INFO 5709
Project 3 - Part 1

COVID-19 Vaccinations and their trends in the United States

by

Venkata Sri Ambica Mokarala
Information science
University of North Texas

Table of Contents

Introduction:

Roll out of vaccinations has given hope to humanity against the deadly unknown virus COVID 19 which started in 2019. Around 6M lives across the world and 900k lives just in the United States were lost because of the virus and around 502M people got affected by COVID 19 so far. Different nations started administering vaccinations to help the world to go back to normal conditions. Did vaccination really end the ongoing pandemic completely? Even though there is a lot of controversy around vaccinations and the side effects caused by vaccination. There is still so much research that shows that there is a positive effect of vaccinations on the COVID 19.

In this paper, we study how the vaccination campaigns helped the world in the fight against the deadly virus. The datasets used for this research will be collected from the official sources that contains reliable data about COVID 19 vaccination administered in the United States that are frequently updated. We find out whether the vaccinations have a reduced effect on the transmission of the virus and help the world by ending the pandemic.

Analyzing the official datasets that provide reliable and updated sites about the covid 19 vaccinations may help in studying how the vaccination affected the COVID cases and COVID deaths. How vaccinations are administered has influenced or affected the transmission of the Coronavirus and save the lives of many people. This paper is organized as follows analyzing and understanding the trends of vaccinations across the different states of the United States, background of the topic, problem statement, limitations of the research, conclusion, and future work of the research.

Background:

The first case of novel coronavirus SARS-CoV-2 (COVID 19) was discovered in Hubei Province of the People's Republic of China and then started spreading across the world in a very less time. Coronavirus had caused havoc and turned the world upside down. COVID 19 affected the world in all the possible ways one could ever imagine. Closing borders between countries, Masks, and quarantining, drastically affecting the world's economy, killing millions of people, disrupted the education system, caused burnouts for healthcare professionals etc. United States is one among the many countries that have been adversely affected by COVID 19. The United States was the country with the topmost number of COVID cases followed by India. This virus has adverse health effects including death which especially is seen in the elderly and people with chronic illness.

Although many preventive measures were started to stop the spread of the pandemic, Scientists, and health professionals all around the world started researching about the virus trying to find cure for this deadly disease and came up with vaccination within no time to set the world back to normal. Vaccines are forever known for helping the human body to fight against the viruses. Vaccinations for Covid 19 had been developed and delivered at a very high pace than any other vaccines had been in the history. Vaccines were approved by CDC and vaccine administration started in phases by the end of December 2020. Vaccinations are administered by government in phases starting with for people with high-risk factors like elderly people and people with chronic illnesses. Mainly two types of vaccines were approved by CDC and administered by the government. We can say that distributions of vaccines improved the living conditions as well as economic conditions of the United States.

As we see several variants are emerging in different parts of the world like Delta, Omicron, etc. There are many conflicting opinions about how effective vaccinations are for the new emerging variants. A lot of misinformation regarding the Coronavirus, coronavirus vaccinations, and side effects caused by the vaccinations was spread on social media making it difficult to differentiate from the reliable information.

Research Problem:

Even though there are many questions and misinformation regarding the effectiveness of vaccination, efficacy of vaccinations against other emerging variants, chances of reinfection after vaccination. Finding out the effects vaccinations so far had on the pandemic may clear some of the misinformation about the COVID vaccinations and encourage people to find the reliable information. The research mainly focuses on providing the clear understanding to the readers about how effective the vaccination campaigns are in stopping the spread of the Coronavirus in different states of the United States. The following are the possible issues related to vaccination that are presented in this research.

1. How are the vaccination trends in different states across the United States?
2. What are the 5 states with the Highest vaccination and least vaccination?
3. What is the percentage of people who are vaccinated in the total vaccines distributed in different states?
4. What vaccine is more distributed or administered (Moderna/Pfizer/Johnson-Johnson) in different states of US? What is most distributed vaccine in United States?
5. Analyzing Country wise distribution of COVID 19 vaccines.

Related Work:

The study of vaccination strategies to help fight the virus is not a new concept. There are many studies conducted to study various types of vaccine distribution strategies that are aligned or accepted by the public. Eshun-Wilson conducted research to find out many aspects associated with the COVID 19 vaccine distribution strategy (Eshun-Wilson et al., 2021). In their research, different variety of problems related to this vaccine distribution and public preferences about the vaccination. Vaccination doses people prefer, opinions about vaccinating annually, and wait times at vaccination sites are some of the research problems that are explained in their paper. Methods such as weighted sampling, mixed logit, etc. are used to find solutions to the research questions of their study.

Another similar study was conducted by Chen to research the effects that are caused at the early stage of vaccination by estimating the reduction of the cases with respect to vaccination and predicting the herd immunity in the United States (Chen X et al., 2022). In this study, the authors used Susceptible-Infected-Recovered (SIR) model to predict herd immunity. This study concluded that herd immunity can be achieved by increasing the vaccination pace and a decrease in the vaccine hesitancy is needed to attain herd immunity.

Using exploratory data analysis combined with visualization to study the different types of covid 19 trends for studying a research problem is also not a brand-new concept. According to Dey S. K, it is important to analyze the data available data on COVID to increase situational awareness among the communities (Dey S. K., 2020). The authors of the research used EDA and visuals to study different aspects of the COVID 19 datasets to provide an analysis of the different vaccination programs going around the world. Different visualizations like bar plots and heatmaps together with different python libraries such as NumPy, Seaborn, etc. are used to generate the visuals and as well as to clean the data that is collected from different sources.

There are also many studies that are done in a short time to analyze the COVID data after the virus outbreak. In research to find out what factors or variables are most effective and least effective to control and contain the pandemic and reduce pandemic effects. The authors Nazir (Nazir et al., 2021) used visual exploratory data analysis(V-EDA) for their research as V-EDA is user-friendly and easy to understand. The data collected was analyzed by using t-tests for finding out statistical differences between the datasets and the spearman test to analyze the correlation between the datasets. The research showed how different factors such as testing rates, restrictive measures, aged population, school closings, public event cancellations, facial coverings, etc. acted as significant variables/factors in controlling the virus. (Nazir et al., 2021)

Methodology

The datasets collected for analysis of the COVID 19 vaccinations in the United States are gathered from different. These datasets are collected from the reliable and official sources like Kaggle, CDC, California Health, and Human Services Open Data Portal, Data.gov.

Exploratory data analysis is performed on the dataset. The tools used for this project are python and PowerBI.

Exploratory data analysis:

Exploratory data analysis is the method used to explore the data, identify patterns in the data, visualize the data and use the data for modelling, identify underlying features in the data and find the relationship between the datasets. The data collected from different sources need to be imported and merged if necessary to visualize the information present in the data.

Python libraries like pandas, NumPy, seaborn, matplotlib are used to perform Exploratory data analysis on the datasets collected. Pandas, library of python is used to work and manipulate the datasets. NumPy library is used for dealing with numerical values present in the data set Seaborn and Matplotlib libraries are used for data visualization. Seaborn provides beautiful default styles and color palettes to make the visualizations more attractive. The functions like drop () are used to drop all the irrelevant columns of the dataset.

