Seniors remain the most vulnerable population group as Toronto enters the 6th wave of COVID-19*

An analysis of the COVID-19 pandemic in Toronto with an emphasis on the past six months

Ka Chun Mo

23 April 2022

Abstract

Ontario is relaxing its COVID-19 measures even though health care experts are predicting an incoming sixth wave of COVID-19. It is crucial for us to continue taking measures to protect ourselves and the people around us and to seek insights to keep ourselves inform. In this paper, we provide an up-to-date analysis of the various aspects of COVID-19 with an emphasis of the past six months. We found that seniors are still the most susceptible population group after two years into the COVID-19 pandemic.

1 Introduction

As Toronto is entering a sixth wave of COVID-19, health care experts have been urging Toronto residents to continue taking measures to protect themselves and the people around them. With the Omicron BA. 2 sub-variant starting to become the dominant version infecting people in Toronto, Toronto residents are seeing a surging of cases as of March 30. Even though Toronto already lifted the mask mandate, there are still people questioning the validity of this decision and its consequences of it. Some people are frustrated by the lifting of the mask mandate and they want the government to bring back the mask mandate. They believe the mask mandate should be kept in place since the COVID-19 pandemic still is not gone yet.

The Ontario government is still taking measures to protect the most vulnerable group of Toronto residents. Ontario plans to offer the fourth dose to residents who are 60 and up. There are already ongoing planning to open fourth-dose access for people in long-term care and retirement home. The second booster shot aims to be an added level of protection for the residents of Ontario. Moreover, Ontario is also providing Paxlovid, an antiviral medication developed by Pfizer as a treatment for COVID-19, to the adults with the highest risk. It is not available to all Toronto residents but it serves to protect the most vulnerable population group. It is both the government's responsibility and our responsibility to do justice to ourselves and also to the vulnerable population.

With the COVID-19 pandemic situation constantly evolving and new studies and information coming out all the time, we aim to investigate the COVID-19 pandemic data in Toronto to have a better understanding of various aspects (e.g. gender, age, number of new cases) of the situation in Toronto. We want to have a better understanding of the overall COVID-19 situation in Toronto so we can be better informed to make the best decision ourselves. We investigated various distributions of the variables of the COVID-19 cases in Toronto data from the Open Data Portal. We found that people who are 60 and up are still the most susceptible people as we are entering the sixth wave of the COVID-19 pandemic in Toronto. To better protect ourselves and according to our analysis, we should still take measures such as wearing masks and social distancing.

In this paper, we will first take an overview of the dataset we got from Open Data Toronto. We will look at the information Open Data Toronto gathered for each COVID-19 case to have a general insight into what we can look into further. We are also going to discuss the methods we use to analyze the data. Then, we will

*Code and data are available at: https://github.com/francomomo/covid_19_toronto.

discuss the limitation of our dataset. We are going to construct a linear regression model to identify the important factors that contribute to the fatality of COVID-19 cases. Towards the end of the paper, we will dedicate a significant portion to discussing the results we found after analyzing the COVID-19 dataset.

2 Data

2.1 Overview

We got the COVID-19 cases in Toronto dataset from Open Data Toronto portal (Gelfand 2020). The dataset contains 18 variables and approximately 32000 observations. A summary statistics table of all the variables can be found in Section B. "id" is a variable that is used as a unique row identifier for Open Data Toronto's database. "assigned_id" is a unique ID assigned by Toronto Public Health for the purpose of processing the data and post it on Open Data for tracking specific cases. "outbreak_associated" are associated with the category of outbreaks of COVID-19 in Toronto. The outbreak can be associated with the long-term care home, hospitals, congregate settings, household contact, and so on. "age_group" is the age group of the person at the time of illness. Age groups are divided into less than or equal to 19 years old, 20-29 years old, 30-39 years old, 40-49 years old, 50-59 years old, 60-69 years old, 70-79 years old, 80-89 years old, 90+, and unknown. "neighbourhood_name" contains 140 distinct neighborhood name in Toronto and they are established to help government and community agencies to identify the meaningful geographic area. "fsa" is the first three characters of postal code of the cases' primary home address. "source_of_infection" is the most likely way that the person got COVID-19. "classification" is categorized as either confirmed or probable. "episode_date" is the best estimate of when the person got COVID-19. "reported_date" is the date on which the case is reported to Toronto Public Health. "client_gender" is the self-reported gender of the case. "outcome" is the outcome of the person who got COVID-19 and it can be either fatal or resolved or active. For the rest of the variables ("currently_hospitalized," "currently_in_icu," "currently_intubated," "ever_hospitalized," "ever_in_icu," "ever_intubated") the variable names are self-explanatory and they can take values of either yes or no. We did not create any new variables for this paper.

There are similar COVID-19 datasets such as datasets from the City of Toronto and the COVID-19 Advisory for Ontario. However, we chose to use the dataset from Open Data Toronto because similar to other datasets and Open Data Toronto provides instructions on how to import their data using R code.

2.2 Methods

In this paper, we use R, a Statistical Computing Language, to analyze this dataset (R Core Team 2021). We use tidyverse for data manipulation (Wickham et al. 2019). We also use dplyr for data manipulation (Wickham et al. 2021). We use janitor for examining and cleaning dirty data (Firke 2021). We created either bar plots or line plots for most of the variables. We also renamed the column names to convert them into lower case and replace spaces with underscores for easier analysis.

2.3 Limitations

The COVID-19 data from Open Data Toronto is subject to change as Ontario's public health investigation receives more reports of the ongoing pandemic. The data might not be up-to-date for the most recent few days. However, the data will be updated completely and get overwritten at 8:30 AM on the Tuesday of every week. Then, the data would be posted on the following Wednesday. Furthermore, the numbers may be different from the numbers on other websites since the data are extracted at different times and from different sources.

3 Model

We created a model to predict fatality. We mutate the data so that the outcome will have value 1 if the outcome is fatal and the outcome will have value 0 if the outcome is resolved. We used age group, source of infection, and gender as independent variable to predict fatality. We create a multiple linear regression with

those variables' factors. We found that the factors that have p-value that are less than 0.05 are the age groups that are older than 60 years old and if the source of infection is in health care institutions. Those significant variables all have parameter estimates larger than 0. We reject the null hypothesis for those variables and we conclude that if the person who got infected with COVID-19 is older than 60 years old or if people who got infected COVID-19 in healthcare institutions, the outcome of those cases will more likely to be fatal. The model statistics is shown in section C.

4 Results

From our analysis, we found that most people who got COVID-19 are those who are 20 to 29 years old as shown in figure 1. People who are 20 to 29 years old are the most active compared to the other age groups. Figure 9 also shows that people who are 20 to 29 years old got COVID-19 the most often from traveling since they go out and travel the most. We also found that most COVID-19 cases are sporadic without any outbreak associated with it. This means that people might have a harder time identifying their sources of infection. As shown in figure 3, there is a significant portion of COVID-19 cases that cannot identify the source of infection. Different age groups also have different significant sources of infection as shown from figure 4 to figure 9. Our analysis also reaffirmed that seniors are still the most vulnerable population group as Toronto is entering into the sixth wave of COVID-19. It is still important for us to wear masks, social distance, and take measures to protect ourselves and the people around us.

5 Discussion

5.1 Most people who got COVID-19 are from 20 to 29 years old

Most people who got COVID-19 are from 20 to 29 years old. Figure 1 shows a distribution of the number of people who got COVID-19 by their ages. The reason why people from 20 to 29 years old got COVID-19 might be because they are usually the most active among all the ages. Compared to other age groups, people from 20 to 29 years old are young adults and they have the most freedom to go out and meet up with friends. They also have a stronger desire to meet people and they have a better chance to recover from COVID-19 compared to other age groups. For the age groups from 30 to 59 years old, they have a similar number of cases. People from 30 to 59 years old usually have their own family or planning to start their own family. Compared to 20 to 29 years old, people from 30 to 59 years old might choose to stay at home and are less active. People who are 19 years old and younger have a lower number of cases compared to the aforementioned age groups. This might be due to people who are 19 years old and younger are usually still under their parents' control. To protect their children better, parents might decide to make sure their kids go out less and stay at home more often. Starting from 60 years old, there is a decreasing number of cases as age increases.

