FORECASTING NUMBER OF RECOVERED COVID-19 PANDEMIC CASES ACROSS THE GLOBE

FATIMAH BINTI MOHD NIZAM

17218825

FACULTY OF COMPUTER SCIENCE AND TECHNOLOGY
MASTER IN DATA SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

Table of Contents

INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The novel coronavirus is related to the virus that causes severe acute respiratory syndrome (SARS). The COVID-19 outbreak had affected 213 countries and it started in the early of December in 2019, when the disease spread rapidly among the residents of Wuhan City, Hubei Province, China (Aina Umaira Md Shah, 2020). Quick early findings showed that number of patients started to being affected at the Huanan seafood market in Wuhan City. It is a market where a lot of exotic animals which are high potential carriers of various viruses and bacteria, due to their eating habits and habitats.

The virus mostly spread through person-to-person contact. This does not only limit to physical contact through skin touch. It can also cause infection when people who with COVID-19 cough. The tiny droplets expelled contains the virus have the potential to affect other people with the virus. Fever, fatigue, dry cough, and upper respiratory symptoms (nasal congestion and running nose), vomiting and diarrhea are the main clinical symptoms of COVID-19 infections (Zare-Zardini, Soltaninejad, Ferdosian, Hamidieh, & Memarpoor-Yazdi, 2020).

The first case reported outside of China was in Thailand on 13 January 2020 and the number of outbreaks is assumed to reach its peak in late of May 2020 and started to drop early July 2020. (Fairoza Amira binti Hamzah, March 2020). China, the United State and other countries have been enforcing instituted temporary restrictions on travel to slow the spread of the disease in their countries and throughout the rest of the world (Anthony S. Fauci, 2020). The mobility restrictions seem to have effectively contribute to save lives by protecting the spreading of the disease (D, et al., 2020).

There are a few drivers that lead to the Covid-19 widely spread across the globe in a rapid rate. The role of climate (temperature and precipitation), region-specific COVID-19 susceptibility (malaria incidence, percentage of population aged over 65 years), and human international mobility that involves the amounts of international visitors in the regions (Yasuhiro Kubota, 2020).

This project is conducted to analyze the impact of strategies and the actions taken to reduce the COVID-19 spread cases and increase the number of recovered cases across the globe. This project is to gain a new insight on whether the spread of this disease has a potential to show a pattern that will leads to a controllable trend.

OBJECTIVES

The objectives of this project are:

1. To analyze the data patterns gained from number of confirmed, deaths and recovered Covid-19 cases in 2020.
2. To predict the number of recovered case results from the Covid-19 pandemic using Polynomial Regression, Holt's Linear Model, Holt's Winter Model and ARIMA Model.

RESEARCH BACKGROUND

There are 3 type of cases that need to be considered when it comes to COVID-19: The number of confirmed cases, number of recovered cases and the number of death cases. The number of active cases can be obtained by subtracting the number of recovered and death cases from the confirmed cases.

