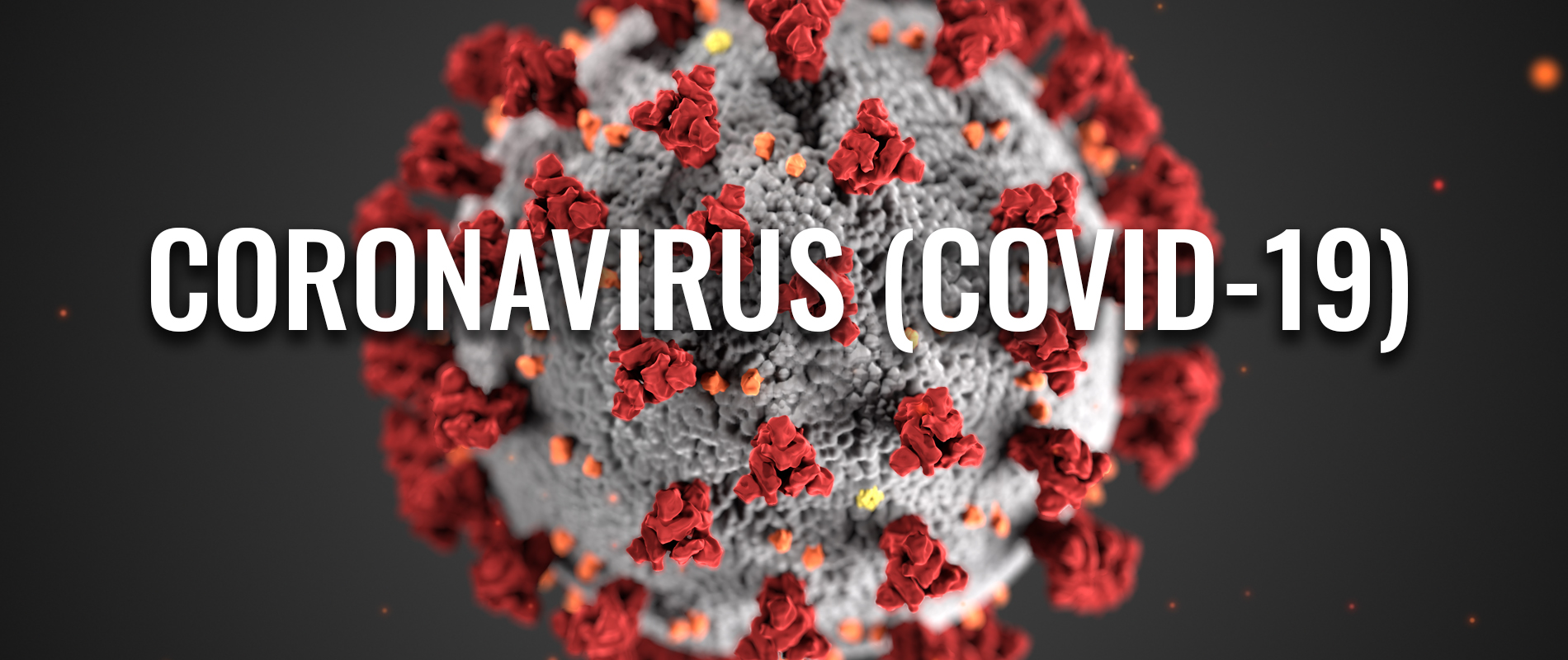
**COVID19 Death and Recovery prediction**

****

**Methods to Predict Deaths and Recoveries from COVID-19**

**Problem Statement**

In this project, our goal is to analyze and decide which individuals who have tested positive for COVID-19 are more likely to recover and die. Our reason for this problem statement is three-fold: having the results from this data will help frontline healthcare workers and government officials to prioritize the limited access to things such as emergency rooms, equipment such as ventilators, and resources such as oxygen. These results will also help to decide which individuals should be advised to take some time away from the public and sequester themselves in quarantine. Lastly, we would like to know which patients would need more attention once admitted to the hospital based on the results of the data concluding who is more likely to die in comparison to the others.

**Background and Literature Review**

All around the world, beginning in late 2019 and now well into 2020, every single one of us are living in a way we never could have ever expected in our lifetime. Some of us had plans for gatherings with loved ones such as wedding and birthday celebrations, while others planned on walking across a stage to receive their diploma in-hand and throw their caps in the air as friends and families clapped and cheered for them in the audience. In order to fully comprehend the implications of the time we are living in now due to the COVID-19 pandemic, Newland shares that “COVID-19 has been the greatest disruption to the movement of people since World War II” (2020).

In November 2019, various reports of a pneumonia-like sickness arose in Wuhan, China and one month later, on December 31, the Wuhan Municipal Health Commission “reported a cluster of cases of pneumonia” where it was also determined that “a novel coronavirus was identified” (World Health Organization, 2020). China reported its first death on January 11, 2020 and shared the genetic sequence of COVID-19 the next day. This public information sharing was considered “of great importance for other countries to use in developing specific diagnostic kits” (World Health Organization, 2020). Just a couple of weeks later, on January 21, the first case of the new coronavirus was confirmed in the United States (Hauck et al., 2020). This first case was discovered in Washington state and was “a man in his 30s [who] returned from Wuhan a week earlier” (Hauck et al., 2020). About two weeks later, the first death in the United States occurred. This was not originally known until months later, after an autopsy was completed.

Much has been learned about the virus along the way and still continues to this day. In late February, the CDC reported community spread after an individual in California who had not visited another country recently or come into contact with an infected patient (Hauck et al, 2020). On March 11, it was announced by the World Health Organization that the spread of COVID-19 had become a pandemic and two days later, U.S. President Donald Trump declared a national emergency. About one week later, the U.S. President then invoked the Defense Production Act, “a wartime authority that [allowed] him to direct industry to produce critical equipment” (Hauck et al., 2020). This same day, a study published by the New England Journal of Medicine found that the virus could be detected on copper up to 4 hours, on cardboard for up to 24 hours, and 2-3 days on plastic and stainless steel (Doremalen et al., 2020, p. 3).

The United States has continued to see a growth in the number of COVID-19 cases. On March 19, 2020, the country surpassed 10,000 cases and on March 26 became “the planet’s most infected nation” with 83,836 cases after surpassing China and Italy (Hauck et al., 2020). The largest stimulus package in U.S. history was signed on March 27 by the U.S. President for the amount of two-trillion dollars which was used to provide many American citizens with a $1,200 check (and more for individuals/families with children) while also helping companies to maintain their payroll throughout the crisis (Hauck et al., 2020). This same day, the United States surpassed 100,000 cases of COVID-19 and just a few days later on April 1, the cases doubled to more than 200,000.

