WQD7006 Machine Learning

Assignment Report

Title : Covid-19 Outbreak Prediction

Name : Fatin Nabilah Abdul Raman (17201487)

Table Contents

1.0 Introduction

Coronavirus disease 2019 (COVID-19) is defined as illness caused by a novel coronavirus now called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly called 2019-nCoV), which was first identified amid an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China[1]. The first cases were initially reported to the WHO on December 31, 2019 and on January 30, 2020, the WHO declared the COVID-19 outbreak a global health emergency.

COVID-19 or Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a new type of virus family that has not been earlier identified in people. This virus was transmitted mostly through the respiratory droplets via coughing, sneezing or when people interact with each other for some time with close proximity. These droplets can then be inhaled or can land on any surfaces that others may touch who's then get contaminate when they contact their eyes, mouth, or nose which is a sensitive area.

Figure 1:COVID-19 life span on different surface

This COVID-19 can live on different surface like 2-3 days for stainless steel and 3 hours for paper and tissue paper surface (refer Figure 1). However, the amount of visible virus declines over time and may not always be present in sufficient numbers to cause infection. According to current evidence, COVID-19 virus is primarily transmitted between people through respiratory droplets and contact routes[2-7]. Even with a small droplet of COVID-19, it can be still be transmitted and the symptoms of this virus can be experienced in between 1 to 14 days or up to 21 days from the day of infection.

The symptoms (refer figure 2) of coronavirus change in severity from having no symptoms at all (being asymptomatic) to having fatigue, cough, fever, general weakness, sore throat, muscular pain and in the most extreme cases, sepsis, severe pneumonia, acute respiratory distress syndrome, and septic shock, all potentially leading to death. COVID-19 affect different people in different ways. Most infected people will develop mild to moderate illness and recover without hospitalization. However, it was strongly affecting the old people and those has a chronic medical background.

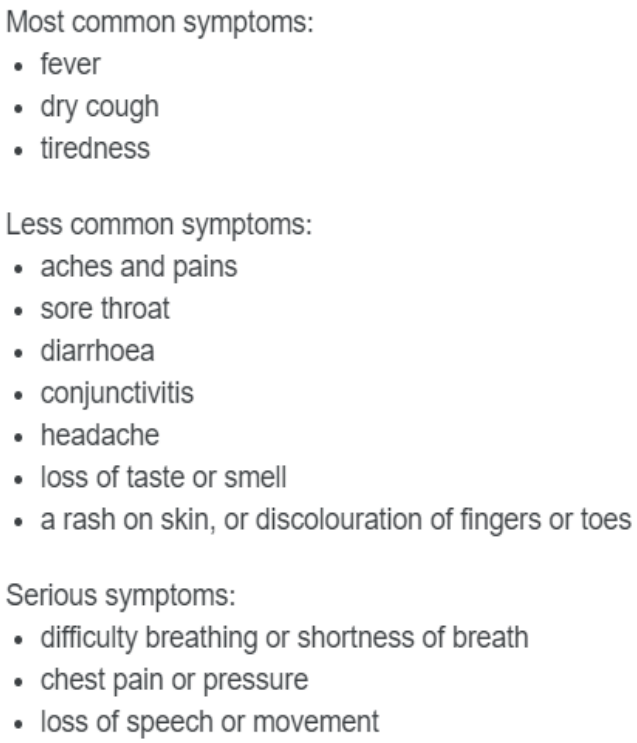

Figure 2: Symptoms of COVID-19

This COVID-19 pandemic caused a major disruption to all people. More than 90 million people have confirmed the diagnosis of the disease and more than 1.9 million people have died from it [8]. It has also had a significant impact on daily life and economic activities [9][10]. Many studies have focused on measuring who is most affected by COVID-19, or which treatments are appropriate at each stage of the disease. However, it is also crucial to understand how the spread and the growth rate of COVID-19 depends on preventive measures such as social distancing and how the reopening may affect the spread.

1.1 Objective of this project

1. to predict the spread of corona virus across the region for China, US and Malaysia
2. to analyses the growth rates and types of mitigation applied based on China, US and Malaysia

1.2 Data Source

The data was obtained from

- Kaggle: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv

The data "covid_19_data.csv" contain of 8 columns and 172481 rows. However, it will be filter out to only China, United State and Malaysia country. The data was extracted from 2 January 2020 to 11 December 2020.

Below are the details of covid_19_data.csv columns:

- Sno - Serial number
- ObservationDate - Date of the observation in MM/DD/YYYY
- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it)
- Confirmed - Cumulative number of confirmed cases till that date
- Deaths - Cumulative number of of deaths till that date

- Recovered - Cumulative number of recovered cases till that date

Data and code source can be obtained from:

- Github: FatinNabilah1/WQD7006-Machine-Learning (github.com)

2.0 Analysis and Design

In this project, it will be based on the CRISP-DM model or "Cross-Industry Process for Data Mining". Figure 3 shows the adapted CRISP-DM model.

