COVID-19 Virus Spread Pattern Worldwide, Vaccine Rollout and Predicting Timelines for Herd Immunity

DS 5110 – Project – Group 01

By:
Narsing Rao Akula
Mu Cheng
Kavya Gajjar
Manas Purohit
Himani Tyagi

Summary

Overview

COVID-19 has completely changed how the world functions as it has run its course over the past year and a half. The virus, which originated in Wuhan, China, quickly spread to all corners of the globe and brought the world to a screeching halt with overrun hospitals, strict lockdowns, and mandatory masks. Finally, we saw glimmers of hope as various vacccines gained emergency use authorization in Decebmer 2020. Since then, countries have been in a race to get their population vaccinated with the United States emerging as a front runner. By looking at the data, we hope to provide predicitions of when the United States will reach “herd immunity” and life can go back to some semblance of normalcy.

Goals

Our goal is to first visualize the spread of COVID-19 across the world using chloropleths. Then, we would like to use the same methods to visualize the vaccination progress all over the world as well as zoom in on the vaccination progess in the United States. Finally, we wanted to predict dates for “herd immunity” using the ARIMA model on our time series data.

Data Description

The COVID-19 cases and vaccination data came from Our World in Data. The cases dataset contained 59 variables including location, date, cases, deaths, ICU patients, tests, etc. We mainly looked at the information on cases from this dataset to visualize the spread of the virus. For vaccination, there were two datasets: one for world-wide vaccinations with 19 variables and one for United States vaccinations with 14 variables. Several of these variables were useful to us such as the location, date, number of vaccinations, number of people vaccinated, etc. The last dataset we used was the US state population dataset to understand how many people would need to be vaccinated to reach “herd immunity”.

Methods

Preprocessing and Data Analysis

On January 9th 2020, WHO (World Health Organization) announced mysterious Corona Virus related pneumonia in Wuhan, China with about 59 cases. At around January 20th, some countries started COVID-19 testing. Hence the data we have starts from January 22 onwards.

Initially, the COVID-19 cases were limited to only few countries, with its origination in China with about 550 cases and other countries like Japan, USA, South Korea, Taiwan, and Thailand reporting 2-4 cases each. This is shown in *Apprendix figure*1 with green color reflecting minimum cases and red as higher cases. *Apprendix* figure2 represents the total cases till January 31st 2020. More cases were reported in India, Australia, Germany, and France and indicates some of the countries that had started early testing. February 2020 data (*Apprendix* figure3) shows that the situation got worst in China with about 80,000 cases. That is when most of the countries took the situation very seriously and started COVID-19 testing. This is reflected in the March 2020 data in *Apprendix* figure4. Almost every country in the world except few had detected some Covid cases in their countries. This showed that the pace with which the new virus was spreading across the world as well as more and more countries performing testing. It can also be noted that the number of cases in USA increased significantly to about 200,000, almost 3 times the number of cases in China. This could be attributed to the extensive testing USA was performing and more number of cases spreading across the country. Similar trends can be seen in coming months.

By August 2020, the 3 countries worst hit by COVID19 were United States, India, and Brazil. As the actual cases for every country ranged from few thousands to millions, for the sake of visualizing COVID-19 spread trend in all countries, a higher limit was put on choropleth legend that made the top 3 countries with most COVID cases gray out. This is shown in *Apprendix* figure6. Cases started building up in other countries like Russia, Europe, Middle Eastern and south American countries. Similar trend can be seen through January, February, March and recently in April 2021. Choropleths for the other months (May 2020 to February 2021) are also attached in the appendix.

In December 2020, several vaccines received emergency use authorization and countries quickly began administering the vaccines to their citizens in a race to get to herd immunity. In Appendix 2, there are several maps showing the average number of total vaccinations in that month as a percentage of the country's population. For December 2020, there are several countries that begin their vaccination programs including countries in North America, western Europe, Russia, China, Argentina, and Chile. In January 2021, several more countries start administering vaccines including India, Brazil, parts of northern Africa and southeast Asia. Feburary 2021 is when we see the US beginning to pull ahead as it is a distinctly darker shade of green. In this month, all the countries in South America and Europe have begun vaccinations. By March 2021, the US is much further ahead, along with Chile and the UK who are all darker shades of green than their neighbors. At this point, almost all of Asia also has some vaccines being adminstered. Finally, we arrive at present day in April 2021, where most of the world has some kind of vaccination program. The US and western Europe are clear front runners here with the rest of the world following suit.

For December 2020, there are several countries that begin their vaccination programs including countries in North America, India, Europe and many others. In Appendix 3, there are 2 plots a) Choropleth showing number of people being vaccinated across different states in the USA, b) Bar plot showing the number of people being vaccinated across United States of America. We notice that many states have started vaccinating their people across USA. There are few states where the people are being vaccinated heavily like Califronia, Texas & Florida, where as there are also few states which are vaccinating their people at a slower pace like Wyoming, Utah, Idoha & Motana. To conclude we aniticipate that very soon most of the people across United States would be vaccinated very soon.