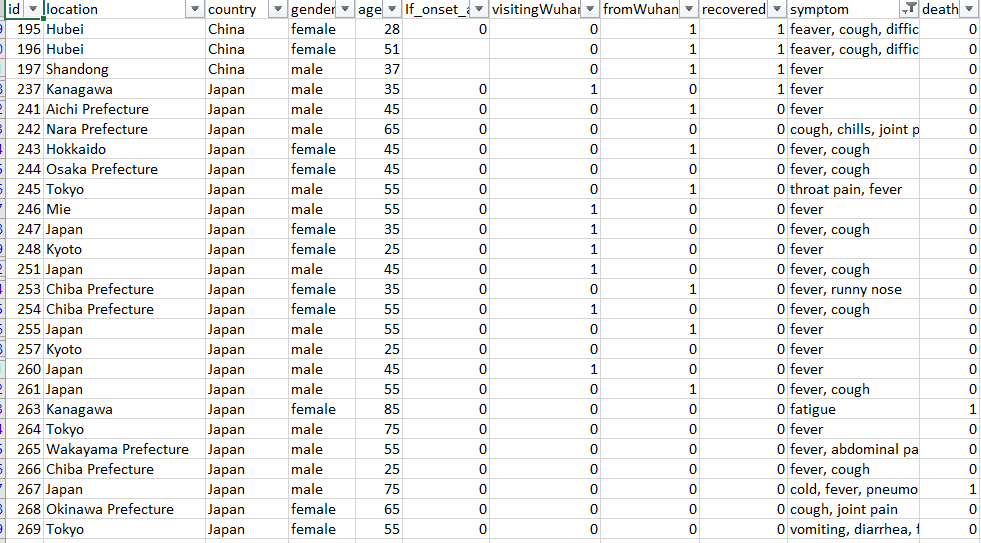
Living through the COVID-19 pandemic has been a challenging time for many. In April, the Department of Labor reported that a record 6.86 million Americans “filed first-time jobless claims” making it an all-time high (Hauck et al., 2020). The U.S. President advised Americans to begin wearing face masks while in public a week before the U.S. experienced it’s deadliest day on April 10 where more than 2,000 people died (Hauck et al., 2020). During this time, the city of Wuhan where the virus was first discovered celebrated the lifting of a 76-day lockdown. Studies regarding the psychological effects of this pandemic have begun to be published. One such study conducted an online cross-sectional survey of the general public in Italy for a period of six days during the time of the highest daily increase of infections (Davico et al., 2020, p. 4). The study concluded that “up to 30% of adults and children in the pandemic area [were] at high risk for post-traumatic stress disturbances…[with a greater] risk...for HCW [health care workers] directly involved in COVID-19 care and their children” (Davico et al., 2020, p. 2).

New information continues to be discovered regarding COVID-19. The most recent includes six new symptoms which the CDC cautioned may be signs of the coronavirus including “chills, repeating shaking with chills, muscle pain, headache, sore throat, and a loss of taste or smell” (Hauck et al., 2020). On April 28, the United States surpassed one-million confirmed cases -- “nearly a third of the world’s cases” (Hauck et al., 2020). Although we certainly acknowledge that the information regarding the pandemic is rapidly evolving, the spread is still ongoing, and investigations are still occuring, we chose to work with some of the data presented thus far regarding COVID-19 in order to gain a more detailed understanding of what we are all currently dealing with.

**Methodology** a) Research Questions  
 In this project, we worked to predict who may die based on the individual's age, gender, whether they visited Wuhan, China or not, whether they live/lived in Wuhan, China or not, and the symptoms they displayed/reported.

b) Features Explanation, Models, and Evaluation Metrics

**Dataset:**



Data Source: John Hopkins GitHub Repository

After data cleaning, there were 20 features and 825 rows present in the dataset.

The dataset contains information from January 13, 2020 up to February 28, 2020.

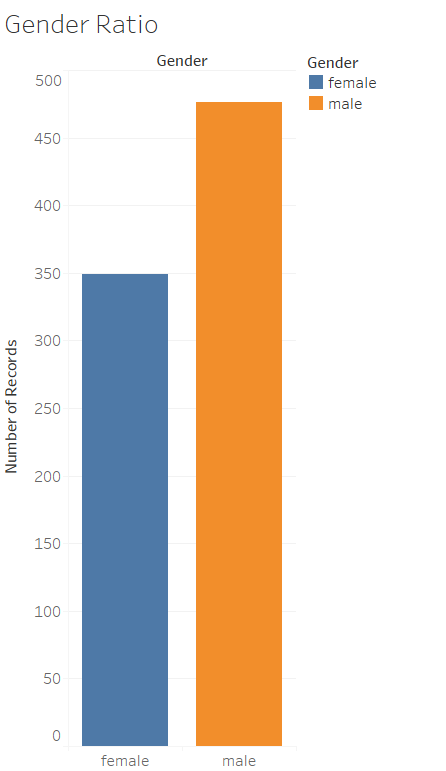
The features which are included in our dataset are “gender,” “age,” “Visit Wuhan,” “From Wuhan,” “mainsymptom,” ”recovered,” and “death.”

* The gender feature is binary and categorized depending on whether the individual is/was Male or Female
* The age range of our dataset includes individuals as young as a 6-month old baby to a 96-year-old
* The “Visit Wuhan” feature categorizes whether or not the individual has visited Wuhan, China (1 meaning they have visited Wuhan and 0 meaning they have not visited Wuhan)
* The “From Wuhan” feature categorizes whether or not the individual is from Wuhan, China (1 meaning they are from Wuhan and 0 meaning they are not from Wuhan)
* The “mainsymptom” is a logical feature and categorizes whether or not a patient has the main symptoms for COVID-19. Main symptoms include cough, fever, throat itching, breathlessness, dyspnea, pneumonia and malaise
* The “death” feature includes 1 for individuals who have died and 0 for individuals who are still alive
* The “recovered” feature includes 1 for individuals who have recovered and 0 for individuals who have not recovered