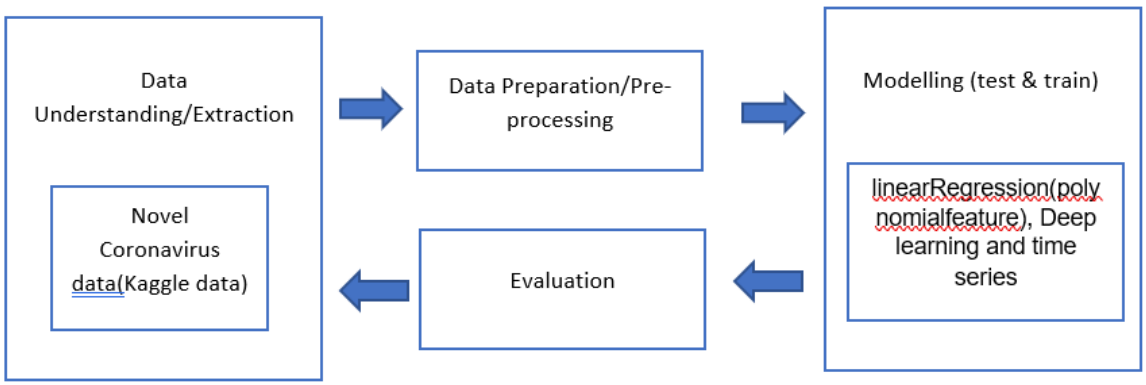

Figure 3: Workflow of the project

Figure 3 shows the four phases of CRISP-DM Model: data understanding/extraction, data preparation/preprocessing, modelling and lastly evaluation. In this project, the novel corona virus 2019 dataset or covid_19_data.csv will be used. Data preparation or data pre-processing will be discussed in Section 2.1, Section 2.2 will be described more on the model that was used for this project. Section 3 will be discussed on the experimental result of this project. The evaluation or discussion of the model will be presented in Section 4.

2.1 Data Preparation/Data Pre-processing

For the data preparation or preprocessing, the data was first observed or check for the null data or missing value. Only Province/State column contains missing value (refer Figure 5) and it was later will be impute. It was then will be group or filter for only China, United

State and Malaysia country. Besides, it was found that most of the missing or null data for the Province/State come from Malaysia data from the selected three country.

```
: # check if there exist a missing value
  mis = covid.isnull().sum()
  mis[mis>0]

: Province/State    47883
  dtype: int64
```

Figure 4:Missing Value

2.2 Model

2.2.1 Deep Learning Method (Gated Recurrent Unit (GRU)

The GRU is a new generation of Recurrent Neural Networks and is very similar to an LSTM. To solve the vanishing gradient problem of a standard RNN, GRU uses the update gate and reset gate. These are two gates decide what information should be passed to the output. These two gates can be trained to keep information from many time steps before the actual time step, without washing it through time, or to remove information which is irrelevant for the prediction. If carefully trained, GRU can perform extremely well even in complex scenarios.

In the Figure 6, a GRU unit is composed of:

- a **reset gate**, that decides how much of the information from the previous time steps can be forgotten;
- an **update gate**, that decides how much of the information from the previous time steps must be saved;
- a **memory**, that brings informations along the entire sequence and represents the memory of the network.

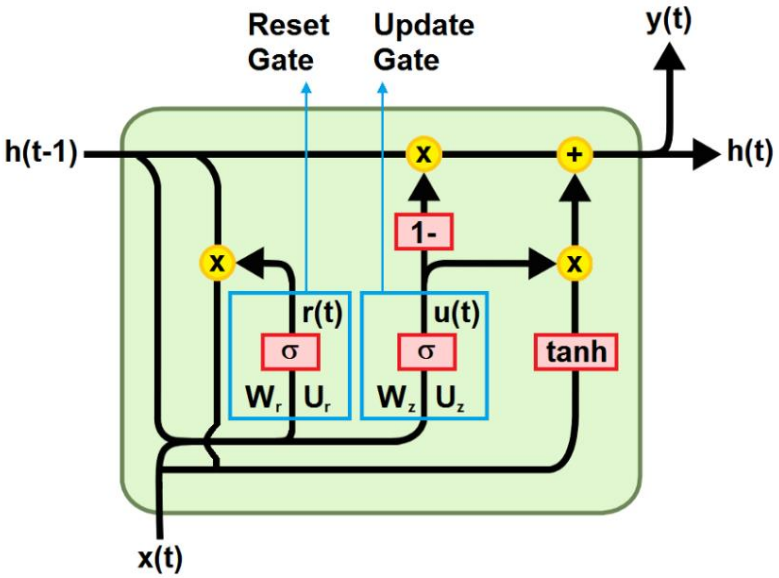

Figure 5: GRU

2.2.1 Linear regression with polynomial feature

Polynomial features are those features created by raising existing features to an exponent. The "degree" of the polynomial is used to control the number of features added, e.g. a degree of 3 will add two new variables for each input variable. A small degree is used such as 2 or 3. Typically, linear algorithms, such as linear regression and logistic regression, respond well to the use of polynomial input variables. Linear regression is linear in the model parameters and adding polynomial terms to the model can be an effective way of allowing the model to identify nonlinear patterns.

3.0 Experimental Result

The dataset used for the two-task consisted of the number of confirmed cases, recoveries, deaths from 2/01/2020 to 11/12/2020 in different Province or State, Country/Region name, and date. However, this project was mostly focused on Mainland China, US, and Malaysia.