Data Tidying

Based on our analysis, we decided to select the variable 'people_fully_vaccinated' as our target variable. We realized that the target variable has around 13% of 113 data points as 'NA' and taking the limited amount of observations in our dataset and requirement of keeping stationarity and linearity for the time series modeling into considerations, we chose to use na.interpolation to replace NAs.

Model Selection and Implementation

Our goal of the project is to forecast the 'herd immunity' timelines therefore finding the best model to interact with time series is the key. Based on our research, 'herd immunity' occurs when a substantial portion of a community (the herd) becomes immune to a disease, making the spread of disease from person to person unlikely, and since Coronavirus is a highly contagious diseases we decided to keep 90% of population as a threshold to measure 'herd immunity' timeline.

After analyzing the data, we realized that our time series data does not involve seasonal pattern, so we chose to implement ARIMA model in our project instead of SARIMA (Seasonal Arima). ARIMA is an acronym for Auto Regressive (AR) Integrated (I) Moving Average (MA) which indicates that an ARIMA model has three components to it namely:

- Auto Regressive Models (AR) models which are similar to a regression model but the regressor in this case is the same dependent variable with a specific lag.
- Differencing (I): For ARIMA to perform at its best it needs the data to be stationary. That means that the mean and variance are constant over the entire set. Differencing is used to transform the data so that it is stationary.
- Moving Average (MA) models which are widely used in time series analysis and is an already well-known concept. It incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

A standard notation for an ARIMA model with the order of p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

p: the number of lag observations in the model (i.e. the lag orders.) (AR)

d: the number of times that the raw observations are differenced. (i.e. the degree of differencing) (I)

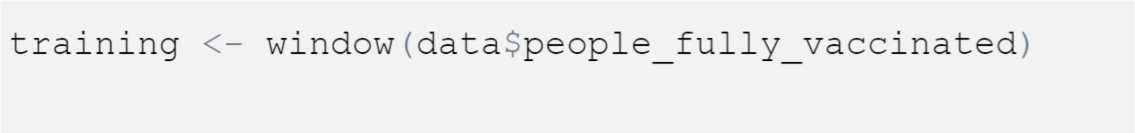
q: the size of the moving average window (i.e. the order of the moving average) (MA)

For our project, we used auto.airma() function of R, which returns best ARIMA model (i.e. best values of p,d,q) according to either AIC, AICc or BIC value. This function conducts a search over all possible model within the order constraints provided.

Our initial approach was to predict the 'herd immunity' timeline for the entire USA country, so we selected two particular variables 'date' and 'people_fully_vaccinated' from the country_vaccinations file only for the country USA.

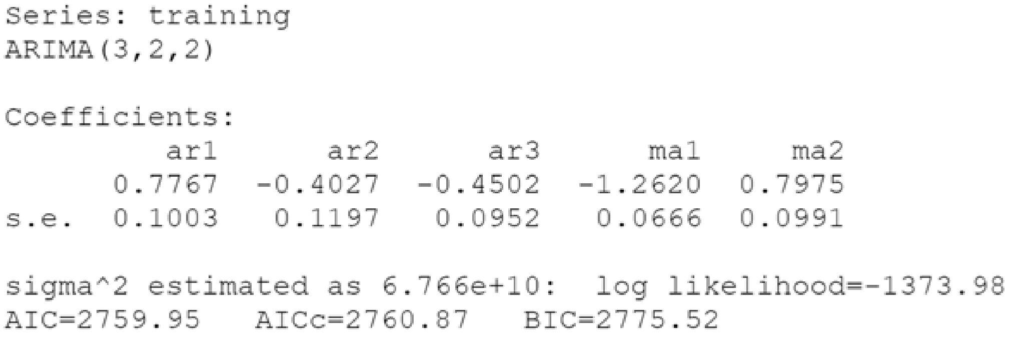
The following procedure was performed while modelling:

1. We converted the continuous data to time series data using the window() function in R



```
training <- window(data$people_fully_vaccinated)
```

2. After converting the data into time series format, we used auto.arima() model to predict the 'heed immunity' timeline. And based on our data, we an ARIMA model of order(3,2,2) i.e. Auto regression with order 3, Integration of order 2 and Moving average of order 2.

