Data 100 Project Report

By: Nei (Alex) Fang, Derrick Sun, Ziyi Ding

Introduction and Part 2 Reflection

For this project, we are going through the process of cleaning, analyzing, modeling, and interpreting data to better understand COVID. This report aims to explain why we chose our model, how we improved on it, the problems we faced and how we overcame them, and what we learned over the course of the project. In part 2, we have completed both a guided modelling assignment as well as implementing the model outlined in our design doc. With some additional datasets, we hope to understand COVID cases per capita with respect to features we believe to have correlations with.

Additional datasets

We ultimately incorporated three additional datasets into our model: walking mobility by county (provided by Apple), COVID vaccination percentage by county (provided by CDC), and deaths due to COVID (CSSEGISandData, the same source as cases provided to us):

- https://covid19-static.cdn-apple.com/covid19-mobility-data/2203HotfixDev14/v3/en-us/applemobilitytrends-2021-12-07.csv
- https://data.cdc.gov/api/views/8xkx-amqh/rows.csv
- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv

Problem

Because the spread of the pandemic can be elevated by numerous factors, it is difficult for the government to coordinate policies to best control the growth of COVID. With the aim to offer governments a statistical perspective to determine the priority of public preventative actions, we hypothesized that there is a relationship between COVID cases per capita and the following factors: mask use, citizen mobility, vaccination, and testing, and of the four, we suspect mask use has the strongest relationship. The negation is that mask use doesn't have the strongest relationship of the four, which would suggest that it would be better to prioritize some other factor when it comes to reducing COVID cases.

However, we learned that state-level data was not as representative as county-level, but there was no publicly available data on testing by U.S. counties. Therefore, we chose to remove testing as a feature from our original hypothesis in order to obtain better representation. Furthermore, we increase our hypothesis complexity beyond simple correlation, resulting in a new hypothesis: COVID cases per capita can be predicted with a multiple linear regression

model using mask use, citizen mobility, vaccination, and death, with a correlation of 0.75 or above. The alternative hypothesis is that these features are insufficient to generate accurate predictions of cases per capita above the correlation threshold of 0.75. We can further test our hypothesis by creating a linear model and analyzing the coefficients.

We added the percentage of deaths as a feature in the modeling due to COVID. This was done because as the features and COVID cases per capita both increase the longer COVID is present in a county, our coefficients give the illusion that our features are related to an increase in COVID cases. Deaths due to COVID also increase the longer COVID is present and scale at roughly the same rate, allowing it to serve as a balance. We observed that once death was factored in, the coefficients for the other features turned negative, which matches our knowledge.

Both versions of our hypothesis have a null hypothesis and a way to be tested, so we are capable of confirming or rejecting them in principle. However, our initial hypothesis included COVID testing as a feature. While this could be tested with state-level data, there is no county-level data for testing, so the hypothesis could not be confirmed or rejected after we switched to county-level data. This means our initial hypothesis could not be confirmed nor rejected using existing data, which is why we chose to omit testing as a feature and change the hypothesis to focus on mask use, citizen mobility, deaths, and vaccinations instead.

Modeling

We chose a multiple linear regression model as our model. A multiple linear regression model is good for using numerical data to predict a numerical statistic, which is exactly what our hypothesis requires. Furthermore, the coefficients assigned to each feature can be used to fact-check.

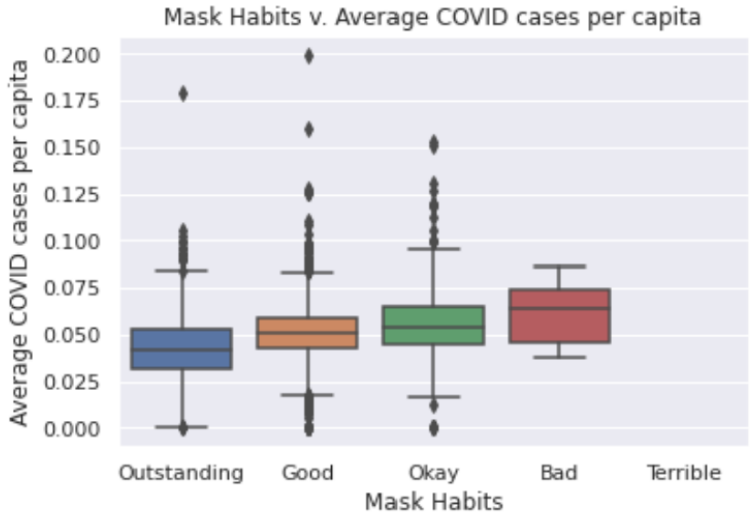
Other models do not fit our hypothesis. Logistic regression does not apply since we are not attempting to perform classification. While it is possible to attempt classification for individuals on whether or not they have COVID, we do not have that data. Decision trees and random forests do not generate the coefficients we are looking for, as they are made through greedy algorithms, which make decisions in steps and thus do not show general trends or relationships.

