Investigating Toronto COVID-19 Cases in 2020: Age group, Source of Infection, and Whether Outbreak Associated*

Yunkai Gu

September 24, 2024

Process up to now: done drafts for data overview; generated graphs needed; done drafts for appendix; done references

1 Introduction

Still working.

Use cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019a). The remainder of this paper is structured as follows. Section 2….

*Code and data are available at: https://github.com/Kylie309/Toronto-2020-Covid-Cases.

2 Data

2.1 Overview

The analysis uses the dataset of "COVID-19 Cases in Toronto" from the Open Data Toronto portal. The raw dataset contains "anonymized, person-level information for all COVID-19 cases reported from the start of the COVID-19 pandemic in January 2020", which is shared by Toronto Public Health (TPH) and lastly refreshed in February 14, 2024 (Toronto Public Health 2024).

The earliest onset of symptoms of COVID-19 was confirmed to be around 1 December 2019 (by The Lancet) to 8 December 2019 (by WHO), and human-to-human transmission was confirmed by the WHO and Chinese authorities by 20 January 2020 (Wikipedia 2023). Therefore, year 2020 was considered as the first year of COVID-19.

Based on the background of COVID-19, the paper found it meaningful to investigate the COVID-19 cases in Toronto during the first year of the pandemic, which is year 2020. The raw data is then downloaded and cleaned before conducting analysis with tables, graphs and plots.

The analyses presented in this paper were conducted using R programming language (R Core Team 2023). The `tidyverse` (Wickham et al. 2019b) packages were used in the process of data simulation, testing beforehand. Original raw data was downloaded from the Open Data Toronto using the `opendatatoronto` (Gelfand 2022) packages and the `tidyverse` (Wickham et al. 2019b) packages. Then, data cleaning process was done by using the `tidyverse` package (Wickham et al. 2019b) and the `dplyr` package (Wickham et al. 2023). Graphs and plots were made by using the `ggplot2` packages (Wickham 2016).

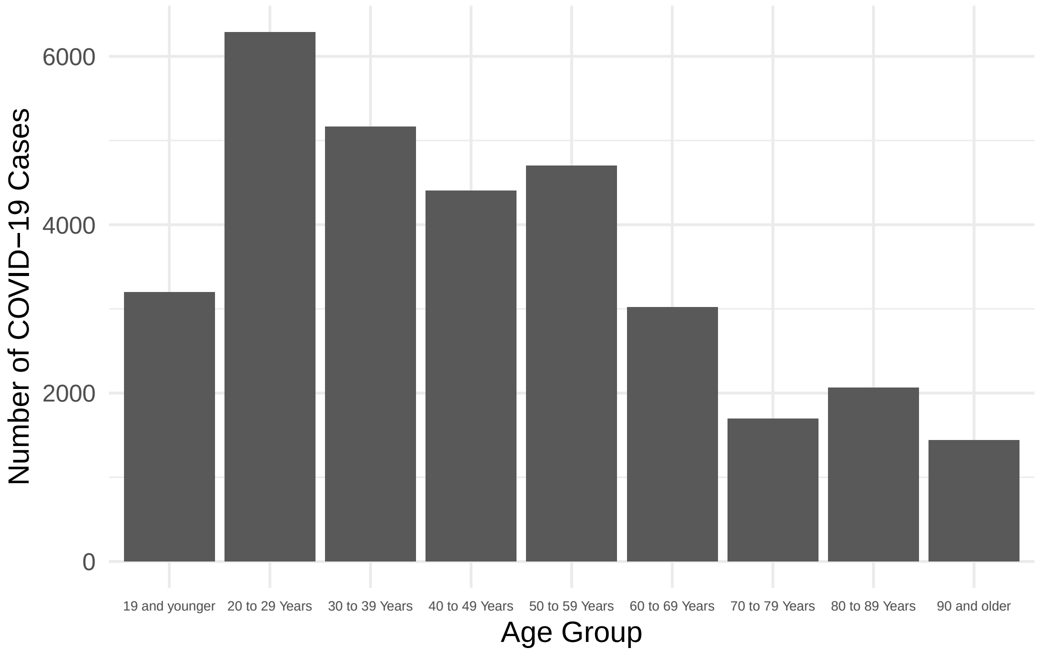


Figure 1: Relationship between covid cases and age groups

Talk way more about it.

