Analysis SARS-CoV-2 Virus Infection

Group 5.9
STAT1005: Essential skills for undergraduates: Foundations of Data Science
Dr. Lau Sau Mui

December 2022

1 Abstract

The outbreak of Coronavirus Disease 2019 (COVID-19) has posed an unprecedented disruption in almost every part of the world, resulting in over 647 million cases of infection and 6.6 million deaths worldwide by the end of November 2022. Not only it represents a scientific crisis, instead, it can also be viewed as a social crisis, which eventually imposes an economic crisis. One of the most significant unmet needs for the virus is its unpredictable clinical course, as the infection rate can be caused by a multifarious range of unascertainable factors. Machine learning models and data science methods undeniably perform quintessential purposes in scrutinizing significant and symbolic infection factors as well as generating precise predictions on future circumstances by diagnosing COVID-19 cases.

The purpose of our project is to investigate and explore COVID-19 infection from three distinct perspectives: (1) To investigate different sources and causes of infection. (2) To analyse the relationship between the number of daily new COVID-19 cases with seasons and further make predictions on the number of daily new cases. (3) To figure out effective solutions to mitigate the spread of COVID-19. In this study, we implemented classification models such as Logistic Regression models, time series models like ARIMA and SARIMAX models, regression models like Ordinary Least Squares regression (OLS), compartmental models such as the SIR model, and data analysis methods such as graph plotting and correlation analysis.

Our study adds some novelty about the identification of unusual causes of infection, and effective health policies that limit virus spread, which can be further premeditated by policy makers for case diagnosis, formation of preventive policies, and infection trend estimation.

Contents

2 Introduction

The Coronavirus Disease 19 (COVID-19) pandemic since January 2020 has turned into a significant public health crisis that has dramatically affected the health and well-being of the general public and, to some extent, reshaped social habits and ideologies. The COVID-19 pandemic alarms people and demonstrates the critical need for relevant and accurate data sources to inform data-driven perspectives on disease surveillance due to its massive impact on global economies. No matter which infectious disease we are dealing with, to reduce or even eliminate the threat as soon as possible, we need to conduct research on the virus itself, on its transmission, and on ways to prevent and control it. It can be said beyond the shadow of a doubt that the collaboration of data scientists and public health scholars can create an unimaginable and significant impact in terms of identification, analysis, and data modeling, and ultimately produce useful insights. Alarmed by the fifth wave of the epidemic, we conducted our COVID analysis.

In the first part, we analyze the effect of age and the immune system on the infection rate, and the change of the source of infection over time, starting from the disease characteristics and the sources of infection.

In the second part, in the fight against the virus, a macroscopic analysis of the epidemic in the past and a forecast of future trends are needed to better understand and interrupt the spread of the virus. Therefore, time series models need to be constructed. With more complete information and accurate predictions, we have more time to prepare and can deal with the epidemic with more confidence.

In the last part, for more scientific, rational, and cost-effective prevention and control, we focused on various measures that are heavily used, including vaccination. We will evaluate policy effectiveness both at the individual level and aggregate level

3 Literature Review

3.1 Studies on the sources and causes of infection

3.1.1 The relationship between the source of infection and time

(Graph Plotting) By: Li Guocheng

As the Chinese government divides the source of infection of the novel coronavirus into patients infected with the novel coronavirus or asymptomatic infected people, the population is generally susceptible to the virus (Wuhu News Network, 2021). We decided to search for the potential factors that be ignored in such conditions. Some researchers have developed similar models to predict COVID-19 cases. Multiple past studies have shown that the time series models ARIMA and SARIMAX work effectively in predicting COVID-19 and give appropriate results. (Jain et al., 2021). Kiwi already has a wealth of statistics on COVID-19 cases, including almost every country, even some with vaccination rates and cumulative deaths (Centers for Disease Control (CDC), WHO, 2022). But in fact, surprisingly, there is little data to analyze the proportion of COVID-19 cases in most countries and how it changes. Based on the New South Wales (Australia) Government having made available datasets for COVID-19 cases and tests in NSW, we decided to make an assumption to the number of cases in each category of likely source of infection in NSW. This paper will show the quantity of several factors in different periods according to the drawing method, use the best fit line and its concavity properties to count, and finally use the Chi-square test in statistics to verify whether our conjecture is correct.

3.1.2 Test relationship between age and infection of COVID-19

(Regression Model) By: Jiang Qingyi

Efficient analysis of age distribution with cases of infection may control the spread of COVID and help to improve the reaction time of the government to find certain patients. It is important to know which age group of people has a higher possibility of getting the virus so that the government may provide policies targeted to them, in order to lower the risk of infection (Edward Goldstein et al. 2020). As the age distribution for COVID-19 infection is determined, a more accurate estimation of the size of the pandemic can be concluded (Houssein H Ayoub, et al.2020). In 2020, COVID-19 infection was highest in persons within the interval of 20–29 years old, more than 20% of the total cases were contributed by people in this group. Younger adults were likely to be the major part of transmission in the COVID-19 pandemic. Focus on the southern United States in June 2020, increases in the percentage of positive COVID-19 test results among adults aged 20–29 years preceded increases among those aged 60 years by 4–15 days (Boehmer, et al. 2020)

3.1.3 Relation between CBC and respiratory viruses against COVID-19

(Classification Model) By: Yeo Kiah Huah

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has swept the globe and claimed the lives of over 3.4 million people as of May 2021.

According to Palladino (2021), decreased platelet, lymphocyte, haemoglobin, eosinophil, and leukocytes counts as well as increased monocyte and neutrophil count have been associated with COVID-19 infection. Rachael Zimlich also discusses the link between a low white blood cell count and COVID and explains how white blood cells are affected by COVID. On the other hand, an analysis led by Ian Brown had proven that one out of five people with COVID-19 is also infected with other respiratory viruses (Conger, 2020).

The results tend to cast doubt on the notion that COVID-19 is unlikely to affect someone who also has another viral respiratory disease. In light of this, we are also going to investigate the relationship between Coronavirus Disease 19 (COVID-19) infection and Respiratory Viruses. It is undeniable that machine learning plays an indispensable role in COVID-19 prediction and classification problems. A number of academics have been working on machine learning models such as the Logistic Regression model, k-Nearest Neighbors, Decision Trees, and Random Forest in order to solve binary-classification problems in the study of COVID-19 infection. Majumder et al. (2021) have proposed a method to identify whether a patient has a risk of COVID-19 using the Logistic Regression model, considering multiple symptoms like pneumonia, diabetes etc. The study by Jawa, T. M (2022) also uses logistic regression analysis to investigate the impact of home quarantine during the COVID-19 pandemic on the psychological stability of people living in Makkah region in Saudi Arabia. In another study, Assaf, D. et al. (2020) utilized Artificial Neural Network, Random Forest, and Regression Tree to accurately predict the outcome of patients with non-critical COVID-19 based on clinical parameters on admission. Another significant and salient example of the implementation of Artificial Neural Networks, Random Forest, Logistic Regression, and KNN is from China scholars, Li Yan et al. (2020), who analyzed blood samples from Chinese patients for the identification of predictive biomarkers of COVID-19. Their model eventually achieved an accuracy of 93%. Fleitas et al. (2021) also proposed a multivariate Logistic Regression Analysis of 16 main symptoms associated with COVID-19, and their diagnostic characteristics to aid in the clinical diagnosis.

3.2 How infection rate of COVID-19 changes though a time interval

3.2.1 Studies on time and number of infections and prediction

(Time series Model) By: Luo Dongyu

Effective prediction of the daily number of new COVID-19 additions can help control the transmission of COVID-19 and thus reduce its adverse effects. Various time series models have been widely used in the prediction of COVID-19. It is evident from multiple past studies that the time series models ARIMA and SARIMAX operate effectively in the forecast of COVID-19 and give appropriate results. (Jain et al., 2021). Up to now, some authors have made valid predictions for COVID-19. They have used ARIMA models to predict COVID-19 for different regions separately. Mustafa & Fareed (2020) used the ARIMA model to predict COVID-19 for Iran and concluded that the ARIMA (2, 1, 5) model gave the best prediction results. Alzahrani et al. (2020) predicted the spread of COVID-19 in Saudi Arabia using an ARIMA model. And the authors found ARIMA (2, 1, 1) to be the best model for the prediction. Gupta & Pal (2020) performed a trend analysis and forecast of the outbreak of COVID-19 in India using a dataset obtained from Johns Hopkins University and an ARIMA model and determined that ARIMA (1, 1, 2) was the best model. Singh et

al. (2020) found that ARIMA (0, 1, 0) was the best method to predict COVID-19 cases in Malaysia. Some authors consider the ARIMA model the best model for making COVID-19 predictions (Anne & Jeeva, 2020; Sulasikin et al., 2020). Therefore, this paper uses the book Time series analysis by Hamilton (2020) as a statistical theoretical basis and the ARIMA model mentioned by Shumway & Stoffer (2017) as a reference. Meanwhile, this paper will further optimize and enhance the ARIMA model using the SARIMAX model. This section refers to the SARIMAX model Solanki & Singh (2021) mentioned.

3.3 Studies on the solutions to COVID and their effectiveness

3.3.1 The relationship between COVID cases, deaths and vaccinations

(Regression Model) By: Liu Qi

Since the object of this section is to find the relationship between COVID cases, deaths, and vaccinations from different regions, we are going to find the relationship between 4 pairs of data (explained further in Chapter 3), so a linear regression or Ordinary Least Squares regression (OLS) is a good method to achieve this goal. According to Rustagi et al. (2022), two variables are used in linear regression: the dependent variable, which is plotted on the y-axis and is the basis for the prediction, and the independent variable, which is plotted on the x-axis and is used to make the prediction. There are two types of variable-based prediction: univariate (based on one variable) and multivariate (based on several variables). A straight line that depicts the relationship between changes in the independent variable and the dependent variable is known as a regression line. In this case, we would like to let our data fit the univariate-based line, $y = mx + c$ where y is the dependent variable, x is the independent variable, and c is an intercept, m is the slope. The coefficient of determination, or m, is determined using Karl Pearson's coefficient (m). This coefficient shows how many variations the predicted variable can account for.

3.3.2 To identify and analyze the relationship between COVID-19 and vaccination

(Graph Plotting) By: Xu Ziqi

The impact of vaccination on COVID-19 outbreak control will be investigated in this object, in which the impact will be discussed through infection, mortality, and hospitalization. It has been proved that vaccines are one of the most successful and cost-effective interventions to improve health outcomes (B.Roberts et al, 2021), and some people believe that vaccination plays an important and positive role in diminishing the hospitalization rate (C.Huang, 2022). Hence, we look into the relationship between the vaccination and infection rate, death rate, and hospitalization rate, which are closely connected to the situation of COVID-19 outbreak. Therefore, we can prove the positive impact of vaccination on COVID-19 and obtain the motivation to promote it in the foreseeable future.

3.3.3 Find to what extent the health policies impact number of covid infections

(SIR Model) By: Jiang Qingyi

Due to the difference of policies against COVID-19, its essential to identify if the policies are valid and how different combinations will affect virus spread, better attempts can be made during the pandemic. The report shows that the if there is any kind of economic support, limiting number of contacts, and containment policies are significant for countries with dense populations, while lock down, economic, and health policies are significant for countries with populations that have smaller populations. Specific analyses for each country show that health, economic and containment policies differ in importance across waves of the pandemic. (Hye Won Chung et al., 2021). SEIR model is developed to simulate the behavior of COVID-19 under different parameters and predict the future direction. Countries around the world have taken various approaches to shut down the spread of the virus since the beginning of COVID-19 pandemic. The three scholars studied the policies of 10 countries. Some countries have adopted very different approaches and get different achievements. from the study, the comprehensive blockade implemented by the United States and Italy is not as effective as the more targeted measures taken by South Korea, Iceland and other countries. At the end of the analysis, the author puts forward some suggestions to the American government(Kevin Dayaratna et al. 2020)

3.3.4 The effectiveness of stringency index on COVID-19 infection rate

(Regression and classification Model) By: Chen Haodong

In the past few years of the Coronavirus pandemic, most countries have taken strict precautionary measures, including lockdowns, school closures, store closures, social distancing, telecommuting, etc.

These measures were designed to slow the spread of the virus, but there is a lack of quantitative research on their effectiveness, leaving the population with a lack of awareness of their importance.

Therefore, researches on the stringency of various COVID-related policies were conducted. An early study gained insights into the trend slopes in Italy and Spain in daily incident cases and intensive care unit admissions during a second lockdown. It revealed that the slopes were reduced when more restrictive measures for mobility were introduced (Tobías et al., 2020). In addition, the scientists used the SIDARTHE mathematical model to study the number of infections and deaths in Italy in 2020 and found that social control measures, combined with epidemiological strategies to protect susceptible populations, are essential to curb the spread of COVID-19. Moreover, in studying non-pharmaceutical interventions, researchers used the two-dose model to examine the impact of vaccination policies on infection rates and found that it provided efficient resistance to disease and substantial protection against infection (Moore et al., 2021). In addition, using time series analysis, Islam et al. (2020) examined the relationship between interventions to maintain physical distance and the incidence of coronavirus disease. Later, by synthesizing studies similar to previous themes, researchers at Oxford provided a real-time database for all to use based on nine data sets: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. Based on the matrixes of the nine datasets, the OxCGRT model is designed to achieve a comprehensive evaluation of each indicator (Mathieu et al., 2020).

Most of these studies were scattered to examine the role of a single epidemic prevention policy

on infection rates and stopped at data from 2020-2021, with many countries adjusting to their policies in 2022, thus reducing the accuracy of previous studies. In addition, most studies examined only one or two countries and could not estimate the effectiveness of epidemic prevention policies globally. It would be more generalized if more countries were included in the study.

Our study also investigated the relationship between the stringency of each policy and the infection rate using the daily stringency index calculated by the OXCGRT model, previous studies may provide data sources and guides for the study to evaluate the accuracy of the model. Taking into account previous models, we choose to apply both Linear Regression Model and Logistics Regression Model.

4 Data Science Methods

4.1 Studies on the sources and causes of infection

4.1.1 The relationship between the source of infection and time

Data source:
We use data from New South Wales in Australia to conduct this project. The data is from:

- https://data.nsw.gov.au/nsw-covid-19-data

The data for all cases in the file:

- https://data.nsw.gov.au/search/dataset/ds-nsw-ckan-c647a815-5eb7-4df6-8c88-f9c537a4f21e/distribution/dist-nsw-ckan-2f1ba0f3-8c21-4a86-acaf-444be4401a6d/details?q=

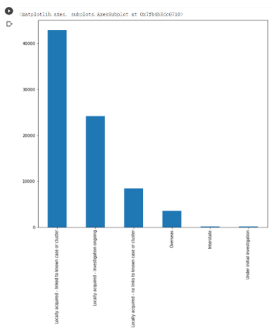
Data pre-processing:
Based on the information provided by NSW, there are five categories of data. We listed them and used a pie chart to present the number of each one. The assumption is that "The proportion of each factor in the total number of cases remains constant." The objective of my project is to test whether the hypothesis is correct.

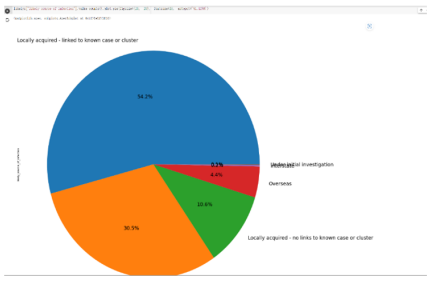
I selected two sources of data to check the number counted by NSW. Use excels to select different numbers of cases with time order and convert them into CSV files (a total of five CSV files).

Methodology
To find and draw a general conclusion about the source of COVID-19 infection, several key influencing factors – time, region and policy – are considered first. Since the latter two are not directly similar in different countries, we decided to use time as a factor in determining COVID-19's likely source and infections in this project; hence we could get a general conclusion.

The graph plotting method can directly show the concavity of best-fit lines, so the relationship between each category can be illustrated without a specific number.

