DASC 5131 Data Programming in Python

U.S. COVID-19 Vaccinations Analysis at County Level

University of Houston-Clear Lake

Instructor: Dr. Yalong Wu
Students: Meet Hiteshbhai Patel, Bipin Bhattarai, Ravi Kishore Marella, Aashlesha Voditelwar

Abstract

The differences in COVID-19 vaccination rates among United States' counties are thoroughly examined in this paper. The research analyzes the ways in which several socio-economic and policy-driven factors impact the vaccine uptake by using a variety of county-level information that include vaccination statistics, demographic profiles, and medical facilities. Python based data-programming methods like visual analytics and correlation matrices are employed Using both statistical and geographic methods, the study identifies important trends and connections, providing useful details to help support more reasonable and effective immunization programs.

Keywords: *COVID-19, vaccination, counties, demographic, immunization*

1. Introduction

1.1 **Project Introduction**

In the United States, the outbreak of Covid-19 drastically changed the public health operations. Since the rollout of vaccines, the utilization has varied significantly across U.S counties. It is influenced by multiple range of factors including population density, race, income, education and access to healthcare.

Vaccinations measures were really crucial in limiting the virus's multiplication, each county's pace and success differed greatly. In this project, we aim to analyze the vaccination of United States' counties and determine the causes of the variations. With that, we hope to reveal the insights that can support national vaccination strategy improvements.

1.2 **Usage of This Report**

The format of this report is designed to make things clear for both technical and non-technical readers. It can be used by researchers, public health professionals, policy makers, and even students, who want to learn more about the trends of vaccination and the imbalances in the U.S and can utilize it. The study provides the framework for using data programming methods in order to deal with practical public health issues.

1.3 **Prerequisite Knowledge**

The study is meant to be easily understood by everyone, but some knowledge of data science concepts like data cleaning, statistical analysis, and visualization can enhance the understanding of the reader. In order to ensure clarity, description is provided in each significant findings.

1.4 **Report Structure**

In the second part of the report, we have the description of the datasets used, the process of data cleaning and the analysis techniques. The third part basically highlights the trends, state comparisons and regional discrepancies. The fourth part summarizes the tools and platforms used. The fifth and sixth section presents final insights and future recommendations and concludes with references and external resources.

2. Design and Implementation

2.1 Datasets Used

- **COVID19_Vaccination_Demographics_in_the_United_States_National_20250206.csv**:

  This dataset contains 29,886 records with 25 columns and presents vaccination data at the national level. It is segmented by various demographic categories like age groups, gender, ethnicity. It has key indicators like the number and percentage of people who received the first dose, received booster doses.

- **Covid-19_Vaccinations_County.csv**:

  This dataset comprises 171,914 entries with 80 attributes, which provides a detailed look at the county level across the United States. It includes extensive information like the number of vaccination doses given to the different age groups, series completion rates, booster dose uptake. It also gives us data such as census population estimates, urban-rural classifications, Social Vulnerability Index categories and FIPS codes, which are the county IDs. This data enables a more thorough analysis of the regional differences in vaccination uptake.

- **County-Hesitancy Estimates.csv**

  This includes the Covid-19 vaccination hesitancy estimates at the county level across the US. It has FIPS codes and the county names along with the estimated percentage of people who were hesitant or unsure, or strongly hesitant to take the vaccine. It covers 3143 counties.

- **State-Hesitancy Estimates.csv:**
  This dataset provides similar hesitancy estimates but at the state level with data for 8 states. It also includes state FIPS codes and state names.

2.2 Data Cleaning Process

Cleaned Datasets

- **Covid-19_Vaccine_Cleaned.csv:**
  Contains cleaned and processed COVID-19 vaccination records by county and state across the U.S. The key contents are county names, vaccination metrics, dates and reporting periods. It also includes FIPS codes for merging with geographic data.

- **Covid-19_Geographic_NAN.csv:**
  Includes national-level demographic and geographic information across counties. The key contents are population figures, socioeconomic indicators, health infrastructures, geographic identifiers like FIPS codes, and also some missing values.

  Conducted in Jupyter notebooks:

- **DataCleaningDPP.ipynb** and **DataCleaningDPP_Aashlesha.ipynb**

  First and foremost, the rows with null values for the key indicators were removed. Then the FIPS and county along with state names were standardized. The demographic and the vaccination data were combined into a single comprehensive dataset. Then the analysis was focused mostly on the most recent and complete entries in order to achieve consistency.

- **C19_Vaccine_Hesitancy_Estimates.ipynb**

  The notebook analyzes COVID-19 vaccination hesitancy in specific US counties and states, using a county-level dataset. It produces graphics to show trends and follow-through rates. The study provides a data-driven view of vaccine hesitancy patterns, aiding in targeted public health campaigns.

- **COVID19_Demographic_Insights (1). ipynb**

  With an emphasis on sex and race/ethnicity, the "COVID19_Demographic_Insights (1). ipynb" notebook examines COVID-19 vaccination trends across several demographic groups. Racial and ethnic differences in vaccine uptake are revealed, and immunization rates are consistently higher among females. The notebook highlights follow-through and booster uptake gaps via visuals.

- **MData_Visualisation_County_Dataset (1). ipynb**

  This notebook visualizes COVID-19 vaccination data at the county level in U.S. states, filtering the dataset to include Arizona, California, and Florida. It analyzes key vaccination parameters, showing top and bottom 10 counties with highest and lowest vaccination rates, enhancing data-driven decision-making.

