COronaVIrus Disease 2019

A PREDICTION OF THINGS TO COME

CHAN DOMINIC
- DSI13 -

TABLE OF CONTENTS

In January 2020, a novel coronavirus, SARS-CoV-2, was identified as the cause of an outbreak of viral pneumonia in Wuhan, China. The disease, later named coronavirus disease 2019 (COVID-19), subsequently spread globally. In the first three months after COVID-19 emerged nearly 1 million people were infected and 50,000 died.

Similar to the severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) of the past.

MERS is a viral respiratory disease that was first reported in Saudi Arabia in September 2012 and has since spread to 27 countries. From its emergence through January 2020, WHO confirmed 2,519 MERS cases and 866 deaths (about 1 in 3).

Infection with SARS coronavirus (SARS-CoV) can cause a severe viral respiratory illness. SARS was first reported in Asia in February 2003, though cases subsequently were tracked to November 2002. SARS quickly spread to 26 countries before being contained after about four months. More than 8,000 people fell ill from SARS and 774 died. Since 2004, there have been no reported SARS cases.

02

PROBLEM STATEMENT

Determining the significance of variables to accurately predict the global COVID-19 cases; **to effectively predict their eventual progression of deaths and cases by means of the Regression models**. This is to ensure that global citizens are made **aware to the situation's progression** and can make the necessary preparations should there be an extended lockdown period. Goverenments and private companies can leverage on this model to make **a informed decision as to growth of the economy and businesses.**

covid19 SG

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Center for Systems Science and Engineering

kaggle

SG RELATED DATA

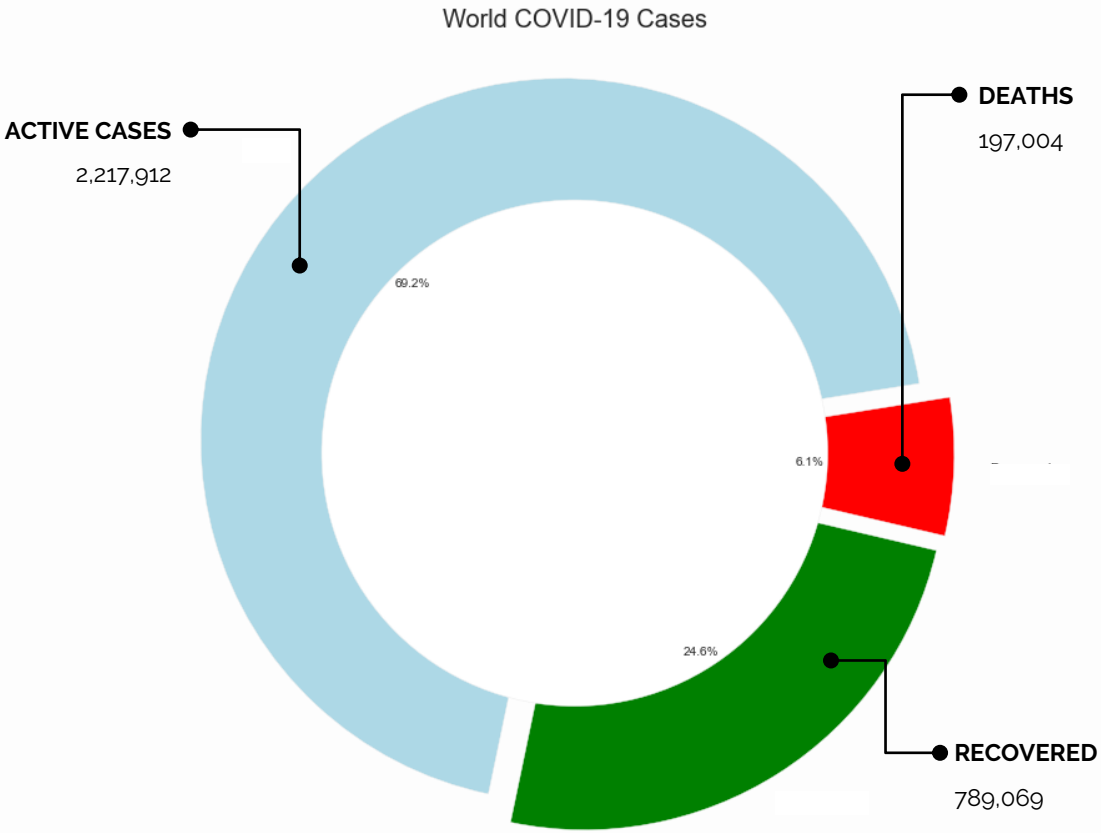
co.vid19.sg provided detailed information of the individual cases up till 19 April 20

