Final Report: COVID-19 Vaccinations in California

Jessica Pak, Saira Mayet, Andrew Nguyen

PH251: Fall Semester 2021

Project Problem Statement

We are monitoring COVID-19 vaccination rates among counties in California and in relation to age. Specifically, we are interested in whether there is any correlation between median age and vaccinated person prevalence at the county level. Utilizing two datasets, the California Census Data and the COVID-19 Vaccine Progress Dashboard, we intend on exploring, analyzing, and visualizing vaccination rates at the county level and the relationship between age and vaccination rate.

Methods

Of the two datasets of interest, the first describes COVID-19 vaccination administration across the state of California. This is sourced from the California Open Data Portal for the California Department of Public Health. Relevant fields include date, zip code, county, and raw counts of vaccination status, among other population information. The dataset starts from January 5th, 2021 and is continually updated to reflect new changes in the data, most recently updated on November 23rd, 2021.

The second dataset describes demographics (e.g. population, race/ethnicity, age, household size, etc.) for each California county from 2012, which may include outdated numbers and demographics. This dataset was rehosted on Avery Richard's GitHub, and is sourced from Census data.

The third dataset, or data object, was a shapefile of the state of California and its county boundaries, available in the USABoundaries R library.

Data Cleaning & New Variable Calculations

Variables kept from each data source: COVID-19 Vaccination Dashboard: Date, Zip Code, County, Population 5+, Number of Persons Fully Vaccinated, Number of Persons Partially Vaccinated, Vaccine Equity Metric Quartile

County Census Data: Name, Median Age

Geographic Shapefile Data: Geometry

After subsetting the data from all sources, the data were cleaned and new variables were created.

Data Cleaning

1. Mean imputation: county-level means of fully and partially vaccinated that will be used to replace NA values in dataset
2. Merging relational data: county demographic dataset with vaccine administration dataset using key variable "county", merging county-level aggregate data with geographic shapefiles

New Variables

1. Percent eligible population partially vaccinated = # of persons partially vaccinated / population 5+ (at county level)
2. Percent eligible population fully vaccination = # of persons fully vaccinated / population 5+ (at county level)

Visualizations: Table

The Table, "Vaccination Rates for California Counties," displays the partial vaccinated rate, fully vaccinated rate, total eligible population, and median age of each county in California. The summarized average rates, total eligible population, and median age for the entire state of California is calculated and presented at the top of the table. Counties with the highest and lowest full-vaccination rates are highlighted in red.

