Albert Tang
Jack Wang
Natasha Wong

Data 100: Final Project Narrative After Exploring COVID-19 Dataset

Abstract
Up until January 2020, most people have never heard of the coronavirus, also known as COVID-19. Due to its contagious nature, the virus has spread like wildfire across the globe, affecting many in terms of health, work, and stability. For this final project, we decided to tackle the COVID-19 dataset in order to predict the total number of confirmed cases and fatalities in the United States. This project also aims to explore COVID-19 through data analysis, projections, and visualizations. Three different models - linear regression, logistic regression, and lasso regression - were evaluated. In the end, the best model, which was linear regression, was able to predict the total number of confirmed cases and deaths with 100% and 100% accuracy, respectively, for the training set. In terms of performance on the validation set, it was able to predict the total number of confirmed cases and deaths with 99.99% and 99.98% accuracy, respectively.

Introduction
For the past three months, the news has been flooded with questions surrounding the status of COVID-19: how many deaths are there, is the lockdown slowing the spread, and when will there be a vaccination ready. During this time of uncertainty, schools have been closed or switched over to remote learning, jobs have been lost among many, and shortages have become more prevalent among the population - whether that be masks, toilet paper, or food. However, the most gruesome effect includes the death toll. The elderly and those who have an underlying medical condition are among those at higher risk of contracting the virus. All in all, efforts such as social distancing, have been enforced to slow down the spread of the virus in order to flatten the curve.

Because of its high presence in society, we decided to explore the COVID-19 dataset in order to *predict the total number of confirmed cases as well as fatalities in the United States at any given point in time*. In order to do so, our group used Google Colab as a means to collaborate. Additionally, we used the provided datasets, with the exception of replacing the 4.18states.csv file with the most up to date version (05/11/2020) provided by Johns Hopkins University's COVID-19 U.S. daily report on github. Using all of the datasets, we were able to create the following visualizations as seen below. As explained later, we will see how these visualizations helped shape our prediction model. We will also show the process of cleaning our data, making the visualizations, and training our model to achieve the best predictions.

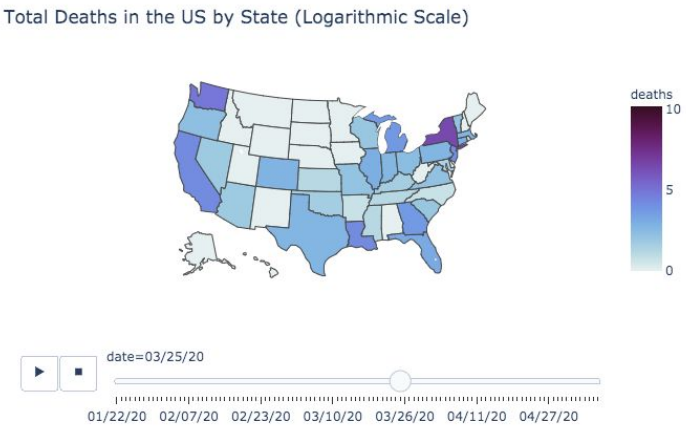

Figure 1: A snapshot of the timelapse displaying the total number of deaths, using a log scale, in the United States on March 25, 2020. As shown above, the number of deaths is captured at the state level.

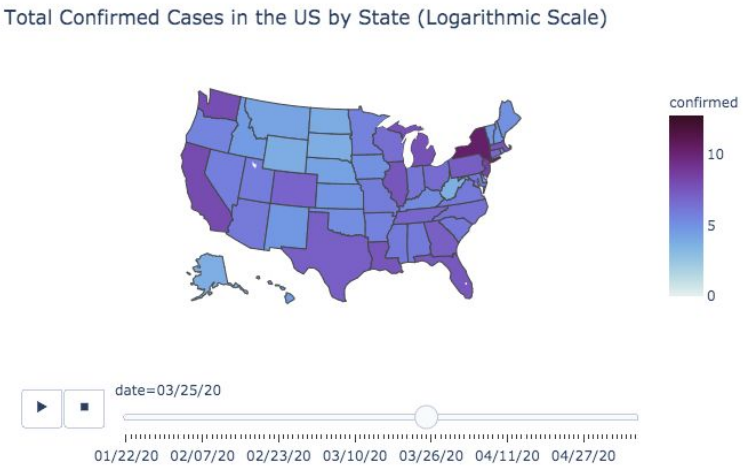

Figure 2: A snapshot of the timelapse displaying the total number of confirmed cases, using a log scale, in the United States on March 25, 2020. As shown above, the number of confirmed cases is captured at the state level.

