Modelling COVID-19 Deaths in the US

by Grégoire Van Thienen & Oscar Chaix

Abstract

In this project, we explore the COVID-19 datasets to better understand the factors behind COVID-19 mortality and to develop an algorithm that can predict the number of county deaths from the disease according to a defined set of features. We find that population density is highly correlated with the number of deaths in a certain county, which supports the federal guideline of social distancing. Our multi-linear regression model further predicts county deaths with a decent level of accuracy, generalizing well to unseen data and scalable for modeling applications.

Introduction

The unprecedented COVID-19 pandemic has upended our lives and created havoc from a health, economic and social perspective throughout the world. A wide array of professions has hence mobilized to both combat the virus and find a way out of the crisis. Notably, data scientists have stepped up and have played an essential role in government and public health decision-making. Researchers at Imperial College for example came up with advanced modelling of the disease' potential developments, warning of the risk of millions of deaths if nothing was done and suggesting that the virus might surface repeatedly in 'waves'. The White House and the CDC have based a number of their policies, and public announcements, on such models predicting number of cases and deaths in the US.

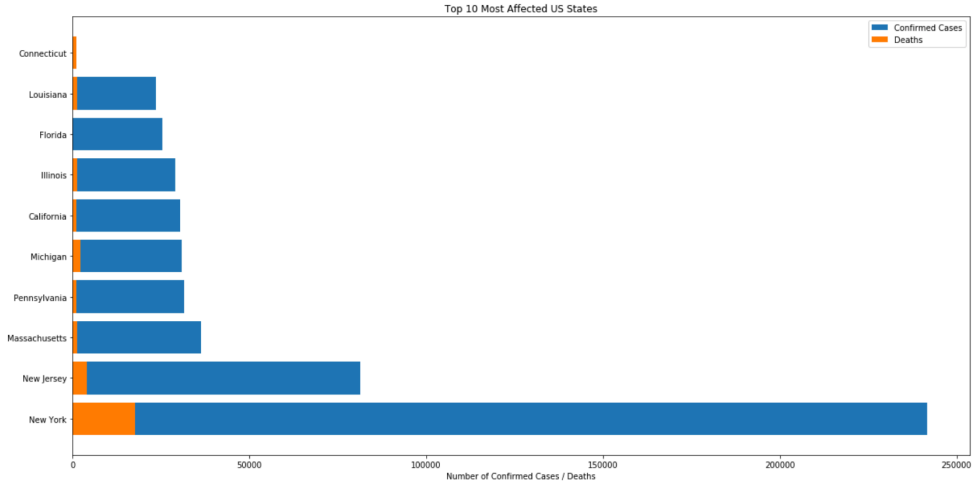
In this project, we seek to both understand how to predict the number of COVID-19-related deaths through such models, and to better understand the factors behind the virus' mortality. By exploring three datasets: '4.18states.csv', 'abridged_couties.csv' and 'time_series_covid19_deaths_US.csv', we hope to have a clearer picture of these predictive algorithms' accuracy and to have a better appreciation of the reasons why some states and counties have much more COVID-19 mortality than others.

Exploratory Data Analysis

We first load the four datasets and try to better understand them by looking at their columns and first five rows. All tables provide interesting information. The state_data table ('4.18states.csv') provides essential metrics around COVID-19, such as the number of confirmed cases, the amount of deaths, and the testing rate for specific regions. The county_data table ('abridged_couties.csv') only provides information on US counties, but gives a lot of useful details on each counties that we can later use to create our prediction algorithm. For example, features around demographics (age and gender) and around the prevalence of certain serious diseases could prove explanatory and significant in predicting the number of cases and deaths from COVID-19. Finally, the last two tables ('time_series_covid19_confirmed_US.csv' and 'time_series_covid19_deaths_US.csv') provide

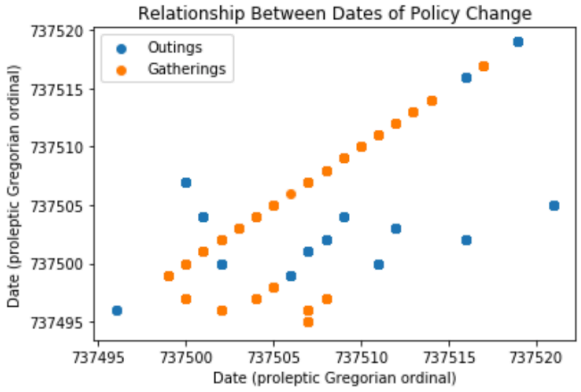
useful information with regards to the development of the disease in each state. Incorporating early and late data could also improve our algorithm's accuracy.

