Example Manuscript Template for a Data Analysis Project

Natalie Cann

# 1. Load Packages

**Authors**

* Natalie Cann

**Author affiliations**

1. Masters in Public Health Candidate, University of Georgia
2. B.S. Microbiology, Virginia Tech

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

# 2. Summary/Abstract

*Add in Summary/Abstract Once Study is Complete*

# 3. Introduction

## 3.1 General Background Information

*Not Complete*

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 illness is Influenza. Seasonal Influenza in humans is caused by influenza viruses that undergo antigenic drift and antigenic shift, resulting in novel flu viruses each year. Antigenic drift is known to occur when mutations develop within the surface proteins hemagglutinin (HA) and neuraminidase (NA) (1). These surface proteins serve as anitgens which are recognized by our immune systems, which will produce antibodies to recognize the surface protein antigens (1). Antigenic shift results when entirely new HA and/or NA proteins develop, creating a novel flu virus (1). According to the CDC, the most common month for a peak in flu cases to be observed is February; 17 of the past 40 flu seasons have reached peaks in February (2).

Another respiratory virus is COVID-19, also known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (3). 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 (3). 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 (4). Therefore, the spike protein is a major antigen of the COVID-19 virus recognized by the immune system.

Many individuals receive vaccinations to protect themselves from these respiratory illnesses. Flu vaccines are designed to protect individuals against the flu strains that are predicted to be most common during the upcoming flu season (5). The flu vaccine results in the production of antibodies within roughly two weeks after vaccination (5). The effectiveness of the flu vaccines varies with strains in circulation and the individual’s health status (5). The flu vaccine will be more beneficial during years which the vaccine matches the strains of that season (5).

There are two different types of COVID-19 vaccinations available in the United States. 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 (6). 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 (6). The Pfizer-BioNTech and Moderna vaccines are both mRNA vaccines (7). The Novavax vaccine is a protein subunit vaccine (7). The Janssen vaccine, also known as the Johnson and Johnson vaccine was disconintued (7). This vaccine utilized an adenvovirus as a vector (8). Adenovirus vaccines use an omnipresent virus that is altered so that it cannot cause illness or integrate into the host’s DNA (8). The adenovirus is engineered to contain a genes involved in making the spike protein COVID-19, which then leads to the immune system producing antibodies for the spike protein (8).

## 3.2 Description of data and data source

*Data Sources*

The datasets used in this analysis are both from the CDC. The first dataset, regarding COVID-19 vaccinations, contains information on the distribution and administration of COVID-19 vaccinations between 2020 and 2023. The second dataset, focusing on influenza vaccinations, contains information on the distribution of flu vaccines between 2018 and 2024. The influenza vaccination dataset is grouped by flu-season. More information on the datasets can be found in the “Methods” section under “Data Acquisition.”

## 3.3 Questions/Hypotheses to be addressed

I aim to assess if the administration of the COVID-19 vaccine is associated with its manufacturer. I will examine whether or not certain manufacturers had a greater number of administered doses. I will also assess whether or not the administration of the COVID-19 vaccine is associated with the administration of the flu vaccine by looking at the number of doses of the vaccines administered in the same year.

# 4. Methods

## 4.1 Schematic of Workflow for Analysis

*Will be completed upon finishing the analysis*

Along with descriptive analyses, I plan on using (simple linear) regression models to assess the potential relationship between the administration of the COVID-19 vaccine and its manufacturer as well as the potential relationship between the administration of the COVID-19 vaccine and the administration of the flu vaccine. I will peform a cross-correlation function to assess whether or not increases in Covid-19 vaccine administration are associated with increases in flu vaccine administration. I will use K-Means Clustering to assess patterns in vaccination trends across different time periods. This may be subject to change as we learn more in the course. Then I will polish up my graphs and tables.

Note: store in ‘assets’ folder; it is ok to add a schematic diagram/figure not created with code

Use Code Chunk below:

## 4.2 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. I downloaded the data directly from the CDC website provided above. This dataset was last updated May 12, 2023.

**Influenza Dataset:** https://data.cdc.gov/Vaccinations/Weekly-Cumulative-Doses-in-Millions-of-Influenza-V/k87d-gv3u/about\_data

This dataset also came from the CDC and contains information regarding the doses of flu vaccines in Millions distributed between 2018 and 2024. According to the CDC website, the dataset was provided by the National Center for Immunization and Respiratory Diseases (NCIRD). This dataset was last updated March 29, 2024.

