COVID-19 in South Korea:
Data Visualization and Analysis

Kathy JaYoung Byun
Rhea Li
Jody Zhu

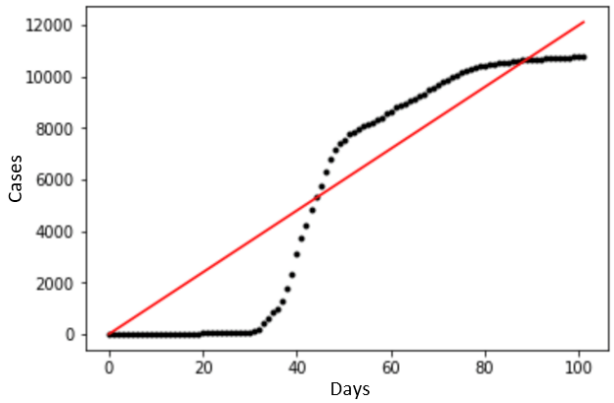
The current pandemic situation is full of unknowns. Some countries including South Korea are keeping detailed records of those affected, and with the aid of data visualizations and analysis tools, many observations can be made about those large datasets. Our work is based on the DS4C dataset retrieved from Kaggle and created using information provided by the Korea Centers for Disease Control & Prevention until the end of April. In this project, we attempt to answer the following three questions:

1. What will be the number of cases in 1 week or in 2 weeks?
2. Which provinces will see the fastest increase or decrease?
3. Which demographics of Koreans are most likely to be confirmed or deceased?

Number of Cases

A huge problem COVID-19 can create is an overwhelming number of patients needing medical care but not enough hospitals and equipment to service them. Having even a rough estimate can improve physical and mental preparations in the coming weeks or months.

We will arrive at a prediction for the next 2 weeks from the total number of confirmed cases up to April 30th. Linear regression produces a preliminary time series model. For simplicity, we use the number of days since the start of the COVID-19 outbreak on January 21, 2020 ("td" in Figure 1b) instead of a year-month-day format for time.



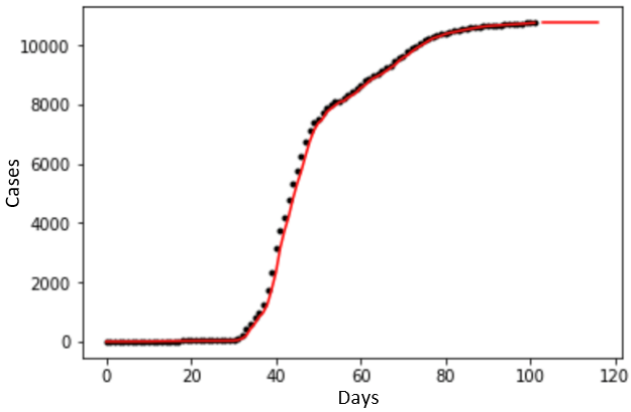
OLS Regression Results

Dep. Variable:	confirmed	**R-squared:**	0.942
Model:	OLS	**Adj. R-squared:**	0.942
Method:	Least Squares	**F-statistic:**	1644.
Date:	Tue, 12 May 2020	**Prob (F-statistic):**	2.60e-64
Time:	13:16:00	**Log-Likelihood:**	-905.67
No. Observations:	102	**AIC:**	1813.
Df Residuals:	101	**BIC:**	1816.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
td	119.9302	2.958	40.550	0.000	114.063	125.797

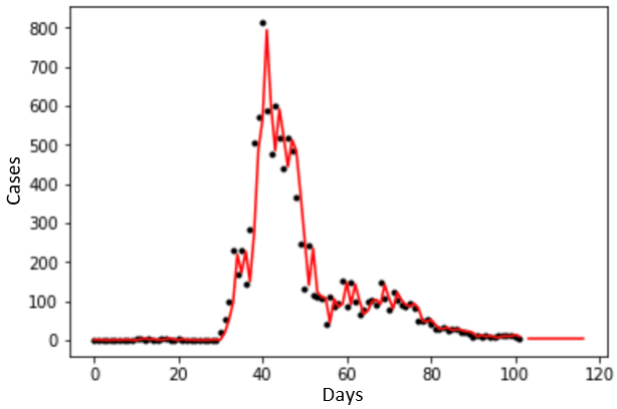
Figure 1a: (left) Cumulative confirmed cases with red linear fit Figure 1b: (right) Model summary

Although the adjusted R-squared is 0.942, which suggests that this model fits the data well, the shape of the graph in Figure 1a clearly indicates that the correlation between the total number of confirmed cases and days passed is not linear. Forecasting is typically better suited for time series and extrapolation into the future; we will further examine simple exponential smoothing, exponential smoothing (with trend), and Holt's method.



