A Data-Driven Investigation of Crime and Arrest Disparities

Author: Ramisa Bhuiyan Raka (23172960)

Project Question: Which neighborhoods or regions have the highest disparities between crime reports and arrests?

Objective:

This project will focus on the locations and types of crime where crime is reported often but arrests are few hence possible under policing or lack of adequate resources. The eventual use of the maps is to help the policymaker to gain informed insights of disparities that exist in the polices implemented by the police.

Data Sources

Description of Data Sources

Crime Data

- **URL:** Crime Data
- **Content:** This dataset contains records of reported crimes in Los Angeles, including descriptions, locations, dates, and areas.

Arrest Data

- **URL:** Arrest Data
- **Content:** This dataset includes records of arrests made in Los Angeles, providing details like arrest types, charges, locations, and dates.

Data Structure and Quality

Crime Data

- **Columns:** Crime description, report date, victim's gender, location, area name, and a unique identifier.
- **Initial Observations:** Contains missing values in critical columns like crime description and area name.

Arrest Data

- **Columns:** Charge description, arrest date, arrest type, location, area name, and a unique report ID.
- **Initial Observations:** Some rows have missing data in charge descriptions or area names.

License Information

Both datasets are published on the Los Angeles Open Data Portal. The license details are available on the respective dataset pages:

- **Crime Data** Open licensed CC0 1.0 Universal
  - **Obligation:** Any use of the crime data must include proper attribution to the Los Angeles Open Data Portal as the source.
  - **Link:** Link
- **Arrest Data** Open licensed CC0 1.0 Universal
  - **Obligation:** Similar to the crime data, proper attribution to the Los Angeles Open Data Portal is required.
  - **Link:** Link

Data Pipeline

Overview

The pipeline is designed to:

- Extract raw data from the provided URLs.
- Transform the data by selecting relevant columns, cleaning missing or invalid entries, and storing the results in a SQLite database.
- Load the cleaned data into two separate SQLite tables for analysis.

Steps in the Pipeline

1. Extraction: Data is fetched directly from the source URLs using Python's pandas library.

2. Transformation:

Crime Data

- **Selected columns:** Crm Cd Desc, Date Rptd, Vict Sex, LOCATION, AREA NAME, DR_NO.
- Rows with missing values in critical fields (Crm Cd Desc, Date Rptd, AREA NAME) are dropped.

Arrest Data

- **Selected columns:** Charge, Arrest Date, Arrest Type Code, Charge Description, Location, Area Name, Report ID.
- Rows with missing values in critical fields (Charge Description, Arrest Date, Area Name) are dropped.

3. Loading:

- Cleaned data is written into two SQLite databases: **crime_data.db** and **arrest_data.db**.

Challenges and Solutions

- **Challenge:** Missing values in critical fields reduced data completeness.
- **Solution:** Rows with missing values were removed to ensure reliability.
- **Challenge:** Large datasets may cause performance issues.
- **Solution:** Optimized data selection and processing using pandas and SQLite.

Error Handling and Adaptability

- The pipeline validates URLs before downloading data to handle network errors.
- Database operations ensure existing data is replaced without duplication.

Results and Limitations

Results

- **Crime Data Output:** A cleaned dataset with columns relevant to understanding reported crimes.
- **Arrest Data Output:** A cleaned dataset focusing on arrests and their descriptions.
- **Output Format:** Data is stored in two SQLite tables:

  - crime_data: Contains cleaned crime data.

  - arrest_data: Contains cleaned arrest data.

Data Structure and Quality

The output data consists of two cleaned SQLite tables: `crime_data` with fields like crime description, date, and location, and `arrest_data` with arrest-related fields. Rows with missing critical values were removed to ensure completeness, and fields were standardized for consistency and usability. However, the data's accuracy depends on the source, and potential biases or geographic limitations must be considered.

Limitations and Potential Issues

- **Representativeness:** The datasets are limited to reported crimes and arrests in Los Angeles. They may not reflect unreported crimes or systemic biases in reporting.
- **Completeness:** Data with missing fields was removed, potentially discarding valuable but incomplete records.
- **Accuracy:** Any errors in the source datasets (e.g., incorrect crime descriptions or misreported locations) could propagate into the analysis.
- **Timeliness:** The data reflects the state of crime and arrests at the time of extraction. Analysis may require frequent updates for ongoing relevance.