CS 5010: Final Project

Aubrey Brockmiller

Mike Wetklow

Bryan DeMarcy

Hemani Choksi

**Introduction**

It is no secret that 2020 has been shaken by the effects of a global pandemic. COVID-19 knows no gender, ethnicity, or race. Not a single person will forget the impacts that the disease has had. Unfortunately, the effects the COVID-19 are far from over. Everyday activities are now being questioned. Is it safe to send my child to school? Can I visit my grandma for Thanksgiving? Will eating at my favorite restaurant put me at risk? Our country, in particular, is struggling to find relief. To answer some of these questions for the future, we must look at what has been done in the past.

For our final project, our group decided to look at where the money is being allocated in the United States. We looked at allocations by state so the data would be organized. We then wanted to see how the money allocations compared to the current COVID rates. This included normalizing the money and COIVD rates by population of each state. Our hope was to be able to summarize information about each state’s money allocation, COVID rates, political affiliation, and population in search of a pattern. To find these patterns, we looked at a series of visualizations as well as data queries. We asked questions such as: is there an amount per person that seems sufficient for COVID related funding? Have any states been successful in bringing their COVID rates down? Most importantly, our group sought to find some hope in the data. In the end, our country will overcome this pandemic. We believe that the actions already taken have a story to tell that can lead to a brighter, COVID free, future. The key to unlocking that story is by connecting the correct datapoints.

**The Data**

There is no shortage of publicly available COVID-19 related data. From small businesses, to government officials, to a mom who is just trying to send her kids to school- everyone has something to gain from analysis on these sets. One downside to so much available data is finding the right set with the right combination of variables. Our group was not able to find just one set that encompassed the variables we wanted to look at. So instead, we compiled key columns from five different datasets to perform our analysis. The sets are as follows:

* Census population data: <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>
* US Governors: <https://en.m.wikipedia.org/wiki/List_of_United_States_governors>
* State Abbreviations: <https://worldpopulationreview.com/states/state-abbreviations>
* COVID-19 cases, testing, and hospitalizations from March through June 2020: <https://covidtracking.com/data/download>
* HHS spending data: https://taggs.hhs.gov/coronavirus

Each source provided a key variable that we felt contributed to answering our main questions. All data was extrapolated from reliable sources. Utilizing this number of sources required a fair amount of data wrangling and cleaning before we could begin any analysis. All sets were saved as downloadable .csv files from their respective sources. Each .csv file was read in and the chosen columns were pulled out. All datasets we utilized had their information broken into state. Therefore, we had to rename all the columns in the files to share a column name, such as ‘stateCode’. We were then able to join the dataframes by state name. Some of the data sets had extra rows for Washington D.C., US Territories, and various regions; however, we were only focusing on the 50 US States.

Our compilation and wrangling resulted in two final dataframes to be used for analysis. The first frame, ‘df\_merged’ contains summary values for the 50 states to include governor name, political affiliation, population, and total HHS funding received. The second frame, ‘covid’, contains the daily COVID-19 information for every state. We did not combine them into a single dataframe because the ‘df\_merged’ data only contained 50 rows for each state and would have been duplicated for every date in the ‘covid’ dataframe. This seemed excessive and could lead to errors when trying to aggregate the data later. However, we felt two dataframes were sufficient in information, but simplistic enough, considering the amount of source information we had to merge. Finally, we printed the dataframes into new .csv files for easy access, as well as an additional visual check on the merge before any analysis took place.

**Beyond Original Specification**

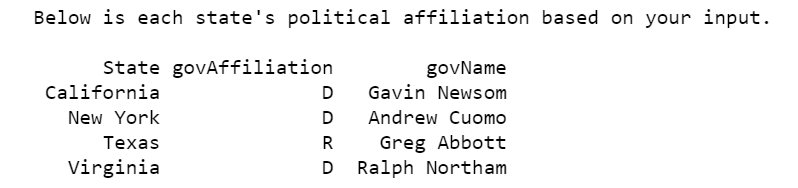
The original specifications of the project called for a single dataset that may have needed some pre-analysis cleaning. Our group went beyond this requirement and essentially made our own dataset. Again, we felt there was wealth of information on COVID-19 and the best way to utilize this information was to take the key variables in each set and form our own dataframes.

