STAA57 Project

Chun Kang Lu 1008161150

2023-03-24

About the Dataset

Available here (https://open.toronto.ca/dataset/covid-19-cases-in-toronto/), the COVID-19 Cases dataset provides weekly updating information about all reported cases of COVID-19 by the Toronto Public Health since the first reported case in January of 2020. The information in this dataset changes every week as new information get updated at 8:30 AM every Tuesday. For consistency in our analysis, we will be using the downloaded `.csv` version of this dataset from March 24 in this report.

This dataset contains 15 columns and 394761 rows, with each row representing a reported individual with COVID-19.

The columns are as follows:

- `_id` - A unique row identifier for Open Data database.
- `Assigned_ID` - A unique ID assigned to cases by Toronto Public Health for the purposes of posting to Open Data, to allow for tracking of specific cases.
- `Outbreak Associated` - Cases associated with outbreaks are identified by their settings and institutions (eg. long-term care homes, retirement homes, hospitals, homeless shelters, etc.).
- `Age Group` - Age at time of illness. Age groups (in years): ≤19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+, unknown (blank)
- `Neighbourhood Name` - 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` - 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 the person contracted COVID-19. These are determined through a few factors of a public health investigator's assessment, association with a confirmed COVID-19 outbreak, and known risk factors (like travel or close contact with another known COVID-19 case).
- `Classification` - Categorization of the cases as confirmed or probable, according to standard criteria of the Ontario Ministry of Health.
- `Episode Date` - Best estimate to when the disease was first acquired.
- `Reported Date` - Date which the case was reported to Toronto Public Health.
- `Client Gender` - Self-reported gender of the individual.
- `Outcome` - 3 scenarios of fatal, resolved, and active. - Fatal means the individual has died directly due to COVID-19 and not an alternative cause of death. - Resolved refers to individuals who have either: - Been marked Fatal but the cause of death was not due to COVID-19. - Are alive and has been more than 14 days

from the episode date and the case is not currently hospitalized, intubated, or in ICU. - Active refers to all other ongoing cases

- `Ever Hospitalized` - Cases that were hospitalized related to their COVID-19 infection (includes cases that are currently hospitalized and those that have been discharged or are deceased).
- `Ever in ICU` - Cases that were admitted to the intensive care unit (ICU) related to their COVID-19 infection (includes cases that are currently in ICU and those that have been discharged or are deceased).
- `Ever Intubated` - Cases that were intubated related to their COVID-19 infection (includes cases that are currently intubated and those that have been discharged or deceased)

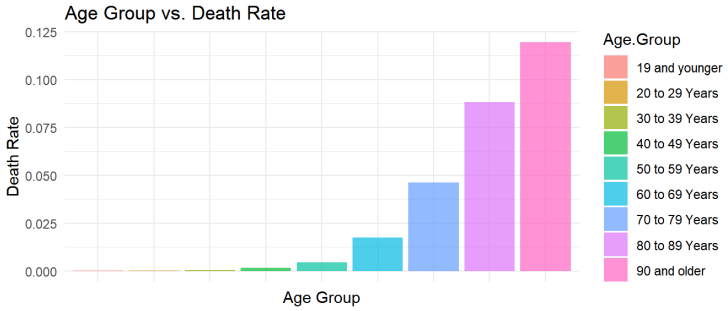
Main Question

We sought to examine what factors seemed to be the most prevalent among cases who are classified *FATAL* as these could serve as predictors for which portion of the population COVID-19 is most dangerous towards.

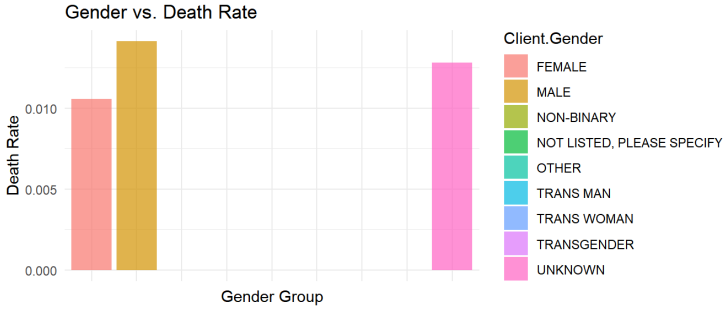
Preliminary Analysis

We chose to discard *ACTIVE* cases as these are people who have COVID-19 at the time of obtaining this CSV file, and we are unable to know if they will become *RESOLVED* or *FATAL* in the future.

