Population Growth and Crime Trends in San Francisco

Group: Anagha Niravane, Gerson Aaron Morales Deras, Sankalp NV, Patrick Connor, Amgad Al-Zamkan

Motivation

As populations grow, communities change. More people mean more demand for housing, transportation, and basic services. In areas where population growth outpaces the development of these 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. Based on the results of our analysis and identified patterns, our goal is to provide policy recommendations for local government and law-enforcement to identify and address some of this criminal activity.

Data Sources

United States Census Bureau:
San Francisco Population and Demographic Census Data (2010 – 2022)

DataSF:
DataSF - Police Department Incident Reports: 2018 to Present
DataSF - Police Department Incident Reports: Historical 2003 to May 2018

Data Cleaning

1. **Geospatial**: In order 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 - the year, whether it is a weekend or weekday, categorizing hours of the day with expected criminal patterns (e.g. higher crime rates at night). Additionally, '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 (~100) of crimes, which were Label and One-Hot encoded. This was useful since there were originally too many different categories to efficiently work with. Further, we mapped them into a lower dimensional space to create a target variable for classification.
4. **General Cleaning**: Null-handling, removal of duplicate values, and sanity and completeness checks.

See Appendix for code details.

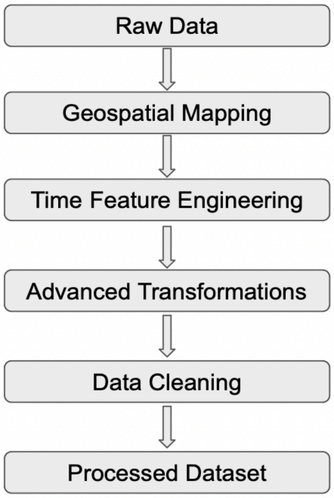
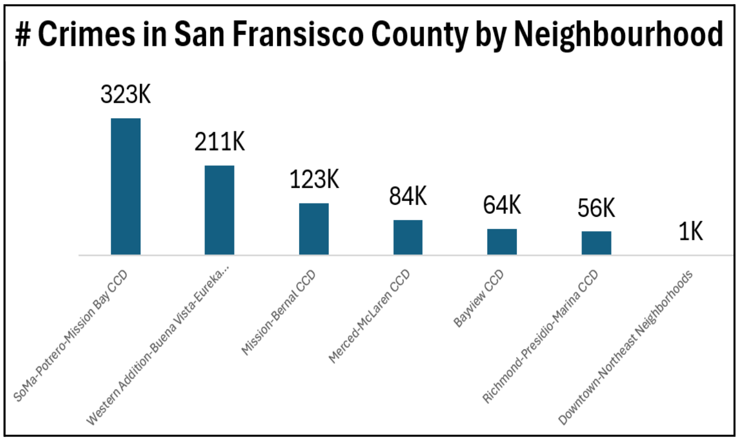


Figure 1: Data Preprocessing
Source: Team creation

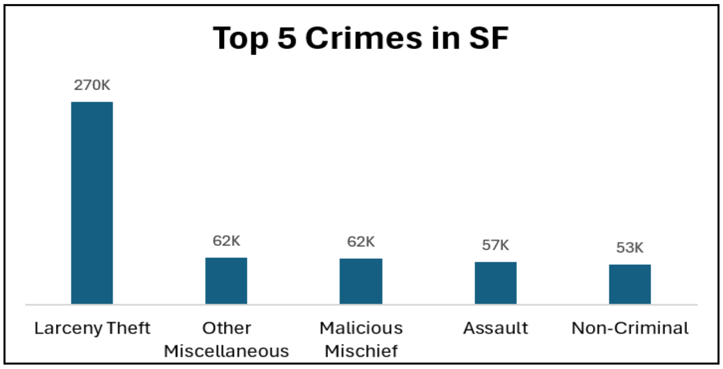
Exploratory Data Analysis

Figures generated from processed data.

Figure 2 (right): Crime count by neighborhood.The most dangerous neighborhoods by crime count are Soma-Potrero-Mission Bay and Western Addition-Buena Vista-Eureka. Meanwhile, the RichmondPresidio-Marina and Downtown-Northeast neighborhoods report fewer crimes in San Francisco County.



