An Analysis of COVID-19 Mortality Rates in U.S. Counties

Indra Mar

DATA 100 Final Project

Abstract/Question Framing

My project is based around tracking and predicting counties' COVID 19 related death rates with regards to the national average. I specifically wanted to know if there were indicators in the data that could be used to predict whether a county would fall above or below the national average. I also wanted to see if I could go so far as to predict the exact death rates, however, I was not very successful in this endeavor. Thus, my two questions are:

Can I accurately predict a county's COVID 19 related deaths?
Can I predict whether a county will fall above or below the national average of 4.3%[1]?

My motivation behind examining these two questions was based around the intensive news coverage surrounding the varying degrees to which different parts of the country had been hit by the pandemic, as well as their differing reactions. I was curious if one could predict the severity of the crisis within different parts of the US (measured by death rate), as well as potentially link that severity with certain characteristics held by the area. This would then allow for both predictions to be made about the severity of the outbreak in the future and analysis of the factors that affect that severity, creating a mitigation tool of sorts for the future.

The average death rate (# dead / confirmed cases) in the United States is approximately 4.3%. I chose to use the national rate as a rough baseline as there was no county that could be used as a base comparison for others. The national rate, while undoubtedly much more complex (in both the number of variables and the interactions of those variables) at least provides a blanket standard against which to measure the individual counties' death rates.

Data Cleaning

I began by loading in the data that I was planning on using. In order to compare death rates with county features, I needed to first merge the Confirmed Cases and the Deaths data frames based on county. I then calculated the death rate by dividing the total number of deaths until 4/18/20 (the end of the data collection) by the total number of confirmed cases by the same date.

After cleaning the data a bit, which included dropping unnecessary columns, renaming confusing columns, as well as removing NaN and INF values, I merged the already combined DF into the Counties DF, allowing me to compare death rates with potential indicators. For removing the NaN and INF values, while I was somewhat concerned about decreasing the accuracy of my model (due to having fewer data points), I

[1] Amy Harmon, "Why We Don't Know the True Death Rate for Covid-19," The New York Times (The New York Times, April 17, 2020), https://www.nytimes.com/2020/04/17/us/coronavirus-death-rate.html.

believed the risk of potential bias created by filling the values in (due to the fact that in this case lack of data does not necessarily mean no COVID deaths or cases) outweighed that concern. I also restricted my analysis to counties with at least 1 death, as there were a large number of counties with 1 or no cases who's data would have a negligible or even negative effect on the model (such few cases do not represent a county's vulnerability, but rather appear as anomalies in the county health).

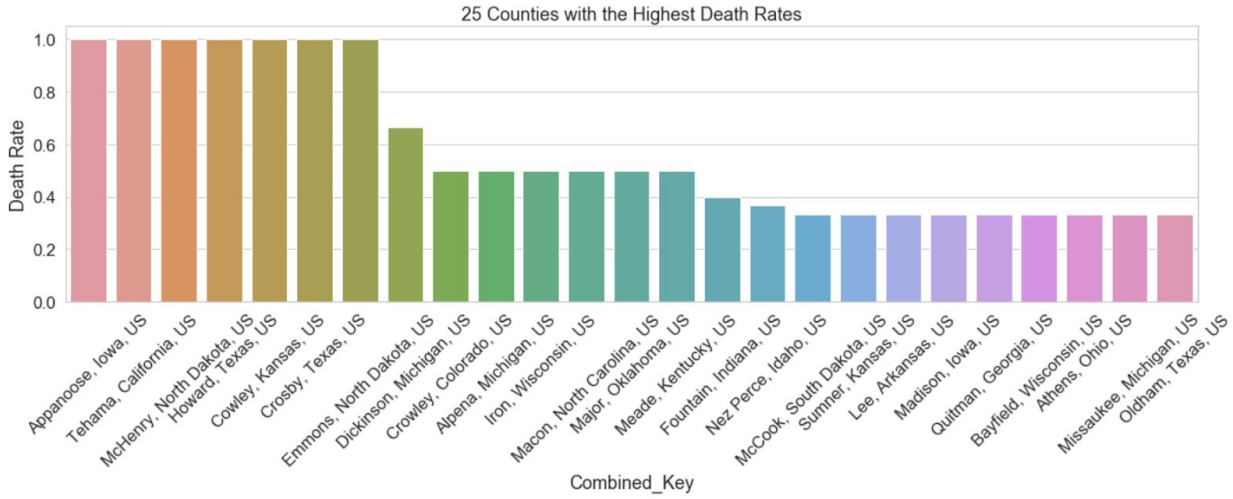
I did run into some problems merging the DF's however, as there are a large number of counties in different states that share names, yet the only common column along which to merge the Counties DF with the others is county names. I believed I had solved this problem by joining the county names with the states they were in, then using that combination as an updated key, however, I then found out that many counties, most notably New York City, and specifically all of those in Puerto Rico and Alaska, as well as a large number in Virginia, were listed without states. Though I attempted to add the state names manually, the number proved unmanageable and so I was forced to cut these counties from the analysis. I see this as a major flaw in my project, as NYC specifically is the epicenter of the epidemic in the US, making its information particularly valuable.

