Analysis of COVID-19 Vaccination Trends: Distribution and Administration Between 2021 and 2023

Natalie Cann

**Authors**

* Natalie Cann

**Author Affiliations**

1. Masters in Public Health Candidate, University of Georgia

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect those of their university affiliations.

# 1. Summary/Abstract

This project presents an analysis of COVID-19 vaccine distribution and administration trends in the United States from 2021 to 2023, utilizing publicly available data from the Centers for Disease Control and Prevention (CDC). The analysis focuses on three primary research questions: (1) How does COVID-19 vaccine distribution and administration differ by region of the United States? Do certain manufacturers of the COVID-19 vaccine have a greater number of administered doses than others? (2) How has the vaccine distribution and administration by region in the United States changed with time? Has vaccine distribution and administration by manufacturer changed with time? (3) Can the proportion of Pfizer vaccines administered be predicted by several synthetic variables and MMWR week? Exploratory data analysis revealed that the Pfizer COVID-19 vaccine led in both the number of doses distributed and administered, followed by Moderna. Regional analysis uncovered that the South consistently had the highest number of doses distributed and administered. The models to predict the proportion of Pfizer vaccines administered from synthetic variables and MMWR week did not perform well. This indicates that these variables are not strong predictors of the proportion of Pfizer vaccines administered. A key limitation in this project was the lack of appropriate predictor variables within the CDC dataset. This underscores the need for better data collection and reporting of vaccine distribution and administration. Further research is needed to assess the factors influencing COVID-19 vaccine manufacturer specific distribution and administration.

# 2. Introduction

## 2.1 General Background Information

Respiratory illnesses are a large burden to both medical and public health systems. These illnesses are known to spike in the colder months, such as December, January, and February. One such respiratory virus is COVID-19, also known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1). COVID-19 was declared a pandemic in 2020. The virus is known to spread through fluid droplets shared between individuals in close contact as well as through the air (1). The spike protein is known to mediate entry of the virus into host cells via its S1 subunit, which binds to a receptor on the host cell, and S2 subunit, which allows the virus to fuse with the host cell (2). Therefore, the spike protein is a major antigen of the COVID-19 virus recognized by the immune system.

There are two different types of COVID-19 vaccinations available in the United States (U.S.). The first is the mRNA vaccine, which uses mRNA created in a laboratory that triggers an immune response within the body that produces antibodies to protect against COVID-19 (3). The second is the protein subunit vaccine, which contains spike proteins from the virus as well as an adjuvant; this enables the body to produce antibodies and better recognize the spike protein in the future (3). The Pfizer-BioNTech and Moderna vaccines are both mRNA vaccines (4). The Novavax vaccine is a protein subunit vaccine (4). The Janssen vaccine, also known as the Johnson and Johnson vaccine was discontinued (4). This vaccine utilized an adenovirus as a vector (5). Adenovirus vaccines use an omnipresent virus that is altered so that it cannot cause illness or integrate into the host’s DNA (5). The adenovirus is engineered to contain genes involved in making the spike protein COVID-19, which then leads to the immune system producing antibodies for the spike protein (5).

Research on the distribution of vaccines as a whole has been relatively limited. Vaccine distribution and administration have not had the best documentation throughout history in both the U.S. and globally (6). Few states have effective systems to track vaccine distribution, administration, inventory, and demand (6). As a result of this, in the U.S., some regions receive more vaccines than are used, while other regions do not receive enough vaccines to fulfill the demand (6). Not only is this an ineffective use of resources, but it also has an economic cost (6). Successful distribution and administration of vaccines, especially in times of emergency, is crucial as it can lower the burden on healthcare systems and lower mortality (7). These issues convey the importance of effective vaccine data collection and reporting in order to improve the efficiency of vaccination campaigns.

While the COVID-19 vaccines were developed in a timely response, the U.S. was not prepared to effectively distribute these vaccines. Healthcare workers and elderly individualized were prioritized for vaccination, which was a great first step. However, there proved to be issues with the distribution of the vaccines. First, vaccines distributed to pharmacies and other small venues were administered at a low rate, indicating poor distribution plans (7). Additionally, the storage and transport of the COVID-19 vaccines was a challenge as the vaccine needed to be kept at a very low temperature. This proved to be difficult to incorporate into planning effective distribution of the COVID-19 vaccine (7). Research must be done in order to better respond to regional vaccine demand in emergency, and non-emergency, situations in the future.

Research has been done to determine ways in which the distribution of the COVID-19 vaccine during the pandemic could be improved. A study by Muckstadt et al. (2023) proposed a new distribution strategy that would’ve allow for more effective distribution of the COVID-19 vaccine in the U.S. during the pandemic (7). Another study by Bertsimas et al. (2021) proposed the use of the DELPHI model to capture the effects of the vaccine and differing mortality rates across age groups to improve the allocation of the COVID-19 vaccination (8). Building on this body of research, this project aims to assess whether the manufacturer-specific proportion of total vaccines administered can be predicted by several synthetic variables. Understanding these patterns may offer additional insight into how vaccine characteristics, supply chain logistics, and demographic or regional factors influence distribution outcomes.

## 2.2 Description of Data and Data Sources

The dataset used in this analysis came from the Centers for Disease Control and Prevention (CDC). This contains information on the distribution and administration of COVID-19 vaccinations between 2020 and 2023. However, the data from 2020 is not as well documented as that of 2021 through 2023; therefore, only data from 2021 to 2023 will be used in this project. The dataset contains both overall distribution and administration data, while also including distribution and administration specific to each manufacturer in the United States. Furthermore, there are variables stating the MMWR week that the data was reported during as well as the state from which the data was reported.

## 2.3 Questions to be Addressed

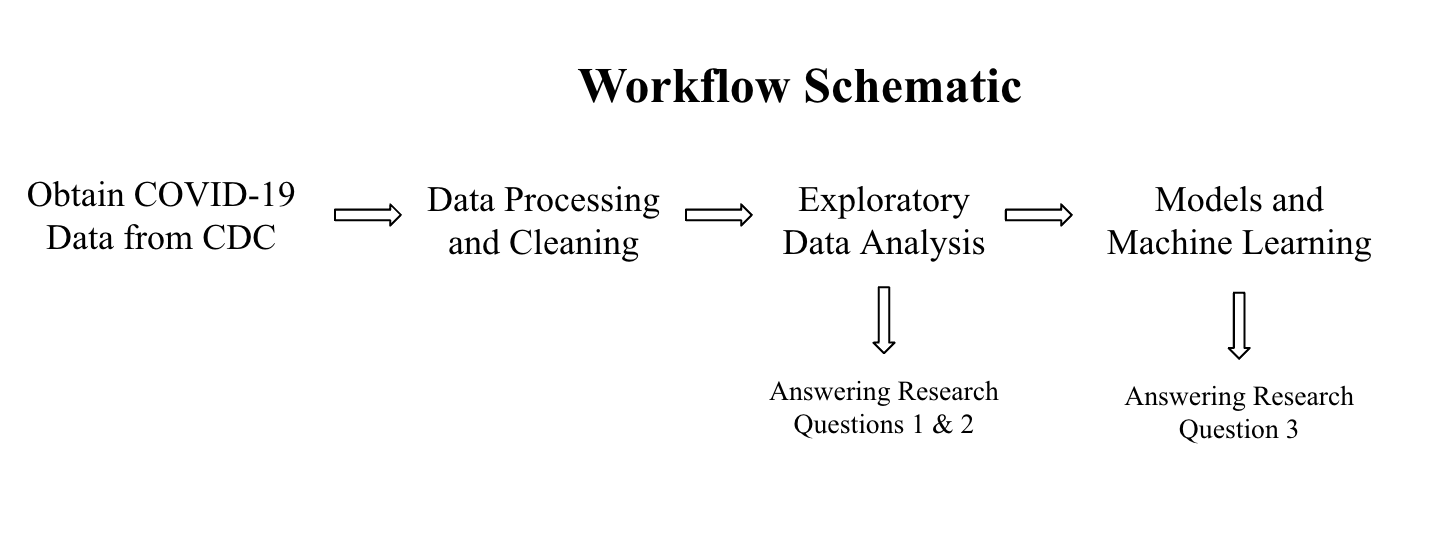
The three questions that will be addressed in this project are as follows:

1. How does COVID-19 vaccine distribution and administration differ by region of the United States? Do certain manufacturers of the COVID-19 vaccine have a greater number of administered doses than others?
2. Furthermore, how has the vaccine distribution and administration by region in the United States changed with time? Has vaccine distribution and administration by manufacturer changed with time?
3. Can the proportion of Pfizer vaccines administered by predicted by several synthetic variables (listed in the modeling section) and MMWR week?

# 3. Methods

## 3.1 Schematic of Workflow for Analysis

First, the data was obtained from the CDC. For more details on the dataset, please see the Data Acquisition section. Then, the data was processed and cleaned. The data was sorted by MMWR week and region. For more details on this, please see the Data Cleaning Process section. Then, exploratory data analysis was performed. From this, the answers to the first two research questions were obtained. Finally, in order to answer the final research question, models were performed. The models used were Simple Linear Regression with Polynomial Linear Regression, LASSO Regression, and Random Forest. The models were evaluated using RMSE (Root Mean Square Error) values and their residual plots were observed.



Schematic of Workflow for Analysis

## 3.2 Data Aquisition

**COVID-19 Dataset:** https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc/about\_data

This dataset came from the CDC and contains information on the distribution and administration of COVID-19 vaccinations between 2020 and 2023. Data is included from all sites that distribute and administer vaccinations in the U.S.; such as jurisdictional partner clinics, retail pharmacies, dialysis centers, long-term care facilities, Federal Emergency Management Agency (FEMA), and federal entity facilities. The CDC states that the dataset was provided by IISInfo. The data was downloaded directly from the CDC website provided above. This dataset was last updated May 12, 2023.

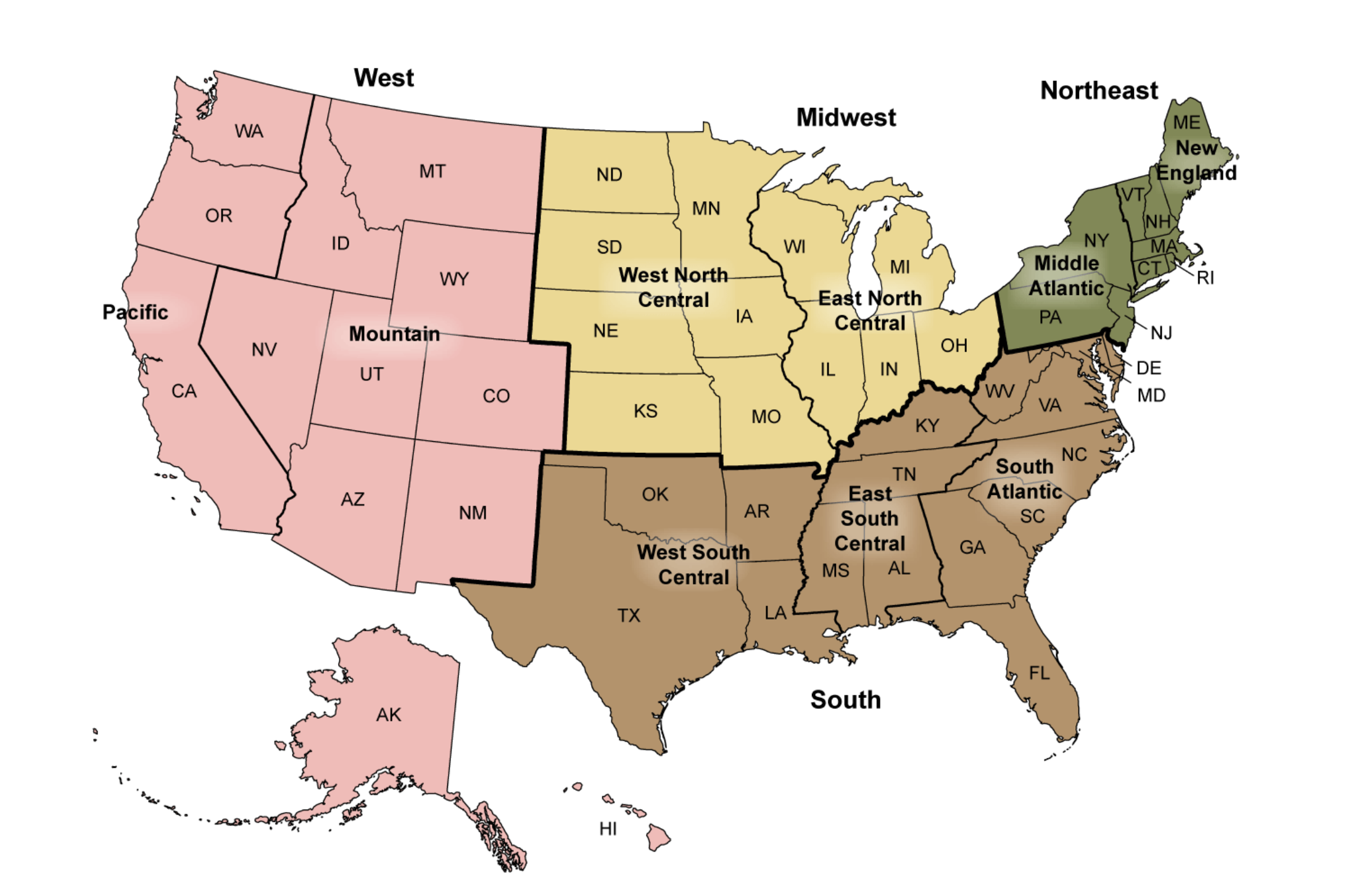
## 3.3 Data Cleaning Process

The COVID\_vaccine dataset contains 109 variables and 38,488 observations. The variables of special interest in this dataset are: Date, MMWR week, Location, Total Distributed Doses, Total Manufacturer-Specific Distributed Doses, Total Administered Doses, and Total Manufacturer-Specific Administered Doses.

