IST 652 : Scripting for Data Analysis

COVID-19 Deaths & Mentions in the United States

Team 6:
Mahitha Chennamadhava
Shubham Patil
Kayla Carleton

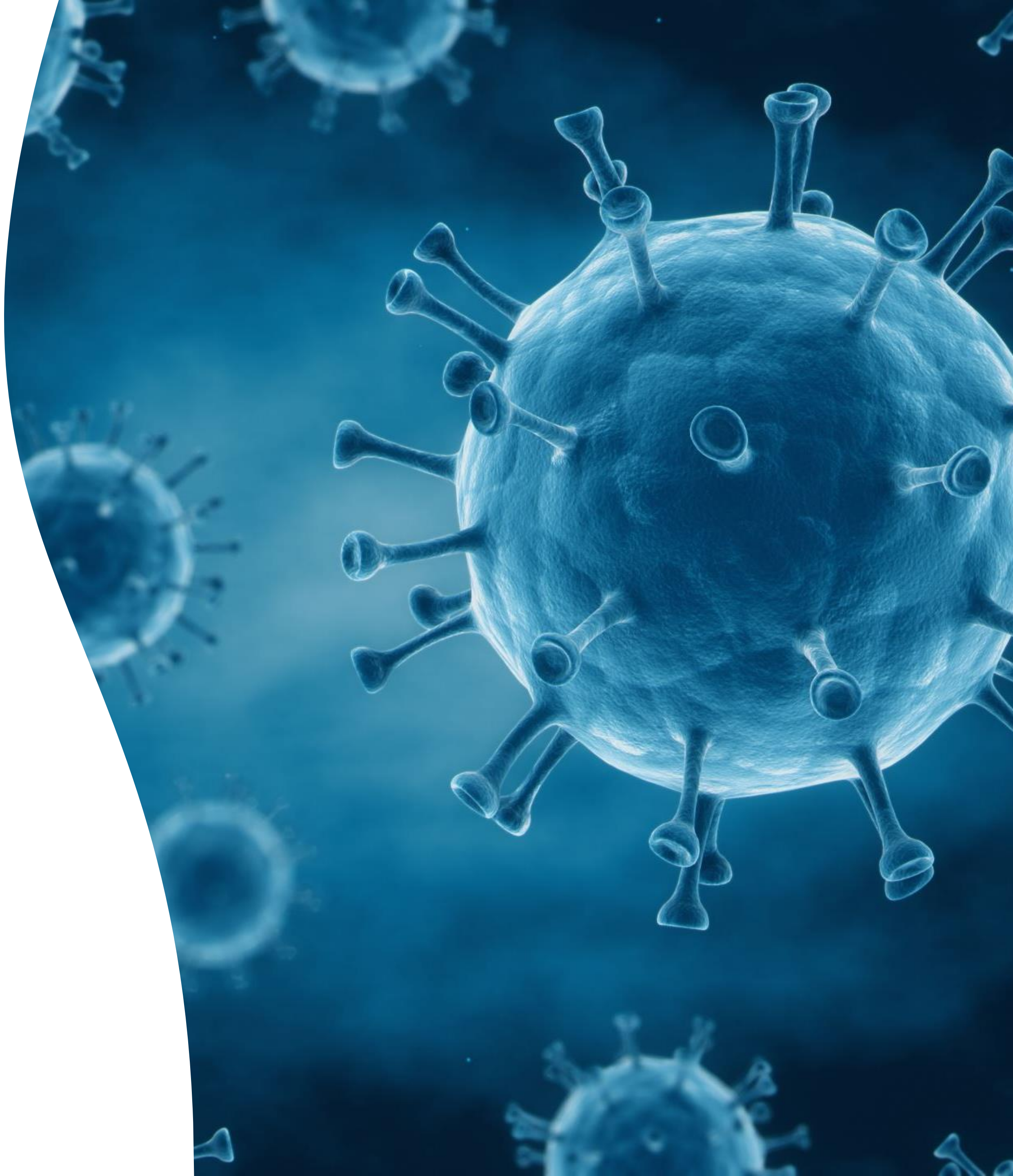


Table of Contents

Introduction

- When we sat down to discuss the dataset for our final project, we all agreed to do something that was health related and something that was recent
- The pandemic left many disturbances today including the underlying health conditions affiliated with COVID. This in return opens a wide range of possibilities in order to understand future precautionary measures
- In this presentation we will cover how COVID-19 deaths impacting a range of factors in the United States

Objectives

Visualization for Communication:

- Use visualizations (such as histograms, line plots, and heatmaps) to effectively communicate trends and patterns to a diverse audience, making the analysis accessible and informative.

Predictive Modeling

- Implement predictive models into our analysis and forecast future trends in COVID-19 deaths based on historical data.

Comparative Analysis:

- Compare COVID-19 death rates across different health conditions, states, or demographics to draw insights into the effectiveness of public health measures and healthcare systems.

Understanding the Impact:

- Analyze the overall impact of COVID-19 on mortality rates, identifying trends and patterns in the number of deaths over the pandemic.

Geospatial Analysis:

- Explore geographic variations in COVID-19 deaths, examining how different states have been affected using heatmaps

Demographic Patterns:

- Investigate demographic factors such as age and underlying health conditions to understand how different populations are affected by COVID-19

Data Description

```
RangeIndex: 621000 entries, 0 to 620999
Data columns (total 22 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   sid                 621000 non-null  object
 1   id                  621000 non-null  object
 2   position            621000 non-null  int64
 3   created_at          621000 non-null  int64
 4   created_meta        0 non-null       object
 5   updated_at          621000 non-null  int64
 6   updated_meta        0 non-null       object
 7   meta                621000 non-null  object
 8   Data As Of          621000 non-null  object
 9   Start Date          621000 non-null  object
 10  End Date            621000 non-null  object
 11  Group               621000 non-null  object
 12  Year                608580 non-null  object
 13  Month               558900 non-null  object
 14  State               621000 non-null  object
 15  Condition Group     621000 non-null  object
 16  Condition           621000 non-null  object
 17  ICD10_codes         621000 non-null  object
 18  Age Group           621000 non-null  object
 19  COVID-19 Deaths     437551 non-null  object
 20  Number of Mentions  443423 non-null  object
 21  Flag                183449 non-null  object
```

- **Source:** The dataset was sourced from Data.gov
- The dataset has 124,200 rows and 22 columns
- **Timeframe**: Covers data from [start date] to [end date], providing a comprehensive view of the COVID-19 impact over time.
- **Scope and Scale:** Comprises over 600,000 records, reflecting a wide range of demographic and clinical data points across the United States.
- **Key Variables**:
- **State**: Geographical location within the United States, excluding entries labeled as 'United States' to focus on individual states.
- **Condition Group and Condition**: Categorization of COVID-19 associated health conditions as per ICD10 clinical codes.
- **COVID-19 Deaths:** Recorded fatalities attributed to COVID-19, requiring conversion from object to numeric data type for analysis.
- **Number of Mentions:** Frequency count of specific conditions or keywords in the dataset, potentially linked to reported deaths.
- **Age Group:** Demographic segmentation of data, which can provide insights into the age-related impact of the pandemic.

