Investigating the Impact of Gender and Age on the Severity of COVID-19*

Shuhan Yang

2024-01-24

COVID-19 has greatly impacted our society globally, as this disease spread around the globe, infecting all population. Some previous widespread viruses typically process certain infectious characteristics, such as male can be more likely to get infected compare to female, or children are more likely be infected etc. Thus, this research serves for the purpose to investigate whether COVID-19 has such characteristics, more specifically, the research will focus on the effects of Gender and Age. Note that this report used generative-AI tools for the purpose of data cleaning and graphical analysis.

Table of contents

1.0. Introduction **2**

2.0. COVID19 Data **2**

3.0. Graphical Analysis **2**

3.1 Distribution Analysis 3

3.2. Severity Analysis 5

3.3. Trend Analysis 6

4.0. Conclusion **7**

5.0. Reference **8**

*Code and data are available at: https://github.com/Jasmineee35/Investigating-the-Impact-of-Gender-and-Age-on-the-Severity-of-COVID-19.git

1.0. Introduction

COVID-19 changes our society in a lot of ways, study around this topic is essential as results can better improve the overall health and economic system for nations. Understanding some trends that this virus pocess may contribute the public health and vaccine development.

The COVID-19 data retrieved from opendatatoronto provides a public record of individuals infected by the virous (tested positive). Analyzing some trends in this dataset in an statistical approach can possibly help the toronto public health system prepare for any further global coronavirus: determine effective prevention strategies, identifying high-risk populations, and provide insights to vaccine development.

In the following paragraphs, I will first introduce the dataset then moving on to the initial statistic analysis of the data. First I will use two bar charts to see the overall distribution of the two main variables I want to study for: Age and Gender. Then, using another bar charts, it will indicate the COVID cases by both age and gender groups, showing the female and male distribution for each age interval side by side. Then moving on to the Severity Analysis Section (will include reference later), where additional analysis will be taken to study the specific impact of age to the severity of COVID infections. This particular section will offer insights to identify high-risk groups as well, preparing the toronto public health system. The final section (will add reference later) will touch on trends analysis to summarize COVID development with the passage of time.

2.0. COVID19 Data

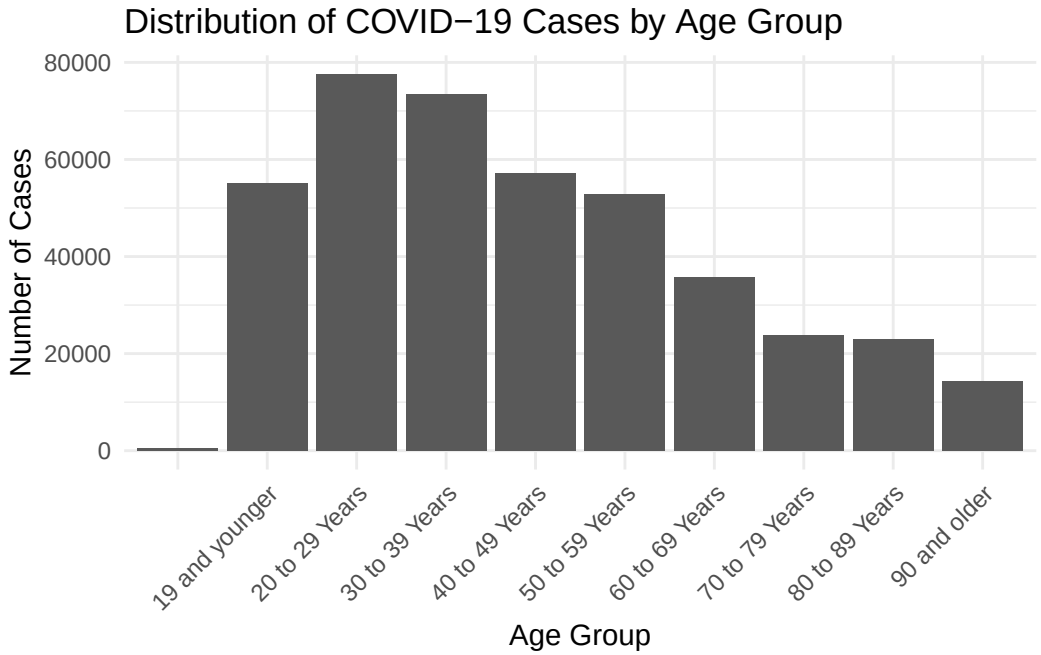
To investigate the COVID infectious trends in Toronto, I retrieved the dataset "COVID19 cases.csv" from the Toronto Open Data Portal (Gelfand 2020). This dataset contains 413474 observations and 15 variables. Among these 15 variables, this report will focus on four majour variables to conduct further analysis: age, gender, severity(ever_hospitalized), date. By using R (R Core Team 2020), and R packages "tidyverse" (Wickham et al. 2019), "dplyr" (**citedplyr?**), "janitor" (Firke 2021), and "knitr" (Xie 2021), I cleaned the dataset by first checking whether there exists missing value, and then convert the date column name to ymd format, finally standardize the column names.

