Tianheng Ma

839931320

UPI:TMA502

Iteration 4

Contents

[Step 1 business and situation understanding 3](#_Toc68767416)

[1.1 Identify the objectives of the business/situation 3](#_Toc68767417)

[1.2 Assess the situation 4](#_Toc68767418)

[1.3 Determine data mining objectives 5](#_Toc68767419)

[1.4 Produce a project plan 6](#_Toc68767420)

[Step 2: business and situation understanding 7](#_Toc68767421)

[2.1 Collect initial data 7](#_Toc68767422)

[2.2 Describe the data 7](#_Toc68767423)

[2.3 Explore the data 9](#_Toc68767424)

[2.4 Verify the data quality 14](#_Toc68767425)

[Step 3 Data preparation 16](#_Toc68767426)

[3.1 Select the data 16](#_Toc68767427)

[3.2 Clean the data 18](#_Toc68767428)

[3.3 Construct the data 19](#_Toc68767429)

[3.4 Integrate various data sources 20](#_Toc68767430)

[3.5 Format the data as required 21](#_Toc68767431)

[Step 4 Data Transformation 21](#_Toc68767432)

[4.1 Reduce the data 21](#_Toc68767433)

[4.2 Project the data 22](#_Toc68767434)

[Step 5 Data-Mining Method 23](#_Toc68767435)

[5.1 Match and discuss the objectives of data mining to data mining method. 23](#_Toc68767436)

[5.2 Select the appropriate data mining method based on discussion. 24](#_Toc68767437)

[Step 6 Data mining algorithms selection 24](#_Toc68767438)

[6.1 Conduct exploratory analysis and discuss 24](#_Toc68767439)

[6.2 Selecting data mining algorithms based on discussion 25](#_Toc68767440)

[6.3 Build/Select appropriate models and choose relevant parameter 27](#_Toc68767441)

[Step 7 Data Mining 29](#_Toc68767442)

[7.1 Create and justify test design 29](#_Toc68767443)

[7.2 Conduct Data Mining 29](#_Toc68767444)

[7.3 Search for Patterns 31](#_Toc68767445)

[Step 8 interpretation 32](#_Toc68767446)

[8.1 Study and discuss the mined pattern 32](#_Toc68767447)

[8.2 Visualize the data, result, models and patterns: 33](#_Toc68767448)

[8.3 Interpret the result, model and patterns 36](#_Toc68767449)

[8.4 Assess and evaluate results, models and patterns 38](#_Toc68767450)

[8.5 Iterate prior step 39](#_Toc68767451)

# Step 1 business and situation understanding

## 1.1 Identify the objectives of the business/situation

Study Objective: Trend of the number of deaths caused by covid-19 in Italy and its effect.

Covid-19, discovered in the Wuhan seafood market in early 2020, has become a global disaster in the past year. It has caused countless people to get sick and lose their jobs a. A large number of enterprises closed down, and the national economy regressed. To fight against this global disease, the 17 sustainable development goals of the United Nations are vital. Seventeen sustainable development goals aim to sustainable development and solve the development problems in three dimensions: society, economy, and environment. (United Nations, n.d.)

Goal 1: No Poverty

According to the latest monitoring of the international labor organization, 114 million jobs were lost in 2020. The loss of working hours of employment was about four times that during the financial crisis in 2009. The International Labor Organization assesses that the loss of working hours in 2020 (compared with pre-pandemic levels) was equivalent to 255 million full-time jobs, resulting in a loss of $3.7 trillion in labor income. This is even higher than the worst estimate made by the international labor organization in the spring of 2020, when it predicted that the annual loss of labor income would be between $860 billion and $3.44 trillion (Richter, 2021). In order to control the epidemic, many cities have to remain closed for a while. This led to the impact of enterprise work, and enterprises have to lay off staff to control operating costs. Even before the occurrence of corvid-19, the goal of eliminating poverty by 2030 could not be achieved. Therefore, how to reduce the impact of covid19 on poverty eradication is a significant research topic.

Goal 3: Good Health and Well-being

Health is one of the vital aspects of human life. The most common symptoms of COVID-19 are fever, dry cough, fatigue, and so on. In some impoverished areas, such as parts of Africa and parts of South America, the government cannot provide all citizens treatment. Poor health conditions also lead to the rapid spread of the disease. Even in developed countries, the virus's spread has not been curbed because effective measures have not been taken at the first time. Because the covid-19 virus is highly infectious, and there is no effective vaccine for it at present, experts from the World Health Organization of the United Nations even warn people to be prepared for the virus not to be eliminated in the short term. As of today, more than 1.3 million people around the world have been infected with covid19. (News, 2021)The United Nations' ultimate goal is to reduce the global maternal mortality ratio to less than 70 per 100,000 live births by 2030. (United Nations, n.d.). The control of the covid-19 epidemic is undoubted of great help to goal 3 of the United Nations.

Goal 16: Peace, Justice, and Strong Institutions

The first confirmed case of covid-19 was found in the Wuhan seafood market and spread worldwide in the following months. To control the epidemic, major cities have been blocked across China, most of Europe and Australia. The city's blockade caused many people to lose their jobs and the unemployed forced to roam the streets. The city's crime rate has been increasing, which may significantly affect the safety of the town. Besides, racial discrimination caused by the epidemic also needs to be considered. Some extreme racists abuse Asians because the virus was found in Asia. Many cases of violence against Asian groups have been reported in many countries around the world. All of these factors will have a negative impact on the goals of the United Nations All of these factors will have a negative effect on the goals of the United Nations, and these also need to be addressed

The data mining goal is to figure out the connection between disease spread rate and mortality.

The critical success criteria: according to the data analysis, results can effectively curb the trend of virus transmission

## 1.2 Assess the situation

The epidemic situation of covid has caused significant losses all over the world. As early as Jan 31, Italy launched a six-month national health emergency, the first country in Europe to stop flights to and from China. The datasets’ resource is Kaggle (Kaggle, 2021), a data modeling and data analysis competition platform. Enterprises and researchers can publish data on it, and statisticians and data mining experts can compete on it to produce the best model.

My assumptions are:

1. In the early stage, with the entry of virus carriers, the number of infected people increases over time. At this point, the mortality rate increases.
2. With the government's attention to the epidemic, the number of infected people has gradually decreased. Medical facilities are steadily established around the epidemic, and patients are isolated. At this time, the growth rate of infected people slowed down, but the mortality rate would not decrease immediately.
3. In the later stage, if the epidemic situation is gradually controlled, the infection rate and mortality rate will decrease simultaneously. On the contrary, if the government does not take adequate measures, the secondary infection rate may increase.

However, there still may exist some risks. For example, the adequacy of funds may limit the plan's progress; due to the epidemic situation, the data communication between different regions is not timely, resulting in more error values.

## 1.3 Determine data mining objectives

The data mining goal is to figure out the connection between disease spread and mortality. Reflected in the specific variables is the relationship between the number of deaths and new infections, the number of intensive care patients, and rehabilitation. Also, data analysis results can help the government predict the trend of the whole epidemic situation and then better allocate medical supplies according to the epidemic situation. Time and quality are critical success criteria for a project. The earlier the government can get the predicted results from the data analysis, the earlier it can take action against the covid-19 epidemic.

## 1.4 Produce a project plan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Duration | Start | Finish | Predecessors |
| Business Understanding | 7 days | 06/01/21 | 06/08/21 | 1 |
| Data understanding | 14days | 06/09/21 | 06/22/21 | 1, 2 |
| Data preparation | 14days | 06/23/21 | 07/06/21 | 1, 2, 3 |
| Data transformation | 14days | 07/07/21 | 07/017/21 | 4, 1, 2, 3 |
| evaluation | 7days | 07/15/21 | 07/21/21 | 5 |
| Data mining method selection | 7days | 07/22/21 | 08/04/21 | 6 |
| Data mining algorithm selection | 14days | 08/05/21 | 08/21/21 | 5, 6, 7, 8 |
| Data mining | 14days | 08/22/21 | 08/28/21 | 5, 6, 7, 8 |
| Interpretation | 7days | 08/29/21 | 09/04/21 | 1, 2, 3, 4, 5, 6, 7, 8 |
| Action | 7days | 09/05/21 | 09/12/21 | 9 |

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | **Resources** | **Risks** |
| Business understanding | One week | All analysts | National conditions gap  Information gap |
| Data understanding | Two weeks | All analysts | Data problems, technology problems |
| Data preparation | Two weeks | Data mining consultant, economist | Data problems, technology problems, |
| Modeling | One week | Data mining cousultant | Data problems, technology problems |
| Evaluation | Two weeks | All analysts | inability to implement results |
| Deployment | Two weeks | All analysts | Economic problems, inability  to implement results |

Figure 1: Project Plan Table

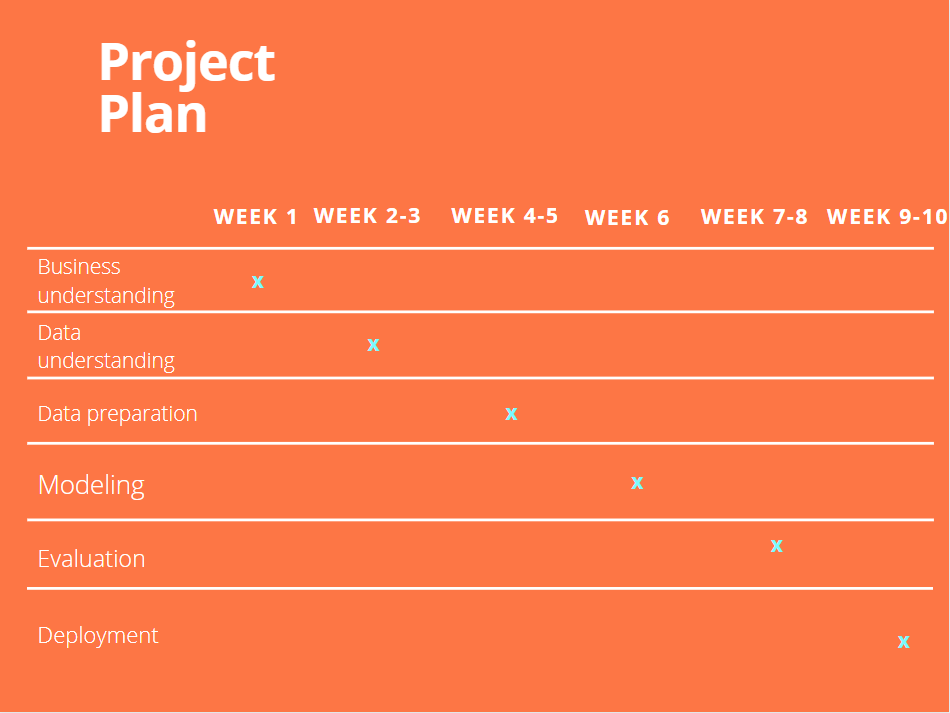


Figure 2: Gantt Chart

# Step 2: business and situation understanding

## 2.1 Collect initial data

**Data collection process:**

The initial data was collected from Kaggle (Kaggle, 2021),(https://www.kaggle.com/sudalairajkumar/covid19-in-italy), a free dataset modeling platform. Detailed data on the Covid-19epidemic is available free of charge on this website. After searching and screening, Italy's dataset in 2020 was selected as the analysis object because Itay can easily represent developed countries' efforts and results to deal with the epidemic.

During the dataset selection part, The date column format of the selected data is month-day-year. So I divided the date format column into three columns: year, month, and day, which can more intuitively reflect the trend of infected people over time.

I use spark.read.csv to collect the initial data. 

**Data privacy issue:**

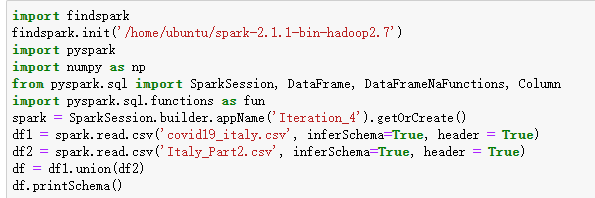
This is an open-source dataset uploaded to Kaggle. So there are no privacy issues. Also, the data source is not the Italian government, so the reliability of the data can not be guaranteed.

