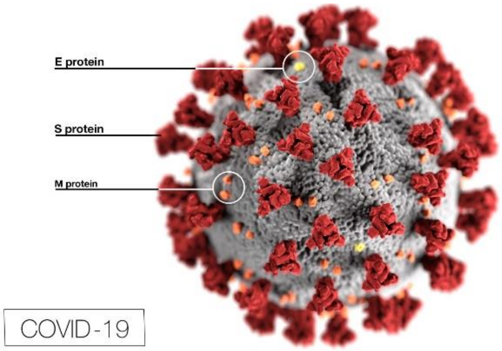
# 1 Cover Page

# COVID-19 Spread



[1]

## Global statistics of infection and mortality over the last

Student:

Rohit Bhingare

Teacher:

Alex Aklson

Course:

IBM Data Science Professional Certificate - Data Science Capstone

[1] Source – CDC https://www.cdc.gov/coronavirus/2019-ncov/index.html

# 2 Table of Contents

# 3 Executive Summary

## Background

Coronavirus disease (COVID-19) is an infectious disease caused by a new virus. The disease causes respiratory illness (like the flu) with symptoms such as a cough, fever, and in more severe cases, difficulty breathing. The new variant 2019-nCoV was first identified in Wuhan, the capital of China's Hubei province. People developed pneumonia without a clear cause and for which existing vaccines or treatments were not effective. The virus has shown evidence of human-to-human transmission

## Description of the problem

A new coronavirus designated 2019-nCoV was first identified in Wuhan, the capital of China's Hubei province. People developed pneumonia without a clear cause and for which existing vaccines or treatments were not effective. The virus has shown evidence of human-to-human transmission. Transmission rate (rate of infection) appeared to escalate in mid-January 2020. As of 30 January 2020, approximately 8,243 cases have been confirmed.

Transmission rate (rate of infection) appeared to escalate in January 2020 and as of 22nd of March, cases of new coronavirus infections have been confirmed in more than 160 countries or regions. Italy has over 50’000 individuals infected, while globally we have over 300’000 individuals infected. There is now a risk of becoming infected with new coronavirus in almost all parts of the world.

# 4 Introduction of Data

## List of datasets used

1. We use the **COVID-19 Complete Dataset** from Kaggle[2]. This file contains the cumulative count of confirmed, death and recovered cases of COVID-19 from different countries from 22nd January 2020.
2. To get the latest COVID-19 statistics, we use the worldometers [3]index data that is refreshed in live.
3. We will also use the world's good country index ranking[4]
4. We will use the world population counts published by United Nations.[5]
5. Further, we will use the world nominal GDP index[6] data.

## Scope of analysis

- Since the Kaggle data is repeated as a time series for each country for the last 2 months, we will use specific data points (e.g. start of recording and last recorded date) to visualise the rate of change in global outbreak at specific times.
- We will look into the outbreak on specific regions and try to explain whether there are specific areas of world where the spread is increasing faster/slower.
- We will attempt to combine the COVID-19 spread with world's good country index statistics, population as well as nominal GDP index to see whether we can explain some of the outbreaks against population/GDP of countries. For this part, we will attempt to use a machine learning model to see how these factors affect the outbreak spread rate and whether we can predict the rate of increase/decrease in the near future.

[2] https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov.html downloaded on 23.March.2020 at 16:31.

[3] https://www.worldometers.info/coronavirus/

[4] https://en.wikipedia.org/wiki/Good_Country_Index

[5] https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)

[6] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)

