Course Project

Aly Abdelwahed

3/17/2022

1. Introduction

The dataset that I will be using for this project is the covid-19 dataset that's available on the **Open Toronto Website** website, specifically **COVID 19 CASES**, which can be downloaded from this **link** and can and I will conduct the statistical analysis as follows

Let's first read the data (Appendix A) and start by looking at some of the data's features (Appendix B)

```
## Rows: 298,912
## Columns: 18
## $ X_id                   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ Assigned_ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ Outbreak.Associated    <chr> "Sporadic", "Sporadic", "Sporadic", "Sporadic",~
## $ Age.Group              <chr> "50 to 59 Years", "50 to 59 Years", "20 to 29 Y~
## $ Neighbourhood.Name     <chr> "Willowdale East", "Willowdale East", "Parkwood~
## $ FSA                    <chr> "M2N", "M2N", "M3A", "M4W", "M4W", "M2R", "M1V"~
## $ Source.of.Infection    <chr> "Travel", "Travel", "Travel", "Travel", "Travel~
## $ Classification         <chr> "CONFIRMED", "CONFIRMED", "CONFIRMED", "CONFIRM~
## $ Episode.Date           <chr> "2020-01-22", "2020-01-21", "2020-02-05", "2020~
## $ Reported.Date          <chr> "2020-01-23", "2020-01-23", "2020-02-21", "2020~
## $ Client.Gender          <chr> "FEMALE", "MALE", "FEMALE", "FEMALE", "MALE", "~
## $ Outcome                <chr> "RESOLVED", "RESOLVED", "RESOLVED", "RESOLVED",~
## $ Currently.Hospitalized <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Currently.in.ICU       <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Currently.Intubated    <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Ever.Hospitalized      <chr> "No", "Yes", "No", "No", "No", "No", "No", "Yes~
## $ Ever.in.ICU            <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Ever.Intubated         <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
```

i. Description of the variables and the data

According to the Open Toronto website,

- **_id**: It is simply a unique row identifier
- **Assigned_ID:** This is a unique ID number that is assigned to each covid-19 cases by the Toronto Public Heath. This variable is created for the purpose of posting this data to the Open Data website to allow for tracking of specific cases
- **Outbreak Associated:** It indicates if the case is associated with an outbreak from Toronto healthcare institutions and healthcare settings (e.g. long-term care homes, retirement homes, hospitals, etc.) and other Toronto congregate settings (such as homeless shelters) or not
- **Age Group:** The Age group that the person belongs to at time of illness. Age groups (in years): <=19 years old, 20-29 years old, 30-39 years old, 40-49 years old, 50-59 years old, 60-69 years old, 70-79 years old, 80-89 years old, 90+ years old, unknown (blank)
- **Neighbourhood Name:** According to the covid-19 dataset on the Open Data website, Toronto is divided into 140 geographically distinct neighborhoods that were established to help government and community agencies with local planning by providing socio-economic data for a meaningful geographic area

- **FSA:** Stands for **Forward Sortation Area** (i.e. first three characters of postal code) based on the case's primary home address
- **Source of Infection:** The most likely way that the COVID-19 cases have acquired their COVID-19 infection (e.g. Household contact, Close contact with a case, Outbreaks, Travel, Community, No information, etc.)
- **Classification:** Determines whether the positive COVID-19 cases are confirmed or probable, according to the standard criteria
- **Episode Date:** This variable best estimates when the COVID-19 infection was acquired
- **Reported Date:** The date on which the case was reported to Toronto Public Health
- **Client Gender:** The self-reported gender by the clients (e.g. Male, Female, etc.)
- **Outcome:** This indicates the outcome of the COVID-19 cases (FATAL, RESOLVED, ACTIVE)
- **Currently Hospitalized:** Patients that are currently admitted to hospital with no discharge date reported
- **Currently in ICU:** Patients that are currently admitted to the intensive care unit (ICU) with no discharge date reported
- **Currently Intubated:** Patients that were intubated from their COVID-19 infection
- **Ever Hospitalized:** Patients that were hospitalized from their COVID-19 infection. This includes the patients that are currently hospitalized and those that have been discharged from the hospital or are deceased
- **Ever in ICU:** Patients that were admitted to the intensive care unit (ICU) from their COVID-19 infection. Again, this includes cases that are currently in ICU and those that have been discharged from the hospital or are deceased
- **Ever Intubated:** Patients that were intubated from their COVID-19 infection. This includes the patients that are currently intubated and those that have been discharged from the hospital or deceased