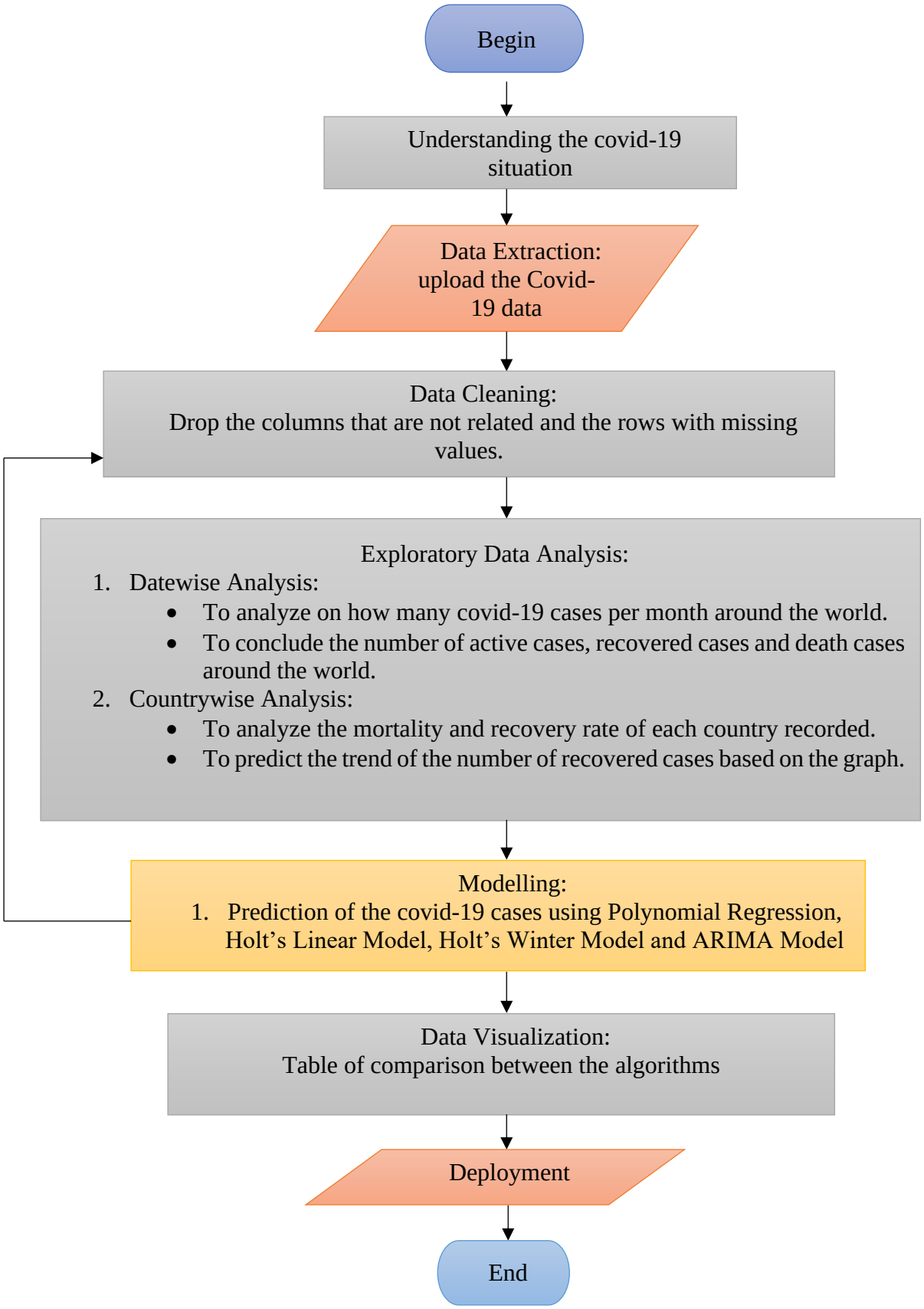
There is a big significance of understanding on how the COVID-19 pandemic is handled around the world. To view the effectiveness of the steps or the strategies taken by the governments to handle the cases, it is essential for us to evaluate the pattern of the number of recovered cases, whether the COVID-19 is handled by taking the right action. If the number of recovered cases is efficiently reduced, the actions taken need to prolong until the transmission of COVID-19 is eliminated.

This can be proven by taking New Zealand as an example. There were no new community transmissions in 17 days and all patients are fully recovered until they were confident to lift the lockdown in the country.

ANALYSIS AND DESIGN

1. Suitable tools for system development based on the objectives given: Python

2. Algorithm (system flow) to process the data according to your objectives.



3. Machine Learning Algorithms

3.1 Polynomial Regression

Regression analysis involves in extracting information from the relationship between a dependent variable and one or more independent variables. Polynomial regression is useful when there is a curvilinear relationship between two numerical variables (Ostertagova, 2012). It is an algorithm for a special case where there is only one independent variable X. Polynomial regression model can be expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + e_i \text{ , for } i = 1, 2, \ldots, n$$

Equation 1: Polynomial Regression Model Equation

Where k is the degree of the polynomial. The degree of the polynomial is the order of the model.

3.2 Holt's Linear Model

Holt's Linear Model or exponential smoothing model is robust in forecasting and it helps in dealing with trending data (Yapar, Capar, Selamlar, & Yavuz, 2017). The goal of an exponential smoothing model is to estimate the final values of the level, trend and seasonal pattern and apply the result to construct forecasts.

Holt's Linear Model involves a forecast equation and two smoothing equations (one for the level and one for the trend):

$$
\begin{aligned}
\text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\
\text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1},
\end{aligned}
$$

Equation 2: Holt's Linear Model Equation

ℓ_t = estimate of the level of the series at time t.

b_t = estimate of the trend (slope) of the series at time t.

α = smoothing parameter for level

β^* = smoothing parameter for trend

3.3 Holt's Winter Model

The Holt's Winter Model was developed to predict trends and seasonality from exponentially weighted averages. The additive version of the model can be described by the following formulas:

$$Lj=\alpha(yj-Sj-s)+(1-\alpha)(Lj-1+bj-1)\ bj=\beta(Lj-Lj-1)+(1-\beta)bj-1$$

$$Sj=\gamma(yj-Lj)+(1-\gamma)Sj-s\ (2)\ Fj+1=Lj+bj+Sj+1-s$$

$$j=s+1,s+2,\ldots,$$

$\alpha, \beta, \gamma \in [0,1]$ = smoothing parameters

Lj = smoothed level at time j

bj = change in the trend at moment j

Sj = the seasonal smooth at moment j

s = number of periods in the season

$Fj+1$ is one step ahead forecasted value

3.4 ARIMA Model

Autoregressive integrated moving average (ARIMA) model has many applications such as to capture the time correlation and possibility of distribution of determined wind-pace time collection records is offered (Almasarweh & Wadi, 2018).

