Visual Exploratory Data Analysis of COVID-19 Pandemic

Sumindar Kaur Saini
Computer Science and Engineering , UIET, Panjab University, Chandigarh, India
sumindarkaursaini@gmail.com

Vishal Dhull
Computer Science and Engineering , UIET, Panjab University, Chandigarh, India
mdhull07@gmail.com

Sarbjeet Singh
Computer Science and Engineering , UIET, Panjab University, Chandigarh, India
sarbjeet@pu.ac.in

Akashdeep Sharma
Computer Science and Engineering , UIET, Panjab University, Chandigarh, India
akashdeep@pu.ac.in

2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 978-1-7281-8867-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICRAIE51050.2020.9358331

***Abstract*— The emerging novel coronavirus (2019-nCoV) caused by a respiratory syndrome coronavirus 2 (SARS-CoV-2) is the lead cause of threat to life worldwide today. It is important to analyze the worldwide pandemic spread so that certain guide strategies can be set for complete situational awareness and application of conventional methodologies to control the impacts caused by it globally. This paper is composed of the visual exploratory data analysis of the countries based on the number of confirmed, recovered and death cases along with the comparative analysis of the mortality and recovery rate for nearly 222 nations worldwide. Also, K-means clustering is used to cluster the countries according to the number of confirmed and death cases. Hence, this study can be used to evaluate the rise of risks in a given area by comparing the count of the cases via visual analysis and work on the set up of some strategies to control its spread globally.**

Keywords - **COVID-19; Pneumonia; Data Analysis; Pandemic; Clustering; Coronavirus;**

I. INTRODUCTION

There have been recent pneumonia outbreaks worldwide due to the deadly 2019 novel coronavirus (2019-nCoV) belonging to the Orthocoronavirinae subfamily. It is entirely distinct from the MERS-CoV and SARS-CoV and is the seventh member of this family [1]. A new form of pneumonia was detected by the Chinese Center for Disease Control and Prevention (CDC) on 12 December 2019 using cell cultures and molecular techniques and it was tested to be non-SARS nCoV. Coronaviridae family is a group of single, large-sized and plus stranded RNA viruses that can cause cold and diarrheal diseases [2-3]. The increase in the number of cases in Wuhan a province in Hubei, China in 2019 reflected the spread of viral pneumonia [4]. The virus has spread to almost all parts of the world. It is supposed to have originated from the contact with local fish and wild animal markets in certain parts of Wuhan leading to the transmission of virus to humans and interhuman transmission [5-6]. The World Health Organization (WHO) renamed this epidemic virus as coronavirus disease (COVID-19). As per the situation report–144 given by the WHO on June 12, 2020, there are 7,410,510 cases of COVID-19 worldwide and 418,294 deaths globally. China confirmed nearly 84,659 cases with 7 total confirmed new cases and 4,645 deaths, whereas in the other parts of the world, the cases are increasing rapidly. 2019-nCoV has been renamed as SARS-CoV-2 by the International Committee on Virus Taxonomy. COVID-19 has been declared a global pandemic so the exploratory data analysis (EDA) is necessary to locate such areas and aware the people in the times to come. The prevention of the pandemic is possible by the cooperation of the governments, health and security officials worldwide. The following section focuses on the study of COVID-19 outbreak with the help of some basic visualizations techniques.

II. DATASET DESCRIPTION

The dataset used is an open dataset of 2019-nCoV contributed by Johns Hopkins University wherein there is daily updation of the total confirmed cases and deaths worldwide. The data from different countries has been shown on a dashboard and is provided in the form of Google sheets, wherein the daily reports of the cases are updated and analyzed helping the data scientists and researchers to work on the data analysis [8]. The dashboard presents the total number of confirmed cases, deaths, and recovery on the cases of COVID-19 on a cumulative basis. We have used the dataset from 22 January, 2020 to 12 June, 2020 for the predictive analysis. The dataset consists of 8 columns and 21866 rows with the data from 222 countries.