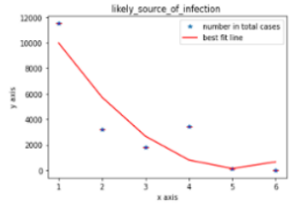
In the first step, we use a histogram and a pie chart to present each category. They can show the overall consideration to the proportion of each category. With the aid of these two graphs,





```
#first 20000 cases of likelysource and best fit line
x20000 = [1, 2, 3, 4, 5, 6]
y20000 = np.array([11504, 3704, 1794, 3404, 94, 6])#y got from first 20000 rows in the table by selecting
z120000 = np.polyfit(x20000, y20000, 2)
p120000 = np.poly1d(z120000)
print(p120000)
yvals=p120000(x20000)
plot1=plt.plot(x20000, y20000, '*',label='number in total cases')
plot2=plt.plot(x20000, yvals, 'r',label='best fit line')
plt.xlabel('x axis')
plt.ylabel('y axis')
plt.legend()
plt.title('likely_source_of_infection ')
dy=np.random.rand(6)
plt.errorbar(x20000, y20000, yerr=dy,fmt='o',ms="1",ecolor='r',color='b',elinewidth=2,capsize=4)
plt.show()
```

```
          2
597 x - 6043 x + 1.545e+04
```

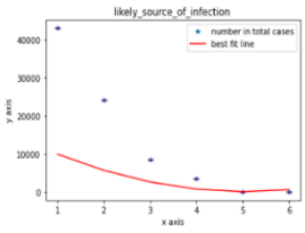


we can make a hypothesis that the most significant factor that contributes to the likely source of infection is “Locally acquired - linked to known case or cluster” and the smallest one is “under initial investigation”

Use matplotlib.pyplot to plot the best-fit line of whole-period data. The graph shows a convex

```
# plot a graph of all cases
xf = [1, 2, 3, 4, 5, 6]
yf=([42930, 24153, 8441, 3513, 150, 92])
z1f = np.polyfit(xf, yf, 2)
p1f = np.poly1d(z1f)
print(p1f)
yvals=p1(xf)
plot1=plt.plot(xf, yf, '*',label='number in total cases')
plot2=plt.plot(x, yvals, 'r',label='best fit line')
plt.xlabel('x axis')
plt.ylabel('y axis')
plt.legend()
plt.title('likely_source_of_infection ')
dy=np.random.rand(6)
plt.errorbar(xf,yf,yerr=dy,fmt='o',ms="1",ecolor='r',color='b',elinewidth=2,capsize=4)
plt.show()
```

```
          2
2553 x - 2.619e+04 x + 6.616e+04
```



function between them and the general trend is decreasing as listed. In the following graphs, we donate the convex line fit to the overall as we assumed, the concave line is not fit the overall.

4.1.2 Test the relationship between age and infection of COVID-19

Data source: US covid patients' age distribution:

- https://covid.cdc.gov/covid-data-tracker/demographics

Data pre-processing:

- The downloaded data set, there contain 11 groups of ages, Compared with other intervals, the population for elderly people takes very small proportion of the data set. In order not to lead to any inaccurate tests, we combined all the people with ages greater than 65 as one group. To simplify the explanatory variable, we ranked interval as 1 to 9(i.e., 1 is 0-4 years old, 2 is 5-11 years old, etc)

Methodology:
In the data set, all the age intervals are tested against count with cases by hypothesis testing to find out if their correlation is statistically significant.

choose to use pearson moment correlation coefficient (PMCC) to test if there is any linear correlation between age interval and infection rate.

denote $PMCC_{x,y} = \rho_{x,y}$

$\rho_{x,y} = \frac{E[(X-\bar{x})(Y-\bar{y})]}{\sigma_X \sigma_Y}$ where:

σ_Y and σ_X are defined as above

- μ_X is the mean of X
- μ_Y is the mean of Y
- E is the expectation.

Since

$$\mu_X = E[X]$$
$$\mu_Y = E[Y]$$
$$\sigma_X^2 = E\left[(X - E[X])^2\right] = E\left[X^2\right] - (E[X])^2$$
$$\sigma_Y^2 = E\left[(Y - E[Y])^2\right] = E\left[Y^2\right] - (E[Y])^2$$
$$E\left[(X - \mu_X)(Y - \mu_Y)\right] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}}.$$

Peason's correlation coefficient does not exist when either σ_X or σ_Y are zero, infinite or undefined.

For r_{xy}:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

```
[ ]  1 import  scipy.stats  as  stats
     2 X  =  scipy.stats.pearsonr(x,  y)
     3 print('PMCC=',X[0],'P-value=',X[1])

     PMCC= 0.6959826442835073 P-value= 0.03731078873370304
```

H_0 : There is no relationship between age interval and the possibility of infect COVID-19

H_1 : There do have relationship between age interval and the possibility of infect COVID-19

```
[ ]  1 if  X[1]<0.05:
     2       print('reject  H0,  there  is  certain  relationship  between  age  interval  and  possibility  of  infect  COVID-19')
     3 else:
     4       print('do  not  reject  H0,  there  is  no  relationship  between  age  interval  and  possibility  of  infect  COVID-19')

     reject H0, there is certain relationship between age interval and possibility of infect COVID-19
```

```
1 import  scipy.stats  as  stats
2 X  =  scipy.stats.pearsonr(x,  y)
3 print('PMCC=',X[0],'P-value=',X[1])
```

```
PMCC= 0.6959826442835073 P-value= 0.03731078873370304
```

After that polynomial fitting is then applied to identify the detailed relationship expressed by certain degree of polynomial, then viewing the curves and loss value to make sure that the function has a good generalization ability and relatively fit the data provided.

4.1.3 To identify and analyze the relationship between COVID-19 and vaccination

Data sources, data cleansing and pre-processing
Data source:

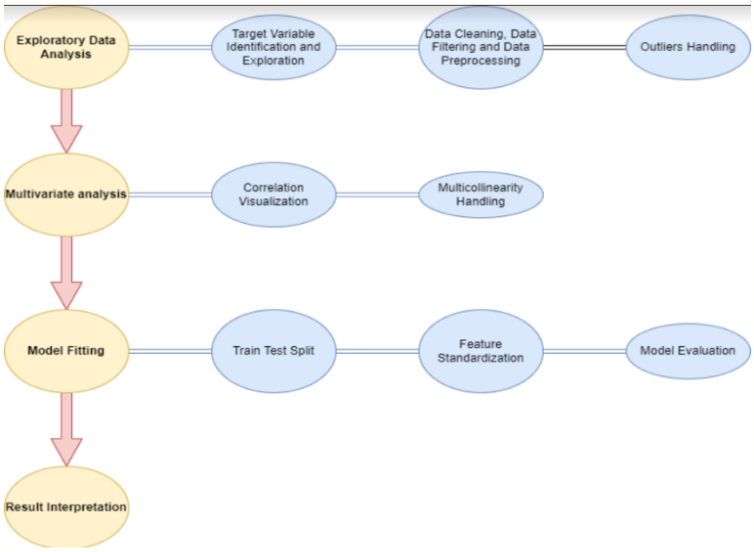
- https://www.kaggle.com/datasets/einsteindata4u/covid19

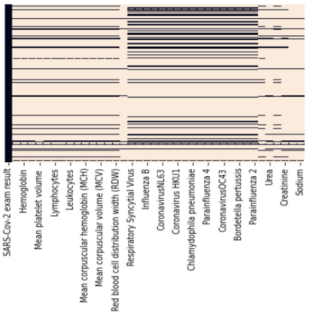
Data Cleaning and Pre-processing:

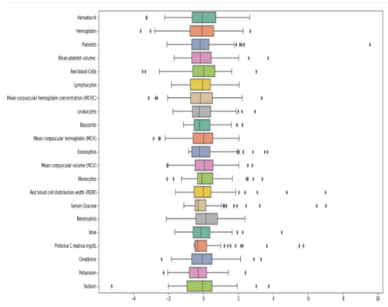
- First and foremost, desired variables and target variable are directly filtered out using the loc() function. In this study, our target variable is the feature "SARS-Cov 2 exam test". In the section of data-cleaning, missing values were visualized using a heatmap and checked using the built-in-function, isnull(). The process is then continued by removing all missing values with dropna() function.

- Then, the process is followed by outliers handling. Outliers can be visualized by plotting box plots. The outliers are then detected by using the Inter-quartile range (IQR) methods and they are eliminated by replacing with the calculated median values.

- Finally, all string binary categorical data needs to be encoded into integer values, 0 and 1.







Models design

We first conduct multivariate analysis by plotting heat map and pair plots to discover the relationships among independent variables in the dataset, then remove one of the variables with correlation >0.9 with each other. The reason of this is because the existence of multicollinearity in the predictors might inflate the standard errors of the regression coefficients and deteriorate the interpretability and reliability of the model.

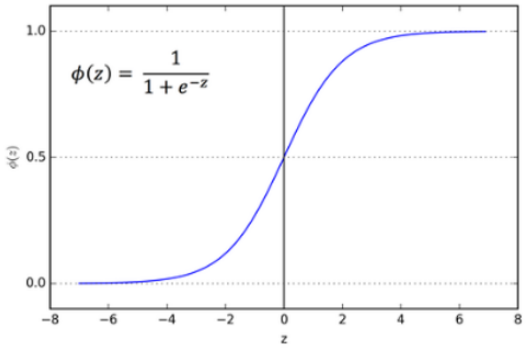
Then, our data is split into training data and test data. Before fitting the data into the logistic regression model, we perform standardization on numerical variables to yield comparable regression coefficients, and our goal is to resize the distribution of values so that the mean of our predictors is 0 and the standard deviation is 0. After standardization, the predictor, Z, which has the largest coefficient is the one that has the most significant impact on our target outcome. After fitting into the model, we can then observe the prediction performance of our model and interpret the results.

```
scaler = StandardScaler()
scaler.fit(ds_train[num])

def scaled(ds, num ,cat, scaler):
    X_numscaled = scaler.transform(ds[num])
    X_cat = ds[cat].to_numpy()
    X = np.hstack((X_cat, X_numscaled))
    y = ds['SARS-Cov-2 exam result']
    return X, y

X, y = scaled(ds_train, num, cat, scaler)
print(X,y)
```

For logistic regression, it solves the following problem:

$$\ln\left[\frac{p}{1-p}\right] = a + BX + e$$

where p is the probability of event Y (SARS-Cov 2 exam test = 1) occurs, while p(1-p) is the log odd ratio, or "logit". Our logistic regression model is simply a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped distribution function which is like the standard normal distribution.



The logit distribution constrains the estimated probabilities to lie between 0 and 1. The instance probability is:

$$p = \frac{1}{1 + \exp(-a - BX)}$$

Variables
We are going to use "SARS-Cov-2 exam result" as our target variable. Predictors or independent variables are separated into two different categories. One of them is test results of other Respiratory Viruses which are all binary categorical variables, another is Complete Blood Count (CBC) Parameters which are all numerical variables.

4.2 How infection rate of COVID-19 changes though a time interval

4.2.1 Studies on time and number of infections and prediction

Data sources, data cleansing and pre-processing
Data source:

- (Form 22/1/2020-21/11/2022, continuously updated) COVID-19 daily

new cases:

- https://ourworldindata.org/covid-cases

Data cleansing and pre-processing:

- In the full covid data set, different COVID-19 indicators (iso code, continent, location, date, total cases, new cases, new cases smoothed, total deaths, new deaths, new deaths smoothed, total cases per million, new cases per million, new cases smoothed per million, total deaths per million, new deaths per million, new deaths smoothed per million, reproduction rate, icu patients, icu patients per million, hosp patients, hosp patients per million, weekly icu admissions, weekly icu admissions per million, weekly hosp admissions, weekly hosp admissions per million, total tests, new tests, total tests per thousand, new tests per thousand, new tests smoothed, new tests smoothed per thousand, positive rate, tests per case, tests units, total vaccinations, people vaccinated, people fully vaccinated, total boosters, new vaccinations, new vaccinations smoothed, total vaccinations per hundred, people vaccinated per hundred, people fully vaccinated per hundred, total boosters per hundred, new vaccinations smoothed per million, new people vaccinated smoothed, new people vaccinated smoothed per hundred, stringency index, population density, median age, aged 65 older, aged 70 older, gdp per capita, extreme poverty, cardiovasc death rate, diabetes prevalence, female smokers, male smokers, handwashing facilities, hospital beds per thousand, life expectancy, human development index, population, excess mortality cumulative absolute, excess mortality cumulative, excess

mortality, excess mortality cumulative per million) are shown. Considering the demand for time series analysis, data and new cases are selected from the data set. And the new cases per day worldwide from 22/01/2020 to 21/11/2022 are selected through the excel method. Then the data and new cases per day worldwide are grouped in a new CSV for further time series analysis. For time series data missing values cannot simply use the whole data mean, median, or mode processing; the most commonly used methods are before and after weighted mean method, linear interpolation method, and n nearest neighbor mean the method of filling, this time using n=2 nearest neighbor mean the method of filling, such as n take 2, then use t-2, t-1, t+1, t+2 moments of the mean to fill the missing t moments of the value, code implementation is as follows.

```
[ ]  def knm(df,n):

            temp = df.isnull().T.any().values
            temp_df = df.copy()
            for i in range(len(temp)):
                    if temp[i] == True:
                            if i < n-1:
                                    temp_df.loc[i,"newcases"] = df.loc[i:i+n,"newcases"].mean()
                            elif i > len(temp) - 1 -n:
                                    temp_df.loc[i,"newcases"] = df.loc[i-n:i,"newcases"]
                            else:
                                    print(df.loc[i-n:i+n,"newcases"])
                                    temp_df.loc[i,"newcases"] = df.loc[i-n:i+n,"newcases"].mean()
                                    print(i-n,i+n+1)
            return temp_df
    not_miss = knm(df[["newcases"]],2)
    df["newcases"] = not_miss.values
```

Raw data:

- https://drive.google.com/file/d/1BhuSXgUNjm-DrZqGOFP3RkHXtauaLlhQ/view?usp=sharing

Processed data:

- https://drive.google.com/file/d/10MJaZo1MLMXurPomPYAT6U1x5ezPAjYn/view?usp=sharing

Models design The first step is data smoothing by drawing the ACF (autocorrelation) and PACF (partial autocorrelation) images separately. After that the objective further run Data Smoothing tests and Unit Root tests to determine the parameter d.

The second step is determining the parameter p and q by using the BIC/AIC index.

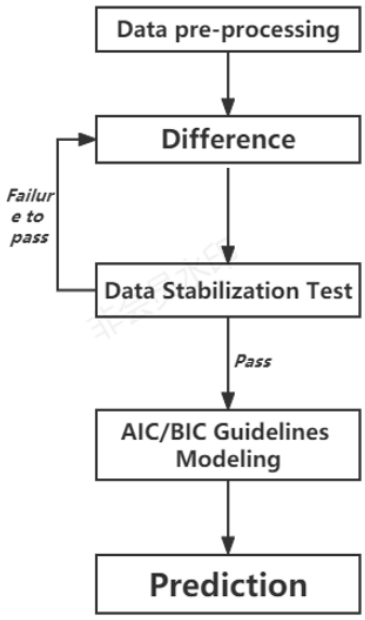
The third step is the model fitting operation. Then do statistical tests to the model. One is Residual Normality Test; another is Residual Series Autocorrelation (whether the residuals are independent).

Afterwards, plot the comparison between the original and predicted data.

After training the model, the final step is the prediction by model ARIMA.

Similarly, considering that the daily new cases of COVID-19 may be seasonality. Therefore, a similar approach is used to construct the SARIMAX model and use the SARIMAX model for prediction.

The abstract modeling process is as follows.



4.3 Studies on the solutions to COVID and their effectiveness

4.3.1 The relationship between COVID cases, deaths and vaccinations

Data sources, data cleansing and pre-processing Data sources:

- WHO COVID-19 data: (viewed on 16/11/2022)
- https://covid19.who.int/WHO-COVID-19-global-table-data.csv
- https://covid19.who.int/who-data/vaccination-data.csv
- https://covid19.who.int/who-data/vaccination-metadata.csv

Data cleaning and pre-processing:

- The above csv files contain information about cumulative cases, deaths and newly reported cases, deaths in different countries, and the vaccinations conditions in different countries, which include total number of vaccinations, number of people fully vaccinated, types of vaccinations used, first time taken vaccinations etc. To make it sample, we combine these files into a new csv file by countries, and then convert it into an excel file for convenient. We also

remove some unused data, like types of vaccinations used and first time taken vaccinations etc. Moreover, the WHO website announced that the data of some Africa countries are incomplete, therefore we also removed data related to Africa countries. Further, we would like to visualize the total number of COVID-19 cases, deaths and number of people vaccinated at least 1 does in different regions, so we group the countries into Americas, Eastern Mediterranean, Europe, South-East Asia and Western Pacific for further processing.

Models Design The data about cases, deaths, vaccinated (at least 1 does) and vaccinated (fully) are selected and compared. So, there are 4 pairs of data in total, they are:

- Cases - Vaccinated (at least 1 does)
- Cases - Vaccinated (fully)
- Deaths - Vaccinated (at least 1 does)
- Deaths - Vaccinated (fully)

These 4 pairs of data are visualized by scatterplots. After that, liner regression lines are added into each graph to show the relationships between corresponding data. Finally, the correlation coefficients (R-values) are calculated and shown.

4.3.2 To identify and analyze the relationship between COVID-19 and vaccination

Data sources, data cleansing and pre-processing Data sources: The data of infected people vaccinated people, deaths and patients:

- https://ourworldindata.org/covid-vaccinations

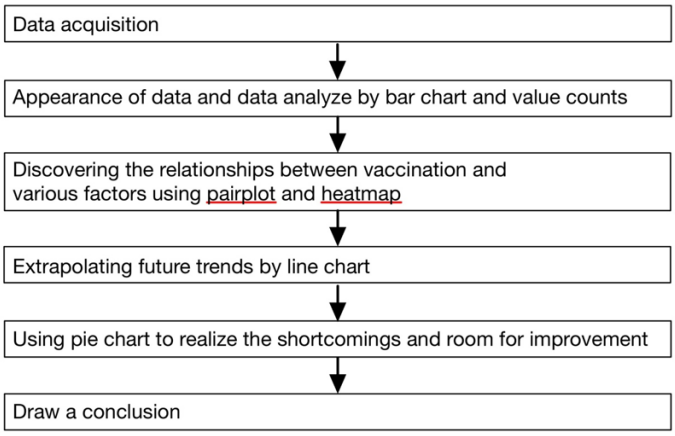
The data of the epidemic situation:

- https://ourworldindata.org/explorers/coronavirus-data-explorer

Data cleaning and pre-processing:

- For the data of infection("total cases"), vaccination ("total vaccinations per hundred"), mortality ("total deaths") and hospitalization ("hosp patients"), the number of it are calculated and the value of it are analyzed.