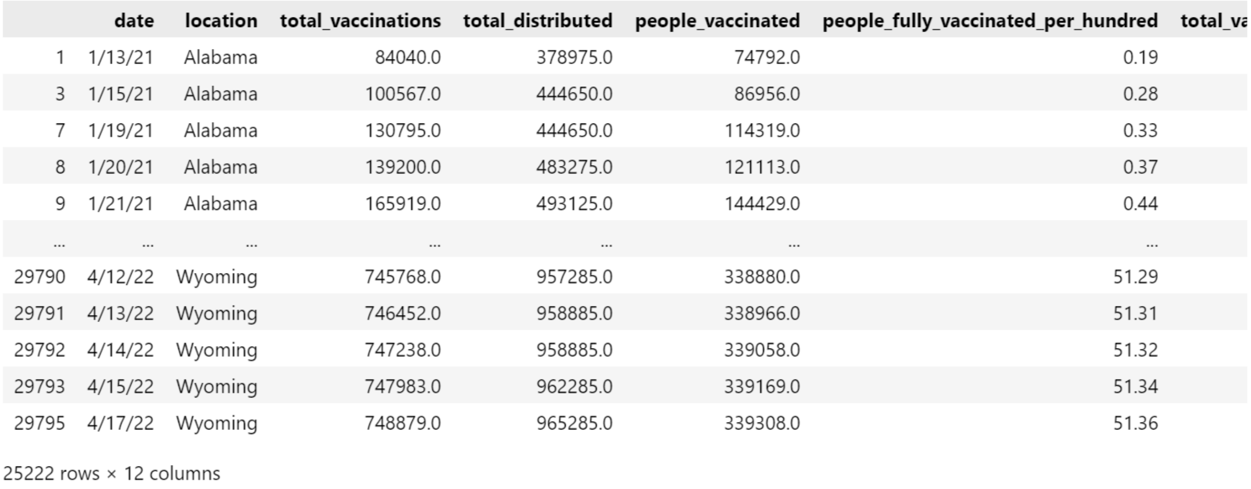
PowerBI provides all the tools and graphs to create different types of visuals which helps in efficient communication of data. PowerBI extracts data from different formats of data such as excel, csv, web etc. The extracted data can be transformed by ETL process which makes the e data more understanding and in return makes the visuals efficient. The visuals combined with EDA provides a efficient way to analyze the information and underlying data. However, it is important that both the simple and complex ideas should be communicating with clarity, precision and efficiency.

Data collection and Data cleaning:

The first step of any Exploratory Data Analysis is to collect data. graphical excellence principle requires telling truth about the data. So, it is important that the data collected should be good and reliable. The datasets collected from the sources contain many missing/null values, empty values, garbage values. It is important to clean the data before using for modelling or visualization. The datasets after cleaning are as shown in the below figs.

Dataset 1

This dataset contains the information about the total distributed vaccinations, total people vaccinated, date by locations which is useful for answering the research questions. There are also other columns in the dataset like people vaccinated by hundred, daily vaccinations etc.



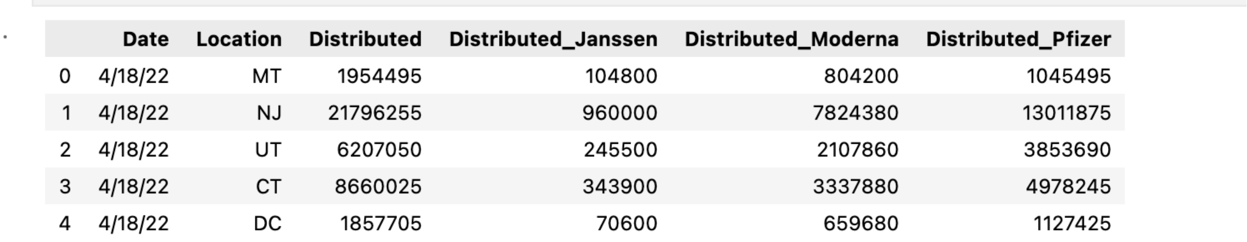
	date	location	total_vaccinations	total_distributed	people_vaccinated	people_fully_vaccinated_per_hundred	total_va
1	1/13/21	Alabama	84040.0	378975.0	74792.0	0.19	
3	1/15/21	Alabama	100567.0	444650.0	86956.0	0.28	
7	1/19/21	Alabama	130795.0	444650.0	114319.0	0.33	
8	1/20/21	Alabama	139200.0	483275.0	121113.0	0.37	
9	1/21/21	Alabama	165919.0	493125.0	144429.0	0.44	
...	...	...	...	...	...	...	
29790	4/12/22	Wyoming	745768.0	957285.0	338880.0	51.29	
29791	4/13/22	Wyoming	746452.0	958885.0	338966.0	51.31	
29792	4/14/22	Wyoming	747238.0	958885.0	339058.0	51.32	
29793	4/15/22	Wyoming	747983.0	962285.0	339169.0	51.34	
29795	4/17/22	Wyoming	748879.0	965285.0	339308.0	51.36	

25222 rows × 12 columns

Fig Dataset 1(vacdetails.csv)

Dataset 2

The dataset that contains type of vaccines **vaccine.csv,** dropped different columns that are not useful for the analysis and only columns with useful information as shown below.



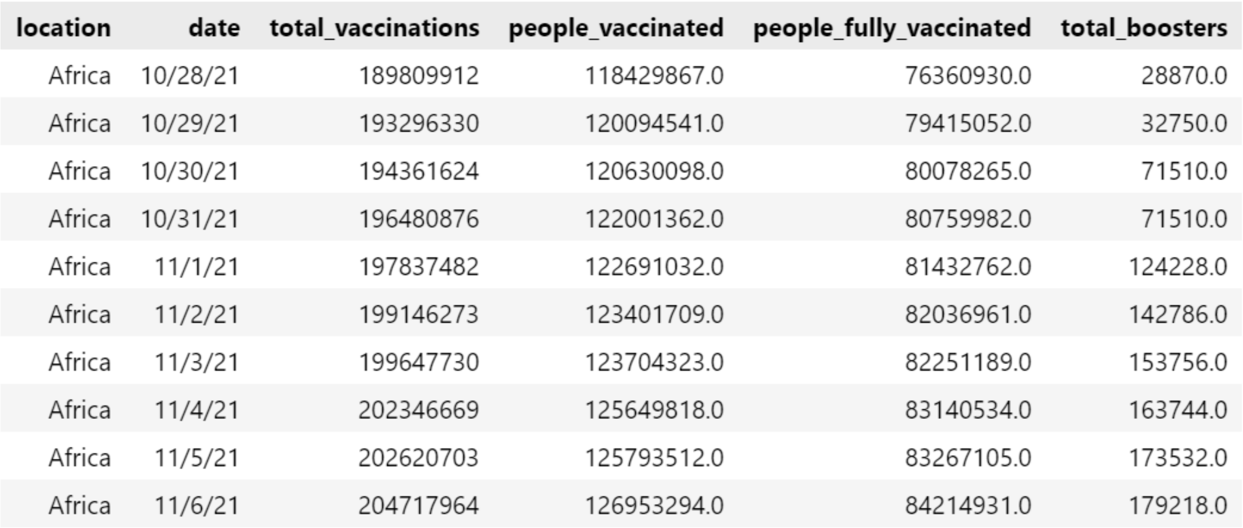
	Date	Location	Distributed	Distributed_Janssen	Distributed_Moderna	Distributed_Pfizer
0	4/18/22	MT	1954495	104800	804200	1045495
1	4/18/22	NJ	21796255	960000	7824380	13011875
2	4/18/22	UT	6207050	245500	2107860	3853690
3	4/18/22	CT	8660025	343900	3337880	4978245
4	4/18/22	DC	1857705	70600	659680	1127425

