**Data Analyses of COVID-19 Case Surveillance Public Use Data with Geography**

Hyunjae Kwon,1 Luke Yee,2 and Aly Kapasi3

*1 Department of Statistics, Student ID: 915084195, Email: kevkwon@ucdavis.edu*

*2 Department of Statistics, Student ID: 914913445, Email: leyee@ucdavis.edu*

*3 Department of Statistics, Student ID: 915223613, Email:* [*amkapasi@ucdavis.edu*](mailto:amkapasi@ucdavis.edu)

**Abstract:**

As soon as the world was struck with a life-threatening pandemic last year, COVID-19, a vast number of datasets and studies have been produced. The current paper performs multiple data analyses on the CDC’s “COVID-19 Case Surveillance Public Use Data with Geography.” Our analyses are centered around perhaps the most important factor of the pandemic: death rate. All the measures that are being taken in real life at the individual and state level are done with the underlying goal of reducing the number of deaths. Similarly, we approach our analyses with the hope that by finding meaningful patterns in the dataset, we can aid in reducing the number of deaths. First, the relationship between the number of cases and level of mask usage in each county in the United States is analyzed. Second, the dataset is transformed to state-level, seperated into 4 different waves , and grouped using K-means clustering. Third, three classification models--logistic regression, linear discriminant analysis, and gradient descent logistic regression--are used to classify whether or not the patient died, was admitted to an intensive care unit, and was hospitalized.

**I. Introduction**

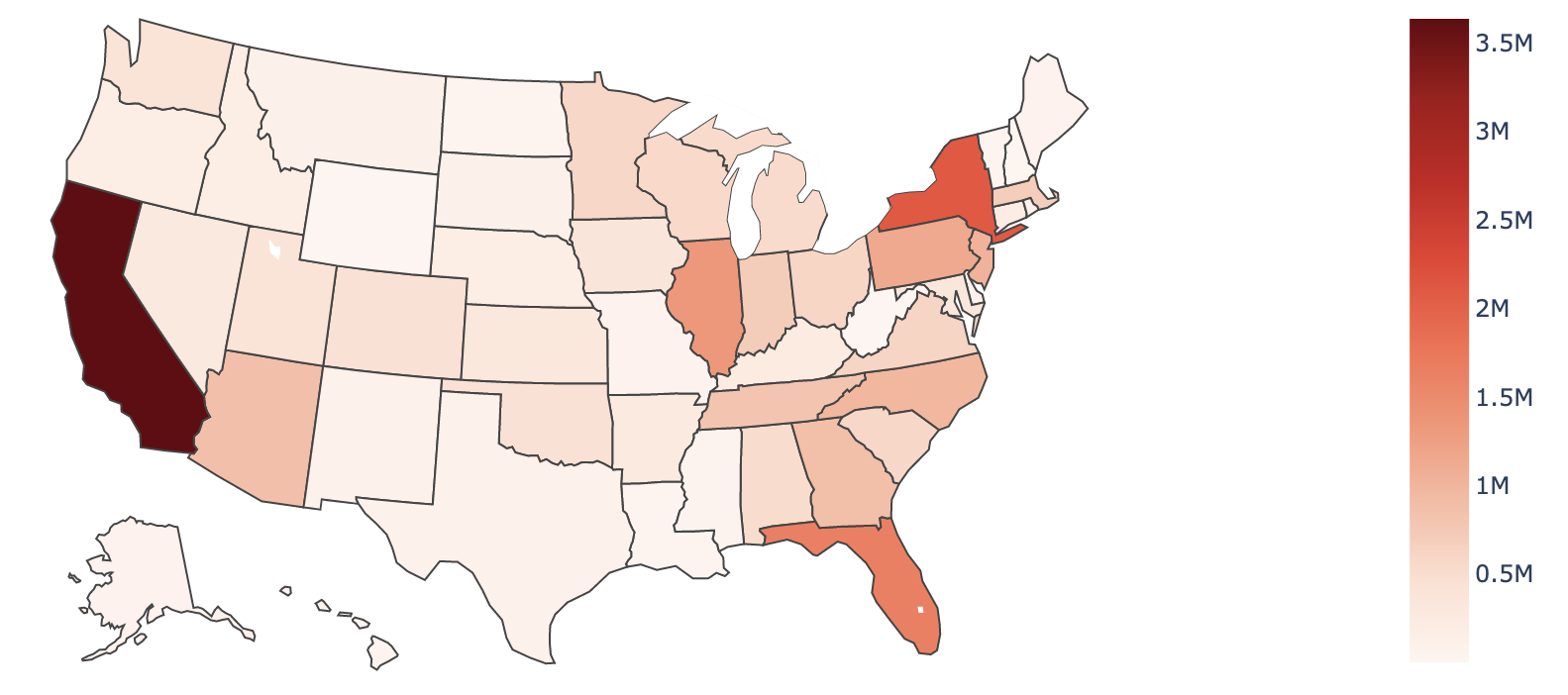
**A. Dataset**

CDC’s “COVID-19 Case Surveillance Public Use Data with Geography” is a public dataset that “includes patient-level data reported by U.S. states and autonomous reporting entities” (CDC) and “includes demographics, geography (county and state of residence), any exposure history, disease severity indicators and outcomes, and presence of any underlying medical conditions and risk behaviors” (CDC). COVID-19 cases reports are submitted frequently to CDC and thus the dataset is considered provisional (last updated on May 24, 2021). For the privacy of patients, all potentially patient exposing information is suppressed. There are 19 variables and 25,009,120 unique patients. The variables are *case\_month, res\_state, state\_fips\_code, res\_county, county\_fips\_code, age\_group, sex, race, ethnicity, case\_positive\_specimen\_interval, case\_onset\_interval, process, exposure\_yn, current\_status, symptom\_status, hosp\_yn, icu\_yn, death\_yn, and underlying\_conditions\_yn*.

