Determining Which is More Concerning: COVID-19 Outbreaks, Close-Contact Exposure, or Travel-Related Incidents

Masheil Mir

June 18, 2021

Abstract

COVID-19 has been an ongoing concern worldwide since 2019. This study aims to determine which is more concerning in Canada: COVID-19 Outbreaks, Close-Contact Exposure, or Travel-Related Incidents. The data obtained was filtered down to remove any cases that were not reported, resulting in only considering Ontario data. The variables of interest were determined and consist of the following: age, gender, hot spot health regions in Ontario, case status, and exposure type. To conduct further analysis, the difference in differences methodology was used to compare COVID-19 cases across the exposure types vs. case statuses. After computing the differences, we saw that recovered close-contact cases had the highest counts. Additionally, a multilevel linear regression model was used to determine significant factors for each exposure type. Through these models, we learned that each exposure type had different significant variables that contributed to case counts across hot spot regions in Ontario. These significant factors included recovered cases and the gender of individuals who contracted the virus being male. Through these findings, we were able to conclude that close-contact cases were of most concern, and the Government of Ontario should begin looking into different ways to ensure close-contact exposure is limited, beyond just a provincial lockdown.

Introduction

COVID-19 has affected millions across the globe since March 2019. Canada itself is experiencing its third wave of the virus with new variants arising and increasing the spread. The government's main concern is to ensure the safety of millions of Canadians during this time and containing the virus [1]. Lockdowns, restrictions, and rapid vaccination roll outs have been in full effect in an attempt to slow down the spread. Unfortunately, the newest variant of the COVID-19 virus, the Delta variant, is now known to double the risk of hospitalization [2] when compared to previous variants. This variant is now apparent across some provinces in Canada, such as British Columbia [3] and Alberta [4]. Although previously not seen as a concern, this new research may constitute for a fourth wave, and perhaps even cause lockdown restrictions to become stricter, despite just being lifted in Ontario.

Due to the airborne nature of the the COVID-19 virus (COVID), the emphasis on social distancing and wearing masks has been heightened since the global pandemic began. Thus, when looking at positive COVID cases, it is important to consider how a person was infected, and in return, determine what measures the government must take in order reduce the spread. They have been monitoring travel-related cases, and as a result have ceased all incoming flights from India and Pakistan [5] due to an increase in the number of COVID cases detected in passengers who arrived in Canada from these countries [6]. In addition, province wide lockdowns have come into affect after regional hotspots, such as Peel, York, and Toronto regions in Ontario continue to see higher case counts [7]. Additionally, there have been several workplace closures across Ontario due to outbreaks in local restaurants, warehouses, etc.. Ontario reported its first outbreak-related

closure on April 23rd, 2021, at a McDonald's fast food restaurant, cosmetics store, and car dealership in Toronto [8].

In order to allow readers to better understand the difference between the terminology, the below table has been created:

Table 1: COVID-19 Relevant Terms

Term	Definition
COVID-19	COVID-19 is a disease caused by a new strain of coronavirus. CO stands for corona, V for virus, and D for disease. Formerly, this disease was referred to as '2019 novel coronavirus' or '2019-nCoV.'
Pandemic	A pandemic is a global disease outbreak often caused by a new virus or a strain of virus that has not circulated among people for a long time. Humans usually have little to no immunity against it. The virus spreads quickly from person-to-person worldwide.
Outbreak	An outbreak is a sudden rise in the number of cases of a disease. An outbreak may occur in a community or geographical area, or may affect several countries. It may last for a few days or weeks, or even for several years.
Close-contact Exposure	A close contact is generally someone who has been near a person with COVID-19 for at least 15 minutes when health and safety measures were not in place or were insufficient. This includes up to two days before someone develops symptoms.
Travel-Related Incident	An individual who has symptoms compatible with COVID-19 and traveled to, or from an affected area (including inside of Canada).

It is imperative to consider all factors when looking at the mass spread of the virus. In essence, the government must consider factors such as regions and provinces with high case counts, date it was reported, how the spread occurred, age of those affected, gender, and the status of the respective cases to get a better perspective.

This study will aim to answer the following research question: **What has a greater effect on the number of COVID-19 cases across Canada: travel-related incidents, outbreaks or close-contact exposure?** To explore this question, the age, gender, province, date it was reported, health region, exposure type, and case status of individuals will be analyzed. The data will be filtered down to look at cases in Ontario, British Columbia and Alberta. This is because Ontario currently holds the highest number of COVID cases across the country [9] and the new Delta variant is present in British Columbia and Alberta. Using these provinces will give a better outlook on how the country should proceed as a whole to reduce the spread of the virus.

The hypothesis is that close-contact cases are of greatest concern as it is harder to ensure that individuals across the country will not interact with each other despite lockdown restrictions. However, when looking at travel-related incidents and outbreaks, they seem to be more easily contained by imposing travel restrictions and shutting down certain facilities.

Data

Data Background and Collection

The data was retrieved through a COVID-19 Canada resource hub outlined in the appendix section of this report.

The original data set contains 12 variables, with 11 of them being categorical and 1 being numerical. For this study, we will be focusing on 7 of the 11 categorical variables: health region, age group, gender, date it was reported, exposure type, case status and province.

This data set is a compilation of multiple other data sets that were put together. Since we will be filtering the data set to hone in on the Ontario, British Columbia and Alberta data, we will outline where this particular information was retrieved from.

The Ontario (ON) data was obtained from the Government of Ontario website [14]. This is an open source data set that compiles daily snapshots of publicly reported data testing for COVID-19 in Ontario. The data is provided by Public Health Units across the province, for example, the Durham Region Health Department, York Region Public Health, etc.. It includes age group, gender, date the case was reported, case information, Public Health Unit, and longitudes/latitudes of the region (note that the longitude/latitude is not for the individuals reported on, but rather the central longitude/latitude of the region).

The British Columbia (BC) data was obtained from the BC Centre for Disease control website [15]. It includes information on cases, recoveries, deaths, hospitalizations and testing. There are overlapping variables with the Ontario data, such as age group, cases reported to Public Health, date it was reported, recovery cases and gender.

The Alberta (AB) data was obtained from the Government of Alberta's website [16]. The Government of Alberta obtains this data through the Provincial Surveillance Information system (PSI), which is a laboratory surveillance system that receives positive results for all Notifiable Diseases and diseases under laboratory surveillance from Alberta Precision Labs [16]. The data obtained includes case data, age group, gender, zones/public health units, recovery cases, and date it was reported. These are also variables that overlap with the BC and ON data sets.

Although other province data is not being used, it is important to note that the rest of the data was obtained through the COVID-19 Canada Open Data Working Group [17]. This is an open source data set collected publicly and updated through a spreadsheet including variables such as cases, age, gender, health region, and case information.