On the other hand, Figure 3 (below) tells us that 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.



Dataset Limitations

Neighborhood-specific tagging is available only for years 2020-2022, which strictly limits our scope with respect to analyzing crime by neighborhood.

Additionally, for the year 2020, Male population statistics are not present for the 5 < age group, which suggests that the data is incomplete and unavailable.

Analytical Model #1: Analyzing Police Response Times

We constructed a logistic regression model using consolidated crime categories (violent crimes, property crimes, etc) to improve performance. Even with these consolidated categories and variable selection, our accuracy still matched the baseline accuracy of 72.39% (predicting a response within 24 hours).

However, the important aspect of this model was not to predict response times necessarily, so accuracy (and also TP and FP rates) were not as important. Rather, the goal was to use the properties of the logistic regression model itself 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. This is an important result, because it is common knowledge that higher drug usage and distribution in an area increases crime rates. If these drug-related crimes have longer response times, it can negatively impact that neighborhood's crime rate in the long run. Police should devote more resources to responding to drug-related crimes, since that is the category of crime (according to our analysis) that is more likely to have a longer response time, and therefore a higher likelihood of the offenders committing further crimes. Despite matching baseline accuracy, this insight can impact crime-reduction policies.

Analytical Model #2: Predicting Yearly Crime Counts by Neighborhood

Using a Random Forest Regression model, we trained on the aggregated crime data, incorporating temporal features such as monthly and hourly cycles, as well as encoded neighborhoods. The baseline model achieved an R^2 of 0.647 and an RMSE of 2442.47, improving to an R^2 of 0.705 and RMSE of 2235.95 after hyperparameter tuning.

This model explained 70% of the variance in crime counts and reduced prediction error. Neighborhood encoding was the strongest predictor at 37.5%, followed by monthly seasonal trends at 45%. Meanwhile, hourly trends had a minimal impact.

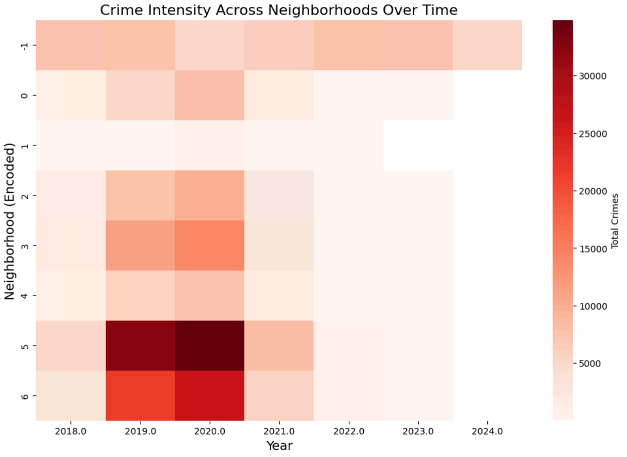


Figure 4: Crime intensity across neighborhoods and years. We can see the critical geographic patterns captured by the model, when noticing the darker colors that indicate areas with consistently high crime. This analysis shows which neighborhoods have experienced increased or decreased crime rates over the years, and therefore indicate how public service and law-enforcement resources can be better distributed.

Analytical Model #3: Predicting the Number of Crimes per Neighborhood

Similar to the previous model, we trained a Random Forest Classifier, using aggregated crime data and incorporating new features such as demographic information. This new data provided us with more information about the generational composition of San Francisco's neighborhoods, like the year the data was collected, gender, age group and population. The baseline model had an accuracy score of 14%, improving to 51% after cross validation.

Using feature importance analysis to improve the model, we discovered that we could remove the Year and Gender features that we previously merged. Figure 5 shows that we could benefit from removing them, and after doing it, our model improved to 53% (again, after cross validation). Please see appendix for code details.

