W200 - Project 2 - Summer 2020
Studying the Health and Social Impact of the COVID-19 Pandemic in California
Daniel Chow, Vishakh Pillai, Sandip Panesar

Introduction/Context

The COVID-19 pandemic has devastated both the health and economic wellbeing of global societies. Though some countries imposed early lockdown/quarantine measures, others, including the United States, took a more relaxed approach. Consequently, the virus has spread unchecked throughout the country with some states, including California, being hit harder than others. Our project therefore seeks to explore the health and social impacts of COVID-19 on California.

Hypotheses

1. We hypothesize that the spread of COVID-19 has been uneven throughout California, given the highly variable geography and population densities of the state.
2. In line with the larger body of literature on COVID-19, we hypothesize that in California, COVID-19 has affected different socioeconomic, age groups, sex and ethnicities differently.
3. We also hypothesize that, like many other places, California was initially caught off guard in terms of preparedness at the start of the pandemic, but local regions have taken necessary measures to alleviate this.

Questions to Answer

1. How has COVID-19 spread through California over time?
2. Which age groups, sexes and ethnicities have been most severely affected by COVID-19?
3. What has the response to the escalating pandemic been in terms of homeless contingencies and personal protective equipment (PPE)?

Dataset

The core datasets used were from the California Department of Public Health (https://data.ca.gov/dataset/covid-19-cases), who release daily updated COVID-19 statistics. As the datasets contained were specific to a certain parameter or demographic, there was no single ‘core’ dataset. Datasets were merged as necessary to help us answer the questions. Various datasets differed in collection start and end dates. In an effort to maintain consistency, however, we chose 7/18/2020 as the cut off date for our respective analyses.

Additional Datasets

1. COVID-19 Age demographics, COVID-19 Sex demographics, and COVID-19 Ethnicity demographics also available from the California Open Data Portal (see above link).
2. California 2018 population census data to calculate incidence, prevalence and mortality rates.
3. California geographical location dataset in SHP format for creation of choropleths.

4. California Homeless Response and PPE county information datasets

Initial Exploration and Data Preparation

Statewide Cases

The shape of the raw dataset was 7325 rows by 6 column variables. The data was arranged with all counts listed per county, per day. The ‘newcountconfirmed’ and ‘newcountdeaths’ columns both contained negative values, which were not explained in the data documentation. We therefore elected to replace all negative values with 0. The ‘county’ frame revealed 60 unique values, when there are only 58 counties in California. Further analysis revealed two instances ‘unassigned’ and ‘out of country’, that were eliminated due to having only single instances, respectively.

In order to prepare the data for creation of choropleth maps, a merge was performed with the geographical SHP dataframe. We only kept the ‘geometry’ column of the county geography dataset, which contained polygon objects to be utilized by the Geopandas package. In order to label the dataframe, a further column containing the centroid object was calculated from the ‘geometry’ column and Geopandas ‘.centroid’ method.

We utilized the raw numbers to analyze the state-wide spread of COVID-19 by grouping the main, cleaned dataframe. Epidemiological metrics (incidence, prevalence, mortality) can be useful in this context. In order to calculate the specific indices, county-wide population information was required obtained in .csv format from the 2018 census (the latest publicly accessible figures). Incidence was thus calculated by dividing the ‘newcountconfirmed’ column by the county population. Mortality was calculated by dividing ‘newcountdeaths’ with population, and prevalence was calculated by dividing ‘totalcountconfirmed’ with population.

In order to study the spread of COVID-19 through California over the study period, the master dataframe was split into 4 smaller dataframes. Time intervals chosen were 3/18 - 4/17, 4/18 - 5/17, 5/18 - 6/17, 6-18 - 7/18. These smaller dataframes were then grouped by county to calculate average metrics for each month long period. These were then utilized to create the final choropleth maps.

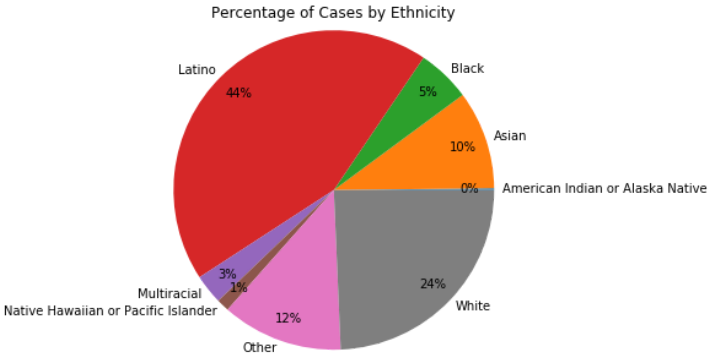
Ethnicities

The shape of the raw dataset was 776 rows by 7 column variables. There were no negative values, nor were there any missing values.

