Coronavirus Cases at the County and Global Level

Written by Kunal Adhia, Sravya Basvapatri, and Shreya Mohanty

Data 100 Spring 2020 Final Project

Final Project Jupyter Notebook Outline

Appendix Jupyter Notebook Outline

I. Introduction

We were interested in understanding factors that affect the spread and fatality of COVID-19, as well as predicting how the pandemic would affect different areas in the coming weeks. We looked at three main questions:

- Which factors affect mortality rates in the United States?
- How can we predict cases per capita in United States counties, and can we make predictions for the next month?
- How can we predict the number of cases in the world, and can we predict the next month?

II. Abstract

We first raise the question of which United States counties are affected the most by coronavirus cases, analyzed in terms of morality rates. This builds an understanding of which counties are hit the hardest by the pandemic. This information would be used to make more vulnerable populations aware of measures they need to take and help us as a society provide more equitable care, bringing resources to areas that need them the most.

We then analyzed the number of cases per 100K population in U.S. counties, creating and training a model to predict the number of cases per 100K population. In the process we identified features that were useful in predicting the number of cases, as well as using time series data to predict the number of cases one month into the future. We used a linear regression model using lasso regularization, but it tended to underpredict the number of cases.

Shifting our focus to a broader global scale, we wanted to predict how the number of cases would change into the future. Fitting a logistic curve, we were able to predict the number of cases in world countries for both one week and one month into the future. This information would help countries that are still at the beginning of their curve understand how to best prepare for the pandemic.

Datasets Used (Sources Linked at the top of Jupyter Notebook)

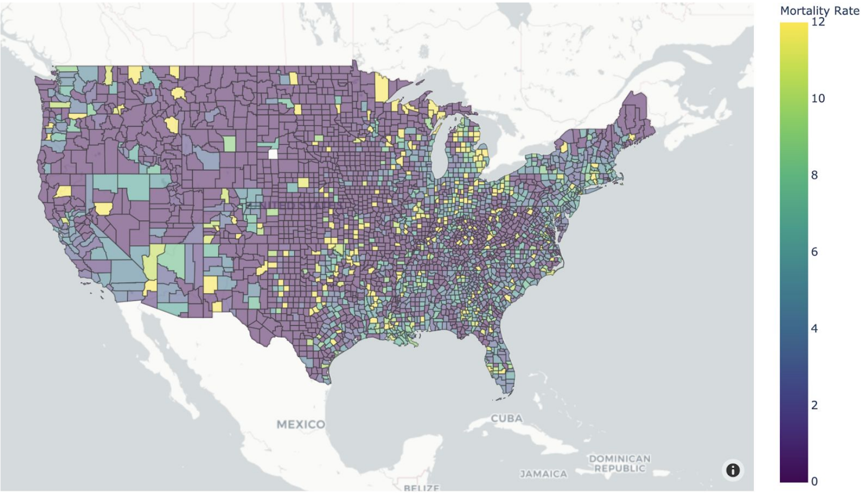
- covid_confirmed_usafacts.csv : time series data for confirmed covid-19 cases in the US
- covid_deaths_usafacts.csv : time series data for confirmed covid-19 deaths in the US
- covid_county_population_usafacts.csv : population counts by county as of 2019 in the US
- csse_covid_19_daily_reports_us/05-04-2020.csv : covid-19 cases information by state in the US
- county_data_abridged.csv : census and demographic information and health risks in US counties
- us-county-health-rankings-2020.csv : health related statistics in US counties
- time_series_covid19_deaths_global.csv : time series data for confirmed covid-19 cases in the world
- time_series_covid19_confirmed_global.csv : time series data for confirmed covid-19 deaths in the world

III. In the United States: Analyzing Mortality Rates in U.S. Counties

Question: Our first question was understanding how mortality rates, measured in the number of deaths divided by the number of cases, varied across the United States. We were curious whether factors such as rurality, health risks, poverty, and population density affected how deadly coronavirus is.

