Covid-19: Analysis and Predictions

Authors: Alan Liu, Dhruv Krishnaswamy, Varun Sewal
Prof. Ani Adhikari and Prof. Joseph Gonzalez
Stat C100

Abstract:

Over the past few months, the novel Coronavirus has become a huge cause of concern for countries across the planet, with massive economic and social costs caused by this virus. With no known cure for the virus as of now, the most important way to protect ourselves from this virus and survive this pandemic is to use our knowledge of data to better understand and analyze factors such as risk of dying and the susceptibility of catching the infection. In order to learn how to prevent deaths from the coronavirus, we attempt to find the correlation between the medical conditions surrounding an individual and their risk of death following a coronavirus infection. Additionally, in order to gauge the magnitude of the crisis, we will also try to predict the number of cases in the United States after a period of time, which will help provide context. Finally, in this project we also analyze how political ideologies play into the coronavirus pandemic and create a model that utilizes our analysis of political ideologies. To answer these questions, we follow the Data Science Lifecycle of Exploration, Analysis and Prediction.

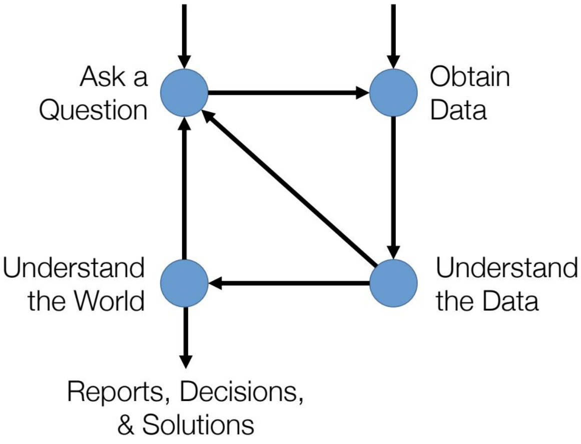

Figure 0: Data Science Lifecycle

Introduction:

To begin we use the datasets provided in the Covid-19 files. Upon looking at the contents of the CSV files in Excel, we notice that to answer our healthcare related questions, it would make sense to use the dataset, abridged_couties.csv as our primary data, referred to as *counties* from hereon, because it provides us with medical information separated by counties. Simultaneously, as we also require information pertinent to the Covid-19 pandemic in the USA, the dataset, confirmed time series dataset for the US proves useful, referred to as *confirmedCases* from hereon, with other provided datasets used as necessary. We notice that the time series dataset contains information for counties and territories belonging to the US, however, due to a lack of uniformity in the data from US territories, we choose to drop that information, and only utilize counties under the 50 states. We also create columns in this dataset which provide information such as total cases, total deaths and the proportion of infected people who have died for each county. In order to use this data frame with our time series, we must merge the *confirmedCases* data with the counties data. After merging, we notice that the new dataframe contains several NaN values, which are a result of the *counties* dataframe containing information for counties outside the USA, which are not present in the other merging dataframe, *confirmedCases,* as a result, we remove these NaN values, while also dropping the column *3-YrDiabetes2015-17,* because most values in this column are NaN and it makes it easier for us to remove the column in its entirety than deal with all the missing values. The resultant dataframe, *countiesWithDeaths,* contains 3075 listed counties and 188 columns.

Our Questions:

Q1) As we discussed above, to better understand coronavirus and the possible risk associated with it, it is important to identify how medical conditions and medical factors impact the death rate attributed to the virus, so that we can identify which people might be at a greater risk of suffering from this virus.

To answer this question, we created a correlation plot using relevant features from the dataframe *countiesWithDeaths* such as the proportion of the population that has diabetes, or heart disease, or are active smokers. We also looked at hospital conditions for the county, which told us if the county had the facilities to improve a person's chance of survival. We made this plot using the *sns.heatmap* feature and the resultant plot looked like this.

