A visual analysis of the security situation and potential crime influencing factors in California‘s cities

Introduction

California is not only the most economically developed state in the U.S., but also the most populous state in the U.S. Its stunning scenery, booming economy, advanced technology and excellent educational resources combine to attract visitors from all over the world. While the security situation in California is an important consideration for people before traveling, California's security situation presents complexity. In this regard, this project puts forward the following questions to study the security situation in California cities and tries to answer them through data engineering methods:

1. What types of crimes are most common in all California cities? Is there a type of crime that is predominant in most cities?
2. Which cities in California have the highest crime rate (per 100000 residents)?
3. Do all California cities have similar ratios of the number of law enforcement officers to the number of crimes? Or are the ratios of the number of law enforcement to the population similar?
4. Do California cities with higher median household incomes or high school graduation rates have lower crime rates?
5. Do cities with high poverty rates have higher crime rates?

The answers to these questions can provide a safety index for people who want to travel or settle in California, and can also help policymakers and government officials develop personalized crime prevention measures in specific cities.

Used Data

This project uses two datasets named ***Crime_and_Law_Enforcement_Data*** and ***Social_Factors_Data***. They record data from 320 cities in California. They are described in detail with examples below:

1. ***Crime_and_Law_Enforcement_Data***
   This dataset is under the US-PD license. It records crime cases, population, and number of law enforcement officers in each California city in 2015. It consists of the following fields:
   - *City*: This field represents the name of the city. It is a string type that identifies the different cities in the table.
   - *Murder_and_nonnegligent_manslaughter*: Non-negligent manslaughter is the intentional infliction of death. This field indicates the number of murders and non-negligent manslaughters that occurred in the corresponding city, which is expressed as an integer.
   - *Rape_(revised_definition)*: This field represents the number of rapes that occurred in the corresponding city according to the revised definition, which is expressed as an integer.
   - *Robbery*: This field indicates the number of robberies that occurred in the corresponding city, which is expressed as an integer.

- *Aggravated_assault*: An aggravated assault usually results in injury to the victim. This field indicates the number of aggravated assault cases that occurred in the corresponding city, which is expressed as an integer.
- *Burglary*: Burglary is the breaking and entering of another person's property in order to steal property. This field indicates the number of burglaries that occurred in the corresponding city, which is expressed as an integer.
- *Larceny-theft*: Larceny is the criminal act of taking another person's property without permission. This field indicates the number of burglaries that occurred in the corresponding city, which is expressed as an integer.
- *Motor_vehicle_theft*: This field indicates the number of motor vehicle thefts that occurred in the corresponding city, which is expressed as an integer.
- *Arson*: Arson is the offense of intentionally creating a fire. This field represents the number of arson cases that occurred in the city, which is expressed as an integer.
- *Population*: This field represents the total population of a city, expressed as an integer.
- *Total_law_enforcement_employees*: Law enforcement employees include police officers and other personnel with law enforcement authority. This field represents the total number of law enforcement employees in the corresponding city, expressed as an integer.

2. ***Social_Factors_Data***
   This dataset adheres to the CC BY-NC-SA 4.0 license, which documents social factors potentially affecting crime rates in various California cities in 2015.It was released by the U.S. Census Bureau and was organized by Karolina Wullum. It consists of the following fields:
   - *City*: This field represents the name of the city. It is a string type.
   - *Median_Income*: This field represents the median household income in the city, which is expressed as an integer. Median income is an important indicators of a city's economic level.
   - *poverty_rate*: This field represents the city's poverty rate, which is the percentage of the total population living below the poverty line. It is expressed as a floating point number.
   - *percent_completed_hs*: This field represents the percentage of residents in the city who have completed high school, which is an important indicator of the level of education and quality of the residents. Floating point numbers are used for percentage representation.
3. **Examples**
   For crime and law enforcement data in 2015, Los Angeles had 282 murders and non-negligent manslaughters, 2,209 rapes, 8,952 robberies, 13,713 aggravated assaults, 16,160 burglaries, 61,191 larcenies. The number of motor vehicle thefts was 16,152, the number of arsons was 1,131, the population was 3,962,726, and the total number of law enforcement officers was 12,665.

City	Murder_and_nonnegligent_manslaughter	Rape_(revised_definition)	Robbery	Aggravated_assault	Burglary	Larceny-theft	Motor_vehicle_theft	Arson	Population	Total_law_enforcement_employe
Los Angeles	282	2209	8952	13713	16160	61191	16152	1131	3962726	12665

Figure 1: Crime and law enforcement data of Los Angeles in 2015

For social factors data in 2015, Los Angeles had a median household income of $50,205, a poverty rate of 22.1%, and a high school graduation rate of 75.5% for residents.