EDA: We used the 'covid_confirmed_usafacts.csv' and 'covid_confirmed_usafacts.csv' and the CSSE state daily report for 5-04-2020. We filled FIPS codes to be appropriate length strings to merge datasets on this primary key. We created columns for 'Mortality_Rate' at both the state and county level, and used these to create visualizations that informed the analysis. Plotting mortality rate by state was fairly homogenous, so we plotted a more granular mortality rate by county. It's difficult to see a geographic trend, so this motivated analysis into factors influencing mortality rates.

Image: Mortality Rates by U.S. counties



Data Cleaning: After EDA, we performed further data cleaning, selecting relevant columns from 'county_data_abridged.csv' and adding an updated 2019 ‘population’ column to replace the 2018 population counts. We found that after joining, there were 16 null values in the dataframe of 3141 counties, mostly from sparsely populated boroughs in Alaska. We filled these values with the average from the state.

Finally, we merged with the earlier dataframe we created with the number of cases, deaths, and mortality rate by county. We added additional features that provided 65+ population estimates, the % of people affected by the virus, and cases & deaths per 100K population.

Data Visualizations & Analysis: We wanted to understand how different features corresponded to higher mortality rates. We first took a look at the Rural Urban Continuum index. We created box plots, treating each value on the index as an ordinal categorical variable.

We see that the overall mortality rates are lower in more rural areas, which makes sense because the virus is less likely to spread in rural areas. These are likely counties with few cases that recovered. However, when we look only at positive mortality rates (counties that have had at least one COVID-19 related death), we see that mortality is positively correlated with the rural-urban continuum code of a county. This hints that cases in rural areas may be more fatal, perhaps due to other factors, such as availability of health care. It could also simply be because these counties had a small number of cases, of which most happened to be fatal.

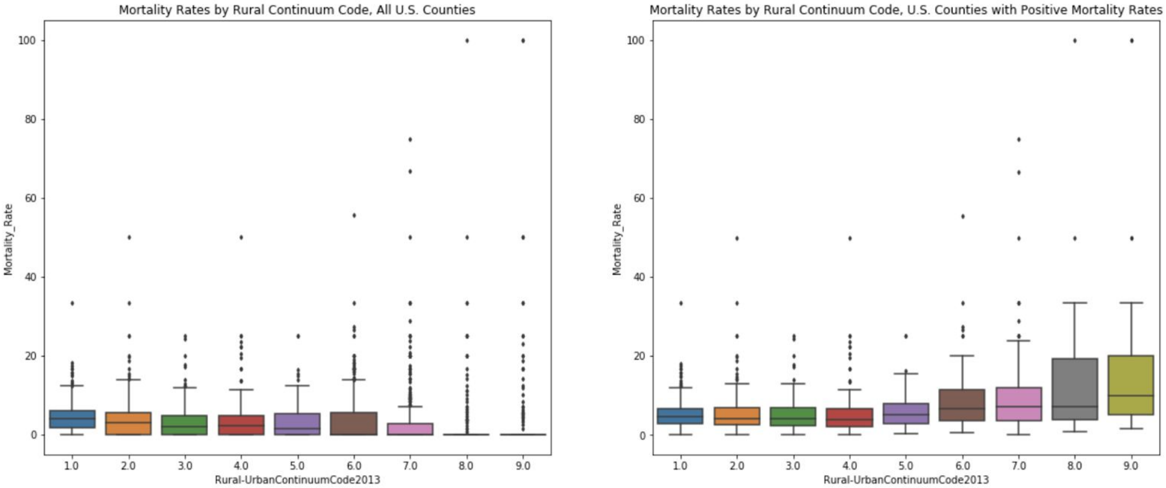


Image: Boxplots of Mortality Rates by County Rural Urban Continuum Index (All & Positive Rates)

Next, we wanted to see if mortality rates were linked with the population density of a county. Each scatterplot encodes the Rural Urban Continuum index with colors, and the number of cases with the scatter plot point size.