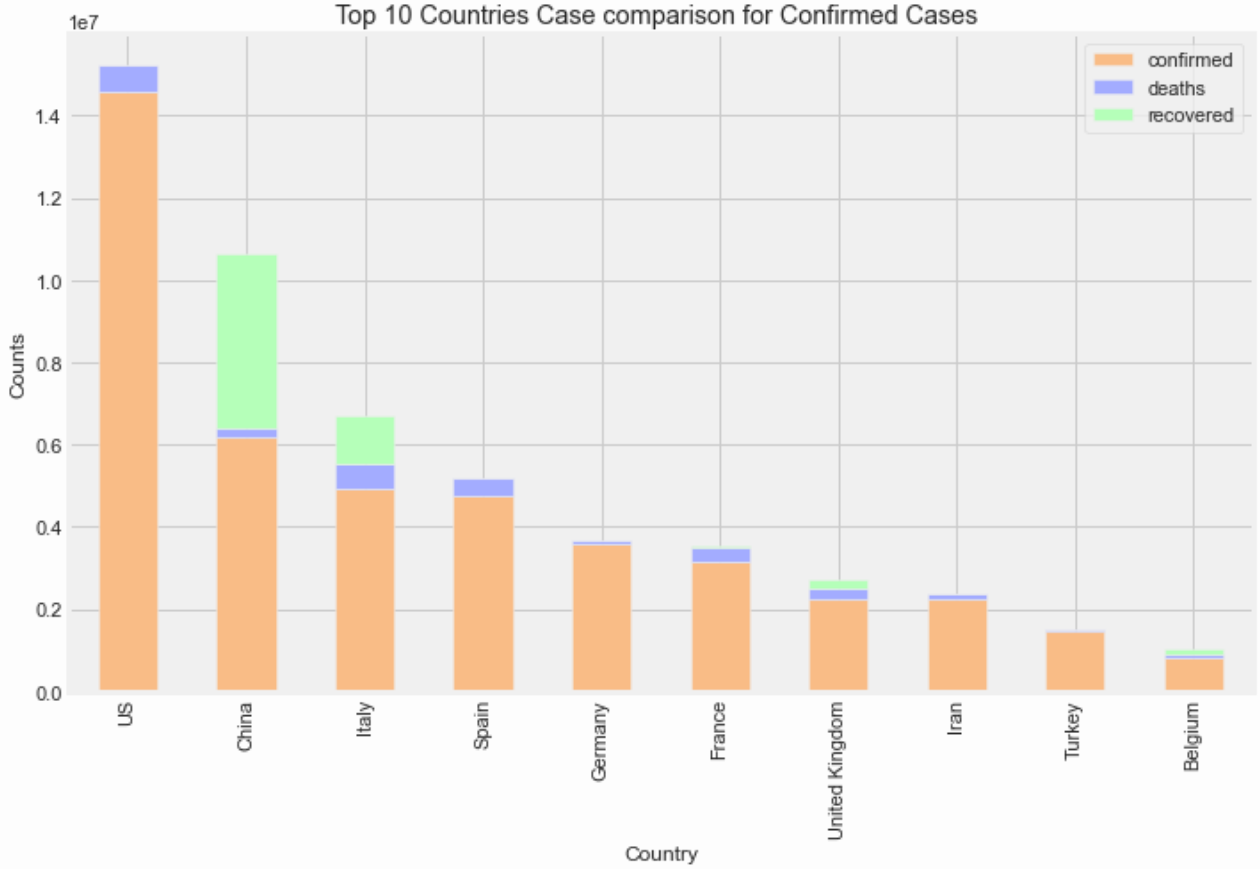
GLOBAL SOURCE

Dataset operated by the Johns Hopkins University Center for Systems Science and Engineering and was updated daily. This was also the source to the Kaggle dataset

