IJCSN - International Journal of Computer Science and Network, Volume 9, Issue 4, August 2020
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Data Analytics and Visualization Techniques of Corona Impact

[1] Abhishek Parikh; [2] Sandeep Shah; [3] Vishvam Bhatt

[1] Optimized Solutions Limited
Ahmedabad, Gujarat 380015, India

[2] Optimized Solutions Limited
Ahmedabad, Gujarat 380015, India

[3] Optimized Solutions Limited
Ahmedabad, Gujarat 380015, India

Abstract - The world right now is facing a pandemic due to novel coronavirus disease (nCOVID) or coronavirus disease -19 (COVID-19). A good explanatory visualization of available dataset will provide insights to understand the behavior of pandemic. In this paper we have worked upon descriptive analysis from the COVID dataset and tried to resolve following problems 1. Behavior of mortality and recovery rate with respect to age and sex 2. Comparison trends for confirmed, active, deaths and recovered cases around the world 3. To Map logarithmic number of cases on World and Indian map. We have taken around 1.5 million points to plot the graphs over the data analytic tools based on python and Knime.

Keywords - *Data Analytic, Python, Knime, Coronavirus visualization, descriptive analytic, machine learning prediction, Linear regression*

1. Introduction

All the countries in the world are trying to cope with the virus such that its impact could be controlled. It is a disease for which symptoms are very vague and highly volatile. But some certain common symptoms are fever, dry cough, headache and tiredness. This disease affects the respiratory system and it is being spread through the droplets of saliva, cough and sneeze of the nose of infected people. COVID-19 was first reported in Wuhan city of Hubei Province of China in December 2019 which later on 11 March 2020 became pandemic declared by the world health organization. This study here deals with effects of corona in various countries with the comparative study of confirmed cases, recovery ratio and death ratio of COVID-19. The comparison is shown for the worldwide cases and it also shows effects in India for different states and districts. With this study, after comparison, it is also known that steps by different countries lead to a better strategy for controlling the cases for coronavirus. Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed. To have great visual understanding, advanced techniques of visualization are lused in python language.

Previous researches on this topic have shown symptomatic comparison with respect to patient number and tried to visualize it [2][3]. Bar graph comparison of male to female number shows direct comparison gender wise. Time series plot of total number of pariet is being shown along with country wise comparison over pie chart and tabular form[2][8]. Some better online blogs have been written on data science which has given a good representation of the data [8]. Recovery rate of including and excluding China in tree diagrams has been shown to get a better view of visualization. Still there are few better ways that we have evaluated that can help to understand the distribution of the Coronavirus spread.

General ways of representation include time series charts or trend plots of patients' condition which we have covered in our representation. The overall research has been divided into four parts 1. Time series plots 2. World map plots 3. Distribution plots 4. Data analytic.

2. Time Series Plot

There are three plots we have represented here. First from it, is comparison of patients condition wise time series plot i.e. active, total, recovered and death of 1.5 million patients shown in below figure. Time series plot shown in Figure 1 clearly represents the behavior of data which is exponential and shows the exponent rise of the cases happened in the world around 15th to 20th March [1].

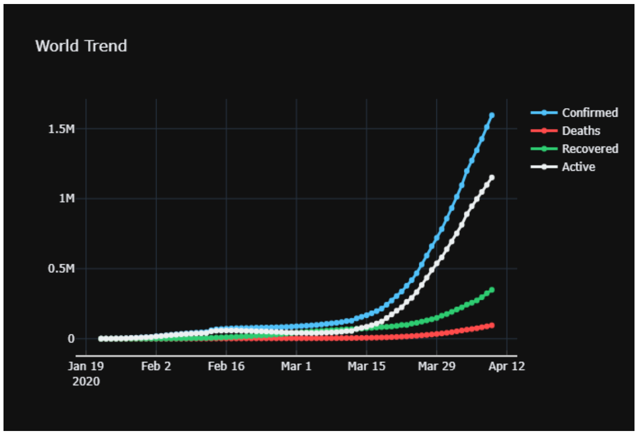


Fig. 1 Time series plot of patient's condition

Once the exponent comes into existence, there becomes a linear rise in the cases same as a typical exponent function has!