Additionally, our group sought to make this project user friendly and beneficial for the average citizen. Therefore, we added a user query were a user can input the states they are interested in finding more information about. This sparks a downstream effect that gives information about the input states as well as visual representations of the data for these states. These features give the user a lot of power and really provide customizable answers given by our code.

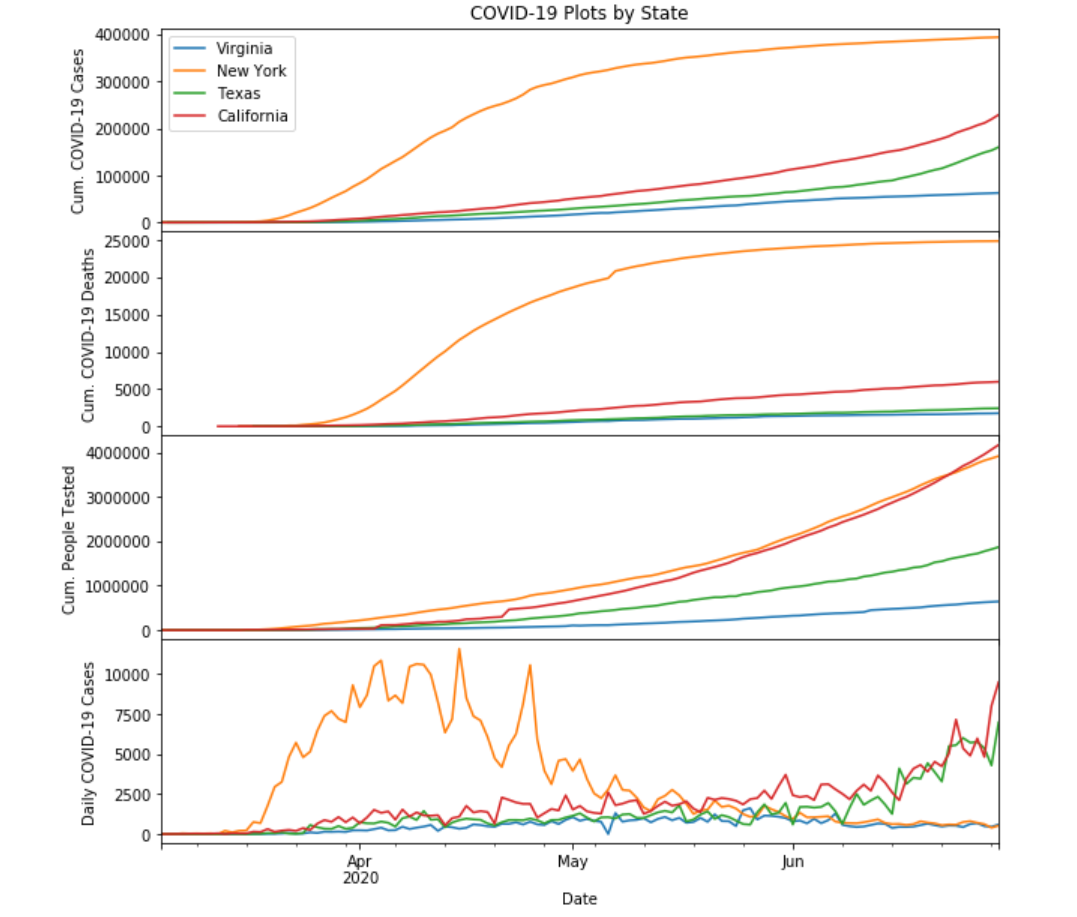
**Results**

Our analysis begins with a series of overall queries taken on the compiled dataframes to get an idea of what these sets could offer without any user input. We chose to look at more positive queries that could result from the data. First, we looked for states that received the most funding per person in that state. This showed Alaska, Vermont, and Wyoming received the most money per resident in the state. Contrarily, the second query showed the states with the least amount of funding per person: Florida, Minnesota, Ohio, Texas, and Virginia. The three queries at the end look at more positive outcomes from the set. Looking at the highest amount of recovered COVID cases, it is apparent that Massachusetts was the first state to hit a high amount of recovered cases (70,000 was the utilized threshold). As of late June, Texas and New York also reached this number. The next query looks at the lowest number of hospitalizations. The tail() function was used so the last date in the dataset was displayed. Thirteen states, including the three with the highest funding per person, have less than 100 people hospitalized as of June 30th. The last query prints the states that had no increases in positive COVID cases. Again, the end of this filter was printed to see the most recent dates.

