Predicting Probability of COVID-19 Fatality among Individuals in Toronto Region with Multiple Factors Utilizing Logistic Regression

Nayoung Kim

Dec. 22nd/ 20

Code and data supporting this analysis is available at: https://github.com/nayounganniekim/FinalProject

Abstract

This report analyzes how COVID-19 has impacted individuals in the Toronto, Ontario region in Canada. Specifically, it examines the fatality count of COVID-19, and scrutinizes if certain factors contribute to fatality from COVID-19. Data is obtained from "COVID19 cases" file, and contains the most updated data regarding information about COVID-19 cases in the Toronto region. This analysis utilizes logistic regression to determine the dependent variable that has two categorical levels- whether an individual who is diagnosed with COVID-19 has recovered, or unfortunately has passed away. Because the fatality rate was much smaller compared to the resolved rate, no definite factors illustrated playing a significant role in influencing the fatality rate. This was nevertheless a meaningful examination, as it could be repeated or closely followed if the data set updates with more cases, and the fatality count continues to increase.

Keywords

COVID-19, Cases, Toronto, Fatality, Resolved, Logistic Regression

Introduction

COVID-19 has undoubtedly been one of the most discussed topics of the year 2020. Suggested to have originated from Wuhan, China in December 2019 (World Health Organization, 2020), the disease acquired its abbreviated name from the "coronavirus disease," hyphenated by "19," to represent the year of 2019 in which the disease was originally discovered (World Health Organization, 2020). Shortly after the virus was introduced, it rapidly spread not just across its derived city and country, but also across the entire world. On March 11th, 2020, the World Health Organization announced COVID-19 to be a pandemic (World Health Organization, 2020), confirming the seriousness of the situation globally. As of today, numerous countries around the world continue to suffer from this virus, and Canada is one of them. And Toronto, being the largest city in Canada, is also experiencing high confirmed number of COVID-19 cases daily.

The new virus COVID-19 is known to cause symptoms such as coughs, shortness of breath, high temperature, fatigue, body aches, and more (Government of Canada, 2020). A frightening fact about COVID-19 is that it is possible to infect other even if one is not displaying any of these symptoms (Government of Canada, 2020). Therefore, it may be challenging to figure out exactly where one may have been exposed to the disease, as the exposure could have come from an asymptomatic individual. Because of reasons like these, the Canadian government has outlined that anyone experiencing symptoms related to COVID-19 is to isolate at home for 14 days to avoid potentially spreading to others, which is called quarantining. Due to so many Canadians going into quarantine, the virus has affected how Canadians perform their daily tasks. A vast majority of events that used to happen in person has moved to being online, including work, school, and even doctor's appointments.

Examining COVID-19 cases in Toronto specifically, this analysis will deal with the most updated data set to analyze the fatality and resolved counts. Active cases will not be counted into consideration for this research, as they do not fit in to the result category of COVID-19 for this paper of being either fatal or resolved. The COVID-19 cases data set provides a fairly large data that includes not just the confirmed number of cases, but also other factors such as age, neighbourhood, gender, and more components about the diagnosed patients. This paper will precisely deal with four independent variables- outbreak association, age group, source of infection, and client gender in order to portray a logistic regression model. The goal is to study if these variables have a relationship with the outcome variable of the result of COVID-19, which is either fatality or recovery. More precisely, it is to predict the probability of COVID-19 fatality among individuals in the Toronto region, utilizing these variables through the logistic regression model.

Methodology

Data

As mentioned briefly in the introduction, the data set of COVID19 cases contains data of COVID-19 cases in Toronto, including additional factors of the infected residents. The output variable is the result of COVID-19, whether it was fatality or recovery- a binary outcome. To assist with building a logistic regression model to obtain this outcome, the four auxiliary variables that were utilized were outbreak association of COVID-19, age group of the patient, source of infection, and gender of the diagnosed. I incorporated this analysis because I wanted to examine if there was any variable that is related to or contributes to the outcome of the disease, and logistic regression was the suitable model. This would be helpful to know since the government of Canada is not just trying to halt or decrease the spread of COVID-19, but also is severely concerned about the number of casualty that continues to increase. From the data set, only data that appeared to be clear were chosen- for instance, neighbourhood was not picked because it was unclear if it meant the neighbourhood of the infected's residency or where the individual was exposed to the virus.