- **Datasets_Median_Cleaned.ipynb**

  This notebook that deals with cleaning and processing COVID-19 vaccine datasets, especially when it comes to median-based imputation or transformations. Finding and managing missing or conflicting data values may be part of it, particularly for county or state-level demographic or immunization measures. The word "median" implies that the notebook contains steps that substitute median values for each pertinent column in place of lacking numerical values. To guarantee that the datasets are prepared for precise analysis and visualization in later phases, this notebook is probably a component of the data preparation workflow.

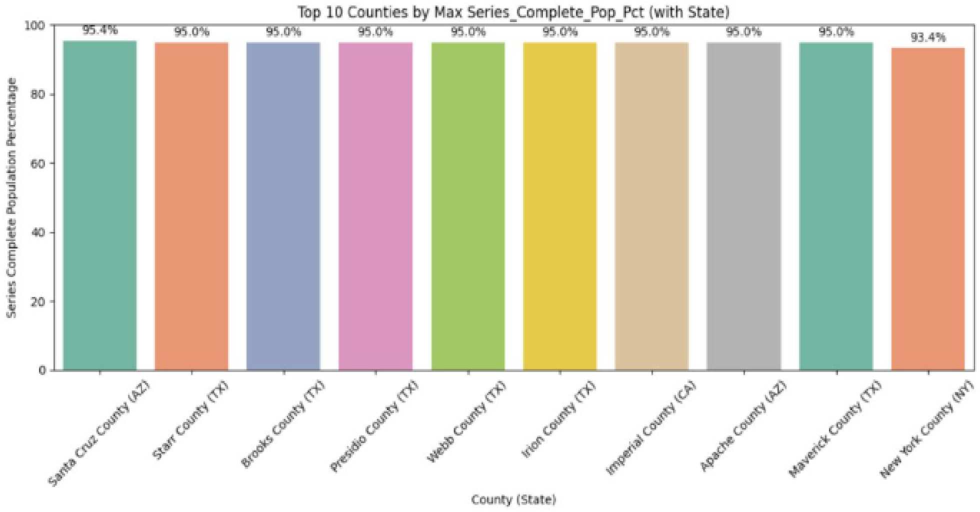
2.3 Variables Considered

- Vaccine doses administered per population
- Full vaccination completion percentage
- Booster dose coverage
- Demographic breakdowns by race and ethnicity
- Urban vs. rural classification
- Household income, education levels, population density

3. Results and Analysis

3.1 National Vaccination Trends

Vaccination rates at the county level are among the highest in the northeastern states. Meanwhile the southern states show major areas of low vaccination uptake. The five boroughs of New York City reflect a strong uptake, whereas the surrounding upstate counties vary more widely.



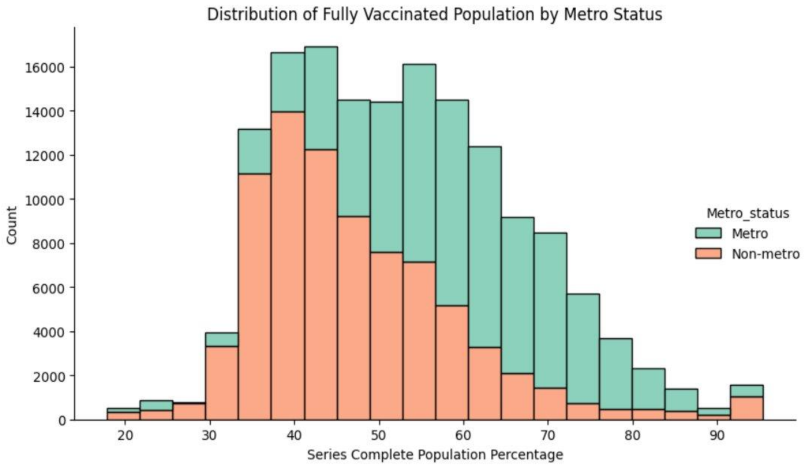
We found out that the counties like Santa Cruz(AZ), Starr(TX), Imperial(CA), have exceptionally high vaccination rates, all at or more than 95%. We can see that the counties in Texas dominate the list. This reflects good public health efforts and community engagement.

3.2 State vs. County Disparities

In the states like California and Texas, the urban counties significantly outperform rural ones in vaccination in the number of vaccination coverage. This reveals the geographic and infrastructural disparities. The five boroughs of New York City reflect a strong uptake, whereas the surrounding upstate counties vary more widely.

