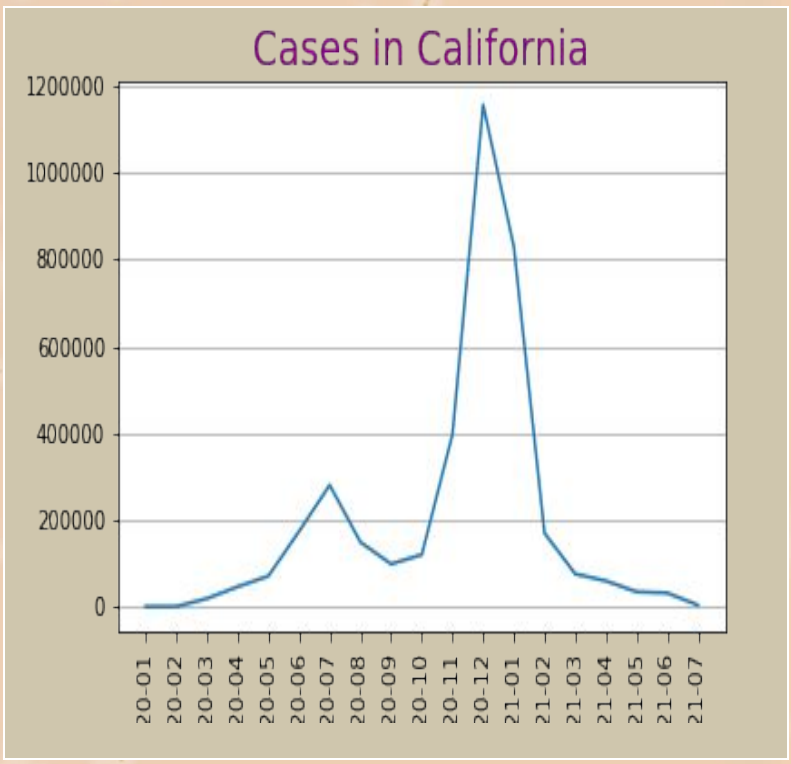
The Boolean Bee

no.203.078 | **The Hive: Erin Kinney, Kevin Gil, Henry Tirado** | **- Since 2020**

COVID-19 Special Edition



Measuring COVID's Impact on CA Sub-Populations

We sent our reporters out to collect reliable datasets for analysis to confirm or refute the news being reported about the COVID pandemic. Are the other news outlets being truthful? Are others downplaying the severity? Sound data analysis is needed.

California was selected as the state of choice to analyze. Our news readers want to know what impact COVID has had on California's sub-populations. In Los Angeles County, we investigated the reports that some race/ethnicity groups were disproportionately impacted by the pandemic. At the state level, we looked into the impact of the pandemic on distinct age groups, and crunched numbers to determine the impact on rural vs metropolitan counties.

Motivation & Summary

The impact of COVID is on the mind of a polarized nation. As people of science, we can access the same data that news outlets and the government use; we can confirm or refute their findings for the State of California.

- Did some races/ethnicities experience a disproportionately higher percent of deaths than others? The news reports that minority people of color were impacted much more than the majority population. We looked at Los Angeles County.

- How are our young, working, and retired ages faring? The State's current economy relies on a strong work force, its future economy depends on our youth, and our retired citizens contribute to the economy. We looked at California.

- Is there a large divide between the case counts from rural counties and those in metropolitan counties? Both types are critical to the State's agricultural and financial wellbeing. We looked at California.

Questions & Data

Our questions required us to understand what our State's population looks like. As the world's 5th largest economy, the State of California has thrived for years due to its population diversity. For this demographic county data, we went to three sources: CDC, County Health Rankings, and State of California.

-CDC COVID Tracker Data/COVID-19 Integrated County View
https://covid.cdc.gov/covid-data-tracker/#county-view

- County Health Rankings & Roadmaps
https://www.countyhealthrankings.org/app/california/2021/overview

- Statewide COVID-19 Cases Deaths Tests
https://data.chhs.ca.gov/dataset/f333528b-4d38-4814-bebb-12db1f10f535/resource/046cdd2b-31e5-4d34-9ed3-b48cdbc4be7a/download/covid19cases_test.csv

We needed disaggregated data about COVID cases in California that could be viewed by county. It helped us determine how disruptive COVID was.
https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data

Data Cleanup & Exploration

The dataset by the CDC contained public data on reported COVID case surveillance and geography across the United States.

