Geospatial Analysis of Crime Hotspots in San Francisco: Focusing on Car Break-Ins and Proximity to City Facilities

Atharva Chouthai
Computer Science
Virginia Tech
Falls Church
atharvachouthai@vt.edu

Anvita Karne
Computer Science
Virginia Tech
Falls Church
anvitakarne@vt.edu

ABSTRACT

Urban safety poses significant challenges, particularly in densely populated cities like San Francisco, where car break-ins remain a persistent issue. Our project explores the interplay between geospatial patterns, urban infrastructure, and temporal crime trends to identify insights that can drive informed decision-making. Utilizing two comprehensive datasets—police incident reports and city facility records—we conducted detailed data preprocessing and exploratory analysis to uncover high-crime rate neighborhoods, assess the impact of urban facilities on crime distribution, and identify underserved areas lacking adequate security infrastructure. To predict future crime trends, we implemented a series of predictive models, including Random Forest, XGBoost, ARIMA, and SARIMAX. Our findings revealed that Random Forest outperformed other models, achieving an RMSE of 11.61 and a MAPE of 15.08%, demonstrating its robustness in forecasting car break-ins when leveraging temporal features like lags. Moreover, geospatial analysis highlighted critical areas requiring targeted interventions, emphasizing the role of data-driven approaches in urban planning. This study underscores the potential of combining predictive analytics and geospatial insights to address urban safety challenges, offering a scalable framework for policymakers to allocate resources efficiently and reduce crime rates in vulnerable communities.

CCS CONCEPTS

Computing methodologies → Machine learning → Supervised learning → Ensemble methods
Information systems → Data mining → Geographic Information Systems
Applied computing → Law, social, and behavioral sciences → Urban studies
Social and professional studies → Social impact of computing → Crime analysis

KEYWORDS

Urban Safety, Crime Analysis, Geospatial Patterns, Temporal Trends, Predictive Analytics, Random Forest, XGBoost, ARIMA, SARIMA, Urban Infrastructure, Data-Driven Decision Making, Crime Hotspots, Urban Planning, Machine Learning, Car Break-ins, Resource Allocation.

ACM Reference format:

Atharva Chouthai, Anvita Karne. 2024. Geospatial Analysis of Crime Hotspots in San Francisco: Focusing on Car Break-ins and Proximity to City Facilities

1. INTRODUCTION

San Francisco, celebrated for its iconic Golden Gate Bridge, bustling Fisherman's Wharf, and rich cultural tapestry, faces a growing concern that is undermining the safety and peace of mind of its residents and visitors alike—car break-ins. This issue has persisted as a significant challenge for the city, earning it a reputation for high property crime rates. In recent years, car break-ins have emerged as one of the most common forms of theft, impacting not only vehicle owners but also the city's reputation as a global tourist destination. According to recent reports, 2023 witnessed over 15,000 incidents of car break-ins, equating to an alarming average of nearly 59 incidents per day. Such a staggering figure highlights the scale of this problem and its far-reaching consequences for the safety of residents, the experience of visitors, and the broader economic implications for San Francisco.

Addressing this pervasive issue demands a comprehensive understanding of its underlying patterns and contributing factors. With this objective in mind, our team embarked on an in-depth analysis of car break-ins in San Francisco. Leveraging two distinct datasets—one detailing reported police incidents and the other cataloging city facilities—we aimed to provide a holistic perspective on the issue. This dual-data approach not only allowed us to examine the spatial and temporal trends of these crimes but also to explore the potential correlation between city infrastructure, such as parking facilities and recreational areas, and the occurrence of car break-ins. Our approach aligns with the broader goal of urban computing: to use data-driven insights to improve city planning and enhance public safety.

Through our analysis, we uncovered significant insights into the nature of car break-ins in the city. Spatially, certain neighborhoods were identified as persistent hotspots, with areas such as the Tenderloin and Bayview-Hunters Point recording disproportionately high incident rates of over 65,000 and 46,000 incidents, respectively, over the dataset's timeframe. Temporally, we observed interesting patterns, such as the higher frequency of break-ins during weekends and particular months of the year, underscoring the seasonal and behavioral factors at play. By analyzing these trends, we sought to understand not only where these crimes occur but also when they are most likely to take place, providing valuable information for preventive measures.

To add a predictive layer to our analysis, we implemented several forecasting models, including Random Forest and XGBoost, to estimate monthly car break-in trends. Among these, Random Forest emerged as the most accurate model, achieving a Root Mean Squared Error (RMSE) of 11 and a Mean Absolute Percentage Error (MAPE) of 15% for monthly predictions. These forecasts are crucial for city planners and law enforcement agencies to anticipate crime trends and allocate resources efficiently. Additionally, our forecasts allow for proactive measures to be taken during high-risk periods, such as holidays or weekends when car break-ins are likely to spike.

