AC 209, Group #25 Nick Normandin, Sam Thau
Modeling COVID-19 in the US Bliss Perry, and Michael Yin

Milestone 3: EDA, Baseline Models, and Revised Project Statement

Project Statement

We have created a simple model to predict the number of new daily COVID-19 cases for a given county based on demographic data and the number of cases recorded in the few weeks prior. Additionally, based on the data we were able to gather and our exploratory data analysis, we have decided to extend our project to analyze whether protests (largely associated with the US Black Lives Matter movement) are associated with higher rates of COVID infection.

Given the tense political climate within the United States, many have been quick to point fingers at groups that seem to be violating social distancing guidelines. For instance, some claim that BLM protests occurring in the wake of George Floyd's killing must have led to the spread of coronavirus. However, a number of economists associated with the National Bureau of Economic Research found that there was a minimal impact on the overall spread[1]. We intend to investigate further using a data-driven and non-partisan approach.

Our analysis may lead to important policy actions: if COVID cases did indeed seem to spike as a result of protests, it will be necessary to evaluate how public health and the freedom of speech and assembly may be reconciled. On the other hand, if COVID cases did not rise, the unique circumstances of the protests and any health precautions taken can be a guide for reducing the risk of coronavirus transmission more generally, specifically when crowds may assemble. Our work could also be important for protest organizers.

Below, we have included descriptions and analysis for the data we've collected and cleaned. Additionally, the key insights we've drawn from our basic county-level COVID prediction model are included. Of course, we hope to improve both this model and our analysis of how protests and COVID are related in our completed project, making them more complex and drawing on additional data as necessary. We are planning to examine both recurrent neural networks as a prediction model, and to utilize difference in difference methods to estimate the impact of protests.

Demographic Data

To properly understand the COVID-19 pandemic in the United States, it's important to analyze the spread of the virus within the context of unique regional demographics. The demographics of a county are important for understanding how much risk the population may be due to comorbidities or age, and may provide insight on likelihood to seek treatment, quality of treatment received, and even quality of data collection.

In the United States, counties are uniquely identified by a Federal Information Processing System (FIPS) code, where the first two digits correspond to a state and the final three correspond to a county. We use the FIPS code as the merge key to combine data from several sources:

Basic census data: County-level demographic data from JHU and the US Census, containing FIPS code, county / state names, male and female populations, median county age, as well as the latitude and longitude of the center of the county.

IHME health data: Some additional health data from the Institute for Health Metrics and Evaluation (IHME), containing time series records (which we cleaned and standardized) of mortality risk and life expectancy measures across every county for several age ranges dating back to the 1980s.

Census income data: County-level income data from the most recent Census survey (2018).

Census land area data: 2010 Census data for the land area (in square miles) of each county

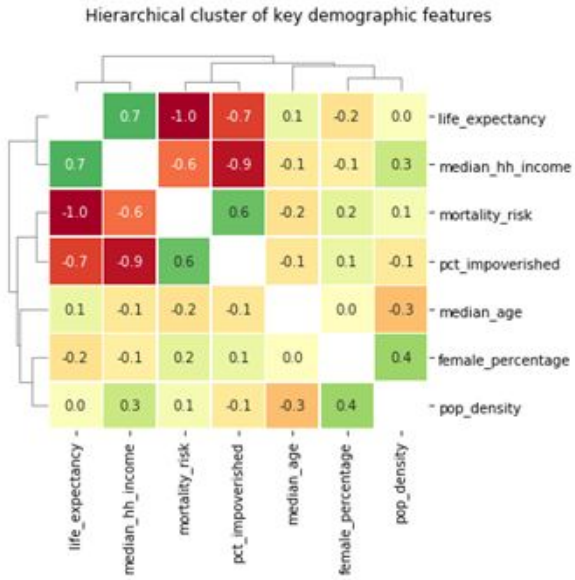
We engineer two features that we believe may have some relationship to the spread of COVID-19. The first is related to the rate at which the population is impoverished. We also use the population variable and the estimated land area variable to determine the population density in terms of inhabitants per square mile.

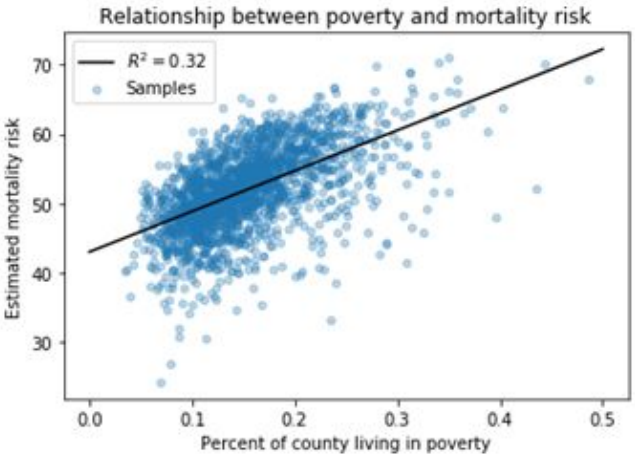
In our data exploration, we conduct a preliminary analysis of seven key features: estimated mortality risk, life expectancy, percent of population impoverished, median household income, median age, female percentage of the population, and population density. We begin our analysis by assessing the correlation between features. Since we expect significant nonlinearities, we use Spearman correlation as our metric of interest rather than the traditional Pearson.