Fig Dataset 2(vacname.csv)

This dataset contains what type of vaccination used in different states. This dataset can be used to study the research question about which vaccination is used most in the United States.

Dataset 3

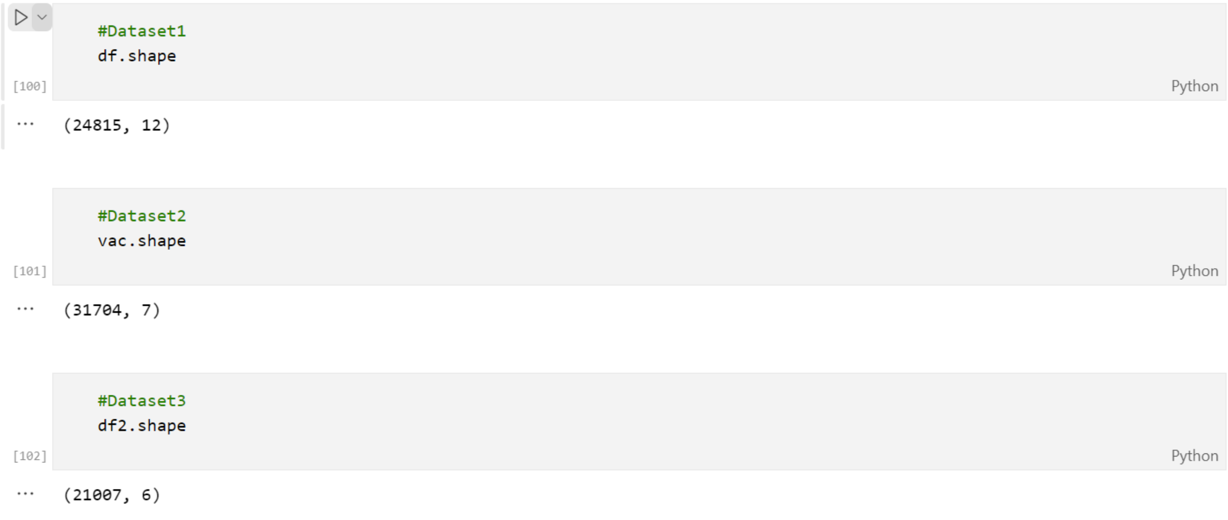
This dataset consists of total vaccinations, date, people vaccinated by location just like the first dataset but for countries across the world instead of states in United States.



location	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters
Africa	10/28/21	189809912	118429867.0	76360930.0	28870.0
Africa	10/29/21	193296330	120094541.0	79415052.0	32750.0
Africa	10/30/21	194361624	120630098.0	80078265.0	71510.0
Africa	10/31/21	196480876	122001362.0	80759982.0	71510.0
Africa	11/1/21	197837482	122691032.0	81432762.0	124228.0
Africa	11/2/21	199146273	123401709.0	82036961.0	142786.0
Africa	11/3/21	199647730	123704323.0	82251189.0	153756.0
Africa	11/4/21	202346669	125649818.0	83140534.0	163744.0
Africa	11/5/21	202620703	125793512.0	83267105.0	173532.0
Africa	11/6/21	204717964	126953294.0	84214931.0	179218.0

Fig Dataset 3(countrydeets.csv)

Python functions such as shape (), info (), describe () etc. can be used to explore the dataset in python IDE. These functions help us to analyze the data about number of rows and columns in the dataset, type of variable in the dataset, names of variables/columns etc. By understanding the data, it becomes easy to select the required tools and techniques for data modelling and visualization.
Using visuals to explore data is much more helpful for analyst to understand the data in short time.



```
#Dataset1
df.shape
```

```
(24815, 12)
```

```
#Dataset2
vac.shape
```

```
(31704, 7)
```

```
#Dataset3
df2.shape
```

```
(21007, 6)
```

Fig. Datasets details

Univariate Data analysis:

Univariate data analysis is analysis of one variable at a time. Single variable of interest is taken and analyzed to find the basic as well as underlying information about the data like mean, standard deviation, outliers missing values, distribution etc. Univariate analysis can be performed in two ways: graphical methods and non- graphical. Non-graphical Univariate data analysis includes using statistics and frequency tables to analyze the variable. Whereas in graphical Univariate data analysis involves using graphs such as histograms and bar plots to analyze the data.

The univariate analysis for individual variables in datasets(vacdetails.csv) are as follows

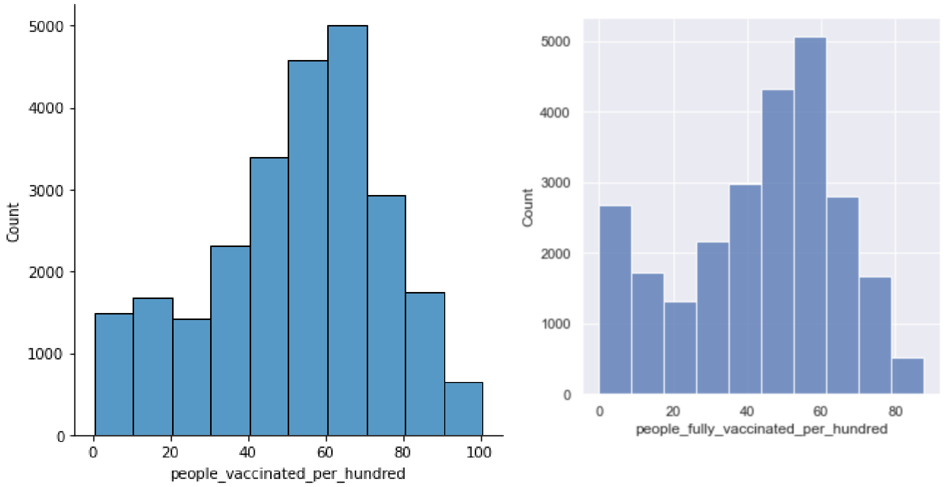


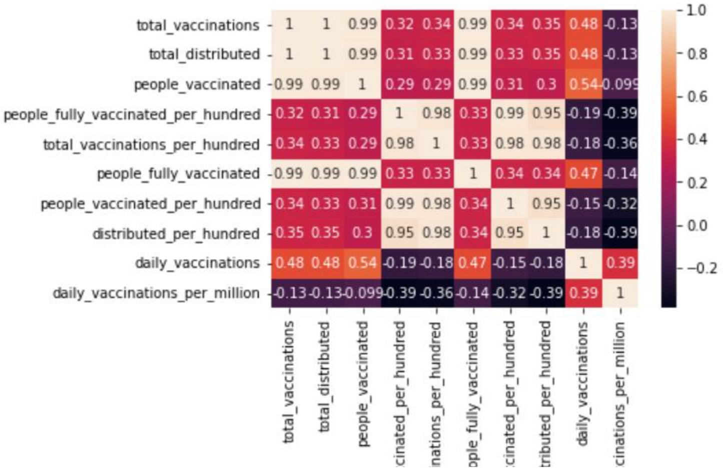
Fig.Histogram plot for variable people fully vaccinated per hundred

