IPI Student assistant tasks

Instructions

- Please complete these tasks within *one hour* of receipt of our email.
- Ideal answers should include reproducible script that goes from the data import to the final output, or as many steps as possible so that someone else may recreate your answers. You are encouraged to include comments to your code if submitting work in R or Stata.
- We strongly encourage you to use `.Rmd` or quorto if you are comfortable with these tools and in any case to include your answers to all questions in a *single* document (this can be .html, .pdf, .doc, .R, .do, .xlsx, etc).

1 Data management

You've been given 2 datasets.

- Dataset *A* ("A_CSES.csv") is a short version of the Comparative Study of Electoral Systems (CSES) survey wave 2. Each observation in this dataset refers to an individual in a certain year in a specific country.
- Dataset *B* ("B_QoG.xlsx") is a short version of the Quality of Government (QoG) data. Each observation in this dataset refers to a certain year in a specific country.

Dataset *A* includes the following variables:

Variable name	Variable label
B1004	Country-year code
B1006_NAM	Country name
B1008	Year
B2001	Age (number of years)
B2002	Gender (1=male; 2=female)
B2005	Union membership (1=is a member; 2=is not a member)
B2020	Income in household (1=low, 5=high)
B3014	It matters who people vote for (1 = it doesn't matter; 5 = it matters a lot)

Dataset *B* includes the following variables:

Variable name	Variable label
cname	Country name
year	Year of measurement
gle_cgdpc	GDP per capita (in current prices)
p_polity2	Polity score (measure of democracy)
undp_hdi	UNDP's Human Development Index

1.1 Read data

Read in both datasets into R using the needed functions for their format. Note that dataset A is a .csv (Comma-Separated Value) file, while dataset B is an .xlsx (Excel) file.

1.2 Clean data

Compare the list of unique countries that are present in datasets A and B. There are 5 countries that are named differently between these 2 datasets. Find these countries and then rename those in dataset B so that they match the names used in dataset A.

1.3 Merge data

Merge dataset B into dataset A using country and year to match observations. The dataset resulting from the merging procedure can be called `merged_df`. Please perform the merging so that the resulting dataset, has the same number of rows as dataset A.

2 Descriptive statistics

From this point onward, please continue your work only with the merged dataset `merged_df`. (We also provide this dataset in case you have not merged properly)

2.1 Summary statistics

For each country-year pair in the merged dataset, please compute the percentage of respondents who report being members of a union, as well as the GDP per capita recorded in that country-year.

Store this resulting country-year data as a data frame in a new R object called `summary_df`.

2.2 Display table

Display `summary_df` as a table.

3 Analysis

The next questions use `summary_df`. If you did not create `summary_df` successfully you can use the copy that we have sent you.

3.1 Compute correlations

Compute the correlation across country-years between the percentage of respondents who are members of a labor union and the GDP per capita.

3.2 Scatterplot 1

Please produce a scatterplot of the relationship between union membership and GDP per capita. Plot union membership on the X-axis, and GDP per capita on the Y-axis.

3.3 Scatterplot 2

Please create a new version of the same scatterplot as above, but here give a different color to the points if the year of the observation is 2004. All other years will have the same color on the plot, except the countries that were measured in 2004.

3.4 Regression

Run two OLS regressions, one of union membership on year (regression 1), and one of GDP per capita on year (regression 2). Display your output.

4 Editing

Please read this text and suggest improvements.

"In the last twelve month's there have been no less protests about the conflict then any time before, this has lead some to wonder if empathy is dead and what affect that might have on future stability (see eg Masterson 2022 and Peters 2021, 2022.)"