ANALYZING COVID-19:A DATA-DRIVEN JOURNEY THROUGH THE PANDEMIC

PROFESSOR NAME : JIAN YANG

Narayan Raval

Student ID: 11614786

INFO 5709.003

Department of Information Science

The University Of North Texas

July 2023

Contents

ABSTRACT

The paper "Analyzing COVID-19: A Data-Driven Journey Through the Pandemic" offers a thorough analysis of the COVID-19 pandemic utilizing data-driven methodologies and visualization tools. This study intends to gather important insights from massive datasets about COVID-19 in the wake of the extraordinary global health catastrophe, revealing light on numerous facets of the pandemic's spread, effect, and response. Throughout the course of the study, we collect and examine a variety of datasets, including epidemiological information, hospitalization records, mobility patterns, and administrative data. We explore patterns, correlations, and trends within the data utilizing advanced data analytic and visualization using Python, allowing for a greater comprehension of the dynamics of the pandemic. In conclusion, this work highlights the significance of data-driven research in comprehending intricate global issues like the COVID-19 pandemic. We provide a thorough and educational picture of the pandemic's trajectory, lessons learned, and prospective routes toward a more resilient and prepared global community by fusing data analytic and visualization tools.

1 Introduction

The World Health Organization (WHO) declared the COVID-19 pandemic on March 11, 2020, following rapid transmission that began with the virus's debut in late 2019 signaled the start of a worldwide health disaster. When an infectious disease spreads widely over several locations at once, it is called a pandemic. The pandemic presented society and people with previously unheard-of difficulties as it spread over the world.

The analysis of COVID-19 data from the DS4C-PPP dataset is the main focus of this study, which offers important insights into how the outbreak affected the Korean population. The DS4C-PPP dataset, in contrast to other COVID-19 datasets, has the distinction of include patient-specific data, such as dates of symptom onset, dates of confirmation, and travel histories. Understanding the epidemiology and infection trends in South Korea, assessing the efficacy of infection control methods, and identifying essential variables influencing the success of containment efforts all depend on having this degree of detail in patient data.

The COVID-19 pandemic has had broad and significant repercussions, with various effects being felt by different people and communities. While some people adjusted to remote employment, online learning, and contactless services, others, such as those who maintain key services, experienced higher infection risks. The socio-economic impact and susceptibility to the virus are significantly influenced by social identities and group memberships.

With a relatively brief incubation period of 15 to 20 days, the disease had a severe mortality rate as the pandemic expanded. Even though the epidemic had an effect on every country, this study focuses on South Korea because it has access to the dataset thanks to the collaboration between the Korean government's Ministry of Science and Technology and MindsLab.

2 Related Work

Topic 1: COVID-19 Visualizers

This section details the different dashboards and visualizers that have been created to display COVID-19 data throughout the epidemic. These visualizers concentrate on various facets of the pandemic, such as COVID-19 cases, the financial impact, and associated research. Examples of this are the COVID-19 Dashboards created by the World Health Organization (WHO), the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, and the European Center for Disease Prevention and Control (ECDC). The visualizers depict spatial and temporal data about COVID-19 cases, deaths, and testing

