The COVID-19 Numbers are just the tip of the Iceberg[*]

The data we simply can't trace

Bo Phi

06 February 2022

Abstract

During the recent pandemic, the datasets surrounding the COVID-19 virus have played an important role in navigating the future to move forward around the pandemic, however, the manner in which cases are reported could have effects on the credibility and integrity of the data. This statistical report uses data gathered by the city of Toronto to (1) examine the ways in which the cases have been recorded over time along with the metadata that is attached to each case, and (2), discuss the ways in which the data has an influence on peoples actions, and how those actions are further reflected within the data. With the number of cases having reached a record amount at the end of 2021, it is important to explore possible explanations for why that is the case. Additionally, with many of the direct sources of cases, and many cases also going unreported, it makes it incredibly difficult to estimate the true magnitude of the pandemic in Toronto.

1 Introduction

Since the beginning of 2020, the world has been struck by a global pandemic. A handful of politicians took the early signs of the pandemic very seriously and went into early preventative measures while many politicians did not. As a result, the virus began to spread like wildfire. Before very long, nearly all countries began the process of shutting down, closing nearly all parts of the economy to protect the world from this virus. In the United States, the President openly expressed control over the situation, assuring the cases were being reported but outbreaks were being contained. Soon after that, hospitals began to be flooded with COVID-19 patients and schools began to close as a result. After the development of the vaccine, people have been in a rush to get their immunization but a reasonable chunk of people still opt out of getting vaccines. There are a multitude of reasons for opting out of the vaccine but one of the most common reasons is the lack of belief in the vaccine. People who opt out of vaccines also have a lack of belief in the data and the governmental bodies that collect the data as well. Are their reasons justified?

Two years later in 2022, the same problems are still present. The pandemic has seen no real end and cases have spiked up to the tail end of 2021. Millions of people have died and there are still skeptical people on the validity of the data associated with COVID-19. Good data would serve as an important tool to inform the public and governing officials on choices they make but as it stands, the effectiveness of the current data doesn't seem to be convincing enough people to work towards ending the pandemic. With new variants emerging and the world itching for things to go back to normal, examining why the current model isn't working in an effort to suggest a better one seems appealing but is it even possible?

The importance of truthful pandemic statistics in informing the public cannot be understated as it seemingly determines the rate at which the world can move past COVID-19. For this report, I will be using an open-access dataset from the Toronto Public Health organization to examine the metadata that is associated with a report of a case. More importantly, I will use the data to discuss the ways in which the data is somewhat incomplete and incapable of telling the full story of the pandemic. The data will be processed and analyzed in R (R Core Team 2020) using the 'tidyverse' (Wickham et al. 2019) amd 'dplyr' (Wickham et al.

---

[*]Code and data are available at: https://github.com/bophiwork/paper-1

2021) packages. The packages 'knitr' (Xie 2021a), 'bookdown' (Xie 2020), and 'tinytex' (Xie 2021b) are also used to generate the final markdown report.

2 Data

```
## # A tibble: 6 x 4
##   Assigned_ID `Age Group`    `Source of Infection` `Episode Date`
##         <dbl> <chr>          <chr>                 <date>
## 1           1 50 to 59 Years Travel                2020-01-22
## 2           2 50 to 59 Years Travel                2020-01-21
## 3           3 20 to 29 Years Travel                2020-02-05
## 4           4 60 to 69 Years Travel                2020-02-16
## 5           5 60 to 69 Years Travel                2020-02-20
## 6           6 50 to 59 Years Travel                2020-02-24
```
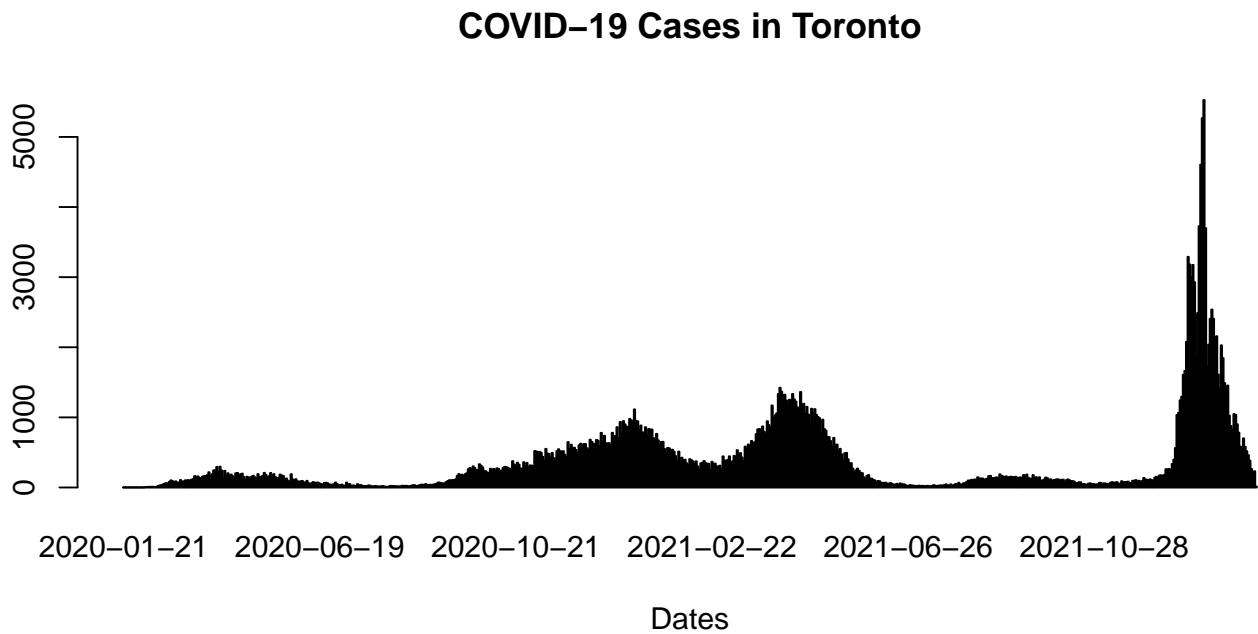
Our data is of penguins (Figure 1).