III. DATA ANALYSIS

The analysis was done on the total active cases, confirmed cases, deaths and recovery worldwide as provided in the dataset from 22 January 2020 to 12 June, 2020. Then the comparison was made with the different types of cases within India. Till now, the count of confirmed cases globally is 7632802.0 and the total count of recovered cases around the World is 3613277.0. The total count of death cases reached a value of 425394.0 globally. This increase in the total number of active cases is an indication that the recovered cases and the total number of global deaths are dropping in comparison to the count of confirmed cases drastically. Figure 1 gives an insight to the weekly increase in the number of death cases and the death cases as these results are significant in analyzing the overall effect on COVID-19 on the total population of 222 countries globally. As it is represented the weekly rise in the number of such cases declined in the 21st week due to the lockdown and other measures taken by different governments worldwide.

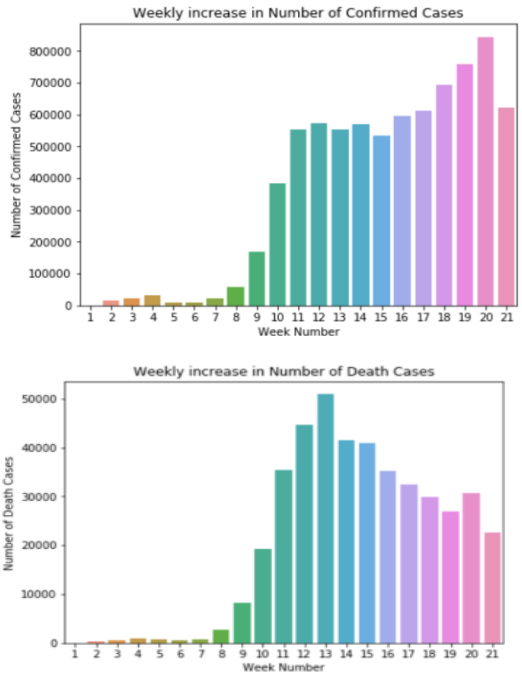

Figure 1: Weekly Increase in number of confirmed COVID-19 cases and number of death cases

Figure 2 represents the total number of confirmed cases (in blue colour), recovered cases (in orange) and deaths (in green) from 22 January 2020 to 12 June 2020 and the average increase in the number of Confirmed Cases every day is 53372.0 whereas for the recovered cases it is nearly 25267.0 on a daily basis and for the death cases, the value is nearly 2975.0.

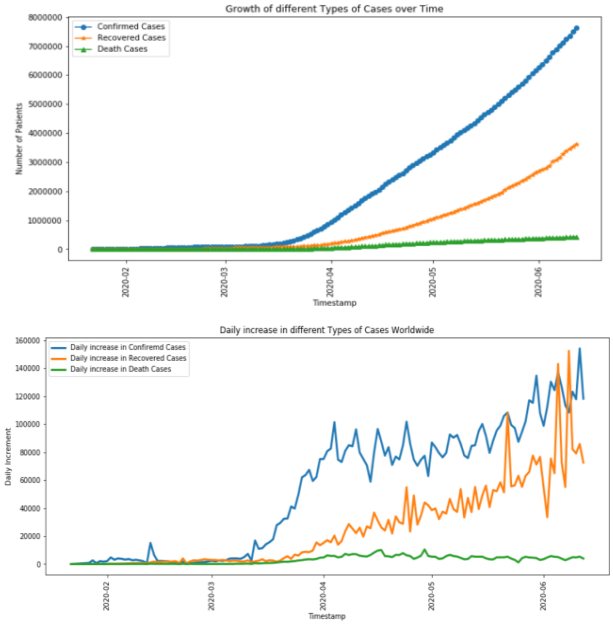

Figure 2: The growth of different types of cases over time and daily increase in different types of cases worldwide

Equation 1 gives the mortality rate and equation 2 gives the recovery rate. The analysis over the dataset gave the value of average mortality rate as 4.8451 and median mortality rate as 5.2287.Whereas the average recovery rate was calculated to be 29.4634 and the value of median recovery rate was calculated to be 30.1022. Figure 3 gives a graphical representation of the overall date wise mortality and recovery rate. There is a significant dip in the mortality rate over the past few days in May signifying a positive review in the deaths due to COVID-19.There has been an improvement in the recovery rate after April leading to an increase in the number of closed cases.