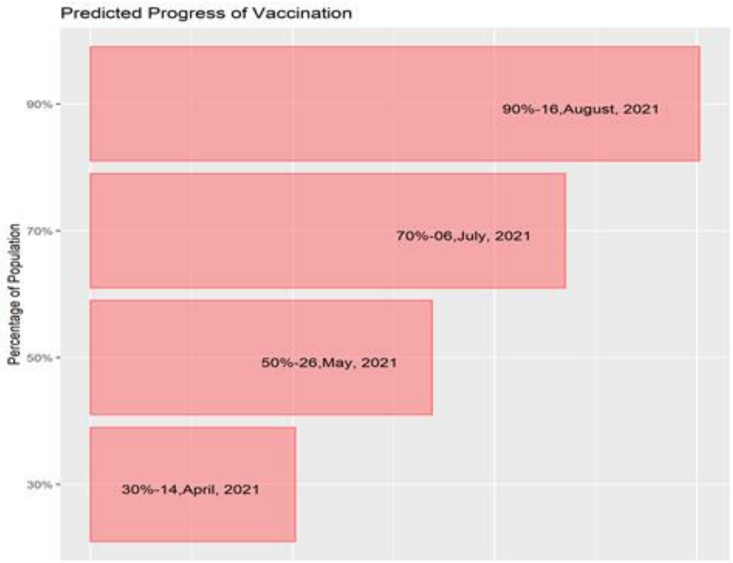


```
Series: training
ARIMA(3,2,2)

Coefficients:
         ar1      ar2      ar3      ma1     ma2
      0.7767  -0.4027  -0.4502  -1.2620  0.7975
s.e.  0.1003   0.1197   0.0952   0.0666  0.0991

sigma^2 estimated as 6.766e+10:  log likelihood=-1373.98
AIC=2759.95   AICc=2760.87   BIC=2775.52
```

3. Now, for predicting the date on which 'herd immunity' will be achieved, we calculated the value of herd population i.e 90% of the entire US (United States) population (~301,084,071), and as soon as the prediction value exceeds the herd population, we can say the 'herd immunity' is achieved.

Results & Discussions

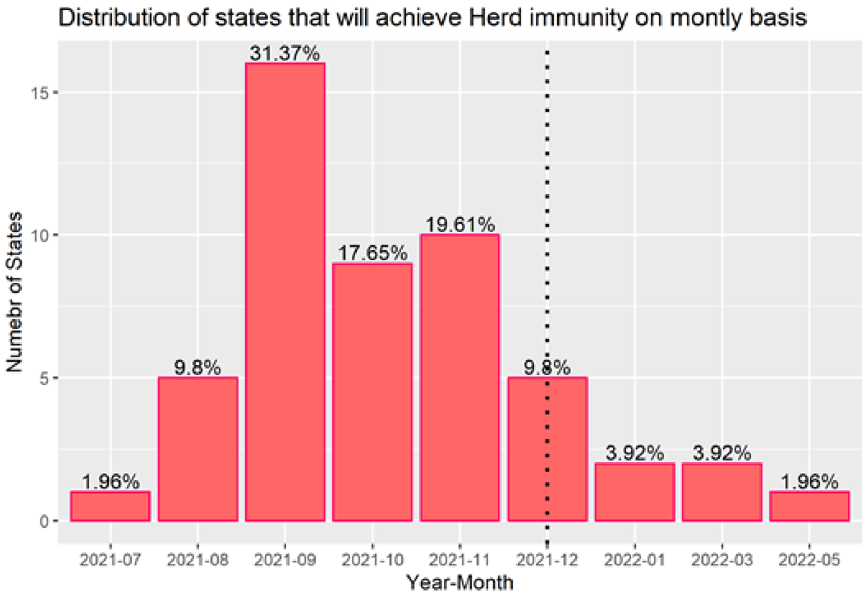
After implementing the ARIMA model on USA country to predict the 'herd immunity' timelines, we got the results shown in Figure 1, basically USA will have 30% of population vaccinated by 14th April, 2021, 50% by 26th May, 2021, 70% by 6th July, 2021 and 90% of population by 16th August, 2021. And as we mentioned earlier, since coronavirus is highly contagious disease, we selected 90% as our threshold for considering 'herd immunity'.



[Figure 1: Progress of covid vaccination rollout in USA with respect to its population]

But, since the vaccination progress in USA varies from state to state due to difference in populations, phases, efficiencies of vaccine distributions, capabilities of healthcare facilities and variuos other factors, we deciced to take a step further and predict the 'herd immunity' timeline for each state using the same procedure mentioned above.

After implementing the same procedure at state level, approximately 43% of states i.e. (21 states) will achieve the 'herd immunity' by the end of August, 2021 and approximately 90% of states i.e. (45 states) will achieve the 'herd immunity' by the end of December, 2021. The above mentioned results are presented in the Figure 2.



[Figure 2: Number of states who achieve the 'herd immunity' on Year-month level]

Limitations of Analysis & Future Work

Considering the recently started COVID-19 vaccine rollout, the limited amount of data points, approximately 3 months (20 Dec,2020 - 11 April, 2020), is not enough to build an accurate time series model to predict herd immunity timelines. Hence, in the furutre we would like to perform the ARIMA model again with the updated data after a few more months of vaccine rollout, it will give a better and more accurate predictions of the herd immunity timelines across the globe. Furthermore, since the vaccination progress varies from state to state, looking into more detailed factors that could possibly impact on each state's rollout plan such as healthcare provider/vaccine register ratio, polulation(number of residents, immigrants, un-documented, homelss), and etc could be beneficial to make better predictions. Therefore, future work also contains research on the aforementioned factors. We believe as we have more data and research findings collected, our ARIMA model will generate more accurate herd immunity timelines.