## 4.3 Data Import and Cleaning Process

Below, I will import the two datasets via read\_csv() function from the readr package.

covid\_vaccine <- read\_csv(here("data", "raw-data", "COVID-19\_Vaccinations\_in\_the\_United\_States\_Jurisdiction\_20250208.csv"))

Rows: 38488 Columns: 109  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (2): Date, Location  
dbl (107): MMWR\_week, Distributed, Distributed\_Janssen, Distributed\_Moderna,...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

flu\_vaccine <- read\_csv(here("data", "raw-data", "Weekly\_Cumulative\_Doses\_\_in\_Millions\_\_of\_Influenza\_Vaccine\_Distributed\_by\_Flu\_Season\_in\_the\_United\_States\_20250208.csv"))

Rows: 193 Columns: 9  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (6): Influenza\_Season, Start\_Date, End\_Date, Reporting\_Timeframe, Curren...  
dbl (3): Week, Week\_Sort\_Order, Cumulative\_Flu\_Doses\_Distributed  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

I will provide a view of the first rows as well as the dimensions of the datasets.

# Covid Dataset  
head(covid\_vaccine) #viewing first rows

# A tibble: 6 × 109  
 Date MMWR\_week Location Distributed Distributed\_Janssen Distributed\_Moderna  
 <chr> <dbl> <chr> <dbl> <dbl> <dbl>  
1 05/10/… 19 NE 5481710 152400 1647380  
2 05/10/… 19 LA 10282120 330500 3807980  
3 05/10/… 19 GA 28727475 869100 9763000  
4 05/10/… 19 WY 1281755 49300 490040  
5 05/10/… 19 CO 17769135 501900 5402640  
6 05/10/… 19 PA 42895735 1569200 13941120  
# ℹ 103 more variables: Distributed\_Pfizer <dbl>, Distributed\_Novavax <dbl>,  
# Distributed\_Unk\_Manuf <dbl>, Dist\_Per\_100K <dbl>,  
# Distributed\_Per\_100k\_5Plus <dbl>, Distributed\_Per\_100k\_12Plus <dbl>,  
# Distributed\_Per\_100k\_18Plus <dbl>, Distributed\_Per\_100k\_65Plus <dbl>,  
# Administered <dbl>, Administered\_5Plus <dbl>, Administered\_12Plus <dbl>,  
# Administered\_18Plus <dbl>, Administered\_65Plus <dbl>,  
# Administered\_Janssen <dbl>, Administered\_Moderna <dbl>, …

dim(covid\_vaccine) #viewing dimensions

[1] 38488 109

# Influenza Dataset  
head(flu\_vaccine) #viewing first rows

# A tibble: 6 × 9  
 Influenza\_Season Start\_Date End\_Date Reporting\_Timeframe Week Week\_Sort\_Order  
 <chr> <chr> <chr> <chr> <dbl> <dbl>  
1 2018-2019 07/29/2018 08/04/2… 2018-07-29 - 2018-… 30 1  
2 2018-2019 08/05/2018 08/11/2… 2018-08-05 - 2018-… 31 2  
3 2018-2019 08/12/2018 08/18/2… 2018-08-12 - 2018-… 32 3  
4 2018-2019 08/19/2018 08/25/2… 2018-08-19 - 2018-… 33 4  
5 2018-2019 08/26/2018 09/01/2… 2018-08-26 - 2018-… 34 5  
6 2018-2019 09/02/2018 09/08/2… 2018-09-02 - 2018-… 35 6  
# ℹ 3 more variables: Current\_Season\_Week\_Ending\_Label <chr>,  
# Cumulative\_Flu\_Doses\_Distributed <dbl>, Current\_Through <chr>

dim(flu\_vaccine) #viewing dimensions

[1] 193 9

The COVID\_vaccine dataset contains 109 variables and 38,488 observations. I am specifically interested in the variables that contain the counts of administered COVID-19 vaccines by manufacturer (Administered\_Janssen, Administered\_Moderna, Administered\_Pfizer, Administered\_Novavax, and Administered\_Unk\_Manuf) as well as the variable “date” (day that the vaccine data was reported).

The flu\_vaccine dataset contains 9 variables and 193 observations (considerably less than the COVID\_vaccine dataset). The variables of interest from this dataset include: year (Influenza\_Season) and the cumulative doses distributed (Cumulative\_Flu\_Doses\_Distributed; Note: this is in millions).

