COVID-19 Data Analysis Project Documentation

Group Number: 4
Team Mates:
Sakshi Patel
Trisha Chandrashekar
Ravya
Sai Surya Teja Reddy Bommepalli

Enrichment data sets taken by individual:

Name	Data Set Name
Sakshi Patel	Census Demographic ACS
Trisha Chandrashekar	Presidential Election Results
Ravya Vangaveti	Employment Dataset
Sai Surya Teja Reddy Bommepalli	ACS Social, Economic, and Housing

Project Overview

This project aims to analyze COVID-19 cases and deaths at the county level in the United States. By integrating datasets related to county populations, COVID-19 confirmed cases, and deaths, the analysis attempts to uncover insights into how the virus spread and affected different counties. The dataset merges these three key variables and serves as a foundation for further analysis.

Datasets Used

1. **COVID-19 County Population Dataset**
   - Provides population data for U.S. counties, essential for calculating case and death rates per 100,000 people.
2. **COVID-19 Confirmed Cases Dataset**
   - Tracks daily confirmed cases of COVID-19 at the county level.
3. **COVID-19 Deaths Dataset**
   - Tracks daily COVID-19-related deaths at the county level.

Steps in Data Processing

1. **Uploading and Displaying Datasets**
   - Imported and displayed the COVID-19 population, confirmed cases, and deaths datasets using pandas. Each dataset is structured at the county level and contains key information such as county name, state, and cases or deaths.
2. **Merging the Datasets**

    - Merged the three datasets (population, confirmed cases, and deaths) using the county FIPS code and the state as the keys. The final dataframe is referred to as the "super COVID-19 data frame."
3. **Final Super COVID-19 Dataframe**
    - The merged dataframe includes:
        - County FIPS code
        - County name
        - State
        - Population
        - Daily confirmed cases
        - Daily deaths

1. COVID-19 County Population Dataset

This dataset typically includes:

- **FIPS Code (County Identifier):** This is a unique numeric identifier for each county. It can be represented as an **integer** or **string** depending on formatting needs.
    - **Data Type:** Integer or String
- **County Name:** The name of the county.
    - **Data Type:** String
- **State Name/Code:** The name or two-letter code of the state.
    - **Data Type:** String
- **Population:** The total population of the county.
    - **Data Type:** Integer

2. COVID-19 Confirmed Cases Dataset

This dataset tracks the number of confirmed COVID-19 cases:

- **FIPS Code (County Identifier):** Numeric code for each county.
    - **Data Type:** Integer or String
- **County Name:** The name of the county.
    - **Data Type:** String
- **State Name/Code:** The name or two-letter code of the state.
    - **Data Type:** String
- **Date:** The date when the cases were recorded.
    - **Data Type:** Date or String (Date format: YYYY-MM-DD)
- **Confirmed Cases:** The number of confirmed COVID-19 cases.
    - **Data Type:** Integer

3. COVID-19 Deaths Dataset

This dataset tracks the number of deaths caused by COVID-19:

- **FIPS Code (County Identifier):** Numeric code for each county.
    - **Data Type:** Integer or String
- **County Name:** The name of the county.
    - **Data Type:** String
- **State Name/Code:** The name or two-letter code of the state.
    - **Data Type:** String

- **Date:** The date when deaths were recorded.
    - **Data Type:** Date or String (Date format: YYYY-MM-DD)
- **Deaths:** The number of COVID-19-related deaths.
    - **Data Type:** Integer

Data Types Summary:

- **Integer:** Used for numeric values like FIPS codes, population counts, confirmed cases, deaths, and other counts (e.g., housing units).
- **String:** Used for text data like county names, state names, and date strings if not in datetime format.
- **Float:** Used for numeric values that require decimals, such as rates (e.g., unemployment rates) or averages (e.g., median income).
- **Date:** This can be represented as a **string** or **datetime** object (in Python or pandas) to store dates.

Combining Data

We combined the **COVID-19 datasets** (cases, deaths, population) with the **FIPS code** as the primary key, since it is the common identifier for each county across all datasets.