Analysis of COVID-19 Vaccination Disparity Across Texas

Project Report

Bobby Bagley, James Freeman-Chambless, Tim McGowan, Sara Ray

University of Houston Clear Lake

DASC 5131 Data Programming in Python

Dr. Yalong Wu

November 23, 2025

Abstract

This study investigates the disparities in COVID-19 vaccination rates across Texas, specifically focusing on the divide between urban (metro) and rural (non-metro) counties. Despite widespread vaccine availability, significant variations in uptake persist, posing challenges for public health preparedness. Following the data science lifecycle, we acquired and preprocessed county-level data from the CDC, utilizing a cumulative snapshot from May 10, 2023, to analyze established vaccination trends. The methodology involved extensive data cleaning, handling of missing values, and feature engineering to quantify categorical predictors.

We employed statistical hypothesis testing to evaluate the relationship between vaccination rates (`Series_Complete_Pop_Pct`) and factors such as the Social Vulnerability Index (SVI) and metropolitan status. Feature ranking identified population size and urban-rural equity metrics as significant differentiators between counties. Consistent with prior literature, our analysis reveals that metro counties maintained higher vaccination rates than non-metro counties across every SVI category, with most counties falling into higher-vulnerability classifications. These findings confirm a persistent urban-rural divide in Texas and suggest that future public health interventions must be specifically tailored to high-vulnerability, non-metro regions to mitigate these disparities.

Keywords: COVID-19, Texas, Social Vulnerability Index (SVI), Metro, Urban-Rural Divide, CDC

1. Introduction

1.1 Project Overview

COVID-19 created one of the largest mobilizations of public health to combat the epidemic. Vaccination is the primary source of immunity used in this mobilization to quell the spread. Texas has been a particularly hesitant stronghold in receiving the vaccine. Uptake varied significantly between the different counties. Rural areas particularly have a lower uptake percentage than urban areas. We want to look into the disparities and gain an understanding.

The data science life cycle is applied to analyze county-level data from the Centers for Disease Control (CDC) and other supplementary sources. The data set we gathered is then cleaned and engineered. After these steps, we visualize the relationships and interpret our findings to later present. We want to perform a data-driven insight into the vaccination disparities among counties to better inform public health.

1.2 Problem Statement

Herd immunity to COVID-19 relies on high vaccination rates within the population. As controversial as COVID-19 was, counties vary widely in their vaccination rates. The Centers for Disease Control and Prevention (2024) claims that individuals are 54% less likely to contract COVID-19 within 14 days of vaccination. Do these claims result in declines in infection rates in highly vaccinated counties? This issue is complex with many contributing factors that will need to be taken into account. These factors include socioeconomic status and vaccine availability, among others. To examine the factors that affect these rates, it will be necessary to apply the data science life cycle to analyze county-level data on COVID-19 vaccinations. A better understanding of the disparities and differences will lead to more effective targeting strategies within public health agencies, ultimately raising vaccination rates and reducing the impact of COVID-19. The following research question will be used to guide this study:

1. **How do vaccination rates differ among rural and urban counties in Texas?**

1.3 Project Objectives

We analyzed county-level COVID-19 vaccination data to identify the primary factors contributing to disparities in vaccination levels among Texas counties. With this research, public health preparedness can be improved through the data-driven insight gained. Data on these counties was gathered and then used for analysis and testing to gain deeper insight into this issue. The analysis covered identifying, classifying, and evaluating the variation in vaccination rates and the factors that contribute to the rates. This study was guided by the following sub-objectives:

1. Identify and analyze demographic, political, and socioeconomic factors to find associations with vaccination rates among counties in Texas. Note any auxiliary factors that may affect these rates.
2. Examine how vaccination completion varies across Social Vulnerability Index (SVI) categories and whether this pattern differs for metro vs. non-metro counties..
3. Develop and test a hypothesis. Then, interpret the data to complete the analysis for recommendations and presentation.

Using these results will better inform standards and practices going forward in the healthcare industry, along with the CDC, in targeting and informing unvaccinated members of the populace.

1.4 Report Structure

In the following sections, this report will cover initial research into the topic to orient ourselves to the problem. Then, cover each phase of the project as it pertains to the data science lifecycle: data acquisition, data cleaning, exploratory data analysis, and hypothesis testing. The report will conclude with a summary of the researchers' findings and recommendations. Reference material is listed at the end of the paper.

2. Literature Review

In Texas, COVID-19 vaccination rates were believed to differ due to several factors, including vaccination rates and hesitancy between rural and urban counties in Texas. Research suggests that the difference in county characteristics, such as poverty, education, health resources, and geographic location, all contribute to COVID-19 vaccination hesitancy. These differences matter because infections continue to spread, yet the growth of vaccinations has slowed.

A Texas county-level study by da Costa et al. (2023) examined all 254 Texas counties and found that metropolitan versus rural status was a significant predictor of vaccination rates. Their research showed statistical data that counties with higher poverty and lower education levels affected vaccination access and hesitancy.

