THE PREDICTION OF CASE'S RECOVERY RATE CORONA VIRUS 2019 (COVID-19) WORLDWIDE

NURUL SYAFIQAH BINTI MD KHAIRI

FACULTY OF TECHNOLOGY AND COMPUTER SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

THE PREDICTION OF CASES OF RECOVERY RATE CORONA VIRUS 2019 (COVID-19) WORLDWIDE.

NURUL SYAFIQAH BINTI MD KHAIRI

FACULTY OF TECHNOLOGY AND COMPUTER SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Nurul Syafiqah Binti Md Khairi

I.C/Passport No: 940208 – 14 – 6034

Matric No: WQD 180090 / 17199335

Name of Degree: Master of Data Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"): The prediction of cases Recovery rate COVID-19 worldwide.

Field of Study: Machine Learning

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 18 January 2021

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: Nurul Syafiqah Binti Md Khairi

No. K.P/Pasport: 940208 – 14 – 6034

No. Matrik: WQD 180090 / 17199335

Nama Ijazah: Master of Data Science

Tajuk Kertas Projek/Laporan Penyelidikan/Disertasi/Tesis ("Hasil Kerja ini"):

The prediction of cases recovery rate COVID-19 worldwide.

Bidang Penyelidikan: Machine Learning

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon Tarikh:

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh:

Nama:

Jawatan:

THE PREDICTION OF CASES BASED ON AVERAGE MOVING DAYS OF COVID-19 WORLWIDE

ABSTRACT

.

Around 96,000 confirmed coronavirus disease cases have been reported in 2019 (COVID-2019) and 3300 reported deaths to date of 05/03/2020. Inhalation or contact with infected droplets transmits the disease and the incubation duration varies from two (2) to fourteen (14) days. This research then aimed to predict the Recovery rate and fatality rate from the confirm, deaths, recovery and active results by using Linear Regression, Random Forest and Decision Tree model to evaluate the prediction of the result. The Clustering method also be applied by clustering the cases based on the high, middle and low cases within days.

ACKNOWLEDGEMENTS

I would like to express my very great appreciation to Dr Woo Chaw Seng for his valuable and constructive teaching during the pandemic when the class need to be handle online. The struggle to make the class for the entire semester to be great and the knowledge received by student perfectly. His passion to conduct the class and the project for the course was really appreciated.

TABLE OF CONTENTS

CHAPTER 2: INTRODUCTION

The discovery and spread of the novel coronavirus virus in 2019 challenge the planet with a new public health crisis (2019-nCoV) or also known as Extreme Acute Respiratory Syndrome (SARS-CoV-2). The virus originated in bats and was transmitted to humans in Wuhan, Hubei province, China in December 2019 through still unknown intermediate animals. Around 96,000 confirmed coronavirus disease cases have been reported in 2019 (COVID-2019) and 3300 reported deaths to date of 05/03/2020. Inhalation or contact with infected droplets transmits the disease and the incubation duration varies from two (2) to fourteen (14) days. Usually, the symptoms are fever, cough, sore throat, breathlessness, weakness, malaise and more. In most individuals, the disease is mild; pneumonia, acute respiratory distress syndrome (ARDS) and multi-organ dysfunction can progress in some, usually the elderly and those with comorbidities. The case fatality rate is estimated to range between 2% and 3%.

The COVID-19 case description is based on symptoms irrespective of travel history or interaction with reported cases. In patients with a recent, continuous cough, fever or loss or a changed sense of normal smell or taste, diagnosis is suspected (anosmia). A diagnostic test has been developed, and suspected cases are quarantined by countries. A critical source of information and expertise has been generated by this sudden burst of cases and their health data. Using various data storage systems, there is an immediate need to store such a vast volume of data in these situations. These data are used for the purpose of research and development on the virus, the pandemic, and initiatives to contain the virus and its aftermath.

2.1 Problem Statement

New cases of COVID-19 (Coronavirus) are rising rapidly at staggering rates worldwide; more than 1.2 billion people have acquired an infection and about 65,000 have died of the disease to date. Every day, new coronavirus outbreaks are announced, related statistics fluctuate quickly, and huge amounts of data are produced at great speed, becoming a challenge for academics and professionals to convert this information into useful information. People around the world are still unaware with the growth rate of the

new cases every day and take it slightly due to lack of understanding the given information such as the causes and future forecasting regarding on result collected.

2.2 Background of Study

As the latest coronavirus (COVID-19) keeps the earth in limbo, it has become an urgent, significant challenge for mankind to end it as soon as possible. Every day, new coronavirus outbreaks are announced, related statistics fluctuate quickly, and huge amounts of data are produced. Big data may potentially help us understand the existence of the new coronavirus, offering a source of preventive and treatment-inspiring knowledge. Big data analysis may also promote the prediction of systemic changes that may arise from the current pandemic in our culture, economy, and lifestyle.

The dataset found in open website, Kaggle.com are provided with different datasets provide Global confirm cases, Global death cases and Global Recover cases. All datasets contains the features such as the list of country with longitude and latitude and the list of dates for every cases until 9th January 2021. The all 3 collection of datasets have 272 of rows and 358 columns.

2.2.1 Objectives

The objectives are listed below to achieve the goal of this project:

1. To improve understanding of COVID19 information through processing and visualization of massive and structured data.
2. To trace the trends and growth rate related to deaths and recovery cases expected during pandemic COVID19 around the world.
3. To predict the growth rate within the time related to data from confirm, deaths, recovery and active cases around the world.

CHAPTER 3: METHDOLOGY

In recent years, with promising results, predictive medical analysis using machine learning techniques has experienced enormous growth. In numerous types of applications in various fields, machine learning algorithms are effectively applied. The study was carried out in several phases.

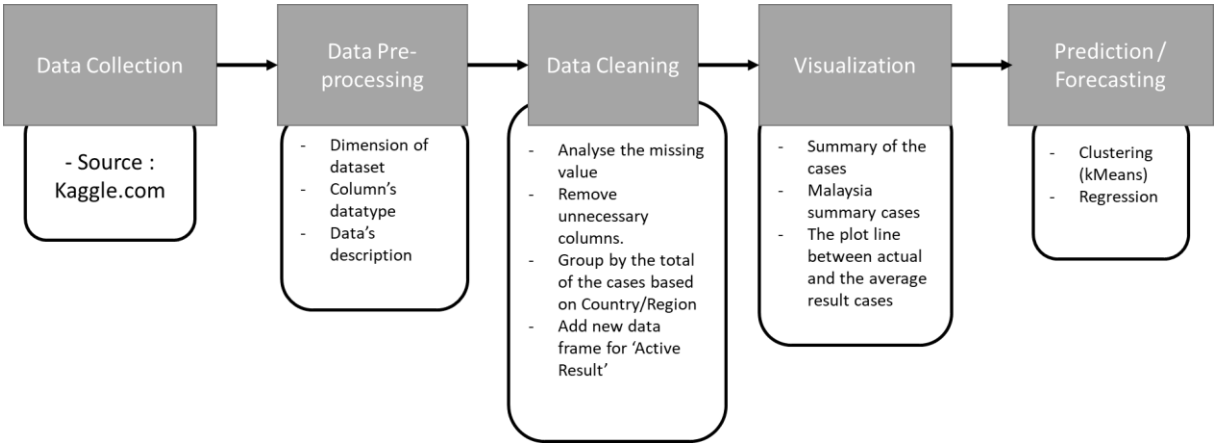


Figure 2.0: *Flowchart of the basic data analysis methodology*

3.1 Data Collection

For this project, the COVID-19 dataset from the Kaggle is taken for predictive analysis. There are 3 different datasets used in this project named as Global Confirmed Cases, Global Deaths Cases and Global Recovery Cases. All the datasets have the same dimension which is 272 of rows and 358 of columns before cleaning process. The features provided in the dataset is the province, country with its longitude and latitude and list of dates up until 9th January 2021.

```
confirmed_data = read_csv("time_series_covid19_confirmed_global.csv")
deaths_data = read_csv("time_series_covid19_deaths_global.csv")
recovered_data = read_csv("time_series_covid19_recovered_global.csv")
#latest_cases_data = read_csv("latest.csv")
```

