Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases

Joanita DSouza
Department of Computer Science and Engineering
Amity University Dubai
Dubai, UAE
joanitaD@amitydubai.ae

Senthil Velan S
Department of Computer Science and Engineering
Amity University Dubai
Dubai, UAE
svelan@amityuniversity.ae

***Abstract*—Exploratory Data Analysis (EDA) is a field of data analysis used to visually represent the knowledge embedded deep in the given data set. The technique is widely used to generate inferences from a given data set. Data set of current pandemic, the COVID-19 is widely made available by the standard dataset repository. EDA can be applied to these standard dataset to generate inferences. In this paper, data visualization technique is applied to the dataset and is used to formulate patterns for better insights on the effects of the pandemic with respect to the variables/ labels given in the dataset. A Web application tool called Jupyter Notebook is used to generate graphs using python language as it consists of libraries which are used for the process of EDA and the visualization is depicted for the attributes showing higher correlation. Based on the graphs obtained, we can draw conclusions from the current situation based on the data available, understand why a certain variable is increasing/decreasing with respect to another and what can be done to improve the drawbacks found.**

Keywords—Exploratory Data Analysis, Big Data Visualization, Python, Seaborn, Matplotlib

I. Introduction

Rich and high volume data is the modern fuel that possess inherent characteristics for driving today's intelligent decision making abilities of smart businesses and services. When comparing with the energy sector, unprocessed raw data is equivalent to the crude oil. The fuel that powers the internal combustion engines is the intelligent information that is processed from the raw data. Similar to the extraction of different products using fractional distillation of crude oil, extraction of intelligent information at different levels will improve the decisions of different levels across the business unit.

Exploratory data analysis (EDA) is a process by which the given data set is analyzed to interpolate useful information. The process commonly depicts the data in a visual form enabling betting understanding and to adept informed decision making of the business entities. Visualization of data is in accordance with us in identifying testing, tendency, and interdependence.

Human comprehension prepares 60,000 times sensitive to perceived visual data than text. Visible knowledge is currently measured at 90% of the instruction transmitted to the brain [1][2]. Today's organizations provide exposure to such an immense amount of information that the company produces from through inside and out of the doors. Visualizing awareness helps to develop a perception of it all. The scanning of various worksheets, tablets or papers is common and wearisome at best, while the inspection of charts and graphs is always simpler enough for the eyes [3].

II. Data Visualization

Visualization of data involves the presence of data of any character in a graphical pattern that addresses the uncertainty and representation that can be handled. It is part of implementing more contemporary visualization procedures to display the relations between the data. Such instances continuously curve from the use of hundreds of lines, specifications, and link to reach a larger, aesthetically detectable information production. But various heterogeneous representations including heat maps and fever charts go far behind traditional corporate graphs, histograms and pie charts, enabling decision-makers to analyze data sets to identify correspondence or accidental trims. Generally when companies demand relations between data, graphs, bars, and charts are applied to do so. They also get a range of colors, words, and figures to help. Data visualization uses more immersive, graphical drawings to represent symbols and create relationships between bits of information, including personalization and animation [2] [4].

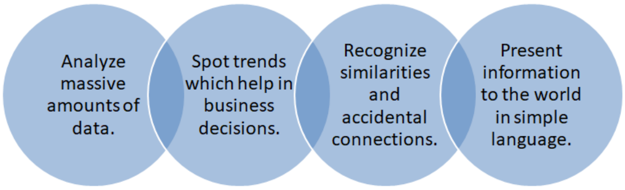

Analyze massive amounts of data.
Spot trends which help in business decisions.
Recognize similarities and accidental connections.
Present information to the world in simple language.

Fig. 1. Benefits of Big Data Visualizations

Visualization of data relies on convincing computer operations to absorb raw corporate data and prepare it to generate graphical diagrams that allow people to capture and grant huge quantities of data in seconds [5].

From Fig. 1, the benefits of data visualization are analyzing the huge and massive amounts of data which otherwise may be meaningless or difficult to comprehend. Big data visualization may also help organizations and businesses that produce an alarming rate of data by spotting trends and eventually helping in decision making. It can also identify the similarities and connections within the data and present the information to others in a simpler language [2][5].

I. Related Work

Availability to aggregated public health data tends to help in the analysis procedure of endemic, epidemic, and pandemics, and provides a means to visualize their behavior patterns and demographic influence [6]. Much work on related topics has already been conducted to provide rich guidance on the basic concept of fusing computer science with medical science. Findings from different polls demonstrated the health-care effect of big data [7]. Public health information, however, is complicated, and multiple. It really isn't useful to attain valuable information from such a massive series of data structured in large tables, or to recognize patterns. Many techniques of data analysis and visualization are mandatory in this context; otherwise a huge amount of data is largely useless.