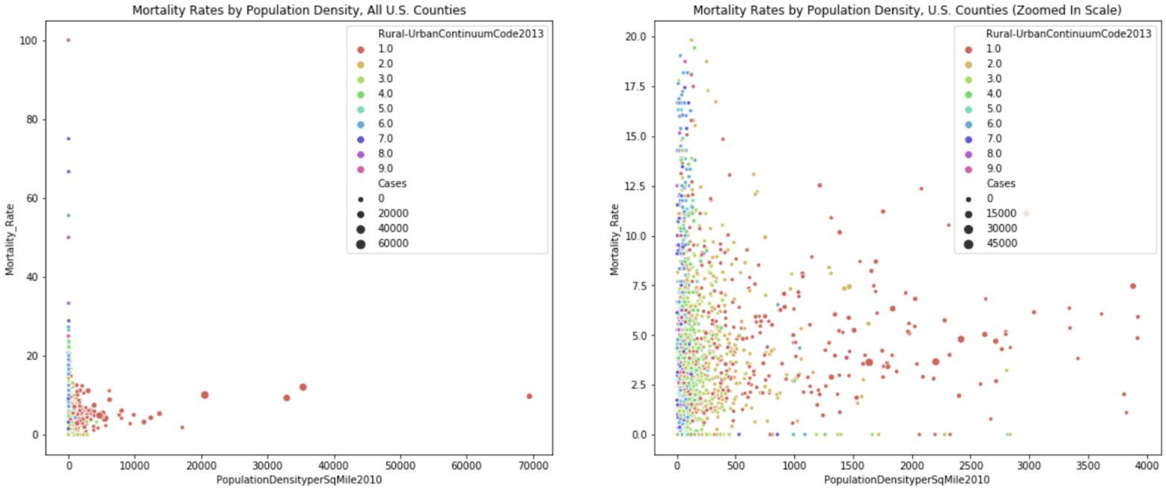


Image: Scatterplot of Population Density by Mortality Rates (All & Zoomed In)

From these scatterplots, we can see two things. First, greater population density is generally positively correlated with a greater number of cases, shown by the growing sizes of the points as we move along the x-axis. Considering the correlation between population density and mortality rates is more complicated. We can see that population density is a poor indicator of mortality rate in more rural areas (shown in blues, purples, and pinks), but a better indicator of the mortality rate in more urban counties (shown in red).

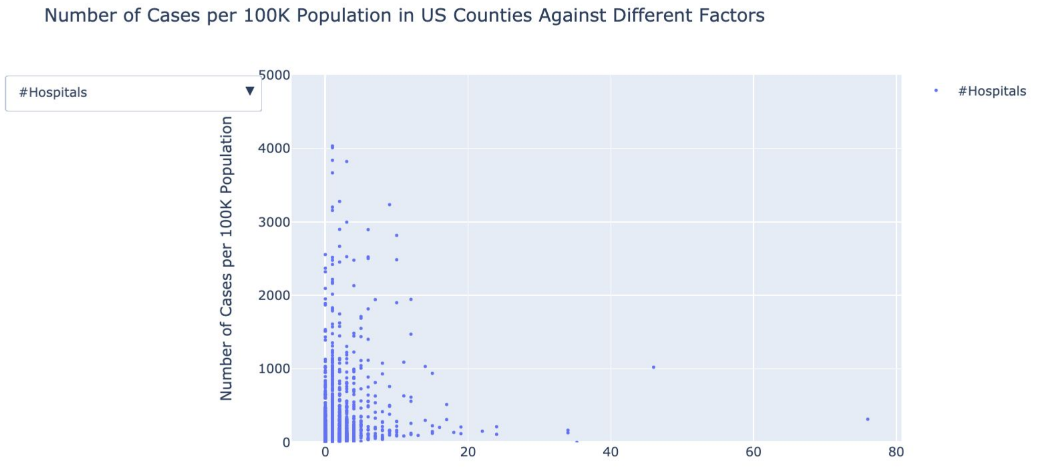
We initially intended on using these features to develop a model to predict mortality rates, but it was unsuccessful due to a large number of counties having 0% or very highly skewed mortality rates. This model can be found in the Appendix notebook.

IV. Predicting Cases per 100K People

Question: In this section of the analysis, we wanted to explore whether certain demographic and health risk factors such as population, rural/urban location, number of hospitals in an area, and percentage of smokers in a population were correlated with the number of coronavirus cases in US counties. We also wanted to incorporate this data with time series data on the number of confirmed cases in each county to predict the number of cases in the future for each county.

