COVID-19 in California: An In-Depth and Transparent Analysis

As the Coronavirus continues to spread it has become ever more apparent that we are dealing with the biggest public health challenge of our generation. Fortunately, California's state and county officials have been working tirelessly for months to help protect and ensure the safety of all citizens. Thus, the purpose of this project is to utilize my data wrangling, visualization, and analytical skills to obtain some insight on the spread of the coronavirus in California and learn about the challenges that our government experts are facing while making these tough decisions. In this report, I will not only share my results but also go in-depth into the decisions taken to come to my conclusions.

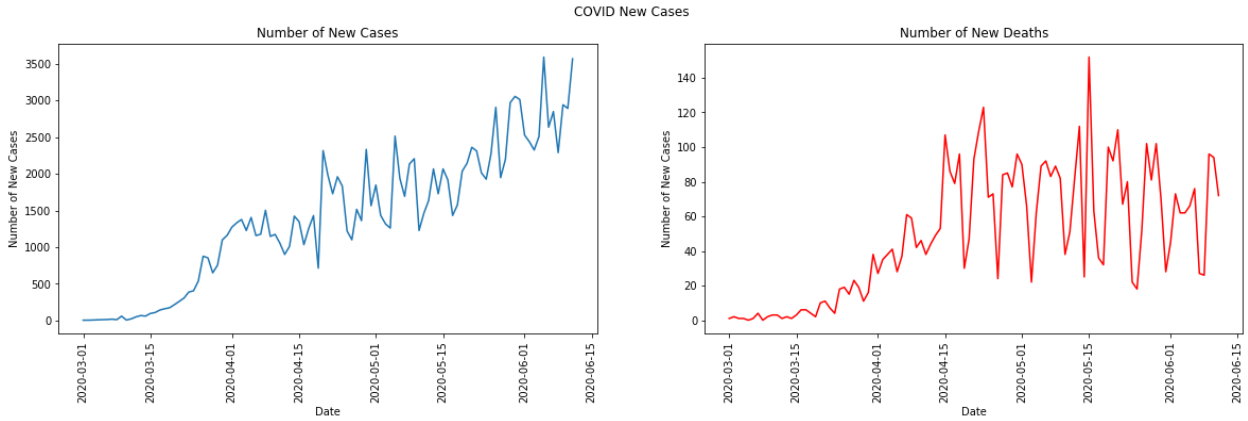
This analysis is divided into three parts:

1. Overview of the Spread of COVID-19 in California
2. Analysis of some Factors that could explain the differences in spread among counties
3. Analysis of Testing Access, and ICU availability across California.

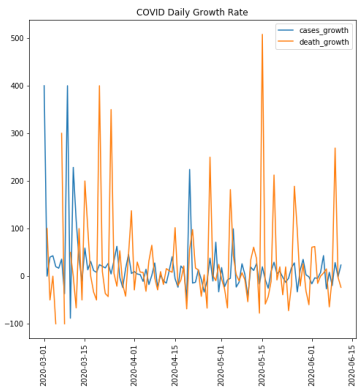
Overview of the Spread of COVID-19 in California

The Coronavirus Data was obtained from the New York Times GitHub, and the California shapefile was obtained from the California Government.

Before I delve into individual counties, I first want to learn about how California has progressed as a whole.



These graphs represent the number of new cases /deaths and the growth rates from March 1st until June 9th. The number of new cases has fluctuated greatly from day to day but has shown a trend of overall increase. However, while the number of new deaths fluctuates a lot more, we do not see this same pattern of overall increase from April-15 to June. Additionally, looking closer at the rate of overall increase for new cases from the COVID new cases graphs, it appears that from April 15 to mid-May (approx. May-15th), the rate of overall increase is smaller than from mid-May to June-9th. This could be due to reopening policies and the protests. The COVID Percentage Daily Growth Rate is interesting as it shows some dates had an unusually high peak. While the early peaks in March could be attributed to small numbers, the large peak in cases in Mid-April, and