In the paper 'Heart Failure Risk Prediction and Medicine Recommendation using Exploratory Data Analysis', the author performed exploratory data analysis on an Austrian medical dataset. The main aim was to help doctors come to a conclusion about the patient with respect to heart conditions. EDA mainly helped analyze and form a general and broad idea about the dataset and the possibilities that occur based on certain circumstances. They performed EDA and data processing after which they generated the most important attributes and rows which helped to attain the best candidates to form a predictive system [6].

From the paper 'Vis-Health: Exploratory Analysis and Visualization of Dengue Cases in Brazil.', the author presented a tool called Vis-Health which is mainly used for analysis and visualizations of the health data of the public with the help of covariance matrix in order to choose the most suitable variables. The study for the covariance matrix is accomplished by the using Principal Component Analysis (PCA), which resulted in a linear orthogonal regression The main goal of the paper is to give the user a better and clear idea and understanding of the occurrence of the disease dengue based on which the government is able to take the necessary measures and decisions [7].

The "Flu Near You" website introduces statistical data comparable to HealthMap on the occurrence of the flu in the United States and Canada. Throughout this system, the user can envision the development of the new cases with intensity of labeled documents for each state through an animation with the passage of time per week [8].

Google created an interactive method called Google Trends that enables study into the occurrence of illnesses, such as flu and dengue, depending on the search words used by users in particular areas. The data is seen on a regional map of country delimitation. Flu incidence is categorized into 5 tiers, in which each stage is given a different color, when the device user clicks on one area, a new screen is displayed with relevant data for that region and graphics are chosen showing the average flu activity in each month in the country [9].

In this paper, I have performed Exploratory Data Analysis on the current undergoing COVID-19 situation. The main aim is to obtain a generalized vision of the current pandemic.

II. Applying EDA For COVID-19 Data

A. *Process Flow of EDA*

EDA is the initial step of deciphering data by first showing the visual representation using different tools available in a data processing tool. The generic process flow of the EDA is shown in Fig 2. Since the input data could be usually raw and unclean the data is first cleaned and trimmed using inbuilt functions with a tool. These tools can identify undefined and incomplete data which can be cleaned to obtain the refined and enriched data. This enriched data can then be used to analyze using visualization techniques. Visualization of the refined data can be used write inferences on the general trend of the data under study.

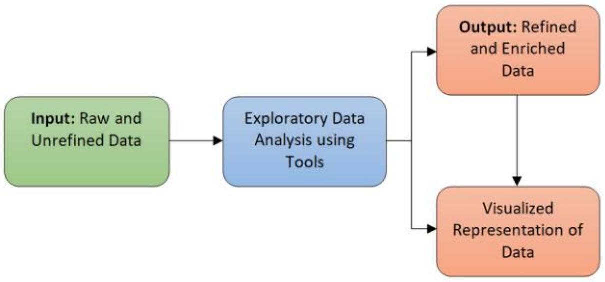

Input: Raw and Unrefined Data
Exploratory Data Analysis using Tools
Output: Refined and Enriched Data
Visualized Representation of Data

Fig. 2. A Generic Process Flow of EDA

B. *Application to COVID-19 Data*

In the current epoch, nearly every field from science and financial matters to designing and showcasing measures, accumulates, and stores information in some advanced structure. Data is being produced at a fast rate and is a trend which will be continuing over time. It is important to make sense of the large amounts of data produced in order to come to conclusions and decisions. By visualizing data, strategies and business models can be modified for it to be beneficial.

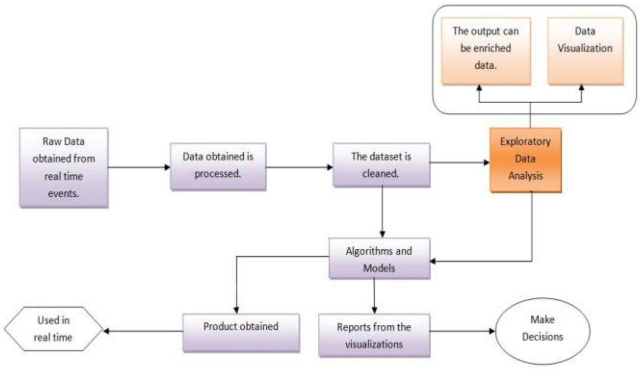

The output can be enriched data.
Data Visualization
Raw Data obtained from real time events.
Data obtained is processed.
The dataset is cleaned.
Exploratory Data Analysis
Algorithms and Models
Used in real time
Product obtained
Reports from the visualizations
Make Decisions

Fig. 3. Process Flow of applying EDA to COVID-19 Data

From Fig 3, the first step in the process of data analysis and science is obtaining the raw data from various surveys or polls, sources or companies. The data obtained should be real data [10]. The data obtained from reliable sources is then processed and prepared for further analysis. Data wrangling is then done on the datasets to obtain clean dataset which is suitable for analyzing.