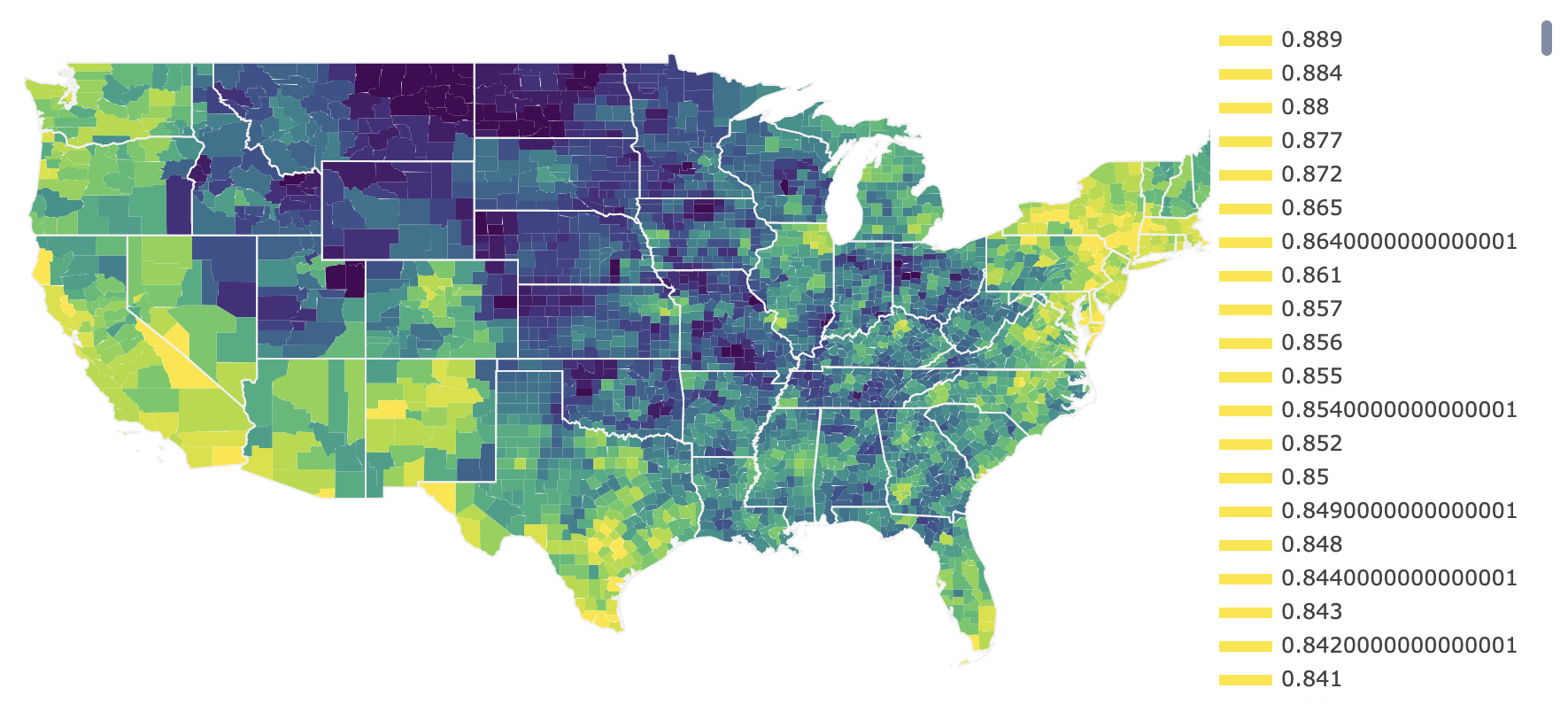
Another dataset used in the current paper is The New York Times and Dynata’s mask-wearing survey dataset. This public dataset was produced for “broad, noncommerical public use including by medical and public health researchers, policymakers, analysts and local news media” (The New York Times & Dynata). Dynata originally asked the following question--*how often do you wear a mask in public when you expect to be within six feet of another person?*--to over 250,000 interviewees between July 2 and July 14, 2020. Using these survey responses, the probability distribution of the 5 levels of mask usage--*never, rarely, sometimes, frequently, or always*--was estimated for each county in the country.

**B. Distribution of Cases**

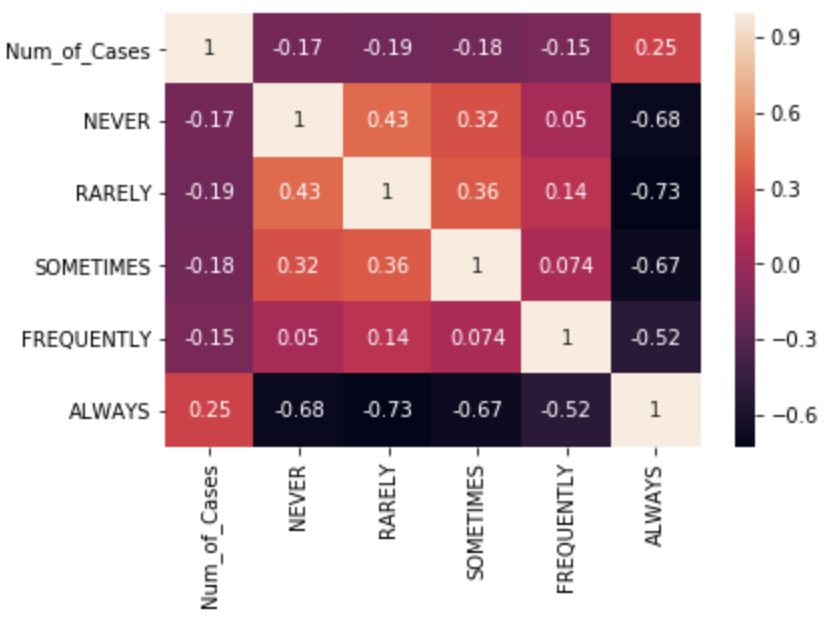
The geographic variables *res\_state, state\_fips\_code, res\_county, and county\_fips\_code* in the CDC dataset provide valuable insight into the geographic distribution of the individual cases. The choropleth map in figure 1 illustrates the distribution of cases. As one can observe, the cases are highly concentrated in California, which is followed by Florida, New York, and Illinois. Population is one factor that could explain the distribution of cases across the states. Having a larger population does not necessarily suggest a higher number of cases, but it is certainly possible because more populous states tend to have the larger metropolitan cities--e.g. Los Angeles, New York, etc.--where the population density is higher. Still, note that, depending on the effectiveness of individual and municipal measures, populous cities are very well capable of maintaining a low number of cases. California has the highest population out of all the states (~ 39 million) according to the 2019 US Census. Similarly, Florida (~21 million), New York (~19 million), and Illinois (~12 million) rank third, fourth, and sixth respectively. The number of cases and population seem to have an association. Texas is an exception, however: it has the second highest population, yet it has one of the lower number of cases (figure 1).

Figure 1. Choropleth of the Total Number of Cases by State

Population is only one of many factors that explains the distribution of cases in the country. What is more important and pertinent than population may be mask usage. From early on in the outbreak to today, masks have been continuously recommended by the CDC to protect oneself and others from the infection. Figure 2 illustrates the estimated share of people who would say “always” to the survey question “how often do you wear a mask”. Comparing Figure 1 and Figure 2, the two do not seem to share much pattern. California has the highest number of cases but has one of the highest estimated share of people who would say “always.” The same applies for Florida and New York. However, Illinois has one of the higher number of cases and one of the lower estimated share of people who would say “always.”

