A7 - Final Report

Alyson Suchodolski

Introduction

For the past two years, we have been inundated by COVID-19 studies, trying to find the missing link that could help cull the virus. Many medical professionals, politicians, and laymen have their ideas on what could work best, or whether we need something to work at all, but the public opinion of this pandemic has been polarizing nonetheless. News outlets, scientific publications, and op-eds offer their insight into this problem, and I feel as though most avenues have been explored when it comes to transmission, protection, and immunity. Despite the myriad of studies available to the general public, including on how vaccinations, mask mandates, and lockdowns affects the spread of COVID and how the pandemic has impacted the normal lives of billions of people, I thought it would be fruitful to run my own analysis on the data of the virus. We often see analysis on a wide scale, such as on a state, national, or global level. I think breaking it down into smaller segments can provide intuition on how mandates and policies affect the population medically and socially, and if different counties receive different benefits from solutions or lack thereof.

I believe looking at whether vaccinations affect transmission in San Diego County can offer the public an idea on how we can help control this pandemic on a smaller scale across the country. On top of this, comparing San Diego County to other counties in California can give us an idea on what's working and what isn't working when it comes to prevention and protection.

Background/Related Work

As seen in the title, my research questions involve whether vaccinations affect transmission rate in San Diego County, and how San Diego County compares to other counties in California. My hypothesis is that vaccinations do decrease transmission rate in San Diego County. On top of this, I believe that San Diego County will have a lower transmission rate as well as a higher vaccination rate when compared to other counties in California.

To address my first hypothesis, I did some research into the studies that already exist analyzing the link between vaccination and COVID risk. On the CDC's page, you can find the many benefits of vaccination, such as "COVID-19 vaccination remains highly effective against COVID-19 hospitalization and death caused by the Delta variant of SARS-CoV-2." This indicates that there is evidence that the vaccine is helpful when it comes to breakthrough cases, but how does it affect transmission between people? Further in the page linked here, we see that "data show fully vaccinated persons are less likely than unvaccinated persons to acquire SARS-CoV-2…" There is literature available on the effectiveness of vaccinations, so I find it to be interesting to look at it on a local level.

For the second hypothesis, I have had more trouble finding scholarly work on comparing counties, let alone counties in California. Despite this, many institutions, such as John Hopkins University and COVID Act Now, offer numbers and graphics at a national, state, and county level. In fact, COVID Act Now is where I pulled my data for analysis from, which will be explained in further detail in my README. What I do know is that, based on Assignment 4 of this class, San Diego County had a mask mandate for most of the pandemic. On top of this, news from California indicates a masked mandate is going back into effect as Omicron-variant becomes more of an issue. San Diego County had a high rate of following mask mandate policy and an economic lockdown was in effect until June 15th, 2021. These facts from San Diego County as well as California lead me to believe San Diego County is better prepared for COVID-19 prevention than other counties in California.

Methodology

I used multiple different methods to explore my data and compare my points of interest. These methods include an exploratory analysis involving time-series data, generating heat maps of vaccination and case rate, as well as hypothesis testing using two-sided t-tests.

Exploratory Analysis - Cases:

A general exploratory analysis was done to create time-series graphs of the cumulative cases, daily cases, and daily case rate of COVID-19 in San Diego County as well as the other Californian counties. I did a similar exploration of total vaccinations, daily vaccinations, and daily rate of vaccinations as well. California has 58 counties which made for cluttered graphs, so I then looked at only the most populated counties in California. The counties that made the list include any one with a population over one million citizens. You can see the results in the timelines below.

This method of analysis was chosen because it provided an overarching theme to what I wanted to look at. There is no better way to look at COVID, in my opinion, than over time. We get to see the ebb and flow of cases as the pandemic progresses, and significant peaks and valleys point us in a direction for further exploration. In this particular study, these peaks and valleys indicate people who have potentially suffered from the virus. We don’t want peaks when it comes to cases, we want to see these people happy and healthy.

Exploratory Analysis - Vaccinations:

A similar exploration of cases was done on completed vaccinations. Once again, timelines give us information on the progression of vaccine rollout. In this case, people who receive the vaccine cannot receive another vaccine (although boosters are now out and about), so this creates a certain bias in my data. The rates might be lower as time goes on, not because less people are getting the vaccine, but because the amount of people who can receive the vaccine is less. This could be remediated if I were to subtract the number of people who were vaccinated from the total population available.

For this section, I decided to gloss over looking at these time plots with all the counties included. If you would like to see those charts, you can find them in the repository under ‘Images’.

Heat Maps:

Heat maps were generated in Tableau and used to compare counties’ mean vaccination rates as well as mean case rates. I chose this method because it gave a look at the geographical implications of this study. Are there clusters of counties with higher or lower rates? On top of this, it gives us a view of all the counties together, which allows for easy comparison when using color. I chose red and green shading for the coloring, which I realize now could cause issues for those who are color-blind, so that is something I will have to keep in mind for the future.

T-Testing:

A t-test is used to infer if there is a significant statistical difference between the means of two groups. I used t-tests to compare the daily rate of covid in San Diego county before and after Vaccine Rollout. I also conducted this same test to take into consideration when a lockdown was in place. I also used t-tests to compare the daily case rates and vaccination rates of San Diego county compared to other counties. T-tests are great to use when comparing means, but they lack the inference of how the means could be different. For example, we might find that San Diego County’s vaccination rate is different from Los Angeles County’s, but we won’t know how it’s different unless we take a look at the numbers. There are also multiple assumptions that must be checked when using a t-test.

