A Data-Driven Analysis of COVID-19's Evolution and Impact

INFO 5709 Sections 001,002 - Data Visualization and Communication (Fall 2023)

Professor name: Dr. Jian Yang

Project 3

Submitted By:

Kamini Suresh Yelamar

Student ID- 11609561

Master’s in Data Science

University of North Texas

TABLE OF CONTENTS

ABSTRACT

The paper titled "Analyzing COVID-19: A Data-Driven Journey Through the Pandemic" presents a comprehensive examination of the COVID-19 pandemic using data-driven approaches and visualization tools. This research aims to extract valuable insights from extensive COVID-19 datasets, shedding light on various aspects of the pandemic's spread, impact, and response during this unprecedented global health crisis.

Throughout our study, we collect and analyze a diverse range of data, including epidemiological data, hospitalization records, mobility patterns, and administrative information. We employ advanced data analysis techniques and Python-based visualization tools to uncover patterns, correlations, and trends within the data, enabling a deeper understanding of the pandemic's dynamics.

In summary, our work underscores the importance of data-driven research in grasping complex global challenges like the COVID-19 pandemic. We offer a comprehensive and informative overview of the pandemic's evolution, the lessons we've gleaned, and potential pathways toward a more resilient and prepared global community by harnessing the power of data analysis and visualization tools.

INTRODUCTION

The World Health Organization (WHO) officially declared the COVID-19 pandemic on March 11, 2020, in response to its rapid global transmission, which commenced late in 2019, marking the onset of a global health crisis of unprecedented magnitude. When an infectious disease spreads extensively across multiple regions simultaneously, it is termed a pandemic. This global health crisis posed unprecedented challenges to society and individuals worldwide as it continued to spread.

This study primarily centers on the analysis of COVID-19 data derived from the DS4C-PPP dataset, offering valuable insights into the impact of the outbreak on the Korean population. The distinctive feature of the DS4C-PPP dataset, setting it apart from other COVID-19 datasets, is its inclusion of patient-specific information, including dates of symptom onset, confirmation dates, and travel histories. A comprehensive understanding of the epidemiology and infection patterns in South Korea, the assessment of the effectiveness of infection control measures, and the identification of key factors influencing the success of containment efforts all hinge on the availability of such detailed patient data.

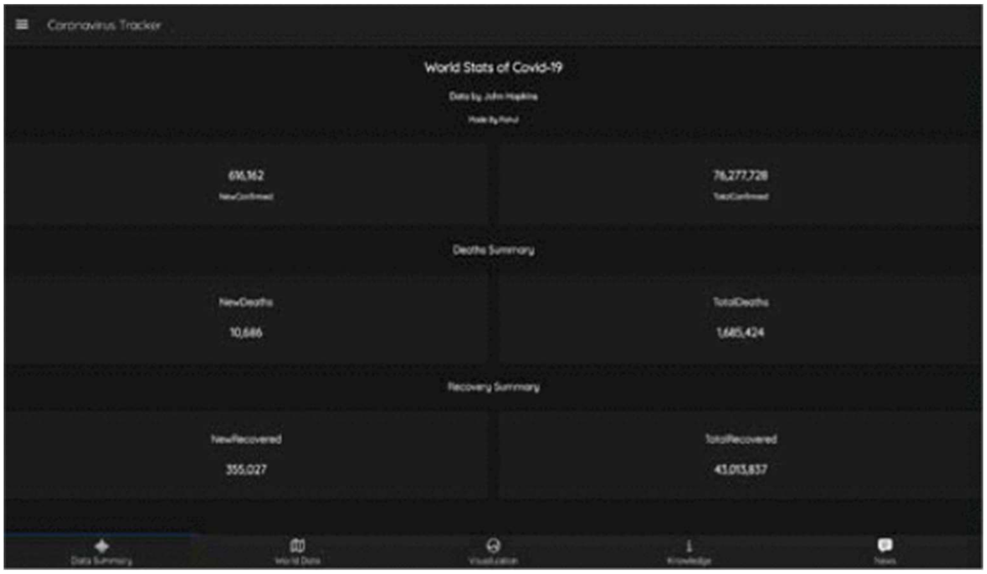
With a relatively short incubation period ranging from 15 to 20 days, the disease exhibited a high mortality rate as the pandemic continued to unfold. Although the epidemic affected every country to some extent, this study specifically concentrates on South Korea due to its access to the dataset facilitated by the collaboration between the Korean government's Ministry of Science and Technology and MindsLab. The consequences of the COVID-19 pandemic have been far-reaching and have varied across different individuals and communities. While some individuals adapted to remote work, online education, and contactless services, others, particularly those involved in essential services, faced heightened risks of infection. Social identities and group affiliations significantly influenced the socio-economic impact and susceptibility to the virus.

