SF Crime Forecasting
CS133 - Data Visualization
San Jose State University
Professor Jessica Huynh-Westfall

May 5, 2025

RuiYuanLi,Tang Ho Lam

1. Abstract

This project investigates the predictability of crime types in San Francisco using historical police incident data from 2018 to 2025. The primary goal is to build a machine learning pipeline that can classify crimes into grouped categories (`crime_type_group`) based on temporal and spatial features. We focus on identifying crime patterns over time (hour, day, season) and space (neighborhoods, police districts, DBSCAN clusters) to support public safety decision-making.

The dataset includes detailed incident records with timestamps, categories, location data, and derived frequency metrics. After preprocessing—including data cleaning, label encoding, and DBSCAN/KMeans clustering—we trained several models: RandomForest, LogisticRegression, DecisionTree, and LSTM. RandomForest delivered the most consistent accuracy and F1-score, while LSTM captured sequential time dependencies using 24-hour rolling windows.

Evaluation was done using classification metrics (accuracy, F1-score) and confusion matrices. The results confirmed that crime type prediction is feasible with proper spatio-temporal features and class balancing. This project showcases the practical impact of data-driven modeling on urban crime analysis and resource planning.

2. Background and Data Exploration

San Francisco is a densely populated urban center with diverse neighborhoods and varying crime dynamics. Historical crime data collected by the San Francisco Police Department provides detailed insight into when and where specific types of incidents occur. Our project aims to leverage this temporal and geographic information to identify consistent patterns and improve crime categorization through machine learning.

Five Key Questions

1. Can we classify crime category groups based on time and spatial patterns?
2. What are the peak hours for different types of crimes?
3. Do neighborhoods show consistent patterns for specific crimes like theft or assault?
4. Are there seasonal trends in crime patterns?
5. What are the geospatial hotspots for crime in San Francisco?

The dataset includes police department incident reports from 2018 to 2025. Each record contains attributes such as `incident_category`, `incident_time`, `incident_day_of_week`, `analysis_neighborhood`, `police_district`, and geographic coordinates (`latitude`, `longitude`). Additional derived features include `hour`, `day`, `month`, `year`, and various frequency-based aggregations, such as crime count by hour or district.

We enriched the dataset using two clustering techniques:

- **DBSCAN** was used to identify geospatial crime hotspots by clustering similar latitude/longitude coordinates.
- **KMeans** was applied to `incident_category` with frequency and time-based features to define higher-level `crime_type_group` labels for classification.

Exploratory Data Analysis (EDA) included heatmaps, bar plots, and log-scaled charts to examine patterns across time, location, and crime types. These visualizations informed feature engineering and revealed class imbalance and seasonal trends, laying the foundation for the machine learning models used later in the project.

3. Organization

This project followed a structured workflow aligned with the project phases (L1–L10):

- **L1–L2:** Data was cleaned, timestamp fields were split, and frequency features were created. Initial EDA with heatmaps and log plots helped reveal key spatial and temporal trends.
- **L3–L5:** DBSCAN and KMeans were used to cluster locations and crime types respectively. The result: two key features – `dbscan_cluster` (spatial group) and `crime_type_group` (classification target).
- **L6:** Visualizations such as PCA, Plotly scatterplots, and cluster-specific histograms were created to support pattern discovery and model decisions.
- **L8:** Four classifiers were trained: RandomForest, LogisticRegression, DecisionTree, and LSTM. Scaling, stratified train-test splits, and class weighting were applied appropriately.
- **L9:** Confusion matrices, F1-scores, and accuracy were used to evaluate and compare performance.
- **L10:** The results were compiled into a written report and presentation, with all visuals and conclusions summarized.

This structured approach ensured a reproducible, modular, and data-informed workflow from cleaning through deployment.

