Did covid encourage assaults in Toronto neighbourhoods in 2020?*

A statistical analysis of how the rise of Covid in 2020 affected assault rates in various neighborhoods in Toronto

Varun Vijay

27 April 2022

Abstract

This paper presents an analysis on the effect of Covid on assault rates in Toronto neighbourhoods in 2020. The paper uses datasets obtained from opendatatoronto to show how covid has a positive relation with assaults in a neighbourhood. The analysis consists of using the statistical programming language R to model the data using regression and constructing various plots and graphs to visually represent the data. The results obtained will allow us to better predict people's behavior. With the better prediction of people's behaviour, the police can better manage the public order and government will be able to make reasonable and effective policy to reduce the crime rate.

Keywords: Covid, Assaults, Crime,Toronto neighbourhoods, residents, open data toronto

Contents

---

*Code and data are available at: https://github.com/Varun1005473462/final_folder-main.git.

1 Introduction

After the first reported outbreak of Covid in December 2019, the entire world changed. Covid was quickly declared to be a health emergency of global concern by the World health organization in March of 2020. Governments took measures to restrict the spread of the disease such as imposing lockdowns, travel bans, restrictions on public gatherings, closing down businesses and advising people to isolate at home. Citizens of countries were essentially restricted to their homes and rarely left. Covid had a huge impact on the global economy and had restricted people's actions and movements. However, this paper wants to analyse how the rise of Covid has affected one of the most common types of crime, assaults, which are often physical in nature. The question this paper seeks to answer is how the Covid pandemic had affected assault rates in various neighbourhoods in Toronto for the year 2020.

In this paper, we use data on Covid cases and crime statistics in Toronto neighbourhoods for the year 2020 to see how the number of Covid cases in a neighbourhood had affected assaults in 2020 and what factors affected the assault rate the most. We constructed a linear model using R (R Core Team 2020) to model covid's effect on assaults. We also used the dataset to construct various plots and summary statistics about the dataset to improve the quality of our analysis. The results obtained seem to imply that the Covid pandemic had a positive effect on assaults in 2020 as there is a positive relation between number of covid cases in a neighbourhood and number of assaults in Toronto.

The results obtained from this paper help contribute to our understanding of how the presence of a common threat that affects everyone such as a pandemic in this case can discourage people from criminal actions. It will allow us to improve our predictions of a person's behavior during such situations. Better predictions of people's behaviors will allow the police to better manage public order and government to make reasonable and effective policy to reduce the crime rate during such situations.

The paper first presents an overview of the datasets and the variables we will be using for this study in the data section. The model section presents the variables chosen for our model and the justification for these variables. The results section discusses the steps taken to generate estimates for our final model and the various conclusions we can draw from both the data and the model and finally the discussion section discusses the strengths and weaknesses of the paper. Any additional details are included in the appendix. An enhancement is included in appendix C.

2 Data

2.1 Data Source and collection

The relevant datasets were obtained using opendatatoronto and show statistics related to Covid cases and to crime figures. The first dataset is about Covid cases in Toronto and records all Covid cases reported since January 2020. The dataset includes all cases related to outbreaks and sporadic cases, those that occur in the community. The data was collected by opendatatoronto from the provincial Case & Contact Management System and the variables in the dataset include the neighbourhood the case was reported in, the date it was reported, the client gender, Age group and the classification of the case, whether it was probable or improbable. The full dataset @ref(table: dataset1) can be found in Appendix B.

The second dataset used for this paper was the dataset related to crime statistics in Toronto neighbourhoods. The dataset contains a list of crime related statistics by neighbourhood and includes figures such as number of assaults, number of car thefts, number of shootings, etc. for the years 2014 to 2020. We are only interested in the number of assaults for 2020. The full dataset @ref(table: dataset2) can be found in Appendix B.

The datasets were extracted using R (R Core Team 2020) and R packages opendatatoronto (Gelfand 2020), tidyverse (Wickham et al. 2019a), dplyr (Wickham et al. 2021) and knitr (Xie 2021a). The data was then processed, manipulated and analysed for the purposes of this paper.
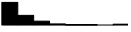
2.2 Data Modification
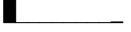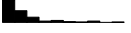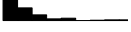
For the purposes of this paper, we are interested only in covid cases that were reported during 2020 since that was when the pandemic first started. So the only cases selected were those reported during 2020. We then focused on the variables Client Gender and the age group of the infected. These variables are the ones that pertain the most to the total number of covid cases in a neighbourhood and so are the ones we will be focusing on for this dataset. New variables were created such as number of female covid cases, number of male covid cases, number of transgender covid cases, number of other genders and unknown genders covid cases and variables related to the various age groups, less than 19, between 20 to 29, 30 to 39, etc. We also created a new variable where we counted the total number of Covid cases in a neighbourhood. However, we will be excluding all age groups greater than or equal to 50 since it is unlikely that anyone of that age will be assaulting people as they are mostly middle aged or old people. All NA cases were also deleted in order to improve the analysis. All of these variables were placed together in a dataset for analysis. This dataset is the covid dataset we will be using.

For the crime dataset, we only focused on the neighbourhoods and the number of assaults for the year 2020 since these pertain directly to my paper. We then joined the crime and covid datasets together and obtained the final_dataset. All 140 neighbourhoods are included in this dataset and it is the dataset we will be using for the rest of this paper.

The code for the dataset modifications are included in scripts in final-data-prep.R.

