STA 160 Group 7 – Final Project

Rishi Bhuva,[1] Aditya Kallepalli,[2] Vishnu Rangiah,[3] and Ivan Yang [4]

[1] *Student ID: 915218518, Email: rdbhuva@ucdavis.edu*

[2] *Student ID: 915079375, Email: arkallepalli@ucdavis.edu*

[3] *Student ID: 916562849, Email: vrrangiah@ucdavis.edu*

[4] *Student ID: 915463046, Email: igyang@ucdavis.edu*

(we will contact the first author for any question about the article)

In this report, our group performed data analysis on the CDC's available data on COVID-19 patients in the US. We specifically looked at deaths and healthcare worker status at a national level, as well as deeper analysis of other variables specifically in California. We also built two models predicting the probability of death as a COVID patient in California. (59 words)

I. INTRODUCTION

We will be analyzing the United States Centers for Disease Control's (CDC) "COVID-19 Case Surveillance Restricted Access Detailed Data".

The CDC is the national public health agency of the United States. This data set contains national patient information ranging from January 1, 2020 to April 26, 2021 with a myriad of patient information consolidated from state and county health department reports. In Section 1, we looked at the frequency of COVID-19 patients' sex, age group, death, healthcare worker status relative to county on a national and state level.

While doing exploratory data analysis, we realized that there were many entries with missingness. Due to the United States' unique separation of powers between federal and state governments, local governments are able to withhold a lot of their patient health information from the federal government due to privacy reasons, as stated on this dataset's accompanying data dictionary. States such as Washington and Kansas may have entire variable columns such as healthcare worker status with no data. Additionally, data collection based on self-reporting results in a lot of inherent missingness of data points. Therefore, this dataset contains a lot of "NA" values indexed under "unknown", "missing", or "none".

Due to the missingness mentioned above as well as many other confounding factors not accounted for in this dataset, we have decided to focus our machine learning portion of data analysis (Section 2) on a particular state. We chose California due to it being the state with the largest subset of complete patient data, as well as our familiarity of California as Californian residents. We constructed linear and hierarchical clustering models. Due to the missingness mentioned above as well as many other confounding factors not accounted for in this dataset, we have decided to focus our machine learning portion of data analysis (Section 2) on a particular state. We chose California due to it being the state with the largest subset of complete patient data, as well as our familiarity of

California as Californian residents. Apart from our CDC dataset, we also got additional data of different factors from external sources. We wanted to look at how political leaning, population per county, and air pollution also affected probability of death by COVID-19. We constructed a linear regression model and a hierarchical geospatial clustering model.

We will be performing our analysis using Python with the packages *pyspark, pandas, numpy, plotly, lasso, sklearn, geopandas, folium,* and *seaborn* to conduct exploratory data analysis and the construction of our machine learning models.

II. EXPLORATORY DATA ANALYSIS

A. Main Data Frames

	abdom_yn	abxchest_yn	acuterespdistress_yn	age_group	cdc_case_earliest_dt	chills_yn	county_fips_code	cough_yn	current_status	death_yn	diarrhea_yn	fever_yn	sfever_yn	hc_work_yn	headache_yn	hosp_yn	icu_yn	mechvent_yn	medcond_yn	myalgia_yn	nauseavomit_yn	pna_yn	race_ethnicity_combined	res_county	res_state	runnose_yn	sex	sob_yn	sthroat_yn
0	Missing	Missing	Missing	10 - 19 Years	2021-02-06	Missing	45001	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	ABBEVILLE	SC	Missing	Female	Missing	Missing
1	Missing	Missing	Missing	10 - 19 Years	2021-01-28	Missing	45001	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	ABBEVILLE	SC	Missing	Female	Missing	Missing
2	Missing	Missing	Missing	10 - 19 Years	2021-01-19	Missing	45001	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	ABBEVILLE	SC	Missing	Female	Missing	Missing
3	Missing	Missing	Missing	10 - 19 Years	2021-01-07	Missing	45001	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	ABBEVILLE	SC	Missing	Female	Missing	Missing
4	Missing	Missing	Missing	10 - 19 Years	2020-07-20	Missing	45001	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	Missing	Missing	Missing	Missing	Missing	Missing	Unknown	ABBEVILLE	SC	Missing	Female	Missing	Missing

FIG. 1. (Main Data Frame – National).

Here we are given 2 Apache Parquet files with all our data points. We first used pyspark and pandas to convert them into dataframes. We then merged them in to one dataframe (Figure 1). Our main dataframe consists of 25,120,922 entries and 29 variables. They are: current_status, cdc_report_dt, cdc_case_earliest_dt, sex, age_group, race_ethnicity_combined, county_fips_code, res_county, res_state, onset_dt, pos_spec_dt, hosp_yn, icu_yn, death_yn, hc_work_yn, pna_yn, abxchest_yn, acuterespdistress_yn, mechvent_yn, fever_yn, sfever_yn, chills_yn, myalgia_yn, runnose_yn, sthroat_yn, cough_yn, sob_yn, nauseavomit_yn, headache_yn, abdom_yn, diarrhea_yn, and medcond_yn. Rudimentary analysis shows that these entries are records of COVID-19 patients in the USA between January 1st, 2020 and April 26th, 2021.

	abdom_yn	abxchest_yn	acuterespdistress_yn	age_group	cdc_case_earliest_dt	cdc_report_dt	chills_yn	county_fips_code	cough_yn	current_status	death_yn	diarrhea_yn	fever_yn	sfever_yn	hc_work_yn	headache_yn	hosp_yn	icu_yn	mechvent_yn	medcond_yn	myalgia_yn	nauseavomit_yn	onset_dt	pna_yn	pos_spec_dt	race_ethnicity_combined	res_county	res_state	runnose_yn	sex	sob_yn	sthroat_yn
7062	Missing	Missing	Missing	0 - 9 Years	2020-10-05	None	Missing	06001	Missing	Laboratory-confirmed case	No	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	None	Missing	None	American Indian/Alaska Native, Non-Hispanic	ALAMEDA	CA	Missing	Female	Missing	Missing
7063	Missing	Missing	Missing	0 - 9 Years	2020-07-27	None	Missing	06001	Missing	Laboratory-confirmed case	No	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	None	Missing	None	American Indian/Alaska Native, Non-Hispanic	ALAMEDA	CA	Missing	Female	Missing	Missing
7064	Missing	Missing	Missing	0 - 9 Years	2020-12-22	None	Missing	06001	Missing	Laboratory-confirmed case	No	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	None	Missing	None	American Indian/Alaska Native, Non-Hispanic	ALAMEDA	CA	Missing	Female	Missing	Missing
7065	Missing	Missing	Missing	0 - 9 Years	2021-04-03	None	Missing	06001	Missing	Laboratory-confirmed case	No	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	None	Missing	None	American Indian/Alaska Native, Non-Hispanic	ALAMEDA	CA	Missing	Female	Missing	Missing
7066	Missing	Missing	Missing	0 - 9 Years	2020-11-13	None	Missing	06001	Missing	Laboratory-confirmed case	No	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	None	Missing	None	American Indian/Alaska Native, Non-Hispanic	ALAMEDA	CA	Missing	Female	Missing	Missing

FIG. 2. (Main Data Frame – California).