Figure 2.1.1: *query in python reading the datasets (csv file)*

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/31/20	1/1/21	1/2/21	1/3/21	1/
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	51526	51526	51526	51526	
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	58316	58316	58991	59438	
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	99610	99897	100159	100408	1
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	8049	8117	8166	8192	
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	17553	17568	17608	17642	
5	NaN	Antigua and Barbuda	17.060800	-61.796400	0	0	0	0	0	0	...	159	159	159	160	
6	NaN	Argentina	-38.416100	-63.616700	0	0	0	0	0	0	...	1625514	1629594	1634834	1640718	16
7	NaN	Armenia	40.069100	45.038200	0	0	0	0	0	0	...	159409	159738	159798	160027	1
8	Australian Capital Territory	Australia	-35.473500	149.012400	0	0	0	0	0	0	...	118	118	118	118	
9	New South Wales	Australia	-33.868800	151.209300	0	0	0	0	3	4	...	4928	4947	4958	4965	
10	Northern Territory	Australia	-12.463400	130.845600	0	0	0	0	0	0	...	75	75	81	81	
11	Queensland	Australia	-27.469800	153.025100	0	0	0	0	0	0	...	1253	1255	1255	1260	
12	South Australia	Australia	-34.928500	138.600700	0	0	0	0	0	0	...	580	580	580	583	
13	Tasmania	Australia	-42.882100	147.327200	0	0	0	0	0	0	...	234	234	234	234	
14	Victoria	Australia	-37.813600	144.963100	0	0	0	0	1	1	...	20376	20388	20391	20395	
15	Western Australia	Australia	-31.950500	115.860500	0	0	0	0	0	0	...	861	863	867	868	

Figure 2.1.2: *A glance of Global Confirmed Cases Dataset*

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/31/20	1/1/21	1/2/21	1/3/21	1/4/21
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	2191	2191	2191	2191	2237
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	1181	1181	1190	1193	1199
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	2756	2762	2769	2772	2777
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	84	84	84	84	84
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	405	405	407	408	408
5	NaN	Antigua and Barbuda	17.060800	-61.796400	0	0	0	0	0	0	...	5	5	5	5	5
6	NaN	Argentina	-38.416100	-63.616700	0	0	0	0	0	0	...	43245	43319	43375	43482	43634
7	NaN	Armenia	40.069100	45.038200	0	0	0	0	0	0	...	2823	2828	2836	2850	2864
8	Australian Capital Territory	Australia	-35.473500	149.012400	0	0	0	0	0	0	...	3	3	3	3	3
9	New South Wales	Australia	-33.868800	151.209300	0	0	0	0	0	0	...	54	54	54	54	54
10	Northern Territory	Australia	-12.463400	130.845600	0	0	0	0	0	0	...	0	0	0	0	0
11	Queensland	Australia	-27.469800	153.025100	0	0	0	0	0	0	...	6	6	6	6	6
12	South Australia	Australia	-34.928500	138.600700	0	0	0	0	0	0	...	4	4	4	4	4
13	Tasmania	Australia	-42.882100	147.327200	0	0	0	0	0	0	...	13	13	13	13	13
14	Victoria	Australia	-37.813600	144.963100	0	0	0	0	0	0	...	820	820	820	820	820
15	Western Australia	Australia	-31.950500	115.860500	0	0	0	0	0	0	...	9	9	9	9	9

Figure 2.1.3: *A glance of Global Deaths Cases Dataset*

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/31/20	1/1/21	1/2/21	1/3/21	1/4/21
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	41727	41727	41727	41727	425
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	33634	33634	34353	34648	349
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	67127	67395	67611	67808	679
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	7432	7463	7463	7517	75
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	11044	11146	11189	11223	112
5	NaN	Antigua and Barbuda	17.060800	-61.796400	0	0	0	0	0	0	...	148	148	148	148	1
6	NaN	Argentina	-38.416100	-63.616700	0	0	0	0	0	0	...	1426676	1426676	1447092	1452960	14580
7	NaN	Armenia	40.069100	45.038200	0	0	0	0	0	0	...	142801	143355	143640	144091	1448
8	Australian Capital Territory	Australia	-35.473500	149.012400	0	0	0	0	0	0	...	114	114	114	114	1
9	New South Wales	Australia	-33.868800	151.209300	0	0	0	0	0	0	...	3197	3197	3197	3197	31
10	Northern Territory	Australia	-12.463400	130.845600	0	0	0	0	0	0	...	71	71	71	70	
11	Queensland	Australia	-27.469800	153.025100	0	0	0	0	0	0	...	1218	1224	1232	1232	12
12	South Australia	Australia	-34.928500	138.600700	0	0	0	0	0	0	...	565	566	566	566	5
13	Tasmania	Australia	-42.882100	147.327200	0	0	0	0	0	0	...	221	221	221	221	2
14	Victoria	Australia	-37.813600	144.963100	0	0	0	0	0	0	...	19538	19539	19539	19539	195
15	Western Australia	Australia	-31.950500	115.860500	0	0	0	0	0	0	...	838	838	838	839	8