- This dataset was selected for its number of columns of supporting data that would allow for multiple questions to be asked from the same sample size.
- The dataset was 3.7GB in size and could not be uploaded to Git with a free account. $5 was donated to Git for additional storage.

The following cleanup processes were made:

- Our analysis focused specifically on cases in California, so all other state data were removed

- Columns were narrowed down to only those relevant to our analysis:
  - Year and month
  - Case status (probable or confirmed)
  - Race & Ethnicity
  - State, county, & FIPS code
  - Death (yes or no)
  - Age group & gender

- The following changes were made to unknown or missing data fields:
  - 'age_group': removed "Missing" rows
  - 'sex': if "Missing" then "Unknown"
  - 'current_status': removed "Probable Cases" as we were interested in lab-confirmed cases
  - 'death_yn': removed "Missing" and "Unknown" case reports

- A unified ‘race/ethnicity’ column was needed as there was no consistency between the two columns. A five-step logic between the ‘race’ and ‘ethnicity’ columns was implemented to populate the unified column.

Data Analysis: Race/Ethnicity

Investigation 1: Impact of COVID-19 on particular race/ethnicity groups in Los Angeles County

- Was the COVID-19 infection fatality rate (IFR)* of a given race/ethnicity group disproportionate compared with the IFR of the whole of Los Angeles County?

- In Los Angeles County, was the percentage of reported COVID-19 deaths of a given race/ethnicity group disproportionate to the given group's percentage of the County population?

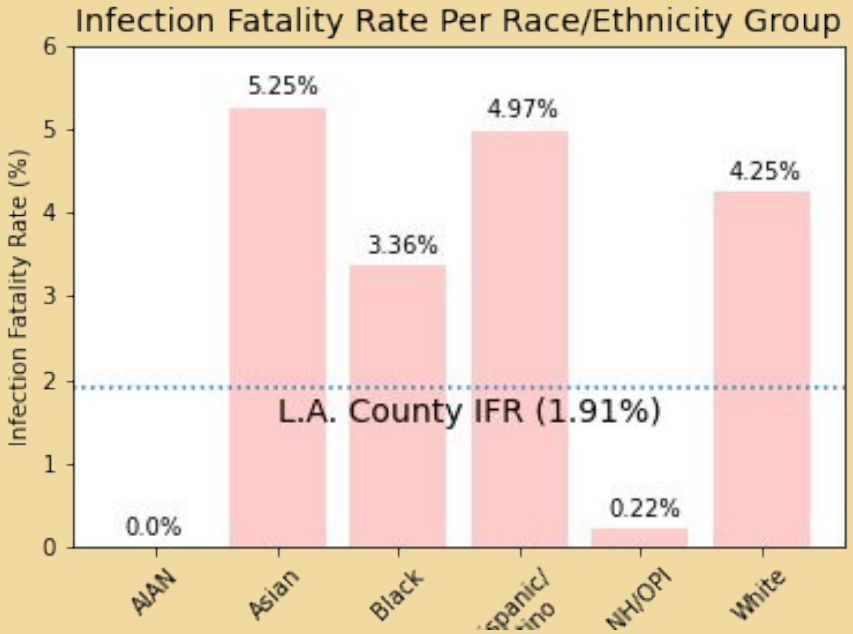
Data Analysis: Race/Ethnicity

Investigation 1: Analysis steps taken

- Calculate total reported cases and deaths in Los Angeles
- Calculate infection fatality rate (IFR) in Los Angeles
- Repeat steps 1 and 2 for six race/ethnicity groups:
  - American Indian & Alaskan Native
  - Asian
  - Black
  - Hispanic/Latino
  - Native Hawaiian/Other Pacific Islander
  - White
- Determine cumulative deaths over time for each race/ethnicity group
- Determine the percentage of deaths contributed by each race/ethnicity group to the L.A. County total
- Plot IFR for each race/ethnicity group against the IFR for L.A. County
- Plot the percentage of deaths per race/ethnicity against the group's population percentage