EXPERIMENTAL RESULT

1. The dataset was cleaned by dropping the null values and the features that are not needed for the analysis.

Table 1: The Number of Confirmed, Deaths and Recovered cases across the Globe

	ObservationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	01/22/2020	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	01/22/2020	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	01/22/2020	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	01/22/2020	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	01/22/2020	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...
172475	12/06/2020	Ukraine	2020-12-07 05:26:14	36539.0	337.0	6556.0
172476	12/06/2020	Netherlands	2020-12-07 05:26:14	6710.0	104.0	0.0
172477	12/06/2020	Mainland China	2020-12-07 05:26:14	1295.0	1.0	1288.0
172478	12/06/2020	Ukraine	2020-12-07 05:26:14	31967.0	531.0	22263.0
172479	12/06/2020	Netherlands	2020-12-07 05:26:14	154813.0	2414.0	0.0

172480 rows × 6 columns

2. The number of cases is analyzed over time.

2.1 The trend for the number of cases per month.

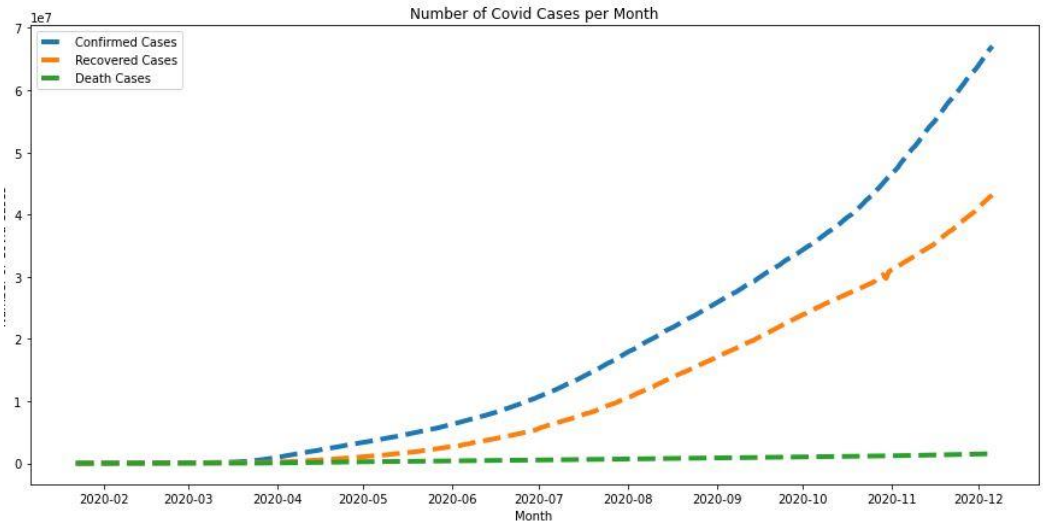


Figure 1: The number of COVID-19 Cases per Month

The trend for the number of cases across the globe per month are still rapidly increasing. However, the number of deaths caused by the COVID-19 pandemic are approximately constant throughout the year. Even though the number of confirmed cases is increasing, the number of covered cases is also increasing.

3. Number of Active Cases

The number of active cases is obtained by subtracting the number of recovered and death cases from the confirmed cases.

Number of Active Cases = Number of Confirmed Cases – (Number of Deaths Cases + Number of Recovered Cases)

Table 2: The Number of COVID-19 Active Cases

ObservationDate	Confirmed	Recovered	Deaths	Active_Cases
2020-01-22	555.0	28.0	17.0	510.0
2020-01-23	653.0	30.0	18.0	605.0
2020-01-24	941.0	36.0	26.0	879.0
2020-01-25	1438.0	39.0	42.0	1357.0
2020-01-26	2118.0	52.0	56.0	2010.0
...	...	...	...	...
2020-12-02	64530517.0	41496318.0	1493742.0	21540457.0
2020-12-03	65221040.0	41932091.0	1506260.0	21782689.0
2020-12-04	65899441.0	42352021.0	1518670.0	22028750.0
2020-12-05	66540034.0	42789879.0	1528868.0	22221287.0
2020-12-06	67073728.0	43103827.0	1536056.0	22433845.0

320 rows × 4 columns

4. Recovery and Mortality Rate per Month

The mortality rate is calculated by dividing the death cases over the confirmed cases. The result of the calculation is then multiplied by 100.

Average Mortality Rate = (Death Cases/Confirmed Cases) * 100

The recovery rate is calculated by dividing the recovered cases over the confirmed cases. The result of the calculation is then multiplied by 100.