Mofleh et al. (2022) analyzed ZIP Code Tabulation Areas and found that higher social vulnerability, particularly related to income, housing, and transportation, was associated with lower vaccination rates. Even when vaccine sites were present, vulnerable areas still had fewer vaccinations. This suggests that access points alone are insufficient and that larger obstacles, such as travel distance, poor internet connectivity, or mistrust, contribute to vaccine rates and hesitancy.

Luningham et al. (2023) surveyed adults across Texas and identified demographic and psychosocial predictors of vaccination. Lower education, reduced trust in the vaccine, and certain personal beliefs were strongly associated with not being vaccinated. These factors can be seen in rural areas, where education levels are lower, healthcare providers are scarcer, and mistrust of government health programs may be more common.

Ekren et al. (2025) analyzed Texas counties and found that rural counties ranked consistently lower in healthcare access and health outcomes compared to urban counties. This matters because counties with fewer providers, weaker health infrastructure, and poorer health outcomes are also less likely to achieve high vaccine coverage.

Some studies have looked at ways to close the gap between rural and urban vaccination rates. One study comes from Houston, where Munoz-Lavanderos et al. (2023) mapped out neighborhoods with low coverage and then teamed up with local groups to reach people directly. Even though Houston is an urban city, the idea is still helpful because using data and community voices made it easier to connect with residents who were hesitant or hard to reach.

Overall, the Texas literature points to three main conclusions: First, rural counties generally fall behind urban ones in vaccination rates, even when controlling for other factors. Second, the reasons such as poverty, education, internet access, and transportation, mixed with personal factors such as trust and beliefs, are the main factors of vaccine hesitancy. Simply adding more vaccine sites does not automatically close the gap. Third, to help reduce vaccine hesitancy, targeted, community-driven strategies are the most effective. Programs that use local data, work with trusted leaders, and reduce everyday barriers are ways to help increase vaccination rates for both urban and rural settings.

---

3. Methodology

In this study, we employed a quantitative research design using a publicly available data set to achieve the above project objectives. At the heart of our methodology, we followed the data science life cycle:

- Ask the question.
- Obtain data.
- Understand the data.
- Understand the world.
- Report our findings.

This study was done in four phases:

1. Data Acquisition.
2. Data Preprocessing and Cleaning.
3. Exploratory Data Analysis.
4. Statistical Analysis.

4. Data Acquisition

4.1 Data Source

- COVID-19 Vaccinations in the United States, County: Overall US COVID-19 Vaccine administration and vaccine equity data at the county level. Data represents all vaccine partners, including jurisdictional partner clinics, retail pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency, and Health Resources and Services Administration partner sites, and federal entity facilities.

For this project, the data was initially filtered (`State == 'Tx'`) and manually downloaded via the web app provided by the CDC. This file would have to be provided along with the project code in order for the data to be processed. Additionally, this file would have to be set up in the appropriate structure to have the code recognize and process the data. This manual file transfer could lead to errors in data processing and failure of the project code.

In an effort to streamline the process and provide consistent results, it was deemed necessary to web scrape the data and prepare it for processing within the project code. Since the web scraping and preprocessing would be performed by the code, more consistent results were attained. By removing the manual transfer process of the dataset, a smaller code package could be transferred while yielding the same results as the data pulled directly from the CDC, as long as it is publicly available.

During the streamlining process, it was found that the column headers provided are different based on how the data was collected. The manually downloaded data had inconsistent column headers. Some column headers were uppercase, while other column headers yielded a mixture of title and camel case. The web-scraped data headers were all lowercase. As some data processing had been completed before the web scraping process, it was necessary to review the project code and make

adjustments. In order to minimize the number of adjustments, it was determined that the column headers will consist of a mixture of title and camel case.

Example:

Acquisition Method	**Column Headers**
Web Filter and Download	`Metro_status, Recip_county`
Web Scrape	`metro_status, recip_county`
Web Scrape and Preprocessing	`Metro_Status, Recip_County`

The above example demonstrates that all the different Acquisition methods affected the column headers.

The web scrape and processing method was chosen as it required the least number of changes to the already defined project code.

4.2 Dataset Overview

Initial data shape: (152177, 80)

Description of key variables:

`Series_Complete_Pop_Pct`	Percent of people who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives
`Metro_status`	"Metro vs. non-metro classification type is an aggregation of the six National Center for Health Statistics (NCHS) Urban-Rural Classification Scheme for Counties. - "**Metro**" counties include Large Central Metropolitan, Large Fringe Metropolitan, Medium Metropolitan, and Small Metropolitan classifications. - "**Non-Metro**" counties include Micropolitan and Non-Core (Rural) classifications."