Simple Exponential Smoothing

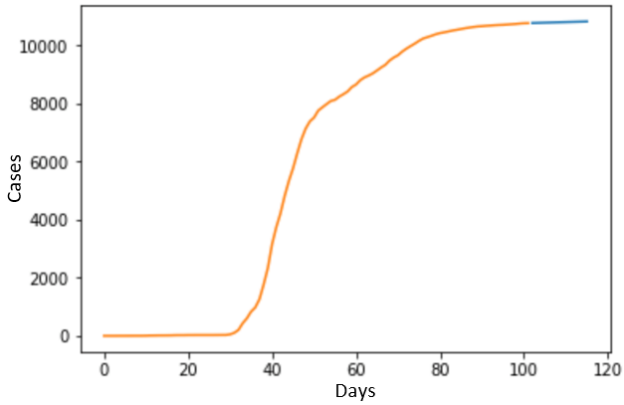
{'smoothing_level': 1.0,
'smoothing_slope': nan,
'smoothing_seasonal': nan,
'damping_slope': nan,
'initial_level': 1.0,
'initial_slope': nan,
'initial_seasons': array([], dtype=float64),
'use_boxcox': False,
'lamda': None,
'remove_bias': False}



{'smoothing_level': 0.9246411035791019,
'smoothing_slope': nan,
'smoothing_seasonal': nan,
'damping_slope': nan,
'initial_level': 0.997486114726252,
'initial_slope': nan,
'initial_seasons': array([], dtype=float64),
'use_boxcox': False,
'lamda': None,
'remove_bias': False}

Figure 2a: (top) Cumulative confirmed cases with red fit Figure 2b: (bottom) New cases each day with red fit

Simple smoothing allows us to consider the data without adding a trend. Using simple exponential smoothing, we get that there will be no more growth in the total number of confirmed cases, leaving us with the prediction that there will be a total of 10765 cases by the end of both week 1 and week 2. Contradictorily, when we forecast the new confirmed cases each day, we get that the number of new confirmed cases will be a constant 4.40 each day for the following 2 weeks. The smoothing level represents the weight of previous data points: a larger value closer to 1 as seen in all our models means predictions are made heavily based on the most recent data.



Exponential Smoothing (with trend)

{'smoothing_level': 0.9694815917289168,
'smoothing_slope': 0.9694678198214725,
'smoothing_seasonal': nan,
'damping_slope': nan,
'initial_level': 1.0000000287212958,
'initial_slope': 0.0,
'initial_seasons': array([], dtype=float64),
'use_boxcox': False,
'lamda': None,
'remove_bias': False}

Figure 3: Cumulative confirmed cases with blue forecast

Using exponential smoothing (with trend), there will be a total of 10794.51 cases by the end of week 1 and 10823.85 cases by the end of week 2. Number of new cases per day are below:

Week 1	4.44	4.47	4.51	4.54	4.57	4.61	4.64
Week 2	4.67	4.71	4.74	4.77	4.81	4.84	4.87

There is approximately a 0.03 increase from day to day, which is reasonable since the increase might be natural from the sheer amount of people now with the virus who can infect others. Although weak plateauing in the short run is not a big issue, we do expect to see a leveling-out based generally on the visualization as well as the pandemic eventually ending. Predicting further out like 1 or 2 months with the exponential smoothing model might be less reliable.

The last model, Holt's method, returns the same result as exponential smoothing (with trend). Rather than repeating the same information, we decided to explore possible impacts on the number of confirmed cases if everyone in South Korea were to get tested.

