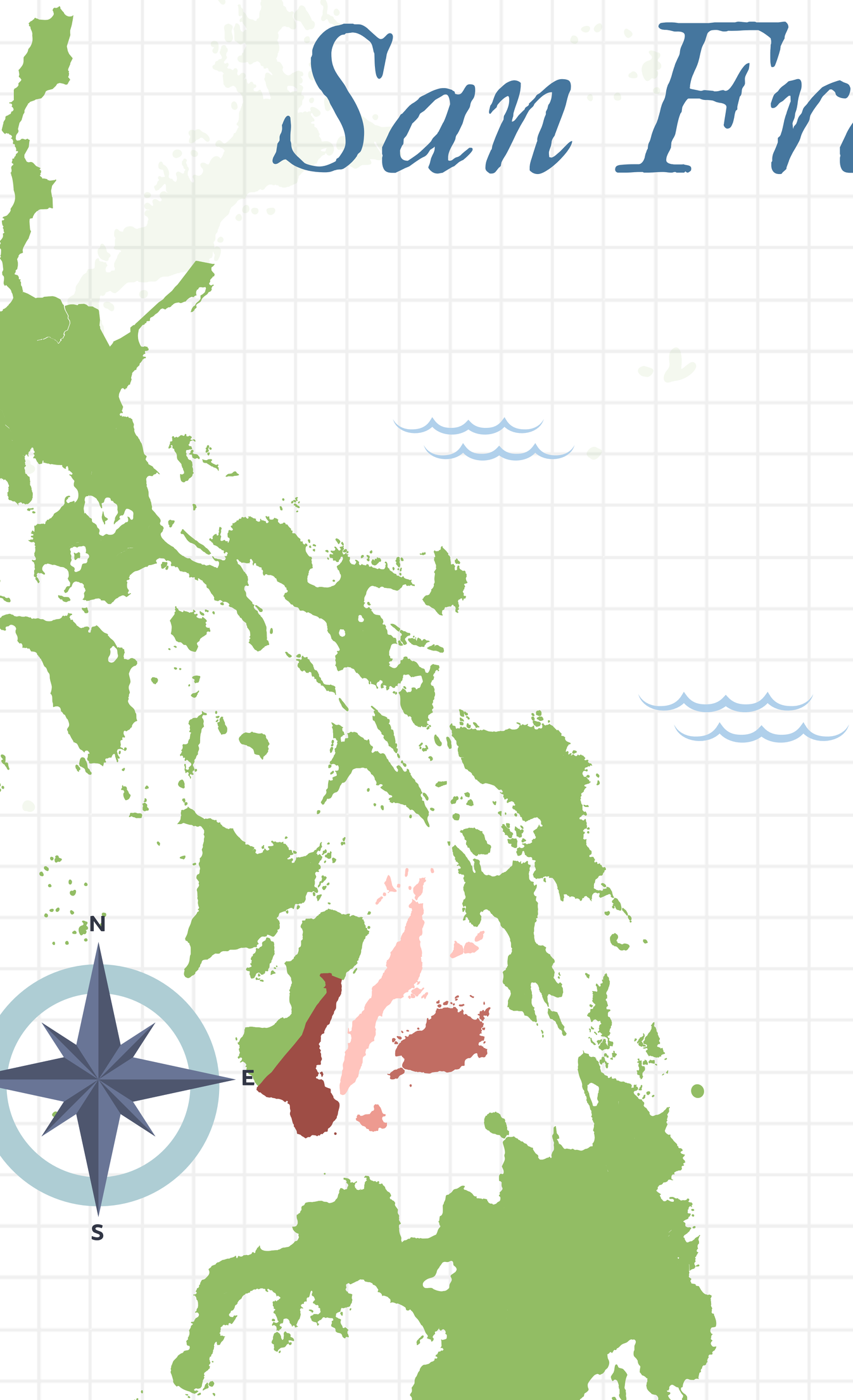
ADVANCED ANALYTICS

Spacial and temporal Crime Analysis in San Francisco



Group 3

Ngocanh Nguyen - BS22DSY029

Dung Nguyen - BS22DSY030

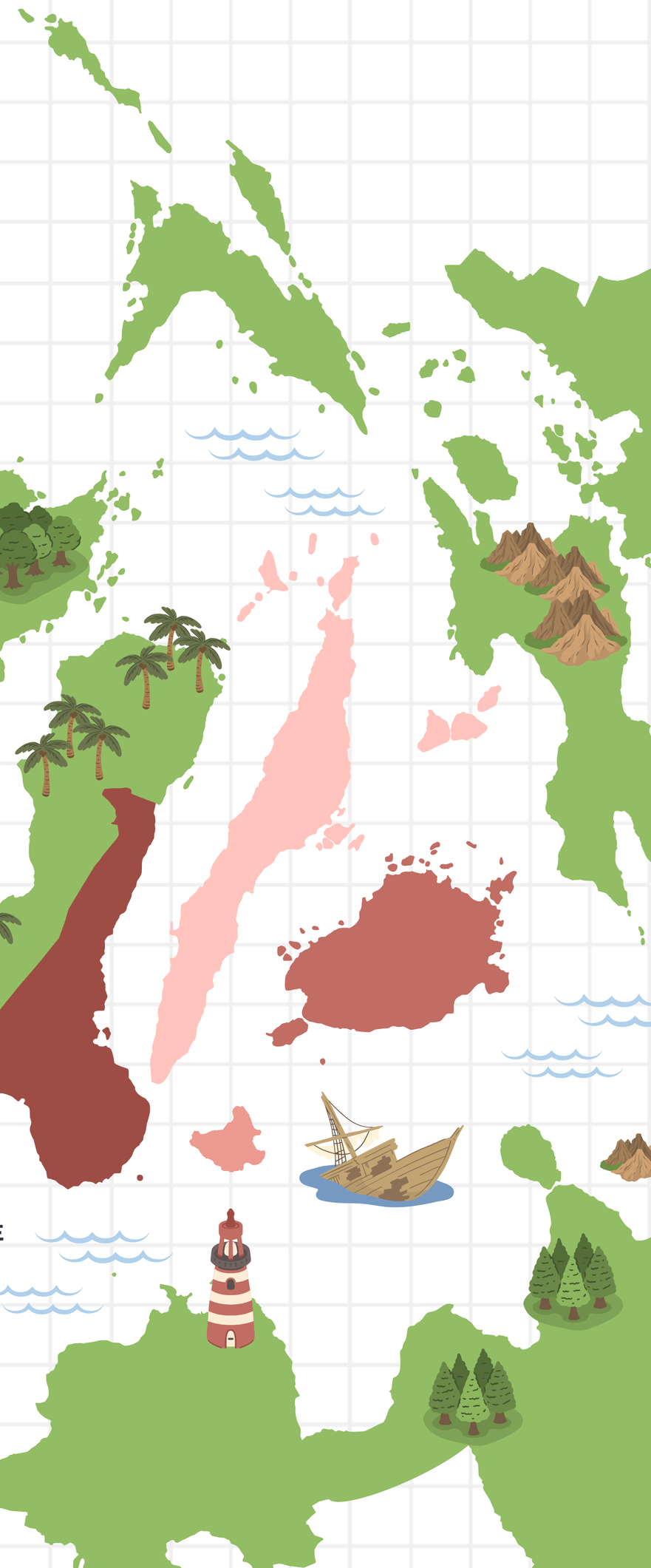
Huyen Truong - BS22DSY032

Contribution

This report, "Advanced Analytics: Spatial and Temporal Crime Analysis in San Francisco," was developed by Huyen Truong, Dung Nguyen, and Ngocanh Nguyen, with each contributing equally to the analysis of crime data from July to December 2012.

- **Huyen Truong**: Data Preprocessing, Exploratory Data Analysis (EDA), and Time Forecasting Analysis – 33.33%
- **Dung Nguyen**: Time Series Visualizations, Spatial Analysis, and Clustering – 33.33%
- **Ngocanh Nguyen**: Spatial Analysis and Predictive Modeling – 33.33%

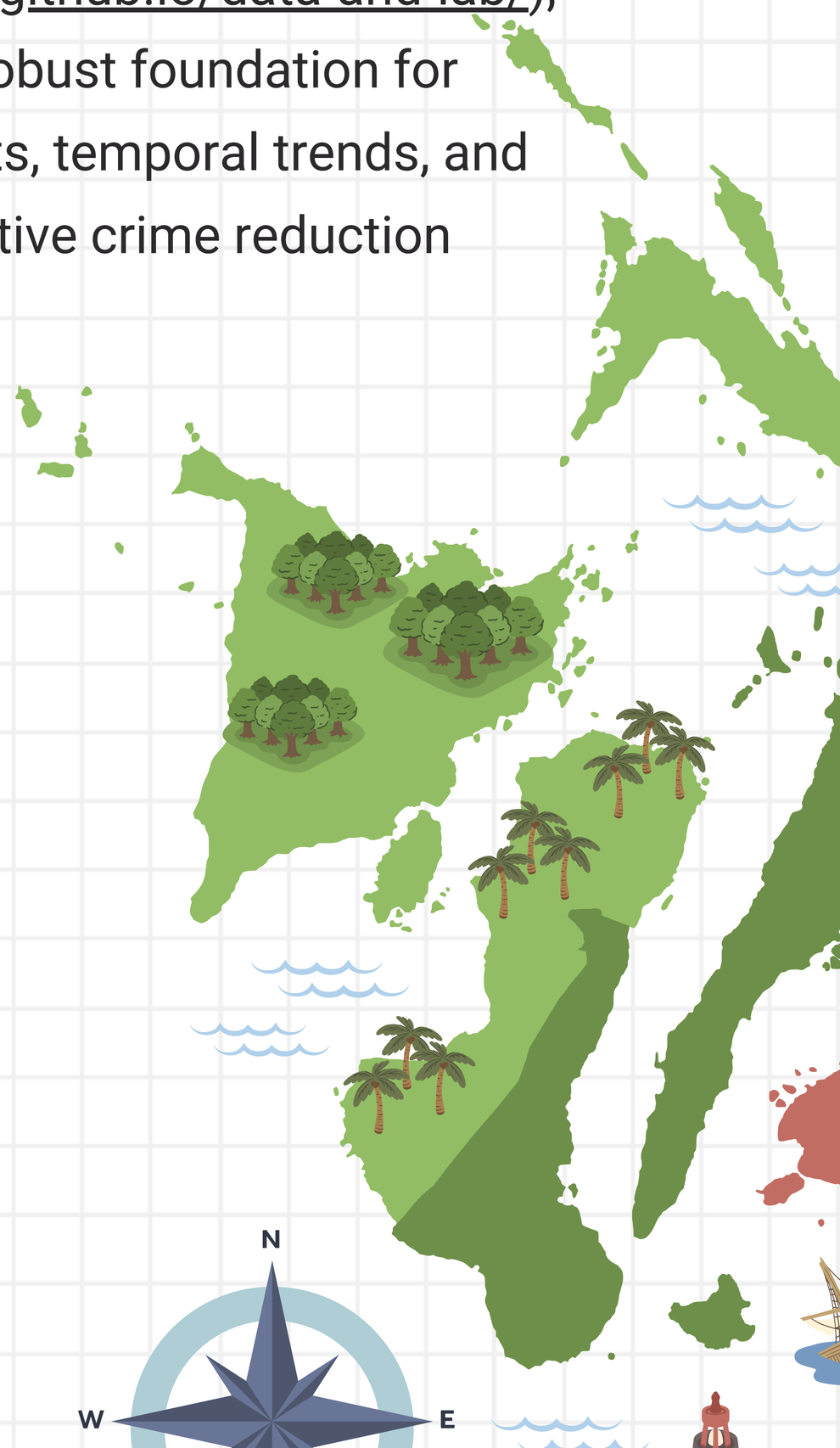
The team collectively cross-checked the work and finalized the report writing together.



Introduction

Crime remains one of the most persistent social challenges confronting urban environments worldwide. Historically addressed through behavioral and sociological approaches, the advent of data analytics has shifted the paradigm toward a more quantitative, evidence-based understanding of crime dynamics. This research analyzes crime data from San Francisco, California, spanning July 1, 2012, to December 31, 2012, focusing on four major crime categories: Vandalism, Vehicle Theft, Robbery, and Drug/Narcotic offenses. The study leverages advanced analytical techniques—exploratory data analysis, time series forecasting, spatial mapping, spatial clustering, spatial analysis and predictive modeling—to uncover hidden patterns and provide actionable insights for optimizing law enforcement resource allocation and enhancing crime prevention strategies.

San Francisco provides a unique case study. Once known for housing notorious criminals on Alcatraz Island, the city has evolved into a technology hub while still contending with crime across its diverse neighborhoods. The dataset, sourced from the GeoDa Data Database (https://geodacenter.github.io/data-and-lab/), contains crime incidents that, after cleaning, offer a robust foundation for analysis. This research aims to identify crime hotspots, temporal trends, and predictive factors to support more targeted and effective crime reduction strategies.



Literature Review

Crime analysis has evolved from traditional behavioral studies to data-driven approaches leveraging modern analytics. Weisburd et al. (2006) demonstrated that crime concentrates in specific geographic "hotspots," a concept central to spatial analysis. Time series forecasting, as explored by Chen et al. (2017), has been used to predict crime trends, while predictive modeling, such as classification algorithms, has improved crime type prediction (Wang et al., 2019). These studies underscore the value of combining spatial, temporal, and predictive techniques—approaches adopted in this analysis. The San Francisco crime dataset provides a rich opportunity to apply these methods, building on prior work to derive localized insights.

Methodology

Dataset and Pre-processing

The dataset for this analysis was acquired from the GeoDa Data Database (https://geodacenter.github.io/data-and-lab/), containing crime incidents reported in San Francisco from July to December 2012. The original data was stored in separate shapefiles for each crime category ("sf_drugs.shp," "sf_cartheft.shp," "sf_vandalism.shp," "sf_robbery.shp"). These files were merged into a single geodataframe (combined_gdf) for comprehensive analysis.

Initial data exploration revealed several issues requiring preprocessing:

1. The dataset contained 13,472 rows and 14 columns initially
2. The Time column contained only a single redundant value (1899-12-30 00:00:00) and was removed
3. 3,319 duplicate records were identified and removed

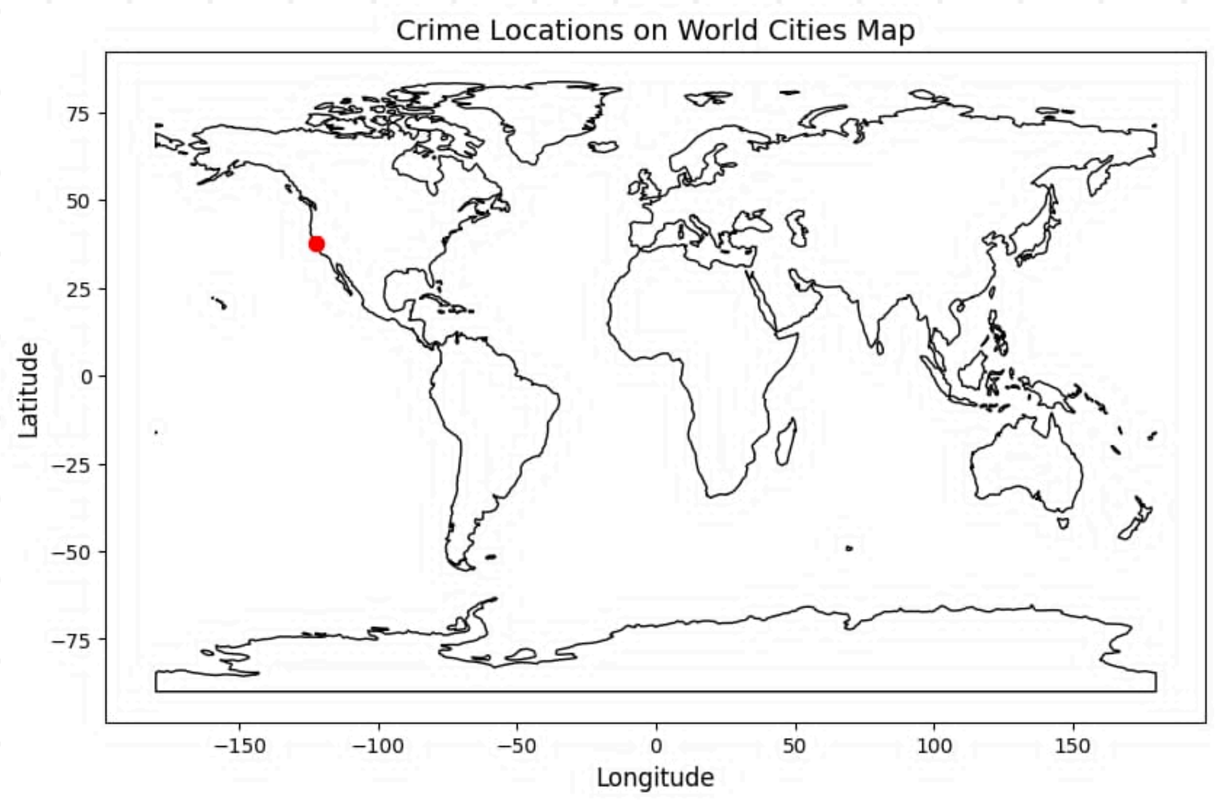
After addressing the variables, missing values, duplicates, and redundancy, the cleaned dataset now consists of 10,153 rows and 11 columns.

Evaluate the position of the data points using the coordinates to identify any misplaced points

The coordinate reference system was identified as EPSG:2227 (NAD83 / California zone 3), which uses feet as units. For certain analyses, the coordinates were converted to the more standard WGS84 (EPSG:4326) system to facilitate mapping and spatial analysis.

The geographical position of the data points was evaluated to identify any misplaced coordinates. The expected coordinate range for San Francisco was defined based on the central latitude and longitude values (37.773972, -122.431297) from https://www.latlong.net/place/san-francisco-ca-usa-594.html. Data points falling outside the defined bounds were flagged as potential errors or outliers, ensuring that the data points are correctly located within the geographical area of interest.

No data points fall outside the expected geographical range for San Francisco.



Analytical Techniques

1. Exploratory Data Analysis (EDA): Summary statistics, categorical distributions, and visualizations (e.g., bar charts) were used to understand crime patterns.

2. Time Series Analysis: Crime patterns were explored across different temporal granularities: Daily trends with anomaly detection using Z-scores (threshold=2.5), Day-of-week patterns, Monthly trends, Moving averages (7-day) to identify underlying patterns, and decomposition (trend, seasonality, residuals) were analyzed, followed by forecasting with ARIMA, Holt-Winters, and Facebook Prophet models.

3. Spatial Analysis: Heatmaps, choropleth maps (Natural Breaks and Fisher-Jenks), spatial lag (Queen contiguity), and clustering (K-Means, Hierarchical) were applied using GeoPandas and LibPySal.

