Problem Set 1 - Q1

Analysis on COVID-19 cases in Toronto

Guemin Kim (1005280946)

October 1, 2020

Question 1

Part a

```
# install.packages("opendatatoronto")
library(opendatatoronto)
library(tidyverse)
library(visdat)
library(skimr)

# get package
covid19_packages <- show_package("64b54586-6180-4485-83eb-81e8fae3b8fe")

#get all resources for this package
covid19_resources <- covid19_packages %>% list_package_resources()

# download the resource
covid19_data <- covid19_resources%>%get_resource()
```

The data set I chose is about the COVID-19 cases in Toronto. According to the Toronto Open Data Portal, this data contains geographic, demographic, and severity of illness information for all reported and confirmed individuals within the city by Toronto Public Health.

This is what our data set looks like:

```
head(covid19_data)
```

```
## # A tibble: 6 x 18
##    `_id` Assigned_ID `Outbreak Assoc~ `Age Group` `Neighbourhood ~ FSA
##    <int>       <int> <chr>            <chr>       <chr>            <chr>
## 1 200103           1 Sporadic         50 to 59 Y~ Willowdale East  M2N
## 2 200104           2 Sporadic         50 to 59 Y~ Willowdale East  M2N
## 3 200105           3 Sporadic         20 to 29 Y~ Parkwoods-Donal~ M3A
## 4 200106           4 Sporadic         60 to 69 Y~ Church-Yonge Co~ M4W
## 5 200107           5 Sporadic         60 to 69 Y~ Church-Yonge Co~ M4W
## 6 200108           6 Sporadic         50 to 59 Y~ Newtonbrook West M2R
## # ... with 12 more variables: `Source of Infection` <chr>,
## #   Classification <chr>, `Episode Date` <chr>, `Reported Date` <chr>, `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>
```

Ever since the coronavirus pandemic broke out, day to day lives of people around the world have greatly affected, in both direct and indirect ways. COVID-19 has not only changed people's lives, business, and many different components of our society, but it also reminded us of the importance of understanding and caring our community. Therefore, I chose this data set to deeply examine and analyze how this ongoing COVID-19 crisis has affected our community in Toronto.

What's interesting about this data set is that it is updated on a weekly basis.
In other words, if we want to analyze anything about the population of people infected by COVID-19 in Toronto, this data set provides the closest list of individuals we can get as having a list of all individuals in the population.
However, on the other hand, any analysis or conclusions made on the data set before the latest update will be considered out-of-date since the data set that was used to analyze could have missed some important individuals who are added after the update. Therefore, as the data is refreshed and overwritten every week, we need to check if our analysis on the previous week still holds for the updated data set.

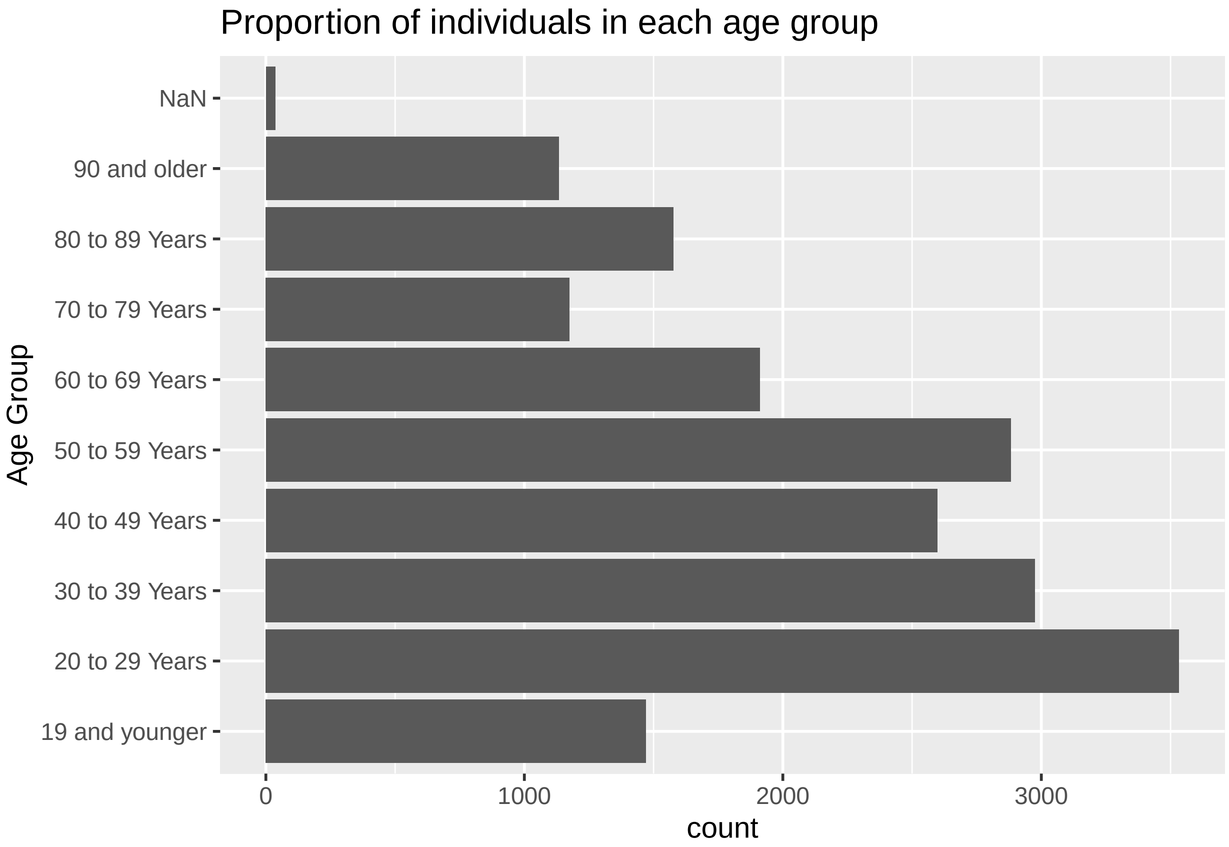
Another characteristic we found is that the majority of variables in the data set are categorical variables. However, a drawback of having categorical data is that, the kinds of statistical analysis that can be performed on this type of data is limited, and hence we may be required to create new numerical or dummy variables in order to perform any summary statistics from the data.

