COVID-19 in Toronto

Kevin Li

3/29/2023

Abstract

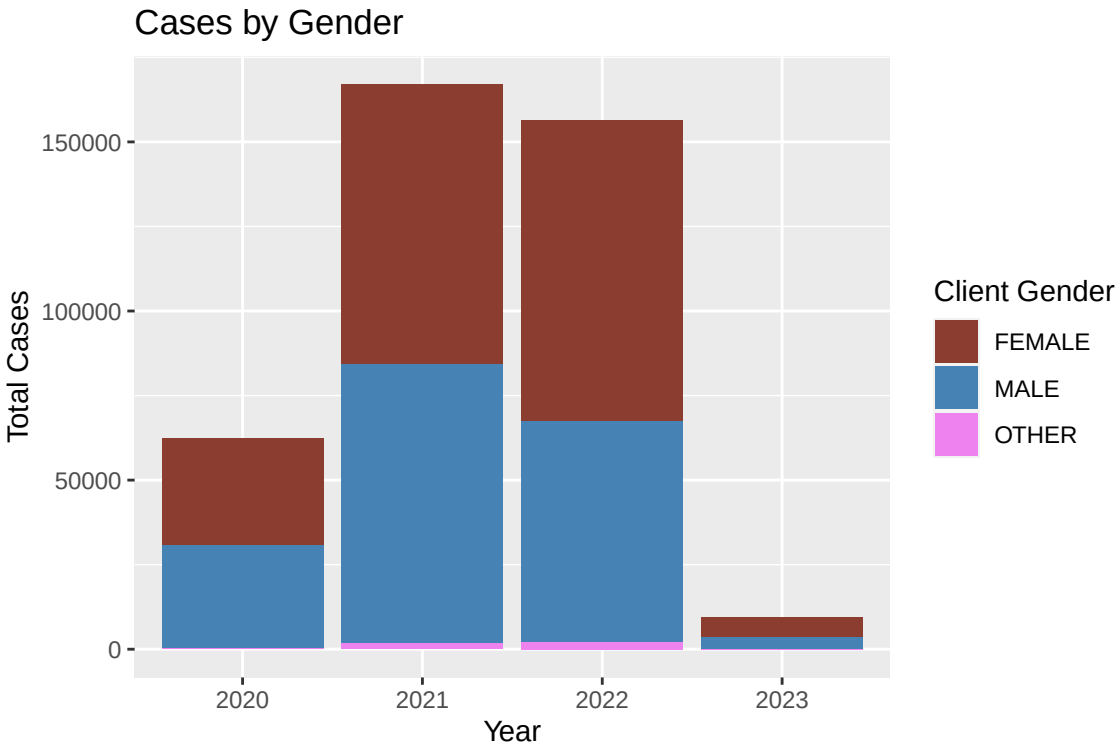
This report aims to establish a relationship between COVID-19 cases in Toronto based on data released by the City of Toronto and fatalities, and will attempt to fit a classifier to determine chance of survival for a given case. The features that will be analyzed are:

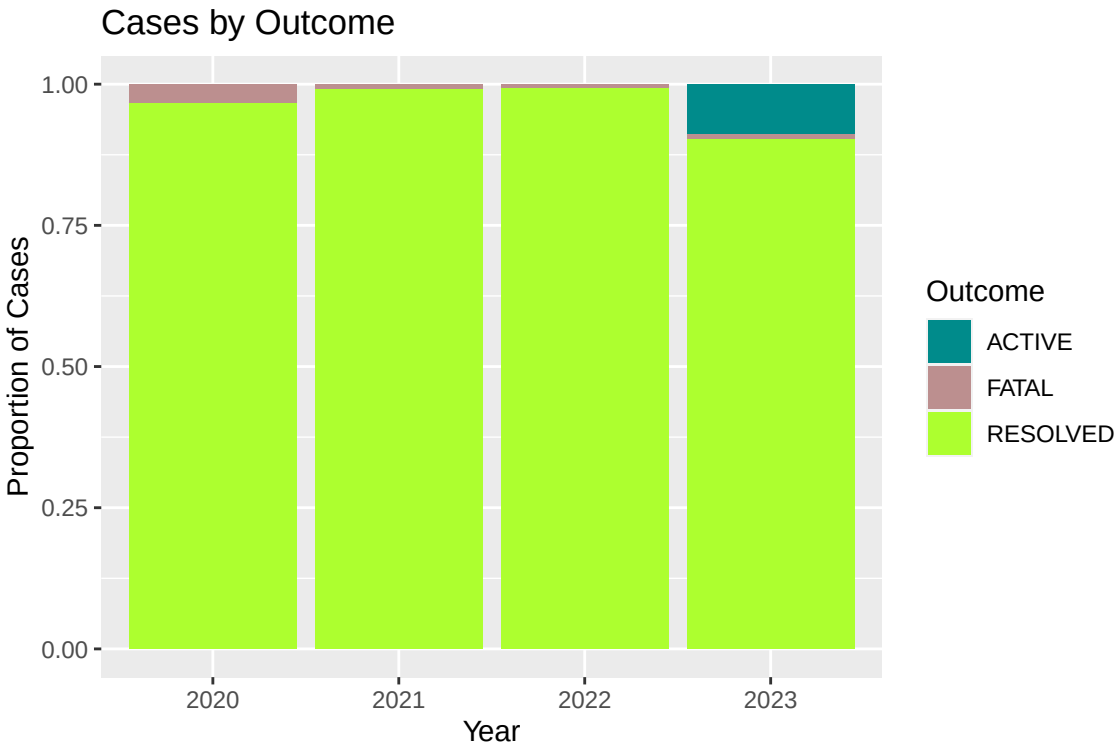
- Age, which is split by decade starting from ≤ 19 to ≥ 90
- Neighborhood, geographically divided into 140 distinct regions
  - Additionally, average income and population by neighbourhood was considered when modeling a relationship - this data was extracted from a different dataset, which can be found here: https://open.toronto.ca/dataset/neighbourhood-profiles/
- Full Street Address: The first 3 characters of the postal code
- Infection source: either travel, household contact close contact, outbreak, or community (none of the former sources)
- Classification: either confirmed or probable
- Gender
- If the patient was ever hospitalized, in ICU, Intubated
- Outcome, the feature to be classified - either active, resolved or fatal

The data was compiled by Toronto Public Health in response to the COVID-19 pandemic and released under the Open Government License, available here: https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Observations

Breakdown of cases by year:





Total fatalities: 4821 (1.22% of cases)

Rate of patients admitted to Hospital/ICU/Intubation:

```
## # A tibble: 4 x 5
## # Groups:   Ever Hospitalized, Ever in ICU [3]
##   `Ever Hospitalized` `Ever in ICU` `Ever Intubated`      N Proportion
##   <lgl>               <lgl>         <lgl>             <int>      <dbl>
## 1 FALSE               FALSE         FALSE            377973    0.956
## 2 TRUE                FALSE         FALSE             14338    0.0363
## 3 TRUE                TRUE          FALSE              1238    0.00313
## 4 TRUE                TRUE          TRUE               1682    0.00426
```

Patients by source of infection:

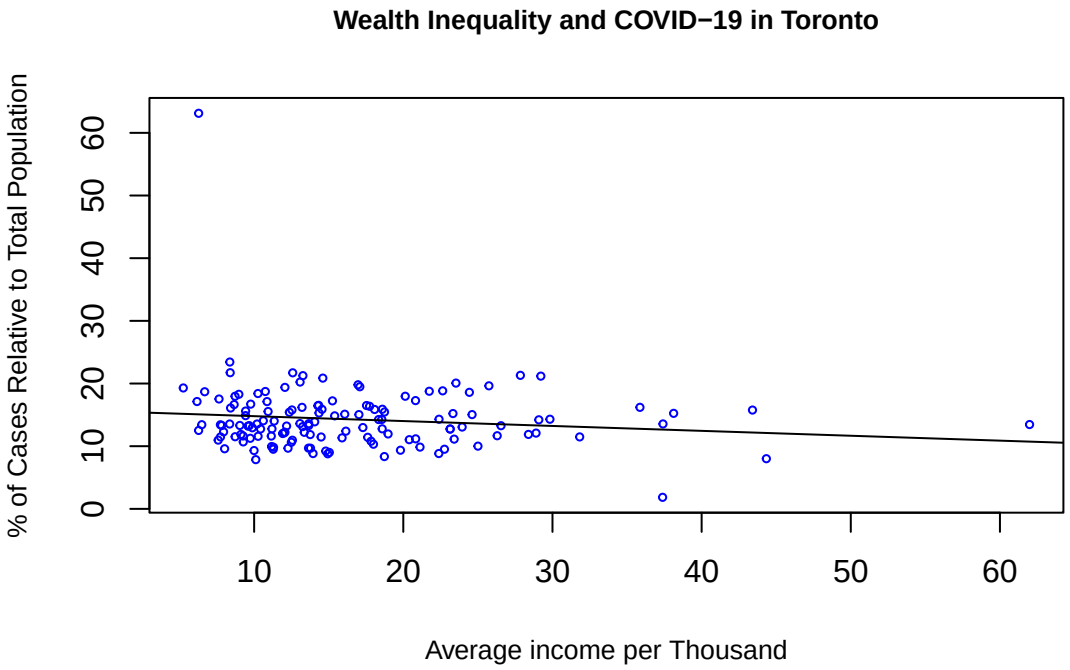
```
## # A tibble: 6 x 3
##   `Source of Infection`      N Proportion
##   <chr>                  <int>      <dbl>
## 1 Close Contact          20665     0.0523
## 2 Community              80781     0.204
## 3 Household Contact      42565     0.108
## 4 No Information        196706     0.498
## 5 Outbreak               49763     0.126
## 6 Travel                  4751     0.0120
```

Average age of infection: since the ages of patients are released by age range, calculating the average age of infection can be done using a confidence interval. A random sample of size 1000 will be taken from the dataset to reduce the running time of the bootstrap function. Assuming that patients within a certain age range (ex. 50-59 years) are distributed uniformly, i.e. $X_{a-b} \sim Uniform(a, b)$, a random age will be assigned to each patient accordingly using `runif`. For the age range of 90+ years, an exponential distribution will be used instead($X_{90+} \sim exp(1) + 90$). A bootstrap confidence interval can then be established using the random ages, which will accurately represent the true random age.
Mean age of patient:

```
##     2.5%    97.5%
## 39.65776 42.47961
```

The relationship between income and number of cases can be determined by fitting a linear model between the average income by neighbourhood and the proportion of cases to total population.
The test hypothesis of the linear model is: $H_0 : Income \perp \frac{cases}{population}$
while the alternative hypothesis is: $H_a : Income \not\perp \frac{cases}{population}$