I performed the following data processing and cleaning steps for the COVID-19 vaccine dataset:

1. Defined variables of interest
2. Assessed missingness of data and variable classes
3. Used the lubridate package to fix the date format
4. Select variables of interest to work with for analysis
5. Created Region variable and sorted U.S. states into their appropriate geographic regions
6. Manipulated dataset to only include data from 2021 to 2023 (data from 2020 was not as well documented)
7. Summed distribution and administration variables by MMWR week and region (each region had its own row for each MMWR week)
8. Created synthetic variables to use in modeling
9. Save final dataset as an RDS file

In order to categorize the states into their appropriate regions, the CDC’s geographic division regions were used (9). The graphic below displays these regions. For simplicity, only the Northeast, Midwest, South, and West regions were used in this project.

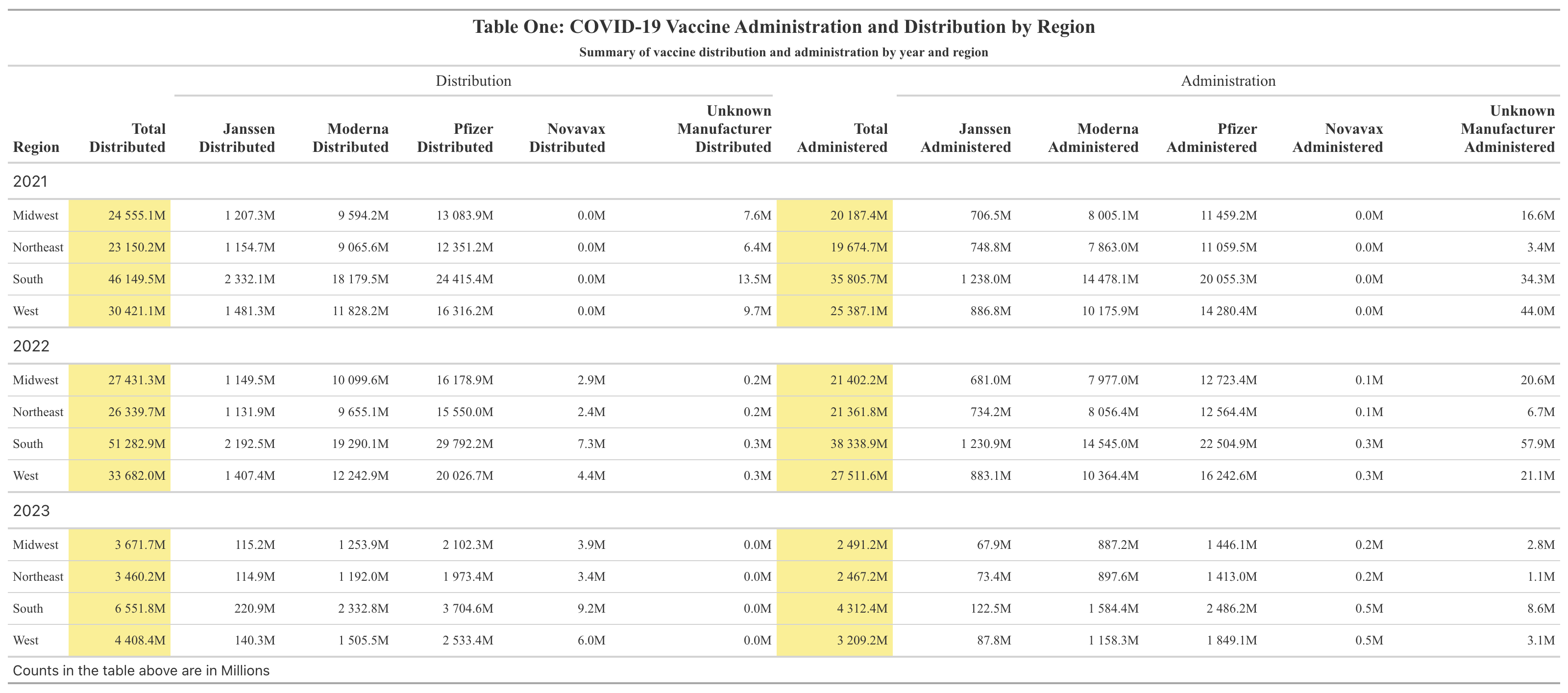


CDC’s Geographic Division Regions

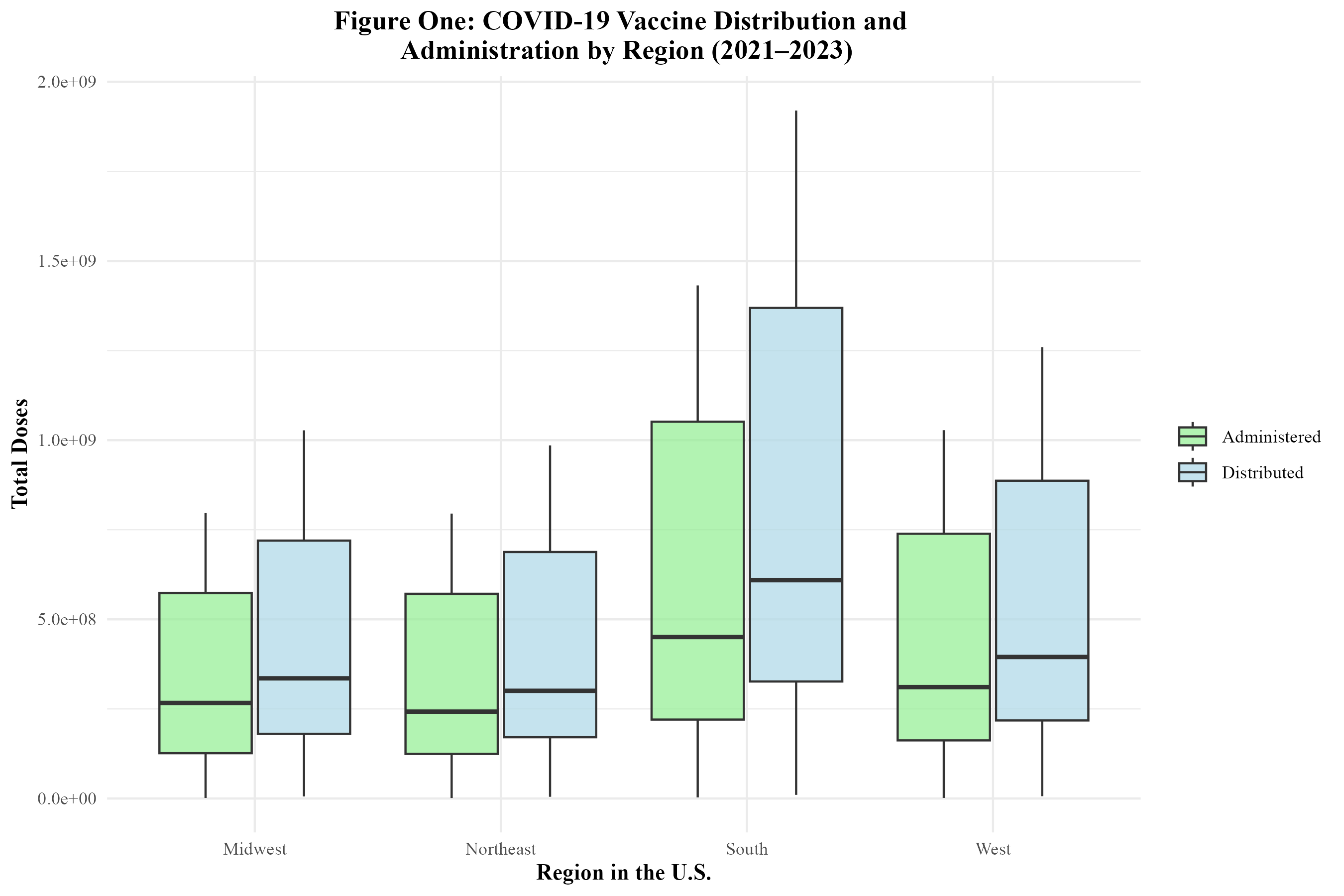
# 4. Results

## 4.1 Exploratory and Descriptive analysis

Table 1 displays the total counts of COVID-19 vaccines distributed and administered in 2021, 2022, and 2023 for each region of the United States. The mean total distributed doses of the vaccine for the United States as a whole between 2021 and 2023 was determined to be 5.667e+08; the mean total administered doses of the vaccine for the United States between 2021 and 2023 was 4.479e+08 (Supplement Table 1). There was a nearly perfect positive correlation of 0.99 between the number of total weekly doses distributed and administered (Supplement Figure 1). Pfizer had the highest number of doses distributed and administered, followed by Moderna, Janssen, and Novavax (Supplement Table 2). It should be noted that the Novavax vaccine was not made available until 2022 and that the Janssen vaccine was discontinued in 2023. Across