COVID-19 Real-Time Tracker and Analytical Report

Jiawei Long

Department of Biostatistics
UCLA Fielding School of Public Health
University of California, Los Angeles

Email: peterljw@g.ucla.edu

***Abstract* -** While the COVID-19 outbreak was reported to first originate from Wuhan, China, it has been declared as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 by WHO, and it has spread to over 180 countries by the time of this paper was being composed. As the disease spreads around the globe, it has evolved into a world-wide pandemic, endangering the state of global public health and becoming a serious threat to the global community. To combat and prevent the spread of the disease, all individuals should be well-informed of the rapidly changing state of COVID-19. In the endeavor of accomplishing this objective, a COVID-19 real-time analytical tracker has been built to provide the latest status of the disease and relevant analytical insights. The real-time tracker is designed to cater to the general audience without advanced statistical aptitude. It aims to communicate insights through various straightforward and concise data visualizations that are supported by sound statistical foundations and reliable data sources. This paper aims to discuss the major methodologies which are utilized to generate the insights displayed on the real-time tracker, which include real-time data retrieval, normalization techniques, ARIMA time-series forecasting, and logistic regression models. In addition to introducing the details and motivations of the utilized methodologies, the paper additionally features some key discoveries that have been derived in regard to COVID-19 using the methodologies.

Index Terms - COVID-19, Real-Time Tracker, Common Symptoms, Data Visualization, Hypothesis Testing, ARIMA Time-Series Forecast, Penalized Logistic Regression

1. Introduction

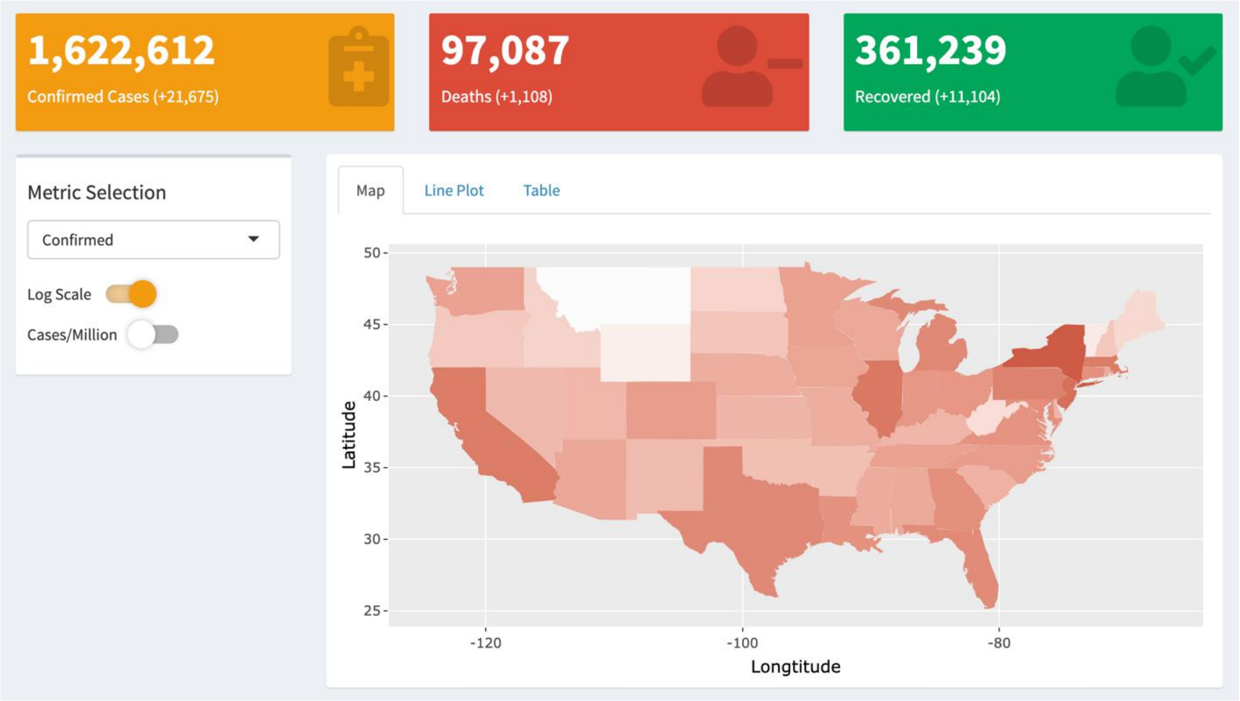
The COVID-19 real-time tracker primary includes features such as odometers of the latest status of COVID-19 cases, trend analysis, and prediction of COVID-19 cases in 185 different countries, informative visualizations of the most common symptoms and risk factors, as well as patient demographic distributions. Subsequent sections will be providing a brief description of every major feature, discussing relevant methodologies behind the feature, as well as highlighting selective findings from the feature.

Link to the COVID-19 real-time tracker: https://peterljw.shinyapps.io/covid_dashboard/

2. Feature: Overview

2.1 Overview

The section of the COVID-19 real-time tracker contains two different pages to separately highlight the most current states of COVID-19 in the states within the U.S. and countries around the globe (See figure 1 and figure 2). The two pages share the same features and elements. The top of the page has three odometer boxes to display the total confirmed cases, total deaths, and total recovered cases along with their respective daily new counts. The bottom half of the page contains a user-interactive control panel and a display window. The users are able to apply population normalization or log transformation to the visualizations in the display window through the widgets in the control panel.



1,622,612
Confirmed Cases (+21,675)

97,087
Deaths (+1,108)

361,239
Recovered (+11,104)

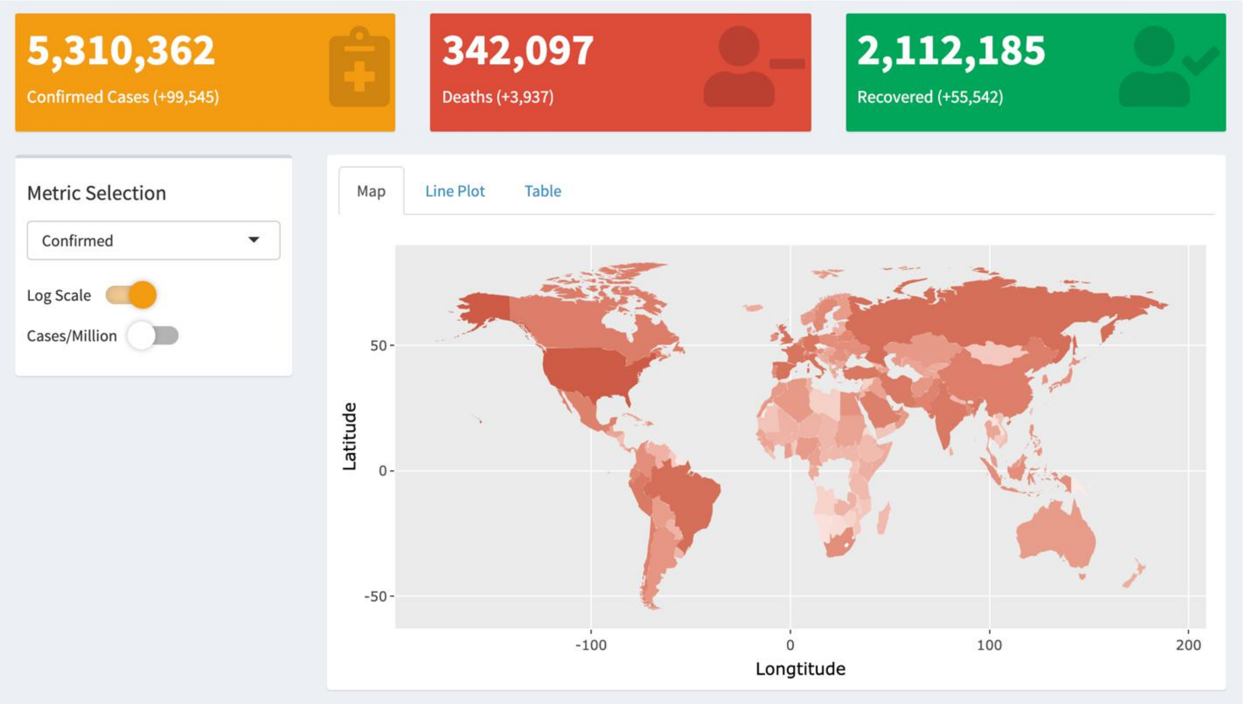
Metric Selection

Confirmed

Log Scale ●

Cases/Million ○

Figure 1. U.S. Overview



5,310,362
Confirmed Cases (+99,545)