SVI_CTGY	"CDC Social Vulnerability Index (SVI) rank categorization where: **A** = 0–0.25 SVI rank **B** = 0.2501–0.50 SVI rank **C** = 0.5001–0.75 SVI rank **D** = 0.7501–1.0 SVI rank"
Census2019	2019 Census Population

5. Data Preprocessing and Cleaning

The raw dataset required extensive processing to make it suitable for analysis. Incorrect datatypes, missing values, and outdated duplicate values were among the issues. Our data engineering process followed a workflow structure to ensure that our data is consistent, accurate, and useful.

5.1 Data filtering

One issue that contributed heavily to the noise in the data was outdated values, which acted as duplicates. The data contained repeated entries for each county, which introduced bias and inflated the sample size for our analysis. For our analysis, we are not doing a time series model, so we removed the entries before May 10, 2023. This ensured that the most recent cumulative values were used as the vaccination status for each county.

5.2 Datatype correction

Some columns had inconsistent datatypes as they were incorrectly typified as objects instead of numeric due to formatting. A simple script to remove the commas or formatting artifacts was applied, then `pd.to_numeric` to convert these columns to our required numeric datatype. Error coercion was utilized to ensure that invalid entries are converted to `NaN`.

5.3 Handling missing values

The missing values required a systematic approach that addressed each type of these missing values. Columns with significant amounts of null values were evaluated to inform our handling methods. Calculations were made using the `isnull.sum()` function to see our sums. We believed the columns with greater than 30% of missing data were not meaningful enough to leave in the data frame. The `.drop()` function was implemented to drop these columns.

Next, we handled missing values in the rows. A missing county was dropped since there was no way to know the actual value of this record. Imputation was used, but it was only for our modeling. The mean was imputed into the missing values in the modeling phase using the SimpleImputer.

Note: Predictive modeling for this project was dropped due to a change in scope after the project proposal.

5.4 Feature Selection

The data requirements for investigating the difference between urban and rural counties were straightforward. The dataset, given its thorough data dictionary and the distribution of missing values, helped us select the appropriate features to be included in our EDA and Statistical Analysis. Below are the features selected:

Feature	Datatype	Description
`Date`	datetime	Date
`FIPS`	object	Federal Information Processing Standards code
`Recip_County`	object	County where vaccine recipient lives
`Metro_Status`	object	"Metro vs. non-metro classification type is an aggregation of the six National Center for Health Statistics (NCHS) Urban-Rural Classification Scheme for Counties. - "**Metro**" counties include Large Central Metropolitan, … and Small Metropolitan classifications. - "**Non-Metro**" counties include Micropolitan and Non-Core (Rural) classifications."
`Census2019`	float64	2019 Census Population
`Completeness_Pct`	float64	Represents the proportion of people with a completed primary series whose Federal Information Processing Standards (FIPS) code is reported and matches a valid county FIPS code in the jurisdiction.
`Series_Complete_Pop_Pct`	float64	Percent of people who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine).
`Series_Complete_Yes`	float64	Total number of people who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives
`Svi_Ctgy`	object	"CDC Social Vulnerability Index (SVI) rank categorization where: **A** = 0–0.25 SVI rank **B** = 0.2501–0.50 SVI rank **C** = 0.5001–0.75 SVI rank **D** = 0.7501–1.0 SVI rank"

6. Exploratory Data Analysis (EDA)

In the Exploratory Data Analysis phase, we used visualizations and summary statistics to get an initial understanding of how COVID-19 vaccination rates vary across Texas counties. We first examined the overall distribution of `Series_Complete_Pop_Pct` to see how fully vaccinated population percentages are spread across all counties. We then compared metro and non-metro counties using histograms with KDE curves and boxplots, and examined how vaccination rates relate to county population size (Census 2019 estimates) and Social Vulnerability Index (SVI) categories. Together, these plots helped us identify patterns, potential outliers, and evidence of an urban–rural gap in vaccination coverage, which informed our later hypothesis testing and interpretation of results.