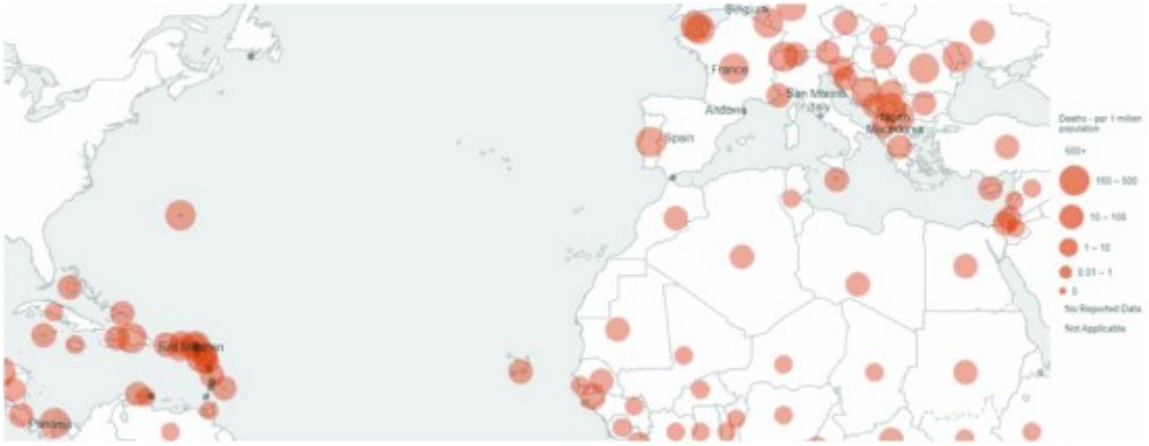

Figure 1: WHO coronavirus disease (COVID-19) dashboard

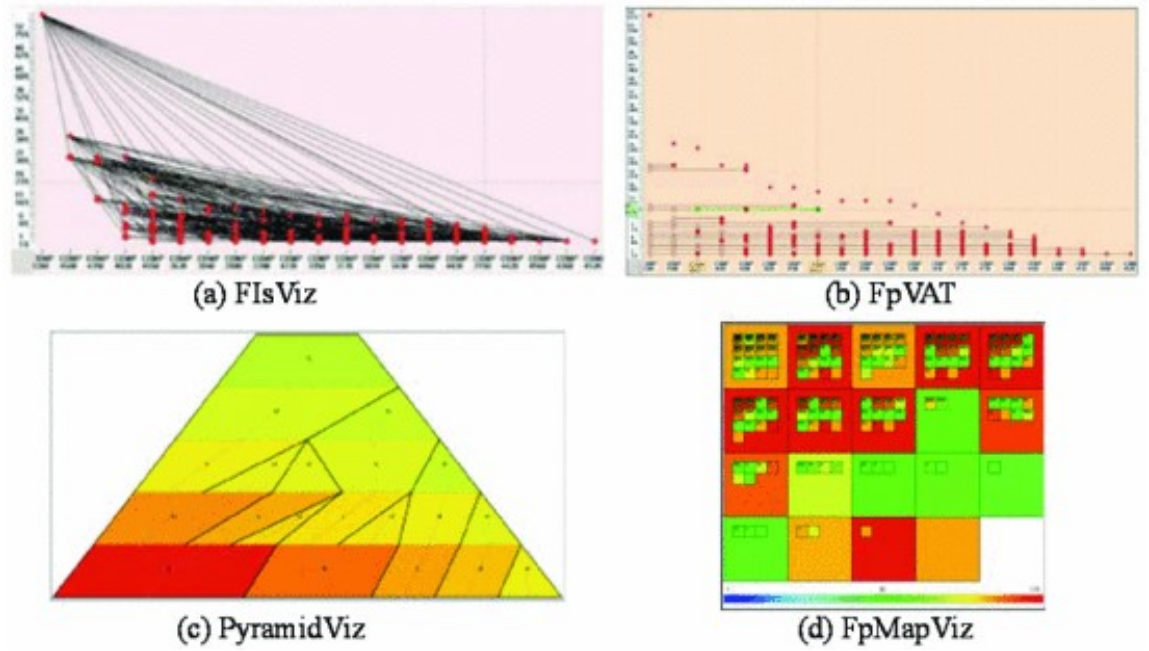

Figure 2: Frquent Pattern visualizer

rates using bubble maps, choropleth maps, and column charts. They struggle to accurately portray tiny nations and populated areas, nevertheless.

Topic 2: Frequent Pattern Visualizers

This section introduces data visualizers for data mining jobs, notably frequent pattern mining, in addition to COVID-19 data visualizers. In datasets, frequent pattern mining identifies pairings of traits and values that frequently occur. For the purpose of visualizing these common patterns, a number of visualizers have been created, including FIsViz, FpVAT, PyramidViz, and FpMapViz. They portray the identified patterns in an intelligible way

using a variety of layouts, including polylines, orthogonal graphs, pyramid layouts, and top-view layouts. It's important to keep in mind, though, that these visualizers weren't created expressly to display COVID-19 epidemiological data. Figure 2 is an example of the Frequent Pattern Visualizers.

3 Methods: Data Visualization using Google Colab and Python

1. Data Preparation:

The COVID-19 data used in this analysis was obtained from public repositories on Kaggle. The datasets were preprocessed to prepare them for visualization. This involved steps like renaming columns to match across datasets, removing null values, and correcting any incorrect country names to enable accurate plotting.

2. Python Libraries and Tools:

The visualizations were created in Python using Google Colab. The key libraries used were seaborn, pandas, and plotly for data manipulation and visualization. In particular, Plotly Express provided capabilities to create interactive bubble maps, choropleth maps, and other visuals. Custom functions were written to generate specific visualizations like treemaps showing total confirmed cases worldwide.

