# Part 1 - Common Analysis

## The Course Project

The course project is broken into four parts:

- Part 1 - Common Analysis sets the stage for the subsequent assignments. In A4 you conduct a base analysis. All of the students in the class will conduct the same analysis, but with a slightly different data subset.
- Part 2 - Extension Plan will require you to ask a human centered data science question that extends the work in A4: Common Analysis.
- Part 3 - Presentation will require you to give a modified (shorter) PechaKucha presentation of your completed project.
- Part 4 - Project Repository, creation of a fully documented repository and also requires the submission of a written project report.

## Common Analysis

During the last three years we all have been experiencing a global pandemic. This has been tragic and disruptive to many countries and has taken a deep personal toll on many individuals and their families.

One aspect that has been hard to miss in the last three years is the datafication of the pandemic. That is, many aspects of the individual toll of the pandemic have been collected, aggregated and re-represented as data. This datafication gives us the privilege to examine the pandemic from potentially many different perspectives to understand how it has changed lives and how it has changed society. To be honest, we are actually at the very beginning of understanding and comprehending these impacts.

During this Course Project you are going to begin taking a look at some of the social aspects of the pandemic by conducting a human centered data science analysis of some available COVID-19 data. In Part 1- Common Analysis, every student in the course will work from the same datasets. Students will be assigned to analyze data for one specific County of the United States.

## Sharing and Collaboration is Allowed

For **PART 1 ONLY** all students in the class **MAY SHARE CODE SNIPPETS**, **STATISTICAL APPROACHES**, and **VISUALIZATION TECHNIQUES**.

**SNIPPETS are OK**. Students are **NOT ALLOWED** to share **THE SOLUTION**. That is, you may not share a specific coded application or collection of subroutines that comprises a mostly complete solution.

We are encouraging **SHARING** but we want sharing to include **COMPREHENSION** of the **METHOD, APPROACH, and TECHNIQUE**. Your mantra for sharing is "I can help you understand this, but I won't do this for you." When sharing a snippet, a statistic, or a technique, it is very helpful to explain what it does. One advantage is that you are all working on the same data with the same structure. The context of your explanation is relatively fixed for this assignment. You are all working in the same context.

When you borrow or reuse a code snippet, or a statistical approach, or some technique that was provided or outlined by one of your classmates, you should keep track of WHO provided it so you can make an appropriate **ATTRIBUTION** in your submission of Part 1.

## Step 0: Data acquisition

The common analysis research question will require several different datasets. You will need:

1. The **RAW_us_confirmed_cases.csv** file from the Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily. You can use any revision of this dataset posted after October 1, 2022.
2. The CDC dataset of masking mandates by county. Note that the CDC stopped collecting this policy information in September 2021.
3. The New York Times mask compliance survey data.

The majority of this data is by US County by Day. The mask compliance is a single shot estimator that gives you a compliance estimate for every County in the US. You should carefully review the data descriptions that accompany these datasets. They each have some interesting caveats. As well, some of them are explicit with regard to the way you should interpret missing data.

Lastly, you have been assigned a specific US County for analysis. You are **NOT** analyzing the entire dataset. You have been assigned one US County that forms the basis for your individual analysis. You can find your individual US County assignment from this Google spreadsheet.

## Step 1: Analyze

The common research question that you are to answer is:

- How did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021?

Answering this question can be a little tricky - and it will be useful for you all (whole class) to discuss this on Slack. Some of the issues that you might consider when conducting your analysis include:

1. What needs to be cleaned and standardized over the three datasets?
2. There is a delay between the time of infection and the time a case is confirmed. Many factors may contribute to such delay. People may not show symptoms right away after infection. It may take a few days for the testing results to become available especially during the early period of the pandemic. Should we model the delay?
3. Masking may simply make it longer to get infected or it may prevent some percentage of infection. How should we consider the effect of a mask?
4. The research question is about how a time series changes. The infection time series is a set of slopes. Therefore the question is about a derivative function. That is, you want to answer a question about the change in slope over time. How can we test the difference in the derivative function?
5. Masking survey data shows probability of compliance in several categories. How can we model different proportions for population compliance?
6. Masking policies varied in their implementation (e.g., size of “crowd” required, different situations, restaurants, bars, clubs ...). How should I handle things when my County implemented two different policies at different times?
7. The County I was assigned did not implement a masking policy! What is a reasonable way to answer this question? That is, how might I model “voluntary” masking?
8. Vaccinations probably impacted the apparent effectiveness of masks. How should we account for different vaccination rates in different populations within the same County?

We note that we did not enumerate all potential issues that you may want to discuss. Further, there are some better and worse ways to handle these questions. We are not looking for the one right answer. We are looking for reasonable solutions. There are aspects of this problem that are very hard to model - and so you will probably want to make simplifying assumptions. For example, you might decide to ignore the impacts of vaccinations or consider pre-vaccine availability as one time series and post-vaccine availability as a totally different time series.

## Step 2: Visualize

In this step we want you to create a graph that visualizes how the course of the disease was changed by masking policies. For your county, you should create a time series showing the changes in the derivative function of the rate of infection. Your graph should indicate days where masking policies were in effect (or not) and whether the difference in the derivative

function was significant. Optionally, you can add a second time series that shows the actual rate of infection.

## Step 3: Write & Reflect

This step has two objectives.

First, for the visualization you created in Step 2, you should write up an explanation of the visualization. Some of the important things you might need to explain include: What does the figure show? How does the viewer "read" the figure? What are the axes, and what do they represent? What is the underlying data and how was it processed? You might think of this explanation as an extended figure caption. This explanation should be no more than one written page. Making a good effort now will make it easier to write your final report for Part 4.

Second, we would like to understand what you got out of the collaborative activities in this assignment. You should write a reflection statement that highlights one or two specific things that you learned from answering the research question posed in this assignment. How did the possibility of collaboration on this Part help, hinder, or change your thinking about the problem? Your reflection statement should include specific attributions for any/all code, methods and techniques that you reused. Your reflection statement should be no more than two written pages.

## Step 4: Submission

The submission for Course Project Part 1 will be four items:

1. A link to a snapshot of your current repository that should minimally include: (a) code/notebook, (b) an appropriate readme including information on the data used for this part of the project, (c) a license file, and (d) a .jpeg, .jpg or .png file of your visualization of your analysis for your assigned county. This should be a stable copy of your repository that can be used for grading.
2. Submit an explanation of your visualization
3. Submit a reflection statement on what you got out of the collaborative assignment

Note, all linked documents should be set to be publicly viewable so that the instructional staff may view and grade your assignment.

Also, we are not asking for an extensive write-up of everything you did in this part. Course Project Part 4 requires that you submit a full write-up of this part as well as the work you do for your Extension (Part 2). If you start drafting the relevant sections of your report now, it will be easier to complete the Course Project.