2.3 Dataset for this paper

```
## Rows: 140 Columns: 12
## -- Column specification -------------------------------------------------
## Delimiter: ","
## chr  (1): Neighbourhood
## dbl (11): Assault_2020, FEMALE_Covid_CASES, MALE_Covid_CASES, transgender_co...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```
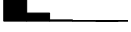
	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Assault_2020	107	0	127.5	127.5	12.0	87.5	761.0	
FEMALE_Covid_CASES	105	0	112.3	104.1	10.0	78.0	502.0	
MALE_Covid_CASES	97	0	112.3	102.6	7.0	75.5	564.0	
transgender_covid_cases	2	0	0.0	0.1	0.0	0.0	1.0	
unknown_covid_cases	6	0	0.8	1.1	0.0	0.0	5.0	
other_covid_cases	2	0	0.0	0.2	0.0	0.0	1.0	
<=19	65	0	32.3	35.7	2.0	20.5	198.0	
20 to 29	75	0	44.7	46.2	0.0	30.0	260.0	
30 to 39	69	0	37.5	32.2	1.0	28.0	170.0	
40 to 49	65	0	32.4	32.0	1.0	21.5	165.0	
Total_CASES	119	0	225.4	206.3	17.0	156.5	1054.0	

```
## # A tibble: 140 x 12
##    Neighbourhood Assault_2020 FEMALE_Covid_CA~ MALE_Covid_CASES transgender_cov~
##    <chr>                <dbl>            <dbl>            <dbl>            <dbl>
##  1 Agincourt No~           69              157              141                0
##  2 Agincourt So~          132              140              155                0
##  3 Alderwood               37               43               50                0
##  4 Annex                  275               96              162                0
##  5 Banbury-Don ~           83              126               88                0
##  6 Bathurst Man~           58              117              122                0
##  7 Bay Street C~          516               48               55                0
##  8 Bayview Vill~           69               54               54                0
##  9 Bayview Wood~           33               39               47                0
## 10 Bedford Park~           51              115              133                0
## # ... with 130 more rows, and 7 more variables: unknown_covid_cases <dbl>,
## #   other_covid_cases <dbl>, '<=19' <dbl>, '20 to 29' <dbl>, '30 to 39' <dbl>,
## #   '40 to 49' <dbl>, Total_CASES <dbl>
```

The final dataset @ref(table: dataset_final) consists of 140 observations and 12 variables: Neighbourhood: Name of the neighbourhood, Assault_2020: number of assaults in that neighbourhood for the year 2020, FEMALE_Covid_CASES: Number of reported covid cases in that neighbourhood for females, MALE_Covid_CASES: Number of reported covid cases in that neighbourhood for males, transgender_Covid_CASES: Number of reported covid cases in that neighbourhood for transgenders, unknown_covid_cases: Number of reported covid cases in that neighbourhood for people of unknown genders, other_covid_cases: Number of reported covid cases in that neighbourhood for people of other genders, <=19: number of reported covid cases for people less than 19 years old, 20 to 29: number of reported covid cases for people between 20 and 29, 30 to 39: number of reported covid cases for people between 30 and 39, 40 to 49: number of reported covid cases for people between 40 and 49, Total_CASES: total number of covid cases in a neighbourhood.

These 12 variables are the ones that are most pertinent to the topic of this paper. And in this dataset, the number of assaults in a neighbourhood will the response variable for this paper.

2.4   Summary statistics

The above table @ref(table: summary 1) shows summary statistics for the dataset. As you can see, on average, the number of assaults per neighbourhood in Toronto was 127.5 for 2020 with a standard deviation

	Assault_2020	FEMALE_Covid_CASES	MALE_Covid_CASES	transgender_covid_c
Assault_2020	1	.	.	
FEMALE_Covid_CASES	0.32	1	.	
MALE_Covid_CASES	0.38	0.98	1	
transgender_covid_cases	0.03	0.16	0.19	
unknown_covid_cases	0.31	0.62	0.62	—
other_covid_cases	0.01	0.11	0.10	
<=19	0.20	0.88	0.88	
20 to 29	0.47	0.91	0.95	
30 to 39	0.46	0.93	0.96	
40 to 49	0.32	0.96	0.96	
Total_CASES	0.35	0.99	0.99	

of 127.5 which means that there was a variation in assaults per neighboruhood. Furthermore, the number of covid cases on average in a neighbourhood is 225.4 with a standard deviation of 206.3 so each neighbourhood has a varying amount of covid cases. So it is possible there is a relation between number of assault cases in a neighbourhood and number of covid cases.