4. Predictive Modeling: Features were engineered (e.g., Day, Block), and a tree-based classifier was trained to predict crime categories based on time and location.

Tools

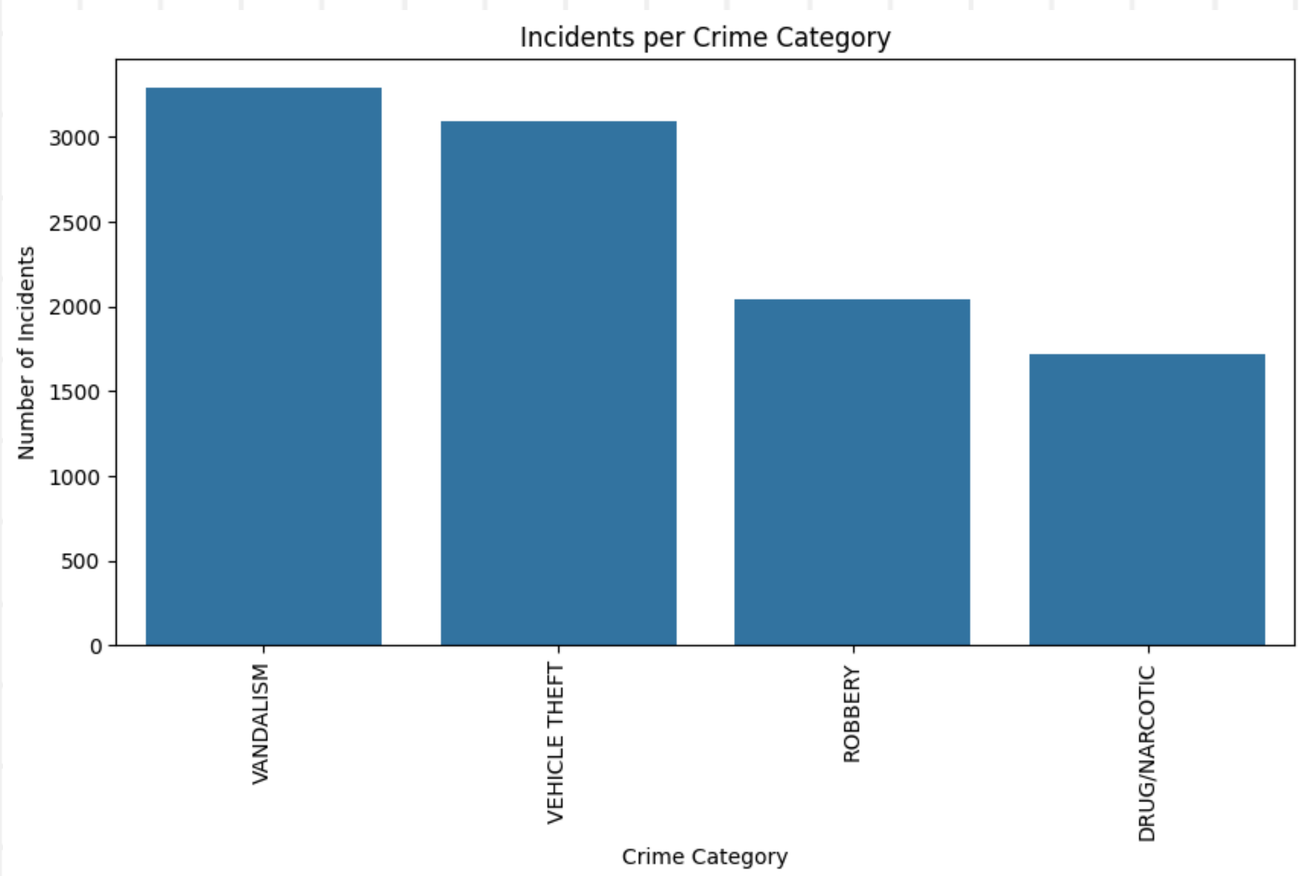
The analysis was implemented in Google Colab, leveraging a comprehensive suite of Python libraries to handle data processing, visualization, statistical modeling, and spatial analysis. These included Pandas for data manipulation, NumPy for numerical computations, Matplotlib and Seaborn for plotting visualizations, Scikit-learn for machine learning and statistical tools, GeoPandas and LibPySal for geospatial analysis, Folium for interactive mapping, Statsmodels for time series decomposition, , and additional utilities like sklearn.model_selection and sklearn.metrics to support robust model training and evaluation.

Results and Discussion

Dataset Exploration

Incidents per crime category

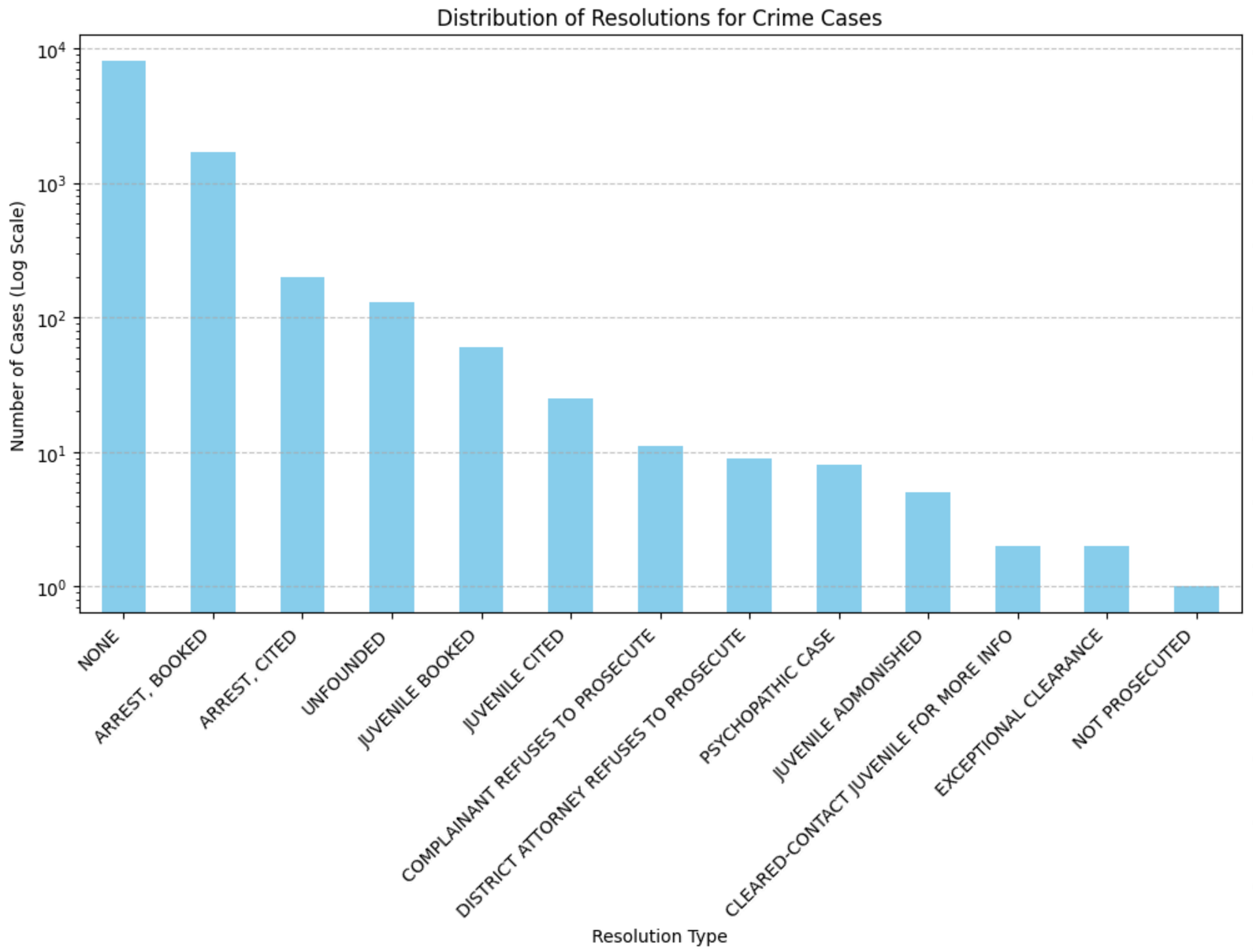
The cleaned dataset revealed Vandalism incidents as the most frequent (3,897 cases), followed by Vehicle Theft and Robbery.



Distribution of Resolutions for Crime Cases

Since "NONE" has a significantly higher count than the others, using a log scale on the y-axis to make the differences clearer.

The majority of cases (8,018 cases) have no resolution ("NONE"), which could indicate unsolved or ongoing cases. "Arrest, Booked" (1,682 cases) is the most common resolution where action was taken. Other resolutions like "Juvenile Booked" (60 cases) or "District Attorney Refuses to Prosecute" (9 cases) occur much less frequently

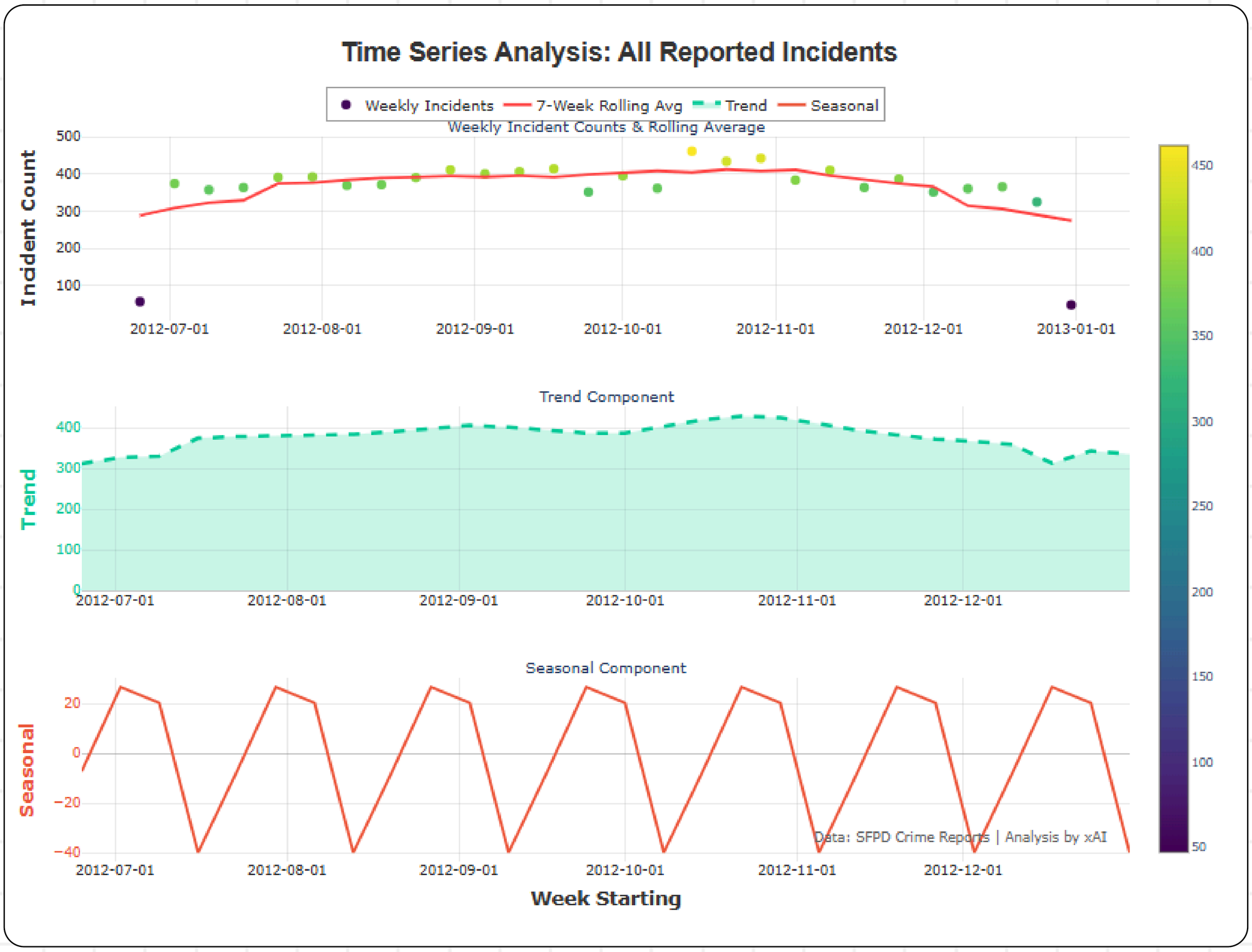


Time Series Analysis

Time series analysis was conducted to explore temporal patterns in the San Francisco crime dataset from July 1 to December 31, 2012. This section examines daily, weekly, and monthly trends, identifies anomalies, and smooths data to reveal underlying patterns, providing a foundation for forecasting and actionable insights.

Time Series Overview

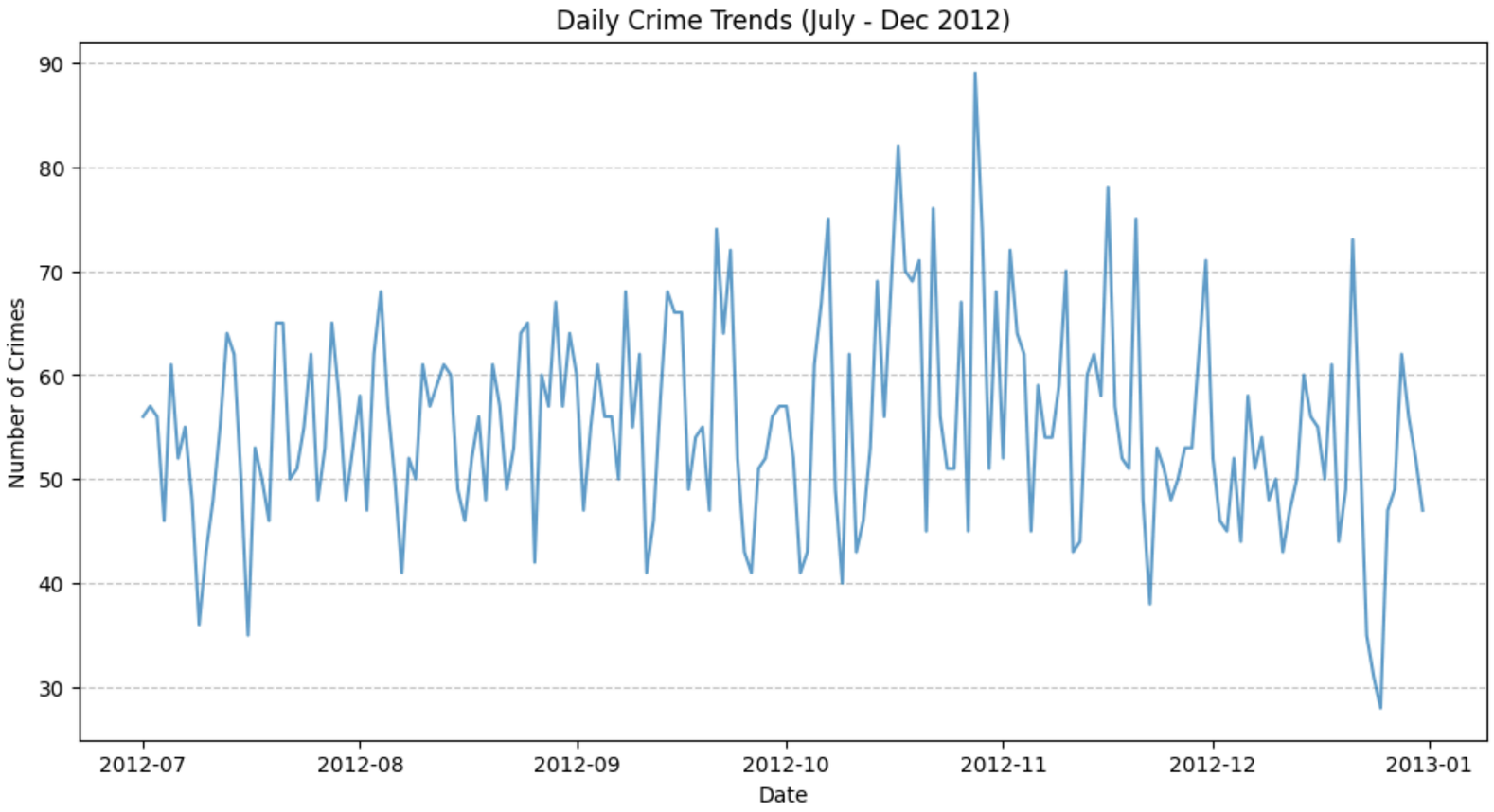
We begin this section with a targeted analysis, presenting a time series breakdown of all reported incidents in San Francisco through three subplots: weekly incident counts with a rolling average, a trend component, and a seasonal component. This visualization provides a clear perspective on the patterns and trends in the incident data.



- The "Weekly Incident Counts & Rolling Average" subplot shows weekly counts as green markers (200–500 incidents), with a red 7-week rolling average. Counts rise from July to a peak of 500 in November, then drop to 300 by December. The Viridis color scale emphasizes intensity, with darker shades for lower counts.
- The "Trend Component" subplot, with a dashed cyan line and light fill, confirms the upward trend from 300 incidents in July to 400 in late 2012, then a decline to 300 by December, filtering out short-term noise.
- The "Seasonal Component" subplot, with a red line, reveals a 4-week cycle, fluctuating between -20 and +20. Dips in mid-August, mid-September, and mid-December suggest monthly influences from social, economic, or environmental factors.

