Data 100: Final Project — COVID-19

Authors: *Mitchell Cazares, Jeremy Martinez, Eduardo Rivera*

05-10-2020

Abstract

COVID-19 research is at the forefront of most Data Science, today, as people strive to tackle all and any questions related to the global pandemic. A lot of the techniques for current research build on fundamental tools introduced over the past semester in Data 100, and in this document the three of us aimed to tackle just a subset of Corona related questions. We began by examining simple trends within the U.S. and began to break it down by region. Without surprise we began to see clearer trends and patterns as we narrowed down the data to regions with more cohesive structure. This report dives into aspects of general Machine Learning algorithms to predict and model the data provided on COVID-19.

1 Introduction

Modeling data is a huge task filled with great subtleties and deep complexities. In some of the Data 100 assignments we were tasked to model data that ranged from predicting house prices to classifying emails as spam/ham. In this document we explain exactly how we dove into the task of predicting the confirmed cases and death cases for COVID-19, mostly focusing our attention to the US. It is important to note that we decided to use the data available as of May 10th, 2020. This report is concentrated on COVID-19 and tries to answer the confirmed cases per country for the next 30 days following this date and the cumulative deaths per US county that occurred on May 10th using two different linear models. It also provides a comparative study as to how different states are being impacted by the virus.

2 Data Description

There are 7 different data sets we accessed namely 'us_confirmed_cases', 'us_confirmed_deaths', 'us_states', 'medical_info', 'global_recoveries', 'global_death', 'global_confirmed'.
The first two data sets serve as time series for confirmed cases and death cases, respectively, related to the infection. 'us_states' provides a detailed summary for each US province/state for the specified date (05/09/20). Similarly, 'medical_info'

provides appropriate information for respective regions. Lastly, we have the information for recoveries, deaths, and confirmed cases on a global scale, also in the simple format of a time series.

2.1 Challenges With Data

One of the problems we experienced was trying to merge all the data frames together. There wasn't an equal number of rows for all the data frames since some of our data was given to us by US counties, US states, and by countries. To get around this we needed to create a new data frame with the relevant information like the number of confirmed cases and death cases. For any missing information we just simply replaced any NaN with 0 so that we could keep our data consistent. Another challenge we experienced was trying to find a way to manipulate the month columns in the following data frames:

- us_confirm_date, us_death_date, world_confirm_date, world_death_date,world_recovery_date

These data frames were crucial to implementing our time series plots as we needed to get the months in order to keep our visuals consistent. The way we got around this was to we simply added the codes from the sorted-months-weekdays and sort-dataframeby-monthorweek packages to the notebook instead as we were having issues installing them in our notebooks.

3 Description of Methods

To be able to make sense of the data, we began to narrow down the features we deemed as relevant to our investigation.

3.1 Data Pre-Processing

Processing of the data played a vital role in understanding the data as well as making the model more robust. Quantitative data may have erroneous characters or values that don't make sense and should be removed from the data for robustness. For example some of the entries in the data frames were 'NaN'. Values such as these can be interpreted in a variety of ways, but from our understanding we treated the 'NaN' values as 0.0 because there just wasn't enough information for some regions. This, however, did not change our investigation because we began to focus on the regions that were heavily impacted, and so 'NaN' values weren't an issue for those regions.

