

Analyzing rate of SARS-CoV-2 Cases and Deaths in Five Different Brazilian States

Aavikumar Bhakta

15 December, 2021

Background and Overview

The World Health Organization (WHO) labeled COVID-19 as a pandemic on March 11, 2021, and since then the number of cases and deaths in countries around the world have continued to rise (Silva and Pena, 2021). Brazil recorded its first case on February 26, 2020, in the city of São Paulo which is the most populous city and state in Brazil (Silva and Pena, 2021). Since then, there have been first waves, second waves, and now third waves as a result of variants of SARS-CoV-2. Lack of proper safety precautions taken within certain countries like Brazil also attributes to the number of cases and deaths (Nikas, 2021). Unfortunately, Brazil is in the top 10 of countries in the world in cases, deaths, and deaths per 100,000 people (Mortality analyses, 2021). These are not categories that any country would want to be associated with being in the top 10 of. Much controversy surrounds the federal response to the COVID-19 pandemic namely by the President of Brazil, Mr. Jair Bolsonaro (Nikas, 2021). According to the a New York Times article, he did not use science as the basis of his decisions in attempting to protect the country. He tried pushing drugs that had not been scientifically proven to be effective, such as hydrochloriquine, rather than implementing social distancing and masks (Nikas, 2021).

I chose to work with this data because I wanted to find out how each state was affected differently within Brazil, and if they were impacted worse by the spread of the P.1 variant or Gamma variant that was detected in November 2020 and labeled as a variant of concern by the WHO on December 29, 2020 (). This variant is able to evade neutralizing activity of mRNA vaccines by 4.5 fold which can be attributed to its 35 nucleotide mutations also making it more transmissible (Silva and Pena, 2021). This dataset selected from the NCBI database analyzes five of Brazil's states, Amazonas, Rio Grande do Norte, Rio de Janeiro, Bahia, and Paraíba. It only included a short timeframe of December 2020 through February 2021, however, there did seem to be more cases in late January into February. I could not tell if that was because of the Gamma variant, but that is what the data showed and I made a hypothesis that that is a possible reason why cases went up then. I would have to conduct another analysis specifically on the SARS-CoV-2 virus and see whether the virus that was detected in Brazil during rise in cases in February had the same mutations as the Gamma variant.

Methods

SARS-CoV-2 Sequencing Data

I downloaded a Brazilian dataset that included data specifically from five different states on November 20, 2021. I found that data in the BioProjects Resources database of NCBI was labeled as Genomic Surveillance of SARS-CoV-2 in different Brazilian states. This data only included data that was read on an Illumina sequencing machine. It included 185 runs and 10.65Gb of data taken from Brazil between December 2020 and February 2021.

Creating tables and plots

Each plot that was created in the figures section was made by parsing the original Brazil dataset file in numerous ways. First, the entire dataframe was cleaned to make the headers and titles all the same format. Figures 2 and 3 were made by removing the states and towns from the `state_in_brazil` column so that only states would be present in the column making it easy to group them together and make a graph that compares them all using appearances in the dataset (Figure 2), and in sex and age (Figure 3). Figures 4 and 5 were made using a parsed data table that separated all the states besides the ones that I wanted to analyze further, Rio de Janeiro and Rio Grande do Norte. The parsing for these plots did not remove the cities because I wanted to see how each one compared to each other in their respective states. Figure 6 uses a data table that Figure 2 and 3 use, but parses it further to only include states and collection dates.

Packages that were used

The packages that I used to parse and categorize the data were `dplyr` and `tidyr` (Wickham *et al.*, 2021; Wickham, 2021) and I also used `janitor` (Firke, 2021) to clean the dataframe making proper headers and titles. `Magrittr`, `kableExtra`, and `stringr` were used to make the data tables and pipe them (Bache and Wickham, 2020; Zhu, 2021; Wickham, 2019). Graphs and colors used to analyze the data tables visually used the packages `hrbrthemes`, `viridis`, and `ggplot2` (Wickham, 2016; Arnold, 2021; Rudis, 2020). `Readr` was used to read in and write the .csv files from the data folder (Wickham and Hester, 2021).