Data Cleaning

Variables changed: year & month: we made sure these variables were numeric, removing all commas or characters that were unnecessary

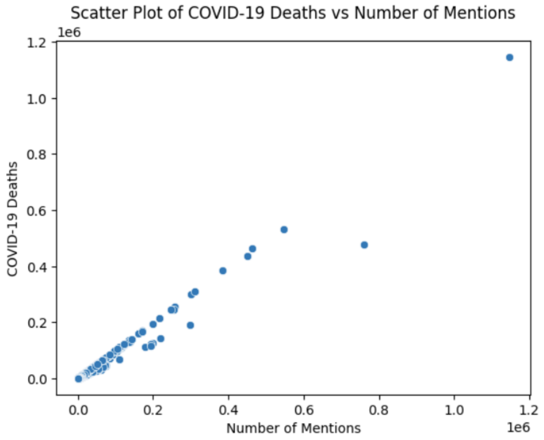
We removed anything labelled "COVID-19" and renamed everything to "deaths"

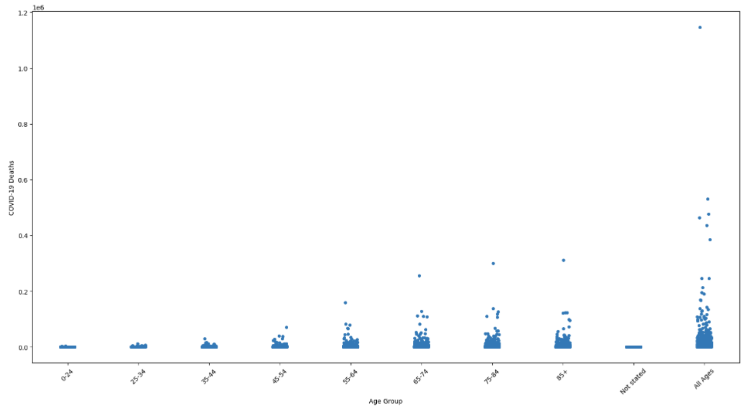
Filled and replaced any values ie. Year with most frequency, etc

Dropped any duplicates

Data Exploration: Scatter Plot & Stripp Plot

- Explored the data with df.head(), df.info(), df.describe()
- Created a scatter plot for Deaths Vs Number of cases to understand the relationship between these variables
- Plotted stripp plot to understand how age group factor has an impact on Covid 19 deaths