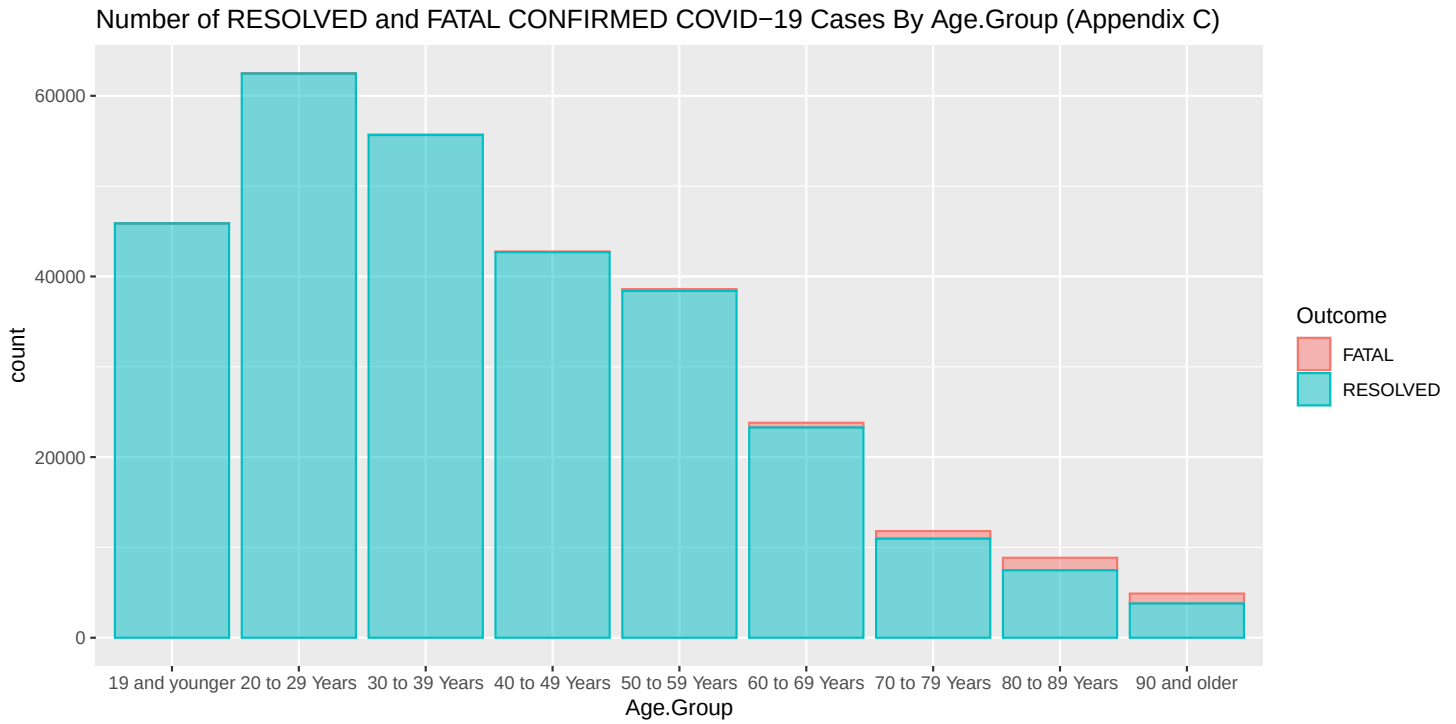
ii. A little background about the data (who collected it in what context)

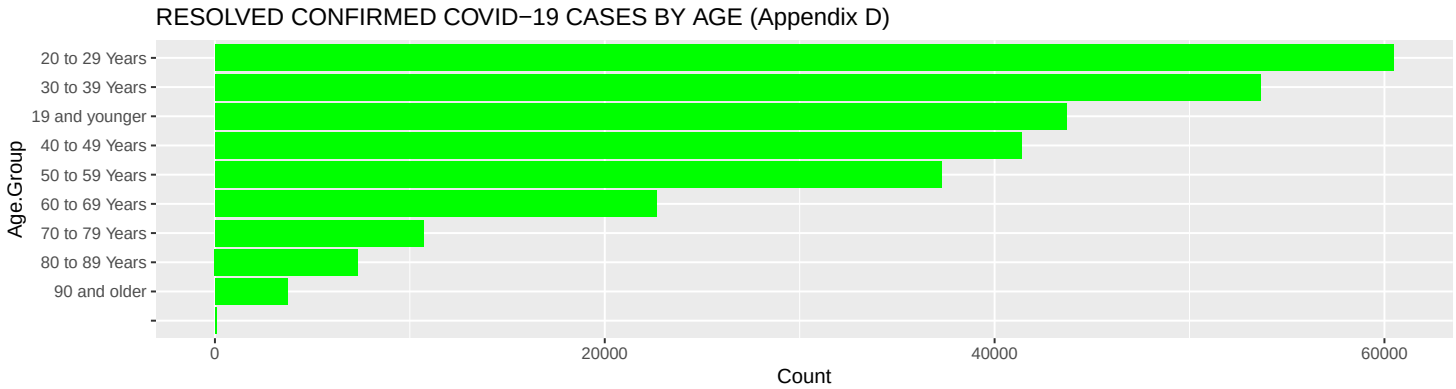
The Toronto Public Health is handling an ongoing COVID-19 outbreak that is in the context of an evoloving global pandemic. This dataset contains numerous types of information about all the confirmed or probable COVID-19 cases that were reported to and managed by the Toronto Public Health since the first case that was reported in January 2020. These types of information vary from demographic and geographic information to severity, start date, fatality and more. The data are extracted from the provincial Case & Contact Management System (CCM)