all three years of interest, the South had both the highest number of COVID-19 vaccines distributed and administered. The West had the second highest number of vaccines distributed and administered. The Midwest had the third highest number of vaccines distributed and administered. The Northeast had the lowest number of vaccines distributed and administered. However, the Midwest and Northeast’s numbers were very similar.



In figure 1, the distribution of the COVID-19 vaccine by region between 2021 and 2023 is displayed. The South’s dominance in the average weekly number of COVID-19 vaccines distributed and administered is visually apparent. The West comes in second. The west actually has the highest correlation (0.9990827) between doses administered and distributed (Supplement Table 2). It can be seen that the Midwest and Northeast have similar numbers. Figure 1 also displays the administration of the COVID-19 vaccine by region between 2021 and 2023. It can be seen that the South also had the highest average weekly number of COVID-19 vaccines administered. The West comes in second. The Midwest and Northeast have similar numbers. The higher numbers of distributed and administered doses within the South is likely due to the fact that South is the most populous region within the U.S., with 128,716,192 residents in 2022 (10).



In figure 2, the distribution of the COVID-19 vaccine by manufacturer in the U.S. is displayed. The Pfizer vaccine had the highest number of doses distributed and administered, followed by Moderna, Janssen, and Novavax. Figure 2 also displays the administration of the COVID-19 vaccine by manufacturer in the U.S. The trends are similar to those of distribution. The Pfizer vaccine had the highest number of doses distributed and administered, followed by Moderna, Janssen, and Novavax.

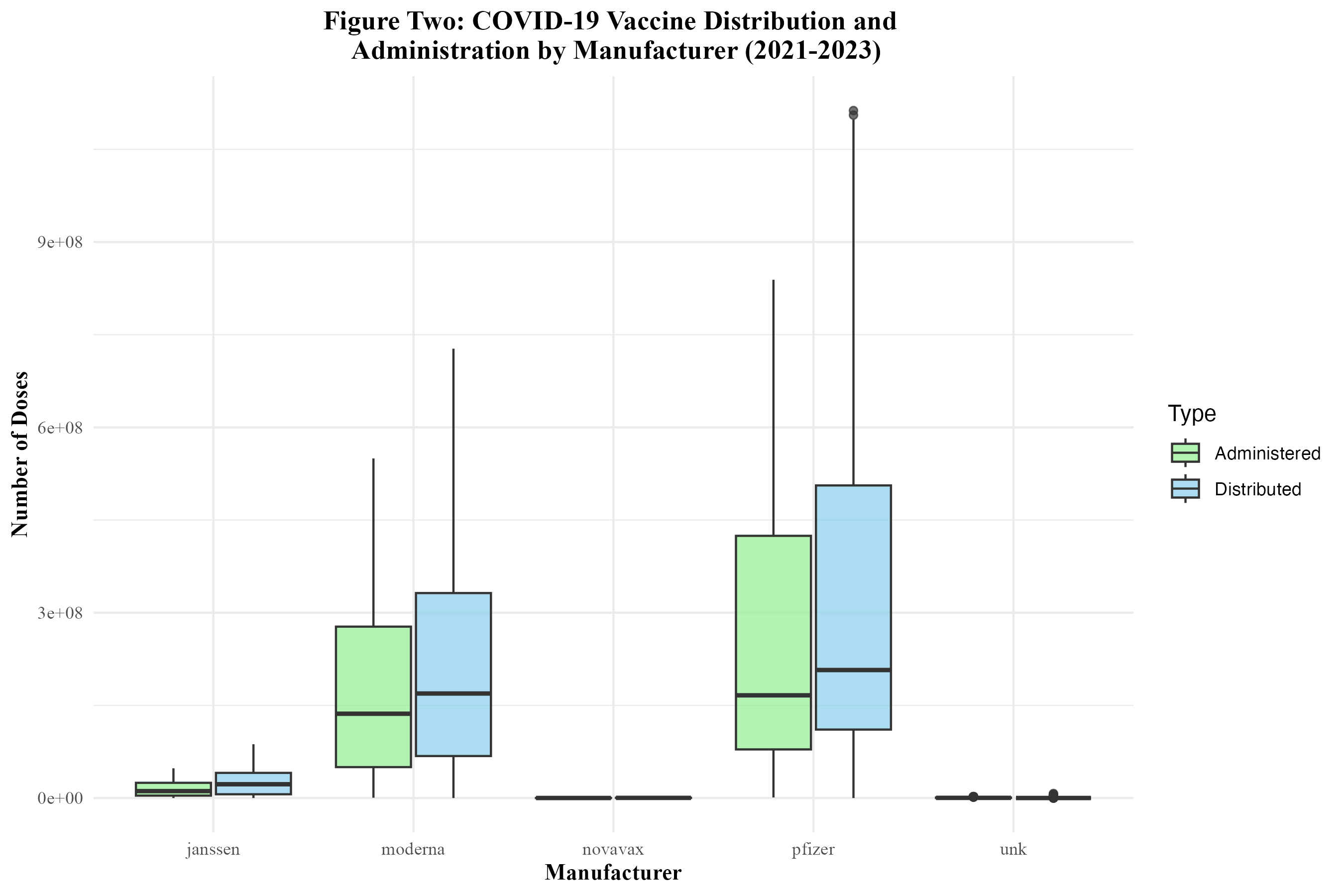


Figure Two: COVID-19 Vaccine Administration and Distribution by Manufacturer in the U.S.

