Alex Schmid
Annabel Lynch
Aditya Kannoth
Tiago Magalhães

SYS 2202 Group 2 Final Project: Coronavirus Tracker

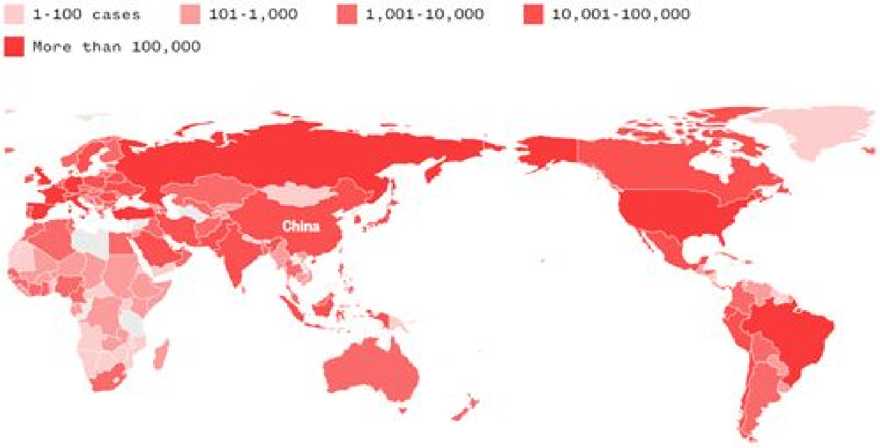
I. Summary of the idea/problem

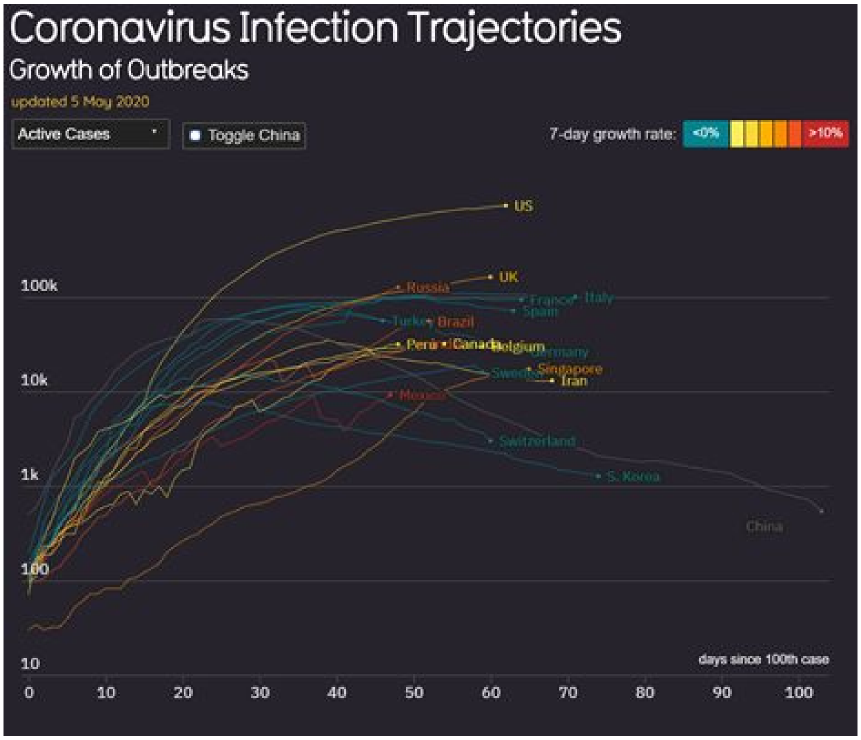
Our topic of research was the Coronavirus and its impact on the United States. Interestingly enough, when we signed up for the project, the epicenter of the virus was still China, and COVID-19 was only just beginning to spread to other countries. By the time we began working on the project, the coronavirus had already spread to every country around the world, with thousands of confirmed cases in the United States. Early during the project, the United States had already consolidated itself as the epicenter of the virus. Given the country's situation and the close impact of the virus in all of our daily lives, we decided it would be most interesting to study the spread of the virus in the United States.

Once the area of study was decided, we had to decide exactly what it was we wanted to focus on. The problems due to the coronavirus are many, and this was clearly reflected in the data sources. Most sources we found had a variety of features such as confirmed cases, deaths, and recoveries, mortality rate, hospitalization rate, tests given and tests confirmed, among others. Not only this, but there were large amounts of data with features for both counties and states. Given this information, it was clear there were many areas we could choose to focus our project on. Ultimately, to guarantee the significance of the project while keeping it unique, we settled on analyzing the infection rate of the virus by state and the ratio of available to positive tests in each state.

II. What else has been done in this domain? Details of the applications their strengths and weaknesses with screenshots and references to sources

Given the international relevance of the virus, we encountered many studies and a variety of different ways to interpret and present the data. The two most common studies were line plots comparing the spread of the virus among countries, and choropleth maps based on confirmed cases. The line plots were simple yet extremely powerful visualizations because by comparing the infection rate in each country they not only showed which countries were most and least affected by the virus, but they also portrayed the stage of the pandemic in each country. For example, if the infection rate was growing exponentially, then it can be assumed that the country is in the initial to middle stages of the virus, whereas if the growth rate is flatlining, then the country may be nearing the end stages. The growth rate could also be interpreted in many other ways when paired with information such as social distancing policies and even the form of government in each country. The choropleth maps were also efficient visualizations because they quickly presented which countries had most cases of the virus. Furthermore, the choropleth maps are easy to understand and are quite visually appealing. Beneath are examples of these visual representation taken from the Information is Beautiful organization and from NBC New1s.[1, 2]



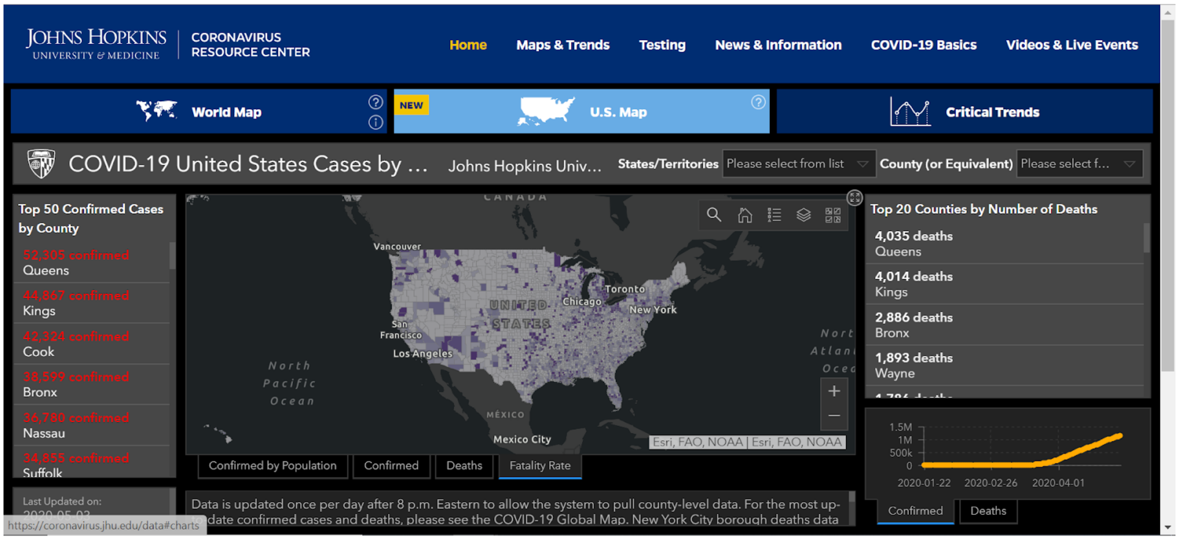


The limits of these visualizations were also a topic of interest to our group, because by identifying them we could better avoid them when doing our own mappings. The most significant limitation with the choropleth maps is mapping data with large disparities. When there is a large difference in data, the choropleth map usually fails to represent accurately. This is because either all locations are of the same tone, with the outlier being significantly colored in, or, in an attempt to solve this issue, a lot of data is left out. This is seen specifically in the shown choropleth map with the "More than 100,000" category. This category is added to avoid the map being entirely white compared to the 1,000,000 cases in the United States, but as a consequence, a viewer cannot tell the difference between 100,000 cases and 1,000,000 cases. The limitation associated with the line plot is that in order to keep it simple and comprehensive, the data pictured must be filtered. If one were to plot every country in the world, the graph would be saturated and difficult to understand. Even in the visualization shown above, which is significantly filtered to show the most affected countries, it is still hard to differentiate among them. Therefore, to create a well designed line plot, it is necessary to determine what is being plotted and have a reasonable justification for it.

