**Time series analysis with ARIMA model and prediction**

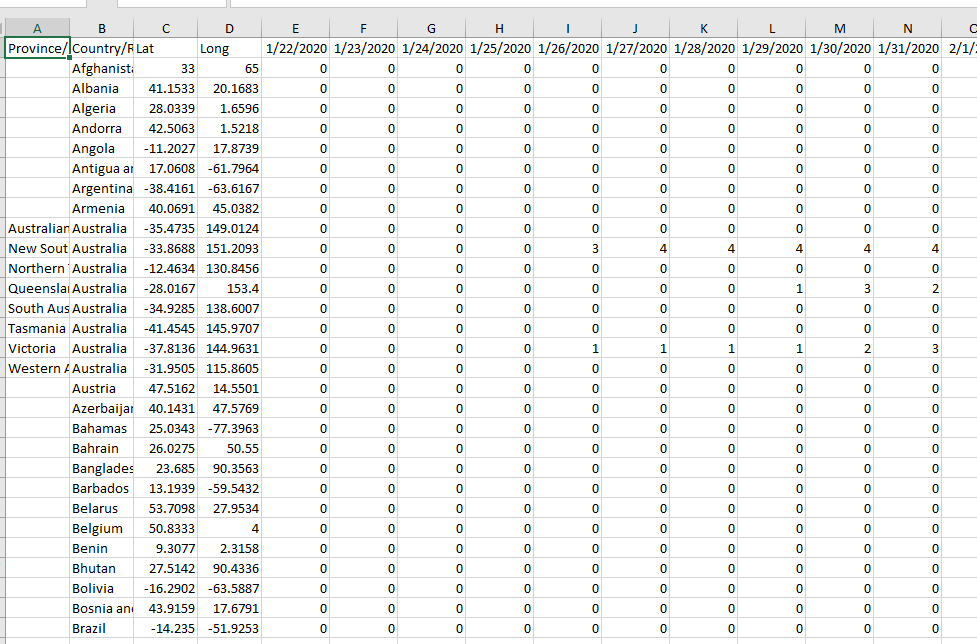
[**Data set**](https://github.com/CSSEGISandData/COVID-19)**:** CSSE, Johns Hopkins University

Tools: Rstudio v1.25

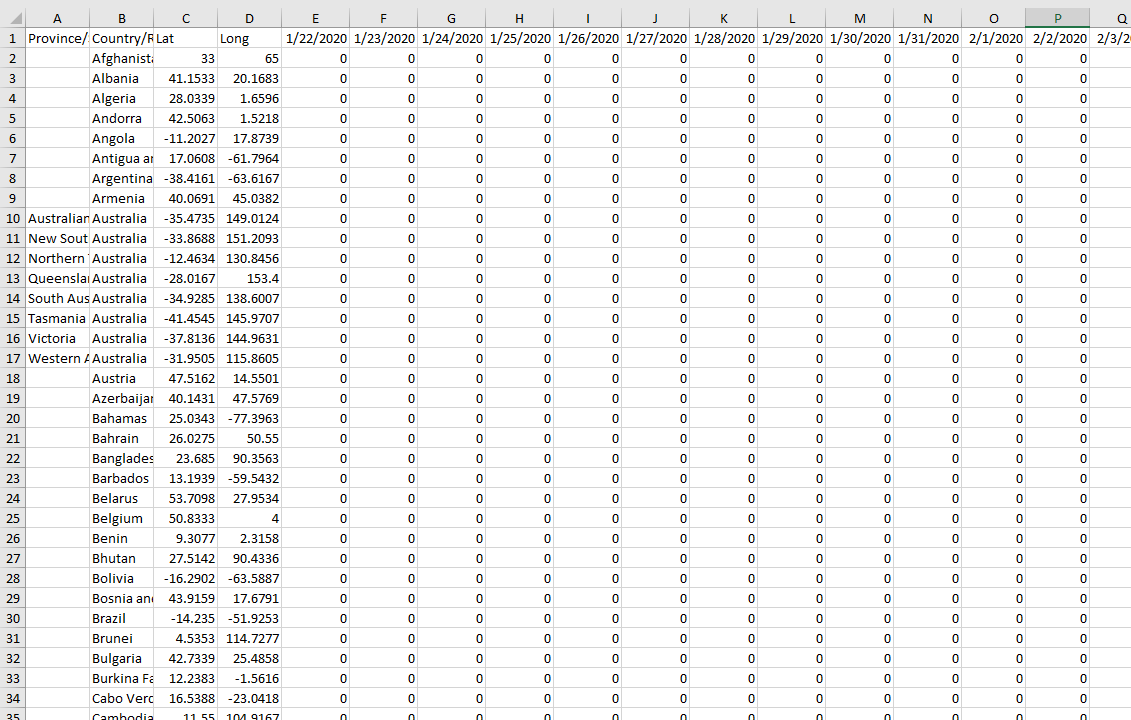
1. Data cleaning and engineering

Raw data:

[time\_series\_covid19\_confirmed\_global.csv](https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv)



[time\_series\_covid19\_deaths\_global.csv](https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv)



***Step 1, Pulling and tidying the Johns Hopkins Covid-19 data to long format***

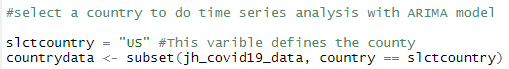
We need to build a function to read .csv files and convert them to long format. Next, we pull official country level indicators from the UN Statistics Division to get country level identifiers. Merging by country name is messy. I start with a fuzzy matching approach by using the [**stringdist**](https://www.rdocumentation.org/packages/stringdist/versions/0.9.5.5/topics/stringdist)package.



