The State of COVID-19 Patients in South Korea

Amy Tan

Pittsburgh, PA

amt147@pitt.edu

***Abstract*—I have chosen to research the outcomes of people who have contracted COVID-19 (Coronavirus Disease 2019). As a Chinese person in the midst of this viral outbreak, I have felt pretty strongly about the virus due to its origin in China and the public’s reaction to it. Since COVID-19 is spreading widely and quickly, I thought it would be useful to look at patterns of who it has been affecting the most. Anyone who is planning on travelling to heavily infected areas or anyone who is simply worried about the virus could somewhat benefit from my model. I plan to look at datasets with patient information and analyze the effects of the virus on the patients based on variables such as proximity to certain areas (location), age, gender, and the date that the patient contracted the virus (telling of how long it has been since the patient was confirmed to have the virus).** (*Abstract*)

Keywords—COVID-19, coronavirus, South Korea

# Introduction

A novel coronavirus broke out in 2019, first identified in Wuhan, China. As a new coronavirus that had not been previously seen circulating among humans, there is much unknown about the long-term effects of the virus. The World Health Organization announced that COVID-19 would be the official name for the disease in February of 2020.

The virus spreads person to person through respiratory droplets that are produced when an infected person coughs or sneezes. COVID-19 has spread quickly and efficiently and has affected some countries more than others. South Korea experienced rapid spread after a patient became a super-spreader (someone who passes the infection to a larger number of people). The patient had travelled to many crowded areas, gotten involved in a minor traffic accident, was admitted into a hospital, and attended church services. Within ten days, the number of recorded cases in South Korea was almost 10 times larger than before.

# Methodolody

## The Dataset

The dataset used gives the epidemiological data of COVID-19 patients in South Korea. It was made based on the report materials of the Korea Centers for Disease Control & Prevention (KCDC) and local governments. The dataset gives:

* *patient\_id*: the ID of the patient
* *global\_num*: the number given by KCDC
* *sex*: the sex of the patient
* *birth\_year*: the birth year of the patient
* *age*: the age of the patient
  + 0s: 0 ~ 9
  + 10s: 10 ~ 19

…

* + 90s: 90 ~ 99
  + 100s: 100 ~ 109
* *country*: the country of the patient
* *province*: the province of the patient
* *city*: the city of the patient
* *disease*: TRUE: underlying disease / FALSE: no disease
* *infection\_case*: the case of infection
* *infection\_order*: the order of infection
* *infected\_by*: the ID of who infected the patient
* *contact\_number*: the number of contacts with people
* *symptom\_onset\_date*: the date of symptom onset
* *confirmed\_date*: the date of being confirmed
* *released\_date*: the date of being released
* *deceased\_date*: the date of being deceased
* *state*: isolated / released / deceased
  + isolated: being isolated in the hospital
  + released: being released from the hospital
  + deceased: being deceased

## Cleaning the Data

*patient\_id*, *global\_num*, *infection\_order*, *infected\_by*, *released\_date*, and *deceased\_date* do not give us much insight on deciding whether the patient is released, isolated, or deceased, so I removed those columns. I also removed *country* because location is given by *province*/*city*.

A close up of a logo

Description automatically generated

Fig. 1. A heatmap showing all of the features. Purple represents data that is not null or missing, and yellow represents data that is missing from the dataset.

*infection\_case*, *contact\_number*, *symptom\_onset\_date* have way too much null data, so I dropped those columns as well. Factors like a patient's sex, age, location, and confirmed date are probably more indicative of a patient's state anyway. *disease* would be a really telling factor for determining the state of a patient, but there is way too much missing data in that column as well (less than 20 real data entries out of over 2000 in that column), so I dropped it. I also removed city because we can use *province* as the factor for location.

Since all the rows that have null ages also have null birth years, I could not fill in this column with precise data. I dropped the *birth\_year* column (since *age* that has the age range by tens can be used, and such precise ages that we could get from *birth\_year* is not so necessary). The proportion of *age* missing was so small that I decided to replace the rows that have null data for *age* with the mean of the column.

The proportion of *sex* missing was so small that I decided to replace the rows that have null data for *sex* with the mode of the column.