We also subsetted a dataframe specifically for our data analysis machine learning on Californian patient data. We generated this subset just by setting res_state as "CA".

B. Data Cleaning

Our first line of action when analyzing this data is to properly clean the dataset. Having clean data will provide us with the highest quality of information needed and therefore will provide us with the most accurate predictions and correlations. Our process for cleaning the dataset can be seen below.

1. Removing Missing and Irrelevant Information

As mentioned in our introduction, this dataset has a lot of missing values recorded as "Missing", "Unknown", or "None" due to the data's self-reporting nature and privacy restrictions. We will first remove the missing values under res_state and county_fips_code as our data analysis is heavily location-related and entries with no location data is not useful to us. We also drop rows with NA values in onset_dt, pos_spec_dt, and cdc_report_dt for better quality data. We end up with 24,362,135 rows. Throughout our EDA, we also clean out missing data points within each subsetted dataframe necessary for generating our plots.

III. GRAPHS

We will plot some graphs with our tidied dataset to gain further insight.

A. National COVID-19 Cases

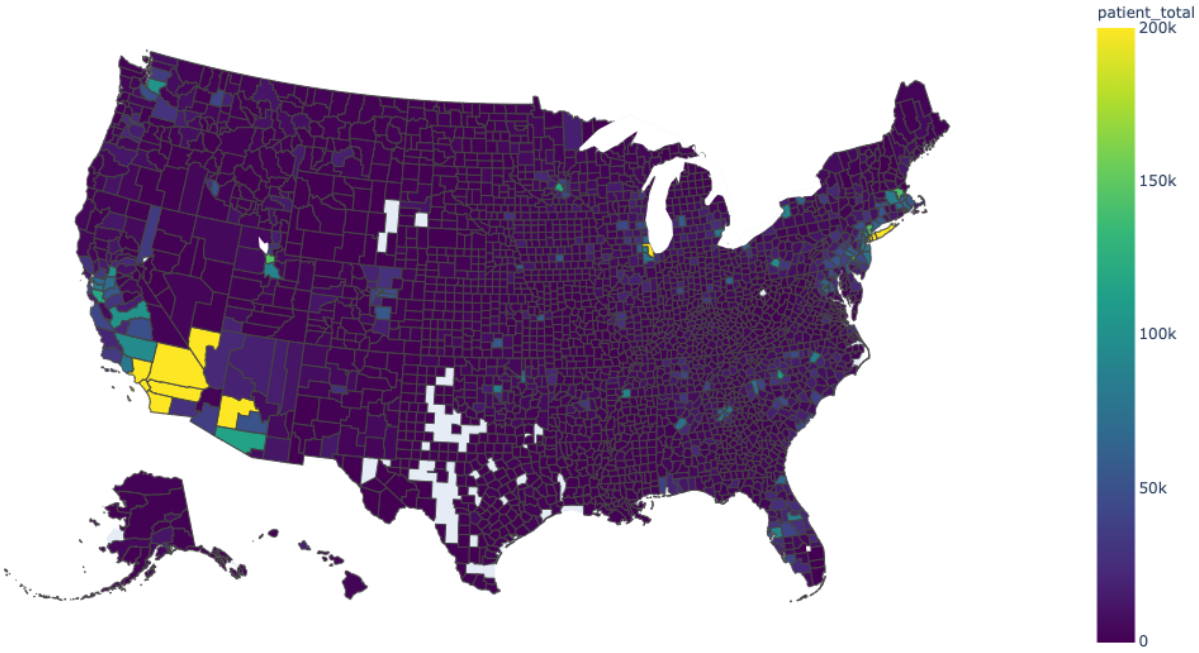


FIG. 3. (Total Number of COVID-29 Cases by County).

Most counties reported at least 1 COVID-19 patient to the CDC, with exceptions mostly in Texas and Wyoming (grey areas in the plot). It is unclear whether these were unreported due to privacy concerns or a non-existence of COVID cases in these counties. We can see that the Southwestern region of the US in Southern California and Arizona had the most COVID-19 patients, with Los Angeles County in California amassing a whopping 1,128,065 patients, which is twice more than Cook County, Illinois, the county with the second most COVID-19 patients.

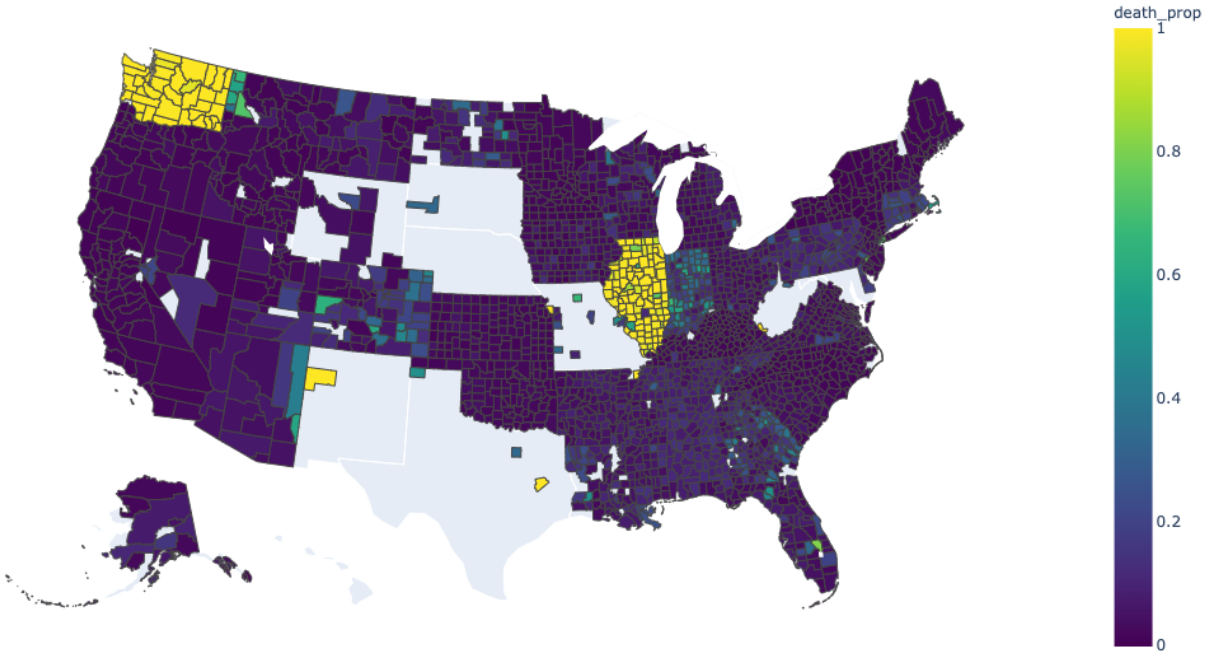


FIG. 4. (Proportion of COVID-19 Deaths Among All Patients Per County).

