Data Report

Audience

Policymakers and government officials, who can develop crime prevention policies and crime-fighting instruments based on differences in the number of crimes committed in each city, as well as on social factors.

The media, residents of California cities, and people who want to travel to California, who can use this project to learn about crime in California cities, and to increase their safety awareness and knowledge of crime.

California law enforcement agencies and police departments can use this project to optimize the number of law enforcement officers deployed in each city, thereby ensuring that adequate law enforcement resources are available to keep residents safe in high risk areas.

Content

Question

1. What are the most common types of crime in each city in California, and is there one type of crime that is most common in all cities?
2. What type of crime is more numerous in California?
3. Do all California cities have the same ratio of the number of law enforcement officers to the number of crimes in that city? Or is the ratio greater in certain cities?
4. Do California cities with higher median household incomes have lower crime rates?
5. Do cities with high poverty rates have higher crime rates?
6. Do cities with higher high school graduation rates have lower crime rates?

Data Sources

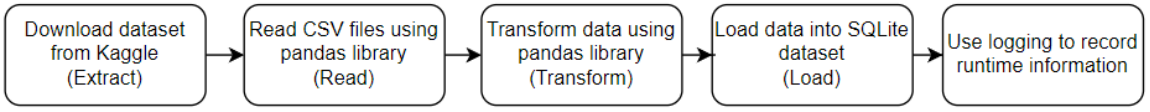
1. California Crime and Law Enforcement
    - This dataset contains crime and the number of law enforcement officers in each California city in 2015. It is published by the FBI and is listed on data.gov, the official website of the us government, which shows the dataset is under the US-PD license. Questions one through three can be answered using this dataset.
    - This dataset is a work that has a US-PD license, which is not protected by copyright, and the public is free to copy, distribute, modify, or use the work as they wish. So this project can perform cleaning and transformation operations on this dataset, of course, these operations would not destroy the accuracy of the data.

- The data in the dataset are all structured data in csv format. The dataset was released by the FBI, so it is accurate and highly related to the questions posed by the project. However, due to limited resources, the dataset is not up-to-date, has a null column and the data format is inconsistent, requiring data cleaning and conversion.

2. Fatal Police Shootings in the US
    - This dataset contains US census data on poverty rate, high school graduation rate, median household income in 2015. It was released by the U.S. Census Bureau and was organized by Karolina Wullum. Programmers can access this dataset on kaggle. The data in this dataset is authoritative, complements and extends another dataset of data, and can be used to answer questions four to six.
    - According to the dataset's release page on Kaggle, the dataset is licensed under the CC BY-NC-SA 4.0 license. Users are free to share and adapt this data as long as they give the correct attribution and indicate if the data has been changed, with no additional restrictions. This project follows the obligations of that license, attributes the author who is Karolina Wullum, and indicates that the data have been changed due to cleaning and transformation.
    - The data in the dataset are all structured data in csv format. The dataset was released by the U.S. Census Bureau, thus it is accurate, complete and closely related to the questions raised by the project. However, the dataset is not up-to-date and there are inconsistencies in the format of the data that require data cleaning and transformation.

Data Pipeline

1. Data pipeline description: This project follows the ETL process by first calling the Kaggle API to download the dataset from Kaggle (Extract), then using the pandas library to read and process the data from the CSV file (Transform), and finally using the sqlite3 library to load the processed data into the SQLite database (Load). At the same time, the logging operation will record the runtime information of the data pipeline.



2. Transformation or cleaning steps:
    - For the California Crime and Law Enforcement dataset, this project removed unnecessary columns, removed special characters from column names, merged tables based on city names, and finally converted the format of the data for each column.
    - For the Fatal Police Shootings in the US dataset, this project removed unnecessary columns, removed data unrelated to California cities, replaced outliers with uniform values, merged tables based on city names, and finally converted the format of the data for each column.

    - These operations are designed to ensure that the data are consistent, accurate, and that the data can answer the questions posed by the project, thus facilitating subsequent data analysis and modeling.
3. Problems and solutions:
    - The dataset on California's economy and education could not be found at the beginning of the program as of 2015, this was finally obtained by looking for other relevant datasets.
    - At the beginning, the dataset Fatal Police Shootings in the US cannot be read properly by the pandas library, this problem was solved after changing the encoding format of the read data to 'ISO-8859-1'
4. Meta-quality measures:
    - Use the logging module to record information about the program's runtime infomation.
    - A retry mechanism is implemented using the @retry decorator, which will automatically retry up to 3 times if the download fails, which helps to cope with download failures due to unstable networks or other temporary problems.
    - Exception handling is used throughout the ETL process to catch and handle exceptions that may occur.
    - Cleaning and transforming data, as well as format specification. Ensure data accuracy and consistency to facilitate subsequent data analysis.

Result and Limitations

Result: Two tables are generated for each of the two databases and are stored together in the database. They are structured data in the database. The process of data cleansing and conversion was successful and the consistency and accuracy of the data was ensured. It can be prepared for later data analysis.

Limitation: The city names of the two data sources are not exactly the same, so the data from the two sources cannot be merged into a single table. The current project uses only a few software engineering techniques, the robustness of the code needs to be strengthened.