$$Mortality\ rate = \frac{(Number\ of\ Death\ Cases)}{(Number\ of\ Confirmed\ Cases)}\ x\ 100 \qquad (1)$$

$$Recovery\ rate = \frac{(Number\ of\ Recovered\ Cases)}{(Number\ of\ Confirmed\ Cases)}\ x\ 100 \qquad (2)$$

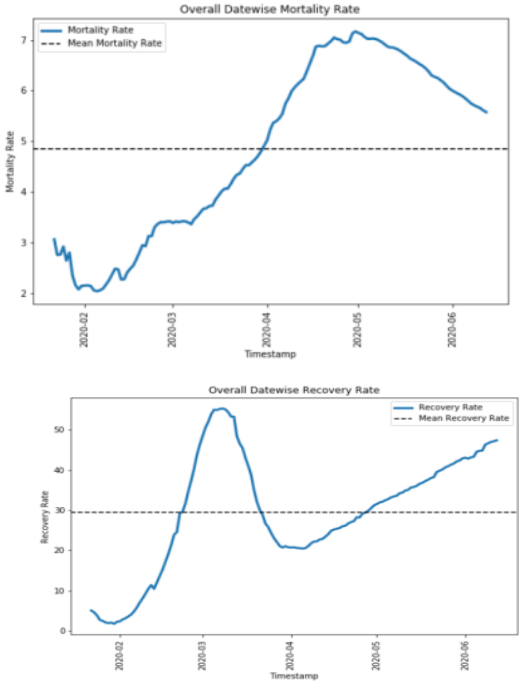

Figure 3: Overall Date Wise Mortality and Recovery Rate

Growth factor is the rate by which a quantity multiplies itself over time. Equation 3 represents the growth factor. It is an ideal approach to determine the pattern of the spread of COVID-19 as it measures the rate based on the confirmed cases, recovered cases and deaths of the two consecutive days. Figure 4 represents the growth factor of different types of cases worldwide.

$$Growth\ factor = \frac{Daily\ (Confirmed, Recovered, Deaths)}{\text{New (Confirmed,Recovered,Deaths) on previous day}} \qquad (3)$$

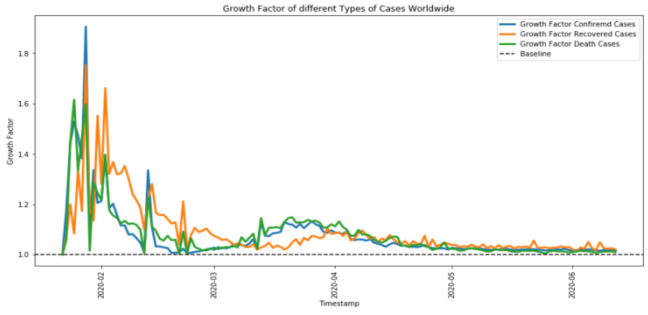

Figure 4: Growth factor of different types of cases worldwide

The visual exploratory analysis has been performed on the dataset of different countries and table 1 gives a vivid description of the total number of confirmed cases, recovered cases, deaths, mortality rate and recovery rate in 5 nations and similarly the analysis for the remaining nations has been done on a global scale.

Table 1: Description of different types of cases in different countries

Country	Confirmed Cases	Recovered Cases	Deaths	Mortality	Recovery Rate
US	2048986	547386	114669	5.5963	26.71497
Spain	828810	445123	41828	5.0467	53.70627
Italy	510761	268862	6705	1.3127	52.63949
UK	297535	147195	8498	2.8561	49.47149
France	294402	1282	41566	14.1187	0.435459