UID	County	Cases as of 4/18/20	Combined_Key	Deaths as of 4/18/20	Death Rate
84036061	New York	135572	New York City, New York, US	13202	0.097380
84036059	Nassau	29180	Nassau, New York, US	1109	0.038005
84036103	Suffolk	26143	Suffolk, New York, US	693	0.026508
84036119	Westchester	23179	Westchester, New York, US	668	0.028819
84017031	Cook	20395	Cook, Illinois, US	860	0.042167
...	...	...	...	...	...
84027125	Red Lake	1	Red Lake, Minnesota, US	0	0.000000
84027127	Redwood	1	Redwood, Minnesota, US	0	0.000000
84050009	Essex	1	Essex, Vermont, US	0	0.000000
84005047	Franklin	1	Franklin, Arkansas, US	0	0.000000
84029075	Gentry	1	Gentry, Missouri, US	0	0.000000

Combined DF

Combined_Key	Cases as of 4/18/20	Deaths as of 4/18/20	Death Rate	PopulationDensityperSqMile2010	MedianAge2010	DiabetesPercentage	HeartDiseaseMortality	StrokeMortality	Smokers_Percentage	TotalM.D.'s/TotNon-FederalFed2017	#Hospitals	#ICU_beds	dem_to_rep_ratio	SVIPercentile	Death Rate over National Avg?
Autauga, Alabama, US	25	2	0.08	91.8	37.0	9.9	204.5	56.1	18.08153718	50.0	1.0	6.0	0.326563944530050	0.4354	1
Baldwin, Alabama, US	109	2	0.018348635853211100	114.7	41.1	8.5	183.2	41.9	17.489032597	536.0	3.0	51.0	0.253255217266029\0	0.2162	0
Calhoun, Alabama, US	66	2	0.030303030303030300	195.7	38.2	15.6	315.8	44.0	20.612560375	228.0	2.0	24.0	0.402921040620731\0	0.8252	0
Chambers, Alabama, US	240	11	0.045833333333333300	57.4	41.5	17.5	196.7	45.2	19.392927053	28.0	0.0	0.0	0.737472906775851\0	0.7382	1
Colbert, Alabama, US	15	1	0.066666666666666670	91.8	41.8	13.5	261.3	55.2	18.658821061000000	91.0	2.0	22.0	0.435641586059400\0	0.4374	1
Coosa, Alabama, US	22	1	0.045454545454545500	17.7	44.2	17.9	237.0	45.6	20.3899653623	1.0	0.0	0.0	0.527062956112689\0	0.5261	1
Covington, Alabama, US	21	1	0.047619047619047600	39.7	42.4	12.9	262.0	51.5	19.249451541	44.0	2.0	15.0	0.179021025599870	0.7723	1
Cullman, Alabama, US	42	1	0.023809523809523800	109.4	39.9	17.2	230.6	52.1	17.901071221	111.0	1.0	12.0	0.115129285511941600	0.5182	0
Dallas, Alabama, US	22	2	0.090909090909090900	44.8	37.7	13.9	308.1	56.2	24.089910070000000	85.0	1.0	18.0	2.217308688927300	0.949	1
Elmore, Alabama, US	58	1	0.017241379310344800	128.2	37.8	14.9	225.1	53.0	17.452501185	42.0	2.0	5.0	0.305529402279366	0.5401	0
Etowah, Alabama, US	93	7	0.075268817204301110	195.2	40.2	13.5	292.3	48.0	20.147398601	246.0	2.0	56.0	0.327521404506540	0.6564	1
Franklin, Alabama, US	17	1	0.058823529411764700	50.0	37.8	13.3	259.0	57.0	19.230583040000000	33.0	2.0	7.0	0.233093804231990	0.8417	1
Hale, Alabama, US	23	1	0.043478260869565200	24.5	40.0	15.5	253.7	52.0	22.077291339000000	5.0	1.0	0.0	1.534884966902900	0.9599	1
Houston, Alabama, US	63	2	0.031746031746031700	175.1	38.5	11.8	195.6	43.7	19.249782785	407.0	2.0	66.0	0.347045040354075\0	0.6838	0