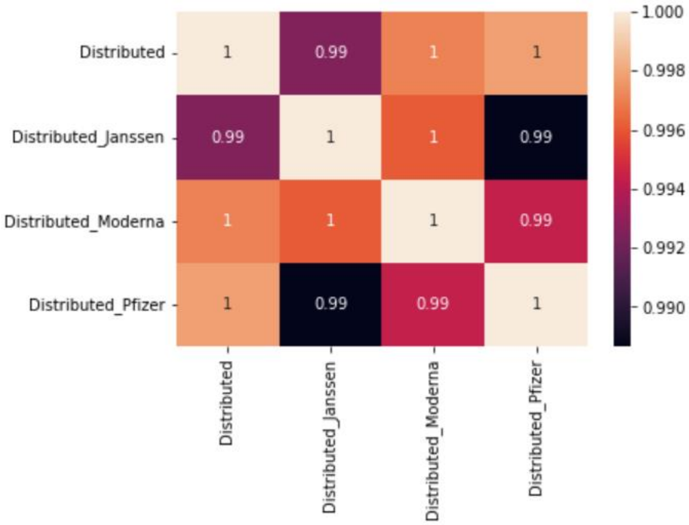
Histograms give the frequency of occurrence on the Y-axis. From above univariate distribution we can see the most people vaccinated fully is around 40-60 people for every hundred according to dataset.
And similarly, people vaccinated (at least one dose) is between 60-80 per hundred. We can also determine details like symmetry, skewness and kurtosis of the variable.

Bivariate Data Analysis:

Bivariate analysis is analysis of two variables in a dataset. Bivariate analysis can be used to determine the relationship between two variables, comparing two variables. Some of the examples for bivariate data analysis are scatter plots, correlation plots etc.

Correlation plots are also called as heat maps. **Heat maps** are used to determine the correlation among the variables in a dataset. The correlation values help to determine the dependency of variables on each other.





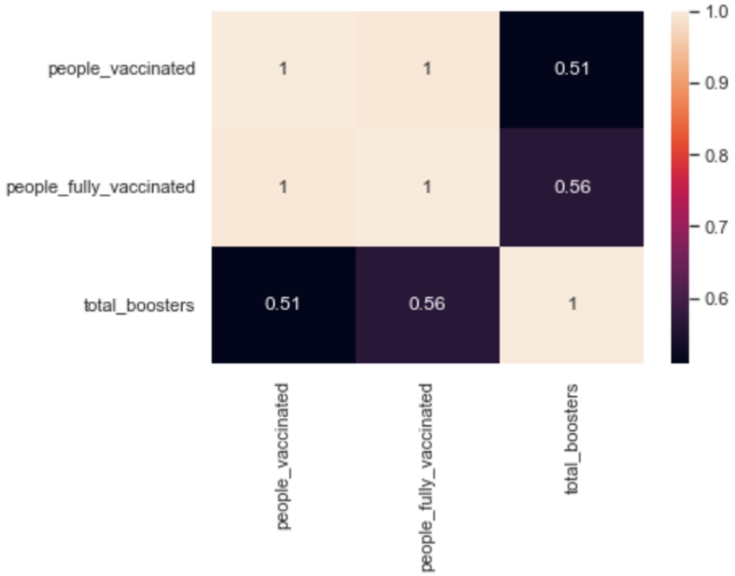


Fig Heatmaps of datasets.

Pair plots of seaborn library is another way of finding relation between variables in a dataset. Pair plot gives the plots between all the variables in the dataset both categorical and interval variables of the dataset.
Pair plots can also be drawn for the determined number of variables in dataset as shown below.

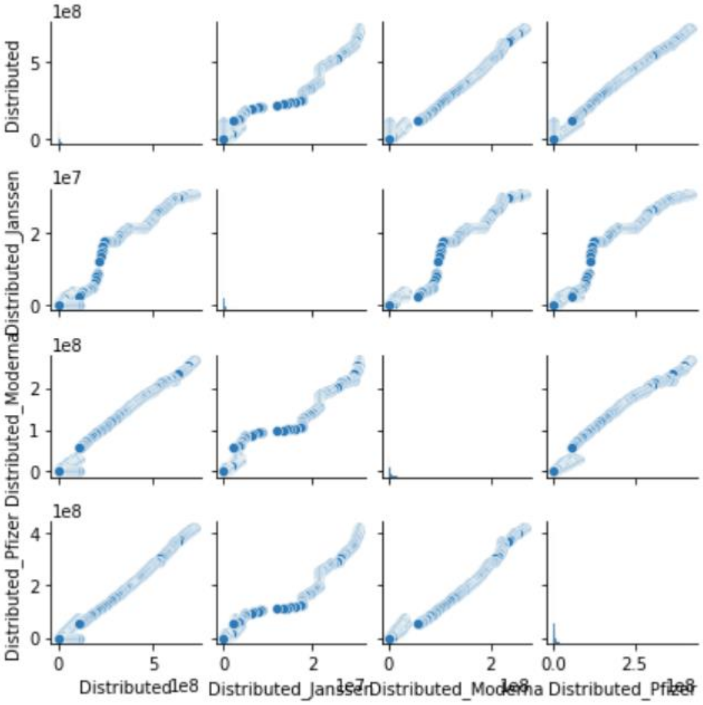


Fig. pair plot of vaccine dataset.