Models Design
To fine the background information, the data analyzed by value counts and visualized by bar chart. And the module pair plot and heatmap will be made to discover the relationships between vaccination and various factors (infection rate, death rate and hospitalization rate in specific) more professionally. After that, the future trend of COVID-19 epidemic will be extrapolated by line chart. At the end, pie chart will be used to help us realize the shortcomings and room for improvement for current vaccination generalization and outbreak control.



4.3.3 Find to what extent the health policies impact number of covid infection

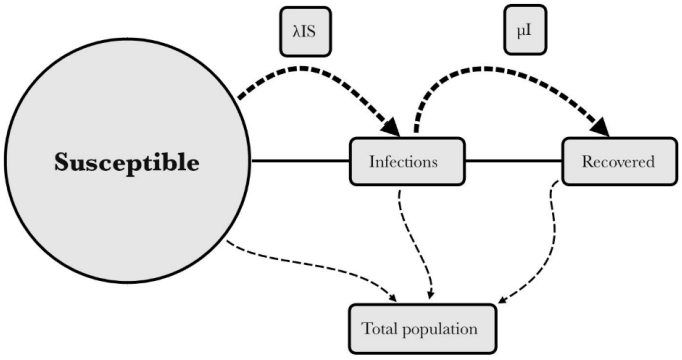
Data sources and data pre-processing Data source:

- https://ourworldindata.org/coronavirus

Data pre-processing:

- In the data set, there are different columns, for example date, total cured, total death, total infected. To fit in the python program, and make it easier for coding, the column 'date' is convert into numbers of days have passed. A column called 'Unnamed:4', in order to make sure that there is no disruption from missing values, that column is then deleted.

Model design
Considering the health policies are categorical data, we think all the policies will affect the cure rate and infection rate, as SIR model contains both cure rate and infection rate. Both rates form a basic reproduction number, which suitable for the situation. The model divides the population within the epidemiological range of infectious diseases into three categories:



S: susceptible, which refers to people who do not have the disease but lack immunity and are susceptible to infection after contact with infected people
I: infectious, which refers to people who have contracted an infectious disease that can be transmitted to members of Class S
R: recovered, which refers to people who are isolated or have immunity due to recovery from the disease.
We have following assumptions:

- This disease can be cured or lead to death
- All the patients after recover will have antibodies and never be infected again.
- The population is fixed
- In a period of time, the ratio between cured and patient is fixed(μ)
- In a period of time, the ratio between infected and population is fixed(λ)

The possibility of being infected exists when a potentially susceptible subject comes into contact with a patient. We assume that each susceptible subject has the same probability of having contact with each patient at any given time, and has some probability of resulting in transmission. Introducing a constant for the probability that contact actually occurs per unit time and leads to transmission is called the infection rate λ, total number of contect is SI.

Thus in time Δt, the expection of infection is:
$SI \times \lambda \Delta t$

The amount of decrease susceptible is:
$\Delta S = -SI \times \lambda \Delta t$

For those infected cases, every period of time,a group of patients whose immune systems are either in full swing or whose medical interventions are effective will recover. Similarly, we introduce a constant probability of recovery per infected patient per unit of time, called the cure rate μ, total number of cured people in a period of time is $I \times \mu \Delta t$ the amount of increased cure number is:
$\Delta R = I \times \mu \Delta t$

For the patient group, there will be some unlucky people who get infected, causing the number of I to rise; but at the same time there will also be some who recover, causing the number of the population to fall.
Taken together, the total patient population changes as:
$\Delta I = \Delta S - \Delta R$

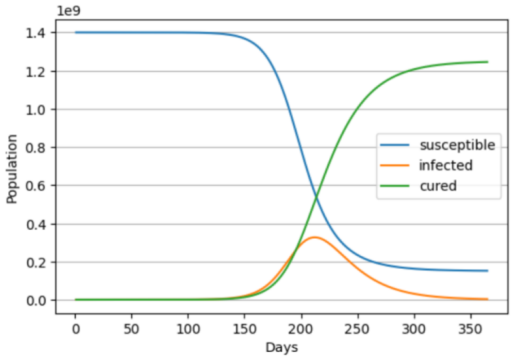
As $\Delta t -> 0$:

$\frac{ds}{dt} = -SI\lambda$

$\frac{dI}{dt} = SI\lambda - \mu I$

$\frac{dR}{dt} = \mu I$

Establish a function with those factors and time, so there exist 3 Partial differential equations By inputting different combinations of cure rate and infection rate (basic reproduction number) This will show how the virus spread out.



4.3.4 The effectiveness of stringency index on the COVID-19 infection rate

Data sources, data cleansing and pre-processing Data source (from 24/01/2020 to 15/11/2022):

- Full COVID-19 dataset: https://github.com/owid/covid-19-data/tree/master/public/data

Stringency index, new cases and total cases:

- https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

Data cleaning and pre-processing:

- In the full covid-19 dataset, the factors associated with the new crown epidemic (new cases, total cases, deaths, hospital admissions, etc.) and the factors associated with the country (location, total population, poverty level, beds per capita) are presented. For the modeling of the second three parts, we specifically chose 'location', 'date', 'total cases', 'new cases', 'stringency index', and 'population' as variables and calculated the total and new infection rates using, 'total cases', 'new cases', 'population'.

Model Design:
In order to understand the approximate level of global preparedness, we first mapped the density distribution of the 'stringency index' globally. We then examined the Q-Q plots of the 'stringency index' and the 'new infection rate' and 'original infection rate' to decide whether to adopt a regression model or a classification model. Besides, to quantify the stringency of policies, this study introduces the Oxford Coronavirus Government Response Tracker (OxCGRT) stringency index model. The model is derived from multiple regressions of school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport;

stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. The accuracy of the parameter was also generally above 70% after the reference adjustment.

$$(6) SI_{legacy} = \frac{1}{7}\left(I_{C1} + I_{C2} + max(I_{C3}, I_{C4}) + I_{C5} + max(I_{C6}, I_{C7}) + I_{C8} + I_{H1}\right)$$

After selecting countries with well-documented data and significantly different government measures of the epidemic: the United Kingdom, the United States and China as representatives. For model selection, scatter plots were first plotted between 'date', 'original infection rate', and 'new infection rate' to observe trends between the individual data. When the rate of change was found to be too large for the China and UK data, we selected the US data and used a Logistic Regression Model to classify strict or loose policies. The Linear Regression targets for regression analysis.

The data obtained from the stringency index model is coherent, while in the Logistic Regression Model, a discrete data set is passed in. So the stringency index needs to be encoded once. This double processing does reduce accuracy. The advantage is that with fewer variables, the observations are straightforward, easy to understand and computationally efficient. However, we will consider a regression model if the accuracy is below 50%.

In the previous scatterplot, we found that there may be a robust linear relationship between the total infection rate and stringency index. We therefore modeled the linear correlation between the 'stringency index' and 'original infection rate', 'stringency index' and 'new infection rate', and used the confusion matrix to check for accuracy

5 Summary and Interpretation

5.1 Studies on the sources and causes of infection

5.1.1 Relation between the source of infection and time

We finally get three convex functions and one concave function. We donate the fit graphs as 1 and unfit one as 0. Our expectation to fit the overall can be written as [1,1,1,1]. Since this, we can carry out a chi-square test to check if the contingency table is good fit for our model. The

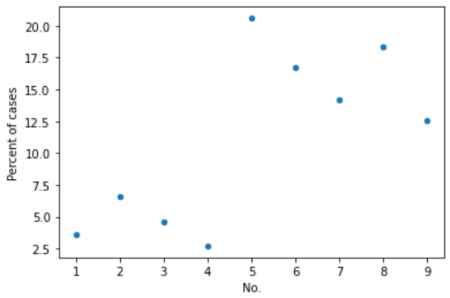
```
from scipy.stats import chi2_contingency

info = [[0, 1, 1, 1], [1, 1, 1, 1]]
print(info)
stat, p, dof, expected = chi2_contingency(info)
print(dof)

significance_level = 0.05
print("p value: " + str(p))
if p <= significance_level:
        print('Reject NULL HYPOTHESIS')
else:
        print('ACCEPT NULL HYPOTHESIS')
```

```
[[0, 1, 1, 1], [1, 1, 1, 1]]
3
p value: 0.831456304592444
ACCEPT NULL HYPOTHESIS
```

chi-square test shows our assumption is correct. This indicates the order of categories from big to small is 1 Locally acquired - linked to known case or cluster 2 Locally acquired - investigation ongoing 3 Locally acquired - no links to known case or cluster 4 overseas 5 interstate 6 under initial investigation. We notice that the trend is not the same as at the beginning it is supposed to be. The peak value of the graph is “overseas”. Therefore, original assumption can be rewritten.

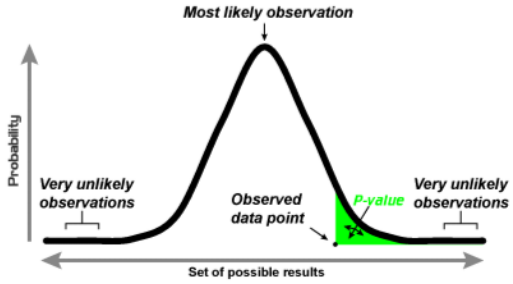
5.1.2 Relation between age and infection of COVID-19

	No.	Age Group	Percent of cases	Count of cases
0	1	0-4 Years	3.6	3,228,143
1	2	5-11 Years	6.6	5,965,126
2	3	12-15 Years	4.6	4,137,206
3	4	16-17 Years	2.7	2,411,405
4	5	18-29 Years	20.6	18,609,230
5	6	30-39 Years	16.7	15,104,529
6	7	40-49 Years	14.2	12,829,647
7	8	50-64 Years	18.4	16,638,745
8	9	65+ Years	12.6	11,364,462



After plotting the graph of age interval against percentage of cases, the scatter diagram is shown. Since we don't know if there is any correlation between those two factors, Pearson moment correlation coefficient (PMCC) and hypothesis testing is then applied.

H_0 : There is no relationship between age interval and the possibility of infect COVID-19
H_1 : There do have relationship between age interval and the possibility of infect COVID-19
$PMCC = 0.696$, $P\ value = 0.037 < 0.05$

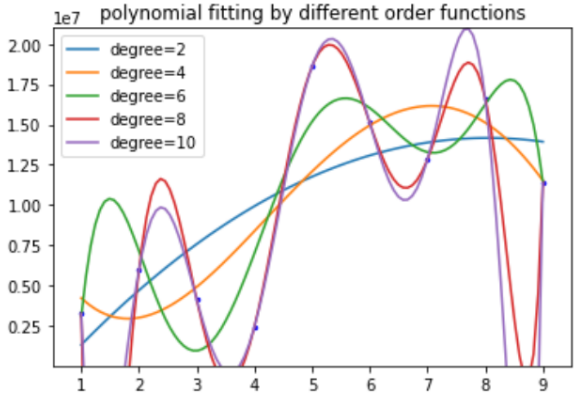


This result implies the result is out of 95% CI, so the correlation is statistically significant.
Then the polynomial fitting for the scatter diagram is used to find with degree have best approximation.
$y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_k x^k = \sum_{i=0}^{k} w_i x^i$

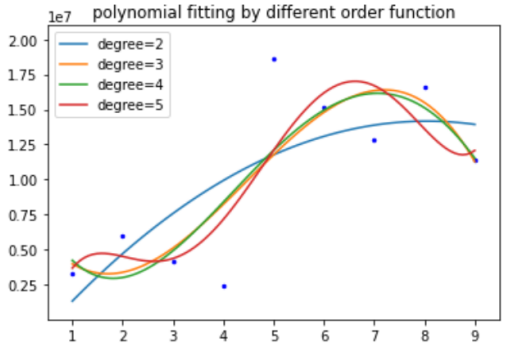
```
x = [1, 2, 3, 4, 5, 6, 7, 8, 9]
y = np.array([3228143, 5965126, 4137206, 2411405, 18609230, 15104529, 12829647, 16638745, 11364462])
x0=np.linspace(1, 9, 100)
def get_model(deg):
        return lambda input_x=x0:np.polyval(np.polyfit(x, y, deg), input_x)
def get_cost(deg, calcu_x, data_y):
        return 0.5*((get_model(deg)(calcu_x)-data_y)**2).sum()
test_set=(2, 4, 6, 8, 10)
for d in test_set:
        print('The loss of degree{} is {}'.format(d, get_cost(d, x, y)))
plt.figure()
plt.title('polynomial fitting by different order functions')
plt.scatter(x, y, c="B", s=6)
plt.xlim(0.5, 9.5)
plt.ylim(1e4, 2.1e7)
for d in test_set:
        plt.plot(x0, get_model(d)(), label='degree={}'.format(d))
plt.legend()
plt.show()
```

```
The loss of degree2 is 69164423282487.15
The loss of degree4 is 50988711826539.3
The loss of degree6 is 22747269944285.73
The loss of degree8 is 2.9845059827948717e-06
The loss of degree10 is 2.3695584161487204e-08
```

In polynomial fitting, the higher degree will show better approximation of data given. "Loss" is used to define how accurate the function is, lower the "Loss" more accurate the fitting is and vice versa.



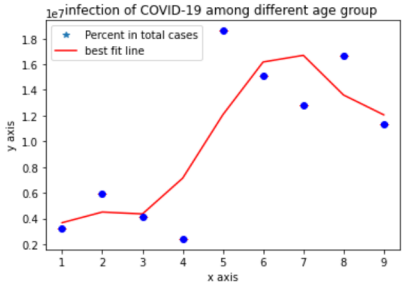
Although the lower generalization ability may lead to more precised model under the data given, when new data set is applied, this model will not be able to fit in and may cause many errors. Thus, to prevent the function from over fitting, the functions with degree more than 6 are rejected.




```
test_set=(2, 3, 4, 5)
for  d  in  test_set:
        print('The  loss  of  degree{}  is  {}'.format(d, get_cost(d, x, y)))

plt.figure()
plt.title('polynomial  fitting  by  different  order  function')

plt.scatter(x, y, c="B", s=6)

plt.xlim(0.5, 9.5)
plt.ylim(1e4, 2.1e7)

for  d  in  test_set:
        plt.plot(x0, get_model(d)(), label='degree={}'.format(d))

plt.legend()
plt.show()
```

```
The loss of degree2 is 69164423282487.15
The loss of degree3 is 51305045808555.53
The loss of degree4 is 50988711826539.3
The loss of degree5 is 46593018488545.07
```

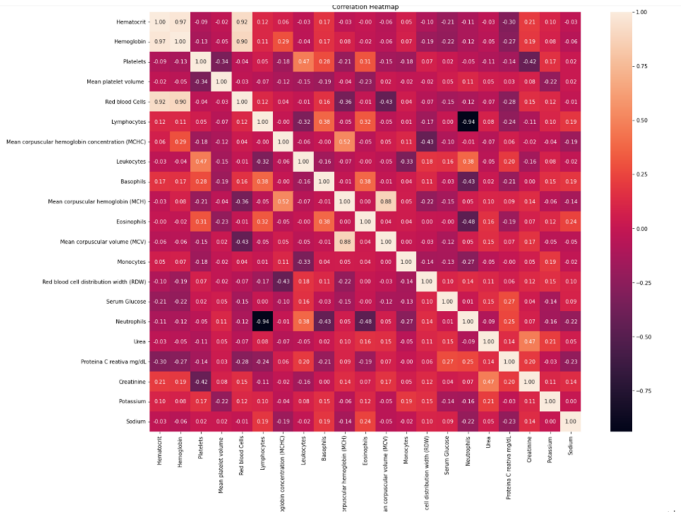
Not only need to pay attention to the generalization ability, also the fitting ability matters. There should have a trade off between those factors. If the model does not show any feature of the data set, although it can fit in any data set, but there is no information given. By viewing functions with lower degrees, degree of 5 shows relatively accurate curve to fit the data.