Average Recovery Rate = (Recovered Cases/Confirmed Cases) * 100

The mortality rate and the recovery rate are plotted:

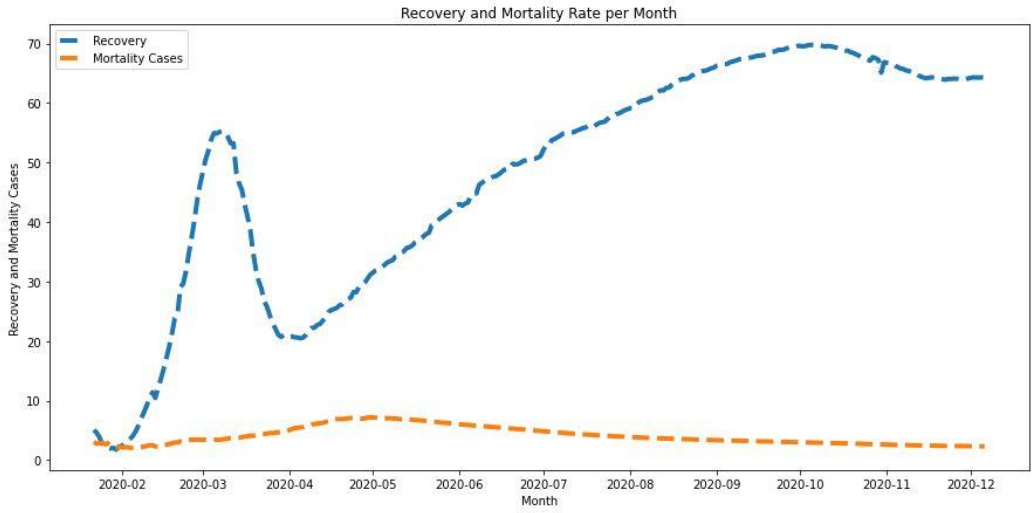


Figure 2: Recovery and Mortality Rate per Month

From the result it is proven that the recovery rate is rising, and the mortality rate is decreasing in a slow manner. This shows that the actions taken, and the awareness developed in the society has impact to the recovery rate of COVID-19 cases. There is a possibility for the transmission to be eliminated if the actions such as social distancing are practiced in a steadfast manner.

5. The number of COVID-19 Cases per Country

5.1 Top 10 Countries with High Number of COVID-19 cases in the last 24 Hours.

Table 3: The Top 10 Countries with High Number of COVID-19 Cases in the Last 24 Hours

	Country Name	Last 24 Hours Confirmed	Last 24 Hours Recovered	Last 24 Hours Deaths
0	US	175663.0	48418.0	1113.0
1	India	32981.0	39109.0	391.0
2	Brazil	26363.0	1819.0	313.0
3	Russia	28701.0	21270.0	447.0
4	France	11022.0	228.0	174.0
5	Italy	18887.0	17186.0	564.0
6	UK	17372.0	86.0	231.0
7	Spain	0.0	0.0	0.0
8	Argentina	3278.0	5907.0	138.0
9	Colombia	8854.0	7708.0	175.0

It shows that US has the highest number of confirmed, recovered and death cases in the last 24 hours. However, the number of cases in Spain are listed in the top 10 even though the results are zero. This is probability because the number of cases in Spain has not yet updated and it was listed in the top 10 due to its previous cases.

5.2 Top 30 Countries with the Highest Number of Confirmed Cases in the last 24 Hours.

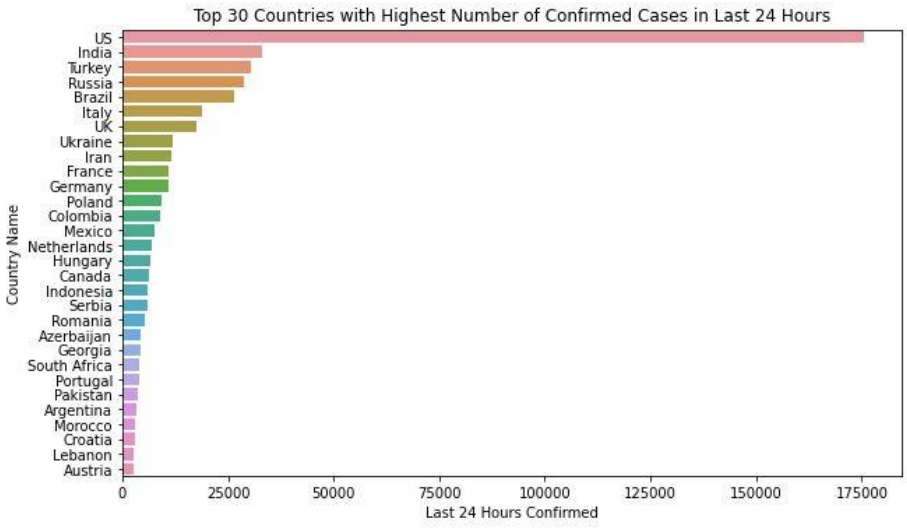


Figure 3: Top 30 Countries with Highest Number of Confirmed Cases in Last 24 Hours

Spain was not shown in the horizontal bar graph. It is shown that US confirmed cases almost achieve to 175000 cases in 24 hours. US has the most rapid growing cases and it has a large difference comparing to other countries. Lebanon and Austria have almost the same number of confirmed cases in the last 24 hours.