5.2 Most COVID-19 cases are sporadic without any outbreak associated

As shown in figure 2, most COVID-19 cases are sporadic without any outbreak-associated. This is interesting because we can tell most COVID-19 cases might not be related to any outbreak associated with a gathering of people. Rather, people might get COVID-19 without going to a place that has been later identified as a sight that caused most people to get sick. Combining with the insight that we can get from 3, there is a significant number of COVID-19 cases that do not have any information on their sources of infection. This shows that there might be a portion of people who were in the sporadic infection category who did not any information about how they got COVID-19.

5.3 Household and Community Contact are the most common source of infection

Community Contact is the most common source of infection as shown in 3. Most COVID-19 cases spread the virus to the people in their community since they might interact the most with people who are close to them. It is also worthy to point out that there are a significant number of COVID-19 cases that could not identify how they got infected with COVID-19. Since it is hard to pinpoint an exact time and place of how people

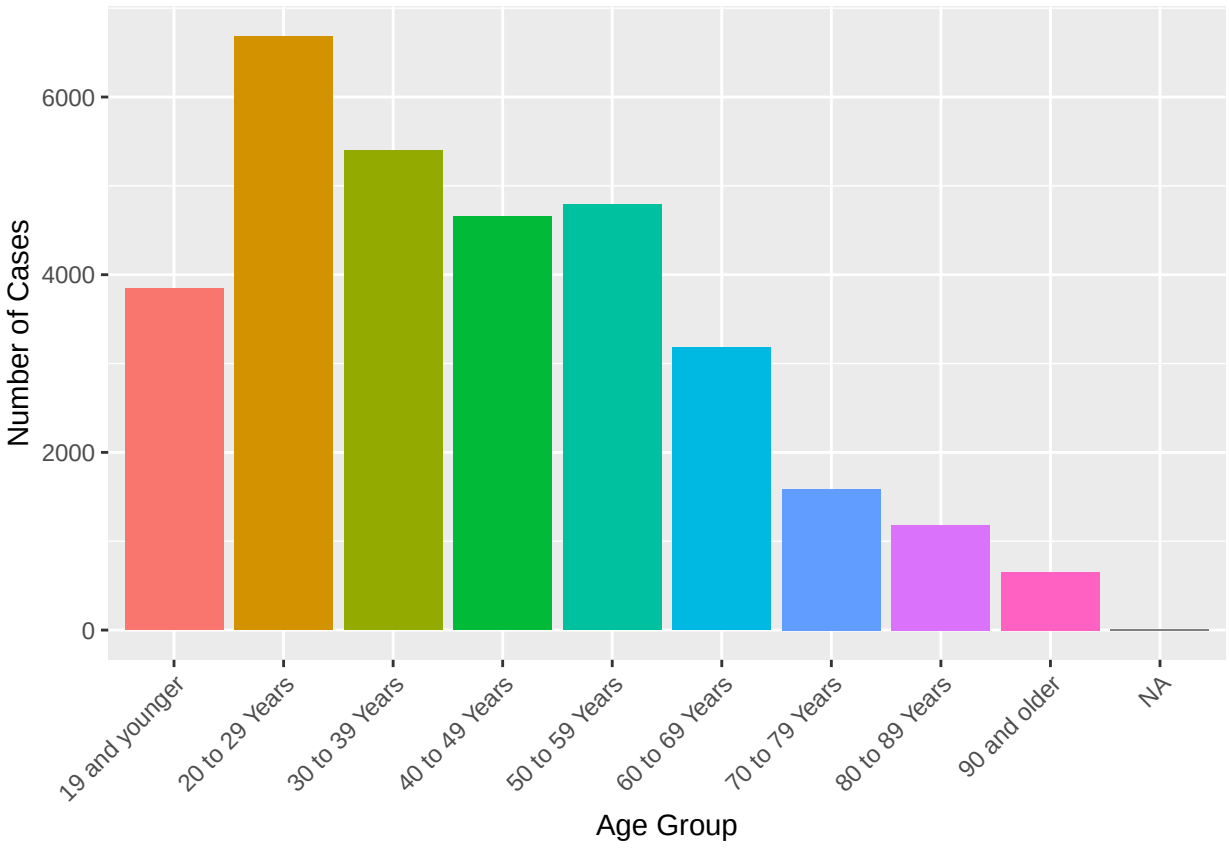

Figure 1: Age Group

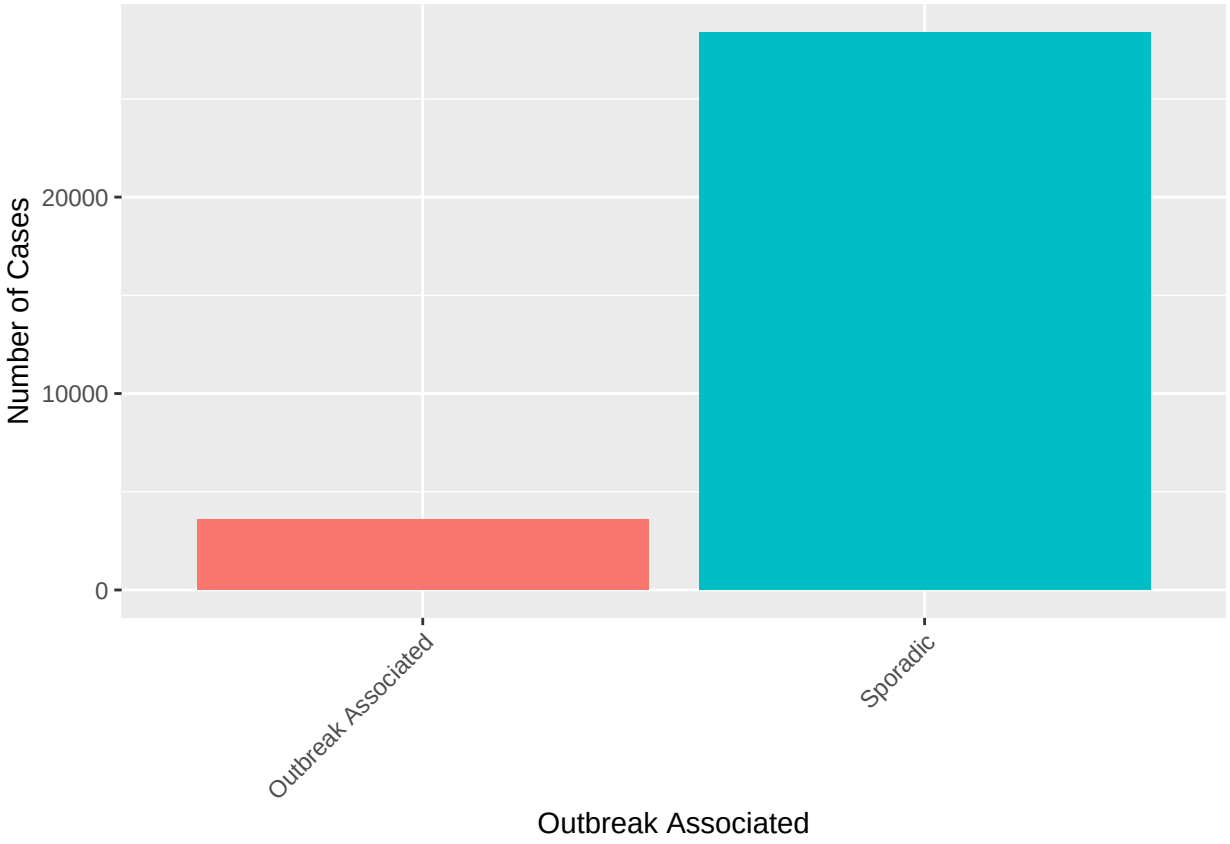

Figure 2: Outbreak Associated

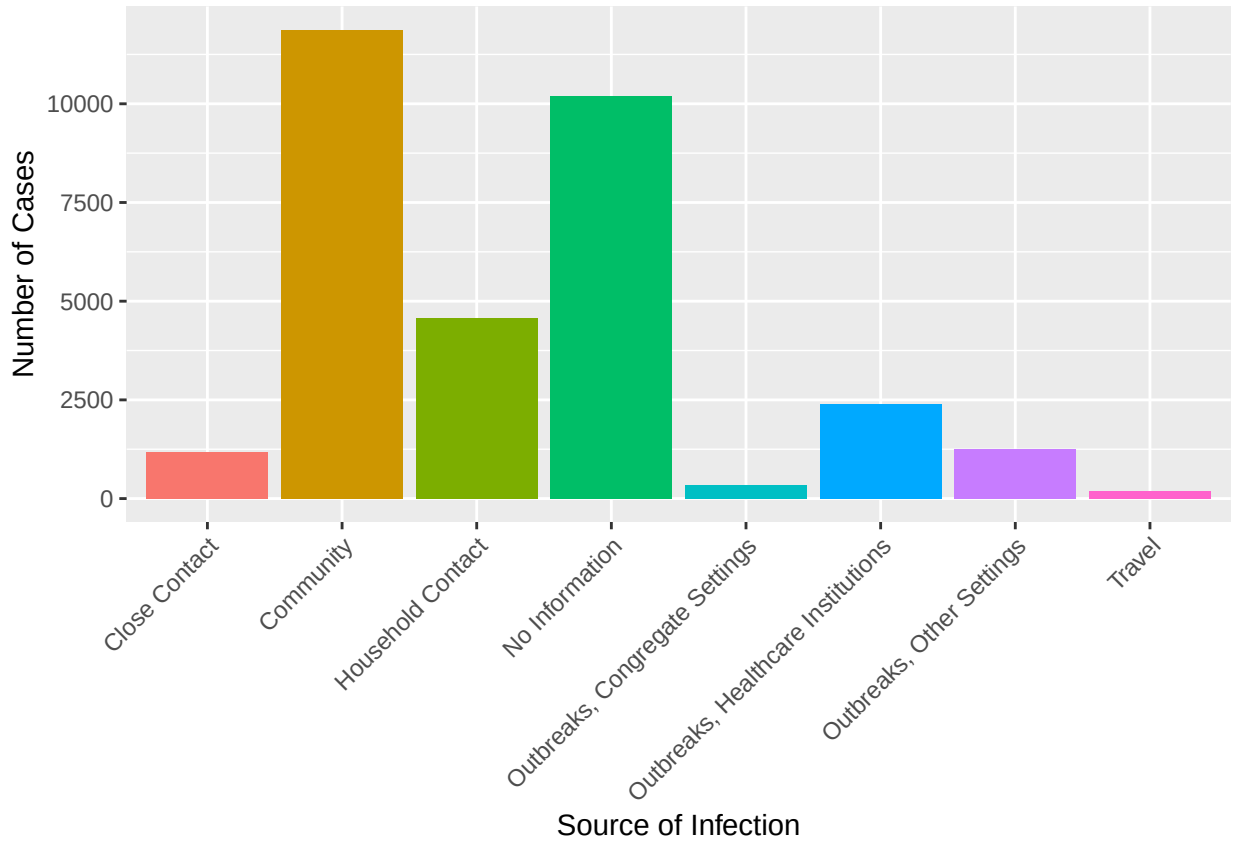

Figure 3: Source of Infection

got COVID-19 and most COVID-19 cases are sporadic, it is indeed hard for people to identify where exactly they got COVID-19. This also reflects how important it is for people to have COVID-19 exposure apps on their phones so that they can be alerted if they ever came across someone who had COVID-19 at least.

Other than Community Contact and No Information, Household contact is the third most common source of infection. It is understandable that once a person in a household got COVID-19, the rest of the household is also very likely to get COVID-19 since they are in close proximity for a huge portion of their days. Healthcare Institutions are the fourth most common source of infection. Staff in Healthcare Institutions might come in close contact with people who are sick. It is very likely that they will have close contact with someone who has COVID-19 since people who are infected with COVID-19 will go to health institutions for help. Surprisingly, Congregate Settings are the second least common source of infection. This might be due to incomplete data, but it is also worth noting that people might do a good job of avoiding going to crowded places during COVID-19. Also, since people travel less during COVID-19, travel is the least common source of infection.