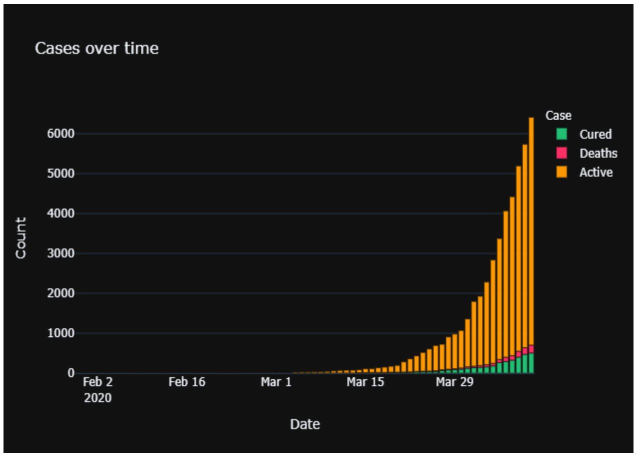


Fig. 2 Indian Map of Cases over the time

The COVID 19 map of Indian Nation's representation has been shown in Figure 2 and Figure 3. Just like in Figure 1, line chart shown for world corona cases, Figure 2 represents time series stacked bar chart of Covid 19 active, cured and death cases

Both of the above figures are time series stacked bar charts for India. Where in the second figure, stacks are state wise total cases with respect to time. Stack bar graph represents the comparison between the stacks as time progresses, hence here it shows spread of Coronavirus with respect to states over time.

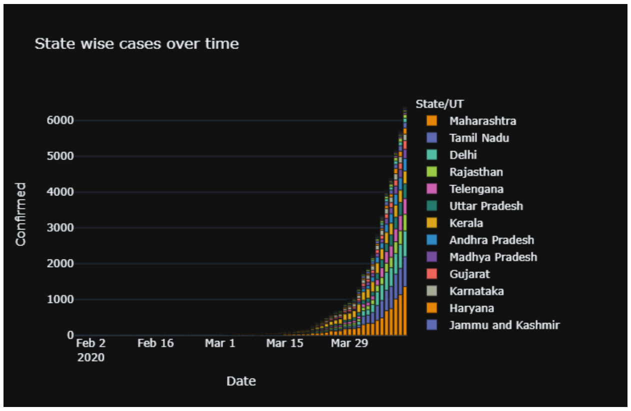


Fig. 3 State Wise Indian map of coronavirus cases

3. World Map Plots

Bubble charts and heat maps on world map is a good way to show the spread of Covid 19 [8]. In the figure below we have represented confirmed, recovered and death cases due to Coronavirus on world map using heat map chart. Bubble charts work well generally with Cartesian coordinate systems but it is pretty bad while shown over the map with smaller regions. If there are more number of points, the bubble chart becomes bigger and overlaps multiple regions. Hence we have used heat map and bubble chart both.

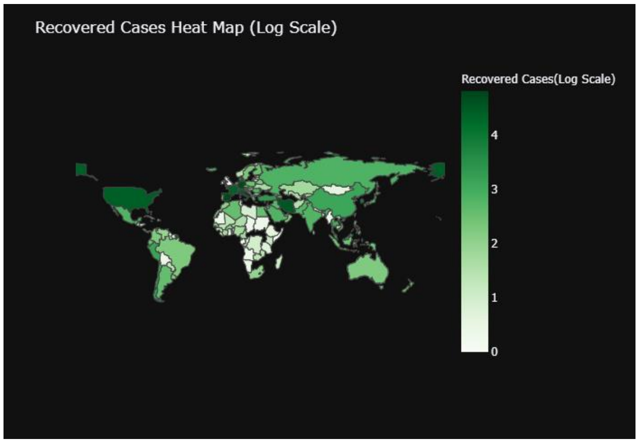

Recovered Cases Heat Map (Log Scale)
Recovered Cases(Log Scale)
4
3
2
1
0

Fig. 4 Recovered cases heat map

Figure 4 represents a recovered coronavirus patients' heat map over the world map. To make the scale we have converted the total number of coronavirus into logarithm scale and then after, we have placed the plot of the log scale of patients. Same as Figure 4, Figure 5 and 6 represents the death cases of the Coronavirus affected patients and the total confirmed Coronavirus patients log scale heat world map.