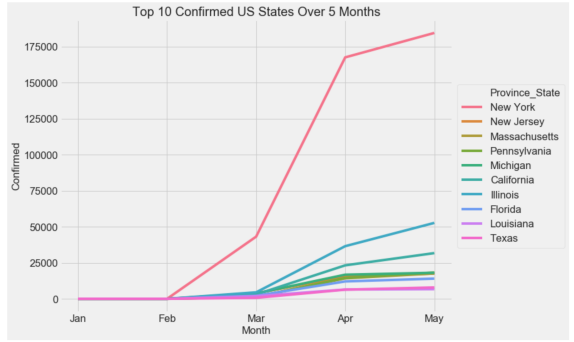
3.2 Gathering Data

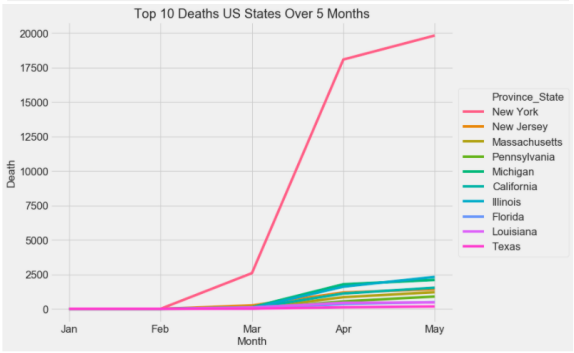
We decided to look at the US, and grouped by the 'Province_State' column. Working accordingly with the other tables, we gathered the useful information into our 'US_Data' data frame. We found this necessary, because we wanted to separate by groups with structure. We understand that the U.S. may be experience some

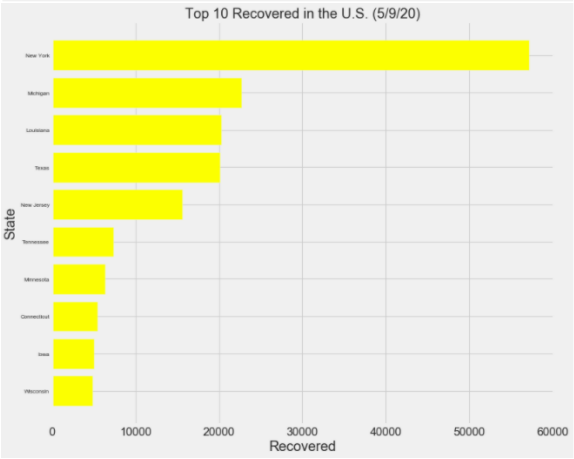
common trends on a higher level, but we were more interested in the intricacies surrounding smaller regions.

3.3 US States

Just to quickly display some of these patterns, we saw how the virus has affected some states in the US. by the confirmed cases, death cases, and recovery cases.

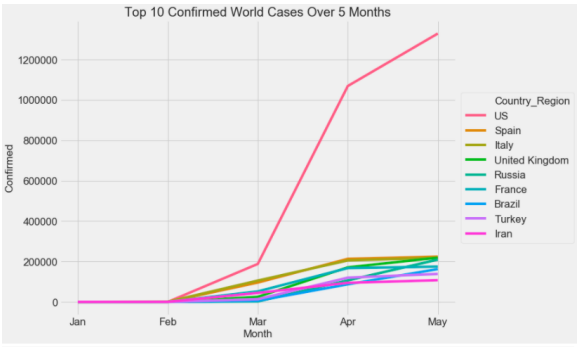


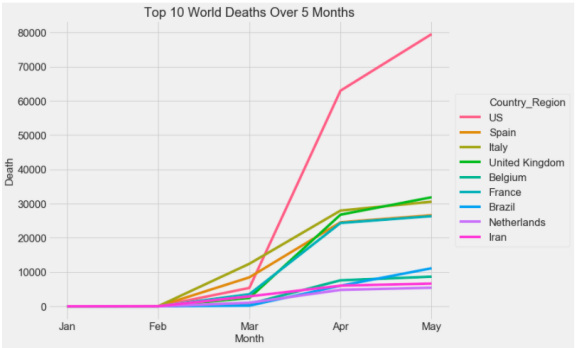


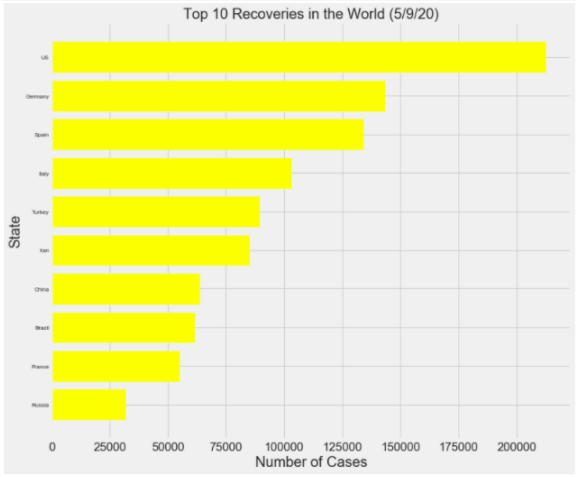


3.4 Global

We decided to extend this to the top 10 nations that are being affected by the confirmed cases, death cases and recovery cases.