**Visualizations:**

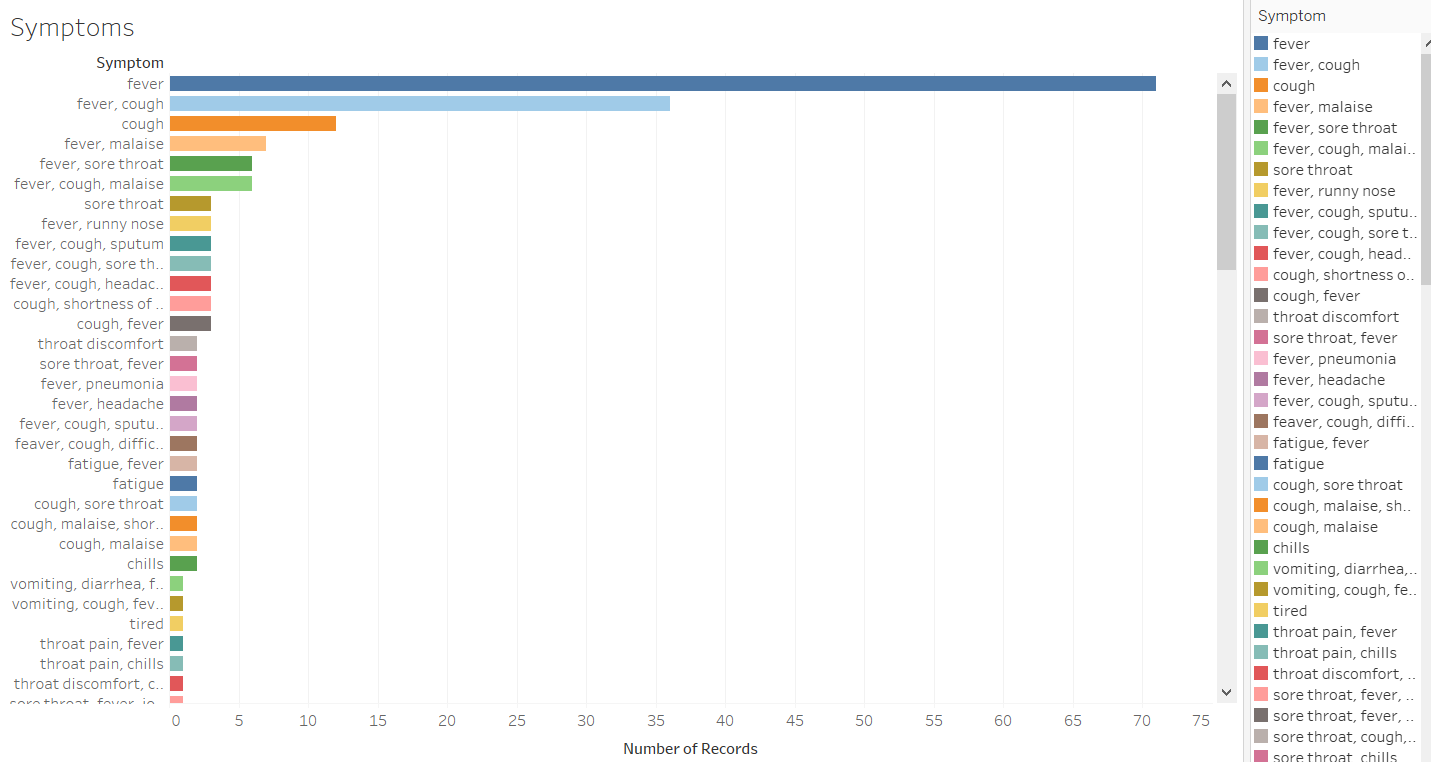


The dataset contained data of patients from 26 countries across 4 continents (Asia, Australia, Europe, and North America).

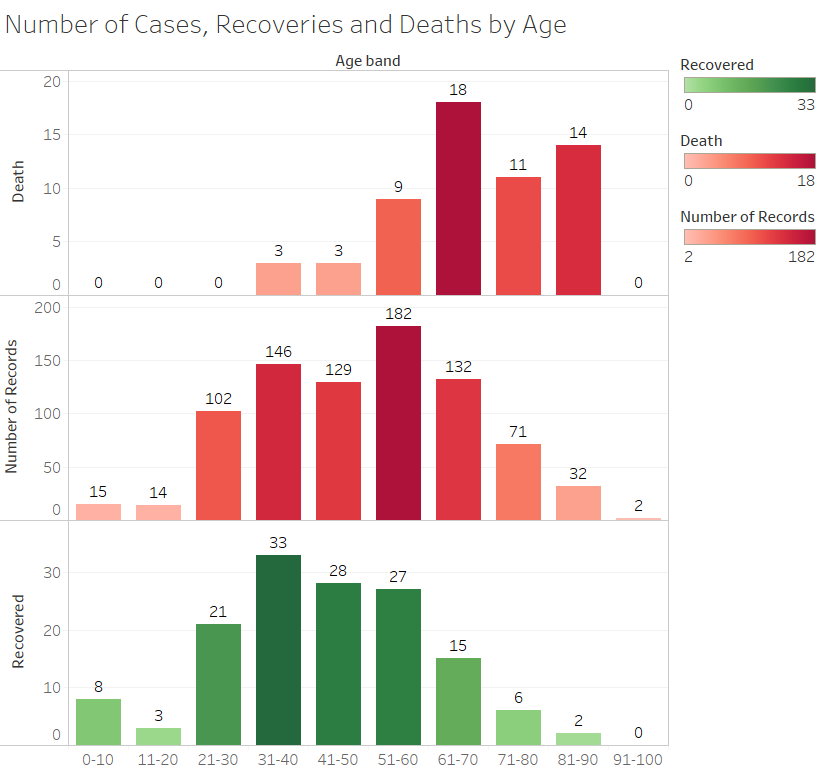


Number of Females = 350

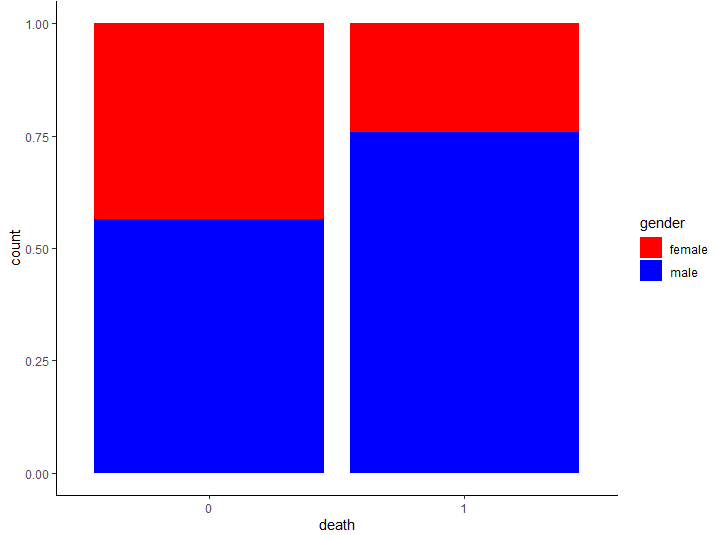
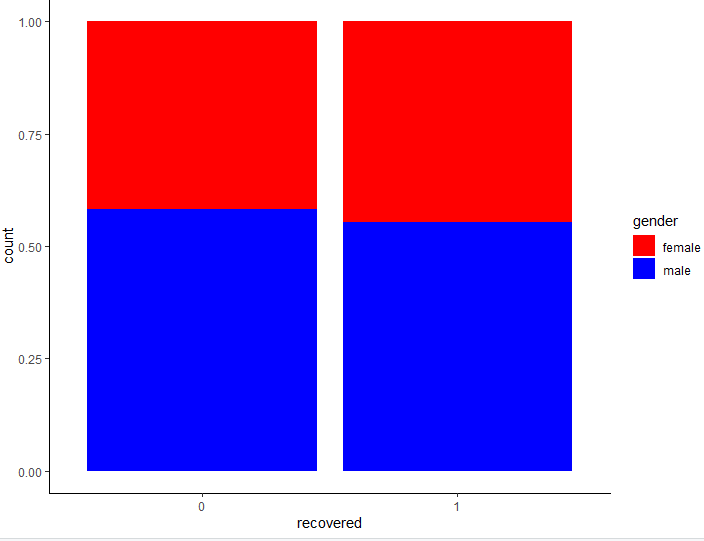
Number of Males = 475