Moreover, FSA, or forward sortation area was not employed for the same reason- the partial postal code was not useful if it could not be determined if the FSA is the residency or exposure of the person. Also, several responses to neighbourhood name and FSA were left unfilled, thus another reason these were not considered for independent variables. For sections of ever hospitalized, ever in ICU, and ever intubated, again it was ambiguous if it signified hospitalization ever due to COVID-19, etc., or including other illnesses as well in life. Eliminating these inexact variables, the four left to work with were the current variables. The remaining factors were still interesting ones, nonetheless, since for example, it would be compelling to find out if age is related to the fatality rate. Furthermore, gender was a great variable to use because it not only included male and female, but also comprised of other options such as other, transgender, and unknown, to represent the gender of the patient more inclusively.

Figures one to four illustrate the visualizations of the raw data. For example, Figure 1 displays a histogram of the distribution of the types of outbreaks associated, which are either outbreak associated, or sporadic. The bar graph portrays the number count vertically, which extends to over 30 000. Similarly, the rest of the figures picture each of the remaining independent variables. An interesting fact to note about the distribution of age groups is the highest count group is ages 20 to 29, while the lowest count group is ages 90 and older. Moreover, female and male count show as almost equal in the distribution of gender types, while other, transgender, and unknown types show very little count. Table 1 indicates the fatal and resolved count of COVID-19 cases.

