Toronto COVID-19 Outbreak Analysis*

Comprehensive analysis of data collected since January 2020

Jingxian Zhai

25/04/2022

Abstract

The hottest topic in the world right now should be COVID-19, and it's the same in Toronto. The covid19 outbreak began in early 2020, during which the city of Toronto also shut down more than once. To overcome the virus, we must first get to know it. Through the data found on the website–open.toronto.ca, we can get a clear situation of the number of infected people in Toronto which began to be collected in January 2020 by Case & Contact Management System (CCM). In this paper, we find the main reasons which could lead to be infected are: (1) age; (2) sex; (3)source of infection.

Contents

*Code and data are available at: https://github.com/JingT13/sta304_final.git

1. Introduction

The New Year 2020 is not just the beginning of a new year, it is the beginning of a very stressful period and this is what the COVID-19 brings. In early 2020, the death toll in the US topped 190,000, and Canada also reached 9,200. Until now, the virus is still constantly mutating. Since we cannot reject the virus, we must find a suitable way to coexist with the COVID-19. The first thing to do is to study the impacts of the virus brings to people and try to find a good way to survive in those effects. Therefore, we found data on the virus on the website of open.toronto.ca for a specific analysis.

In the open.toronto.ca, we choose the dataset, COVID-19.cases, to analyze. The data began recording COVID-19 cases in Toronto in January 2020. The variables include age group, infection status, gender, family information, and the route of transmission of the virus, etc. This dataset was collected by Case & Contact Management System we also call it CCM. In this paper, we will focus on the aspects of age group, source of infection and sex.

This paper is organized as follows: First we will create a model to see the connection with each variable. Then we will make a short introduction to the raw data in the data section. We will demonstrate the data with tables, line graphs, radar charts and histograms in the paper and make an analyzation about the data. In the part of discussion, we will talk about whether these three variables is related to the confirmed cases and show the limitations and expectations.

2 Data Section

To get a better understanding of covid-19, I used data of covid-19 cases provided by the Provincial Case & Contact Management System (CCM) on the opentoronto website. In this dataset, the raw data consisted of 32,000 variables, then, we cleaned and extracted important data to start our analysis. In the analysis, we will use R statistical language (R Core Team 2020), tidyverse packages (Wickham et al. 2019), devtools (Wickham et al. 2021), dplyr (Wickham et al. 2022), fmsb (Nakazawa 2022), janitor (Firke 2021), kableExtra (Zhu 2021), gt (Iannone, Cheng, and Schloerke 2022), readr (Wickham and Hester 2020), ggplot2 (Wickham 2016).

First, we selected from the raw data all the variables that we will use for the analysis in relation to hospitalizations due to covid-19, age group, gender and source of infection. Since this dataset does not have a specific quantitative expression, before analyzing this set of data, we will first count the quantitative values of each variable to facilitate our graphing and analysis. Therefore, we have new data representations, we count the number of people who were hospitalized for men and women; various age groups and different sources of infection. Then the new dataset removes NA values and some unknown values or values which do not match the purpose of this report.

In the data section, first, we used gt to draw tables to make a brief summary of this set of data. In addition, we used ggplot2 to draw one vertical bar chart to describe the number of hospitalizations for men and women and one horizontal bar chart to describe the number of hospitalizations by age groups. In the end, we used fmsb to build a radar chart to describe hospitalization data for each source of infection. From the table in Figure 1, we found that after the data of no information was taken out, the most likely source of infection to be hospitalized is travel (13.92%). Moreover, men are more likely to be infected than women and the older you are, the more likely you are to get infected.