CSV's were extracted by the publisher for all of these data sets and then linked together to form this larger data set that is being used in this study.

The drawbacks and limitations of this data set is that it does not delve into several other reasons that could result in individuals contracting the virus. One factor that would be interesting and important to have considered is race. Race has played a large factor in the demographic of those being infected by the virus. According to the Centers for Disease Control and Prevention (CDC), non-Hispanic Black people and Hispanic or Latino people were 4.7 times more likely to have been hospitalized as a result of testing positive for COVID when compared to non-Hispanic white people [18]. Although there is no direct evidence, it is believed that this is because people of colour are more susceptible to having underlying health conditions [18] and in return, COVID targets those have weaker immune systems. There is no data on if any of these individuals have underlying health conditions or what their racial backgrounds are, thus we are only able to make generalized conclusions off of the data provided without granular details.

Another drawback of this data set is that it does not include vaccination data, which plays a large role in whether or not individuals may recover once they have the virus. Although being vaccinated does not guarantee immunity, it does increase the chances of an individual recovering.

Lastly, this dataset does not include case information from the beginning of the COVID-19 pandemic. Although the dates range from March 2020 to June 2021, there were cases prevalent that were not integrated

into the data set. Thus, this may also result in inaccuracies that cannot be accounted for.

Despite these drawbacks, the data still contains important and useful information that will aid in this study, and thus we will continue to use it for further analysis.

Data Cleansing

In order to ensure that the data is clean and free of unwanted variables and values, the following steps were performed:

1. After reading in the data, unnecessary variables such as Object ID, row ID, date reported, latitude, longitude, and province abbreviations were removed from the data set, thus leaving only the following variables: health region, age group, gender, date reported, exposure type, case status, and province.
2. After selecting the desired columns, it was important to ensure that all the variable types were as wanted. The date column was converted from a variable being read into the data as a character into a date.
3. All entries in age group, gender, exposure type, and case status that were "Not Reported" were treated as "NA's" in the data, and thus were removed. This is because this study is dependent on having these variables to determine which exposure type, given age, case status and gender, is of most concern. It is important to note that performing such filtering resulted in all cases from British Columbia and Alberta being removed, leaving only data for Ontario.
4. Since the data only consisted of Ontario cases, the province column was removed as it is now obsolete and does not provide significant information (because we know all of the cases are from Ontario). The new variables of interest are now the following: date reported, health region, age group, gender, exposure, and case status.
5. To make the data more valuable and hone in on regions of interest, the health region column was further filtered down into hotspot regions in Ontario. In this data set they are the following: Toronto Public Health, Peel Public Health, York Region Public Health Services, Durham Region Health Department, Halton Region Health Department, Hamilton Public Health Services, Niagara Region Public Health Department, Simcoe Muskoka District Health Unit, Region of Waterloo Public Health, Wellington-Dufferin-Guelph Public Health, Windsor-Essex County Health Unit and Ottawa Public Health.
6. Finally, since we want to make comparisons between the exposure type of the individuals, three new data frames will be created and sorted by exposure type for later analysis.

Since the data cleansing is complete, we can now move onto giving a complete breakdown of the variables in the data frame alongside important characteristics. This will allow us to see the significance of the variables we are using in the study.

Data Description

Table 2: Variable Descriptions

Variable	Type	Description
Health Region	Categorical	This variable represents the public health region that reported the case. It consists of several health regions, for example Peel Public Health or Toronto Public Health
Age Group	Categorical	This variable outlines the age group of the individual that has tested positive for COVID-19. It ranges from ages >20 and increments in intervals of 10 (20-29, 30-39, etc.) up to ages 80+. For example, if the person who contracted the virus was 87, they would be classified under the 80+ category, whereas if someone who was 34 contracted the virus, they would fall under the 30-39 age category.
Gender	Categorical	This variable represents whether the individual identifies as a male or female. It only has two options, thus the response given is either female or male.
Exposure Type	Categorical	This variable describes how the individual contracted the virus. It consists of three options: outbreak, close-contact, and travel-related. For example, if a person travelled abroad and came back to Canada and tested positive for COVID, their exposure type would be classified as travel-related.
Case Status	Categorical	This variable describes the case status of an individual who contracted the COVID-19 virus. The three options for this variable are: recovered, deceased, and active. For example, if someone had COVID and they have been isolating for 14 days and no longer have symptoms, they are recovered
Date Reported	Categorical	This variable outlines the date that the case was reported to the health region. It consits of dates from January 1st, 2020 to June 2nd, 2021.

Since these are categorical variables, we can assign numerical values to each "group" within each variable and conduct numerical summaries. We can then tie those back to the original data set to make our inferences. In the tables below, the different bins for the categorical variables in the data set have been created and explained for clarity.

The following table represents the numerical values that will be tied to the hotspot health regions in Ontario. The values range from 1 to 12 and will be used to create numerical summaries of the data.

Table 3: Health Region Numerical Representation

Data Set Value	Health Region
1	Durham Region Health Department
2	Halton Region Health Department
3	Hamilton Public Health Services
4	Niagara Region Public Health Department
5	Ottawa Public Health
6	Peel Public Health
7	Region of Waterloo, Public Health
8	Simcoe Muskoka District Health Unit
9	Toronto Public Health
10	Wellington-Dufferin-Guelph Public Health
11	Windsor-Essex County Health Unit
12	York Region Public Health Services

The next table represents the numerical values that will be tied to the age group that individuals who tested positive for COVID belong to, in Ontario. The values range from 1 to 8 and will be used to create numerical summaries of the data.

Table 4: Age Group Numerical Representation

Data Set Value	Age Group
1	<20
2	20-29
3	30-39
4	40-49
5	50-59
6	60-69
7	70-79
8	80+

The next table represents the numerical values that will be tied to the exposure type of individuals who tested positive for COVID in Ontario. The values range from 1 to 3 and will be used to create numerical summaries of the data.

Table 5: Exposure Type Numerical Representation

Data Set Value	Exposure Type
1	Close Contact
2	Outbreak
3	Travel-Related

The next table represents the numerical values that will be tied to the case status of individuals who tested positive for COVID in Ontario. The values range from 1 to 3 and will be used to create numerical summaries of the data.

