DS 804: Communication of Data
Final Project: An Analysis on Covid-19
Shawn Bedard, Srija Gandhesree, Vishwas Sabbani, Sharmishta Tallapally

---

Introduction

Data visualization has recently become a major tool for data scientists and has allowed for data to be represented in a simple yet effective manner. Over the course of two years, the covid-19 epidemic has played a crucial role in affecting global economies, increases in hate crimes, and destroying social interactions. In this study, the goal is to visually represent the timeline of covid and its effects using a data set of the United States covid statistics and Covid-19 vaccine manufacturer information to better understand the causes and effects of the global pandemic.

In this analysis, we are also trying to answer our research question, which is now that Covid-19 vaccines are widely available, what are the public perceptions regarding vaccination? To better understand how public sentiment has evolved since vaccinations have become widely available, we scraped twitter for specific tweets regarding the Covid-19 vaccine from 11/1/2021 to 12/7/2021. The variables collected are a collection of the keywords that were searched, which resulted in twelve variables with one thousand tweets per keyword. To determine whether sentiment has changed over time, our results are being compared to a previous study titled, *COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA* by Sattar & Arifuzzaman (2021).

Methodology

To achieve our goal in our analysis, our methodology consisted of gathering and collecting data, being able to properly clean the datasets and finally understanding the data gathered to create a dashboard with visualizations that clearly explain our dataset.

Beginning by cleaning the data, we collected three datasets that were in csv formats regarding the United States covid statistics and information on the vaccine manufacturers. The data was already in a tidy state and resulted in some simple cleaning of dates, texts, and the running totals. The data cleaning tool that was used was excel as well as Power Query since the amount of data was not too large and the cleansing that needed to be done was simple. For our dataset involving tweets regarding the pandemic, the data collection was done through R Studio which allowed us to use numerous libraries such as Rtweets to collect the data, tidyverse for data cleansing, and formatting libraries such as syuzhet, textstem, and SnowballC. The data was first collected by the key word we specifically intended to search and then by grouping by the general variable we wanted information on. Once we had gathered datasets of tweets for each of our key variables, we then combined all the variables into one large dataset in order to begin cleaning and preparing the data for analysis. The cleaning process was completed in R Studio due to the complexity of the cleaning that needed to be completed as well as the size of the dataset.

The next step that needed to be completed before moving on was to understand each dataset was connected to each other and to better understand the variables that we have gathered. Before going straight into visualizations, we took the time to comprehend the data and come up with possible visualizations that

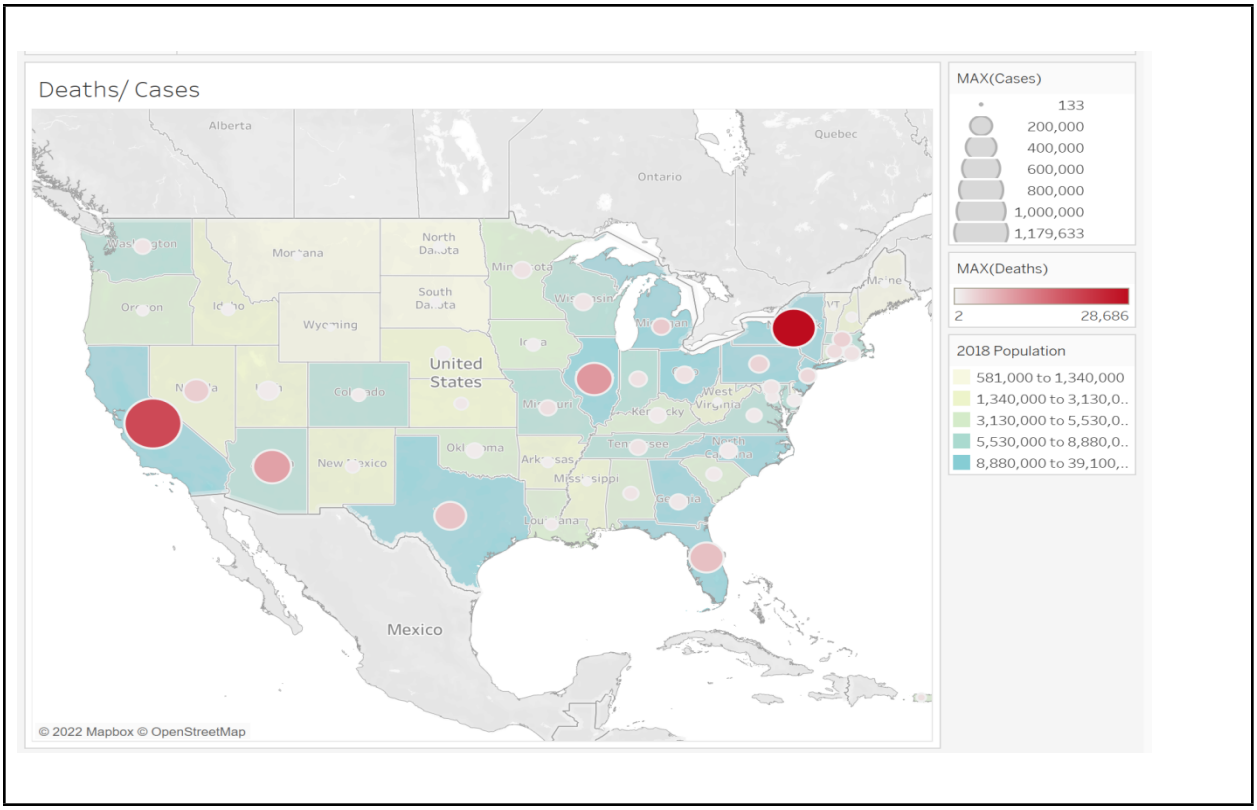
we can tie together in order to explain our analysis. This was a point in our analysis where, as a group, we decided on which direction to lean towards for our dashboard and gave us more confidence when moving forward with the creation of the visualizations.

The construction of the Covid-19 dashboard began with the creation of the visualizations separately in order to allow us to have numerous options and allow us to pick the most related and valuable graphs that would eventually end on our dashboard. The construction of the graphs revolve around locating where cases and deaths are occurring across the United States, how did the vaccinations affect the trend of positive cases, which vaccines were most popular by country, and others that help tell the story of overall pandemic.

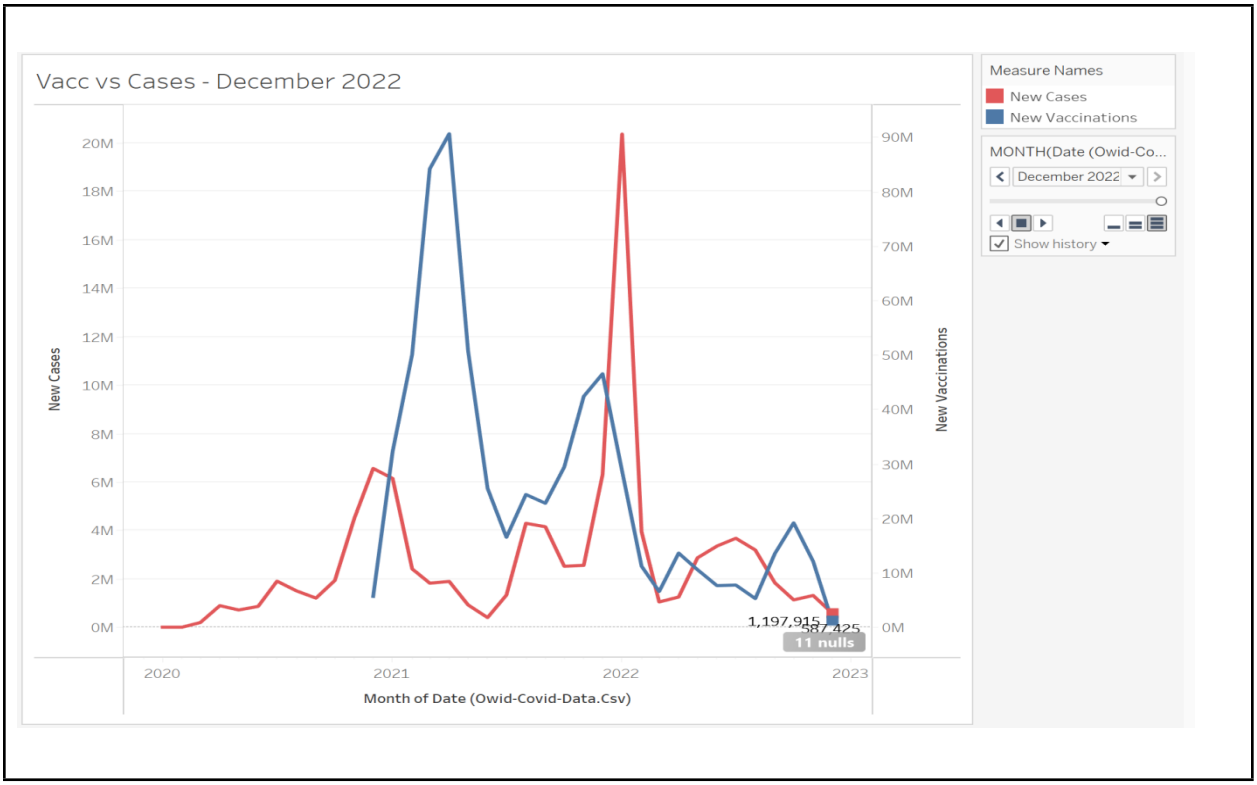
Covid-19 Analysis:

To begin our analysis, the first visualization created was intended to display the total number of positive Covid-19 cases across the United States. This visualization was extremely useful to understand which states were hit the hardest and where most of the deaths were occuring. Over the course of the pandemic, a majority of positive cases appear in major U.S. hotspots that have very large populations. According to the map, California and New York were the two worst states then followed by other largely populated states such as Illonois, Florida, Texas, and Arizona. This visualization shows how disproportionate the virus has affected individual states since most of the middle and southern states clearly have much less cases but population size does play a factor.

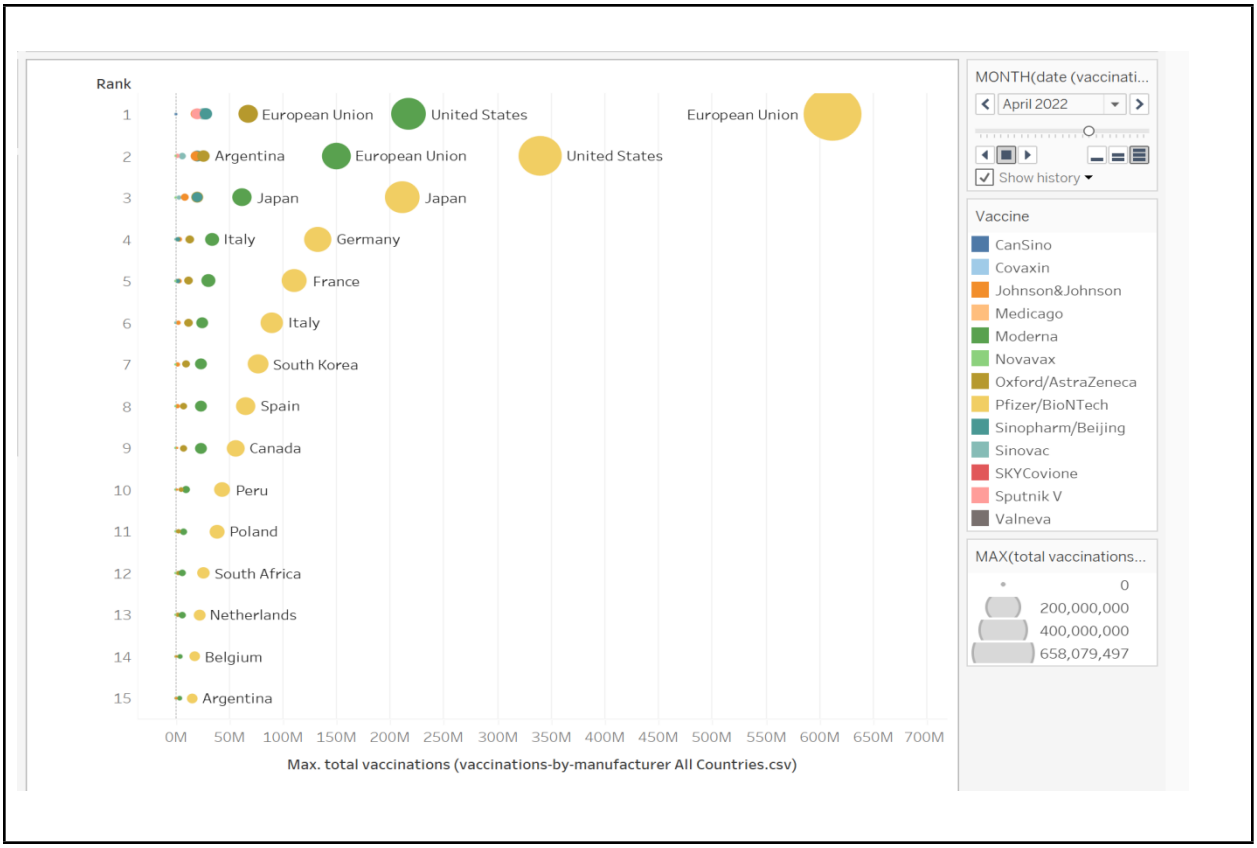
The second aspect of this visualization is that we represent the number of deaths as a color scale of red where darker colors represent a larger number. It is expected that the locations with the most cases would also have a greater death toll which is true in our analysis. California and New York both have a much higher death toll than any other state in the United States, while Nevada has a large number of deaths especially for the size of their population. Along the same line as the number of cases, the number of deaths seem to be focused in the northrn east coast while the middle and northern west coast experienced very little deaths.