# 5 Methodology

## Data collection

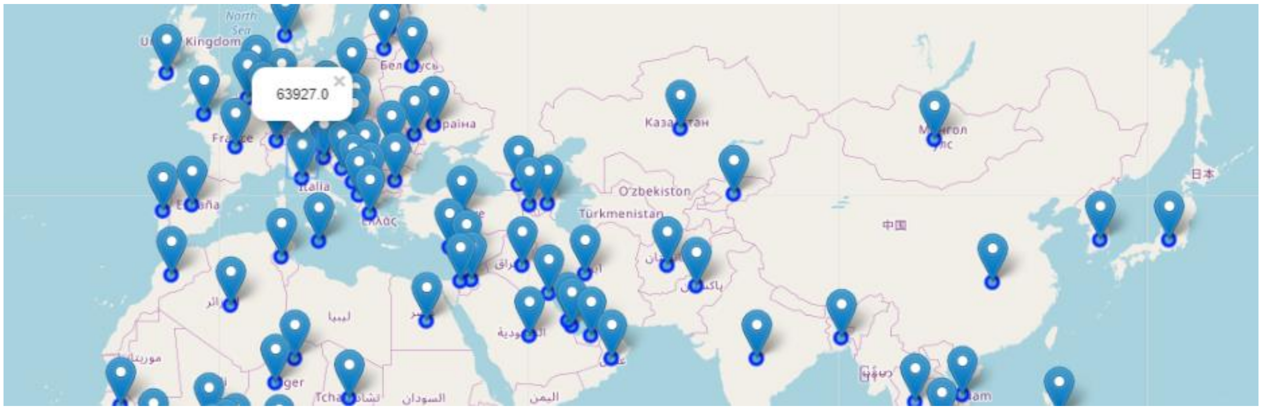
In the first step, we gather the data sources.  The KAGGLE data is directly available for download. For the latest statistics, we use the Worldometers data mentioned previously. This data gives one row per country and is much easier to process for our exploratory analysis. Additionally, we use the good country index, population of countries as well as GDP information from Wikipedia. Note that when more than one sources were available on Wikipedia, I have prioritized the World Bank Index data over others.

## Data preparation

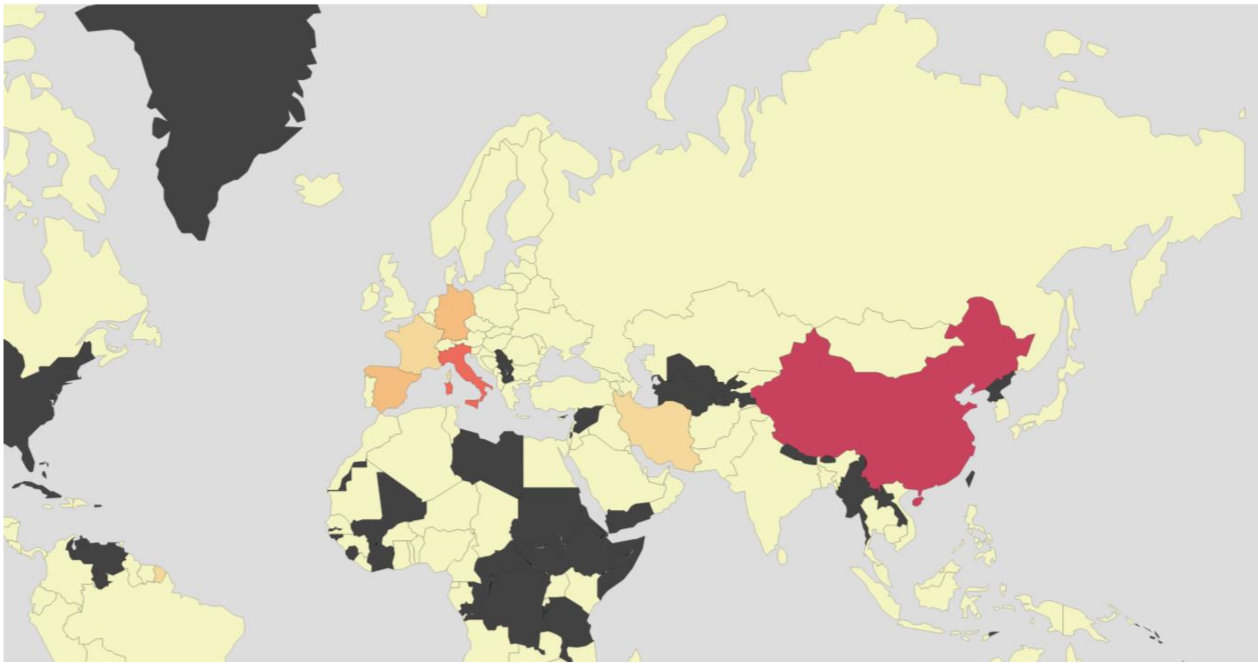
Since the data was coming from varied sources, the index (country name) was changing for each source. The cleanup in Python was somewhat cumbersome. Hence, I used Excel to perform some of the country name cleanup. e.g. depending upon the source, some sources state “United States”, while others state “United States of America”. This has also caused some conflict in the GeoJSON file being able to correctly map Choropleth maps in some cases.