2.5 Correlation statistics for the final dataset

From the above table @ref(table: summary 2) we can see that the variables with the strongest correlation with assaults in 2020 are number of female covid cases, number of male covid cases, number of covid cases for age groups 20 to 29, 30 to 39 and 40 to 49. It makes sense for these variables to relate to number of assaults in a neighbourhood since they are the most important ones since age and gender are incredibly important when building a model for crime and assaults. We also see that age groups and the number of male and female covid cases are strongly related.

2.6 Assault cases by neighbourhood

Fig @ref(fig: figure 1) shows us the number of assaults in 2020 per neighbourhood. There does seem to be some variation in the number of assaults per neighbourhood with some high outliers. The plot obtained matches the results shown in the summary table.

2.7 Covid cases by neighbourhood

As we can see from figure @ref(fig: figure 2), the number of covid cases per neighbourhood varies and are not similar to each other. Additionally, the variation pattern in the number of covid cases seems to be similar to the pattern in the number of assaults per neighbourhood from figure @ref(fig: figure 1.) So, the plots support our hypothesis that there is a relation between covid cases and number of assaults in a neighbourhood.

3 Model

The model we will be using for this paper is:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \beta_5 * x_5$$

y= number of assaults in a neighbourhood, x_1 = number of female covid cases in a neighbourhood, x_2 = number of male covid cases in a neighbourhood, x_3 = number of covid cases for people between ages 20
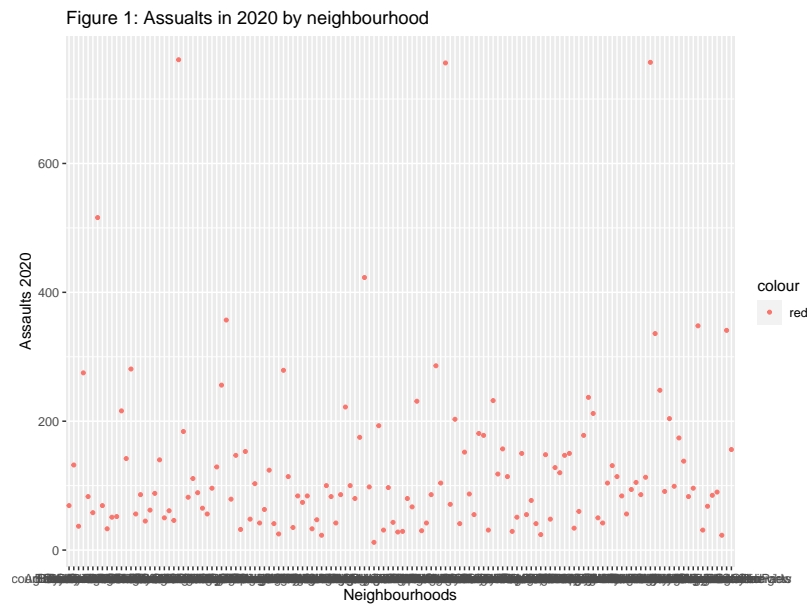
Figure 1: Assualts in 2020 by neighbourhood
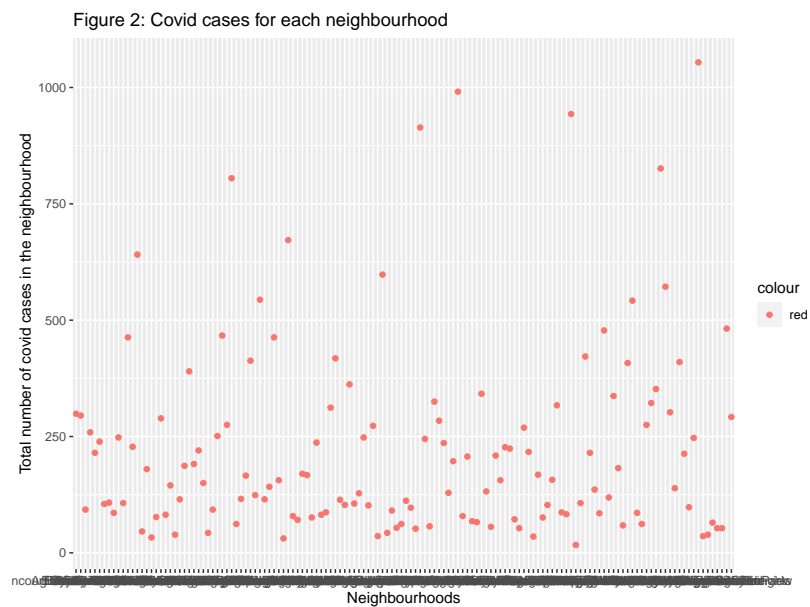
(#fig:figure 1)

Figure 2: Covid cases for each neighborhood in 2020

(#fig:figure 2)

and 29, x_4 = number of covid cases for people between ages 30 and 39, x_5 = number of covid cases for people between ages 40 and 49.

Here, the number of assaults is the response variable and the number of female covid cases, number of male cases, number of covid cases for age group 20 to 29, number of covid cases for age group 30 to 39 and number for 40 to 49 are the explanatory variables.

The reason I have chosen these 5 variables is because from the correlation table @ref(table: summary 2) in the data section, these variables have some of the strongest correlation with the number of assaults in 2020.

Gender is one of the most important factors in today's world and is an important variable to include. It is always important to take gender into account as it can often be a confounding variable if you do not include it in the model. I have excluded the transgender and unknown variable from the model because those entries were far less than the number of entries for male and female. Since the male and female cases compose the majority of the dataset it is likely that they will have a higher effect on the response variable so we will be using these variables for the model.

People older than 50 would not normally be commiting assaults so we will exclude them from the dataset and also exclude people younger than 20 since they would be teenagers and children. So we will only be focusing on the age groups in the range of 20 to 49 for this model.

4   Results

Using modelsummary (Arel-Bundock et al. 2022) and AIC selection process, we were able to obtain the final model and the estimates for the model such as intercept, R^2, estimates for independent variables etc which are included in the table **??**table: model estimates).

The AIC process and the code for the estimates are located in tables **??**eq: model). By the process of forwards AIC, we were able to obtain the final model by removing variables from the model to see if it reduced the AIC value. Once we obtained the lowest AIC value possible and could not add more variables, we had obtained my final model.

4.1   Model estimates

