“The Americas”
(Projects with relation to North-, Middle- or South-America)

Project Context:

The project explores key issues related to North, Central, and South America. This specific analysis focuses on trends within the United States, particularly examining the relationship between population growth and unemployment rates over recent years.

Main Question:

Is the unemployment rate in the USA growing at a faster rate than the population from 2020 to 2023? If so, what factors might be contributing to this trend?

Data Sources:

- **Data Source 1**: U.S. Census Bureau Population Data (2020-2023)
  - **Metadata URL**: Census Bureau Metadata
  - **Data URL**: Population Data CSV
  - **Description**: This dataset provides annual population estimates for U.S. states, capturing factors like births, deaths, and migration to understand population trends.
- **Data Source 2**: Department of Labor and Iowa State University Unemployment Data (2020-2023)
  - **Metadata URL**: Iowa State Metadata
  - **Data URL**: Unemployment Data XLSX
  - **Description**: This dataset contains annual unemployment rates for U.S. states, helping analyze changes in the job market over time.

Data Pipeline:

The data pipeline was implemented using Python, Pandas, SQLite, and Matplotlib to automate the process of extracting, transforming, and loading data. The following steps and challenges were encountered during the project:

Technologies Used:

- **Python** for scripting and data processing.
- **Pandas** for data manipulation and cleaning.
- **SQLite** for storing cleaned data.
- **Matplotlib** for visualizing results.
- **Jupyter Notebook** for documenting and running analyses interactively.

Overview of the Pipeline:

1. **Data Extraction**:
   - Loaded data from two sources:
     - A **CSV file** containing population estimates from the U.S. Census Bureau.
     - An **Excel file** containing unemployment rates from the Department of Labor.

2. **Data Cleaning and Transformation**:
    - **Merging Issues**:
        - The Excel file had merged cells, making it challenging to extract structured data.
        - Different formats between the CSV and Excel files required custom parsing logic.
    - **Column Selection and Renaming**:
        - Extracted only the necessary columns (e.g., population estimates for specific years and unemployment rates).
        - Renamed columns to consistent names (state, 2020, 2023, rate_2020, rate_2023) to simplify downstream analysis.
    - **Validation and Data Integrity Checks**:
        - Ensured that extracted columns matched the expected schema to avoid data inconsistencies.
        - Implemented validation to drop rows with missing or malformed data.
        - Verified that states were correctly aligned between the two datasets before merging.

3. **Data Integration**:
    - Merged the cleaned datasets on the state column to create a unified dataset for analysis.
    - Calculated new metrics such as **population growth** and **changes in unemployment rates** between 2020 and 2023.

4. **Data Storage**:
    - The cleaned and transformed data was stored in an **SQLite database** to enable efficient querying and data retrieval.
    - The database tables created were:
        - population_usa_2020_2023 for population estimates.
        - unemployment_rates_usa_2020_2023 for unemployment rates.

Key Challenges and Solutions:

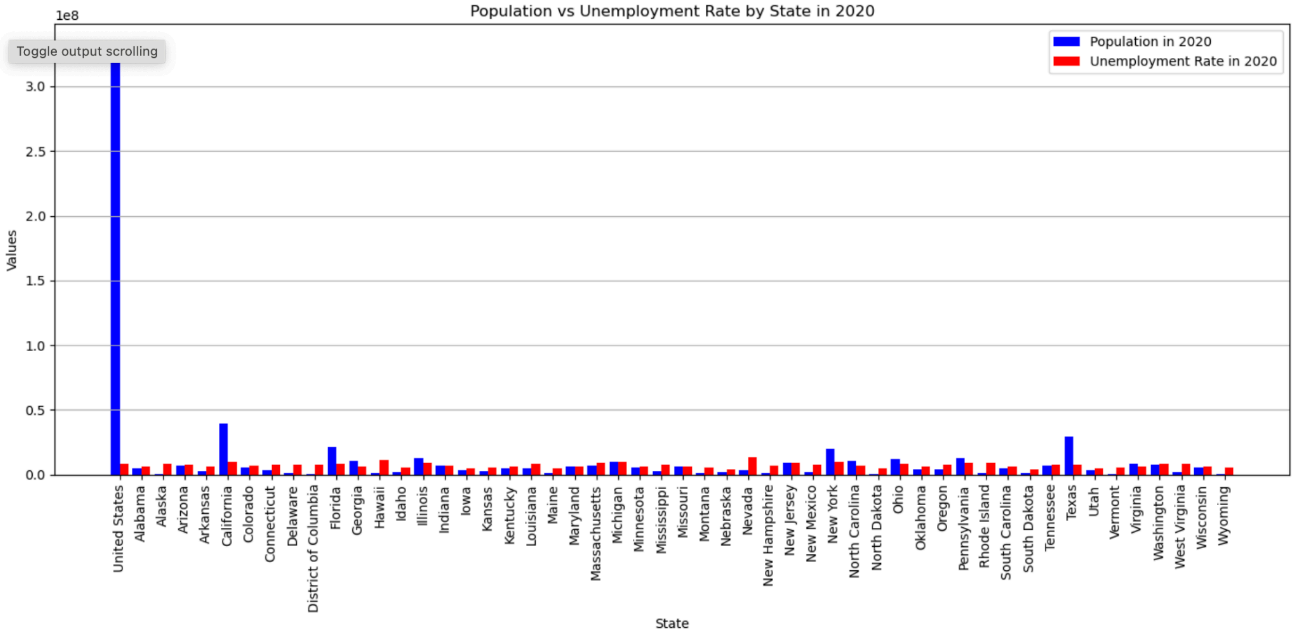
1. **Different File Formats (CSV and XLSX)**:
    - Handling the differences between CSV and Excel files was challenging, especially with the merged cells in the Excel file.
    - Solution: Utilized Pandas with the openpyxl engine to dynamically detect and extract the relevant rows and columns.

2. **Column Selection and Renaming**:
    - Extracting specific columns from both datasets and renaming them was necessary for consistency.
    - Solution: Implemented custom column selection and renaming functions to standardize the data before merging.

3. **Data Validation and Quality Assurance**:
    - Ensuring the integrity of the data was crucial, especially when merging datasets with different structures.
    - Solution: Performed checks to validate that columns contained the expected data types and values. Rows with missing or invalid data were dropped to prevent errors during analysis.
4. **Handling Missing Values**:
    - The datasets had missing entries, especially in the unemployment rates for certain states.
    - Solution: Implemented strategies to handle missing values, including dropping rows or filling with default values where appropriate.

Data Analysis:

- **Population vs. Unemployment Trends**:
  - Analyzed whether the unemployment rate increased at a faster rate than population growth from 2020 to 2023.
- **Key Findings**:
  - Potentially identified states where unemployment rates increased significantly despite stable population growth.
- **Visualizations**:
  - Included dual bar charts, scatter plots, and waterfall charts for visual comparison.



Results and Limitations:

- **Output Format**:
  - Data stored in an SQLite database and visualized using Matplotlib.
- **Data Structure**:
  - Cleaned tables: population_usa_2020_2023, unemployment_rates_usa_2020_2023.
- **Limitations**:
  - The analysis relies on the structure of the input datasets; changes in data formats may require pipeline adjustments.
  - Limited by the availability and granularity of data on state-level unemployment rates.
  - Differences in data collection methodologies between sources can lead to inconsistencies, especially when comparing population data from the Census Bureau to unemployment data from state labor departments.
  - The visualizations focus on aggregate data, which might miss nuances like age, income, or sector-specific unemployment rates.
  - Although the unemployment rate data is sourced from reputable institutions, it is challenging to verify the precise methodologies and criteria used to calculate these rates.