Population Growth and Crime Trends in San Francisco

Anagha Niravane, Gerson Morales, Sankalp NV, Patrick Connor, Amgad Al-Zamkan
University of California, Berkeley

Abstract

As populations grow, communities face increased demands for housing, transportation, and public services. When these demands outpace development, social challenges such as high crime rates may arise.

This project examines the relationship between population growth and crime in San Francisco. By analyzing trends and applying machine learning models, we identified key patterns influencing crime, including temporal, geographic, and demographic factors. Our findings aim to start discussions about new policies to enhance resource allocation, improve response times, and support preventative measures in high-risk areas. These insights aim to guide urban planners and law enforcement in addressing crime effectively and proactively.

Motivation

In areas where population growth outpaces the development of basic resources, stress can occur, potentially leading to social issues, including increased crime.

By analyzing the relationship between population growth and crime, urban planners and local governments can better anticipate areas that might need additional investment in infrastructure, public services, and safety measures before issues escalate.

The data sources for this project were:
San Francisco Pop. and Demographic Census Data (2010-2022)
Police Department Incident Reports: 2018 to Present
Police Department Incident Reports: Historical 2003 to May 2018

Exploratory Data Analysis

The most dangerous neighborhoods by crime count are Soma-Potrero-Mission Bay and Western Addition-Buena Vista-Eureka. Meanwhile, the Richmond-Presidio-Marina and Downtown-Northeast neighborhoods report fewer crimes in San Francisco County.
On the other hand, the top crime by count in the city is Larceny Theft, while Assault and Non-Criminal crimes such as local ordinance violations are less frequent

We also noticed some dataset limitations. Specifically, that neighborhood-specific tagging, such as demographics, was only available for years 2020-2022, which strictly limits our scope with respect to analyzing crime by neighborhood.

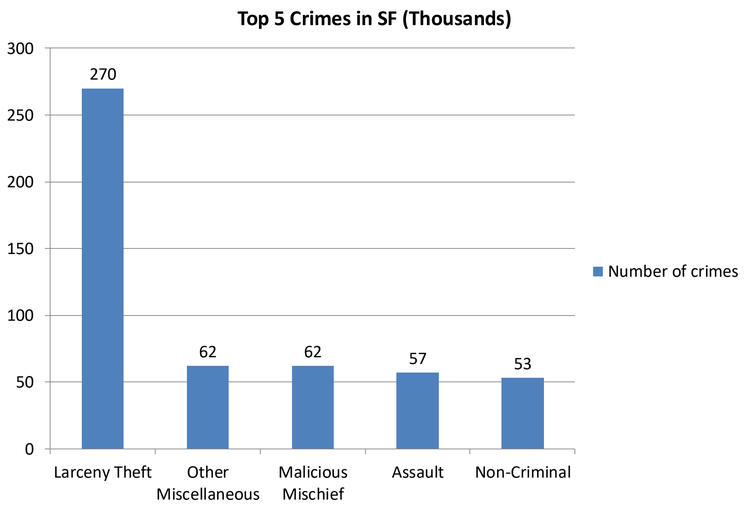

Chart 1. Top Crimes in San Francisco

Results

We built 5 analytical models, each trying to answer a different question. For model #1, the goal was to use the properties of the logistic regression model to inform police response policy decisions. After variable selection, we found that only drug-related crimes were statistically significant in predicting if a response time would exceed 24 hours.

We used Random Forest Regression model for #2, where we trained on the aggregated crime data, incorporating temporal features such as monthly and hourly cycles, as well as encoded neighborhoods. This model explained 70% of the variance in crime counts and reduced prediction error. Similarly for model #3, we used the aggregated crime data but this time incorporating new features such as demographic information, to predict the number of crimes per neighborhood.

To enhance law enforcement's capacity to prevent violent crimes by optimizing resource allocation and coverage in critical areas, we explored machine learning techniques including Random Forest, LightGBM and XGBoost, incorporating hyperparameter tuning for improved performance (model #4). XGBoost demonstrated the highest accuracy and was selected to predict the occurrence of violent crimes.

Finally, in another logistic regression model, our team identified statistically significant categories of crimes that can be used to predict whether a case will be closed. Feature analysis of the significant categories show that sexual and financial crimes are mor likely to result in a case closure, whereas property, drug, traffic and other miscellaneous crimes are less likely to result in a case closure.

Data Cleaning

1. Geospatial: To align the geographic regions from each data source, the crime locations are mapped to the nearest San Francisco neighborhood from the Census dataset.
2. Time Feature: The Incident Datetime column is processed into more meaningful features – year, whether it is a weekend or weekday, categorizing hours of the day with expected criminal patterns (e.g. higher crime rates at night). 'Incident hour' and 'incident month' were encoded as their sine and cosine values in order to preserve the cyclical nature of time.
3. Crime categorization: Aggregation for incidents involving multiple types of crimes (approx. 100), which were Label and One-Hot encoded.
4. General Cleaning: Null-handling, removal of duplicate values, and sanity and completeness checks.

Table 1. Model comparison.

Name	Model	Baseline	Model
#1 Analyzing Police Response Times	Logistic Regression	72.39% (accuracy)	72.39% (accuracy)
#2 Predicting Yearly Crime Counts by Neighborhood	Random Forest	64.7% (R^2)	70.5% (R^2)
#3 Predicting the Number of Crimes per Neighborhood	Random Forest	14% (accuracy)	51% (accuracy)
#4 Predicting Occurrences of Violent Crimes	XGBoost	91.55% (accuracy)	95.64% (accuracy)
#5 Predicting Case Closure	Logistic Regression	NA	99.5% (accuracy)

Impact

After analyzing our final models, we believe that policy makers and police departments could benefit from our insights. Some neighborhoods have a higher crime rate than others and therefore should receive a higher allocation of law-enforcement, as well as preventative social services.

Some policies and recommendations that could be implemented after sharing our results could be:
1. Increasing the severity of the charges for frequent violent crimes as a deterrent.
2. Expanding police presence in unsafe neighborhoods during times of the day when crimes are expected to increase (model #4)
3. Creating social programs catered towards specific demographic neighborhoods (model #3).

Click here to read the report