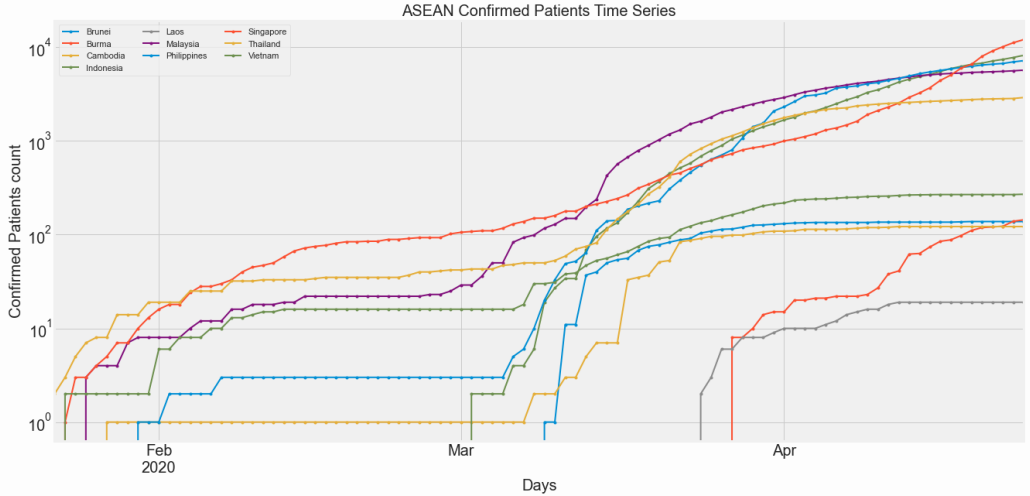
COMPETITION WEEK 4

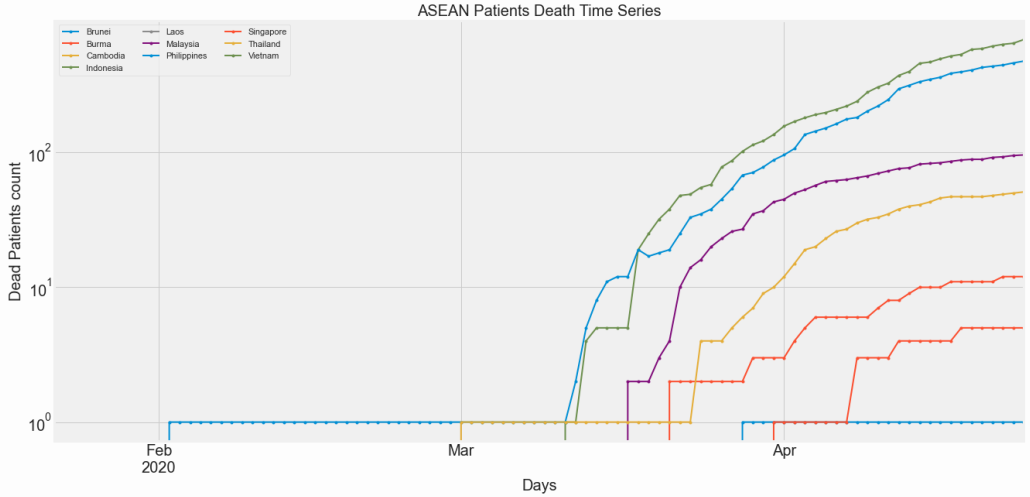
The challenge involves forecasting confirmed cases and fatalities between April 15 and May 14. Training set contain features from prior.

