Achievement 6 – Task 6.1

DATA SOURCE

1. Dataset: "Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023"

a. Data Source Summary:

- **Source:**https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group
- **Data Type:** External data
- **Owner:** Centers for Disease Control and Prevention (CDC)
- **Reliability:** The data is highly reliable, given that it is provided by the CDC, a well-respected government health agency.

b. Data Collection Method:

- **Data Type:** Administrative data
- **Collection Process:** Data was automatically gathered through the National Vital Statistics System (NVSS) from death certificates submitted to vital registration offices across the United States.
- **Time Lag:** The dataset experienced a time lag due to its provisional status. Updates were made as records were processed and verified, but no further updates have been made since September 27, 2023.

c. Overview of Data Contents:

i. Variables Included:

1. **Data As Of:** The date when the data was last updated.
2. **Start Date:** The beginning of the data collection period.
3. **End Date:** The end of the data collection period.
4. **Group:** The level of aggregation (e.g., total, by month, by year).
5. **Year:** The year of death.
6. **Month:** The month of death.
7. **State:** The U.S. state where the death occurred.
8. **Condition Group:** The category of conditions contributing to death.
9. **Condition:** The specific conditions that contributed to death.
10. **ICD10 Codes:** The International Classification of Diseases codes.
11. **Age Group:** The age range of the deceased.
12. **COVID-19 Deaths:** The number of deaths attributed to COVID-19.
13. **Number of Mentions:** The number of times a condition is mentioned on death certificates.
14. **Flag:** Indicates if data in the row is suppressed for confidentiality reasons.

d. Why this dataset was chosen

Working with a COVID-19 dataset is a strategic choice that leverages my background in environmental science and my experience as a food safety coordinator. The pandemic's intersection with public health and environmental factors provides a unique opportunity for me to apply my expertise in understanding how variables like air quality, climate, and safety protocols impact virus transmission and outcomes. By analysing this data, I can

contribute valuable insights that inform public health strategies while also enhancing my skills in data analysis, particularly with large and complex datasets. This experience not only strengthens my analytical capabilities but also diversifies my portfolio, positioning me for broader career opportunities in public health, epidemiology, and environmental health.

Data Profile

1. Data Cleaning Process

a. Initial Data Exploration:

- The dataset was inspected to understand its structure and basic information.
- Numerical statistics were reviewed, and distributions were visualized using histograms and box plots.

b. Handling Missing Values:

- Missing values were identified.
- The DataFrame was divided based on the “Group” column to focus on data aggregated by month. iii. Missing values in the COVID-19 Deaths and Number of Mentions columns were imputed with random integers between 1 and 9 for rows with suppression.

c. 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 State was "United States" or "Puerto Rico", keeping “New York City” and “District of Columbia”. New York state data does not include the data from New York City.

d. Data Type Conversion:

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

e. Addressing Duplicates and Mixed-type Data:

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

f. 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).
- Outliers identified and retained as the values appeared reasonable in context.

g. Final Checks and Exporting Data:

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

2. Data Profile of final cleaned dataset:

a. General Information

- Total Entries: 430,560
- Columns: 13
- Categorical Variables: 6
- Numerical Variables: 4
- Date Variables: 3

b. Column Descriptions

i. Data As Of

- Type: Date
- Description: The date when the data was last updated.
- Missing Values: 0

ii. Start Date

- Type: Date
- Description: The star/ng date for the data record.
- Missing Values: 0

iii. End Date

- Type: Date
- Description: The ending date for the data record.
- Missing Values: 0

iv. Year

- Type: Numerical (Float)
- Description: Year of the record.
- Range: 2020 - 2023
- Missing Values: 0

v. Month

- Type: Numerical (Float)
- Description: Month of the record.
- Range: 1 (January) - 12 (December)
- Missing Values: 0

vi. State

- Type: Categorical
- Description: U.S. state or territory.
- Unique Values: 52
- Missing Values: 0

vii. Condition Group

- Type: Categorical
- Description: Broad group of conditions contributing to COVID-19 deaths.
- Unique Values: 12
- Missing Values: 0

