COVID-19 Data Analysis

Abstract

For our final Data100 project, our group chose to work on the COVID-19 dataset. In order to break down the project, we came up with four major questions that we wanted to explore which we will refer to by its corresponding number throughout the paper:

(1) Which factors most strongly influence death rates?
(2) Which states are more drastically affected than others?
(3) Are areas with more health resources more likely to have lower death or hospitalization rates?
(4) Which features best predict death rates?

To do this, we performed exploratory data analysis, cleaned the data, and built a model to predict death cases. We approached Question (1) with hypothesis testing and found that hospitalization rates affect death rates most. We tackled Question (2) by manipulating the datasets and focusing on determining mathematically and graphically which states seemed to have higher deaths and case counts over time. This fueled our response to Question (3) as our results motivated us to explore and question if it was the diversity of health resources that may have affected counties differently with the death and confirmed case rates. We did this by focusing on features that could tell us the most about health resources and hospital accessibility to see if we could reveal a potential trend. We approached Question (4), the modeling component through a variety of methods. Our best one uses a multiple linear regression model with K-Folds cross-validation and found that the best features to predict future death rates were hospitalization rates and testing rates. In this paper we will be doing in depth to share our findings and process to draw these conclusions.

Introduction

We chose to work on this dataset since we are currently facing a global pandemic, so we wanted to learn more about the world we are currently living in and see if we can use data science to find interesting trends in the COVID-19 data. We hear a lot about the virus in the news, but unfortunately there is such vast misinformation spread about fake cures and trends and remedies that it’s difficult to determine what’s real vs what isn’t. We wanted to see if we could confirm some of the intuition we had when we first got our hands on the dataset as well as identify any conflicts with some of our previous assumptions about the spread and impact of COVID-19. We found it interesting to try to draw some of our own conclusions and explore the impacts from a state level to a county level using our knowledge of data science from this class, and attempt to predict trends.

Description of Data

The data that we were given access to use came in four separate CSV files as well as a README file that were helpful in telling us what the several columns found in the data meant. We utilized all four notebooks for different components of our exploratory data analysis (EDA) and models. The abridged counties file gave us the most information about data and demographics of counties from every state and territory of the US, which was helpful for us when exploring correlations at a county and state level. The death and case rates files helped us analyze death and case confirmations over time per county and state and were super important when mapping against the counties’ file data. The last data file gave us more high level analytical information about the US and even other country's rates such as testing, hospitalization, and mortality. We mostly focused on US data as other country data we found to be mostly missing and unreported.

Description of Methods for Data Cleaning, EDA, and Modeling

Data Cleaning

For EDA, we carefully approached data cleaning to make sure that we preserved the numerical data; it was different for each EDA visualization we were focusing on. For clarity's purpose it was important for us to renaming columns into useful titles when making sub-dataframes and manipulating and joining data across common columns in datasets. We were able to effectively utilize data that came with NaNs by dropping them when appropriate. For example, when looking at HPSA underserved populations, it was important to not convert NaNs to 0 as that would be inappropriate and would skew a potential trend. By creating and reusing intermediate dataframes, we dropped NaNs and only used reliable, reported dates and focused on counties and states with the most data for the features that tended to have many NaNs. We also found it useful to drop miscellaneous, extraneous rows and columns when joining frames for clarity and working with only the information that is relevant for a particular graph or potential correlation. For example, when laying out counties within California, two of the rows were just not valid counties and offered miscellaneous information so it was important to drop those rows to make sure the data we think we are using for a feature is as accurate as possible. Fortunately, for the time series deaths and cases data, the numerical data was very reliable as there were no apparent gaps or missing data for deaths or cases per county or state per date.

We also had to clean the data more to create our model. We wanted to work with the recovered cases column, but there were too many NaNs. Initially we replaced them with mean 0 but that didn't work, so we deleted the column entirely. We also didn't use the active cases column since the number of active cases is essentially the same as the number of confirmed cases, and we realized this after we got a suspicious model with 100% accuracy. We had to omit certain places, like the cruise ships and certain states/territories, due to the overwhelming number of NaNs in those rows. We also learned that when trying various models some features can contradict each other and hold redundancies, so we had to pick and choose which redundant column to remove per model. For example, we would have to pick one feature between people tested vs testing rate. For the model we split the data into training and testing data, 80% training and 20% testing.