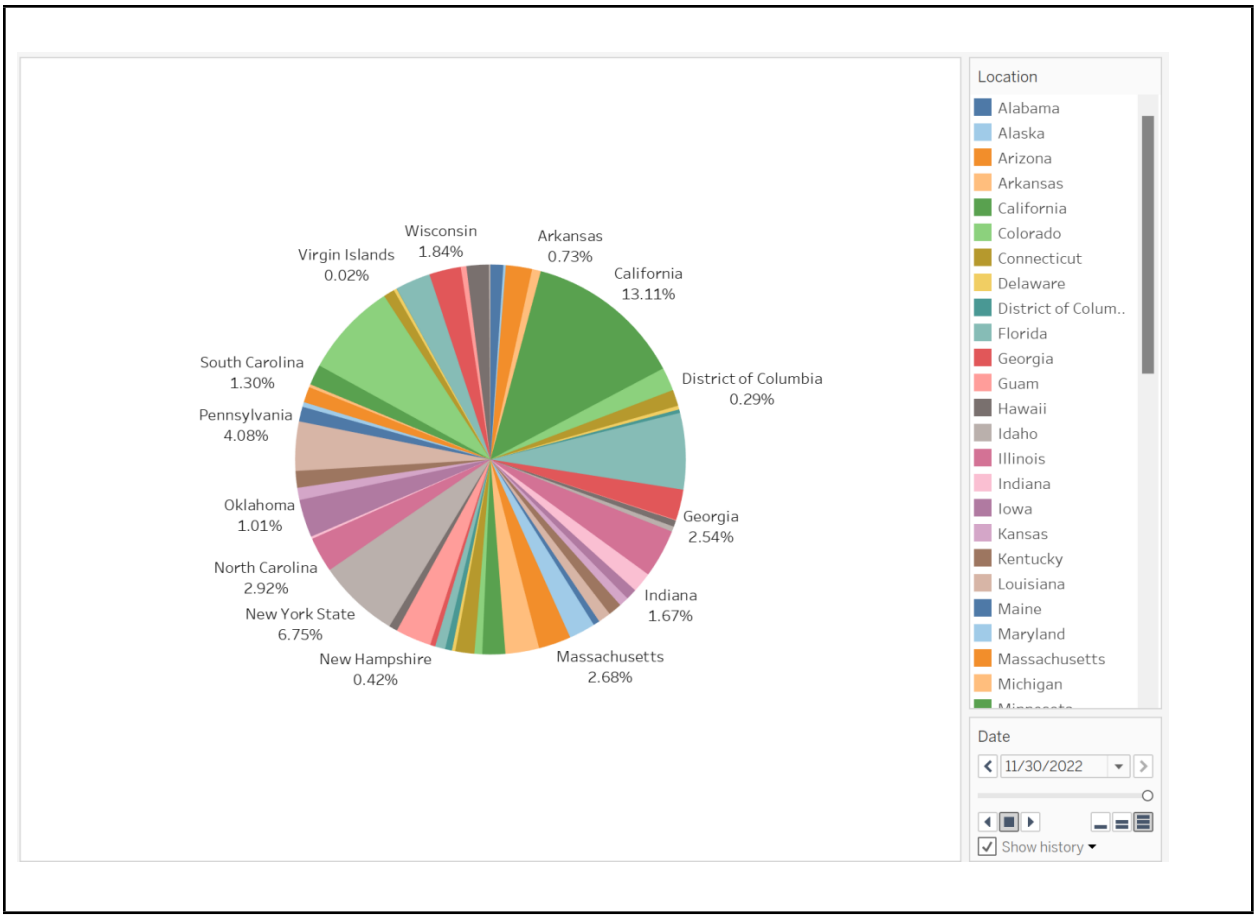
The next chart visualizes the trends of the number of positive cases compared to the trend of new vaccinations occurring from 2020 to 2023. As we can visually see in the Vacc vs Cases chart, the number of positive cases slowly began to increase as it was still largely undetected and no vaccine has been created yet. At the end of 2020, we see a large spike in the number of cases jumping from two million to over six million. As the Covid-19 vaccine gets introduced the number of vaccinations greatly increase as people rush to be vaccinated and at the same time the number of cases begins to drop mostly due to the introduction of social distancing as well as the increase in vaccinated people. As the main rush of vaccinations slows down in mid 2021, we can see the number of new cases remaining constant, slightly increasing and decreasing until the end of 2021 where we see the largest spike of new cases throughout the entire dataset. This can be explained by the relaxing of social distancing laws and the holiday season which caused more people to leave their households and socialize amongst the population. As the pandemic slowly begins to end in 2022 we can see rises in cases as vaccination rates decline which is met with a spike of vaccinations bringing the number of cases back down.



The third visualization created involved determining which type of vaccine was most used in countries across the globe. The data is broken down having each country aligned with their ranking and the points on the chart show the number of vaccines by the size and which manufacturer by the color. According to the visualization, it is clear across all rankings that Pfizer was the favorite or at least more accessible of a vaccine for many if not all countries compared to any other manufacturer. The second most used vaccination was created by Moderna and is a clear cut second when comparing vaccine types. From the beginning of our time period, both of the initial vaccines to become available to the public were obviously more prepared to become the dominant vaccines on the market and would not allow for any manufacturer to create any sort of market share in any of the global countries.

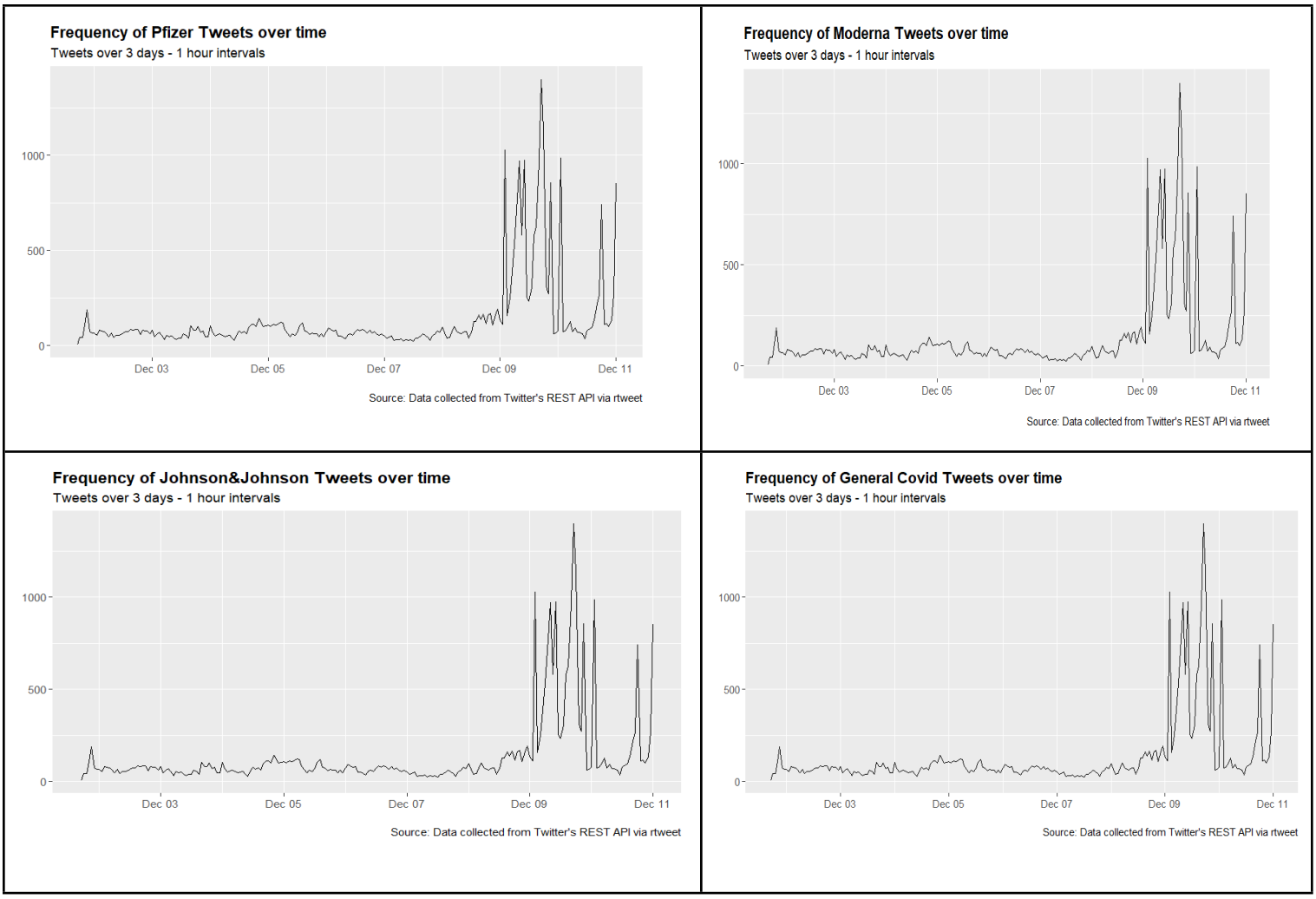


Knowing how Covid-19 vaccinations affected the overall outcome of the pandemic, we decided to create a visualization of the number of vaccinations by each state over the course of the covid time period. At the end of our dataset timeline, we can see the major outliers from the country and the most prominent states appear to be California with 13% of the country's population followed by Texas, Florida and New York. This would further harden our initial theory that higher populated states which have a higher number of cases will ultimately result in having a high number of vaccinations given out. From the beginning of 2020, the states with the largest percent of the country's vaccinations remained at the top through the entire pandemic with no other states with much smaller populations coming near the top percentage of vaccinations.

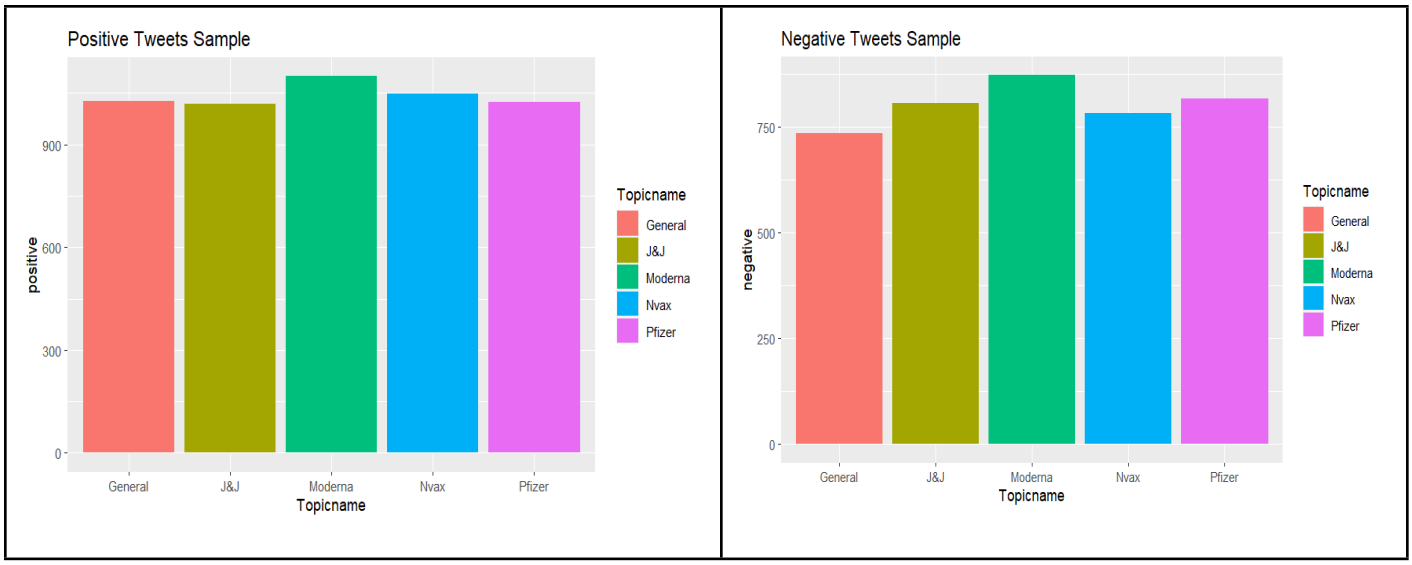


Twitter Sentiment Analysis:

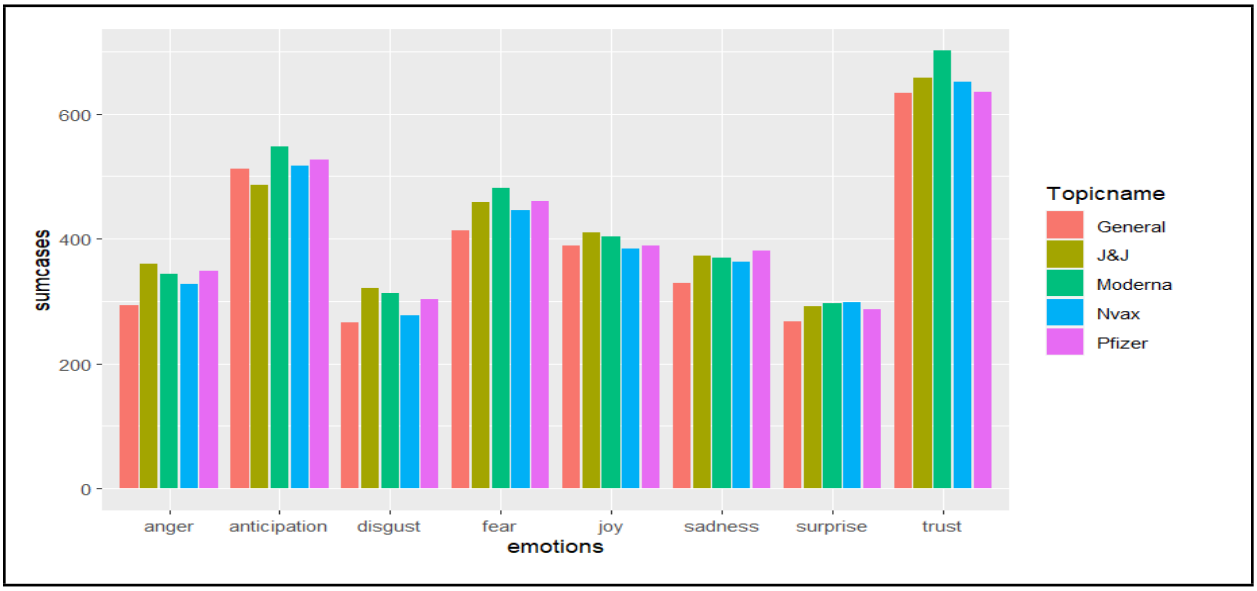
To begin our understanding of the twitter dataset, we began by comparing the frequency of tweets about different vaccine different covid-19 manufacturers and the number of general covid tweets over a period of time. The goal of this analysis, alongside comparing our results to a previous study, is to understand when people are tweeting about a certain topic and to match trends with real life events. According to the visualizations below, the number of tweets before December 9th were somewhat constant only increasing or decreasing slightly on certain days but following the 9th there is a clear spike in the number of tweets which correlate with the timing of the FDA approval of the Pfizer vaccination. After the announcement of the vaccine, it is clear that people almost immediately began expressing their opinions on the topic as it was becoming a major topic in the news, social events, and even in the government. The visualizations help demonstrate the public concerns regarding the vaccines whether it be negative or positive and can be used to determine the point in time where viewpoints began to divide the US population.



The following text analysis that was performed was to understand the intentions behind the twitter data that was collected in order to see which vaccinations were being talked about the most in a positive or negative form. From our two sentiment visualizations, we can determine that moderna was the most talked about vaccine receiving the highest amount of positive and negative tweets. When comparing the other vaccine manufacturers, we can see that they have roughly an even amount of positive tweets but when looking at the negative sentiment chart we can see that NVAX had the lowest amount of negative tweets followed by Johnson and Johnson with the second least. When looking at the general column that contains the other vaccinations can be seen having a large number of positive tweets but much less negative tweets as they were not as popular around the world. The overall sentiment of the dataset can be seen as positive since we can see the number of positive tweets collected contain around 200 tweets more across vaccine brands compared to the amount of negative tweets

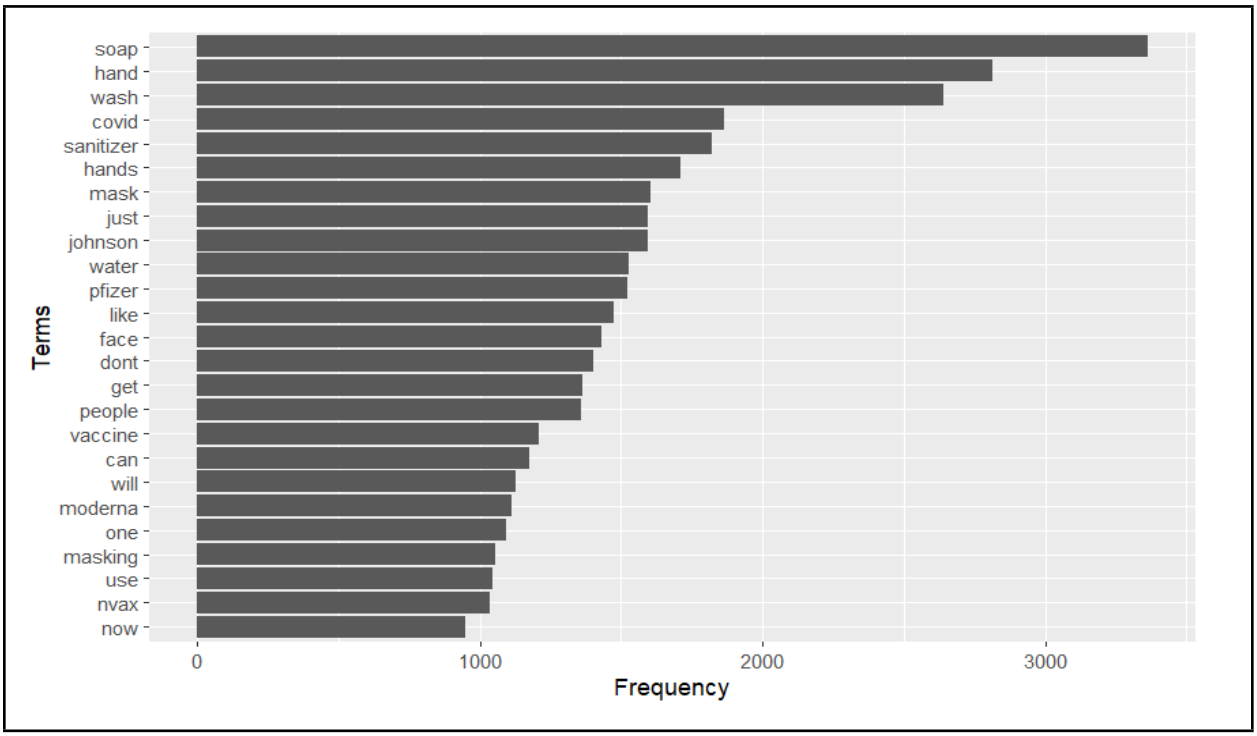


Continuing on sentiment analysis, the following chart visualizes the number of tweets about each of the vaccines across the most popular emotions in our dataset. From looking at the graph, we can determine that trust was the most popular emotion across all of the different types of vaccines. It is an odd result due to the fact that vaccine hesitancy was a major problem within the United States. The following emotion with the most number of tweets is anticipation which is most likely due to the fact that vaccines at the beginning of the pandemic were not widely available for everyone and left many age groups and non high risk people anxious about being able to be vaccinated. Looking at the more negative emotions, with fear being the third most tweeted about emotion for all vaccines. This could represent the other side of the vaccine debate that distrusting of a vaccine that has been rushed through production in order to aid during a crisis. Following fear, we can see that anger and sadness both in the lower side of the count of tweets which shows that Twitter as an overall social media platform heavily favored people becoming vaccinated.

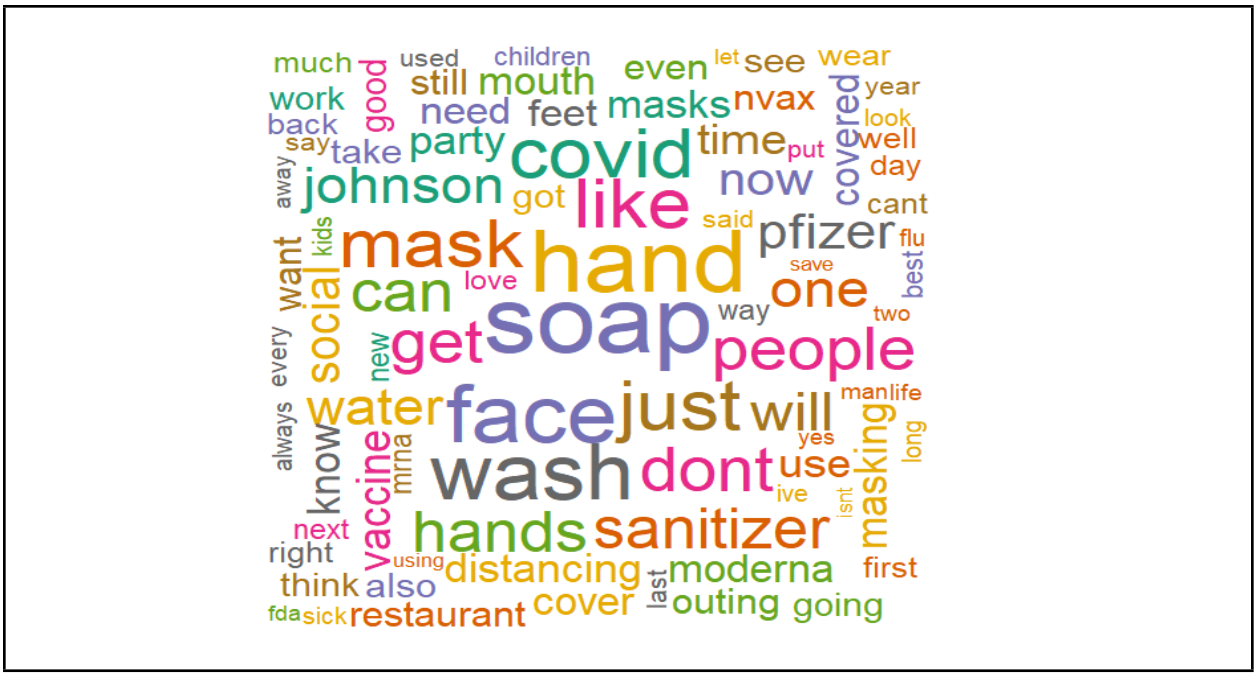


Another aspect of our twitter analysis was to break down the main words used in tweets throughout our entire dataset. We created a frequency chart that shows the distribution of each of the key terms. From

the visualization, we can clearly see that a large portion of the tweets involved hygiene related terms such as soap, hand, wash, and sanitizer which dominate the top terms. Following the hygiene related terms, we can start seeing terms such as mask, and both the covid-19 vaccines Johnson and Johnson along with Pfizer that finish off the top ten terms. We can clearly see from the visualization that most people's biggest concerns were not about lockdowns or covid restrictions but rather more focused on the habits of individuals that would also help prevent the spread of the virus. This breakdown allows us to better understand what twitter users are actually talking about and allows us to point out individual key words that are being used and not just analyze the overall sentiment of a given tweet.



Finally, the last visualization for our twitter analysis was a wordcloud of the most commonly used terms in our dataset. Again, we can see some of the most prominent words being hygiene related with terms such as soap, wash, water, and wash. Having Covid-19 spreading across the world, it is no surprise that hygiene is one of the biggest topics of discussion since the virus greatly emphasized the need to do the simple things such as washing your hands, using soap, and just generally being cleaner. The next biggest group of terms in the word cloud charts revolving around masking and social distancing. This also includes words such as travel, party and restaurants since throughout the pandemic locations that involve socializing had to go through phases of lockdowns leaving people talking about the precautions taken such as wearing masks and distancing themselves from others. It is surprising that a majority of the terms in the word cloud appear to be positive or neutral words and does not contain many negative terms even with the amount of people that shared their opinions against the covid vaccines.



Summary of findings and implications:

Through our analysis of the Covid-19 pandemic, our initial hypothesis was correct being that the higher populated states would have a proportionally larger number of cases as well as number of deaths and our analysis confirmed it. Throughout all of our visualizations there was a clear correlation with population size and the effects of the virus across the United States. What we can claim is that the North Eastern states such as New York, Pennsylvania, Connecticut, Massachusetts, and even as far inland as Illinois seem to have been the most impacted region in the United States with states such as California, Nevada, and Arizona being the most affected states on the west coast. It has become clear that only certain states have been majorly affected by the covid-19 virus which shows that federal restrictions were not required and restrictions should have been left to individual states since only a few were severely impacted. The use of federal level restrictions would have been unfair to a majority of the population since each state had to deal with different amounts of positive cases and those with very few cases would be unfairly forced to follow restrictions that do not need to be applied in their own region. The percentage of total vaccines spread across the country also helps support this argument since we can see that the states that had more positive cases were also taking more of the country's vaccination. In some locations vaccinations were not necessary while in others it was almost needed due to the amount of spreading that was occurring and it should have been left to each individual state in order to better manage the allocation of vaccines in order to better utilize them in major US hot spots rather than in unneeded areas.

When it comes to the usage of the actual vaccines, we can claim that although it did not immediately solve the pandemic, mostly due to constant new variants, the number of vaccines being taken does seem to have an effect on the number of positive cases with a lag of a few weeks. It can also be said that the number of vaccines were following the trend of cases since we can see instances where after a large spike in case numbers, we then see a large spike in vaccinations as a response which is usually followed with a decline

in covid cases. The clear winner of the vaccine manufacturers was Pfizer, as we can see from our visualizations it has taken the largest percent of the market share in almost all of the countries within our dataset. Over the course of time, Pfizer who was one of the first vaccines to be approved, was never in the position where they fell behind in production and never allowed another company to come near the top. By far the second most used vaccine was Moderna which was also one of the very first vaccines but they were never able to become the dominant vaccine in a specific country even though it dominates any of the other competitors. This was most likely due to vaccine production issues and even some distrust in their product since at the start of the vaccination period, people had the option between Pfizer and Moderna and it is possible that people's bias played a role across social media in order to make Pfizer seem as the safer option.