3.0. Graphical Analysis

Having a foundation of what the dataset look-like, we will now come to the graphical analysis part. We will look at the direct distribution of the Age Group and Gender data using a bar chart. A bar chart can provide clear view on the total distribution of the data: mean, median, skewness, shape.

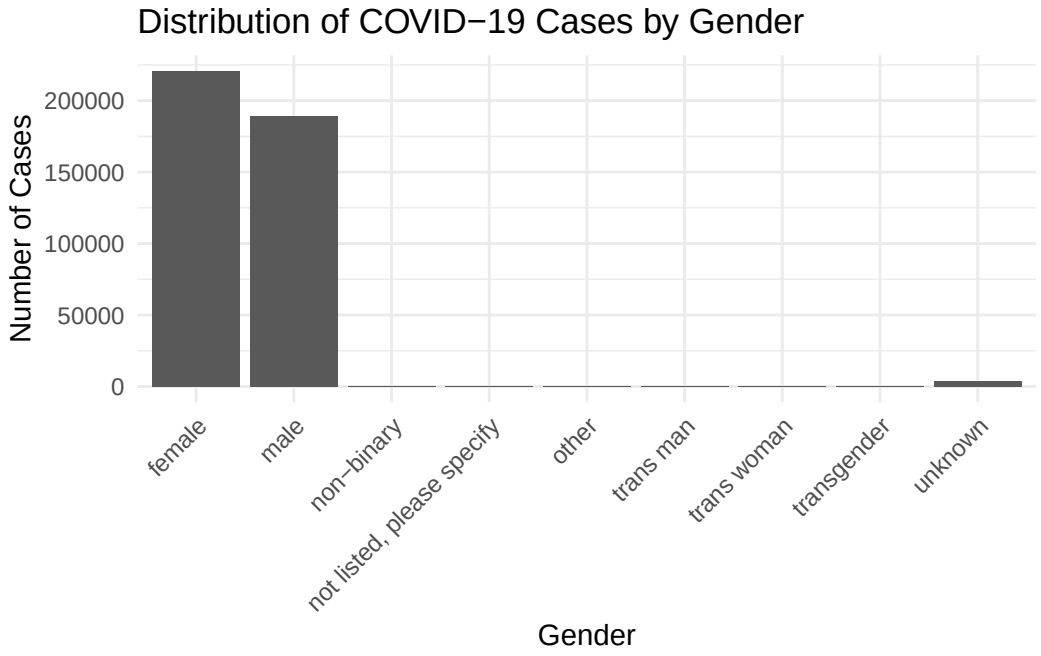
3.1 Distribution Analysis

Bar Chart for Age Group Distribution



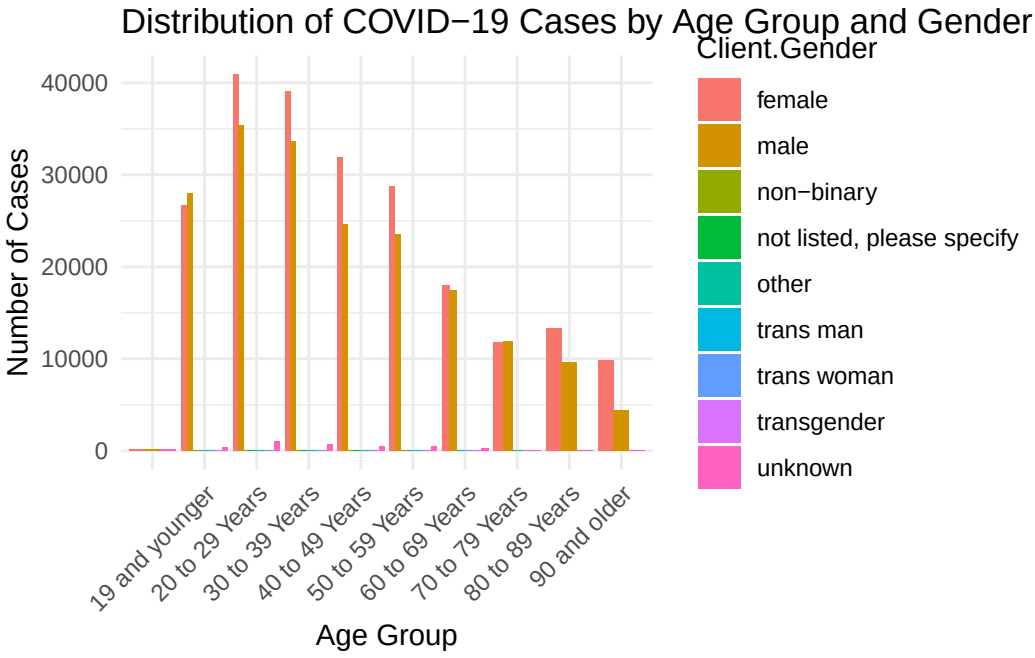
This graph displays the distribution of COVID-19 cases across different age groups.From the bar chart above, the data is positively skewed