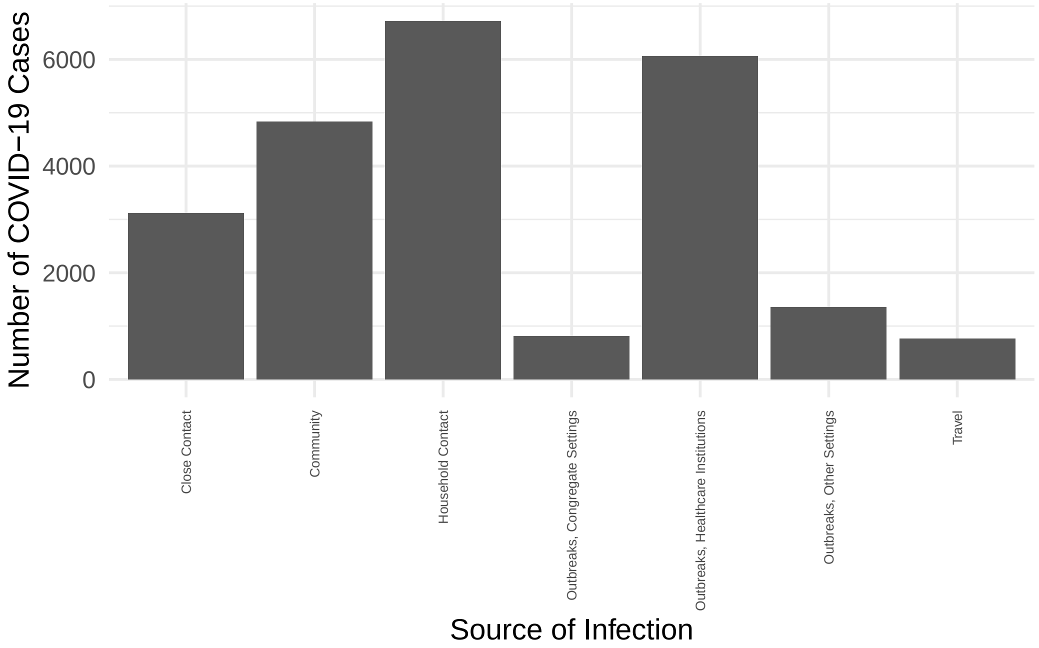


Figure 2: Relationship between Covid cases and source of infection

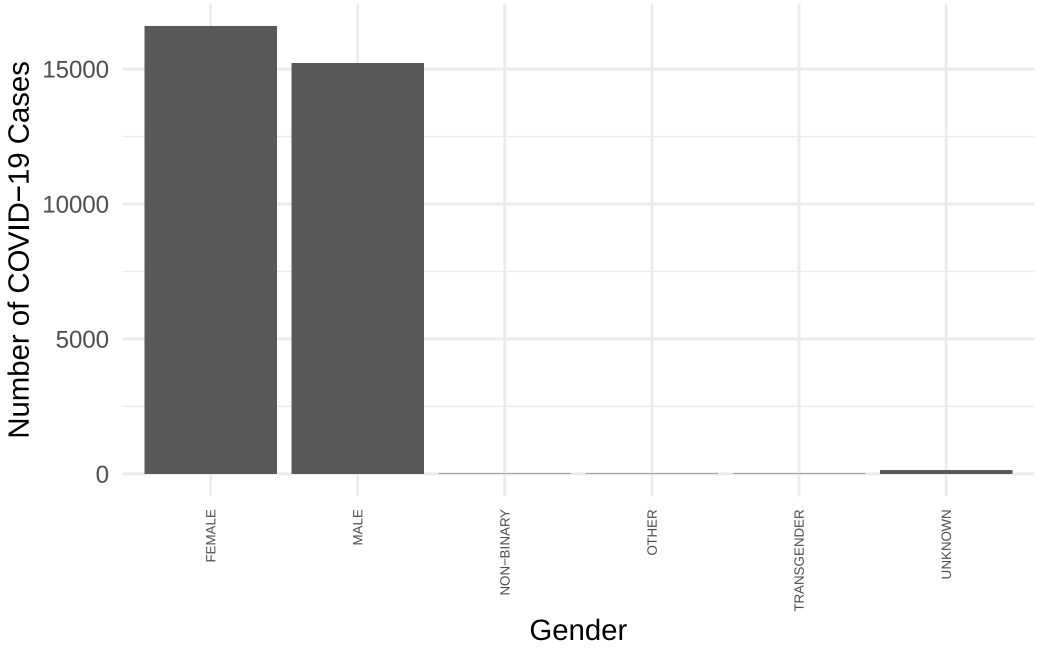