RELATED WORK

1. Understanding COVID-19 using Data Visualization [1]

The article focuses on the design and implementation of a data visualization dashboard tailored for COVID-19 data analysis. The abstract outlines the primary objective of the dashboard: to support researchers and enthusiasts in exploring diverse COVID-19 trends, patterns, and facets. This is achieved through the utilization of an innovative interactive dashboard system named "Covid Dashboard," which was specifically developed to operate within the R-Shiny framework. Its overarching goal is to function as an informative and analytical platform for comprehensively examining pandemic-related data.

The paper's content is organized into several sections, with one section dedicated to an in-depth analysis and visualization of COVID-19 data for any selected country. The dashboard facilitates interactive user engagement, allowing users to access various visualizations illustrating daily COVID-19 statistics, including cases, recoveries, and fatalities, both in cumulative and daily formats. Additionally, users can explore metrics such as recovery rate trends, active cases, and more. The paper also includes supplementary information such as recovery rates, total cases, total deaths, and active cases. The latter part of the paper plays a crucial role in addressing misinformation by providing information about the virus, safety measures, and a news section to keep readers updated with the latest verified information.



World Stats of Covid-19

Data by John Hopkins

Made By Rahul

616,162
NewConfirmed

76,277,728
TotalConfirmed

Deaths Summary

NewDeaths
10,686

TotalDeaths
1,685,424

Recovery Summary

NewRecovered
355,027

TotalRecovered
43,013,837

Fig 1

Fig 1 contains a summary of covid-19 data. This summary includes new confirmed, totally confirmed cases, new deaths, total deaths, people recovered last day, and total people recovered till then.

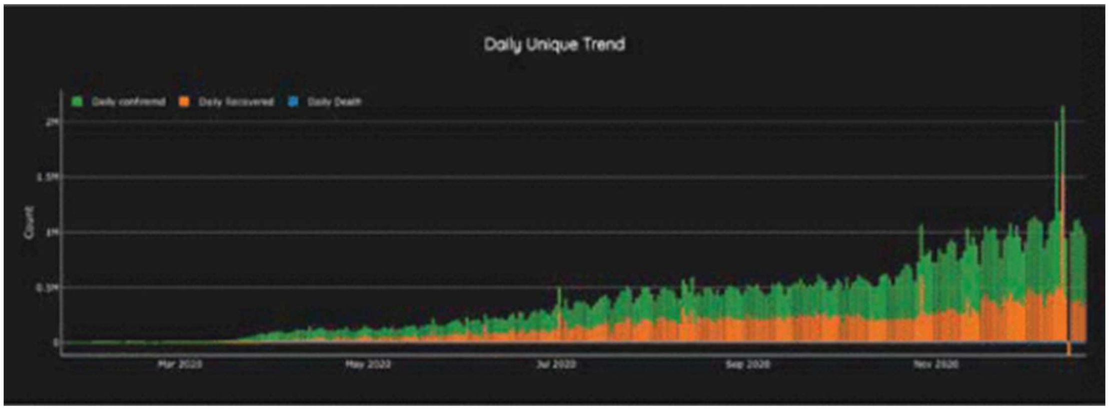


Fig 2

Fig 2 plot was drawn to show the worldwide trend of covid19. On X-axis a date and on the Y-axis, a total count on that date.