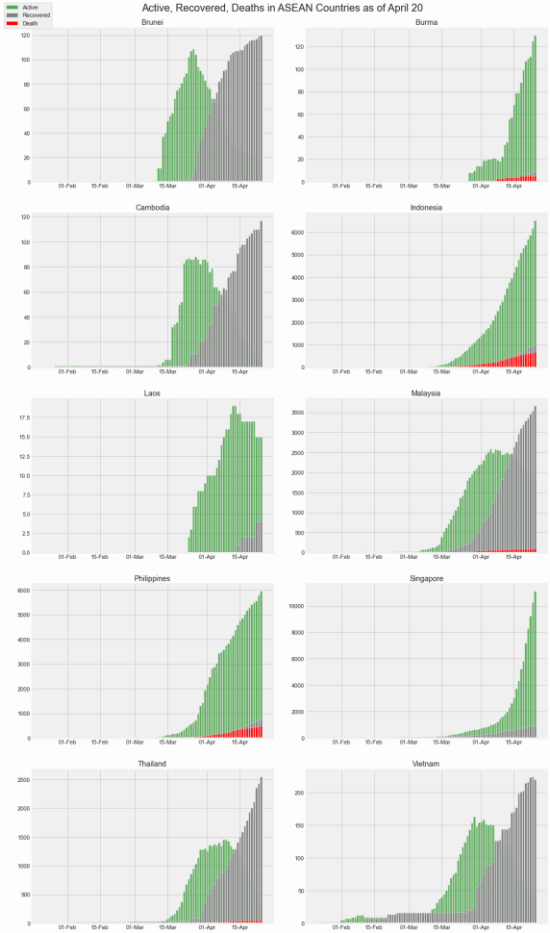

World COVID-19 Cases
ACTIVE CASES
2,217,912
DEATHS
197,004
RECOVERED
789,069
69.2%
6.1%
24.6%

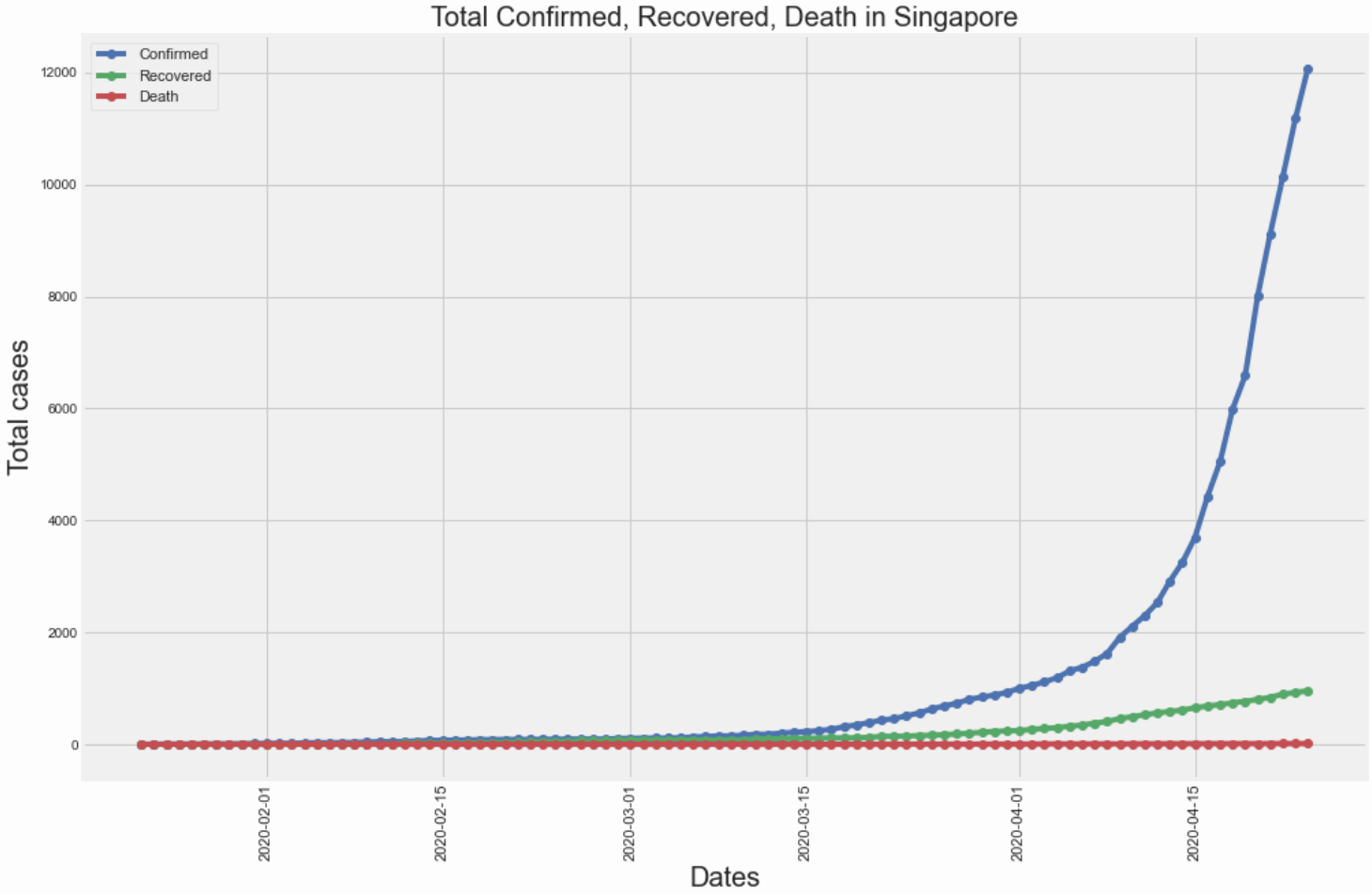

Top 10 Countries Case comparison for Confirmed Cases
1e7
confirmed
deaths
recovered
Counts
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
US
China
Italy
Spain
Germany
France
United Kingdom
Iran
Turkey
Belgium
Country

