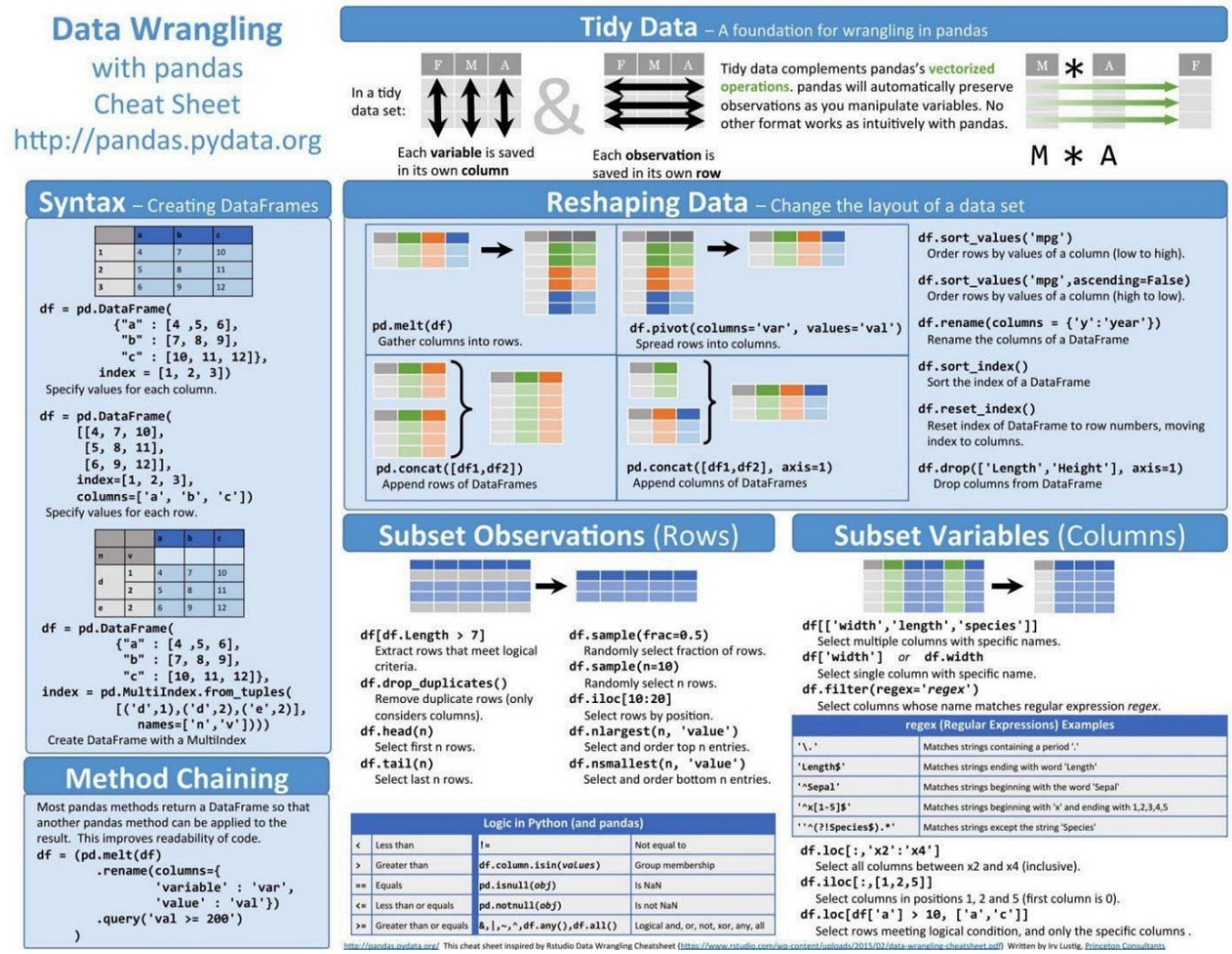

Group Delivery

23/07/2020 to 31/08/2020

—

Group A

Natalio Altube, Tomas Lizarazo, Rosario Marianeschi, Elsa Toribio.

Data Science

The Bridge

Table of Contents

GENERAL VISION 2

GOALS 2

SPECIFICATIONS 2

SOFTWARE 2
HARDWARE 3
REQUIREMENTS 3

STEPS 3

I. RESEARCH THE CONTEXT 3
II. GET DATA 4
III. DATA WRANGLING 4
IV. DATA MINING / CLEAN DATA 4
V. OTHERS. 5
VI. ANALYSIS & CONCLUSIONS 5

General Vision

Our vision is to complete the required project as if we were carrying out a real project “at work”. We aim to work as a group, assigning roles and tasks to each other, maintaining communication and cooperation among ourselves while using the tools and processes learnt during the first 7 weeks of DataScience BootCamp at The Bridge.