III. How novel is your idea/solution? What are you doing differently that hasn't been done? Provide evidence.

We came up with this idea back in February, so our initial idea was simply to create a tracker that would visualize the spread of cases. As institutions and companies realized how dire the situation was getting, they quickly invested time and money into multifaceted, interactive visualizations to communicate the spread of the disease. As a result, we decided to focus on testing data rather than the spread of confirmed cases and deaths, like Johns Hopkins was, or the important breakdown of the virus' effects by race, as the CDC is.[3, 4] We did this because we felt that testing data was not being accurately represented in the media and its implications were being misrepresented at times. The idea was to add to existing work on the subject by going further in-depth on the amount of tests that states had done. This was relevant because in lieu of a federal response, social distancing restrictions were mostly coordinated by states. This meant that even as the US had a significant difference in how different states would be affected by the virus in the first place, it would further be complicated by the difference in implementation of distancing restrictions by state.

In addition, organizations like Johns Hopkins also included visualizations about the fatality rate.



COVID-19 United States Cases by ... Johns Hopkins Univ...

States/Territories: Please select from list

County (or Equivalent): Please select f...

Top 50 Confirmed Cases by County

52,305 confirmed
Queens

44,867 confirmed
Kings

42,324 confirmed
Cook

38,599 confirmed
Bronx

36,780 confirmed
Nassau

34,855 confirmed
Suffolk

Last Updated on:
2020-05-03

Data is updated once per day after 8 p.m. Eastern to allow the system to pull county-level data. For the most up-to-date confirmed cases and deaths, please see the COVID-19 Global Map. New York City borough deaths data

Top 20 Counties by Number of Deaths

4,035 deaths
Queens

4,014 deaths
Kings

2,886 deaths
Bronx

1,893 deaths
Wayne

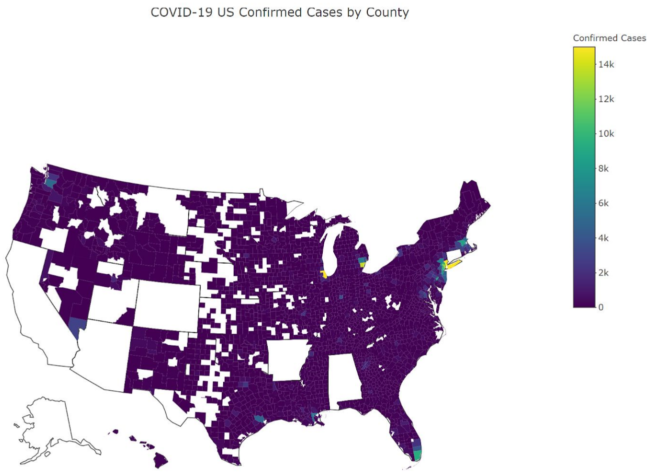
We felt that visualizations like this could be misleading to many people when presented without explanation. At the locality level, this visualization depicts the number of deaths divided by the number of confirmed cases, yielding the map above. However, as we now know, localities in states like New York have experienced a far higher number of deaths than places in the states in the middle of the country, but the rate gives an inexperienced viewer the idea that the severity of the virus in deep purple areas is dire, when these localities could just have a lack of available tests for the population.

Our static visualizations have several different facets, and we settled on presenting several main graphs, created after cleaning the external COVID Tracking Project Data and extracting specific features, in order to clarify our purpose. These static graphs include a stacked bar graph of positive vs negative tests in each state, percentage of positive tests in each state with a color fill based on absolute number of positive tests, and ratio of positive tests to state population with a color fill based on the absolute population. We found that Politico's tracker, the data journalism that was most similar to ours, presented virus trends with different scales based on the different states, a problem that we ran into early in our process.[5] Our static visualizations dealt with this issue by putting them on the same scale and managing to integrate ratios and absolute statistics into singular graphs, as demonstrated in our presentation.

As the Johns Hopkins website is one of the most popular in the world for its visualization of statistics down to a county and city level, we wanted to create a product with a different purpose - providing a visualization of testing data by state, which was very relevant given the importance of testing to states' reopening strategies.[6] The interactive visualization presented a US-based R Shiny app that demonstrated how the number of coronavirus tests and associated positive results increased exponentially over the course of three months, representing the possible strain on our healthcare systems across the nation.

IV. How did you design your solution? Details of the thought process

Our thought process for designing this solution began with the primary goals of analyzing state testing data in a comprehensive manner. In order to accomplish this, we wanted to not only create an interactive map as suggested in the prompt, but include a variety of visualizations and features extracted, which would create a more holistic representation of the data. We wanted to integrate novel aspects into our design, such as a simple and direct interactive map. One important aspect of this was the idea of scale. For example, we wanted to scale individual state graphs of positive cases over time to the scale of New York. This scale would add more information to the viewer as they have a better scale of comparison to understand the magnitude of impact in each state. Additionally, scale was an important concept when determining how we wanted to design our interactive map. We initially wanted to try using a choropleth map, but with the limitations we saw from other trackers, the scale was difficult to create. It ranged from hundreds of thousands of positive cases to only a few hundred cases which proved difficult to address in a choropleth map and did not represent the testing day how we designed. Below is an initial image of our interactive map. We changed our design to better represent the scale of testing data by using circles for each state. The size of the circles corresponds to the number of tests and is much easier for the viewer to understand the scale, and therefore the magnitude of the outbreaks in each state.



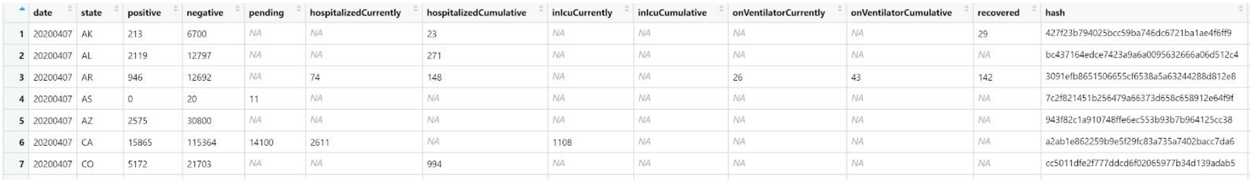
V. How did you implement your solution? Details of data sources and processing steps

Data Source:

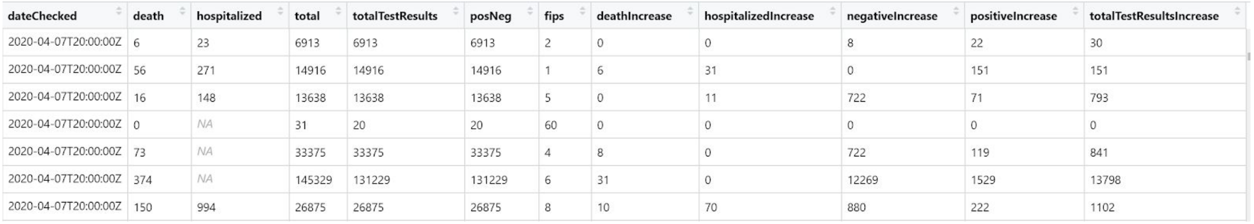
We gathered our data from the COVID Tracking Project's state daily dataset. This data is updated everyday at 4pm and is collected from state and district health officials and occasionally supplemented by news reports, press conferences or tweets when appropriate. All data sources are cited in their datasets. During our research we also found that a few other coronavirus dashboards and trackers used the COVID Tracking Project's data, suggesting that others found that it is a reliable source.

