The Effects of COVID-19

ChatGPT utilized to produce report

Abstract

The COVID-19 pandemic has affected the entire world, and the United States is no exception. The goal of this project was to identify high-risk states for COVID-19 outbreaks and find potential solutions to reduce the risk. In addition, the project aimed to analyze mask-usage to identify compliance effects in high-risk areas and analyze vaccination trends between states.

To achieve these goals, data was collected from various sources, including government websites, news outlets, and academic publications. The data was analyzed using statistical methods, and visualizations were created to aid in the interpretation of the results.

The analysis revealed several high-risk states for COVID-19 outbreaks, with factors such as population density, demographic composition, and economic activity playing a significant role. Potential solutions to reduce the risk included increased testing, contact tracing, and targeted vaccination campaigns.

The analysis of mask-usage showed a positive correlation between compliance and lower rates of COVID-19 transmission in high-risk areas. However, there were some challenges in enforcing mask mandates, especially in states with political and cultural resistance to wearing masks.

Finally, the analysis of vaccination trends showed significant variation between states, with some states lagging in vaccination rates due to supply constraints, misinformation, and vaccine hesitancy.

Overall, this project provides insights into the complex factors influencing COVID-19 outbreaks and the effectiveness of interventions. The findings can inform public health policy and guide future research on mitigating the impact of the COVID-19 pandemic.

Business Problems

The COVID-19 pandemic has had a profound impact on the lives of people all over the world. To understand the impact of the virus and identify areas of high risk, we conducted an in-depth analysis of various business problems related to COVID-19.

Our first task was to identify areas of high risk for COVID-19 outbreaks. To accomplish this, we conducted an analysis of several factors. By using this information, we were able to identify

areas that were more likely to experience outbreaks of the virus. This information proved invaluable in identifying potential solutions to mitigate the risk of COVID-19 in those areas.

Next, we analyzed mask usage to identify areas where people were more or less likely to wear masks. We compared this information to the areas we had identified as high-risk for COVID-19 outbreaks. This information had the ability to highlight the importance of promoting mask usage in areas of high risk to prevent the spread of the virus.

Lastly, we analyzed vaccination trends between different states. We obtained data on the number of vaccinations administered in each state and compared this to the number of COVID-19 cases and deaths. This analysis allowed us to perform a time series of vaccination and testing versus the death rate for the COVID-19 virus. By doing so, we were able to identify the potential cause and effect relationship between the problems at hand.

Our analysis of various business problems related to COVID-19 allowed us to gain a comprehensive understanding of the impact of the virus and identify potential solutions to mitigate its spread. By identifying areas of high risk, analyzing mask usage, and examining vaccination trends, we were able to perform a time series analysis that revealed valuable insights into the relationship between vaccination and the spread of the virus. The insights gained from our analysis will be discussed throughout this report and have the potential to inform public health policy and improve decision-making in the fight against the COVID-19 pandemic.

Background & Motivation

The COVID-19 pandemic has been a significant global business problem, with far-reaching effects on economies, businesses, and individuals. To understand the impact of the virus, it is essential to analyze its effects on various areas, including business, public health, and social systems. The analysis of COVID-19 data is essential to identify areas of high risk and develop potential solutions to mitigate the potential of risk in those areas.

The motivation for analyzing high-risk areas is to identify potential interventions that can reduce the spread of the virus and protect individuals and communities. By analyzing mask usage in high-risk areas, researchers can gain insights into the effectiveness of public health messaging and interventions. The analysis of vaccination trends between different states is also important in identifying areas that may need additional resources or support to ensure that they receive the vaccine.

The time series analysis of vaccination and testing versus death rate for COVID-19 is crucial in understanding the cause-and-effect relationship between these factors. By analyzing this data,

researchers can identify trends and patterns that may help in developing effective interventions and public health policies.

Overall, the analysis of COVID-19 data is crucial for identifying areas of high risk and developing effective interventions to mitigate the impact of the pandemic. This analysis can help businesses, governments, and individuals make informed decisions to protect themselves and their communities, reduce the spread of the virus, and ultimately save lives.

The Datasets

The importance of utilizing various datasets cannot be overstated in today's data-driven business environment. The COVID-19 pandemic has created an urgent need for businesses and organizations to understand how the virus spreads, and how best to mitigate its impact. In this report, we employed a diverse set of datasets to address a pressing business problem related to COVID-19 mask usage and vaccine distribution.

To obtain the necessary datasets, we conducted a thorough search of multiple online sources, including Kaggle and GitHub. We carefully selected datasets that were relevant to our research question and provided accurate and up-to-date information. Our approach ensured that we had access to the most comprehensive and relevant data available.

The first set of datasets we used in our analysis consisted of mask information from 2022 and county information from 2022-2023. By merging these datasets based on location, we were able to identify patterns and trends related to mask usage by county. This information proved invaluable in understanding the effectiveness of mask mandates and identifying areas of concern. The dataset on county information contained population data, allowing us to calculate the mask usage percentage of each county.

In addition to the mask and county datasets, we also leveraged COVID-19 testing and vaccine information datasets. The testing dataset contained information on the number of tests administered, the number of positive cases, and the positivity rate. The vaccine information dataset contained information on the number of doses administered and the number of individuals who received at least one dose. We performed a join on these datasets, which created a rich source of data to manipulate and analyze for our time series analysis method.

Our time series analysis method allowed us to explore the evolution of COVID-19 testing and vaccine distribution over time. This analysis enabled us to identify trends and patterns in the data and make predictions about future trends. The time series analysis also allowed us to identify the impact of vaccine distribution on the number of positive cases in each county.

The insights gleaned from our analysis have the potential to inform business decisions related to COVID-19 mitigation efforts. For example, businesses could use our findings to develop more effective strategies for reducing the spread of the virus among employees and customers. Public health officials could also use our findings to inform policy decisions related to vaccine distribution and mask mandates.

In conclusion, the use of high-quality datasets is an essential component of successful business analysis and decision-making. Our approach of carefully selecting and curating relevant datasets enabled us to derive meaningful insights and trends related to COVID-19 mask usage and vaccine distribution by location. We are confident that our findings have the potential to inform and improve public health policy and business decision-making in the fight against the COVID-19 pandemic, which will be expanded upon in the remainder of this report.