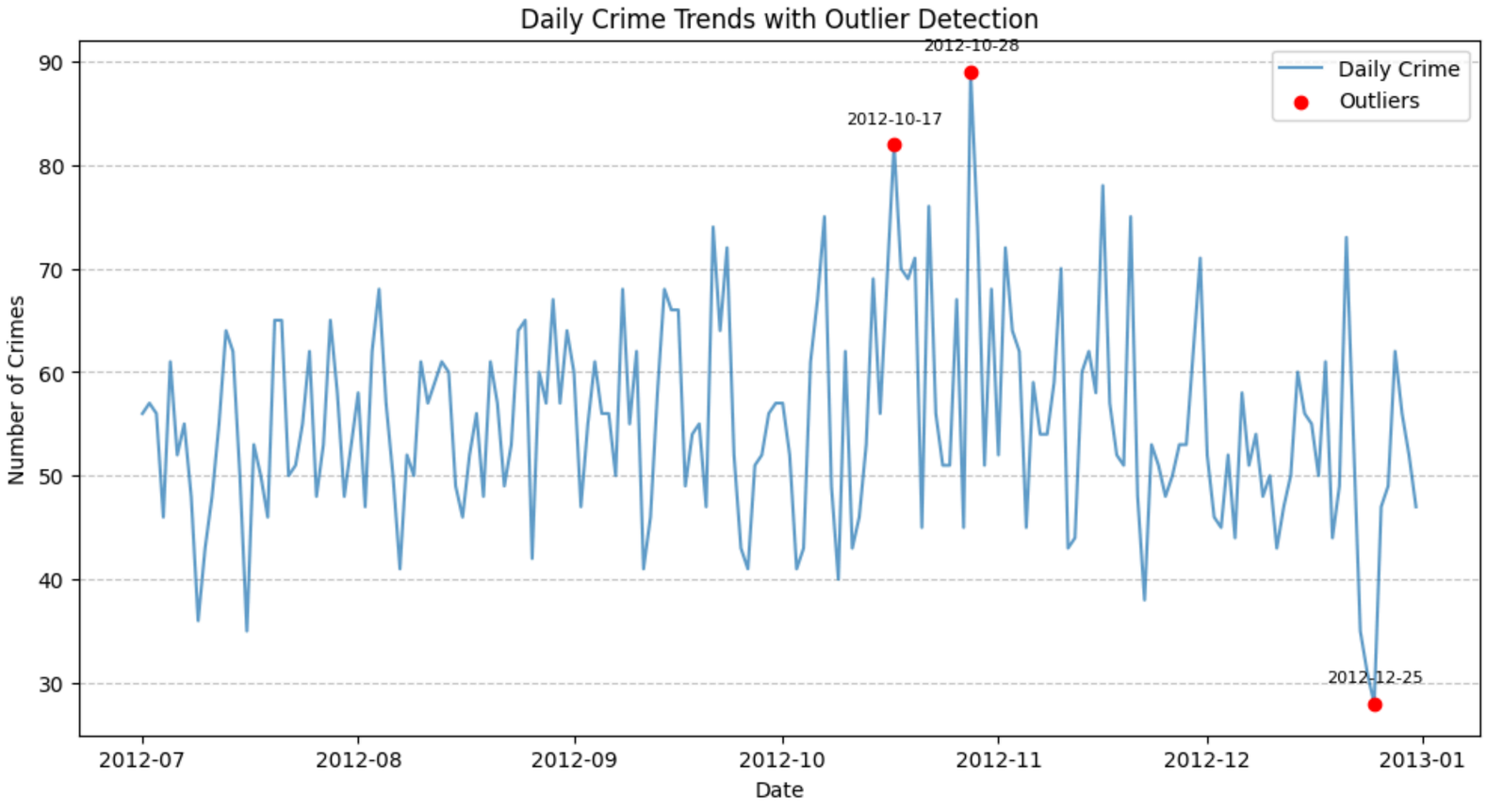
Crime Trends Over Time (Daily)

The graph clearly shows a fluctuating pattern throughout the six-month period, indicating that the number of crimes varies significantly from day to day. The noticeable spike around mid-October marks the highest crime volume, while a noticeable decrease occurs towards late December 2012. These observations suggest potential influences such as seasonal factors or specific events, warranting further investigation.



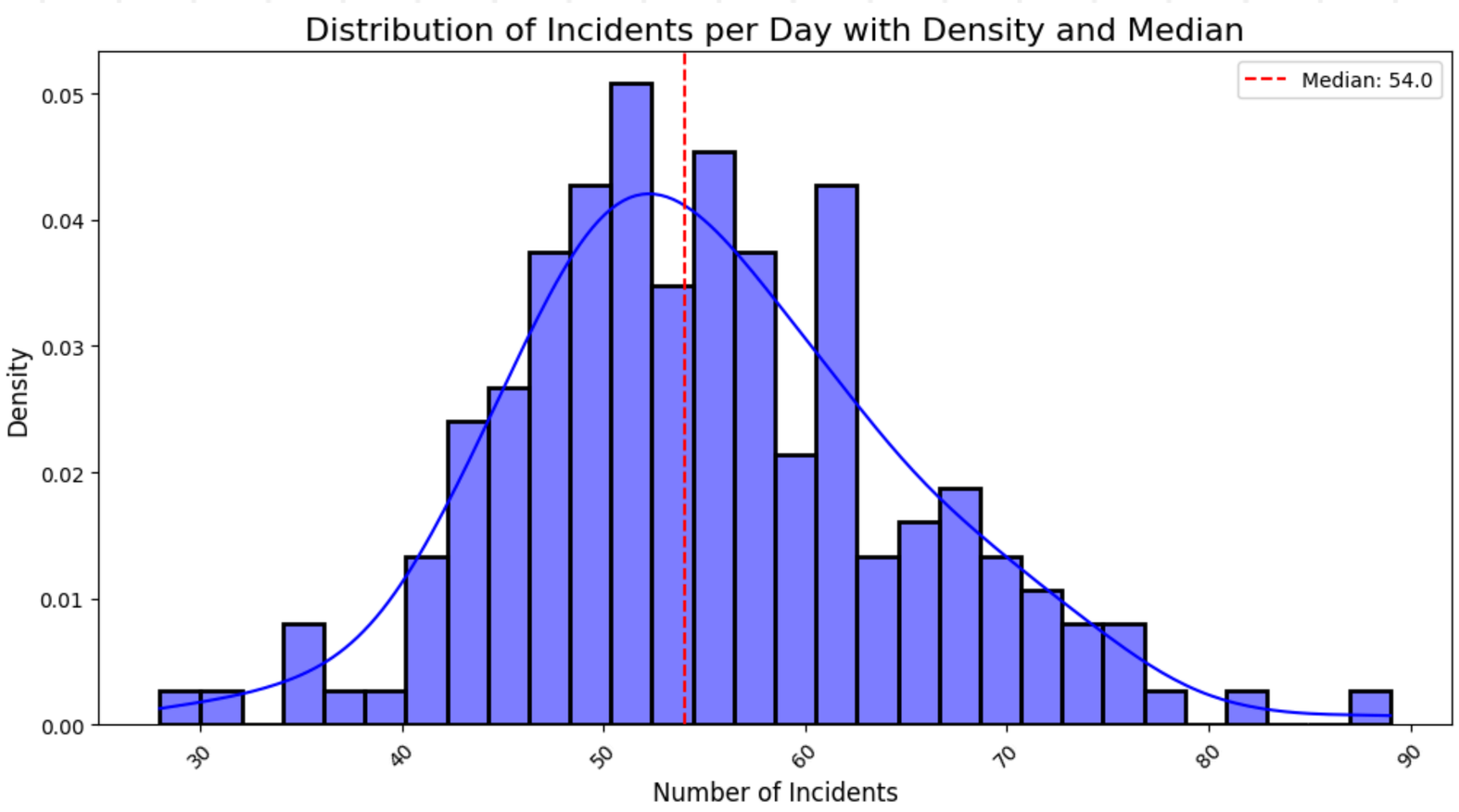
Anomaly Detection (Identifying Unusual Crime Spikes)

To identify significant deviations, anomaly detection was performed using a Z-score threshold of 2.5. This method flagged September 14, 2012, and October 17, 2012, as unusually high-crime days, and December 25, 2012, as a significant dip. These outliers may reflect external triggers—e.g., holidays for the December dip or unreported events for the September and October spikes—requiring contextual investigation beyond the dataset's scope.



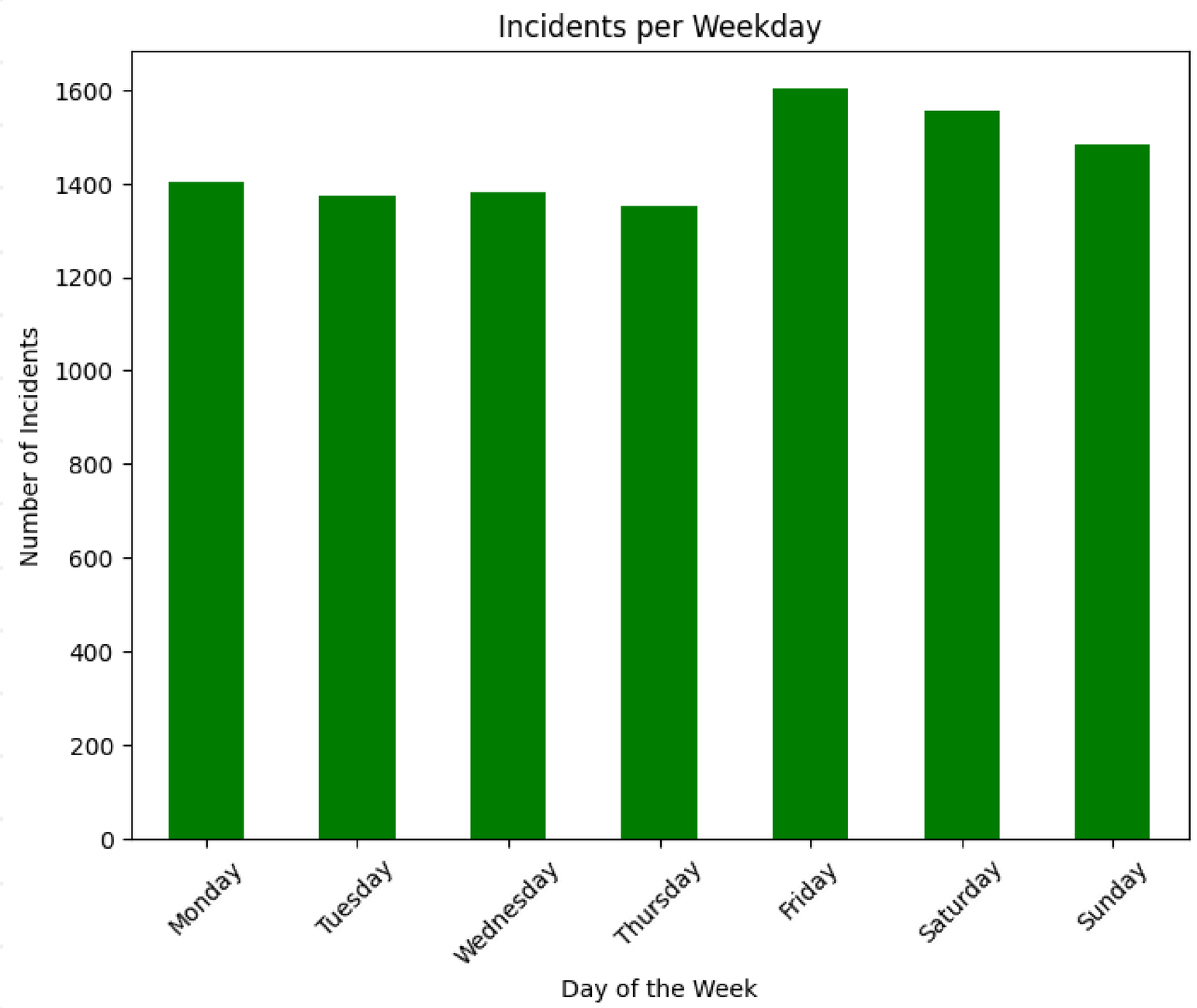
Distribution of Incidents per Day with Density and Median

The distribution of daily incident counts approximates a normal curve, centered around a median of 54 incidents per day. A slight right skew indicates occasional days with elevated counts, consistent with the identified anomalies. This distribution provides a baseline for understanding typical crime volumes and deviations.



Crime Trends by Day of the Week

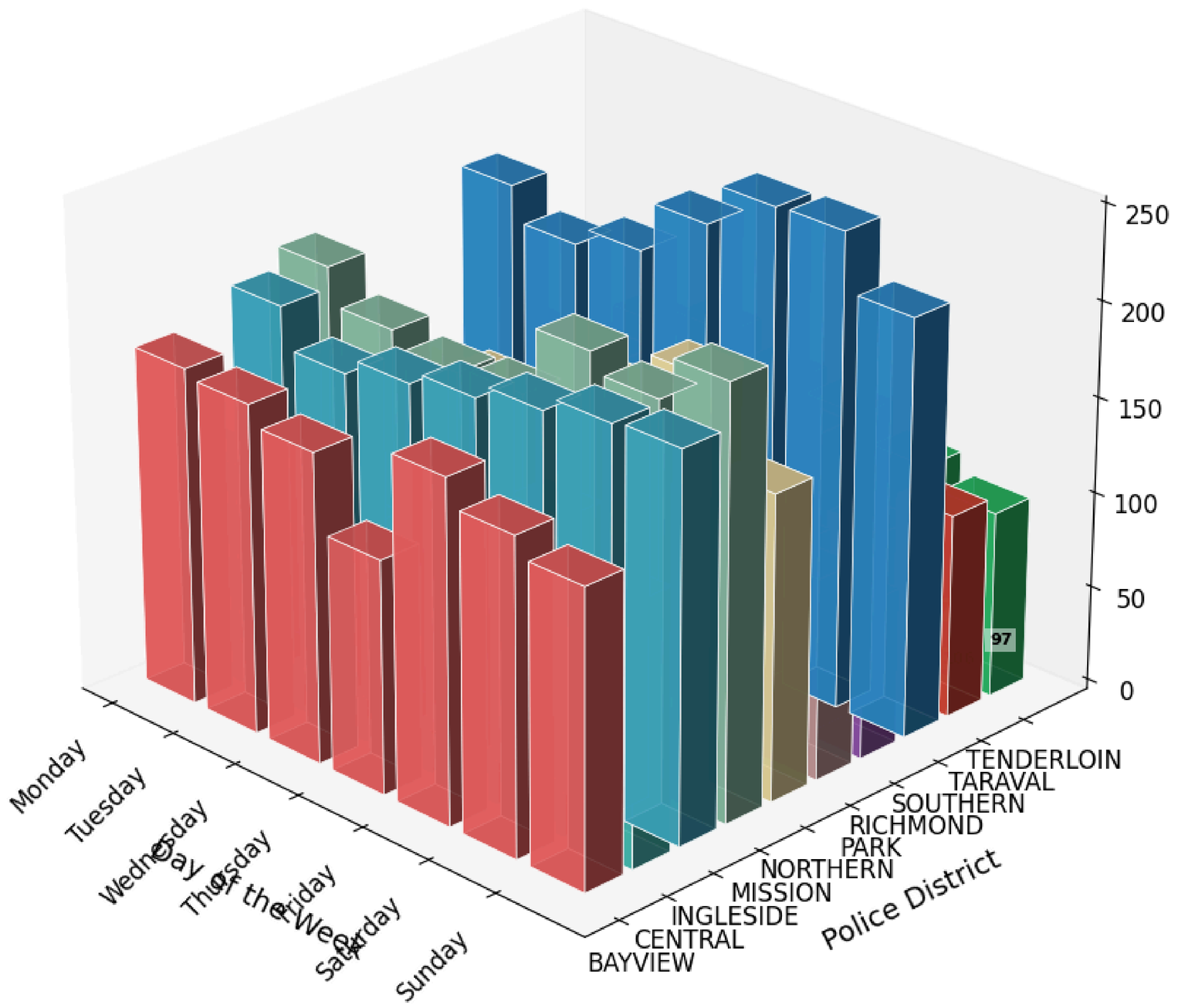
Breaking down incidents by day of the week reveals Friday as the peak day for crime, with a slight uptick on Saturday and Sunday compared to other weekdays. This pattern suggests a weekend effect, possibly linked to social activities or reduced policing, though the increase is modest.



Crime Trends by Day of the Week and Police District

Having identified Friday as the peak day for crime with a modest weekend uptick in the "Crime Trends by Day of the Week" analysis, it's crucial to explore how these patterns vary across different areas.

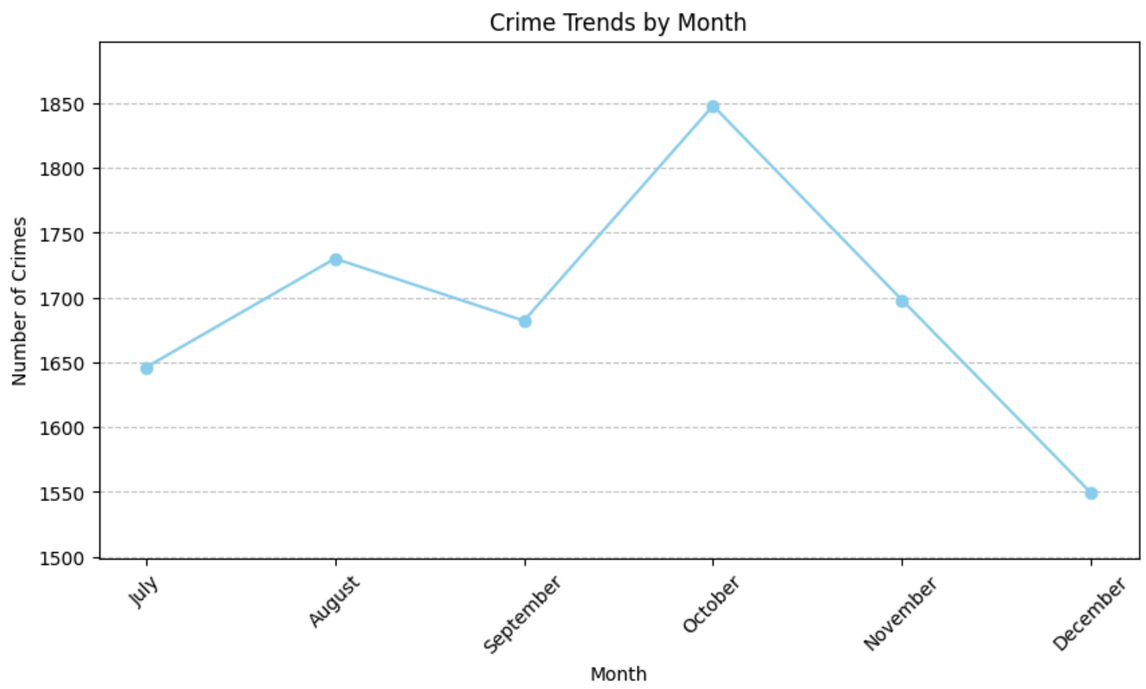
3D Incidents by Day of the Week and Police District

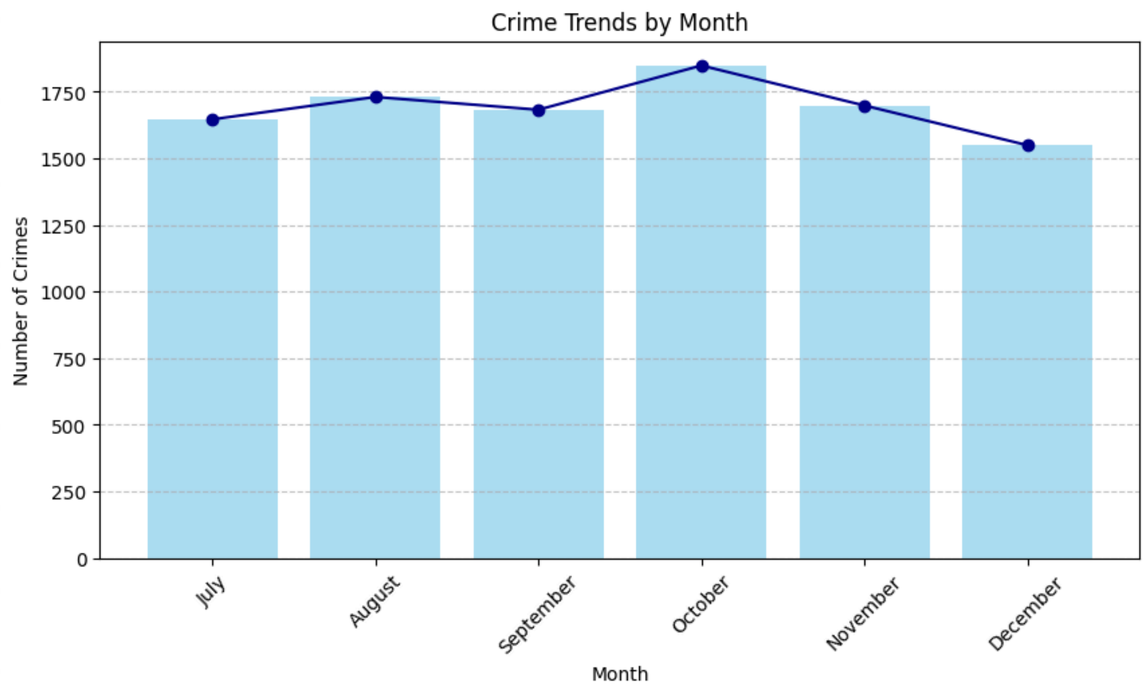


The 3D chart reveals that Tenderloin, Southern, and Central districts experience the highest incident rates, with Tenderloin peaking at 97 incidents on Saturday, indicating a weekend hotspot. Saturday stands out as the most incident-heavy day across multiple districts, while Sunday shows lower activity. Districts like Bayview, Taraval, and Richmond report fewer incidents, suggesting safer areas or underreporting. Midweek spikes on Wednesday and Friday in Southern and Central may tie to commuting or events, highlighting the need for targeted policing in high-incident districts during peak times. Adding a color legend and adjusting bar transparency could enhance clarity.