## Data visualization

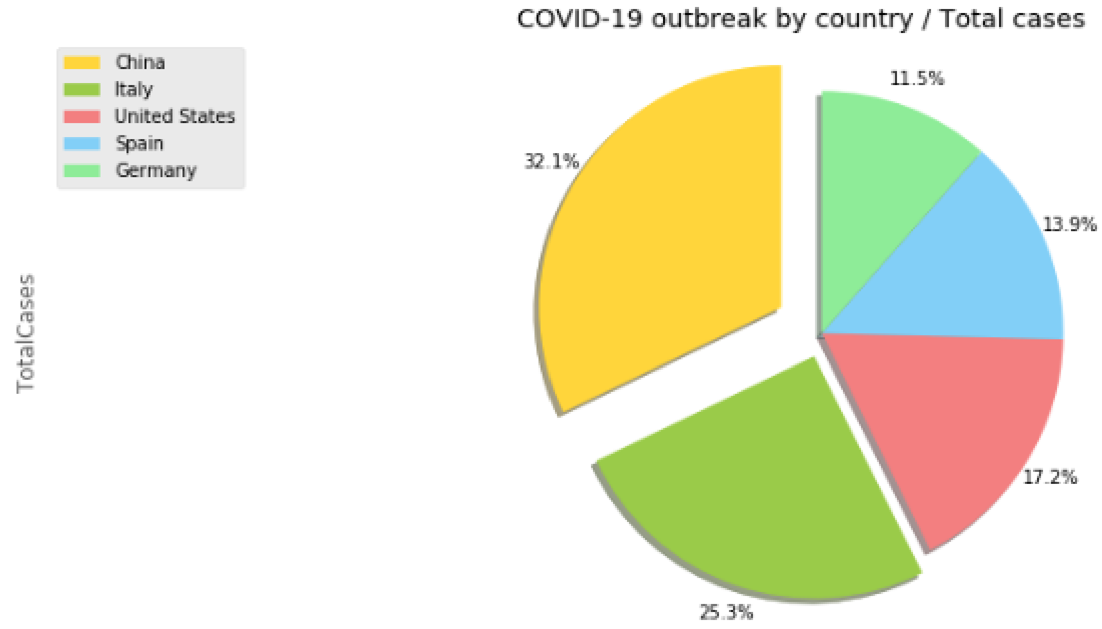
- We performed a top-down statistics to see how the COVID-19 outbreak looks using a world map with circle markers and labels per country.



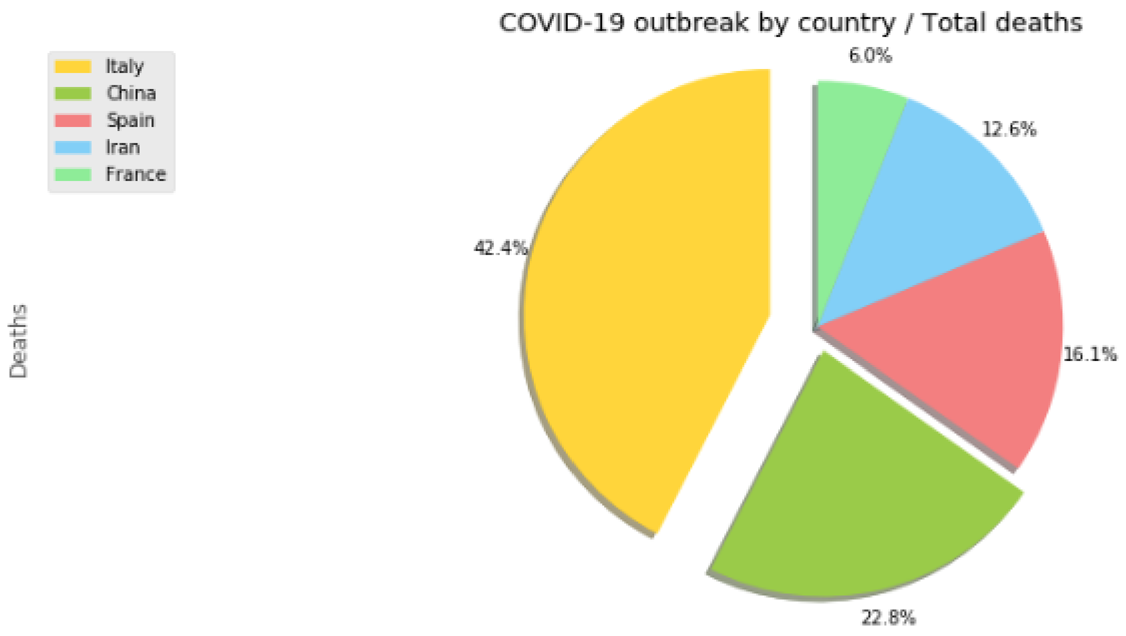
- We used Choropleth world map to see the intensity of outbreak on the most affected countries.