342,097
Deaths (+3,937)

2,112,185
Recovered (+55,542)

Metric Selection

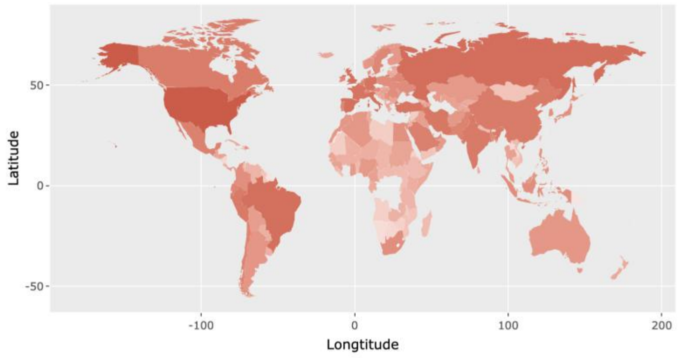
Confirmed

Log Scale ●

Cases/Million ○

Figure 2. World Overview

The display window has the viewing options of heat map visualizations (Figure 3), time-series line plots (Figure 4), or data tables (Figure 5). All visualizations have interactive features such as tooltip and zooming, and all tables can be interactively sorted by clicking on column names. The heat map visualization allows the users to quickly assess the severity of COVID-19 in different geographical locations while the time-series line plot shows a comparison of the most affected regions on a standardized time scale. The data table provides the users with the flexibility to explore and search for data of their interest.



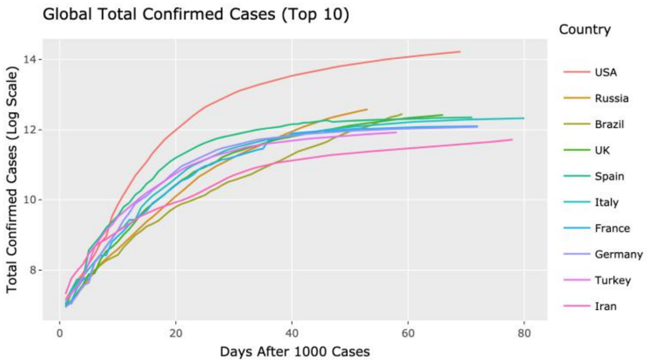
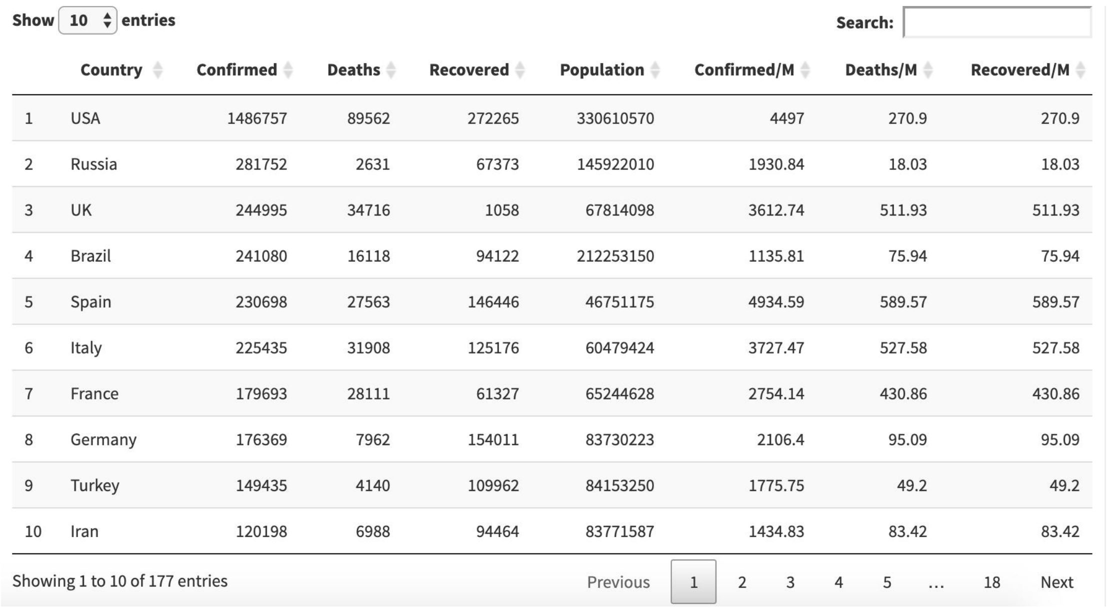


Figure 3. World Confirmed Cases Heat Map

Figure 4. World Confirmed Cases Time-Series Line Plot



Show 10 entries

Search:

	Country	Confirmed	Deaths	Recovered	Population	Confirmed/M	Deaths/M	Recovered/M
1	USA	1486757	89562	272265	330610570	4497	270.9	270.9
2	Russia	281752	2631	67373	145922010	1930.84	18.03	18.03
3	UK	244995	34716	1058	67814098	3612.74	511.93	511.93
4	Brazil	241080	16118	94122	212253150	1135.81	75.94	75.94
5	Spain	230698	27563	146446	46751175	4934.59	589.57	589.57
6	Italy	225435	31908	125176	60479424	3727.47	527.58	527.58
7	France	179693	28111	61327	65244628	2754.14	430.86	430.86
8	Germany	176369	7962	154011	83730223	2106.4	95.09	95.09
9	Turkey	149435	4140	109962	84153250	1775.75	49.2	49.2
10	Iran	120198	6988	94464	83771587	1434.83	83.42	83.42

Showing 1 to 10 of 177 entries

Figure 5. World Data Table

The purpose of this feature is to provide the audience with an aggregated view of the severity of COVID-19 in different locations and inform the audience of the latest status of the disease at a first glance. The options of applying log transformation and population normalization allow the audience to observe the state of COVID-19 from different perspectives while the interactive table allows the audience to explore specific metrics of their interest.

2.2 Real-Time Data Retrieval

To ensure the accuracy and the reliability of the tracker's content, the website's server retrieves the newest data from the COVID-19 data repository by the Center for Systems Science and Engineering at Johns Hopkins University when any user tries to load the web page. The data repository is regulated by the Johns Hopkins University Center for Systems Science and Engineering and supported by the ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab. According to the documentation of the repository, the data source is being updated in real-time numerous times throughout the day, and the validity of the data is verified by researchers at Johns Hopkins University. The content displayed in the overview feature is therefore derived from a real-time and reliable data source.

To achieve real-time data retrieval, the webserver contains a protocol to download and ingest the data source from the CSSE data repository by Johns Hopkins University when a request is sent to the server when a user tries to access the web page in a browser. When the server receives such requests, it will attempt to download the data by getting the current date and accessing the data source with an updated URL. Once the data file is downloaded successfully, it will be ingested and stored temporarily on the server. Subsequently, the pre-specified R script will read the data and preprocess it into different data frames to support the insights to be derived on the web page. If the download were to be unsuccessful due to unforeseen circumstances, the web server will load up the most recent data file that it has ingested previously to support the content on the web page. The server will also log such errors so that they could be handled to improve the robustness of the tracker. Figure 6 provides a visual summary of the real-time data retrieval process.

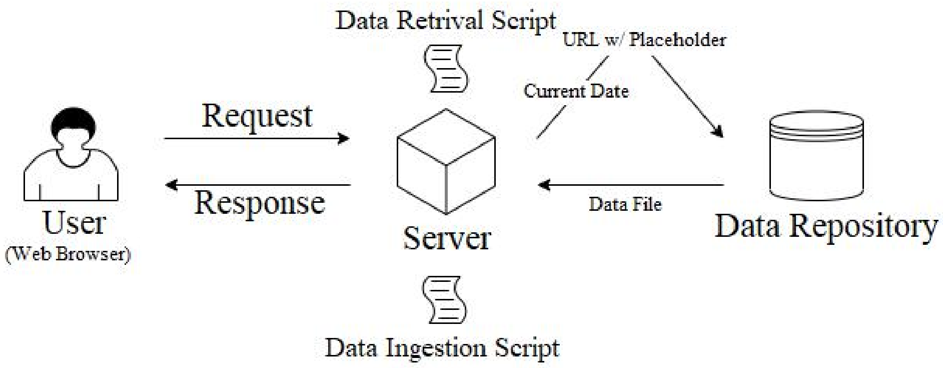
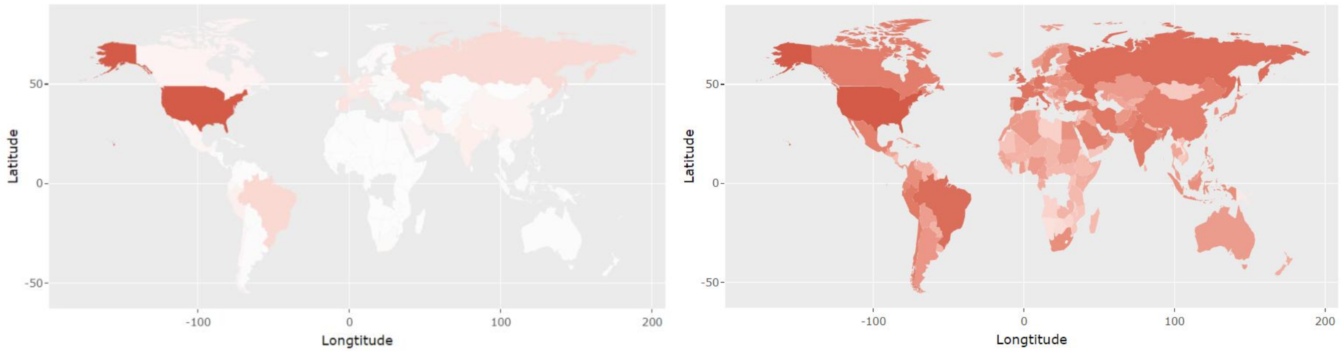