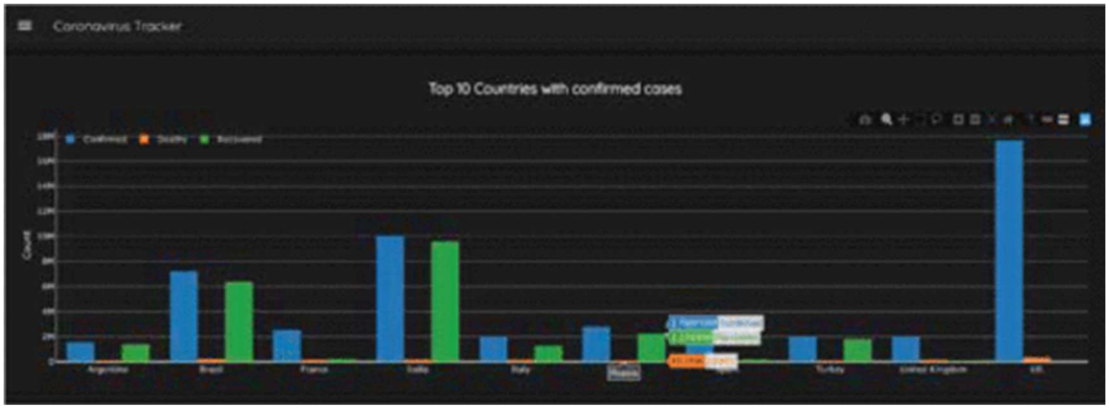


Fig 3

Fig 3 shows Top 10 most affected countries and their data on bar graphs.

2. **Data interpretation and visualization of COVID-19 cases using R programming. [2]**
   The study focuses on the utilization of R programming as a formidable tool for the effective interpretation and visualization of COVID-19 cases. Its central aim revolves around harnessing R's capabilities to facilitate a more profound comprehension of COVID-19 data, emphasizing the significance of programming languages like R in the context of a global health crisis like COVID-19. The paper is expected to offer insights into various aspects of COVID-19 data analysis, encompassing data cleaning, transformation, and visualization techniques, which can be invaluable for researchers and data analysts seeking to extract meaningful insights and trends from COVID-19 datasets.

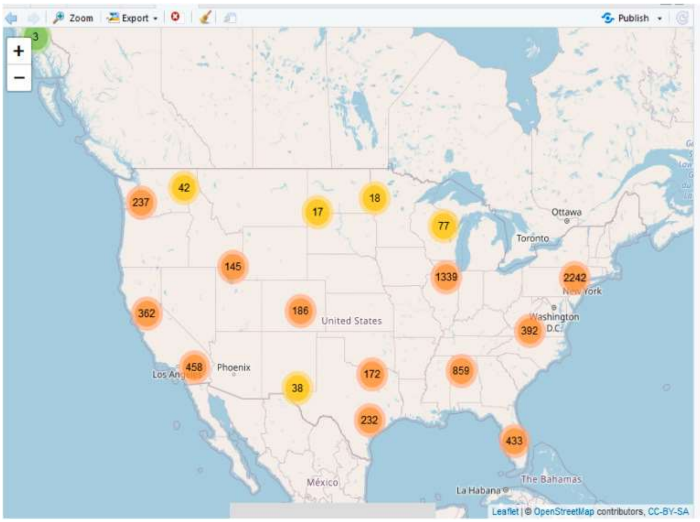


Fig 4

Fig 4 displays Covid maps with respective data of the U.S.

