**COVID - 19 Vaccination in California**

**Intro:**

Immunization with a safe and effective COVID-19 vaccine is a critical component of the strategy to reduce COVID-19-related illnesses, hospitalizations, and deaths and to help restore societal functioning. COVID- 19 vaccines are being administered by Department of Public Health, Medical Associations and their partner clinics. Currently the vaccines are available for all aged 12 and up. The vaccine provider will record the vaccinations.

In this project, we obtained the record data from CDC websites, USAfacts.org, and California Department of Public Health, United States Census Bureau, and California Secretary of State as sources.

**Method:**

Once we obtained the data,

1. Imported and prepared them for the analysis by slicing and aggregating.
2. Decided on and calculated a good measure for analysis.
3. Merged datasets and regressed for correlations among the data.
4. Visualized the analysis results using Pandas and Matplotlib.
5. Prepared report and presentation

**Questions:**

We considered the following questions and attempted find the solutions.

1. How is the vaccination trend in overall CA as well in counties?
2. How strong is the relationship between COVID cases, deaths, and COVID vaccination?
3. Do demographics affect vaccination?
4. Does location and time affect vaccination?

**Data Preparation and Cleaning:**

The data obtained from various sources were cleaned for any inconsistencies and blank cells (Na) using Pandas in Jupyter Notebook

**Findings:**

Once the vaccination became available in December 2020 and was later made available to a wider range of the population, the number of COVID cases started showing a downward trend and afterwards remained low.

This indicated that the higher the vaccinated percentage of the population, the faster the recovery from the COVID-19 pandemic.

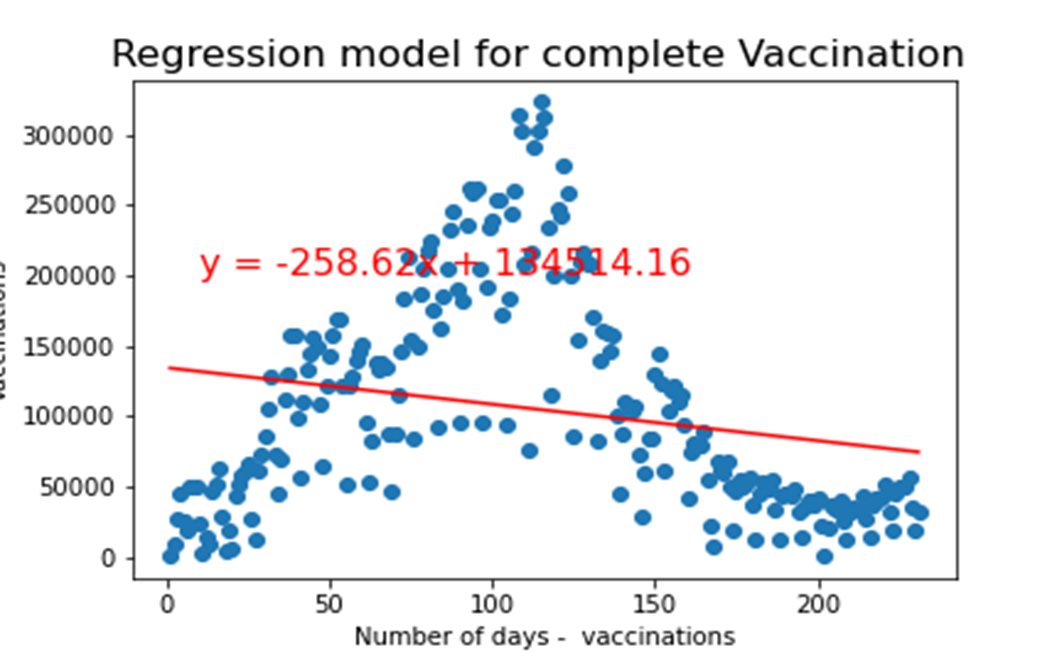
In California, at the end of July 2021, a total of 53.5% of the population were fully vaccinated: meaning two shots of the Moderna or Phizer vaccine, or one dose of the Johnson and Johnson vaccine. Comparatively, 56.5% of the total population have received at least one dose of the vaccines with two dose requirements.