Describe the columns of the data:

Daily State Schema:



	date	state	positive	negative	pending	hospitalizedCurrently	hospitalizedCumulative	inIcuCurrently	inIcuCumulative	onVentilatorCurrently	onVentilatorCumulative	recovered	hash
1	20200407	AK	213	6700	NA	NA	23	NA	NA	NA	NA	29	427f23b794025bcc59ba746dc6721ba1ae4f6ff9
2	20200407	AL	2119	12797	NA	NA	271	NA	NA	NA	NA	NA	bc437164edce7423a9a6a0095632666a06d512c4
3	20200407	AR	946	12692	NA	74	148	NA	NA	26	43	142	3091efb8651506655cf6538a5a63244288d812e8
4	20200407	AS	0	20	11	NA	NA	NA	NA	NA	NA	NA	7c2f821451b256479a66373d658c658912e64f9f
5	20200407	AZ	2575	30800	NA	NA	NA	NA	NA	NA	NA	NA	943f82c1a910748ffe6ec553b93b7b964125cc38
6	20200407	CA	15865	115364	14100	2611	NA	1108	NA	NA	NA	NA	a2ab1e862259b9e5f29fc83a735a7402bacc7da6
7	20200407	CO	5172	21703	NA	NA	994	NA	NA	NA	NA	NA	cc5011dfe2f777ddcd6f02065977b34d139adab5



dateChecked	death	hospitalized	total	totalTestResults	posNeg	fips	deathIncrease	hospitalizedIncrease	negativeIncrease	positiveIncrease	totalTestResultsIncrease
2020-04-07T20:00:00Z	6	23	6913	6913	6913	2	0	0	8	22	30
2020-04-07T20:00:00Z	56	271	14916	14916	14916	1	6	31	0	151	151
2020-04-07T20:00:00Z	16	148	13638	13638	13638	5	0	11	722	71	793
2020-04-07T20:00:00Z	0	NA	31	20	20	60	0	0	0	0	0
2020-04-07T20:00:00Z	73	NA	33375	33375	33375	4	8	0	722	119	841
2020-04-07T20:00:00Z	374	NA	145329	131229	131229	6	31	0	12269	1529	13798
2020-04-07T20:00:00Z	150	994	26875	26875	26875	8	10	70	880	222	1102

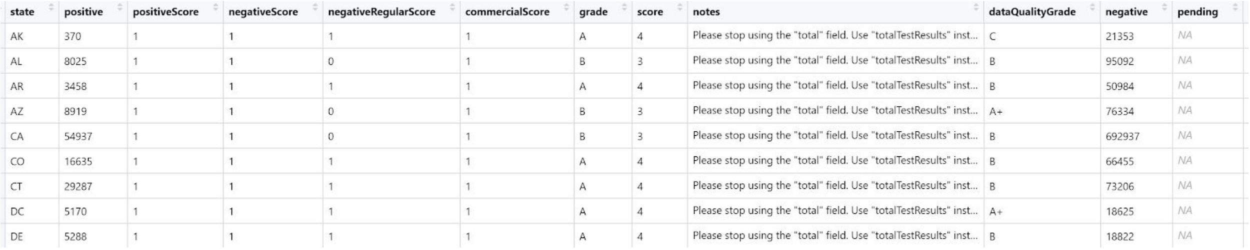
Daily State Columns:

#	Column Name:	Date Type:	#	Column Name:	Date Type:
1	date	int	14	dateChecked	Factor w/ 35 levels
2	State	Factor w/ 56 levels	15	death	int
3	positive	int	16	hospitalized	int
4	negative	int	17	total	int
5	pending	int	18	totalTestResults	int
6	hospitalizedCurrently	int	19	posNeg	int
7	hospitalizedCumulative	int	20	fips	int
8	inIcuCurrently	int	21	deathIncrease	int
9	inIcuCumulative	int	22	hospitalizedIncrease	int
10	onVentilatorCurrently	int	23	negativeIncrease	int
11	onVentilatorCumulative	int	24	positiveIncrease	int
12	recovered	int	25	totalTestResultsIncrease	int
13	hash	Factor w/ 1821 levels			

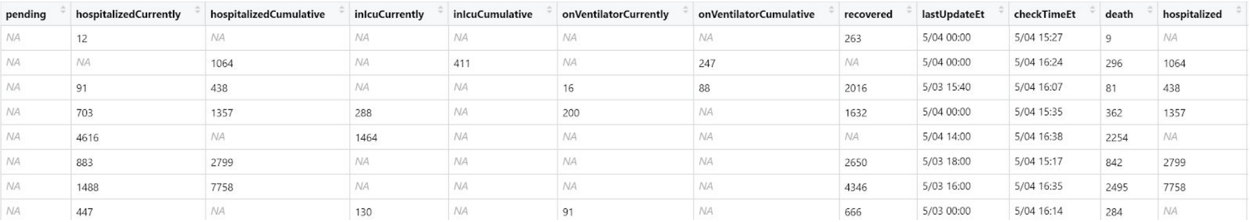
Table 1: Daily State Data Columns

As displayed in the schema and the table above, this dataset has a large number of attributes. When looking at the schema to better understand the data, we found that even in the first few rows, there are a large amount of NA values in many columns.

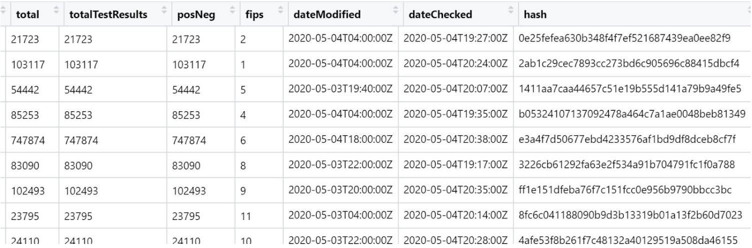
Current Data Schema:



state	positive	positiveScore	negativeScore	negativeRegularScore	commercialScore	grade	score	notes	dataQualityGrade	negative	pending
AK	370	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	C	21353	NA
AL	8025	1	1	0	1	B	3	Please stop using the "total" field. Use "totalTestResults" inst...	B	95092	NA
AR	3458	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	B	50984	NA
AZ	8919	1	1	0	1	B	3	Please stop using the "total" field. Use "totalTestResults" inst...	A+	76334	NA
CA	54937	1	1	0	1	B	3	Please stop using the "total" field. Use "totalTestResults" inst...	B	692937	NA
CO	16635	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	B	66455	NA
CT	29287	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	B	73206	NA
DC	5170	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	A+	18625	NA
DE	5288	1	1	1	1	A	4	Please stop using the "total" field. Use "totalTestResults" inst...	B	18822	NA