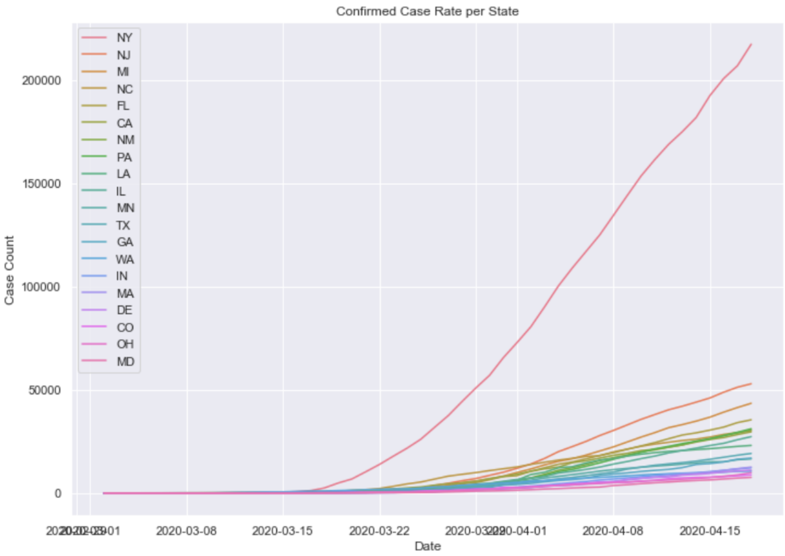
EDA

To begin our EDA, we wanted to familiarize ourselves with the overwhelming amount of data points we were given across 4 files. The first thing we wanted to do once we figured out that we had data on deaths-per-day over a period of time for every county and state was to visualize the rate of how fast death counts were rising over the given time period. We decided to graph this out first because it is hard to see any upward trends or rates just staring at columns and columns of values. We chose to do this with a line plot and by sorting by the top 20 states with the highest overall total number of deaths. This would get us closer to answering our major question number (2), which was to figure out if some states were much more impacted by the virus than others. In order to generate this graph, we were able to carefully manipulate the data by dropping unnecessary columns and using groupby, join, and transposing functions in order to gather all the death counts per state per date given into a useful dataframe (called death_plot). Our result is shown in Graph 1 below.

We then created a similar visualization with the rate of confirmed cases. We wanted to see that if over time, some states had more total deaths than total cases, or if some states had more cases than deaths. This could lead to interesting analysis to identify states that may be practicing better health and safety precautions and have better hospital facilities if they have less fatalities resulting from cases. This case rate graph is shown in Graph 2. Looking at these graphs, and answering major question (2), it is evident that New York, New Jersey, and Michigan are the most impacted in terms of sheer numbers of deaths and confirmed cases. We thought it was surprising that Michigan and New Jersey were 2nd and 3rd place as we would have expected a larger state like California or Florida to be higher because of the diversity in populations of cities like San Francisco, LA, and Miami. We were not surprised by New York taking the most hits and impact by this virus as the state has many crowded, large cities like NYC which attract people from all over the world that could have been carriers of the virus. This was a fascinating conclusion. It goes to show how the dynamics of population per state and the state's diversity seems to play a large role in seeing how detrimental a pandemic will be for that state.

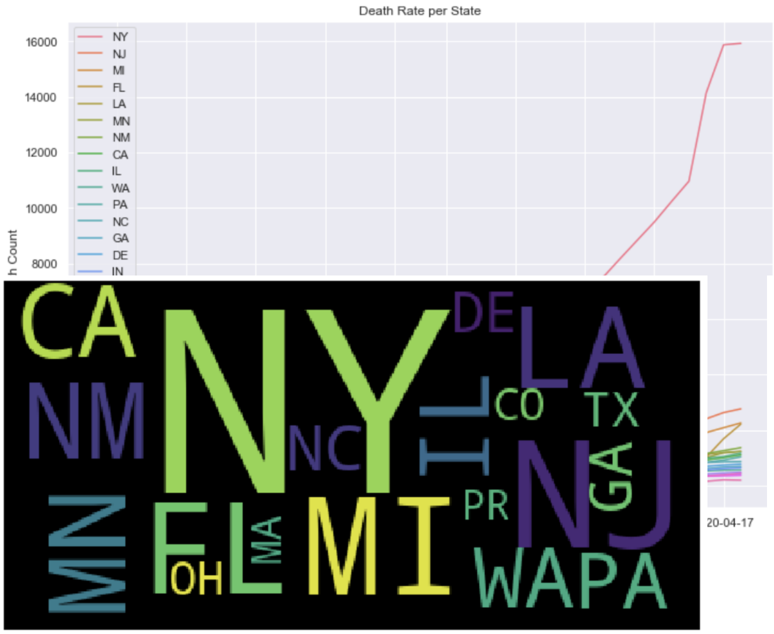
Graph 1: Death Rate per US State/Terr

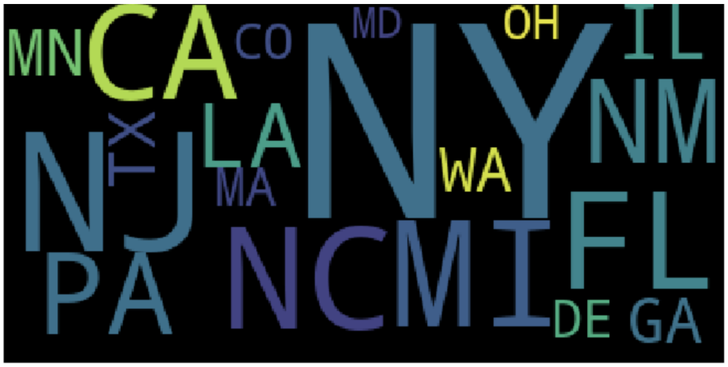
Graph 2: Confirmed Case Rate per US/TerrState