Results and Discussion

People Infected with SARS-CoV-2 from Different Brazilian States

I graphed all the states against each other to determine whether the information given in the dataset from December 2020 through February 2021 was correct compared to what more general data on Brazil says from Our World In Data (OWID) that was acquired from GitHub. The states of Rio de Janeiro and Bahia are two of top five most populous states in Brazil and according to the data the age of most people in the dataset who were infected was around 40 years old (Figure 3) (2021). This is much lower than the age of people from Rio Grande de Norte and Paraiba, and higher than Amazonas. These three countries are in the bottom 15 of populous states in Brazil (2021). When broken down even further between the least and most population states, Rio Grande do Norte and Rio de Janeiro respectively, have drastic difference in age groups that were infected with the virus (Figures 4 and 5). Overall between all the states sex is not something that was a factor in infections while age was. Age and population is a connection found throughout these figures. The last figure, figure 7, uses data from OWID and parses it to only include Brazil data then groups the table into only total cases and deaths per million to compare the two data against each other on the same graph. Total deaths per million seems to slowly increase, but it only appears that way because the Y-axis scale is so large so that total cases per million can fit on it. They both increase, and seem to increase faster once late 2020 comes around possibly because of the Gamma variant.

Affect of the P.1 (Gamma) Variant in Brazilian States

The P.1 variant was declared as a variant of concern (VOC) by the WHO on December 29, 2020 (). However, it was first detected in the capital of Manaus in the Amazonas state on January 12, 2021 (Silva and Pena, 2021). I was curious about how the introduction of this variant would affect the number of cases. It was difficult to come to a concrete conclusion because the dataset only included a small timeframe, December 2020 through February 2021. According to the data, most of the cases became more prevalent at the beginning of January once the Gamma variant was deemed a VOC (Figure 6). This correlates with the higher transmission rate of the Gamma variant resulting in more cases (Silva and Pena, 2021). The deaths

per million began to rise quicker as the variant came along (Figure 7) according to data acquired from the OWID database ().

Mutations in the SARS-CoV-2

There are a couple of significant mutations in the SARS-CoV-2 sequences found in Brazil during the time that the data was collected late in 2020 and early 2021. The mutations, or SNPs, are within the N and S genes of the virus that (Figure 1). The N and S genes are also the longest which makes sense because the longer the gene is the more likely that there will be mutations within it as it goes through the process of transcription and translation into a protein (Table 1).