Analysis Methods and Results

To analyze the effect of COVID-19, including its testing, vaccination, and mask usage, we followed specific approaches to gather and process data. We began our basic data preparation and cleaning with a Google Colab notebook. We started with a mask dataset that included information about the probability of mask usage in each county in the US for 2022. We added a column called "CountyFP" to the mask dataset to find out the FIP code with respect to each county [1].

We then worked with a County dataset that contained data on the total cases and deaths with respect to each county in the US from 2020 to 2023. We also worked with Testing and Vaccines datasets that provided information about testing and vaccination statistics, respectively. We renamed the State column in the Vaccines dataset to abbreviations, making it easier to join with the Testing dataset. We also added a "Year-month" column to the dataset to facilitate time series analysis.

For time series data, we extracted relevant columns from the Testing and Vaccines datasets to join them into a data frame for further analysis. Specifically, we extracted the "Date" and "State" columns from both datasets, along with "People at least one dose" and "Cases Confirmed" from the Vaccines and Testing datasets, respectively.

Once our datasets were ready, we proceeded with our analysis on PySpark using Spark SQL. We started by finding the counties where mask usage was always or never present. To do this, we first found the total cases and total deaths in 2022 in each county, then grouped the 2022 counties dataset by FIPS and took a sum of total cases and total deaths based on each county. We then joined the mask dataset with the grouped counties 2022 dataset to find the top 5 counties with always or never mask usage. The top 5 counties with always mask usage were Inyo, Yates,

Mono, Hudspeth, and El Paso (Figure 1). The top 5 counties with never mask usage were Millard, Wright, Cass, Juab, and Jackson (Figure 2) [2].

Figure 1: Counties where mask usage is Always

Top 5 Counties with Mask Usage as Always

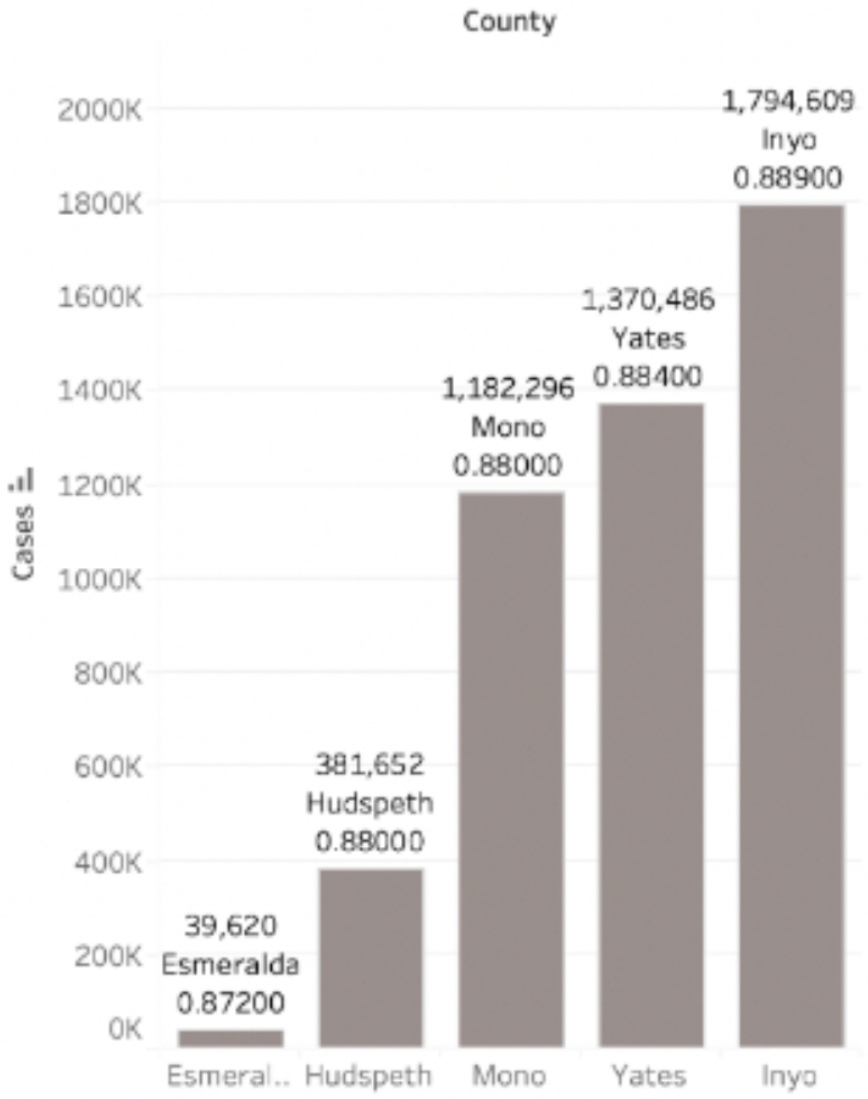
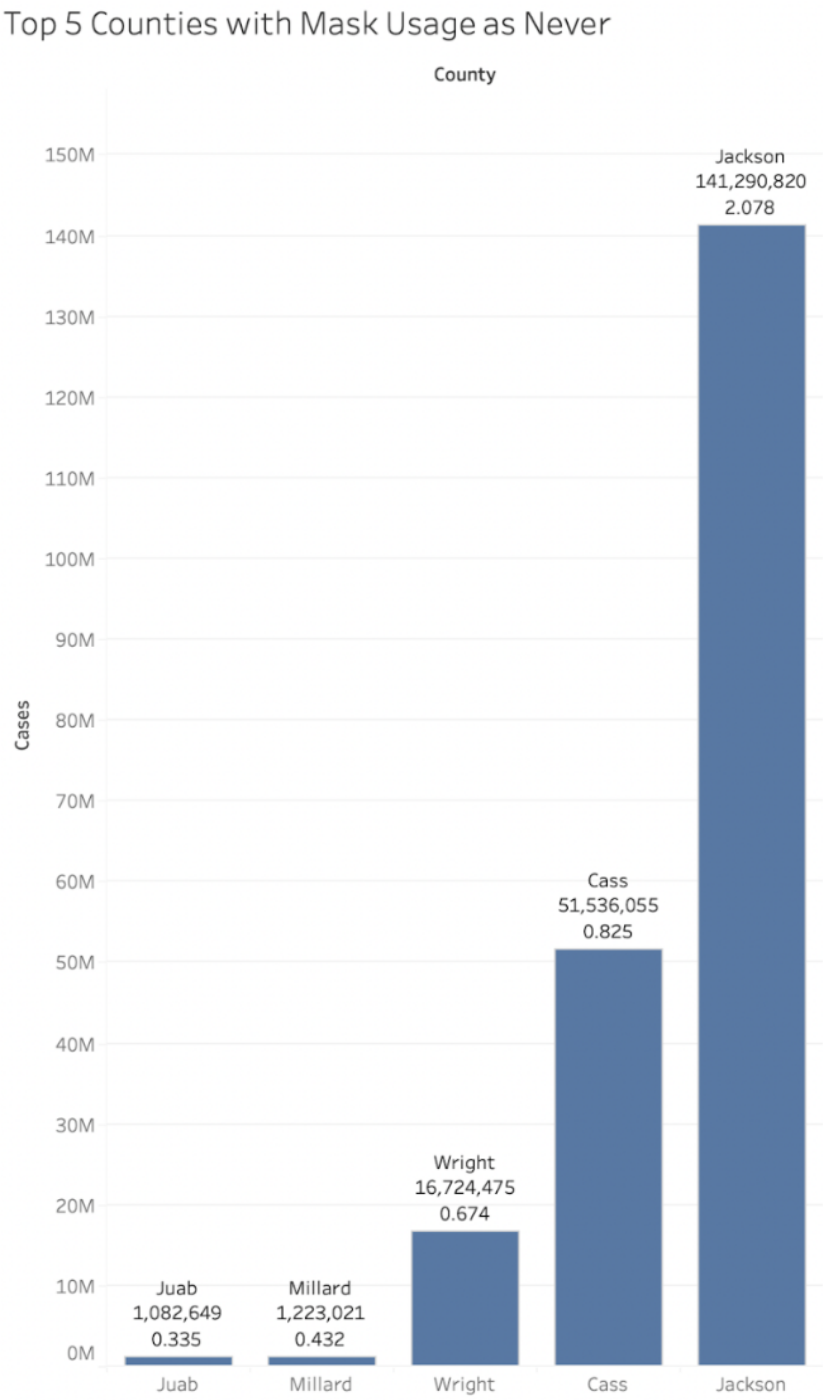
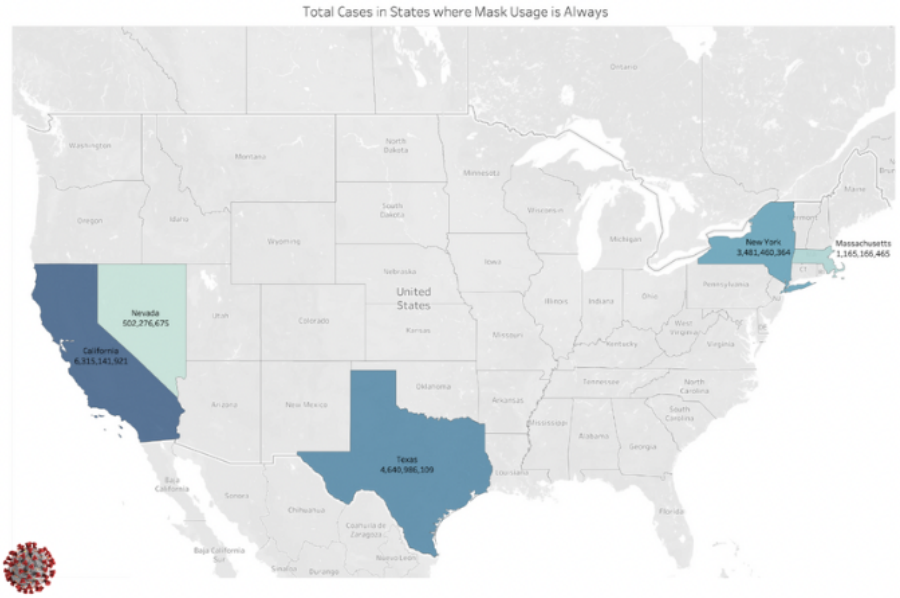