Regarding specific ethnicity groups, the raw dataset contained 10 unique ethnicities ('Latino', 'White', 'Asian', 'Black', ‘Multiracial', 'American Indian or Alaska Native', 'Native Hawaiian or Pacific Islander', 'Other', 'Multi-Race', 'Native Hawaiian and other Pacific Islander'). It became apparent that there were duplicate classifications (i.e. ‘Native Hawaiian…’) and an ‘Other’ category. The Native Hawaiian categories were collapsed into a single category.

Upon further exploration, it became apparent that several of the ethnicities were underrepresented by the data (see **Figure 1**). Even though 'other' represented 12% of the cases, it would be impossible to include this in our analysis which requires dividing the case percentages for each ethnicity by their percentage population representation. It was decided that 'American Indian/Alaska Native', 'Multiracial' and 'Native Hawaiian or Pacific Islander' ethnicities, due to very low subject counts, would be excluded from the main analysis. The highest-represented groups remaining: Asian, Latino, White and Black were kept for further analysis. Once the dataframe was sliced to only include the chosen ethnicities, the values in the 'case_percentage' column were normalized by subtracting their population proportions.

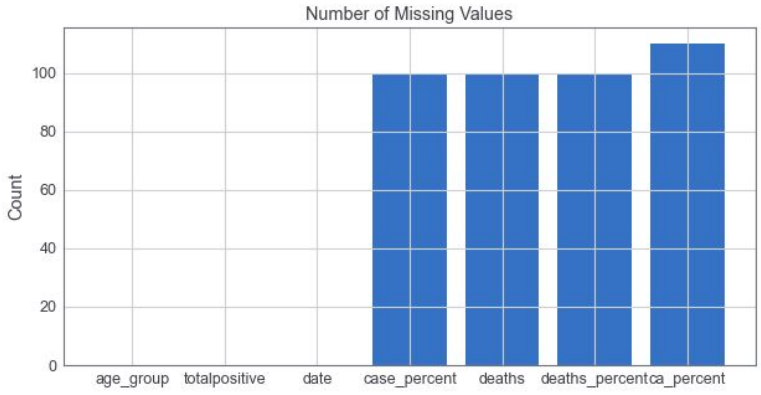
Figure 1: Percentage of Cases by Ethnicity



Age Groups

The age dataset contained 570 observations and 7 features. The dataset was well maintained and little cleaning was needed. The missing numbers shown in **Figure 2** are from the beginning of the dataset, where they had not begun recording. The NA values were dropped from the dataset.

Figure 2: Number of Missing Values in Age Groups



The dataset switched naming conventions for two age groups in the middle. We modified the naming conventions to unify this discrepancy. Finally, we removed all Missing/Unknown labels from the dataset because no information about their age was provided.

Sex Demographics
The raw dataset was relatively small; the shape was 328 rows by 7 columns. There were instances in all columns from 2020-04-02 till 2020-04-21 of 'NaN'. It can be assumed that prior to 2020-04-22, data was simply just not collected for this specific demographic. This represented roughly 20% of the dataset. These rows were omitted from the final plots.

The column 'ca_percent' appears to represent the population information of 'Male' and 'Female', and 'Unknown' demographics. A sanity check showed that these data points did not change significantly over time, and normalization based on sex populations was not necessary. We were unclear about the meaning of 'totalpositive2', so it was not used. Likewise, we the significance of the 'Unknown' sex category was not identified and it was dropped.

Homelessness
The shape of this raw dataset was 4833 rows by 7 columns. There were no 'NaN' values in the original dataset. A sanity check showed that the 'rooms_occupied' parameters never exceeded the 'rooms' parameter.

The trailer data and rooms data represent California Gov. Newsom's 'Project RoomKey', a state government effort in which over 15,000 motel rooms, and 1,300 trailers were serviced to extremely vulnerable individuals experiencing homelessness during this pandemic. All parameters have non-uniform distributions, which may represent that every county may be considerably different in how they followed this project.

