COVID-19 Analysis and Visualization

ABSTRACT

In the current era of big data, a huge amount of data has been generated and collected from a wide variety of rich data sources. Embedded in these big data are useful information and valuable knowledge. An example is healthcare and epidemiological data such as data related to patients who suffered from epidemic diseases like the coronavirus disease 2019 (COVID-19). Knowledge discovered from these epidemiological data helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. As "a picture is worth a thousand words", having methods to visualize and visually analyze these big data makes it easily to comprehend the data and the discovered knowledge. In this paper, we present a big data visualization and visual analytics tool for visualizing and analyzing COVID-19 epidemiological data. The tool helps users to get a better understanding of information about the confirmed cases of COVID-19. Although this tool is designed for visualization and visual analytics of epidemiological data, it is applicable to visualization and visual analytics of big data from many other real-life applications and services.

***Keywords* – COVID-19, Data Visualization, Analysis, Pandemics, Decision-making**

I. INTRODUCTION

As we are living an era of big data, big data are everywhere. To elaborate, with advances in technology, a huge amount of data has been easily generated and collected from a wide variety of rich data sources at a rapid rate. These big data can be of different levels of veracity (e.g., precise data, imprecise and uncertain data [1]-[3]). Examples of big data include social network data [4]-[8], financial time series [9]-[11], omic data (e.g., genomic data) [1], [12], disease reports [13], [14], as well as epidemic data and statistics.

Embedded in these big data are useful information and valuable knowledge. This calls for data science, which aims to discover knowledge from these big data via data mining algorithms, machine learning tools, mathematical and statistical models, data analytics, and visual analytics. The discovered knowledge is useful. For instance, knowledge discovered from these epidemiological data helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, prevent, and/or control diseases such as viral diseases. Examples of viral diseases include:

- severe acute respiratory syndrome (SARS), with outbreak in 2002–2004;
- Middle East respiratory syndrome (MERS), with outbreak in 2012–2015; and
- coronavirus disease 2019 (COVID-19), with outbreak started in 2019 and became pandemic in 2020.

knowledge . A majority of the existing visualizers onthe COVID-19 epidemiological data focused on showing the numbers of confirmed cases and mortality spatially and/or temporally. In other words, they show

- spatial differences among different continents, countries, regions, or sovereignties; and/or
- temporal differences among weeks or days along the timeline—e.g., to show the effects of public health strategies and mitigation techniques such as social/physical distancing or stay-at-home orders in "flattening the (epidemic) curve".

As the numbers of inhabitants and tests both play roles in the data and their analyses, they help in the computation of figures like (a) the numbers of confirmed cases and mortality per thousand/million inhabitants and (b) the number of tests per thousand inhabitants.

However, in addition to the number of cases or mortality, there are other important knowledge that can be discovered from the epidemiological data via data mining. For instance, frequent pattern mining finds relationships among attributes (or features) associated with confirmed COVID-19 cases. Moreover, a visual representation of this discovered knowledge gives a more comprehensive representation, which in turn leads to a better insight and understanding of the data and discovered knowledge. Hence, in this paper, we present a tool for big data visualization and visual analytics of COVID-19 epidemiological data. Due to the nature of these data, it is not unusual to have NULL values for some of the attributes (e.g., unstated transmission methods of disease).

II. LITERATURE SURVEY

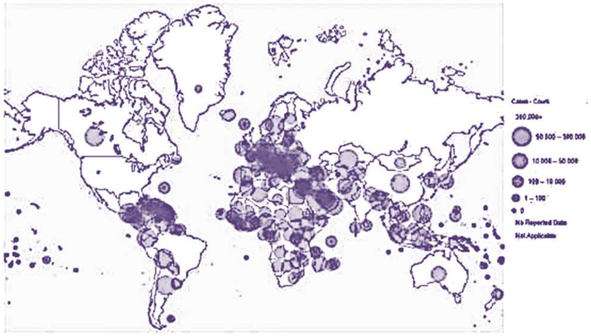
Due to the COVID-19 pandemic, many visualizers and dashboards have been developed over the past few months. Some of them - visualized literatures related to COVID-19 research, and some others visualized economic impact of COVID-19. However, a majority of them focused on the actual COVID-19 cases. Globally notable visualizers include (a) World Health Organization (WHO) Coronavirus Disease 2019 (COVID-19) Dashboard , (b) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)[1], and (c) COVID-19 dashboard by European Center for Disease Prevention and Control (ECDC)[2]. In addition, Government of Canada[3], provincial and territorial governments, major news channels/media/networks (e.g., TV[4], newspaper), and Wikipedia[5] capture data and provide dashboards on information about the COVID-19 pandemic situations in Canada. There are several commonality among these visualizers and dashboards. For instance, they mostly focused on the total numbers of new cases, confirmed cases, and deaths.