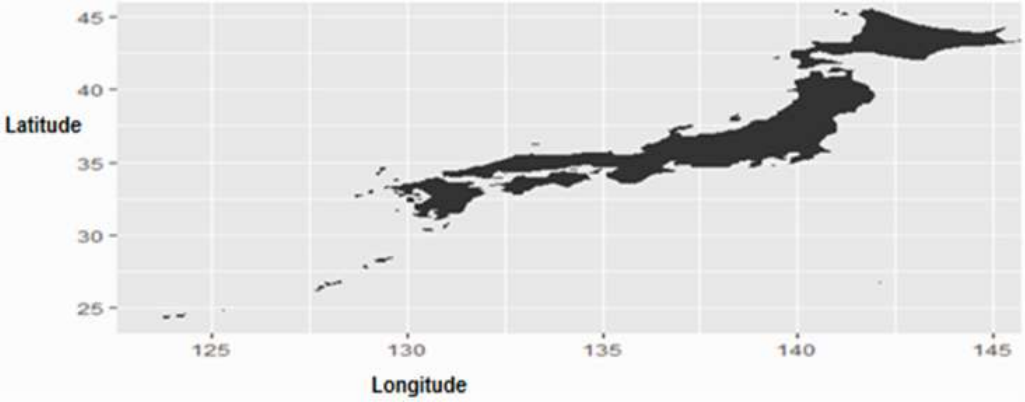


Fig 5

Fig 5 shows Japan Country map. The research utilized application libraries such as ggfortify, mapdata, maps, and ggplot2.

METHODS

1. **Data Preparation**

   The COVID-19 data utilized in this analysis was acquired from publicly accessible repositories hosted on Kaggle. Prior to visualization, a preprocessing phase was undertaken to ensure data readiness. This encompassed procedures such as standardizing column names across datasets, eliminating instances of missing data, and rectifying inaccuracies in country names to facilitate precise graphical representation.

2. **PowerBI, Python Libraries and tools**

   Data carpentry has been done in google colab using Python. Mapping Continents, data labeling and a few data processing tasks were performed. For dataset named Trends, Power Query Editor of PowerBI is used to aggregate and perform operations on data.

3. **Types of Visualizations**

   Total four types of visualization have been addressed as below:

   - Dashboard is created as a combination of pie charts, column charts and area charts for Covid-19 Data.
   - A stacked bar chart is employed to illustrate the proportion of fatalities among different age groups, providing insight into the intensity of the cases by age.
   - The world map graph is developed to demonstrate the geographical spread of cases over time, with the size of bubble effectively depicting the change in case distribution across different areas.
   - Pie charts are used to answer yearly changes in symptoms of different diseases.

   Each style of visualization served a distinct purpose, aiming to highlight facets of the data, such as cumulative totals, geographical dispersion, possible correlations, detailed breakdowns, and emerging trends. Collectively, these visualizations contribute to a comprehensive perspective on the pandemic.

4. **Interactive Visualizations**

   All visualizations created are interactive. On hovering, tooltip displays legends/details selected which makes visualization easy to read. To facilitate in-depth examination, every visualization includes pan and zoom functionalities.

5. Graphical Excellence

- Picked colors that everyone, including those with color-blindness, could easily distinguish, ensuring that our visualizations are accessible and well-crafted.
- To make it simple and to follow along as data changes or transitions occur, added interactive highlighting, making all visuals clear and easy to understand.
- Created labels, legends, and markers on charts that can be read easily, whether we're zoomed in or out, ensuring that all the visuals are user-friendly.
- Maintained a balance between packing in information and offering interactive details when we need them, preventing data overload while still giving a comprehensive view.
- By using a consistent visual style and theme throughout, made it easy to connect the dots between different charts, providing a seamless and excellent graphical experience.