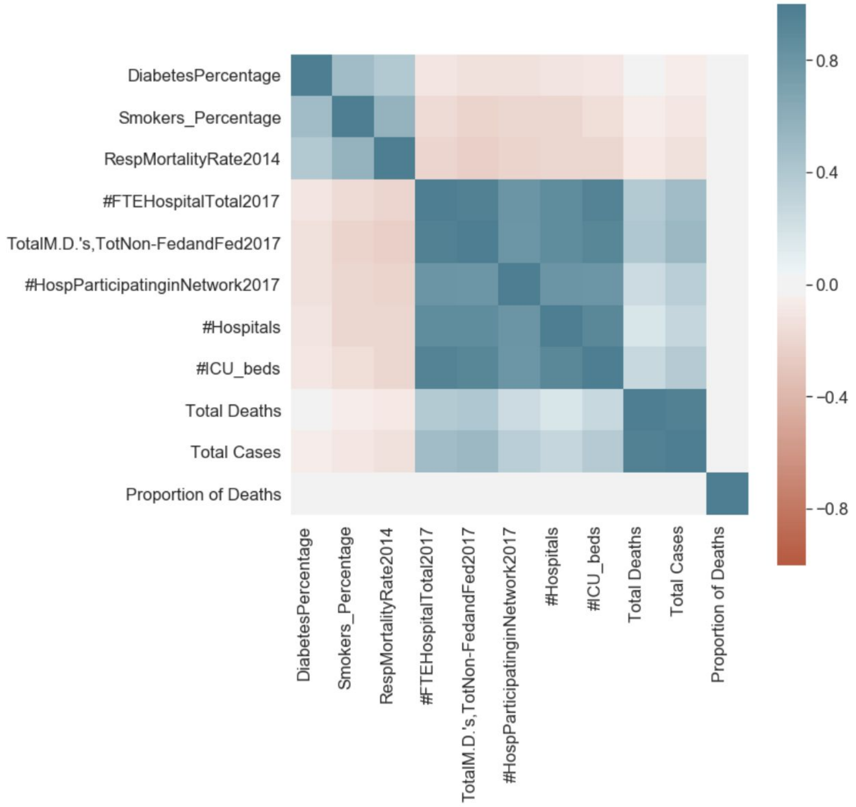

Figure 1: Correlation Plot between Medical Features and Coronavirus Death Rates

From Figure 1, we can make some very interesting observations. When we look at the correlation between Total Deaths/Cases and the number of Hospitals in the County and even the number of ICU beds, there is a positive correlation. This means that as the number of hospitals go up, so do the number of cases and deaths. However, this seems counterintuitive and is actually a great example of correlation not implying causation. We look at it further, by analyzing the case of New York. The average number of hospitals in counties in New York State is 2.66, while the average population density per square mile is 3014. In comparison, NYC, with the highest number of cases for a county in New York State, has 12 hospitals. If we simply look at the correlation, it appears that having more hospitals means there will be more cases. However, if we look at more information, such as the population density of the city, we can tell that NYC, with a density of 69484 per square mile, is much higher than the state average. And with studies backing the idea that COVID-19 transmits faster in urban counties due to the proximity between people, NYC being a hotbox for cases makes sense (Florida). Additionally, the high number of hospitals in the county can be attributed to the fact that the same area needs more hospitals to deal with the much higher population density. To sum this up, a large number of hospitals possibly means that the county has a higher population density, so it required more hospitals within the same space to deal with this density. The high population density in the case of COVID-19 leads to easier transmission of cases. So while it appears that a high number of hospitals implies more cases, it is not necessarily the case. Another interesting observation was the seeming lack of correlation between the number of deaths and the percentage of the population that smokes. Logically it may seem that smoking may make an individual prone to respiratory diseases, like COVID-19. However, recent studies have shown that this may not be the case, arguing that individuals who smoke, may have a stronger immunity to the virus (Economist). As a result, it is interesting to see an almost negligible correlation between Proportion of Cases that Die and Smokers Percentage for a county.

Q2) The coronavirus situation led to several states locking down, with California being the first state to shut down, on the 19th of March. Will shutting down the state early make a difference to the proportion of people in the population that got affected by coronavirus?

To do this, we plot a boxplot using the *countiesWithDeaths* dataframe. We notice that there are some states which have NaN as their stay at home, which implies that as of 18th April, when this data was published, they had not been shut down. To make the process of visualization easier, we provide a placeholder date to these NaN values. We pick the 18th of April since it does not coincide with any other lockdown date and doesn't skew our analysis, while also being the last reported day. We create a boxplot, with the different dates that states in the USA shut down as the different categories, and the Cases Per Capita as the values depicted on the y-axis. The plot that we see is depicted below in Figure 2.

