The Effects of the Covid-19 Pandemic on Groups at Risk in California

David Mao[1, a]

Student ID: 914766922

The Covid-19 pandemic has had a profound impact on the state of California. According the the CDC, California has suffered 3.6 million cases and 63 thousand deaths by the start of June 2021. Being a large and diverse state, California has many different vulnerable populations that reside within its borders. This includes large populations of racial minorities, the economically underprivileged, the uninsured, and the elderly. These groups may be under-equipped to deal with such a devastating global disaster, and therefor more likely to suffer the adverse effects of the pandemic. Using data to understand the impact of the pandemic on these groups at risk is key to understanding the level of impact, and ensuring the Californian government sends the appropriate aid and resources to these groups to help mitigate the pandemic's negative impacts.

[a]dlmao@ucdavis.edu

I. INTRODUCTION

The dataset we will be using for this analysis is the Covid-19 Case Surveillance with Geography, which denotes all CDC confirmed Covid-19 cases in the United States. In our report, we seek to determine the impact of the pandemic on vulnerable populations, specifically racial minorities, the impoverished, the uninsured, and the elderly, in the state of California. We will be looking at the impacts of the pandemic on these vulnerable groups first at the state level, then more in-depth at a county level basis and over time. More specifically, we will check for overrepresentation of Covid-19 cases for certain racial groups, then check if counties with a higher rate of poverty, uninsured populace, and elderly populace have a significant correlation with infection rates. We will then employ hierarchical clustering to find similar counties in their composition of groups at risk and Covid-19 infection rates. Some important insights that we find is that African Americans and Native Hawaiian/Pacific Islanders are the most vulnerable races to the Covid-19 pandemic, and are overrepresented by 1.4 times and 1.8 times their population makeup respectively. Additionally, we find that there is a high positive correlation between counties with high poverty rates and high Covid infection rates, as well as high insured rates and high Covid infection rates.

II. DATA ANALYSIS

A. Data Preparation

An initial glance at the dataset reveals a few major problems with our dataset. For one, there are many missing values in many of the columns in our dataset. In particular, there

are many missing values in the race column and res_county columns, both of which we will be using throughout our data. Many of the values in the race column are denoted missing or unknown, furthur compounding this missing data issue. Additionally, the length of the dataset is only 24 million long, while CDC's own website states that there have been 32 million cases in the United States. This problem is mitigated when we restrict the dataset to only California (3.6 million cases vs 3.8 million reported on CDC's website), but there is still a large portion of the missing values in the data. Too much of the data is missing for us to handle all at once at this stage, so instead, we will deal with the missing values then discuss what doing so entails when we require certain variables later on.

B. Impact of Covid-19 in California

Before we examine the effects of the pandemic on groups at risk, we first observe the impact of the Covid-19 pandemic on California. Observing a time-series plot of the number of cases overtime in Fig. (1), we can see a small spike in cases at around July of 2020, then a major spike occurring around December of 2020. We will take note of these periods, as certain groups at risk may be especially vulnerable during the worst periods. Meanwhile, Fig. (2) shows which counties Covid-19 cases occur, and Fig. (3) shows the infection rate of each county. These two maps show that unsurprisingly, the most cases occur in the most populated areas (see Fig. (4)), but percentage-wise the hardest hit regions are Southern and Central California, while the Bay Area counties, despite being heavily populated, have low infection percentages.

C. Racial Composition of California

To start our analysis on the effects of Covid on different racial groups in California, we first observe the overall Racial Composition in California. To do this, we will use United States Census information from 2019. We will be looking at the characteristics of 5 race groups in California, White, Asian, Black, Native Hawaiian Pacific Islander, and Native American/Alaskan Native. Fig. (5) shows the distribution of different race groups in California, while Fig. (6), Fig. (7), Fig. (8), Fig. (9), Fig. (10), shows the percentage populace makeup for different race groups in California Counties. The most important things to take note of are that the population centers are where most ethnic minorities (Asian, Black, and Hawaiian/Pacific Islander) reside, with the exception of Native Americans/Alaskan, who reside mainly in the isolated very Northern sections and Eastern most parts of California.

D. Impact of Covid-19 on Racial groups in California

Going into the analyzing the effect of the Covid-19 pandemic on different racial groups in California, we need to first do some data preparation on the Race column. We find that 58% of the entries in the race columns are either NaN values, Missing, or Unknown. An additional 8% percent of data is labeled Multiple/Other, which we will not consider due to this group being too broad to give any insightful analysis. In total, we are only left with 32% of the data that contains labels of a specific race, which corresponds to the five races we mentioned earlier: White, Asian, Black, Native Hawaiian Pacific Islander, and Native

American/Alaskan Native. Given we only have a small portion of the data labeled, we need to make assumptions that the labels that we do have are a good representation of the data.

Using the data that we have, we will calculate two values for each of our race groups: Population Percentage, which is found by dividing the population of each race by the total population of every race, and Covid Infection Percentage, which is found by dividing the number of Covid Cases from each race and dividing it by the total number of cases. The results of this are found in Fig. (11). We will then find the ratio of overrepresentation, which is found by dividing the Covid Infection Percentage of each race by its corresponding Population Percentage, shown in Fig. (12). From these two results, we can see that Black and Native Hawaiian/Pacific Islanders heavily overrepresented, by 1.4 and 1.8 times the number of expected cases respectively. Meanwhile, Whites and Asians are underrepresented. Surprisingly, Native American/Alaskans are well below their expected representation. However, this number may be slightly misleading, as most Native Americans in America are now of mixed race, which we are not considering in our analysis. However, We can also find the Ratio of Overrepresentation for each month of the pandemic, which is plotted in Fig. (13). We can see at the height of the pandemic, during December 2020, white representation drops while representation of minority groups increases. In particular, Native Hawaiian/Pacific islanders increase from slightly underrepresented to nearly 2.5 times their expected representation.

Next, we take a deeper look into the the ratio of overrepresentation for California Counties. Fig. (14), Fig. (15), Fig. (16), Fig. (17), Fig. (18), show the ratio of overrepresentation for each California county. The main takeaway from these figures are that Whites and

Asians have low representation in the population centers of California, leading to their overall low representation, while conversely Blacks and Native Hawaiian/Pacific Islanders have very high representations in these areas, leading to their overall high representation.

E. Impact of Covid-19 on the Impoverished in California

For this section of analysis, we will use poverty rate of each California County as calculated by the Public Policy Institute of California. The poverty rates for each county are shown in Fig. (19). In particular, we notice that Central California has the highest rates of poverty, while the Bay Area has the lowest rates of poverty. We will now check to see if the poverty rate has any relationship with the percentage of people infected in each California County. The correlation between the poverty rate and infection percentage was 0.408032, indicating that there is a positive correlation between poverty rate and infection percentage. Additionally, a scatter plot and best fit line shown in Fig. (20) visualizes this relationship. We can conclude that high rates of poverty leads to higher infection rates of of Covid-19. This is likely due to underprivileged residents unable to have the luxury of taking precautions against Covid-19.

F. Impact of Covid-19 on the Uninsured in California

For this section of analysis, we will use the uninsured rate of each California County as calculated by the California Health Care Foundation. The uninsured rates for each county are shown in Fig. (21). We can see that the Bay Area has very low rates of uninsurance, while the whole of Southern California has relatively high rates of uninsurance. To test the

relationship between uninsurance rate and Covid infection percentage, we will repeat the what we have done in the previous section. We find that the correlation between uninsurance rate and Covid infection percentage is 0.310675, which is again a positive relationship. A scatter plot and best fit line shown in Fig. (22) again affirms this positive relationship, leading us to conclude that the uninsured are likely more susceptible to contracting Covid. Again, this is likely due to the uninsured not having the proper resources to combat the negatives of the pandemic, leading to higher infection rates.

G. Impact of Covid-19 on the Elderly in California

For this section of analysis, we will be using United States Census information about the age distribution of each California County. In particular, we will define the elderly as any person over the age of 65. We will use the percentage of people over 65 and see if there is a relationship between infection rate and an elderly population. Fig. (23) shows the percentage of people over 65 for California Counties. The plot implies that the populated areas of California have relatively low population percentages of people over 65. Again, to show the correlation between elderly population and Covid infection rate, we will compute the correlation, which we find to be -0.571717, showing that a high percentage of people over 65 usually leads to a lower infection rate. A scatter plot and best fit line found in Fig. (24) shows this negative relationship. However, as we noticed earlier, the counties with a high proportion of people over the age of 65 are also the counties with a lower population, which may mean that although there is correlation, the high population of elderly people

may not be the cause of lower infection rates, but rather the sparse population of these counties.

A better way to check the impact of Covid-19 on the elderly is to bring back the overrepresetation ratio. We will use the United States Census data to calculate the percentage of people over the age of 65, then divide that value by the number of Covid cases in each county from people over the age of 65 over ever age group. Fig. (25) shows the overrepresentation ratio for every California County. We observe that people 65+ have very low representations for all California Counties, with the exception of Lassen county. Given both the negative correlation between counties with high elderly populations and Covid infection rates as well as the underrepresentation of elderly people in Covid infection rates, it seems that although the virus is deadliest among the elderly population, their overall cases are low. This may be due to a combination of good policy decisions in California to protect the elderly as well as the elderly generation understanding to dangers the virus poses especially to them.

H. Clustering California Counties on Covid Impact and Vulnerable Groups

In this section, we will be using hierarchical clustering to cluster California Counties based on indicators of groups at risk that we found to be indicative of higher Covid infection rates, namely racial composition, poverty rates, and uninsured rates, along with the overall population infected. In doing so we hope to find counties or regions with similar makeup of groups at risk that are similarly more susceptible to the Covid-19 pandemic. We will be using the farthest point algorithm. Additionally, we will not scale our features, as every feature is a percentage and scaling our data would put unnecessary weights into features with low

representations. Fig. (26) shows the dendrogram of the resulting hierarchical clustering. We can tell just from this dendrogram that the orange cluster on the left hand side denotes less urban California Counties while the right hand sides contain the urban California Counties. We can furthermore go into more detailed clusters by restricting our hierarchical cluster to five groups. Fig. (27) plots the clusters on a map of California clusters while Tab. (I) has the explicit county names for each cluster. We can see that group 1 corresponds to a group of mainly less urban Northern Counties, group 2 corresponds to a group of mainly Central and Inland Counties, group 3 corresponds to an odd outlier in Alpine County, group 4 corresponds to two distinct regions of urban counties around Los Angeles and Sacramento, and group 5 corresponds to a group of urban counties in the Bay Area. We can check the averages of each of our features within each of the five clusters, as shown in Tab. (II). Group 1 has relatively high number of Whites and Native Americans, and also the lowest average infection rate, which goes in hand with the group's less urban characteristics. The second group has the highest infection rates, and a corresponding highest uninsurance and poverty rates. The third group is the outlier county Alpine, due to it having a Native American population much higher than any other county. The fourth group has a very low infection percentage, especially given its urban makeup, and is also characterized by its very high Asian percentage, and also has the lowest poverty and uninsurance rates of all the clusters. The fifth group has the highest black population, but also has large percentages of all race populations, and is one of the more diverse groups.

III. CONCLUSION

In conclusion, we found that some racial minorities, namely Black Americans and Native Hawaiians/Pacific Islanders, were very overrepresented in Covid-19 cases. Additionally, we found that at the height of the pandemic, racial minorities had much higher representation of Covid infection rates compared to other times. We additionally learn that Whites and Asian Americans are highly underrepresented in heavy urban areas, a major contributing factor to their overall low representations. Additionally, we find strong positive relationships between counties with high poverty and uninsurance rates with each county's Covid infection percentage. Contrary, counties with high elderly population had a high negative correlation with Covid infection rate, and the elderly were also well underrepresented in Covid infections. Finally Hierarchical clustering revealed four distinct regions that have similar characteristics in terms of its at risk groups and Covid infection rates. We have a region consisting of the non-urban Northern Counties with low Covid rates, a relatively low income region consisting of Central and Inland California with high Covid rates, an urban Bay Area region with low Covid rates despite its urban make, and an diverse urban region that has high Covid rates consisting of Urban regions around Los Angeles and Sacramento.