The features I ultimately worked with are *sex*, *age*, *province* (location), and *confirmed\_date*. (gives insight on the time around which the patient contracted the virus, and therefore, how long it has been since). I picked these features to use to predict a patient's *state*, which could help someone determine how severe their condition is. Released might indicate that the patient is healthy or had mild symptoms and can go home. Isolated might indicate that the patient's condition is more severe and needs to be isolated and treated in the hospital. Deceased might indicate that the patient's condition was extremely severe.

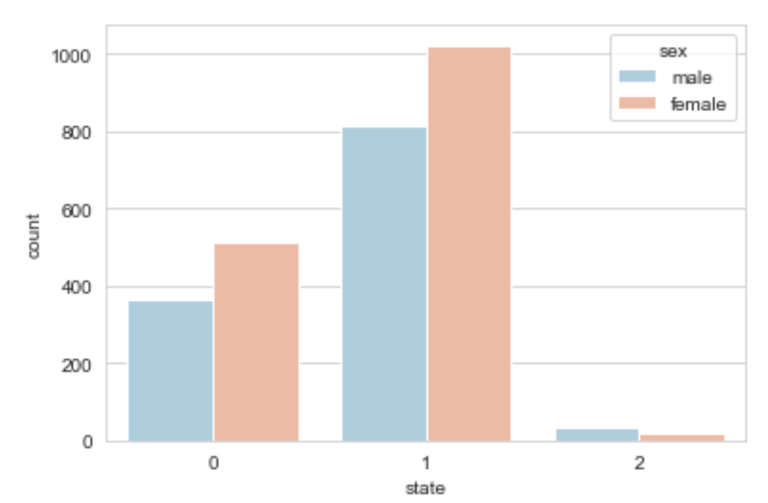


Fig. 2. A graph showing how many patients are in each state, separated by sex. 0 = released, 1 = isolated, 2 = deceased.

As evident from Fig. 2, there are more female patients than male. This could indicate that females are more susceptible to contracting COVID-19. Being female may not definitively mean a higher chance of contraction, but rather, it is a trend. Consider that the estimated male to female ratio in South Korea is currently 0.995, meaning that there are more females than males in the population.

Fig. 2 also shows that more patients are isolated than released or deceased. This could indicate that the virus is not extremely deadly, but it is also not something so easily treated at home.

A picture containing screenshot

Description automatically generated

Fig. 3. A boxplot showing the age distributions, separated by gender.

Fig. 3 shows that the female patients tend to be slightly older than the male patients. This could be an indication of a patient’s state of health prior to contracting COVID-19.

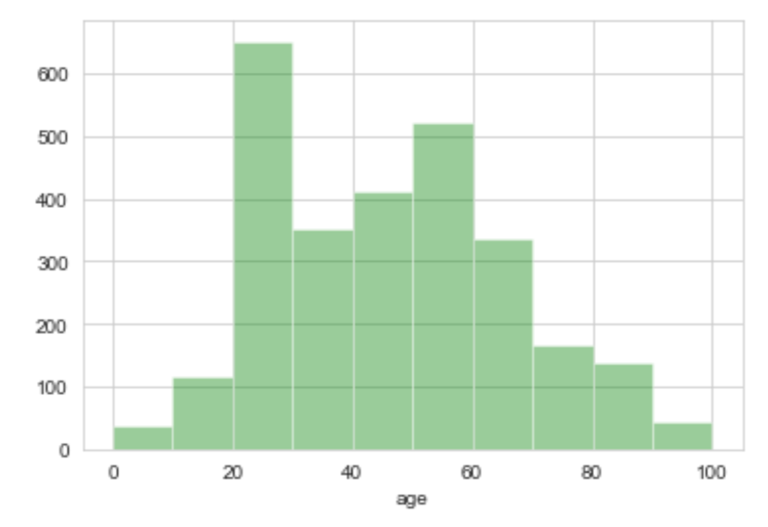


Fig. 4. A distribution plot showing the distribution of ages of patients.

Fig. 4 shows that patients’ ages are mostly in the 20s. Aside from that, the distribution is relatively normally distributed. Consider that age generally affects the actions and health of a person.