Figure 6. Real-Time Data Retrieval Visualization

2.3 Transformation and Normalization Techniques

The control panel on the page allows the users to apply log transformation and population normalization (i.e. cases per million) to the data, which interacts with the corresponding visualizations of heat map and time-series line plot. When the user turns log scale switch on, the logarithmic function with a base of 10 will be used as a deterministic mathematical function to be applied to each point in a data set. That is, for every data point x_i, its value will be replaced by $y_i = \log_{10}(x_i)$. Such transformation significantly improves the interpretability and the appearance of visualizations. The choice of using the logarithmic function is based on the nature of exponential growth associated with pandemic and the relatively large differences in the raw counts of cases across different locations in the later stages of a pandemic. The effect of log transformation is demonstrated in figure 7.



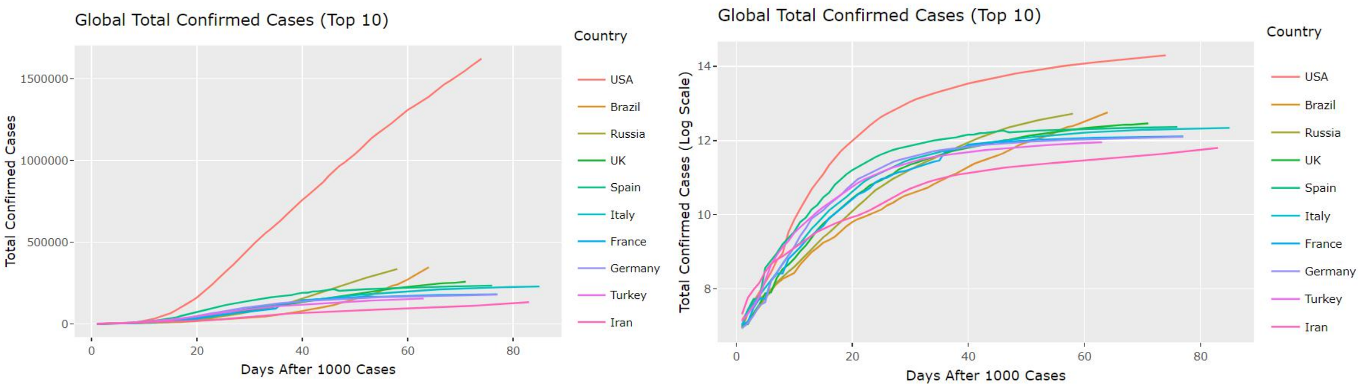


Figure 7. Before (Left) and After (Right) Log Transformation

On the other hand, while population normalization does not necessarily improve the appearance of the visualizations, it alters the interpretation of the visualization by accounting for the population of each region. Such a perspective is beneficial because each country or state could vary significantly in its population. Assessing the number of cases per million provides a more robust estimation of the severity of the COVID-19 in each region rather than solely observing the total counts. To achieve population normalization, global country-level population data and state-level population data of the U.S. are preprocessed and stored on the server, and they will be joined to the retrieved data to produce the corresponding visualizations. Precisely, the normalization is applied in the following manner, for every data point x_i, its value will be replaced by $y_i = (x_i / p_j) \cdot 1{,}000{,}000$, where p_j denotes the population of country$_j$. The effect of log transformation is demonstrated in figure 8.

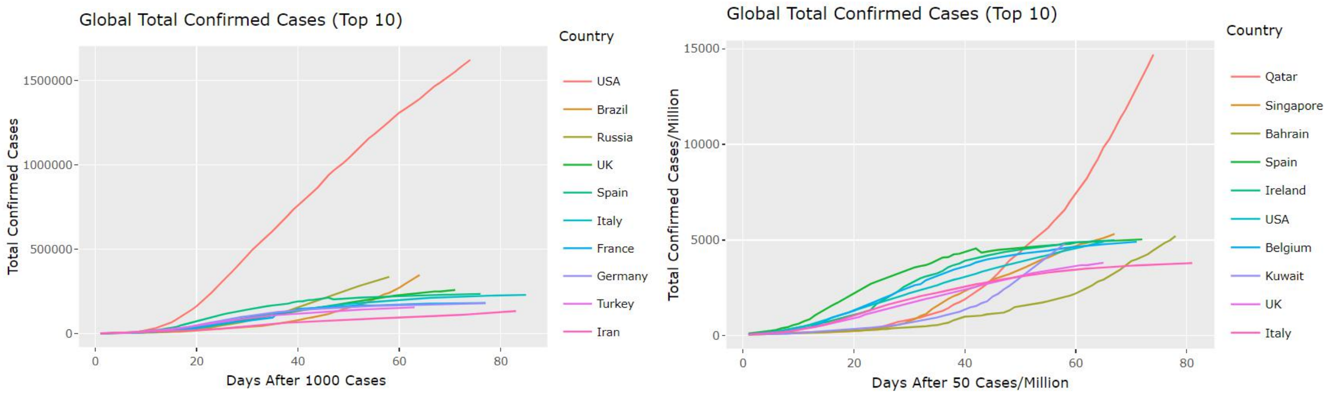


Figure 8. Before (Left) and After (Right) Population Normalization

In addition, the time-series line plots have built-in timescale standardization. Rather than comparing the time-series data with respect to date, the plot compares them with respect to the number of days after the spread of the disease reaches a certain magnitude. Since the time frames of outbreaks are different in every region, it will be hard to compare the severity of the disease in each region in separate time frames. Hence, the application of timescale standardization will help to standardize the time-series data into a universal time scale. In conjunction with population normalization, the audience will be able to compare regions which have the fastest spread of COVID-19.

2.4 Key Findings

There are a number of noticeable findings which surface from the visualizations and data presented from the overview feature. In spite of the fact that the U.S. has the most confirmed cases of COVID-19 around the globe and accounts for approximately 30% of the global total confirmed cases, it no longer tops the chart with the application of population normalization. Among countries which have over a million population, the most severe countries are Qatar, Singapore, Bahrain, Spain, and Ireland. Similarly, if we sort the world data table by confirmed cases per million, we can see the U.S. is not in the top 10 countries. As shown, there are numerous smaller countries that are having tougher struggles with COVID-19 as they have higher confirmed cases per million and tend to have less advanced medical supplies to combat the disease. Such countries gain much less exposure and discussion on mainstream news coverage due to their limited presence in the global economy, but they may be suffering from much greater severity of COVID-19. Similarly, within the United Sates, the severity of COVID-19 has been quietly climbing in some smaller states. For example, Rhode Island, Connecticut, Delaware, Louisiana, and Nebraska are among the top 10 states with the highest confirmed cases per million. While California has the fourth highest confirmed number of cases in the United States and has received relatively high amount of media coverage, it is only standing at the 32th position among all the states while measuring confirmed cases per million. While it is true that regions with higher population is more vulnerable to bigger outbreaks, regions with small populations cannot be overlooked and also deserve some amount of attention.

3. Feature: Trend By Country

3.1 Overview

This section of the COVID-19 real-time tracker contains a user-interactive control and a display window , where the display window shows visualizations of a bar plot of the total cases, a line plot of daily new cases with respect to time along with its moving average, and a 5-days interval estimation of daily new cases (Figure 9). The visualization is presented in a dual-axis plot. The gray line in the back represents the number of new cases while the dotted black line represents the moving average of the number of new cases, and they correspond to the left vertical axis. The orange bar plot represents the accumulated total cases, and it corresponds to the right vertical axis.