Our inputs are mask use, average mobility, percentage vaccinated of each county, and percentage of deaths due to COVID. Each variable is represented by a normalized column in our dataframe. Mask use is assessed by first assigning weight to each mask-wearing frequency from the given mask_use dataframe: -1 for proportions of never wearing masks, -0.5 for proportions of rarely wearing masks, 0 for proportions of sometimes wearing masks, +0.5 for proportions of frequently wearing masks, + 1 for proportions of always wearing masks. Then, we sum those proportions, multiplied by their weight for each county to generate a mask_use score, with a greater score indicating more residents are wearing masks more frequently. Average mobility is measured by Apple's mobility trend report dataset (Apple *Covid-19 - mobility trends*

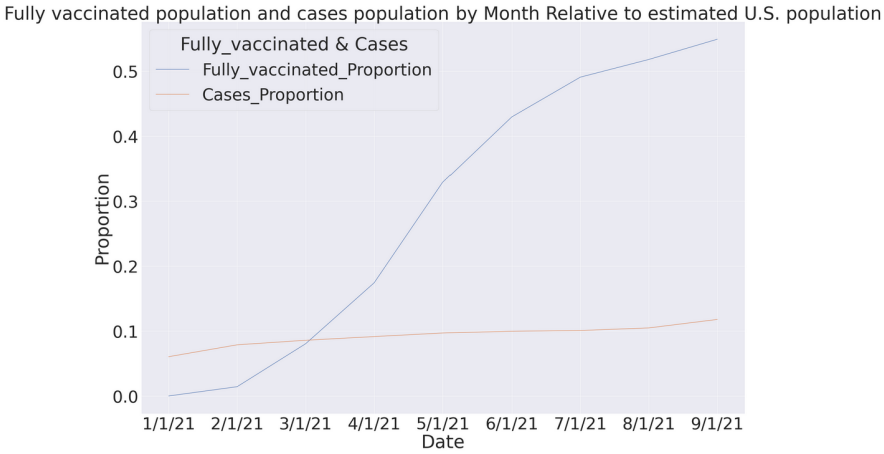
reports). It is a reflection on the number of daily direction requests by Apple users. We use groupby command to aggregate the sum of each county's mobility trends. The percentage vaccinated is calculated based on Centers for Disease Control and Prevention (CDC)'s provided dataset (Centers for Disease Control and Prevention *COVID-19 Vaccinations in the United States,County*). Since it is the biggest government agency in the U.S. that monitors and controls this pandemic, we expect this data to yield representative figures about vaccination population. The death data comes from the same source as the cases, with percentage deaths due to COVID.

Our output was the normalized number of COVID cases confirmed per capita. We had the data for the population for each county, so in order to find the cases per capita, we used the population estimate in 2020 for each county, dividing the number of cases by population. Since we were already standardizing our features, we also chose to standardize the COVID cases per capita. As for what date we chose to get our information from, we chose September 12, 2021, since that is the last date in our data and is thus the most recent.

Mask choice was selected because of our work on it in the EDA. It was found that there did seem to be a relationship between mask use and COVID cases per capita, as our boxplot showed:



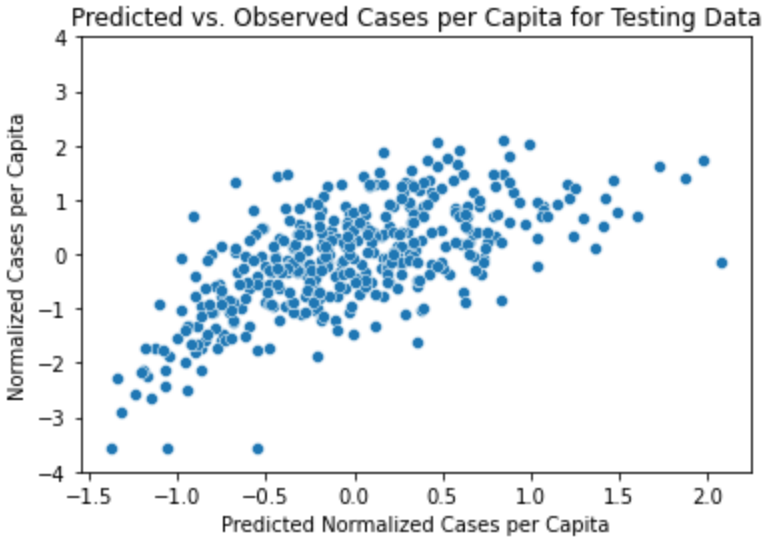
We also investigated the relationship between vaccinations and COVID cases per capita. It seemed that there was also a relationship there, as the slope of the cases proportion decreased slightly just as vaccination proportion started rising, but it was slight:



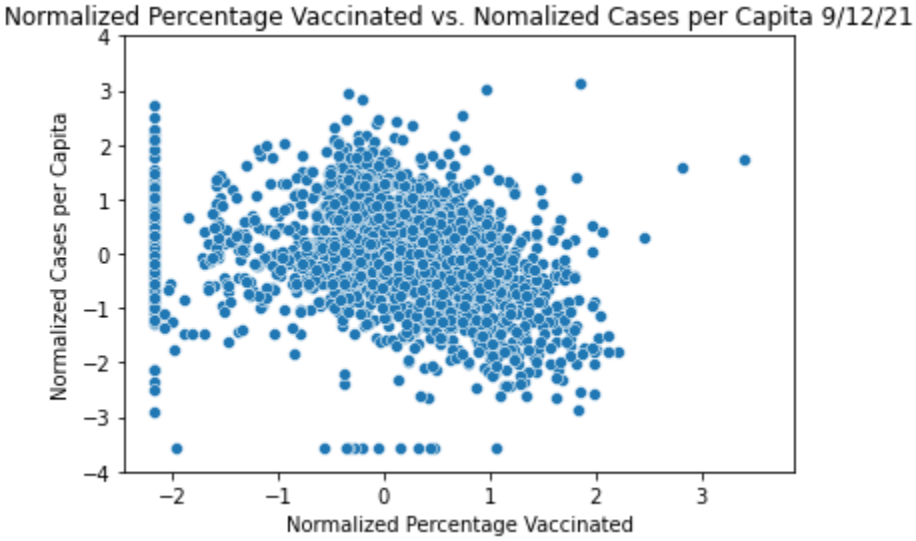
Thus, we had two factors that may be related to the COVID cases per capita, but we did not know how strong the relationships were. Hence the two were included as features in our hypothesis that we sought to compare. Mobility was introduced as a possible third factor in COVID cases per capita, given our personal knowledge suspected that there was a relationship, though we didn't know how strong or if the correlation was positive or negative. Percentage of deaths due to COVID was added later as another possible way to predict COVID cases per capita. All four features were normalized for the coefficients to be on the same scale, which we could use for fact-checking and gain deeper insight into the relationships between the features and the cases per capita.

Model Evaluation

With our model, we reached a cross-validation error of 0.756 and the correlation between reality and our prediction is 0.632 for the training set and 0.564 for the testing set. Thus, there is a moderately-strong correlation between our predictions and the observed values. We mainly evaluated the model through the cross-validation error and the correlation between our predictions and the observed values. We also plotted each feature against the COVID cases per capita in order to verify that a linear regression line fit. If it did not, we would have had to perform feature engineering to find a better way to model using the features. We believe that we have created a decent model with a correlation of around 0.6 between our predictions and the observed values. Additionally, it has a lower error than our initial model, which had a cross-validation error of approximately 0.89. Thus, we believe we do have a decent model. The following are the plots we generated while modeling.



This plot plots the relationship between the predicted and the observed cases per capita for the testing data. As we can see, there is a positive correlation between the two values. The correlation coefficient is about 0.564, and is roughly linear. The correlation coefficient indicates that there is a moderately strong association between our predictions and the actual cases per capita, so using a linear model fits and our model is strong enough to be used to draw inferences from.



This plot plots the relationship between the normalized percentage vaccinated against the normalized cases per capita on 9/12/2021. The correlation coefficient is around -0.22 with a clear negative linear pattern. However, the scatters below -2 for the x-value (normalized percentage vaccinated) does not follow the linear pattern: Instead, it forms a vertical line, suggesting that normalized percentage vaccinated have no correlation with cases per capita. This is rational because a county with an extremely low vaccination percentage can cause pandemic spread. Alternatively, it may be attributed to the fact that little COVID cases are present so citizens are less incentive to take the vaccination. The statistic confirms our hypothesis: as more citizens get vaccinated, there is a decrease in cases per capita.

Nevertheless, the general linear trends validates that a linear model is appropriate to explore the relationship between vaccination, our predictor, and cases per capita.