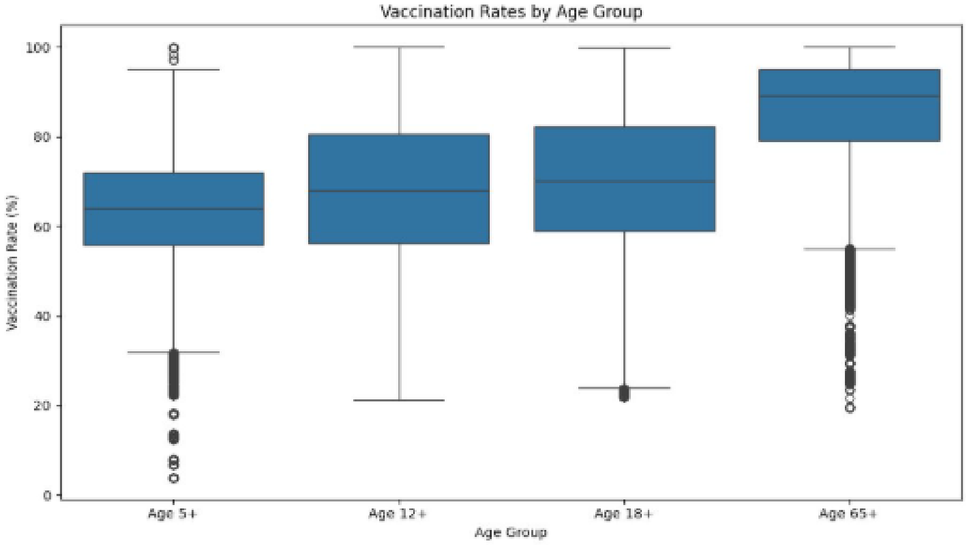
3.3 Visualization Highlights

- **Distribution by Metro status:**



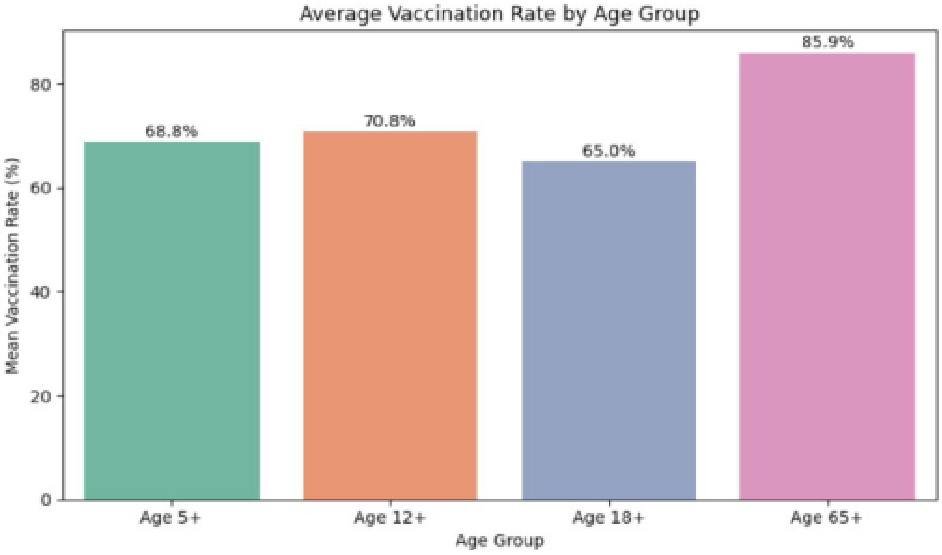
We can find a noticeable difference in the Covid-19 immunization rates in the metro and non-metro counties. Metro counties, in green, have higher vaccination rates, where many of them have more than 50% completely vaccinated. On the other hand, non-metro counties suggest a lower rate of vaccination overall. This shows that in comparison to metros, rural communities have more difficulties in terms of healthcare access, outreach initiatives, and more vaccine hesitancy.

- **Boxplot by age group:**



This boxplot shows that age 65 and older have the most stable vaccination rates, of more than 85%, whereas the younger age exhibits lower median values and is volatile especially for age 5+. Variation in the vaccination uptake in some counties is seen in the age 5+ and age 65+ categories. Basically, this shows the prioritization of elderly people during the vaccine rollout and some potential hesitation among the younger population can be seen in this diagram.

- **Average vaccination by age:**



This bar chart shows us that the elderly age people have the highest average vaccination rate. The age group 18+ has the lowest average which could possibly be due to reduced outreach. Younger people show a high uptake but still trail behind the older adults.

4. Tools, Software, and Resources

- Python 3.9+
- Jupyter Notebook and Google Colab for scripting, collaboration and documentation
- pandas, seaborn, matplotlib, numpy
- CSV datasets from CDC and U.S. Census Bureau

5. Conclusion

5.1 Summary

This project looked at the vaccination variances across the U.S. counties by using the national datasets that are available to the general public. We came to find out that the socioeconomic and the characteristics according to the regions are highly associated with the high or low vaccination uptake. We were able to achieve this by the means of data cleaning, analysis and visualization.