PPE
The shape of this raw dataset was 795,968 rows by 5 columns. A sanity check showed that county names repeat in this dataset with each time data point. There were instances of 'NaN' interspersed throughout the dataset, all of which were disregarded.

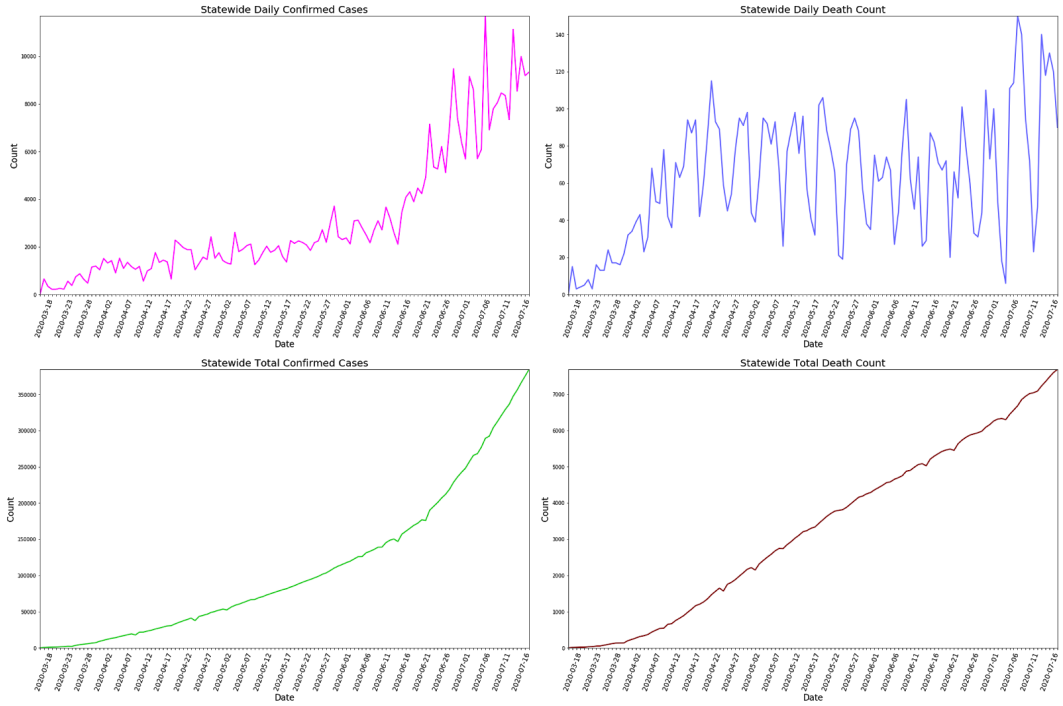
To note, the 'quantity filled' parameter, according to the CDPH website, represents the quantity of product sent to the warehouse or agency for fulfillment. The column 'shipping_zip_postal_code' was not necessary for any of our analyses, so it was dropped. There were close to 27 entities in the 'product_family' column. The categories pertinent to our study were: 'N-95 Respirators', 'Surgical Masks', 'Surgical or Examination Gowns', and 'Examination Gloves'.

Data Stories, Text, and Figures

Statewide Cases
In concordance with data released from official public health bodies, our analysis shows that for the state of California, the number of new daily cases, total confirmed cases and the total death count continued to increase. The rate of increase for both the new daily cases and total confirmed cases began accelerating around Mid-June (**Figure 5**). Interestingly, the state-wide daily death count demonstrated an initial period of increase until around Mid-April, before

assuming a widely fluctuating ‘sawtooth’ pattern, the oscillations of which have continued to increase in magnitude.