5.4 Different age groups have different significant source of infection

From figure 4 to figure 9, we have several bar plots that show the age distribution of COVID-19 cases by the source of infection. Indeed, we see that there is a different age distribution for different sources of infection. For Community Infection as shown in figure 4, people who are 20 to 29 years old are the people who got COVID-19 most often from community infection since they are the most active age group among all the age groups. Whereas, people who are above 80 years old are the people who got COVID-19 from healthcare institutions as shown in figure 5 since there are more older people who stay in healthcare institutions and COVID-19 might also be spreading in healthcare institutions.

For Household Infection as shown in figure 6, most people who got COVID-19 from their household are people who are 19 and younger. There is a significant portion of people who are 19 and younger who are not adults yet. They might stay at home more often than other age groups as ordered by their parents. The only way for those people to get COVID-19 is for their family members to bring COVID-19 home. Close Contact also reflects the same scenario as what we have just talked about as shown in figure 7. People who are 19 and younger are the people who got the most infected from close contact. The number of COVID-19 cases from close contact decreases as age increases.

For Congregate Settings as shown in figure 8, the number of COVID-19 cases is approximately the same for the age groups between 30 and 59 years old. Those are the age groups that go to congregate settings such as public transportation areas and offices the most. Whereas the other age groups will either stay at home or avoid crowded places. For Travel Infection as shown in figure 9, people who are 20 to 29 years old again are again the age group that might get COVID-19 the most from traveling since they are more active than other age groups.

5.5 Seniors remain the most vulnerable among the COVID-19 pandemic

As shown in figure 10, fatality increases as age increases. Seniors remain the most vulnerable age group among all the age groups. Seniors might have weaker immune systems than other age groups. They are the most vulnerable when facing COVID-19. It is our responsibility to protect ourselves and continue practicing social distancing so that we can protect ourselves as well as the vulnerable population in our society. Even though, we only see the numbers from our data analysis. But even if one fatal case is serious enough, we should do our best to spread truthful information and protect the people around us.