We use hierarchical clustering of the correlation vectors (with the self-correlation masked) to arrive at a "cluster map" of the features which reveals several intuitive relationships and some surprising ones. Of major note is the fact that median household income and life expectancy have a strong positive correlation, and that those two features are

[1] *Black Lives Matter Protests, Social Distancing, and COVID-19*. Dhaval M. Dave, Andrew I. Friedson, Kyutaro Matsuzawa, Joseph J. Sabia, and Samuel Safford. NBER Working Paper No. 27408. June 2020, Revised August 2020

negatively correlated with mortality risk and the percent of the county living in poverty. We also see some correlation suggesting that high population density counties have a higher percentage of female residents, higher income, and lower median age.





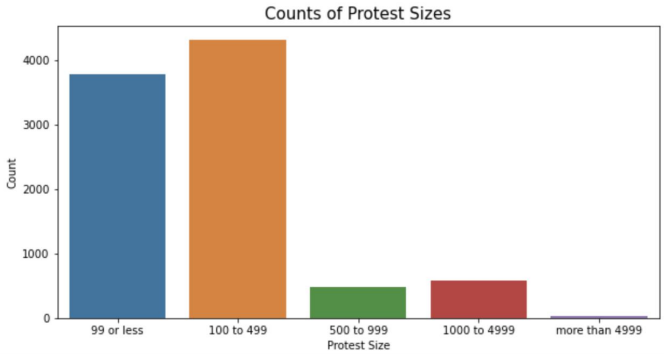
Of these findings, we expect that features related to regional poverty and mortality risk will be critical to understanding both the impact of the disease and the efficacy of interventions. We find a relatively strong relationship between poverty and mortality risk with an R^2 value of 0.32, which we believe will be an important contextual feature in understanding the impact of mitigating protocols for slowing the spread of COVID-19.

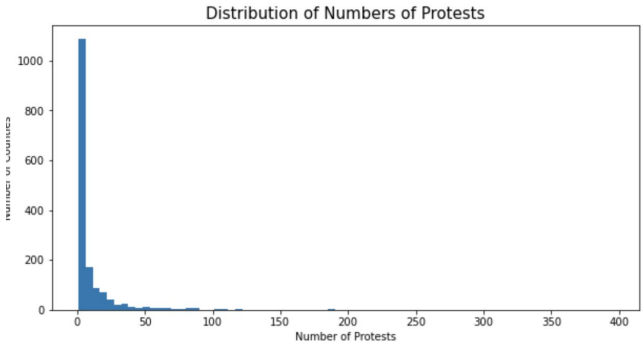
Protest Data

The Armed Conflict Location and Event Data Project publishes a continually updated dataset with detailed information about the vast majority of protests and demonstrations in the United States. We worked with their most recent dataset from the summer of 2020, that spans May 24th through November 7th of this year, with nearly 17,000 entries compiled from news stations. The data contains three types of information: location (city, state, latitude, longitude), type of protest (organizations, if there was violence), and a general notes category (data sources, qualitative descriptions).

To make this dataset compatible with the rest of our data, we used the latitude and longitude to attach a FIPS code via a Federal Communications Commission API. From the qualitative notes, we extracted estimates of protest size. Of the 16,880 records, 6,603 explicitly indicated there was no event size associated with the entry. Of the remaining 10,277, we were able to categorize 9,201 by finding numbers or keywords (like “few hundred” or “dozens”).

We can start with some histograms to get a sense of protest sizes and frequencies.