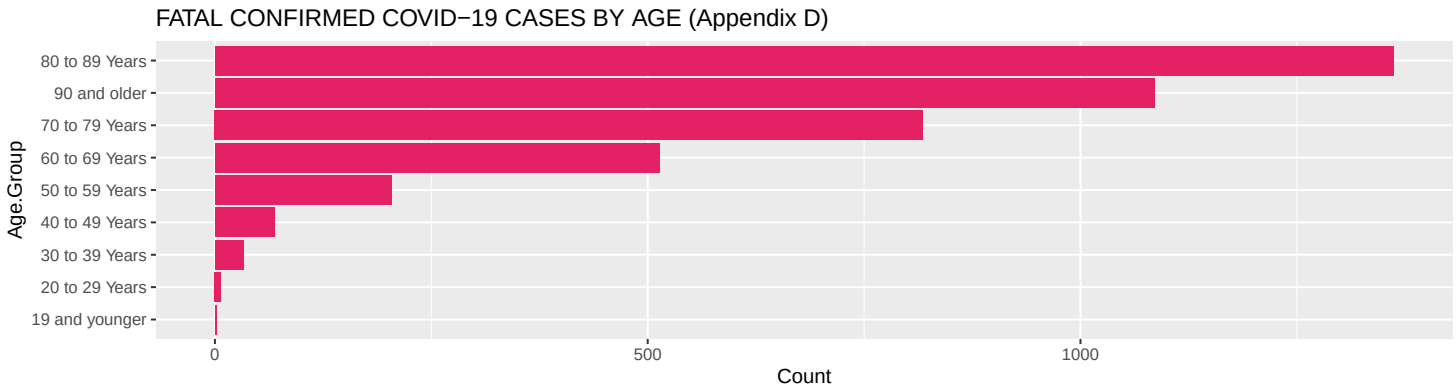
iii. What is the overall research question that you are trying to answer in your report

Are people in the higher age groups (e.g. 70 to 79 Years, 80 to 89 Years, 90 and older) who have been confirmed of contracting the COVID-19 infection in a higher risk of dying compared to those in the lower age groups (e.g. 20-29, 30-39, 40-49)?

2. Graphical Summaries







Comments: As seen from the plots above we can see the following:

- People who have been confirmed of contracting COVID-19 in age groups from the age group of "60 to 69 Years" and above (i.e. "60 to 69 Years", "70 to 79 Years", "80 to 89 Years", "90 and older") have much higher proportions of COVID-19 fatal cases

3. Summary Tables

i. Amount of Confirmed Cases Resolved By Grouped By Age (Appendix E)

```
##        Age.Group NUMBER_OF_RESOLVED_CASES PROPORTION_OF_AGE_GROUP
## 1 50 to 59 Years                    37290              0.9554679
## 2 40 to 49 Years                    41380              0.9549304
## 3 20 to 29 Years                    60480              0.9544400
## 4 30 to 39 Years                    53655              0.9489070
## 5 60 to 69 Years                    22648              0.9408832
## 6 19 and younger                    43679              0.9406482
## 7 70 to 79 Years                    10698              0.8945564
## 8 80 to 89 Years                     7340              0.8190136
## 9   90 and older                     3733              0.7558210
```

Comments:

- The people that have been confirmed of contracting COVID-19 that have the lowest recovery proportion (i.e. the people that have the lowest recovery rate with respect to their age group) are the people that belong in the highest age group (90 and older)
- The people that have been confirmed of contracting COVID-19 that have the highest recovery proportion with respect to their age group (i.e. the people that have the highest recovery rate from their respective age group) are the people that belong in the "20-29 Years" age group
- As the age group gets larger than 60-69 years old, the recovery rate with respect to their age group gradually decreases. In other words, if one person is part of the "70 to 79 Years" age group, his/her chances of recovery decrease considerably, and they keep gradually decreasing from their on onwards as that person grows older and joins the higher age groups ("80-89 Years" and "90 and Older Years")

ii. Amount of Confirmed Cases That Were Fatal Grouped By Age (Appendix F)

```
##        Age.Group NUMBER_OF_FATAL_CASES PROPORTION_OF_AGE_GROUP
## 1   90 and older                 1086            0.2233189389
## 2 80 to 89 Years                 1362            0.1547199818
## 3 70 to 79 Years                  818            0.0701664093
## 4 60 to 69 Years                  514            0.0219489282
## 5 50 to 59 Years                  204            0.0053847169
## 6 40 to 49 Years                   69            0.0016435615
## 7 30 to 39 Years                   33            0.0006057602
## 8 20 to 29 Years                    7            0.0001141590
## 9 19 and younger                    2            0.0000452284
```