Figure 1: Distribution of the Types of Outbreak Associated

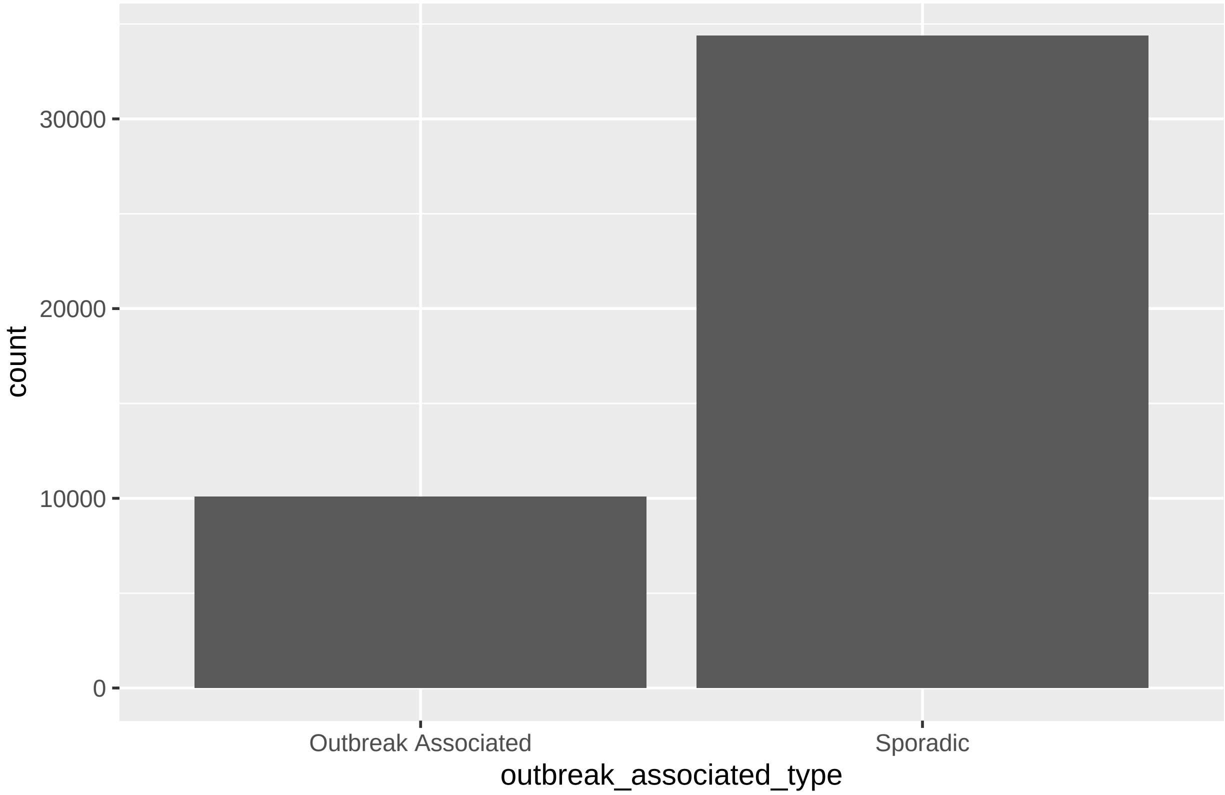