Figure 2.1.4: *A glance of Global Recovered Cases Dataset*

3.2 Data Pre-processing

Using data preprocessing and data cleaning methodologies, the considered dataset was cleaned, then the resulting dataset was considered for multiple experiments using various classification algorithms. The datasets contain the information of the cases within the dates. With a glance through datasets, the datasets have the dimension of 272 of rows and 358 of columns before cleaning. This pre-processing part is the way to determine the data type of every columns, the description summary of the datasets, came out with the unique keys of every strings (object) columns and determine the correlation among the features of the datasets.

```
corr_matrix = confirmed_data.corr()
```

```
print(corr_matrix)
```

```
              Lat      Long   1/22/20   1/23/20   1/24/20   1/25/20   1/26/20  \
Lat      1.000000 -0.116701  0.030553  0.035551  0.040513  0.045399  0.048170
Long    -0.116701  1.000000  0.092918  0.109523  0.126105  0.138323  0.146237
1/22/20  0.030553  0.092918  1.000000  0.998413  0.994030  0.990475  0.986139
1/23/20  0.035551  0.109523  0.998413  1.000000  0.997986  0.995347  0.992209
1/24/20  0.040513  0.126105  0.994030  0.997986  1.000000  0.998392  0.996448
...           ...       ...       ...       ...       ...       ...       ...
1/5/21   0.065263 -0.100064 -0.010974 -0.013685 -0.015036 -0.017841 -0.017577
1/6/21   0.065446 -0.100393 -0.010983 -0.013692 -0.015039 -0.017843 -0.017575
1/7/21   0.065583 -0.100782 -0.010990 -0.013697 -0.015042 -0.017847 -0.017571
1/8/21   0.065835 -0.101206 -0.010991 -0.013694 -0.015032 -0.017835 -0.017554
1/9/21   0.065892 -0.101446 -0.010989 -0.013688 -0.015020 -0.017822 -0.017534

          1/27/20   1/28/20   1/29/20  ...  12/31/20    1/1/21    1/2/21  \
Lat      0.049588  0.038710  0.042425  ...  0.063884  0.064185  0.064371
Long     0.150103  0.116398  0.128139  ... -0.099095 -0.099211 -0.099366
1/22/20  0.984274  0.996681  0.992690  ... -0.010981 -0.010995 -0.010961
1/23/20  0.990233  0.997908  0.996338  ... -0.013710 -0.013724 -0.013679
1/24/20  0.994997  0.996874  0.997385  ... -0.015084 -0.015096 -0.015040
...           ...       ...       ...  ...       ...       ...       ...
1/5/21  -0.019242 -0.015396 -0.017065  ...  0.999815  0.999865  0.999939
1/6/21  -0.019241 -0.015403 -0.017071  ...  0.999734  0.999795  0.999888
1/7/21  -0.019238 -0.015409 -0.017074  ...  0.999633  0.999704  0.999816
1/8/21  -0.019223 -0.015407 -0.017068  ...  0.999479  0.999565  0.999703
1/9/21  -0.019204 -0.015400 -0.017058  ...  0.999360  0.999455  0.999612
```

Figure 2.2.1: *A glance correlation matrix of Global Confirmed Cases Dataset*

3.3 Data Cleaning

Data cleaning part is the crucial parts where it is the part to handle the missing value of every columns. Every column is in numerical and easy to compute the missing value with the mean or medium value. The summary of the total missing values was imputed by the percentage. The highest percentage represent the column that contains higher result of missing value.

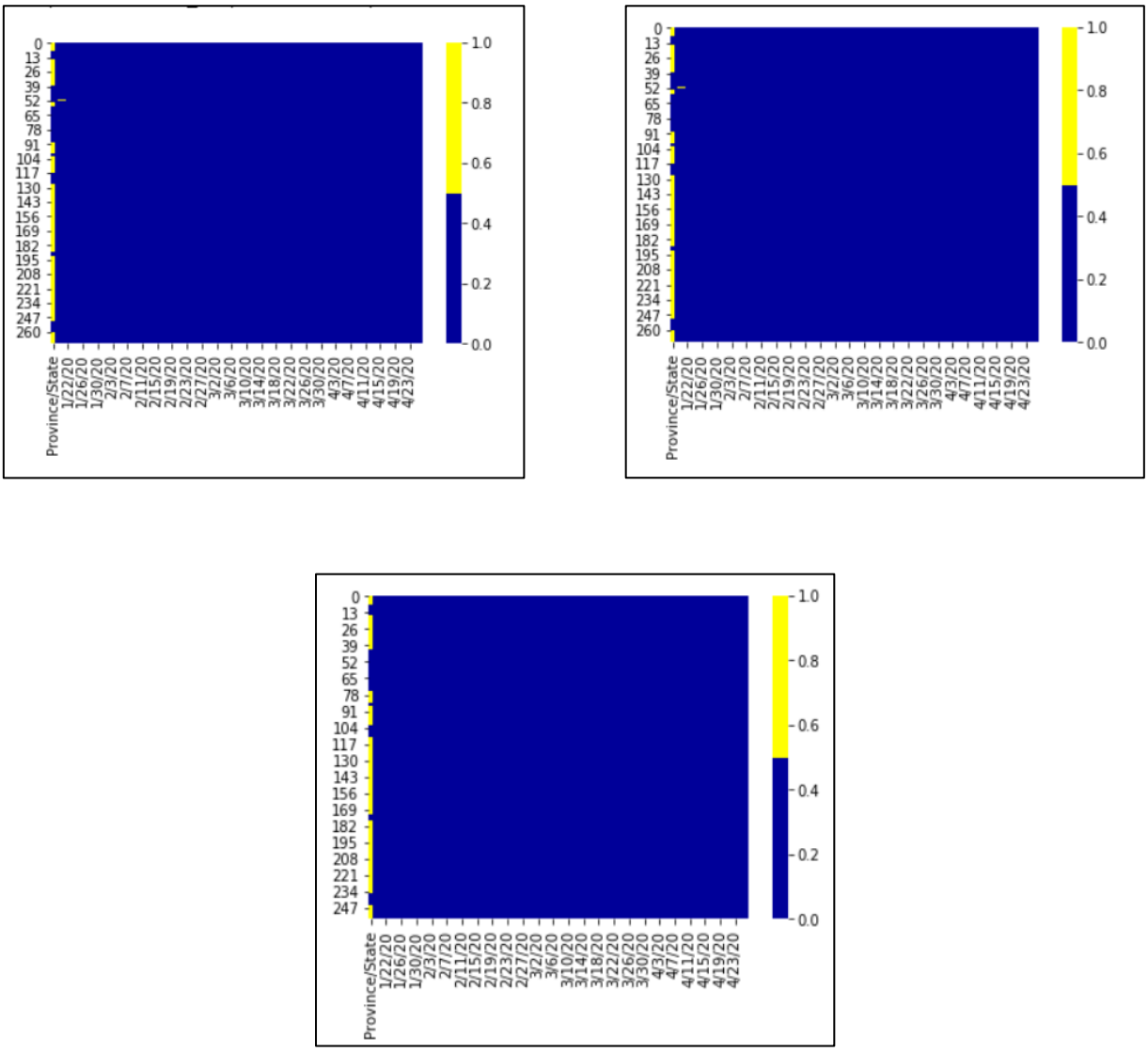


Figure 2.3.1: *summary of total missing values for every dataset*

As shown in figure 2.1.6 above, the columns of 'province/State' of each dataset have the most missing values. For this project, the column is removed. After that, every dataset was group by the country and sum the cases value for every column based on same country. It is much easier to visualize and further analysis. The dimension of the dataset's changes to 191 of rows and 357 of rows.

```
confirmed_data = confirmed_data.groupby(by = 'Country/Region', as_index = False).sum()

deaths_data = deaths_data.groupby(by = 'Country/Region', as_index = False).sum()

recovered_data = recovered_data.groupby(by = 'Country/Region', as_index = False).sum()
```

Figure 2.3.2: *'Group by' python query based on similar country*

After handling the missing value of the datasets, the new data frame was created named ‘Active Results’ by compute using formula below:

Active Result = Confirm Cases - (Deaths Cases + Recover Cases)

The result used the provided datasets that contained the details of the cases within dates. The country, longitude and latitude column remain same.

```
active_result.head(5)
```