For our analysis on our twitter dataset, we have come to the conclusion that most of the data collected was in favor of vaccinations compared to against. This could possibly be due to the random data collected being positively skewed, that the during our date range covid hysteria had slowed down, or even possibly due to the fact that twitter was known for it liberal and left sided beliefs within the organization and its users so it would be interesting to see sentiments of post for various other social media platforms to see if there is any bias within the users that use the specific platform. It is clear that Covid-19 and the vaccine was a topic that was discussed on a small scale but once time went on there are clear events that we can attribute to the large spikes in the number of tweets that sparked a long lasting upward trend in tweets. From the text analysis, we continue to prove that many of the sentiments behind our data were in fact positive but there was still a decent amount of negative sentiments that we clearly expected but not at the frequency that we initially believed.

Lessons learned:

Throughout our analysis, some of the major challenges that we took on made our team have to rethink and reevaluate our process of creating visualizations since we were no longer doing them ourselves but rather as a group that can have different views on the direction of the dashboard and different methods of creating visualizations. Going through the analysis, a major aspect of it was trying to get all of the pieces working correctly in order to make sure our visualizations were accurate and a true representation of the dataset. In order for our team to complete our goal, we had to first decide on the tools that were going to be the most successful at completing a specific task which led us to utilizing multiple tools in combination to reach our conclusion. Understanding how to prepare the information for each tool was something that had to be planned for in order to not run into issues down the line when using a tool intended for a different purpose. This challenge was a great example of how the life of a data analyst would consist of since a large portion of the time spent was on the preparation of the data rather than the actual creation of the visualizations. Preparing the data was not only the most time consuming but also the area where we needed to think the most in order to have our dataset ready to accurately and properly visualize since none of our charts would have been useful if the original data was not up to the task due to uncleanliness.

This carried over to our twitter analysis where we quickly realized the incredible amount of work that is needed in order to prepare scraped twitter information. Once the data was able to be collected, it was fairly easy to run the code but without the ability to extract the useful aspects of scraped information then it would have been useless. Overall, we as a team believe that this analysis forced us to think outside of simply repeating a task completed in class and to think about the overall lessons we have learned regarding how to create a clean dataset that can easily be used, visualizations that clearly depicts an idea, and finally

being able to combine everything together to create a story that is comprehensible for any audience with varying levels of data management skill.

References:

"CDC Museum Covid-19 Timeline." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 16 Aug. 2022, https://www.cdc.gov/museum/timeline/covid19.html.

Mathieu, Edouard, et al. "Coronavirus (COVID-19) Vaccinations." *Our World in Data*, 5 Mar. 2020, https://ourworldindata.org/covid-vaccinations.

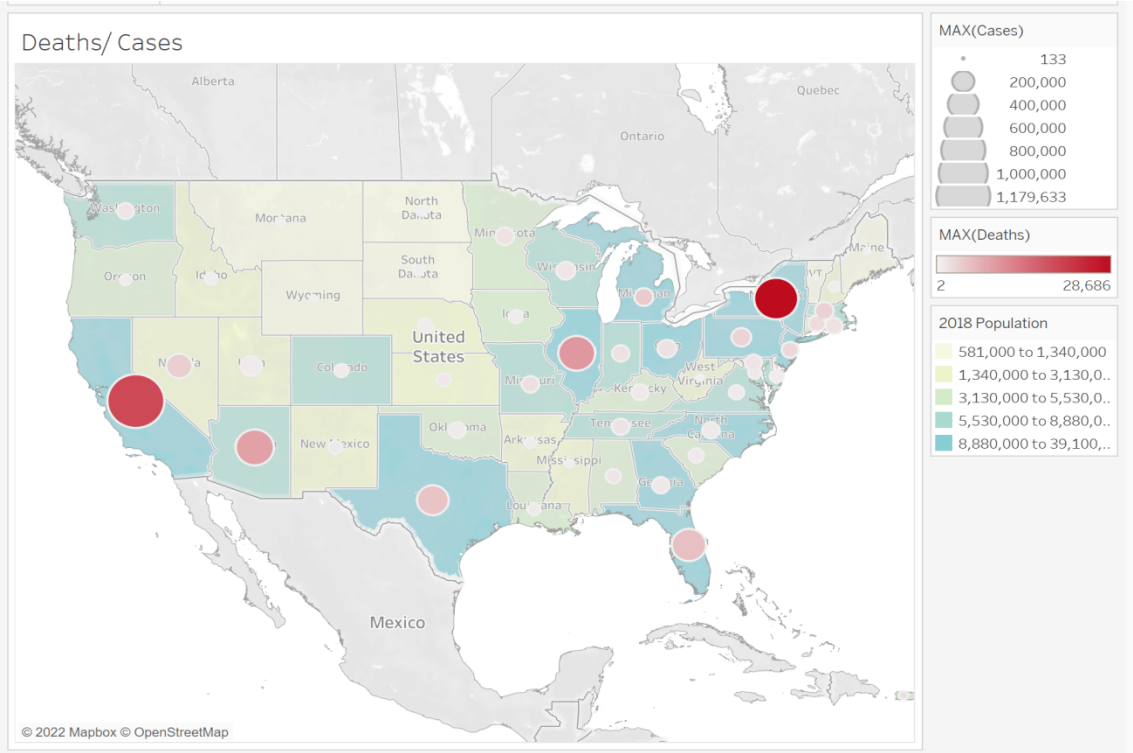
Sattar, Naw Safrin, and Shaikh Arifuzzaman. "Covid-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA." *Applied Sciences*, vol. 11, no. 13, 2021, p. 6128., https://doi.org/10.3390/app11136128.

Appendix:

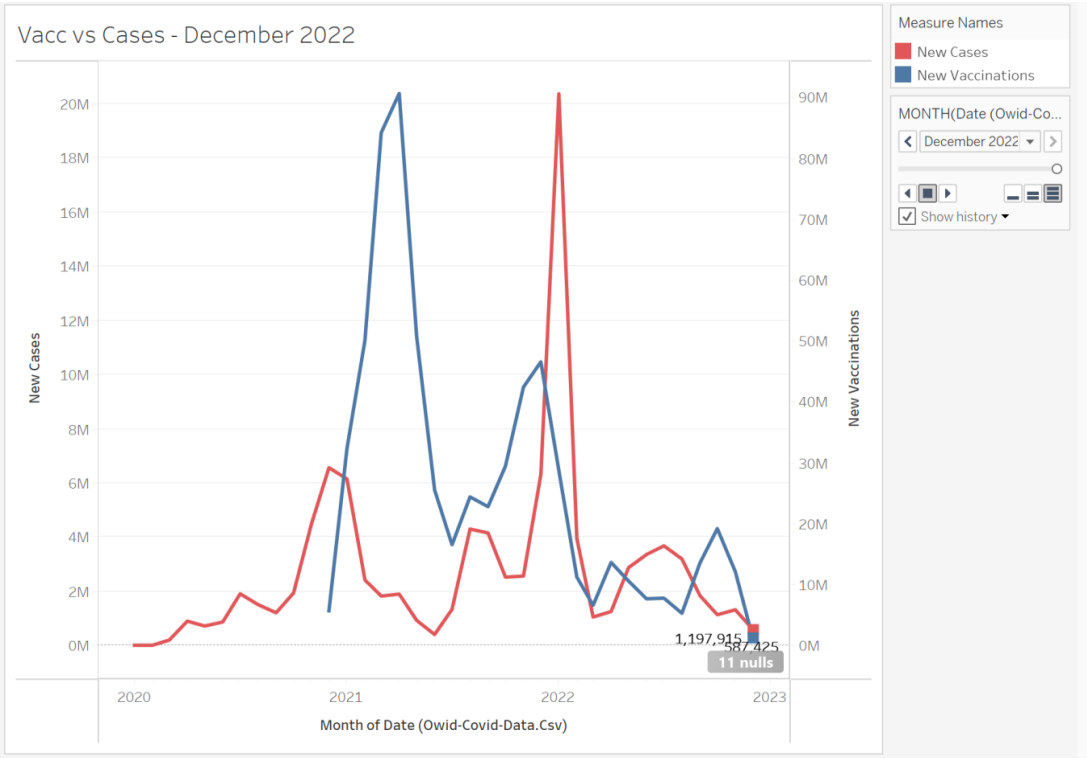
1.

Twitter Topic Variables	Include Keywords
Pfizer	pfizer, Pfizer-BioNTech, BioNTechpfizer
Moderna	Moderna, moderna_tx, Moderna-NIAID, NIAID, NIAID-Moderna
Johnson&Johnson	Johnson & Johnson, Johnson and Johnson, Janssen, Janssen Pharmaceutical, J&J
Oxford-AstraZeneca	OXFORDVACCINE, Oxford-Astraeneca, OxfordAstraZeneca, AstraZeneca, Vaxzevria, Covishield
SputnikV	Sputnik V, sputnikv, sputnikvaccine
Covaxin	covaxin, BharatBiotech
Sinovac	coronavac, sinovac
Hygiene	hand sanitizer, sanitizer, wash hands, wash face, soap, soap water, hand soap, sanitize
Wear Mask	mask, wearamask, masking, N95, face cover, face covering, face covered, mouth cover, mouth covering, mouth covered, nose cover, nose covering, nose covered, cover your face, coveryourface
Travel	travel, outing, camping, air-travel
Social Distancing	social distancing, physical distancing, 6 feet, social distance, physical distance
Social Gathering	social gathering, gathering, party, restaurant

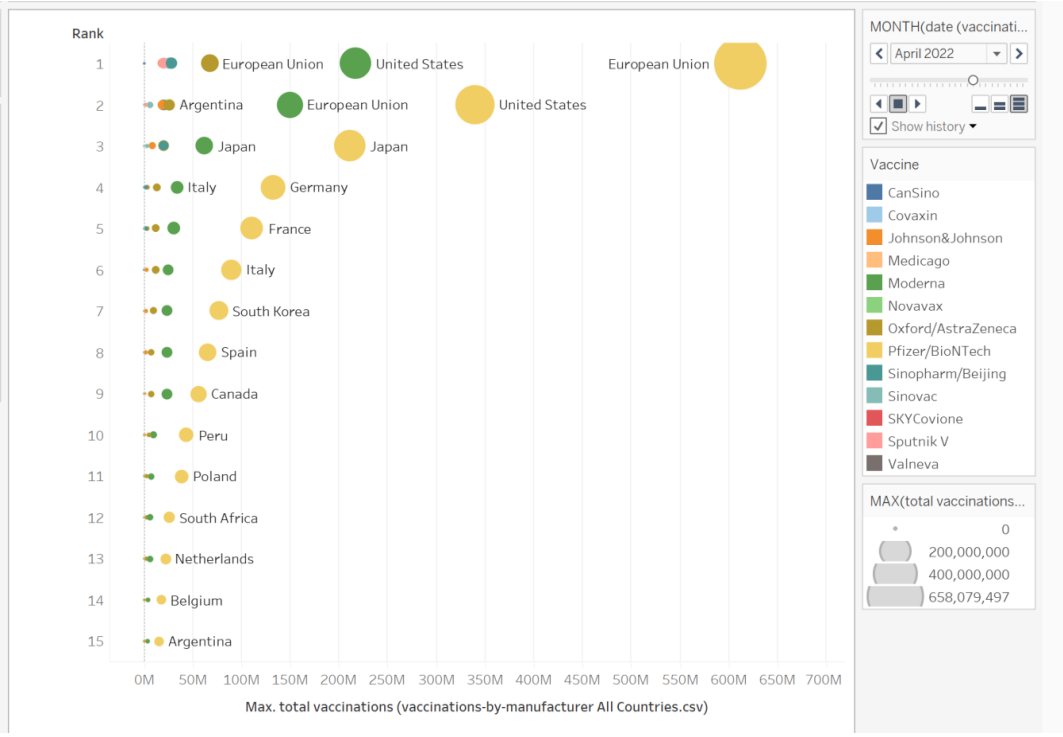
2.


Deaths/ Cases
Alberta
Ontario
Quebec
United States
Mexico
© 2022 Mapbox © OpenStreetMap
MAX(Cases)
133
200,000
400,000
600,000
800,000
1,000,000
1,179,633
MAX(Deaths)
2
28,686
2018 Population
581,000 to 1,340,000
1,340,000 to 3,130,0..
3,130,000 to 5,530,0..
5,530,000 to 8,880,0..
8,880,000 to 39,100,..