Crime Trends by Month

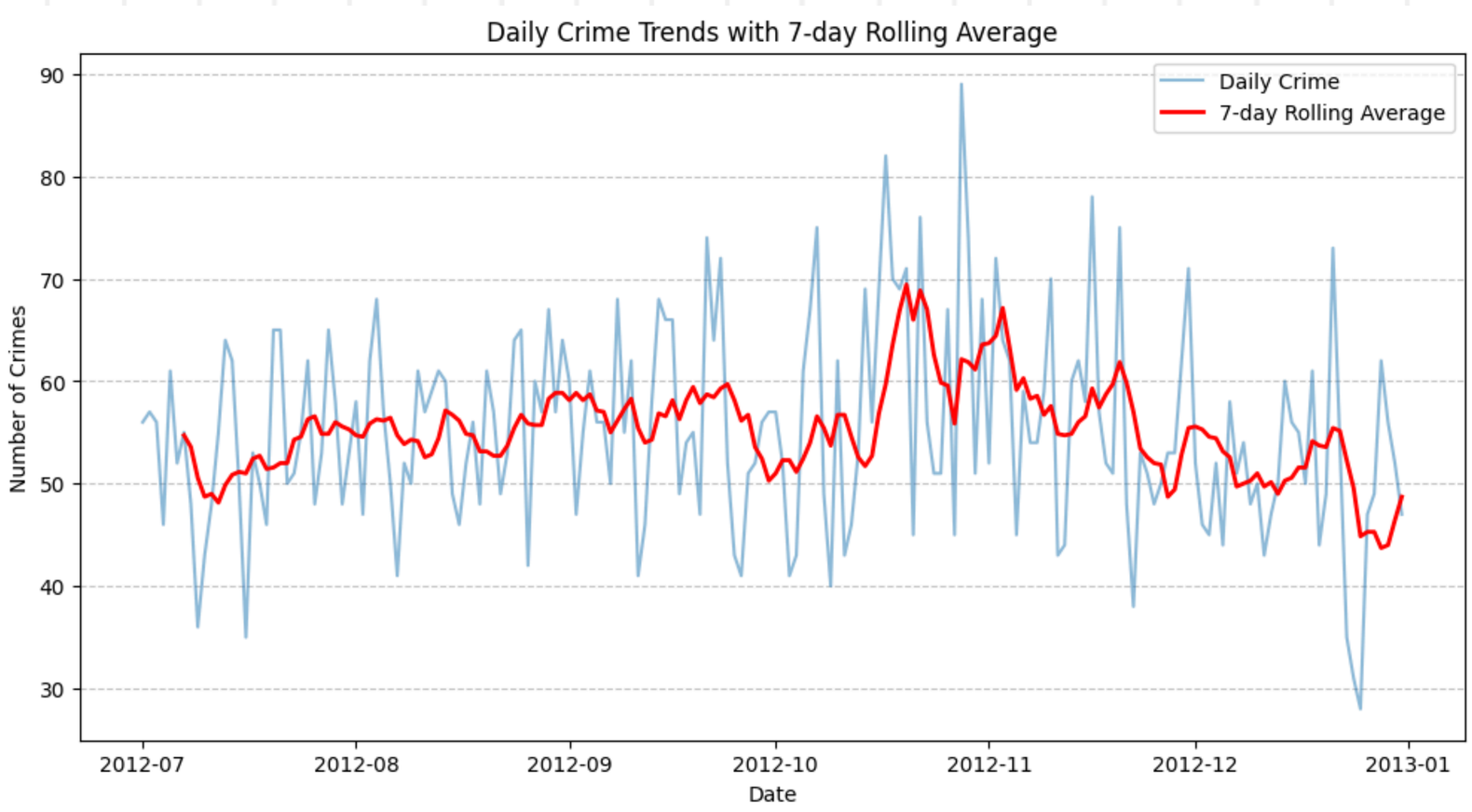
Monthly analysis shows October as the peak month for crime, followed by a downward trend through December. To emphasize relative changes, the y-axis does not start at zero, focusing attention on trends rather than absolute counts. This decline may reflect seasonal influences, such as holiday-related behavioral shifts, or end-of-year reporting variations.





Moving Average for Crime Trends (Smoothing)

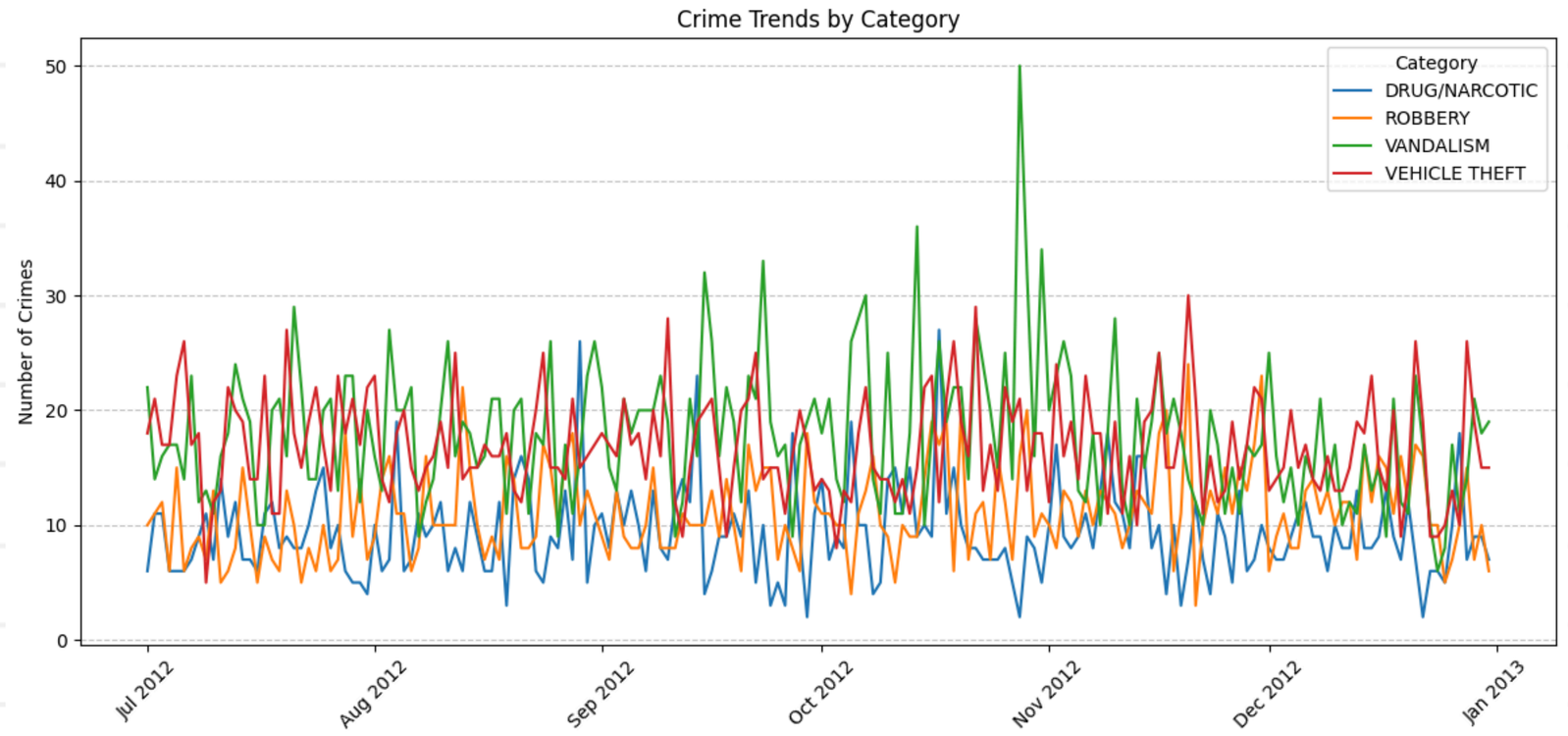
A 7-day rolling average was applied to smooth daily fluctuations and reveal broader trends. While raw daily data exhibits significant volatility, the smoothed trendline indicates a relatively stable baseline with a gradual rise culminating in a late October/early November peak. This approach minimizes noise, offering a clearer picture of underlying patterns.

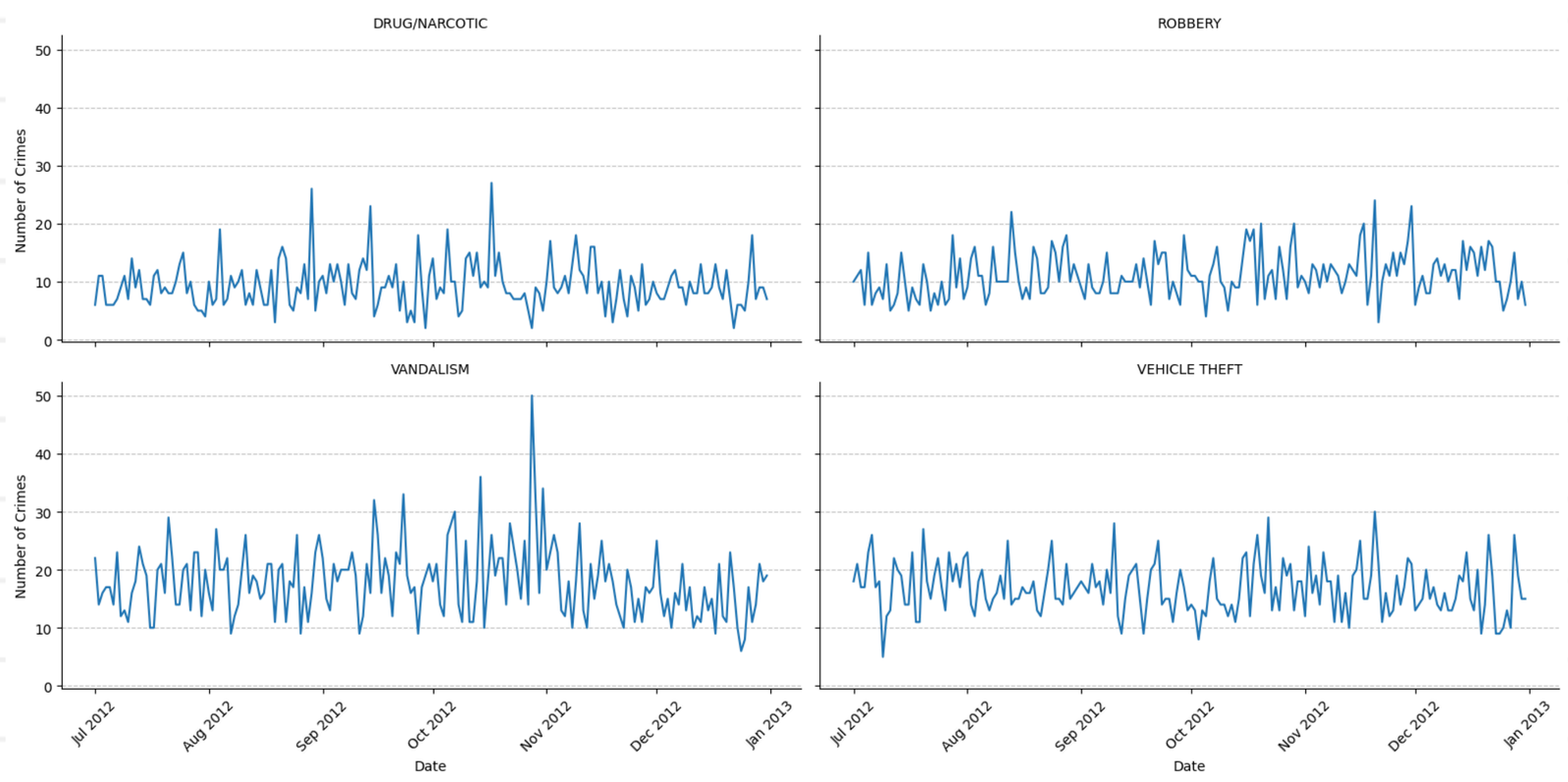


Comparing Crime Trends by Category

Crime trends were further analyzed by category—Drug/Narcotic, Vehicle Theft, Robbery, and Vandalism. All categories fluctuate, but Vandalism exhibits the most pronounced variability. Drug/Narcotic and Vandalism peak sharply around October, aligning with the overall trend, while Robbery and Vehicle Theft remain comparatively lower and more stable.

Daily crime trends fluctuated, with a notable spike in mid-October and a dip on December 25, 2012. Crime trends by category are integrated in a single line chart, while a 2x2 grid of separate line charts provides a detailed breakdown, highlighting distinct category-specific patterns.



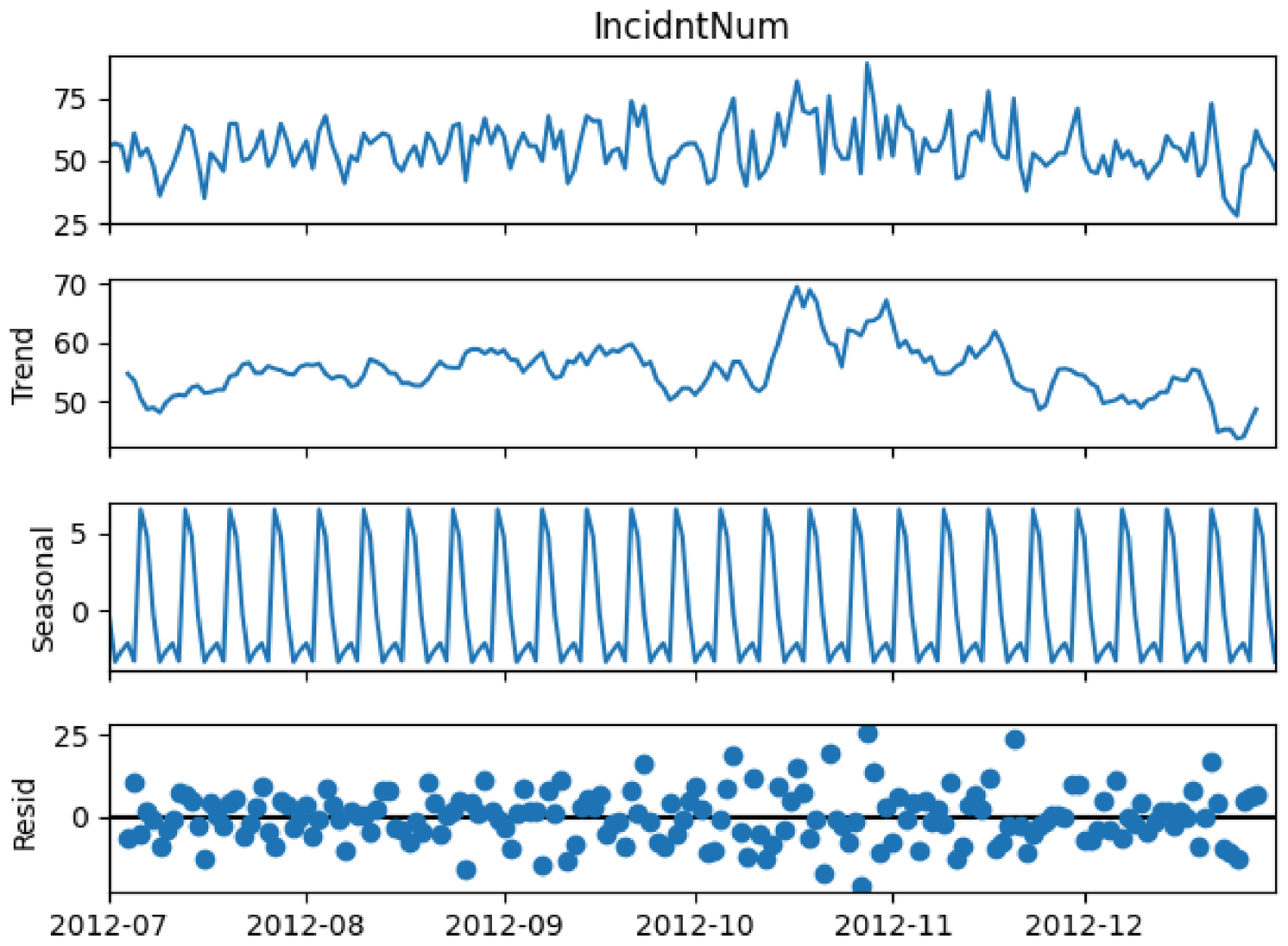


Time Series Forecasting

Time series forecasting was employed to predict future crime incidents based on historical patterns in the San Francisco dataset from July to December 2012. Given the dataset's five-month span, capturing long-term trends or strong seasonal effects is limited, so the focus shifted to short-term daily and weekly fluctuations. Data preparation involved selecting 'Date' and 'IncidntNum' columns, aggregating incidents daily, and setting 'Date' as the index. This structured approach ensures that the dataset is appropriately formatted for time series decomposition and forecasting..