Age.Group	TotalCases	deathRate
90 and older	10810	0.1193340
80 to 89 Years	18308	0.0882128
70 to 79 Years	20653	0.0462403
60 to 69 Years	33369	0.0175312
50 to 59 Years	51253	0.0045071
40 to 49 Years	55909	0.0014667
30 to 39 Years	72173	0.0005127
19 and younger	54337	0.0001104
20 to 29 Years	76618	0.0001044



Client.Gender	TotalCases	deathRate
MALE	180963	0.0141355
UNKNOWN	3583	0.0128384
FEMALE	208602	0.0105704
NON-BINARY	186	0.0000000
NOT LISTED, PLEASE SPECIFY	4	0.0000000
OTHER	14	0.0000000

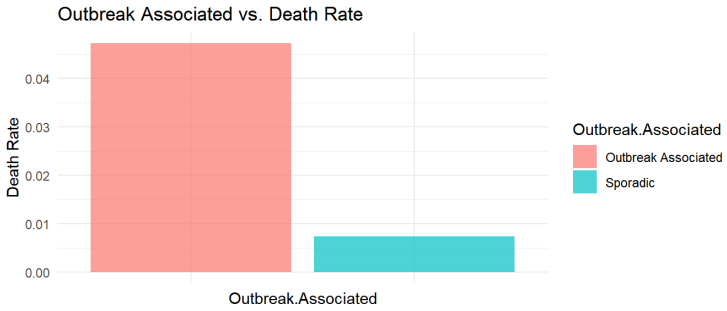


Client.Gender	TotalCases	deathRate
TRANS MAN	29	0.0000000
TRANS WOMAN	25	0.0000000
TRANSGENDER	24	0.0000000

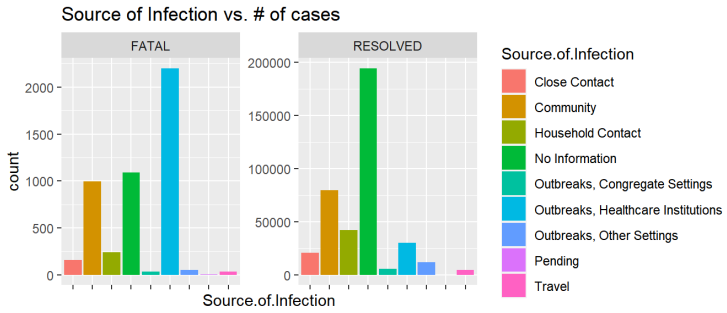
Client.Gender	Age.Group	TotalCases	deathRate
UNKNOWN	90 and older	31	0.4193548
UNKNOWN	80 to 89 Years	53	0.2452830
MALE	90 and older	3306	0.1454930
MALE	80 to 89 Years	7732	0.1122607
FEMALE	90 and older	7473	0.1065168
FEMALE	80 to 89 Years	10523	0.0697520
MALE	70 to 79 Years	10385	0.0581608
FEMALE	70 to 79 Years	10179	0.0342863
UNKNOWN	70 to 79 Years	87	0.0229885
MALE	60 to 69 Years	16362	0.0229801

location	TotalCases
M1B - Malvern	7072
M9V - Mount Olive-Silverstone-Jamestown	6460
M3J - York University Heights	5285
M5V - Waterfront Communities-The Island	4999
M2R - Westminster-Branson	4934
M1P - Dorset Park	4473
M3N - Black Creek	4293
M9W - West Humber-Clairville	4288
M1G - Woburn	4218
M3A - Parkwoods-Donalda	4191

Outbreak.Associated	TotalCases	deathRate
Outbreak Associated	48067	0.0472049
Sporadic	345363	0.0073546