Jackson, Alabama, US	38	2	0.052631578947368400	49.4	41.4	13.9	284.1	54.4	19.958920884	59.0	1.0	10.0	0.223035090596930	0.5726	1
Jefferson, Alabama, US	671	25	0.037257824143070000	592.5	37.1	11.7	190.9	59.6	17.622045263	4626.0	8.0	471.0	1.164029616649000	0.6621	0
Lauderdale, Alabama, US	23	4	0.173913043478260900	138.9	40.4	11.9	219.7	49.9	16.6560043	195.0	1.0	48.0	0.356715294549980	0.4468	1
Lee, Alabama, US	305	14	0.045901639344262300	230.9	29.5	9.5	180.0	45.9	17.560140407000000	249.0	1.0	28.0	0.613282491261519\0	0.6602	1
Macon, Alabama, US	24	2	0.083333333333333300	35.2	36.3	21.3	277.1	47.9	21.41499395	21.0	0.0	0.0	5.287211740041930	0.8971	1
Madison, Alabama, US	197	4	0.020304568527918800	417.7	37.3	12.4	197.2	42.4	16.8201737701	1188.0	2.0	100.0	0.731764968720730	0.3076	0
Marengo, Alabama, US	25	1	0.04	21.5	40.4	14.9	301.9	54.3	21.356000347000000	15.0	1.0	5.0	1.072998280145230\0	0.7621	0
Marion, Alabama, US	60	5	0.083333333333333300	41.5	42.8	18.1	299.0	54.1	19.060209807000000	12.0	1.0	4.0	0.127317971733191400	0.6710000000000000	1
Marshall, Alabama, US	138	4	0.028985507246376800	164.4	38.2	10.2	262.4	56.2	21.017853481	125.0	1.0	20.0	0.168200321554408\0	0.7261	0
Mobile, Alabama, US	637	25	0.039246467817896640	335.9	36.6	12.6	231.5	49.1	19.3699619\9	1410.0	4.0	157.0	0.758925943059020	0.7532	0
Monroe, Alabama, US	8	1	0.125	22.5	40.1	15.7	183.7	54.8	23.257389630000000	19.0	1.0	8.0	0.747540936866570	0.7481	1
Montgomery, Alabama, US	215	5	0.023255813953488400	292.5	34.9	13.0	179.7	54.8	19.68315291	807.0	3.0	88.0	1.733670646776400	0.8427	0
Randolph, Alabama, US	51	4	0.078431372549019\6	39.5	41.2	13.8	295.5	48.9	18.5983375\26	7.0	1.0	0.0	0.297335990066499\0	0.7191	1
Shelby, Alabama, US	257	7	0.027237354085603100	248.5	36.9	9.6	162.2	42.7	15.561157468	658.0	1.0	32.0	0.314667214417910	0.1169	0
Talladega, Alabama, US	44	1	0.022727272727272700	111.7	39.3	14.3	245.9	48.4	23.159862261	68.0	2.0	14.0	0.587998476569320	0.8736	0
Tallapoosa, Alabama, US	179	11	0.061452513966480400	58.1	42.4	13.3	227.1	45.6	18.180967185	65.0	2.0	6.0	0.405987935854053\0	0.8863	1
Washington, Alabama, US	15	1	0.066666666666666670	16.3	39.7	14.2	238.6	56.7	20.790093076	7.0	1.0	0.0	0.392916252896392\0	0.7303	1
Apache, Arizona, US	169	4	0.023668639053254400	6.4	32.4	15.4	142.8	31.7	21.513107022	64.0	4.0	0.0	2.073179611650490	0.9885	0
Coconino, Arizona, US	314	29	0.092356687898089800	7.2	31.0	7.7	124.3	30.4	15.990186028	436.0	3.0	41.0	1.535152548799600	0.7131	1
La Paz, Arizona, US	5	1	0.2	4.6	53.9	9.5	182.7	31.4	15.521010320000000	16.0	2.0	3.0	0.393454908819399\0	0.9238	1
Maricopa, Arizona, US	2491	70	0.028101164191087900	414.9	34.6	8.2	130.3	29.0	13.686396251000000	13059.0	34.0	1004.0	0.940518705155230\0	0.6354	0
Mohave, Arizona, US	52	1	0.019230769230769200	15.0	47.6	10.1	230.9	35.6	19.35284\3037	312.0	4.0	80.0	0.299489212449813\00	0.8624	0