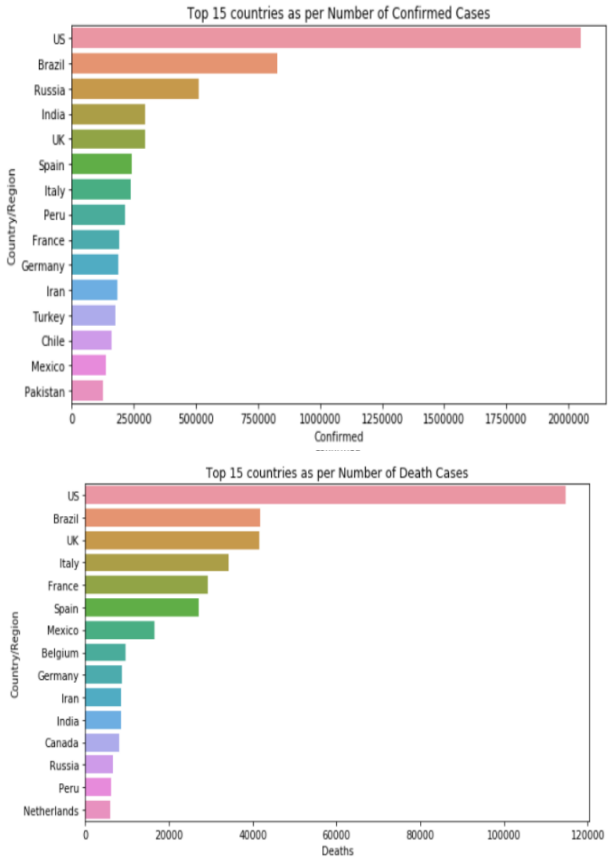

Figure 5: Top 15 countries as per the total number of confirmed cases and death cases

Figure 5 represents the top 15 countries with respect to the total number of confirmed and death cases of COVID-19. Table 2 represents that MS Zaandam had nearly 9 confirmed cases, 2 deaths with the mortality rate 22.22 whereas Sweden had nearly 49684 confirmed cases, 4854 deaths with the mortality rate as 9.76974.These nations had no recovered patients, but comparatively had a low number of confirmed cases than the other nations.

Table 2: Description of nations with no recovered patient but with lesser number of confirmed cases

Country/Region	Confirmed Cases	Deaths	Mortality Rate
MS Zaandam	9	2	22.2222
Sweden	49684	4854	9.76974

Figure 6 represents the position of countries with respect to high recovery rate and low recovery rate.

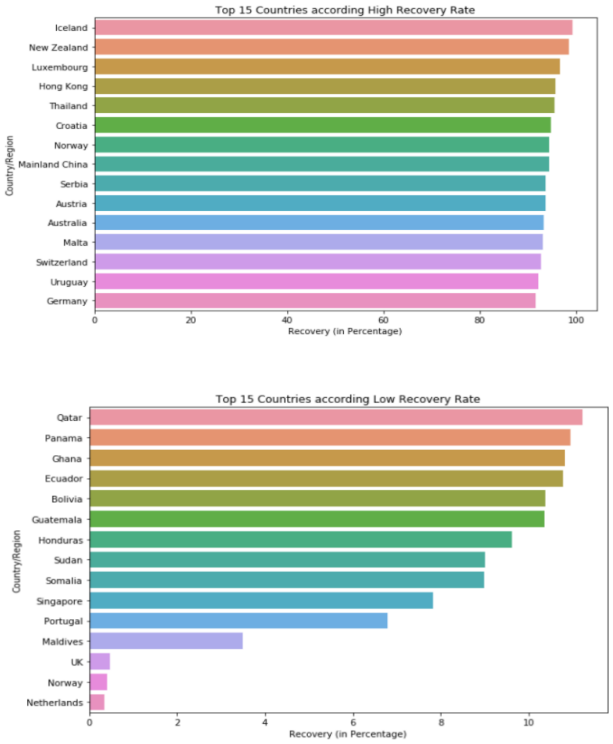

Figure 6: Top 15 countries according to high recovery rate and low recovery rate

There are certain nations that have more than 100 confirmed cases and 0 deaths with a considerably high recovery rate. The table 3 represents that Cambodia, Vietnam, Mongolia and Uganda have been able to control the transmission of COVID-19 in a positive way with 0 deaths recorded until 12 June 2020 with a healthy recovery rate. Vietnam was the first country to inform the World Health Organization (WHO) about Human to Human Transmission of COVID-19.The analysis represented in the table signifies that Vietnam and Cambodia will soon be free from this deadly virus.