Let's explore which states are the most affected by the pandemic, both in terms of confirmed cases and in terms of deaths. To do so, we filter the state_data table to contain only US states and relevant columns. We then sort these states by descending count of confirmed cases and deaths, and take the top 10. To visualize their relative importance, we finally plot a horizontal bar chart that displays both metrics:



It's important to note here that the data has only been updated on April 18th at latest, and hence these metrics are today much outdated. The number of both confirmed cases and deaths has sadly much increased since April 18. Nevertheless, we can see first hand with this bar-chart that some states are disproportionately more affected than others. The state of New York is the most affected, with nearly a quarter-million confirmed cases and a high number of deaths. However, states like Connecticut and those not shown on the plot have much smaller number of confirmed cases and deaths. Why were some states like New York and New Jersey disproportionately affected by the crisis? Hopefully, this project should shed some light on this question, by highlighting factors that contribute to high number of confirmed cases and deaths (such as previous conditions, demographics, policy changes, etc.).

We can then explore in more details the county_data table and its features. The list of its features is extensive (see code for detailed list). Going through the dataset's documentation, we learn that features such as '>50 gatherings' and 'restaurant dine-in' correspond to dates, given as a proleptic Gregorian ordinal. We can convert them to normal dates by using the date.fromordinal() function: for example, executing the code 'date.fromordinal(737503)' outputs the date 2020-03-19. Although we will keep these column features in their proleptic Gregorian ordinal format for computational and plotting purposes, it is important to keep in mind that they correspond to dates. Plotting the correlation between some of these features provides interesting insights. We can for example plot the correlation between the dates when dining out was banned and the dates when entertainment and gyms were banned; we can also plot the correlation between the dates when gatherings of more than 50 were banned and the dates when gatherings of more than 500 were banned. These correlations are plotted in purple and orange, respectively:



We can observe a clear correlation between outings bans and between gatherings bans. Intuitively, this makes sense since states will usually close restaurants the same time they will close down gyms; similarly, there should be a strong relationship between when states ban gatherings of more than 500 people and when they ban gatherings of more than 50 people. Although there is some over-plotting in the above chart, which explains why a minority of counties are displayed (points overlap on top of each other), this visualization provides a clear picture of how policy changes strongly follow each other in time. There should therefore be significant collinearity from these variables, and using only one or an aggregate measure of them could prove sufficient for our algorithm.