5.2 Future Work

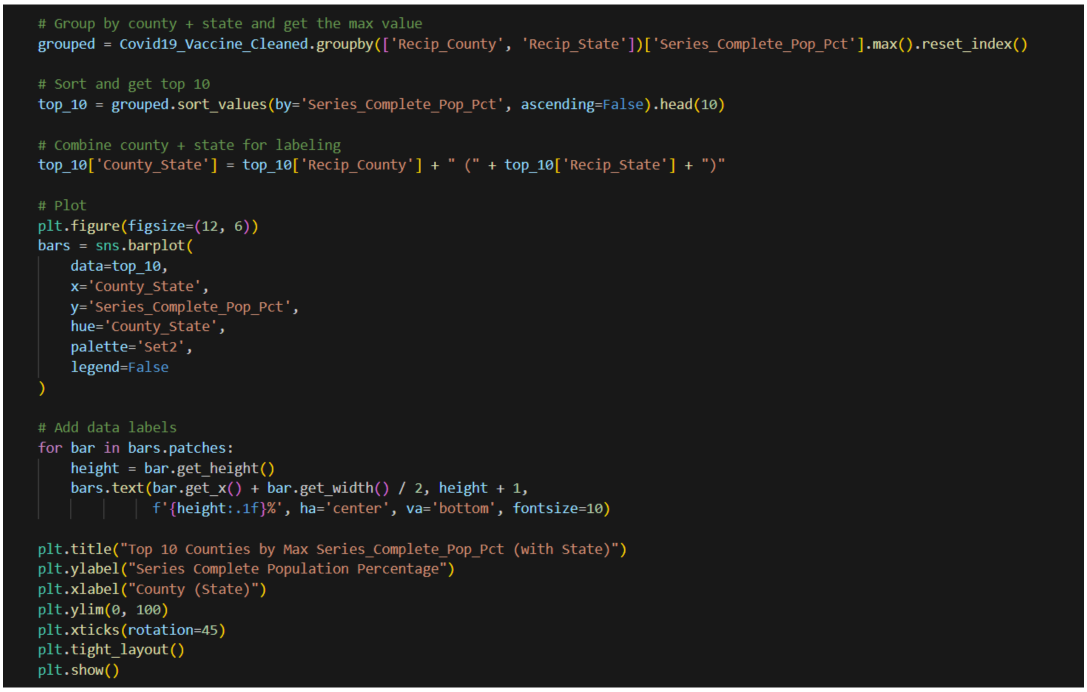
In the future, this project can be expanded into a national scale by using the county-level data across all 50 states. This can enable the comparisons between the trends amongst the states. Also, some machine learning techniques like clustering and classification can be employed to forecast the low uptake zones and also predict the outcomes due to delays.

Furthermore, integrating the real-time dashboards would help out the public health officials with insights that could lead to actions, allow them to monitor the progress, identify the lagging areas and respond immediately. This can come into more use with the use of strategies like social media sentiment analysis and some surveys. The insights that we generate from this approach can help find out the patterns of vaccine hesitancy. Not only that, it also helps guide health campaigns and support collaboration with different government agencies.

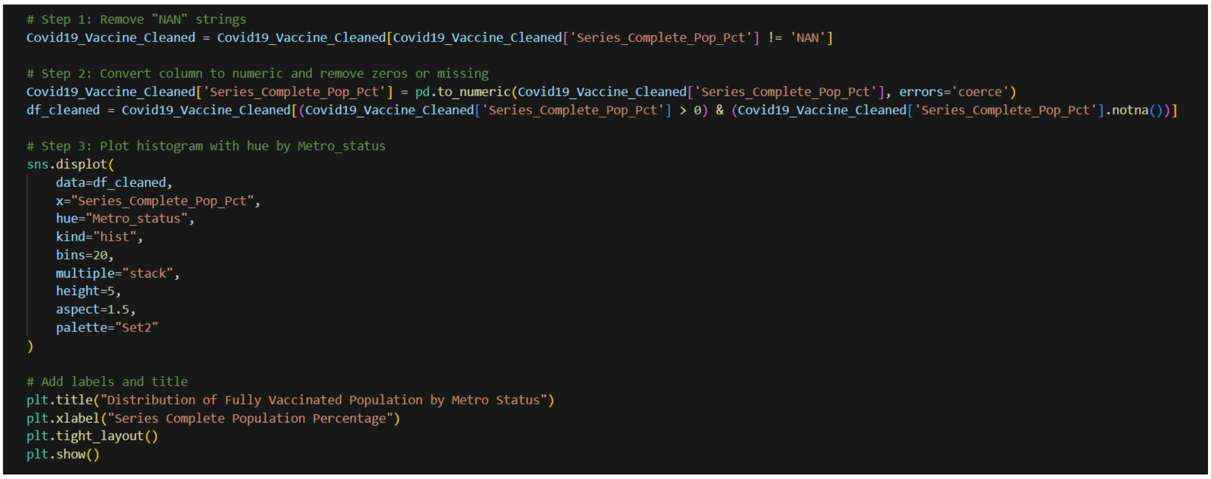
6. References

1. Centers for Disease Control and Prevention (CDC). *COVID-19 Data Tracker*. Retrieved from https://covid.cdc.gov/covid-data-tracker
2. U.S. Census Bureau. *Data Explorer*. Retrieved from https://data.census.gov
3. Arizona Department of Health Services. *COVID-19 Vaccination and Health Data*. Retrieved from https://www.azdhs.gov
4. Python Software Foundation. *Python Documentation (v3)*. Retrieved from https://docs.python.org/3/
5. The Pandas Development Team. *pandas: Python Data Analysis Library*. Retrieved from https://pandas.pydata.org
6. Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org
7. Centers for Disease Control and Prevention (CDC). (2021). *COVID-19 Vaccinations in the United States, County* [Dataset]. Retrieved from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/about_data
8. Centers for Disease Control and Prevention (CDC). (2021). *COVID-19 Vaccination Demographics in the United States, National* [Dataset]. Retrieved from https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Demographics-in-the-United-St/km4m-vcsb/about_data
9. U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation (ASPE). (2021). *Predicted COVID-19 Vaccine Hesitancy for Counties and Local Areas* [Data file]. Retrieved from https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//200826/Predicted-Vaccine%20Hesitancy-by%20State-PUMA-County.xlsx

