Analysis of COVID-19 Cases in Toronto During 2020: Trends by Age, Gender, and Sources of Infection*

Yunkai Gu

September 27, 2024

This paper analyzes COVID-19 case data from Toronto in 2020, specifically investigating the trends by age group, gender, and sources of infection. The study finds that younger adults (aged 20 to 29) had the highest case counts, and that there were no significant differences in counts between males and females. The three primary sources of infection of COVID-19 was household contact, community spread, and healthcare institution outbreaks. These results support public health planning, and provide guidelines for future policy decisions and community re-establishment during aftermath of the pandemic.

Table of contents

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

1 Introduction

Since a novel coronavirus was identified after Wuhan Municipal Health Commission in China reported a cluster of pneumonia cases in December 2019, the coronavirus disease 2019 (COVID-19) pandemic has rapidly spread worldwide, resulting in profound consequences in communities, societies and public health systems worldwide. World Health Organization issued the first Disease Outbreak News on 5 January 2020, and declared COVID-19 a pandemic on 11 March 2020, indicating the the alarming levels of spread and severity of the incident (WHO 2020).

In January 2020, the first "presumptive" case of the new coronavirus in Canada was reported in Toronto, which marked the beginning of its spread within the country. As the most most populous city in Canada, Toronto's COVID-19 case data became a critical point of reference, providing insights not only for the province of Ontario but for the country as a whole. Although the pandemic has moved into the recovery stage, analysis of historical COVID-19 case data of Toronto, could support prevention for potential incidents in the future, and contribute to new health policy development and community re-establishment during aftermath of the pandemic.

To support improvements in health system of Toronto, specific analysis on demographic trends in case data is crucial. However, a gap remains in understanding how key factors - age, gender, and sources of infection - contributed to the spread of COVID-19 in Toronto during 2020.

This paper aims to fill this gap. To do this, the paper leverages the dataset of COVID-19 cases in Toronto in 2020 from Open Data Toronto. After cleaning the raw dataset, the analysis of the case data starts from the examination of distribution of cases by age group, gender and source of infection. The findings show that there was higher case counts of younger adults (particularly those aged 20 to 29), no significant differences in counts between female and male cases, and the three main sources of infection were household contact, community and outbreaks in healthcare institutions. A detailed examination of outbreak-related cases shows that healthcare institutions were the most common source of outbreak-related infections.

Next, the study compares case counts of different age groups within the same source of infection. The main findings include that younger teenagers (those aged 19 and younger) represented a significantly high proportion of cases that were infected due to household contact, and that individuals aged 20 to 29 years were the leading group within the cases with infection sources

of “Community”, “Close Contact” and “Travel”. A similar comparison of counts of different infection sources within the same age group is then conducted. It reveals that there was a decreasing trend between number of cases with the source of infection “Household Contact” and the age, while case counts linked to healthcare institution outbreaks increased with age.

The paper structured as follows: Firstly, Section 2.1 provides a description of the dataset. This includes introduction of data source and data cleaning process (Section 2.1.1), and explanations of variables and measurement (Section 2.1.2). Then, Section 2.2 displays all visualizations of the data, including distribution by age groups (Section 2.2.1), distribution by gender (Section 2.2.2), and distribution by sources of infection (Section 2.2.3). Then, Section 3 presents the results of data analysis. Section 3.1 discusses how counts of COVID-19 cases were associated with outbreaks; Section 3.2 compares the counts of COVID-19 cases of different age groups within the same source of infection; and Section 3.3 compares the counts of cases of different sources of infection within the same age group. Finally, Section A includes supplementary information on data cleaning process (Section A.1) and summarized statistics of each variable (Section A.2). References of the paper are listed at the end.

2 Data

2.1 Overview

2.1.1 Data Source and Cleaning

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 this 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. Specifically, what groups of people were more vulnerable to coronavirus (which gender, which ages stratum), and which source was most likely to cause infection. 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` packages (Wickham et al. 2019) were used in the process of data simulation, testing beforehand. Original raw data was downloaded from the Open Data Toronto using the `opendatatoronto` package (Gelfand 2022) and the `tidyverse` package (Wickham et al. 2019). Data cleaning process was done by using `tidyverse` package (Wickham et al. 2019), `dplyr` package (Wickham et al. 2023) and `lubridate` package (Grolemund and Wickham 2011). Then, tables were made with `knitr` package (Xie 2024), and graphs were made with `ggplot2` package (Wickham 2016) and `tibble` package (Müller and Wickham 2023).

Table 1 shows the sample of dataset after cleaning. See Section A.1 for detailed steps.

