Project Work

Faizan Naseer, Chang Ma, and Lei Lim

3/18/2022

Description of the variables of the data

The data we use concludes the cases of covid-19 since January 1st, 2020, and the last update is on March 25th, 2022. There are total of 18 variables in the dataset which describes the demographic, geographic, and severity information for all confirmed and probable cases.

Variable	Description
id	Unique identifier of each row
Assigned_ID	Unique ID assigned to cases by Toronto Public Health
Outbreak Associated	outbreaks of COVID-19 in Toronto
Age Group	Age of the person who got COVID
Neighborhood Name	The name of one of the 140 geographically distinct areas in Toronto
FSA	Forward sortation area (first three characters of postal code)
Source of Infection	The most likely way of how the COVID is acquired
Classification	The identification of either the case is confirmed or probable
Episode Date	A derived variable that best estimates when the disease was acquired
Reported Date	The date which the case is reported
Client Gender	Gender of the person
Outcome	Fatal/Resolved/Active
Currently Hospitalized	Cases that are currently admitted to hospital
Currently in ICU	Cases that currently admitted to the ICU
Currently Intubated	Cases that were intubated related to their COIVD infection
Ever in ICU	Currently that were admitted to ICU because of COVID infection
Ever Intubated	Cases that were intubated because of COVID infection.
Reported Date	The date of the case was reported

Background of the data

The data is reported and managed by Public Health of Toronto, and it includes the cases that are sporadic, and outbreak associated. Furthermore, the data are extracted from the provincial Case & Contact Management System (CCM) which is a is a central data repository for COVID-19 case and contact management, and reporting in Ontario.

Research Goal

Our goal for this research project is to observe and analyze the data and find the pattern of how the number of COVID cases change as time goes further. In addition, we will predict what will happen to the growing rate of the number of cases in the future. We are going to accomplish our goal by using different static tests and techniques such as the t test.

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.1.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ─────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

Tables

The number of COVID cases in each month

```
##    Reported.Date number
## 1        2020-01      2
## 2        2020-02      7
## 3        2020-03    974
## 4        2020-04   5465
## 5        2020-05   5129
## 6        2020-06   2718
## 7        2020-07   1015
## 8        2020-08    785
## 9        2020-09   3836
## 10       2020-10   9049
## 11       2020-11  13787
## 12       2020-12  19730
## 13       2021-01  25391
## 14       2021-02  10381
## 15       2021-03  16337
## 16       2021-04  35007
## 17       2021-05  17881
## 18       2021-06   2443
## 19       2021-07    909
## 20       2021-08   3662
## 21       2021-09   4133
## 22       2021-10   2328
## 23       2021-11   2407
## 24       2021-12  43172
## 25       2022-01  53613
## 26       2022-02   9758
## 27       2022-03   8993
```

In this table, we can see that the number of cases keeps increasing from January 2020 to May 2021, and then the number starts to decrease. However, the number of cases suddenly went up in January 2022, and then it went down again after that month.

The Number of COVID cases of each gender

```
##                Client.Gender number
## 1                     FEMALE 153280
## 2                       MALE 142958
## 3                 NON-BINARY    134
## 4 NOT LISTED, PLEASE SPECIFY      2
## 5                      OTHER     15
## 6                  TRANS MAN     19
## 7                TRANS WOMAN     12
## 8                TRANSGENDER     22
## 9                    UNKNOWN   2470
```

From the table above, we notice that the number of female cases is more than the number of male cases, but the ratio of the two categories are very close together. /

The Number COVID cases of each age Group

```
##        Age.Group number
## 1 19 and younger  45553
## 2 20 to 29 Years  61177
## 3 30 to 39 Years  54700
## 4 40 to 49 Years  42304
## 5 50 to 59 Years  38255
## 6 60 to 69 Years  23606
## 7 70 to 79 Years  11763
## 8 80 to 89 Years   8874
## 9   90 and older   4889
```

From this table, we can see that most of the cases are from the age group of 20 to 29. /

The number of COVID cases from each source of Infection

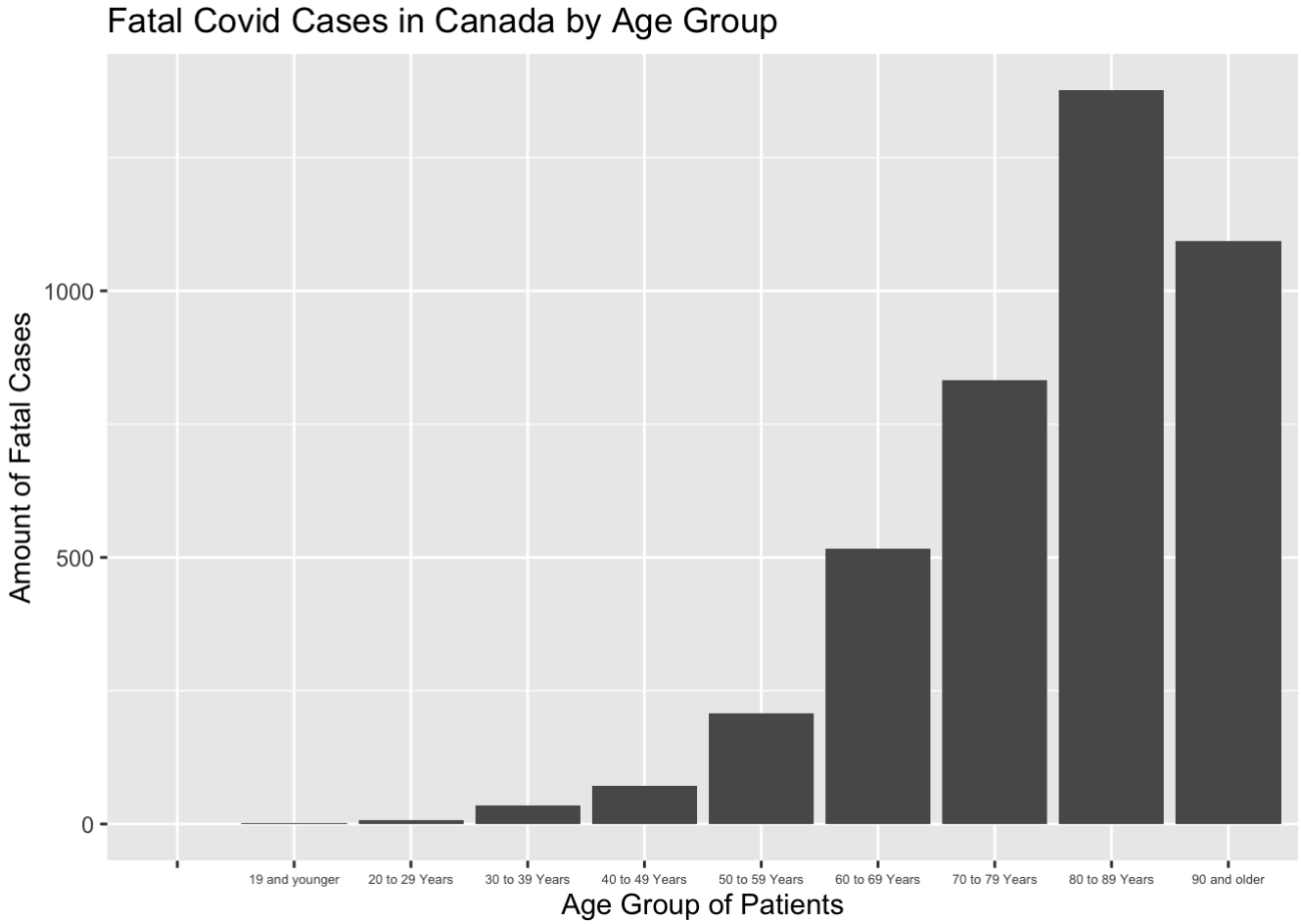
```
##                  Source.of.Infection number
## 1                       Close Contact  16321
## 2                           Community  65755
## 3                   Household Contact  37523
## 4                      No Information 135068
## 5     Outbreaks, Congregate Settings   4054
## 6 Outbreaks, Healthcare Institutions  18032
## 7          Outbreaks, Other Settings  10485
## 8                            Pending     61
## 9                             Travel   3822
```