I performed the following data processing and cleaning steps for the COVID-19 vaccine dataset: - Defined variables - Assessed missingness of data and variable classes - Used the lubridate package to fix the date format - Select variables of interest to work with for analysis - Create separate datasets for 2021, 2022, and 2023 - Create a new variable that calculates the cumulative doses distributed and administered for each manufacturer - Save final datasets as RDS files

I performed the following data processing and cleaning steps for the influenza vaccine dataset: - Defined variables - Assessed missingness of data and variable classes - Used the lubridate package to fix the date format - Create a new variable that contains the number of flu vaccine doses administered per week - Create two new variables that display the number of doses of the flu vaccine administered in millions (one variable for cumulative doses and the other for weekly doses) - Create separate datasets for 2021, 2022, and 2023 - Save final datasets as RDS files

## 4.4 Statistical analysis

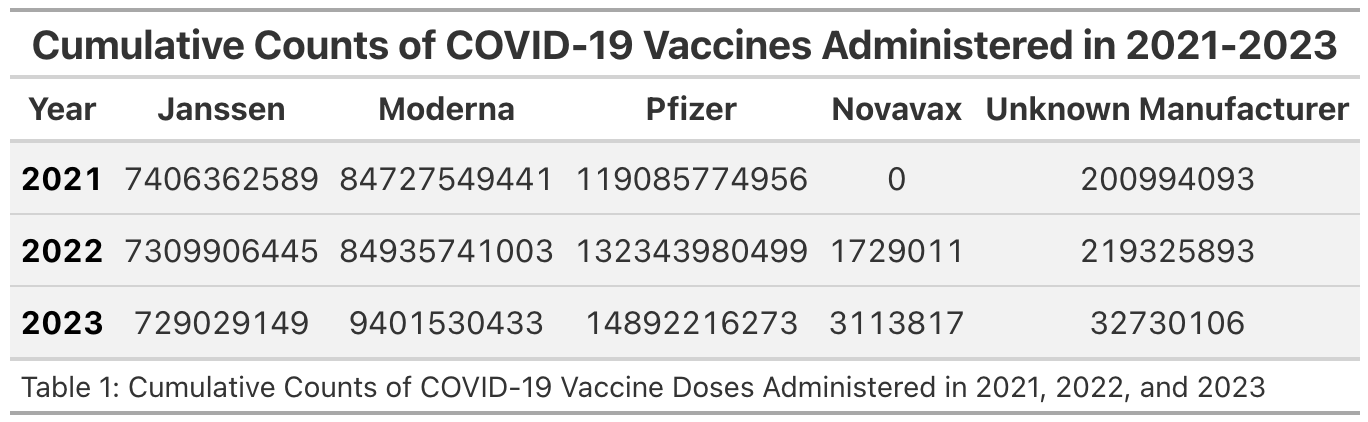
*Not Fully Complete Yet* I will perform simple linear regressions, k-means clustering, and cross-correlation functions to assess the relationships between the administration of the COVID-19 vaccine and its manufacturer as well as the administration of the flu vaccine. I will change this section as more is covered in the course.

