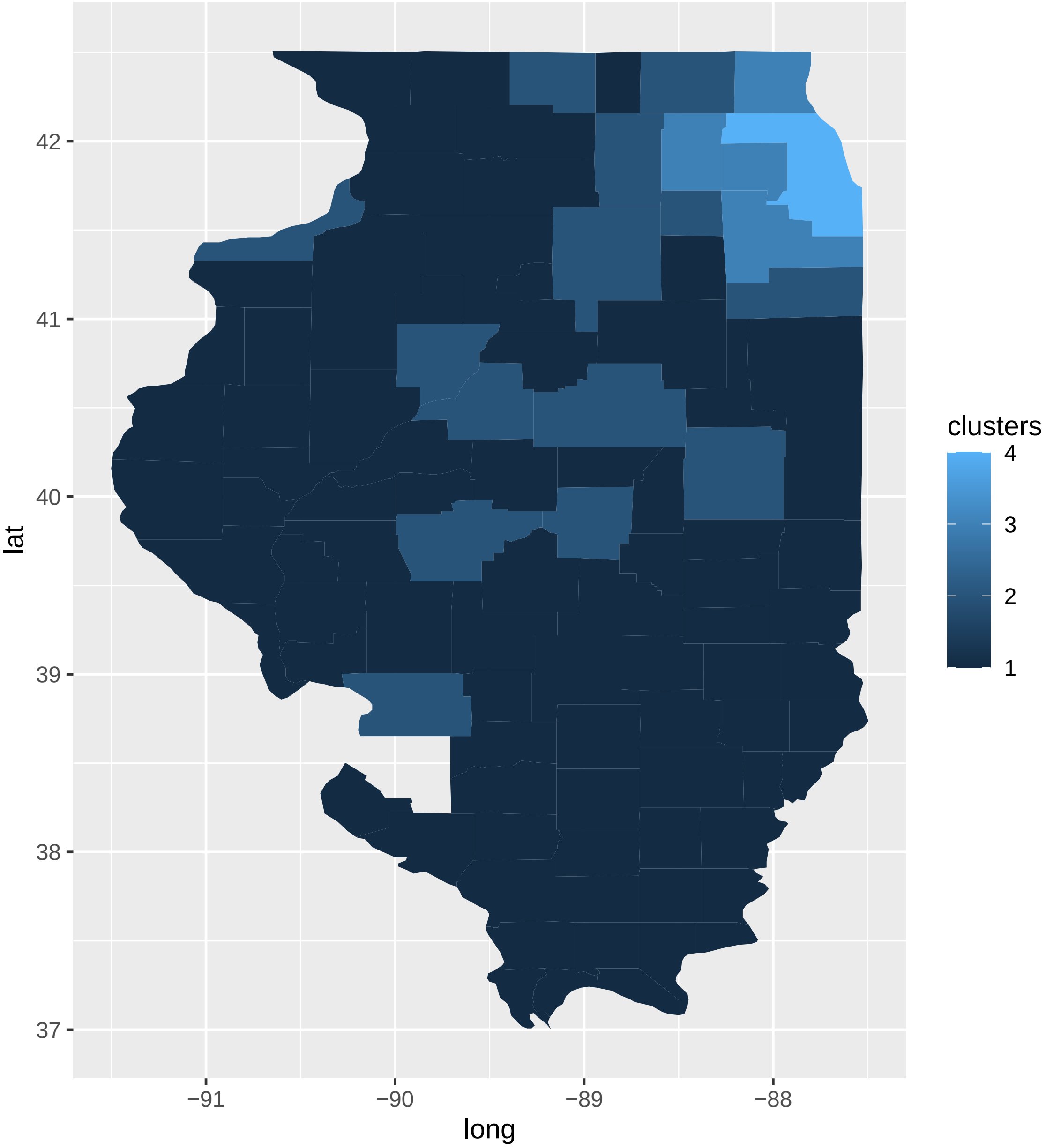


```
clusters <- cutree(hc_ward, 4)
```

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`.

```
a <- acs_map %>%
  mutate(cluster = cutree(hc_ward, 4))
```

```
ggplot(a) +
geom_polygon(aes(x = long, y = lat, fill = clusters, group = group))
```



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 = 'e3ef32e30f690ff26b243cea3315af9cc1ce6ede') %>%
            mutate_all(list(~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