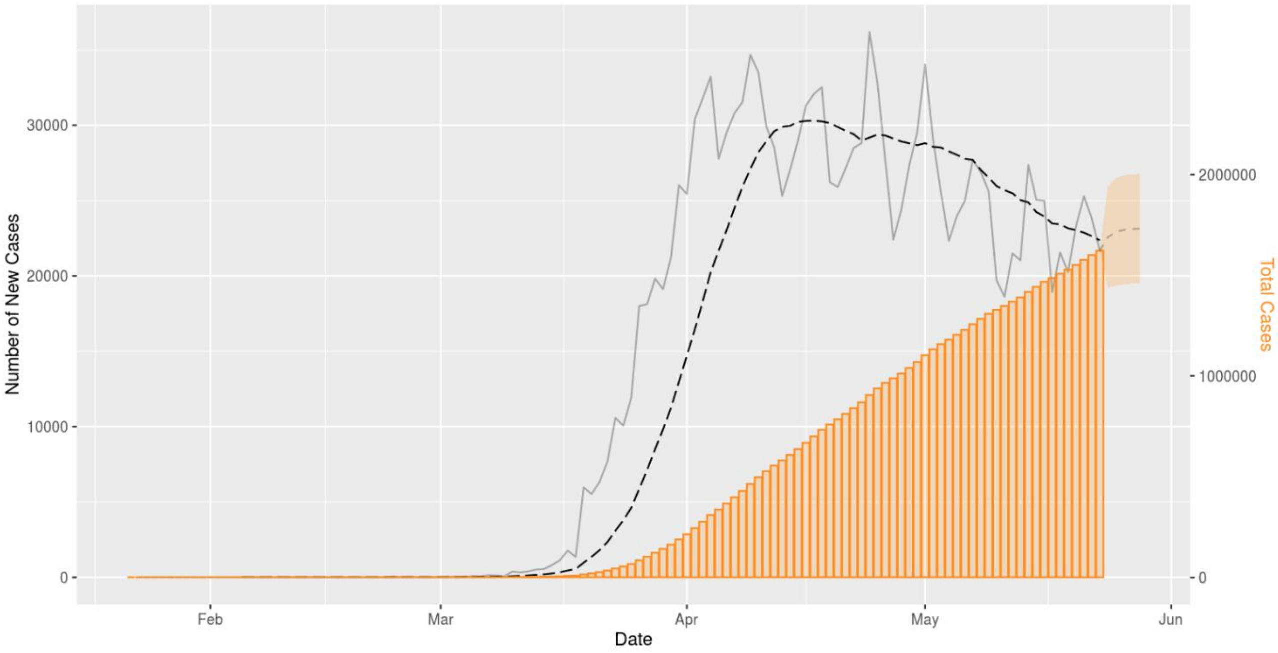


Figure 9. Trend in U.S.

The purpose of this feature is to inform the audience with insights into the trend of the spread of the disease in each individual country. In other words, the plot aims to answer the question of whether the curve has flattened. Since the number of daily new cases have a substantial amount of fluctuations, applying a moving average aggregation will help unrevealing the underlying direction of the curve of the number of daily new cases. In addition, the fluctuations in the number of daily new cases display some degree of short-term patterns that can serve as the basis for time-series forecasting, which will provide the audience of an estimation of the trend in the near future.

3.2 Moving Average

The control panel on the page allows the users to specify the period of days which the moving average aggregation uses to draw the moving average curve. Moving average is an aggregating calculation to analyze data points by creating a series of averages of different subsets of the full data set. In this case, the method of simple moving average is used to compute the values of the moving averages. To calculate a simple moving average, let X_t be the number of new cases at time t, then a simple moving average at $t = m$ is computed as

$$SMA_m = \frac{X_m + X_{m-1} + X_{m-2} + \ldots + X_{m-(n-1)}}{n} = \frac{1}{n} \sum_{i=m-(n-1)}^{m} X_i$$

By computing a series of simple moving averages, we are able to smooth out short-term fluctuations in the number of daily new cases and highlight longer-term trends or cycles. This is especially useful in determining the constantly changing state of the COVID-19 outbreak in a particular region.

3.3 ARIMA Model Forecast

Auto ARIMA model (i.e. auto.arima() function in the forecast R package) is used to implement a 5-days time-series prediction on the number of daily new cases. In summary, for every given time-series, the script will automatically fit the best ARIMA model to the data by using AIC, AICc, or BIC scores as the basis of judgment. The following discussion briefly introduces the notion of the ARIMA model and provides the necessary background information for the subsequent discussion of the prediction mechanism.

Suppose that X_t is a time-series data, where X_t are real numbers and t is an integer, then the $ARIMA\ (p^{!}q)$ model can be written as follows,

$$X_t - \propto_1 X_{t-1} \ldots \propto_{p^{!}} X_{t-p^{!}}$$

$$= \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

It can also be written as

$$\left(1 - \sum_{i=1}^{p^{!}} \propto_i L^i\right) X_t$$

$$= \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

, where L represents the lag operator, ε_t represents the error terms, θ_i represents the moving average part parameters, and $\propto_i$ represents the autoregressive part parameters. However, the general assumption about the ε_t error terms is that they are:

i. Sampled from a normal distribution with zero mean;
ii. Identically distributed variables;
iii. Independent variables.

Assuming that $\left(1 - \sum_{i=1}^{p^{!}} \propto_i L^i\right)$, which is a polynomial, has a unit root, i.e. a factor $(1 - L)$ whose multiplicity is d, it can be written as follows:

$$\left(1 - \sum_{i=1}^{p^{!}} \propto_i L^i\right) = \left(1 - \sum_{i=1}^{p^{!}-d} \Phi_i L^i\right) (1 - L)^d$$

The above polynomial factorization property is expressed by an $ARIMA\ (pdq)$ process with $p = p^{!} - d$. It is given by the following,

$$\left(1 - \sum_{i=1}^{p} \Phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

doi:10.20944/preprints202006.0063.v1

Looking at the above output, one can think of it as an $ARMA\ (pdq)$ process case that has the autoregressive polynomial with the unit roots being equivalent to d, and the above is an indication that there does not exist a wide sense stationary for an $ARIMA$ model with $d > 0$.

Generalizing the above, the definition of $ARIMA\ (pdq)$ obtained is:

$$\left(1 - \sum_{i=1}^{p} \Phi_i L^i\right)(1 - L)^d\, X_t = \delta + \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t$$

The drift of the above defined $ARIMA\ (pdq)$ process is:

$$\frac{\delta}{1 - \sum \Phi_i}$$

There are three major parameters (pdq) of the ARIMA model, in which case p represents the lag order or rather the number of lag observations that are included in the $ARIMA\,(pdq)$ model, d represents the degree of differencing or rather the number of times that the raw observations are differenced, and q refers to the order of moving average or rather the size of the moving average window.

In addition, the ARIMA model assumes that the input time-series data is univariate and stationary. Stationarity implies that the time series' properties are independent of the time when they were captured. In other words, the data has a constant mean and variance. If not, the data needs to be transformed before one can use the ARIMA model. The auto.arima() function automatically determines the appropriate order of differencing using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

As mentioned earlier, auto.arima() performs model selection based on the AIC, AICc or BIC scores. AIC is an abbreviation of the Akaike Information Criterion, and AICc refers to the corrected AIC. BIC is known as Schwarz information criterion, and it is the acronym of the Bayesian information criterion. All of them are very useful model evaluation metrics.

AIC provides a means for model selection. This is particularly because it estimates the quality of each model, from a collection of data models, relative to each of the other models. In other words, it serves as an estimator of out-of-sample prediction error, thus estimating the statistical models' relative quality for a given data set. Given a particular statistical data model with k being the model's estimated parameters number and $\hat{L}$ the maximum likelihood function value. Then the model's AIC value is given by

$$AIC = 2k - 2\ln\left(\hat{L}\right)$$

Thus, as assessed by likelihood function, AIC rewards goodness of fit.

On the other hand, BIC is closely related to the AIC model because it is also partly based on the likelihood function. Equally, BIC is also a criterion for model selection, particularly amongst a "finite" set of models. The most preferred model is one with the lowest BIC. The formal definition of BIC is as follows,

$$BIC = k\ln(n) - 2\ln\left(\hat{L}\right)$$

In both settings,
k = the number of parameters that the model has estimated;
n = the sample size, or number of observations, or number of data points in x;
L = the model's likelihood function's maximized value.