Figure 5: Numbers of Statewide Cases



We then moved on to analyzing the spread of COVID-19 by county. The averaged daily incidence, mortality and prevalence rates for each county increased gradually between months 1-3. However, beginning month 3, the rate of all three metrics accelerated. Due to large size, the choropleth maps demonstrating county spread can be found here:

https://drive.google.com/file/d/1NDAXpKG2HA3NGbKrdfzficARubKU7up-/view?usp=sharing

Table 1: Incidence, Mortality and Prevalence by Month

Month	Incidence	Mortality	Prevalence
1	0.0000142	0.0000004	0.0001873
2	0.0000190	0.0000006	0.0006833
3	0.0000396	0.0000006	0.0015333
4	0.0001320	0.0000010	0.0040050

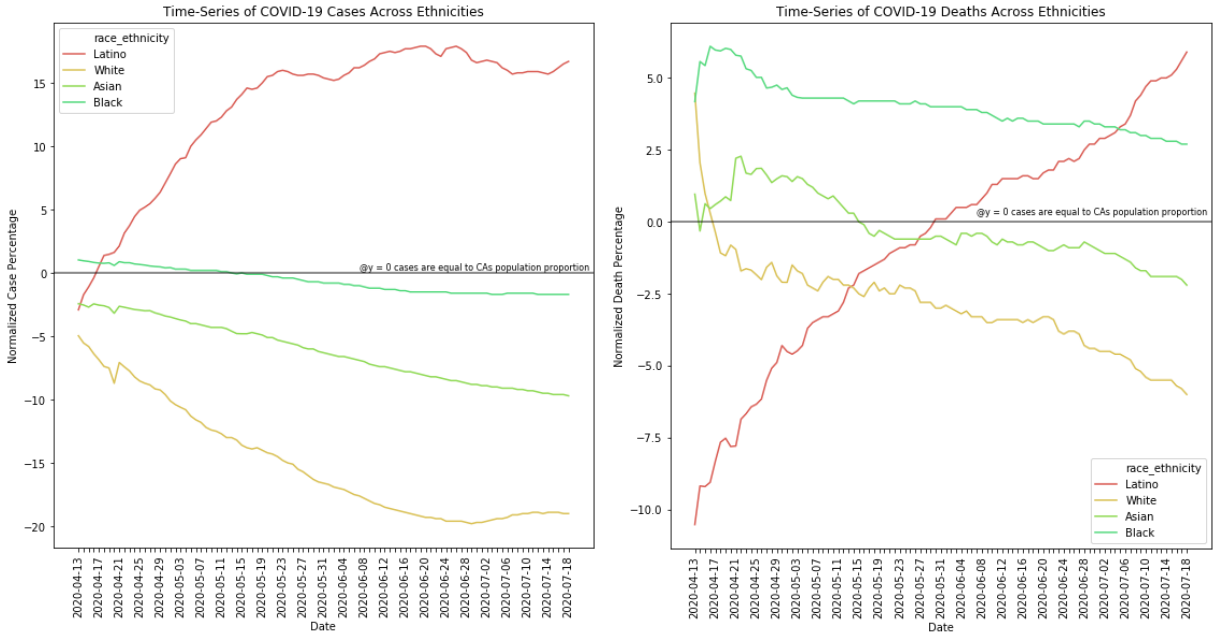
It was clear from these analyses that the overall rates of all 3 metrics increased over the study period. The virus has established a foothold in California with the most populous counties in the

south of the State being hit much harder than those in the north. Nevertheless, even in the north, San Francisco and its surrounding counties have also been particularly severely affected by the virus. Though California was one of the first states to implement shelter-in-place measures on March 19th, these were relaxed towards the end of April. This relaxation may be responsible for the discernable uptick in case metrics observed between months 3 and 4.

Ethnicities

It became apparent during our initial analysis that there was ethnic variation in the impact of COVID-19 in California, the effect of which was further studied by subtracting the particular ethnicities statewide population percentage from the daily case values. Our analyses **(Figure 6a, b)** show that though cases continue to rise among Latinos, they are declining among the other ethnic groups. Moreover, blacks continue to die at rates above their normalized population level with little relative change, and deaths among Latinos continue to increase. Deaths among Asians and whites have decreased to below-population proportions. This evidence is in line with that from major health organizations, pointing to an ethnic variation in COVID-19 impact.