Comments:

- The people that have been confirmed of contracting COVID-19 that belong in the largest age group (90 and older) have the highest fatal cases proportion (i.e. The people that are at least 90 years old have the highest fatal cases with respect to their age group)
- The people that have been confirmed of contracting COVID-19 that belong in the smallest age group (19 and younger) have the lowest fatal cases proportion (i.e. The people that are at most 19 years old have the lowest fatal cases with respect to their age group)

4. Hypothesis Test and Confidence Interval

i. "90 and Older" Vs "80 to 89 Years" Proportion Test of Fatal Cases (Appendix G)

Now we want to conduct the proportion test to see if the proportion is the same between the "90 and Older" sample and "80 to 89 Years" sample or if they are different. And if they are different, we want to estimate the difference between them

Therefore we want to test the following:

$$H_0 : \pi_{\text{90 and older}} - \pi_{\text{80 to 89 Years}} = 0 \text{ (Proportion Of Fatal Cases Is The Same Among Both Age Groups)}$$

$$H_a : \pi_{\text{90 and older}} - \pi_{\text{80 to 89 Years}} > 0 \text{ (Proportion Of Fatal Cases is Greater in the } \mathbf{90 \text{ and older}} \text{ Age Group)}$$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(number_of_success_90, number_of_success_80_89) out of c(number_of_trials_90, number_of_trials_80_89
## X-squared = 99.784, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05450797 0.08268994
## sample estimates:
##    prop 1    prop 2
## 0.2233189 0.1547200
```

Since the p-value is $1.6995245 \times 10^{-23}$ which is very small and certainly much less than 0.05 we reject the null hypothesis, that is, we reject the idea that the proportion of fatal cases is the same among both age groups and we now have very strong evidence that the proportion of fatal cases of patients that have been confirmed of contracting COVID-19 is higher among the **90 and older** age group than **80 to 89 Years** age group from the positive confidence interval that we received from the test and that the difference is big enough to be statistically significant.

Confidence Interval Interpretation:

The confidence interval that we received from the proportion test is: [0.054508, 0.0826899]

Since we put the **90 and older** inputs in the proportion test first and the confidence interval was positive (None of the values in the interval were negative). This means that we are 95% confident that the proportion of fatal cases of confirmed COVID-19 patients in the **90 and older** age group is higher than the proportion of fatal cases of confirmed COVID-19 patients in the **80 to 89 Years** by a percentage value that lies between 5.450797% and 8.2689944%.

ii. "90 and Older" Vs "70 to 79 Years" Proportion Test of Fatal Cases (Appendix H)

Now we want to conduct the proportion test to see if the proportion is the same between the "90 and Older" sample and "70 to 79 Years" sample or if they are different. And if they are different, we want to estimate the difference between them

Therefore we want to test the following:

$$H_0 : \pi_{\text{90 and older}} - \pi_{\text{70 to 79 Years}} = 0 \text{ (Proportion Of Fatal Cases Is The Same Among Both Age Groups)}$$

$$H_a : \pi_{\text{90 and older}} - \pi_{\text{70 to 79 Years}} > 0 \text{ (Proportion Of Fatal Cases is Greater in the } \mathbf{90 \text{ and older}} \text{ Age Group)}$$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(number_of_success_90, number_of_success_70_79) out of c(number_of_trials_90, number_of_trials_70_79
## X-squared = 787.88, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1404167 0.1658883
## sample estimates:
##     prop 1     prop 2
## 0.22331894 0.07016641
```

Since the p-value is $2.327046 \times 10^{-173}$ which is very small and certainly much less than 0.05 we reject the null hypothesis, that is, we reject the idea that the proportion of fatal cases is the same among both age groups and we now have very strong evidence that the proportion of fatal cases of patients that have been confirmed of contracting COVID-19 is higher among the **90 and older** age group than **70 to 79 Years** age group from the positive confidence interval that we received from the test and that the difference is big enough to be statistically significant.

Confidence Interval Interpretation:

The confidence interval that we received from the proportion test is: [0.1404167, 0.1658883]

Since we put the **90 and older** inputs in the proportion test first and the confidence interval was positive (None of the values in the interval were negative). This means that we are 95% confident that the proportion of fatal cases of confirmed COVID-19 patients in the **90 and older** age group is higher than the proportion of fatal cases of confirmed COVID-19 patients in the **70 to 79 Years** by a percentage value that lies between 14.0416711% and 16.5888348%.

iii. "80 to 89 Years" Vs "70 to 79 Years" Proportion Test of Fatal Cases (Appendix I)

Now we want to conduct the proportion test to see if the proportion is the same between the "80 to 89 Years" sample and "70 to 79 Years" sample or if they are different. And if they are different, we want to estimate the difference between them