Our work represents a significant effort to address the challenge of car break-ins in San Francisco through data-driven insights and forecasting. By integrating spatial, temporal, and predictive analyses, we aim to provide actionable recommendations to stakeholders. Whether it involves deploying additional law enforcement personnel to high-crime areas, enhancing lighting and surveillance around parking facilities, or improving the design of city infrastructure to deter theft, our findings offer a foundation for targeted interventions. Ultimately, we hope our analysis contributes to making San Francisco a safer city for its residents and visitors, restoring its image as a vibrant and secure destination.

2. RELATED WORK

Urban crime studies have increasingly adopted spatio-temporal analytical methods to identify and predict crime hotspots, leveraging advancements in statistical, geospatial, and machine learning techniques. Andresen and Malleson (2011) emphasize the importance of understanding the stability of crime patterns, highlighting implications for both theoretical frameworks and policy interventions [1]. Grubesic and Mack (2008) explore the spatio-temporal interactions of urban crimes, offering insights into how crime fluctuates based on spatial and temporal dynamics [2]. Recent works have incorporated crowd-sensed and multisource data to enhance predictive accuracy. For instance, Jiang and Zeng (2019) utilize urban data streams to analyze crime patterns, demonstrating the power of integrating diverse data types for spatio-temporal analysis [6]. Butt et al. (2020) provide a systematic review of methodologies in crime hotspot detection, identifying machine learning as a transformative tool for real-time prediction and mapping of crime trends [4]. Moreover, modern approaches emphasize the integration of shape and concentration characteristics in crime mapping, as discussed by Lai et al. (2022), who compare various hotspot detection techniques for their effectiveness [9]. Advances in machine learning have further enhanced predictive capabilities, with Nair and Gopi (2019) contrasting different algorithms to uncover their applicability in crime hotspot prediction [8]. Additionally, Kapoor and Gopi (2020) highlight the potential of combining geospatial and statistical methods to visualize crime distributions and improve decision-making processes [7]. Publicly accessible tools like ArcGIS StoryMaps (2023) and platforms like Medium (2023) demonstrate how visualizations contribute to understanding and addressing urban crime challenges [10] [11]. These studies collectively underscore the significance of interdisciplinary methods in improving crime prediction and prevention strategies.

3. METHODOLOGY

3.1 Dataset Description and Preprocessing

The study leverages two distinct datasets to address the problem of car break-ins in San Francisco and their relationship with city infrastructure. The San Francisco Police Department Incident Reports (2018 to Present) dataset provides comprehensive information about reported incidents, including geospatial and temporal attributes, while the City Facilities Dataset offers details about public facilities, including their locations and types. Together, these datasets enable a detailed analysis of car break-ins and their spatial correlation with city infrastructure.

To prepare the data for analysis and modeling, we undertook a rigorous preprocessing process. For the San Francisco Police Incident Dataset, we began by selecting only the most relevant columns: Incident Datetime, Incident Category, Incident Subcategory, Latitude, Longitude, and Analysis Neighborhood. This ensured that our analysis would focus specifically on spatial and temporal crime patterns. The Incident Datetime column was converted into a proper datetime format, allowing us to extract key temporal features, such as the year, month, day of the week, and hour. To maintain the integrity of our spatial analyses, we removed all rows with missing latitude or longitude values. Next, we filtered the dataset to include only car break-ins, focusing on incidents described as "Theft, From Locked Vehicle,>$950" or "Theft, From Locked Vehicle, $200-$950." This step ensured that our study remained targeted to the research objectives.

For the City Facilities Dataset, we selected critical columns such as common_name, latitude, longitude, and jurisdiction to better understand the spatial relationship between facilities and crime. Similar to the incident dataset, rows with missing latitude or longitude values were removed to maintain accuracy in geospatial analyses. These steps ensured that the facilities dataset was clean and ready to be integrated with the crime dataset.

To enable a detailed analysis, we integrated the two datasets by identifying the nearest city facility for each crime incident. Using the KDTree algorithm, we computed the closest facility for every incident based on geographic coordinates and added two key features to the dataset: Nearest Facility ID and Distance to Facility. Furthermore, we aggregated crime counts by facility and neighborhood to highlight high-crime areas and underserved regions in terms of infrastructure. These preprocessing steps created a robust foundation for our subsequent analysis and modeling, ensuring the data was clean, relevant, and aligned with our research objectives.

