Noor Gill
Natalia Ramirez

# An Exploratory Data Analysis of the Role of Public Policy During the COVID-19 Pandemic

## 1. Abstract:

This study was undertaken in order to examine the effects of public policy measures and the availability of resources such as healthcare workers and hospitals on virus spread and mortality rate amidst the current COVID-19 pandemic, within the United States. Using the publicly available covid19 dataset, the focus question for this exploratory data analysis is whether or not we can use prior information of precautionary measures and policies enacted in response to COVID-19 in various counties to accurately predict the spread of the virus, defined as the number of confirmed cases. We also examine the optimum availability of hospital equipment or number of frontline hospital workers in the patient's county in order to minimize the death rate. Both of these topics are trending in political media and news outlets. In order to investigate these inquiries, an extensive 5-step data cleaning was performed, numerous distinct and descriptive visualizations were developed and analyzed such as density mapping and by PCA, in addition to creating, fitting, and utilizing a logistic regression model and gradient descent algorithm. Through logistic regression, data on social distancing and precautionary measures was used to predict the spread of the virus within a given county. While gradient descent for optimizing the number of full-time hospital workers to minimize the mortality rate did not converge to the MSE, associations between features such as population served by healthcare practitioners within a HPSA and the number of hospitals or ICU beds were exhibited. This analysis, based on the data science life cycle, examines not only the role of measures such as shelter-in-place and restrictions on gatherings to "flatten the curve," but also reveals discrepancies in the allocation of resources and healthcare jobs across United States counties, which emphasizes the importance of public policy during a pandemic.

## 2. Introduction:

Our first primary question was: can we use prior information of precautionary measures and policies enacted in response to COVID-19 in various counties to predict the spread of the virus? This question was aimed at determining whether or not data of policies enacted by counties (such as shelter-in-place, limits on the number of individuals in public places, remote instruction of public schools, and restrictions on entertainment/recreational services) as well as state guidelines and federal travel bans, are sufficient to predict the number of confirmed cases within a county accurately, which we assume to be the "spread." This is a predictive question that involves regression analysis and it is of interest because, as citizens become impatient to exit “quarantine,” and return back to “normal,” the question of whether social distancing and other precautionary policies are even working becomes very important to address, especially under the sharp eyes of news and media outlets.

Our second primary question was: what is the optimum availability of hospital equipment or the number of frontline hospital workers in the patient's county in order to minimize the death rate? This question was posed to examine whether it is possible to minimize mortality and quantify the optimal availability of hospital equipment or workers necessary with features such as the number of medical doctors, healthcare workers, ICU Beds, or hospitals within a county. Doing so, allows us to look into how current economic and geographic structures and policies contribute to differences in these factors between counties. This is an optimization question that involves principal component analysis. It is of interest because the nation is undergoing a shortage of full-time healthcare workers and hospitals during this pandemic, especially in “hot spots” such as New York City. Hence, it is important to consider differences in the availability of such resources in varying countries and whether or not there is a minimum number needed in order to decrease mortality due to COVID-19, so that the appropriate public policy measures can be conducted. Through our exploratory data analysis, we also utilized guiding questions based on our visualizations and observed patterns including, how does the number of confirmed cases vary based on geographic region? Or which features account for the greatest fraction of variance in this dataset?

## 3. Description of Methods:

After intense exploration of the data through table manipulation and developing visualizations to view associations between variables, we were able to develop a step-by-step approach to answer each of our questions. In order to determine whether or not we can use prior information of precautionary measures and policies enacted in response to COVID-19 in various counties to predict the spread of the virus, we first created a merged DataFrame where each row corresponds to a different county with information about different lockdown policies as well as the day-to-day spread of the virus, measured as the count of confirmed cases (Table 1). How does the number of confirmed cases vary based on geographic region? To better understand our data through this exploratory question, we created a Geospatial Hexbin plot in order to visualize the number of confirmed cases per county in the United States. This information was helpful in that it allowed us to view the general distribution of confirmed COVID-19 cases across the United States with points representing the density of the number of confirmed cases for each county. However, we wanted to create a more specific and concrete visualization to better examine the spread of the virus - we created a line plot that depicts the 25 counties with the largest average number of confirmed cases since Jan 25th 2020. Since it may not be as uniform or impartial to determine trends in the data based on the average number of confirmed cases, we decided to quantify the number of confirmed cases with respect to the population size of the county by deriving the proportion, p = x / n, where x represents the number of successes (average number of confirmed cases) and n represents the sample size (population size of the county).

After viewing the data in many different ways, we decided that the best method for approaching this inferential question would be to perform predictive analysis with a logistic regression model since this would allow us to train a model that predicts the growth of infection in a county based on public policy data and precautionary measures. First, we allocated 10% of our data as test and 90% as training. Then, we made a function that would calculate two features, Growth Factor and Initial Value, by performing linear regression and calculating the slope and y-intercept respectively. These are the most crucial values when trying to determine the spread of a contagious virus. The growth factor suggests how many people are infected by each person who gets the virus, and the initial values refers to how many people were initially infected. We applied this function only to our training data, and this ultimately allowed us to fit our data, save our model, and predict the growth rate and initial value for our testing data. Finally, we created some functions that would allow us to predict the virus growth while comparing this to the real coronavirus cases per day that were confirmed for that county.

In an attempt to determine optimum availability of hospital equipment or number of frontline hospital workers in the patient's county to minimize the death rate, we performed table manipulation and use the pandas library to combine refined hospital worker and equipment data into a single DataFrame (Table 2), which was used to develop scatter plot visualizations of how key features affect the mortality rate across counties. But which features account for the greatest fraction of variance in this dataset? Through PCA, we hoped to find the underlying principal components that best differentiate our data points in regards to mortality rate, in addition to visualizing the relationships between these features and identifying which ones to consider in further analysis. After creating an interactive scatterplot with px.plot, we noticed an association between the average number of MDs in 2017, average number of hospitals, and population served by healthcare practitioners within a HPSA, especially since these variables were clustered in the no-noise plot. Upon developing a scree plot with this data, we noticed that the majority of the variance in the data was accounted for with three components: the average number of full-time employees at hospitals in 2017, the average number of MDs in 2017, and the average number of ICU Beds.

