Final Project

**Topic: The impact of state government action on the severity of the COVID-19 pandemic in the United States**

**INTRODUCTION**

With the continued COVID-19 outbreak changing the fundamental nature of our how we conduct our lives, there has been great interest in the development of new approaches to expanding citizen understanding of the crisis and what actions are being taken. One group, COVID ACT NOW, has harnessed the collaborative efforts of multiple data scientists and engineers to build simple, easily digestible views for topline information on COVID at the state and county level. Inspired by their work and their approach to gathering the most up-to-date information on the crisis, I chose to use Python as my primary method to answer the following research questions:

* How many states have taken specific government action to limit COVID spread?
* What types of actions have been taken across the country (by % of total US)?
* Which government actions appear to be effective in reducing the rate of new cases?

**About the Data**

The first data set used for my analysis is a state-level tally of confirmed cases of and deaths related to COVID-19 collected by the New York Times from the beginning of the outbreak in the US in late January 2020 until the most recently recorded totals.

The second data set is **state\_intervention** – a data frame derived from information scraped from the web, structured to represent a dataset collected by AEI ([American Enterprise Institute](https://www.aei.org/covid-2019-action-tracker/)) on government actions taken in response to the pandemic.

I know that the data available from AEI is not immediately available in a format that can be ported into python directly for analysis. Rather than attempt to transform the accompanying .xls files for this data in a series of scheduled pulls, I created a program to scrape the embedded HTML code that builds their visualizations and pull the data in a more usable format (JSON). I consulted [techniques used](https://github.com/covid-projections/covid-projections/blob/develop/scripts/interventions-data.js) by the group “COVID Act Now” to convert this HTML into JSON.

**Program Description**

The script I developed does the following:

1. Scrapes the Web for HTML code used to build embedded visualizations on the AEI site.
2. Converts this HTML code into a ‘beautiful soup’ object and searched for tags specific to the embedded visualization
3. Uses RegEx parsing functions to extract a JSON object containing a nested list of states and information on government actions enforced or not taken.
4. Restructures the data to make states the relevant unit for my analysis
5. Converts “yes/no” fields in the data frame into a binary 0/1 format for easier aggregation
6. Merges this newly created data frame with the dataset of cases and deaths gathered by the New York Times
7. Calculates “new cases” and “new deaths” for each state over time
8. Groups the data into the following units of analysis:
   1. State
   2. Type of Government Action
9. Calculates fields for “new case rate” and “new death rate” at each unit of analysis, over time
10. Generates plots to answer the main research questions
    1. Map plot of States w/ current government action at the state level
    2. Bar plot by government action, showing % of Total US States employing the method
    3. Line plot by government action, showing rate of new cases and deaths over time

This program uses the following libraries: *json, pandas, urllib.request, beautifulsoup, plotlyexpress, us*  and *re*.

**DATA PREPROCESSING**

Though the data itself can also be readily available in csv form or by direct calls to public APIs, I was interested in utilizing an approach that gathered data from embedded visuals. That meant that before any analysis could be done, I needed to gather, restructure, and transform the data available.

Using the embedded visual found at this URL (<https://e.infogram.com/_/bo5pjUi7dprQAvs1l6oZ?src=embed>), I used the *urllib.request* library to scrape the HTML code that builds the linked site. I then converted this html object into a ‘beautiful soup’ object and searched for tags specific to the embedded visualization. After finding all ‘script’ objects, I used RegEx parsing to derive a JSON object, \*graphic\_json\*. Using a series of nested for loops, I was finally able to extract the list of states and their associated interventions.

At this stage, the remaining steps were to convert this list of states into a pandas data frame, \*state\_intervention\*, add state as the index for the data frame, and convert the strings in each row of the data frame into a 0/1 binary field with column labels associated with each type of government intervention.