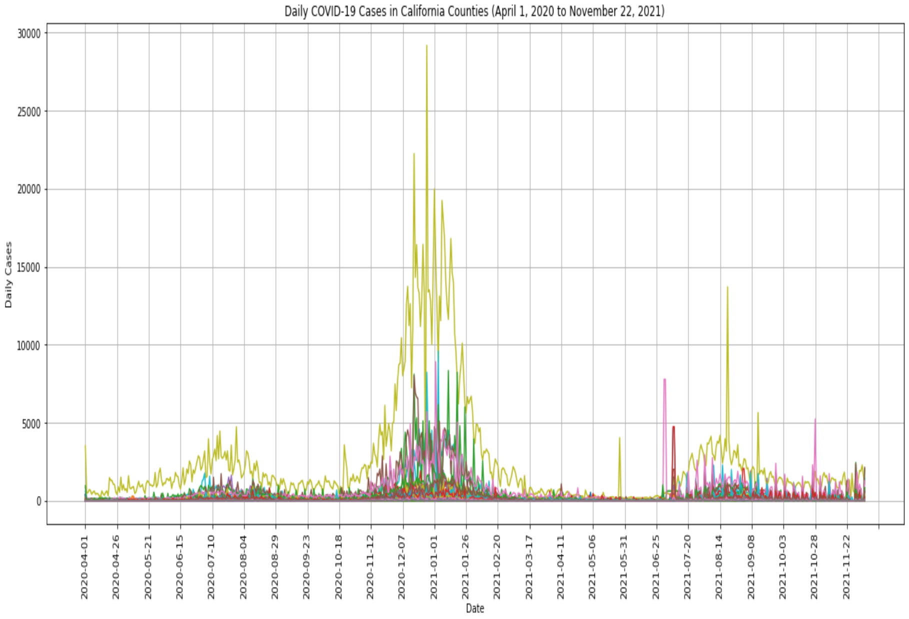
1. The scale of the data from the t-test must be continuous or ordinal. Our data is continuous.
2. The data must be a random sample of the population. Random samples were taken from the necessary features when performing all my t-tests.
3. It is assumed that the data is normal.
4. The random sample must be large enough, which will then ensure a normal bell-shaped curve. A sample size of 50 was taken for each t-test.
5. Finally, variance must be equal. The function I used for my t-tests does take into consideration unequal variances and can be toggled accordingly.

Findings

Exploratory Analysis

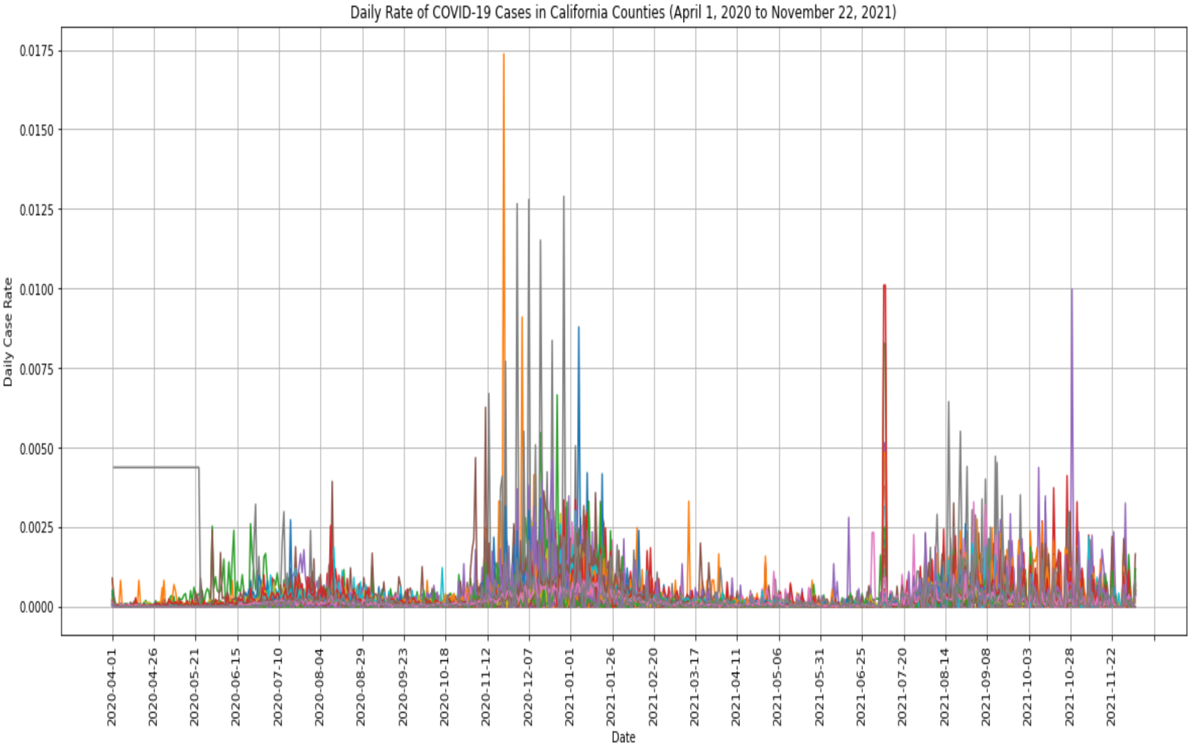
Daily Covid-19 Cases in California Counties:

The yellow line is representative of Los Angeles county, which is the most populated county in California and the United States, with about 10 million citizens. San Diego County is the second most populated county in California, with about 3.3 million citizens. While interesting, this graph does not provide relevance between counties, so I decided to divide these daily numbers by the population of the respective county to create our next graph.



