INFO370 Problem Set 2: Data manipulations (100 pt)

June 29, 2021

Instructions

This is the second problem set. It is noticeable more complicated than the first one, so start early!

This problem set asks you to analyze and manipulate a world development dataset. You'll be using filtering, grouping, and related functions. This requires you to extensively use pandas library. The background is provided by McKinney (2018), chapters 4, 5 (numpy and pandas), 7 (data cleaning). The basics is also explained in python notes http://faculty.washington.edu/otoomet/machinelearning-py/numpy-and-pandas.html

General requirements:

- All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licensed as Creative Commons (CC-BY-SA). This means you have to attribute any code you pick from SO (a link to the question/answer webpage will normally do).
- As the final submission, you should submit a) code; b) output; and c) explanations. If you are working with jupyter notebooks, all this will be included automatically but you still have to submit both your original file (so you grader can actually run the code), and an html version of it (which is much faster to check).
- Working together is fun and useful but you have to submit your own work. Discussing the solutions and problems with your classmates is all right but do not copy-paste their solution! First understand it, and thereafter create your own solution. Please list all your collaborators!
- Your results will only count if accompanied with sufficiently and clear explanatory text. Just plain output, with no explanation, will not count.
- Be sure that each visualization (graph or table) adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
- Don't output irrelevant, or too much of relevant information. A few figures is helpful. A few thousand figures is useless. In particular, don't print thousands of lines of data!

1 Setup (10pt)

In this problem set you will work with gapminder data. The data is complied from https://www.gapminder.org/data/ but you use the ready-made file on canvas. There is an accompanying documentation file that you should consult for the meaning of the variables.

1. (2pt) Load the data
2. (8pt) Do basic sanity checks:

(a) How many variables (columns) is there in the data? Ensure you know the variables in the data. Keep the documentation nearby.

(b) How many rows of data is there?

(c) print the first few lines of data. Does it look reasonable?

Through the course we expect you *always* to do similar sanity check every time you load data.

2 Wealth (30pt)

First, let's do some data exploration. Answer the following questions: show the code, the computation results, and comment the results in the accompanying text.

1. (2pt) How many different countries are there in the data?
2. (3pt) What is the earliest and the most recent year in the dataset?

Now let's define wealth as GDP per capita and let's explore countries by average wealth.

3. (3pt) For which year do we have the most recent GDP data?

   Hint: You can remove all missing GDP data to answer this question
4. (3pt) What is the average wealth on this planet as of 2019? Let's just compute average GDP across all countries for 2019 and ignore the fact that countries are of different size.
5. (4pt) But not all countries may have this final year. Which 5 countries countries have most recent years missing? Till which year do they have data?

   Hint: you may group by country and find max value for the year. In the resulting series, find the min/max. Check out the `nlargest` method.
6. (4pt) Now let's compare the continents. We'll do it easy again and just compute the average wealth (i.e. GDP) for each continent in 2019, and we use *region* as continent. We disregard the fact that countries are of different size. Print the continents, and the corresponding GDP in a decreasing order.

   Remember to use only the most recent data!

   Hint: check out methods `groupby` and `sort_values`.
7. (6pt) But this was just about the average numbers. Now for each continent let's also find the richest and poorest country, the corresponding GDP, and population (for 2019). Print these in a readable form.

   Note: While this gives a hint about inequality, we still completely ignore the intra-county inequality. Quite likely the rich in the poor countries earn more than the poor in the rich countries. But these measures unfortunately do not let us to assess this.

   Hint: while you can extract the values using construct like `data.gdp == data.gdp.min()`, you may also check out methods `idxmin` and `idxmax`. If this seems overwhelming, then just loop over continents, and for each continent find the richest and poorest country as of 2019 (check out methods `nlargest` and `nsmallest`).
8. (4pt) Comment the list of poorest and richest countries. What do you think about these lists. Did you know that Bermuda is the richest country in Americas? Do you know why? Why do most of the rich countries have small population?

3 Health (30pt)

Health is a complex concept, but fortunately we can proxy health with life expectancy (LE). It is a natural index of health that has been measured rather well for a long time already.