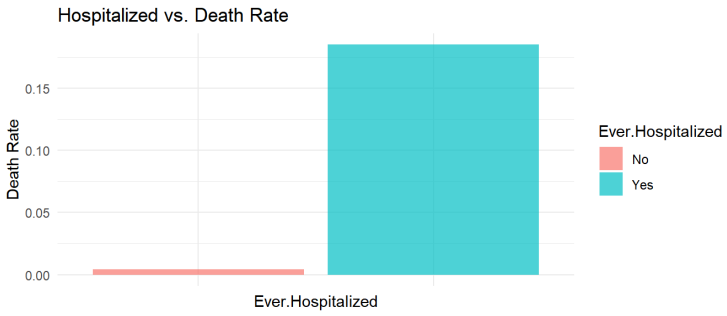


Source.of.Infection	TotalCases	deathRate
No Information	195525	0.0055645
Community	80602	0.0123198
Household Contact	42499	0.0056943
Outbreaks, Healthcare Institutions	32110	0.0685456
Close Contact	20613	0.0076651
Outbreaks, Other Settings	11689	0.0044486
Outbreaks, Congregate Settings	5607	0.0062422
Travel	4733	0.0076062

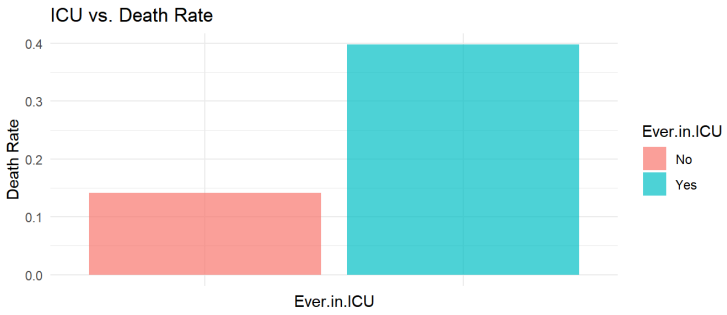


Source.of.Infection	TotalCases	deathRate
Pending	52	0.0769231

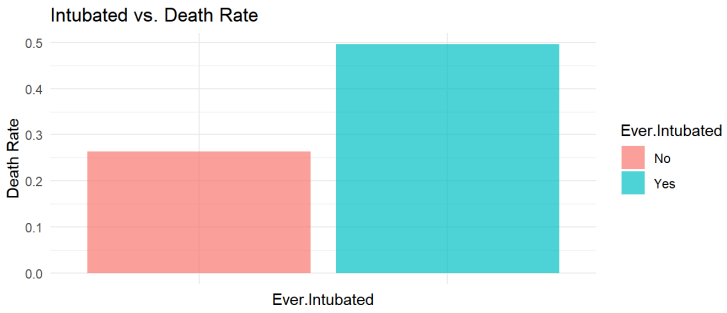
Ever.Hospitalized	TotalCases	deathRate
Yes	17126	0.1849819
No	376304	0.0043608



Ever.in.ICU	TotalCases	deathRate
Yes	2900	0.3975862
No	14226	0.1416421



Ever.Intubated	TotalCases	deathRate
Yes	1674	0.4958184
No	1226	0.2634584



Regression Analysis

As we can see from the figure below, there are magnitudes more RESOLVED cases than FATAL cases, creating a really imbalanced dataset.

Outcome	Count
FATAL	4809
RESOLVED	388621

From such, we use boostrapping to oversample the FATAL cases and undersample the RESOLVED cases to create a more balanced dataset and allow for our logistic regression model to fit to the data.

Using 8000 cases from each category, we fit a model on *Outbreak Associated, Age Group, Client Gender, Episode Date, Ever Hospitalized, Ever in ICU, and Ever Intubated* to determine which seem to have the highest correlation to Outcome. *Client Gender* was limited to only Male and Female as the other categories do not contain enough samples to draw conclusions. Both ID columns were not used as they are not relevant. *Neighbourhood name* and *FSA* both contain too many categories to draw significant correlations. *Reported Date* and *Episode Date* were not included as a logistic regression model will not be able to capture the time series data very well.