**Issue encountered during data collection:**

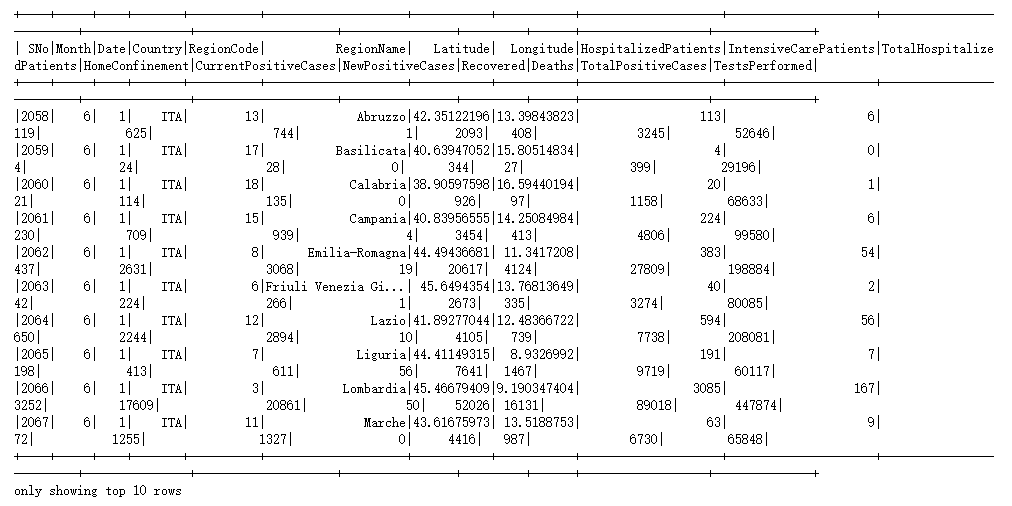
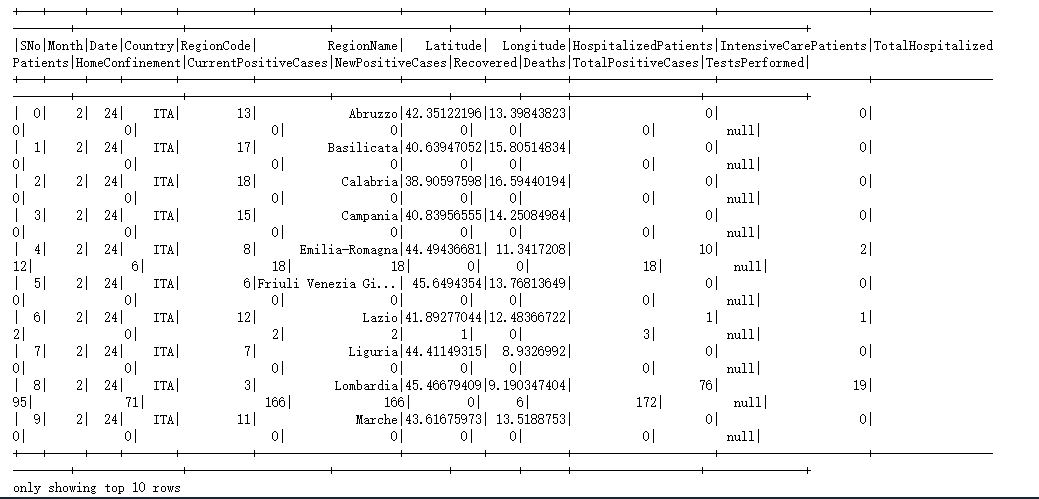
The date form in the original dataset is recorded as mm/dd/yy. It is not convenient to classify. So I separate the data into three different columns that represent month, date, and year.

## 2.2 Describe the data

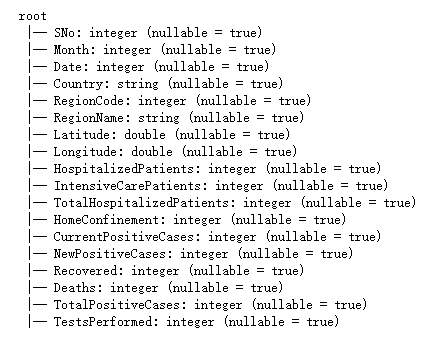
In this step, I use python to explore the collected data. At first, I use spark.read.csv to import the two datasets. Then, count() and printSchema() functions are used to show the type of values in the two datasets.



And here are the result screenshots:



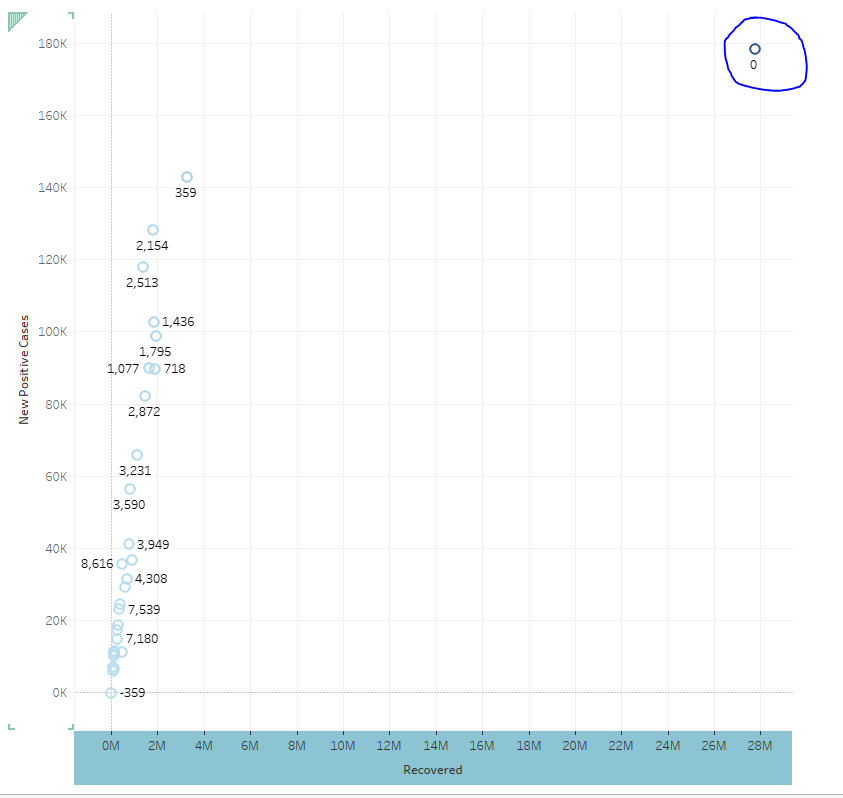
This dataset includes 18 followed columns:



1. Sno: the number of the infected individual, Numerical unit.
2. Month: the month in which the infected person was detected, Numerical unit.
3. Date: the day on which the infected person was found, Numerical unit.
4. Country: Country name. It’s categorical.
5. RegionCode: the code of the living area of the infected person. It’s categorical.
6. RegionName: the name of the living area of the infected individual. It’s categorical.
7. Latitude: the latitude value of the region, Numerical unit.
8. Longitude: the longitude value of the region, Numerical unit.
9. HospitalizedPatients: the number of hospitalized patients, Numerical unit.
10. IntensiveCarePatients: the number of intensive care patients, numerical unit.
11. TotalHospitalizedPatients: the counts of hospitalized patients and intensive care patients, numerical unit.
12. HomeConfinement: Number of individuals in the home confinement stage, not sure if they are infected, numerical unit.
13. CurrentPositiveCases: the number of current positive patients, numerical unit.
14. NewPositiveCases: the number of new confirmed cases, numerical unit.
15. Recovered: the number of recovered patients, numerical unit.
16. Deaths: the number of deaths, numerical unit.
17. TotalPositiveCases: total count of positive cases, numerical unit.
18. TestsPerformed: the number of performed tests, numerical unit.

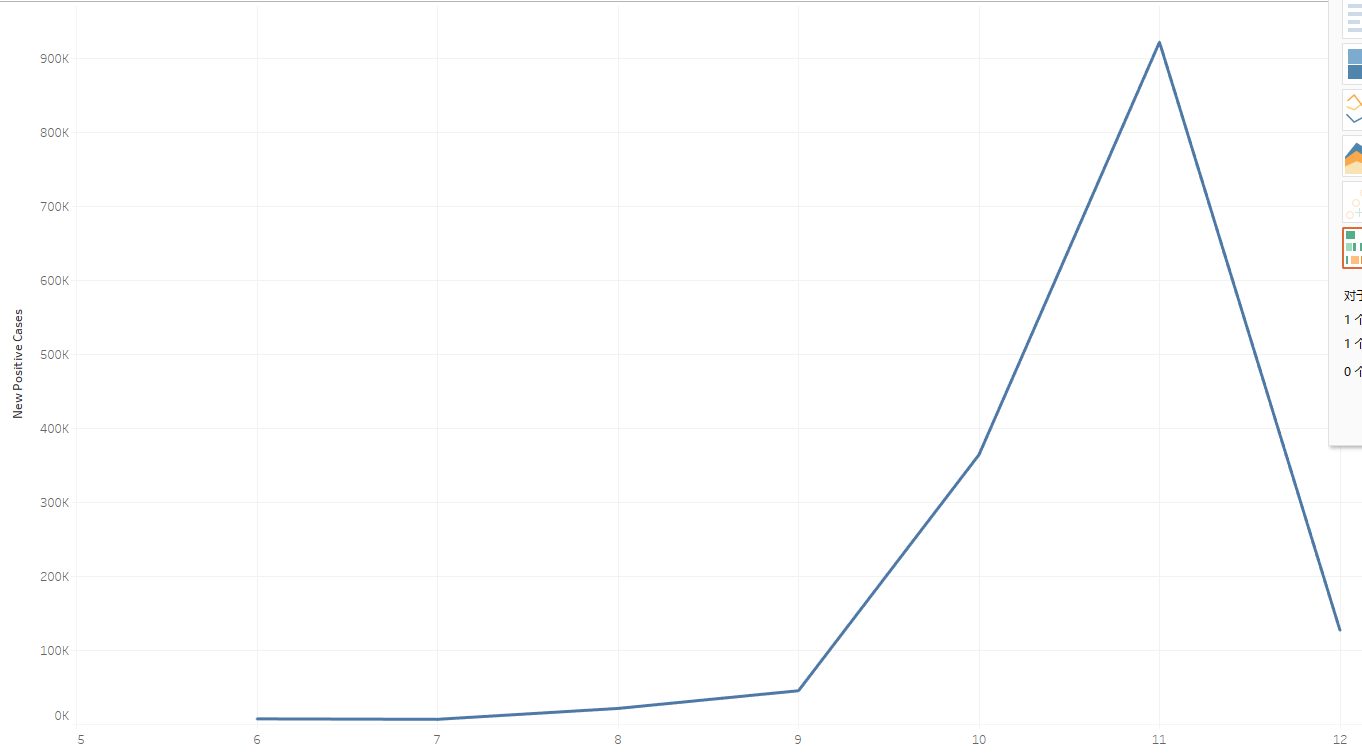
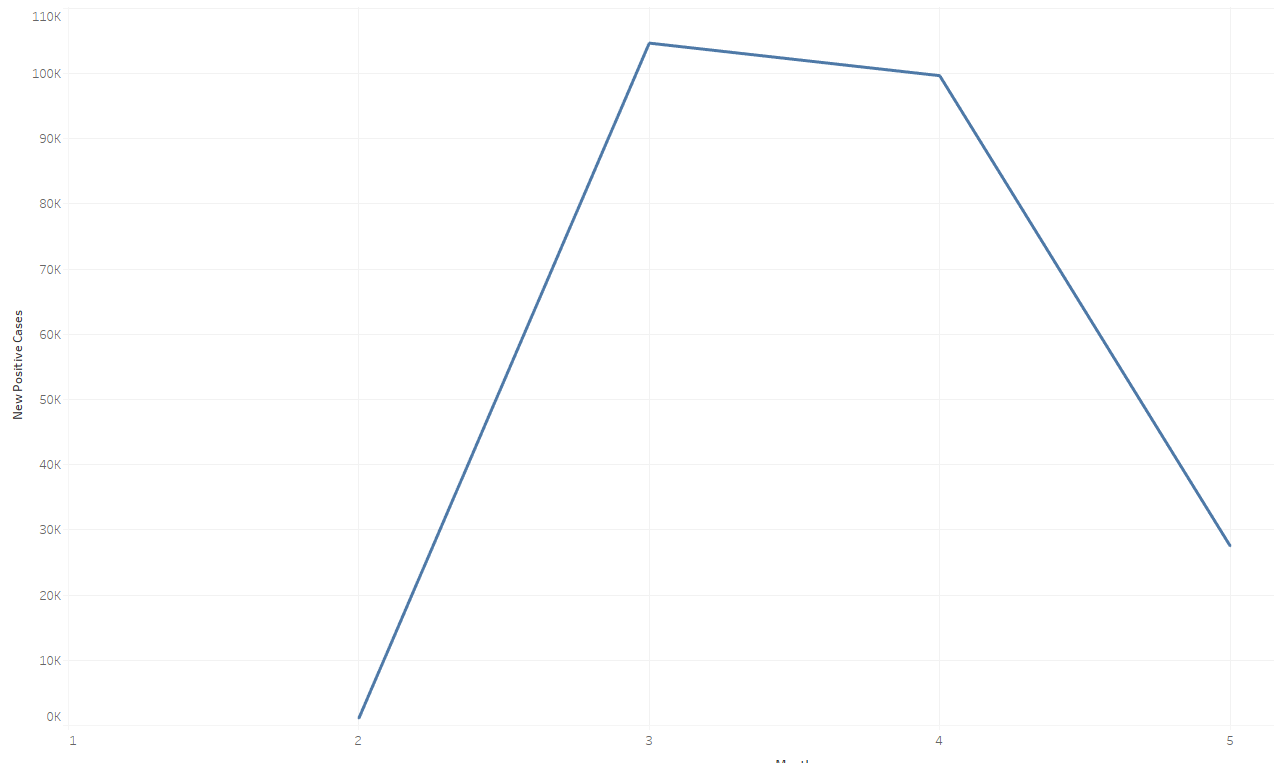
**Extreme and outliers:**