3. Types of Visualizations:

Four primary visualization types were developed to depict various facets of the COVID-19 data.

1. The total number of cases worldwide was displayed on an interactive tree map. This enables seeing case details by diving down from the continent to the country level.

2. Animations of choropleth maps were created to show how cases spread over time. The expanding shaded areas successfully show how the cases have changed geographically.

3. In order to examine potential connections between climatic variables and case growth, a scatter plot compared cases to weather.

4. To understand the severity of the cases by age group, a stacked bar chart was used to display the percentage of deaths across various age groups.

5. The percentage of particular coronavirus symptoms over time was plotted against the number of confirmed cases in the final graph. This indicates patterns in the prevalence of symptoms as cases increased.

The purpose of each style of visualization was to draw attention to specific aspects of the data, including aggregate totals, geographic distribution, potential correlations, breakdowns, and trends. Together, they offer a rounded viewpoint of the pandemic.

4. Interactive Visualizations:

1. On the tree map, you may click to zoom in on specific areas and scroll over boxes to get tooltips for case numbers.

2. The timeline sliders and play/pause buttons on the choropleth maps allow users to animate through the distribution of cases. A hover displays the case counts.

3. The tooltip hovers for the bar chart and line graph facilitate on-demand viewing of the details.

4. For concentrated analysis, all visualizations have pan and zoom options.

5. Graphic Excellence

1. Color schemes were chosen to be legible to color-blind viewers.

2. In order to make it simple to follow marks during animation and transitions, interactive highlighting was implemented.

3. The labels, legends, and markers on the charts were created to be easily readable at both high and low zoom levels.

4. In order to prevent information overload, information density was balanced while yet offering interactive detail when needed.

5. The adoption of a unified visual language, styles, and themes enabled the intuitive connection of various charts.

4 Results and Research Questions

Data visualization was used to examine four main COVID-19 trend hypotheses

https://colab.research.google.com/drive/1gGnq62YcLX8LNsZHK_6jKoyW8HlEvCDu?usp=sharing

Create a Treemap representation for a COVID-19 dashboard. Users can hover over boxes to see case numbers and click to zoom in on particular regions. Confirmed instances, nation names, and parent continents are displayed on the dashboard. Based on the overall number of confirmed instances, the dashboard's color changes, getting darker for more cases.

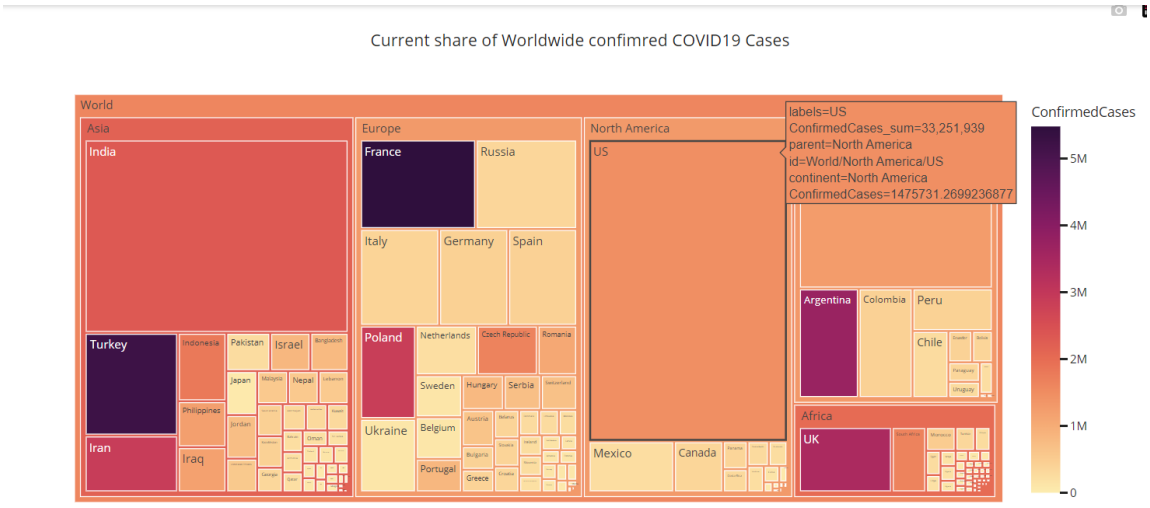

Figure 3: Dashboard

4.1 Research Question 1

What was the overall spread across the globe?

