Nicole Tang
PM566
Final Project

Data Analysis for Covid 19 Vaccine Hesitancy and Possible Demographic and Geographic Correlations

Introduction

Background

COVID-19 vaccine hesitancy refers to the reluctance or refusal to get vaccinated despite the availability of vaccines. Vaccination plays a crucial role in controlling the pandemic by reducing the spread of the virus, preventing severe illness, and decreasing hospitalization and death rates. The COVID-19 vaccines have been proven to be highly effective in boosting immunity and protecting not only individuals but also communities by contributing to herd immunity. However, hesitancy has been influenced by factors such as misinformation, distrust in healthcare systems or government authorities, concerns about the speed of vaccine development, and fears about potential side effects. Social, cultural, and political contexts have also shaped people's attitudes toward vaccines. Addressing vaccine hesitancy requires comprehensive public health strategies that include transparent communication, community engagement, and efforts to build trust by addressing the specific concerns and barriers faced by different populations.

The CDC has published a data set about vaccine Hesitancy for COVID-19 in 2021. The data was collected from the US Census Bureaus Household Pulse Survey form May 26, 2021-June 7, 2021. This specific dataset had 2862 observations and 21 variables. In order to determine hesitancy levels, people were surveyed "Once a vaccine to prevent COVID-19 is available to you, would you…get a vaccine?" and the following options were: 1) "definitely get a vaccine"; 2) "probably get a vaccine"; 3) "unsure"; 4) "probably not get a vaccine"; 5) "definitely not get a vaccine". his data set also looks into varying levels of hesitancy: hesitant, hesitant or unsure, or strongly hesitant. People who responded "probably not" or "definitely not" were categorized as hesitant.

This data set has various demographic information showing information by county, state, ethnicity, COVID-19 vaccine coverage (CVAC) and social vulnerability index (SVI). CVAC measures supply and demand challenges to vaccine rollout based on healthcare accessibility barriers, sociodemographic barriers, and historic undervaccination. CVAC Index is categorized as follows: Very Low (0.0-0.19), Low (0.20-0.39), Moderate (0.40-0.59), High (0.60-0.79), or Very High (0.80-1.0) Concern. SVI is measure of how much a community is socially vulnerable to disaster. Characteristics of this include education, family, language, and vehicle access. The SVI is categorized as follows: Very Low (0.0-0.19), Low (0.20-0.39); Moderate (0.40-0.59); High (0.60-0.79); Very High (0.80-1.0). The objective of this project is to observe any possible correlations between demographic and geological factors and the rates of vaccine hesitancy.
Data set origin:
https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw/about_data

Questions of Interest

Nicole Tang
PM566
Final Project

1. Are there any correlations between state and region and the rates of vaccine hesitancy?
2. Are there any correlations between the social vulnerability index and rates of vaccine hesitancy?
3. Are there any correlations between COVID-19 vaccine coverage and rates of vaccine hesitancy?
4. Are there any correlations between ethnicity and rates of vaccine hesitancy?

Methods

Acquiring Data

A csv file downloaded to my files from the CDC website was read into a data frame. 280 observations with NA were removed to clean the data. One of the data columns held latitude and longitudinal information in the data type “Point”. So, I coded two new variable columns for latitude and longitude so it is in a more usable form for future visualizations. I also added a column for region (Northeast, Midwest, South, West) based on the state. The regions were picked based on Census Bureau designated regions.

Cleaning and Wrangling

I created a new data frame with State as the primary key and get the mean “hesitancy”. I created a new data frame with State as the primary key and get the mean “not hesitant”. The mean “not hesitant” was calculated by subtracting hesitant, hesitant or unsure, and strongly hesitant from 100. I join these two data frames to have a data frames with both hesitancy variables. I made another data frame with state as primary key and added latitude and longitude variables. I merged the data frames with the hesitancy table to have a table with pk: state and variables: estimated hesitant, not hesitant, lat, long. I broke these down into multiple steps and data frames to help organize the steps and run each of the means separately as a way to check their functionality before moving to the next step. The summary statistics are tabulated as follows to help serve as a good baseline for preliminary analysis and understanding.

Summary Statistics for Vaccine Hesitancy

Mean Hesitant (%)	SD Hesitant	Min Hesitant (%)	Max Hesitant (%)	Mean Not Hesitant (%)	SD Not Hesitant	Min Not Hesitant (%)	Max Not Hesitant (%)
12.37017	5.2458	4.026429	25.13857	61.69957	15.01524	27.73661	85.12571

The data set has columns for each ethnicity (Hispanic, non-Hispanic American Indian/Alaska Native, non-Hispanic Asian, non-Hispanic Black, non-Hispanic Native Hawaiian/Pacific Islander, non-Hispanic White) and the percentage of that ethnicity in the region. To wrangle the ethnicity data, I made a new categorical variable column that is the predominant ethnicity of that location. I also created a new data frame with a variable of ethnicity and the hesitancy to determine the average estimated hesitancy percentage by ethnicity. This hesitancy value is averaged the estimated hesitancy grouping by the predominant ethnicity.

