**NTRODUCTION:**

You have a choice of two datasets for the final project.

1. The first dataset is COVID related and the observations are the states (and DC). There are 45

variables (columns) that are mostly related to the pandemic, including the number of cases, whether

they have mandatory quarantine for travelers etc, but also include the prevalence of depression, the

air quality index and others.

2. The second dataset is from the World Bank. The rows are countries, and for each country we have

various data, some of them straightforward (GDP) and others which are not so clear (number of

endangered bird species). There are 37 variables.

You are **not** expected to use *all* the variables in your analysis. We have provided more variables than you

will use and you can pick the variables according to the story you want to investigate and analyze.

**PROJECT COMPONENTS:**

• Introduction: Background, and describe the questions that you are asking, with some background

• Exploratory Data Analysis: Variable descriptions (categorical or quantitative, what the variable

measures, etc.), visualization, summary graphics, and summary statistics. In this part, we expect each

member of your group to contribute at least one graph (the caption should include the name of the

person who created the graph.

• Data Analysis and Inference: Any inference you perform - including regression, hypothesis tests,

parameter estimation. Discussion of why you chose your methods and plots. We would like you to

include at least one regression and one hypothesis test in your analysis.

• Conclusion: What story does your data tell you? Provide the results of your final analysis (not all of the

intermediate steps). Support your analysis with plots and tables. Discuss limitations and draw

conclusions from your analysis. Compare results to other published work (if desired; and please cite all

work used).

• References, if you looked at any other sources. This is not mandatory, but may help you think of

questions to ask.

**HOW TO WORK ON YOUR PROJECT:**

• Discuss which dataset you want to work with (if you disagree, you may want to switch groups).

• Intro: Brainstorm as a group about what you are curious about in the dataset. Take notes, the intro can contain questions that you don’t end up focusing on. You’ll probably write a draft as you meet and edit more carefully when you’ve pretty much completed the project. It might be good to assign group

members to look at the data with particular questions in mind.

• **EDA:** **As individuals**, look at plots, summary stats, think about questions you brainstormed about in

the intro but be open to new questions or comparisons.

• **Data Analysis**: (Hopefully) after everyone has completed EDA, get together and discuss. For some

questions that have produced promising EDA plots, consider a regression line, hypothesis testing, and

confidence intervals as ways to use what you’ve learned. Every project should have at least one

hypothesis test (whichever kind(s) are appropriate for your analysis) and a linear regression, and this

may mean more plots in this section. Between the EDA and more formal data analysis, **everyone** in

the group should be responsible for **at least one** plot along with its explanation.

**• Conclusion:** Together you should reflect on what you learned and think about a coherent story to be

told throughout the steps of the project. This can include some dead ends that you thought would be

interesting but turned out not to be.

**• References:** Along the way you should poke around online and see what other people have done with

data like this. Refer to anything that you thought was interesting and/or influenced what you did.

**GUIDELINES:**

• Avoid describing or reproducing your code in the report. We can always look at your RMarkdown file,

and we will be knitting it to html. (You will lose points if it doesn’t compile.)

• Your conclusion should include a concise description of your findings and provide supporting

evidence of any claims that you make. You will tell the story of your data in this section.

• Numbers in tables should have only the significant digits needed to establish your point.

• Use the present tense. Use action verbs rather than passive verbs.

• Refer to the rubric (which will be posted) for further details.

**WHAT YOU WILL SUBMIT:**

• An report (in pdf or html) This is a guideline. It’s difficult to imagine a successful report with

fewer than 5 pages, and if you’re getting much above 12 you should consider paring it down. We’re

grading on quality, not quantity, so if you find that you’re just repeating the same plots and analysis,

you should leave it out.

• Your R markdown file

• We’ll give more specific submission requirements as we get closer to the due date.

* I am planning to work on prevention measures i.e Non Essential Business Closures (column J) that the US government has taken in July 2020 and October 2020 and trying to find out which is the most effective based on the total number of cases in June and October (column AO and AP).
* There are like 6 different values in column J and I want to see how they impacted the total number of new cases in column AO and AP.
* For each value in J, I need to compare the values in June and October. Probably like a graph wherein we have both the values shown in one graph.
* Finally conclude with which of the 6 was proven to be most effective in stopping the spread of covid-19(i.e the least number of new cases).
* I am not sure what other analysis can be done, please do whatever can be done and give it to me in a doc file and also a R markdown file with detailed comments.