The final product of this transformation produces the following dataset:

|  |  |  |
| --- | --- | --- |
| **Column** | **Type** | **Definition** |
| State | Index | Name of the US State |
| Latlon | Object | Latitude and Longitude of given state |
| State\_emp\_travel | Object | Binary indicator for whether official travel was restricted for state employees |
| School\_closure | Object | Indicator whether schools are closed |
| National\_guard | Object | Indicator whether National Guard was activated |
| Ne\_business\_closed | Object | Indicator whether non-essential businesses are closed |
| Bar\_res\_closed | Object | Indicator whether bars or restaurants are closed (outside of carryout) |
| Curfew | Object | Indicator whether a curfew is in effect |
| Gather\_limit | Object | Limit of gathering sizes |
| Hospital\_cap\_inc | Object | Indicator whether hospital capacity has been increased |
| Relax\_licensure | Object | Indicator whether state relaxed medical licensure |
| Mand\_vis\_quar | Object | Indicator whether visitors have mandatory quarantine |
| Postponed\_primary | Object | Indicator whether primaries have been postponed |
| Stay\_at\_home | Object | Indicator whether a ‘stay at home’ order has been issued |
| Elective\_surgery\_post | Object | Indicator whether elective surgeries have been postponed |

The next step in my data processing was to merge this newly created data frame with a dataset of cases and deaths at the state level gathered by the New York Times. I was able to access this dataset directly from a public repository on github I found from Kaggle. After conversion of this data from directly from the web using the pandas read\_csv function and a quick transformation of the ‘date’ field to a date object, I produced this dataset:

|  |  |  |
| --- | --- | --- |
| Fields | Description | Example |
| Date | Date case or death was recorded | 2020-02-02 |
| County | US County Name | Sublette |
| State | US State/Territory Name | Wyoming |
| Fips | State/County FIPS code—used to join other geographic data if necessary | 53061 |
| Cases | # of recorded cases for the county on the given date | 64 |
| Deaths | Number of recorded deaths | 12 |

I then used the concatenation function from *pandas* to combine my datasets into one large data frame using all these fields—the final step to making this data ready for analysis.

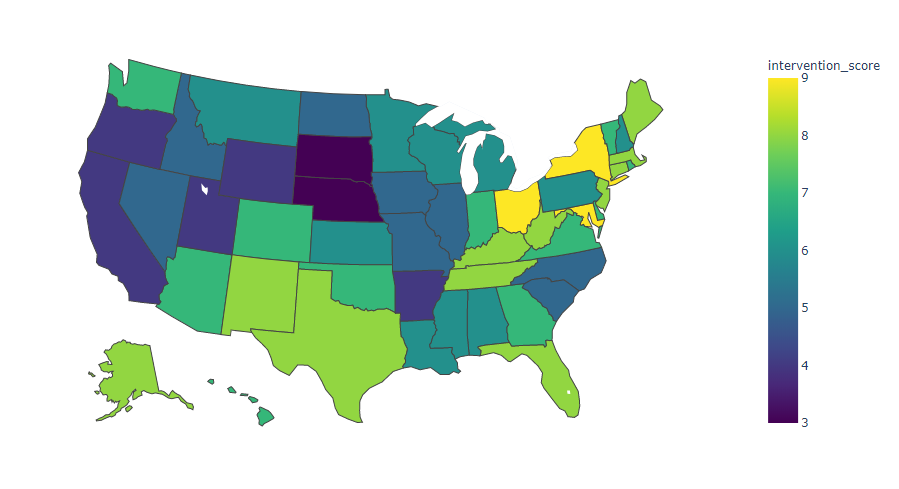
**DATA ANALYSIS**

1. **Which states have taken specific government action to limit COVID spread?**

To answer this question, I focused only on my first generated datasets of states and their government actions, grouping my data by state and calculating a sum of government action to create a new ‘score’ for each state where a higher score was associated with a higher level of government action. Using this grouped data, I then used the *plotly* library to generate a choropleth map of US States by score.

