INFO370 Lab: data programming

April 4, 2022

Instructions

This lab is about pandas–python data frames. Consult python notes Ch 3: Numpy and pandas and McKinney's book.

- You should answer the questions (type the code) in a jupyter notebook, in *code cells*. Please also mark which question you are answering (you may copy-paste text to the notebook, but just question number will aslo do). When explaining and commenting on something, do this in a *markdown cell*. See Python notes, Ch 2.2.3 for a brief explanation.
- Please submit your solutions in two forms:
  - the notebook: the `.ipynb` file (this can be run)
  - html. You can get html from jupyter notebook file menu: *File – Export Notebook As... – HTML* if you are using jupyterlab, and *File – download as – HTML* if you are using jupyter notebook.
- 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, and thereafter create your own solution. Please list all your collaborators on the solution.
- Don't be scared. We are here to help you learn. :)

In this exercise you will get hands-on experience in importing data into structured format, summarizing data using descriptive statistics (e.g. sum, average, etc.) and manipulating data including indexing, slicing and grouping. throughout The data you will be working with is the pulled from Johns Hopkins University COVID-19 Data Repository, and we will be focusing on confirmed COVID-19 deaths throughought the US. It is a daily record, collected from local and state health departments. You can find additional information and updated data on github.

Some of the variables that are not quite obvious:

FIPS US only. Federal Information Processing Standards code that uniquely identifies counties within the USA. For instance, King County has fips code 53033 where "53" stands for Washington.

Admin2 County name

UID Similar to Admin2, but includes locations like Diamond Princess, unassigned, etc. (Not needed for this activity)

"6/16/2020" (and other dates): cumulative number of confirmed COVID-19 deaths as of June 16, 2020

1 Data import and summary

1. Download the file *time-series-covid19-deaths-us.csv.bz2* from canvas (or directly from the GH page linked above) and load it into a pandas dataframe.

   Hint: use `pd.read_csv`, but make sure you use the right separator or the import won't look right (`sep=...`). Print the first few lines of it as a quick check.

   Note: `pd.read_csv` can read compressed files directly, you do not have to decompress it.

2. It's time to get to know your data! Report the number of rows and columns in the dataset.

3. What variables does this dataset have? Report the variable names along with the data type of each variable.

   Hint: check out method `dtypes`

If you did this correctly, you ended up with hundreds of variables. This is too much for this lab. Let's take a subset.

4. create a sub-dataframe that contains all observations but only variables *FIPS*, *Admin2*, *Province_State*, *3/1/20*, *3/15/21*, and the most recent date (*3/26/22* for now). Check that you did this correctly!

   Hint: you should check the number of rows and columns, and print out a few lines of the subset.

Below, we only work with this subset.

---

2 Explore deaths in Washington

1. Take the subset you created above and filter results for Washington state only. Store it into a new variable.

2. Add a new variable, "growth", to the WA data. It should be growth in the cumulative number of deaths through the last year, between 3/26/22 and 3/15/2021.

3. You may notice a warning: "A value is trying to be set on a copy of a slice...". Why do we get this warning? How can we get rid of it? Repeat the previous steps in a way that you do not get the warning!

   Hint: see Python notes Ch 3.3.5 "Modifying data frames".

4. Extract the newly created variable "growth" as *series*. Show that it is a series!

5. What is index of the growth series? What do you think, is it an useful index? What might be a better index? Why?

6. Replace the original index with one you think is better. Show that it worked (you may just want to print a few lines).

   Hint: you may also consider questions below.

---

3 Data interpretation

1. Sort your series of growth numbers starting with the counties with the largest growths down to the smalles. Which 10 counties did experience most growth through 2021-2022?

   Hint: check out `Series.sort_values` method.

2. If you did the previous question correctly, you found that King county has the largest number of growth (1215). What does this tell you about the efficacy of social distancing measures in Seattle?

   Note: Seattle is (mostly) located in King county.