Creators of the Data Visualizations: Sanjum Sahni (ss3873), Grace Wei (gtw25), Arushi Aggarwal (aa2555)

Picture of Data Visualizations:

Data Description:
We found two of our datasets on Our World In Data. This is the git data from Our World In Data: https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/README.md. Within this repository, there were multiple datasets pertaining to different COVID-19 data in the United States and also, around the world. We ended up choosing two datasets to focus on within this repository. The first dataset is called us_state_vaccinations.csv and the second dataset is called vaccinations_by_manufacturers.csv.

The reason we chose this data was because we were interested in looking at COVID-19 data in the United States to see the different trends in all the states. We were also interested to see how these trends have changed over time.

The us_state_vaccinations.csv contains different COVID-19 vaccination data from all the different states for each month from 2021 to 2023. The specific variables we focused on in this dataset were date, location, total_vaccinations, total_distributed, people_fully_vaccination, and total_boosters. These variables are pretty self explanatory. The location variable contains a state. The total_vaccinations variable contains the total_vaccinations for that date in that state. The people_fully_vaccinated variable looks at how many people became fully vaccinated in that state on that date. The total_boosters variable looks at how many booster vaccines were distributed on that date in that location. The total_distributed is how many vaccines were distributed in that state. In our code, we only wanted to focus on the month and year for the dates. For this reason, we had to aggregate the data per month for each year for each state; this required a series of dictionaries and arrays in order to properly store the date and aggregate it.

The vaccinations_by_manufacturers.csv contained manufacturer vaccinations data for all the countries. The variables we wanted to focus on in this dataset were location, date, vaccine, and total_vaccinations. The location variable contained the name of the country. The vaccine variable contained the manufacturer of the vaccine. The total_vaccinations variable contained the total number of vaccinations given to individuals in that country on that date. This dataset had data for all of the countries around the world. However, we just wanted to focus on the United States. For this reason, we did some data pre-processing to select only the rows pertaining to the United States. This new csv was renamed to us_vaccine_manufacturer_data.csv.

Below are screenshots from the data pre-processing:

```
import pandas as pd
```

```
df = pd.read_csv("/content/sample_data/vaccinations-by-manufacturer.csv")
```

```
def process_us_data(df, output_path='us_vaccine_data.csv'):
    # Filter for United States data
    us_data = df[df['location'] == 'United States'].copy()

    # Sort by date and vaccine
    us_data = us_data.sort_values(['date', 'vaccine'])

    # Save to CSV
    us_data.to_csv(output_path, index=False)

    # Also save to Excel if needed
    excel_path = output_path.replace('.csv', '.xlsx')
    us_data.to_excel(excel_path, index=False)

    return us_data
```

```
def main():
    # Create sample DataFrame

    # Process and save locally
    us_data = process_us_data(df)

    # Print preview of the filtered data
    print("\nPreview of US vaccine data:")
    print(us_data.head())

    # Print summary statistics
    print("\nSummary of US vaccine data:")
    print(f"Total rows: {len(us_data)}")
    print("\nVaccines included:")
    print(us_data['vaccine'].unique())
```

```
if __name__ == "__main__":
    main()
```

```
Preview of US vaccine data:
          location        date          vaccine  total_vaccinations
51645  United States  2021-01-12          Moderna             3835859
51646  United States  2021-01-12  Pfizer/BioNTech             5488697
51647  United States  2021-01-13          Moderna             4249795
51648  United States  2021-01-13  Pfizer/BioNTech             6025872
51649  United States  2021-01-15          Moderna             5122662

Summary of US vaccine data:
Total rows: 1530

Vaccines included:
['Moderna' 'Pfizer/BioNTech' 'Johnson&Johnson' 'Novavax']
```

Our last dataset was a topoJSON (us-smaller.json) file that we used to create the outlines for the shapes in our map visualization. This dataset came from the course repository. In order to connect our dataset (which only included the full names of states) to the state IDs in the topoJSON file, we had to create a dictionary mapping the state names to the state IDs, which we loaded in through the state_codes.csv file. This data was also sourced from the course repository, and modified to fit our dataset.