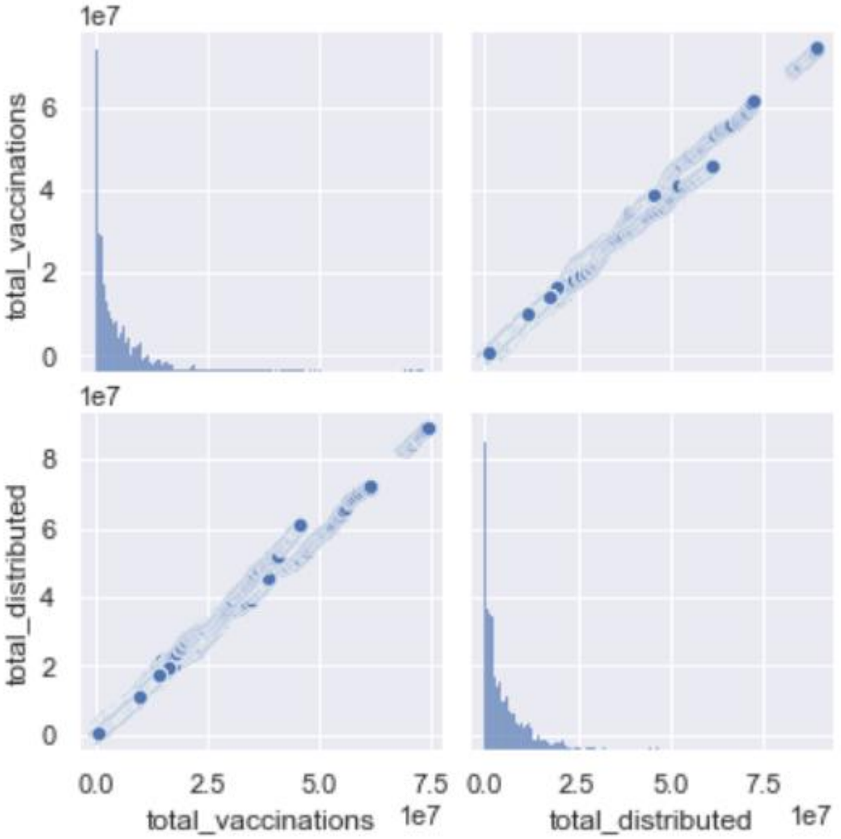


Fig. pair plot for determined variables of dataset.

Multivariate Analysis

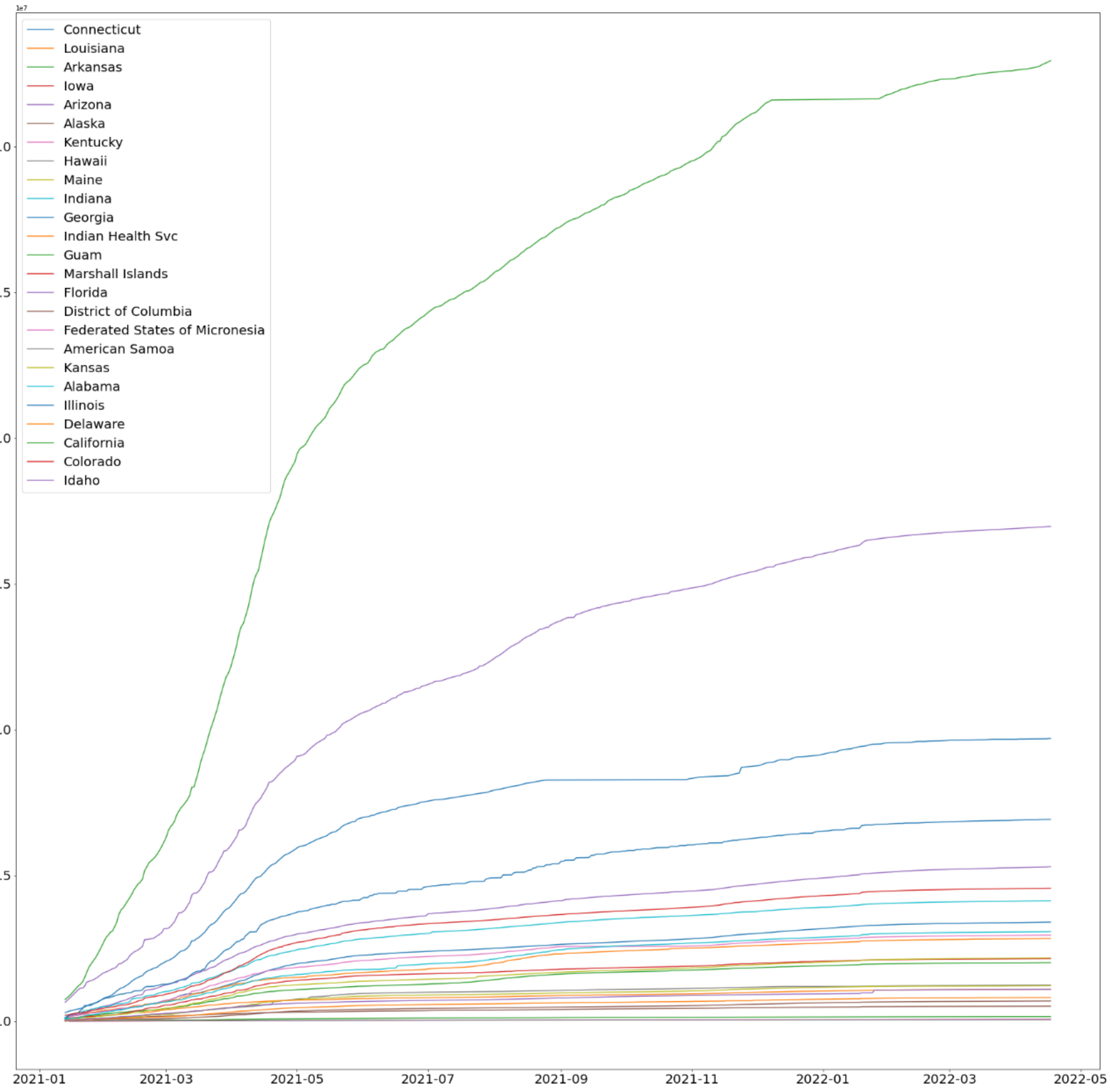
Multivariate analysis is the analysis of more than two variables at a time. As the variables or dimensions of the data increases, it becomes difficult to analyze the data. Multivariate data analysis provides an even complex and deep understanding of the data variables, their relationship with each other etc. Some of the complex examples of Multivariate analysis are cluster analysis, Principal component analysis. According to Bertin's theorem, only 3 dimensions can be effectively percieved by human brain at a time.

Multivariate data visualization is also used to compare the differences between variables and observe and analyze the spatial patterns of the variables. There are several different approaches for multivariate data visualization. The approach depends on the type of analysis we need. There are numerous applications for multivariate visualizations. Graphic excellency is mostly about the multi variate analysis. Scatter plot matrix, parallel coordinates plots are some more examples of multivariate data visualization.