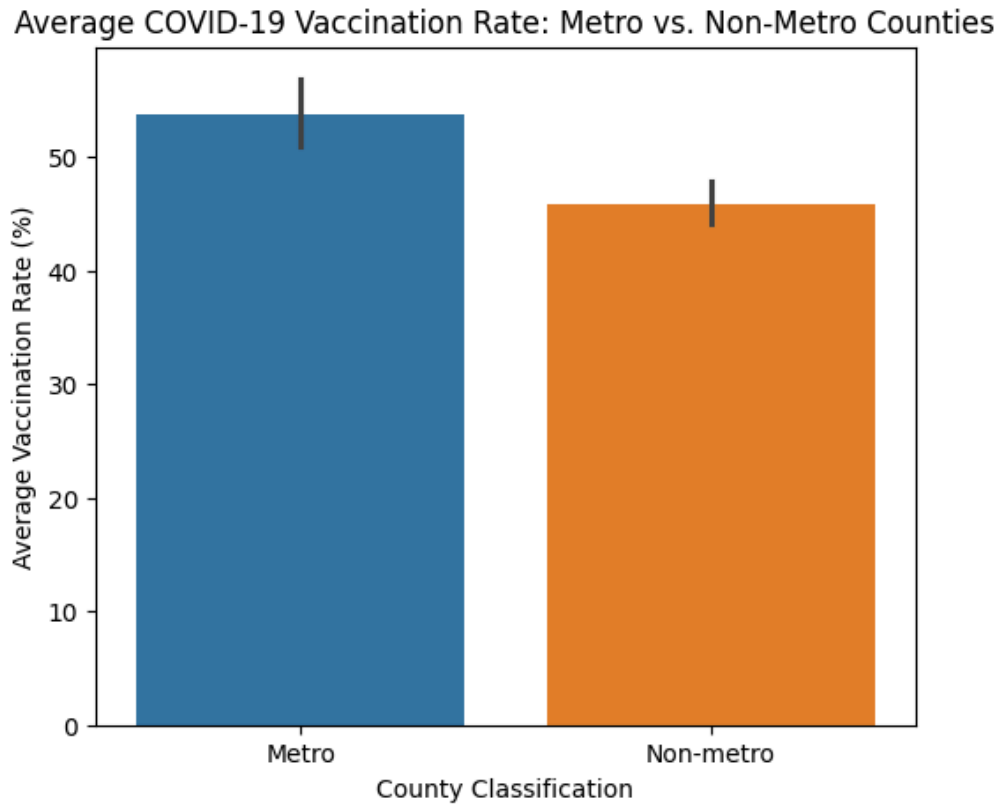

Figure 1. Average Vaccination Rate: Metro v. Non-Metro

Figure 1

This bar plot compares metro and non-metro average vaccination rates. The plot shows that metro counties in Texas have a higher vaccination rate than non-metro counties. A statistical analysis will be done to see if the difference is significant.