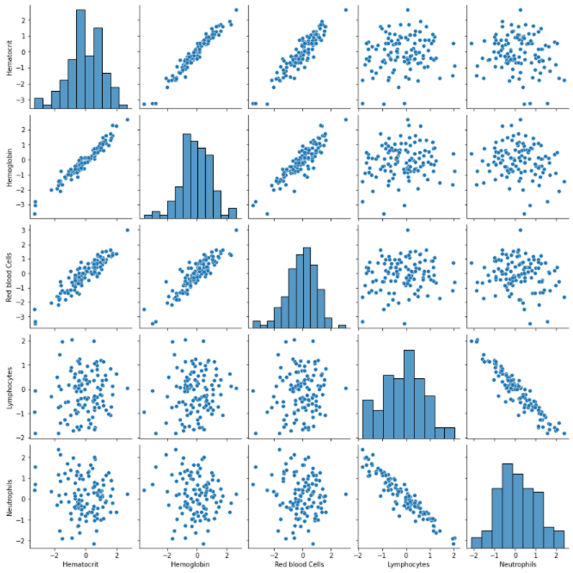


The simulated function is:
$f(x) = 2.056 \times 10^4 x^5 - 5.036 \times 10^5 x^4 + 4.351 \times 10^6 x^3 - 1.586 \times 10^7 x^2 + 2.486 \times 10^7 x - 9.205 \times 10^6$

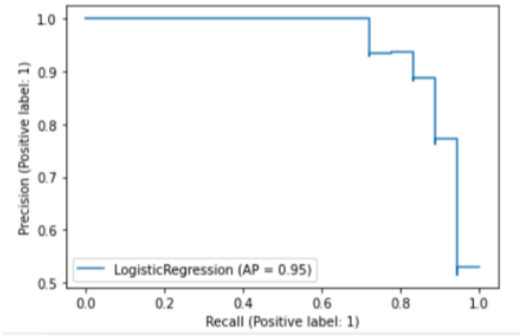
5.1.3 Relation between CBC and respiratory viruses against COVID-19

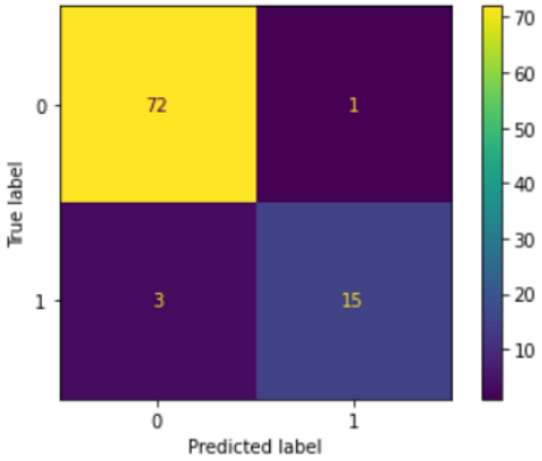
By conduct multivariate analysis with heat map and pair plot, we can find out the correlation between variables. Few pairs of variables have very significant correlations with each other. They are Haemoglobin, Red blood Cells, Hemotocrit, Lymphocytes and Neutrophils.





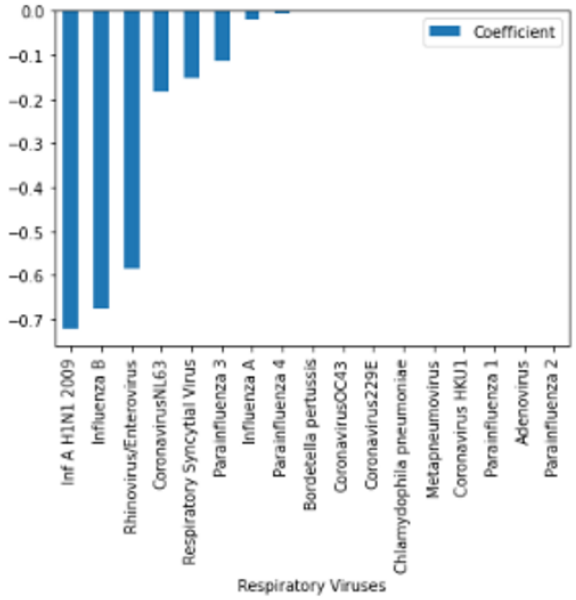
To estimate the generalization accuracy of our model on the future data, we are going to use a series of model evaluation metrics to evaluate the prediction accuracy of the model. ROC Curve is first plotted and AUC Score of our model is 0.98, this indicates our model can predict classes correctly at most of the time. A piece-wise recall curve is also plotted and a larger area under curve indicates a better performance. Precision is a parameter that test accuracy of predictors when the model predicts positive, and the precision score of our model is 0.94. Meanwhile, Recall Score is used when we want to investigate the accuracy where the model predicts negative, and our model has a high recall score of 0.83. Most importantly, our accuracy score is 0.96, it indicates that the model can make very high number of predictions over the total predictions. Also obtain F1 score of 0.88, this indicates a very high preciseness and robustness of out model. Lastly, we use scikit-learn's confusion matrix methods for computing the confusion matrix and classification report.





Given the reality that the model can make predictions with high accuracy and performance, we then proceed to analyze the odd ratios and coefficient, β of the predictors. The logistic regression coefficient β associated with a predictor X is the expected change in log odds of having the outcome per unit change in X. So, increasing the predictor by 1 unit (or going from 1 level to the next) multiplies the odds of having the outcome by $e\beta$. We first interpret coefficient of "Respiratory Viruses" which are all binary categorical data.

	Respiratory Viruses	Coefficient
0	Respiratory Syncytial Virus	-0.152351
1	Influenza A	-0.018557
2	Influenza B	-0.674223
3	Parainfluenza 1	0.000000
4	CoronavirusNL63	-0.184827
5	Rhinovirus/Enterovirus	-0.583133
6	Coronavirus HKU1	0.000000
7	Parainfluenza 3	-0.115506
8	Chlamydophila pneumoniae	0.000000
9	Adenovirus	0.000000
10	Parainfluenza 4	-0.004338
11	Coronavirus229E	0.000000
12	CoronavirusOC43	0.000000
13	Inf A H1N1 2009	-0.722014
14	Bordetella pertussis	0.000000
15	Metapneumovirus	0.000000
16	Parainfluenza 2	0.000000

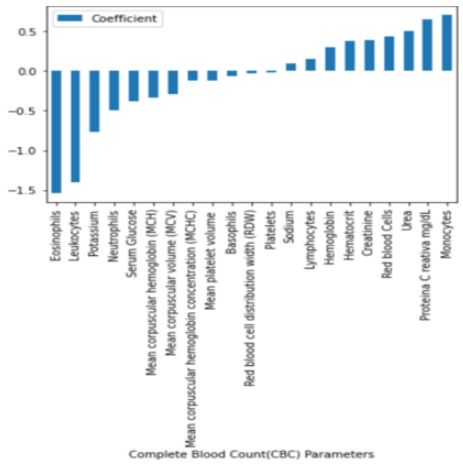


Eventually, our finding is that "Rhinovirus/Enterovirus", Influenza B", "Inf A H1N1 2009" are the three most significant independent variables.

The finding can be interpreted in this way:
an Influenza B carrier has 0.51 times the odds of a non-carrier of being detected positive for SARS-Cov-2 exam test; a Rhinovirus/Enterovirus carrier has 0.56 times the odds of a non-carrier of being detected positive for SARS-Cov-2 exam test; a Inf A H1N1 2009 carrier has 0.49 times the odds of a non-carrier of being detected positive for SARS-Cov-2 exam. Besides, it is also worth noting that variables like "Adenovirus" and "Parainfluenza 2" with 0 coefficient have little or no effect with our target variable.

On the other hand, for numerical variables, remember we have all our numerical variables standardized by StandardScaler, we need to interpret the coefficients in terms of standard deviations. We can visualize the coefficients with a bar chart. It is obvious that "Monocytes", "Proteina C reativa mg/dL" and "Urea" have the most positive coefficients, β. "Leukocytes", "Eosinophils" and "Potassium" have the most negative coefficients, β. "Sodium", "Platelets" and "Red blood cell distribution width (ROW)" have the smallest coefficients among all variables.

	Complete Blood Count(CBC) Parameters	Coefficient
0	Hematocrit	0.371984
1	Hemoglobin	0.300197
2	Platelets	-0.021256
3	Mean platelet volume	-0.123064
4	Red blood Cells	0.430981
5	Lymphocytes	0.150259
6	Mean corpuscular hemoglobin concentration (MCHC)	-0.125904
7	Leukocytes	-1.400785
8	Basophils	-0.058657
9	Mean corpuscular hemoglobin (MCH)	-0.341284
10	Eosinophils	-1.536975
11	Mean corpuscular volume (MCV)	-0.286300
12	Monocytes	0.701722
13	Red blood cell distribution width (RDW)	-0.030026
14	Serum Glucose	-0.386201
15	Neutrophils	-0.493447
16	Urea	0.501174
17	Proteina C reativa mg/dL	0.653861
18	Creatinine	0.385790
19	Potassium	-0.767935
20	Sodium	0.097522



We then convert coefficient into odds ratio by the equation mentioned above.

The findings can be interpreted in this way:

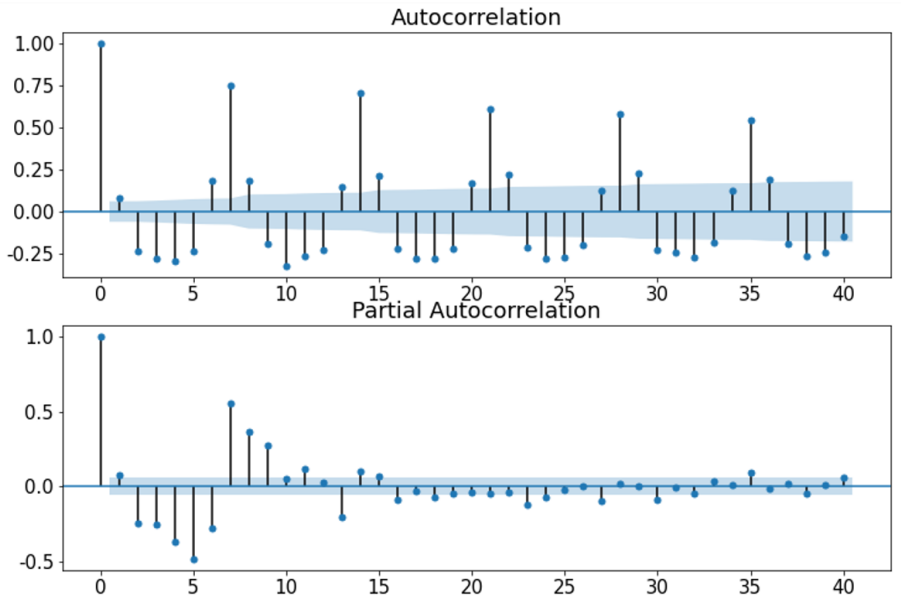
- an increase in 1 standard deviation in **Monocytes** ratio, is associated with **102% increase** in the odds of being detected positive for SARS-Cov-2 test.
- an increase in 1 standard deviation in **Proteina C reativa mg/dL**, is associated with **93% increase** in the odds of being detected positive for SARS-Cov-2 test.
- an increase in 1 standard deviation in **Urea**, is associated with **65% increase** in the odds of being detected positive for SARS-Cov-2 test.
- an increase in 1 standard deviation in **Eosinophils**, is associated with **78% decrease** in the odds of being detected positive for SARS-Cov-2 test.
- an increase in 1 standard deviation in **Leukocytes**, is associated with **75% decrease** in the odds of being detected positive for SARS-Cov-2 test.
- an increase in 1 standard deviation in **Potassium**, is associated with **54% decrease** in the odds of being detected positive for SARS-Cov-2 test.

5.2 How infection rate of COVID-19 changes though a time interval

5.2.1 Studies on time and number of infections and prediction

The results of ARIMA model are as follow (only key results are presented).

the result of drawing the ACF (autocorrelation) and PACF (partial autocorrelation) images separately, both first-order autocorrelation and partial correlation plots show the characteristics of censored, and it is not possible to determine p and q from these two plots.



The results of the data smoothing test and the unit root test are as follows.

```
[ ] from statsmodels.tsa.stattools import adfuller as ADF
    print(u'The results of the ADF test for the original series are.', ADF(df["newcases"]))

    The results of the ADF test for the original series are. (-3.8060230480309123, 0.002846289003967209 6, 16, 1018, {'1%': -3.436789946800891 6, '5%': -2.8643833180472744, '10%': -2.568283908970553

[ ] from statsmodels.tsa.stattools import adfuller as ADF
    print(u'The results of the ADF test for the first-order difference series are.', ADF(df["diff_1"][1:]))

    The results of the ADF test for the first-order difference series are. (-6.4847475092083515, 1.2690844655835432e-08, 22, 1011, {'1%': -3.436834649927693, '5%': -2.864403033735098, '10%': -2.56
```

The p-value of the first-order difference unit root test and the p-value of the original series were both less than 0.05, so the parameter d in ARIMA is set to 1.

By using the BIC/AIC index, the result of the parameter p and q are 5 and 4 separately. In summary, we determined that the model fitting ARIMA (5, 1, 4).

```
pmax = 5
qmax = 5
bic_matrix = []  #bic matrix
for p in range(pmax+1):
    tmp = []
    for q in range(qmax+1):  #There is a partial error, so use try to skip the error.
        try:
            tmp.append(ARIMA(df["newcases"], order=(p,1,q)).fit().bic)
        except:
            tmp.append(None)
    bic_matrix.append(tmp)
bic_matrix = pd.DataFrame(bic_matrix)  # from which the minimum value can be found
p,q = bic_matrix.stack().idxmin()
print(u'The minimum p and q values of BIC are: %s、%s' %(p,q))
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/arima_model.py:472: FutureWarning:
statsmodels.tsa.arima_model.ARMA and statsmodels.tsa.arima_model.ARIMA have
been deprecated in favor of statsmodels.tsa.arima.model.ARIMA (note the .
between arima and model) and
statsmodels.tsa.SARIMAX.  These will be removed after the 0.12 release.

statsmodels.tsa.arima.model.ARIMA makes use of the statespace framework and
is both well tested and maintained.

To silence this warning and continue using ARMA and ARIMA until they are
removed, use:

import warnings
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARMA',
                        FutureWarning)
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARIMA',
                        FutureWarning)

  warnings.warn(ARIMA_DEPRECATION_WARN, FutureWarning)
The minimum p and q values of BIC are: 5、4
```

Minimum p and q of BIC shows above.

```
pmax = 5
qmax = 5
aic_matrix = []  #bic matrix
for p in range(pmax+1):
    tmp = []
    for q in range(qmax+1):  #There is a partial error, so use try to skip the error.
        try:
            tmp.append(ARIMA(df["newcases"], order=(p,1,q)).fit().aic)
        except:
            tmp.append(None)
    aic_matrix.append(tmp)
aic_matrix = pd.DataFrame(bic_matrix)  # from which the minimum value can be found
p,q = bic_matrix.stack().idxmin()
print(u'The minimum p and q values of AIC are: %s、%s' %(p,q))
```

```
The minimum p and q values of AIC are: 5、4
```

Minimum p and q of AIC shows above, the result of modeling fitting is as follows.

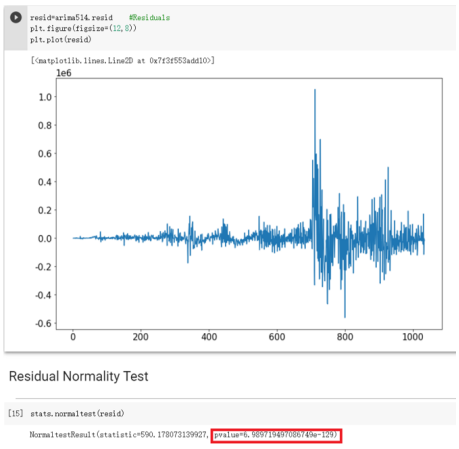
```
arima514 = ARIMA(df["newcases"],order=(5,1,4)).fit()
arima514.summary2()
```

Model:	ARIMA	BIC:	26875.2064
Dependent Variable:	D.newcases	Log-Likelihood:	-13399.
Date:	2022-11-23 19:09	Scale:	1.0000
No. Observations:	1034	Method:	css-mle
Df Model:	10	Sample:	1
Df Residuals:	1024		5
Converged:	1.0000	S.D. of innovations:	102444.399
No. Iterations:	48.0000	HQIC:	26841.478
AIC:	26820.8533		

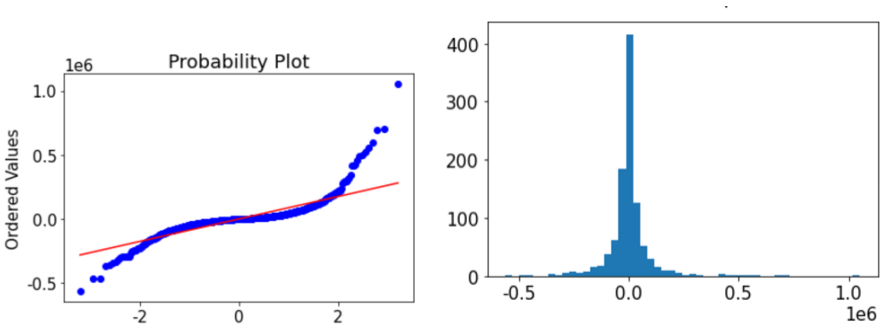
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	343.7012	1589.8501	0.2162	0.8288	-2772.3477	3459.7500
ar.L1.D.newcases	0.3528	0.0413	8.5474	0.0000	0.2719	0.4337
ar.L2.D.newcases	-1.0465	0.0328	-31.8851	0.0000	-1.1108	-0.9822
ar.L3.D.newcases	0.1159	0.0554	2.0912	0.0365	0.0073	0.2245
ar.L4.D.newcases	-0.5853	0.0292	-20.0611	0.0000	-0.6424	-0.5281
ar.L5.D.newcases	-0.4520	0.0367	-12.3184	0.0000	-0.5239	-0.3801
ma.L1.D.newcases	-0.7981	0.0421	-18.9751	0.0000	-0.8805	-0.7156
ma.L2.D.newcases	1.1580	0.0374	30.9625	0.0000	1.0847	1.2313
ma.L3.D.newcases	-0.6450	0.0437	-14.7767	0.0000	-0.7306	-0.5595
ma.L4.D.newcases	0.5890	0.0282	20.8856	0.0000	0.5338	0.6443

	Real	Imaginary	Modulus	Frequency
AR.1	0.6299	-0.7844	1.0060	-0.1423
AR.2	0.6299	0.7844	1.0060	0.1423
AR.3	-0.2167	-0.9918	1.0151	-0.2842
AR.4	-0.2167	0.9918	1.0151	0.2842
AR.5	-2.1213	-0.0000	2.1213	-0.5000

By residual normality test the objective obtains the p-value¡0.05, so the objective accept the alternative hypothesis that the residuals are normal.