Figures

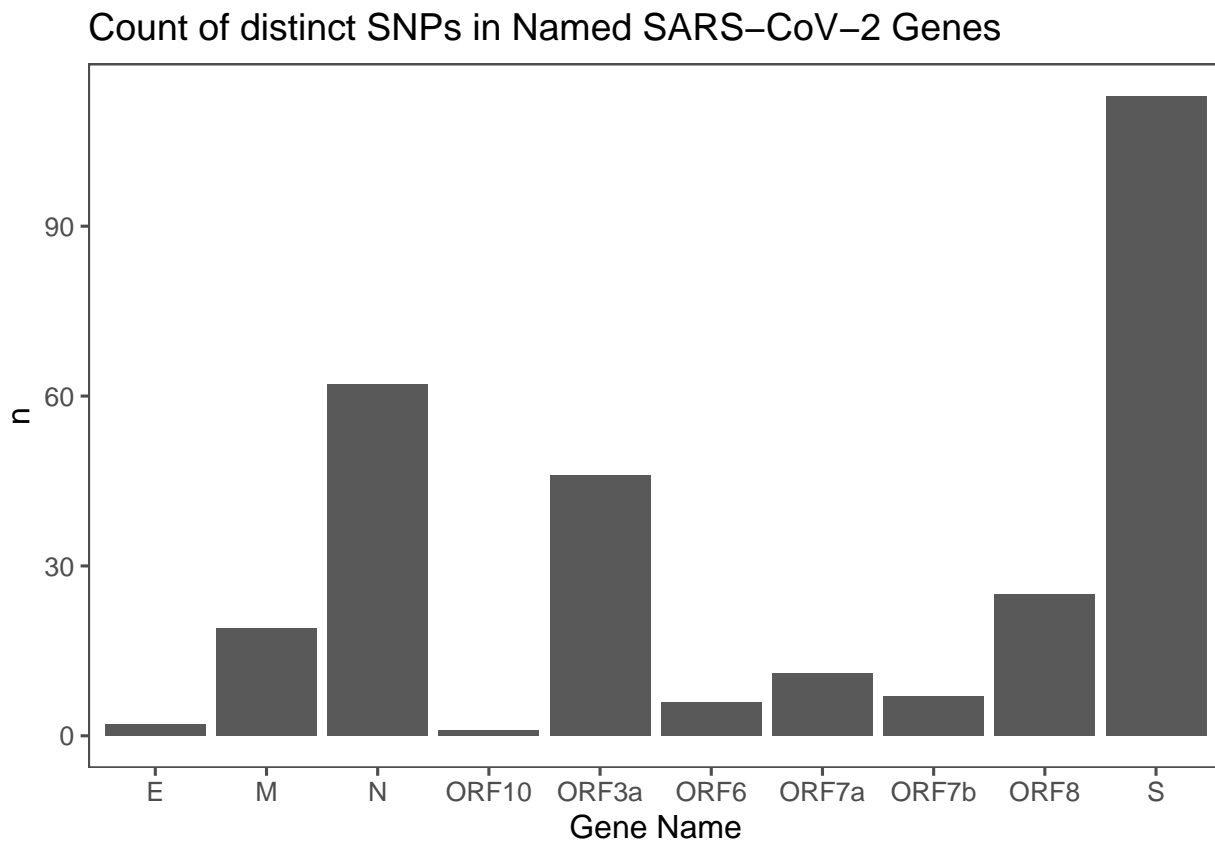


Figure 1: N and S genes have the most unique SNPs in the set of samples analyzed.