Time Series Decomposition

To better understand the underlying patterns in the data, we performed a time series decomposition to separate the trend, seasonality, and residual components. Despite the relatively short time frame, the decomposition reveals a clear weekly seasonal pattern, as evidenced by the repeating peaks and troughs in the seasonal component. This suggests that the number of incidents follows a weekly cycle, likely influenced by external factors such as workdays, weekends, or reporting behaviors.

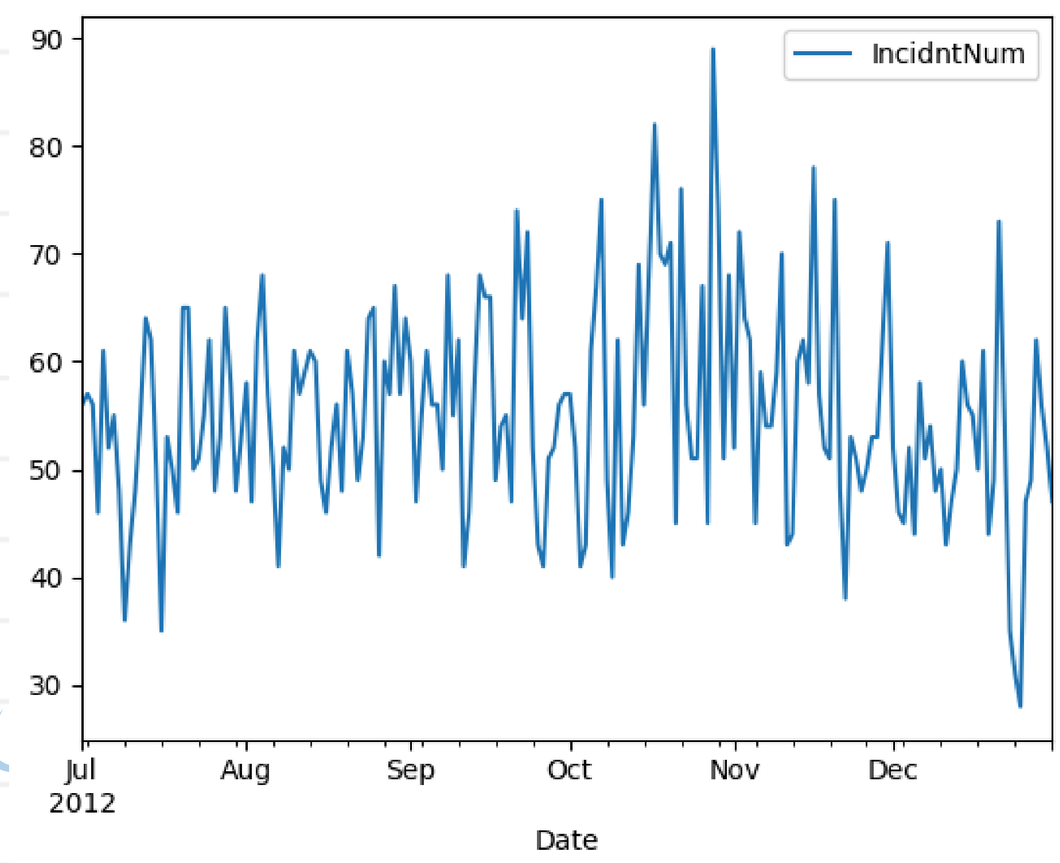


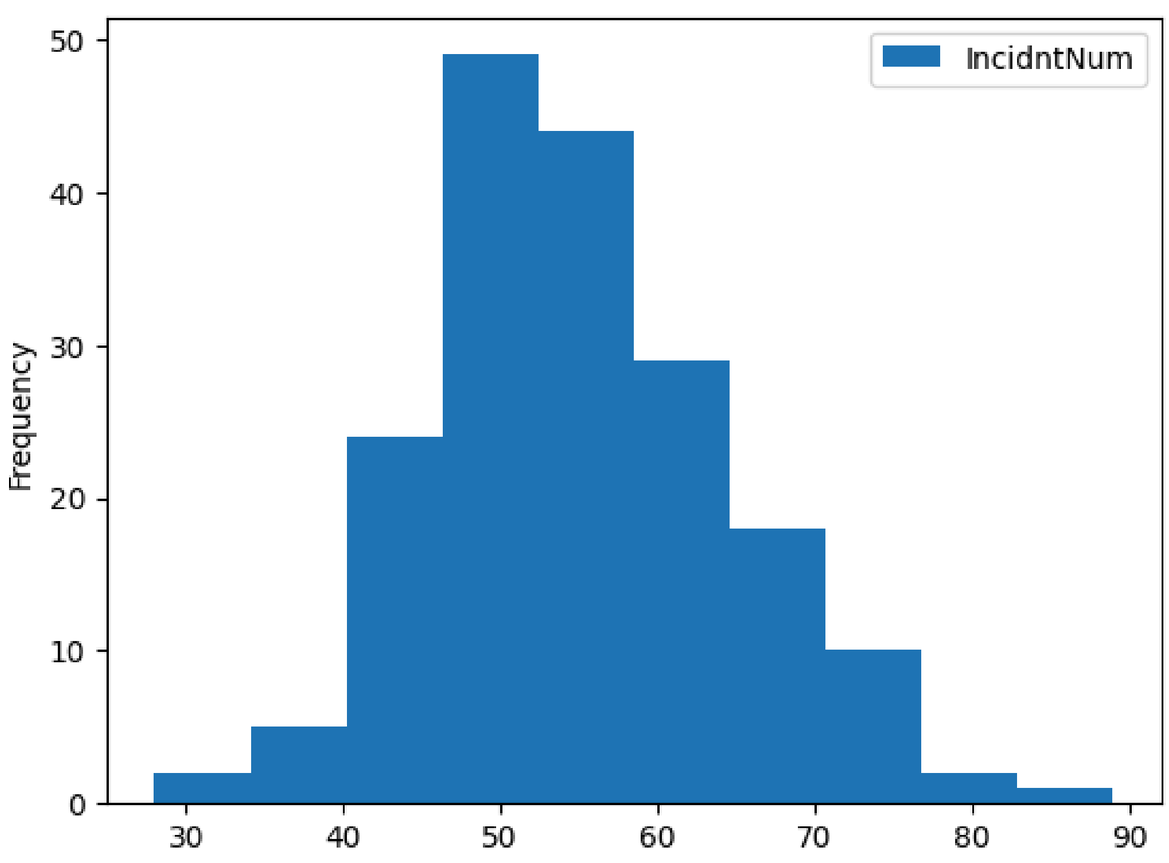
Stationarity Check

Before building forecasting models, we assessed whether the dataset was stationary—a crucial property for many time series models. We applied four different tests to evaluate stationarity: Visual plot test, summary statistic, AD Fuller Test. and KPSS test.

1. Visual Plot Test

A histogram of the incident data was plotted to observe its distribution. While the histogram slightly deviates from a perfect Gaussian bell curve, it remains relatively centered, suggesting potential mean stationarity. This observation is further validated by statistical tests.





2. Summary Statistics (Mean and Variance Comparison)

The dataset was split into two halves to compare their means and variances. The means were found to be relatively close, while the variances exhibited some differences. However, the overall stability of the mean, reinforced by the results from the ADF and KPSS tests, suggests that the dataset is stationary.

```
mean1=54.97826086956522, mean2=55.380434782608695
variance1=62.10822306238186, variance2=127.88787807183365
```

3. Augmented Dickey-Fuller (ADF) Test

The ADF test is used to detect the presence of a unit root, which indicates non-stationarity. The test was conducted with the following hypotheses:

- Null hypothesis (H_0): The series has a unit root (non-stationary).
- Alternative hypothesis (H_1): The series is stationary.

The test statistic was found to be significantly lower than the critical values, leading to the rejection of the null hypothesis. This confirms that the dataset is stationary.

```
Results of Aug. Dickey-Fuller Test:
Test Statistic                  -1.042171e+01
p-value                          1.697915e-18
#Lags Used                       0.000000e+00
Number of Observations Used      1.830000e+02
Critical Value 1%               -3.466598e+00
Critical Value 5%               -2.877467e+00
Critical Value 10%              -2.575260e+00
dtype: float64
-10.421714527848824 -3.466598080268425
Stationary
```

4. KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test

The KPSS test provides a complementary check for stationarity. The results indicated that the test statistic was lower than all critical values, further confirming that the dataset is stationary.

```
Results of KPSS Test:
Test Statistic          0.286101
p-value                 0.100000
Lags Used               3.000000
Critical Value 10%      0.347000
Critical Value 5%       0.463000
Critical Value 2.5%     0.574000
Critical Value 1%       0.739000
dtype: float64
Stationary (Null Hypothesis Accepted)
<ipython-input-127-cc44d72aac11>:6: InterpolationWarning: The test statistic is outside of the range of p-values available in the
look-up table. The actual p-value is greater than the p-value returned.

  kpsstest = kpss(timeseries, regression='c', nlags="auto") # Added nlags="auto"
<ipython-input-127-cc44d72aac11>:22: InterpolationWarning: The test statistic is outside of the range of p-values available in the
look-up table. The actual p-value is greater than the p-value returned.

  kpss_results = kpss(df['IncidntNum'], regression='c', nlags="auto")
```

Splitting the Data into Training and Testing Sets

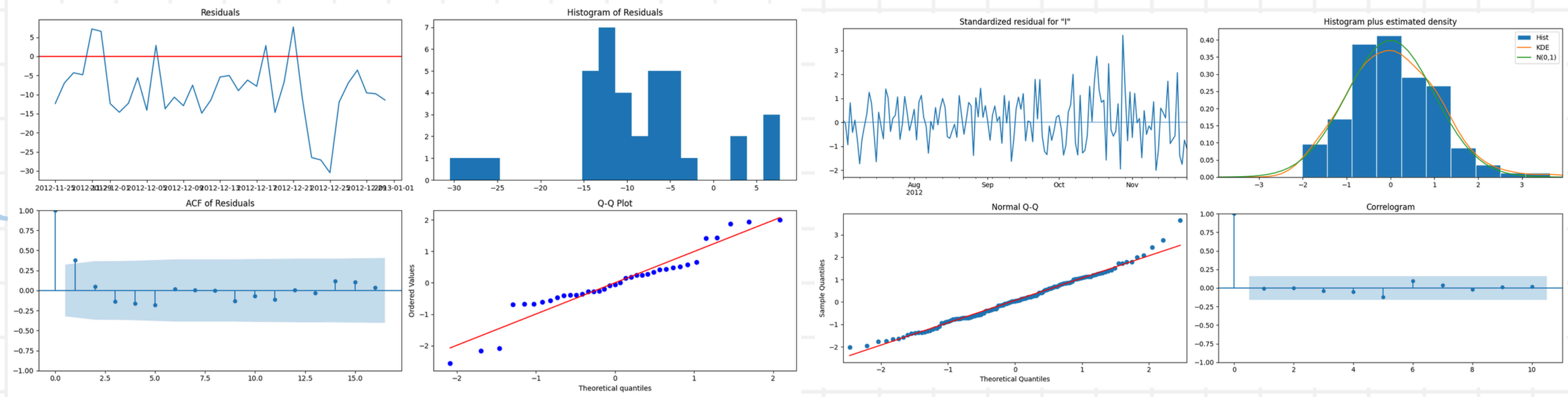
The dataset was divided into 80% training and 20% testing subsets to fit models and assess forecasting accuracy, ensuring reliable performance evaluation.

Model Development and Forecasting

Three different forecasting models were evaluated to predict the number of incidents- **Holt-Winters Exponential Smoothing, ARIMA, and Facebook Prophet**—. To determine the best-performing model, we compared the following performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE) metrics. After comparing the results, ARIMA demonstrated the best performance across all evaluation metrics with RMSE of 11.05, MAE of 9.20, and MAPE of 20.95%, despite fit limitations across all models due to the short data span. While all models had limitations in fit, ARIMA consistently provided the most accurate forecasts for this dataset. Therefore, ARIMA is selected as the recommended model for future incident number predictions.

Model Diagonstics

Residual diagnostics were conducted for all three models, including plots for randomness, histograms for normality, autocorrelation function (ACF) for correlation, and Q-Q plots for distribution, as shown in the model diagnostics charts. For ARIMA, residuals exhibited the closest approximation to a normal distribution, with minimal autocorrelation and reasonable randomness, reinforcing its suitability for forecasting. However, imperfections remain, as the residuals are not perfectly normal, reflecting the dataset's limited duration. In contrast, Facebook Prophet's diagnostics revealed less favorable outcomes. While Prophet effectively captures weekly seasonality and the general trend in the training data, its residuals show greater deviation from normality and higher autocorrelation, indicating a weaker fit for this specific dataset.

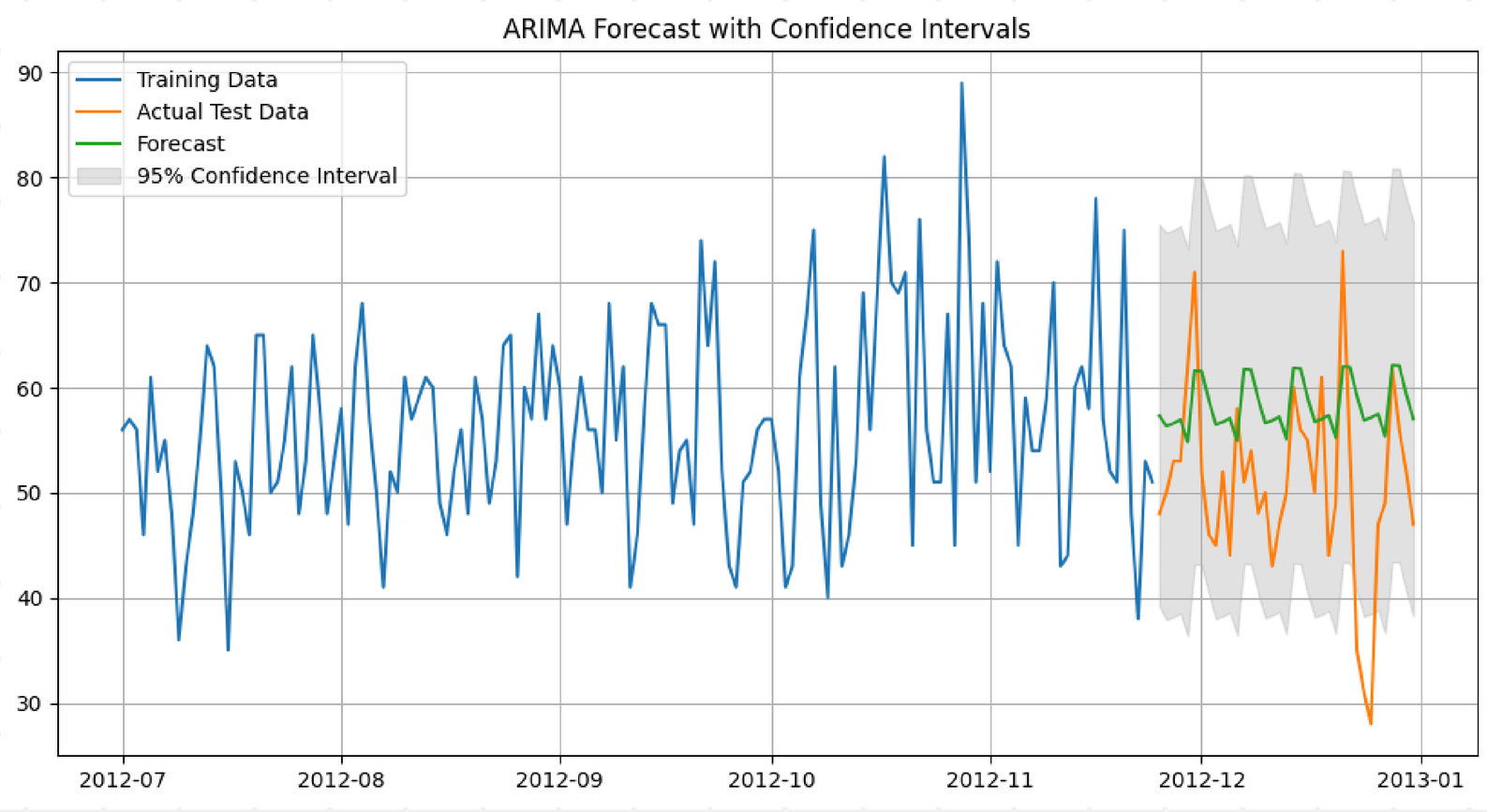


Facebook Prophet model (left) - ARIMA model (right)

Forecasting Future Incidents

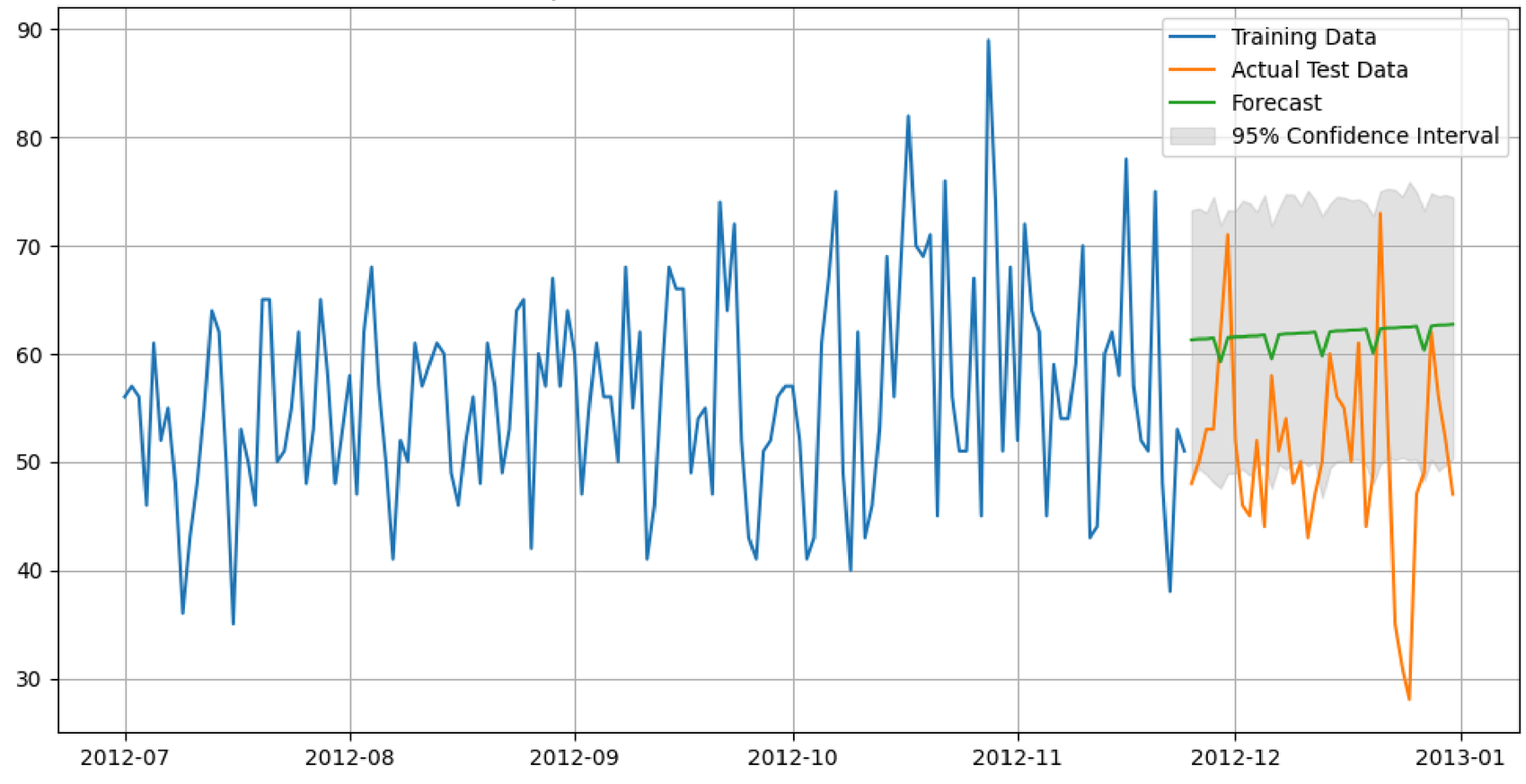
To assess forecast accuracy, predicted crime incidents were compared to actual test data, with confidence intervals included to account for uncertainty. For ARIMA, forecasts were generated with 95% confidence intervals, as depicted in the ARIMA future forecast chart. These intervals remain relatively narrow during the test period (late December 2012), reflecting moderate uncertainty and aligning closely with short-term observed trends, though some divergence occurs due to the dataset's brevity. This supports ARIMA's capability for short-term predictions despite its limitations.