3.1 Task 1: to predict the spread of corona virus across the region for China, US and Malaysia

The epidemic of Covid-19 is turning into a big international crisis and is beginning to have an impact on crucial facets of everyday life. For example, bans have been imposed on hotspot countries, multinational manufacturing activities have also had to slow down production, and many products manufactured exclusively in China have been entirely halted. People are beginning to stock up on critical commodities in the heavily impacted countries. In order to determine how the virus will spread through various countries and continents, GRU models were used to predict the overall number of positive cases.

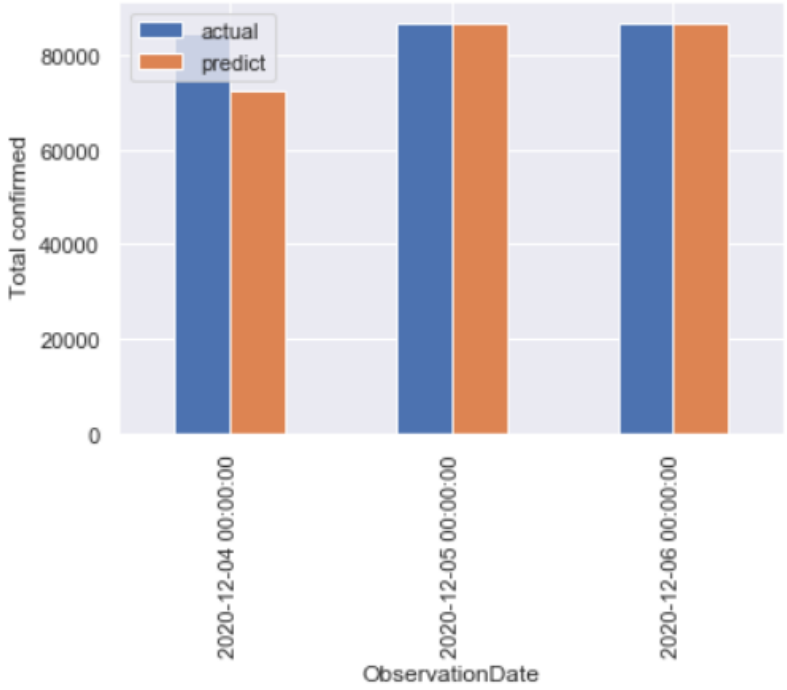

Figure 6: China Total Confirmed Actual and Predict

Result of RMSE for the test set: 10623.4615

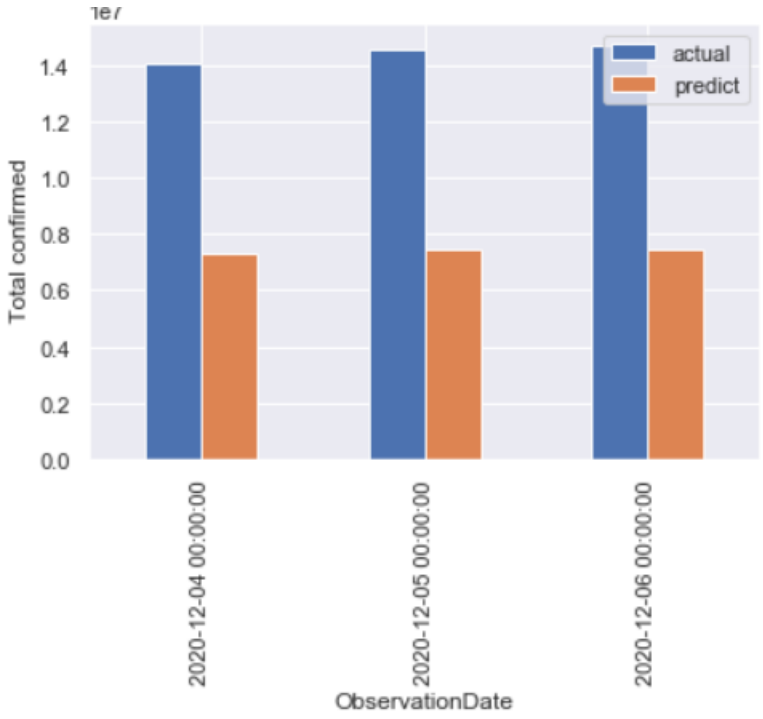

Figure 7: US Total Confirmed Actual and Predict

Result of RMSE for the test set: 317110.3204

From Figure 8, we can see from April up to Dec, China have a rapid increasing of confirmed case instead of US and Malaysia. While both US and Malaysia, remain stable from April to Dec. We have seen qualitatively, how the COVID-19 is spreading in these 3 countries. So, confirmed feature is an important feature in this data. we can make a model based only on that feature. Confirmed feature depend on time.