Table 3: Description of nations with zero deaths and a considerable high recovery rate

Country/Region	Confirmed Cases	Recovered Cases	Deaths	Recovery Rate
Cambodia	126	125	0	99.2063
Vietnam	333	323	0	96.997
Mongolia	197	95	0	48.2234
Uganda	686	161	0	23.4694

Figure 7(a) represents the variation of mortality rate in Mainland China, Italy, US, Spain and rest of the world with respect to time. It is clearly visible from the graph that Italy stands at the top among the mentioned countries with 14.48% followed by Spain 11.15%, US with 5.9%, Mainland China with 5.5% and for rest of world overall its value is 4.87% as of June 12,2020 which shows that Mainland China is using better control method as compared to the rest of the nations.

Figure 7(b) represents the variation of recovery rate in Mainland China, Italy, US, Spain and rest of the world with respect to time. It is clearly visible from the graph that Mainland China stands at the top among the mentioned countries with 94.33% followed by Spain 61.82% ,US with 26.71% and for the rest of the world overall it's value is 53.056% as of May 5, 2020 . Taking off the recovery rate of Spain, it is a good sign but it's nowhere in comparison to the mortality rate. Its alarming sign for the US as the recovery rate is dropping down as compared to the rise in the mortality rate.. The total number of closed cases in a country reflect that either the recovery rate is high or the number of deaths due to COVID-19 have increased. So it is important to analyze the top 15 countries with the most number of closed cases as represented in the figure 8. The U.S. tops the list followed by Brazil, Russia, Italy, Germany, Spain, India, Turkey and Iran. So the most number of closed cases have been observed in the US due to higher number of deaths and moderate mortality rate. A country with the greater number of confirmed cases per day is unable to control the spread of the COVID-19 virus due to which there is a significant rise in the number of deaths and the mortality rate.

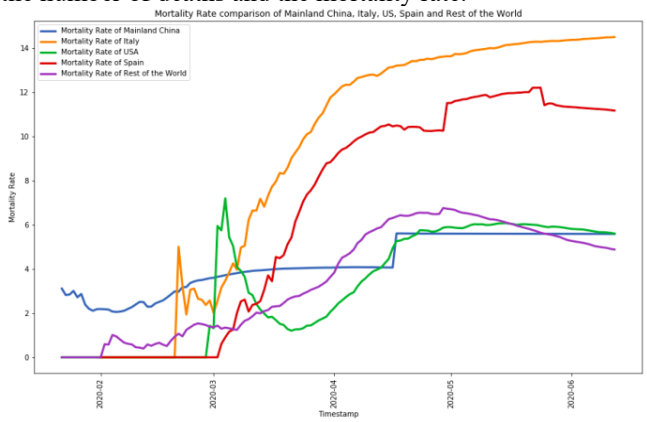

Figure 7(a): Mortality rate comparison

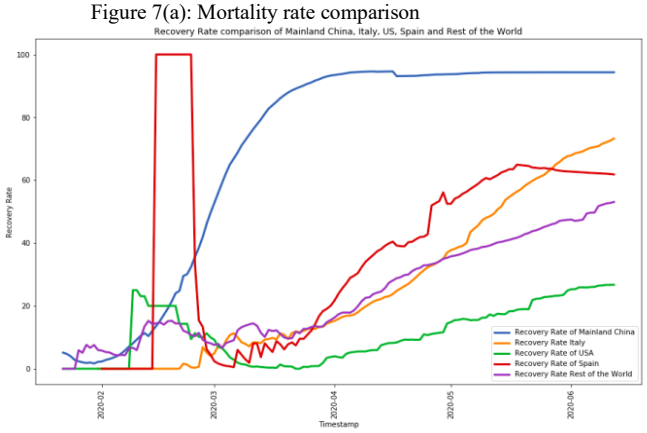

Figure 7(b): Recovery rate comparison

Figure 9 represents the top 15 countries with respect to the higher number of confirmed cases per day wherein the US is having the maximum number of such cases followed by Brazil, Russia, India, UK , Peru and Turkey. The number of confirmed cases per day were nearly 14328.573427 in the USA,7603.761468 in Brazil, 3811.649254 in Russia, 2203.962963 in India and 2197.029851 in the UK. This is evident from the analysis that the situation of the COVID-19 in the US is very poor and out of control and will definitely take much more time than the other countries to eradicate this virus out of the nation.