```
##                           Terms     values
## 1                     Intercept     64.112
## 2   Female_covid_cases estimate     -0.918
## 3     Male_covid_cases estimate     -0.343
## 4             20 to 29 estimate      2.060
## 5             30 to 39 estimate      4.370
## 6             40 to 49 estimate     -1.573
## 7        Number of observation    140.000
## 8                           R^2      0.397
## 9                  R^2 adjusted      0.375
## 10                          AIC   1696.900
## 11                          BIC   1717.500
## 12                      Log.Lik   -841.440
## 13                            F     17.648
```

4.2   Final model

So, using the estimates obtained, the model we have is:

$$y = 64.112 - 0.918 * x_1 + -0.343 * x_2 + 2.060 * x_3 + 4.370 * x_4 - 1.573 * x_5$$

The estimates for our model show us that there seems to be a positive relationship between the number of covid cases in a neighbourhood and the number of assaults. When there is an increase in the number of male or female cases in a neighbourhood, then there is a small decrease in the number of assaults. However, if the age group affect was 20 to 29 or 30 to 39, the positive increase in the number of assaults is greater than the decrease due to x1 and x2. Only when the age group is of 40 to 49 does the assaults actually decrease but since two of the age groups have a positive effect on number of assaults, we can say that the number of cases of covid in a neighbourhood will have a positive effect on the number of assaults in that neighbourhood.

So when number of covid cases in a neighbourhood increase, so do the number of assaults. From our analysis of the dataset and the model, this means that in neighbourhoods which had higher covid cases in 2020 also had a higher number of assaults for that year.

4.3   Final model graphs

We will use graphs to confirm our results by modelling covid's effect on assaults in a neighbourhood.
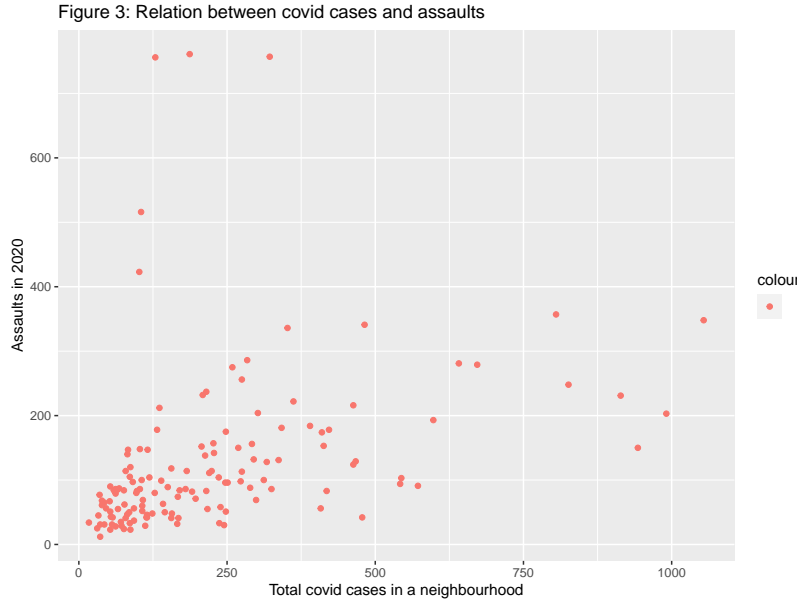
Figure 3: Relationship between Covid and Assault rates in 2020

(#fig:figure 3)

From figure 3 @ref(fig:figure 3), we can see that there appears to be a sort of positive relation between the total number of cases in a neighbourhood and the number of assaults in that neighbourhood for the year 2020. Except for a few outliers, it seems that when cases are low, assaults are low and when cases increase, assaults also increase.

Figure 4: Relation between number of assaults in 2020 and the number of female covid cases
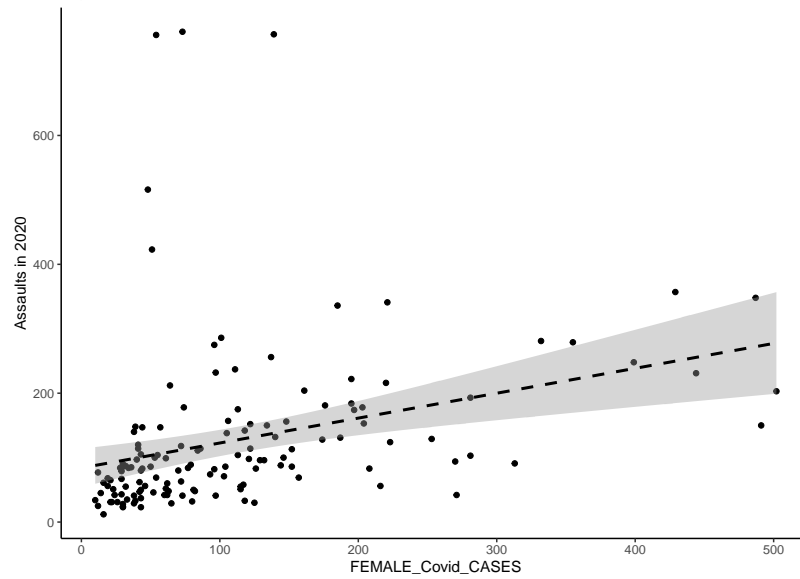
Figure 4: Relation between number of assaults in 2020 and the number of female covid cases
(#fig:figure 4)

Figure 5: Relation between number of assaults in 2020 and number of male covid cases
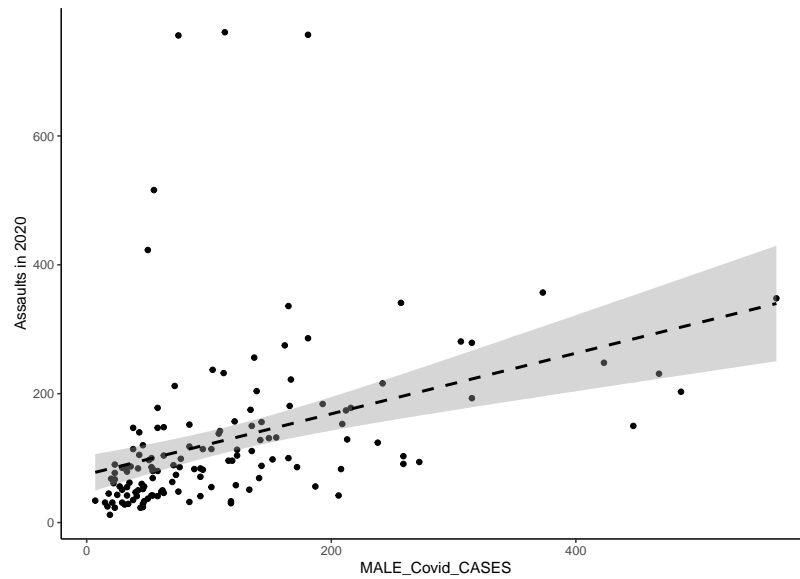
Figure 6: Relation between number of assaults in 2020 and number of covid cases for age gro
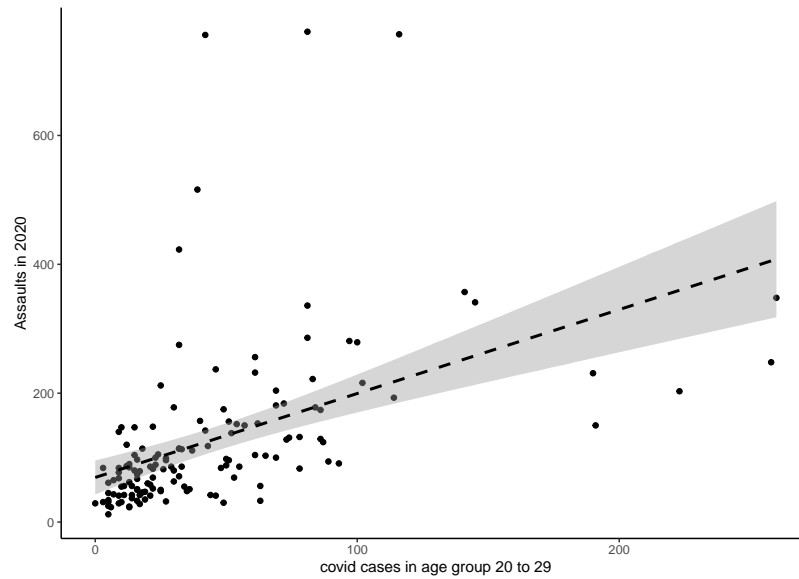
Figure 7: Relation between number of assaults in 2020 and number of covid cases for age gro
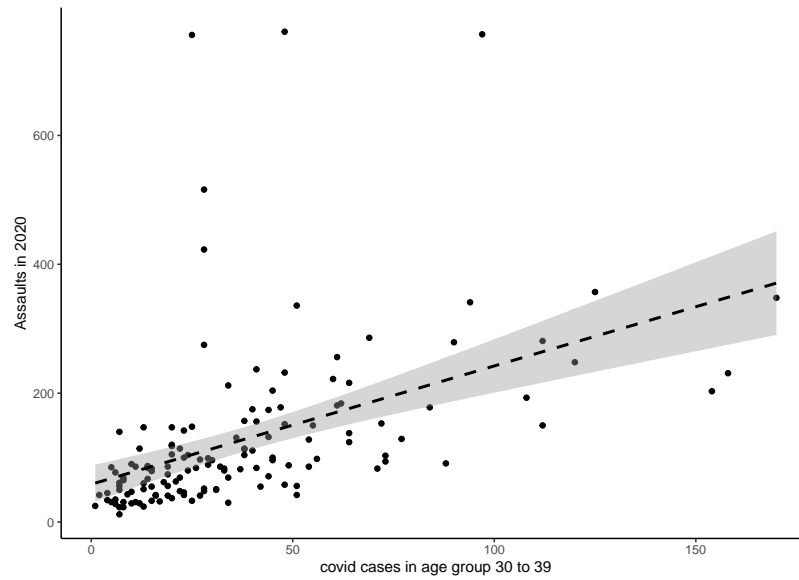
Figure 8: Relation between number of assaults in 2020 and number of covid cases for age gro
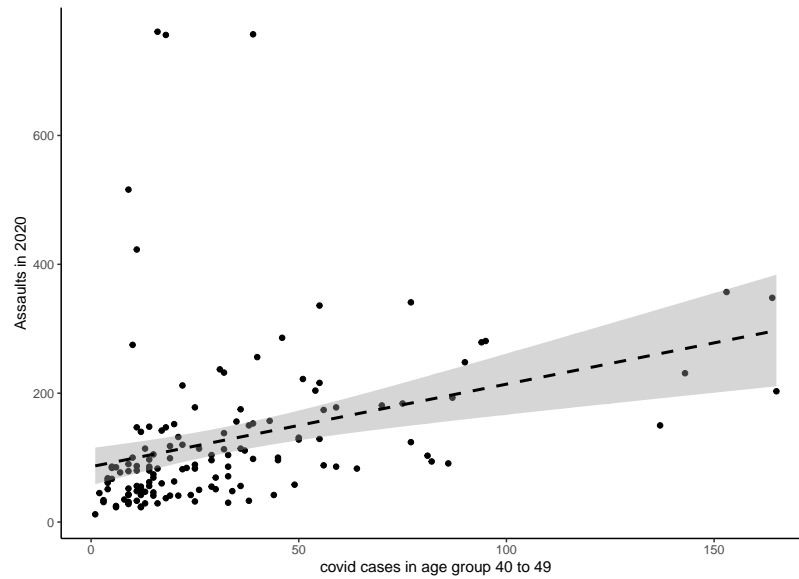
Figure 9: Relation between number of assaults in 2020 and the final model obtained
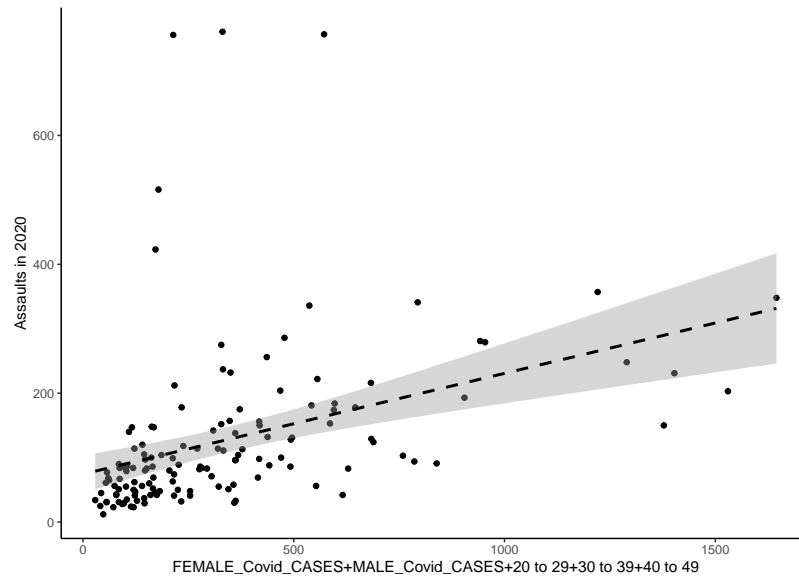