Figure 3 displays the distribution of the COVID-19 vaccine by region over time. It can be seen the distribution of the vaccine rises between 2021 and 2022, but then begins to rapidly fall. All regions follow this trend. However, the doses distributed in the South appears to be significantly higher than the other regions. Between 2022 and 2023, there was an 87.2% decrease in doses distributed in the South; an 86.9% decrease in doses distributed in the West; an 86.6% decrease in doses distributed in the Midwest; and an 86.9% decrease in doses distributed in the Northeast (Supplement Table 3). Figure 4 displays the administration of the COVID-19 vaccine by region over time. The administration of the vaccine is also on the rise between 2021 and 2022, but then rapidly drop off as 2022 goes on. The each region’s trend is similar for both the distribution and administration of the vaccine. Between 2022 and 2023, there was an 88.8% decrease in doses administered in the South; an 88.3% decrease in doses administered in the West; an 88.4% decrease in doses administered in the Midwest; and an 88.5% decrease in doses administered in the Northeast (Supplement Table 3).

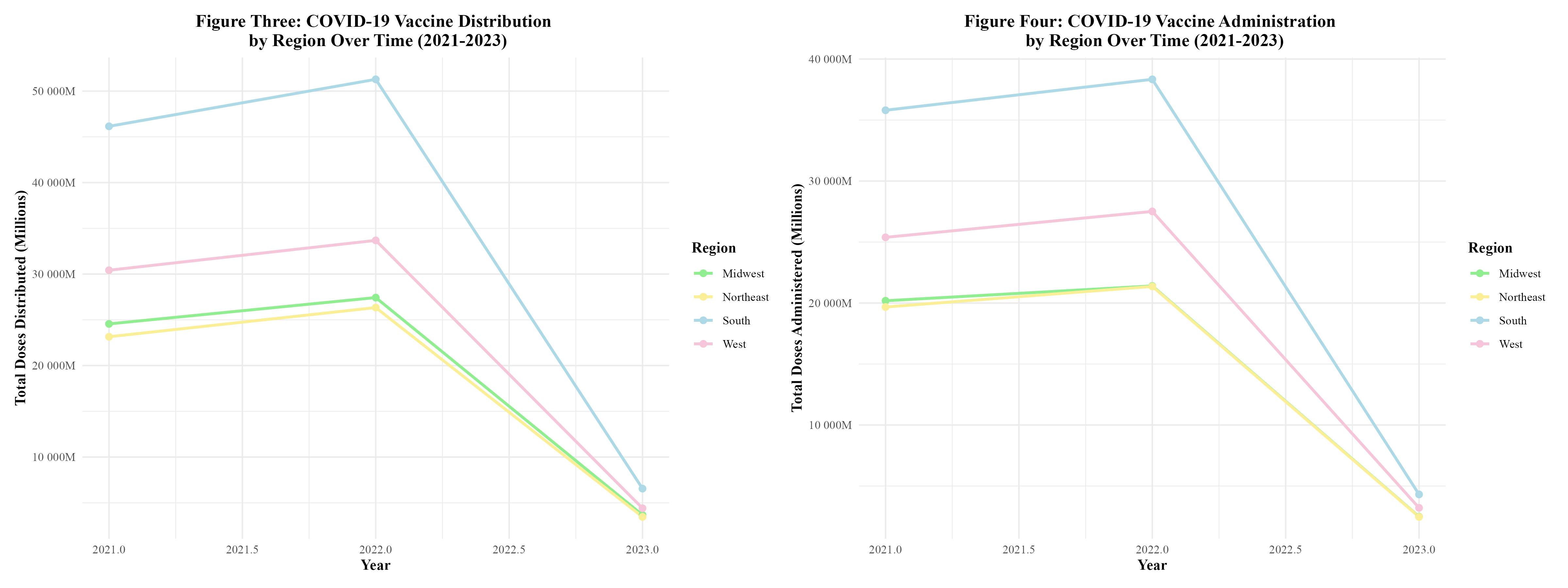


Figure 3 and 4: COVID-19 Vaccine Distribution and Administration by Region Over Time (2021-2023)

Figure 5 displays the distribution of the COVID-19 vaccine by manufacturer over time. It can be seen that the distribution of the vaccine rises between 2021 and 2022, but then begins to rapidly fall. All manufacturers follow this trend. This trend was observed in the distribution and administration of vaccines by region (figure 3) as well, which makes sense. Between 2022 and 2023, Janssen observed a 90.0% decrease in doses distributed; Moderna observed an 87.8% decrease; Pfizer observed an 87.4%; and Novavax observed a 32.4% decrease (Supplement Table 4.1). Novavax’s percent decrease in doses distributed is much lower due to the fact that this vaccine did not start rolling-out until 2022. By then, Moderna and Pfizer had been dominanting the distribution field. In figure 6, the administration of the COVID-19 vaccine by manufacturer over time can be seen. The administration of the vaccine is also on the rise between 2021 and 2022, but then rapidly drop off as 2022 goes on. This trend was also observed in the distribution and administration of vaccines by region (figure 4). In both figure 5 and 6, Pfizer and Moderna in the distribution and administration of the vaccine. This goes along with what was seen in figure 2. Between 2022 and 2023, Janssen observed a 90.0% decrease in doses administered; Moderna observed an 88.9% decrease; Pfizer observed an 88.8%; and Novavax observed a 97.2% decrease (Supplement Table 4.2).