Thus, for ARIM models, the evaluation metrics can be computed as follows,

$$AIC =- 2\log L + 2(p + q + k)$$

$$AICc = AIC + \ [2(p + q + k)(p + q + k + 1)]/(T - p - q - k - 1)$$

$$BIC = AIC + \ [\log T - 2(p + q + k)]$$

, where
k = the ARIMA model's intercept
q = the moving average part order
p = the autoregressive part order
L = the model's likelihood function's maximized value.

For every given time-series, auto.arima() chooses the parameters which give the lowest AIC, AICc, or BIC, and forecast the values for the next 5 days. As a part of the output from the auto.arima() function, the 95% intervals are taken to plot the transparent orange ribbon around the mean prediction. The 95% confidence intervals for ARIMA forecasts are computed as

$$\hat{y}_{T+h|T} \pm 1.96\sqrt{v_{T+h|T}}$$

, where $v_{T+h|T}$ refers to the variance of $y_{T+h|y_1, \ \dots, y_T}$.

3.4 Key Findings

From this feature, we can observe various trends and patterns as countries around the globe have been reacting differently to mitigate the spread of COVID-19. For example, European countries such as Spain, Italy, and Germany have implemented a relatively strict country-level lockdown policy, and it is being reflected from their trend plots (Figure 9).

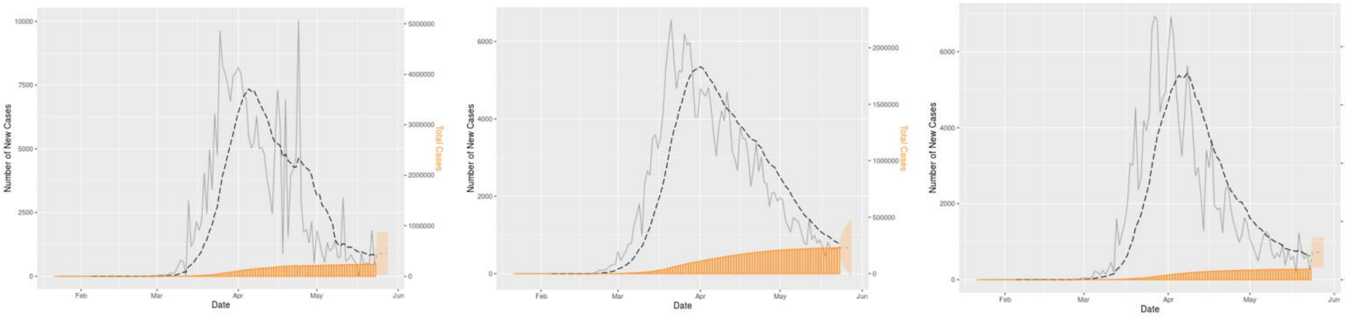


Figure 9. Trends in Spain, Italy, and Germany

Iran also implemented similar lockdown policies, and the country has begun to reopen throughout April after seeing a significant drop in its number of new cases. However, the country has been hit by a new surge of COVID-19 cases in May (Figure 10).

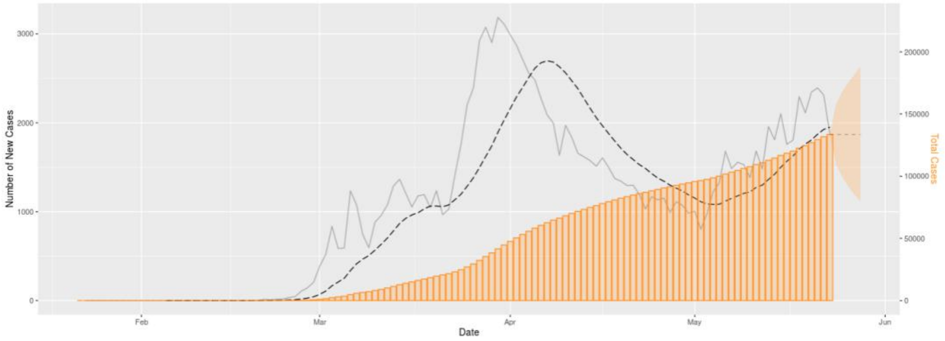


Figure 10. Trend in Iran

In contrast, the U.S.'s reaction at the federal level has been relatively slow-moving and incoherent. Despite that the number of new cases has been gradually decreasing, the rate of decrease is relatively small to countries that implement strict lockdown policies. We may be able to expect a similar new surge of cases in the U.S. if the country were to reopen without cautions in the near future.

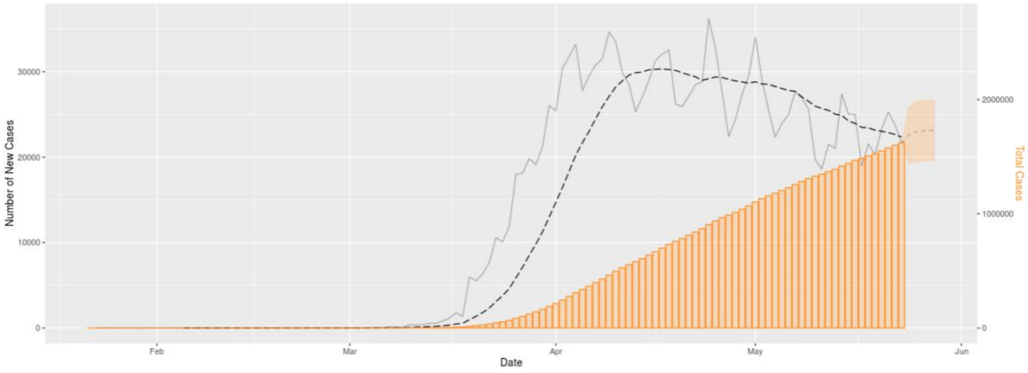


Figure 11. Trend in U.S.

4. Feature: Common Symptoms

4.1 Overview

This section of the COVID-19 real-time tracker contains an interactive visual summary of the most common symptoms associated with the disease (Figure 12). Due to data quality issues and the uncertain nature of the disease, it is difficult to estimate the true prevalence of the symptoms among infected patients. Hence, their prevalence measure is standardized to an 0-to-10 scale, represented by the horizontal axis of the plot.

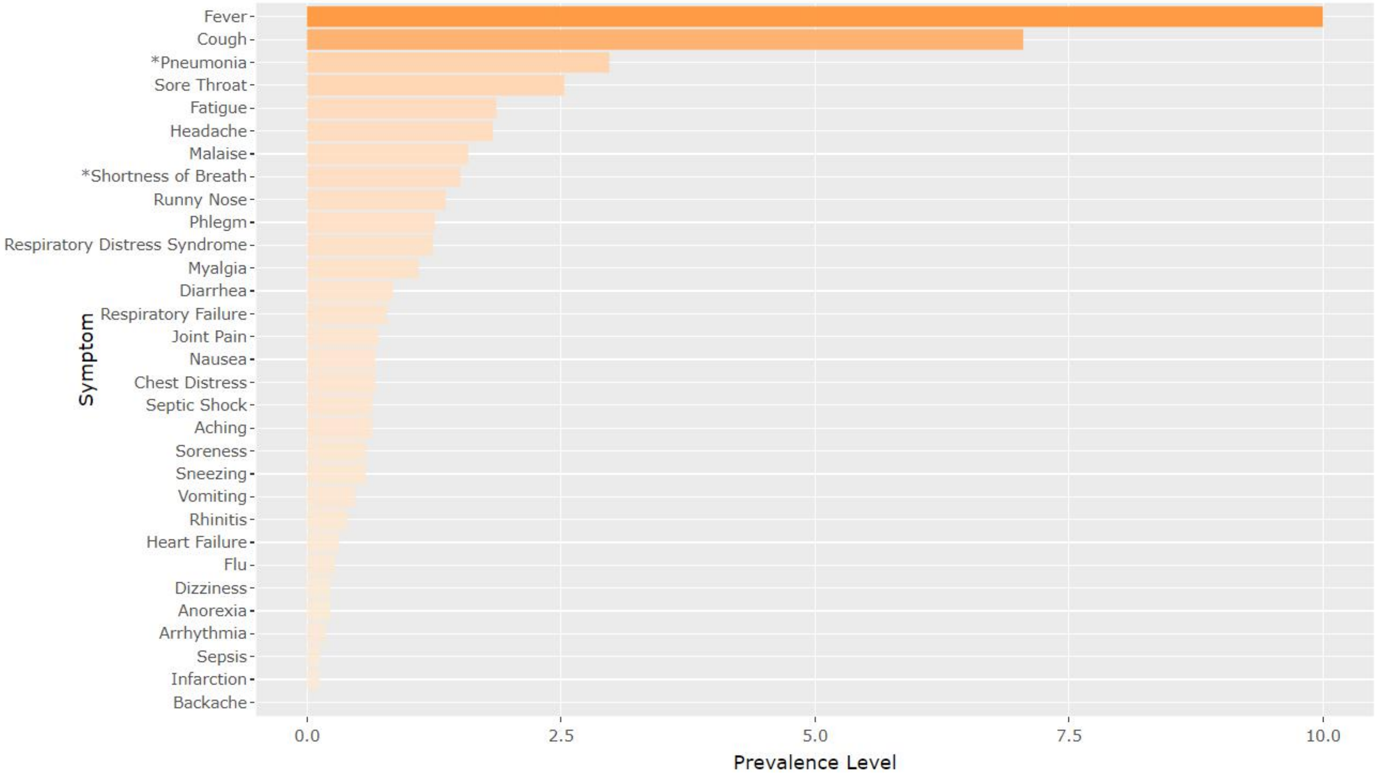
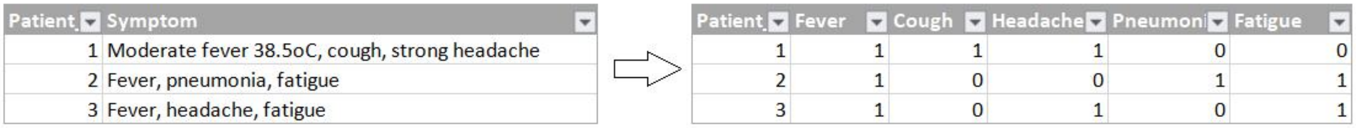