Figure 1: Episode Dates in Toronto

Talk more about it.

Also bills and their average (Figure 2).
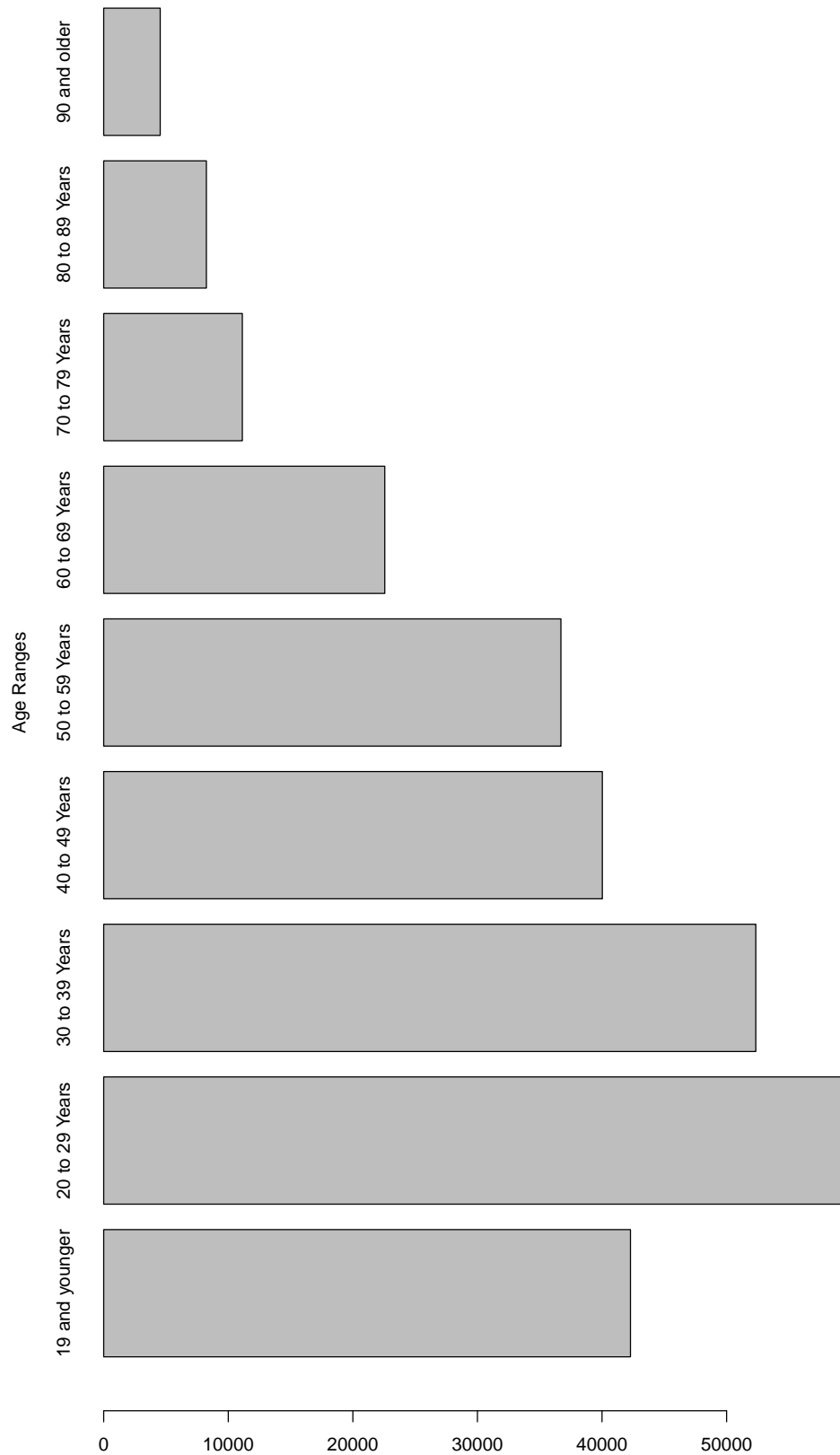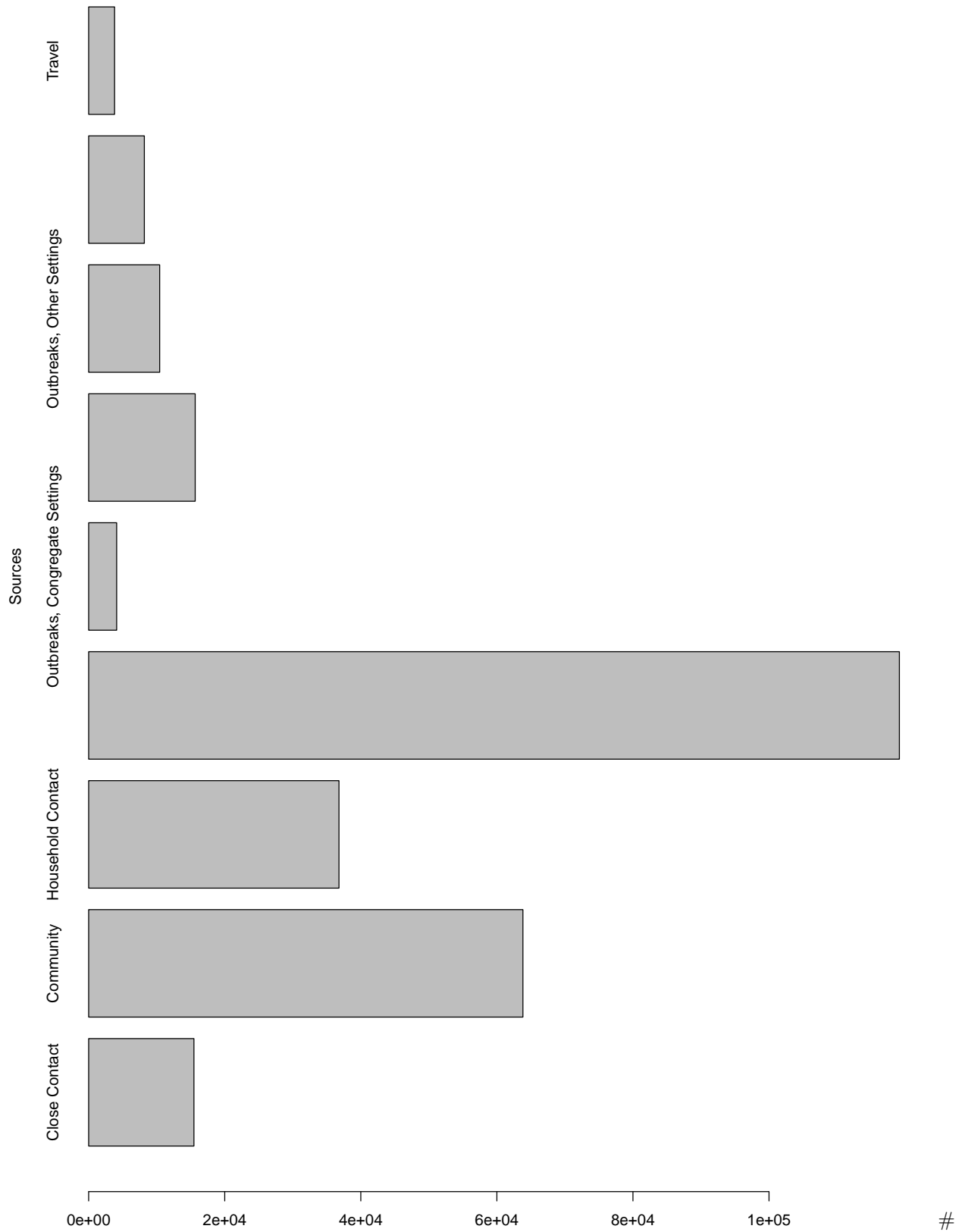
Talk way more about it.

Age Groups Infected by COVID−19

Figure 2: Age Ranges of Infected People

Source of COVID−19 Cases

Sources

Travel

Outbreaks, Other Settings

Outbreaks, Congregate Settings

Household Contact

Community

Close Contact

0e+00   2e+04   4e+04   6e+04   8e+04   1e+05

\#

Model

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \qquad (1)$$

Equation (1) seems useful, eh?

Here's a dumb example of how to use some references: In paper we run our analysis in `R` (R Core Team 2020). We also use the `tidyverse` which was written by (**thereferencecanbewhatever?**) If we were interested in baseball data then (**citeLahman?**) could be useful.

We can use maths by including latex between dollar signs, for instance θ.

3 Results

4 Discussion

4.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

4.2 Second discussion point

4.3 Third discussion point

4.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A   Additional details

References

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, 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.

———. 2021a. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

———. 2021b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents.* https://github.com/yihui/tinytex.