Visualization and Analysis of trends in COVID-19 Deaths by State, HHS Region, Race, and Age in the United States

1st Pranit Kumbhar *Computer Science Department Pace University* New York, USA [pranit.kumbhar@pace.edu](mailto:pranit.kumbhar@pace.edu)

2nd Vivek Karhale *Computer Science Department Pace University* New York, USA vivek.a.karhale@pace.ed

**Abstract - The variety of responses to and conditions caused by the COVID-19 pandemic in the United States has provided researchers with a wealth of data to study, especially as the pandemic progresses. With more than a year of data available in various regions and at various granularities, methods of analysis that require larger datasets are now worth investigating or refining. Furthermore, as the United States moves away from national and state-wide policies and toward approaches centered on individual communities, open and easily readable data must be made available at both the state and county levels. Using google colab and python, a comprehensive visualization encompassing COVID-19 data and a large body of related data is proposed in this paper.**

1. INTRODUCTION

The novel coronavirus (COVID-19) outbreak began in Wuhan, China, in late 2019. Studies on a patient admitted to the hospital on December 12th, 2019, resulted in the discovery of a new coronavirus in January of 2020. Since then, the virus has spread throughout the world, with COVID-19 declared a pandemic on March 11th, 2020, and cases reported in 192 countries as of the date of writing, according to Johns Hopkins University CSSE data. There have been over thirty million cases and over five hundred thousand deaths in the United States.

Notably, there was an increase in scientific research on all aspects of pandemic science, from epidemiological to societal and psychological, possibly influenced by the ubiquity of experience provided by the pandemic. The WHO's Global Literature on Coronavirus Disease database contained over 2,000 papers. These studies' large amounts of data provide critical research areas, but at the expense of the extensive time and effort required to collect data from disparate sources and formats. This procedure is ongoing.

Many scientists will be involved, delaying critical research and insights. The proposed database was created to address this problem by providing a curated collection of data covering the most common research areas.

1. RELATED WORKS

One of the first and possibly most well-known sources of COVID-19 data was the Johns Hopkins CSSE Dataset. The popular Johns Hopkins COVID-19 dashboard is powered by this data. At the federal, state, and even international levels, it has served as a reliable source of case and death statistics. Another early source of data was the COVID Tracking Project, an effort to manually gather, cross-check, and publish COVID-19 data from 56 US states and territories in three primary areas: testing, hospitalization, and patient outcomes.

This dataset was made available in March 21 by the U.S. Department of Health and Human Services; since then, it has been updated on a regular basis.

Data Analysis:

"Provisional COVID-19 Deaths by HHS Region, Race, and Age" (Centers for Disease Control and Prevention): This dataset provides provisional counts of COVID-19 deaths in the US, categorized by HHS region, race, and age group. This can be a valuable source for both data analysis and visualization.

"COVID-19 Provisional Counts - Weekly Updates by Select Demographic and Geographic Characteristics" (Centers for Disease Control and Prevention): This website provides weekly updates on provisional COVID-19 deaths, categorized by age, sex, race, Hispanic origin, and comorbidities. It also offers an index of state-level and county-level data for download.

Both the datasets have data as of 09/27/2023 starting from 12/29/2019.

Data Visualization:

"Mapping COVID-19 Mortality Risk Factors in US Counties" (Nature Medicine, 2020): This study uses interactive maps to visualize the geographic distribution of COVID-19 mortality risk factors, including age, race, and underlying health conditions.

"COVID-19 Racial and Ethnic Disparities Dashboard" (NPR): This interactive dashboard presents data on COVID-19 cases, hospitalizations, and deaths by race and ethnicity across the US.

"COVID-19 Data Tracker" (Johns Hopkins Coronavirus Resource Center): This website provides a comprehensive set of COVID-19 data visualizations, including maps, charts, and graphs, covering cases, deaths, vaccinations, and other metrics.

1. DATASET

For the visualization and analysis, we have used the datasets provided by the US department of Health and Science on their website and can be downloaded easily. This dataset covers number of deaths due to COVID-19 and total deaths in the United States collected weekly from start of pandemic until September 2023. The datasets were obtained from the US Department of Health and Human Services and the National Center for Health Statistics.

1. *Geographic Data*

The project contains COVID-19 data related to the United States. The United States provides a unique opportunity for insight based on its size, by states and HHS regions, further summarized by race, Hispanic origin, age group.. Each of these jurisdictions plays a role in crafting response and mitigation measures, providing directly comparable regions for analyzing the trends over these criteria during the spread of COVID-19. The dataset captures these hierarchical relationships and geographical adjacencies. Geographic characteristics like location (e.g. state and region) have been considered.

A graph of covid-19 deaths

Description automatically generated

These states have also been categorized and grouped into regions.

Deaths involving COVID-19 reported to NCHS by time-period, HHS region, race and Hispanic origin, and age group.

United States death counts include the 50 states, plus the District of Columbia and New York City. The ten (10) United States Department of Health and Human Services (HHS) regions include the following jurisdictions. Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont; Region 2: New Jersey, New York, New York City, Puerto Rico; Region 3: Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, West Virginia; Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee; Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin; Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, Texas; Region 7: Iowa, Kansas, Missouri, Nebraska; Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming; Region 9: Arizona, California, Hawaii, Nevada; Region 10: Alaska, Idaho, Oregon, Washington.

Below is the distribution of Covid-19 deaths by HHS Regions. This graph shows that HHS region was affected due to COVID-19 region as this is the region of dense population and having more influx of tourist and migrants. This can be a contributing factor to it. But after that this trend drastically reduced. HHS region in South of United States face the drastic variations in number of COVID-19 deaths.