While building visualizations, it is also important to note that to factor the ordinal categorical variables.

Nicole Tang
PM566
Final Project

Tools for Data Exploration

I used R packages ggplot2, plotly, and kable to analyze the data and create interactive data visualizations.

Results

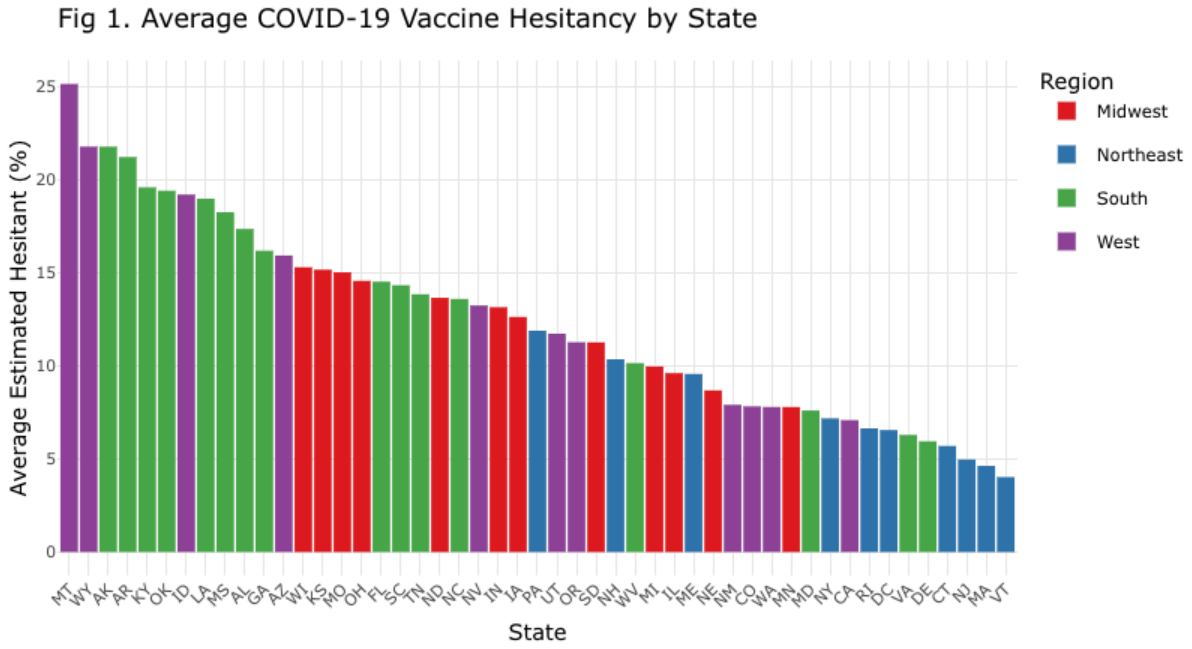

Figure 1. Average Hesitancy Rates by State and Region Analysis

I created a bar chart to observe which states have the highest rates of hesitancy. Montana has the highest rates of hesitancy at 25.14%. It is followed by Wyoming (21.78%) and Arkansas (21.77%) who have similar rates of hesitancy. Vermont has the lowest rates of hesitancy (4.03%). Based on this visualization and the difference from maximum to minimum of estimate rates I would conclude that hesitancy rates and states are correlated. The color coded grouping of region also allows to see if there is a possible correlation between region and vaccine hesitancy. Based on the figure the South appears to be more hesitant and the Northeast on average is less hesitant.