Table 1: Sample of Cleaned Data of COVID-19 Cases in Toronto During 2020

Age Group	Source of Infection	Gender
50 to 59 Years	Travel	FEMALE
50 to 59 Years	Travel	MALE
20 to 29 Years	Travel	FEMALE
60 to 69 Years	Travel	FEMALE
60 to 69 Years	Travel	MALE
50 to 59 Years	Travel	MALE

One additional notes have to be made on the variables: the data cleaning process does not involve removal of all rows with missing values. Specifically, in the analysis dataset, there are still missing values in "Age Group" column (marked as NA), "Source of Infection" column (marked as "No Information") and "Gender" column (marked as "UNKNOWN"). This could be seen in the below summary tables, where there are still counts of cases which are lack of information. This is because only the rows that lack information in all three columns are removed while cleaning data. The purpose is to maintain data integrity. The trends of COVID-19 cases by age, gender and source of infection are analyzed separately before making cross analysis in the paper, and therefore every case which has value in at least one variable is kept to ensure more complete and precise analysis of pattern with strong statistical power. Removing cases with simply one or two variable values missing discards a significant portion of the dataset, which could lead to biased results.

Following tables present the summarized counts of each value of different variables. These tables provide a basic overview of the data and simplifies the analysis. See Section A.2 for the summarized statistics of each variable.

Firstly, Table 2 shows the counts for cases of each age group.

Table 2: Summary of Counts for Cases of Each Age Group

Age Group	Number of Cases
19 and younger	7616
20 to 29 Years	12378
30 to 39 Years	10254
40 to 49 Years	8764
50 to 59 Years	9164
60 to 69 Years	5874
70 to 79 Years	3105
80 to 89 Years	3177
90 and older	2112
NA	38

Table 3 shows the counts for cases of each gender classification.

Table 3: Summary of Counts for Cases of Each Gender

Gender	Number of Cases
FEMALE	31779
MALE	30462
NON-BINARY	1
OTHER	12

Table 3: Summary of Counts for Cases of Each Gender

Gender	Number of Cases
TRANSGENDER	10
UNKNOWN	218

Finally, Table 4 shows the counts for cases associated with each source of infection.

Table 4: Summary of Counts for Cases with Each Source of Infection

Source of Infection	Number of Cases
Close Contact	4800
Community	9895
Household Contact	11779
No Information	21885
Outbreaks, Congregate Settings	1089
Outbreaks, Healthcare Institutions	8724
Outbreaks, Other Settings	3392
Travel	918

2.1.2 Variable and Measurement

Table 1 shows the sample of cleaned data for the analysis. Each row represents one COVID-19 case reported in Toronto during 2020. "Age group" column refers to the patient's age at the time of illness, divided into 10-year range (19 and younger, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90 and older). If unknown, it would be marked as blank (NA). "Gender" column refers to the self-reported gender of the patient, classified as male, female, non-binary, transgender, other. If unknown, it would be marked as "UNKNOWN". "Outcome" column records whether the case has been resolved or has left fatal consequences (for example, death). "Source of Infection" column identifies the most likely way that the case acquired COVID-19 infection, which includes: close contact, community, household contact, travel, outbreaks (congregate settings), outbreaks (healthcare institutions), outbreak (other settings). If unknown, it would be marked as "No Information".

To avoid ambiguity, each source of infection is defined as below (Toronto Public Health 2024):

- Close Contact: Case who acquired infection from a close contact with a confirmed or probable COVID-19 case (e.g. co-worker).
- Community: Cases who did not travel outside of Ontario, did not identify being a close contact with a COVID-19 case, and were not part of a known confirmed COVID-19 outbreak.

- Household Contact: Case who acquired infection from a household contact with a confirmed or probable COVID-19 case (e.g. family member, roommate).
- Travel: Case that travelled outside of Ontario in the 14 days prior to their symptom onset or test date, whichever is the earliest.
- Outbreaks, Congregate Settings: confirmed outbreaks in Toronto in shelters, correctional facilities, group homes, or other congregate settings such as hostels or rooming houses.
- Outbreaks, Healthcare Institutions: confirmed outbreaks in Toronto in long-term care homes, retirement homes, hospitals, chronic care hospitals, or other institutional settings.
- Outbreaks, Other Settings: confirmed outbreaks in Toronto in workplaces, schools, day cares, or outbreaks outside of Toronto.
- No information: Cases with no information on the source of infection.

Note that according to Toronto Public Health (Toronto Public Health 2024), the most likely way that cases acquired their COVID-19 infection is determined by examining several data fields including: (1) A public health investigator's assessment of the most likely source of infection; (2) Whether being associated with a confirmed COVID-19 outbreak; (3) Reported risk factors such as contact with a known case or travel. If the public health investigator's assessment is absent, then the other data fields are used to infer source of acquisition using the following hierarchy: Travel > Outbreak > Household Contact > Close Contact > Community > No information (Cases with episode dates before April 1 2020) or Outbreak > Household Contact > Close Contact > Travel > Community > No information (Cases with episode dates on or after April 1 2020). The measurement of other variables (age, gender) are self-reported.

It could be seen that the assessment by public health investigators plays a significant role in determining the most likely source of infection. This type of subjective judgement introduced potential bias to the dataset, and could be affected by multiple aspects such as limited information. Similar ethical challenges also occurred due to the rigid hierarchy set by TPH, where the role of different sources in spreading COVID-19 would be overestimated or underestimated. This could largely influence public responses and lead to unfair legislation, policies and restrictions to vulnerable groups, negatively affect public trust to the health system of Toronto. Moreover, though the dataset was anonymized before publishing, privacy patient was always a main concern. The raw dataset published by TPH included the neighborhood name of each case, and such geographic data made it possible to track more private information of the patients, creating privacy leak issues.