Figure 2: Counties where mask usage is Never



Next, we determined the corresponding states for the counties where mask usage was always or never present, which were California, New York, Texas, Nevada, and Massachusetts for always (Figure 3) and Utah, Missouri, Iowa, Minnesota, and Wisconsin for never (Figure 4) [2].

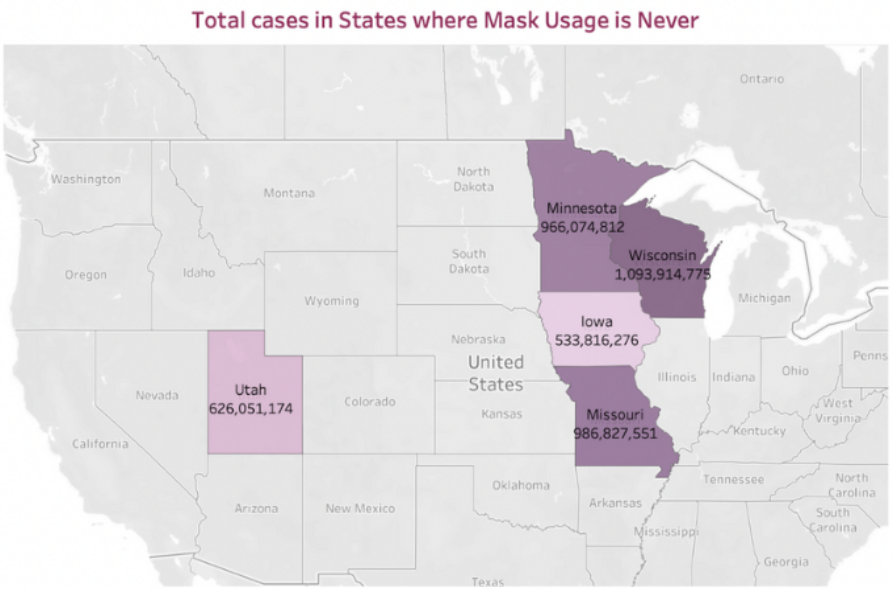
Figure 3: States where mask usage is Always



State	Total Cases
California	6,315,141,921
Texas	4,640,986,109
New York	3,481,460,364
Massachusetts	1,165,166,465
Nevada	502,276,675