Figure 4 shows that across most counties, 10% or lower of the COVID-19 patients with known death status died. This density map is a great example of the problem of selective reporting of COVID data. For example, states like Washington and Illinois that are almost completely yellow does not equate to complete death of patients but rather an interesting fact that these counties only publicly shared the death status of dead patients. As seen in the next figure, the death proportion compared to all patients in the county, including ones with unknown death statuses, are all quite low. We can also see that almost all counties in states such as Kansas, Texas, South Dakota, and New

Mexico did not report the death status of their patients. Due to this discrepancy generating low quality results, we will generate a new map with the newly calculated proportions of deaths with all patients per county, including patients with unknown death statuses.

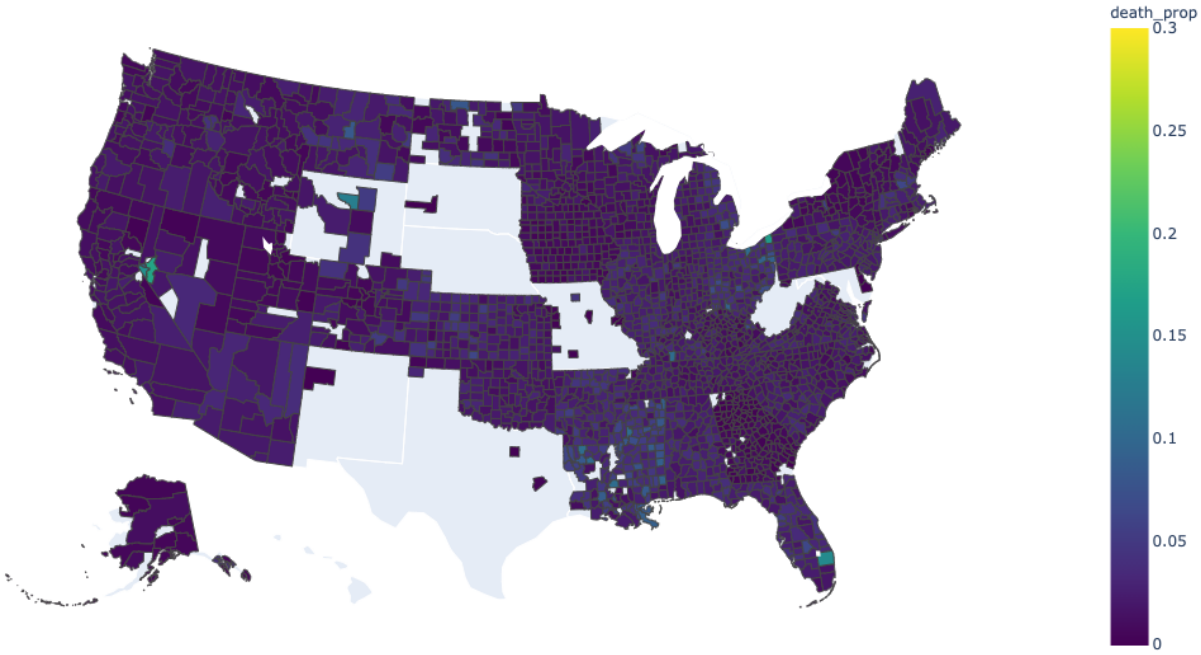


FIG. 5. (Proportion of COVID-19 Deaths Among All Patients Per County).

Figure 5 shows the proportion of deaths of COVID-19 patients relative to all COVID-19 patients in each county. We can see that most counties have a pretty low death proportion of below 5%. The highest death proportion reflected in this map seems to be in Nevada with a little bit lower than 20%. There also seems to be a higher death proportion around the Mississippi River Delta and Mississippi, as well as the counties near Lake Erie in Ohio.

Although our analysis shows that the maximum death proportion per county is around 29%, we do not observe any yellow-colored counties in this density map, and the highest death proportion reflected in this map seems to be in Nevada with less than 20%. Further analysis shows that the 14 counties with highest death proportions are all in outlying US territories not included in this map, where 13 of them are in Puerto Rico and 1 is in the Northern Mariana Islands. The county with the

highest proportion of deaths reflected in this map is Lyon County, Nevada with around 18%, which is only the 15th highest among our entire dataset of known deaths. This is a high indicator of resource bias for fighting COVID-19 when comparing unincorporated US territories with the mainland.