2.2 Visualization

2.2.1 Distribution by Age Groups

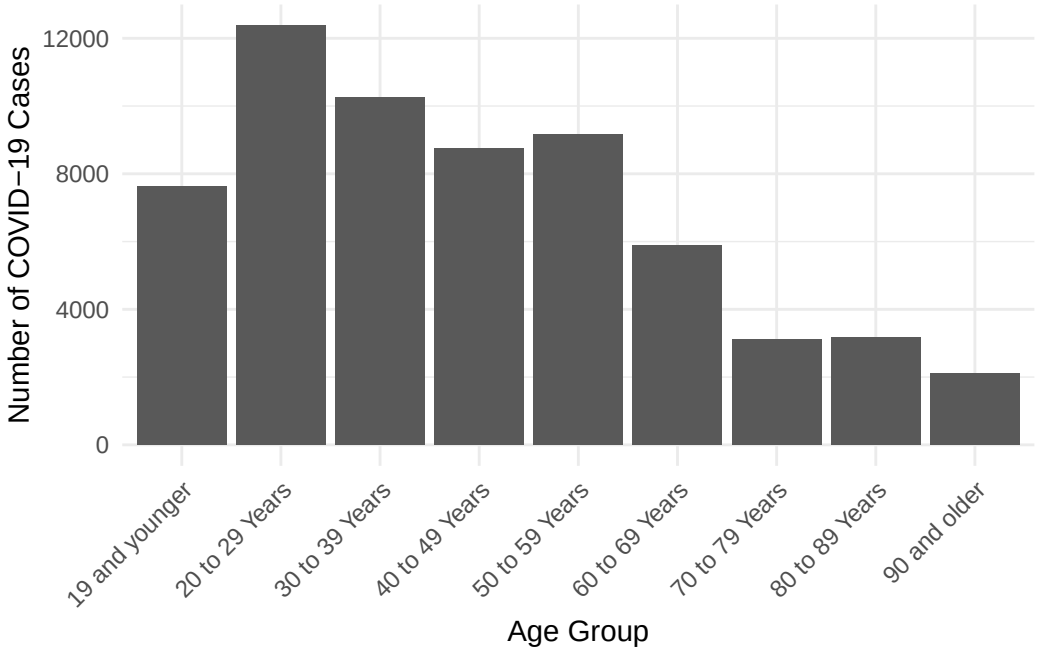

Figure 1: Distribution of COVID-19 Cases by Age Groups

Figure 1 illustrates the distribution of COVID-19 cases in Toronto during 2020 by age groups. The age group with the highest number of COVID-19 cases is 20 to 29 years, with specifically 12378 counts (approximately 19.8% of all cases), followed by the age group of 30 to 39 years (16.4%) and the age group of 50 to 59 years (14.7%). There is a noticeable decreasing trend in case numbers as age increases beyond 50 years, with the age group of 90 and older having the lowest proportion of cases.

2.2.2 Distribution by Gender

Figure 2 illustrates the distribution of COVID-19 cases in Toronto during 2020 by gender. The proportion of female and male cases are relatively similar (with specifically 30462 male cases and 31799 female cases). There is a total of 23 non-binary, transgender and other cases, representing a tiny proportion of the total number of COVID-19 cases.