IV. APPENDIX A: PLOTS

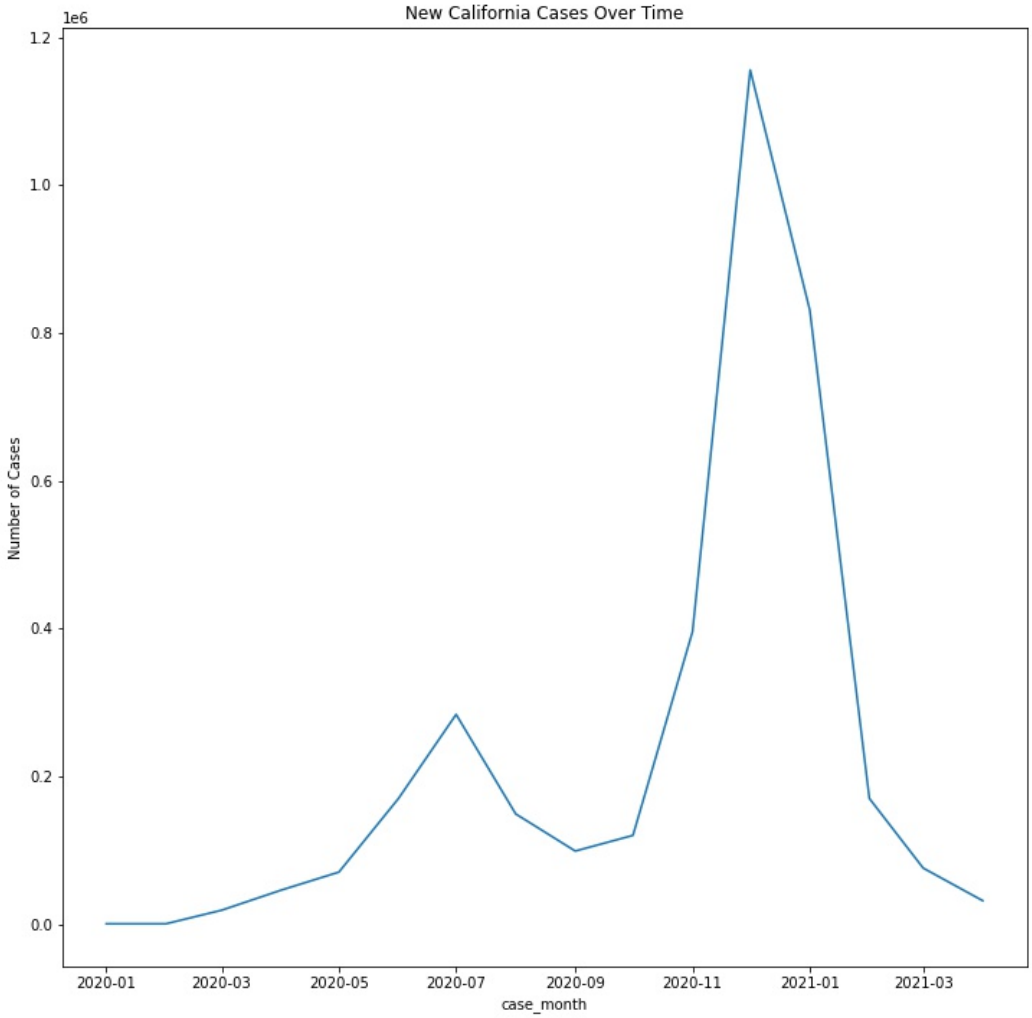

FIG. 1. Time-Series Plot of New Covid-19 Cases in California

We can observe from this plot that the height of the pandemic occurs around the time December 2020. It would be interesting to know if certain groups were more vulnerable during this period of time.

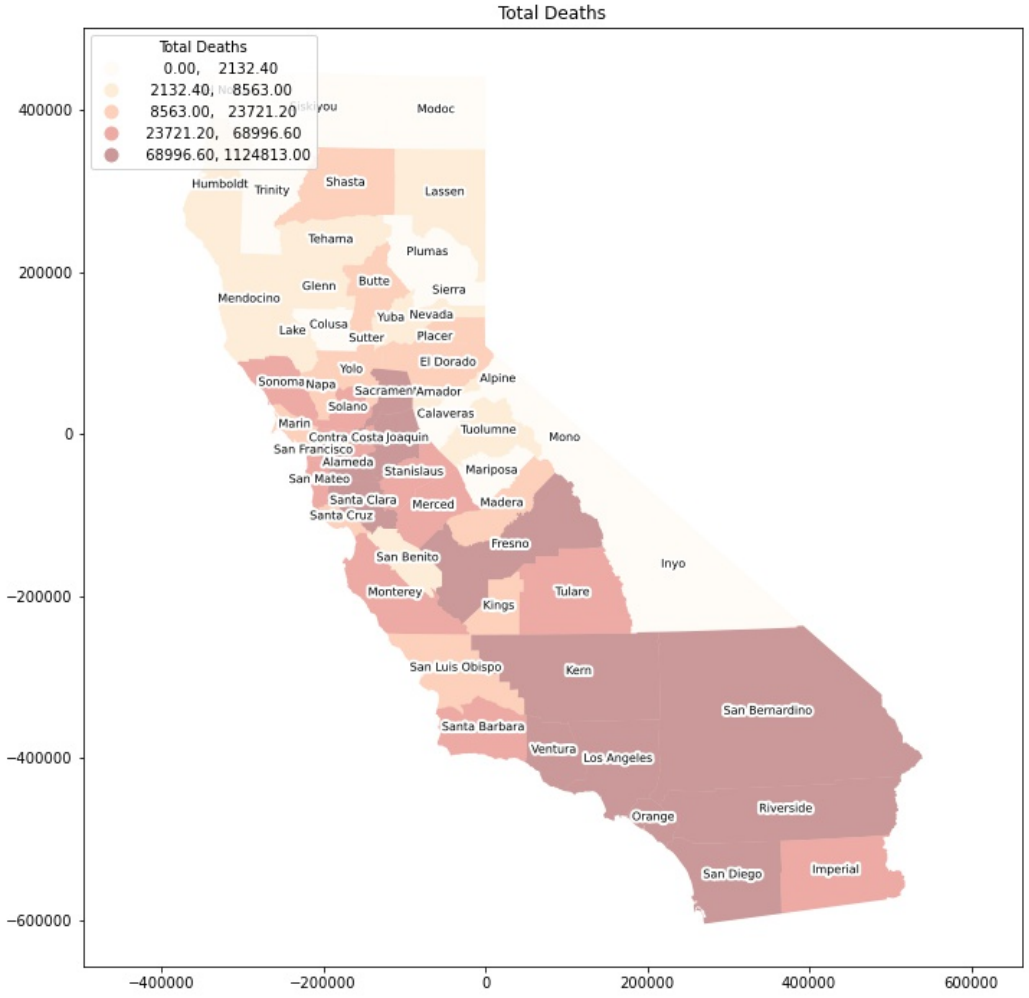

FIG. 2. Map of Covid Infection Across California Counties

We can see that Southern California, Fresno, and the Bay Area are the hardest hit region, which is directly due to the population centers being there as shown in Fig. (4).

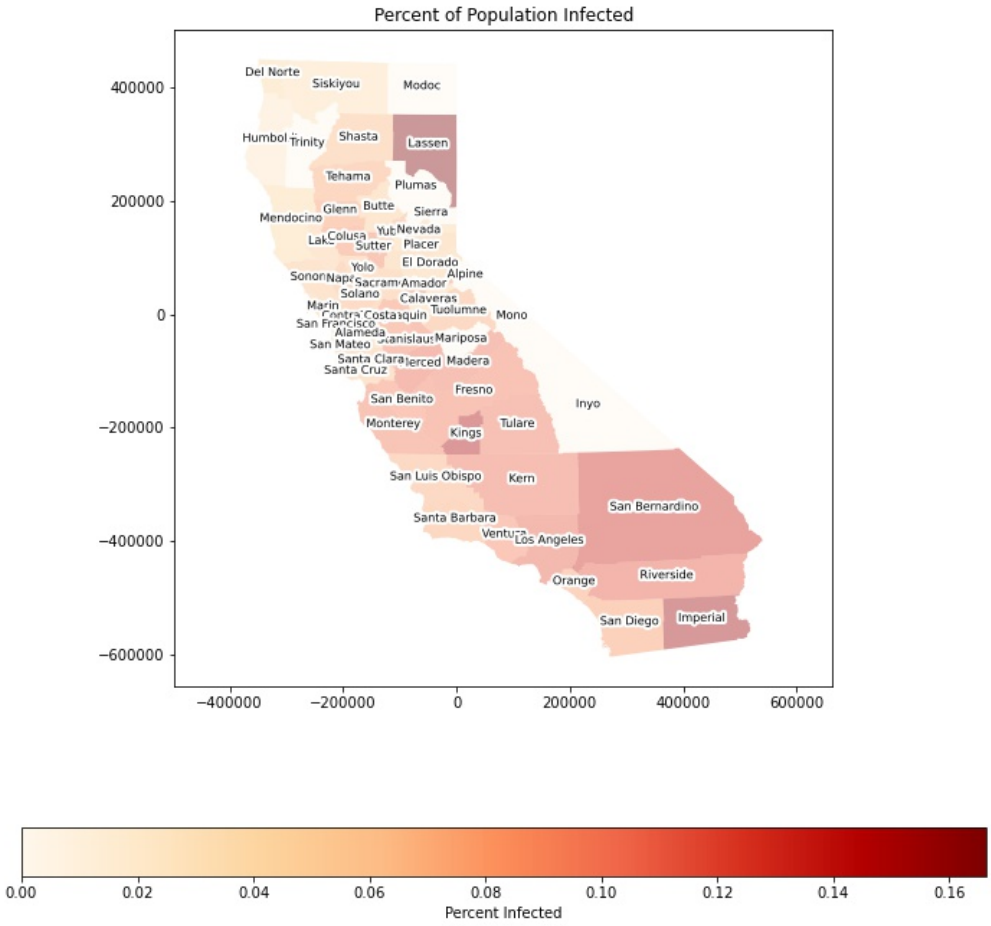

FIG. 3. Map of Percent of Population Infected by California Counties

By percentage, Southern and Central California are the hardest hit regions. Meanwhile, the very north and inland are less affected.

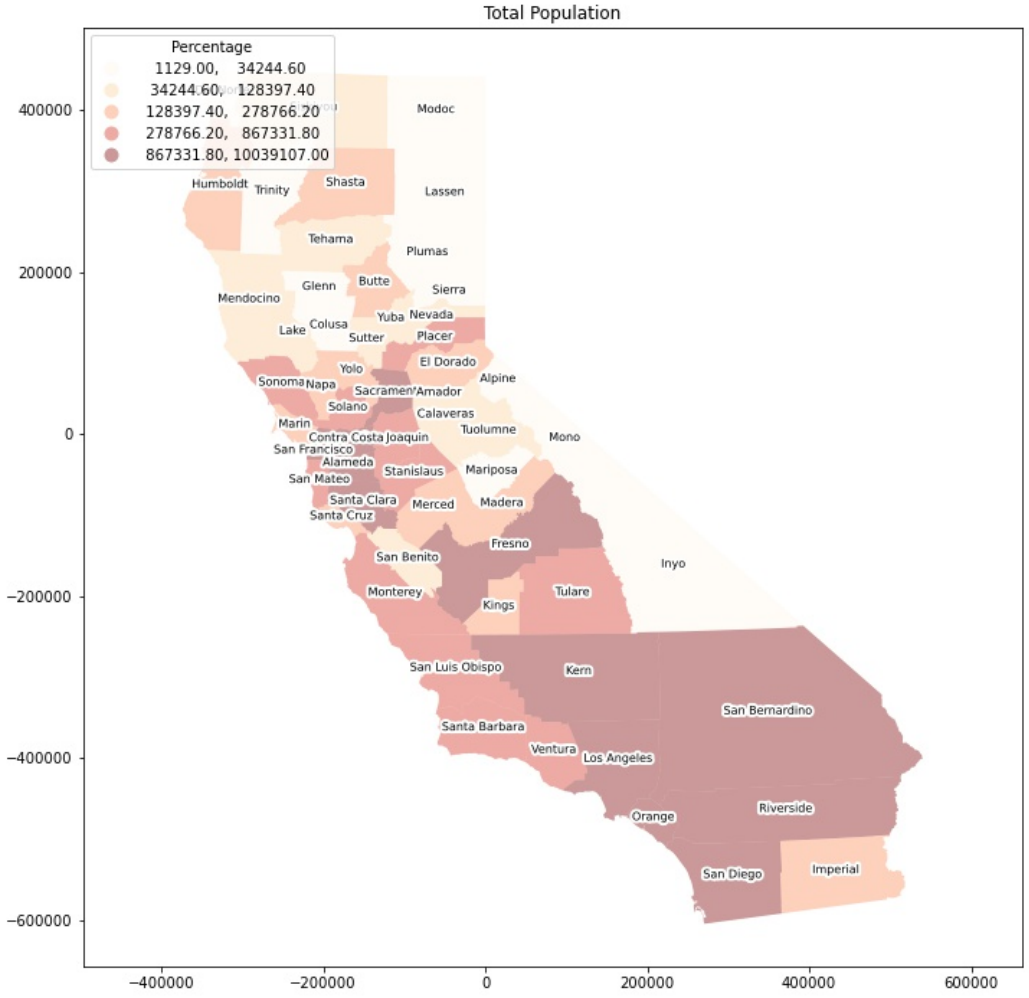

FIG. 4. Map of Total Population of California Counties

We can observe from this plot that the most populous counties in California are in the South in the Inland Empire, Fresno in Central California, and in the Bay Area. Meanwhile the northern parts and inland counties are less populated. As Covid spreads faster in populated areas, it is important to keep track of where California's population centers are.