Figure 12. Common Symptoms of COVID-19

4.2 N-Gram Tokenization

Since the symptom variable from the patient-level data contains descriptive sentences of a patient's symptoms (e.g. "Moderate fever 38.5^{o}C, cough, strong headache"), we would have to apply natural language processing techniques such as n-gram tokenization to transform and preprocess the data. The goal is to convert the descriptive sentences into a set of binary indicator variables, as shown by the simple example in figure 13.



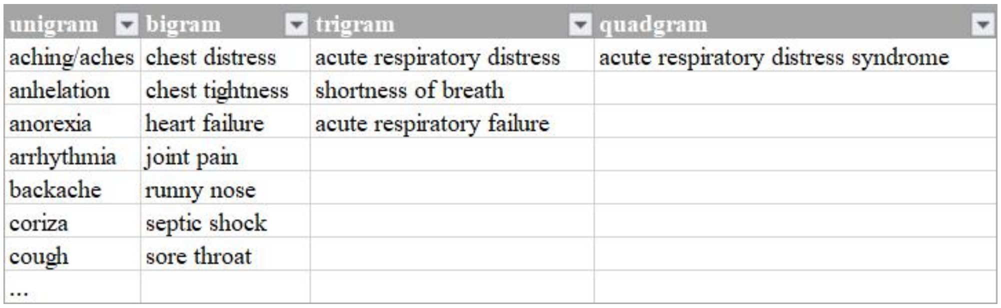
Patient	Symptom
1	Moderate fever 38.5oC, cough, strong headache
2	Fever, pneumonia, fatigue
3	Fever, headache, fatigue

Patient	Fever	Cough	Headache	Pneumoni	Fatigue
1	1	1	1	0	0
2	1	0	0	1	1
3	1	0	1	0	1

Figure 13. Example of Converting Sentences to Binary Indicators

The process of word tokenization refers to splitting a sample of text into words or phrases. In addition, n-gram tokenization refers to tokenization that splits the text into phrases which contain n words. For example, unigram tokenization will turn the sentence *"he has shortness of breath"* into *[he, has, shortness, of, breath]* while trigram tokenization will turn the sentence into *[he has shortness, has shortness of, shortness of breath]*.

As an attempt to collect all of the recorded symptoms in the dataset, we can apply n-gram tokenization to every descriptive sentence and compute the frequency of each token for n = {1, 2, 3, 4}. As anticipated, we can obtain a list of the most common symptoms from the symptom records by looking through the processed output from n-gram tokenization (Figure 14).



unigram	bigram	trigram	quadgram
aching/aches	chest distress	acute respiratory distress	acute respiratory distress syndrome
anhelation	chest tightness	shortness of breath	
anorexia	heart failure	acute respiratory failure	
arrhythmia	joint pain		
backache	runny nose		
coriza	septic shock		
cough	sore throat		
...			

Figure 14. Examples of N-Gram Output

After obtaining a comprehensive list of symptoms, we can then create a dictionary of phrases for each symptom and loop through all descriptive sentences to see if they contain any phrase in any dictionary. For example, the dictionary for cough is *[cough, coughing]*, and any sentence that contains *cough* or *coughing* will take the value of 1 for the cough's binary indicator variable and 0 if otherwise. By the end of the loop, we would have finished converting the descriptive sentences into a set of binary indicator variables in the format that is shown in figure 12.

4.3 Min-Max Normalization

After the application of n-gram tokenization to create all the necessary binary indicator variables, we can then obtain the aggregated count of patients for every symptom by calculating the columnar sums of the binary indicator variables. To better communicate the level of prevalence of each symptom, we can apply min-max normalization to the columnar sums to standardize each data point into a scale of 0 to 10. For any symptom's columnar sum, S_i, its scaled value could be computed as follows,

$$S_{i,scaled} = \frac{S_i - S_{\min}}{S_{\max} - S_{\min}} \cdot 10$$

The scaled value is an abstract representation of the symptom's prevalence relative to other symptoms, and it does not reflect the true prevalence of the symptom among patients who have been infected with COVID-19.

4.4 Logistic Regression

A logistic regression model is built to identify risk factors that could potentially increase a patient's likelihood of dying from COVID-19. Once we have formed all the binary indicator variables for symptoms, we can use them along with other variables as predictors to build a logistic regression model on a patient's outcome, which is either active/recovered or death. The following discussion briefly introduces the notion of the logistic regression model and provides the necessary background information for the subsequent discussion of hypothesis testing of the model's coefficients.

Let y be a binary output variable, taking on values $\in (0,1)$, analogous to a patient's outcome, and we would like to model the output y as a linear function of the input variables, $x = (x_1, ..., x_p)$. As a way to represent $E(y | x)$ so that its value $\in (0,1)$, we can apply the sigmoid function as follows,

$$P(y = 1 | x, \beta) = \frac{e^{\beta^T x}}{1 + e^{-\beta^T x}}$$

$$P(y = 0 | x, \beta) = \frac{1}{1 + e^{\beta^T x}}$$

We can invert the transformation above to obtain the logit function,

$$g(x | \beta) = \log\left(\frac{P(y = 1 | x, \beta)}{1 - P(y = 1 | x, \beta)}\right) = \beta^T x$$

Suppose we are fitting a logistic regression model to a dataset of n observations, $D = \{(x^1, y^1), ..., (x^n, y^n)\}$, we can express the condition likelihood of a single data observation as

$$P(y^i | x^i, \beta) = P(y^i = 1 | x^i, \beta)^{yi} P(y^i = 0 | x^i, \beta)^{1-yi}$$

, where $P(y^i = 1 | x^i, \beta) = \frac{e^{\beta^T x^i}}{1 + e^{-\beta^T x^i}}$, and $P(y^i = 0 | x^i, \beta) = \frac{1}{1 + e^{\beta^T x^i}}$

This then gives the conditional log-likelihood

$$l(\beta \mid X,Y) = \sum_{i=1}^{n} y^i \log(P(y^i = 1 \mid x^i, \beta)) + (1 - y^i)\log(P(y^i = 0 \mid x^i, \beta))$$

To find the maximum likelihood estimators of β, we would take the gradients of the expression above and set them equal to 0 to find the solutions.

$$\frac{\partial l(\beta \mid X,Y)}{\partial \beta_k} = \sum_{i=1}^{n} y^i \frac{1}{P(y^i = 1 \mid x^i, \beta)} \frac{\partial P(y^i = 1 \mid x^i, \beta)}{\partial \beta_k} - (1 - y^i)\frac{1}{P(y^i = 0 \mid x^i, \beta)} \frac{\partial P(y^i = 0 \mid x^i, \beta)}{\partial \beta_k}$$

$$\frac{\partial l(\beta \mid X,Y)}{\partial \beta_k} = \sum_{i=1}^{n} x_i^k (yi - P(y^i = 1 \mid x^i, \beta)$$

Because of the nonlinear nature of the parameters, there is no closed-form solution to these equations and they must be solved iteratively using numerical methods such as the Newton-Raphson method. The details of the method will not be discussed in this report as we will focus on the problem of testing hypotheses of the coefficients. Suppose we have successfully estimated all the coefficients, $\hat{\beta}$, using numerical methods, we can then use hypothesis testing to evaluate if the predictors have statistically significant associations with the output variable.

