Understanding COVID-19
Last Updated: May 14, 2020

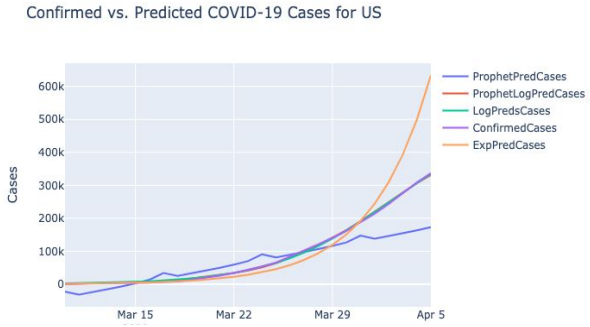
I. Part A – COVID - 19 Global Forecasting

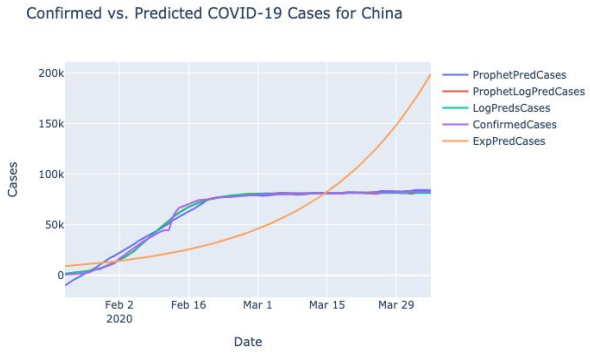
Preprocessing

The goal of the *COVID - 19 Global Forecasting* Kaggle challenge was to forecast the number of daily cases and deaths from COVID-19. Preprocessing the data first involved replacing any null states and regions with "NA" and converting all dates to timestamp values. We chose a cut off date of 03/31/20 to predict the last five days (04/01/20 – 04/05/20). Each data frame was thus split into a test and training set using Pandas according to this cut off date. We chose to forecast cases and deaths by country as opposed to by date in order to address the diverse approaches adopted by each country in response to COVID-19. As a result, we grouped the data by country and date and then summed ConfirmedCases. If a country did not have any reported cases for a given date, we removed that date and its null value. We also created a new column for the number of days since the first confirmed case in every country. In order to forecast total world cases, we grouped the data by date and summed ConfirmedCases over all countries.

Model

We fit multiple models, because we realized that some countries were rapidly increasing in cases (Ex: U.S., Italy), while some were slowing down (Ex: China). Thus, we chose the curve that fit best for each given country. We created exponential and logistic curves on the training data and using sklearn and numpy functions and also experimented with some time series models (Ex: ARIMA, Prophet). We found a logistic growth curve and Prophet's logistic time series model to give us the lowest RMSE (1.1185e6 and 1.1186e6) and MAPE's (38.8 and 30.9) in the U.S. (and China's) predictions, which was significantly lower in comparison to the other models. In order to train the Prophet logistic model, we computed a "maximum capacity" as suggested by Andrej Baranovskij's blog post, by finding the largest rate of change in the training data. If the max rate of change was after April 1st (the training cut off date) that meant that the country was still following an exponential growth model, and the curve was yet to flatten. Otherwise, we expected that the country had hit its "maximum capacity" and the curve was soon to flatten. This helped improve our Prophet model significantly.





Results and Conclusion

The nature of the data is inherently challenging to work with; every country has approached COVID-19 management in fairly unique manners, and even with consistencies among policies, there are unique political situations and limitations per country that make this data less reliable. For example, Singapore had extensively comprehensive testing in contrast to the United States, where testing has been incredibly sparse. As a result, case numbers are likely a far more reliable reflection of reality in Singapore as opposed to the United States, where many more people were likely exposed than the numbers reflect. To try and mitigate some of the variations, we chose to forecast cases and deaths by country instead of by certain time points, as each country had unique time trajectories as to when the first case was discovered, when public health decisions were made, and how the public has responded accordingly.