Our analysis continues by giving the user the power to "Enter a list of state abbreviations separated by a space: ". We wanted the user to care about the information they were getting, and we felt like this allowed the user to input the state(s) that are most relevant to their lives. It also allows the code to be run multiple times and give a high amount of state comparison combinations. For our reporting purposes, let us say the user input “VA NY TX CA”. The code will then output a chart with the state name, the governor of that state, and the governor’s political affiliation.



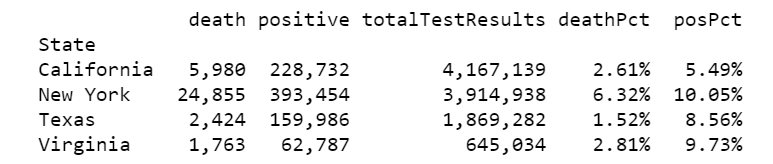
The code will also output a series of line charts. These four charts show the cumulative COVID-19 cases, cumulative COVID-19 deaths, cumulative people tested, and daily COVID-19 cases in each state from March 1, 2020 through June 30, 2020. This provides the user with an easy visual comparison of each of their input states. But it also allows them to view how each of these variables have changed over the course of the pandemic. An example of these charts is shown below.



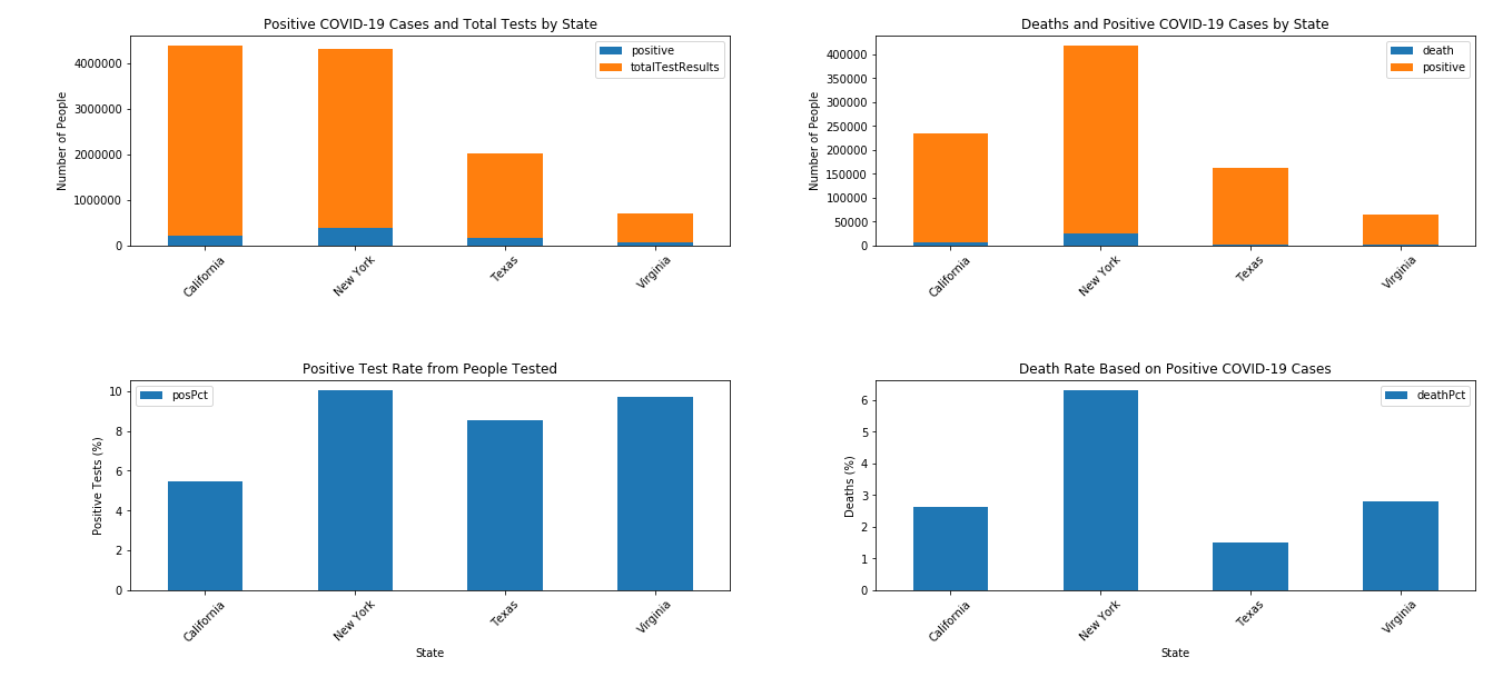
For this example, specifically, you can see how New York’s cases (shown in orange) were spiking and climbing until May. However, you can then see the cumulative cases flatline. And, supporting, you can see the daily cases drop as well. Looking at the blue line, representing Virginia, you can see a slow and steady increase in cumulative cases. But dropping to the bottom chart you can see the daily cases are also starting to decrease. However, both Texas and California seem to have a rising amount of daily cases.