5.3 Top 30 Countries with the Highest Number of Recovered Cases in the Last 24 Hours

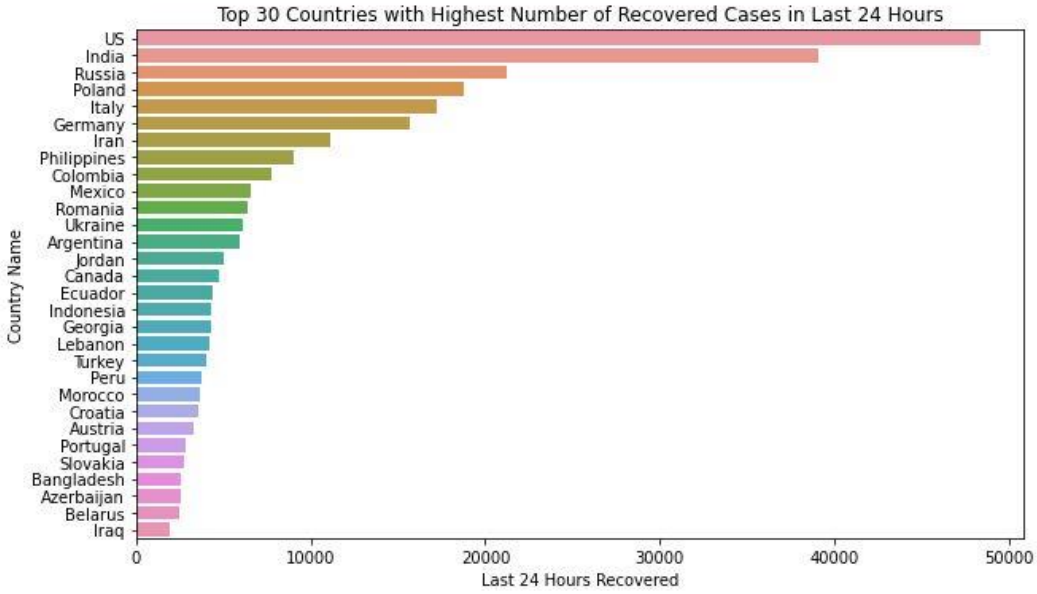


Figure 4: The Top 30 Countries with Highest Number of Recovered Cases in Last 24 Hours

From the result is shows that US is also the top in having the highest number of recovered cases. However, Iraq, Belarus, Azerbaijan, Slovakia, and some of the other countries show high recovered cases without being listed in the top countries that have confirmed cases in the last 24 hours. These countries probably are having good progresses in handling the COVID-19 cases in their countries recently.

5.4 Top 30 Countries with the Highest Number of Death Cases in Last 24 Hours

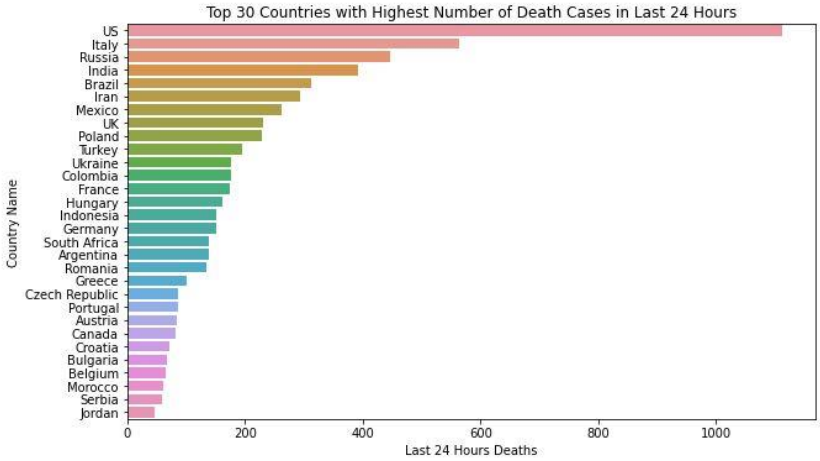


Figure 5: Top 30 Countries with the Highest Number of Death Cases in Last 24 Hours

Jordan was probability having some difficulties in handling the COVID-19 Cases in the last 24 hours. The number of death cases has made the country to be in the Top 30 with the Highest Number of Death Cases in the last 24 Hours. However, it shows in the previous graph that the recovery cases in Jordan was around 4000 cases in the last 24 hours. The recovery cases are higher than the number of death cases, which is less than 200 cases.