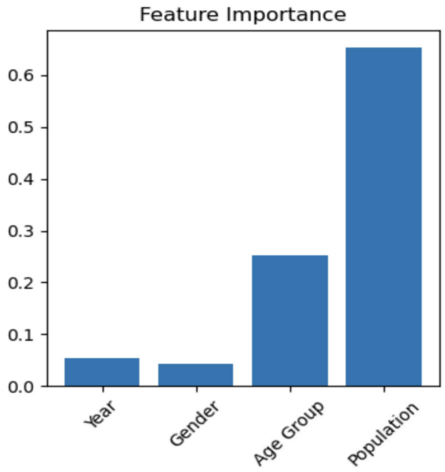


Figure 5: Features important for Model 4. See code for figure generation details.

Analytical Model #4: Predicting Occurrences of Violent Crimes

To enhance law enforcement's capacity to prevent violent crimes by optimizing resource allocation and coverage in critical areas where lives are at stake i.e. violent crimes (such as Homicide, Human trafficking, etc.) we explored machine learning techniques including Random Forest, LightGBM, and XGBoost, incorporating hyperparameter tuning for improved performance. Among these models, XGBoost demonstrated the highest accuracy and was selected to predict the occurrence of violent crimes.

For this analysis, we focused on a subset of crimes identified as particularly violent and used features such as neighborhood, police district, time of day, weekend indicator, response time, and population.

The final model achieved an accuracy of approximately 95.64%, surpassing the baseline accuracy of 91.55%. This improvement highlights the model's enhanced predictive capability where saving even a single life matters.

Analytical Model #5: Predicting Case Closure

In another logistic regression model, our team identified statistically significant categories of crimes that can be used to predict with 99.5% accuracy whether or not a case will be closed. Feature analysis of the significant categories show that sexual and financial crimes are more likely to result in a case closure, whereas property, drug, traffic and other miscellaneous crimes are less likely to result in a case closure (see code referenced in appendix for the sub-crimes in these categories). Notably, violent crimes were not statistically significant - closure in these cases seems to vary on a case-by-case basis. The impact of this result is that, in order to close more cases (and catch more criminals), law-enforcement needs to dedicate more resources to solving and closing property, drug, traffic and other specific types of cases. The fact that drug-related crimes are indicated both in this model, as well as in the analysis of model #2, indicate that this is a significant area of police procedure that needs improvement.

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, such as SoMa-Potrero-Mission Bay, and therefore should receive a higher allocation of law-enforcement, as well as preventative social services. By analyzing factors such as the time of day, the police district where the crime took place, or if it was during the weekend, we can equip stakeholders with tools that allow them to protect law-abiding citizens.

For example, some policies and recommendation 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 (based on results of model #4).
3. Reorienting budget efforts to allocate more resources towards types of crimes that typically have high response times and low case closure rates, such as drug-related crimes (based on analysis in model #2 and model #5).
4. Creating social programs catered towards specific demographic neighborhoods (identified by our feature importance analysis in model #3) as a preventative measure.
5. Allocate more resources and law-enforcement personnel towards neighborhoods where crime intensity is predicted to increase (based on model #2 predictions).
6. Increased scrutiny and oversight on unresolved police reports (overall rate of case closure is low).

Implementing even a subset of the above recommendations should, in the long run, lead to a reduction in San Francisco crime rates and lead to a more efficient allocation of local government and law enforcement resources.

APPENDIX

Data & Code Location (Google Drive):
https://drive.google.com/drive/folders/1EDhVA4UimMfp5v7yQeKK04NJWnf2hcqg?usp=drive_link

To reproduce, make sure all CSVs and notebook files from the above folder are in the same location on your computer to run locally. File names corresponding to each model are below.

Data Cleaning Code:
-Crime Data: *crime_data_processing.ipynb*
-Population Data: We cleaned this dataset using Excel functions and Powery Query, because the original structure didn't allow us to merge it to the Crime Data. *per_Community_2.csv*

Model #1 Code: *crime_response_and_closure_analysis.ipynb*

Model #2 Code: *dataprocessing_and_crimecounts_per_neighborhood.ipynb*

Model #3 Code: *crimes_per_neighborhood.ipynb*

Model #4 Code: *predicting_violent_crimes.ipynb*

Model #5 Code: *crime_response_and_closure_analysis.ipynb*