Project Narrative
SARS-COV2/COVID-19 Exploration
Coefficient Analysis of Growth Factors and Estimation Methods for New Cases

Jack Stehn
jack.stehn@berkeley.edu

Lubah Nelson
lubah@berkeley.edu

Erica Zhu
ericaz@berkeley.edu

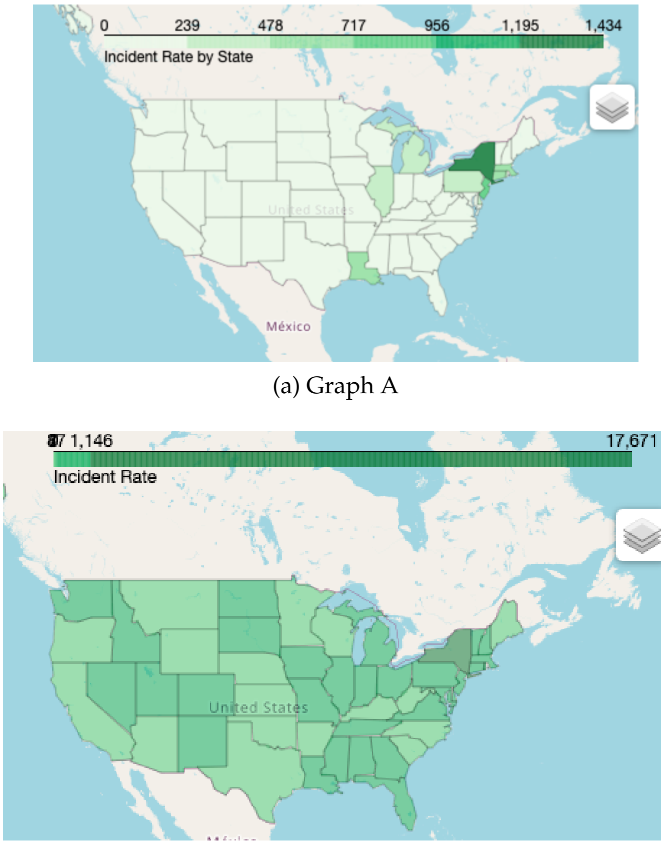
May 13, 2020

1 Introduction

When we were approaching the exploration of COVID-19 data set, we had two large questions that we wanted to address. With the knowledge that confirmed cases are not an accurate reflection of true infections, we wondered: Can we use the data in this set to find a better way of measuring the spread of the disease? Secondly, which factors played the largest role in determining how quickly a disease would spread in a region?

2 Exploratory Data Analysis for Question One

The first step in EDA was just to begin looking at some graphs and seeing if there are some patterns that emerge. As our intention is to check for patterns of the outbreak, the following plots measure the incident rate of every state.



(a) Graph A

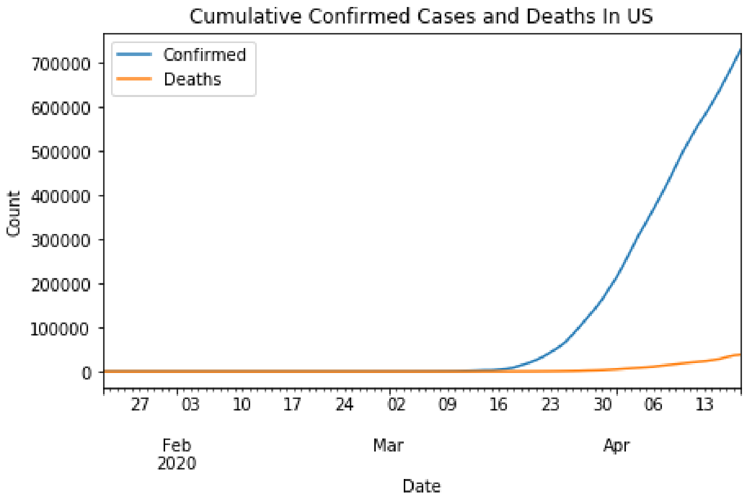
(b) Graph B

Figure 1: Distribution of Confirmed Case

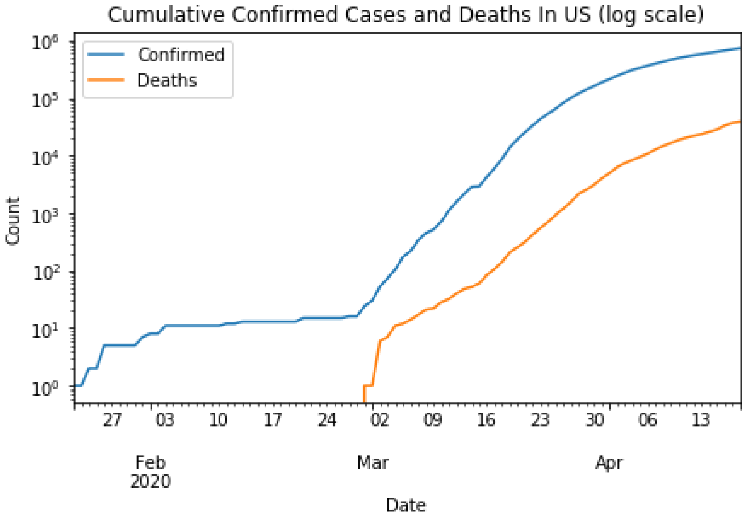
We noticed in Graph A that the data is very skewed. The plot implies that New York is the only state truly affected by the epidemic. We could use a logarithmic transformation to solve this problem, but it lost its human readable meaning and introduced problems at very low incident rates. We decided to examine the incident rates colored according to the quantile (Graph B). Using the quantiles to bin the data gives us more information about the distribution of infections and more understandable numbers. While it leads to a loss of granularity in color, a more understandable graph is much better for EDA.