Statements of Contribution

Narsing Rao Akula: US state wise covid vaccination progress analysis and data tidying

Mu Cheng: Time series forcast (ARIMA) model, Time series background Research & implementation

Kavya Gajjar: Time series forcast (ARIMA) model, Time series background Research & implementation

Manas Purohit: Global covid vaccination progress and data tidying

Himani Tyagi: Covid-19 spread world wide analysis and data tidying

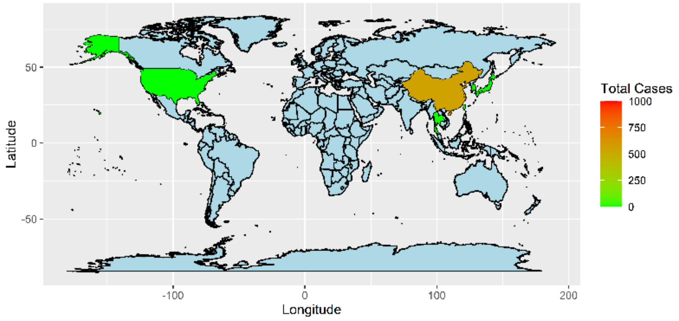
References

1. Covid-19 progress data: https://covid.ourworldindata.org/data/owid-covid-data.csv
2. Covid-19 Global Vaccination data: https://www.kaggle.com/gpreda/covid-world-vaccination-progress
3. Covid-19 USA state wise Vaccination data: https://www.kaggle.com/bumjunkoo/us-vaccination-progress
4. USA state wise 2021 Population data: https://worldpopulationreview.com/states

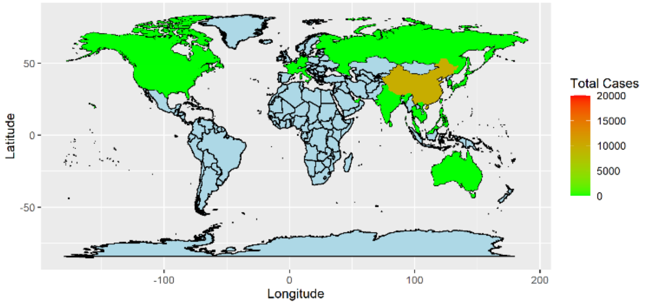
Appendix

Appendix 1:

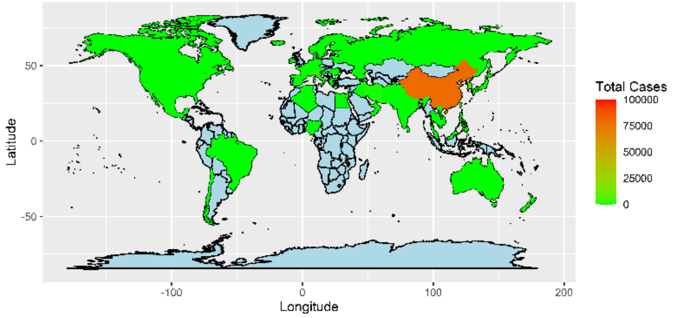
Total Covid-19 Cases - **22 January 2020**



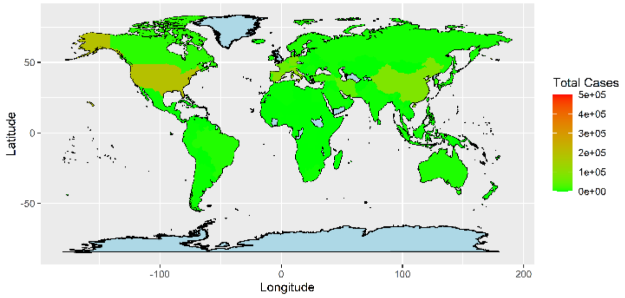
Total Covid-19 Cases - **31 January 2020**



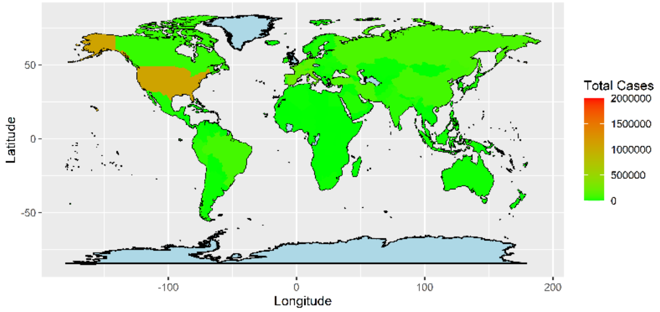
Total Covid-19 Cases - **February 2020**