H_0 : Not consistent with normal distribution
H_1 : Consistent with normal distribution
Also, can accept the alternative hypothesis that the residuals are normal by observing the Q-Q Plot scatter is basically on a straight line, and the histogram is also normal.

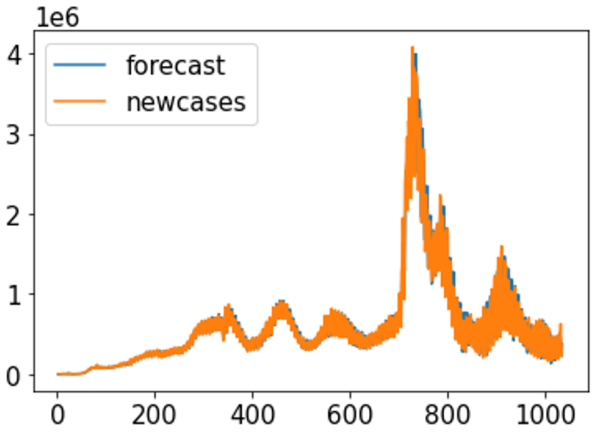


The result of Residual Series Autocorrelation, the DW value is very close to 2, indicating that the series is not correlated.

```
from statsmodels.stats.stattools import durbin_watson
durbin_watson(arima514.resid.values)          #DW test: near 2 - normal; near 0 - positive autocorrelation; near 4 - negative autocorrelation
```

```
2.0085085767199637
```

The two tests above show that the model is reasonable. Afterwards, plot the comparison between the original and predicted data.



The images reveal that the original data largely overlap with the predicted data. This is followed by a training step for the model.

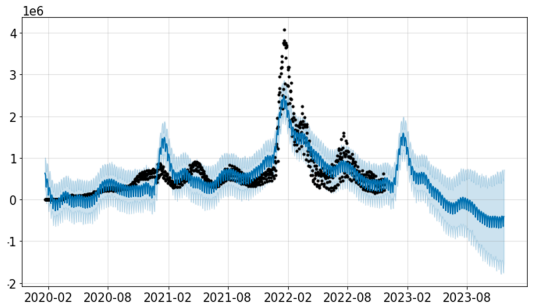
```
# Check the length of the dataset to see how much data we want to split between the test and train set
len(df)
```

```
1035
```

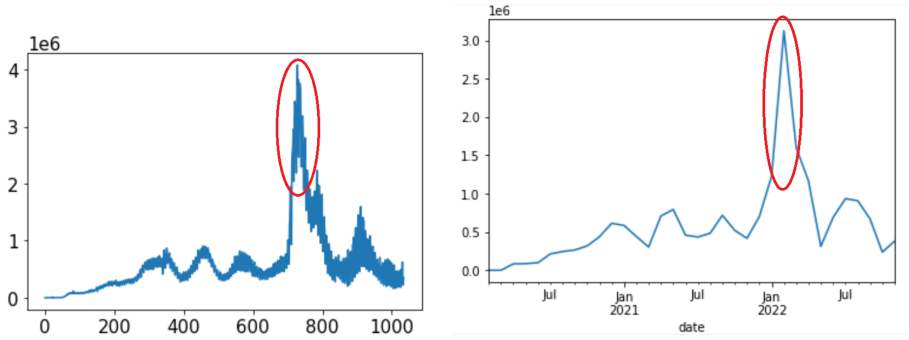
```
# Split the data amongst training and test sets
train = df.iloc[:1027]
test = df.iloc[1027:]

# define start and end variables for .predictstart = len(train)
start = len(train)
end = len(train) + len(test) -1
predictions = results.predict(start=start,end=end,typ='levels').rename('ARIMA Predictions')
```

Finally, make the prediction by model ARIMA (5, 1, 4)

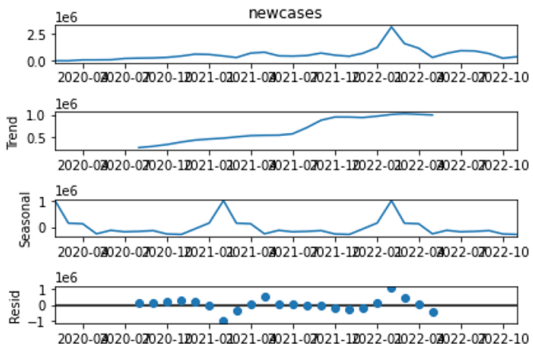
```
predictions = results.predict(start=train.shape[0],end=(train.shape[0]+test.shape[0]-1), dynamic=False).rename('ARIMA FORECAST')
df['newcases'].plot(legend=True,figsize=(14,8))
predictions.plot(legend=True)
```



The results of the SARIMAX model are as above (omit some results similar to the ARIMA model and keep only the key results), these graph shows the result of peaks appearing in the data



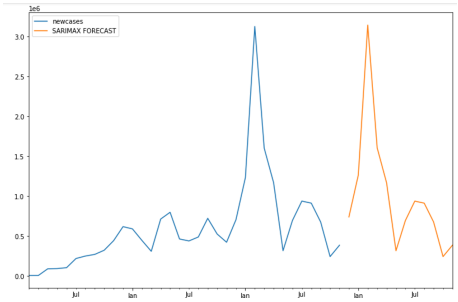
The results of the grouping are as follows.



The result of error is as follow

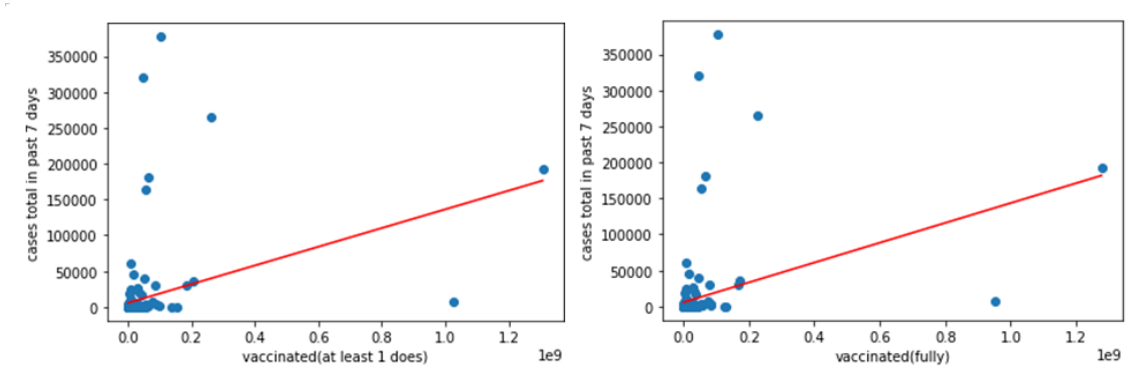
```
error = rmse(test['newcases'],predictions)
error
```

```
115939.40872191709
```

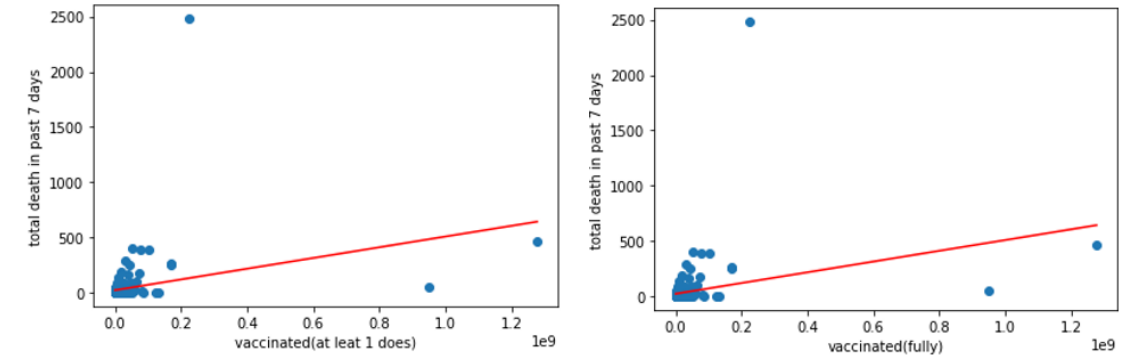


5.3 Studies on the solutions to COVID and their effectiveness

5.3.1 The relation between COVID cases, deaths and vaccination




```
intercept: 15524312.14165673          intercept: 13912623.874437854
slope: [931.04573683]                 slope: [891.66304937]
r value:  0.8134035027365248          r value: 0.905190081845558
```




```
intercept: 17552497.289977666         intercept: 16124421.921014994
slope: [199053.98424847]              slope: [181973.46817711]
r value: 0.6939757216975355           r value: 0.6538537168258398
```

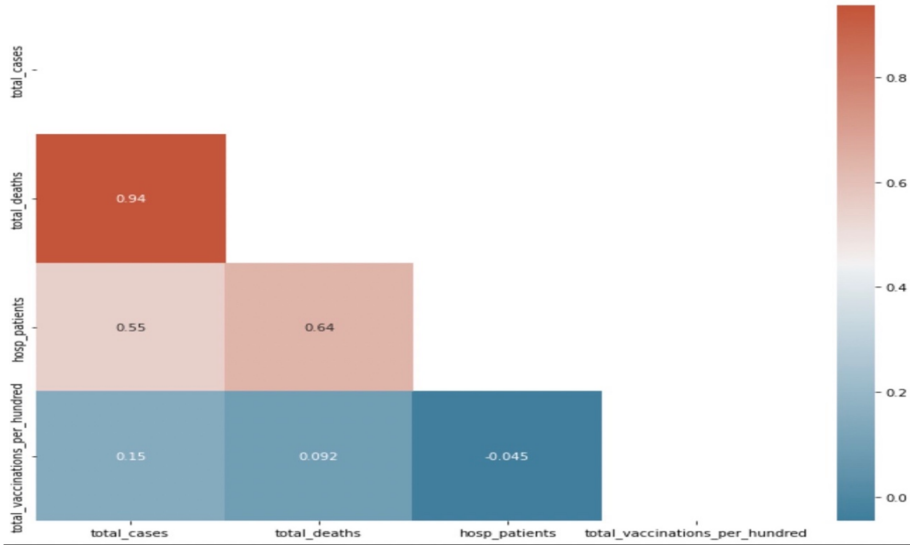
These scatter plots and linear regression lines show a positive relationship between number of people vaccinated at least 1 does/fully and cases/deaths in pat 7 days.

From the correlation coefficients (R-values), we can see that cases in last 7 days have a strong positive correlation to vaccinated at least 1 does/fully (0.81, 0.91), while deaths have a weaker correlation to vaccinated at least 1 does/fully (0.69, 0.65).

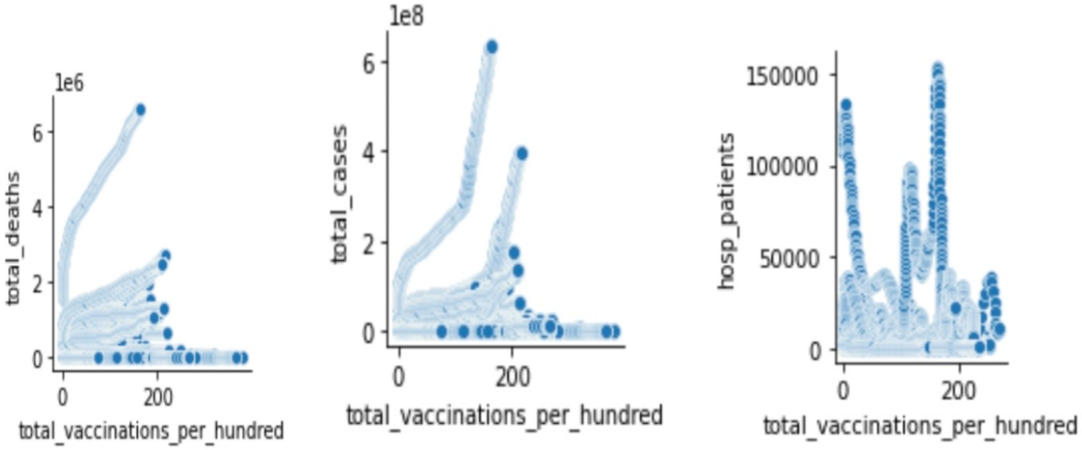
From the gradient of the regression lines, the gradients of vaccinated fully is smaller than the

gradients of vaccinated at least 1 does, so we conclude that taking fully vaccinated have a big impact on reducing number of cases and deaths and it is essential.

5.3.2 Identify and analyze the relation between COVID-19 and vaccination



The value between each variable is determined using python and visualized using heat map. Each square shows the correlation between the variables on each axis which ranging from -1 to +1. The negative value between vaccination and hospitalization("hosp patients") indicate that vaccination is negative correlated with hospitalization. Hence, we can prove that the promoting vaccination can benefit to diminish the possibility of becoming severe patients and decreasing hospitalization rate.



The pairs plot builds on two basic figures, the histogram and the scatter plot. The above scatter

plots on the upper and lower triangles show the relationship between two variables. Under a rule of correspondence:

- Graph 1: higher number of vaccination ("total vaccinations per hundred") set assigned to lower number of deaths("total deaths"). It provides a comprehensive expression to the negative correlation between vaccination and deaths. Specifically, it indicates that higher rate of vaccination may help to decrease the mortality.

- Graph 2: higher number of vaccinations set assigned to lower number of infections("total cases"). We can learn the negative correlation between vaccination and infection. However, according to the heat map shows above, the vaccination is positive correlated with infection in a low value. Combining two graph and module, it reflects that the vaccination is not an influential factor in reducing infection rates but can lead to a positive indirect impact to infected people.

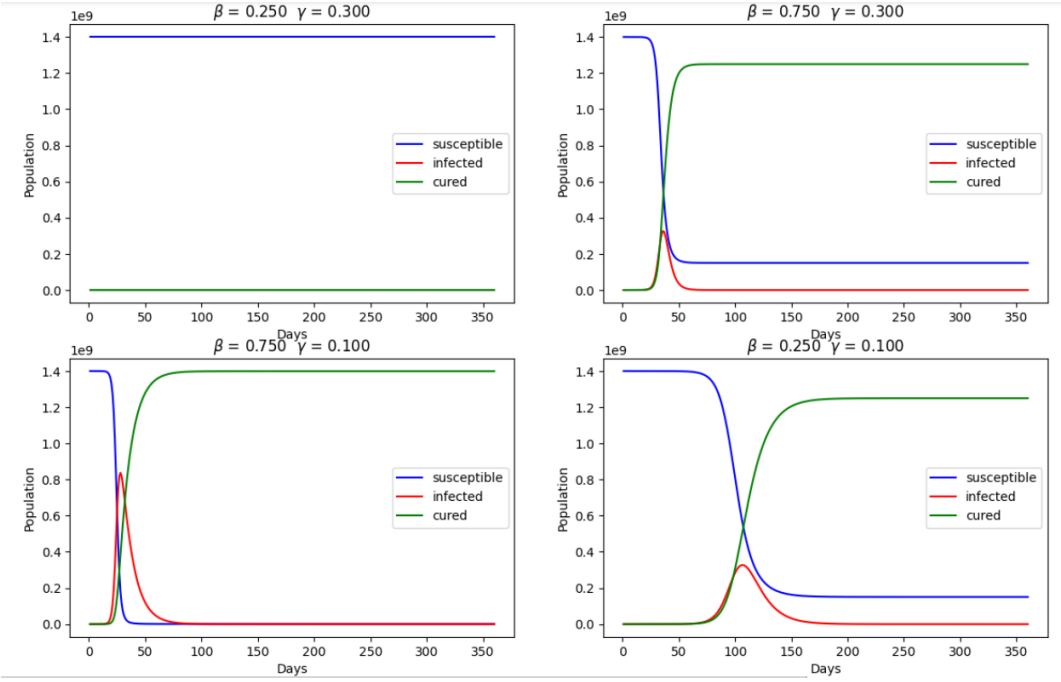
- Graph 3: higher number of vaccinations set assigned with lower number of patients in hospital("hosp patients"). It shows the negative correlation between these two variables. After combining the above heat map, we can effectively prove that vaccination bring a positive impact to reducing the likelihood of the disease becoming severe. In other words, relieving pressure on the social health care system.