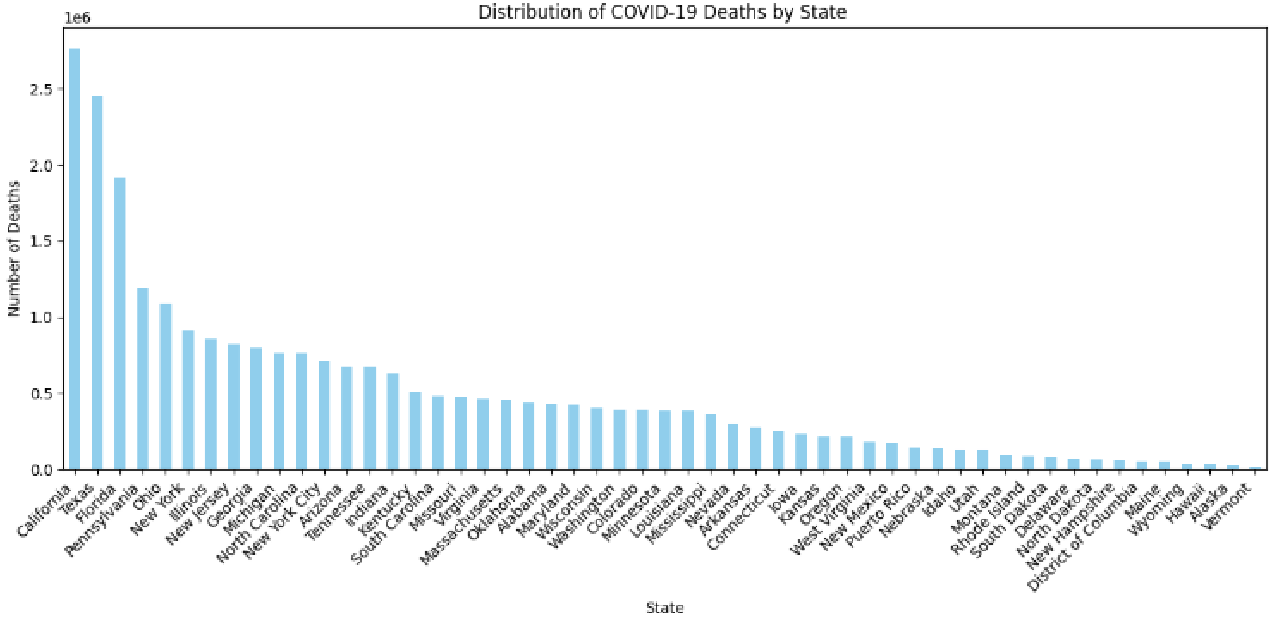
Age Filtering

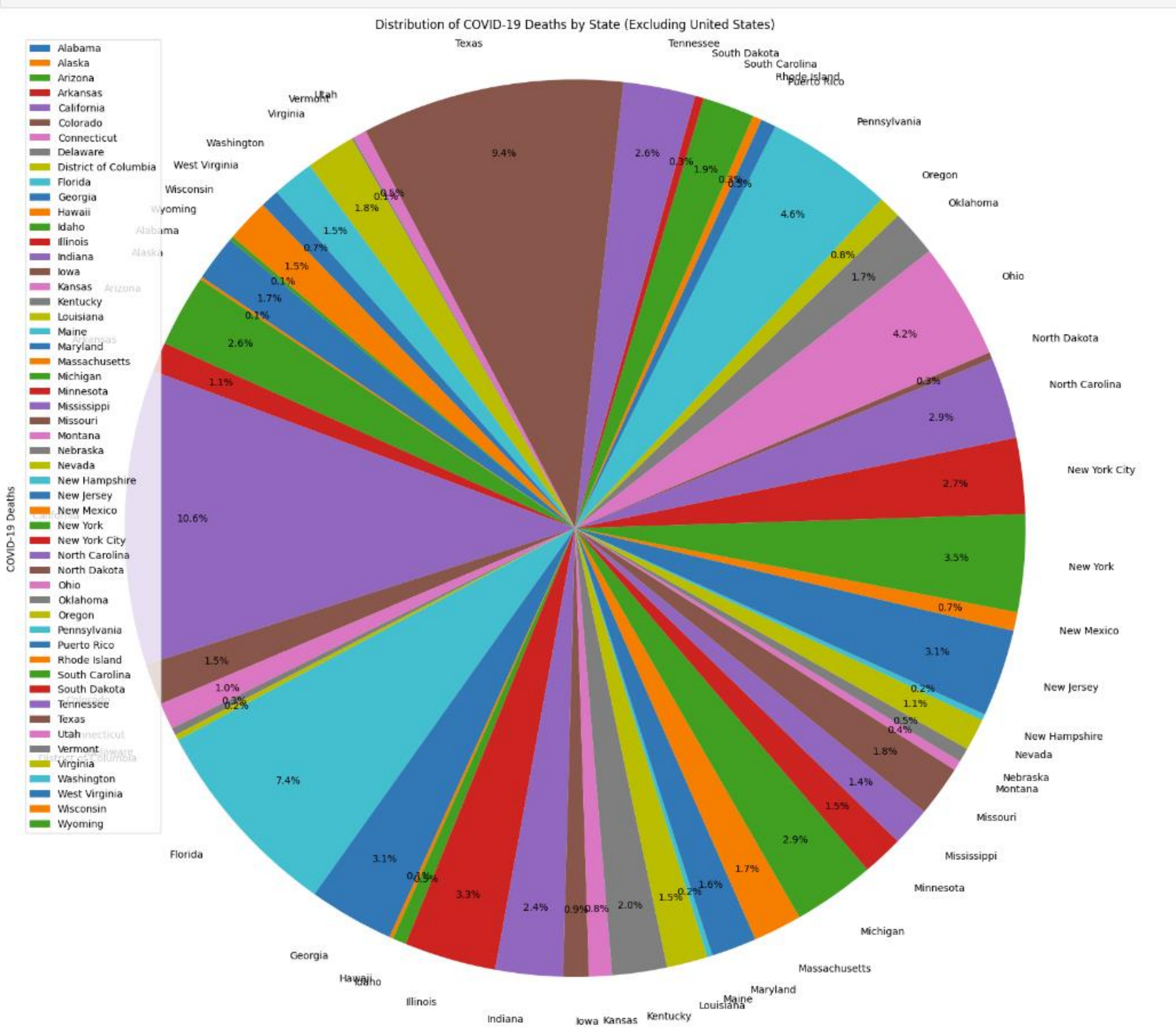
- Performed filtering on the Age group variable
- Age filtering for above age of 65

_meta	updated_at	updated_meta	meta	Data As Of	Start Date	...	Year	Month	State	Condition Group	Condition	ICD10_codes	Age Group	COVID-19 Deaths	Number of Mentions	Flag
None	1695825684	None	{ }	2023-09-24T00:00:00	2020-01-01T00:00:00	...	None	None	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	65-74	129005.0	133088	None
None	1695825684	None	{ }	2023-09-24T00:00:00	2020-01-01T00:00:00	...	None	None	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	85+	121119.0	123018	None
None	1695825684	None	{ }	2023-09-24T00:00:00	2020-01-01T00:00:00	...	None	None	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47	65-74	27920.0	29359	None
None	1695825684	None	{ }	2023-09-24T00:00:00	2020-01-01T00:00:00	...	None	None	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47	85+	27866.0	28796	None
None	1695825684	None	{ }	2023-09-24T00:00:00	2020-01-01T00:00:00	...	None	None	United States	Respiratory diseases	Adult respiratory distress syndrome	J80	65-74	30138.0	30138	None

Pie Chart & Bar Graph

- Pie chart and bar graph shows the distribution of Deaths by states.
- California, Texas, followed by Florida recorded most number of deaths. Where as Alaska , Vermont have least number of Covid 19 deaths.





Grouping Data

Counted Condition Group

```
Condition Group
All other conditions and causes (residual)                                27000
Alzheimer disease                                                         27000
COVID-19                                                                  27000
Circulatory diseases                                                     189000
Diabetes                                                                  27000
Intentional and unintentional injury, poisoning, and other adverse events 27000
Malignant neoplasms                                                       27000
Obesity                                                                   27000
Renal failure                                                             27000
Respiratory diseases                                                     162000
Sepsis                                                                    27000
Vascular and unspecified dementia                                         27000
dtype: int64
```

Number of Mentions Per Age Group

```
Age Group
0-24           1.556827
25-34          5.398871
35-44         15.637929
45-54         43.152662
55-64        107.697316
65-74        178.955195
75-84        203.862396
85+          196.574813
All Ages     622.197288
Not stated     0.005286
Name: Number of Mentions, dtype: float64
```