AIRMA Forecast with Confidence Intervals



Facebook Prophet Forecast with Confidence Intervals

Residual Conversely, Facebook Prophet struggled significantly in forecasting future incidents. While Prophet captures the general trend of the training data (July to November 2012), its forecasts diverge markedly from the actual test data in December 2012 and into January 2013. This poor performance, illustrated in the Prophet forecast chart, suggests that Prophet is less effective for this dataset, likely due to its reliance on longer-term seasonality that the five-month span cannot adequately support.

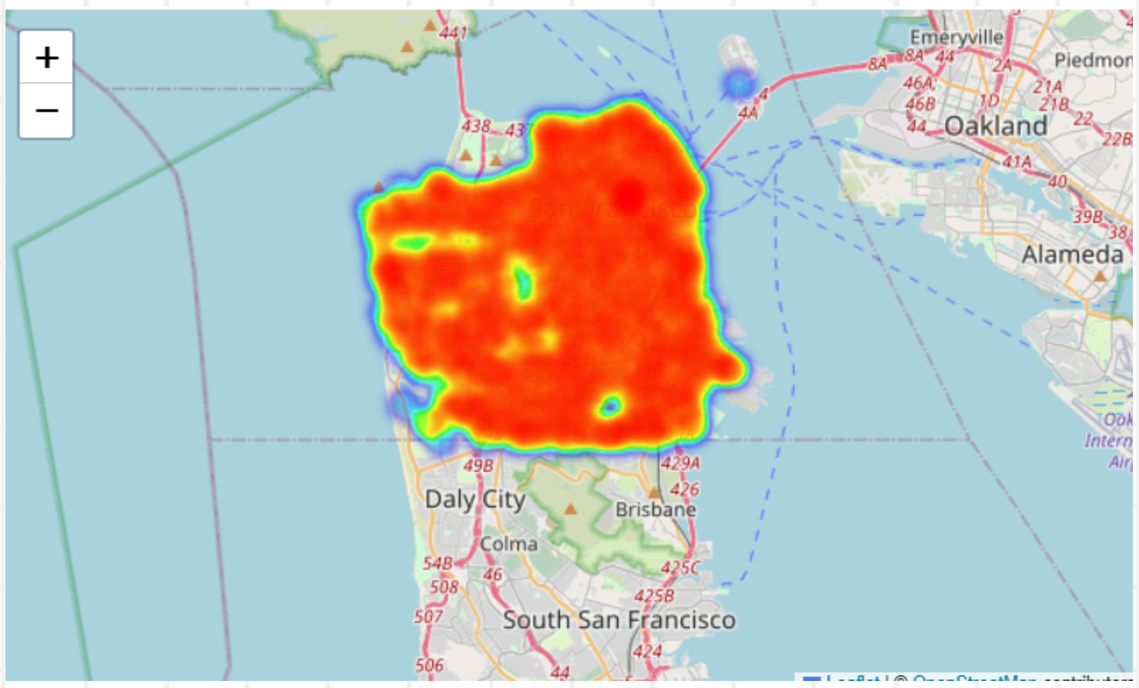


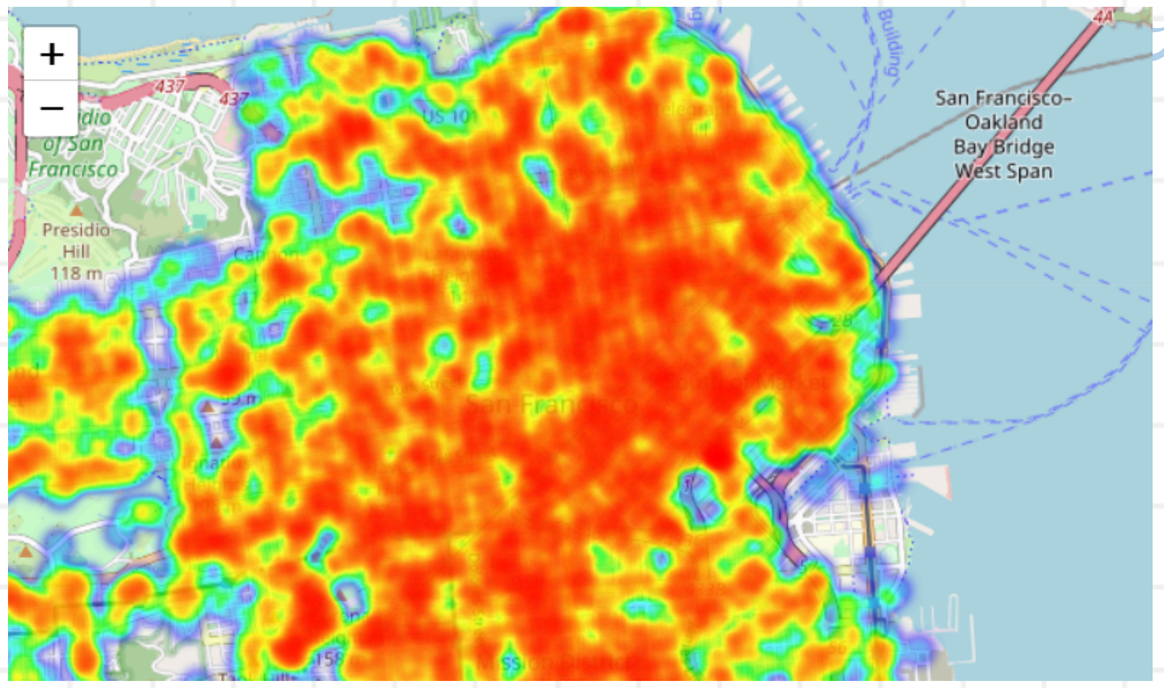
Spatial Analysis and Clustering

Spatial analysis was conducted to examine the geographic distribution of crime incidents in San Francisco from July to December 2012, using tools like GeoPandas, Folium, and LibPySal. This section explores crime density, district-level patterns, spatial relationships, and block-level insights to inform targeted interventions.

Heatmap of Crime Locations

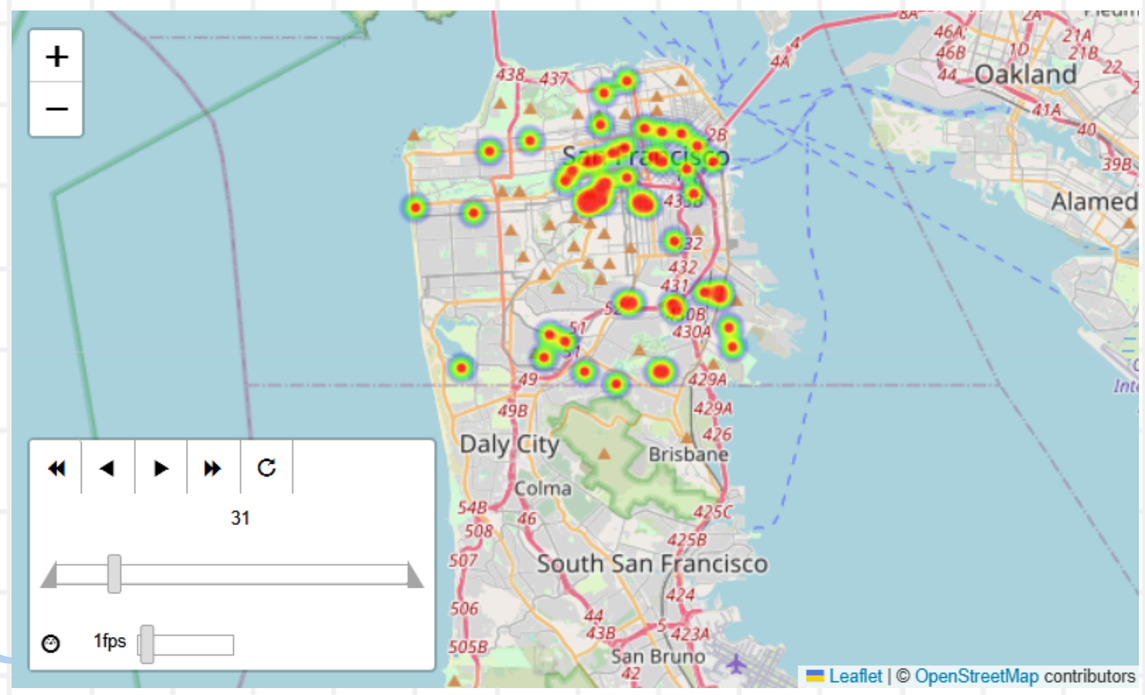
An interactive Folium heatmap visualizes crime density across the city, highlighting the Southern district as a major hotspot. A concentrated peak around coordinates (-122.42, 37.78) indicates significantly higher incidence density compared to other areas, as shown in the heatmap of crime locations.

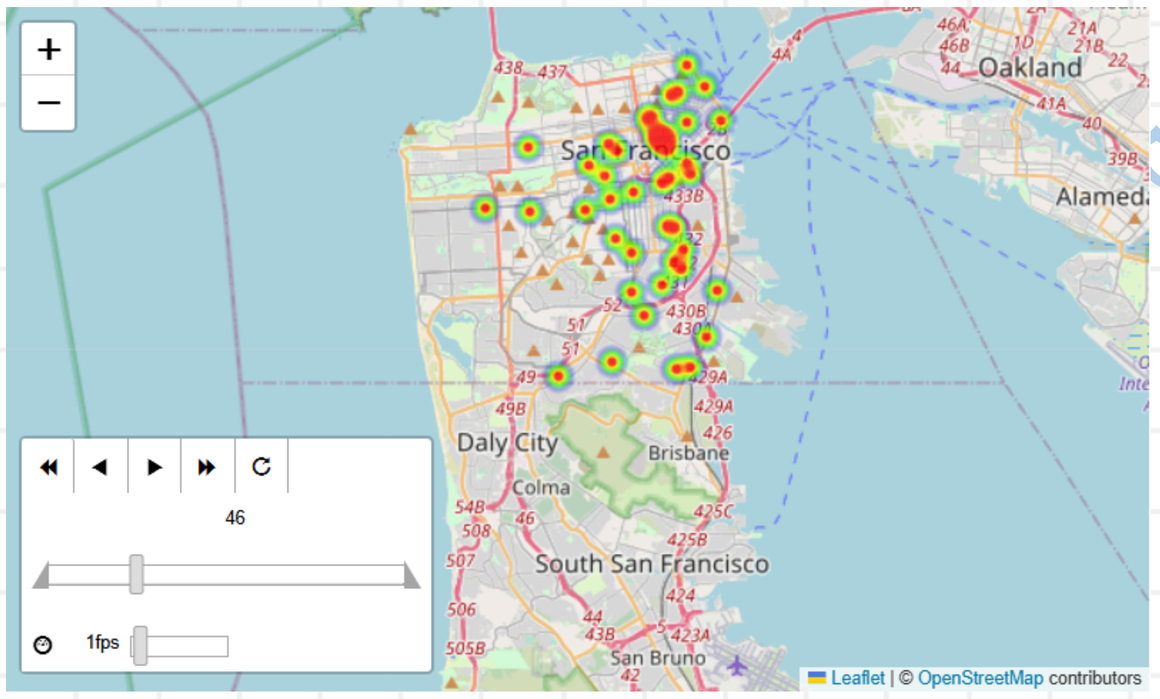




Animated Crime Map (Folium + Timestamp)

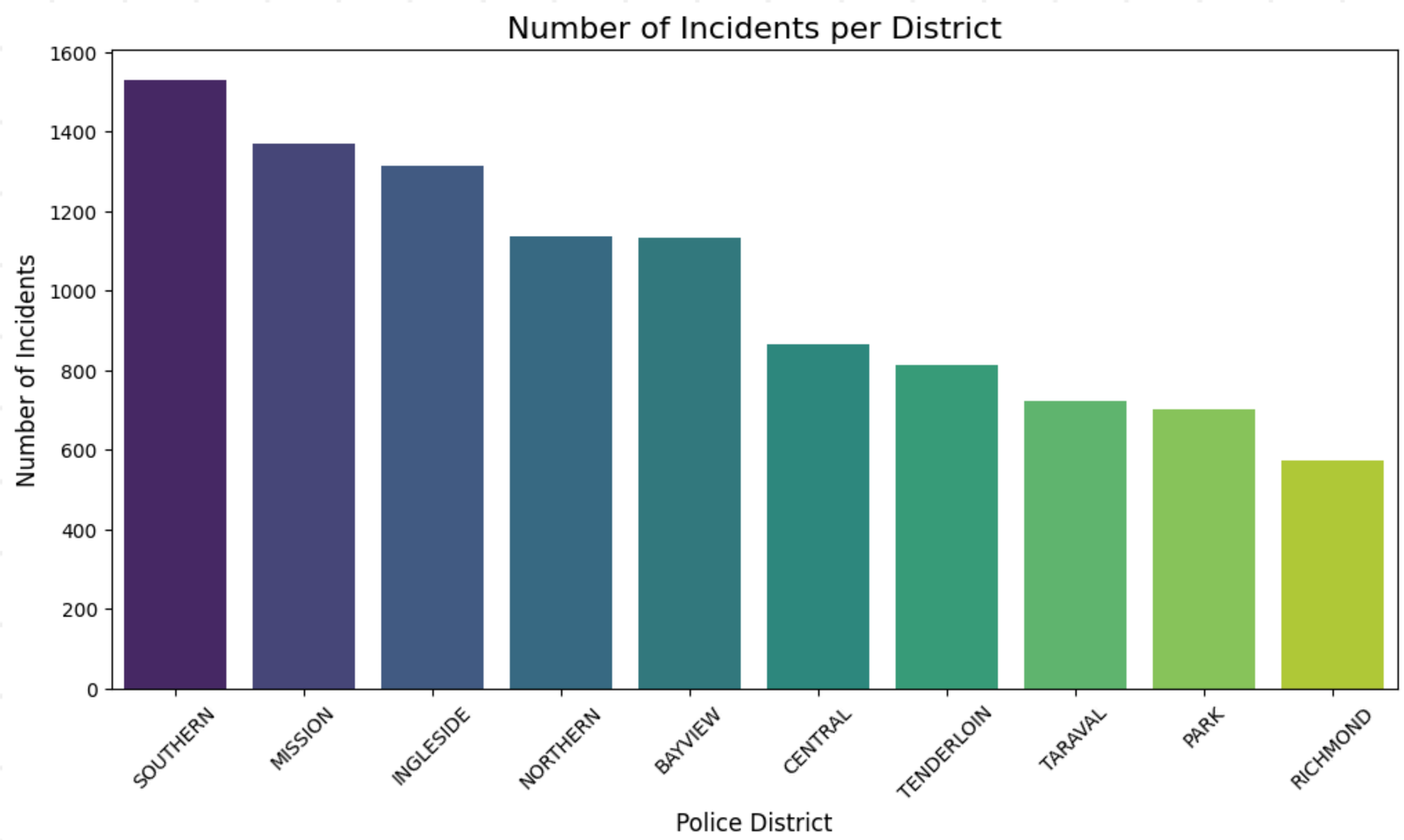
An animated Folium map with timestamps illustrates the evolution of crime patterns over time, offering dynamic insights into spatial-temporal shifts, as depicted in the animated crime map.





