Correlations between

Bao  
*School of Information Science*  
*University of Pittsburgh* Pittsburgh, United States hab123@pitt.edu

# Abstract

For my final project, I decided to explore the features that may affect Coronavirus disease 2019 (COVID-19) death rate. The purpose of my project is to discover if a person’s age, gender, geographical location, and time he got COVID-19 will affect the COVID-19 death rate. The reason why I chose this topic is because this global pandemic affects me a lot. The COVID-19 makes the plane ticket much more expansive than usual, which stops me from going back home during the summer vacation. Even now, China is still insisting a “zero-COVID policy”. I am wondering if this policy is actually necessary and whether COVID-19 is that severe and fatal after the vaccines are widely distributed.

My plan of analyzing the COVID-19 death data is to get my sample data from Centers for Disease Control and Prevention (CDC) and use the features of “Sex”, “States”, “Year”, “Month”, and “Age Group” to predict the COVID-19 Death rate. With the result, I should be able to predict a patient’s death rate with his personal information.

# Introduction

COVID-19 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2. Symptoms of COVID-19 contains fever, cough, headache, fatigue, breathing difficulties, loss of smell, and loss of taste. Of people who got COVID-19, 81% of them only have mild to moderate symptoms, while 14% of them have severe symptoms and 5% of them have critical symptoms. Up to now, there are 578 million COVID-19 cases and 6.4 million people died because of COVID-19 around the world.

I chose five features as my predictors. They are “Sex”, “States”, “Year”, “Month”, and “Age Group”. I chose the feature “Sex” and “Age Group” because I have heard that males have a higher infection and death rate compared to females, and older people have a higher death rate after being infected than younger people. I chose the feature “State” because I want to see if a state with greater performance of health care has a smaller COVID-19 death rate. For example, if Hawaii, the top state for health care, has a smaller death rate compared to Mississippi, the state that is considered as the worst state for health care. I chose the features “Year” and “Month” because I want to check if the development and distribution of COVID-19 vaccine helps the COVID-19 death rate. In other words, if the year of 2022 has a smaller death rate compared to the year of 2020, when the COVID-19 just started spreading.

# Methodology

The data set includes COVID-19 deaths, pneumonia deaths, pneumonia and COVID-19 deaths, influenza deaths, pneumonia, influenza, or COVID-19 deaths, and total deaths. Since I only want COVID-19 deaths as my response value, I will get rid of all other columns. I deleted “Data As If”, “Start Date”, “End Date”, and “Group” as well and only leave my predictors because I am not interested in these data. I thought of getting rid of the “Month” column and only leave the “Year” column, but I figured it would be good to see how the death rate changes specifically after the vaccine came out and was distributed, so I decided to separate the time by months.

Since the time was separated by three groups, “by total”, “by year”, and “by month”, the state was separated by two groups, a “United States” group that contains all the states and each different states separately, the sex was separated by two groups, “All gender” that contains all genders and two general genders, and the age group was separated by two groups, “All Ages” that includes all range of ages and different range of ages, so there are many repeated data. To get rid of the repeated data, I deleted all rows that contains “by total”, “by year”, “All Ages”, “All gender”, and “United States”.

After getting rid of all repeated data and data that I do not want for my model, I check for missing values. It turns out that the COVID-19 Deaths column has about half of the data that are missing. From the footnotes, I noticed that missing values have counts between 1 to 9, and these values are hide because of the NCHS confidentiality standards. I cannot delete them since these values are not missing at random, so I will replace the missing values with median and mean, since the median and the mean are the same, which is 5.

To find outliers, I deleted all the commas in COVID-19 Death columns and changed the data type to integer. Then I use the box plot to check for outliers. I did find some outliers that seems extremely high compared to other data. For example, 75 – 84 years old males have a COVID-19 deaths of 2944, which is the highest number of deaths among all months and years, in January, 2021. I looked up another dataset called “COVID-19 Weekly Update” made by Country of San Diego and I confirmed that the outliers that I found in my data frame are not statistical error, but real data, so I will keep all the outliers.

Since all states have different population, state with greater population will more likely to have greater number of COVID-19 death. For example, California, the state that has the greatest population, will have more COVID-19 death number compared to Wyoming, the state that has the lowest population, while the COVID-19 death rate may not have a large difference. In order to calculate the death rate, I duplicate the COVID-19 Death column and create a new column called “COVID-19 Death Rate”, then divide the numbers by the number of populations for each state.

