Data Visualization

for

COVID19

#### By Kiran Kumar

#### Date: 16-Apr-2020

Table of Contents

[OBJECTIVE 3](#_Toc38035849)

[DATA 3](#_Toc38035850)

[NEW TERMINOLOGY 3](#_Toc38035851)

[NUMBER OF TOTAL CASES 3](#_Toc38035852)

[NUMBER OF DAILY NEW CASES 4](#_Toc38035853)

[DAILY GROWTH FACTOR 4](#_Toc38035854)

[CASE FATALITY RATE (CFR) 4](#_Toc38035855)

[OTHER 4](#_Toc38035856)

[THINGS I WANT TO CHECK 5](#_Toc38035857)

[**HYPOTHESIS 1**: IF PEOPLE LIVE VERY CLOSE TO EACH OTHER (POPULATION DENSITY), IT CAUSES HIGHER SPREAD (TOTAL CASE/1MPPL) 5](#_Toc38035858)

[CHART: TOTAL CASES/1MPPL VS POPULATION/SQKM OF LAND AREA 5](#_Toc38035859)

[CHART OBSERVATION: 5](#_Toc38035860)

[**HYPOTHESIS 2**: SPREAD RATE IS SIMILAR IN ALL THE COUNTRIES 6](#_Toc38035861)

[CHART: TOTAL CASES BY REPORT DATE 6](#_Toc38035862)

[CHART OBSERVATION 6](#_Toc38035863)

[CHART: TOTAL CASES BY #DAYS FROM 100 CASES 7](#_Toc38035864)

[CHART OBSERVATION 7](#_Toc38035865)

[**HYPOTHESIS 3**: IS IT A DISEASE OF RICH PEOPLE? 8](#_Toc38035866)

[Chart: Total Cases by World bank Income Group 8](#_Toc38035867)

[CHART OBSERVATION 8](#_Toc38035868)

[**HYPOTHESIS 4**: IS COUNTRY-WISE LOCKDOWN IS AN EFFECTIVE STRATEGY? 9](#_Toc38035869)

[Chart: Side by side comparison of Daily growth factor for top 8 counties by total cases 9](#_Toc38035870)

[CHART: SIDE BY SIDE COMPARISON OF TOTAL CASE PER 1M PEOPLE FOR TOP 8 COUNTIES BY TOTAL CASES 9](#_Toc38035871)

[CHART OBSERVATION: 10](#_Toc38035872)

[DISCLAIMER 11](#_Toc38035873)

[RESOURCES 11](#_Toc38035874)

[DATA SOURCES LINKS 11](#_Toc38035875)

[POWERBI REPORT 11](#_Toc38035876)

[SOURCE CODE FOR POWERBI REPORT 11](#_Toc38035877)

# OBJECTIVE

There is need not to say how Corona is infecting the whole world. Scientists are racing to understand the coronavirus disease, Covid-19, which is now a fast-growing global pandemic.

For the last couple of weeks, I have been collecting data about coronavirus spread across the world to make some sense of its complexity**.**

My objective was not to create a real-time tracker. I guess Google and Wikipedia authors have done really great work here.

* I wanted to understand how those curves are forming up.
* I want to compare two or more countries side by side.
* Are there more patterns to look for?

It was a challenge on multiple fronts.

# DATA

Within the initial hours of searching, I realized that only reliable data source for multi-country data is situation reports being published by WHO daily.

However, these reports are in PDF format, and their structure varies within a couple of days.

I used PowerBI to parse the tabular data inside these PDFs. There were some obvious challenges like page breaks, line breaks, white spacing issues. It was also difficult to map these names to ISO country codes. I wanted to combine this data with data country-level data and wanted to plot these on a map. It was a tedious job to parse a new PDF every day and combine with previously collected data.

I was not alone, many people were doing the same thing and struggling with similar challenges. Luckily someone in ECDC (European Centre for Disease Prevention and Control) has solved this problem for all. They are uploading the cleaned-up version of daily time-series data, along with ISO country codes.