Therefore we want to test the following:

$$H_0 : \pi_{\text{80 to 89 Years}} - \pi_{\text{70 to 79 Years}} = 0 \text{ (Proportion Of Fatal Cases Is The Same Among Both Age Groups)}$$

$$H_a : \pi_{\text{80 to 89 Years}} - \pi_{\text{70 to 79 Years}} > 0 \text{ (Proportion Of Fatal Cases is Greater in the } \mathbf{80 \text{ to } 89 \text{ Years}} \text{ Age Group)}$$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(number_of_success_80_89, number_of_success_70_79) out of c(number_of_trials_80_89, number_of_trials
## X-squared = 375.81, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07558997 0.09351718
## sample estimates:
##     prop 1     prop 2
## 0.15471998 0.07016641
```

Since the p-value is 1.018448×10^{-83} which is very small and certainly much less than 0.05 we reject the null hypothesis, that is, we reject the idea that the proportion of fatal cases is the same among both age groups and we now have very strong evidence that the proportion of fatal cases of patients that have been confirmed of contracting COVID-19 is higher among the **80 to 89 Years** age group than **70 to 79 Years** age group from the positive confidence interval that we received from the test and that the difference is big enough to be statistically significant.

Confidence Interval Interpretation:

The confidence interval that we received from the proportion test is: [0.07559, 0.0935172]

Since we put the **80 to 89 Years** inputs in the proportion test first and the confidence interval was positive (None of the values in the interval were negative). This means that we are 95% confident that the proportion of fatal cases of confirmed COVID-19 patients in the **80 to 89 Years** age group is higher than the proportion of fatal cases of confirmed COVID-19 patients in the **70 to 79 Years** by a percentage value that lies between 7.5589967% and 9.3517178%.

Final Comments On Hypothesis Test:

As seen from the previous three hypothesis tests, as the age group a person belongs into is grows, if he/she is confirmed of contracting COVID-19, then his/her chances of dying from COVID-19 are also gradually higher than that of the people that belong in the younger age groups as evidenced by how the proportion of fatal cases of confirmed COVID-19 patients was larger in the people that belong in the higher age groups in the 3 hypothesis tests that we conducted. Note that we would also see the same results every time we run a proportion test between any two combination of age groups.

5. Hypothesis Test Using Bootstrap Sampling

i. Test the true proportion of the FATAL COVID-19 cases in the "90 and older" age group (Appendix J)

$$H_0 : \pi_{\text{90 and older}} = 0.20$$
$$H_a : \pi_{\text{90 and older}} > 0.20$$

```
## [1] "Mean of the observed sample proportion:"
```

```
## [1] 0.2233189
```

```
## [1] "Test Stat:"
```

```
## [1] 3.908268
```

```
## [1] "P-value:"
```

```
## [1] 4.648005e-05
```

Since the p-value of the test statistic that we had is `pnorm(test_stat, lower.tail = FALSE)` which is smaller than 0.05 which means that we reject the null hypothesis H_0, that is, we reject the idea that the true proportion of the FATAL COVID-19 cases in the "90 and older" age group is 0.20. And now, we have strong evidence that true proportion of the true proportion of the FATAL COVID-19 cases in the "90 and older" age group is greater than 0.20

ii. Test the true proportion of the FATAL COVID-19 cases in the "80 to 89 Years" age group (Appendix K)

$$H_0 : \pi_{\text{80 to 89 Years}} = 0.20$$
$$H_a : \pi_{\text{80 to 89 Years}} < 0.20$$

```
## [1] "Mean of the observed sample proportion:"
```

```
## [1] 0.15472
```

```
## [1] "Test Stat:"
```

```
## [1] -11.81474
```

```
## [1] "P-value:"
```

```
## [1] 1.637826e-32
```

Since the p-value of the test statistic that we had is `pnorm(test_stat)` which is smaller than 0.05 which means that we reject the null hypothesis H_0, that is, we reject the idea that the true proportion of the FATAL COVID-19 cases in the "80 to 89 Years" age group is 0.20. And now, we have **very strong evidence** that true proportion of the true proportion of the FATAL COVID-19 cases in the "80 to 89 Years" age group is less than 0.20 which automatically means that it is less than the true proportion of the FATAL COVID-19 cases in the "90 and older" age group

iii. Test the true proportion of the FATAL COVID-19 cases in the "70 to 79 Years" age group (Appendix L)

$$H_0 : \pi_{70 \text{ to } 79 \text{ Years}} = 0.10$$
$$H_a : \pi_{70 \text{ to } 79 \text{ Years}} < 0.10$$