Data Analysis: Race/Ethnicity

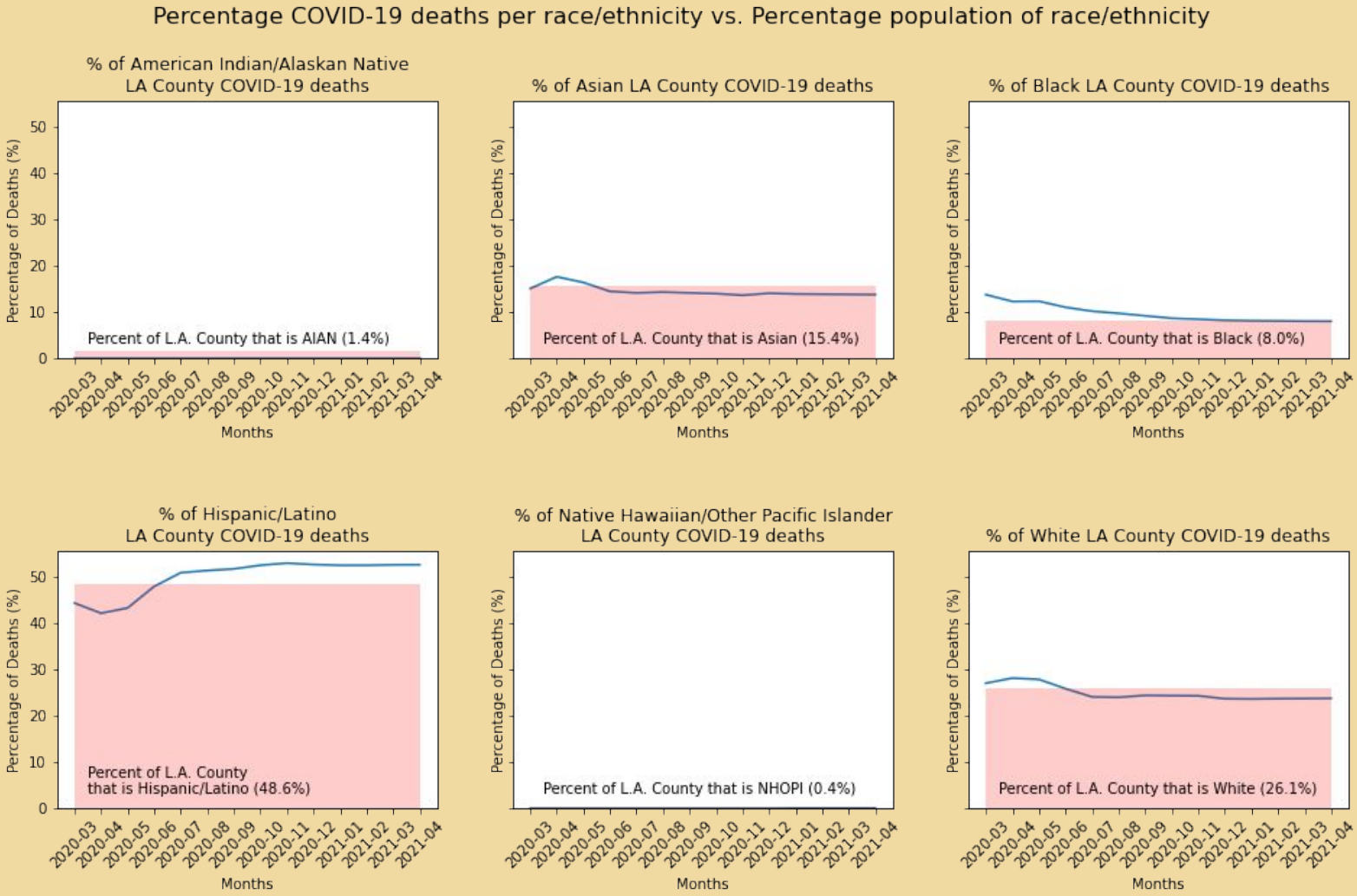
Investigation 1: Figure 1



** IFR is calculated on reported cases; once reported on actual cases, the rates will most likely be lower*

Data Analysis: Race/Ethnicity

Investigation 1: Figure 2



Discussion: Race/Ethnicity

Investigation 1: Races/Ethnicities

- The infection fatality rates (IFR) of the Asian, Black, Hispanic/Latino, and White populations are higher than the IFR of Los Angeles County; however, it is not clear if this measure alone indicates a greater impact from COVID-19 on these groups.