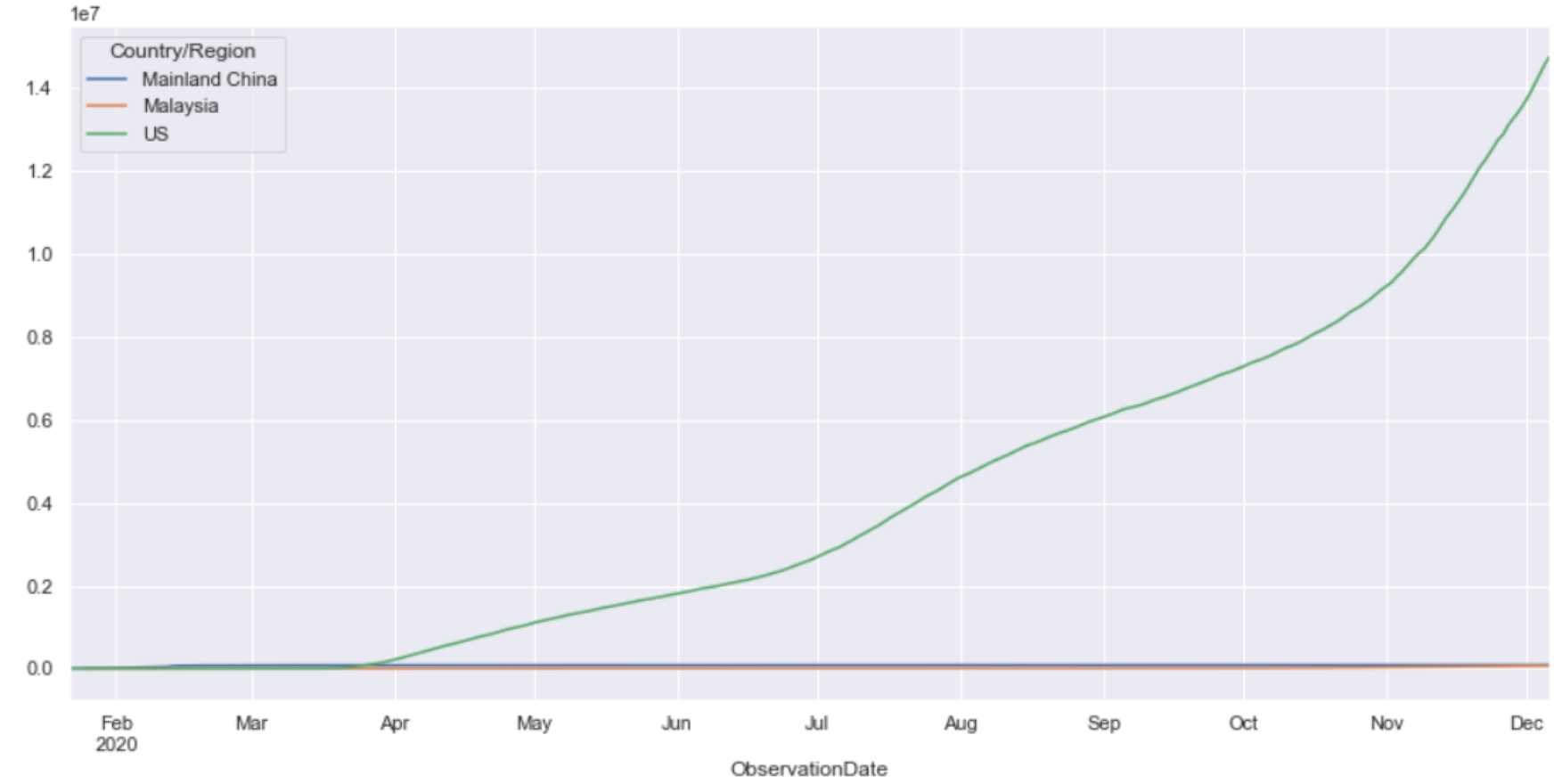

Figure 8: Graph between Observation Date and Confirmed for the 3 Selected Countries

3.2 Task 2: to analyses the growth rates and types of mitigation applied based on China, US and Malaysia

Over the last few months dt countries have performed different forms of mitigation to prevent the spread of COVID 19. These mitigations involved like the ban of large gatherings, closing of schools, banned flights, and other transportation, put cities in lockdown, etc. Such that, in order to see the effects of mitigation, this task was performed. Thus with this mitigation would it impact on the active case or grow rate of the COVID-19.

China

Statement by the Chinese authorities in Hubei province: "The measures to prevent and control the epidemic in our province have entered their decisive phase, but the situation remains extremely serious. And it is to effectively isolate the source of infection and to curb the risks of contagion that it has been decided to further tighten prophylaxis and control measures. We count on the understanding and cooperation of the inhabitants of the province, as well as on their active participation in the fight against the epidemic."

United State (US)

Trump issues ‘Coronavirus Guidelines’ for next 15 days to slow pandemic
San Francisco Mayor London Breed said, “Effective at midnight, San Francisco will require people to stay home except for essential needs. Necessary government functions & essential stores will remain open.”

Malaysia

On the 25th January 2020, the first case of COVID-19 was detected in Malaysia and traced back to 3 Chinese nationals who previously had close contact with an infected person in Singapore. Coronavirus: Malaysia in partial lockdown from March 18 to limit outbreak,:

- The lockdown will take effect for two weeks from March 18, with all businesses shut except shops selling food and daily necessities
- Malaysians will be barred from travelling overseas, while mass gatherings across the country have been prohibited

