Accurate prediction and review of the COVID-19 development trend

Robin Chiang

January 18, 2021

1. Introduction

1.1 Background

Coronavirus is a family of viruses that are named after their spiky crown. The novel coronavirus, also known as SARS-CoV-2, is a contagious respiratory virus that first reported in Wuhan, China. On 2/11/2020, the World Health Organization designated the name COVID-19 for the disease caused by the novel coronavirus. Now the virus is sweeping the world, a serious threat to human health and well-being and life. As of 2 January 2021, more than 83.9 million cases have been confirmed, with more than 1.82 million deaths attributed to COVID-19. Therefore, it is beneficial for all human beings to accurately predict the trend of covid-19 in the future and make appropriate prevention. This report aims at exploring COVID-19 through data analysis and projections.

1.2 Problem

Symptoms of COVID-19 are highly variable, ranging from none to severe illness. The virus spreads mainly through the air when people are near each other. It leaves an infected person as they breathe, cough, sneeze, or speak and enters another person via their mouth, nose, or eyes. It may also spread via contaminated surfaces. People remain infectious for up to two weeks and can spread the virus even if they do not show symptoms. Data that might contribute to determining COVID-19 development trend might include it performance in different countries, daily confirmed cases and daily deaths cases. This project aims to predict whether COVID-19 will slow down or intensify in the future based on these data.

1.3 Interest

Obviously, this virus is attacking people all over the world. Therefore, the World Health Organization, government authorities, public health experts and ordinary citizens are very interested in whether they can accurately predict the development trend of COVID-19 and effectively prevent the spread of the epidemic.

2. Data acquisition and cleaning

2.1 Data sources

This is a daily updating version of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The data updates every day at 6am UTC, which updates just after the raw JHU data typically updates.

2.2 Data cleaning

The data in the data table is very simple and does not need to be cleaned up too much. However, there are several problems with the datasets. First, the new ***Date*** column are all string with ***mm/dd/yy*** format, therefore we have to convert it to datetime values. Second, replacing missing value ***NaN***. We can find a lot ***NaN*** in the ***Province/State*** by running the test, and that makes sense as many countries only report the ***Country/Region*** data. However, there are 1,602 ***NaN*s** in ***Recovered*** and let's replace them with 0. Third, there are COVID-19 cases reported from 3 cruise ships: Grand Princess, Diamond Princess and MS Zaandam. These data need to be extracted and treated differently due to ***Province/State*** and ***Country/Region*** mismatch over time.

2.3 Feature selection

After data cleaning, there were 95,472 samples and 9 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features such as, ***Province/State, Lat*** and ***Long***. Let's aggregate data from ***Province/State*** into the total number of those countries and remove ***Lat*** and ***Long*** of non-critical data. Next, aggregate data into ***Country/Region*** wise and group them by ***Date*** and ***Country/Region***, and the total count of ***Confirmed***, ***Deaths***, ***Recovered***, ***Active*** for the given ***Date*** and ***Country/Region*** will be summarized one by one. Now add day wise ***New cases***, ***New deaths*** and ***New recovered*** by deducting the corresponding accumulative data on the previous day.

Table 1. Simple feature selection during data cleaning.

Kept features	Dropped features	New features
Country/Region	Province/State	
Confirmed, Deaths, Recovered, Active	Lat, Long	
Date		New cases, New deaths, New recovered

3. Exploratory Data Analysis

3.1 Calculation of target variable

The cumulative confirmed, deaths, recovered and active number of each country were not a feature in the dataset, and had to be calculated. I chose to sum up the daily confirmed, deaths, and other items of each country as the target variable. And calculate the new project based on the daily change difference of each project, including new cases, new deaths and new recovered cases. The cumulative number is the easiest to explain. After all, to better understand the COVID-19 and effectively control the epidemic situation, we need to analyze the development trend and future changes led by the cumulative number of viruses.

3.2 Relationship between top ten economies and COVID-19

The study of this subtopic will help to explore the potential impact of the virus on the world's major economic powers, and select the 2020 top ten economies in the world, including the United States, China, Japan, Germany, the United Kingdom, France, India, Brazil, Italy and Russia (Figure 1).

3.3 Countries with excellent results in fighting COVID-19

In addition to the world's top 10 economies, I also selected some countries with outstanding results in fighting the epidemic, so as to compare their epidemic prevention policies and differences. This subtopic includes Australia, New Zealand and Taiwan (Figure 1).