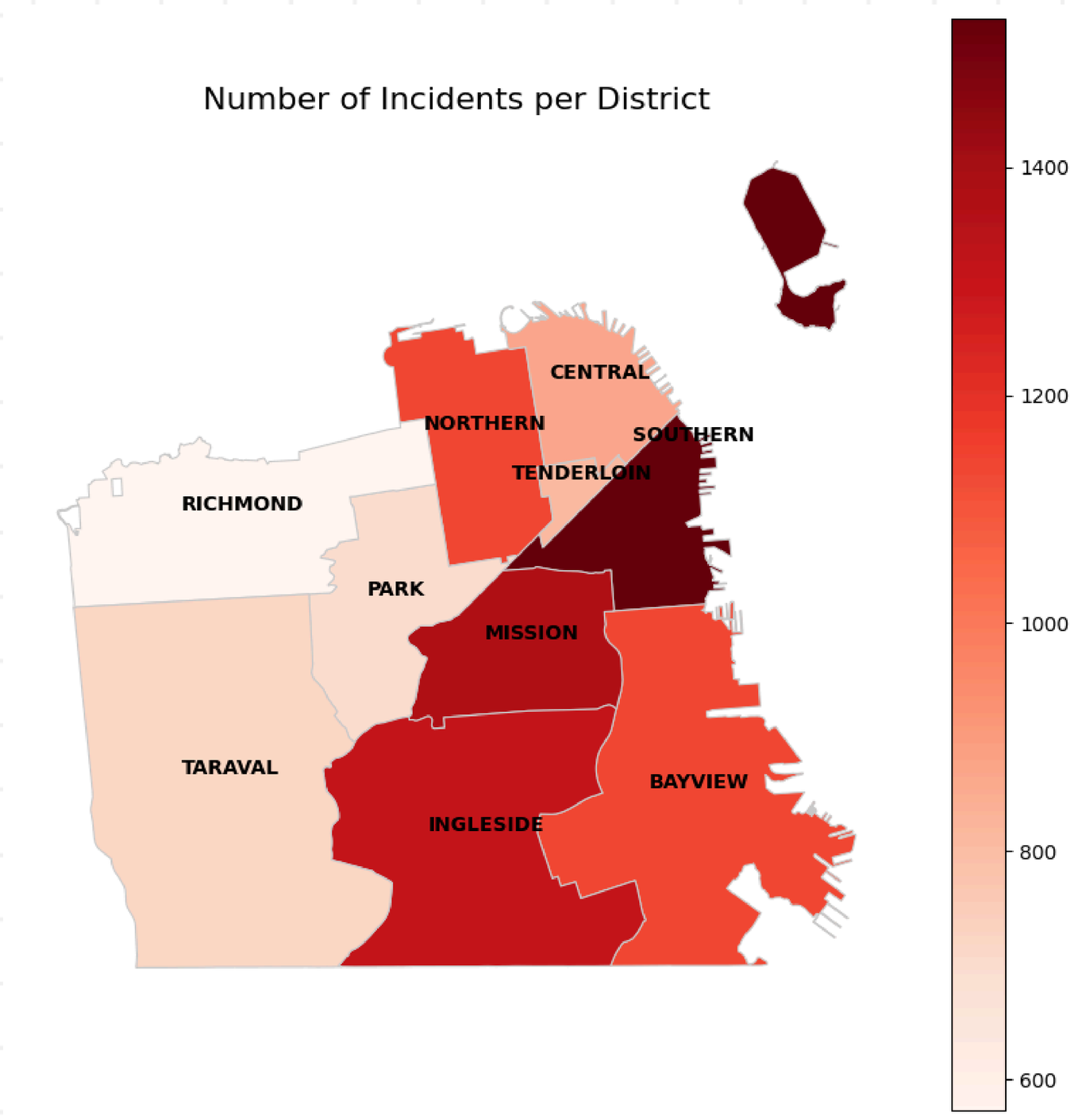
Bar Chart of Crime Counts per Police District

Crime counts vary widely across police districts, with Southern reporting the highest (2,006 cases), followed by Mission and Ingleside, while Richmond has the lowest, as presented in the bar chart of crime counts per police district.



Choropleth Map of Crime Incidents by Police District

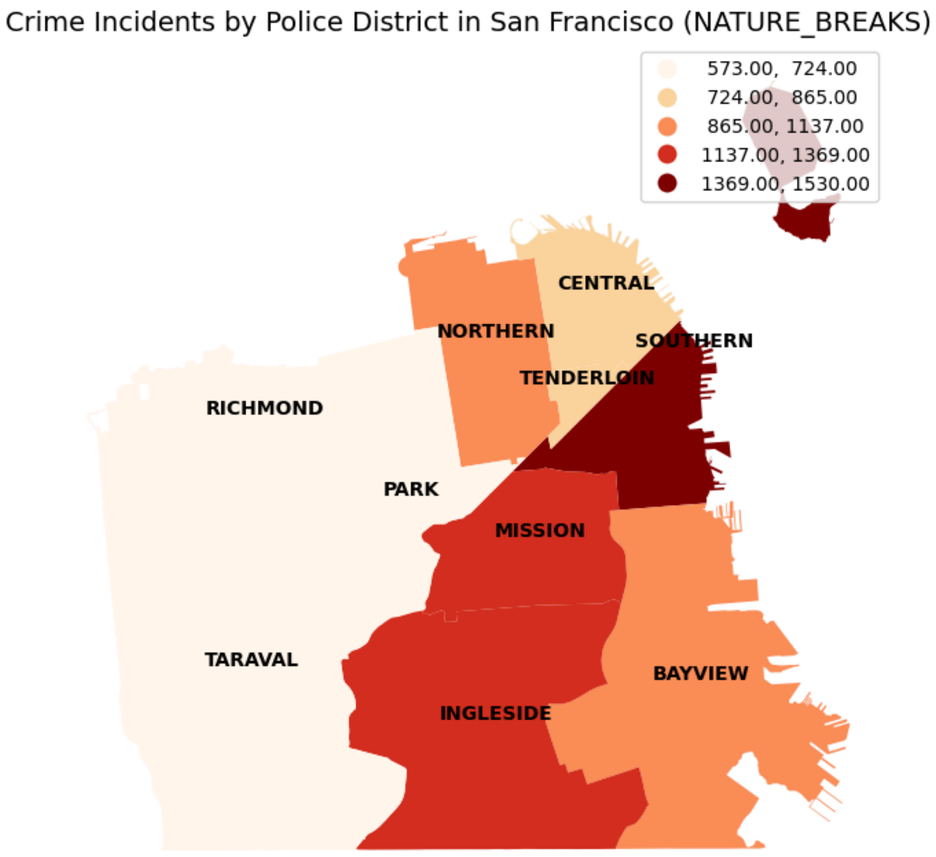
Merging district GeoDataFrame with combined_gdf via the ‘PdDistrict’ column enabled a choropleth map displaying incidents per district. This visualization, shown in the choropleth map of crime incidents by police district, enhances spatial representation.

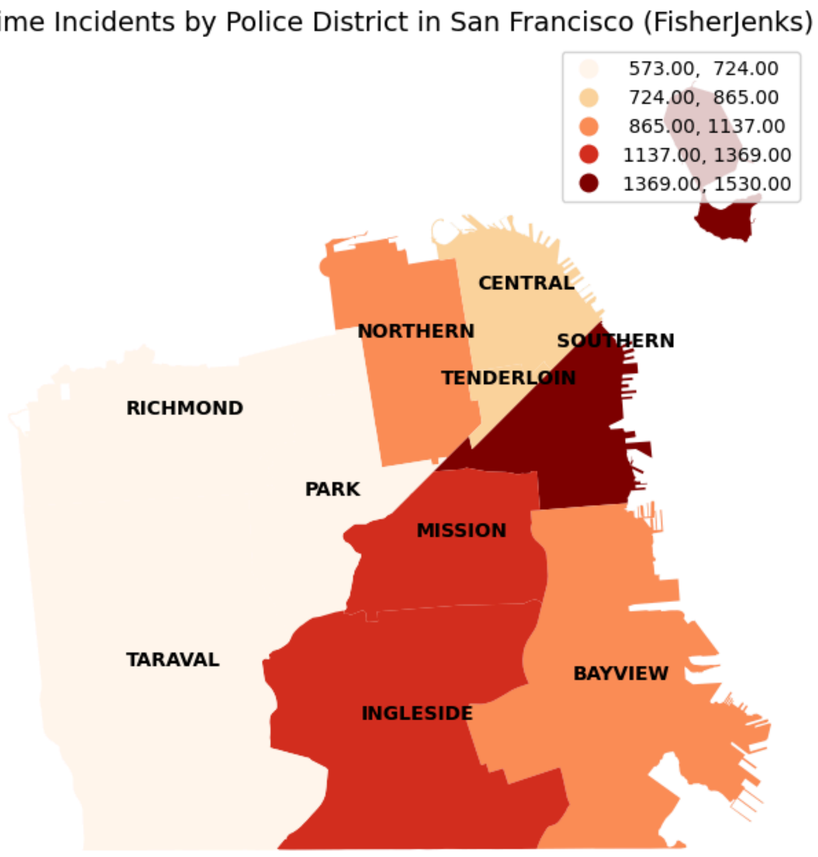


Crime Incidents by Police District in San Francisco (Natural Breaks and Fisher-Jenks Classification Methods)

Building upon the previous choropleth visualization, we further refined our analysis by applying specific classification techniques to enhance the representation of crime distribution. To ensure the robustness of our findings, both Natural Breaks and Fisher-Jenks methods were employed. To maintain visual consistency with the previous choropleth map, the same 'Reds' color scheme was used for the classification plots.

Notably, both Natural Breaks and Fisher-Jenks classification methods yielded identical results, reinforcing the spatial patterns observed in the initial choropleth map. The map clearly highlights the Southern district as having the highest concentration of crime incidents, as indicated by the darkest shade. Other districts, such as Mission and Ingleside, also exhibit relatively high crime rates compared to areas like Richmond and Taraval, which show significantly lower incident counts. The consistency in results between the classification methods underscores the robustness of the observed spatial crime patterns, confirming the findings from the previous choropleth representation.

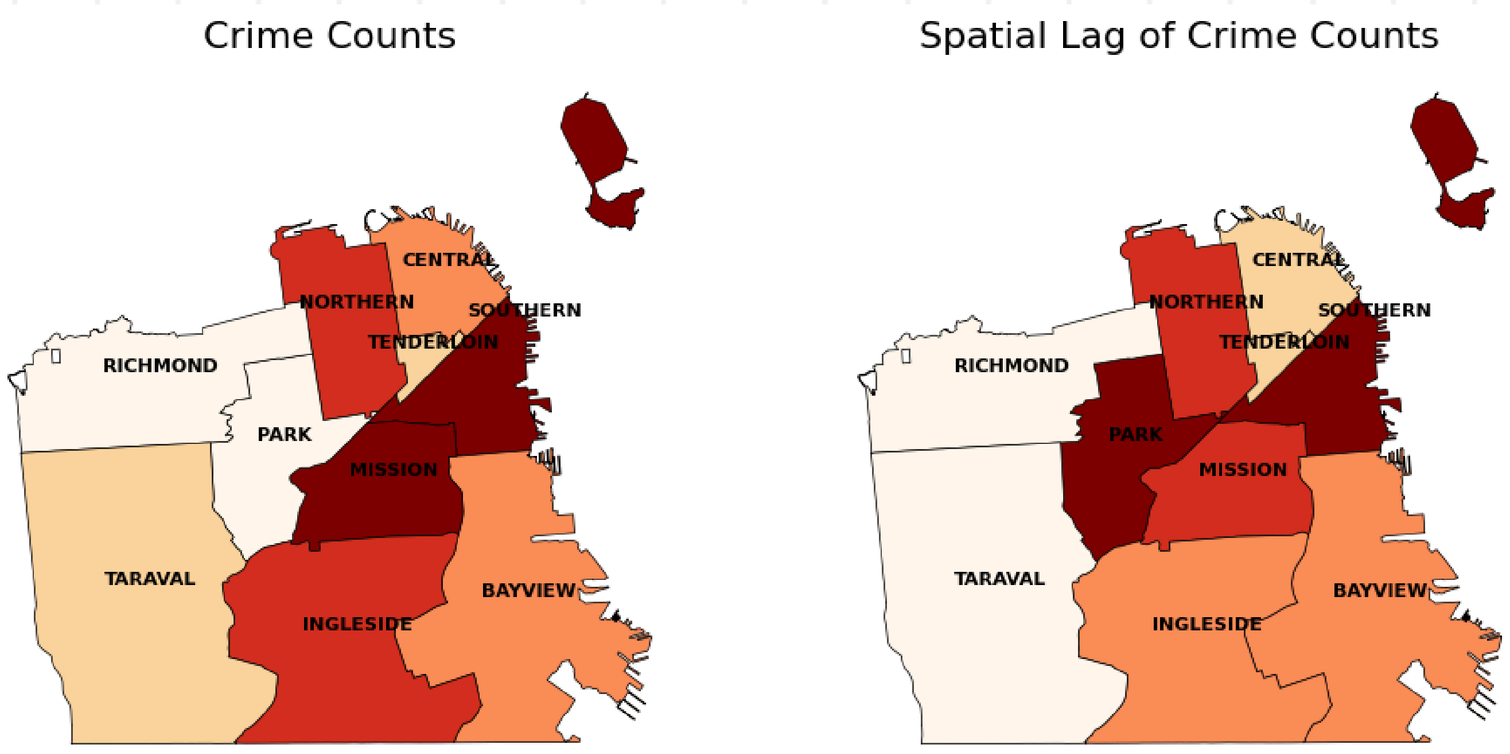




Spatial Lag Analysis of Crime Incidents in San Francisco

This analysis examines how crime incidents in San Francisco are influenced by neighboring districts. **A Queen contiguity approach** was used to calculate the spatial lag of crime counts for districts like Southern, exploring the relationship between crime levels in adjacent areas.

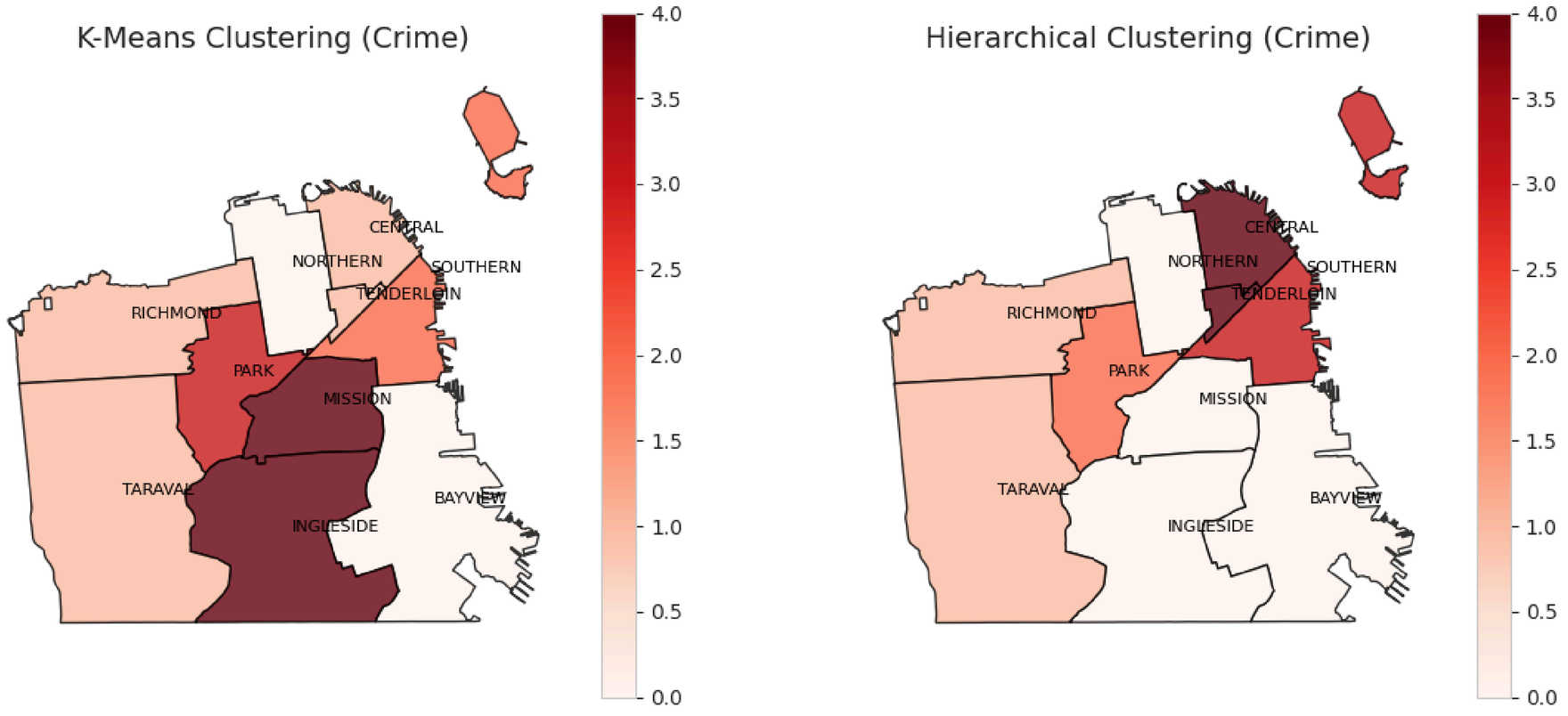
The spatial lag of crime counts reveals clear spatial dependence. Districts with higher crime counts are typically surrounded by areas with similarly elevated crime rates, indicating a clustering effect, particularly pronounced in the Southern district, as evidenced in the spatial lag analysis results.



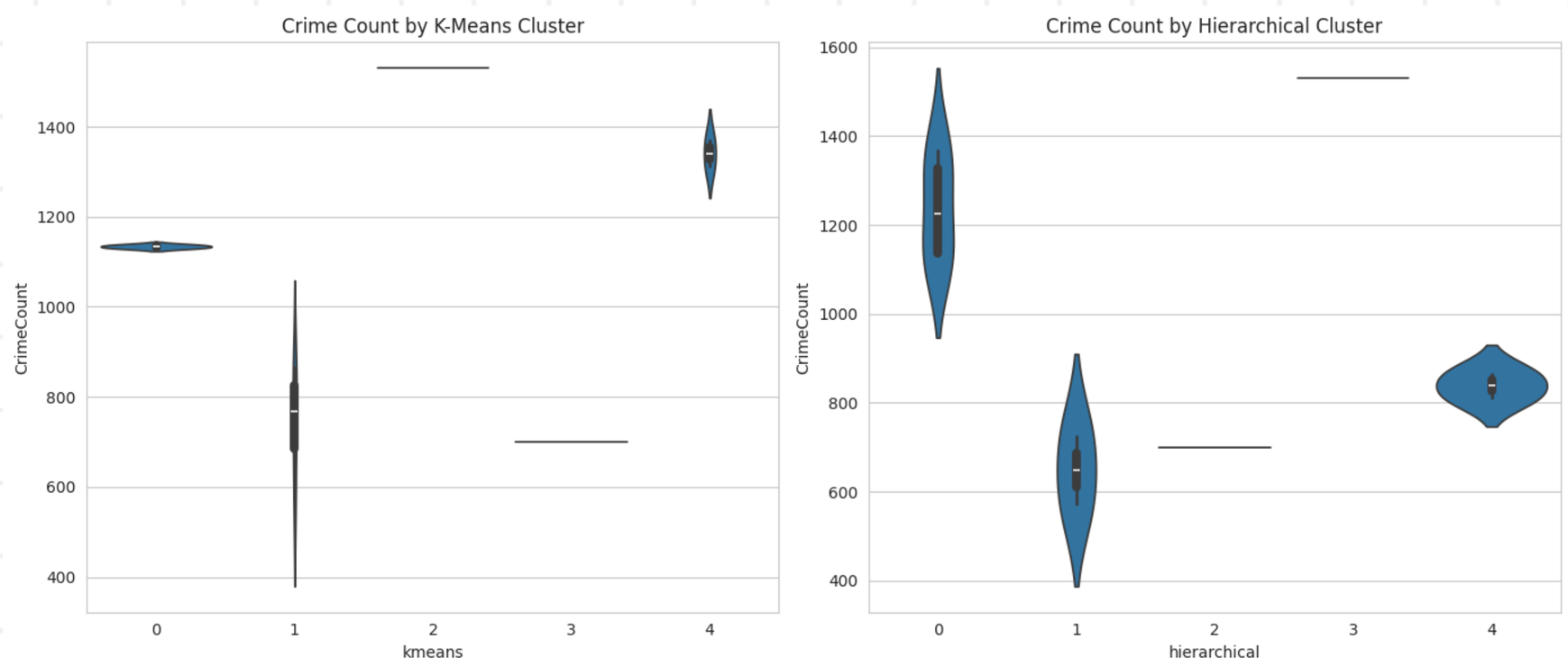
Clustering Analysis on Crime Data

Unsupervised clustering with K-Means (k=5) and Hierarchical (Agglomerative) methods analyzed crime patterns using normalized ‘CrimeCount’ and ‘CrimeLag’ features. K-Means (silhouette score 0.37) showed distinct central clusters, while Hierarchical (0.46) revealed gradual transitions, particularly around Southern. Geospatial maps of K-Means and Hierarchical clustering and violin plots of cluster distributions highlight these differences.