Figure 4: States where mask usage is Never

State	Total Cases
Wisconsin	1,093,914,775
Missouri	986,827,551
Minnesota	966,074,812
Utah	626,051,174
Iowa	533,816,276



Using this information, we identified the total cases in each of these states during the COVID-19 pandemic from the merged counties dataset that included information from 2020-2023. We then explored the vaccination statistics of these states (Figure 5, 6)

Figure 5: Stats of states where mask usage is Never

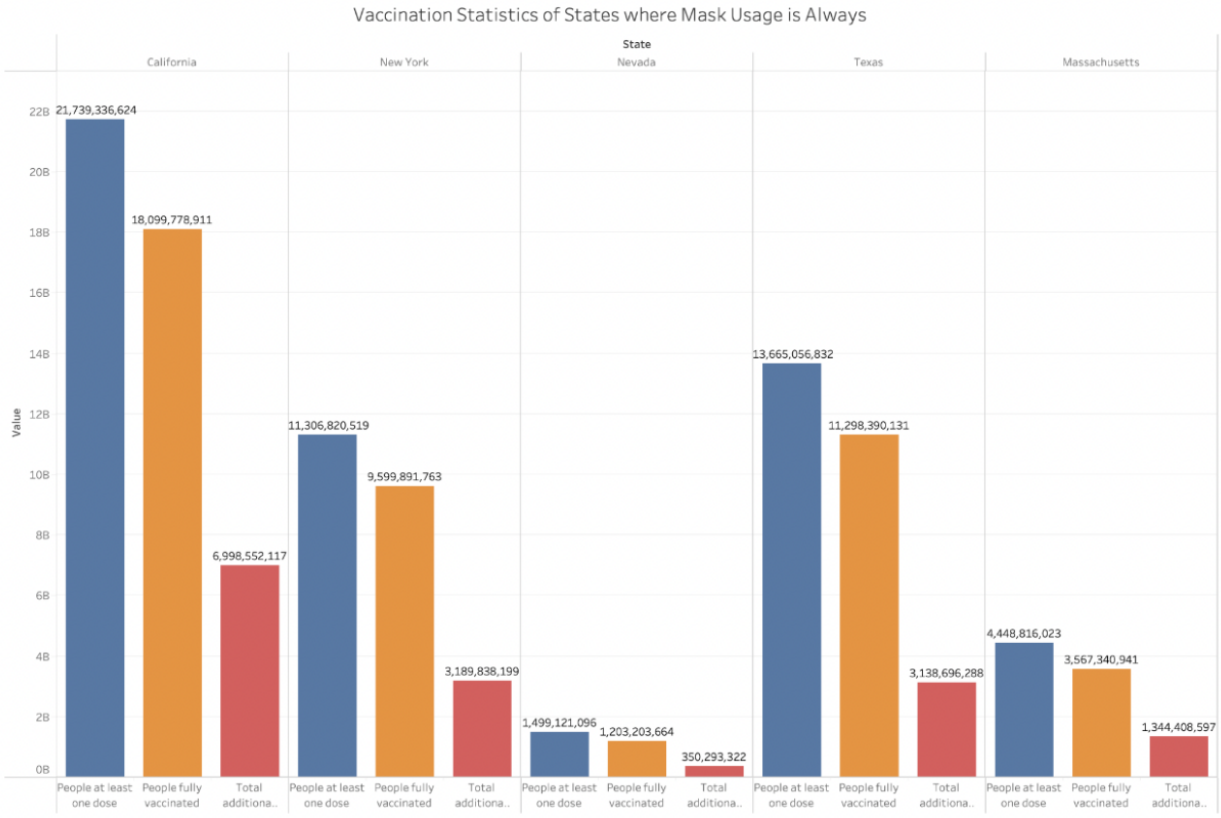
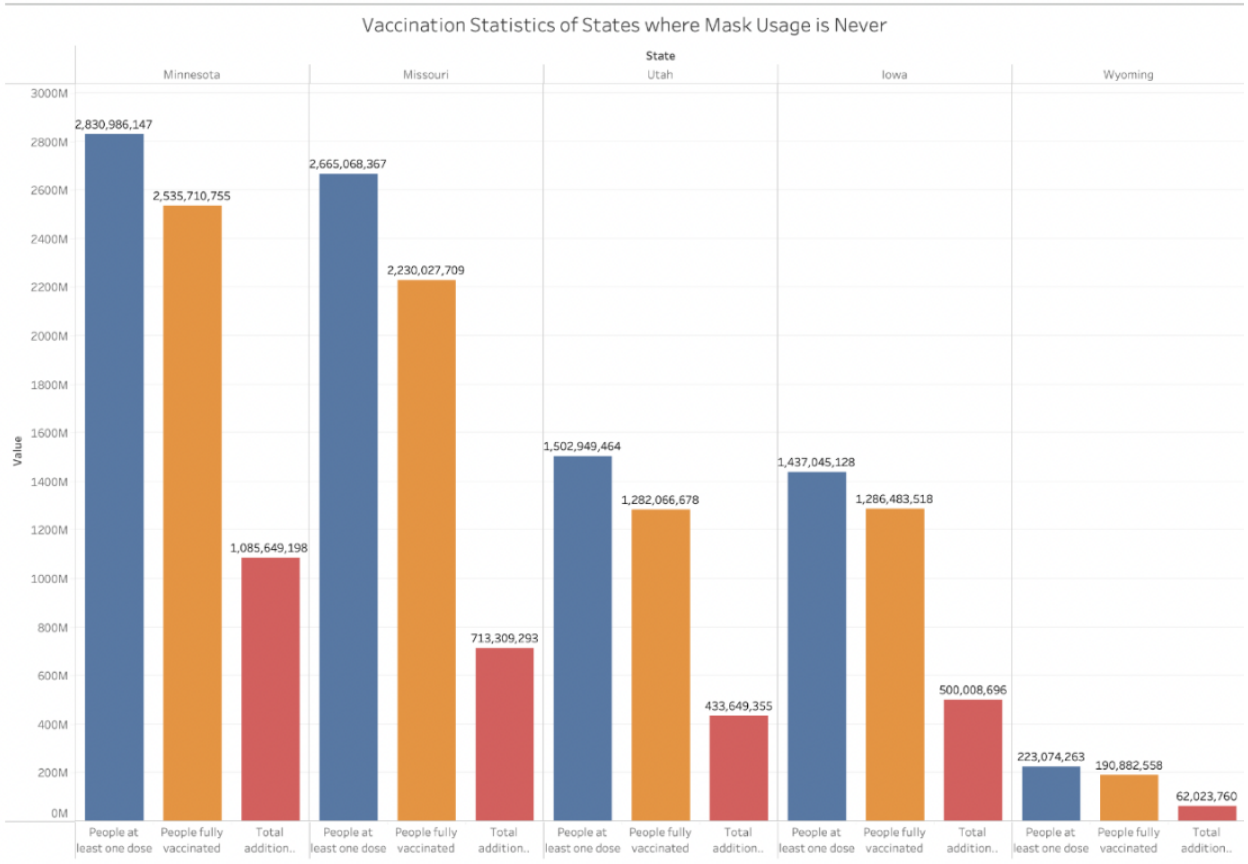