Results:

Vaccination Trends in Different States Across the United States-Multiple Line Graph

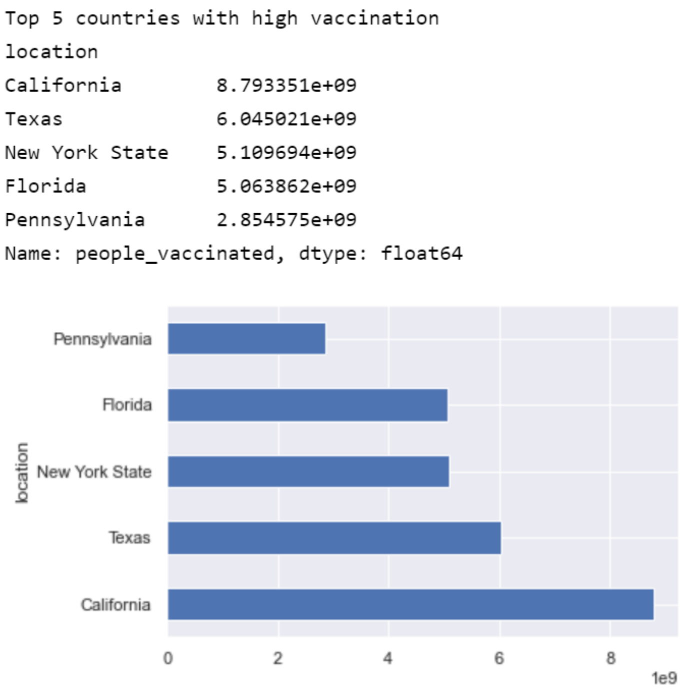
The best graph to explain the overtime changes or trends of any variables. Multiple line graph is used to visualize trends of the vaccination in United states. The dataset contains records from 2021 January to 2022 April. Line graphs are best for analyzing the trends or overtime changes in any variable.



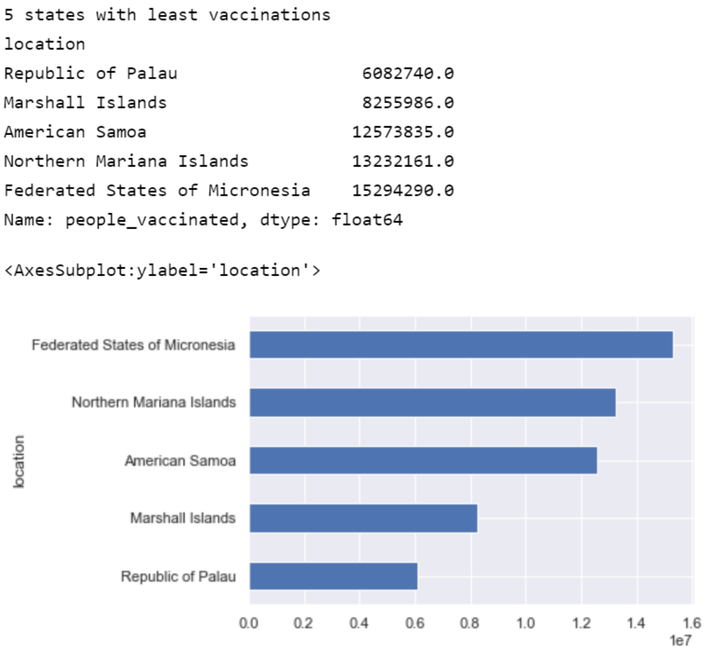
The line graph above has X-axis coordinates as time (year-month format) and y- axis coordinates as people vaccinated (in Billions). From the above graph we can assume that almost most of the states has an exponential increase in the vaccinations from January 2021 to present. States like Arkansas, Arizona, Illinois, Alabama are some of the states with significant exponential increase in the vaccinations/vaccinated people from mid-2021.

Top 5 States with The Highest Vaccination and Least Vaccination-Horizontal Bar Charts

Bar chart is best graph to represent comparisons and show differences between the quantities. Horizontal bar graph contains horizontal bars. Horizontal bar charts are used to answer this part of the research question.



```
Top 5 countries with high vaccination
location
California        8.793351e+09
Texas             6.045021e+09
New York State    5.109694e+09
Florida           5.063862e+09
Pennsylvania      2.854575e+09
Name: people_vaccinated, dtype: float64
```



```
5 states with least vaccinations
location
Republic of Palau                  6082740.0
Marshall Islands                   8255986.0
American Samoa                    12573835.0
Northern Mariana Islands          13232161.0
Federated States of Micronesia    15294290.0
Name: people_vaccinated, dtype: float64

<AxesSubplot:ylabel='location'>
```

Fig. Bar charts

One axis of the horizontal bar contains a categorical variable and other axis is a numerical variable. The length of the bars/bins goes from left to right. Here two horizontal bars are used to represent the top 5 states with highest vaccination and top 5 states with least vaccination. The length of the bar gives the corresponding value to the categorical value.

From the above fig, we can see the top 5 states with highest vaccination record so far is California, Texas, New York, Florida and Pennsylvania. And the states /territories that has the least vaccination rate are American Samoa, Northern Mariana Islands, Federated states of Micronesia, Marshall Islands and Republic of Palau. From above we can say that most of the populated states have the highest vaccination percentage. And Islands and territories with less population are the least vaccinated places in the United States.

Percentage Of People Who Are Vaccinated In The Total Vaccines Distributed In Different States- Dashboard

According to the graphic excellency principles, it is most important to convey greatest number of ideas within a short time and limited space. Dashboard is one of the best practices that follows the graphic excellency principle which conveys different types of information using a limited space and its interactive nature helps in engaging the user and thus provides efficiency.

PowerBI dashboards are a single page which uses multiple charts in a single page. Dashboards are the most effective visualization tools. Dashboards centralizes all the important information of the data in a single page and tell the story about the dataset/information. To analyze this part of the research questions, Microsoft PowerBI dashboard is used. In the dashboard, three types of graphs are utilized which are stacked bar charts, Line graphs and area chart.