Because of the colors and matching key to the lines, it was unfortunately difficult to see which states were more affected than others apart from the top couple who led in total deaths or cases. We came up with using word clouds (Graphs 3 and 4) to show the prevalence of the amount of deaths and the amount of cases per state. It is much easier to grasp with this visualization as the bigger the state name is, the more deaths or cases there are in that state over the time period. Another interesting thing we found was that the leading 3 biggest names are the same in both word clouds. However, states like NC seem to have less deaths than cases comparatively, and states like Louisiana and territories like Puerto Rico had more deaths than cases over time. This got us thinking about what might make this happen? We agreed that diversity in health resources intuitively should make a difference across states and territories. It would make sense that areas with more access to hospitals, testing, and health coverage would be less affected overall in terms of fatalities. We cannot say this for sure without data, so we decided to analyze health resource data to address our major question (3)!

Graph 3: Total US Death Prevalence
Graph 4: Total US Case Prevalence

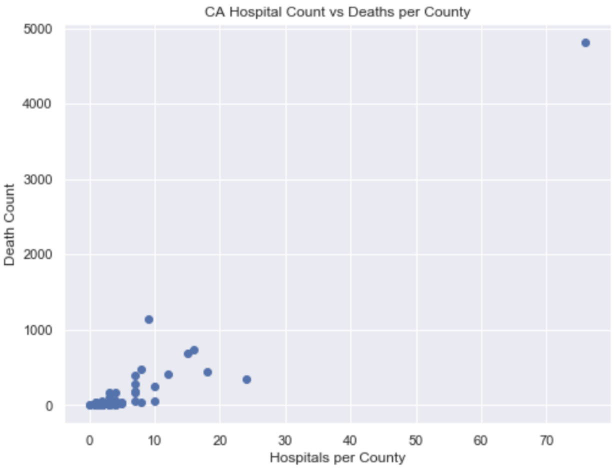




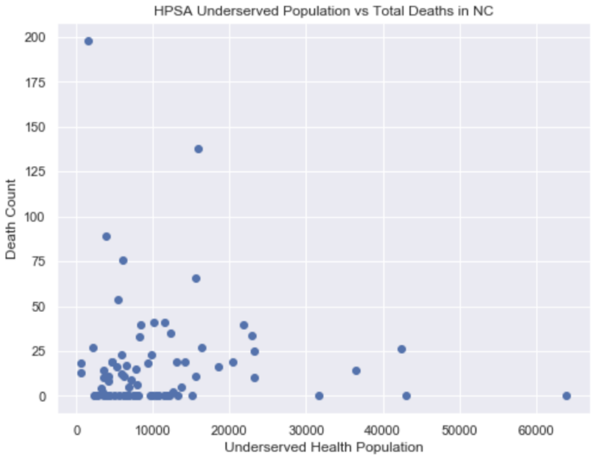
To dive into health resource accessibility data, we decided to use the features that described HSPA underserved population (the amount of people in an area with inadequate healthcare coverage) as well as hospital and medical employee count per county to see if this data had any correlations with increasing death or case rates. We focused on California because it was one of the states we had the most data reported for these features across all counties. Further, California is a diverse state that has regions of different socioeconomic status. It is also our home state, so we were curious about the data!

Our prediction was that with more hospitals or more medical employees in an area, the less deaths there would hopefully be as people would recover better or have access to medicines and certified health professionals. We were actually proven wrong by our graphs. From Graph 5, there oddly seems to be a positive association between the death counts and amount of hospitals in a county. This was surprising. We justified this by thinking that the more people filling up these hospitals the more deaths there are likely to be because the virus so far has had no effective cure. We also thought that areas with less medical coverage would lead to higher death rates, but as we can see from Graph 6 there isn't really any clear association. We tried this for multiple states that had enough HPSA data and were disappointed in seeing no helpful correlations across the board.