Figure 3: Relationship between Covid cases and gender

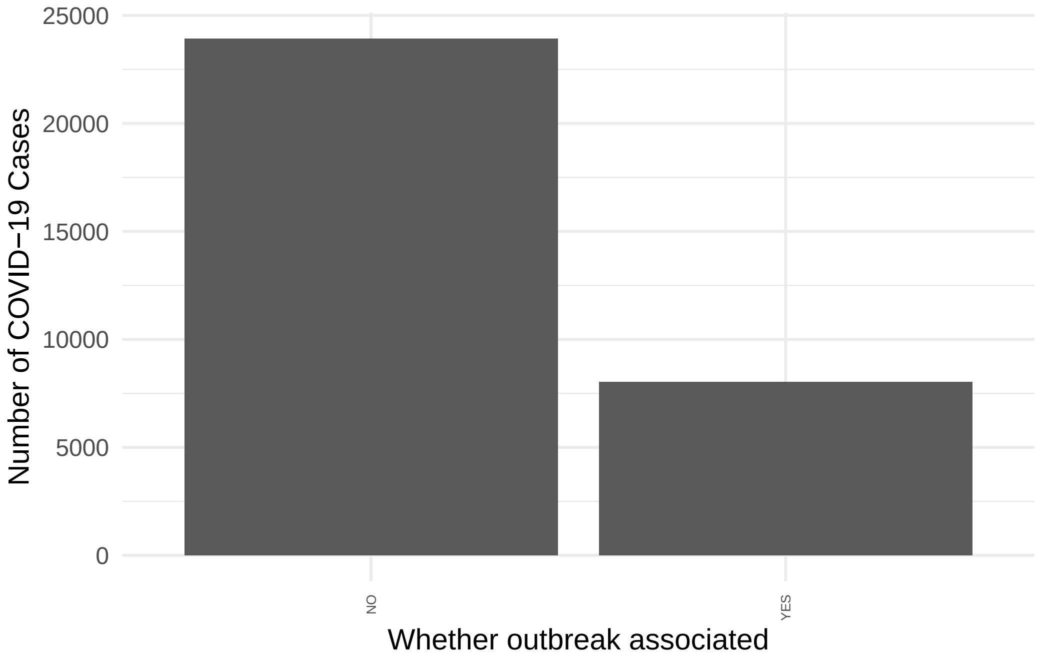
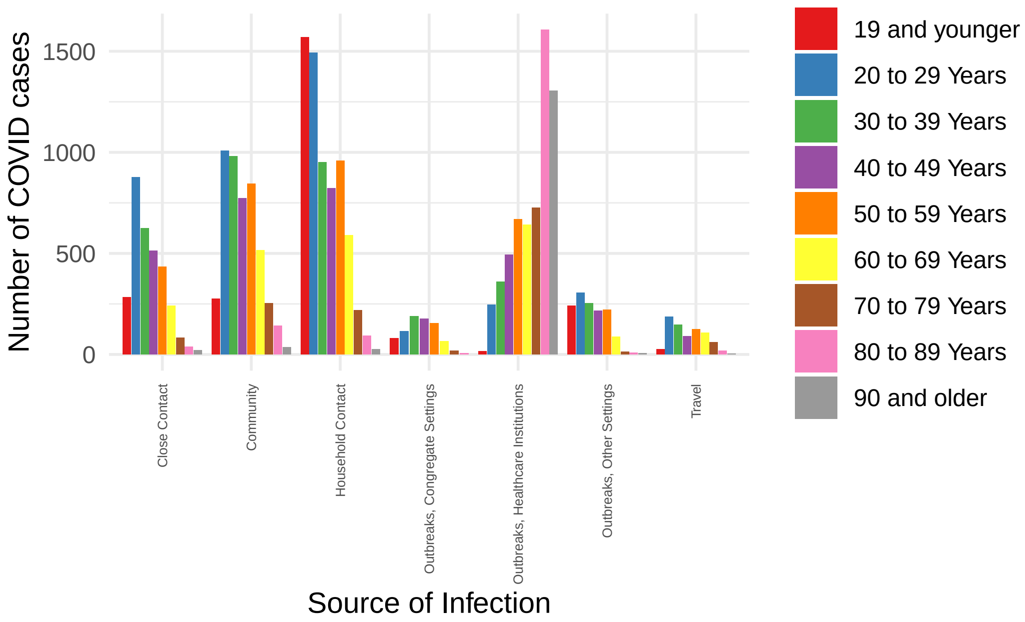
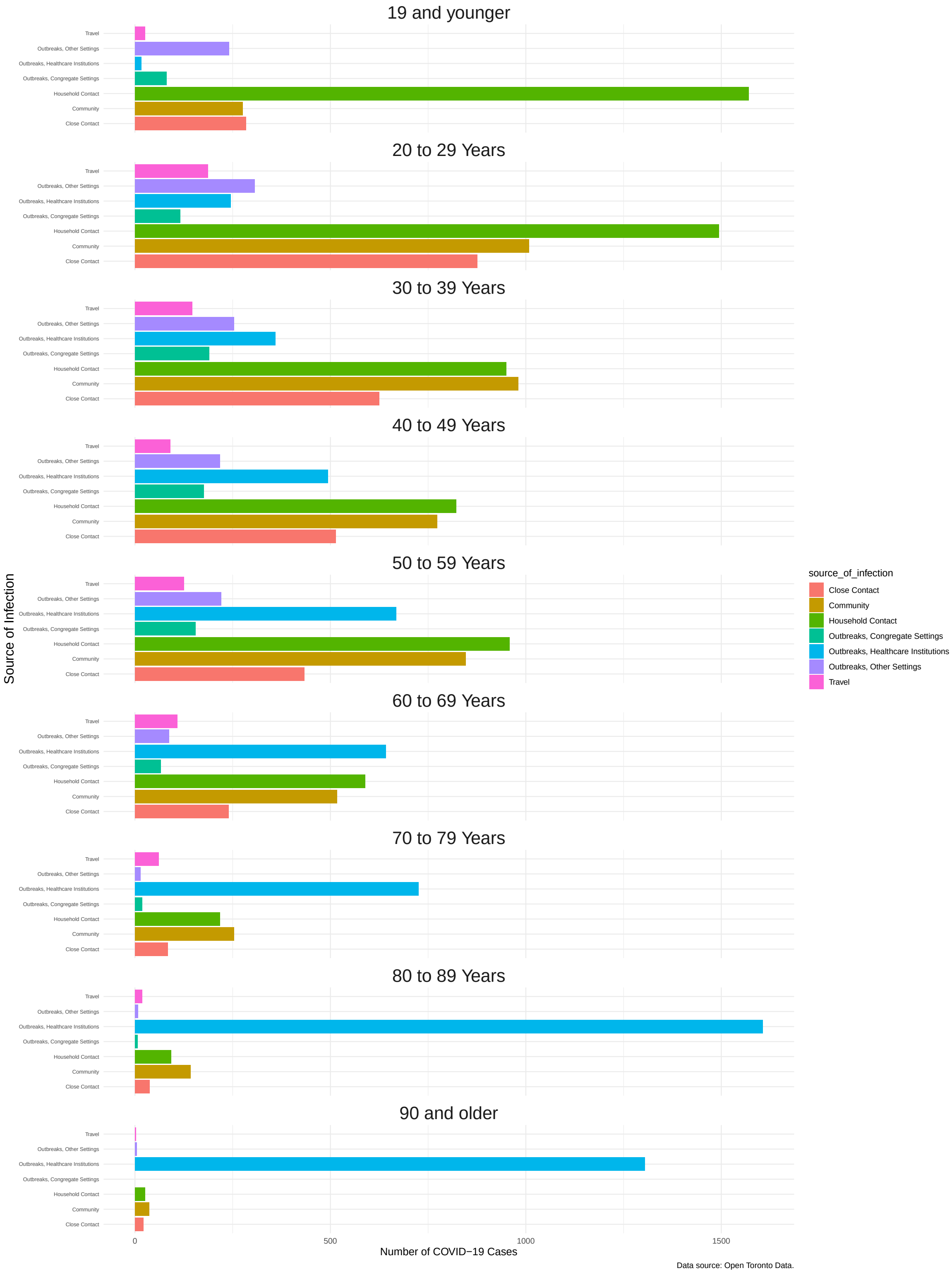