3.2 Exploratory Data Analysis (EDA)

We conducted a detailed Exploratory Data Analysis (EDA) to uncover meaningful patterns in the car break-in data and identify trends across temporal, spatial, and categorical dimensions. Our EDA began with a temporal analysis, where incidents were aggregated by year, month, and day of the week to identify seasonal and weekly trends. Using visualizations such as bar charts and line plots, we explored how the frequency of car break-ins fluctuated over time, revealing potential correlations with factors like holidays or tourist activity. This analysis helped us understand whether certain periods experienced spikes in incidents, providing valuable context for predicting future occurrences.

In addition to temporal analysis, we examined the spatial distribution of car break-ins using geospatial techniques. Neighborhood-wise crime counts were calculated, and heatmaps were created to visualize high-crime areas. To further investigate the relationship between crime prevalence and infrastructure, we overlaid crime locations with city facility data, identifying underserved areas with high crime rates but limited resources. Finally, categorical analysis was performed by grouping data by Incident Subcategory and Incident Description to highlight the most common types of incidents. These insights formed the foundation for modeling efforts and allowed us to target areas and periods with higher vulnerability to car break-ins. The combination of descriptive statistics, geospatial mapping, and visualizations provided a holistic view of the data and guided our subsequent steps.

3.3 Feature Selection

For our analysis, we carefully selected features from both the City Facilities and Police Incidents datasets to ensure that our study was both meaningful and efficient. From the Facilities dataset, we focused on fields like facility_id, common_name, latitude, longitude, and jurisdiction, which helped us understand the types and locations of facilities across the city and their proximity to crime hotspots. Similarly, from the Police Incidents dataset, we chose features such as Incident Datetime, Incident Category, Incident Subcategory, Latitude, Longitude, and Analysis Neighborhood to capture when, where, and what types of crimes were occurring.

To prepare the data, we worked to address missing values by inputting categorical fields with "Unknown" and removing rows with incomplete geographical coordinates. Duplicates were carefully handled by leveraging unique identifiers like Incident ID. These steps ensured that the datasets were clean and ready for analysis, allowing us to draw accurate and impactful insights. This process was a vital part of setting the stage for the exploratory and predictive analyses that followed.

3.4 Model Description

The modeling phase of our project was built upon the preprocessed car break-ins dataset, where we aggregated the data into monthly counts of break-ins to create a time series. This aggregation provided us with a comprehensive view of trends and patterns over time. To prepare the dataset for modeling, we began by performing a stationarity check using the Augmented Dickey-Fuller (ADF) test, which indicated the need for differencing to make the series stationary. This step was crucial for ensuring the validity of our statistical models.

We also examined the autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify relationships within the data and determine suitable parameter values (p, d, q) for our ARIMA and SARIMA models. For the Random Forest and XGBoost models, we transformed the time series into a supervised learning format by creating lag features, allowing the models to leverage temporal dependencies for prediction. By combining statistical methods and machine learning approaches, we ensured a robust and diverse exploration of predictive modeling.

3.4.1 ARIMA Model

The ARIMA (AutoRegressive Integrated Moving Average) model was applied to forecast monthly car break-ins by analyzing past patterns and trends. The model uses three components:

- Autoregression (AR): The relationship between the number of car break-ins in each month and its preceding months.
- Integration (I): The differencing of data points to remove trends and make the series stationary.
- Moving Average (MA): The dependency between the number of car break-ins and residual errors from previous predictions.

The ARIMA model for our problem can be represented as:

$$yt = c + \sum_{i=1}^{p} \phi i yt - i \, + \sum_{j=1}^{q} \theta j \epsilon t - j + \epsilon t$$

Where yt is the monthly car break-ins, ϕi and θj are coefficients for AR and MA components, and ϵt is the error term.

3.4.2 SARIMA Model

The SARIMA (Seasonal ARIMA) model extends ARIMA by adding seasonal components ($P,\ D,\ Q$) to capture annual cycles. The SARIMA equation for our data is:

$$yt = \Phi(B^s)yt - s + \Theta(B^s)\epsilon t - s + non - seasonalterms$$

Where s = 12 represents the seasonal period (months in a year), $\Phi(B^s)$ and $\Theta(B^s)$ captures the seasonal autoregressive and moving average components respectively, and $\epsilon t - s$ represents seasonal error terms.