- The Black and Hispanic/Latino populations of Los Angeles County had the longest history (11 months and 10 months respectively) of having their percentage of COVID-19 deaths above their population percentages, indicating a greater impact from COVID-19 than other race/ethnicity groups.

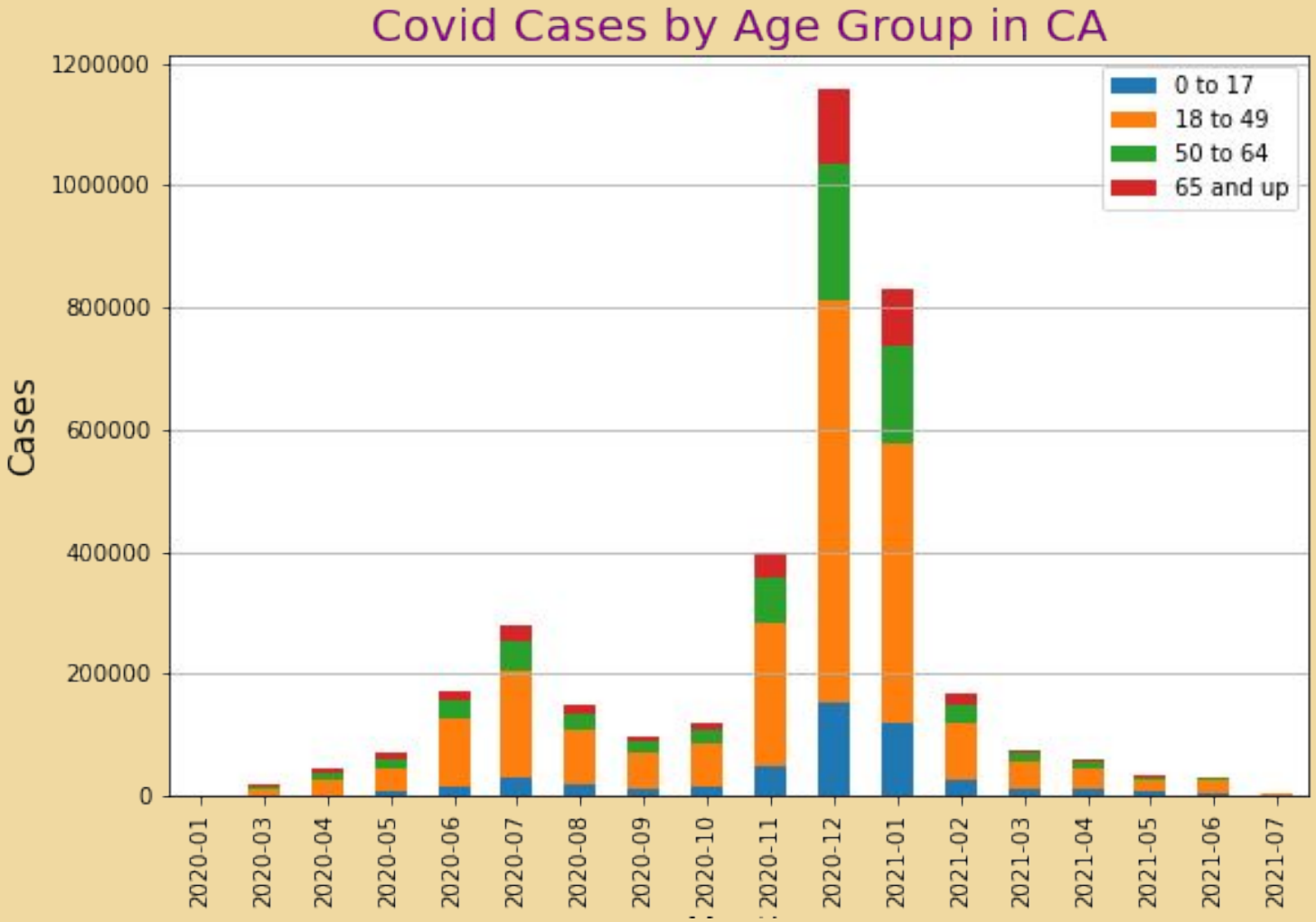
- The higher IFR coupled with the disproportionate percentage of COVID-19 deaths among the Los Angeles County Black and Hispanic/Latino populations, indicate a greater impact on these populations from COVID-19 than on other race/ethnicity groups.