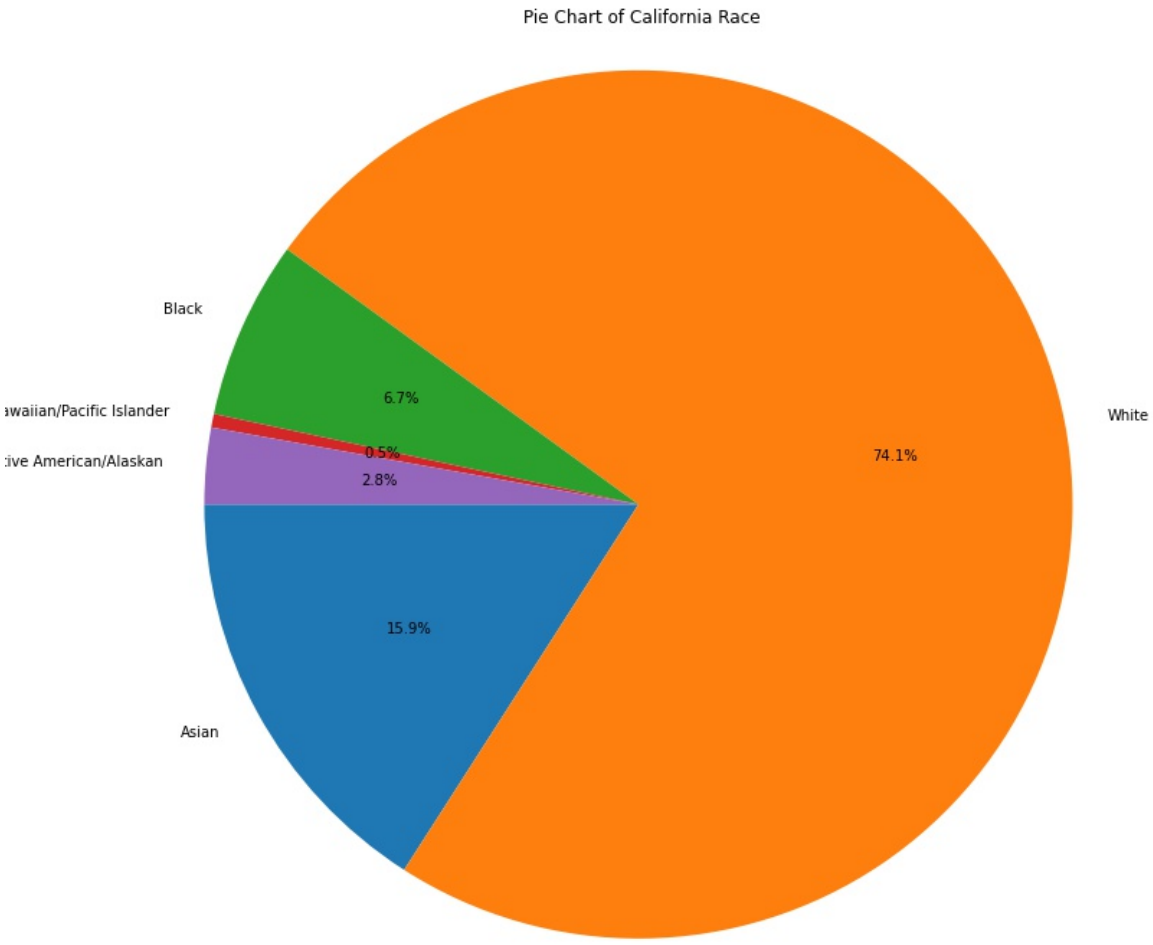

FIG. 5. Map of Total Population of California Counties

We can observe from this plot that the most populous counties in California are in the South in the Inland Empire, Fresno in Central California, and in the Bay Area. Meanwhile the northern parts and inland counties are less populated. As Covid spreads faster in populated areas, it is important to keep track of where California's population centers are.

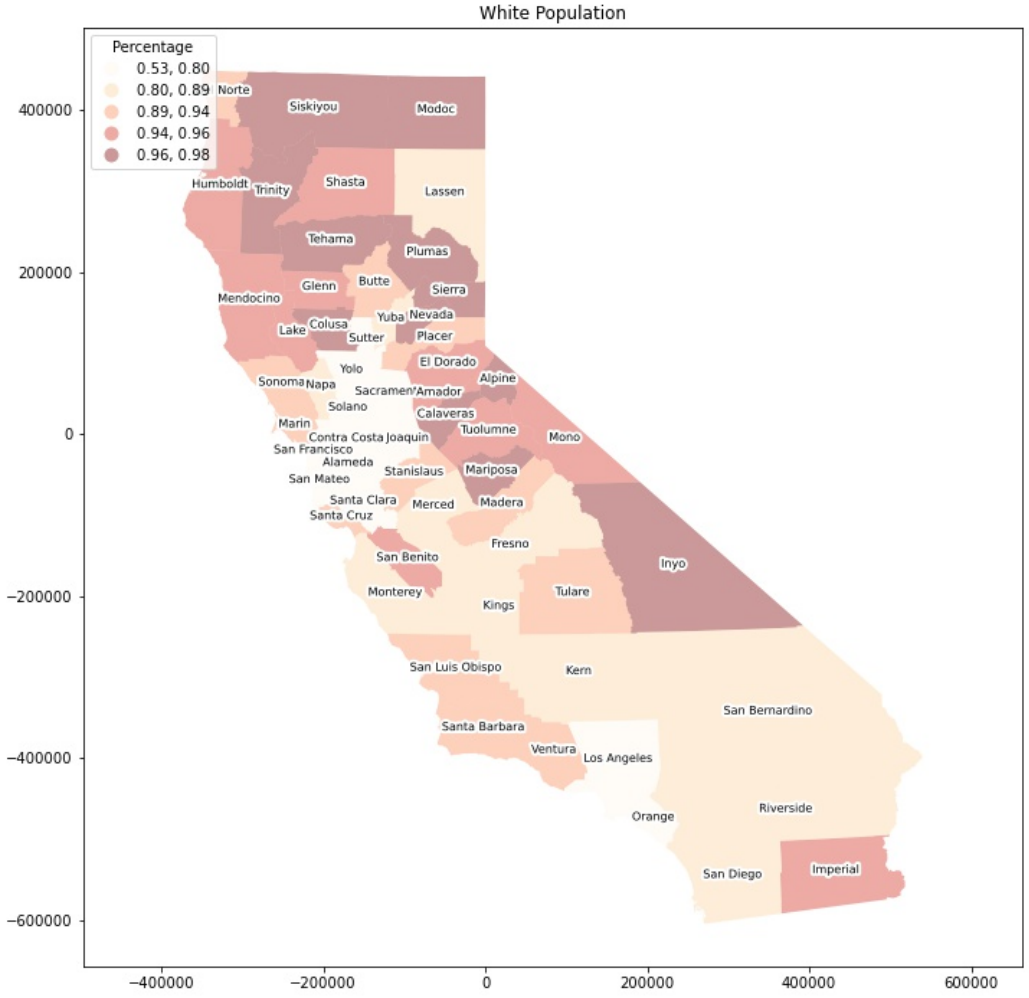

FIG. 6. Makeup of White Populace in California

We can observe that whites make up large portions of the population in every county, but less so in the population centers, indicating that the population centers are the most diverse places.

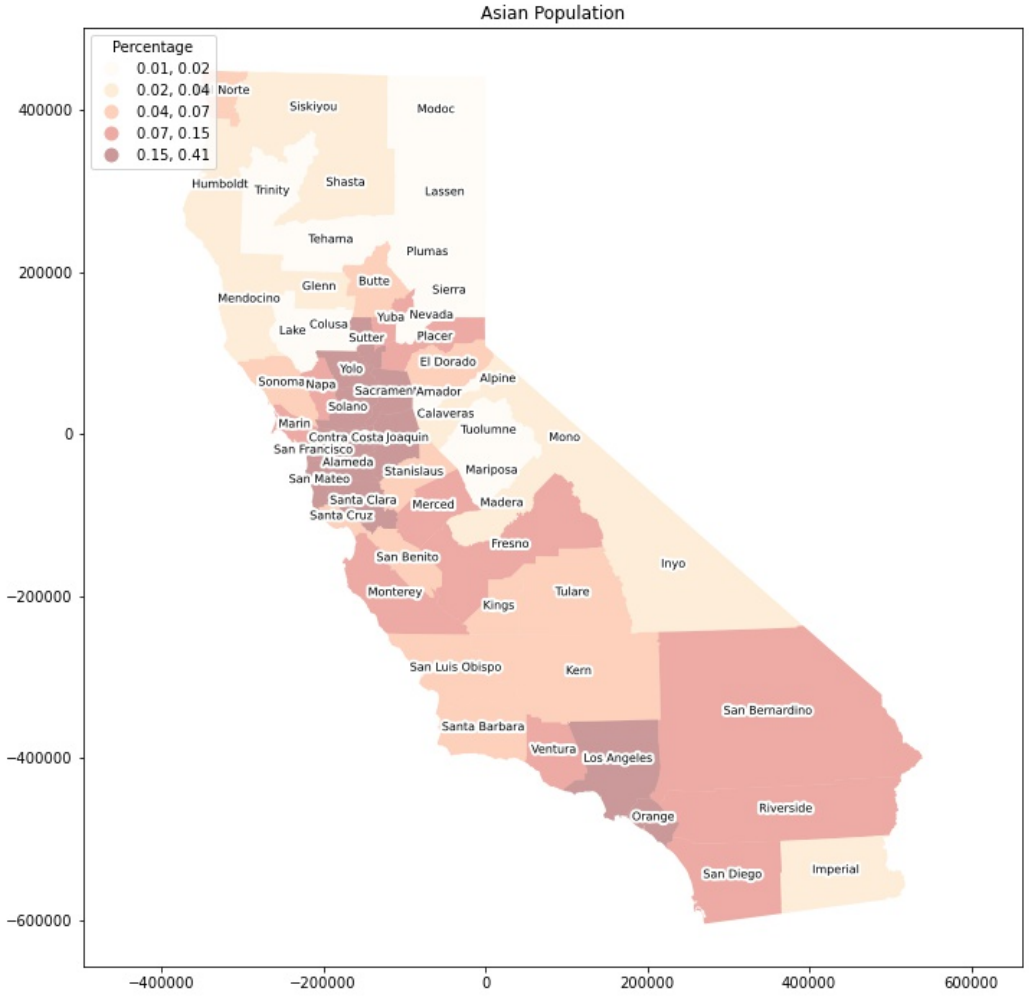

FIG. 7. Makeup of Asian Populace in California

We can observe that most of the Asian populace lives in the population centers of California, especially in the Bay Area.

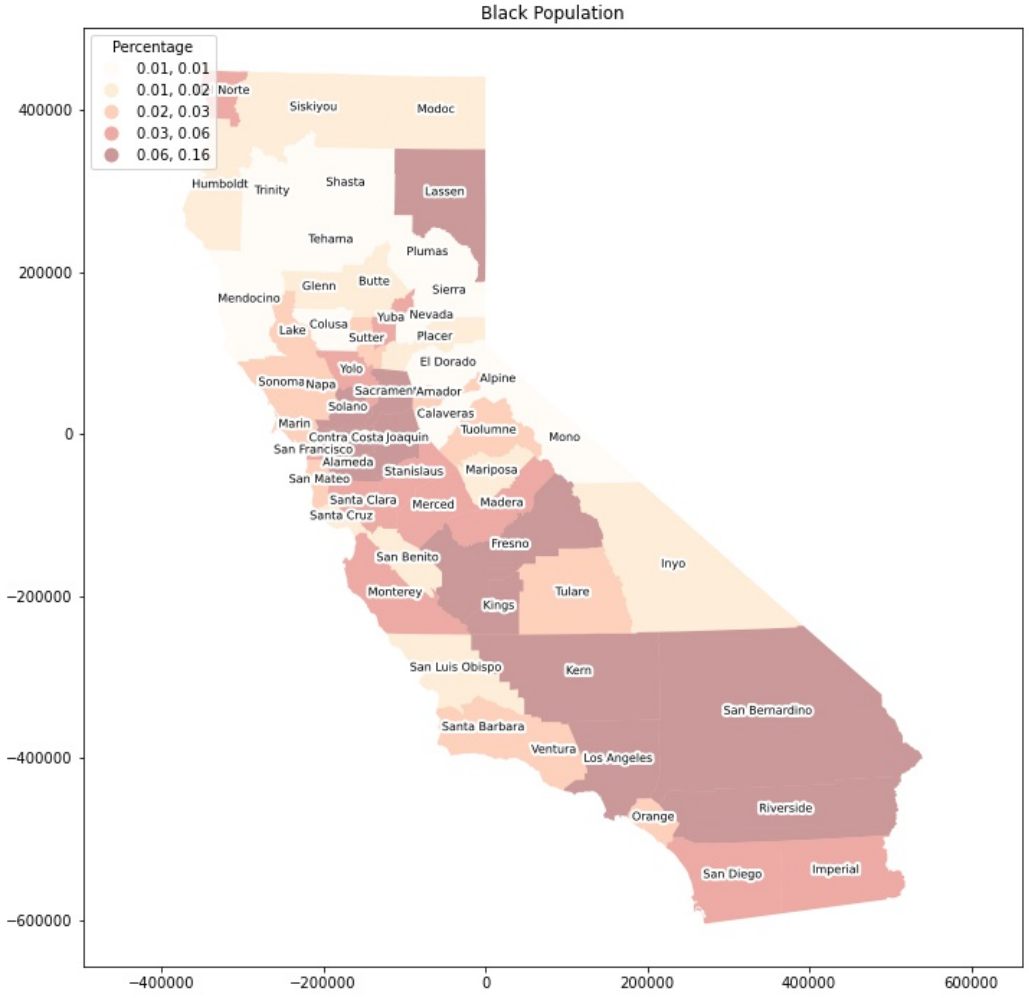

FIG. 8. Makeup of Black Populace in California

We can observe from this map that most of the black population lives in the population centers of California.