**RESULTS**

Presented as a map, here is what I found:

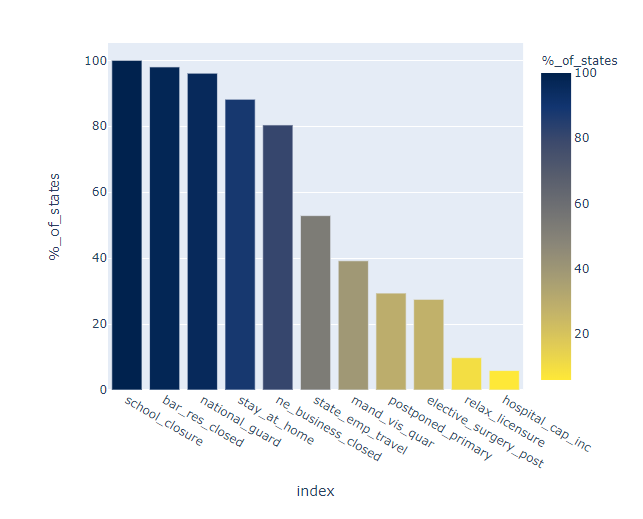


We learn from this map that relatively few states have implemented what we’d call a “high” level of intervention in comparison to other states, but the ones that did coincide with areas of higher population and outbreak – New York and Maryland/DC area. What is surprising is to see Texas, Florida, and Tennessee higher on this list. That’s likely because the \*severity\* of these restrictions was more limited in these states, and that these actions have already since been significantly relaxed.

1. **What types of actions have been taken across the country (by % of total US)?**

I answered this question by first generating column sums for each government intervention, and second by dividing these column sums by the total number of states (+ District of Columbia) in the dataset (51). Using plotly I produced a bar chart that indicated the % of total states employing each intervention:

**RESULTS**



This bar plot shows that although all states initiated school closures, most closed bars and restaurants and deployed the national guard, the percentage that issued a stay-at-home order drops to 88%, and only 80% closed non-essential businesses. The least-practiced restrictions included an ordered increase of hospital capacity, relaxed licensure requirements for healthcare practitioners, postponed primaries, and mandatory quarantine of visitors to the state.

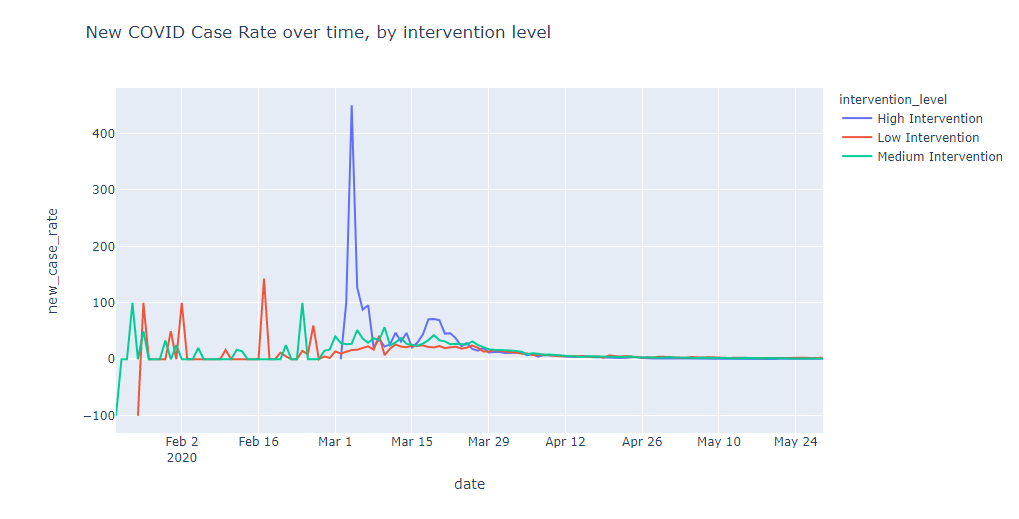
1. **Do government actions appear to be effective in reducing the rate of new cases and deaths?**