```
## [1] "Mean of the observed sample proportion:"
```

```
## [1] 0.07016641
```

```
## [1] "Test Stat:"
```

```
## [1] -12.54422
```

```
## [1] "P-value:"
```

```
## [1] 2.138056e-36
```

Since the p-value of the test statistic that we had is `pnorm(test_stat)` which is smaller than 0.05 which means that we reject the null hypothesis H_0, that is, we reject the idea that the true proportion of the FATAL COVID-19 cases in the "70 to 79 Years" age group is 0.20. And now, we have **very strong evidence** that true proportion of the true proportion of the FATAL COVID-19 cases in the "70 to 79 Years" age group is less than 0.20 which automatically means that it is less than the true proportion of the FATAL COVID-19 cases in the "80 to 89 Years" age group which is already less than the true proportion of the FATAL COVID-19 cases in the "90 and older group"

Final Comments On Bootstrap Proportion Test:

As seen from the previous three bootstrap tests, as the age group a person belongs into is grows, if he/she is confirmed of contracting COVID-19, then his/her chances of dying from COVID-19 are also gradually higher than the chances of people from a younger age group as evidenced by how the proportion of fatal cases of confirmed COVID-19 patients decrease as the age group that the patient belongs to decreases. Note that we would also see the same results every time we run a bootstrap proportion test between any two combination of age groups.

6. Regression Analysis

Logistic Regression (Appendix M)

```
##
## Call:
## glm(formula = Outcome_Dummy_Variable ~ Age.Group, family = binomial,
##     data = covid_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4839   0.0159   0.0347   0.1034   0.7073
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                10.0526     0.7070  14.219  < 2e-16 ***
## Age.Group20 to 29 Years    -1.0755     0.7905  -1.361 0.173640
## Age.Group30 to 39 Years    -2.6368     0.7275  -3.625 0.000289 ***
## Age.Group40 to 49 Years    -3.6543     0.7167  -5.099 3.42e-07 ***
## Age.Group50 to 59 Years    -4.8235     0.7104  -6.790 1.12e-11 ***
## Age.Group60 to 69 Years    -6.2336     0.7084  -8.800  < 2e-16 ***
## Age.Group70 to 79 Years    -7.4593     0.7079 -10.538  < 2e-16 ***
## Age.Group80 to 89 Years    -8.3455     0.7076 -11.795  < 2e-16 ***
## Age.Group90 and older      -8.7945     0.7078 -12.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43666  on 298637  degrees of freedom
## Residual deviance: 28375  on 298629  degrees of freedom
## AIC: 28393
##
## Number of Fisher Scoring iterations: 12
```

Interpretation Of Parameters:

- The log(odds) that a person who is in the **19 and younger** age group dies from COVID-19 is 10.0526184
- The log(odds) that a person who is in the **20 to 29 Years** age group dies from COVID-19 is -1.0754877
- The log(odds) that a person who is in the **30 to 39 Years** age group dies from COVID-19 is -2.636806
- The log(odds) that a person who is in the **40 to 49 Years** age group dies from COVID-19 is -3.6542777
- The log(odds) that a person who is in the **50 to 59 Years** age group dies from COVID-19 is -4.8234656
- The log(odds) that a person who is in the **60 to 69 Years** age group dies from COVID-19 is -6.2336103
- The log(odds) that a person who is in the **70 to 79 Years** age group dies from COVID-19 is -7.459321
- The log(odds) that a person who is in the **80 to 89 Years** age group dies from COVID-19 is -8.3454947
- The log(odds) that a person who is in the **90 and older** age group dies from COVID-19 is -8.794511

7. Cross Validation

K-Fold Cross Validation (Appendix N)

```
## [1] "c.index:"
```

```
##  [1] 0.9295924 0.9464317 0.9524792 0.9404979 0.9371945 0.9337376 0.9413041
##  [8] 0.9394017 0.9357840 0.9381510 0.9397300 0.9429329 0.9499355 0.9403970
## [15] 0.9417939 0.9312096 0.9401303 0.9464124 0.9419880 0.9272822
```

```
## [1] "Mean of c.index:"
```

```
## [1] 0.9398193
```

Since the average summary measure of our 20 different summary measures of the model performance is 0.9398193, this indicates that our model's predictive performance of how well someone is likely to die from COVID-19 given the age group he/she belong in is very good

8. Final Summary Of The Findings

The question that we were trying to answer was the following:

Are people in the higher age groups (e.g. 70 to 79 Years, 80 to 89 Years, 90 and older) who have been confirmed of contracting the COVID-19 infection in a higher risk of dying compared to those in the lower age groups (e.g. 20-29, 30-39, 40-49)?

And we can indeed say that yes, the older the COVID-19 patients are, the higher their risk of dying becomes. This was evidenced by the following:

- The proportion tests that we have conducted in Part 4 of this report where it was obvious that every time we have compared one age group to another using the proportion test (Of course, one age group will be older of the other age group), we always found that the higher age group always has the higher proportion of confirmed COVID-19 cases that were fatal
- The bootstrap proportion test that we have conducted in Part 5 of this report where it was obvious that every time the age group of the confirmed COVID-19 patients decreases and the proportion value of confirmed COVID-19 cases that were fatal which we conduct the test with decreases, the bootstrap proportion test proves that the real proportion value of confirmed COVID-19 cases that were fatal is even lower than the reduced value that we already conducted the bootstrap test with.
- The logistic regression model that we have created in part 6 of the report which clearly shows that every time the people who have been confirmed of contracting COVID-19's age group increases, then for every one unit increase in that age group, the log odds of dying from COVID-19 increase (e.g. the log odds of dying of COVID-19 for people in the **40 to 49 Years** age group increase by 3.6672618, whereas the log odds of dying of COVID-19 for people in the **80 to 89 Years** age group increase by 8.1974721)

As such, we have more than enough evidence to conclude that the people in the higher age groups (e.g. 70 to 79 Years, 80 to 89 Years, 90 and older) who have been confirmed of contracting the COVID-19 infection are indeed in a higher risk of dying compared to those in the lower age groups (e.g. 20-29, 30-39, 40-49)

9. Appendix

i. Appendix A