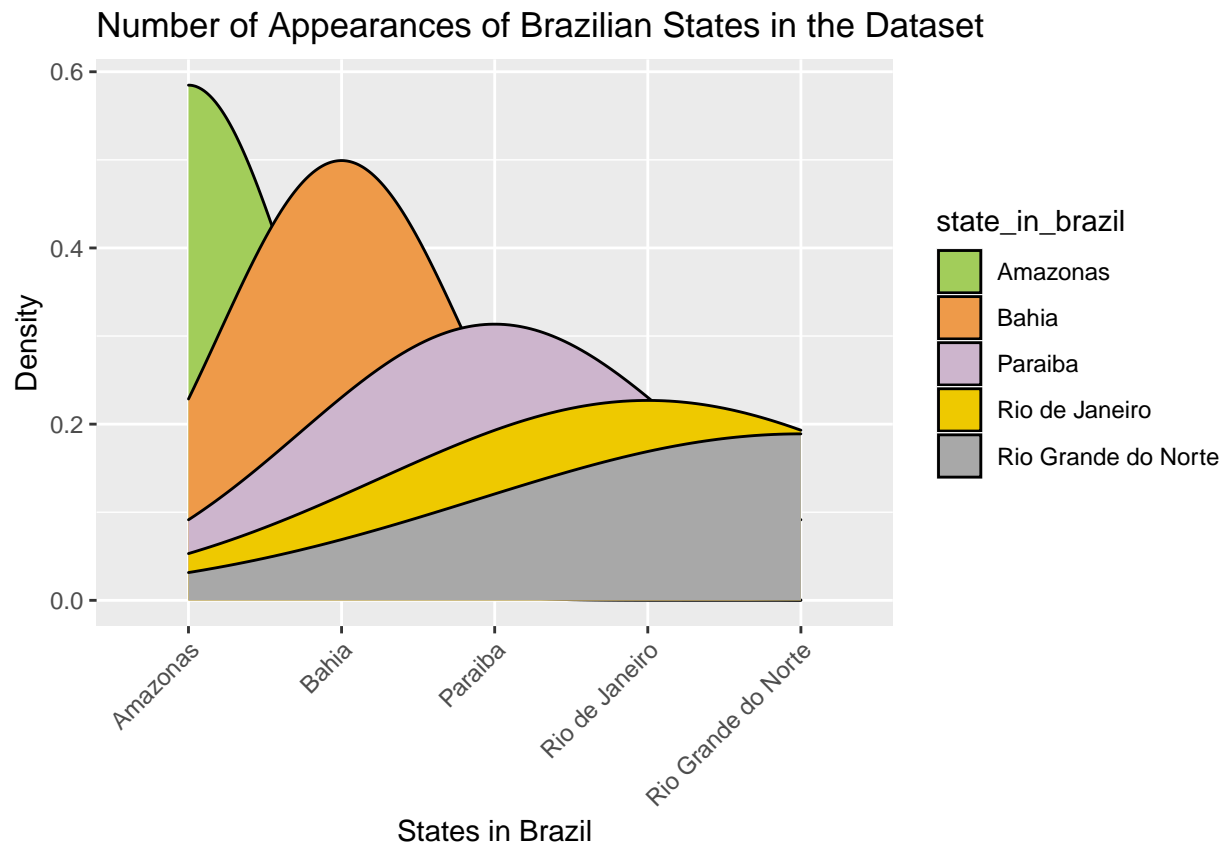


Figure 2: Amazonas and Bahia represent most of the state data from the dataset.

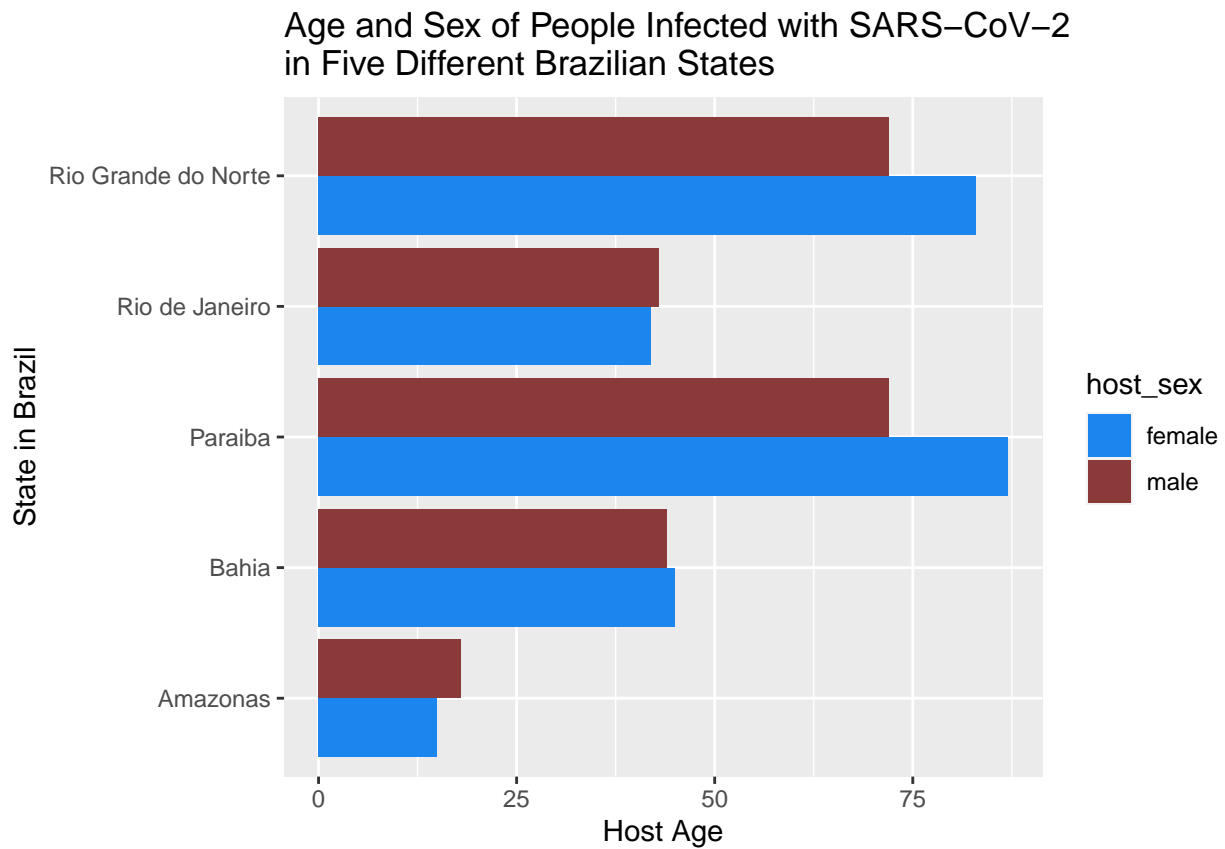


Figure 3: The most populous states, Rio de Janeiro and Bahia, have middle aged populations that were infected with SARS-CoV-2 and least populous states have either an average high or low age of people infected with SARS-CoV-2.