We can also use this opportunity to test the correlation between the each independent variable and their Outcome, as well as examine the confidence interval of the correlation.

```
##
## Call:
## glm(formula = out ~ Outbreak.Associated + Age.Group + Client.Gender +
##     Ever.Hospitalized + Ever.in.ICU + Ever.Intubated, family = binomial,
##     data = boot_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6388  -0.1912  -0.0251   0.2604   3.4888
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -4.49871    0.42407 -10.608  < 2e-16 ***
## Outbreak.AssociatedSporadic -1.58498    0.07181 -22.072  < 2e-16 ***
## Age.Group20 to 29 Years       0.34437    0.52468   0.656  0.51160
## Age.Group30 to 39 Years       1.39118    0.45546   3.054  0.00225 **
## Age.Group40 to 49 Years       2.21887    0.44188   5.021 5.13e-07 ***
## Age.Group50 to 59 Years       3.12796    0.43016   7.272 3.55e-13 ***
## Age.Group60 to 69 Years       4.03796    0.42685   9.460  < 2e-16 ***
## Age.Group70 to 79 Years       5.04258    0.42556  11.849  < 2e-16 ***
## Age.Group80 to 89 Years       5.93513    0.42510  13.962  < 2e-16 ***
## Age.Group90 and older         6.54917    0.42906  15.264  < 2e-16 ***
## Client.GenderMALE             0.50610    0.06708   7.544 4.54e-14 ***
## Ever.HospitalizedYes          3.12492    0.09098  34.347  < 2e-16 ***
## Ever.in.ICUYes                1.78208    0.25987   6.858 7.01e-12 ***
## Ever.IntubatedYes             1.35474    0.34484   3.929 8.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22180.7  on 15999  degrees of freedom
## Residual deviance:  6545.1  on 15986  degrees of freedom
## AIC: 6573.1
##
## Number of Fisher Scoring iterations: 7
```

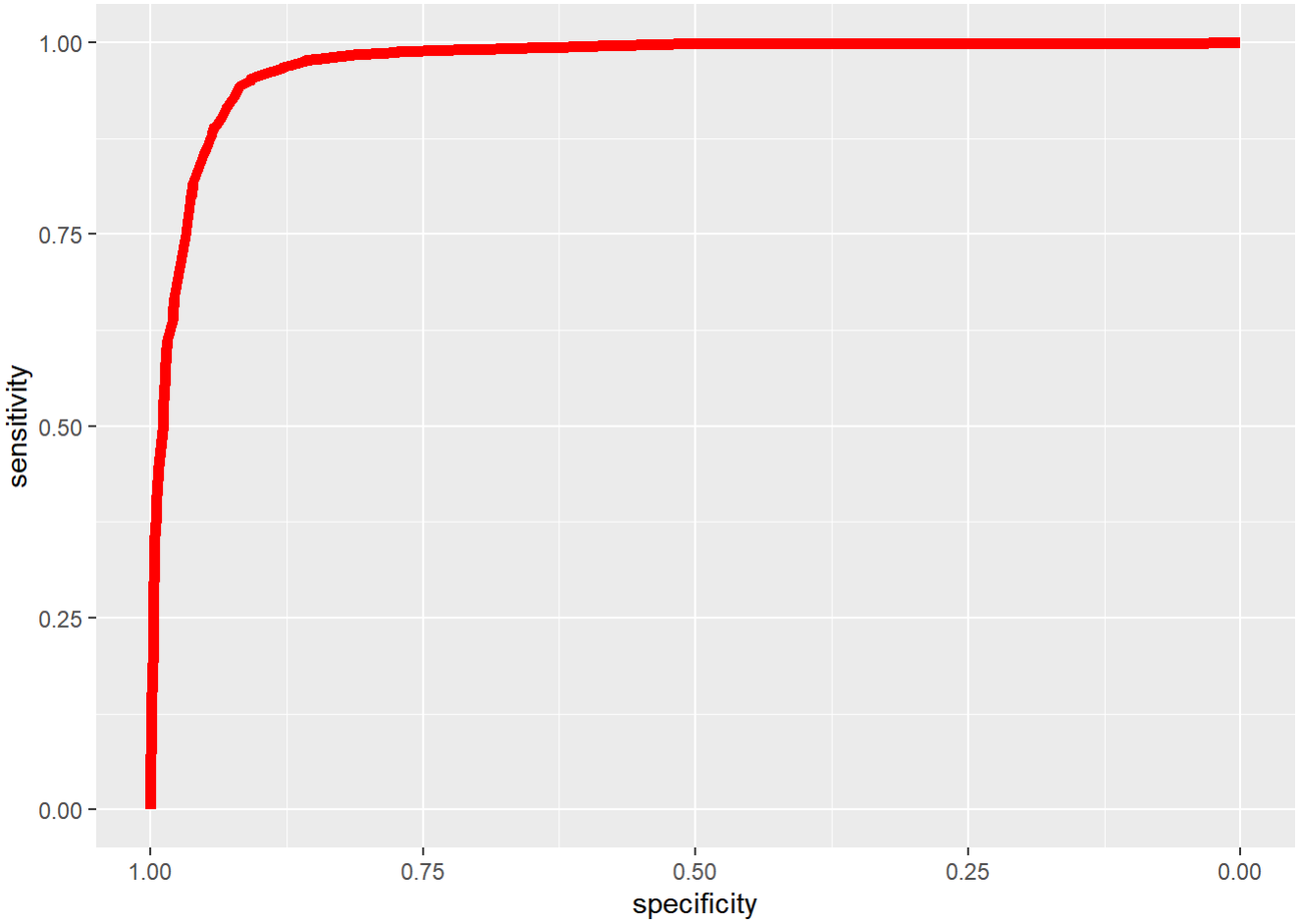
Following analysis on prior tables and graphs, we can see that FATAL cases are most prevalent in those who were in an outbreak, older age, male, as well as being hospitalized in ICU, being intubated. The highest correlation being those in the Age Group of 90 and older as that category contains the hightest log-odds at 6.54917 with respect to a FATAL outcome out of all variables.