Table 6: Case Status Numerical Representation

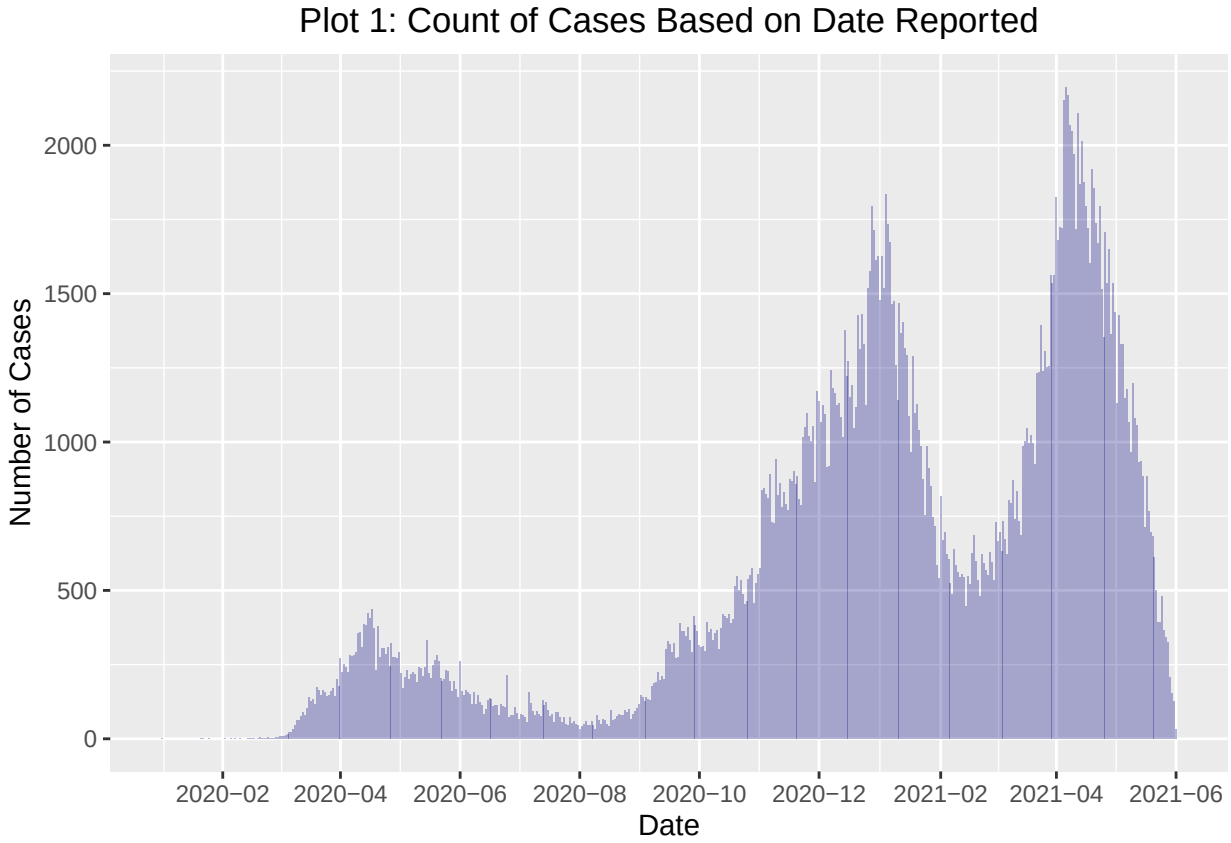
Data Set Value	Case Status
1	Active
2	Deceased
3	Recovered

Now that our variables are identified and explained, we can proceed with creating meaningful summaries. These summaries will provide important insight that will aid on our analysis and model creation.

Numerical Summaries and Plots

Below we have obtained numerical summaries that will include the center and spread of variable. The center and spread includes the following values that will also be calculated as a result: mean, median, mode, standard deviation, interquartile range (IQR), minimum and maximum. It is important to note that not all of the values make sense with our variables as they are categorical, thus only relevant information will be displayed in the plots tables below.

a. Date Reported



When looking at *Plot 1*, the count of cases by date, it is interesting to see that the number of cases are more than double what they were at the beginning of the pandemic (March 2020). It seems that the spike in cases began after September 2020 and reached a high in January 2021, despite the lockdown order being in place since October 7th, 2020 [26]. The dip in the number of cases is seen in February 2021, which was shortly follow by an all time high in case counts in April 2021. This drastic increase in cases resulted in another lockdown that was placed on April 7th, 2021 [26] with stricter restrictions than lockdowns that had been in place previously. Since the lockdown came into effect, we can see a somewhat rapid decline in the

number of cases. From this graph we can safely see the effectiveness of lockdown orders and how they have helped keep case counts down.

b. Health Region

Table 7: Health Region Numerical Summary

Numerical Measure	Health Region
Mean	6.970665
Median	6.000000
Min	1.000000
Max	12.000000
Standard Deviation	3.116897
Interquartile Range	4.000000

The table above shows the categorical results of the health region in which individuals tested positive for COVID-19. We will only be focusing on the mean for this variable since the rest of the values do not provide relevant information for the study. Notice that the mean for health region is **6.9706648**. Since this is not actually representing numerical data, we can round this to 7.00. If we reference *Table 4*, notice that the value 7 represents Region of Waterloo, Public Health. This means that the Region of Waterloo, Public Health has reported more cases than the other regions. The standard deviation value tells us that the data is spread out by **3.116897** standard deviations. The standard deviation is relatively low, thus we can safely conclude that the overall data is close to the mean. The Interquartile Range, **4** represents how spread out the middle 50% of the data points are. Overall, the Region of Waterloo is one that should be of most concern when the government begins to hone in on hotsopt regions in Ontario.