****

201 patients had symptoms related to COVID-19 such as Fever, cough, breathlessness, dyspnea, pneumonia, and malaise.



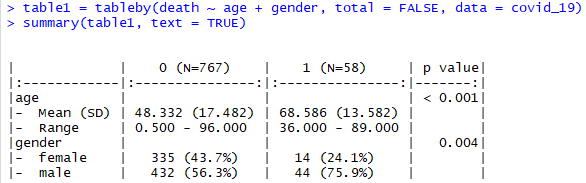
Elder patients (ages 60 and above) are more susceptible to die since they have a weak immunity, with patients in the age group 81-90 having the highest ratio of deaths. Younger patients recover more than older patients as they have better immunity and patients in the age group 0-10 and 21-30 have the highest ratio for recoveries



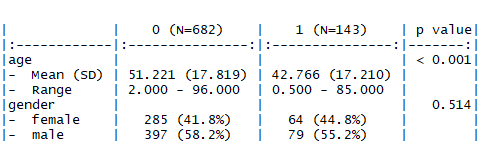
Men and women recover at the same rate. However, when it comes to death men are more likely to die (75% chance) than women (25% chance)

For our models, we chose to utilize logistic regression and random tree forest. The threshold for the logistic model to categorize a patient as dead is a probability of >0.5 or 50%. For both of our models, our data has been split in 75 percent for training and 25 percent for testing. The predictions are based on features -- the percentage of death for each person is given.

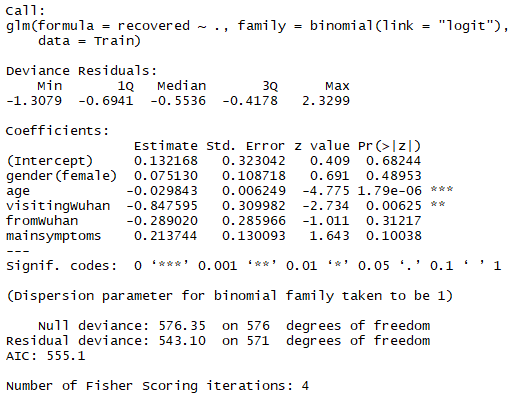
**Results (Prediction Results, Evaluation)**