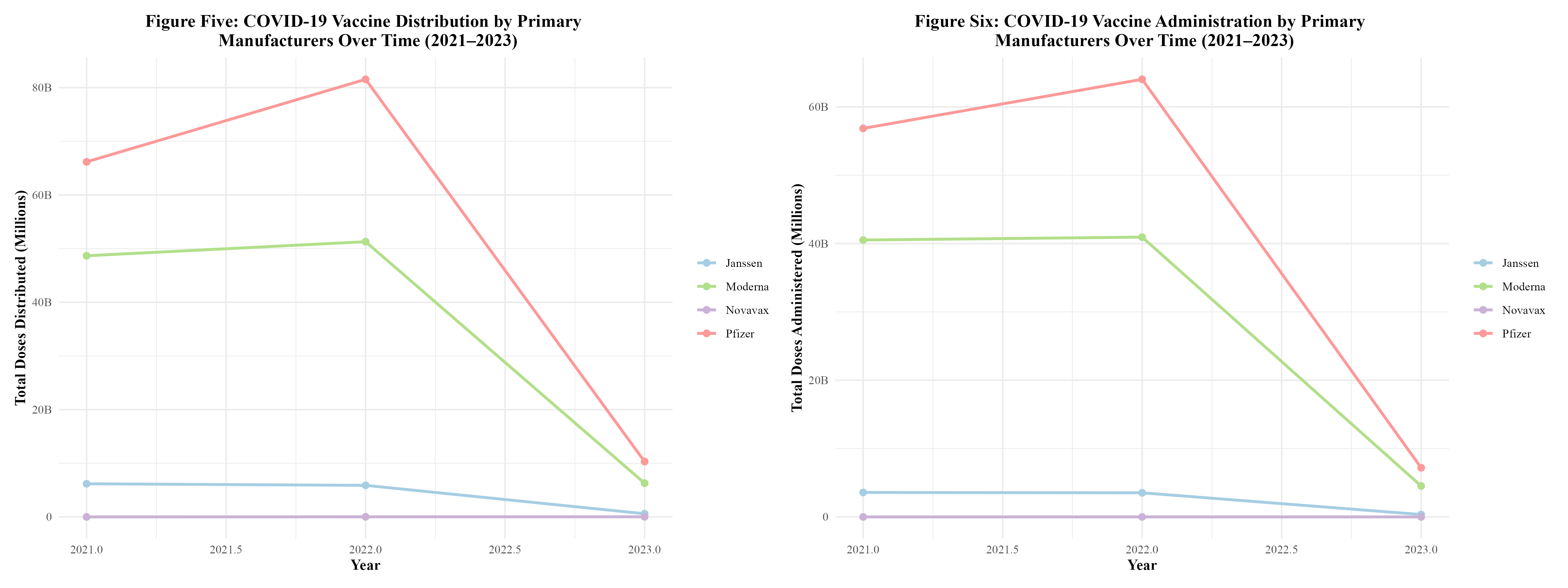
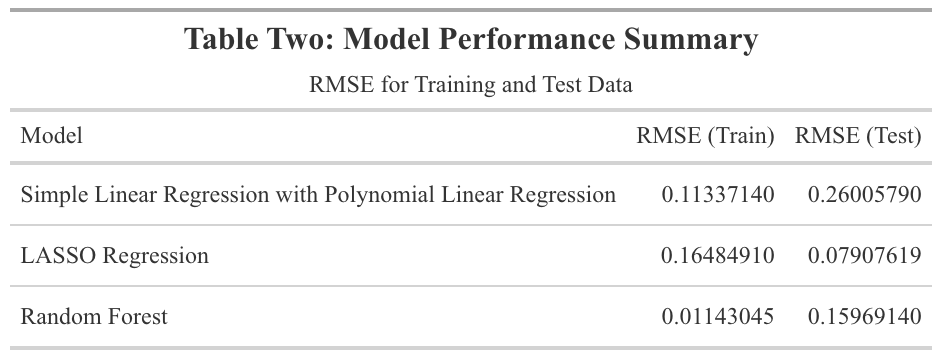


Figure 5 and 6: COVID-19 Vaccine Distribution and Administration by Primary Manufacturers Over Time (2021-2023)

## 4.2 Modeling

Three modeling methods were utilized to predict the proportion of Pfizer vaccines administered from several synthetic predictors and MMWR week. The synthetic predictors were: doses per 100k administered in the region during the MMWR week (based on the average United States population between 2021 and 2023), the average age of those vaccinated in the region that week, the COVID-19 vaccine hesitancy index for the region that week, the COVID-19 accessibility index for the region that week, and the proportion of vaccines administered to males in the region that week. First a null model was fit. The null model had an RMSE of 0.1693047 (Supplement Table 5). Table 2 displays the models that were used during the project’s analysis as well as the root mean square error (RMSE) when the model was fitted to both the train data and the test data. The RMSE value was obtained for both the train and test data. The Simple Linear Regression with Polynomial Linear Regression model had a Train RMSE of 0.11337140 and a Test RMSE of 0.26005790. The LASSO Regression model had a Train RMSE of 0.16484910 and a Test RMSE of 0.07907619. The Random Forest model had a Train RMSE of 0.01143045 and a Test RMSE of 0.15969140.



Below are the model residual plots. In figure 5, for the Simple Linear Regression Model (including the polynomials), The residuals appear to have a downward trend, indicating that the residuals decrease as the predicted values increase. This suggests that the model might be underestimating the values at higher predictions and overestimating at lower predictions. Figure 6 contains the LASSO Regression Model residual plot. The residuals are relatively concentrated around the predicted values, suggesting that the model’s predictions are fairly consistent within this range. The Random Forest Model residual plot, in figure 7, the downward trend in residuals suggests that the model might be overestimating at lower predicted values and underestimating at higher predicted values. Overall, none of the models performed well - indicating that the predictors used in this analysis are not good predictors of the proportion of Pfizer vaccines administered.