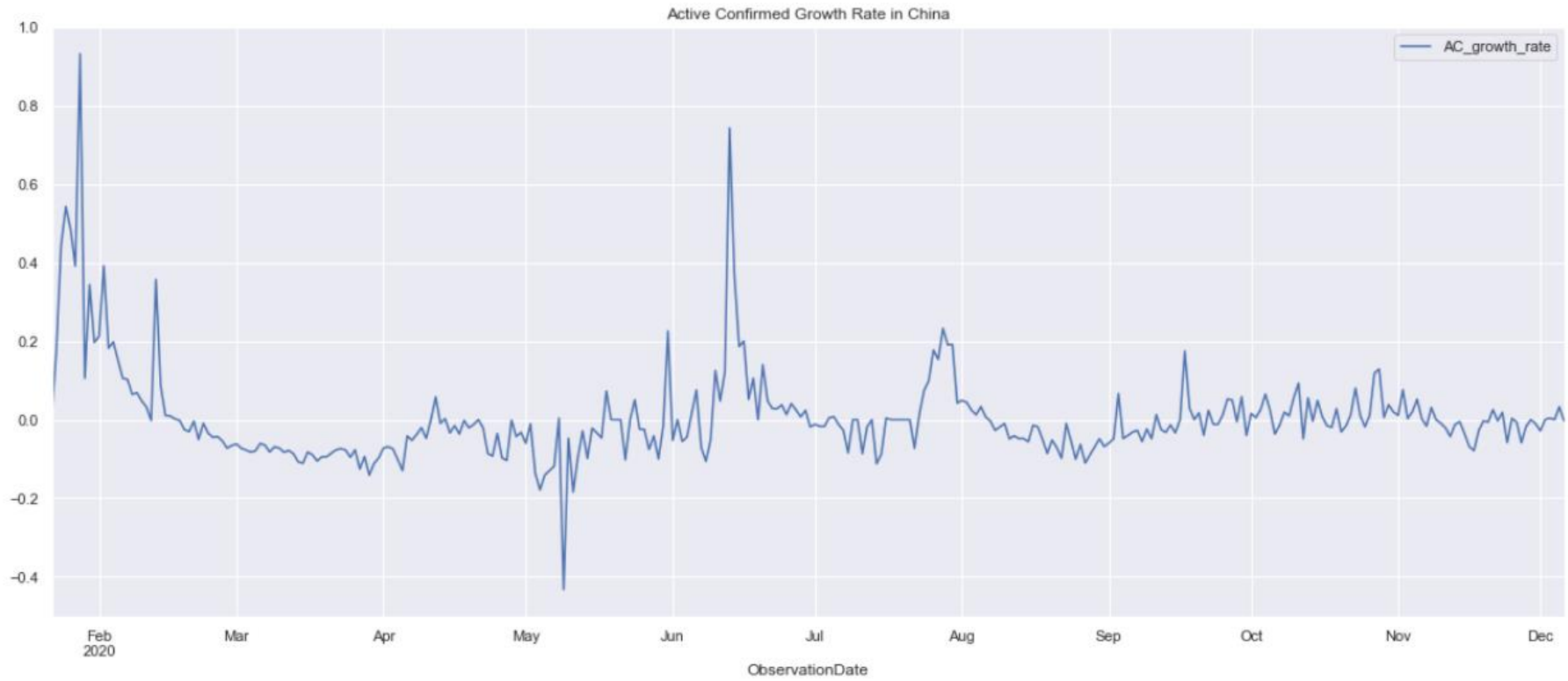

Figure 9:Active Confirmed Growth Rate in China

From Figure 9, we can see that it shows the growth rate of the active confirmed cases in Mainland China region. As mention previously, China started its complete lockdown from early January, and the total number of active cases of Mainland China is decreasing at a satisfying rate until the at the end of December 2020.