****

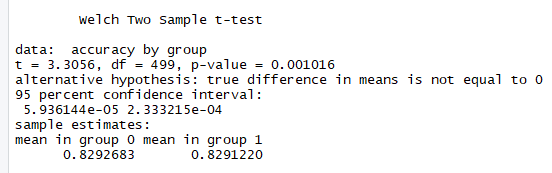
**Table 1 - Deaths vs AGE and Gender**

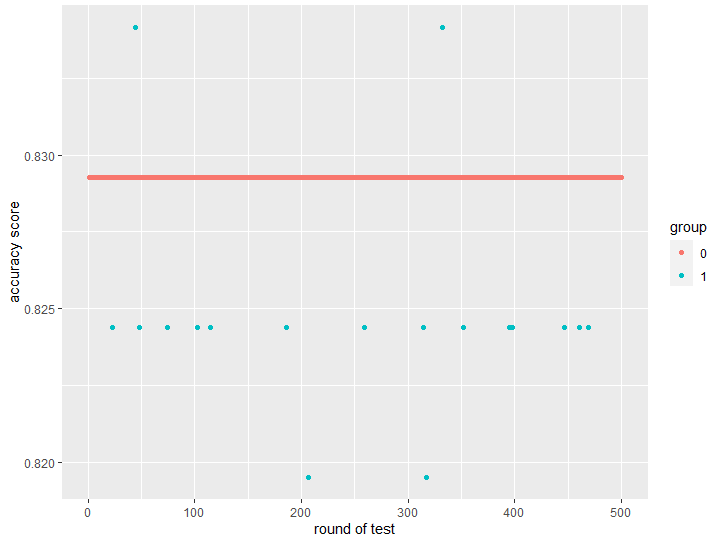
****

**Table 2 - Recoveries vs AGE and Gender**

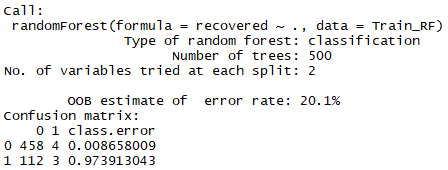
**Logistic model for Recovered:**



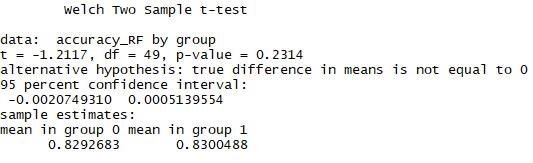


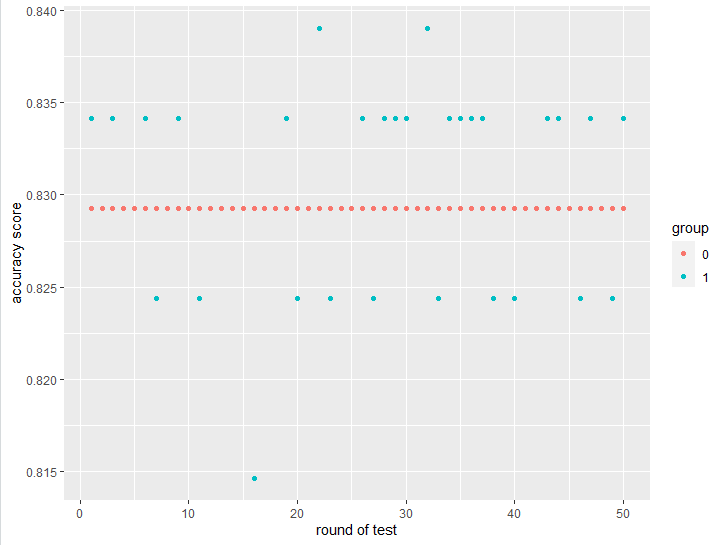


**Random Forest for Recovered:**

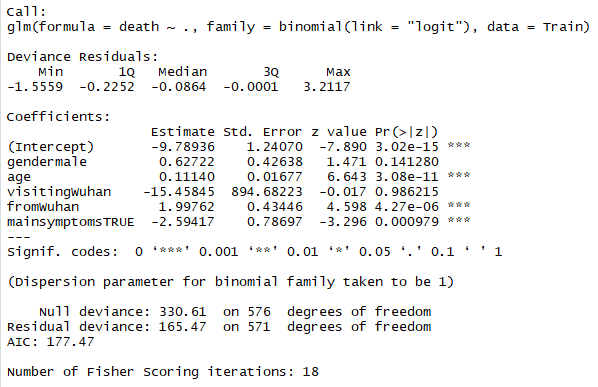




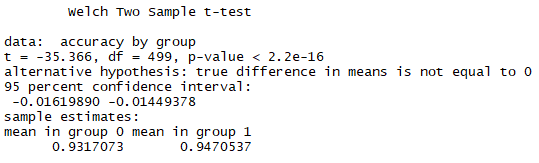


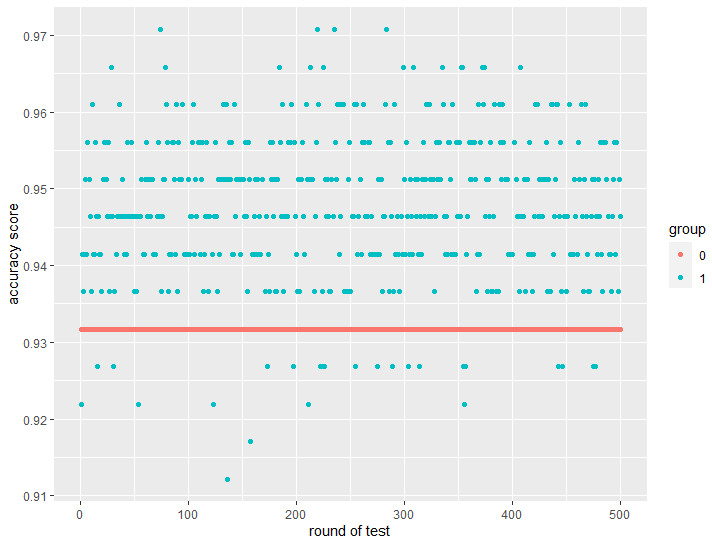


**Logistic regression for Death:**

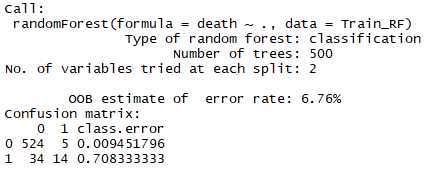


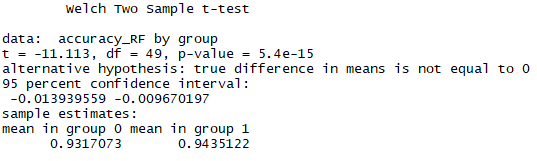


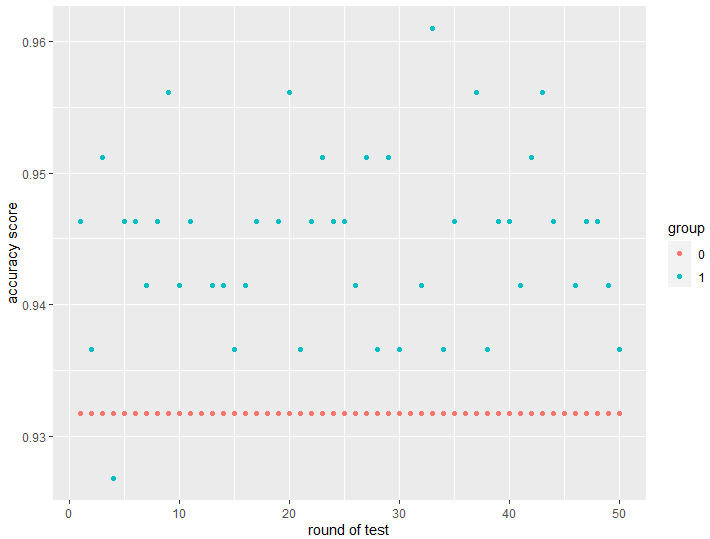


**Random Forest for Death:**

****

** **

****



**Discussion**

**Logistic model for Recovered:**

* Modeled for ‘recoveries’ (0- not recovered, 1- recovered)
* The p-values indicate that age and visitingWuhan are the most significant features
* Null deviance = 576.35 on df = 576
* Residual deviance = 543.1 on df = 571
* AIC = 555.1
* The value of Deviance/df ~ 1, which is very close to unity
* Hence the model is an adequate fit to the data
* Average accuracy of the model is 82.91%
* The baseline model consists of visitingWuhan as the only independent feature and has an average accuracy of 82.93%
* From the Welch two sample t-test we can state that there is no clear difference between the average accuracy scores for the baseline model and the original model

**Random Forest model for Recovered:**

* Number of trees is 500
* Error rate = 20.1%
* Confusion matrix indicates that the number of TP = 3, TN = 458, FP = 4 and FN = 112
* The average accuracy of the model is 82.99%
* The baseline model consists of visitingWuhan as the only independent feature and has an average accuracy of 82.93%
* From the Welch two sample t-test we can state that there is no clear difference between the average accuracy scores for the baseline model and the original model

**Logistic model for Death:**

* Modeled for deaths’ (0- didn’t die, 1- died)
* The p-values indicate that age, fromWuhan and mainsymptomsTrue are the most significant features
* Null deviance = 330.61 on df = 576
* Residual deviance = 165.47 on df = 571
* AIC = 177.47
* The value of Deviance/df ~ 0.3, which is less than unity
* Hence the model is an adequate fit to the data
* Average accuracy of the model is 94.7%
* The baseline model consists of visitingWuhan as the only independent feature and has an average accuracy of 93.17%
* From the Welch two sample t-test we can state that the initial model is more significant than the baseline model

**Random Forest model for Death:**

* Number of trees is 500
* Error rate = 6.76%
* Confusion matrix indicates that the number of TP = 14, TN = 524, FP = 5 and FN = 34
* The average accuracy of the model is 94.35%
* The baseline model consists of visitingWuhan as the only independent feature and has an average accuracy of 93.17%
* From the Welch two sample t-test we can state that the initial model is more significant than the baseline model

**Conclusion**

After running both the models for recovered and death cases we can conclude that there is not much difference between the performance of Logistic Regression Model and Random Forest Model for both the cases. Although the models have much higher accuracy predicting the deaths than recoveries.

