Data Science with COVID-19

David Hay

Introduction

This is a free resource for teachers and students, and is part of the Callysto Project, a federally-funded initiative to foster computational thinking and data in Canadian Grade 5-12 classrooms.

This "Introduction to Data Science with COVID-19 Data" activity uses COVID-19 statistics from Johns Hopkins University CSSE and population data from Gapminder to create some data visualizations using Python code in a Jupyter notebook.

There are student and instructor versions of the notebook. The instructor version includes all of the code and examples, the student version has some code removed. Those links will create a copy of the lesson plans repository in your Callysto Hub account, and open the (student or instructor) notebook.

To log in to the Callysto Hub you and your students will need a Google or Microsoft account. This can be a school division-provided account or a personal account. Callysto does not collect any personal information about accounts.

Outcomes

- Managing Information
  - Accessing online data
  - Data tables
  - Graphing
- Communication
- Problem Solving
- Critical Thinking
- Cultural and Global Citizenship

Python Libraries

A library in Python is a collection of code that can be imported and used in your code. In this notebook we are using pandas (panel data) for data manipulation, plotly for plotting visualizations, and cufflinks for linking pandas to plotly.

Using the Notebook

If your students are comfortable with following directions and have some experience with coding, you may want to just provide a brief introduction and then let students work through the notebook on their own. I'm a proponent of allowing students to develop their competence with computational thinking and data science by using examples before they fully understand everything about those examples.

If data science, computational thinking, and Python code in a Jupyter notebook are all new experiences for your students, you may prefer to provide direct instruction on the functioning of the notebook.

Getting Started

The first part of the notebook imports the data. Run the code by clicking inside the code "cell" and then clicking the Run button in the toolbar at the top.
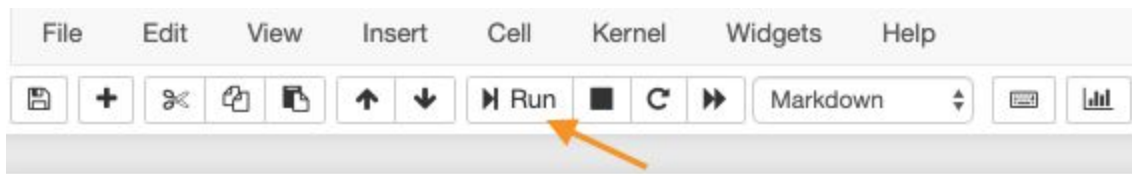
You should see the first five and last five rows of COVID-19 statistics from the date specified in the code cell. You can also change the date and re-run the cell. To see the most recent date for which data are available, check here.

Data Cleaning

The next code cell discards some of the unnecessary data and summarizes the data by country. If you are not interested in including all of the countries from the data set, put a **#** character in front of the line **country_list = covid_stats['Country_Region'].unique()** and remove the **'''** symbols from before and after the list of countries. You can, of course, edit that list of countries.

Graphing Data

Now that we have a clean data set, there are instructions for graphing using the cufflinks library. This is the first time in the notebook that students will need to enter code, they can copy either example and paste it into the code cell. The statement **cf.go_offline()** means that we will be using the library in offline mode and rendering the graphs in this notebook rather than on plotly.com.

Hopefully the rest of the notebook is expressive enough to get you started with data science.

Next Steps

For more information, you can check out our [YouTube videos](#), [online courses](#), or [callysto.ca](#) for learning modules, tutorials, lesson plans, or perhaps coding contests.

Contact

If you encounter any issues or have any suggestions, please get in touch with us at [contact@callysto.ca](#) or [twitter.com/callysto_canada](#).