Data Analysis: Age Groups

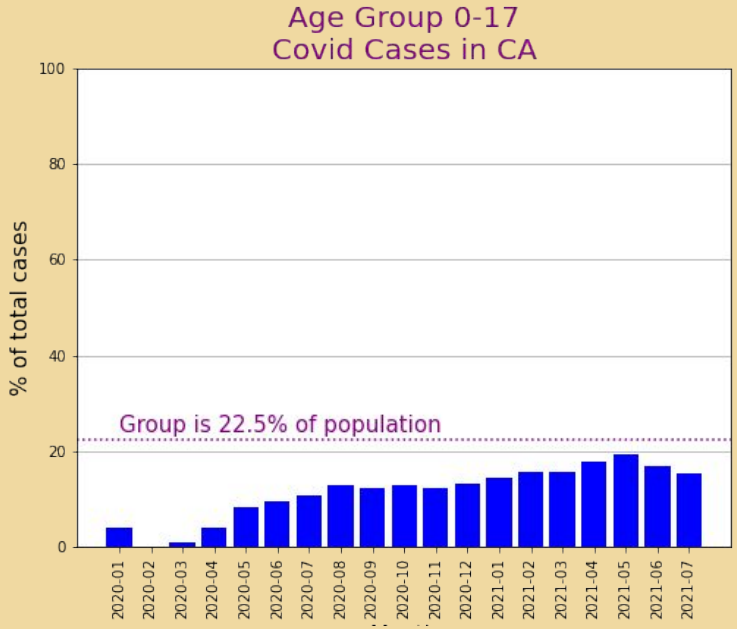
Investigation 2: Analysis steps taken

- Demographics dataset
  - Convert percentages of population into actual quantities
  - Extrapolate population quantities grouped by county and age groups
  - Get sum and percentages for age groups
- COVID Cases dataset
  - Group 4 age groups into 3 to match demographics set
  - Group the cases by month, age group; count the cases
  - Calculate percentage of total cases per month for each age group
  - Unite into one percentage dataframe suitable for plotting
- Merge resulting data
  - Match cases to demographics statistics by age group
- Plot % of cases per age group by month
  - Each group cases by percentage of total cases per month. Add constant line referencing the percentage of population.