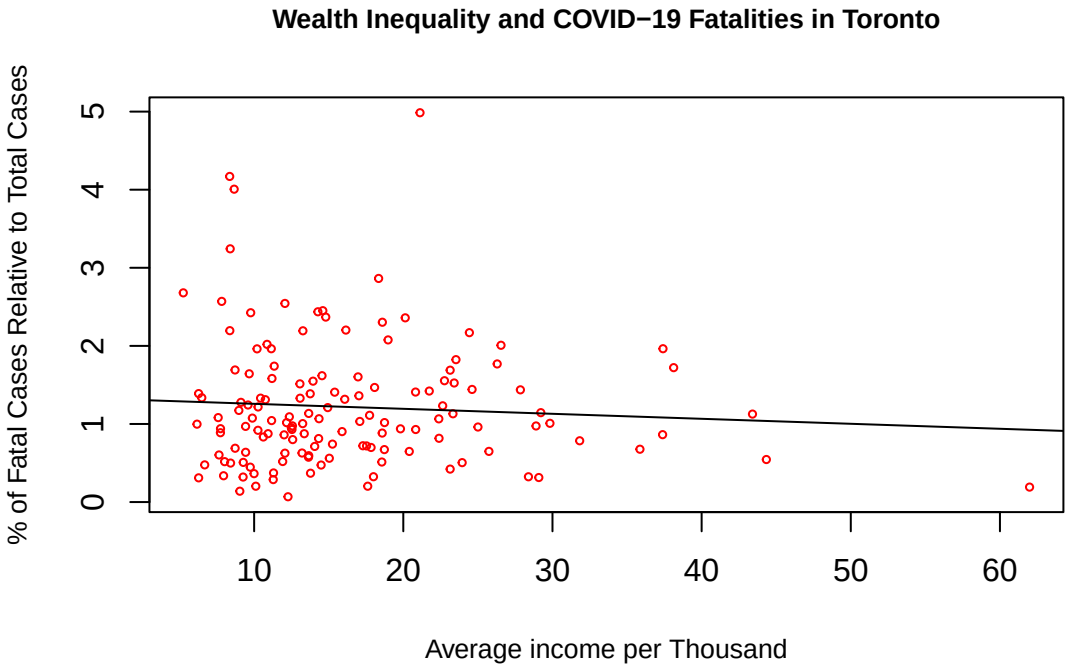
```
##
## Call:
## lm(formula = data_neighbour$Prop_cases_pop ~ average_income_thousands)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.810  -2.976  -1.028   2.043  48.035
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                15.57416    0.98460  15.818   <2e-16 ***
## average_income_thousands -0.07813    0.05304  -1.473    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.468 on 138 degrees of freedom
## Multiple R-squared:  0.01548,    Adjusted R-squared:  0.008345
## F-statistic:  2.17 on 1 and 138 DF,  p-value: 0.143
```



The intercept coefficient of 15 implies that when average income is 0, the proportion of cases to neighbourhood population is 15%. The average_income coefficient of -0.07 reveals that the rate of cases drops by 7% per $1000 in additional average income.
A p-value of 0.159 and R-squared value of 0.014 imply that the correlation between average income and rate of cases by neighbourhood is extremely weak. There is insufficient evidence to reject the null hypothesis of income being unrelated to the probability of contracting COVID-19.

Similarly, the relationship between average income and chance of fatality per case can be fitted with a linear model using the following hypotheses: $H_0 : Income \perp \frac{Fatalities}{TotalCases}$
$H_a : Income \not\perp \frac{Fatalities}{TotalCases}$

```
##
## Call:
## lm(formula = data_neighbour_fatal ~ average_income_thousands)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1756 -0.5224 -0.1841  0.3373  3.7986
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.322105   0.144628   9.141 7.08e-16 ***
## average_income_thousands -0.006396   0.007791  -0.821    0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8032 on 138 degrees of freedom
## Multiple R-squared:  0.00486,    Adjusted R-squared:  -0.002351
## F-statistic: 0.674 on 1 and 138 DF,  p-value: 0.4131
```

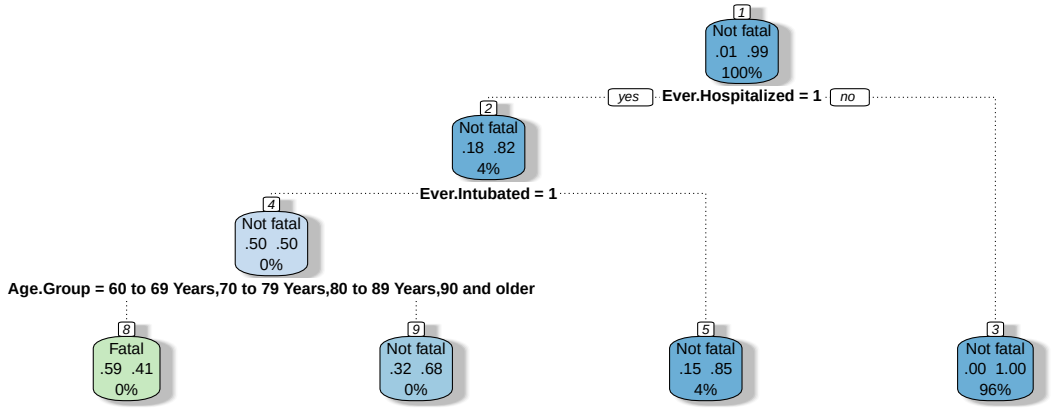


An intercept coefficient of 1.32 implies that when the average income is $0, the fatality of cases is approximately 1.32%. The average_income_thousands coefficient of -0.006 implies that the lethality of COVID-19 drops by around 0.6% per additional thousand dollars of average income.
The correlation is even lower than before, signalling that there is almost no relationship between fatality rate and average income of the patient. Likewise, a p-value of almost 0.5 does not provide sufficient evidence to reject the null hypothesis of fatality rate being unrelated to income.