3.4.3 Random Forest Regression

While ARIMA and SARIMA provided robust statistical foundations for understanding and forecasting time series data, we recognized their limitations, especially when capturing non-linear relationships in complex datasets like ours. Given the intricate nature of factors influencing car break-ins, such as temporal, spatial, and seasonal patterns, we decided to explore machine learning techniques to better capture these dynamics. Random Forest stood out as a suitable candidate due to its ability to handle non-linearities, incorporate lagged features, and manage high-dimensional data effectively. This ensemble method leverages decision trees and aggregation, providing both robustness and interpretability for our forecasting problem. The Random Forest model aggregates predictions from multiple decision trees. For each tree Ti the ensemble, the prediction $y^{\wedge i}$is generated as:

$$y = \frac{1}{N}\sum_{i=1}^{N} Ti(X)$$

Where y^: Final forecasted value (e.g., monthly car break-ins), N: Total number of decision trees, Ti (X): Prediction of the i-th tree for input features X.

3.4.4 XGBoost

While Random Forest demonstrated robustness and the ability to handle feature-rich datasets effectively, we recognized its limitations in capturing temporal dependencies and handling very high-dimensional data efficiently. To address these challenges, we explored XGBoost, a gradient boosting framework known for its scalability, speed, and precision in predictive modeling. XGBoost is particularly well-suited for sequential data like time series due to its ability to optimize both gradient descent and decision tree learning, making it an ideal choice for our forecasting problem.

The XGBoost algorithm minimizes the following regularized objective function for a dataset with n samples and m features:

$$L = \sum_{i=1}^{n} l(yi, y^{\wedge} i) + \sum_{k=1}^{K} \Omega(fk),$$

$l(yi, y^{\wedge} i)$ is a differentiable convex loss function (e.g., mean squared error) for the prediction $y^{\wedge} i$ compared to the true value yi .$\Omega(fk) = \gamma T + \frac{\aleph}{2}\sum_{j=1}^{T} wj2$, is the regularization term, controlling the complexity of the treefk , where T is the number of leaves and wj is the weight of each leaf.

4. EXPERIMENTS

This study undertakes a systematic investigation into urban dynamics, addressing critical challenges through the spatial and temporal analysis of city facilities and police incident datasets. The objective was to identify patterns, relationships, and actionable insights that could guide urban planning and policy formulation. The experimentation process entailed thorough data preparation, establishing a structured experimental framework, and extracting meaningful conclusions through detailed analysis. This section details the datasets employed, the experimental design, and the results achieved.

4.1 Data

The foundation of our experiments lies in the San Francisco Police Incidents Dataset and the City Facilities Dataset, which provided a robust framework for analyzing car break-ins across the city. These datasets underwent rigorous preprocessing to ensure consistency, reliability, and relevance for our analysis. The Police Incidents Dataset initially comprised 843,053 records with 17 columns. After removing duplicates, the dataset was reduced to 697,556 unique records, ensuring the integrity of the data for further analysis. To focus specifically on car break-ins, we filtered the data for relevant incident descriptions, yielding a refined dataset of 97,261 records with six relevant features. This targeted dataset captured essential spatial, temporal, and categorical dimensions of car break-ins, such as incident descriptions, timestamps, and neighborhood locations. The City Facilities Dataset, containing 1,729 records and 14 columns, required minimal preprocessing as it exhibited no duplicates. The dataset provided critical details such as facility names, locations, and jurisdictional information, which were vital for proximity and neighborhood-level analyses. By merging the datasets, we created a comprehensive view of car break-ins across neighborhoods. This dataset was then aggregated monthly, enabling effective temporal analysis and forecasting. For our experiments, we divided the data into a training set (January 2018 to March 2024) and a test set (April 2024 to September 2024), ensuring a realistic setup for evaluating model performance.

4.2 Experimental Setup

The experiments that we conducted were in systematic manner to compare the performance of different models in forecasting monthly car break-ins. The workflow comprised the following key steps:

- **Stationarity Check:** Before applying time series models, we ensured that the data adhered to stationarity requirements, as this is a critical assumption for many time series algorithms. Using the Augmented Dickey-Fuller (ADF) test, we assessed the stationarity of the data and applied differencing where necessary to achieve a stationary series. This step was crucial in preparing the dataset for effective modeling.
- **Autocorrelation and Partial Autocorrelation Analysis:** To determine the parameters for ARIMA and SARIMA models, we relied on autocorrelation (ACF) and partial autocorrelation (PACF) plots. These visualizations guided us in identifying the lag values and seasonal dependencies in the data, ensuring that the models were tailored to capture both short-term and seasonal patterns.
- **Modeling Framework:**
- **ARIMA and SARIMA Models:** We implemented these traditional statistical approaches to understand the seasonal and non-seasonal trends present in car break-ins. These models allowed us to explore the temporal dynamics and seasonal behaviors inherent in the dataset.
- **Random Forest and XGBoost Models:** We incorporated these machines learning models for their ability to handle non-linear relationships and utilize feature-rich datasets. By creating lagged features, we ensured that temporal dependencies were effectively captured, enabling these models to make robust predictions.
- **Evaluation Metrics:** To objectively assess the performance of our models, we used RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error). These metrics provided a clear indication of how well the models performed on the test data, helping us compare the strengths and limitations of each approach.
- **Tools and Libraries:** We conducted our experiments using Python and a suite of powerful libraries. Statsmodels was utilized for statistical modeling, while scikit-learn and xgboost were instrumental in implementing machine learning models. For data visualization and analysis, matplotlib and other libraries were employed, allowing us to present insights effectively. Throughout this process, we leveraged these tools to streamline our workflow and ensure accuracy in our analyses.