5.3.3 How health policies impact number of COVID-19 infection

λ : infection rate, μ : cure rate, $R_0 = \frac{\lambda}{\mu}$: basic reproduction sumber

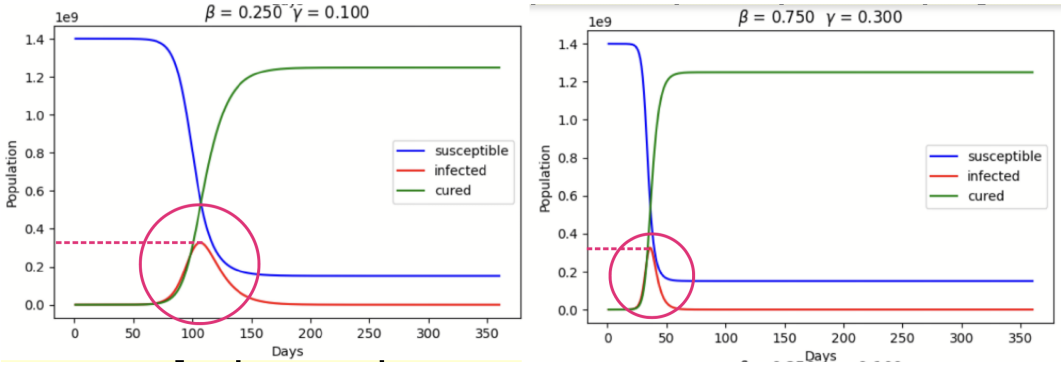
index	λ	μ	R0	feature
0	0.25	0.30	1.25	wearing masks, government pay for medical expenses
1	0.750	0.30	2.50	not wearing masks, government pay for medical expenses
2	0.750	0.10	5.00	not wearing masks, government don't pay for medical expenses
3	0.25	0.10	2.50	wearing masks, government don't pay for edical expenses

SIR model shows the different combinations of policies and how will the spread of COVID-19 have been affected with no other interruption. We pretend not wearing mask will lead to increase of infection rate cause the virus may spread via droplet and government pay for medical expenses will motivate patients to go hospital, there are more people recovered.



The graphs show the number of infection, cured, susceptible people under different policies.When the government developed strict policies, less chance for the virus infect susceptible people.The peak of infection line will then lower and takes longer time to rich the peak As more subsidy provided to lower down the stress from finical side, those who are infected are more willing to go hospital, thus cure line on the graph will rapidly increase till the peak value. By the contraction of graphs, there is huge difference in infection number based on different health policies

- Graph 1: All the citizens wear masks, and once they are sick, all of them will immediately go to hospital to stop the separation of virus. As a result, there is no pandemic at all.
- Graph 2: No citizen are asked to wear mask but when they are sick, government will pay for their medical expenses. This situation caused a short period of pandemic with medium size.
- Graph 3: Citizens are neither wear masks nor go to hospital for illness, since government choose not to pay for the cost of medicine. This region then caught in a vicious cycle, people do not go to hospital for treatment, they will sick for longer time until death or recover. During this period, virus is spread in larger area, more susceptible are infected, more people choose not to go to hospital, etc. The result is all the citizens are infected by COVID-19, the epidemic will last till Herd immunization.
- Graph4: This situation is similar to Graph 2, people do not wear masks, so the virus spread out, but as they all go to hospital for treatment, cure number increased as well. By looking at the time, epidemic starts at about 70th day, which is much late than previous cases.



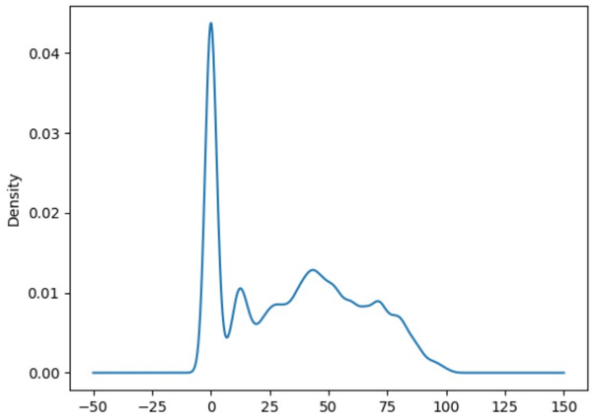
Notably, even though the two graphs above have same R0, there is still have a different outcomes of COVID-19 infection. This represent that different policies will not only lead to difference in number of infection, but also the start time and duration.

So different combination of policies may lead to significant difference of how COVID spread in many aspect:

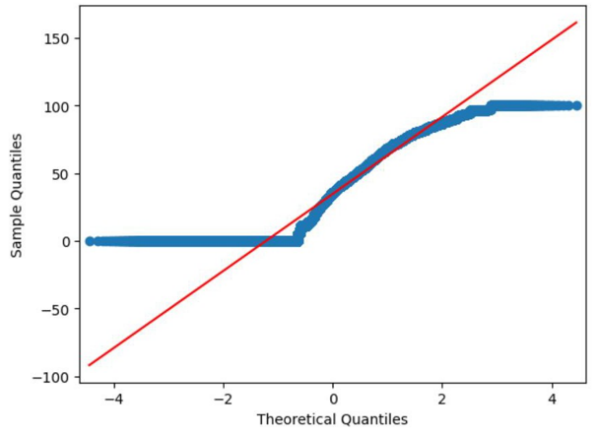
- The duration of COVID depends on the proportion of citizens take treatment. As more people to go hospital for medical treatment on time, there would be less sources of infection, those people who choose to go to hospital will quickly being cured.That lead the epidemic quickly being limited.
- The peak of a pandemic may be decided by wearing mask or not with influence of confounding variables. Insofar as face mask can decrease the possibility of infection from patient to healthy people. Then the total infection number will decrease.

With combining different policies, the peak of pandemic tend to be determined by R_0. Different strategy will lead to difference in duration.

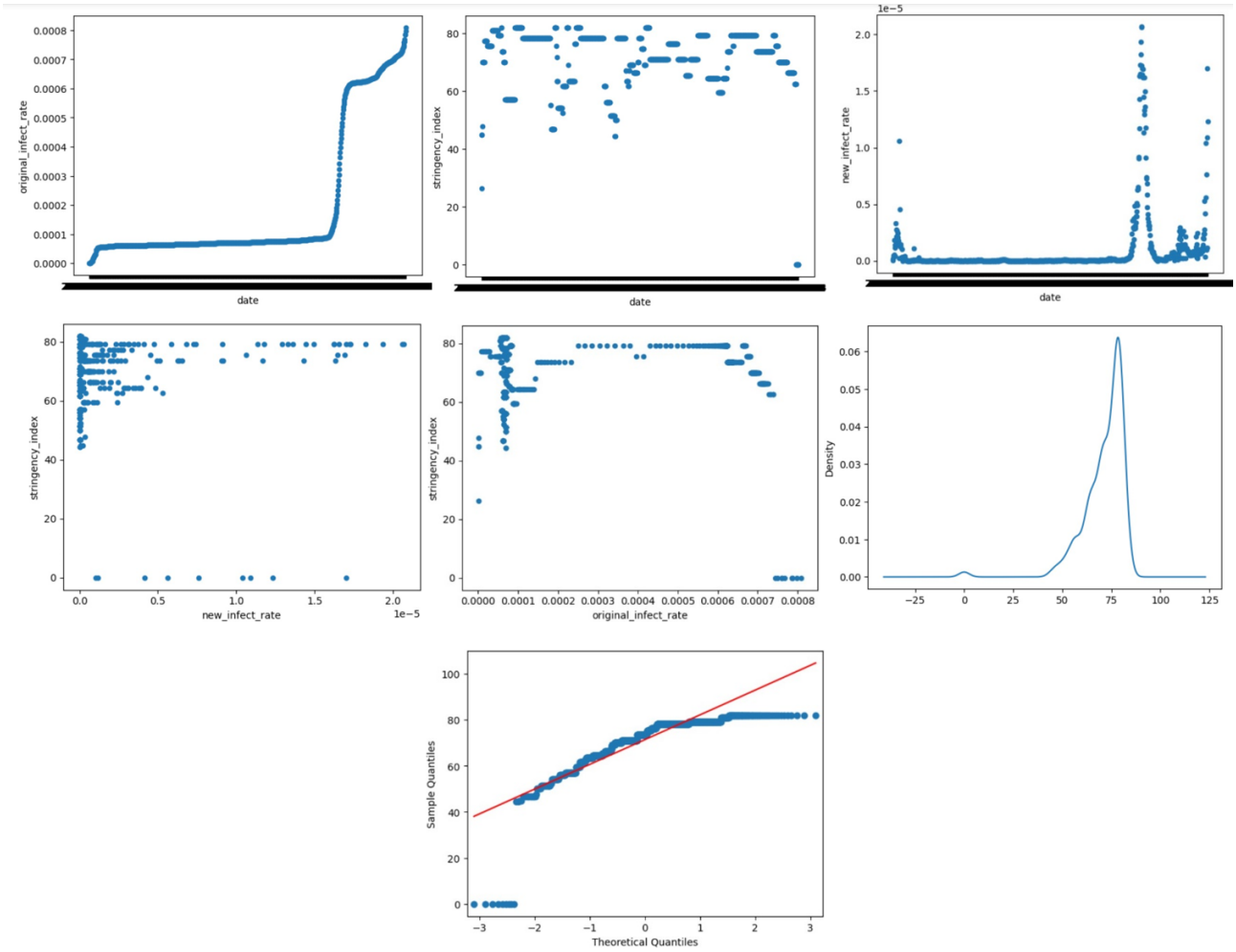
5.3.4 The effectiveness of stringency index on COVID-19 infection rate



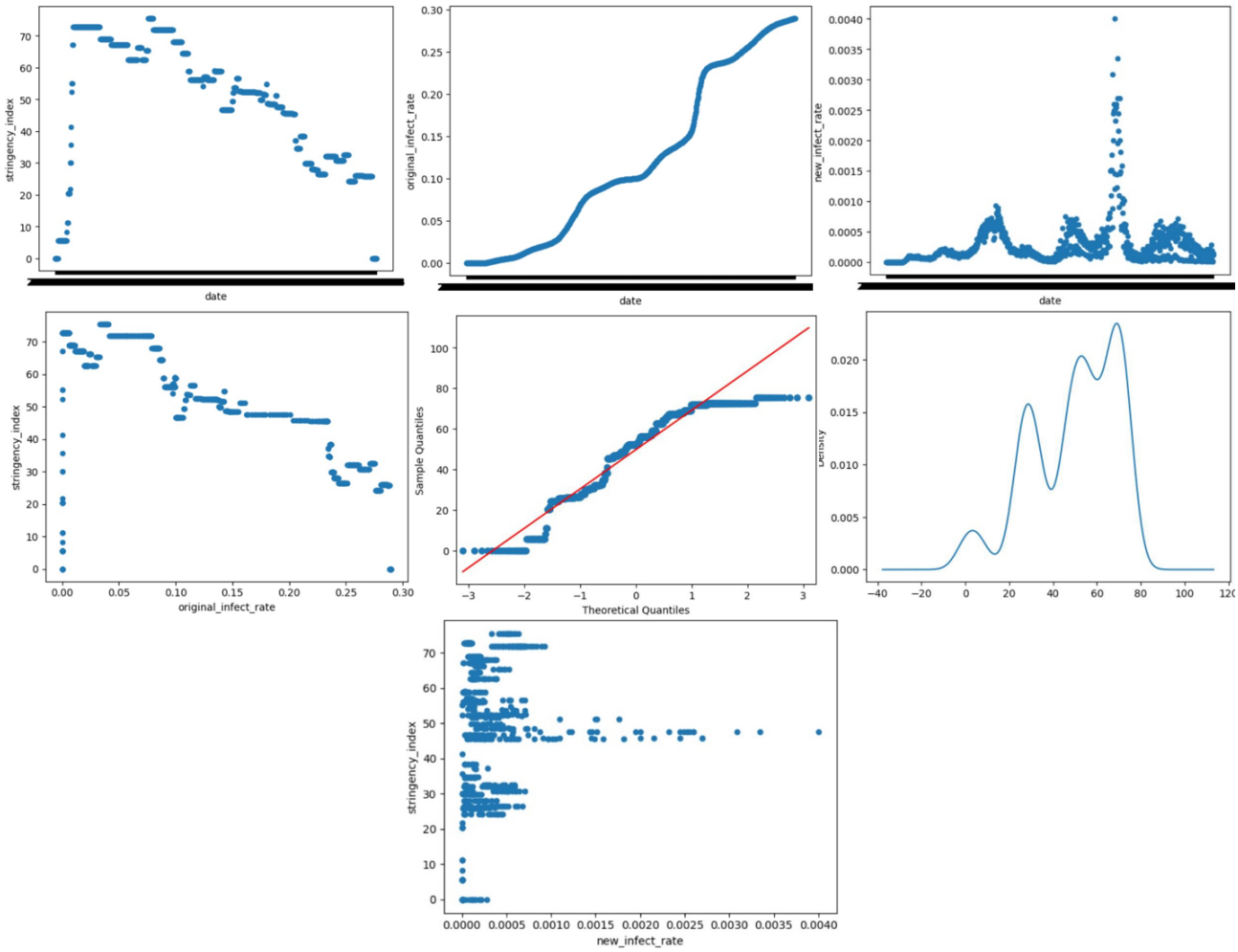
The density plot indicates that most countries in the world take loose epidemic prevention policies.



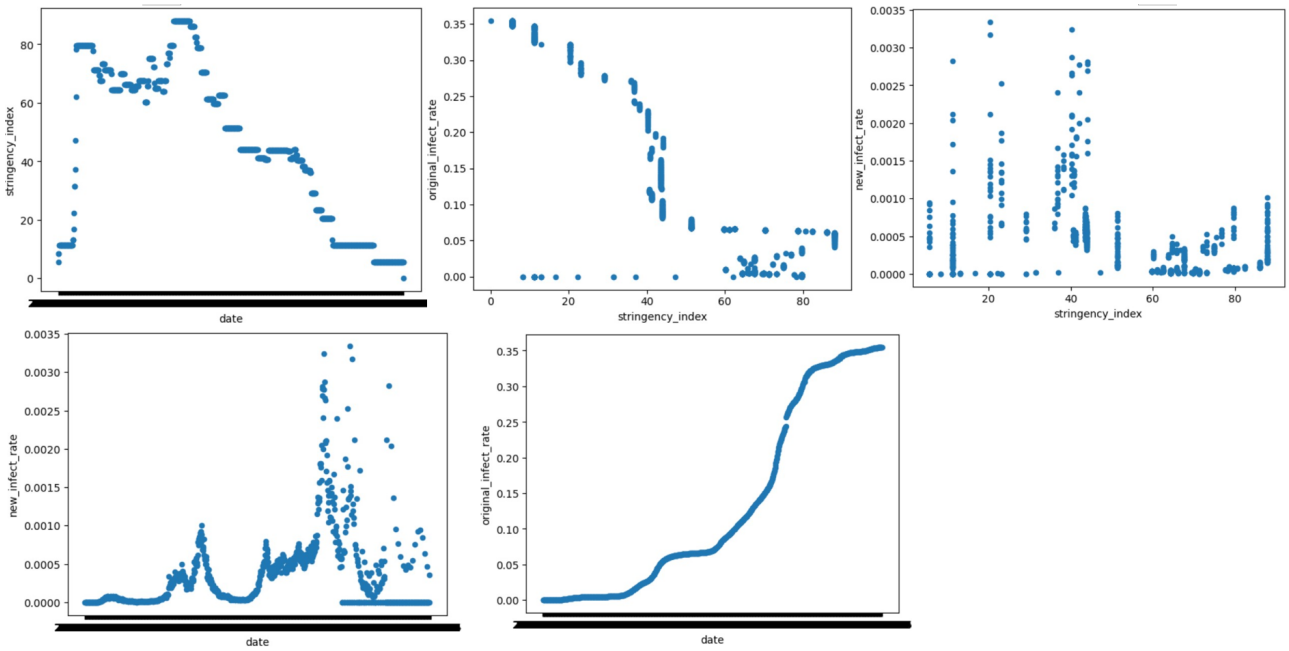
The Q-Q plot indicates that the global stringency index does not obey normal distribution. Moreover, the previous plot indicates that the global stringency index and the infection rate are not linear related. So, we need to be specific to a country for analysis.



These images show the relationship between the data collected from China, where the stringency index is consistently high, but the infection rate is less correlated with it.



These images show the relationship between the data collected from the US, where we find that the stringency index is essentially in decline, and the infection rate is probably inversely proportional to it.