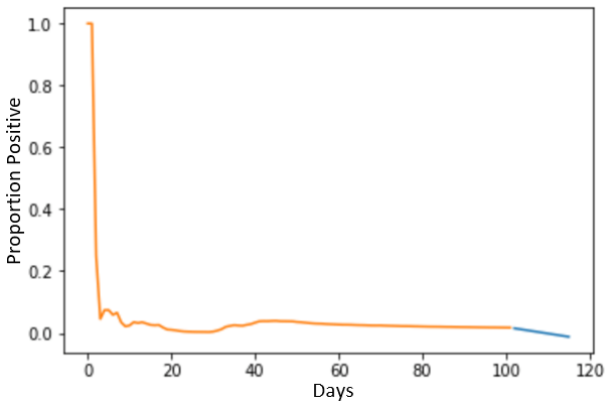


Figure 4: Proportion of tests that are positive over time

This is a graph of the proportion of all tests conducted that turned out to be positive cases. We see that in the very beginning, almost everyone who got tested received a positive result. However, this ratio decreases to a lower, steady value as testing capacity increases. The blue forecast line predicts that this ratio will keep declining. With South Korea tending towards zero new confirmed cases, it is almost certain that the ratio will drop to hover around zero.

Number of Cases by Province

The nationwide numbers above are composed of those of 9 provinces and 8 special/metropolitan cities, which we will group and loosely call provinces. By breaking down into regions, it will be more informational for individuals in understanding the situation closer to home.

Figure 5 (below) depicts the number of weekly confirmed cases for each province starting where there is an overall fall in cases. Most provinces see a decrease. A steeper negative slope would indicate a faster decrease rate. Daegu has the fastest decrease while Gyeonggi-do and Seoul are more or less tied in second. Based on the trend, Daegu will have the fastest decrease. Because a decrease cannot occur past 0 cases, the long term slopes might not be good predictors, so we resort to forecasting.

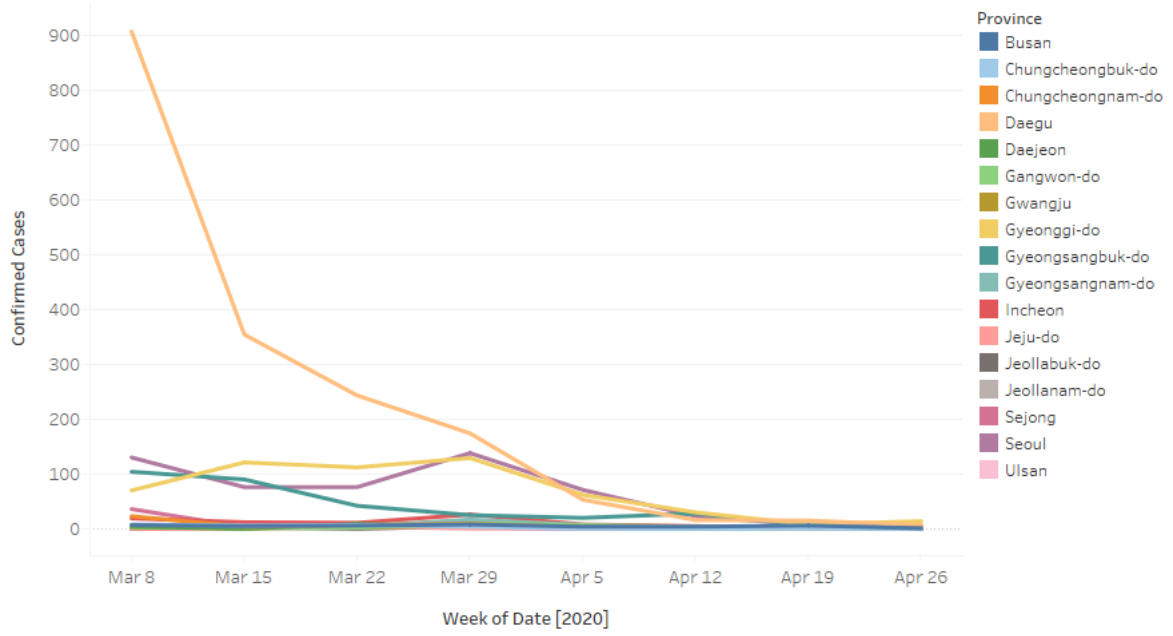


Figure 5: Line graph: Number of new confirmed cases each week by province after March 8