4.Data Cleaning & Preparation

The original dataset contained over 900,000 police incident reports with varying levels of completeness and format consistency. To prepare this data for modeling, we implemented several preprocessing steps:

- **Datetime Parsing:** Combined and split `incident_date` and `incident_time` into `hour`, `day`, `month`, and `year`.
- **Missing Values:** dropped rows with missing `incident_category`, `analysis_neighborhood`.

```
Rows before drop: 952287
Rows after drop: 899225
Rows dropped: 53062
5.57% of rows have been lost
```

	0
incident_category	0
analysis_neighborhood	0
police_district	0
incident_date	0
incident_day_of_week	0
longitude	0
latitude	0
hour	0
day	0
month	0
year	0
date_of_week	0

- **Label Encoding:** Converted categorical columns like `incident_category`, `analysis_neighborhood`, and `police_district` into numerical labels.
- Grouping Similar/Redundancy Crime:Combined Alike crime in to one categories

Weapons Carrying Etc, Weapons Offense, Weapons Offence to Weapons Offense

Original	Cleaned
`'Weapons Carrying Etc'`, `'Weapons Offense'`, `'Weapons Offence'`	`'Weapons Offense'`
`'Motor Vehicle Theft'`, `'Motor Vehicle Theft?'`	`'Motor Vehicle Theft'`
`'Human Trafficking (A), Commercial Sex Acts'`, `'Human Trafficking, Commercial Sex Acts'`	`'Human Trafficking – Commercial Sex Acts'`
`'Human Trafficking (B), Involuntary Servitude'`	`'Human Trafficking – Involuntary Servitude'`
`'Forgery And Counterfeiting'`	`'Forgery & Counterfeiting'`
`'Suspicious Occ'`, `'Suspicious'`	`'Suspicious Activity'`

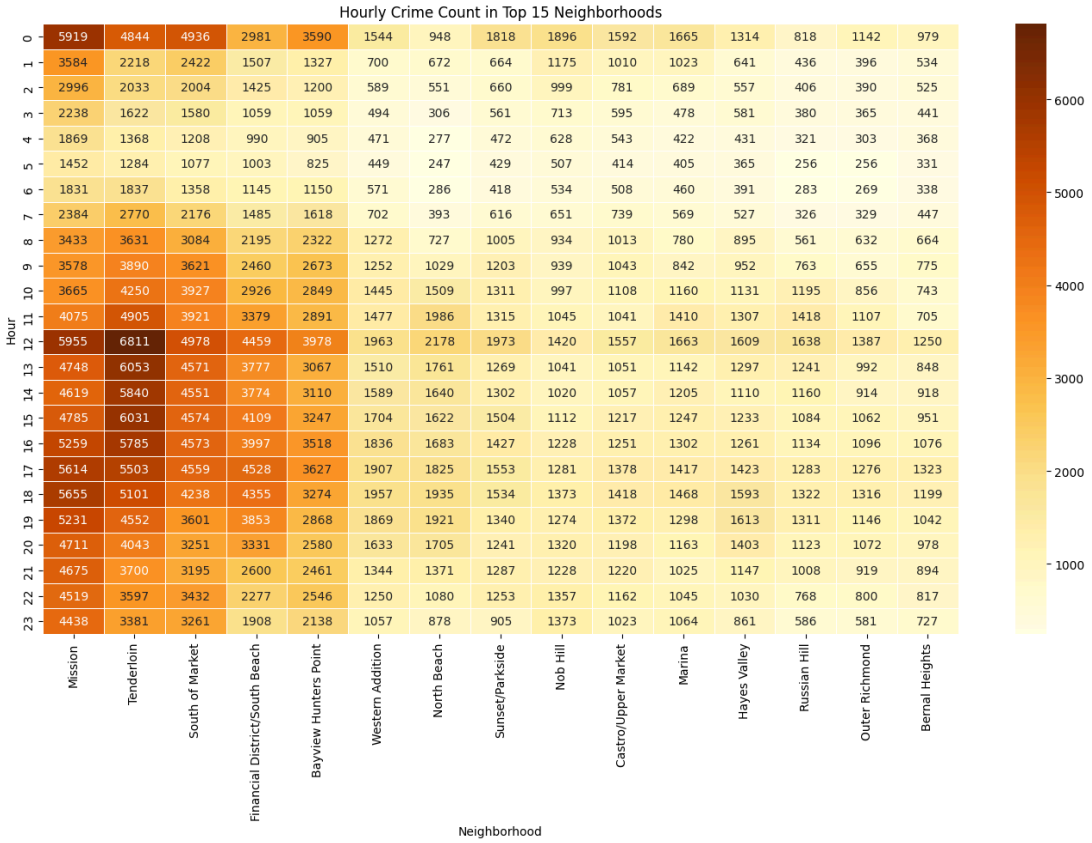
- **Derived Features:** Engineered additional columns such as:
  - `crime_count_hourly`
  - `crime_count_neighborhood`
  - `crime_count_police_district`
  - `crime_count_incident_category`
- **Spatial Clustering:** Used DBSCAN to identify geospatial crime hotspots and added the resulting `dbscan_cluster` as a spatial feature.
- **Crime Type Grouping:** Applied KMeans to create the `crime_type_group` target based on temporal and frequency characteristics of each incident type.