In this table, we can see that the source of most cases can not be determined, but we can still see that community inflection is the biggest source of infection from the knowing cases.

Graphs

Graph 1

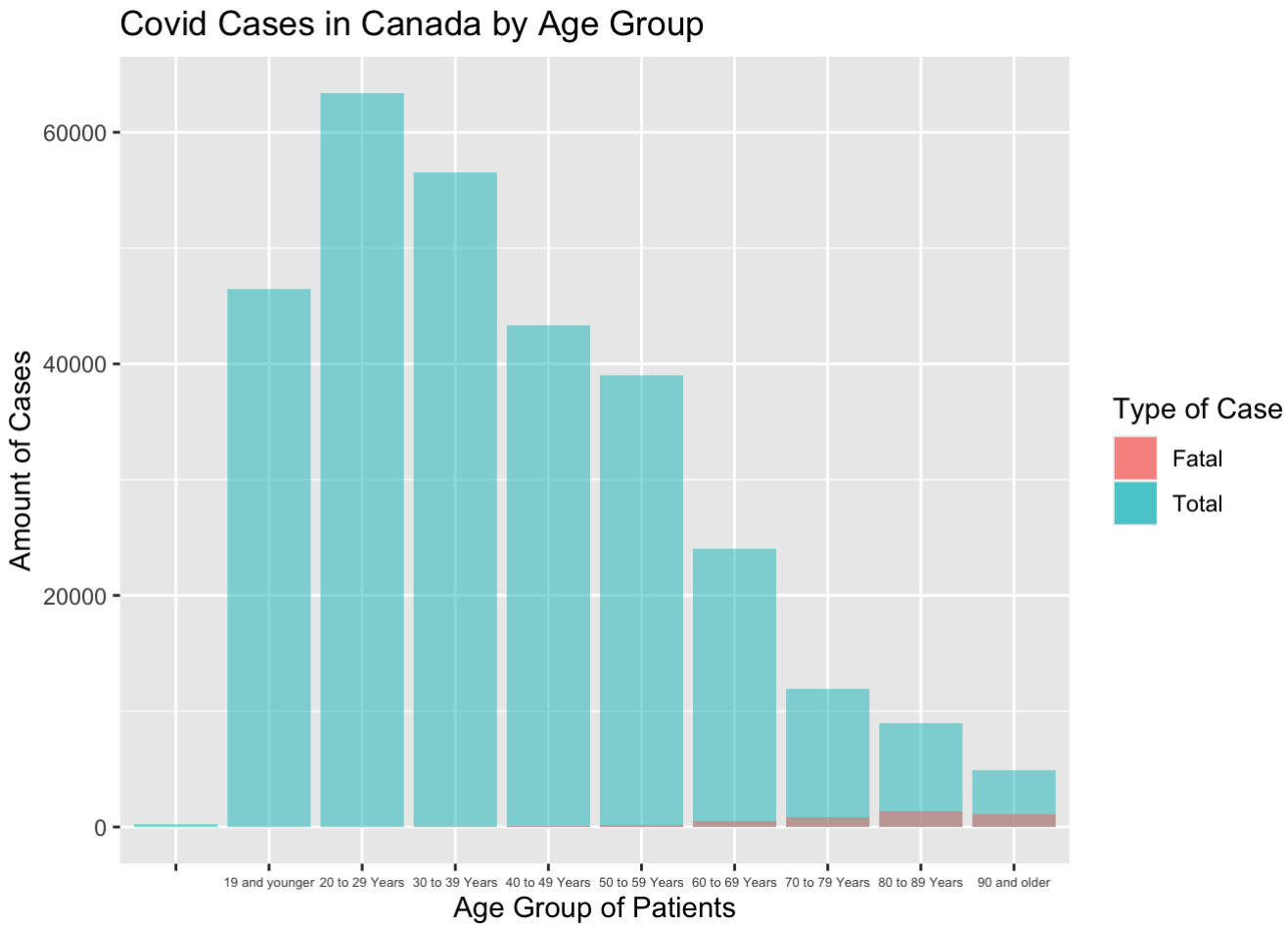
In this first graph, a bar graph will be plotted with the age group of the patients on the x-axis and the amount of fatal cases on the y-axis. Fatal cases in this situation means that the covid patient has died due to contracting the virus.



As it can be seen from the graph, it seems like fatal cases exponentially increases as the age group of the patients increase, right until the end, at 90 and older. However, I believe that is due to the much lower population of 90 and older people, leading to a much lesser amount of them contracting the virus, and thus lower amounts of fatal cases.