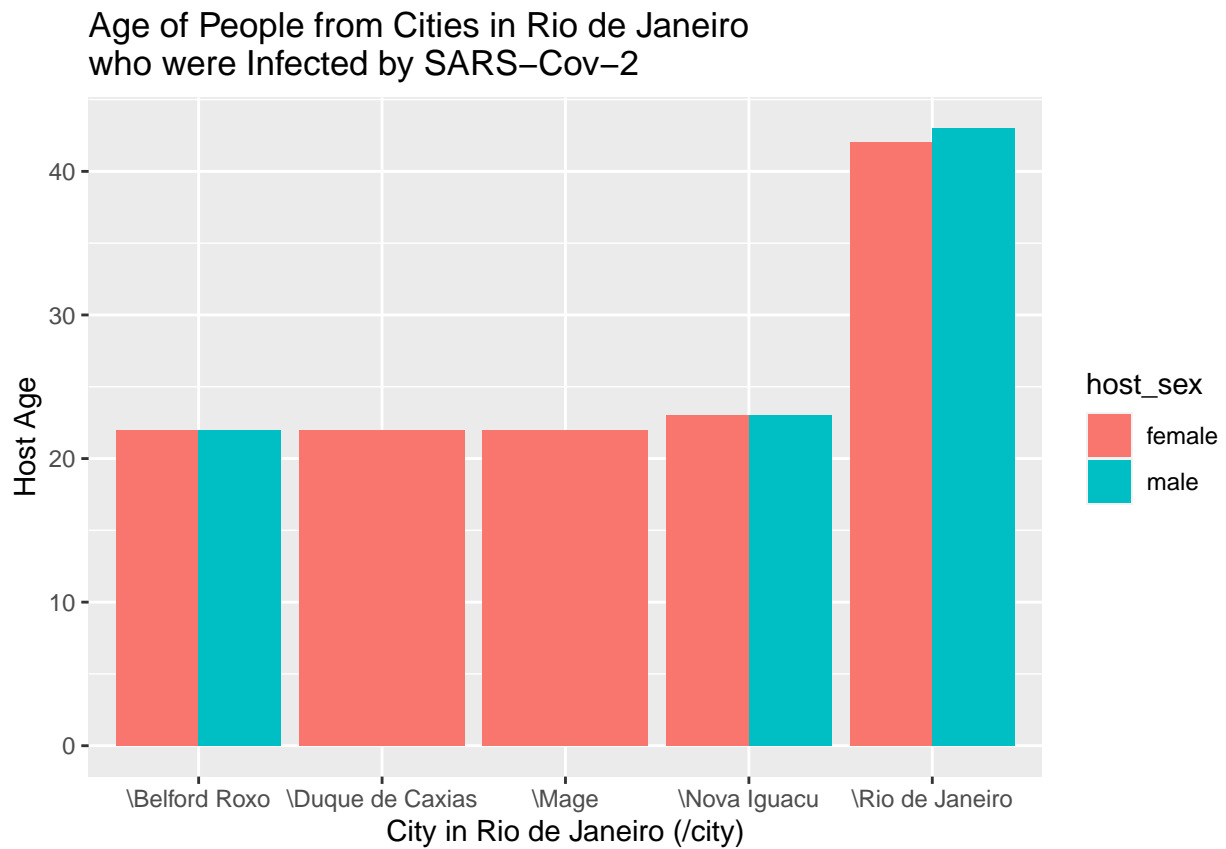


Figure 4: Rio de Janeiro has the highest age of people infected with the virus, but on average the cities within Rio de Janeiro have an age of around 20 years that were infected.