Data Cleaning: We used the cleaned version of the ‘county_data_abridged.csv’ dataframe from the mortality rates analysis. Then, we selected columns from ‘us-county-health-rankings.csv’ and joined those with the abridged county dataframe. We again dealt with null values by replacing them with the average for the state. We also merged ‘covid_confirmed_usafacts.csv’, which contained comprehensive county time series data.

EDA and Visualizations: We plotted each demographic and health factor against the most recent number of cases for each county (5/10/2020), with a dropdown list by factor. When initially plotting the factors against the raw number of confirmed cases, we found that urban areas tend to have a higher number of cases and the number of hospitals, ICU beds, and population factors seem correlated to the number of cases, while health risk factors do not. We switched our visualization to plotting against cases per 100k for each county, and found that rural urban classification was no longer correlated, which indicates that many urban areas have a high number of cases due to higher populations, but this does not necessarily translate to having a high density of cases. We decided to use the correlated factors to guide feature selection for our model, described in the next section.



Model Methodology: Our model aims to use demographic and health factors along with past data on the number of cases per 100k in an area to predict the number of cases for a specific date. We initially used data from 5, 6, and 7 days ago but found that there was some overlap between the features and the numbers we were trying to predict. Thus, when training our model, we switched to using data from 28, 29, and 30 days ago to predict for the dates 5/1/2020, 5/4/2020, 5/7/2020, and 5/10/2020. When feeding the data into the model, we took out county identifier variables (FIPS code, county name, and state) with the idea that this model could be generalized to any area where information on the features used in the model were available. In processing our data for the model, we found that in some counties, the number of confirmed cases decreased from one day to the next, indicating inconsistent reporting. This would sometimes lead our model to predict a negative number of confirmed cases, so we dropped these counties from the data.

We trained 3 linear regression models. The first was a regular linear regression model using the correlated features from our EDA. Next, we trained a model using lasso regularization with all the features to see which features the model selected. Finally, we trained another regular linear regression model using the features that the lasso model had selected. We also tried normalizing the data before running the models, but found that it had no effect, so we kept the unnormalized data.

After training the first linear regression model, we found that the model tends to underestimate the data. This might be because each county's confirmed case growth rate is different, and in a month, the virus might have spread much faster in some counties as opposed to others. It could also be because the number of cases reported a month ago may not have been as accurate as testing and reporting is today.

When training the lasso model, we found that the model initially did not converge and would stop training at the maximum number of iterations. We tried increasing the maximum number of iterations, but this significantly lengthened training time, especially when performing cross validation to find an optimal alpha value for the model. We decided to keep the default number of iterations in favor of training time due to our limited computational resources. When examining the features selected for by the model, we decided to keep features which had an absolute value coefficient greater than or equal to 0.1. We found that the model selected several of the features we used in the first model, but also included some more health risk factors. We used these features to train the next linear regression model.