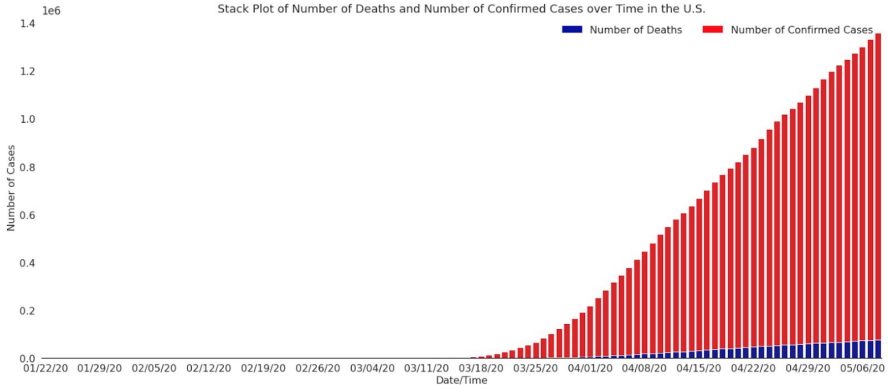

Figure 3: A stack plot of the number of deaths and confirmed cases over time in the United States, where blue indicates the number of deaths and red indicates the number of confirmed case

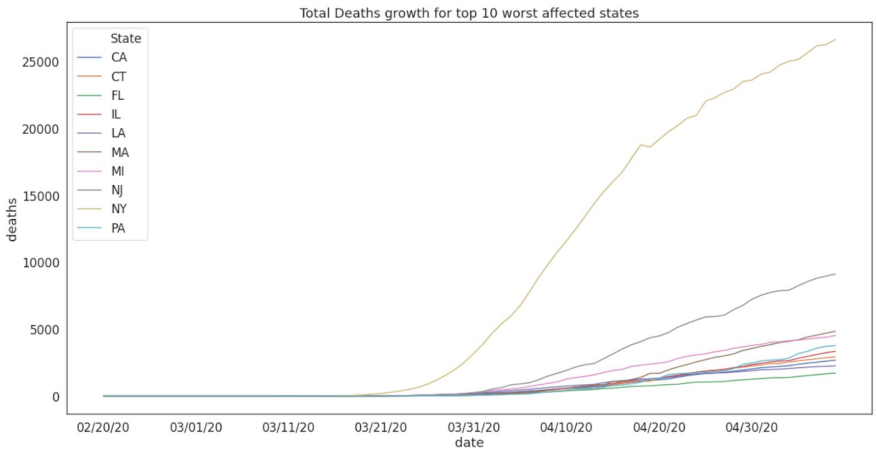

Figure 4: A line plot of the growth of total deaths for top 10 worst affected states over time in the U.S.

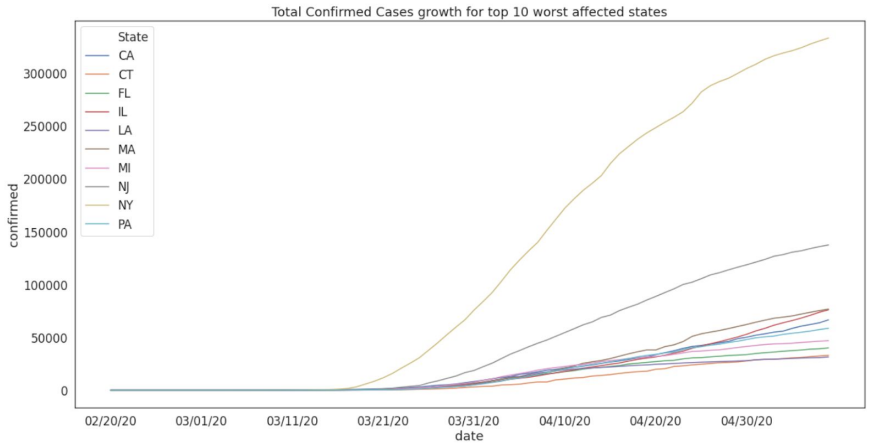

Figure 5: A line plot of the growth of total confirmed cases for top 10 worst affected states over time in the United States.