```
url <- "https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/64b54586-6180
-4485-83eb-81e8fae3b8fe/resource/fff4ee65-3527-43be-9a8a-cb9401377dbc/download/
COVID19%20cases.csv"
covid_data <- read.csv(url)
```

ii. Appendix B

```
glimpse(covid_data)
```

iii. Appendix C

```
ggplot(aes(x=Age.Group, color=Outcome, fill=Outcome), data= covid_data %>%
         filter(Outcome != "ACTIVE",Age.Group!="")) +
  geom_histogram(alpha=0.5, stat="count") +
  ggtitle("Number of RESOLVED and FATAL CONFIRMED COVID-19 Cases By Age.Group (Appendix C)")
```

iv. Appendix D

```
fatal_cases_data <- covid_data %>%
  filter(Outcome =="FATAL", Classification=="CONFIRMED")
resolved_cases_data <- covid_data %>%
  filter(Outcome =="RESOLVED", Classification=="CONFIRMED")
fatal_cases_data_count <- fatal_cases_data %>%
  group_by(Age.Group) %>% select(Age.Group) %>%
  summarise(Count=n()) %>% arrange(desc(Count)) %>%
  mutate(Age.Group = fct_reorder(Age.Group, Count, .desc=FALSE))
fatal_plot <- ggplot(data = fatal_cases_data_count, aes(x= Count, y=Age.Group)) +
  geom_col(fill = "#e52165") + ggtitle("FATAL CONFIRMED COVID-19 CASES BY AGE (Appendix D)")
 resolved_cases_data_count <- resolved_cases_data %>%
  group_by(Age.Group) %>% select(Age.Group) %>%
  summarise(Count=n()) %>% arrange(desc(Count)) %>%
  mutate(Age.Group = fct_reorder(Age.Group, Count, .desc=FALSE))
resolved_plot <- ggplot(data = resolved_cases_data_count,
                        aes(x= Count, y=Age.Group)) +
  geom_col(fill = "Green") + ggtitle("RESOLVED CONFIRMED COVID-19 CASES BY AGE (Appendix D)")
ggarrange(resolved_plot, fatal_plot, labels = c("", ""), ncol = 1, nrow = 2)
```

v. Appendix E

```
resolved_covid_cases <- covid_data %>%
  group_by(Age.Group) %>%
  filter(Age.Group!= "", Outcome=="RESOLVED", Classification=="CONFIRMED")
resolved_covid_cases_summarised <- resolved_covid_cases %>%
  summarise(NUMBER_OF_RESOLVED_CASES= n())
grouped_covid_data <- covid_data %>%
  group_by(Age.Group) %>%
  summarise(TOTAL_AMOUNT_OF_PEOPLE= n())
```

```
subset <- merge(x= resolved_covid_cases_summarised,
                y= grouped_covid_data,by="Age.Group",all=TRUE) %>%
  mutate(PROPORTION_OF_AGE_GROUP = format(NUMBER_OF_RESOLVED_CASES/TOTAL_AMOUNT_OF_PEOPLE,
                  scientific =FALSE))
drop <- c("TOTAL_AMOUNT_OF_PEOPLE")
df <- subset[,!(names(subset) %in% drop)]
df <- df %>% arrange(desc(PROPORTION_OF_AGE_GROUP))
df <- na.omit(df)
df
```

vi. Appendix F

```
fatal_covid_cases <- covid_data %>%
  group_by(Age.Group) %>%
  filter(Age.Group!= "", Outcome=="FATAL", Classification=="CONFIRMED")
fatal_covid_cases_summarised <- fatal_covid_cases %>%
  summarise(NUMBER_OF_FATAL_CASES= n())
grouped_covid_data <- covid_data %>% filter(Classification=="CONFIRMED") %>%
  group_by(Age.Group) %>%
  summarise(TOTAL_AMOUNT_OF_PEOPLE= n())
subset <- merge(x= fatal_covid_cases_summarised,
                y= grouped_covid_data,
                by="Age.Group",all=TRUE) %>%
  mutate(PROPORTION_OF_AGE_GROUP =
           format(NUMBER_OF_FATAL_CASES/TOTAL_AMOUNT_OF_PEOPLE,
                  scientific =FALSE))
drop <- c("TOTAL_AMOUNT_OF_PEOPLE")
df <- subset[,!(names(subset) %in% drop)]
df <- df %>%
   arrange(desc(PROPORTION_OF_AGE_GROUP))
df <- na.omit(df)
df
```

vii. Appendix G