All cleaned and engineered features were stored in a master DataFrame `crime_count_df`, used for both static models and LSTM sequence generation.

3. Data Exploration & Visualization

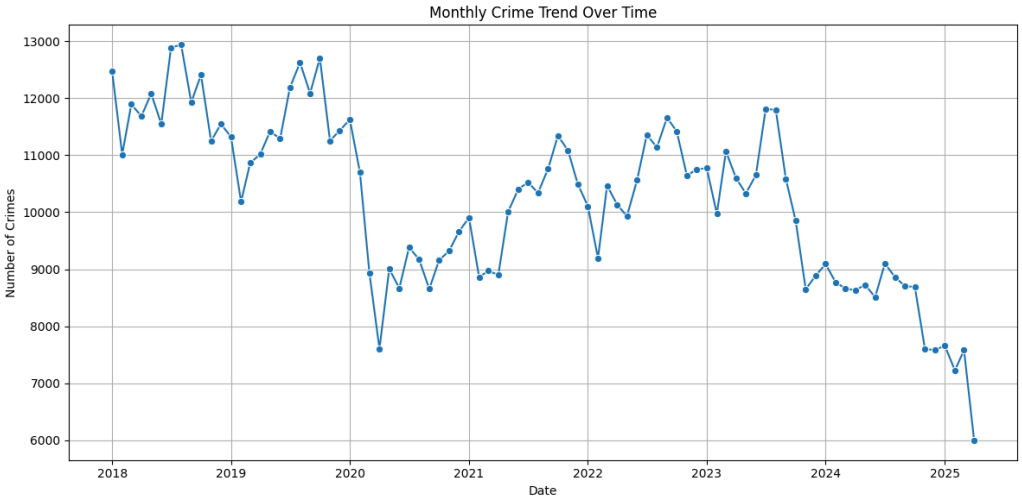
Exploratory Data Analysis (EDA) was conducted to understand trends in time, location, and crime category. The following visualizations and insights guided feature engineering and modeling:

- **Hourly and Weekly Heatmaps:** Showed that certain crimes peaked during late-night and weekend hours. Theft and assault were especially concentrated in evening hours.



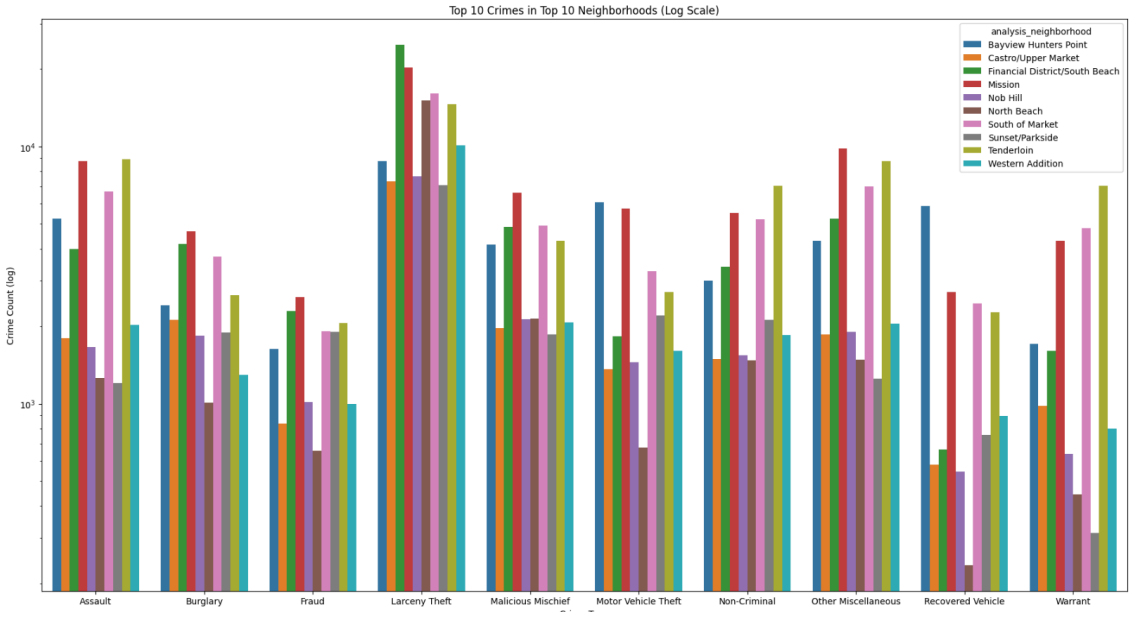
(The Heatmap showed clear trends of increased crime between late afternoon and late evening, with Mission and Tenderloin having the highest hourly counts)

