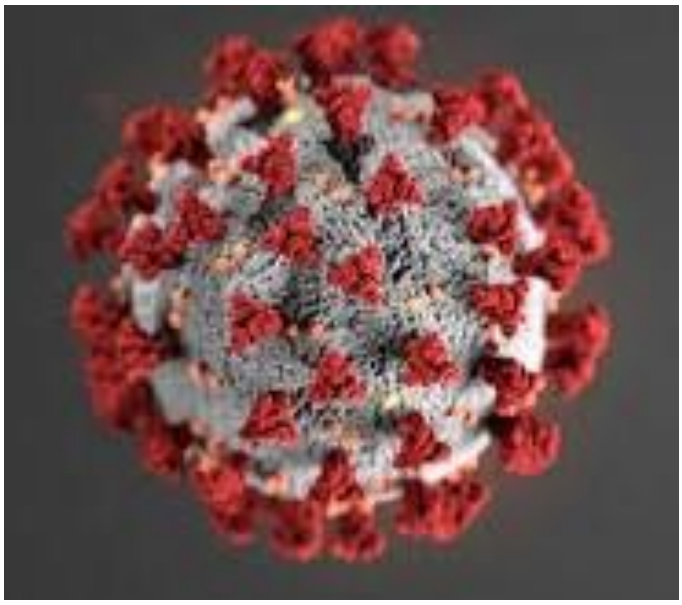
Coursera Capstone

IBM Applied Data Science Capstone

Understanding the impact of Pandemic & Recovery Status

Coronavirus Disease 2019 (COVID-19)



By:

Purnachand Kollapudi

May 2020

Contents:

Introduction:

The Chinese country office of the World Health Organization (WHO) on 31.12.2019 confirmed cases of unknown cause pneumonia found in Wuhan City, Hubei Province of China. The Chinese authorities have described a new form of coronavirus which was detected by laboratory experiments on 07.01.2020. This is a new strain that had not been previously found in humans until the epidemic in Wuhan, China was identified. Currently officially known as Coronavirus Disease 2019 (COVID-19), this "novel" coronavirus. It is from the virus family that causes illness ranging from common cold to more serious diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Extreme Acute Respiratory Syndrome (SARS-CoV).

CoronaVirus:

Coronaviruses (CoV) derive their name from the fact that under electron microscopic examination, each virion is surrounded by the corona. Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS - CoV) and Severe Acute Respiratory Syndrome (SARS -CoV). So far, seven types of coronavirus are infecting people.