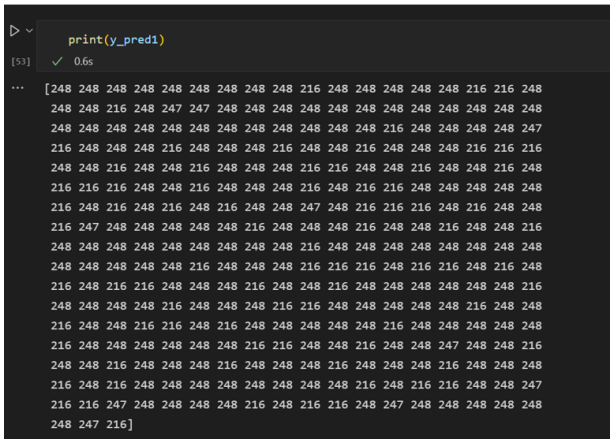
These images show the relationship between the data collected from the UK, where the stringency index is more fluctuating, with an overall decreasing trend but a low correlation between infection rates and itself.

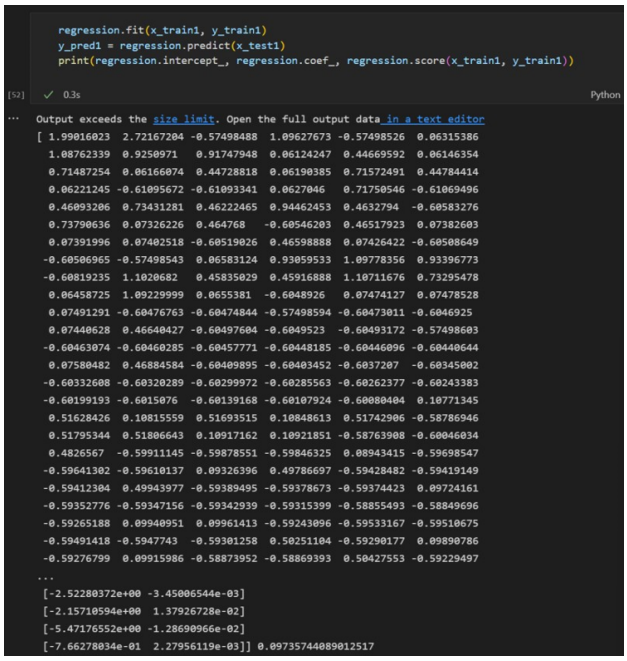
```
UnitedStates['strictness']=0
UnitedStates.loc[UnitedStates['stringency_index']>=50,['strictness']]=1
# UnitedStates[UnitedStates['strictness']==1]
UnitedStates
```

```
from sklearn import preprocessing
process = preprocessing.LabelEncoder()
US_x=UnitedStates[['original_infect_rate','new_infect_rate']]
US_y=UnitedStates['strictness']
US_y = process.fit_transform(US_y)
x_train,x_test,y_train,y_test = train_test_split(US_x,US_y,test_size=0.3,random_state=0)
```

```
    regression.fit(x_train, y_train)
    y_pred = regression.predict(x_test)
    print(regression.intercept_, regression.coef_, regression.score(x_train, y_train))
[22]  ✓ 0.3s                                                                    Python
...  [1.48182449] [[-9.67267155e+00 -7.53811057e-03]] 0.8734353268428373
```

We then chose to analyze the data for the USA. Firstly, by setting index=50 as the cut- off, we divided the strict and the loose boundaries. Then logistic regression was used to calculate its accuracy. The fit was found to be around 0.88, indicating a good fit.





After passing the data into the Logistic Regression Model, the model's accuracy is only about 0.1, which means that the stringency index and the correlation between the new and existing infection rates are extremely low. The prediction results were also very different from the actual results, which shows that the logistic regression model is unsuitable.

Therefore, we used a linear regression model instead. We obtained model A by using 'original infection rate' as the explanatory variable and 'stringency index' as the dependent variable, and model B by using 'new infection rate' as the independent variable and 'stringency index' as the dependent variable, respectively. The correlation score for model A was around 0.42, and the difference between the predicted and actual values was still not small, showing that the fit was still low, while the fit score for model B was only 0.0026, with almost no relationship.

```
import numpy as np
from sklearn.model_selection import GridSearchCV
hyperparameters = {
 'C': [0.1, 1, 100, 1000],
 'gamma': [0.0001, 0.001, 0.005, 0.1, 1, 3, 5],
 'kernel': ('linear', 'rbf')
}
grid = GridSearchCV(
 estimator=model,
 param_grid=hyperparameters,
 cv=5,
scoring='f1_micro',
n_jobs=-1)
grid.fit(x_train1,y_train1)
print(f'Best parameters: {grid.best_params_}')
print(f'Best score: {grid.best_score_}')
```

Finally, after a goodness-of-fit calculation, we chose Linear Support Vector Regression for prediction. The results indicate that a stringency of at least 216 needs to be maintained to achieve the most outstanding possible suppression of the emergence of new cases.

```
model = svm.SVC(C=500, gamma = 0.1, kernel = 'rbf')
model.fit(x_train1,y_train1)
predictions = model.predict(x_test1)
print(sklearn.metrics.f1_score(predictions, y_test1, average='micro'))
print(predictions)
```

```
array([248, 216, 248, 248, 248, 248, 248, 248, 248, 216, 248, 248, 248,
       248, 248, 216, 216, 248, 248, 248, 216, 248, 216, 216, 248, 248,
       248, 248, 248, 248, 248, 248, 248, 248, 248, 248, 248, 248, 248,
       248, 248, 248, 248, 248, 248, 248, 248, 216, 248, 248, 248, 248,
       248, 216, 216, 248, 248, 248, 216, 248, 248, 248, 216, 248, 216,
       216, 248, 248, 248, 216, 216, 216, 248, 248, 216, 248, 248, 216,
       248, 248, 248, 216, 216, 248, 248, 216, 248, 248, 216, 248, 216,
       216, 216, 248, 248, 216, 248, 248, 248, 216, 248, 216, 216, 248,
       216, 248, 248, 248, 216, 248, 216, 248, 216, 248, 216, 248, 248,
       216, 248, 216, 216, 216, 248, 216, 248, 248, 216, 216, 248, 248,
       248, 248, 248, 216, 248, 248, 248, 216, 248, 248, 216, 248, 248,
       216, 248, 248, 248, 216, 248, 216, 248, 248, 248, 216, 248, 248,
       248, 248, 248, 248, 248, 248, 248, 248, 248, 248, 216, 248, 248,
       248, 248, 216, 216, 216, 248, 216, 216, 248, 216, 248, 216, 248,
       216, 216, 248, 248, 248, 216, 248, 248, 216, 248, 248, 248, 248,
       248, 248, 216, 216, 248, 248, 248, 216, 216, 248, 248, 216, 216,
       248, 248, 248, 248, 248, 216, 216, 248, 216, 248, 248, 216, 216,
       248, 216, 248, 248, 248, 248, 248, 216, 248, 248, 248, 248, 248,
       216, 248, 248, 248, 248, 248, 248, 216, 216, 248, 248, 216, 248,
       248, 216, 248, 248, 216, 248, 248, 216, 248, 248, 248, 216, 248,
       248, 248, 216, 248, 248, 248, 216, 248, 248, 248, 216, 248, 216,
       216, 248, 248, 248, 248, 248, 248, 248, 216, 248, 216, 216, 248,
       248, 216, 216, 216, 216, 248, 248, 248, 248, 216, 248, 216, 216,
       248, 216, 248, 248, 248, 248, 248, 248, 216, 216], dtype=int64)
```

6 Data interpretation and Discussions

6.1 Studies on the sources and causes of COVID-19 infection

6.1.1 The relationship between the source of infection and time

The virus spreads primarily through person-to-person contact, and many cases were imported from abroad when policies were not strictly targeted at any customs administration. But because of the rapid spread of the novel coronavirus, the virus that spreads into the country will be quickly introduced into the main cities. The local cases will dramatically increase.

The convexity of the function image is an important property to describe the bending direction of the function image. By using it, we can directly draw a draft sketch of functions. The Chi-square statistic is a non-parametric (distribution-free) tool designed to analyze group differences when the dependent variable is measured at a nominal level (Mary L., 2013). Using chi-square testing, we can infer whether the contingency table containing the data fits our assumptions. As a result, the category "overseas" is initially dominant, before shifting to "locally acquired - linked to a known case or cluster." Because of the rapid spread of COVID, the number of local cases of unknown and unidentified cases is increasing.

The main source of infection is known cases or clusters, and the initial step is from overseas. Therefore, governments should strictly check whether international passengers are carrying the virus to control the first stage of COVID outbreaks. The suggestion to residents is to try to avoid contacting any known COVID cases or places.

6.1.2 Test relationship between age and infection of COVID-19

From the scatter plot of age and percent of cases, from 0-17 years old, the infection rate fluctuates in [6.6%, 2.7%]. From 18 to 65 years old and above, the percentage of cases still show a decreasing trend [20.6% , 12.6%] but much higher than the percentage in 0-17 years old. So overall the whole graph shows an increasing percentage as age increases. This phenomenon can be explained by working age, as the people from 0 17 years old, most of them are in homes or schools during epidemic, there are very few people they will contact, so there is less chance for them to be infected. For people above 17 years old, they are able to work, so the people they contact increased rapidly, much higher chance for them to be infected. For adults, as their age increase, the number of retirements increased, so they don't meet people as much as before.

Therefore, for adults, the trend shows a decreasing pattern. Overall, there are more people an adult will meet than kids, so we conclude that as age increases, there will be more possibility to be infected by COVID-19.

6.1.3 Relation between CBC and respiratory viruses against COVID-19

Overall, the results of Eosinophils, Leukocytes, Monocytes, and Urea tend to align with the studies that we mentioned earlier in Chapter 1. Increases in Eosinophils and Leukocytes reduce the risk of COVID-19, whereas increases in monocytes and Proteina C reativa mg/dL increase the risk of COVID-19. In short, a simple and quick test like CBC is useful in discovering major abnormalities in COVID-19-infected patients and predicting the risks of COVID-19 infection. Monitoring these predictors can help in the early detection of potentially serious cases, allowing for early and efficient medical intervention.

On the other hand, our findings on respiratory viruses support the notion that COVID-19 is unlikely to affect someone who carries another viral respiratory disease such as "Influenza B", "Rhinovirus/Enterovirus", "Inf A H1N1 2009". Due to the fact that the number of respiratory viruses is extremely large, investigating the effect of other respiratory viruses on COVID-19 infection has been difficult. Nevertheless, an increasing body of laboratory evidence provides some reassurance: SARS-CoV-2 and other respiratory viruses cannot get along very well, and this statement has been proven by Richard Webby, an influenza researcher at St. Jude Children's Research Hospital. It is highly improbable that they will circulate simultaneously in one's body.

Our study had several limitations:

- As for the weak coefficients of some predictors such as "Adenovirus"," Parainfluenza 2", limited datasets being used could be blamed. Small sample size might reduce the statistical power.

- The samples are limited to a single region, Sao Paulo, Brazil. This might cause the analysis to be limited in detecting specific co-infection patterns potentially predictive of SARS-CoV-2.
- With the usage of logistic regression, it is slightly more difficult for us to obtain complex relationships, and more powerful algorithm like Neural Networks can easily outperform this algorithm.
- Logistic regression requires zero multicollinearity between independent variables and this has forced us to remove some of the predictors which may be informative in further research.
- Medication may have relevant factor in these parameters of different patients, and we do not consider them in our study.

While the effects of other respiratory viruses and haematological factors on COVID-19 infection are still being debated, our study provides a fundamental and fruitful insight. To make the analysis more informative, we can consider other influential factors such as blood types, the severity of different respiratory viruses, and the medication of patients. To strengthen our argument, further analysis on larger datasets in different hospitals and regions can also be carried out and the results compared with our study.

6.2 Studies on time and number of infections and prediction

6.2.1 Analysis on the solutions to number of infections and prediction.

The graph of the results predicted by the ARIMA model shows a certain degree of fluctuation in the number of COVID-19 daily new cases. And it predicts that the next peak of daily new cases will come in January and February 2023. And the number of COVID-19 daily new cases will gradually decrease after that. However, the credibility of the model's predictions gradually decreases over time.

At the same time, we can clearly see from the graph that the number of COVID-19 daily new cases has indeed increased in the past few winter seasons compared to the other three seasons. Therefore, further predictions are made by the SARIMAX model.

Both from the line chart of daily new cases and from the line chart of COVID-19 daily new cases at the end of each month. The relationship between the daily new cases and the season is largely affected by the sharp increase in daily new cases in January and February 2022. Therefore, it is doubtful whether the SARIMAX model can effectively predict COVID-19 daily new cases. From the following set of graphs, it is also uncertain whether the SARIMAX model can properly predict the number of COVID-19 daily new cases. Next, the objective determines that the SARIMAX model does not correctly predict the number of COVID-19 daily new cases by calculating the error in the prediction results. Since the error is very large, the model is unsuitable for predicting the number of COVID-19 daily new cases. Therefore, the objective concludes that the SARIMAX model is not applicable to predict the number of COVID-19 daily new cases.

6.3 Studies on the solutions to COVID and their effectiveness

6.3.1 The relationship between COVID cases, deaths and vaccinations

Initially, we only used the data grouped into regions and cumulative cases and deaths to find the relationships, which gave us extremely awful results: the correlation coefficient is small, which shows a weak correlation. Moreover, the result did not pass our hypothesis testing, the p-value is too large. The cause may come from different aspects.

Firstly, maybe most cases and deaths happened before the popularity of vaccines; in our data set, we use data about cumulative cases and deaths, so the impact of vaccines may have been largely reduced.

Secondly, there are only 5 pairs of data when we combine raw data into regions, and as the sample size is small, the model we made may not be accurate enough. After that, we improved our method. We found the relationship between vaccinations and cases or deaths in the last 7 days. Do not group countries into regions so that we can have more samples to implement our model. This time, we had a relatively good result, but this can still be improved further.

According to our findings, immunizations have a significant influence on preserving people's health. The information above allows us to make a clear conclusion about the significance of vaccinations: Vaccinations are essential for reducing the risk of death and the infection rate; hence, governments should promote the process of vaccinations so that people can get immunized, in order that the government's COVID-19 policies will have less restriction on human beings' daily lives. Therefore, the world's economic growth can develop further.

6.3.2 To identify and analyse the relationship between COVID-19 and vaccination

As shown in the scatter plot and heatmap, vaccination has a negative relationship with infection, death, and hospitalization. Studies have shown that COVID-19 vaccination has substantially altered the course of the pandemic while saving tens of millions of lives globally (O.J. Watson et al., 2022). Besides, some research also suggests that a high rate of COVID-19 vaccine coverage is negatively correlated with the reproduction rate and the number of ICU patients per million (C. Huang et al., 2022).

We are able to convince people that vaccination plays an important role in COVID-19 outbreak control. Although it cannot prevent the public from being infected directly, it can raise people's awareness of epidemic prevention, reducing mortality and hospitalization rates. In the long run, it is one of the most effective ways to protect public health from the outbreak. Hence, as some developing and low-income countries are experiencing shortages in vaccination and vaccination rates (P. Winskill, 2022), reinforcing the need for global vaccine equity and coverage is necessary.

At the present time, the COVID-19 outbreak has not been completely contained. However, fortunately, vaccination, one of the most effective ways to control the epidemic, has been promoted as it plays an important role in diminishing infection, mortality, and hospitalization.

After the investigations on the objective, some possible suggestions are given based on the re-

sults obtained. It is suggested that vaccination should be promoted in developing countries by developing the technique to produce the vaccine and reducing the marginal cost of production. Besides, as vaccine hesitancy is the main reason why the public resists vaccination, it is necessary to put vaccines in perspective.

6.3.3 Investigate to what extend the health policies impact number of COVID-19 infection.

As different policies lead to two identical results, either the policy is effective to limit COVID-19, or the policy is not functioning as excepted. Turn these results into numerical data, then it should be an increase or decrease in infection and cure rate. SIR model gives simulation under different infection and cure rate.

By viewing the graphs of the SIR model given:

- R0 is the threshold value, if R0 is close to 1, there is no epidemic, and the spread of infectious diseases is very limited. However, once R0 ¿ 1, the total number of infected people explodes as R0 increases. Lower cure rate led to longer period of epidemic, higher infection rate lead to more rapid outbreak patients. So lower down R0 is essential for limiting pandemic, strict policies like instruct people wearing masks takes significant rule during this period.

- Since the shape of the infection curve is an exponential function, the slope of the function will vastly increase as time goes on. As a result, shorter time for patients contact with others can largely slow down the speed of virus spread.

Insofar as the SIR model is purely based on mathematics, as the parameters are inserted, the model is then fixed. In order to make it closer to reality, some arbitrary variable should be considered for example how people's sentiment will affect their action.