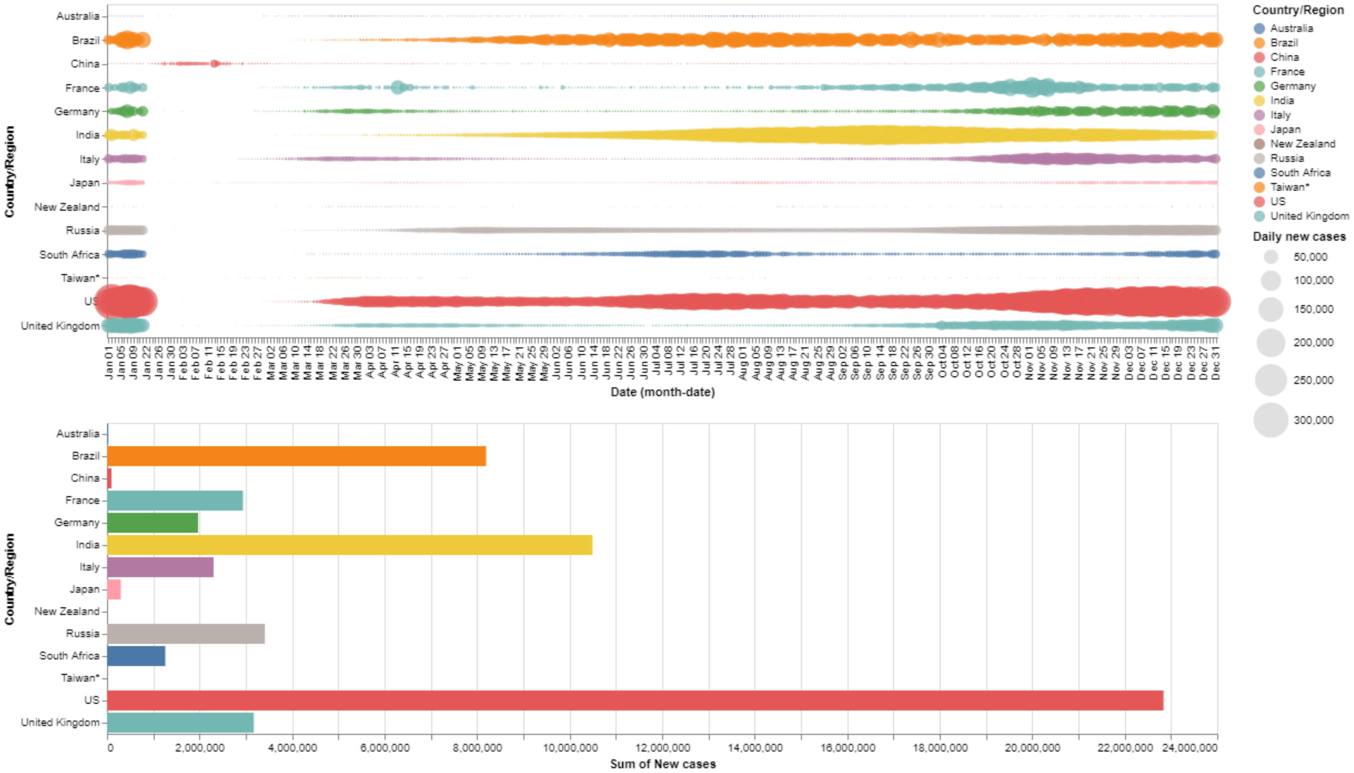

Figure 1. Changes in the number of confirmed COVID-19 per day in the top ten economies and countries with outstanding anti epidemic achievements in 2020.

3.4 Search the top 10 countries with COVID-19 confirmed.

I selected the 10 countries with the most confirmed cases of epidemic in the world from the database to analyze which countries belong to the top 10 economies and which are not. The number of confirmed cases from high to low were the United States(23,489,378), India(10,512,093), Brazil(8,326,115), Russia(3,495,816), the United Kingdom(3,260,258), France(2,851,670), Turkey(2,364,801), Italy(2,336,279), Spain(2,211,967) and Germany(2,003,985). However, Turkey and Spain are not among the world's top 10 economies (Figure 2).

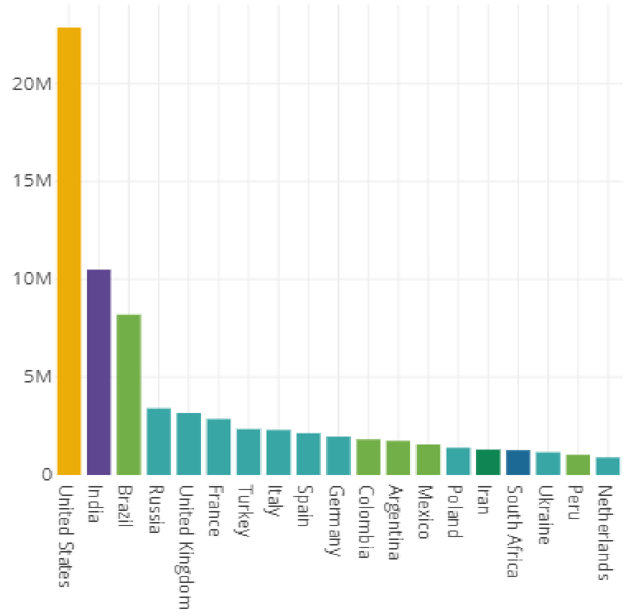

Figure 2. Some countries with the most confirmed cases in the world.

3.5 Search the top 10 countries with COVID-19 deaths

In this subtopic I selected the 10 countries with the most deaths of epidemic in the world from the database to analyze which countries belong to the top 10 economies and which are not. The number of deaths from high to low were the United States, Brazil, India, Mexico, the United Kingdom, Italy, France, Russia, Iran, Spain. However, Mexico, Iran and Spain are not among the world's top 10 economies (Figure 3).

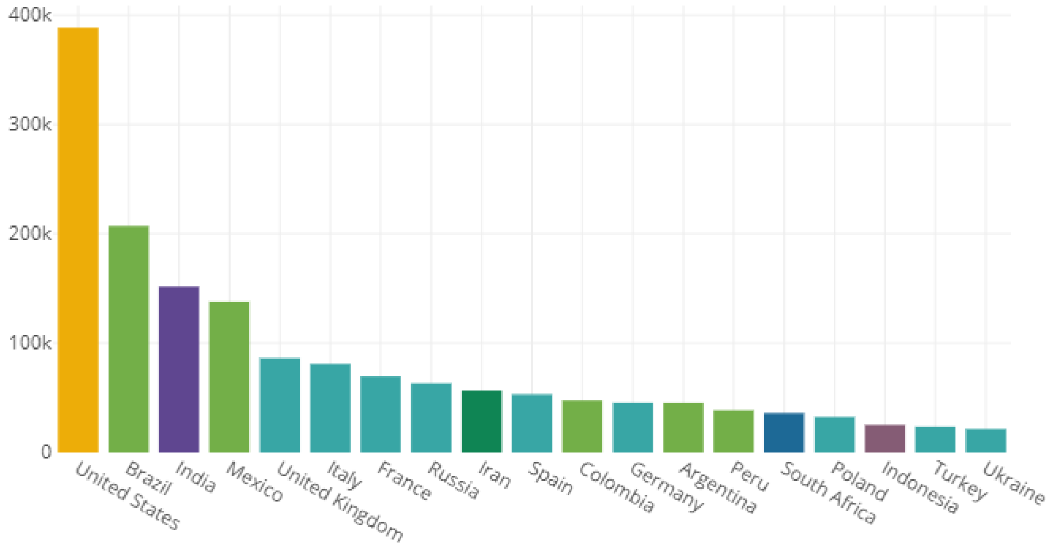

Figure 3. Some countries with the most deaths in the world.

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict COVID-19 development trend. Regression models can provide additional information on the amount of improvement. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. However, the linear regression model can accurately predict the future trend of the epidemic. Therefore, linear regression model was used in this study.

4.1 Regression models

4.1.1 Analysis and prediction of COVID-19 trend in the USA

As the world's first largest economy, the United States reflects its difficulty in fighting COVID-19. In addition to the different epidemic prevention measures of the state governments, due to a large number of passengers, business passengers and import and export goods, it is very difficult to control the border at the initial stage of the epidemic. Therefore, the United States has become the country with the highest number of confirmed cases and deaths in the world. Using the cumulative number of confirmed cases and daily cases, I used the predictive linear model to calculate the development trend of the epidemic situation in the United States in the next 90 days (Figure 4).

