# INFO370 Lab: data programming

January 9, 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). When explaining and commenting on something, do this in a *markdown cell*.
- 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.

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 thse into a pandas dataframe.

   Hint: use `pd.read_csv` and check what is the right separator (`sep=...`). Print the first few lines of it as a sanity 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*, and the most recent date (*1/9/22* for now). Check that you did this correctly!

   If you pick a different date, then adjust all the questions/answers below accordingly.

   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.

5. What is the number of NULL/NA values in each column of the dataframe?

# 2 Explore deaths by state

1. What are the 'states/provices' with the most confirmed deaths as of 1/9/2022?

   Hint: there are no state level data, so you have to aggregate all counties within a state. Check out `DataFrame.groupby` method, `Series.sum` method, and `Series.sort_values` method.

2. What is the data structure you got? What is its index and what is its value?

   Hint: you can use function `type` for data type, and methods `.index` and `.values` for the latter.

3. What is the total number of confirmed deaths in Washington State as of 1/9/2022?

4. What is the average number of deaths for a county in Washington State at that date? How does it compare to average number of cases for a county nationally?

   Hint: the answers should be 244 and 250.

# 3 Data interpretation

1. If you did the previous question correctly, you found that the average county in Washington has fewer deaths than the national average. What does this tell you about the efficacy of social distancing measures in Washington?

2. In Question 1, we found that some columns have NULL/NA values. Briefly look at those values (you may just print the corresponding data rows). Can you hypothesize why those rows are included as separate entries? Googling some entries may help, as well as referring to the Github data source for variable definitions. Furthermore, how does this hypothesis affect how you would go about analyzing the dataset?