City	Median_Income	poverty_rate	percent_completed_hs
Los Angeles	50205	22.1	75.5

Figure 2: Social factors of Los Angeles in 2015

Analysis

With the above data, this project uses data visualization methods and statistical analysis with the help of $Tableau$ to try to answer the questions previously posed:

1. Using a pie chart to count the most common types of crime in each city, it's found that 96.67% of the cities have $Larceny$ as the most common type of crime, 2.73% of the cities have 2.73% of $Burglary$ as the most common, and $Motor\ Vehicle\ Theft$ and $Aggravated\ Assault$ are respectively most common in 0.3% of the cities. That said, the most common type of crime across cities in California is $Larceny$, which dominates most cities.
2. A horizontal bar graph is used to visualize the cities with the highest crime rates in California, where the crime rate is calculated as $(Total\ Crime\ cases\ /Total\ Population) * 100000$, which represents the number of crimes occurring per $100,000$ residents. A reference line is added for comparison, where the green line represents the median and the blue line represents the Average. In this case, $Vernon$ has an unusually high crime rate, which may be due to the fact that it is an outlier or due to its specificity. It will be excluded from subsequent analyses.

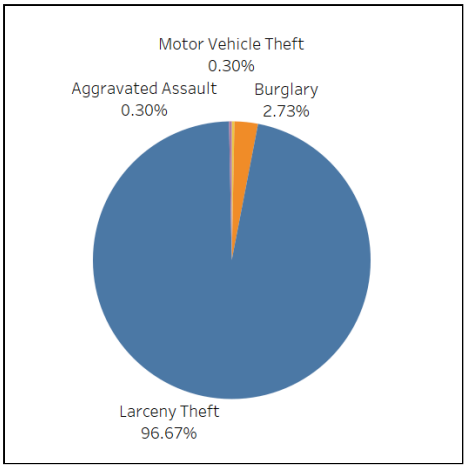

Figure 3: Most common crimes

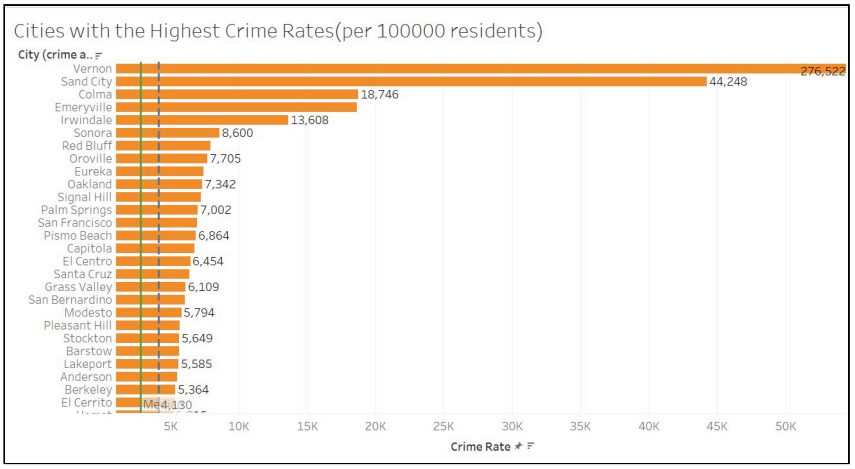

Figure 4: Cities with high crime rate

3. Treemaps are used to reflect the ratio of the number of law enforcement to the number of crimes in each city, with the larger the ratio, the darker color of a city's rectangle. As well as the ratio of the number of law enforcement to the population, the larger the ratio the larger of a city's rectangle. It turns out that the ratios are not similar for all cities and that the number of law enforcement officers does not depend on the number of crimes or population. According to the treemaps, $Etna$ has the highest ratio of the number of law enforcement to the number of crimes, but $Sand\ City$ has the highest ratio of the number of law enforcement to the population.

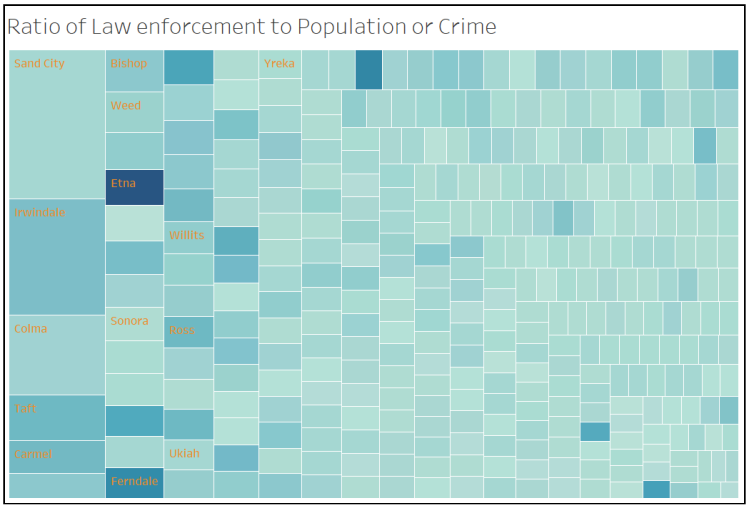

Figure 5: Treemaps reflect two ratios

4. A scatterplot is used to explore whether cities with high median incomes or high high school graduation rates have lower crime rates. The black trend line reflects the general trend. It can be