Figure 4: Relationship between covid cases and whether outbreak associated



Data source: Open Toronto Data.

Figure 5: Comparison of different age groups among the same source of infection



Data source: Open Toronto Data.

Figure 6: Comparison of different age groups among the same source of infection

Appendix

A.1 Dataset and Graph Sketches

Target cleaned dataset and processed graphs are available in the Github Repository. See "other/sketches".

A.2 Data Cleaning

The data cleaning process involves the following steps to ensure the integrity and usability of the dataset:

1. Modification of Column Names: The column names are standardized by `janitor::clean_names()` function. The column names are converted into lower cases, and the spaces between words are replaced with underscores, thereby enhancing readability and consistency across the dataset.
2. Separation of Date Columns: The "reported_date" column, originally formatted as "YYYY/MM/DD," is separated into three new columns: "reported_year", "reported_month" and "reported day". The same cleaning process is done to the "episode_date" column. This step improved readability and simplicity for data processing by allowing for more straightforward presentation of temporal variables.
3. Removal of Probable Cases: Any rows where the "classification" column is marked as "probable" are excluded from the cleaned dataset. This step removes any cases that are categorized as probable COVID-19 cases according to standard criteria, thereby enhancing the precision and reliability of analysis by focusing on confirmed cases.
4. Exclusion of Incomplete Age Data: Any rows containing the column "age_group" equals"NA" is removed. This step ensures that the data analysis is not compromised by incomplete information.

A.3 Additional data details

References

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto Public Health. 2024. "COVID-19 Cases in Toronto." https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org.

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

———, et al. 2019a. "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. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org/.

Wikipedia. 2023. "COVID-19 Pandemic." https://en.wikipedia.org/wiki/COVID-19_pandemic.