- **Seasonal Trends:** Bar plots across months revealed an uptick in certain crime types during the summer, indicating potential seasonality.

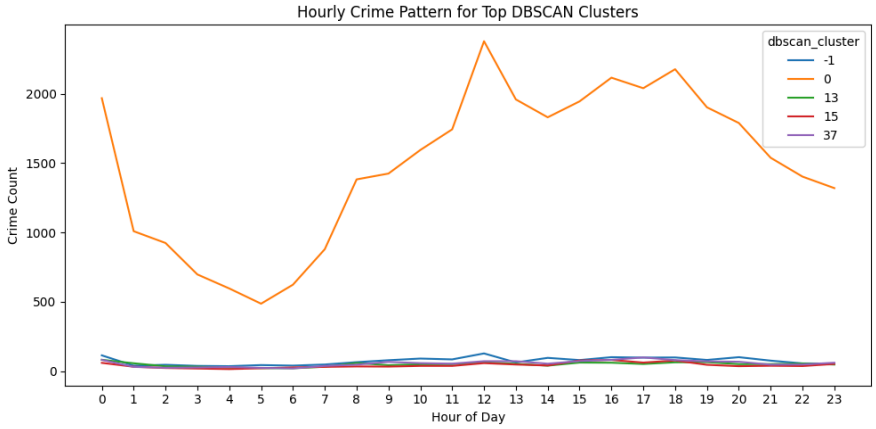


(The line chart displayed monthly crime counts from 2018 to 2025, revealing seasonal patterns and long-term trends. We observed a sharp decline around 2020, likely due to external factors like COVID-19)

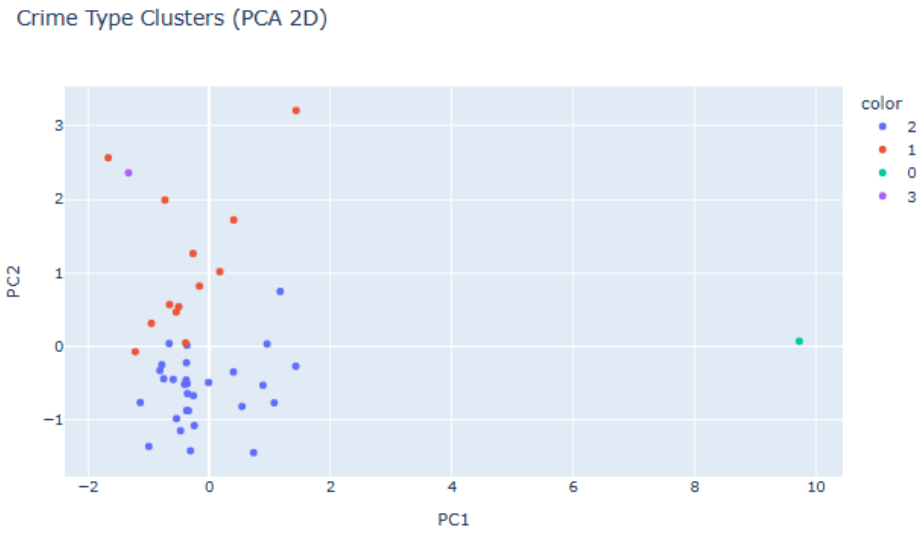
- **Neighborhood Log Distributions:** Used to identify crime-prone neighborhoods and support decisions to include `crime_count_neighborhood`.



- **Category Frequencies:** Showed extreme imbalance in `incident_category`, motivating KMeans grouping into `crime_type_group`.



- **PCA & Plotly (KMeans Visualization):** Implemented interactive map - Confirmed that clustering was effective in grouping crime categories with similar temporal behavior.