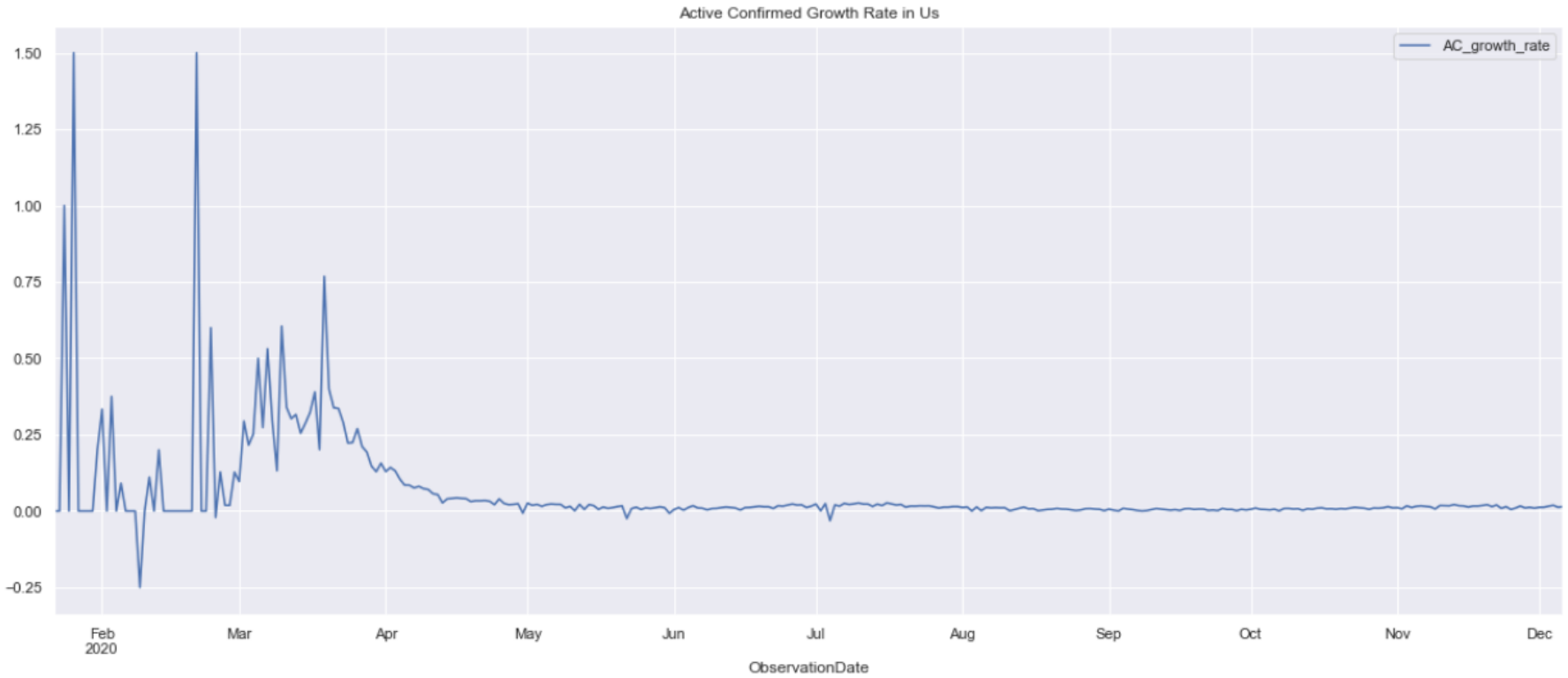

Figure 10: Active Confirmed Growth Rate in US

From Figure 10, we can see that it shows the growth rate of the active confirmed cases in United States region. As we know that United States did not do a full lock down, however, they decided to lock or apply restriction at some region and the number of active cases of United State is remain decreasing at 0 rate from May 2020 until the at the end of December 2020.

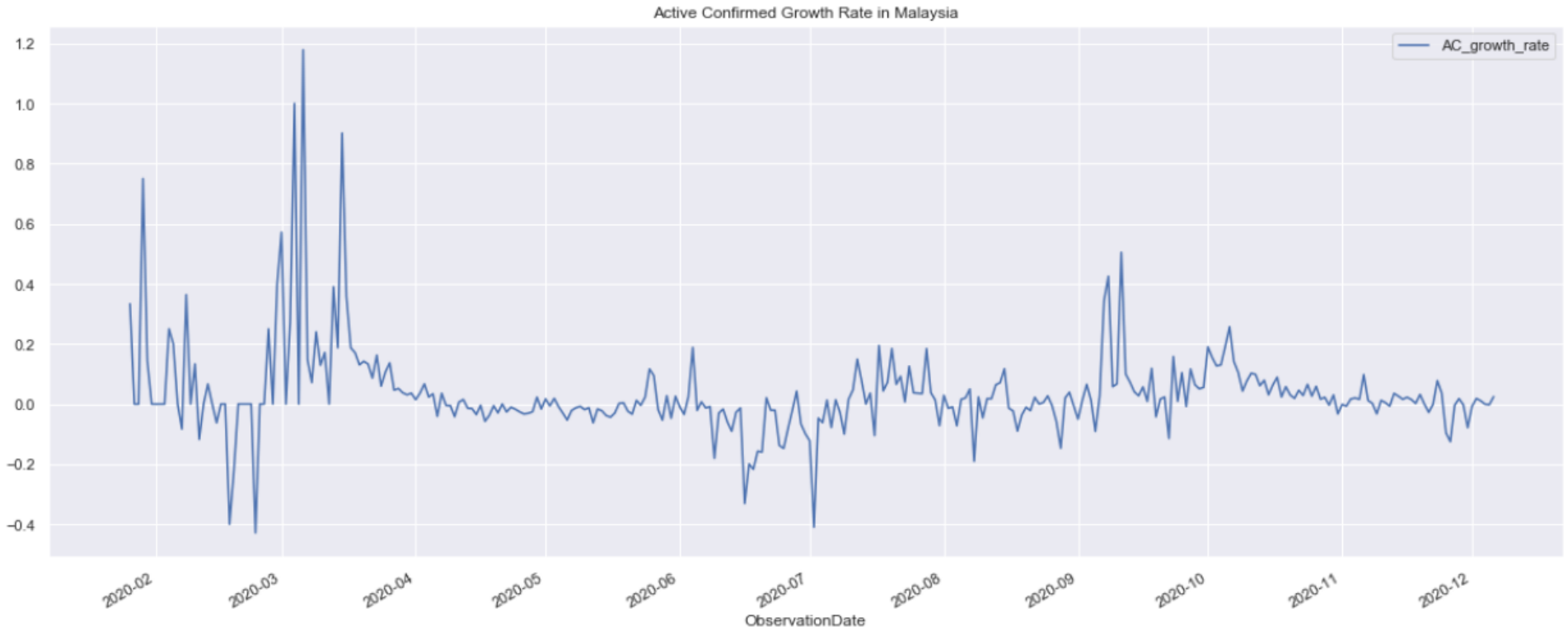

Figure 11: Active Confirmed Growth Rate in Malaysia

From Figure 11, we can see that it shows the growth rate of the active confirmed cases in Malaysia region. For your information, Malaysia did start the lockdown from 18 March 2020 after the spreading of COVID which started from the first detected casesAnd from Figure above, we can see that the growth rate of active confirmed case was going up and down during the lockdown with peak rate at March and September 2020.