All these maps show the effect of the virus over different regions of the world in a single view. Also it gives the behavior of the virus with respect to region over mortality. This type of map helps us to understand the death rate between the different regions of the world and further leads to understand why a particular region of the earth is more affected.

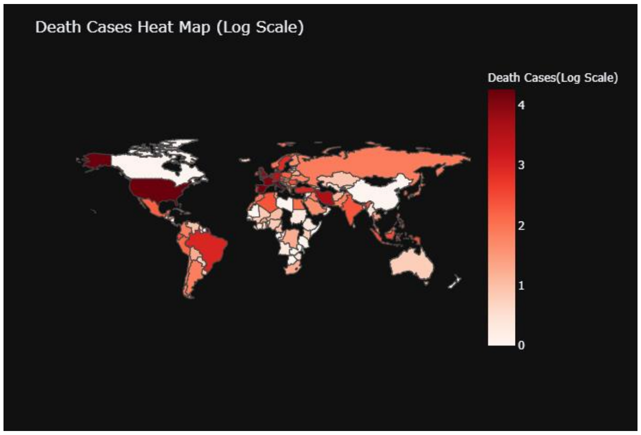


Fig. 5 Dearth cases heat world map

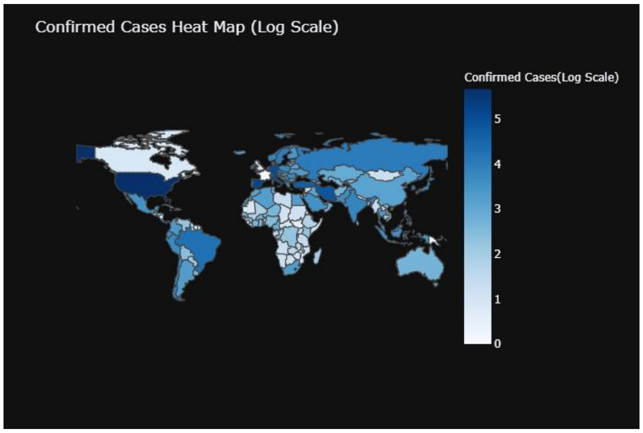

Confirmed Cases Heat Map (Log Scale)
Confirmed Cases(Log Scale)
5
4
3
2
1
0

Fig. 6 Confirmed cases heat world map

All of these eventually help us to get prepared for those countries or regions. Recovered cases are represented in green color as above, whereas death cases are represented in red color and the total number of cases are represented in blue color. Darker the color shades Red, the more affected area it represents.

We have used the Bubble chart on a 3D movable world map shown in Figure [7] and Figure [8]. This chart is zoomable and shows time-lapse. Coronavirus spread, it’s growth rate and effect comparison can be easily measured and visualized using time-lapse based bubble charts as visualized below.