Daily Rate of Covid-19 Cases in California Counties:

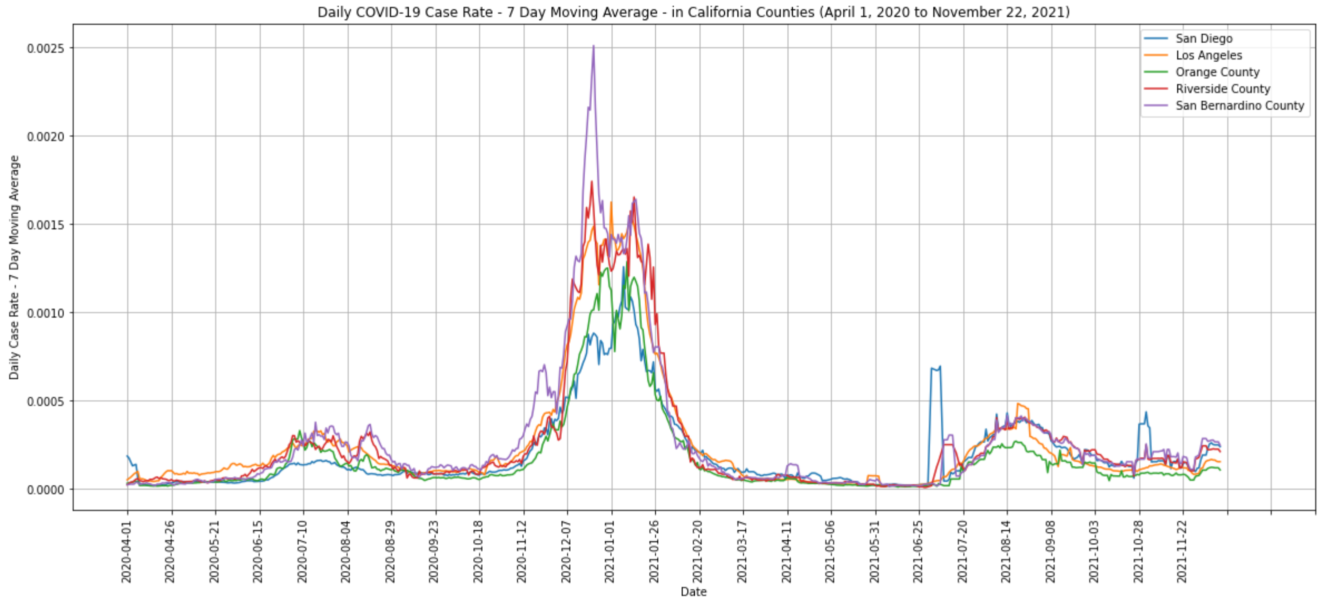
Below is a graph of the daily rate of COVID cases and the lines become a lot less distinguishable. This leads me to believe the rates of COVID cases in each county is a lot closer than I thought in spite of the cumulative numbers. Of course Los Angeles county would previously seem like a rough county to be in for COVID because of its high cumulative and daily cases, but we had to take into consideration the population. This creates rates we can more readily compare. But the graph is still messy and unreadable, so I took a look at the most populated counties.

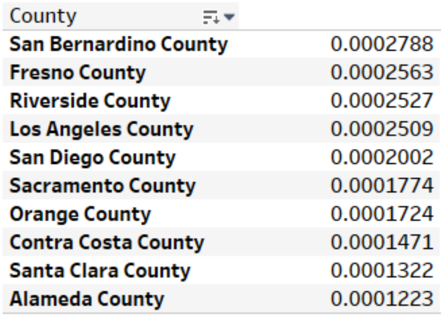


Daily Rate of COVID-19 Cases in California Counties (April 1, 2020 to November 22, 2021)

Daily Rate of Covid-19 Cases in California's Most Populated Counties:

While we look at 10 counties during our statistical analysis of the data, the focus of this graph is on the top 5 most populated counties in California. A 7-day moving average was taken of the daily case rate to smooth out the lines. We see in blue San Diego County, a line that falls below the others in the graph during the period with the largest spike. That purple spike is San Bernardino County, topping off this chart with a 0.0025 daily case rate at the end of 2020. Below the graph, you can find a table of the mean daily rate of cases for each county, listed in descending order. San Diego County comes in fifth.

