Alexandria University
Faculty of Engineering
Department of Electrical Engineering
Communications & Electronics Program

Second Year 2020

Course: Mathematics IX (Probability)

Covid-19 Data Analysis II

Submitted To: Prof.Dr / Yasmine Abouelseoud

Submitted From: Ismail Mohamed Farahat
Section: 2
ID: 51
Date: 16/4/2020

Content:

1.Data Wrangling ----------4

1.1. Info about Covid-19 ----------4

1.2. Gathering the data ----------4

1.3. Cleaning the data ----------4

1.4. Factors that will be studied ----------5

1.4.1. Temperature ----------5

1.4.2. Wind Speed ----------5

1.4.3. Quarantine Days ----------6

1.4.4. Arrivals Number ----------6

1.4.5. Population ----------6

1.4.6. Health Efficiency Index ----------6

2. Factors Analysis for Every Country ----------7

2.1. China ----------7

2.1.1. Temperature ----------8

2.1.2. Humidity ----------8

2.1.3. Wind Speed ----------8

2.1.4. Quarantine Days ----------8

2.2. USA ----------10

2.2.1. Temperature ----------11

2.2.2. Humidity ----------11

2.2.3. Wind Speed ----------11

2.2.4. Quarantine Days ----------12

2.3. Spain ----------13

2.3.1. Temperature ----------14

2.3.2. Humidity ----------14

2.3.3. Wind Speed ----------14

2.3.4. Quarantine Days ----------15
2.4. Italy ----------16
2.4.1. Temperature ----------17
2.4.2. Humidity ----------17
2.4.3. Wind Speed ----------17
2.4.4. Quarantine Days ----------18
2.5. The World (All Countries) ----------19
2.5.1. Temperature ----------20
2.5.2. Arrivals Number ----------20
2.5.3. Population ----------20
2.5.4. Health Efficiency Index ----------20
3. Conclusion ----------21
3.1. Temperature ----------21
3.2. Quarantine Days Number ----------21
3.3. Arrivals Number ----------22
4. Jupyter Workspace (ALL CODES) ----------23

1.Data Wrangling

1.1. Info about Covid-19

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and has since spread globally, resulting in the ongoing 2019–20 coronavirus pandemic. Common symptoms include fever, cough, and shortness of breath. Other symptoms may include fatigue, muscle pain, diarrhea, sore throat, loss of smell, and abdominal pain. The time from exposure to onset of symptoms is typically around five days but may range from two to fourteen days. While the majority of cases result in mild symptoms, some progress to viral pneumonia and multi-organ failure. As of 16 April 2020, more than 2.09 million cases have been reported across 210 countries and territories, resulting in more than 139,000 deaths. More than 528,000 people have recovered.

1.2. Gathering the data

In this step, I gathered the data from multiple sources of trusted organizations and websites:

1- **World Health Organization (WHO)**
2- **European Centre for Disease Prevention and Control**
3- **Data world website**
4- **Kaggle website**
5- **Worldometers**
6- **www.accuweather.com**
7- https://www.indexmundi.com/facts/indicators/ST.INT.ARVL/rankings

I used only the data that will be benefit in this report and I used the other sources to make sure that I gathered trusted data.

1.3. Cleaning the data

After gathering the data, we should clean the data first before doing any analysis and that's because the data have null values, repeated values, useless columns and discrete data that should be in one column.

I did clean the data using python programming language and excel to make the data useful in my analysis…but this not the point of this report so I am not going to discuss the techniques of this step.

In the end, we have five datasets that will be used in our analysis,

For the part 1 of the assignment, there are four datasets:

1- COVID-19-daily-data-2020-04-12_china
2- COVID-19-daily-data-2020-04-12_Italy
3- COVID-19-daily-data-2020-04-12_spain
4- COVID-19-daily-data-2020-04-12_USA

For the part 2 of the assignment, there is one dataset:

5- COVID-19-countries-data-2020-04-04_total

1.4. Factors that will be studied

Important note: while our studding of any environmental factor, we are going to use the moving average of this factor and the reason for that is that the cases discovered in a given day doesn't relate to the factor value in this day but related the factor average value in the past days because the virus have an incubation period between 2 to 14 days which means by high possibility the discovered cases in a given day are already infected in the past days.

1.4.1. Temperature

According to doctors and science of biology, temperature have an effect on the DNA & RNA of all living things and non-living things like virus. So, we are going to study temperature effect on the spread of the virus.

We expected to find a strong negative correlation (as temperature increases the virus dies easily reducing the number of cases according to doctors).

1.4.2. Wind Speed

According to scientists, the virus could stay at air for three hours at least and with high speed of the wind there is a possibility that the virus spread rapidly.

We expected to find a strong positive correlation (as wind speed increases the virus will spread easily causing increasing of cases number).

1.4.3. Quarantine Days

The purpose of quarantine procedures is to flatten the curve of the virus spread (makes the spread goes more linearly instead of exponentially gross) not to reduce the cases so we are going to study this factor to see the possible relations.

We expected to find a strong positive or negative correlation (as quarantine days increases the virus spread will slow down and the curve of spread will be linear more than exponential but the number of cases will stay increases or decreases according to other factors).

1.4.4. Arrivals Number

According to logic, the virus began at china so the virus spreads globally by the travelers around the world practically the travelers by air so there is a possibility that the arrivals numbers that reached every country this year have an effect on the virus spread in this country.

We expected to find a strong positive correlation (as arrivals number increases the virus will spread easily causing increasing of cases number).