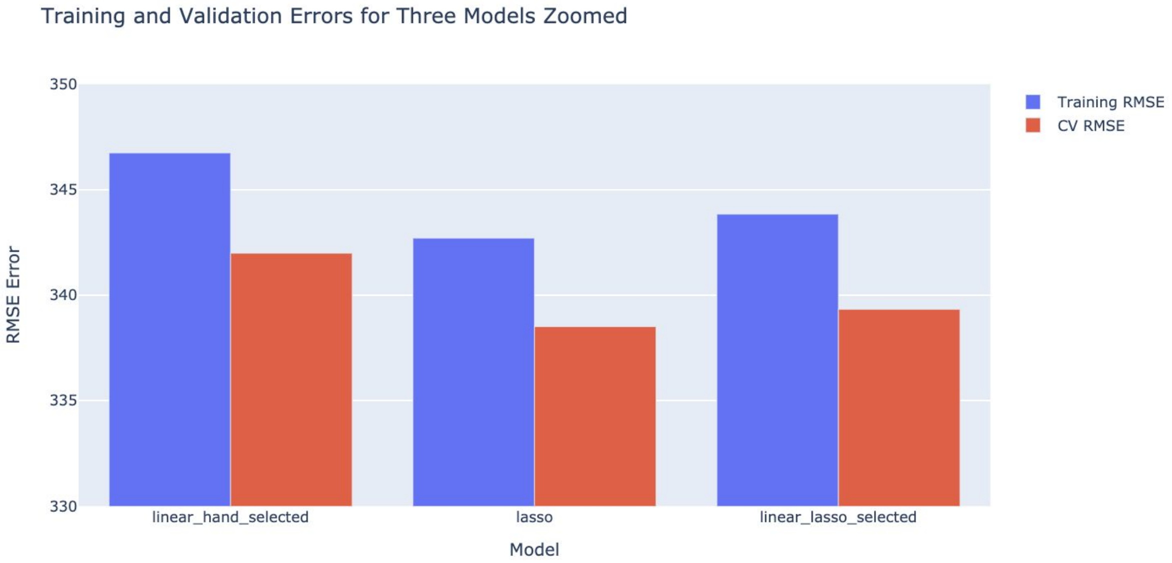


Image: RMSE error for each of the three models

Model Evaluation and Analysis: When evaluating our three models, we looked at both the training error and average error from 5-fold cross validation. Our models tended to have errors in the range of 335 - 350 cases per 100k, which is reasonable given the metric. We found that the lasso model had the least training and cross validation error, so we selected this as our final model and ran it on the test data. The test error was 334.9 cases per 100k, and we found that the model again underestimated the number of cases. Although we trained the model with data from April and May, each county is at a different point in its rate of spread of the infection. Therefore, the data should be varied enough to translate to predicting for June. We ran this model with data from 5/1/2020, 5/2/2020, and 5/3/2020 to predict the number of confirmed cases on June 1. We found that the model did predict negative numbers for some counties, but this was mostly for counties that had seen no cases so far.

V. A Global View: Predicting Growth of Number of Cases by Country

Question: Countries, especially large ones, may have multiple coronavirus hotspots that emerge at different times, which makes it important to analyze both the granular and broader level. In order to visualize and predict cases in countries, we attempted to fit the given time-series data to a logistic curve, which has been known to model virus spread with sufficient accuracy.

EDA and Visualizations: To initially visualize the data, we plotted the time series over time for major world countries. We noticed that curves had a similar shape, and could modelled with the logistic equation. We also saw that mortality rates were lower in the United States than in other major countries like Spain, Italy, and France.