Different color schemes were used to distinguish land, water bodies, and COVID-19 case numbers in the animated choropleth map. To promote visual clarity and prevent confusion, each element was given a distinct hue. A particular color scheme was used to present land areas in order to emphasize geographic characteristics and borders. Oceans and other bodies of water were given their own color to help them stand out from the land masses. A sequential single-hue color scheme was used to represent COVID-19 case numbers in order to visually express severity ranging from low to high. This highlighted geographic regions with greater infection rates. The map provided visual consistency and made it easier to grasp the geographic spread of the pandemic by utilizing distinct colors for cases of COVID-19 on land, in water, and in other surfaces.

The animated choropleth map efficiently illustrates the geographical progression of COVID-19 cases, which begin in China and spread throughout the world. The visual representation of the temporal evolution answers the study issue about transmission patterns.

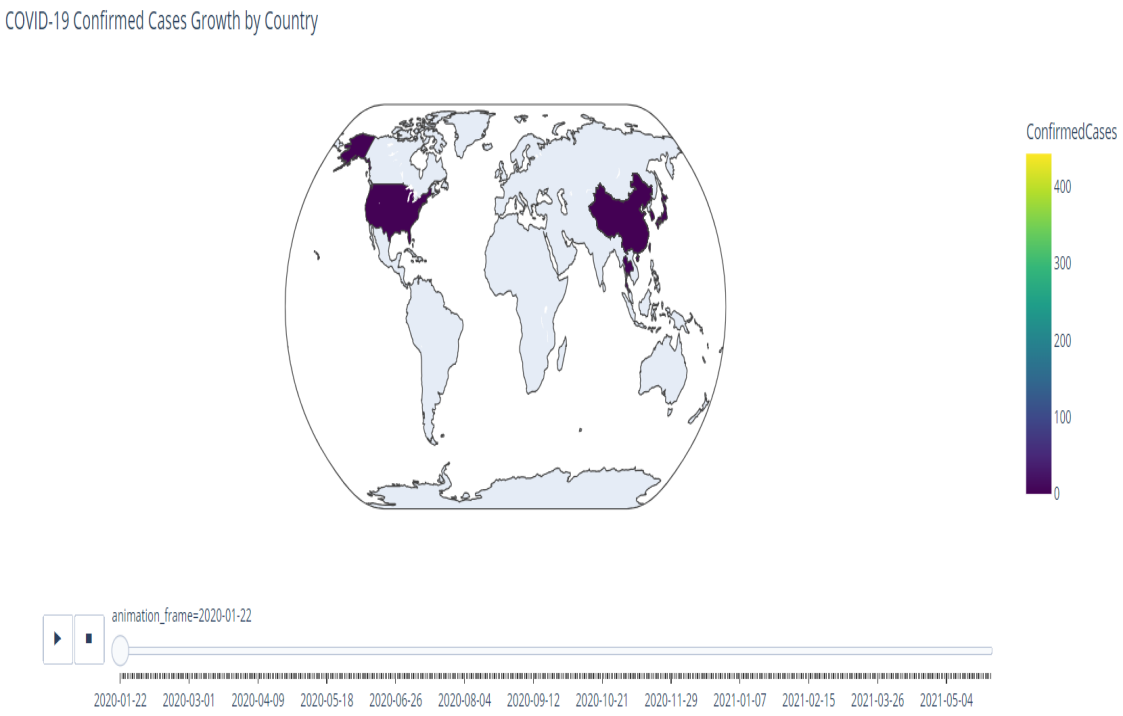

Figure 4: Covid19- Confirmed Cases Growth

4.2 Research Question 2

Did the weather have any impact on the number of confirmed Covid-19 cases?

The scatter plot shows that areas with lower wind speeds had higher case growth, which provides important insights into how weather variables affect COVID-19 transmission. Correlations between wind speeds and case growth are evident and clearly discernible due to the plot's ability to highlight particular subsets of data points.

The plot uses separate colors for various data subsets in accordance with design principles to help with visual separation and clarity. Assuring accurate depiction, the axis scales are correctly calibrated to meet the range of data values. Additionally, the plot displays the entire dataset in its original form, keeping the accuracy of the data.

