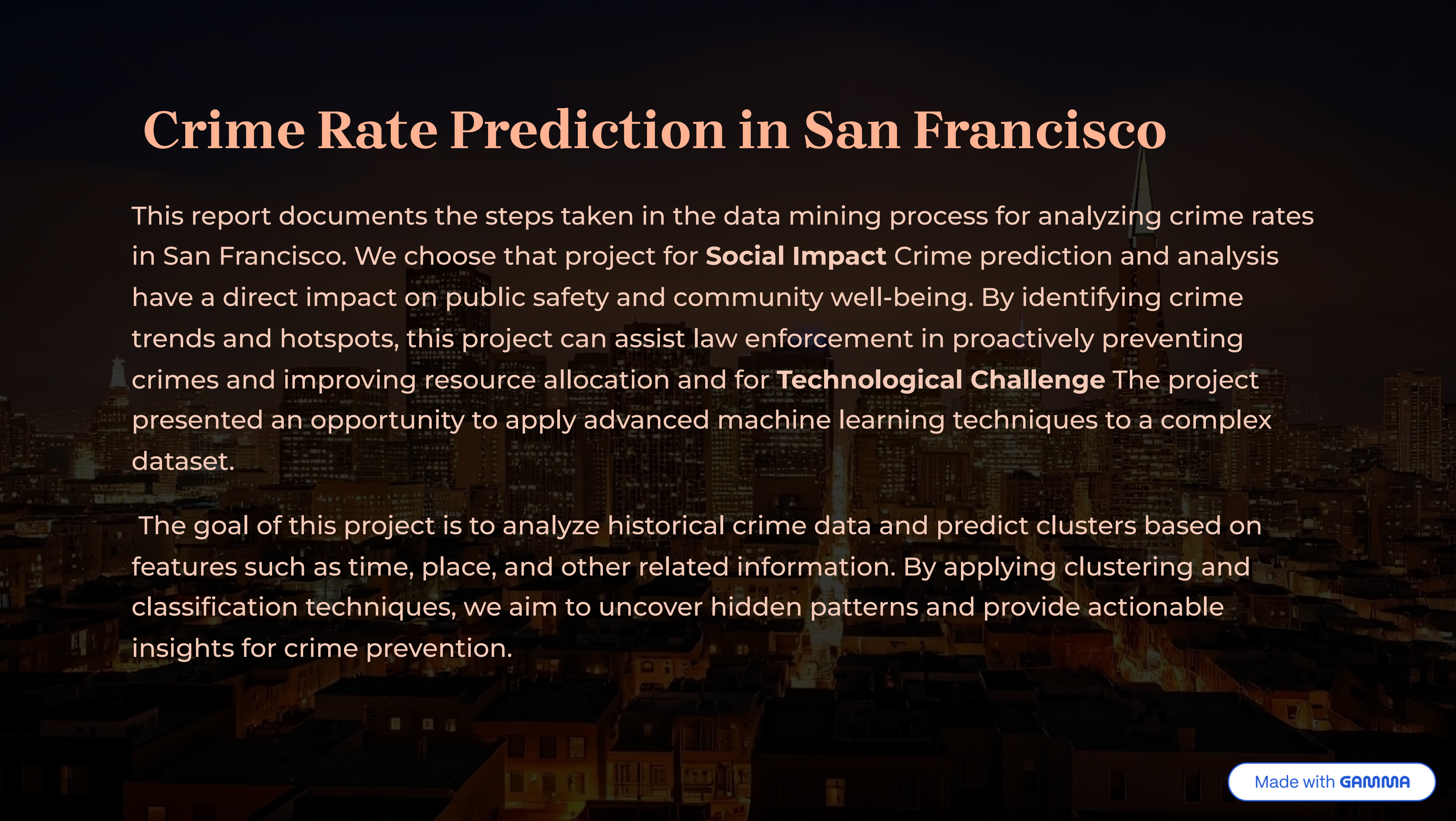
Crime Rate Prediction in San Francisco



This report documents the steps taken in the data mining process for analyzing crime rates in San Francisco. We choose that project for **Social Impact** Crime prediction and analysis have a direct impact on public safety and community well-being. By identifying crime trends and hotspots, this project can assist law enforcement in proactively preventing crimes and improving resource allocation and for **Technological Challenge** The project presented an opportunity to apply advanced machine learning techniques to a complex dataset.