1.4.5. Population

Large population doesn't have an effect on the virus spread but this factor is going to give us an indication how fast the virus spreads so we need to study this factor to know how dangerous the spread of the virus will be.

We expected to find a positive correlation.

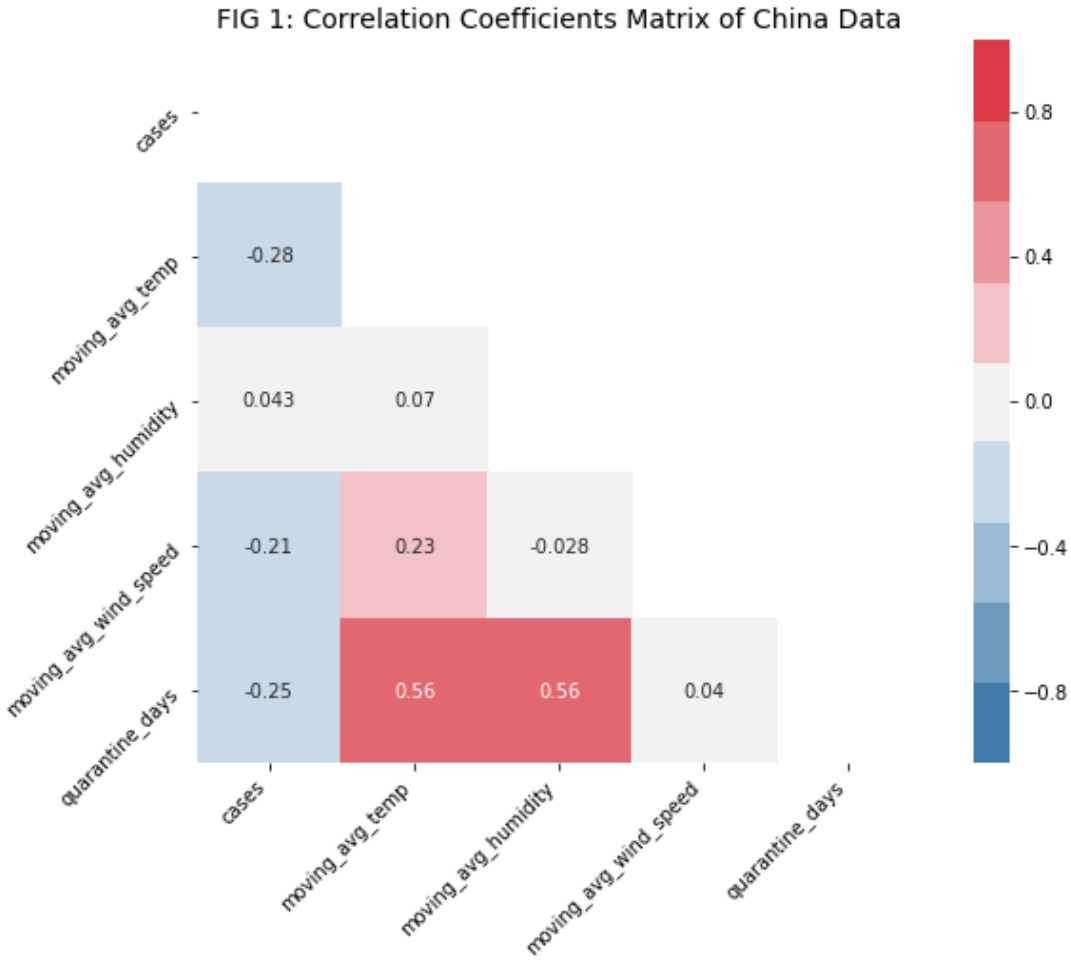
1.4.6. Health Efficiency Index

Health Efficiency Index is an indication for how good the health system is in a country. So, we need to study the effect of this factor on death cases number and also active cases to estimate world position in this crisis and to know if the world has the ability to stop the virus or not.

We expected to find a strong negative correlation. (with death cases number)

2. Factors Analysis for Every Country

2.1. China



	cases
cases	1.000000
moving_avg_temp	-0.279222
moving_avg_humidity	0.042989
moving_avg_wind_speed	-0.214802
quarantine_days	-0.252984

Correlation Coefficients between the number of cases and some possible factors for China

2.1.1. Temperature

Correlation coefficient of temperature is - 0.28 (negative, weak correlation).

Temperature has negative correlation which meets with our expectation but we didn't expect to see a weak correlation value so according to this data, temperature doesn't have a great effect on reducing the number of cases in China but we need to study the same factor on other countries to be sure.

2.1.2. Humidity

Correlation coefficient of humidity is 0.04 (positive, negligible correlation). There is no correlation between humidity and number of cases in China.

2.1.3. Wind Speed

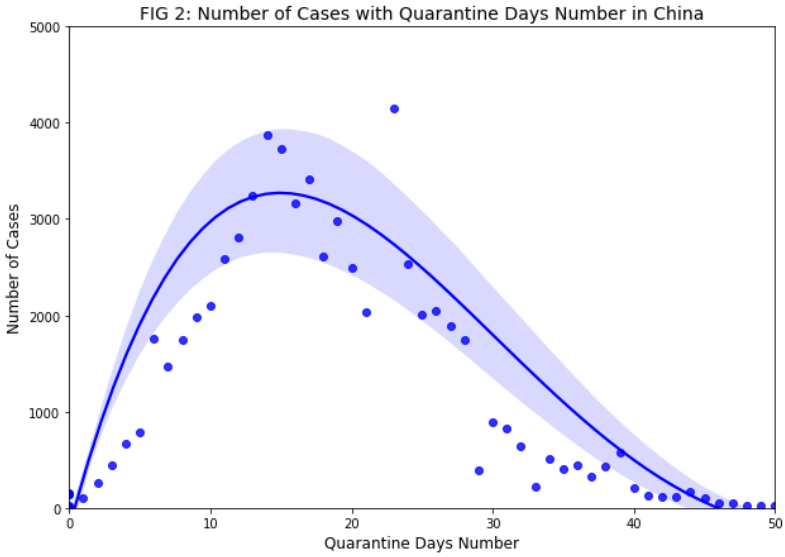
Correlation coefficient of wind speed is - 0.21 (negative, negligible correlation). There is no correlation between wind speed and number of cases in China.