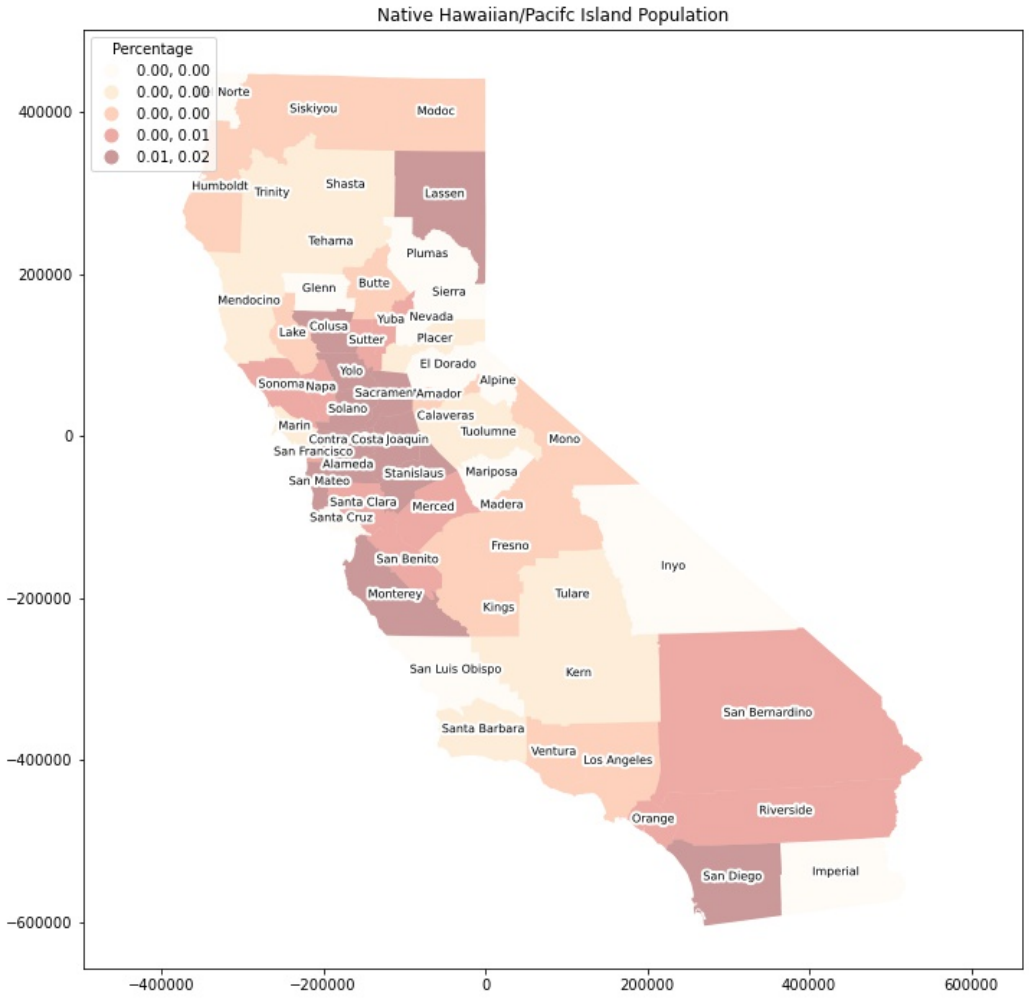

FIG. 9. Makeup of Hawaiian Populace in California

Our map indicates most of the Hawaiian/Pacific Islander populace live near the Bay Area.

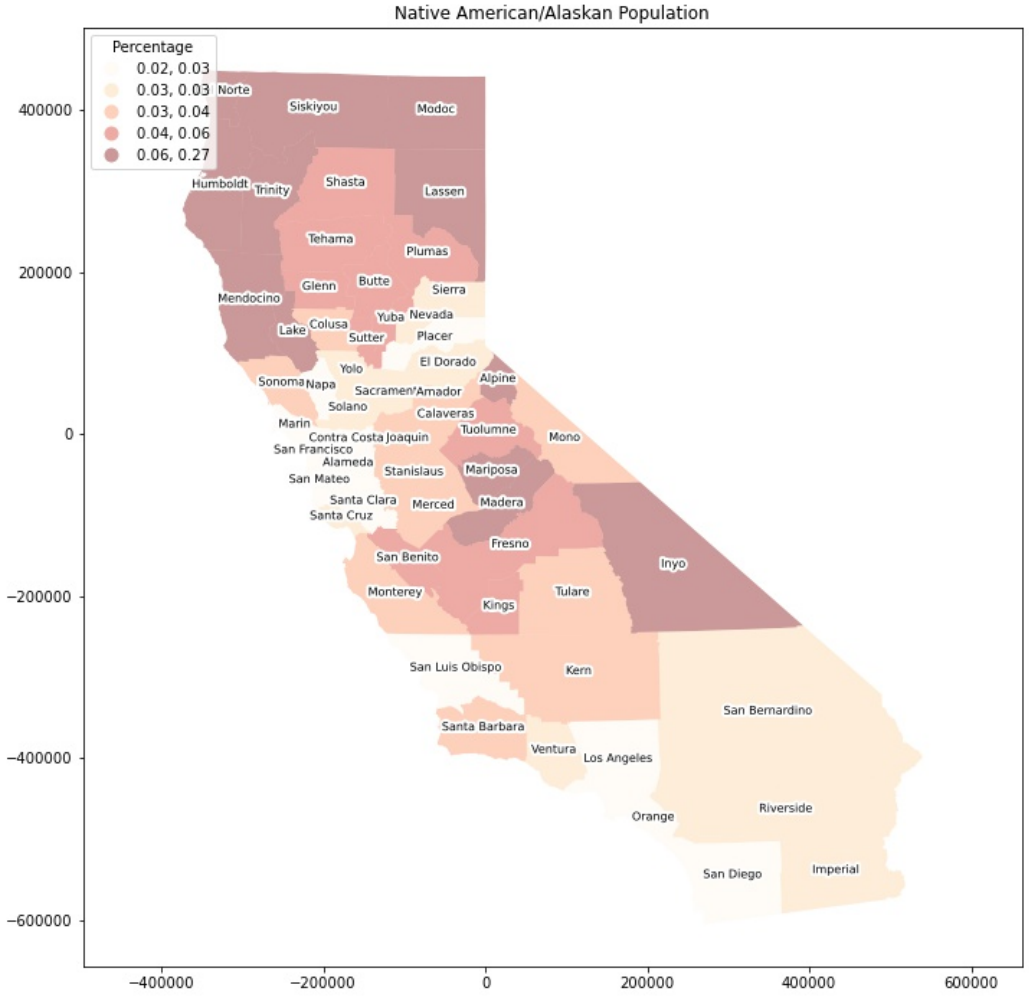

FIG. 10. Makeup of Native American/Alaskan Populace in California

The low populated counties in Northern California and inland have higher proportions of Native American/Alaskan populations

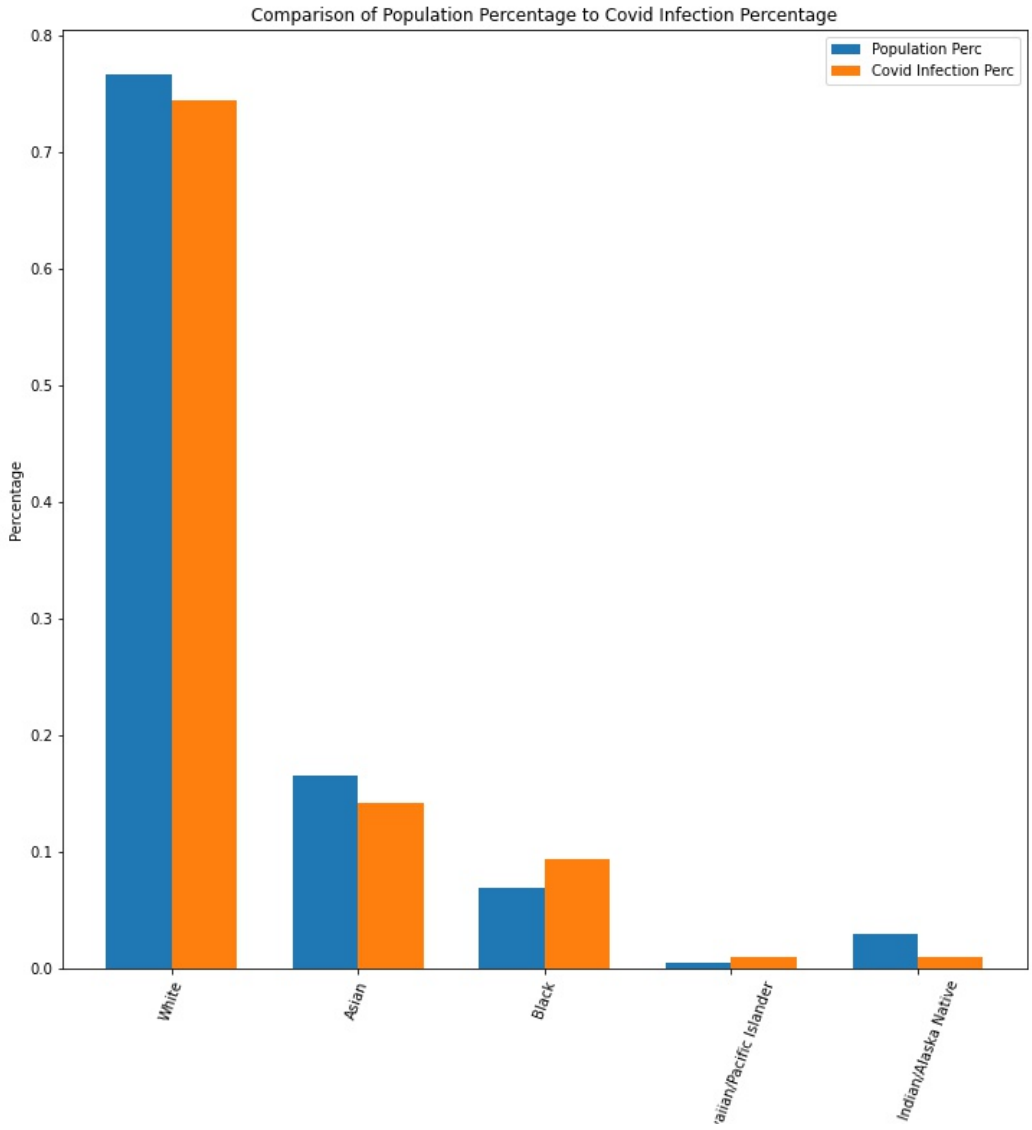

FIG. 11. Population Percentage Makeup of Each Ratio Group Compared to the Percentage Makeup of Covid Cases

A comparison of population percentage makeup and percentage makeup of Covid cases shows that White and Asians are underrepresented in Covid Cases while Black and Hawaiian/Pacific Islanders are overrepresented.

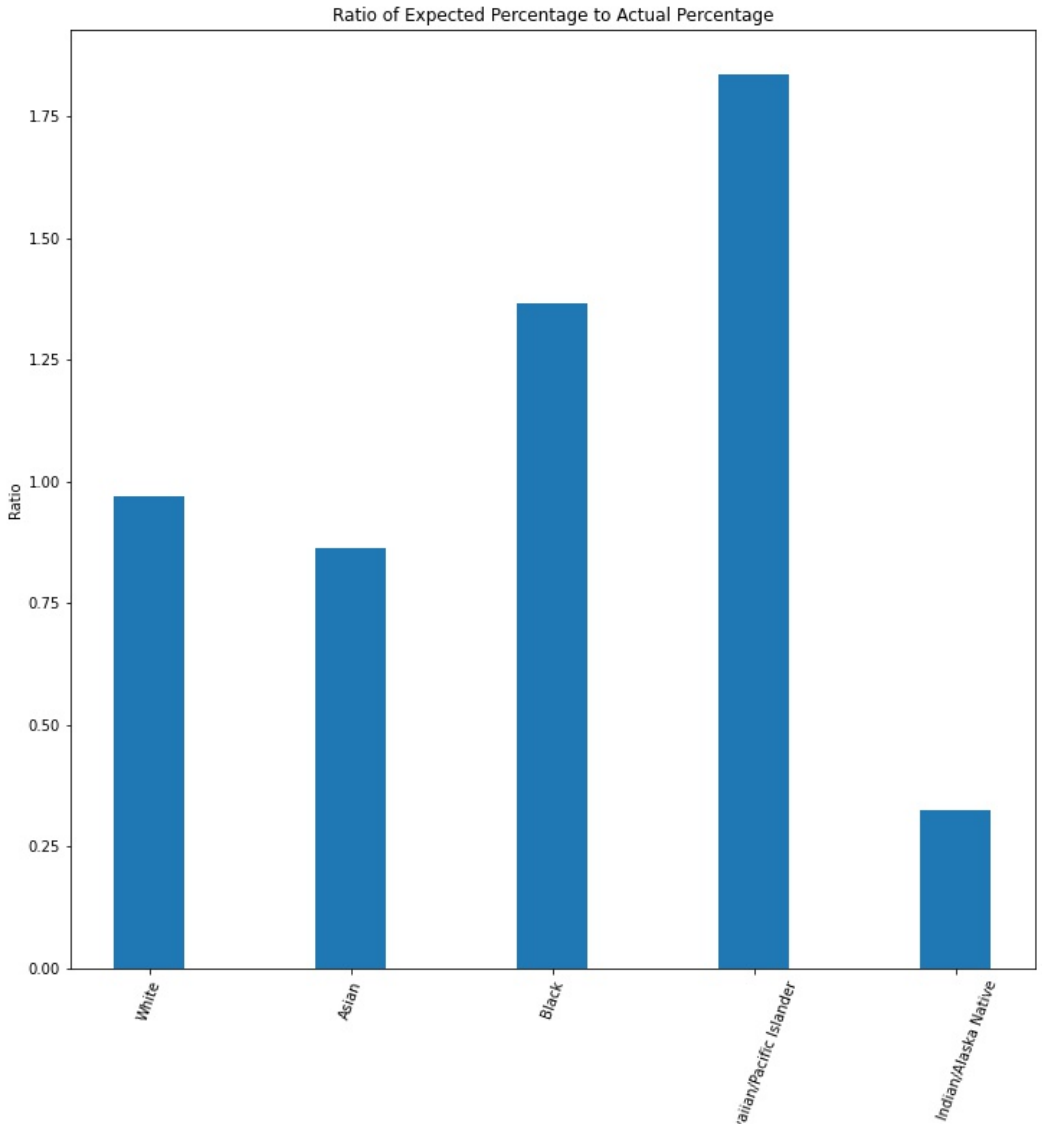

FIG. 12. Ratio of Overrepresentation

The ratios show that African Americans are overrepresented by 1.4 times the expected number of cases and Native Hawaiian/Pacific Islanders are overrepresented by 1.8 times the expected number of cases. Native Americans/Alaskans are notably low, but this may have to do with most Native Americans/Alaskans being of mixed race origin, which we do not analyze.