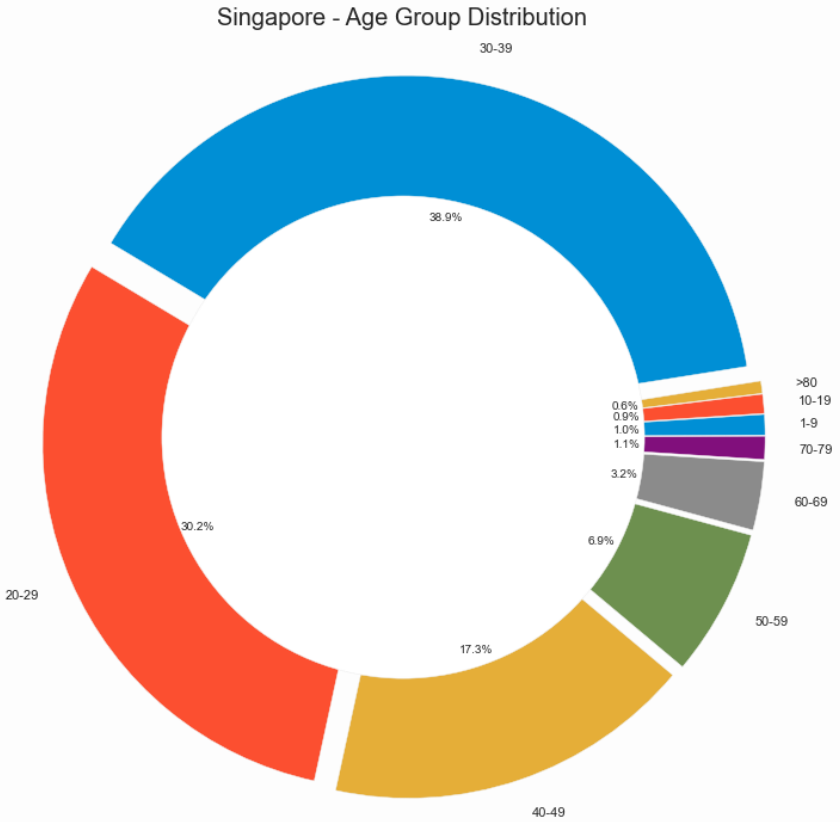
04 EXPLORATORY DATA ANALYSIS

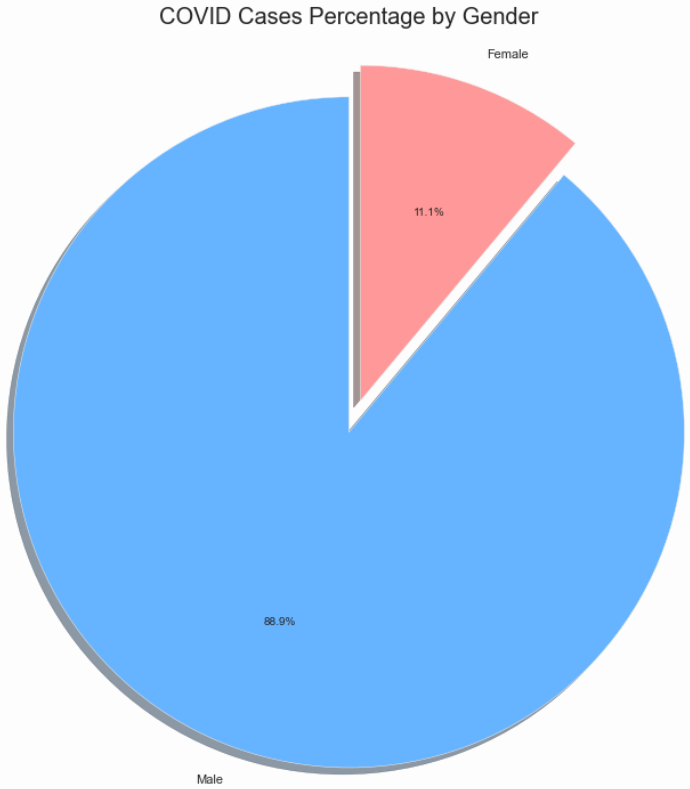


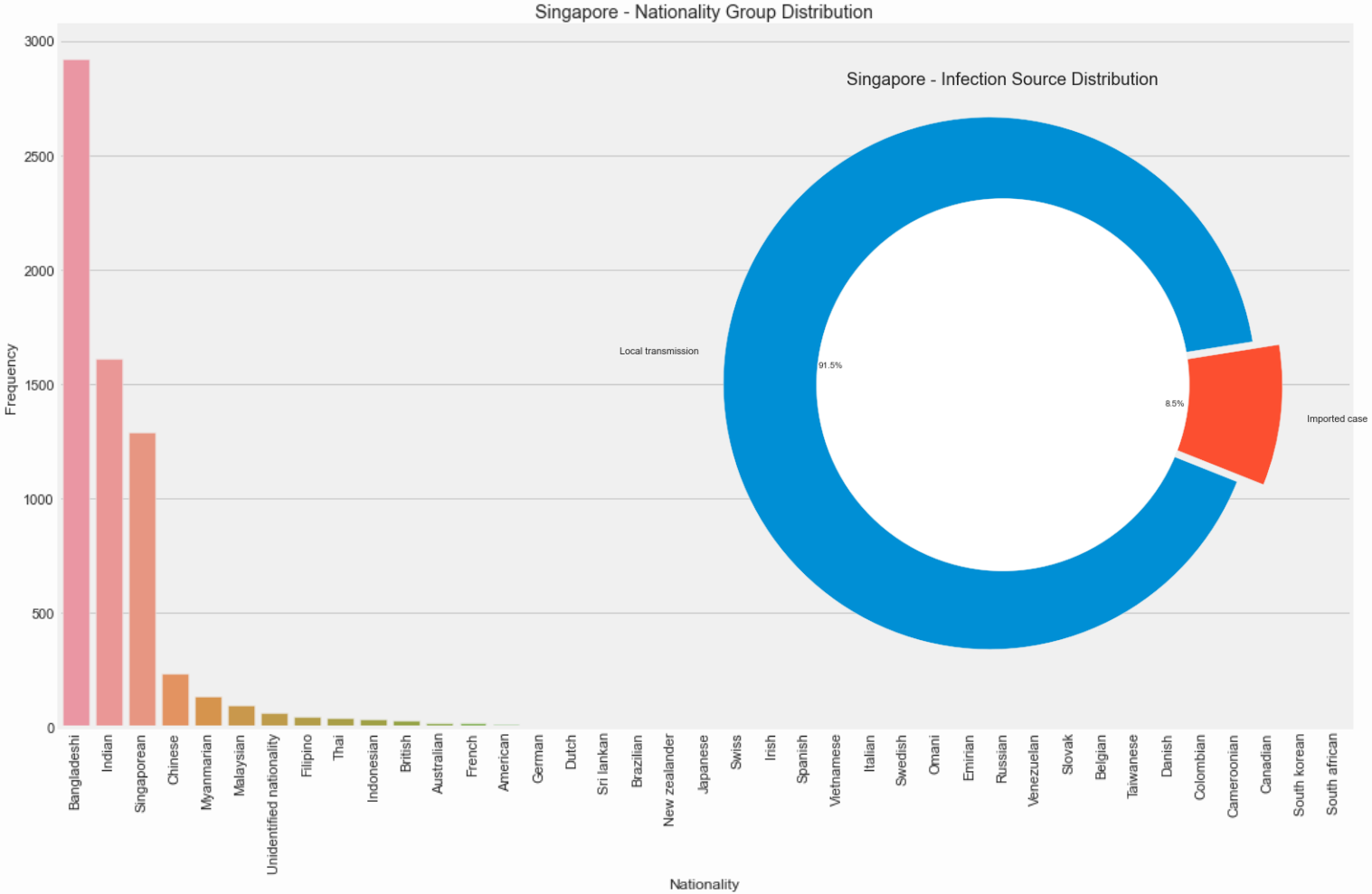


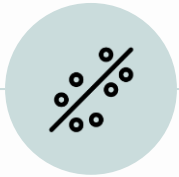



Total Confirmed, Recovered, Death in Singapore
Confirmed
Recovered
Death
Total cases
12000
10000
8000
6000
4000
2000
0
2020-02-01
2020-02-15
2020-03-01
2020-03-15
2020-04-01
2020-04-15
Dates


Singapore - Age Group Distribution
30-39
38.9%
20-29
30.2%
40-49
17.3%
50-59
6.9%
60-69
3.2%
70-79
1.1%
1-9
1.0%
10-19
0.9%
>80
0.6%


COVID Cases Percentage by Gender
Female
11.1%
Male
88.9%


Singapore - Nationality Group Distribution
Frequency
3000
2500
2000
1500
1000
500
0
Bangladeshi
Indian
Singaporean
Chinese
Myanmarian
Malaysian
Unidentified nationality
Filipino
Thai
Indonesian
British
Australian
French
American
German
Dutch
Sri lankan
Brazilian
New zealander
Japanese
Swiss
Irish
Spanish
Vietnamese
Italian
Swedish
Omani
Emirian
Russian
Venezuelan
Slovak
Belgian
Taiwanese
Danish
Colombian
Cameroonian
Canadian
South korean
South african
Nationality
Singapore - Infection Source Distribution
Local transmission
91.5%
8.5%
Imported case



For scraped data, the linear regression model performed the worst on the training set,, though it performed the best on the validation set.

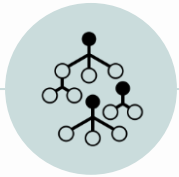
For Kaggle dataset, linear regression was significantly the worst performing model.

LINEAR REGRESSION

For scraped data, the RF model performed average on the training set, though it performed worst on the validation set.

For Kaggle dataset, RF was the best perform model.