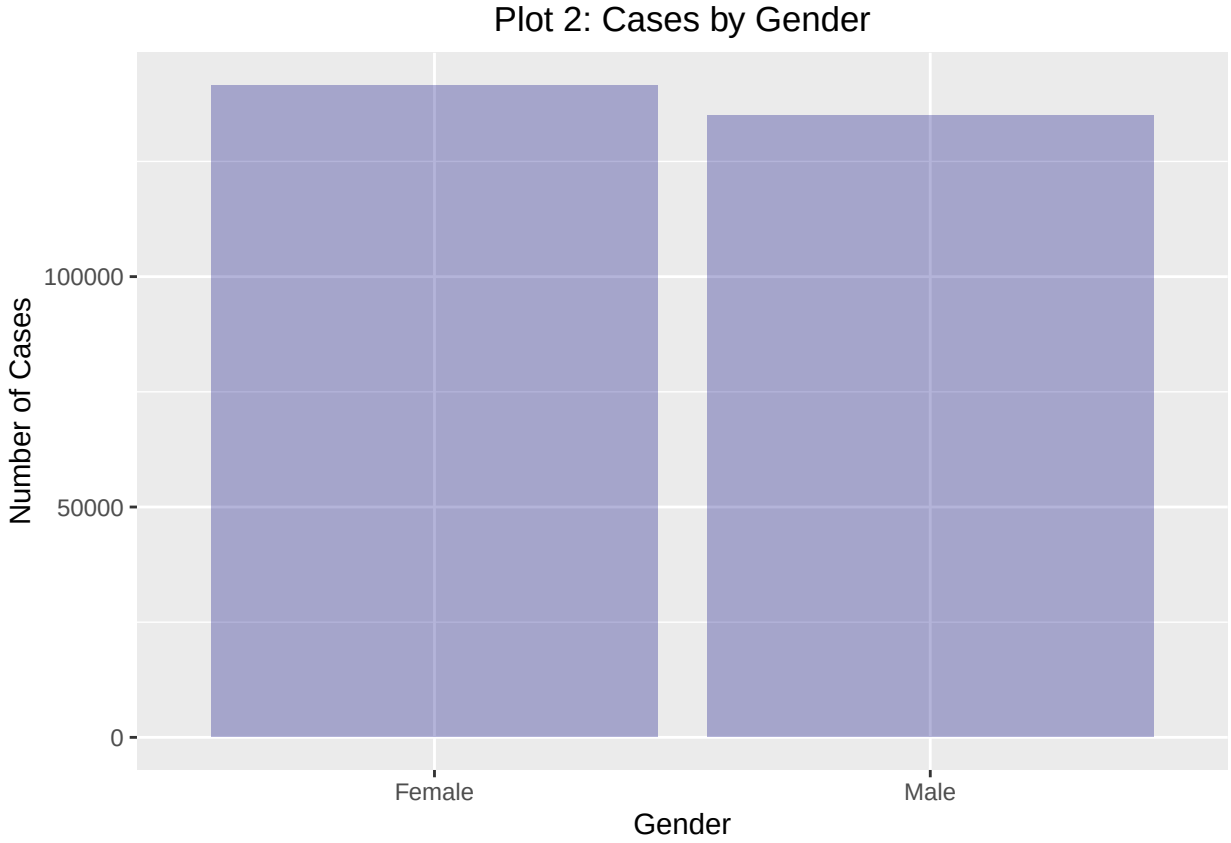
c. Age Group

Table 8: Age Group Numerical Summary

Numerical Measure	Age Group
Mean	3.397218
Median	3.000000
Min	1.000000
Max	8.000000
Standard Deviation	1.995525
Interquartile Range	3.000000

The table above shows the categorical results of the Age Group of individuals who have tested positive for COVID-19. The min and max values represent the minimum and maximum age of the individuals, with **1** representing the age group of 20 years old or younger and **8** representing the age group 80 years old and older, respectively. Looking at the mean value, which is not actually representing numerical data, we see that it is **3.3972184**. We can round this number down to 3.00. If we reference *Table 5*, notice that the value 3 represents individuals between the age of 30-39. This means that the majority of individuals that tested positive were in the age range of approximately 30 to 39 years old. The standard deviation value tells us that the data is spread out by **1.9955246** standard deviations. The standard deviation is relatively low, thus we can say that the overall data is close to the mean. The Interquartile Range, **3** represents how spread out the middle 50% of the data points are. Overall, the age group of most concern can be concluded to be those aged 30-39 years old.

d. Gender



Looking at *Plot 2: Cases by Gender*, we can see the difference between males and females that tested positive for COVID-19. We can see that the female case count seems to be greater than the male case count. That being said, this graph does not provide significant information other than what has already been mentioned and seen from the above, but it can be paired with the rest of our data to possibly provide important insight.

e. Exposure Type

Table 9: Exposure Type Numerical Summary

Numerical Measure	Exposure Type
Mean	1.2794618
Median	1.0000000
Min	1.0000000
Max	3.0000000
Standard Deviation	0.5027489
Interquartile Range	1.0000000

The table above shows the categorical results of the Exposure Type of individuals that have tested positive for COVID-19. Since these numbers are referencing 1 of 3 types, 1 being Close Contact exposure, 2 being exposure by an Outbreak, and 3 being travel related exposure (refer to *Table 6* in the Data Description section), we can see that close-contact cases make up majority of the data. This is derived from the numerical mean which is **1.2794618**. Rounding this down to 1, and referencing *Table 6* gives us the prior conclusion. This aligns with our hypothesis as travel restrictions have been in place for the majority of the pandemic and exposure by outbreaks are not as common since many communities have had appropriate safety measures in place to reduce the chances of a communal outbreaks. On the contrary, it is harder to control a population

from socializing despite lockdowns, and thus this value seems to give significant insight. The standard deviation value tells us that the data is spread out by **0.5027489** standard deviations. The standard deviation is relatively low, thus we can say that the overall data is close to the mean. The Interquartile Range, **1** represents how spread out the middle 50% of the data points are.

f. Case Status

Table 10: Case Status Numerical Summary

Numerical Measure	Case Status
Mean	2.9537691
Median	3.0000000
Min	1.0000000
Max	3.0000000
Standard Deviation	0.2754134
Interquartile Range	0.0000000

The table above shows the categorical results of the case statuses of individuals that have tested positive for COVID-19 Since these numbers are referencing 1 of 3 types, 1 being Active Cases, 2 being Deceased, and 3 Individuals that have recovered, (refer to *Table 7* in the Data Description section), we can see that most of the individuals who tested positive for COVID-19 recovered. This is derived from the numerical mean which is **2.9537691**. Rounding this value up to 3 and referencing *Table 7* gives us the prior conclusion. This value is not surprising since most cases are not usually active for a prolonged period of time and the chances of death are not as common for the average age group we have discussed above (30-39). The standard deviation value tells us that the data is spread out by **0.2754134** standard deviations. The standard deviation is relatively low, thus we can say that the overall data is close to the mean. The Interquartile Range, **0** represents how spread out the middle 50% of the data points are.

These numerical summaries alone only provide some insight on the variables. To further our analysis and delve deeper into the data, we will be conducting Difference in Differences analysis between the three exposure types

Methods

Difference in Differences

The Difference in Differences method will be used to further analyze the data. The number of cases per exposure type cases based on their status will be compared using the following steps:

1. From the data set, we will sum all the cases based on exposure type and case status. For example, all recovered outbreak cases will be summed, recovered travel related cases, recovered close contact cases, etc., until the sums for 3 categories for both variables are satisfied.
2. Next, we will compute the differences for each of these sums. For example, the recovered outbreak sum will be compared to the active outbreak sum to determine the difference between both of these values. This value with tell us which case count is greater and of more concern. This will be done for all 9 sums generated. The 9 sums generated are as follows:

a. Recovered Outbreak Sum
b. Active Outbreak Sum
c. Deceased Outbreak Sum
d. Recovered Travel Related Sum

e. Active Travel Related Sum
f. Deceased Travel Related Sum
g. Recovered Close Contact Sum
h. Active Close Contact Sum
i. Deceased Close Contact Sum

3. Now we can calculate the average overall difference between all of the exposures by case status. This entails taking the differences computed in step 2, grouping them by exposure type, and averaging the differences of each exposure type individually. For example, the Outbreak differences will be averaged independently of the Travel-Related and Close-Contact cases.

