



SAN FRANCISCO CRIME PREDICTION SYSTEM



Problem Statement

- High crime rates globally
- Need for predictive systems to support law enforcement
- Goal: Build a system to analyze and predict future crime patterns using historical data

Objectives

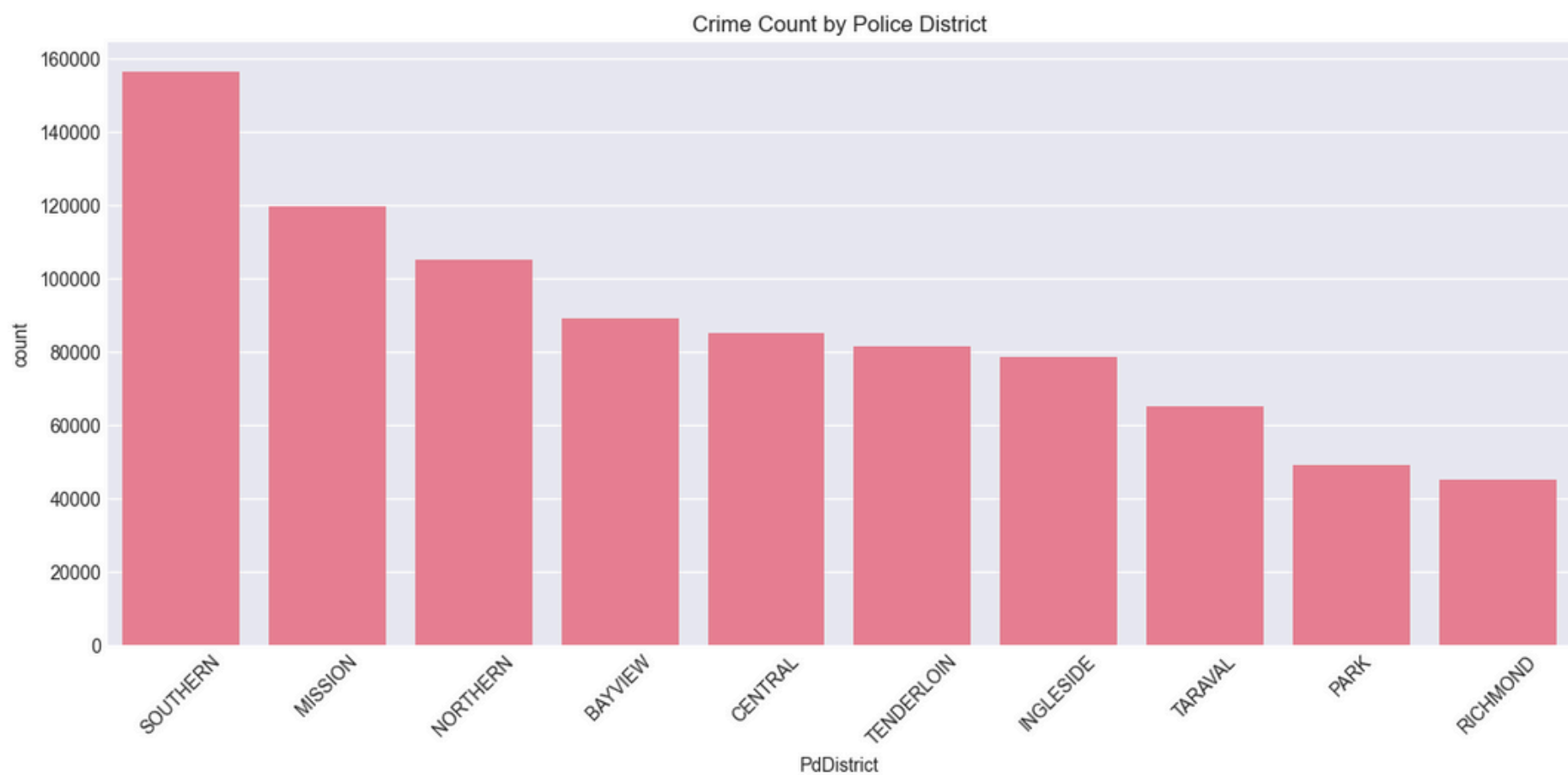
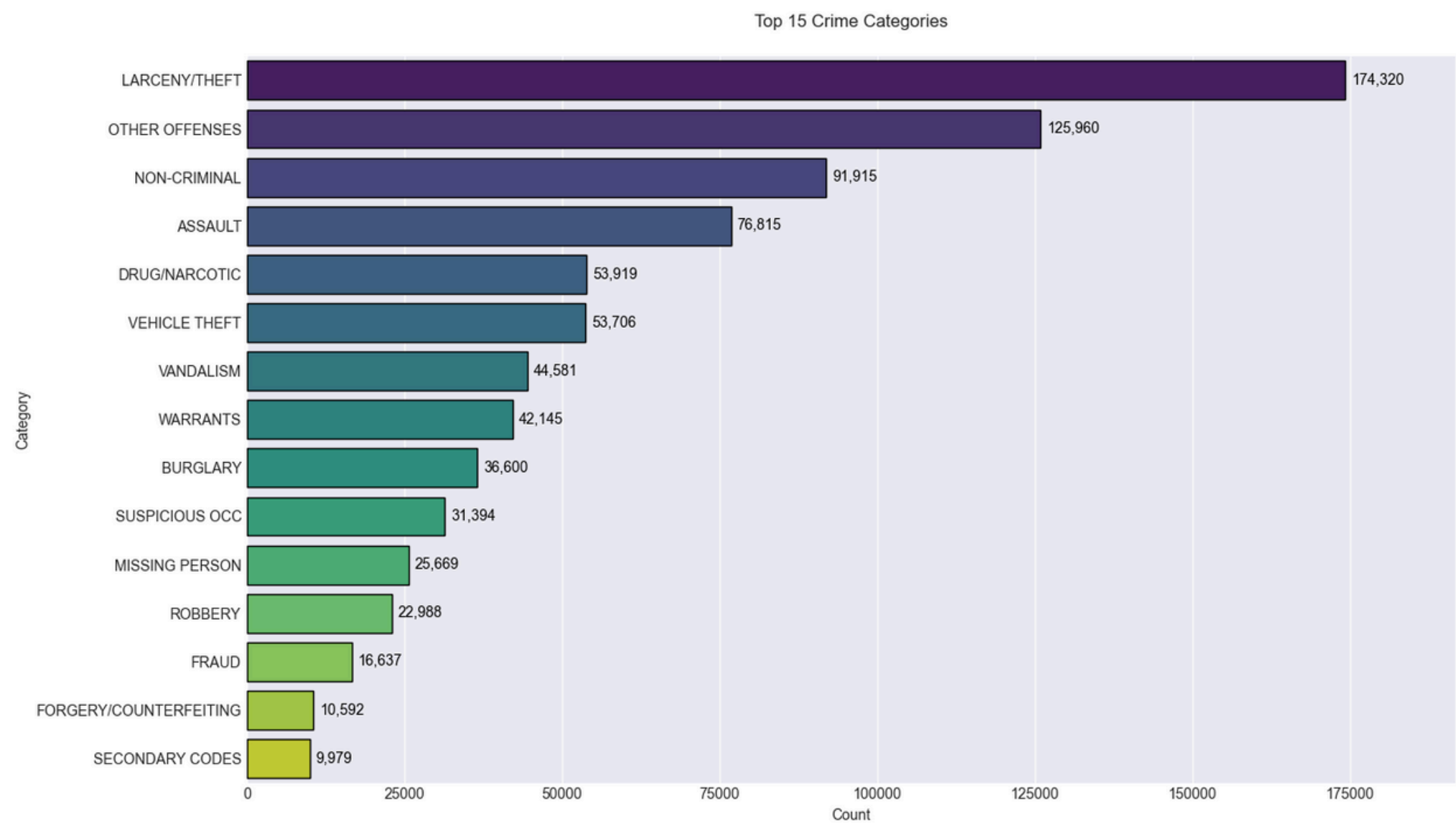
- Analyze historical crime data
- Engineer features (time, location, patterns)
- Predict crime categories using ML
- Visualize crime trends via heatmaps and dashboards