Based on our visualizations, assuming an association between the variable with the steepest slope in our scree plot - the average number of full-time employees at hospitals - and the mortality rate, we decided to use gradient descent with pytorch as an iterative method for optimizing the objective function relating these two variables. Starting from a random point on a function, gradient descent is useful in cases where the optimal points cannot be found by equating the slope of the function to 0, which is relevant in this case since we want to find the minimized death rate based on the number of employees in hospitals, or worker availability. After plotting and observing the overall distribution of the data with no evident trend, we developed a simple linear model and implemented the algorithm with step size as 100 and the learning rate as 1 / t. In an attempt to improve this model and decrease the loss, we reimplemented the model with step size being 500 (higher), and the learning rate as 1 / (1+t) (lower), which minimized our loss. We used a contour plot to visualize the gradient descent process and created a scatterplot with a corresponding regression line based on the predicted y values. While we were able to explore the data in depth and learn more about the features that contribute most to the variation in the death rate across counties, the gradient descent algorithm attained a local minimum and was not able to reach the global minimum in this case.

## 4. Summary of Results:

The first step of the data science life cycle after question formulation is data acquisition. The dataset, covid19, contains a README.md file with links to Github resources that describe attributes and units of measurement and four csv files contain the necessary DataFrames for this analysis: “counties,” “states,” “confirmed,” and “deaths.” However, before performing any exploratory data analysis, we performed basic data cleaning after our immediate observations. In terms of *structure*, we observed that for each of the relations, the index is set at unique numerical indices for each row. However, attributes such as "UID" or "FIPS," as seen in the states and confirmed DataFrames, could also potentially be used as primary keys. It appears that there is quantitative continuous data in columns such as "Incident_Rate," "Lat," and "Long_" as well as quantitative discrete data in columns such as "Population," "Confirmed," and "Recovered." There is also qualitative nominal data in columns such as "Country_Region" and "Province_State." In terms of the *granularity*, it can be noted that in the counties DataFrame, each record represents a county. In the states DataFrame, each record represents a state. In the confirmed and deaths DataFrames, each record represents a geographic region characterized as either a county in the United States or United States territory such as Puerto Rico and Guam. In all 4 DataFrames, quantitative values are aggregated by count. In terms of *scope*, the data covers the area of interest, which is the United States in this case. The data is also expansive as it incorporates United States territorial regions. The data covers roughly the correct time frame since this is a fairly recent dataset from January to the present day since it is continuously being updated for May 2020. The sampling frame is the United States of America for this dataset. In terms of *temporality*, the states DataFrame includes a "Last_Update" attribute, which represents the time that the data was inputted and/or updated. This is compliant with Unix/POSIX Time formatting. The confirmed and deaths DataFrames include attributes corresponding to specific dates, with times that are not unique based on index. In terms of *faithfulness*, upon observing the updated states data available online, it is evident that some NaN recovered data from 4/18 is filled in with 0 values in the updated sets, so we assume this to be the most reliable data provided. In one-hot encoding, we replace NaN values with 0, which allows us to disregard the impact of that cell in the case of calculating np.mean or np.sum. For the confirmed and death DataFrames, substituting 0 for NaN values allows us to make the assumption that no confirmed case or death was reported in that particular instance, which is reasonable considering the adaptive nature of current COVID-19 data collection.

The next step is exploratory data analysis. In creating a merged DataFrame of social distancing policies and precautionary measures as well as number of confirmed cases per day (Table 1), we noticed there were counties with null values as names so we decided to remove these counties as there was no form of identification for them and the incident could potentially represent data falsification. Moreover, for counties such as New York City and Kansas City, the values for all policy attributes were NaN, hence we also removed these indices from the DataFrame since further analysis would not be effective due to this missing data. We also noticed duplicated records or fields for counties with different numbers reported for quantitative data columns for confirmed cases, so we chose to group by county and calculate the mean of the reported quantitative values for future analysis. Since, as data scientists, we are unable to assume which of the data is more accurate or updated, we decided to aggregate it with the mean as a summary statistic. We noticed that counties with NaN data in the 'stay at home' column also have null values for the other attributes we require for analysis, so we removed these insignificant counties from the DataFrame due to this data collection discrepancy. Using this data, we created a Geospatial Hexbin plot to visualize the number of confirmed cases per county in the United States. Note that we also removed all rows for which the value of the confirmed cases was null or negative, in order to make logical sense and ensure that the visualization is concise and valid.

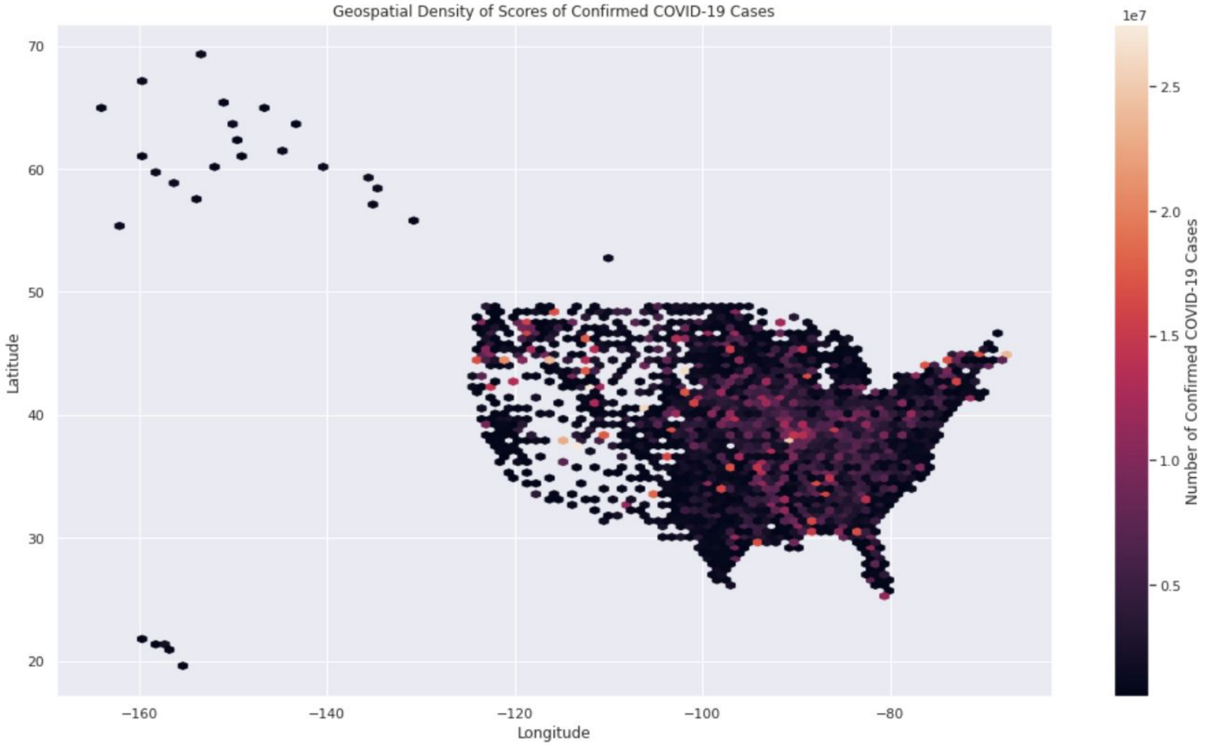

Geospatial Density of Scores of Confirmed COVID-19 Cases
Latitude
70
60
50
40
30
20
Longitude
−160
−140
−120
−100
−80
1e7
2.5
2.0
1.5
1.0
0.5
Number of Confirmed COVID-19 Cases

Fig 1.1: Geospatial Density of Scores of Confirmed COVID-19 Cases - Using data from Table 1, the density of the count of confirmed cases is depicted with data points based on United States counties.