# 5. Results

## 5.1 Exploratory/Descriptive analysis

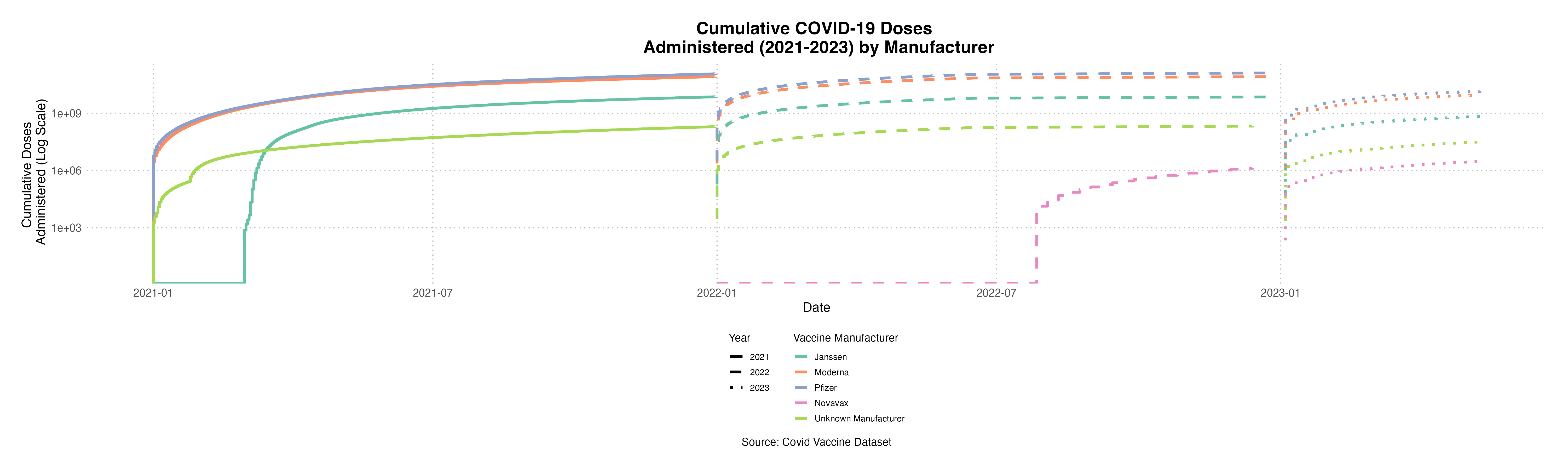
*Not complete yet - I have not yet polished up my graphs and tables*

The table below displays the cumulative counts of flu vaccine doses administered in 2021, 2022, and 2023.