Since p-values for all categories aside from *Age Group 20 to 29 Years* are very low, we can safely reject the null hypothesis in that those categories have zero correlation to the *Outcome* of the case. We can similarly obtain this conclusion by seeing that 0 is not within the confidence intervals of these categories.

```
##                                   2.5 %     97.5 %
## (Intercept)                 -5.4473922 -3.7556329
## Outbreak.AssociatedSporadic -1.7261080 -1.4445823
## Age.Group20 to 29 Years     -0.6579384  1.4401986
## Age.Group30 to 39 Years      0.5697437  2.3864762
## Age.Group40 to 49 Years      1.4315476  3.1940199
## Age.Group50 to 59 Years      2.3696913  4.0856592
## Age.Group60 to 69 Years      3.2878123  4.9906873
## Age.Group70 to 79 Years      4.2957264  5.9934429
## Age.Group80 to 89 Years      5.1894866  6.8853377
## Age.Group90 and older        5.7940888  7.5054180
## Client.GenderMALE            0.3749579  0.6379655
## Ever.HospitalizedYes         2.9487903  3.3055350
## Ever.in.ICUYes               1.2945104  2.3165144
## Ever.IntubatedYes            0.6784200  2.0386924
```

Model Prediction Accuracy

Using the same format of data, we perform cross-validation with a 70/30 train-test split to measure the accuracy of the model.




```
## Area under the curve: 0.9734
```

The model performs very well. Choosing a cutoff point of 0.9, we get the following confusion matrix.

```
##
##        0    1
##   0 2368   50
##   1  834 1528
```

And and overall accuracy of:

```
## [1] 0.8150628
```

Results

The COVID-19 pandemic has affected people of all ages and genders across the world, and people are striving to develop effective public health strategies to combat the COVID-19 pandemic. To achieve this, we need to understand how factors such as age, gender, and exposure to outbreaks can affect the death and resolution rates of COVID-19 patients.

One significant discovery in our project is that age plays a crucial role in determining the death rate. The elderly, particularly those between the ages of 80 and 89, have a higher risk of mortality compared to other age groups, possibly due to their weakened immune systems. On the other hand, young adults aged 20 to 29 tend to have a higher chance of recovery and a lower risk of death, suggesting that they are less susceptible to the severe impact of the virus.

We have also analyzed the impact of gender on death rates due to COVID-19. Although the death rate is slightly higher for males than females, this difference is not significant enough to be considered a major determinant of fatality from the disease. Thus, when we then analyze the outcome of the disease by grouping age and gender, we found that the role of gender is not significant; the death rate is mainly determined by age but not gender.

Furthermore, exposure to outbreaks is a crucial factor that affects the death rate from COVID-19. Those who have been present in an outbreak have a significantly higher death rate, almost seven times higher than those who were not. Moreover, the highest death rates occur during outbreaks and in some community settings, highlighting that most people contract COVID-19 in the community.