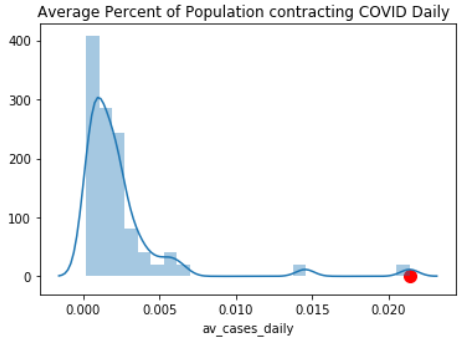


the large peak in deaths in Mid-May is interesting. There could be many possible reasons. One reason could be delays in hospital/ other sources reporting, which could erroneously attribute more cases to a certain day. Another reason could be abnormal social events. For instance, it may not be a coincidence that the spike in growth in Mid-April (approx.- 15th-20th) occurred during the increase in stay at home order protests (https://www.foxnews.com/us/california-protest-erupts-over-states-coronavirus-stay-at-home-rules). However, I do not know the exact reason and it could still be something else.

Now that we understand how the Coronavirus has progressed in California as a whole, lets analyze the growth in specific counties and uncover which counties have been hit the hardest. To do this, I will be utilizing two metrics- average number of people who contract the coronavirus daily as a percentage of total population and the ratio of total death/cases. I chose to observe the percent of total population instead of the number of people, because I did not want to bias my results against small counties. Also, I decided to use the ratio of death/cases to understand where the virus is more deadly since the total percentage of population is very small and the death/cases ratio is a better representation off the killing efficiency of the virus.

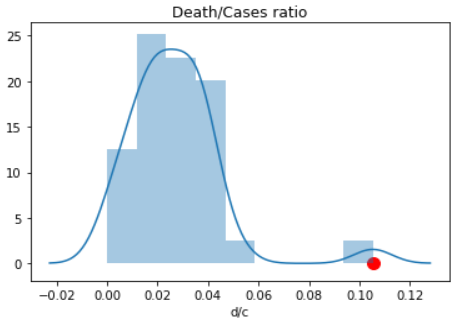
Average number of new COVID-19 Cases as a percent of population

The distribution seems to be right skewed. Most of the percentages are between .001% and .002%. The outliers are Imperial (shown in red), with a percentage of approximately 0.0213%, and Kings County with a percentage of approximately 0.0145%. These are more than 10x higher than the median California county! Following these counties are Los Angeles County with approx. 0.0065%, Tulare with approx. 0.0058%, and Modoc with approx. 0.0055%. It is important to note that these are not the percentage of people with COVID-19 cases (to come later), but instead represent the percentage of people we can expect to obtain COVID-19 on a typical day.



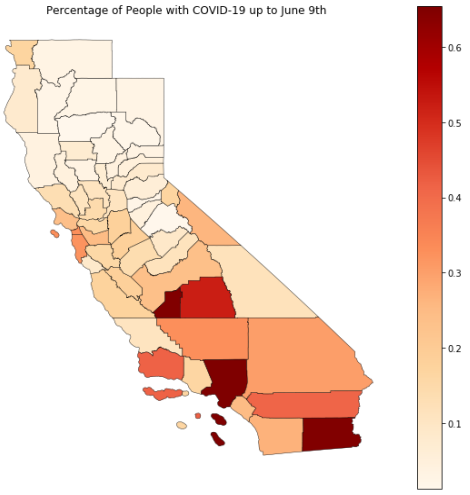
Death: Cases Ratio

Like the percentage distribution, there is an outlier in the total death/cases ratio. However, this time it is Yolo county, with a death: cases ratio of approximately 0.11 which is approximately double the next largest county (San Diego), which has a ratio of 0.05. Following San Diego is Los Angeles County with 0.041, Tulare 0.039, and San Mateo 0.038. Yolo Counties position on top could be due to a combination of its low total number of cases (only 228), and the "riskiness" of the population who have contracted the virus. Also, it is interesting that Los Angeles and Tulare have maintained high positions on both lists, while Kings and Imperial counties are not even in the top 25 when considering death: cases ratio. Furthermore, because some counties have very few cases and deaths, I decided to limit the distribution to counties with 100 or more cases. Also, it is apparent from the histogram and the Shapiro–Wilk test, that this distribution is normal when Yolo County is removed.