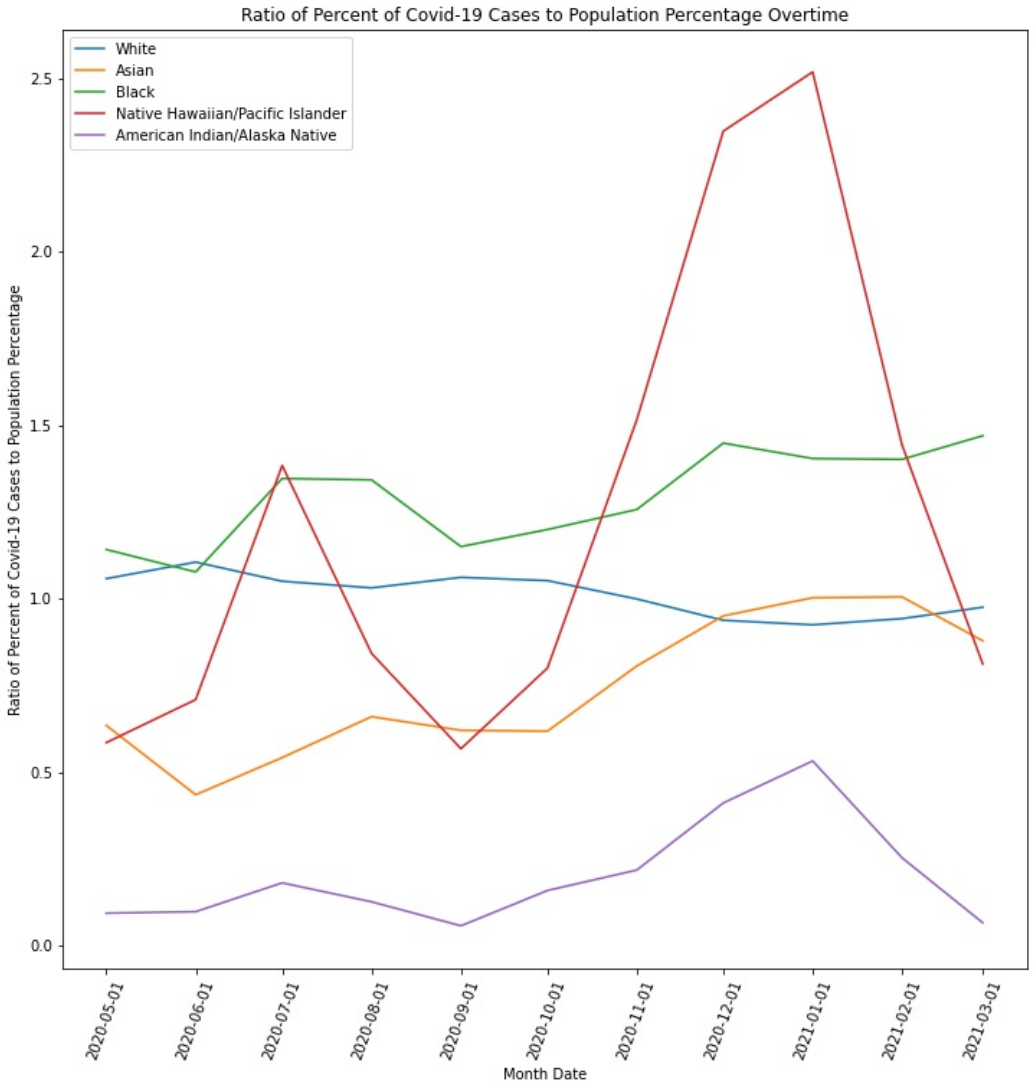

FIG. 13. Time Series of the Ratio of Overrepresentation

As we can see, during the heighten times of the pandemic noted by Fig. (1), the ratios of overrepresentation of the ethnic minority groups increases dramatically, while white representation falls. Notably, Native Hawaiian/Pacific Islanders dramatically increase from being slightly underrepresented to having more than 2.5 times their expected representation.

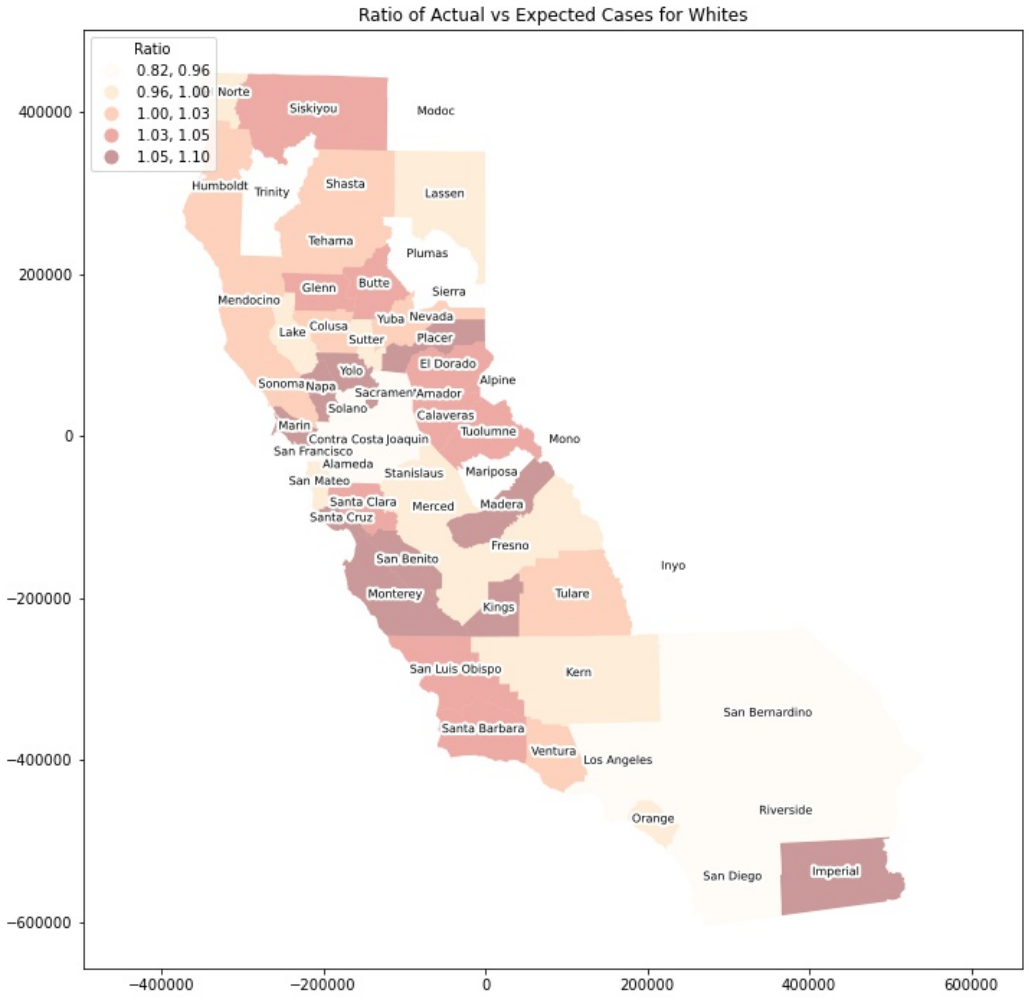

FIG. 14. Ratio of Overrepresentation for Whites

Notably, whites are extremely underrepresented in the population centers in Southern California and the Bay area, leading to an overall low representation.

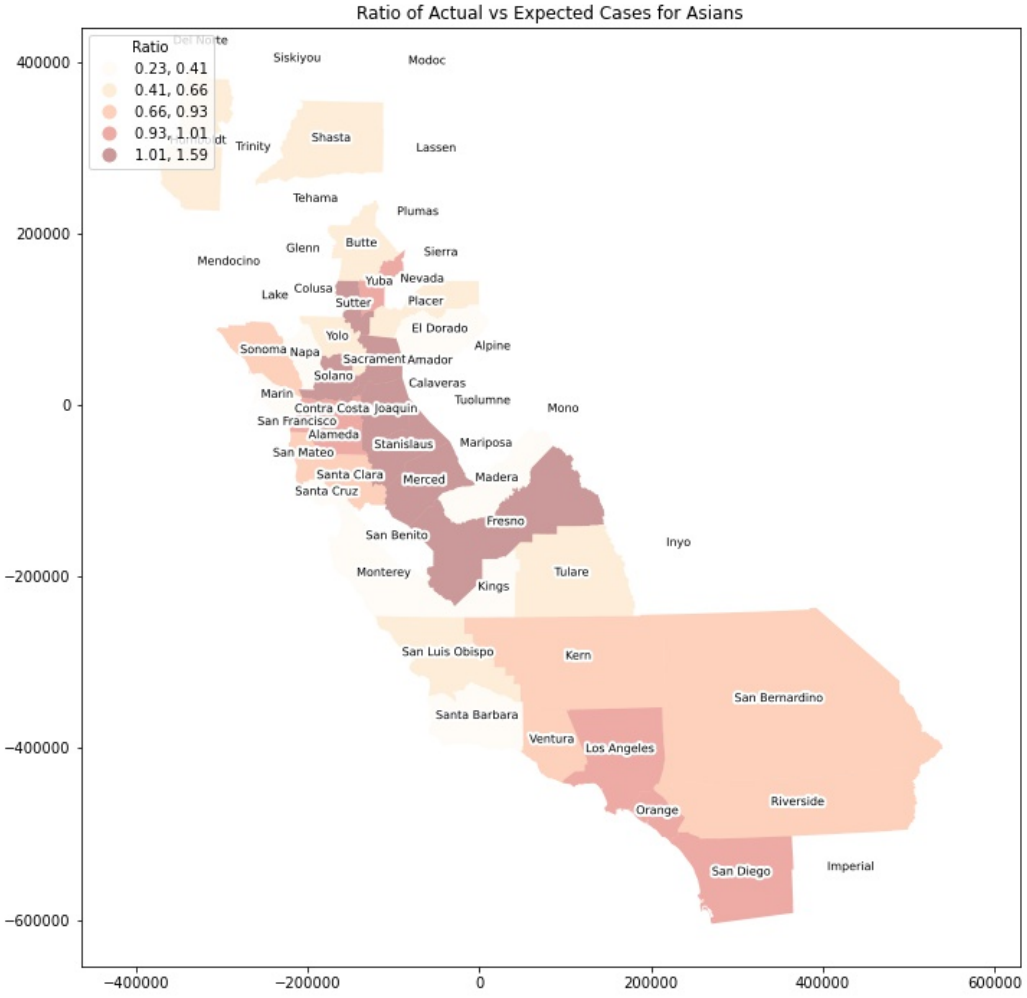

FIG. 15. Ratio of Overrepresentation for Asians

Asians have high overrepresentation in the Central Parts of California, but have low representations in the places where high numbers of Asians reside, such as in the Bay Area and Southern California, resulting in an overall low representation.

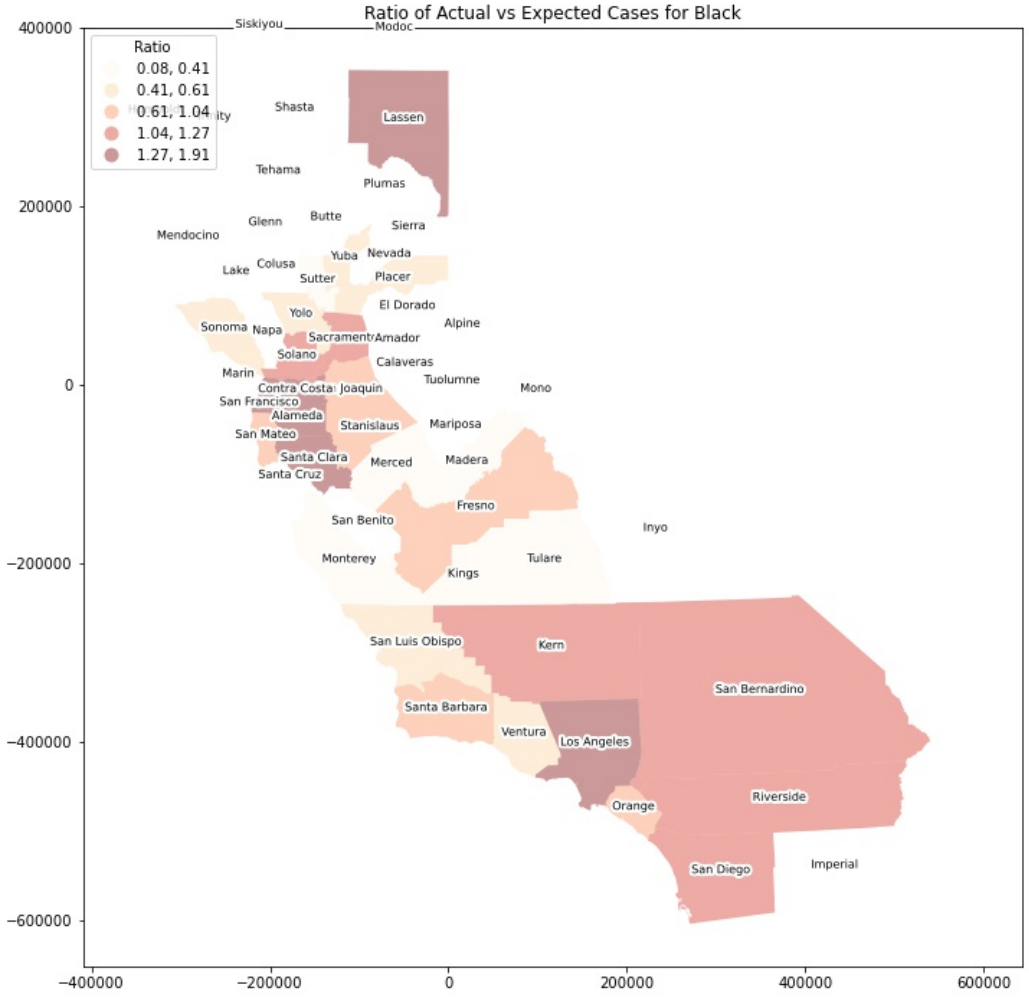

FIG. 16. Ratio of Overrepresentation for Blacks

Blacks have very high representation in the population centers in California, likely the cause of the overall high overrepresentation if this group.