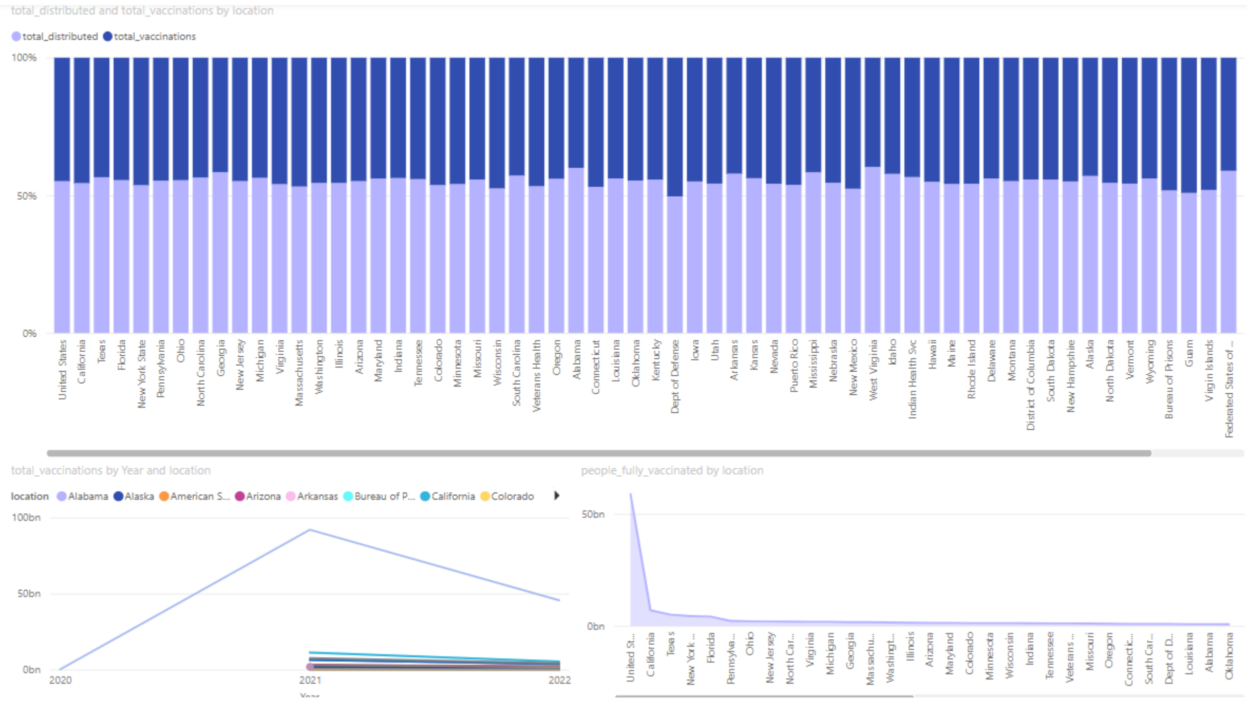


Fig Dashboard

Dashboards ae interactive in nature and allows the viewer to hover over the chart or graph and click on the interested data t see the changes in the corresponding variable from other charts. Dashboards play a crucial role in analyzing the business data and make important decisions.
The above dashboard provides different information like vaccine distributed to people vaccinates and total vaccinations trends in time and location wise details about people fully vaccinated. All these are together presented in the dashboard to effectively communicate the information.

World-Wide COVID Distribution Vaccine Distribution - Choropleth Map

Choropleth maps are the statistical thematic maps that uses intensity of colour to correspond to geographical characteristics. Choropleth maps help in analyzing distribution of a variable around a geographical area. The changes in the intensity of color helps analyzing the variations in the variable. It is also possible to represent more than one variable (two or three variables) at a time in the choropleth maps. The difference between the variables can be distinguished by allocating color fir each variable.

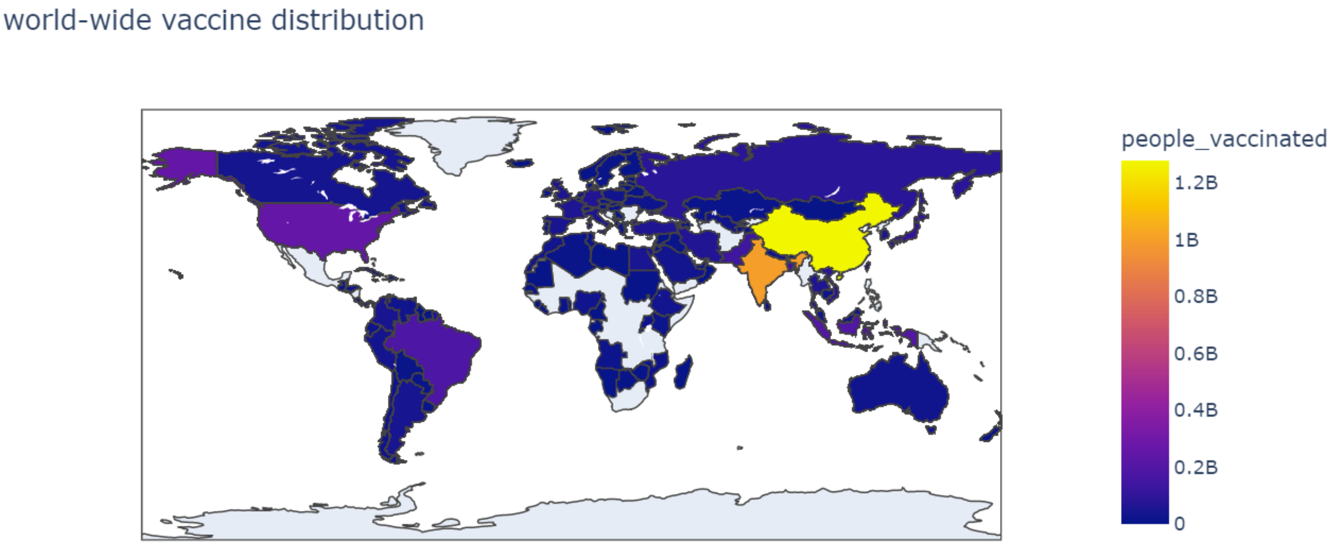


Fig. Choropleth Map

Choropleth Maps is used to study the worldwide vaccination of Covid 19. COVID 19 hit the world at the end of 2019 and different vaccines were developed out within less than a year and most of the countries started administering the vaccines.

The dataset with data about countries and people vaccinates is collected and performed EDA and then plotted using choropleth maps. Choropleth maps provide the geographical coordinates of any part of the world. A Json file called Countries.geojson is mapped to our collected dataset and then used to plot the people vaccinated across the world. The color is used to represent the number range of people vaccinated in that part of the world. Choropleth maps are interactive in nature and when hovered over the geoplot we can see the location details and as well as the number of people vaccinated in that region.

The color that is not available in the legend and that’s on the map is the data that is not present in the collected dataset. We can see most of the countries vaccination rate ranges below 0.4 B.

Vaccine More Distributed or Administered (Moderna/Pfizer/Johnson-Johnson) In The US-Clustered Column Chart and Pie chart

With roll out of different types of vaccines, there are mixed opinions on what vaccine is more effective, side effects of vaccines, hesitancy. There are mainly three types of vaccines administered in United states. Moderna, Pfizer and Johnson-Johnson are the three vaccines that are mostly distributed in different states of America.

For this project, we try to find out what vaccine among the three vaccines is highly administered and used across different states. To answer this question, we collected a dataset with different states and data about the three vaccine distributions.