Graph 5: CA Hospitals vs Death Count



Graph 6: HPSA Underserved Pop. vs Deaths in NC

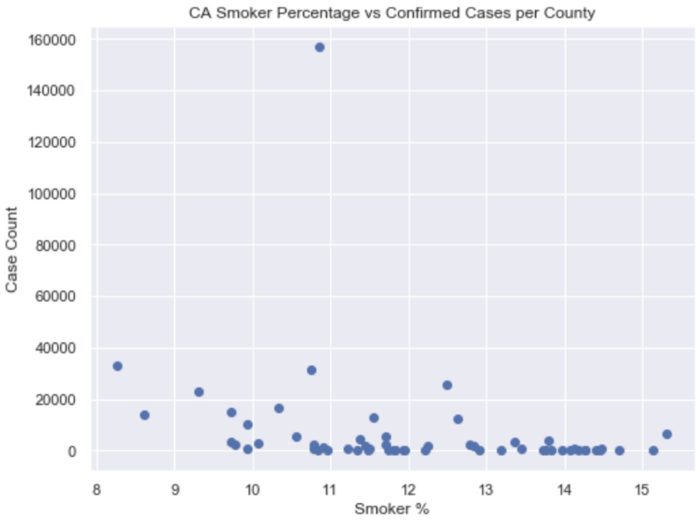


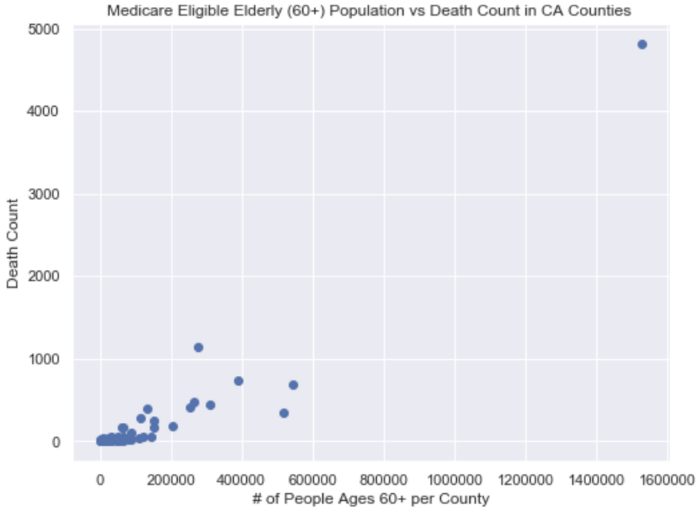
Another feature that we thought would be more helpful was when we began to analyze smoker demographics. We have heard that people with predisposition to respiratory problems are more likely to be more seriously impacted by the virus. So, we were curious to see if areas with more smokers and people with worse lungs would have an association with the amount of cases confirmed of the virus. It turns out there really isn't any correlation as the smoker percentage rises and there is no trend in case count. This feature was deemed unhelpful as shown in Graph 7.

However, we did confirm a trend we had a previous belief about. Media tells us that seniors are at high risk for the virus and more likely to have more severe, and potentially fatal, health reactions to COVID-19. We wanted to see if this was true so we graphed the Medicare-eligible population (seniors ages 60 and above) in CA per county to the death rate per county in CA in Graph 8 and found a pretty strong linear correlation! This is one of the most reliable correlations we found.

Graph 7: Smoker % in CA vs Confirmed Cases

Graph 8: CA Senior Population vs Death Count

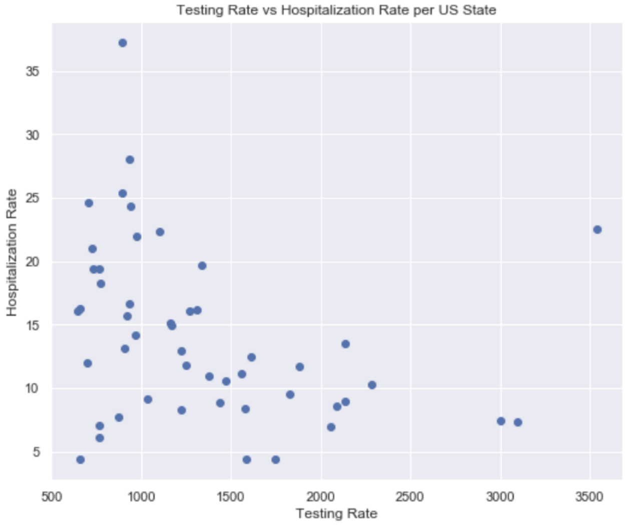

CA Smoker Percentage vs Confirmed Cases per County


Medicare Eligible Elderly (60+) Population vs Death Count in CA Counties

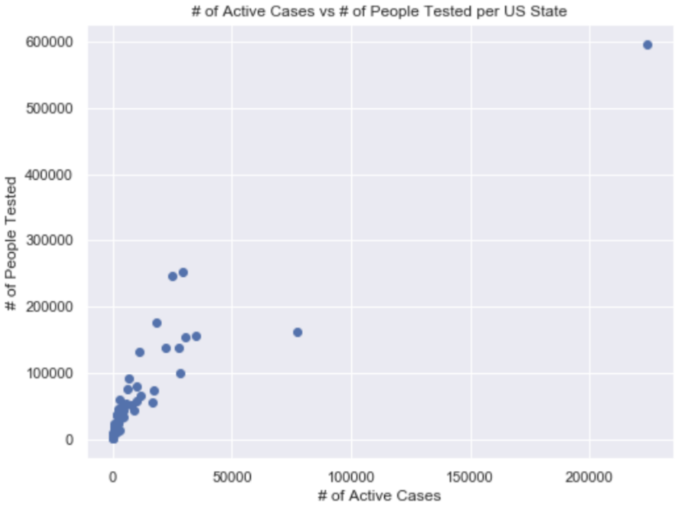
To further explore more health resource trends to address our major question (3), we were motivated to use the April 18 data file with rates on mostly US data to see if testing rates and the amount of people able to be tested led to any interesting insights. We found that by graphing the testing and hospitalization rates in the US, states with higher testing rates tended to have lower hospitalization rates and that states with high hospitalization rates tended to have lower testing rates. We cannot default to any strong explanation, however it would make sense that areas with greater testing capabilities allowed people to catch their symptoms early and isolate themselves. This would lead to lower hospital entries for that area. Areas with less testing capabilities would lead to exacerbated conditions with people not knowing how long they've had the virus, therefore infecting more people, and leading to higher hospitalization. This was an interesting, yet logical conclusion to make.