While both K-Means and Hierarchical clustering reveal areas of high crime concentration, particularly in the Southern and central districts, the spatial patterns differ. K-Means clustering shows a more distinct cluster in the central area, whereas Hierarchical clustering results in a more gradual transition of crime rates across districts.



The K-Means clustering shows relatively compact and distinct crime count distributions across clusters, while the Hierarchical clustering exhibits wider ranges and more pronounced outliers, particularly in cluster 0.



Block-Level Analysis

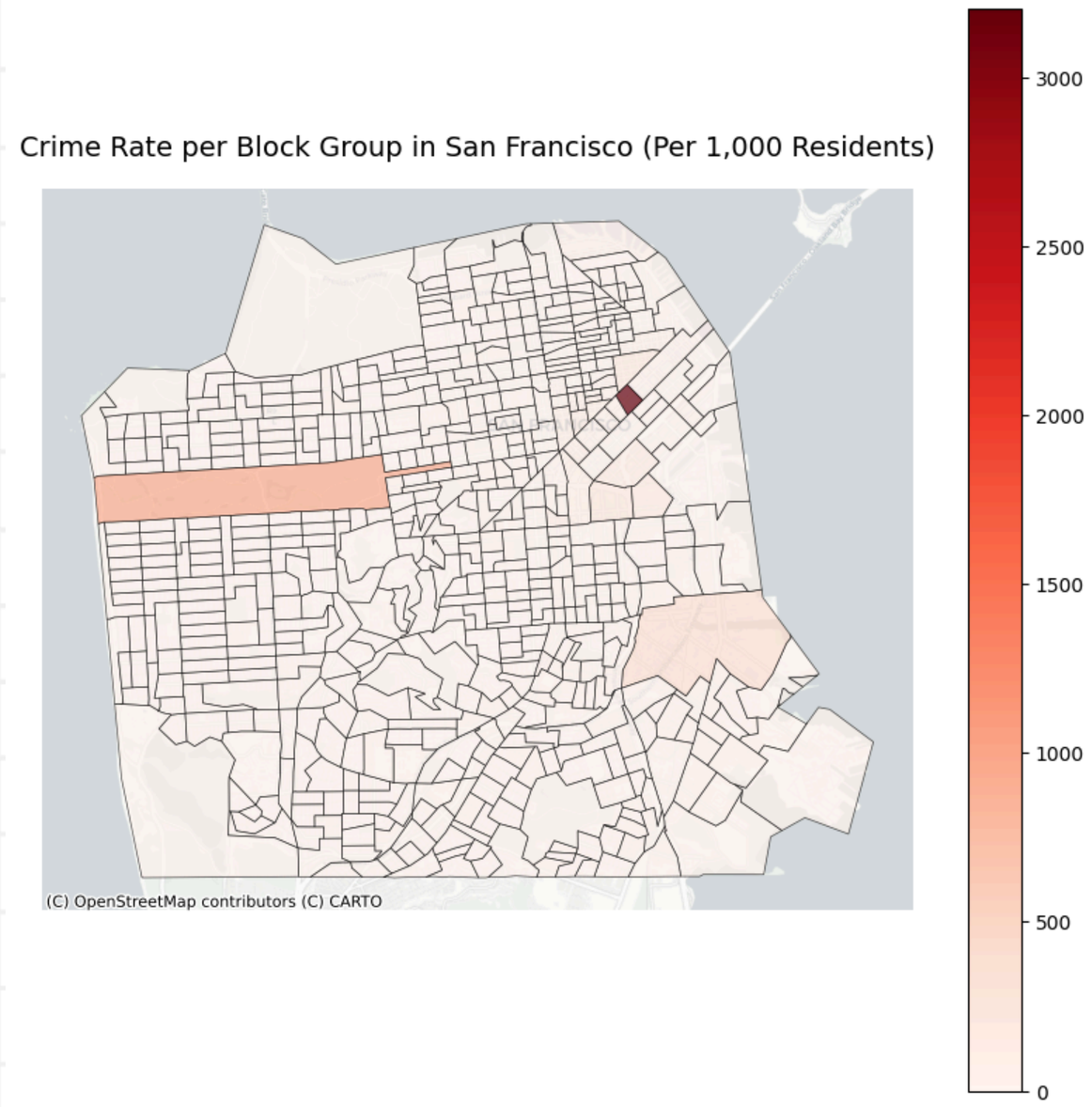
Following the district-level analysis, which highlighted the Southern district as having the highest crime incidents, a more granular examination at the block level was conducted to identify specific high-crime areas within these districts, enabling targeted interventions and resource allocation **using "SFCrime_blocks.shp".**

Crime Incidents by Block

While the block labels do not provide readily interpretable geographic information, the color gradient clearly indicates areas with higher crime incidents. Vandalism and Robbery show similar patterns with high concentrations in the central-eastern area and a long block on the west for drugs, while Car Theft have more scattered distributions. Though Car Theft maintains a significant concentration in the central-eastern area, it has the highest density of crime in part in Bayview.

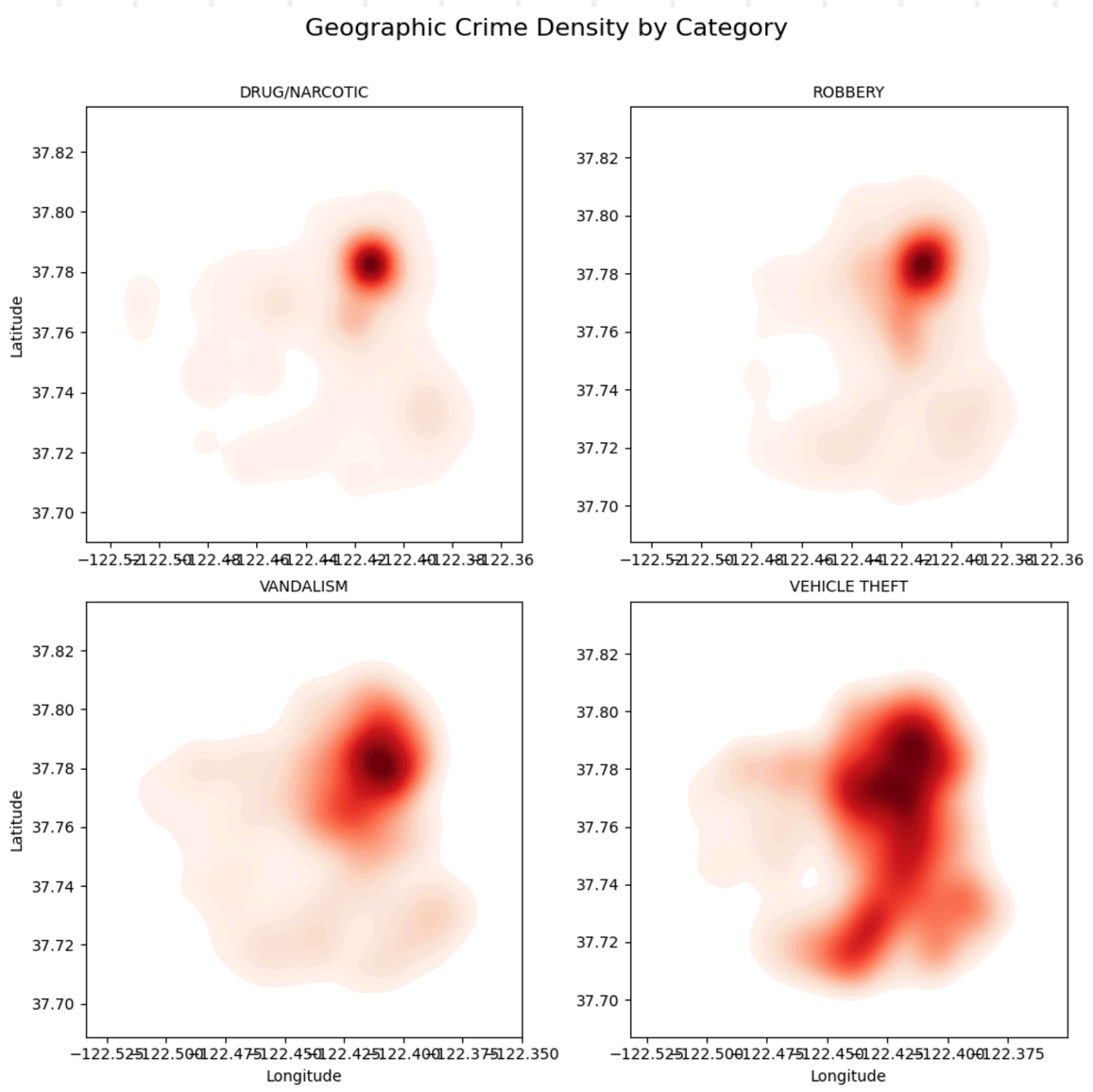
Crime Rate per Block Group in San Francisco (Per 1,000 Residents)

Adjusting for population, the crime rate per block group map reveals a central-eastern hotspot with elevated rates, contrasting with lower rates elsewhere, as depicted in the block group crime rate visualization.



Geographic Density of Different Crimes (KDE Plot)

(KDE) plots visualizes the spatial distribution and density of different crime categories across the city, revealing potential hotspots and variations in geographic patterns. While a central hotspot area, located in the northeast region of the city, exhibits high crime density across all categories, each crime type displays a distinct density pattern throughout the city.



This variation underscores the significant influence of location (coordinates/Police Districts) on crime distribution, which will be crucial for further analysis and forecasting. Drug/Narcotic and Robbery show a more localized high-density pattern compared to Vandalism and Vehicle Theft, which have slightly broader areas of elevated density. Vehicle Theft exhibits the most dispersed density among the four categories.

Predictive Modelling

This section focuses on predicting the most likely crime type in San Francisco given location and time, using data from July to December 2012. A classification approach was developed to leverage temporal and spatial features for actionable crime type predictions.

Crime Type Prediction: Given a Location & Time

The objective was to predict crime categories (Drug/Narcotic, Vehicle Theft, Robbery, Vandalism) based on location and time inputs, enhancing preventive strategies through data-driven insights.

Data Preparation

Feature Engineering

To prepare the dataset, new features were engineered to capture temporal and spatial patterns. The day of the month and days since the dataset's start were extracted from the 'Date' column, adding granularity beyond 'DayOfWeek.' The 'Location' column was processed to create a 'Block' feature, categorizing incidents as 'Block' (True) or 'Other' (False) based on the presence of "block of" in the description. These, alongside 'Month' and 'DayOfWeek,' formed eight input features for modeling.

Feature Scaling

A tree-based algorithm (LightGBM) was chosen, eliminating the need for feature scaling due to its insensitivity to unscaled data.

Feature Selection

Permutation Importance assessed the eight features' contributions by shuffling each feature's values and measuring the impact on model performance. All features showed positive importance, indicating no need for removal, as detailed in the feature importance table.

Model Training and Evaluation

The dataset was split into 80% training and 20% testing sets (random_state=42) to ensure robust evaluation. Seven models were tested: SGD Classifier, KNN, Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Each model was trained on the training set (8,122 samples, 8 features) and evaluated on the test set (2,031 samples) using accuracy and classification reports (precision, recall, F1-score).

LightGBM, a gradient-boosting model, was initially trained with a multiclass objective, achieving optimal performance. Permutation Importance confirmed feature relevance, with results displayed in the feature importance table. Comparative evaluation across all models used accuracy as the primary metric, supplemented by detailed classification metrics.

Results

LightGBM outperformed others with an accuracy of 0.4653, followed by XGBoost (0.4609) and Random Forest (0.4495). SGD Classifier scored lowest at 0.2757, reflecting poor fit. LightGBM's classification report showed strong precision for Drug/Narcotic (0.62) and recall for Vandalism (0.56), though Vehicle Theft had lower recall (0.26), as presented in the model performance summary. Feature importance highlighted 'X' (longitude), 'Y' (latitude), and 'DayOfWeek' as key predictors, aligning with spatial and temporal influences.

Model: SGD
Accuracy: 0.2757
Classification Report:

	precision	recall	f1-score	support
0	0.23	0.65	0.34	369
1	0.23	0.14	0.18	388
2	0.00	0.00	0.00	648
3	0.36	0.42	0.39	626
accuracy			0.28	2031
macro avg	0.20	0.30	0.23	2031
weighted avg	0.20	0.28	0.21	2031

Model: RandomForest
Accuracy: 0.4495
Classification Report:

	precision	recall	f1-score	support
0	0.56	0.45	0.50	369
1	0.33	0.28	0.30	388
2	0.45	0.50	0.47	648
3	0.46	0.50	0.48	626
accuracy			0.45	2031
macro avg	0.45	0.43	0.44	2031
weighted avg	0.45	0.45	0.45	2031

Model: XGBoost
Accuracy: 0.4609
Classification Report:

	precision	recall	f1-score	support
0	0.60	0.46	0.52	369
1	0.40	0.31	0.35	388
2	0.43	0.47	0.45	648
3	0.46	0.55	0.50	626
accuracy			0.46	2031
macro avg	0.47	0.45	0.45	2031
weighted avg	0.47	0.46	0.46	2031

Model: LightGBM
Accuracy: 0.4653
Classification Report:

	precision	recall	f1-score	support
0	0.62	0.44	0.51	369
1	0.38	0.26	0.31	388
2	0.45	0.51	0.48	648
3	0.46	0.56	0.50	626
accuracy			0.47	2031
macro avg	0.48	0.44	0.45	2031
weighted avg	0.47	0.47	0.46	2031

Discussion

LightGBM’s superior accuracy and balanced metrics make it the recommended model for crime type prediction. Its ability to handle categorical data and spatial coordinates enhances its practical utility. However, the moderate accuracy (0.47) suggests limitations from the dataset’s size and feature set. Incorporating additional variables (e.g., weather, socioeconomic data) could improve performance. These predictions enable targeted policing—e.g., focusing on Drug/Narcotic hotspots—demonstrating the value of predictive analytics in crime prevention.

Conclusion

This comprehensive analysis of San Francisco crime data from July to December 2012 has uncovered significant spatial and temporal patterns that offer valuable insights for enhancing law enforcement strategies. By integrating exploratory data analysis, time series forecasting, spatial mapping, spatial clustering, spatial lag analysis and predictive modeling, this study provides a multi-dimensional perspective on crime dynamics across the city.

Key findings:

- **High Prevalence of Vandalism incidents**: Vadalism incidents emerged as the most frequent crime category (3,897 cases), underscoring their dominance in the dataset, followed by Vehicle Theft and Robbery.Most cases (8,018) had no resolution ("NONE"), while "Arrest, Booked" (1,682) was the primary actionable outcome
- **Temporal Patterns with Distinct Peaks**: Daily crime trends showed significant variability, with a notable spike in mid-October and a dip on December 25, 2012. Weekly analysis identified Friday as the peak day, while monthly trends highlighted October as the highest-crime month, suggesting potential seasonal or event-driven influences.
- **ARIMA as the Best Forecasting Model**: Among forecasting models (ARIMA, Holt-Winters, Prophet), ARIMA outperformed with an RMSE of 11.05 and MAPE of 20.95%, effectively capturing short-term fluctuations and weekly seasonality despite the dataset’s limited five-month span.

Key findings:

- **Spatial Clustering in the Southern District**: Spatial analysis revealed a pronounced concentration of crime in the Southern police district (2,006 cases), with heatmaps and block-level KDE plots pinpointing a central-eastern hotspot, particularly for Drug/Narcotic and Robbery incidents, while Vehicle Theft showed broader dispersion.
- **Predictive Power of Spatial and Temporal Features**: The LightGBM model achieved the highest accuracy (0.4653) in predicting crime categories, with longitude ('X'), latitude ('Y'), and 'DayOfWeek' as key predictors, highlighting the critical role of location and timing in crime type classification.

These findings provide a robust foundation for evidence-based decision-making, enabling law enforcement to optimize resource allocation, refine patrol scheduling, and develop targeted intervention strategies. By leveraging insights into when and where specific crimes are most likely to occur, agencies can enhance operational efficiency and strengthen crime prevention efforts in San Francisco.

Limitations and Future Work

This study provides valuable insights into San Francisco's crime patterns from July to December 2012, but several limitations must be acknowledged. The dataset's six-month duration restricts the ability to detect long-term trends or complete seasonal cycles, potentially missing broader patterns. Socioeconomic and demographic variables, such as income or population density, were excluded, limiting contextual understanding of crime drivers. Additionally, the models do not account for policy shifts or law enforcement strategies that could influence crime rates, reducing their adaptability to external changes.

Future research could enhance this study by incorporating longer time series data, integrating socioeconomic, demographic, and environmental variables, developing advanced models with spatial autocorrelation, assessing policing strategy impacts, and exploring real-time prediction for dynamic resource allocation. By addressing these gaps, future efforts could deepen the understanding of urban crime dynamics and support safer communities through more robust, data-driven crime prevention and law enforcement strategies.

References

1. Blumstein, A., & Wallman, J. (2006). The crime drop in America. Cambridge University Press.
2. Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal, 21(1-2), 4-28.
3. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), 588-608.
4. Cohn, E. G., & Rotton, J. (2003). Even criminals take a holiday: Instrumental and expressive crimes on major and minor holidays. Journal of Criminal Justice, 31(4), 351-360.
5. Felson, M., & Poulsen, E. (2003). Simple indicators of crime by time of day. International Journal of Forecasting, 19(4), 595-601.
6. Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115-125.
7. Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399-1411.
8. Ratcliffe, J. H. (2010). Aoristic analysis: The spatial interpretation of unspecific temporal events. International Journal of Geographical Information Science, 14(7), 669-679.
9. Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. American Journal of Sociology, 94(4), 774-802.
10. Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. Criminology, 27(1), 27-56.
11. Weisburd, D. (2015). The law of crime concentration and the criminology of place. Criminology, 53(2), 133-157.
12. Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. Journal of Quantitative Criminology, 37(2), 445-480.

Thank you