(Group0 :Human Trafficking - Involuntary Servitude
Group1 :Other Miscellaneous, Warrant, Drug Offense…

Group2 :Malicious Mischief, Assault, Burglary…

Group 3: Larceny Theft…)

These visual patterns informed feature selection and confirmed that both temporal and spatial factors play a key role in crime classification.

San Francisco is a densely populated urban center with diverse neighborhoods and varying crime dynamics. Historical crime data collected by the San Francisco Police Department provides detailed insight into when and where specific types of incidents occur. Our project aims to leverage this temporal and geographic information to identify consistent patterns and improve crime categorization through machine learning.

The dataset includes police department incident reports from 2018 to 2025. Each record contains attributes such as `incident_category`, `incident_time`, `incident_day_of_week`, `analysis_neighborhood`, `police_district`, and geographic coordinates (`latitude`, `longitude`). Additional derived features include `hour`, `day`, `month`, `year`, and various frequency-based aggregations, such as crime count by hour or district.

We enriched the dataset using clustering techniques:

- **DBSCAN** was used to identify geospatial crime hotspots by clustering similar latitude/longitude coordinates.
- **KMeans** was applied to `incident_category` with frequency and time-based features to define higher-level `crime_type_group` labels for classification.

Exploratory Data Analysis (EDA) included heatmaps, bar plots, and log-scaled charts to examine patterns across time, location, and crime types. These visualizations informed feature engineering and revealed class imbalance and seasonal trends, laying the foundation for the machine learning models used later in the project.

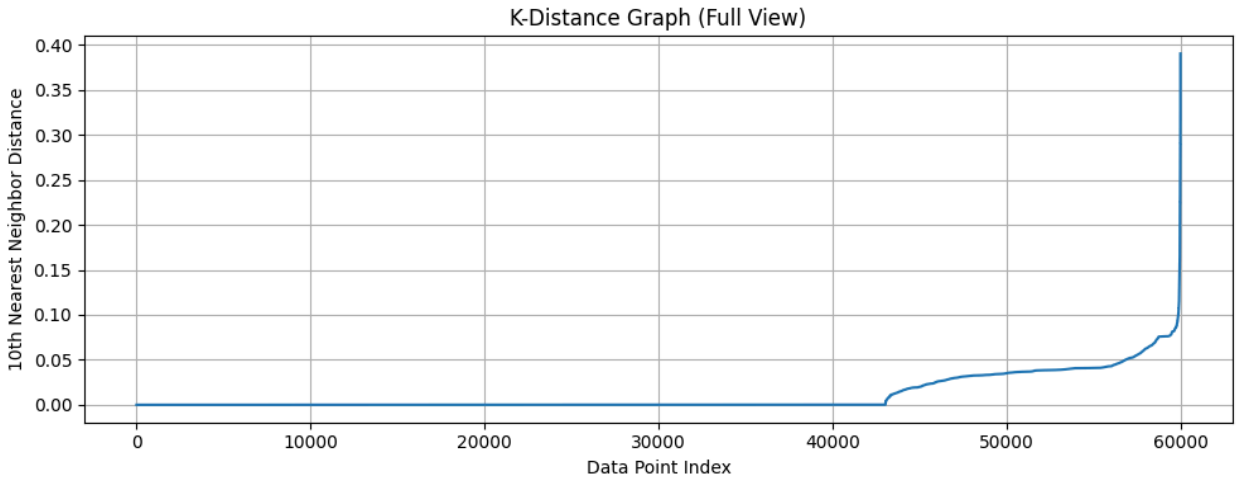
6. CLUSTERING

K-Distance for DBSCAN Parameter Tuning

We used a K-Distance Graph to determine the optimal `eps` value for the DBSCAN clustering algorithm. This graph plots the distance to each point's 10th nearest neighbor, helping identify the distance where points transition from dense clusters to noise.

The elbow point (around 0.05) indicates the best threshold for `eps`. This ensures that the DBSCAN algorithm effectively groups nearby crime incidents while filtering outliers.

This method was crucial for detecting meaningful spatial crime clusters in the San Francisco dataset.