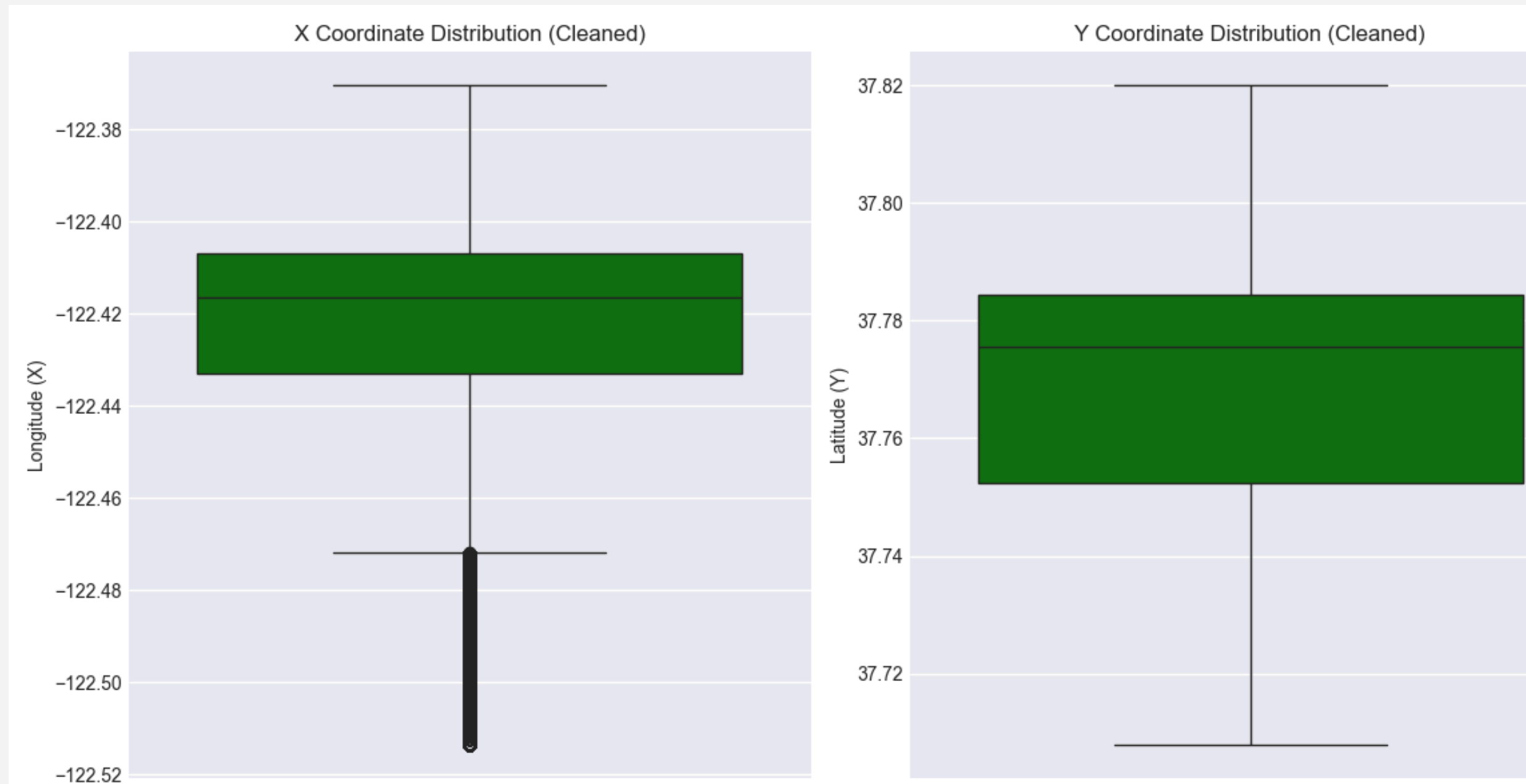
-