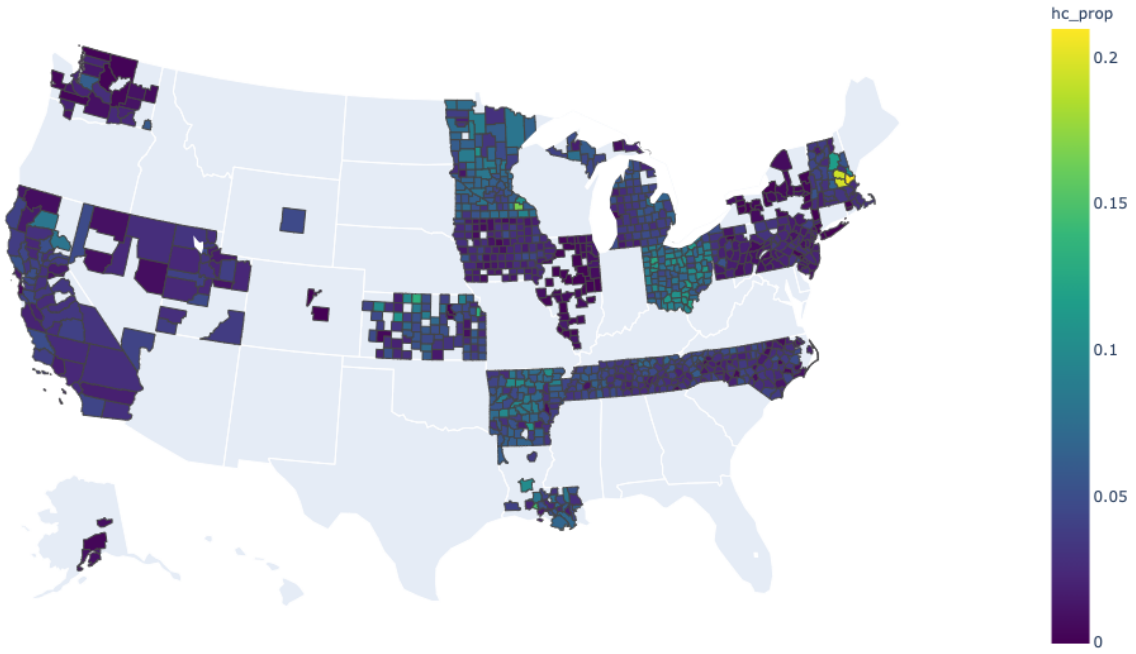
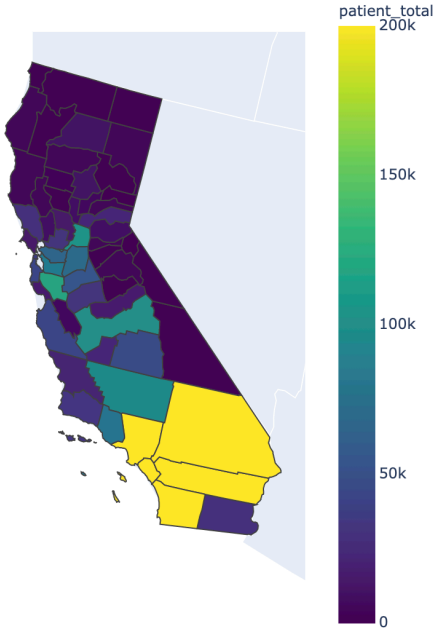


FIG. 6. (Proportion of COVID-19 Patients Who Were Healthcare Workers).

According to the map plot above, we can see the proportion of COVID-19 patients who were healthcare workers. Many states did not disclose their patient's healthcare worker status. Different states seem to have different abilities in keeping their healthcare workers safe. States such as California, New York, and Washington have low healthcare worker COVID contraction rates, while states such as Minnesota, Arkansas, Ohio, and New Hampshire had high healthcare worker patient proportions, with Merrimack, Hillsborough, and Rockingham Counties in New Hampshire having between 19.5% to 20.8% patients who were healthcare workers.

As seen in our above 4 graphs, there are many states that don't have much data. Due to such large missingness as well as technological difficulties wrangling more than 24 million entries, we have decided to focus the rest of our project's data analysis specifically on California.

B. Californian COVID-19 Cases



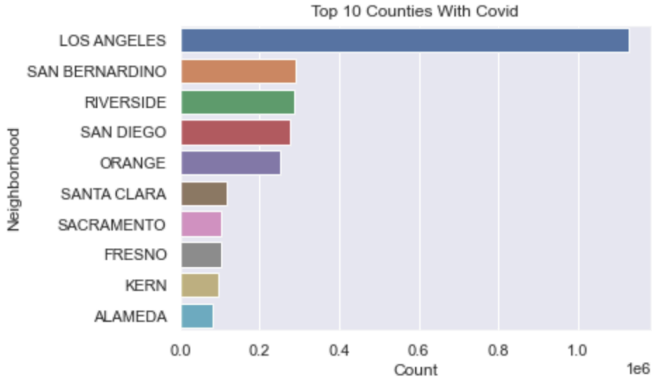


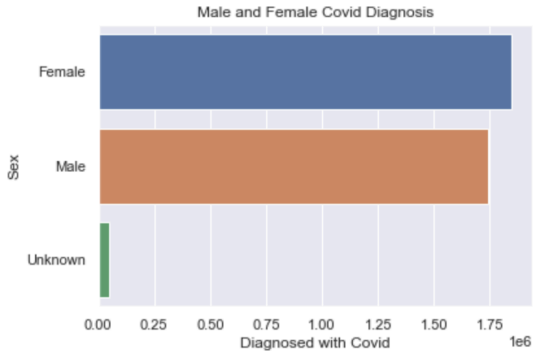
FIG. 7. (Density Map of Total Number of COVID Cases by County).

FIG. 8. (Bar Chart of Total Number of COVID Cases by County).

As already noted above, the Los Angeles region has some of the highest number of COVID-19 patients per county. This would make sense as Los Angeles County has the largest county population in California We can also see a clear difference between Los Angeles and the rest of the counties. San Bernardino, Riverside, San Diego and Orange county had similar numbers to one another but none nearly as close to Los Angeles.

We can also clearly see that Northern California has much less total COVID-19 patients and that there might be a general increasing trend from north to south of California. As Northern

California was earlier in imposing COVID-19 related mandates, we can also make an inference that they have some sort of correlation with the number of cases.



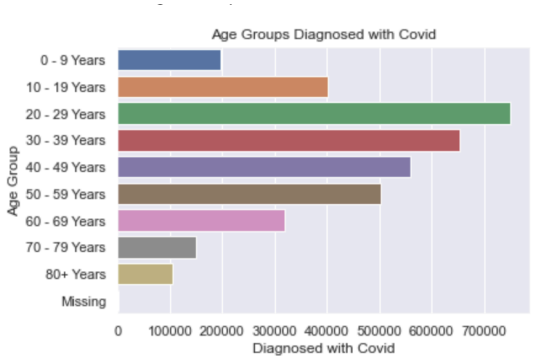
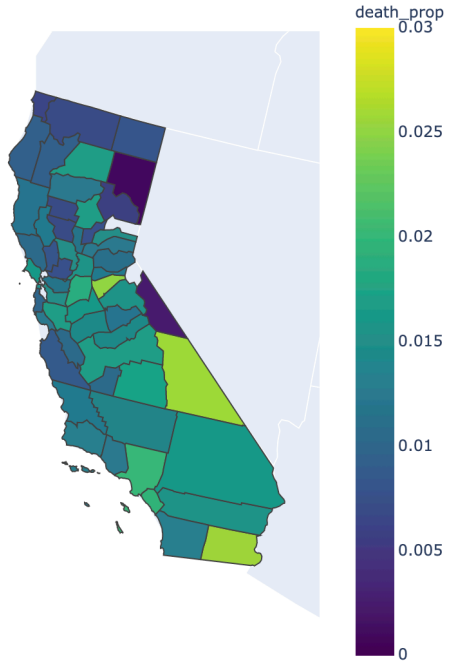


FIG. 9. (COVID-19 Patients Based on Sex). FIG. 10. (COVID-19 Patients Based on Age Group)

Another interesting feature to this dataset was the sex of the individual who was diagnosed with COVID-19. As we mainly dealt with data only regarding those of California, here we can see the difference of sex to those who pertained in California. As we can see, more females were diagnosed rather than males. Along with this, we can also see that a good percentage of values were unknown. This could possibly skew the data that we see regarding males and females, although it seems quite small and negligible. The difference between male and female does seem quite small, although as our scale is in terms of a million, this means that the difference between female and males is quite large.