Number of Deaths by State

```
[14]: State
Alabama                   432004.0
Alaska                     25563.0
Arizona                   675395.0
Arkansas                  277702.0
California               2765450.0
Colorado                  388689.0
Connecticut               248369.0
Delaware                   69242.0
District of Columbia       52580.0
Florida                  1915568.0
Georgia                   797183.0
Hawaii                     35550.0
Idaho                     127624.0
Illinois                  857240.0
Indiana                   630686.0
Iowa                      238231.0
Kansas                    214046.0
Kentucky                  512253.0
Louisiana                 381519.0
Maine                      49061.0
Maryland                  423271.0
Massachusetts             454212.0
Michigan                  766221.0
Minnesota                 383921.0
Mississippi               365180.0
Missouri                  477690.0
Montana                    92001.0
Nebraska                  134276.0
Nevada                    297838.0
New Hampshire              59426.0
New Jersey                818458.0
New Mexico                174399.0
New York                  912528.0
New York City             713151.0
North Carolina            761941.0
North Dakota               65317.0
Ohio                     1090446.0
Oklahoma                  438130.0
Oregon                    213144.0
Pennsylvania             1190005.0
Puerto Rico               142259.0
Rhode Island               84425.0
South Carolina            483097.0
South Dakota               81556.0
Tennessee                 673816.0
Texas                    2453758.0
United States           26501616.0
Utah                      126856.0
Vermont                    18419.0
Virginia                  460649.0
Washington                390447.0
West Virginia             175517.0
Wisconsin                 402985.0
Wyoming                    36806.0
Name: COVID-19 Deaths, dtype: float64
```

- Grouping data by 'State' and calculating total COVID-19 deaths per state
- Grouping data by 'Condition Group' and counting the occurrences of each group
- Grouping data by 'Age Group' and calculating average number of mentions per age group

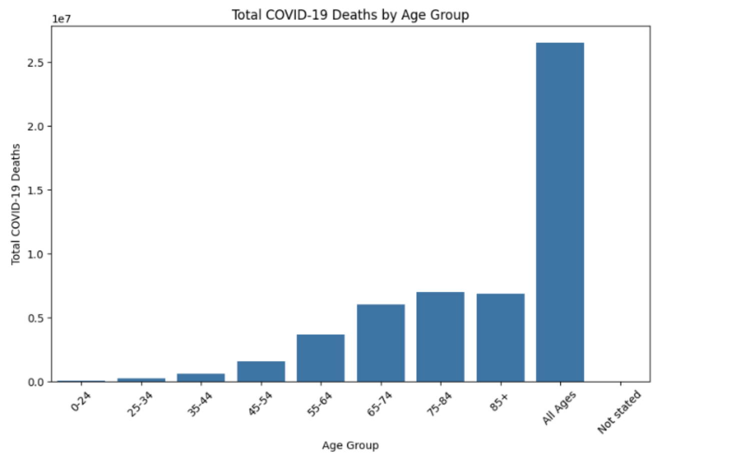
Analysis

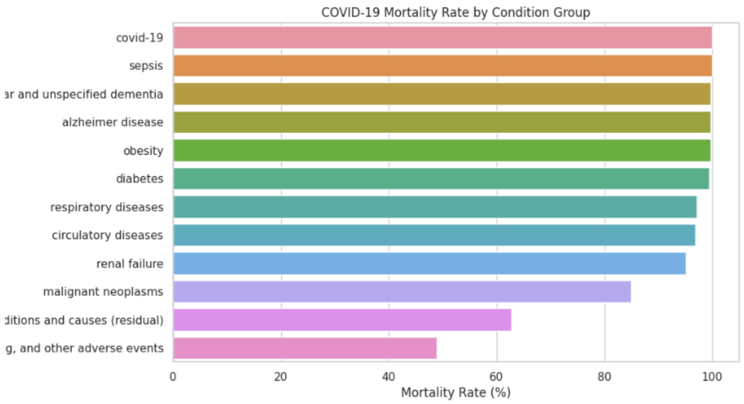
Bar Chart: Total COVID-19 Deaths by Age Group

- This bar chart illustrates the distribution of COVID-19 related deaths across different age groups.
- The data suggests that the impact of COVID-19 on mortality rates increases with age, with the highest number of deaths occurring in the oldest age bracket.

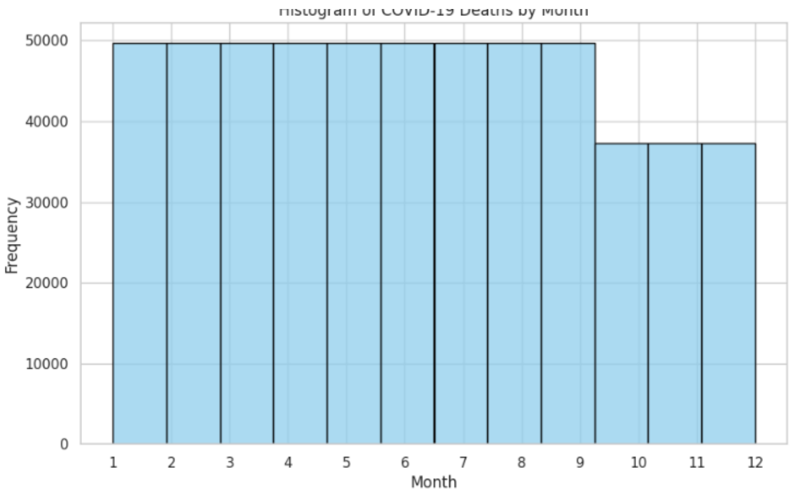
Stacked Bar Chart: COVID-19 Mortality Rate by Condition Group

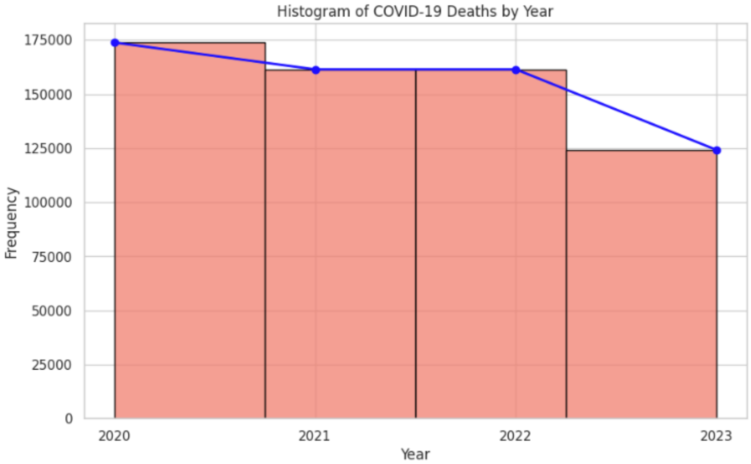
- The stacked bar chart presents the mortality rate of patients with COVID-19 in conjunction with various underlying health conditions or comorbidities.
- The length of each bar represents the percentage of the mortality rate attributed to each condition group, providing insight into which health issues are most commonly associated with fatal COVID-19 outcomes.
- This visualization emphasizes the significant risk factors, such as cardiovascular diseases and diabetes, contributing to COVID-19 mortality.