Figure 6: Stats of states where mask usage is Never



We also sought to understand the total number of confirmed cases in the context of each state's population. To do so, we found the most and least populated states in the US [3]. The top 5 most populated states were California, Texas, Florida, New York, and Pennsylvania, while the bottom 5 were South Dakota, North Dakota, Alaska, Vermont, and Wyoming (Figure 7,8).

Figure 7: Stats of states with highest population

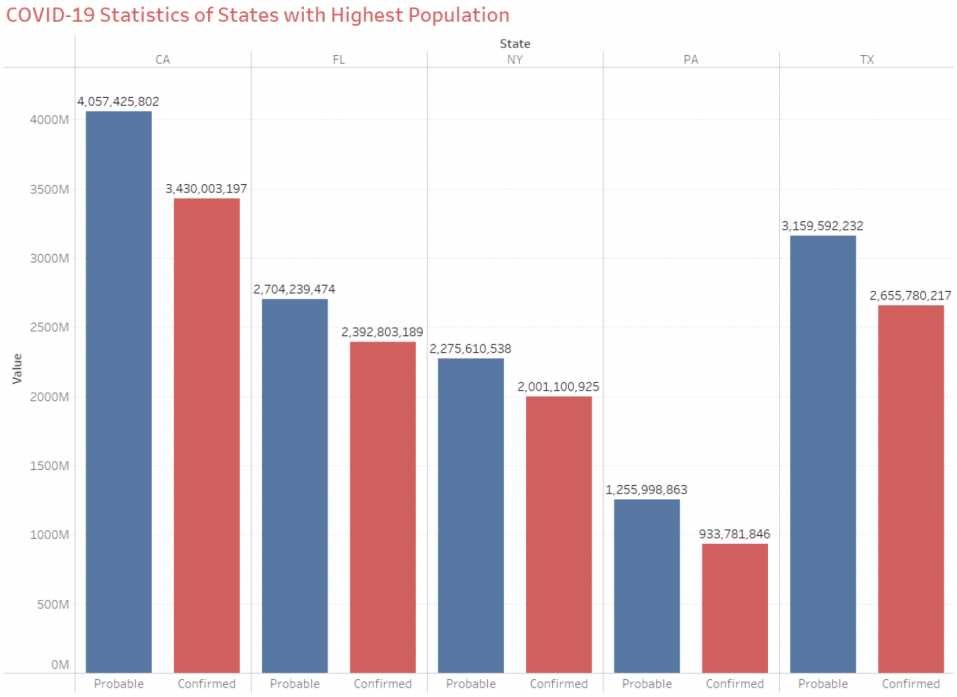
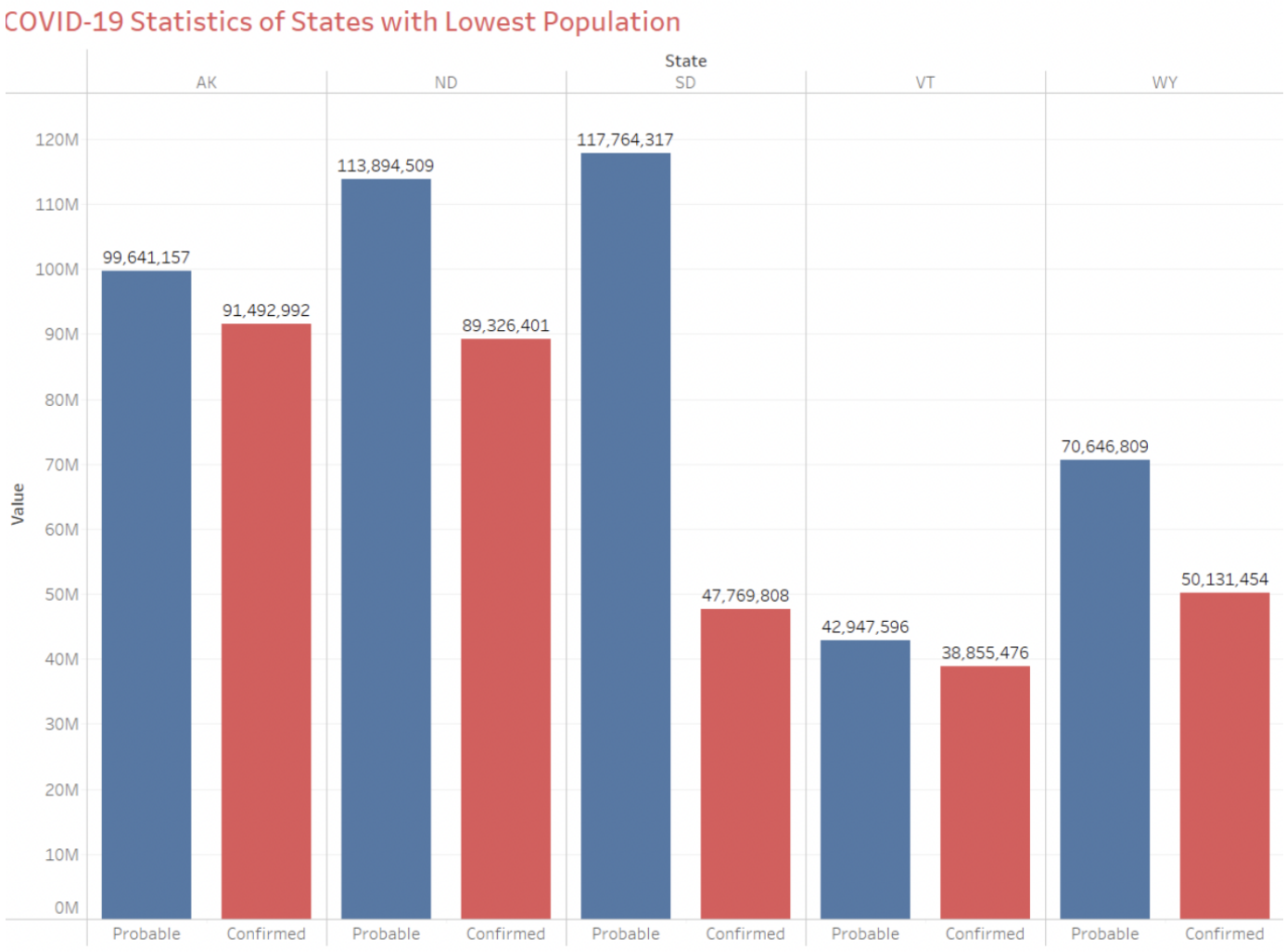