II. Part B – UNCOVER COVID-19 Challenge

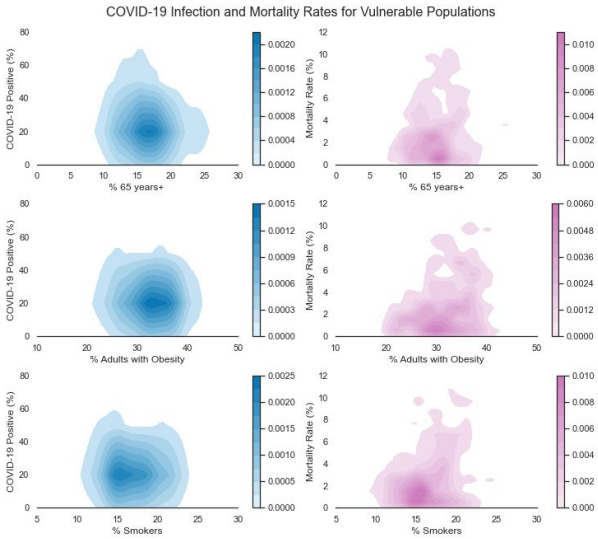
Preprocessing

For this Kaggle challenge, we chose to focus on investigating which populations are at risk of contracting or dying from COVID-19. One great challenge for this question was deciding what data to use and what specifics to focus on. We were provided with a large variety of data to work with and decided to limit our investigation to the United States, specifically at the county level, since many medical protection policies, such as shelter in place orders, were implemented county-wide. Given that several news sources have linked immuno-compromised and minority populations with the virus, we decided to use the provided data to investigate this pattern and see what we could find.

We worked with several datasets for this part. We used the New York Times' (NYT) county-level dataset to observe the number of COVID cases per county. We removed points where the counties were labeled "Unknown" except for states/ regions with only one county (Ex: Guam, District of Columbia). We also matched some differing county names using string correction (upper case, removed "City" suffix from county names, etc.). Since the dataset indicated the number of new reported cases and deaths per day, we grouped the dataframe in Python by FIPS county codes (a unique identifier for each county) and summed the cases and deaths per county. We also knew from prior knowledge that the global mortality rate for COVID patients was 2%. Thus, in order to see if certain populations are more vulnerable, we calculated a mortality rate by dividing the total number of deaths by cases. Because the number of cases is also linear with respect to the population of the county, we calculated a percentage of infected residents by dividing the total number of cases by the population for each county (found from ESRI's social vulnerability data).

For HIFLD's "hospitals" dataset, we similarly removed hospitals with "Unknown" FIPS codes and grouped by FIPS. We then divided the total bed count per county by each county's population to get the expected bed count per person. Since we knew of several risk factors, such as obesity, diabetes, age, and prior lung issues, we wanted to observe if counties with higher percentages of these risk populations exhibited a higher mortality or infection spread rate. We inner joined the cleaned data with the New York Times cleaned data. Finally, we were interested in socially vulnerable populations, such as minorities, disabled, low-income, or high unemployment rates. Thus, we also inner-joined ESRI's social vulnerability dataset with NYT's case data.

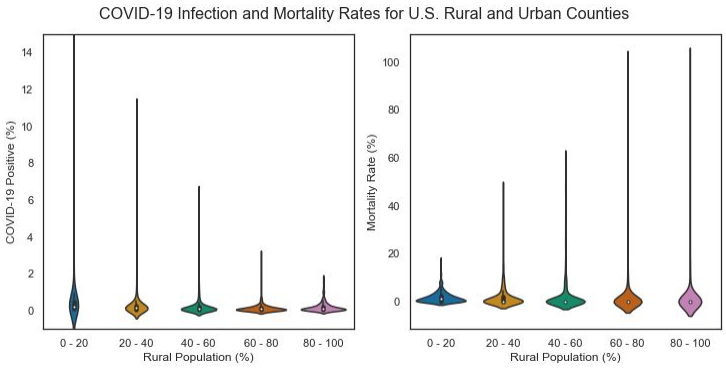
Lastly, we were interested in comparing trends in rural (cities with less than 2,500 people) and urban areas. This wasn't provided in the Kaggle dataset, so we pulled the "PctUrbanRural_County" from the Census 2010 website, as the Kaggle challenge suggested looking into outside datasets as well. Since FIPS codes were not provided, we joined the (county, state) name pairs with the NYT data and string matched differing county names. We specifically looked at the rural and urban population percentages & densities. Because the population percentages were continuous, we used Pandas cut() function to bin the rural/urban percentages into categorical labels (from 0-20%, 20-40%, etc.), which was easier to analyze and visualize. Our overall dataset ended up consisting of 2,400 counties, out of the 3,007 total in the U.S. (according to the Census in 2016). The remaining counties were lost due to pre-processing steps or simply lack of data.