Table 1: Vaccination Rates for California Counties

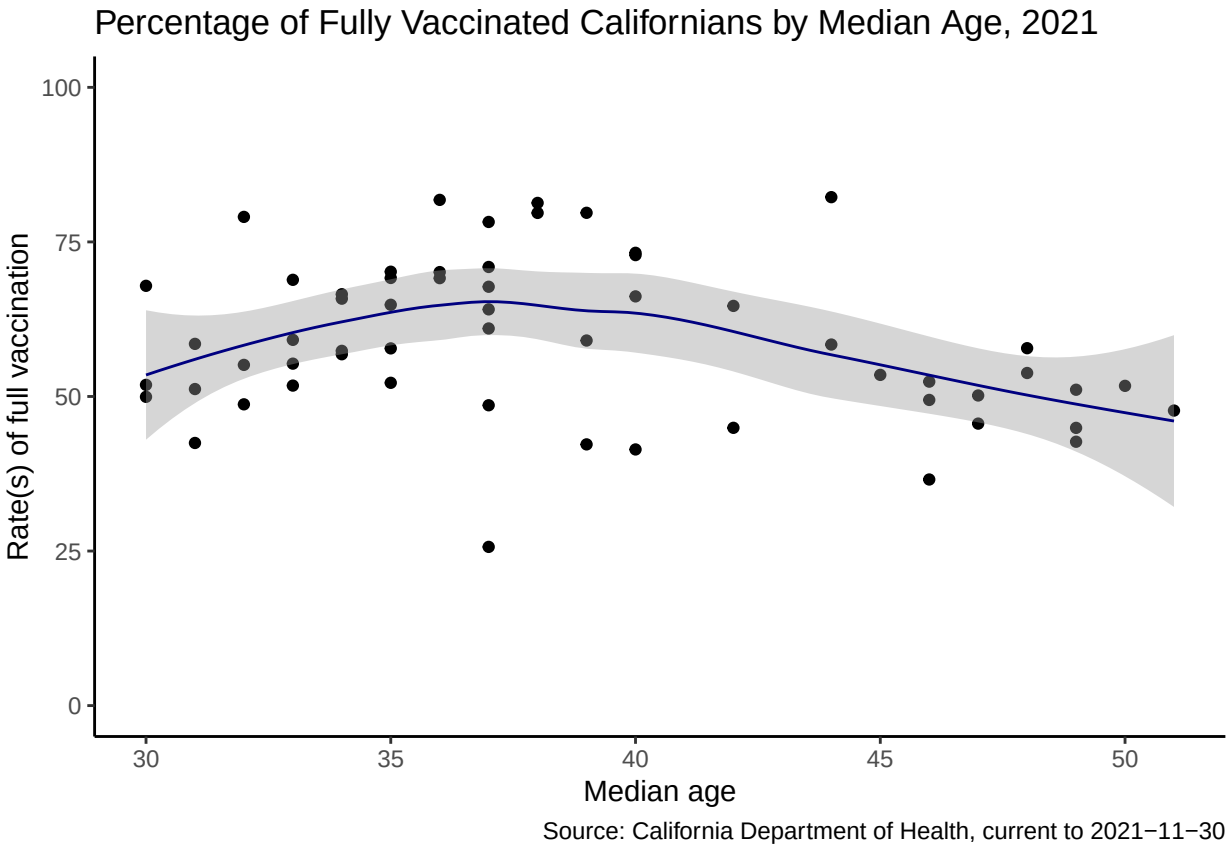
California	**6.97**	**62.8**	**33,330,578**	**38**
County Name	Partially Vaccinated Rate	Fully Vaccinated Rate	Total Eligible Population	Median Age
Alameda	7.74	78.23	1,565,553	37
Alpine	6.07	49.44	890	46
Amador	9.57	53.80	37,067	48
Butte	5.04	48.58	213,817	37
Calaveras	7.84	51.09	43,656	49
Colusa	6.68	56.83	20,153	34
Contra Costa	6.69	79.69	1,071,086	38
Del Norte	5.88	42.25	25,963	39
El Dorado	6.16	58.40	179,821	44
Fresno	7.53	58.51	911,080	31
Glenn	5.03	52.21	26,204	35
Humboldt	6.95	61.00	128,806	37
Imperial	19.75	79.05	161,453	32
Inyo	6.85	52.40	18,268	46
Kern	6.74	51.20	818,823	31
Kings	6.40	42.48	137,655	31
Lake	6.40	53.49	60,336	45
Lassen	3.08	25.67	26,919	37
Los Angeles	8.56	69.17	9,463,365	35
Madera	6.89	51.76	143,316	33
Marin	9.30	82.24	246,959	44
Mariposa	20.38	44.92	15,319	49
Mendocino	8.46	64.65	81,751	42
Merced	11.29	49.94	248,786	30
Modoc	3.24	36.57	9,384	46
Mono	7.80	67.75	12,259	37
Monterey	8.40	68.87	387,591	33
Napa	9.21	73.24	133,221	40
Nevada	7.47	57.81	92,519	48
Orange	6.95	70.12	2,986,910	36
Placer	6.22	66.19	367,860	40
Plumas	5.29	51.72	19,548	50
Riverside	6.84	57.38	2,255,664	34
Sacramento	6.83	64.83	1,427,122	35
San Benito	8.28	66.54	55,001	34
San Bernardino	6.10	55.11	1,991,511	32
San Diego	12.40	70.18	3,101,086	35
San Francisco	7.35	81.31	835,425	38
San Joaquin	9.48	59.16	688,728	33
San Luis Obispo	6.71	59.05	268,922	39
San Mateo	8.17	79.71	696,222	39
Santa Barbara	8.06	65.82	417,143	34
Santa Clara	7.00	81.80	1,833,854	36
Santa Cruz	6.70	70.95	278,046	37
Shasta	6.17	44.93	161,537	42
Sierra	3.19	47.72	2,632	51
Siskiyou	6.30	45.61	40,245	47
Solano	9.55	64.09	414,116	37
Sonoma	7.24	72.87	475,030	40
Stanislaus	9.02	55.31	505,820	33
Sutter	6.45	57.78	90,595	35
Tehama	4.61	41.43	66,959	40
Trinity	6.02	42.68	12,260	49
Tulare	7.05	51.88	420,906	30
Tuolumne	7.07	50.16	52,564	47
Ventura	6.66	69.14	801,586	36
Yolo	8.28	67.90	204,789	30
Yuba	5.55	48.73	69,526	32