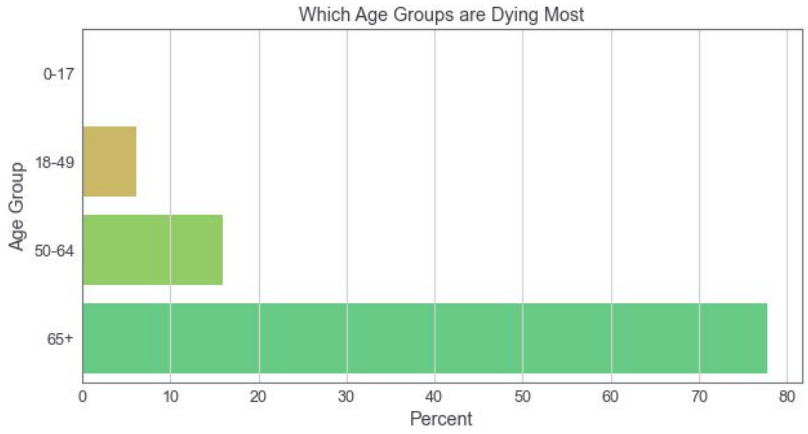
Figure 6a, b: Time series of COVID-19 Cases; Deaths



Age groups

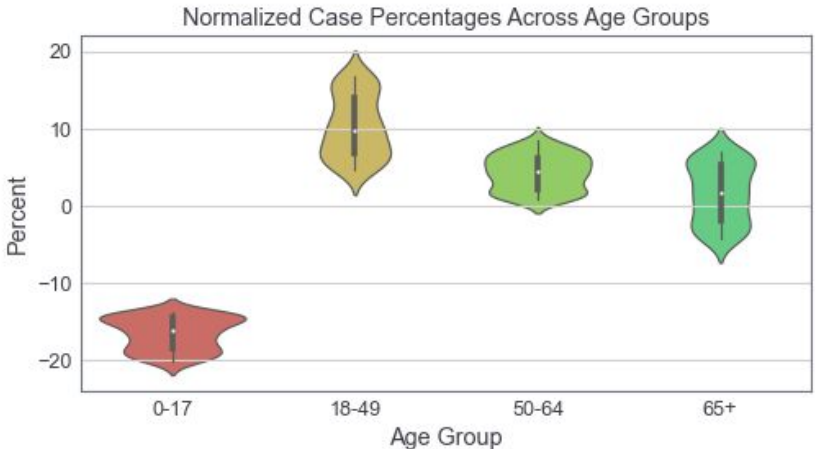
Figure 7 below shows that almost 4 out of 5 COVID-19 deaths have occurred in people in the high-risk 65 and older age group. This confirms the idea that age is one of the major factors in determining the lethality of COVID-19.

Figure 7: Percent of Deaths by Age Groups



The next question is the infection rate of COVID-19 across various age groups. In California, the 65+ age group represents around 15% of the population. **Figure 8** below shows the under or over representation of a specific age group. The graph is normalized for perfect representation of an age group at 0.

Figure 8: Case Percentages Since Shelter in Place



The two tallest distributions are the 18-49 and 65+ age groups. **Figure 9** shows the change in representation of COVID-19 in the various age groups over time.

Figure 9:

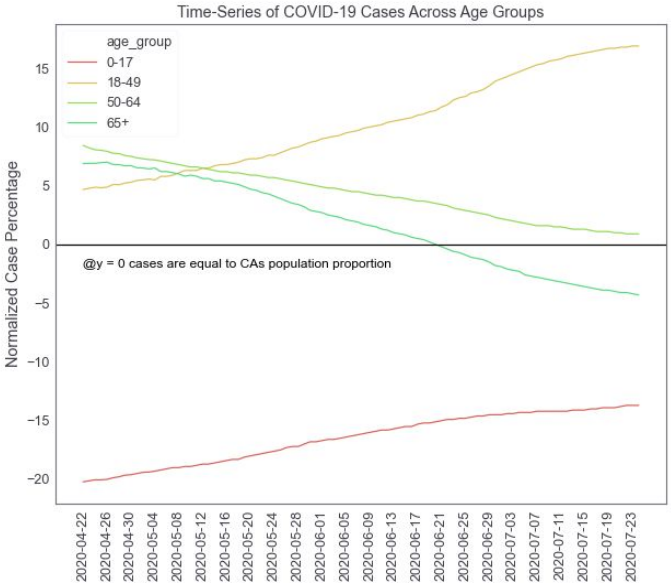


Figure 9 demonstrates that the 65 and older age group was one of the hardest hit in the beginning. However, this decreased by mid-late July to eventually become one of two groups underrepresented in their age category. The second highlight is that the 18-49 age group has increased as much as the 65+ age group decreased.