6. The Total Number of Cases per Country.

Table 4: The Total Number of Cases per Country

Country/Region	Confirmed	Recovered	Deaths	Mortality_Cases	Recovery_Cases
US	14757000.0	5624444.0	282299.0	1.912984	38.113736
India	9677203.0	9139901.0	140573.0	1.452620	94.447755
Brazil	6603540.0	5866657.0	176941.0	2.679487	88.841091
Russia	2439163.0	1920744.0	42675.0	1.749576	78.746029
France	2345648.0	175220.0	55247.0	2.355298	7.470004
Italy	1728878.0	913494.0	60078.0	3.474970	52.837389
UK	1727751.0	3736.0	61342.0	3.550396	0.216235
Spain	1684647.0	150376.0	46252.0	2.745501	8.926262
Argentina	1463110.0	1294692.0	39770.0	2.718183	88.489040
Colombia	1371103.0	1257410.0	37808.0	2.757488	91.707917

Spain was listed as one the highest number of cases recorded for COVID-19 cases. US is still placed as the first country with the highest total number of cases in 2020.

7. Modelling using Machine Learning Algorithms

7.1 Linear Regression

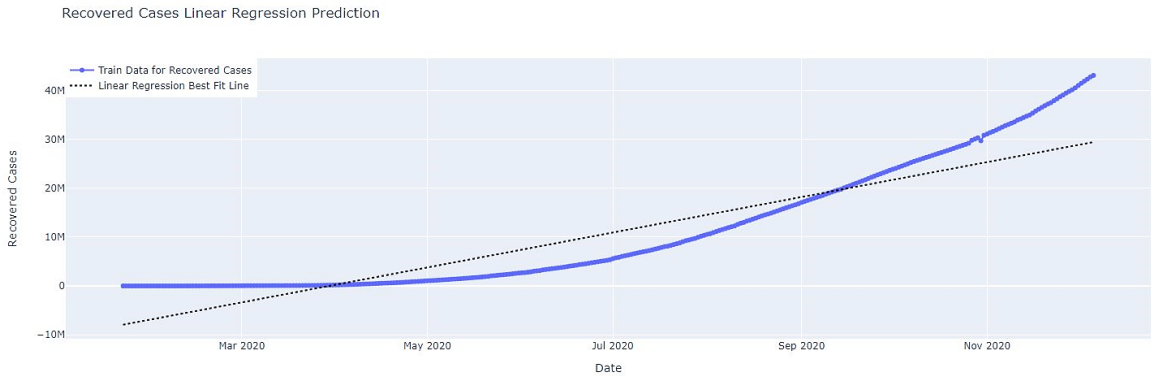


Figure 5: The Prediction of Recovered Cases obtained from Linear Regression Model

The graph shows that the result of the recovered cases is not linear. Linear regression algorithm probably provides a less accurate result in predicting the number of recovered cases from the COVID-19 pandemic.

7.2 Polynomial Regression

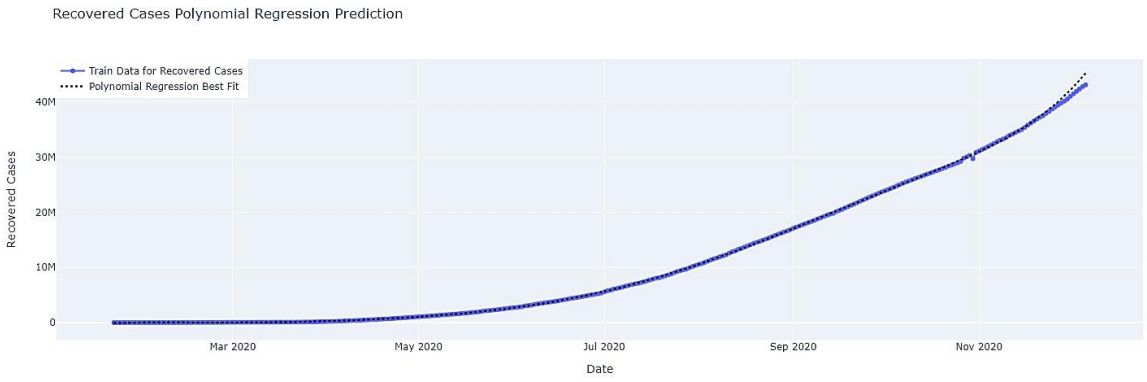


Figure 6: The Forecasting obtained from the Polynomial Regression Model

The polynomial regression shows a better result than linear regression.