I was able to easily combine this data, other country-level stats from Wikipedia and WHO.

# NEW TERMINOLOGY

Once I was over with parsing and engineering challenges, I was able to see the data on maps and started seeing the curves. But these were trying to take off at their own wish, these were not taking off in a coordinated manner like the ones I have seen in "flatten the curve" videos. Here, I realized that I need to bake more features into the data.

We should not compare two countries data on the report date basis. So, I created two more offset numbers specific to each country - #Days from 100th Case, # Days from 10th Death. Now when I started comparing the curves using these offsets, it was giving me similar results like I was seeing the news.

As I start looking more into the data, I realize that I need more data features.

## NUMBER OF TOTAL CASES

This number for a country can be used to understand the pressure on the health infrastructure.

## NUMBER OF DAILY NEW CASES

It can help understand the speed of spread and effectiveness of the preventive strategies deployed by an admiration. It also helps calculate the daily growth factor.

## DAILY GROWTH FACTOR

It is the ratio of new cases reported today by new cases reported yesterday. A value of DGF > 1 means signals exponential growth, whereas DGF < 1 indicates whatever strategy the country is using is working.

## CASE FATALITY RATE (CFR)

Again, it was not a high number of positive cases, which creates panic. It is the number of hospitalizations required to reduce the permanent impact on health and avoid death. I could not get the concrete data on the case to hospitalization ratio, however, deaths to cases can be easily calculated. This number is called the **Case Fatality Rate (CFR).** It is the proportion of deaths from a compared to the total number of confirmed cases for a certain period. A CFR is conventionally expressed as a percentage and represents a measure of disease severity.

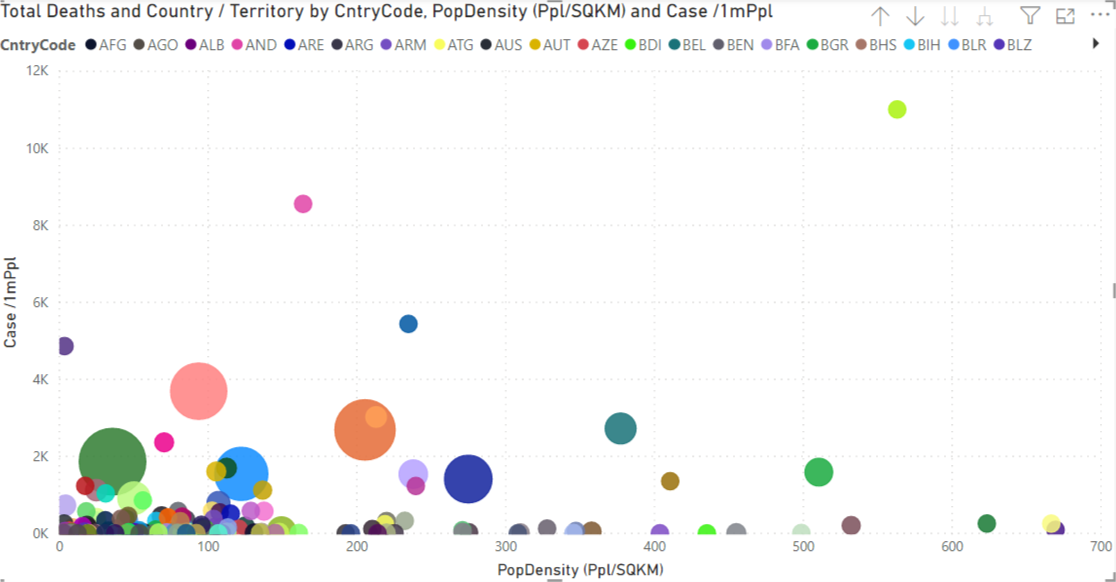
## OTHER

If you look further, you will say that a country with a higher population may have a higher number of total cases. Hence it is important to look at the cases per 1 million people. I could not find a proper name for it, so I called it **Total** **Spread** andrepresented it as **Cases/1mPpl**. Also created **Deaths/1mPpl**.