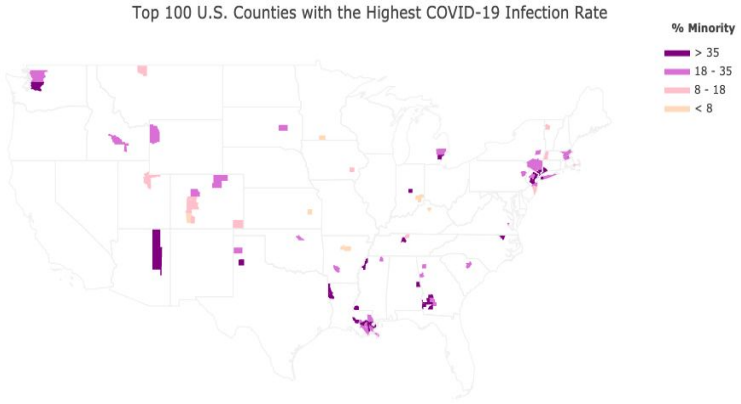
Exploratory Data Analysis

Since we were mainly interpreting the data, we did not build any models for this part. However, we experimented with the data to explore our hypothesis. We first focused on immuno-compromised or vulnerable populations and noticed

that while the infection rate (percentage of the county's population who were confirmed with COVID-19) is overall constant amongst the groups, there is a slight positive correlation between mortality rate and counties with higher obesity rates and smokers. These findings are in line with current public knowledge regarding risk factors for COVID-19 patients.



We also found an inverse correlation between infection and mortality rates in primarily rural vs. urban counties. As indicated in the violin plots, counties with higher rural populations have far less residents declared as COVID-19 positive (the violin is fatter at 0%), while having a much higher mortality rate (skinnier violin). The tails also indicate outliers in the distribution. We think that the long mortality tails and short infection tails show that rural populations are either more vulnerable, have less access to medical resources, or aren't testing as much as urban counties. To investigate this, we compared the number of beds per 10,000 people and overall health scores in rural vs. urban counties and found that rural areas actually have more beds per resident and are just as healthy as urbanites. Thus, we think that the discrepancy might be due to less testing in rural cities, where infection rates are unrealistically low and mortality rates are inflated because patients are only tested once they are too sick to recover.



Finally, we found a high correlation between minority populations and high infection rates. Nationally, the mean minority percentage is 18. However in the top 100 counties with the highest infection rates, we found that 82 had higher than average minority populations, with a mean of 37%. Without significant differences in county incomes, we thought that this might be due to higher exposure in minority labor populations, since laborers can't social distance if they are in essential services. These counties are mapped with the national percentiles.

Results and Conclusion

Throughout this exploratory data analysis, we came across a lot of interesting finds. Recent current events related to Tyson Foods brought to light issues that are occurring at the county level, in rural areas, and with high minority populations. In our investigation, we were able to confirm a slight increase in risk for populations that are obese and smoke. We also found interesting correlations that rural areas had a higher mortality rate, and minority populations are exhibiting a higher risk of exposure to COVID-19. It is imperative for rural areas and areas dense with minority populations to become part of the conversation and involved in managing COVID-19 effectively, in order to prioritize their resident's health.

References for Part A & B

Baranovskij, Andrej. "COVID-19 Growth Modeling and Forecasting with Prophet." *Medium*, Katana ML, 8 Apr. 2020, medium.com/katanaml/covid-19-growth-modeling-and-forecasting-with-prophet-2ff5ebd00c01.

US Census Bureau. "2010 Census Urban and Rural Classification and Urban Area Criteria." *The United States Census Bureau*, 2 Dec. 2019, www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.