Sex demographics

No normalization was performed on the initial sex dataset as differences between the two populations were negligible. As demonstrated in **Figure 10**, case and death percentages were plotted with respect to time from May - July of 2020. There was no substantial difference between males and females in terms of cases.

Figure 10a, b: Cases; Deaths Percentages for Males/Females vs. Time

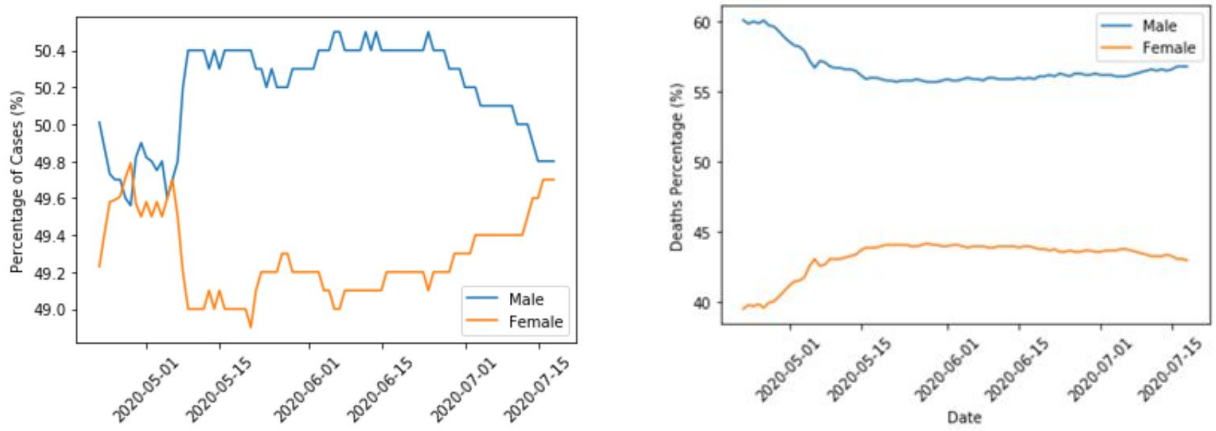
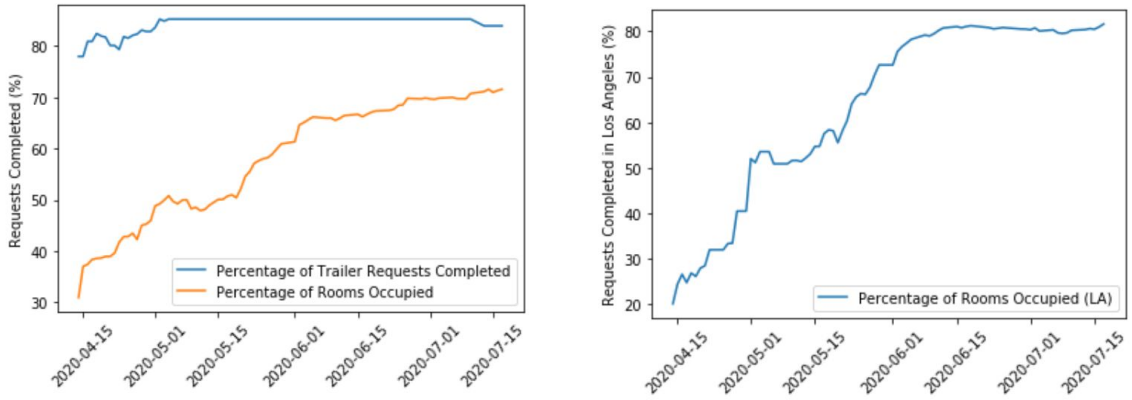


Figure 10b demonstrates a 10-15% difference in deaths between males and females. This may be due to biological factors and warrants further investigation.

Homelessness

In response to the crisis California Gov. Newsom created *Project Roomkey*, aimed at helping the homeless. According to **Figure 11a, b**, the project has been particularly successful with trailers. There was initially a strong uptake of hotel room occupancy from the middle of April until June. This plateaued from June to July. The slower rate may be aligned with reports that federal agencies have been struggling to find permanent solutions for the homeless after the end of the project.

