**Group 2: MIS 6380.002**

**Covid Vaccinations in The United States**

**Akshay Kamoji**

**Arun Vasa**

**Neha Vijayananda**

**Niharika Gogate**

**Shreya Patel**

**Viveshnaraj Arulrajah**

**Executive Summary**

Since the outbreak of Covid-19 in the United States in late December 2019, the country has been working diligently to create a vaccine and distribute it to the public. To explore the insights of vaccinations, distributions, and booster status, we investigated a primary dataset that included data on vaccinations and distributions by date and by state. We also used two secondary datasets to understand demographic data regarding vaccinations and distributions and political affiliation. Through our analysis, we learned the number of vaccinations distributed by state, the percentage of shots distributed vs. vaccinations fully taken, the role of the population on vaccinations, the growth in boosters, and the number of full vaccines by political party and vaccine hesitancy. Date, state, total vaccinations, boosters, and total dispersed vaccinations are all variables in our dataset. Our primary goal was to visualize and prove if there was an increase or decrease in total vaccinations distributed vs. vaccinations taken between January and December of 2021 and expand to include location, population, boosters, political affiliation, and hesitancy.

Our five main hypotheses are expressed in detail in our introduction. Our hypotheses included whether there was a correlation between location and vaccination status, an increase in the number of booster shots after September 2021, vaccination among democrats and republican states, percentage of vaccinations by state based on population size, distribution, and hesitancy of the vaccines month over month across all states. Tableau was used to visualize the dataset and R and excel to clean the data. Each hypothesis is visualized through Tableau to find insights from the data.

**Index**

[Data Description](#_c0qzcf6ze749) **4**

[Data Cleaning](#_2p0q6rcrxvkv) **5**

[General Introduction](#_n77f0cnw90bh) **6**

[Insights and Findings](#_w4o7bqfxwehd) **7**

[Conclusion](#_dlc88215bvme) **10**

**Data Description**

The Primary Dataset has information regarding covid 19 vaccinations in the United States from January 2021. We approached this data to prove five hypotheses using the tool Tableau. We mostly used the Max of each state value because the data is cumulative, and maximum gives the cumulative number at the end of each month. The columns like the number of vaccinations and percentage are cumulative, increasing with successive additions. The value on Dec 31st gives all the accumulated data for the year 2021 in each column. The various columns in the dataset provide the following information:

1. Date: The data were collected between Jan 12th to Dec 31st of 2021
2. State: The data were collected for all the 50 states in the USA
3. Total vaccinations: Cumulative number of vaccines, including first and second doses.
4. Total vaccines distributed: Cumulative numbers of vaccines distributed.
5. People vaccinated: Cumulative number of first dose vaccinations.
6. People fully vaccinated per hundred: Cumulative percentage of people who are fully vaccinated i.e., first and second dose.
7. Total vaccinations per hundred: Cumulative percentage of people vaccinated including both partially vaccinated people and fully vaccinated people.
8. People fully vaccinated: Cumulative number of vaccinations of people who took the first and second dose.
9. People vaccinated per hundred: Cumulative percentage of people who are only partially (only first dose) vaccinated.
10. Distributed per hundred: Cumulative percentage of distributions of vaccinations.
11. Daily vaccinations: Number of vaccinations daily, including both doses.
12. Share doses used: Total vaccinations divided by total vaccinations distributed.
13. Booster shots: The cumulative number of booster shots that were taken starting from September 21, 2021.
14. Total boosters per hundred: Cumulative percentage of boosters.

> Secondary dataset - US Presidential election results for Biden and Trump

> Secondary dataset - Population of states in the USA. This dataset provides information on the US population by region, state, and month in 2021

**Data Cleaning**

The data was cleaned using R and Excel. The missing rows in the dataset were imputed using the time series imputation package in R since all data is a function of time. The states were also mapped to their respective political ruling parties (i.e., republican and democratic). For the booster columns, all missing rows after September 21, 2021, were imputed since booster shots started in the month of September. The original dataset also has 2020 and 2022 data, the rows of which were removed since the analysis was only for the year 2021.

**General Introduction**

As mentioned in our Executive Summary, Covid-19 hit the United States in late December 2019. We have seen the country work hard and fast to develop a vaccine. The first available vaccine from Pfizer and Moderna was available in late 2020. Looking at the history of Covid in the United States and the actions of our political leaders, we can see that 2021 was an important year not only for vaccine distribution but also for US citizens getting fully vaccinated. In searching for Covid-related data, we discovered a dataset that allowed us to explore US vaccinations, specifically the rate of distribution and the rate of vaccinations taken by day, month, year, and state. Our primary dataset includes vaccine information filtered by date and by state. Our objective was to understand the correlation between location (state), date, and vaccination status vs. vaccination distribution. We included secondary datasets to dive deeper into the data and understand all variables around US vaccinations. Our first secondary dataset included census information to understand if population by state played a role in the volume of vaccinations. Our second secondary dataset included information on the 2020 election. Together with our primary dataset and our secondary datasets, we wanted to explore the following hypotheses:

First, we wanted to answer whether there was an increase in the distribution of vaccines month over month across all states using our primary dataset.