# THINGS I WANT TO CHECK

## **HYPOTHESIS 1**: IF PEOPLE LIVE VERY CLOSE TO EACH OTHER (POPULATION DENSITY), IT CAUSES HIGHER SPREAD (TOTAL CASE/1MPPL)

### CHART: TOTAL CASES/1MPPL VS POPULATION/SQKM OF LAND AREA



*Chart Note: Data as of 15th April 2020. Excluded countries with Population density more than 1200 people per SQKM.*

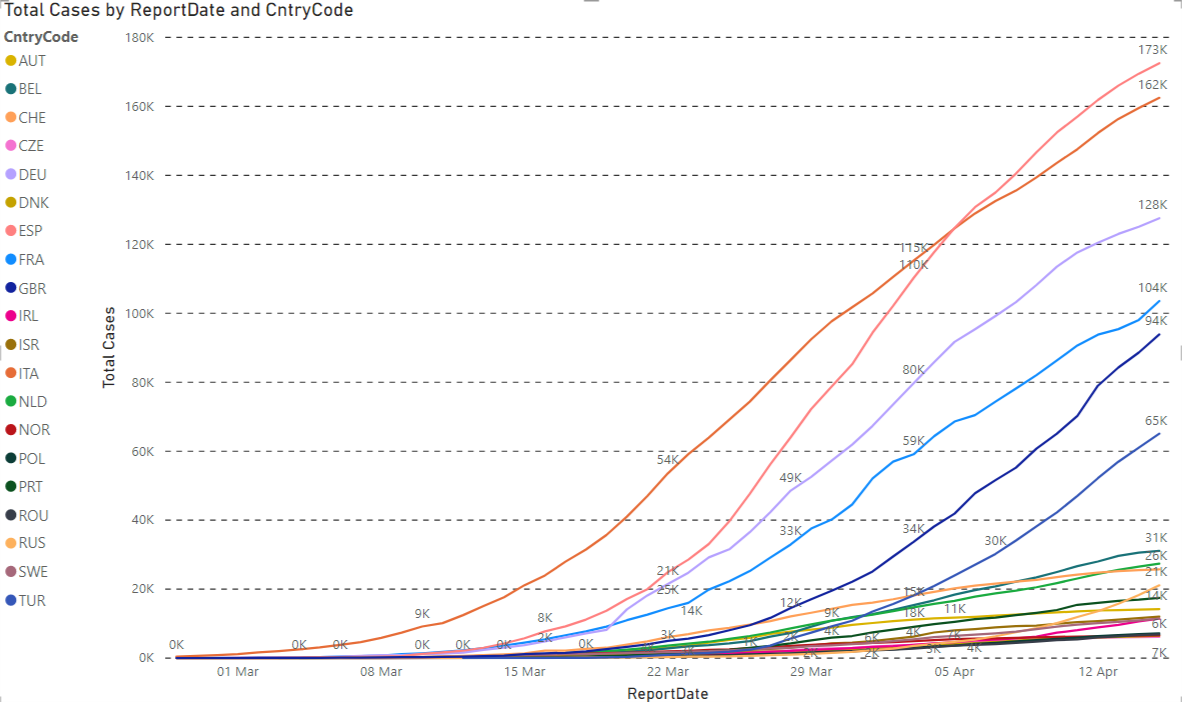
### CHART OBSERVATION:

After excluding some of the outliers (population density > 1200 ppl/SQKM) like Singapore, Bahrain, Malta, etc. what I see does not confirm the hypothesis. Rather see the reverse of the hypothesis.

## **HYPOTHESIS 2**: SPREAD RATE IS SIMILAR IN ALL THE COUNTRIES

Observation: When plotting for all the world, it is getting difficult to look for any pattern. So, I switched to the region view. In my example it is Europe. And I selected the top 20 countries in terms of the total cases.