Additionally, individuals requiring hospitalization, ICU admission, or intubation have a higher death rate. This is not surprising, as these cases typically represent more severe forms of the disease. When examining hospitalized cases, the death rate is higher among those who were admitted to the ICU or required intubation than those who were not. These observations imply that early medical intervention and hospitalization could be vital in enhancing the chances of recovery from COVID-19.

Appendix

```
library(tidyverse)
library(knitr)
library(gridExtra)
library(kableExtra)
library(pROC)
library(leaflet)
# We have to note the data was downloaded on March 24, and the dataset is updated every week, so future downloads may get slightly different results.
data = read.csv("COVID19 cases.csv")
# Turn Episode.Date and Reported.Date into Date type
data = data %>% mutate(Episode.Date = as.Date(Episode.Date, format = "%Y-%m-%d"),
                       Reported.Date = as.Date(Reported.Date, format = "%Y-%m-%d"))
data = data %>% filter(Age.Group != "")
main_data = data %>% filter(Outcome!="ACTIVE") %>% select(-c(X_id, Assigned_ID))
gen_filtered = main_data %>% filter(Client.Gender %in% c("MALE", "FEMALE"))
```

```
# Tables & Graphs
# Relation between Age and death rate
data_byAge = main_data %>% filter(Age.Group != "") %>% group_by(Age.Group) %>% summarise(TotalCases=n(), deathRate = mean(Outcome == "FATAL"))%>%arrange(desc(deathRate))

kbl(data_byAge) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byAge, aes(x = Age.Group, y = deathRate, fill=Age.Group)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Age Group vs. Death Rate", x = "Age Group", y = "Death Rate") +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
# Relation between Gender and death rate
data_byGender = main_data %>% filter(Client.Gender!="") %>% group_by(Client.Gender)%>%summarise(TotalCases=n(),deathRate = mean(Outcome == "FATAL"))%>%arrange(desc(deathRate))

kbl(data_byGender) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byGender, aes(x = Client.Gender, y = deathRate, fill=Client.Gender)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Gender vs. Death Rate", x = "Gender Group", y = "Death Rate") +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
# Relation between gender and age vs. death rate
data_byGender_Age = main_data %>% filter(Client.Gender!="") %>% group_by(Client.Gender,Age.Grou
p)%>%summarise( TotalCases=n(),deathRate = mean(Outcome == "FATAL")) %>% arrange(desc(deathRat
e))
kbl(data_byGender_Age[1:10,]) %>%
    kable_styling(full_width = F, font_size = 10)

#Relation between location and total case
data_byLocation  = main_data %>% filter(FSA != "") %>%filter(Neighbourhood.Name != "")%>% mutate
(location = paste(FSA, Neighbourhood.Name, sep = " - "))%>% group_by(location) %>% summarise(Tot
alCases=n())

topCaseLocation = data_byLocation %>% arrange(desc(TotalCases))
kbl(topCaseLocation[1:10,]) %>%
    kable_styling(full_width = F, font_size = 10)
```

```
# Relation between Outbreak Associated and death rate
data_byOutbreakAssociated = main_data %>% filter(Outbreak.Associated!="") %>% group_by(Outbreak.
Associated)%>%summarise( TotalCases=n(), deathRate = mean(Outcome == "FATAL"))%>% arrange(desc(d
eathRate))

kbl(data_byOutbreakAssociated) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byOutbreakAssociated, aes(x = Outbreak.Associated, y = deathRate, fill=Outbreak.Asso
ciated)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Outbreak Associated vs. Death Rate", x = "Outbreak.Associated", y = "Death Rate"
) +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
# Relation between Source of Infection and Number of cases
data_byInfection = main_data %>% filter(Source.of.Infection!="") %>% group_by(Source.of.Infectio
n)%>%summarise( TotalCases=n(), deathRate = mean(Outcome == "FATAL"))%>% arrange(desc(TotalCase
s))