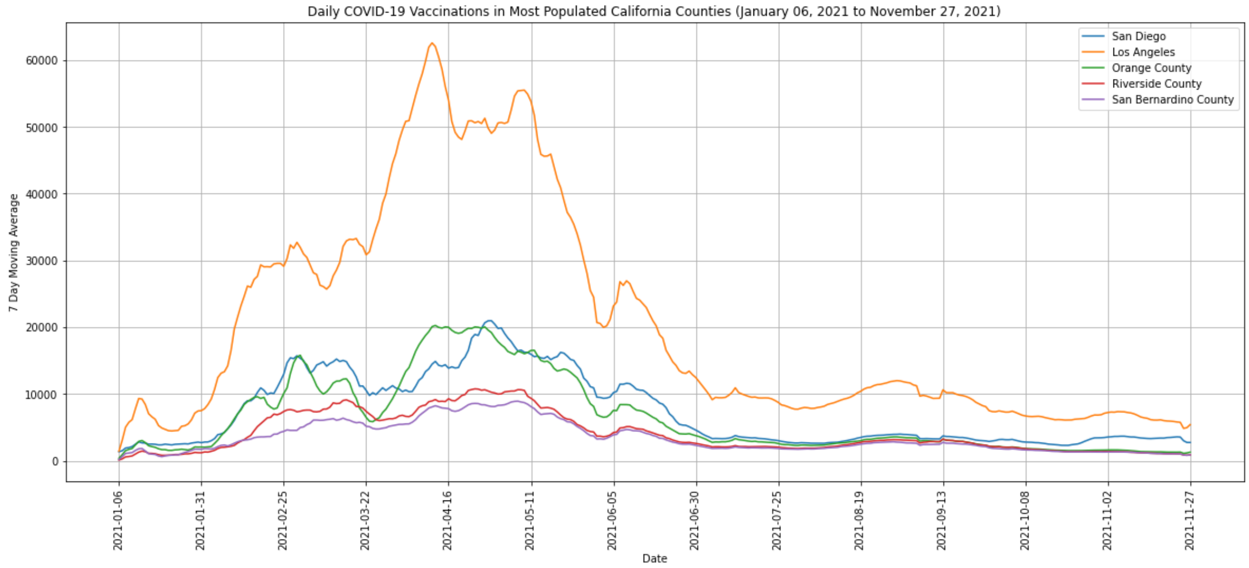




County	
San Bernardino County	0.0002788
Fresno County	0.0002563
Riverside County	0.0002527
Los Angeles County	0.0002509
San Diego County	0.0002002
Sacramento County	0.0001774
Orange County	0.0001724
Contra Costa County	0.0001471
Santa Clara County	0.0001322
Alameda County	0.0001223

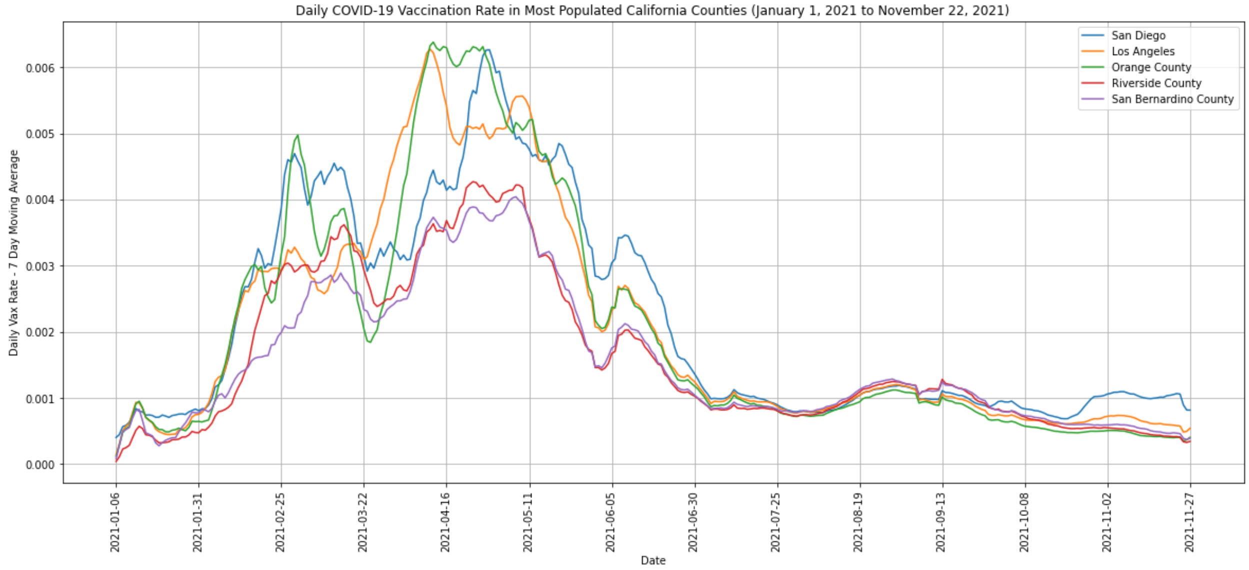
Daily Covid-19 Vaccinations in California’s Most Populated Counties:

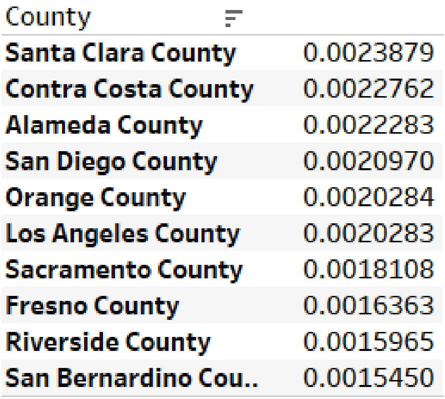
We see that Los Angeles County has a high number of daily covid vaccinations daily, as compared to the other highly populated counties. In this case, we have to take into consideration each county’s population so that the number can be standardized.



Daily Rate of Covid-19 Vaccinations in California's Most Populated Counties:

Now we see that Los Angeles' daily rate of COVID vaccinations fall in line with the rates from the other counties. San Diego County seems to pick up yet again toward the end of this year, and seems to be relatively in the middle when it comes to vaccination rates. Below this chart, I have created a list of the means of the daily vaccination rates of each county. San Diego county comes in fourth.