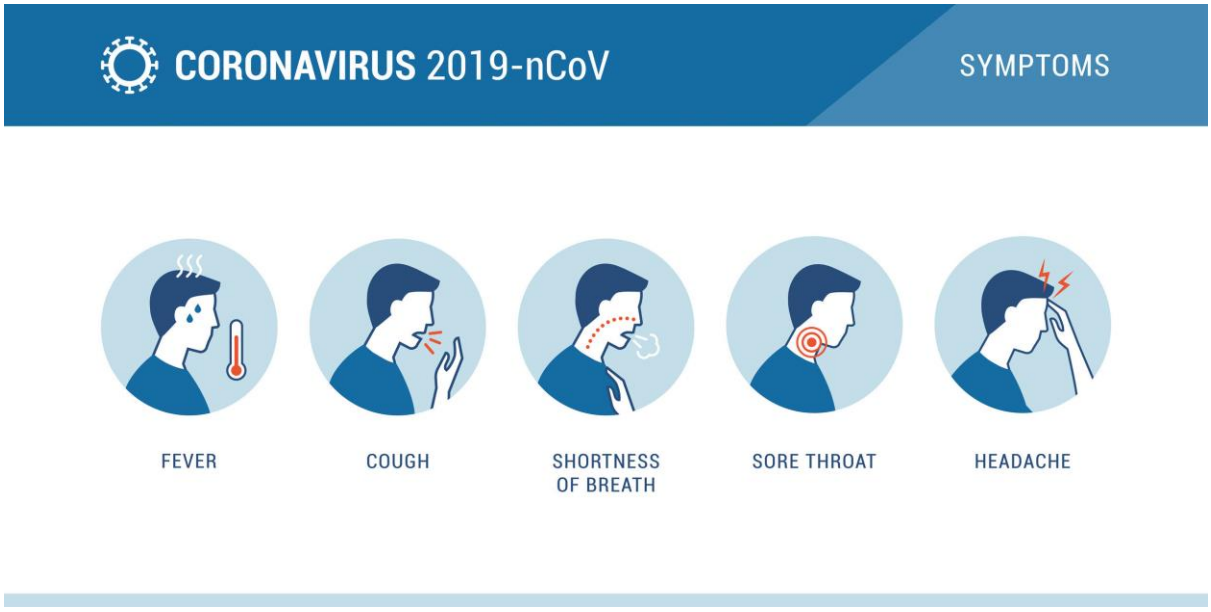


Fig.1.Corona Virus Symptoms

Direct Transmission: Person-to-Person

COVID-19 causes respiratory disease and is mainly transmitted in person-to-person. It can happen in the following circumstances:

- Between people who are in close contact with one another (within about 6 feet)
- Through respiratory droplets produced when an infected person coughs or sneezes
- These droplets can land in the mouths or noses of people who are nearby or possibly be inhaled into the lungs

Indirect Transmission: Other Causes

Contact with Infected Surfaces or Objects A person can possibly get COVID-19 by touching a surface or an object (e.g. doorknobs and table) that has the virus on it and then touching his own mouth, nose, or eyes.

Influential Survey:

	COVID-19	SARS	Influenza	Common Cough
Clinical Manifestations	Excessive fatigue; coughs; shortness of breaths; coughing up yellow or green mucus; chest X-ray shows scattered opacities in the lung	Coughs; breathing difficulties; fatigue; headache and diarrhea; fever	Running nose; sneezing; coughs; high temperature; muscle pain; diarrhea; vomiting	Nasal congestion; coughs; sore throat; throat discomfort; sneezing
Incubation Period	7-14 days	2-7 days	1-4 days	1 day
Ways of Transmission	Short distance droplets spread; close contact; contacts with animals	Short distance droplets spread; close contact	Coughs; sneezing and droplets spread; contact with secretions of an infected person	Droplets spread; contact with infected nasal secretions
Preventive Measures	Regular and frequent hand washing; check body temperature; use alcohol-based disinfectant; wear a surgical mask; enhance airflow; avoid contacts with animals or eat game meat	Cover mouth and nose when sneezing and coughing; regular and frequent hand washing; do not touch nose and mouth; wear a surgical mask; enhance airflow	Vaccination (flu shot); keep hands clean; wear a surgical mask; improve airflow	Regular hand wash, wear a surgical mask, boost your immune system

Statistics-as-on-date:

As per the statistics given by **Center for Systems Science and Engineering (CSSE)** is a research collective housed within the Department of Civil and Systems Engineering (CaSE) at **Johns Hopkins University (JHU)**. The team of CSSE works on a range of complex and interdisciplinary problems, united by the goal to better understand and improve societal, health, and technological systems for everyone**.**