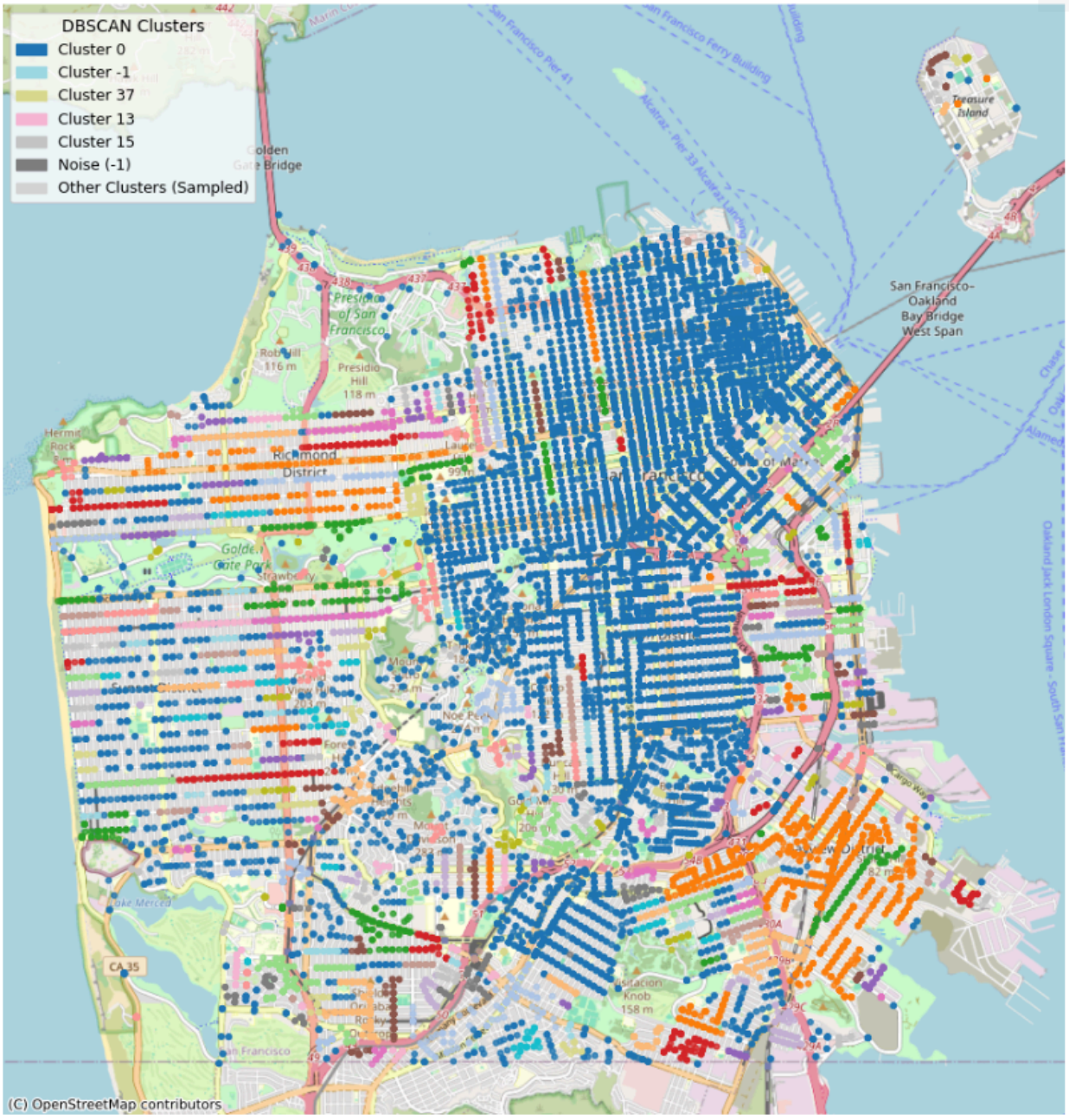
Hourly Crime Pattern for Top DBSCAN Clusters

We analyzed the hourly crime counts for the top DBSCAN clusters to understand their temporal patterns.

- Cluster 0 (the largest cluster) showed the highest crime activity, peaking around noon and early evening.
- Other clusters (13, 15, 37) had much lower and flatter patterns, indicating smaller or less active hotspots.

This analysis confirms that different crime clusters not only vary spatially but also have unique time-based trends, which is valuable for understanding when and where crime is most likely to occur.