A trend of increased case growth is seen in regions with lower wind speeds, according to the scatter plot comparing wind speed with COVID-19 cases. This raises the possibility that wind may contribute to viral transmission and sheds light on the scientific subject of how meteorological conditions affect the development of new cases. Less windy areas may have infected droplets hanging in the air for longer, accelerating the spread of the disease.

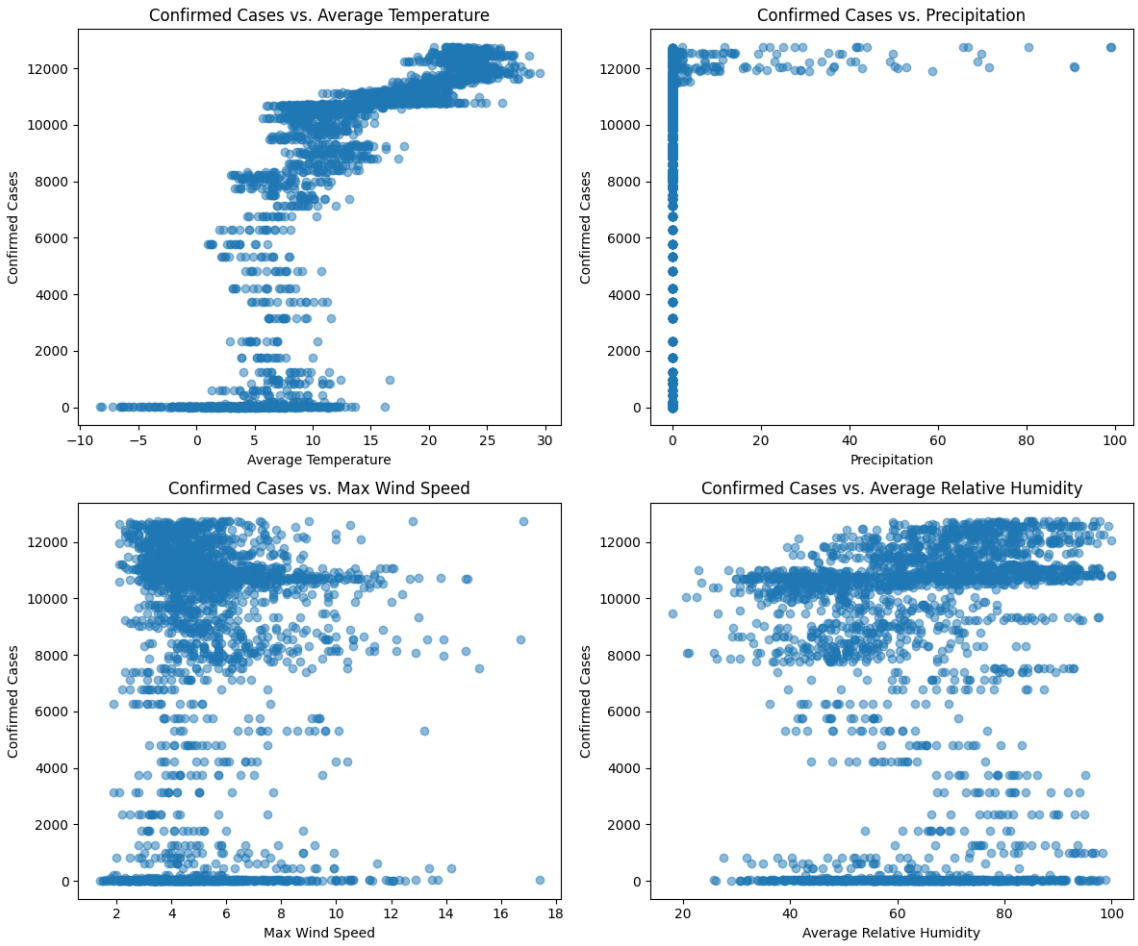

Figure 5: Scatter Plot of Weather vs Confirmed cases

4.3 Research Question 3

Does age play an important role in Covid-19?

The stacked bar graph clearly shows how the mortality rate significantly increases with age. Death rates in the oldest age groups are over 2.0, while those in the youngest age groups are less than 0.5. This answers the research question of demographic effects by emphasizing that advancing age is the biggest risk factor for infection-related death. The illustration successfully draws attention to the increased risks that older populations confront.

We used careful graphic design decisions to improve interpretability. It is simpler to understand the information displayed because to the categories' logical arrangement, the constant usage of colors, and the plain data labels on the bars. The visualization ensures the integrity of the data display by accurately representing the entire dataset.