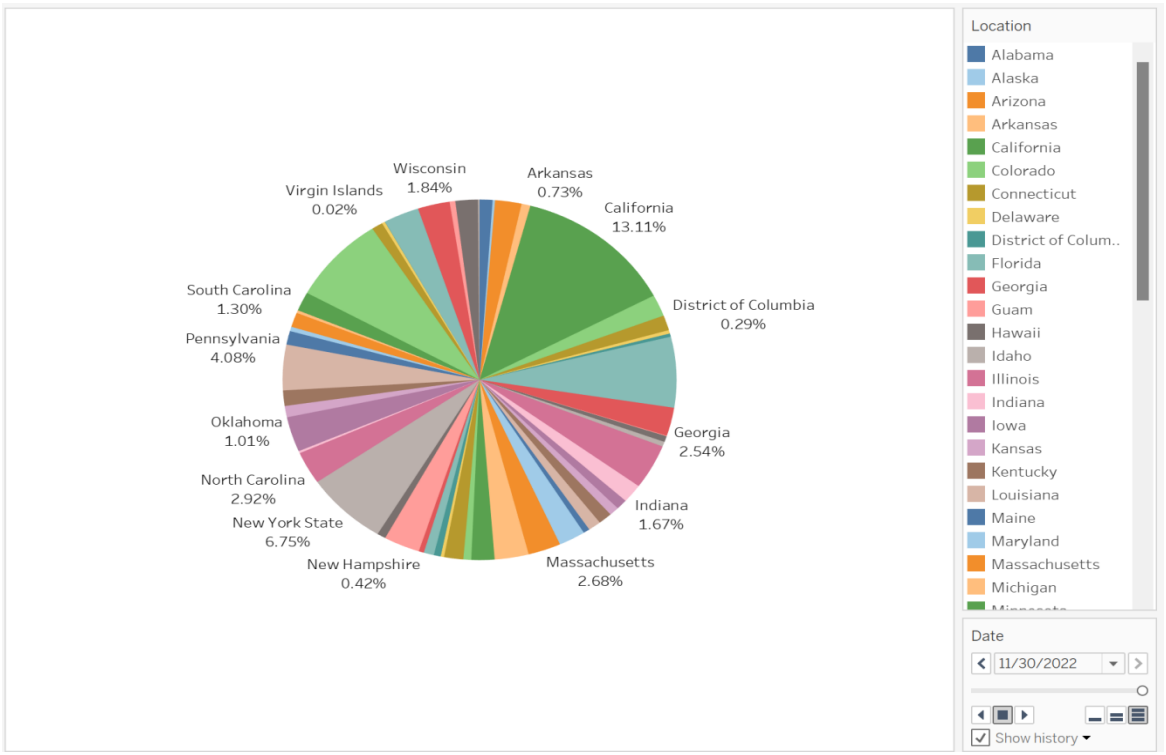
3.


Vacc vs Cases - December 2022
New Cases
New Vaccinations
Month of Date (Owid-Covid-Data.Csv)
2020
2021
2022
2023
1,197,915
587,425
11 nulls
Measure Names
New Cases
New Vaccinations
MONTH(Date (Owid-Co...
December 2022
Show history

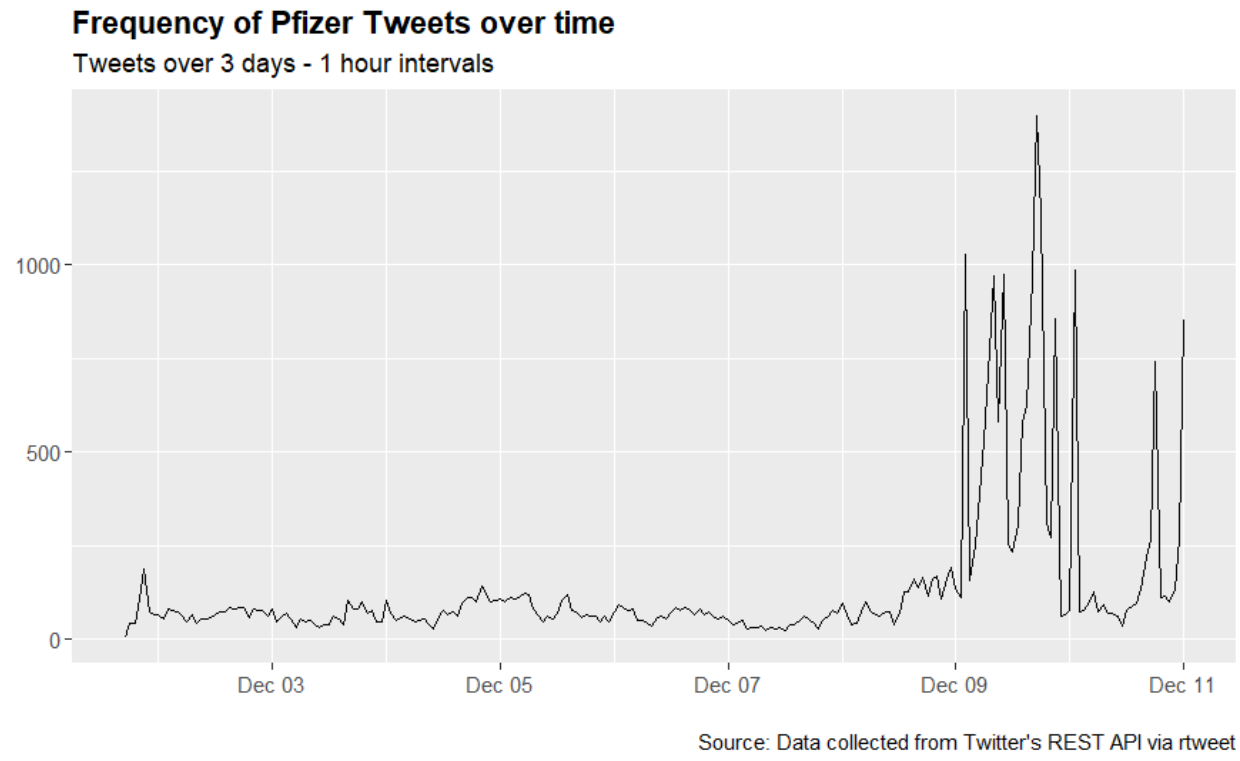
4.


Rank
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
European Union
United States
European Union
Argentina
European Union
United States
Japan
Japan
Italy
Germany
France
Italy
South Korea
Spain
Canada
Peru
Poland
South Africa
Netherlands
Belgium
Argentina
0M
50M
100M
150M
200M
250M
300M
350M
400M
450M
500M
550M
600M
650M
700M
Max. total vaccinations (vaccinations-by-manufacturer All Countries.csv)
MONTH(date (vaccinati...
April 2022
Show history
Vaccine
CanSino
Covaxin
Johnson&Johnson
Medicago
Moderna
Novavax
Oxford/AstraZeneca
Pfizer/BioNTech
Sinopharm/Beijing
Sinovac
SKYCovione
Sputnik V
Valneva
MAX(total vaccinations...
0
200,000,000
400,000,000
658,079,497

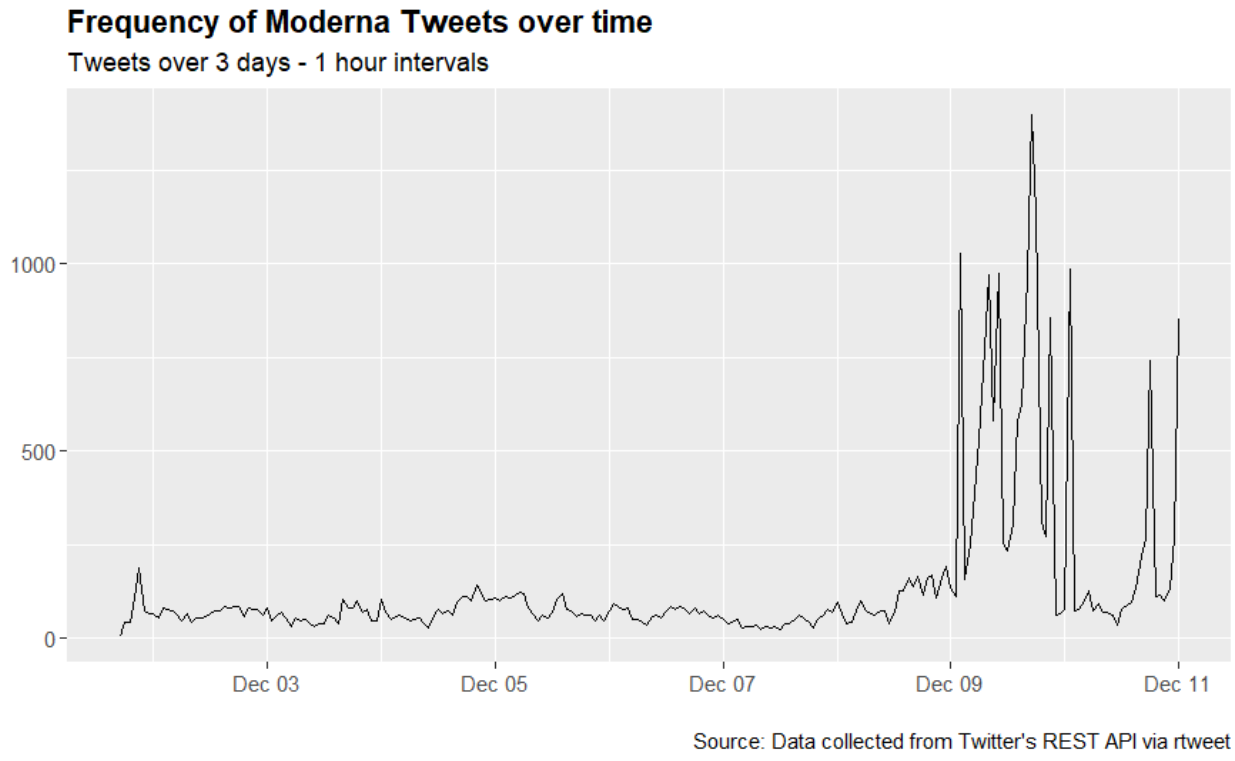
5.


Wisconsin
1.84%
Arkansas
0.73%
Virgin Islands
0.02%
California
13.11%
South Carolina
1.30%
District of Columbia
0.29%
Pennsylvania
4.08%
Georgia
2.54%
Oklahoma
1.01%
Indiana
1.67%
North Carolina
2.92%
New York State
6.75%
New Hampshire
0.42%
Massachusetts
2.68%
Location
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Colum..
Florida
Georgia
Guam
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Date
11/30/2022
Show history

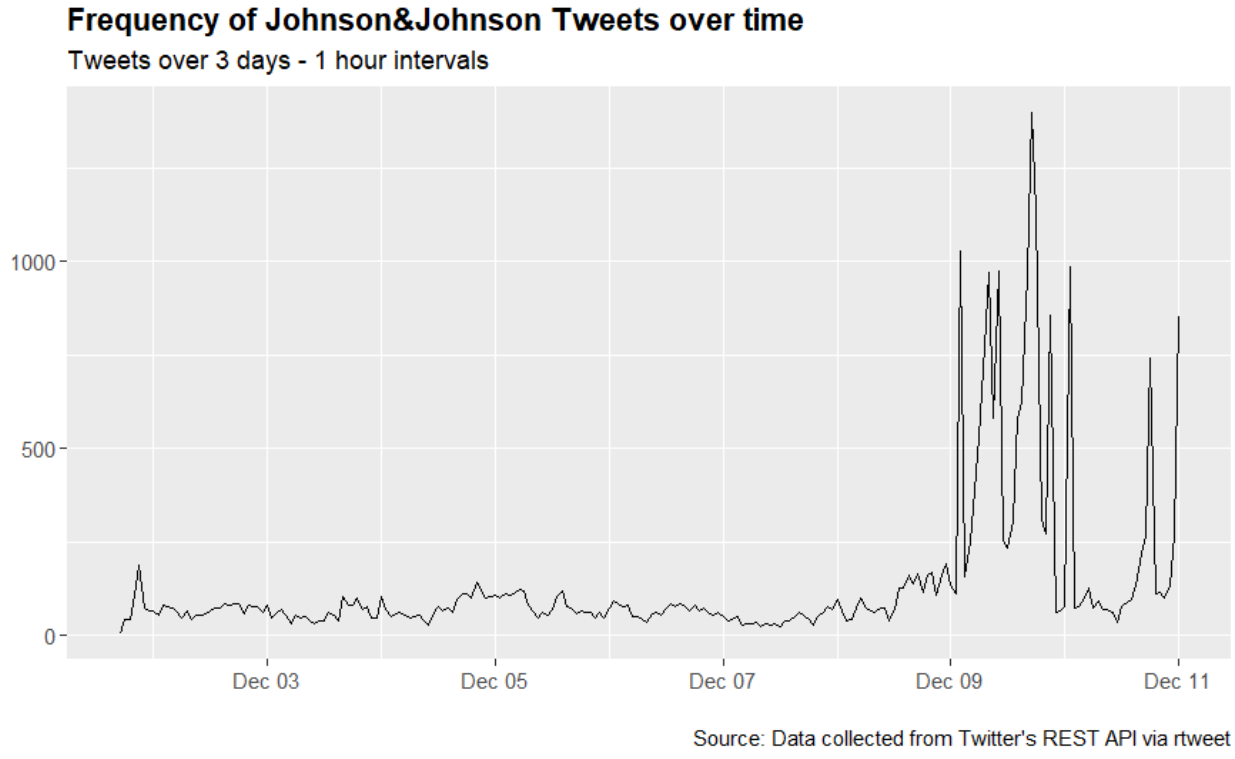
6.