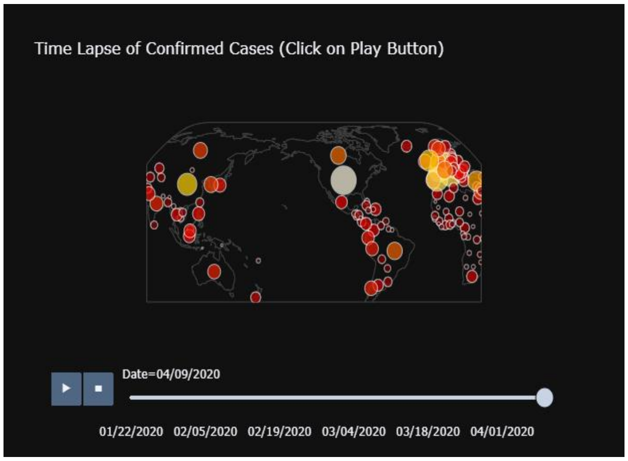


Fig. 7 Time-lapse of Confirmed cases Covid 19

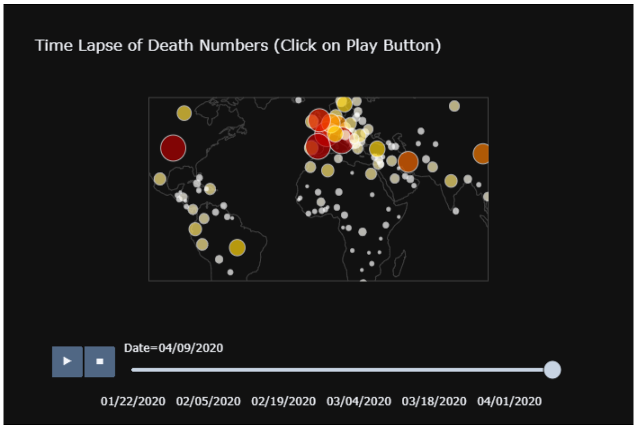


Fig. 8 Time-lapse of death cases Covid 19

We have used Plotly library for representation of this chart which provides downloadable report options also. We have also made a State wise heat map and we have developed a city wise heat map which is not available anywhere, shown in figure 9 and figure 10 respectively.

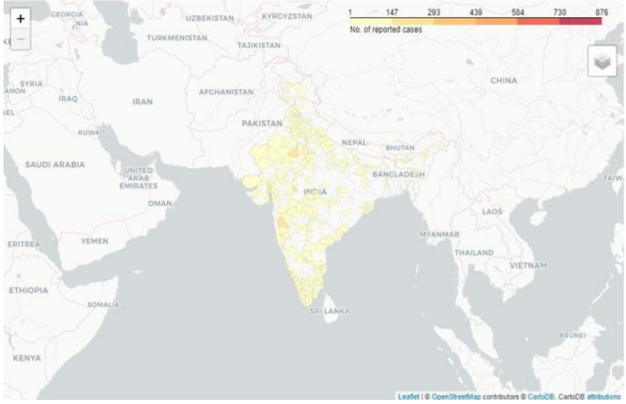


Fig. 9 Citywise Coronavirus heat map India

Upper right corner shows the meaning of the colors that we have in every state or city. This information helps us to make decisions on where lockdown shall be extended and how much relaxation shall be given in some particular area.

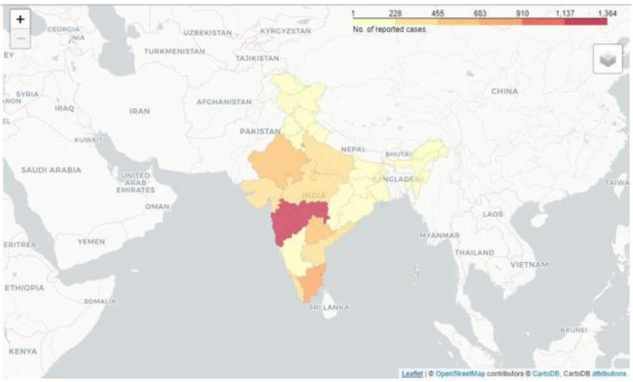


Fig. 10 State wise Coronavirus heat map India

This chart allows us to stop spread in those regions where already cases are lower by not enabling transportation between the respective source and destination regions and also can help to make decisions on starting up the transport between two green zone regions.

4. Distribution Plots

Time series plot shown in the above sections help us to understand the time domain behavior of a dataset.