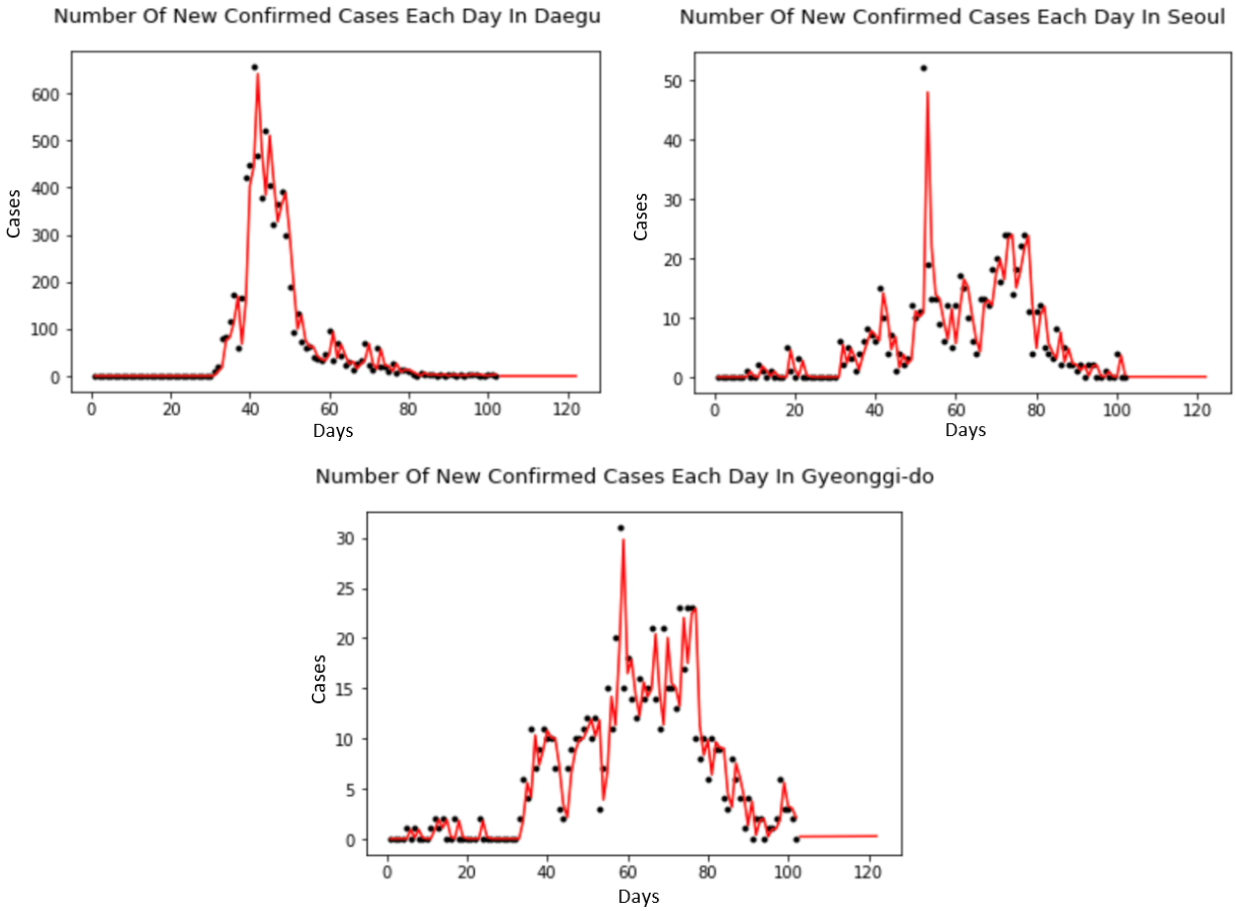


Figure 6a: (top left) Daily new confirmed cases in Daegu Figure 6b: (top right) Seoul Figure 6c: (bottom) Gyeonggi-do

We analyze the number of new confirmed cases each day for different regions to predict which region will see the fastest decrease. We once again use linear regression, simple exponential smoothing with and without trend, and Holt's method to predict the number of confirmed cases. Holt's method model performed best in forecasting based on the sum of squared error (SSE) value and autocorrelation plot. Figure 6a is Daegu, Figure 6b is Seoul, and Figure 6c is Gyeonggi-do. Each graph illustrates the number of new confirmed cases each day from January 21 to April 30.