### CHART: TOTAL CASES BY REPORT DATE

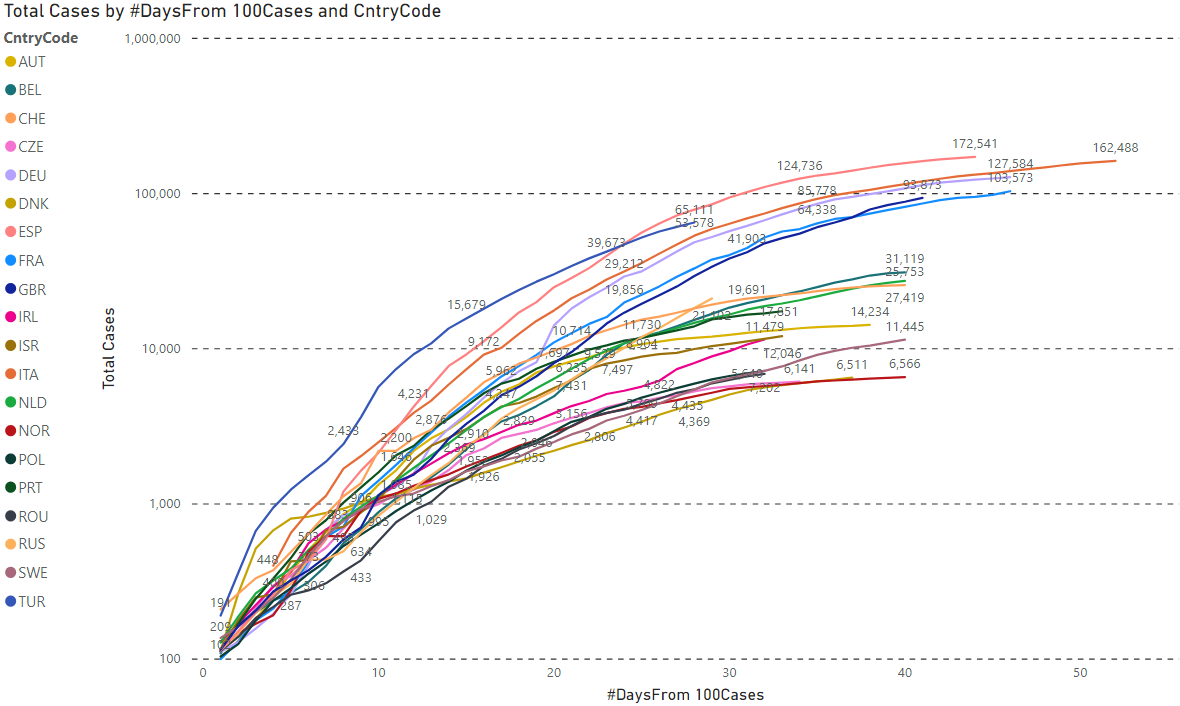


*Chart Note: Data as of 15th April 2020. Showing the top 20 countries with total cases countries in the WHO region Europe.*

### CHART OBSERVATION

When analyzing the total case across the report date on a linear scale, it shows rapidly inclining lines for each country.

### CHART: TOTAL CASES BY #DAYS FROM 100 CASES



*Chart Note: Data as of 15th April 2020. Showing the top 20 countries with total cases countries in the WHO region Europe.*