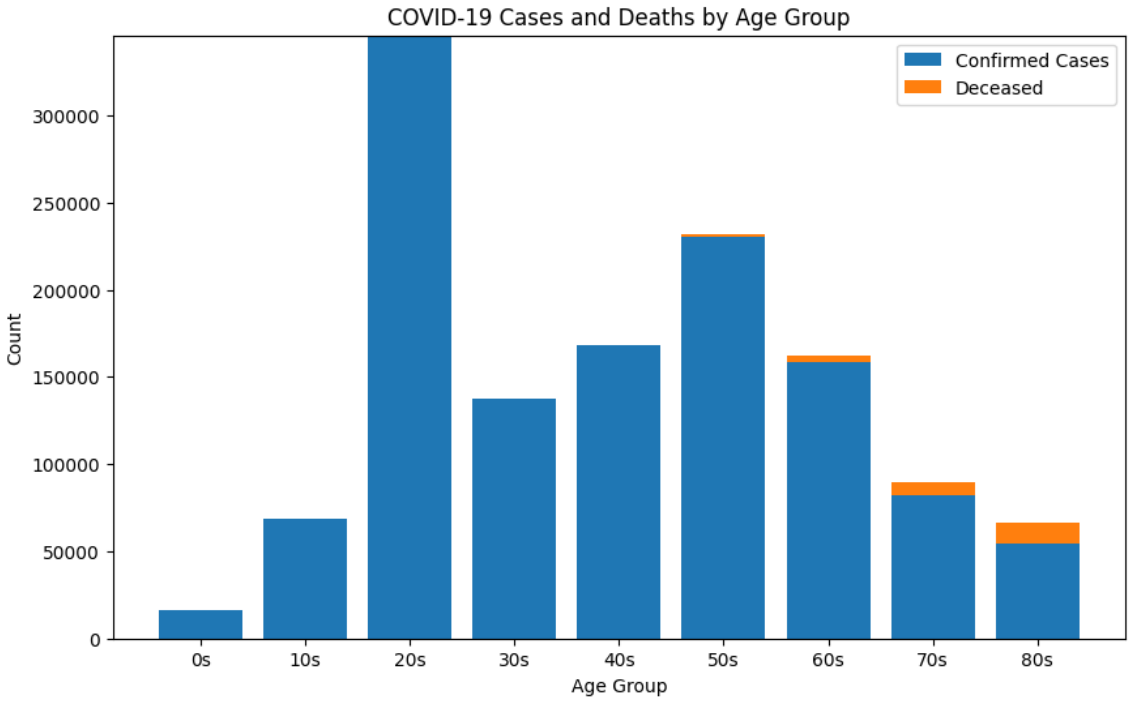

Figure 6: Enter Caption

4.4 Research Question 4

Has there been any variation or evolution in the symptoms of COVID-19 over time?

The line graph shows a substantial increase in COVID-19 instances that have been confirmed while the percentage of positive tests fell from 2 to 0.5. The relationship between instances and viral load is connected, which answers the research question. Notably, despite a rise in the virus's dissemination as more mild and asymptomatic cases appeared, positive percentages fell.

The line chart's superior visual encoding makes clear how excellent its graphics are. The readability and interpretability of the chart is improved by distinctive styles, data point markers, gridlines, and unambiguous labels. The visualization maintains correctness and reliability in communicating the observed link between transmission and viral load by presenting the entire dataset without any alterations.