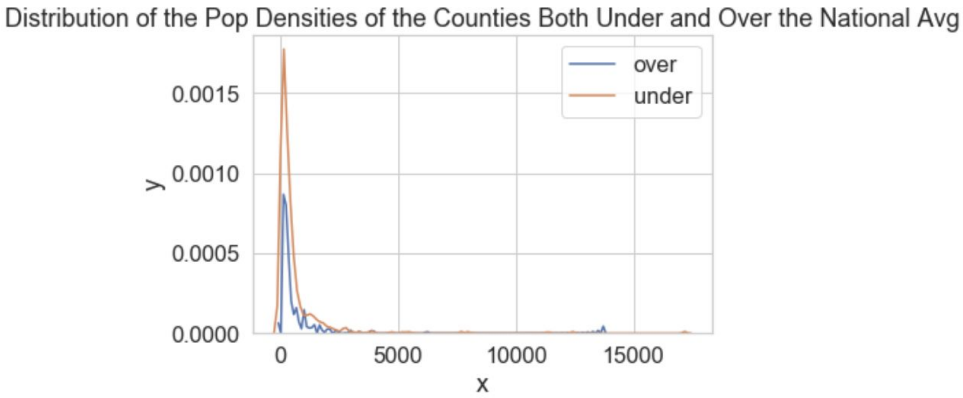
EDA and Data Visualization



For my EDA, I attempted to isolate the columns in the DF that would best serve as predictors for my model. I first gave each county a score of either 1 (over the national average) or 0 (under the national average), then analyzed both this measure and the county's death rate against the potential predictors in the DF. After plotting the relationship each category had with the county's death rate, as well as sorting the table in multiple ways (comparing the relative maximums and minimums for each category) I decided on Population Density (population density per square mile in the county in 2010), Percentage of Smokers (estimated percentage of adult smokers in the county (2017)), SVIPercentile (the county's overall percentile ranking indicating the CDC's Social Vulnerability Index (SVI); higher ranking indicates greater social vulnerability) and Median Age (median age of county in 2010) as the 4 predictors I would use for my model.