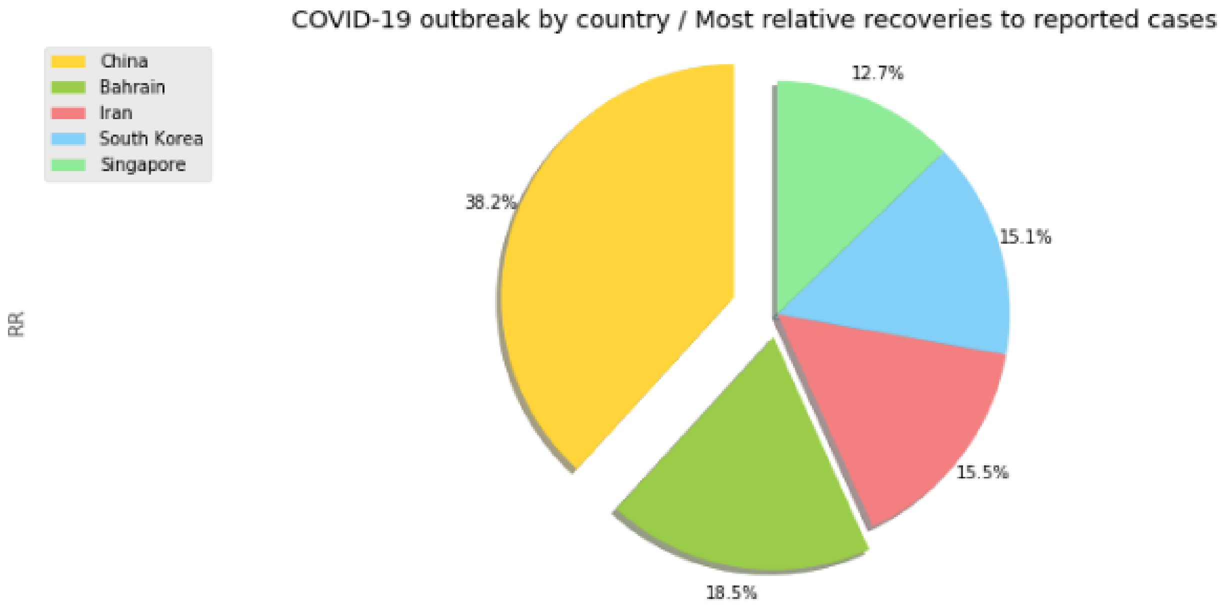
- We used pie-charts to assess the nature of outbreak using various factors.


COVID-19 outbreak by country / Total cases

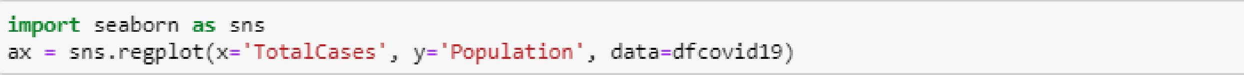
We see that China and Italy have the highest spread, followed by USA, Spain and Germany



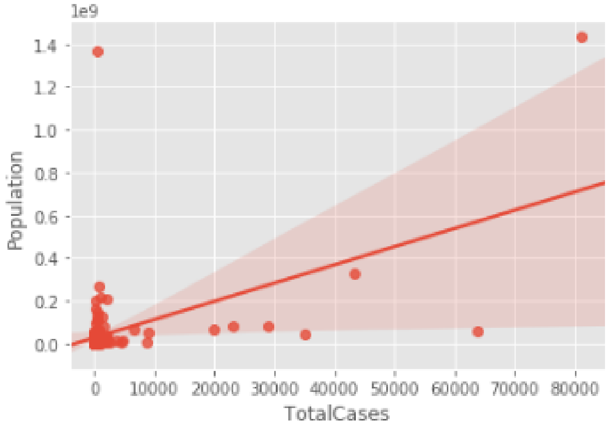
We see that Italy has the highest reported deaths so far, followed by China, Spain, Iran and France



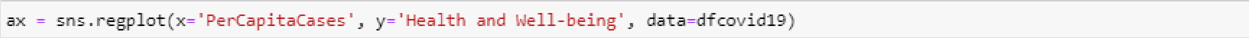
- We used seaborn regression plots to use some of the socioeconomic factors and see whether they directly or indirectly affect the outbreak.