We observed an interesting pattern in the New York metropolitan area: the infection did not seem to respect state borders. This led us to question of what models can we create that would identify these patterns that were not simply captured by looking at state level data.

Next, we focused on investigating how the disease grows over time. The following two graphs are plotted from the same data points and depict the cumulative confirmed cases and deaths in the US over time. Graph C is shown on a linear scale, while graph D is on a logarithmic scale.



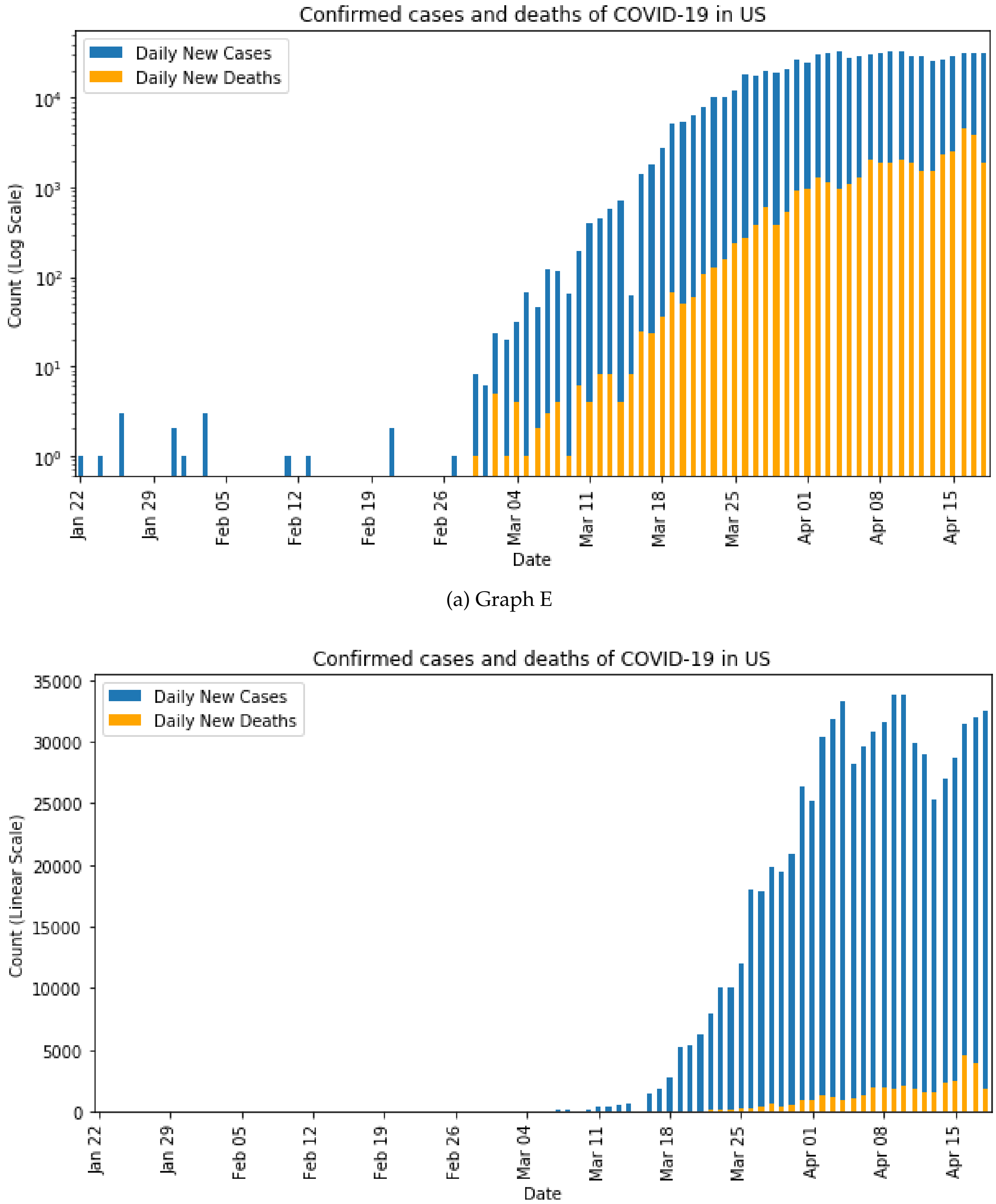
(a) Graph C



(b) Graph D

Figure 2: Cumulative Confirmed Cases and Deaths

From graph D, we were able to gain some insights into the nature of the spread. The first 100 cases, while peculiar, is expected. With low numbers of infections and low testing rates, small changes in infection count would create huge fluctuations on a log scale. Additionally, at the start of the spread of COVID-19, the country had very few tests. As we further inspect and go down the timeline, we see a steady arc for both deaths and confirmed cases. This suggests that, while logarithmic scales are more appropriate than linear scales, the spread of the virus is not strictly exponential. This could be due to community activities or just the natural spread of the disease (as exponential growth would only be valid in infinite and non-closed systems).



(a) Graph E

(b) Graph F

Figure 3: Daily New Cases

To gain further insights into the nature of infectious spread, we continued our EDA using daily new infections and deaths rather than using cumulative cases as in graph E . From this graph we gained insights on the nature of the spread. We were able to see the exponential growth has slowed down. On a logarithmic scale, the curve of the graph seems to be flattening. While it's useful for understanding the rate of growth for an epidemic that grows exponentially, that curve is misleading. The linear scale in graph F of cumulative cases provides clearer insight.

The data lacked information about the number of active infections. We had to estimate it ourselves. The World Health Organization (WHO) report suggests that the virus exists on its host between 2-6 weeks, depending on the severity of the case. While the most common duration was 2 weeks, we chose to be pessimistic and assume that it lasts for 3 weeks. Although an inaccurate measure, this is our best attempt to estimate active cases as the law of large numbers suggests it would look roughly similar. Graph G illustrates the result.