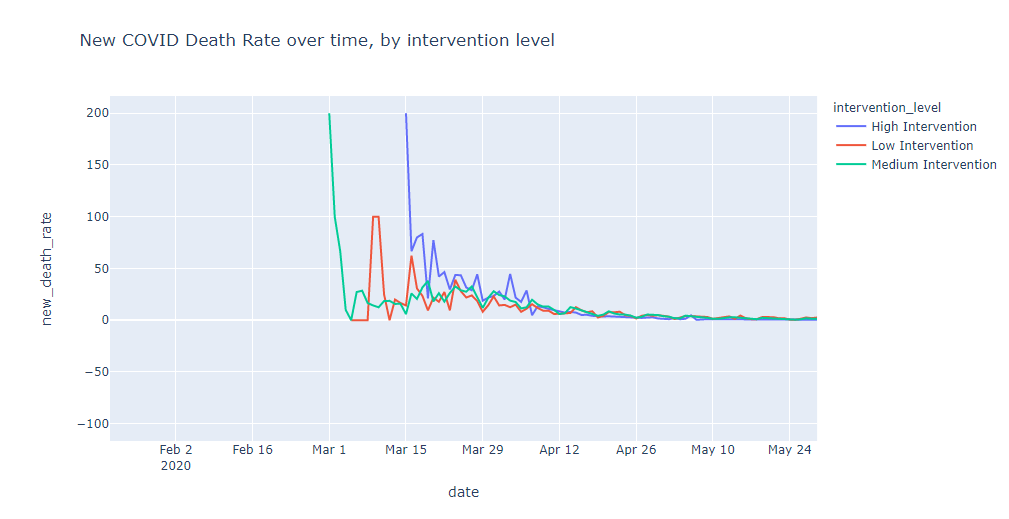
To answer this question, I needed to create some new fields. First, I created calculated fields for “new cases” and “new deaths” for each state over time. I then used a similar approach from question 1 and calculated an “intervention score” for each state. This time however, I grouped the time series dataset by these scores to produce the following units of analysis in a field \*intervention\_level\*:

* No/Low Intervention (Score of 0 – 4)
* Medium Intervention (5 – 8)
* High Intervention (9 – 12)

Using a data frame of state and intervention level, I joined the time series data with new cases and deaths. This allowed me to create calculated fields for “new case rate” and “new death rate” for each intervention level, over time. With one more use of *plotly*, I finally produced two charts, one for new case rate by intervention level, and one for new death rate.

**RESULTS**





We learned from these views over time that

* There are some potential errors in the earliest collected data, showing dramatic decreases in new cases in the early stages—however these level out quickly as the volume of cases increases.
* States with higher intervention saw the sharpest jumps in new cases, and the highest jumps in new deaths.
* States with medium interventions initially saw large increases in new deaths, and remained higher than states with low intervention
* States with low intervention saw the smallest variation in new cases, but saw more frequent jumps in new deaths, especially in March.

**FINAL CONCLUSION**

By analyzing these two datasets at various levels of analysis and different views, we were able to better understand the following about the COVID-19 crisis:

* There was wide geographic variation in the severity of interventions taken at the state level to combat COVID-19
* Of the interventions taken, School closures, restaurant/bar closures, and deployment of the national guard were taken by virtually all states.
* Though the rate of new cases and deaths saw significant variation in the early stages of the crisis (especially among states with high levels of state action taken), at this stage the rate of new cases and deaths have largely converged.

From these views alone, I don’t necessarily conclude that a lower level of state government action is associated strongly with a dramatic increase or decrease in cases and deaths when compared to states with more severe action. The next step in this analysis would be to quantify the true *impact* of these interventions, which could likely be done by including movement data provided by cellphone service providers and developing a measure for said impact. I also believe it’s still too early to determine how many deaths can be attributed to actions taken or lifted more recently--this initial analysis opens the door to a revisited look when CDC counts are updated in upcoming months.

This project was a valuable exercise in demonstrating that Python is a powerful tool for collecting data directly from the internet, and in a format that can be later revisited without significant rebuild. I look forward to exploring these trends further using what I’ve learned in this class.