since it has a longer tail on the right hand side. The skewness aligns with the shape of the distribution that majority of the dataset is concentrated on the left hand side, meaning that more COVID cases are younger generations.

Looking at another bar chart that shows the distribution of COVID-19 cases between different types of genders, the graph shows that within this dataset, more COVID cases are distributed in the female group.



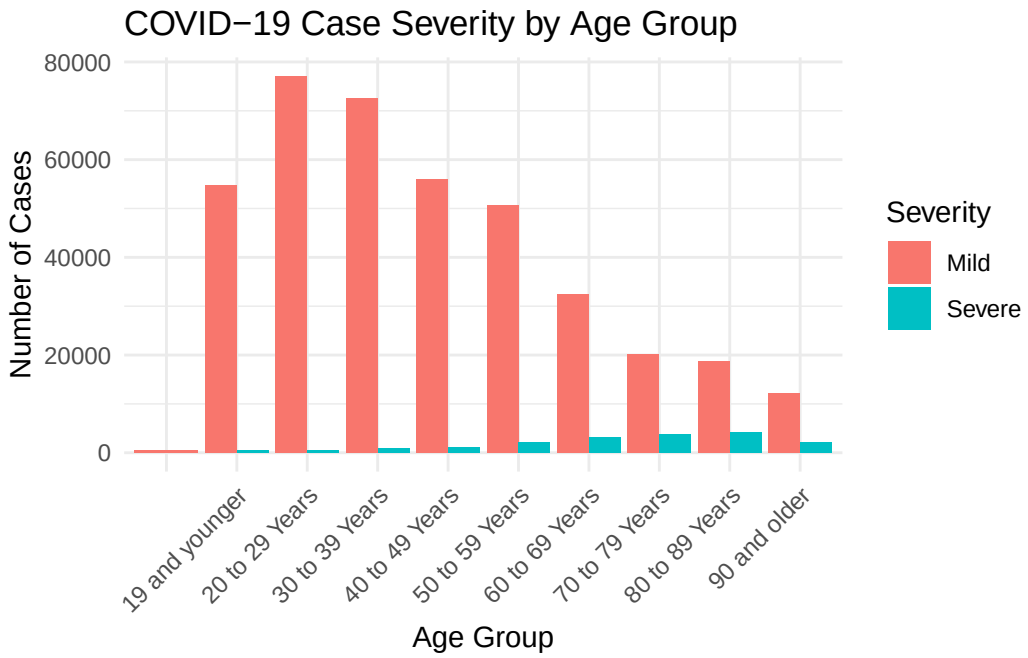
Now we can summarize that based on the initial analysis, female and younger individuals are more likely to get infected by COVID.