```
number_of_trials_80_89 <- nrow(covid_data %>%
                                 filter(Classification=="CONFIRMED", Age.Group=="80 to 89 Years"))
number_of_success_80_89 <- nrow(covid_data %>%
                                  filter(Age.Group == "80 to 89 Years", Classification=="CONFIRMED",
                                         Outcome== "FATAL"))
number_of_trials_90 <- nrow(covid_data %>%
                              filter(Classification=="CONFIRMED", Age.Group == "90 and older"))
number_of_success_90 <- nrow(fatal_cases_data %>%
                               filter(Age.Group == "90 and older", Classification=="CONFIRMED",
                                      Outcome==  "FATAL"))
proportion_test <- prop.test(
  x=c(number_of_success_90, number_of_success_80_89), # number of success
  n = c(number_of_trials_90, number_of_trials_80_89)) # number of trials
proportion_test
```

viii. Appendix H

```
number_of_trials_70_79 <- nrow(covid_data %>%
                                 filter(Classification=="CONFIRMED", Age.Group=="70 to 79 Years"))
```

```r
number_of_success_70_79 <- nrow(covid_data %>%
                                  filter(Age.Group == "70 to 79 Years", Classification=="CONFIRMED",
                                         Outcome== "FATAL"))
proportion_test <- prop.test(
  x=c(number_of_success_90, number_of_success_70_79), # number of success
  n = c(number_of_trials_90, number_of_trials_70_79)) # number of trials
proportion_test
```

ix. Appendix I

```r
proportion_test <- prop.test(
  x=c(number_of_success_80_89, number_of_success_70_79), # number of success
  n = c(number_of_trials_80_89, number_of_trials_70_79)) # number of trials
proportion_test
```

x. Appendix J

```r
a <- covid_data %>% filter(Age.Group=="90 and older", Classification=="CONFIRMED")
x = as.numeric(a$Outcome=="FATAL")
print("Mean of the observed sample proportion:")
mean(x)
boot_function=function(){
  s= sample(x,
            size= length(x),
            replace=T)
  return(mean(s))
}
boot_prop = replicate(5000,boot_function())
test_stat = (mean(x)-0.20)/sd(boot_prop)
print("Test Stat:")
test_stat
print("P-value:")
pnorm(test_stat, lower.tail = FALSE)
```

xi. Appendix K

```r
a <- covid_data %>% filter(Age.Group=="80 to 89 Years", Classification=="CONFIRMED")
x = as.numeric(a$Outcome=="FATAL")
print("Mean of the observed sample proportion:")
mean(x)
boot_function=function(){
  s= sample(x,
            size= length(x),
            replace=T)
  return(mean(s))
}
boot_prop = replicate(5000,boot_function())
test_stat = (mean(x)-0.20)/sd(boot_prop)
print("Test Stat:")
test_stat
print("P-value:")
pnorm(test_stat)
```

xii. Appendix L

```
a <- covid_data %>% filter(Age.Group=="70 to 79 Years", Classification=="CONFIRMED")
x = as.numeric(a$Outcome=="FATAL")
print("Mean of the observed sample proportion:")
mean(x)
boot_function=function(){
  s= sample(x,
            size= length(x),
            replace=T)
  return(mean(s))
}
boot_prop = replicate(5000,boot_function())
test_stat = (mean(x)-0.10)/sd(boot_prop)
print("Test Stat:")
test_stat
print("P-value:")
pnorm(test_stat)
```

xiii. Appendix M

```
covid_data <- covid_data %>%
  mutate(Outcome_Dummy_Variable = ifelse(Outcome == "FATAL", 0, 1)) %>%
  filter(Age.Group != "")
m = glm(Outcome_Dummy_Variable ~ Age.Group, family = binomial, data = covid_data)
summary(m)
```

xiv. Appendix N

```
library(pROC)
k=20
fold.ind = sample(c(1:20), size=nrow(covid_data), replace=T)
c.index = vector()
for (i in 1:20){
  d.train = covid_data[ fold.ind != i , ]
  d.test = covid_data[ fold.ind == i , ]
  logit.mod = glm(Outcome_Dummy_Variable ~ Age.Group, family = binomial, data = d.train)
  pi_hat = predict(logit.mod, newdata=d.test, type = "response")
  m.roc=roc(d.test$Outcome_Dummy_Variable ~ pi_hat)
  c.index[i]=auc(m.roc)
}
print('c.index:')
c.index
print('Mean of c.index:')
mean(c.index)
```