7.3 Time Series Forecasting: Holt's Linear Model

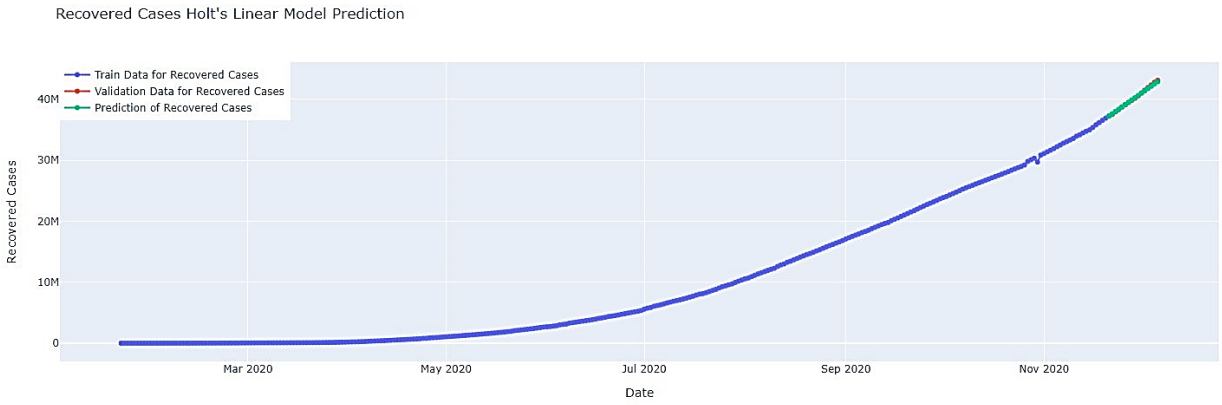


Figure 7: Prediction of Number of Recovered Cases using Holt's Linear Model

7.4 Time Series Forecasting: Holt's Winter Model

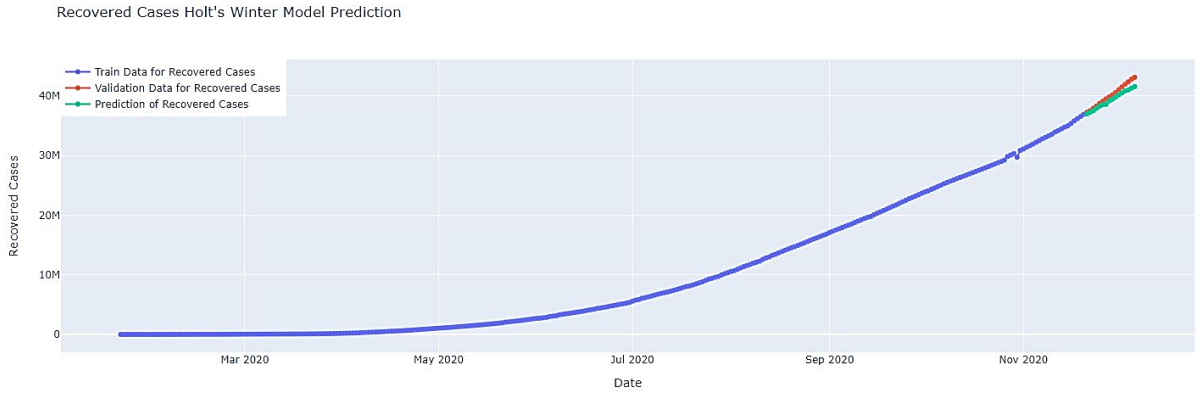


Figure 8: Predicted of the Number of Recovered Cases using Holt's Winter Model

7.5 Time Series Forecasting: ARIMA Model

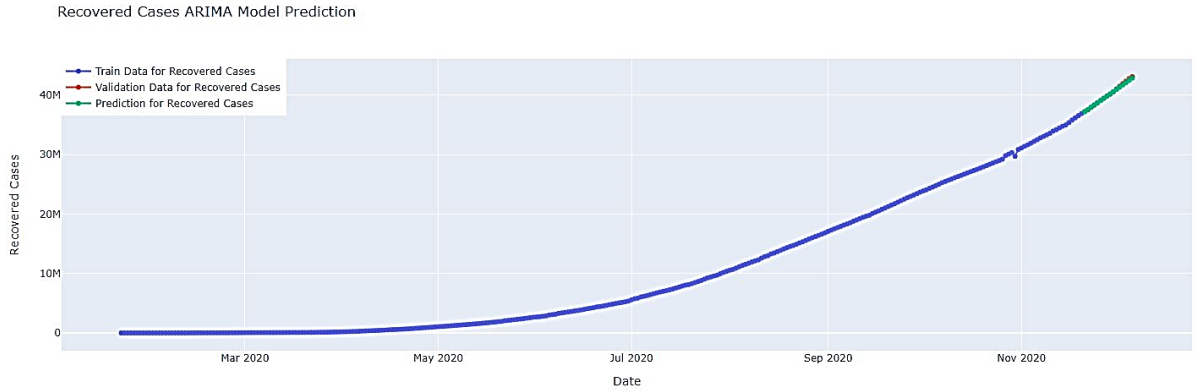


Figure 9: The Prediction of Number of Recovered Cases using ARIMA Model

8. Machine Learning Algorithms Comparison based on the number of Recovered Cases predicted.

Table 5: Comparison between the Machine Learning Algorithms used

	Polynonmial Regression Prediction	Holt's Linear Model Prediction	Holt's Winter Model Prediction	ARIMA Model Prediction
0	45925646.659647	43256913.210901	41822016.738689	43246407.284313
1	46628850.922760	43632855.494483	42231587.787107	43632974.680695
2	47357227.707281	44008797.778065	42614762.311094	44020758.304184
3	48111769.050104	44384740.061647	42870364.428532	44409686.599797
4	48893494.848253	44760682.345229	42873441.721326	44799797.483467

The number of recovered cases is predicted on each day. It shows that there is not much difference in the number of prediction when Holt's Linear Model and Arima Model are used.

However, there is a huge difference when it comes between the Polynomial Regression and Holt's Winter Model algorithms. The number of recovered cases is around 4 000 000 cases in difference.

9. Evaluation Metrics using Root Mean Squared Error

Root Mean Squared Error (RMSE) is a standard method used to measure the error of a model in quantitative analysis.

i. `Root Mean Squared Error for Polynomial Regression: 1090114.218156747`

ii. `Root Mean Square Error Holt's Linear Model: 1170724.9761738346`

iii. `Root Mean Square Error for Holt's Winter Model: 1138567.8326810487`

iv. `Root Mean Square Error for ARIMA Model: 1196612.5819942027`

Arima Model has the least root mean square error.

DISCUSSION

The strength of this project is that it analyzed the whole data including the list of countries around the world throughout the year, from January 2020 to December 2020. The result or the patterns obtained are based on the data collection that has been done. The data is collected by the Centre for Systems Science and Engineering (CSSE) at John Hopkins University that updates the COVID-19 record from time to time. Besides that, the project is approached by conducting quantitative research that involves a large sample of data that represents the world population. The result from the project can concludes the number of cases of COVID-19 around the world.

