# Correlation Analysis of Fatality Rate through COVID-19 Data Visualization

**Yuho Jeong***
yuho8437@unist.ac.kr
UNIST
School of Mechanical, Aerospace and Nuclear Engineering
Ulsan, Republic of Korea

**Daehyeon Nam***
ndh8392@unist.ac.kr
UNIST
School of Electrical and Computing Engineering
Ulsan, Republic of Korea

## ABSTRACT

It has already been six months since the COVID-19 began, but the disease is still unabated. There are a number of differences in different countries for this terrible disease, which accounts for 390,000 cumulative deaths worldwide. Among them, we wanted to know why COVID-19 fatality rates vary from 1 to 18% in each country.

For this, factors predicted to be related to COVID-19 were selected and correlation analysis was conducted. Factors were classified into point data type and time-series data type considering differences between data types. Visualization was also divided into two types. For the point data type factor, the factor value is set on one axis, and the factor of each country is expressed on the 3D scatter plot. For the time series data type factor, how the fatality rate varies according to the value of the factor for each country is expressed on the world map.

The result of our study showed that 5 factors which are ratio of female population, age, ratio of female smokers, day length and precipitation have high correlation value with fatality rate of COVID-19.

## KEYWORDS

COVID-19, Fatality rate, Correlation analysis, Point data type, Time-series data type, Information visualization

## 1 INTRODUCTION

Since the outbreak of COVID-19 in Wuhan in December 2019, it began to spread to China's neighboring Asia in the early days and has now became into a pandemic disease that has spread all over the world. Accordingly, the WHO declared a "pandemic" on March 11.

The fatality rate of covid-19 varies from country to country. Representatively, countries such as Korea, Japan and China have relatively low fatality rates of 2 to 5%, while countries such as the United Kingdom, Italy and France have high fatality rates of over 14% (Figure 1).

So, after we checked the difference in fatality rates between countries, we wondered what factors affect to fatality rate, in other words, why each country shows the difference in fatality rate. We thought if we can identify factors that affect fatality rate, we will be able to provide hints to lower the fatality rate and predict countries with potential risks in the future that are just increasing the number of confirmed people.

Thus, in this research, we aim to investigate the correlation between multiple factors and fatality rates in different countries and infer factors that increase mortality.

The article is organized as follows. In the section 2, we introduce the related researches which have done, followed by the method how we analyzed and the reason why we chose the specific visualization to show in the section 3. Section 4 introduces how we implemented visualization including the data analysis and, in section 5, we show how users will be able to use our visualization. In section 6, we discuss about the result we brought and finally, in section 7, we present future work and conclusion.

## 2 RELATED WORK

We wanted to know the relationship between COVID-19 fatality rate and factors. This requires methodological knowledge of correlation analysis and analysis of fatality rate factors of COVID-19. We should also be aware of analysis of the other diseases. We gained these insights from the materials mentioned in the reference.

### Analysis the cause of death from COVID-19

The papers in this group analyze the cause of death from COVID-19 based on the data from regions which have a lot of data about COVID-19, such as Wuhan(China) or Italy, or where serious problems occur. They talk about the effects of complications, BCG vaccination, and the age on the deaths of patients. They identify key factors affecting fatality rate and talk about what people need to be aware of.

The goal of our project is to identify factors that have a major impact on the fatality rate for COVID-19. We will perform correlation analysis on the fatality rate for various factors. So, we need to select several factors initially that are expected to be important. That's why we need to know what factors the medical community says have a major impact on fatality rate. For example, Mohammad Pourhomayoun

*Both authors contributed equally to this research.