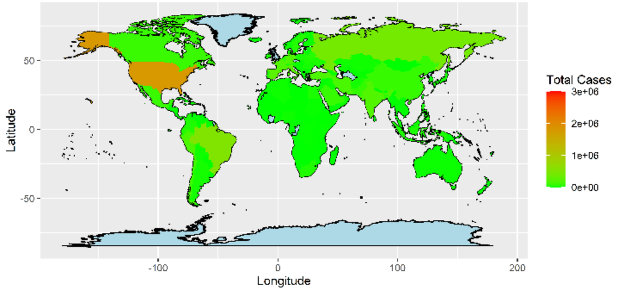
Total Covid-19 Cases - **March 2020**



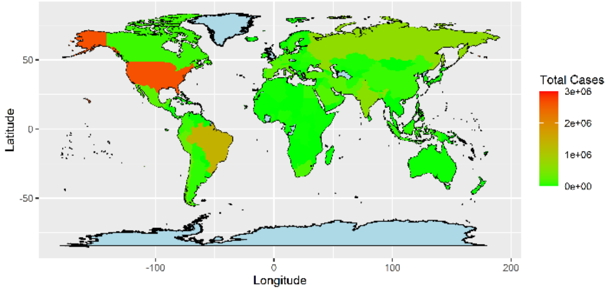
Total Covid-19 Cases - **April 2020**



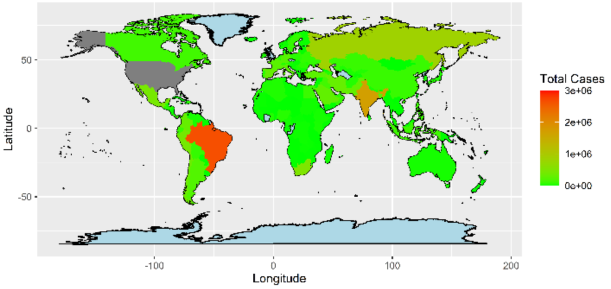
Total Covid-19 Cases - **May 2020**



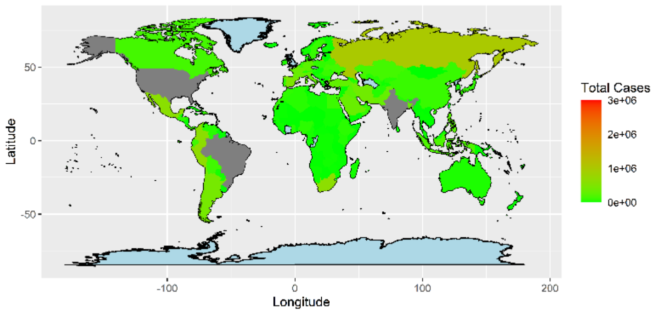
Total Covid-19 Cases - **June 2020**



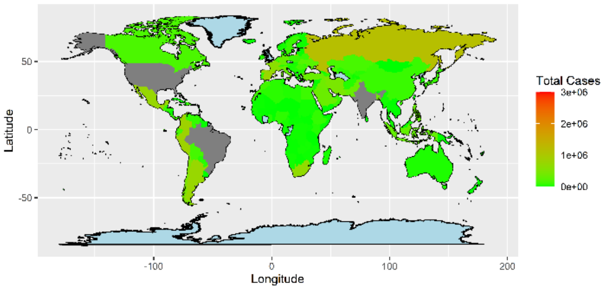
Total Covid-19 Cases - **July 2020**



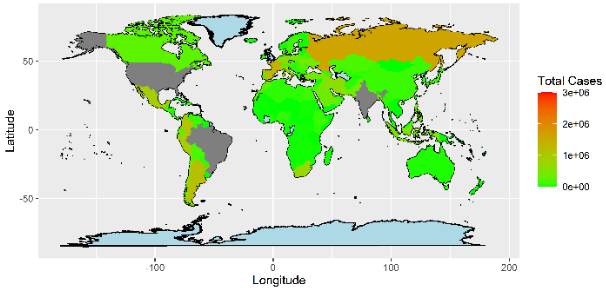
Total Covid-19 Cases - **August 2020**



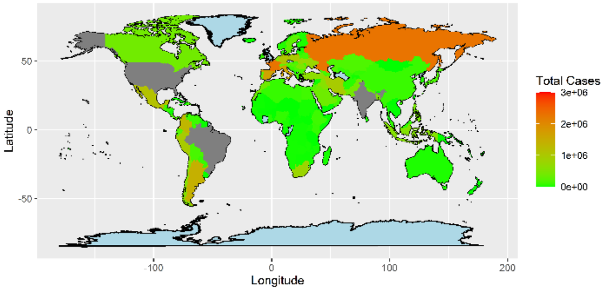
Total Covid-19 Cases - **September 2020**