Part b

While exploring the data set, some interesting facts about age groups and outcomes were found.

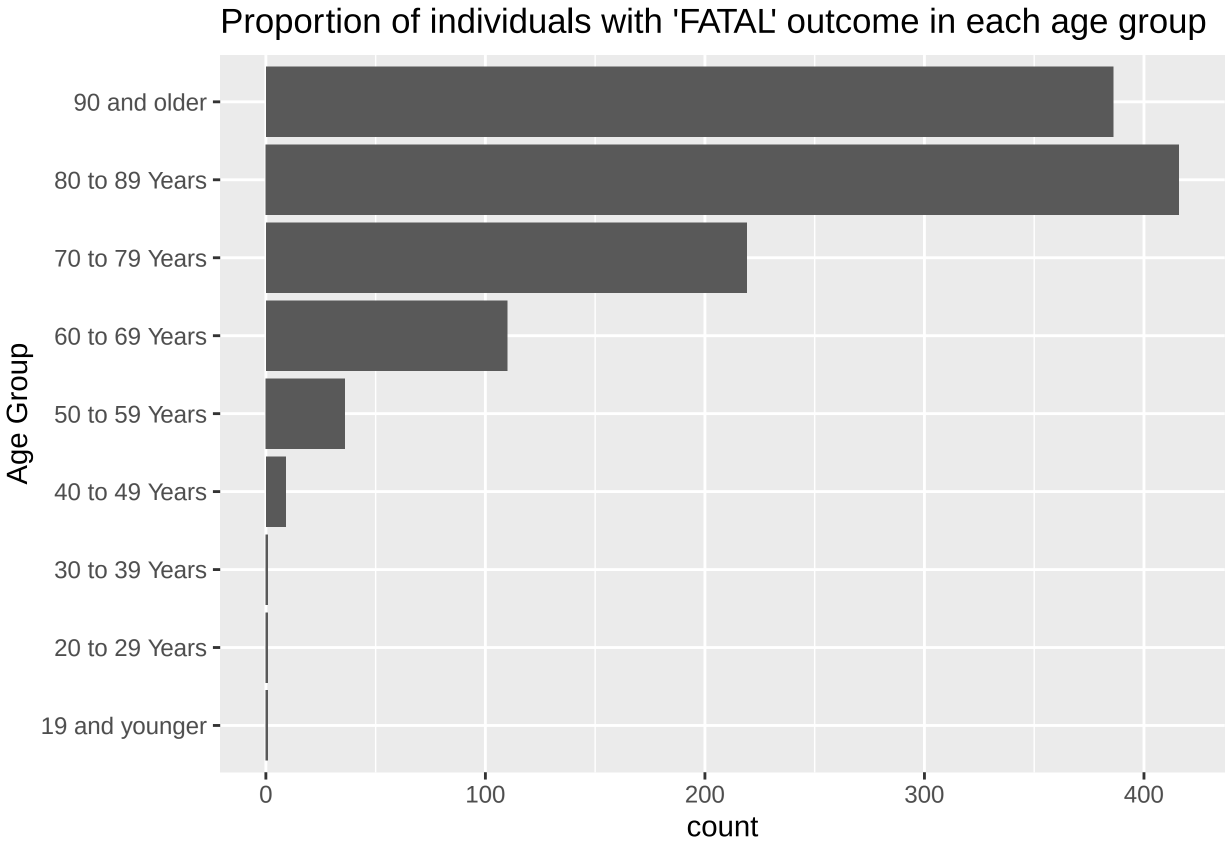
```r
age_group<- covid19_data %>%
  ggplot(aes(x=`Age Group`)) +
  geom_bar() +
  coord_flip() +
  ggtitle("Proportion of individuals in each age group")

age_group
```



