Introduction to Data Science with R

Homework Assignment #3

HA #3 is analyzing the age structure of African countries. The questions themselves only state the required outcome – try to think for each of then what are the required steps and progress as much as you can. At the end of the form there are additional hints and more detailed steps that you can use, by I recommend to first try without looking at the hints.

Please read the instructions carefully before solving the assignment:

- Create an R script that performs the required operations and submit the **R script file only**.
- Do not use absolute paths in reading or saving files. Assume a data subfolder called "Data" and a results subfolder called "Results". Assuming that the data file resides in the Data subfolder, **your script should run on any computer without errors** and with no need for any change.
- Whenever there is a numerical question, ("What is the median of…"), running your code should calculate the required answer, assign it into a variable and answer the question with a full sentence (i.e. the code should print "The median of something is 17").
- When asked to remove variables or observations, you should continue the assignment with the reduced dataframe.
- Make sure that your file has **meaningful variable names** and is **properly documented** with comments, like in the R scripts presented in class.
- Make sure that all the plots have a proper title, that the x/y axis titled are properly named, or removed if redundant, and that all the labels are clearly shown (i.e. not overlapping)
- Please use only the R libraries presented in class and try to use pipelines as much as possible.

---

1. Load the data from "population-by-broad-age-group.xlsx" and use the list of African countries in "African countries.xlsx" to keep only African countries. Make sure that you **do not miss any country** even if spelled differently. Keep the subregion in the joined dataframe.
2. Create a plot, for 1950, which shows the share of each age group for eastern African countries. Make sure the age groups are ordered correctly.
3. Create a plot for the total population of countries, ordered by population, in the latest available year. The plot should also include a bar for the simple average of countries and a bar for each subregion average, highlighted in a different color (one color for the average of all countries and one for the subregion averages).

4. Calculate the average population growth rate per subregion between the first and last year of data and show on a plot, with the average rate written above each bar in percentage format (i.e. 1.8%).
5. Create a dataframe with the average growth rate for each country per decade (1950-1960, 1960-1970, etc), in a tidy format (i.e. with 3 variables – Country, Decade and Growth).
6. Spread the dataframe to a human-readable wide format (each country in a row and each decade in a column) and save it to an Excel file.
7. What is the current population of Africa? Assuming population growth rates per country will stay as they are in the last decade, what will be the total population of Africa in 2021?
8. Harder: Still assuming population growth rates per country will stay as they are in the last decade, when will Africa's population be double what it was in 2020?

More elaborate instructions:
Q2: First filter the data to keep only the countries and year relevant for the plot, then gather the data so you will have an age_group variable to use for the "fill" mapping.

Q3:

a. Create two separate summaries, one for the average of all and one for subregion averages and bind each one to the dataframe of the latest year data. Think: why should you add the average for all first and not the subregions first?
b. During the preparation of these summaries, also prepare a geography variable for the highlighting in the plot
c. When creating the averages for the subregions, note the names of the variables before binding and change as needed for the bind_rows to create the dataframe you need.

Q4:
This is very similar to what we did in class for GDP growth rates, just note that the average growth per subregion is not the average of the countries in each subregion, but the average growth of the entire subregion, so you need to create a dataframe for the total population of each subregion and calculate growth rates on that

Q5:
The question requires you to loop on the 7 decades and create a dataframe with the growth rates for each decade, then bind the 7 dataframes together using bind_rows. It will probably be simpler to code the calculation for one specific decade (for example the decade starting at 1950 and ending at 1960) and see that the calculation works, before creating the loop. Once this is done you can go for the loop with the instructions below:

a. Prepare an empty list.

b. Loop on the 7 decades of data by creating a vector of the first year of each decade, and filter for the start and end years of the decade (i.e year is either first year of decade or first year + 10).
c. Using data for these two years, calculate the average population growth rate like we did in class for GDP (the formula for growth rate is the same).
d. Add a variable to the dataframe with the first year of the decade.
e. Keep the resulting data frame in the list.
f. At the end, you can use bind_rows on the entire list of dataframes, see ?bind_rows for details.

Q7:
Join the dataframe you created in Q6 with the growth rate per decade with the dataframe with total population per country in 2020. Then, calculate a new variable with expected population for 2021 per country. The sum of this new variable is the total population.

Q8:
Continuing with the dataframe you created in Q7, you can calculate expected population from a certain year to the following year from the year and the 2010 decade growth rates. Do that in a while loop, after each iteration check the total population to see if it already exceeds twice the 2020 level. Count the number of required iterations to check at which year this will happen