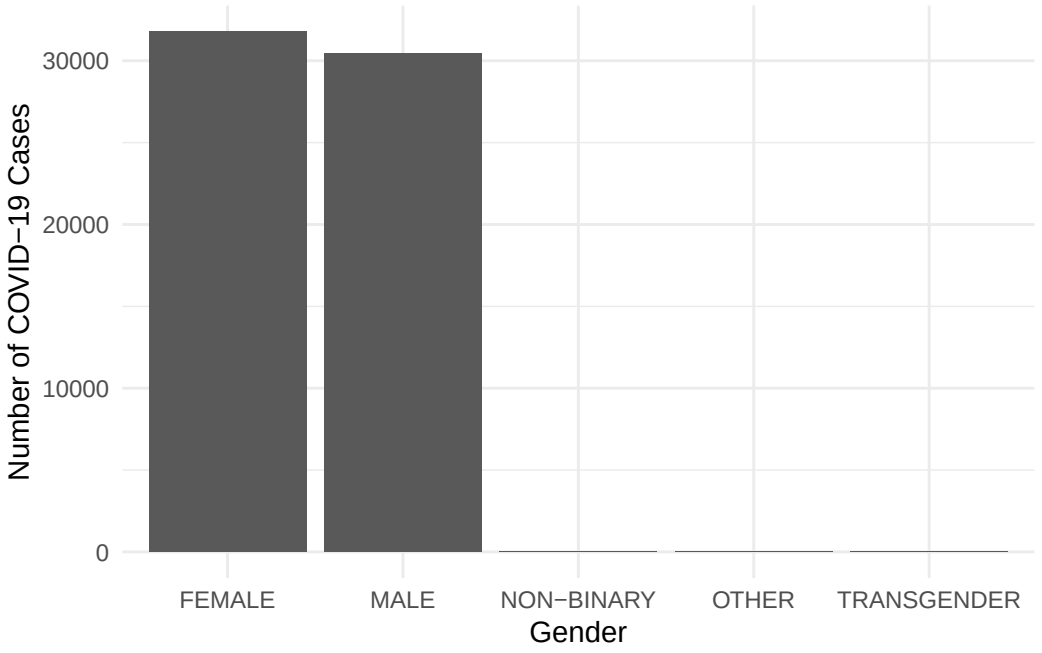

Figure 2: Distribution of COVID-19 Cases by Gender

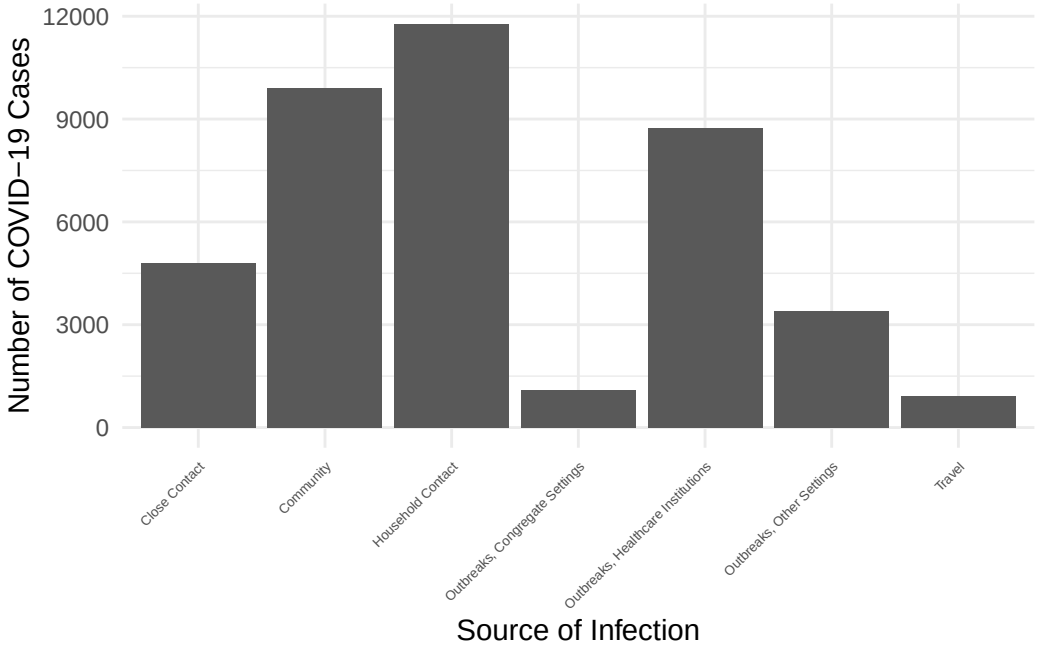

Figure 3: Distribution of COVID-19 Cases by Source of Infection

2.2.3 Distribution by Sources of Infection

Figure 3 illustrates the distribution of COVID-19 cases in Toronto during 2020 by source of infection. "Household Contact" is the most common source, with specifically 11779 cases (29.0%). The second common source is "Community" (9895 cases, 24.4%), followed by "Outbreak in healthcare institutions" (8724 cases, 21.5%). These three leading sources of infection have significantly higher proportion than others, and almost three-fourths of all cases were driven by them. On the other hand, the travel-related cases, with fewer than 1,000 cases, are notably less significant.