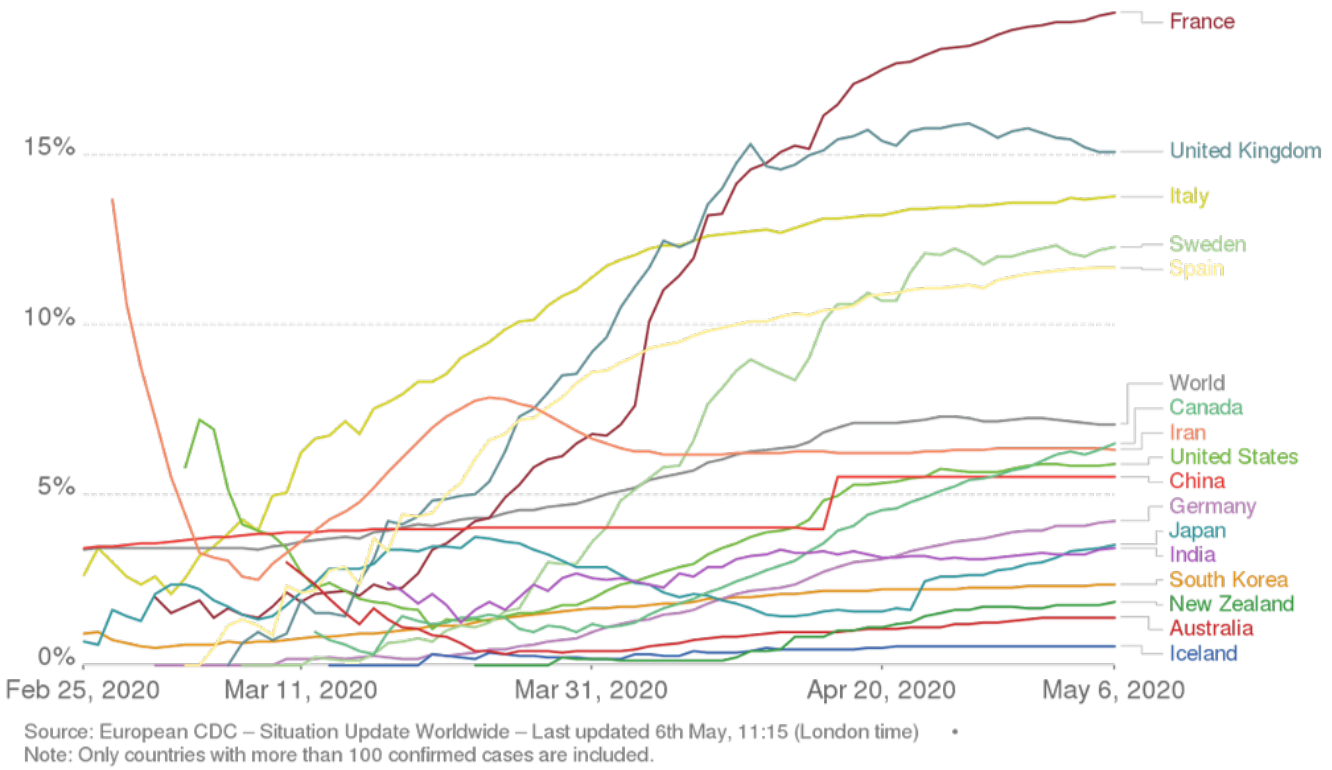


**Figure 1: Case fatality rate of the ongoing COVID-19 pandemic: The table shows the fatality rate(percentage of deaths by COVID-19 / confirmed by COVID-19) by country over time. The world average fatality rate is 7.3% and mortality rate varies greatly from country to country.**

computed correlation coefficient between complications of patients and fatality rate for prediction fatality risk in patients with COVID-19 [12]. They analyzed that the correlation with fatality risk was high for age and chronic disease. We refer to these results in project implementation.

### Correlation and Visualization

The papers in this group provides us some methods to analyze and visualize information. Since we will be using the correlation for extracting potential factors for fatality rate of COVID-19, the method for extracting the key factors, and how to visualize this information are important. In [14], it describes how we can effectively select the features in high-dimensional data. As we will use diverse and large number of factors for find the key factors, we can use this method for selecting some of key features among them. And in visualization papers, it showed diverse manners of showing correlation and clustering. We got inspired by the interactive visualization of correlation introduced in [17], and we will use visualization that shows information when the cursor is over a data point by applying the introduced method.