Figure 11a, b: *Project Roomkey* requests in California; Los Angeles county

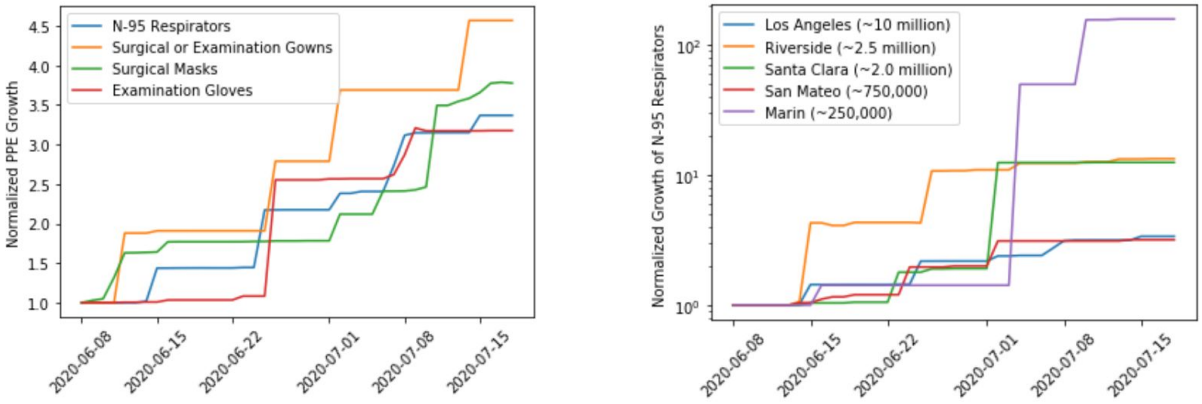


Los Angeles has one of the highest populations of homelessness in the entire country. From room occupancy data seen in **Figure 11b**, *Project Roomkey* has been mostly unsuccessful throughout the pandemic with only a rise of 10%. This could be because many hotels chose not to participate in the program, despite public subsidies as an incentive.

PPE

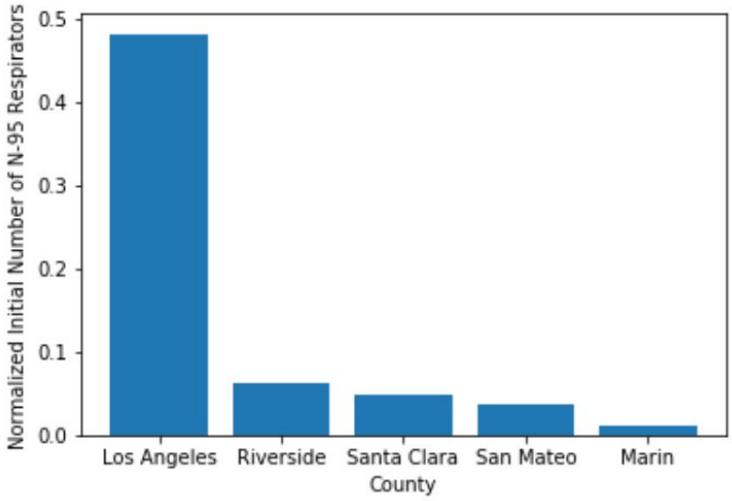
Both charts in **Figure 12a, b** are normalized by the amount of PPE in that county at the start of the pandemic. **Figure 12a** demonstrates that from June until July 2020, there was increased procurement of respirators, gowns, surgical masks, and gloves. This response may have been reactionary to the surge in cases in California over this period. Compared to gowns, there were fewer respirators ordered, potentially due to cost or supply chain issues.

Figure 12a, b: PPE in California; N-95 Respirators by county over time



We elected to study the PPE procurement by a selection of counties from June to July (**Figure 12b)**. It was apparent that the greatest growth of N-95 Respirators occurred in the least populous county (Marin), while the smallest growth occurred in a relatively more populous county (Los Angeles). For medium-sized populations, Santa Clara and Riverside both had similar final growths to Los Angeles.

Figure 13: Normalized Initial Amount of N-95 Respirators by county



From **Figure 13** it is clear that Los Angeles was far more prepared than any other county at the start of June. At this time, Marin county appeared to be far less prepared than San Mateo, which can explain the both large increase seen in Marin and the steady increase for San Mateo.

Conclusions

Our analyses of publicly-available COVID-19 data has demonstrated the widespread, yet varied toll of this devastating virus in California. It is clear that there are socioeconomic, geographical and demographic factors affecting in its spread and impact. This knowledge may potentially be exploited to devise public health response strategies.