CMSC 201 Section 40
Spring 2020
Project #30 - Using Jupyter Notebooks and Graphics Routines to Analyze Data

---

Value: 80 points

Due Date: May 11, 2020 at 11:59:59 pm

High level description:

This project is different from the previous two in that you will do the programming on your own computer. If you do not have access to a computer that is capable of running Jupyter notebooks, please let Prof. Arsenault know immediately so that arrangements can be made.

Details:

Step 1: Install pip

Pip – the Package Installer for Python – usually comes included with Python 3 releases. If you have Python 3.7 or 3.8 installed on your computer, you most likely already have this step done. To verify, on Windows open a command window and type

```
pip list
```

On a Mac, open a terminal window and type the same string.

If you get a list of loaded modules, move on to Step 2. Otherwise, you'll have to install pip.

Step 2: Install the Jupyter Notebook

Download and install Jupyter Notebooks on your computer.

1. Go to https://jupyter.org/

2. Go to the section labeled “The Jupyter Notebook”.
3. Click the link labeled “Install the Notebook.”
4. This will take you to a page with a lot of different installation instructions. Open a command window or terminal and type “pip install notebook” (without the quotes!). This will install Jupyter Notebook on your system.
5. To start Jupyter Notebook, type the command “jupyter notebook” (without the quotes) in a terminal or command line window.
6. This will start Jupyter Notebook running in your default browser.

Bookmark the following link: https://jupyter.readthedocs.io/en/latest/running.html#running

That will answer a great many of your questions.

3:54

Step 3: Install plotnine and pandas

In a command window (Windows) or terminal (Mac) type pip install plotnine

And

pip install pandas

3:58

Step 4: Create a new Jupyter notebook.

In a command window (Windows) or terminal (Mac) type jupyter notebook. This should cause a new tab or window to open in your default browser. In that tab, select "New". Select "Python 3" from the list of options that pops up. This will open a new tab in your default browser, containing a new, untitled Notebook.

3:59

Step 5: Title

Titlethis notebook "Project 3" by clicking on the name field and typing "Project 3" when you're asked for a new name.

4:00

Click "File', then "Save as" and then give this file a name you will remember on your computer.

Step 6: start your notebook!!

Click in the one cell that shows by default. Click "Cell" on the menu above and click "Type". Then click "Markdown."

4:08

Now it's time to write your program heading as Markdown. Here's a tip-sheet on how to do that: https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Working%20With%20Markdown%20Cells.html

4:08

Using Markdown, in that cell, write CMSC 201, your section number, Spring 2020, Project 3; the file name; and the date.

4:09

Then add another cell below. It should be a code cell. Start by importing the plotnine module that you installed. Then import pandas as pd.

Step 7: Get the data file

The data file used for this project is called "P3_data.csv." It is available for copying from Prof. Arsenault's pub directory on gl at

/afs/umbc.edu/users/a/r/arsenaul/pub/P3_data.csv

It can also be obtained on the Github site, on the class Slack channel, or on Blackboard

This data file is a subset of the data that Johns Hopkins University publishes daily on the number of confirmed COVID-19 cases globally.

This is a .csv – Comma Separated Value – file with 12 rows and 61 columns. The first row is a header row; the rest start with the name of a country and then the number of confirmed cases of COVID-19 reported in that country each day. The header row and the country column are STRINGS; all other values should be INTEGERS. These are the 11 countries with the highest number of reported cases as of April 29..

Step 8: read in the data file

Add a markdown cell that explains what you're going to do. Then add a code cell and write a function to read in the file and create a 2D list from it. Convert the number of cases to integers.

Here's the neat part: you don't actually have to do that work this time! Pandas has a "read_csv" command that reads the file and converts for you. Your statement should be something like

dataframe = pd.read_csv("P3_data.csv")

or whatever you want to call your 2D list variable.

Step 9: Data analysis and visualization

Add another markdown cell that tells what you're going to do and why. Then, add a code cell that prints out the dataframe you created in Step 8.

Now some graphics. You already imported the plotnine module in the code cell from Step 6. Now time to use it.

Create a new ggplot object and draw a bar chart with country on the x axis, with the y-axis showing the number of cases on 4/29/20. Your statement should involve something like "geom-bar(aex(x="country", 'y=4/29/20")

Step 10: Cases over time

Make a line plot(geom_line()) showing the increase in the number of cases in each country over the months of March and April.