6.3.4 The effectiveness of stringency index on the COVID-19 infection rate

A density plot of the world's stringency index and a study of the stringency cut-offs show that most countries around the globe adopted more lenient policies during the epidemic. In the scatter plot of date versus stringency index, there has been some degree of fluctuation in the stringency of policies in different countries, suggesting that epidemic policies were not static. The two models studied for the US and the predictions suggest that the stringency index has little correlation with infection rates. Still, the most consistent is the linear regression model. These models can only be used to speculate that stricter prevention and control measures in some regions are probably ineffective in suppressing virus transmission. There are, of course, areas for improvement in this study. The first is data selection. Since The index and variables should be revised repeatedly until the model is closest to the test data. Besides, the classification of infection rate, or the encoding part, should be improved in further analysis. The third is a cross-sectional comparison with other data, such as age and health care per capita, that may be more useful in determining how the stringency index affects infection rates.

This research aims to investigate the effectiveness of implementing strict COVID-19 control policies. The result shows that

1. Stringency = 50 is a perfect separatrix for dividing the strictness of COVID control policies.

2. The appropriateness or effectiveness of a country's response to COVID-19 is apparently not measured by the stringency index.

3. It is necessary to adapt the stringency according to outbreak accelerations, and react expeditiously.

Considering the reality of COVID-19 prevention and control, a higher score does not necessarily mean that a country's response is 'better' than others lower on the index.

7 Conclusions

After investigating the following objectives (1Studies on the sources and causes of infection 2 time and number of infections and prediction, 3 the solutions to COVID and their effectiveness), we have made several valuable conclusions and suggestions about covid-19. Through the data science method, we hope our results can better understand the virus itself and contribute to people's understanding of its transmission and prevention and control methods.

First, we analyzed the reasons for COVID infection. For a particular region, the sources of infection approach local causes because of the virus's fast-spreading character. In terms of age, the graph shows that the percentage increases generally with age as the division is 17 years old. This phenomenon can be explained by working age (adults have more social contacts). To Complete Blood Count and Respiratory Viruses; simple, rapid tests like the CBC are useful in detecting significant abnormalities in patients with COVID-19 infection and predicting the risk of COVID-19 infection and COVID-19 is unlikely to affect people with other viral respiratory diseases.

Second, we made a prediction to the number of cases. The number of daily new cases confirmed by COVID-19 diagnoses was predicted by ARIMA and SARIMAX models. COVID-19 was found to be seasonal, with a high number of confirmations in winter. The number of daily new cases was predicted to decrease after March 2023. However, the error is very large, the SARIMAX model is unsuitable for predicting the number of COVID-19 daily new cases.

Third, we estimated the relationship between COVID-19 and ongoing solutions. By showing the relationship between vaccination and cases of deaths over the past 7 days using scatter plots and heat maps, it was demonstrated that vaccination was inversely associated with infection, death and hospitalization. Vaccination has been promoted as one of the most effective ways to contain outbreaks, playing an important role in reducing infection, mortality and hospitalization rates. Immunization has a major impact on keeping people healthy. We believe governments should promote the vaccination process so that people can get vaccinated.

Last, the relationship between governmental policies and COVID-19 infection was analyzed. SIR Model was used to give the simulation results under different infection and cure rates and made notes. Although most countries in the world adopted relatively relaxed policies during the epidemic, the strictness of policies in different countries fluctuated to a certain extent, indicating that the epidemic policies were not immutable. Two models and projections from a US study showed little correlation between the rigor index and infection rates. Therefore, the policy is not the only important factor in determining the course of the epidemic.

8 References

8.1 Studies on time and number of infections and prediction

8.1.1 Analysis on the solutions to number of infections and prediction.

McHugh, M. L. (2013). The chi-square test of independence. Biochemia Medica, 23(2), 143-149.

WHO Coronavirus (COVID-19) Dashboard. Retrieved November 16, 2022, from https://covid19.who.int/data

8.1.2 Test relationship between age and infection of COVID-19

Ayoub, H. H., Chemaitelly, H., Seedat, S., Mumtaz, G. R., Makhoul, M., & Abu-Raddad, L. J. (2020). Age could be driving variable SARS-CoV-2 epidemic trajectories worldwide. PLoS One, 15(8), e0237959.

Boehmer, T. K., DeVies, J., Caruso, E., van Santen, K. L., Tang, S., Black, C. L., ... & Gundlapalli, A. V. (2020). Changing age distribution of the COVID-19 pandemic—United States, May–August 2020. Morbidity and Mortality Weekly Report, 69(39), 1404.

Davies, N. G., Klepac, P., Liu, Y., Prem, K., Jit, M., & Eggo, R. M. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics. Nature medicine, 26(8), 1205-1211.

Goldstein, E., Lipsitch, M., & Cevik, M. (2021). On the Effect of Age on the Transmission of SARS-CoV-2 in Households, Schools, and the Community. The Journal of infectious diseases, 223(3), 362-369.

8.1.3 Relation between CBC and respiratory viruses against COVID-19

Assaf, D., Gutman, Y. A., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., ... & Tirosh, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine, 15(8), 1435-1443.

Conger, K. (2020). COVID-19 patients often infected with other respiratory viruses, preliminary study reports. Fleitas, P. E., Paz, J. A., Simoy, M. I., Vargas, C., Cimino, R. O., Krolewiecki, A. J., & Aparicio, J. P. (2021). Clinical diagnosis of COVID-19. A multivariate logistic regression analysis of symptoms of COVID-19 at presentation. Germs, 11(2), 221.

Jawa, T. M. (2022). Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia. Alexandria Engineering Journal, 61(10), 7995-8005.

Majumder, A. B., Gupta, S., Singh, D., & Majumder, S. (2021, February). An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. In

Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012011). IOP Publishing.

Palladino, M. (2021). Complete blood count alterations in COVID-19 patients: A narrative review. Biochemia medica, 31(3), 0-0.

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. MedRxiv.

8.2 Studies on time and number of infections and prediction

8.2.1 Analysis on the solutions to number of infections and prediction.

Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. Journal of infection and public health, 13(7), 914-919.

Anne, W. R., & Jeeva, S. C. (2020). ARIMA modelling of predicting COVID-19 infections. medRxiv.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 29, 105340.

Bhangu, K. S., Sandhu, J. K., & Sapra, L. (2021). Time series analysis of COVID-19 cases. World Journal of Engineering.

Chintalapudi, N., Battineni, G., & Amenta, F. (2020). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. Journal of Microbiology, Immunology and Infection, 53(3), 396-403.

Gupta, R., & Pal, S. K. (2020). Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv. Hamilton, J. D. (2020). Time series analysis. Princeton university press.

Jain, A., Sukhdeve, T., Gadia, H., Sahu, S. P., & Verma, S. (2021, March). COVID19 prediction using time series analysis. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 1599-1606). IEEE.

Kumar, N., & Susan, S. (2020, July). COVID-19 pandemic prediction using time series forecasting models. In 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-7). IEEE.

Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel medicine and infectious disease, 37, 101742.

Mustafa, H. I., & Fareed, N. Y. (2020, August). Covid-19 cases in Iraq; forecasting incidents

using box-Jenkins Arima model. In 2020 2nd Al-Noor International Conference for Science and Technology (NICST) (pp. 22-26). IEEE.

Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. In Time series analysis and its applications (pp. 75-163). Springer, Cham.

Singh, S., Sundram, B. M., Rajendran, K., Law, K. B., Aris, T., Ibrahim, H., ... & Gill, B. S. (2020). Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. The Journal of Infection in Developing Countries, 14(09), 971-976.

Solanki, A., & Singh, T. (2021). COVID-19 epidemic analysis and prediction using machine learning algorithms. In Emerging Technologies for Battling Covid-19 (pp. 57-78). Springer, Cham.

Sulasikin, A., Nugraha, Y., Kanggrawan, J., & Suherman, A. L. (2020, September). Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta. In 2020 IEEE International Smart Cities Conference (ISC2) (pp. 1-6). IEEE.

8.3 Studies on the solutions to COVID and their effectiveness

8.3.1 The relationship between COVID cases, deaths and vaccinations

Rustagi, V., Bajaj, M., Singh, P., Aggarwal, R., AlAjmi, M. F., Hussain, A., ... & Singh, I. K. (2022). Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models. Frontiers in Cellular and Infection Microbiology, 1380.

World Health Organization. (n.d.). WHO Coronavirus (COVID-19) Dashboard. Retrieved November 16, 2022, from https://covid19.who.int/data

8.3.2 To identify and analyse the relationship between COVID-19 and vaccination

de Albuquerque Veloso Machado, M., Roberts, B., Wong, B. L. H., van Kessel, R., & Mossialos, E. (2021). The relationship between the COVID-19 pandemic and vaccine hesitancy: a scoping review of literature until August 2021. Frontiers in public health, 9, 747787.

Subramanian, S. V., & Kumar, A. (2021). Increases in COVID-19 are unrelated to levels of vaccination across 68 countries and 2947 counties in the United States. European journal of epidemiology, 36(12), 1237-1240.

Watson, O. J., Barnsley, G., Toor, J., Hogan, A. B., Winskill, P., & Ghani, A. C. (2022). Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. The Lancet Infectious Diseases, 22(9), 1293-1302.

8.3.3 Investigate to what extend the health policies impact number of COVID-19 infection.

Chung, H. W., Apio, C., Goo, T., Heo, G., Han, K., Kim, T., ... & Park, T. (2021). Effects of government policies on the spread of COVID-19 worldwide. Scientific reports, 11(1), 1-10.

Dayaratna, K., & Vanderplas, A. (2021). A Statistical Analysis of COVID-19 and Government Protection Measures in the US (No. 243). Heritage Foundation Special Report. Gol, S., Pena, R. N., Rothschild, M. F., Tor, M., & Estany, J. (2018). A polymorphism in the fatty acid desaturase-2 gene is associated with the arachidonic acid metabolism in pigs. Scientific reports, 8(1), 1-9.

Kennedy, D. M., Zambrano, G. J., Wang, Y., & Neto, O. P. (2020). Modeling the effects of intervention strategies on COVID-19 transmission dynamics. Journal of Clinical Virology, 128, 104440.

Law, K. B., Peariasamy, K. M., Gill, B. S., Singh, S., Sundram, B. M., Rajendran, K., ... & Abdullah, N. H. (2020). Tracking the early depleting transmission dynamics of COVID-19 with a time-varying SIR model. Scientific reports, 10(1), 1-11.

Liu, P. Y., He, S., Rong, L. B., & Tang, S. Y. (2020). The effect of control measures on COVID-19 transmission in Italy: Comparison with Guangdong province in China. Infectious diseases of poverty, 9(1), 1-13.

Valcarcel, B., Avilez, J. L., Torres-Roman, J. S., Poterico, J. A., Bazalar-Palacios, J., & La Vecchia, C. (2020). The effect of early-stage public health policies in the transmission of COVID-19 for South American countries. Revista Panamericana de Salud Pública, 44.

8.3.4 The effectiveness of stringency index on the COVID-19 infection rate

Bajra, U. Q., Aliu, F., Aver, B., & Čadež, S. (2022). COVID-19 pandemic–related policy stringency and economic decline: was it really inevitable?. Economic Research-Ekonomska Istraživanja, 1-17.

Cheng, H., & Tsang, R. (2021). The potential impact of the COVID-19 pandemic on global poverty and income disparity: A literature review.

Dergiades, T., Milas, C., Mossialos, E., & Panagiotidis, T. (2022). Effectiveness of government policies in response to the first COVID-19 outbreak. PLOS Global Public Health, 2(4), e0000242.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature medicine, 26(6), 855-860.

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., ... & Roser, M. (2020).

Coronavirus pandemic (COVID-19). Our World in Data.

Moore, S., Hill, E. M., Tildesley, M. J., Dyson, L., & Keeling, M. J. (2021). Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. The Lancet Infectious Diseases, 21(6), 793-802.

Tatlow, H., Cameron-Blake, E., Grewal, S., Hale, T., Phillips, T., & Wood, A. (2021). Variation in the response to COVID-19 across the four nations of the United Kingdom. Blavatnik Sch Gov Work Paper.

Wibbens, P. D., Koo, W. W. Y., & McGahan, A. M. (2020). Which COVID policies are most effective? A Bayesian analysis of COVID-19 by jurisdiction. PloS one, 15(12), e0244177.

9 Appendix

9.1 Studies on time and number of infections and prediction

9.1.1 Analysis on the solutions to number of infections and prediction.

Python:
https://colab.research.google.com/drive/1jTMF1WJShBn _U5X9z-EGeg9EiORKphxO?usp=share _link
HTML:
https://drive.google.com/file/d/1i-cn1uO5BVkZrwIbIUcIhZQaVN7nbIai/view?usp=share _link Dataset:
https://drive.google.com/drive/folders/1nUw1WgJbCRpGiBOZvXjrc8XUBYojLzS5?usp=sharing

9.1.2 Test relationship between age and infection of COVID-19

Python:
https://colab.research.google.com/drive/1-4Ovyf8piyX _bgdpUiPDnXMdMfApeaJ2?usp=share _link
HTML:
https://drive.google.com/file/d/1kEoyK06jSk15BcgrVYCFSpaQI5ZQuNCk/view?usp=share _link
Dataset:
https://drive.google.com/file/d/1ZIqmvrUGuyprktqNiJQaVEivT _xSRKpW/view?usp=share _link

9.1.3 Relation between CBC and respiratory viruses against COVID-19

Python:
https://colab.research.google.com/drive/1laG07FhvuMAx5zQR2HbyjpD1mHuIhw-i?usp=sharing
HTML:
https://drive.google.com/file/d/1BpaEjqtDpnUuwXzorJEJcp6E6Xnjt3hW/view?usp=share _link
Dataset:
https://docs.google.com/spreadsheets/d/1nnQHGMuTRcTElb6kePIHmKOuFbi5RSea/editgid=773891069

9.2 Studies on time and number of infections and prediction

9.2.1 Analysis on the solutions to number of infections and prediction.

Python:
https://colab.research.google.com/drive/10hkA2uPEsvNfIz0vd–zCD7yAGdS8172?usp=share _link
https:
//colab.research.google.com/drive/1CuKw1 _TfagBxa _vcCis99UcI3MZ3Fdhx?usp=share _link
HTML:
https://drive.google.com/file/d/18a7 _IHEUuCCQdSSHEJj-knInguvPpdk7/view?usp=share _link
https:
//drive.google.com/file/d/1vrEjjUR-0qyDsKNKhkJ1yE _j-BfiIXwY/view?usp=share _link
Dataset:
https:

//drive.google.com/file/d/1BhuSXgUNjm-DrZqGOFP3RkHXtauaLlhQ/view?usp=sharing
https:
//drive.google.com/file/d/10MJaZo1MLMXurPomPYAT6U1x5ezPAjYn/view?usp=sharing

9.3 Studies on the solutions to COVID and their effectiveness

9.3.1 The relationship between COVID cases, deaths and vaccinations

Python:
https:
//colab.research.google.com/drive/1U-sZJacnN3nFPHZeKaa975TAYfRKn0vt?usp=sharing
HTML:
https:
//drive.google.com/file/d/1N3n _0noOR0MRhV _s1rKVjC4 _RhJkWKRV/view?usp=share _link
Dataset:
https://drive.google.com/file/d/11qD8OOBhHgEamlREI _6jB1ATMrzepkvL/view?usp=share _link
https:
//drive.google.com/file/d/1iC9vnUpINc51z6OmoX44sYrtMTXSPvX0/view?usp=share _link

9.3.2 To identify and analyse the relationship between COVID-19 and vaccination

Python:
https:
//colab.research.google.com/drive/1bXt0bvyMgo4ySz2HVnQ9XUTbCr9Tojj9?usp=share _link
HTML:
https://drive.google.com/file/d/1yN42GYrzHtL9Lg5RyK79Uhb89rLxQ0KI/view?usp=share _link
Dataset:
https:
//drive.google.com/drive/folders/1g1AQB4rKOxlJPEUS2YJJ _yJMz7GFqnx5

9.3.3 Investigate to what extend the health policies impact number of COVID-19 infection.

Python:
https:
//colab.research.google.com/drive/1SOLwv4dXaOVfg1YBPpVWRYDSUaGJNULR?usp=sharing
HTML:
https:
//drive.google.com/file/d/1A4uFDuljTq4uFTnCVZeZf _loS-GufCyz/view?usp=share _link
Dataset:
https:
//drive.google.com/file/d/1LJgsX0MEEdK0gQDkjPi8JpHIZ07Jihka/view?usp=share _link

9.3.4 The effectiveness of stringency index on the COVID-19 infection rate

Python: https:
//colab.research.google.com/drive/1E6nFjeQeRCjr9L9vAVz0ly2jRIvIGuHW?usp=sharing
HTML:
https:
//drive.google.com/file/d/1-raGWbAiFzjlSAz _D9Ht-GXPvpo3XNyV/view?usp=share _link
Dataset:
https:
//drive.google.com/file/d/1VqnivbJyPnL2OiJl0sNGqbgWDUV _G _rR/view

10 participation

Luo Dongyu 3035974597

Chen Haodong 3035974030

Li Guodong 3036098211

Jiang Qingyi 3036103858

Xu Ziqi 3036063060

Yeo Kiahhauh 3036103286

Liu Qi 3036094629