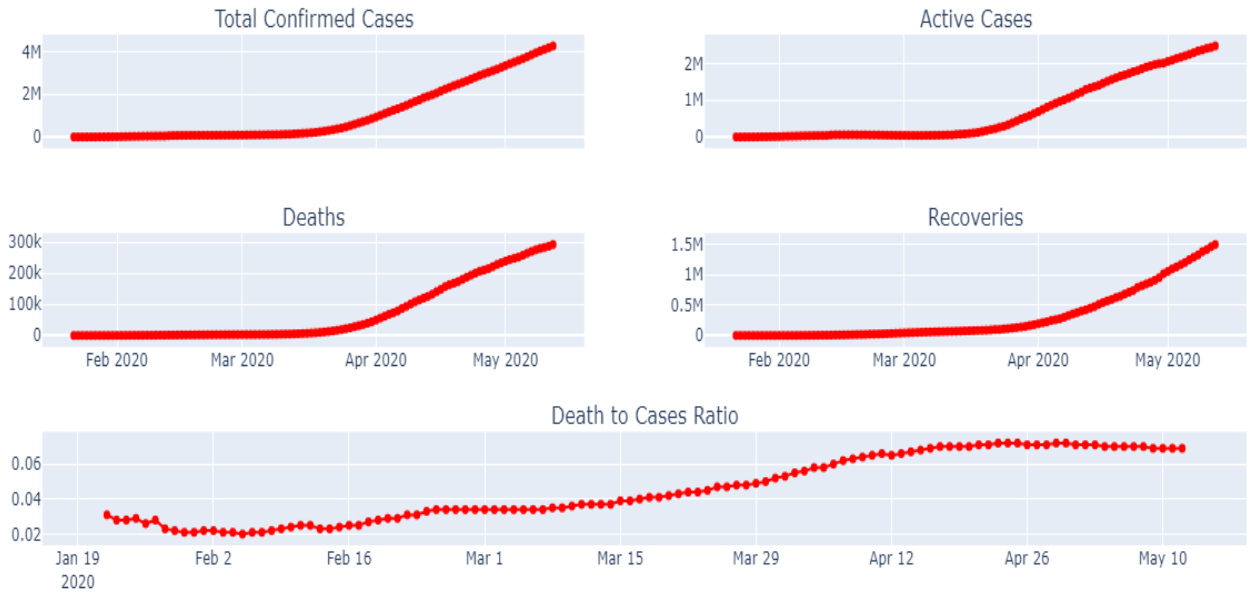


Fig.2. Statistics-As-On-Date by CSSE, JHU

Problem Statement:

With respect to the COVID-19 outbreak, the WHO Secretariat works with Taiwanese health experts and authorities, following established procedures, to facilitate a fast and effective response and ensure connection and information flow.

The innovators who are leveraging disruptive technologies to work on it and find unique and decisive solutions to improve the management of the pandemic and contain further outbreaks. The new ideas that emerge will help us and our countries to step back and observe the changes and figure out ways of taking advantage of a horizon of innovative opportunities that are emerging.

Research Objectives:

In this capstone project, the spread of the COVID-19 pandemic across large number of nations is in recent times is collected and analysed to identify the average recovery status of each country and visualized the same using geo maps to identify the clustered zones to predict the cause for recovery.

The focus areas for this project are as follows:

- Easy detection of infected persons in each country and recovery status
- Regular monitoring of the spread of the virus and predict outcomes
- Identifying the clustered zones to predict the reason for recovery
- Low cost and easy to implement

Dataset Description:

The dataset has been collected from an interactive web-based dashboard hosted by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, to visualize and track reported cases in real-time. The dashboard, first shared publicly on January 22, illustrates the location and number of confirmed COVID-19 cases, deaths and recoveries for all affected countries. It was developed to provide researchers, public health authorities and the general public with a user-friendly tool to track the outbreak as it unfolds. Further, all the data collected and displayed is made freely available, initially as google sheets, now in a GitHub repository, along with the feature layers of the dashboard, which are now included in the ESRI Living Atlas.

Additional data sources are relied upon for reporting on regions outside China. These include U.S. county and state health departments, multiple national government health departments, as well as data aggregating websites including 1point3acres, Worldometers.info, BNO and the COVID Tracking Project (testing and hospitalizations), which rely on a combination of reporting from local health departments and local media reports. The full list of sources is maintained on our **CSSE COVID19 GitHub Repository**. All dashboard data curation and updates are coordinated by a team at JHU.

Methodology:

This capstone project has been collected "**time_series_covid19_recovered_global.csv**", it consists of **256 Countries** information represented as **rows** and **116 fields**(attributes) including **id, province/state, country/region, latitude and longitude** and also day wise recovery status from **22nd January,2020 to 12th May,2020** etc.

This project is aimed to address the following.

- Day-wise recovery status of each country and/or province or State.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to countries to find out the clustering on the neighbourhoods.
- Understanding the correlation between the attributes
- This analysis helps all others to predict the reasons for healthy recovery like lockdown, maintaining social distance, work from home etc.