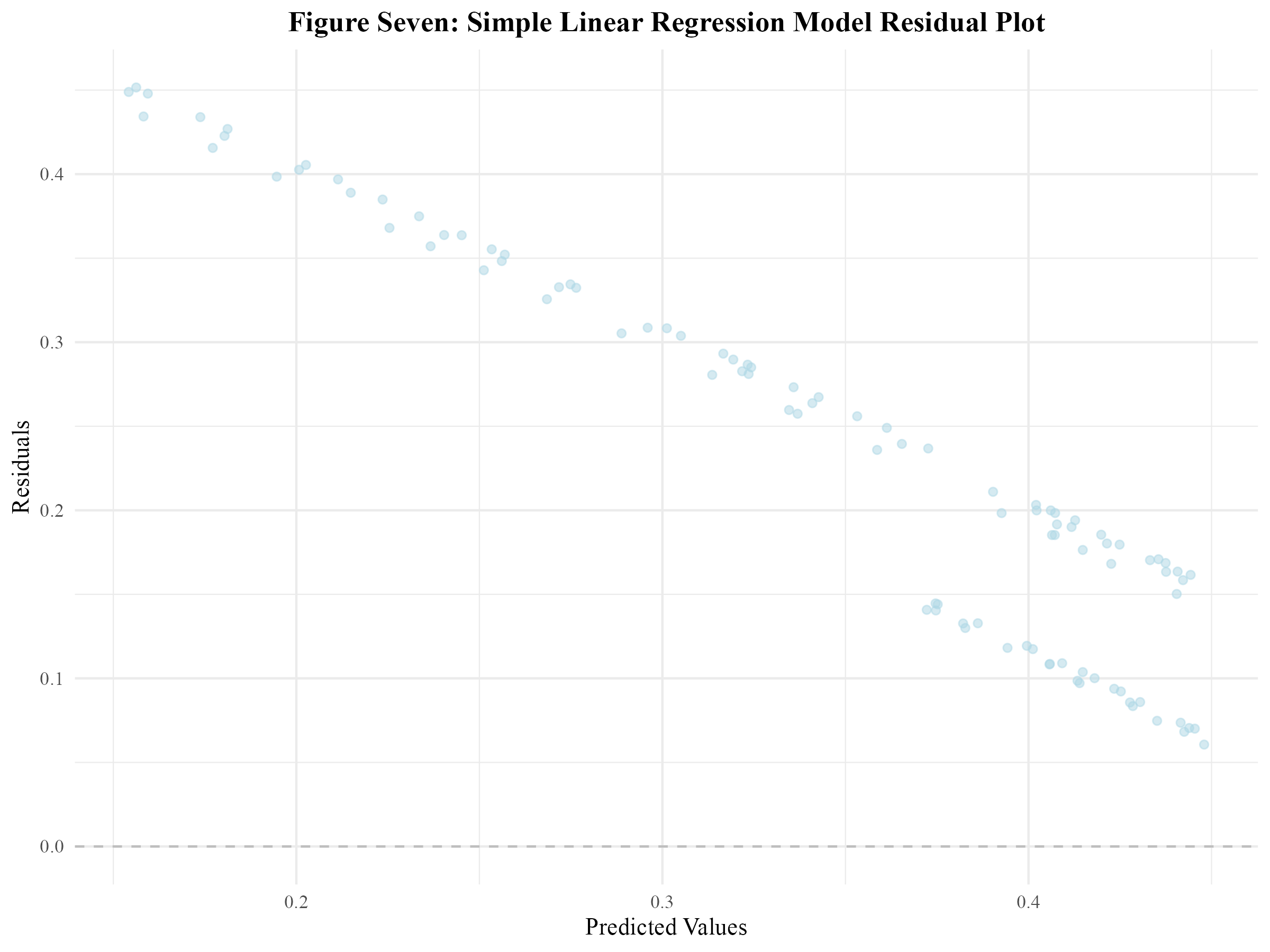


Figure Seven: Simple Linear Regression Model Residual Plot

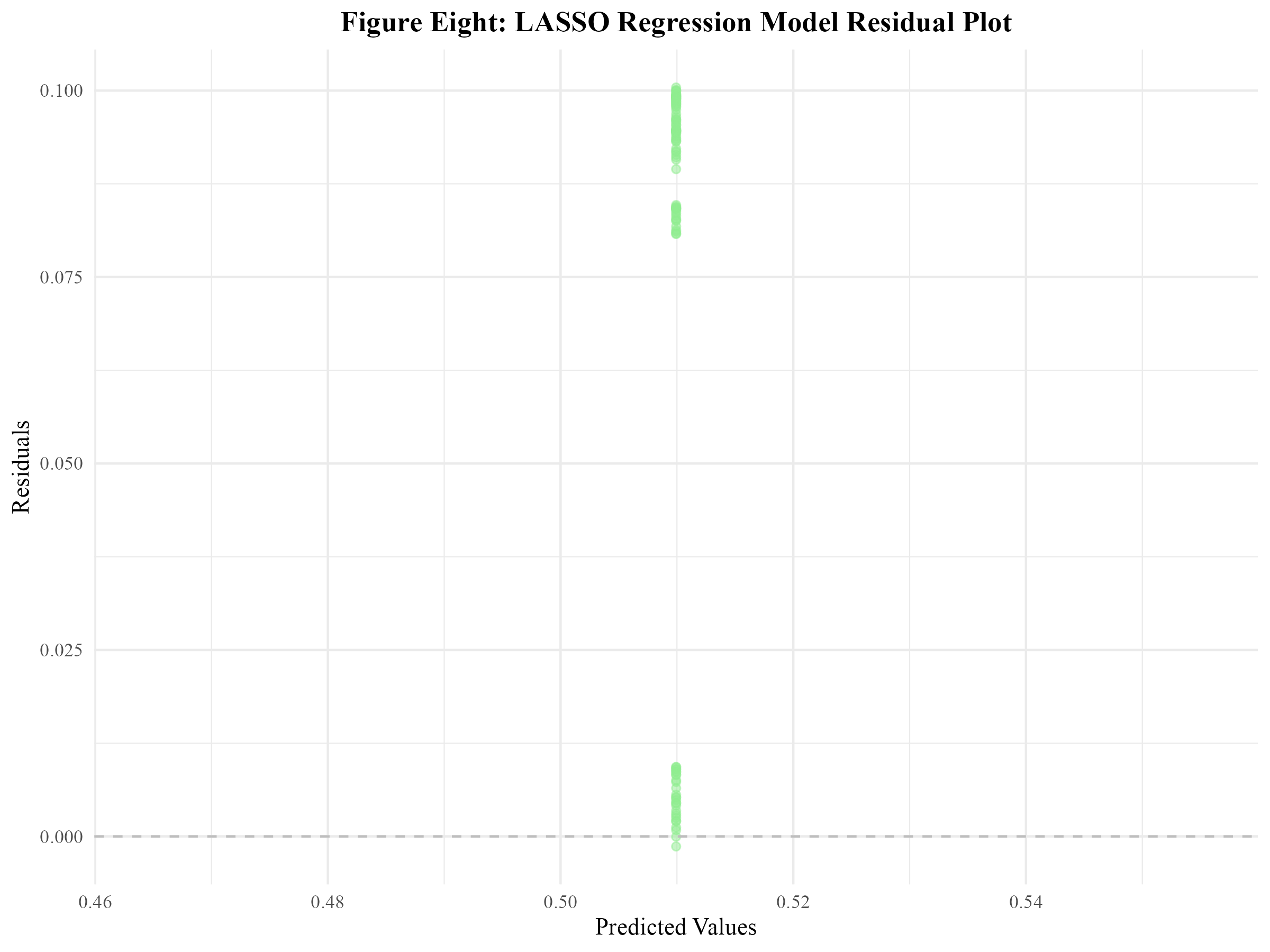


Figure Eight: LASSO Regression Model Residual Plot

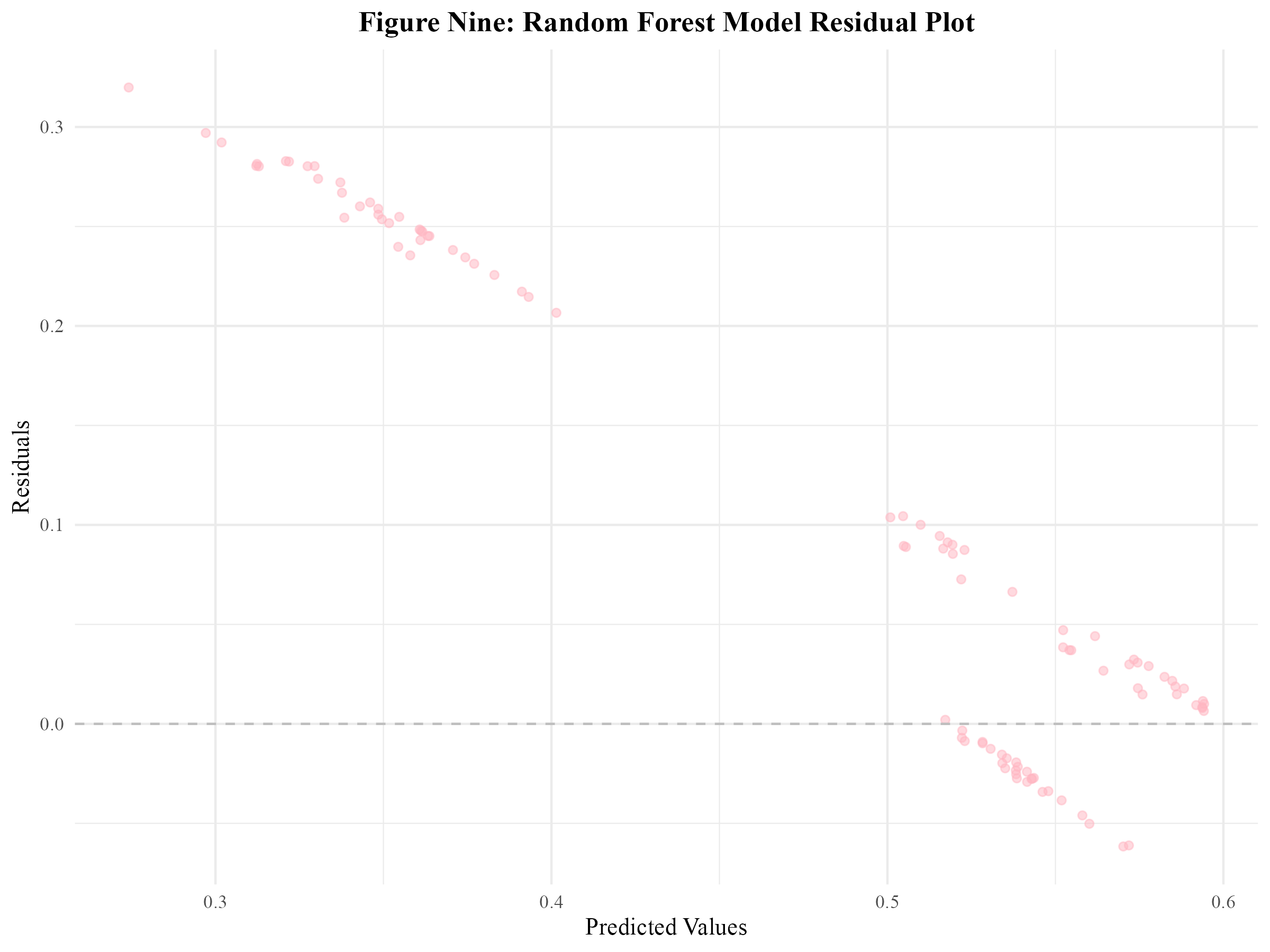


Figure Nine: Random Forest Model Residual Plot

# 5. Discussion

## 5.1 Summary and Interpretation

This project examined the distribution and administration of COVID-19 vaccines across regions of the United States from 2021 to 2023. Three key research questions were answered: (1) How does COVID-19 vaccine distribution and administration differ by region of the United States? Do certain manufacturers of the COVID-19 vaccine have a greater number of administered doses than others? (2) How has the vaccine distribution and administration by region in the United States changed with time? Has vaccine distribution and administration by manufacturer changed with time? (3) Can the proportion of Pfizer vaccines administered be predicted by several synthetic variables and MMWR week? The synthetic variables created for this model were: doses per 100k administered in the region during the MMWR week (based on the average United States population between 2021 and 2023), the average age of those vaccinated in the region that week, the COVID-19 vaccine hesitancy index for the region that week, the COVID-19 accessibility index for the region that week, and the proportion of vaccines administered to males in the region that week.

Descriptive analyses exhibited that Pfizer had the highest number of doses distributed and administered, followed by Moderna, Janssen, then Novavax between 2021 and 2023. It should be noted that the Novavax vaccine was not made available until 2022 and that the Janssen vaccine was discontinued in 2023. Furthermore, it was revealed that the Southern U.S. consistently had the highest number of both distributed and administered doses across 2021 to 2023, followed by the West, the Midwest, and lastly, the Northeast. This pattern is likely reflective of regional population differences, as the South is the most populous region within the U.S. Temporal analyses revealed that the distribution and administration of COVID-19 vaccines peaked between 2021 and 2022, followed by a sharp decline. This trend may be associated with public health campaigns, initial rollout efforts, and changes in public perception of the vaccine and the COVID-19 pandemic.

To assess whether the proportion of Pfizer doses administered could be predicted by the previously described synthetic variables and MMWR week, three models were evaluated: Linear Regression (including Polynomial and Interaction terms), LASSO Regression, and Random Forest. All three models had both low RMSEs for the train and test data. However, the residual plots revealed that the models were not well-fitted. This suggests that the synthetic variables used in this analysis were not strong predictors of the proportion of Pfizer vaccines administered.

## 5.2 Strengths and Limitations

A strength of this research project is that the data was obtained from the CDC, which is a reliable source. The data contained a large number of observations from all states within the U.S., which enhanced the representativeness of the study. The use of standardized temporal units (MMWR weeks) allowed for consistent analysis of distribution and administration trends over time.