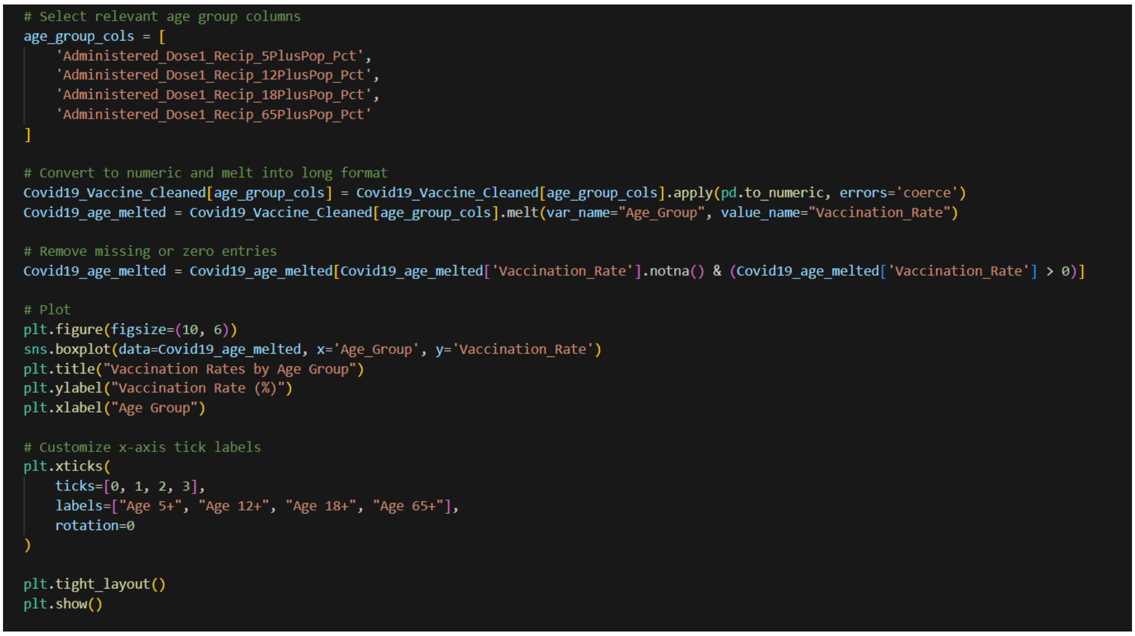
Appendix



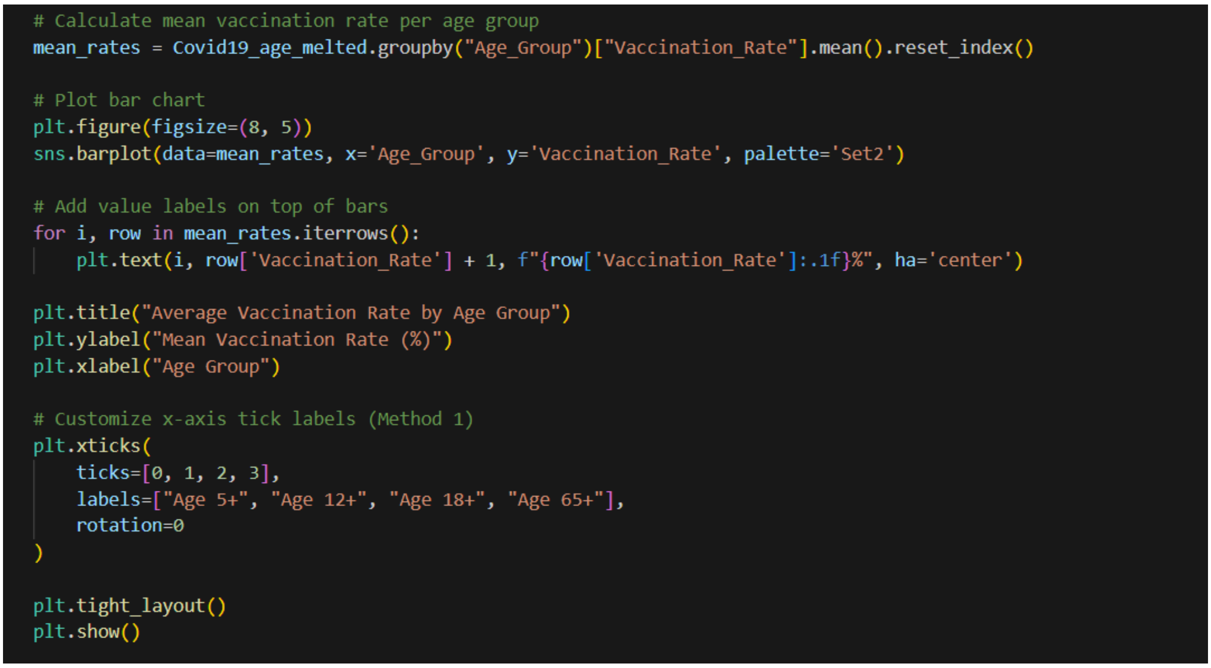
```python
# Group by county + state and get the max value
grouped = Covid19_Vaccine_Cleaned.groupby(['Recip_County', 'Recip_State'])['Series_Complete_Pop_Pct'].max().reset_index()

# Sort and get top 10
top_10 = grouped.sort_values(by='Series_Complete_Pop_Pct', ascending=False).head(10)

# Combine county + state for labeling
top_10['County_State'] = top_10['Recip_County'] + " (" + top_10['Recip_State'] + ")"

# Plot
plt.figure(figsize=(12, 6))
bars = sns.barplot(
    data=top_10,
    x='County_State',
    y='Series_Complete_Pop_Pct',
    hue='County_State',
    palette='Set2',
    legend=False
)

# Add data labels
for bar in bars.patches:
    height = bar.get_height()
    bars.text(bar.get_x() + bar.get_width() / 2, height + 1,
              f'{height:.1f}%', ha='center', va='bottom', fontsize=10)

plt.title("Top 10 Counties by Max Series_Complete_Pop_Pct (with State)")