Based on Fig 1.1, it appears that while most counties have a similar range of confirmed cases in this time period, the counties located in the Northeast United States have a higher overall number of confirmed cases. It can be noted that, in the Western United States, there are counties present with almost no cases confirmed, most likely in rural regions. There are also outlier counties with very high numbers of confirmed cases dispersed throughout the plot. Then, we aggregated the DataFrame to have each date represented per county along with the number of total confirmed cases for that date; we convert the Date column to be DateTime for uniformity (Table 3). In order to plot the 25 counties with the largest number of confirmed cases since Jan 25th 2020, we quantified the number of confirmed cases relative to the population size of the county to account for variations between counties.

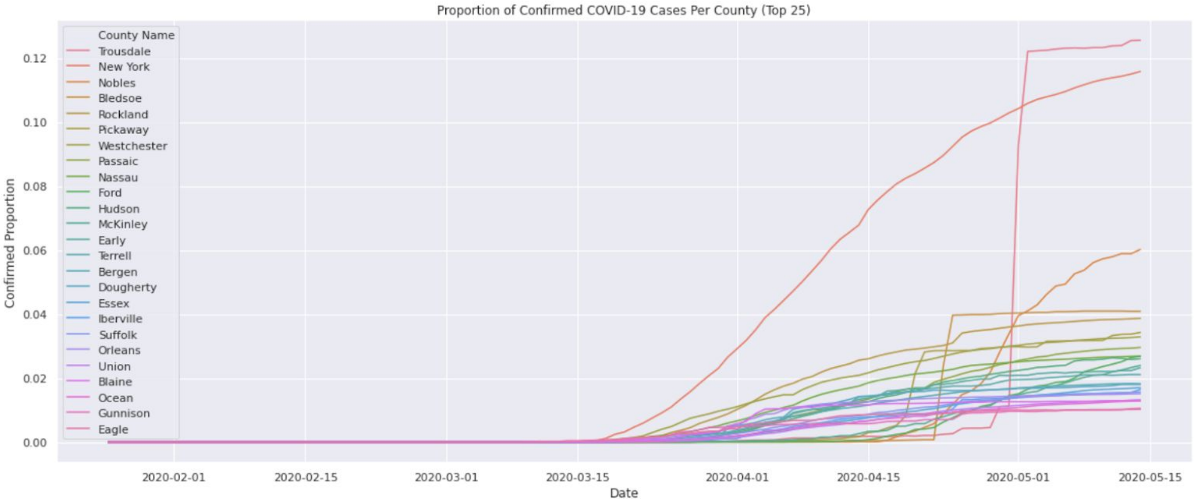


Fig 1.2: Proportion  of Confirmed COVID-19 Cases Per County (Top 25) - Using Data from Table 3, the confirmed proportion of Coronavirus cases, with respect to each county's population size, are plotted against the date for each county record. The top 25 counties with the greatest number of confirmed cases are reported for brevity.

Based on Fig 1.2, it is evident that New York County, which consists of the Manhattan district of New York, has the greatest number of confirmed COVID-19 cases out of all American counties with a drastic difference. Los Angeles county has the second greatest number of confirmed cases, followed by Westchester County. Manhattan is the most densely populated of New York City's 5 boroughs and Los Angeles is also a crowded urban region, so transmission is more likely to occur, and may occur at a faster rate due to overcrowding and congestion, which is why these results make logical sense. However, the data portrays a peculiar trend with Trousdale County taking the lead as of May 1st, 2020. This sharp increase in the proportion of confirmed cases leads us to believe this may be a discrepancy in data collection since this drastic change in the pattern of the data does not make logical sense. However, Trousdale County is a fairly small county in Tennessee, with a population size of about 7,816, which may justify this pattern since a smaller value for population size would lead to a larger proportion value with the roughly the same value for number of confirmed cases. In the logistic regression, we make sure that the same counties, or indices, that appear in countyCASES are also in countyDEATHS and vice versa so that each row corresponds to one unique county, for accuracy purposes prior to the test-train split. Before creating our model, we add a data visualization that will help us better understand the spread of this virus. The original "cases" data contained a group of columns that shows how total cases per county increased over the past few months. We graphed this in both log scale and linear scale.

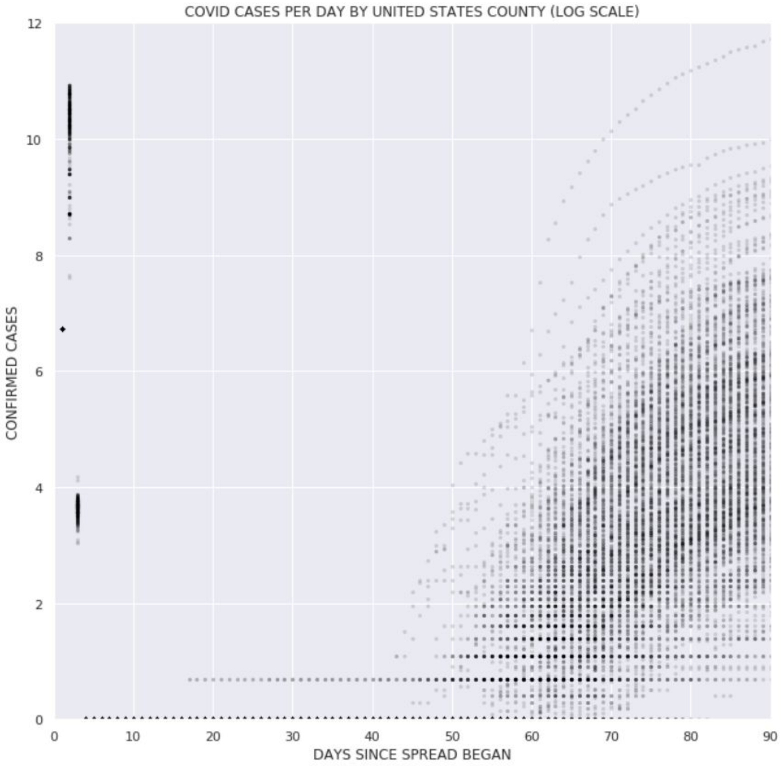


Fig 1.3: COVID-19 Cases Per Day By United States County - This is a plot of the cases per day for each county, plotted on a logarithmic scale which we achieved by applying np.log to every cell in the dataframe.