2.1.4. Quarantine Days

Correlation coefficient of quarantine days number is - 0.25 (negative, weak correlation).

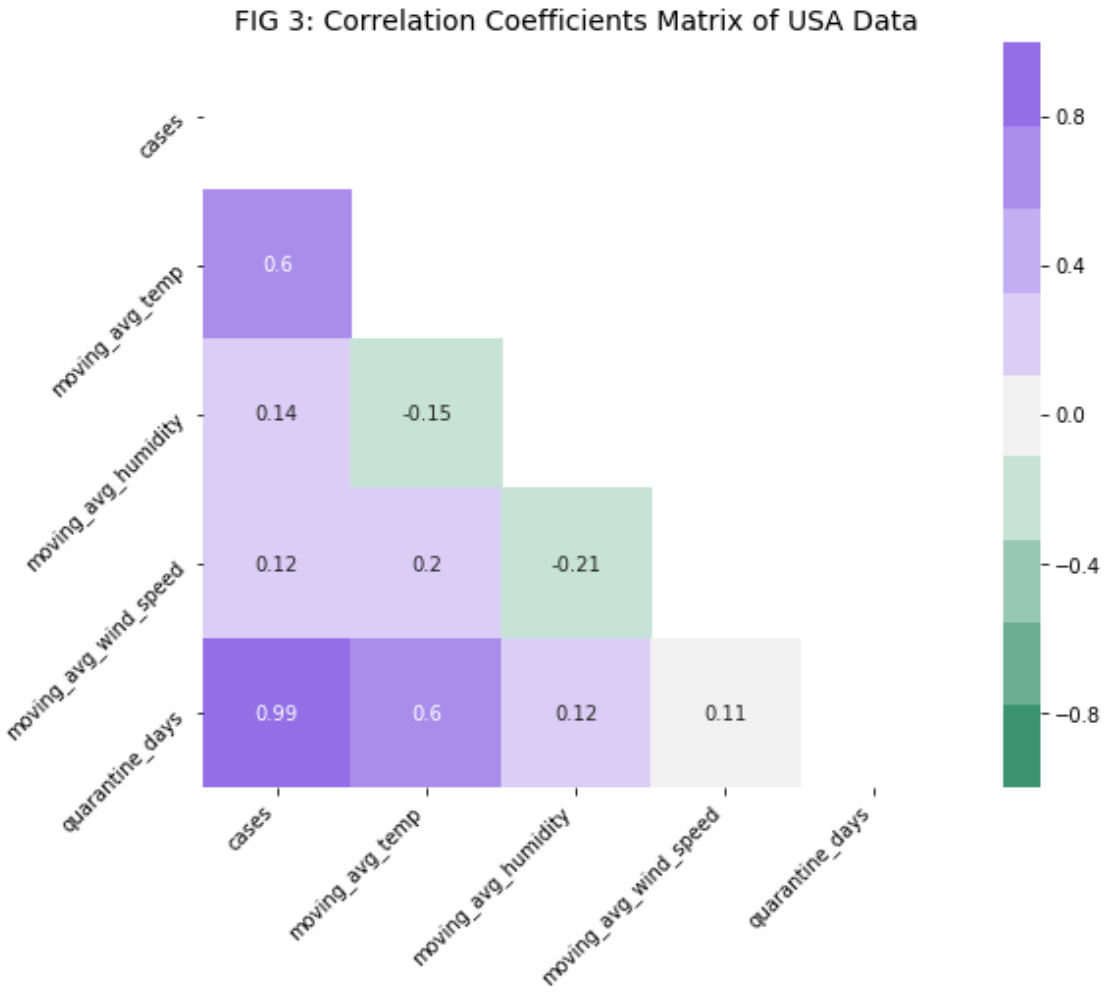
Quarantine days number has negative correlation which meets with our expectation but with a weak correlation value so according to this data, quarantine days number doesn't have a great effect on reducing the number of cases in China but we need to plot the scatter plot to see if there is another non-linear relation to be sure from our analysis.

To make sure if there is another relation or not, we need to plot the scatter plot of the data with regression model.



As we could see from the figure, the relation is close to be quartic curve (or bell curve) between the quarantine days number and the number of cases which is needed to flatten the curve and we also could see that at first days of quarantine, the cases number increases fast but as the quarantine days increases the number of cases decreases.

2.2. USA



	cases
cases	1.000000
moving_avg_temp	0.604013
moving_avg_humidity	0.142973
moving_avg_wind_speed	0.118098
quarantine_days	0.986090

Correlation Coefficients between the number of cases and some possible factors for USA

2.2.1. Temperature

Correlation coefficient of temperature is 0.6 (positive, moderate correlation).

