Yongru Chen, Chuman Xie, Yitian Liang

Professor Felix Muzny

DS2001

April 13th 2021

Report

**Problem Statement and Background**

Countries in the world need to recover from the pandemic. Vaccines are the key element for rapid recovery. The vaccine is one of the most popular topics today and it has been distributed since December 2020. Each state would get vaccines on a weekly basis. There were three types of covid-19 vaccines in the United Stated but we will focus on Moderna only which needed people to get two shots for vaccination. We will be presenting on how many covid-19 vaccines have been distributed to each State? And which States would get the most of them in total? In the data set, there are clearly four columns. The first one is states' names. The second column is the date of vaccine distribution. The third column is the amount of how many the first dose they received. The fourth column is the amount of how many the second dose they received. In addition, there are no ethical issues that our group identifies around the manner in which the data was collected.

**Introduction and Description of the Data**

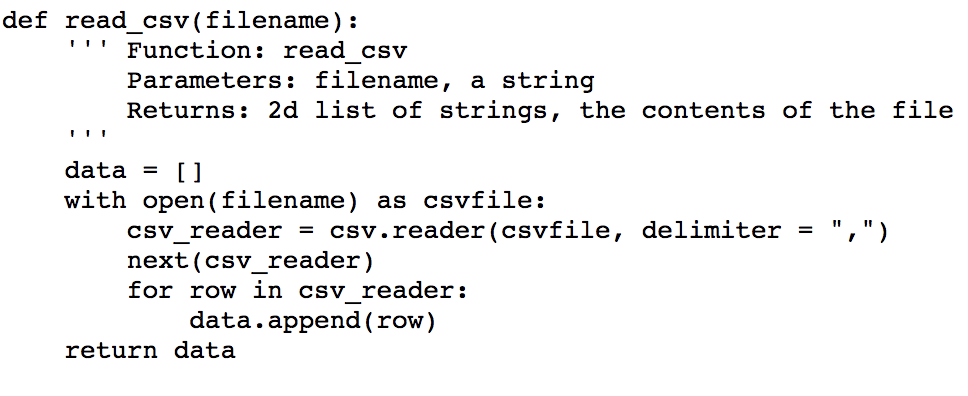
At this time, most people wanted to get vaccinated to prevent themselves from getting coronavirus. Especially, people are eager to get back to work and return to school after one year of suffering from the pandemic. Businesses and companies also need to reopen again and they wanted to come back safely. Moreover, we just got an email from the school telling us to get vaccinated before the first day of school in the fall semester. All those things made us realize how important vaccines are. Therefore, we wanted to know more about this topic. At the same time, while our group presents these data and graphs to the whole class, we could also play a part in telling them how important vaccines are. The data set that we used in our project is from a website and we download it in csv format.

Every single country in the world has been focusing on this pandemic. This pandemic has been badly affecting us for more than a year. If we could upload the data set and analyze it perfectly, we could inform other students in class how serious this is and how important vaccines are.

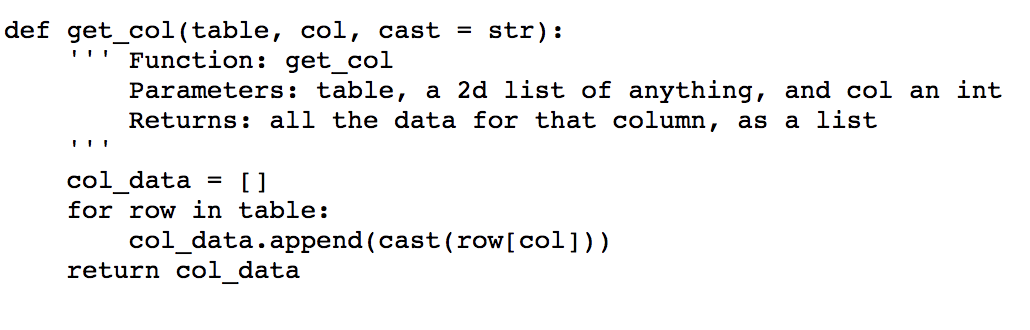
**Methods**

Approaching our question, we used several methods to interpret our desired data by utilizing the concepts that we have learned from courses DS 2000 and DS 2001, which include conditionals, dictionaries, functions, lists, while loops, variables, and data visualization with Matplotlib etc..