Visual Design Rationale:

For all three of our data visualizations, we wanted to see the trends in vaccinations in the United States during the peak Covid-19 years. For this reason, we chose three different types of graphs that could display these trends.

The first data visualization is a map that shows the vaccination data per state over the years. The reason we displayed this data as a map was so the user could easily visualize which states have more individuals who got vaccinated. We were able to show which states had more vaccinations by our color scale; the lighter colored states had less vaccinations and the darker colored states had more vaccinations. We used a sequential blue scale to show the numerical differences in the values; these were the quantile colors we use: "#caf0f8", "#90e0ef", "#00b4d8", "#0077b6", "#03045e". Additionally, for this data visualization, we created a color scale legend so the user would be able to tell that a darker colored scale is associated with more vaccinations. For this data visualization, we used three datasets: us_state_vaccinations.csv, the topoJSON file, and state_codes.csv. We first created the outlines of the states following the method shown in class. To populate the map with the data, we aggregated our data by summing up the values for each state (filtering on the years/feature that is selected), and then filled in the state with the right color. We used the dictionary we created from state_codes.csv to fetch the state that we are coloring. This visualization also included interactivity that will be explained in the next section. A drawback of this visualization is that the user has to click on buttons to see all the data that our visualization can display; it would have been nice if the user was able to see all the data at once in order to more easily compare the data.

The second visualization we had was a line graph showing the distribution of vaccines by state. The reason we chose a line graph was because the user would easily be able to see the trends (peaks, declines, dips) of which states had more distribution of vaccinations at which time; in order to display this we had the months and years displayed on the x axis and the number of vaccinations of the y axis. We had to tilt the x axis labels so all the different dates could fit on the axis. In order to help the user distinguish between the different lines representing the states, we used an ordinal scale with d3.schemeCategory10. This made each line a different color. The reason we used d3.schemeCategory10 was because it outputted colors that were easy to distinguish from for the user. We used the total_distributed variable from us_state_vaccinations.csv for the visualization. Similar to what we did for the map visualization, we had to aggregate the data per month per state. This visualization also included interactivity that will be explained in the next section. The drawback of this visualization is that you can either see one line on this visualization or you see all the state lines on the visualization; there is no in between. This is due to how we did our interactivity which will be explained in the next section.

The third visualization we created was a bar graph that showed distribution by manufacturer over time. We chose a bar graph so the user could easily compare the different COVID-19 manufacturers and their distribution over time. The y axis is total vaccinations and the x axis the date in month and year. We had to tilt the x axis labels so all the different dates could fit on the axis. The bars were colored based on the manufacturer. One again, we use d3.schemeCategory10 as our color scale to color in the different bars because we believe it offers the user a range of colors that are easy to distinguish. We also added a legend on the right side of the graph so the user could see which color represents which manufacturer in the bar graph. For this visualization, we used the us_vaccine_manufacturer_data.csv. For this dataset, we had to aggregate by month for each manufacturer; we also used dictionaries for this visualization. Then, we used this aggregated data to create the bar graph. The drawback of this visualization is that the Novavax manufacturer data is so miniscule compared to the other manufacturers that you cannot really see the data on the bar graph; the scaling extent worked out that way.

Interactive Elements and their Design Rationale:
In order to determine which interactivities we should have for our visualizations, we all sat down together to discuss what we would like to see. We decided that we wanted four buttons that controlled all the 3 data visualizations in order to show how each visualization changes over time. We also decided that we wanted hover features in our code so that users could gain more insight from our graphs. The rest of our interactivities are explained below.