1. DBSCAN Spatial Clusters of Crime in San Francisco
Focusing on Largest Clusters + Sampled Noise

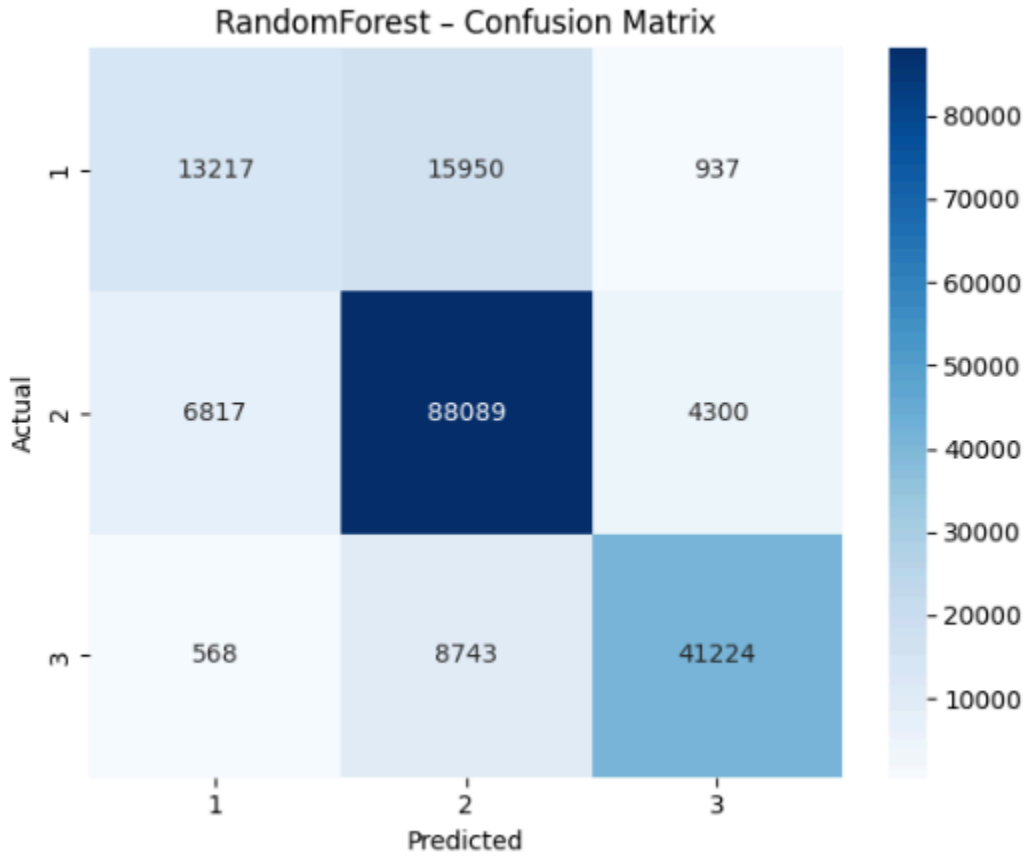


The first map gives location context—helping you understand which neighborhoods you're looking at.The second map uses DBSCAN to highlight crime hotspots based on actual clustering of incidents.This allows you to see both the crime distribution and how it relates to specific city regions, giving a clear visualization of where crime is most concentrated.

8. Model Training

RandomForest

	precision	recall	f1-score	support
1	0.64	0.44	0.52	30104
2	0.78	0.89	0.83	99206
3	0.89	0.82	0.85	50535
accuracy			0.79	179845
macro avg	0.77	0.71	0.73	179845
weighted avg	0.79	0.79	0.78	179845



We tested LogisticRegression, DecisionTree, LSTM, and RandomForest gave the best results. Based on accuracy and F1-scores, RandomForest is the best-performing model, achieving 79% accuracy and the highest scores across all classes.

This confusion matrix shows how well the RandomForest model classified the crime type groups. Most class 2 cases were correctly predicted. There is some confusion between class 1 and class 2, and between class 3 and class 2.

9. Challenges

Parameter Tuning for DBSCAN:
Choosing the optimal eps value was tricky. We used a K-Distance Graph to guide our selection, but it required trial and error to correctly interpret the elbow point.

Data Imbalance:
Some crime categories and clusters had very few data points, making it difficult to build balanced and accurate models.

Finding the Right Approach:
One of the biggest challenges was figuring out the best way to approach the problem. We explored different clustering methods, visualizations, and machine learning models before deciding what was most effective for our data.