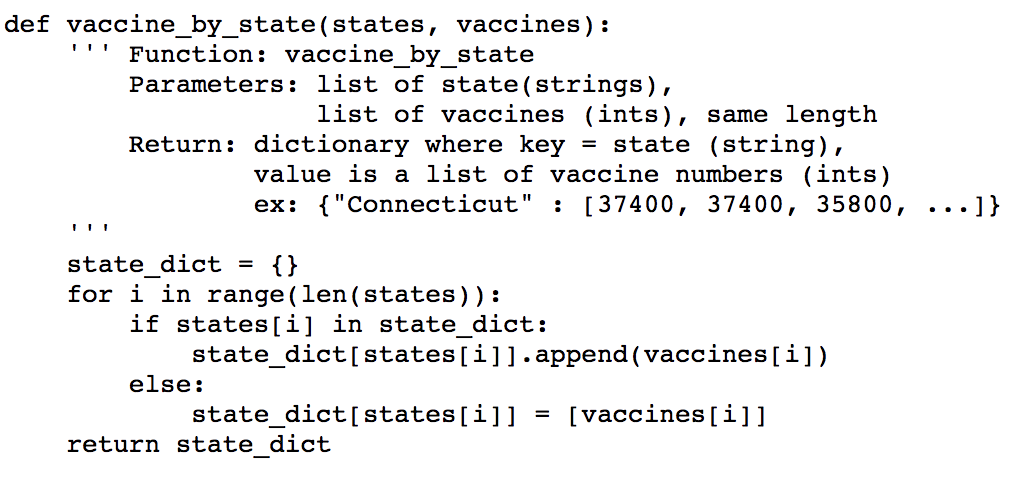
As mentioned before, our data set is a csv file containing data for “states”, “dates”, and “numbers of 1st & 2nd doses vaccines”. Therefore, we wrote a function called “read\_cvs” to make csv file be read into Python as below:



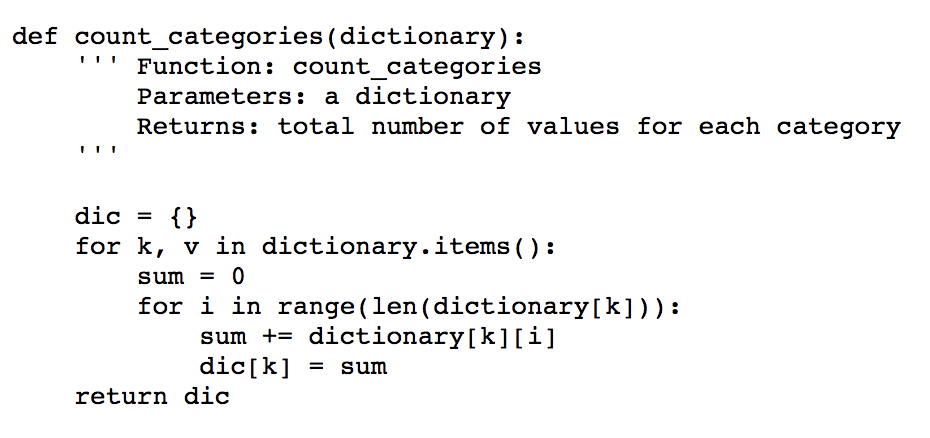
Next, in order to get the data from the first, second, and third columns of csv file, we wrote a function called “get\_col” to get all data from each column as below:



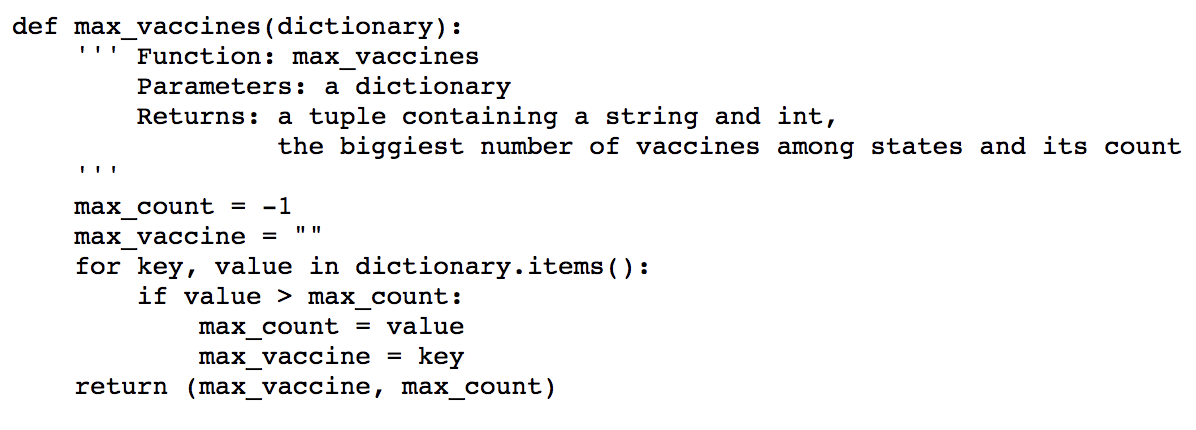
Then we had to apply the dictionary concept to get our third function done called “vaccine\_by\_state”. This function allowed us to get a list of states and a list of numbers of vaccines that both have the same length. The key of the dictionary would be the state (strings) while the value would be a list of the numbers of vaccines (integers). The return would be the new dictionary that we just created. If the numbers of vaccines came from the same state, they would be appended together to become a list, which would be the value for the key (= state).



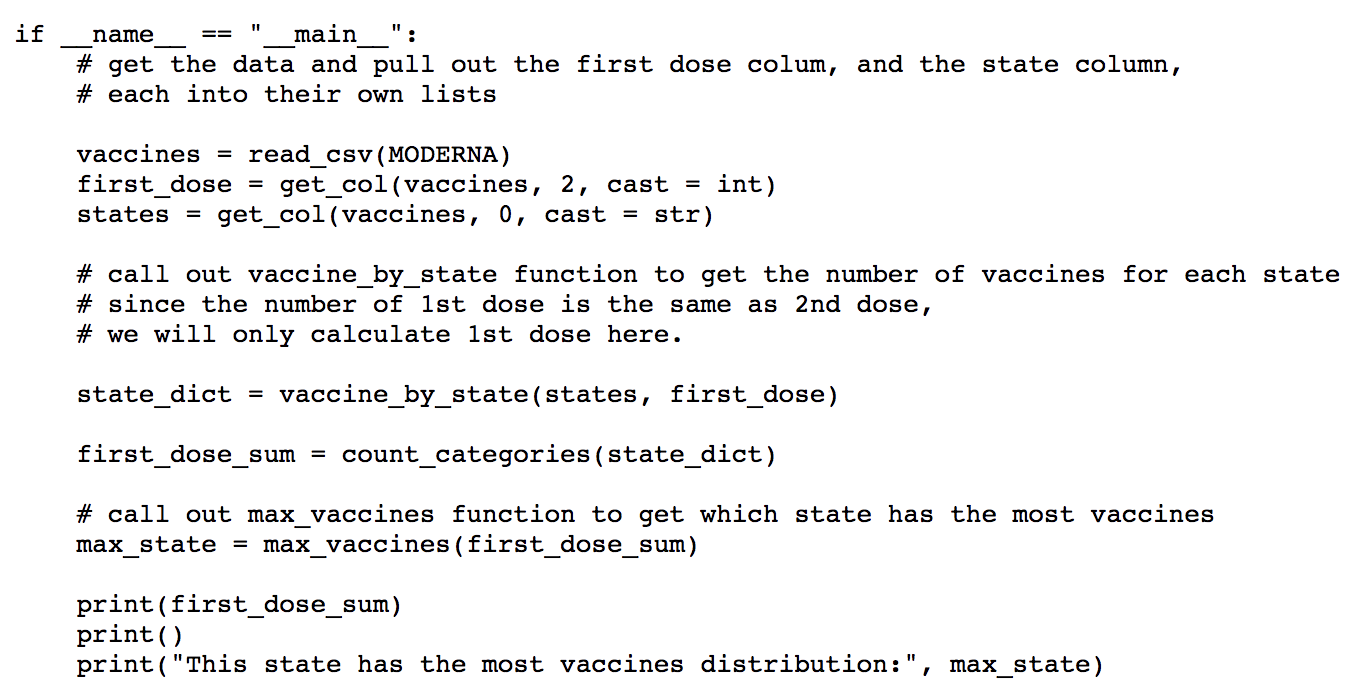
Our fourth function, “count\_categries” would be counting the sum of vaccines that were distributed for each state from 12/21/2020 to 4/12/2021. We combined the dictionary and “for” loop functions here to make it work. First, we created a new dictionary. Then we set the value as the sum of numbers of vaccines and the key as the state. The return would be the dictionary that we just created.