4.3 Results

Model	**RMSE**	**MAPE (%)**
Random Forest	11.61	15.08
XGBoost	22.46	20.15
ARIMA	226.35	89.80
SARIMA	234.19	82.04

To evaluate the effectiveness of different predictive models for forecasting monthly car break-ins in San Francisco, we compared the performance of Random Forest, XGBoost, ARIMA, and SARIMAX. The models were assessed using two key metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). These metrics provided insights into the accuracy and reliability of the predictions.

Random Forest emerged as the best-performing model, achieving an RMSE of 11.61 and a MAPE of 15.08%, making it significantly more accurate than the other models. The Random Forest's ability to handle non-linear relationships and capture temporal patterns through lagged features proved crucial in forecasting.

XGBoost, while not as precise as Random Forest, performed reasonably well, with an RMSE of 22.46 and a MAPE of 20.15%. The model’s gradient boosting approach allowed it to capture trends effectively, though it fell short of Random Forest in terms of fine-grained accuracy.

ARIMA, a statistical model, struggled with the complexity of the dataset. It produced an RMSE of 226.35 and a MAPE of 89.80%, indicating that the model was unable to handle the non-stationarity and non-linearity inherent in the data. This result highlights the limitations of purely statistical methods for this use case.

SARIMA, which extends ARIMA with seasonal components and external regressors, performed slightly better than ARIMA in terms of MAPE (82.04%) but exhibited a comparable RMSE of 234.19. While it accounted for seasonal patterns, the model still faced challenges with the dataset's complexity and variability.

In our case, the dataset primarily contained historical crime data without any external factors, such as economic conditions, population density, or city-level policies, which could significantly impact crime trends. ARIMA and SARIMA are not equipped to infer relationships beyond the observed series. As a result, they likely struggled to adapt to the non-linear and complex patterns present in the data, leading to poor performance. In future, may be incorporating relevant external factors, such as socioeconomic indicators or weather conditions, could enhance the predictive power of SARIMA and ARIMA models, enabling them to capture a more comprehensive view of the factors influencing car break-ins.

5. DISCUSSIONS

The findings of our study offer critical insights into the patterns and trends of car break-ins in San Francisco, providing a foundation for actionable strategies in urban planning and crime mitigation. Through detailed exploratory data analysis (EDA) and model predictions, we aimed to uncover key drivers of car break-ins and the geographical and temporal distributions of these incidents. This section discusses our key observations and their implications, focusing on how the results can inform decision-making and urban management.

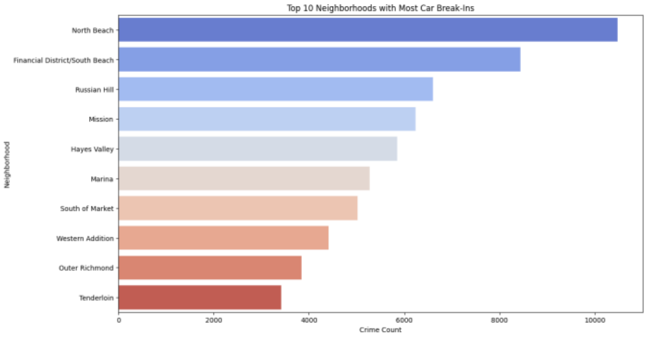

Fig 1: Top 10 Neighborhoods with Most Car-break ins

One of the most striking findings comes from the "Crime Distribution Across Top 10 Facilities" plot shown above. This visualization highlights the frequency of car break-ins and other crimes across facilities such as garages, navigation centers, and shelters. Notably, Ellis-O'Farrell Garage, which is essentially a parking garage, stands out as a major hotspot with over 14,000 incidents. Such findings underscore the need for focused interventions in these facilities, such as enhanced surveillance, improved lighting, and increased security personnel. Additionally, this analysis can assist policymakers in prioritizing resource allocation to high-crime facilities to prevent future incidents effectively.