	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	...	12/31/20	1/1/21	1/2/21	1/3/21	1/4/21	1/5/21	1/6/21
0	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	0	...	7608	7608	7608	7608	8244	8195	8195
1	Albania	41.15330	20.168300	0	0	0	0	0	0	0	...	23501	23501	23448	23597	23428	23522	23689
2	Algeria	28.03390	1.659600	0	0	0	0	0	0	0	...	29727	29740	29779	29828	29869	29906	29951
3	Andorra	42.50630	1.521800	0	0	0	0	0	0	0	...	533	570	619	591	617	639	649
4	Angola	-11.20270	17.873900	0	0	0	0	0	0	0	...	6104	6017	6012	6011	6010	5970	5974

Figure 2.3.3: *A glance of Global Active Cases Dataset*

Next, with all datasets, the summary of global result was computing, and the new data frame created to place all the values of the result. The data frame was created to easily make a prediction model later.

	index	dates	Confirmed	Deaths	Recovered	Active
0	0	1/22/20	555	17	28	566
1	1	1/23/20	654	18	30	666
2	2	1/24/20	941	26	36	951
3	3	1/25/20	1434	42	39	1431
4	4	1/26/20	2118	56	52	2114
...	...	...	...	...	...	...
349	349	1/5/21	86464101	1868779	48464481	133059803
350	350	1/6/21	87241950	1883761	48777336	134135525
351	351	1/7/21	88104210	1898639	49098418	135303989
352	352	1/8/21	88925739	1913902	49396101	136407938
353	353	1/9/21	89690533	1926624	49729445	137493354

354 rows × 6 columns

Figure 2.3.4: *A glance of Global Cases Dataset*

3.4 Data Visualisation

Visualizations is needed in data processing to view the outcome or information in an interesting and insightful, that will help clients and management to understand what the data is trying to show and what action can be taken with a second glance of the dashboard. sing visual elements in dashboards view, people can easily monitor and view the pattern or even the prediction trends.

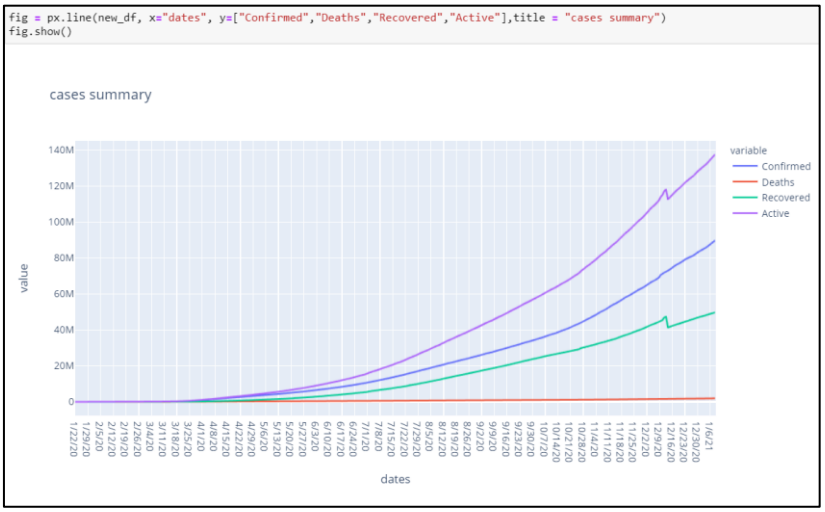


Figure 2.4.1: *Plot of summary case all over the world*

The figure 2.4.1 above shows the plot line of the cases summary Data Frame that compute the sum of the cases for all country for every stated date. The plot line is nearly linear where the cases increasing every day for confirm, deaths and active cases but the deaths result line almost flatten where the deaths reported is small value.

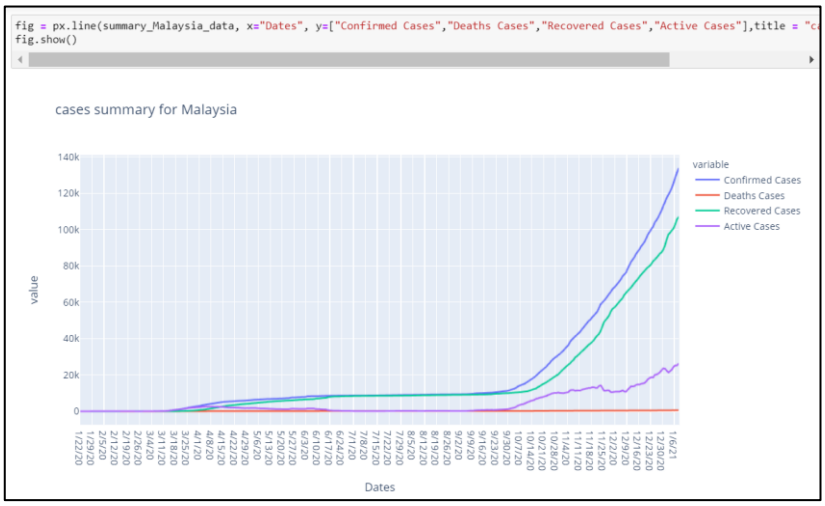


Figure 2.4.2: *Plot of summary case of Malaysia*

Figure 2.4.2 above shows the plot line of Malaysia cases summary. The data of Malaysia's cases from datasets are extract and group into new data frame named as 'summary_Malaysia_data' where it is contains the data of the total confirmed cases, deaths cases and recovered cases. From the graph, the 'confirmed cases' and 'recovered cases' are almost linear as the data increasing within continuous dates. The deaths cases are almost flat because the number of cases reported as death are small value and consistent.

