Milestone 5

Problem Statement

(Rubric: Include 1-2 paragraphs describing the scenario that your team is aiming to address. This can be similar to the scenario text, but please phrase in your own words.)

There is a growing concern that counties with younger median ages are less likely to have higher counts of vaccinated persons. To investigate this, we explored the California Census Demographics and the California COVID-19 Vaccinations by Zip Code datasets to compare the median age and proportions of vaccinated persons on a county level. By comparing these two variables, we can determine if there is any relationship between median age and vaccination rate in California.

Methods

(Rubric: Describe data sources, years and/or dates of data, methods used, strategies for data cleaning, creating new variables, decisions team made about the data, and additional wrangling done for visualizations.)

Our two main datasets are the COVID-19 Vaccine Administered by Zip Code and the California County Demographics. The vaccine dataset is from the California Immunization Registry and the American Community Survey’s 2015-2019 5-Year data, and the demographics data is from the California Census. Before the analysis, both datasets needed to be cleaned. Within the vaccine dataset, we selected the date, zip code, county, persons fully vaccinated, and age 12+ population columns, then filtered the date to keep only the most current data from 9/14/21.

Certain zip codes were missing counts of persons fully vaccinated so we replaced those ‘NA’ values with the county level average. We accomplished this by grouping the data by county, computed the county level mean, then created a new column for the count of fully vaccinated without missing values using an if else function where the county mean would replace the missing value or else keep it as is. The county vaccination rate was calculated by dividing the number of persons fully vaccinated by the age 12+ population per county.

The median age and county columns were the only needed variables from the demographics dataset, so the dataset was condensed by only selecting for those two variables. The county name was changed to uppercase to create a key where the demographic dataset could be joined to the vaccine dataset. After joining both datasets, we performed an analysis on the resulting data. We used the datatable function to produce a print quality table of the COVID-19 vaccination rate per California county and the median age of each county. We used the ggplot function to produce a scatterplot of the relationship between county level median age and vaccination rate. Lastly, we also used the ggplot function to visualize a vertical bar chart of the median age per county.

Results

(Rubric: Describe the results of your analysis, including what is displayed in the visualizations. (Note: no statistical testing is needed)

The first table presents the total eligible population, persons fully vaccinated, vaccination rate, and the median age for each county in California. The vaccination rate varies by county, ranging from 40% to 125%. A possible explanation for counties with rates greater than 100% is that residents could have traveled to other counties to get vaccinated.

We created a scatter plot to determine if there is a strong correlative relationship between median age and vaccination on a county level. Each point in the plot represents a county. The scatterplot shows a weak correlation because the points are scattered between 40-80% rate with no particular pattern. There are 4 outliers with 3 that have a rate greater than 100%. The blue line of best fit is not very linear, further supporting the weak association we see between the variables.

This vertical bar chart presents the median age of each county in California. Overall, in California as a whole, the median age is less than 50 years old with each county's median age ranging from 30 to 50 years old.

Discussion

(Rubric: Provide interpretation of results and what they tell your team about the given scenario.)

Based on the scatterplot and table, we can sort median age compared to vaccination rates.

The data does not support the concern that counties with younger median ages are less likely to have higher counts of vaccinated persons. Yolo County also has a median age around 30 yet has almost twice the vaccination rate at 72%. The median age of Imperial County is 32 years old yet it’s vaccination rate is 84%, nearly 2.5 times the age. Mariposa County has the 4th oldest median age (49%) with the 3rd lowest vaccination rate (40%).

As of September, the vaccines were not approved for children less than 12 years old.

Perhaps, it appears that counties with younger median ages have lower vaccination rates because the ineligible children are included in the calculations. To get a more accurate view of the relationship, we used the age 12+ population as the denominator to calculate our vaccination rate. The weak correlation in the scatter plot provides evidence that we should not be too concerned about the counties with younger median ages.