Graph G

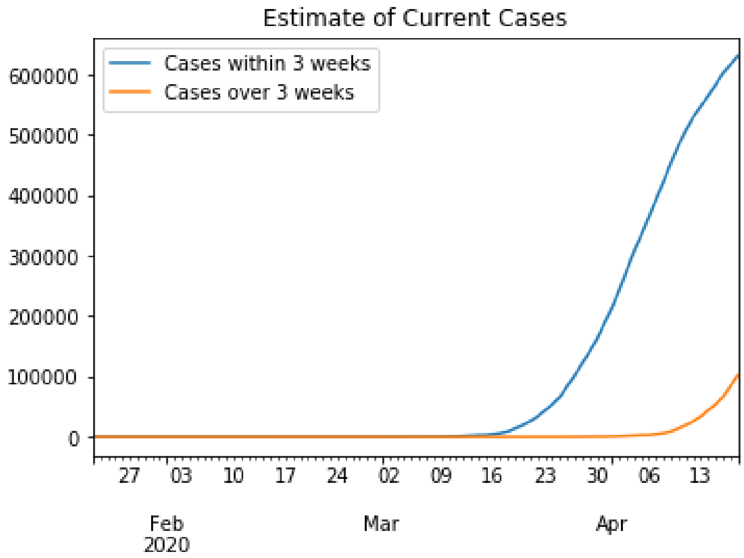


Figure 4: Estimation of Confirmed Cases

This is a very rough estimation and largely relies on when tests are confirmed. It presumes an average duration of 21 days as well as accurate testing, which has been a problem worldwide (though especially in the United States). We can see the curve is slightly decreasing, but the number of active cases is still rising.

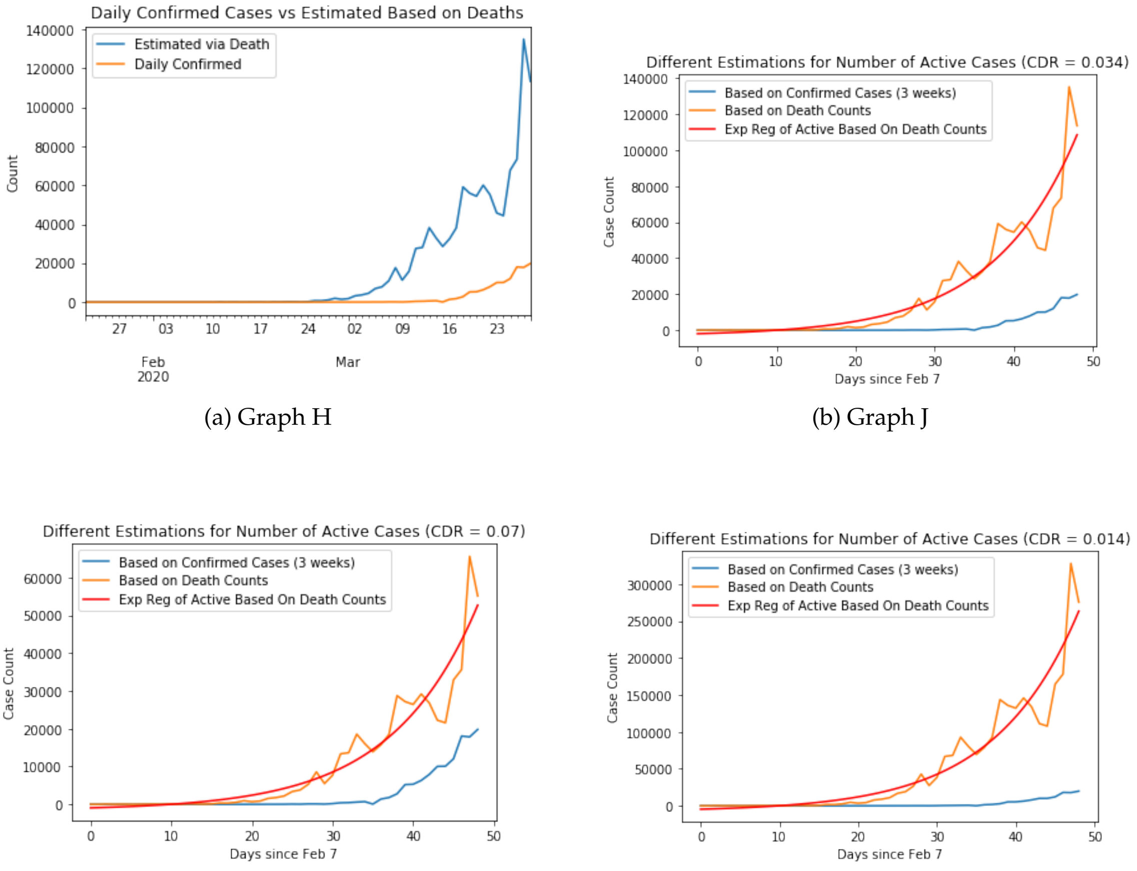
3 Modeling Question One

After the EDA, we proceeded to answer the first question of the project: estimating daily new infections. We began looking for a way to estimate active cases based on the death count. While death count is also not a perfect measure, our hypothesis is that the death count is closer to the true number of deaths than the confirmed cases is to a true number of cases. This was inspired by the article: Estimating cases of Covid-19 from Daily Death Data in Italy (Raheem, 2020).