Table Data Source:
Number of vaccinations and eligible population by zipcode sourced from CDPH COVID-19 Vaccination Dashboard (live dashboard). Median age data sourced from California Census demographic data (2012)

Visualizations: Scatterplot

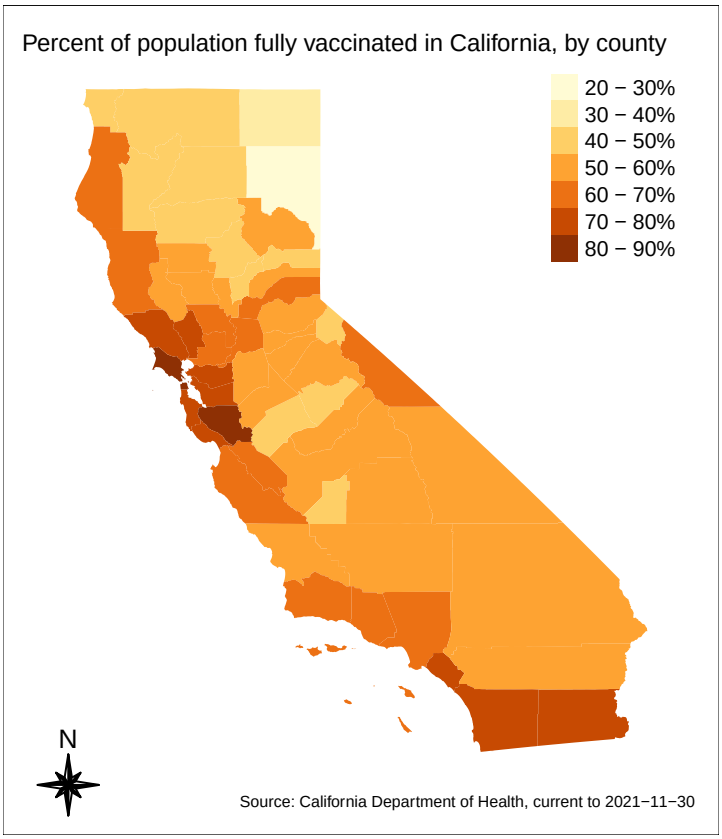
The scatterplot, "Percentage of Fully Vaccinated Californians by Median Age, 2021," shows the rate of fully vaccinated individuals stratified by median age of California residents, as measured from California Department of Public Health's latest online data repository for COVID-19. A trend line was fitted to follow the points on the plot, and a general trend of increasing fully vaccinated rates can be seen near the median age of 36-38.

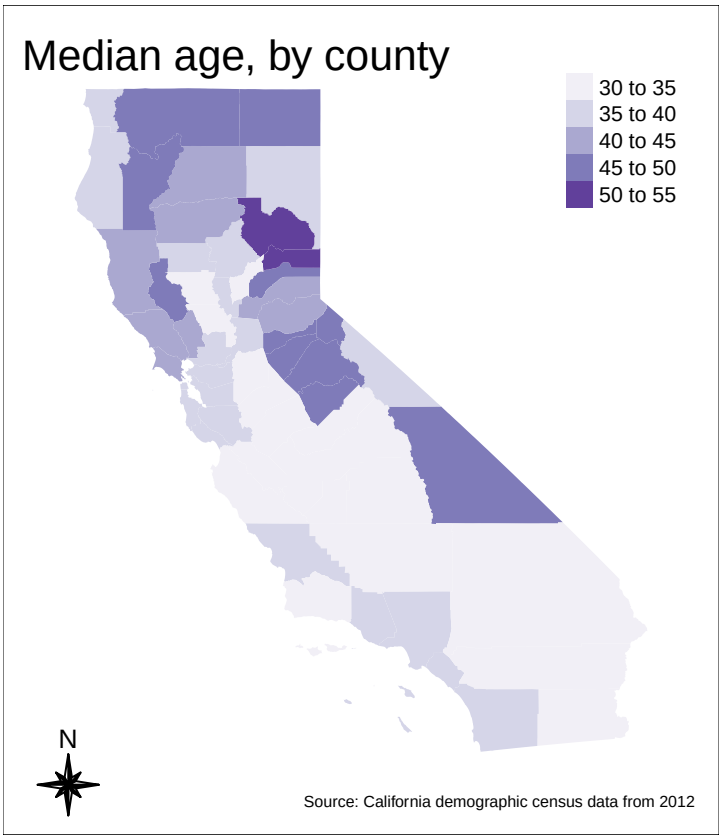
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Visualizations: Maps

The map shows the rates of full vaccination, by county, symbolized by a color ramp (from low to high rates of vaccination going from yellow to dark red). A color ramp was made with classes of 10 percent each. A second map shows the median age of each country, symbolized by a purple color ramp (younger to older median age going from light purple to darker purple). The five-step color ramp ranges from 30 to 55, with each step covering five years of age.