Using Holt's model to predict, we found that Gyeonggi-do actually had the greatest *increase* with 1.27 new cases per day. Seoul is second with 0.65 new cases per day. Daegu is third with 0.23 new cases per day. The numbers of new cases are close to 0, implying that there should not be huge changes in total confirmed case numbers.

Previous predictions based on visualization are decreasing, yet predictions based on forecasting are increasing. Big picture-wise, we observed downward trends. When we zoom in on the most recent segment of the time series graph, there does appear to be minor fluctuations or noise around 0. The forecast model focuses more on recent days, so the positive values it outputted are sensible. Both visualization and forecasting suggest that the number of confirmed cases will eventually plateau for all regions.

A major factor determining the number of cases for each region is the location of major outbreak events rather than population or other demographics, which will be discussed in the next section. There were two major events that left huge aftermaths. In February, patient 31 participated in a religious ceremony in Shincheonji Church in Daegu with thousands of people. As of May 4th, 5212 confirmed cases are related to Shincheonji Church, which is about 48.3 percent of all confirmed cases. Just this week on May 9th, a man living in Yongin went to clubs in Itaewon (Seoul). He exposed the virus to approximately 1500 people; 86 people tested positive.

Vulnerable Demographics

Knowing COVID-19's vulnerable populations can lead to further precautions for them. Location, age, and sex are three demographics available for a sample of patients in the dataset.

Continuing the discussion of provinces, we examine a possible correlation with population density. The darker red and oranges signify higher population density in the left map of Figure 7 (below), and they seem to match the darker purple patient hotspots in the right map. This means more confirmed cases might occur where more people reside. Due to the lack of data on individuals without the virus, a model cannot be generated. We can predict a poor fit from the following contradicting observation: the percentage of residents who are confirmed cases in Daegu (pop. density 2818/km^2) is 0.278% while in Seoul (pop. density 17000/km^2) it is 0.007%. This leads back to the major factors mentioned previously, especially "super-spreader" patient 31 in Daegu.

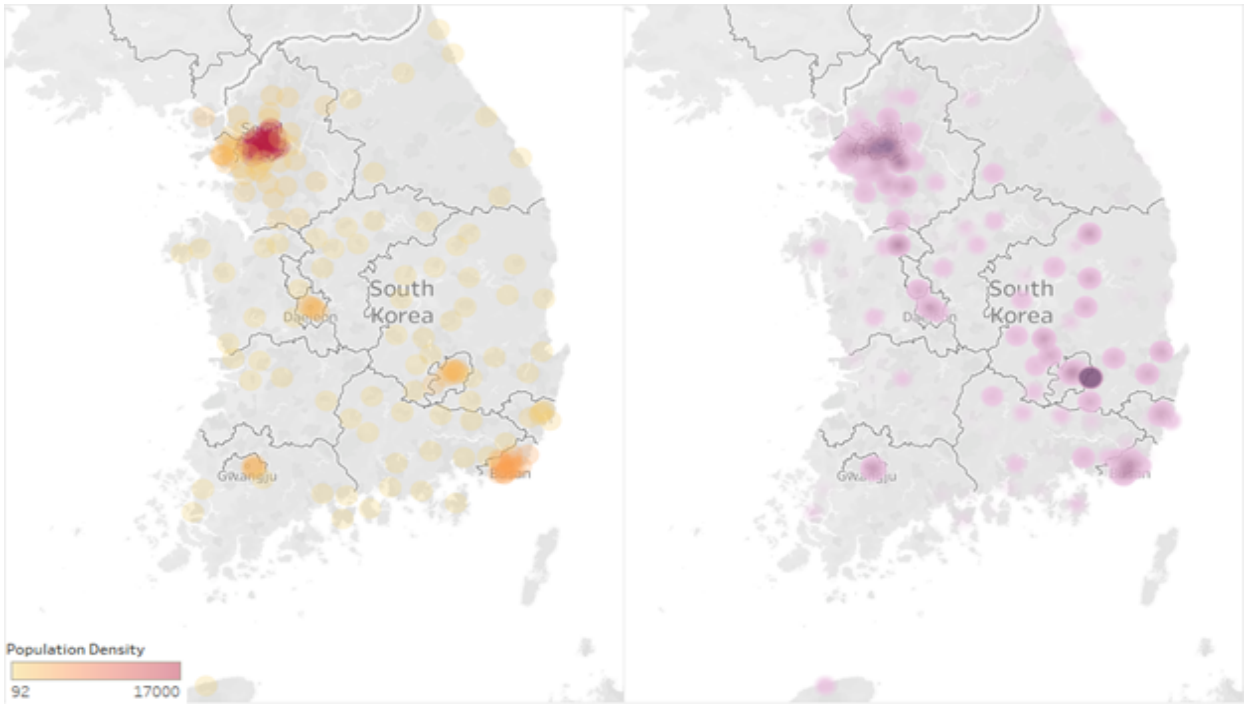


Figure 7: Density maps: Left is population density and right is number of patients