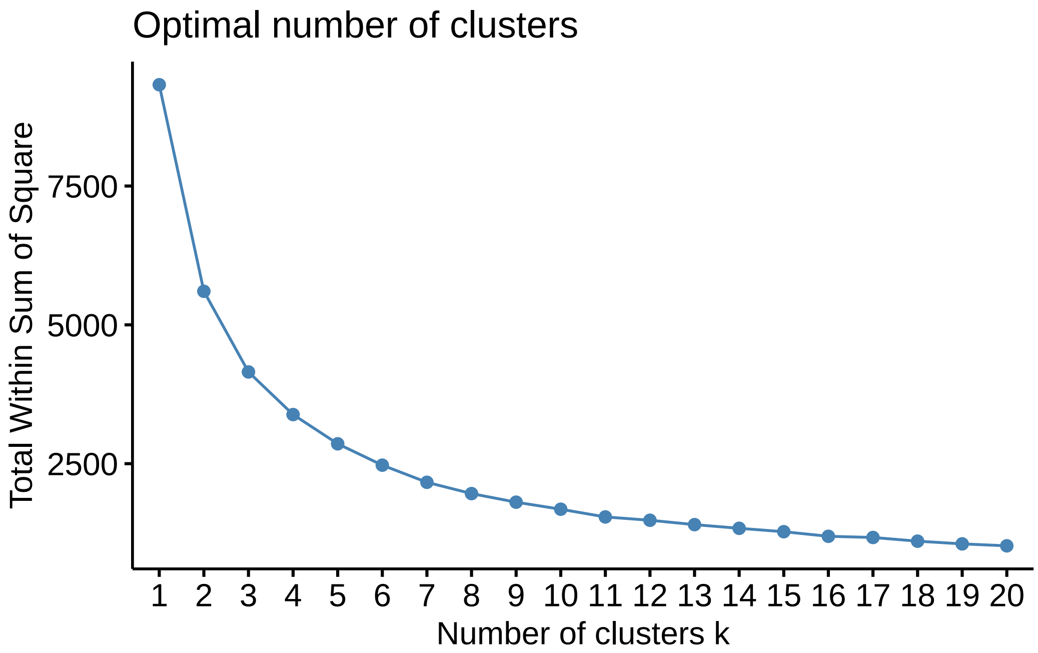
As before, clean our data for clustering census tracts based on population, average household income and per capita income.

```
acs_il_t$pop <- as.numeric(acs_il_t$pop)
acs_il_t$hh_income <- as.numeric(acs_il_t$hh_income)
acs_il_t$income <- as.numeric(acs_il_t$income)
```

```
# Remove missing data
acs_il_t <- na.omit(acs_il_t)
```

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).

```
hclust_data2 <-
  acs_il_t %>%
  select(pop, hh_income, income) %>%
  mutate_all(scale)

fviz_nbclust(hclust_data2,
             kmeans,
             method = "wss",
             k.max = 20)
```



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

```
# 8 looks like optimal cluster number due to the elbow criterion
kmean <- kmeans(hclust_data2, 8, nstart = 20)
```

```r
# Make cluster number in dataset to be 8
acs_il_t$cluster <- kmean$cluster
```

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.

```r
# Create county name variable with just name
acs_il_t <- acs_il_t %>%
  mutate(county_name = sapply(strsplit(NAME, ", "), `[`, 2))

# Create summary table
summary_cluster <- acs_il_t %>%
  group_by(cluster) %>%
  summarize(
    mean_pop = mean(pop),
    mean_hh_income = mean(hh_income),
    mean_income = mean(income),
    most_frequent_county = names(which.max(table(county_name)))
  )
summary_cluster
```

```
# A tibble: 8 x 5
  cluster mean_pop mean_hh_income mean_income most_frequent_county
    <int>    <dbl>          <dbl>       <dbl> <chr>
1       1    3120.          53255.      27136. Cook County
2       2    3947.          78524.      38004. Cook County
3       3    2484.          29696.      16257. Cook County
4       4    5518.          47532.      22220. Cook County
5       5   14738.          88459.      40134. Kane County
6       6    3892.         108732.      57768. Cook County
7       7    7098.          82426.      36679. Cook County
8       8    4056.         149179.      86846. Cook County
```

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
km_func <- function(K) {
  km <- kmeans(hclust_data2, centers = K, nstart = 20)
  return(km)
}
```

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.

```
for (K in 2:10) {
  km_result <- km_func(K)
  cluster_col_name <- paste("cluster ", K)
  acs_il_t[cluster_col_name] <- km_result$cluster
}
```

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

```
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 cluster county_name cluster  2 cluster  3 cluster  4
1     56975  23750       4 Cook County          2          3          4
2     53769  25016       4 Cook County          2          3          4
3     62750  30154       1 Cook County          2          2          3
4     53583  20282       1 Cook County          2          2          1
5     40125  18347       3 Cook County          2          2          1
6     63250  31403       1 Cook County          2          2          3
  cluster  5 cluster  6 cluster  7 cluster  8 cluster  9 cluster  10
1          3          1          4          1          9           9
2          3          1          4          1          9           9
3          5          4          3          3          6           5
4          4          4          1          3          6           5
5          4          2          1          6          4           8
6          5          4          3          3          6           5
```