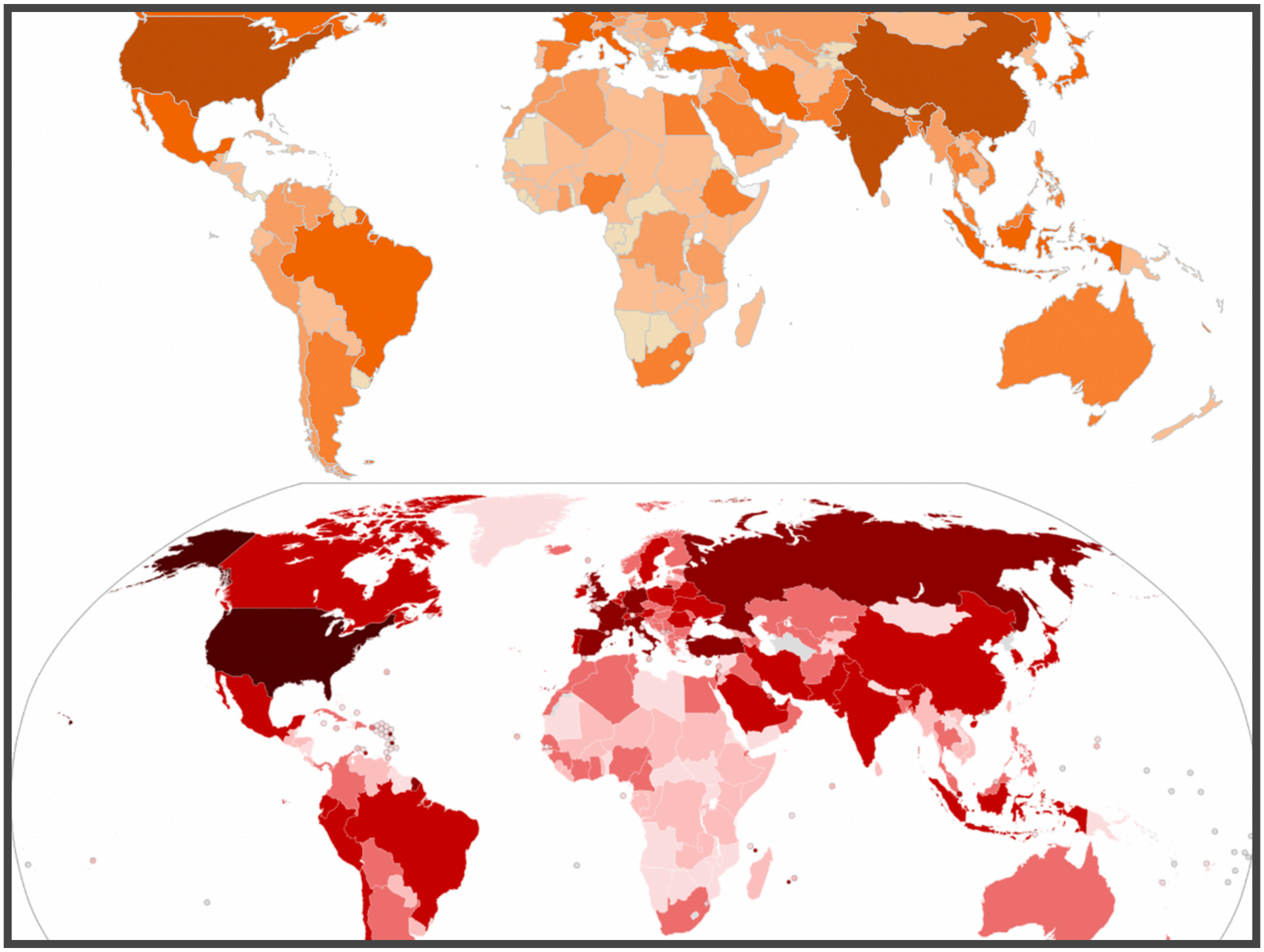
Christopher Schutte

INFO 2201

May, 2020

# climate and covid



## Initial Approach:

The initial goal was to use two datasets, one from https://www.quandl.com/tools/api that had JSON files for the DOWJ Industrial Average and the COVID-19 dataset from https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset in CSV format to compare the outbreak to the economic downturns and see if a correlation could be found. This was going to be accomplished like this:

- Import requests, pprint, pandas, numpy, and seaborn
- Use request.get to pull a json file out of Quandl
- Use pprint to isolate the DOWJ information by day and create a dictionary of day:value tuples
- Populate a new CSV file with this new DOWJ dictionary starting in Jan, 2020 and ending in April, 2020
- Download the CSV files for COVID from kaggle
- Use pandas to sort through the information and isolate infection numbers and days
- Import the DOWJ CSV I created and use pandas to open and parse through to make sure its day metrics are comparable with the other dataset
- Use numpy and seaborn to create graphs detailing infection rates over time vs the DOWJ
- Write and algorithm to isolate the events when an uptick in infections was followed by an event in the DOWJ
- Do the same for the world infection rate data and see if there is any correlation
- Throw this information and these new CSV’s into Numbers on Mac to make really nice data visualizations

## Problems and Re-planning:

Although Quandl had great and fairly extensive free datasets, the one I needed which is the DOWJ Industrial Average from Jan - April, 2020 was premium membership only and I didn’t want to pay 60 dollars for something I intended to use once. So, I have now changed the plan to have a similar data comparison, but use an API I am more familiar with that I barely got into during HW 2.

## The Actual Plan:

- Import requests, pprint, pandas, numpy, and seaborn
- Use request.get to pull a json file out of the Global Footprint Network
- Use PPrint to find where the data is stored in the GFN for Countries, Year, and Carbon Emissions and write some code to extract Carbon Emissions for the last 30 years in ten year increments for the top 50 countries by GDP (I'll just use the wikipedia GDP list)
- Format this information into a dictionary of Country: Carbon Footprint and write into a CSV
- Download the CSV files for COVID from kaggle
- Use pandas to sort through the information and isolate the most recent infection numbers and format this information into its own dataset with Country and Infections as the first two columns
- Use pandas to open the CSV with the Carbon Footprint data and turn that into a dataset as well, combine the datasets into a master dataset which has the columns Country Name/ Number of Infections / Carbon Footprint of Nation (Most recent year)
- Use seaborn to start to plot this information in different ways
- Create an Algorithm that finds the countries with the most reductions in Carbon Footprint in the last 30 years and see if those countries have higher or lower infection rates then countries with little, no, or increased Carbon Footprint change

## Sources:

https://www.footprintnetwork.org (a great api with extensive datasets on the carbon footprint of nearly every nation on earth dating back decades, the api is a little bit of a

pain to work with due to a data structure that relies on a template that looks a little crazy in PPrint, but the data is free and high quality)
https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset (free, updating daily, great formatting, already clean, this is a dream CSV to work with and SRK is a legend on Kaggle already which lends much credibility to the data)

## Reason for Libraries Used:

requests - I need to use an API that outputs JSON files
pprint - I need to be able to read those JSON files
numpy - This is the cornerstone of the entire CSV, plotting, and data organizing side
pandas - I am never using CSV_reader again
seaborn - Really excellent plotting library that its going to make the data analysis much easier

## Findings:

After creating a coefficient to demonstrate the relationship between infections and emissions, no clear pattern emerged. There was no bell curve and no areas of concentration worth paying attention to. Most countries had a difference between the two placements in the mid teens to mid twenties apart and that distance is too far to create any real evidence. Over all, even though there was no relationship there, the lack of findings is in itself a finding and the data did yield interesting insights info emissions, and COVID cases.