Temperature has positive correlation which doesn't meet with our expectation so according to this data,

There are two possibilities:

1- Temperature doesn't have any effect on the number of cases in USA like China.
2- Temperature has not increased yet in USA to the level that kill the virus so at the moment the temperature doesn't have any effect on stopping the spread of the virus but in the next months could have a great effect on slowing the virus spread

So, we need to study the same factor on other countries to be sure which possibility is right. (**Note:** In China data, we assumed only the first possibility and that is because the virus spread had already stopped in China)

2.2.2. Humidity

Correlation coefficient of humidity is 0.14 (positive, negligible correlation). There is no correlation between humidity and number of cases in USA like China.

2.2.3. Wind Speed

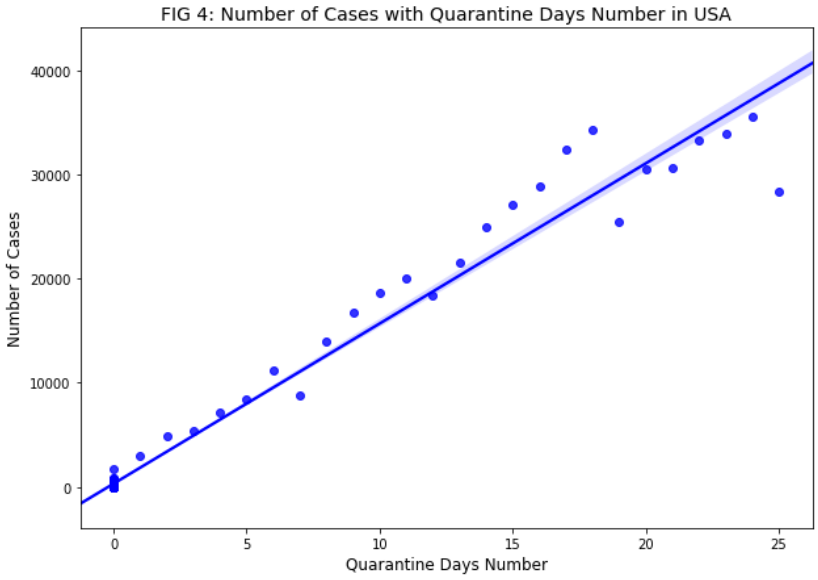
Correlation coefficient of wind speed is 0.12 (positive, negligible correlation). There is no correlation between wind speed and number of cases in USA like China.

2.2.4. Quarantine Days

Correlation coefficient of quarantine days number is 0.99 (positive, very strong correlation).

Quarantine days number has positive correlation with a very strong correlation value which meets with our expectation. so according to this data, quarantine days number have a great effect on reducing the number of cases in USA (convert the growth rate from exponential to linear).

To make sure that there is a linear relation, we need to plot the scatter plot of the data with regression model.

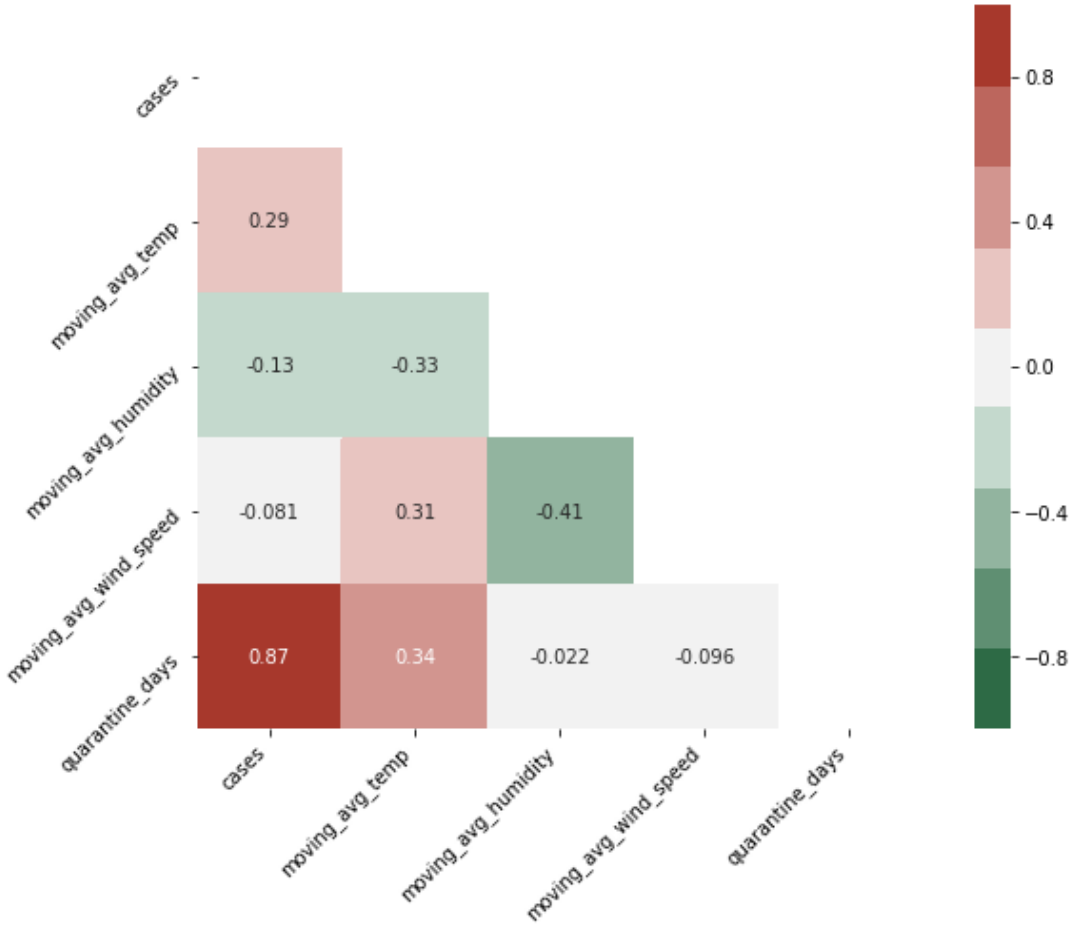


As we could see from the figure, the relation is completely linear between the quarantine days number and the number of cases which is needed to flatten the curve (converting the curve from exponential growth to linear one) and we also could see that at zero days of quarantine , the cases number increases very fast.