The *spatial information* about the total of numbers of confirmed cases and deaths in different countries (or regions/sovereignties) is usually represented by (a) the bubble map or (b) the choropleth map. To elaborate, in a *bubble map*, the total number of confirmed cases for each country is indicated by the radius of the bubble representing the country. See Fig. 1. Similarly, the bubble map can also show the number of new cases or deaths, in an absolute value (e.g., N_1 newly reported cases) or a relative figure with respect to population (e.g., N_2 deaths per one million population). See Fig. 2.

While the severity of COVID-19 in many countries can be representing by the sizes or radii of the bubbles representing these countries, many bubbles overlap. The overlapping and/or containment of bubbles makes it challenging to visualize the severity of the disease in countries in dense regions such as Eastern Caribbean and Southeastern Europe as shown in Fig. 2.

Fig. 1.

A snapshot of a bubble map showing the total number of confirmed cases among different countries in the world as of August 07, 2020.



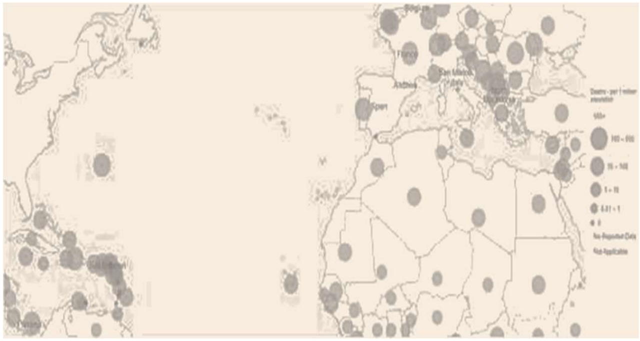


Fig. 2.

A zoom-in view of a bubble map —with overlapping bubles (especially in Eastern Caribbean and Southeastern Europe)—showing the total number of deaths per 1M population among different countries in the world as of August 07, 2020.

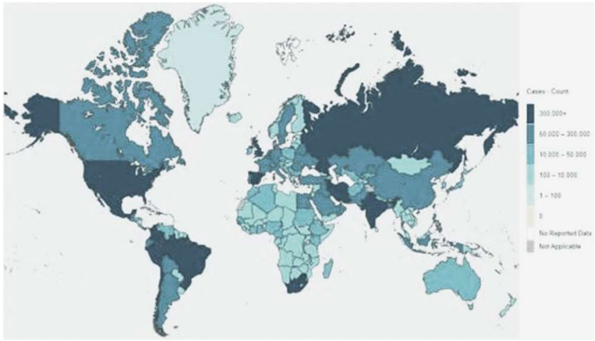


Fig. 3.

A snapshot of a choropleth map showing the total number of confirmed cases among different countries in the world as of August 07, 2020.

As an alternative, a *choropleth map* uses differences in shading, coloring, or the placing of symbols within predefined areas to indicate the average values of a property or quantity in those areas. For instance, Fig. 3 shows a choropleth map for the COVID-19 epidemiological data, in which differences in shading to represent the total number of confirmed cases. The darker the shading, the more severity is the COVID-19 situation, which means the higher is the number of confirmed cases. Note that the choropleth map in Fig. 3 shows the same information as the bubble map in Fig. 1 but with an alternative representation.

III. Our Visualizer for Analyzing COVID- 19 Epidemiological Data

To help users and researchers to get a better understanding of COVID-19 disease, we design and develop a big data visualization and visual analytics tools for COVID-19 epidemiological data. In this section, we illustrate our idea on COVID-19 epidemiological data for Canada .