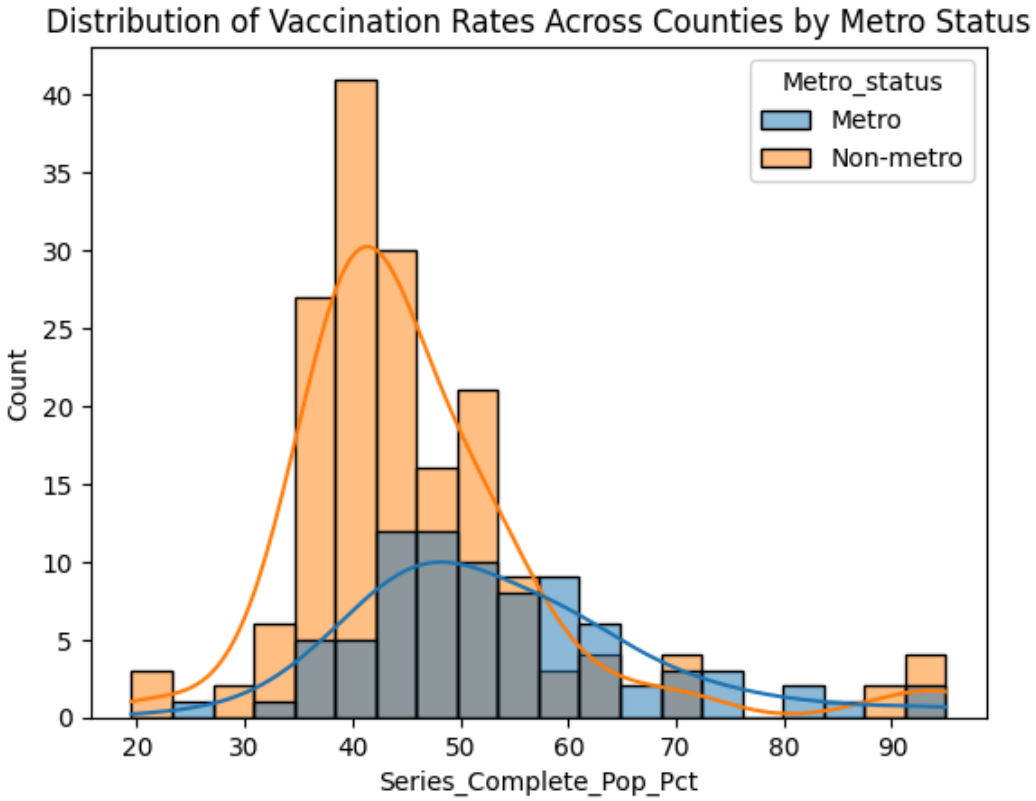

Figure 2. Distribution of Vaccination Rates Across Counties

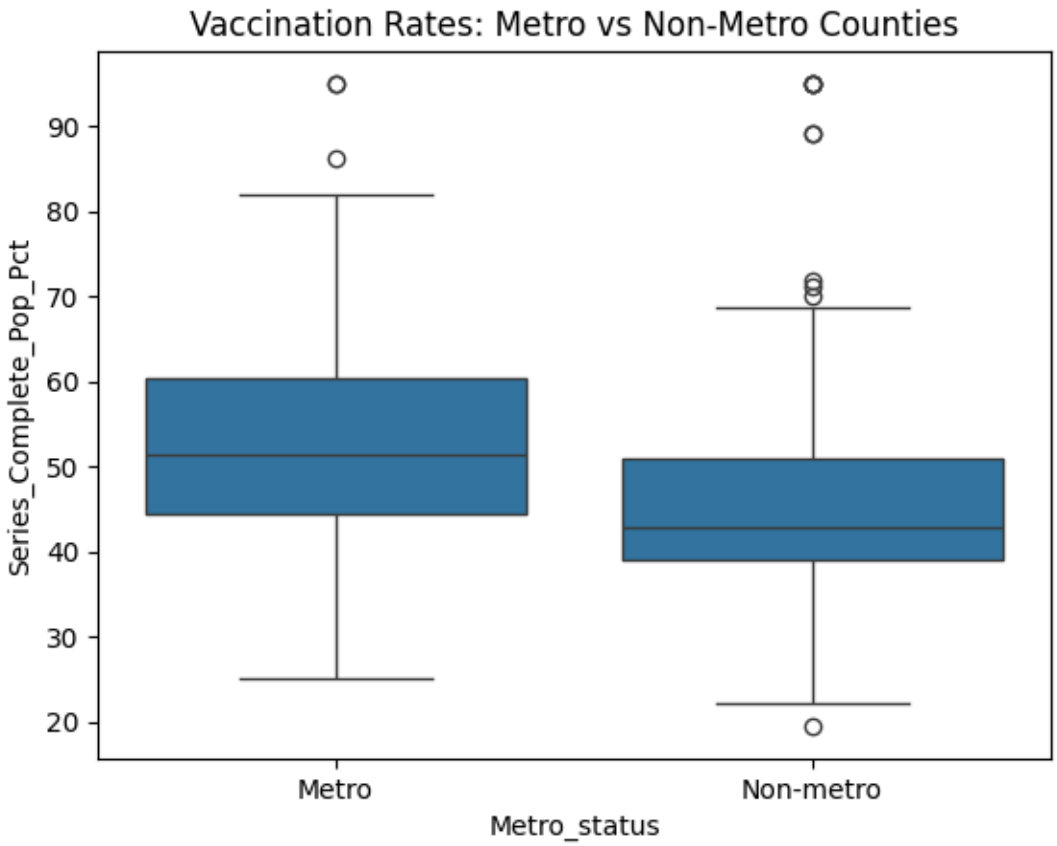

Figure 3. Vaccination Rates: Metro vs. Non-Metro Counties

Figure 2-3

These plots looked at the distribution of `Series_Complete_Pop_Pct` (fully vaccinated population %) using a histogram and a boxplot to compare Texas counties, focusing on metro vs. non-metro areas. The histogram with the KDE line showed the overall shape of vaccination rates for both groups, so we could visually compare how their rates are spread out. The boxplot then gave a quick summary of the median, spread, and any outliers for each group. Together, these plots show that metro counties tend to have higher median vaccination rates and a different overall distribution than non-metro counties, suggesting a clear gap in vaccination completion based on whether a county is metro or non-metro.