RANDOM FOREST

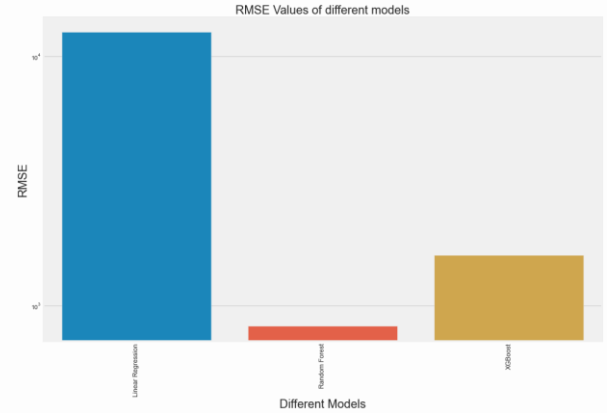


For scraped data, the XGB model performed best on the training set, though it displayed average performance on the validation set.

For Kaggle dataset, XGB was the 2nd best performing model.

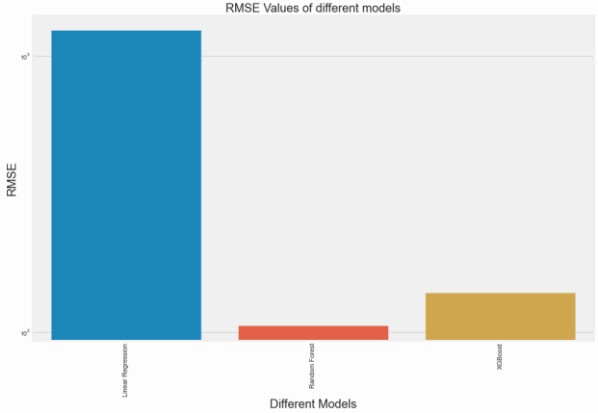
XGBOOST

ConfirmedCases



model	mse	rmse
Linear Regression	157470458.21	12548.72
Random Forest	**687457.82**	**829.13**
XGBoost	2543786.22	1594.93

Fatalities



model	mse	rmse
Linear Regression	1529577.74	1236.76
Random Forest	**11173.06**	**105.7**
XGBoost	19406.37	139.31

Overview Data Notebooks Discussion **Leaderboard** Rules Team My Submissions **Late Submission**

⚠ This competition has completed. This leaderboard reflects preliminary final standings. The result will become final after the competition organizers verify the results.

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	just now	0 seconds	0 seconds	0.68772

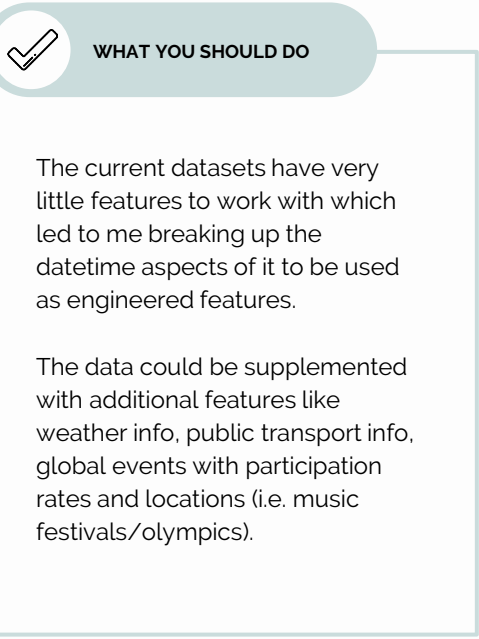
Complete

Jump to your position on the leaderboard ▾

Recommendation

The model can serve as a **prediction model for countries and businesses**. Additionally, citizens can be more discerning of the potential for number of cases and deaths to rise (or fall) with time.

Notwithstanding this, more data, as mentioned above, will serve to optimise the model**. The model can also serve as a starting point, to be extrapolated to individual countries with additional research** to be done marrying the features together.



THANKS!

Does anyone have any questions?
domchanjl@gmail.com
+65 8112 7772
yourcompany.com

RESOURCES

https://github.com/CSSEGISandData/COVID-19
https://co.vid19.sg/singapore/dashboard.
https://www.niaid.nih.gov/diseases-conditions/covid-19