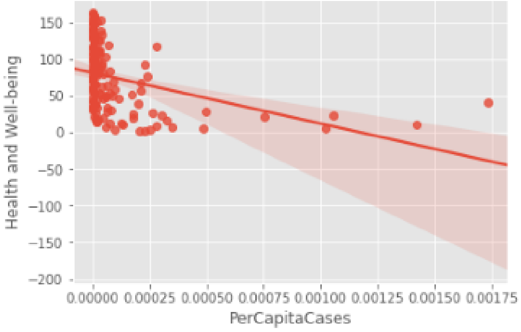
```
import seaborn as sns
ax = sns.regplot(x='TotalCases', y='Population', data=dfcovid19)
```



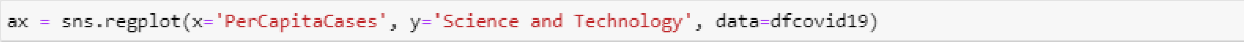
Observation 1 - Fairly obvious, but we see a somewhat linear relationship between population and total cases reported.



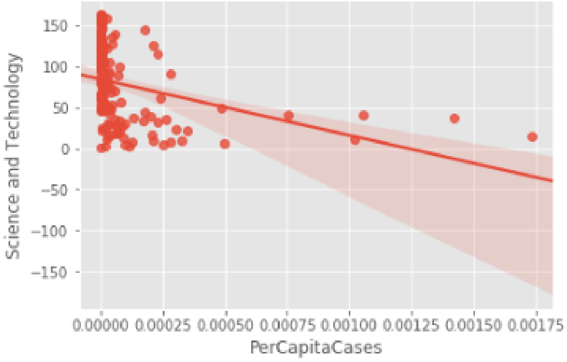
```
ax = sns.regplot(x='PerCapitaCases', y='Health and Well-being', data=dfcovid19)
```



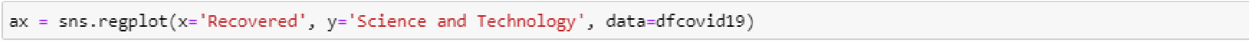
Observation 2 - Again expected, but we see a inverse linear relationship between economic health and well-being index and per capita cases reported.



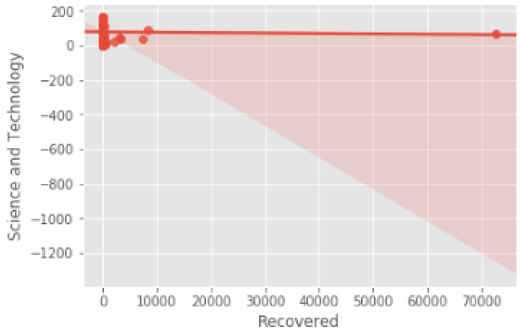
```
ax = sns.regplot(x='PerCapitaCases', y='Science and Technology', data=dfcovid19)
```

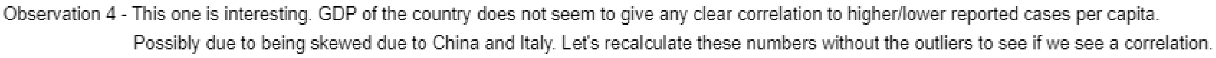


Observation 3 - Similar to 2, we see a inverse linear relationship between Science and Technology index and per capita cases reported.



```
ax = sns.regplot(x='Recovered', y='Science and Technology', data=dfcovid19)
```

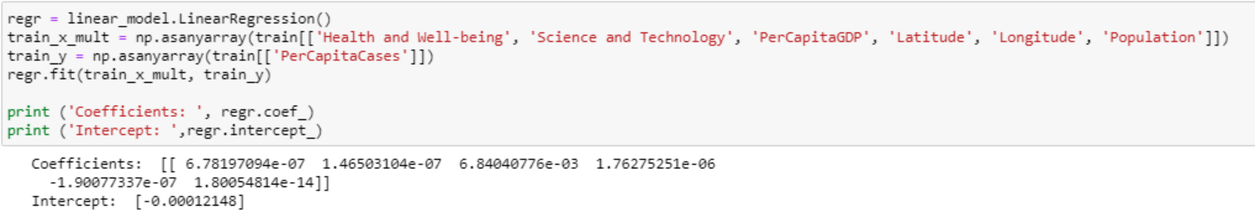




Observation 4 - This one is interesting. GDP of the country does not seem to give any clear correlation to higher/lower reported cases per capita. Possibly due to being skewed due to China and Italy. Let's recalculate these numbers without the outliers to see if we see a correlation.