Most of the data were prepared and cleaned in the EDA phase when we created new data for daily new infections and deaths. If we know the death rate and the number that have died, we know how many were infected. If we know how long it takes the disease to be fatal, we use this to estimate how many new infections occurred at a given point. This is demonstrated in the math below with D_T being deaths at a given date T, I being the number of infections, and CDR being the cumulative death rate.

$$D_T = I_{T-21} * CDR \implies I_{T-21} = \frac{D_T}{CDR}$$

For our first estimations, we used a cumulative death rate of 3.4%, an estimation made by the CDC (CDC COVID-19 Response Team, 2020). Additionally, we applied an exponential regression to view a smooth curve as early numbers would be susceptible to chance when it comes to the date of death (seen in the Graph H and Graph I below). With these plots, we were able to see which death rates seem reasonably accurate. While these are all rough estimations, this served only as a gauge in evaluating our model or assumptions. True values of cases are impossible to know, but so wildly diverging estimations would be an indicator of one of the models or assumptions being incorrect. To gain an idea of estimating the death rate, we looked at various numbers from different institutions:



(a) Graph H

(b) Graph J

(c) Graph I

(d) Graph K

Figure 5: Cumulative Death Rate: X%

Observing the different death rates, we pick out the ones that seem reasonable. For example, a death rate of 7% seems unrealistically pessimistic, although technically possible. The values 3.4% and under seem much more consistent with our knowledge on testing rates and missing data. It also seems to be consistent with a death rate lower than 3.4% and the recent research estimating the death rate as between 0.39% and 1.33% (Mahase, 2020). Note that this research is based on data from China and may not apply well to the United States. The death rate also may shift over time as we hit medical care capacity. The accuracy of these models depends very much on proper data collection and knowledge of the virus, though the number of confirmed cases acts as a floor to the number of infections. It is widely assumed that this is a gross under count of people that are affected largely due to failures in testing and asymptomatic people. Graph M depicts regression lines based on various assumptions of the death rate. We created an estimate of daily new cases using the information that we have which presumes uncertainty about the disease. It assumes that the count of deaths caused by COVID-19 is more accurate than the number of confirmed cases as a measure of actual cases. It is also modeled off of data primarily before stay-at-home orders were made. This model may be good at predicting what would have happened had the orders not been made, but the actions taken by the people of the United States have changed the trajectory of this virus. Finally, the exponential regression, although adequate at the start of any pandemic, fits fairly well, will diverge compared to the traditional SIR Model for spread of disease as a significant portion of the population has had the disease.

Graph M

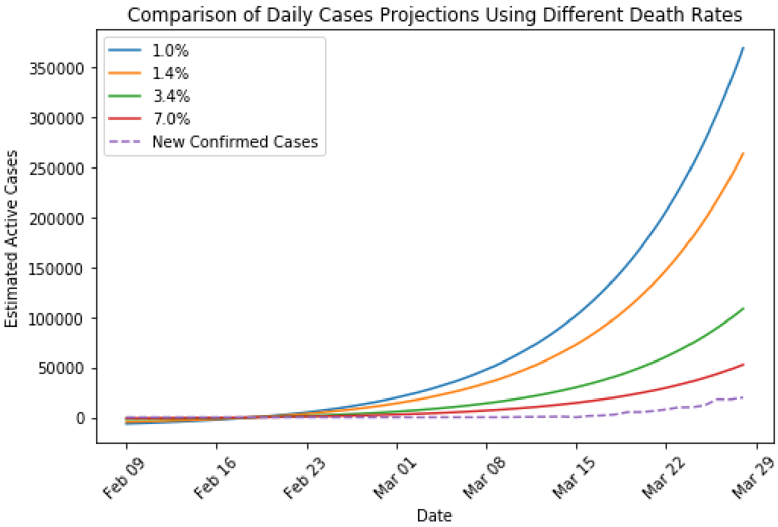


Figure 6: Exponential Regression

We conclude that this model has potential, however, there are some fundamental flaws with using it so early in the pandemic. First, the model relies on hyperparameters which we do not know. It relies on the length of time between infection and death and an assumed death rate. For both, we have rough estimations from John Hopkins or the CDC, but the range of possible values is so wide at this point that the results are incredibly varied. Additionally, the choice of an exponential regression is only adequate for the early stage of the disease but ultimately will fail later on. We cannot see the effects or changes due to actions of states as, because this model tells us infections from 3 weeks ago, we do not get to see how those actions have affected the spread of the disease. As an opportunity for more research, we could adjust this model as new data comes out and switch to logistic regression or a linear regression with many more features engineered. We would also adjust the hyperparameters as we begin to understand the patterns of this disease better.

4 Exploratory Data Analysis for Question Two

The second goal in this project was to find which environmental factors most strongly correlated with the spread of the disease.The EDA phase made us develop a caution in creating our model. We had to find a way to geographically border our data that didn't rely on state borders. Referring to graph B, we were able to see interesting patterns emerging in areas such as New Jersey and New York; they seemed to coincidentally heat up. This implies that the disease doesn't respect state borders, and spreads on a community level. We then compared looking at data on the state level and then on a county level.

Graph B

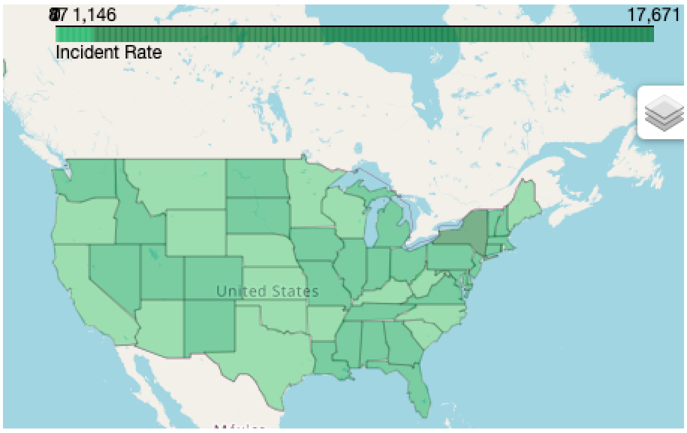


Figure 7: Confirmed Cases by State

We ultimately decided that it was much more reasonable to look at metropolitan areas alone. Looking at states leaves us with a small sample for training a data set and problems with train-test splits. The overwhelming majority of counties also do not have many infections and such model building would be based on relatively small integers. Looking at metropolitan areas combines similar counties, gives us aggregation on larger numbers, ignores state lines, and gives us a set of data that is much closer than the vast differences between rural and urban states/counties.