Frequency of Pfizer Tweets over time
Tweets over 3 days - 1 hour intervals
1000
500
0
Dec 03
Dec 05
Dec 07
Dec 09
Dec 11
Source: Data collected from Twitter's REST API via rtweet

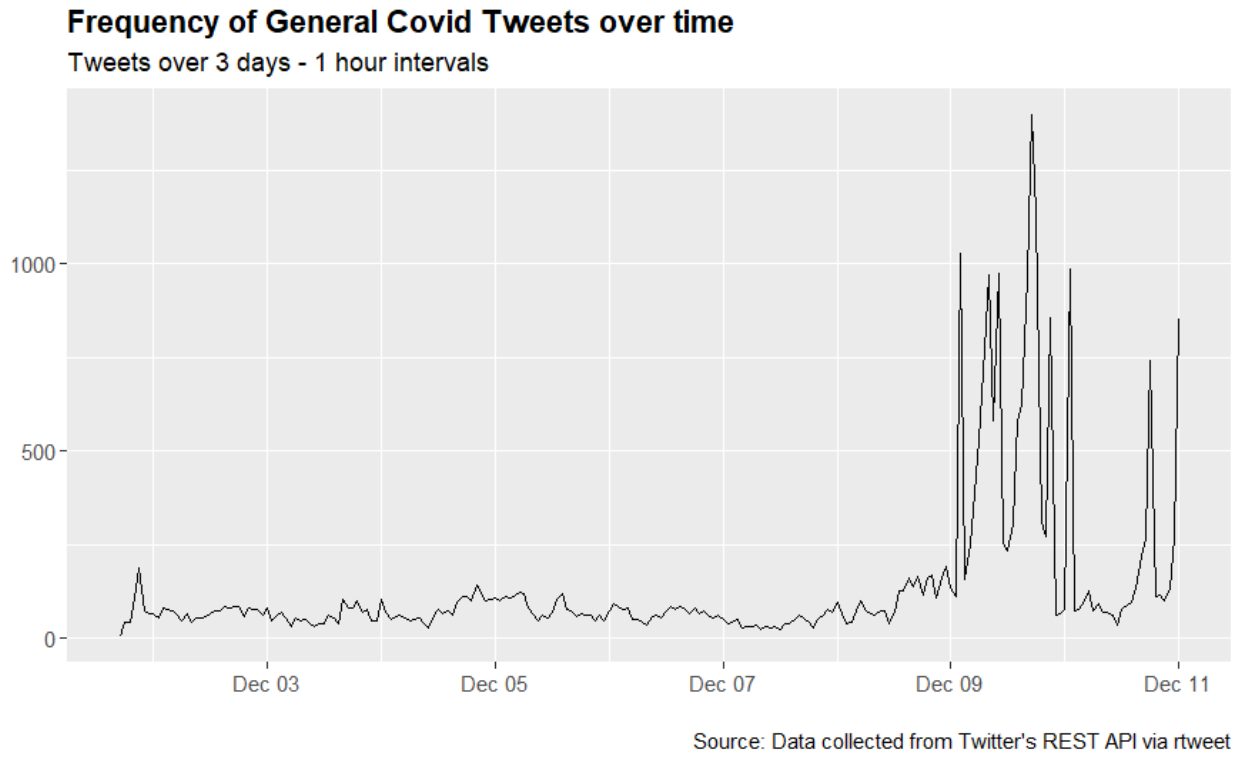
7.


Frequency of Moderna Tweets over time
Tweets over 3 days - 1 hour intervals
1000
500
0
Dec 03
Dec 05
Dec 07
Dec 09
Dec 11
Source: Data collected from Twitter's REST API via rtweet

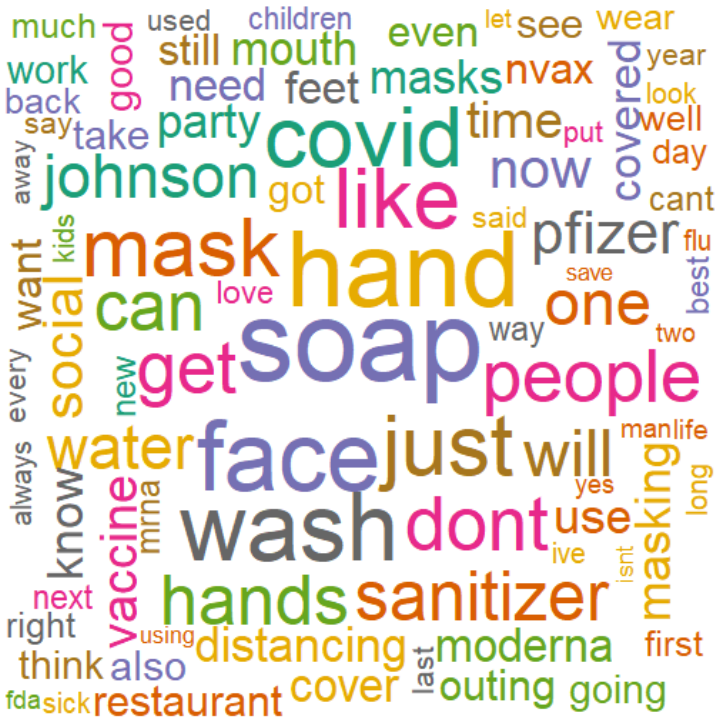
8.


Frequency of Johnson&Johnson Tweets over time
Tweets over 3 days - 1 hour intervals
1000
500
0
Dec 03
Dec 05
Dec 07
Dec 09
Dec 11
Source: Data collected from Twitter's REST API via rtweet

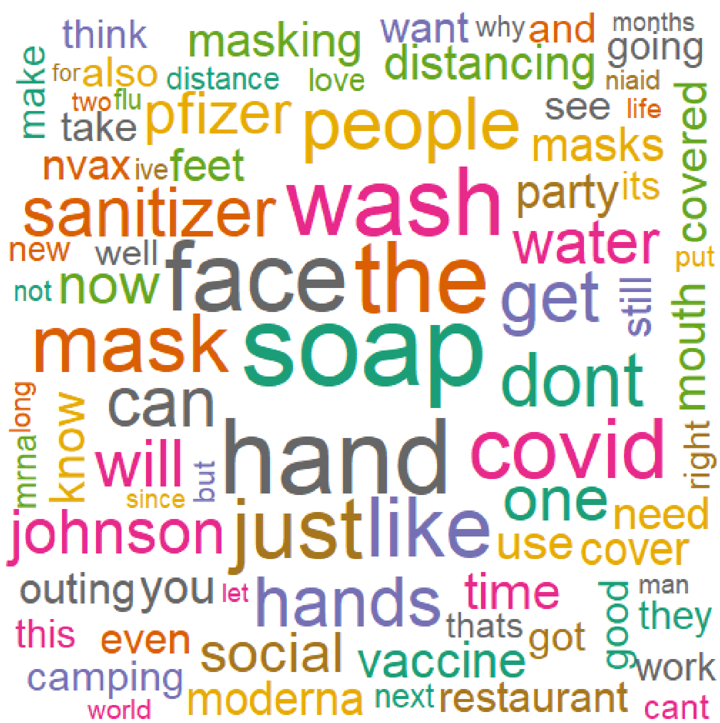
9.


Frequency of General Covid Tweets over time
Tweets over 3 days - 1 hour intervals
1000
500
0
Dec 03
Dec 05
Dec 07
Dec 09
Dec 11
Source: Data collected from Twitter's REST API via rtweet

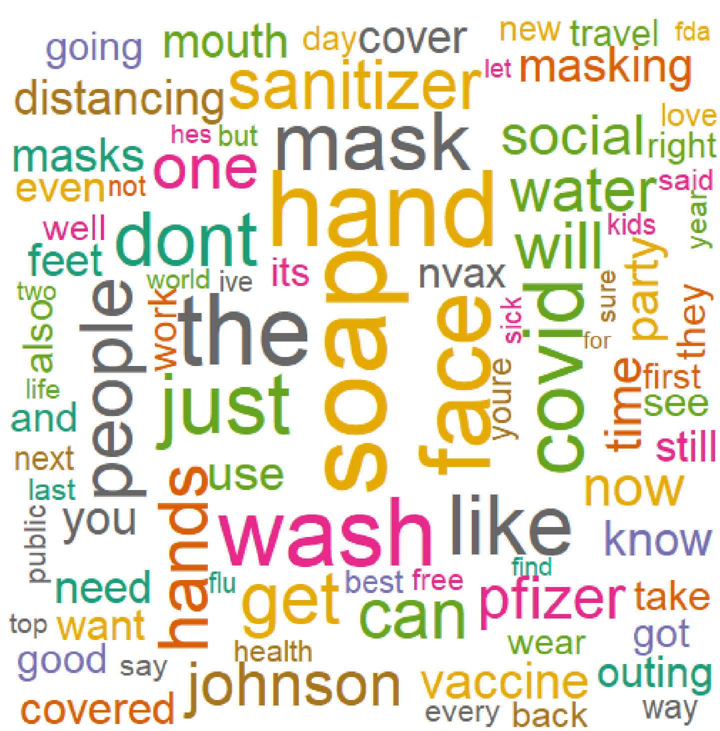
10.


much
used
children
even
let
see
wear
work
still
mouth
good
need
feet
masks
nvax
year
back
look
say
take
party
covid
time
put
well
away
johnson
got
like
now
covered
day
cant
said
pfizer
flu
kids
want
mask
hand
save
best
social
can
love
one
every
soap
way
two
get
people
new
water
face
just
will
manlife
always
know
vaccine
mrna
wash
dont
use
yes
long
masking
ive
isnt
next
hands
sanitizer
right
using
distancing
moderna
first
think
also
last
cover
outing
going
fda
sick
restaurant

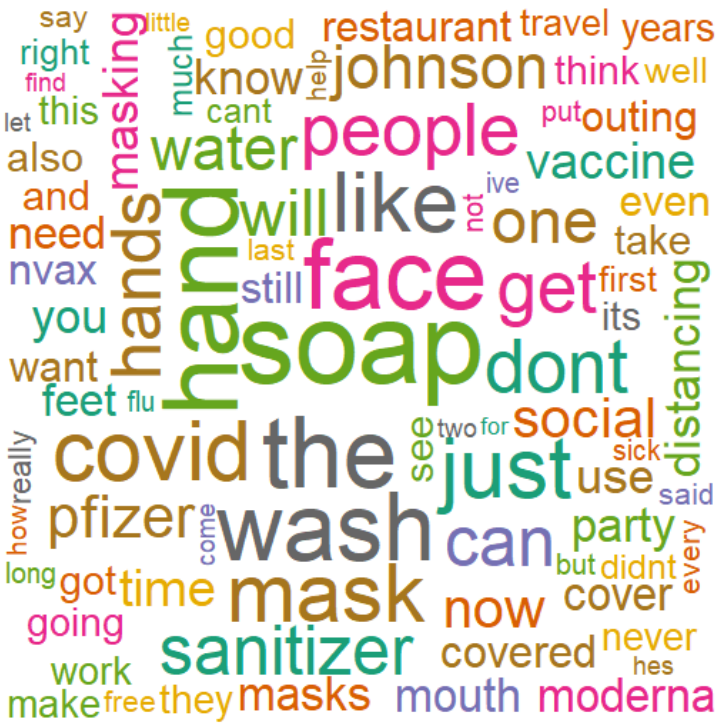
11.


think
masking
want
why
and
months
going
make
for
also
distance
love
distancing
niaid
two
flu
take
pfizer
people
see
life
masks
covered
nvax
ive
feet
sanitizer
wash
party
its
new
well
water
put
not
now
face
the
get
still
mouth
mask
soap
dont
long
mrna
know
can
will
but
since
hand
covid
right
johnson
just
like
one
need
use
cover
outing
you
let
hands
time
man
good
they
this
even
thats
got
social
vaccine
camping
work
moderna
next
restaurant
world
cant

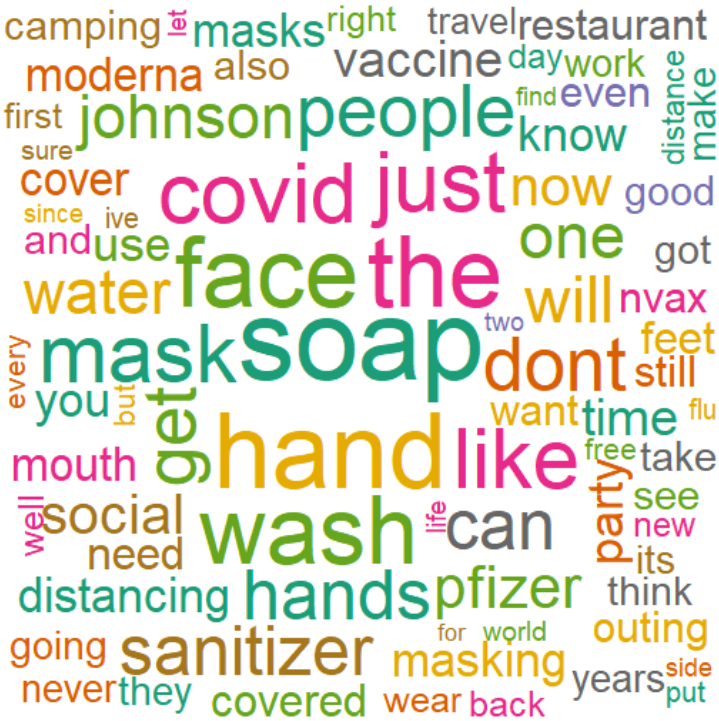
12.


going
mouth
day
cover
new
travel
fda
sanitizer
let
masking
distancing
hes
but
mask
love
social
right
masks
one
even
not
water
said
well
hand
year
kids
dont
will
feet
two
world
ive
its
nvax
also
people
work
the
sick
sure
party
for
they
soap
face
covid
life
youre
time
first
just
see
and
still
next
use
now
last
public
you
hands
wash
like
know
find
need
flu
best
free
get
pfizer
take
top
want
can
wear
got
good
say
health
johnson
vaccine
outing
covered
every
back
way