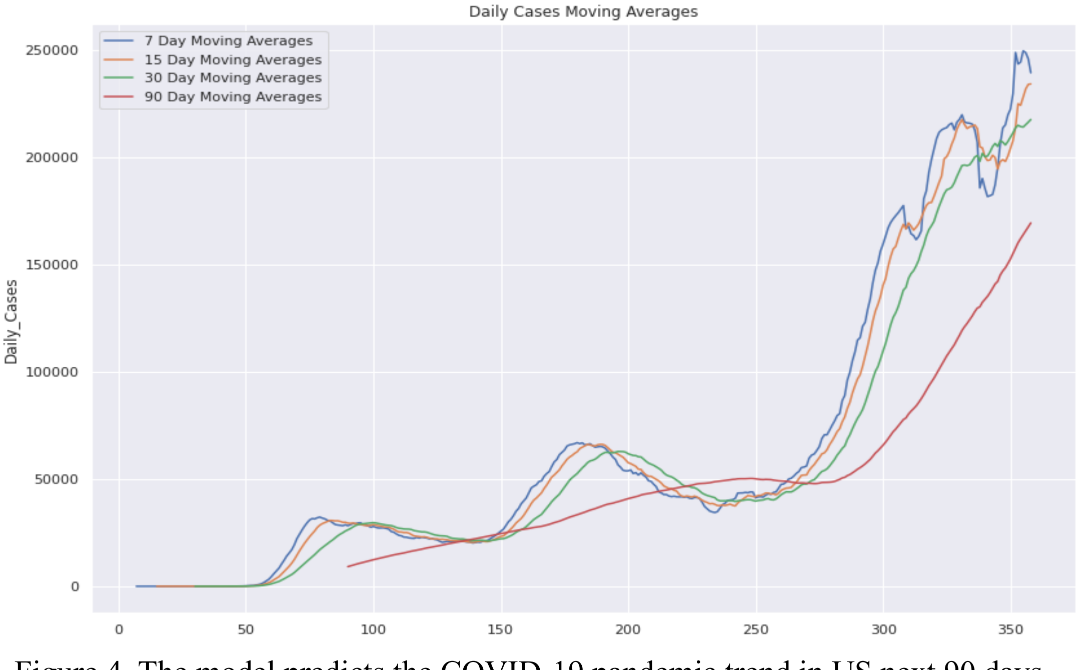

Figure 4. The model predicts the COVID-19 pandemic trend in US next 90 days.

4.1.2 Analysis and prediction of COVID-19 trend in India

India is the country with the second-largest number of confirmed cases, the seventh-largest economy and the second-largest population in the world. India has serious loopholes and mistakes in epidemic prevention. India has suffered severe economic losses during the epidemic, with GDP shrinking by 23.9% and unemployment rate exceeding 23%. This makes the Indian authorities rush to lift the city closures and various control measures before the epidemic slows down, and the number of confirmed cases will surpass Brazil in September 2020. Furthermore, the majority of citizens neglect the epidemic prevention regulations, which leads to the accelerated deterioration of the epidemic situation. The sultry weather makes it impossible for people to wear masks all day. I also used India's cumulative confirmed cases and daily confirmed cases to make a linear prediction model in the next 90 days (Figure 5).

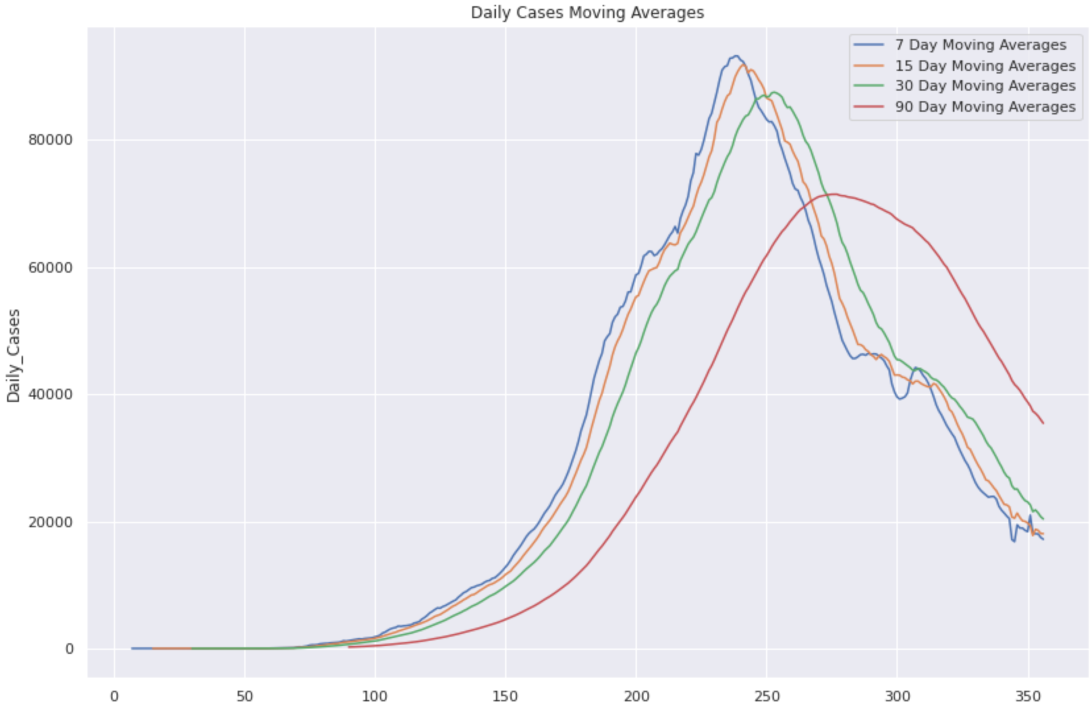

Figure 5. The model predicts the COVID-19 pandemic trend in India next 90 days.

4.1.3 Analysis and prediction of COVID-19 trend in Brazil

Brazil is the eighth largest economy in the world and the sixth most populous country in the world. It is a big country and a major economy in South America. But the authorities ignored the infectivity of the virus and downplayed the severity of the epidemic, and even the president and the people were not willing to wear masks to block the virus. As a result, the number of COVID-19 confirmed cases in Brazil has soared to the third-highest, and the deaths are also the second largest in the world. The development trend of the epidemic in Brazil in the next 90 days is shown in the figure below (Figure 6).