After the EDA, we had to clean the dataset and apply transformations to create our features. First, we cleaned up the data frame by dropping all the missing values and merged a few tables to associate time series data with county information, and then metropolitan area. We as well converted the FIPS column into integer values as we wanted to graph the predicted values by county. We then needed to combine the data frames together based on (Census.gov, 2020) metropolitan area by choosing appropriate aggregate functions for each column.

5 Feature Engineering for Question Two

Once we made our data frame, we proceeded with a series of feature engineering. Our first goal was to find some way to measure the spread of the disease. To account for different times that community spread began, we decided to come up with some measurement of growth. This is the growth that we used in our model to predict and analyze coefficients. First, we checked the first date that a metropolitan area reached a threshold number of confirmed cases. We then looked at the cumulative number of cases after a set time period. This gives us a sense of growth speed after the disease is left for a while. After fiddling with the parameters, we selected 50 cases and 10 days as our time period. Other values either allowed for too much variability or eliminated too many metropolitan areas.

Some more feature engineering was done on variables that measured relative information about when measures were taken. We presented them as dummy variables. Because we are concerned about relative importance, we needed to make sure each variable was on similar scales. We created a function that took in data frames and normalized the data as appropriate. We as well created a feature that checked to see the percentage of the population that was eligible for medicare and a feature that changes the dates to be the number of days since the start of the data.

6 Modeling Question Two

Once we created a full data frame with features, we started to work on our model using coefficient analysis. Because we were going to be analyzing coefficients, it was essential that we minimize colinearity. Since we just wanted the strength of the correlation, we used the absolute value of the correlation matrix.

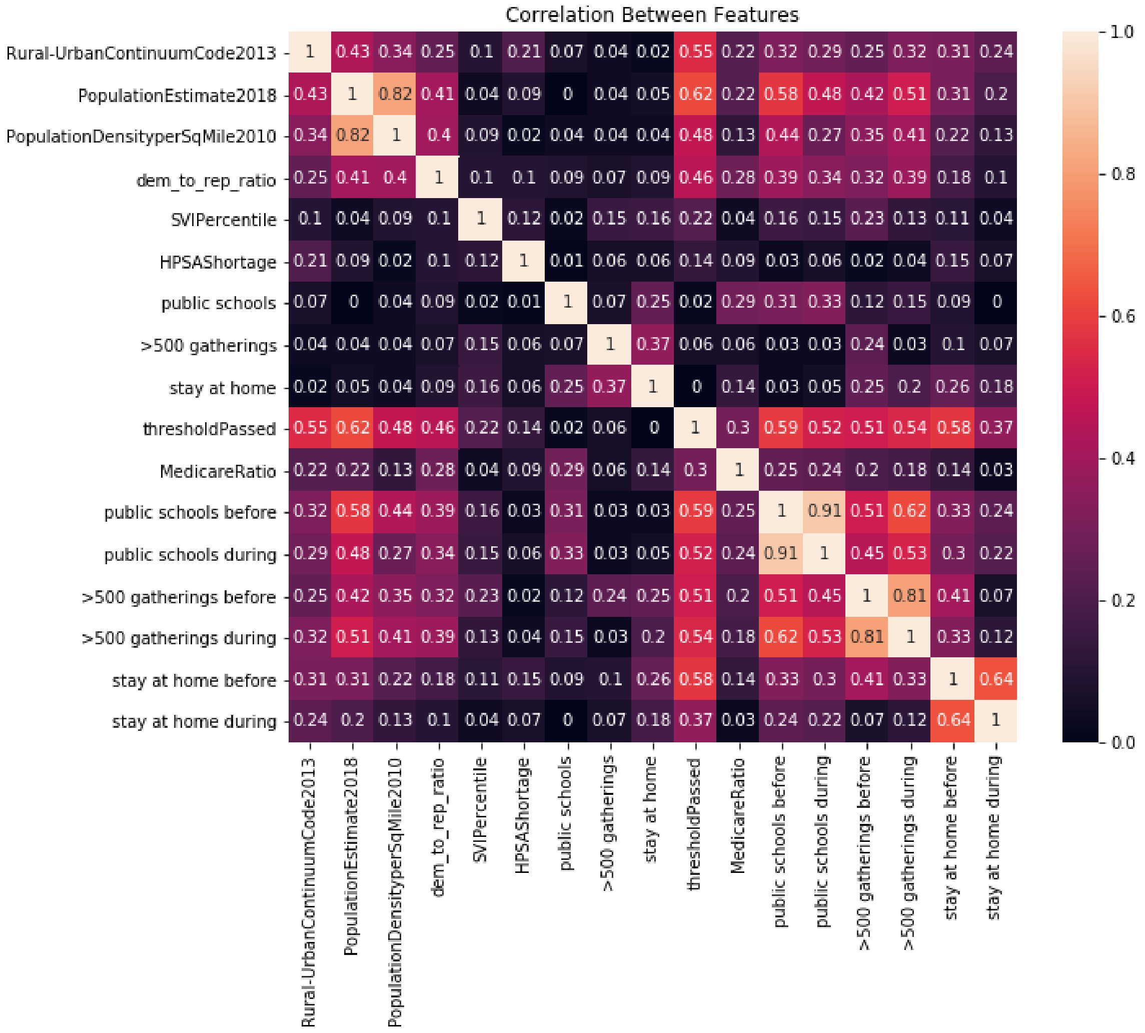


Figure 8: Correlation Matrix

As can be seen in figure 8, there exist strong correlations between the categorical variables (as they are related to each other). We dropped the values that shared strong relationships accordingly to reduce multicollinearity. The final graph shows much less correlation between the variables.