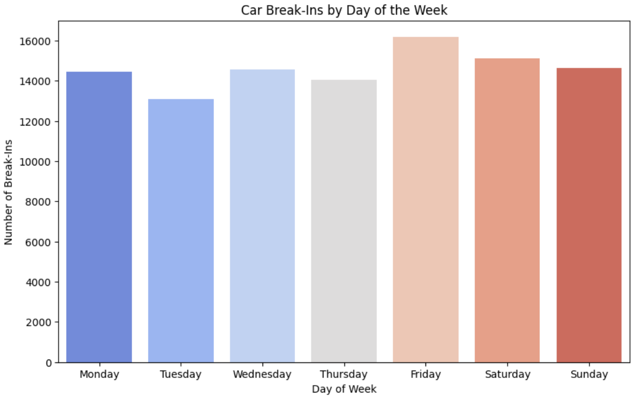

Fig 2: Temporal Analysis: Car Break-ins by Day of the Week

Building upon our earlier spatial insights, we next delved into the temporal patterns of car break-ins to understand how crime activity fluctuates throughout the week. The above bar plot reveals the distribution of car break-ins across the seven days. Interestingly, Friday emerges as the day with the highest number of incidents, closely followed by Saturday and Monday, while Tuesday shows the lowest count. This pattern could reflect the dynamics of urban activities. Fridays and weekends might witness heightened activity around public places, recreational areas, or nightlife hubs, correlating with increased opportunities for vehicle-related crimes. Conversely, lower activity on Tuesdays aligns with reduced mobility and fewer crowded events. These insights are particularly valuable for law enforcement agencies and policymakers in scheduling targeted patrols or implementing preventive measures around peak crime days. By focusing resources effectively, the overall impact of such interventions could be significantly amplified.

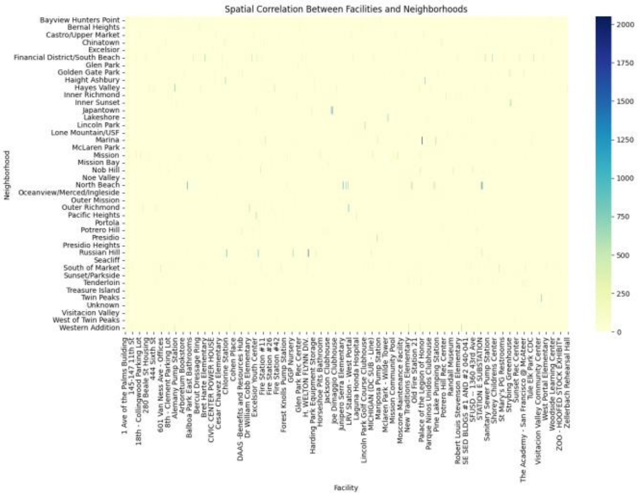

Fig 3: Spatial Correlation Analysis Between Facilities and Neighborhood

The heatmap illustrates the spatial correlation between city facilities and neighborhoods, highlighting how crimes are concentrated around specific facilities. Notable high-crime neighborhoods, such as Tenderloin and Mission, show significant activity near facilities like parking garages and transportation hubs. These insights reveal how urban infrastructure can inadvertently become crime hotspots, emphasizing the need for targeted interventions like improved lighting, surveillance, or urban redesign to enhance community safety.

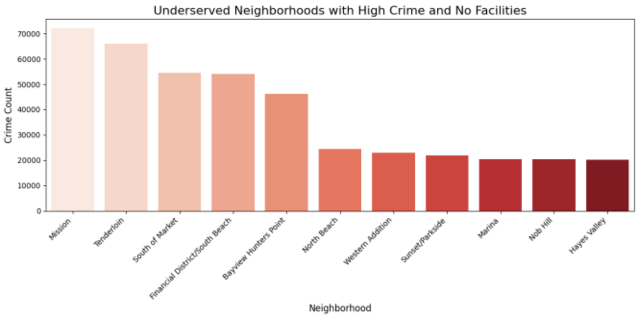

Fig 4: Underserved Neighborhoods with High Crime and Number of Facilities

An essential insight from our analysis is the identification of underserved neighborhoods, which experience high overall crime rates but lack adequate facilities or security measures. As shown in the plot above, neighborhoods like Mission, Tenderloin, and South of Market exhibit the highest crime counts across all categories. This trend underscores the need for targeted interventions in these areas to reduce crime and enhance urban safety. We think that these underserved areas provide a more comprehensive view of the city's crime distribution. It highlights how resource allocation and facility placement can influence crime mitigation strategies. These findings can inform urban planners and policymakers to prioritize infrastructure development and safety measures in these high-crime neighborhoods.