VISUALIZATIONS

1. COVID-19 Dashboard

The dashboard displays various charts and graphs related to COVID-19 data analysis:

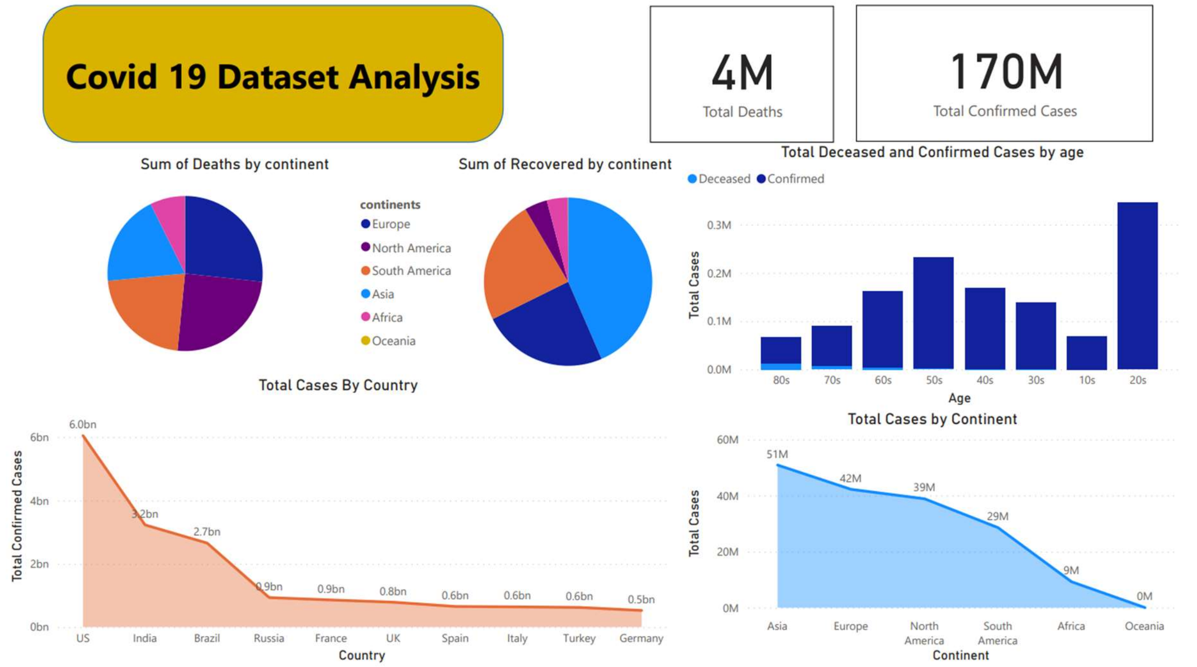


Fig 6

- ***Total Deaths and Confirmed Cases:*** These boxes highlight key figures: 4 million total deaths and 170 million total confirmed cases globally.
- ***Sum of Deaths by continent**:* This pie chart shows the distribution of COVID-19 related deaths across different continents, with each slice of the pie representing a continent's share of total deaths.
- ***Sum of Recovered by continent:*** Similar to the first pie chart, this one represents the distribution of recovered COVID-19 patients across continents.
- ***Total Deceased and Confirmed Cases by age:*** This bar chart illustrates the distribution of confirmed cases and deaths by different age groups, showing higher numbers in some age brackets.
- ***Total Cases by Country:*** This area chart shows the total confirmed cases by country, with countries like the US, India, and Brazil having the highest numbers.
- ***Total Cases by Continent:*** This area chart depicts the total confirmed cases across continents, with Asia having the highest count, followed by Europe, North America, South America, Africa, and Oceania.

2. Research Questions

Q1. What was the overall spread across the globe?

The animated global world map effectively demonstrates the geographic spread of COVID-19 cases, originating in China and subsequently spreading globally. This visual depiction of the temporal evolution addresses the research question regarding the patterns of transmission.

Design Rationale:

The aim is to create a design that is not only informative and accurate but also engaging and intuitive for the user, regardless of their background or expertise.

The use of bubbles of varying sizes on a world map provides an immediate visual cue about the relative number of cases in each region. Larger bubbles draw attention to areas with more severe outbreaks. Placing the bubbles on a world map allows for geographic context to be instantly understood, helping viewers to locate areas of interest or concern quickly. By providing a global view, the map facilitates comparative analysis between different regions and countries, highlighting disparities in the pandemic's impact.