Figure 2. Choropleth of the Estimated Share of People Who Would Say “Always”

For a deeper analysis, the relationship between the number of cases by county and the types of mask usage is visualized using a heatmap (Figure 3). The correlation coefficients between the number of cases, “always,” “frequently,” “sometimes,” “rarely,” and “never” are 0.25, -0.15, -0.18, -0.19, and -0.17, respectively. Since the correlation coefficients are low, it can be suggested that there is almost no linear relationship between the number of cases and the types of mask usage.

Figure 3. Heatmap of Number of Cases (county-level) & Mask Usage

Another worthwhile exploration is the effects of politics on the distribution of Covid-19 cases. Covid-19 occurred during an election year and though the outbreak is a medical issue, media on both sides capitalized on the chance to politicize the disease and an individual's choice to wear a mask. These effects can be seen most clearly in a recent survey taken during July 2020 while the Pandemic was still growing rapidly. Both Republicans and Democrats were surveyed and the percentage who reported wearing masks from “Always” to “Never” is recorded.

Figure 2: Mask Usage among Democratic Citizens

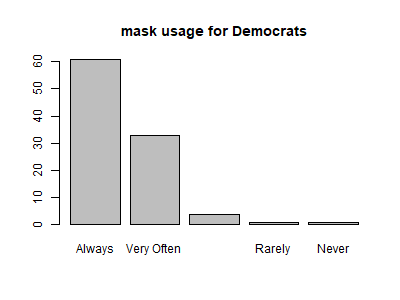
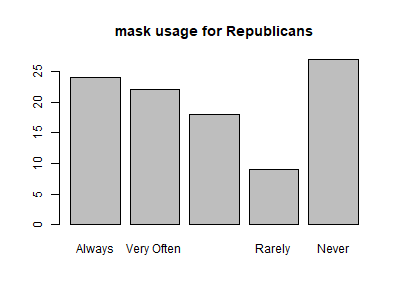


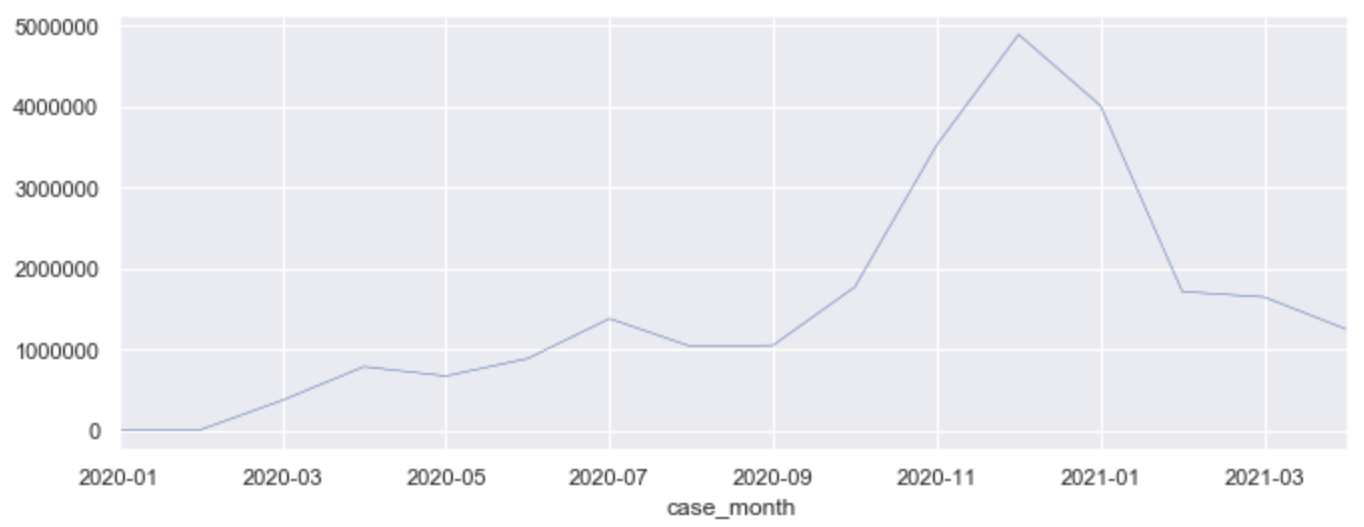
Figure 3: Mask Usage among Republican Citizens



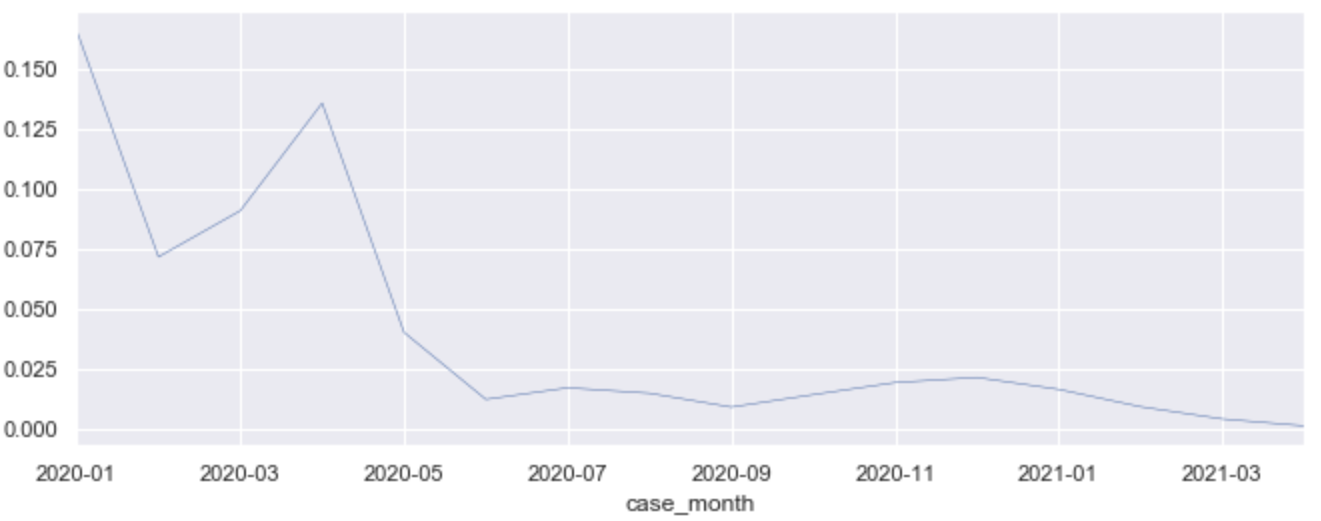
Over 90% of Democatic citizens report wearing a mask “Always” or “Very Often'' compared to just 45% of Republicans. If this is the case, why do Blue States like California and New York have some of the highest number of covid cases, while red states like Texas and Wyoming have relatively low numbers of covid cases? There are several contributing explanations for this alleged discontinuity. First, population density, which is a large factor in Covid transmission, is greater in blue states which includes large cities such as New York and San Francisco. Red states, due to their primarily rural and spread population, could have had a lower number of covid cases *despite* not wearing masks, while blue states wore masks and still suffered due to their high population density. If red states had the higher population density, we could have seen a much higher number of covid cases. Another explanation that must be considered are the limitations of surveys themselves. While political leaning citizens may answer surveys in one way, how reality plays out could be very different. One may answer that they “never” wear masks, but when they aren’t being let into stores, eventually realize that they have to wear masks. Surveys can only represent the reality that the individual construes inside their own head, while large scale reports and covid case recordings look at hard data, and the disconnect can be seen.