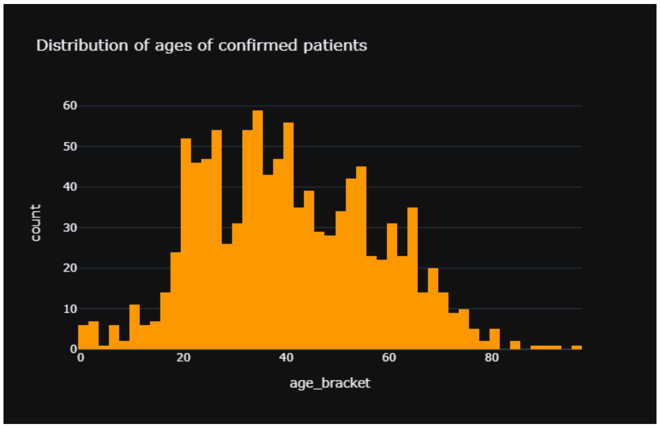


Fig. 11 Confirmed patient age distribution

Though it is a good representation, it does not give complete analysis of the data. To fetch a description from data like what is gender ratio or age bracket distribution, time series chart won't be useful

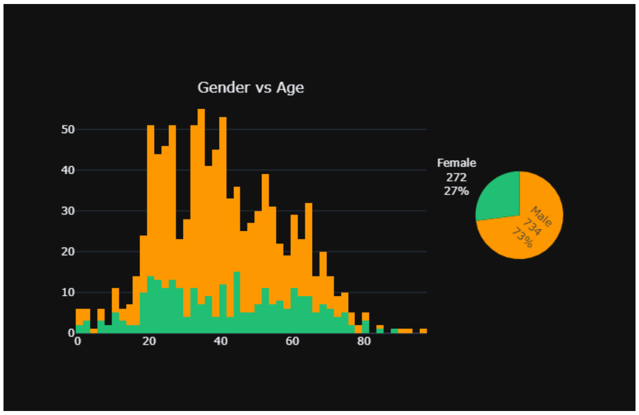


Fig. 12 Confirmed patient age and gender distribution

Tree map is the best suitable map for frequency representation with respect to multiple data as it is able to show immense amount of data with respect to variables. For our development, we have plotted a tree map of patient numbers with respect to the city and state.

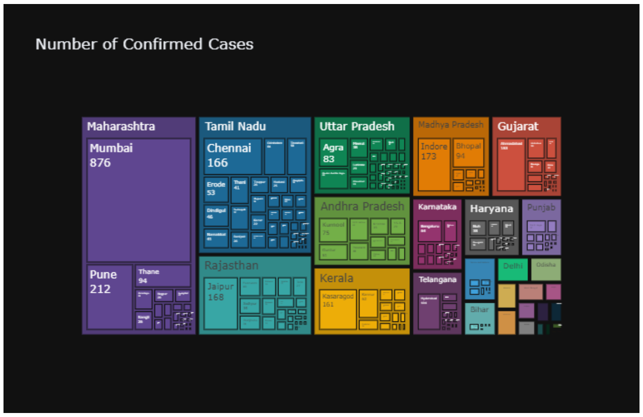
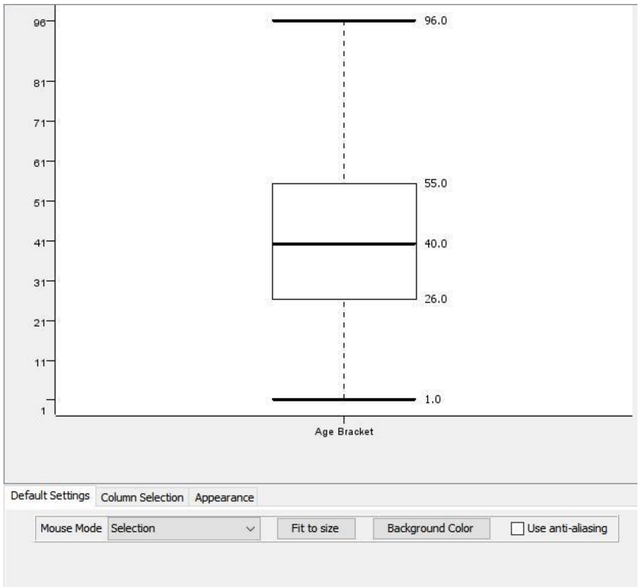