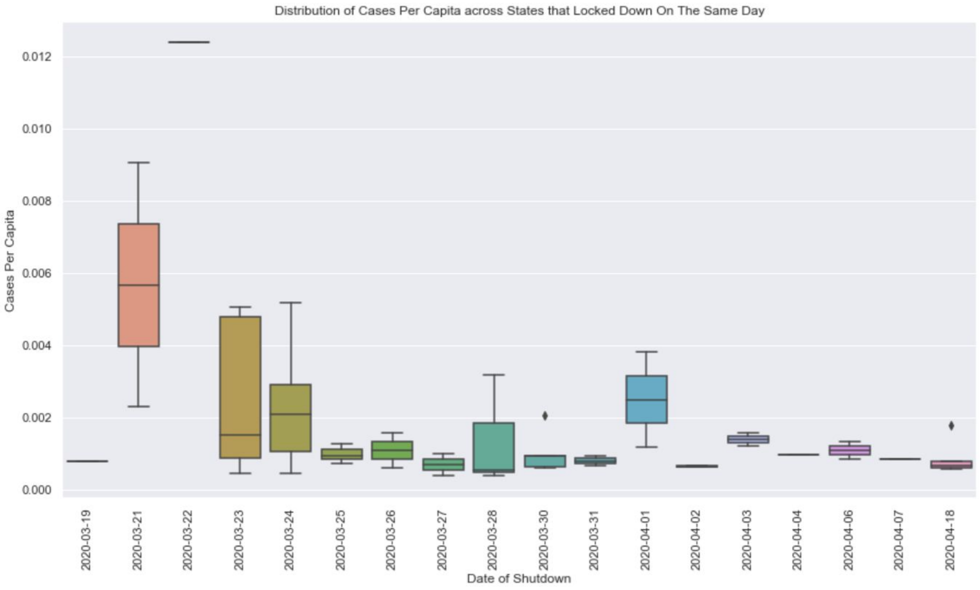

Figure 2: Distribution of Cases Per Capita Across States That Locked Down On The Same Day

This plot is interesting because it appears to say that states that shut down earlier had more cases per capita as compared to states that shut down a few days later . We see that although some states shut down early, they still had one of the higher cases per capita. States that shutdown before the 23rd of March included California, Illinois, New Jersey, Washington and New York. Although we might believe that shutting down early should result in a lower value of Cases Per Capita, some of the states that shutdown were already hotbeds for the spread of coronavirus, and it is possible that these lockdowns were enforced in response to the growing concern caused by the spread of the virus. Additionally, some states, such as Indiana, given their population density did not have as high a risk of spreading the virus, and as a result they shut down later, while also having a lower number of cases per capita than something like, New York, which sits with the highest Cases Per Capita in the US, while being one of the first states to shut down. Another analysis of the above boxplot distribution is that states that shut down later have lower cases per capita also because of the fact that once coronavirus started spreading in the country, and the first state went into lockdown, people living in the remaining states started practicing social distancing and avoided going out. As a result, by the time they were officially locked down, they had been practicing lockdown measures for a few days prior.

The final conclusion that we get from this data is when we look at the states that did not shutdown by the 18th of April, which was 11 days after the previous states had shut down. These states are: Oklahoma, Iowa, North Dakota, Arkansas, South Dakota and Nebraska. From data from the 2016 elections, all six of these states voted Republican. This is an interesting observation, because these 6 states make up 20% of the total 30 Republican states in 2016. This could be representative of the political ideologies of the United States and in fact, some reports by agencies such as Reuters have shed light on this, reporting that from a bipartisan survey of Americans, 88% of Democrats are in favor of shelter-in-place laws while only 55% of Republicans can say the same (Martina, Renshaw, Reid). This definitely calls for more analysis and is the basis for our third question

Q3) Given our analysis above which indicates that the response to coronavirus differs between the two parties, can we use this theory to make some classifications?

Using information relating to coronavirus and medical care for each county, we try to use this information to predict whether a county is democratic or republican.

We begin by converting the column dem to rep ratio provided in the counties dataframe into a column of 1s and 0s, where 1s represent Democrats and 0s represent Republican counties. To be able to test our future model, we make a training and validation split of our *countiesWithDeaths* dataframe. For our model, since we are attempting binary classification, we utilize a Logistic Regression model.We split this section into three cases, we begin Case I by using the 2 arbitrary features, Male Fraction of Population and Hospitals per person. These features are slightly arbitrary, with our intuition behind using them based on an article from The Atlantic. The article talks about the difference in political ideology as a result of the gender gap and each party's manner of dealing with this gender gap (Thompson). The Hospitals Per Person column is based on the analysis earlier, where less densely populated places had less hospitals. Since populations in rural areas tend to be more Republican, we used this as the second feature. For each case, we also use the non-default value for C, which is the inverse of regularization strength. We do so because it helps reduce variance in our model, as a result helping us deal with overfitting. Additionally, since our data is being regularized, we also have to standardize the data passed into the model. This is done to help with regularization, because now all of our weights and features are on the same magnitude, as a result our regularization penalty is better suited for our model.