In Figure 10, we can see the difference in age groups that were diagnosed with COVID-19 in California. As we can see, individuals from the age group 20-29 were the most prominent to be diagnosed with COVID-19. This makes sense as individuals within those age groups would be the ones to leave the house more and not follow proper pandemic precautions. Along with this, we surprisingly have a good amount of individuals from the age group 0-9 who were diagnosed with COVID. Lastly, we can also see the smallest number of people who were diagnosed ranged in the

80+ age group. This also does make sense as we would expect those who are 80+ to properly follow guidelines as this disease could be very detrimental to their health.



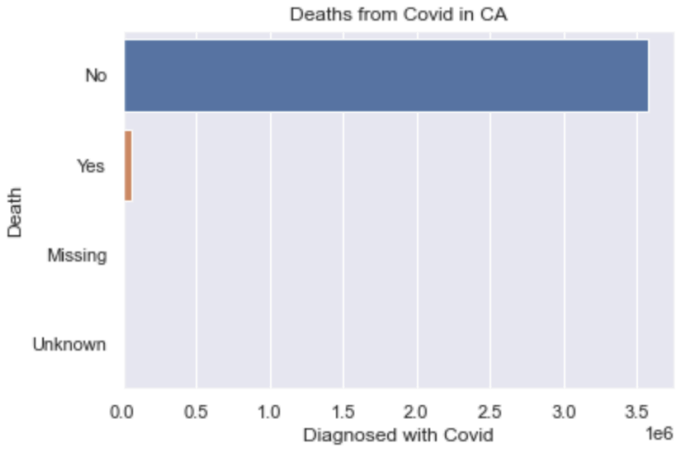
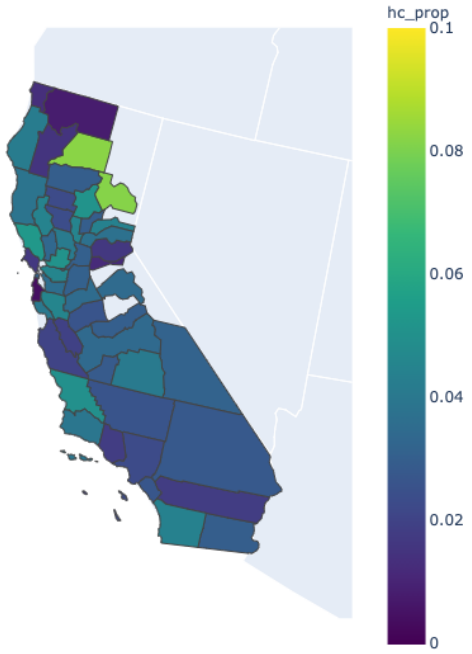


FIG. 11. (Proportion of COVID-19 Deaths Among All Patients Per County).

FIG. 12. (Number of COVID-19 Patient Deaths in California).

In Figure 12, we can see a count plot regarding the number of deaths specifically in California. As seen from the plot, the majority of individuals who were diagnosed with COVID-19 did not seem to die. Although, as the X-axis is in terms of 1e6, meaning in terms of one million, we can see that in general, lots of people in California have gotten affected by this pandemic. This also shows that, prior to what it seems in the plot above, many have died in California. Fortunately, we can see that there is a massive turnout of less deaths than more deaths. Along with this, we can see data regarding the 'Missing' and 'Unknown' values. As discussed earlier, due to the political platform of the United States and local government, is why these values are within our dataset. It can be seen that not much of death data in California was deemed missing or unknown, although there was a minor amount.

Around 5%-20% of all COVID patients died per county in California. It seems like the death proportion increases the more south the county is, with Calaveras County being an exception. Further research shows that Calaveras County experienced a COVID surge during the fall and winter of 2020. We are also missing data from Sierra County and Alpine County.



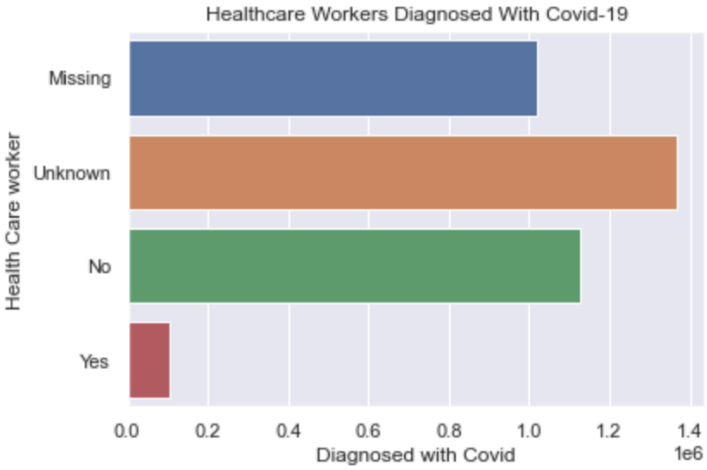


FIG. 13. (Proportion of COVID-19 Patients Who Were Healthcare Workers).

FIG. 14. (Count Plot of COVID-19 Patients Who Were Healthcare Workers.).

We are missing data for 7 Californian counties. As seen in Figure 14, the majority of our data tends to have a lot of missing and unknown values. This makes logical sense, as the identity of a healthcare worker would naturally be more susceptible to being anonymous by the government. Along with that those with known healthcare worker status generally did not get diagnosed. We believe that simply due to the high precision of PPE, a majority of healthcare workers combated COVID in their designated workplace.

Looking deeper into the proportions per county, most counties' patients were 2%-5% healthcare workers, with Shasta and Plumas counties having the highest healthcare worker

proportions of over 8%. Apart from Siskiyou and Amador counties, it seems like there is a higher proportion of healthcare worker patients in less densely populated counties.