Figure 2: Distribution of Age Groups

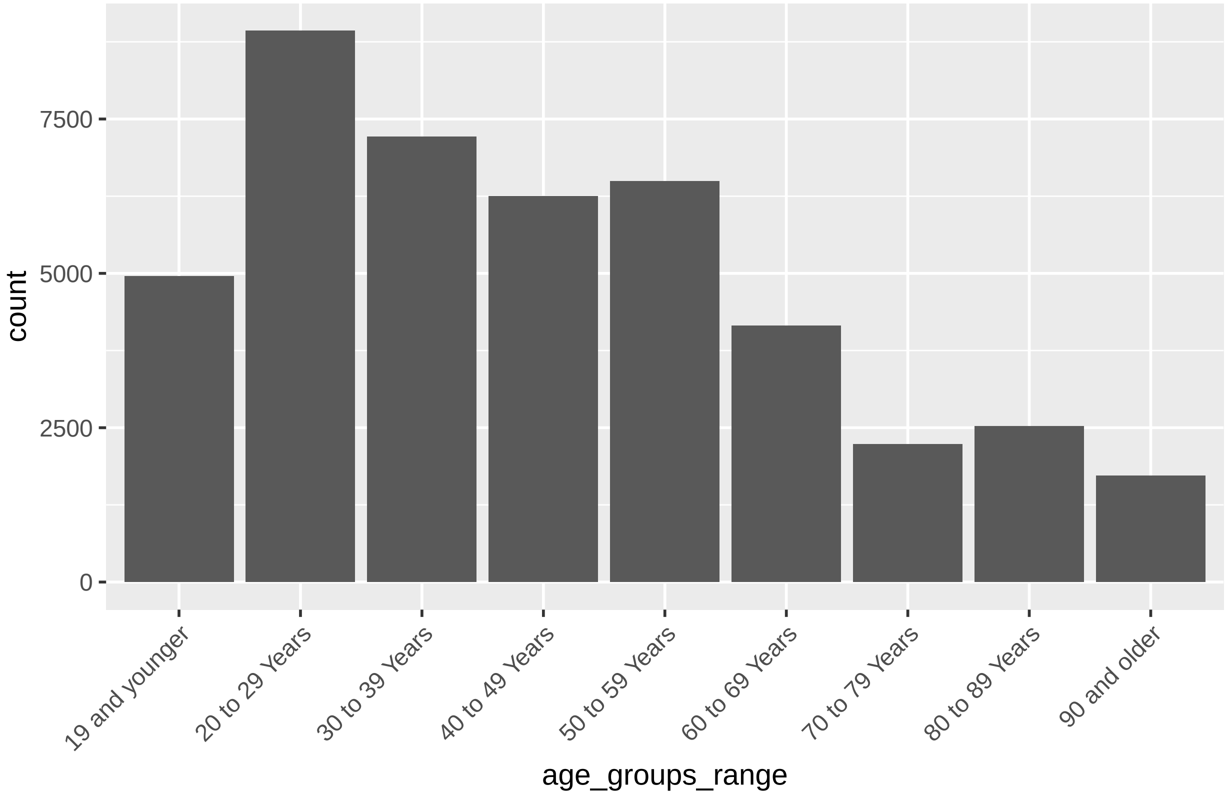
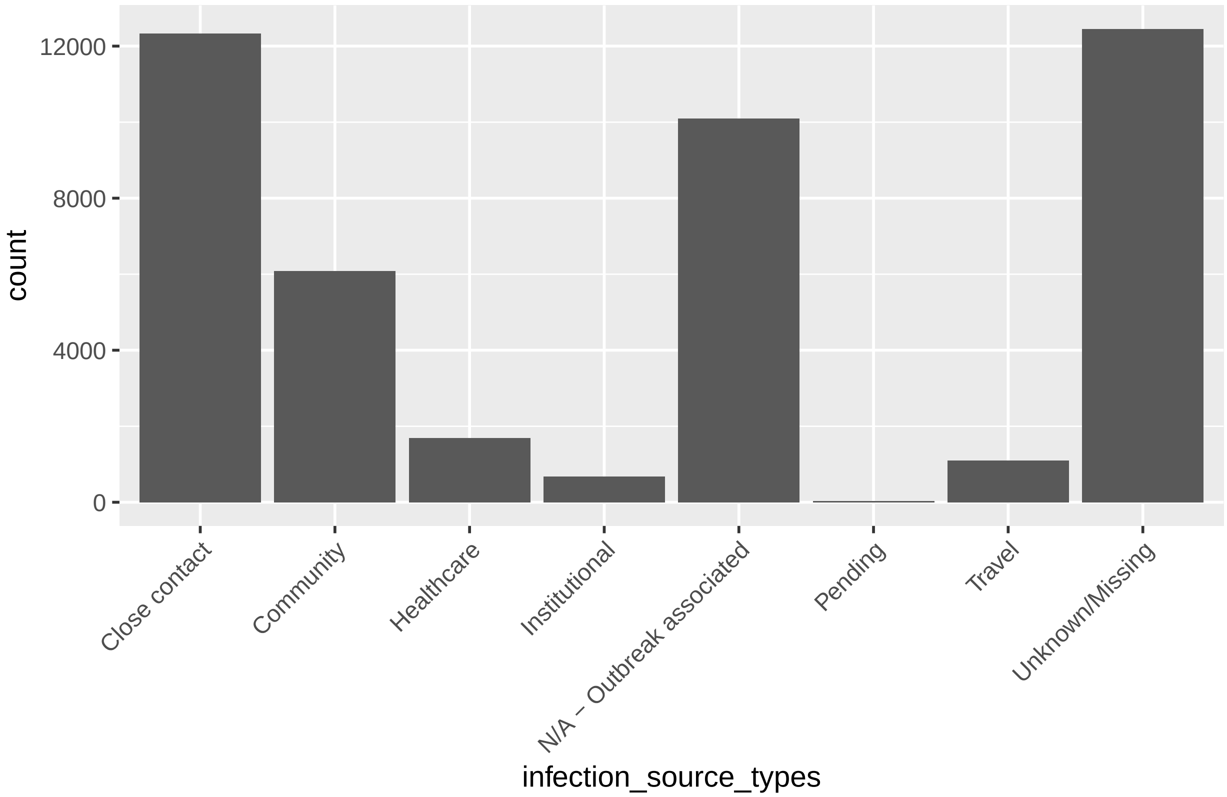