We added multiple interactive components in our map visualizations. The first component we added was a hover feature. When the user hovers their mouse over a specific state, the state has a dark bold outline around it and a small black box shows up with the state name and the total number of vaccinations in it. This feature was added so the user could easily see which state they were focusing on and what the statistics for that state exactly are. Another feature we added for the map was a click option. When the user clicks on a specific state, it zooms into that state; this is done so the user can get a more clear image of the state. The third feature we added for the map were buttons. There are a total of 7 buttons that a user can use to further filter the map. The buttons are: 2021, 2022, 2023, All Years, Total Vaccinations, People Fully Vaccinated, and Boosters. The user can choose if they only wanna see the map of Booster vaccine distribution for the year 2021. The buttons give the user leeway to decide what portion of the data they want to focus on.

For the line graph visualization, there are also multiple interactions. The first interaction is that the user can hover over a specific line and the line will become bold while also displaying the name of the state in the rightmost corner of the svg. The second interaction we have is that the user can select which specific state they would like to select in the dropdown menu; we added this interaction so the user can get a more focused view of a singular state instead of having to look at all of the colorful lines at one moment. The last interaction we added is that this

visualization is also controlled by the 2021, 2022, 2023, and All Years buttons that the map visualization is controlled with. For example, if the user selects New Jersey in the dropdown for the year 2022 then the line graph only displays that respective data.

For our third visualization, the only interaction we added was that this visualization is also controlled by the 2021, 2022, 2023, and All Years buttons that the map visualization is controlled with. For example, if the user selects the year 2022 then only the manufacturer data for 2022 is displayed.

Since the 2021, 2022, 2023, and All Years buttons control all of the visualizations, we decided to add these buttons to the top of the page so they act like a navigation bar. This means that even when the user scrolls down, the buttons are still visible in a navigation bar at the top of the page. For all of the buttons, we also added a mechanism where when the button is clicked it turns blue; this is so the user is aware of the buttons currently selected.

For all of our interactivities, we made them discoverable by giving the user feedback. For example, for the buttons, when the user hovers over the button, it turns light blue thus showing the user that they can click on it. Similarly, for the map, when the user hovers over the graph, the state outline becomes bold and the state name shows up in a black box. All of our interactivities provide feedback which indicates to the user that an interactivity exists. We believe all these interactivities make our data visualization unique and interesting.

The Story:

Map Graph:

This map visualizes the total number of COVID-19 vaccinations distributed across the U.S. by state from 2021 to 2023 and all years. The states are shaded in a gradient, with darker shades representing higher total vaccination numbers and we have a legend to show this on a more detailed view. The map provides a clear geographic representation of vaccine distribution, highlighting state-by-state disparities in total vaccinations administered with a breakdown of different data values we represented, such as the total, people that were fully vaccinated, and the boosters given. One thing that was surprising to us was how Florida is one of the darker shades of blue with a higher number or vaccines for all the three data points even though they have a lot of anti-vaccine sentiment. Overall, the insights we want to convey to the viewer of our visualization are the geographic disparities in vaccination totals indicate varying levels of vaccine distribution success and challenges across states. These differences may be tied to factors such as population size, state policies, healthcare infrastructure, or public demand for vaccines. The map allows viewers to identify regional trends, such as higher vaccination totals in states with large populations or those with proactive public health measures, and see which states those are.

Line Graph:

The visualization showcases the monthly distribution of COVID-19 vaccines across the United States, with the ability to filter by specific years (2021, 2022, 2023, or all years combined). Between all the states the chart shows the fluctuations in vaccine distribution over time. It highlights the peaks, plateaus, and declines in vaccine rollout efforts. One surprising observation is that the highest actually came later in the vaccine distribution, when we thought it would be the initial rollout of vaccines when demand and urgency were at their highest. Also, looking at the all years chart, the rapid decline after the initial distribution afterward could suggest several factors, including saturation of eligible populations, logistical challenges, or vaccine hesitancy with the first rollout. The visualization also shows that certain states had disproportionately higher distributions, as indicated by the more prominent curves in the graph, which was an interesting insight to us as it correlates to population and the opinions of that state based on the vaccine rollout. Overall, we wanted to analyze to see the impact of the vaccine rollout on all the states in a more visual (line) format with all the years and each year the rollout changes.