3 Result

3.1 Whether Outbreak Associated

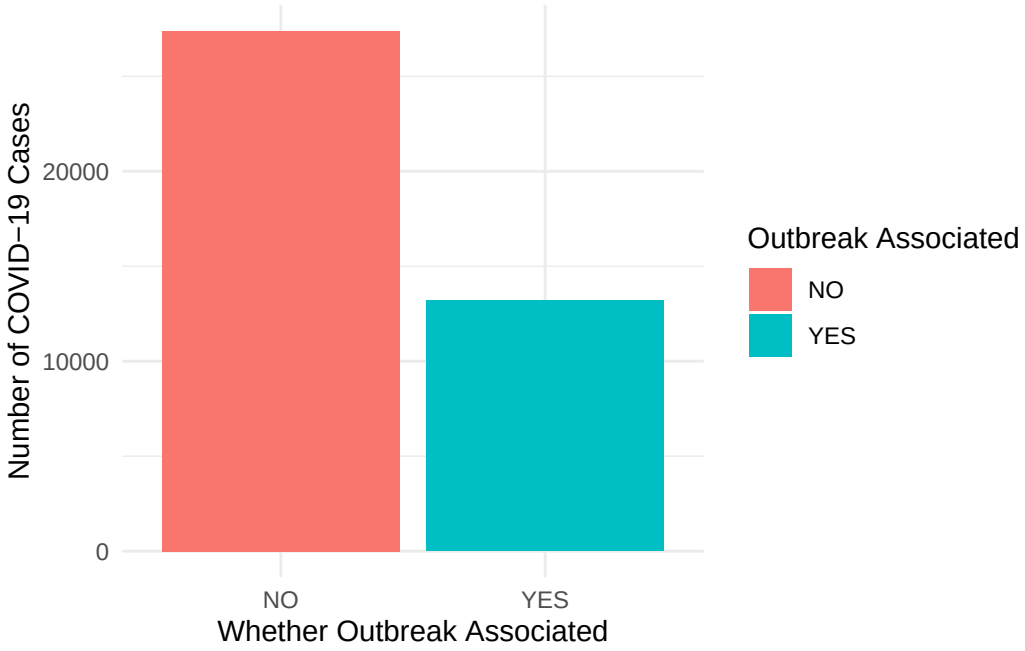

Figure 4: Counts by Whether COVID-19 cases Were Associated with Outbreaks or Not

Figure 4 compares the number of cases that were associated with outbreaks (i.e. "Outbreaks, Congregate Settings", "Outbreaks, Healthcare Institutions" or "Outbreaks, Other Settings") to the number of cases that were not. The figure shows a smaller proportion of cases that were associated with outbreaks (with 13205 cases, 32.5%), approximately half of the number of non-outbreak-related cases.

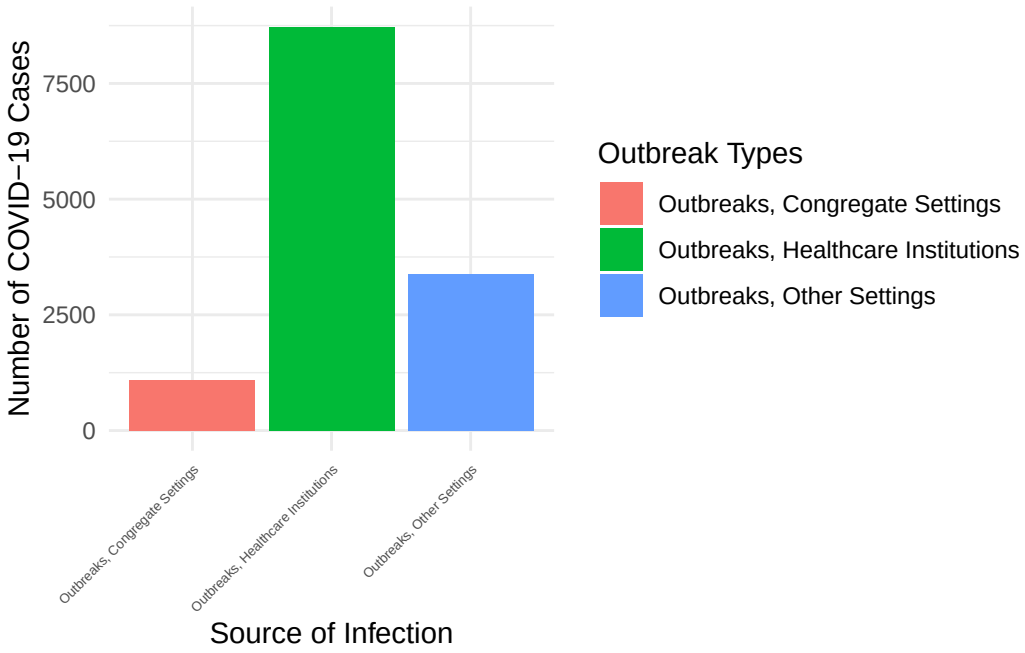

Figure 5: Distribution of COVID-19 Cases by Outbreak Types

Recall the distribution of cases by sources of infection shown in Section 2.2.3 above. By selecting out the three distinct sources of infection that are associated with outbreaks, Figure 5 highlights the distribution of COVID-19 cases by outbreak types. Among all outbreak-related cases (13205 in total), the outbreaks that happened in healthcare institutions, such as long-term care homes, retirement homes, and hospitals, were the most common sources of infection (8724 cases, 66.1%), accounting for more than half of all cases. Outbreaks that happened in congregate settings, such as shelters, correctional facilities, and group homes, were the least common ones (1089 cases, 8.24%).

3.2 Different Age Groups Within the Same Source

Figure 6 compares the counts of cases across different age groups within the same source of infection. One of the key observations is that among all cases with "Household Contact" as their source of infection, individuals aged 19 and younger represented significantly high proportion (3587 cases, 30.5%). The age group of 19 and younger was also the most likely to get infected by COVID-19 due to "outbreaks in other settings", such as outbreaks in workplaces, schools, and day cares in Toronto. Additionally, the age group of 20 to 29 years acted as the leading group for the sources of infection of "Community", "Close Contact" and "Travel". The source of infection "Outbreaks, Health Institutions" had age group of 80 to 89 years to be the primary group, followed by that of 90 and older.