We could conclude that the relation between the quarantine days number and the number of cases could be non-linear (quadratic curve) like china (that the virus spread had already stopped on it) and could be linear relation like USA (which at the moment still at the beginning of the virus spread)

2.3. Spain

FIG 5: Correlation Coefficients Matrix of Spain Data



	cases
cases	1.000000
moving_avg_temp	0.290919
moving_avg_humidity	-0.131268
moving_avg_wind_speed	-0.081034
quarantine_days	0.867385

Correlation Coefficients between the number of cases and some possible factors for Spain

2.3.1. Temperature

Correlation coefficient of temperature is 0.29 (positive, weak correlation).

Temperature has positive correlation which doesn't meet with our expectation so according to this data,

There are two possibilities:

3- Temperature doesn't have any effect on the number of cases in Spain.
4- Temperature has not increased yet in Spain to the level that kill the virus so at the moment the temperature doesn't have any effect on stopping the spread of the virus but in the next months could have a great effect on slowing the virus spread

So, we need to study the same factor on other countries to be sure which possibility is right.

2.3.2. Humidity

Correlation coefficient of humidity is -0.13 (negative negligible correlation). There is no correlation between humidity and number of cases in Spain like USA and China.

2.3.3. Wind Speed

Correlation coefficient of wind speed is -0.08 (negative negligible correlation). There is no correlation between wind speed and number of cases in Spain like USA and China.

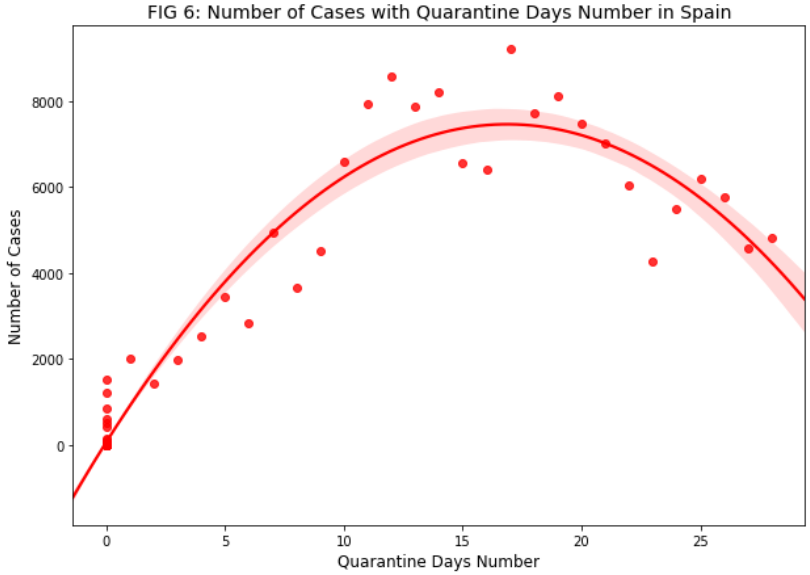
2.3.4. Quarantine Days

Correlation coefficient of quarantine days number is 0.87 (positive, very strong correlation).

Quarantine days number has positive correlation with a very strong correlation value which meets with our expectation. so according to this data, quarantine days number have a great effect on reducing the number of cases in Spain (convert the growth rate from exponential to linear).

To make sure that there is a linear relation, we need to plot the scatter plot of the data with regression model.