We almost had everything, but our ultimate goal was to find out “which state has got the biggest number of vaccine distribution”. We wrote a function called “max\_vaccines” that set a dummy number for max\_count firstly, then whenever a bigger number came out, it would replace the old number to become a new value. The return would be a tuple containing a string (max\_vaccine) and an integer (max\_count).

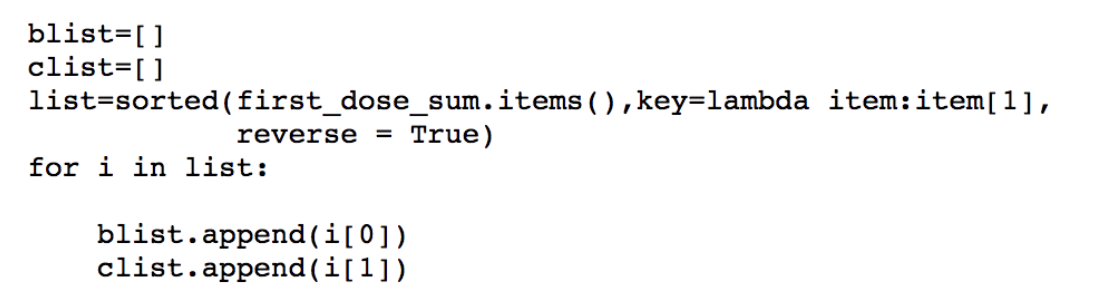


After we called out all the functions above and obtained all the data that we wanted (see below), we created three graph visualizations by using Matplotlib - two of them were bar charts that shared many similarities with only one difference (one was out of order and the other was in order), and one of them was a pie chart.



The image output is the easiest part of the project because it doesn't require a lot of thought. Our initial intention was to use the chart to reflect the sum of vaccines distributed for each state from 12/21/2020 to 4/12/2021. The only thing worth thinking about is how to communicate our message to everyone better.