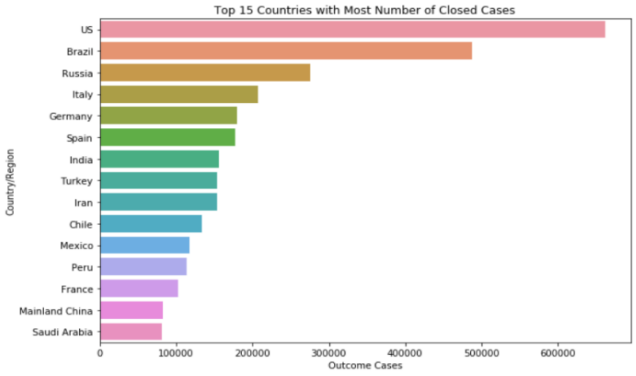

Figure 8: Top 15 countries with the most number of closed cases

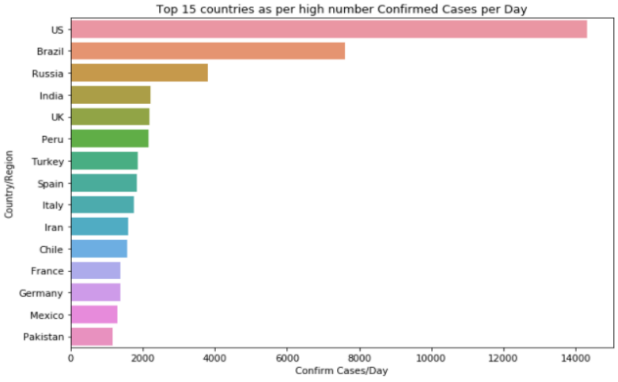

Figure 9: Top 15 countries as per high number of confirmed cases per day

There are some countries wherein the total number of confirmed cases until 12 June, 2020 are more than 1000 however the number of confirmed cases per day is the lowest in number as represented in figure 10. Slovakia is on the top of such nations followed by Kosovo, Zambia, Sierra Leone, Slovenia, Albania, Madagascar, Equatorial Guinea, New Zealand and Sri Lanka. This signifies that these countries have been able to control the spread of this deadly virus and there will be a significant reduction in the mortality rates in these nations. It is pretty evident that the disease is spreading in a similar manner everywhere in the world, but if a particular country is following the pandemic controlling practices rigorously then the results are proved to be better as evident from the figure 11 . Most of the countries follow the same trajectory as that of US, which is "Uncontrolled Exponential Growth". The classic example is Germany wherein the graph shows a sharp dip, which is an evidence that these nations are able to control the spread of the COVID-19. Italy seems to be heading towards that similar dip, showing the pandemic control practices are working in favour.

The figure 12(a) represents the daily increase in the number of confirmed cases in China, Italy, US, Spain, and the rest of the world with respect to time. It is clearly visible from

the graph that while in other countries like the US, Italy and Spain the number of confirmed cases are increasing exponentially. Mainland China has achieved a flatten curve with less than 10 number of increase in cases per day when observed from the past few days.

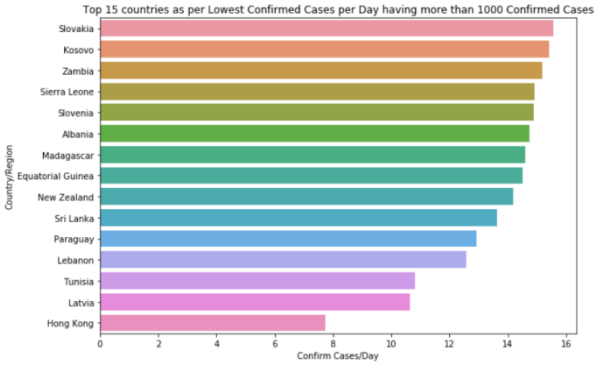

Figure 10: Top 15 countries as per lowest confirmed cases per day having more than 1000 confirmed cases