Graph 2

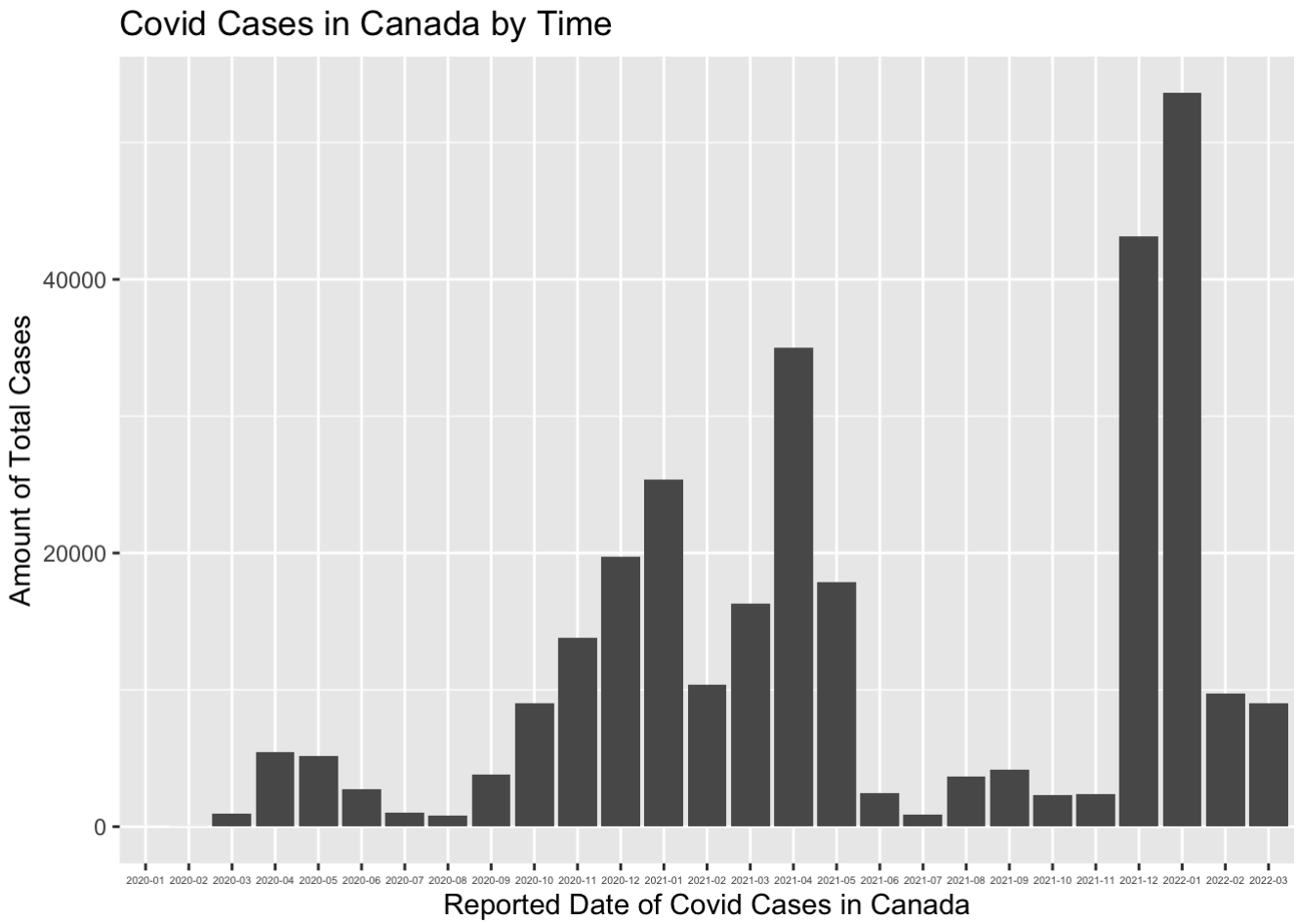
If the graph of fatal cases were to be plotted in the same graph as total cases by age group, the following would show:



As it can clearly be seen from the graph, there is much more covid patients in the age groups of 19 to 40, but the amounts of fatal cases are almost not visible. On the other hand, even when there are much less cases in the older ages, they are still more susceptible to having their covid case be fatal. This clearly shows that the older people are more likely to die from the virus due to their weaker immune system, while even when thousands of young people contract the virus the amount of fatal cases are almost negligible.

Graph 3

In the final graph, the graph of total covid cases in Canada will be plotted against time. The graph will start at 2020 January when covid first surfaced and will end at March 2022, as it is the current time and no further data can be retrieved for the future.



From the graph, it can be seen that there really isn't a correlation in the data. There isn't really an increasing or decreasing trend in the reported covid cases in relation to time. However, it can be seen that there are periods of times when cases are very stable and low, such as from 2020-03 to 2020-09 and from 2021-06 to 2021-11, and much higher spikes of cases in the other months. However, this data is not conclusive enough for us to predict a trend and such.

Prop Testing

In this prop test, we are testing whether the true proportion of covid-19 patients whose outcome is FATAL is 0.013 or not.

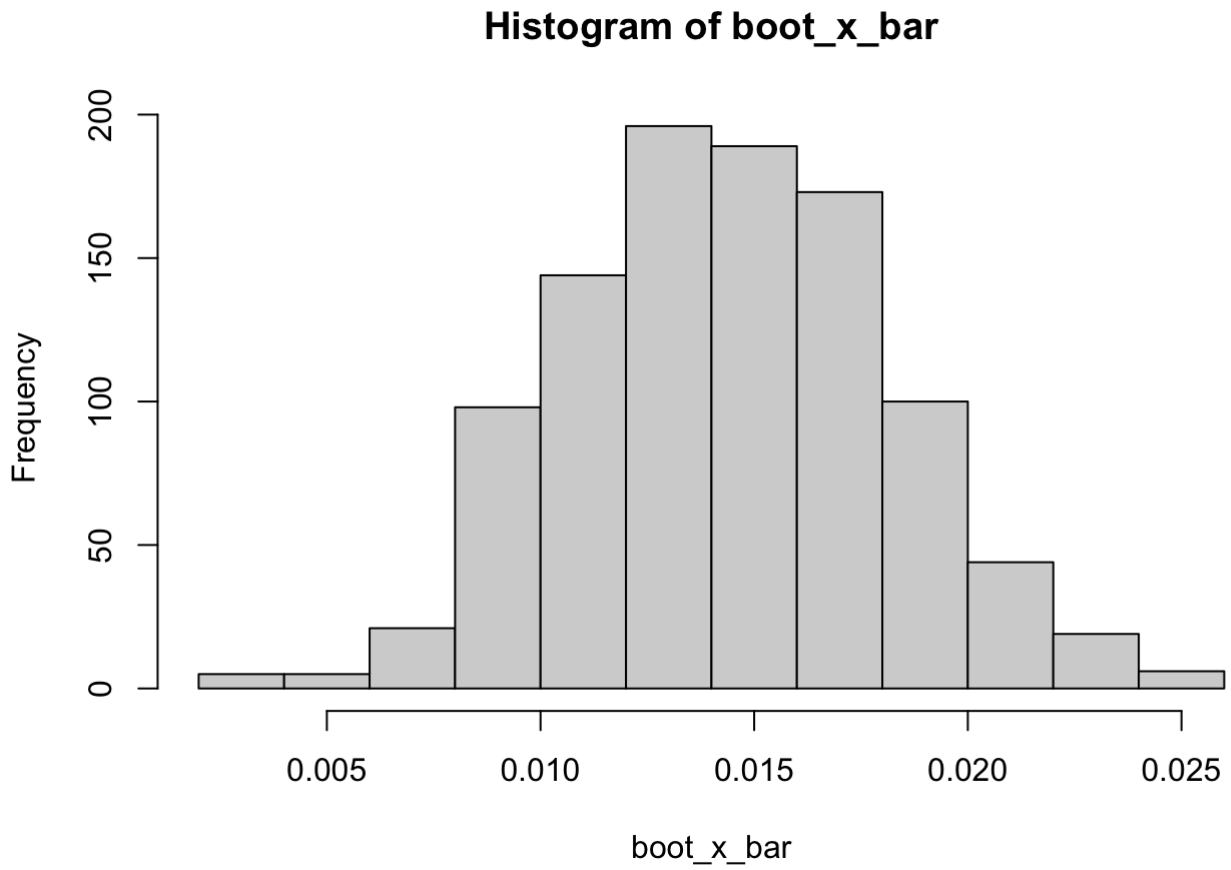
Suppose π is the true proportion. We are interested in: $H_O : \pi = 0.013$ vs $\pi \neq 0.013$

```
##
##  1-sample proportions test with continuity correction
##
## data:  d2$x out of d2$n, null probability 0.013
## X-squared = 17.04, df = 1, p-value = 3.66e-05
## alternative hypothesis: true p is not equal to 0.013
## 5 percent confidence interval:
##  0.01384185 0.01387201
## sample estimates:
##          p
## 0.01385692
```

