Data Source Summary

1. **Data source summary:**

   The source of the Data-Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023 - Catalog (data.gov)

   Data Type - External Data

   Owner Of The Data - CDC

   Trustworthiness of the Data - The CDC is a reputable government health organization, so it is highly trustworthy.

   Data Collection Method:

   Data Type – Administrative data

   Method of Collection – The Data is collected through NVSS(National Vital Statistics System)

   Time – It was updated as records were processed and verified but the dataset is no longer being updated as of September 27, 2023

   Overview of Data contents:

   Variables:-

   1. Data As Of: Date when the data was last updated.
   2. Start Date: Beginning date of the data collected/on period.
   3. End Date: Ending date of the data collected/on period.
   4. Group: Aggregation level (by total, by month, by year).
   5. Year: Year of death.
   6. Month: Month of death.
   7. State: U.S. state where the death occurred.
   8. Condition Group: Category of conditions contributing to death.
   9. Condition: Specific conditions contributing to death.
   10. ICD10_codes: International Classification of Diseases codes.
   11. Age Group: Age range of the deceased.
   12. COVID-19 Deaths: Number of deaths due to COVID-19.
   13. Number of Mentions: Number of times a condition is mentioned on death certificates.
   14. Flag: Indicates whether data in the row is suppressed for confidentiality.

   Why this dataset was chosen:

   I chose this dataset because it meets all project requirements such as allowing for advanced analytical techniques- regression, clustering, and geospatial analysis. The dataset is ethically sound, with no personal information such as name, address, or phone number.

2. **Data cleaning process:**

   Data Exploration:-

   Inspected the dataset to check the structure and basic information. And checked the numerical statistics and visualized distribution with histograms and box plots.

   Handling Missing Values:-

   Separated the data frame according to the “Group” column to continue working only with the data aggregated by month. And Imputed missing values in COVID-19 Deaths and Number of Mentions columns with random integers between 1 and 9 for rows with the suppression.

Filtering and Dropping Data:

Dropped the Group column as it only contained "By Month" and was no longer relevant to the analysis. Removed rows where Age Group was "All Ages" or "Not stated" as these represented aggregated or non-informative data. Filtered out rows where the State was ' United States' or 'Puerto Rico', keeping ' New York City' and ' District of Columbia'. New York State does not include the data from New York City.

Converted columns:-

Converted date columns (Data as Of, Start Date, End Date) to date/me format. And converted categorical columns to the 'category' type for improved memory efficiency.

Addressing Duplicates and Mixed-Type Data:-

Checked for and confirmed no duplicate rows. And ensured no mixed-type data within columns.

Checking and Handling Outliers:-

Defined and used a function to identify outliers using the IQR method for numerical columns (Year, Month, COVID-19 Deaths, and Number of Mentions).

Exporting Data:-

Conducted final checks on the cleaned data frame for structure, statistics of numerical columns, and unique values in categorical columns. And Exported the cleaned data frame to a CSV file for further analysis.

3. **Data Profile:**

Information regarding Cleaned data:-

There are 430,560 total entries.
There are 13 columns.
There are 6 categorical variables.
There are 4 Numerical variables.
There are 3 Date variables.

Column Description:-

Column Name	Type	Description	Range	Unique Value
Data As Of	Date	When the data was last updated		
Start Date	Date	The starting date for the data record		
End Date	Date	The ending date for the data record		
Year	Numerical(Float)	Year of the record	2020-2023	
Month	Numerical(Float)	Month of the record	1-12(Jan-Dec)	
State	Categorical	US State or Territory		52
Condition Group	Categorical	Broad group of conditions contributing to COVID-19 deaths		12
Condition	Categorical	Specific conditions contributing to COVID-19 deaths		23

ICD10_codes	Categorical	ICD-10 codes for the condition		23
Age Group	Categorical	The Age Group of the individuals		8
COVID-19 Deaths	Numerical(Float)	Number of deaths attributed to COVID-19		0-5,094
Number of Mentions	Numerical(Float)	The number of times the medical condition is mentioned on the death certificate		0-5,094
Flag	Categorical	If the data cells have counts between 1-9 & have been suppressed for confidentiality		1

Summary Statistics of Numerical Columns:

	Year	Month	Covid-19 Deaths	Number of Mentions
Count	430,560	430,560	430,560	430,560
Mean	2021.4	6.2	10.91	11.78
Std Dev	1.08	3.35	53.96	57.09
Min	2020	1	0	0
25%	2020	3	0	0
50%	2021	6	1	1
75%	2022	9	7	8
Max	2023	12	5,094	5,094

4. Limitations and Ethical Considerations:

Limitations:-

- Reporting delays can range from 1 week to 8 weeks or more, meaning the data for recent periods may be incomplete. However, data for 2020 and 2021 are based on final data.
- Different states may have varying standards for reporting COVID-19 deaths and contributing conditions, which can make comparisons across states less reliable.
- Deaths involving multiple conditions are counted in each relevant category, so numbers for different conditions should not be summed to avoid counting the same death multiple times.

Biases present in the dataset:-

- Measurement Bias:- Deaths with multiple conditions are counted in each category leading to measurement Bias.
- Selection Bias:- Certain demographic groups may be underrepresented, impacting the accuracy of analysis.

- Reporting Bias:- Varying state standards for reporting ccccovid-19 deaths and conditions leading to reporting Bias.

Ethical Considerations:-

- Privacy: The dataset does not contain personally identifiable information, ensuring privacy and compliance with data protection regulations.
- Sensitivity: The data is sensitive as it pertains to causes of death, necessitating careful handling to avoid misinterpretation, stigmatization of individuals with pre-existing conditions, and ensure fair representation of all demographic groups to prevent biased public health interventions.

.

5. Questions for the further Analysis:

1. What are the most common conditions contributing to COVID-19 deaths in different age groups?
2. Which states have the highest and lowest prevalence of specific conditions contributing to COVID-19 deaths?
3. Are there any notable seasonal patterns or trends in COVID-19 deaths or in the prevalence of specific conditions contributing to COVID-19 deaths?
4. What factors are most predictive of COVID-19 death rates?