Figure 4 **??**fig:figure 4),5 **??**fig:figure 5),6 **??**fig:figure 6),7 **??**fig:figure 7),8 **??**fig:figure 8) and 9 **??**fig:figure 9) all show that there seems to be a positive relation between the number of covid cases in a neighbourhood and the number of assaults in that neighbourhood except for a few outliers. When the number of covid cases in a neighbourhood are low so are the number of assaults and when the number of covid cases increase so do the number of assaults increase with it.

We can analyse the goodness of fit for the model from the above 4 plots in figure 10 **??**fig:figure 10). From the residuals vs fitted plot, we see that linearity for the model does hold and that the independent variables do enter in a linear way. However, from the spread of the residuals we see that variance of our residuals is not constant which means that the model has heteroscedasticity. From the Normal Q-Q plot, we see that our data is normally distributed since the points mostly lie on the dotted line except for a few outliers.
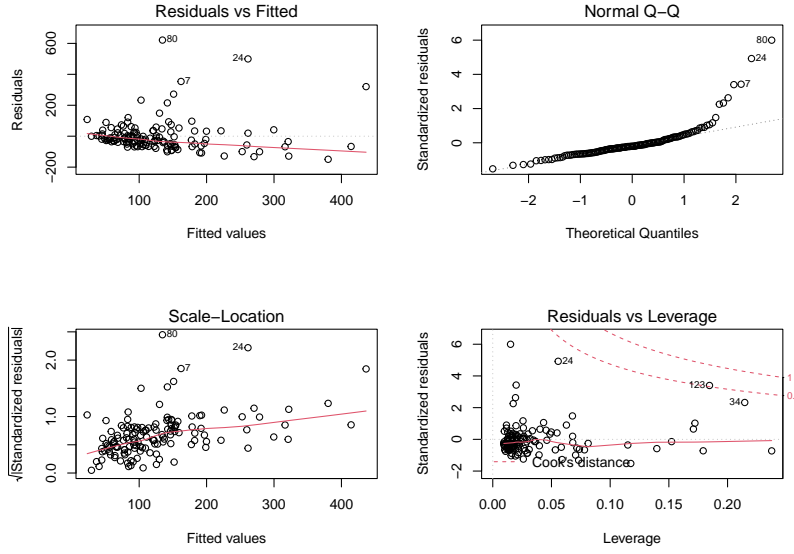
Figure 5: Figure 9: Goodness of fit plots

(#fig:figure 10)

4.4 Conclusion of results

The results obtained from our model and analysis of the dataset seem to imply that the number of covid cases in a neighbourhood had a positive relationship with the number of assaults in that same neighbourhood in the year 2020. The final model we obtained is:

$$y = 64.112 - 0.918 * x_1 + -0.343 * x_2 + 2.060 * x_3 + 4.370 * x_4 - 1.573 * x_5$$

where y= number of assaults in a neighbourhood x_1 = number of female covid cases in a neighbourhood x_2 = number of male covid cases in a neighbourhood x_3= number of covid cases between ages 20 and 29 x_4 = number of covid cases between ages 30 and 39 x_5 = number of covid cases between ages 40 and 49

The number of assaults were higher in neighbourhoods with higher covid cases and were lower in neighbourhoods with lower number of covid cases.

5 Discussion

5.1 First discussion point

The paper analysed datasets about neighbourhood crime statistics and covid cases reported since January 2020 obtained using opendatatoronto in order to see how number of covid cases in a neighbourhood affect the number of assaults. Here the number of assaults in a neighbourhood was the response variable. The paper focused only on the data during 2020 since that was the year covid was first reported and we were interested in it's initial effects on crime. A linear regression model with several explanatory variables was built using the data obtained and several graphs were created based on this model. Using AIC we were able to build a model and obtain estimates for the various variables.

We found that the variables that affected the response variable were the number of females covid cases in a neighbourhood, number of male covid cases, number of covid cases for age group 20 to 29 and age group

30 to 39. We also found that the number of covid cases in a neighbourhood had a positive relation with the number of assaults in a neighbourhood. Neighbourhoods with higher number of covid cases also had a higher number of assaults. Various graphs plotted using the data confirmed these results for us as they showed a positive and linear relation between number of covid cases in a neighbourhood and number of assaults.