A. Data Collection

Our tool first collects different categories of big COVID-19 epidemiological data from differentsources (e.g., federal and provincial/territorial governments). These include:

- administrative information, which includes an unique privacy-preserving identifier for each case, its region and episode week (i.e., symptom onset week or its closest week);
- case details, which include gender, age group, and occupation of the cases (e.g., health care workers, school or daycare workers, long-term care residents);
- symptom-related data, which include additional information for the case who is not asymptomatic (i.e., symptomatic case) such as onset week of symptoms, as well as a collection of symptoms (e.g., cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms);
- clinical course and outcomes, which include hospital status (e.g., hospitalized in the intensive care unit (ICU), non-ICU hospitalized, not hospitalized, unstated). For recovered case, it also includes additional information such as the recovery week. For the case who has not recovered, it indicates that the case died while infected by COVID-19; as well as
- exposures, which include transmission methods (e.g., domestic acquisition via contact of COVID-19 case or contact with traveller; international travel).

B. Data Preprocessing

After gathering related data from heterogeneous sources, our tool preprocesses the data. Given the nature of these cases, it is not unusual to have missing, unstated or unknown information (i.e., NULL values). For example, for some "Boolean" attributes (e.g., attribute "asymptomatic"), we observed three values: asymptomatic, symptomatic, and unstated (i.e., NULL value). Our tool links all values (including NULL) for each attribute. Moreover, our tool also detects and flags any data inconsistency.

C. Visualization of Frequent Patterns of Cardinality 1

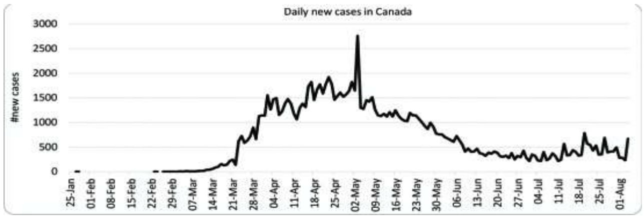
Once the data are preprocessed, we integrate the datamining capability into our tool. Specifically, via frequent pattern mining, our tool first discovers frequently occurring domain attributes (i.e., 1-itemsets). Our tool uses a *pie chart* or sunburst diagram to represent values for each frequently occurring domain attribute.

Recall that existing frequent pattern visualizers mentionedin Section II-B display all frequent patterns. In contrast, our tool displays frequent patterns one-by-one. For a frequent 1-itemset, we show not only the frequent ⟨*attribute, value*⟩-pair but also pairs associated with other values for that *attribute.* This enables users to get insight about theportion of that *value* for the *attribute.*

Example 1 As a preview, when applying our tool to the COVID-19 epidemiological data for Canada up to August 06, 2020, our tool finds a frequent 1-itemset ⟨ transmission, domestic acquisition⟩-pair with a frequency of 97,052 out of 107,916 cases (90%). Our tool also displays other pairsfor attribute "transmission" in the pie chart: ⟨transmission, international travel⟩ (4%) and ⟨transmission, NULL⟩(6%).

In addition, our tool gives users an option to ignore the NULL value for any attribute. If the user select this option, our tool displays not only the frequent ⟨ *attribute, value*⟩-pair but also pairs associated with other non-NULL values for that *attribute.* This enables users to get insight about the portions of all non-NULL values for the *attribute.*

Example 2 Continue with our preview, without NULL values, our tool finds a frequent 1-itemset ⟨ transmission, domestic acquisition⟩-pair with a frequency of 96% of cases with stated transmission methods. Our tool also finds and displays another pair with a non-NULL value for attribute "transmission": ⟨ transmission, international travel ⟩ (4%).



IV. RESULT AND DISCUSSION

Trends in COVID-19 Cases and Recoveries

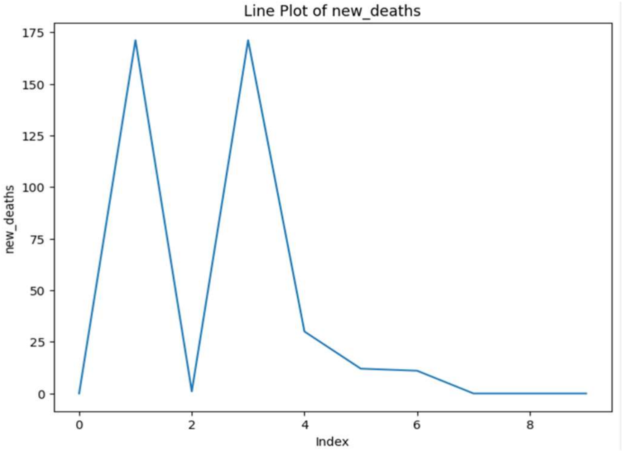
Using **line plots**, the dataset revealed clear trends in the progression of COVID-19:

a. A **sharp increase in cases** was observed during specific time periods, likely corresponding to waves or outbreaks.