In Fig 1.3, it is apparent that each individual county is growing in terms of total cases daily, but more importantly, they seem to be concaving down. In other words, the second derivative of the function that models this growth seems to be growing less than exponentially; if the cases per day growth were perfectly exponential, plotting it on a logarithmic scale would produce linear functions with second derivatives of 0. This goes against our prior knowledge that viruses spread exponentially, however it is not immediately clear why this is. One explanation could be that we simply have not allowed COVID-19 enough time to spread and display its growth behavior. Another example could be a lack of testing -- while it is very convenient that we are working with such raw and new data, the downside is that our model will be limited in the same way that the data collection process for this dataset was limited. In other words, scarce testing in the United States could be a prominent reason why the virus’ exponential spread is not evident.

As we know, viruses grow exponentially. therefore, we apply log to every element of the dataframe when you call slopeandinterceptfeature on it. Then, the slopeandinterceptfeature function uses linear regression to calculate the slope and intercept. However, we want to undo the log we applied on the data earlier in this cell, we write a short function that undoes the log we applied to our data earlier, x(t) = a * b^t.  This model is effective to predict how the virus will spread throughout a given county. This does not necessarily equate to “reported cases per day”, as was given. Instead, our model focuses on predicting two things: the Growth Factor, and the Initial Value. In the function x(t) = a*b^t, the Growth Factor is b and the initial value is a.

In the creating a combined DataFrame of mortality rate, hospital equipment data, and hospital worker data (Table 2), we rename variables for easier identification, ensure that all county indices are unique (there are no duplicate county entries), round all values to the nearest whole number as it does not make logical sense to have decimal values for the mean of these attributes, add proportion columns, and make sure to only include data with mortality values greater than 0 to examine. In using PCA to determine which features account for the greatest fraction of variance in this dataset, we first  "center" the data so that the mean of each feature is 0. After we created the scatter plot corresponding to the first two principal components, we saw that there are overlapping observations in this scatterplot. To get a better visualization, we want to introduce some noise into the plot, which will distinguish points without affecting the overall structure of the plot and use the plotly library to label these points.

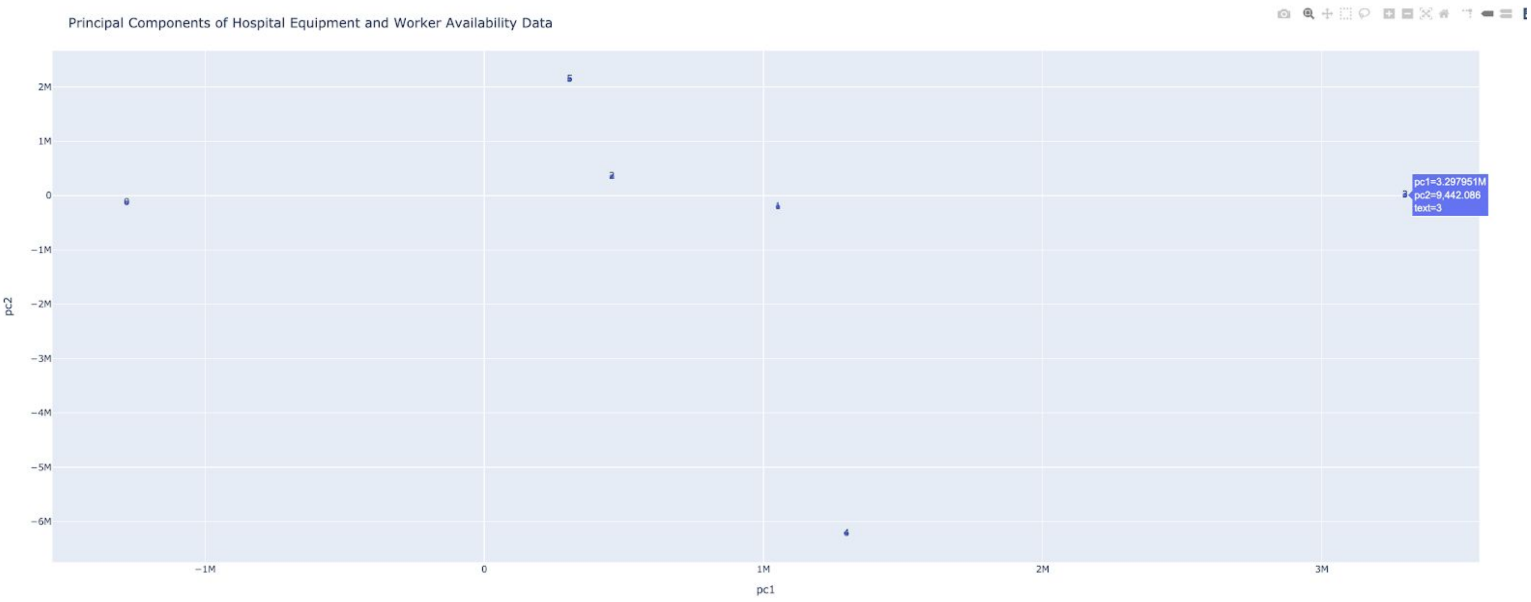


Fig 2.1: Principal Components of Hospital Equipment and Worker Availability Data - Starting with the index 0, each number label corresponds to an attribute specified in our data features array, or 'Avg Num of Full-time Employees at Hospitals (2017)', 'Avg Num of MDs (2017)',

'Avg Num of ICU Beds', 'Avg Num of Hospitals', 'Population Unserved by Full-time Equivalent Healthcare Practitioners within a Health Professional Shortage Area', and 'Population Served by FTE Healthcare Practitioners within a HPSA,' respectively.