COVID Cases as a Percent of Total Population

Now, let's consider the number of people with COVID-19 as a percent of population as of June 9th:

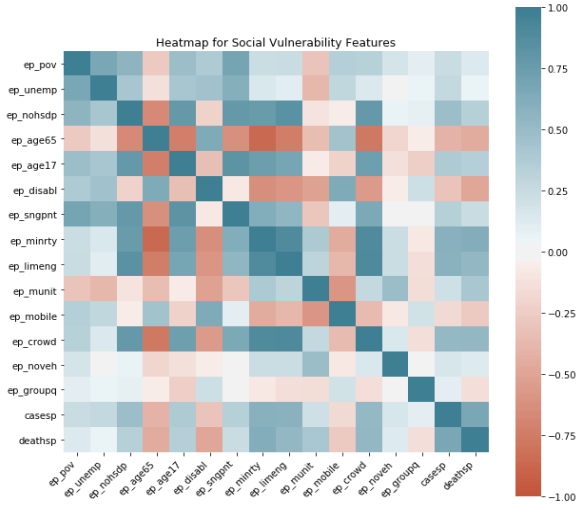


This map clearly shows that the counties with a higher percentage of COVID-19 cases are mostly in Southern California, with the number of lightly impacted counties increasing as we progress north. Furthermore, the county with the highest percentage is Imperial with approx. 1.75% infected, which is more than 10x higher than the median county. Following Imperial is Kings with approx. 1.08% infected, Los Angeles with approx. 0.65% infected, Tulare with approx. 0.52% infected and Santa Barbara with approx. 0.41 % infected. Taking all three of these statistics into account, I believe that the county hit the hard by COVID-19 has been Los Angeles county. Even though Imperial county has a very high percentage of cases, LA county's top three presence when it comes to cases and deaths makes it the hardest hit county.

Analysis of Risk Factors

Source: CDC Social Vulnerability Score 2016

The source I will be analyzing is the Center for Disease Controls Social Vulnerability score data, which I obtained from Kaggle. This dataset, which is on the county level, comprises of features that can be grouped into four categories- Socioeconomic Status, Household Composition & Disability, Minority Status & Language, and Housing & Transportation. Furthermore, in the dataset, one feature is represented in three different formats- as a value, as a percent, and as a percentile. In my analysis, I will only look at the values as a percentage.



Preprocessing:

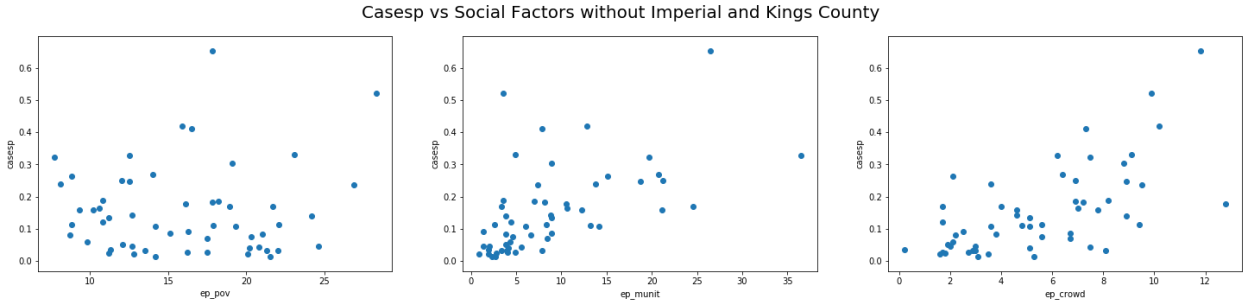
To determine which features are most relevant, I will perform a multivariate regression with the social vulnerability variables as the features and the number of cases as a percent of county population as the label. However, first, I did some preprocessing. The first step of the preprocessing was to choose the best features that minimize co-linearity. While there are statistical tools that can do this like the Variation Inflation Factor, I decided to create a heatmap of the correlation coefficients and choose features that intuitively related to the spread of COVID-19 and are not that similar to increase transparency. Here are the features I picked and why:

ep_pov: I chose the Percentage of persons below poverty, as poverty may influence social distancing adherence.

ep_crowd: I chose the percentage of occupied housing units with more people than rooms. This seems vital to the spread of COVID-19 as this means that there would be more frequent human interactions. This feature seems to be correlated with the Age65 and minorities which is why I did not include these features in my model, as I believe crowd could explain why these communities could experience any disparities.

ep_munit: I chose the percentage of housing in structures with 10 or more units estimate for the same reason as ep_crowd. Interestingly, these two features are not linearly correlated, which makes it an appropriate choice.