This initial model was fairly accurate, providing us an accuracy of 79% on the Validation Set, however, upon looking at the precision and recall plot for the graph, we notice that there is a fair bit of drop off in our Precision, and as a result, we could improve upon this.

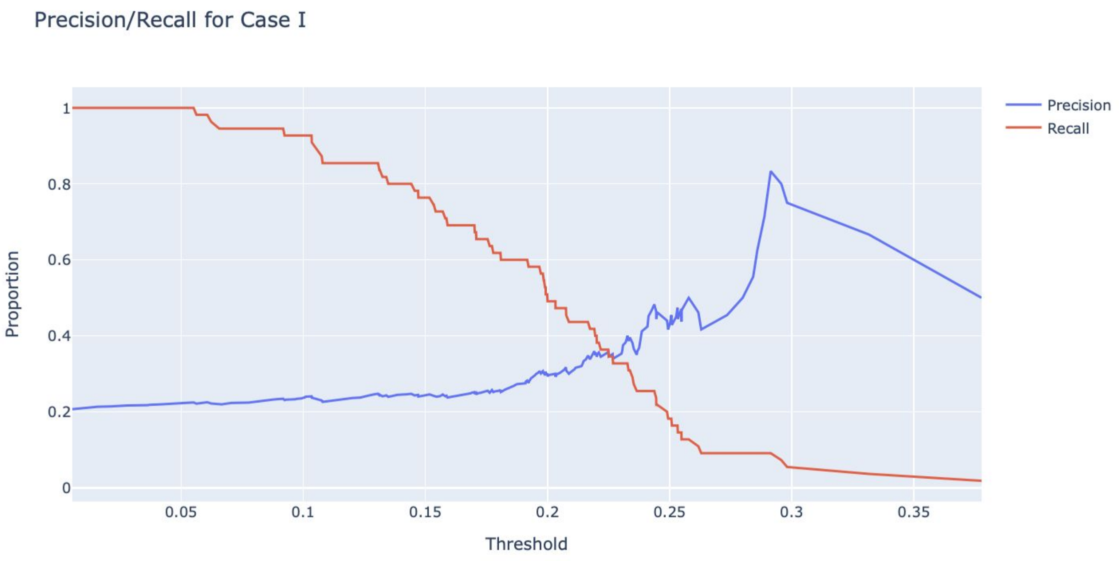

Figure 3.1: Precision/Recall Curves for Logistic Regression Model 1

We improve our model by taking into account more features in Case II, adding in the proportion of population above 60, because Republican voters normally tend to be older than Democrat voters. Additionally, we generalize the hospitals per county feature from the previous case and instead look at the population density, which proves to be helpful, taking our accuracy to a 82.33% on the Validation Set. Additionally, our precision and recall plot is significantly better, and now our precision has improved as well. Additionally, our cross validation accuracy on the training set is 86%.