1. (3pt) For how many countries we do not have LE for 1960 and 2019?
2. (4pt) What is the lowest and highest LE in data? Which years/countries does this correspond?
3. (4pt) If you did this correctly, you notice that the shortest LE is less than 20 years. What historical events does it correspond to? (You may consult e.g. Wikipedia).
4. (5pt) Find the country with longest and shortest LE for each continent.
5. (6pt)Which countries had the fastest and slowest growth rate of LE? You can compute the growth rate (pct per year) as

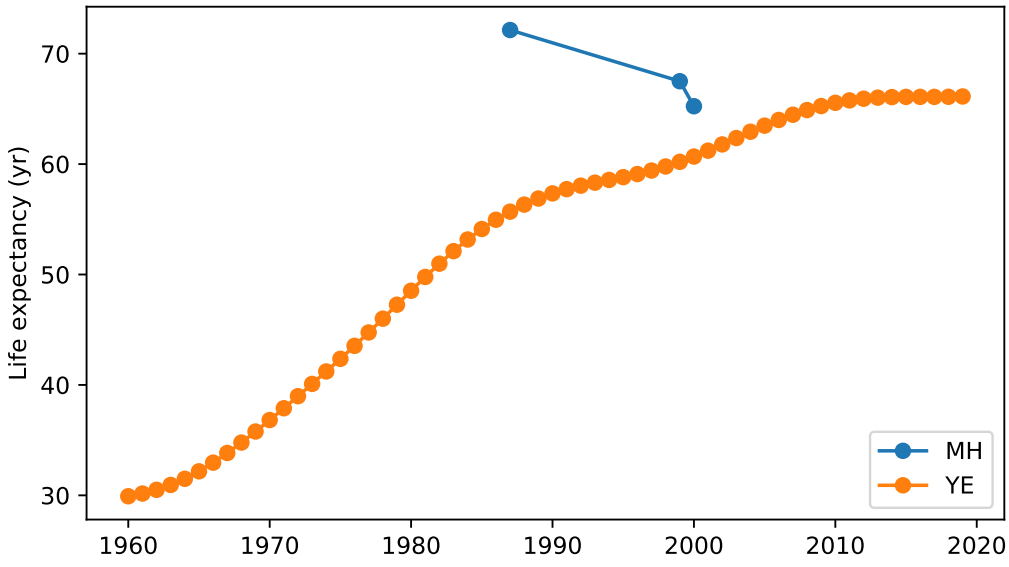
$$g = 100 \left[\left(\frac{LE_1}{LE_0} \right)^{\frac{1}{n}} - 1 \right] \tag{1}$$

   where LE_0 is the life expectancy at the beginning of the period, LE_1 is it at the end of the period, and n is the length of the period in years.

   Hint: for each country, compute the first valid year of life expectancy, last valid year of life expectancy, and find their life expectancies for the corresponding years. If the result is a series, just order it, if it is not a series, make it into a series and order it.
6. (4pt) Do you see a pattern (or multiple patterns) here? Remember: you are looking at growth of life expectancy over an extended period.
7. (4pt) Now pick the top and bottom countries in terms of life expectancy growth you identified above, and make a plot where you show how life expectancy has changed over time in these two countries.

   Example results provided below, but you need to come up with your own solution with a comparable result!

Example results:



One can see that life expectancy in Yemen (YE) has been growing fast, but has flattened out during the civil war in 1990-s and during the current wars and cholera epidemics. For Marshall Islands (MH) we have very few data points. Also, the islands have large problems with diabetes and obesity (wikipedia), this may have resulted in falling life expectancy.

4 Gender disparities (30pt)

Let's see how has gender disparity changed over time. We use difference between male and female youth literacy rate as a measure of gender disparity. Let's define it as *male* − *female*, i.e. positive numbers indicate disparity in favor of males.

1. (3pt) How many valid male/female youth literacy rate values do we have? How many are missing?
2. (5pt) How many missing cases do we have by year? How does the data quality change over years? Make plot to demonstrate this!

   Hint: you can get this done using `.apply` method, but if you feel this overwhelming use loops: create an empty list, loop over all years, and for each year filter the data and add the number of missings to the list.
3. (8pt) How has the world gender disparity developed through years? This time compute the weighted average where weights are the corresponding total population size (well, should use the corresponding gender/age group size but let's stay simple).

   Hint: you can do it like this:

   (a) Select only cases where literacy data is not missing
   (b) For each year, compute the total world population (you can just add new variables to the data frame).
   (c) For each country-year (i.e. each observation), compute the country population times literacy for both males, females; and divide this by world population.
   (d) Now group the above by year and sum. This is the weighted average.
4. (4pt) Make a plot that shows how has the disparity changed over time.
5. (4pt) If your result is like mine, you see that the disparity is volatile, and fluctuating b/w 0 and 25 pct pt, mostly in favor of males. However, the trend is clearly downward.

   Does this result indicate that gender disparity is an issue that the world in recent years has mostly overcome?
6. (5pt) Find the countries with the disparity more than 5%, in favor of either boys or girls, as of 2018.

   What are these places? Do you know why some of those are in this list?
7. (5pt) What are the 10 countries with larges disparity *in favor* of girls?

5 How much time did you spend?

And finally-finally, tell us how much time (how many hours) did you spend on this PS!

References

McKinney, W. (2018) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, O'Reilly Media, 2nd edn.