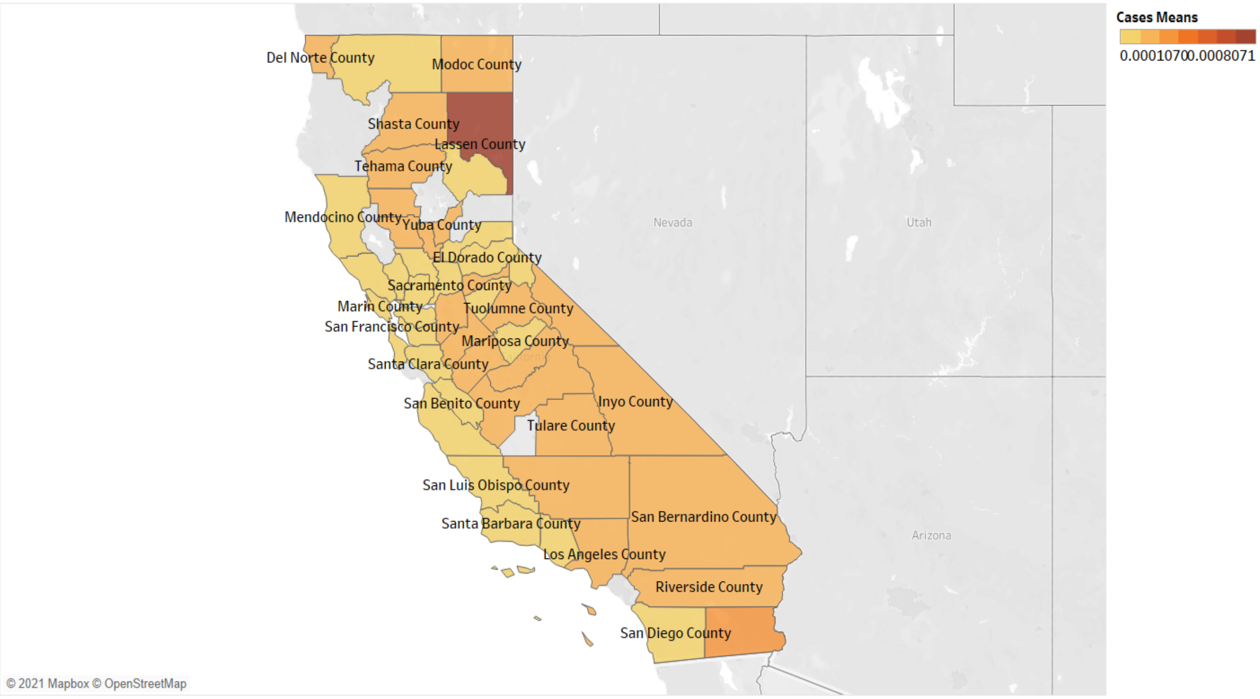
County	
Santa Clara County	0.0023879
Contra Costa County	0.0022762
Alameda County	0.0022283
San Diego County	0.0020970
Orange County	0.0020284
Los Angeles County	0.0020283
Sacramento County	0.0018108
Fresno County	0.0016363
Riverside County	0.0015965
San Bernardino Cou..	0.0015450

Heat Maps:

Case Rate:

The heat map below provides information on the daily case rate of each county in California. In the bottom left is San Diego County, which is colored yellow. This means that it is one of the counties that falls lower on the scale of the case rate. An interesting tidbit is Lassen County, which seems to have a high rate. This county's population is made of 3 prison systems and is mostly a ghost-town.