**C. Time Series of Cases and Death Rate**

Although the total number of cases by state reveal valuable insight about the dataset, it does not take into account the time dependent patterns of the pandemic. The growth and decline pattern of cases throughout time is first examined. Figure 4 illustrates the development of cases from January 2020 to March 2021--a little more than a year. Overall, the number of cases increased from January to December 2020 with multiple ups and downs in between. Then, the number of cases started to drop after December 2020.

Figure 4. Time Series Plot of Cases

Looking at the time series plot of cases in Figure 4, one can detect broad waves of the COVID-19 pandemic. First wave lasts from January to May 2020; Second wave begins in May and ends roughly

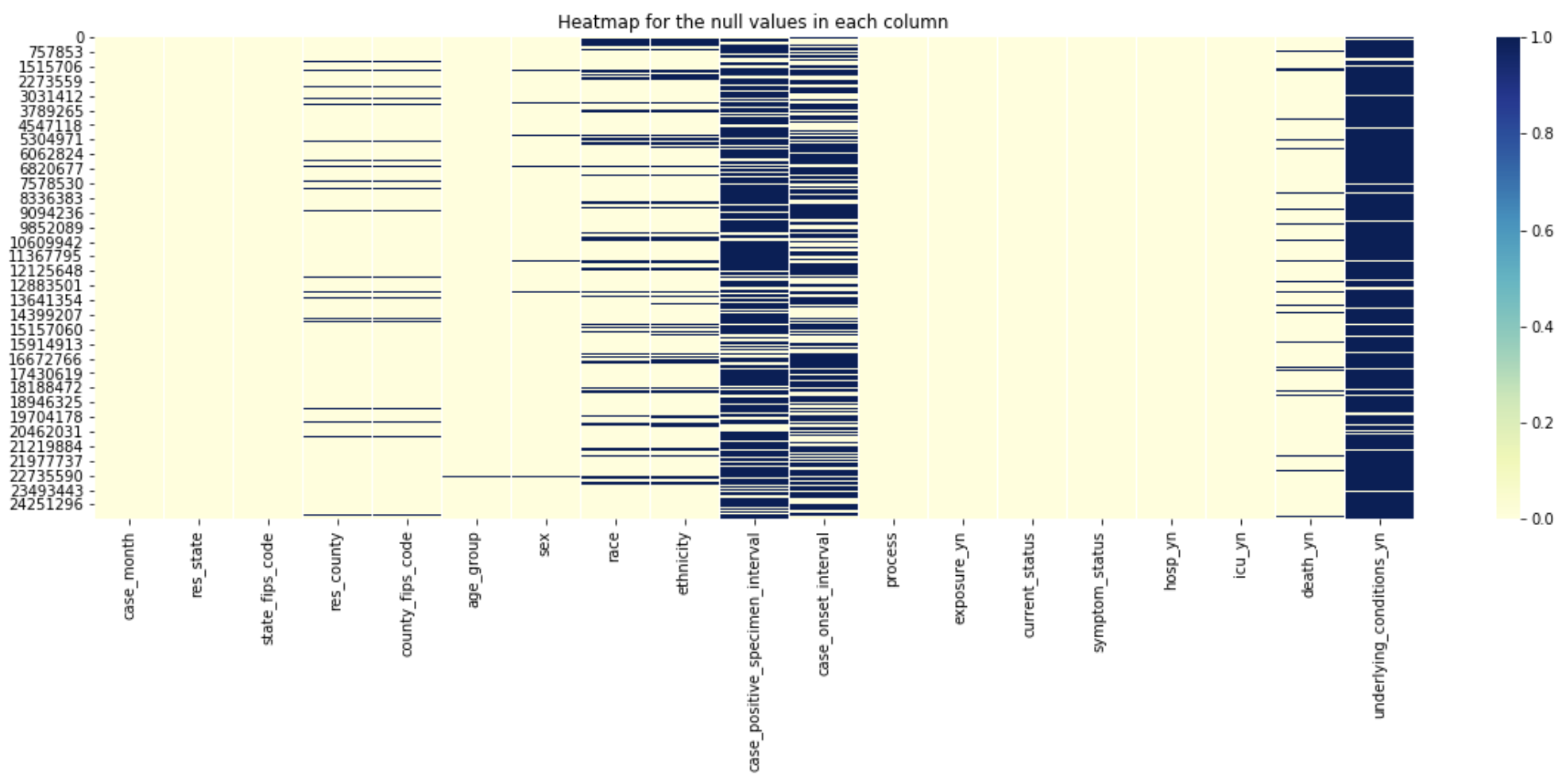
Figure 5. Time Series Plot of Mortality Rate

around July 2020; Third wave starts in August and lasts until December 2020; Fourth wave is from January to April 2021.

The plot of mortality rate in Figure 5 clearly shows that the distribution of mortality rate is denser towards the beginning of the pandemic. However, similar to the number of cases, mortality rate has several waves.

**D. Missing Values**

Several variables in the dataset consist of a large number of missing values. The variables *casee\_positive\_specimen\_interval* (72.85%)*,* *case\_onset\_interval* (53.42%), and *underlying\_conditions\_yn* (93.51%) have NA values more than 50 percent. Figure 3 provides a visualization of the NA values in the variables. Each row in the heatmap represents a case while each column represents a variable. NA values, however, are not the only missing values in the dataset. There are also “Missing” and “Unknown” values that must be dealt with. Summing the values of NA, “Missing,” and “Unknown,” there are more variables with more than 50 percent missing values (Table 2): *case\_positive\_specimen\_interval* (72.85%), *case\_onset\_interval* (53.42%), *process* (95.59%), *exposure\_yn* (93.24%), *symptom\_status* (55.50%), *hosp\_yn* (59.70%), *icu\_yn* (93.92%), *death\_yn* (50.78%), and *underlying\_conditions\_yn* (93.51%).