4 Modeling

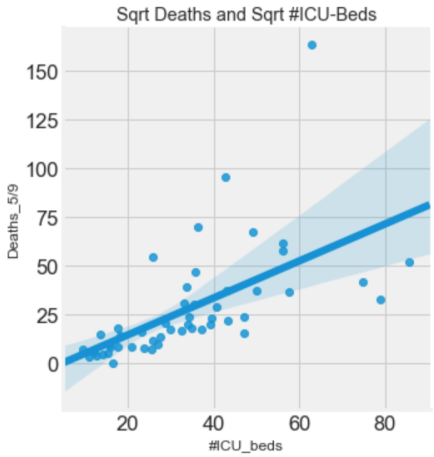
4.1 Linear Regression Predicting Deaths

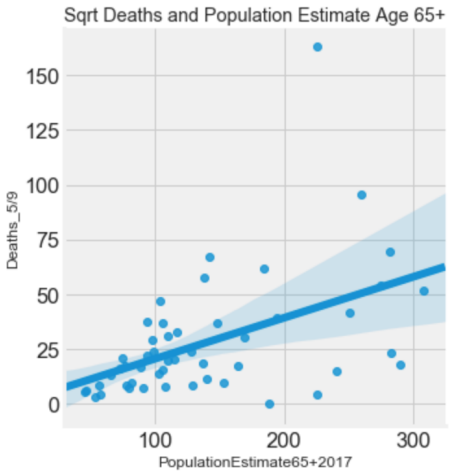
Using the 'US_Data' data frame, we created a linear regression model with L2 regularization. At this point we had the data grouped by US counties. We tried a simple model to predict the cumulative number of deaths on May 10th using the data up to May 9th. We split the data into training and testing sets and trained our model with different regularization values to determine the best cross validation root mean square error. Our accuracy on our testing set was about 62 which tells us

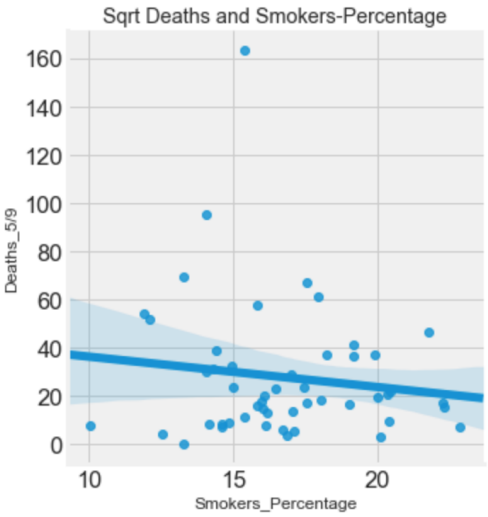
our model is not to far off from the actual cumulative death that occurred on May 10.

4.2 Features

We examined several relationships and visualized some of those. We were pretty surprised by some of our results. They are best understood through visual representations. Visualizing our data helped us understand the correlation between different features and helped us select features that were best for predicting the cumulative death per US county up to May 10th.
(i) Two features that were interesting to look at were the number of icu-beds and the population estimate of ages 65+. Like the visuals show, generally the more icu-beds a county has the more deaths occur due to COVID-19 and generally the more elderly there are in a county the more death occur due to COVID-19. We must take measures to protect the elderly.
(ii) One feature that we thought would be useful to predict the deaths was smoker percentage per county but as the visual shows, there is not to much correlation between the two.
(iii) Some challenges that we experienced while modeling the cumulative deaths per US county was that some information like the mortality rate and incident rate was only available per US states. So our data frame had those specific values set to their state values. For example , all California counties had the same mortality rate and incident rate which pertained to California.
(iv) Thus, we were limited to the amount of data that was available and our assumptions of replacing our NaNs with 0 could be affecting our results for this model but we thought this would be the most reasonable value since there were not too many NaNs.
(v) Some ethical dilemmas that we faced with this data was that some of the features we chose come from data about medical information and patients with previous medical conditions are generally more likely to not survive COVID-19 if infected.
(vi) If we were able to get all the data and most recent data per county and have no NaNs, our model would perform better. Also, we could have checked which communities are more affected if given the data and compare the death rates between communities.
(vii) Some ethical concerns doing this though is that data could be bias. Also, some people would like to have their medical information private.