Second, we wanted to answer whether there was a correlation between regional location and vaccination status. To understand and analyze this hypothesis, we explored vaccination data per hundred by state and by region from January 2021 to December 2021.

Third, we wanted to answer whether there was a greater percentage of vaccinations fully taken by state based on the overall population size for each state from January 2021 to December 2021. We used vaccination data by state and a secondary dataset with information on the US census to analyze this hypothesis.

Fourth, we wanted to answer whether there was an increase in booster shots after September 2021. We used data on boosters from our primary dataset to analyze this hypothesis.

Fifth, we wanted to answer whether there was a correlation between the number of full vaccinations taken and political parties from January 2021 to December 2021. We used a secondary dataset with information from the most recent election.

To supplement hypothesis five, we also looked into vaccine hesitancy based on political parties. Together, we analyzed and answered each hypothesis through our datasets and visualizations and presented our findings below.

**Insights and Findings**

1. Our first hypothesis was to determine if there would be an increase in the distribution of vaccinations month over month across all the states in the US throughout the year 2021. From the primary dataset, Total Distributions and Distributions per Hundred individuals were used to test the hypothesis. The distributions were analyzed using parameters like states, regions, months, and years. When data about distributions were plotted, it can be seen by the changes in color that as the months go by, the distributions are increasing significantly in California and Texas. California has the highest vaccine distribution at the end of the year, followed by Texas, Florida, and New York. The same analysis can be inferred from the bubble chart, which shows distributions in the six regions for two years-2021 and 2022. The above states had the highest distributions in the West Pacific, South, South (Atlantic), and Northeast for both years. Other regions like Midwest (North Central), Illinois, Ohio, and Michigan are among the top few states, and in the West (Mountain) region, Arizona is the highest, followed by Colorado.

The line graph of Distributions per Hundred against Months depicts that the distributions spiked dramatically in the initial months of 2021. This rise suggests that there was quite a need for vaccines since people were suffering in 2020, so using vaccines was the only respite they had. After the month of May, there is a slight and steady increase. This can be attributed to the fact that most of the population was already vaccinated. However, the distributions have escalated after September 2021 and have been on a constant rise since then. Overall, we can say that the distribution is increasing across all the states month by month.

1. Our second hypothesis involved determining whether there was a correlation between location and vaccination status. The number of fully vaccinated individuals per hundred in the year 2021 was considered when analyzing the data and creating visualizations. It is important to note that the “fully vaccinated” status consists of two vaccine doses. The findings from the visualizations created are both by state and by region. When looking at the data for fully vaccinated individuals by state and by month, as the months go by, it is shown that this number gets larger and larger for each state. However, we can see those states along the coasts - such as California, Washington, Maine, and New York - exhibit a greater ratio with values up to 77.46 people fully vaccinated per hundred by December 2021. The contrast is evident as opposed to the mid-region states, such as Wyoming or Missouri, with numbers around 40 and 50 as the number of people fully vaccinated per hundred by December 2021. Additionally, the data was analyzed by grouping states into regions. The regions were split up as the Northeast, West Pacific, Midwest, South Atlantic, West, and South. With this divide, it is evident that the average number of individuals fully vaccinated is highest in the Northeast (48.86 people per hundred) and West Pacific (44.58 people per hundred) while other regions like the Midwest, South Atlantic, West, and South had significantly lower averages closer to 35 to 40 people. Again, this shows that the greatest amounts of people fully vaccinated are along the East and West coasts.
2. Our third hypothesis was to determine if there was a relation between the number of vaccinations and the population of each state across the United States. Using our primary dataset and our secondary dataset on the US census, we were able to answer that the greater the population of the state, the greater the number of vaccinations. In our story, we compared fully vaccinated people to the population of each state by the end of 2021. California has the highest population in the US and the highest number of fully vaccinated people, followed by Texas, Florida, New York, etc. By also analyzing month-wise data across each state, California also had the highest number of fully vaccinated people by the end of December 2021. By the end of the year 2021, California had 64 million total vaccinations; this includes the first dose, second dose, and boosters. In the treemap, we included total vaccinations, population, and fully vaccinated status in the text labels so that we can easily correlate between those three values by state. Based on the comparisons of metrics in the charts, we fail to reject the hypothesis.

In addition, to supplement Hypothesis 3, we investigated the vaccination rates by month.