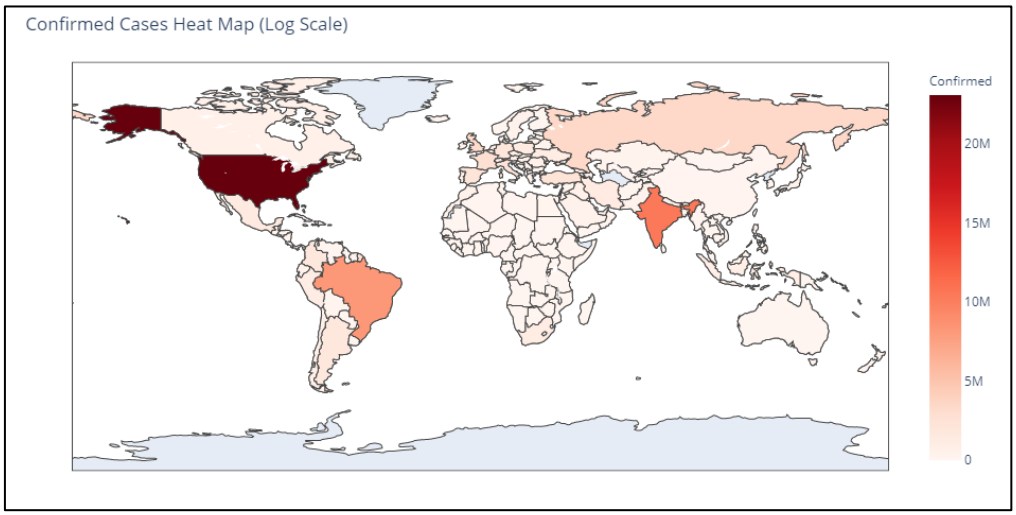


Figure 2.4.3: *Heat Map based on Confirmed Cases*

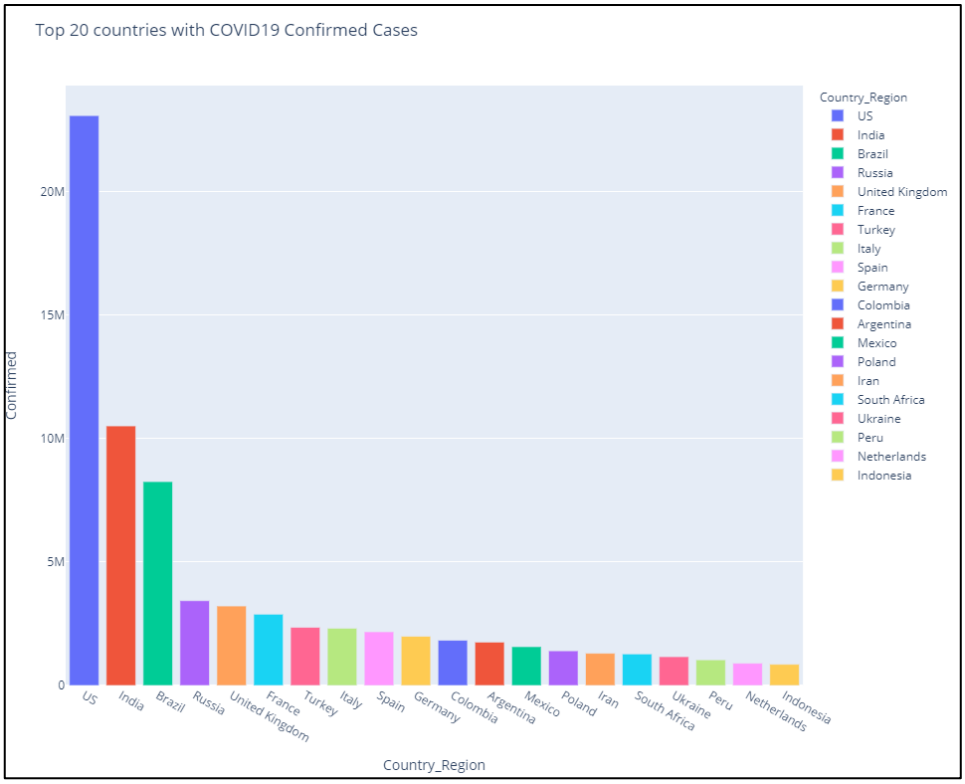


Figure 2.4.4: *Bar Chart based on Top 20 Countries Confirmed cases COVID-19*

Figure 2.4.3 is the heat map based on confirmed cases and figure 2.4.4 is the bar chart of top 20 Countries confirmed cases of COVID-19. Both figures were extract from the date 13th January 2021 of dataset. From heat map figure, the darker color of red was the country had the highest confirm cases and the result shows United States (US) had the most highest cases among countries with total cases more than 23 million.

Figure 2.4.4 shows the bar chart that extract the 20 countries that have the highest number of confirmed cases recorded within COVID 19 epidemic. The result shows US lead the chart among other countries and followed by India that has total cases more than 10 million.

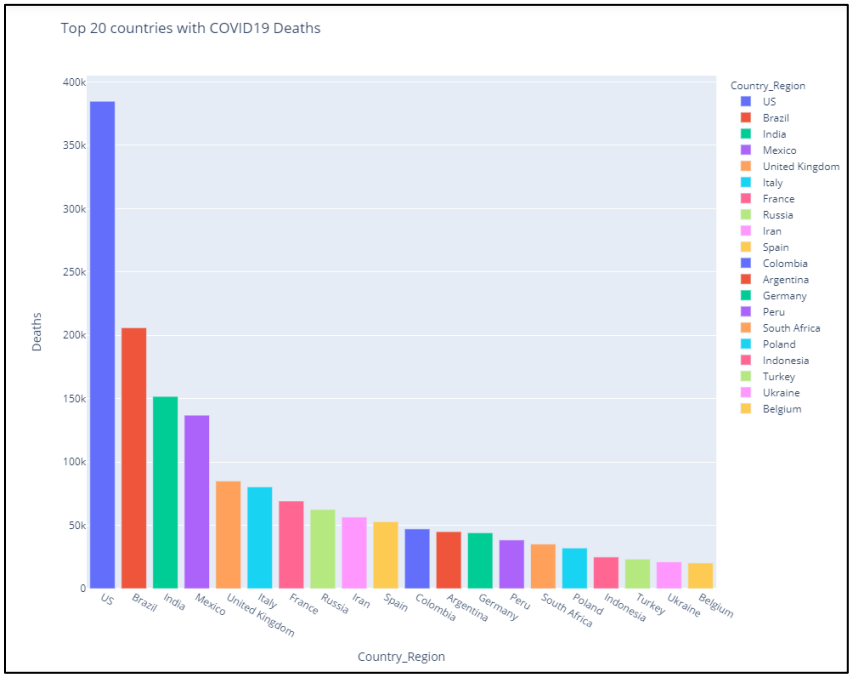
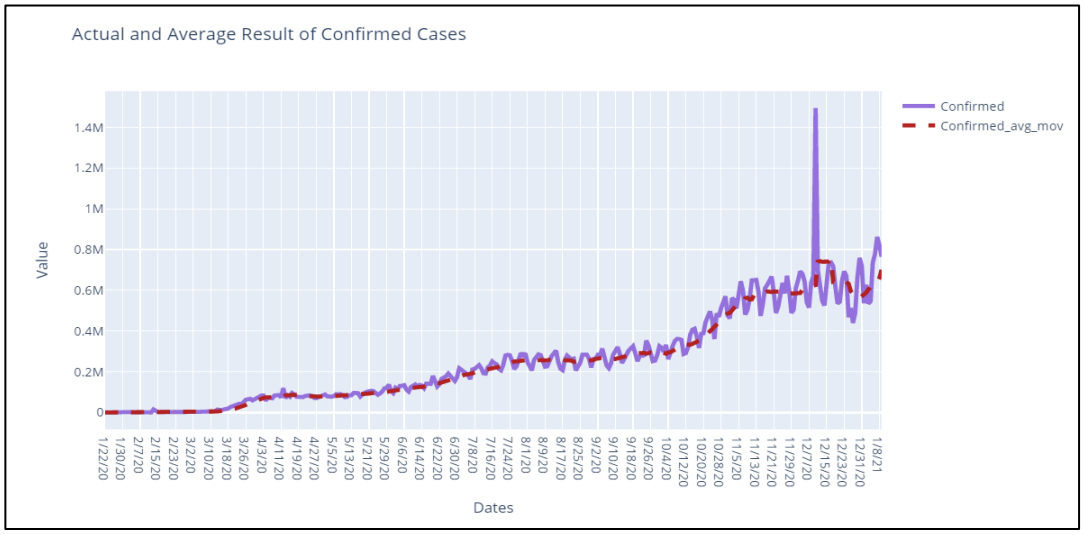