Sqrt Deaths and Sqrt #ICU-Beds
Deaths_5/9
150
125
100
75
50
25
0
20
40
60
80
#ICU_beds


Sqrt Deaths and Population Estimate Age 65+
Deaths_5/9
150
125
100
75
50
25
0
100
200
300
PopulationEstimate65+2017


Sqrt Deaths and Smokers-Percentage
Deaths_5/9
160
140
120
100
80
60
40
20
0
10
15
20
Smokers_Percentage

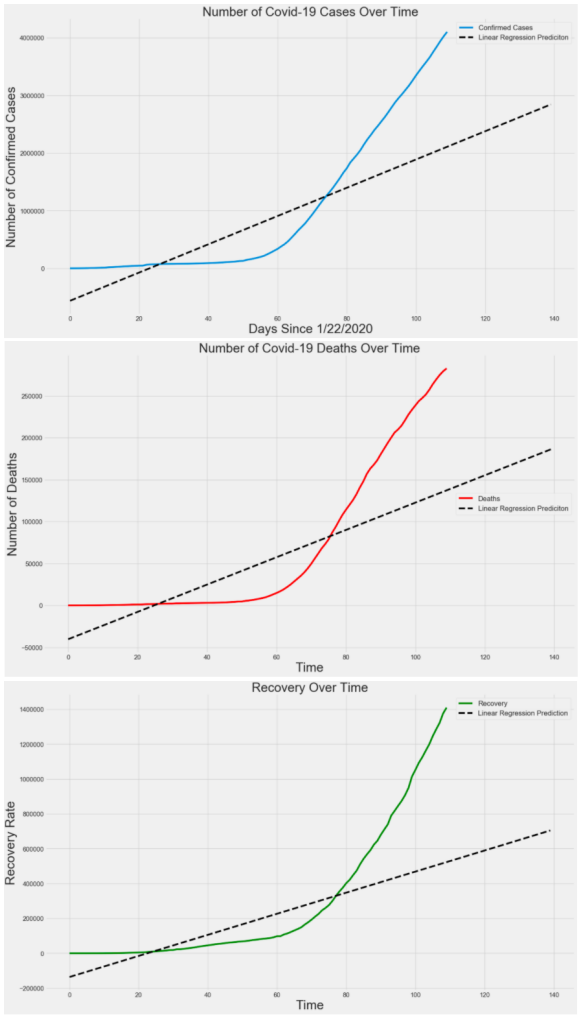
4.3 specific questions to answer about 30 Day Prediction

(i) From the data we had, the only features that were useful was the the country/region and the column dates that contained the information for each date by country.
(ii) We thought the latitude and longitude would be useful but weren't used in the final results
(iii) For this part of our model, we didn't run into any challenges with creating the model, as the data we needed was readily available for us.
(iv) One of the limitations that we faced was that some of the data frames that we had didn't contain all the necessary information that was needed to do a full analysis on COVID-19. For example, we found that there were a lot of NaNs in certain data frames that included the hospitalization rate or mortality rate of other countries just to name a couple. This limited information made it much more difficult to explore additional things we wanted to do within our analysis. The assumptions that we made for this was that the data wasn't available or that it wasn't reported at all. In the end, we googled to see if any of the information was available and found out that it was reported. The issue with the information was that it wasn't being reported correctly.
(v) For the data that was used to create the linear regression model in predicting the the number of confirm, death and recovery cases, we didn't run into any ethical dilemmas as the data was discreate.
(vi) Within creating our 30 day prediction model, we felt that if we had the information for every country such as the testing rate and the number of people that have been tested, we could have done a comparison of the USA and other countries to see how much of a change there would have been with this specific information over a 30 day period from the most up to date information that we have.
(vii) It doesn't seem that we wouldn't encounter any problems with this data, as the data we had for this model was relatively straight forward

4.4 30 Day Prediction: Global

After changing the format of some of the data, such as the date/time, we began to gather all the data together, including the confirmed cases, deaths, and recovery. We utilized this information to predict the next 30 days, following May 10th, and we were able to provide a linear model for this.

We noticed that in some regions of the plots, there may be some under fitting, but we definitely believe that within a few weeks this should not be as much of an issue, because it should only be reasonable for the number of cases to begin to plateau. Currently the data seems to be following exponential growth, but this should begin to level out gradually.