Table1: Summary of Source ofinfection
Data source: provincial Case & Contact Management System (CCM)

Source of Infection	number of People	Percetage	Total
Travel	103	13.92%	740
Community	613	10.29%	5959
Outbreaks Congregate Settings	108	13.17%	820
Household Contact	306	4.91%	6228
No Information	300	3.41%	8788
Outbreaks Healthcare Institutions	763	13.81%	5526
Outbreaks Other Settings	33	3.07%	1076
Pending	5	11.36%	44
Close Contact	153	4.91%	3119

Table2: Summary of Client Gender
Data source: provincial Case & Contact Management System (CCM)

Client Gender	number of People	Percetage	Total
FEMALE	760	4.49%	16936
MALE	901	6.09%	14816
NON-BINARY	0	0.00%	4
OTHER	0	0.00%	6
TRANSGENDER	0	0.00%	5
UNKNOWN	0	0.00%	233

Table3: Summary of Age Group
Data source: provincial Case & Contact Management System (CCM)

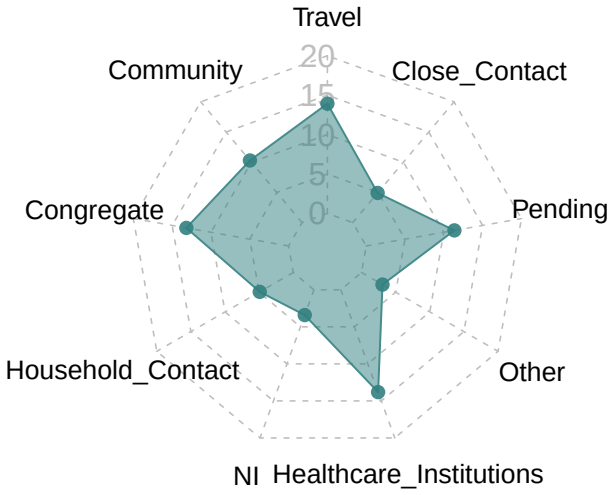
Age Group	number of People	Percetage	Total
19 and younger	23	0.74%	3112
20 to 29 Years	43	0.71%	6047
30 to 39 Years	64	0.12%	5526
40 to 49 Years	102	2.21%	4613
50 to 59 Years	167	3.69%	4529
60 to 69 Years	295	10.12%	2914
70 to 79 Years	347	19.41%	1787
80 to 89 Years	413	20.23%	2042
90 and older	211	15.49%	1362

3 Result

3.1 Source of Infection

In this radar chart we could find that there are 8 main different types of sources of infection. They are travel, close contact, community, congregate, household contact, healthcare institutions, pending and others. Depending on table 1 and figure 1, we find the the percentage of people who were hospitalized related to covid-19 due to travel and heathcare institutions are about the same number, 13.92% for travle while 13.81% for heathcare institutions. Another large percentage is aggregation and the two very small percentages areclose contact and household contact. From here we can see that interacting with a large number of different strangers increases the probability of being discharged from the hospital due to infection.