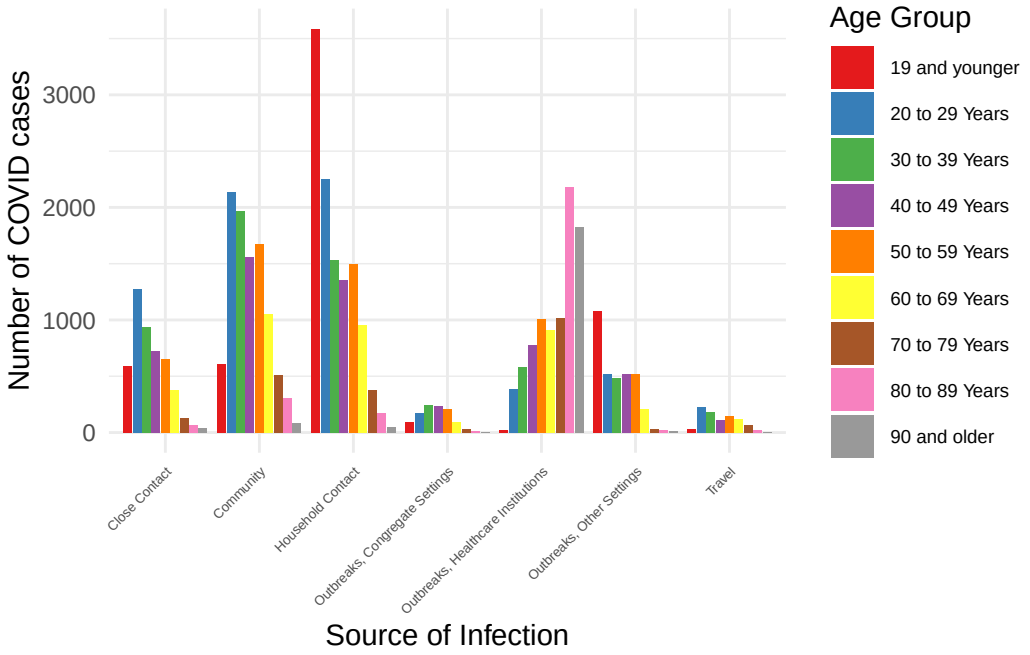

Figure 6: Comparison of Case Counts Across Different Age Groups Within the Same Source of Infection

Figure 7 shows a clearer comparison by plotting separate bar plot for each source of infection vertically.

