Exercise 4: Using US Census Bureau Data

ECON 256 Data Analysis and Visualization

Objective

Learn to import Census Bureau data into R. Clean up the data and plot a relationship.

[Provide answers to comments in red with a comment (#) in your code]

1 Download County Level Data

- Go to Social Explorer (socialexplorer.com)
- Set up an account with your UH email address (if you haven't already)
- Go to Tables
- Go to American Community Survey (5-year Estimates) and click on "Begin Report" for the 2019-2023 version
- Pick **one state** in the US, and download **county level** data for that state. Include the following three variables: Total Population, Median Household Income (In 2023 Inflation Adjusted Dollars) and Highest Educational Attainment for Population 25 Years and Over
- Under the "Data Download" tab, download the data as a csv to a folder on your computer (your working directory). Also download the "Data Dictionary."

2 Setup Your R Workspace

Setup a R script in the normal way, by assigning a working directory, with the `setwd()` function, and initializing the TidyVerse with the `library()` function. Make sure the data you just downloaded is in your working directory folder.

3 Open the Data in R

Load your data into an object in R using the `read_csv()` function.

4 Clean Up the Data

Consult the data dictionary to determine what the variable names mean.

Take a look at the data set. There will be a lot of empty variables that you don't need.

Let's keep a couple variables to identify the county (Geo_FIPS,Geo_NAME) as well as all the ACS variables (Note the "FIPS" codes). The ACS variables all start with "SE". Select only the variables you need with a `select()` function similar to the following:

```
mydata2<-select(mydata,Geo_FIPS,Geo_NAME,starts_with("SE"))
```

note we can use the `starts_with()` function to make a list of all the variables that start with certain characters. (alternatively we could just list all the variables individually)

Rename your 10 ACS variables with intuitive names with the `rename()` function. The meaning of all of the ACS variables are given in the Data Dictionary that you downloaded (note the data dictionary omits the SE_ prefix).

For example:

```
mydata3<-rename(mydata2, "population" = SE_A00001_001)
```

You can add additional rename arguments to the same rename() function, eg:

```
mydata3<-rename(mydata2, "population" = SE_A00001_001,"over25" = SE_A12001_001)
```

Create one long rename function that gives all 10 of your ACS variables intuitive names.

5 Generate a Variable

Let's calculate the share of over 25 year olds in each county that have at least a bachelors degree. To do this, you will need to use the mutate() function and generate a new variable that is equal to the total number of people in these groups:

Bachelor's Degree

Master's Degree

Professional School Degree

Doctorate Degree

Divided by the number of people who are over 25.

6 Summary Stat

Use the `summary()` function. For the education variable you just created, what is the maximum value among your counties? what is the minimum value?

What county in your state has the highest share of people with at least a bachelor's degree?

7 Make a Plot

Create a scatter plot (`geom_point()`) using `ggplot()` that shows the relationship between Median Household Income and the share of over 25 year olds with at least a bachelor's degree.

8 Send me Your Code

Save your R script. Name it with your last name, followed by the exercise number.

Submit it on Laulima.