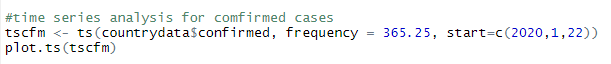
As you can see, there are 2 main columns: confirmed (confirmed cases) and deaths (death cases) from 22nd Jan. 2020 to 19th Apr. 2020.

***Step 2, selecting a country to analyze.***

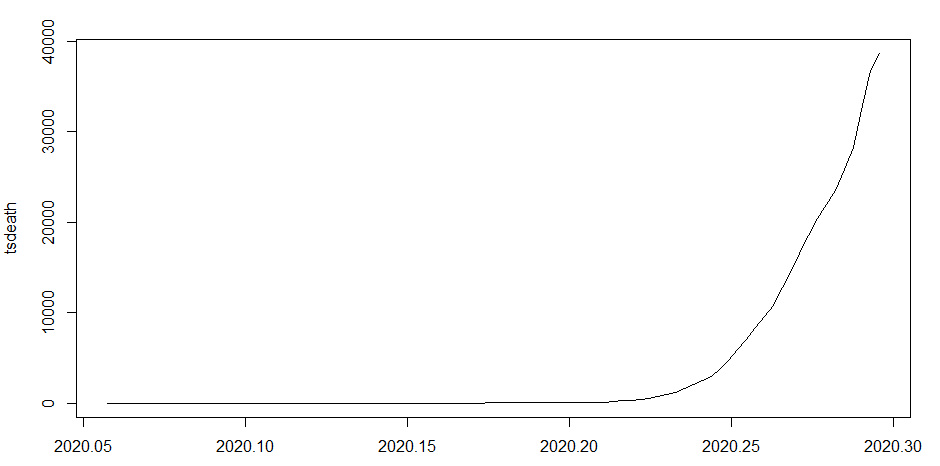
In order to take a country's data from the DF for analysis, we need to first define a country name variable, and then use this variable, called an indicator, to create a subset and "sample" the country.



***Step 3, Converting to ts format and using*** [***auto.arima***](https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/auto.arima) ***function to analyze***

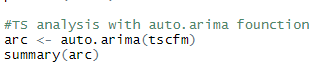


We converts([ts founction](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/ts)) the data frame into time series by using frequency 365 (in days), and it will start with 22nd Jan. 2020. Then we plot it:

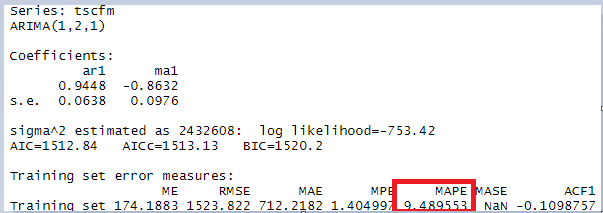


As we can see from the plot, take the death case of US for instance, the number of deaths began to increase significantly around the 25th day.

Then we can feed them into ARIMA (Autoregressive Moving Average Model) models:



By summarizing the model, we learn that the error rate (**MAPE: Mean absolute percentage error**) is 9.489553, which is not too bad. And the ARIMA model here is ARIMA(1, 2, 1).



*ARIMA (p, d, q):*

*p- Represents the lags of the time series data itself used in the prediction model, also called the AR/Auto-Regressive term.*

*D--Represents the time series data that needs to be differentialized by several orders of magnitude to be stable, also called the integrated term.*

*q - Represents the lagged number (lags) of prediction errors used in the prediction model, also called the MA/Moving Average term.*

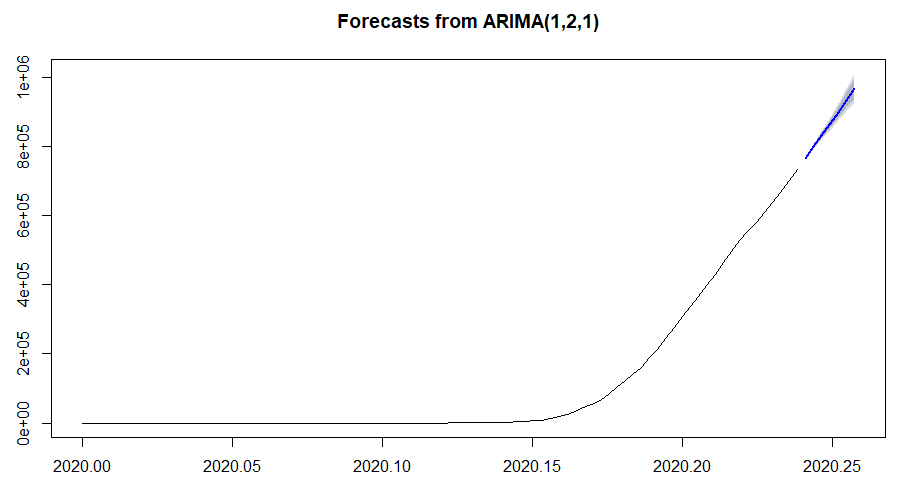
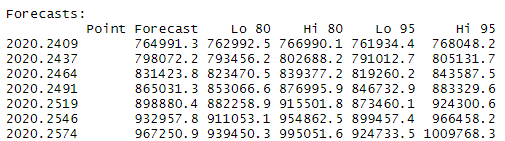
**So here the ARIMA (1, 2, 1) model means: The level of the process, measured as the most resent value, plus a trend measured as the most recent change in the process.**

**More specifically: d = 2, adaptive trend in addition on the level.**

***Step 4, Prediction***



Use the current model to do prediction for the next week (7 days).

As we can see from the forecast matrix, on day one, there is a 95% probability that the predicted value will be between 761934.4 and 768048.2, and an 80% probability that the predicted value will be between 762992.5 and 766990.1...

***Step 4, Residual analysis：***