My next preprocessing step was to remove outliers. This is a tough task since this is a small dataset. While removing lots of large data points will increase our r^2 performance, we also might lose important information. Therefore, to balance this I decided to just remove Kings and Imperial County.



Here are scatterplots which compare our features. From the first plot, poverty is not correlated. Thus, I decided not to use it for the regression. By observing the plots for crowd, I decided that adding a square polynomial feature would be beneficial due to the fact that the cases percentage seems to grow faster for 8-12 percent than for 2-6 percent. I decided to leave ep_munit alone. Initially, it may appear that ep_munit could benefit from the addition of a fractional polynomial feature since it appears to level off after 20%. However, the larger datapoints are sparse, and therefore there is reason to believe that this phenomenon could just be caused by a lack of data. The final step was to min-max scale the variables.

The Regression

I decided to preform two least squares regressions. The first regression is the model with all our features. The P-values show ep_munit is an extremely significant variable as it has a very small p-value. Crowd 2 also has a small p-value of 0.069, though it is not significant in this model as this is above the custom 0.05 threshold. Furthermore, it appears that ep_crowd is not significant in this model. The model explained 58.4% of the variation(r^2). The second model is a simpler model, as it contains just the variables without the squared feature. In this model, both variables are statistically significant with low p-values. Also, the R-Squared and adjusted R-squared values are only about .02 lower than the more complex model. This increases favorability for the simpler model as it is more transparent and almost as effective.

Dep. Variable:	casesp	**R-squared:**	0.584
Model:	OLS	**Adj. R-squared:**	0.559
Method:	Least Squares	**F-statistic:**	23.83
Date:	Sun, 14 Jun 2020	**Prob (F-statistic):**	8.85e-10
Time:	17:30:42	**Log-Likelihood:**	56.512
No. Observations:	55	**AIC:**	-105.0
Df Residuals:	51	**BIC:**	-97.00
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	0.0321	0.040	0.795	0.430	-0.049	0.113
ep_munit	0.2788	0.062	4.514	0.000	0.155	0.403
ep_crowd	-0.0744	0.204	-0.365	0.716	-0.483	0.335
crowd2	0.3937	0.212	1.855	0.069	-0.032	0.820

Omnibus:	11.816	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	17.754
Skew:	0.681	**Prob(JB):**	0.000140
Kurtosis:	5.427	**Cond. No.**	27.6

OLS Regression Results

Dep. Variable:	casesp	**R-squared:**	0.556
Model:	OLS	**Adj. R-squared:**	0.538
Method:	Least Squares	**F-statistic:**	32.50
Date:	Sun, 14 Jun 2020	**Prob (F-statistic):**	6.97e-10
Time:	17:51:39	**Log-Likelihood:**	54.716
No. Observations:	55	**AIC:**	-103.4
Df Residuals:	52	**BIC:**	-97.41
Df Model:	2		
Covariance Type:	nonrobust		