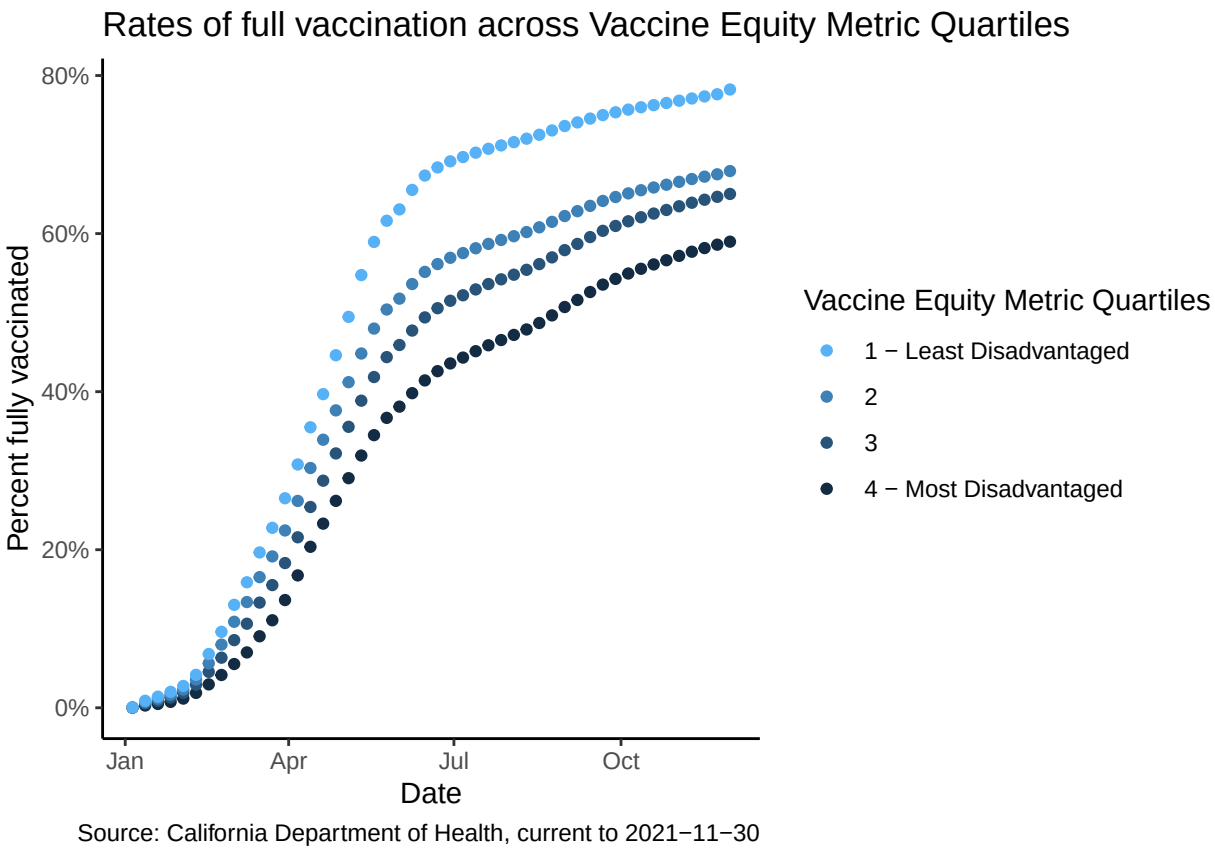



Median age, by county
30 to 35
35 to 40
40 to 45
45 to 50
50 to 55
N
Source: California demographic census data from 2012