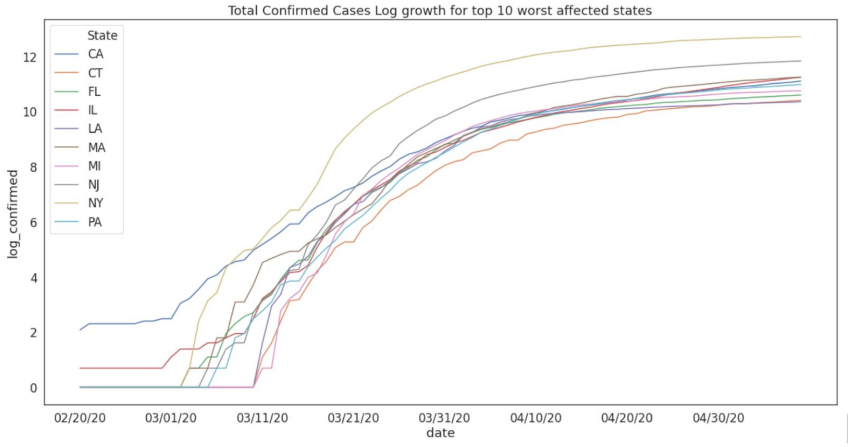

Figure 6: A line plot of the log growth of the total number of confirmed cases for top 10 worst affected states over time in the United States

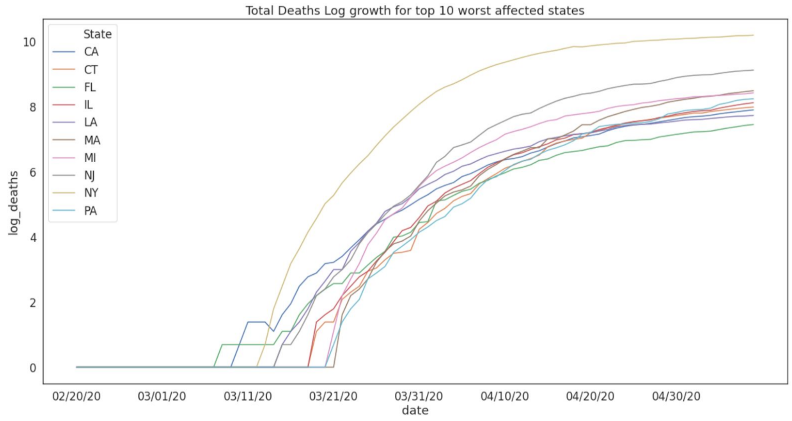

Figure 7: A line plot of the log growth of the total number of deaths for top 10 worst affected states over time in the United States

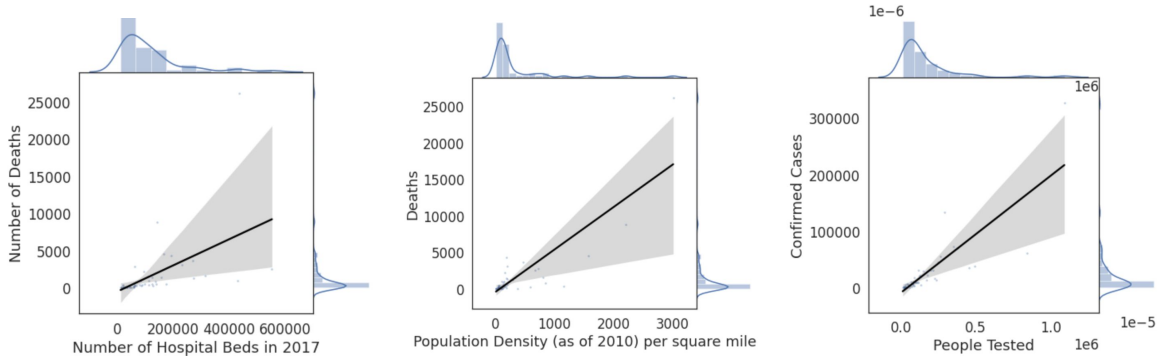

Figure 8, 9, 10: Joint plots used to determine feature applicability. Figure 8 demonstrates the correlation between the number of hospital beds and deaths. Figure 9 shows the correlation between the population density and number of deaths. Figure 10 displays the correlation between the number of people tested and confirmed cases