Extremes and outliers exist in some attributes, and they still need to be considered. Because all extreme and outlier data are the actual number of the patients, it depends on the number of newly detected patients, and it cannot be replaced by other values like mean value.

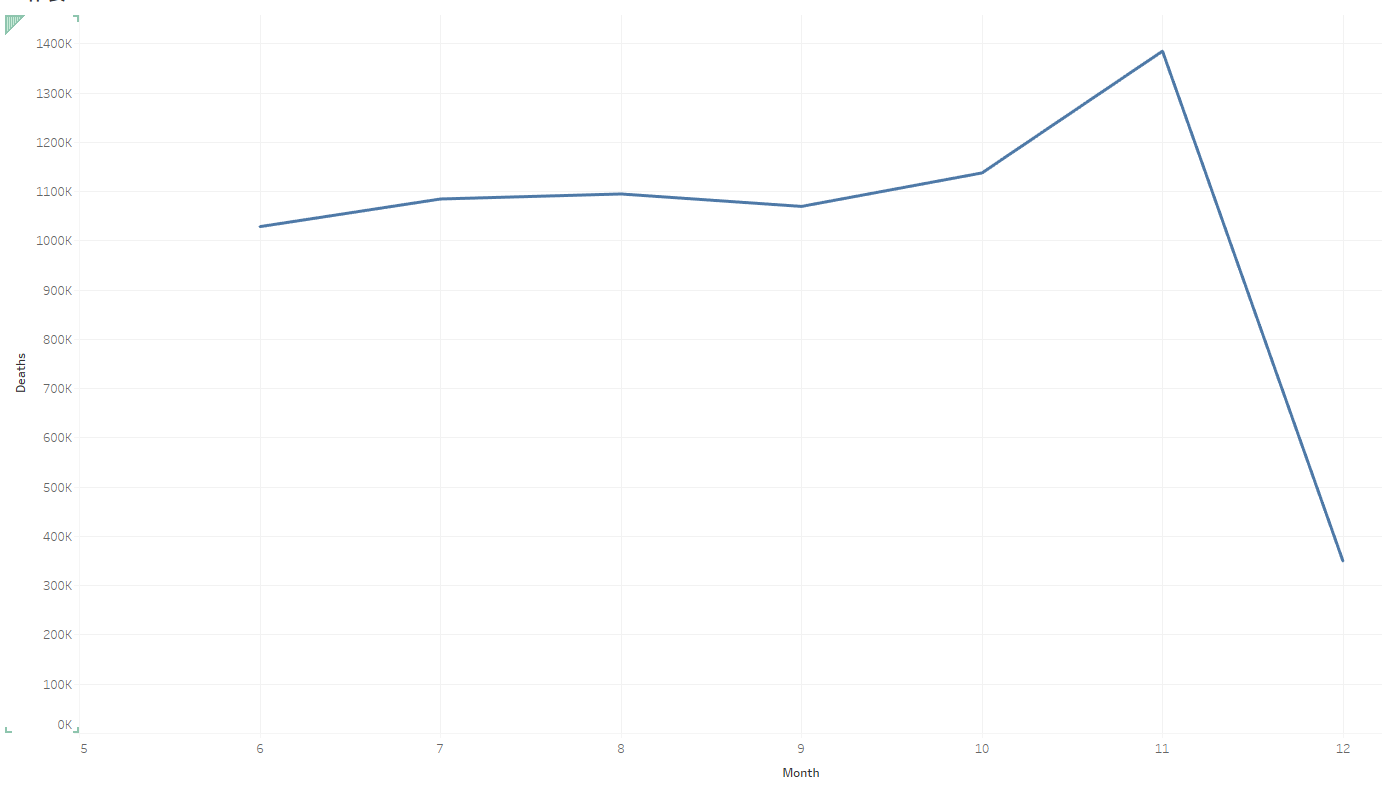
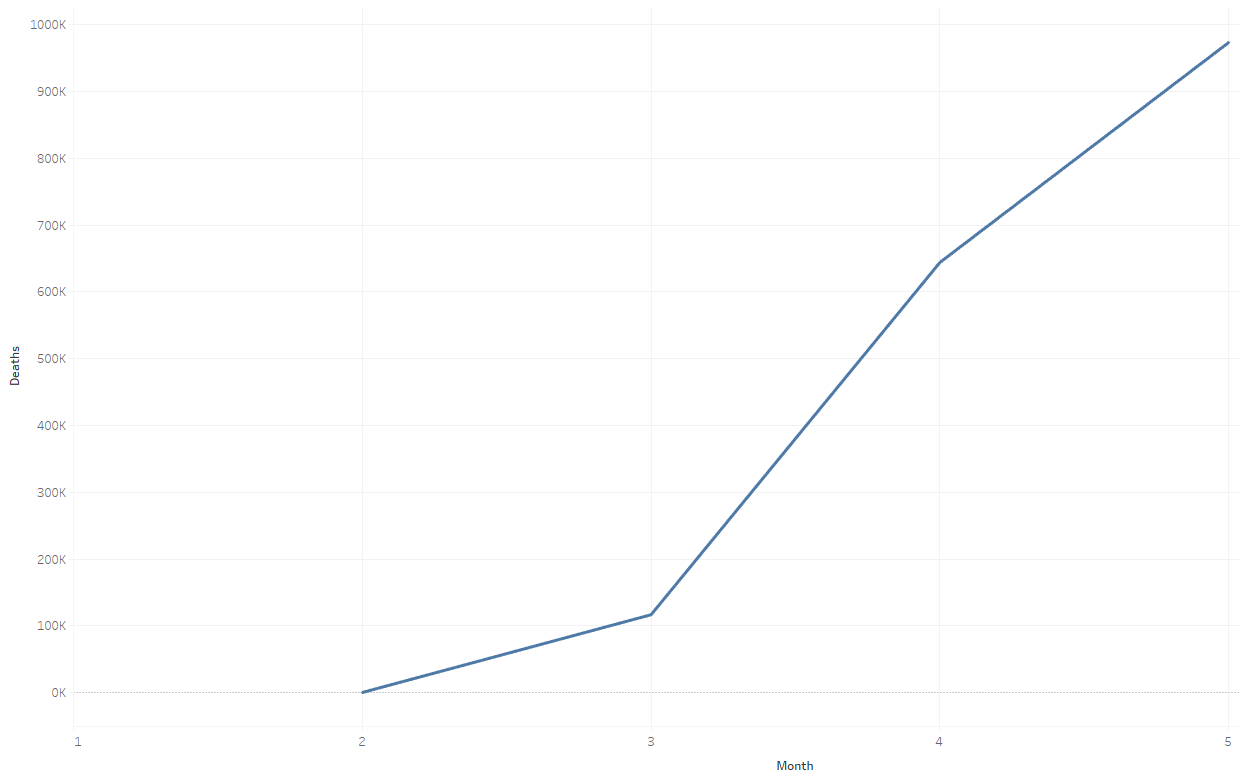


## 2.3 Explore the data

Some graphs are used to visualize the raw data from the two datasets. And they can analyze the data more intuitively and easily. The relative attributes that need to be figured are new positive cases, recovered, deaths, and tests performed. These aspects can well reflect the changes in the epidemic situation over time

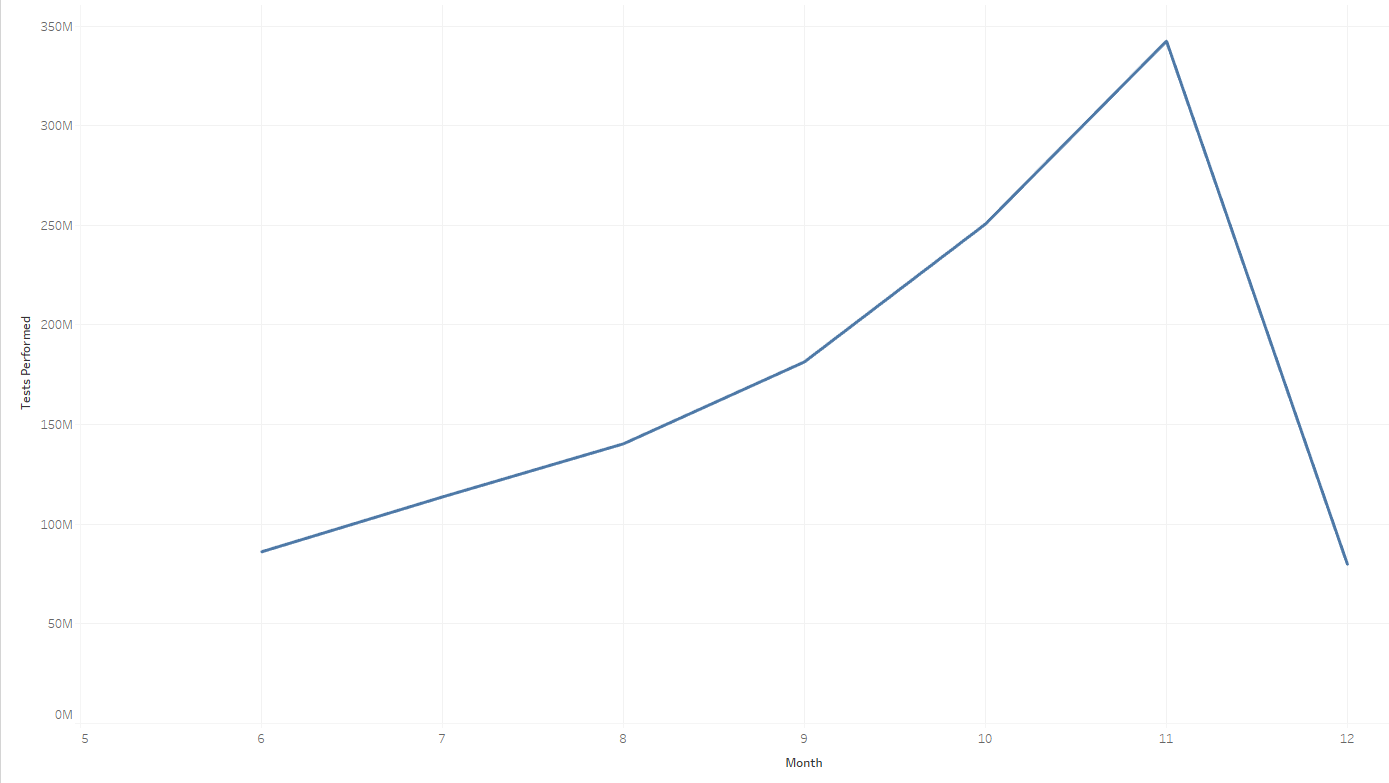
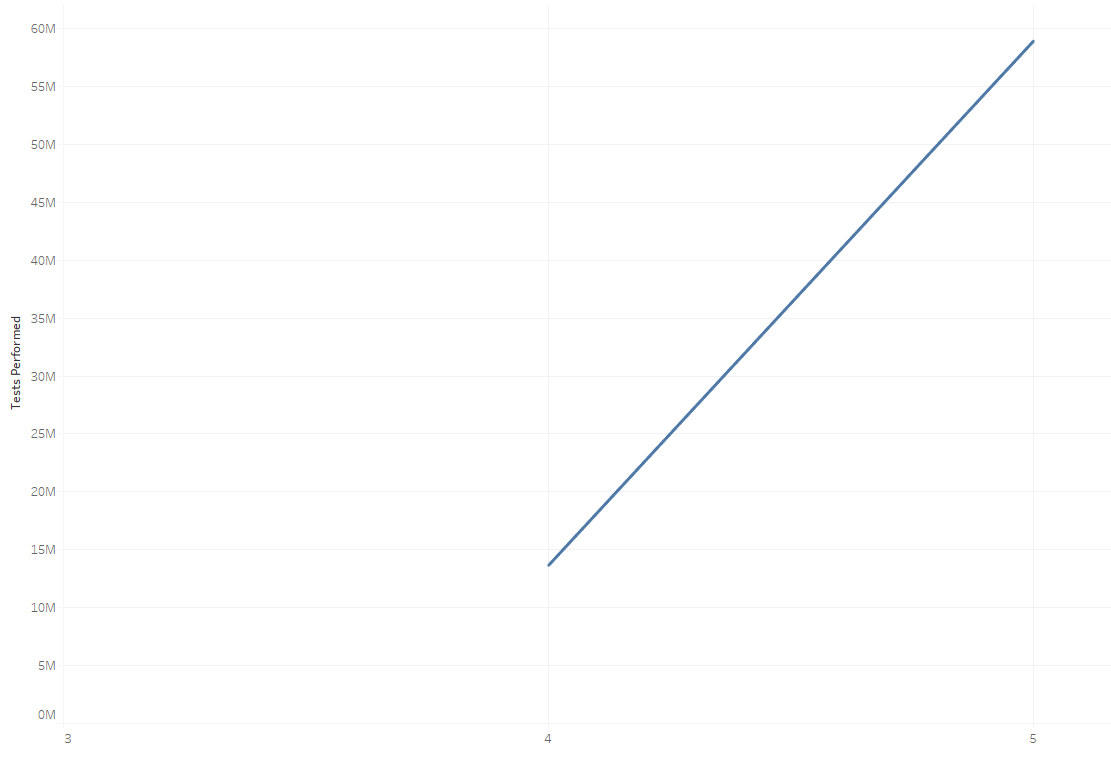