Based on Fig 2.1, we can conclude that variables 1, 2, 3, and 5, or the average number of MDs in 2017, average number of hospitals, population served by healthcare practitioners within a HPSA appear to be most closely related, especially since these variables displayed clustered in the no-noise plot. These variables seem to have a high pc2 value but correspond to low pc1 values. On the other hand, the variables 0 and 4, or average number of full-time employees at hospitals in 2017 and the population unserved by healthcare practitioners within a HPSA seem to have higher pc1 values and the average number of full-time employees at hospitals even has a high value for pc2. Since the pc1 passes through the average, a high pc1 value means that there is a low value for largest possible explained variation in the data. Since the pc2 is interpreted in the same way, the pc2 is the best of what's left. To visualize the weight of each principal component, we created a scree plot.

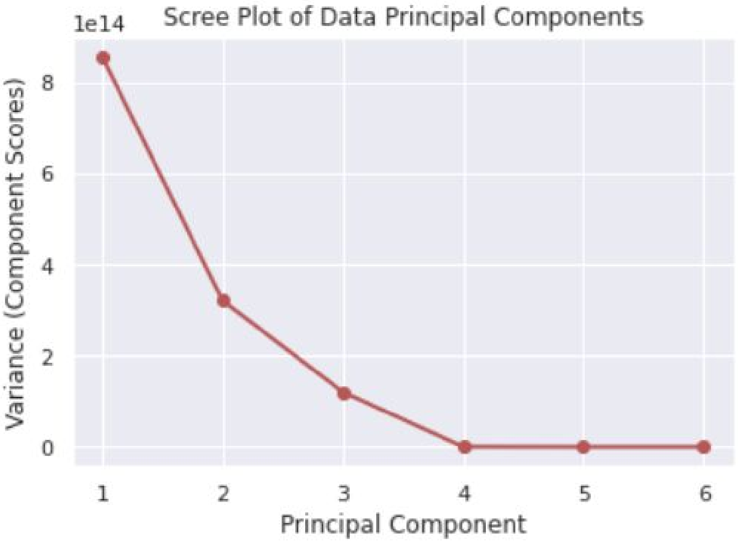


Fig 2.2: Scree Plot of Data Principal Components - The variance, or component score, that is described is plotted against each of our 6 principal components in focus based on Table 3. The steepest slope is represented by pc1, followed by pc2 and pc3, until there is an evident plateau for pcs4, pc5, and pc6.

This scree plot entails that the variance is captured the greatest increase by the first three components, before it tapers off with the inclusion of the fourth component. Therefore, we want to focus on the first three components. When performing the gradient descent, to account for any variations and make the data more consistent, converting the tx and ty to a standard unit of measurement provides a reference point for describing objects and patterns. After our first iteration, in order to decrease the loss, we increased step size and decreased learning rate with a revised model, which resulted in a minimized loss. We used a contour plot in order to visualize this gradient descent process.

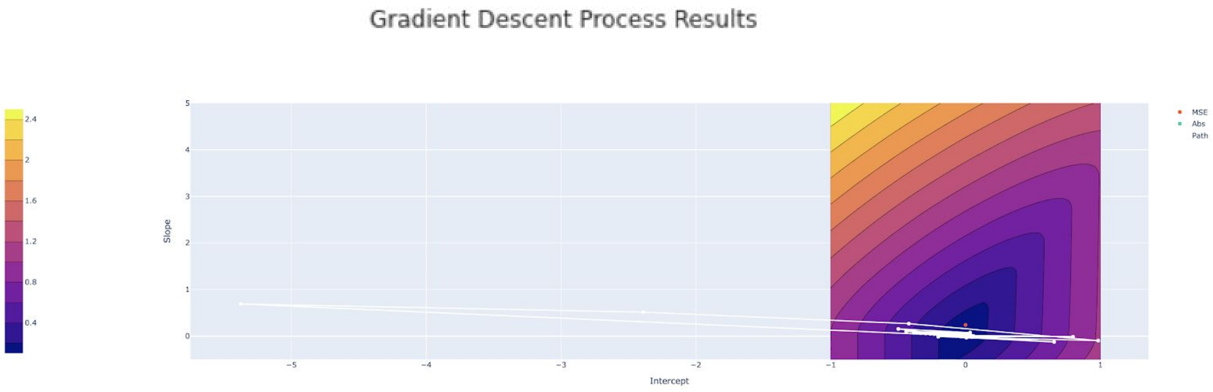


Fig 2.3: Gradient Descent Execution - The contour plot depicts the gradient descent algorithm and its path with respect to the slope and intercept. As visible, the gradient descent fails to converge to the MSE value as it reached a local minima rather than the global minima.

As per Fig 2.3, it appears that the gradient descent algorithm did not converge to the MSE value. Based on this gradient descent, we can conclude that the minimized loss is equal to about 0.034963201320518514, which is a fairly low value. The intercept is about -0.02237895 and the slope is about 0.00049639. While this does not directly answer our question, it leads us to conclude that the value at which the death rate is minimized, as per gradient descent, is 0.02237895 number of full-time hospital employees per county in 2017. This value does not make logical sense. We recall that if a gradient descent algorithm attains local minimum, it is nearly impossible to reach global minimum, which is most likely to have happened in this case, also since the scatterplot above depicts a regression line that does not seem to accurately account for the variations in the data. While we did not necessarily answer this question we were able to explore the data in depth and learn more about the features that contribute most to the variation in the death rate across counties.

## 5. Discussion:

The next step in the data science life cycle is prediction and inference. Variations in the shelter-in-place restrictions between counties within a state, and even differences between cities are important to consider during the building of our model. In fact, we were interested in seeing how successful lockdown regulations were in slowing down the spread of Coronavirus, so we used several government regulation and lockdown related features while fitting our model, so that we could predict the spread of COVID-19 based only on these features. Building this model took several nested functions. First, we defined slope and intercept feature, which takes in our training data frame in which every row is a county. Then, it looks *only* at the reported cases per day. First, we take the log of every element in the training data. Then, we use np.polyfit to fit the growth of the cases over time. After fitting the data row by row, we attain "slope" and "intercept" from the polyfit method. Then, the function returns another function called apply exponential, which exponentiates the dataframe, hence undoing the logs we took earlier. We rename slope and intercept as "Growth Factor" and "Initial Value" respectively. With these two values, we can model the growth of any county in our training data. This function returns our dataframe, with our original features plus our two new features.

The last step to our model is predicting the Growth Factor and Initial Value of counties that we do not have empirical case per day data for. Thus, we use go on to the last step, predicting using our testing data. We define graphmodelprediction, which takes in a single county for which we will be predicting the spread of COVID-19. Perhaps most importantly, this function takes in a list of columns that we believed would provide our model with the most influential features. This makes logical sense, given that lockdown procedures are put in place with the intention of slowing down viral growth, so these features are likely relatively influential and relevant compared to others. After we have predicted our Growth Factor and Initial Value, we must simply input it into the viral spread function we defined earlier, (initial_value * (growth_factor ** x)).