Data Cleaning

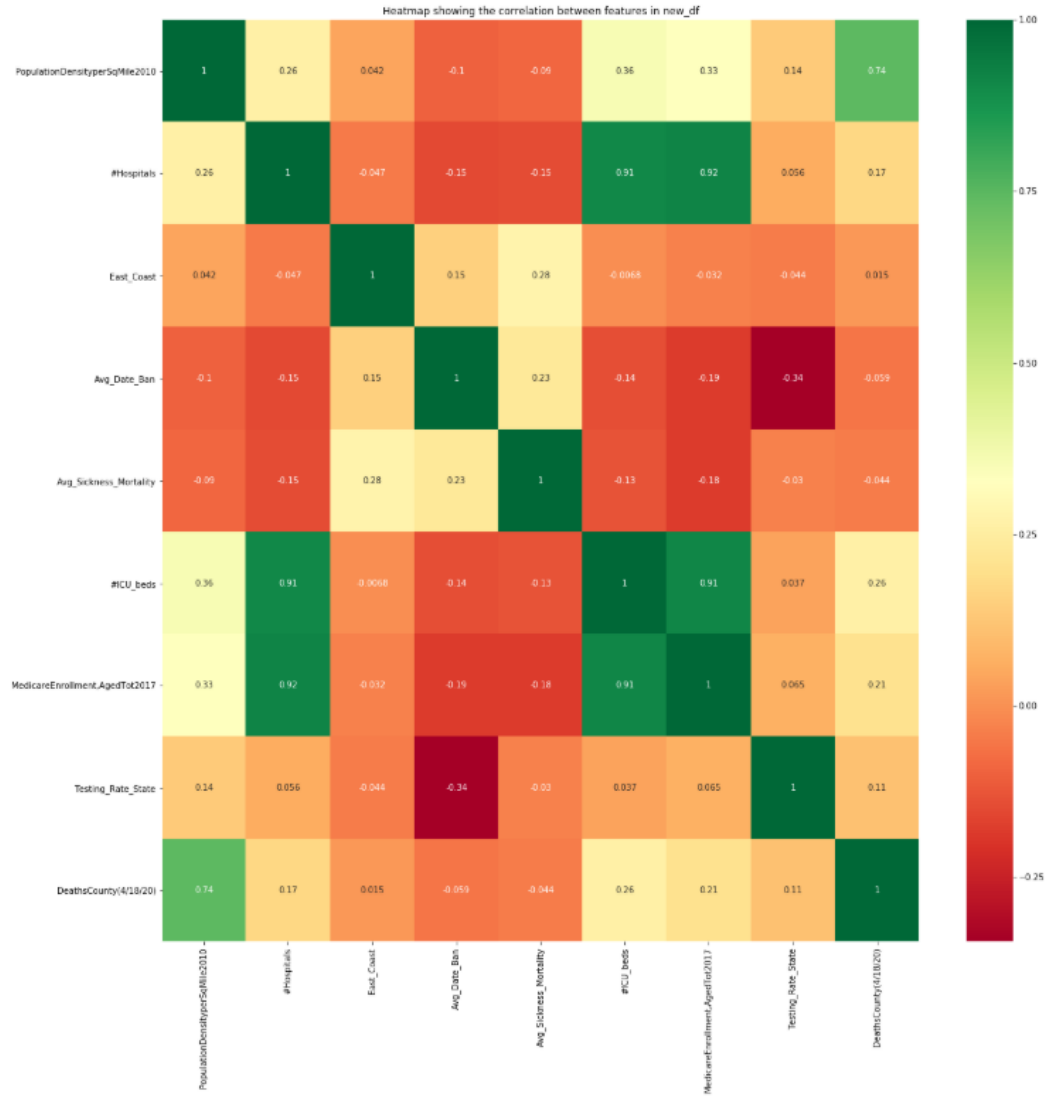
We clean all datasets to only keep relevant columns and limit the amount of missing values. For example, we only keep rows corresponding to US states in state_cleaned because these rows seem to have a very small number of NAN values and we are interested in US states. After filtering out non-US states/locations, we see that the number of NAN values drops significantly. Overall, the reported data across US states is consistent. For example, the last update dates back 2020-04-18 for all rows, which is convenient if we want to go on to compare cases or death rates across states at the same point in time. Concerning the county_data table, we decided to filter out some columns and essentially keep those we believe could help us predict the number of cases or deaths in the US. Public schools, restaurants dine-in, federal guidelines, and foreign travel ban have relatively few NAN values, so they could potentially be good predictors. We create a list ('list_of_columns') which contains some other selected features, as well as some unique identifiers like STATEFP for instance, which we will need to merge `state_cleaned` and `county_cleaned` into one dataframe. Therefore, the rows where STATEFP had a NAN value were removed from our dataframe. Doing so also helps us get rid of NAN values in some other columns as well. We finally merge state_cleaned and county_cleaned into one main dataframe. We use an inner merge so that only common values between the left and right dataframes are retained. Hence, `county_state_merged` has 3142 rows, whereas `county_cleaned` has 3221.This particular kind of merge helps reduce the number of NAN values in our final output.

Model Choice, Methods & Assumptions

We decide to create a multi-linear regression model that predicts the number of deaths by county according to a specific set of features. We initially wanted to predict the number of deaths at the state level, but soon realized that this would provide us too little data points to work with: having only 50 states limits the number of rows to split between training, validation and test sets, as well as limits the granularity of our model and the level of details in the features we hope to include in our model. We decided to use a multi-linear regression because most variables of interest are quantitative, and we are trying to predict a quantitative variable: number of deaths in a county. We normalize all variables so as to have all features on the same scale. We then select features for our feature matrix, split between training, validation and test set, fit our model on the training data and measure the prediction root-mean-squared-error for our training and validation set. We hold out our test error for later, when we will want to evaluate our model's ability to generalize to unseen data. We test our model accuracy only on the training and validation sets, as we don't want to use our test error to make modifications in our model.