A picture containing timeline

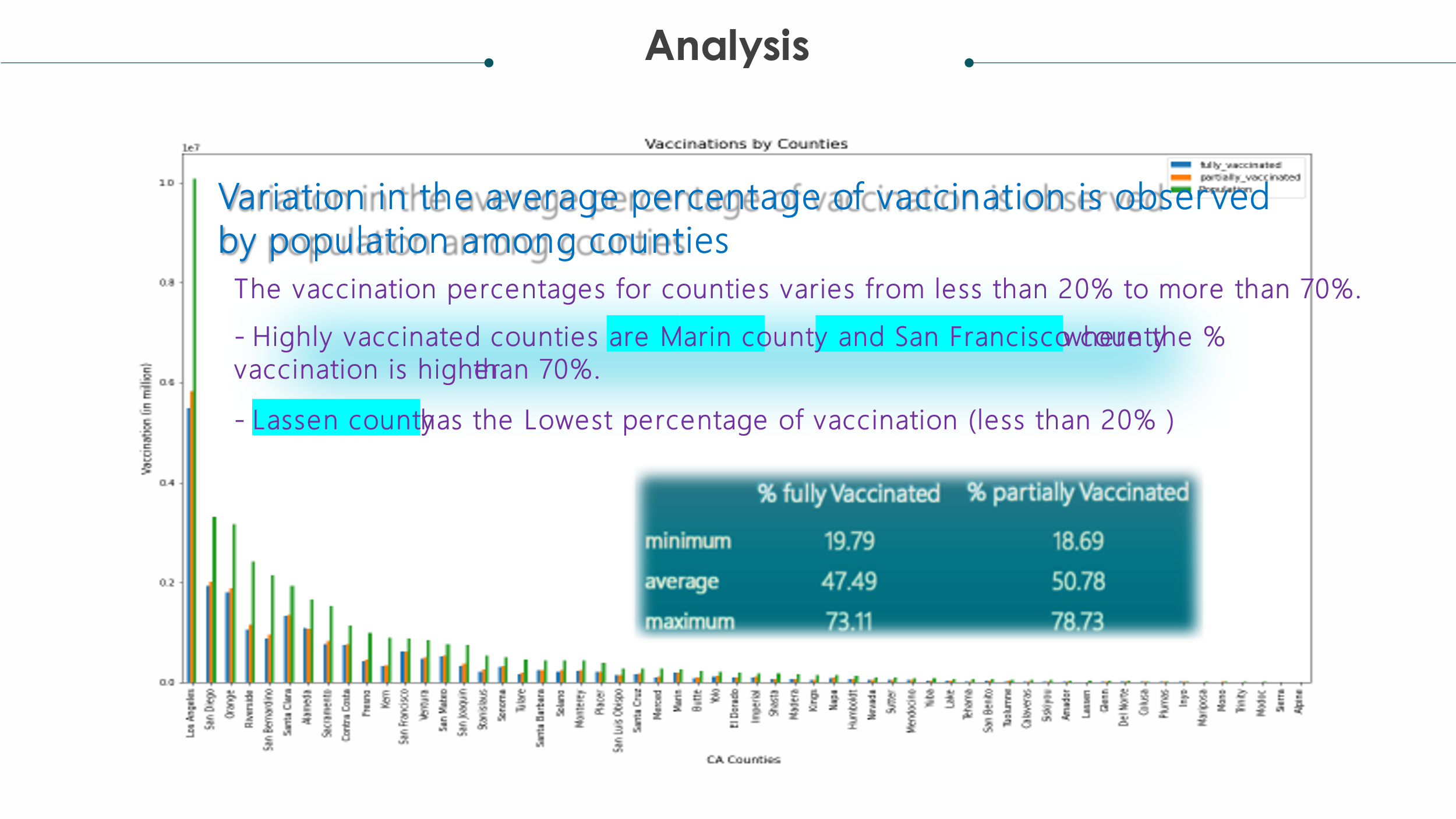
Description automatically generated

The vaccination on daily basis was analyzed over the time of 7 and half months (228 days). The vaccination rate was inclining for the first few months of availability and before steeply dropping as more people got fully vaccinated.

Analyzing the fully vaccinated data by daily vaccination using linear regression method it is found that the estimated completion of vaccination in California will take approximately 521 days from the start date, if all other compounding factors remain unchanged.