pending	hospitalizedCurrently	hospitalizedCumulative	inIcuCurrently	inIcuCumulative	onVentilatorCurrently	onVentilatorCumulative	recovered	lastUpdateEt	checkTimeEt	death	hospitalized
NA	12	NA	NA	NA	NA	NA	263	5/04 00:00	5/04 15:27	9	NA
NA	NA	1064	NA	411	NA	247	NA	5/04 00:00	5/04 16:24	296	1064
NA	91	438	NA	NA	16	88	2016	5/03 15:40	5/04 16:07	81	438
NA	703	1357	288	NA	200	NA	1632	5/04 00:00	5/04 15:35	362	1357
NA	4616	NA	1464	NA	NA	NA	NA	5/04 14:00	5/04 16:38	2254	NA
NA	883	2799	NA	NA	NA	NA	2650	5/03 18:00	5/04 15:17	842	2799
NA	1488	7758	NA	NA	NA	NA	4346	5/03 16:00	5/04 16:35	2495	7758
NA	447	NA	130	NA	91	NA	666	5/03 00:00	5/04 16:14	284	NA



total	totalTestResults	posNeg	fips	dateModified	dateChecked	hash
21723	21723	21723	2	2020-05-04T04:00:00Z	2020-05-04T19:27:00Z	0e25fefea630b348f4f7ef521687439ea0ee82f9
103117	103117	103117	1	2020-05-04T04:00:00Z	2020-05-04T20:24:00Z	2ab1c29cec7893cc273bd6c905696c88415dbcf4
54442	54442	54442	5	2020-05-03T19:40:00Z	2020-05-04T20:07:00Z	1411aa7caa44657c51e19b555d141a79b9a49fe5
85253	85253	85253	4	2020-05-04T04:00:00Z	2020-05-04T19:35:00Z	b05324107137092478a464c7a1ae0048beb81349
747874	747874	747874	6	2020-05-04T18:00:00Z	2020-05-04T20:38:00Z	e3a4f7d50677ebd4233576af1bd9df8dceb8cf7f
83090	83090	83090	8	2020-05-03T22:00:00Z	2020-05-04T19:17:00Z	3226cb61292fa63e2f534a91b704791fc1f0a788
102493	102493	102493	9	2020-05-03T20:00:00Z	2020-05-04T20:35:00Z	ff1e151dfeba76f7c151fcc0e956b9790bbcc3bc
23795	23795	23795	11	2020-05-03T04:00:00Z	2020-05-04T20:14:00Z	8fc6c041188090b9d3b13319b01a13f2b60d7023
24110	24110	24110	10	2020-05-03T22:00:00Z	2020-05-04T20:28:00Z	4afe53f8b261f7c48132a40129519a508da46155

#	Column Name:	Date Type:	#	Column Name:	Date Type:
1	state	Factor w/ 56 levels	16	inIcuCumulative	int
2	positive	int	17	onVentilatorCurrently	int
3	positiveScore	int	18	onVentilatorCumulative	int
4	negativeScore	int	19	recovered	int
5	negativeRegularScore	int	20	lastUpdateEt	Factor w/ 34 levels
6	commercialScore	int	21	checkTimeEt	Factor w/ 47 levels
7	grade	Factor w/ 4 levels	22	death	int

8	score	int	23	hospitalized	int
9	notes	Factor w/ 1 level	24	total	int
10	dataQualityGrade	Factor w/ 6 levels	25	totalTestResults	int
11	negative	int	26	posNeg	int
12	pending	int	27	fips	int
13	hospitalizedCurrently	int	28	dateModified	Factor w/ 34 levels
14	hospitalizedCumulative	int	29	dateChecked	Factor w/ 47 levels
15	inIcuCurrently	int	30	hash	Factor w/ 56 levels

Table 2: Current State Data Columns

Whereas the daily state data provides data for each day on record, the current dataset provides the most up to date totals for each of the columns. It additionally provides details regarding the quality of data and dates it has been updated and modified.

Another dataset we used was state population data from the Census in order to execute our analyses with state population and testing.

The last dataset we used was a dataset from Kaggle that had a column for the state abbreviation and columns for latitude and longitude values corresponding to a centralized location in each state.

Cleaning:

To clean the daily state data, we changed the data type of the state column from a factor to a character using as.character() so it would be easier to work with in the data processing stage. We additionally used the transform() function and as.Date() to convert the integer date column to a date type. As many columns had a large amount of NA values and were therefore unusable and unnecessary, we used the select() function to focus on the date, state, positive, negative, pending, death, total, and totalTestResult attributes.

In this decision we kept the pending attribute because although many of its values are NA, it demonstrated that a couple states were struggling to report their testing results in large volumes, often having thousands of pending tests at a time.The presence of pending tests is the reason why we kept both columns “total” and “totalTestResults”. “totalTestResults” does not include pending tests, while “total” includes pendings tests, resulting in a difference of thousands of tests for specific states. California specifically, was one state that had an outstanding number of pending tests. As time progressed, tests became more abundant and turnaround time was decreased. We primarily focused on using “totalTestResults” because the outcomes of tests are clear and defined, as opposed to the uncertainty of using the “total” data. We made a few initial visualizations with the “total” data to explore the differences between the two, however, overall the differences were quite small. In regards to further cleaning, we then converted the NA values in pending to zero and lastly, we removed US territories to focus on state trends.

For the current data frame, we approached cleaning in the same way. We changed the appropriate data types, selected the necessary columns, replaced NA values with zero, and removed US territories.

For the state population data, we removed US territories with the subset() function, changed the state name to an abbreviation using state.abb and the which() function, and arranged the rows in alphabetical order with the arrange() function.

For the state latitude and longitude date, we removed one unnecessary column, arranged the states in alphabetical order with the arrange() function, and replaced specific states’ latitude and longitude values with new values that map to a more central location in the state so it would appear clearer on the R Shiny interactive map.

Transformation and Feature Extraction:

In order to evaluate what percentage of the population has been tested, we joined the current data frame with the state population data frame using left_join() at the state attribute. We extracted a new feature, percent_tested, and added it to the current data frame by dividing totalTestResult by the population for each state. This feature establishes how much of the population has been tested for COVID-19 and can provide insights into a state’s ability to test its people effectively. We created a new data frame, max_test_pop, which selected the state and percent_tests attributes and arranged the states from highest percentage of population tested to lowest percentage of population tested using the order() function.

Next we wanted to identify the states with the highest number of positive cases and the highest ratio of positive test results to total test results. The new dataframe, max_pos, was created with the same attributes as current but ordered so that the states with the highest number of positive results are at the top. Next, we extracted the feature percent_pos and added it to the current data frame by dividing the number of positive results by the total number of tests, using the totalTestResults column, not the total column. We then created a new data frame, max_ratio, which ordered the states with the highest percentage of positive results at the top.

Once we established the top states in regards to the number of positive cases and the percentage of positive cases, we wanted to evaluate the top states in isolation from all 50 states.

We used the subset() function to select NY, NJ, MA, CA, CT, PA, MI, GA, IL and FL, as they were the top states at the time, and create a new dataframe called top_daily. We additionally used the subset function to create individual data frames for specific top states, such as NY, NJ, MA, IL, and CA. These individual data frames give the ability to evaluate a single state's progress over time with visualizations.

Methods for Analyzing Data:

To analyze our data we compiled extracted features, graphs, and an interactive map to provide a more holistic image of states' testing abilities.

State	**Number of Positive Test Results**
NY	318,953
NJ	128,269
MA	69,087
IL	63,840
CA	54,937

State	**Percentage of Positive Test Results**
NJ	46.26%
NY	31.66%
CT	28.57%
DE	21.93%
MA	21.31%

We extracted the top five states that had the highest raw number of positive cases and the highest percentage of positive cases to understand their testing needs and how severe the outbreaks in these states are. New York and New Jersey are the top two on both tables, although

New Jersey has a higher percentage of positive test results than New York by about 15% and about 200,000 fewer positive cases. This contrast suggests that New Jersey's testing capacity is too low for the severity of their outbreak. Although New York's number of positive cases is high, around 300,000, that only makes up about 30% of their tests. As we have been working on this presentation the percentage of positive tests has continually decreased, suggesting an improving testing capacity that may be coupled with a decreasing number of positive cases.