kbl(data_byInfection %>% arrange(desc(TotalCases))) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(main_data, aes(x=Source.of.Infection, fill=Source.of.Infection)) +
    geom_bar(position = "dodge") +
    facet_wrap(~Outcome, scales="free_y") +
    labs(title = "Source of Infection vs. # of cases")+
    theme(axis.text.x=element_blank())
```

```
# Relation between hospitalized and death rate
data_byHospitalized = main_data %>% filter(Ever.Hospitalized!="") %>% group_by(Ever.Hospitalize
d)%>%summarise( TotalCases=n(),deathRate = mean(Outcome == "FATAL")) %>%arrange(desc(deathRate))

kbl(data_byHospitalized %>% arrange(desc(deathRate))) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byHospitalized, aes(x = Ever.Hospitalized, y = deathRate, fill=Ever.Hospitalized)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Hospitalized vs. Death Rate", x = "Ever.Hospitalized", y = "Death Rate") +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
# Relation between ICU and death rate
# data_check = data %>% filter(Ever.in.ICU =="Yes" && Ever.Hospitalized== "No") gives 0 obs whic
h means in icu must be in hospital
# death rate is computed by only considering the case that Ever.Hospitalized== "Yes"
data_byICU = main_data %>% filter(Ever.in.ICU!="") %>%  filter(Ever.Hospitalized== "Yes") %>% gr
oup_by(Ever.in.ICU) %>%         summarise( TotalCases=n(),deathRate = mean(Outcome == "FATAL"))

kbl(data_byICU %>% arrange(desc(deathRate))) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byICU, aes(x = Ever.in.ICU, y = deathRate, fill=Ever.in.ICU)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "ICU vs. Death Rate", x = "Ever.in.ICU", y = "Death Rate") +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
# Relation between intubated and death rate
# data_check = data %>% filter(Ever.in.ICU =="No" && Ever.Hospitalized== "Yes") gives 0 obs whic
h means intubated must be in icu
data_byIntubated = main_data %>% filter(Ever.Intubated!="") %>% filter(Ever.in.ICU=="Yes") %>%
group_by(Ever.Intubated)%>%summarise( TotalCases=n(),deathRate = mean(Outcome == "FATAL"))

kbl(data_byIntubated %>% arrange(desc(deathRate))) %>%
    kable_styling(full_width = F, font_size = 10)

ggplot(data_byIntubated, aes(x = Ever.Intubated, y = deathRate, fill=Ever.Intubated)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Intubated vs. Death Rate", x = "Ever.Intubated", y = "Death Rate") +
  theme_minimal()+
  theme(axis.text.x=element_blank())
```

```
outcome_count = main_data %>% group_by(Outcome) %>% summarize(Count=n())
kbl(outcome_count) %>%
    kable_styling(full_width = F, font_size = 10)
```

```
# Regression Analysis
main_res = gen_filtered %>% filter(Outcome=="RESOLVED")
main_fat = gen_filtered %>% filter(Outcome=="FATAL")
set.seed(5)
boot_res = main_res %>% sample_n(8000, replace=T)
boot_fat = main_fat %>% sample_n(8000, replace=T)
boot_data = rbind(boot_res, boot_fat)
boot_data = boot_data %>% mutate(out = ifelse(Outcome == "FATAL", 1, 0))
model = glm(out ~ Outbreak.Associated + Age.Group + Client.Gender +  Ever.Hospitalized + Ever.i
n.ICU + Ever.Intubated, family=binomial, data = boot_data)
summary(model)
confint(model)
```

```
# Cross Validation
boot_data = boot_data %>% mutate(group_ind = sample(c("train","test"),
                size=nrow(boot_data),
                prob = c(0.7,0.3),
                replace = T))
train = boot_data %>% filter(group_ind == "train")
test = boot_data %>% filter(group_ind == "test")
model2 = glm(out ~ Outbreak.Associated + Age.Group + Client.Gender + Ever.Hospitalized + Ever.i
n.ICU + Ever.Intubated,
            family=binomial, data = train)
test = test %>% mutate(pred = predict(model, newdata = test, type="response"))
```

```
# ROC, AUC and Analysis
roc_logit = roc(test$out ~ test$pred)
ggroc(roc_logit, color="red", size = 2)
auc(roc_logit)
test = test %>% mutate(pred_out = ifelse(pred >= 0.9, 1, 0))
conmat = table(test$out, test$pred_out)
conmat
sum(diag(conmat))/sum(conmat)
```