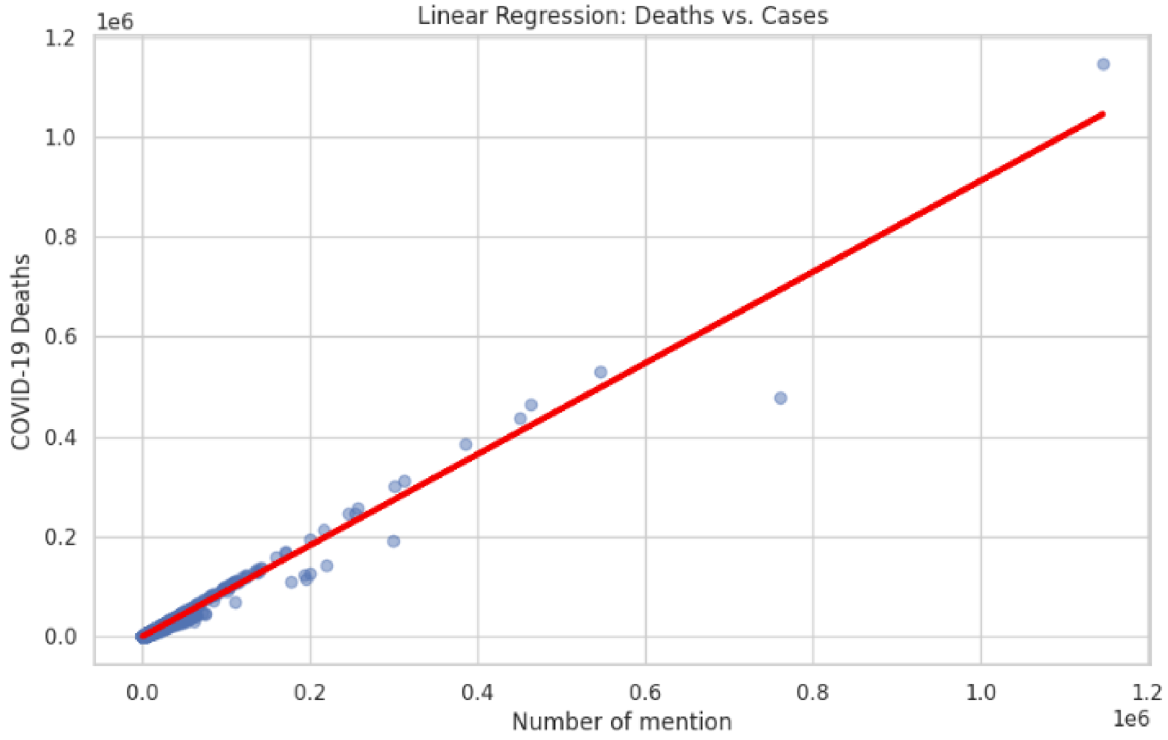
Histogram Analysis





- For the "Histogram of COVID-19 Deaths by Month," you can observe the distribution of COVID-19 deaths across different months. The x-axis represents the months, and the y-axis represents the frequency of deaths.
- For the "Histogram of COVID-19 Deaths by Year," the histogram gives an overview of the distribution of deaths across different years, while the line plot shows the trend or pattern of COVID-19 deaths over the years. The line plot connects the high points of the histogram, providing a visual representation of the variation in the number of deaths each year.