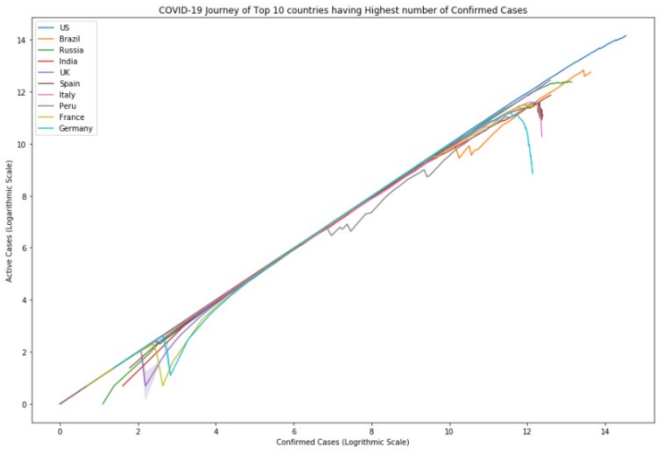

Figure 11: Journey of top 10 countries having the highest number of confirmed cases

The figure 12(b) represents the daily increase in the number of death cases in China, Italy, US, Spain, and the rest of the world with respect to time. It is clearly visible from the graph that while in other countries like the US, Italy and Spain the number of death cases are increasing. Mainland China has achieved a flatten curve when observed from the past few days.

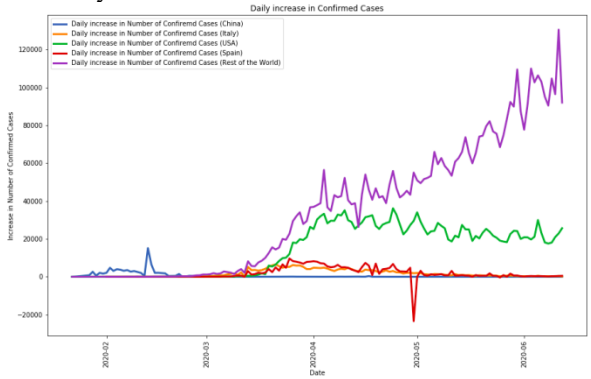

Figure 12(a): Daily increase in number of confirmed ases

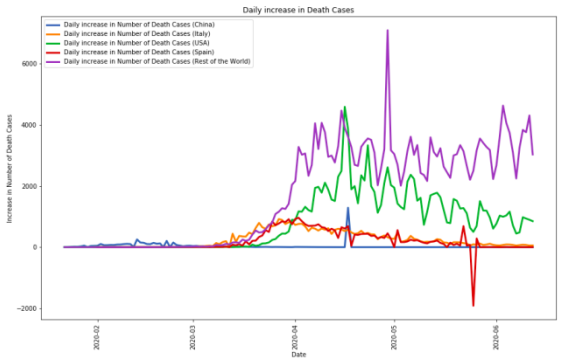

Figure 12(b): Daily increase in number of Death cases

IV. K-MEANS CLUSTERING

The countries were clustered according to the number of their confirmed cases and number of deaths using K-means Clustering with cluster initialisation. In order to find the appropriate number of clusters by grouping the countries with similar conditions, two methods namely, Elbow and Silhouette scores have been used.

1. ***Elbow method:*** Dataset was clustered for a range of values for k using k-means clustering and then for each value of k, an average score was computed for all clusters as represented in table 4. Here, Inertia is the sum of squared error for each cluster. Therefore the smaller the inertia the denser the cluster (i.e. closer are all the points) as represented in figure 13.