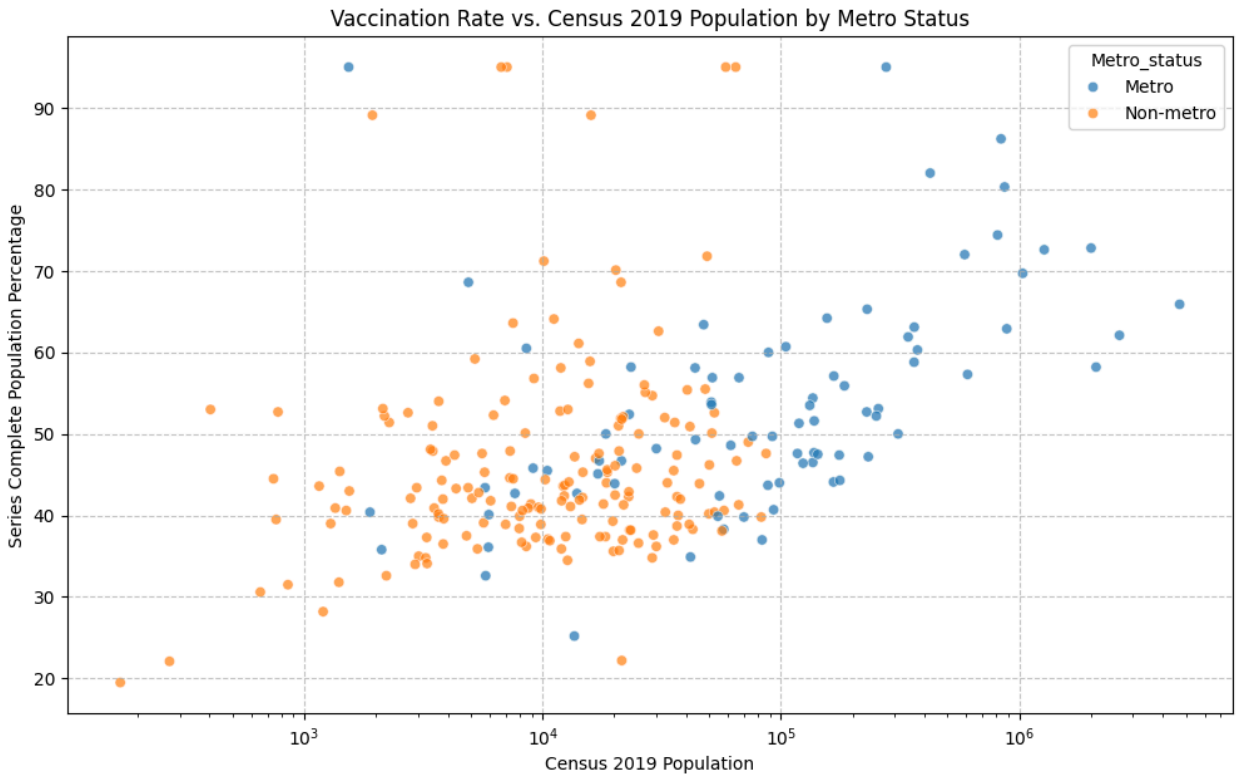

Figure 4. Vaccination Rates: Census 2019 by Metro Status

Figure 4

In this plot, we see that non-metro counties (orange) are mostly the smaller-population counties and have vaccination rates spread between about 30–60%. Metro counties (blue) dominate the higher-population range and tend to have somewhat higher vaccination rates on average, though there is still a lot of overlap. Overall, there's a loose pattern where larger, metro counties often reach higher vaccination rates, but local variation is still significant.