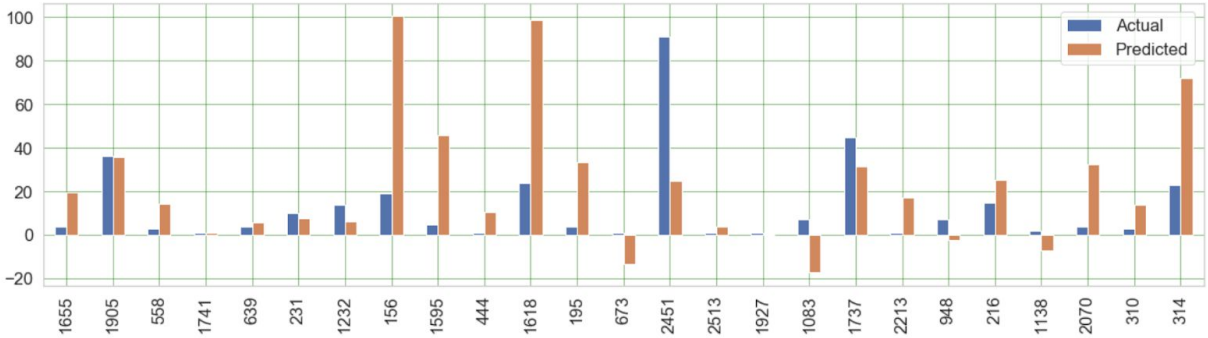
I originally was interested in analyzing the relationship party affiliation had with death rate (ratio of democrats to republicans in the county population), however, while definitely an interesting potential avenue of future research (especially given more recent data), it served only to decrease the accuracy of my model.

I used multiple visualizations in determining my preferred predictors, however, I have elected to only show those belonging to the predictors I ended up choosing, as I do not have the room for over 30 different visualizations. I primarily used Barplots, Lineplots, and Distribution plots in order to make my comparisons.

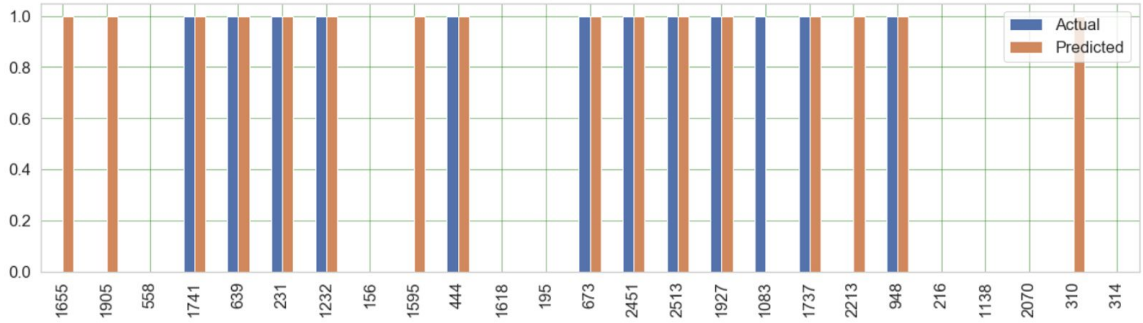


Method and Experiments

For my experiments, I first attempted to predict the total deaths in each county based on the 4 predictors I chose. I used them to fit a training set I had split from 80% of the data using the SKLearn linear regression model (due to the continuous nature of the outcome I was looking for), then tested my model on the remaining 20% of the data (standard train/test split). Though my model was pretty inaccurate, as demonstrated by the bar graph below, it still proved to be an interesting analysis. Though the majority of my predicted values were substantially off from the actual values, through multiple iterations of the model, I was able to, I believe, determine that the variables given in the dataset are insufficient to perform this sort of prediction. The actual number of deaths in each county, as well as the counties themselves, vary too greatly to create such an accurate prediction.



I next attempted to predict, using the same predictors but fitted using logistic regression (due to the binary nature of my target: either 1 for over or 0 for under the national avg), whether a county would fall above or below the national average death rate of 4.3%. In this, I was substantially more successful, and I was able to bring my model from a training accuracy of 12.25% all the way up to 63.43% (I was performing EDA while building my models, thus I was able to bring the accuracy up by going back and forth substituting in different predictors) which I consider a massive improvement. I believe I was more successful because the variables given in the dataset serve as better indicators of severity, rather than actual deaths. They highlight the potential weaknesses or vulnerabilities each county has in relation to one another, which, when given a standard against which to measure (national average), allows us to create a relatively accurate severity prediction. The increased accuracy of this model is demonstrated by the bar graph below.