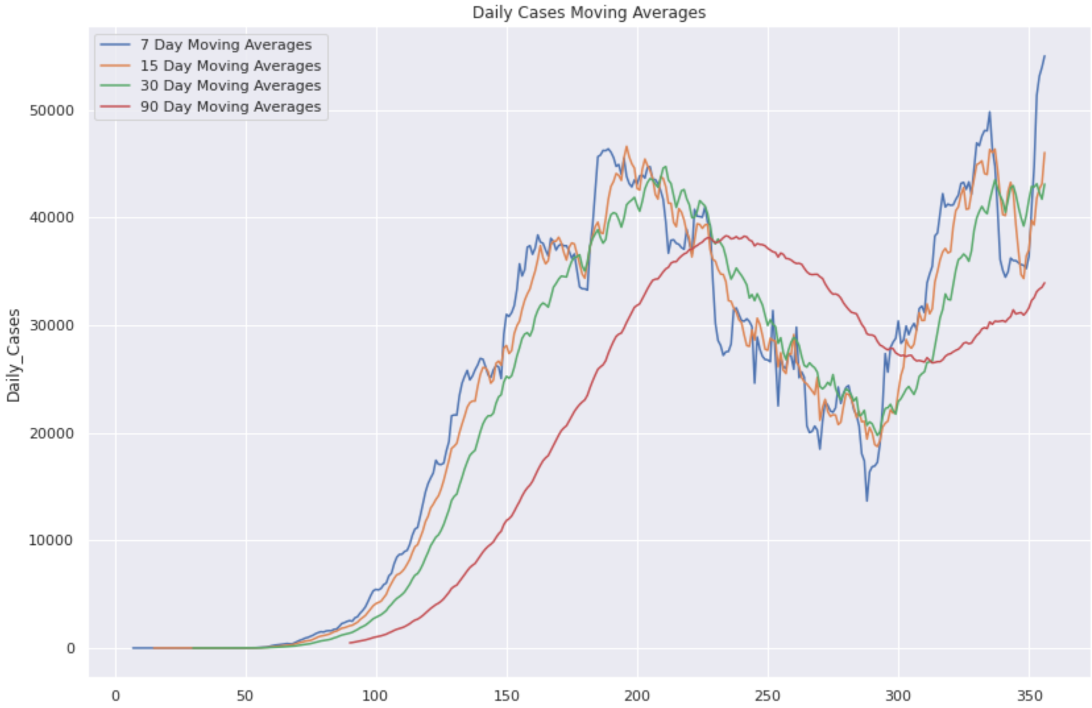

Figure 6. The model predicts the COVID-19 pandemic trend in Brazil next 90 days.

4.1.4 Analysis and prediction of COVID-19 trend in Russia

Russia is the world's tenth-largest economy and the ninth most populous country, but it is also the country with the fourth-largest number of confirmed cases and the eighth largest number of deaths. This is due to the fact that the Russian government played down the epidemic situation in the early days. Apart from rendering the infection of the COVID-19 pandemic as a foreign problem, the authority also claimed that there was no real epidemic in Russia. As a result, people get together as usual and even go to the beach for the holiday, which leads to a sharp increase in the number of COVID-19 confirmed cases. The development trend of the epidemic in Russia in the next 90 days is shown in the figure below (Figure 7).

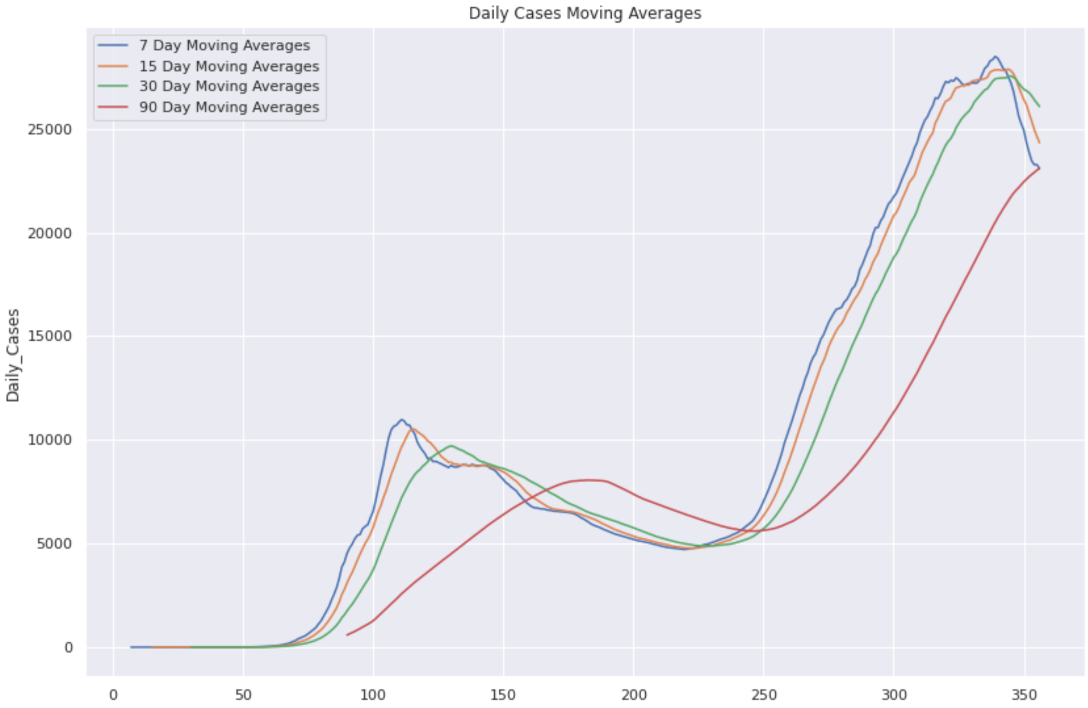

Figure 7. The model predicts the COVID-19 pandemic trend in Russia next 90 days.

4.1.5 Analysis and prediction of COVID-19 trend in UK

As the world's fifth-largest economy, Britain is still in a state of failure in the fight against the virus. Like most countries, the British government neglected the task of epidemic prevention in the early stage. Even Prime Minister Johnson was diagnosed with the virus in March 2020. Although the UK government has strengthened its anti-epidemic measures since then, they still cannot turn the tide. A new variant of the COVID-19 appeared at the end of 2020 and is now spreading to the world. This situation undoubtedly worsens the current global virus pandemic. In addition, according to the World Bank forecast, the UK will decline from the world's fifth largest economy to the seventh largest economy in 2025.The development trend of the epidemic in the United Kingdom in the next 90 days is shown in the figure below (Figure 8).

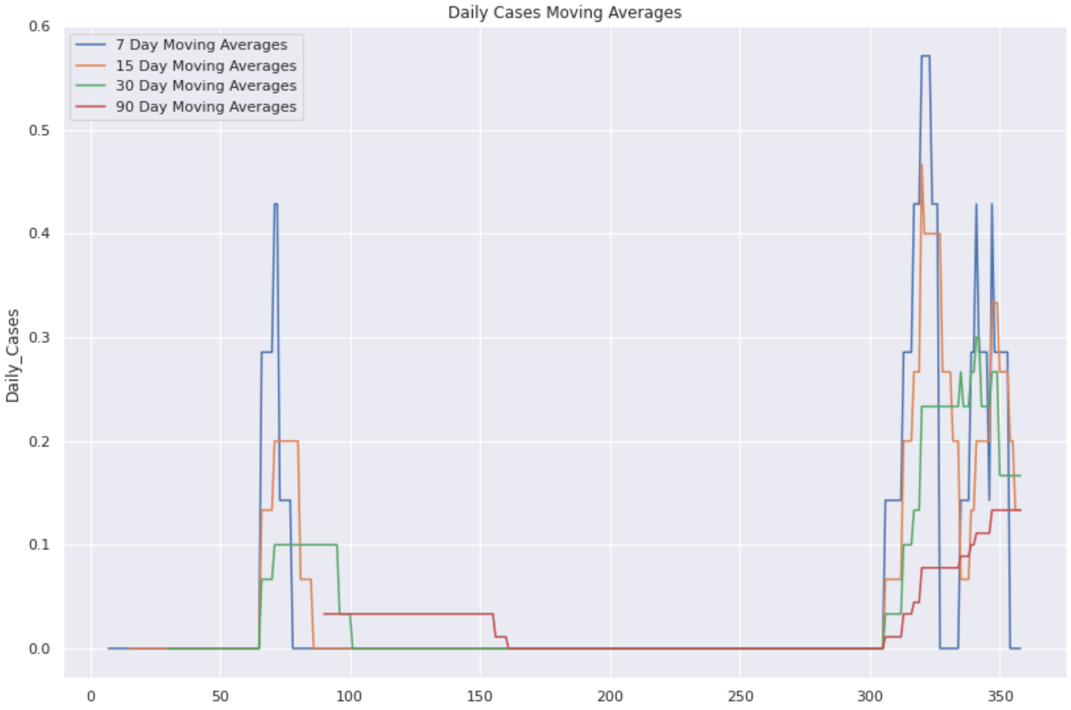

Figure 8. The model predicts the COVID-19 pandemic trend in UK next 90 days.

