Data from open data Toronto on Covid-19 and the amount of affected people in Toronto*

Swarnadeep Chattopadhyay

Feb 6, 2022

Abstract

This paper is regarding the Covid-19 cases in Toronto. The city of Toronto has been able to collect data on a daily basis and present it to the public for their use. This paper portrays figures such as graphs and tables in order for the readers to have a visual of the data. The results were rather astonishing as it provides an in depth visual of who or where the most cases occurred. The paper concludes by discussing the data and provides any biases that resulted from the data.

1 Introduction

Covid-19 is a global pandemic which started back in 2019 and has been evolving since. In order to measure the people affected by this disease, the City of Toronto has launched a data set which provides how many people have been affected and other factors such as gender, age, neighborhood etc. This data can be retrieved from Gelfand (2020). R Core Team (2020) and other pakages such as Wickham et al. (2019), Wickham et al. (2021), Horst, Hill, and Gorman (2020) and Kassambara (2020) all helped with making sure the codes ran smoothly. I was able to create visuals using Wickham (2016) which helps the readers understand the data better. Xie (2020) was used to help with the project as the book contained all the information needed.

In this paper, we look closely at how Covid has affected in terms of age groups, neighborhoods and which year had the highest cases recorded. In terms of age group, 20 to 29 years saw the highest amount of cases recorded. In terms of neighborhoods, Woburn saw the highest number of cases. As for which year, it was 2021. These will help us create an understanding of who has been affected the most or which area has been hit the hardest. In the next section, we will look at these data and discuss in detail as to what the visuals are telling us and what it could mean.

In terms of omitting any bias, I decided to include all the data to make sure the results were accurate. This however took long as loading the data was difficult at first but it eventually got faster as the more times I ran the code. In terms of bias in the data itself like age groups, neighborhoods and year, we will discuss what could have caused them.

2 Data

The data used in this analysis consists of the daily and cumulative incidence (confirmed cases) of COVID-19 in Toronto neighborhoods. This data covers 140 neighborhoods since the start of the pandemic, Jan, 2020 to Feb 2nd, 2022 for all confirmed and probable cases reported to and managed by Toronto Public Health, including cases that are sporadic (occurring in community) and outbreak-associated.

The data analyzed in this report was obtained in csv format from the City of Toronto Open Data Portal using the R Core Team (2020) package Gelfand (2020).

*https://github.com/WanteXecutioN/Sta304-Assignment-

Let's take a quick look at our data.

```
## [1] 277473     18
```

```
##  [1] "X_id"                   "Assigned_ID"            "Outbreak.Associated"
##  [4] "Age.Group"              "Neighbourhood.Name"     "FSA"
##  [7] "Source.of.Infection"    "Classification"         "Episode.Date"
## [10] "Reported.Date"          "Client.Gender"          "Outcome"
## [13] "Currently.Hospitalized" "Currently.in.ICU"       "Currently.Intubated"
## [16] "Ever.Hospitalized"      "Ever.in.ICU"            "Ever.Intubated"
```

```
##                   X_id            Assigned_ID    Outbreak.Associated
##                 277473                 277473                      2
##              Age.Group     Neighbourhood.Name                    FSA
##                     10                    141                    181
##    Source.of.Infection         Classification           Episode.Date
##                      9                      2                    717
##          Reported.Date          Client.Gender                Outcome
##                    711                      9                      3
## Currently.Hospitalized       Currently.in.ICU    Currently.Intubated
##                      2                      2                      2
##      Ever.Hospitalized            Ever.in.ICU         Ever.Intubated
##                      2                      2                      2
```

```
##   Outbreak.Associated      Age.Group    Neighbourhood.Name Source.of.Infection
## 1            Sporadic 50 to 59 Years       Willowdale East              Travel
## 2            Sporadic 50 to 59 Years       Willowdale East              Travel
## 3            Sporadic 20 to 29 Years     Parkwoods-Donalda              Travel
## 4            Sporadic 60 to 69 Years Church-Yonge Corridor              Travel
## 5            Sporadic 60 to 69 Years Church-Yonge Corridor              Travel
```

As seen, there are 277,473 cases recorded and 18 different colums containing key information. This includes 10 age groups, 141 neighborhoods and 9 source of infection. The data also includes other factors such as gender, ICU etc. but we will be focusing on the 3 mentioned in the above sentence.

A simple exploratory analysis of the incidence data is provided in this report.

