INFO370 Problem Set 2: Data exploration, descriptive Statistics

January 22, 2023

Instructions

This problem set asks you to analyze real data, the gapminder dataset of world development indicators. Your task is to check for missings and certain inconsistencies in data, and do only basic descriptive statistical analysis. 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
  ```

- Data Science requires technical expertise, and pairing with contextual knowledge. Several of the questions below link to historical events, or geography - we expect you to not know everything and to Google for some results! You will need to cite your sources for full credit!

- 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. (1pt) Load the "gapminder" dataset (the same we used in class)
2. (2pt) Do a quick check: how many rows and columns do you have? Does the first few lines of data look reasonable?
3. (2pt) How many missing values are there in each variable? Comment the results–which variables are good, which ones nearly unusable?
4. (4pt) You should notice that "time" is missing in a number of cases. This seems surprising. 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. (2pt) Describe what you see there. Why do you think these observations are missing?
6. (4pt) 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
7. (3pt) If you did the previous question correctly, then you saw that one of the countries with missing name is Namibia. Can you figure out what is two-letter country code for Namibia? Why do you think, out of all countries, it is Namibia that has its iso-2 code missing?

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

8. (1pt) How many different countries are there in these data?
9. (1pt) What is the earliest and the most recent year in the dataset?

2 Wealth (20pct)

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

1. (2pt) 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. (2pt) 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. (9pt) 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?

   Explanation: imagine the dataset only contains GDP for two countries–*Funan* and *Kmer Empire*, and the following years:

	Funan	Khmer Empire
2010	300	300
2011	400	400
2012	n/a	500
2013	n/a	600
2014	n/a	n/a

   As you see, Funan has years 2012-2014 missing, while Khmer Empire has only 2014 missing. Hence Funan has a larger number most recent years missing (3) versus Khmer Empire (1).

   What is the most recent GDP data for these countries? Try to create a table like this:

Country	Last year of GDP
Funan	2011
Khmer Empire	2013
...	

   What do you think, why do these countries have issues with newer 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. (7pt) 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. Order the continents by wealth so that the richest one is at top. 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.

---

3 Descriptive Stats (30pt)

Now let's take a closer look at the world wealth. Unfortunately, these data are not good for analyzing inequality–we only have values for country averages, and we cannot tell how much inequality there already is inside countries. But we use what we have. Also, as above, we ignore the fact that countries are of different size and analyze just the distribution of *country averages*, using GDP per capita as the wealth measure. First we compute the simple descriptive figures and thereafter two inequality measures: *quintile share ratio* and *Pareto ratio*. See Lecture notes 1.2.2 "Describing data", toward the end of the section (p 14 for now).

Let's look at the world both in 1960 and 2019. We'll analyze how the global wealth has changed over that period.

1. (2pt) First the simple descriptive statistics: compute minimum, maximum, median and mean wealth both in 1960 and 2019.
2. (5pt) What do these figures suggest? Has the world become richer? Has it become more equal/inequal? Anything else you notice here?
3. (3pt) Plot histograms of world wealth for 1960 and 2019.
4. (5pt) Compare these histograms. Comment what do you see.
5. (5pt) Compute the *quintile share ratio* of GDP per capita for 1960 and 2019. What does it tell you–has inequality grown over time?
6. (7pt) Now compute the *Pareto ratio* for the same two years. Will your conclusion be the same?
7. (5pt) What do you think–are these data useful to tell something about how the world inequality has changed over the last 60 years?

---

4 Health (15 pct)

This is a more demanding problem. Health is a complex concept. Here we measure health through child mortality (CM), calculated as how many children, out of 1000, do not live till their 5th birthday. As you can see below, this has been rapidly falling through the past 50 years in most of the world.

1. (3pt) How many countries do not have CM data for year 1960? How many countries do not have this information for year 2019? So is data improving over time?

2. (3pt) What is the largest and smallest CM 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. (9pt) For each continent, find the country with smallest and largest CM (as of 2019).

   Hint: you can use different methods (e.g. for-loop), but you can also use `groupby` and get the index of smallest/largest CM out. Check out methods `idxmax`/`idxmin`.

---

5 Graphical Analysis (15 pct)

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

1. (15) Make a plot of child mortality over time, where you display the following countries: a) the country with largest child mortality (as of 2019), b) the country with smallest child mortality (as of 2019), c) Egypt, d) two (or more) countries of your choice.

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

   Comment your results.

   Hint: Consult Python notes Ch 4. You can achieve this with matplotlib, but it is much easier with seaborn.

How much time did you spend on this PS? (I spend 3h on updating it).