The map is interactive, users can engage with the data more deeply, perhaps clicking on a bubble to receive more detailed information about the cases in that region. The use of color and size makes the data accessible to a broad audience, including those who may not be able to understand more complex statistical representations.

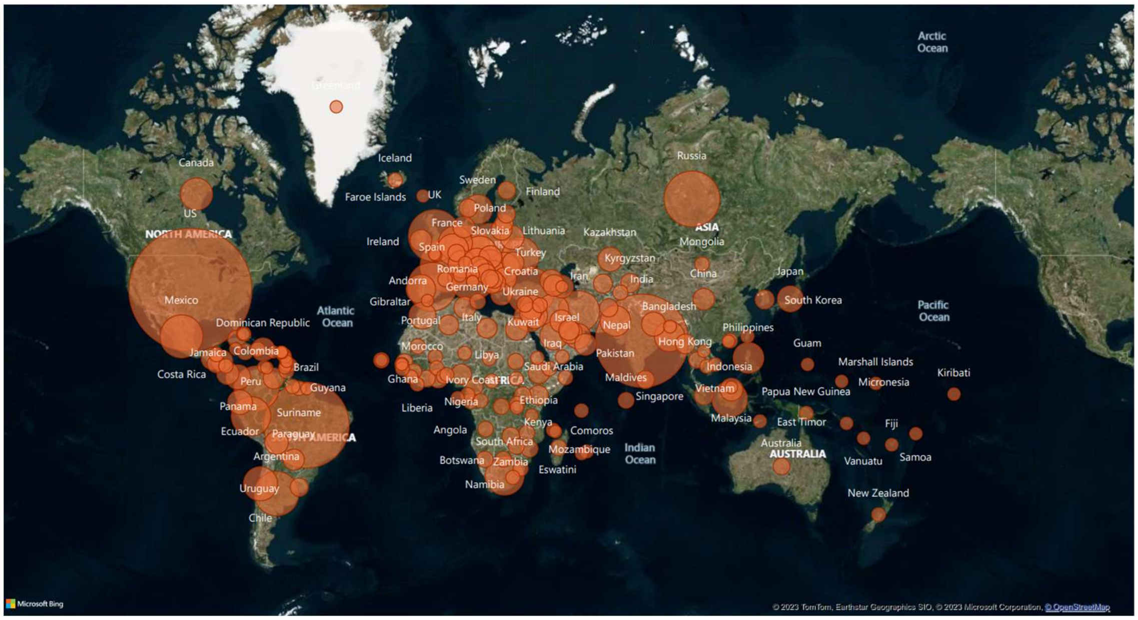


Fig 7

Q2. Does age play an important role in Covid-19 Death?

The stacked bar graph provides a clear representation of how the mortality rate experiences a significant rise as individuals age. In the oldest age groups, the death rates surpass 2.0, whereas in the youngest age brackets, they remain below 0.5. This addresses the research question regarding demographic impacts by underscoring that advancing age constitutes the most substantial risk factor for infection-related fatalities.

Design Rationale:

The bar chart is chosen for its straightforward representation of data. Age groups are segmented by decades, a common method for demographic studies, providing clear differentiation between the data points. By aligning the bars side by side, the design enables easy comparison across age groups, highlighting trends such as which age groups have higher or lower-case counts. It effectively highlights the heightened vulnerability of older populations to such risks. The graph focuses on both confirmed and deceased cases, two critical metrics for understanding the impact of the pandemic across different age demographics. The scaling on the Y-axis is selected to accommodate the range of the data, ensuring that even age groups with relatively low case counts are visible and comparable. The design avoids unnecessary elements that could complicate interpretation, making it universally understandable regardless of the viewer's background.

The overall design is effective in communicating the essential information about COVID-19's impact on different age groups, allowing for immediate visual analysis and interpretation.