4.0 Discussion or Evaluation

China

From this Figure 12, we can see that Hubei have the highest number of confirmed, death and recovered cases reported with 68149, 4512, 63633.

	Province/State	Confirmed	Deaths	Recovered	active_confirmed
171912	Anhui	992.000000	6.000000	986.000000	0.000000
171946	Beijing	952.000000	9.000000	939.000000	4.000000
171996	Chongqing	590.000000	6.000000	583.000000	1.000000
172038	Fujian	500.000000	1.000000	453.000000	46.000000
172043	Gansu	182.000000	2.000000	180.000000	0.000000
172061	Guangdong	2004.000000	8.000000	1960.000000	36.000000
172062	Guangxi	263.000000	2.000000	260.000000	1.000000
172065	Guizhou	147.000000	2.000000	145.000000	0.000000
172068	Hainan	171.000000	6.000000	165.000000	0.000000
172074	Hebei	373.000000	6.000000	367.000000	0.000000
172075	Heilongjiang	949.000000	13.000000	936.000000	0.000000
172076	Henan	1295.000000	22.000000	1266.000000	7.000000
172085	Hubei	68149.000000	4512.000000	63633.000000	4.000000
172087	Hunan	1020.000000	4.000000	1016.000000	0.000000
172095	Inner Mongolia	336.000000	1.000000	308.000000	27.000000
172109	Jiangsu	680.000000	0.000000	674.000000	6.000000
172110	Jiangxi	935.000000	1.000000	934.000000	0.000000
172111	Jilin	157.000000	2.000000	155.000000	0.000000
172159	Liaoning	289.000000	2.000000	287.000000	0.000000
172243	Ningxia	75.000000	0.000000	75.000000	0.000000
172308	Qinghai	18.000000	0.000000	18.000000	0.000000
172353	Shaanxi	502.000000	3.000000	478.000000	21.000000
172354	Shandong	857.000000	7.000000	841.000000	9.000000
172355	Shanghai	1366.000000	7.000000	1294.000000	65.000000
172356	Shanxi	222.000000	0.000000	218.000000	4.000000
172360	Sichuan	812.000000	3.000000	779.000000	30.000000
172392	Tianjin	301.000000	3.000000	292.000000	6.000000
172393	Tibet	1.000000	0.000000	1.000000	0.000000
172463	Xinjiang	980.000000	3.000000	977.000000	0.000000
172471	Yunnan	221.000000	2.000000	210.000000	9.000000
172477	Zhejiang	1295.000000	1.000000	1288.000000	6.000000

Figure 12: China report on Covid-19

United State

From this Figure 13, we can see that California, Texas and Florida have the highest number of confirmed cases reported with 1366673, 1322738, and 1058074. While New York, Texas and California have the highest number of deaths cases reported with 34958, 23137, and 19928. It also show that there the highest number of recovered cases reported was 5624444 in Recovered Province/State

	Province/State	Confirmed	Deaths	Recovered	active_confirmed
171896	Alabama	269877.000000	3889.000000	0.000000	265988.000000
171898	Alaska	37036.000000	143.000000	0.000000	36893.000000
171923	Arizona	364276.000000	6950.000000	0.000000	357326.000000
171924	Arkansas	170924.000000	2660.000000	0.000000	168264.000000
171967	California	1366673.000000	19928.000000	0.000000	1346745.000000
172002	Colorado	260581.000000	3356.000000	0.000000	257225.000000
172003	Connecticut	127715.000000	5146.000000	0.000000	122569.000000
172013	Delaware	39912.000000	793.000000	0.000000	39119.000000
172016	Diamond Princess cruise ship	49.000000	0.000000	0.000000	49.000000
172017	District of Columbia	23136.000000	697.000000	0.000000	22439.000000
172033	Florida	1058074.000000	19177.000000	0.000000	1038897.000000
172046	Georgia	501405.000000	9806.000000	0.000000	491599.000000
172054	Grand Princess	103.000000	3.000000	0.000000	100.000000
172059	Guam	6959.000000	113.000000	0.000000	6846.000000
172073	Hawaii	18842.000000	262.000000	0.000000	18580.000000
172091	Idaho	110510.000000	1035.000000	0.000000	109475.000000
172092	Illinois	787573.000000	14116.000000	0.000000	773457.000000
172093	Indiana	381617.000000	6242.000000	0.000000	375375.000000
172096	Iowa	244691.000000	2717.000000	0.000000	241974.000000
172123	Kansas	171364.000000	1786.000000	0.000000	169578.000000
172128	Kentucky	200631.000000	2072.000000	0.000000	198559.000000