Various algorithms and models can be used on the cleaned dataset in order to gain insights to make better decisions. Preventive Maintenance can also be used to predict the negative impacts which may be caused due to certain measure or decisions taken [11]. However, exploratory data analysis is an essential step before performing algorithms and models. Exploratory data analysis provides an output which can be an enriched form of data and also provide data visualizations. These outputs are implemented along with various algorithms and models to make decisions or obtain a product which when used in real –time will be advantageous.

III. Data Visualization features and classification

Confronted with the complicated high-dimensional data, people sometimes argue that we often don't understand the implications of the data. By incorporating the innovation of statistical data visualization to render complicated data into the form people will acknowledge [12]. Therefore, data visualization technology has the characteristics as the following:

Data visualization technology indeed has the following elements:

(1) *Graphical fidelity:* effectively organize the information.

(2) *Multidimensional:* get data from multiple aspects (characteristics, features) of the information, but in a one-dimensional manner in front of the people.

(3) *Visibility:* the data is eventually presented in the form of graphs, diagrams, patterns and graphics and its correlations and interdependencies are envisioned. Through the increasing evolution of technological advances, only single-scale information can be referenced, and vast-scale and higher-dimensional data and information are now being interpreted and expressed seamlessly [13].

Graphical systems display a large degree of usability, fast and easy usual features, influential visualization abilities and valuable techniques for handling and utilizing data resources. Data mining visualization methods and technologies as a hybrid term derive from the combination of data mining techniques and visualization methods as the creative achievement in the creation and exploitation of data resources.

Data mining may help workers find knowledge of interest in the data field more easily, or draw some new conclusions[4]. Data visualization technology can visualize and simplify complex data, and make understanding the underlying data laws simple for researchers.

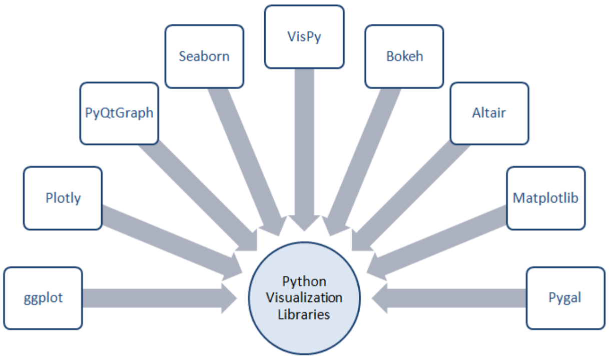


Fig. 4. Python libraries used for visualizations

Currently, visualization techniques based on the Python data analysis libraries NumPy, Pandas, Matplotlib, Seaborn, ggplot, Plotly, PyQtGraph, VisPy, Bokeh, Altair and Pygal (Fig 4) are widely used and can be used for various dataset analyses. The visualization system also has the advantages of being easy to incorporate and having good interactivity [14].

I. Methodology

In this section, we discuss how to obtain patterns using the Exploratory Data Analysis (EDA) for the current COVID-19 pandemic.

The first step is loading the dataset and comprehending the nature of the dataset [15]. The dataset used is with regards to the COVID-19 cases. Once the nature of the data is understood, a correlation matrix can be formed in order to understand the relationship between the variables using the correlation function. Knowledge about the relationship between these features is needed as a step to help further analyze the data [16]. This dataset will serve as a catalyst for people to generate additional outbreak data, not just numbers but also new interpretations, policy responses, and so forth [17].

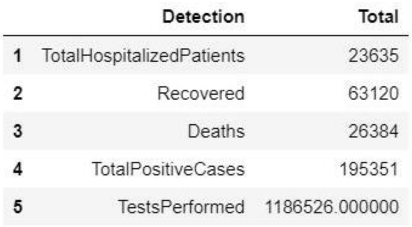
Python Pandas is perhaps the simplest and easiest base mapping process. Python Seaborn is perfect for color-including visually pleasing statistical charts. Bar charts are an important method for the representation of values in various categories. By adjusting the x- and y-axes, a bar chart may be structured horizontally or vertically, based on the kind of data or classifications the graph requires to express. Seaborn is a Matplotlib-based Python data visualization library, which offers a high-level framework to generate approachable and informative statistical graphics.