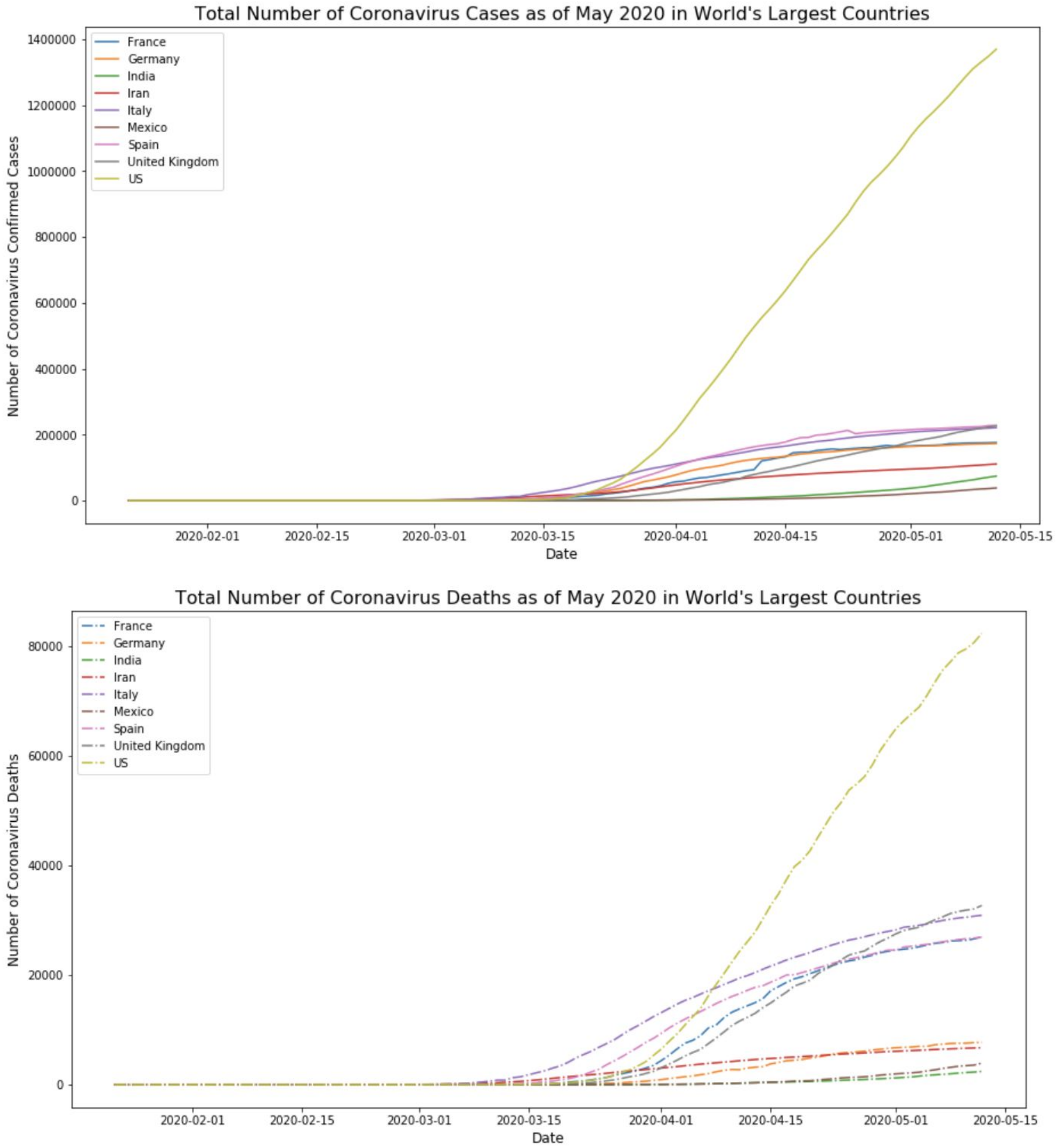


Image: Coronavirus Cases Overtime in Largest World Countries

Data Cleaning and Transformations: From our initial analysis, we observed that while some countries appeared as a single row in the dataset, others were broken down by state or province. To account for this, we grouped the DataFrame by country and used a sum aggregate to get country totals. Otherwise, the global time series data was clean and we could proceed with our analyses and model. In addition, in order to more easily create visualizations for our model, we changed the date columns in the time-series DataFrame to the number of days since 1/22/2020. For example, the 1/27/2020 column name was changed to 5.

Methodology for Fitting our Logistic Equation: After looking through some research and observing the trends in cases over time through our initial visualizations, we decided to use a logistic curve to fit our data. The logistic curve has the equation $f(x) = L/(1 + e^{-k(x-x_0)})$. The parameters each represent important aspects of the curve, which helps us understand how the number of cases is evolving in countries:

- L represents the upper y-asymptote of the curve, and signifies the total number of cases a country will have.
- X_0 represents the midpoint of the logistic curve. This is the number of days since 1/22/2020 that the curve changes concavity and when the number of cases recorded daily begins to decrease. This is considered to be the "peak day".
- K represents the growth rate of the logistic curve.

To fit our data to the logistic curve, we followed the following steps:

1. *Defining the model equation:* We created a function that takes in the logistic equation parameters and returns the output of the function. We need this function to pass into our curve fitting optimization.
2. *Curve Fitting:* We used scipy.optimize.curve_fit to fit our data and obtain the L, x_0, and k parameters for each country's logistic curve. We put this into its own DataFrame for easy analysis, along with the RMSE to show how well the logistic curve fits the data.
3. *Plotting:* We plotted the data as a scatterplot alongside the logistic equation curve for China, Italy, and India, which are at different stages in dealing with coronavirus, to show that our logistic equation accurately models the data. Any country can be plotted by using the dropdown selection.