The weakness of this project is that it does not go deeply than identifying the number of cases across the countries. The data has some large missing values when it comes to the province/state feature. This does not allow the study to go deeply into counting the cases that occur in each state/province. It is advisable to search for a better dataset with a more accurate record number of COVID-19 cases.

CONCLUSION

The insights or patterns that are discovered from the study shows that the number of recovered cases is also growing even though the number of confirmed cases soaring from time to time. It is proven that the actions or steps such as social distancing, avoiding contact, avoiding gathering in a crowded place, washing hands, regular temperature check and many more assist in reducing the chances for the disease to be transmitted to another human.

These actions or steps should be a normal habit in our daily life. There are still no signs that the number of cases would soon decrease to zero, but the awareness among the society must lever up so that everyone feels responsible in fighting the pandemic together. A lot research has been conducted to create a vaccine for the disease. There is a high hope that the number of recovered cases could rocket up soon if the creation of the ideal vaccine succeeds.

REFERENCES

Aina Umaira Md Shah, S. N. (2020). Covid-19 outbreak in Malaysia: Actions taken by the Malaysian Government. *International Journal of Infectious Diseases 97*.

Almasarweh, M., & Wadi, A. (2018). ARIMA Model in Predicting Banking Stock Market Data.

Anthony S. Fauci, M. D. (2020). Covid-19- Navigating the Uncharted.

D, T., s, S., Vespe, M., LAcus, S., Santamaria, C., & Sermi, F. (2020). *How Human Mobility explains the Initial Spread of COVID-19.* EU Publications.

Fairoza Amira binti Hamzah, C. H. (March 2020). CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction.

Ostertagova, E. (2012). Modelling using Polynomial Regression.

Yapar, G., Capar, S., Selamlar, H. T., & Yavuz, I. (2017). Modified Holt's Linear Trend Method.

Yasuhiro Kubota, T. S. (2020). Multiple drives of the Covid-19 Spread: The roles of Climate, International Mobility and Region-specific Conditions.

Zare-Zardini, H., Soltaninejad, H., Ferdosian, F., Hamidieh, A. A., & Memarpoor-Yazdi, M. (2020). Coronavirus Disease 2019 (COVID-19) in Children: Prevalence, Diagnosis, Clinical Symptoms and Treatment. *International Journal of General Medicine*.