Linear Regression Analysis

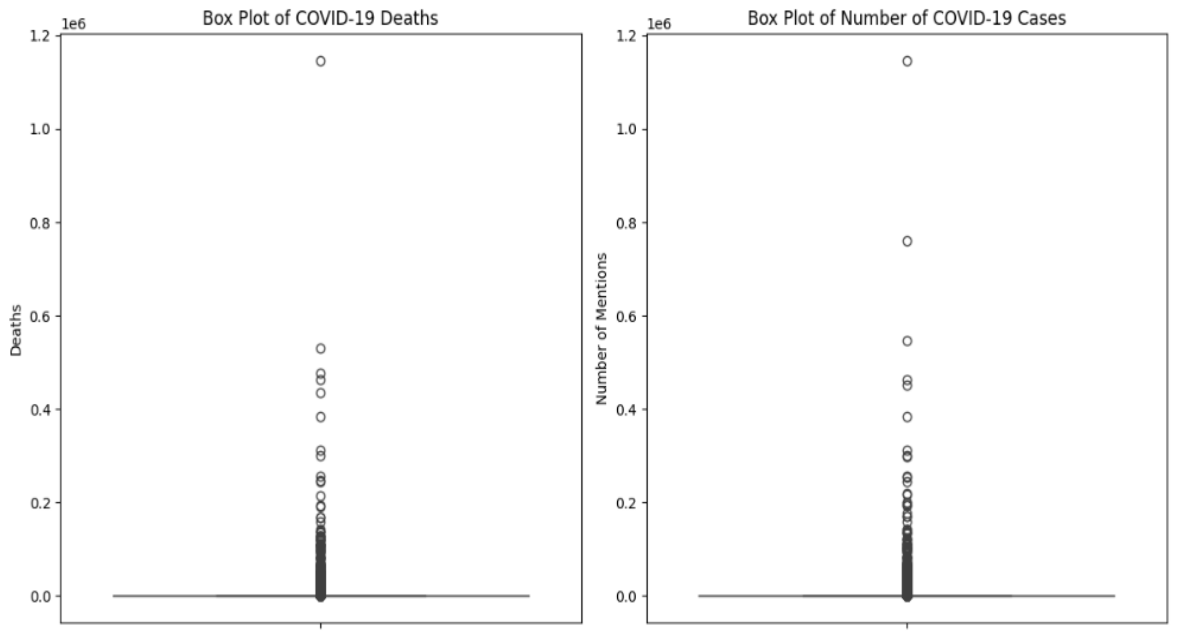


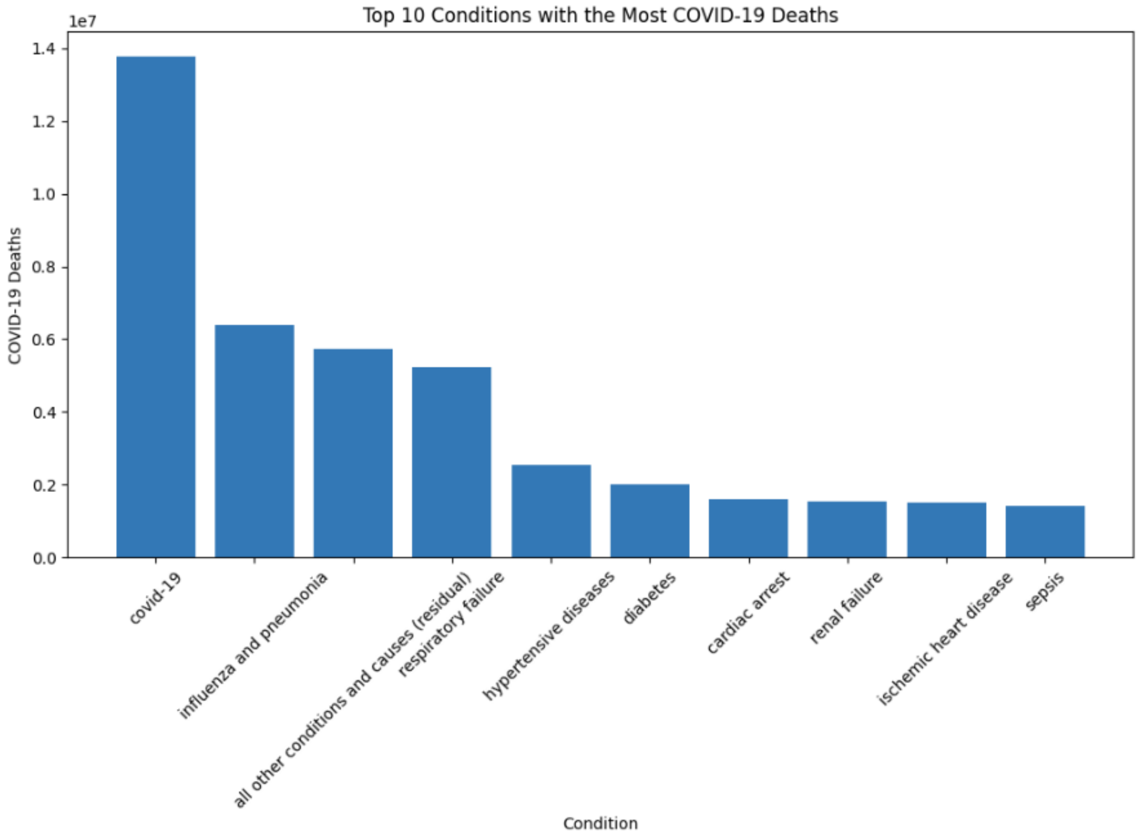

```
Intercept: 0.6028922878045648
Coefficient: 0.9117763436901557
```

- Predictive model that showcases how deaths and mentions are closely related to the data's average
- As deaths increase, the number of mention increases (and vice versa), creating a positive linear regression line

Outlier Analysis-Box Plot

- Both COVID-19 deaths and cases have a skewed distribution with most data points concentrated at the lower end.
- The median values are closer to the bottom of the data range, indicating a lower central tendency for both deaths and cases.
- There are numerous outliers for both deaths and cases, signifying instances of very high numbers that deviate from the typical values.
- The variance in the data is substantial, with the bulk of observations being low, but with some significantly high values.



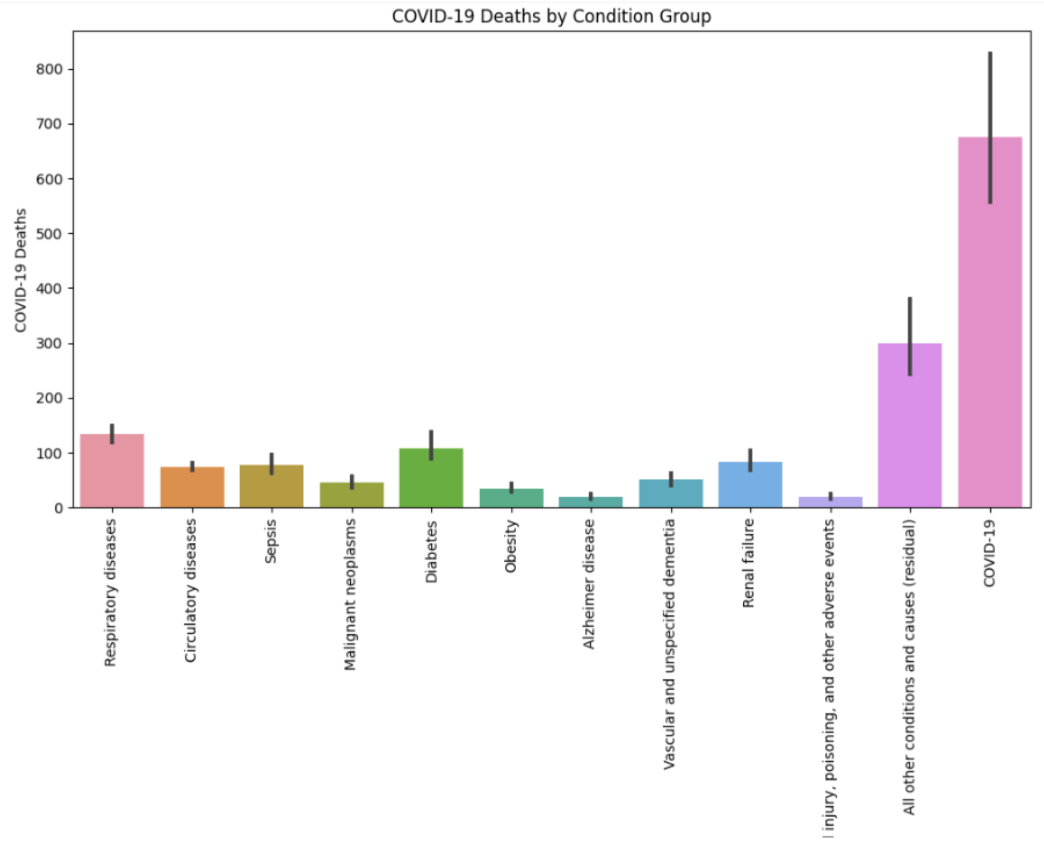


Hypothesis Testing & Frequency Distribution

- Comparing the COVID-19 deaths to condition variable in the dataset
- The ANOVA results show a high F-statistic of approximately 53.98 and a very low p-value (around 3.49×10−2373.49×10−237). This indicates a statistically significant difference in COVID-19 deaths across different condition groups in our dataset, allowing to reject the null hypothesis that there is no difference between group means.