The vaccination rates for various states and time periods can also be obtained by the following formula: Vaccination Rates = amount of vaccinations/ number of vaccinations distributed.

The higher the number, the higher the rate of vaccinations for that state and a designated time.

Using animation and the vaccination rates formula, the vaccination rates in the month of January were the highest (a sudden spike can be seen in the vaccination rates in the month of January). After January, the vaccination rates pretty much remain the same though the rest of the year. This was because covid vaccinations in the US began in the month of December 2020. Since many people were eager to get vaccinated, the vaccination rate in the month of Jan was the highest.

1. Our fourth hypothesis was to confirm if there was an increase in booster shots after September 2021. We visualized the number of boosters administered each month and by state. Boosters are taken after an individual is fully vaccinated and is the main reason why there was no data from January 2021 until September 2021, when booster shots distribution started. In October 2021, New Mexico had the highest number of boosters shots taken, and by the end of December, the states of California and Texas became first and second-highest booster shots taken. By the beginning of 2022, New York replaced Texas to be the second-highest state to have taken boosters. When it comes to Boosters per hundred, New Mexico had the highest in October 2021, and later California took over. Hence from the visualization, there is an increasing trend of booster shots from October 2021.
2. Our fifth hypothesis was to determine whether there was a correlation between the number of full vaccinations taken and political parties in the United States (Democratic or Republican). Using our primary dataset and our secondary dataset on the 2020 election, we were able to answer whether there was a higher number of vaccinations between democratic or republican states. We used a secondary dataset that showed the traditional political lines for Democratic and Republican states. For example, California is a Democratic state, shown in blue in our story. Texas is traditionally a Republican state, shown as red in our story. We can see the yearly vaccination growth from 1/2021 to 12/2021. Overall, 201M Americans took both the first and second vaccine. California had the most vaccines taken for a democratic state, and Texas had the most vaccines taken for a republican state. Through our analysis and visualizations, traditional democratic states had a higher number of full vaccinations, meaning citizens took both the first and the second vaccine more than citizens from republican states. There was a gap of about 45M vaccinations between democratic and republican states. Our analysis shows that based on the traditional political parties in the United States, citizens of democratic states had a higher rate of vaccinations fully taken, which aligns with the public perception of Covid from the perspectives of democrats and republicans.

To supplement hypothesis 5, we looked into vaccine hesitancy by states and political parties. Using the data regarding the number of vaccinations and the number of vaccines distributed, a measure of vaccine hesitancy was calculated by the ratio of the number of vaccinations to the number of vaccines distributed.

Degree of hesitancy = 1- [number of vaccinations/ number of vaccinations distributed].

The higher the value of the Degree of hesitancy, the higher the hesitancy and vice versa.

The following insights were obtained:

The top 10 most and the least hesitant states were found using the degree of hesitancy formula. Alabama was found to be the most hesitant state with a degree of hesitancy measure of 0.3, while New Mexico was found to be the least hesitant state with a degree of hesitancy measure of 0.09.

Amongst the 10 least hesitant states, it was found out that all 10 were democratic states whereas, amongst the 10 most hesitant states, it was found out that 8 of them are republican states while the other 2 are democratic states. This follows up with hypothesis five, where it was found that democratic states had a higher volume of full vaccinations than the republican states.

The states were also divided by various zones such as North, East, West, and South. An average degree of hesitancy measure was taken for those states; it was found that southern states, on average, are the most hesitant, while the northern states are the least hesitant when it comes to taking covid vaccines.

**Conclusion**

To conclude, the following hypotheses were proven correct:

1. **There is an increase in the distribution of vaccines month over month in all the states.** Moreover, California and Texas have been consistently at the top concerning distributions. Hence, we fail to reject the hypothesis.
2. **There is a relationship between location and vaccination status.** Findings based on the number of individuals fully vaccinated per hundred splits by both state and region showed that areas on the East and West coasts had a greater amount of vaccinations than the other areas. Therefore, we fail to reject the hypothesis.
3. **There is a relationship between the number of vaccinations and state population.** It was found that a greater state population also showed a greater number of vaccinations, with states like California, Texas, and Florida leading in both factors. Therefore, we fail to reject the hypothesis.
4. **There is an increase in booster shots after September 2021.** There was an increasing trend of booster shots from October 2021. Based on our analysis, we fail to reject the hypothesis.
5. **There is a correlation between full vaccinations and political affiliation.** Citizens of Democratic states had a higher number of fully vaccinated individuals. Therefore, we fail to reject the hypothesis.

Supplementary data: **Vaccine hesitancy**

All top 10 least hesitant states were Democratic, while 8 out of 10 of the most hesitant states were Republican. The southern states were most hesitant out of the other areas.