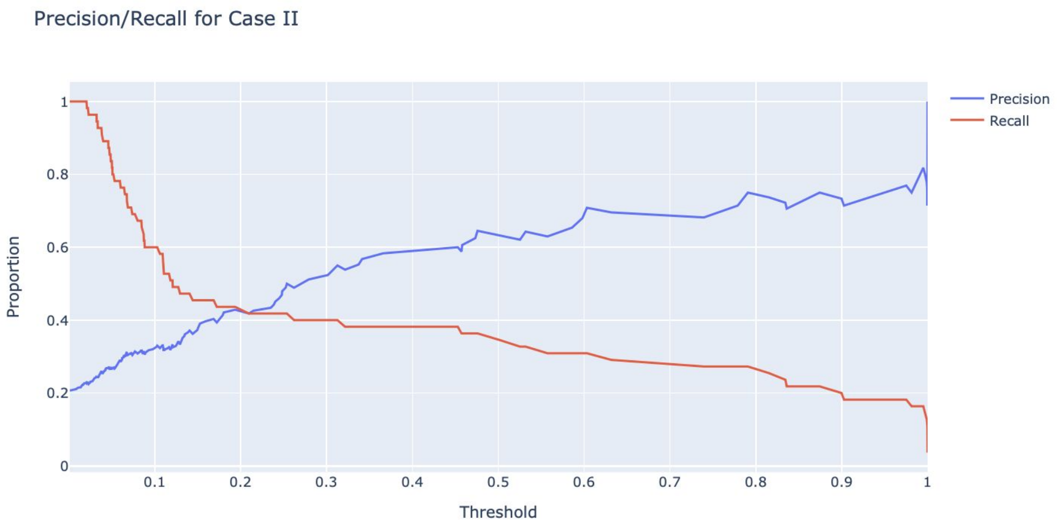

Figure 3.2: Precision/Recall Curves for Logistic Regression Model 2

Since we are trying to use information regarding a counties response to the Covid-19 pandemic to predict whether the county leans more Democratic or Republican, we now include information regarding lockdown dates into our model. To do so we One Hot Encode the values in columns such as stay at home, >500 gatherings ban and public schools shutdown. We one hot encoded this data to better express our categorical variables, which in this case are the varying dates that these laws were enforced on. Using this model, we are able to improve our accuracy a lot more, with a new cross validation accuracy of 89.39% and an accuracy of 85.33% on our Validation Set. Additionally, our precision and recall plot has improved as compared to case II as well. As a result, this is our most accurate classification.

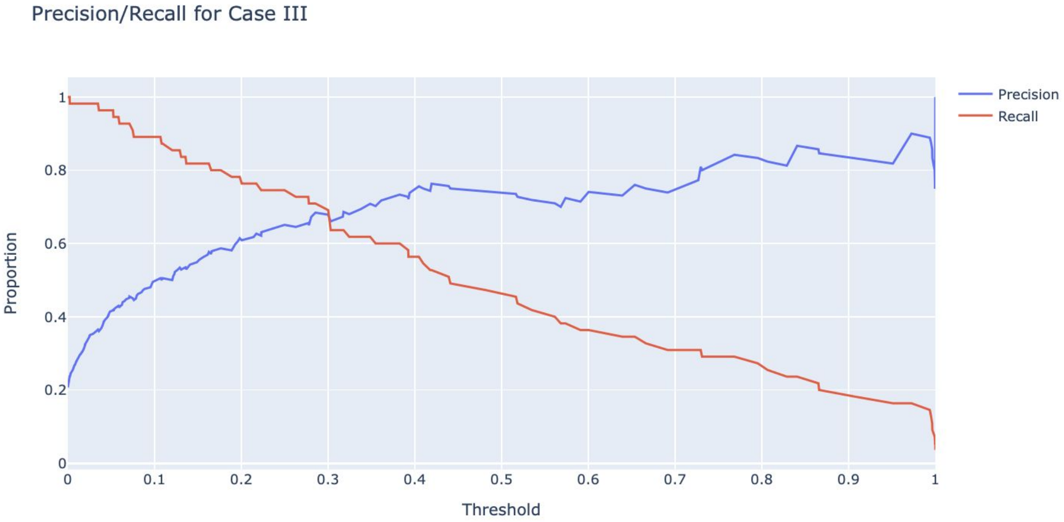

Figure 3.3: Precision/Recall Curves for Logistic Regression Model 3

Q4) Our final question attempted to see if we could use the health data that we have to predict the growth of Coronavirus cases in the United States.

For this aspect, we attempt to forecast the future number of cases in the United States using the time series dataset, *confirmedCases.* The growth of coronavirus cases in the US has been rapid and as of the 18th of April, this is what the trend looked like for the nation.

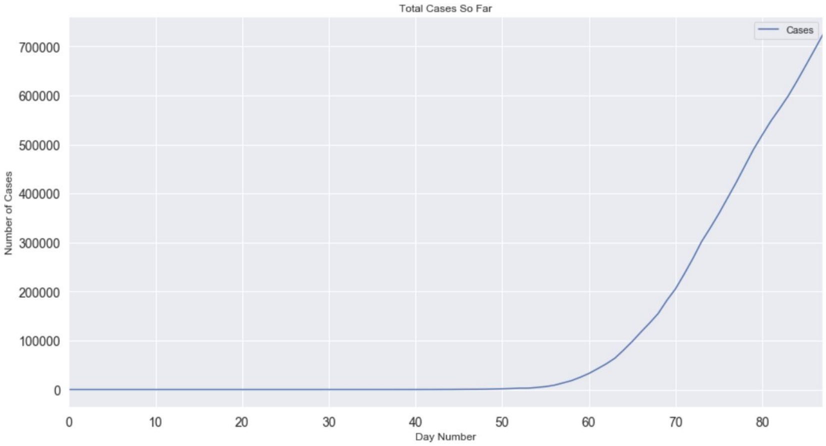

Figure 4.1: Total Cases per Day for USA (as of 18th April)