This leads us to be confident that testing kits and allowing people to test more is very important when a highly contagious pandemic is occurring. In Graph 10, we can see that there is a steep positive association with the number of active cases of COVID-19 and the number of people tested. This goes to show that the more people that can get tested if they have doubts is better because it will let them know for sure if they do have the virus and can start treating it. We hear and see of a lot of places denying people tests but this will only make the situation worse as people react to the virus in different ways; some people barely show any symptoms and could have affected dozens of people without even knowing.

Graph 9: US Testing vs Hospitalization Rate

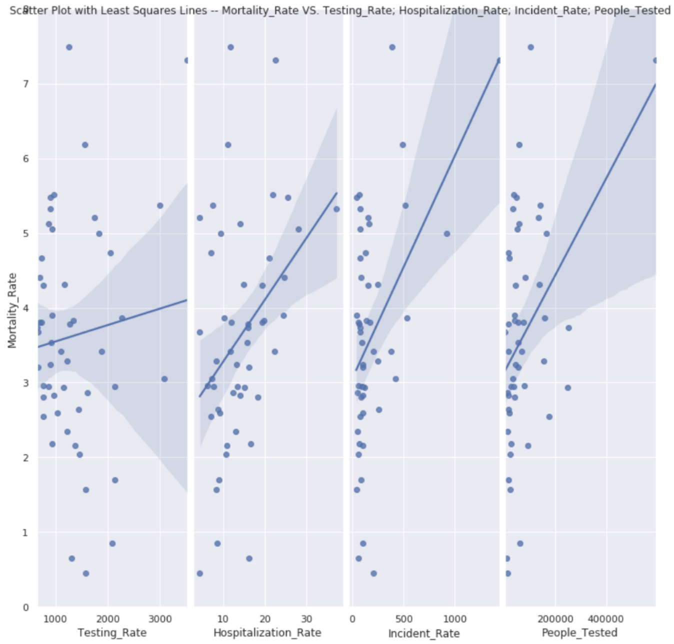


Graph 10: Active Cases vs People Tested in US

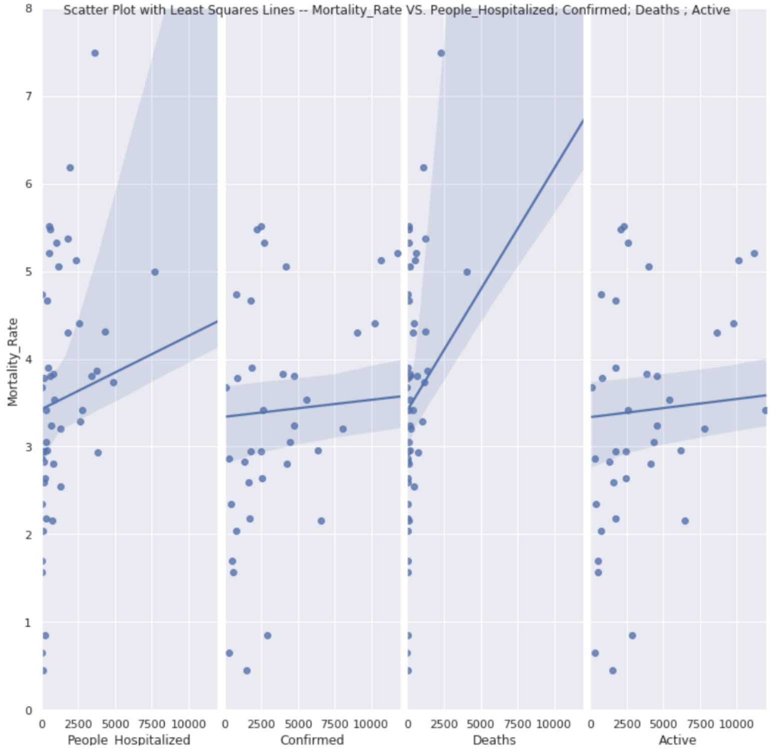


After familiarizing ourselves with the data and all the features, we decided to try using least-squares lines to figure out which factors could have a linear relationship with mortality rate or death cases. We graphed a number of scatter plots with least square lines to see what stronger linear relationships between features exist (Graph 11 and 12 are Mortality vs features, and Graphs 13 and 14 are Death Cases vs features). We found that death cases against various features (*Hospitalization Rate, Incident Rate, People Tested, People Hospitalized, Confirmed Cases, People Tested, Active Cases)* have a linear relationship, as seen in Graphs 13 and 14.