As it is shown in this plot, quite a number of infected individuals in Toronto were concentrated in the age group from 20's to 50's. Among all of the age groups, '20 to 29 years' had the greatest number of infected people, and '90 and older' age group had the least.

```
outcome_by_age <- covid19_data %>%
  filter(`Outcome`=="FATAL") %>%
  ggplot(aes(x=`Age Group`)) +
  geom_bar() + coord_flip() +
  ggtitle("Proportion of individuals with 'FATAL' outcome in each age group")
outcome_by_age
```



On the other hand, if we look at this plot which presents the number of individual with “Fatal” outcome in each age group, there are striking disparities in the number of ‘fatal’ individuals between elders and young or middle-aged adults.

*Note: Here, ‘fatal’ individuals indicate people in the cases where a fatal outcome is reported.

Together, these findings suggest that, although the majority of the infected people in Toronto are young or middle-age adults in age groups from ‘19 and younger’ to ‘50 - 59 years’, elders (or people older than 60) are at higher risk for severe illness from COVID-19.

Part c

Since we found out that the outcome of coronavirus is much critical for older adults from the plots in Part b, now we want to see if the coronavirus is actually less fatal for younger adults.

To be more specific, we will test if the proportion of the 'resolved' patients in younger age groups is equal to the proportion of the 'resolved' patients in older age groups with hypothesis testing to see if COVID-19 affects people in both groups equally .

- Note: Here, 'resolved' patients indicate those who are reported as recovered or whose reported date of infection is more than 14 days but not currently hospitalized.
  Also, We will classify individuals in the age groups from '19 and younger' to '50 - 59 years' as "Younger adults" and '60 - 69 years' to '90 and older' as "Older Adults" for brevity.

Hypothesis Testing:

$$H_0 : P_Y = P_O$$

where P_Y represents the total proportions of the 'resolved' patients who are classified as "Younger Adult", and P_O indicates the total proportion of the 'resolved' patients who are classified as "Older Adult". Alternatively, our H_1 is $P_Y > P_O$.

Before starting the hypothesis testing, we first need to tidy our data set.

```r
# create a new data to use for the hypothesis testing

# Here, we created a new column called "New Age Group"
# and classified individuals in the original age group
# to either "Younger Adult" or "Older Adult" using mutate function.
# Then, we  select only the "Age Group", "New Age Group" and the corresponding "Outcome"
# to be shown using select function.

data <- covid19_data %>%
  mutate(`New Age Group` = ifelse(`Age Group` %in%
                                    c("19 and younger","20 to 29 Years",
                                      "30 to 39 Years", "40 to 49 Years",
                                      "50 to 59 Years"),
                                  "Younger Adult", "Older Adult")) %>%
  select(`Age Group`,`New Age Group`, `Outcome`)
```

With this newly formatted data set, We will begin our hypothesis testing:

```r
# number of patients in Younger Adult age group
num_younger <- nrow(filter(data, `New Age Group` == "Younger Adult"))

# number of patients in Older Adult age group
num_older <- nrow(filter(data, `New Age Group` == "Older Adult"))

# number of 'RESOLVED' patients in Younger Adult age group
num_younger_res <- nrow(data %>% filter(`New Age Group` == "Younger Adult") %>%
                          filter(`Outcome` == "RESOLVED"))

# number of 'RESOLVED' patients in Older Adult age group
num_older_res <- nrow(data %>% filter(`New Age Group` == "Older Adult") %>%
                        filter(`Outcome` == "RESOLVED"))

# test whether the proportion of 'RESOLVED' patients in Younger Adult age group
# is equal to those of Older Adult age group;
# alternative hypothesis tests whether the proportion of 'RESOLVED' patients
# in Younger Adult age group is greater than those of Older Adult age group
result <- prop.test(x = c(num_younger_res, num_older_res),
                    n = c(num_younger, num_older), alternative = "greater")

result
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(num_younger_res, num_older_res) out of c(num_younger, num_older)
## X-squared = 335.49, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.09572725 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.8690839 0.7629007
```