3 Methodology

The data set consists of over 277,000 cases, and the first thought was to create a smaller sample in order to efficiently execute the data in the statistical programming language, R Core Team (2020). For this, stratified sampling, a method that involves dividing a population into smaller groups, was taken into consideration. Each subgroup would be adequately represented within the whole sample population. In our case, the data set includes an "Age Group" category. The only thing required, now, was to preserve the proportion of "Age Groups" in our sample set. However, during a test-run of the original data, it was observed that the results were computed rather quickly than expected. As such, the idea of sampling the original data was dropped and the analysis was conducted on the entire data set.

Lets get into some data visualization.

4 Results

Neighbourhood Name	Number of Cases
Woburn	6448
	6384
Downsview-Roding-CFB	6117
Waterfront Communities-The Island	6046
Malvern	5554
Rouge	5492
West Humber-Clairville	5457
Mount Olive-Silverstone-Jamestown	5268
Glenfield-Jane Heights	4729
York University Heights	4411

Here we can look at the top 10 neighborhoods in Toronto with the most amount of cases. As seen, Woburn tops the list with 6448 cases while Islington-City Centre West is at the bottom with 4130.

We now look at some more visuals that will show a detailed version of the year, month, age group that got hit the most alongwith the leading factors for the infection.

Fig1 shows total cases annually below

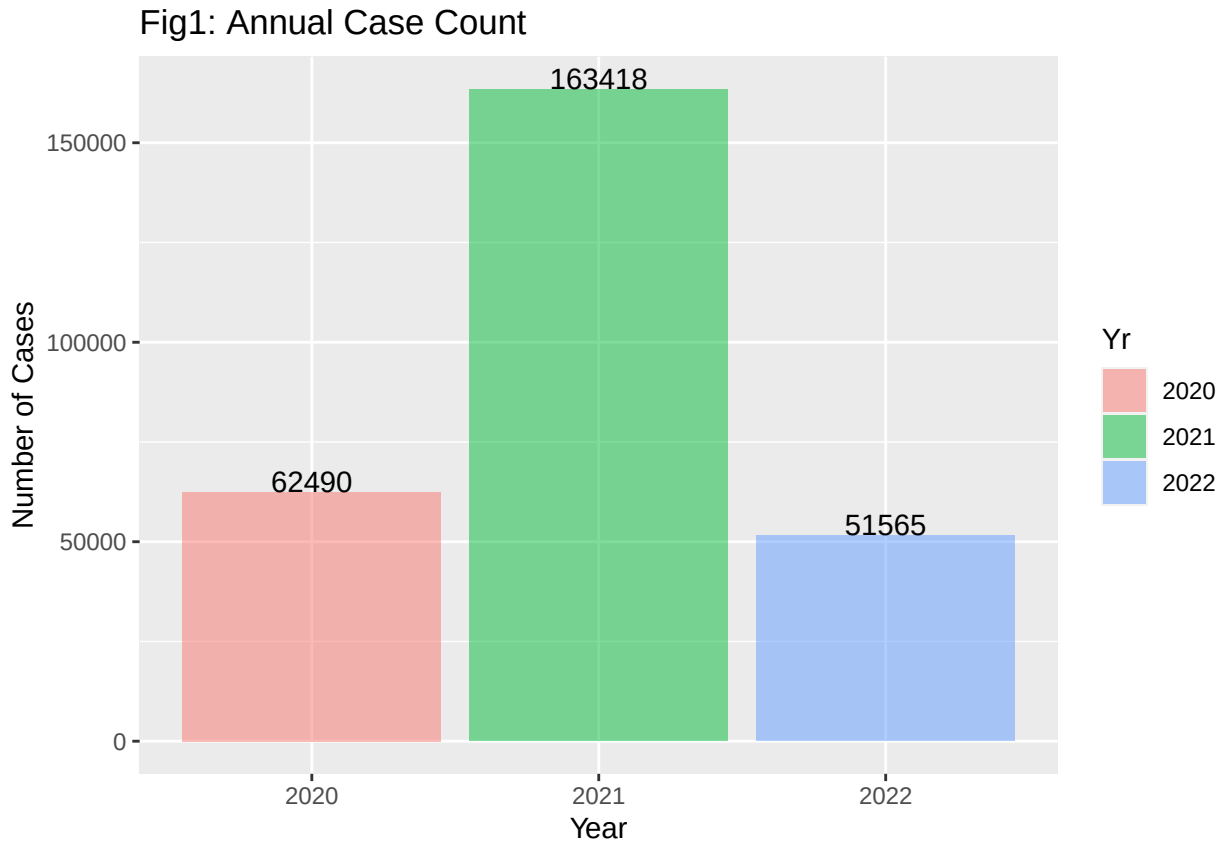


Fig2 shows monthly numbers for each year below

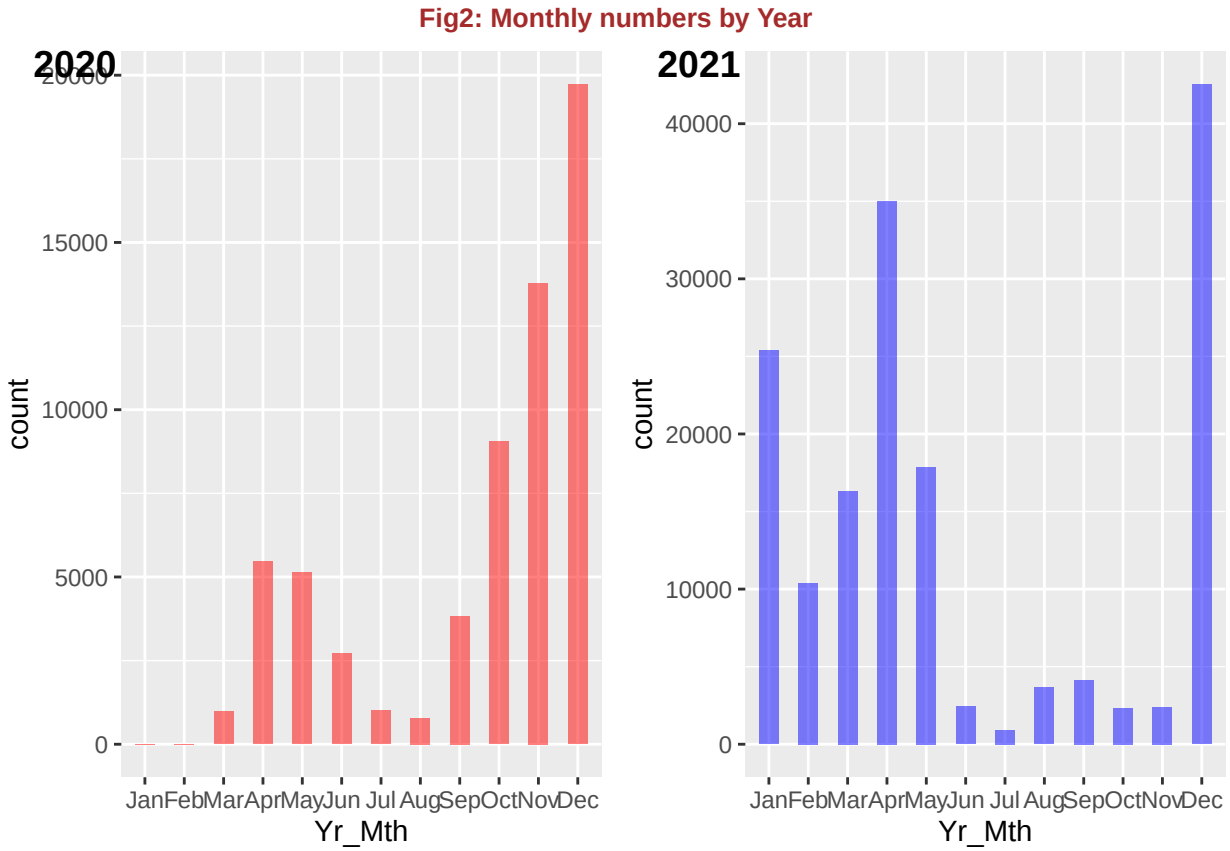


Fig3 shows the case count by age groups below

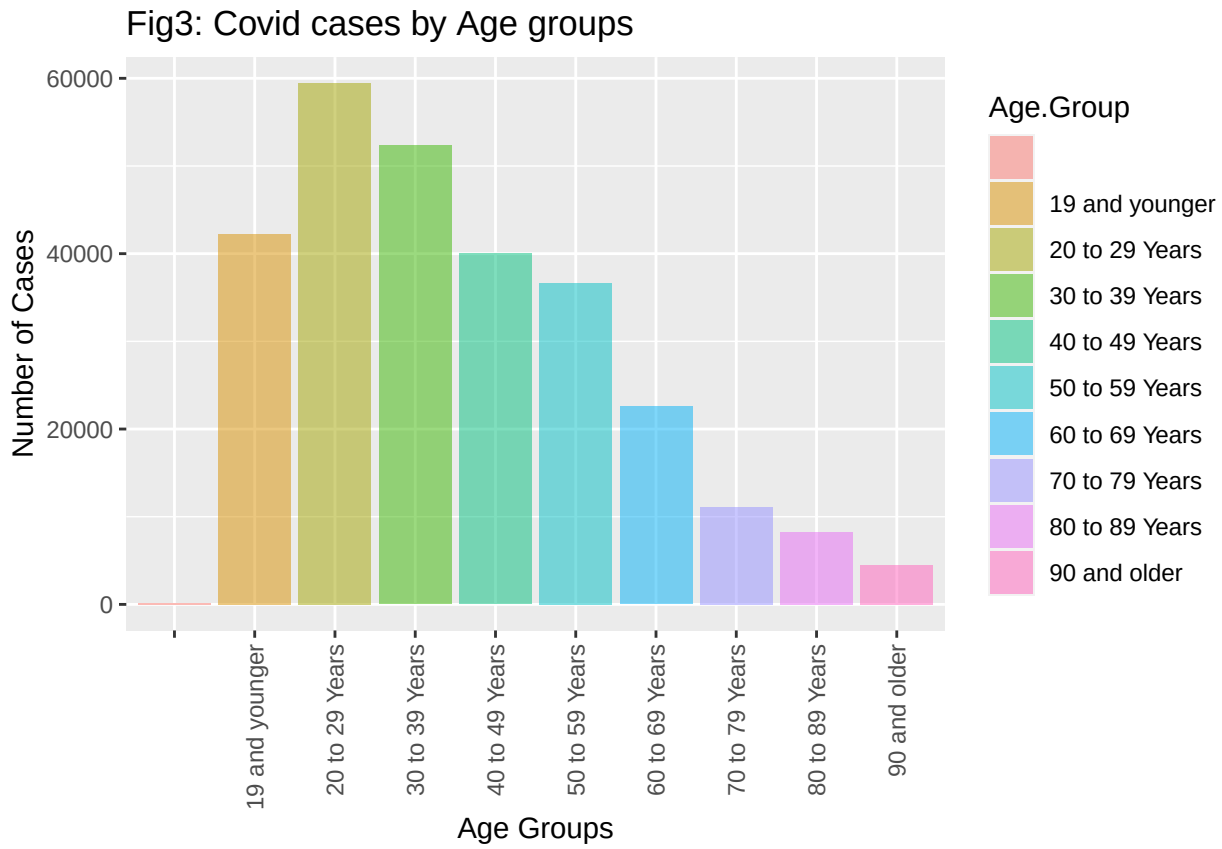
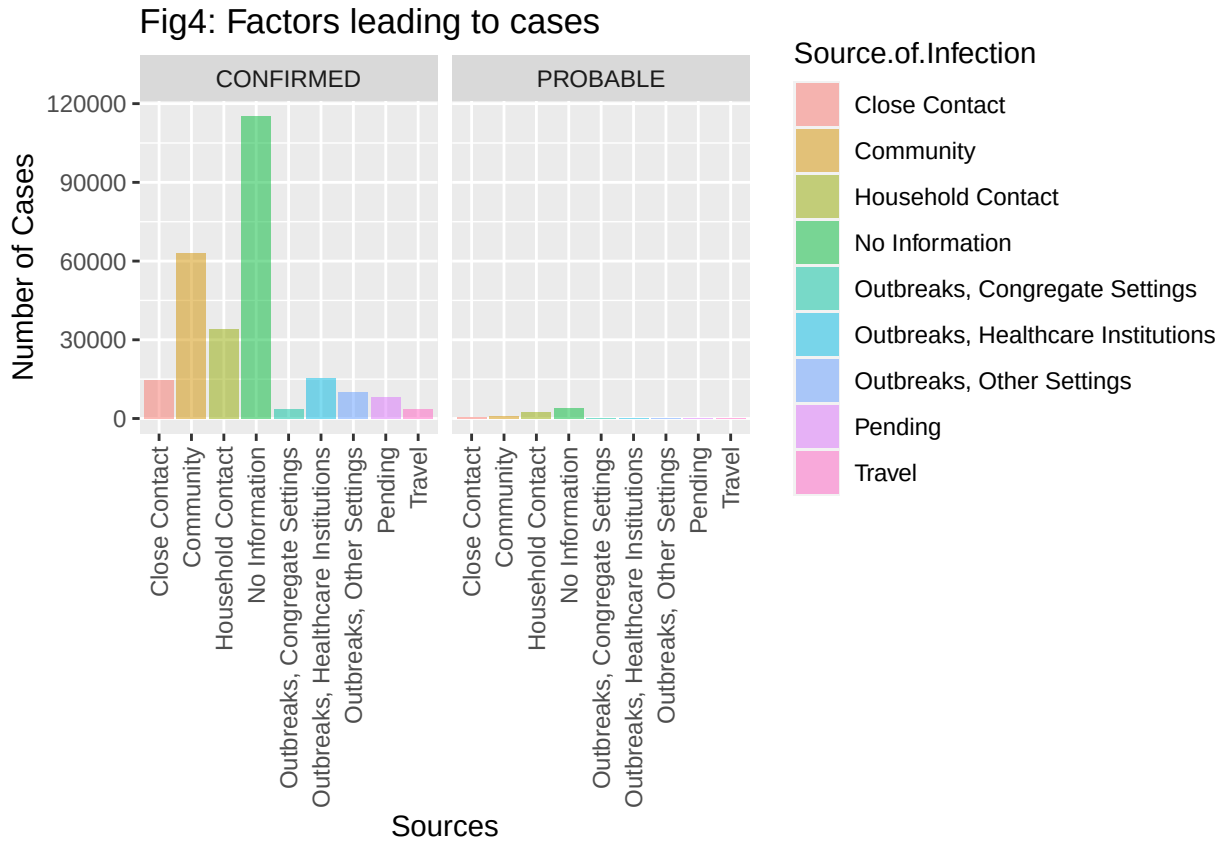