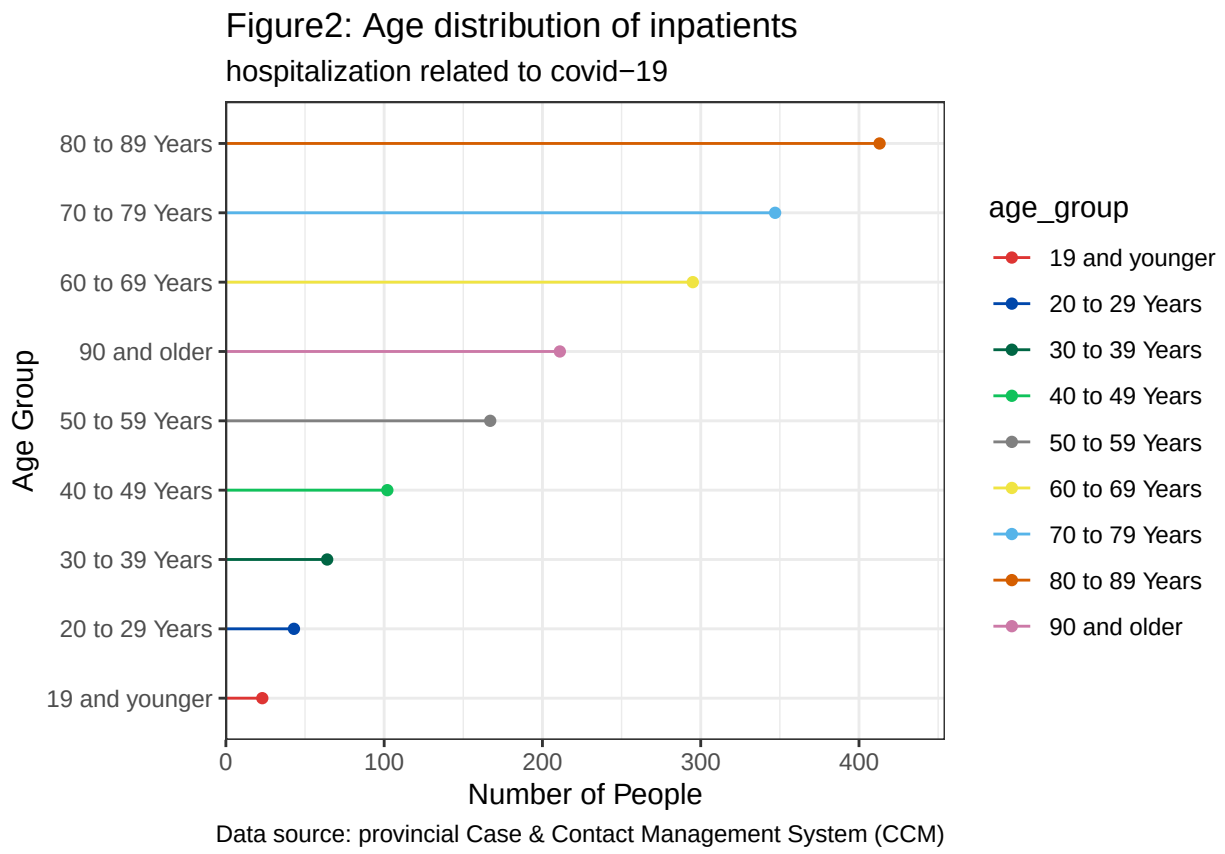
Figure1: Source of infection



Note: NI = No Information

3.2 Age

In this figure we could see that the biggst numnber of people of age group sho were hospitalized due to covid-19 is from 80 to 89 (413). In fact, there was no big difference in the number of people over 60 being hospitalized for infection, with older people being more likely to be hospitalized than younger people. We found that among the 19-year-old and younger population, only 23 people were hospitalized with the infection. According to table 3, the percentage is only 0.74% while the the percentage of the infected peopel in 80 to 89 age group is 20.23% which is the largest propotion.



This table 4 shows the proportion of older adults hospitalized with an infection who were admitted to the ICU. We found that the age group above 60 accounted for the largest proportion (29.44%). From this, we can see that if the elderly are infected, they are likely to be critically ill patients who need to be admitted to the ICU for treatment.