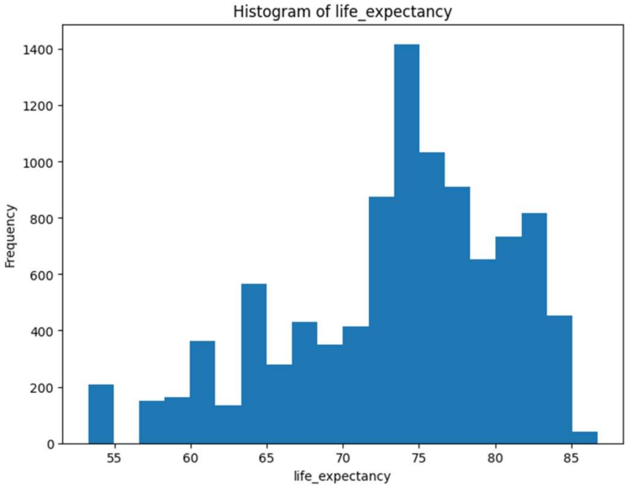
b. **Recovery trends** often lagged behind the spikes in cases, highlighting the time required for treatment and recovery.

c. **Lockdown periods** and vaccination campaigns were reflected in reduced case numbers during certain phases.

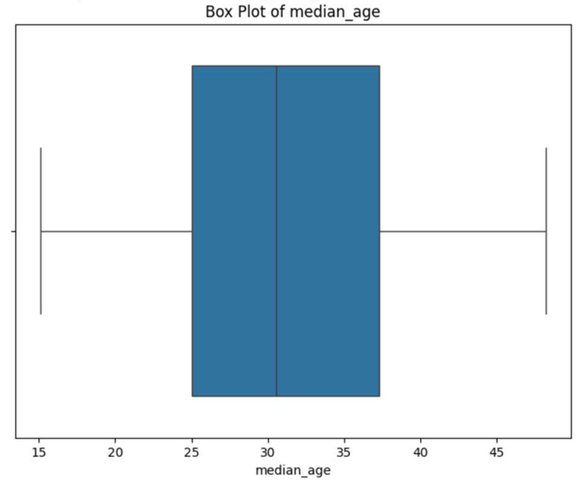
- **Line Plot**: Visualizing trends over time for cases, recoveries, and deaths.



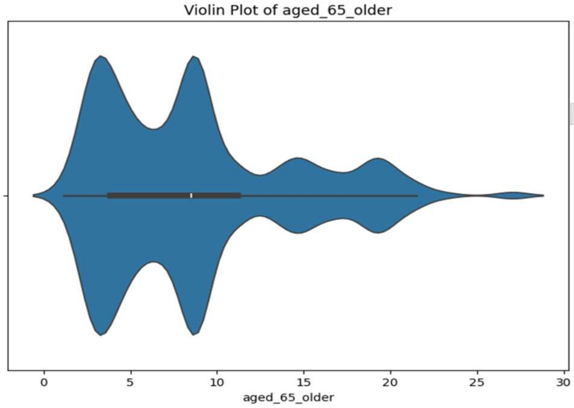
A **histogram plot** is a graphical representation of the distribution of a dataset. It groups continuous data into intervals, called **bins**, and shows how many data points fall into each bin.



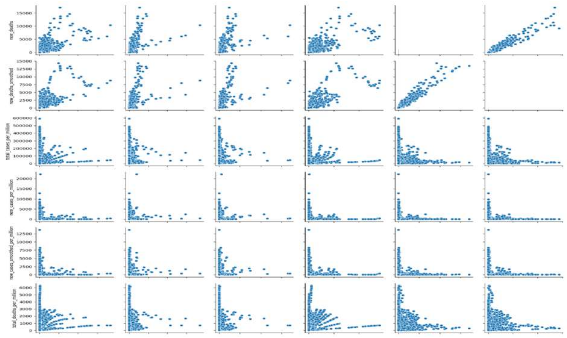
Box Plot: Analyzing the spread of data, e.g., recovery times across age groups.