Visualization Techniques:

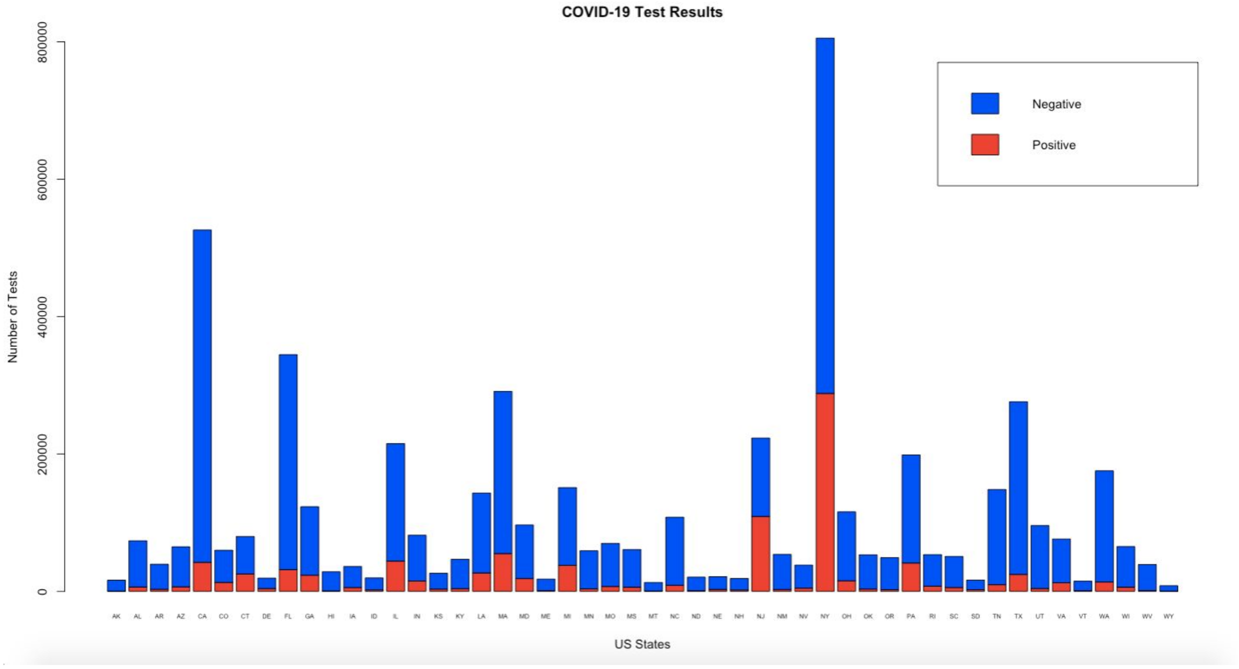


Figure 1: Stacked Bar Plot of State Positive and Negative Cases of COVID-19

This first graph below shows the 50 states and the total number of tests in each state. The bars are filled with blue and red to demonstrate the ratio of positive and negative results from those tests. It was interesting to see how over the course of a few weeks, the number of tests given in each state changed. California is more recently on the rise with an increase in testing. New York remains to be a standout with over 800,000 tests. For all 50 states, it is clear that a majority of the test results are negative.

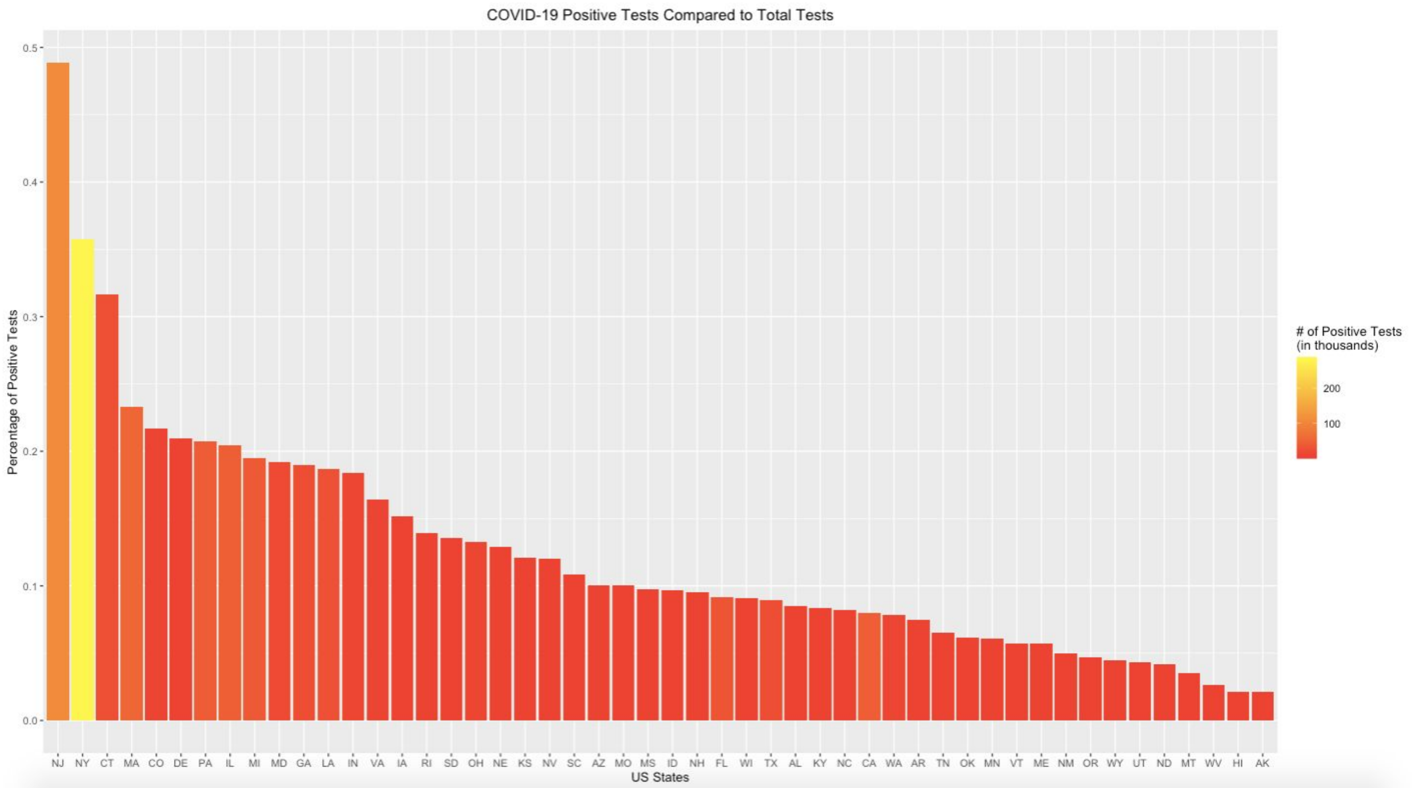


Figure 2: Bar Plot of Percentage of Positive Cases and Number of Positive Cases

Figure 2 further reflects on the information in the stacked bar graph. This bar graph compares the US states and the percentage of their test results that are positive. Furthermore, the bars are filled relative to the number of positive tests in each state.

As you can see, although New York has tested the most positive cases in the country, New Jersey has a higher percentage of positive test results. Many states have ramped up their testing since the spread of coronavirus across the country. It was interesting to see that when less tests were readily available, states had a much higher percentage of positive results. States were prioritizing tests for people who were most likely to have it. Factors such as illness symptoms, recent travel, and possible exposure to a confirmed COVID-19 patient were heavily considered. As described above, testing became much more readily available as weeks passed. This resulted in a lower percentage of positive test results.