The next step was to check for multicollinearity by using the variance inflation factor for measurement. We were successful in this as there was little multicollinearity. Furthermore, no VIF went above 5, which would have been a critical point for removal. The following can be observed in the table below:

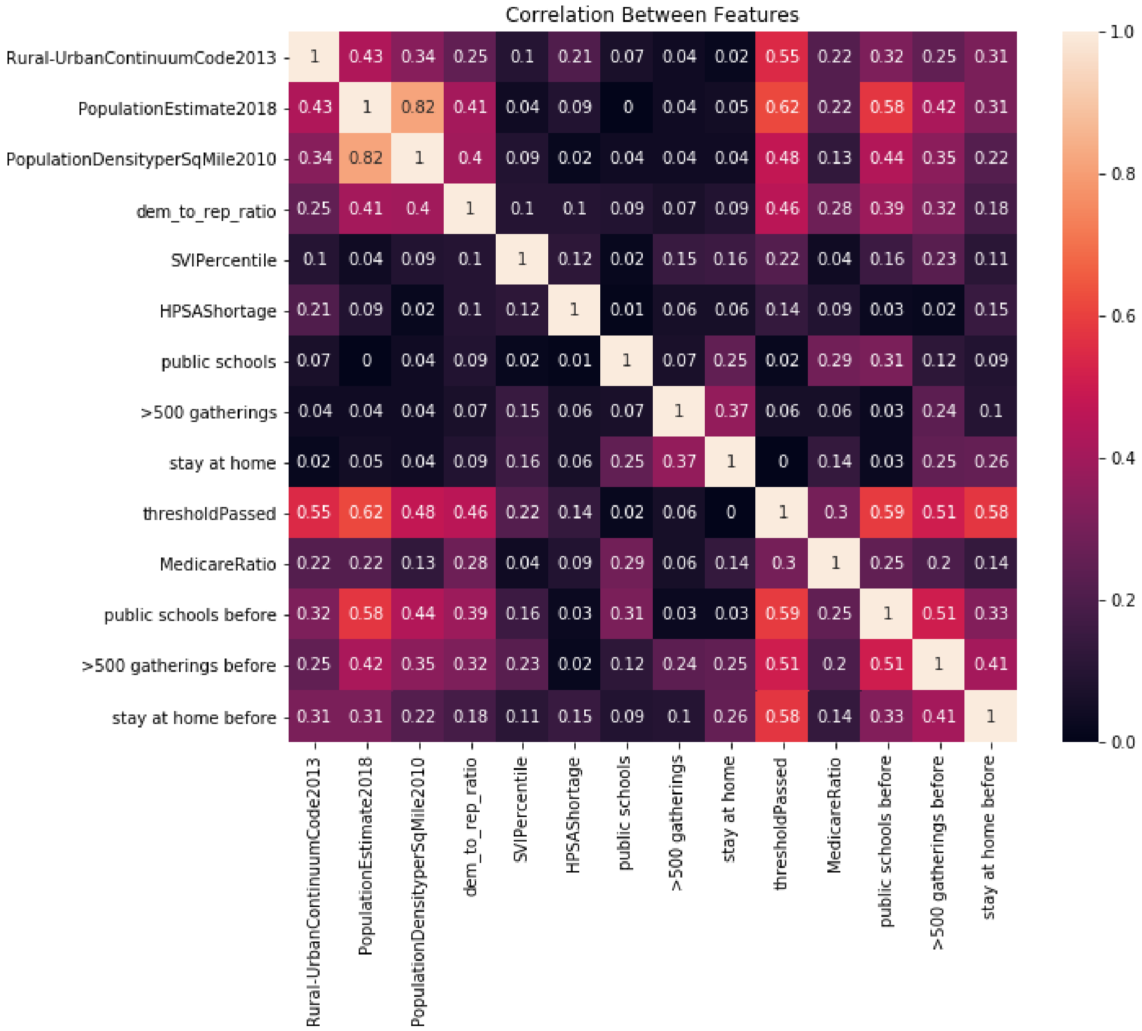


Figure 9: Adjusted Correlation Matrix

Feature	Variance Inflation Factor
threshold-Passed	3.61
Stay at home before	1.8
Rural-UrbanContinuumCode2013	1.53
Dem_to_rep_ratio	1.45
MedicareRatio	1.35
HPSAShortage	1.14
public schools	1.5
stay at home	1.43
>500 gatherings	1.25
SVI Percentile	1.19
PopulationEstimate2018	1.98
public schools before	2.17

After we carefully selected our features, we were ready to create our model. We first created a train-test-split. We then selected a linear regression model to fit our data as we wish to compare all of our coefficients. With the training data, we received a score of roughly 0.63. With the complexities of the pandemic and limitations on data, we could not expect too great of accuracy. Although this score is not perfect, there exists a correlation between the results. We then used 5 fold cross-validation to ensure that we are not overfitting the data. Yielding a score of 0.66, we concluded through the cross-validation that our model generalizes fairly well and is not overfit to our data.

The next step was to focus on the relative influence of different factors by doing the analysis of the coefficients of the model. First, we wanted to make sure there was no pattern in the residuals as seen in the graph below. We then finally were able to check the model on the test set yielding a score of 0.72.