FIG 6: Number of Cases with Quarantine Days Number in Spain

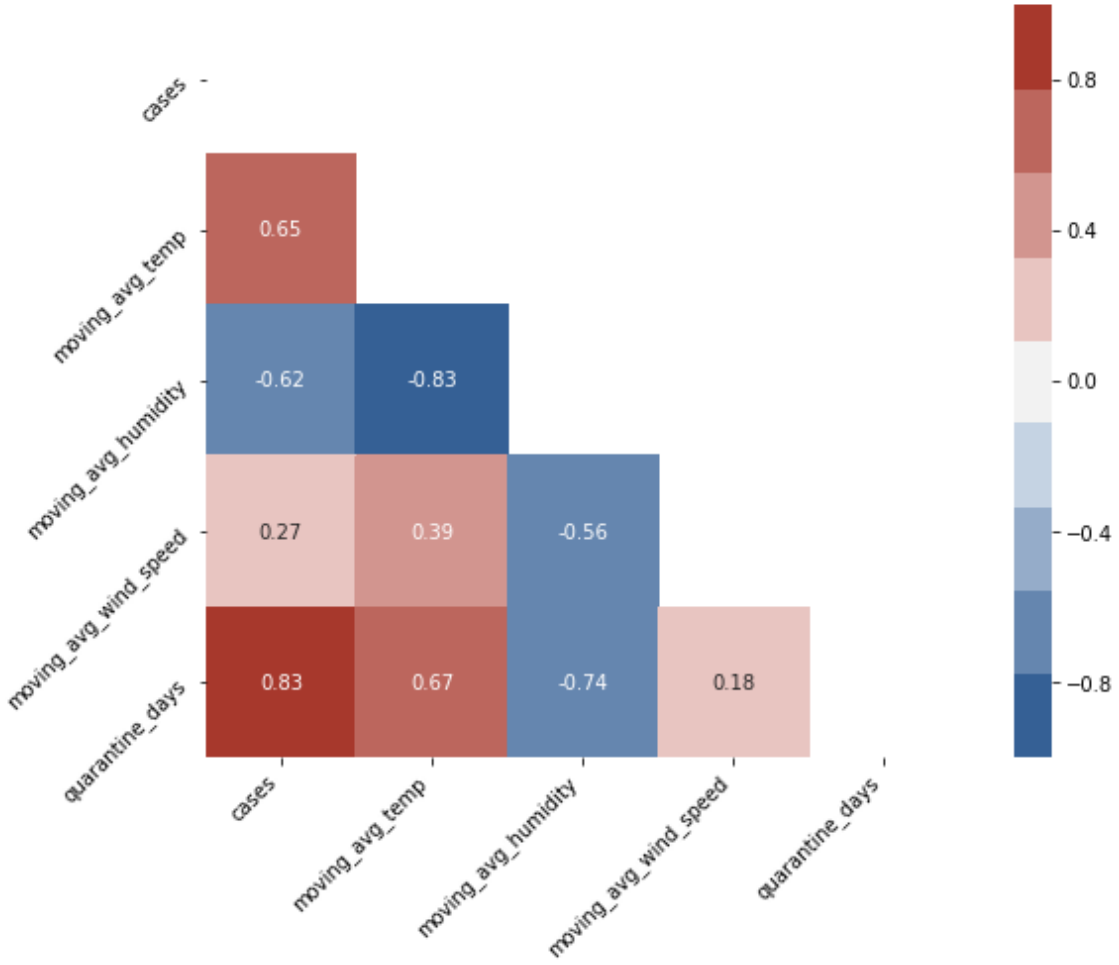


As we could see from the figure, the relation is close to be quadratic curve (non-linear but not exponential) between the quarantine days number and the number of cases which is needed to flatten the curve and we also could see that at first days of quarantine , the cases number increases very fast then decreases as the days of quarantine increases (after 16 days of quarantine).

(**Note:** In China, the curve was close to be quartic curve or bell curve because the virus spread had stopped in China but in Spain case the curve is close to be quadratic and the curve doesn't reach zero cases because the virus still spreads in Spain)

2.4. Italy

FIG 7: Correlation Coefficients Matrix of Italy Data



	cases
cases	1.000000
moving_avg_temp	0.649391
moving_avg_humidity	-0.624167
moving_avg_wind_speed	0.273921
quarantine_days	0.831733

Correlation Coefficients between the number of cases and some possible factors for Italy

2.4.1. Temperature

Correlation coefficient of temperature is 0.65 (positive, moderate correlation).

Temperature has positive moderate correlation which doesn't meet with our expectation so according to this data,

There are two possibilities:

1- Temperature doesn't have any effect on the number of cases in Italy.
2- Temperature has not increased yet in Italy to the level that kill the virus so at the moment the temperature doesn't have any effect on stopping the spread of the virus but in the next months could have a great effect on slowing the virus spread.

So, we need to study the same factor on other countries to be sure which possibility is right.

2.4.2. Humidity

Correlation coefficient of humidity is -0.65 (negative, moderate correlation). There is a moderate correlation between humidity and number of cases in Italy contrary to expected from USA, China and Spain data. (as humidity increases in Italy, the number of cases decreases but we don't find this relation in other countries so we couldn't say that there is a relation between the humidity and the cases number)

2.4.3. Wind Speed

Correlation coefficient of wind speed is 0.27 (positive, negligible correlation). There is no correlation between wind speed and number of cases in Italy like Spain, USA and China.

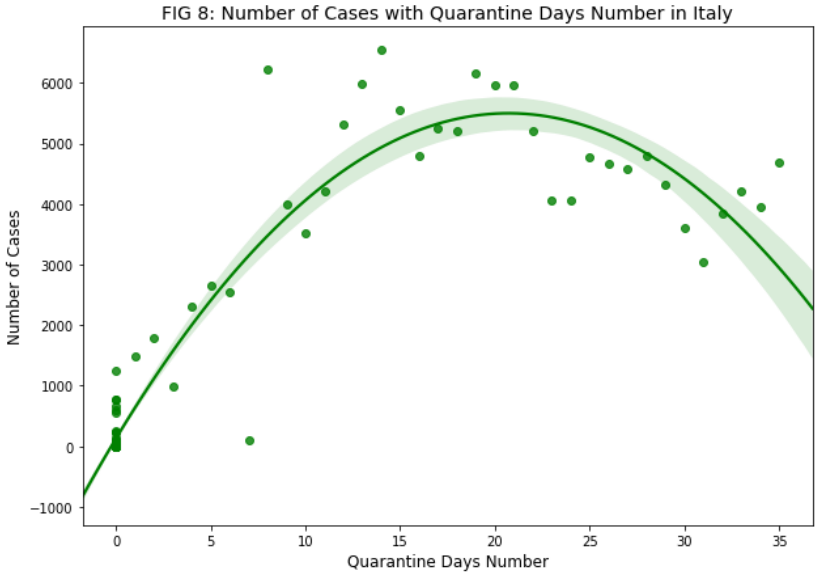
Now, we are completely sure there is **no relation between the wind speed and the number of cases.**

2.4.4. Quarantine Days

Correlation coefficient of quarantine days number is 0.83 (positive, very strong correlation).