From Table 1 in Results we can conclude that:

* Males are at a higher risk of death - close to 75% are male
* The mean age for death is about 68.5 and the minimum age was 36 years and maximum age was 89 years
* Hence, we can also conclude that older men are the most susceptible to die from COVID-19

From Table 2 in Results we can conclude that:

* Males and females both recover at a similar rate
* The mean age for recoveries is about 43 and the minimum age was 6 months and maximum age was 85 years
* Hence, we can also conclude that young patients are the most likely to recover from COVID-19

**References**

Davidco, C., Ghiggia, A., Marcotulli, D., Ricci, F., Amianto, F., Vittiello, B. (2020).

“Psychological Impact of the COVID-19 Pandemic on Adults and their Children in

Italy.” The Lancet Psychiatry.

Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson,

B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., Lloyd-Smith, J.

O., Wit, E., Munster, V. J. (2020). “Aerosol and Surface Stability of HCoV-19

(SARS-CoV-2) Compared to SARS-CoV-1.” MedRxiv. Retrieved from

<https://www.medrxiv.org/content/10.1101/2020.03.09.20033217v1.full.pdf>.

Huack, G., Gelles, K., Bravo, V., & Throson, M. “Three Months In: A Timeline of How

COVID-19 Has Unfolded in the US.” Retrieved from

[https://www.usatoday.com/in-depth/news/nation/2020/04/21/coronavirus-updates](https://www.usatoday.com/in-depth/news/nation/2020/04/21/coronavirus-updates-how-covid-19-unfolded-u-s-timeline/2990956001/)

[-how-covid-19-unfolded-u-s-timeline/2990956001/](https://www.usatoday.com/in-depth/news/nation/2020/04/21/coronavirus-updates-how-covid-19-unfolded-u-s-timeline/2990956001/).

Newland, K. “Lost in Transition.” *Science Magazine.* Vol 368, Issues 6489, p. 343. DOI:

10.1126/science.abc3197.

World Health Organization. (2020). “Novel Coronavirus - China.” Retrieved from

<https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/>.

Word Health Organization. (2020). “WHO Timeline - COVID-19.” Retrieved from

<https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19>.

**Appendix**

**R-code:**

**Logistic and Random Forest for Deaths:**

setwd("P:\\Spring 2020\\ML\\Final project")

#Loading all the required packages

library(caret)

library(tidyverse)

library(arsenal)

library(GGally)

library(party)

library(stringr)

library(dplyr)

library(ggplot2)

library(randomForest)

library(corrplot)

#Reading the file and inserting it into a variable

covid\_19 = read.csv('COVID19\_line\_list\_data.csv')

View(covid\_19)

summary(covid\_19)

## Extract the main symptoms of COVID-19

covid\_19$mainsymptoms = str\_detect(covid\_19$symptom, 'cough|fever|throat|breathlessness|dyspnea|pneumonia|malaise')

#Creating a subfile with only the essential variables

covid\_19 = covid\_19 %>% select(death,reporting.date, country, gender, age, visitingWuhan, fromWuhan, mainsymptoms, symptom) %>%

mutate(

death = ifelse(death == '0', 0, 1),

country = factor(country),

gender = factor(gender),

death = factor(death, label = c('0','1')),

reporting.date = as.Date(reporting.date, format = c('$d/$m/$Y')),

mainsymptons = factor(mainsymptoms)

)

#Plotting Age vs Risk of Death

ggplot(covid\_19, aes(death, age, fill = death))+

geom\_boxplot()+

theme\_classic()+

scale\_fill\_manual(values = c('blue','red'))

#Plotting Gender vs Risk of Death

ggplot(covid\_19, aes(death, fill = gender))+

geom\_bar(position ='fill')+

theme\_classic()+

scale\_fill\_manual(values = c('red','blue','black'))

#Most of Deaths were in older and male sex

table1 = tableby(death ~ age + gender, total = FALSE, data = covid\_19)

summary(table1, text = TRUE)

#Creating a dataset with only the essential columns used to create the logistic model

covidLR = covid\_19 %>% select(death, gender, age, visitingWuhan, fromWuhan, mainsymptoms)

#Creating a train and a test set in the ratio 75:25 respectively

Train = covidLR[1:577,1:6]

Test = covidLR[578:825,1:6]

#Building the logistic regression model

log\_model = glm(formula = death~., family = binomial(link = "logit"), data= Train)

summary(log\_model)

#Predicting the death for the Test dataset and comparing it with the Test dataset to determine accuracy of the logistic model

result = predict(log\_model, newdata=Test[,2:6], type="response")

result

result = ifelse(result>0.5,1,0)

result

accuracy = mean(result == Test$death)

accuracy

#Two models training/testing/comparison

num\_iterations = 500

acc\_history = list(num\_iterations)

acc\_history\_visitingwuhan = list(num\_iterations)

for (i in 1:num\_iterations) {

inTrain = createDataPartition(y=covidLR$death, p=0.75, list=FALSE)

X\_train = covidLR[inTrain, ]

X\_test = covidLR[-inTrain, ]

model = glm(formula = death~., family = binomial(link = "logit"), data= X\_train)

result = predict(model, newdata=X\_test[,2:6], type="response")

result = ifelse(result>0.5,1,0)

accuracy = mean(result == X\_test$death)

acc\_history[[i]] = accuracy

model\_1 = glm(formula = death~visitingWuhan, family = binomial(link = "logit"), data= X\_train)

result\_visitingwuhan = predict(model\_1, newdata=X\_test[,2:6], type="response")

result\_visitingwuhan = ifelse(result\_visitingwuhan>0.5,1,0)

accuracy\_visitingwuhan = mean(result\_visitingwuhan == X\_test$death)

acc\_history\_visitingwuhan[[i]] = accuracy\_visitingwuhan

}

##Printing average accuracy for 500 iterations

sum\_acc = 0

for (i in 1:num\_iterations) {

sum\_acc = sum\_acc + acc\_history[[i]]