plt.ylabel("Series Complete Population Percentage")
plt.xlabel("County (State)")
plt.ylim(0, 100)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



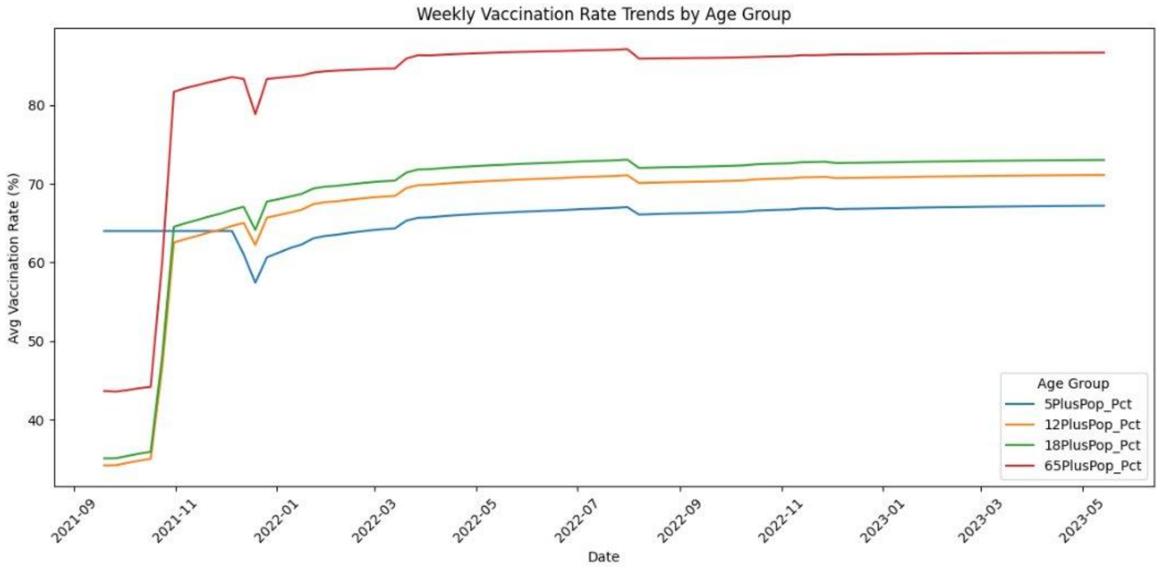
```python
# Step 1: Remove "NAN" strings
Covid19_Vaccine_Cleaned = Covid19_Vaccine_Cleaned[Covid19_Vaccine_Cleaned['Series_Complete_Pop_Pct'] != 'NAN']

# Step 2: Convert column to numeric and remove zeros or missing
Covid19_Vaccine_Cleaned['Series_Complete_Pop_Pct'] = pd.to_numeric(Covid19_Vaccine_Cleaned['Series_Complete_Pop_Pct'], errors='coerce')
df_cleaned = Covid19_Vaccine_Cleaned[(Covid19_Vaccine_Cleaned['Series_Complete_Pop_Pct'] > 0) & (Covid19_Vaccine_Cleaned['Series_Complete_Pop_Pct'].notna())]

# Step 3: Plot histogram with hue by Metro_status
sns.displot(
    data=df_cleaned,
    x="Series_Complete_Pop_Pct",
    hue="Metro_status",
    kind="hist",
    bins=20,
    multiple="stack",
    height=5,
    aspect=1.5,
    palette="Set2"
)