Finally, a 4-fold cross validation technique will be implemented to develop a classifier. The features used for the classifier will be:

- Age Group
- Source of Infection
- Client Gender
- Ever Hospitalized/in ICU/Intubated

It is likely that source of infection and gender may become statistically irrelevant as the information gained from these features is very small.



Rattle 2023−Apr−06 09:42:49 likev

```
## [1] "4-fold cross-validation results: "
## [1] 0.9880169 0.9877081 0.9885163 0.9877662
## [1] "Accuracy of decision tree classifier: "
## [1] 0.9880019
```

Conclusion

COVID-19 has had detrimental effects in the city of Toronto. While the overall fatality rate of the disease is lower than its predecessor SARS, its overwhelming prevalence has resulted in many more deaths. COVID is mostly benign, since under 5% of confirmed or suspected infections require hospitalization. The number of infections between women and men was approximately equal in 2020 and 2021, but more women were infected than men in 2022. The mean age of infection in Toronto is likely somewhere between 40 and 43, and a weak negative correlation between average income and total cases could be found - but there was no evidence to suggest fatalities were linked to income. Finally, decision tree analysis reveals that intubated patients over the age of 60 are more likely to die, while intubated patients under the age of 60 were more likely to survive.

```r
# Initial setup
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width = 6, fig.height = 4)
library(tidyverse, warn.conflicts = FALSE)
library(lubridate, warn.conflicts = FALSE)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
library(rattle, warn.conflicts = FALSE)
library(rpart, warn.conflicts = FALSE)
data <- read_csv("COVID19 cases.csv")

# Basic information (Client cases by Gender/Outcome)
data <- subset(data, select=-c(`Outbreak Associated`))
rows = nrow(data)
data <- data %>% mutate(`Client Gender` = ifelse(
  `Client Gender` %in% c('NON-BINARY', 'TRANS MAN',
                         'TRANS WOMAN', 'TRANSGENDER', 'UNKNOWN',
                         'NOT LISTED, PLEASE SPECIFY'), 'OTHER', `Client Gender`))
data <- data %>% mutate(`Source of Infection` = ifelse(
  `Source of Infection` %in% c('Pending'), 'No Information', `Source of Infection`))
data <- data %>% mutate(`Source of Infection` = ifelse(
  `Source of Infection` %in% c('Outbreaks, Congregate Settings',
                               'Outbreaks, Healthcare Institutions',
                               'Outbreaks, Other Settings'),
  'Outbreak', `Source of Infection`))
data <- data %>% mutate(`Ever Hospitalized` = (`Ever Hospitalized`=="Yes")) %>%
  mutate(`Ever in ICU` = (`Ever in ICU`=="Yes")) %>%
  mutate(`Ever Intubated` = (`Ever Intubated`=="Yes"))
data_table <- data %>% group_by(Year=floor_date(`Reported Date`, "year")) %>%
  mutate(Year = format(Year, "%Y"))
ggplot(data_table, aes(x=Year, fill=`Client Gender`)) + geom_bar(position = "stack") +
  scale_fill_manual(values=c('coral4','steelblue','violet')) +
  labs(title = "Cases by Gender", y="Total Cases")
ggplot(data_table, aes(x=Year, fill=Outcome)) + geom_bar(position = "fill") +
  scale_fill_manual(values=c('cyan4','rosybrown','greenyellow')) +
  labs(title = "Cases by Outcome", y="Proportion of Cases")

# Rate of Hospitalization
patients_hosp <- data %>% group_by(`Ever Hospitalized`,