Table4: Proportion of infected people admitted to ICU

Note: elderly people

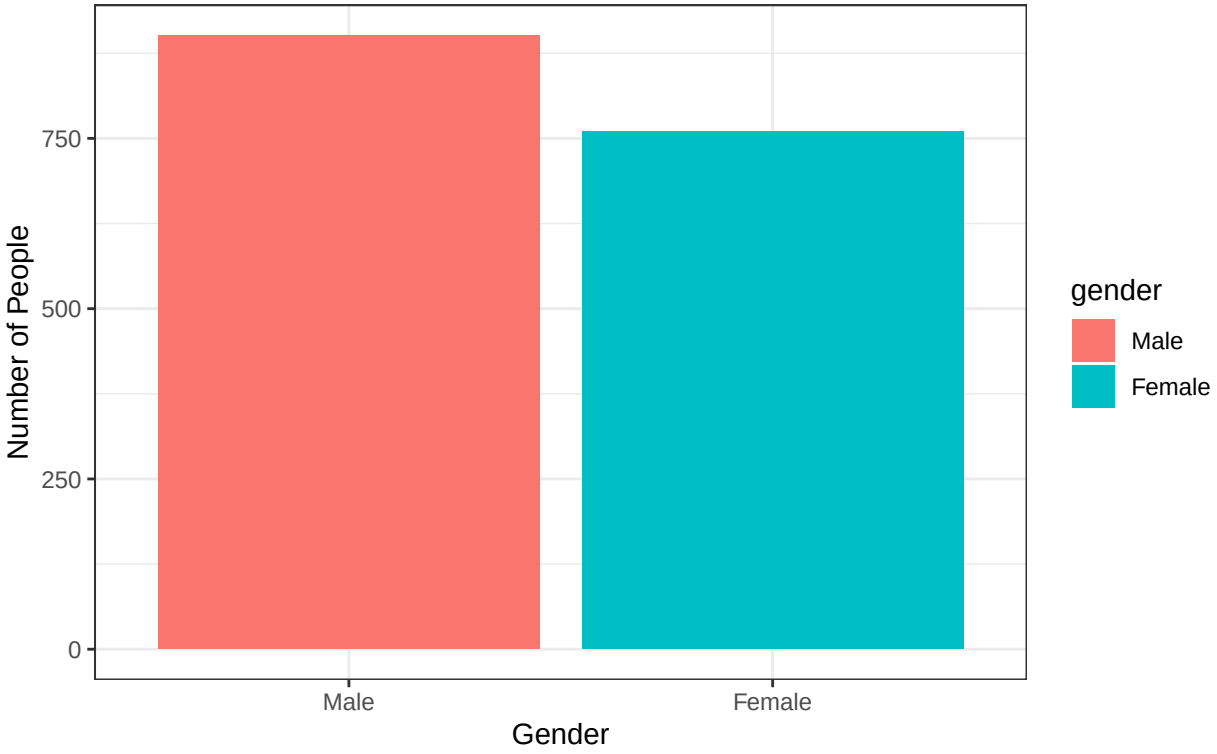
Age	number of People	Percetage
50 to 59 Years	44	23.61%
60 to 69 Years	83	29.44%
70 to 79 Years	77	22.96%
80 to 89 Years	51	11.04%
90 and older	17	4.06%

3.3 Sex

In this figure we find that the number of female who were hospitalized due to COVID-19 (901) is bigger than male (760). Then we also calculated the percentage, the result is the same: the percentage of female ()6.08% is bigger than male (4.49).

Figure3: Gender Distribution of Inpatients Related to COVID−19

Source: provincial Case & Contact Management System (CCM)



4 Discussion

4.1 Source of Infection

What we found in the data section is that travel has the largest proportion of people infected with COVID-19, followed by healthcare institutions, aggregation. According to the data, during the outbreak period, around May 2020, we found that the spread of COVID-19 was largely through air travel. At the time, more than 200 countries were affected and more than 280,000 people died. It has been found that managing closed areas and screening all travelling passengers is the right response (Rahman et al. 2020).

4.2 Age Groups

In the data section we learned that older people over 60 are more likely to be infected with COVID-19 than younger people. Also, people over the age of 65 have an 80% chance of needing hospitalization after being infected and are 23 times more likely to die than younger people. What puts people at risk after being infected are some of the complications that come with it. One of the things that makes patients critically ill with COVID-19 is immune aging of the adaptive immune system. From this point of view, the elderly do greatly increase the probability of being hospitalized, and the possibility of entering the ICU for treatment will also increase (Mueller, McNamara, and Sinclair 2020).

4.3 Gender

We have analyzed in the data section that the number of men hospitalized due to COVID-19 infection is greater than that of women. In fact, some research also mentioned that men are more susceptible to infection than women. There are many factors, such as biological differences between men and women, occupational differences, and whether or not they smoke etc.There is also a different perception of COVID-19. According to data, 59% of women think the virus is very serious, while only 48.7% of men agree with this view (Galasso et al. 2020).

4.4 Limitation and Future

This paper has two limitations: the first is that there are many results such as no information in this set of data, which will affect the final analysis results. For example, in the first part of the infection source, there are 8788 results with no information alone. The second is some factors that are not considered in the dataset. Maybe the infection source has other directions that have not been considered.

Through the above analysis, we found that travel increases the chance of contracting COVID-19. The need for people to reduce the number of trips or even suspend their travel plans during the peak period of the epidemic will have a large degree of protection. In addition, the elderly over 60 years old must take protective measures during the high epidemic period, such as wearing masks when going out and washing hands frequently when returning home.