4. Then, we will compute the difference within these differences. We will take the difference values calculated in step 2 above and compare them to each other to determine the differences between those differences. For example, we will take the difference value of recovered outbreak cases vs. the active outbreak cases computed in step 2, and subtract that with the difference of the recovered close contact cases vs. active close contact cases also computed in step 2. This will allow us to find the differences between the exposure cases whose status is active, recovered, or deceased to determine which exposure type is prevalent.

5. Next, we can compute the average for the difference in differences to obtain the average difference between cases by exposure type. This, again, entails calculating the differences independent of each other.

6. Finally, we can take the average difference values found and subtract the difference of differences values. This will provide us with the final average difference values we want.

It is imperative to understand that all differences and averages are estimates that are being used to validate the hypothesis presented in the introduction section of this report. We hypothesized that close-contact cases are of greatest concern as it is harder to ensure that individuals across the country will not interact with each other despite lockdown restrictions. We are using case statuses to determine which exposure type is the most relevant in each status, as that will help us see how most exposure type cases affect individuals.

Our parameters of interest include Exposure Type and Case Status as these are the variables we will be using to delve further into the difference in differences analysis. The steps outlined above will ensure that we obtain an appropriate difference in differences value that will allow us to determine which exposure type, related to case statuses, is of most concern.

To find the significant values that affect exposure types, we will revisit the following variables: age group, health region, and gender, on top of exposure type and case status. We will build an ordinary least squares (OLS) model with these dummy variables to aid with determining significance. In specific, we will take the exposure-specific data sets created in the data cleansing section, and build the model. For example, we will run the OLS model on the outbreak data set and determine which factors are significantly affecting this exposure type (this will be repeated for all 3 exposure types). The significant values will come from the estimates generated by the model, and they will further be analyzed and presented in order to reach a conclusion.

That being said, we must ensure that the parallel trends assumption, alongside a few other assumptions are met in order to conduct the difference in differences methodology.

Assumptions

a. Parallel Trends

The parallel trends assumption says that even though treatment and comparison groups may have different outcomes before the treatment begins, the trends should be the same or close to each other before the

treatment begins. It is important to note that parallel trends cannot be proven, but instead they are determined by how much the reader is convinced by it.

The comparison group we are using for the data set is exposure type. Before introducing the exposure type, individuals who test positive for COVID-19 usually have the same symptoms. Some of the symptoms include fevers, chills, cough, difficulty of breathing, etc. [27]. This means that the trends apparent in COVID-19 patients are consistent to each other before the comparison groups are considered. Therefore, we can conclude that the parallel trends assumption is met.

b. Additional Assumptions

A few other assumptions that can be made are the following:

1. Are there factors that have only affected one specific group of people and not another? This may be likely given the exposure type of the cases and where an individual contracted the virus.
2. How likely is it that everybody was infected the same way? This is very unlikely as many people contracted the disease in various ways, and it is not only limited to the exposure groups we have in our data set.
3. How do we know who was vaccinated and who was not? A way to look at this is by creating an assumption. The assumption could be that the cases for those aged 60+ after January 2021 received the vaccine, whereas all age groups before January 2021 did not, since the vaccine was not available to Canadian residents until January 2021.

Since the parallel trends assumption holds and we have outlined a few other assumptions that are being made, we can now create and present the model that will aid us in determining significant factors.

Model

Using the difference in differences method required us to ensure that certain assumptions were met so we could use the appropriate model. Initially, the Ordinary Least Squares (OLS) model seemed like it would be the most appropriate as well as simplest model to use, but due to some of the assumptions made, we decided to use a Multiple Level Regression (MLS) instead. This is because OLS only compares a dependent variable given changes in some other variables, but our variables of interest are not solely dependent on one other variable. For example, exposure type may vary by region, cases may vary by region, age groups may vary by region, case counts may vary by age and exposure type, etc.. Therefore, it makes more sense to use a MLS model as it will consider the effect of multiple variables on our variable of interest.

Below is a breakdown of the MLS model with a multilevel random intercept. We have also provided a detailed overview of the Level 1 and Level 2 models.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{gender} + \beta_2 x_{casestatus}$$

Level 2:

$$\beta_{01} = r_{01} + r_{01}\mathcal{W}_1 + u_{01}$$
$$\beta_{02} = r_{02} + r_{02}\mathcal{W}_2 + u_{02}$$

The terms in the model are the following:

Level 1 Terms:

- Y_{ij} is the health region.

- β_{0j} is the intercept which will be the age group variable.
- β_1 is the slope of the gender variable
- β_2 is the slope of the case status variable

Level 2 Terms:

- β_{01}: Intercept of Exposure Type by Case Status
- β_{02}: Intercept of Exposure Type
- r_{0i} $i \subset [1, 2]$: Intercept of the intercepts of each exposure type
- $\mathcal{W}_i$ $i \subset [1, 2]$: Weight
- u_{0i} $i \subset [1, 2]$: Error term with respect to each exposure type

In this model, the significant factors that affect health regions based on exposure type will be analyzed and discussed in the results and conclusion. Additionally, with the model, parameters of interest, assumptions and differences in differences methodology explained, the results of the methods can now be shown.

Results

Difference in Differences

Using the difference in differences method and the steps outlines above, the treatment groups will be split by the following:

1. Case status based on exposure type
2. Exposure type based on case status

Below are the results of the steps used to calculated the differences using this methodology:

Table 11: Outbreak Cases Based on Status

Status	Count
Recovered	59362
Active	569
Deceased	3121

Table 12: Close-Contact Cases Based on Status

Status	Count
Recovered	201726
Active	3763
Deceased	815

Table 13: Travel-Related Cases Based on Status

Status	Count
Recovered	6981
Active	58
Deceased	65

The above tables give the sums of the cases statuses, grouped by exposure type. By just looking at these tables, we can see that *Table 12: Close Contact Cases Based on Status* seems to have the highest number of cases in total, as well as recovered and active, whereas *Table 11: Outbreak Cases Based on Status* seems to have the highest number of deceased cases. Overall, *Table 13: Travel Relates Cases Based on Status* seems to have the lowest counts of all three tables.

To further analyze the values, we will delve into the difference values.

The difference values for each case status type can be found in the Appendix section of this report. It is important to note that if we have positive values, that means the count for the case status mentioned first is higher than what it is being compared to. **Tables 17-19** in the appendix section include the tables that we will be analyzing below.

To begin, if we refer to *Table 17* in the appendix section, we will see that it shows us that the recovered outbreak cases were greater than the active outbreak cases by **58793** cases and greater than deceased outbreak cases by **56241**, but deceased outbreak cases were greater than active outbreak cases by **-2552** cases. The average difference between the outbreak cases is **37494**. This value tells us that the recovered outbreak cases outweigh the deceased and active outbreak cases.