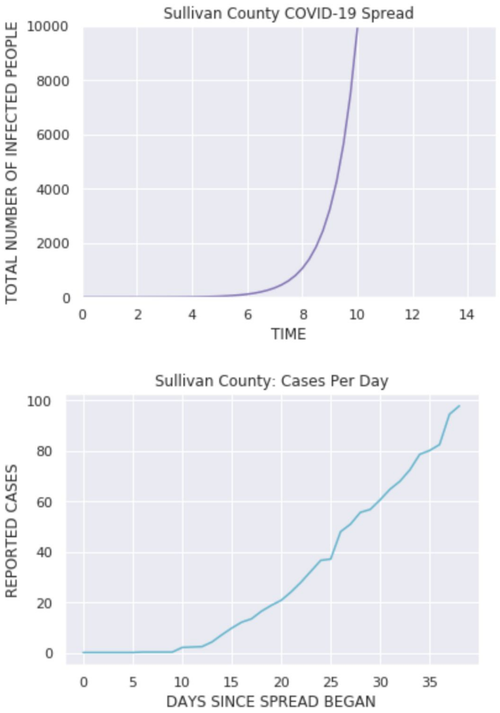


Fig 2.4: For this visualization, we developed a function named graphmultiple that graphed two functions side by side: our COVID-19 model prediction, as well as the daily case data, so we can see how the empirical data looks when compared to the theoretical growth model.

As of May 4th, 6 Bay Area counties have jointly agreed to ease some restrictions of the ongoing shelter-in-place orders during the pandemic by allowing businesses such as construction, retail, nurseries, and landscaping to resume as lower risk so there are differences between counties within a state. Also, some counties may have more data science professionals or volunteers that collect data, while rural regions may have fewer such professions, resulting in missing (NaN) inputs, which decreases the accuracy as well. Moreover, since the CDC "draws from a combination of data sources from existing influenza and viral respiratory disease surveillance...ongoing research platforms, and other new systems designed to answer specific questions," it is possible that pre-existing biases or inaccuracies in these datasets and systems are perpetuated in this dataset, affecting the output of our predictive model. We were able to answer our initial question by creating a model that predicts the estimated number of confirmed cases - the spread - based on data of precautionary measures and quarantine policies. Perhaps in the future, adding mask or glove-use data would allow us to refine our model and increase accuracy if there is a statistically significant association between use of these resources and the number of cases within counties.

In the first model, the most pivotal feature was ironically not related to lockdown regulations-- it was likely the 'PopulationDensityperSqMile2010' feature. While we did want to keep our features almost exclusively related to government social distancing rules, we found that it was necessary to include a couple features related to population, and most importantly, population density, as this is ultimately the biggest indicator of how fast a contagious virus like COVID-19 will spread. During the construction of this model, we ran into several issues. For instance, because the features we were trying to predict are not categorical, we could not calculate accuracy the way we would have been able to if we had built a classifier. Therefore, the best we could do to confirm we were on the right track was to compare the predicted features to one another, and to further attributes of the county, and determine if it makes sense. Other difficulties during this stage in our program were choosing the right type of regression to do (logistic vs linear), and knowing when and what to take the log of and exponentiate. One of the wrong assumptions  that was made while writing the model was that the graph we were predicting should look similar to the cases per day data that was provided to us. Hence, it was very frustrating when the model appeared so different from the data we had been looking at earlier. One of my first mistakes is that I was looking at the X axis as "days" rather than "time". This leads to one of our model's weaknesses, which is that it is unclear how we can quantify these results, since there is no specific way to translate "time" in our model into "days", although this would probably be very useful if it was possible, especially amidst a pandemic.

In the second model, while we were unable to determine the optimum availability of hospital equipment or the number of frontline hospital workers in the patient's county in order to minimize the death rate, we were able to minimize the death rate with respect to the number of full-time healthcare workers in a given county. Based on our PCA analysis, we were able to observe some clustering between the average number of MDs in 2017, the average number of hospitals, and the population served by healthcare practitioners within a HPSA. When the value of one of these variables increases or decreases, the value another variable tends to change in the same way. Since the variance in the data is captured by the greatest increase by the number of full-time healthcare workers in hospitals, we used this as the explanatory variable where the mortality rate served as the dependent variable in conducting gradient descent. Consequently, we were able to tailor our model and obtain a minimized loss of about 0.034963201320518514, which is a lot lower than our previous attempt by about 0.04 units. The slope and intercept corresponding to this loss are -0.02237895 and 0.00049639, respectively. While this does not directly answer our question, it leads us to conclude that the value at which the death rate is minimized, as per gradient descent, is 0.02237895 number of full-time hospital employees per county in 2017. This value does not make logical sense. We recall that if a gradient descent algorithm attains local minimum, it is nearly impossible to reach global minimum, which is most likely to have happened in this case, also since the scatterplot of the relationship between the independent and dependent variable depicts a regression line that does not seem to accurately account for the variations in the data. Since the mean squared error term has a minimum at an intercept of about 100, we can conclude that about 100 is the average squared difference between the estimated values and what is estimated, in terms of the intercept.

In terms of the most interesting features we came across for this question, I did not expect for there to be any sort of association present between the average number of MDs in and average number of hospitals and the population served by healthcare practitioners within a HPSA since one would expect that counties with smaller populations have fewer hospitals and fewer doctors so the proportion of the population of a county that is served is not affect much. One feature that we thought would be useful is the number of hospitals in a county, but this turned out to be ineffective relative to the other variables in the DataFrame since it did not account for a significant portion of the variance in the results. We were slightly stuck in the process of comparing units of measurement and quantitative values in our columns, especially after computing the mean, but we referred to the dataset description to resolve this confusion. A challenge that we found with our data was ensuring that all units of measurement are accounted for and, in order to resolve this, we used proportions for EDA rather than numbers since we want to compute results with respect to county population to have some sort of standardization. However, if redone, we would also use proportions in executing the gradient descent algorithm to account for this difference between counties. Some of the limitations of the analysis performed include that it was not effective at computing the optimal value of the explanatory variable in order to minimize the dependent variable, as only the latter portion of this aim was completed. Also, since the gradient descent algorithm reached the local minimum rather than the global minimum, utilizing stochastic gradient descent would have possibly been more effective as it introduces noise. An assumption that we made that proved to be incorrect was that replacing the NaN values with 0 would still provide an accurate result. No data is not the same as the value

0. This could have led to skewed results. In this model, however, we did not include counties where the death toll was equal to 0 so this may not be as reflective in our results.