Description of Data

The provided datasets contain different information about COVID-19, ranging from daily cases and fatalities to hospital level data. Specifically, the time_series_covid19_deaths_US.csv dataset provides information on the total number of confirmed deaths in the United States, reported at the county level. It is also important to note that the reporting begins on January 22, 2020 and is updated daily to reflect present data. Additionally, the time_series_covid19_confirmed_US.csv dataset provides information on the total number of confirmed cases in the United States, reported at the county level. Similar to the time_series_covid19_deaths_US.csv dataset, the reporting begins on January 22, 2020 and is being updated daily to reflect present data. For both of the time series datasets, they would be useful for seeing the rate of change in terms of the total number of cases confirmed and deaths over time. Additionally, we can analyze how the numbers respond to stay-at-home orders being enforced over time.

As specified above, we decided to use the 05-11-2020.csv dataset rather than the provided 4.18states.csv dataset in order to have a more up-to-date depiction of what is currently happening. Additionally, there were many missing values in the provided dataset, so we used the most current one in the hopes that these missing values will be filled in. This dataset contains an aggregation of data at the state level. Information includes the number of confirmed cases, deaths,

and recovered as well as the different rates to consider for hospitalization per state. This dataset may be useful for seeing an overall snapshot of how well the country is doing at that point of time. Incident rates and hospitalization rates may be important features to consider when creating our prediction model as we will see later.

Lastly, there is the abridged_couties.csv dataset that includes information on county classifications, health professions, health facilities, utilization, expenditures, population, and environment which may be useful for determining whether or not a certain population is more vulnerable to the virus. Additionally, we can see how limitations in medical aid can be correlated to an increasing number of deaths.

Description of Methods

As a first step, we decided to look at the documentation for each of the datasets, understand what each of the columns represented, and formulate the question we wanted to answer. To reiterate, the question we wanted to address was predicting the number of confirmed cases and deaths in the United States at any point of time.

We quickly realized that the datasets contained many missing values while we were cleaning the data. We did the following to each of the datasets as an effort to do data cleaning: for the time_series_deaths and time_series_confirmed dataset, we converted the string dates to timedate objects in order to make it easier to group for visualizations. Since there were many missing values in the 'Admin2' column, we decided to drop it since it referred to a country identifier and was not needed in our analysis. The 'FIPS' column also had a couple of missing values. We decided to manually identify the locations for the 10 missing values by researching the FIPS code.

As for the states_summary dataset, if a province state had a missing 'Last_Updated' value, we decided to drop the entire row since it was more likely than not that the same row would have additional missing values. In the analysis, we also decided to drop all rows where the province state was not an official state in the United States, so that included dropping data on territories and foreign countries. For the remaining columns that had missing values, which were 'Recovered', 'People_Hospitalized', and 'Hospitalization_Rate', we used the mean of the column and used that to fill the missing value for that same column. For example, if there was a missing value in the 'Recovered' column, we would use the mean number of people recovered as the substitute value. This is because we wanted to avoid using 0 as that may negatively impact our analysis. We thought imputing them with the mean will provide some consistency with the existing data.

Lastly, for the abridged_counties dataset, we converted all the ordinal dates to timedate objects to make it easier to group for visualizations. For missing values in the columns that contain ordinal dates, we treated it such that the county has not implemented a policy yet. Thus, we decided to replace the missing values with a future date - December 31, 2020. Additionally, we dropped all rows that had more than 10 missing values, especially since this dataset contained the most missing values. In order to justify this action, we believed that having a lot of missing data in a single row was similar to the case where that row did not exist at all. After this initial phase of data

cleaning with the datasets, it was decided that for all remaining missing values, we would replace it with 0. We chose to use 0 as the replacement value because we believed that these columns served least important in training our prediction model.