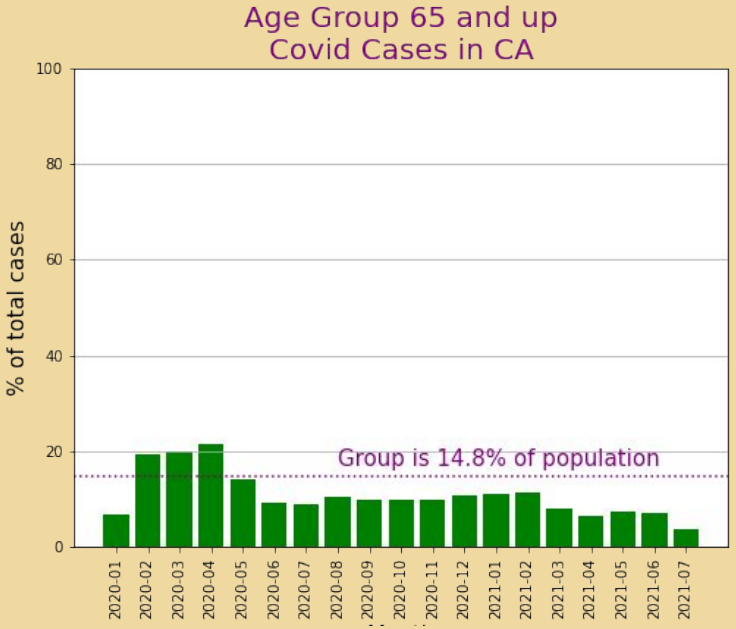
Data Analysis: Age Groups



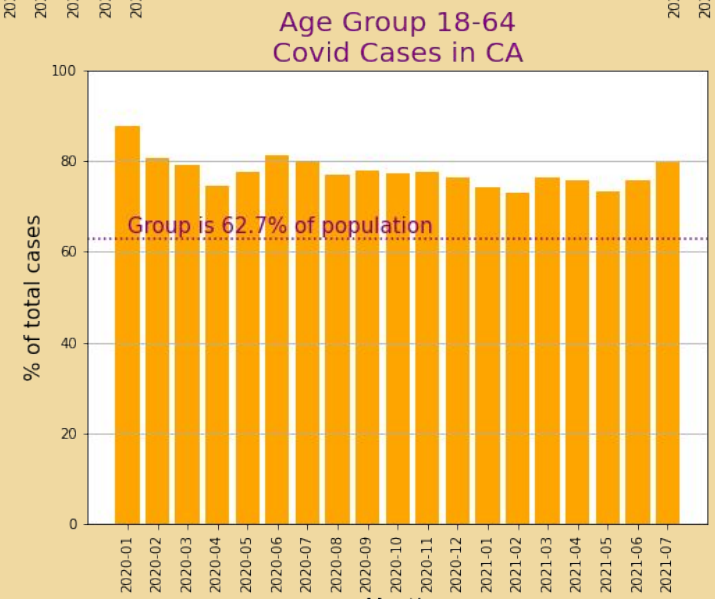
Data Analysis: Age Groups



Least impacted but rising until vaccine availability



Initially heavily impacted but dropped with safety guidelines



Most impacted continuously

Discussion: Age Groups

Investigation 2: Age Groups

- The State's working age group 18-64 was and remains the most disproportionately impacted over the youth 0-18 years and seniors over 65. California cases have fallen since February '21; but working age group bears the brunt of cases.

- In early 2020, seniors over 65 were initially disproportionately impacted for 2 months before measures were enacted to reduce their monthly case counts.

- The youth 0-18 group has been least impacted but was slowly trending up. When the vaccine became available to the older members of their age group, that rising trend began to show a decrease in their percentage of monthly cases.

Data Analysis: County Density

Investigation 3: Is there a large divide between the case counts from rural counties and those in metropolitan counties?

Analysis steps taken: Our demographic dataset consisted of counties and their metropolitan classification as either being "Metro" or "Non-metro". We also had the number of cumulative COVID cases through 7/31/2021 by county. A choropleth map of California could identify which counties classified as metro or non-metro, as well as identify where most of the cases were found.

On a pure number basis, we clearly see that most cases are found in metro counties

Classification	# of Cases
Metro	3,825,761
Non-metro	54,094

Data Analysis: County Density

Top 15 metro counties

County	Classification	Cases	Percent
Los Angeles	Metro	1,260,502	32.49%
San Bernardino	Metro	302,942	7.81%
Riverside	Metro	302,671	7.80%
San Diego	Metro	298,999	7.71%
Orange	Metro	265,978	6.86%
Santa Clara	Metro	123,617	3.19%
Sacramento	Metro	117,162	3.02%
Fresno	Metro	105,445	2.72%
Kern	Metro	99,539	2.57%
Alameda	Metro	96,932	2.50%
Ventura	Metro	84,114	2.17%
Contra Costa	Metro	77,490	2.00%
San Joaquin	Metro	77,403	1.99%
Stanislaus	Metro	59,114	1.52%
Tulare	Metro	50,626	1.30%

Top 15 non-metro counties

County	Classification	Cases	Percent
Nevada	Non-metro	5,604	0.14%
Lassen	Non-metro	5,338	0.14%
Tehama	Non-metro	5,260	0.14%
Humboldt	Non-metro	5,155	0.13%
Tuolumne	Non-metro	4,522	0.12%
Mendocino	Non-metro	4,508	0.12%
Lake	Non-metro	4,372	0.11%
Amador	Non-metro	3,925	0.10%
Glenn	Non-metro	2,476	0.06%
Calaveras	Non-metro	2,344	0.06%
Colusa	Non-metro	2,016	0.05%
Siskiyou	Non-metro	1,991	0.05%
Del Norte	Non-metro	1,648	0.04%
Inyo	Non-metro	1,455	0.04%
Mono	Non-metro	1,295	0.03%