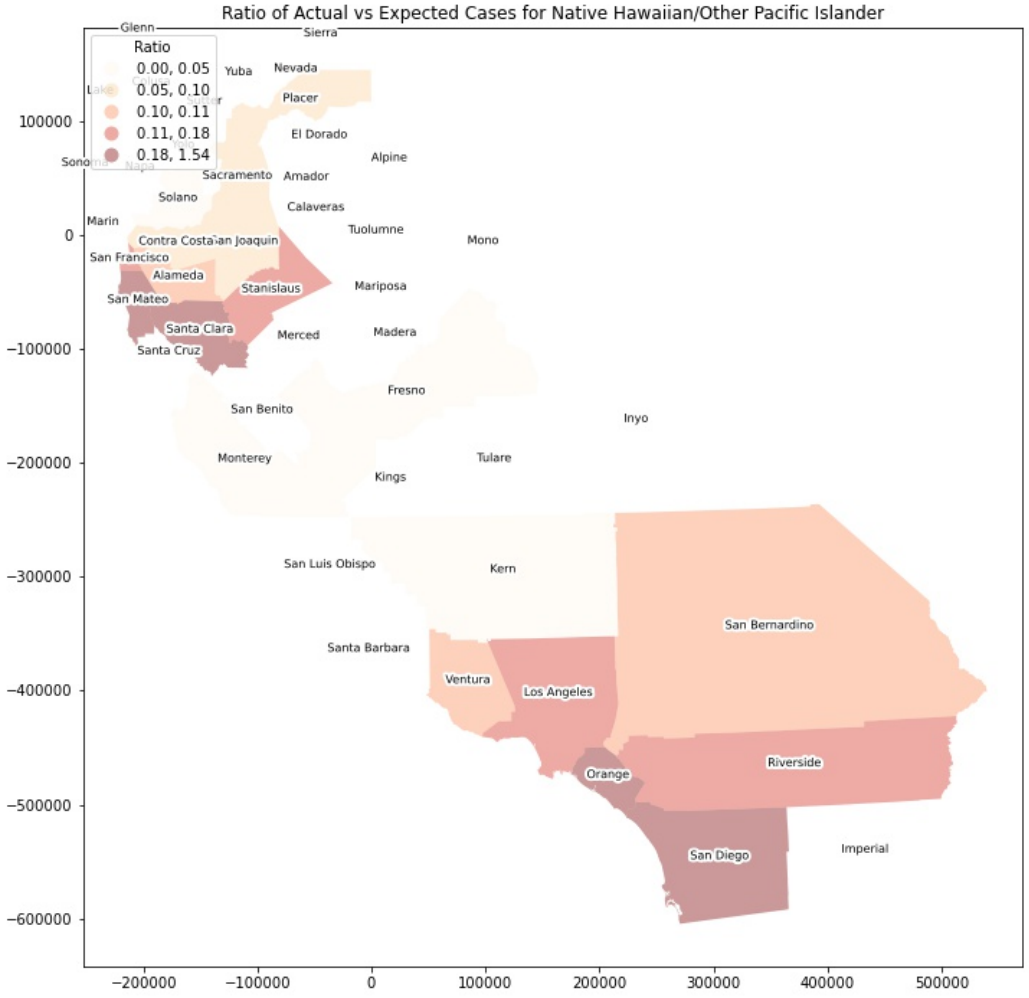

FIG. 17. Ratio of Overrepresentation for Native Hawaiian/Pacific Islander

Native Hawaiian/Pacific Islanders have very high representation in the population centers in California, likely the cause of the overall high overrepresentation if this group.

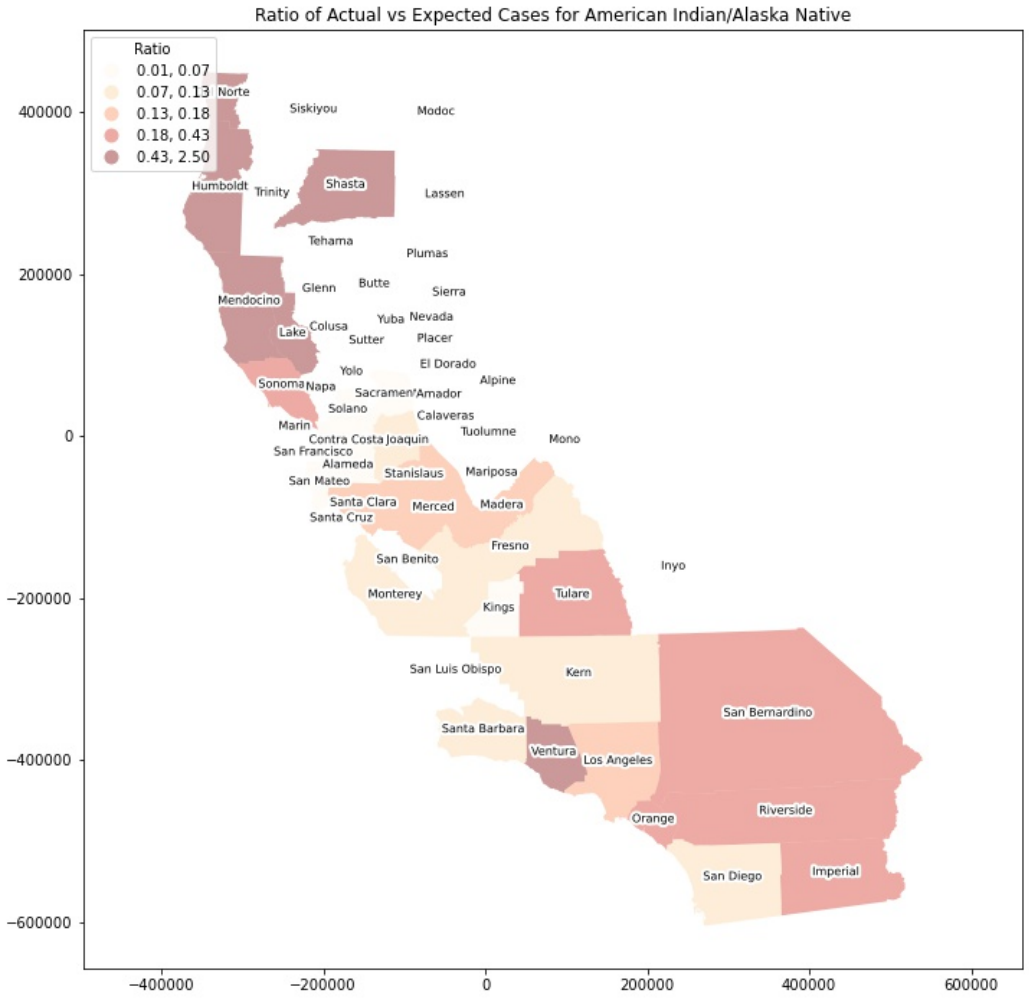

FIG. 18. Ratio of Overrepresentation for Native Americans/Alaskans

Again, Native American representation is surprisingly low everywhere, but as mentioned before, this may be misleading due to a high number of Native Americans being mixed race.

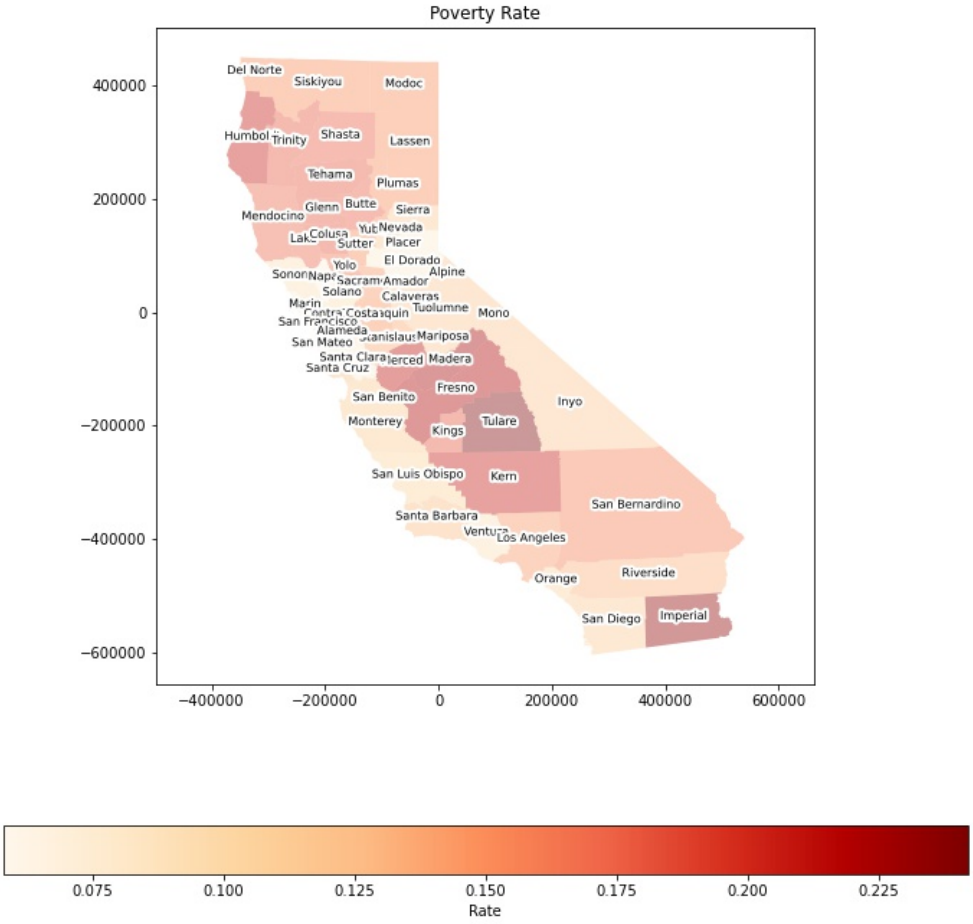

FIG. 19. Poverty Rates for California Counties

We can see that Central California has the highest rates of poverty, while the Bay Area has the lowest rates of Poverty.

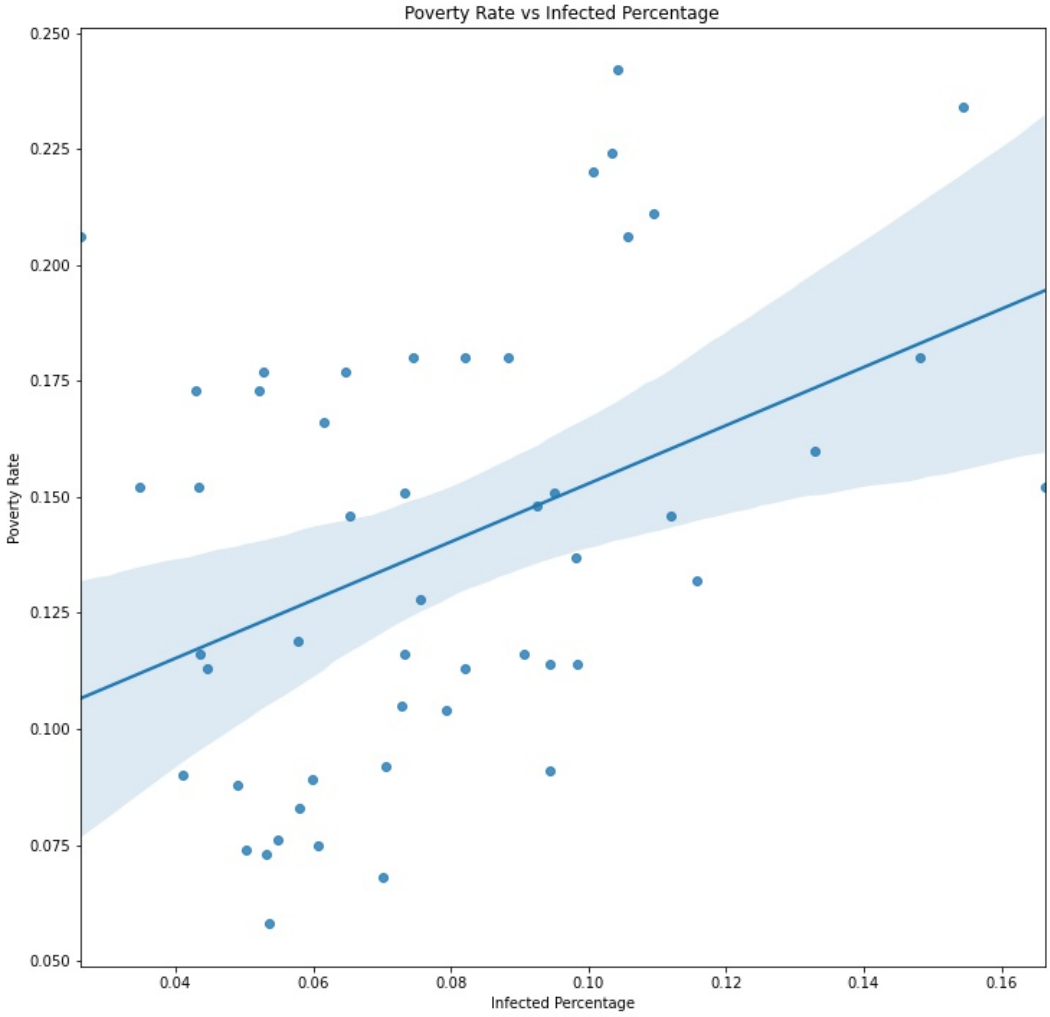

FIG. 20. Scatter Plot of Poverty Rate vs Infected Percentage With Regression Line and Confidence Interval

A scatter plot with a regression line shows a positive relationship between the Covid Infection Percentage and Poverty Rate.