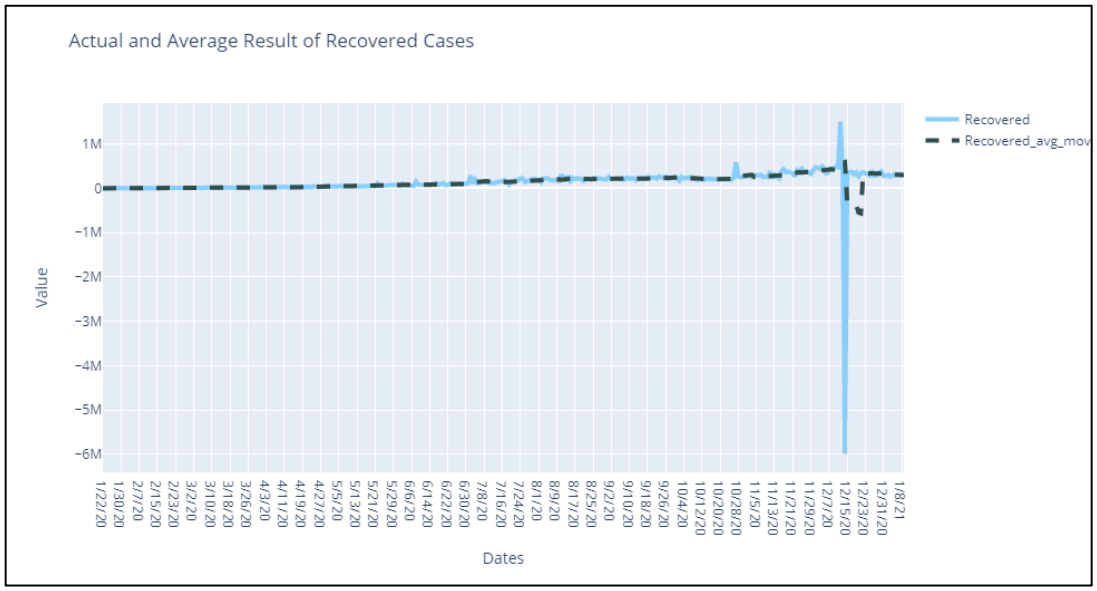
Figure 2.4.5: *Bar Chart based on Top 20 Countries Deaths cases COVID-19*

Figure 2.4.5 shows the bar chart that extract the 20 countries that have the highest number of deaths cases recorded within COVID 19 epidemic. The result shows the same as confirmed cases which US lead the chart with 384,764 cases among other countries and followed by India that has total cases 205, 964 cases recorded every day.

```
dates = new_df['dates']

confirmed_cases = new_df['index'].apply(lambda x : new_df['Confirmed'][x]-new_df['Confirmed'][x-1:x].sum())
avg_moving_cases1 = new_df['index'].apply(lambda x: (new_df['Confirmed'][x-7:x].sum()-new_df['Confirmed'][x-8:x-1].sum())/7 if x

active_cases = new_df['index'].apply(lambda x: new_df['Active'][x]-new_df['Active'][x-1:x].sum())
avg_moving_cases2 = new_df['index'].apply(lambda x: (new_df['Active'][x-7:x].sum()-new_df['Active'][x-8:x-1].sum())/7 if x>0 els

recovered_cases = new_df['index'].apply(lambda x: new_df ['Recovered'][x]-new_df['Recovered'][x-1:x].sum())
avg_moving_cases3 = new_df['index'].apply(lambda x: (new_df['Recovered'][x-7:x].sum()-new_df['Recovered'][x-8:x-1].sum())/7 if x

deaths_cases = new_df['index'].apply(lambda x: new_df['Deaths'][x]-new_df['Deaths'][x-1:x].sum())
avg_moving_cases4 = new_df['index'].apply(lambda x: (new_df['Deaths'][x-7:x].sum()-new_df['Deaths'][x-8:x-1].sum())/7 if x>0 els
```

Figure 2.4.3: *queries of computation the average of moving cases within 7 days*





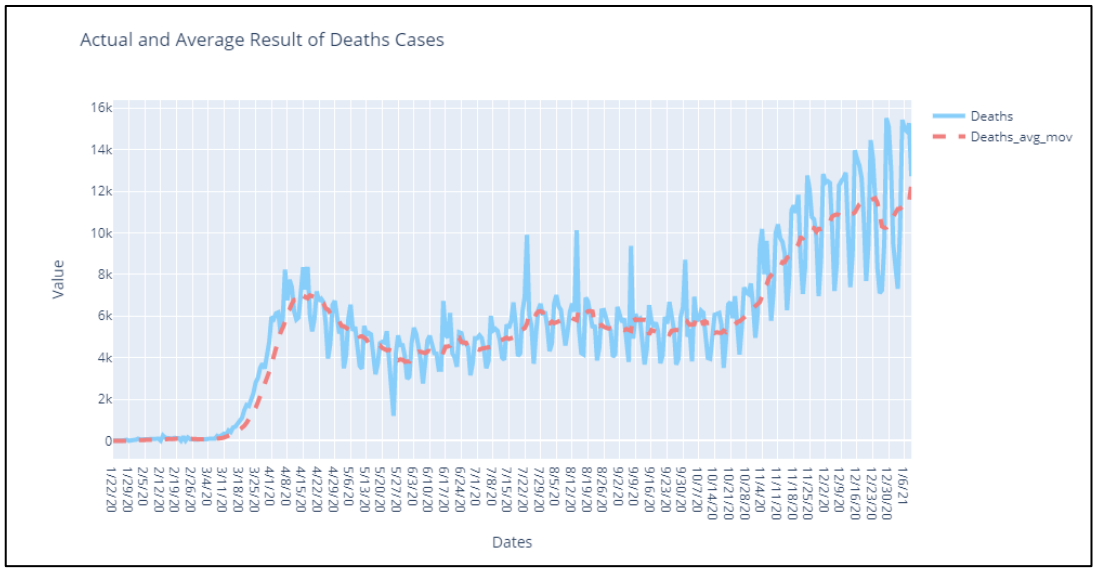


Figure 2.4.4: *Actual and Average moving result of confirmed cases, deaths cases and recovered cases.*

In the graphs above, it is show the relationship of the total cases of confirmed, deaths and recovered with their average moving. A moving average implies that numbers take the previous days, take the average of those days, and graph it on the graph. It takes the last 7 days for a 7-day moving average, adds them up, and splits it by 7. The last 14 days will take an average of 14 days.

So, the datasets have data on COVID starting on 22nd February 2020 for the project planned. 7 days of COVID cases are required for the 7-day moving average. Between 22nd January 2020 with 30th February 2020, it added all the cases together, splitting them by 7. That point is then plotted. The cases started extremely growth is by the end of March where the virus start attack all over the world after its origin happen in Wuhan, China. Over time, it gives viewer an average line, and over a period of time it knocks out these wide peaks and valleys to the average.