### Analysis of the causes of other diseases

The authors of the papers in this group basically wanted to know the association between mortality with the diverse factors. And they used correlation coefficient for indicating the association between the fatality and factors that presumed to be the cause. Given that similar studies have done in advance, it shows that it is feasible for using correlation coefficient to investigate the association with factors using correlation coefficient for COVID-19. And there were some visualization methods that we can refer for our research. In [19], they visualized the correlation coefficient on the map so that users can compare the impact of the factors among the different regions in America. We might be able to use this visualization method for showing correlation between fatality rate of COVID-19 and factors among the countries. But the main distinction of our research will come from finding association from diverse factors. We not only use the factors which is point data as the papers above done, but also use the time-series data as a factors to consider diverse factors. And also analyzing method is distinct in our research. Even

though we will use correlation concept, we will analyze the data using interactive visualization of data to show it easily.

## 3 APPROACHES AND METHODS

We classified the factors for correlation analysis into two types. One is point data type and the other is time-series data type. Point data type factors include BCG, GDP per capita, CVD death rate, and time-series data types include the number of tests, the number of cases, UV radiation, and temperature. We calculated the correlation coefficient with fatality rate for two factors. The analysis was conducted only for countries with more than 3,000 confirmed cases (approximately 60 countries). This is because the analysis of the country where a significant amount of confirmed cases exist would lead to meaningful results. For point data type factors, they were classified into countries with high fatality rates and countries with low fatality rates, and the control point of case-fatality-rate was set at 7.5%. Referring to the graph below (Figure 2), when classified into two, the interval between them was set to be significant.

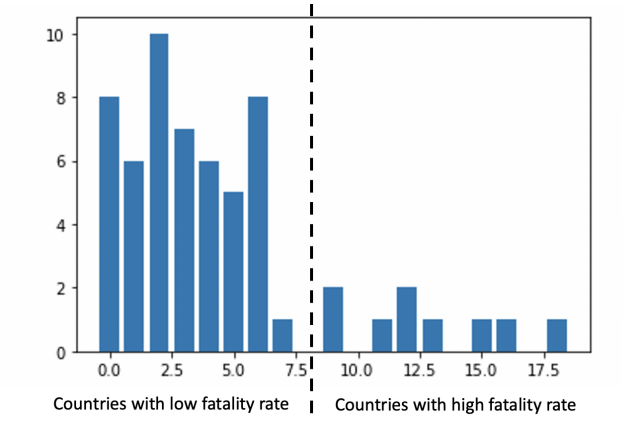


**Figure 2: Bar charts visualization of case fatality rate (COVID-19):The bar charts shows the global case fatality rate. Vertical axis is about the number of countries and horizontal axis is about the case fatality rate (%)**

### Visualization methods

About point data type factors, the users guess which factor has the greatest correlation with the fatality rate. And select three of the factors. The three are represented by the axis of the 3D graph, and each country is output as a point of the 3D graph. Countries with high fatality rates are represented by red dots, and countries with low fatality rates are represented by blue dots. Whenever the users change the factor, the points on the graph change interacting with users. The view of graph can be changed by dragging the cursor. Using this visualization, we would like to easily identify which of the point data type factors are highly correlated with the fatality rate. And based on this result, we will determine whether predicting the fatality rate using only the point data type factors has a meaningful output. In our expectation, if users pick the three most influential factors, the red dots will converge to one space.

About time-series factors, we wanted to let users try to search factors that have high correlation with fatality rate by themselves. We used world map with heat map that shows the current fatality rate of each country. So, countries with darker color have higher fatality rate than other countries with lighter color. The visualization has empty box on the left side which shows scroll box filled with various time-series factors when a user clicks it. And on the bottom, there is a slider, so that user can change the value of factor. And on the map, it shows countries that have been had the factor value with color according to the current fatality rate. Thus, if a user choose one factor and slides the value of it from lower value side to the higher side, and if a user can see the map is filled with darker color, it means the factor the user chose has positive relationship with fatality rate. So, by using this, users can guess which factors have high co-relationship with fatality rate and the mechanism that increases the fatality rate of COVID-19.

### Analysis methods

Correlation analysis is a method of analyzing what linear or nonlinear relationship exists between two variables. The strength of the relationship between the two variables is called the correlation coefficient. The result is a value for -1 to 1, and the higher absolute value means the stronger relationship between the two variables. We calculated the correlation coefficient for the fatality rate with the point data type factors and the time-series data type factors. We used the Pearson method which is built-in function in python pandas for calculation. For point data, we mapped each of country with its own fatality rate and factor value. And we calculated the correlation value between these fatality rates and factor values. For the time-series data, we divided a range from minimum to maximum value of a factor into 40 ranges. And from the time-series data we collected the number of confirmed cases and that of death based on the days with each factor value range. After that we calculated correlation value between factor range values and fatality rate. Through these results, we tried to identify factors that have a significant effect on COVID-19.