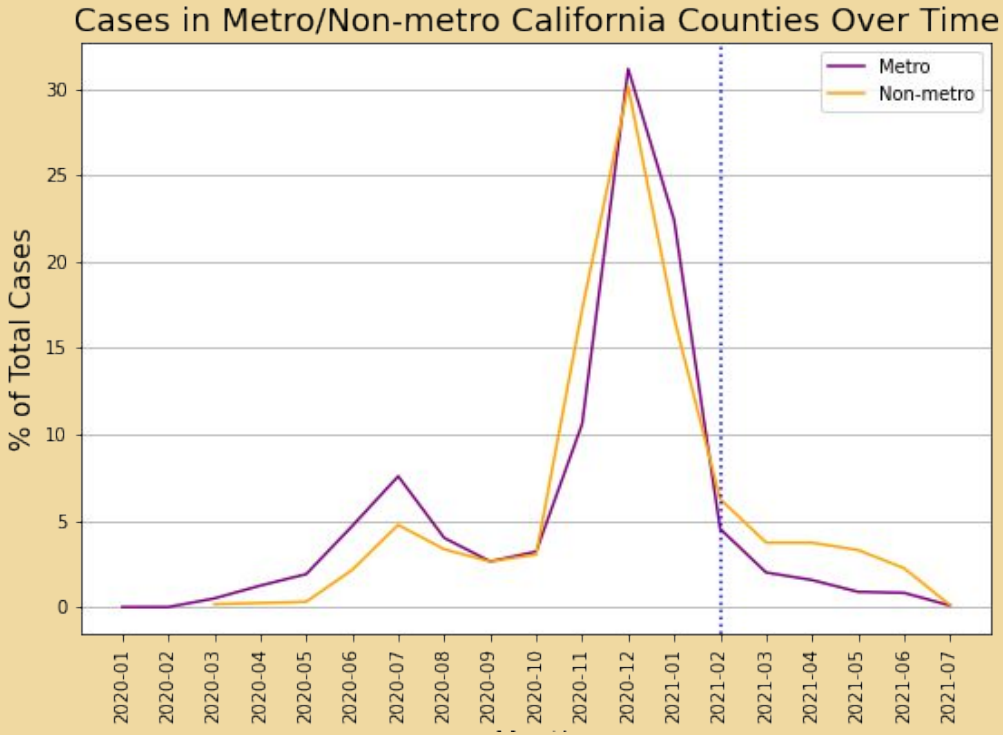
Classification	Total Population	Cases	Cases %
Metro	38,674,939	3,825,761	9.89%
Non-metro	837,284	54,094	6.46%

- When we look at the individual metro and non-metro case counts based on their respective county populations, we see a relatively even disbursement amongst the county classifications.
- Metro cases were only about 3.5% higher than those of more rural counties.

Data Analysis: County Density

Investigation 3: What if we looked at cases in metro/non-metro counties over time? Could we see either showing more COVID infection rates throughout different parts of the pandemic?

Analysis steps taken:
Our main dataset contains cases by county over the span of 2020 through 7/31/2021.
By merging our other demographic dataset with the metropolitan classifications into our main dataset, we would be able to see COVID impacts over time for metro and non-metro counties.



- COVID cases across metro and rural areas have been relatively consistent throughout the pandemic.
- There have been slightly more cases in metro areas, but both have seen similar spikes in the last quarter of 2020.

- One observation was that case counts in metro areas have always been equal to or greater than those in rural areas, except for beginning February 2021.
- This was roughly the same time the first vaccines were made available to the public, and we see for the first time that % of cases in metro areas were actually below those of rural areas.

- Were vaccines more readily available in metro areas compared to rural?

Discussion: County Density

Investigation 3: County Density

- Our initial thoughts on the case counts across metro and rural county analysis was that infection rates would have been significantly higher in metro compared to rural. What we found was that the disbursement of COVID cases between the two were relatively even, with metro showing slightly higher infection rates.

- Additionally, with rural counties showing a higher infection rate starting February 2021, it seems to allude to the overall availability of COVID vaccine sites in more population dense areas as well as the effectiveness of the COVID vaccines.

Post Mortem

Difficulties:

- **Investigation 1:** We did not have demographic information for the race/ethnicity groups of Mixed/Other, Unknown, and Non-Hispanic

- **Investigation 2:** We had to consolidate 4 age groups in the cases dataset to 3 age groups to match the demographics. That prevented us from looking at 18-49-age and 50-64-age groups separately.

Additional questions:

- We wanted to further explore each county's death rates among different races/ethnicities to see if there are any trend differences.
- We wanted explore whether median household income for race/ethnicity groups had a correlation to death rates per county.
- Our research could extend to comparing California with other States like Florida, with its different approach to containing the spread.
- By adding data relating to the vaccination availability and rate, we could have analyzed the vaccine's effectiveness in flattening the curves for each of the questions we posed.

Questions