# Add labels and title
plt.title("Distribution of Fully Vaccinated Population by Metro Status")
plt.xlabel("Series Complete Population Percentage")
plt.tight_layout()
plt.show()
```

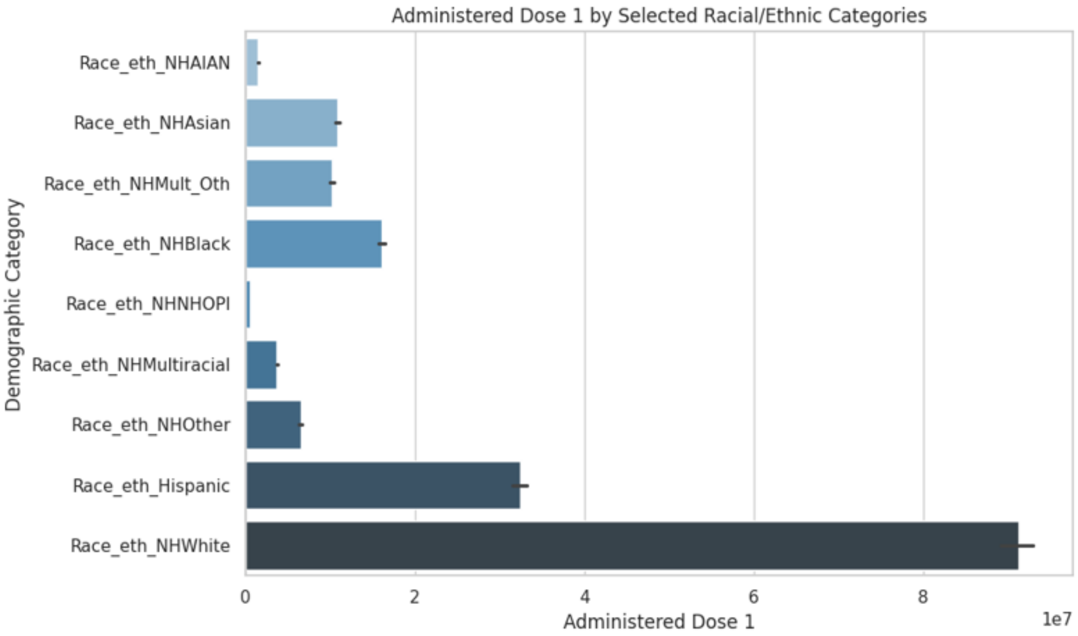


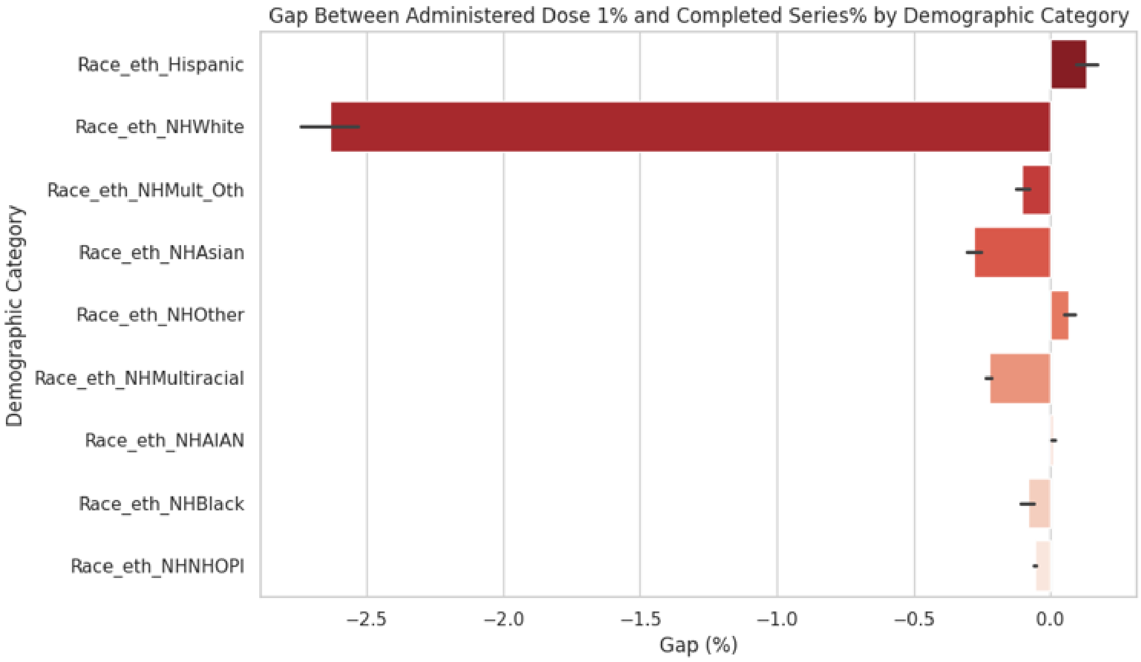
```python
# Select relevant age group columns
age_group_cols = [
    'Administered_Dose1_Recip_5PlusPop_Pct',
    'Administered_Dose1_Recip_12PlusPop_Pct',
    'Administered_Dose1_Recip_18PlusPop_Pct',
    'Administered_Dose1_Recip_65PlusPop_Pct'
]

# Convert to numeric and melt into long format
Covid19_Vaccine_Cleaned[age_group_cols] = Covid19_Vaccine_Cleaned[age_group_cols].apply(pd.to_numeric, errors='coerce')
Covid19_age_melted = Covid19_Vaccine_Cleaned[age_group_cols].melt(var_name="Age_Group", value_name="Vaccination_Rate")

# Remove missing or zero entries
Covid19_age_melted = Covid19_age_melted[Covid19_age_melted['Vaccination_Rate'].notna() & (Covid19_age_melted['Vaccination_Rate'] > 0)]

# Plot
plt.figure(figsize=(10, 6))
sns.boxplot(data=Covid19_age_melted, x='Age_Group', y='Vaccination_Rate')
plt.title("Vaccination Rates by Age Group")
plt.ylabel("Vaccination Rate (%)")
plt.xlabel("Age Group")