4.1.6 Analysis and prediction of COVID-19 trend in France

France is also one of the world's top ten economies, ranking sixth. The country with the sixth-largest number of confirmed cases and the seventh-largest number of deaths is in the world. France's poor epidemic prevention is attributed to the negligence of the government, the people's misconception of the disease and the strong romantic liberalism of its citizens. Many French claims that wearing masks will deprive them of their freedom to breathe and that the policy of closing cities will lead to their loss of freedom. Therefore, their citizens took to the streets to protest and even turned into riots. All kinds of protest actions challenge the French authorities, which also makes the epidemic situation in France more serious. The development trend of the epidemic in France in the next 90 days is shown in the figure below (Figure 9).

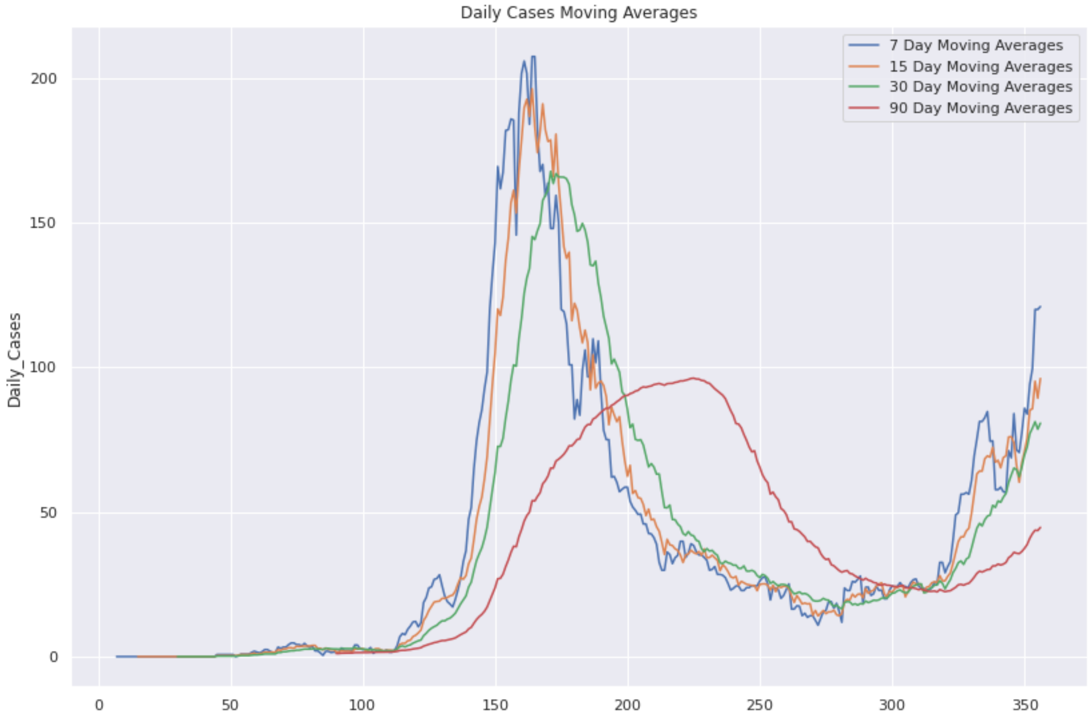

Figure 9. The model predicts the COVID-19 pandemic trend in France next 90 days.

4.1.7 Analysis and prediction of COVID-19 trend in Italy

Italy is the ninth-largest economy in the world. It is also the first country in Europe to break out COVID-19 confirmed cases. Italy, which has long relied on tourism revenue, is now facing severe challenges. Since the financial tsunami in 2008, the economy of the euro area has been damaged, especially Italy, which has experienced a financial crisis and sovereign debt crisis. The euro area has implemented a strict budget deficit policy. Various factors forced Italy to cut its budget, leading to the collapse of the medical system, which is one of the important reasons for the epidemic out of control. The development trend of the epidemic in Italy in the next 90 days is shown in the figure below (Figure 10).

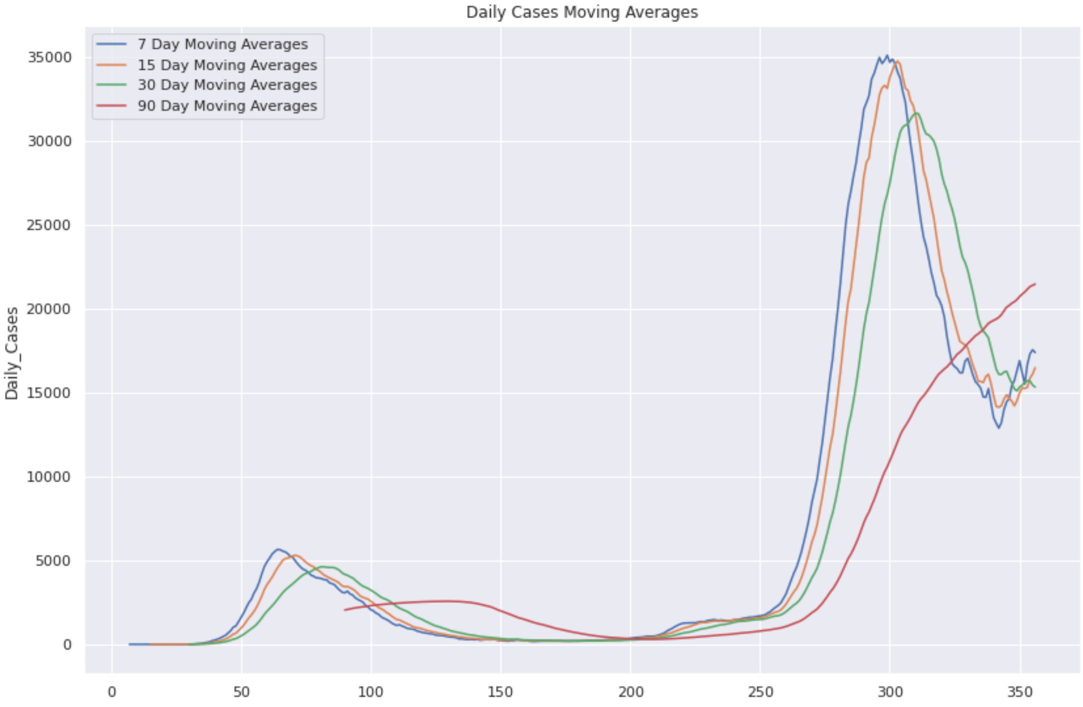

Figure 10. The model predicts the COVID-19 pandemic trend in Italy next 90 days.