Figure 3: Distribution of Types of Infection Sources



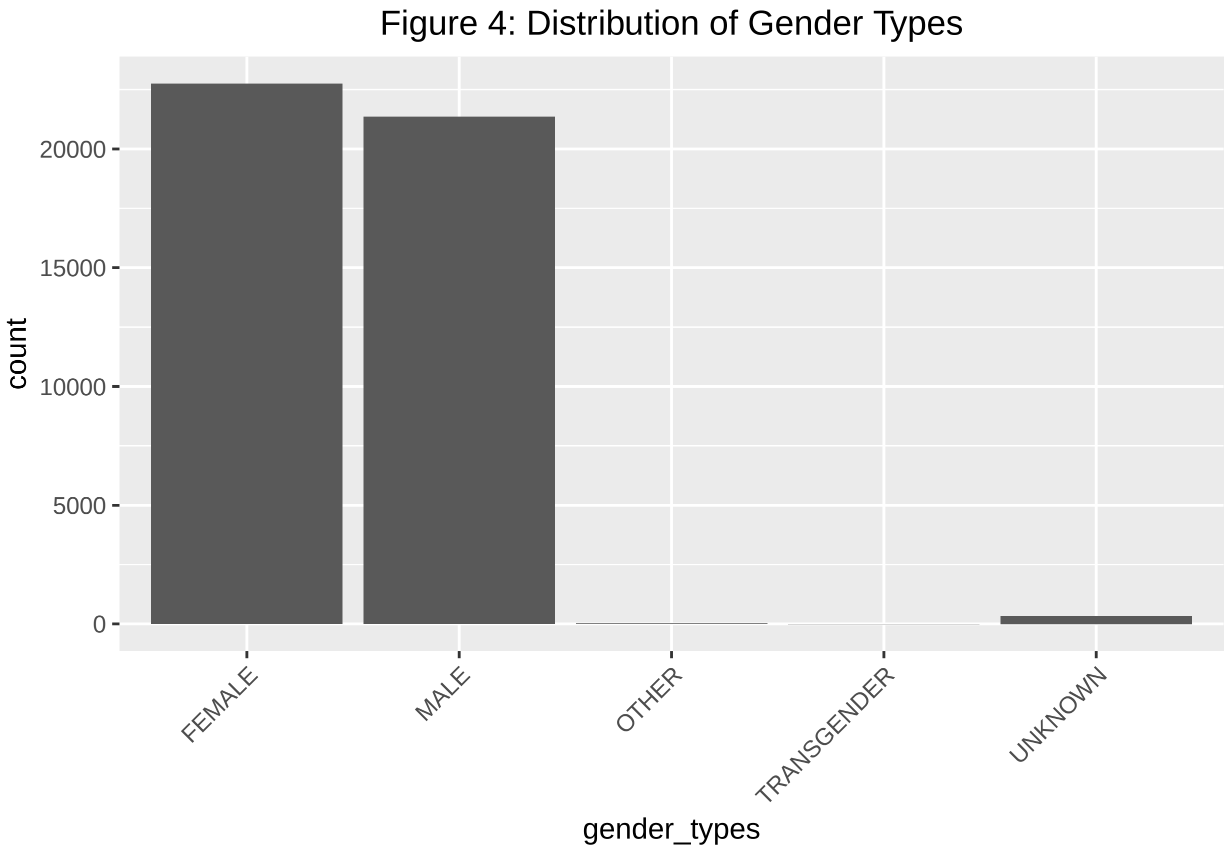


Table 1: Total Count of Fatal and Resolved

	Count
Fatal	1722
Resolved	42738

Model

The probability of a person resulting in fatality outcome is defined as in the below equation. As well, a logistic regression model of the analysis is portrayed below. The functional form of the logistic model outputs the logarithm of the odds of the outcome variable. In this case, the binary variable is `fatality` (f). For each of the categorical variables, dummy variable coding is placed with the variable at the top representing the baseline, in order to be able to assess what effect, if any, moving to a category would have compared to the baseline. A hypothetical case of the model is that if age is higher, the patient may have more underlying health issues that may worsen with COVID-19, therefore fatality rate would be higher.Another hypothesis would be that if source of infection is from close contact, the virus may be more deadly than if the virus was caught in a more open environment such as travel, or more protected environment such as health care.

$$\text{Prob}(f) := \begin{cases} 1, & \text{fatality} \\ 0, & \text{otherwise (resolved)} \end{cases}$$