Next, if we refer to *Table 18* in the appendix section, we will see that it shows us that the recovered close-contact cases were greater than the active close-contact by **197963** cases and were greater than deceased close-contact cases by **200911** cases. Additionally, active close-contact cases cases were greater than active close-contact cases by **2948**. The average difference between the close-contact cases is **133940.6666667**. This value tells us that the recovered close-contact cases outweigh the deceased and active close-contact cases.

Finally, if we refer to *Table 19* in the appendix section, we will see that it shows us that the recovered travel-related cases were greater than the active travel-related by **6923** cases and were greater than deceased travel-related cases by **6916** cases. Additionally, deceased travel-relate cases cases were greater than active travel-related cases by **-7**. The average difference between the travel-related cases is **4610.6666667**. This value tells us that the recovered travel-related cases outweigh the deceased and active travel-relate cases.

Overall, the average difference between the case statuses based on exposure type difference is **58681.7777778**. Since this is a positive value, we can see that the recovered cases still outweigh the other case status types.

Next, we can delve into the difference values of the differences computed above. This will allow us to see which exposure type had the most prominent number of cases in each difference.

The difference values for each exposure type can be found in the Appendix section of this report. It is important to note that if we have positive values, that means the count for the exposure type mentioned first is higher than what it is being compared to. **Tables 20-22** in the appendix section include the tables that we will be analyzing below.

To begin, if we refer to *Table 20* in the appendix section, we will see that it shows us that the active close-contact cases were greater than the close-contact Active cases. The value of **-139170** tells us that recovered outbreak cases were **-139170** less than active close-contact cases. Deceased close contact cases were also greater than recovered outbreak cases, differing by **-144670**. Additionally, deceased close contact cases were also greater than active outbreak cases by **-5500**. The average difference between the close outbreak cases and close contact cases differs by **-96446.6666667**. Since the average difference value is negative, we

can safely assume that close-contact cases are of more concern when compared to outbreak cases, as they outweigh the case counts in all three categories.

Next, we will refer to *Table 21* in the appendix section. This table shows us the outbreak cases vs travel-related cases. Notice that the recovered outbreak cases are greater than the active travel-related cases by a factor of **51870**. The recovered outbreak cases are also greater than the deceased travel-related cases by **49325**. On the other hand, the active outbreak cases are less than the deceased travel-related cases, and they differ by **2955**. This means that the outbreak cases outweigh travel-related cases for the number of recovered cases, but travel-related cases have more deaths in comparison to outbreak cases. The average difference between outbreak cases and travel-related cases is **32883.3333333**. Since this value is positive, we can safely assume that although travel-related cases had more deaths than the outbreak cases, outbreak cases were far greater and had a larger impact.

Lastly, we will compare the close-contact cases with the travel-related cases. If we refer to *Table 22* in the appendix section, we can notice that all the values are positive. In particular, we can conclude that recovered close-contact cases are greater than the active and deceased travel-related cases, and the active close-contact cases are greater than the deceased travel-related cases. These are by factors of **191040**, **193995**, and **2955**, respectively. The average difference between close-contact cases and travel-related cases is **129330**. Since this value is positive and quite large, we can safely assume that close-contact cases are of much more concern than travel-related cases.

Overall, the average difference between the outbreak differences vs. close-contact difference vs travel-related difference is **21922.2222222**. Since this is a positive value, we can conclude that the close-contact cases still outweigh the other exposure types.

Lastly, if we compare the overall differences from both of our findings, we can see that we have difference value of **36759.5555556**. This tells us that recovered cases, and in particular, close-contact cases are of most concern as the value is still positive, and our previous findings confirmed these results.

All analysis for this report was programmed using `R version 4.0.5`.

Models

Since we have 3 exposure types, 3 models were created to determine the significant factors for each exposure type. These factors are outlined below.

Table 14: Outbreak Model Statistics

Factor	Estimate	Pvalue	Significance
as.factor(gender)Male	0.19962	0.0000000237	***
as.factor(case_status)Deceased	-0.38133	0.0956	
as.factor(case_status)Recovered	-0.41930	0.0468	*

The table above models the outbreak cases. We can see that there are two significant factors that affect the health region of outbreak cases, based on our intercept, age group. These factors are the following: individuals who are male and whose case status is recovered. Notice that being male is more significant than case status being recovered. This means that COVID-19 cases across health regions have higher male counts than female counts. Although our data earlier in this report showed us that female case counts were higher overall, when broken down by health region, we can now see that male case counts are of more significance. This is not shocking as we discussed, in the introduction section of this report, how males are more prone to getting sick.

Case status being recovered is also of significance. This is because most COVID-19 cases result in an individual fully recovering after testing positive for the virus. Keeping track of this metric between the

health regions is important as it allows us to see that although cases may spike, the measures being taken are probably working as the government aims to reduce the amount of deceased and active cases overall.

Table 15: Travel-Related Model Statistics

Factor	Estimate	Pvalue	Significance
as.factor(gender)Male	0.2493	0.0237	*
as.factor(case_status)Deceased	-1.4243	0.0822	
as.factor(case_status)Recovered	-0.3552	0.6224	

The table above models the travel-related cases. We can see that there is only one significant factor that affects the health region of travel-related cases, based on our intercept, age group. That factor consists of individuals who are male. This means that travel-related COVID-19 cases across health regions have higher male counts than female counts. This may mean that there are more men who are travelling within and outside of Ontario, whether this be for business trips or leisure. Again, this significance is not shocking as we discussed, in the introduction section of this report, how males are more prone to contracting diseases and getting sick in general.

Table 16: Close-Contact Model Statistics

Factor	Estimate	Pvalue	Significance
as.factor(gender)Male	0.03919	0.030296	*
as.factor(case_status)Deceased	-0.26869	0.066723	
as.factor(case_status)Recovered	-0.23261	0.000956	***

The table above models the close-contact cases. We can see that there are two significant factors that affect the health regions of close-contact cases, based on our intercept, age group. These factors consist of the following: individuals who are male and individuals who have recovered from COVID-19. Notice that the recovered category is more significant than being male. This means that close-contact COVID-19 cases across health regions have higher recovery rates than the other exposure cases. As mentioned earlier, this is because most COVID-19 cases result in an individual fully recovering after testing positive for the virus. Keeping track of this metric between the health regions is important as it allows us to see that although cases may spike, the measures being taken are probably working as the government aims to reduce the amount of deceased and active cases overall.

Individuals who are male and contract the virus are also a significant measure as it align with our earlier hypothesis that males are more prone to getting sick and contracting diseases or viruses. As mentioned earlier, although our data in this report showed us that female case counts were higher overall, when broken down by health region, we can now see that male case counts are of more significance.