5.6 Weakness and Looking Forward

In this paper, we mostly look at the distributions of the variables we have from the COVID-19 dataset from Open Data Toronto. We have drawn a lot of insights and interesting observations from those distributions. However, there might be a hidden relationship between all these variables. If we dive deeper and start making more regression between all the variables in our datasets, we might understand more than what we have now. Furthermore, as mentioned in section 2.3, the data is constantly updating as time goes by. The data

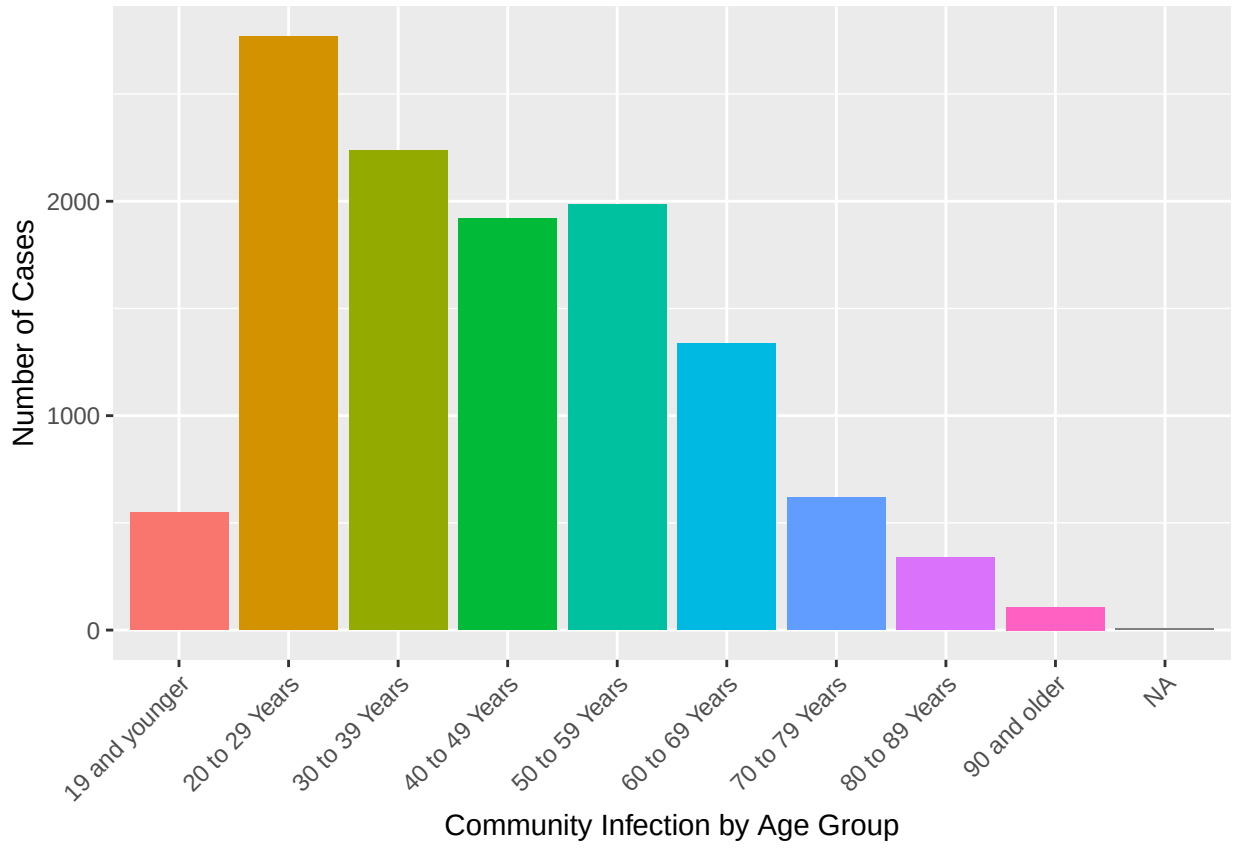

Figure 4: Community Infection by Age Group

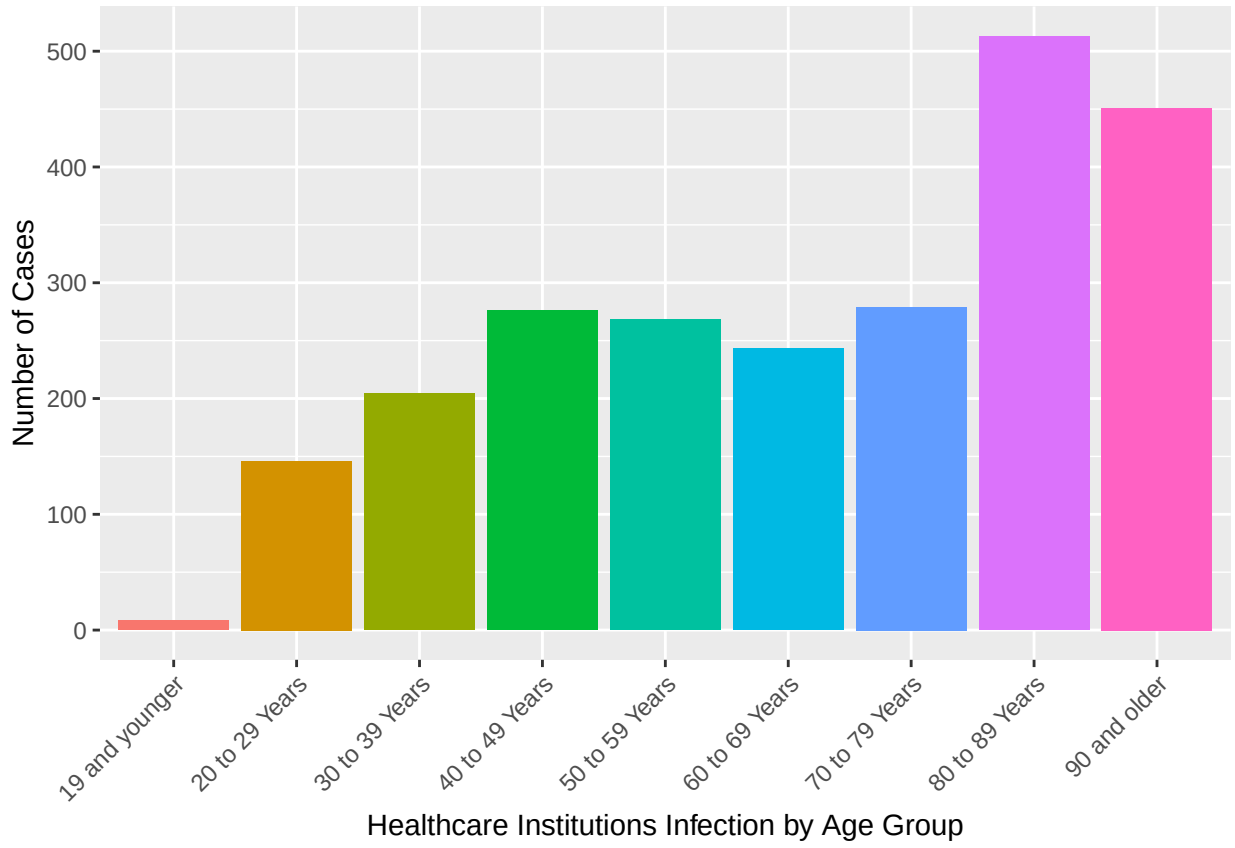

Figure 5: Healthcare Institutions Infection by Age Group

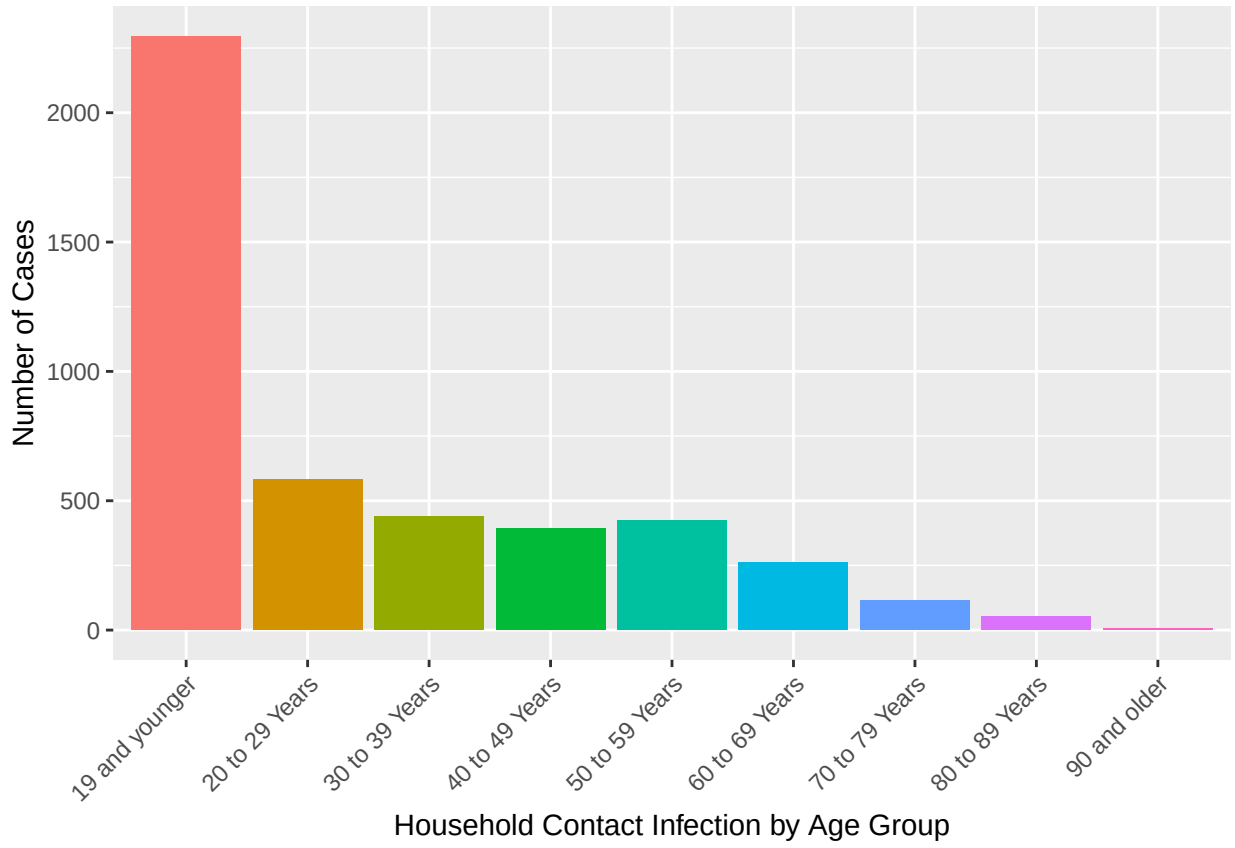

Figure 6: Household Contact Infection by Age Group

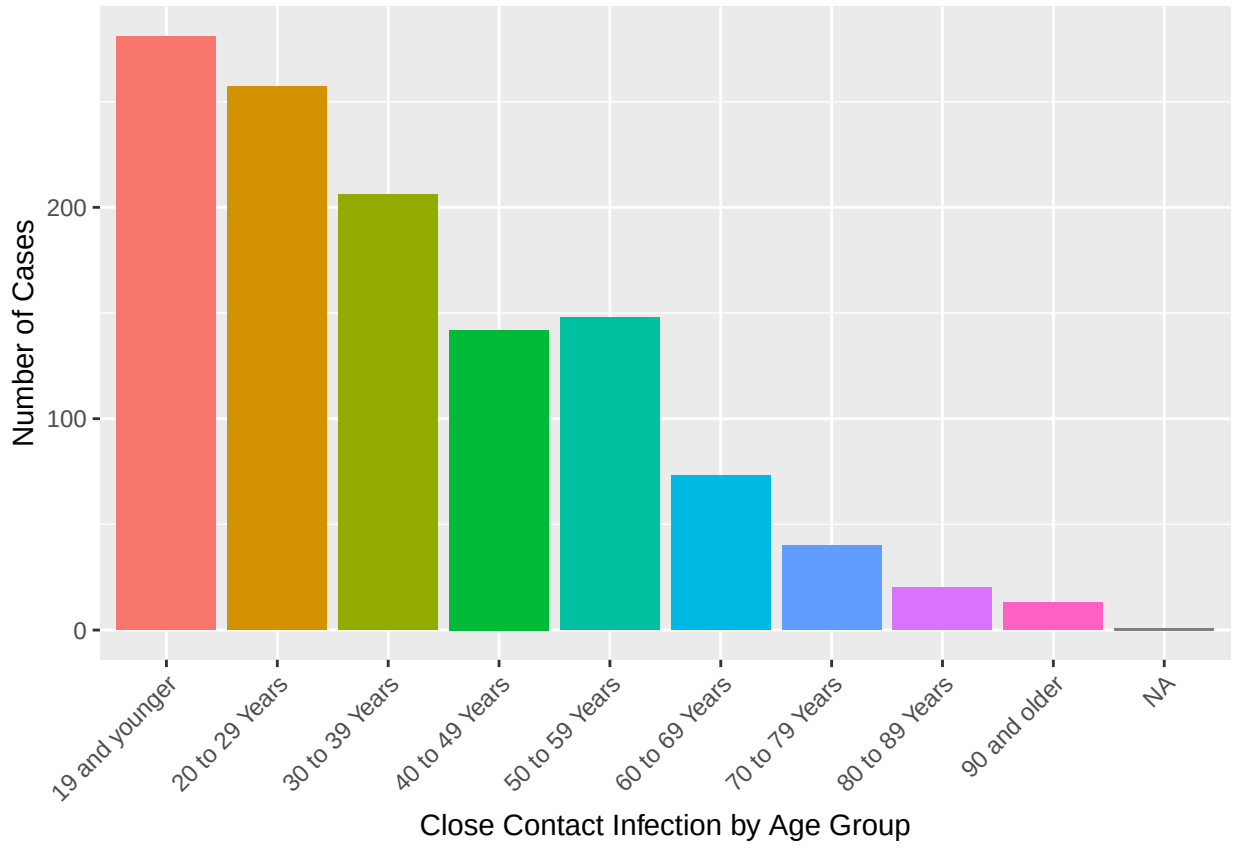

Figure 7: Close Contact Infection by Age Group

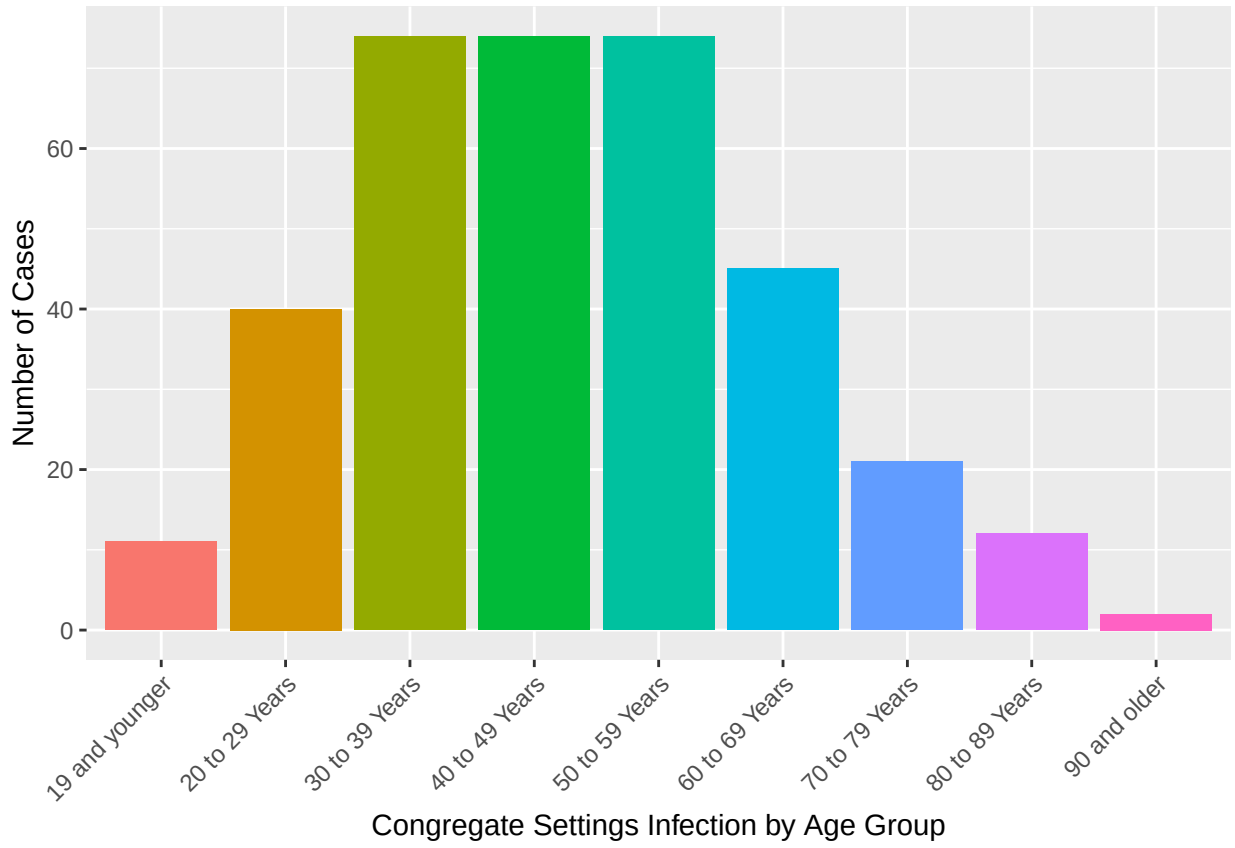

Figure 8: Congregate Settings Infection by Age Group