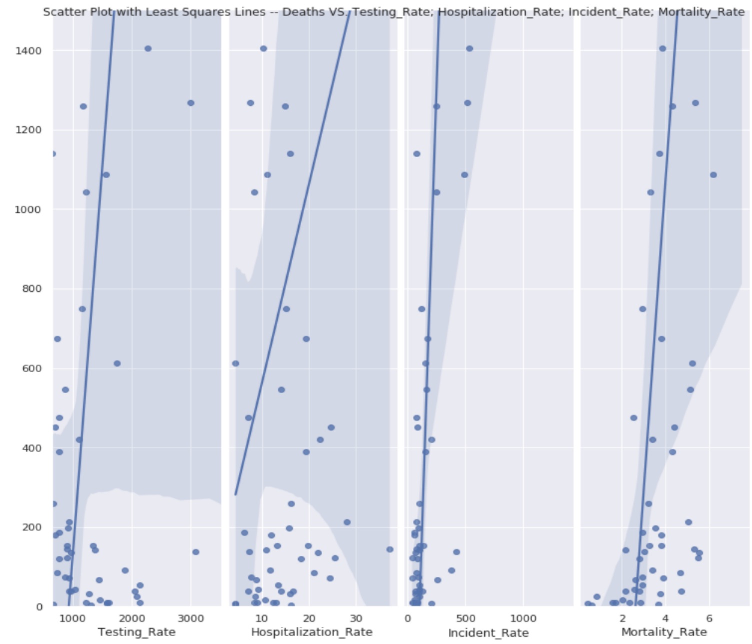
Graph 11: Scatter Plot with Least Squared Lines -- Mortality Rate VS. Testing Rate; Hospitalization Rate; Incident Rate; People Tested



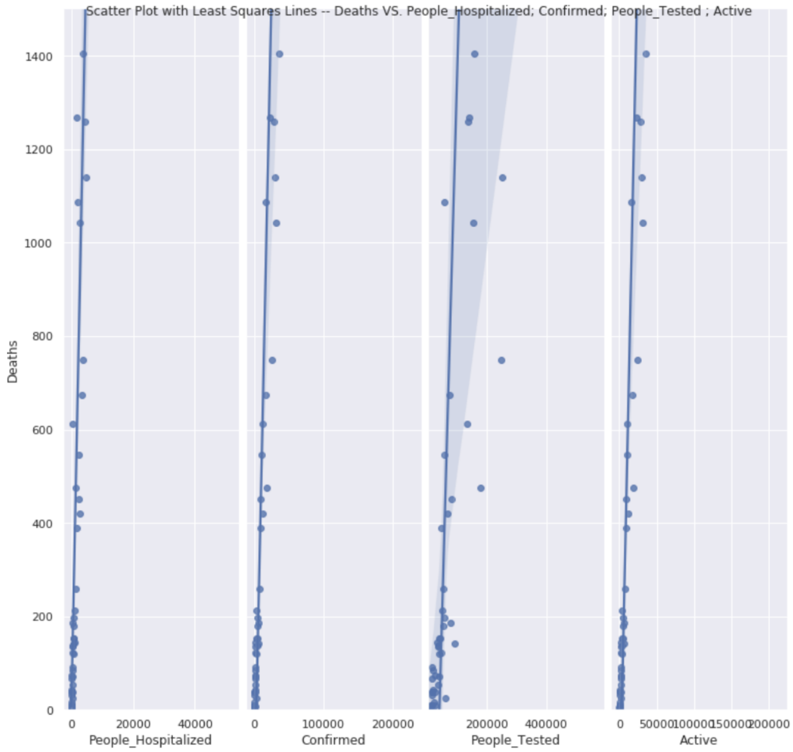
Graph 12: Scatter Plot with Least Squared Lines -- Mortality Rate VS. People Hospitalized; Confirmed Cases; Death Cases; Active Cases



Graph 13: Scatter Plot with Least Squared Lines -- Death Cases VS. Testing Rate; Hospitalization Rate; Incident Rate; People Tested



Graph 14: Scatter Plot with Least Squared Lines -- Death Cases VS. People Hospitalized; Confirmed Cases; People Tested; Active Cases



Limitations & Challenges with Data and Analysis

A limitation for our project was that there were not enough reported values or too many NaNs for many states on features we were interested in such as HPSA data per county. It did not seem reported well and therefore, it was difficult for us to draw reliable conclusions analyzing that data. This also means that our current model does not take in information from as many states or territories as it could have been informed upon. A lot of the columns that were listed in the README file were unfortunately nowhere to be found in the columns of the actual data csv file. Data such as poverty percentages and asthma prevalence and median income would be data points we would have been interested in digging into further to visualize and find trends and better indicators of death rates. Comparing countries was also not possible as the dataset with multiple countries awkwardly only selected a few other countries like China, Australia, and Denmark that have no obvious connection to each other. Most of the data on those countries was not accounted for so we were forced to focus on US data. Almost all the quantitative social distancing data that the README said was also unfortunately not in the data files except for the proleptic Gregorian calendar dates for when local and federal bans on social distancing rules and large gatherings were passed. We got stuck as it was hard to meaningfully transform this data or really visualize or compare anything with just ban dates other than timelines. This was something we found very unhelpful and challenging to work with. So, we decided to focus on other features we found more promising instead.