Once the datasets were cleaned, visualizations and exploratory data analysis were conducted to better grasp the big picture. Additionally, these outputs allowed us to determine whether or not a certain feature would be appropriate for our prediction model. First, we tried finding correlations between the number of deaths and confirmed cases based on features related to hospitals and populations. We found that the population density proved to be a feature that we could use as there was a positive trend between that and death. We also found that there was a slight correlation between the number of hospital beds and deaths, which may have been enough to be considered a feature despite the amount of variation. Another feature we analyzed was the number of people tested against the total number of confirmed cases. The plot indicated that the number of confirmed cases was dependent on how aggressive a particular state was testing against the virus, and thus useful as a feature. Additionally, we tested 'PopulationEstimate65+2017', 'PopulationEstimate2018', and others as features, but they all had too much variation to be considered useful. Through trial and error, we found the following features to be the most useful in creating our prediction model: number of people tested, total population, 65+ population, and the number of deaths and confirmed cases each day per state.

After deciding the features, we split the time_series_confirmed and time_series_deaths datasets into their respective training and validation sets. The training set was used to help train our prediction model whereas the validation set was used once in order to see how well our model performed in the end after training. A thing to note is that our validation set serves a similar purpose as a test set would. We used three different prediction models: linear regression, logistic regression, and lasso regression. After creating all three models, we used the cross validation score to confirm that the accuracies found from the training set would uphold. Taking into account the cross validation accuracy as well as the training set accuracy for all three models, we concluded that the linear regression model was the best one to move forward with.

Something to note during this process was that we went through many trials and errors. In the beginning, we wanted to create a clustering model in order to determine which state counties were similar. If they were in the same cluster, then we could fit a specific model to them. We first used the affinity propagation model from sklearn since it learns how many clusters one should have. However, the model was not able to converge, even with an increase in max iterations. As a next step, we decided to use the k-means clustering method. Through this, we found that with a higher number of clusters, say 10, the clusters would fluctuate and remain inconsistent whereas with a lower number of clusters, say 3, the clusters would remain stable. However, there would be very small clusters present, providing not enough information to draw a good model from it.

Summary of Results

As previously stated, we decided to make three different models to experiment with: the linear regression model, the logistic regression model, and the lasso regression model. As found in Table 1 and 2, the accuracies for confirmed cases and deaths show that the linear regression model performed best, which is not too surprising. Based on the data provided, we can see that there is a linear trend in the total number of confirmed cases and deaths for most if not all states.

Table 1: Accuracies for Confirmed Cases

	Training Accuracy (%)	CV Accuracy (%)	Test Accuracy (%)
Linear Regression	100%	99.94%	99.991%
Logistic Regression	57.5%	NA	0%
Lasso	99.996%	42.23%	98.38%

Table 2: Accuracies for Deaths

	Training Accuracy (%)	CV Accuracy (%)	Test Accuracy (%)
Linear Regression	100%	99.98%	99.80%
Logistic Regression	17.5%	NA	0%
Lasso	99.997%	99.55%	99.90%

Analysis and Conclusion

Based on the experimental results, we can conclude that states in the US experience similar linear growth rates thus far, due to the near-perfect performance of the linear regression model. This is surprising because we expected the growth rate to be logistic in nature, but due to limitations of the logistic model, we could not verify this hypothesis. Data surrounding COVID-19 is still in progress and ongoing, so there may be inaccurate data as well as missing crucial ones. Stay at home orders are in the process of being lifted, so this may alter the number and rate of confirmed cases and deaths. Another key factor is that the disease has been experiencing constant and similar conditions for the past month due to the nationwide stay-at-home order, but as the population develops immunity or the disease mutates, we may see different orders of growth, whether it be exponential or a plateau.

Addressing the following 7 questions:

1. **What were two or three of the most interesting features you came across for your particular question?**

Two of the most interesting features that we came across for our particular question was the population density and the total number of people tested. We did not expect population density and the total number of people tested to have a somewhat decent positive correlation with the number of deaths and confirmed cases.