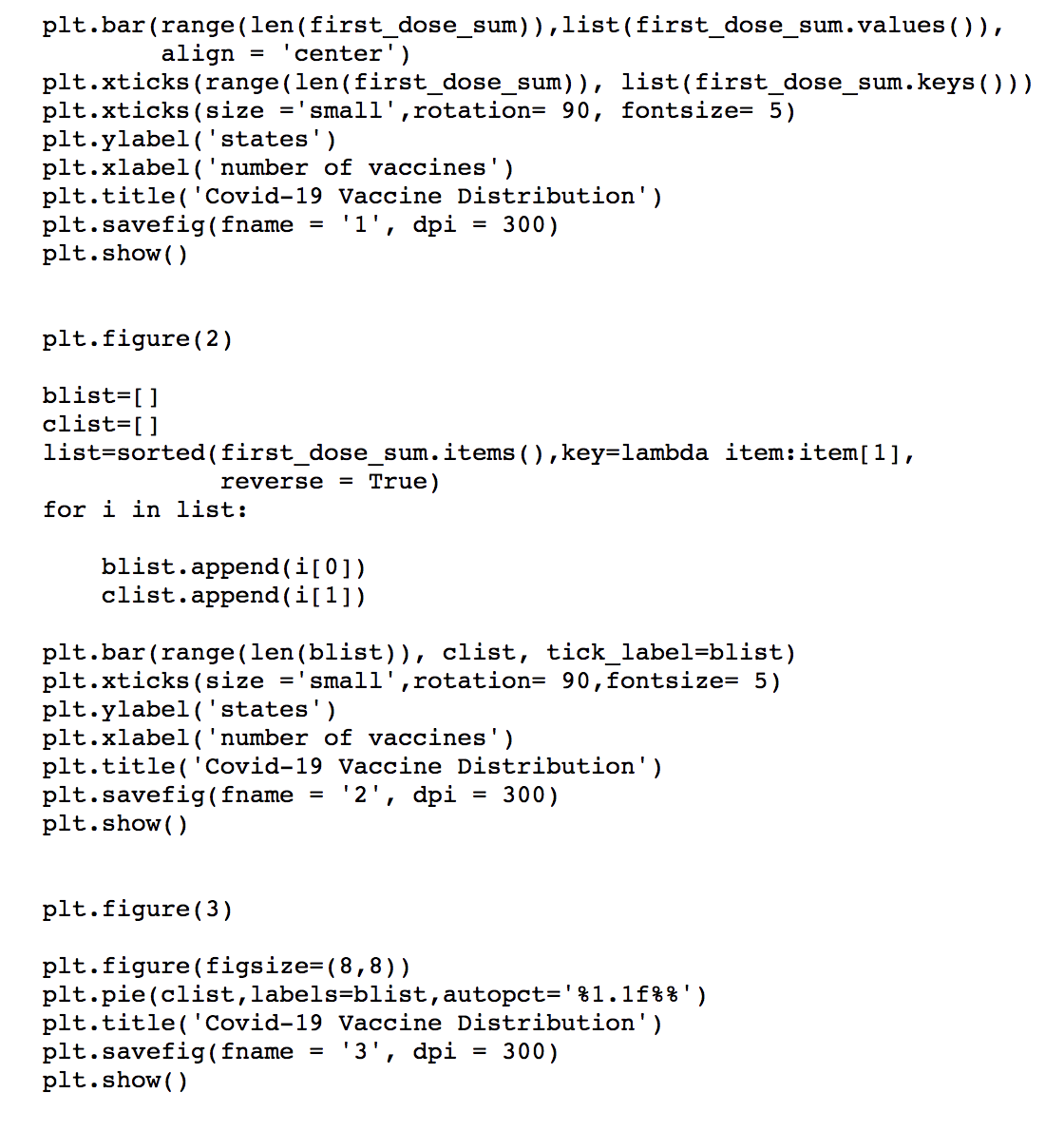
We thought a lot when we wrote Figure 1 and Figure 2. For the first two charts, we can see how many vaccines each state received and how they ranked. This data is necessarily related to many aspects, such as the population of each state and its economic strength and so on, facilitating our discussion in the future.



Since we have too much data, reacting with histogram will cause all sticks to overlap each other, so we choose to rotate them by ninety degrees for more convenient viewing.I went to find a way to divide the histogram, creating two new lists, and reorder them. The leader of the team has made more explicit images and improved the resolution of the pictures.