Goals

We want to complete parts C and B as a minimum objective. So our specific goals will be:

- Create an API that returns a .json with the logic explained for your group
- The flask server to be executed running the src/api/server.py file
- Use Trello as a “TO DO LIST”
- The creation of the .json,
- Visualize the data assigned to our Group A and practice the structure of a real ‘project’, with all the folders and Python files very well developed.
- Answer the questions proposed by the teachers in the assignment.
- Analyse outliers or some rare data.
- Draw, in different colours and vertically, the moments when the daily death curve increases and decreases.
- Visualize with bars, lines, points and pie charts the daily deaths and infected.
- Write conclusions.

Specifications

To achieve the goals and make the most out of the delivery, below you can see the detailed specifications.

Software

Minimal Software needed for execution is MacBook Air – iOS Mac HD, OS X, 10.10.5

Platforms:

VS Code has been tested on the following platforms:

- OS X Yosemite
- Windows 7 (with .NET Framework 4.5.2), 8.0, 8.1 and 10 (32-bit and 64-bit)
- Linux (Debian): Ubuntu Desktop 14.04, Debian 7
- Linux (Red Hat): Red Hat Enterprise Linux 7, CentOS 7, Fedora 23

Hardware

Minimal hardware needed for visualizing this work:

Visual Studio Code is a small download (< 100 MB) and has a disk footprint of 200 MB. VS Code is lightweight and should easily run on today's hardware.

It is recommended:

- 1.6 GHz or faster processor
- 1 GB of RAM

Requirements

Visual Studio Code

Additional Windows requirements:

Microsoft .NET Framework 4.5.2 is required for VS Code. If you are using Windows 7, please make sure .NET Framework 4.5.2 is installed.

Additional Linux requirements:

- GLIBCXX version 3.4.15 or later
- GLIBC version 2.15 or later

Python 3.6 – Libraries to import: matplotlib, pandas, seaborn, plotly.express, plotly.graph_objects as go, plotly.figure_factory as ff, from plotly.colors import n_colors, from plotly.subplots import make_subplots, nbconvert, nbformat, %matplotlib inline, plot as pltt and pycountry.

Steps

I. Research the context

There is plenty of information about COVID19 nowadays. The majority of the Search engines will show COVID19 on the top places of the ranking, as well as the main headlines in the news.

In summary, COVID-19 is the infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019. COVID-19 is now a pandemic affecting many countries globally. In most cases, COVID-19 causes mild symptoms including dry cough, tiredness and fever, though fever may not be a symptom for some older people. Other mild symptoms include aches and pains, nasal congestion, runny nose, sore throat or diarrhoea. Some people become infected but don't develop any symptoms and don't feel unwell. Most people recover from the disease without

needing special treatment. Around 1 out of every 6 people who gets COVID-19 becomes seriously ill and has difficulty breathing.1

For this particular assignment, we used information found on the website selected by the school program. Our group was designated to show the new COVID19 cases detected in Argentina, Chile, Colombia, Russia and Spain since January 2020 to present.

II. Get Data

The url we used to get the data from, is the following:

https://ourworldindata.org/coronavirus-source-data

III. Data Wrangling

We selected the file that is in .csv format and always up-to-date, to import it in Python and turn it into a dataframe. The information shows all the numbers since the beginning of the pandemic, when the first cases were detected until present. As it had the information about all the countries, and all the categories (such as: total deaths per million, new cases per million, new tests, population, median age, gdp per capita, poverty index, etc.) we arranged it in order to obtain the information we needed.

First, we regrouped it by the date, so we had the information day by day for the 5 countries we needed. We came up with a summarized dataframe, showing the new cases for the countries we had for our analysis: Argentina, Chile, Colombia, Russia and Spain. On that basis, we worked out many functions in pandas to finally obtain the graphs, trends and figures.

In addition, we used the generic dataframe for all the countries of the world to get the total cases, total deaths and, using an operation, find out the recoveries.

IV. Data Mining / Clean Data

The delivery is divided in folders: api, documentation, resources (where all the graphs are included in picture format) and rsc.

In the documentation, we designed a PowerPoint presentation, using a nice template we found online. We included the documentation of the API, and the guidelines for this work.

As far as resources go, we saved different graphs, such as:

- Daily count of deaths for each country
- Daily count of new cases for each country
- Situation for each country
- Outliers for each country

---

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/coronavirus-disease-answers?query=history+of+covid+19

- For the group of 5 countries, we've created the graphs in different formats, e.g. bars, lines, pie chart, scatter.
- We also kept there the graph received from Group D, regarding their countries: Portugal, Venezuela, Turkey, UK and Spain.

In the 'rsc' folder, we've included 'main' subfolder with Options C and B from the Project) and 'utils' subfolder with the Python files, cotaining the functions and coding to be run in the Terminal.