However, there were a significant number of limitations to this project. First and foremost, there are always drawbacks to not collecting your own data. An issue that stemmed from this was that there were not enough predictor variables to use in the modeling analysis. Therefore, synthetic variables had to be created. These synthetic variables may not have fully reflected the influences on manufacturer-specific uptake (such as the proportion of Pfizer vaccines administered). Additionally, many distributed and administered doses were reported with an “unknown manufacturer”, which may have limited the ability to obtain accurate estimates for each manufacturer. Furthermore, the dataset was limited to the United States, which may limit the generalizability of the findings to other countries.

## 5.3 Conclusions

In conclunsion, this project revealed distinct regional and temporal trends were observed in both the distribution and administration of the COVID-19 vaccine across the United States from 2021 to 2023. However, attempts to predict the proportion of Pfizer vaccines administered using synthetic predictors and MMWR week were unsuccessful. Further research needs to be done to assess the factors the influence COVID-19 vaccine manufacturer specific uptake. The collection of open-ended qualitative data from the public would be of great benefit to obtaining a deeper understanding of these factors.

# 6. References

1. Coronavirus disease 2019 (COVID-19) - Symptoms and causes. (<https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963>). (Accessed February 21, 2025)

2. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol* [electronic article]. 2016;3(1):237–261. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5457962/>). (Accessed February 21, 2025)

3. CDC. COVID-19 Vaccine Basics. 2024;(<https://www.cdc.gov/covid/vaccines/how-they-work.html>). (Accessed February 21, 2025)

4. Comparing the COVID-19 Vaccines: How Are They Different? (<https://www.yalemedicine.org/news/covid-19-vaccine-comparison>). (Accessed February 21, 2025)

5. The Johnson & Johnson adenovirus vaccine explained. (<https://www.mayoclinic.org/johnson-johnson-adenovirus-vaccine-explained/vid-20510091>). (Accessed February 21, 2025)

6. Medicine NA of, National Academies of Sciences E, Division H and M, et al. Vaccine Distribution and Delivery. In: *Globally Resilient Supply Chains for Seasonal and Pandemic Influenza Vaccines*. National Academies Press (US); 2021 (Accessed April 18, 2025)(<https://www.ncbi.nlm.nih.gov/books/NBK580009/>). (Accessed April 18, 2025)

7. Muckstadt JA, Klein MG, Jackson PL, et al. Efficient and effective large-scale vaccine distribution. *International Journal of Production Economics* [electronic article]. 2023;262:108921. (<https://www.sciencedirect.com/science/article/pii/S0925527323001536>). (Accessed April 18, 2025)

8. Bertsimas D, Digalakis Jr V, Jacquillat A, et al. Where to locate COVID-19 mass vaccination facilities? *Naval Research Logistics (NRL)* [electronic article]. 2021;69(2):179–200. (<https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.22007>). (Accessed April 18, 2025)

9. Geographic division or region - Health, United States. 2024;(<https://www.cdc.gov/nchs/hus/sources-definitions/geographic-region.htm>). (Accessed April 17, 2025)

10. Bureau UC. Growth in U.S. Population Shows Early Indication of Recovery Amid COVID-19 Pandemic. (<https://www.census.gov/newsroom/press-releases/2022/2022-population-estimates.html>). (Accessed April 17, 2025)