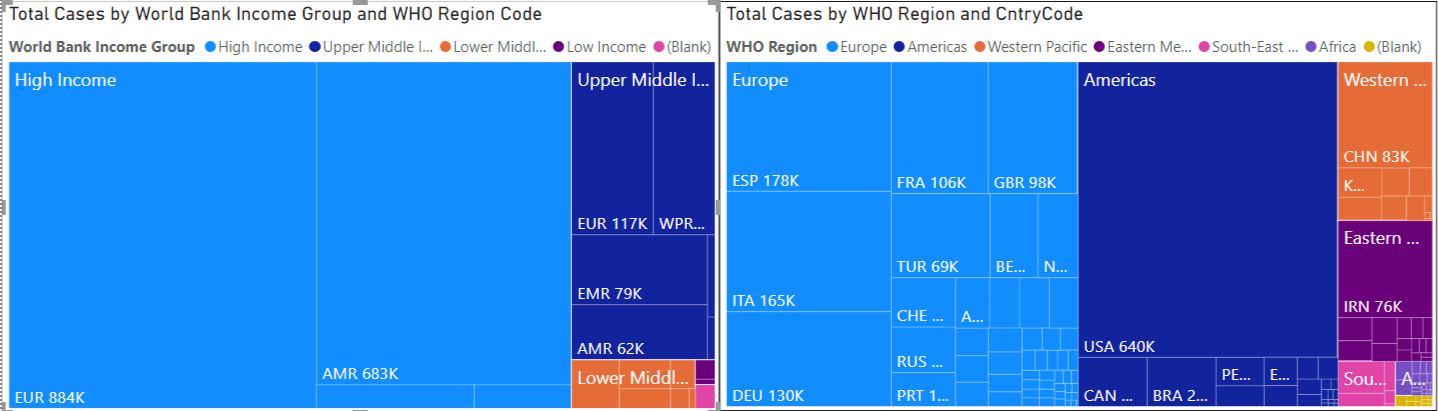
### CHART OBSERVATION

When analyzing the total case across 100 Case offset on a log scale, a mathematical symmetry of the exponential curve.

## **HYPOTHESIS 3**: IS IT A DISEASE OF RICH PEOPLE?

It was a difficult statement to test. There was no data available about the income group of the infected people. Hence, I used the proxy variable of Income group ranking by the world bank of the country. This is still very coarse grain, as within a country the spread is very different from province to province.

### Chart: Total Cases by World bank Income Group



*Chart Note: Data as of 16th April 2020.*

### CHART OBSERVATION

It does seem like high-income group countries are most impacted. Although it seems less to do with their immunity and more to do with their exposure and access to test facilities.