13.


say
little
good
restaurant
travel
years
right
masking
much
know
help
johnson
think
well
find
this
cant
put
outing
let
water
people
vaccine
also
hand
will
like
ive
not
one
even
and
hands
need
last
face
get
take
nvax
still
first
you
its
soap
dont
want
distancing
feet
flu
social
see
two
for
sick
covid
the
just
use
said
howreally
pfizer
come
wash
can
party
long
got
time
mask
but
didnt
every
now
cover
going
sanitizer
covered
never
work
hes
make
free
they
masks
mouth
moderna

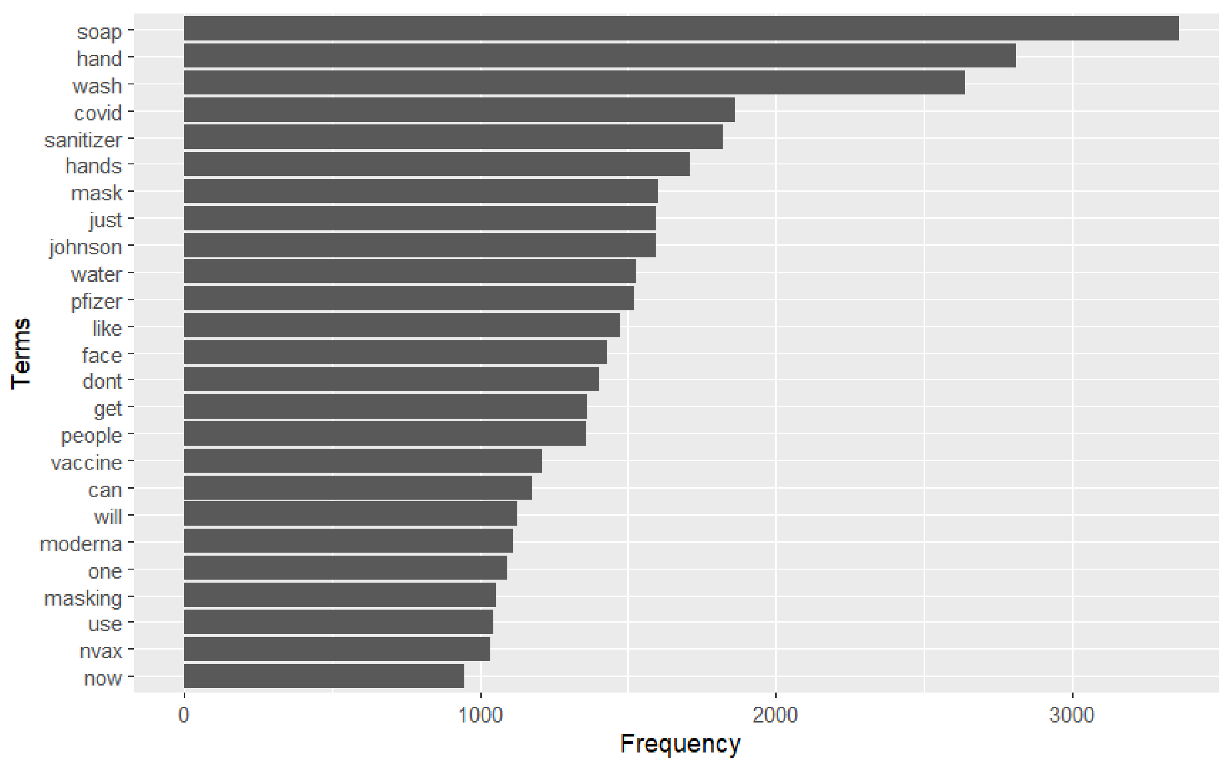
14.


camping
let
masks
right
travel
restaurant
moderna
also
vaccine
day
work
distance
make
first
johnson
people
find
even
know
sure
cover
covid
just
now
good
since
ive
one
and
use
got
face
the
water
will
nvax
two
feet
every
mask
soap
dont
still
you
but
get
want
time
flu
hand
like
free
take
mouth
party
see
social
wash
can
life
new
well
need
its
distancing
hands
pfizer
think
for
world
outing
going
sanitizer
masking
years
side
neverthey
covered
wear
back
put

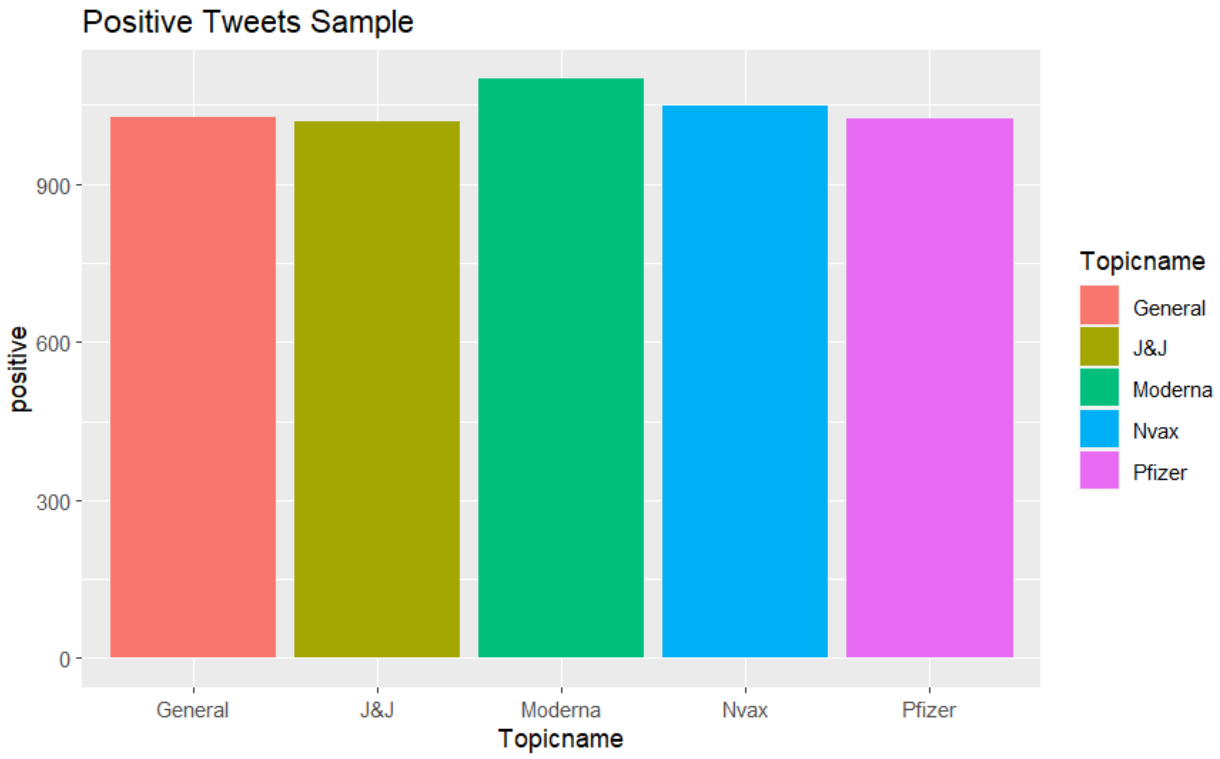
15.


right
booster
covered
cover
even
this
also
next
since
good
going
sanitizer
pfizer
need
novavax
not
masks
ive
side
hes
world
can
mouth
want
but
will
like
just
time
use
find
for
let
nvax
flu
face
the
mask
social
its
yes
distance
people
soap
now
see
last
one
get
fda
week
masking
kids
two
party
hand
covid
life
work
vaccine
wash
dont
know
got
you
moderna
said
never
outing
made
hands
water
still
first
say
think
johnson
top
what
big
day
camping
feet
put
distancing
take
mrna

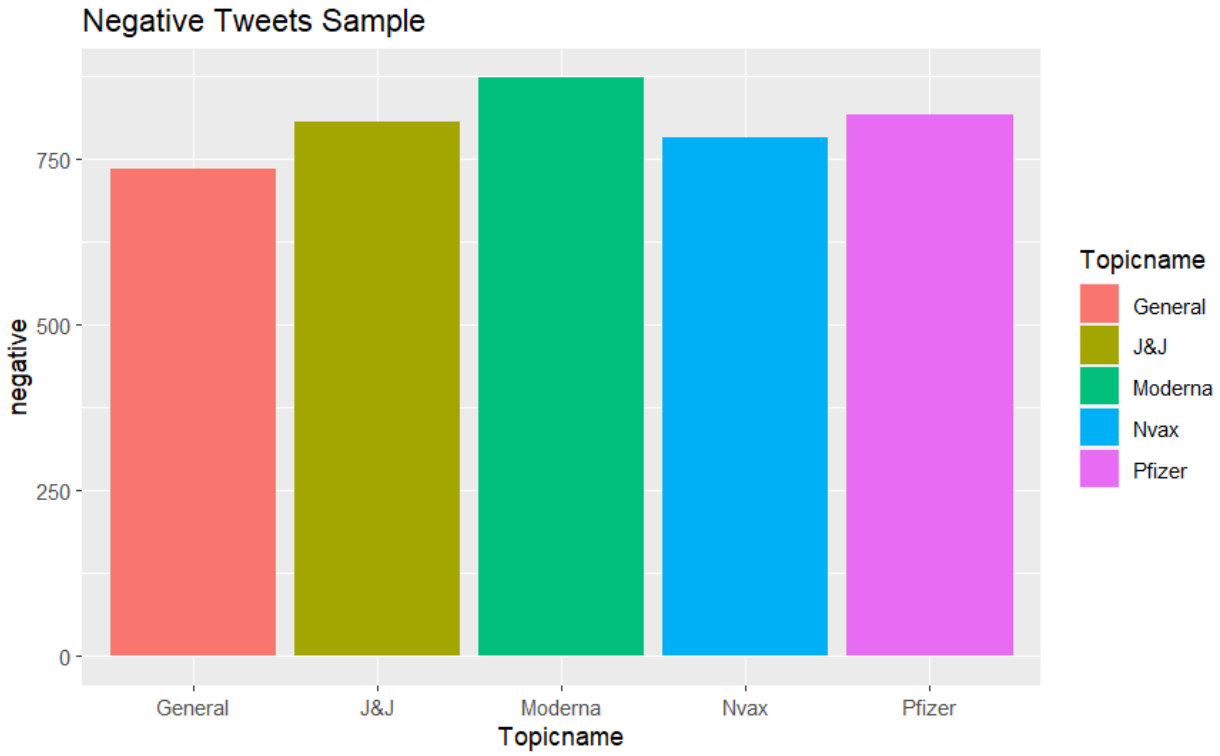
16.


soap
hand
wash
covid
sanitizer
hands
mask
just
johnson
water
pfizer
like
face
dont
get
people
vaccine
can
will
moderna
one
masking
use
nvax
now
Terms
0
1000
2000
3000
Frequency

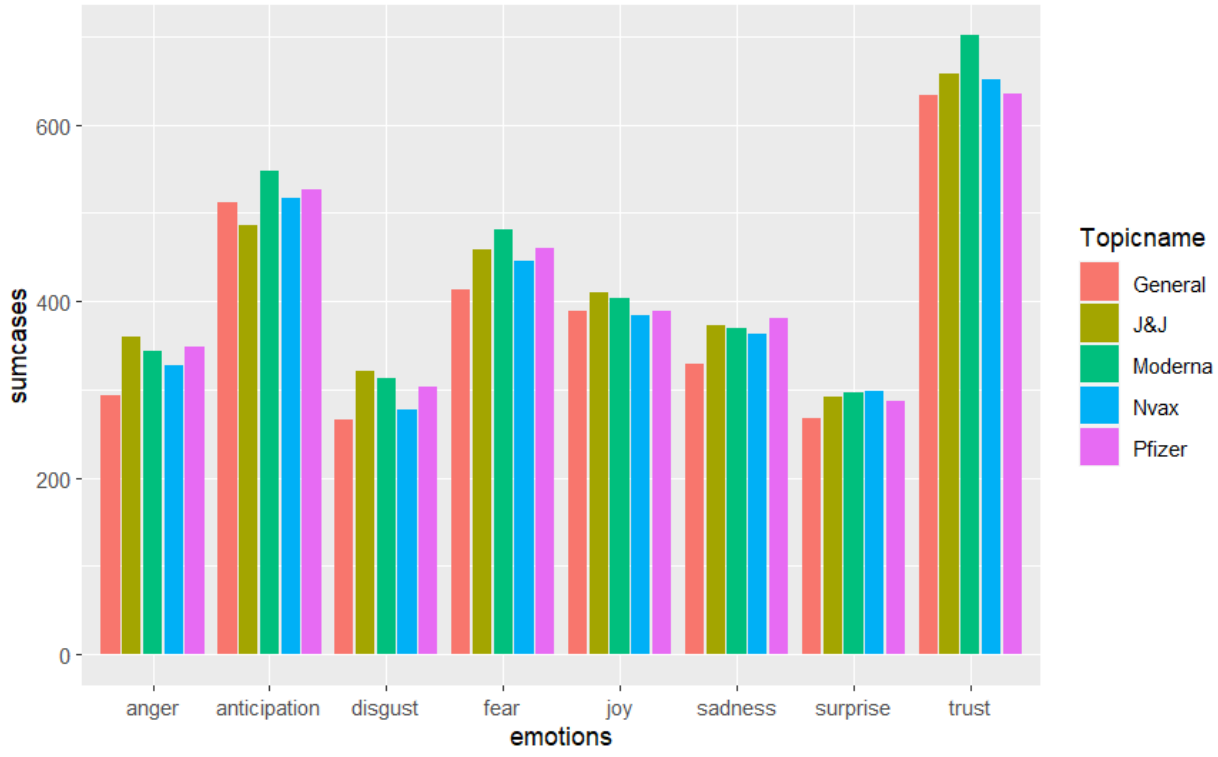
17.