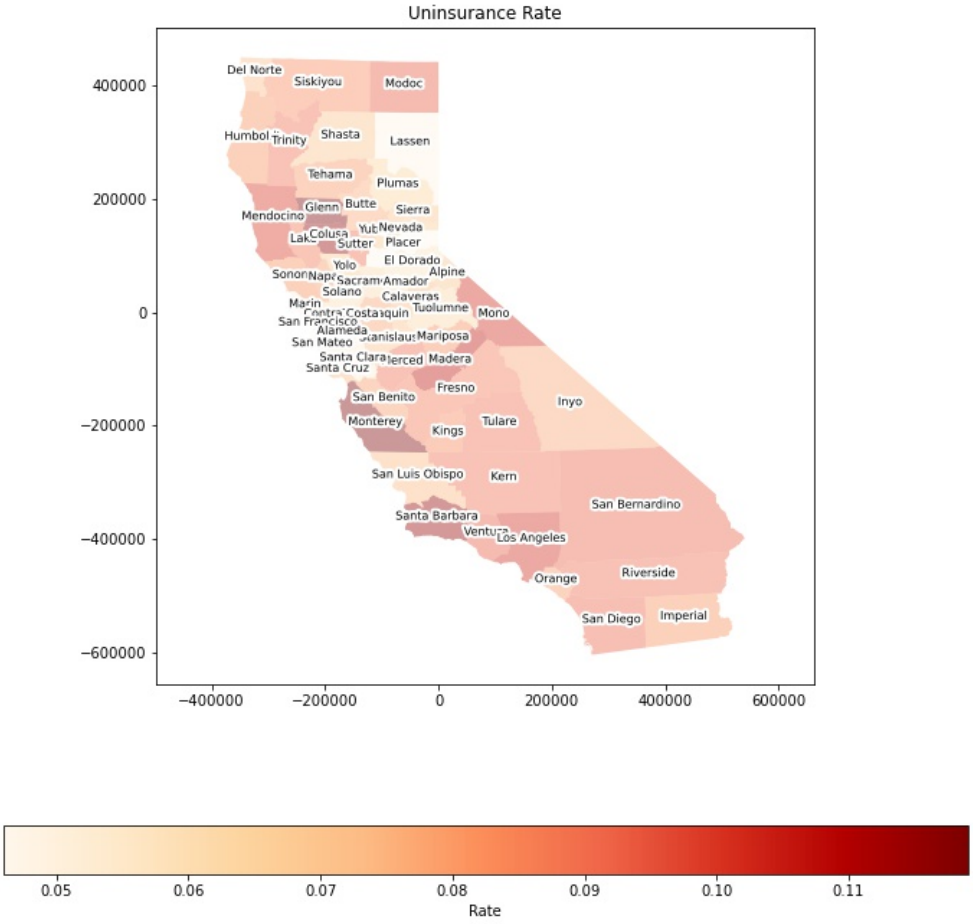

FIG. 21. Uninsured Rates for California Counties

We can see that the Bay Area has relatively low percentage of uninsured people, while Southern California seems to have high rates of uninsurance.

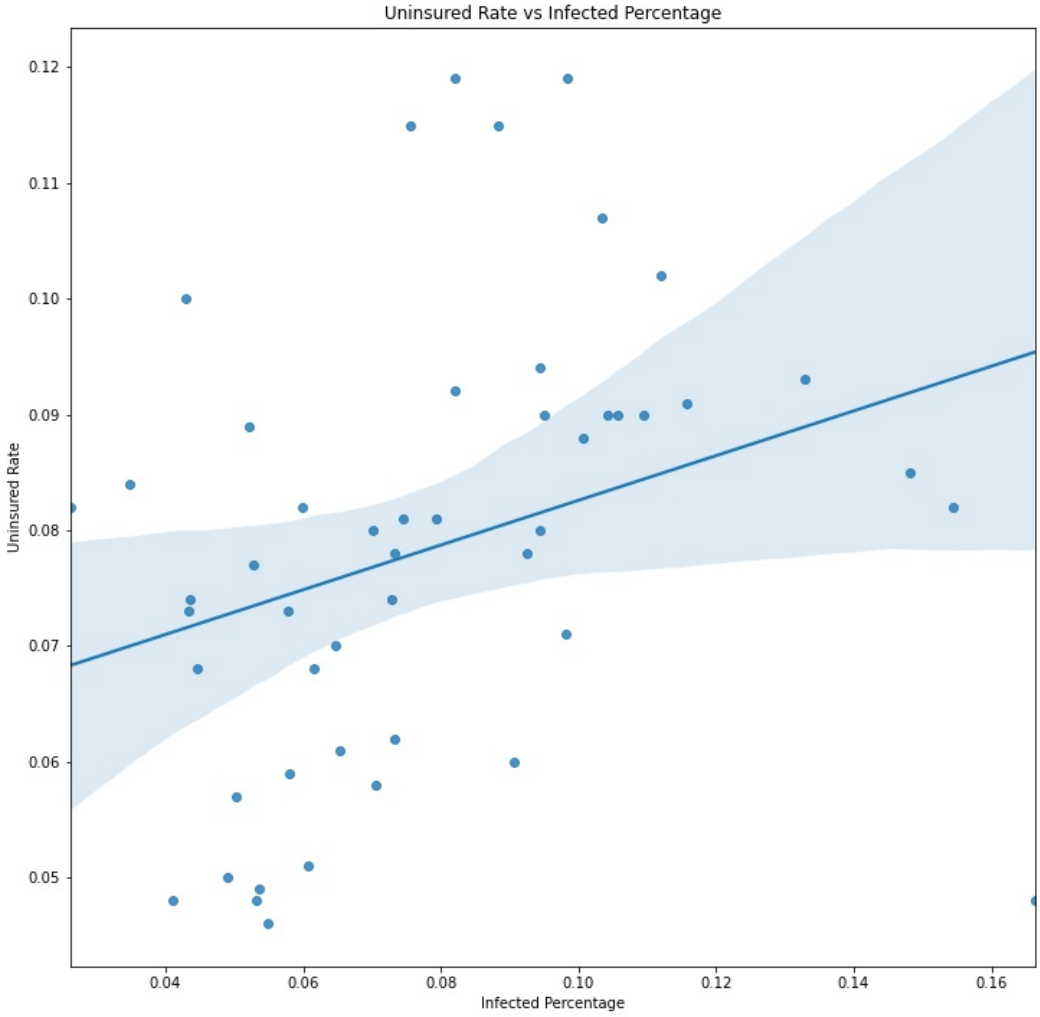

FIG. 22. Scatter Plot of Uninsurance Rate vs Infected Percentage With Regression Line and Confidence Interval

A scatter plot with a regression line shows a positive relationship between the Covid Infection Percentage and Uninsurance Rate.

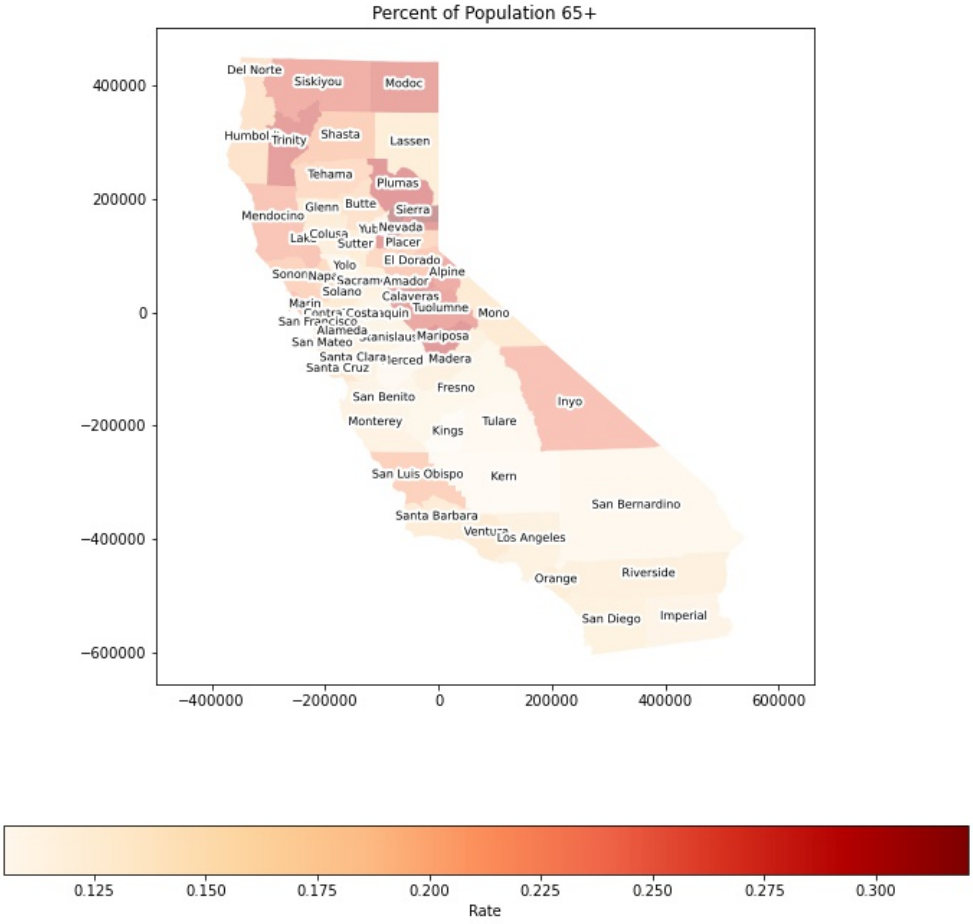

FIG. 23. Map of Percentage of People of 65 for California Counties

We can see that Southern California has low rates of elderly people. The highest rates of elderly people are found the the far Northern Counties.

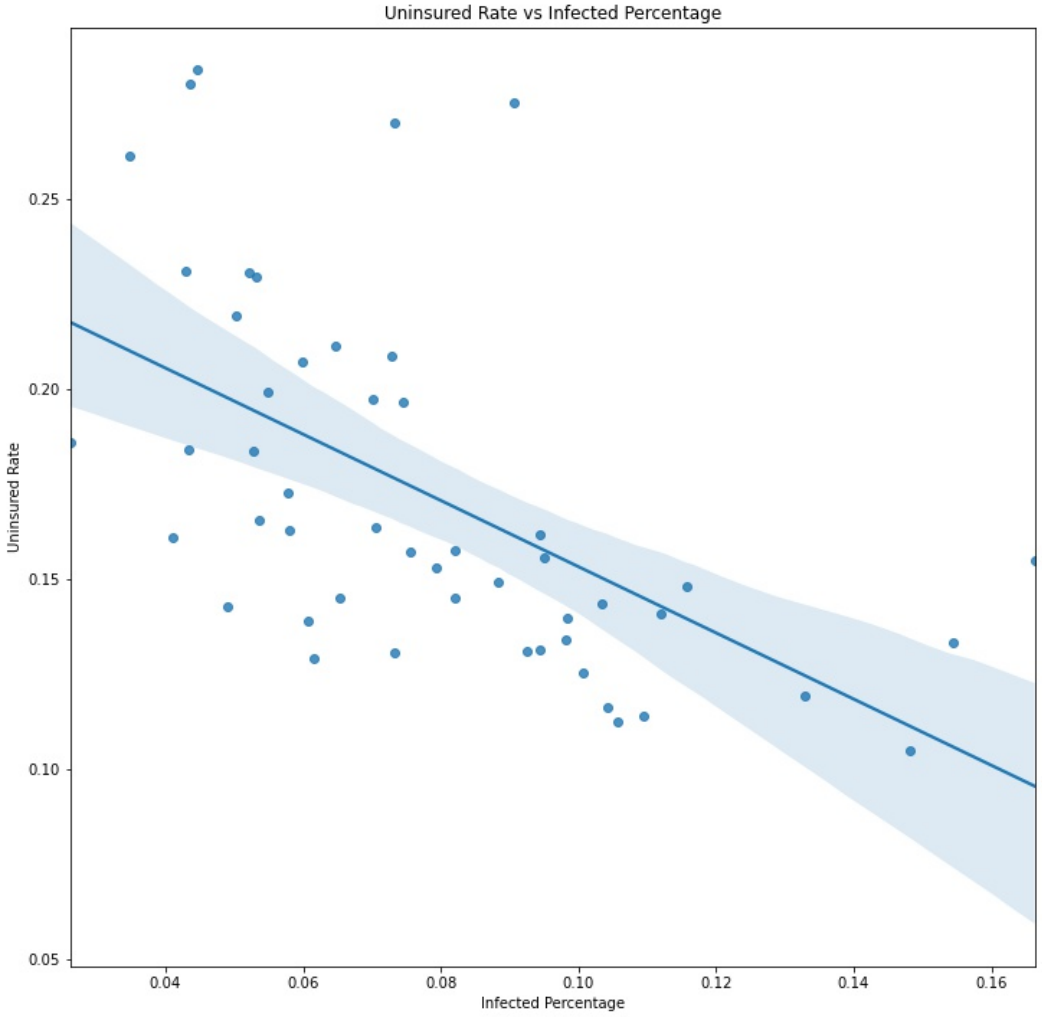

FIG. 24. Scatter Plot of Elderly Rate vs Infected Percentage With Regression Line and Confidence Interval

A scatter plot with a regression line shows a negative relationship between the Covid Infection Percentage and Elderly Rate.