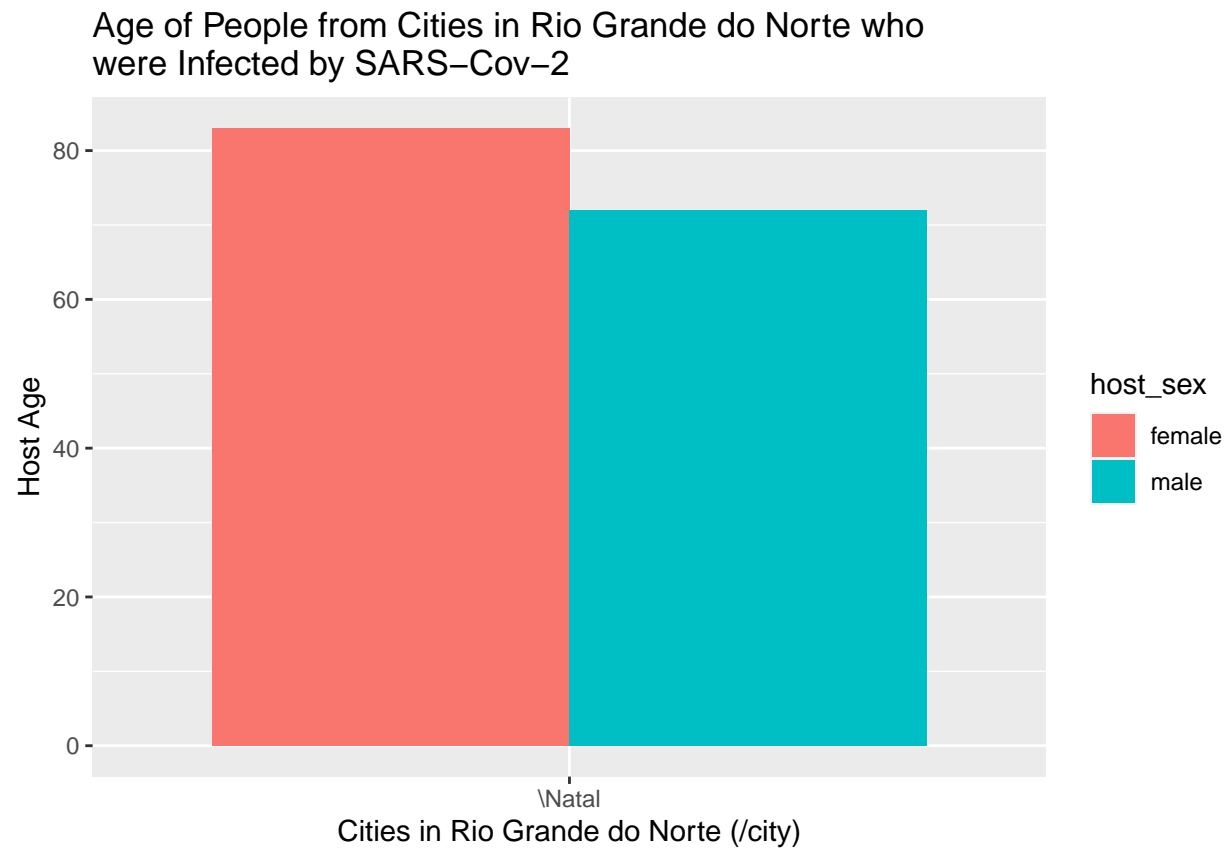


Figure 5: Rio Grande do Norte has only one city in the dataset and the age of people infected with the virus in the city was high compared to the higher populous cities and states.