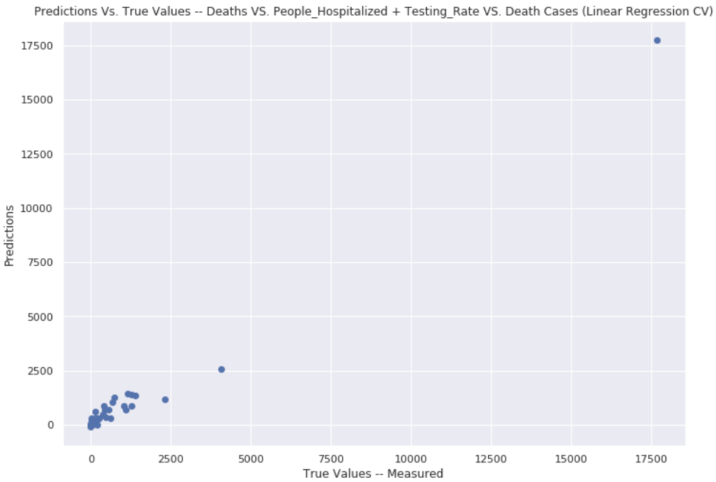
Model & Predictive Analysis

We chose to use a linear regression model to predict future deaths within the US and incorporate as many features as we can. The initial plan was to try a linear regression model, and incorporate k-fold cross validation to help its performance if needed. If neither worked, we would try a decision tree regression model. We wanted to also try one with logistic regression, but after data cleaning it didn't make much sense since there weren't sufficient binary variables. We assumed linear regression would be best over other types of models since we suspected a linear relationship between various features and disease spread. First we did single linear regression, then we did multiple, and then we did single/multiple with k-fold cross validation. We then tried to use a decision tree to see how this model could perform, however it didn't work particularly well and went with linear regression. When evaluating the model we first used the train-test-split methodology to separate training and testing data from a randomized perspective. We then performed cross validation to see how our model could perform when used against an independent dataset. To answer key question (4), we learned that the factors that predict death cases the best are people hospitalized and testing rate, and using these features on a multiple linear regression model with k-fold cross validation yields our best result with highest accuracy of around 98%.

Graph 15: Predictions vs True Values of Death Cases vs People Hospitalized + Testing_Rate

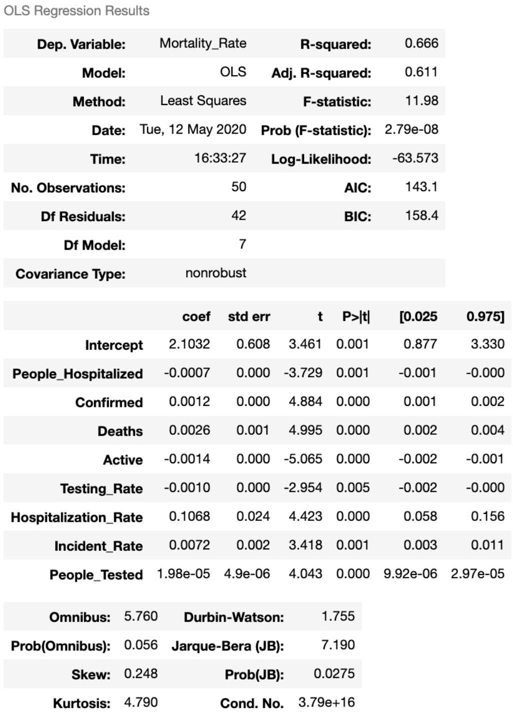
Hypothesis Testing

After approaching the model, we wanted to address key question (1) which is: which factors most strongly influence death rates? Based on hypothesis testing, we discovered that the hospitalization rate seems to influence death rates the strongest. You can see results in the table below. Our null hypothesis was that there are no relationships between mortality rate and other factors, and our alternative was that there are relationships between the two. All of the p values were less than 5%, thus supporting the alternative hypothesis.



We learned that the number of confirmed deaths cases, and incident rate have slightly positive correlation with mortality rate, and that hospitalization rate has a relatively strong correlation with the mortality rate. The number of people hospitalized, active cases, testing rate have slightly negative associations with the mortality rate. As the testing rate increases, mortality rate decreases. This reflects that the current method of identifying disease carriers is effective. As the number of people hospitalized increases, mortality rate decreases. This shows that social distancing (enforced by most of the state governments) is effective in preventing further spread of COVID-19, thus protecting people. As the number of active cases increases, mortality rate decreases. Since we can get the number of active cases by subtracting the number of recovered and death cases from the number of confirmed cases, it is safe to assume that the number of recovered cases slowly increases. However, it is important to note that all the negative associations are relatively small. It is not yet the same to make the assumption that the guidelines enforced by the government and the CDC are significantly efficient. There is still a significant part of the US population who are severely sick, reflecting from the coefficient of the hospitalization rate is around 0.1068, which is much more significant than the other coefficients. With the coefficient of incident rate (confirmed cases per 100,000 persons) of 0.0072 tailing behind, it is evident that the US medical system faces enormous challenges.