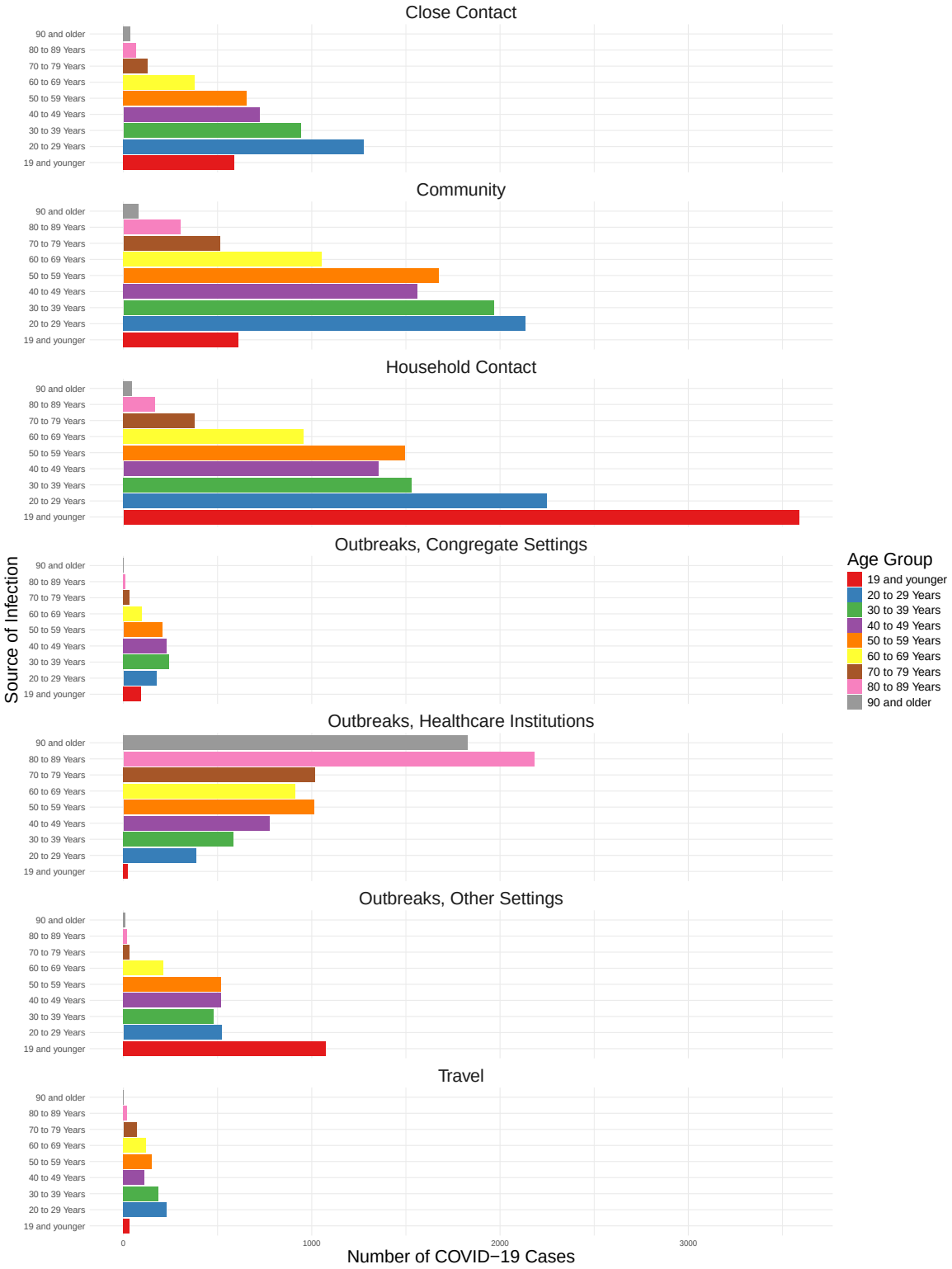

Figure 7: Comparison of Case Counts Accorss Different Age Groups Within the Same Source of Infection

3.3 Different Sources Within the Same Age Group

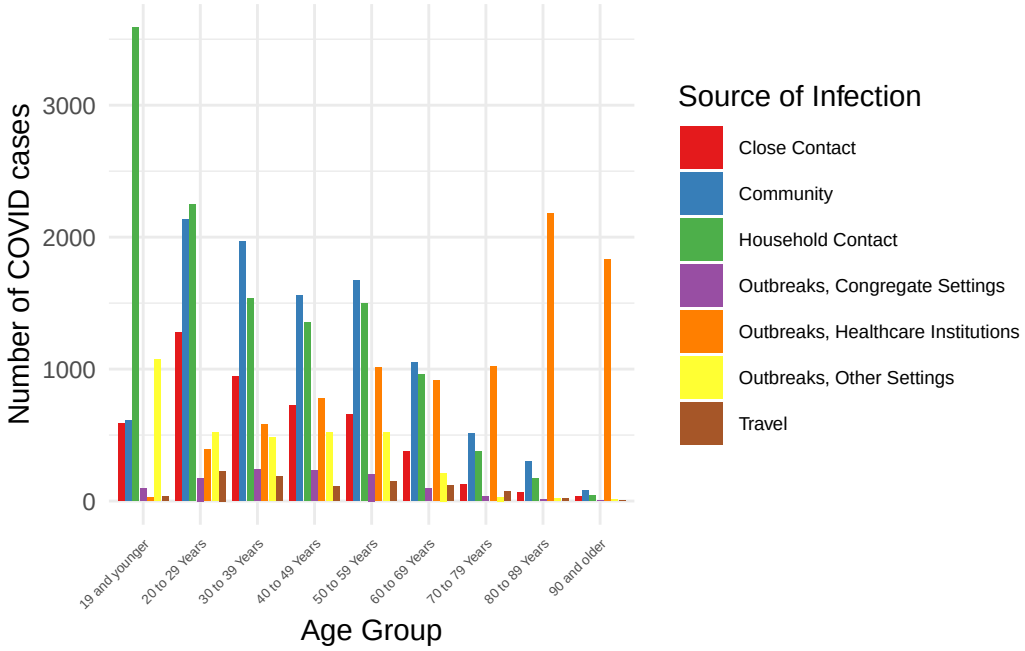

Figure 8: Comparison of Case Counts Across Different Sources of Infection Within the Same Age Group

Figure 8 compares the counts of cases from different sources of infections within the same source of infection. One of the key observations is that the source "Household Contact" dominates across a majority of age groups, particularly the younger ones. For age group of 19 and younger, the case counts and the proportion of cases related to the source is significantly high (3587 cases, 47.1%). There is a downward trend between the number of cases from "Household Contact" and the age, as the decreasing length of the green bar in Figure 8 shows: as age increases, the number of cases with "Household Contact" as a source of infection decreases. A similar decreasing trend is discovered between number of "Close Contact" cases and the age, and between the number of "Community" cases and the age, especially among people aged 20 to 80.

On the other hand, the trend for cases counts related to "Outbreaks, Healthcare Institutions" and the age reveals to be upward in the graph. Starting from the age group of 19 and younger, the length of orange bar increases with the age. For people over 80, the source dominates (with 2181 cases, 68.6% for 80-89 age group and 1828 cases, 86.6% for 90 and older age group).

Figure 9 shows a clearer comparison by plotting separate bar plot for each source of infection vertically.