Scatterplots can also be used to specify if there is a relationship amongst two continuous parameters evaluated on the basis of the proportion or interval. The x-and y-axis plots the two variables. Each point that appears on the scatterplot is a single observation. The point position shall be determined by the value of the two variables. Every point appearing on the scatterplot is one single observation. The point position of the two variables shall be determined by their value. Scatterplots allow you to see the relationship type between possible two variables.

A stable or a positive association occurs when in the first variable higher values correlate with higher values in the second variable, and in the first variable lower values correlate with lower values in the second variable (points in the graph are upward movements from left to right). Negative relationships arise when higher values correlate with lower values in the second variable in the first variable, while lower values in the first variable correlate with higher values in the second variable (points pass from left to right down). Often essential is the essence of relationships-linearity or non-linearity. A linear relation exists when a second variable changes proportionally in response to changes in the first variable. A nonlinear relationship is drawn as a curve indicating the shift is not identical in the second variable, when the first variable shifts. It is clear that in recording an outbreak activity of this serious disease, processing such data in real-time is extremely useful. We assume that this type of data analysis can definitely improve situational awareness as well as notify strategies [17].

II. Inferences From the Dataset

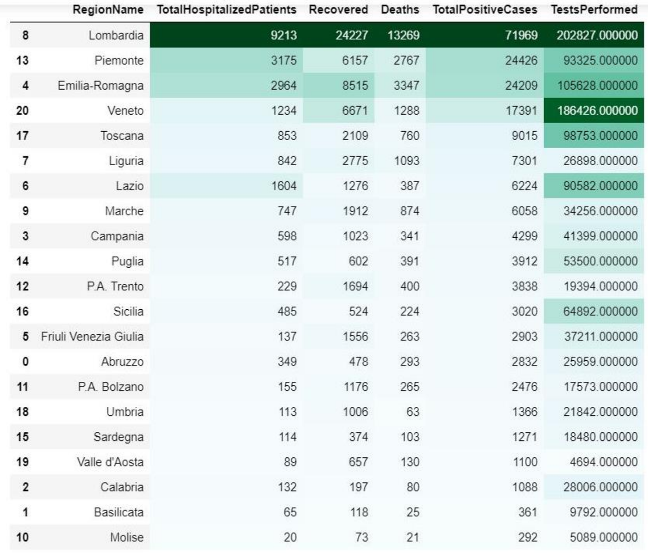
A. Introductory Infereences



	Detection	Total
1	TotalHospitalizedPatients	23635
2	Recovered	63120
3	Deaths	26384
4	TotalPositiveCases	195351
5	TestsPerformed	1186526.000000

Fig. 5. Total number of rows for a particular variable

From Fig 5, we can obtain a brief statistics about the total number of instances in each of the important columns. The total number of deaths is more than the total hospitalized patients but is comparatively less than the patients who have recovered.



	RegionName	TotalHospitalizedPatients	Recovered	Deaths	TotalPositiveCases	TestsPerformed
8	Lombardia	9213	24227	13269	71969	202827.000000
13	Piemonte	3175	6157	2767	24426	93325.000000
4	Emilia-Romagna	2964	8515	3347	24209	105628.000000
20	Veneto	1234	6671	1288	17391	186426.000000
17	Toscana	853	2109	760	9015	98753.000000
7	Liguria	842	2775	1093	7301	26898.000000
6	Lazio	1604	1276	387	6224	90582.000000
9	Marche	747	1912	874	6058	34256.000000
3	Campania	598	1023	341	4299	41399.000000
14	Puglia	517	602	391	3912	53500.000000
12	P.A. Trento	229	1694	400	3838	19394.000000
16	Sicilia	485	524	224	3020	64892.000000
5	Friuli Venezia Giulia	137	1556	263	2903	37211.000000
0	Abruzzo	349	478	293	2832	25959.000000
11	P.A. Bolzano	155	1176	265	2476	17573.000000
18	Umbria	113	1006	63	1366	21842.000000
15	Sardegna	114	374	103	1271	18480.000000
19	Valle d'Aosta	89	657	130	1100	4694.000000
2	Calabria	132	197	80	1088	28006.000000
1	Basilicata	65	118	25	361	9792.000000
10	Molise	20	73	21	292	5089.000000