We found that the logistic equation very accurately fits the time-series data for countries at any stage in coronavirus case count. However, it is important to note that the curve fitting did fail for some countries due to overflow errors. After looking at each of these countries' data, we attributed this to inconsistent reporting or virtually no coronavirus cases so far, meaning that there is not enough significant data to fit a logistic curve. We have printed out these countries in our notebook for reference.

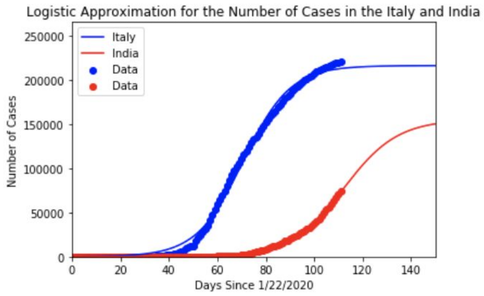


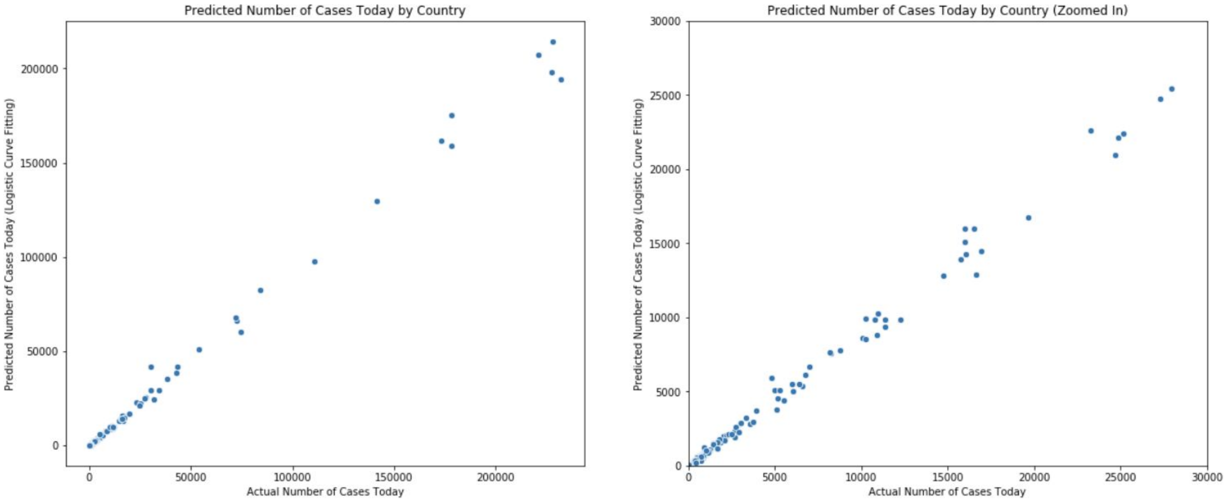
Image: Logistic Curve Fitting for Italy and India

Visualizing Cases Over Time: We can use our logistic equation model to predict cases in the future by simply observing how the curve behaves for dates beyond our given data. In addition, we can use past data to predict today's data in order to see how accurately the predictor works.

Our Model and Assumptions: We simply used the least-squares logistic equation optimal curve in order to predict future cases. We can input any date into the logistic equation for a given country, and the result will be our prediction for coronavirus cases on that day. This allows us to predict any day in advance, although we know that the further in advance, the less accurate our model will be for countries that are still seeing a significant rise in cases. This is because our model has not been allowed to develop sufficiently enough for these countries. Our assumptions are as follows:

- The time-series data alone is sufficient to create relatively accurate predictions for future number of confirmed cases.
- The logistic curve is an accurate predictor for future cases for all countries.

Results and Visualizations: We first decided to use data for up to a week before today in order to predict the number of cases today using our logistic equation model. For each country, we plotted the predicted versus the actual number of cases for today. We observed this plot to be very close to the line y = x, meaning the observed and predicted cases were very similar. This tells us that our model is a very accurate predictor for cases for one week in the future. We have provided both 7 and 28 day future predictions in our results DataFrame for reference and analysis.


Predicted Number of Cases Today by Country

Predicted Number of Cases Today by Country (Zoomed In)

Image: Scatter plot showing actual cases today vs predicted cases today (All & Zoomed In)

VI. Further Questions and Conclusion

(i) What were two or three of the most interesting features you came across for your particular question?