Nicole Tang
PM566
Final Project

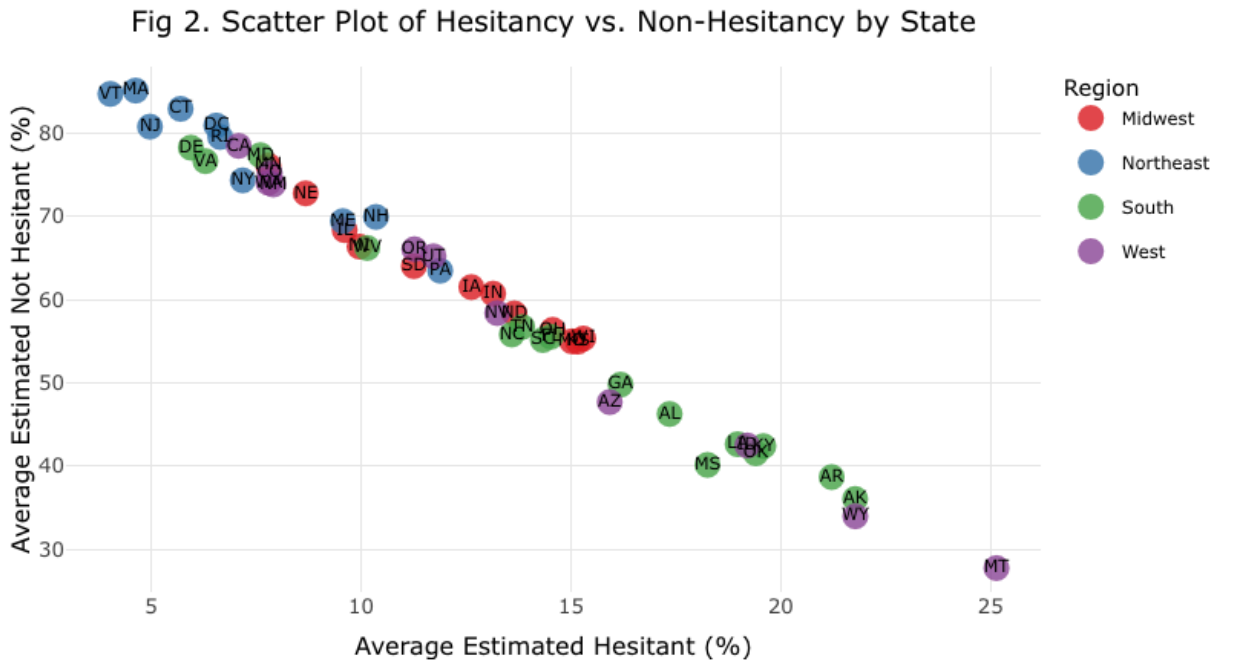

Figure 2. Average Hesitancy Rates and Non Hesitancy Rates by State

I created a bar chart to observe which states have the highest rates of hesitancy and also displays their non hesitancy rates. Montana has the highest rates of hesitancy (25.14%) and the lowest rates of non hesitancy (27.74%). It is followed by Wyoming (21.78%) and Arkansas (21.77%) who have similar rates of hesitancy. However, Arkansas (36.05%) has higher rates of non hesitancy than Wyoming (33.96%). Vermont (4.03%) has the lowest rates of hesitancy and Massachusetts (4.63%) has the second lowest rates of hesitancy. However Massachusetts (85.13%) has the highest rates of non hesitancy. Based on this visualization I would conclude that hesitancy and non hesitancy appear to be inversely related and the higher the rate of hesitancy, the lower the rate of non hesitancy. I would also conclude that the variance of mean hesitancy rates and non hesitancy rates differ by states indication a correlation between state and hesitancy rates. By region, the south appears to be clustered together with higher hesitancy rates and lower not hesitant rates. The Northeast appears to be clustered with low rates of hesitancy and high rates on non hesitancy.