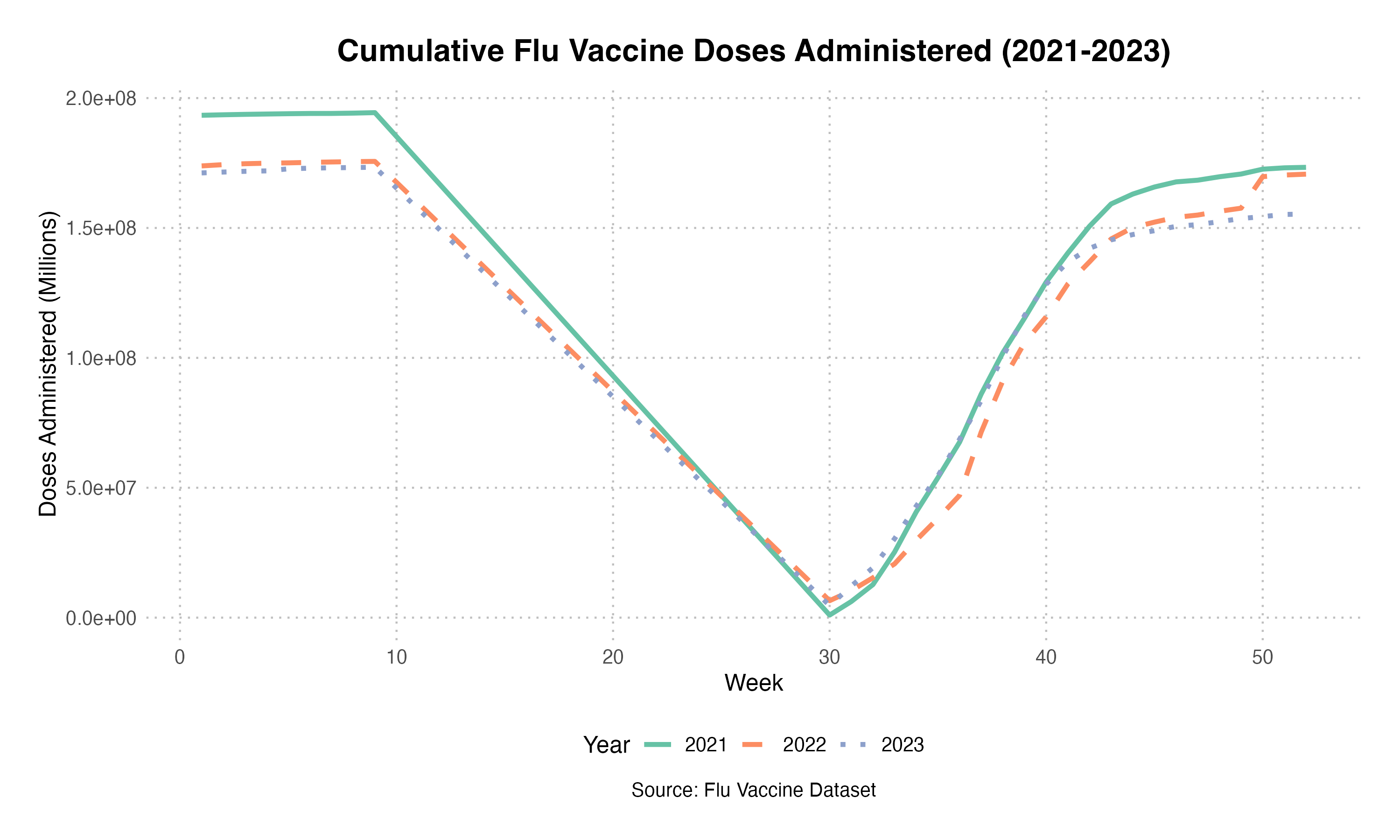


As seen in the table above, each year Pfizer had the greatest number of vaccines administered. Followed by Moderna, Janssen, Unknown Manufacturer, and Novavax. It is important to note that Novavax was not available in 2021; it became available in late 2022.

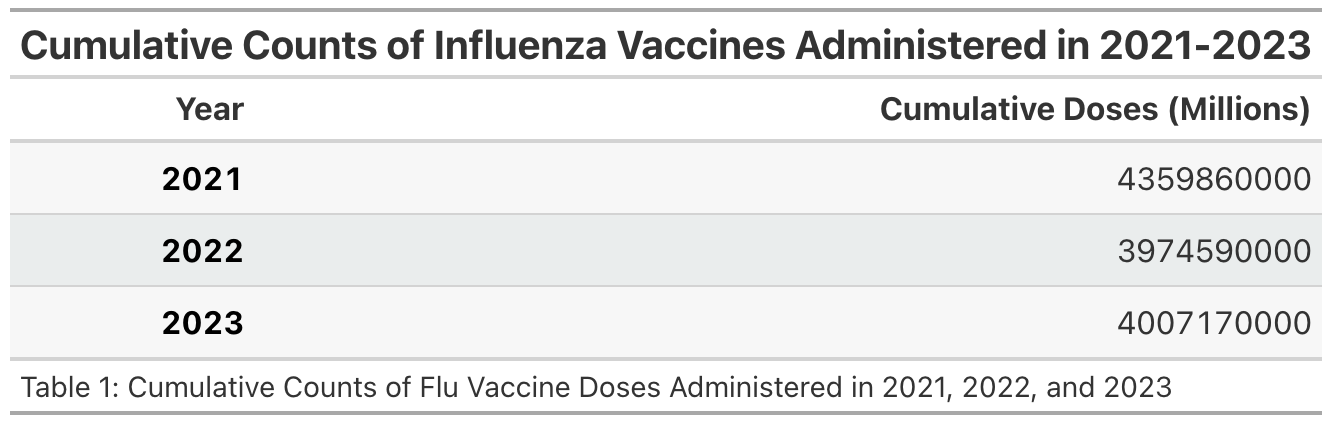
The plot below displays the cumulative counts of COVID-19 vaccine doses administered in 2021, 2022, and 2023 on a logarithmic scale for better visibility.



The plot below displays the cumulative counts of flu vaccine doses administered in 2021, 2022, and 2023.



The table below displays the cumulative counts of flu vaccine doses administered in 2021, 2022, and 2023.



## 5.2 Basic statistical analysis

*Not completed yet - simple statistics (simple models with 1 predictor); look for associations b/w outcomes and each individual predictor variable; unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

Use Code Chunk below:

## 5.3 Full analysis

*Not completed yet - use one or several statistical/machine learning methods to analyze your data and make figures/tables*

Use Code Chunk below:

# 6. Discussion

## 6.1 Summary and Interpretation

*Not completed yet*

## 6.2 Strengths and Limitations

*Not completed yet*

## 6.3 Conclusions

*Not completed yet - include main take home messages*

# 7. References

Note: reference formatting is determined by the CSL file specified in the YAML header

1. CDC. How Flu Viruses Can Change: "Drift" and "Shift". 2024;(<https://www.cdc.gov/flu/php/viruses/change.html>). (Accessed February 21, 2025)

2. CDC. Flu Season. 2024;(<https://www.cdc.gov/flu/about/season.html>). (Accessed February 21, 2025)

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

4. 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)

5. CDC. Key Facts About Seasonal Flu Vaccine. 2025;(<https://www.cdc.gov/flu/vaccines/keyfacts.html>). (Accessed February 21, 2025)

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

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

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