## 4 IMPLEMENTATION

### Visualization implementation

We used JavaScript library to simplify visualization of data. And for user interaction, we used HTML and React.

*Point data visualization.* In order to scatter point data in 3D space, "3D Scatter Plots" provided by Plotly were used. The functions of this library are follows. It projects points in 3D space just by putting the coordinate data of the x, y and z axis of the points. It allows a user to adjust the viewing angle of the 3D scatter plot by dragging and to zoom the view in and out through scrolling. And finally, it shows the information about the point when the user moves the cursor over the point data of the 3D scatter. The user interface, such as scroll bar, except 3D scatter Plots provided by Plotly, was implemented using another JavaScript library, React. We implemented a scroll bar beside to the 3D scatter visualization where there is a list of factors so that user can directly select the types of factors which will be used in visualization. Thus, when users select 3 factors from the scroll bar as shown in figure 3, the 3D scatter visualization right next to it will show users the location of point data according to the correlation value of selected factor.

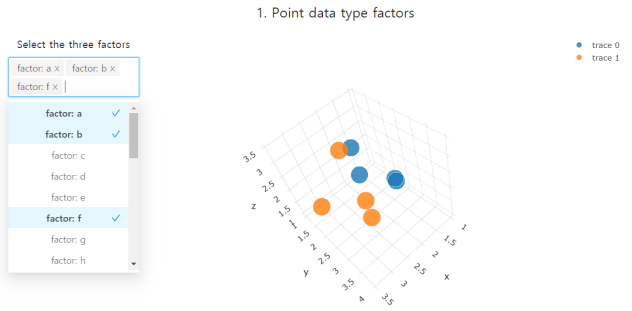


**Figure 3: 3D scatter plot visualization example: It visualizes the correlation value of selected factor (a,b and c) on the 3D space using 3D scatter plot from Plotly.**

*Series data visualization.* "Choropleth maps" provided by Plotly was used to express time-series data on a world map with heat map. The functions that this library provides are follows. It visualizes the data as a form of world map just by putting the data of each country. And it also allows the user to adjust the view of the world map by dragging and to zoom the view in and out through scrolling. Furthermore, it shows the information about the fatality rate of the country when a user moves the cursor over the country on the world map. The user interface except world map was implemented using React. When a user click a empty box on the left side, scroll bar shows the list of factors that are available to be selected. And we have another slider on the bottom which is also implemented using React. After a user selects a factor, user will be able to select factor's degree using the slider. If a user changes the value of the factor, the map shows countries that fall under the category with the color that shows fatality rate of each country. For example, if a user selects temperature, and selects 20°C, the map shows countries where the temperature had been 20 degrees with a color of each country.

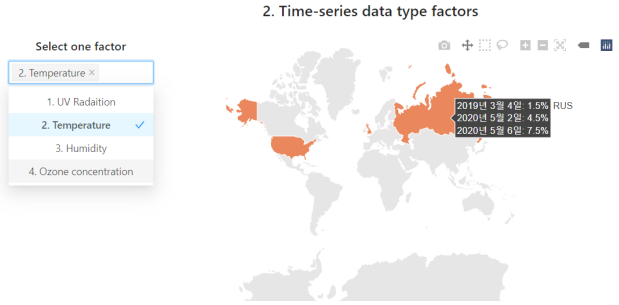


**Figure 4: World map with heat map visualization example: It visualizes the correlation value of selected factor (2) on the world map with heat map from D3.**

## Analysis implementation

The data we got from various sources usually includes data that we did not need. So, to leave the data that we need only, we mainly used a module named Pandas from Python to process the raw data.