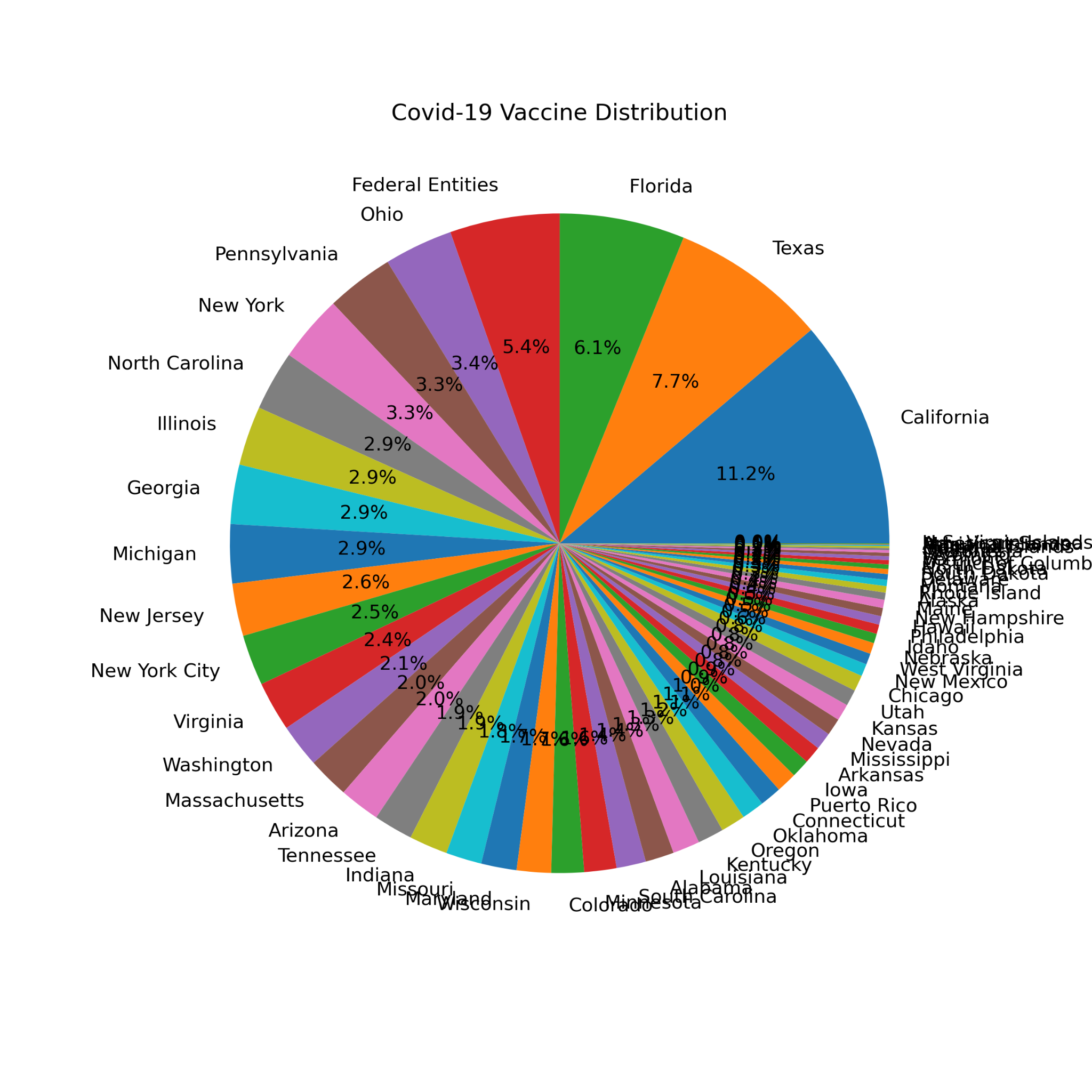
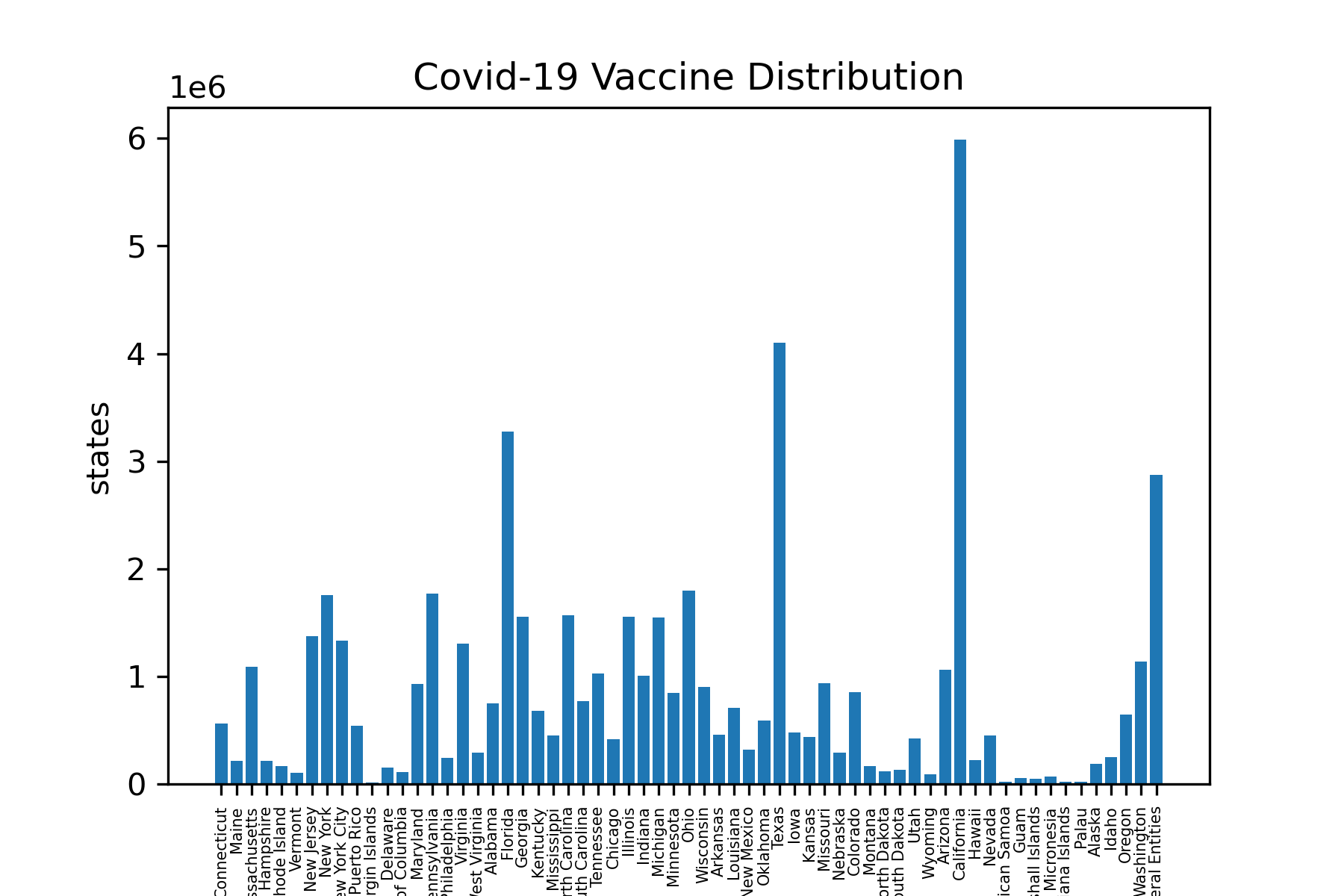
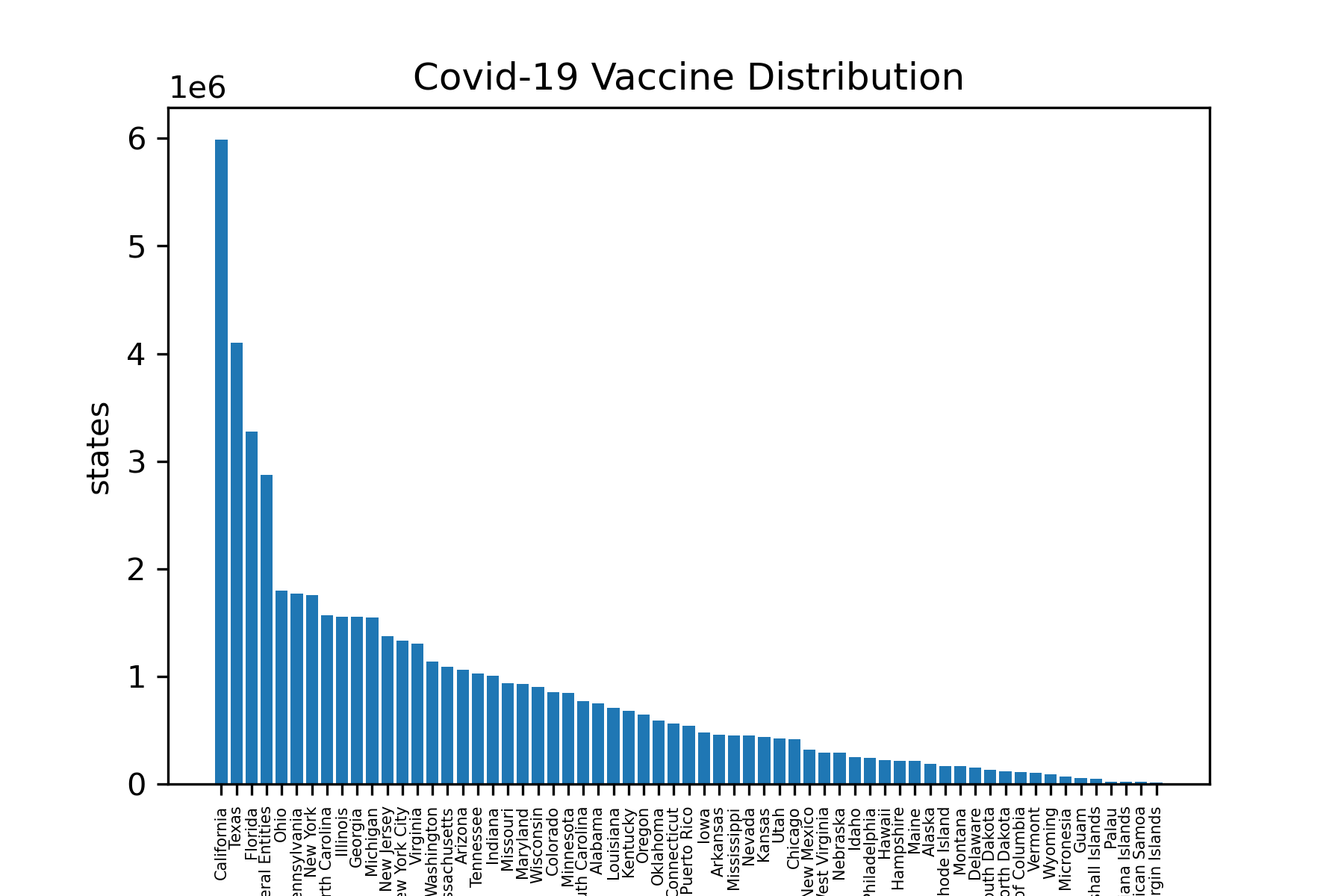
We agree that we need to reflect the proportion of the total number of vaccines received by each state, so the fan chart is an excellent choice.This is the inspiration for our figure 3.

