Exercise 5: Joins

ECON 256 Data Analysis and Visualization

Objective

Learn to combine data sets using the "join" functions in R.

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

1 Set up Your R Workspace

Set up a R script with the normal preface. eg:

```
setwd("___________")

library(tidyverse)
```

2 Load Data Sets

Lets explore who won the most medals at the Tokyo 2020 Summer Olympics.

Download the three data sets provided on Laulima to your working directory: medalcount2020.csv ; imf_gdp.csv ; UNpopulations.csv.

The first data set contains the number of medals won by each country. The second data set contains the national GDP for every country in the world, according to data from the International Monetary Fund (IMF) and the third data set is the population of every country (in millions) according to the UN.

Load each data set into a separate object in R using the `read_csv` function.

Once loaded, take a look at the variables in each data set.

3 Join Data Sets Together

Within TidyVerse there is a set of commands to join different data sets together. Type `?join` in the conosle and hit enter to view the help file. From viewing the three data sets, you may have noticed they all have a variable in common: "country." We will merge all three data sets together based on the variable "country."

First, lets merge together the Medals data set and the GDP data set.

The command will look something like this:

```
medalsgdp <- left_join(medals,gdp,by="country")
```

Where "medalsgdp" is the new object containing the combined data set, "medals" is the object that contains the medals data and "gdp" is the object that contains the GDP data. We use the `by` argument to tell R we want to join the objects by the variable "country." Take a look at your newly created object to see if the join went as you expected.

Now join the new data set (`medalsgdp`) to the population data set in a similar way to create a new object that contains all the variables from the three data sets, and one observation per country.

Now take a look at your final data set and see if all the variables are there.

Note that some observations have missing data (represented by an NA). This is because not all three input data sets contained the exact same selection of countries. We could create a data set that only contains observations with complete data. To do this, replace your `left_join()` functions with `inner_join()` functions. This should reduce the number of observations in your final data set.

Why does using `inner_join()` result in a data set with fewer observations as compared to `left_join()`?

4 Create a New Variable

The data set of medals won contains the number of gold, silver and bronze medals won. Using the `mutate()` function, create a new variable that equals the total number of medals won by each country (gold + silver + bronze).

Create an other new variable that is equal to the number of medals won per million population for each country.

ie: `total_medals_won / population_mils`

Create an other new variable that is equal to GDP per person for each country.

ie: `(gdp_billions * 1000)/population_mils`

5 Make a Plot

Write a ggplot() command to make one scatter plot (`geom_point()`) showing the relationship between total medals won per million population and GDP per person.

How many countries have a GDP per person below $10,000 AND won more than 3 medals per million people?

6 Send me Your Code

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

Submit it on Laulima.