The above two screenshots show a linear relationship between new cases and time. The number of new cases increased sharply from February and gradually decreased from March to September. But after September, the second outbreak came and was brought under control a month later.



With the outbreak of the epidemic, the number of cured people has been increasing, and it increased significantly in November.

With the epidemic outbreak, the number of deaths from Covid-19 has been increasing and increased significantly in November. And the number of deaths is almost stable at about one-tenth of the number of cured



Before September, the virus infection was controlled at a low level, but for some reason, the virus broke out again in September. Italy increased the number of tests, but the number of new infections and deaths increased significantly. We can see from the screenshot that the Italian government did not conduct the test before March. Until the outbreak of the epidemic, the test was gradually promoted. The number of tests was reduced after the epidemic was brought under control.

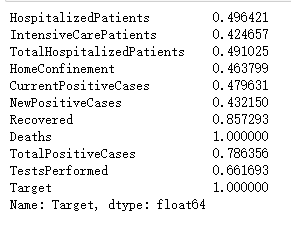


This picture shows the provinces where infected people have been found. However, the data source provides the longitude and latitude of the center of the province. So we can't do a detailed statistics of the specific location of infection cases. And because it can only be detected in the hospital or fixed location, it is unrealistic to analyze the particular discovery location.

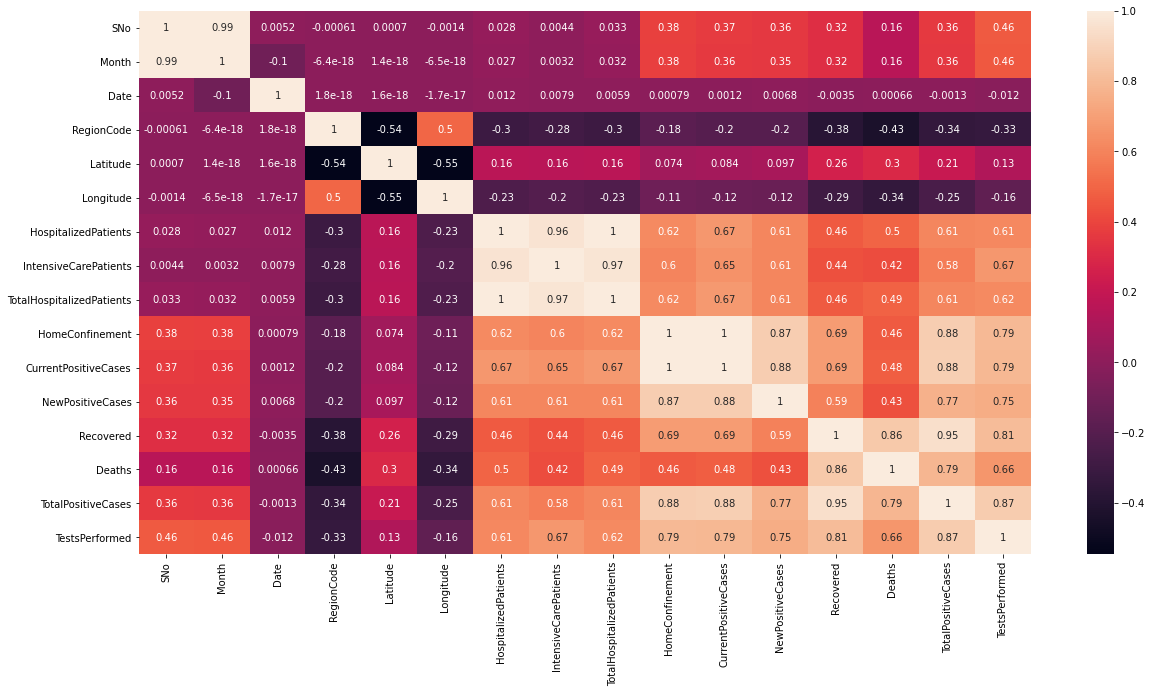
To understand the correlation of the values inside the dataset. I use the corr() function. Which is shown below:



And here is the result table:

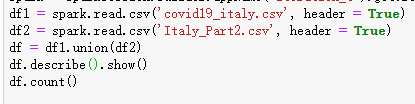


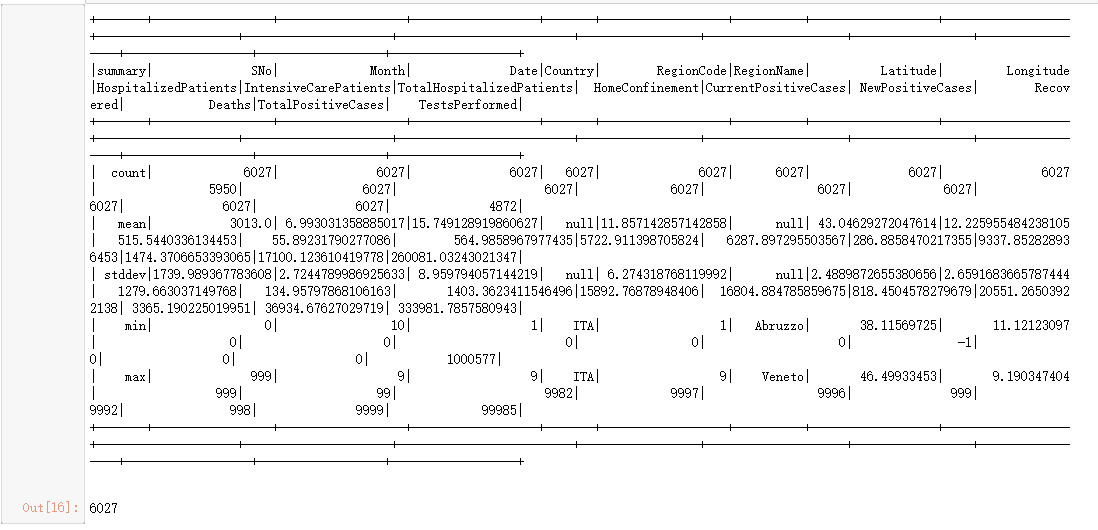
However, the console only provides relevant figures, which is not easy to analyze. So I tried to make a chart to show the correlation by the depth of the color:



## 2.4 Verify the data quality

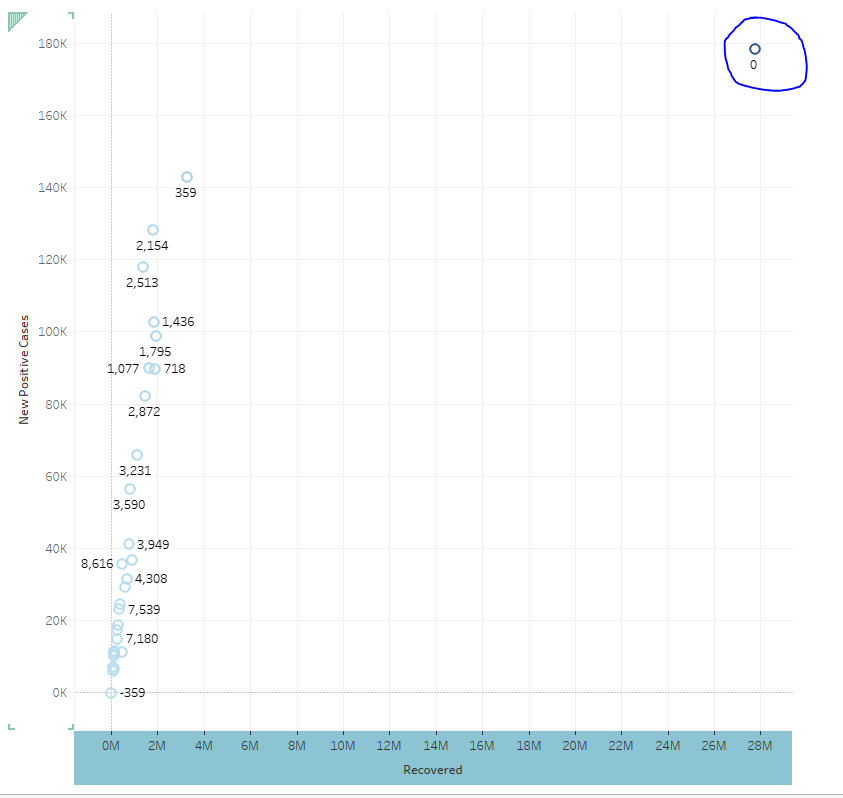
To verify the data quality, the ‘df.describe().show()’ function is used here to present the information of the datasets.