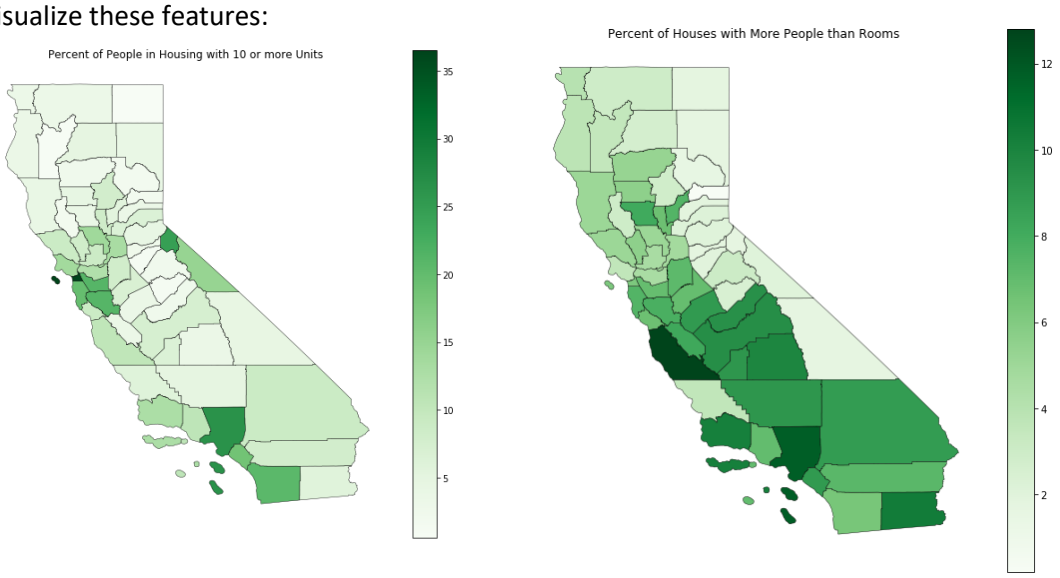
	coef	std err	t	P>|t|	[0.025	0.975]
intercept	-0.0256	0.026	-0.975	0.334	-0.078	0.027
ep_munit	0.2698	0.063	4.284	0.000	0.143	0.396
ep_crowd	0.2898	0.056	5.197	0.000	0.178	0.402

Omnibus:	14.147	**Durbin-Watson:**	1.946
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	16.557
Skew:	1.022	**Prob(JB):**	0.000254
Kurtosis:	4.744	**Cond. No.**	6.15

From this regression analysis, it is clear that the ep_munit and ep_crowd characteristics are very important characteristics that could explain the difference in spread of COVID-19 in California. Just these two variables explain more than a majority (55.6%) of the variation in total cases as a percentage as a population. Intuitively, this provides evidence to the theory that counties with more apartment/dorm-like structures, should have a higher COVID case percentage since people are in more frequent contact with each other.

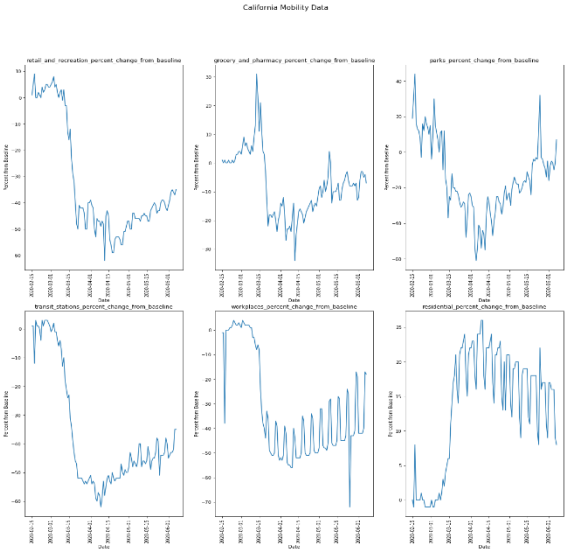
Analysis of EP_MUNIT and EP_CROWD

Lets visualize these features:

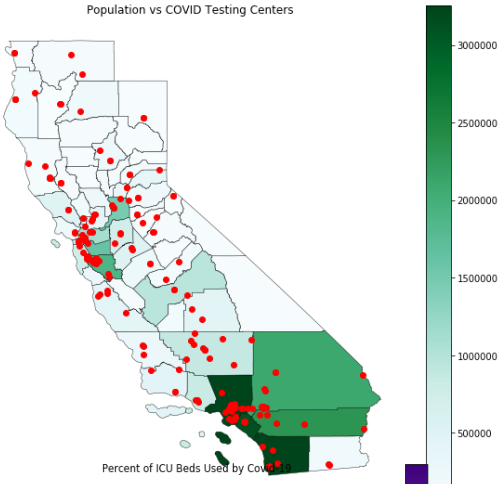


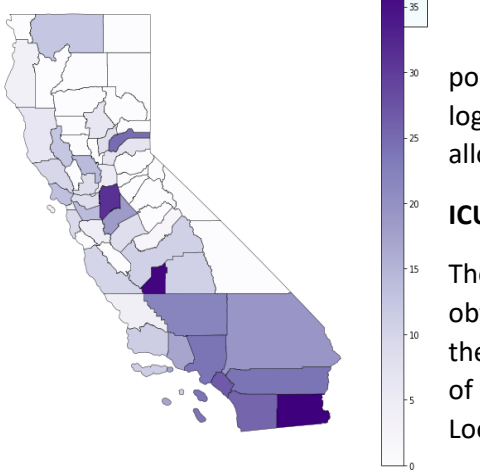
The first map represents MUNIT while the second represent CROWD. Regarding the first map, it shows that a majority of this type of apartment housing is concentrated in urban, high populated counties. In fact, San Francisco County leads this category with 36.5% of its housing in this manner. This is followed by Los Angeles County (26.5%), Alpine(24.6%), Alameda(21.2%), and Santa Clara(21.1%). This is to be expected of these regions as they contain a high volume of people in a confined area. However, the percentages depicted by the second map of overcrowded houses seems to be distributed more evenly across California. Central California seems to possess lots of crowding relative to the other counties, which differs from the second map. The top crowded counties are Monterey(12.8%), Los Angeles (11.8%), Imperial(10.4%), Santa Barbara(10.2%), Tulare(9.9%). It is especially noteworthy that Los Angeles county, which I deemed the hardest hit by COVID-19 is near the top on both lists.

Furthermore, the strong relationship between the percentage of COVID-19 cases in a county and housing characteristics may be due in part to California's social distancing policies. The time series graphs represent Google Mobility for grocery/pharmacy, parks, transit, retail/recreation, residential and workplace. While Mobility in places like parks and grocery have returned to base levels recently, overall, the mobility in public places has been drastically reduced during the period. This trend and the drastic increase in residential mobility are strong indicators that the social distancing efforts have been efficient. However, the trade-off with this policy is that it leads people in overcrowded and compact living spaces at risk, thus providing a potential hypothesis to explain the results of the regression.



Analysis of Testing Access, and ICU availability across California





Testing Access

The data for testing access was obtained on Kaggle by the organization Coders against COVID-19. This dataset is a crowdsourced list of testing centers across the United States. There is a total of 182 centers in California in this dataset. However, only 170 centers screen for COVID-19. The question of whether there are certain groups of people that are restricted access to testing can only truly be answered at a county level. However, to get a general sense if there were any high population areas that did not have any adequate testing centers, I decided to plot the COVID testing centers on top of a population map of California. It appears that the testing centers are concentrated in high population counties with lower population counties having relatively fewer testing centers. This is logical as the more populous areas should have more resources allocated towards them.

ICU Availability

The data for ICU bed counts and COVID patients with ICU where obtained by the California Department of Public Health. To calculate the number of COVID-19 ICU patients, I decided to add the number of ICU COVID positive patients with ICU COVID Suspected Patients. Looking at the map, it is clear that Southern California ICUs are

filling fast due to COVID 19. This is probably due to the higher number of COVID cases as a percent of population. Additionally, the county with the highest Percentage is Imperial (approx. 64%), which is nearly 1.8x higher the next largest county, Kings County (approx. 37%). Following these counties is San Joaquin (approx. 31%), Orange (approx. 27%) and San Diego (approx. 26%). Los Angeles County is ranked 8th on the list with approx. 24%. Based on this data and previous findings, Imperial County and Kings County are the counties that could be most at risk. While they have been able to keep the death: cases ratio low, the low ICU availability and the high percentage of the population with the virus may be a strong indication that these counties might experience a higher death/cases ratio in the coming weeks. Therefore, more resources should be provided to these two counties to ensure that this does not happen.

Key Findings:

1. From April 15 to mid-May (approx. May-15th), the rate of overall increase is smaller than from mid-May to June-9th.
2. Even though Imperial county has a very high percentage of cases, LA county's top three presence when it comes to cases and deaths makes it the hardest hit county.
3. While Imperial and Kings counties have been able to keep the death: cases ratio low, the low ICU availability and the high percentage of the population with the virus may be a strong indication that these counties might experience a higher death/cases ratio in the coming weeks.
4. There is a strong relationship between a counties COVID-19 cases as a percentage of population and the metrics: percentage of people with housing with 10 or more units and percentage of houses with more rooms than people. This may be due in part to California's social distancing policies.