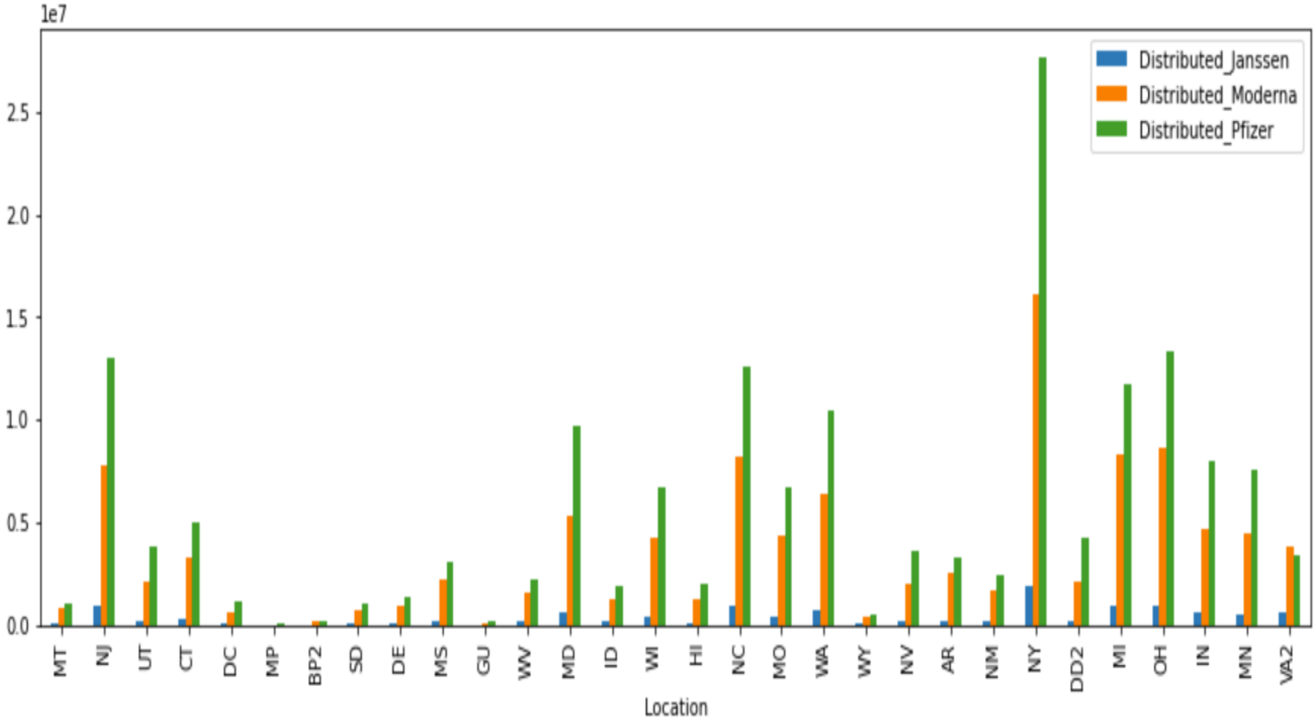


Fig. Column Chart

The graph used to visualize this data is column chart. In this chart the distribution of three types of vaccines for a given state are plotted using the bars. The height of the bar represents the vaccinated number of the vaccine. This chart makes it easy to compare the three types of vaccines and find out the most used vaccine type. Different colors are used for different types of vaccines to avoid confusion and provide clarity to the viewers. As the legend indicates green color bar is used for Pfizer, Orange for Moderna, blue for Johnson-Johnson (named as Janssen). X-axis indicates the code of the state and Y-axis indicates the vaccine distributed.

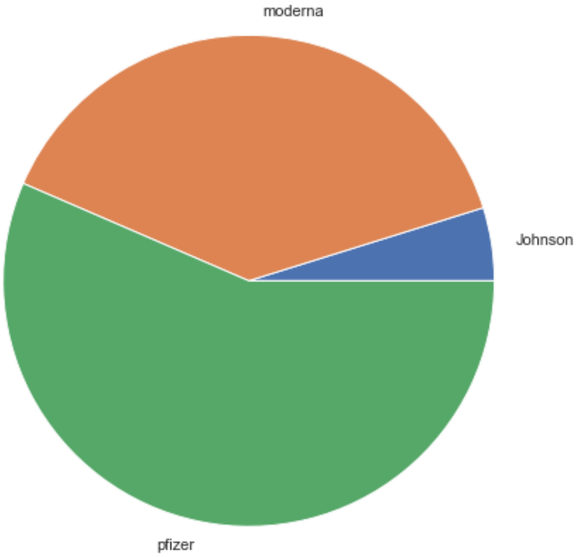


Fig Pie Chart

From the column graph, we can see that for most of the cities the green color bar is dominant compared to the orange and blue bars. So, we can conclude that Pfizer is highly distributed than Moderna and Johnson-Johnson.

Pie Chart is best visual representation when comparing parts and whole of any variable. In this project different types of vaccines are compared to find out which among the three vaccines is highly distributed in United States. From the above pie graph created with the types of vaccines data, we can conclude that Pfizer vaccine is distributes more than the other two vaccines (Moderna and Johnson). From the Pie chart we can see portion for Pfizer vaccines takes up almost half of the total vaccinations. Different colors for different types or labels of data is used to easily distinguish the types.

From above clustered column chart and the pie chart we can say that Pfizer is the most distributed vaccine across the United States and as well as in Most of the cities in the United States.

Conclusion & Future work:

This research helps the readers in understanding the vaccination trends across different states of the United States and around the world, vaccines distributed, and people vaccinated, types of vaccines distributed.
This study might help all the hesitant people to clear their doubts about vaccinations and effects of vaccinations and how effectively did the vaccination help in fighting with the virus. This research can also be helpful in administering for any future vaccination programs.

References

Chen, X., Huang, H., Ju, J. *et al.* Impact of vaccination on the COVID-19 pandemic in U.S. states. *Sci Rep* **12,** 1554 (2022). https://doi.org/10.1038/s41598-022-05498-z

Dey, S. K., Rahman, M. M., Siddiqi, U. R., &amp; Howlader, A. (2020). Analyzing the epidemiological outbreak of Covid-19: A visual exploratory data analysis approach. Journal of Medical Virology, 92(6), 632–638. https://doi.org/10.1002/jmv.25743

Eshun-Wilson I, Mody A, Tram KH, Bradley C, Sheve A, Fox B, et al. (2021) Preferences for COVID-19 vaccine distribution strategies in the US: A discrete choice survey. PLoS ONE 16(8): e0256394. https://doi.org/10.1371/journal.pone.0256394

Nazir, A., Ulusoy, S., &amp; Lotfi, L. (2021). Visual exploratory data analysis of COVID-19 pandemic: One year after the outbreak. https://doi.org/10.1101/2021.05.04.21256635

L. Harper, N. Kalfa, G.M.A. Beckers, M. Kaefer, A.J. Nieuwhof-Leppink, Magdalena Fossum, K.W. Herbst, D. Bagli, The impact of COVID-19 on research,2020. https://www.sciencedirect.com/science/article/pii/S1477513120304125

Mathieu, E., Ritchie, H., Ortiz-Ospina, E. *et al.* A global database of COVID-19 vaccinations. *Nat Hum Behav* **5,** 947–953 (2021). https://doi.org/10.1038/s41562-021-01122-8

Mohamed, K., Rzymski, P., Islam, M. S., Makuku, R., Mushtaq, A., Khan, A., Ivanovska, M., Makka, S. A., Hashem, F., Marquez, L., Cseprekal, O., Filgueiras, I. S., Fonseca, D. L., Mickael, E., Ling, I., Arero, A. G., Cuschieri, S., Minakova, K., Rodríguez-Román, E., … Rezaei, N. (2021). Covid-19 vaccinations: The unknowns, challenges, and hopes. Journal of Medical Virology, 94(4), 1336–1349. https://doi.org/10.1002/jmv.27487

Wikimedia Foundation. (2022, February 12). Template: COVID-19 pandemic data. Wikipedia. Retrieved April 14, 2022, from https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data