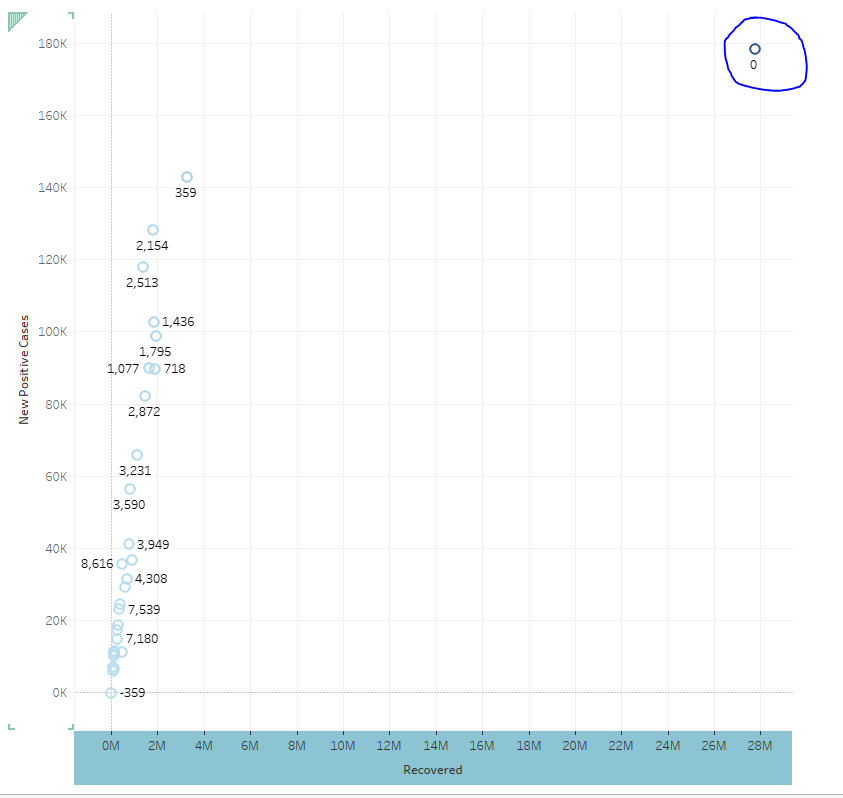




The quality of this data is very high, with almost no errors, missing values. The apparent wrong values showed in the diagram are NewPositiveCases and tests performed. However, in the first few months without testing, the values of TestPerformed were all missed. The missing data were recorded as null values, which will be handled in the following step.

**Null values:** as shown in the above screenshot, There are some null values inside the dataset. It may because that there are some missing values during the data collection part.

**Missing values:** missing values exist in the ‘testperformed’ value. The possible reason is the test are not broadly performed during the first few months. 



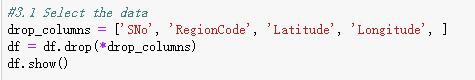
**Extremes and outliers** exist in some attributes, and they still need to be considered. Because all extreme and outlier data are the actual number of the patients, it depends on the number of newly detected patients, and it cannot be replaced by other values like mean value.

# Step 3 Data preparation

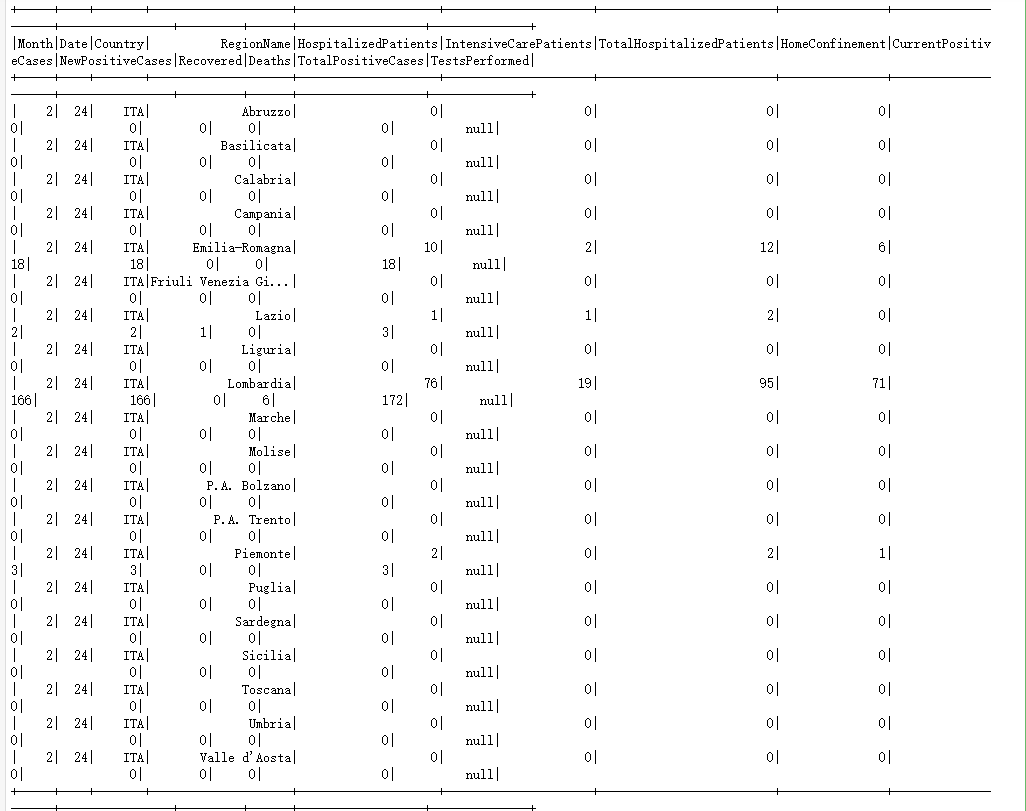
## 3.1 Select the data

This topic’s data mining goal is to determine the connection between time and the Covid-19 epidemic situation in Italy. After analyzing all the data, I figure that all the above attributes are essential in the research.

However, Sno, longitude, latitude, and RegionCode are removed from the data since those three attributes are irrelevant to the data mining goal. Detailed geographic information is not helpful for data filtering and analysis of this topic. To finish this goal, drop() function is used here.

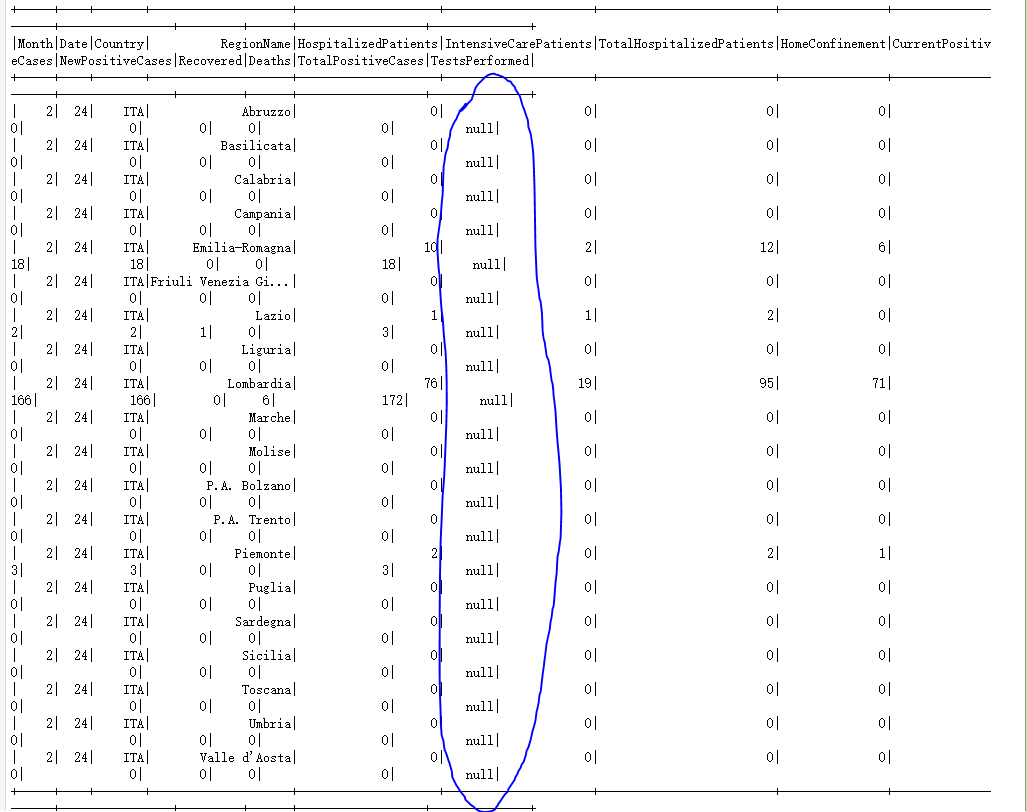


And here is the final cleaned dataset.



## 3.2 Clean the data

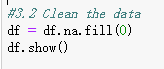
From 3.1, we can know that there are still some null values in the dataset.

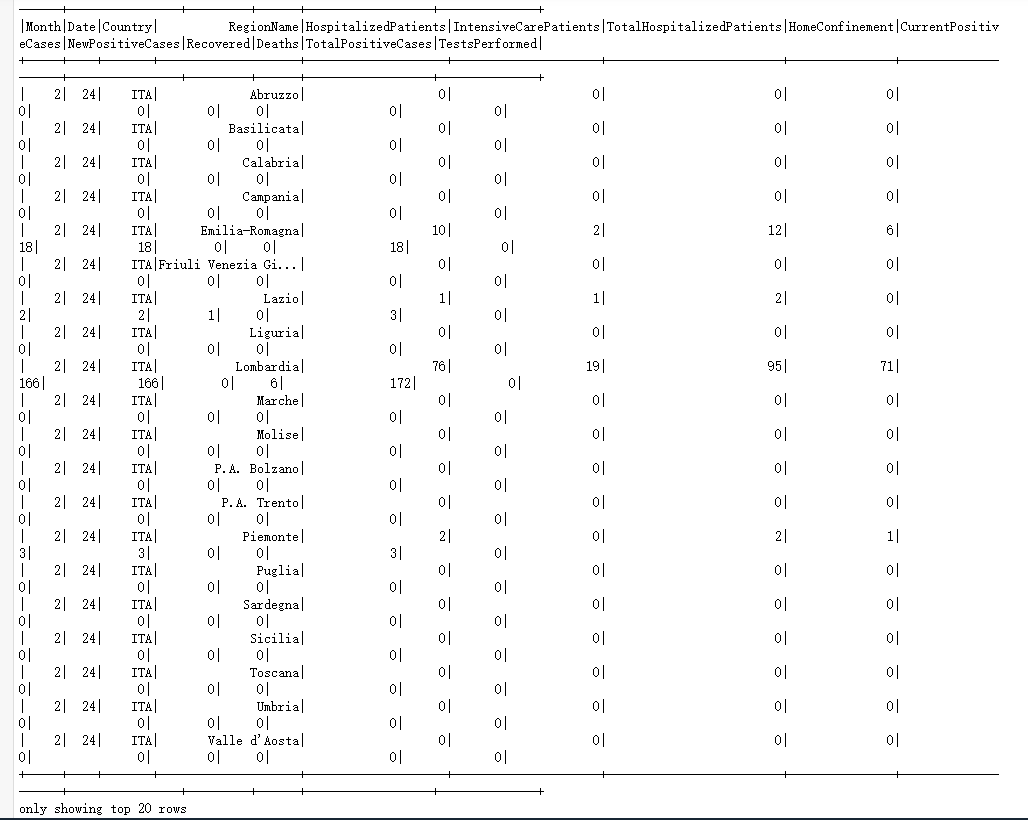


We can see that there are null values in the ‘Testperformed’ value and ‘HospitalizedPatients’ value.

Numeric data cleaning: Considering the missing value is described as null value, and all of the missing values are numeric, I decided to replace those missing values with the number 0.

To finish this section, na.fill(0) function replaces the null value with value ‘0’.

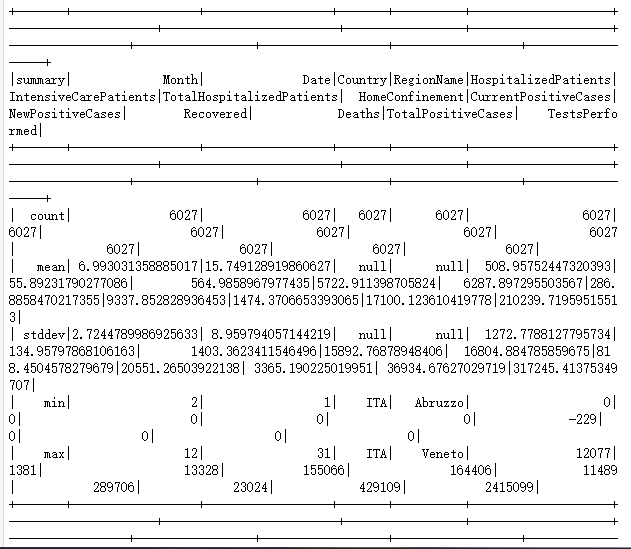


And here is the processed dataset: 

Data in this dataset represents the real-time number of individuals, so all the extremes and outliers cannot be replaced.

Then, df.describe().show() is used.