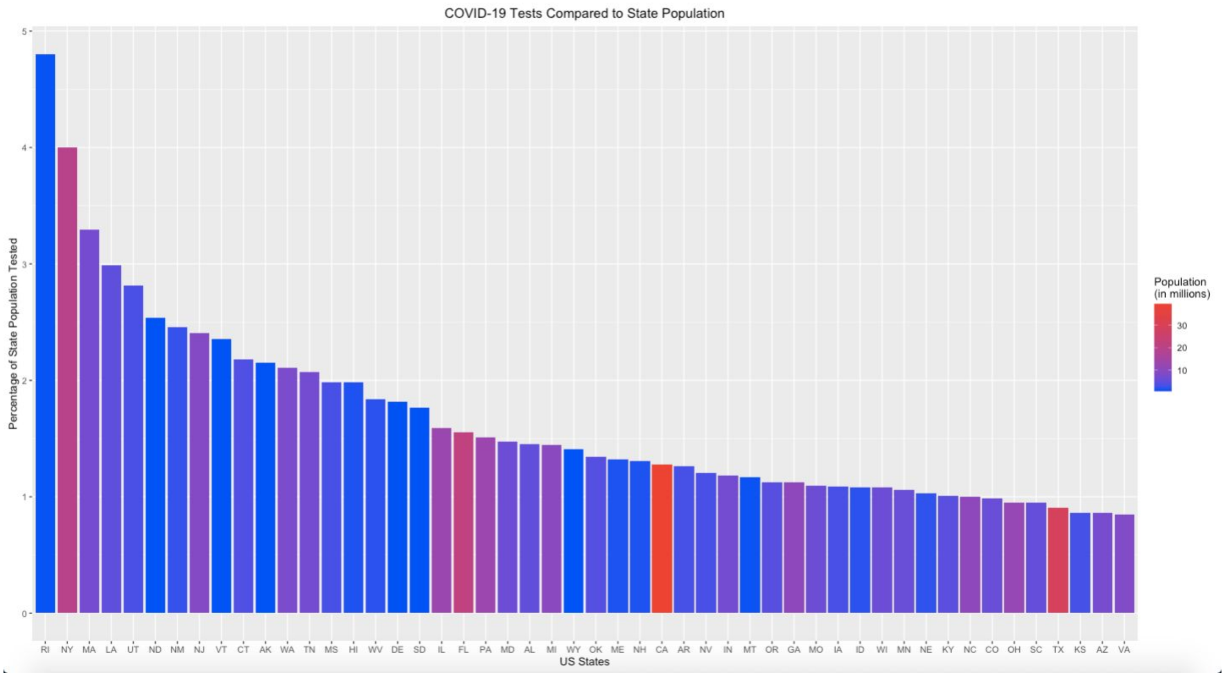


Figure 3: COVID-19 Testing Compared to State Population

The final graph shows the US states and what percentage of their population has been tested. Prior to making this graph, the group analyzed two possible scenarios that the data could have reflected. On one hand, it could be assumed that states with larger populations have a higher number of tests. More resources and funding may allow for a greater production of tests, and therefore a greater percentage of their population tested. On the other hand, it could be assumed that states with smaller populations can more easily test their residents because they have significantly less people. From the graph, we can see that the latter assumption is more accurate. When analyzing California and Texas, the two most populated states, their percentage of population that has been tested falls in the bottom half of the 50 US states. Surprisingly, Rhode Island beats out New York for the percent of the population tested. Rhode Island ranks 44/50 when it comes to highest state populations.

It's notable to mention that almost 3/4 of states have tested less than 2% of their populations. As some states consider to begin lifting policy restrictions and start the re-open of their economies, the number of available tests will play a huge factor. Regular testing on a large scale will keep people safe. It is not necessary for everyone to be tested, but society becomes safer as larger portions of communities are tested. The exact numbers are not known, but many people infected with COVID-19 could be asymptomatic carriers. These cases cannot be tracked or known without testing. Some experts say that millions of tests must be performed a day before safely returning back to normal life. According to the CDC, there have only been around 4 million tests given in the US since March 1, 2020.

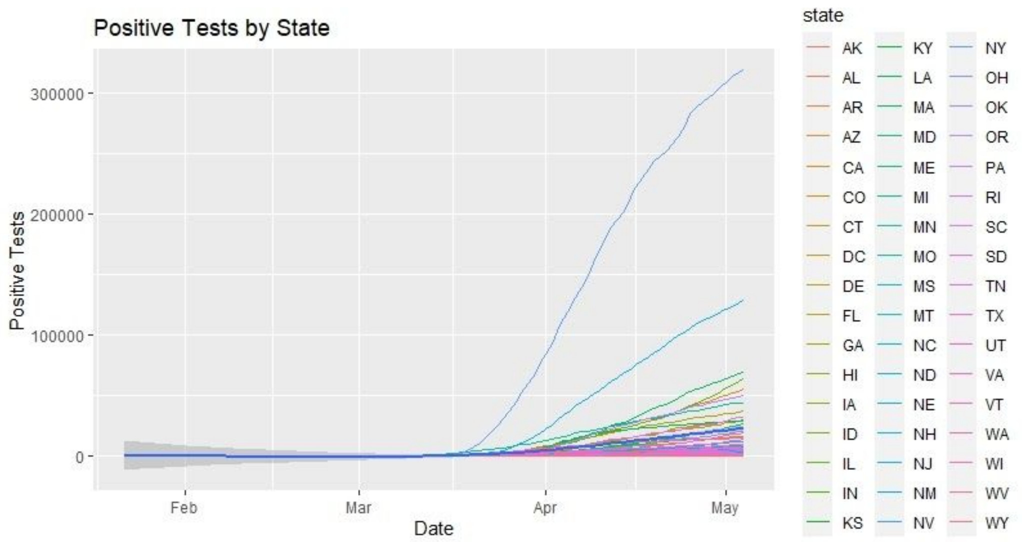


Figure 4: Number of Positive Tests by State Over Time

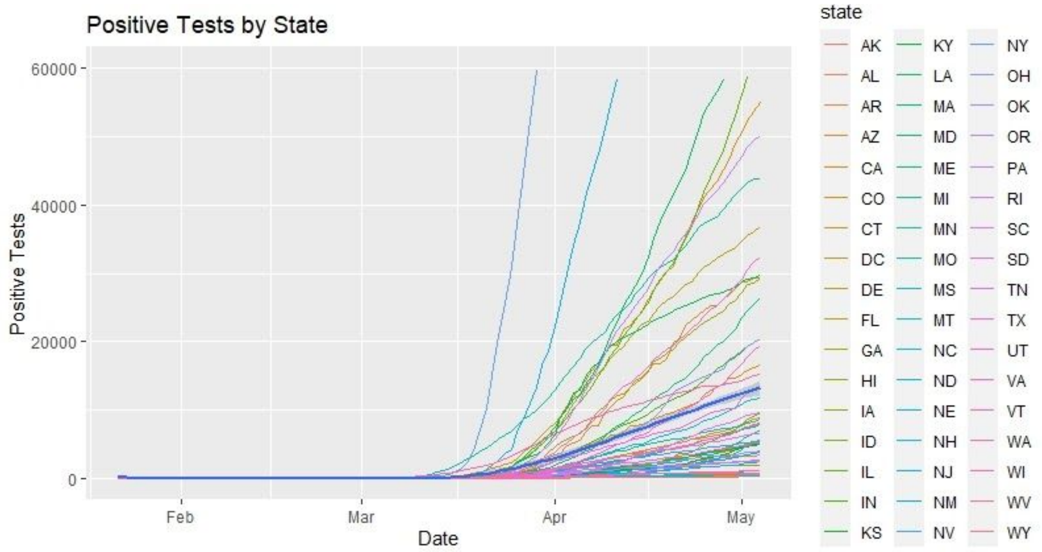


Figure 5: Number of Positive Tests by State Over Time, Scaled

In order to look at how the number of positive cases in each state has progressed over time, we created two line graphs, Figure 4 and Figure 5. These graphs identify states with outbreaks, such as New York and New Jersey, and provide an image of their immense severity in comparison to all other states. We wanted to see what the overall trend of positive cases in the United States was so we used geom_smooth(), which is the blue line above. This line provides a rough idea of state trends in COVID-19 cases, mainly contributing to the dramatic rise in early April, that appears to be slowly, although it is unclear at this time.