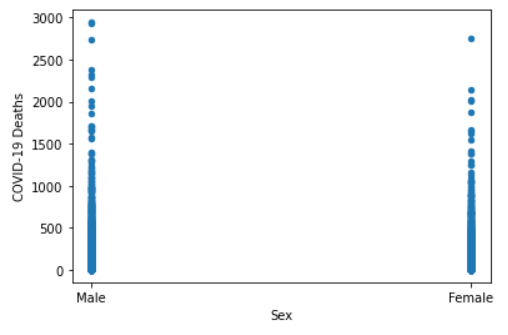
Finally, I categorized all of my predictor columns. I categorized the State column by their rank of health care. I found the ranking list in US news website. The top half on the best states for health care list, which are the states with good health care quality, will be 0, and the rest will be 1, meaning these states have a poor health care quality. I categorized the time by the releasing date of COVID-19 vaccine. The vaccine came out on August 23, 2021. Since it came out on late August, I will set all time before September, 2021 as 0, which means no vaccine, and after September, 2021 as 1, which means vaccine was released. I categorized the Sex column by gender. Males will be 0 and females will be 1. And I categorized the Age Group column by young and old. I will set under 1 year to 30-39 years as 0, which means young, and 35-44 years to 85 years and over as 1, which means old.

For the final form of my data, I chose to use the random forest algorithm to predict my data set because random forest algorithm has a generally high accuracy, and it is more efficient when dealing with large data set. Since I'm dealing with 52576 data sets, and I am also interested in the importance score ranking of each predictor, I believe random forest algorithm can be useful.

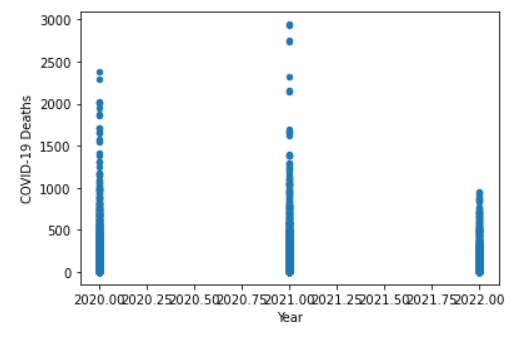
# Result

Before starting the random forest algorithm, I create four scatter plots for my predictors. They are Sex, Age Group, Year vs. COVID-19 Deaths, and State vs. COVID-19 Death Rate. The scatter plots are listed below.

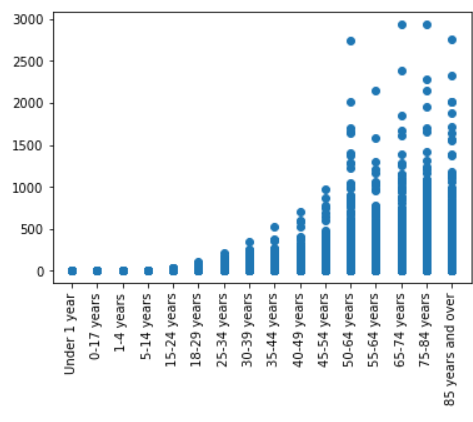
Sex vs. COVID-19 Deaths



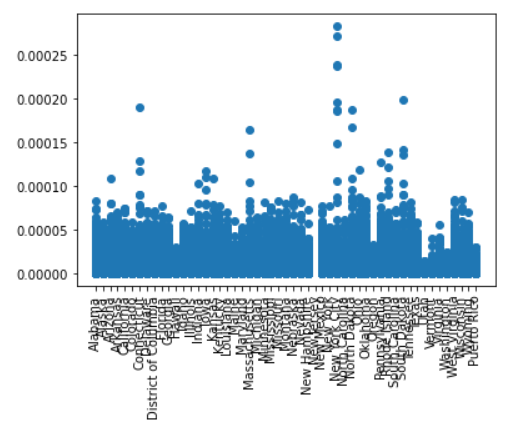
Year vs. COVID-19 Deaths



Age Group vs. COVID-19 Deaths



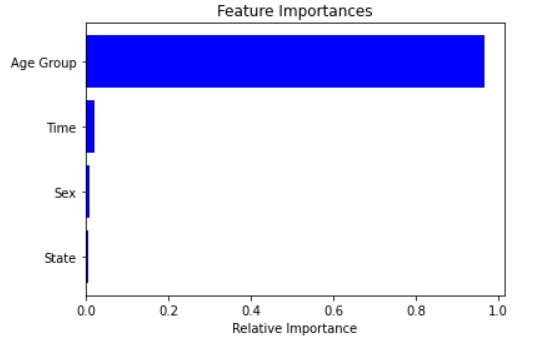
State vs. COVID-19 Death Rate

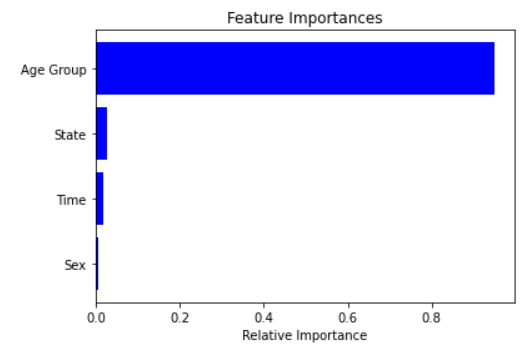


Just by looking at the scatter plot, I was assuming that both Sex and Time will affect the COVID-19 Death a little, and Age Group will strongly affect the COVID-19 Death. The State should affect the COVID-19 Death Rate a bit. For example, the graph shows that Hawaii, the state that is considered as the best state for health care, has a fairly small death rate. New York City, where is considered one of the worst places for health care, has two highest COVID-19 Death Rate dot that are above 0.00025.