Transitioning to looking at the relationship between deceased and population density, logistic regression can be performed. Logistic regression is a statistical method that models probabilities for binary dependent variables, in this case, released or deceased. The outputted p-value of 0.078 is greater than the 0.05 significance level, meaning the predictor of population density does not alter the chance of dying from the virus.

The insignificance of population density is also apparent during model selection. The process first randomly splits the entire dataset into a training set and a test set and then continuously runs logit command, which is related to logistic regression, until all independent variables with large p-values are eliminated. The new subset of variables are used for the test set. Starting with the three demographics of population density, age, and sex, only age and sex remained in what was thought of as the best model.

Age is the next predictor to analyze. In Figure 8 (below), it is depicted by horizontal bars. The number of confirmed cases, which is the summation of cases from all the states, is the total length of the horizontal bars. Currently, the 20s are in the lead followed by the 50s. In general, those who frequently commute to work using public transportation (20s-50s) have greater numbers of confirmed cases than those who are too young or retired. This supports the stay at home campaigns to stop the spreading. If more data was available to us, then we can model to see whether the observations are statistically supported.

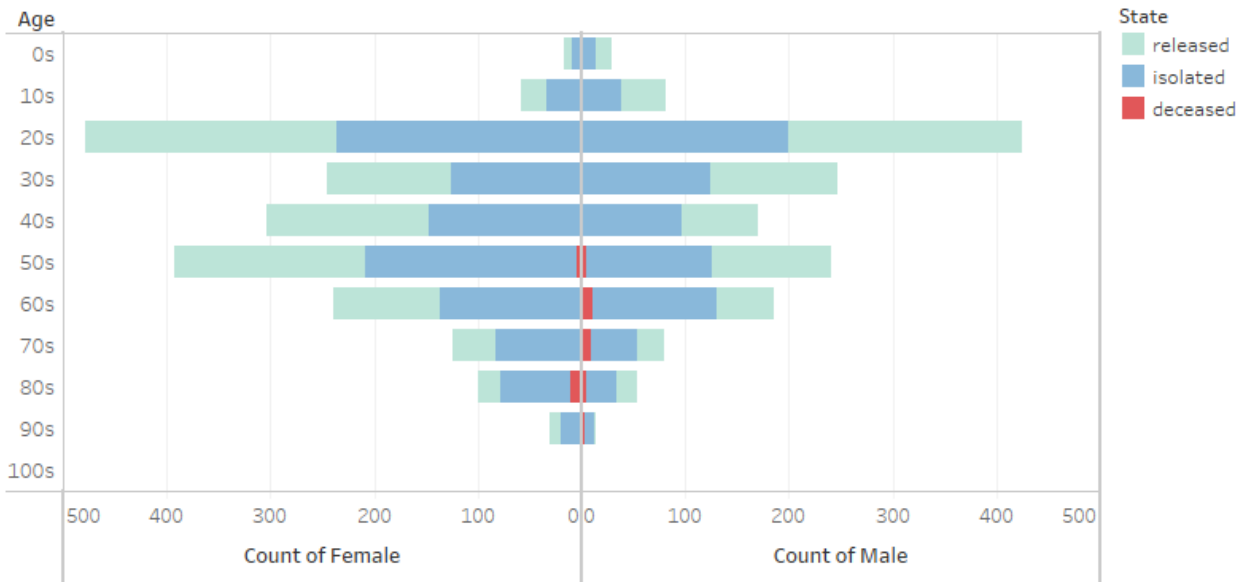
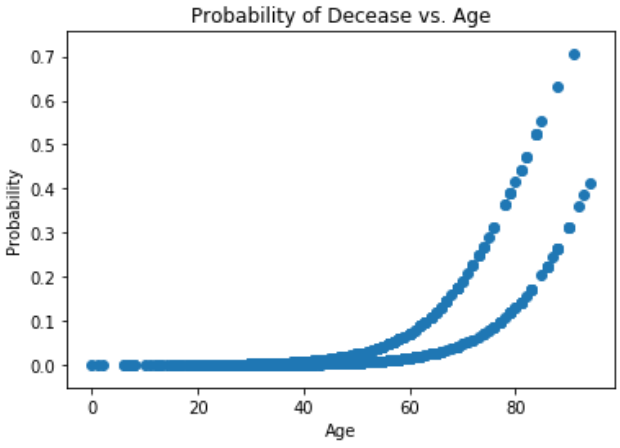