Since the p-value is very very small, it shows that it is strongly against the null hypothesis, and therefore we have to reject it. Therefore the alternative hypothesis is right, where $\pi \neq 0.013$.

Bootstrapping

In this bootstrapping model, we will sample data where the outcome of the virus is “FATAL”, and replicate it 1000 times and plot out the histogram.



This shows the histogram created from replicating the boot function. It can be seen that the mid point is around 0.014.

To find the 95th percentile, the following code has to run:

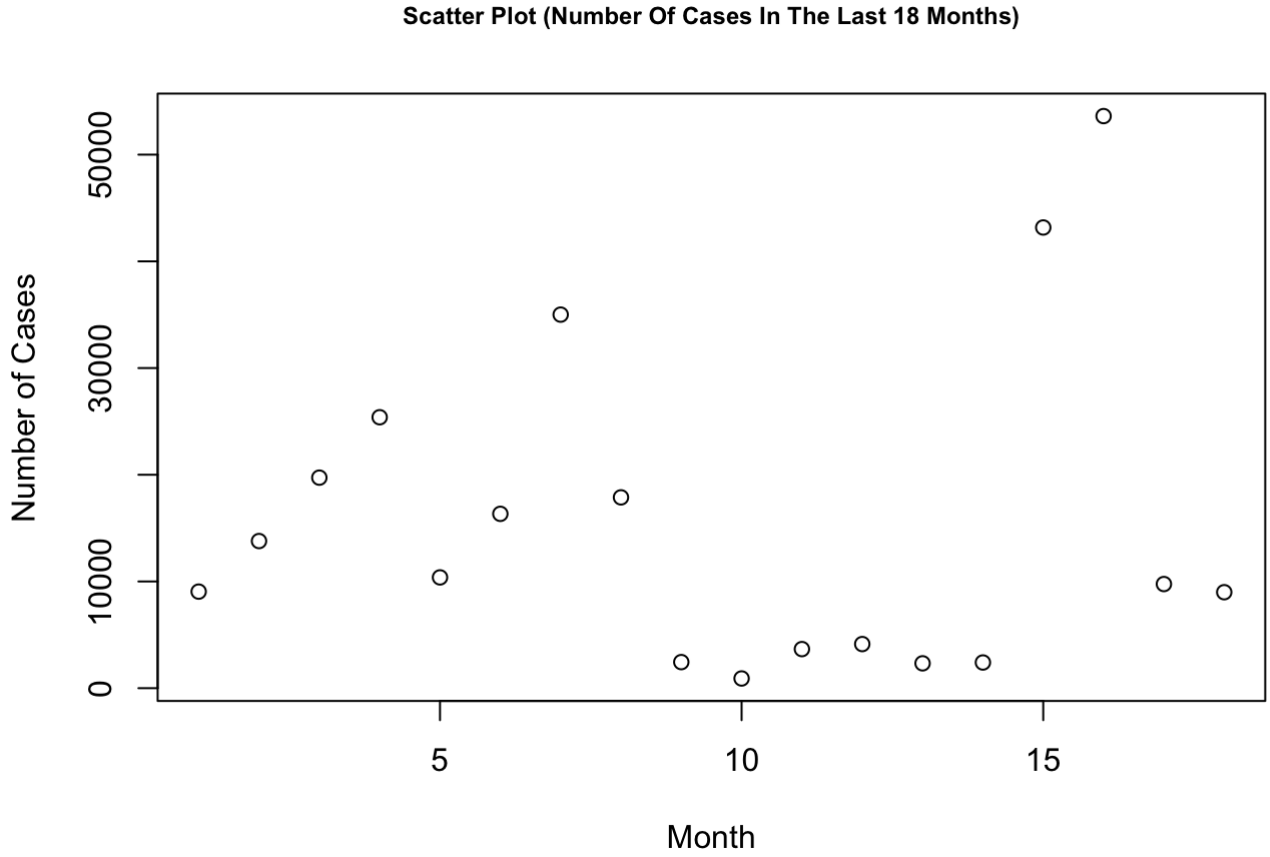
```
##     2.5%    97.5% 
## 0.008000 0.022025
```

As it can be seen, these are the values where 95% of the times, the percentages will fall in between.

Regression Analysis

To predict how the COVID-19 pandemic will affect Toronto in the future using regression, we can plot a graph of how the overall number of cases have fluctuated in the last 18 months, as the months have passed by. Subsequently, an attempt at mapping a linear model to the graph can also be used, to help aid our predictions.