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset aims to collect, analyze the data of demographic, geographic and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020..
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by the provincial Case & Contact Management System (CCM).
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The creation was funded by Toronto Public Health.
4. *Any other comments?*
    - TBD

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The instances represent the the COVID-19 cases in Toronto. The types are: sources of infection, age groups and gender.

2. *How many instances are there in total (of each type, if appropriate)?*

- There are 32000 instances

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- No.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

- In the raw data, the instance consists of 18 variables.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

- None.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- There is no missing individual instances.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- There are no relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- There are no recommended data splits.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- There are no errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

- There is no confidential data, and the dataset is publicly available.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

- No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- The dataset entirely comprises differnet age groups and differnt sex (men and women) and different sources of infection.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- It is not possible to identify individuals in any way.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- None.

16. *Any other comments?*

- None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data is what people report to the public health department.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Manual human curation.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- None.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- The data are extracted from the provincial Case & Contact Management System (CCM).

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- The data was collected since 2020.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Ethical review processes were not conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- We obtained the data via the website: open.toronto.ca.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- The data is what people report to the public health department. .

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- The individuals consented to the collection and use of their data. The exact language to which consent was granted is not available.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- The mechanism to revoke their consent was not provided.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- The analysis of the potential impact of the dataset and its use on data subjects has not conducted.

12. *Any other comments?*

- None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- We obtained r code by copying the r code on the open.toronto.ca website.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

- The raw data obtained from the PDF is saved in inputs/data/raw_data.csv.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- R Software is avalaible at https://www.R-project.org/

4. *Any other comments?*

- None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

- The dataset has not been used for other tasks yet.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

- TBD

3. *What (other) tasks could the dataset be used for?*

- The dataset can be used for analyze the COVID-19 cases.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- The process of cleaning data is specific to only this table in the original PDF report. This is not suitable in other tables.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- The dataset is not suitable for any other purposes except the COVID-19 cases in the aspects of age group, sex and sources of infection.

6. *Any other comments?*

- None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

- No, this dataset if openly available.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset will be distributed using Github.

3. *When will the dataset be distributed?*

- The dataset will be distributed in April 2022.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- The dataset will be released under the MIT license.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- There are no restrictions.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- None.

7. *Any other comments?*

- None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- Jingxian Zhsai.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- This can be contacted by Github.

3. *Is there an erratum? If so, please provide a link or other access point.*

- There is no erratum available.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- No, the dataset will not be updated.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- This dataset was collected by Toronto Public Health. There are no applicable limits.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

- The older versions of the dataset are not hosted. The dataset somsumers could be able to check the dataset by github.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- None.

8. *Any other comments?*

- None.

Reference

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

Galasso, Vincenzo, Vincent Pons, Paola Profeta, and Martial Foucault. 2020. *Gender Differences in Covid-19 Attitudes and Behavior: Panel Evidence from Eight Countries*. https://doi.org/10.1073/pnas.2012520117.

Iannone, Richard, Joe Cheng, and Barret Schloerke. 2022. *Gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

Mueller, Amber L., Maeve S. McNamara, and David A. Sinclair. 2020. *Why Does Covid-19 Disproportionately Affect Older People?* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7288963/.

Nakazawa, Minato. 2022. *Fmsb: Functions for Medical Statistics Book with Some Demographic Data*. https://CRAN.R-project.org/package=fmsb.

Rahman, Heshu Sulaiman, Ridha Hassan Hussein Masrur Sleman Aziz, Hemn Hassan Othman, Shirwan Hama Salih Omer, Eman Star Khalid, Nusayba Abdulrazaq Abdulrahman, Kawa Amin, and Rasedee Abdullah. 2020. *The Transmission Modes and Sources of Covid-19: A Systematic Review*. ScienceDirect. https://doi.org/10.1016/j.ijso.2020.08.017.

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. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2020. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2021. *Devtools: Tools to Make Developing R Packages Easier*. https://CRAN.R-project.org/package=devtools.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.