4.1.7 Analysis and prediction of COVID-19 trend in Germany

As the fourth-largest economy in the world, Germany is also the largest economy in Europe. Although the implementation time of epidemic prevention measures is slow, the epidemic prevention work is still rigorous. The number of confirmed cases is the tenth in the world and the number of deaths in the twelfth. Compared with other economic countries, the number of deaths is relatively small. But the government's challenge comes from the voice of citizens. Like France, many German citizens feel that they are deprived of breathing freedom by wearing masks. Tens of thousands of citizens took to the streets to protest absurdly, challenging the decision-making ability of the government authorities. The development trend of the epidemic in Germany in the next 90 days is shown in the figure below (Figure 11).

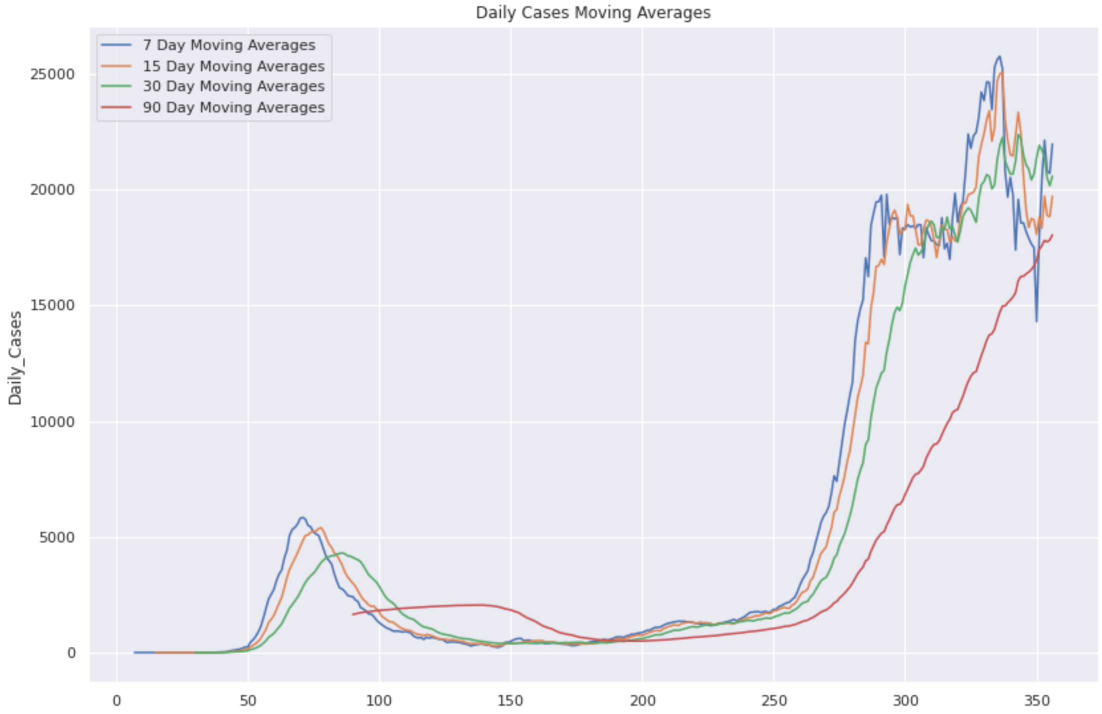

Figure 11. The model predicts COVID-19 pandemic trend in Germany next 90 days.

4.1.8 Analysis and prediction of COVID-19 trend in China

China is the second-largest economy and the most populous country in the world. However, as the world's recognized birthplace of the virus, the number of confirmed cases and deaths in this country is far lower than that in other countries, which is doubtful. It is not hard to see from the observation of the epidemic prevention measures in this country that due to the autocratic totalitarian regime, the epidemic prevention measures are very strict. At the beginning of the epidemic, China built many mobile cabin hospitals to receive and treat many confirmed patients, and strictly blocked the information about the virus. This has made it impossible for the international community to have a clear understanding of the epidemic situation in China. The development trend of the epidemic in China in the next 90 days is shown in the figure below (Figure 12).

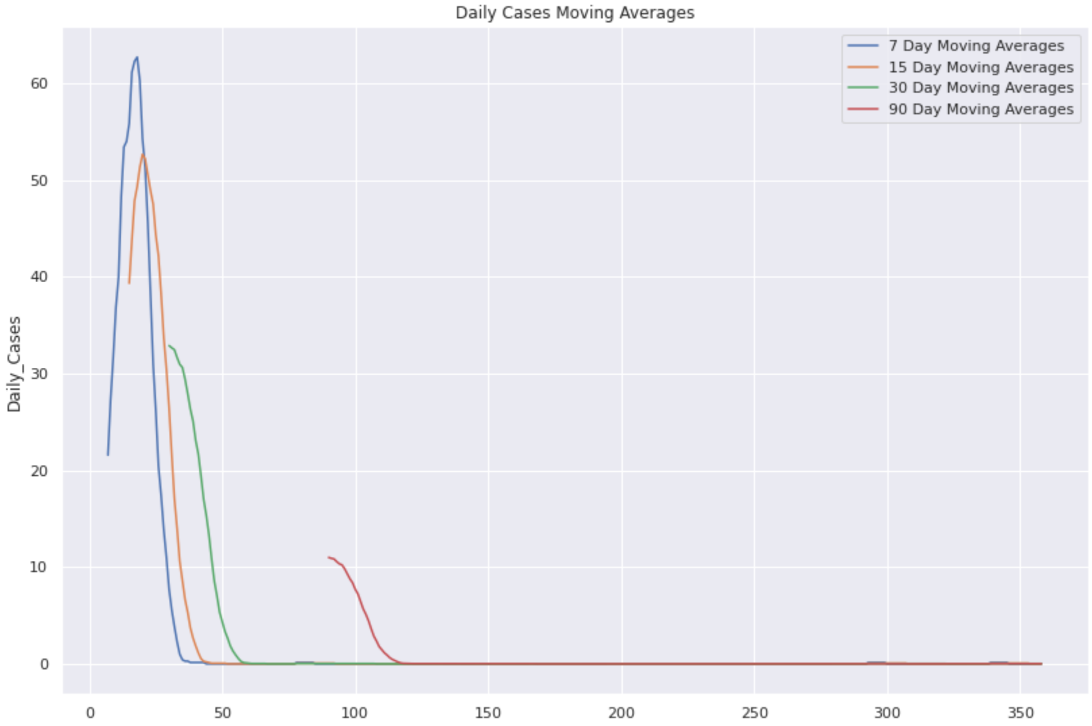

Figure 12. The model predicts COVID-19 pandemic trend in China next 90 days.