The results obtained supports the theory that pandemics such as covid increase the number of assaults in an area that has been severely afflicted in such situations. Desperation leads to panic which might cause people to inflict violence and assault others. It shows us that when people are placed in situations of high stress, they might panic and resort to violence especially when the stress start to increase and people around them fall ill to some affliction. Our model also showed us that people who were younger and were young adults were more susceptible to panic than older people

The results obtained in this paper could be used to inform police actions in the future at the beginning of situations like Covid where the entire city is impacted. Public order could be maintained by having police increase policing in neighbourhoods that are suffering the most to help protect citizens or have governments make better policy decisions during such situations so that people do not feel stressed and react violently. As an example, they could ensure better care for those in neighbourhoods that are suffering the most from the situation and could focus on younger age groups.

5.2   Weaknesses

There are a few weaknesses in this paper, the datasets used did not include factors such as race of the person with covid which could become a limitation as race might be a confounding factor. However, it might be unethical to try and have race in the model.

Another limitation for the paper is that we were not able to obtain economic data for Toronto neighbourhoods for the year 2020. Economic data such as income of households or families is one of the most important variables to account for when analysing crime since income of a person often dictates whether they would turn to crime or not. Since, we were not able to obtain economic data for the year 2020, it is possible that our paper will be biased in its analysis.

5.3   Next steps

The next steps to improve the model obtained here is to increase the data used for this model. Economic data for Toronto neighbourhoods should be obtained for the year 2020 in order to include these variables in the model. If we are able to obtain such data and include it in the future, we will be able to avoid the effect of confounding variables and will be able to obtain an unbiased model.

Another limitation that could be accounted for is racial data but profiling such data might not be ethical as we will be including race in a model on crime which will lead to racial profiling. This paper recommends that race not be included in such models since it is unethical to model it.

So, in order to fully comprehend people's behaviours in such situation more data is required in order to obtain an unbiased model that will help contribute to out understanding of people's behaviours during high stress situations.

	Model 1
(Intercept)	127.457
	(10.773)
Num.Obs.	140
R2	0.000
R2 Adj.	0.000
AIC	1757.7
BIC	1763.6
Log.Lik.	−876.854
F	

	Model 1
(Intercept)	84.059
	(15.122)
FEMALE_Covid_CASES	0.386
	(0.099)
Num.Obs.	140
R2	0.100
R2 Adj.	0.093
AIC	1745.0
BIC	1753.8
Log.Lik.	−869.512
F	15.262

Appendix

A   Appendix A

A.1   AIC selection

For this paper, AIC selection using modelsummary was used to obtain the final model. Several models were constructed and then using modelsummary, the model with the lowest AIC value was selected.