Fig 3. Box Plot of COVID-19 Vaccine Hesitancy by Social Vulnerability

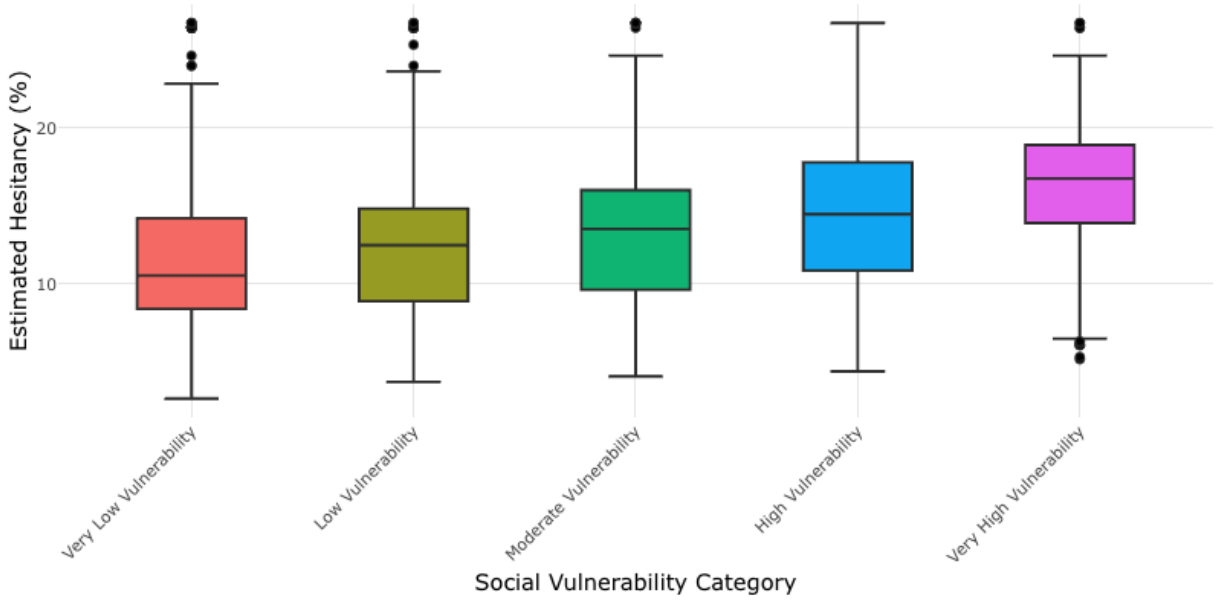

Figure 3. Average Hesitancy Rates by Social Vulnerability Index (SVI)

To examine the relationship between SVI category and hesitancy rates, I made a box plot so you can also see maximum, minimum, and median by ordinal category. Very high vulnerability has the highest median estimated hesitancy (16.76%). Very low vulnerability has the lowest median rates of hesitancy (10.55%). This is interesting because you would think that the higher vulnerability would not be quite so hesitant.

Fig 4. Box Plot of COVID-19 Vaccine Hesitancy by CVAC Level Of Concern

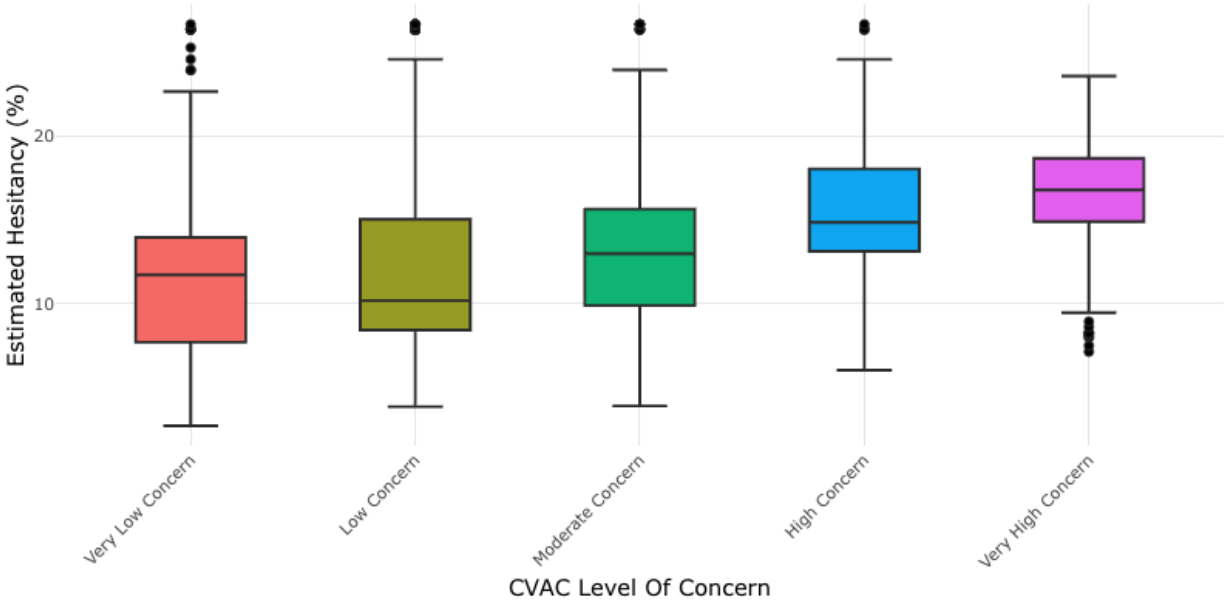

Figure 4. Average Hesitancy Rates by COVID-19 Vaccine Coverage (CVAC)

To examine the relationship between CVAC category and hesitancy rates, I made a box plot so you can also see maximum, minimum, and median by ordinal category. Very high concern has the highest median estimated hesitancy (16.80%). Low concern has the lowest median rates of hesitancy (10.18%). This makes sense that areas where there is very high concern of vaccine rollout challenges could be high levels of hesitancy. For example, misinformation could be the cause of high levels of hesitancy and cause challenges to vaccine rollouts.