An ethical dilemma we faced with this data was assuming that data in all counties were collected uniformly with no bias introduced by volunteer/paid data collectors/analysts. However, as an ethical issue, it could be plausible that in some counties, no confirmed cases could be reported as 0, while in others they could be reported as NaN, which could lead to discrepancies in the validity of the predictive model. In studying this problem, there may be ethical concerns in that the number of hospital workers or the quantity of hospital equipment could be controlled by political structures and economic barriers such as licensing restrictions or budget cuts in the healthcare industry, so not every county can be judged on the same scale. In order to address these concerns, it is important to propose efforts to increase the availability of these resources on larger state-wide or federal scales. This can be done by lobbying and virtually protesting for public policies in favor of the healthcare industry in battling against this pandemic and "flattening the curve," in this case, the mortality curve.

## 6. Conclusion:

Since the data we used for hospital worker availability, from 2017, is roughly 3 years old, our analysis would be stronger with the use of 2020 data on the number of MDs within counties and the number of full-time healthcare workers per county. From 2010 to 2016, the actively licensed US physician-to-population ratio increased from 277 per 100,000-population to 295 per 100,000-populations and we expect this increasing trend to have continued past this time period with all other variables held constant. This additional data would allow us to make more valid conclusions for the optimization of hospital workers and equipment in minimizing mortality, with a sample that is more representative of the population at this time. Moreover, since the data was collected in an observational manner, we can only determine associations between variables, and cannot claim causality, which is an inevitable restriction in using this dataset.

The logistic regression method used to answer our first question proved to be a very powerful yet limited method. The benefit to this method was that we were relying on a very accurate model of viruses in general, x(t) = a * b^t, which most contagious viruses conveniently follow. Unfortunately, though, it relies partly on estimation to determine the accuracy, since we cannot find RMSE or exact accuracy since it is not exactly a classifier or a regressor.

On the other hand, the use of the gradient descent algorithm in minimizing the mortality rate and quantifying an optimal value for the number of full-time hospital workers within a county did not prove as effective since it is difficult to obtain specific data such as the optimal value with such a convoluted dataset with large quantities of missing, nullified data, hence it is harder to make accurate assumptions. While it is tougher to obtain optimized hospital equipment or workers counts in compliance with our focus question, the gradient descent method would be effective for determining the minimized mortality rate based on the number of full-time hospital workers if we incorporated the stochastic gradient algorithm which behaves like a simulated annealing algorithm, such that the learning rate of the SG is related to the temperature of SA. The randomness or noise introduced by SG would allow us to escape from local minima to reach a better minimum, which was a limitation of our current model.

The final step of the data science life cycle returns to question and issue formulation. For the future, we could elaborate upon the role of the public sphere and politics in the spread of a virus such as COVID-19 through questions such as, how does the political party of the authoritative figures in states, such as the governor, affect confirmed case count? To answer this question, we can use additional data on the political party with which each governor is associated with, as well as corresponding actions and measures in order to predict the spread of the virus through a regression model. Although the political realm is a controversial one, this would be an interesting question to pose that could affect the results of the upcoming state-level or federal elections, since historically, citizens tend to vote conservative during times of national crises or uncertainty. Or we can utilize SVI Percentile data to investigate, how does a county's socioeconomic composition relate to death rate within that county? For this analysis, we can utilize SVI data such as SVIPercentileSEtheme and SVIPercentileHDtheme from the counties DataFrame in order to quantify the composite socioeconomic status of a particular county relative to others so we can develop a regression that determines whether or not there is an association between these variables and mortality rate. This has been a trending topic in the current media, as underprivileged citizens from minority group backgrounds claim that there is prejudice present in-hospital treatment for critical condition COVID-19 patients.

## Bibliography

CDC. “FAQ: COVID-19 Data and Surveillance.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 17 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html.

Reyes, Kris. “6 Bay Area Counties Relax Some Shelter-in-Place Restrictions, Certain Businesses to Reopen May 4.” *ABC7 San Francisco*, 4 May 2020, abc7news.com/bay-area-counties-extend-shelter-in-place-sf-dates-6-issue-extende/6148415/.

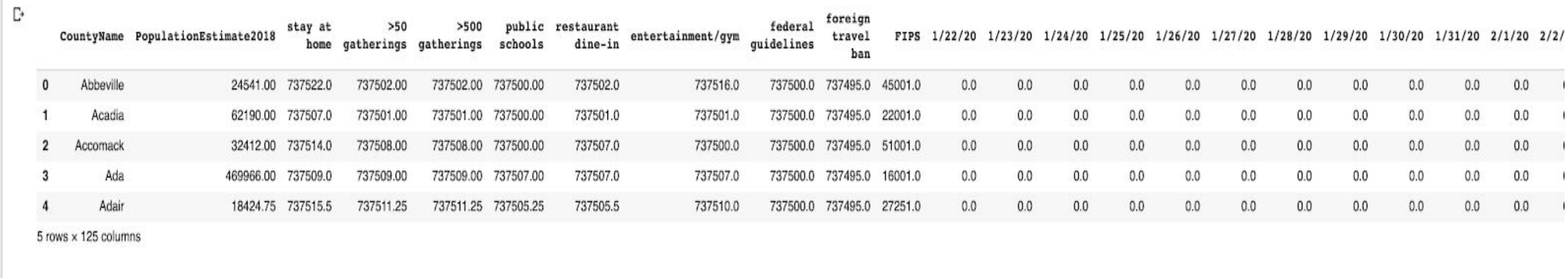
Saint Jacques , Al. “There Are Now 953,695 Actively Licensed Physicians in US, According to Physician Census.” *MDLinx*, www.mdlinx.com/article/there-are-now-953-695-actively-licensed-physicians-in-us-according-to-physician-census/lfc-994.

“U.S. Census Bureau QuickFacts: Trousdale County, Tennessee.” *Census Bureau QuickFacts*, www.census.gov/quickfacts/trousdalecountytennessee.

“World Population Review: Manhattan Population 2020.” *Manhattan Population 2020*, worldpopulationreview.com/boroughs/manhattan-population/.

