ECON 5 / POLI 5D

Homework 3

Due Monday, May 20, at 11:59 PM

Problem Set 3 is due on 5/20 at 11:59 PM. Please submit your work through TritonEd. Use R to complete the following problems. Download the .R script from TritonEd and save it as `LastName_FirstName.R`. Write *all* of your code and answers in this .R script, under the appropriate heading. Your completed .R script is the *only* thing that you need to upload to TritonEd when you are finished.

There are many ways to do anything in R. Unless we specifically tell you to do something a particular way, you are free to choose whichever way makes the most sense to you.

This problem set will use the following datasets:

MexicanCongress.csv

This dataset contains information on the share of seats in Mexico's Chamber of Deputies (the lower house of Congress) controlled by different political parties between 1964 and 2018. This dataset includes only the parties that won at least one seat in the election.

Variable	Description
`Year`	Election year
`Party`	The initials of the political party
`Seat.Share`	The percentage of seats in the Chamber that this party controlled after this election

AfricaDemocracy.csv

This dataset contains information on the regime type (level of democracy) and colonial legacy of different African countries from the 1960s to the 1990s. Higher democracy scores indicate that the country's political system was more democratic. Missing values for the democracy score indicate that the country was not yet independent.

Variable	Description
`Country`	Country name
`CCode`	3-letter country code
`Democracy.1960s`	The country's average Polity IV score in the 1960s
`Democracy.1970s`	The country's average Polity IV score in the 1970s
`Democracy.1980s`	The country's average Polity IV score in the 1980s
`Democracy.1990s`	The country's average Polity IV score in the 1990s
`Colonizer`	The European colonial power that governed this country prior to its independence

Question 1: Data

(A) The datasets for Questions 1 through 3 are posted online at the following links:

https://raw.githubusercontent.com/cjsells/ECON-5-POLI-5D-HW3-Data/master/MexicanCongress.csv
https://raw.githubusercontent.com/cjsells/ECON-5-POLI-5D-HW3-Data/master/AfricaDemocracy.csv

Use the `read.csv()` function to read both datasets into memory. Do not download them first. Instead, just copy and paste the entire link into your .R script and wrap the link first in quotation marks (`""`) and then in the `read.csv()` function. You should be able to import the datasets directly from the web into R. Store the first dataset as an object called `Congress`, and store the second dataset as an object called `Africa`.

(B) Which of these datasets is in "wide" format? Which of these datasets is in "long" format?

(C) Using the `Africa` dataset, convert (coerce) the variable "Country" into a character (string) variable. Write code that displays the names of the first 10 countries in the dataset.

(D) Subset the `Congress` dataset to create a new dataframe that contains only the observations whose party is the PRI. Store this new dataframe as an object called `pri`.

(E) Add a new variable to the `pri` dataset called `PRI.Majority` that equals `TRUE` if the PRI held at least 50% of the seats in Congress after that election, and equals `FALSE` otherwise. In which year did the PRI lose its congressional majority?

Question 2: Subsetting

Use the `Africa` dataset and subsetting to answer the following questions. In your .R script, write the *code* that produces the correct answer as its output. You do *not* need to copy and paste the output into the .R script. See the .R script for an example.

(A) Which countries in the dataset were former British colonies?
(*Your output should be a character vector of length 16*)

(B) Which countries had a democracy score of -9 in the 1970s?
(*Your output should be a character vector of length 4*)

(C) Which former French colonies had a positive (greater than 0) democracy score in the 1990s?
(*Your output should be a character vector of length 5*)

(D) Which country had the highest democracy score in the 1980s?
(*Your output should be a character vector of length 1*)

(E) Which countries were *more* democratic in the 1990s compared to the 1980s?
(*Your output should be a character vector of length 36; do not include countries that were equally democratic in both decades*)

(F) Which countries had a democracy score less than -5 in the 1980s *and* a democracy score greater than +5 in the 1990s?
(*Your output should be a character vector of length 2*)

(G) What was the average democracy score among former French colonies in the 1990s?
(*Your output should be a numeric vector of length 1 (i.e., a number)*)

Question 3: For Loops

For this question, you are going to use a for loop to iterate through each election year between 1964 and 2018 and count up the number of parties that held at least one seat in the Mexican Chamber of Deputies after each election. We have already set up most of the loop for you in the .R script. Your job is to add some code that will count the number of parties that won seats in year `y` and store that number as an object called `n`. Some functions that you might use to count the number of parties are `length()`, `nrow()`, `dim()`, and `sum()`. When you are done, run the entire loop all at once in order to verify that it produces the output that you expected. The first eight lines of output should look like this:

```
[1] "There were 4 parties in Congress in 1964"
[1] "There were 4 parties in Congress in 1967"
[1] "There were 4 parties in Congress in 1970"
[1] "There were 4 parties in Congress in 1973"
[1] "There were 4 parties in Congress in 1976"
[1] "There were 7 parties in Congress in 1979"
[1] "There were 6 parties in Congress in 1982"
[1] "There were 9 parties in Congress in 1985"
```

Question 4: Variables for the Final Project

Using R, begin analyzing the Final Project dataset that you chose in the last homework, and answer the following questions:

(A) What is the unit of analysis in your dataset?

(B) How many observations are in your dataset? How many variables?

(C) Restate your research question (if your question has not changed, you can just copy and paste what you wrote last time). What is your hypothesis?

(D) What is your dependent variable (outcome variable; the outcome of interest that you are trying to explain) and what is your main independent variable (explanatory variable; the variable that you think *explains* variation in your dependent variable)? What *types* of variables are they?

(E) Write code that produces summary statistics for *each* of the variables that you intend to use in your analysis (including control variables). If the variable is numeric, calculate the variable's mean, standard deviation, median, min, max. If the variable is categorical, calculate the variable's frequency distribution

A Note on the Final Project Datasets

In order to answer Question 4 of this problem set, your data will need to be in "analysis-ready" format, meaning that you have already done all of the data cleaning, merging, aggregation, transformation, subsetting, variable creation, and variable recoding necessary in order to use your dataset to answer your research question. Depending on your research question and the dataset that you chose, your dataset may already be analysis-ready. If it is not, you will have to use the data-wrangling tools that we covered earlier in this course to get your dataset in the format that you intend to use for the analysis. Since we have not covered data-wrangling in R yet, you **may** do this part in STATA this time. However, we expect you to use R to complete Question 4 of this problem set. We also expect you to use R to perform all analyses for the Final Project.

If your dataset is stored as a STATA file (.dta) or an Excel spreadsheet (.xlsx), the simplest way to load the dataset into R is by using RStudio's "Import Dataset" menu:

1. In RStudio, click on the "Import Dataset" button in your Environment window (top-right panel)
2. Select "From Stata" or "From Excel"
3. If RStudio asks whether it can install the required package, click "Yes" and wait for the package to finish installing (you will need an Internet connection for this part)
4. Click the "Browse" button
5. Find the Stata file or Excel spreadsheet saved on your computer
6. Look at the Data Preview window and verify that the dataset is formatted correctly
7. Click "Import"
8. The dataset should now be stored as a dataframe object. If you look in your console window, you should also see the code that RStudio used to import the dataset. If you copy and paste this code into your .R script, you can just run this code next time.

Another option is to first load the dataset in Stata/Excel, export it as a comma-separated values file (.csv), and then use `read.csv()` to load it into R. Something is more likely to go wrong this way, so use this solution only as a last resort.