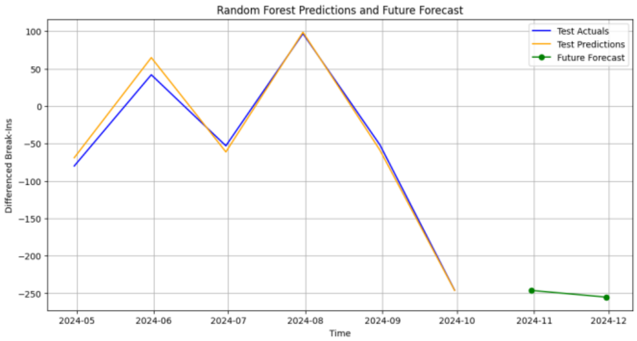

Fig 5: Forecast for Car Break-ins using Random Forest

The Random Forest model demonstrated strong predictive capabilities during our evaluation, as shown in the plot above. The model effectively captured the trends in the test dataset, producing predictions that closely align with the actual test values. When used for forecasting future monthly car break-ins, the Random Forest model predicted a significant decrease in incidents for October and November 2024, with values of -246.36 and -255.34, respectively. This suggests that the model has identified a downward trend in car break-ins. These results underline the robustness of Random Forest in handling complex temporal dependencies, making it a reliable tool for forecasting in scenarios involving fluctuating crime data.

6. CONCLUSION

Urban safety is a pressing concern, and we believe innovative, data-driven approaches are essential for tackling challenges like persistent car break-ins in major cities such as San Francisco. Through this project, we set out to explore how crime data and geospatial analysis can uncover the spatial and temporal factors driving these incidents. By identifying crime hotspots and their connection to urban infrastructure, we aim to provide insights that are not just theoretical but directly actionable. By offering a framework that urban planners, policymakers, and law enforcement agencies can use to allocate resources more effectively and implement targeted interventions, we hope to contribute to safer communities. We are also mindful of the need to address underserved areas that face heightened crime risks, striving to promote equity and build trust within urban populations. Through this effort, we aspire to help create cities where everyone feels secure and supported.

Our findings highlight the value of applying Machine Learning and Time-Series forecasting models to understand and predict urban crime, with each model offering unique insights. Among the models we evaluated, Random Forest emerged as the most effective, achieving an RMSE of 11.61 and a MAPE of 15.08%. This performance indicates its strong capability to predict crime occurrences based on the features we utilized. The XGBoost model also showed promising results, with an RMSE of 22.46 and a MAPE of 20.15%. In contrast, the time-series models, ARIMA and SARIMAX, underperformed, with SARIMAX recording a notably high MAPE of 82.04%. We believe this underperformance is due to the absence of external influencing factors, such as economic trends, weather conditions, or population density, which could have provided additional context to temporal crime patterns.

These results emphasize the importance of selecting models that align with the richness of the available data. Machine learning models like Random Forest and XGBoost proved effective in leveraging the spatial and temporal aspects of our dataset, delivering actionable insights. However, we recognize the potential of time-series models like SARIMA, and incorporating additional external variables in future iterations could significantly improve their predictive power. This project not only validates the potential of data-driven approaches in understanding urban crime but also lays the groundwork for enhancing these methodologies to better support efforts in creating safer, more resilient cities.

In the future, we believe that leveraging advanced neural network architectures, such as LSTM or WaveNet, could significantly enhance the accuracy and performance of our predictive models. However, implementing these techniques effectively would require a more extensive and diverse dataset to fully capture the complex temporal patterns and dependencies inherent in the data.

7. ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to the San Francisco Government for providing access to the datasets that formed the foundation of this study. The San Francisco Police Incident Reports dataset and the San Francisco City Facilities dataset were instrumental in conducting the spatio-temporal analysis and identifying crime hotspots. Their comprehensive and well-maintained open data initiatives have greatly facilitated research aimed at improving urban safety and policy planning.