172170	Louisiana	251123.000000	6584.000000	0.000000	244539.000000
172182	Maine	13348.000000	227.000000	0.000000	13121.000000
172189	Maryland	215027.000000	4846.000000	0.000000	210181.000000
172190	Massachusetts	256844.000000	11004.000000	0.000000	245840.000000
172201	Michigan	426576.000000	10321.000000	0.000000	416255.000000
172205	Minnesota	350862.000000	4043.000000	0.000000	346819.000000
172206	Mississippi	164931.000000	3961.000000	0.000000	160970.000000
172207	Missouri	329420.000000	4284.000000	0.000000	325136.000000
172212	Montana	67875.000000	736.000000	0.000000	67139.000000
172230	Nebraska	139834.000000	1205.000000	0.000000	138629.000000
172232	Nevada	168140.000000	2315.000000	0.000000	165825.000000
172235	New Hampshire	24888.000000	564.000000	0.000000	24324.000000
172236	New Jersey	368016.000000	17321.000000	0.000000	350695.000000
172237	New Mexico	108088.000000	1749.000000	0.000000	106339.000000
172239	New York	705827.000000	34958.000000	0.000000	670869.000000
172250	North Carolina	394990.000000	5543.000000	0.000000	389447.000000
172251	North Dakota	82981.000000	1019.000000	0.000000	81962.000000
172254	Northern Mariana Islands	109.000000	2.000000	0.000000	107.000000
172267	Ohio	475024.000000	6959.000000	0.000000	468065.000000
172271	Oklahoma	216486.000000	1896.000000	0.000000	214590.000000
172275	Oregon	84496.000000	1033.000000	0.000000	83463.000000
172288	Pennsylvania	423100.000000	11255.000000	0.000000	411845.000000
172302	Puerto Rico	56671.000000	1192.000000	0.000000	55479.000000
172315	Recovered	0.000000	0.000000	5624444.000000	-5624444.000000
172319	Rhode Island	62137.000000	1413.000000	0.000000	60724.000000
172371	South Carolina	232099.000000	4566.000000	0.000000	227533.000000
172372	South Dakota	85991.000000	1110.000000	0.000000	84881.000000
172388	Tennessee	400594.000000	4943.000000	0.000000	395651.000000
172390	Texas	1322738.000000	23137.000000	0.000000	1299601.000000
172427	Utah	215407.000000	939.000000	0.000000	214468.000000
172442	Vermont	5015.000000	79.000000	0.000000	4936.000000
172446	Virgin Islands	1633.000000	23.000000	0.000000	1610.000000
172447	Virginia	255053.000000	4200.000000	0.000000	250853.000000
172456	Washington	177447.000000	2925.000000	0.000000	174522.000000
172459	West Virginia	54997.000000	838.000000	0.000000	54159.000000
172461	Wisconsin	441067.000000	3952.000000	0.000000	437115.000000
172462	Wyoming	36218.000000	266.000000	0.000000	35952.000000

Figure 13: United State report on Covid-19

Malaysia

There are no details on the Province/State for Malaysia in the data.

	Province/State	Confirmed	Deaths	Recovered	active_confirmed
171815	missing_value	72694.000000	382.000000	61273.000000	11039.000000

Figure 14: Malaysia report on Covid-19

4.1 Limitation of the project

In this project, there are a few limitations. Firstly, the data, the data that was obtain from Kaggle seems to be outdated where it only covered up to early December 2020 or 11 December 2020. Beside that, this data also seems have some anomaly or incorrect data. Example: Province or State for US country like "Recovered". This "Recovered" Province or State cannot be found under United State list of province or state. Besides that, this "Recovered" Province or State has the highest Cumulative number of recovered cases with 500000++. Meanwhile, the data was not a complete data. Example, in this research, we will filter out or group and select China, US and Malaysia for the selected country. So, we found that Malaysia data did not have any Province or State details or value. It would be better if the province or state for Malaysia have a reasonable value instead of "missing_value".

5.0 Conclusion

COVID-19 causes illness in humans and creates severe damage in the lungs. However, COVID-19 has killed many people in the entire world. In this paper, there are two tasks objective to achieves. The main advantage of doing the first task would be that, it will give the idea about the level of spread, and in accordance to that, the government and the citizens can make proper plans to handle the situation by taking measures to minimize the virus spread by various mitigation and other necessary actions. With the help of second task, it will have an idea about how well the mitigations are working, and the actions that are taken

till date how effective are they, or how many cases have been prevented by this and we can see the grow rate from the start of lockdown to the end of year.

6.0 Reference

[1]. Zhu N., Zhang D., Wang W., Li X., Yang B., Song J. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020

[2] Liu J, Liao X, Qian S et al. Community transmission of severe acute respiratory syndrome coronavirus 2, Shenzhen, China, 2020. Emerg Infect Dis 2020 doi.org/10.3201/eid2606.200239

[3] Chan J, Yuan S, Kok K et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 2020 doi: 10.1016/S0140-6736(20)30154-9

[4] Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020; doi:10.1056/NEJMoa2001316.

[5] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020; 395: 497–506.

[6] Burke RM, Midgley CM, Dratch A, Fenstersheib M, Haupt T, Holshue M,et al. Active monitoring of persons exposed to patients with confirmed COVID-19 — United States, January–February 2020. MMWR Morb Mortal Wkly Rep. 2020 doi : 10.15585/mmwr.mm6909e1external icon

[7] World Health Organization. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) 16-24 February 2020 [Internet]. Geneva: World Health Organization; 2020 Available from: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

[8] World Health Organization coronavirus disease (COVID-19) dashboard. https://covid19.who.int/. Accessed on January 12, 2021.

[9] The Opportunity Insights economic tracker. https://tracktherecovery.org/. Accessed on January 12, 2021.

[10] Google COVID-19 community mobility reports.
https://www.google.com/covid19/mobility. Accessed on January 12, 2021.