*point data analysis.* We used 15 types of point data which include population density [PD], GDP per capita [GDP], CVD death rate [CVD], the availability of handwashing facilities of home [HW], the number of hospital beds per 100k [HB], doctor consultations per capita [DC], voluntary health spending ratio(% of GDP) [HS], gender ratio about confirmed case(% male) [SEX], BCG(% of one-year-old children) [BCG], air pollution [AIR], the ratio of age 65 years and older [A65], the ratio of age 70 years and older [A70], the rate of female smokers [FSM] and the rate of male smokers [MSM]. Using Python, we left only the Iso code and each of factor values for 68 countries, and deleted all other data we didn't need. After this data processing, we calculated the correlation value between these factor values of each country and fatality rate values of each country using pandas correlation function. The result of calculation of absolute correlation values of each factor is shown in figure 5. We can notice that relatively the absolute correlation values of SEX, A70, FSM are high. While SEX has negative values, A70, FSM showed positive values of correlation value.

*series data analysis.* We used 8 types of series data which include day length [DL], precipitation [PRI], pressure [PRE], wind speed [WIN], ozone density [OZ], temperature [TEM], humidity [HUM], ultraviolet [UV]. We only left factor data,

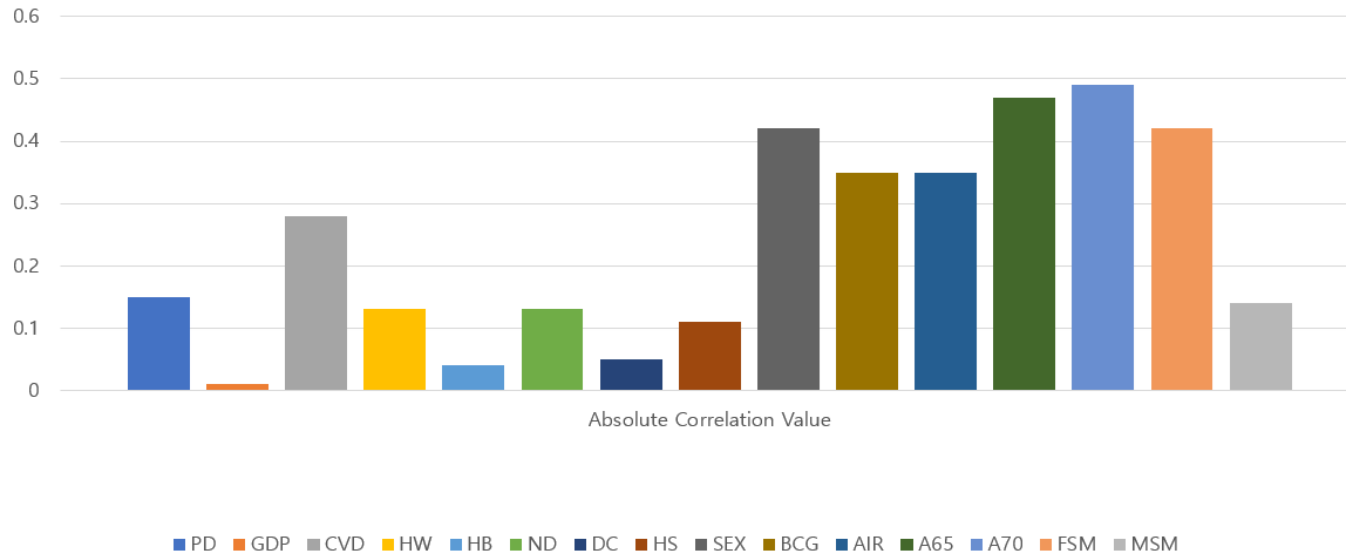


**Figure 5: Bar chart of absolute correlation value of point data: The bar chart shows the calculated absolute correlation value of factors. We can notice that gender ratio, BCG rate, air pollution, age, and the ratio of female smoker have relatively high absolute correlation value.**

iso code and date for 68 countries without any other confusing data. After we processed these data, we calculated the minimum value and maximum value of factor data and We divided minimum to maximum value to 40 pieces. And for each of piece, we searched which country's factor value on which day corresponds to this piece. After that we calculated correlation of average daily fatality rate of each piece from according countries with average factor value of each piece. The result is shown in figure 6. The figure 6 shows the absolute value of correlation. As we see, day length, and precipitation have relatively high absolute correlation value which are 0.48 and 0.42 each. The correlation value of day length is positive while that of precipitation is negative.