}

ave\_acc = sum\_acc/num\_iterations

print(ave\_acc)

sum\_acc\_1 = 0

for (i in 1:num\_iterations) {

sum\_acc\_1 = sum\_acc\_1 + acc\_history\_visitingwuhan[[i]]

}

ave\_acc\_1 = sum\_acc\_1/num\_iterations

print(ave\_acc\_1)

#Welch test to comare the 2 models

df1 = data.frame(matrix(unlist(acc\_history), nrow=length(acc\_history), byrow=T))

df1$group = 1

df1$i = seq.int(nrow(df1))

names(df1)[1] = "accuracy"

df2 = data.frame(matrix(unlist(acc\_history\_visitingwuhan), nrow=length(acc\_history\_visitingwuhan), byrow=T))

df2$group = 0

df2$i = seq.int(nrow(df2))

names(df2)[1] = "accuracy"

df3 = rbind(df1, df2)

ggplot(data = df3, aes(x=i, y=accuracy, color=factor(group))) + xlab("round of test") +

ylab("accuracy score") + geom\_point() + labs(color="group")

t.test(accuracy~group, data=df3)

##################### Random Forest #########################

#Creating a dataset with only the essential columns used to create the RandomForest model

covid\_RF = covid\_19 %>% select(death, gender, age, visitingWuhan, fromWuhan, mainsymptoms)

#Creating a train and a test set in the ratio 75:25 respectively

Train\_RF = covid\_RF[1:577,1:6]

Test\_RF = covid\_RF[578:825,1:6]

##Building the RandomForest model

model\_RF = randomForest(death~., data=Train\_RF)

model\_RF

#Predicting the death for the Test dataset and comparing it with the Test dataset to determine accuracy of the RandomForest model

prediction\_RF = predict(model\_RF, newdata=Test\_RF)

prediction\_RF

accuracy\_RF = mean(prediction\_RF == Test\_RF$death)

accuracy\_RF

#Two models training/testing/comparison

num\_iterations\_RF = 50

acc\_history\_RF = list(num\_iterations\_RF)

acc\_history\_1\_RF = list(num\_iterations\_RF)

for (i in 1:num\_iterations\_RF) {

inTrain\_RF = createDataPartition(y=covid\_RF$death, p=0.75, list=FALSE)

X\_train\_RF = covid\_RF[inTrain\_RF, ]

X\_test\_RF = covid\_RF[-inTrain\_RF, ]

model\_RF = randomForest(death~., data=X\_train\_RF)

model1\_RF = randomForest(death~visitingWuhan, data=X\_train\_RF)

prediction\_RF = predict(model\_RF, newdata=X\_test\_RF)

prediction1\_RF = predict(model1\_RF, newdata=X\_test\_RF)

accuracy\_RF = mean(prediction\_RF == X\_test\_RF$death)

accuracy1\_RF = mean(prediction1\_RF == X\_test\_RF$death)

acc\_history\_RF[[i]] = accuracy\_RF

acc\_history\_1\_RF[[i]] = accuracy1\_RF

}

for (i in 1:num\_iterations\_RF) {

print(acc\_history\_RF[[i]])

}

sum\_acc\_RF = 0

for (i in 1:num\_iterations\_RF) {

sum\_acc\_RF = sum\_acc\_RF + acc\_history\_RF[[i]]

}

##print average accuracy for 50 iterations

ave\_acc\_RF = sum\_acc\_RF/num\_iterations\_RF

print(ave\_acc\_RF)

sum\_acc\_RF\_1 = 0

for (i in 1:num\_iterations\_RF) {

sum\_acc\_RF\_1 = sum\_acc\_RF\_1 + acc\_history\_1\_RF[[i]]

}

ave\_acc\_RF\_1 = sum\_acc\_RF\_1/num\_iterations\_RF

print(ave\_acc\_RF\_1)

#Welch test to comare the 2 models

df1\_RF = data.frame(matrix(unlist(acc\_history\_RF), nrow=length(acc\_history\_RF), byrow=T))

df1\_RF$group = 1

df1\_RF$i = seq.int(nrow(df1\_RF))

names(df1\_RF)[1] = "accuracy\_RF"

df2\_RF = data.frame(matrix(unlist(acc\_history\_1\_RF), nrow=length(acc\_history\_1\_RF), byrow=T))

df2\_RF$group = 0

df2\_RF$i = seq.int(nrow(df2\_RF))

names(df2\_RF)[1] = "accuracy\_RF"

df3\_RF = rbind(df1\_RF,df2\_RF)

ggplot(data = df3\_RF, aes(x=i, y=accuracy\_RF, color=factor(group))) + xlab("round of test") +

ylab("accuracy score") + geom\_point() + labs(color="group")

t.test(accuracy\_RF~group, data=df3\_RF)

**Logistic and Random Forest for Recoveries:**

setwd("P:\\Spring 2020\\ML\\Final project")

#Loading all the required packages

library(caret)

library(tidyverse)

library(arsenal)

library(GGally)

library(party)

library(stringr)

library(dplyr)

library(ggplot2)

library(randomForest)

library(corrplot)

#Reading the file and inserting it into a variable

covid\_19 = read.csv('COVID19\_line\_list\_data.csv')

View(covid\_19)

summary(covid\_19)

## Extract the main symptoms of COVID-19

covid\_19$mainsymptoms = str\_detect(covid\_19$symptom, 'cough|fever|throat|breathlessness|dyspnea|pneumonia|malaise')

#Creating a subfile with only the essential variables

covid\_19 = covid\_19 %>% select(recovered,reporting.date, country, gender, age, visitingWuhan, fromWuhan, mainsymptoms, symptom) %>%

mutate(

recovered = ifelse(recovered == '0', 0, 1),

country = factor(country),

gender = factor(gender),

recovered = factor(recovered, label = c('0','1')),

reporting.date = as.Date(reporting.date, format = c('$d/$m/$Y')),

mainsymptons = factor(mainsymptoms)

)

#Plotting Age vs Recoveries

ggplot(covid\_19, aes(recovered, age, fill = recovered))+

geom\_boxplot()+

theme\_classic()+

scale\_fill\_manual(values = c('blue','red'))

#Plotting Gender vs Recoveries

ggplot(covid\_19, aes(recovered, fill = gender))+

geom\_bar(position ='fill')+

theme\_classic()+

scale\_fill\_manual(values = c('red','blue','black'))

#Most recoveries were in young patients

table1 = tableby(recovered ~ age + gender, total = FALSE, data = covid\_19)

summary(table1, text = TRUE)

#Creating a dataset with only the essential columns used to create the logistic model

covidLR = covid\_19 %>% select(recovered, gender, age, visitingWuhan, fromWuhan, mainsymptoms)