Fig. 13 Tree map of Coronavirus spread over India

5. Data Analytics

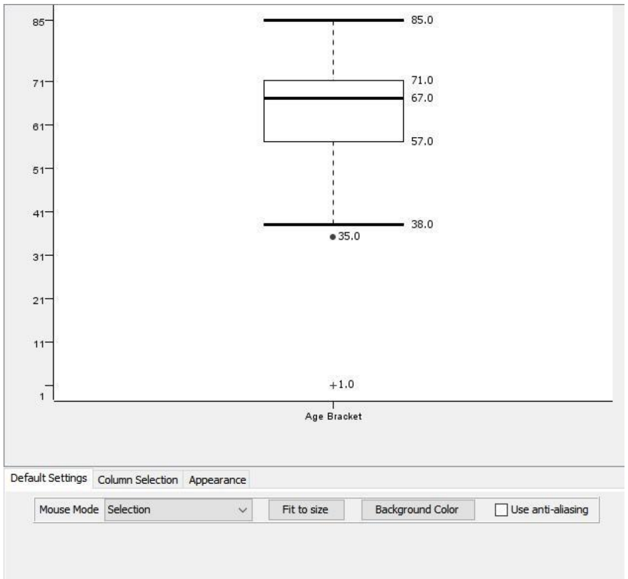
To get into more detail and to understand the important insights like mortality and recovered patient's features, we have used Knime workspace and have populated some charts. The data we have captured from open source available dataset [6] Covid 19 India. There were very few number of entries with complete information and hence we have rejected null entries and counted only those entries which have complete details. We have used box plot to represent the dataset visualization where we have considered all available details including outliers also. For age wise mortality and recovery rate, we have used histogram. We have plotted two box plots and two histograms whose details are given in subsequent sections below.



Mouse Mode: Selection

Fig. 14 Recovered patient box plot

Mean of the recovered patient from Coronavirus is 40 where interquartile ranges of the recovered patient from Coronavirus are, 26 lower quartiles and 55 upper quartile range. Lowest outlier is 1-year-old child and highest age of the recovered patient is 96 which is shown in Figure 14. So from the box plot we can assume that from the age spread of the virus, recovery chances of the individual are greater if the age of the patient is less than 55 years. This information is not sufficient enough to justify the complete behavior of the virus and hence we have to look for the other plots also. Therefore, we have plotted patients who were not able to survive in the below portion.



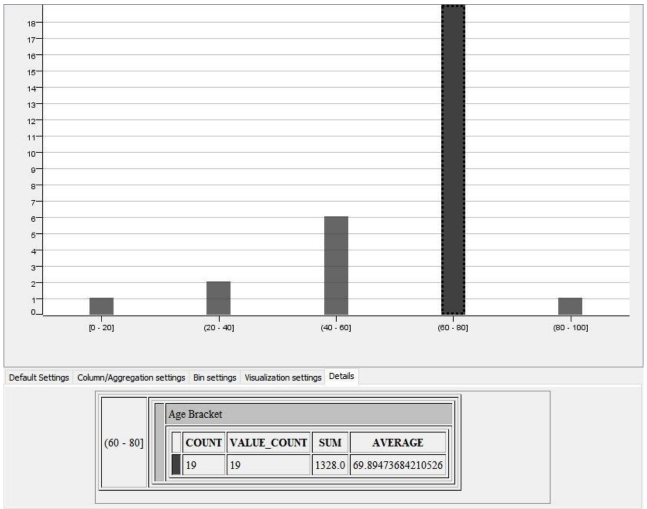
Mouse Mode: Selection

Fig. 15 Deceased patient box plot

Mean of the deceased patient from Coronavirus is 67 where interquartile ranges of the deceased patient from Coronavirus are, 57 lower quartile ranges and 71 upper quartile range.

Lowest outliers death reported is 1 year old child and highest age of the recovered patient is 85 which is shown in Figure 15. So from the box plot now we can surely say that from the age spread of the virus, deceased chances of the individual are greater if the age of the patient is greater than 57 years.