In this pie chart, we can well focus on the top ten states for research. Because it visually represents the data as a small part of the whole. The audience sees the data comparison at a glance, enabling the audience to analyze immediately or quickly understand the information. On making the pie chart, we can also use python to manipulate the data fragments in the circle of the pie chart to keep them away from the center of the circle to emphasize the points we want to express.



**Results, Conclusions, and Future Work**

Our result shows that California has had the most vaccination distribution (5,987,400 for the 1st dose) since 12/21/2020 followed by Texas (4,099,900 for the 1st dose) and Florida (3,275,100 for the 1st dose). Even though New York and New York City have the highest prevalence of COVID-19 infection among states, they seem not to have enough vaccine distribution (1,752,200 and 1,330,200 for the 1st dose) compared to California and Texas. Massachusetts is rated in the 15th position (1,086,100 for the 1st dose). It is interesting to note that the data only comes from 38 states while some other states in the U.S. do not participate in the data collection of Moderna allocation. The reason may be that these states use different brands of vaccines such as Pfizer or Johnson & Johnson, or the number of vaccinations is too small (<1000) compared to other states so it could be ignored in the data set. Based on the news on April 1st, 2021, California has administered more than 18 million doses as of Thursday and 6.7 million people are fully vaccinated (U.S. News, 2021). This result is different from our result because we only count the Modena allocation but exclude other brands of vaccines such as Pfizer and Johnson & Johnson.



The strengths of our project include understandable functions written, concise code languages, and clear visualization design. We use only five functions to obtain our final results and each function has a brief coding. Readers can easily understand these graphs.

However, we also have some limitations in our project. First, our graph is not clean enough. We tried to make an American map showing numbers of vaccines in different states but we failed. In the bar chart, we can’t interpret the exact number of vaccines for each state since the number is too large. We can only estimate the number of vaccines to 10^6. Also, in the pie chart, there are too many states so the states with smaller numbers are overlapping together, which makes the chart hard to be read. If we had more time given, we might address these shortcomings by attending Dr. Muzny’s tutoring section (we attended one of the TA tutoring sections but the person was not familiar with making the map that we wanted) and make a highly visualized representation such as an American map to show the data.

Now, more than 99.6 million people, or 30% of the U.S. population, have received at least one dose of a coronavirus vaccine, according to the Centers for Disease Control and Prevention. Some 56 million people, or 16.9% of the population, have completed their vaccination (U.S. News, 2021). By conducting this final Python project, we are seeing the hope of getting herd immunity in the upcoming months soon and we are all excited to look forward to having life back to “Normal”.

References:

1. COVID-19 vaccine Distribution allocations by jurisdiction - MODERNA. (n.d.). Retrieved April 13, 2021, from <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws>
2. The latest: California has given over 18m vaccine doses. (2021, April 1). Retrieved April 13, 2021, from <https://www.usnews.com/news/health-news/articles/2021-04-01/the-latest-south-korea-mulls-rapid-tests-for-home-use>
3. Strange L. (2021). *DS2000 Spring 2021\_Handout*
4. Muzny F. (2021). *DS2001 Spring 2021\_Handout*