INFO370 Problem Set 2: Data exploration, descriptive Statistics

April 5, 2022

Instructions

This problem set asks you to analyze real data, the gapminder dataset of world development indicators. We do not do any statistical analysis here, but you should check missings and look for certain inconsistencies in data. We also ask you a few questions about what this data implies, and linking it to historical events–why do you see such results as what you see.

- Comment and explain your results! Just numbers with no explanation will not count! Remember: your task is to convince us that you understand, not just to produce correct results!
- Include the question numbers in your solution! You may leave out the text of the question you are answering.
- Ensure your submission is readable. Depending of the complexity of your code and the choice of variable names you may need more or less explanations. For instance, if you are asked to find largest income, then code

  ```
  print(largestIncome)
  ```

  needs no additional explanations. But if you choose to call the variable "maxy", then you may need to add a comment:

  ```
  print(maxy)  # 'maxy' is the largest income
  ```

- Make sure your solutions are your own! It is all well to work together, to talk to other students, help them and let them to help you. But at the end you have to understand their suggestions and write your own solution. Please list the other students you worked together with–this helps to avoid too many red flags when graders find solutions surprisingly similar.

Good luck!

List your collaborators here:

1. ...
2. ...

1 Cleaning data (20pct)

In the first problem you are asked to do basic technical data exploration. Show your code, the computation results, and comment the results in the accompanying text.

1. Load the "gapminder" dataset (the same you used in PS1)
2. Do a quick check: how many rows and columns do you have? Does the first few lines of data look reasonable?
3. How many missing values are there in each variable? Comment the results.
4. You should notice that "time" is missing in a number of cases. Analyze the cases with missing time.

   You can do it in the following manner: print out a small selection of cases where time is missing. To make the print-out manageable, include only the most basic variables, e.g. name, iso2, and total population.
5. Describe what you see there. Why do you think these observations are missing?
6. But what about Taiwan? It is a highly developed country of 24 million. How many records do you find for Taiwan? Why do you think you find this result?

   Hint: use `Series.str.contains` method to search for country names that contain "taiwan". See Python notes, Ch 4.6, and the respective documentation.
7. You may also be surprised that in names/iso-2 codes have different number of missing cases. How many cases do you find where

   - name is missing but iso-2 code is there?
   - iso-2 code is missing but name is there?

   Provide examples in both cases
8. If you did the previous question correctly, then you saw that one of the countries with missing name is Namibia. Can you find two-letter country code for Namibia?

Finally, let's get a few basic descriptive facts about these data:

9. How many different countries are there in these data?
10. What is the earliest and the most recent year in the dataset?

---

2 Wealth (35pct)

Now let's go to some more serious data exploration. For simplicity, let's define wealth as GDP per capita and let's explore countries by wealth.

1. For which year do we have the most recent GDP data?

   Hint: You can remove all cases where GDP is missing and see what is left.

2. What is the average wealth on this planet as of 2019? Let's just compute average GDP per capita across all countries in 2019 and ignore the fact that countries are of different size, and the fact that some of those are not countries at all.

3. But not all countries may have the same most recent year where GDP data is present. Which 5 countries countries have the largest number of the most recent years missing?

   (For instance, imagine there is GDP data for Funan for 2015, 2016, and 2017; but for Khmer Empire only for 2015. Hence Khmer Empire has two of the most recent years missing, while Funan has everything present.)

   Till which year do these countries have data? What do you think, why do these countries have issues with more recent 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.

   Hint2: two of these countries are Lichtenstein and Faroe Islands.

4. Now let's compare the continents. We'll make it easy again and just compute the average wealth (i.e. average GDP per capita) 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 average GDP per capita in a decreasing order. Do you think this order is reasonable?

   Remember to use only the most recent data!

   Hint: check out methods `groupby` and `sort_values`.

   Hint2: you should see value around 11,800 for Oceania.

5. But this was just about the average numbers. Now for each continent let's also find the richest and the poorest country, the corresponding GDP, and population (2019).

   Note: While this gives a hint about inequality, we still ignore the intra-county inequality. Quite likely the rich in the poor countries earn more than the poor in the rich countries. Unfortunately we cannot tell based on these data.

   Hint: if you can find a good solution yourself, go for it! But here is a suggestion how you can do it, step-by-step.

   (a) Check out indexing with `.loc` and `.iloc` (McKinney, p 143)
   (b) Check out methods `.idxmax` and `.idxmin` (McKinney p 158-159)
   (c) Find the index of the largest GDP value in data (use `.idxmax`).
   (d) Extract the row of dataframe that corresponds to the index you computed in 5c
   (e) Now repeat the two previous steps either with `.groupby`, or if you cannot figure this out then with a for-loop over all continents.

   Alternatively, you can extract the values using construct like `data.gdp == data.gdp.min()`. You can also just loop over continents, and for each continent find the richest and poorest country as of 2019 (check out methods `nlargest` and `nsmallest`).

6. Finally, let's put all this into a single data frame. The data frame should contain continent name as index, and for each continent 6 variables: name, GDP and population for the poorest country, and the same for the richest country. Pick suitable variable names!

   The result might look something like:

```
          poorest    GDP          population  richest     GDP           population
region
Africa    Burundi    208.07473    11530580.0  Seychelles  15048.74693   97625.0
...
```

   Hint: consult Python notes 4.6: Combining data into data frames. You can use `.set_index` to convert a variable into index.

7. (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 (35 pct)

This is a more demanding problem. 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 decades, and in some places even centuries!

1. How many countries do not have LE data for year 1960? How many countries do not have this information for year 2019?

2. What is the shortest and longest LE in data? Which years/countries does this correspond to?

   Hint: you can use methods `nlargest` to extract dataframe rows, and `idxmax` to extract index of row with the maximum value.

3. 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 Wikipedia).

4. Find the country with longest and shortest LE for each continent.

   Hint: you can use the same approach as for GDP by continent.

5. Now lets find, for each country, the first valid LE. Find the corresponding year, and LE value.

   Hint: you can use the same approach as when finding the largest GDP on each continent: find the index for the minimum year of each country, and then extract the rows based on the index.

6. Now repeat this with the most recent valid LE.

7. Finally, let's compute the growth rate. 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.

Show the countries, drowth rates, and the corresponding life expectancies for 10 fastest and 10 slowest growers.

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. You can do it in a fairly similar way as in the previous question. Now use this formula for each country. At the end just order the result.

Hint2: depending how did you do the computations, you may find a bunch of countries with growth rate exactly 0.000. Where are these numbers coming from? What do these mean?

8. Do you see a pattern (or multiple patterns) here? Remember: you are looking at growth of life expectancy over an extended period.

---

4 Graphical Analysis (10 pct)

Finally, it is time to make a nice plot of your previous results.

1. Make a plot of life expectancy over time, where you display the following countries: a) the country with largest life expectancy growth, b) the country with smalles life expectancy growth, c) Ukraine, d) two (or more) countries of your choice.

   Ensure that the countries are displayed in different colors, and clearly labeled. Do not overload the figure with too much data, ensure it can be easily understood.

   Comment your results

How much time did you spend on this PS?