# Customize x-axis tick labels
plt.xticks(
    ticks=[0, 1, 2, 3],
    labels=["Age 5+", "Age 12+", "Age 18+", "Age 65+"],
    rotation=0
)

plt.tight_layout()
plt.show()
```

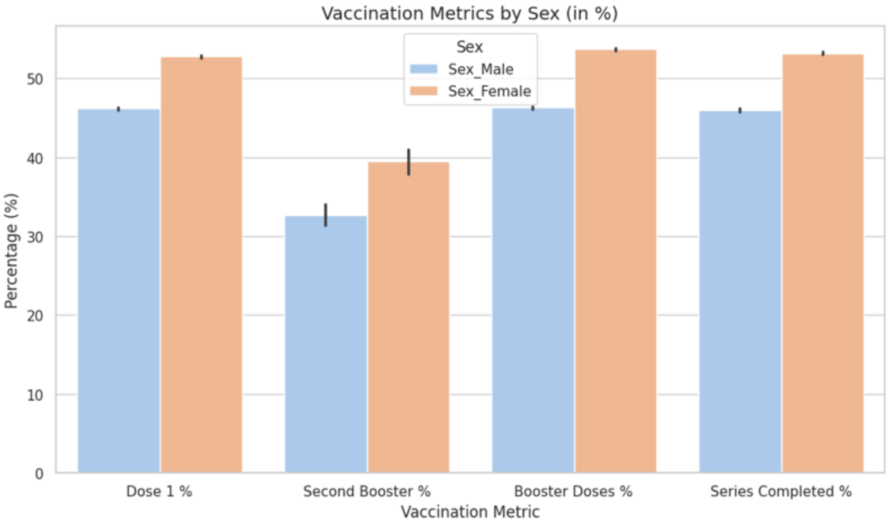


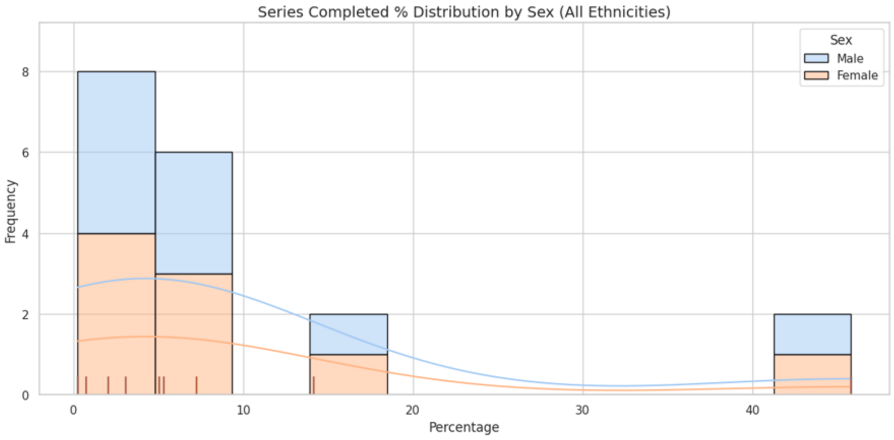
```python
# Calculate mean vaccination rate per age group
mean_rates = Covid19_age_melted.groupby("Age_Group")["Vaccination_Rate"].mean().reset_index()

# Plot bar chart
plt.figure(figsize=(8, 5))
sns.barplot(data=mean_rates, x='Age_Group', y='Vaccination_Rate', palette='Set2')

# Add value labels on top of bars
for i, row in mean_rates.iterrows():
    plt.text(i, row['Vaccination_Rate'] + 1, f"{row['Vaccination_Rate']:.1f}%", ha='center')

plt.title("Average Vaccination Rate by Age Group")
plt.ylabel("Mean Vaccination Rate (%)")
plt.xlabel("Age Group")

# Customize x-axis tick labels (Method 1)
plt.xticks(
    ticks=[0, 1, 2, 3],
    labels=["Age 5+", "Age 12+", "Age 18+", "Age 65+"],
    rotation=0
)

plt.tight_layout()
plt.show()
```


Weekly Vaccination Rate Trends by Age Group
Avg Vaccination Rate (%)
40
50
60
70
80
Date
2021-09
2021-11
2022-01
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
2023-03
2023-05
Age Group
5PlusPop_Pct
12PlusPop_Pct
18PlusPop_Pct
65PlusPop_Pct


Administered Dose 1 by Selected Racial/Ethnic Categories
Demographic Category
Race_eth_NHAIAN
Race_eth_NHAsian
Race_eth_NHMult_Oth
Race_eth_NHBlack
Race_eth_NHNHOPI
Race_eth_NHMultiracial
Race_eth_NHOther
Race_eth_Hispanic
Race_eth_NHWhite
0
2
4
6
8
1e7
Administered Dose 1


Gap Between Administered Dose 1% and Completed Series% by Demographic Category
Demographic Category
Race_eth_Hispanic
Race_eth_NHWhite
Race_eth_NHMult_Oth
Race_eth_NHAsian
Race_eth_NHOther
Race_eth_NHMultiracial
Race_eth_NHAIAN
Race_eth_NHBlack
Race_eth_NHNHOPI
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
Gap (%)


Vaccination Metrics by Sex (in %)
Sex
Sex_Male
Sex_Female
Percentage (%)
0
10
20
30
40
50
Dose 1 %
Second Booster %
Booster Doses %
Series Completed %
Vaccination Metric


Series Completed % Distribution by Sex (All Ethnicities)
Sex
Male
Female
Frequency
0
2
4
6
8
0
10
20
30
40
Percentage

****