Still based on the states that were input by the user, the code then creates a new dataframe with only the max death, number of tests, and positive case values. A series of calculations are then performed to calculate death percentage in people testing positive for COVID-19 within that state, calculate positive test percentage in people getting tested, and to calculate HHS funding per capita. These values are saved in the dataframe and displayed visually.

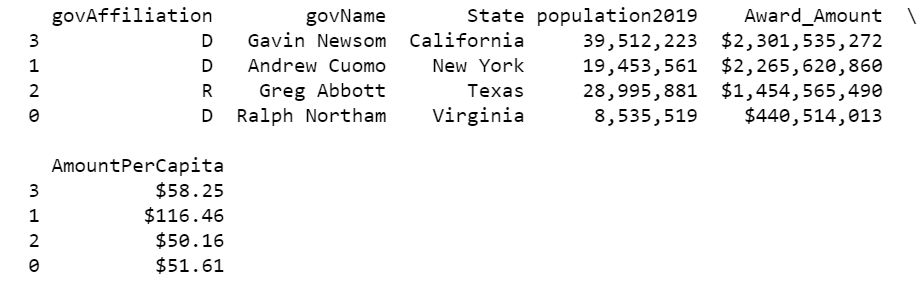
First a chart is output summarizing the calculations that are preformed and stored in the new datefame. The output can be seen below:



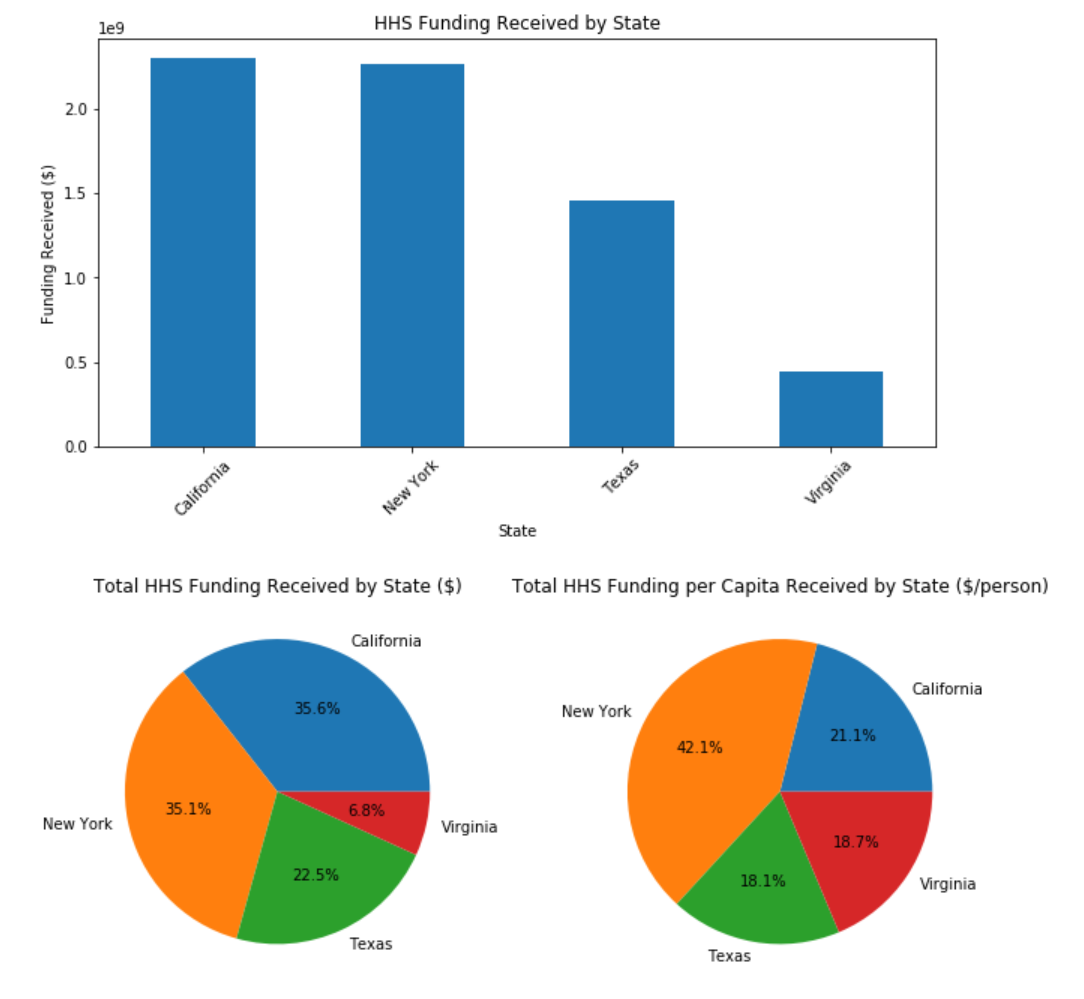
These calculations are also displayed using four bar charts. The first chart is a stacked bar chart showing the amount of positive cases in comparison to the number of tests performed in that state. The second chart shows the amount of deaths that have occurred out of those positive cases. It should be noted that even though these charts are beside each other, and the bars look to have similar heights, the y-axis uses different scales. The bottom two charts display the same information by utilizing percentages, again the y-axis scales should be noted.



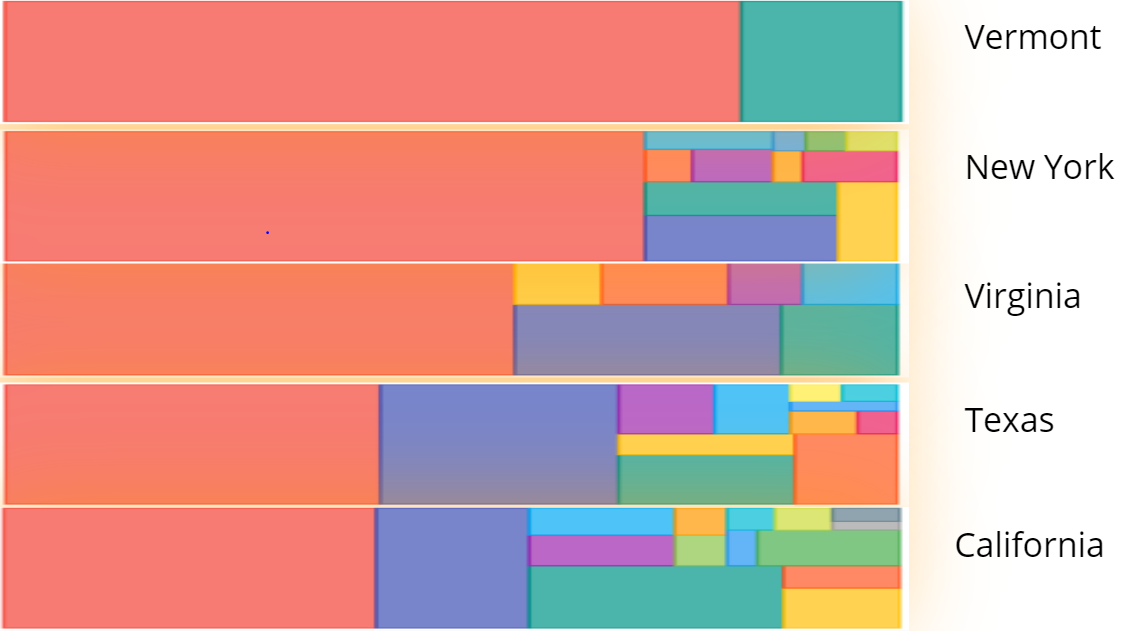
The next analysis section looks at the population size of each state and how much funding those states received. A summary chart is first output:



This again displays the governors name and political affiliation as well as the population of the state, amount of federal funding that state has received, and a calculation of how much money is essentially being allocated per person in each state. Visual representations of this information then follows. A bar chart summarizes how much each state is given. Then two pie charts show how much money is given before and after the population calculation is taken. We feel the difference in these charts can really tell an important story. It may seem like some states are receiving a ridiculous amount of money, but when taking the population into perspective the numbers quickly begin to add up. An example of these charts is on the following page.

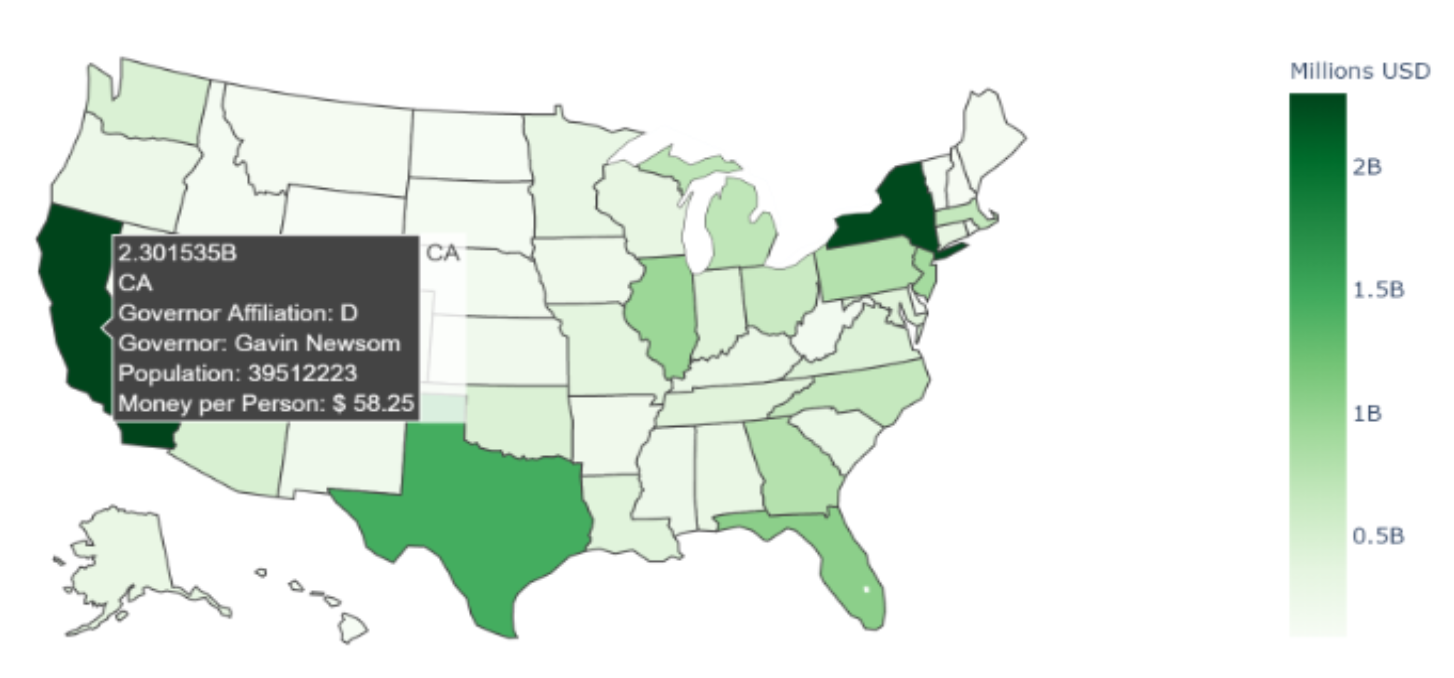


After comparing the funding by state, we decided to take a deeper look into the HHS funding with a series of queries as well as a treemap visual for key states (VA, NY, VT, CA, TX). A compilation of the five treemaps can be seen on the following page. It should be noted we only included spending avenues that were above 100 million dollars. Any spending sources below this threshold are not represented in the charts.