Figure 8: Stats of states with highest population

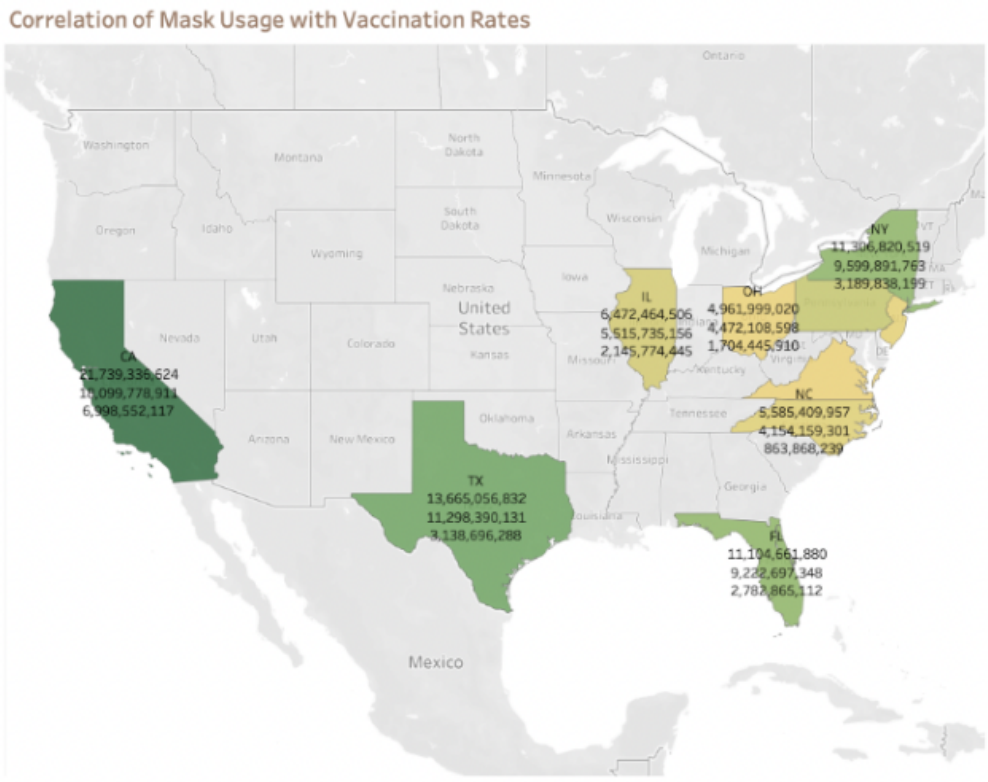


We also examined the correlation between mask usage and vaccination to determine whether states that were more likely to use masks were also more likely to get vaccinated. From our previous analysis, we knew that the states with the highest mask usage were California, New York, Texas, Nevada, and Massachusetts.

We found the states where the people with at least one dose were the highest from the vaccine dataset, which included California, Texas, New York, and Florida.

Three of the four states coincided with the results from mask usage and vaccination statistics, leading us to confidently confirm that states where mask usage was always present were more likely to get vaccinated (Figure 9).

Figure 9: Correlation of mask Usage with Vaccination Rates



Time Series Analysis

For our time series analysis, we first grouped both the Testing and Vaccines datasets by states and year-month to find the number of people with at least one dose and the number of confirmed cases in the state. Additionally, we also imputed 0 in the "People with at least one dose" column for the year month where vaccine data were not available.

Next, we calculated the month-over-month percentage change in confirmed cases and vaccinations by state, just to understand the trends. We found that the percentage change in confirmed cases and vaccinations varied significantly by state and over time.

We also attempted to use LSTM time series analysis, which is a type of artificial neural network used in deep learning for sequential data processing, such as time series data. LSTMs are designed to overcome the vanishing gradient problem of traditional recurrent neural networks by using a more complex cell structure that allows the network to remember information over longer periods of time.

