ETL Project: Police, Are They Understaffed in CA?

Joseph Scott

A group of people in uniform

Description automatically generated

# **Project Description**

In the great state of California are our police understaffed. To answer this question, we will locate datasets from California that contain information such as city/location, population, police department size, and the amount of violent crimes committed. We will combine these data sets to help provide an ease of access and comparable table to help analyze the data.

**Databases**

We utilized 3 datasets from the public platform Kaggle. Each data set is based on each city located in the great state of California. Since we live in California, I believe it has been interesting seeing the sizes and crimes reported to each police department. The datasets are below:

* <https://www.kaggle.com/camnugent/california-housing-feature-engineering>
* <https://www.kaggle.com/fbi-us/california-crime?select=ca_offenses_by_city.csv>
* <https://www.kaggle.com/fbi-us/california-crime?select=ca_law_enforcement_by_city.csv>

**Transformation**

The first step in cleaning the data is to remove any unwanted columns listed in the csv files. For the ca\_law.csv we dropped the “Total law\reinforcement\employees” columns.

A screenshot of a social media post

Description automatically generated

Next we focused on the offenses dataset and began to clean it up. Similarly, to the law dataset we started by removing columns that were unwanted. To help clean and focus in on the data needed for the project.

A screenshot of a cell phone

Description automatically generated

Finally, we took our third dataset and renamed one of the columns to help merge the 3 datasets seamlessly. By renaming one of the columns this allowed each dataset to have a section that can match up and allow for a seamless join.

A screenshot of a social media post

Description automatically generated

**Load**

The final step to our project was to transfer my final output into an actual Database. To accomplish this goal, we created a database and the tables to match the data frame from the final Panada’s document using MySQL. We then connected the database using SQLAlchemy to load the results.