```r
## the first model constructed using lm with no variables
first_model <-
  lm(Assault_2020 ~1,
     data = final_dataset)
modelsummary(first_model)
```

The AIC obtained from this model is 1757.7, so we will add a variable to see if we can reduce it.

```r
second_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES,data = final_dataset)
modelsummary(second_model)
```

The AIC obtained from this model is 1745.0 which is less so we can keep the female variable.

	Model 1
(Intercept)	76.098
	(14.316)
FEMALE_Covid_CASES	−1.521
	(0.442)
MALE_Covid_CASES	1.979
	(0.449)
Num.Obs.	140
R2	0.212
R2 Adj.	0.200
AIC	1728.4
BIC	1740.2
Log.Lik.	−860.215
F	18.381

	Model 1
(Intercept)	75.591
	(14.333)
FEMALE_Covid_CASES	−1.565
	(0.445)
MALE_Covid_CASES	2.040
	(0.454)
transgender_covid_cases	−63.524
	(68.180)
Num.Obs.	140
R2	0.217
R2 Adj.	0.199
AIC	1729.5
BIC	1744.2
Log.Lik.	−859.770
F	12.532

```
third_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES,data = final_dataset)
modelsummary(third_model)
```

AIC obtained is 1728.4 so we can keep going.

```
fourth_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+transgender_covid_cases,data = final_dataset)
modelsummary(fourth_model)
```

The AIC obtained is 1729.5 so we don't include transgender covid cases

```
fifth_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`20 to 29`,data = final_dataset)
modelsummary(fifth_model)
```

We keep the 20 to 29 variable since the AIC is lower.

	Model 1
(Intercept)	84.747
	(13.634)
FEMALE_Covid_CASES	−1.102
	(0.428)
MALE_Covid_CASES	0.383
	(0.563)
'20 to 29'	2.764
	(0.644)
Num.Obs.	140
R2	0.306
R2 Adj.	0.290
AIC	1712.6
BIC	1727.4
Log.Lik.	−851.324
F	19.951

```
sixth_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`20 to 29`+`30 to 39`,data = final_dataset)
modelsummary(sixth_model)
```

So we move on with this model since the AIC is 1697.2

```
seventh_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`20 to 29`+`30 to 39`+`40 to 49`,data = final_da
modelsummary(seventh_model)
```

We proceed with this model and will try to remove the 20 to 29 to see if we get a lower AIC.

```
eight_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`30 to 39`+`40 to 49`,data = final_dataset)
modelsummary(eight_model)
```

The AIC is not reduced so we keep 20 to 29 and we try and remove another variable.

```
seventh_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`20 to 29`+`40 to 49`,data = final_dataset)
modelsummary(seventh_model)
```

This time the AIC is higher so we do not go with this model.

So the final model we obtain is the seventh model

Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+20 `to 29`+ 30 to 39+40 to 49'

A.2   Model

These are the estimates obtained for the final model using modelsummary.

	Model 1
(Intercept)	68.019
	(13.455)
FEMALE_Covid_CASES	−1.148
	(0.404)
MALE_Covid_CASES	−0.525
	(0.573)
'20 to 29'	2.205
	(0.622)
'30 to 39'	3.970
	(0.939)
Num.Obs.	140
R2	0.387
R2 Adj.	0.369
AIC	1697.2
BIC	1714.9
Log.Lik.	−842.623
F	21.287

	Model 1
(Intercept)	64.112
	(13.639)
FEMALE_Covid_CASES	−0.918
	(0.430)
MALE_Covid_CASES	−0.343
	(0.583)
'20 to 29'	2.060
	(0.626)
'30 to 39'	4.370
	(0.972)
'40 to 49'	−1.573
	(1.041)
Num.Obs.	140
R2	0.397
R2 Adj.	0.375
AIC	1696.9
BIC	1717.5
Log.Lik.	−841.440
F	17.648

	Model 1
(Intercept)	54.075
	(13.768)
FEMALE_Covid_CASES	−1.139
	(0.440)
MALE_Covid_CASES	0.680
	(0.511)
'30 to 39'	5.150
	(0.976)
'40 to 49'	−2.097
	(1.065)
Num.Obs.	140
R2	0.348
R2 Adj.	0.329
AIC	1705.8
BIC	1723.4
Log.Lik.	−846.879
F	18.040

	Model 1
(Intercept)	84.331
	(13.764)
FEMALE_Covid_CASES	−1.058
	(0.458)
MALE_Covid_CASES	0.434
	(0.595)
'20 to 29'	2.747
	(0.649)
'40 to 49'	−0.295
	(1.070)
Num.Obs.	140
R2	0.306
R2 Adj.	0.285
AIC	1714.6
BIC	1732.2
Log.Lik.	−851.285
F	14.881

	Model 1
(Intercept)	64.112
	(13.639)
FEMALE_Covid_CASES	−0.918
	(0.430)
MALE_Covid_CASES	−0.343
	(0.583)
'20 to 29'	2.060
	(0.626)
'30 to 39'	4.370
	(0.972)
'40 to 49'	−1.573
	(1.041)
Num.Obs.	140
R2	0.397
R2 Adj.	0.375
AIC	1696.9
BIC	1717.5
Log.Lik.	−841.440
F	17.648

```
final_model<-
  lm(Assault_2020 ~FEMALE_Covid_CASES+MALE_Covid_CASES+`20 to 29`+`30 to 39`+`40 to 49`,
     data = final_dataset)
modelsummary(final_model)
```

B Appendix B

B.1 Covid Cases

This is a preview of the covid_cases dataset obtained from opendatatoronto (Gelfand 2020)

```
## Rows: 32000 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): Outbreak Associated, Age Group, Neighbourhood Name, FSA, Source o...
## dbl   (2): _id, Assigned_ID
## date  (2): Episode Date, Reported Date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.


## # A tibble: 32,000 x 18
##    `_id` Assigned_ID `Outbreak Associated` `Age Group`    `Neighbourhood~` FSA
##    <dbl>       <dbl> <chr>                 <chr>          <chr>            <chr>
## 1 62978       65427 Sporadic              30 to 39 Years Dovercourt-Wall~ M6H
## 2 62979       65428 Sporadic              30 to 39 Years York University~ M3J
## 3 62980       65429 Sporadic              50 to 59 Years Glenfield-Jane ~ M3N
## 4 62981       65431 Sporadic              30 to 39 Years Dufferin Grove   M6H
```

```
##  5 62982      65432 Sporadic            30 to 39 Years Malvern        M1B
##  6 62983      65433 Sporadic            19 and younger Black Creek    M3N
##  7 62984      65434 Outbreak Associated 30 to 39 Years L'Amoreaux     M1T
##  8 62985      65435 Sporadic            40 to 49 Years Bendale        M1P
##  9 62986      65436 Sporadic            50 to 59 Years Malvern        M1B
## 10 62987      65437 Outbreak Associated 80 to 89 Years Guildwood      M1E
## # ... with 31,990 more rows, and 12 more variables:
## #   'Source of Infection' <chr>, Classification <chr>, 'Episode Date' <date>,
## #   'Reported Date' <date>, 'Client Gender' <chr>, Outcome <chr>,
## #   'Currently Hospitalized' <chr>, 'Currently in ICU' <chr>,
## #   'Currently Intubated' <chr>, 'Ever Hospitalized' <chr>,
## #   'Ever in ICU' <chr>, 'Ever Intubated' <chr>
```

B.2   Neighbourhood crime rates

This is a preview of the neighbourhood crimes dataset obtained from opendatatoronto (Gelfand 2020)

```
## Rows: 140 Columns: 104
## -- Column specification ------------------------------------------------------
## Delimiter: ","
## chr   (3): Neighbourhood, Hood_ID, geometry
## dbl (101): _id, OBJECTID, F2020_Population_Projection, Assault_2014, Assault...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.