Figure 16: Hypothesis Testing for Mortality Rat VS. All the Other Dependent Variables



```
OLS Regression Results

       Dep. Variable:   Mortality_Rate        R-squared:    0.666
              Model:              OLS    Adj. R-squared:    0.611
             Method:    Least Squares       F-statistic:    11.98
               Date: Tue, 12 May 2020 Prob (F-statistic): 2.79e-08
               Time:         16:33:27    Log-Likelihood:  -63.573
   No. Observations:               50               AIC:    143.1
       Df Residuals:               42               BIC:    158.4
           Df Model:                7
    Covariance Type:        nonrobust

                        coef  std err       t  P>|t|    [0.025    0.975]
          Intercept   2.1032    0.608   3.461  0.001     0.877     3.330
 People_Hospitalized -0.0007    0.000  -3.729  0.001    -0.001    -0.000
          Confirmed   0.0012    0.000   4.884  0.000     0.001     0.002
             Deaths   0.0026    0.001   4.995  0.000     0.002     0.004
             Active  -0.0014    0.000  -5.065  0.000    -0.002    -0.001
       Testing_Rate  -0.0010    0.000  -2.954  0.005    -0.002    -0.000
 Hospitalization_Rate  0.1068    0.024   4.423  0.000     0.058     0.156
      Incident_Rate   0.0072    0.002   3.418  0.001     0.003     0.011
      People_Tested 1.98e-05  4.9e-06   4.043  0.000  9.92e-06  2.97e-05

       Omnibus: 5.760   Durbin-Watson:    1.755
 Prob(Omnibus): 0.056 Jarque-Bera (JB):   7.190
          Skew: 0.248        Prob(JB):   0.0275
      Kurtosis: 4.790        Cond. No. 3.79e+16
```

Summary of Results

EDA

We were able to answer our major question (2) (Which states are more drastically affected than others?) that we were very interested in because we wanted to know how our nation is holding up to the virus. From our line plots in graphs 1 and 2 we were able to graph and identify the top states that had the most impact in terms of death counts and confirmed case counts. New York stuck out as the state with the most accelerated developments in death rate and confirmed case rate. The next highest in both deaths and cases was Michigan and New Jersey which was surprising to us. We were able to find interesting observations of how some states had higher death rates than case rates and vice versa. This was shown better and more easily through our word clouds in graphs 3 and 4.

For addressing our major question (3) (Are areas with more health resources more likely to have lower death or hospitalization rates?), we did not expect to come up with the conclusion we did when trying to analyze the effects of accessibility to health resources. We found that the more hospitals in an area correlated to higher death counts because this virus is spread by mere contact and population so this makes sense that more hospitals would be in areas with more people and thus infected people.

We could not find any significant association between those underserved and receiving inadequate medical coverage and the death rate as we thought we would. This feature was also difficult to find states with few missing values. Better conclusions on how access to medical resources with respect to death and case rates is left to further study. However, we did find that testing rates are important to taking precaution as more testing opportunities could prevent higher hospitalization rates for a county as we found in graph 9. We thought that analyzing smoker populations would lead to a better indicator to predict higher numbers of cases but we ended not seeing any correlation at all in graph 7. We were able to confirm our previous assumptions that the virus causes higher fatalities in areas with higher elderly population ages 60+ as we could see in graph 8.

Model

We went over most of these results in the prior section. After modeling we found that people hospitalized and testing rate predict death cases the best out of all of the features, answering key question (4) (Which features best predict death rates?). We then did hypothesis testing for mortality rate because we already used death cases in the model in order to answer the key question (1) (Which factors most strongly influence death rates?). We learned that there is a strong positive correlation between hospitalization rate and mortality rate. Which states are more drastically affected than others?

Discussion

Ethical Analysis

The data reported did not include a vast majority of the columns that the README said it would have regarding diversity in socioeconomic status or poverty levels or income levels of households. This is an ethical concern because models trained without this data would have an inherent bias towards well-funded or more privileged regions. They would not take into account areas or populations that may be differently affected from the pandemic due to differences in urban and suburban regions, lack of healthcare accessibility, or lack of financial resources.

Further Steps

We think that if we had more time, some interesting next steps would be to see if we could gather more information regarding the effects of social distancing through finding other datasets. We think this would be more helpful than just the calendar dates provided from bans on social distancing legislation. We also believe that data on the socioeconomic status of populations and regions, and poverty rates have the potential to have correlations with case rates or death rates as the rate of infection increases in areas where populations may be more crowded or less sanitary. Homelessness data and airport data in urban cities over the 2020 time frame could also be great indicators as places that are more concentrated with people from different places are just more likely to spread the virus more. This data could help build better predictive models!

Additionally, there are currently developers building a self-reporting tool as an online data science resource. We think that incorporating data from other resources and other public datasets would help give us more interesting and better informed results. If the current data on states/territories/places was more detailed we wouldn’t have to omit certain rows. For example, more information on the Grand Princess, could have given us better results. The self-reporting tool is helpful as well, it’s not necessarily as accurate but we could incorporate information about how people suspect they have COVID-19 and draw some conclusions about the lack of testing. If we had more time and could gather more up-to-date data, we could explore more indicators and correlations and make predictions with higher accuracy!

Acknowledgements

In all, this was a very fun final project for us. Thank you to the TAs, GSIs, and professors for teaching us this semester given the difficult circumstances that arose due to the COVID-19 pandemic. It was very rewarding for our project group to come together and have this opportunity to apply the skills we learned in Data100 to analyze a current, real-world dataset!