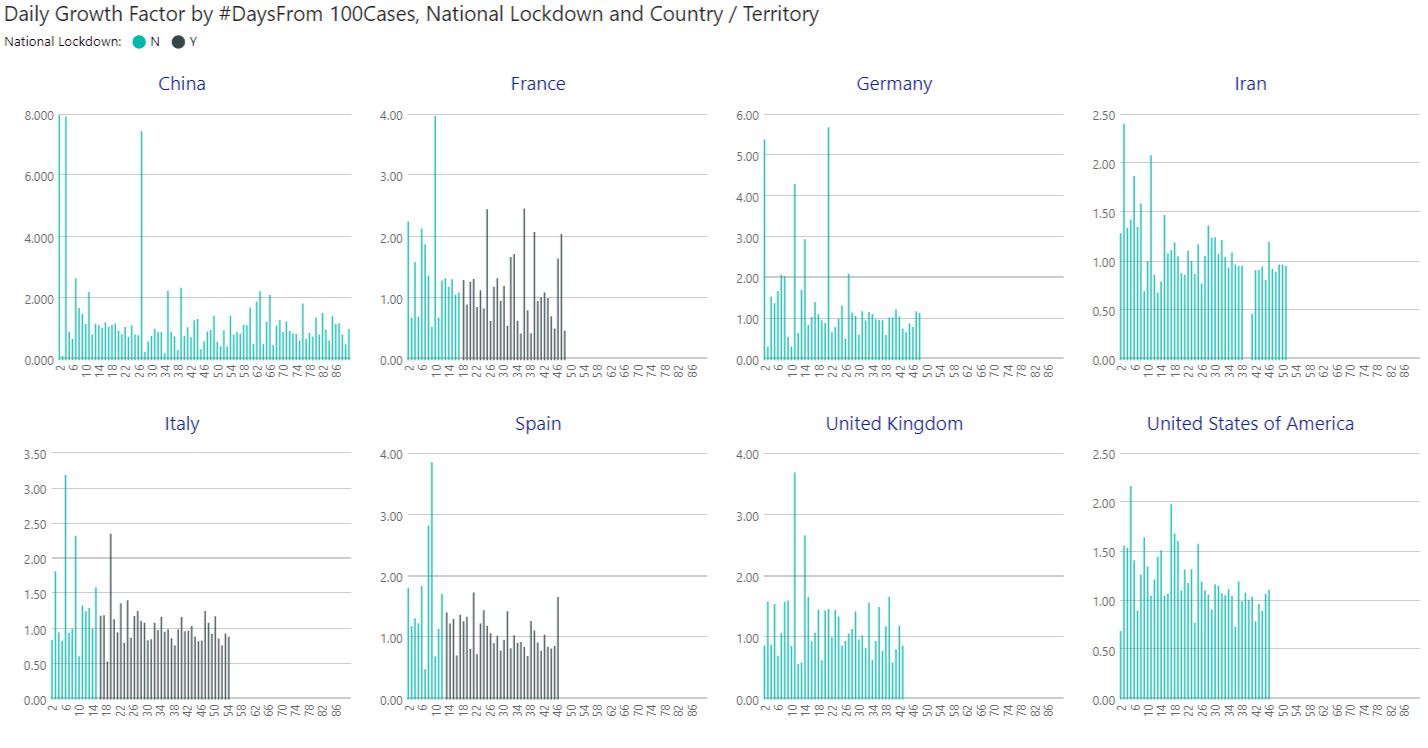
Other possible factors could be

1. Access to international commute - Regions within Europe and the USA are well connected by flights, and this may be cause for widespread.
2. Total cases within rich countries might be quite close to the actual number of people infected. Whereas in low-income group countries people may not be aware of the situation's gravity or do not have access to be tested. Hence, a number coming from those countries might represent only a small fraction of actually infected people.