To predict the future number of cases, we needed to aggregate our *confirmedCases* data which had a time series for each county in the US, into a singular entry representing a time series for the entirety of the country. To predict the future number of cases, we use a model known as Holt's Linear Trend Method. Holt's is a model that uses linear exponential smoothing to forecast data which has a trend.

$$Forecast\ \hat{y}_{t+h|t} = \ell_t + hb_t$$

$$Level\ \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \qquad Trend\ b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$

For our case, we used the Holt's Linear Trend Model in the statsmodels package, which allowed us to forecast easily. In addition, we also used another model, known as the Additive Dampened model to forecast future coronavirus cases. This model takes into account seasonality and dampened trends, and we utilized it in order to capture possible non-linear variations in the time series data.

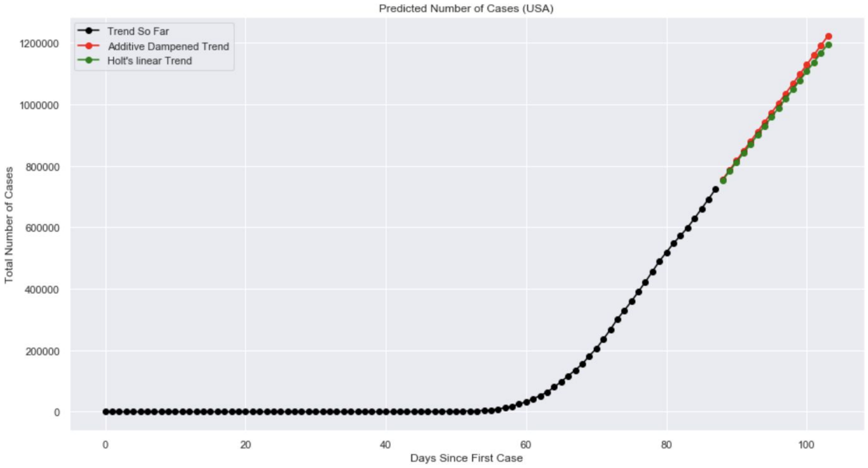

Figure 4.2: Prediction of Cases till May 4th for USA

To verify the accuracy of our prediction, we downloaded the updated time series from, https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series, and we use Root Mean Squared Error as our method for measuring error between our predictions and the actual reported cases in the USA, up until May 4th. Using this method, our errors for Additive Dampened and Holt's Linear Trend were, 20768.2 and 6515.4 respectively. This is a low error value, particularly using the Linear Trend method, and from the plot below, we can see that both the models chosen by us were fairly close to the actual coronavirus case count.