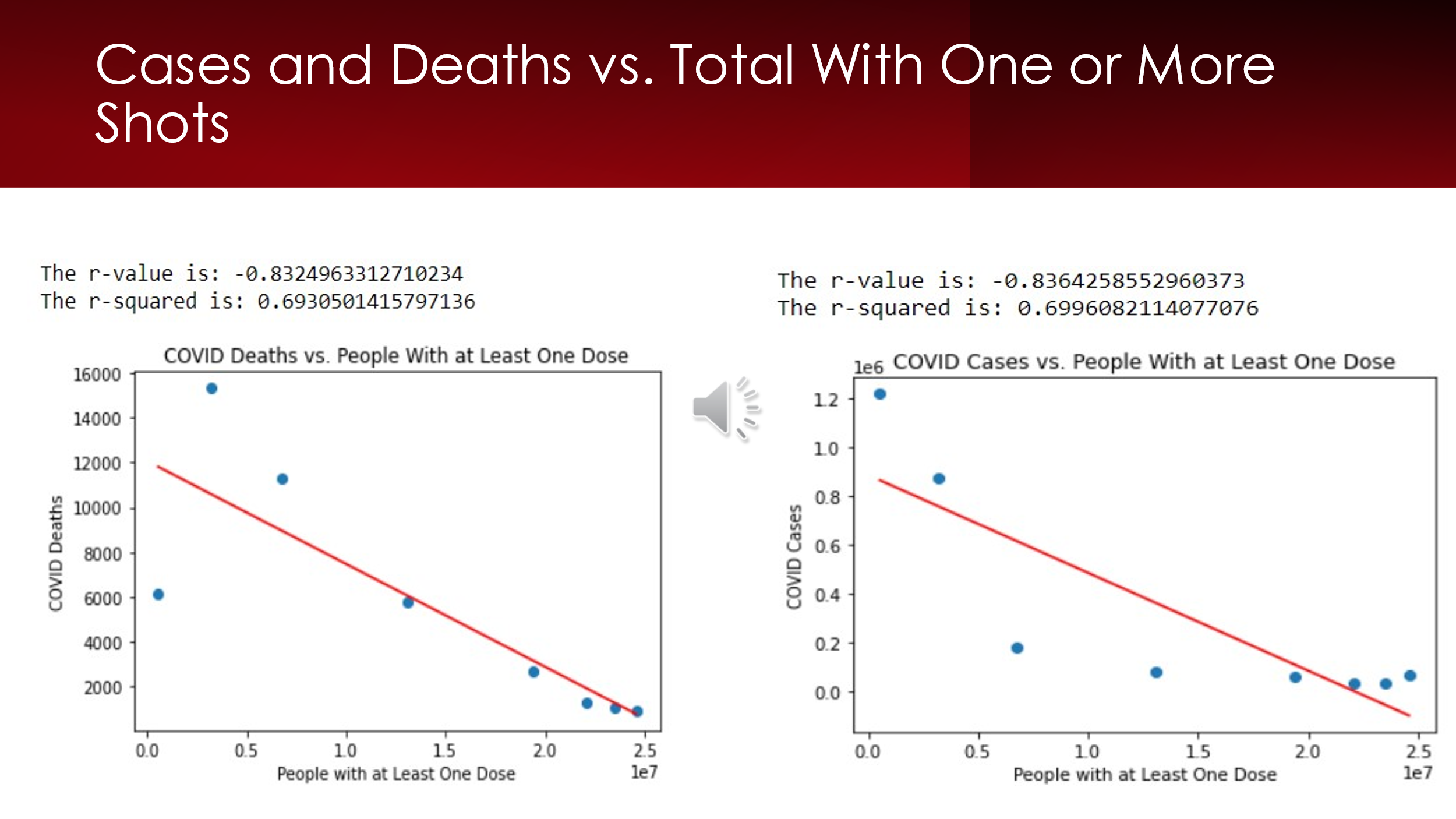


Looking at the data county wise, it is found that the vaccination percentage for counties varies from less than 20% to more than 70%. Marin county and San Francisco counties are the ones with a highly vaccinated population whereas Lassen County data shows the lowest rate of complete vaccinations.