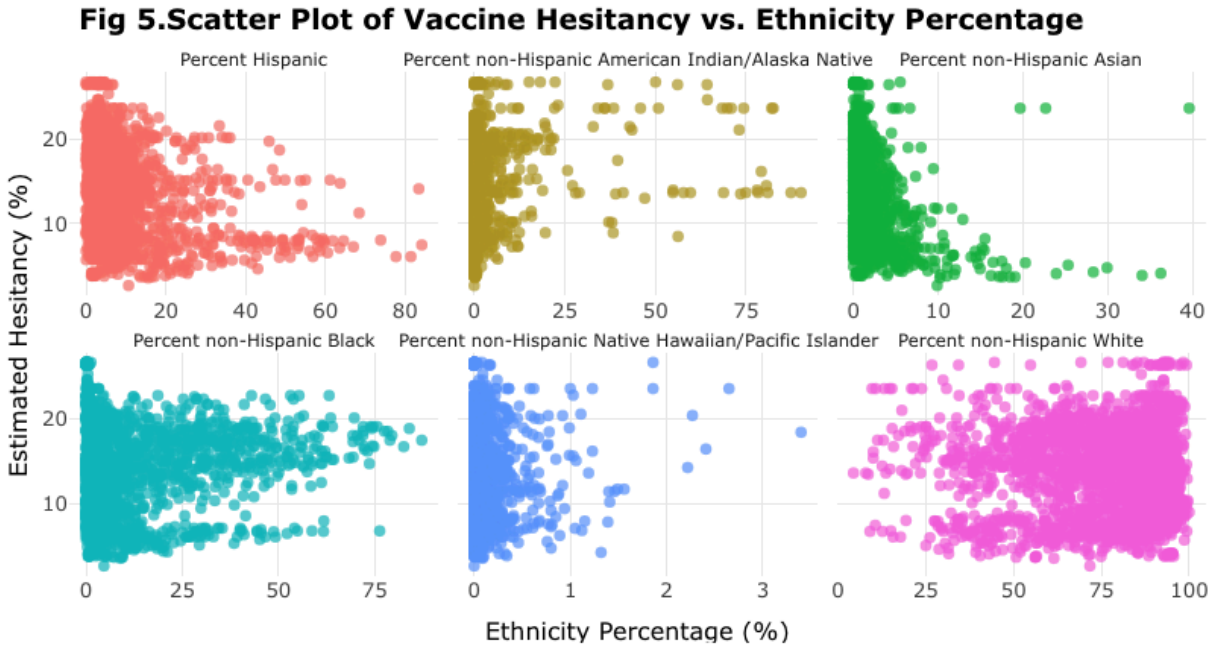

Figure 5. Average Hesitancy Rates by Ethnicity

I also wanted to look at hesitancy rates by ethnicity in a scatter plot to see each individual dot and then compare across ethnicity.

Conclusion

The different variables we looked into when looking into various variables and there possible relationship with vaccine hesitancy rates are geographical location, social vulnerability index, vaccine coverage index, and ethnicity.

Geography: Based on these visualizations and the difference from max to min of estimate rates I would conclude that hesitancy rates and states have a correlation. I would also conclude that hesitancy and non hesitancy appear to be inversely related and the higher the rate of hesitancy, the lower the rate of non hesitancy.

Social Vulnerability Index (SVI): Very high vulnerability has the highest average estimated hesitancy. Very low vulnerability has the lowest rates of hesitancy. This is interesting because you would think that the higher vulnerability would not be quite so hesitant.

Nicole Tang
PM566
Final Project

COVID-19 Vaccine Coverage (CVAC): Higher levels of concern for vaccine rollout challenges is correlated with higher levels of hesitancy. For example, if misinformation could be the cause of high levels of hesitancy and cause challenges to vaccine rollouts.

Ethnicity: Based on this visualization Percent non-Hispanic American Indian/Alaska Native had the highest rates of hesitancy and Percent non-Hispanic Native Hawaiian/Pacific Islander had the lowest rates of hesitancy.

Interventions

To address vaccine hesitancy, interventions should focus on strategies targeting high-hesitancy states and regions. This can be done with local messaging, education, and community engagement. More vulnerable populations should be prioritized. One thing that could be improved upon is accessibility to clinics not just economically but physically with transportation. Combatting misinformation is essential to counter vaccine myths effectively. When working with diverse populations, interventions should involve collaborations with community leaders and culturally tailored messaging to build trust and address unique barriers faced by diverse groups. Culturally relevant interventions, like Es Tiempo, a campaign raises awareness of cervical cancer prevention among Latinas, has proven to be successful. More data collection and evaluation will help in sustaining vaccination rates across all communities.