A graph of different colored lines

Description automatically generated

California and Texas were the most affected states from COIVD-19. This data separates New York City from New York State. Considering that New York is the third largest affected state in the United States

The next bar graph shows the most affected 10 states in the United State.

A graph of the number of states

Description automatically generated with medium confidence

1. *Demographic Data*

To further study the relationship between socioeconomic factors and COVID-19, our project has incorporated demographic data from the same source. It mainly includes Age and race.

From the below chart we can see that,

The total number of deaths from COVID-19 in the United States is over 311,000 in age group of 85 and more years. The number of deaths in the 65-74 age group is the second highest, followed by the 75-84 age group. The number of deaths in the 30-39 and 40-49 age groups is relatively low.

A graph of covid-19 deaths

Description automatically generated

We have also classified the data and distributed it over race and age groups. Used scatter plots as they

offer a powerful and flexible way to visually explore and analyze relationships between two continuous variables. Their versatility makes them a valuable tool for scientists, engineers, data analysts, and anyone looking to gain insights from data.

A graph with a number of dots and numbers

Description automatically generated with medium confidence

IV. RESULTS AND DISCUSSIONS

Looking at the 1st chart (Number of Deaths in thousands vs Age group) People over 65 are by far the most likely to die from COVID-19. The number of deaths in the 85+ age group is more than 10 times higher than in any other age group.

The number of deaths has been steadily increasing for all age groups since the beginning of the pandemic. However, the rate of increase has slowed down in recent months.

The youngest age groups (0-4 and 5-17) have the fewest deaths. However, there has been a slight uptick in deaths in the 0-4 age group in recent months.

In the 2nd chart, it shows a clear trend of increased COVID-19 deaths with age, across all racial and ethnic groups.

Non-Hispanic Whites have the highest overall number of deaths, followed by Hispanics and Non-Hispanic Blacks.

Age groups 65 and above account for the vast majority of deaths in all racial and ethnic groups.

Specific Age Groups:

Non-Hispanic Whites: The 85+ age group has the highest number of deaths, followed by the 75-84 age group. The death rate then gradually declines for younger age groups.

Hispanics: Similar to Non-Hispanic Whites, the 85+ age group has the highest number of deaths, followed by a gradual decline with younger age groups.

Non-Hispanic Blacks: The 75-84 age group has the highest number of deaths, followed closely by the 85+ age group. The death rate then declines for younger age groups.

Non-Hispanic Asian: The 85+ age group has the highest number of deaths, followed by a steeper decline compared to other groups for younger age groups.

Non-Hispanic American Indian or Alaska Native: Data is limited for this group, but the 65-74 age group appears to have the highest number of deaths, followed by a decline for younger age groups.

Native Hawaiian or Other Pacific Islander: Data is limited for this group, but the 75-84 age group appears to have the highest number of deaths.

Disparities:

While Non-Hispanic Whites have the highest overall number of deaths, Black, Hispanic, and American Indian/Alaska Native populations have higher death rates when adjusted for age. This suggests underlying social and economic factors contributing to these disparities.

COVID-19 specific deaths compared to total deaths and COVID deaths over the year are shown below:

A graph of a number of deaths

Description automatically generated

A graph of a number of deaths

Description automatically generated

COVID-19 deaths initially surge, exceeding total deaths in the early stages of the pandemic. This highlights the significant impact of COVID-19 on mortality rates.

Over time, COVID-19 deaths gradually decline, approaching total deaths. This suggests a combination of factors, including:

Public health interventions: Measures like mask-wearing, social distancing, and vaccination campaigns likely helped reduce transmission and deaths.

Improved medical treatment: Advancements in treatment protocols may have improved survival rates for COVID-19 patients.

In the 2nd chart, COVID-19 deaths have fluctuated throughout the year, with several peaks and valleys. This suggests ongoing waves of infection and varying effectiveness of control measures.

The highest peak in deaths appears to be in Week 25, which could correspond to a specific event or variant surge.

Deaths have generally declined since the peak, although there are smaller upticks throughout the remaining weeks. This suggests some control over the pandemic but continued risk.

1. CONCLUSIONS

In this work, a comprehensive visualization of COVID-19 specific and related data was proposed. The proposed aggregates relevant data from verified sources. They are classified into the following categories: State, Region, Age and Race.

The reasons for the disparities in the Results section are complex and multifaceted, but may include:

Access to healthcare: People of color may face barriers to accessing quality healthcare, leading to delayed diagnosis and treatment.

Underlying health conditions: People of color are more likely to have underlying health conditions, such as diabetes and heart disease, which can increase the risk of complications from COVID-19.

Social determinants of health: Factors such as poverty, housing insecurity, and occupational risks can contribute to poorer health outcomes for people of color.

It is important to continue investigating the reasons for racial and ethnic disparities in COVID-19 deaths to inform public health interventions and policies aimed at reducing these disparities.

More research is needed to understand the specific factors contributing to these disparities in different communities.

Culturally competent and targeted interventions are needed to address the social and economic determinants of health that contribute to COVID-19 disparities.

REFERENCES

<https://catalog.data.gov/dataset/provisional-weekly-deaths-by-region-race-age-997d6>

<https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab/about_data>

<https://ieeexplore.ieee.org/document/9990151>

<https://ieeexplore.ieee.org/document/9725751>

https://ieeexplore.ieee.org/document/9853266