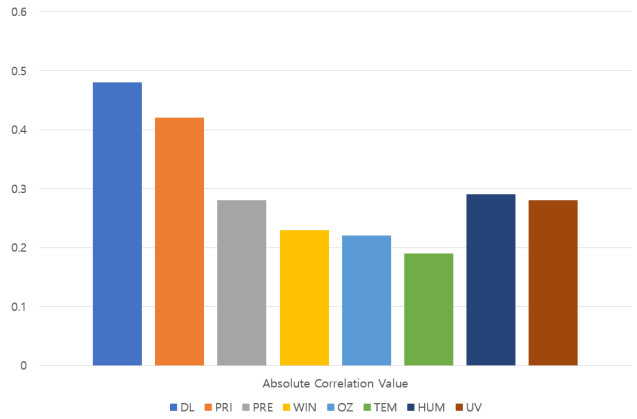


**Figure 6: Bar chart of absolute correlation value of series data The bar chart shows the calculated absolute correlation value of factors. We can notice that specifically day length and precipitation showed relatively high absolute correlation value.**

## 5 USE CASE

You can check our work at the following online link: https://yuho8437.github.io/covid-correlation/ (it takes a few seconds for the entire content to be loaded).

*point data visualization.* We assumed that, after finding the factors with high correlation coefficient, and showing the results on the 3D scatter plot, countries with high fatality rate and countries with low fatality rate will be well clustered. So the user can play a game of finding the factors with the highest correlation efficiency among the factors. For example, if a user selects three factors that are expected to have the highest correlation coefficient with the fatality rate, and if countries with high fatality rate and countries with low fatality rate were not clustered well, the user can select other factors to find the best clustering through trial and error. Through the process, the factors that make the best clustering are finally determined and the result can be compared by pressing the 'correlation result' button.

If a user selects the three factors with the highest correlation coefficient as shown in the figure, clustering will be done well. However, if a user selects factors with low correlation coefficient, clustering will not be done well. If the clustering is fine, we can infer the characteristics of the well-clustered

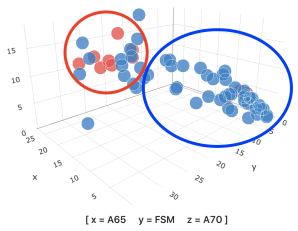


**Figure 7: Output when three factors with the highest correlation coefficients are selected**

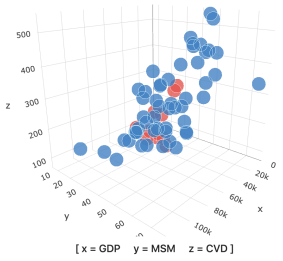


**Figure 8: Output when three factors with a low correlation coefficient are selected**

countries based on factors. Also, even if there are countries with low fatality rate in the cluster of high fatality rate, we can predict that these countries will have a higher fatality rate in the future. A user can also use hover to easily figure out countries information. In addition, the graph is free to enlarge or reduce, and it is easy to move and rotate through the mouse.

*series data visualization.* We calculated the correlation coefficient of the time series data type factor in a different way than the point data type factor. Therefore, visualization was implemented in a different way. We made the assumption that if a user selects factors with high correlation coefficients, the factor values and the color of the fatality rate heat map will have the same or opposite tendency. So, when a user adjusts the value of the factor using the slider at the bottom, the color of the fatality rate heat map also changes accordingly. Through our visualization, a user can infer the factors with the most similar tendencies of the value of factors and the fatality rate heat map. The actual calculation result can be checked by pressing the 'correlation result' button.

## 6 DISCUSSION

We found several meaningful factors to explain fatality rate pattern of global. Since we considered any factors which have less than 0.4 correlation value is meaningless, we excluded them. The factors left from point data are SEX which is gender ratio, A70 which is the rate of age over 70, and FSM which is female smoking ratio. Except factor A70 which is self-evident for the reason why it showed high correlation, we cannot explicitly explain the reason why those factors show high correlation which are higher than 0.4. However, since both of factors are related to the gender ratio, we could guess that the gender ratio matters for fatality rate. Additionally, as the correlation value of SEX is negative, this result means the higher the proportion of men, the higher the mortality rate. However since the correlation value of FSM is positive, it means the higher the proportion of women who smoke, the higher the mortality rate.