#Creating a train and a test set in the ratio 75:25 respectively

Train = covidLR[1:577,1:6]

Test = covidLR[578:825,1:6]

#Building the logistic regression model

log\_model = glm(formula = recovered~., family = binomial(link = "logit"), data= Train)

summary(log\_model)

#Predicting the recoveries for the Test dataset and comparing it with the Test dataset to determine accuracy of the logistic model

result = predict(log\_model, newdata=Test[,2:6], type="response")

result

result = ifelse(result>0.5,1,0)

result

accuracy = mean(result == Test$recovered)

accuracy

#Two models training/testing/comparison

num\_iterations = 500

acc\_history = list(num\_iterations)

acc\_history\_visitingwuhan = list(num\_iterations)

for (i in 1:num\_iterations) {

inTrain = createDataPartition(y=covidLR$recovered, p=0.75, list=FALSE)

X\_train = covidLR[inTrain, ]

X\_test = covidLR[-inTrain, ]

model = glm(formula = recovered~., family = binomial(link = "logit"), data= X\_train)

result = predict(model, newdata=X\_test[,2:6], type="response")

result = ifelse(result>0.5,1,0)

accuracy = mean(result == X\_test$recovered)

acc\_history[[i]] = accuracy

model\_1 = glm(formula = recovered~visitingWuhan, family = binomial(link = "logit"), data= X\_train)

result\_visitingwuhan = predict(model\_1, newdata=X\_test[,2:6], type="response")

result\_visitingwuhan = ifelse(result\_visitingwuhan>0.5,1,0)

accuracy\_visitingwuhan = mean(result\_visitingwuhan == X\_test$recovered)

acc\_history\_visitingwuhan[[i]] = accuracy\_visitingwuhan

}

##Printing average accuracy for 500 iterations

sum\_acc = 0

for (i in 1:num\_iterations) {

sum\_acc = sum\_acc + acc\_history[[i]]

}

ave\_acc = sum\_acc/num\_iterations

print(ave\_acc)

sum\_acc\_1 = 0

for (i in 1:num\_iterations) {

sum\_acc\_1 = sum\_acc\_1 + acc\_history\_visitingwuhan[[i]]

}

ave\_acc\_1 = sum\_acc\_1/num\_iterations

print(ave\_acc\_1)

#Welch test to comare the 2 models

df1 = data.frame(matrix(unlist(acc\_history), nrow=length(acc\_history), byrow=T))

df1$group = 1

df1$i = seq.int(nrow(df1))

names(df1)[1] = "accuracy"

df2 = data.frame(matrix(unlist(acc\_history\_visitingwuhan), nrow=length(acc\_history\_visitingwuhan), byrow=T))

df2$group = 0

df2$i = seq.int(nrow(df2))

names(df2)[1] = "accuracy"

df3 = rbind(df1, df2)

ggplot(data = df3, aes(x=i, y=accuracy, color=factor(group))) + xlab("round of test") +

ylab("accuracy score") + geom\_point() + labs(color="group")

t.test(accuracy~group, data=df3)

############################## Random forest ###############################

#Creating a dataset with only the essential columns used to create the RandomForest model

covid\_RF = covid\_19 %>% select(recovered, gender, age, visitingWuhan, fromWuhan, mainsymptoms)

#Creating a train and a test set in the ratio 75:25 respectively

Train\_RF = covid\_RF[1:577,1:6]

Test\_RF = covid\_RF[578:825,1:6]

##Building the RandomForest model

model\_RF = randomForest(recovered~., data=Train\_RF)

model\_RF

#Predicting the recoveries for the Test dataset and comparing it with the Test dataset to determine accuracy of the RandomForest model

prediction\_RF = predict(model\_RF, newdata=Test\_RF)

prediction\_RF

accuracy\_RF = mean(prediction\_RF == Test\_RF$recovered)

accuracy\_RF

#Two models training/testing/comparison

num\_iterations\_RF = 50

acc\_history\_RF = list(num\_iterations\_RF)

acc\_history\_1\_RF = list(num\_iterations\_RF)

for (i in 1:num\_iterations\_RF) {

inTrain\_RF = createDataPartition(y=covid\_RF$recovered, p=0.75, list=FALSE)

X\_train\_RF = covid\_RF[inTrain\_RF, ]

X\_test\_RF = covid\_RF[-inTrain\_RF, ]

model\_RF = randomForest(recovered~., data=X\_train\_RF)

model1\_RF = randomForest(recovered~visitingWuhan, data=X\_train\_RF)

prediction\_RF = predict(model\_RF, newdata=X\_test\_RF)

prediction1\_RF = predict(model1\_RF, newdata=X\_test\_RF)

accuracy\_RF = mean(prediction\_RF == X\_test\_RF$recovered)

accuracy1\_RF = mean(prediction1\_RF == X\_test\_RF$recovered)

acc\_history\_RF[[i]] = accuracy\_RF

acc\_history\_1\_RF[[i]] = accuracy1\_RF

}

for (i in 1:num\_iterations\_RF) {

print(acc\_history\_RF[[i]])

}

sum\_acc\_RF = 0

for (i in 1:num\_iterations\_RF) {

sum\_acc\_RF = sum\_acc\_RF + acc\_history\_RF[[i]]

}

##print average accuracy for 50 iterations

ave\_acc\_RF = sum\_acc\_RF/num\_iterations\_RF

print(ave\_acc\_RF)

sum\_acc\_RF\_1 = 0

for (i in 1:num\_iterations\_RF) {

sum\_acc\_RF\_1 = sum\_acc\_RF\_1 + acc\_history\_1\_RF[[i]]

}

ave\_acc\_RF\_1 = sum\_acc\_RF\_1/num\_iterations\_RF

print(ave\_acc\_RF\_1)

#Welch test to comare the 2 models

df1\_RF = data.frame(matrix(unlist(acc\_history\_RF), nrow=length(acc\_history\_RF), byrow=T))

df1\_RF$group = 1

df1\_RF$i = seq.int(nrow(df1\_RF))

names(df1\_RF)[1] = "accuracy\_RF"

df2\_RF = data.frame(matrix(unlist(acc\_history\_1\_RF), nrow=length(acc\_history\_1\_RF), byrow=T))

df2\_RF$group = 0

df2\_RF$i = seq.int(nrow(df2\_RF))

names(df2\_RF)[1] = "accuracy\_RF"

df3\_RF = rbind(df1\_RF,df2\_RF)

ggplot(data = df3\_RF, aes(x=i, y=accuracy\_RF, color=factor(group))) + xlab("round of test") +

ylab("accuracy score") + geom\_point() + labs(color="group")

t.test(accuracy\_RF~group, data=df3\_RF)