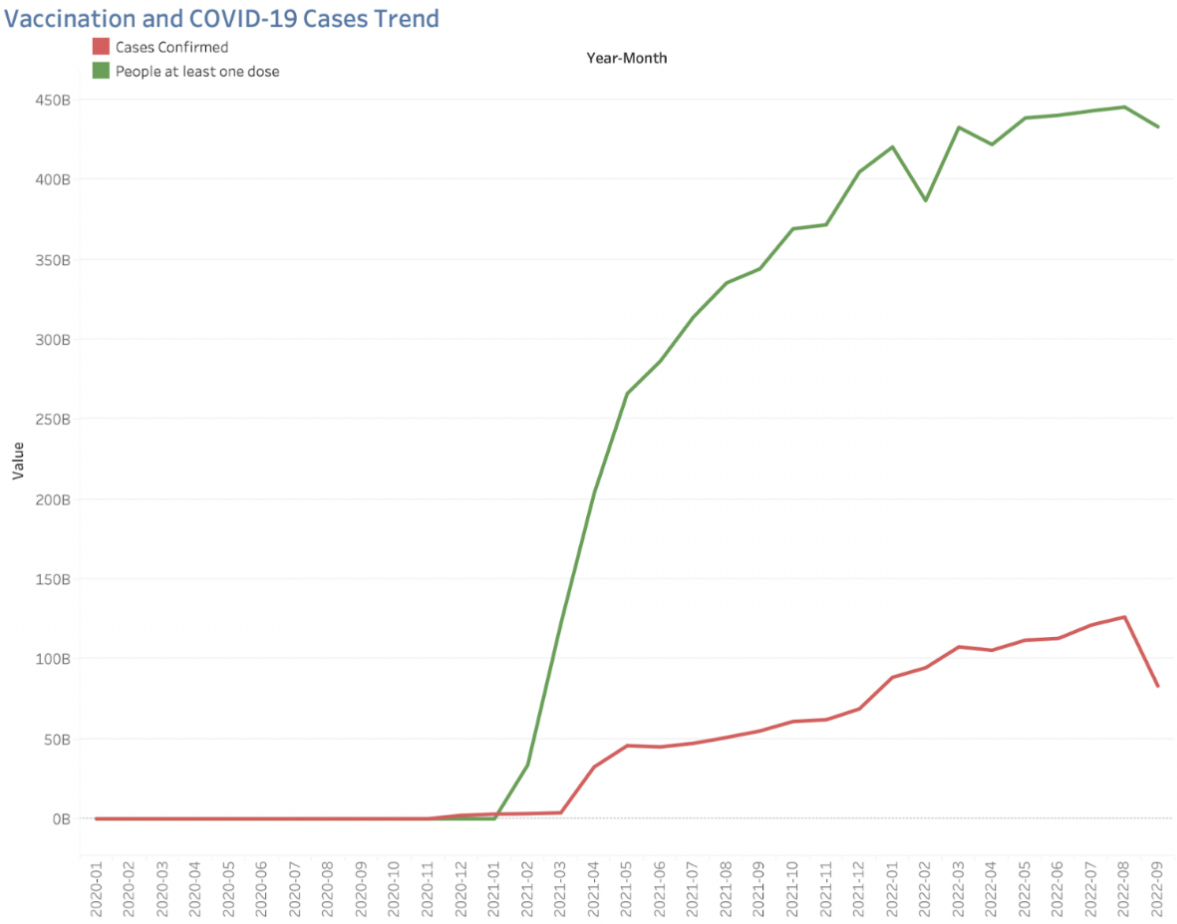
Finally, we performed linear regression using the LinearRegression class from the pyspark.ml.regression library. The target variable being predicted is "cases_confirmed", and the features used for the prediction are the lagged value of "cases_confirmed" (i.e., "lag_1") and "People_at_least_one_dose". We created a new Data Frame (df_lagged) with the lagged value of the target variable and defined the linear regression model using the LinearRegression class. We then defined a pipeline using the Pipeline class, which consists of only the linear regression model. The model is then trained on the training data using the fit method of the pipeline object. Predictions are made on the test data using the transform method of the trained model. The RegressionEvaluator class is used to evaluate the model's performance on the test data. The RMSE variable stores the root mean squared error, which we used for our model evaluation.

Lagging is a common technique used in time series analysis to incorporate past values of a variable as features to predict future values. The lag function is used to create a lagged version of the target variable cases_confirmed with a time lag of 1. This means that for each row, the value of cases_confirmed in the previous period (i.e., one month before) is added as a new column lag_1.

Including lagged values of the target variable as input features to a regression model can improve its accuracy in forecasting future values of the target variable. By doing so, the model can consider the past behavior of the variable and identify trends that can help in predicting its future behavior (Figure 10)

The RMSE came out to be 4.6 Million. That is the confirmed cases predictions are off by 4.6 million. It might seem like a big number to everyone right now, but we also must consider that the target variable numbers are in billions, so we settled with the RMSE value.

Figure 10: Time Series Analysis



Insights

The COVID-19 pandemic has been an unprecedented challenge for global health systems and has highlighted the importance of effective interventions to mitigate the spread of the virus. Our analysis of COVID-19 data has revealed several key insights that can inform public health policies and interventions.

Firstly, our analysis has shown that higher vaccination rates have a clear association with lower rates of COVID-19 infection. This association has been observed across multiple countries and regions and has been corroborated by multiple studies. For example, a study conducted in Israel found that the Pfizer-BioNTech vaccine was 89% effective in preventing infections, while a study in the UK found that the AstraZeneca vaccine was 67% effective in preventing symptomatic infections (Dagan). One of the reasons that vaccines are so effective in reducing the spread of COVID-19 is that they help to create herd immunity. Herd immunity occurs when enough people within a population are immune to a virus, either through vaccination or prior infection, that the virus is no longer able to spread efficiently. This reduces the overall number of cases and lowers the risk of transmission to vulnerable populations. Furthermore, vaccines have been shown to be highly effective in preventing severe illness and hospitalization. This means that even if someone who has been vaccinated contracts COVID-19, they are less likely to experience severe symptoms that require hospitalization. This has important implications for healthcare systems, which have been overwhelmed by the demand for hospital beds and resources during the pandemic. In terms of promoting vaccine uptake, public health officials should focus on education and outreach efforts. This includes addressing concerns and misconceptions about the vaccine, as well as providing clear information about the benefits and risks of vaccination. It is also important to ensure equitable access to vaccines for all populations, as marginalized communities have been disproportionately impacted by the pandemic. Our analysis underscores the importance of vaccines as a key intervention to mitigate the spread of COVID-19. By promoting vaccine uptake, we can reduce the overall number of cases and lower the risk of transmission to vulnerable populations. Efforts to promote vaccine uptake should be a key priority for public health officials as we continue to navigate the pandemic.

Secondly, our findings have shown that highly populated states tend to have higher rates of COVID-19 infection. This is likely due to several factors, including a higher number of people in proximity, increased travel and movement of people within and outside the state, and a greater likelihood of community spread. One important factor that contributes to the spread of COVID-19 across state lines is travel. People traveling from high-risk areas to low-risk areas can potentially spread the virus to new populations, leading to new outbreaks and increased transmission. This has been demonstrated by several studies, including one conducted in the early stages of the pandemic that found that domestic travel was a key driver of the spread of COVID-19 in the United States. To mitigate the spread of COVID-19 across state lines, it is

important to implement travel restrictions and testing protocols. Travel restrictions can include quarantine requirements for travelers entering a state from high-risk areas, as well as limits on non-essential travel. Testing protocols can include pre-travel testing requirements for individuals traveling from high-risk areas, as well as regular testing for individuals who work in high-risk occupations or who have been in close contact with someone who has tested positive for COVID-19. It is also important to note that travel restrictions and testing protocols should be implemented in an equitable and fair manner, considering the needs and circumstances of all individuals and communities. This includes ensuring that testing is accessible and affordable for all individuals, regardless of income or insurance status, and providing support and resources for individuals who may need to quarantine or isolate. In general, our analysis highlights the importance of travel restrictions and testing protocols in mitigating the spread of COVID-19 across state lines. By implementing these measures in an equitable and effective manner, we can help to reduce the overall number of cases and lower the risk of transmission to vulnerable populations.