Bar Graph:

This visualization illustrates the distribution of COVID-19 vaccines over time by manufacturer, showcasing the total number of doses administered for each manufacturer (Moderna, Pfizer/BioNTech, Johnson & Johnson, and Novavax) across the months from 2021 to 2023 and also one for all years. The data highlights how different manufacturers contributed to the vaccine rollout efforts and how their distribution changed over time. We found a couple of things surprising to us and one is how Pfizer/BioNTech appears to have consistently led the vaccine distribution effort, with significantly higher numbers of doses administered compared to Moderna, Johnson & Johnson, and Novavax even though J&J is such a big and well-known company. Also, there was a sharp decline in vaccine distribution across all manufacturers after early peaks in late 2021 and early 2022, which could show diminishing demand or the achievement of certain vaccination milestones. Overall, the insights we want to convey to the viewer of our visualization is how the reliance on specific manufacturers, particularly Pfizer/BioNTech and Moderna, underscores their critical roles in the vaccination campaign and we want to correlate the vaccine rollout from the previous graphs above into more how the companies were affected by the decline in vaccine distribution after the initial surge. This suggests a transition from mass vaccination efforts to targeted campaigns, possibly due to reaching population saturation or decreased urgency and we have been able to show that through our continuous story.

Team Contributions:
All of us contributed significantly to this project. We met up each week to discuss plans of what we wanted to accomplish each week. Initially, we all worked together to come up with a project idea we wanted to focus on. Then, we all worked together to find datasets that we could

potentially be interested in using for our project. Then, we all planned out how we wanted to create our data visualizations and what exactly we wanted in terms of interactivity. The specific contributions of what each team member contributed will be listed below. However, we all worked on many portions of this project together through Liveshare on VScode. The parts of the project that took the most time were actually finding a dataset that we wanted to work with, creating the data visualizations and their respective interactivities. We were initially unsure about which dataset we wanted to work with, but ultimately we chose one that interested all of us. We had some ideas of what we wanted to do with our interactions. However, we did not necessarily know how to go about carrying through with these ideas. For example, we wanted to make a slider for the dates so the user could slide the slider to which specific date they wanted to look at; however, we ran into multiple errors with this approach and decided to scratch the idea. Instead, we went for a button approach. Additionally, we ran into some trouble with displaying our data in the map. In class and for homeworks, we were always given very clean data to work with; however, the data for this assignment required a lot of nested dictionaries and arrays in order to organize the data. We also did not know how to get the map to change along with the buttons; we eventually figured this out through the use of function calls and global variables. These were our most time consuming portions of the project.

Sanjum's Contributions and Time Spent:

- Sanjum helped come up with an idea for the project and helped look for datasets. Sanjum also helped plan out an outline for the data visualizations and their respective interactivity. She also contributed to creating the first data visualization, the map. Sanjum created the different buttons for the map so the map updates when the buttons are pressed. She also created the black box on the map that shows the name of the state and the total. Sanjum also created the dropdown menu for the line graph. Sanjum wrote this report which contains a summary of our data visualization.
- Time Spent on Project Overall: 15 hours

Grace's Contributions and Time Spent:

- Grace helped come up with an idea for the project and helped look for datasets. Grace also helped plan out an outline for the data visualizations and their respective interactivity. She also contributed to creating the first data visualization, the map. Grace also figured out the zoom in and zoom out feature for the map. Grace helped fix some bugs in the line graph. She also created the legend for the bar graph. She also made it so the buttons are connected to all three of the graphs.
- Time Spent on Project Overall: 15 hours

Arushi's Contributions and Time Spent:

- Arushi helped come up with an idea for the project and helped look for datasets. Arushi also helped plan out an outline for the data visualizations and their respective interactivity. She also contributed to creating the first data visualization, the map. Arushi also figured out the hovering feature of the map. Arushi helped create the bar graph and

helped with the line graph also. She created the hover feature of the map. She also did the majority of the css/aesthetics for the project.

- Time Spent on Project Overall: 15 hours