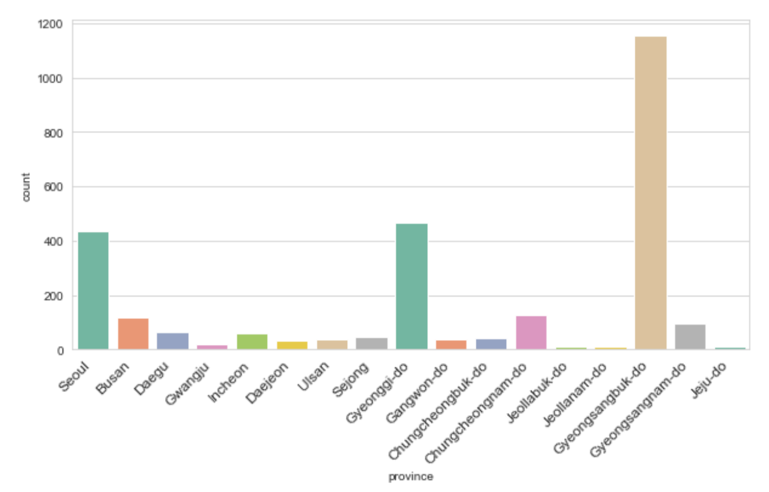


Fig. 5. A plot showing the counts of patients in provinces of South Korea. Some major cities are represented as well.

Fig. 5 shows that a large number of COVID-19 patients are located in Gyeongsangbuk-do. We can deduce that this might be the region where the super-spreader was located. In provinces with more cases, the rate of exposure to the virus is likely higher. An influx of patients might affect an individual’s accessibility to hospital resources, healthcare, or treatment, in turn, affecting an individual’s state.

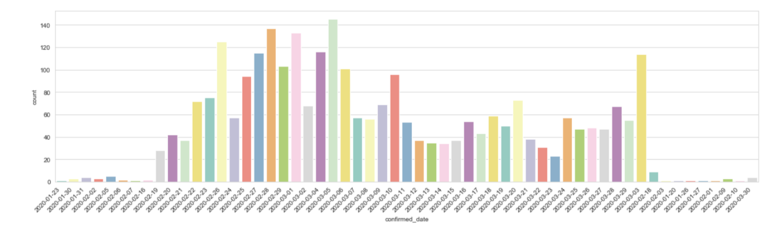


Fig. 6. A plot showing the counts of patients confirmed on a certain date.

Fig. 6 gives an indication of how long it has been since a patient was confirmed to have COVID-19. Considering that many of the cases were confirmed in the first half of the represented dates and that the majority of the patients are being isolated in the hospital, we can hypothesize the severity of the virus and its impact on patients.

## Data Analysis

I chose to use the Naive Bayes algorithm to train and test the data due to it being able to perform well in multiclass class prediction and with categorical input variables.

I also chose to use the k-nearest neighbors (KNN) algorithm, as it naturally handles multiclass cases and is flexible to feature choices.

# Results

The Naive Bayes algorithm resulted in an accuracy score of approximately 0.717 and a k-fold cross-validation score of approximately 0.643.

The KNN algorithm resulted in an accuracy score of 0.778 and a k-fold cross-validation score of 0.654.

# Discussion

Both of the algorithms resulted in cross-validation scores that are lower than the original accuracy scores, indicating that the models may not be the best at determining a patient’s state. However, the models would still be able to predict a patient’s state to some degree of accuracy. One contributing factor to that might be the distribution of *state* in the original dataset. The vast majority of the patients have a state of either “isolated” or “released,” and very few have a state of “deceased,” likely making it difficult to accurately predict a “deceased” state based on the features.

Whether or not a patient had any previous underlying diseases would likely be a large deciding factor in a patient’s *state*. Had there not been so much unknown data for this factor, the models would have probably been more accurate, given that every patient in the dataset that was indicated to have had an underlying disease had a state of “deceased.”

# References

“Coronavirus (COVID-19).” *Centers for Disease Control and Prevention*, U.S. Department of Health & Human Services, www.cdc.gov/.

Fleming, Sean. “South Korea’s Foreign Minister Explains How the Country Contained COVID-19.” *World Economic Forum*, 31 Mar. 2020, www.weforum.org/.