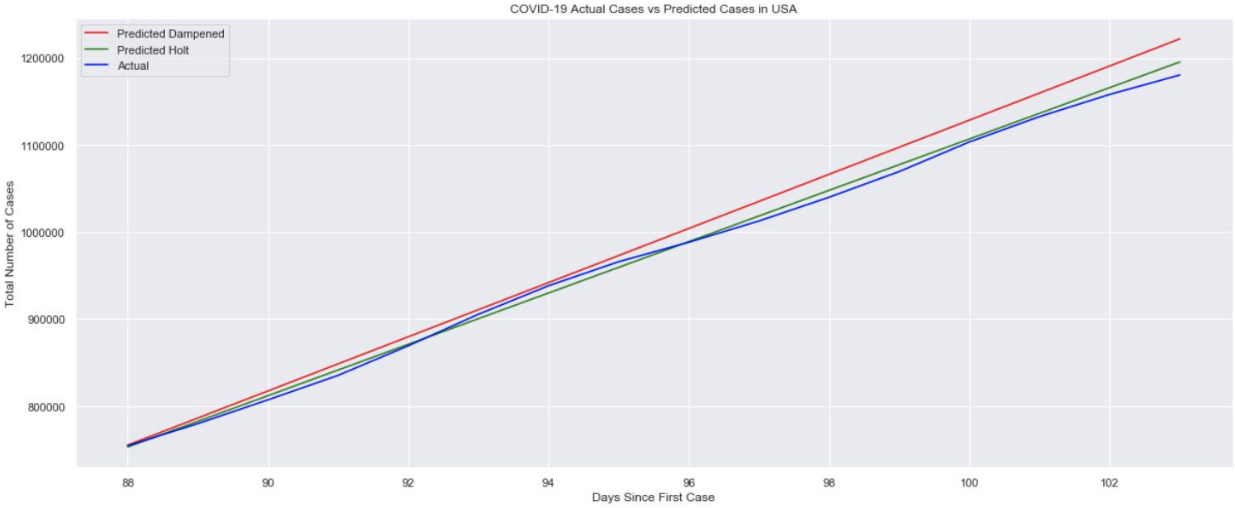

Figure 4.4: Comparison between Predicted Cases and Actual Cases between 04/18 and 05/04

Analysis of Our Models and Methodology:

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

The two most interesting features we came across were the shutdown dates and proportion of people aged above 60. Using a one hot encoding for the different shutdown dates of states proved to be beneficial in increasing our model accuracy in the Q3 classification problem. The proportion of people aged 60 and beyond was a feature created by us. It proved extremely useful in our logistic regression model used to answer Q3

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

Given the recent media reports regarding smokers having a higher immunity to the coronavirus, we believe we could use this aspect in our analysis to possibly predict the sort of people who were more susceptible to dying from the virus. Unfortunately, when we actually plotted this out, we were unable to see a strong negative correlation, so it was not useful to build a model out of.

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

The most difficult task was that of cleaning the data. Particularly dealing with NaN values. Considering the fact that some of the columns had a large number of NaN values, which could not work in some of our modelling and visualizations meant they had to be dealt with. Additionally, converting the data from the csv file we download to actual, useful data was a long process. Each question that we asked required us to manipulate the data in a new manner which was tricky.

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

Although we try to avoid it, in the analysis where we attempt to understand the correlation between a high number of hospitals and proportion of coronavirus patients who die, we possibly generalize the situation too much. Even though it seems logically accurate, that states with a higher population density have a higher overall demand for healthcare, and as a result more hospitals, we do not have any proof to back this theory. Our ideas here are based on our intuition, possibly naive, and as a result our theory about the correlation in a high number of hospitals, and a corresponding high proportion of deaths could prove to be incorrect.

Additionally, since our Logistic Regression model does not like to take NaN values, we must appropriately deal with these values, without throwing that aspect of the data away, because that would be inefficient. So, we deal with these values by replacing NaN values in columns with the average of all the other values in the corresponding columns. While this is not a terrible assumption to make, it is a generalization, and in a situation where we have a considerably large number of NaN values, bu replacing those values with the average of the remaining values, we might end up with far too many identical values, which will as a result not help us as much as it should in the task of classification.

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

Apart from generalizing NaN values as the average of the remaining values in the column, which as we discussed cause a disparity in the data and leads to error in prediction.

More importantly, I believe that some of the questions that we attempted reinforced our biases. We noticed that the 6 states that had not shut down as of the 18th of April were all Republican, and as a result we began digging further into this. While this may not be a false conclusion to reach, it is possible that the states remained open for factors other than the political beliefs, but we made these conclusions which support our inherent political bias.

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

If we had more information specific to counties regarding testing rates, we would be able to make better predictions, because reports and our data have shown us that not every state tests the same way. Additionally, using the current testing information which we have for each state and applying the same value to each county within a state could work, however it is more likely to be an overgeneralization, with testing unfortunately dependent on the socioeconomic status of the county, with the privileged more likely to get tested early.(Hickok) As a result, for large states, like California, where there is a disparity in socioeconomic conditions within the same state, generalizing data would be a massive error. However, if we had the data to back our idea, we could use it to back the hypotheses that counties with a higher working class proportion, vote Republican, and due to the fact that a larger proportion belongs to the working class, the county might have lower socioeconomic conditions, leading to a lower chance of getting tested. (Florida)

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

Since we attempted to identify the political ideologies of a state using their response to Covid-19, we could be misclassifying counties by misrepresenting their beliefs when it comes to the Covid-19 pandemic. We are essentially generalizing the coronavirus beliefs of all democratic counties into one bucket and similarly all republican counties into one. This could be problematic if misused by someone with much more authority than us, because it will leave people who do not agree with the ideology but fall under these buckets by not taking account of their beliefs regarding the coronavirus. To address this, we should realize that this data is simply not enough to decide government response off, people's sentiment needs to be considered before responding to the pandemic.