In series data, we could figure out that day length and precipitation data showed absolute correlation values which are higher than 0.4. As correlation value of day length is positive, it can be interpreted that the longer the daytime length, the higher the car accident rate. However, since correlation value of precipitation is negative, this means that the less rain, in other words, the clearer the weather, the higher the mortality rate. Also, we could not accurately explain the reason why these factors showed higher absolute correlation value. However, what we can carefully guess as a result of this is that the fatality rate was high on days when the weather was clear or the length of the day was great for people to do outside activities.

## 7 FUTURE WORK AND CONCLUSION

We have analyzed factors high relation on fatality rate through correlation anlysis in this study. However, we should not believe only the results of correlation analysis. We need to determine whether the result of correlation analysis is a significant value for our hypothesis. To do this, we need to go through a process of re-validation using methods such as statistical significance or analysis of variance (ANOVA). Through this processes, we get a estimate of the factors dominant to the fatality rate. Also, causalty cannot be explained by correlation analysis. If there are any two values A and B, the correlation coefficient between A and B is high, so it cannot be concluded that A is the cause of B or B is the cause of A. In this case, a causality between the two need to be analyzed using a method such as granger causalty. These are the future work of our study.

The high correlation coefficients for point data type factors include age, female smoker rate, and gender ratio. The results of the age factor were expected, but the results of the female smoker rate and gender ratio were unexpected. Since both factors are related to gender, it can be assumed that there is a relationship between gender and fatality rate. So it would be better if there were more studies analyzing the relationship between COVID-19 and gender.

In the time-series data type factor, high correlation coefficients include day length and precipitation. In general, it is known that COVID-19 spreads between people with saliva

or droplets. Therefore, we expected that the correlation coefficient of the factor which is implicitly implying human contact will be high. An example is the assumption that short day lengths and high precipitation will result in fewer people going out, which will have less risk of COVID. It would be a good to further analyze people's outdoor activities according to day length and precipitation.

## 8 REFERENCE


### Analysis the cause of death from COVID-19

[1] MD. Fei Zhou, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, March 2020
[2] G. Onder, Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy, March 2020
[3] DD. Rajgor, The many estimates of the COVID-19 case fatality rate, March 2020
[4] H. Martha, Coronavirus: Why death and mortality rates differ, April 2020, BBC
[5] M.Aaron, Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study, March 2020
[6] X.Wu, Exposure to air pollution and COVID-19 mortality in the United States, April 2020
[7] Constantine I. Vardavas, COVID-19 and smoking: A systematic review of the evidence, March 2020
[8] CDA Analytics Team, COVID-19 Data Analysis, Part 1: Demography, Behavior, and Environment, March 2020
[9] Unknown author, Age, Sex, Existing Conditions of COVID-19 Cases and Deaths, February 2020
[10] N.Wilson, Case-Fatality Risk Estimates for COVID-19 Calculated by Using a Lag Time for Fatality, March 2020
[11] L.Wang, Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm, April 2020
[12] M.Pourhomayoun, Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making, April 2020
[13] L.Jia, Prediction and analysis of Coronavirus Disease, March 2020

### Correlation and Visualization

[14] L.Yu, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, March 2003
[15] S.George, The 5 Clustering Algorithms Data Scientists Need to Know, Feburary 2018
[16] SK.Lodha, Visualizing Health Determinants in a Global Context, January 2008
[17] R.Veronika, Best Charts to Show Correlation Analysis of the causes of other diseases, July 2019

### Analysis of the causes of other diseases

[18] H.Murphy, Differences in predictions of ODE models of tumor growth, February 2016
[19] C.Tencza, Factors responsible for mortality variation in the United States: A latent variable analysis, July 2014
[20] LI.Horwitz, Correlations among risk-standardized mortality rates and among risk-standardized readmission rates within hospitals, August 2012

### Datasets

[21] Out World in Data, https://github.com/owid/covid-19-data/tree/master/public/data
[22] OECD, https://data.oecd.org/
[23] GlobalHealth5050, https://globalhealth5050.org
[24] WHO, https://www.who.int/data/gho/data/
[25] Kaggle, https://www.kaggle.com/juanjodd/uv-radiation-and-covid-19-global-confirmed-cases/data