Phase-1: Data Preparation:

Importing of the required packages and integrating dataset for analysis.

```
# Data Preparation
```

In [ ]:
```
import folium
import pandas as pd
```

In [2]:
```
country_geo = 'world-countries.json'
```

In [3]:
```
data = pd.read_csv('recovered.csv')
data.shape
```

Out[3]: (252, 116)

In [4]:
```
data.head()
```

Out[4]:

ntry/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/3/20	5/4/20	5/5/20	5/6/20	5/7/20	5/8/20	5/9/20	5/10/20	5/11/20	5/12/20
Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	345	397	421	458	468	472	502	558	558	610
Albania	41.1533	20.1683	0	0	0	0	0	0	...	531	543	570	595	605	620	627	650	654	682
Algeria	28.0339	1.6596	0	0	0	0	0	0	...	1936	1998	2067	2197	2323	2467	2546	2678	2841	2998
Andorra	42.5063	1.5218	0	0	0	0	0	0	...	493	499	514	521	526	537	545	550	550	568
Angola	-11.2027	17.8739	0	0	0	0	0	0	...	11	11	11	11	11	11	13	13	13	13

Phase-2: Data Analysis

In this phase the mean recovery rate of each country is calculated and sorted according to the highest mean to the lowest mean (Descending order). It give us the information about the countries their average recovery rate is good.

```
In [5]: data['mean'] = data.mean(axis=1)
```

```
In [6]: sorted_df = data.sort_values(by='mean', ascending=False)
        print(sorted_df)
```

```
                       Province/State  Country/Region      Lat      Long  \
53                             Hubei           China  30.9756  112.2707
225                              NaN              US  37.0902  -95.7129
112                              NaN         Germany  51.0000    9.0000
199                              NaN           Spain  40.0000   -4.0000
127                              NaN            Iran  32.0000   53.0000
..                               ...             ...      ...       ...
250                              NaN         Comoros -11.6455   43.3333
235                         Anguilla  United Kingdom  18.2206  -63.0686
238                              NaN      MS Zaandam   0.0000    0.0000
245          Saint Pierre and Miquelon          France  46.8852  -56.3159
242  Bonaire, Sint Eustatius and Saba     Netherlands  12.1784  -68.2385

     1/22/20  1/23/20  1/24/20  1/25/20  1/26/20  1/27/20  ...  5/4/20  \
53        28       28       31       32       42       45  ...   63616
225        0        0        0        0        0        0  ...  187180
112        0        0        0        0        0        0  ...  132700
199        0        0        0        0        0        0  ...  121343
127        0        0        0        0        0        0  ...   79379
..       ...      ...      ...      ...      ...      ...  ...     ...
250        0        0        0        0        0        0  ...       0
235        0        0        0        0        0        0  ...       3
238        0        0        0        0        0        0  ...       0
```

Duplicates removed as shown below.

```
In [7]: sorted_df.T.drop_duplicates().T
```

Out[7]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/4/20	5/5/20	5/6/20	5/7/20	5/8/20	5/9/20
53	Hubei	China	30.9756	112.271	28	28	31	32	42	45	...	63616	63616	63616	63616	63616	63616
225	NaN	US	37.0902	-95.7129	0	0	0	0	0	0	...	187180	189791	189910	195036	198993	212534
112	NaN	Germany	51	9	0	0	0	0	0	0	...	132700	135100	139900	141700	141700	143300
199	NaN	Spain	40	-4	0	0	0	0	0	0	...	121343	123486	126002	128511	131148	133952
127	NaN	Iran	32	53	0	0	0	0	0	0	...	79379	80475	81587	82744	83837	85064
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
250	NaN	Comoros	-11.6455	43.3333	0	0	0	0	0	0	...	0	0	0	0	0	0
235	Anguilla	United Kingdom	18.2206	-63.0686	0	0	0	0	0	0	...	3	3	3	3	3	3
238	NaN	MS Zaandam	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
245	Saint Pierre and Miquelon	France	46.8852	-56.3159	0	0	0	0	0	0	...	0	0	0	0	0	0
242	Bonaire, Sint Eustatius and Saba	Netherlands	12.1784	-68.2385	0	0	0	0	0	0	...	0	0	0	0	0	0

252 rows × 117 columns

The top 100 countries with highest mean recovery rate is considered and their latitude and Longitude positions are gathered and moved to another variable i.e. **Location (Loc).**

```
In [8]: import matplotlib.pyplot as plt
        data_to_plot = sorted_df[['Country/Region','Lat','Long','mean']]
        data_to_plot = data_to_plot.head(100)
        loc = sorted_df[['Lat','Long']]
        data_to_plot.rename(columns = {'Country/Region':'Region'}, inplace = True)
        data_to_plot.plot(kind ='hist')
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x22839abc288>
```