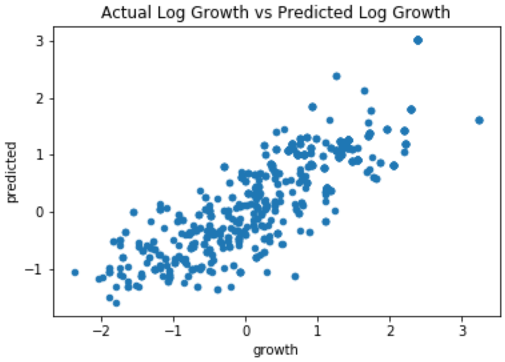


Figure 10: Actual Log Growth vs. Predicted Log Growth

The final step was graphing and analysis of the final model. We plotted the growth measurements for each metropolitan area and compared it to the predicted growth of our model. We as well wanted to look at the difference between the two to see if there were any clear patterns and failures in the model geographically. However, from the maps, we can see that there is no discernible geographic pattern to over or underestimation, suggesting that not including regional features did not have a huge effect on the model.

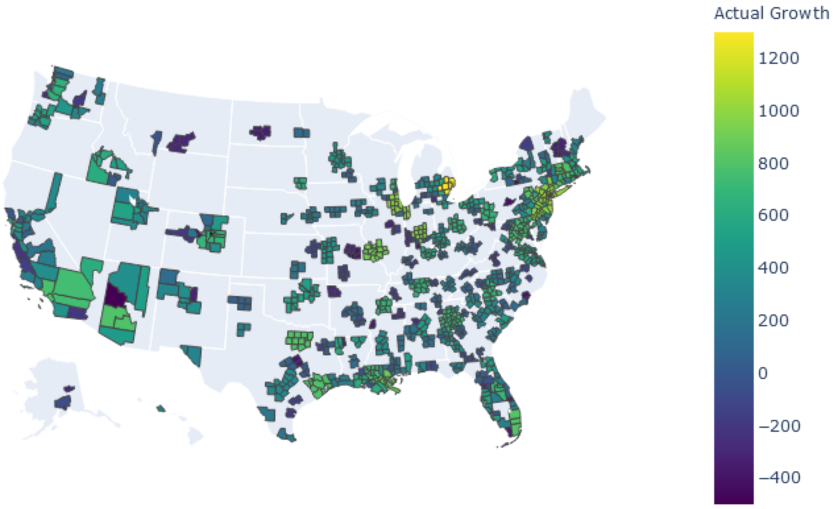


Figure 11: Actual Growth

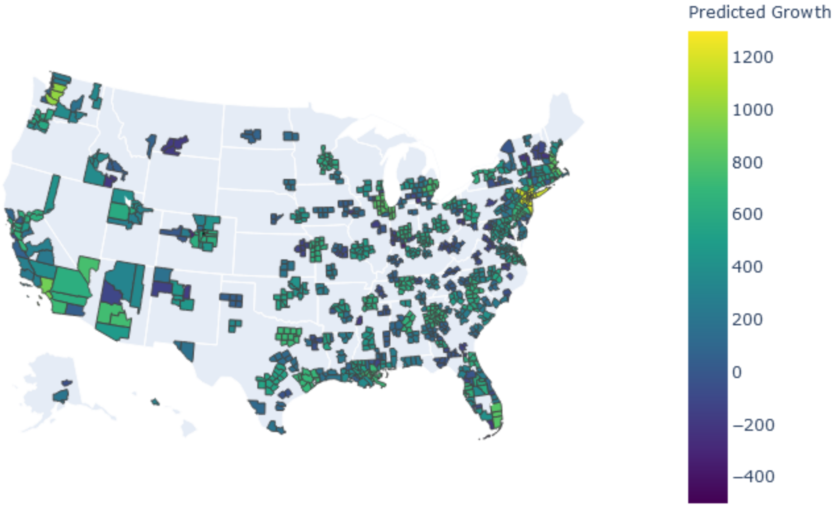


Figure 12: Model Predicted Growth

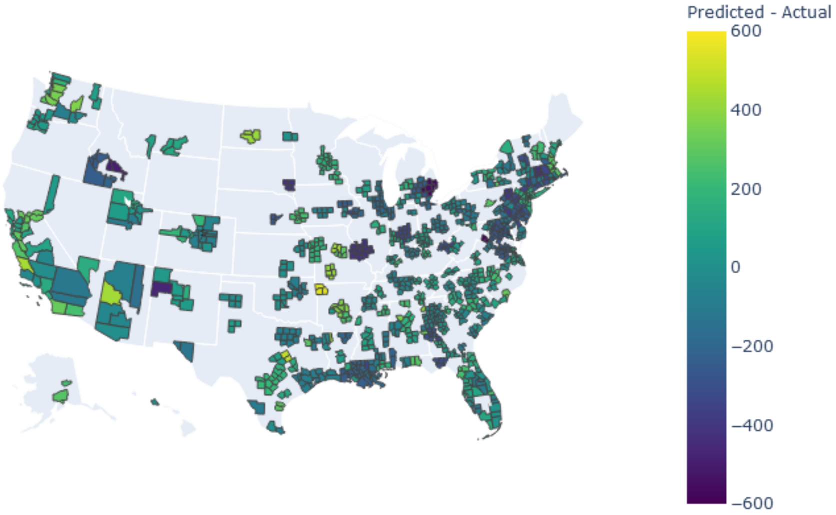


Figure 13: Residuals

The model is complete with the resulting resulting coefficients and their corresponding features in the table below.