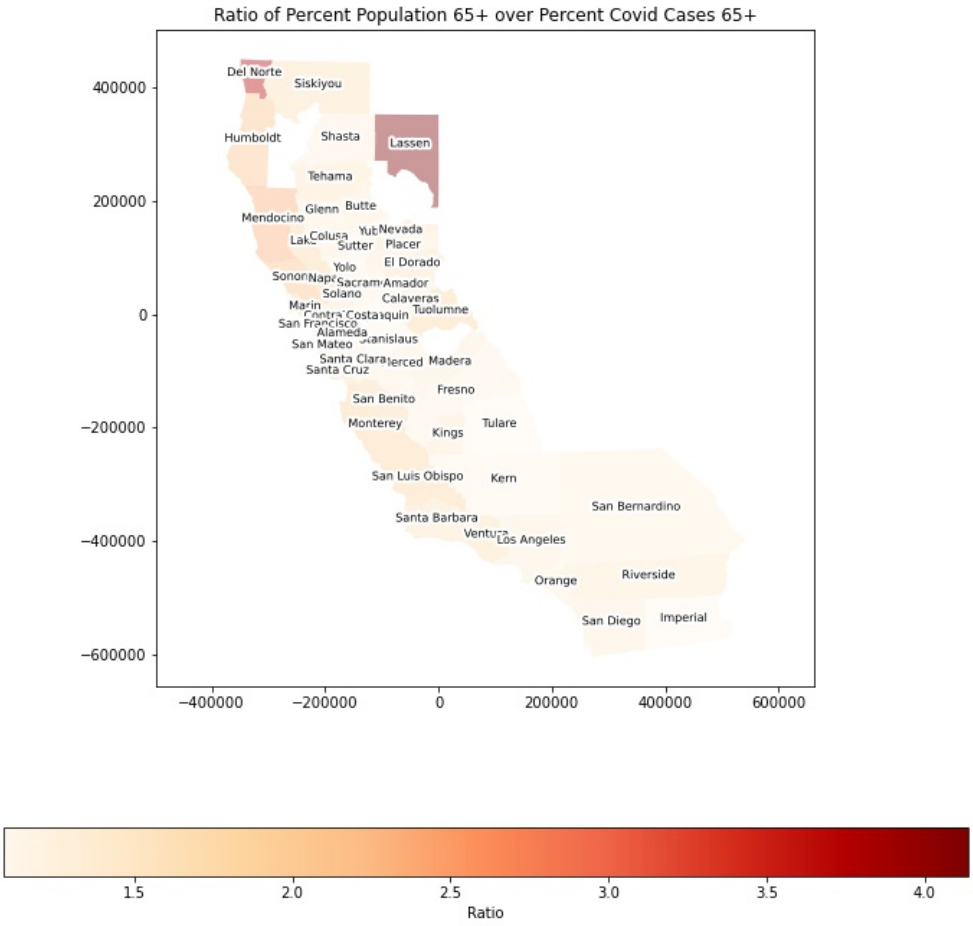

FIG. 25. Overrepresentation Ratio for Covid Cases for People 65+

Other than Lassen County, people age 65+ have low representation in Covid cases.

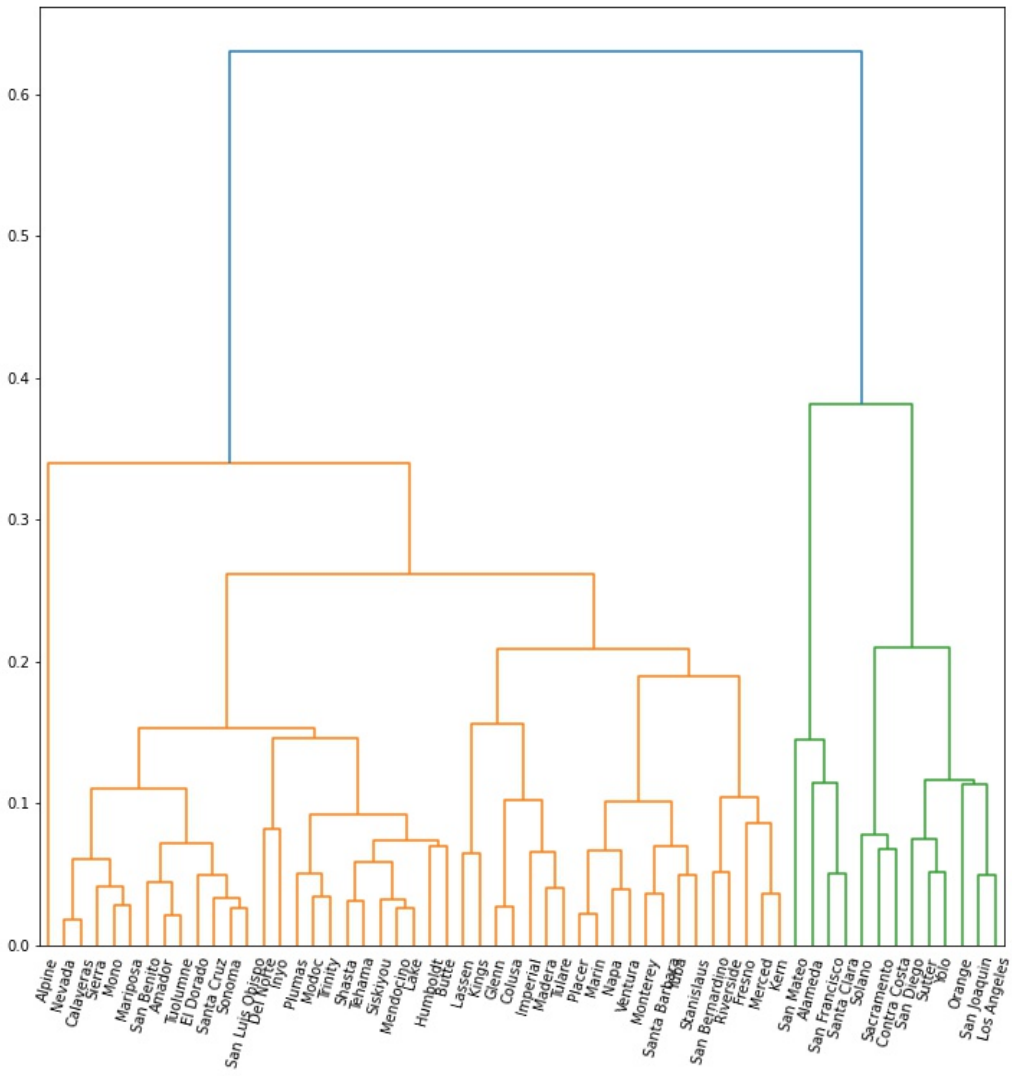

FIG. 26. A Dendrogram for Hierarchical Clustering on Indicators of Groups at Risk and Covid Infection Rate for California Counties

This Hierarchical Cluster was performed using the Farthest Point Algorithm on the the percent population of Whites, Asians, Blacks, Native Hawaiian/Pacific Islander, and Native Americans, poverty rate, uninsurance rate, and Covid infection rate.

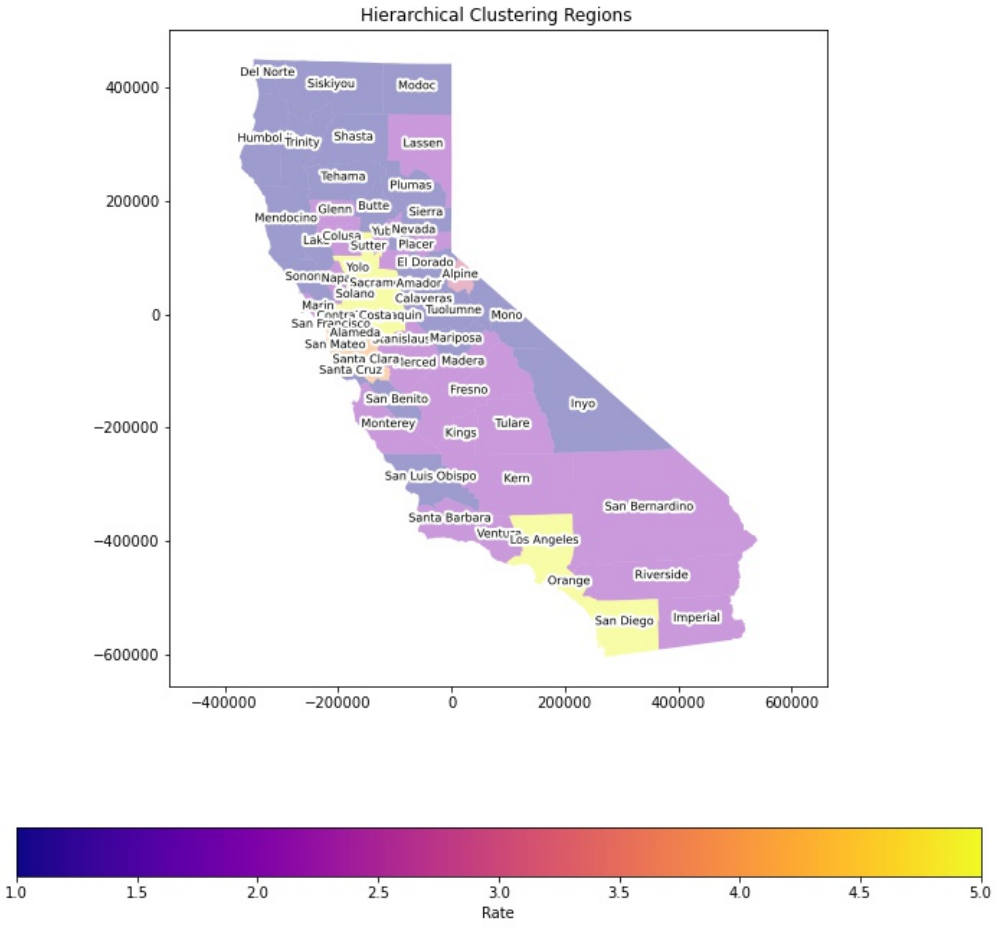

FIG. 27. Map of the Hierarchical Clusters when Restricted to Five Clusters

We notice that there are four distinct groups and one outlier county. The four groups correspond to a group of Bay Area Counties, a group of Northern California Counties, a group of Urban Counties around Sacramento and Los Angeles, and a group of Central and Inland Counties.

V. APPENDIX B: TABLES

TABLE I. Members of each Cluster from Hierarchical Clustering

Group Label	Members
Group 1	Del Norte','Siskiyou','Modoc','Humboldt','Trinity','Shasta','Tehama' 'Plumas','Butte','Mendocino','Sierra','Lake','Nevada','El Dorado' 'Sonoma','Mono','Amador','Calaveras','Tuolumne','Mariposa','Inyo' 'Santa Cruz','San Benito','San Luis Obispo'
Group 2	'Lassen','Glenn','Yuba','Colusa','Placer','Napa','Marin', 'Stanislaus','Madera','Merced','Fresno','Monterey','Tulare','Kings', 'San Bernardino','Kern','Santa Barbara','Ventura','Riverside','Imperial'
Group 3	'Sutter','Yolo','Sacramento','Solano','San Joaquin','Contra Costa' 'Los Angeles','Orange','San Diego'
Group 4	'Alpine'
Group 5	'San Francisco', 'Alameda', 'San Mateo', 'Santa Clara'

TABLE II. Mean Values of Each Cluster Within Features

Group	Inf %[a]	AA %[b]	WA %[c]	BA %[d]	NH %[e]	IAC %[f]	Pov %[g]	Unins %[h]
1	0.04	0.03	0.95	0.01	0.002	0.06	0.07	0.14
2	0.10	0.06	0.89	0.05	0.004	0.039	0.09	0.16
3	0.00	0.02	0.73	0.005	0.00	0.27	0.07	0.16
4	0.05	0.36	0.57	0.05	0.008	0.019	0.045	0.07
5	0.08	0.18	0.75	0.08	0.006	0.028	0.08	0.13

[a] Mean Covid Infection Percentage

[b] Mean Asian American Populace Percentage

[c] Mean White American Populace Percentage

[d] Mean Black American Populace Percentage

[e] Mean Native Hawaiian/Pacific Islander Populace Percentage

[f] Mean Native American/Alaskan Populace Percentage

[g] Mean Poverty Rates

[h] Mean Uninsurance Rates

VI. SUPPLEMENTARY MATERIAL

The Python code, as well as an in-depth explanation of the methods and packages used in the analysis, is provided in 'Covid19Analysis.ipynb'.

Figures used in report are found in the figures folder.

California County shapefile can be found in the data folder, courtesy of UC Berkeley GeoData Repository (https://geodata.lib.berkeley.edu/)

Census Data on Race and Age Statistics are found in cc-est2019-alldata-06.csv, courtesy of US Census (https://data.census.gov/cedsci/).

Poverty Rates for California Counties are found in poverty.csv, courtesy of Public Policy Institute of California (https://www.ppic.org/).

Ininsurance Rates for California Counties are found in Health-Insurance-Coverage.csv, courtesy of California Health Care Foundatation (https://www.chcf.org/).