Linear Model 1 Summary




```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14696 -12092  -4400   4886  36730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13476.5     7621.0   1.768   0.0961 .
## x              212.9      704.1   0.302   0.7663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15500 on 16 degrees of freedom
## Multiple R-squared:  0.005682,   Adjusted R-squared:  -0.05646
## F-statistic: 0.09143 on 1 and 16 DF,  p-value: 0.7663
```

Interpretation Of Regression Parameters

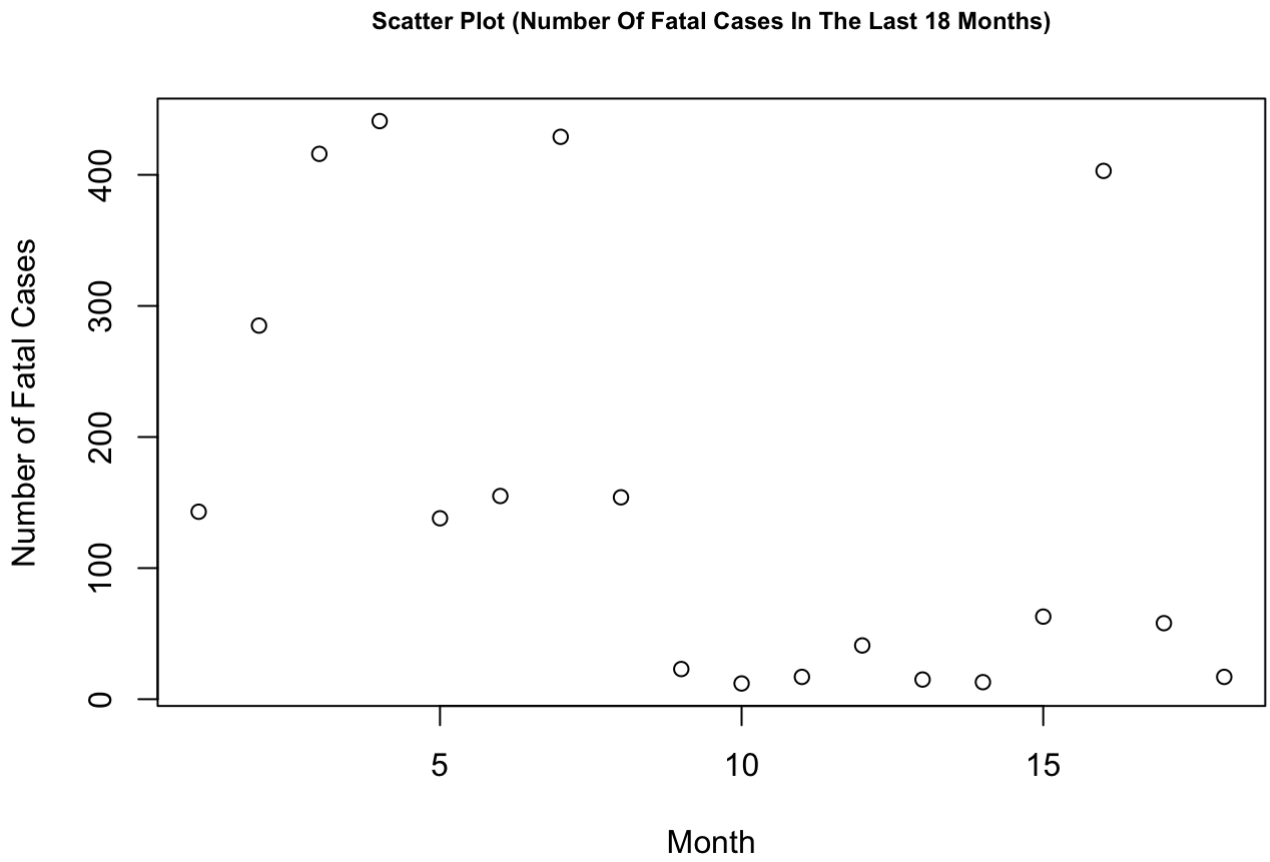
β_1 according to the linear model created is 13476.5 - this is the number of cases at the start of the 1st month.

β_2 according to the linear model created is 212.9 - this means that the number of cases increase by approximately 213 cases, as each month passes by.

For both these parameters, however, the p-value is greater than 0.005 and thus the null hypothesis is mostly accepted (β_1 and β_2 are equal to 0) - this is because of the fact that there is no clear, moderate or strong correlation between the last 18 months and the number of cases, deeming the linear model inaccurate.

To gain a better linear model on the number of cases and how they change as the time passes by, the number of fatal cases were then looked into.

Linear Model 2 Summary



```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141.23  -88.88  -39.33   15.92  342.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  297.431     72.132   4.123 0.000796 ***
## x            -14.800      6.664  -2.221 0.041142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.7 on 16 degrees of freedom
## Multiple R-squared:  0.2356, Adjusted R-squared:  0.1879
## F-statistic: 4.932 on 1 and 16 DF,  p-value: 0.04114
```

Interpretation Of Regression Parameters

This time, β_1 according to the linear model is 297.431 - this is the number of fatal cases at the start of the first month.

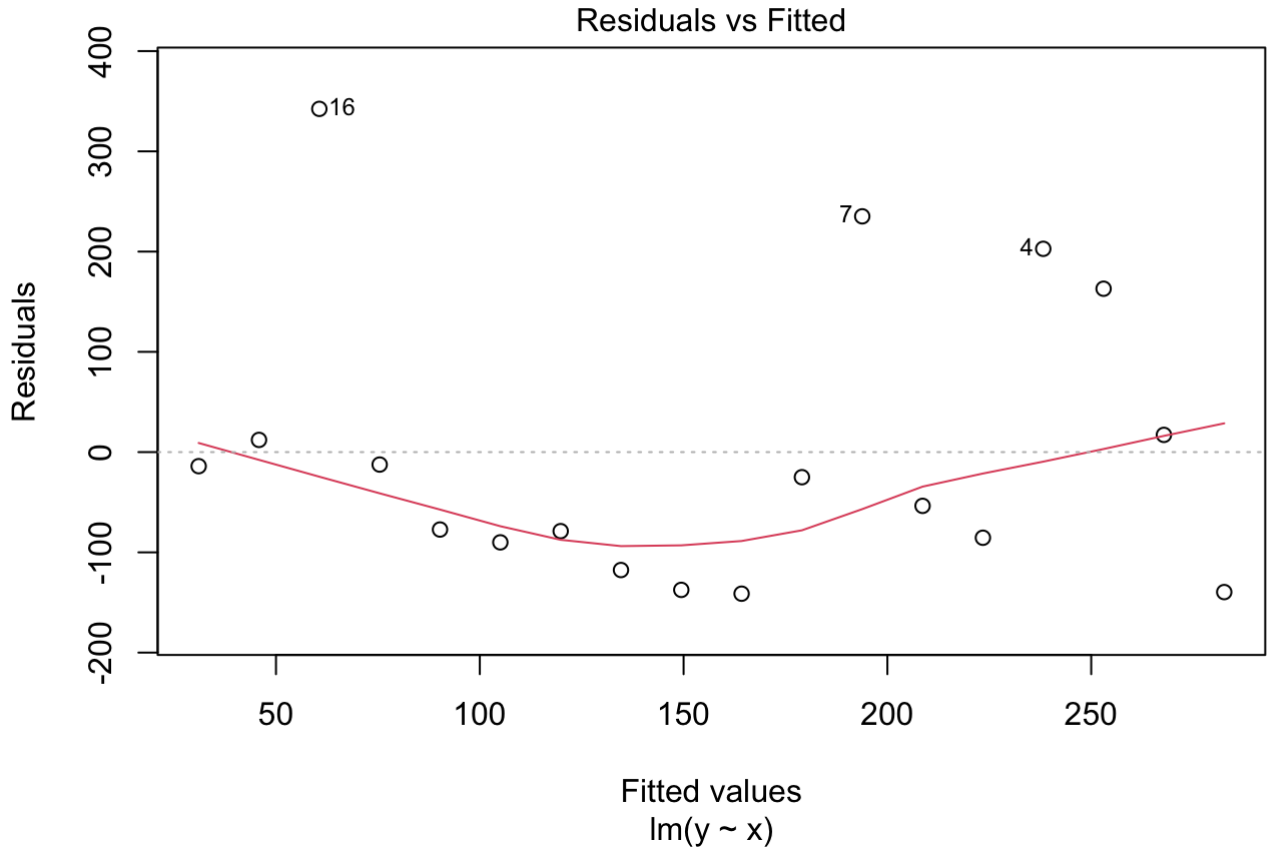
β_2 according to the linear model is -14.8 - this means the number of fatal cases decrease by approximately 15 cases, as each month passes by.