5 Summary

Based on our results, As for the linear model that predicts the cumulative deaths per US county for May 10th, we observed that the number of deaths are increasing at an alarming rate. Our visualisations show that there are certain groups of people who are most vulnerable to COVID-19 and we must do something to protect them. We can see that in our 30 day prediction that there will be a reported 2,843,315 more total confirmed cases, 186,163 more total deaths, and 704,410 more total recoveries on June 9th, which is a huge leap from the May 10th total confirmed cases of 4024009, total death cases of 279,311, and total recoveries of 1,344,278. This is very alarming as the virus will have spread to over another 2.8 million people and the USA is currently the leader of the most cases confirmed cases of COVID-19, which accounts for approximate a little over 25%. But not all is bad! We can also see the survival rate of the virus has also increased and the number of

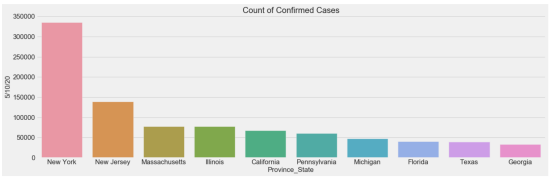
deaths isn't as prominent as well! This is clearly evident in our linear regression plots that was provided in section 4.3 and section 6.3.

6 Discussion

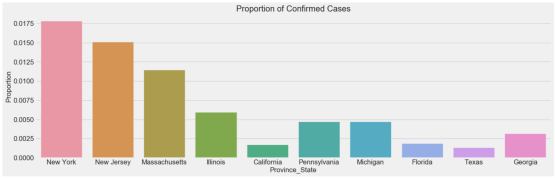
6.1 Additional Observations

We have a lot of other observations that we gathered from the data and some of those are actually surprising. For instance, we looked at that states in the US that had the most cases of COVID-19. We specifically looked at the top 10, and fortunately that data matched up with the public media. We see that New York has taken the biggest hit of course, followed by New Jersey and Massachusetts. California even made it onto the list at 5, but when we readjusted the data, the numbers weren't quite the same.

Below we have a barplot of the number of cases on 5/10/20 by province/state.



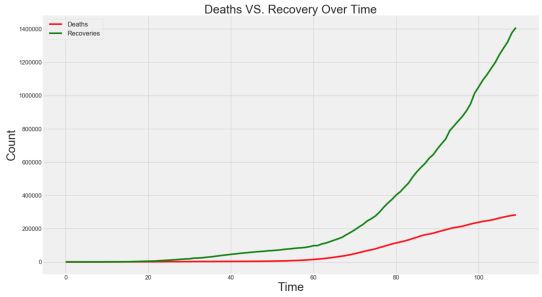
And below we have the proportion of cases relative to the total population by province/state.



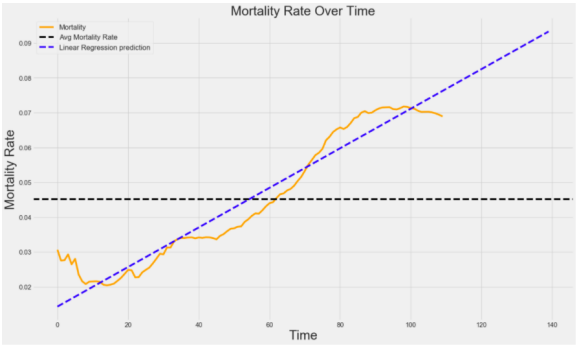
We noticed that when we look at the proportions, states like California aren't as affected relatively by population of states. This must of course be because California is a lot bigger than most states. We bring this up because it has huge implications of state policies. Often, it's easy to blame the state for how this pandemic has panned out, but we see that relatively California is handling better than other states. We understand that there can be other ways to interpret this observation, and we are open to discussion!

6.2 Death vs Recovery

We want to end this on a positive note. Right now it seems that the world is being destroyed by this pandemic. Needless to say, we are certainly in an unprecedented time, but signs are promising. For instance, we noticed in our data that the rate of recovery has surpassed the rate of deaths.



To continue with this we can also see the mortality rate has also increased since the virus was introduced!



But we highly urge all people to combat this virus together. Stay home and wear your masks!