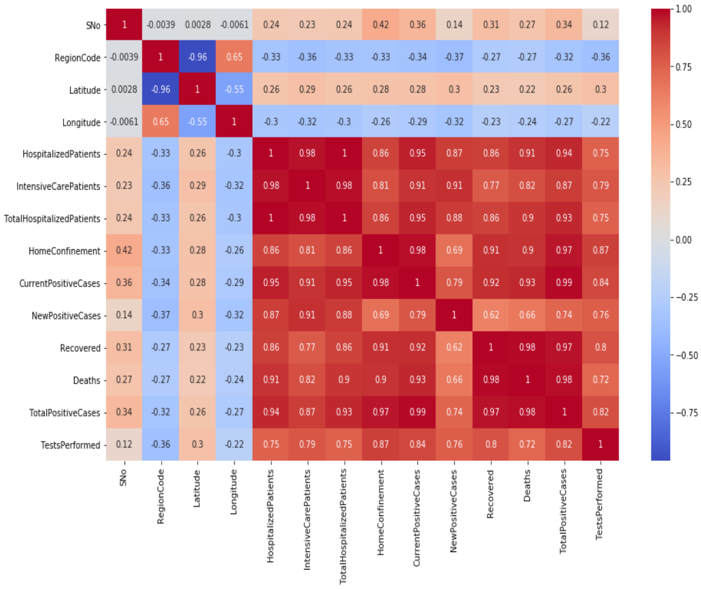
Fig .6. Display based on highest to lowest total positive cases region wise.

From Fig 6, the table has been obtained by selecting the region name and arranging it in descending order with respect to total positive cases.

B. Correlation Matrix

From Fig 7 it can be seen that the color correlation coefficient of the features that are dark red are more positive than the lighter shades of red. Whereas blue and the shades of blue indicate a negative relationship which means that there is no correlation between the variables. The shades ranging from negative to positive depicts the variation of correlation coefficients present in the graph.

The correlation coefficients provide the association between the variables in the dataset. The above correlation depicts the interdependency between the variables. The highest correlation between variables is 1 and the lowest is -1. By observing the correlation output, selection of variables which are to be placed in the X and Y axis is based on the dependency between them. For a better graph, selecting the correlation coefficients which are closed to 1 is more advisable than the coefficient being closer to 0.



	SNo	RegionCode	Latitude	Longitude	HospitalizedPatients	IntensiveCarePatients	TotalHospitalizedPatients	HomeConfinement	CurrentPositiveCases	NewPositiveCases	Recovered	Deaths	TotalPositiveCases	TestsPerformed
SNo	1	-0.0039	0.0028	-0.0061	0.24	0.23	0.24	0.42	0.36	0.14	0.31	0.27	0.34	0.12
RegionCode	-0.0039	1	-0.96	0.65	-0.33	-0.36	-0.33	-0.33	-0.34	-0.37	-0.27	-0.27	-0.32	-0.36
Latitude	0.0028	-0.96	1	-0.55	0.26	0.29	0.26	0.28	0.28	0.3	0.23	0.22	0.26	0.3
Longitude	-0.0061	0.65	-0.55	1	-0.3	-0.32	-0.3	-0.26	-0.29	-0.32	-0.23	-0.24	-0.27	-0.22
HospitalizedPatients	0.24	-0.33	0.26	-0.3	1	0.98	1	0.86	0.95	0.87	0.86	0.91	0.94	0.75
IntensiveCarePatients	0.23	-0.36	0.29	-0.32	0.98	1	0.98	0.81	0.91	0.91	0.77	0.82	0.87	0.79
TotalHospitalizedPatients	0.24	-0.33	0.26	-0.3	1	0.98	1	0.86	0.95	0.88	0.86	0.9	0.93	0.75
HomeConfinement	0.42	-0.33	0.28	-0.26	0.86	0.81	0.86	1	0.98	0.69	0.91	0.9	0.97	0.87
CurrentPositiveCases	0.36	-0.34	0.28	-0.29	0.95	0.91	0.95	0.98	1	0.79	0.92	0.93	0.99	0.84
NewPositiveCases	0.14	-0.37	0.3	-0.32	0.87	0.91	0.88	0.69	0.79	1	0.62	0.66	0.74	0.76
Recovered	0.31	-0.27	0.23	-0.23	0.86	0.77	0.86	0.91	0.92	0.62	1	0.98	0.97	0.8
Deaths	0.27	-0.27	0.22	-0.24	0.91	0.82	0.9	0.9	0.93	0.66	0.98	1	0.98	0.72
TotalPositiveCases	0.34	-0.32	0.26	-0.27	0.94	0.87	0.93	0.97	0.99	0.74	0.97	0.98	1	0.82
TestsPerformed	0.12	-0.36	0.3	-0.22	0.75	0.79	0.75	0.87	0.84	0.76	0.8	0.72	0.82	1

Fig. 7. Correlation between the variables in the dataset.