- In Frequency Distribution , the bar chart displays the top 10 conditions associated with COVID-19 deaths, with "Pneumonia" accounting for the highest number of deaths, followed by and "Influenza".
- The decreasing order indicating the relative impact of each condition on COVID-19 mortality.

```
ANOVA F-statistic: 53.980576930153816
p-value: 3.498008065053823e-237
```



Plot of Deaths By Condition Group

- **Chart Analysis**: Demonstrates COVID-19 fatalities by comorbid conditions.
- **Insights**:
- 'COVID-19' is the predominant cause of recorded deaths.
- 'Alzheimer's' and 'dementia' are notable for their high impact.
- Significant fatalities also occurred in patients with 'Respiratory Diseases', 'Diabetes', and 'Heart Conditions'.

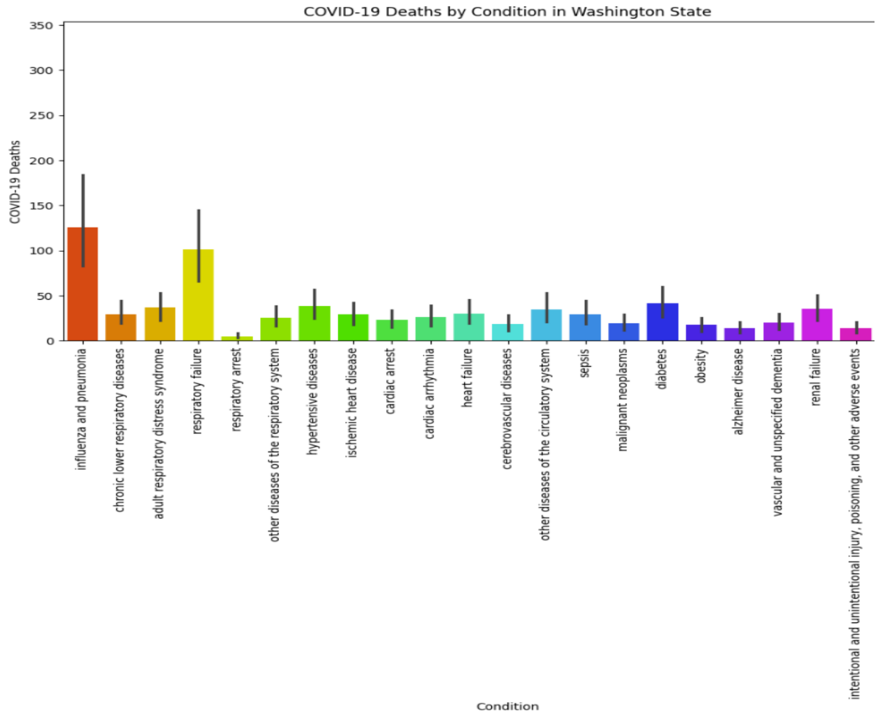
Analysis of Washington State

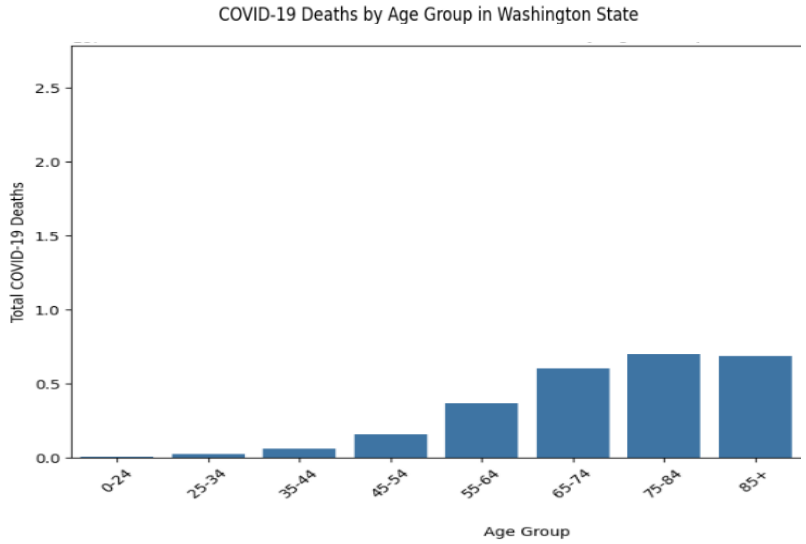
First Plot: COVID-19 Deaths by Condition in Washington State

- This bar chart provides an in-depth look at the comorbidities or underlying conditions associated with COVID-19 fatalities in Washington State.
- Each bar represents a different medical condition and is color-coded for clear differentiation.
- The plot highlights which conditions have been most frequently associated with death cases, with the tallest bars indicating the highest number of deaths.
- Notably, conditions such as 'pneumonia' and 'sepsis' show significantly higher mortality figures, suggesting their severe impact on COVID-19 patients.

Second Plot: COVID-19 Deaths by Age Group in Washington State

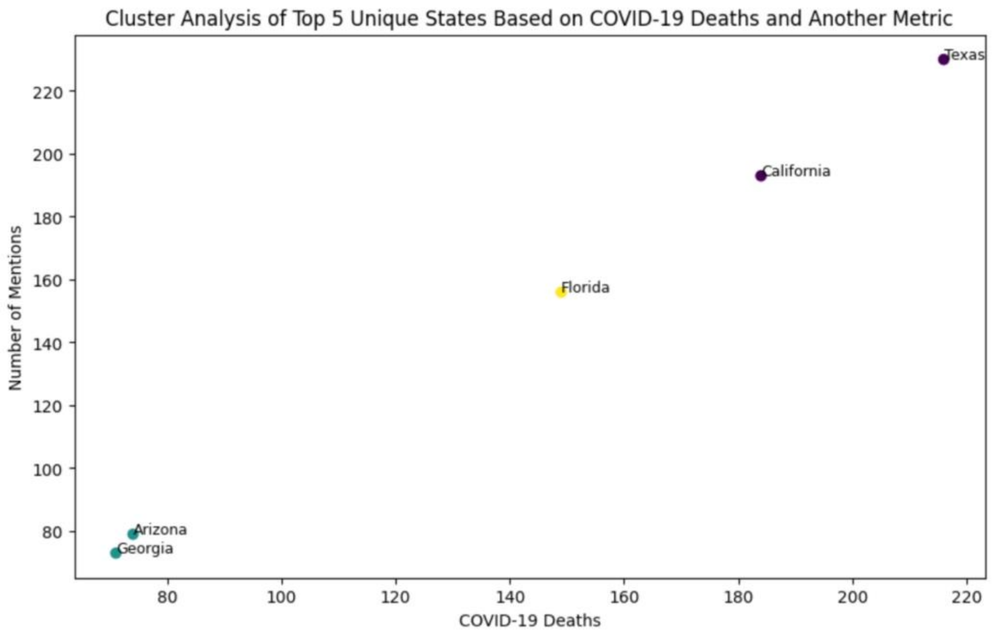
- This histogram breaks down the total number of COVID-19 deaths by age group within the state, offering a clear demographic perspective.
- A stark increase in deaths is observed among the older populations, with the highest number occurring in the '65+' age category.
- The distribution signifies the heightened vulnerability of the elderly to COVID-19, underscoring the need for targeted protective measures for senior citizens.