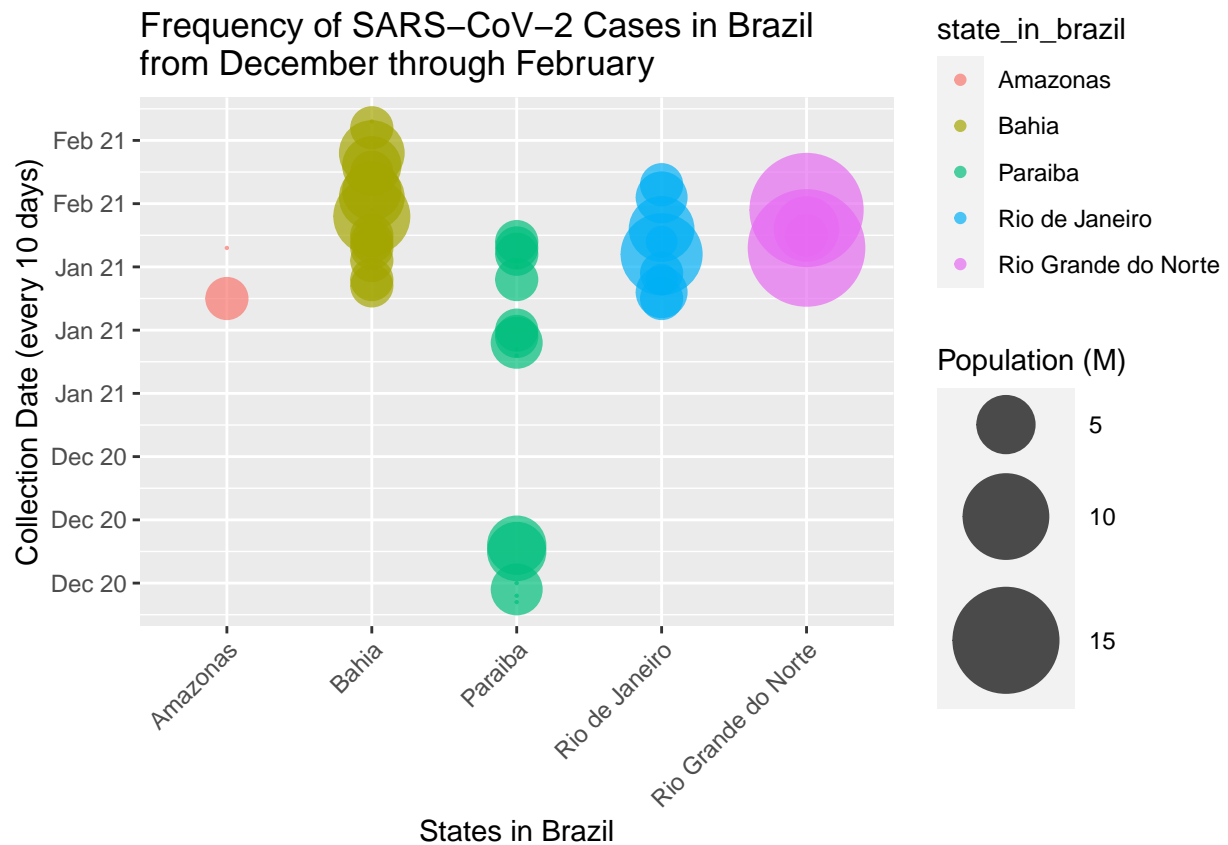


Figure 6: Infections began to rise after the Gamma Variant was detected in Manaus on January 12, 2021.