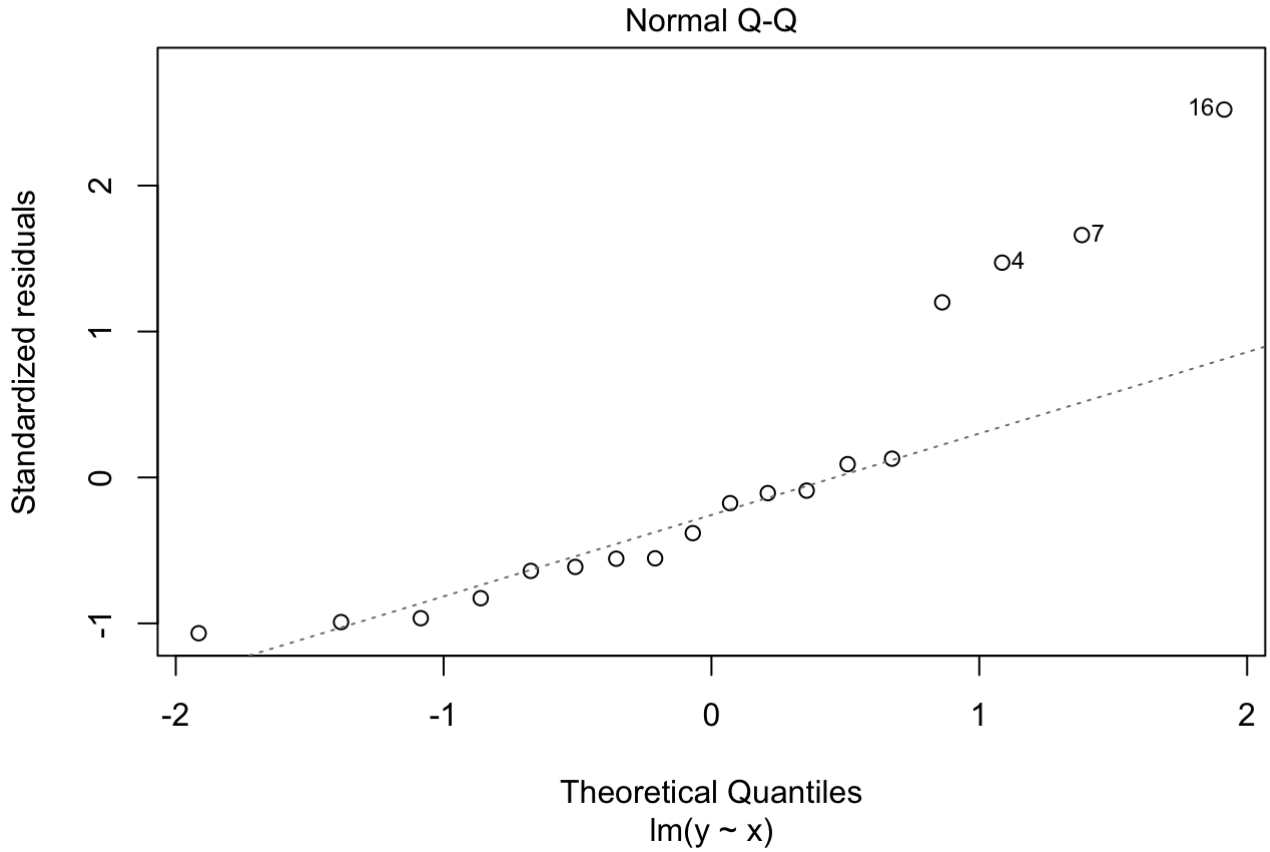
For both these parameters, the p-value is less than 0.005 and thus the null hypothesis is rejected as well (β_1 and β_2 are not equal to 0).