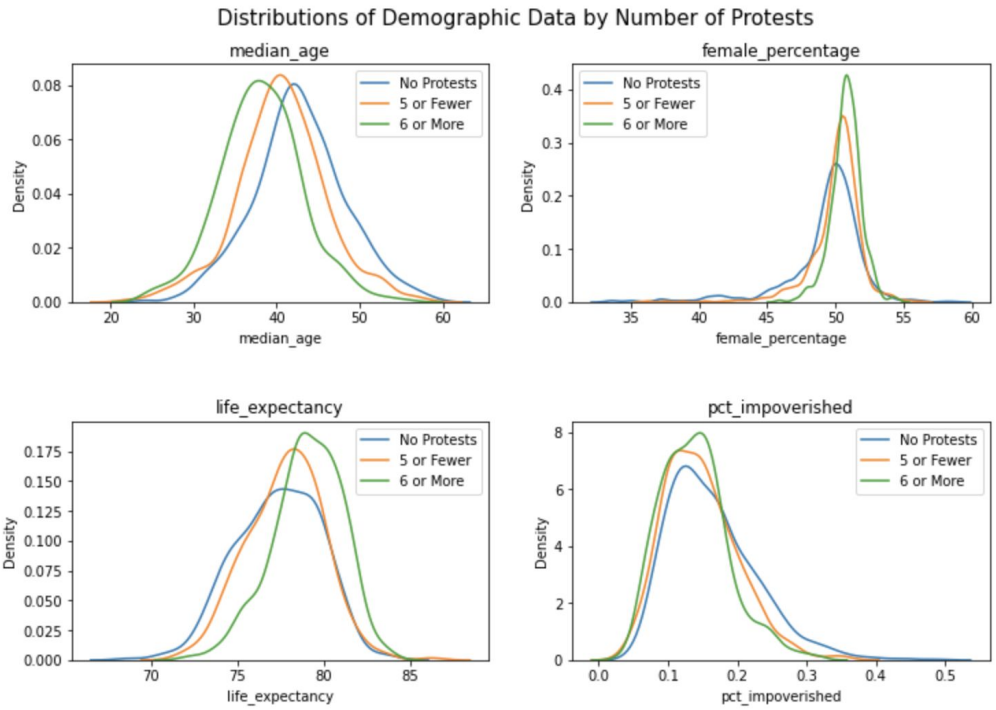




One thing to note is that by far the most common category is actually “unknown,” with a total of 7675 compared to the 100 to 499 category value of 4321. The vast majority of counties that had at least one protest had an extremely small number of protests.

We can make some comparisons between three groups of counties (as categorized by FIPS code): counties with no protests, counties with a small number of protests, and counties with many protests. There were only 271 counties

with no protests. We can split counties that had 6 or more protests from those that had 5 or less. We then joined the data with the demographic data to look at the characteristics of each group.



We can see that counties with more protests were more likely to be younger, more female, have longer life expectancies, and have a lower rate of poverty. This makes some intuitive sense, as those are all traits associated with urban areas which tend to be more liberal and have higher concentrations of minorities, and thus are more likely to protest.

COVID Data

For our COVID case data, we drew from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, which itself aggregates data from a variety of US and international sources. In particular, starting with March 23rd (when almost all US counties were included and the data for COVID cases by county settled into a consistent format), we pulled the daily data for confirmed cases and filtered out everything but US counties to get confirmed cases by FIPS code. Differencing the counts for consecutive days, then, we obtained new confirmed cases by day for each county.

Model

We trained a simple linear regression model to serve as a baseline indicator for developing more complex COVID-19 prediction models. This model is designed to predict the number of daily new cases on a county level given the number of new cases from the previous fourteen days. In addition, we selected as features the median age, male/female ratio, life expectancy, poverty rate, and median household income per county, intuitively assessing that these health-based and socio-economic statistics might have an impact on the spread of the virus. Future models will factor in protest data as well to further illustrate the specific effect those events have on the virus' spread.

We trained our data on an aggregated dataset comprised of the COVID-19 data above and the demographic data, with any example with a null case count within the last 14 days dropped, as well as any example with one particular outlier (the result of a re-categorization of jurisdiction of New York County on August 31st)[2] in its 14-day history dropped as well. Furthermore, we standardized each feature to have mean/variance 1.

Our baseline model produced the following results:

[2] https://github.com/CSSEGISandData/COVID-19/issues/3103

Mean Squared Error

Train	2288.5313
Test	1410.6052

Coefficient Values

1_before:	9.1863
2_before:	11.2791
3_before:	11.6582
4_before:	6.7203
5_before:	3.9931
6_before:	14.4918
7_before:	15.5512
8_before:	1.1444
9_before:	0.7076
10_before:	0.2446
11_before:	4.0721
12_before:	0.4449
13_before:	-3.4698
14_before:	1.7375
median_age:	-0.9844
female_percentage:	0.4895
life_expectancy:	0.7505
pct_impoverished:	0.4115
median_hh_income:	0.6003

Given the relative simplicity of this model, any analysis of these results should be utilized mostly to get a high level picture of the data. Nevertheless, two key trends stand out:

- By a substantial margin, the most significant coefficients seem to be the number of cases within the last seven days of the date in question. This intuitively makes sense given that a high spread of cases reported in recent days will usually lead to a continued high number of reported cases.
    - There is some fluctuation in the coefficient values of the 14-day history (with the 13_before value even taking a negative value), yet this is likely due to the fragility of our single, non-bootstrapped model and correlation between the values in these columns in the training set (because we created our training data by sliding a 14-day window over the entire COVID history of each county).
- Furthermore, the coefficient values are typically more significant for the case history features than for the demographic data. This makes sense, given that the primary determiners of the number of new cases in a day are the current number of local cases from previous days, rather than demographic information.