viii. Condition

- Type: Categorical

- Description: Specific condition contributing to COVID-19 deaths.
- Unique Values: 23
- Missing Values: 0

ix. ICD10_codes

- Type: Categorical
- Description: ICD-10 codes for the condition.
- Unique Values: 23
- Missing Values: 0

x. Age Group

- Type: Categorical
- Description: Age group of the individuals.
- Unique Values: 8 (0-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+)
- Missing Values: 0

xi. COVID-19 Deaths

- Type: Numerical (Float)
- 2. Description: Number of deaths attributed to COVID-19.
- Range: 0 - 5,094
- Missing Values: 0

xii. Number of Mentions

- Type: Numerical (Float)
- Description: Number of /times the medical condition is mentioned on the death certificate.
- Range: 0 - 5,094
- Missing Values: 0

xiii. Flag

- Type: Categorical
- Description: Indicates if the data cells have counts between 1-9 and have been suppressed for confidentiality.
- Unique Values: 1
- Missing Values: 279,731 representing unsuppressed counts.

c. Summary Statistics

Summary Statistics	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	5094	5094

2. Limitations and Ethical Considerations

a. Limitations:

i. Provisional Nature of Data: The data is considered provisional, and any conclusions drawn from it may need revision once finalized data becomes available.

ii. Reporting Delays: Reporting delays, which can range from 1 to 8 weeks or more, may result in incomplete data for recent periods. However, the data for 2020 and 2021 is based on finalized information.

iii. Inconsistent Reporting Standards: The standards for reporting COVID-19 deaths and contributing conditions vary by state, which can reduce the reliability of cross-state comparisons.

iv. Data Suppression: To protect privacy, counts ranging from 1 to 9 are suppressed.

v. Multiple Conditions: On average, four additional conditions are associated with each death, complicating the analysis.

vi. Double Counting Risk: Deaths involving multiple conditions are counted in each relevant category, so summing the numbers across different conditions could result in counting the same death multiple times.

b. Potential Biases in the Dataset

i. Reporting Bias

1. Inconsistent Standards: Variability in state standards for reporting COVID-19 deaths and conditions.

2. Data Suppression: Counts are suppressed for confidentiality purposes.

ii. Selection Bias

1. Demographic Disparities: Certain demographic groups may be underrepresented, affecting the accuracy of analyses.

2. Geographic Variations: Differences in reporting between urban and rural areas can introduce geographic biases.

iii. Measurement Bias

1. Multiple Conditions Reporting: Deaths associated with multiple conditions are reported in each relevant category, which may lead to an overestimation of condition prevalence.

2. Non-Summation Rule: Summing conditions across different categories should be avoided to prevent inaccuracies.

c. Ethical Considerations

i. **Privacy**: The dataset is devoid of personally identifiable information, ensuring privacy and compliance with data protection regulations. However, there remains a potential risk of re-identification, particularly in smaller populations or when the data is combined with other datasets.

ii. **Sensitivity**: Given that the data relates to causes of death, it is sensitive and requires careful handling to avoid misinterpretation, stigmatization of individuals with preexisting conditions, and to ensure fair representation of all demographic groups, thereby preventing biased public health interventions.

Questions to Explore with the Analysis:

a. **Demographic Analysis:**

i. Which conditions are most commonly contributing to COVID-19 deaths across different age groups?

ii. How does the distribution of conditions contributing to COVID-19 deaths vary by age group?

b. **Geographical Analysis:**

i. Which states exhibit the highest and lowest prevalence of specific conditions contributing to COVID-19 deaths?

ii. Are there regional patterns in the types of conditions associated with COVID-19 deaths across the United States?

c. **Temporal Analysis:**

i. How have COVID-19 death rates and contributing conditions evolved over time?

ii. Are there any significant seasonal patterns or trends in COVID-19 deaths or in the prevalence of specific conditions contributing to these deaths?

d. **Predictive Modeling:**

i. Can a predictive model be developed to identify high-risk populations based on the presence of certain conditions, demographics, and geographic data?

ii. Which factors are most predictive of COVID-19 death rates?