Until we evaluate our test error, we seek to make our model more sophisticated through better feature selection and through feature engineering. For example, we create a binary variable that indicates if the county is in the West or the East of the US. We saw that Eastern states were somewhat more affected by the virus, so having this binary variable could improve our model as well as simplify it (we won't need to add specific longitude and latitude measurements). To do this, we assign 0 to states to the west of the 100th Meridian West and 1 to states to the east of that same longitude, at -100. We can also create averages for sickness prevalence and for policy changes and add them to our dataframe. These three new features allow us to simplify our model by removing redundant variables like longitude & latitude, or policy bans, replacing them with these aggregate measures.

We finally make a heatmap to show correlations between our response variable, 'DeathsCounty(4/18/20)', and some of the features in our data table. It helps us determine which features to keep, and which to discard from our features matrix. It's interesting to note that population density is especially correlated with CountyDeath, with a correlation of 0.74. Intuitively, the most dense places should have more contagion and more deaths from COVID-19. There is also a high correlation between testing rate in a state and deaths in the county, although the direction of causation is unclear. This is probably because the most affected states will roll out more tests. This being said, these are definitely variables we could keep in our updated feature matrix. Surprisingly, our engineered variable Avg_Date_Ban does not explain too well the variation in DeathsCounty. That’s probably because most counties enacted these bans around the same time, as they aligned with state or/and federal regulations. This is an observation that we made during our EDA. We will nevertheless keep this engineered feature, as well as East_Coast and Avg_Sickness_Mortality in our updated feature matrix because they allow us to simplify our model by removing redundant variables (we can replace longitude and latitude, some disease mortalities, and date bans by these features). Lastly, we can also check for multicollinearity with this heatmap visaulization. It's not surprising to observe that #ICU_beds is highly correlated with #Hospitals. This indicates that it’s preferable not include both variables in our updated features matrix.



Summary of results

Here are the features that were included in our second and final model, along with their respective coefficients.

```
X_features_updated.columns
```

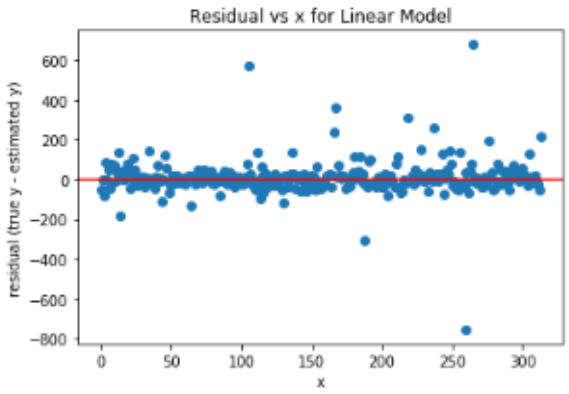
```
Index(['MedianAge2010', 'FracMale2017', 'PopulationDensityperSqMile2010',
       'MedicareEnrollment,AgedTot2017', '#ICU_beds', 'SVIPercentile',
       'DiabetesPercentage', 'Smokers_Percentage', 'dem_to_rep_ratio',
       'Confirmed_State', 'Deaths_State', 'Incident_Rate_State',
       'Mortality_Rate_State', 'Testing_Rate_State',
       'Hospitalization_Rate_State', 'Population', 'East_Coast',
       'Avg_Sickness_Mortality', 'Avg_Date_Ban'],
      dtype='object')

array([  26.20576818,   -6.89069883,  110.18968214, -454.59031461,
          8.95557096,    3.62351513,   -4.96651394,   11.01397064,
        -26.02078705,   60.11494523,  -79.93119154,   29.39895936,
          6.55908747,    4.48309284,    1.81519794,  482.24906454,
         -4.50744845,   -8.0253201 ,    2.19525294])
```

- We observed that the number of cases and deaths from covid-19 could greatly vary depending on the state/county under consideration. As of mid-April, Connecticut had almost no deaths whereas New York had already recorded well over 1000 deaths.
- These kinds of spatial discrepancies are rather surprising, especially since most states seem to have imposed restrictions on outings and gatherings around a similar period.
- The actual effects of these bans on deaths are difficult to measure and interpret at this stage. Perhaps, social distancing measures are still too recent to pick up any significant differences in deaths across states and/or counties.
- Nevertheless, our analysis at the county level shows that deaths increase with population density and total population. These results suggest that social distancing could help reduce the number of deaths, and particularly so in densely populated areas which supposedly have higher risks of transmission.
- Moreover, we found a negative coefficient for medicare enrollment in the county in 2017, indicating that higher medicare enrollment translates to fewer deaths on average.