The overcrowding factor for each county ended up being very helpful to our final linear model that predicted cases per 100K in each county. We brought this factor in from a county health rankings dataset. Out of all the factors that our model used, the case counts at prior dates and overcrowding were given the highest weights. In addition, the rural urban continuum code was very interesting to understand in terms of mortality rates in urban vs. rural areas, and there were clear trends in how coronavirus affected rural areas differently.

(ii) Describe one feature you thought would be useful, but turned out to be ineffective.

We initially had thought that countermeasures such as the implementation of shelter in place, restrictions on large groups, and travel bans would have a measurable impact on the spread of coronavirus within counties. However, we found that it was very difficult to identify a trend between when these restrictions were put into place and a change in how the virus spread. We believe that part of this inconclusive result was that countermeasures because counties hadn't started reporting their confirmed cases until after regulations, and many had implemented regulations in response to nearby outbreaks, without having any local cases. We left this analysis within the appendix, but we believe that with better tools to look at how the rate of change was specifically affected, we might be able to conclude more about the effectiveness of countermeasures.

(iii) What challenges did you find with your data? Where did you get stuck?

We found it challenging to fit every country to a logistic curve given the time series data. We kept encountering overflow errors when we called the curve fit optimization method for certain countries, and spent a lot of time trying to use certain subsets, cutting off leading entries with no reported cases, and determining if incorrect reporting was done for these countries. In the end, we found that either these countries had almost no reported data, or were reporting statistics inconsistently. We decided to not model logistic curves for these countries and just list them in the notebook for reference, as the optimizer consistently failed for this small subset of countries.

(iv) What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

One of the major limitations of our global time series analysis and predictions was that we solely used the time-series data in order to predict future cases in countries. We stated one of our assumptions as being that all countries will follow a general logistic curve, which can be proven to be incorrect if countries quickly subdue the virus, or in cases such as the United States where different hotspots emerged at different times. To make more accurate predictions, we would either need more data, or we would need to create another model taking into account other factors, such as number of major cities and the country's response.

(v) What ethical dilemmas did you face with this data?

In this data, ethical dilemmas we faced were knowing that our two models didn't accurately encompass the spread of coronavirus. For example, we created a linear model to predict the number of cases per 100k population, and ended up ignoring outlier counties in our visualization. However, deeper exploration showed that these high case counts per 100k were not a result of misreporting, but rather implications about how the virus spreads. We found that the highest case count per capita occurred in Trousdale, TN due to an outbreak within a prison. Our model wasn't built to consider such localized outbreaks, but to not include and consider such cases would be dismissing the reality of the virus and the way it spreads.

(vi) What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?

When creating our model at the county level, we wanted to add additional information about the age and gender divisions at each county, because we believe that may have had an impact on cases per capita. In addition, information on how many colleges, prisons, airports, and other institutions were located in each county may also be helpful in strengthening our analysis.
We also struggled with counties misreporting or not reporting data until recently. If earlier data and testing had been available, we believe this information could be used to build stronger analyses.

(vii) What ethical concerns might you encounter in studying this problem? How might you address those concerns?

The accuracy of our models might pose an ethical dilemma in the real world. For example, we could use our models' predictions about the number of cases in the future to drive decisions about shelter in place and economic regulations. The accuracy of our models would then play an important factor in the effects of these regulations on people and businesses. We can address these concerns by consulting outside experts to determine what other factors might be important in our analysis and to gain an external informed perspective on the results of our model.

Opportunities for Further Research

After completing the analysis, we would like to further research different levels of regulation, looking into the effectiveness of methods such as closing schools, colleges, public spaces, and workplaces partially and entirely. We began this analysis on the county level, but weren't able to make solid conclusions partially due to incomplete reporting, and left this analysis in the appendix notebook. Such insights would be useful in evaluating the effectiveness of these regulatory methods, both at the granular county level, and the state and country levels.

On a global and on a country level, this information would be extremely helpful to lawmakers and the general public to understand which practices are best at "flattening the curve," and how other uncontrollable factors may put certain populations at risk.