Cases

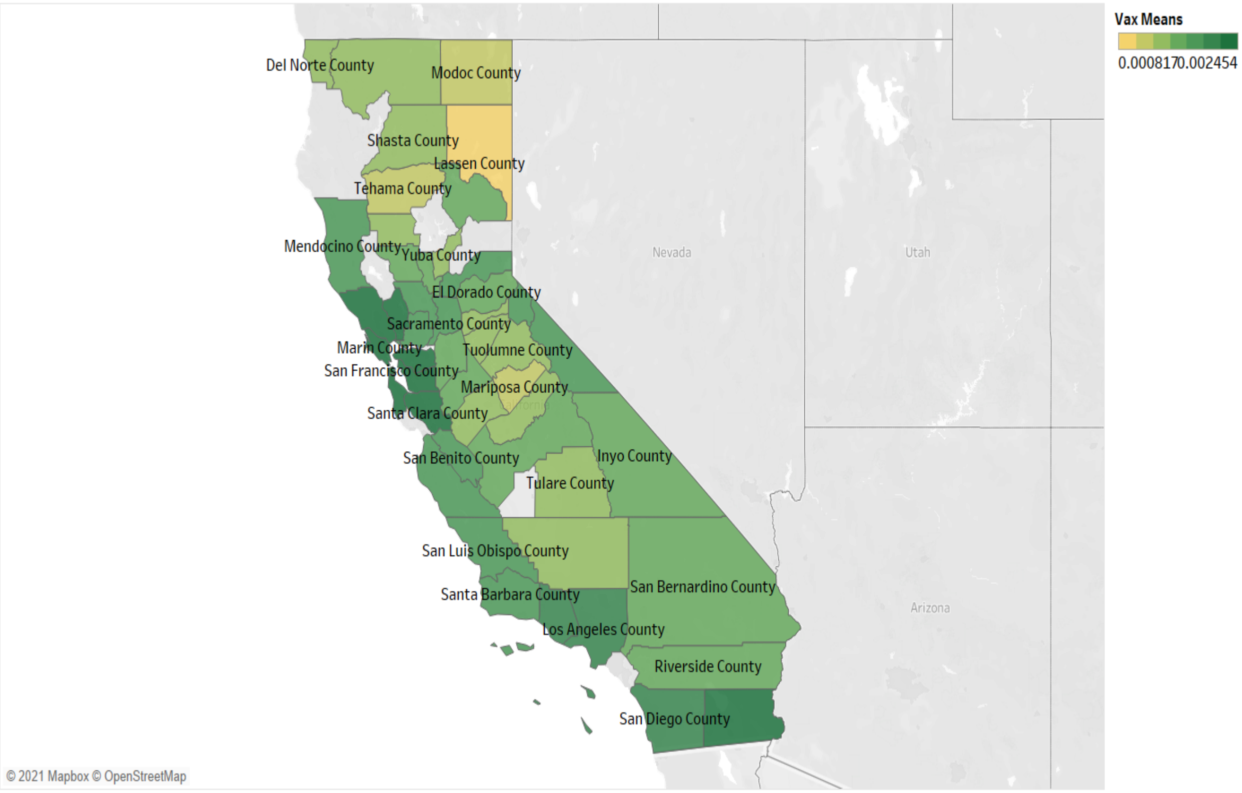


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Cases Means. The marks are labeled by County. Details are shown for County.

Vaccination Rate:

Once again, you can see San Diego County in the southwest corner of the map. A darker green area indicates a higher daily vaccination rate, so San Diego County is doing pretty well relative to other counties in that regard.

Vaccinations



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Vax Means. The marks are labeled by County. Details are shown for County.

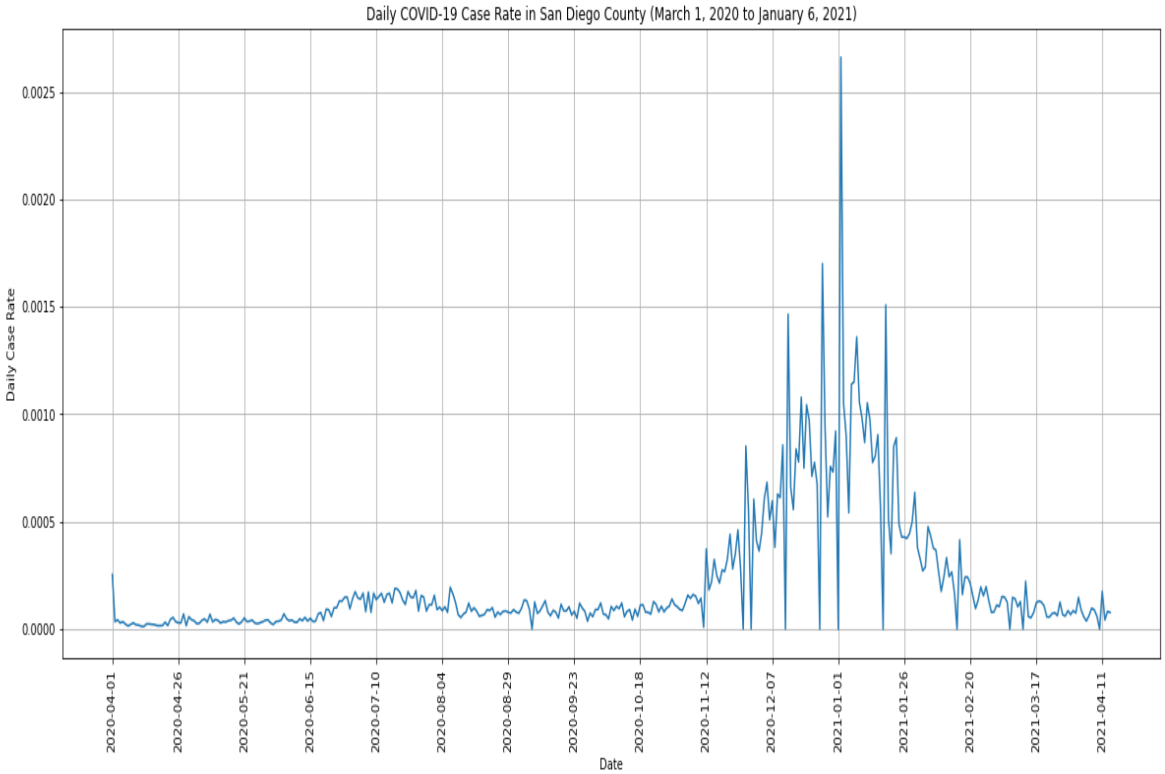
T-Testing:

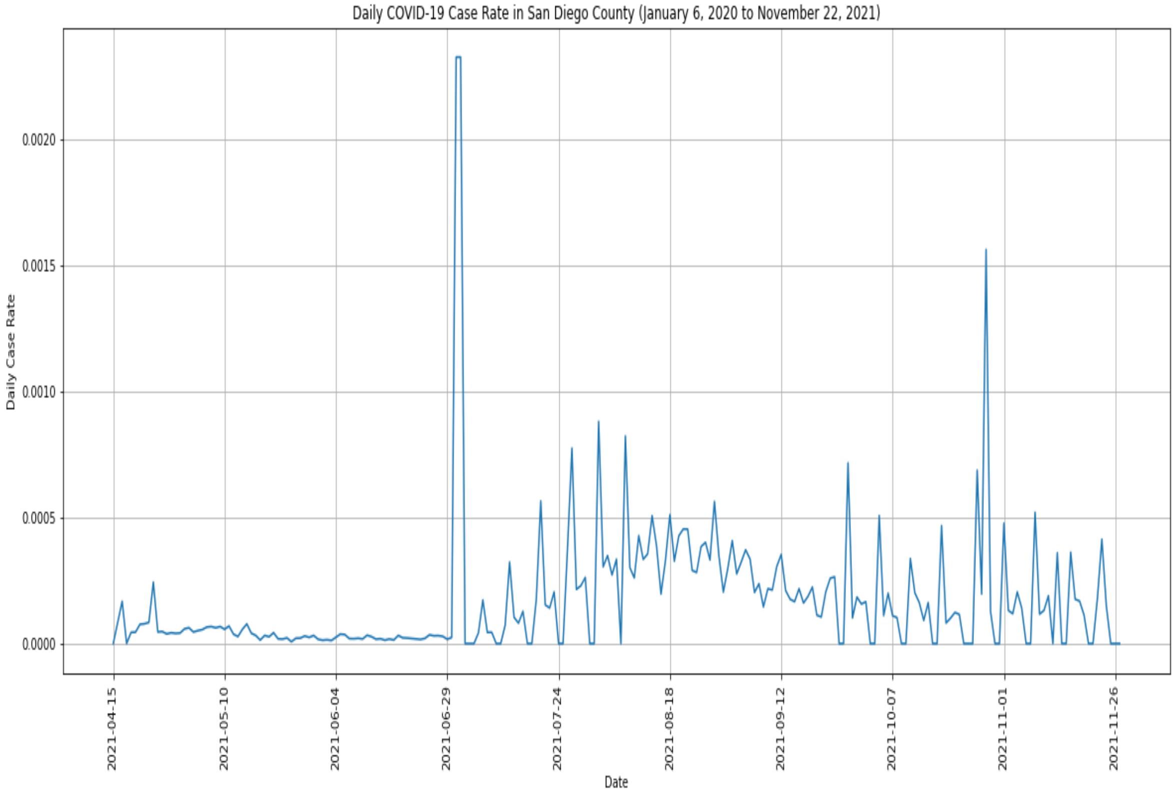
Daily Rate of Covid Cases Before and After Vaccine Rollout:

Null Hypothesis: The mean case rate of San Diego County is the statistical the same before and after the vaccine rolled out April 15, 2021’

Result: We would not reject the null hypothesis.
This was a surprising find because based on the literature I read, I was sure that the vaccine would have an effect on Covid infection rate. This made me question the integrity of the tests as well as the data. I have total faith in Covid Act Now to accurately report information, so I dug deeper into the data to get a better look at the information

before me. I created graphs highlighting before and after the Vaccine Rollout in San Diego County.


Daily COVID-19 Case Rate in San Diego County (March 1, 2020 to January 6, 2021)


Daily COVID-19 Case Rate in San Diego County (January 6, 2020 to November 22, 2021)

From these two graphs, we can compare the rates. I saw there is a flat portion before some volatility in the graph portraying vaccine rollout. I did some research and found that San Diego County opened back up its economy on June 15, 2021. If we take into account a week of lag time for the virus to start spreading again, it lines up with that huge spike in the second graph. I then conducted t-tests based on three different time periods: Before Vaccine Rollout and During Lockdown, After Vaccine Rollout and During Lockdown, and After Vaccine Rollout and After Lockdown.

Daily Rate of Covid Cases Before and After Vaccine Rollout + Lockdown In Place:

If we take a look at the three scenarios presented above, I did a piecewise comparison of the daily case rate means. I was looking to see if there was any statistical difference in means of the scenarios. I highlight the hypotheses and the results in the table below:

Null Hypothesis	**Result**
No Vaccine + Lockdown vs Vaccine + Lockdown	Reject the null hypothesis.
No Vaccine + Lockdown vs Vaccine + No Lockdown	Do not reject the null hypothesis.
Vaccine + Lockdown vs Vaccine + No Lockdown	Reject the null hypothesis.

These results indicate that there would still not be a difference in covid cases whether a vaccine was in place without a lockdown or if there was no vaccine but a lock down was in place. I find this to be a telling result that will be explained later in ‘Implications’.

Case Rate in San Diego Counties vs. Other Counties:

I compared San Diego County’s mean case rate to the other nine most populated counties in California. I was looking to find if there was a statistical difference between the two counties, and what that could mean when taken into consideration with the other facts and figures. A table of the null hypotheses and the results can be found below:

Null Hypothesis - What County We Are	**Result**

Comparing	
Los Angeles	Reject the null hypothesis.
Orange	Do not reject the null hypothesis.
Riverside	Do not reject the null hypothesis.
San Bernardino	Reject the null hypothesis.
Santa Clara	Do not reject the null hypothesis.
Alameda	Reject the null hypothesis.
Sacramento	Do not reject the null hypothesis.
Costa Contra	Do not reject the null hypothesis.
Fresno	Reject the null hypothesis.