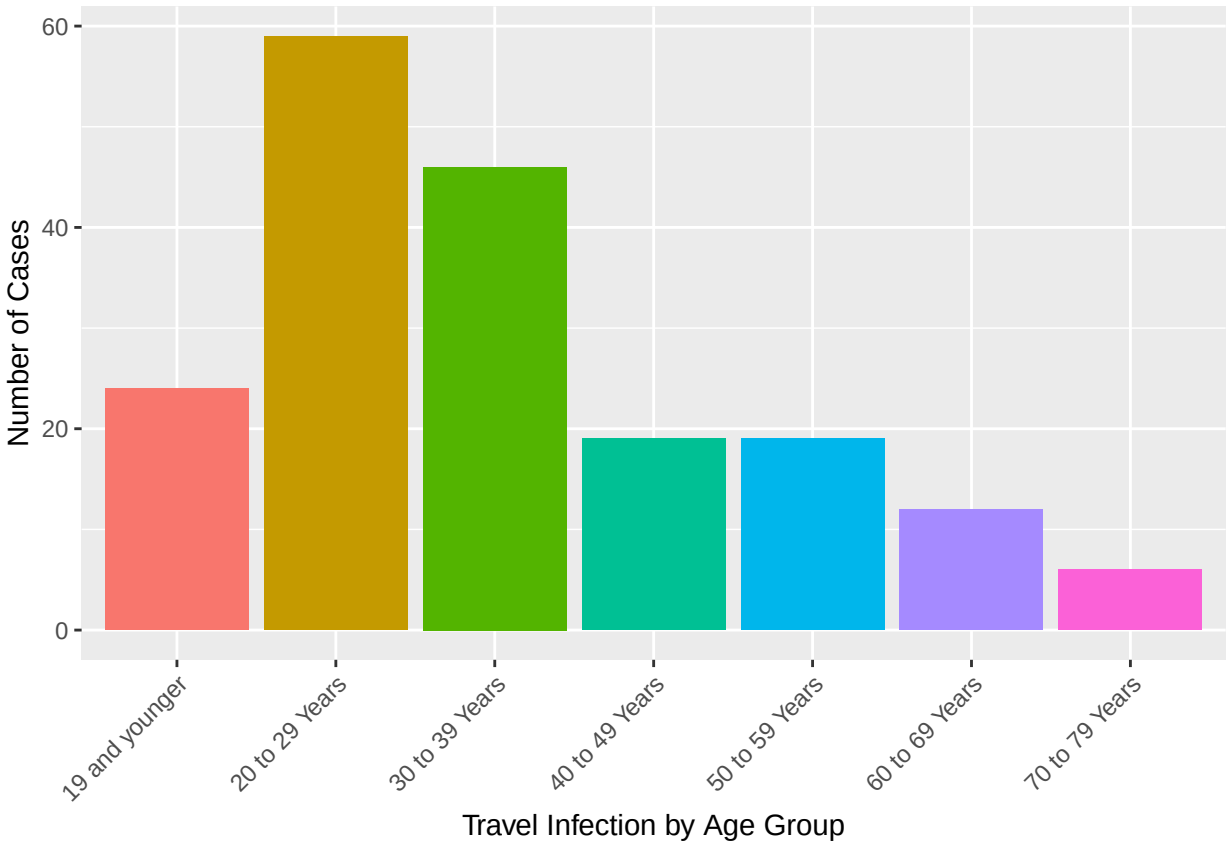

Figure 9: Travel Infection by Age Group

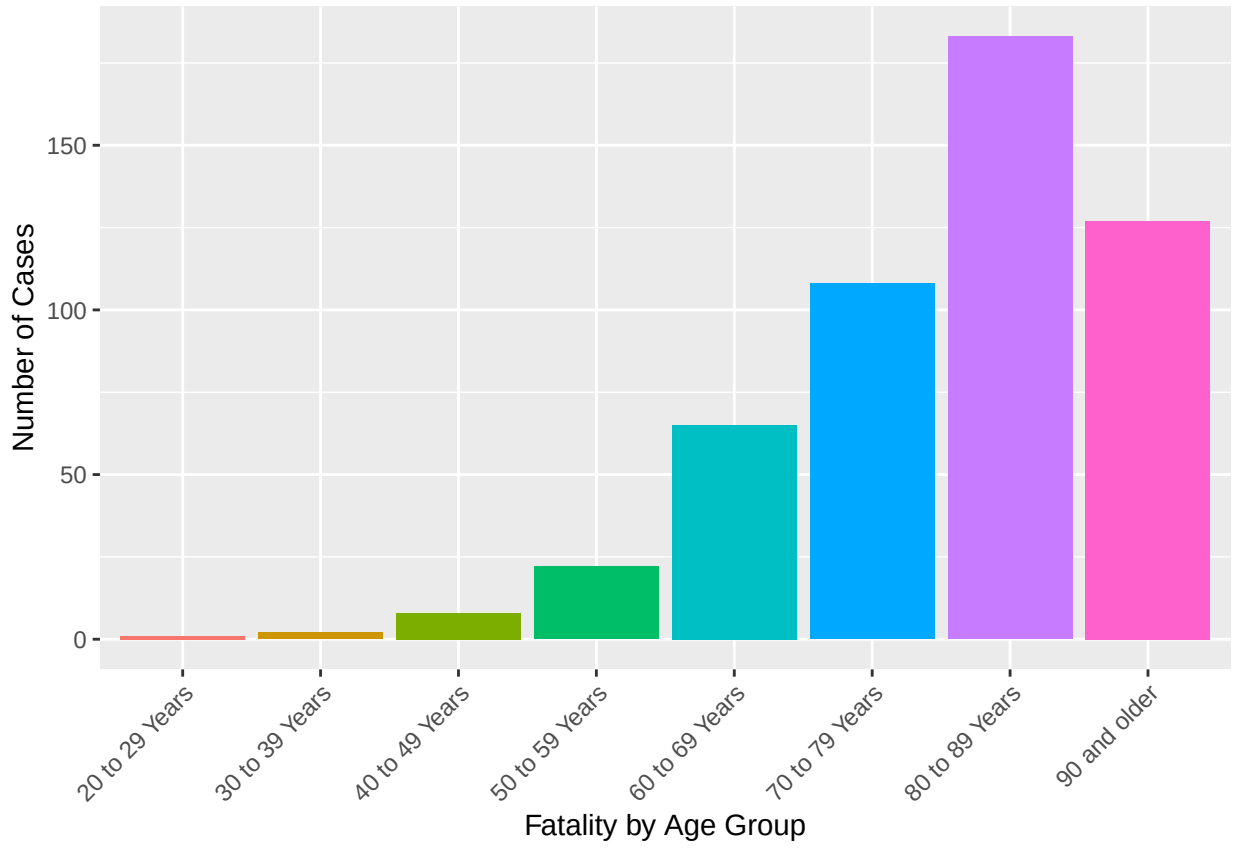

Figure 10: Fatality by Age Group

that we have might not be up-to-date. If we can create some sort of web application that can automate the visualization of the distributions of the variables and also draw relationships between variables automatically, this will allow us to understand and monitor the COVID-19 situation more promptly and draw insights faster.