Fig4 shows Leading factors for infections below



5 Discussion

As we can see clearly, the number of cases raised rapidly in 2021, compared to 2020, mainly because of Delta and Omicron variants, first detected in early 2021, and late 2021 respectively. Both variants were identified as a "variant of concern" by WHO (World Health Organization) with both being highly contagious. Multiple testing centers were also built which caused more people to get testing done resulting in more cases reported.

Fig2 depicts this issue fittingly but in terms of monthly numbers. As we can see high case numbers in the early and late months of 2021. This was again due to the reason that WHO detected more variants which saw a hike in case numbers. Toronto went into lock down after the early months which is the reason why the cases are lower. As things started to open up again during the last few months of the year, cases started to rise again.

Even though COVID-19 does not discriminate by age, the most affected age group is "20 to 29 years" with cases amassing to almost 22% of the total case count, as shown in Fig3. One reason could be that this age group is ideally young and meeting people on a constant basis due to their job requirements. Age 20 - 29 years like to travel more than the other age group as they have the time and freedom to do so. Doing this allows this particular age group to catch the symptoms more compared to someone who is young staying home for online studies or someone old staying home due to old age.

When it comes to determining the leading cause for the infections shown in Fig4, about 44% of total cases have no information. This makes it problematic to conclude a single leading factor. Having no information means that a solution cannot be deducted which makes it difficult to stop the spread. The 2nd and 3rd highest reason is Community and household contact respectively. This is rather surprising as Toronto going

into lock down meant people to stay home and only leave for essential. However, this result shows people were leaving their homes and meeting others in the community who happens to spread the virus. This results in the household members to also get the virus as they are breathing the same air as them and living with each other.

6 Bias

In order to reduce any bias, the entire data set was utilized instead of taking a sample. However, one can expect to see unfairness when it comes to leading causes for infection, as a lot of data falls under "No information" category implying we simply do not know the reason for the confirmed case. This in turn will hinder us to reach a conclusion about the aforementioned category.

Furthermore, a substantial amount of people will not always get tested at the first sign of any covid symptoms. People usually wait out a couple of days, hoping it isn't covid and some random flu infection as the symptoms are identical. This causes a delay in the reported number of cases, or they don't get reported at all.

7 Conclusion

In conclusion, the ethics of providing information regrading the Covid-19 cases to the public become a requirement. This paper provides the readers with a detailed information regarding which year and specifically month saw the most cases and which age group it has affected the most. It also shows the factors that could have resulted in the infection however with 44% of the data containing "No information," it is rather hard to tell. Overall, with this information, readers will be able to understand Toronto better and how the pandemic has affected the city.

References

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://allisonhorst.github.io/palmerpenguins/.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.

Xie, Yihui. 2020. "Bookdown: Authoring Books and Technical Documents with R Markdown." https://github.com/rstudio/bookdown.