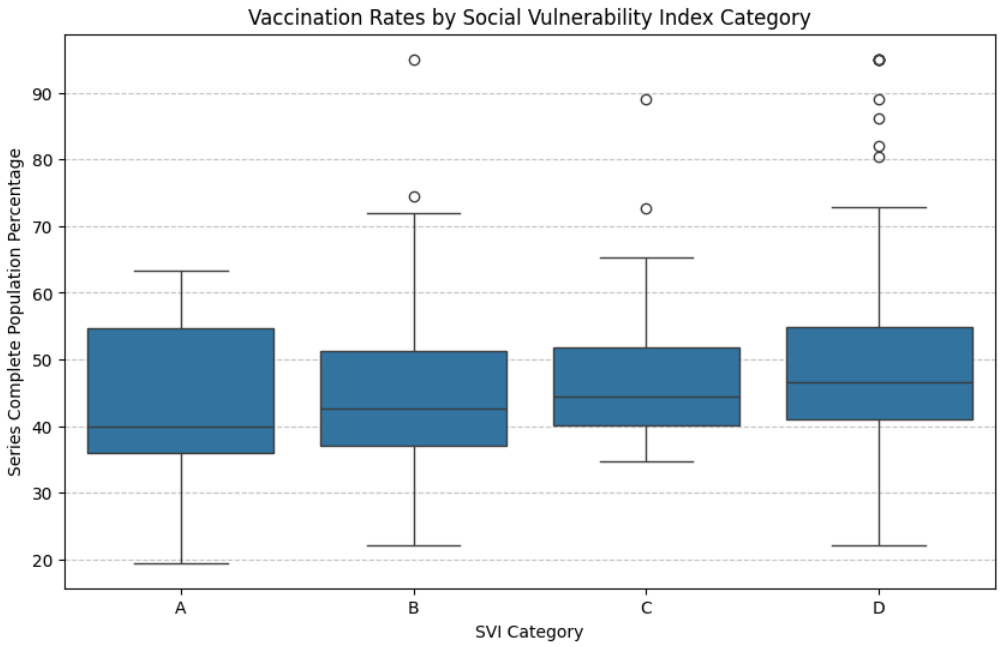

Figure 5. Vaccination Rates by SVI

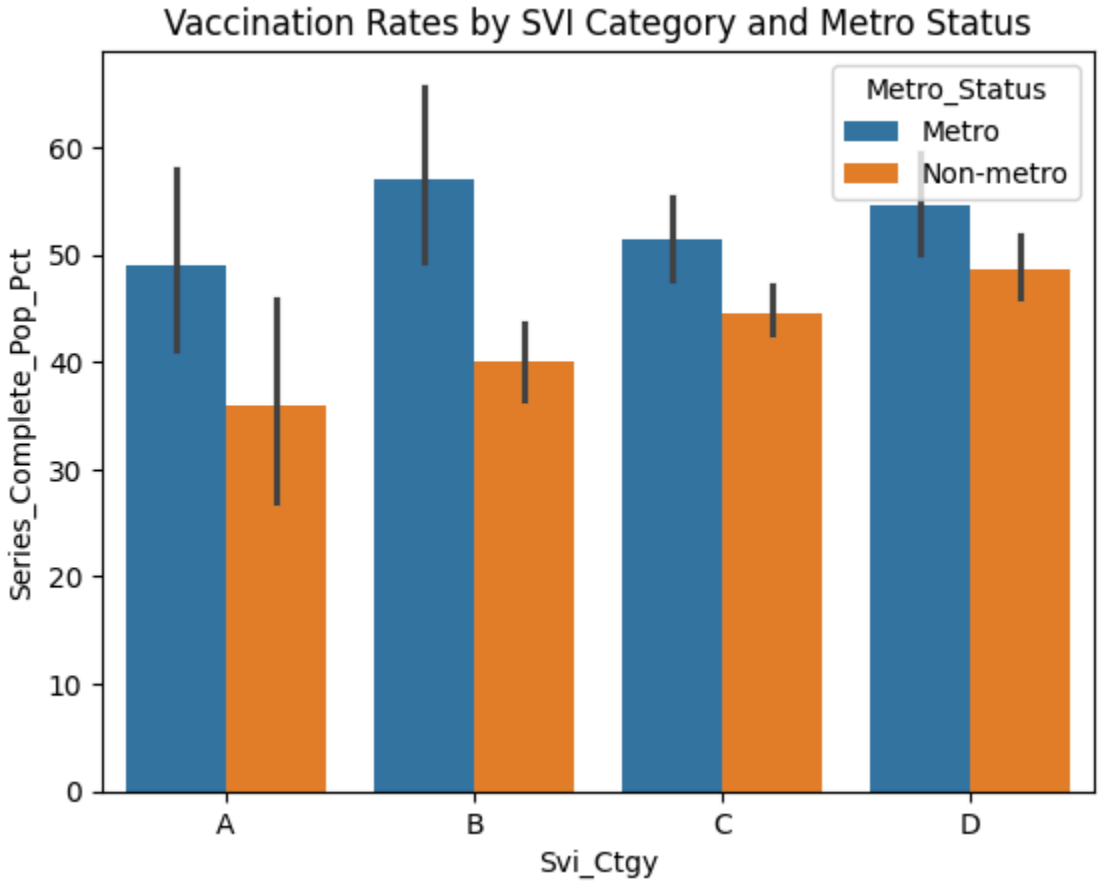

Figure 6. Vaccination Rates: Metro vs. Non-Metro Counties

Figures 5-6

The boxplot shows the overall spread and middle (median) of vaccination rates for each SVI category, combining both metro and non-metro counties. It's good for seeing how much the data varies and spotting any extreme values. The bar chart then breaks this down by metro vs. non-metro, showing that metro status makes a big difference. For example, even if SVI Category D has a higher median in the boxplot, the Metro D counties still have higher average vaccination rates than Non-metro D counties. Together, the two plots show that social vulnerability matters, but metro status has a strong and consistent impact on vaccination rates across all SVI levels, creating clear gaps between metro and non-metro counties.

6. Statistical Analysis

Welch's T-Test

A statistical analysis was performed to compare the mean vaccination rates of metro and non-metro counties in Texas.

Null hypothesis (H_0): There's no difference in mean vaccination completion between metro and non-metro counties.

Alternative hypothesis (H_1): The means differ (one group has higher or lower vaccination rates).

A Welch's t-test was performed to compare the mean rates of metro and non-metro counties. The t- test resulted in:

```
t-statistic: 4.43566240596738 and
p-value: 1.7729395313182087e-05.
```

Because the p-value ($p=1.77\times10-5$) is much smaller than $\alpha=0.05$, we reject the null hypothesis and conclude that mean vaccination rates differ significantly between metro and non-metro counties.

7. Results

After exploring our dataset, we identified patterns that are similar to those found in the studies reviewed in our literature. We plotted metro and non-metro counties in Texas by their SVI category and saw that most counties fall into the higher-vulnerability categories. When we broke vaccination rates down by both SVI category and metro status, metro counties had higher vaccination rates than non-metro counties in every SVI category. This shows that vaccination coverage is consistently higher in metro areas and aligns with the findings reported in prior Texas studies.

8. Recommendations

With the aforementioned results, significant differences were found in each of our categories. A strong focus will need to be made in non-metro areas to increase their uptake values. More outreach resources, mobile clinics, and partnering with rural organizations local to the area may lead to a higher willingness to vaccinate in a future epidemic. SVI was not a big factor in non-metro counties, but was found to have significance in metro counties. Counties with higher social vulnerability were found to have higher vaccination rates. This is likely due to initiatives and the populace knowing their vulnerability to the impact of the epidemic. Mitigating the vulnerability was a strong initiative taken during COVID-19 to increase uptake. Continuing the outreach initiative in future epidemics is recommended. An emphasis on rural counties should be made since SVI was not able to explain much variance within this demographic.