## Machine learning

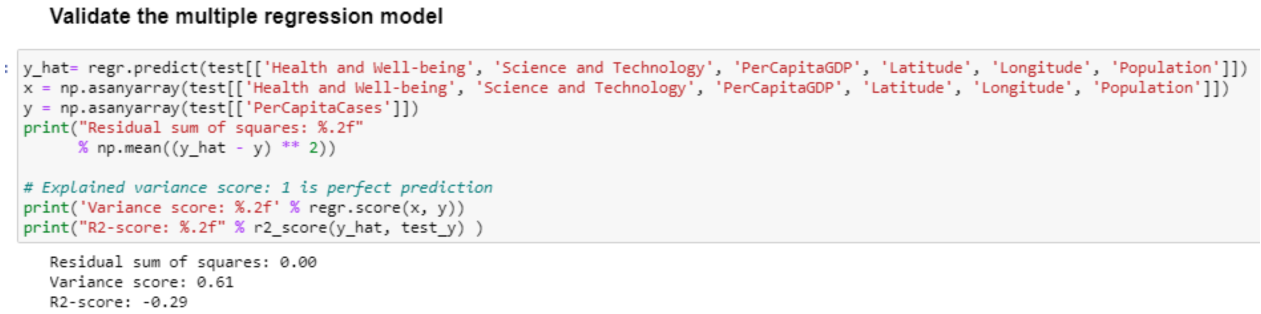
- We attempted to use machine learning to see whether we can explain the outbreak using some of the socioeconomic parameters.



```
regr = linear_model.LinearRegression()
train_x_mult = np.asanyarray(train[['Health and Well-being', 'Science and Technology', 'PerCapitaGDP', 'Latitude', 'Longitude', 'Population']])
train_y = np.asanyarray(train[['PerCapitaCases']])
regr.fit(train_x_mult, train_y)

print ('Coefficients: ', regr.coef_)
print ('Intercept: ',regr.intercept_)
```

```
Coefficients:  [[ 6.78197094e-07  1.46503104e-07  6.84040776e-03  1.76275251e-06
  -1.90077337e-07  1.80054814e-14]]
Intercept:  [-0.00012148]
```



### Validate the multiple regression model

```
y_hat= regr.predict(test[['Health and Well-being', 'Science and Technology', 'PerCapitaGDP', 'Latitude', 'Longitude', 'Population']])
x = np.asanyarray(test[['Health and Well-being', 'Science and Technology', 'PerCapitaGDP', 'Latitude', 'Longitude', 'Population']])
y = np.asanyarray(test[['PerCapitaCases']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat, test_y) )
```

```
Residual sum of squares: 0.00
Variance score: 0.61
R2-score: -0.29
```

# 6 Results/Discussion

1. We observe that though China has the highest outbreak, it also has the highest recoveries per outbreak compared to the other countries.
2. We observe that Italy has the highest deaths (total and per capita) compared to the rest of the countries.
3. We observe From the Choropleth World Map that the major outbreak of COVID-19 has occurred in China and Europe only.
4. We saw that there are many countries where there is no recovery. Based on the statistics, it appears that majority of these countries are in Africa.
5. There is a clear correlation between population of country to the total cases that are reported, both statistically and visually (from the data as well as seaborn graphs).
6. We looked into providing a correlation of corona virus outbreak from various economic factors for a country. We were able to see a clear correlation for some factors.
7. There appears to be an inverse correlation between the good countries index against the outbreak (though not absolutely clear), e.g. We saw that Science and Technology as well as Health and Well-being indices showed a correlation visually, though it is somewhat complex to fit when a machine learning model was created.

# 7 Further extension scope

1. Analyse the time series version of COVID-19 data from KAGGLE. This data provides a daily statistic for last 2 months. It will be interesting to further explore the theme of "rate of recovery" across the months for the outspread per country against the good country KPIs.
2. Analyse the COVID-19 outbreak from a cause-effect perspective, i.e. what effect it has on other socioeconomic factors such as financial indices, situation in hospitals as well as job market. This would be an interesting theme to explore deeper.
3. Analyse the mortality/illness rates in countries in the previous years during the same/similar times to see how much effect COVID-19 outbreak has on them.
4. Drill-down further on the COVID-19 outbreak individuals per country and address whether the individuals had previous history of illnesses.