Quarantine days number has positive correlation with a very strong correlation value which meets with our expectation. so according to this data, quarantine days number have a great effect on reducing the number of cases in Spain (convert the growth rate from exponential to linear).

To make sure that there is a linear relation, we need to plot the scatter plot of the data with regression model.



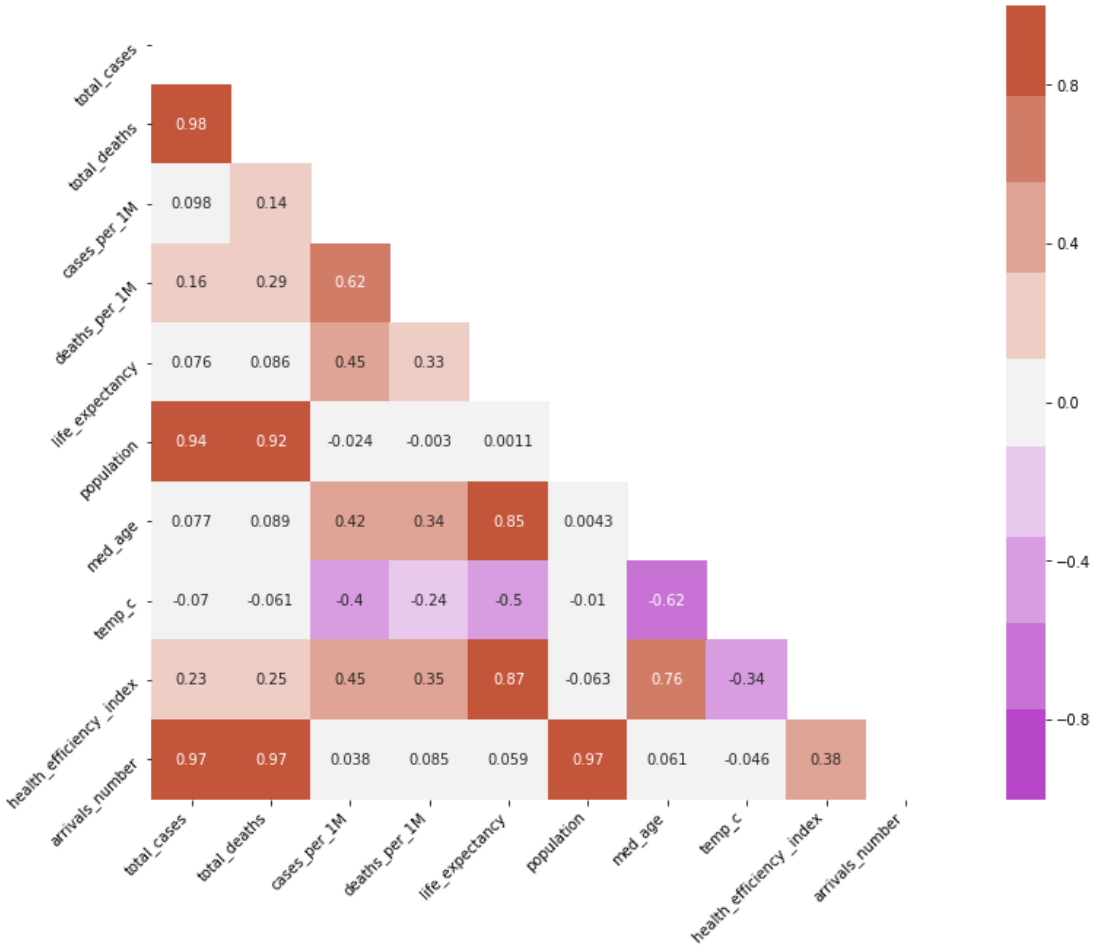
As we could see from the figure, the relation is close to be quadratic curve (non-linear but not exponential) between the quarantine days number and the number of cases which is needed to flatten the curve and we also could see that at first days of quarantine , the cases number increases very fast then decreases as the days of quarantine increases (after 20 days of quarantine).

(**Note:** The curve of Italy is the same curve we found in Spain data and china data which means that the quarantine days number has a great effect on the number of cases)

(**Note:** In China, the curve was close to be quartic curve or bell curve because the virus spread had stopped in China but in Italy case the curve is close to be quadratic and the curve doesn't reach zero cases because the virus still spreads in Italy)

2.5. The World (All Countries)

FIG 9: Correlation Coefficients Matrix of World Data



	total_cases	total_deaths	cases_per_1M	deaths_per_1M
total_cases	1.000000	0.978273	0.098226	0.162677
total_deaths	0.978273	1.000000	0.137816	0.292704
cases_per_1M	0.098226	0.137816	1.000000	0.621053
deaths_per_1M	0.162677	0.292704	0.621053	1.000000
total_tests	0.987175	0.961565	0.062391	0.089096
life_expectancy	0.075501	0.086308	0.451322	0.327950
population	0.939474	0.922750	-0.024058	-0.002980
med_age	0.077200	0.089417	0.417527	0.343832
temp_c	-0.069533	-0.061235	-0.404740	-0.242239
health_efficiency _index	0.226652	0.253657	0.451953	0.345150
arrivals_number	0.974395	0.968043	0.037665	0.084545

Correlation Coefficients between the number of cases and some possible factors for all countries

2.5.1. Temperature

Correlation coefficient of temperature and total cases is -0.06(negative, negligible correlation).

Correlation coefficient of temperature and total cases per 1 million is -0.4 (negative, moderate correlation). (**Note:** we use total cases per 1 million (total cases over population) because it represents the spread of the virus better than the population)

Temperature has negative moderate correlation which meets slightly with our expectation so according to this data,

There are two possibilities:

1- Temperature could have some effect on the number of cases in the world.
2- Temperature has not increased yet in any place in the world to the level that kill the virus so at the moment the temperature doesn't have any effect on stopping the spread of the virus.