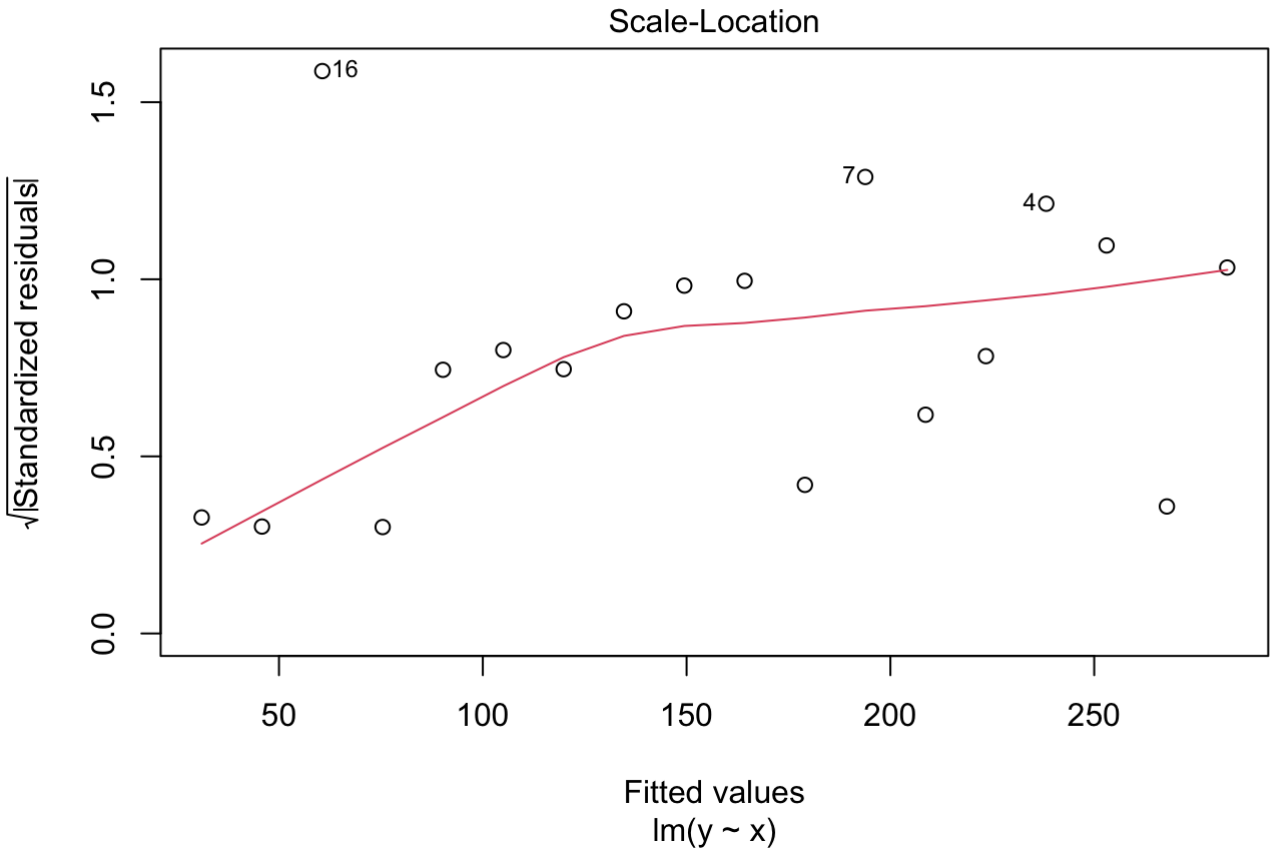
Additionally, the least square regression line predicts that the number of fatal cases will equal 0 after approximately two months. 297.431 -14.8x = 0 => x equals approximately 20 => 20 - 18 = 2

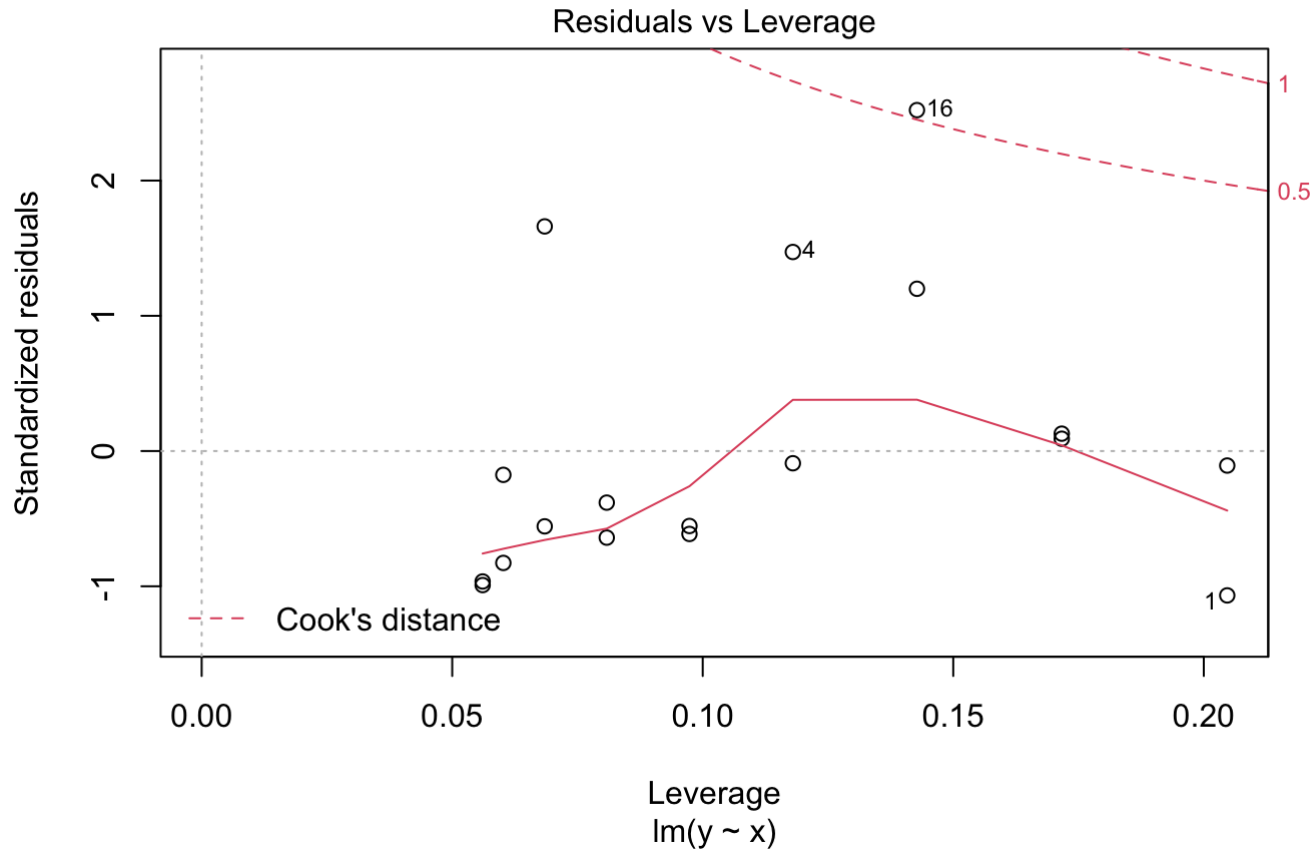
Linear Regression Model Plots

4 Types:






Scale-Location
√|Standardized residuals|
16
7
4
0.0
0.5
1.0
1.5
50
100
150
200
250
Fitted values
lm(y ~ x)


Residuals vs Leverage
Standardized residuals
16
4
1
1
0.5
Cook's distance
-1
0
1
2
0.00
0.05
0.10
0.15
0.20
Leverage
lm(y ~ x)

Mean Squared Error of the Linear Model

```
## [1] 19124.59
```

Cross Validation

The data can be split up into a 60% - 40% proportion where the model is trained and built based on the 60% proportion, and is then externally validated by checking through the 40% proportion.

Training MSE:

```
## [1] 20942.97
```

Test MSE:

```
## [1] 5300.957
```

As the mean squared error value is quite far from 0, the linear model is still not reliable to an extent - this has been checked through cross-validating the data values.