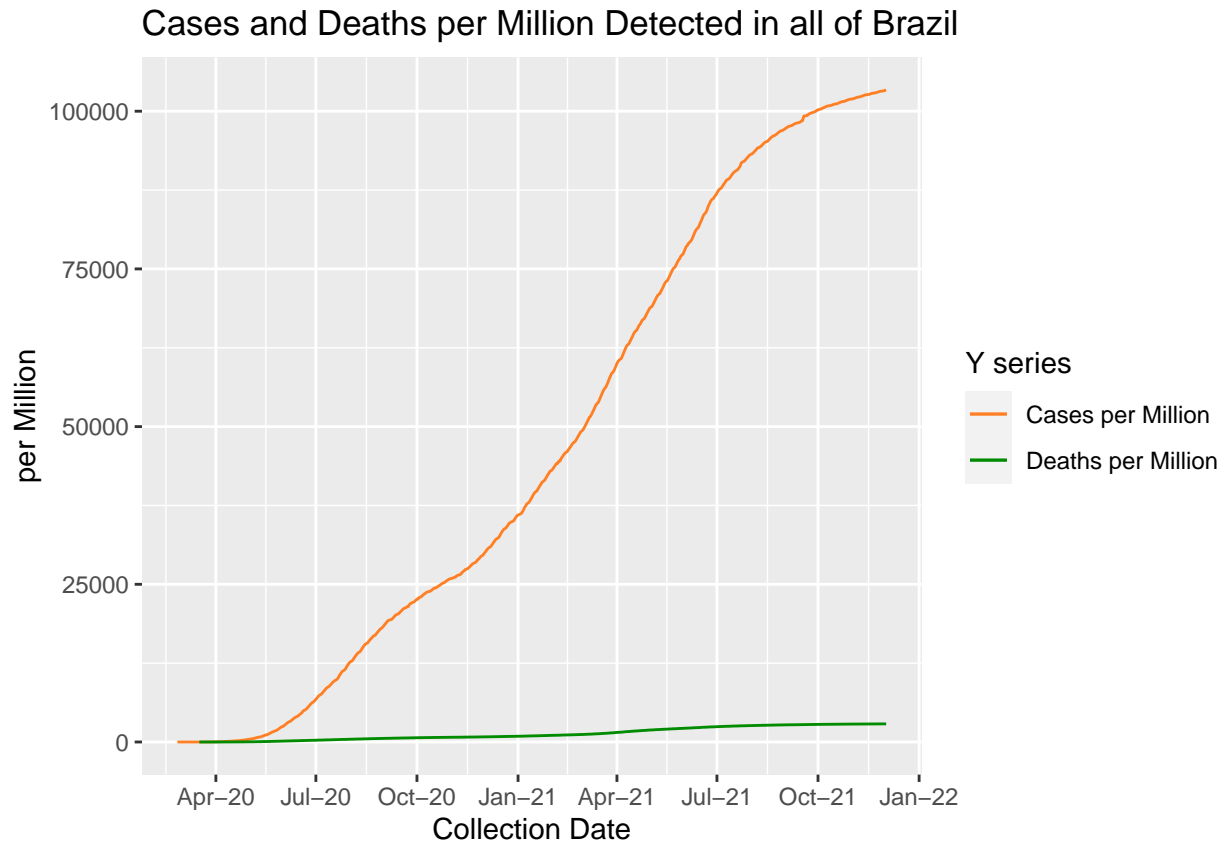


Figure 7: Using outside data from Our World In Data (OWID) comparing cases and deaths per million people in Brazil. Both variables continue to increase every month without plateauing.

Tables

Gene Name	Start	End	Length
S	21563	25384	3822
ORF3a	25393	26220	828
E	26245	26472	228
M	26523	27191	669
ORF6	27202	27387	186
ORF7a	27394	27759	366
ORF7b	27756	27887	132
ORF8	27894	28259	366
N	28274	29533	1260
ORF10	29558	29674	117

Table 1: Genes S and N have are the longest genes found in the SARS-CoV-2 in Brazil and they also have the highest SNP's according to Figure 1.

Sources Cited

Arnold,J.B. (2021) Ggthemes: Extra themes, scales and geoms for 'ggplot2'.

Bache,S.M. and Wickham,H. (2020) Magrittr: A forward-pipe operator for r.

- Firke,S. (2021) Janitor: Simple tools for examining and cleaning dirty data.
- Mortality analyses (2021) *Johns Hopkins Coronavirus Resource Center*.
- Nikas,J. (2021) Brazilian leader accused of crimes against humanity in pandemic response. *The New York Times*.
- Rudis,B. (2020) Hrbrthemes: Additional themes, theme components and utilities for 'ggplot2'.
- Silva,S.J.R. da and Pena,L. (2021) Collapse of the public health system and the emergence of new variants during the second wave of the covid-19 pandemic in brazil. *One Health*, **13**, 100287.
- Wickham,H. (2016) Ggplot2: Elegant graphics for data analysis Springer-Verlag New York.
- Wickham,H. (2019) Stringr: Simple, consistent wrappers for common string operations.
- Wickham,H. (2021) Tidyr: Tidy messy data.
- Wickham,H. *et al.* (2021) Dplyr: A grammar of data manipulation.
- Wickham,H. and Hester,J. (2021) Readr: Read rectangular text data.
- Zhu,H. (2021) KableExtra: Construct complex table with 'kable' and pipe syntax.
- Centers for Disease Control and Prevention*.
- (2021) *Statista*.