From the above matrix, a sample of the strong and weak correlations are shown in the below table. Note: The variables which are the same have the highest correlation of 1.

TABLE I. Correlation of Variables

S. No.	Variable on the X-axis	Variable on the Y-axis	Correlation Analysis
1	Total Positive Cases	Home Confinement	Strong Correlation of coefficient of 0.97
2	Hospitalized Patients	Total Positive Cases	Strong Correlation of coefficient of 0.94
3	Total Positive Cases	Deaths	Strong Correlation of coefficient of 0.98
4	Hospitalized Patients	Tests Performed	Comparatively weaker correlation with coefficient of 0.75
5	Tests Performed	Deaths	Comparatively weaker correlation with coefficient of 0.72
6	New Positive Cases	Recovered	Comparatively weaker correlation with coefficient of 0.62

From Table I, it can be seen that the first 3 serial numbers show strong correlations which mean that there is a high dependency of variable x on variable y. Similarly the last 3 serial numbers show a weak correlation compared to the high available correlations in the correlation matrix.

For instance, the variables ‘Total Positive Cases’ and ‘Deaths’ have a correlation coefficient of 0.98. It can be inferred that there is a strong association between these two variables. In general, when the number of COVID-19 positive cases increase, there is a possibility of the number of deaths also increasing due to various factors such as age, economic class, access to medical help etc.

However, while considering an example of comparatively weaker correlation between variables ‘New Positive Cases’ and ‘Recovered’. It can be inferred that the dependency between the variables is weak with a correlation coefficient of 0.62. This may be because the number of new positive COVID-19 cases does not play any role in the recovery of patients who are tested positive.

C. Tests performed in each Region

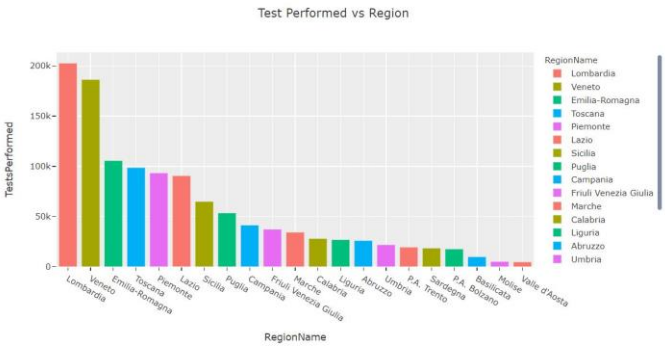


Fig. 8. . Test performed vs Region

Fig 8 depicts the states which have the highest number of tests performed. It can be inferred that Lombardia in Italy has performed the highest number of tests where as Valle d'Aosta has performed the least.

D. Tests performed with respect to Date

In the current COVID-19 crisis, it is advantages if the tests performed on a daily basis keep increasing. This can help treat the patients tested positive and curb the spread of the virus. For analyzing the relationship, seaborn library using line plot of python is used. Hence, from the below Fig 9, the numbers of tests performed are increasing per day.

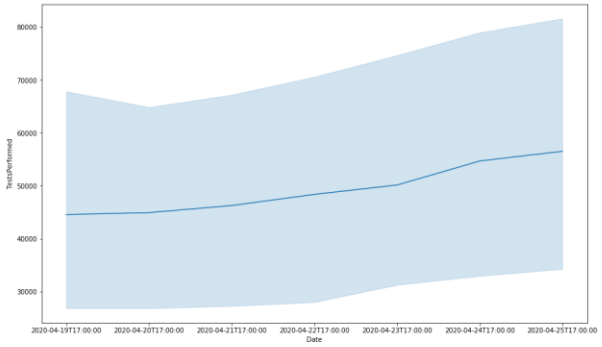


Fig. 9. Represenation of the tests performed to check patients for COVID-19 with respect to dates.

E. Hospitalized patients, Recovery, Death vs Count

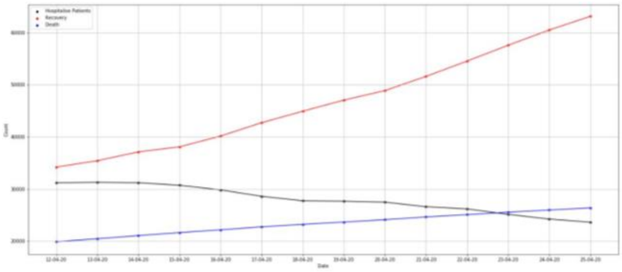


Fig. 10. Lineplot on the number of hospitalized, recovered and death patients with respect to date.