                                   `Ever in ICU`, `Ever Intubated`) %>%
  summarize(N = n(), Proportion=n()/rows)
patients_hosp

# Source of transmission
patients_source <- data %>% group_by(`Source of Infection`) %>%
  summarize(N = n(), Proportion=n()/rows)
patients_source

# Mean age
age <- sample_n(subset(data, select=c(`Age Group`)), 1000)
age <- na.omit(age)
rand_age = function(name){
  if(str_equal(name[1], '19 and younger')){
```

```
    return(runif(1, min=0, max=19)[1])
  }
  if(str_equal(name[1], '90 and older')){
    return(rexp(1)[1]+90.0)
  }
  a = as.numeric(substr(name[1], 1, 2))[1]
  b = as.numeric(substr(name[1], 7, 8))[1]
  return(runif(1, min=a, max=b))
}
random_age <- apply(age, 1, rand_age)
boot = function(){
  boot_sam = sample(random_age, size=1000, replace=TRUE)
  return(mean(boot_sam))
}
boot_mean <- replicate(1000, boot())
quantile(boot_mean, c(0.025, 0.975))

# Regression: average income and COVID-19 cases
data_neighbour <- data %>% group_by(`Neighbourhood Name`) %>%
  summarize(N = n()) %>% head(n=140)

neighbour_prof <- read_csv("neighbourhood-profiles-2016-140-model.csv")
average_income <- neighbour_prof[945,] %>% t() %>%
  tail(n=140)
average_income_thousands <- as.numeric(gsub(",", "", average_income))/1000
total_pop <- neighbour_prof[3,] %>% t() %>% tail(n=140)
total_pop <- as.numeric(gsub(",", "", total_pop))
data_neighbour <- data_neighbour %>% mutate(Prop_cases_pop = N/total_pop*100)
lin_model <- lm(data_neighbour$Prop_cases_pop~average_income_thousands)
summary(lin_model)
plot(average_income_thousands, data_neighbour$Prop_cases_pop, col="blue", pch=1, cex=0.5,
     main="Wealth Inequality and COVID-19 in Toronto", cex.main=0.8,
     xlab="Average income per Thousand",
     ylab="% of Cases Relative to Total Population", cex.lab=0.8)
abline(lin_model)

# Regression: average income and fatal cases
data_neighbour_fatal <- data  %>% filter(str_equal(`Outcome`, "FATAL")) %>%
  group_by(`Neighbourhood Name`) %>% summarize(N = n()) %>% head(n=140)
data_neighbour_fatal <- data_neighbour_fatal$N/data_neighbour$N * 100
lin_model_fatal <- lm(data_neighbour_fatal~average_income_thousands)
summary(lin_model_fatal)
plot(average_income_thousands, data_neighbour_fatal, col="red", pch=1, cex=0.5,
     main="Wealth Inequality and COVID-19 Fatalities in Toronto", cex.main=0.8,
     xlab="Average income per Thousand",
     ylab="% of Fatal Cases Relative to Total Cases", cex.lab=0.8)
abline(lin_model_fatal)

# Classifier: Fatal cases
data <- data %>% mutate(`Outcome` = ifelse(str_equal(`Outcome`, "FATAL"), "Fatal", "Not fatal")) %>%
  mutate(group_ind = sample(c(1,2,3,4), size=nrow(data), replace=T)) %>% na.omit()
names(data) <- make.names(names(data))
accuracy_vec <- vector()
```

```
for(i in 1:4){
  data_train <- data %>% filter(group_ind != i)
  Dec_tree <- rpart(`Outcome` ~ `Age.Group` +
                    `Source.of.Infection` + `Client.Gender` +
                    `Ever.Hospitalized` + `Ever.in.ICU` + `Ever.Intubated`,
                    data=data_train, method = "class")
  data_test <- data %>% filter(group_ind == i)
  result <- predict(Dec_tree, data_test, type="class")
  conf_mat <- table(result, data_test$`Outcome`)
  accuracy_vec[i] <- sum(diag(conf_mat))/sum(conf_mat)
  if(i == 1){
    fancyRpartPlot(Dec_tree)
  }
}
print("4-fold cross-validation results: ")
print(accuracy_vec)
print("Accuracy of decision tree classifier: ")
print(mean(accuracy_vec))
```