Taking the factors of significance from our models into consideration with our difference in differences methodology, we can safely derive a conclusion related to our hypothesis and research question.

Conclusion

Recall that the goal of this study was to determine what exposure type has a greater effect on the number of COVID-19 cases across Canada by looking at the differences between case status counts based off of each exposure type. To delve deeper into the data, the factors that contributed to higher case counts were analyzed and their significance was also determined. We had hypothesized that the close-contact cases would

be of of greatest concern as it is harder to ensure that individuals across the country will not interact with each other despite lockdown restrictions.

To determine which exposure type is of greatest concern based off of the case status, the difference in differences methodology was applied to the data. After completing the difference in difference method, a multilevel logistic regression model was created. This model was used to deduce which factors, such as gender, case status, age, etc., were significant contributors to COVID-19 cases across hot spot regions in Ontario. Recall that the reason multilevel logistic regression (MLS) was used is because we are able to compare the dependent variable based on changes in other variables, and COVID-19 cases based on health regions and exposure types are associated with various other factors and not just one.

The models that were created, as seen in **Table's 23-25**, were then analyzed with their respective summaries. The results found from the MLS model can be compared to our results from the difference in differences method, and we can see that there is some overlap. In particular, we can conclude that our hypothesis was correct, as close-contact cases make up the greatest number of cases overall, and in parallel, make up the highest number of recovered cases. High recovery rates correlate with high contraction rates, which was also shown in our difference in differences models. We saw that recovered close-contact cases are greater when compared to recovered, active, and deceased close-contact, travel-related, and outbreak cases. This in itself already tells us that close-contact cases are very high. All of our difference in difference data showed very high close-contact results, whether that was through comparing case statuses or comparing case differences between exposure types.

Additionally, when we ran our MLS model's, we found that the most significant factor in two out of three of our exposure types was recovered cases and the individuals gender being a male. Recovered cases were significant in the close-contact model which aligns with our findings from the difference in differences method. The significance of males having higher case counts across hot spot regions aligns with our research presented in the introduction of this report, which outlines that males are more susceptible to contracting diseases and viruses as they are more prone to having underlying health conditions.

That being said, there were a few drawbacks that affected the results of this study, including but not limited to the following:

1. Lack Individual Health Data

As mentioned various times throughout this report, underlying health conditions make an individual more prone to contracting COVID-19. Additionally, we found that people of colour are more often affected by these underlying health conditions. Unfortunately, the data we used did not provide a any ethnic related/racial background or whether or not individuals had underlying health conditions. This data would have aided our study since it would provide us more reasoning as to why certain case statuses had more or less case counts than others. It would provide more concrete reasoning for why or how an individual contracted the disease and how it affected their recovery or resulted in their death, and in return would give more accurate results.

2. Lack of Vaccination Data

Additionally, there was no vaccination data present in our data set which could have been the reason for the high recovery cases present. Not knowing if an individual was vaccinated leaves room for errored judgement as it is one of the prime factors currently contributing to keeping individuals out of the intensive care unit's (ICU's) and being hospitalized in general. We had very high recovery cases, but we do not know if these are a result of individuals being able to fight off the virus, or if it is a result of being vaccinated. Vaccine roll out in Ontario began early 2021, which is where we saw spikes and dips in overall case counts. The question then arises is how would the data look difference if the vaccine was not being administered at all? Would this affect the case counts? Have individuals in this data set been vaccinated or is this data independent of individuals who received the vaccine? These are all important questions that could not be answered due to the lack of data present in the data set.

3. Inconsistency of Data Across Provinces

Lastly, after we had cleansed the data, we found that all entries from British Columbia and Alberta had been removed. This was a result of inconsistencies in the data. These inconsistencies made it difficult to do further analysis since a lot of the variables our study depended on did not have the information needed for those provinces. Thus, the study only ended up looking at COVID-19 cases in Ontario, which may have skewed our results as we were not able to hone in on all areas of interest.

In conclusion, with the data that we had, we can safely say that the exposure type of most concern is close-contact exposure. Although the Ontario government continues to attempt decreasing the case counts by introducing new lockdowns, it seems that they are not as effective as counts have reached their peak in late 2020 and early 2021. It is important for the Government of Ontario to begin considering alternative measures, inclusive of the already existing lockdowns, in order to ensure the reduction of cases due to close-contact exposure.

Bibliography

1. Public Health Agency of Canada. (2021, June 15). *Coronavirus disease (COVID-19): Outbreak update.* Canada.ca. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

2. Staff, Reuters. (2021, June 14). Delta variant doubles risk of Covid-19 hospitalization, Scottish study finds. *CTV News.* https://www.ctvnews.ca/health/coronavirus/delta-variant-doubles-risk-of-covid-19-hospitalization-scottish-study-finds-1.5469812

3. Judd, A. (2021, June 15). *COVID Delta variant a 'concern' but not spreading rapidly: B.C. officials.* Global News. https://globalnews.ca/news/7948442/covid-delta-variant-bc-concerns/

4. CBC/Radio Canada. (2021, June 14). *Outbreaks of delta variant spark concern at Calgary hospital | CBC News.* CBCnews. https://www.cbc.ca/news/canada/calgary/foothills-delta-variant-outbreak-calgary-covid-1.6063802

5. (2021, June 2). *Government of Canada / Gouvernement du Canada.* Government of Canada, Canada Border Services Agency. https://www.cbsa-asfc.gc.ca/services/covid/menu-eng.html

6. Turnbull, S., & Dunham, J. (2021, April 22). Canada imposes ban on passenger flights from India, Pakistan for 30 days. *CTV News.* https://www.ctvnews.ca/politics/canada-imposes-ban-on-passenger-flights-from-india-pakistan-for-30-days-1.5397839

7. Katawazi, M. (2021, April 8). Full list of Ontario neighbourhoods where the Covid-19 vaccine will be available to those 18+. *CTV News.* https://toronto.ctvnews.ca/full-list-of-ontario-neighbourhoods-where-the-covid-19-vaccine-will-be- available-to-those-18-1.5379755

8. Freeman, J. (2021, April 26). Toronto announces first workplace closures due to Covid-19 outbreaks of five or more. *CTV News.* https://toronto.ctvnews.ca/toronto-announces-first-workplace-closures-due-to-covid-19-outbreaks-of- five-or-more-1.5402149

9. Public Health Agency of Canada. (2021, May 28). *COVID-19 daily epidemiology update.* Government of Canada. [https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html](https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html

10. *Outbreaks, epidemics and pandemics—what you need to know - APIC.* (2014). APIC. https://apic.org/monthly_alerts/outbreaks-epidemics-and-pandemics-what-you-need-to-know/.