2. **Describe one feature you thought would be useful, but turned out to be ineffective.**

One feature that we thought would be useful, but turned out to be ineffective was the total number of hospital beds in 2017 ('FTEHospitalTotal2017'). We believed that the less hospital beds there were, the more deaths there would be. This comes from the idea of having ICU bed shortages and not being able to accomodate to every infected patient. However, this was proved false based on the correlation plot we graphed between the number of beds and death toll. Instead, the more beds there were, the more deaths there would be. This could be the case because a place with a higher population requires more hospital beds, and with a higher population means more people infected, even with the same or lower infection rate. Nonetheless, the correlation was dispersed and somewhat constricted to a certain region.

3. **What challenges did you find with your data? Where did you get stuck?**

Because the COVID-19 datasets are constantly being updated as it is an ongoing pandemic, it was difficult trying to address the missing values as well as provide justification for the actions we took to solve them. In the end, we decided to drop columns or rows, as appropriate, if the missing value did not provide information or lacked being a feature that could be used when training our prediction model. However, if it did provide value, we would use the mean of the column as the substitute rather than filling it with 0. We decided against using 0 because if the information was informative and important, we did not want it to negatively impact our analysis. For example, the 'recovered' column that had many missing values, but it gave insight and can be related to the question we were trying to answer. If we were to fill in these missing values with 0s, it wouldn't make sense for a state like California to have no patients recovering from the pandemic. Although inaccurate, by using the mean, there is slightly more consistency and its negative impact on our model would be less significant. An additional challenge we faced was being immersed in the pandemic. In other words, it was hard to focus on this project with everything going on in the world and living in this constant fear of loved ones contracting the virus. It was also hard to find time in between all of our schedules to collaborate on this project.

4. **What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?**

Some limitations of the analysis that occurred included the data itself. The provided datasets as well as existing datasets on github did not include information that we wanted, such as reopening dates. With the features that we chose, we cannot ensure that the model we built will be versatile enough to hold for future iterations in predicting later dates. Additionally, the missing values in the datasets seem to be a recurring problem. Because there were multiple debates and ideas on how to address these values, the assumptions that we made could be proved to be incorrect. As previously stated, we decided to use the mean as the filling value for important columns, and although slightly more consistent, the data that we imputed does not provide a true representation of the numbers that we are seeing in the world.

5. **What ethical dilemmas did you face with this data?**

An ethical dilemma we faced was addressing the missing values in the datasets. Although the decisions we made were justified, it was still hard trying to decide if we should drop or use the mean as a substitute. Another dilemma we faced with this data was the growing number of cases and deaths each day. Being outside poses a huge threat to the higher risk population, and with this data, we are able to see a true snapshot of what is happening around the world. Lastly, an ethical dilemma that we faced was the fact that it is still ongoing and present. The data that we have today may be completely different in the future based on unreported cases and deaths. In addition, it is also highly variable as there are many factors that can contribute to changing numbers, such as the number of people outside each day as well as the number of people who are still working.

6. **What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?**

Additional data that would be helpful in strengthening our analysis would be the true values of the missing ones in the provided table. Another would be the fatality composition. In other words, the demographic among those who have passed away due to the virus since it is claimed that the eldery and immuno-compromised patients are among those at higher risk. It would be nice to see what percentage of each is present in the death toll to better predict the total number of deaths. Additionally, it would be nice to have data on the reopening dates since many states are starting to reopen businesses and public places. This could be important to consider since this may play a role in the spread of the virus and thus the total number of confirmed cases and deaths.

7. **What ethical concerns might you encounter in studying this problem? How might you address those concerns?**

Some ethical concerns that we have encountered was the fact that it is still ongoing and present. The data that we have today may be completely different in the future based on unreported cases and deaths. In addition, it is also highly variable as there are many factors that can contribute to changing numbers, such as the number of people outside each day as well as the number of people who are still working. In order to address these concerns, we have to make sure we stay true to what we know and that is the current data that we have in our hands. An additional ethical concern while studying this problem was the additional data given each day. This ties back to the idea that it is still ongoing and that the numbers we have today may change tomorrow. Through this data we are able to see which states have the most deaths and confirmed cases. Following this information, another ethical concern we encountered while studying this problem was how the government will allocate the scarce resources. We addressed these concerns by looking at the numbers and the spread of infection among these high-risk states.