Based on the large-sample distribution of the maximum likelihood estimator, we can apply the Wald test for this problem. For any coefficient, we have the following hypothesis testing set up,

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Concerning the significance of the coefficient, we can calculate the ratio of the estimate to its standard error as follows,

$$z = \frac{\hat{\beta} - \beta_{j0}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1)$$

, where $SE(\hat{\beta})$ is calculated by taking the inverse of the estimated information matrix.

Going back to our case and applying logistic regression to the COVID-19 patient dataset, we have the following logit function,

$$\begin{aligned} &\log it(\pi_i) = \beta_0 + \beta_1 age + \beta_2 sex.female + \beta_3 chronic.disease.1 + \beta_4 respiratory.distres.syndrome.1 + \\ &\beta_5 respiratory.failure.1 + \beta_6 chest.distress.1 + \beta_7 shortness.of.breath.1 + \beta_8 heart.failure.1 + \\ &\beta_9 runny.nose.1 + \beta_{10} spetic.shock.1 + \beta_{11} sore.throat.1 + \beta_{12} anorexia.1 + \beta_{13} arrhythmia.1 + \\ &\beta_{14} cough.1 + \beta_{15} diarrhea.1 + \beta_{16} dizziness.1 + \beta_{17} fatigue.1 + \beta_{18} fever.1 + \beta_{19} headache.1 + \\ &\beta_{20} \inf arction.1 + \beta_{21} malaise.1 + \beta_{22} mya \lg ial.1 + \beta_{23} phlegm.1 + \beta_{24} pneumonial.1 + \\ &\beta_{25} sepsis.1 + \beta_{26} soreness.1 \end{aligned}$$

Before we interpret the model, we would need to first ensure the quality of the model. To evaluate the quality and accuracy of the logistic regression model, we can use k-fold cross-validation. The goal of cross-validation is to evaluate the model's ability to generalize and predict unseen data, in order to flag potential problems such as overfitting or selection bias. It provides insights on the model's level of robustness and generalization on new data that is not a part of its training data. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (i.e. the training set), and validating the analysis on the other subset (i.e. the validation set or testing set). To reduce variability, we can repeat this procedure k times by initially partitioning the data into k subsets. Figure 15 demonstrates a visual summary of the process when k = 5.

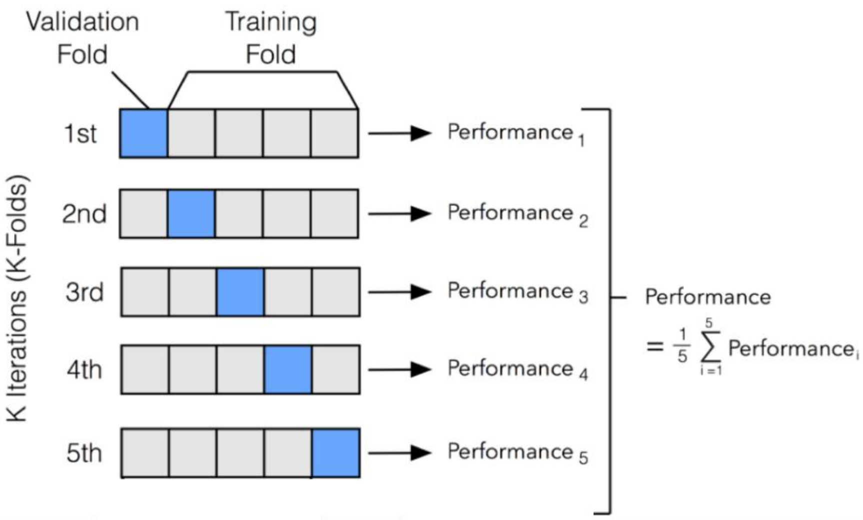


Figure 15. 5-Fold Cross Validation

Using the caret and glmnet packages in R, we are able to perform a 5-fold cross-validation to compute the overall accuracy and the ROC curve of the logistic regression model. After repeating the same procedures for logistic regression with L1 and L2 regularization, it was found that the regular logistic regression had the best performance. According to the results, the model has an overall accuracy of 0.900 with a standard deviation of 0.030. The ROC curve is shown in figure 16.

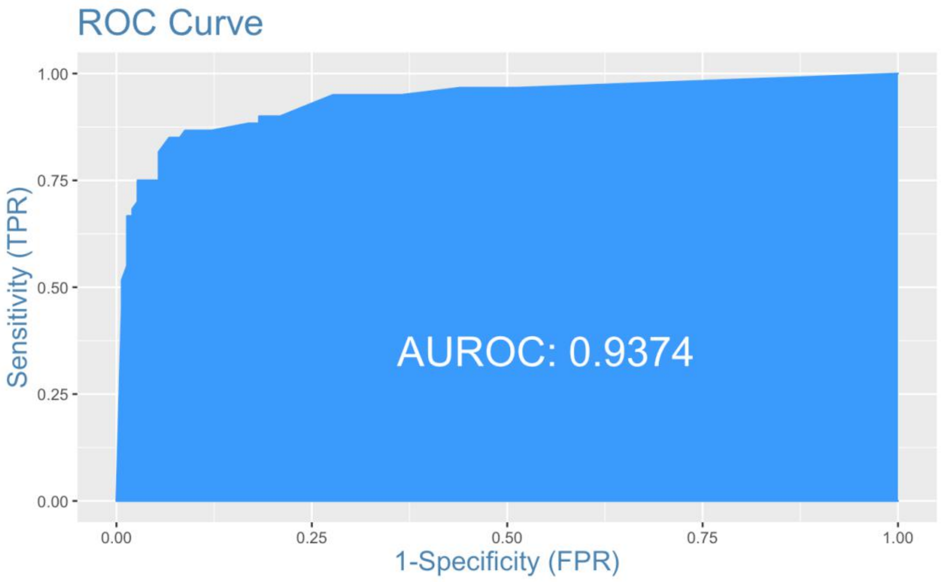


Figure 16. ROC Curve

After confirming the quality of the model, we can apply the same model onto the whole dataset and interpret the coefficient table from the output. The output coefficient table is displayed in the table below.

Variable	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-8.10	0.73	-11.11	**<0.001**
age	0.11	0.01	10.09	**<0.001**
sexFemale	-0.37	0.30	-1.23	0.22
chronic_disease_binary1	0.51	0.42	1.23	0.22
respiratory_distress_syndrome1	19.53	1402.78	0.01	0.99
respiratory_failure1	19.46	1832.13	0.01	0.99
chest_distress1	18.38	4619.90	0.00	1.00
shortness_of_breath1	2.56	1.11	2.30	**0.02**
heart_failure1	17.76	3812.91	0.00	1.00
runny_nose1	-17.40	3196.51	-0.01	1.00
septic_shock1	16.09	2113.25	0.01	0.99
sore_throat1	-0.04	1.15	-0.04	0.97
anorexia1	17.33	7604.24	0.00	1.00
arrhythmia1	12.15	2892.88	0.00	1.00
cough1	0.69	0.60	1.15	0.25
diarrhea1	2.42	9.78	0.25	0.80
dizziness1	18.96	10754.01	0.00	1.00
fatigue1	1.43	1.06	1.35	0.18
fever1	0.79	0.50	1.59	0.11
headache1	0.94	4.41	0.21	0.83
infarction1	19.96	3750.52	0.01	1.00
malaise1	-18.27	5054.53	0.00	1.00
myalgia1	-17.88	4579.83	0.00	1.00
phlegm1	-15.06	4911.26	0.00	1.00
pneumonia1	2.87	1.07	2.70	**0.01**
sepsis1	13.85	4111.91	0.00	1.00

4.5 Key Findings

From this section, we can derive that fever and cough are the two major symptoms associated with COVID-19. Moreover, pneumonia and shortness of breath as discovered to be significant risk factors as they potentially increase one's likelihood of dying from the disease, controlling for other factors.

5. Patient Demographics

5.1 Overview

This section of the COVID-19 real-time tracker shows a summary visualization of the distributions of demographic characteristics of selective patients (Figure 17).

