Introduction to Data Science with R

Moed A, 1.7.2021

Dr. Avihai Lifschitz

The excel file "pwt100.xlsx" contains information on countries from the Penn World Table database. The file includes the data and a legend explaining the variables. The file "country_categories.xlsx" contains additional data required for the exam.

Please read the instructions carefully before solving the exam:

1. You should submit a single .R file with code that answers the questions below. Note – the grading is based on the code only – there is no need to submit the actual plots or dataframes that you need to prepare.
2. Add your ID in a comment in the first line of the file. Do not add your name.
3. Do not use absolute paths when reading the data files.
4. Make sure to name the variables reasonably, use good indentations, separate each question with a clear comment, and comment on non-trivial code. Some (small) part of the grade will be based on these style requirements.
5. Make sure that all plots have a proper title, that the x/y axis titles are properly named, or removed if redundant, and that all the labels are clearly shown (i.e. not overlapping)
6. Whenever a question is not for a specific value ("based on the last year" rather than "based on 2019"), it is better to generalize the code, but if you are not successful use hardcoded values in order not to be blocked

---

Part 1: Descriptive statistics (45%)

1. Load the data from "pwt100.xlsx". Out of the two GDP variables, keep only the output side GDP and name it "gdp". Create a new variable for GDP per capita.
2. Load the income level data and the continent data from "country_categories.xlsx" and join them to the data loaded in Q1. Use the correct join function such that only countries where both the income level and the continent are available are kept in the joined dataset.
3. Create a dataframe which shows, per year, the number of countries for which the population data is available. Which is the first year where population data is available for at least 100 countries?
4. For the year you found in Q3, create a scatter plot, which shows the GDP per capita as a function of the human capital index, for all the coutnries. Add to the plot the relevant linear regression line without the confidence interval, **separately** for each income level (i.e. 3 separate regression lines).
5. Create, in one pipe, a column chart showing for each continent its share in the world GDP, for the last available year, in percentage terms (i.e 37, not 0.37). Show above each column the share in

percentage syntax (i.e. “37%”). Remove the x axis title of the chart. Assign the chart to a variable named “gdp_share_chart”.

Part 2: Human capital advancement (35%)

6. Create a dataframe with the change in the human capital index, for each country, between 2000 and the last available year. The dataframe should also have a variable with the continent of each country.
7. From the dataframe created in Q6, create a dataframe that will hold just the countries with the highest and lowest change per continent (i.e. two countries from each continent), plus one row with the average change for all the countries in the sample (global average).
8. Show a column plot of the change per country (plus the average for all countries) for the dataframe created in Q7, with the countries ordered by the change level. The plot should have 3 colors – one for countries for which the human capital index between 2000 and the last available year has increased, one for countries for which it has decreased and one for the global average change.

Part 3: growth (20%)

9. Loop on all income levels, and for each income level create a yearly line plot of the average GDP per capita for people living in countries with this income level, by continent (i.e. one line for the GDP per capita of each continent). Within the loop, add the plot that you created to a list. After the loop, print the number of plots you have in the list.
10. Create a line plot that will show, per year, the ratio of the GDP per capita of low and emerging economies, in comparison to advanced economies. The line plot should include 2 lines – one for the ratio of low income economies and one for the ratio of emerging income economies, both in comparison to advanced economies (i.e. if the GDP per capita in 1955 for emerging economies was 1500$ and for advanced economies it was 10000$, the ratio for 1955 for emerging economies is 0.15). **Note** – this question is not a continuation of Q9 in any way and should be done independently.