## **HYPOTHESIS 4**: IS COUNTRY-WISE LOCKDOWN IS AN EFFECTIVE STRATEGY?

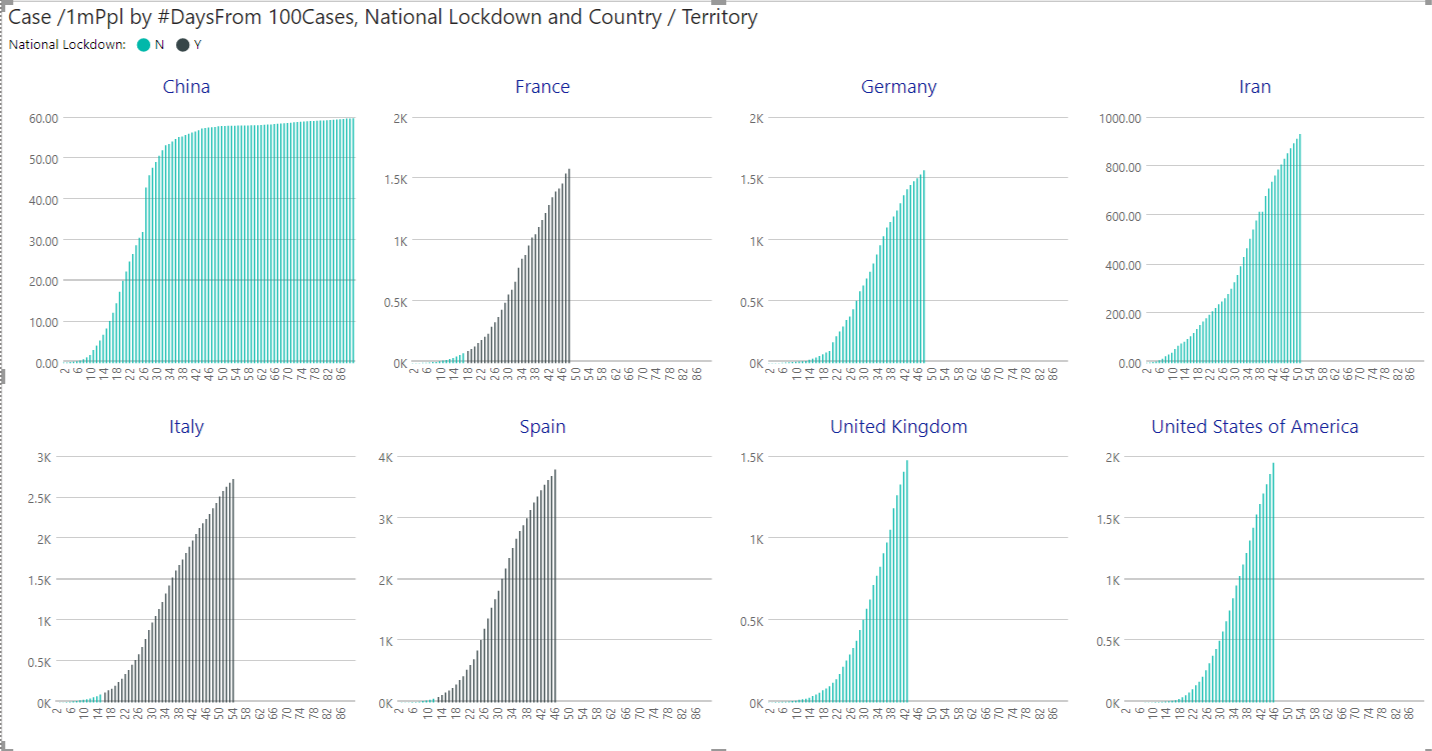
It is really difficult to verify. All administrations are trying their best to control the situation and deploying the strategy based on the extent of the spread and resources available to them.

### Chart: Side by side comparison of Daily growth factor for top 8 counties by total cases



*Chart Note: Data as of 16th April 2020.*

### CHART: SIDE BY SIDE COMPARISON OF TOTAL CASE PER 1M PEOPLE FOR TOP 8 COUNTIES BY TOTAL CASES



*Chart Note: Data as of 16th April 2020.*

### CHART OBSERVATION:

It is difficult to conclude based on the data we are seeing here. It feels like lockdown is more a reaction rather than a deciding factor.

# DISCLAIMER

I am not an expert on coronavirus or pandemic. What I am explaining here is what I am seeing in data. It is possible that I am missing some critical factors while formulating my observations.

# RESOURCES

## DATA SOURCES LINKS

|  |  |
| --- | --- |
| ECDC dataset | <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide> |
| WHO Situation Reports | <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> |
| Coronavirus pandemic by country and territory | <https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic_by_country_and_territory#covid19-container> |
| WHO API for country stats | [https://apps.who.int/gho/athena/data/GHO/WHS9\_86,WHS9\_88,WHS9\_89,WHS9\_92,WHS9\_96,WHS9\_97,WHS9\_90?filter=COUNTRY:\*;REGION:\*&amp;format=xml&amp;profile=excel](https://apps.who.int/gho/athena/data/GHO/WHS9_86,WHS9_88,WHS9_89,WHS9_92,WHS9_96,WHS9_97,WHS9_90?filter=COUNTRY:*;REGION:*&amp;format=xml&amp;profile=excel) |
| Wikipedia flatten the curve | <https://en.wikipedia.org/wiki/Flatten_the_curve> |

## POWERBI REPORT

<https://app.powerbi.com/groups/43e9d8e4-e1f8-409d-96b8-fa864bd08154/reports/57741b26-6749-4238-acb3-09793cba9573?ctid=6b94db52-3791-432c-b97e-871411cd202e>

## SOURCE CODE FOR POWERBI REPORT

<https://github.com/goelkiran/covid19>