Analysis/Questions

(i) Two of the most interesting features I came across when attempting to answer my two questions were the importance of both the percentage of smokers and the percentage of the county population that ranked highly on the CDC's Social Vulnerability Index in determining how severely a county would be hit by COVID deaths. The smokers' percentage would seem to be outweighed in importance by other pre-existing conditions, such as diabetes, heart conditions, etc. or even by the outright mortality rates from respiratory illnesses, yet, when using those variables as features in the models, I achieved far lower standards of accuracy. While I'm not sure why this is the case, I hypothesize that it is possible that smokers take fewer precautions from the virus (from simply sheltering indoors, to using a mask when outside) when compared to those with other pre-existing conditions (though their objective medical vulnerability may be the same), making them experience higher death rates. The SVI percentile's importance demonstrates how the virus affects the poor and needy much more severely than those of affluence. Poorer counties generally had higher death rates (which coincided with lower numbers of hospitals and ICU beds), making the case for a more class conscious examination of the virus-related death rates.
(ii) I thought that perhaps a political party based analysis could yield some information as to the differences in the responses of local leadership and citizen bodies with regards to the virus and its death toll, especially in light of the attention the polarized response has gotten in the media. However, further analysis and model testing revealed that there was very little correlation between the two, making the feature unsuitable for model fitting.
(iii) I had some challenges in data cleaning, as the data was often incomplete, meaning I had to decide where it was appropriate to fill in data (almost never), where my model could, with a mall accuracy penalty, still function despite the missing data, and where I needed to remove or cull the data due to missing values. I worry that missteps in this process affected both the accuracy of and the levels of bias within my model. I also wonder if there are both more recent and more useful data to be found, especially with regards to predicting the actual death tolls or rates (as opposed to just classifying them against the national average).
(iv) The fundamental limitation in my analysis is the lack of accuracy in my model predicting actual death tolls. The variables I had available, while useful in determining relative severity and/or vulnerability, are not as insightful when it comes to the actual quantitative data. I also think that an analysis of the variables given on a national level could (as opposed to solely county level) could provide some insight into how to more accurately create my logistical regression model.
(v) The primary ethical problem I had with the data was the lack of HPSA (Health Professional Shortage Area) data, as this would have allowed me to perform a more

detailed analysis of the effects of medical resources in the area, and thus class effects on the area. The number of hospitals and ICU beds, while useful, did not provide objective or controlled measures of vulnerability, as the effectiveness of those numbers depend on a host of other variables within the county, whereas HPSA measures are scaled to the county, making them more useable as standards of care. There also appeared to be a lot of useful information in the unabridged counties CSV (which was not included in the dataset, but was visible on the Github page that the set came from), which makes me wonder how the original publisher, and even the course staff, chose this particular information.

(vi) As stated previously, more recent data (this set only goes to the 18th of April), as well as more comprehensive data (on a national scale as well as on a county scale) would help strengthen my analysis. The information on pre-existing conditions, for instance, could include conditions such as asthma and various types of cancer, both of which are extremely common and are affected radically by COVID. The national averages for the information given could also be included, which would provide a better standard of comparison. Finally, a dataset on adherence to anti-viral government measures (social distancing, masks, shelter in place orders, etc.) would help determine population created causes for the differences in severity between counties.

(vii) The primary ethical concern I had in analyzing these questions was the importance I discovered in the SVI index in determining vulnerability to COVID. Poorer communities, in general, were hit harder than wealthier ones, making the class-based effects of COVID the next obvious avenue of future study. Poverty should not condemn people to higher risk, yet, all too often in the modern world (not just with regards to COVID), it does.

Concluding Thoughts

Though I was only partially successful in answering my questions, I believe that this work, especially the classification model, serves as an interesting baseline for future COVID research. While I was originally interested in the political implications of the virus and its death rate, through this work, I saw that perhaps a more relevant and revealing path of analysis would be the economic and class-based implications of the two. I also realized that qualitative data about viruses is much harder to predict than other aspects of society (thinking back to two of the bigger examples in class, tips, or the price of diamonds for instance). I believe this to be due to the unpredictable (and often unknown) means of viral spread, as well as the lack of understanding of how the virus operates. It is difficult to accurately predict death tolls because our knowledge and understanding of the virus is constantly changing. Despite this, however, predictions such as this one create new potential questions for examination and perspectives of thought, making analysis such as this one valuable, if somewhat outdated, or inaccurate.