A Graphs

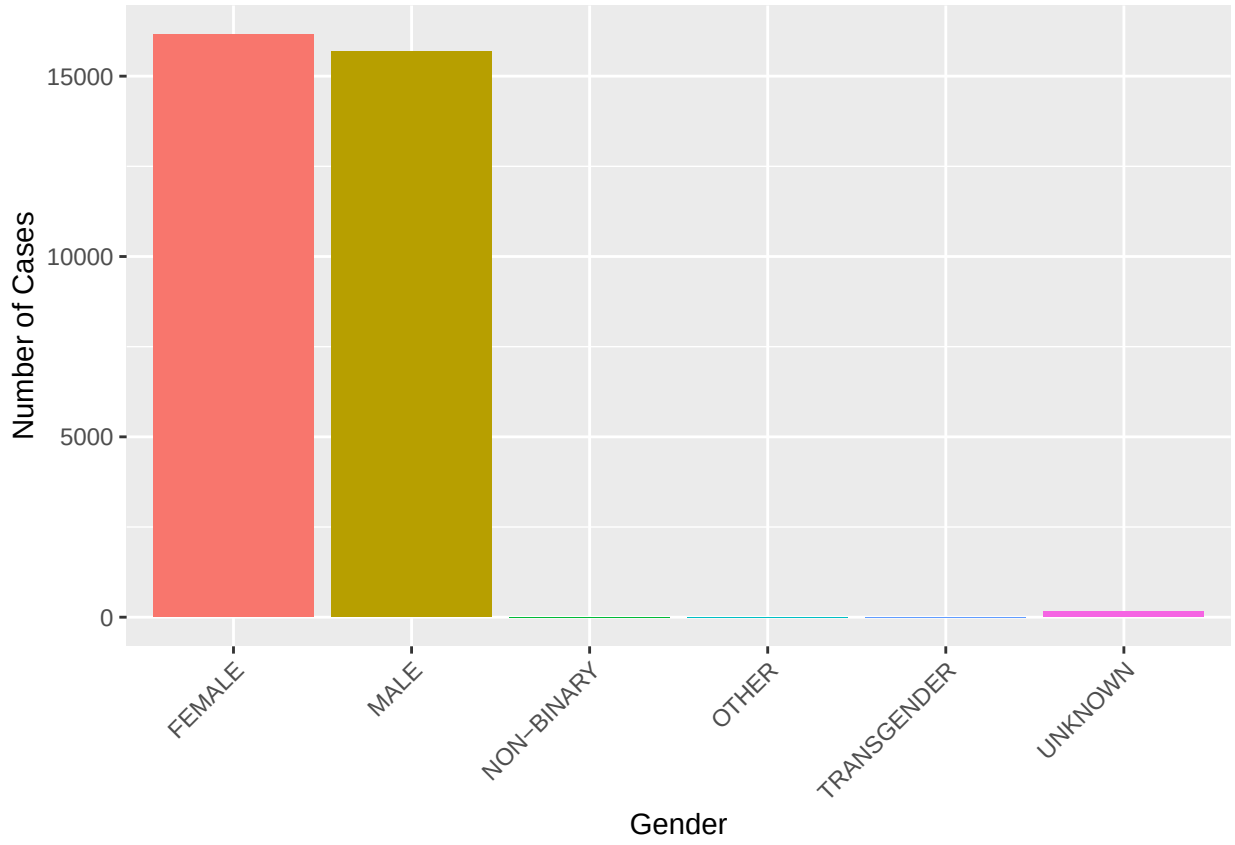

Figure 11: Gender

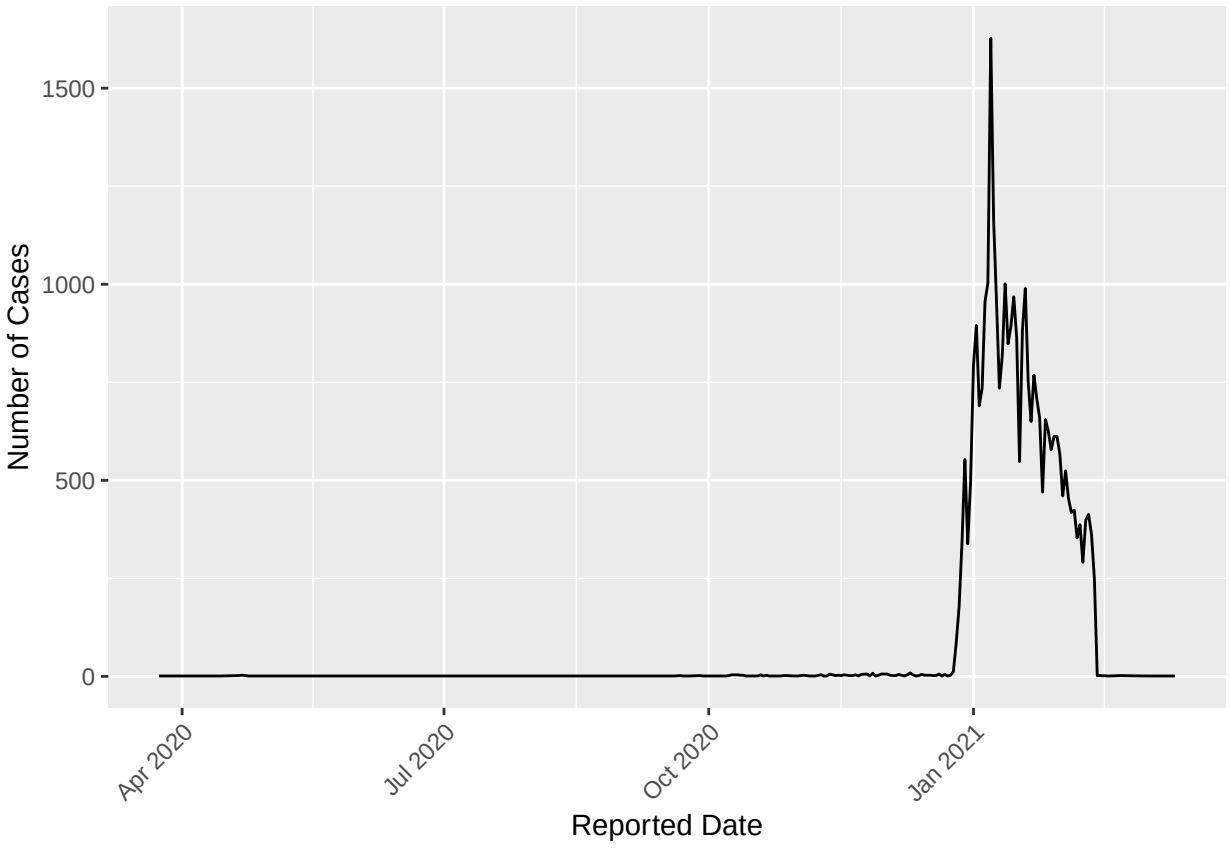

Figure 12: Reported Date

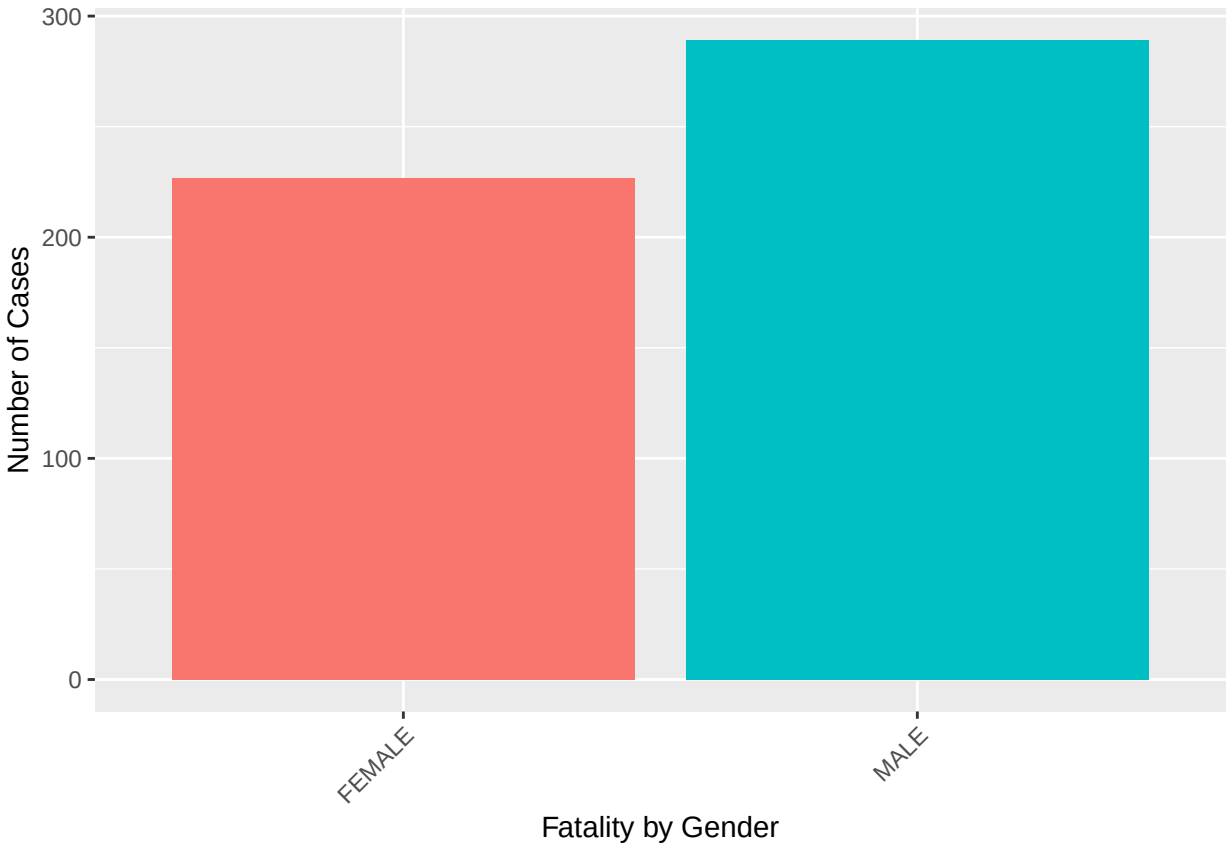

Figure 13: Fatality by Gender

B Summary Statistics

```
##        id          assigned_id    outbreak_associated  age_group
##  Min.   :59565   Min.   :61907   Length:32000        Length:32000
##  1st Qu.:67571   1st Qu.:70118   Class :character    Class :character
##  Median :75570   Median :78320   Mode  :character    Mode  :character
##  Mean   :75570   Mean   :78417
##  3rd Qu.:83570   3rd Qu.:86784
##  Max.   :91570   Max.   :95102
##  neighbourhood_name     fsa            source_of_infection classification
##  Length:32000       Length:32000       Length:32000        Length:32000
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##
##
##
##   episode_date        reported_date        client_gender
##  Min.   :2020-03-22   Min.   :2020-03-24   Length:32000
##  1st Qu.:2021-01-02   1st Qu.:2021-01-07   Class :character
##  Median :2021-01-11   Median :2021-01-15   Mode  :character
##  Mean   :2021-01-11   Mean   :2021-01-16
##  3rd Qu.:2021-01-22   3rd Qu.:2021-01-26
##  Max.   :2021-02-13   Max.   :2021-03-12
##    outcome          currently_hospitalized currently_in_icu
##  Length:32000       Length:32000           Length:32000
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##  currently_intubated ever_hospitalized  ever_in_icu        ever_intubated
##  Length:32000        Length:32000       Length:32000       Length:32000
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
##
##
##
```

C Model Summary

```
model_data <- covid_data %>%
  mutate(fat=
           case_when(outcome=='FATAL' ~ 1,
                     outcome=='RESOLVED' ~ 0))

model <- lm(fat ~ age_group + source_of_infection + client_gender, data=model_data)
summary(model)
```

```
##
## Call:
## lm(formula = fat ~ age_group + source_of_infection + client_gender,
##     data = model_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20452 -0.01026 -0.00312  0.00138  1.00764
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -0.0062816  0.0039704
## age_group20 to 29 Years                           -0.0004437  0.0026670
## age_group30 to 39 Years                           -0.0001890  0.0027643
## age_group40 to 49 Years                            0.0012140  0.0028411
## age_group50 to 59 Years                            0.0040648  0.0028215
## age_group60 to 69 Years                            0.0191084  0.0030845
## age_group70 to 79 Years                            0.0662069  0.0037671
## age_group80 to 89 Years                            0.1511879  0.0043061
## age_group90 and older                              0.1897048  0.0055362
## source_of_infectionCommunity                       0.0053412  