To further testify this conclusion, analysis for the data that include both variable: age, gender is conducted. This bar chart will show individuals' gender side by side within each age group interval. Analysing the graph, I found that female typically have more chance to get infected compare to male especially in younger and middle age: 20-59. It also worth noticing that female aging 90 years or older are approximately twice as likely to get COVID than male. Thus if the nation encounter another similar type of virus, the health system can implement provision strategy with a focus on old age females as well.

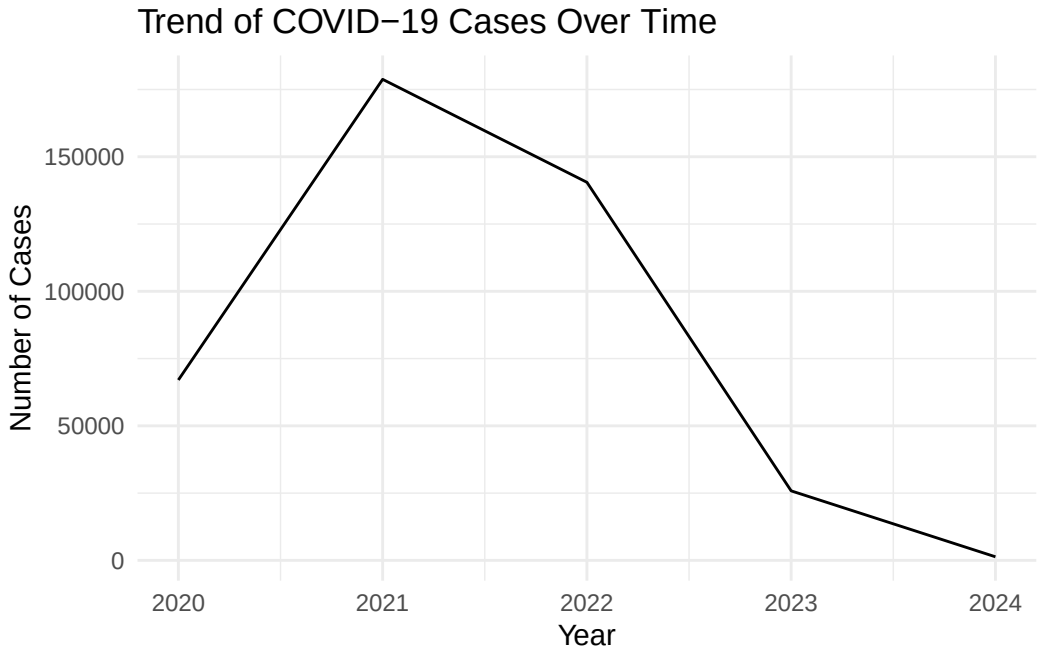


3.2. Severity Analysis

Having a foundation knowledge about the distribution of our dataset, we will now look at how will age group affect the severity of COVID infections. This analytical segment aims to find which type of generation can get more severe if infect COVID. By developing the group most at risk, the public health system know how to allocate resources efficiently, such as the number of beds in hospital needed, to prepare any future virus outbreaks.As when COVID first spread during 2021-2022, hospitals worldwide are full with patients and run out of beds so many individuals are not able to get the treatment they need. For this analysis, I assume that severity is based on hospitalization rates. Observing from the graph, older age group especially 70 to 79 and 80 to 89 generation are severely affected by Covid infections. They exhibit the most hospitalized rate. This aligns with the study (need to find a study and reference later) indicating older generations' immune system and somatic function are weaker than younger generations.