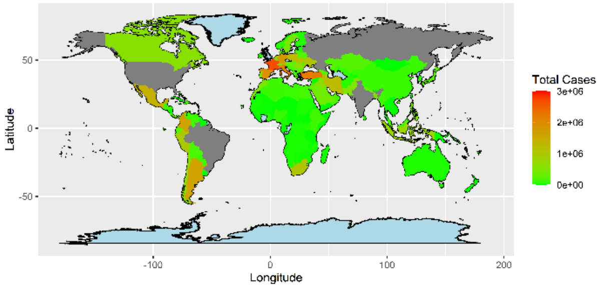
Total Covid-19 Cases - **October 2020**



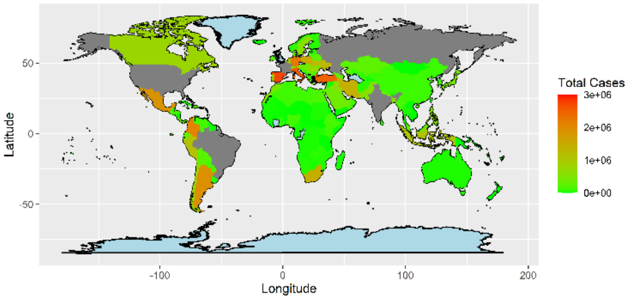
Total Covid-19 Cases - **November 2020**



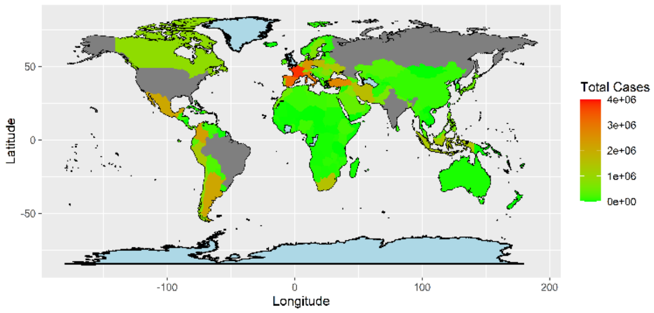
Total Covid-19 Cases - **December 2020**



