**MIS 6380**

**DATA VISUALIZATION**

**GROUP - 4**

**Names of Members:**

Sushmitha Madiraju

Vaishnavi Arni Lakshmana Rao

Neha Raghukula

Shubham Shekhar

**Index**

Executive Summary 1

Data Description 2

Data Cleaning 3

General Introduction 4

Insights and Findings 5

Conclusion 8

**Executive Summary**

With Covid-19, the world is being confronted with multiple pandemics. Covid-19 has provoked a poly pandemic that is undermining development progress, exacerbating state fragility, and potentially further eroding international cooperation. The Covid-19 pandemic is a global and multi-faceted crisis. No region of the world has been spared its dire toll on lives and well-being. And no part of the globe has remained unaffected by the socio-economic shock produced by Covid-19 and its grim effects on human livelihoods. Face masks, including that of surgical masks and cotton masks, were used to prevent the transmission of SARS-CoV-2 during the COVID-19 epidemic. Their use in clinical and community settings is aimed for both source control to restrict viral transmission and personal protection to avoid infection. Similarly, COVID-19 vaccination requirements have been imposed by a number of states, municipalities, and commercial businesses in the United States. President Joe Biden said in September 2021 that the federal government will take steps to require COVID-19 immunization for some organizations under the authority of the federal government or federal agencies.

In order to acquire understanding of how each county of Texas state was affected by covid 19, we have used a primary dataset that consists of information on rate of vaccination for each county with fields like By County, By County-Age, By Age-Gender-Race, By County-Race and By Vaccination Date. We haven’t used any secondary data set to complement our visualization.

We have added five main hypotheses that are expressed in detail in our introduction. Our hypotheses include testing if black ethnicity has received the least number of booster vaccinations, if people that have been vaccinated with one dose have also been vaccinated with a second dose and a booster dose and so forth. Excel and Python were used for data cleaning and Tableau was used to visualize the data. Tableau is used to depict each hypothesis in order to get insights from the data.

**Data Description**

The Primary Dataset has information regarding confirmed COVID19 cases county-wide for the state of Texas. We approached this data to prove five hypotheses using the tool Tableau.

> The primary dataset County-wise vaccination rate. The sheets are:

1. About The Data - explains every column and every unique heading in the data to understand the data better
2. By county - 15 columns that segregates data as per doses administered, health condition, age, health care worker, resident, and other uniqueness
3. By Age, Gender, Race - explains data by gender, age group, race/ethnicity, and doses administered
4. By Age, Day - number of doses administered per day in all counties by age group
5. By County, Age - Vaccines administered at county, age, and doses number level
6. By County, Race - Break-up of data by race/ethnicity, county, and dose number
7. By Vaccination Date - Break-up of data by vaccination date and doses administered

**Data Cleaning**

The data was cleaned using Excel and Python. Pandas’ library in python was mainly used for cleaning and Excel’s built-in function was used to remove duplicates. Our original primary dataset consists of information on COVID-19 vaccine data by county. We’ve preprocessed the data and used only the information that was needed for our visualizations and our hypothesis.

For the booster vaccination doses records after July 2021 were imputed since the booster shots have started in the month of August 2021. The COVID 19 vaccine data by County dataset records have the count of the people who received booster doses and immunocompromised people who received additional doses.

Median value has been substituted wherever appropriate when the values were missing as this would not affect the distribution of the data.

**General Introduction**

Covid-19 struck the US in late December 2019, as stated in our Executive Summary. We have witnessed the nation work diligently and quickly to produce a vaccine. At the same time the usage of masks like N95 were also made mandatory for everybody. We can see from Covid's history in the US and the acts of our government leaders that 2021 was a significant year for both vaccine distribution and full vaccination of US people. We have found a dataset related to covid 19 vaccinations that helped us inspect the number of vaccines administered filtered by date, race, age, profession, etc. for each county of the state of Texas.

We used primary datasets to give information about the vaccination rate. Our primary dataset consists of information related to vaccination rate for each county segregated by age, race, county, race, and vaccination date which was used to check if vaccination played an important role in decrement of covid 19 cases. In conjunction with our primary dataset, we tried to explore hypothesis listed below.

1. We sought to check whether People of Black Ethnicity have received the least number of booster shots.
2. We sought to check whether there are a few common counties where we see least vaccinated areas by ethnicity.
3. We sought to check whether there is an increasing trend in the amount of booster shots taken by individuals from Aug 2021 to Jun 2022
4. We sought to check if overall in all counties, we are seeing a high booster shot intake among youth and mid-age group
5. We sought to check if everyone that has taken a single vaccine dose has also taken the second vaccine dose and the booster dose.

**Insights and Findings**

**1)Null Hypothesis:** People of Black Ethnicity have received the least number of booster shots **Alternate Hypothesis:** People of Black Ethnicity have not received the least number of booster shots.