11. *Close contacts.* (2019). Bccdc.ca. [http://www.bccdc.ca/health-info/diseases-conditions/covid-19/self-isolation/close-contacts](http://www.bccdc.ca/health-info/diseases-conditions/covid-19/self-isolation/close-contacts

12. *Case Definition -Coronavirus Disease (COVID-19) A. Probable Case. (n.d.).* Retrieved June 20, 2021, from https://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/2019_case_definition.pdf

13. Bjørn Kallerud. (2019, September 24). *Format captions in kableExtra().* Stack Overflow. https://stackoverflow.com/questions/58087931/format-captions-in-kableextra

14. *Confirmed positive cases of COVID-19 in Ontario - Ontario Data Catalogue.* (2019). Ontario.ca. https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario

15. *BC COVID-19 Data.* (2019). Bccdc.ca. http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data

16. *COVID-19 Alberta statistics.* (2021). Alberta.ca. https://www.alberta.ca/stats/covid-19-alberta-statistics.htm

17. ccodwg. (2021, June 20). *ccodwg/Covid19Canada.* GitHub. https://github.com/ccodwg/Covid19Canada

18. *Why is COVID-19 more severely affecting people of color?* (2020). Mayo Clinic; https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802

19. John, C. (2016, November 18). *Center Plot title in ggplot2.* Stack Overflow. https://stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2

20. cs0815. (2018, October 11). *avoid scientific notation x axis ggplot.* Stack Overflow. https://stackoverflow.com/questions/52758313/avoid-scientific-notation-x-axis-ggplot

21. CDC. (2020, February 11). *Older Adults and COVID-19.* Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html#:~:text=Increased%20Risk%2%20Older%20adults%20are&text=The%20risk%20increases%20for%20people,having%20certain%20underlying%20medical%

22. Sophiah Ho. (2020, April 26). *Ontario COVID-19 data visualizations of confirmed cases.* COVID-19 Dashboards. https://covid19dashboards.com/ontario-confirmed-cases-per-region/

23. Bwire G. M. (2020). Coronavirus: Why Men are More Vulnerable to Covid-19 Than Women?. *SN comprehensive clinical medicine,* 1–3. Advance online publication. https://doi.org/10.1007/s42399-020-00341-w

24. Zhu, H. (n.d.). *Create Awesome LaTeX Table with knitr::kable and kableExtra.* Retrieved June 20, 2021, from https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_pdf.pdf

25. amices. (2020, October 2).* Error: No tidy methods for objects of class glmerMod · Issue #274 · amices/mice.* GitHub. https://github.com/amices/mice/issues/274

26.Nielsen, K. (2020, April 24). *A timeline of COVID-19 in Ontario.* Global News; Global News. https://globalnews.ca/news/6859636/ontario-coronavirus-timeline/

27. Public Health Agency of Canada. (2020). *Coronavirus disease (COVID-19): Symptoms and treatment - Canada.ca.* Canada.ca. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html

Appendix

Below is a glimpse of the data and additional plots and summaries that are used throughout the report but not significant enough to keep in the main sections.

a) Glimpse of Original Data

```
## Rows: 1,384,881
## Columns: 13
## $ ObjectId      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ row_id        <int> 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 5~
## $ date_reported <chr> "2020/03/23 12:00:00+00", "2020/04/02 12:00:00+00", "202~
## $ health_region <chr> "Peel Public Health", "Peel Public Health", "Kingston, F~
## $ age_group     <chr> "20-29", "40-49", "30-39", "<20", "50-59", "60-69", "20-~
## $ gender        <chr> "Female", "Female", "Male", "Female", "Male", "Female", ~
## $ exposure      <chr> "Close Contact", "Close Contact", "Close Contact", "Clos~
## $ case_status   <chr> "Recovered", "Recovered", "Recovered", "Recovered", "Rec~
## $ latitude      <dbl> 43.76161, 43.76161, 44.67675, 44.67675, 43.76161, 44.676~
## $ longitude     <dbl> -79.81357, -79.81357, -76.88425, -76.88425, -79.81357, -~
## $ province      <chr> "Ontario", "Ontario", "Ontario", "Ontario", "Ontario", "~
## $ province_abbr <chr> "ON", "ON", "ON", "ON", "ON", "ON", "ON", "ON", "ON", "O~
## $ hr_uid        <int> 3553, 3553, 3541, 3541, 3553, 3541, 3541, 3553, 3553, 35~
```

This data set can be found using the following link: https://resources-covid19canada.hub.arcgis.com/datasets/covid19canada::compiled-covid-19-case-details-canada/explore

b) Glimpse of Cleansed Data

```
## Rows: 276,460
## Columns: 6
## $ date_reported <date> 2020-03-23, 2020-04-02, 2020-03-30, 2020-03-19, 2020-03~
## $ health_region <chr> "Peel Public Health", "Peel Public Health", "Peel Public~
## $ age_group     <chr> "20-29", "40-49", "50-59", "80+", "50-59", "50-59", "70-~
## $ gender        <chr> "Female", "Female", "Male", "Female", "Male", "Male", "F~
## $ exposure      <chr> "Close Contact", "Close Contact", "Close Contact", "Trav~
## $ case_status   <chr> "Recovered", "Recovered", "Recovered", "Deceased", "Reco~
```

c) Case Differences Between Case Statuses (Grouped by Exposure Type)

Table 17: Outbreak Case Differences Between Case Status

Status	Count
Recovered vs. Active	58793
Recovered vs. Deceased	56241
Active vs. Deceased	-2552

Table 18: Close-Contact Case Differences Between Case Status

Status	Count
Recovered vs. Active	197963
Recovered vs. Deceased	200911
Active vs. Deceased	2948

Table 19: Travel-Related Case Differences Between Case Status

Status	Count
Recovered vs. Active	6923
Recovered vs. Deceased	6916
Active vs. Deceased	-7

d) Case Differences Between Exposure Types

Table 20: Outbreak vs Close-Contact Case Differences

Exposure	Status	Count
Outbreak vs Close-Contact	Recovered vs. Active	-139170
Outbreak vs Close-Contact	Recovered vs. Deceased	-144670
Outbreak vs Close-Contact	Active vs. Deceased	-5500

Table 21: Outbreak vs Travel-Related Case Differences

Exposure	Status	Count
Outbreak vs Travel-Related	Recovered vs. Active	51870
Outbreak vs Travel-Related	Recovered vs. Deceased	49325
Outbreak vs Travel-Related	Active vs. Deceased	-2545

Table 22: Close-Contact vs Travel-Related Case Differences

Exposure	Status	Count
Close-Contact vs Travel-Related	Recovered vs. Active	191040
Close-Contact vs Travel-Related	Recovered vs. Deceased	193995
Close-Contact vs Travel-Related	Active vs. Deceased	2955