Figure 6. Heatmap of NA values in each variable

**E. Clustering**

K-means clustering is used to group similar states based on the variables in the dataset. Since the K-means algorithm requires numerical input and the dataset is of mixed type, categorical variables are transformed to numerical dummy variables. Then, the individual case-level dataset is grouped by states to form a state-level dataset. To examine whether the membership of a state in a specific cluster changes throughout waves, we perform clustering on the state-level dataset for each wave.

Quickly glancing through the four maps, one can see that several states undergo a switch in cluster membership. For example, in the first three waves, California is clustered with, to name a few, Oregon and Florida, but in the fourth wave, it is clustered with Oklahoma, North Carolina, and Pennsylvania.

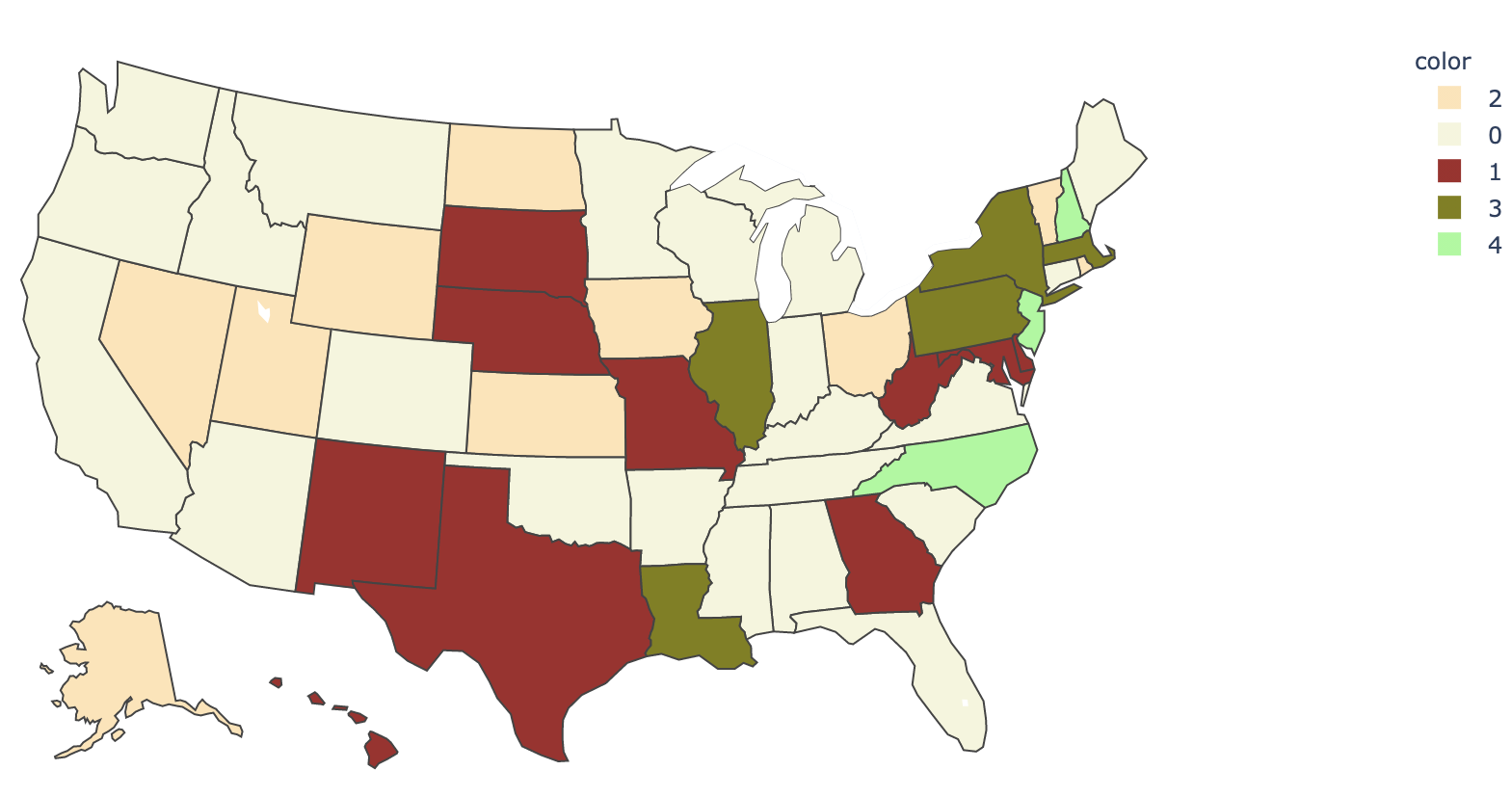
Figure 7. Clusters of States in the First Wave (January to April 2020)