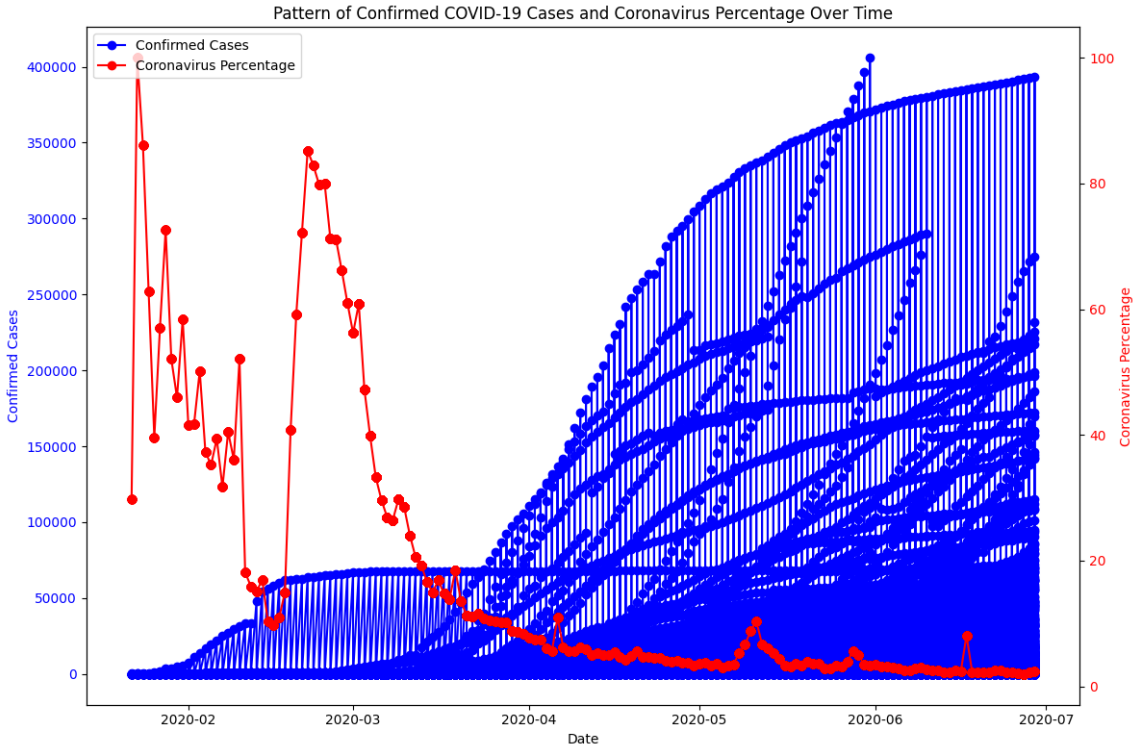

Figure 7: Enter Caption

4.5 Result

Visualization Performance:

To evaluate the system's efficiency, the average running times for creating the scatter plot and the animated choropleth map were recorded. For the animated choropleth map, it took about 2 seconds on average to create a full representation for a particular time frame. Users were guaranteed fluid animation and real-time interaction as a result.

The average time to generate and refresh the visualization for a new dataset for the scatter plot was roughly 1.5 seconds. Users were able to efficiently explore multiple subsets of data and track trends in real time as a result.

5 Discussion

This visualization study illustrates a number of excellent practices for expressing ideas clearly and drawing conclusions from data. Maps, scatter plots, and time series charts are examples of different visualization formats that can be creatively used to illustrate complex patterns and relationships. Users can interact with the data by zooming in on specifics and animating through time thanks to the use of interactive components. Careful visual

design decisions might further improve comprehension while preventing misunderstandings. Conscious use of color, content richness, and sensible design all assist the focus audience's attention. While visualizations should be clear, they also greatly benefit from the context that has been established by thoughtful annotations. One can create information-rich data tales like those given here by fusing visualization expertise with topic knowledge in order to educate audiences and encourage reasoned action.

6 Future Work

- Include fresh information about COVID-19 variants, vaccines, etc. to illustrate current trends.
- Create dynamic data-querying interactive dashboards specifically for public health officials.
- Automated insights can be produced by using machine learning for anomaly detection and pattern recognition.
- Investigate cutting-edge visualizations including networked graphs, 3D maps, and augmented reality.
- Apply comparable visualization and machine learning techniques to fields of science other than public health.

7 Reference

1. https://ieeexplore.ieee.org/document/9373130

2. https://covid19.who.int

3. https://www.kaggle.com/code/shutupandsquat/covid19-explained

- C.K. Leung, P.P. Irani and C.L. Carmichael, "FIsViz: a frequent itemset visualizer", PAKDD, pp. 644-652, 2008.
- C.K. Leung and C.L. Carmichael, "FpVAT: a visual analytic tool for supporting frequent pattern mining", ACM SIGKDD Explor, vol. 11, no. 2, pp. 39-48, 2009.
- C.K. Leung, V.V. Kononov and A.G.M. Pazdor, "PyramidViz: visual analytics and big data visualization of frequent patterns", IEEE DASC-PICom-DataCom-CyberSciTech 2016, pp. 913-916.
- C.K. Leung, F. Jiang and P.P. Irani, "FpMapViz: a space-filling visualization for frequent patterns", IEEE ICDM 2011 Workshops, pp. 804-811.