CHAPTER 4: RESULT AND DISCUSSION

The considered COVID-19 datasets which are Confirmed Cases, Deaths Cases and Active Cases contains 272 records of cases all over the world. The datasets contain features of cases result such as Country/Region, Longitude, Latitude and dates. The data preprocessing and cleaning process remove the missing and outlier's data values from the dataset. The resulted result dataset after processing is reduce to 191 rows and 357 columns required relevant features of the cases details.

The new data frame was created for prediction result that summarize the total cases of Confirmed, Deaths, Recovered and Active data frame with sum up cases of all country and divided by dates. From the summarize data frame, the new columns which are the percentage of Fatality Rate and Recovery rate are compute based on the calculation of (Recovered/Confirmed) x 100% for Recovery rate and (Deaths/Confirmed) x 100% for Fatality rate. The result of the new data frame is shown below.

	Index	dates	Confirm Result	Deaths Result	Recovery Result	Active Result	Recovered Rate	Fatality Rate
0	0	1/22/20	555	17	28	566	5.045045	3.063063
1	1	1/23/20	654	18	30	666	4.587156	2.752294
2	2	1/24/20	941	26	36	951	3.825717	2.763018
3	3	1/25/20	1434	42	39	1431	2.719665	2.928870
4	4	1/26/20	2118	56	52	2114	2.455146	2.644004
...	...	...	...	...	...	...	...	...
349	349	1/5/21	86464101	1868779	48464481	133059803	56.051564	2.161335
350	350	1/6/21	87241950	1883761	48777336	134135525	55.910415	2.159238
351	351	1/7/21	88104210	1898639	49098418	135303989	55.727664	2.154992
352	352	1/8/21	88925739	1913902	49396101	136407938	55.547586	2.152248
353	353	1/9/21	89690533	1926624	49729445	137493354	55.445590	2.148080

354 rows × 8 columns

Figure 3.0: *New Data Frame of cases with new columns*

4.1 Clustering Method

An unsupervised machine learning task is cluster analysis, or clustering. It requires the discovery of natural grouping in data automatically. Clustering algorithms, unlike supervised learning (like predictive modelling), only interpret the input data and find natural groups or clusters in the feature space. There are many kinds of algorithms to cluster.

The most recognized clustering algorithm is K-Means Clustering, which involves assigning examples to clusters to minimize the variance within each cluster. To perform

the K-Means Clustering, find the values of ‘K’ which the optimum number of cluster by using The Elbow Method and Silhouette value Method.

4.1.1 The Elbow Method

```
X = ml_table[["Fatality Rate","Recovered Rate"]]
X = std.fit_transform(X)
```

```
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X)
    #appending the WCSS to the list (kmeans.inertia_ returns the WCSS value for an initialized cluster)
    wcss.append(kmeans.inertia_)
```

```
plt.plot(range(1, 11), wcss,marker ='o')
plt.title('The Elbow Point Graph')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

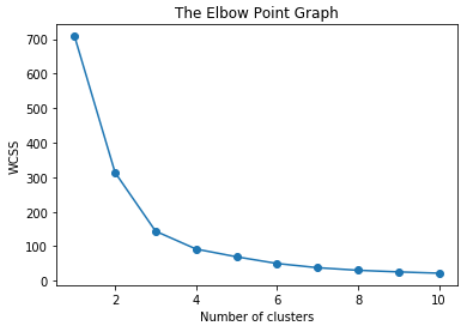


Figure 3.1.1: *The query of elbow method*

The variance (within-cluster sum of squares) decreases as the number of clusters increases. The elbow at 3 or 4 clusters is the most parsimonious balance between minimizing the number of clusters and minimizing the variance within each cluster, the k value can be 3 or 4.

4.1.2 Silhouette value Method

```
cluster_range = range(2,12)
results=[]
for c in cluster_range:
    clusterer = KMeans(init='k-means++', n_clusters = c, n_init=100, random_state = 0)
    cluster_labels = clusterer.fit_predict(X)
    silhouette_avg = silhouette_score(X, cluster_labels)
    results.append([c,silhouette_avg])

result = pd.DataFrame(results, columns = ['n_clusters','silhouette_score'])
pivot_km = pd.pivot_table(result, index = 'n_clusters',values  = 'silhouette_score')

plt.figure()
sns.heatmap(pivot_km, annot=True, linewidths=.5, fmt ='.3f', cmap = sns.cm.rocket_r)
plt.tight_layout()
```

Figure 3.1.2: *The query of Silhouette value Method*

From observation, the optimum number of clusters at n = 3 and the result of the chosen for k values is 3.

4.1.3 Fitting K Means algorithm

```
#Fitting K-Means to the dataset
kmeans = KMeans(n_clusters=3,init='k-means++',random_state=32)
kmeans.fit(X)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=32, tol=0.0001, verbose=0)
```

```
ml_table["Clusters"]= kmeans.predict(X)
```

Figure 3.1.3: *Fitting K Means algorithm for 3 clusters*

Next, all 3 clusters are group by into pivot table with Fatality rate and Recovery rate data and group by with Confirm cases, Deaths Cases and Recovered Cases with other pivot table.

```
table1 = pd.pivot_table(ml_table, values=['Confirm Result', 'Deaths Result','Recovery Result'],
                        index=['Clusters'], aggfunc=np.sum)
table1.style.background_gradient(cmap='Reds').format("{:.2f}")
```

Clusters	Confirm Result	Deaths Result	Recovery Result
0	472916596.00	28148923.00	196412680.00
1	7941489591.00	211995372.00	4980960523.00
2	1376320.00	36399.00	241484.00