Spending Treemap charts were developed to represent data in the form of rectangles. The area or size of each rectangle corresponds to its numerical value. The chart is best read with a detailed legend displaying what the colors indicate. For all states it is apparent the red color is a large consumer of funds. This represents the Epidemiology and Laboratory Capacity for Infectious Diseases (ELC) program. It is the nation’s support system to state and local health agencies for general infectious disease threats.

As a final, overall representation of the compiled data, our group developed a choropleth map. This map is designed to be shaded a darker green in areas of high overall COVID-19 allocated funds. The map is also interactive. You can scroll over each state and a display box will show information such as the overall funds, governor name and affiliation, state population, and federal funding per person. We felt this chart was a good summary of our compiled dataframe. It can be seen on the next page.



**Testing**

The primary piece of output that required unit testing was the user input section. This is the only piece of code that had the potential to change because each user can supply a different state input. For this reason, we wrote a unit test that tested for a user supplying ‘all’ to make sure the code will run all 50 states and produce the related charts. We also supplied code to test for single and multiple states. Lastly, we tested if the user supplied no states, or an abbreviation that is not one of the 50 states abbreviations. Each time the user supplied an unsatisfactory input, the code will output an error and like the 50 states and abbreviations that are appropriate for the input.

**Conclusion**

Of course, numbers and graphs can only tell so much of the story. Without any broader contextual information, our project is simply a conversion of data into information that is displayed in pretty charts. Our group sought to take this information and turn it into usable knowledge. Over the course of our analysis there are some states that stuck out as particularly good or bad performers by our variable’s standards. We took some of these states and looked at why they performed the way they did in hopes of finding patterns or processes that could be used in the lower performing states.

Oddly enough, Vermont seemed to have a consistently low amount of COVID cases and a high amount of funding per person. We found the low cases to be interesting because many of the highest COVID-19 case states have been in the northeast- New York, Pennsylvania, etc. Looking thought the Vermont travel guide, it is evident that Vermont does not openly accept travelers from other states without severe restrictions (Vermont Department of Health, 2020). All travelers that do not reside in Vermont permanently and are entering the state for any reason other than essential work are required to quarantine in the state of Vermont for 14 days before utilizing any public facilities in the state. The entire state is also enacting a required mask policy. You can be ticketed for nor wearing a mask or maintaining a 6-feet social distance. As far as spending is concerned, Governor Phil Scott is currently fighting for control of the state’s funds granted by the CARES Act (Landen, 2020). So far, the money has been split into three appropriate fields- state funding to boost the economy, funding for health related COVID expenses, and a final pool left up for debate. With the health funds, Vermont has said to have stocked up on medical supplies and well as increased the pay for many essential workers. This money allocation is thought to be about $150 million dollars of the $1.25 billion the state has received. Because the money is granted from the government there is a series of approvals that seem to have to take place before the money can be used.

On the other hand, states such as Texas seem to be seeing an increase in their cases. As of May 21st, the required quarantine for incoming travelers to the state of Texas was lifted (Texas Department of Public Safety, 2020). People in the state are required to wear a mask in counties that have more than 20 positive COVID cases. How this number is communicated to the citizens and travelers within Texas is a bit hazy. The mask requirement does not seem to be enforced by police officers at this time. In fact, Texas is moving along on its phases to reinvigorate their state economy. Which begs the question, where is there federal funding going? As of July 15th, the first allocation of funds, totaling $7 million dollars was to be allocated for COVID related relief (Office of the Texas Governor, 2020). Texas has been granted about $1.45 billion dollars in funding. The publicly available information on this spending seems to be harder to find than a state like Vermont. Because Vermont is having such a hard time getting approval to spend the billions of dollars they have been granted by the federal government, it should not be surprising that Texas is encountering a similar problem. Vermont seems to be more successful at liquidating these funds as a faster rate than Texas.