## # A tibble: 140 x 104
##    '_id' OBJECTID Neighbourhood        Hood_ID F2020_Populatio~ Assault_2014
##    <dbl>    <dbl> <chr>                <chr>              <dbl>        <dbl>
## 1      1        1 Yonge-St.Clair       097                14083           16
## 2      2        2 York University Heights 027             30277          273
## 3      3        3 Lansing-Westgate     038                18146           42
## 4      4        4 Yorkdale-Glen Park   031                17560          106
## 5      5        5 Stonegate-Queensway  016                27410           91
## 6      6        6 Tam O'Shanter-Sullivan 118              29970          103
## 7      7        7 The Beaches          063                23364           88
## 8      8        8 Thistletown-Beaumond He~ 003            10948           61
## 9      9        9 Thorncliffe Park     055                23518           86
## 10    10       10 Danforth East York   059                18427           68
## # ... with 130 more rows, and 98 more variables: Assault_2015 <dbl>,
## #   Assault_2016 <dbl>, Assault_2017 <dbl>, Assault_2018 <dbl>,
## #   Assault_2019 <dbl>, Assault_2020 <dbl>, Assault_Rate2014 <dbl>,
## #   Assault_Rate2015 <dbl>, Assault_Rate2016 <dbl>, Assault_Rate2017 <dbl>,
## #   Assault_Rate2018 <dbl>, Assault_Rate2019 <dbl>, Assault_Rate2020 <dbl>,
## #   AutoTheft_2014 <dbl>, AutoTheft_2015 <dbl>, AutoTheft_2016 <dbl>,
## #   AutoTheft_2017 <dbl>, AutoTheft_2018 <dbl>, AutoTheft_2019 <dbl>, ...
```

B.3   Appendix C

The datasheet below is an enhancement for the paper and we will be constructing the datasheet for the covid cases database

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 was created to allow people to track the number of reported covid cases in Toronto and contains demographic, geographic and severity information for all cases of covid that were reported whether or not they were probable or confirmed.

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 Toronto Public Health organization on behalf of the city.

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.*

   - Toronto Public Health organization is a public organization so the creation of the dataset was funded by the city.

4. *Any other comments?*

   - NA

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 that comprise the dataset represent the neighbourhoods in Toronto. This includes downtown Toronto, Scarborough neighbourhoods, etc.

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

   - There are 140 instances since there are 140 neighbourhoods.

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).*

   - The dataset contains all possible instances since it includes all 140 neighbourhoods in TOronto.

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.*

   - Each instance consists of the details about a reported covid case, whether the case was probably or confirmed, whether the client gender was male, female, trans etc, whether it was from an outbreak or sporadic and whether or not the case was resolved.

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

   - There is no target associated with any of the instances.

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 are some NA cases where information in the data features is missing, possibly because they were unable to collect that information when it was reported.

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 made explicit between 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, source 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.*

    - TThe dataset updates on a weekly basis since the pandemic is still ongoing.

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.*

    - It does not contain confidential data

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.*

    - It does not contain such data.

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 identifies the gender of the client and the age group of the reported individual with covid.

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

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.*

    - It does not contain such sensitive data.

16. *Any other comments?*

    - NA

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 was extracted from the provincial Case & Contact Management System (CCM).

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?*

   - The data was collected by recording the features for the person who was reported to have covid in each instance.

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

   - NA

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 provincial Case & Contact Management System was involved and their compensation is not recorded

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 collects all covid cases from January 2020 to the present date.

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.*

   - No ethical review processes appear to have been 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)?*

   - I obtained it using opendatatoronto, a public website.

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.*

   - No

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.*

   - Since it was a public website, I did not require their consent.

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).*

    - NA

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.*

    - NA

12. *Any other comments?*

    - NA

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.*

   - NA

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.*

   - NA

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

   - NA

4. *Any other comments?*

   - NA

Uses

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

   - The dataset has been used for no 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.*

   - https://github.com/Varun1005473462/final_folder-main.git

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

   - It could be used for modelling the spread of covid in Toronto neighbourhoods since 2020.

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?*

   - TBD

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

   - TBD

6. *Any other comments?*

   - TBD

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.*

   - The dataset will not be distributed to third parties outside of the entity on behalf of which the dataset was created

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

   - NA

3. *When will the dataset be distributed?*

   - NA

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.*

   - No

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.*

   - No

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.*

   - No

7. *Any other comments?*

   - NA

Maintenance

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

   - Varun Vijay

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

   - through mail at varun.vijay@mail.utoronto.ca

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

   - No

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)?*

   - The dataset is updated weekly by the city to account for new covid cases. Any updates can be communicated through github.

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.*

   - No

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.*

   - No. Its obsolescence is communicated via commit history on 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.*

   - f others want to extend/augment/build on/contribute to the dataset, they can do s using the github link above. These contributions will not be validated/ verified since I am not responsible for such data extensions

8. *Any other comments?*

   - N/A

C References

Arel-Bundock, Vincent, Joachim Gassen, Nathan Eastwood, Nick Huntington-Klein, Moritz Schwarz, Benjamin Elbers, and Grant McDermott. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://cran.r-project.org/package=modelsummary.

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

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal.* https://CRAN.R-project.org/package=opendatatoronto.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto, Open Data. 2022a. "About COVID-19 Cases in Toronto." City of Toronto Open Data Portal. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

———. 2022b. "About Neighbourhood Crime Rates." City of Toronto Open Data Portal. https://open.toronto.ca/dataset/neighbourhood-crime-rates/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Tidyr: Tidy Messy Data.* https://tidyr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "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. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, Romain Francois, Jennifer Bryan, Shelby Bearrows, Rstudio, Jukka Jylänki, and Mikkel Jørgensen. 2022. *Readr: Read Rectangular Text Data.* https://cran.r-project.org/package=readr.

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

———. 2021b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents.* https://github.com/yihui/tinytex.

(Wickham et al. 2019b)

(Wickham 2016)

(Wickham et al. 2021)

(Wickham 2022)

(Firke 2021)

(Wickham 2007)

(Müller 2020)

(Xie 2021b)

(Wickham et al. 2022)

(Toronto 2022b)

(Toronto 2022a)