Visualizations: Longitudinal Plot

This last visualization features points plotted by month and percent of full vaccination in aggregate ZIP codes falling under four categories of the Vaccine Equity Metric (VEM) quartiles, which is based on demographic indicators of socioeconomic disadvantage and health vulnerability (where 1 is least disadvantaged/vulnerable and 4 and most disadvantaged/vulnerable).

```
## `summarise()` has grouped output by 'as_of_date'. You can override using the `.groups` argument.
```



Discussion

Our analyses and data visualizations show that the highest vaccination rates were seen among the age group between 35-40. The general trend seen for this age group exhibited a full vaccination rate of approximately 70-73%, with rates seen above 75% as well. Outliers were seen with higher rates of vaccination, approximately 85% for a median age of 44. Outliers were also found for lower rates of vaccination of approximately 25% for the median age of 37. According to our results, the minimum median age of persons receiving COVID-19 Vaccination in the state of California is 30, for Merced county. The maximum median age was found to be 51 in Sierra county.

Overall, in acknowledgement to our original problem statement, we were not able to conclude that there seems to be a strong linear relationship between median age and rates of full vaccination across the aggregation of all California counties. A stronger relationship may exist among more localized geographies, by county, or by urban centers (i.e. singular cities) that may have more similar sociodemographic and health-cultural characteristics – this warrants additional stratified research for geography-specific purposes.

If specific counties are interested in their populations' vaccination uptake in relation to age, we recommend utilizing the code used in this report to subset the data and re-analyze the relationship between median age and rates of full vaccination. It should provide an overarching exploratory analysis that could generate other hypotheses and analyses plans. We also recommend that this same report/study design be redone with ZIP-level data, which would be more specific and robust (if data are available). Overall, further analysis and statistical testing would be needed to establish a robust correlative relationship between the two variables for the state of California within this study period (if any).

We are not confident that these findings would be generalizable to other geographies and populations outside of California. Even if other counties, states, or countries had similar sociodemographic characteristics, there may exist a variety of unknown cultural and geographic-specific factors that may affect health behaviors such as vaccination and thus render these aggregate results incompatible for application. Furthermore, a large limitation in our data is in how the median age data is sourced from 2012. Although it is plausible that the age structure of the state did not drastically change in the past decade, it is still undeniably outdated.

We explored other associations that may be interesting to follow up in future studies, including geography of vaccination rates in counties such as San Francisco, Santa Clara, and Marin county with high full vaccination rates at or above 80% and median ages of 38, 36, 44, respectively. In addition, it can be seen that rates of vaccination follow, from highest to lowest, the order of vulnerability derived from the Vaccine Equity Metric (VEM), from least to most – suggesting areas of improvement for equitable vaccine delivery. The aggregate of the least disadvantaged areas (VEM group 4) sees vaccination rates of 77%, whereas the most disadvantaged areas (VEM group 1) trails behind at 59%.

Bonus discussion

Challenges

Understanding and addressing confusing data

Some ZIPs had abnormally high vaccination totals compared to their population totals (sometimes having rates higher than 100%); we sought to better understand where this data anomaly came from (e.g. some people chose to vaccinate outside of their home ZIP, therefore seemingly inflating numbers) and how to address it. In the end, we did not remove any observations due to this issue. Rather, we aggregated vaccination numbers at the county level, acknowledging that the vaccinated population is still captured accurately at a larger geography in comparison to the total population.

Working with a live dataset

Changes in column names, vaccination eligible age, and temporality of data provided an extra layer of consideration when we were devising code that could be replicable and current to the updating dataset without maintenance. We addressed this by creating variables that returned the latest date using the *lubridate* package to ensure the results and visualizations would be up-to-date every time the .rmd was ran without manual user intervention.

Something new we learned

Working with GitHub

Although this was not an entirely new process, we appreciated the ability to implement best practices for reproducibility and collaboration. Although there were challenges in the beginning with reconciling code changes when there were multiple commits, the process emphasized the importance of team communication and scheduling.

Making maps with R

The process of making maps with R is greatly simplified with the ability to treat geographic objects (shapefiles) as dataframes and/or dplyr-compatible tibbles. This meant that county-level data was easily appended to geographic geometry. The process of visualizing this data drew heavily upon the coursework in visualization (e.g. ggplot, table creation) and felt intuitive.

Making HTML slides with R

Creation of visually-appealing and easy to navigate slidedecks through .rmd files with the slidy-presentation preset was another highlight of htis project. This process was extremely intuitive and satisfying – we feel especially empowered to make more R-based presentations and reports in the future that are both reproducible, accessible, and aesthetically-pleasing.