So, our question stands: does federal spending and amount of money granted have a direct relationship with the COVID rates in these states? Our analysis seems to support this suggestion. But a different finding we did not originally anticipate is not just how the money is being spent, but rather can the money be spent fast enough. According to the National Treasury website, the CARES Act is meant to provide relief for any expenses incurred from March through December 2020 (US Treasury, 2020). It does not specify if those funds can continue to be spent after December of 2020 or where the money would go if it is not allocated. We originally believed our visuals, such as the treemap, represented spent money. But after further research, we discovered those funds have been budgeted for those spending avenues. Those numbers do not necessarily indicate the money has already been spent.

Additionally, social factors such as wearing masks and enforcing social distancing seems to have a large effect. States such as Texas and South Carolina may have been too quick to reopen their business doors in hopes of jump-starting their economy with the summer tourists. Speculation that the virus would spread slower in the summer months may have something to do with these state’s choices (Collins, 2020). Whatever it was, social practices seem to still play a big part in stopping the spread of the virus. In an article published by the NY Times, a random analysis of people wearing their masks was taken (Katz, 2020). Unsurprisingly, Vermont was purple; purple was indicative that most people are wearing their masks. Looking at a state like Texas or Florida, they covered the entire color spectrum. We did not do any analysis on social distancing or mask wearing, but the trends all seem to follow a similar pattern.

COVID-19 is not a simple problem to solve. There are so many factors that influence the spread of the virus, many of which no one is probably aware of. There is no hard and fast solution to this problem. However, our data analysis in combination with research and other external analysis paints a clear picture. Funding to restart the economy as well as spending for medical supplies is essential to finding relief. The granting of funds is only the beginning. Governors need to find ways to quickly allocate this money. States that have been able to finalize and spend some of this money are quickly seeing positive results such as lower hospitalizations and higher recovery rates. Arguably, the money per person in the state is irrelevant if the money is not being utilized. Such small amounts have been spent; it will be hard to assess which states received enough funding until a much later date. Additionally, social factors are not to be ignored. It may be difficult to quantify social distancing and mask wearing, but states with stricter travel restrictions are also seeing their numbers decline. Opening businesses too early may be tempting to boost the economy, but it is not worth the spike in cases that it results in.

Move over, our analysis shows there is hope. There are states that are not seeing much of an increase if an increase at all in COIVD-19 cases. Many states have figured out ways to begin to spend the mass amount of funds that have been dropped. We hope these states and their Governors would serve as an example to other states. The key to overcoming this tragic time may sound cheesy and cliché; but we will only overcome this pandemic by utilizing all resources provided and working together as one united front.

Citations:

Collins, Francis. (2020, June 2). *Will Warm Weather Slow Spread of Novel Coronavirus?*NIH Directors Blog. <https://directorsblog.nih.gov/2020/06/02/will-warm-weather-slow-spread-of-novel-coronavirus/>

Katz, Josh. (2020, July 17). *A Detailed Map of Who Is Wearing Masks in the U.S.* NY Times. <https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html>

Landen, Xander. (2020, April 27). *Scott seeks more control over $1.25 billion federal Covid-19 package.* VTDigger. <https://vtdigger.org/2020/04/27/scott-seeks-more-control-over-1-25-billion-federal-covid-19-package/>

Office of the Texas Governor. (2020, July 15). *Governor Abbott Announces $41 Million in Federal COVID-19 Emergency Funding For Local Governments.* <https://gov.texas.gov/news/post/governor-abbott-announces-41-million-in-federal-covid-19-emergency-funding-for-local-governments>

Texas Department of Public Safety. (2020). *COVID-19 Travel Updates.* <https://www.dps.texas.gov/covidtravel/>

US Treasury. (2020, June 30). *Coronavirus Relief Fund Guidance for State, Territorial, Local, and Tribal Governments.* <https://home.treasury.gov/system/files/136/Coronavirus-Relief-Fund-Guidance-for-State-Territorial-Local-and-Tribal-Governments.pdf>

Vermont Department of Health (2020, July 29). *Traveling to Vermont.* HealthVermont. <https://www.healthvermont.gov/response/coronavirus-covid-19/traveling-vermont>