Summary:

In this project, we answered questions such as the correlation between medical factors and coronavirus death rates, where we noticed that contrary to latest medical evidence, we have not been able to find a correlation between Covid related deaths and the percentage of smokers, as seen in Figure 1. We also identified how political ideologies may shape a state's response to the pandemic, which was done using a boxplot in Figure 2, and using this analysis, we aimed to classify counties into democratic or republican using what we know about their coronavirus response and general information about the demographic. Our model successfully predicted this with a validation accuracy of 85.33%, which was achieved with a combination of cross-validation, regularization and precision/recall analysis. At the end of it, we also developed a model that accurately forecasts the future number of coronavirus cases in the USA, and our prediction till June 6th is in Figure 5.

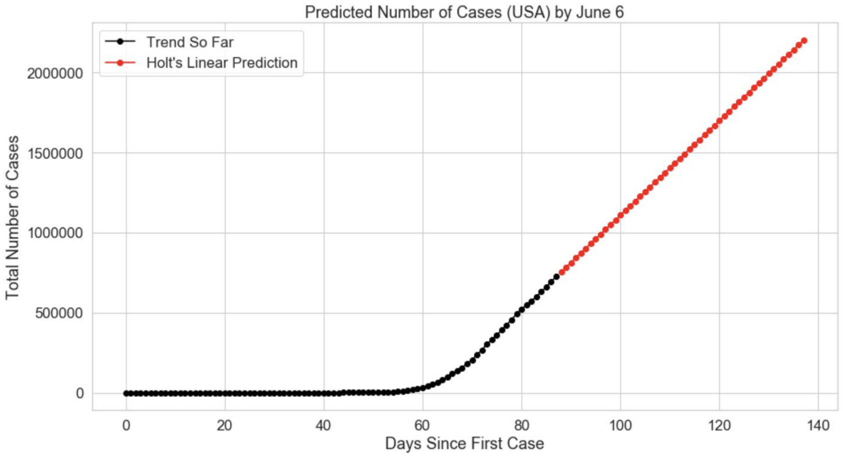

Figure 5: Predicted Number of Cases in the USA by June 6

The situation is pretty dismal, with our prediction set to cross 2 Million by then, and as a result, it is important to continue using our data science tools to improve our decision making when it comes to this pandemic. To achieve this, our idea of using political ideologies to understand coronavirus response can be used by the government at a large scale, with a closer eye to detail, in order to create a plan that undertakes a state-specific response, while also taking into account public sentiment.

Works Referenced:

1. Thompson, Derek. “Why Men Vote for Republicans, and Women Vote for Democrats.” *The Atlantic*, Atlantic Media Company, 9 Feb. 2020, www.theatlantic.com/ideas/archive/2020/02/how-women-became-democratic-partisans/606274/.
2. Florida, Richard. “What We Know About Density and the Spread of Coronavirus.” *CityLab*, 17 Apr. 2020, www.citylab.com/equity/2020/04/coronavirus-spread-map-city-urban-density-suburbs-rural-data/609394/.
3. Martina, Michael, et al. “How Trump Allies Have Organized and Promoted Anti-Lockdown Protests.” *Reuters*, Thomson Reuters, 22 Apr. 2020, www.reuters.com/article/health-coronavirus-trump-protests/how-trump-allies-have-organized-and-promoted-anti-lockdown-protests-idINKCN2240J5.
4. “Smokers Seem Less Likely than Non-Smokers to Fall Ill with Covid-19.” *The Economist*, The Economist Newspaper, www.economist.com/science-and-technology/2020/05/02/smokers-seem-less-likely-than-non-smokers-to-fall-ill-with-covid-19.
5. Florida, Richard. “How Occupational Class Influences U.S. Voting Patterns.” *CityLab*, 29 Nov. 2018, www.citylab.com/life/2018/11/state-voting-patterns-occupational-class-data-politics/575047/.
6. Hickok, Kimberly. “States Aren't Testing Uniformly for Coronavirus. That's Creating a Distorted Picture of the Outbreak.” *LiveScience*, Purch, 27 Mar. 2020, www.livescience.com/coronavirus-testing-us-states.html.
7. “Oracle® Hyperion Planning Predictive Planning in Smart View User's Guide.” *Moved*, 3 Nov. 2016, docs.oracle.com/cd/E57185_01/CBPPU/damped_trend_additive_seasonal_method.htm.
8. “Forecasting: Principles and Practice.” *7.2 Trend Methods*, otexts.com/fpp2/holt.html.