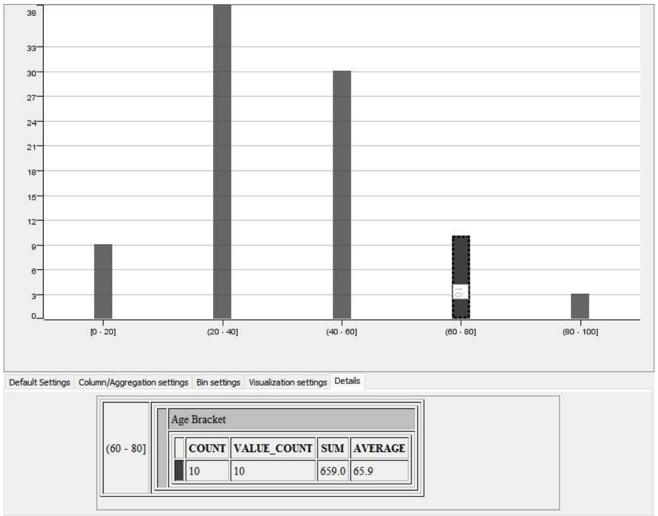
This information is sufficient enough to justify the complete behavior of the virus but in actual, the age wise ratio of recovery and death rate will provide more details.



(60 - 80]

Age Bracket				
	COUNT	VALUE_COUNT	SUM	AVERAGE
■	19	19	1328.0	69.89473684210526

Fig. 16 Histogram of deceased patients



(60 - 80]

Age Bracket				
	COUNT	VALUE_COUNT	SUM	AVERAGE
■	10	10	659.0	65.9

Fig. 17 Histogram of recovered cases

6. Conclusion

For representation over here we have taken 20 as a bracket and plotted the histograms. Histogram comparison of the recovered and deceased patient shown in figure 16 and 17 proves that "Elder patients have less chances of recovery if affected by Coronavirus" in other language we can say that "Coronavirus affected Elder patients have more risk of death" Hence we have related the age group with death cases. As in certain areas of the world cases have been significantly reduced future scope of the data visualization is to have predictions on when and where Coronavirus would stop and which countries should care more from now.

Acknowledgments

This paper and research behind it is only possible with open sourced dataset get available from covid19india, Worldometer team and Keggle. This research is supported by Optimized Solutions Limited. We are thankful to Python and it's open-source worldwide community and other repositories Sci-kit, Plotly, Numpy and Knime.

References

[1] Worldometers website charts and data points https://www.worldometers.info/coronavirus/#countries

[2] Khanam, Fahima & Nowrin, Itisha. Data Visualization and Analyzation of COVID-19. Journal of Scientific Research and Reports.26-3 (2020) pp. 42-52.

[3] Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Biorxiv; 2020.

[4] World Health Organization. Middle East respiratory syndrome coronavirus (MERS- CoV). Available:https://www.who.int/emergencies /mers-cov/en.

[5] Covid 19 Dataset references https://www.kaggle.com/covid19

[6] Indian Patient Coronavirus database https://www.covid19india.org/

[7] Yi W, Wang Y, Tang J, Xiong X, Zhang Y, Yan S. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2020;32(3):279286. doi:10.3760/cma.j.cn121430-20200225-00200

[8] Data science blog on spread of Covid-19 https://towardsdatascience.com/the-impact-of-covid-19-data-analysis-and-visualization-560e54262dc

Authors -

Abhishek Parikh received B.E. (Electronics and Communication) from GTU University Ahmedabad in 2014, M.E. (Microprocessor System and Application) from MSU University Baroda in 2016 and pursuing his Ph.D. in bio-medical signal processing from GTU. He has got 5 years of working experience as Product Development Lead in product engineering services at Optimized Solutions Limited.

Sandeep Shah received his B.E. (Instrumentation and Control) from GU in 2001 and PGDM from IIM Calcutta. He is currently working as Managing Director of Optimized Solutions Limited, Ahmedabad. His area of interest is Monitoring, Automation and Data acquisition systems.

Vishvam Bhatt received his B.C.A. from DDIT in 2018 and M. Sc. In Information and Technology from DAIICT Gandhinagar. His area of interest is Data Analytics and Big Data visualization. He is currently working as Data Scientist at Optimized Solutions Limited.