Figure 8: Butterfly stacked bar chart: Number of patients in each state categorized by age and split by sex

Meanwhile, deceased cases are marked in red and only appear for 50s and older; the numbers appear to be growing until it reaches the max in the 80s. This distribution suggests that the elderly face a higher case fatality rate even though fewer of them are infected. According to logistic regression, there indeed is a significant positive correlation with p-value of 0. The coefficient is relatively small, meaning a drastic influence by age on probability of death might only occur at really "old" ages: in Figure 9 where age is plotted against predicted probability of death, the few points that surpass 50% fatality likelihood rate are past 80. Nevertheless, there is still an increase in age. An interesting pattern in the graph is the presence of two upward trends that is actually because of differences in sex.



	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.5145	1.005	-8.468	0.000	-10.485	-6.544
age	0.0983	0.014	6.839	0.000	0.070	0.126
sex	-1.1273	0.419	-2.687	0.007	-1.950	-0.305

Figure 9: Predicted probability of becoming deceased plotted against age

Sex is shown side by side in Figure 8 (previous page). For the number of confirmed cases, the bars on the right of 0 roughly mirror the shape of the bars on the left. The discrepancies between female and male are small enough where the observations made about age above still hold for both. It is worthwhile to note that there have been more female patients overall and in 8 of the 10 age groups; however, this does not translate to more deceased female patients. In fact, there are much fewer compared to males that there exists a significant difference between the sexes with females having lower probability of dying. Revisiting the butterfly chart, it shows more red on the male side, and combined with a smaller total number of confirmed cases in the denominator, the last statement does not seem far-fetched.

We perform a two proportion z-test to decide whether to reject the null hypothesis of equal population proportions. The z score can be calculated where p is proportion and n is sample size:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

The corresponding p-value is 0.000989 which is less than 0.05, deeming it significant. From the three demographic predictors analyzed, working-aged women are more susceptible to getting the virus out of any group, but elderly men with the virus are in the most danger.

While data visualizations and computing tools have allowed us to complete a more in-depth analysis of the COVID-19 situation in South Korea, forecasts should be taken with a grain of salt. Some models only perform well in the short run, and excessive extrapolation can be inaccurate. Another key fact of life to remember is that unexpected events happen; models that generalize trends over long periods of data could not have predicted the shock in the number of cases that occurred in Itaewon this past week. Many people also speculate a second wave of the virus, so all that being said, much more can still be done in analyzing COVID-19. Stay safe everyone.

Bibliography

"Comparing Two Proportions." *Comparing Two Proportions*, online.stat.psu.edu/stat414/node/268/.

Fuqua School of Business. *Moving Average and Exponential Smoothing Models*. people.duke.edu/~rnau/411avg.htm.

Kim, Jihoo. "Data Science for COVID-19 (DS4C)." *Kaggle*, 14 Apr. 2020, www.kaggle.com/kimjihoo/coronavirusdataset.

"Provinces - Republic of Korea." *Knoema*, knoema.com/atlas/Republic-of-Korea/Provinces-profiles.

Rashid, Raphael. "Being Called a Cult Is One Thing, Being Blamed for an Epidemic Is Quite Another." *The New York Times*, The New York Times, 9 Mar. 2020, www.nytimes.com/2020/03/09/opinion/coronavirus-south-korea-church.html.

"이태원 클럽발 코로나 확진자 86명...'13일까지 확진 많을듯'" *조선일보*, 11 May 2020, news.chosun.com/site/data/html_dir/2020/05/11/2020051102157.html?utm_source=naver&utm_medium=original&utm_campaign=news.