IV. MODELLING COVID-19 PATIENT DATA IN CALIFORNIA

A. Lasso and Ridge Model Predicting Probability of Death

We conducted a Lasso test to predict the probability of death from COVID-19 nationwide. We included all 20 symptom parameters, sex, race_ethnicity_combined, and age_groups and converted all variables to dummy variables. We cleaned up our main nationwide data frame and removed all entries with any null value and were left with 24,847 patients. We used the 80-20 method to create the model. We end up with a Lasso and Ridge model and obtained a higher Ridge score of 0.963, which is only 0.11% more accurate than the Lasso model. Based on our analysis, we posit that we can very accurately predict any patient's fate of death in the US if provided with these symptoms, age group, sex, and ethnicity.

B. Geospatial Clustering Model

We want to examine the effects of COVID-19 and different factors on the counties in California. We want to adjust the counties for population, pm2.5 values. We aim to cluster the counties using the K-means algorithm and identify any relationships between the clusters.

We have used multiple sources of information to obtain the above dataframe. We have chosen to include pm2.5 values as part of our study because COVID-19 primarily affects a person's lungs and a study conducted by Harvard researchers we determined that pm2.5 air values may have any impact on the covid-19 case numbers per county.

We also included the political leanings of each county in California where in the dataframe '0' represented if a particular County had a majority of Democratic registered voters and '1' represented if a County had a Republican voter majority. We chose to include this feature in our dataset since it is apparent that people of different political parties chose to adhere and follow COVID-19 guidelines differently. Such as the use of masks in public areas, which has an impact of COVID-19 numbers.

	state	county	lat	long	pm2.5	cases	deaths	pop	polit
0	CA	Alameda	37.647139	-121.912488	10.115767	89204.0	1822.0	1662323	0
1	CA	Butte	39.665336	-121.603209	11.821581	12583.0	218.0	212744	1
2	CA	Calaveras	38.191068	-120.554107	10.667939	2311.0	62.0	46308	1
3	CA	Colusa	39.177739	-122.237563	10.895770	2334.0	18.0	21558	1
4	CA	Contra Costa	37.919479	-121.951543	10.022710	70041.0	845.0	1152333	0

FIG. 15. (Main Dataframe – California with all above mentioned external factors).

I. Exploratory Data Analysis of PM2.5 and County Population Data

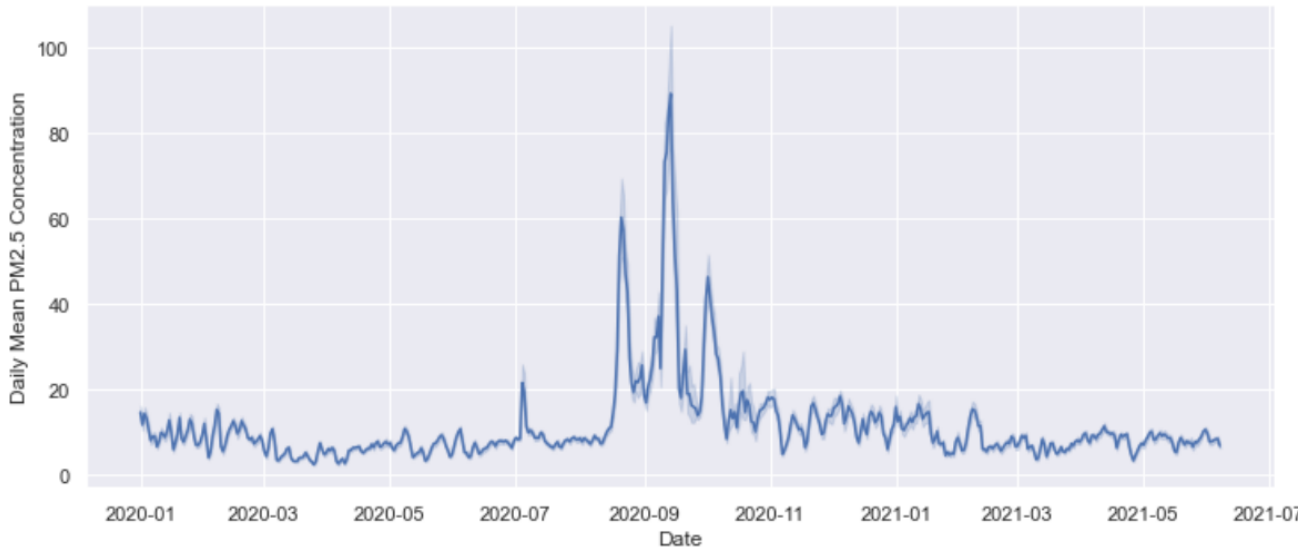


FIG. 16. (Time Series Plot of the Mean PM2.5 Air Concentration Over All Counties in CA)

In Figure __, we can see that there is a spike in the later months of California when the major wildfires took place.

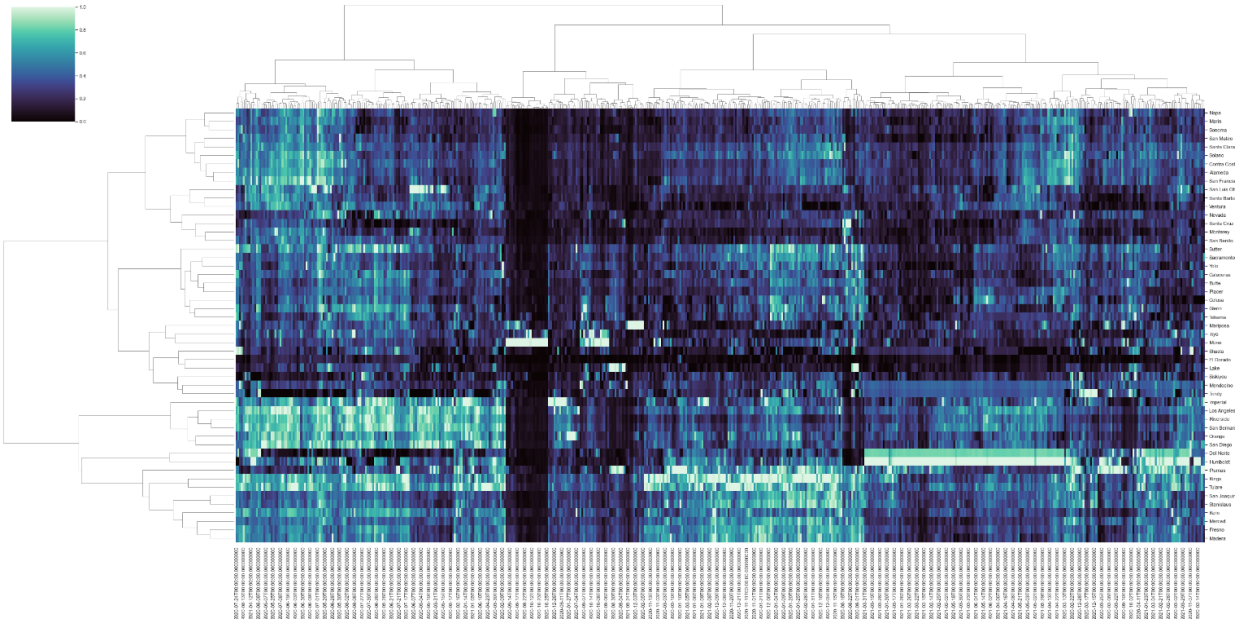


FIG. 17. (PM2.5 Readings Clustered By Date and County in CA.)