$$\log\left(\frac{f}{1-f}\right) = \beta_0 + \underbrace{\beta_{1a}x_{1a}+}_{\text{dummy coding for outbreak associated}}$$
$$+\ \underbrace{\beta_{2a}x_{2a} + \beta_{2b}x_{2b} + \beta_{2c}x_{2c} + \beta_{2d}x_{2d} + \beta_{2e}x_{2e} + \beta_{2f}x_{2f} + \beta_{2g}x_{2g} + \beta_{2h}x_{2h}}_{\text{dummy coding for age groups}}$$
$$+\ \underbrace{\beta_{3a}x_{3a} + \beta_{3b}x_{3b} + \beta_{3c}x_{3c} + \beta_{3d}x_{3d} + \beta_{3e}x_{3e} + \beta_{3f}x_{3f} + \beta_{3g}x_{3g}}_{\text{dummy coding for source of infection}}$$
$$+\ \underbrace{\beta_{4a}x_{4a} + \beta_{4b}x_{4b} + \beta_{4c}x_{4c} + \beta_{4d}x_{4d}}_{\text{dummy coding for gender}}$$

Results

In Table 2, the coefficients fitted using the logistic regression model are illustrated. The Variable column contains the categorical independent variables, Estimate column gives the estimate coefficients, and P column provides the p-values. Precisely, the intercept term has no meaningful interpretation but it is statistically significant with a very small p-value. With the estimates, odds ratios can also be calculated by the exponentiation of the estimate values. Beta can then be calculated, as Beta gives the logarithm of the odds ratio and corresponding standard error of the estimate. Coefficients that have p-value of 0.05 or less would be statistically significant.

It was originally hypothesized that higher age may be related to the fatality outcome of COVID-19, or close contact as well. However, among the total result counts, the fatality counts are too few compared to the resolved counts to determine if these variable play a role. Because of this potentially, the p-values were also in a wide range. Some were below 0.05, whereas others were above. Age group of 90 and older indeed had a very small p-value of 1.62e-13, so the association is statistically significant, although the fatality count data may be too small. On the other hand, age group of 20 to 29 had a p-value of 0.74, so the association is not statistically significant.

Regarding the types of outbreak associated, the sporadic type had a tiny p-value of less than 2e-16. Ages 30 to 39 and 40 to 49 had a p-value greater than 0.05, showing statistical significance. Conversely, the remaining higher age groups all fell into the category of having a p-value of less than 0.05. Therefore, the variables did not mean statistically significant. Some source of infection type groups depicted having a statistically significant p-value, whereas others did not, such as institutional. Male gender showed to have a statistically significant p-value, and other, transgender, and unknown did not.

Table 2:

	Variable	Estimate	P
B0	Intercept	8.18	3.43e+16
Outbreak Baseline	Outbreak Associated	-	-
B1a	Sporadic	0.97	<2e-16
Age Baseline	19 and Younger	-	-
B2a	20-29	0.47	0.74
B2b	30-39	0.35	0.81
B2c	40-49	-1.99	0.06
B2d	50-59	-3.68	3e-04
B2e	60-69	-5.18	2.51e-07
B2f	70-79	-6.42	1.57e-10
B2g	80-89	-6.99	3.17e-12
B2h	90 and Older	-7.4	1.62e-13
Infection Baseline	Close Contact	-	-
B3a	Community	0.4	0.009
B3b	Healthcare	-0.96	7.39e-09
B3c	Institutional	-0.29	0.25
B3d	N/A - Outbreak Associated	NA	NA
B3e	Pending	10.73	0.96
B3f	Travel	-0.58	0.01
B3g	Unknown/Missing	0.52	7e-04
Gender Baseline	Female	-	-
B4a	Male	-0.66	<2e-16
B4b	Other	10.56	0.98
B4c	Transgender	6.99	0.99
B4d	Unknown	-0.08	0.75

Discussion

Summary

Often, hypotheses made before an analysis may not align with the actual results. In this analysis, for instance- although the age group 90 and older was associated to have a statistically significant p-value, the fatality count was very small compared to the resolved count, the rate being around four percent of the total counts of the two. However, these assumptions were still meaningful in that it allowed to construct a logistic regression model about COVID-19 fatality rate. There is still probably more research to be performed, as there may have been confounding variables in the data. Although not all of the categorical variables were found to be associated with the fatality rate, some definitely may be more related than others. A plan to make a method to eliminate potential confounding variables may be revisited in the next steps section.

