Case Study I Group I Report

Cathy Shi, Mariana Izon, Yuxuan Chen, Peyton Chen

9/9/2021

Introduction

According to World Health Organization, more than 222 million COVID-19 cases are confirmed and more than 4 million people died due to COVID-19 globally as of September 6, 2021 (WHO Coronavirus (COVID-19) Dashboard). Due to the highly contagious nature of COVID-19, people have been negatively impacted across every major aspect of their lives. The U.S. Food and Drug Administration issued an emergency use authorization for the Pfizer-BioNTech vaccine and the Moderna Vaccine in December 2020 (Kriss JL, Reynolds LE, Wang A, et al.). However, according to a survey conducted by AP-NORC in July 2021, about 30% of adults in Americans are still not very confident or not confident at all about the vaccines (AP-NORC Center for Public Affairs Research).

In this project, we conducted a preliminary exploration of the relationship between COVID-19 cases, deaths due to COVID-19, and vaccination rates on the global level. Specifically, we built an interactive R Shiny app that allows users to visualize COVID-19 cases, death cases, and the population that has been fully vaccinated for each country on the world map. Users would be able to visualize the data on a specific date from January 1, 2020, to August 1, 2021. With the visualization, users would be able to explore the temporal trend of COVID-19 cases, death cases due to COVID-19, and their correlation with vaccination rate.

Data Acquisition and Data Cleaning

To investigate this question, we found a data set from *Our World in Data*, which is a project collaborated by researchers at the University of Oxford and Global Change Data Lab with a mission to make knowledge and data more accessible for research. Global Change Data Lab has received grants from the Quadrature Climate Foundation, the Bill and Melinda Gates Foundation, the World Health Organization, and the Department of Health and Social Care in the United Kingdom (Our World in Data - About). The data set contains time-series data of new COVID-19 cases, cumulative COVID-19 cases, new death cases, cumulative death cases, the population that has been fully vaccinated, and other covariates that can be helpful for our analysis for each country on each date since the outbreak of COVID-19 in 2020. The data set is updated daily, so we can obtain timely data about COVID-19 from this data set. We chose this data set because it has contains data related to COVID-19 and certain country-specific data for each country on each date. We also used the *world map* data set in the *ggplot2* package in R for visualization in the R Shiny app.

As we started to explore the COVID-19 data set, we found many missing data from the beginning of 2020 towards the end of the first quarter of 2020. This is reasonable since COVID-19 might not have been transmitted to certain regions or the local government was unaware of COVID-19 cases in the communities. To handle this situation, we filled the rows as 0 for new COVID-19 cases and new deaths due to COVID-19. For columns that show cumulative numbers, such as `total_cases` and `people_fully_vaccinated`, we filled the rows as 0 for each country up to the date that does not contain missing data.

Next, we separated the data at the country level from the aggregated data at the continent level. We used the country-level data to visualize COVID-19 cases, death cases, and the vaccination population on the world map; we used the continent-level data to visualize the trend of COVID-19 cases and death cases at the continent level.

We also used the *world map* data set in the *ggplot2* package in R. To visualize the COVID-19 data on the world map, we need to merge the COVID-19 data set to the world map data set. However, we have more than one row of data to represent a country on the world map since the world map data set contains the latitude and longitude for different regions within the countries. This poses a challenge for merging the data sets since the COVID-19 data set contains COVID-19 data for each country at different dates, and therefore, it also contains more than one row of data for each country. With the suggestion of our teaching assistant, Bo Liu, we overcame this challenge by filtering the COVID-19 data set by date based on the user's input and left join the data set to the world map.

In addition, we found mismatches of country names between the COVID-19 data set and the world data set. For example, the country name for the United States of America in the COVID-19 data set is "United States", but in the world map data set is "USA". This poses another challenge for joining the data sets. We manually renamed countries that have reported a large number of COVID-19 cases or have a relatively large geographic area on the world map in the COVID-19 data set before joining it to the world map data set.