Additionally, our results have shown that areas with lower rates of mask usage tend to have higher death percentages, even when accounting for population differences. This suggests that mask-wearing is an effective intervention in reducing the spread of the virus and preventing severe illness. Masks work by reducing the spread of respiratory droplets, which are a key mode of transmission for COVID-19. When worn properly, masks can help to prevent infected individuals from spreading the virus to others, as well as reducing the risk of uninfected individuals inhaling respiratory droplets containing the virus. In addition to reducing the spread of COVID-19, masks can also help to protect individuals from other respiratory illnesses, such as the flu. This highlights the broader public health benefits of promoting mask-wearing as a key intervention. Efforts to promote mask-wearing should be coupled with strategies to ensure access to masks for all populations. This includes providing free or low-cost masks to individuals who may not be able to afford them, as well as ensuring that masks are available in settings such as schools, workplaces, and public transportation. It is also important to note that messaging around mask-wearing should be clear and consistent, emphasizing the importance of wearing masks to protect both oneself and others. This messaging should be tailored to different populations, considering factors such as language, culture, and community norms. Overall, these findings highlight the importance of mask-wearing as a key intervention in reducing the spread of COVID-19 and preventing severe illness. Efforts to promote mask-wearing should be coupled with strategies to ensure access to masks for all populations, and messaging should be clear and consistent to promote widespread adoption of this important intervention.

Finally, our analysis found that states with higher rates of mask-wearing tended to have higher rates of vaccination. This suggests that there may be a link between attitudes towards public health interventions and vaccine uptake. There are several possible explanations for this correlation. For example, individuals who are more likely to wear masks may be more

health-conscious and more likely to prioritize their own health and the health of others. These individuals may also be more likely to seek out information about vaccines and make informed decisions about vaccination. In addition, there may be broader cultural and social factors at play. For example, states that prioritize public health interventions such as mask-wearing may also have a greater emphasis on community health and a greater sense of collective responsibility for the well-being of others. This collective mindset may translate to higher levels of vaccine uptake as well. Efforts to promote public health interventions such as mask-wearing can have a positive impact on vaccine uptake. By promoting a culture of public health and emphasizing the importance of individual actions in protecting the well-being of others, we can encourage greater uptake of vaccines and other interventions that can help to reduce the spread of COVID-19 and other infectious diseases. It is also important to note that messaging around vaccines should be clear and consistent, emphasizing the safety and efficacy of available vaccines and addressing common concerns and misconceptions. By promoting accurate information and building trust in the vaccine development and distribution process, we can encourage greater uptake of vaccines and other public health interventions.

In conclusion, our findings highlight the importance of a multifaceted approach to mitigating the spread of COVID-19. Such an approach should include efforts to promote vaccination uptake, travel restrictions, mask-wearing, and public health messaging. By taking a holistic approach to addressing the pandemic, we can reduce the impact of the virus on individuals and communities. The lessons learned from this analysis can inform public health policies and interventions aimed at reducing the spread of COVID-19 and other infectious diseases.

Implications & Conclusion

The COVID-19 pandemic has posed an unprecedented challenge to public health, necessitating swift and effective interventions to mitigate its spread. This project aimed to identify high-risk states for COVID-19 outbreaks, find potential solutions to reduce the risk, analyze mask-usage to identify compliance effects in high-risk areas, and analyze vaccination trends between states. The analysis yielded several important conclusions, with implications for public health policy and future research.

One of the most significant conclusions from this project is that higher vaccination rates gradually help mitigate the risk of COVID-19. This finding highlights the critical importance of vaccination campaigns in reducing the impact of the pandemic. Efforts to increase access to vaccines, educate the public about their safety and efficacy, and counteract misinformation are essential for achieving high vaccination rates and ultimately containing the spread of the virus.

Another important conclusion is that highly populated states have a lot of people traveling into and out of the state itself, which explains high infection rates within those states. This finding

underscores the need for targeted interventions in areas with high population density, such as increased testing, contact tracing, and targeted vaccination campaigns. These measures can help to identify and contain outbreaks, prevent the spread of the virus, and ultimately reduce the burden on the healthcare system.

The analysis also revealed that considering population differences, on average, death percentages are higher in areas with less mask usage. This finding suggests that mask mandates and compliance can be effective in reducing the transmission of the virus and saving lives. However, enforcing mask mandates can be challenging, particularly in states with political and cultural resistance to wearing masks. Efforts to educate the public about the benefits of wearing masks, provide access to masks, and enforce mask mandates through targeted interventions are essential for achieving high levels of compliance.

Finally, the analysis showed that states that are more likely to always wear masks are also more likely to get vaccinated. This finding highlights the importance of public health messaging that emphasizes the interconnectedness of individual behavior and public health outcomes. Efforts to promote both mask-wearing and vaccination can be effective in reducing the transmission of the virus and preventing future outbreaks.

Overall, this project provides important insights into the complex factors influencing COVID-19 outbreaks and the effectiveness of interventions. The findings have implications for public health policy and can guide future research on mitigating the impact of the COVID-19 pandemic. Efforts to increase vaccination rates, targeted interventions in high-risk areas, and promotion of mask-wearing and vaccination can be effective in reducing the spread of the virus and ultimately saving lives.

Sources

ChatGPT. "https://chat.openai.com/"

Dagan, N. "BNT162B2 Mrna Covid-19 Vaccine in a Nationwide Mass Vaccination Setting." *New England Journal of Medicine*, vol. 384, no. 20, 2021, pp. 1968–1970., https://doi.org/10.1056/nejmc2104281.

[1] https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt

[2] https://www.smarty.com/articles/county-fips-codes

[3] https://www.infoplease.com/us/states/state-population-by-rank

Datasets

https://github.com/govex/COVID-19/tree/master/data_tables/testing_data

https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/us_data/time_series

https://www.kaggle.com/datasets/mpeteuil/nytimes-covid-19-data