Feature	Coef
threshold-Passed	-0.674068
Stay at home before	-0.249376
Rural-UrbanContinuumCode2013	-0.149160
Dem_to_rep_ratio	-0.119084
MedicareRatio	-0.049057
HPSAShortage	-0.043161
public schools	-0.034388
stay at home	-0.008328
>500 gatherings	0.008217
SVI Percentile	0.113961
PopulationEstimate2018	0.128582
public schools before	0.472725

Interesting patterns emerge from the table. First, the dates that the community spread began is the largest factor in this model. The later the spread started, the better the outlook was for growth. This can be attributed to better awareness and preparedness given more time. The next largest factor pertains to schools being closed before the spread really began. Interestingly, it seems that early school closures showed higher growth. Our suspicion is that there are confounding factors that are represented in this data point, which relates to things like properties of the metropolitan area or testing rates. Strangely, earlier dates for those orders were correlated with lower growth rates. It is the relative relationship between the beginning of the spread and those dates that acts as a predictor for a bad growth rate.

The features that had little weight on the models were health professional shortages, the percentage of people eligible for medicare, and population. This contradicted our assumptions as we thought that these features would have a profound impact on our model.

7 Limitations

Although our overall model is not accurate, we did the best we could with the limitations of our data. The dataset itself was not an accurate presentation of the intensity of the condition the country is currently facing. It lacked a way to gauge how many people were currently sick as testing was very limited. There were copious amounts of missing and inconsistent data, thus making the datasets unreliable. We resorted to novel methods to best estimate the status of the disease, however this is no substitute to accurate data.

Furthermore, there were various limitations in the analysis we did which limited our model. For example, we picked a short time-frame while filtering out rows, and areas that did not meet the threshold of 50 cases. As a result, we went from having 1841 to 300 metropolitan areas to use in training our model. An assumption that we made that had proven to be incorrect, was encountered when handling the data for South and North Dakota. The dem_to_rep_ratio feature in the dataset appeared to be missing. To compensate, we filled them with the value with 0 on the standardized dataset which indicated an equal ratio of republican to democrat. This proved to be incorrect as it held a significant difference in ratios, which directly impacted the growth rate of the virus.

There was limited availability in accuracy and quantity of the data, and thus limitation in the analysis we did. There are also ways we could improve our analysis. Having additional data on more dates could have provided us with more clarity in creating our model, as well as data points to our model. Information on how well the population is following the stay at home orders and testing rates of counties or metropolitan areas would help strengthen our analysis as they can provide a greater distinction between how fast the virus is spreading in different areas.

8 Ethical Dilemmas and Concerns

Finally, the data and its analyses are ethical issues themselves. There were various instances when we faced an ethical dilemma. Our model is designed to generalize high population sets, thus concentrating on metropolitan areas. This excludes rural areas by design, thus under-representing them. Another ethical dilemma was that because we visualized the counties and states, we have our own assumptions of why the data may be skewed or why the model is being overfitted based on our existing perspectives on these areas. We are drawing conclusions based on evidence that we do not have.

Drawing conclusions based on the evidence we don't have is a great ethical concern as we can fill the gaps of lack of evidence with our biases. As many minimum wage workers are essential workers as well, especially in lower-income areas, they are more likely to be exposed to the virus but are not represented in the data due to the limited testing and due to this and as impoverished areas have a larger lack of resources, the data collected may be underrepresenting its true severity of the area. Alternatively, the way the data and analysis was done assumes that the causes and blame belong to factors in the community itself rather than larger socio-economic contexts. We do not have the full details on the socioeconomic context of which the data is collected and even without these details this data belongs to humans and is intrinsically tied to identity, wealth, class, and a large number of other factors that are not explicitly stated. We may be creating situations of representational harm.

Another ethical concern that we ran into is the decontextualization of the project itself. When we are looking at the data we are extracting numeric values and forgetting the people tied to these values. We find ourselves to be disconnected to how this data is created. The deaths we used in our analysis are real people. As for decontextualization, being mindful is simple and yet important.

9 Evaluation and Limitations of Approach

Because of the difference in areas and when they met the threshold, we standardized growth to ensure consistency and this helped us determine what the growth values meant in terms of standard deviations from the mean. We chose a linear regression model because we have both continuous and categorical variables; however, as growth is not categorical itself, a limitation to this model is our low accuracy (on the training data) of 66% and (on the testing data) 72%. But because of the limited amount of data, we were not expecting a "good" score, thus we decided if our model was well fit through the cross validation root mean squared error and a scatterplot of its residuals. We were able to determine that our data had some issues when our initial cross validation rmse on the training data was 109 and had a few outliers on the residuals plot. From there, we were able to see how for some rows, their thresholdPassed and growth was highly skewed and fixed these issues accordingly to result in a cross validation rmse of 25. By standardizing, checking model scores and cross validation rmse scores and plotting the residuals, we were able to see complications in our feature engineering; however, our accuracy remains low due to lack of features and based on metropolitan areas

10 Future Improvements

If we could improve on this model in the future, we would use more recent data to include metropolitan areas that have met the threshold after April 18th and research another data set to use for more features to create a model for urban areas. Although a high accuracy was not our primary goal, in the future it would be great to utilize and create more features towards building a more accurate model for all areas in the United States.

References

CDC COVID-19 Response Team. (2020). *Severe outcomes among patients with coronavirus disease 2019 (covid-19) — united states, february 12–march 16, 2020y.* Retrieved 2020-05-12, from `https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm#contribAff`

Census.gov. (2020). *Delineation files: Core based statistical areas (cbsas), metropolitan divisions, and combined statistical areas (csas).* Retrieved 2020-05-12, from `https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html`

Mahase, E. (2020). *Covid-19: death rate is 0.66% and increases with age, study estimates.* Retrieved 2020-05-12, from `https://www.bmj.com/content/369/bmj.m1327`

Raheem, A. (2020). *Estimating cases of covid-19 from daily death data in italy.* Retrieved 2020-05-12, from `https://www.medrxiv.org/content/10.1101/2020.03.17.20037697v3`