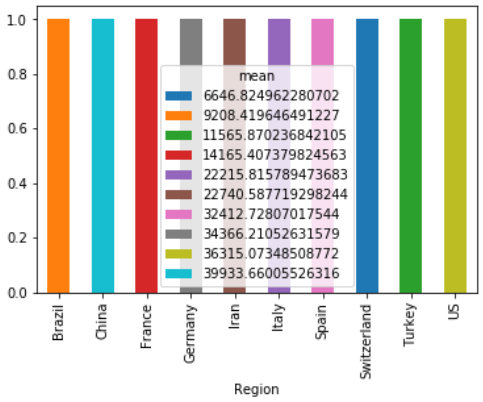
```
In [9]: loc.head(10)
```

Out[9]:

	Lat	Long
53	30.9756	112.2707
225	37.0902	-95.7129
112	51.0000	9.0000
199	40.0000	-4.0000
127	32.0000	53.0000
131	43.0000	12.0000
108	46.2276	2.2137
213	38.9637	35.2433
29	-14.2350	-51.9253
204	46.8182	8.2275

The countries with the highest mean recovery rate is plotted by **groupby()** function between the **region** and **mean** attributes.

```
In [15]: data_to_plot1= data_to_plot.head(10)

         data_to_plot1.groupby(['Region','mean']).size().unstack().plot(kind='bar',stacked=True)
         plt.show()
```



Phase-3: Visualizing the maps and Clustered Neighbourhoods

The following is the choropleth map to visualise the countries with highest mean recovery rates according to their geo coordinates

Visualising the Map

In [16]:
```
map = folium.Map(location=[100, 0], zoom_start=1.5)
country_geo = 'world-countries.json'
```

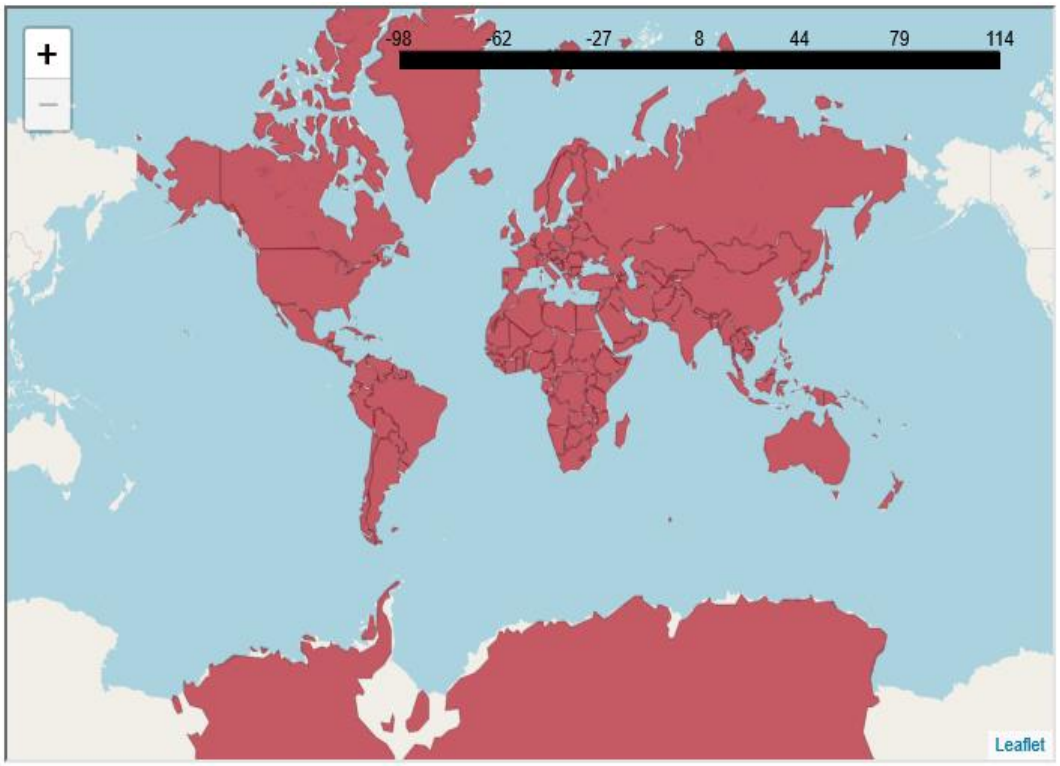
In [18]:
```
map.choropleth(geo_data= country_geo , data=data_to_plot1,
               columns=['Lat','Long'],
               key_on='mean',
               fill_color='RdBu', fill_opacity=0.7, line_opacity=0.2)
```