The Random Forest output after I trained the final form of my data is posted below.

Age Group, Sex, State, Time vs. COVID-19 Deaths



Age Group, Sex, State, Time vs. COVID-19 Death Rate

Apparently, for both COVID-19 Deaths and COVID-19 Death Rate, the most important feature is Age Group. In fact, the Age Group has a relative importance of 0.965119 for COVID-19 Deaths and 0.94856 for COVID-19 Death Rate. From the feature importance graph and the Age Group vs. COVID-19 Deaths scatter plot, we can say that the age hugely affects the COVID-19 death, and older people have a higher death rate compared to younger people. In the COVID-19 Death Rate graph, we can also see that State has a greater relative importance compared to time and sex, showing that the state with better health care has a slightly less death rate. In the COVID-19 Deaths graph, we can see that Time has a greater relative importance compared to sex and state, showing that the release of vaccine slightly affect the COVID-19 deaths. Sex barely affects the COVID-19 deaths and COVID-19 death rate. Sex has the smallest relative importance in COVID-19 death rate model and is the second to the last relative importance in COVID-19 deaths. However, different states have a large difference of population. For example, California has a rank of 5 on the list of health care. And the graph of State vs. COVID-19 Death Rate shows that California has a relatively small rate of death. Nevertheless, California has four of the highest COVID-19 Deaths number of 2944, 2939, 2756, and 2739 due to its huge population base. So the prediction of COVID-19 Deaths number by the feature of state will not be accurate. Sex, in fact, has the smallest relative importance in both COVID-19 Deaths and COVID-19 Death Rate model. My prediction before training the model is basically correct.

# Discussion

The initial visualization results were got before any machine learning process. The initial results I got from the visualization mostly agreed the result I got after running the machine learning, which is the random forest algorithm: the Age Group will strongly affect the COVID-19 Deaths, both State of whether having a good health care condition and Time that is separated by having vaccine or not will slightly affect the COVID-19 Deaths and COVID-19 Death Rate, and Sex will not affect the COVID-19 Deaths at all.

Before using the Random Forest Algorithm, I tried the Decision Tree as well because decision tree has an advantage that the missing values won't affect the process of building a decision tree very much. I thought decision tree will give me a great prediction since I have about half of the values are missing on the COVID-19 Deaths column. Despite I replaced them with the mean, which is 5, there is probably still a large error. But decision tree did not give me a good prediction, so I finally chose to use the Random Forest Algorithm to train my data frame.

Although the prediction and the result from Random Forest algorithm fitted, I found the accuracy score of both of the machine learning process are disappointed. Accuracy is the fraction of predictions the model got right. I got a 0.4330543933054393 for random forest that predicts the COVID-19 Death Rate and 0.5111259033853176 for random forest that predicts the COVID-19 Deaths. Both of the accuracy scores are way less than 70%, meaning I got a poor model. Despite of the disappointed accuracy score, I can still find and take some of the meaningful results from the model. For example, the Age Group is about 96.51% of relative importance to COVID-19 Death. The relative importance is so high so we can ignore the poor accuracy for this feature. The Age Group is definitely worthy of further investigation.

My result has a low accuracy score might because I did not choose my features well. For example, after running the random forest, I noticed that I may should not use Sex as my predictors, and this feature, in fact, has no relation with my response variable.

# References

National Center for Health Statistics, “Provisional covid-19 deaths by sex and age,” Centers for Disease Control and Prevention, 01-May-2020. [Online]. Available: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku. [Accessed: 03-Aug-2022].

County of San Diego, “Covid-19 watch - San Diego County, California,” COVID-19 Weekly Update, 28-Jul-2022. [Online]. Available: https://www.sandiegocounty.gov/content/dam/sdc/hhsa/programs/phs/Epidemiology/COVID-19%20Watch.pdf. [Accessed: 03-Aug-2022].

B. Ziegler, “Best states for healthcare | US news best states,” Health Care Rankings Measuring: how well states are meeting citizens' health care needs, 2022. [Online]. Available: https://www.usnews.com/news/best-states/rankings/health-care. [Accessed: 03-Aug-2022].

1keydata.com, “Alphabetical list of 50 states,” Alphabetical List Of 50 States, 2020. [Online]. Available: https://state.1keydata.com/. [Accessed: 03-Aug-2022].

U.S. Census Bureau, “U.S. Census Bureau quickfacts: New York City, New York,” 2022. [Online]. Available: https://www.census.gov/quickfacts/newyorkcitynewyork. [Accessed: 04-Aug-2022].