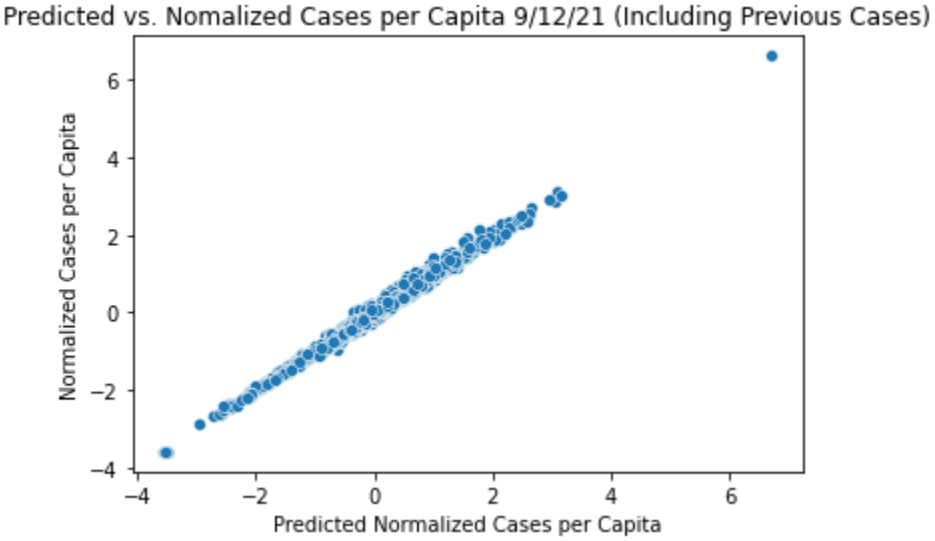
Model Improvement

Our model initially only included mask use, mobility, and percentage vaccinated. This presented two problems. First, our model was too weak - the correlation was 0.4, which was on the weaker end of the scale. And second, since the cases per capita increase over time, there was a positive correlation between our factors and the cases per capita. This flies in the face of common sense - mask usage and vaccinations should decrease the spread of COVID, not increase it. This is because different counties are in different stages of infection - some have had COVID for a long time and have already introduced every measure they can, while others have only just gotten COVID.

To solve this problem, we introduced death as a factor. Since death scales with COVID cases, it can counter the effects of how long COVID has been in a county, allowing for a better representation of how our features each affect the spread of COVID while also improving our model. After introducing death, correlation between predicted and actual number of COVID cases per capita increased to approximately 0.63, a large jump from the previous 0.4. And while we had no issue with error before, since we had nothing to compare it to, the cross-validation error of the new model decreased from 0.89 to 0.76. Our improved model gave us a prediction equation of:

$$f(x_1, x_2, x_3, x_4) = -0.2245x_1 - 0.0059x_2 - 0.0349x_3 + 0.5178x_4 + 0.0107$$

Where x_1 is mask use, x_2 is mobility, x_3 is percent of population vaccinated, and x_4 is the percentage of deaths due to COVID. When drawing our conclusion, we ignore death and its coefficient, as it is only there to correct the effect of time, which leaves masks as the feature with the greatest absolute value of the coefficient, at 0.2245.



Correlation = 0.997

Additionally, we include a previous day's cases per capita to boost the model accuracy. Above is the scatterplot of normalized and predicted COVID cases per capita on 9/12/21, using the exact same features with additional features of percentage death and normalized COVID cases per capita on 9/1/21. This resulted in the new prediction equation:

$$f(x_1, x_2, x_3, x_4) = -0.0056x_1 - 0.0067x_2 - 0.0338x_3 - 0.0057x_4 + 0.9882x_5 + 0.0003$$

Where x_1 is mask use, x_2 is mobility, x_3 is percent of population vaccinated, x_4 is the percentage of deaths due to COVID, and x_5 is the number of cases per capita on 9/1/21. The scatter plot outlines a roughly linear line with a correlation of 0.997. Therefore, combining our hypothesized model with a knowledge on a relatively recent day's COVID cases per capita, we can generate an extremely accurate prediction. However, since this improved model includes additional data - namely, the cases per capita on a previous date, it cannot apply to our hypothesis.

Conclusion and Future Work

In the end, we are able to conclude that based on our model, our hypothesis was rejected, since the model correlation is below our expectation. Thus, the features mask use, mobility, percent of population vaccinated, and deaths due to COVID are not sufficient to obtain predictions with correlation of greater than 0.75.

There are many improvements that could be made to our model. Including a previous day's COVDI cases per capita in the model, as discussed in the model improvement section, will significantly increase our model accuracy. Therefore, combining our hypothesized model with knowledge on a relatively recent day's COVID cases per capita, we can generate relatively accurate predictions. Moreover, due to our lack of access to information, many features that could affect the model are left out. For instance, data on testing by county, a more well-rounded walking mobility dataset including not just information from the apple users, data on population by county, as well as information on geopolitical factors including political affiliation and policy enforcement can all contribute to a more accurate model. All of these features can add to the accuracy of the model, helping people to make more informed decisions. Nevertheless, our current model still sheds light on the importance of some of the factors on COVID cases per capita, which can provide guidance to not only the government but also the citizens on how to better protect themselves.

Bibliography

Apple. "Covid-19 - Mobility Trends Reports." *Apple*, 2020, covid19.apple.com/mobility.

Centers for Disease Control and Prevention. "COVID-19 Vaccinations in the United States,County." 2020, apple mobility trends. Accessed 11 Dec. 2021.