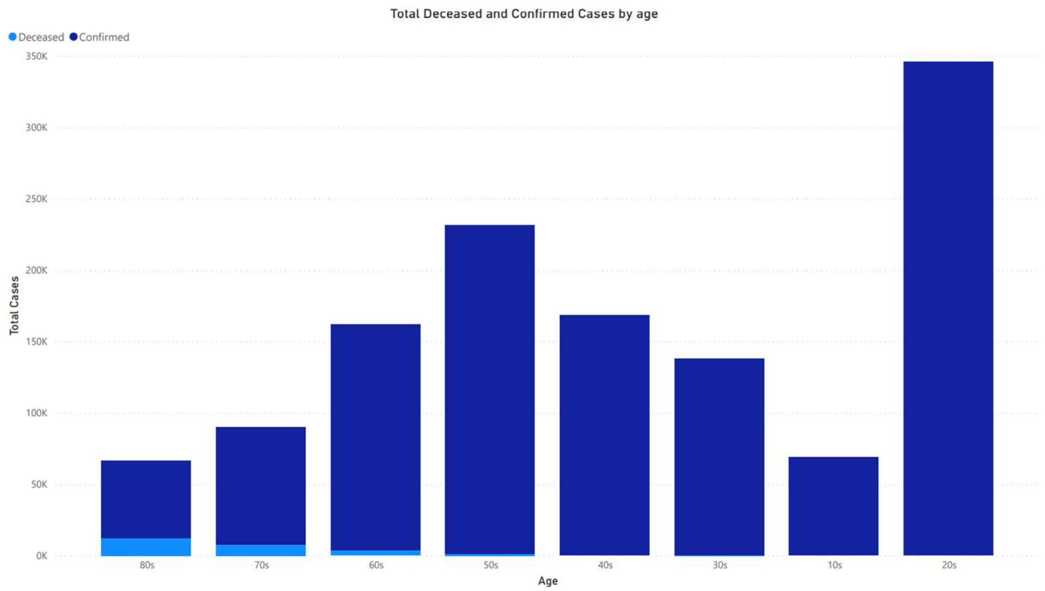


Fig 8

Q3. Has the symptomatology of COVID-19 exhibited any changes or developmental progressions throughout the course of the pandemic?

The years are labeled with the last chart showing a dominant segment for coronavirus, indicating a surge in cases, or reported symptoms for that year. The preceding years show a more balanced distribution among the cold, flu, and pneumonia. This set of pie charts visually demonstrates how the landscape of respiratory illnesses has shifted over the specified period. And answers that COVID-19 has had a significant impact on the overall landscape of respiratory illnesses since its emergence.

Design Rationale:

The visualization is designed to communicate the shifts in the prevalence of respiratory illnesses over time, with a particular emphasis on the stark increase in coronavirus symptoms in 2020, likely due to the COVID-19 pandemic.

Pie charts were chosen for their ability to show parts of a whole, allowing for immediate visual comparison of the data for each year. Different colors likely represent different illnesses to differentiate data points clearly and aid in visual distinction when comparing the charts. Presenting the data in a series of pie charts facilitates the observation of trends over time. The viewer can easily compare each year's data to see the changes in the prevalence of each illness. The pie chart's proportional segments visually encode the data, making it clear when one category (like coronavirus in 2020) becomes predominant. The design is straightforward, making it accessible to a broad audience, regardless of their ability to understand complex data visualizations.