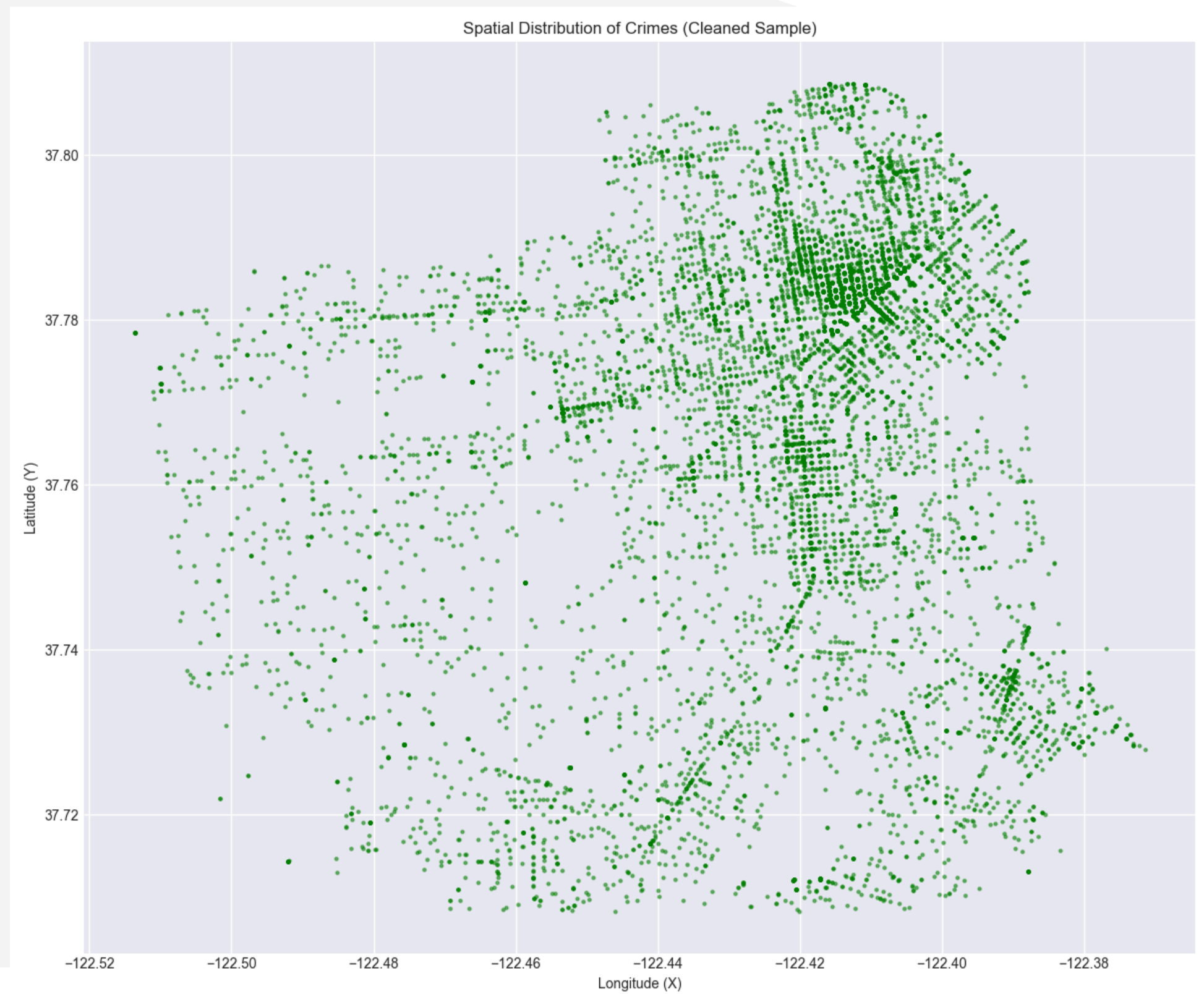