# Appendix

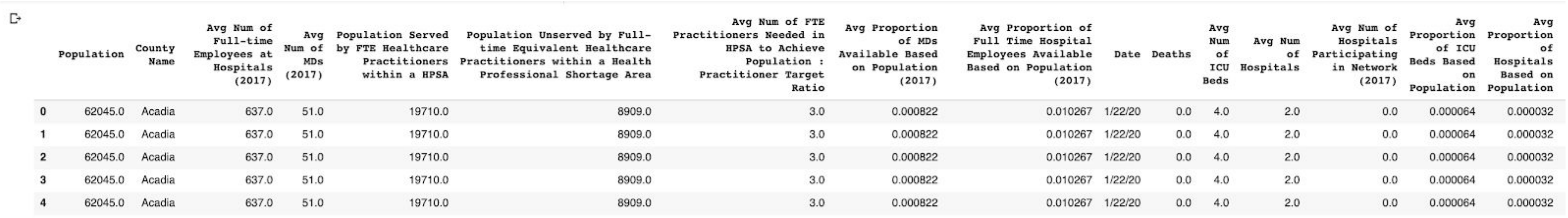
Table 1: Social Distancing and Precautionary Policies Alongside Count of Confirmed Cases Per Day



| | CountyName | PopulationEstimate2018 | stay at home | >50 gatherings | >500 gatherings | public schools | restaurant dine-in | entertainment/gym | federal guidelines | foreign travel ban | FIPS | 1/22/20 | 1/23/20 | 1/24/20 | 1/25/20 | 1/26/20 | 1/27/20 | 1/28/20 | 1/29/20 | 1/30/20 | 1/31/20 | 2/1/20 | 2/2/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Abbeville | 24541.00 | 737522.0 | 737502.00 | 737502.00 | 737500.00 | 737502.0 | 737516.0 | 737500.0 | 737495.0 | 45001.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| 1 | Acadia | 62190.00 | 737507.0 | 737501.00 | 737501.00 | 737500.00 | 737501.0 | 737501.0 | 737500.0 | 737495.0 | 22001.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| 2 | Accomack | 32412.00 | 737514.0 | 737508.00 | 737508.00 | 737500.00 | 737507.0 | 737500.0 | 737500.0 | 737495.0 | 51001.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| 3 | Ada | 469966.00 | 737509.0 | 737509.00 | 737509.00 | 737507.00 | 737507.0 | 737507.0 | 737500.0 | 737495.0 | 16001.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| 4 | Adair | 18424.75 | 737515.5 | 737511.25 | 737511.25 | 737505.25 | 737505.5 | 737510.0 | 737500.0 | 737495.0 | 27251.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |

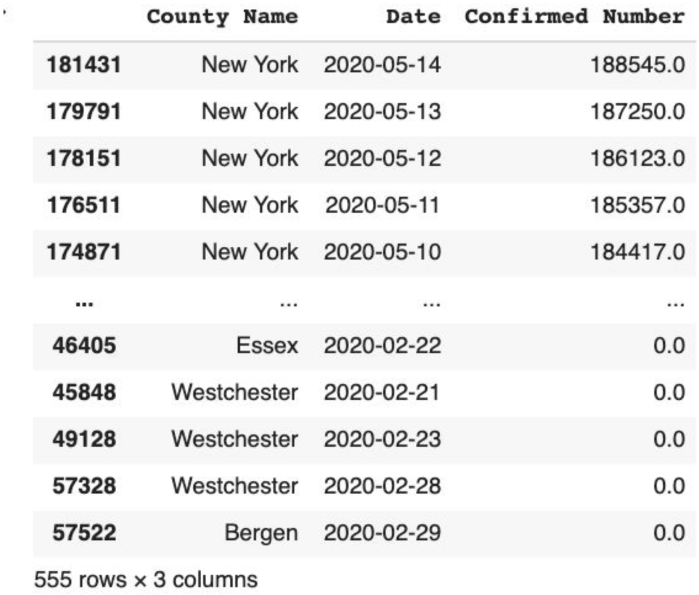
5 rows × 125 columns

Table 2: Hospital Worker and Equipment Data Alongside Death Toll



| | Population | County Name | Avg Num of Full-time Employees at Hospitals (2017) | Avg Num of MDs (2017) | Population Served by FTE Healthcare Practitioners within a HPSA | Population Unserved by Full-time Equivalent Healthcare Practitioners within a Health Professional Shortage Area | Avg Num of FTE Practitioners Needed in HPSA to Achieve Population : Practitioner Target Ratio | Avg Proportion of MDs Available Based on Population (2017) | Avg Proportion of Full Time Hospital Employees Available Based on Population (2017) | Date | Deaths | Avg Num of ICU Beds | Avg Num of Hospitals | Avg Num of Hospitals Participating in Network (2017) | Avg Proportion of ICU Beds Based on Population | Avg Proportion of Hospitals Based on Population |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 62045.0 | Acadia | 637.0 | 51.0 | 19710.0 | 8909.0 | 3.0 | 0.000822 | 0.010267 | 1/22/20 | 0.0 | 4.0 | 2.0 | 0.0 | 0.000064 | 0.000032 |
| 1 | 62045.0 | Acadia | 637.0 | 51.0 | 19710.0 | 8909.0 | 3.0 | 0.000822 | 0.010267 | 1/22/20 | 0.0 | 4.0 | 2.0 | 0.0 | 0.000064 | 0.000032 |
| 2 | 62045.0 | Acadia | 637.0 | 51.0 | 19710.0 | 8909.0 | 3.0 | 0.000822 | 0.010267 | 1/22/20 | 0.0 | 4.0 | 2.0 | 0.0 | 0.000064 | 0.000032 |
| 3 | 62045.0 | Acadia | 637.0 | 51.0 | 19710.0 | 8909.0 | 3.0 | 0.000822 | 0.010267 | 1/22/20 | 0.0 | 4.0 | 2.0 | 0.0 | 0.000064 | 0.000032 |
| 4 | 62045.0 | Acadia | 637.0 | 51.0 | 19710.0 | 8909.0 | 3.0 | 0.000822 | 0.010267 | 1/22/20 | 0.0 | 4.0 | 2.0 | 0.0 | 0.000064 | 0.000032 |

Table 3: Number of Confirmed Cases Per Date Per County



| | County Name | Date | Confirmed Number |
|---|---|---|---|
| 181431 | New York | 2020-05-14 | 188545.0 |
| 179791 | New York | 2020-05-13 | 187250.0 |
| 178151 | New York | 2020-05-12 | 186123.0 |
| 176511 | New York | 2020-05-11 | 185357.0 |
| 174871 | New York | 2020-05-10 | 184417.0 |
| ... | ... | ... | ... |
| 46405 | Essex | 2020-02-22 | 0.0 |
| 45848 | Westchester | 2020-02-21 | 0.0 |
| 49128 | Westchester | 2020-02-23 | 0.0 |
| 57328 | Westchester | 2020-02-28 | 0.0 |
| 57522 | Bergen | 2020-02-29 | 0.0 |

555 rows × 3 columns