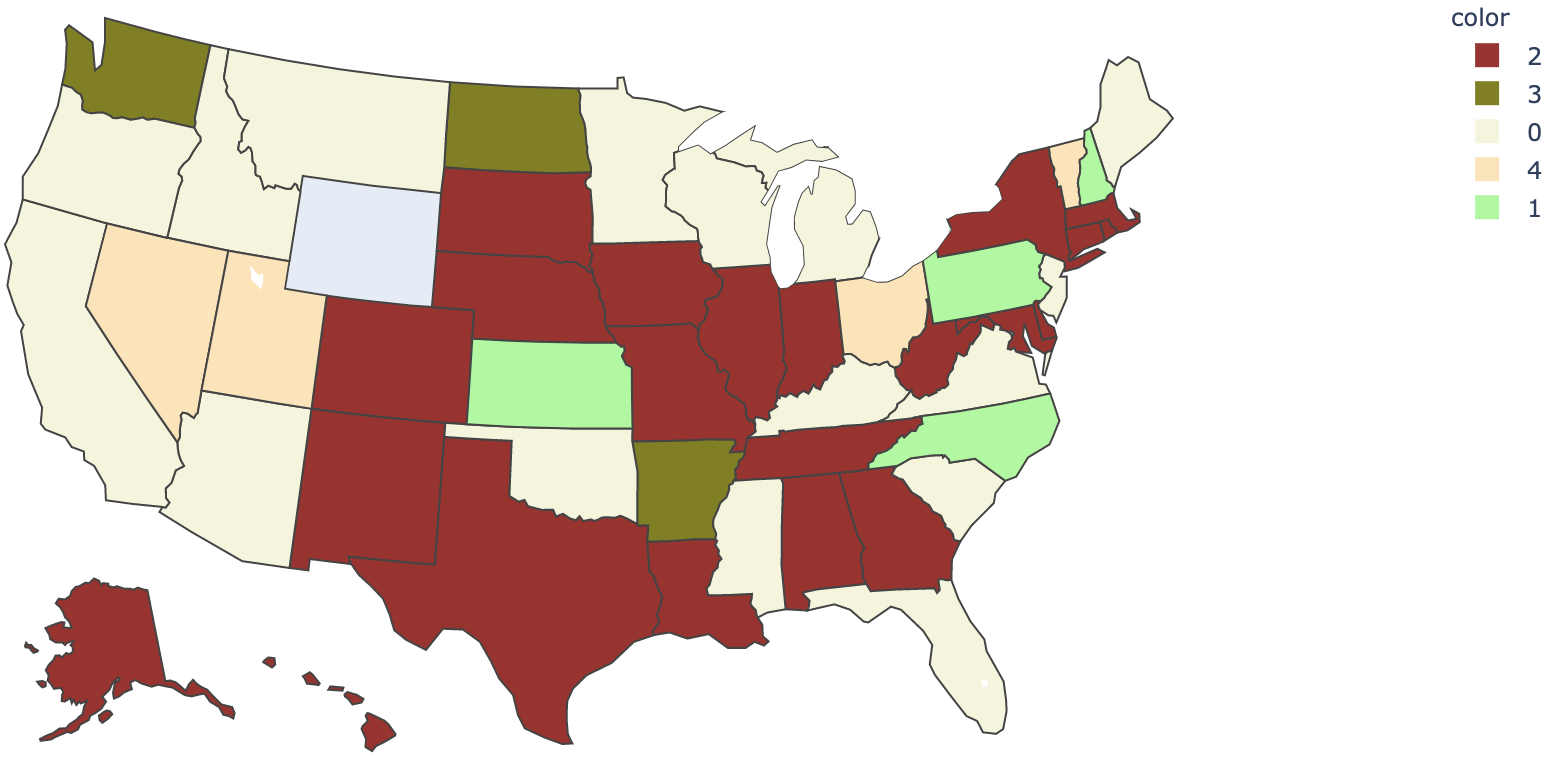
Figure 8. Clusters of States in the Second Wave (May to July 2020)

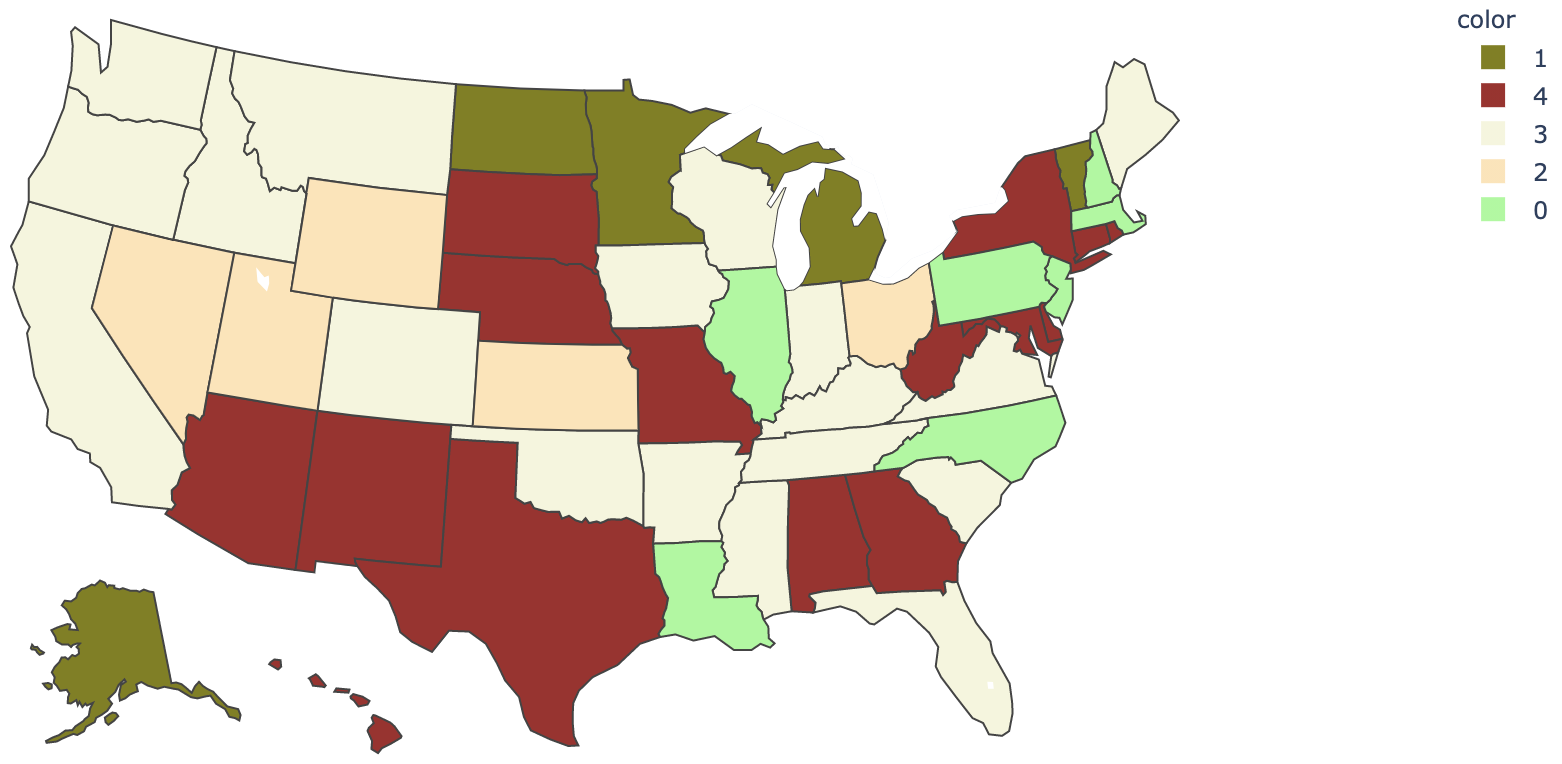
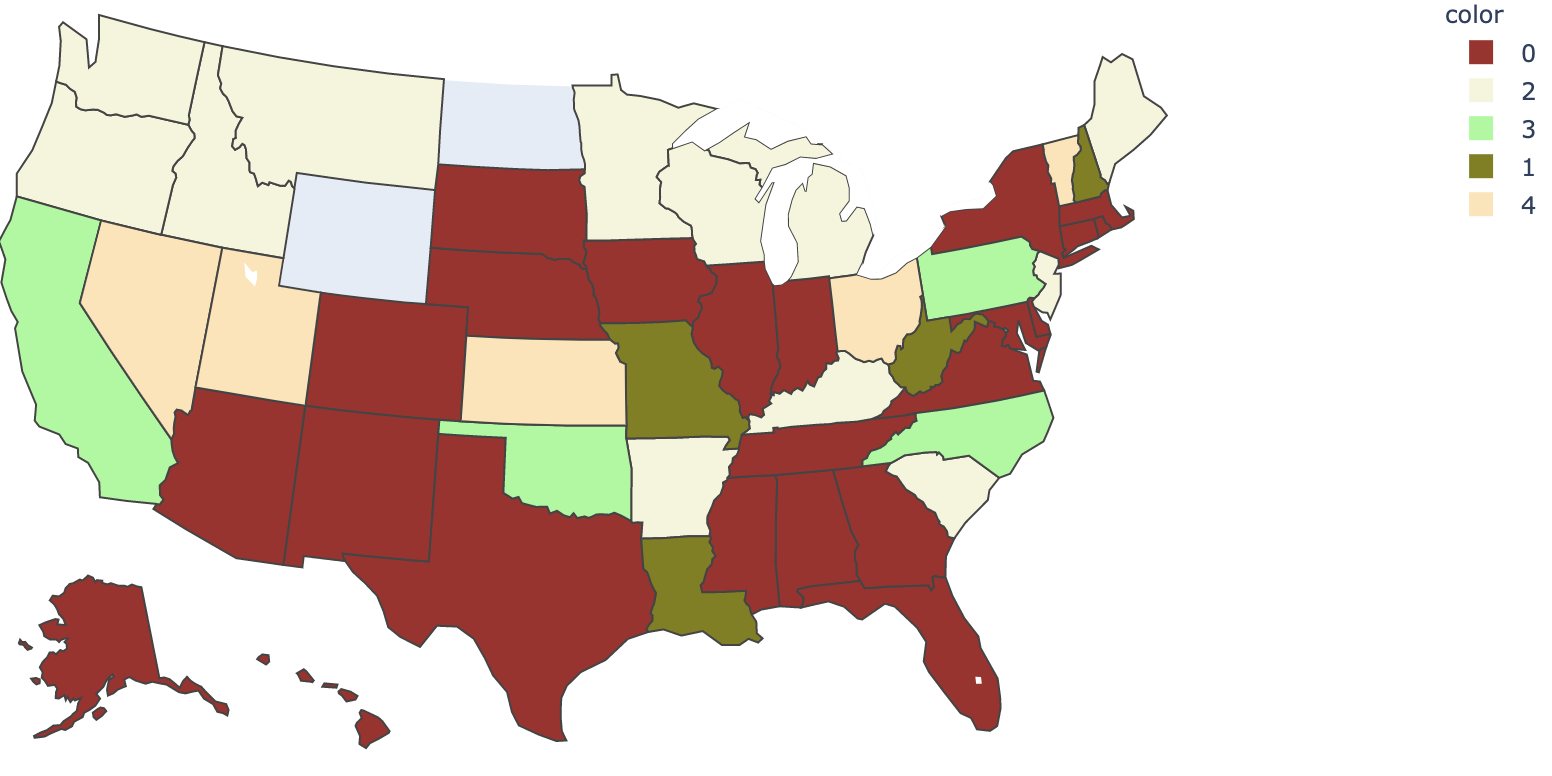
Figure9 . Clusters of States in the Third Wave (August to December 2020)