In the R shiny app, we can see that there are still missing values between January 2020 and April 2020. This might be because we do not have COVID-19 related reports for certain regions in early 2020. But we can assume those missing values to be zeros and that would not affect our analysis.

Visualization and Insight

In the first tab, we have maps that illustrate the number of new cases, new deaths, and people fully vaccinated in different parts of the world. Users can zoom in to a specific continent they are interested in by selecting the continent in the sidebar panel. Users can also move the slider to select the date they wanted to visualize on the map. We put these three maps side by side so that users can see the changes in new cases, death cases, and vaccinations simultaneously as they move the date slider. The table below the maps provides the actual number of new cases, new deaths, and vaccinations. Users can view this table in ascending order or descending order or search for a certain country in the world or the selected continent. The table allows users to look at the data in detail in a specific region on a particular date.

The trend plots show the new cases, new deaths, and vaccination counts over time, though we excluded the plots for number of people hospitalized due to a large amount of missing data in places other than North America. For the new cases, there were three main peaks worldwide across the span of the timeline since Jan 2020, while the continents' peaks varied a lot in time and magnitude. The death counts varied even more across locations and we couldn't observe an obvious pattern in the aggregated worldwide plot. On the other hand, the trend lines for people fully vaccinated across different continents were similar, where most of the curves took off shortly after Jan 2021. In Asia, South America and Europe, we saw that after vaccination number increased, new cases decreased as expected, but worldwide we saw no correlation between new cases and vaccination; and for continents such as North America and Australia, there was even a positive correlation between the two variables towards the end of the time variable.

The proportion plots illustrate the proportion of new cases versus the proportion of the population vaccinated, and the proportion of new deaths versus the proportion of the population vaccinated. Worldwide, we saw an initial decrease in both the ratio of new cases and new deaths as vaccination increased, however both curves ended in a spike upwards. Zooming in on continents, we observed similar patterns in both Asia, North America, and Europe. In South America, we observed a steady decrease in both cases and deaths as more people became vaccinated. However, in Australia, we observed an increase. Africa proved to have too little data in order to observe an interpretable graph. Overall, we believe this to indicate that as a greater percentage of the population became vaccinated, the proportion of new cases and deaths decreased - up until the delta variant appeared in certain continents.

Takeaway

For our project, we used worldwide COVID-19 related data collected from various sources to create a shiny app in order to visualize the correlation between variables such as new COVID cases, death numbers, and

vaccination rate. The app allowed us to see the trends of the different variables by continent as well as worldwide. We thought that the increase of vaccination would lead to a decrease in new cases and new deaths, however this only held true on a general level until the delta variant hit. Hence, via visualization alone, we are not sure how vaccination helps with mitigating the pandemic. Furthermore, if we wanted to assess the vaccination's effectiveness, much more rigorous statistical analysis would come into play.

In the future, we could let the shiny app take live data and work on animating the app, so we could see the map change over time. We could also incorporate more accurate hospitalization data with less missing values.

R Shiny App

The R shiny app can be found in the `app` folder of our Github repository: https://github.com/STA540-21Fall/Project1-Group1

Reference

AP-NORC Center for Public Affairs Research. (July, 2021). "Many Have Doubts about COVID-19 Vaccine Effectiveness against New Strains" https://apnorc.org/projects/many-have-doubts-about-covid-19-vaccine-effectiveness-against-new-strains/

Kriss JL, Reynolds LE, Wang A, et al. COVID-19 Vaccine Second-Dose Completion and Interval Between First and Second Doses Among Vaccinated Persons - United States, December 14, 2020-February 14, 2021. MMWR Morb Mortal Wkly Rep 2021;70:389–395. DOI: http://dx.doi.org/10.15585/mmwr.mm7011e2external icon.

"Our World in Data - About." *Our World in Data*, ourworldindata.org/about.

"WHO Coronavirus (COVID-19) Dashboard." World Health Organization, *World Health Organization*, covid19.who.int/.