The results indicate that 5 of the 9 counties could potentially have a statistically similar case rate to San Diego County. This means we may not know which of these counties is ‘safer’ than the other. For 4 of the counties we would reject the null hypothesis. When I take a look at the table provided in the exploratory analysis section, 3 of them seem to have a higher overall mean than San Diego, while Alameda county has a lower mean.

Vaccination Rates in San Diego County vs. Other Counties:

I compared San Diego County’s mean vaccination rate to the other nine most populated counties in California. I was looking to find if there was a statistical difference between the two counties, and what that could mean when taken into consideration with the other facts and figures. A table of the null hypotheses and the results can be found below:

Null Hypothesis - What County We Are Comparing	**Result**
Los Angeles	Do not reject the null hypothesis.
Orange	Do not reject the null hypothesis.
Riverside	Reject the null hypothesis.

San Bernardino	Reject the null hypothesis.
Santa Clara	Do not reject the null hypothesis.
Alameda	Do not reject the null hypothesis.
Sacramento	Do not reject the null hypothesis.
Costa Contra	Do not reject the null hypothesis.
Fresno	Do not reject the null hypothesis.

The results indicate that 8 of the counties, when compared to San Diego County, may have a statistically similar mean vaccination rate. Of the two counties who we would reject the null hypothesis for, they fall below the mean rate in San Diego County, so we can probably say that SD has a better vaccination rate than them.

Discussion/Implications

My results involving comparing the California counties were pretty inconclusive. Some counties seem to be equivalent or ‘safer’ than San Diego county, and some counties seem to be in a poorer place in terms of COVID than San Diego County. It is hard to compare, and making a contest out of it probably takes away from the human-centeredness of this analysis. Each county is different and should take the precautions they feel are necessary. I know from personal experience at my job, that competition-like studies often don’t get the results we want and can be counterproductive to our cause. It creates a false sense of superiority or insecurity that does not need to be piled onto the daily stressors of life. I would take this anecdote and my further evaluation of California counties as a loss for my general analysis. While it was interesting to get an understanding of State’s precautions and metrics, it did not offer insight that was not already available to the public. I am assuming politics and societal pressures play a bigger role in this as well and must be taken into consideration, lest we bias the conclusions we make.

On the other hand, the analysis of the effect of the vaccine in San Diego County was very compelling. I found that Covid Daily Case Rate dropped rapidly when the vaccine rolled out and the lockdown was still in place. After the lockdown as well as before the vaccine rollout seemed to have statistically similar means, while the time during vaccination and lockdown had statistically different means from both of them. This can really explain the breadth of the pandemic and its danger to society. It also portrays how the solution isn’t just vaccines or mask mandates separately. It really illustrates how many factors must work together to combat this virus and we may need to make some sacrifices in the meantime for it to truly go away. This may include another lockdown as well as getting the vaccine. Doing one or the other in San Diego County didn't seem as fruitful as doing both, and that could be a lesson for the rest of the United States.

That being said, I must remember that San Diego County is not a monolith for this country. It has high density urban populations as well as suburban areas. It is hard to compare this kind of lay of the land to a rural county in the Midwest or the dense areas of the East Coast. Like I said, each delegation must take into consideration their community's needs and put in place the proper precautions.

Limitations

One limitation of my project was the equal variance assumption made during t-testing. I was hard-pressed to be able to assure there was equal variance between my samples, so I toggled that parameter when using scipy.stat's independent t-test function. I also made the assumption that the population stayed the same when looking at vaccination rates. While for case rates, one could reasonably conclude that someone could get Covid again, you can only get the vaccine once. This means you would want to subtract any of those who were vaccinated from the available population. I did not do this simply for ease of use. Despite that, Covid Act Now did provide numbers on those who received a completed vaccination sequence, so the cumulative number does only include those who have been fully vaccinated. There were some missing values when it came to cases and vaccinations, so I interpolated them to fill the missing values. I tried to choose a method that would consider the cumulative amount before and after the date that was missing so that it can find a middle stepping stone between the two.

Conclusions

Within this study, I took a look at the the following research questions:

1. Did the vaccine have an effect on COVID transmission rates in San Diego County?
2. Was there a difference in COVID case rates and vaccination rates between San Diego County and other counties in California?

With these questions, I hypothesized that the vaccine did have an effect on COVID transmission in San Diego County and that San Diego County did have a different mean case and vaccination rate compared to other counties in California. I was only half right in my hypotheses, since there was only a change in case rate in San Diego County when the vaccine was rolled out during a lockdown and few counties were similar and few counties were different in case and vaccination rate when compared to San Diego County. This project elevated my understanding of human-centered data science because it gave me a look at a real world problem that affects millions of people everyday. The data I was looking at was not just data, but people with lives who are changing because of the pandemic. I want to be careful in not creating a monolith or general assumptions of different areas or groups of people because everyone handles these trying times differently. I am only a graduate student, I do not know what the solution to man's problems are, but I can get a better understanding of these problems through analysis and a scientific approach. With that said, it is extremely important to learn through listening to others and their stories. Sometimes data is able to paint a story almost anyone can read, and sometimes you have to dive into the weeds to pull out a resolution or a plotline. And sometimes you make mistakes. I am sure my t-testing isn't perfect or to the rigor of a professional, but I was truly invested in breaking apart some of this COVID data to get a better understanding of the world around me and what other people might be thinking.

References

Science Brief: COVID-19 Vaccines and Vaccination: Link

California Counties by Population: Link.

Lassen County, California: Link

Data Sources

Covid Act Now API | Covid Act Now - Must register to be able to access