Figure 10. Clusters of States in the Fourth Wave (January to April 2021)

**F. Regression Analysis**

We noticed three variables in the dataset which we felt we could explore through statistical modeling, those variables were *death\_yn*, *icu\_yn* and *hosp\_yn*, since these were binary variables for all three variables logistic regression, linear discriminant analysis and gradient descent logistic regression were applied; there were a total of 12 regression models, as there was also a model run to test variable importance, for each y.

Before any analysis could be conducted, the dataset needed to be prepared for the regression models, so first the variables and how many NA values they have were listed, the variables with the most missing values and some redundant variables were excluded. The columns that were removed were *case\_positive\_specimen\_interval, case\_onset\_interval, underlying\_conditions\_yn, state\_fips\_code, county\_fips\_code, res\_county, exposure\_yn*. The data was still not ready, as there were many entries labeled ‘Missing’ or ‘Unknown’, these were also removed as they added nothing significant to the analysis and would impact the parameters of the model.

After the data was ready 3 independent variables and 3 response variables were created. The first response variable was *death\_yn*, indicating whether the patient died from covid-19, the rest of the dataset were the x variables. The second y variable was *icu\_yn* and the third response variable being *hosp\_yn* referencing whether the patient went to the ICU and if they were admitted to a hospital respectively. When running the first models for *death\_yn*, we discovered that all logistic regression, linear discriminant analysis and gradient descent were very accurate, scoring 0.984, 0.975 and 0.981 respectively. The strongest predictors for death were ICU and hospitalization, the third best indicator being the state of Massachusetts, almost all other variables had none, or negative likelihoods to die from covid-19. The second model was used to determine whether the individual may need to go to the ICU, the models here were quite accurate returning scores of 0.981, 0.971, 0.978. The most important variables in predicting True for this model are death, hospital, and some states. The final model has scores of 0.940, 0.939, 0.914.

**II. Conclusion**

This study investigated both individual- and state-level COVID cases provided by the CDC. Five clusters of states were found based on all the variables in the dataset. Thus, the states within the same cluster have similar demographics, exposure history, disease severity indicators and outcomes, and presence of any underlying medical conditions and risk behaviors. Further research can explore the factors behind the clustering. Why are states clustered in the same group? Why do some states have a more similar death rate than other states? Political ideologies, cultural ideologies, and social policy are probable candidates for explanation. Political preferences and cultural beliefs have been known to play a large role dictating our behaviors. Thus, states with similar cultural and political ideologies might also have a similar COVID-19 profile.