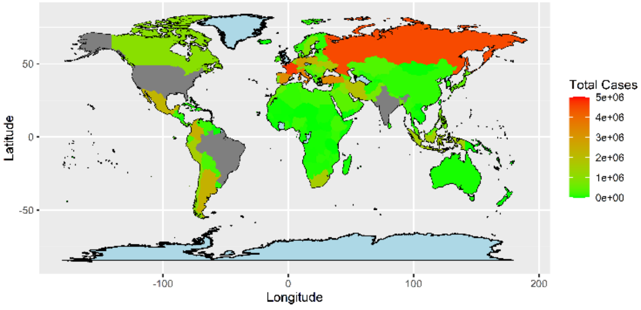
Total Covid-19 Cases - **January 2021**

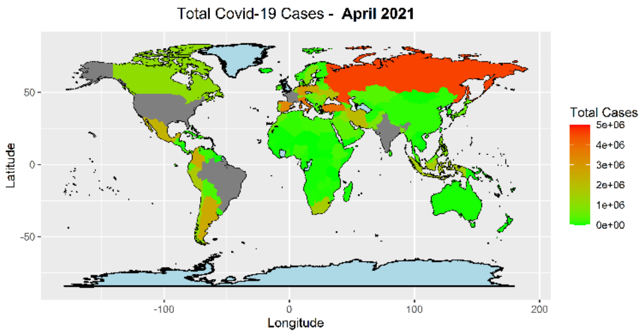


Total Covid-19 Cases - **February 2021**



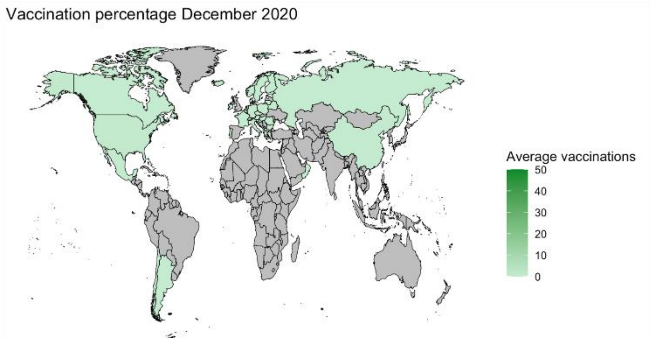
Total Covid-19 Cases - **March 2021**

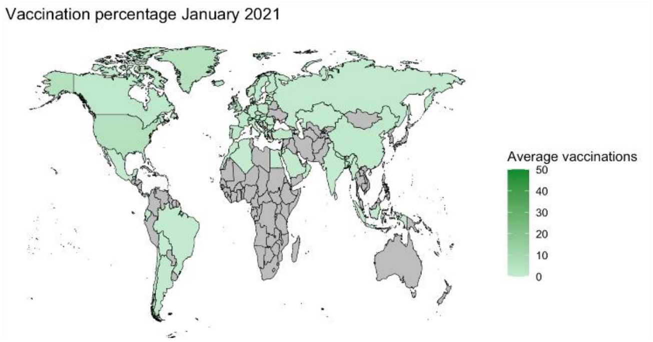


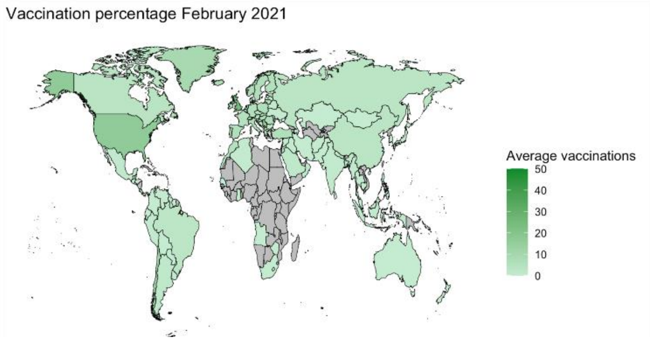


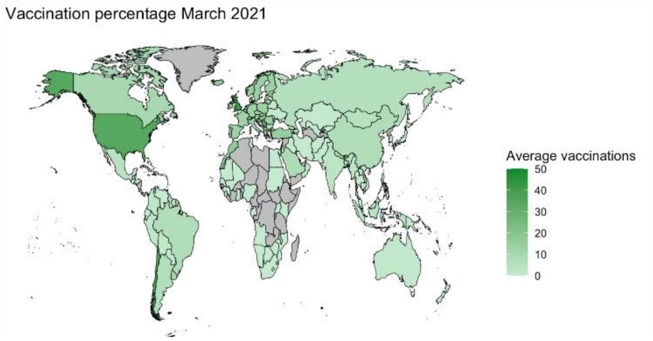
Choropleths depicting Total Number of COVID-19 cases across the world.

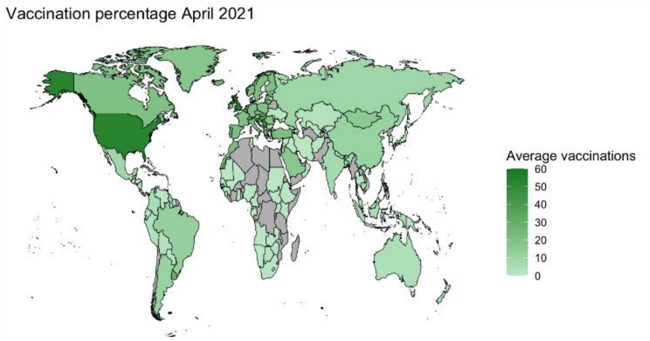
Appendix 2:



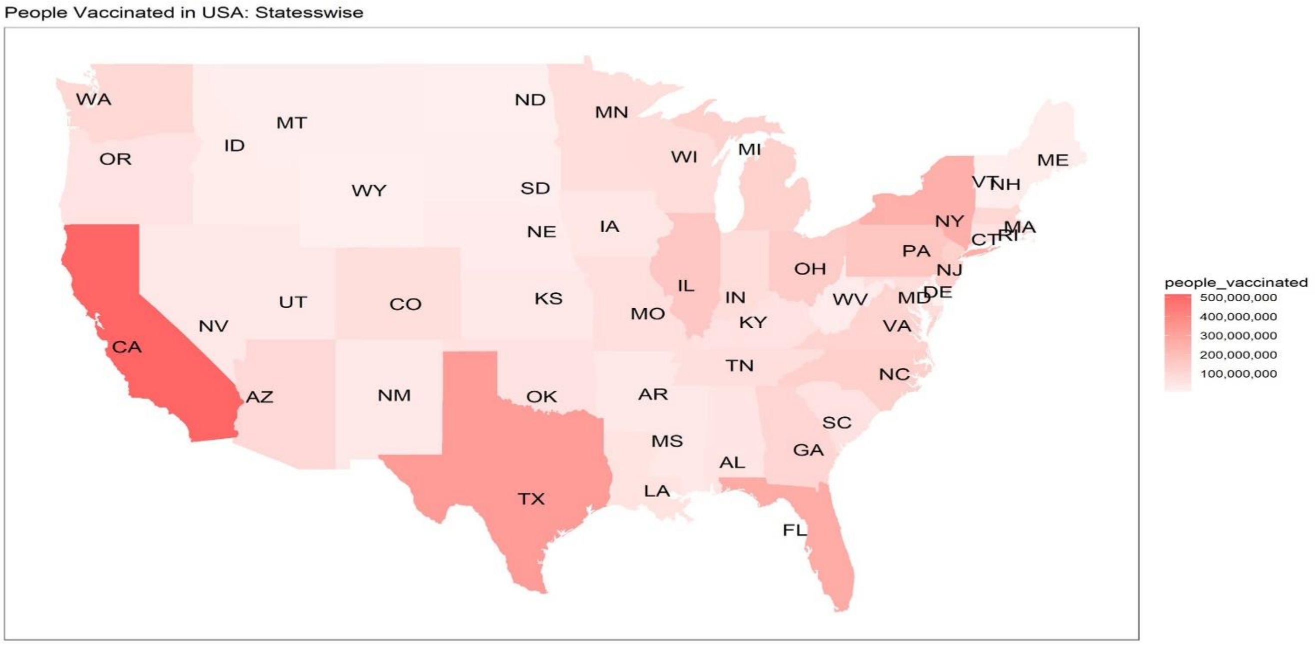


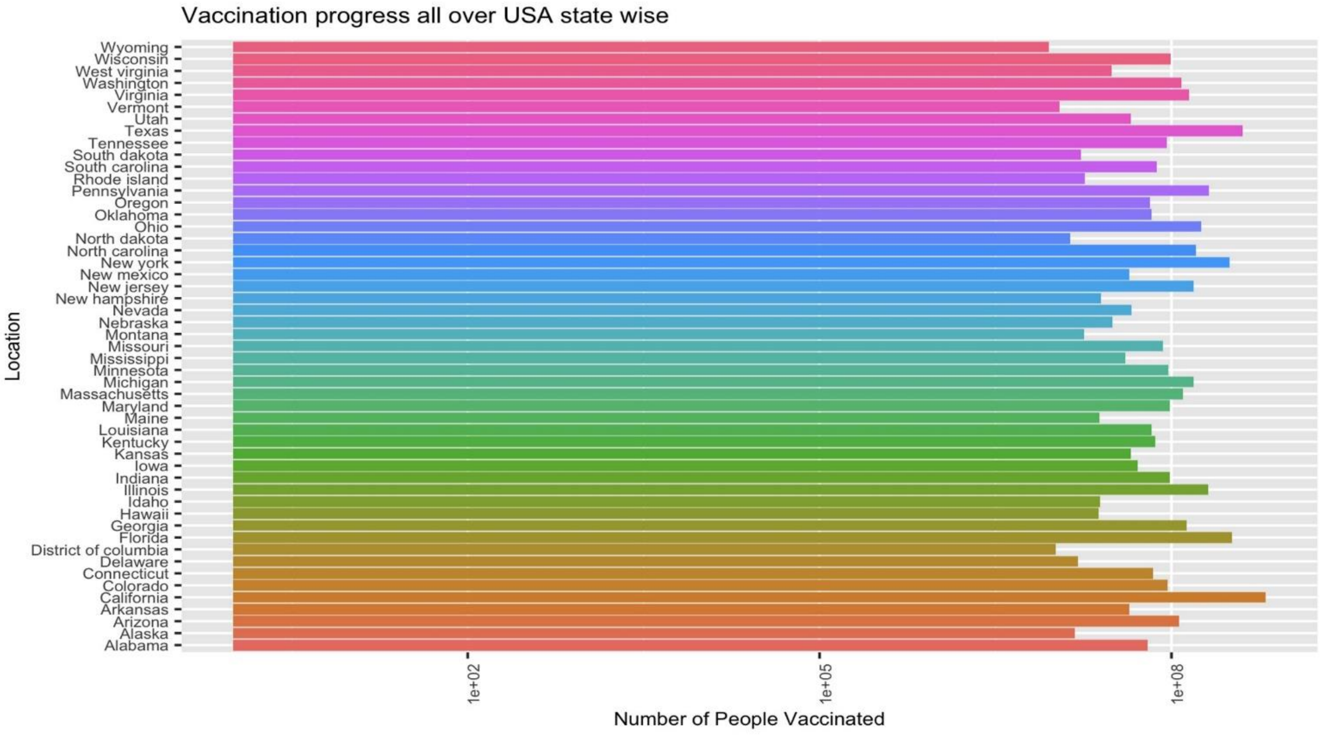






Appendix 3:


People Vaccinated in USA: Statesswise
WA
MT
ID
OR
WY
ND
MN
SD
WI
MI
NE
IA
NY
VT
NH
ME
MA
CT
RI
PA
NJ
OH
IL
IN
WV
MD
DE
VA
KY
UT
CO
KS
MO
NV
CA
TN
NC
AZ
NM
OK
AR
SC
MS
GA
AL
TX
LA
FL
people_vaccinated
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000


Vaccination progress all over USA state wise
Wyoming
Wisconsin
West virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South dakota
South carolina
Rhode island
Pennsylvania
Oregon
Oklahoma
Ohio
North dakota
North carolina
New york
New mexico
New jersey
New hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
District of columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama
Location
1e+02
1e+05
1e+08
Number of People Vaccinated