2.5.2. Arrivals Number

Correlation coefficient of arrivals number and total cases is 0.97 (positive, very strong correlation).

As we expected that the arrivals number that entered the country have a great and powerful effect on the number of cases and the virus spread. As the more travelers enter the country, the more cases will appear and the virus will be spread rapidly.

2.5.3. Population

Correlation coefficient of population and total cases is 0.94 (positive, very strong correlation).

We could conclude that the virus spreads very rapidly between the people in one nation which makes social distancing very important to be applied in all countries.

2.5.4. Health Efficiency Index

Correlation coefficient of health efficiency index and total deaths per 1 million people is 0.35(positive, weak correlation). There is no powerful correlation between health efficiency index and number of deaths which means the virus spread is above all country's health abilities. So unfortunately, there is no ability to save a lot of lives in one time and all countries are going to have shortage in medical supplies.

3. Conclusion

As we found in the previous analysis that there are some factors could have a great effect on the virus spread and there are other factors that have no effect on the spread like wind speed and humidity. And there are some other factors we studied to just understand the nature of the virus spread and its effect like population and health efficiency index.

So, after this analysis the most important factors that could have effect on the virus spread are temperature, quarantine days number and arrivals number.

3.1. Temperature

From our previous analysis, we found that **temperature has negative correlation sometimes and positive value sometimes and the correlation value was between 0.25 to 0.4 in all data which meets very slightly with our expectations** so according to the data we have,

There are two possibilities:

1- Temperature could have some effect on the number of cases in any country in the world.
2- Temperature has not increased yet in the entire world or in any country to the level that kill the virus so at the moment the temperature doesn't have any effect on stopping the spread of the virus but in the next months could have a great effect on slowing the virus spread.

According to science and doctors, the second possibility is the closest one to logic so with the data we have now there no big effect from the temperature on the virus spread but with analyzing the data that will be collected on the next weeks that will have high temperatures, we might find some effect on the virus spread. So, **at the moment temperature doesn't have any effect on the virus spread.**

3.2. Quarantine Days Number

From our previous analysis, we found that quarantine days number has positive correlation with a very strong correlation value in all countries data we studied (correlation value between 0.83 to 0.99) which meets with our expectations. so according to this data, quarantine days number have a great effect on reducing the number of cases in any country.

But as we found that in USA that the relation between the quarantine days number and the number of cases is linear (which is a good curve instead of exponential curve) but in China, Spain and Italy the relation was quadratic curve (which is the best curve we want).

FIG 3: Number of Cases with Quarantine Days Number in USA

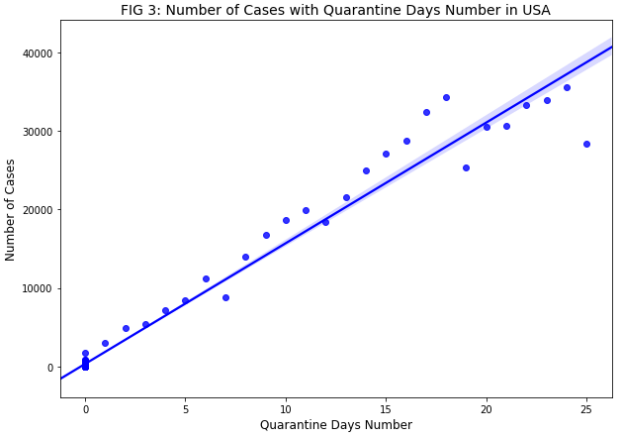
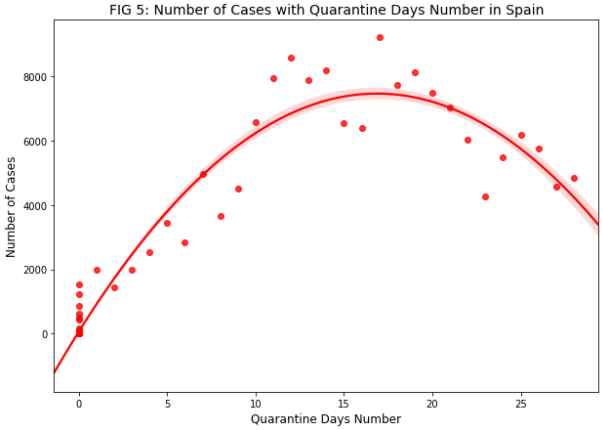


FIG 5: Number of Cases with Quarantine Days Number in Spain



We could explain these relations by the first date of fast spread, the virus had spread very rapidly in China since 21 January 2020 and in Spain and Italy since 28 February 2020 before USA virus fast spread that had begun since 7 march 2020 (almost weak after Spain and Italy and after two months from china). So, we could predict that the USA virus spread might have the same quadratic curve like China, Italy and Spain in the next few weeks.

3.3. Arrivals Number

From our previous analysis, we found that arrivals number has positive correlation with a very strong correlation value in almost all countries (value of 0.97) which meets with our expectations. so according to this data, arrivals number have a great effect on spread the virus all over the world so taking the decision of shutdown airports was good one to reduce the number of cases.

In the end, we could conclude that the quarantine procedures are the best way to stop the virus spread at the present. And we could see the success of Germany, South Korea, japan and some other countries to control the spread of the virus by taking quarantine procedures earlier than most countries that now have a great crisis in dealing with the virus spread like USA, Italy, Spain, UK and many other countries.