```
table2 = pd.pivot_table(ml_table, values=['Recovered Rate','Fatality Rate'],
                        index=['Clusters'], aggfunc=np.mean)
table2.style.background_gradient(cmap='Oranges').format("{:.2f}")
```

Clusters	Fatality Rate	Recovered Rate
0	6.03	35.13
1	3.08	61.61
2	2.53	10.97

Figure 3.1.4: *Pivot tables with the clusters*

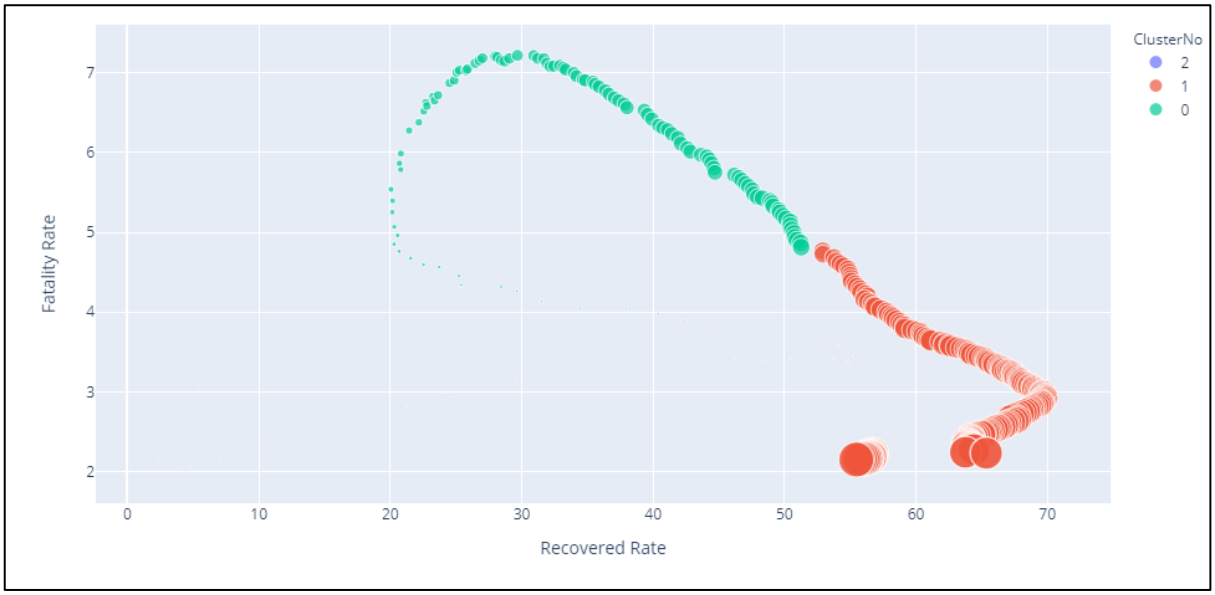


Figure 3.1.5: *Scatter plot of the clusters*

Clusters	Index	Confirm Result	Deaths Result	Recovery Result	Active Result	Recovered Rate	Fatality Rate
0	108.5	4461477.3	265555.9	1852949.8	6048871.3	35.1	6.0
1	236.5	37283988.7	995283.4	23384791.2	59673496.4	61.6	3.1
2	17.0	39323.4	1040.0	6899.5	45183.0	11.0	2.5

Figure 3.1.6: *Interpretation of the cluster summary*

The final analysis from clustering method are :

- **Cluster 0 :** The cases of all countries are around million cases with the average value of Recovered rate and Fatality rate and the second highest cases recorded.
- **Cluster 1 :** The total cases of the world are more than 20 million and the highest cases recorded.
- **Cluster 2 :** The total cases of the world are around thousands and the lowest cases recorded.

4.2 Regression

In this project, the multiple variables or features are used to predict one output. The variables used are result of confirm, deaths, recovery and active. The output that will be predict using the model is Recovery rate.

4.2.1 Multiple Linear Regression, MLR

For the model, the data frame was created before that contains independent variables mark as ‘x’ and dependent variable state as ‘y’. The both was fitted by using the fit function and use the model to make predictions. The result of the prediction was shown in figure below.

```
[22.48117278 22.48100034 22.48094969 22.48104633 22.48047717 22.48062164
 22.47794387 22.476585   22.47443765 22.47377418 22.47230819 22.46906725
 22.46672505 22.46228467 22.45917124 22.45763174 22.4565855  22.45817676
 22.46016815 22.46405227 22.46890735 22.46918883 22.45331022 22.45286226
 22.46246504 22.46879702 22.47550819 22.48642836 22.4988416  22.51256342
 22.51269798 22.53440728 22.53540236 22.55173541 22.56199142 22.56983589
 22.5759208  22.58445375 22.59177754 22.59618835 22.60464372 22.60970102
 22.61793083 22.62408007 22.6276353  22.6309626  22.64623211 22.65549427
 22.67063637 22.68576577 22.69993222 22.70917653 22.72141012 22.75469512
 22.77899363 22.81347474 22.84576767 22.87470429 22.92779317 23.00251231
 23.06949807 23.12674582 23.23758979 23.3594171  23.47443904 23.62605371
 23.79079484 23.95770221 24.1715319  24.41172593 24.74635251 25.07444293
 25.42180175 25.8507525  26.14138481 26.50133998 27.08684735 27.51709142
 28.01790139 28.4842947  28.88809553 29.12095028 29.5094324  29.94951537
 30.55315824 31.00164629 31.55713065 31.92947221 32.25271054 32.61198286
 33.11310482 33.54929174 33.98376945 34.44844735 34.78162717 34.99728825
 35.27819126 35.71962919 36.19181674 36.62851333 36.92458962 37.28254155
 37.45153544 37.67547398 38.05221248 38.48261764 38.79197498 39.10697466
 39.36573909 39.55109046 39.74768696 40.08462161 40.41850769 40.6994748 
```

Figure 4.2.1.1: *Prediction of the data*

The predict function, lm.predict(), predicts the y (dependent variable) using the linear model creation. The score of the model was **0.7974.** The model are compared with other model such as Random Forest and Decision Tree model.

4.2.2 Random Forest

The variables are same as MLR model. The 'RandomForestRegressor' is used to solve regression problems on Random Forest. The important of the RandomForestRegressor is the n_estimators parameter that defines the number of trees in random forest. The n_estimator stated as 100 to see the algorithms performance. The score of the Random Forest is high than linear regression is **0.9999.** The last step is to evaluate the performance of the algorithm. The metrics used to compute the mean absolute error, Mean Square Error and Root Mean Squared Error as shown in figure below.

```
from sklearn import metrics

print('MAE:',metrics.mean_absolute_error(y,y_head_rf))
print('MSE:',metrics.mean_squared_error(y,y_head_rf))
print('RMSE:',np.sqrt(metrics.mean_absolute_error(y,y_head_rf)))
```

```
MAE: 0.08549350367700549
MSE: 0.04575650631538538
RMSE: 0.2923927216553201
```

Figure 4.3.1.1: *Prediction of the data*

4.2.3 Decision Tree

The object was created using DecisionTreeClassifier class and store the address in dt_reg variable so it is easy to be access. The tree was fitted with selected features. The score of the Random Forest is high than linear regression is **0.9999** almost same as random forest.

CHAPTER 5: CONCLUSION

The corresponding model is developed by analyzing the current Hubei epidemic situation data, and then the simulation is carried out. New COVID-19 cases were significantly predicted by the number of days and new cases all over the world, and the number of days and new recoveries significantly predicted new COVID-19 fatalities. According to this analysis, if the cases keep increasing within days, the predicted values of COVID-19 new cases and new deaths will be high as well. Among all the model, the best fitted model to predict the recovery rate and fatal rate was random forest and decision tree with score of 0.9999.

REFERENCES

Singhal, T. (2020). A Review of Coronavirus Disease-2019 (COVID-19). The Indian Journal of Pediatrics, 87(4), 281-286. doi:10.1007/s12098-020-03263-6

Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS CoV-2) case. Transportation Research Part E: Logistics and Transportation Review, 136, 101922. doi:10.1016/j.tre.2020.101922

Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons & Fractals, 139, 110017. doi:10.1016/j.chaos.2020.110017

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. Plos One, 15(3). doi:10.1371/journal.pone.0230405

What is a moving average, and why is it useful? (n.d.). Retrieved from https://www.georgiaruralhealth.org/blog/what-is-a-moving-average-and-why-is it-useful/

Coronavirus disease 2019 (COVID-19): situation report. World Health Organization; 2020. p. 70.

Syazali M., Putra F., Rinaldi A., Utami L., Widayanti W., Umam R., Jermsittiparsert K. Partial correlation analysis using multiple linear regression: impact on business environment of digital marketing interest in the era of industrial revolution 4.0. Manag Sci Lett. 2019;9(11):1875–1886. 2019.

**Githhub link : https://github.com/syafiqahkhairi/Machine-Learning.git