**III. Appendix**

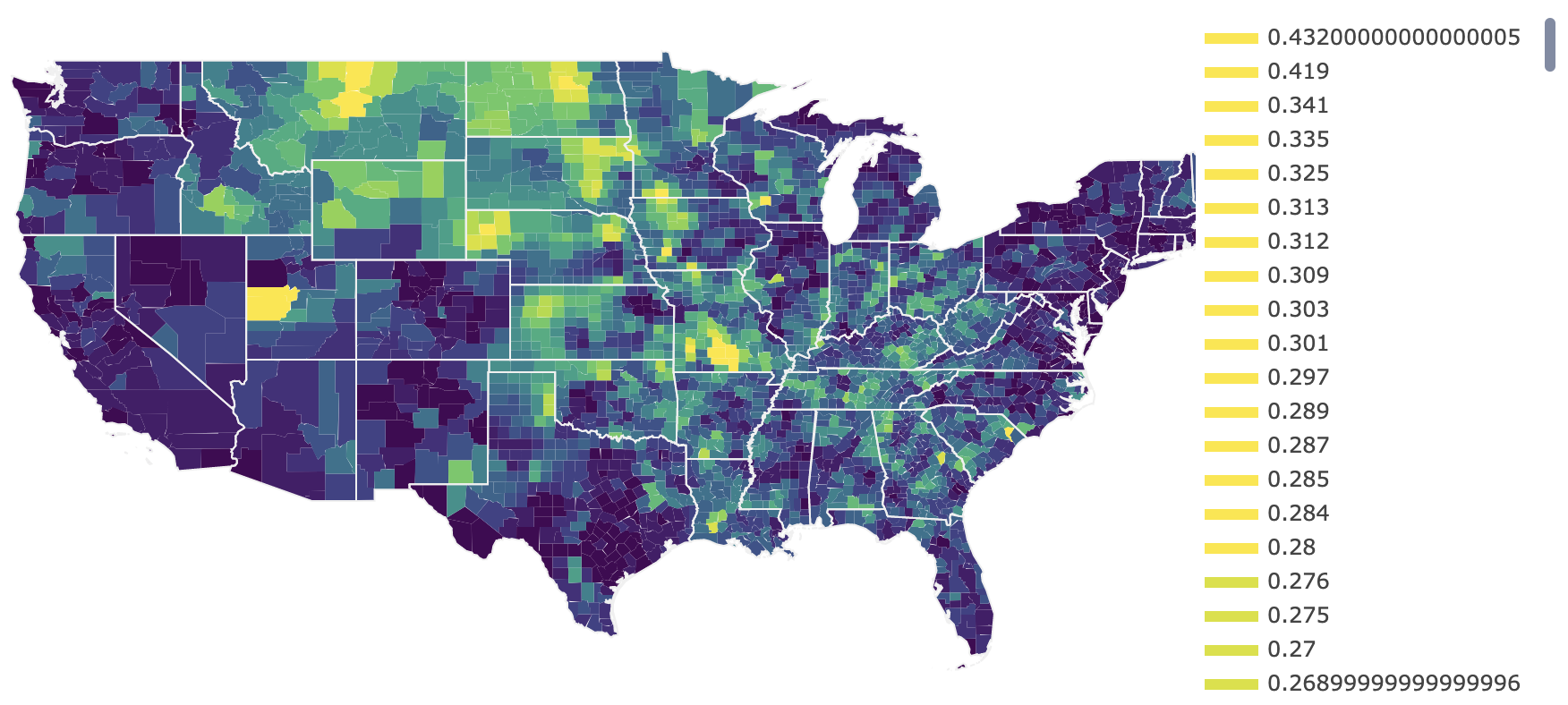
Table 1. Columns in the Dataset

|  | **Column Name** | **Description** |
| --- | --- | --- |
|  | case\_month | The earlier of month the Clinical Date or the Date Received by CDC |
|  | res\_state | State of residence |
|  | state\_fips\_code | State FIPS code |
|  | res\_county | County of residence |
|  | county\_fips\_code | County FIPS code |
|  | age\_group | Age group |
|  | sex | Sex |
|  | race | Race |
|  | ethnicity | Ethnicity |
|  | case\_positive\_specimen\_interval | Weeks between earliest date and date of first positive specimen collection |
|  | case\_onset\_interval | Weeks between earliest date and date of symptom onset |
|  | process | Under what process was the case first identified |
|  | exposure\_yn | In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel, international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/airplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case? [Yes, Unknown, Missing] |
|  | current\_status | What is the current status of this person? |
|  | symptom\_status | What is the symptom status of this person? |
|  | hosp\_yn | Was the patient hospitalized? |
|  | icu\_yn | Was the patient admitted to an intensive care unit? |
|  | death\_yn | Did the patient die as a result of this illness? |
|  | underlying\_conditions\_yn | Did the patient have one or more of the underlying medical conditions and risk behaviors: diabetes mellitus, hypertension, severe obesity (BMI>40), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other. [Yes, No, blank] |

Source: <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4>

Table 2. Missing Values in the Dataset (NA, “Missing,” & “Unknown”)

|  | **Column Name** | **Percentage** |
| --- | --- | --- |
|  | case\_month | 0.000000 |
|  | res\_state | 0.000028 |
|  | state\_fips\_code | 0.000028 |
|  | res\_county | 0.073708 |
|  | county\_fips\_code | 0.073708 |
|  | age\_group | 0.017185 |
|  | sex | 0.036406 |
|  | race | 0.414287 |
|  | ethnicity | 0.489605 |
|  | case\_positive\_specimen\_interval | 0.728504 |
|  | case\_onset\_interval | 0.534206 |
|  | process | 0.955912 |
|  | exposure\_yn | 0.932431 |
|  | current\_status | 0.000000 |
|  | symptom\_status | 0.549958 |
|  | hosp\_yn | 0.597001 |
|  | icu\_yn | 0.939207 |
|  | death\_yn | 0.507849 |
|  | underlying\_conditions\_yn | 0.935073 |

Figure 11. Choropleth of the Estimated Share of People Who Would Say “Never”

**IV. Supplementary Material**

CDC’s “COVID-19 Case Surveillance Public Use Data with Geography” dataset is accessible on the “COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography.csv” file from the following link: <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4>. The dataset is not included in the submission because the size is too large (3.53 GB). The New York Times and Dynata’s mask-wearing survey dataset is accessible on the following link: <https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv>. Finally, the programming is available on the “STA160\_Final.ipynb” file.

**V. References**

*COVID-19 Case Surveillance Public Use Data with Geography*. Centers for Disease Control and Prevention. (2021, May 24). <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4>.

*Mask-Wearing Survey Data*. The New York Times and Dynata. (2020, July 28). <https://github.com/nytimes/covid-19-data/blob/master/mask-use/README.md>.

*State Population Totals and Components of Change: 2010-2019*. United States Census Bureau. (2021, April 20). <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>.