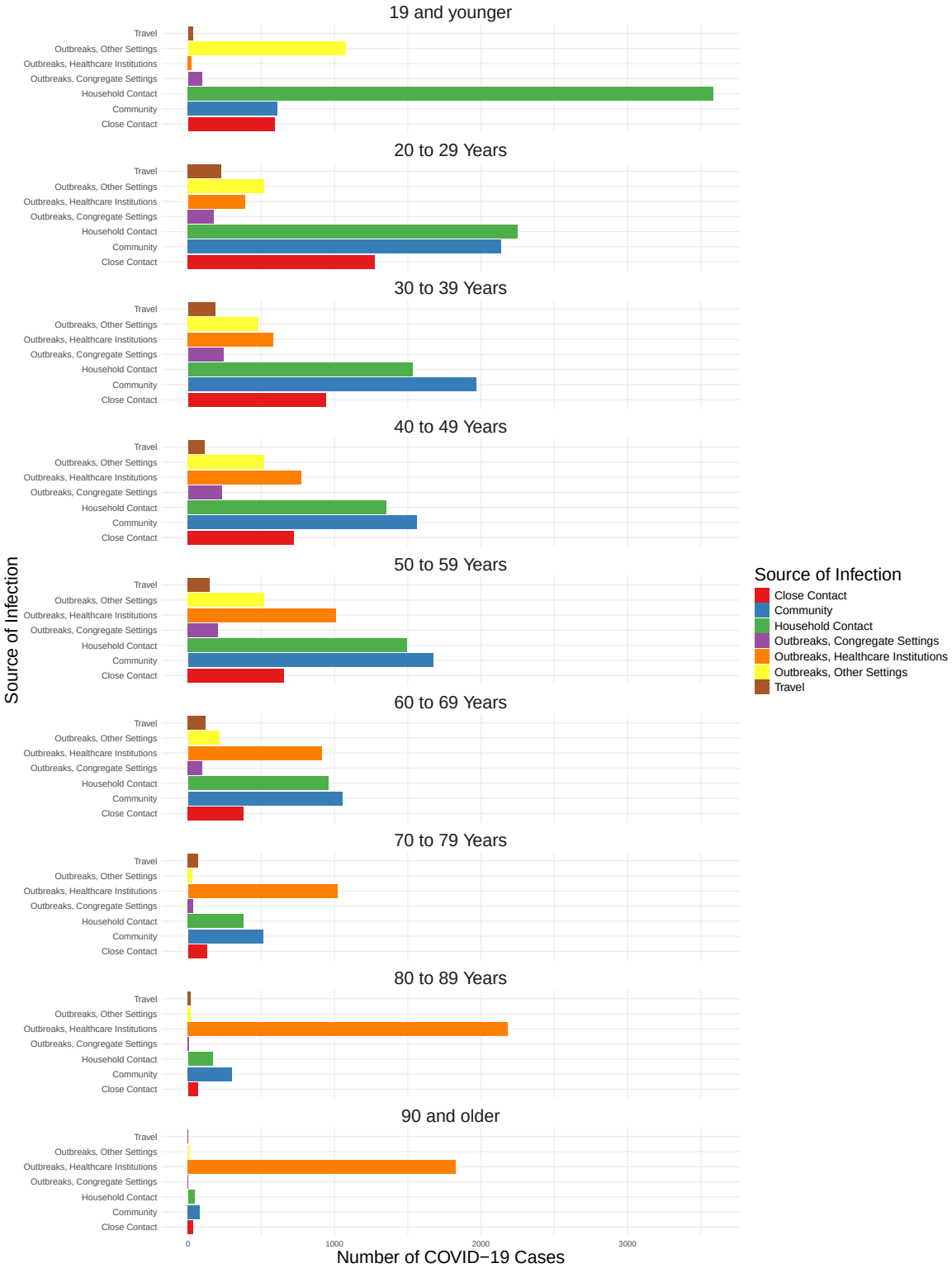

Figure 9: Comparison of Case Counts Across Different Sources of Infection Within the Same Age Group

A Appendix

A.1 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. Specifically, 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". 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. Selection of Required Variables: The data analysis conducted in the paper only focuses on the COVID-19 cases in 2020, and examines 3 key variables: age group, gender and source of infection. The cleaning process removes all irrelevant rows and columns, improving the clarity of the cleaned dataset used for analysis.
5. Exclusion of Incomplete Data: As mentioned in Section 2.1.2, the data cleaning process does not involve removal of all rows with missing values. Instead, only the rows that lack information in all three of the key variables (age group, gender, and source of infection) are removed. The purpose is to maintain data integrity and to minimize the potential biases appeared in results.
6. Standardization of Values in the Gender Column: After reviewing the dataset cleaned by the previous five steps, it is identified that one entry in the "gender" column writes "TRANS WOMAN". To avoid redundancy and simplify the subsequent analysis and plots, this value is recategorized under "TRANSGENDER", so that the "gender" column includes only 5 distinct categories with no overlap or ambiguity in definitions.

A.2 Summarized Statistics for Each Variable

Following tables show the summarized statistics for counts of each variable. Table 5 is for "Age Group", Table 6 is for "Gender", and Table 7 is for "Source of Infection".

Table 5: Summarized statistics for Counts of Cases of Each Age Group

Mean_Age	Median_Age	Min_Age	Max_Age	sd_Age
6248.2	6745	38	12378	4020.083

Table 6: Summarized statistics for Counts of Cases of Each Gender

Mean_Gender	Median_Gender	Min_Gender	Max_Gender	sd_Gender
10413.67	115	1	31779	16045.06

Table 7: Summarized Statistics for Cases with Each Source of Infection

Mean_Source	Median_Source	Min_Source	Max_Source	sd_Source
7810.25	6762	918	21885	6972.308

References

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

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/.

WHO. 2020. "Archived: WHO Timeline - COVID-19." https://www.who.int/news/item/27-04-2020-who-timeline---covid-19.

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, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

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

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.