Table 4: Description of formation of clusters

K (number of clusters)	**WCSS (Within Cluster Sum of Squares)**
1	113.95
2	37.36
3	21.16
4	10.95
5	6.93
6	5.16
7	4.07
8	3.12

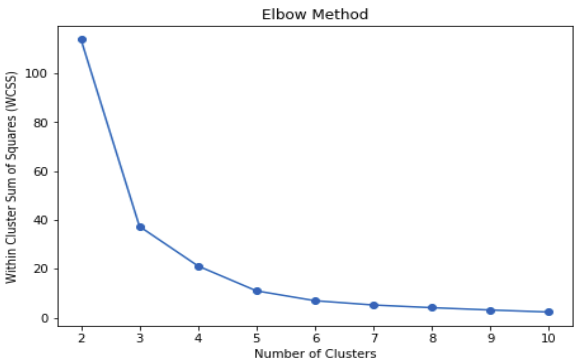

Figure 13: Inertia value with increasing number of cluster

2. ***Silhouette Score:*** The value ranges from -1 to 1 and shows how close or far the clusters are from each other and how dense the clusters are.

Table 5: Description of formation of clusters using Silhouette score

K (number of clusters)	Silhouette Score
1	0.964
2	0.912
3	0.859
4	0.863
5	0.846
6	0.794
7	0.762
8	0.730

Figure 14 represents the variation of Inertia (within clusters the sum of squared distances has been used) and Silhouette score with increasing number of clusters chosen. The best value for the number of clusters as the value of k was found to be 3.

Figure 14: Variation of Inertia

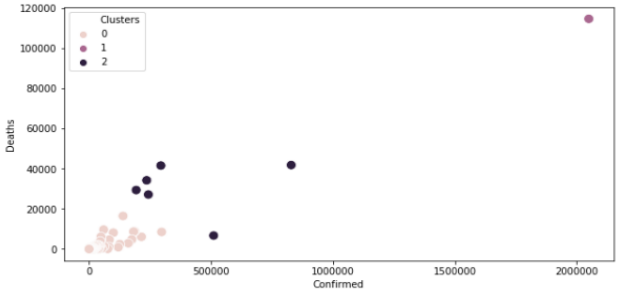

Figure 15: Graphical representation of clustering

Figure 15 shows the results of the graphical k-means clustering algorithm by dividing the 222 nations in the groups of 3 namely (0, 1, 2). The cluster 0 is a set of countries which are moderately or very less affected, few of these countries might have a very high number of confirmed cases but have comparatively very low number of deaths as in the case of Bangladesh, South Africa and Belgium. The cluster 1 is a set of countries which are severely affected, having a really high number of confirmed and death cases as in the case of the US. The cluster 2 belongs to those countries which are worst affected having a moderately high number of confirmed and death cases, but lower than countries belonging to Cluster 1 as in the case of Italy, UK and France.

V. Conclusion And Future Scope

The pandemic caused by COVID-19 is responsible for the higher mortality rate and lower recovery rate. In this study, timely patterns of the rise and fall of confirmed, deaths and recovery cases have been presented visually wherein the top 15 countries for such types of cases have been highlighted. The mortality and recovery rate of Mainland China, Italy, US, Spain has been done with the rest of the world. The journey of the top 10 countries having the highest number of confirmed cases has been highlighted to predict the scenario of the rest of the world in coming days if proper measures are not taken. The 222 nations have been divided into three clusters using k-means clustering based on the similarity in the different types of cases.The results have been evaluated using the Elbow and Silhouette Score.This evaluation of the patterns of COVID-19 is necessary to calculate the risk factor and optimize the strategies for the effective treatment of the patients.This analysis is to be fed into machine learning models for forecasting the number of confirmed cases, recovery cases and deaths across the globe by analyzing this COVID-19 dataset using time series methods.

References

[1] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382:727-733. https://doi.org/10.1056/NEJMoa2001017

[2] Drosten C, Günther S, Preiser W, et al. Identification of a novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med. 2003;348:1967-1976.

[3] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92:418-423. https://doi.org/10.1002/jmv.25681

[4] WHO. Novel Coronavirus—China January 12, 2020. http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/. Accessed 19 January 2020.

[5] Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan China: the mystery and the miracle. J Med Virol. 2020;92:401-402. https://doi.org/10.1002/jmv.25678

[6] Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020;92:433-440. https://doi.org/10.1002/jmv.25682

[7] World Health Organization. Coronavirus disease 2019 (COVID-19): situation report, 144.

[8] Lauren G. Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE January 23, 2020.