Taking into account three important variables which are differentiated based on colors in the garph. As seen in Fig 10, black is for hospitalized patients, red is for recovered patients and blue is for death rates. It can be seen that the range of observation is between 12-04-20 to 25-04-20. Considering the 14 days period, the number of recoveries kept increasing each day as well as the number of deaths. However, on 25-04-20, the number of recovered patients are approximately 6,200 being the highest and the number of deaths were approximately 2,800 being in the middle and the number of hospitalized patients were approximately2,400 being the least. An intersection between the number of hospitalized patients and death rate can be seen between 22-04-20 and 23-04-20.

F. Test and Confirmed cases with respect toRegion

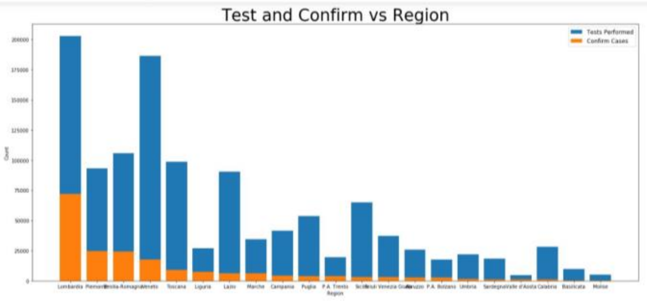


Fig. 11. Test and confirmed cases vs Region

In order to generate the graph in Fig 11, differentiating factors for two variables are used with respect to region name. Blue is used for tested cases and blue is used for confirmed cases. It can be observed that in all the regions the number of tests performed is more than the number of cases confirmed.

G. Total Hospitalized patients with respect to Total Positive Cases

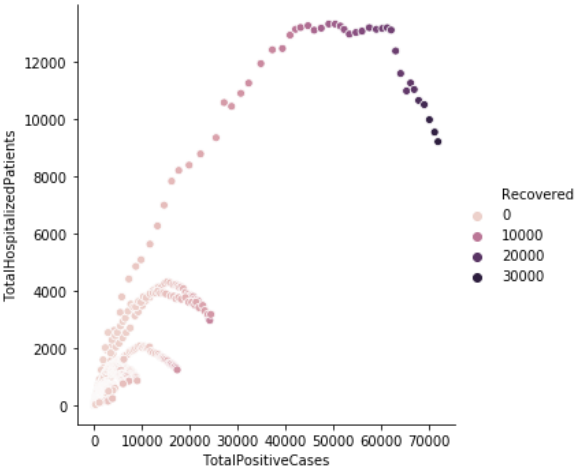


Fig. 12. Relationship between the total positive COVID-19 cases and the Total Hospitalized patients with a hue of the vairable Recovered.

The symptoms of COVID-19 once contracted can be less, moderate and severe. In the case of fewer symptoms, self healing is possible without the need of being hospitalized. People who have moderate to severe symptoms need to be hospitalized and under constant observation.

The variables 'Total Hospitalized Patients' and 'Total Positive Cases' have a correlation coefficient of 0.93 thus depicting strong association. From Fig 12, it can be seen that at certain sections the number of hospitalized patients decreases although there is an increase in the total positive cases. This can be due to the lesser symptoms and self healing of people who have contracted the virus, omitting the need for being admitted in the hospital. The graph then gradually increases then becomes relatively stable with slight variations and then drops. With respect to recovery, the number of patients recover initially are less and then increases and the most recoveries are seen when the drop occurs in the graph. Thus the number of hospitalized patients have dropped and recovered.

H. New positive case/ Hospitalized Patients/ Total Positive Cases and the number of ICU patients

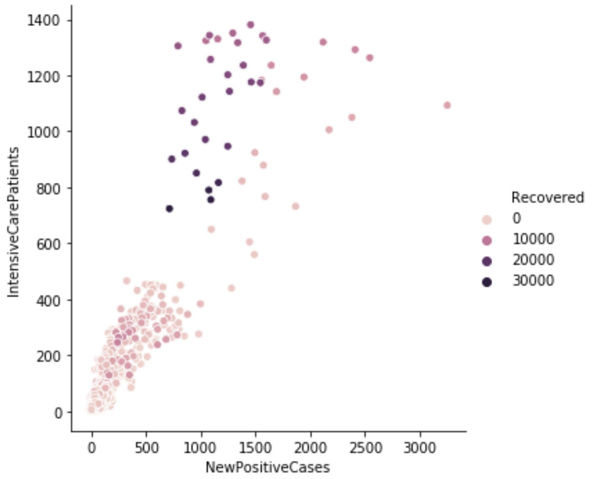


Fig. 13. Relationship between the new positive COVID-19 cases and the number of ICU patients woth a hue or recovered.