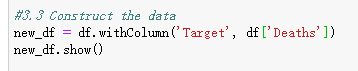




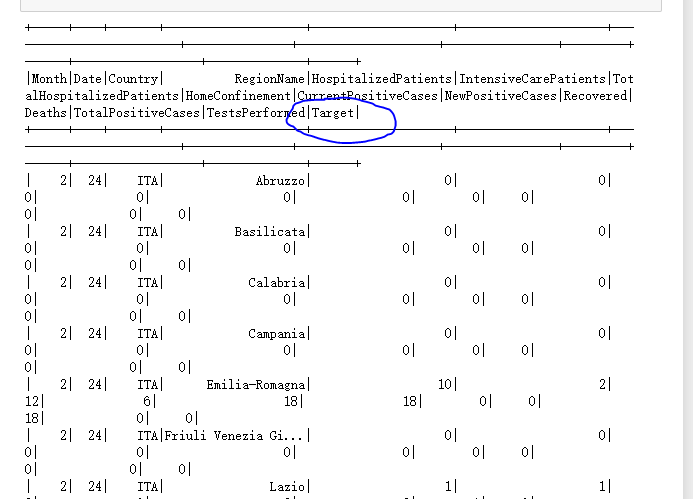
By doing this, we can see the count of all values become 6027, which means that all null values are replaced with 0. The remaining null values in mean and stddev are calculated with country and RegionName. Country and RegionName are strings that cannot be calculated.

## 3.3 Construct the data

To finish this step, withColumn() is performed to set a new column named ‘target.’ This confirms that the target of follow-up data analysis is “Deaths," introducing various algorithms based on this.

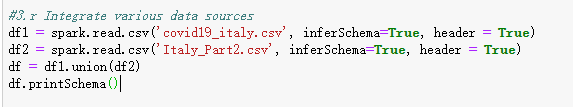


And by doing this, we can see that a new value named “target’ is added to the dataset. And this is the target value for data mining.

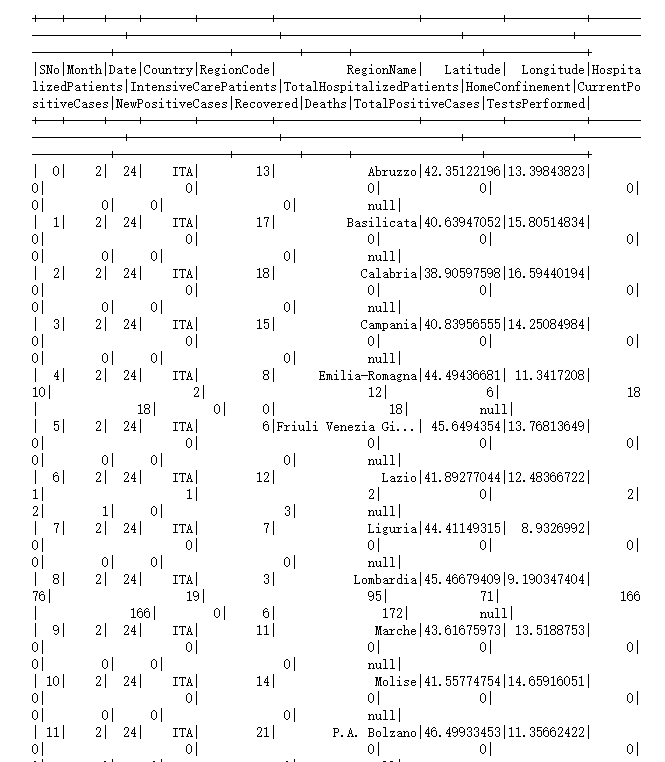


## 3.4 Integrate various data sources

Two independent datasets were integrated using the ‘append() function.



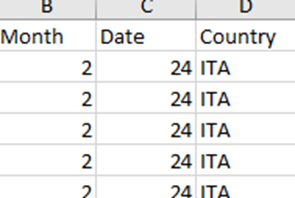
And here is the appended dataset:



Dataset1 contains the patient’s records from Jan to May. Dataset2 includes the patient’s records from June to December. In order to get complete data for analysis, we combine two data sources into one dataset.

## 3.5 Format the data as required

As explained in the above session, I reformatted the ‘date’ attribute. I change the format from mm/dd/yy to separated month, day, and year attributes. All of the collected data are recorded in 2020, so I also deleted the year column.



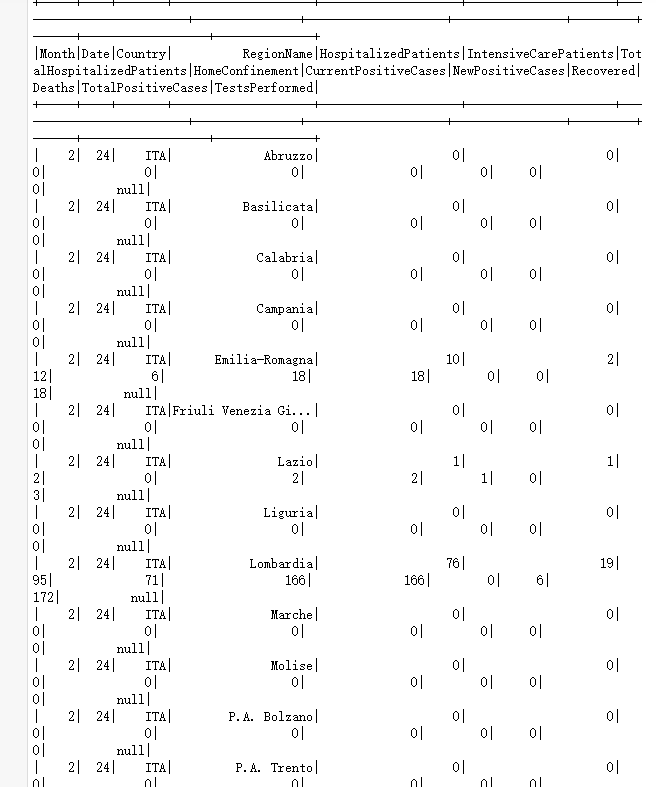
# Step 4 Data Transformation

4.1 Reduce the data

I removed three attributes irrelevant to the topic in the above section: RegionCode, Latitude, and Longitude. RegionCode represents the area where the patient was found. Latitude and longitude represent the Geographic location information. None of the above factors can affect the results. And then, I realize Sno only represents the code number of the patients. And this is irrelevant as well. So the “SNo” value is removed here by using the drop function:

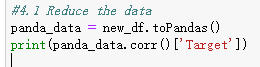


To visualize the dataset, I use the ‘.describe().show()’ function. And here is the processed dataset:

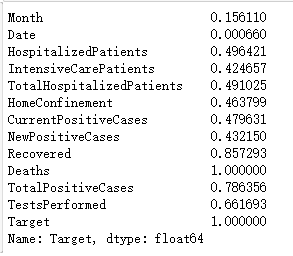


So I decide to remove SNO, RegionCode, Latitude, and Longitude.

Vertically, I need to remove values with low correlation. So panda\_data.corr() is used here:

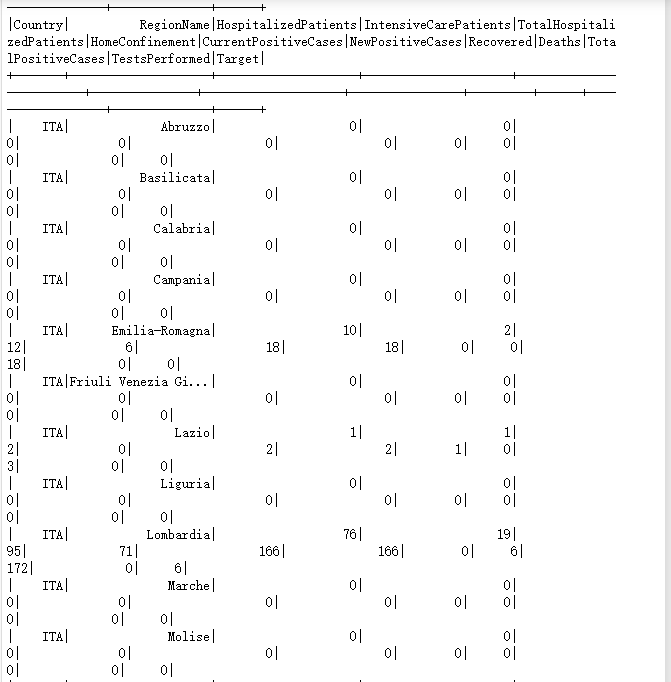


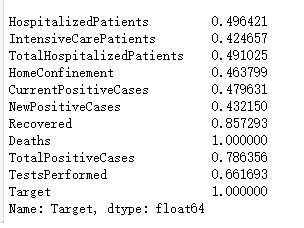
And here the correlation result:



We can find out that value ‘Month’ and ‘Date’ have low correlation with the target from the correlation result. So they should also be removed here.

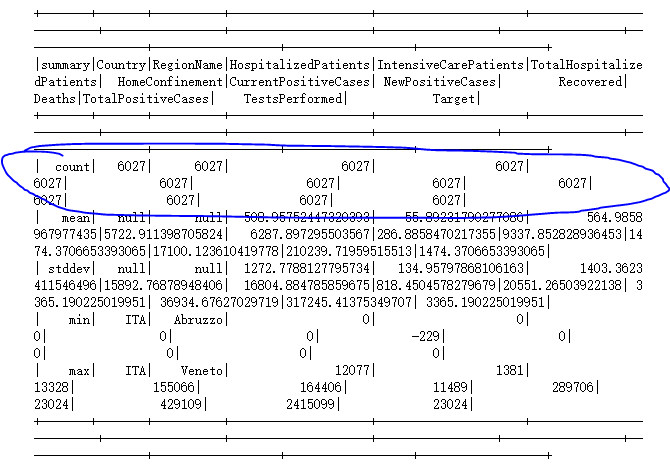


And here is the final result table and correlation result: 

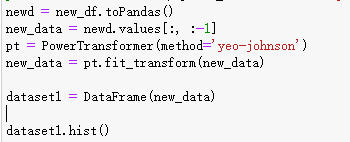


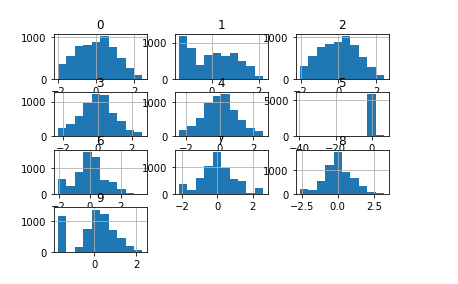
## 4.2 Project the data

I use new\_df.describe().show() to show the description of the dataset. We can see from the figure that our values are 6027. This is the cleaned data. Null values have been replaced by 0. So we know that all the values in this dataset are exactly the same. So there is no need to use the boost method to relocate the number of the values.



I will use statistical transformation to find the logarithm of the distribution due to the uncertainty of some attribute variables in the air. In all, there are two statistical transformation methods: Box-Cox and Yeo-Johnson. For statistical transformation in this project, I shall employ the Yeo-Johnson transformation approach. Because Box-Cox only supports input variables, Because the Yeo-Johnson transformation method does not need the input variables to be precisely positive, it can handle zero and negative values. As a result, the Yeo-Johnson transformation method is best for this project. The statistical transformation method's code and output result are shown below.





# Step 5 Data-Mining Method

## 5.1 Match and discuss the objectives of data mining to data mining method.

In this topic, the data mining objective is the connection between disease spread rate and mortality. The transmission rate of the epidemic can be reflected in the disease's daily increase rate, the number of cured people, and the number of deaths. The more new people said every day, the faster the epidemic spread and the more serious the epidemic

In this data mining object, we can both have the input data and the output data. And we can figure out the connection between input and output, so this is a kind of supervised learning.

**Classification** involves gathering the qualities from various data sets, observing the data value to establish which group the data belongs to, and then predicting the classification label. Rather than classification, numerical prediction is used to estimate the number of deaths. As a result, this approach isn't appropriate for this set of data.

**Clustering** is an unsupervised data mining strategy for categorizing data that is comparable. However, in order to predict the change in death numbers, we must examine the input and output values, so clustering is not an option..

**Regression** is a technique for predicting the outcome of a situation. When other variables are known, this method is generally used to determine the likelihood of a variable's existence. The fundamental purpose of regression is to show how closely two variables in a dataset are related. As a result, we can forecast changes in the number of deaths by using numerical input variables and provided outputs.

The data classification method can not fit this objective. Because target data in this dataset are numeric, but the output value is usually categorical. So classification is not the proper method for this task

Both input and output data are essential in this objective, so the clustering method is also not the proper method.