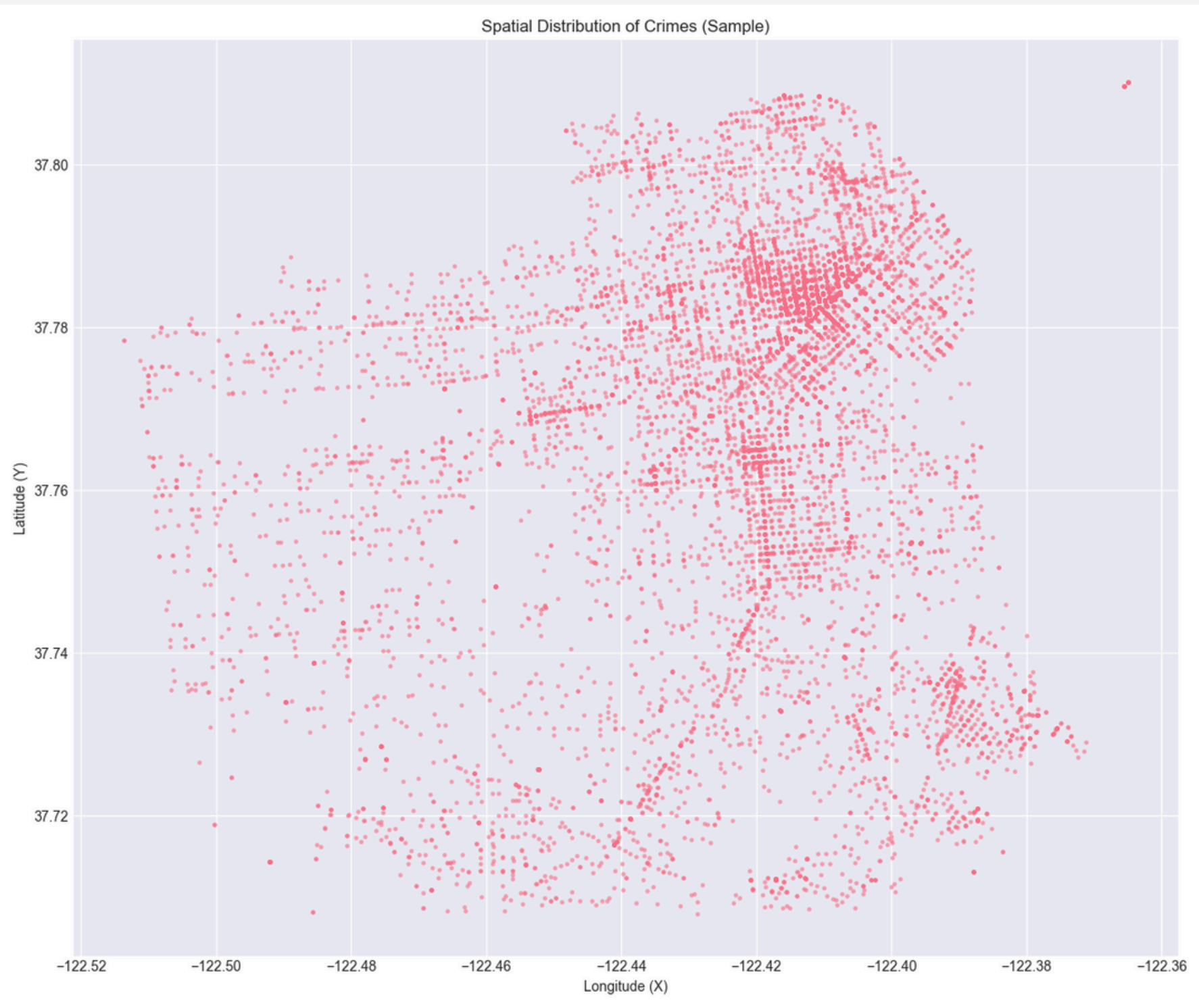