seen that the negative correlation between crime rates and median income or high school graduation rates is not significant.

5. The scatterplot is continued to be used to explore whether crime and poverty rates are positively correlated, in the same way as above. It shows that crime rates have some positive correlation with poverty rates, but it is not significant

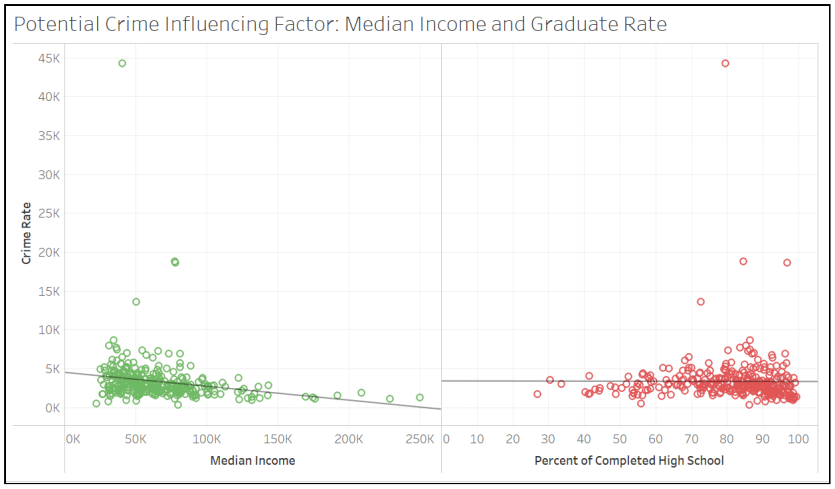

Figure 6: Factor of Median income/Graduate rates

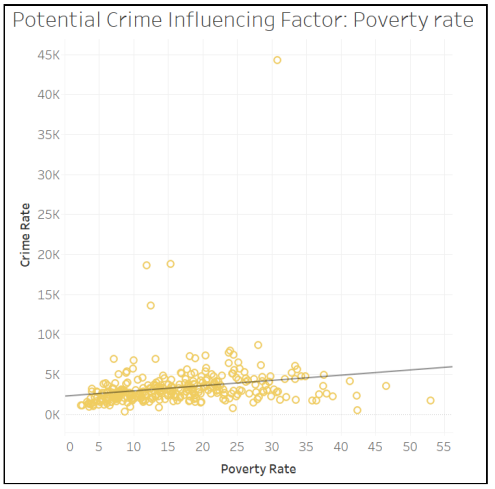

Figure 7: Factor of Poverty Rates

6. The coefficient of determination is used to more visually demonstrate whether crime rates are correlated with median city income, high school graduation rates, or poverty rates, respectively. Its formula is $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$. If R^2 is close to 0, it indicates a low correlation between the two variables and if it's close to 1 it indicates a high correlation. Regardless of the fact that the coefficient of determination of the poverty rate is greater than that of the other two variables. The three variables are all close to 0, suggesting that their impact on crime rate is minimal.

	Social factors		
Variable	High School Gr..	Median Income	Poverty Rate
Crime Rate	0.00001	0.03598	0.03935

Figure 8: Coefficient of determination of three variables

Conclusions and Limitations

Through the above analysis, this project not only reveals the most prevalent types of crime in California cities, and the cities with the highest crime rates in California. It also reveals, through treemaps, that the ratio of the number of law enforcement officers to the number of crimes or to the ratio to the population can vary greatly from city to city. This project also found, through scatterplots and coefficients of determination, that crime rates in the cities are not significantly negatively correlated with median income or high school graduation rates. And there is no significant positive correlation between crime rate and poverty rate. However, due to the fact that the dataset used in the project is from 2015, things may have changed dramatically today. And due to technical means, the project was not able to explore the joint impact of multiple variables on crime rates. Thus the project has limitations.

> Disclaimer: This article was partially translated or grammar-checked by AI tools like ChatGPT and Deepl. Although efforts have been made to ensure accuracy, the content may not perfectly reflect the original meaning.