In [19]:
```
map.save('plot_data.html')
```

In [20]:
```
# Import the Folium interactive html file
from IPython.display import HTML
HTML('<iframe src=plot_data.html width=700 height=450></iframe>')
```

```
C:\Users\babby\Anaconda3\lib\site-packages\IPython\core\display.py:694: UserWarning: Consider using IPython.display.IFrame inst
ead
  warnings.warn("Consider using IPython.display.IFrame instead")
```

Out[20]:



Final Results:

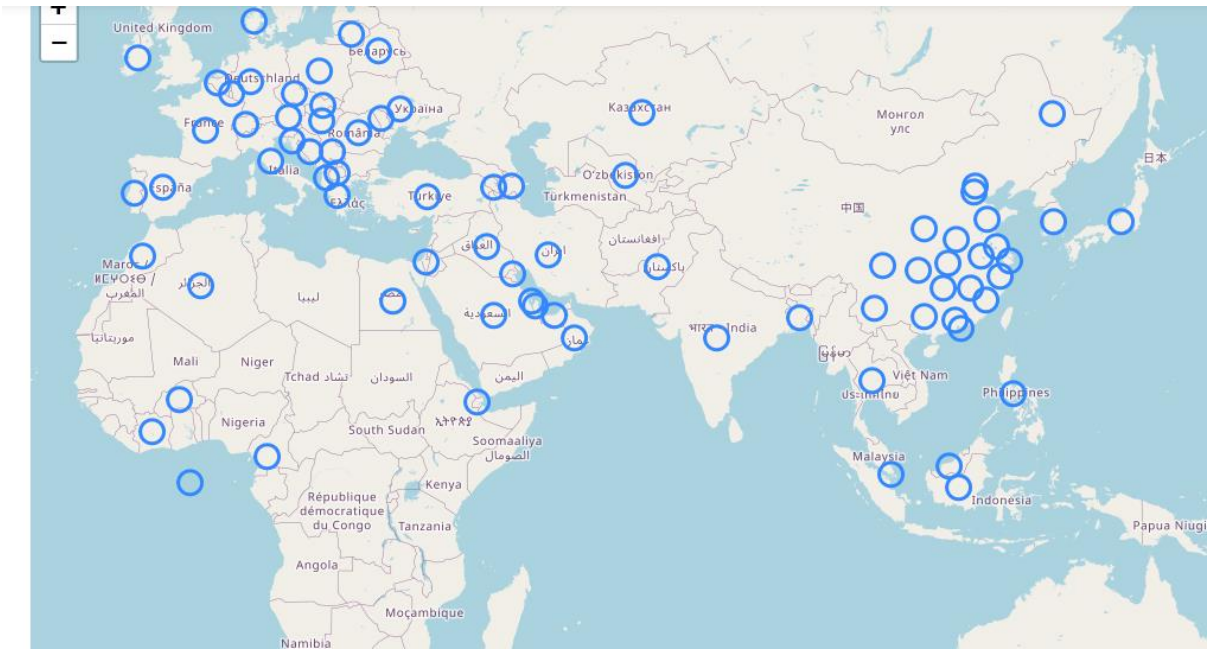
Clustering Neighbourhoods identification with Folium Maps:

Clustering Neighbourhood & Visualization

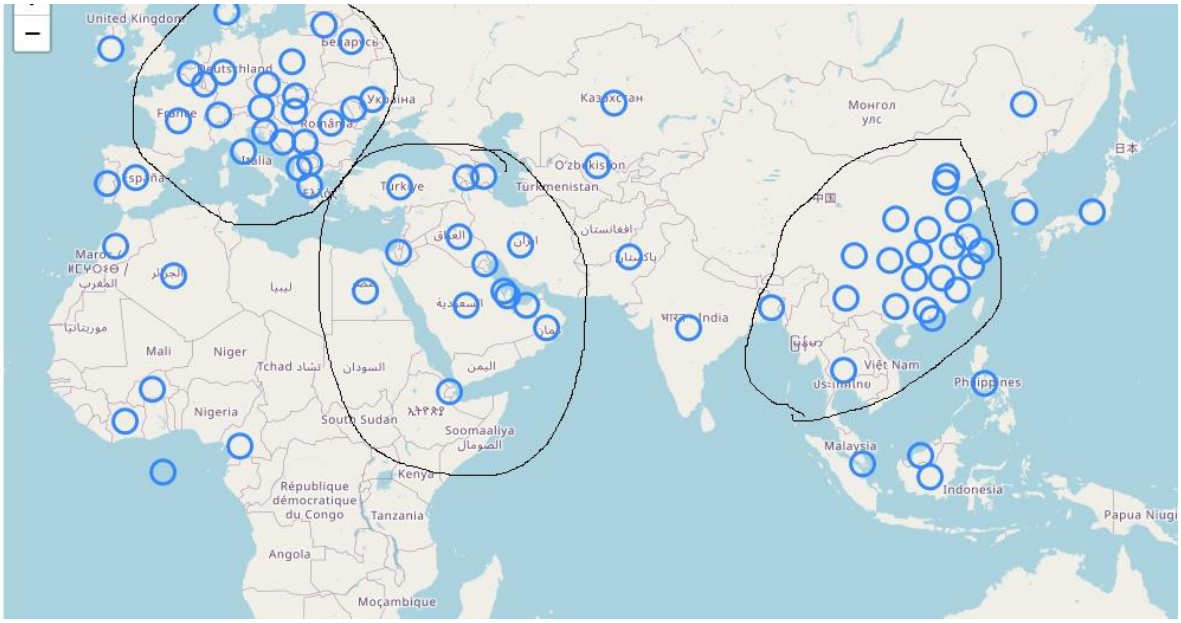
In [21]:
```
import folium
from folium.plugins import MarkerCluster

coord = [10, 0]

map1 = folium.Map(location=coord, zoom_start=12)
marker_cluster = MarkerCluster().add_to(map1)

for each in loc[0:100].iterrows():
   folium.CircleMarker(location = [each[1]['Lat'],each[1]['Long']],
    clustered_marker = True, tiles='Covid 19').add_to(map1)
```

In [22]:
```
map1.save('map.html')
```

In [23]:
```
# Import the Folium interactive html file
from IPython.display import HTML
HTML('<iframe src= map.html width=1000 height=750></iframe>')
```



The countries/regions which are blue coloured circles indicates the locations where mean recovery rate is high.



Cluster the nearby geographical locations to understand/predict the reason for highest recovery like as mentioned lockdown, social distance and Work-from-home

Conclusion:

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS - CoV) and Severe Acute Respiratory Syndrome (SARS -CoV). So far, seven types of coronavirus are infecting people. WHO suggested the innovators who are leveraging disruptive technologies to work on it and find unique and decisive solutions to improve the management of the pandemic and contain further outbreaks.

This capstone project had been built on “time_series_covid19_recovered_global.csv” dataset, it consists of 256 Countries information represented. Initially, this project analysed the Day-wise recovery status of each country and/or province or State. Next, geo coordinate information of those neighbourhoods are separated to plot the map and also to get the venue data. Finally, it visualises the neighbourhoods information to identify the clustered regions and to predict the cause for sustainable health maintenance factors.

Future Enhancements:

This project visualised the geo coordinates of the neighbourhood countries. One can apply a better clustering algorithm to group the countries and classify them as **continent_based**, **weather_based** and **living_style** based clusters. They can also predict the clustered reasons and causes for their sustainable and average growth in the recovery from Corona Virus Disease.

References:

- https://www.mohfw.gov.in/pdf/DGSOrder04of2020.pdf
- https://www.mygov.in/hi/covid-19/
- https://www.codechef.com/COVDHACK?itm_campaign=contest_listing
- https://systems.jhu.edu/research/public-health/ncov/