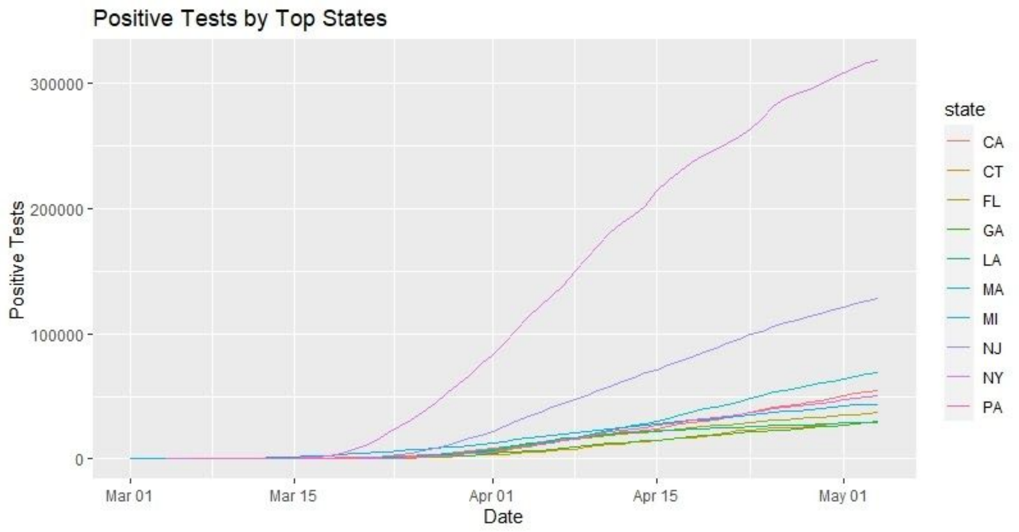


Figure 6: Number of Positive Tests in Top States

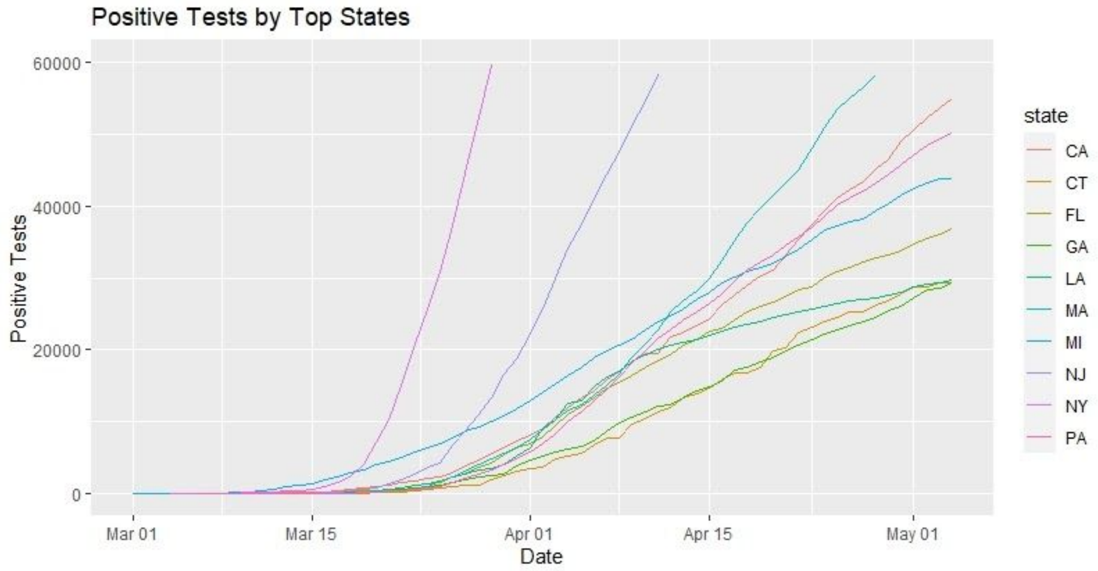


Figure 7: Number of Positive Tests in Top States, Scaled

In Figures 4 and 5, individual states were difficult to discern, so we created another plot with the ten states that had the highest number of positive tests or percentage of positive tests. This plot allows a better image of how states' cases have increased. For example, Michigan, in the blue, cuts across New York and New Jersey's exponential increases, with initially higher numbers than the two in early March which may suggest rapid initiatives to test residents and an early outbreak.