We also allocated a folder for the graphs of each of our countries, specifying the beginning and ending of the alarm state.

V. Others.

As part of the activity, we created a flask server file, to build an API and enable the other work group to get the necessary information from us. That's how they could access our .json file.

Hence, we converted the file containing the mean of new cases detected in the 5 countries we were assigned into a .json. We made the dictionary (.json) available for the other group on the screen of the API. So they accessed through our IP, endpoint and entered our token code.

Creating a clean dictionary, the API and the Flask server where the core of the activity.

VI. Analysis & Conclusions

After reviewing all the dataframes and graphs, we could answer some of the questions and come down to some conlcusions.

We got the .json file from Group D. They represented "t_d_averages", which is the mean of the "total_deaths" of all of their countries, as well as their graph. Needless to say, it shows different trends for each column compared to our dataset, because they had information from different countries. Their group of 5 countries has only 1 country in common with our group of countries – Spain- and even so, their graph shows the mean of total deaths due to COIVD19, whereas we calculated the mean of New Cases detected.

The procedure to create the flask server, the API and the Python module was a great learning experience and could be completed in person. Bearing in mind we've used local IPs from other people's computers. It wouldn't be possible to access remotely from a different network area.

-Considering the actual news and omnipresent topic of COVID19, data analysis was interesting and real. Looking at the situation of each country, we got an idea of how they handle the issue.

- Argentina: The mean of cases detected in 1 day: 1,970 for the whole period, while in Aug 25th the maximum peak was detected as 14,065. The alarm state started on march 18th and keeps going. They are partially locked down and unable to resume their regular activity. Still, during that period and knowing it's wintertime, they experienced the highest contagion rate. In this case, the alarm state could show a negative correlation with the number of new cases detected, however it is due to many other factors that are extremely difficult to consider in this analysis.

- Chile: The mean of cases detected per day is 2,150, and Chile has been showing a regular trend, except for Jun 18th when the surprising peak of 36,179 cases destabilized both the graph and the news. It is definitely the farthest outlier and as such, is not considered in the regular trend. It could have been a "testing error" or data manipulation. That's another factor that's not within our power to control. They're alarm state has been declared from March 30th to July 8th, and the evolution of cases has shown a regular curve.
- Colombia: The mean of cases per day is 3,095. Colombia's highest number of cases was 13,056, detected on Aug 20th. The growing rate of their graph is higher as the slope is always increasing with time. Their alarm state started May 30th and it's still going nowadays.
- Russia: This country showed the highest number of cases, however it also has the highest population, so it's not representative unless we take a percentage rate. The mean detected per day in Russia is 5,190 cases. Even though they have a huge contagion rate, their peak was not as high as in Chile or Colombia. It was detected on July 18th and there were 12,640 cases. However, the number plummeted dramatically the following day. So it can be said that there was a "testing issue" or a mistake in the information. In general, Russia had a regular evolution of new cases and deaths per day with just that specific outlier. They have a usual pattern of cases and recoveries, keeping deaths rate to the minimum. They have stated a short alarm state from the end of March until April 4th. In spite of augmenting their cases during that period, they have seen the maximum by mid May, and then started to slow down and decreasing the cases.
- Spain: In Spain the average of cases per day is 2,285, showing the peak in Aug 24th = 19.382. Nevertheless, Spain had drastically hiked up the cases the first weeks of March. The number of cases was amplifying in a way that it was very hard to assist. Similarly, the number of deaths. Spain has gone through a very tough period and declared alarm state on March 13th. After April, the situation subsided and a relieving stage came up. The slope of the graph plummeted abruptly and the alarm state finished. Notwithstanding, the maximum peak is shown as in August 24th, due to the fact that the contagion rate picked up again and started to expand the virus. Nevertheless the deaths are not as high as they were in March/April and the alarm state has not come back.

	Total NEW CASES	TOTAL DEATHS
Argentina	370,175	7,839
Chile	402,365	10,990
Colombia	572,270	18184
Spain	419,849	28,971
Russia	970,865	16,683

Total reported cases of COVID19 in our 5 countries: 2,735,524

Total reported deaths of COVID19 in our 5 countries: 82,667

Mortality percentage: 0,03%

The alarm state shows a little improvement on the daily-infected rate, at the beginning, but it's not very consistent and it had different effects in each county. If we had to mention a correlation: it would be negative, nevertheless we lack of information to draw this conclusion.
When the alarm state starts: Cases are at their highest. Then they start decreasing as the alarm state is carried out. However it is hard to unify the situation since each country had a different behaviour. In some cases, they tend to stabilize, the alarm state is released and cases start to climb up again.

What scares the most is the lock down, not the virus itself. It is shown that the mortality rate is just 0,03% so the best way to "beat it" is by strengthening the immune system of the whole population.