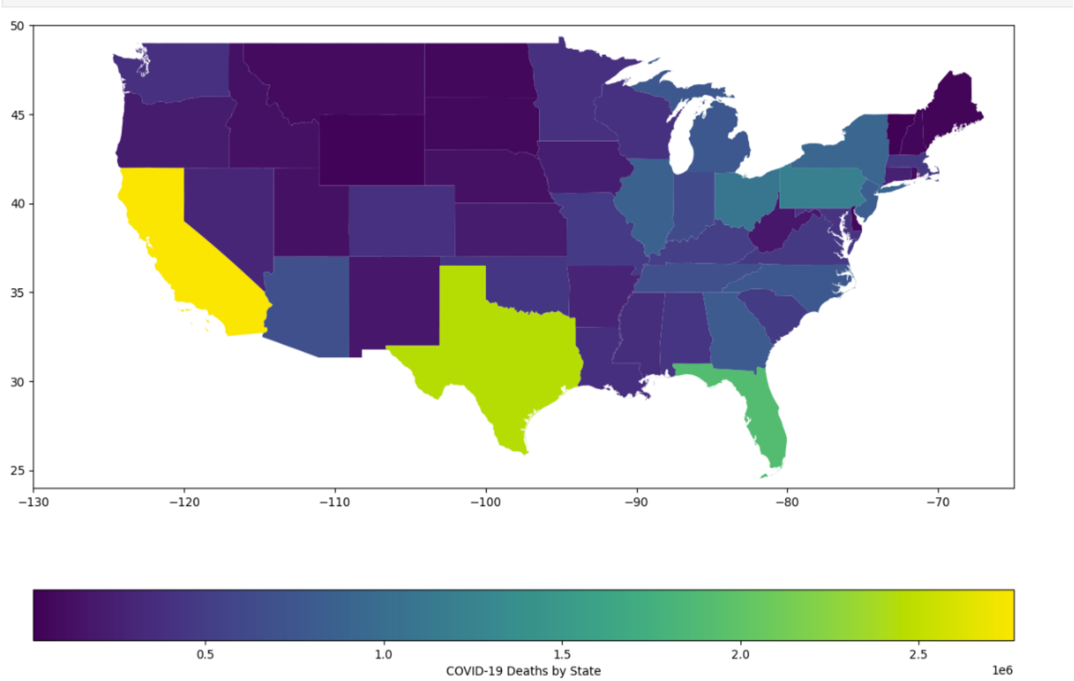




Cluster Analysis

- **Visual Overview**: The scatter plot displays a comparison of the top five states by COVID-19 deaths against the number of times COVID-19 is mentioned in official records.
- **Axes Interpretation**:
- **X-Axis**: Represents the count of COVID-19 deaths.
- **Y-Axis**: Denotes the number of mentions of COVID-19 in documents.
- **Insights**:
- **Texas & California**: High impact, with large numbers of deaths and mentions.
- **Florida**: Notable for a high number of mentions relative to deaths.
- **Georgia & Arizona**: Lower impact, with fewer deaths and mentions.





Heatmap

- **Purpose of the Heatmap**: This map visualizes the intensity of COVID-19 deaths across the United States, using color coding to represent varying levels of impact.
- **Color Interpretation**:
- Dark Purple: Lower numbers of reported deaths.
- Yellow-Green: Higher numbers of reported deaths.
- The scale at the bottom translates the color gradient into the actual number of deaths.
- **Insight**:
- States with darker shades show relatively fewer deaths, while those in yellow-green have reported higher death counts.
- The color gradient provides an at-a-glance understanding of geographical trends in COVID-19 mortality rates.
- **Usage**: Such a heatmap can quickly convey regional differences and identify areas with the most significant health impact from the pandemic, which can be crucial for targeted public health responses.

Challenges

Large Dataset: With over 600,000 entries, processing and analyzing the dataset can be computationally intensive and may require optimization techniques or more robust hardware.

Data Type Discrepancies: Columns like 'COVID-19 Deaths' and 'Number of Mentions' being in an object data type instead of numeric, necessitating type conversion and error handling.

Missing Values: Significant numbers of missing entries in crucial columns, such as 'COVID-19 Deaths' and 'Number of Mentions', require careful handling to avoid bias.

Geospatial Data Integration: Difficulties in integrating geospatial data due to format issues (e.g., shapefiles not being directly accessible or compatible).

Clarity of Visualizations: Ensuring that complex data is represented in a way that is understandable and visually clear to the audience.

Choice of Appropriate Visuals: Determining the most effective type of visualization for the data at hand, such as choosing between heatmaps, bar charts, or cluster diagrams.

Conclusion

- Our analysis revealed distinct temporal trends, highlighting periods of heightened mortality rates. Additionally, we observed intriguing geographic variations, emphasizing the importance of localized interventions and healthcare strategies.
- By showcasing the demographic factors, we identified specific populations that faced elevated risks. Understanding the impact of age, gender, and underlying health conditions is crucial for targeted public health efforts.
- Our use of Python's visualization capabilities not only facilitated a deeper understanding of the data but also made complex information accessible to a broader audience. Effective communication through visuals is crucial for informed decision-making.

Questions?

THANK YOU!