We find that southern counties show higher pm2.5 values. The values were normalized by date

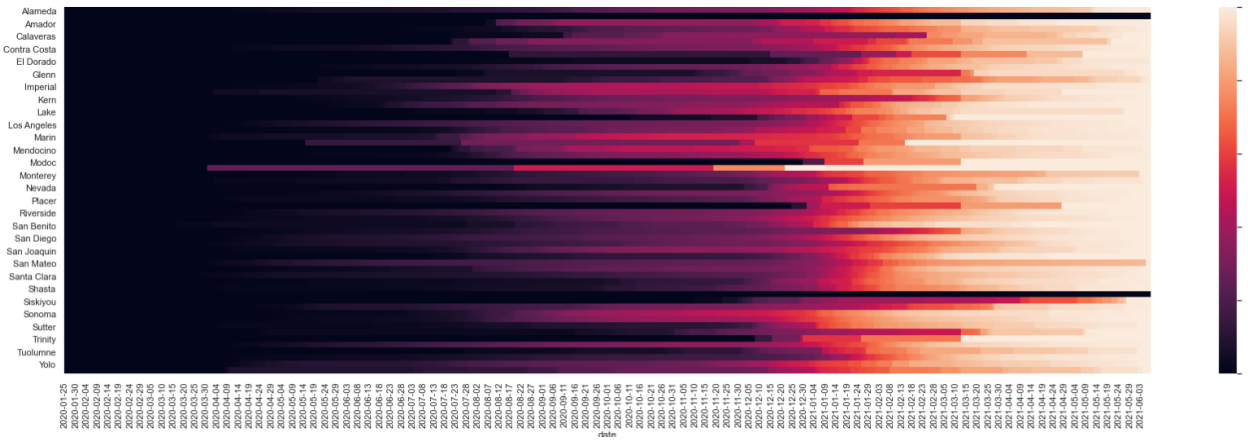


FIG. 18. (Cumulative Sums of COVID-19 Deaths Over Time)

The graph shows cumulative sums of deaths, the cumulative sums were normalized per county to clearly show the change in death counts over time. The numbers were obtained from the New York Time Repository on County Level Covid-19 information.

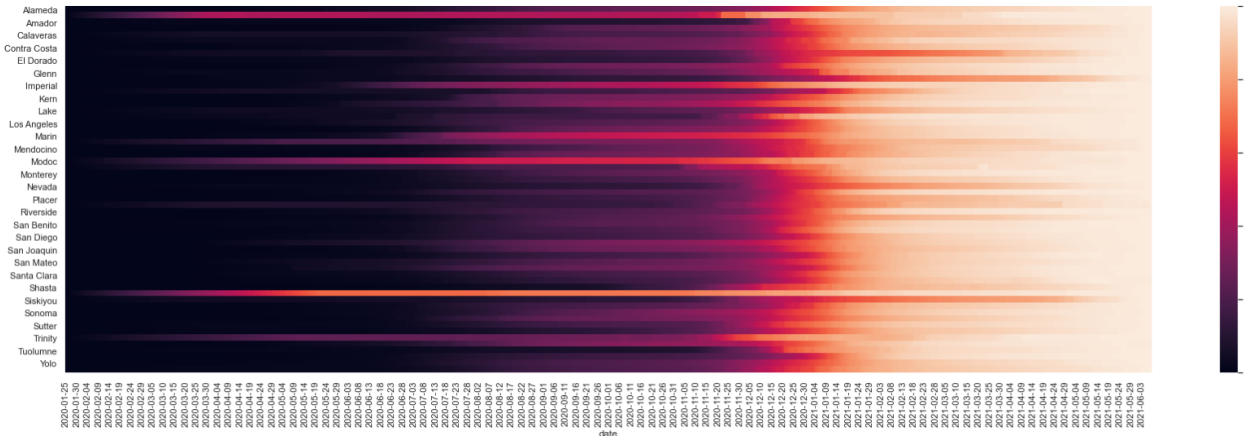


FIG. 19. (Cumulative Sums of COVID-19 New Cases Over Time.)

The graph shows cumulative sums of new cases, the cumulative sums were normalized per county to clearly show the change in new case counts over time.

II. Model Result

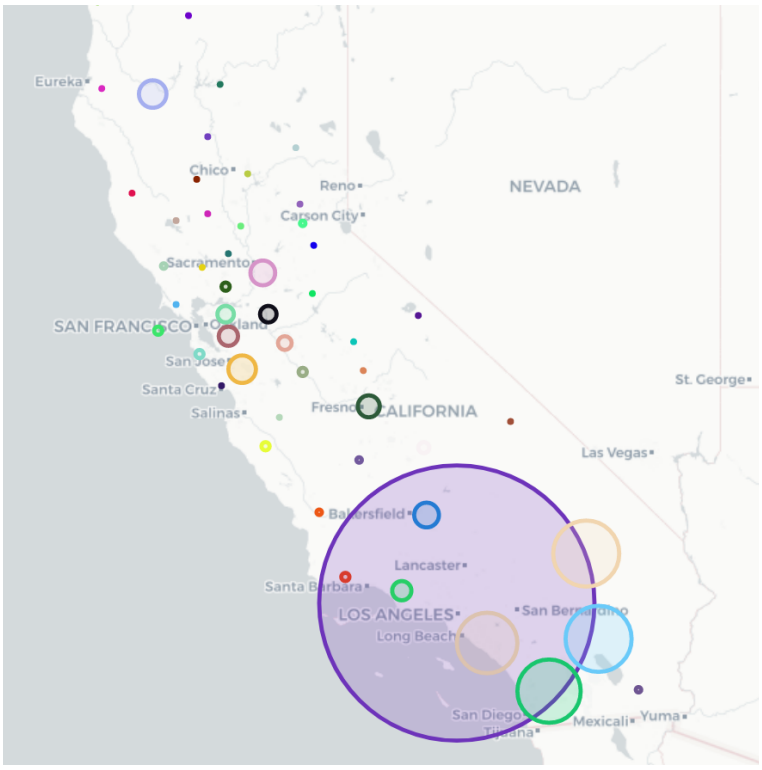


FIG. 20. (Initial Map Plot of Cumulative COVID-19 Case Numbers Per CA County Until June 2021)

We observe a greater number of cases in southern counties, with Los Angeles taking up a significantly larger density circle.