0.0036643
## source_of_infectionHousehold Contact               0.0007914  0.0039386
## source_of_infectionNo Information                 -0.0025697  0.0036869
## source_of_infectionOutbreaks, Congregate Settings -0.0024514  0.0072465
## source_of_infectionOutbreaks, Healthcare Institutions  0.0107177  0.0044719
## source_of_infectionOutbreaks, Other Settings      -0.0009357  0.0048599
## source_of_infectionTravel                          0.0022132  0.0094193
## client_genderMALE                                  0.0103811  0.0013470
## client_genderNON-BINARY                            0.0013841  0.0841877
## client_genderOTHER                                 0.0056792  0.1190609
## client_genderTRANSGENDER                           0.0050676  0.1190958
## client_genderUNKNOWN                              -0.0028562  0.0097724
##                                                   t value Pr(>|t|)
## (Intercept)                                        -1.582   0.1136
## age_group20 to 29 Years                            -0.166   0.8679
## age_group30 to 39 Years                            -0.068   0.9455
## age_group40 to 49 Years                             0.427   0.6692
## age_group50 to 59 Years                             1.441   0.1497
## age_group60 to 69 Years                             6.195 5.90e-10 ***
## age_group70 to 79 Years                            17.575  < 2e-16 ***
## age_group80 to 89 Years                            35.110  < 2e-16 ***
## age_group90 and older                              34.266  < 2e-16 ***
## source_of_infectionCommunity                        1.458   0.1450
## source_of_infectionHousehold Contact                0.201   0.8408
```

```
## source_of_infectionNo Information                       -0.697   0.4858
## source_of_infectionOutbreaks, Congregate Settings       -0.338   0.7352
## source_of_infectionOutbreaks, Healthcare Institutions    2.397   0.0165 *
## source_of_infectionOutbreaks, Other Settings            -0.193   0.8473
## source_of_infectionTravel                                0.235   0.8142
## client_genderMALE                                        7.707 1.33e-14 ***
## client_genderNON-BINARY                                  0.016   0.9869
## client_genderOTHER                                       0.048   0.9620
## client_genderTRANSGENDER                                 0.043   0.9661
## client_genderUNKNOWN                                    -0.292   0.7701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.119 on 31966 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.1079, Adjusted R-squared:  0.1073
## F-statistic: 193.3 on 20 and 31966 DF,  p-value: < 2.2e-16
```

D Reference

https://globalnews.ca/news/8735445/covid-6th-wave-game-plan-canada/

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.