A common issue in rural areas is the availability of reliable internet and cell coverage. Building a communication network to increase outreach will help the populace to know what is available, along with access to those new resources. Telehealth is a strong resource for many of the individuals in a high SVI bracket, but only if they have access to it. This creates an additional hurdle in rural areas that will require an outreach initiative to address. Community-based engagement, along with localized campaigns, will handle this. An issue that is always inherent to data collecting is the process of reporting. Data reporting will require a standard process to ensure that the data is of quality and that it is consistently being reported. This will help with reducing the missing values and inconsistent formatting that affects the processing to analysis of the data.

9. Conclusion

In conclusion, we analyzed county-level data for the COVID-19 vaccination initiative across the state of Texas. Our focus was on comparing the different county types along with any underlying factors that help explain the discrepancies in vaccination levels. The data science lifecycle guided the structure of our workflow and analysis. Metro counties had consistently higher vaccination rates than non-metro areas throughout our project. Higher SVI values were associated with higher uptake in metro areas, but this did not carry over to non-metro counties. This suggests that rural counties face more barriers than just social vulnerability alone, and these barriers must be addressed to improve vaccination efforts.

The statistical tests conducted demonstrate that there is a significant difference between the two distinct areas. Overall, this project shows that these differences are meaningful to better inform public health initiatives. Stronger engagement in rural areas can lead to higher vaccination engagement closer to what we see in high SVI metro areas. Increasing engagement, improving communication channels, and implementing a standardized process for data reporting will help improve future vaccination outcomes. These strategies will lead to better outcomes in any future events. Higher-quality data leads to clearer insights and better-informed decision-making, which is always crucial to public health. In the end, this analysis highlights that addressing rural barriers will be essential in improving vaccination outcomes in Texas.

References

Centers for Disease Control and Prevention. (2024, February 1). COVID-19 vaccine effectiveness. https://www.cdc.gov/ncird/whats-new/covid-19-vaccine-effectiveness.html cdc.gov

da Costa, N. S. S., de Lima, M. d. C. S., & Cordeiro, G. M. (2023). Analyzing County-Level COVID-19 Vaccination Rates in Texas: A New Lindley Regression Model. COVID, 3(12), 1761-1780. https://doi.org/10.3390/covid3120122

Ekren, E., Maleki, S., Curran, C., Watkins, C., & Villagran, M. M. (2025). Correction: Health differences between rural and non-rural Texas counties based on 2023 County Health Rankings. BMC health services research, 25(1), 542. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-024-12109-2

IBM. (2021, September 29). What is data modeling? IBM. https://www.ibm.com/think/topics/data-modeling

Luningham, J. M., Akpan, I. N., Taskin, T., Alkhatib, S., Vishwanatha, J. K., & Thompson, E. L. (2023). Demographic and psychosocial correlates of COVID-19 vaccination status among a statewide sample in texas. Vaccines, 11(4), 848. doi:https://doi.org/10.3390/vaccines11040848

Mofleh, D., Almohamad, M., Osaghae, I., Bempah, S., Zhang, Q., Tortolero, G., . . . Sharma, S. V. (2022). Spatial patterns of COVID-19 vaccination coverage by social vulnerability index and designated COVID-19 vaccine sites in texas. Vaccines, 10(4), 574. doi:https://doi.org/10.3390/vaccines10040574

Munoz-Lavanderos C, Oluyomi A, Rosales O, Hernandez N, Mensah-Bonsu N, Badr H. Development, Implementation, and Evaluation of Three Outreach Events to Improve COVID-19 Vaccine Uptake Among Racial and Ethnic Minority Communities in Houston, Texas, 2022. Public Health Reports®. 2023;139(1_suppl):71S-80S. doi: https://doi-org.uhcl.idm.oclc.org/10.1177/00333549231213848

Murel, J., & Kavlakoglu, E. (2024, January 20). What is feature engineering? IBM. https://www.ibm.com/think/topics/feature-engineering

Zhang, K., Hunyadi, J. V., de Oliveira Otto, M. C., Lee, M., Zhang, Z., Ramphul, R., Yamal, J. M., Yaseen, A., Morrison, A. C., Sharma, S., Rahbar, M. H., Zhang, X., Linder, S., Marko, D., Roy, R. W., Banerjee, D., Guajardo, E., Crum, M., Reininger, B., Fernandez, M. E., … Bauer, C. (2025). Increasing COVID-19 Testing and Vaccination Uptake in the Take Care Texas Community-Based Randomized Trial: Adaptive Geospatial Analysis. JMIR formative research, 9, e62802. https://doi.org/10.2196/62802