Conclusions

Relating the results to the current global problem of pandemic situation of COVID-19, this means that Toronto is among the various cities that are suffering from the virus, and that it should be taken seriously. Age groups that resulted to be associated with fatality rate should possibly be more careful when leaving the house, and not forget to practice safety measures. The Canadian government and the city of Toronto are also both assisting in ensuring their citizens maintain safe, enforcing set rules for them to follow. Whether it is social distancing, only being engaged in essential activities, or entering into lockdown, Canadians are asked to abide by the laws that are in place to keep each other safe. The logistic regression model in the analysis served a role of depicting any dependent variables that contributed to the outcome variable, which was a more advanced model than a regular linear regression model. A more advanced model that could be applied is multilevel regression for the future.

Weaknesses & Next Steps

There exist some weaknesses in the study, and one of them is actually regarding the obtained data set. Explained in one of the earlier sections of the paper, the data set has some ambiguous variables that are not fully explained whether or not a particular variable is related to COVID-19 or the infected individual. Furthermore, the data may not potentially capture some earlier cases of COVID-19, when the disease was not known. Similarly, the data does not take into account of individuals who might have had COVID-19, but never were tested and diagnosed as they just quarantined and recovered. The study itself could be improved in the case that it may have more exact results with higher number of counts. Also, once the data set updates, additional factors could be added that might play as great variables to work with as well.

In the next steps, as stated in the above paragraph too, it would be interesting to repeat this analysis using a larger updated data set that may come out in the near future. With more numbers to work with, the results should be more accurate. Also, if some of the data that were left out could be clarified in terms of if they are information related to COVID-19 or the patient, these data may actually play as vital contributing factors. In order to update the data set, it could be useful to perform a survey on the COVID-19 patients as well. For instance, a questionnaire that has multiple questions in regards to COVID-19, such as major symptoms that the infected individuals experienced, could be helpful in determining if certain symptoms are more deadly than others. Through more research like this analysis, it is hoped that the spread of COVID-19 will decrease and eventually come to an end.

References

Bibliogrpahy:

1. Government of Canada. (2020). Coronavirus disease (COVID-19): Symptoms and treatment. Retrieved from Canada website: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html

2. Sheather, S. J. (2009). A modern approach to regression with R. Springer. https://books-scholarsportal-info.myaccess.library.utoronto.ca/en/read?id=/ebooks/ebooks0/springer/2010-02-11/1/9780387096087#page=1

3. Van Domelen, D. R. (2019). tab: Create summary tables for statistical reports. rdrr.ido. https://rdrr.io/cran/tab/

4. Wickham, H., Averick M., Bryan J., Chang W., McGowan, L. D., Francois R., Grolemund G., Hayes A., Henry, L., Hester J., Kuhn M., Pedersen T. L., Miller E., Bache, S. M., Muller, K., Ooms J., Robinson, D., Seidel, D. P., Spinu, D.,… Yutani, H. (2019). Welcome to the Tidyverse. The Journal of Open Source Software. https://joss.theoj.org/papers/10.21105/joss.01686

5. Wiley, J. F., & Pace, L. A. (2015). Beginning R: An introduction to statistical programming. Apress. https://books-scholarsportal-info.myaccess.library.utoronto.ca/en/read?id=/ebooks/ebooks3/springer/2017-08-17/1/9781484203736#page=1

6. World Health Organization. (2020). Coronavirus disease 2019 (COVID-19) Situation Report – 94. Retrieved from World Health Organization website: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf?sfvrsn=b8304bf0_2#:~:text=Retrospective%20investigations%20by%20Chinese%20authorities,%2C%20some%20did%20not.

7. World Health Organization. (2020). Naming the coronavirus disease (COVID-19) and the virus that causes it. Retrieved from 2020 WHO website: https://www.who.int/emergencies/diseases/novel-

coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

8. World Health Organization. (2020). Rolling updates on coronavirus disease (COVID-19). Retrieved from 2020 WHO website: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen
9. Wu, C., & Thompson, M. E. (2020). Sampling theory and practice. Springer.
10. Xie, Y. (2020). knitr: A general-purpose package for dynamic report generation in R. rdrr.io. https://rdrr.io/cran/knitr/