Final Summary

A data set consisting of COVID-19 information in Toronto was used to thoroughly investigate and analyze how the pandemic has affected different age groups, and how it has changed as the months have passed by. A set of tables were created on the sources of infection, and on how the number of cases fluctuated between different time spans - an increase through Jan 2020 to May 2021, then a decrease and then a sudden spike and fall through Jan 2022.

Graphs related to the relation between age groups and fatality, cases and time, etc. then followed, portraying a relation between older age groups and fatal cases. Proportion tests were also utilized to find the proportion of fatal cases overall, using p-values and null hypotheses accordingly.

Finally, a linear regression model was applied to several aspects of the data set, e.g. overall cases, fatal cases, etc. to help predict the decline of COVID-19 cases in Toronto, where the accuracy, reliability and suitability of the models were gauged using p-values and cross-validation.

Appendix (Code)

```
Variable<-c("id",
"Assigned_ID",
"Outbreak Associated",
"Age Group",
"Neighborhood Name",
"FSA",
"Source of Infection",
"Classification",
"Episode Date",
"Reported Date",
"Client Gender",
"Outcome",
"Currently Hospitalized",
"Currently in ICU",
"Currently Intubated",
"Ever in ICU",
"Ever Intubated",
"Reported Date"
)
Description<-c(
"Unique identifier of each row",
"Unique ID assigned to cases by Toronto Public Health",
"outbreaks of COVID-19 in Toronto ",
"Age of the person who got COVID",
"The name of one of the 140 geographically distinct areas in Toronto",
"Forward sortation area (first three characters of postal code)",
"The most likely way of how the COVID is acquired",
"The identification of either the case is confirmed or probable",
"A derived variable that best estimates when the disease was acquired",
"The date which the case is reported",
"Gender of the person",
"Fatal/Resolved/Active",
"Cases that are currently admitted to hospital",
"Cases that currently admitted to the ICU",
"Cases that were intubated related to their COIVD infection",
"Currently that were admitted to ICU because of COVID infection",
"Cases that were intubated because of COVID infection.",
"The date of the case was reported"
)
library(kableExtra)
df <- data.frame(Variable,Description)
kable(df)
```

```
library(tidyverse)
pop <- read.csv("/Users/faizannaseer/Downloads/COVID19 cases.csv")
```

```
pop1 = pop %>% mutate(Reported.Date = str_sub(Reported.Date,start=1,end=7))
data.frame(pop1 %>% group_by(Reported.Date)%>% summarise(number=n()))
```

```
data.frame(pop %>% group_by(Client.Gender) %>% summarise(number=n()))
```

```
pop [pop == ""] <- NA
data.frame(pop %>% group_by(Age.Group) %>% na.omit() %>% summarise(number=n()))
```

```
pop [pop == ""] <- NA
data.frame(pop %>% group_by(Source.of.Infection) %>% na.omit() %>% summarise(number=n
()))
```

```
library(tidyverse)
library(ggplot2)

pop <- read.csv("/Users/faizannaseer/Downloads/COVID19 cases.csv")
```

```
pop = pop %>% mutate(Age = as.factor(pop$Age.Group))

pop %>% group_by(Age) %>% summarize(n=n(), fatal = sum(Outcome == "FATAL")) %>%
  ggplot(aes(Age, fatal)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(size = 5)) +
  ggtitle("Fatal Covid Cases in Canada by Age Group") +
  xlab("Age Group of Patients") +
  ylab("Amount of Fatal Cases")
```

```
fatal = pop %>% group_by(Age) %>% summarize(n = sum(Outcome == "FATAL"))
total = pop %>% group_by(Age) %>% summarize(n = n())

ggplot(NULL, aes(Age, n)) +
  geom_bar(stat = 'identity', aes(fill = "Total"), data = total, alpha = 0.5) +
  geom_bar(stat = 'identity', aes(fill = "Fatal"), data = fatal, alpha = 0.5) +
  theme(axis.text.x = element_text(size = 5)) +
  ggtitle("Covid Cases in Canada by Age Group") +
  xlab("Age Group of Patients") +
  ylab("Amount of Cases") +
  guides(fill=guide_legend(title="Type of Case"))
```

```
pop = pop %>% mutate(Date = as.factor(str_sub(pop$Reported.Date, start = 1, end =
7)))

pop %>% group_by(Date) %>% summarize(n=n()) %>%
  ggplot(aes(Date, n)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(size = 4)) +
  ggtitle("Covid Cases in Canada by Time") +
  xlab("Reported Date of Covid Cases in Canada") +
  ylab("Amount of Total Cases")
```

```
d2 = pop %>% summarize( x= sum(Outcome == "FATAL"), n = n())

prop.test(x = d2$x,
          n = d2$n,
          p = 0.013,
          alternative="two.sided",
          conf.level = 0.05)
```

```
set.seed(999)
fatalmean = pop %>% summarize(fatalmean = mean(Outcome == "FATAL"))

boot_function=function(){

  d2 = pop[ sample(c(1:250000),size=1000,replace=T) , ]

  prop = d2 %>% summarize(prop = mean(Outcome == "FATAL"))

  return(prop[1,1])

}

boot_x_bar = replicate(1000,boot_function())
hist(boot_x_bar)
```

```
quantile(boot_x_bar, c(0.025,0.975))
```

```
library(tidyverse)
pop <- read.csv("/Users/faizannaseer/Downloads/COVID19 cases.csv")
pop = pop %>% mutate(Date = as.factor(str_sub(pop$Reported.Date, start = 1, end =
7)))
pop_dates = pop %>% group_by(Date) %>% summarise(n = n())
x = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)
y = tail(pop_dates$n, 18)

plot(x, y)

m = lm(y~x)
summary(m)
```

```
pop_dates = pop %>% filter(Outcome == "FATAL") %>% group_by(Date) %>% summarise(n = n
())
x = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)
y = tail(pop_dates$n, 18)

plot(x, y)

m = lm(y~x)
summary(m)
```

```
plot(m)
```

```
y.hat = predict(m)
mse = mean((y-y.hat)^2)
mse
```

```
pop_dates = pop_dates[-c(1:7), ]
pop_dates = pop_dates %>% mutate(group_ind = sample(c("train", "test"),
                                                    size = nrow(pop_dates),
                                                    prob = c(0.6, 0.4),
                                                    replace = T))
pop_dates = pop_dates %>% mutate(x = x)
m = lm(n~x, data = pop_dates %>% filter(group_ind == "train"))

y.hat = predict(m)
mean((pop_dates$n[pop_dates$group_ind == "train"] - y.hat)^2)
```

```
y.hat = predict(m, newdata = pop_dates %>% filter(group_ind == "test"))
mean((pop_dates$n[pop_dates$group_ind == "test"] - y.hat)^2)
```