The variables ‘New Positive Cases’ and ‘Intensive Care Patients’ have a correlation coefficient of 0.91. The hue used is for obtaining the recovered patients from the new cases who are in the Intensive Care Unit. The ICU mainly consists of patients who are severely infected by the virus. Therefore, from Fig 13, it can be inferred that the all the new positive cases are not severely infected by the virus as the plots are not continues. However, initially the new positive cases were in taken to the ICU. With respect to the recovery, there is a distribution among various points in the graph. Initially very few of the new patients recovered but as the number of new patients increased the recovery rate also increased. This may be due to the increase in time for recovering.

III. Conclusion

This paper has presented an exploratory data analysis (EDA) process to formulate patterns related to the current COIVD-19 pandemic. The purpose of this study is to analyze the dataset and obtain important insights form it. As visual representations are appeasing and easy to understand, the results or outputs produced in the form of graphs can help people comprehend the current situation insights easily. The dataset used may not be an updated version; hence the inferences may vary from time to time as graphs can be generated as the data increases. As the amount of data increases, the trends may change and lead to different inferences and solutions.

References

[1] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. Patel, R. Ramakrishnan and C. Shahabi, "Big data and its technical challenges", Communications of the ACM, vol. 57, no. 7, pp. 86-94, 2014.

[2] M. Mani and S. Fei, "Effective Big Data Visualization", Proceedings of the 21st International Database Engineering & Applications Symposium on - IDEAS 2017, 2017.

[3] W. Yafooz, S. Abidin, N. Omar and S. Hilles, "Interactive Big Data Visualization Model Based on Hot Issues (Online News Articles)", Communications in Computer and Information Science, pp. 89-99, 2016.

[4] D. Keim, H. Qu and K. Ma, "Big-Data Visualization", IEEE Computer Graphics and Applications, vol. 33, no. 4, pp. 20-21, 2013.

[5] S. K. A. Fahad and A. E. Yahya, "Big Data Visualization: Allotting by R and Python with GUI Tools," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, pp. 1-8, 2018.

[6] N. Kaieski, L. P. L. d. Oliveira and M. B. Villamil, "Vis-Health: Exploratory Analysis and Visualization of Dengue Cases in Brazil," 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, pp. 3063-3072, 2016.

[7] Patel, J. A. and Sharma, P., Big data for better health planning., International Conference on Advances in Engineering and Technology Research (ICAETR), pages 1– 5, 2014.

[8] Freifeld, C.; Brownstein, J. About HealthMap. [Online] Available: http://www.healthmap.org/about

[9] Google. Google Tendências da Gripe. [Online] Available: http://www.google.org/flutrends

[10] Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, Second Edition. Glenn J. Myatt and Wayne P. Johnson.John Wiley & Sons, Inc. Published by John Wiley & Sons, Inc, 2014

[11] J. Dsouza and S. Velan, "Preventive Maintenance for Fault Detection in Transfer Nodes using Machine Learning," *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, pp. 401-404, 2019

[12] Fiorini P, Inselberg. A Configuration Space Representation in Parallel Coordinates [C]. International Conference on-Robotics and Automation, CA, USA: Jet Propulsion Lab, 1989.

[13] Y. Li and S. Hou, "Methods and Techniques in Data Visualization Model," 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, pp. 71-74, 2017.

[14] Xiong Zhongyang, Chen Ruotian, Zhang Yufang. An effective Kmeans clustering Heart initialization method [J]. Computer Application Research, 28 (11):4188-4190.1963, pp. 271-350, 2011.

[15] https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni

[16] T. Purwoningsih, H. B. Santoso and Z. A. Hasibuan, "Online Learners' Behaviors Detection Using Exploratory Data Analysis and Machine Learning Approach," 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, pp. 1-8, 2019.

[17] Dey, Samrat K., Md Mahbubur Rahman, Umme R. Siddiqi, and Arpita Howlader. "Analyzing the epidemiological outbreak of COVID‐19: A visual exploratory data analysis approach." Journal of Medical Virology 92, no. 6: 632-638, 2020.