The goal of this project is to analyze historical crime data and predict clusters based on features such as time, place, and other related information. By applying clustering and classification techniques, we aim to uncover hidden patterns and provide actionable insights for crime prevention.

Data Overview and Preprocessing

A sample of 15,000 rows is selected randomly to reduce computational overhead.

1 Inspecting the Data

- Initial exploration of the dataset revealed **9 columns**:
  - **Dates**, **Category**, **Descript**, **DayOfWeek**, **PdDistrict**, **Resolution**, **Address**, **X**, and **Y**.
- No missing values or duplicates were found in the sampled dataset.

2 Feature Engineering

- The **Dates** column is split into two new features:
  - **Date**: Extracted as a date component.
  - **Time**: Extracted as a time component.
- The original **Dates** column is dropped.

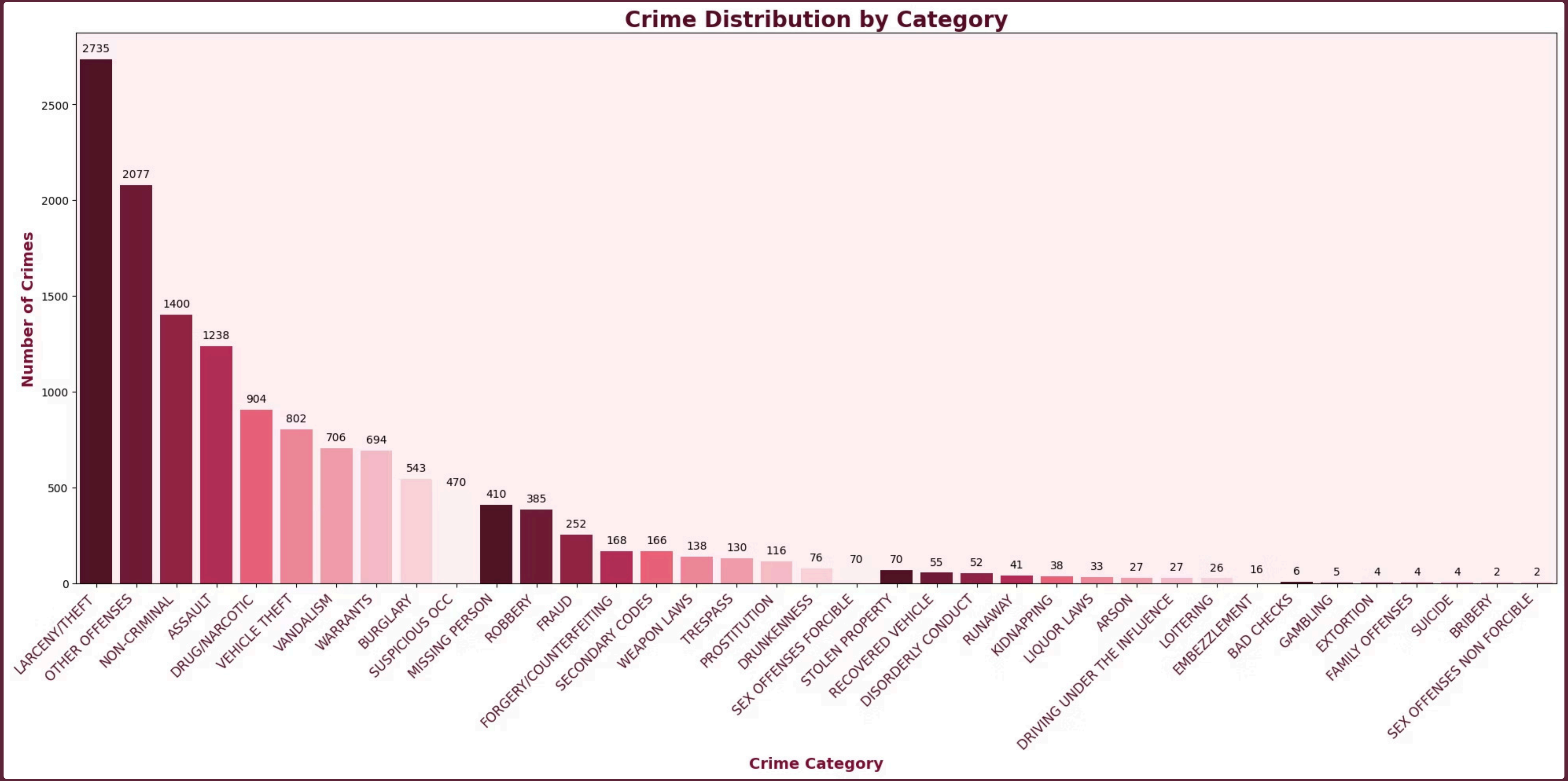
3 Outlier Detection and Removal

- Outliers in the geographical coordinates (**X** and **Y**) are detected using the **Interquartile Range (IQR)** method.
- Rows with outlier values are filtered out to ensure clean data for analysis.

Visualization

1. Crime Distribution:

- Key observation:
  - **LARCENY/THEFT** is the most frequent crime category.



2. Resolution of Crimes

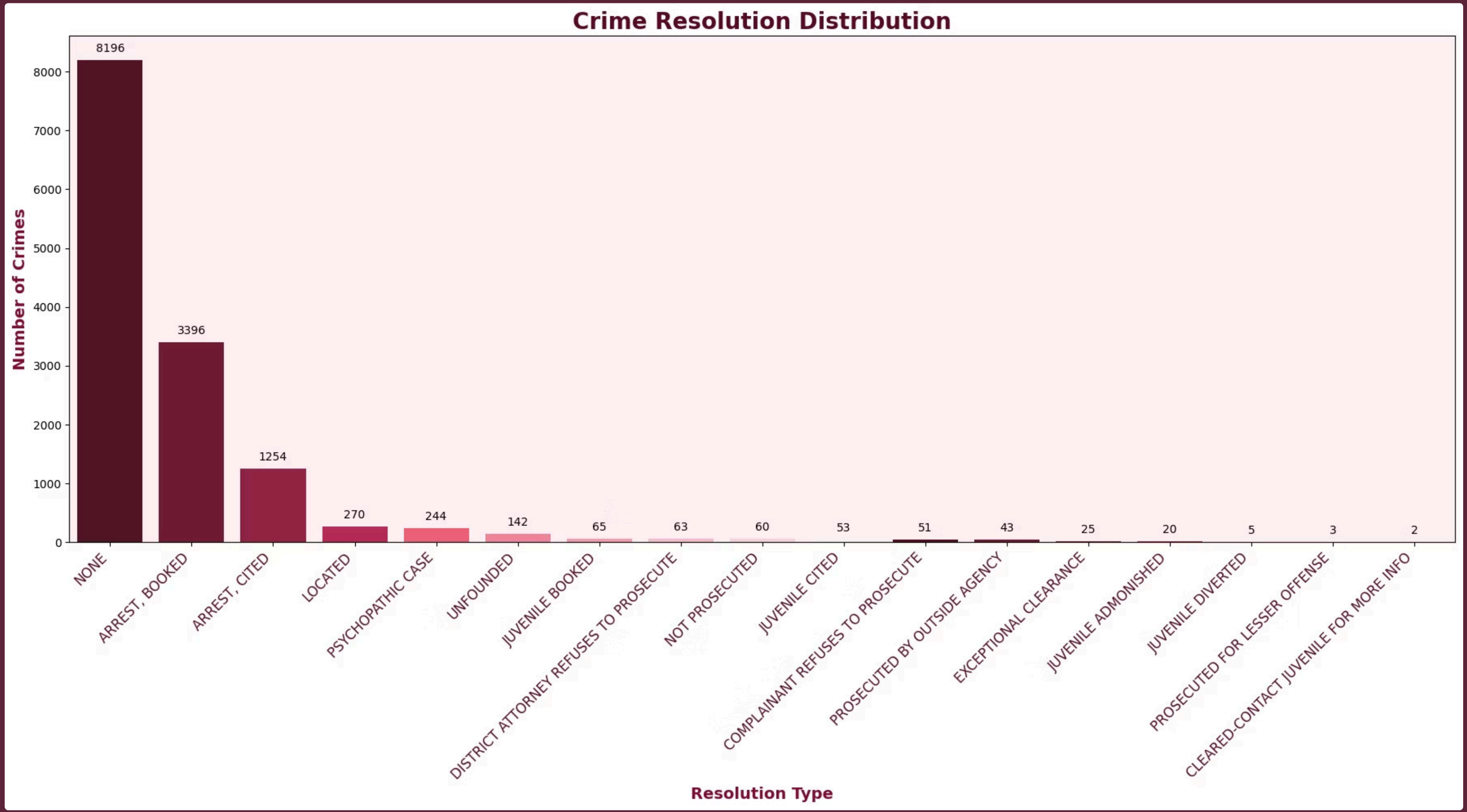
Key Observations:

a. **Resolution Types**:

- The most common resolution type is **"NONE"**, representing 8,196 instances, indicating unresolved cases.

b. **Less Frequent Resolutions**:

- Other resolution types, such as **"LOCATED" (270)**, **"PSYCHOPATHIC CASE" (244)**, and **"UNFOUNDED" (142)**, have significantly lower frequencies.
- Rare resolutions include **"CLEARED-CONTACT JUVENILE FOR MORE INFO" (2)** and **"PROSECUTED FOR LESSER OFFENSE" (3)**.



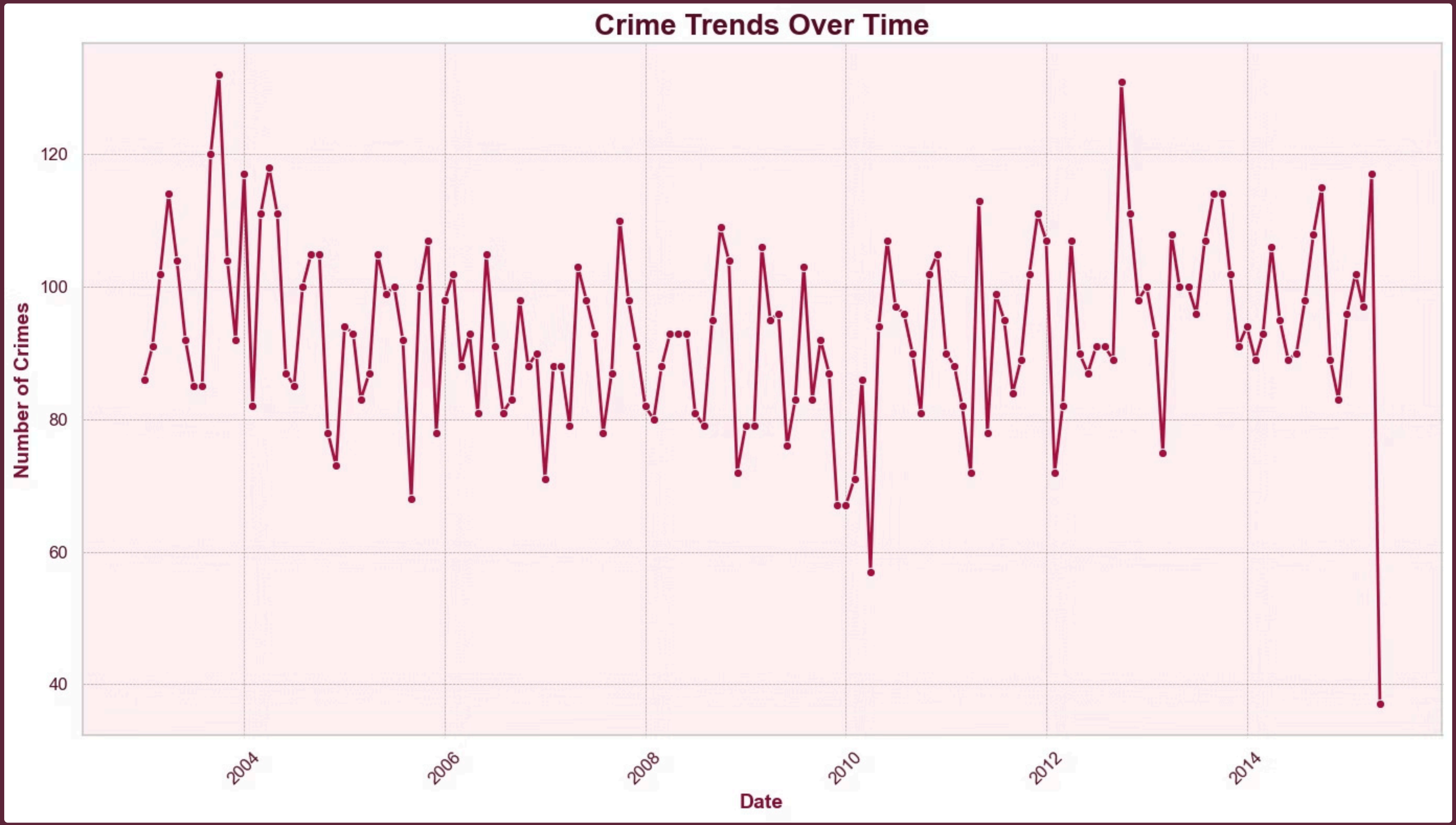
3. Crimes over time

a. **Trend Overview**:

- The number of crimes fluctuates between approximately **40** and **120** per month across the timeline.
- There is no consistent upward or downward trend, but there are periodic spikes and dips indicating variability in crime rates over time.

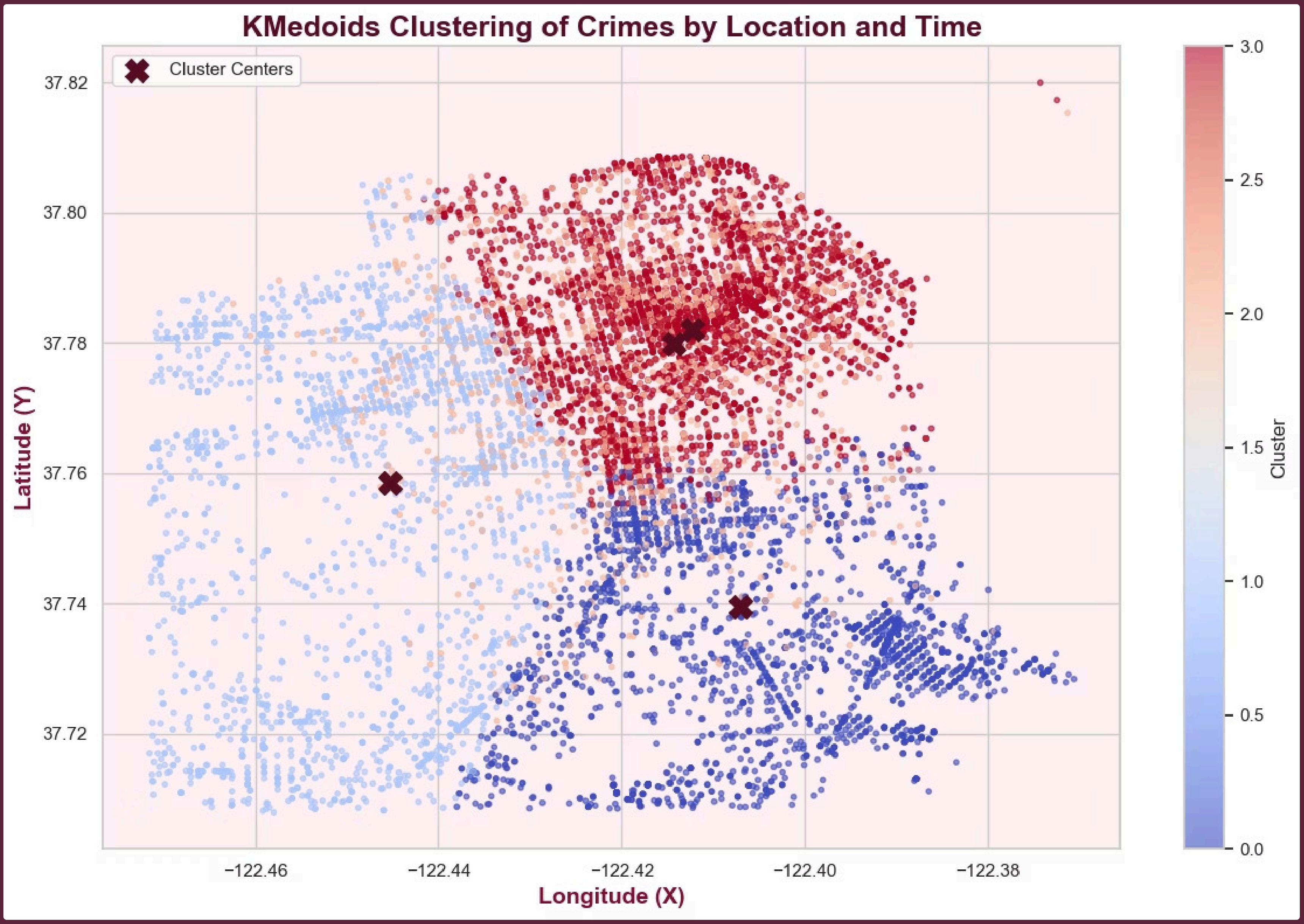
b. **Peaks and Valleys**:

- Some months show significant peaks in crime rates (around **2004**, **2007**, and **2013**).
- There are sharp declines during certain months, creating valleys in the chart.



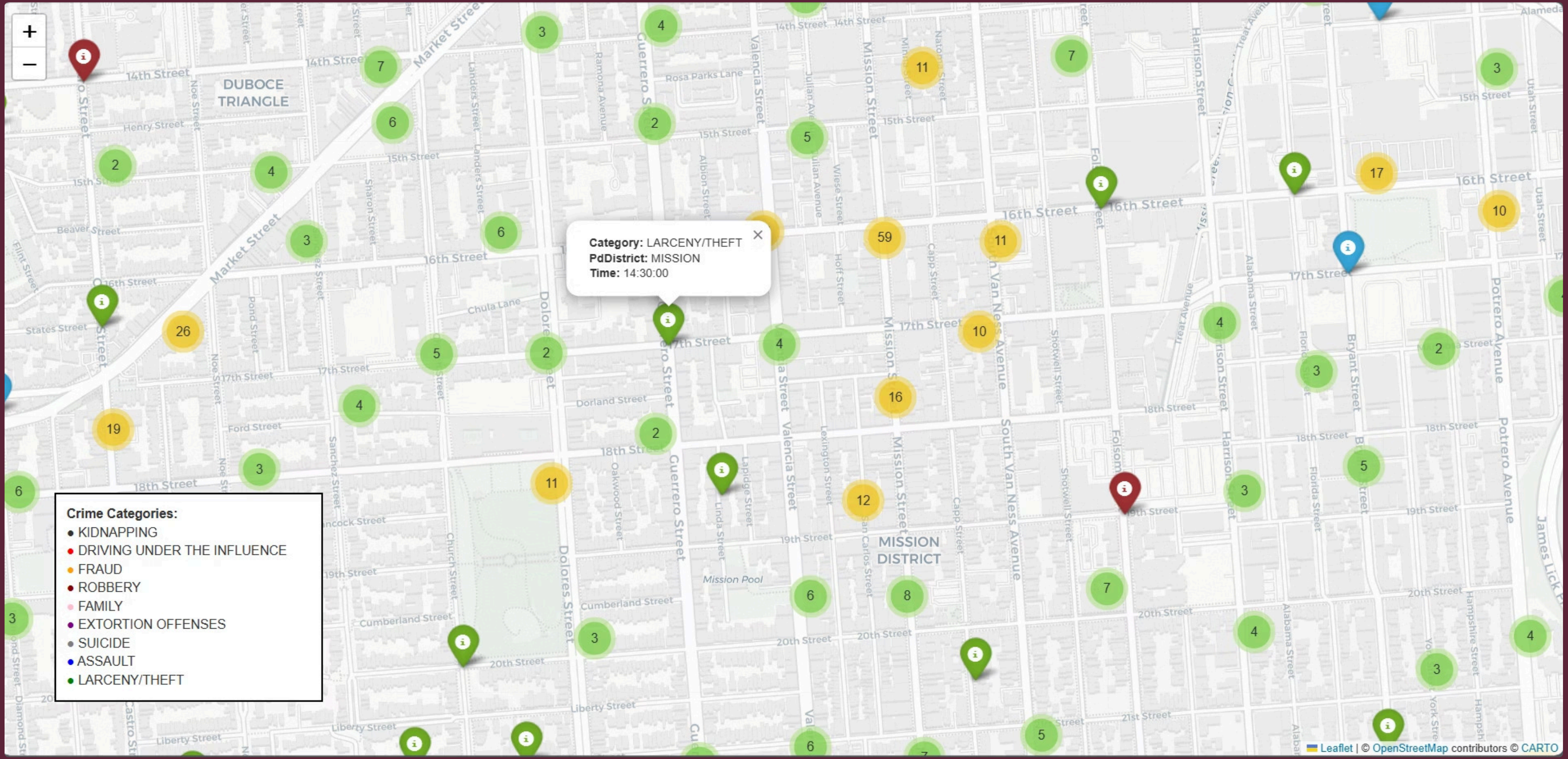
Geographical Clustering

- **Clustering Results**:
  - Using **K-Medoids** clustering, crime locations were grouped into 4 clusters based on geographical coordinates.
  - Cluster centers were identified in high-density areas, particularly in regions like the **Tenderloin** and **Mission** districts..
- **Geographical Crime Patterns**:
  - The clustering reveals distinct geographical areas where crimes are concentrated.
  - For example, the **red cluster** indicates a densely populated crime hotspot, likely corresponding to a high-crime area in the city.
- **Insight**:
  - Urban planners can use this information to improve lighting, surveillance, and community programs in high-crime areas
- **Application**:
  - This clustering can be used to allocate law enforcement resources more effectively by targeting high-crime clusters.
  - Urban planners can also use this information to improve safety measures in specific areas.



Cluster Map Visualization

interactive map visualization showing the **geospatial distribution of crimes** in a specific area, likely part of San Francisco. The map uses markers, colors, and tooltips to provide detailed information about crime incidents and categories.



Category: LARCENY/THEFT
PdDistrict: MISSION
Time: 14:30:00

Crime Categories:
- ● KIDNAPPING
- ● DRIVING UNDER THE INFLUENCE
- ● FRAUD
- ● ROBBERY
- ● FAMILY
- ● EXTORTION OFFENSES
- ● SUICIDE
- ● ASSAULT
- ● LARCENY/THEFT

Metric Evalution For Clustering

1. Silhouette Score :

- **Purpose:** Measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).
- **Usage**: It helps determine the **best number of clusters**. Higher silhouette scores indicate better-defined clusters.