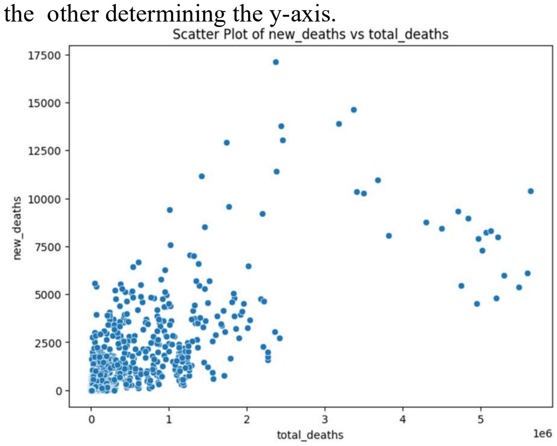
Violin Plot: Distribution of recovery status by demographics.



A **pair plot** is a visualization that shows pairwise relationships between multiple numerical features in a dataset. It plots a grid of scatterplots.



A **scatter plot** is a type of data visualization used to display the relationship between two numerical variables. Each data point is represented as a dot, plotted along two axes, with one variable determining the x-axis and the other determining the y-axis.



I. Conclusion

Over the past few months, there have been works on visualizing and analyzing different aspects of big data related to COVID-19. However, a majority of existing visualizers focus on showing the temporal and/or spatial trends on the numbers of cases and mortality. In this paper, we focus on epidemiological aspects of COVID- 19 data. Our key contributions include our design and development of a big data visualization and visual analytics tool for COVID-19 epidemiological data. By incorporating big data mining and data analytics into our tool, we discover frequent patterns, together with their related patterns. We provide users with flexibility to include or exclude NULL values. Evaluation results show benefits of our tool in providing a comprehensive and effective visual representation of these important patterns, which in turn helps researchers to get a better understanding of the COVID-19 disease and thusenables them to combat the disease. Moreover, our tool can be applicable to other real-life applications with NULL values. As ongoing and future work, we explore possibility to incorporate other techniques - intoour tool.

References

[1] Scribbr. (n.d.). How to cite a website | MLA, APA & Chicago examples. Retrieved December 14, 2024, from https://www.scribbr.com/citing-sources/cite-a-website/

[2] Indiana University of Pennsylvania. (n.d.). *How to cite online material from a computer network in APA*. Retrieved December 14, 2024, from https://www.iup.edu/writingcenter/writing-resources/research-and-documentation/apa-style/citing-online-with-apa.html

[3] UNSW Australia. (2013). *Harvard referencing - electronic sources*. Retrieved December 14, 2024, from https://www.student.unsw.edu.au/how-do-i-cite-electronic-sources

[4] LibGuides at St. Petersburg College. (2019). *APA citation guide: Internet resources examples*. Retrieved December 14, 2024, from https://sfcollege.libguides.com/apa/internet

[5] BibMe. (n.d.). *How to cite a website - MLA citation guide*. Retrieved December 14, 2024, from https://www.bibme.org/citation-guide/mla/website/

[6] Study.com. (n.d.). *Citing online sources | Importance, format & examples*. Retrieved December 14, 2024, from https://study.com/learn/lesson/citing-online-sources-types-examples-how-to-cite-online-sources.html

[7] American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC: Author.

[8] Purdue Online Writing Lab. (n.d.). *MLA formatting and style guide*. Retrieved December 14, 2024, from https://owl.purdue.edu/owl/research_and_citation/mla_style/mla_formatting_and_style_guide/general_format.html

[9] University of California, Berkeley. (n.d.). *Citing sources: APA style*. Retrieved December 14, 2024, from https://guides.lib.berkeley.edu/c.php?g=4472&p=30355

[10] Harvard University Writing Center.

(n.d.). *Citing sources: A quick guide*. Retrieved December 14, 2024, from https://writingcenter.fas.harvard.edu/pages/citing-sources

[11] University of Southern California Libraries. (n.d.). *Citing online sources in APA format*. Retrieved December 14, 2024, from https://libguides.usc.edu/c.php?g=235205&p=1551750

[12] The Writing Center at UNC-Chapel Hill. (n.d.). *How to cite sources*. Retrieved December 14, 2024, from https://writingcenter.unc.edu/tips-and-tools/citing-sources/

[13] The Chicago Manual of Style Online. (n.d.). *Citing electronic sources*. Retrieved December 14, 2024, from https://www.chicagomanualofstyle.org/tools_citationguide/citation-guide-1.html

[14] The Modern Language Association of America. (n.d.). *MLA handbook* (8th ed.). New York: Author.

[15] American Medical Association Manual of Style. (2016). *AMA manual of style* (10th ed.). New York: Oxford University Press.