Positive Tweets Sample
positive
900
600
300
0
General
J&J
Moderna
Nvax
Pfizer
Topicname
Topicname
General
J&J
Moderna
Nvax
Pfizer

18.


Negative Tweets Sample
negative
750
500
250
0
General
J&J
Moderna
Nvax
Pfizer
Topicname
Topicname
General
J&J
Moderna
Nvax
Pfizer

19.



R Code:

```
library(rtweet)
library(tidyverse)
library(tidytext)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(syuzhet)
library(maps)
library(textstem)
library(webr)
library(systemfonts)

appname<-"DS804Final"
apikey<-"dYMvPjgpHfpjYAgfClcjmTCUo"
apisecret<-"Q5gdpHp5ChzjEhWFxPzvz8hfJVTftCZICgChTsJtRSoGNnwwvP"
access_token<-"158247622382075084 8-iU4Pv778Dtfg0RArnau2aXU5fk9EZo"
access_secret<-"ZtLszenFjtgYvMtqwCEg2CmKC2AJiXhYSqGN0ZwDFvz5b"
```

```
AuthTwitter<- create_token(
  app=appname,
  consumer_key =apikey,
  consumer_secret=apisecret,
  access_token =access_token,
  access_secret =access_secret
)

pfizer <- search_tweets(q="pfizer", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

Pfizer_BioNTech <- search_tweets(q="Pfizer-BioNTech", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

BioNTechpfizer <- search_tweets(q="BioNTechpfizer", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

Moderna <- search_tweets(q="Moderna", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

moderna_tx <- search_tweets(q="moderna_tx", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

Moderna_NIAID <- search_tweets(q="Moderna-NIAID", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

NIAID <- search_tweets(q="NIAID", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

NIAID_Moderna <- search_tweets(q="NIAID-Moderna", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

JohnsonndJohnson <- search_tweets(q="Johnson & Johnson", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

JohnsonandJohnson <- search_tweets(q="Johnson and Johnson", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

Janssen <- search_tweets(q="Janssen", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

JanssenPharmaceutical <- search_tweets(q="Janssen Pharmaceutical", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

JnJ <- search_tweets(q="J&J", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

Novavax <- search_tweets(q="Novavax", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)
```

```
Nvax <- search_tweets(q="Nvax", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

socialgathering <- search_tweets(q="social gathering", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

gathering <- search_tweets(q="gathering", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

party <- search_tweets(q="party", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

restaurant <- search_tweets(q="restaurant", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

socialdistancing <- search_tweets(q="social distancing", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

physicaldistancing <- search_tweets(q="physical distancing", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

sixfeet <- search_tweets(q="6 feet", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

socialdistance <- search_tweets(q="socialdistance", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

physicaldistance <- search_tweets(q="physical distance", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

travel <- search_tweets(q="travel", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

outing <- search_tweets(q="outing", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

camping <- search_tweets(q="camping", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

air_travel <- search_tweets(q="air-travel", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

hand_sanitizer <- search_tweets(q="hand sanitizer", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

sanitizer <- search_tweets(q="sanitizer", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

wash_hands <- search_tweets(q="wash hands", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)
```

```
wash_face <- search_tweets(q="wash face", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

soap <- search_tweets(q="soap", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

soap_water <- search_tweets(q="soap water", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

hand_soap <- search_tweets(q="hand soap", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

sanitize <- search_tweets(q="sanitize", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

mask <- search_tweets(q="mask", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

wearamask <- search_tweets(q="wearamask", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

masking <- search_tweets(q="masking", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

N95 <- search_tweets(q="N95", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

face_covered <- search_tweets(q="face covered", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

mouth_cover <- search_tweets(q="mouth cover", n=1000, include_rts =FALSE, lang = "en", token = AuthTwitter)

pfizertwts = rbind(pfizer, BioNTechpfizer, Pfizer_BioNTech)

modernatwts = rbind(Moderna, moderna_tx, Moderna_NIAID, NIAID, NIAID_Moderna)

jnjtwts = rbind(JnJ, JohnsonandJohnson, JohnsonndJohnson, Janssen, JanssenPharmaceutical)

novavaxtwts = rbind(Novavax, Nvax)

generaltwts = rbind(hand_sanitizer, sanitizer, wash_hands, wash_face, soap, soap_water, hand_soap, sanitize, mask, wearamask, masking, N95, face_covered, mouth_cover, travel, outing, camping, air_travel, socialgathering, socialdistancing, socialdistance, physicaldistance, physicaldistancing, sixfeet, party, restaurant)

df = rbind(pfizertwts, modernatwts, jnjtwts, novavaxtwts, generaltwts)

#Text Analysis
```

```
text <- df$text
text[1:5]

text<-tolower(text)
text[1:5]

text<-gsub('[^\x20-\x7E]',",text)
text[1:5]

text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text[1:5]

text<-gsub("&amp", "and", text)
text[1:5]

text<-gsub("[[:punct:]]", "", text)
text[1:5]

text<-gsub("[[:digit:]]", "", text)
text[1:5]

text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text[1:5]

text<-str_replace_all(text, " ", " ")
text[1:5]

text<-removeWords(text, stopwords("english"))
text[1:5]

text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
text[1:5]

df_clean <- bind_cols(df, text, id=NULL)

df_cleaned <-df_clean %>%
  rename(text_updated = ...44)

pfizertwts%>%
  select(text)
```

```
text <- df$text
text<-gsub('[^\x20-\x7E]',",text)
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", text)
text<-gsub("@\\w+", "", text)
text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text<-gsub("&amp", "and", text)
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]", "", text)
text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text<-str_replace_all(text, " ", " ")
text<-removeWords(text, stopwords("english"))
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
pfizertwts1 <- bind_cols(df, text, id=NULL)

df_cleaned1 <-pfizertwts1 %>%
  rename(text_updated = ...44)

modernatwts%>%
  select(text)

text <- df$text
text<-gsub('[^\x20-\x7E]',",text)
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", text)
text<-gsub("@\\w+", "", text)
text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text<-gsub("&amp", "and", text)
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]", "", text)
text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text<-str_replace_all(text, " ", " ")
text<-removeWords(text, stopwords("english"))
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
modernatwts1 <- bind_cols(df, text, id=NULL)

df_cleaned2 <-modernatwts1 %>%
  rename(text_updated = ...44)

jnjtwts%>%
  select(text)
```

```
text <- df$text
text<-gsub('[^\x20-\x7E]',",text)
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", text)
text<-gsub("@\\w+", "", text)
text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text<-gsub("&amp", "and", text)
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]", "", text)
text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text<-str_replace_all(text, " ", " ")
text<-removeWords(text, stopwords("english"))
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
jnjtwts1 <- bind_cols(df, text, id=NULL)

df_cleaned3 <-jnjtwts1 %>%
  rename(text_updated = ...44)

novavaxtwts%>%
  select(text)

text <- df$text
text<-gsub('[^\x20-\x7E]',",text)
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", text)
text<-gsub("@\\w+", "", text)
text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text<-gsub("&amp", "and", text)
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]", "", text)
text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text<-str_replace_all(text, " ", " ")
text<-removeWords(text, stopwords("english"))
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
novavaxtwts1 <- bind_cols(df, text, id=NULL)

df_cleaned4 <-novavaxtwts1 %>%
  rename(text_updated = ...44)

generaltwts%>%
  select(text)
```

```
text <- df$text
text<-gsub('[^\x20-\x7E]',",text)
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", text)
text<-gsub("@\\w+", "", text)
text<-str_replace_all(text, "#[a-z,A-Z]*", " ")
text<-gsub("&amp", "and", text)
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]", "", text)
text<-gsub("http\\w+", " ", text)
text<-gsub("[ \t]{2,}", " ", text)
text<-gsub("^\\s+|\\s+$", " ", text)
text<-str_replace_all(text, " ", " ")
text<-removeWords(text, stopwords("english"))
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
generaltwts1 <- bind_cols(df, text, id=NULL)

df_cleaned5 <-generaltwts1 %>%
  rename(text_updated = ...44)

ts_plot(df_cleaned1, by = "hours")+
  theme(plot.title = element_text(face="bold")) +
  labs(
    x = NULL, y = NULL,
    title = "Frequency of Pfizer Tweets over time",
    subtitle = "Tweets over 3 days - 1 hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
  )

ts_plot(df_cleaned2, by = "hours")+
  theme(plot.title = element_text(face="bold")) +
  labs(
    x = NULL, y = NULL,
    title = "Frequency of Moderna Tweets over time",
    subtitle = "Tweets over 3 days - 1 hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
  )

ts_plot(df_cleaned3, by = "hours")+
  theme(plot.title = element_text(face="bold")) +
  labs(
    x = NULL, y = NULL,
    title = "Frequency of Johnson&Johnson Tweets over time",
    subtitle = "Tweets over 3 days - 1 hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
```

```
  )

ts_plot(df_cleaned4, by = "hours")+
  theme(plot.title = element_text(face="bold")) +
  labs(
    x = NULL, y = NULL,
    title = "Frequency of Novavax Tweets over time",
    subtitle = "Tweets over 3 days - 1 hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
  )

ts_plot(df_cleaned5, by = "hours")+
  theme(plot.title = element_text(face="bold")) +
  labs(
    x = NULL, y = NULL,
    title = "Frequency of General Covid Tweets over time",
    subtitle = "Tweets over 3 days - 1 hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
  )

df_cleaned %>%
  count(in_reply_to_screen_name, sort = TRUE) %>%
  mutate(in_reply_to_screen_name = reorder(in_reply_to_screen_name, n)) %>%
  top_n(10) %>%
  ggplot(aes(x = in_reply_to_screen_name, y = n))+
  geom_col()+
  coord_flip()+
  labs(x= "Username", y = "No. of Posts", title = "Users who posted frequently - Top-10")

wordcloud(df_cleaned$text_updated, min.freq = 2, scale = c(4,0.5),
          colors = brewer.pal(8, "Dark2"),
          random.color = TRUE, random.order = FALSE, max.words = 150)

wordcloud(df_cleaned1$text_updated, min.freq = 2, scale = c(4,0.5),
          colors = brewer.pal(8, "Dark2"),
          random.color = TRUE, random.order = FALSE, max.words = 150)

wordcloud(df_cleaned2$text_updated, min.freq = 2, scale = c(4,0.5),
          colors = brewer.pal(8, "Dark2"),
          random.color = TRUE, random.order = FALSE, max.words = 150)

wordcloud(df_cleaned3$text_updated, min.freq = 2, scale = c(4,0.5),
          colors = brewer.pal(8, "Dark2"),
          random.color = TRUE, random.order = FALSE, max.words = 150)
```

```
wordcloud(df_cleaned4$text_updated, min.freq = 2, scale = c(4,0.5),
        colors = brewer.pal(8, "Dark2"),
        random.color = TRUE, random.order = FALSE, max.words = 150)

wordcloud(df_cleaned5$text_updated, min.freq = 2, scale = c(4,0.5),
        colors = brewer.pal(8, "Dark2"),
        random.color = TRUE, random.order = FALSE, max.words = 150)

test = head(df_cleaned,20000)

df.corpus <- Corpus(VectorSource(test$text_updated))
tdm <- TermDocumentMatrix(df.corpus)
term.freq <- rowSums(as.matrix(tdm))

dataframe.term.freq <- data.frame(term = names(term.freq), freq = term.freq)

dataframe.term.freq %>%
  top_n(25)%>%
  ggplot()+
  geom_col(aes(x = reorder(term,freq), y = freq))+
  xlab("Terms")+
  ylab("Frequency")+
  coord_flip()

df_cleaned1 = df_cleaned1 %>% add_column(Topicname = "Pfizer")
df_cleaned2 = df_cleaned2 %>% add_column(Topicname = "Moderna")
df_cleaned3 = df_cleaned3 %>% add_column(Topicname = "J&J")
df_cleaned4 = df_cleaned4 %>% add_column(Topicname = "Nvax")
df_cleaned5 = df_cleaned5 %>% add_column(Topicname = "General")

test1 <- rbind(df_cleaned1, df_cleaned2, df_cleaned3, df_cleaned4, df_cleaned5)

test2 <- sample_n(test1,1000)
emotions <- get_nrc_sentiment(test2$text_updated)

test3 <- bind_cols(test2, emotions, id=NULL)

test3%>%
  ggplot()+
  geom_col(aes(x=Topicname, y=positive, fill=Topicname))+
  labs(title = "Positive Tweets Sample")

test3%>%
```

```
  ggplot()+
  geom_col(aes(x=Topicname, y=negative, fill=Topicname))+
  labs(title = "Negative Tweets Sample")

test3%>%
  pivot_longer(c(anger:trust), names_to = "emotions", values_to = "cases")%>%
  group_by(Topicname, emotions)%>%
  summarise(sumcases=sum(cases))%>%
  ggplot(aes(x=emotions, y=sumcases, fill=Topicname))+
  geom_bar(stat = "identity", position = position_dodge2())
```