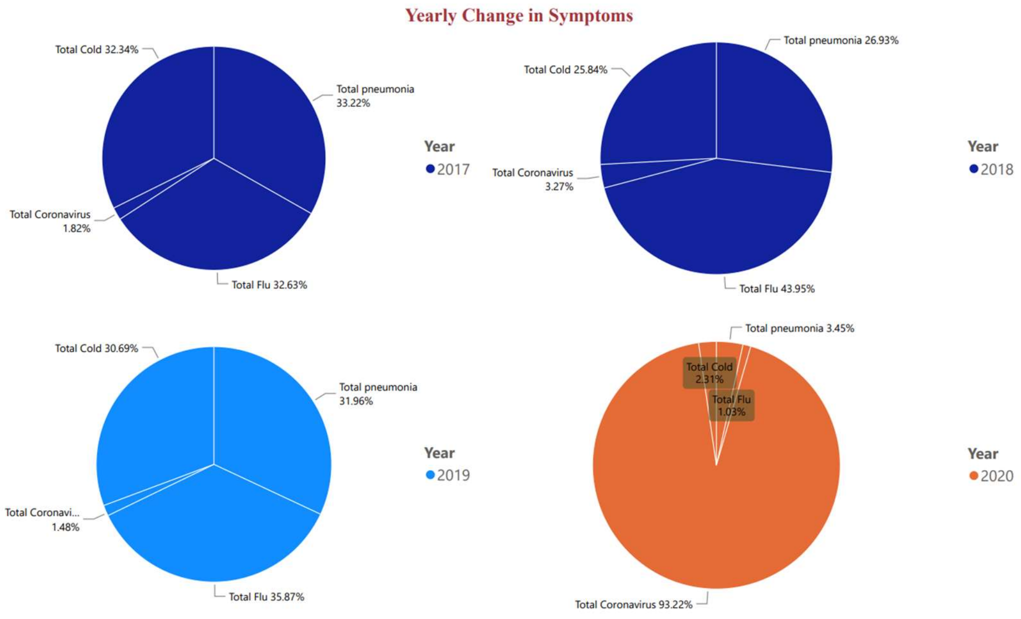


Fig 9

RESULTS:

The use of PowerBI has made plotting all the data easy and fast. Within a few seconds visualizations were created. And PowerBI also gives flexibility to modify the data to check updated results/visualization. Answering research questions became simple. Briefly, these results underscore the significant impact of the COVID-19 pandemic on public health, with implications for healthcare resource allocation, disease monitoring, and preventive health strategies. Also, age played an important role during the pandemic. Since immunity and strength are aligned with it.

If ever user wants to generate new visualization, PowerBI gives ability to update the data in GUI and fetch new visualizations in a minute.

DISCUSSION:

This study on data visualization demonstrates several best practices for presenting ideas with clarity and deriving insights from data. It showcases how various forms of data representation such as maps, column charts, and pie graphs can be effectively utilized to highlight intricate patterns and correlations. Interactive elements enable users to delve into specific details and observe trends over time through animations. Thoughtful visual design choices can enhance understanding and prevent misinterpretation. The judicious use of color, detailed content, and intentional design all serve to guide the viewer’s attention. Clear visualizations are significantly enriched by the context provided through strategic annotations. By integrating visual elements with narrative, one can craft compelling and informative data stories similar to those presented in this study.

FUTURE WORK:

- Incorporate up-to-date data on COVID-19 variants, vaccines, etc., to reflect the latest developments.

- Develop interactive dashboards with dynamic data querying capabilities tailored for use by public health authorities.

- Utilize machine learning for automated anomaly detection and pattern recognition to generate insightful findings.

- Explore innovative visualization methods, such as networked graphs, 3D maps, and augmented reality, to enhance data representation.

- Apply similar visualization and machine learning approaches to domains beyond public health, expanding their utility in various scientific fields.

REFERENCES

1. R. Chauhan, P. Goel, V. Kumar, N. Soni and N. singh, "Understanding Covid-19 using data visualization," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 555-559, doi: 10.1109/ICACITE51222.2021.9404700.
2. Rimal Y, Gochhait S, Bisht A. Data interpretation and visualization of COVID-19 cases using R programming. Inform Med Unlocked. 2021;26:100705. doi: 10.1016/j.imu.2021.100705. Epub 2021 Aug 30. PMID: 34485681; PMCID: PMC8404394.