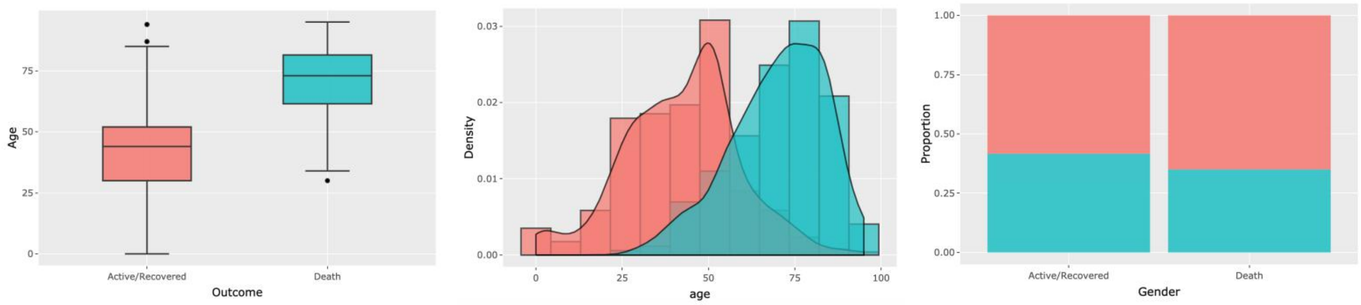


Figure 17. Summary Visualizations of Demographics Characteristics of Selective Patients

5.2 Two-Sample T-Test

To determine if there is statistically significant difference in the age of two patient groups, active/recovered or death, we can conduct a two-sample t-test. Two-sample t-test is a hypothesis testing method to compare two continuous-data distributions. More precisely, it tests to determine if the means of two continuous-data distributions are equal. There are a number of assumptions that need to be satisfied in order to use the two-sample t-test properly, and they are listed as follows,

i. The data are continuous (not discrete)
ii. The data follow the normal probability distribution
iii. The variances of the two populations are equal, if not, use un-pooled variances to calculate the test statistics
iv. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other
v. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample

Assumption (i) is satisfied as the value of age is continuous. However, assumptions (iv) and (v) may not be valid due to potential data quality issues such as missing data. We will presume they are satisfied and proceed with cautions. For assumptions (ii), we can validate the data's normality using QQ plots as follows,

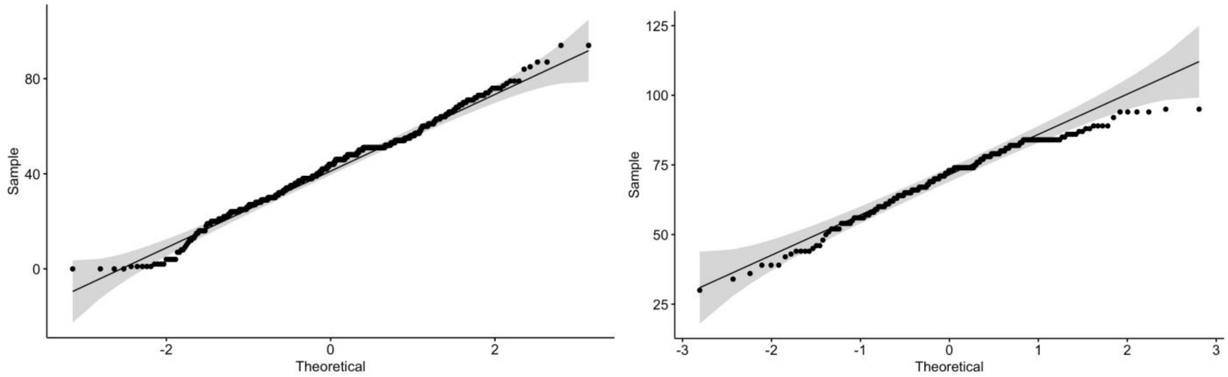


Figure 18. QQ Plots of Ages of Active/Recovered Patients (Left) and Dead Patients (Right)

The data points appear to be decently consistent with the quantiles of a normal distribution. For assumption (iii), we can apply the F-test of equality of variances as follows,

$$H_0 : \sigma_X = \sigma_Y$$
$$H_1 : \sigma_X \neq \sigma_Y$$
$$F = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1)$$

, where S_X denotes and sample standard deviation of age of active/recovered patients, S_Y denotes and sample standard deviation of age of dead patients, n and m denote the sample sizes of two groups. After conducting the hypothesis test, we obtained a p-value of 0.00097. Thus, we have sufficient evidence to reject the null hypothesis and conclude the variances of the two groups are unequal at the alpha level of 0.05.

Consequently, we proceed to conduct a two sample t-test with un-pooled variances. The steps are demonstrated below,

$$H0: \mu_X = \mu_Y$$
$$H1: \mu_X \neq \mu_Y$$
$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t(v)$$
$$,\text{where } v = \frac{(\frac{S_X^2}{n} + \frac{S_Y^2}{m})^2}{\frac{(\frac{S_X^2}{n})^2}{n-1} + \frac{(\frac{S_Y^2}{m})^2}{m-1}}$$

As a result, we obtained a p-value that is approximately 0 that allowed us to reject the null hypothesis at the alpha level of 0.05. Hence, we have sufficient evidence to conclude that the average age of patients who are active or recovered is different from the average of patients who have died from COVID-19.

5.3 Chi-Square Test

To determine if there is statistically significant association between a patient's gender and a patient's outcome, we can conduct a Chi-Square test as a test of association on a 2x2 contingency table. There are a number of assumptions that need to be satisfied in order to use the two-sample t-test properly, and they are listed as follows,

i. The data in the cells should be frequencies, or counts of cases
ii. The levels (or categories) of the variables are mutually exclusive
iii. Each observation is independent of all the others
iv. The value of the expected values should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one

Assumptions (i) and (ii) are met since we are observing counts of patients who are either male or female, and either active/recovered or deceased. In addition, assumptions (iv) is satisfied as shown by the 2x2 contingency table below. We will presume assumption (iii) to hold true and proceed.

After filtering the data to create a subset of patient data with recorded genders and outcomes, we can form the following 2x2 contingency table,

	Active/Recovered	Death
Male	299	132
Female	213	71

We cam calculate the Chi-square test statistic as follows,

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim X^2(r-1, c-1)$$
$$,\text{where } Ei, j = (R_i \cdot C_j) / n$$

As a result, we obtained a p-value of 0.1025, which does not provide sufficient evidence in rejecting the null hypothesis. Thus, we failed to reject the null hypothesis at the alpha level of 0.05 and conclude that a patient's gender has no statistically significant association the patient's outcome.

5.4 Key Findings

From this section, we can derive that the average age of patients who are active or recovered is different from the average of patients who have died from COVID-19, and older populations are more vulnerable to a negative outcome of the disease. In contrast, gender does not appear to have a strong association with the outcome. Both findings are consistent with the results of the logistic regression model from the previous section as both age and gender are incorporated as a part of the model.

6. Conclusion

This research presented the latest trends of COVID-19 across different regions and insights of COVID-19's symptoms and patient demographics as visualized in the real-time COVID-19 tracker. In addition, the research dives into deeper details of the methodologies behind the real-time COVID-19 tracker, which include real-time data retrieval, data transformation and normalization, time-series forecast with ARIMA model, text mining techniques, and logistic regression model. However, we need to be cautious about accepting the conclusions as there are potential data quality issues, in which case the patient-level data has a substantial amount of missing data and erroneous entries. To verify the findings in this research, we can try reproducing the derived insights when we have access to an updated dataset towards the end of the pandemic.

During a global-level pandemic such as COVID-19, it is paramount for the public to have access to the latest status of the outbreak and be well-informed of relevant insights of the disease. A platform such as a real-time COVID-19 tracker will assist the public community to disseminate accurate and reliable insights into the spread of COVID-19. The research and effort behind the tracker are motivated by the social responsibility to spread awareness to the common public by providing scientific-based data analysis, prediction, and relevant findings. This paper and research project is still ongoing research as many more investigations regarding COVID-19 can be carried out. It will serve as an initial step to unravel the many uncertainties that revolve around this global pandemic.

References

[1] Nau, Robert. "Introduction to ARIMA Models." Introduction to ARIMA Models, people.duke.edu/~rnau/Slides_on_ARIMA_models--Robert_Nau.pdf.

[2] "Estimation and Hypothesis Testing for Logistic Regression." Estimation and Hypothesis Testing for Logistic Regression, courses.washington.edu/b515/l13.pdf.

[3] "Two-Sample T-Test." Two-Sample T-Test, ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Two-Sample_T-Test.pdf.

[4] McHugh, Mary. "The Chi-Square Test of Independence." 2013.

[5] Shumway, Robert H, and David S. Stoffer. Time Series Analysis and Its Applications: With R Examples. New York: Springer, 2006.

[6] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.