This analysis explores crime patterns in categorical features, focusing on the most frequent crime types and their distribution across police districts. It helps identify which crimes are most common and where they occur most frequently, guiding targeted interventions.



This spatial analysis identifies and removes outlier coordinates to ensure crimes are within realistic San Francisco boundaries. It uses IQR to detect abnormal latitude and longitude values, then visualizes both raw and cleaned data through scatter plots and box plots. This helps eliminate noise from incorrectly recorded locations and highlights the geographic distribution of crimes more accurately.







✓ 1. VERIFIED COORDINATE INTEGRITY

- WE DOUBLE-CHECKED THAT ALL CRIME RECORDS FALL WITHIN REALISTIC SAN FRANCISCO LATITUDE AND LONGITUDE BOUNDS.
- ANY REMAINING OUTLIERS WERE REMOVED TO ENSURE ONLY VALID, GEOGRAPHICALLY ACCURATE ENTRIES ARE USED IN ANALYSIS.

✓ 2. GENERATED AN INTERACTIVE HEATMAP

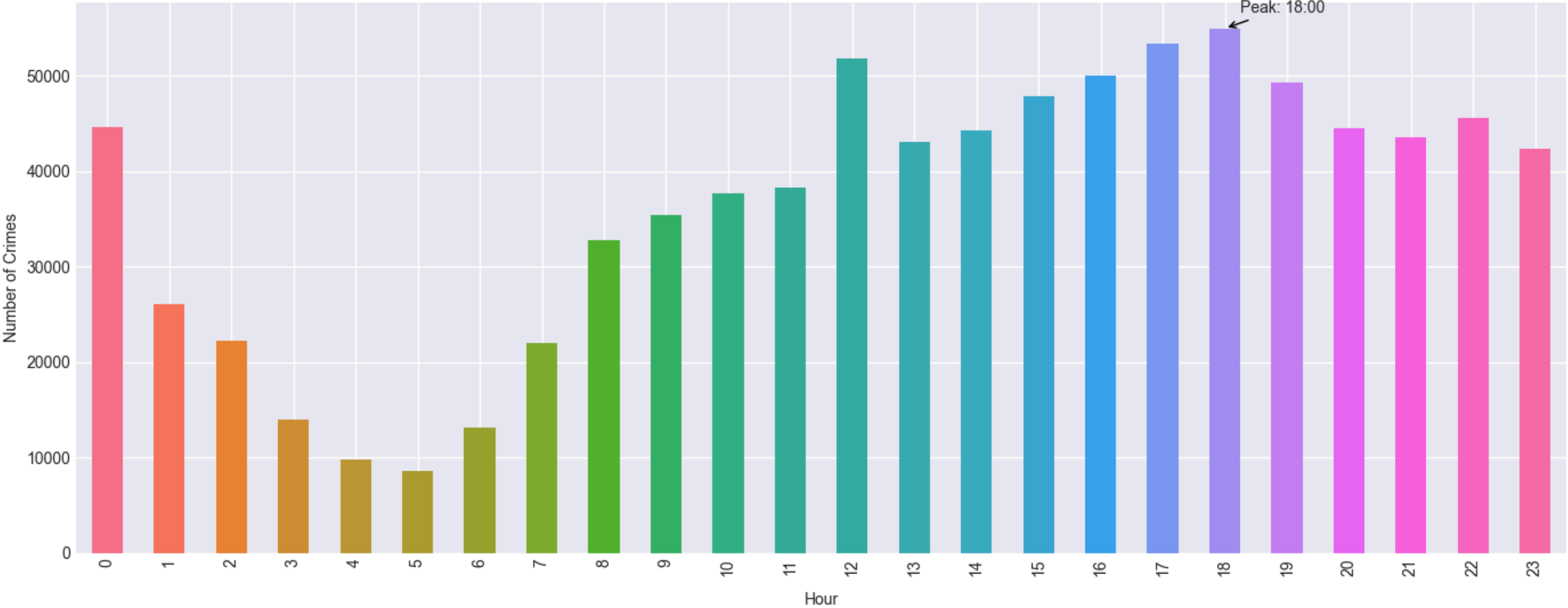
- A SAMPLE OF 50,000 CLEANED RECORDS WAS USED TO CREATE A FOLIUM-BASED HEATMAP SHOWING AREAS WITH HIGH CRIME DENSITY.
- THIS VISUAL HELPS QUICKLY IDENTIFY CRIME HOTSPOTS ACROSS THE CITY.