8. AUTHOR CONTRIBUTIONS

Phase	**Teammate A (Atharva Chouthai)**	**Teammate B (Anvita Karne)**
Phase 1: Topic Research &	Researched potential topics and explored urban	Led final topic selection and refined the scope to include

Finalization	safety challenges, focusing on crime hotspots and car break-ins.	spatial and temporal analysis of crime patterns, focusing on high-crime areas and underserved neighborhoods.
Phase 2: Data Collection & Preprocessing	Collected and cleaned the dataset, merging police incident reports with urban facility data.	Preprocessed the data by handling missing values, encoding categorical variables, and scaling numerical features for analysis.
Phase 3: Dataset Integration & Feature Engineering	Integrated additional external data sources such as transit hubs and parking lots to enhance the dataset.	Focused on feature engineering, deriving new variables like crime density per area and time-specific features to improve model input.
Phase 4: Model Selection & Development	Evaluated and implemented ARIMA and Random Forest models to capture temporal and spacial trends in crime data.	Implemented SARIMA and XGBoost models, focusing on spatial and temporal crime prediction.
Phase 5: Model Evaluation	Evaluated ARIMA and SARIMAX models, analyzing RMSE and MAPE values to assess accuracy.	Evaluated Random Forest and XGBoost models, focusing on model performance metrics like RMSE and MAPE. Visualized results for better interpretation.
Phase 6: Evaluation Metrics & Analysis	Conducted performance analysis using RMSE and MAPE, comparing model results across different methods.	Assisted in performance analysis and contributed to visualizing crime hotspots and trends.
Phase 7: Report Writing	Drafted sections on data collection, preprocessing, and model evaluation.	Drafted sections on introduction, results, conclusions, and implications of findings.

9. DATA AND CODE AVAILABILITY

9.1 Datasets

1. San Francisco Police Incident Reports. Dataset. Available at https://data.sfgov.org/Public-Safety/Map-of-Police-Department-Incident-Reports-2018-to-/jq29-s5wp

2. San Francisco City Facilities Dataset. Dataset. Available at https://data.sfgov.org/City-Infrastructure/Map-of-San-Francisco-City-Facilities/kjx8-bbpd

9.2 Code Availability

We have submitted the code along with the original and processed dataset in zip file in the assignment section as we have coded in single python notebook with proper section headings.

10. REFERENCES

[1] Andresen, M. A., and Malleson, N. 2011. Testing the stability of crime patterns: Implications for theory and policy. Journal of Research in Crime and Delinquency, 48, 1 (2011), 58–82. DOI: 10.1177/0022427810384136.
[2] Grubesic, T. H., and Mack, E. A. 2008. Spatio-temporal interaction of urban crime. Journal of Quantitative Criminology, 24, 3 (2008), 285–306. DOI: 10.1007/s10940-008-9047-5.
[3] Spatial–Temporal Analysis of Hotspot Crime in a Large City: A Case Study of San Francisco. Accessed online at https://rdcu.be/dVCKX.
[4] Butt, U. M., et al. 2020. Spatio-Temporal Crime Hotspot Detection and Prediction: A Systematic Literature Review. IEEE Access, 8, 166555–166575. DOI: 10.1109/ACCESS.2020.3022349.
[5] Leong, J., and Sung, K. 2019. Predictive Crime Hotspot Detection: A Spatial Analysis Approach. Springer Briefs in Criminology, Springer. DOI: 10.1007/978-3-030-28629-7.
[6] Jiang, X., and Zeng, C. 2019. Spatio-temporal analysis of urban crime leveraging multisource crowd-sensed data. Journal of Urban Computing, 7(4), 320–336. DOI: 10.1007/s10940-019-09234-2.
[7] Kapoor, K., and Gopi, R. 2020. Crime Mapping Using Statistical and Geospatial Methods. Applied Spatial Analysis and Policy, 13, 585–601. DOI: 10.1007/s12061-019-09334-
[8] Nair, V., and Gopi, R. 2019. Machine Learning for Crime Hotspot Prediction: Comparative Techniques and Applications. Advances in Urban Safety Analytics, Springer. DOI: 10.1007/978-3-030-13829-9.
[9] Lai, R., et al. 2022. Comparative Study of Approaches for Detecting Crime Hotspots with Considering Concentration and Shape Characteristics. International Journal of Environmental Research and Public Health, 19, 14350. DOI: 10.3390/ijerph192114350.
[10] ArcGIS StoryMaps. 2023. Examining Crime Mapping Using Hotspots. StoryMaps. Available at https://storymaps.arcgis.com/stories/39186dc6973342d4821bef2e53438a39
[11] Medium. 2023. Crime Rates in Urban Areas: A Data-Driven Visualization. Medium. Available at https://medium.com
[12] BioMed Central. 2021. Spatio-Temporal Crime Hotspots and the Ambient Population. Crime Science. DOI: 10.1186/s40163-021-00145-y.
[13] https://abc7news.com/san-francisco-board-of-supervisors-car-break-ins-police-department-in-epidemic/13811744/?
[14] https://machinelearningmastery.com/how-to-grid-search-sarima-model-hyperparameters-for-time-series-forecasting-in-python/