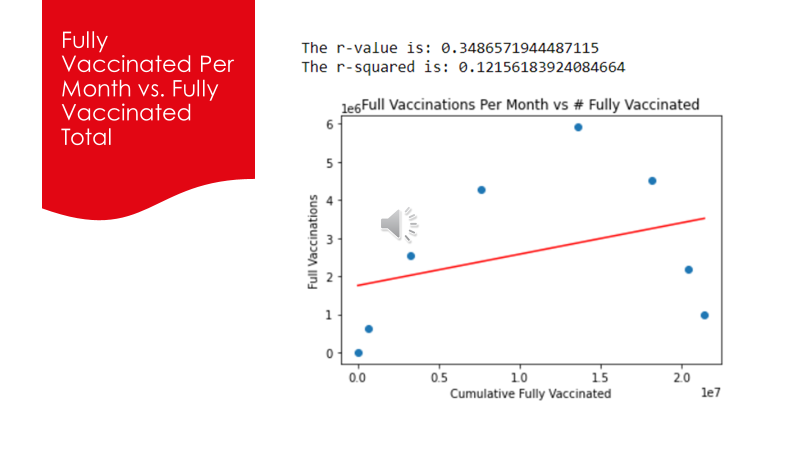


**Analysis of Cases, Death, and Vaccination:**

As pertains to our second question, a strong level of correlation was found between the level of vaccination in the population and the decrease in COVID cases and deaths. Additionally, the correlation coefficient was nearly identical with regards to the strength of the relationship between COVID cases and vaccines as well as COVID deaths and vaccines. This suggests that are equally efficient at reducing the percentage of both cases and deaths.



Linear regression showed substantially weaker levels of correlation between deaths/cases and factors such as partial vaccination/full vaccination per month, which is what was expected. Linear regression was not a fitting statistical analysis for all of our data however, and was seen when comparing full vaccination numbers per month to the cumulative total of vaccinated individuals in the population (shown on next page).



In this instance the r-value is weak enough it is not indicative of a strong relationship between the two factors. Yet, when examining the plot in more detail, it becomes clear that the largest leaps in the cumulative number of people vaccinated corresponds to the highest y-axis values (number of people completing vaccination schedules per month). This discrepancy is likely caused by vaccination rates decreasing sharply after their peak in 2021, after which most people who wanted a vaccine had taken it. The sharp upward and downward curve in the data decreased the slope of the graph as well as the r-value, even though the two factors are closely related. The comparison serves as an example of how one method of statistical analysis does not work for all datatypes.

When analyzed by location and mapped using Google Maps, data showed that urban and highly populated counties have higher vaccinations rates compared to the rural and less populated counties.

**Analysis of Demographics:**

We used data from the United States Census Bureau to collect demographic information. We used the 5 years estimate American Community Survey (ACS) data that creates population estimates based on 60 months of data beginning in 2015. The other choice was the 1 year estimate ACS which has the advantage of being more current, but only reliable for analysis on large populations. The 5 years ACS seemed mandatory because some counties in California are much smaller than the recommended 65,000 limit. Scatter plot regressions based Highest Education Attained showed significant correlation between education level and vaccination rates within a county.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, bar chart

Description automatically generated

We used r values from the above scatter plots to create a divergence chart showing the education levels that have increasing relationships with vaccination rates in a county against the education levels that have a decreasing relationship.

We also used data from the California Secretary of State to analyze demographic relationships with counties’ vaccination rates. We created a series of plots based on political party preference that fit the same mold as the education data.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

We used r values from the above scatter plots to create a divergence chart showing the Political Party Preferences that have increasing relationships with vaccination rates in a county against the party preferences that have a decreasing relationship.

Based upon strong r values, well over 50% for some categories, it seems clear that demographics do influence vaccination rates in counties.

**Limitations:**

The Census American Community Survey data we used is a little dated, from 2019. The Census Bureau released data from the 2020 Census literally the day after we finished our coding. Said data would have been incredibly interesting to examine, and not having access to it greatly limited the recent trends we could have found with our collected data. Additionally, while demographic information such as race, ethnicity, sex, and age were collected in the CDC data on COVID cases, this information was not collected for COVID deaths or vaccinations. Hence, it was not feasible to do an analysis of this information across deaths/cases/vaccinations.

The sample size is limited to 58 because that is the number of counties within California.

**Recommendations and further questions:**

It seems like policy encouraging people to take college courses *may* increase vaccination rates slightly in a county. Correlation obviously does not equal causation, however.

How does the rate and r-value of COVID deaths/cases/vaccinations change across CA counties with different population sizes?

With more time, it would be very interesting to group CA counties by population size, run linear regressions for cases/deaths/vaccinations and see the difference in their relationship as county population size changes.