✓ 3. SAVED CLEANED DATASET AND VISUALIZATIONS

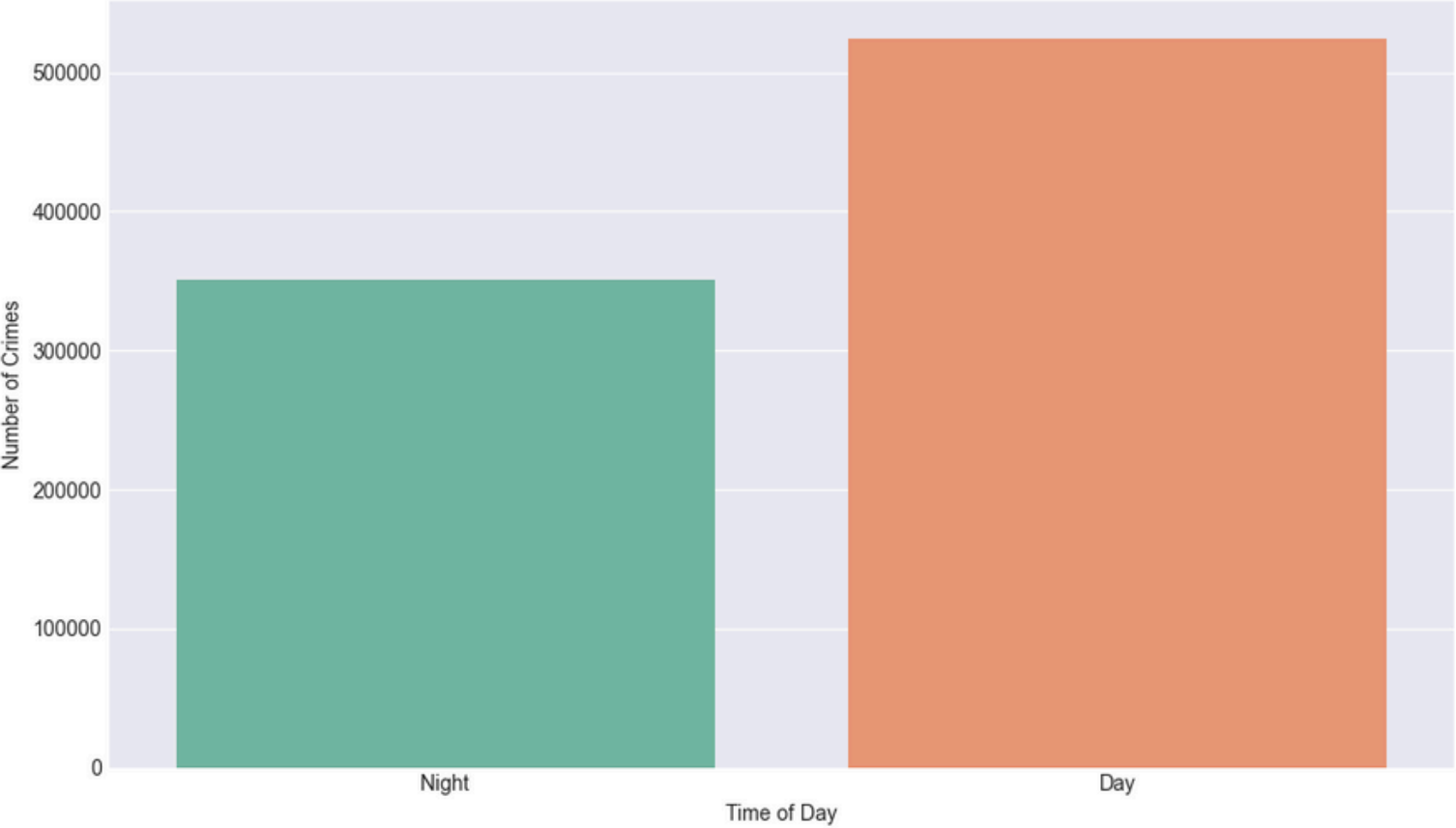
- THE CLEANED DATA WAS SAVED TO DATA/SF_CRIME_CLEANED.CSV.
- ALL SPATIAL VISUALIZATIONS (SCATTER PLOTS, BOX PLOTS, HEATMAP) WERE STORED IN THE VISUALIZATION FOLDER FOR REPORTING AND DASHBOARD USE.

02 May, 2024

Crime Frequency by Hour of Day





Crime Distribution by Time of Day





Dataset Summary:

- Original dataset size: 878,049 records
- Unique crime categories: 39
- Focused categories:
 - LARCENY/THEFT
 - OTHER OFFENSES
 - NON-CRIMINAL
 - ASSAULT
 - DRUG/NARCOTIC
- Filtered data size: 524,233
- Modeling sample size: 100,000

- 
- ## FEATURE ENGINEERING:
- ### SELECTED 12 FEATURES INCLUDING:
- TEMPORAL: HOUR, MONTH, ISWEEKEND, ISNIGHT, SEASON_ENCODED
 - SPATIAL: X, Y, X_AREA, Y_AREA, ISHIGHCRIMELOCATION
 - CATEGORICAL ENCODINGS: PDDISTRICT_ENCODED, DAYOFWEEK_ENCODED
- 



Decision Tree Classifier

- Accuracy: 99.95% (nearly perfect)
 - Strengths:
 - Very high precision, recall, and f1-score across almost all crime categories.
 - Excellent at capturing complex decision boundaries.
 - Key Insight:
 - Handles imbalanced data well in this context.
 - Near-perfect classification on the test set shows strong model fit.
-

K-Nearest Neighbors (KNN)

- Accuracy: 97.37%
- Strengths:
 - High precision and recall for most common crime categories such as ASSAULT, LARCENY/THEFT.
- Observations:
 - Struggles with minority classes
 - Performance varies with class distribution and neighborhood size.

Random Forest Classifier

- Accuracy: 91.66%
- Strengths:
- Robust ensemble method combining many decision trees.
- Good generalization ability.
- Challenges:
- Lower precision and recall on some rare categories (e.g., categories with few samples have near-zero recall).
- Weighted average metrics still high, indicating good overall performance.

COMPARISON