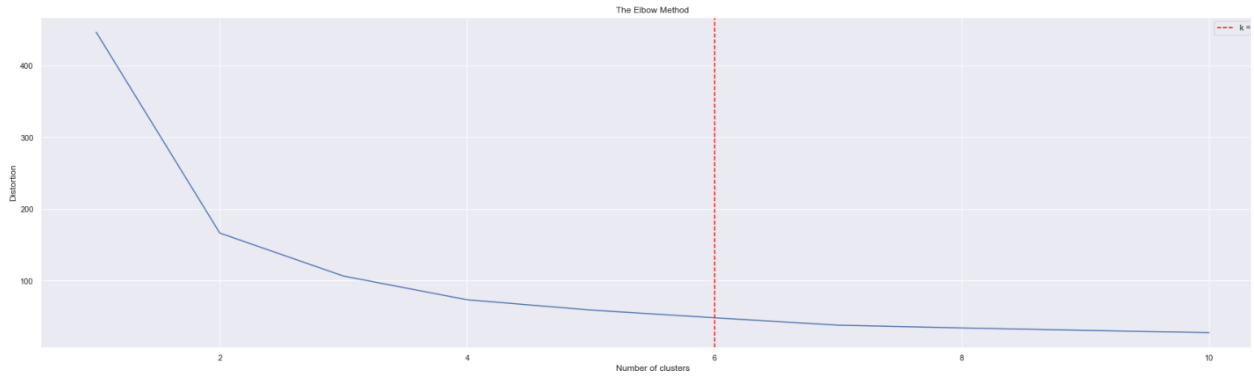


FIG. 21. (Elbow Plot to Determine Number of Cluster for K-Means Clustering Analysis)

Next using K-means clustering we group counties to reveal a pattern among them. Using the elbow plot we find that a possible number of clusters among the California counties is 6. We will choose 6 clusters for our algorithm and plot the results.

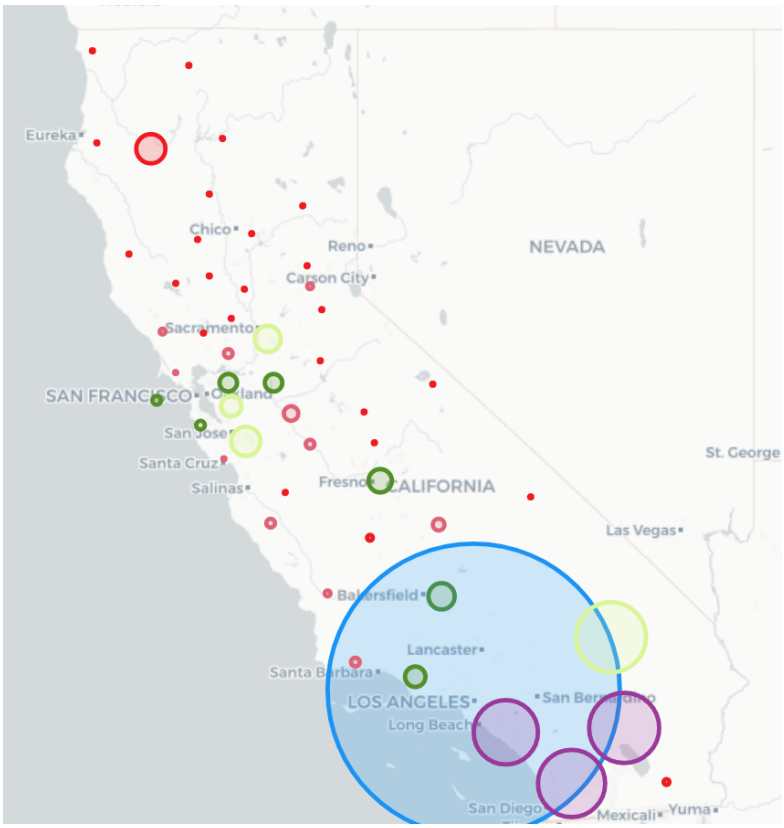


FIG. 22. (Map Plot of Cumulative COVID-19 Case Numbers Per CA County Until June 2021 With K = 6.)

In our final map plot of clusters, we are able to see that Los Angeles due to its pm2.5 concentration, high population, case, and death numbers have clustered separately than the rest of California. We observe interestingly that San Bernardino is grouped into the same cluster as some bay area counties. This is because the Bay Area Counties and San Bernardino share similar average pm2.5 concentration as well as population numbers. We can also see that many Northern and

Central Californian urban areas such as Fresno and San Francisco are clustered together, most likely due to population size.

V. CONCLUSION

In our report, we found different correlations between contrasting features of our dataset in regards to COVID-19. Through our process of exploratory data analysis, we found interesting trends on a regional and local scale. Our regional scale mainly consisted of us looking into our own state, California. In terms of looking at a regional scale, our data analysis consisted of plotting maps of the United States and looking at the effects COVID-19 had within that area. We found area's in the US that had up to 200,000+ cases of COVID-19 along with area's where not too much reporting was done. Along with this, we also plotted a map of California to take a closer look at the counties and see which features in this dataset directly affected these counties. We looked at features such as age group, death, sex and majority of counties with prevalence in COVID. Through this, we concluded that age groups 20-29 seem to be the most prone to COVID-19 diagnosis along with Los Angeles county having by far the highest number of county cases in California. We also concluded that females are more likely to be diagnosed than males in California.

Our models show that COVID-19 affected California similarly based on geographical region and population size. Some outliers we noted include Los Angeles County, which is its own cluster group due to its extreme COVID-19 case count and Imperial County. We also conclude that the symptoms and other personal information of a patient in America is pretty indicative of their probability of death due to COVID-19 contraction.

APPENDIX

REFERENCES (BIBLIOGRAPHIC)

"County Population Totals: 2010-2020." United States Census Bureau, https://www.census.gov/programs-surveys/popest/technical-

documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-counties-total.html

nytimes. "Coronavirus (Covid-19) Data in the United States." Github, https://github.com/nytimes/covid-19-data

"Outdoor Air Quality Data." United States Environmental Protection Agency, epa.gov/outdoor-air-quality-data/download-daily-data.

wxwx1993. "Air pollution and COVID-19 mortality in the United States." GitHub, https://github.com/wxwx1993/PM_COVID.