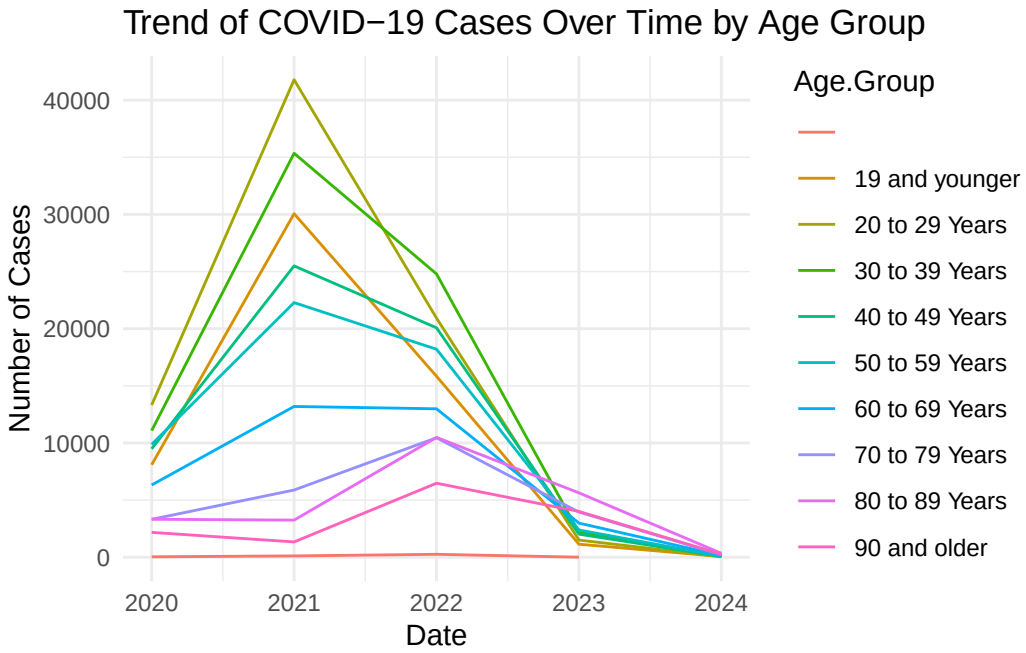


3.3. Trend Analysis



The above graph shows COVID-19 started in 2020, with the infection number peaked around year 2021: approximately 175,000 cases. With the world getting adept to the virus, several

vaccines were developed thus we see in the graph there is a sharp decrease from 2023 to 2024. In the current year, 2024, COVID cases nearly vanished as the World Health Organization announced the end of COVID emergency (cite reference later).



To have a more detail look at each age segment contributing to the overall trend line that we previously discussed, this graph shows a trend line for each age group with the passage of time on the x-axis. Individuals aging between 19 to 39 exhibit the most change in shape. The trends actually align with the overall trend line since the younger generation also consist the majority of COVID cases data (like the discussion in the Distribution Analysis mentioned).

4.0. Conclusion

In conclusion, this investigation on the impact of gender and age on the severity of COVID-19 has yielded several key findings. Firstly, the data indicates younger generations and female are more probable to get COVID infections. While examining the severity analysis section, the outcome reveals that older age group typically can have more severe illnesses as evidenced by their higher hospitalized rate. Additionally, the distribution analysis indicates a notable trend: females with age greater than 90 are almost twice as likely to infect COVID as the male counterparts. Therefore, the most risky group in terms of severity for the past COVID-19 crisis in Toronto is elderly female based on this specific dataset's analysis.

5.0. Reference

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/, https://github.com/sharlagelfand/opendatatoronto/.

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.

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