- **Range:** Between -1 and 1.

  1: Perfect clustering where points are well-separated and tightly packed within their clusters.

  0: Overlapping clusters.

  -1: Poor clustering where points are closer to other clusters than their own.

2. Davies-Bouldin Index (DBI)

- **Purpose**: Measures the average similarity ratio of each cluster with its most similar cluster.
- **Range**: 0 to ∞ (lower is better).
  Smaller values indicate better clustering with well-separated, compact clusters.
- **Usage:** A low DBI value indicates compact, well-separated clusters.

3. Calinski-Harabasz Index (CH Index)

- **Purpose**: Measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion.
- **Range**: **Higher values are better**, indicating well-defined clusters.
- **Usage:** Helps evaluate how well-separated and compact the clusters are. Higher values suggest better clustering.

Classification with CatBoost

To predict cluster membership for new crime incidents, a supervised machine learning approach was implemented. CatBoost model was trained on features including day of week, time of day, police district, and incident type, aiming to classify incidents into one of the K-Medoids derived clusters.

The classification process utilized stratified training and testing splits to ensure representative model evaluation. Performance metrics **accuracy** with the final model achieving an impressive overall accuracy of 98%, validating the feature set's predictive power.

Conclusion

This project was chosen due to its potential to make a tangible difference in public safety, the challenging nature of the data analysis and modeling tasks, and its ability to showcase the power of data-driven insights in solving real-world problems. It combines social responsibility with technical innovation, making it a highly impactful and relevant endeavor.