TheDataOpen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the 2021 Citadel Boston Regional Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

In the last days of the year 2019, the World Health Organization (WHO) was alerted of several cases of pneumonia in patients living in Wuhan, a province of China. Those cases were especially concerning because the virus causing the infection did not match any other virus ever known by humankind. The virus, now known as the 2019 Novel Coronavirus (or COVID-19), spread rapidly in China and the rest of the world and in a matter of weeks was characterized by WHO as a pandemic due to the rapid increase in the number of cases.

COVID-19 is an infectious disease spread primarily through droplets generated when an infected person coughs, sneezes, or speaks. People can become infected by touching a contaminated surface and then touching their eyes, nose or mouth before washing their hands. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment; however, the virus impacts the elderly and those with pre-existing health conditions most severely.

China was the very first country to impose severe restrictions to its citizens, like lockdowns, aiming to reduce the contact between people. The use of face masks, working from home whenever possible, closure of non-essential businesses, and social distancing (staying 6+ feet apart from other people) were also advised by many countries in an attempt to reduce contagion. After China, other countries started to follow these guidelines as they noticed the increase in cases; however, much of Europe only took firm action once the continent had become the epicenter of the pandemic. Italy was the second country after China to go into lockdown, but only after the virus had already spread across the north of the country. The same late response was observed in North and South America, resulting in countries like Brazil and the United States to be among the most affected by the pandemic.

With the rapid spread of the disease, there also came disinformation, fear, and anger. Disinformation and misinformation about COVID-19 was quickly and widely disseminated across the Internet, reaching and potentially influencing many people. Individuals protesting the social

distancing rules, the closure of commerce, and the mandatory use of masks most likely contributed to the number of new cases. As effective new treatments and vaccines become available, disinformation could further hinder uptake and jeopardize countries' efforts to overcome the COVID-19 pandemic.

Your Task

Your goal is to use COVID-related data in order to discover and analyze patterns related to the spread of the disease. More broadly, you should aim to highlight any discovery or hidden patterns in the data that could be used to help prevent/fight the disease. As you are looking through the data and preparing your analysis, please remember that the data on confirmed cases only becomes meaningful when interpreted in light of how many tests have been done.

We have curated COVID data at three different levels (and from three different sources). The first dataset is from "Our World in Data" and it contains information about the daily change in cases for countries worldwide. The second collection of datasets is from the European Centre for Disease Prevention and Control (ECDC) and contains data for a selected number of European countries. The granularity of the data in this group varies - some of the data files contain rows for each day per country, others for day and region, and some at country-per-week level. The third dataset comes from the covid tracking project and contains data related to the United States at two levels: country and state. The country-level file is being provided as well as the file for the state of Alabama (first one alphabetically) as an example. If you wish to pursue any state-level analysis, all the other 49 files can be easily be found at the projects data download page.

If you would like to enhance your analysis, feel free to use any other dataset you may find – just follow the guidelines in the "Additional Datasets" section below and be aware that the quality of your analysis will also be judged by the reliability of the data being used.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is valued over breadth of the question posed**.

Submissions may be predictive, using machine learning and/or time series analysis to investigate your research topic. Submissions may also be illuminating, through the use of data visualizations or through sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigour, which are far more important.**

Sample Question 1: How did the specific response measures adopted by the countries in Europe affected the number of new cases and\or deaths? Was any specific measure (like lockdown or closing of the schools) particularly efficient?

Sample Question 2: How do the number in the US compare with the one in Europe? Has the US used any lessons learned from their European counterparts considering that the epidemy manifested itself several weeks earlier in Europe?

Sample Question 3: How does the COVID evolution differs within each state in the US, each country in Europe or each region in the European countries? Can you find any particular outline in the behaviour, for example, was Sweden's well-known coronavirus strategy of not enforcing lockdowns since the beginning, a good one?

Sample Question 4: Can you build a time series model (or any other model) to predict how the disease will behave in the upcoming weeks? Can this model be applied globally or is it a country (or even stat\region) model necessary due to the many differences in the way the pandemic is being approached?

Sample Question 5: How reliable is the data being presented? Are there discrepancies on the many sources used by ECDC dataset for example (or any other dataset of your choosing)? Is the data collection process valid?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive.

The datasets are provided in three groups based on the source: "Our World in Data" (1_owid), the "European Centre for Disease Prevention and Control" (2_ecdc) and the "Covid Tracking Project" (3_covidtracking). Your team should only use the datasets that are relevant to your chosen question/topic. In some cases, the source of the datasets and information on how they were built is noted in case you want to enhance or recreate the datasets.

1_owid/owid-covid-data
Worldwide covid-related statistics
58,154 rows & 52 columns. Size: 19MB. Source: ourworldindata.

2_ecdc/notification
14-day notification rate of newly reported COVID-19 cases per country/week
18,896 rows & 10 columns. Size: 2MB. Source: ECDC.

2_ecdc/dailynotificationeu
14-day notification rate of newly reported COVID-19 cases per region/day
173,347 rows & 6 columns. Size: 17MB. Source: ECDC.

2_ecdc/weeklynotificationeu
14-day notification rate of newly reported COVID-19 cases per region/week
15269 rows & 6 columns. Size: 2MB. Source: ECDC.

2_ecdc/admissionrates
Hospitalization and Intensive Care Unit (ICU) admission rates per day/country
13,469 rows & 7 columns. Size: 2MB. Source: ECDC.

2_ecdc/testing
Testing volume for COVID-19 by week and country
1,311 rows & 9 columns. Size: <1MB. Source: ECDC.

2_ecdc/country_response_measures
National response measures over time
1,339 rows & 4 columns. Size: <1MB. Source: ECDC.

2_ecdc/agerangenotificationeu
14-day notification rate of newly reported COVID-19 per age group/week/country
7,008 rows & 8 columns. Size: <1MB. Source: ECDC.

3_covidtracking/national-history
Daily data on the COVID-19 pandemic for the US at country level
358 rows & 18 columns. Size: <1MB. Source: covidtracking.

3_covidtracking/STATENAME-history
Daily data on the COVID-19 pandemic for all the states in the US. State of Alabama being provided as a sample
306 rows & 42 columns. Size: <1MB. Source: covidtracking.

Additional Datasets

Participants are welcome to scour the Web for their own custom datasets to supplement their analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's technical product team via Slack if you believe your idea is worthy of an exception).

Other Materials

We will provide you the schema for each of the data tables in another packet.

Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

The competition will have multiple rounds of evaluation. Your Report will be judged as follows:

- **Non-Technical Executive Summary**
    - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
- **Technical Exposition**
    - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
    - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?

- *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 5:00PM EST on Sunday, March 7th, 2021. Any submissions received after that time will NOT be evaluated by the judges**.

Tips & Recommendations

This will be a weeklong event, however, you should try to complete as much of your work as possible before the weekend. The extra time may lull you into a false sense of security. Additionally, with your extra time, you should really think about what problem you want to solve. The outcome of this Datathon for you will likely be decided by how well you planned your work.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard “terminal + text editor” environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 4 hours on your report to ensure strong communication through visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile

Ask for Help

Correlation One's technical product team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.