As we can see from the result, the p-value of the test with $H_0 : P_Y = P_O$ is 3.067745×10^{-75}, and this is much less than the significance level alpha = 0.05.
Hence, the null hypothesis ($P_Y = P_O$) is rejected, and we can conclude that our data supports the alternative hypothesis which states that the proportion of 'RESOLVED' patients in Younger Adult age group is greater than the proportion of 'RESOLVED' patients in Older Adult age group.

Part d

Throughout the analysis, we observed the current state of the coronavirus pandemic in the city of Toronto, and how it influenced us and our neighbors using the data set provided by Toronto Open Data Portal.

To be more specific, we examined how the proportion of infected individuals differ within each age group, and how the proportion of infected individuals with 'fatal' outcome differ within each age group using plots in part b.
Here, we found out that younger age groups from 20's to 50's constitute more than half of all reported cases in Toronto, however, individuals who are older than 60 were found to be more vulnerable to severe illness from COVID-19 than those in the younger age groups.

Furthermore, in order to collect more evidence to support the finding that the severity of illness differs between older and younger age groups, we conducted a hypothesis testing in part c. However, unlike part b where we examined the individuals with 'fatal' outcomes, our target in part c was those with 'resolved' outcomes.
In the hypothesis testing, we tested whether the proportions of infected individuals with 'resolved' outcome are the same for younger and older age groups; the alternative hypothesis was that the proportion of 'resolved' patients in younger age groups was greater than the proportion of 'resolved' patients in older age groups.
From the result of the testing, we obtained the p-value that was much less than the significance level, and hence, we rejected the null hypothesis that the proportions of 'resolved' patients of younger and older age groups are the same.
In other words, from the result, we found out that our data favors the alternative hypothesis which states the proportion of 'resolved' patients in younger age group is greater than the proportion of 'resolved' patients in older age group.

Weaknesses and next steps

Since we haven't gone over every single variables provided in the data, there could be some important information that we missed from those variables. For example, one of the important variables that we didn't go over was 'Source of Infection'.

```
# number of infected cases in each source of infection
count_source<- covid19_data %>%
  group_by(`Source of Infection`) %>%
  tally()

count_source
```

```
## # A tibble: 8 x 2
##   `Source of Infection`         n
##   <chr>                     <int>
## 1 Close contact              6943
## 2 Community                  2652
## 3 Healthcare                 1186
## 4 Institutional               367
## 5 N/A - Outbreak associated  6111
## 6 Pending                      54
## 7 Travel                      885
## 8 Unknown/Missing            1094
```

If we look into the table that shows the number of each source of infection, there are 622 individuals with unknown infection routes. This is surprising, because it means that there are some possibilities that people who infected these 622 patients are still in our community without being identified. If this is the case, then these unidentified infected individuals might infect more people and put our community at greater risk. Likewise, there could be more variables that are very significant, but not analyzed.

Another weakness in our analysis is that we need to check if our findings still make sense for a newly updated data on a weekly basis. Going over the entire analysis process with the updated data is time consuming but very necessary, because we are still in the middle of the COVID-19 crisis, and depending on how we treat and analyze our data, the consequence it will bring on our community could change significantly.

Hence, as the next steps, we must put a lot of efforts to keep up with the updates in our data, as well as identify more interesting findings to help us understand how COVID-19 affects our community in Toronto and to protect it.

Part e

Bibliography:

1. Open Data Dataset. (2020, September 23). Doors Open Toronto – City of Toronto. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/
2. ifelse function | R Documentation. (n.d.). Www.Rdocumentation.Org. Retrieved September 30, 2020, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/ifelse
3. prop.test function | R Documentation. (n.d.). Www.Rdocumentation.Org. Retrieved September 30, 2020, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prop.test
4. tally function | R Documentation. (n.d.). Www.Rdocumentation.Org. Retrieved September 30, 2020, from https://www.rdocumentation.org/packages/dplyr/versions/0.5.0/topics/tally