4.1.9 Analysis and prediction of COVID-19 trend in Japan

Japan is also one of the world's top ten economies and ranks third. The number of confirmed cases and deaths in this country is not as large as that in other big countries. However, Japan has serious loopholes in the fight against COVID-19, not only in the early stage of the epidemic, the government showed a loose response, worse still, Japan has a bureaucratic system that dare not face failure. All these make Japan the most serious country in Northeast Asia. The development trend of the epidemic in Japan in the next 90 days is shown in the figure below (Figure 13).

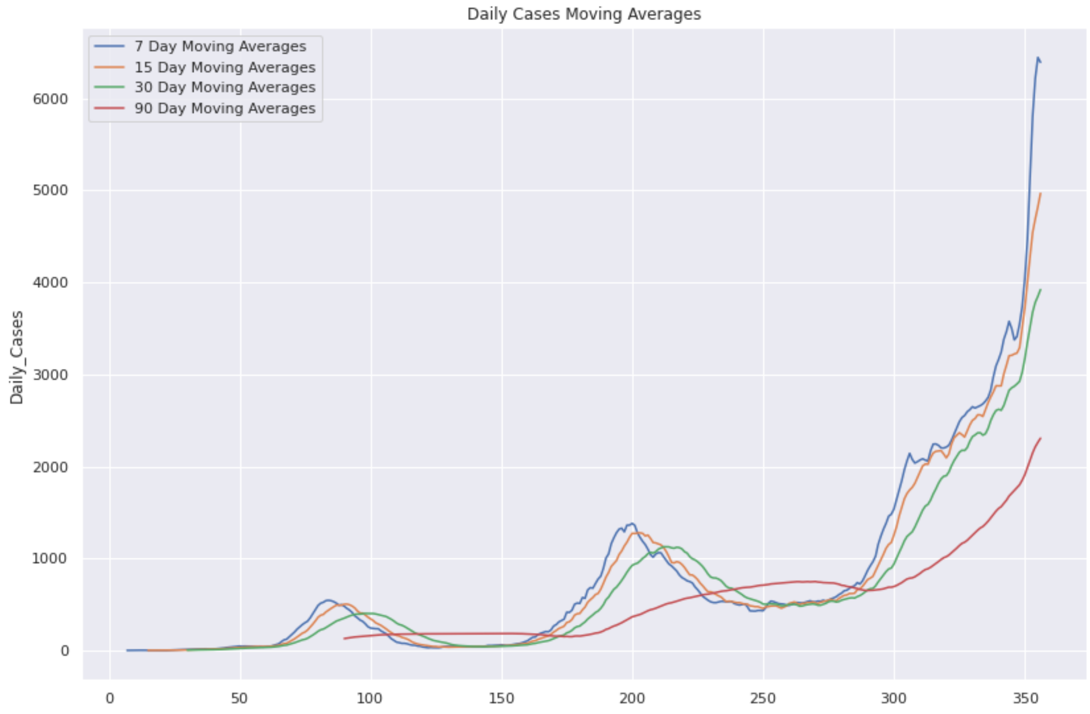

Figure 13. The model predicts the COVID-19 pandemic trend in Japan next 90 days.

4.1.10 Analysis and prediction of COVID-19 trend in the countries with outstanding anti epidemic ability

Australia, New Zealand and Taiwan are the countries that I found in my analysis of the data to be particularly effective in fighting the COVID-19. In particular, in the early stage of the epidemic, Taiwan citizens and the government cooperated to prevent the epidemic. First of all, the government has recruited mask manufacturers to make full use of masks and introduced the policy of purchasing masks under the real-name system. Every citizen can buy a quota of masks every week. Second, strict border control and virus screening measures can effectively prevent and treat the confirmed cases of immigration from abroad. Outstanding protest achievements have made Taiwan frequently appear in the international media, and also caused many governments to carry out protest cooperation plans with Taiwan. For example, New Zealand has ordered a large number of mask production machines like Taiwan. The Australian government also has some strong anti-epidemic measures. First, the borders between provinces should be strictly controlled to effectively prevent asymptomatic infected people from spreading the virus. Second, restrict the entry of non-Australian citizens and non-permanent residents into Australia. As a result, these three countries have achieved outstanding results in epidemic prevention in the world. The development trend of the epidemic in Australia, Taiwan and New Zealand in the next 90 days is shown in the figure below (Figure 14, 15, 16).

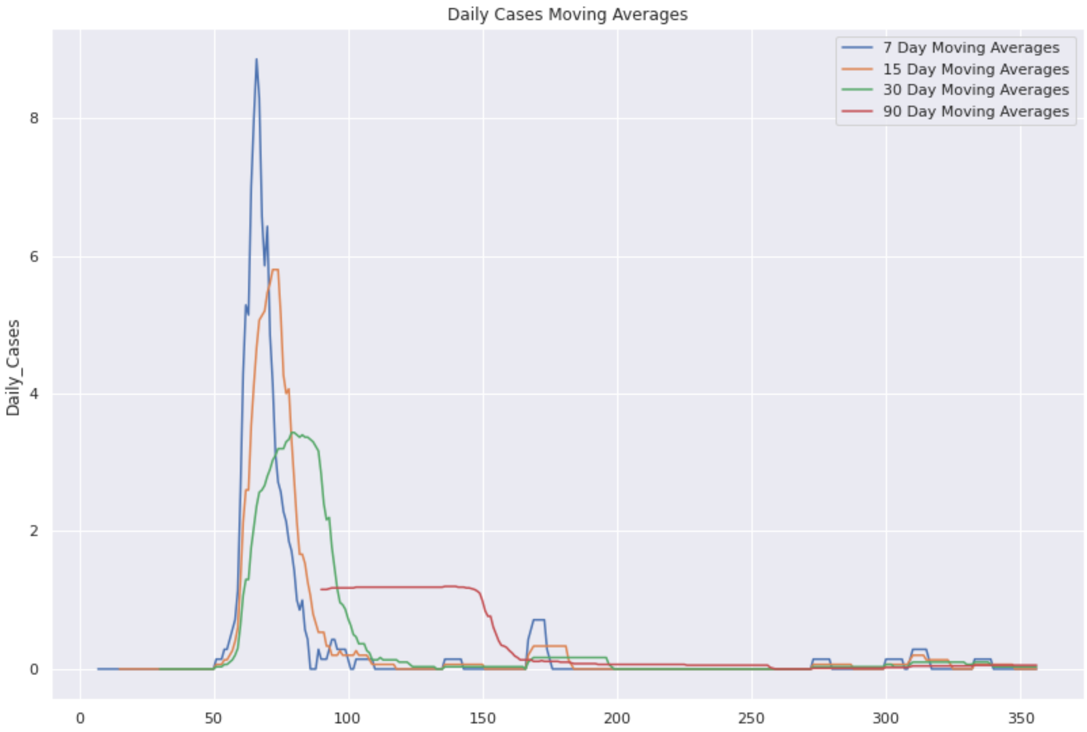

Figure 14. The model predicts COVID-19 pandemic trend in Australia next 90 days.