The visualization has considered 4 major ethnic groups: Asian, Hispanic, Black, and White. With the visualization, we can reject the null hypothesis as we see that the Asian ethnicity has received the least number of booster shots– 427,753. Before we make any further inferences about this visualization about the ethnicities here are a few things to consider:

1. Asian communities are the least populous group in the counties that we have considered
2. We need to look at the percentage of people that belong to a particular ethnicity to get a better idea about the data
3. There are 2 other groups considered which are “others” and “unknown”. These total up to 1.7 million people. This makes 24% of the data. “others” and “unknown” can include people who are biracial, mixed, etc.

**2)Null Hypothesis:** There are a few common counties when we see least vaccinated areas by ethnicity.

**Alternative Hypothesis:** There are no common counties when we see least vaccinated areas by ethnicity.

When we divide the data by county and by race, we see that the counties Austin, Angelina, Archer, and Bailey have the least full vaccination rates for most ethnicities. Using this information, we can research why these counties have lower vaccination rates. Could socio-economic status, standard of living, vaccine availability, religion, and culture be the reason behind low vaccination rates? This demographic data can give us a good base for further research.

Here, we fail to reject the hypothesis.

**3) Null Hypothesis**: There is an increasing trend in the amount of booster shots taken by individuals from Aug 2021 to Jun 2022

**Alternate Hypothesis:** There is no increasing trend in the amount of booster shots taken by individuals from Aug 2021 to Jun 2022

With the visualization we can see that when the vaccine boosters were rolled out in Aug. 2021, there was a month over month increase in the boosters administered. However, we see a change in trend in 2022. Since Jan 2022, there has been a decline in the number of boosters administered every month. Therefore, we reject the Null Hypothesis.

We can also see that from August 2021, there is a declining trend in the full vaccine doses administered as well. Using this information, researchers can investigate the reason for this sudden decline which could be:

a) Decline in the number of COVID cases may have left people confident that they do not require the booster

b) People believe that covid boosters may not be effective in reducing the severity of the disease

c) most people have already been vaccinated

d) Vaccine hesitancy among the people could have left them not wanting to complete their vaccination

**4)** **Null Hypothesis:** Overall in all counties, we are seeing a high booster shot intake among youth and mid-age group

**Alternate Hypothesis**: Overall in all counties, we are not seeing a high booster shot intake among youth and mid-age group

As per the visualization 39% of the booster doses that have been administered in all the counties have been to the age group 16-49 which is the age group that we are testing in our visualization. The age group 50-64 is seeing the next highest number of booster dose intake. Among this we see that the counties Harris, Travis, Dallas, and Bexar have taken the highest boosters. This information can be further used to study why we’re seeing a lower vaccination rate among other age groups as well as in other counties. We fail to reject the Null Hypothesis.

**5)** **Null Hypothesis:** Everyone that has taken a single vaccine dose has also taken the second vaccine dose and the booster dose.

**Alternate Hypothesis:** Everyone that has taken a single vaccine dose has not followed up with the second dose and the booster dose.

We have taken data over 3 years, 2020, 2021, and 2022 to test our hypothesis. We see that there is a downward trend after a single vaccine dose is administered. The total number of first dose vaccines administered is 42.2 million, fully vaccinated is 35.5 million, and the total number of booster doses administered is 14 million. Based on this information we will reject the null hypothesis. These insights can be used to research why people have not followed up with the second and booster doses. Some of this can be:

1. Lack of availability of vaccines
2. Lack of trust in vaccines
3. Post covid 60-day immunity
4. Decline in the number of covid cases

**Conclusion**

To quickly recap, there were five hypotheses we have tested of which we have failed to reject only two null hypotheses and that is, overall, in all counties, we are seeing a high booster shot intake among youth and mid-age group and we are seeing a few common counties irrespective of the ethnicity where the vaccinate rate is low. To draw inferences and make decisions, there are other factors that are to be considered like socio-economic, cultural, geographical, and supply chain which would have influenced the data that we are visualizing today. To truly make decisions based on these visualizations would require us to research and study the causes behind the results of this data.

**1.** **People of Black Ethnicity have received the least number of booster shots:** The visualization has considered 4 major ethnic groups: Asian, Hispanic, Black, and White. From the visualizations, we can draw a conclusion that we can reject the null hypothesis as we see that the Asian ethnicity has received the least number of booster shots– 427,753.

**2.There are a few common counties where we see least vaccinated areas by ethnicity:** When we divide the data by county and by race, we see that the counties Austin, Angelina, Archer, and Bailey have the least full vaccination rates for most ethnicities. From the visualizations, we can draw to a conclusion that we fail to reject the hypothesis**.**

**3.There is an increasing trend in the amount of booster shots taken by individuals from Aug 2021 to Jun 2022:** We can see that when the vaccine boosters were rolled out in Aug. 2021, there was a month over month increase in the vaccines administered. Since Jan 2022 there is a decline in the number of vaccines administered every month. Therefore, we reject the Null Hypothesis.

**4. Overall in all counties, we are seeing a high booster shot intake among youth and mid-age group:** 39% of the booster doses that have been administered in all the counties have been to the age group 16-49 which is the age group that we are testing in our visualization. Hence, we fail to reject the null hypothesis.

**5. Everyone that has taken a single vaccine dose has also taken the second vaccine dose and the booster dose:** We see that there is a downward trend after a single vaccine dose is administered. Based on this information we will reject the null hypothesis.