Thus, we choose the regression method, which can directly reflect the linear relationship between input and output.

## 5.2 Select the appropriate data mining method based on discussion.

As mentioned in the 5.1 session, **the regression method** is chosen as the data mining method. The reason can be described as follows: 1) the objective is supervised learning; 2）both input and output values are numeric.

**Classification**: the attributes in different data sets are gathered together. The data value is observed to determine which group the data belongs to, and then the classification label is predicted. The prediction of the number of deaths is based on numerical prediction rather than classification. Therefore, this method is not desirable. For example, this method is suitable for classifying high, medium, and low risks according to customers' historical credit ratings.

**Clustering**: This is an unsupervised data mining method that categorizes similar data. However, we need to analyze the input and output values to predict Pm2.5 values, so clustering is inappropriate.

**Regression**: Used primarily to determine the likelihood of the existence of a variable when other variables are known. For example, use it to predict a price. More specifically, the main function of regression is the close relationship between two variables in a given dataset.Therefore, we can use numerical input variables and given outputs to predict changes in the number of deaths.

# Step 6 Data mining algorithms selection

## 6.1 Conduct exploratory analysis and discuss

The following three data mining algorithms are selected in this objective:

1. Linear Regression

Linear Regression Algorithm is an algorithm to determine the dependence and correlation between two variables through calculation and analysis. It can be seen from the figure that the objective has a high degree of linear correlation, so the linear regression algorithm is very suitable for this case.

In the Data Mining Objective:

The linear regression algorithm should be the best choice due to the high correlation between the set target value and other related variables.

Advantages:

It is easy to update, more intuitive, and easy to understand to reflect the relevance of data, faster operation speed, and more advantageous when dealing with small databases.

Shortcoming:

Not suitable for nonlinear data, Data accuracy may be low

1. Random Forest

Random forest algorithm is an algorithm that samples different trees to process data. After testing many trees, we can analyze and calculate the forest composed of these trees and draw a conclusion. According to the importance degree, the random forest algorithm sorts all the features, removes some features with low importance degree in the forest, and obtains a new feature set. After the continuous process, we get the final results according to their importance.

In the Data Mining Objective:

Random forest algorithm can also deal with linear correlation data through layer-by-layer comparison to screen out the most relevant two groups of variables.

Advantages:

High accuracy; can process a large number of data; when there are too many missing values in the data set, the random forest algorithm can deal with them.

Shortcomings:

When more decisions need to be made, random forest algorithm is time-consuming

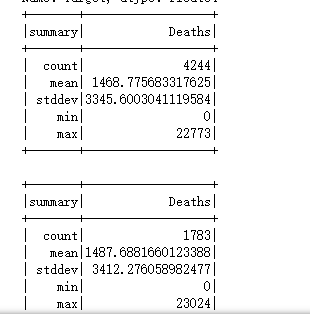
## 6.2 Selecting data mining algorithms based on discussion

After running the three algorithms in weka, The following results are presented. Since all three models’ correlation is high, we can conclude that these three models' data reliability is very high. The degree of linear regression is the highest in the two algorithms. After considering, the following three models are selected to produce the final data mining result: neural network, linear regression, and random forest.

## 6.3 Build/Select appropriate models and choose relevant parameter

First, I split the data into 70% and 30% as train data and test data repectively.





Model 1: linear regression

I set featuresCol as ‘features.’ Then, I set labelCol as ‘Deaths’, which is the value we need to predict. Then, predictionCol is set as ‘prediction.



My prediction target is ‘Deaths’, and the input values are 'HospitalizedPatients', 'IntensiveCarePatients', 'TotalHospitalizedPatients', 'HomeConfinement','CurrentPositiveCases', 'NewPositiveCases', 'Recovered', 'TotalPositiveCases', 'TestsPerformed'

Model 2: Random Forest



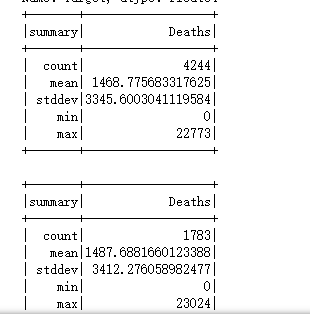
I set featuresCol as ‘features’. Then, I set labelCol as ‘Deaths’, which is the value we need to predict. My prediction target is ‘Deaths’, and the input values are 'HospitalizedPatients', 'IntensiveCarePatients', 'TotalHospitalizedPatients', 'HomeConfinement','CurrentPositiveCases', 'NewPositiveCases', 'Recovered', 'TotalPositiveCases', 'TestsPerformed'

# Step 7 Data Mining

## 7.1 Create and justify test design

Test model

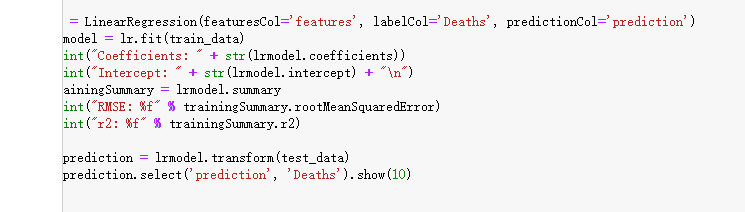
The dataset is usually set 70/30 separately to avoid overfitting. 



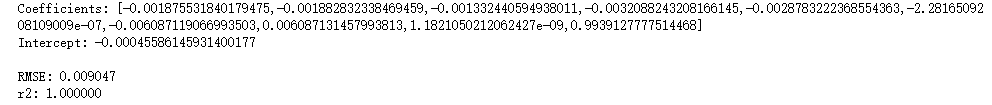
## 7.2 Conduct Data Mining

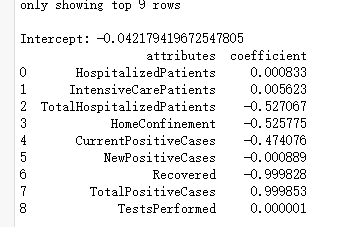
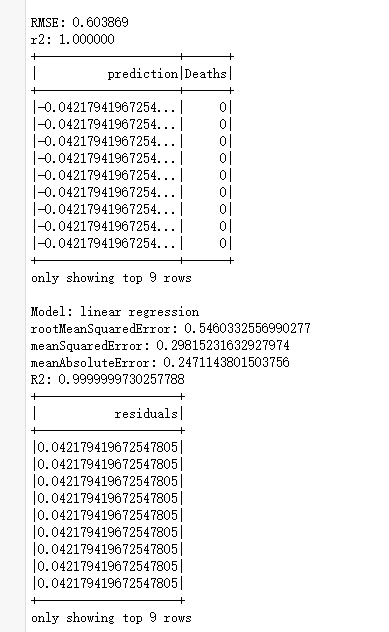
The results of the above three algorithms are in the acceptable range, and the accuracy of the Linear Regression algorithm is the highest, reaching almost 1.

**Model 1: Linear regression**

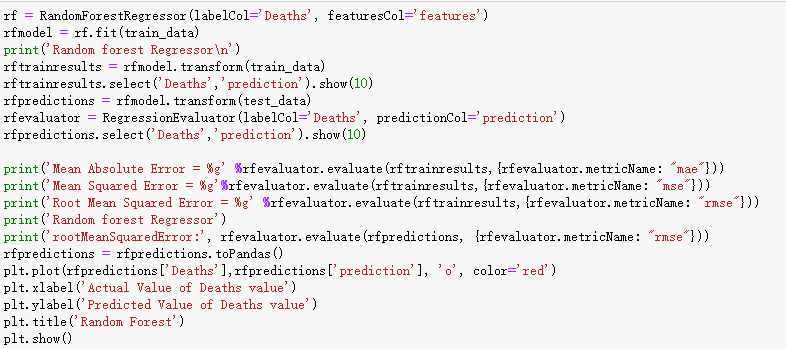




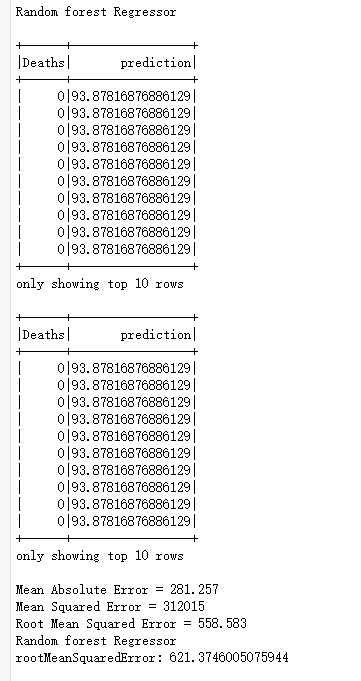
And here is the presented result: 



**Model 2: Random Forest**



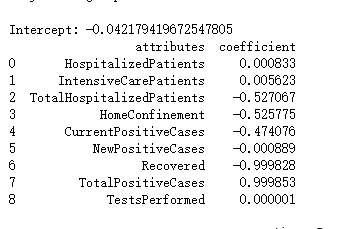
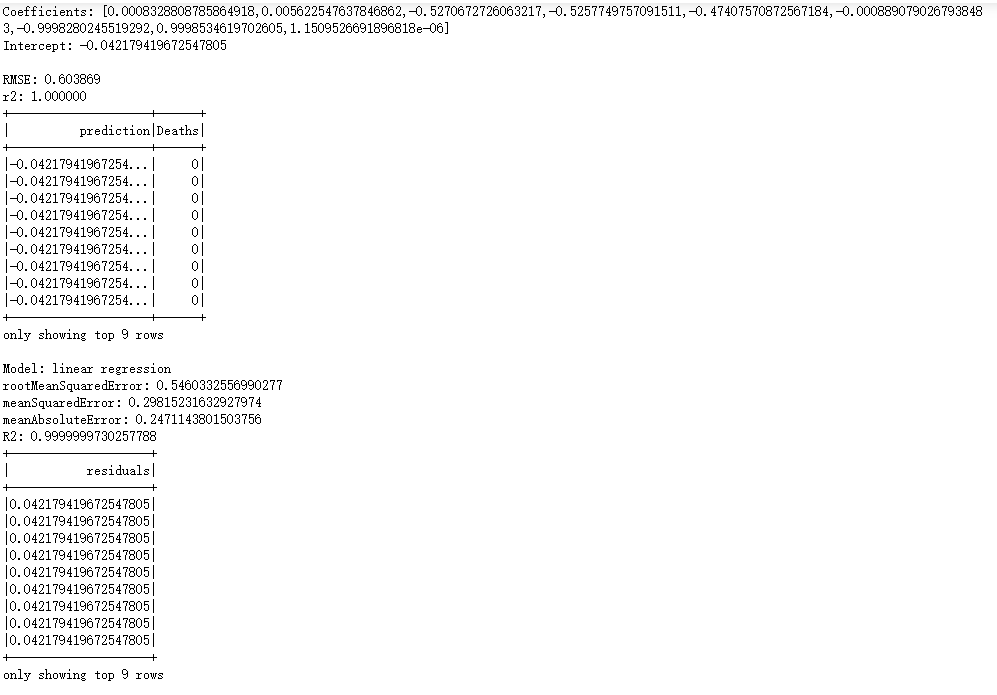
And here is the presented result:

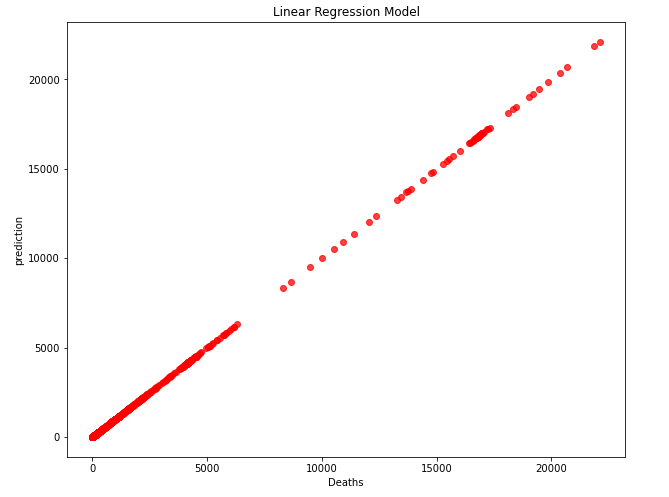


## 7.3 Search for Patterns

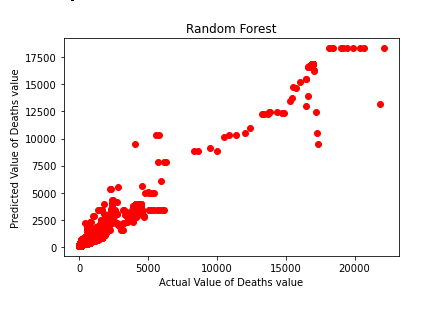
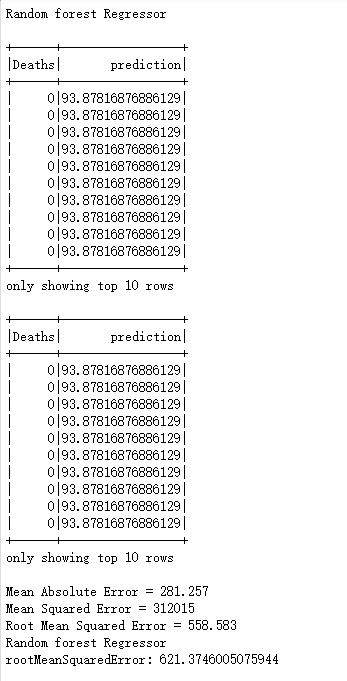
The result of the three data mining algorithms are shown below:

Model 1: Linear regression





Model 2: Random Forest



# Step 8 interpretation

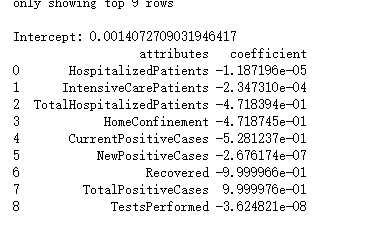
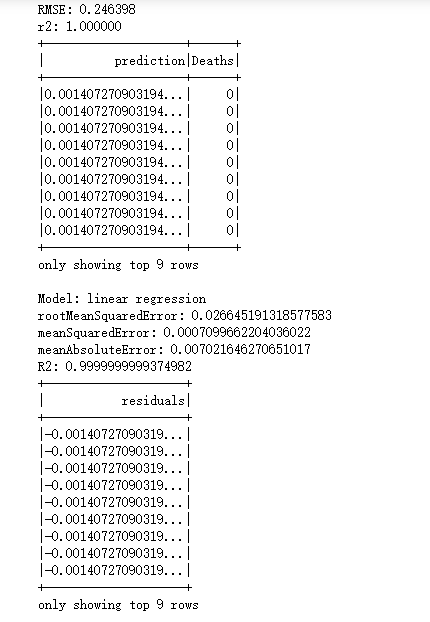
## 8.1 Study and discuss the mined pattern

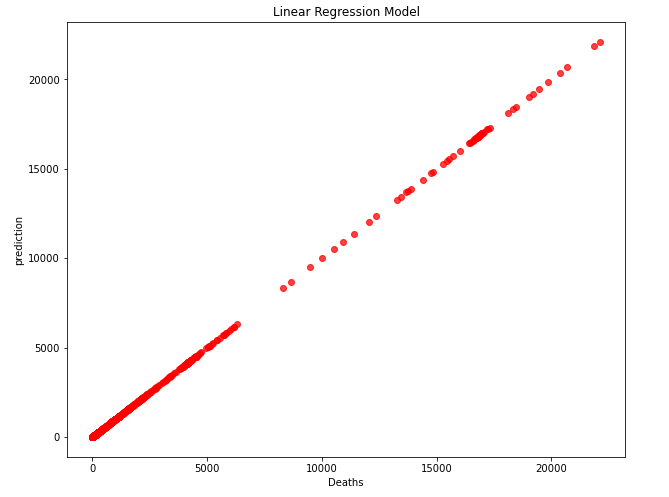
1. Because some algorithms do not support blank values, replacing null values with integer 0 in the clean data session can expand the selection range of algorithms.

The auto-numeric function shows that this objective has a high correlation of 0.99. the data should be absolutely linear in the linear regression algorithm. And our data results confirm this conjecture. But different from iteration 3, in iteration 3, the correlation value of the linear regression algorithm is not high. In this iteration 4, the results of data analysis for the same goal confirm that the linear correlation between the number of deaths and the month in this dataset is exceptionally high. Also, due to some reason the deaths in the linear prediction remains 0.

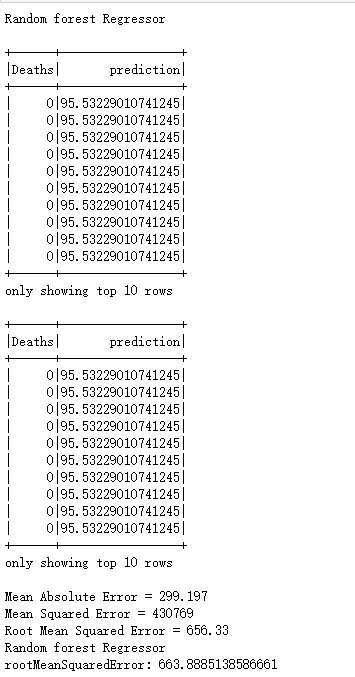
## 8.2 Visualize the data, result, models and patterns:

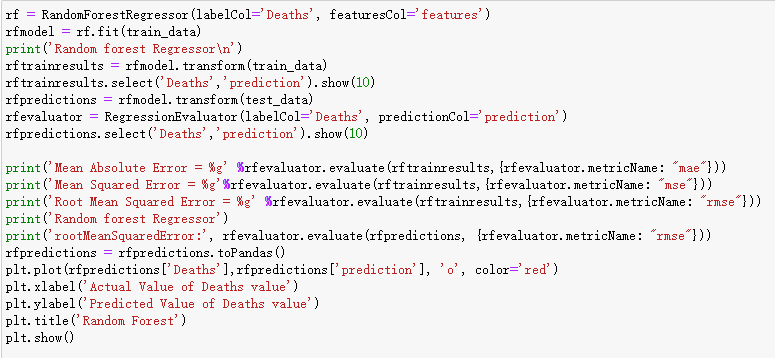
The following screenshots are results for Linear regression:

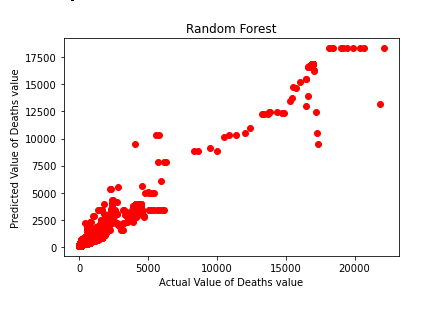




For Random Forest:







## 8.3 Interpret the result, model and patterns

1.Linear aggression model:

Linear regression algorithm is a simple method to verify the correlation between the input value and output value. The Linear regression algorithm is a simple algorithm to verify the correlation between the input value and output value. But in this paper's results, the accuracy of linear regression algorithm is lower than the predicted value. The second-highest proportion of attributes is NewPositiveCases. This is easy to understand because under the condition that the death rate remains unchanged, the more newly infected people are, the more people die of illness.

In this objective linear regression algorithm, the more people recover, the more people die. But I don't think the relationship between the two is direct. From the chart in 2.1, we can see that the increase in the number of rehabilitation patients is a sharp increase in infected patients. With the same infection rate and mortality rate, the more significant the base number, the larger the number of both

Since irrelevant attributes are cleaned, so all attributes like new positive cases are found inside the city. In the Data Mining Objective:

The linear regression algorithm should be the best choice due to the high correlation between the set target value and other related variables.

Shortcoming:

The accuracy of this linear algorithm is much higher than the expectation, which proves that the correlation between deaths and time is close.

2.Random Forest

In the random forest algorithm, we get almost the same conclusion. In the random forest algorithm, the recovered attribute proportion is higher than that in the linear regression algorithm, which means that the correlation between healed patients and deaths is more robust in the random forest algorithm results.

Also, we are surprised that in the random forest algorithm. Test formed accounts for a more significant proportion of the whole table. The ratio of home confinement in this table is also lower than that in the linear regression table

Shortcomings:

When more decisions need to be made, random forest algorithm is time-consuming

Pattern analysis

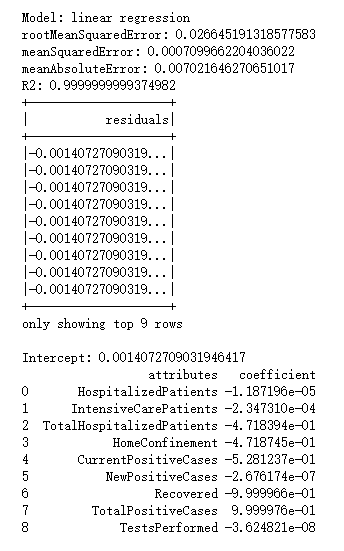
1. Because some algorithms do not support blank values, replacing null values with integer 0 in the clean data session can expand the selection range of algorithms.

2. The auto-numeric function shows that this objective has a high correlation of 1.0. When the value is 1.0, the data should be absolutely linear in the linear regression algorithm. But the running results of the linear regression algorithm show that this algorithm's accuracy is not high. So I think this model is not a perfect linear regression model as shown in the previous auto-numeric function’s result.

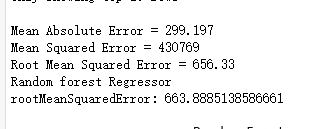
3. Compared with the linear regression algorithm, the neural network algorithm has higher accuracy, up to 97.3%. I think that is because the Neural network can find the optimal solution at high speed. Finding the optimal solution to a complex problem often requires a lot of calculation. By using a feedback artificial neural network designed for a certain issue and giving full play to the computer's high-speed computing ability, the optimal solution can be found quickly.

## 8.4 Assess and evaluate results, models and patterns

Linear model analysis:



Random Forest Analysis:



As shown in the three screenshots above, the correlation coefficient values of both linear regression and neural network are 1. The value of the random forest is also 0.999, which is very high. Compared with the results shown in iteration two, the linear regression is more linear in iteration 2. If the cause of operation error can be ruled out, different software can get different data results for the same target. The linear regression model has almost no mean absolute error, relative absolute error, and root relative squared error. The result also shows a strong positive linear correlation between the number of deaths and the time. While Ann and linear regression can get a high level of data accuracy, the error value of the random forest algorithm is relatively large. The mean absolute error is 35.2386, and the Root mean squared error value is 57.5215, which is significantly higher than the other two algorithms.

ANN algorithm is the most time-consuming algorithm, while random forest and linear regression algorithms are very time-consuming. The following three screenshots record the processing time of ANN, linear regression, and random forest algorithm. This also proves that although ANN is time-consuming, it has higher accuracy than the random forest algorithm. The linear regression algorithm is unique in this data set. Because there is an elevated linear relationship between the selected target and variables, the linear regression algorithm takes the shortest time and has the highest accuracy. Although the procession time of the random forest algorithm is relatively straightforward, compared with the other two algorithms, the error is also higher.



## 8.5 Iterate prior step

Step 1. Iterate business/situation understanding

Step 2 data understanding

Nothing changed. Because the dataset collected by Kaggle won't change in the following iteration

Step 3. Data preparation

I delete SNo in the dataset because I found it irrelevant to the following steps

Step 4 Data Transformation

Nothing changed

Step 5 Data-mining method selection

Nothing changed

Step 6 Data-mining algorithms selection

I choose random forest instead of random trees in the second loop. But all those three algorithms I decided cannot generate a visualized tree because the relation with the two values is linear.

Step 7 Data-mining

Nothing changed

Reference List

1. Richter, F. (2021, Feb 4). World Economic Forum. Retrieved from World Economic Forum: <https://www.weforum.org/agenda/2021/02/covid-employment-global-job-loss/>

2. News, G. (2021, 4 6). Google News. Retrieved from <https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US%3Aen>

3. ReviewBusinessHarvard. (2020). Retreiving Source: https://hbr.org/2020/03/lessons-from-italys-response-to-coronavirus

4. Kaggle. (2021). Retreiving Source: <https://www.kaggle.com/>

5. OrganizationHealthWorld. (12-oct-2020). Retreiving Source: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19#:~:text=symptoms>

6. Statista. (2021). Retreiving Source: https://www.statista.com/statistics/1102808/coronavirus-cases-by-status-italy/

7. (n.d.). Retrieved from United Nations: <https://www.un.org/sustainabledevelopment/health/>