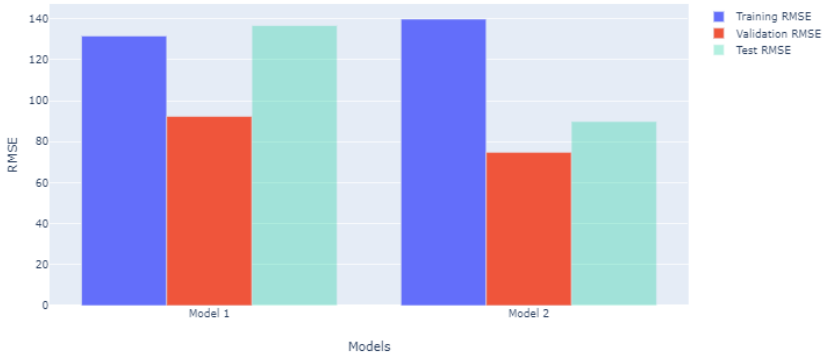
Method evaluation, limitations, and future work

We plotted the residual to visualize the error of our updated model:



The residuals are clustered around the line at 0 and we are not seeing any kind of patterns, indicating that our linear regression model fits the data quite well. The residual for most of our observations is found in the 0 +- 200 range, except for a few points located further away from the horizontal red line, for which we are overestimating/underestimating Y.

We also created a bar plot to compare the training, validation, and test RMSE across our two linear models. 'Model 1' corresponds to our first model, and 'Model 2' to our final model with the updated features matrix (see next page). The changes in RMSE from Model 1 to Model 2 are quite significant. We observe a slight increase in the training RMSE, as well as a decrease in the validation RMSE. More importantly, we see an important drop in the test RMSE for model 2. These results are rather encouraging, as our second model seems to better generalize to unseen data.



Furthermore, we must stress that our approach has room for improvement, as we are ultimately only testing two different combinations of features and comparing the errors associated with each resulting model. The original feature selection process was somewhat arbitrary as it was shaped by both our expectations of each variable's explanatory power and insights gathered from our EDA. However, that gave us a basic model to work with, which we were then able to fine tune by adding new features and discarding existing ones. We had a go at implementing cross-validation to get a sense of what the optimal number of features to include would be, but in all honesty our results were somewhat ambiguous (we found that the optimal number features to reduce the cross-validation error was equal to 1). Therefore, we chose not to include this code and opted for a simpler (but perhaps more tedious) method. It's certainly not perfect, but it fulfills its intended purpose in that it allowed us to identify which factors can be used to better predict covid-19 related deaths at the county level.

We also faced some important challenges with the data itself. At first, we struggled to define the exact scope of our analysis, as we were not sure if we were going to focus solely on U.S. states, counties, or both. It seemed like the provided datasets generally had more information on counties than states, hence our decision to predict county deaths. That said, we feel that the county data was also lacking useful metrics like testing rates, hospitalizations, etc. which would probably have helped strengthen our predictions. We had to settle for the state-wide numbers, although we are not entirely certain of how relevant they are to predict deaths within specific counties.

We are also aware that the study of covid-19 may give rise to several ethical concerns. The fact that this health crisis is still ongoing calls for extreme caution, as people are still being affected by this virus, and we do not want to rush to conclusions at such an early stage. Covid-19 is still relatively new, meaning we don't have enough hindsight or data to fully understand or predict its impact in the future. The world is now being flooded with data on covid-19, and it is our responsibility to verify that our sources of information are up to date and reliable. Finally, we would like to stress that the results expressed in this report are not reflective of our own personal views and may not constitute an accurate representation of reality. With all the aforementioned considerations in mind, we tried to make this project as self-contained as possible and rely on the datasets made available to us to think about a certain set of questions.