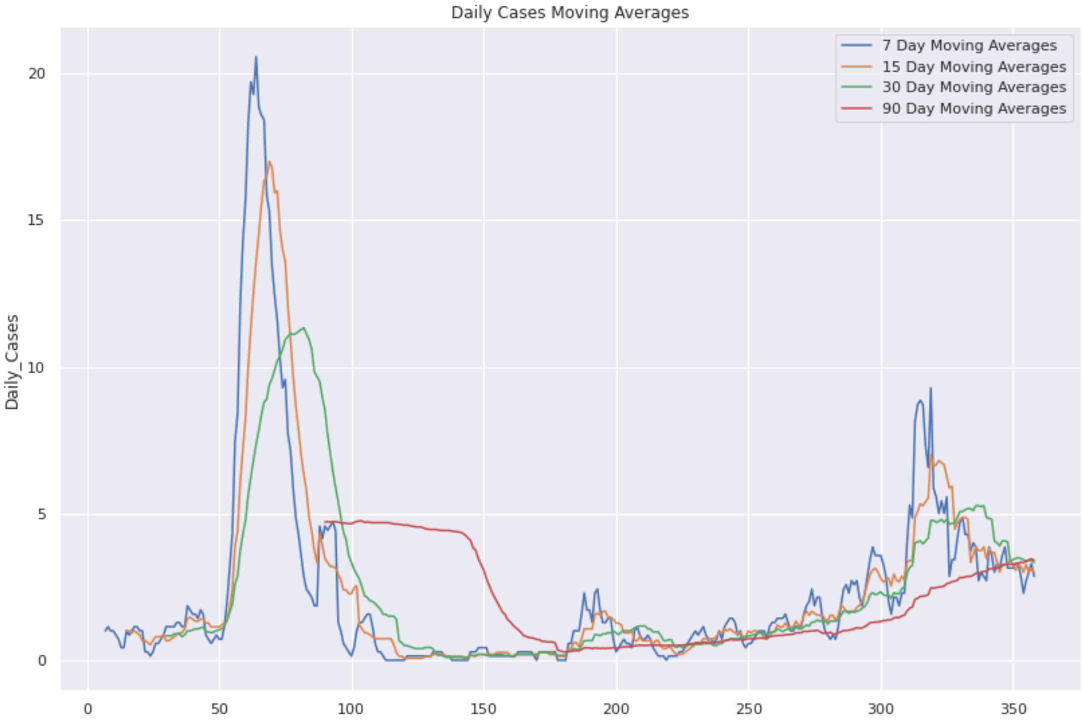

Figure 15. The model predicts the COVID-19 pandemic trend in Taiwan next 90 days.

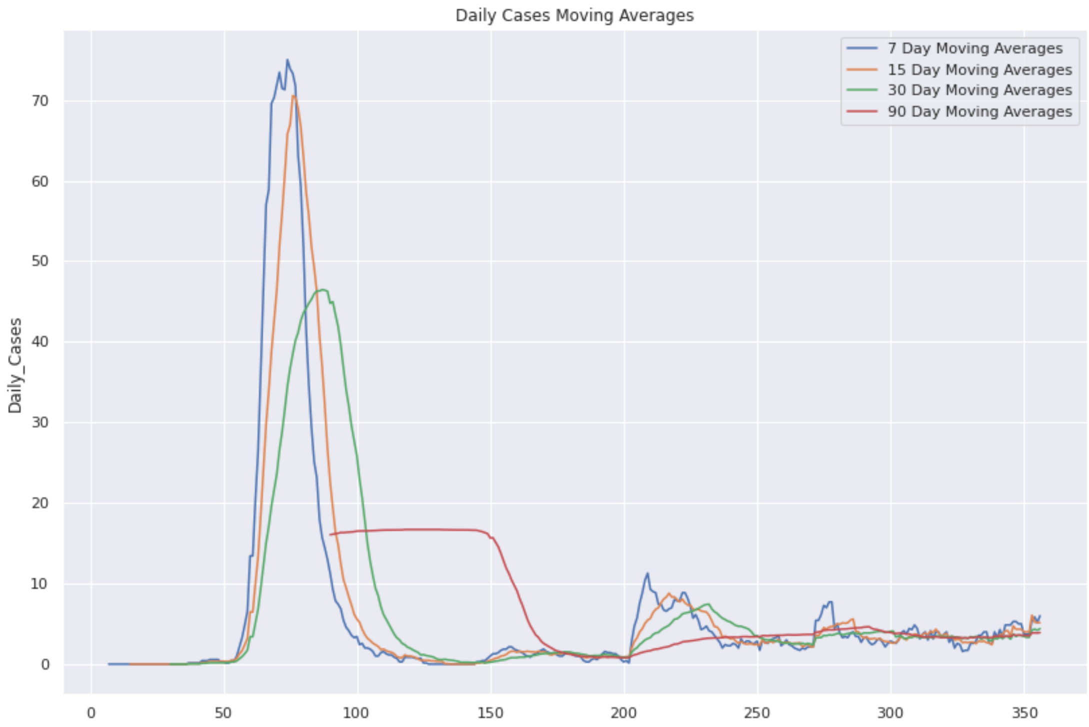

Figure 15. The model predicts the COVID-19 pandemic trend in New Zealand next 90 days.

5. Conclusions

In this study, I analyzed whether the world's top 10 economies and their populations are directly proportional to the severity of the COVID-19 epidemic. I identified the economic, population structure, government regulations and epidemic prevention measures of various countries as the important factors influencing the COVID-19 development trend. I built linear regression models to predict whether and how much the COVID-19 epidemic situation would increase or decline. These models can be very useful in helping the governments and citizens against coronavirus in a number of ways. For example, it could help the government to know the development trend of the COVID-19 ahead of time and make the right decision on whether to tighten or loosen the control measures.

6. Future directions

At present, the coronavirus is still rampant around the world. There was still significant variance that could not be predicted by the models in this study. Hoped that more data scientists will invest in studying virus trends in the future, to provide relevant strategies for governments to effectively combat COVID-19.

7. Data Sources

1. COVID-19 data from John Hopkins University:
   https://github.com/CSSEGISandData/COVID-19
2. World Bank Open Data:
   https://data.worldbank.org/
3. COVID-19 pandemic data on Wikipedia:
   https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data
4. World economy on Wikipedia:
   https://en.wikipedia.org/wiki/World_economy