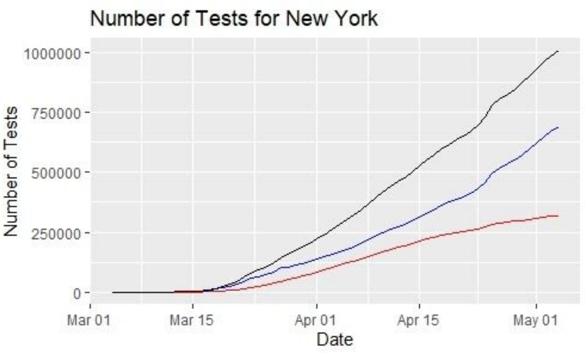


Figure 8: New York Tests Over Time

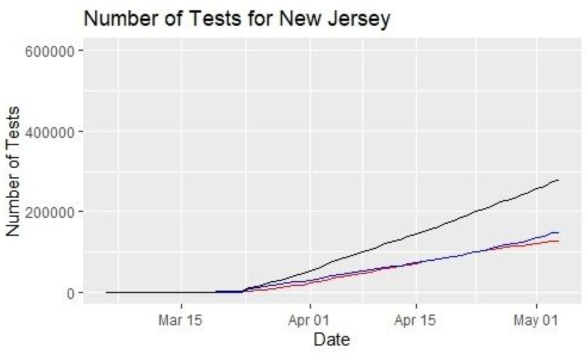


Figure 9: New Jersey Tests Over Time

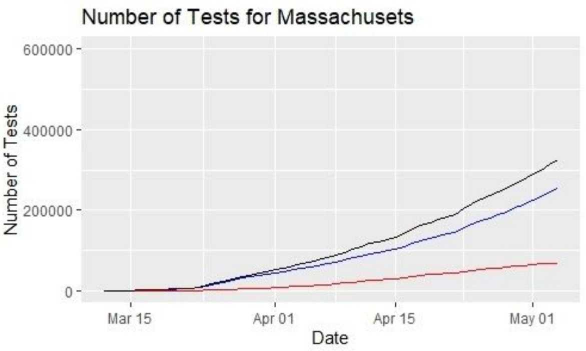


Figure 10: Massachusetts Tests Over Time

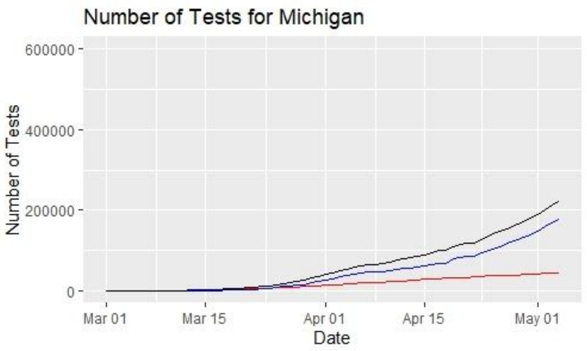
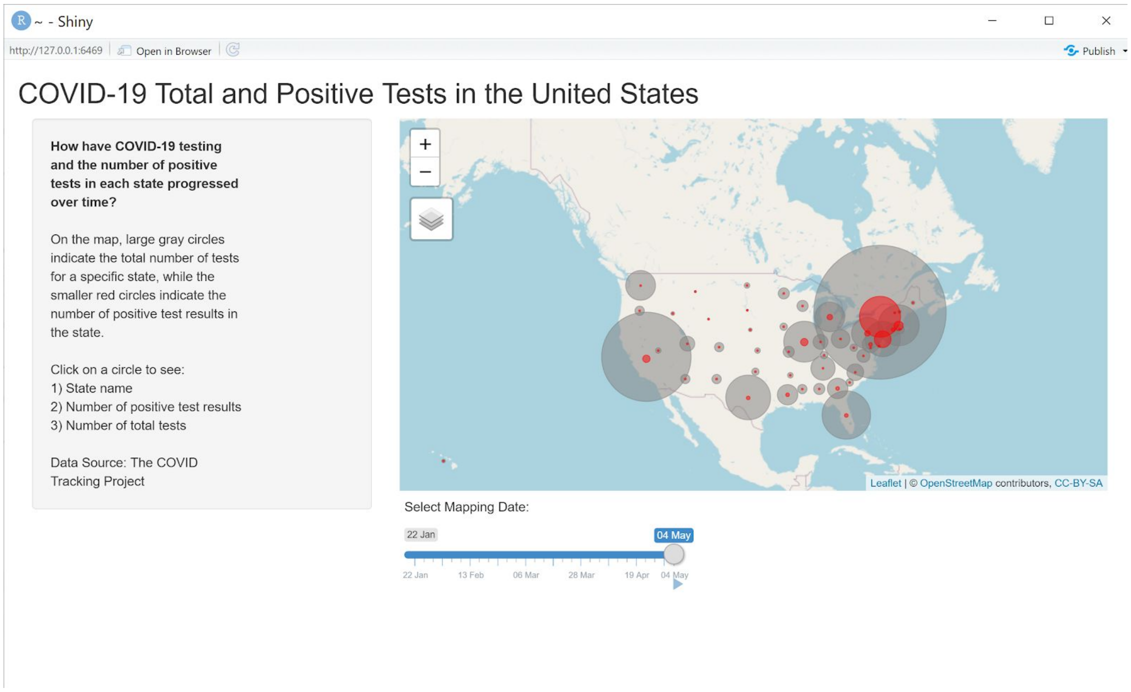


Figure 11: Michigan Tests Over Time

Figures 8 through 11 display the trends in individual states through their number of total tests (black), negative tests (blue) and positive tests (red). As discussed previously, New Jersey has had an approximately equal amount of positive and negative tests, which is demonstrated by the closeness of the red and blue lines in Figure 9, which greatly contrasts the other states with large gaps between their respective red and blue lines. In Figures 8, 10, and 11, the blue and black lines appear to be curving upward more drastically since the end of April, while the red lines are slowing or beginning to flatten. These trends suggest increased testing capacity, more widespread negative results, and a slowing of daily positive results.

Interactive R Shiny Map:

As previously mentioned, we wanted to differentiate our COVID-19 map by focusing on state testing capacity as a state's ability to test its population is one indicator of whether a state can gradually reduce social distancing measures and control the spread of coronavirus. For our R Shiny map, we wanted to visualize this progression of testing over time, so that users can understand the rapid increase in positive cases and watch as the ratio between positive and total tests changes. We initially attempted to create a choropleth map by state, however we ran into the limitations of previous trackers in which. Due to the gradient scale based on hundreds of thousands of cases in New York, all other states appeared to be the same color which made it difficult to discern the variation in state testing. We shifted from displaying data in color to displaying it in shapes and sizes through different sized circles over a centralized location in each state.



COVID-19 Total and Positive Tests in the United States

How have COVID-19 testing and the number of positive tests in each state progressed over time?

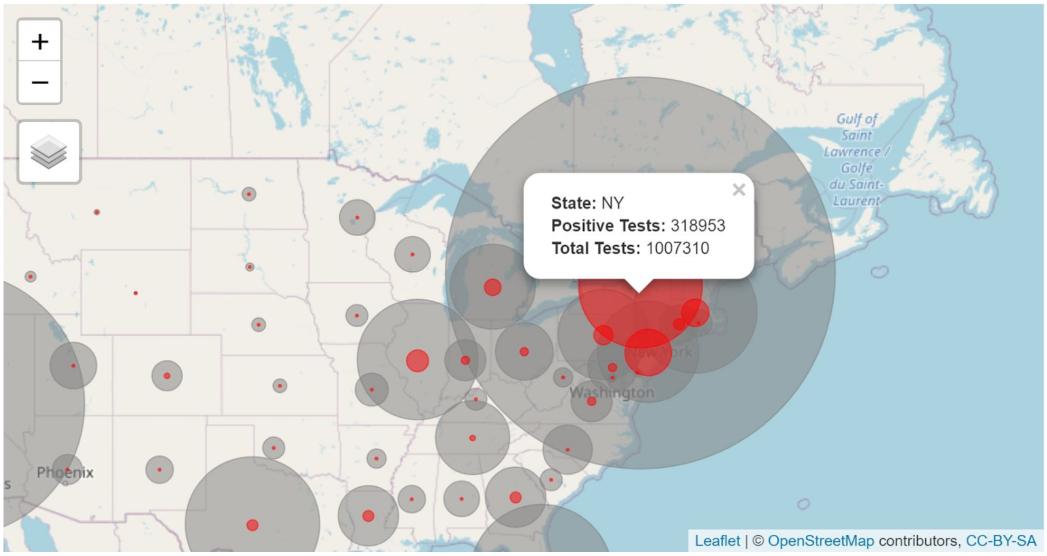
On the map, large gray circles indicate the total number of tests for a specific state, while the smaller red circles indicate the number of positive test results in the state.

Click on a circle to see:
1) State name
2) Number of positive test results
3) Number of total tests

Data Source: The COVID Tracking Project

Select Mapping Date:

Figure 12: User Interface of R Shiny Interactive Map


+
−
Gulf of Saint Lawrence / Golfe du Saint-Laurent
New York
Washington
Phoenix
Leaflet | © OpenStreetMap contributors, CC-BY-SA

State: NY
Positive Tests: 318953
Total Tests: 1007310

Figure 13: R Shiny Popup

The size of the grey circles correspond to the number of total tests in a state, while the red circle inside corresponds to the number of positive tests. The user can use the slider at the bottom to change the date and zoom into different locations in the United States. If you click on a circle it will tell you the number of total tests and positive cases in the state. This method of visualization easily allows users to comprehend the testing ability, the severity of outbreaks in each state, and how these factors have changed over time.

From the large grey circles over California and New York, these states have tested a large number of people and have vastly different numbers of positive cases for their numbers of total tests. Additionally, the outbreaks are concentrated on the east coast, with a severe outbreak in New England, in stark contrast to the northwest region with little testing and positive cases. We chose to keep the user interface simple, and not convoluted as some current dashboards can be overwhelming to use, however we may want to include more information into the interactive map in the future. For example, we could combine our previous analyses with the map by adding the ratio of positive test results to total tests and/or the percentage of state population tested to the label. When you click on one of the large circles, then you could have a better understanding of how these raw values compare to the state's other statistics.

VI. Documented and error-free code available on Github

Github Link: https://github.com/annabellynch/COVID-19-Tracking

References:

1. https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/
2. https://www.nbcnews.com/health/health-news/coronavirus-map-confirmed-cases-2020-n1120686
3. https://coronavirus.jhu.edu/us-map
4. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/data-visualization.htm
5. https://www.politico.com/interactives/2020/coronavirus-testing-by-state-chart-of-new-cases/
6. https://www.nbcnews.com/politics/politics-news/trump-lays-out-new-coronavirus-testing-blueprint-states-weigh-reopening-n1193771