CRIME RATE IN SAN FRANCISCO

San Francisco Crime Data Analysis

A dataset containing information on crimes in San Francisco was analyzed using machine learning techniques to recognize patterns and predict crime types. The data includes details such as crime type, location, time, and region, allowing for analysis of geographic and temporal patterns of crime. This data was selected because it provides insights into crimes and can be used to build predictive models to improve the allocation of security resources. The analysis included data cleaning, the use of clustering techniques such as hierarchical clustering and K-medoids, and crime type prediction using decision trees, with the resulting results evaluated

Why We Chose This Dataset?

Real-World Relevance: Crime is a serious social issue with direct impact on public safety. Predictive analysis can help authorities allocate resources more effectively and prevent crime

Rich and Structured Data: The dataset contains detailed information such as category of crime, location, time, and police district. This makes it ideal for applying various data mining techniques like classification and clustering

Educational Value: This dataset allowed us to apply the full data mining pipeline from preprocessing to visualization—giving us practical hands-on experience with important concepts in the course

Geospatial Analysis: The inclusion of location coordinates (latitude and longitude) made it possible to explore spatial patterns, which added depth to the analysis

Difficulties We Faced:

Data Imbalance: Some types of crimes were much more frequent than others which affected the performance of classification models

Time Series Complexity: Predicting category crime by time of day or day of week involved additional complexity and required careful feature engineering

High Dimensionality: The dataset had many categorical features (e.g., crime categories, police districts) that had to be encoded, which increased the data's dimensionality and model complexity

Interpretability: Some models performed well but were hard to interpret, making it challenging to extract actionable insights for law enforcement

Data Cleaning and Exploration

The data cleaning and processing phase is a critical step to ensure the quality and reliability of subsequent analyses. In this project, special attention was paid to preparing the training and test datasets, **train_df** and **test_df**, to ensure the highest levels of accuracy and consistency

The initial structure of both datasets was examined. The training dataset, **train_df** contains **878,049 entries and 9 columns**, while the test dataset, **test_df** contains **884,262 entries and 7 columns**. Initial examination showed that none of the columns contained missing values. **The Dates column** in both datasets was converted to datetime format, enabling the extraction of new temporal features such as **year**, **month**, **day**, **hour**, **minute**, and **second**. These extracted features enhance the ability to analyze seasonal and daily patterns of crime

```
train_df.head()
```

Python

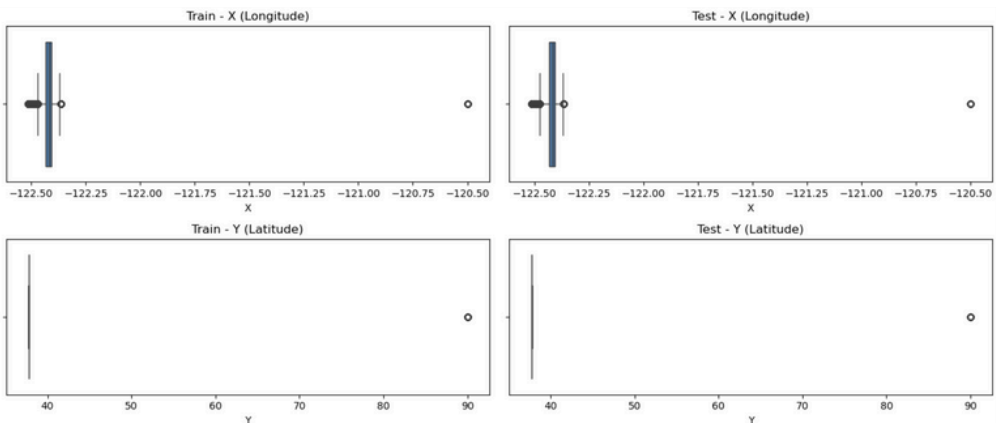
	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y	Year	Month	Day	Hour	Minute	Second
0	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599	2015	5	13	23	53	0
1	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599	2015	5	13	23	53	0
2	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414	2015	5	13	23	33	0
3	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873	2015	5	13	23	30	0
4	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.438738	37.771541	2015	5	13	23	30	0

```
test_df.head()
```

Python

	DayOfWeek	PdDistrict	Address	X	Y	Year	Month	Day	Hour	Minute	Second
0	Sunday	BAYVIEW	2000 Block of THOMAS AV	-122.399588	37.735051	2015	5	10	23	59	0
1	Sunday	BAYVIEW	3RD ST / REVERE AV	-122.391523	37.732432	2015	5	10	23	51	0
2	Sunday	NORTHERN	2000 Block of GOUGH ST	-122.426002	37.792212	2015	5	10	23	50	0
3	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.437394	37.721412	2015	5	10	23	45	0
4	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.437394	37.721412	2015	5	10	23	45	0

Next, the focus was on addressing duplicate rows and outliers that could negatively impact the analysis. **2,323 duplicate rows** were detected in **train_df** and **188,352duplicate rows** in **test_df**, which **were removed** to ensure data uniqueness Furthermore, **outliers in the geographic location coordinates (X and Y)** were identified using box plot analysis and the interquartile range method. Extreme outliers were observed in both columns for both datasets. **67 outliers** were identified in **train_df** and **54** in **test_df**, where both X and Y coordinates were extreme Interestingly, **3,618 of these outliers** were **common to both the training and test sets** These **outliers were removed**, **resulting** in two clean datasets (**train_clean** and **test_clean**) free of these outliers, improving the accuracy of any future spatial analysis



Hierarchical Clustering

In this stage of the project, Hierarchical Clustering was employed as a method of unsupervised learning to uncover latent structures in the spatial and temporal patterns of crime occurrences in San Francisco. This method offers a tree-like model of data similarity, allowing for clear visualization of how data points group together at different levels of granularity

Training Data: Preparation & Training and Visualitaion

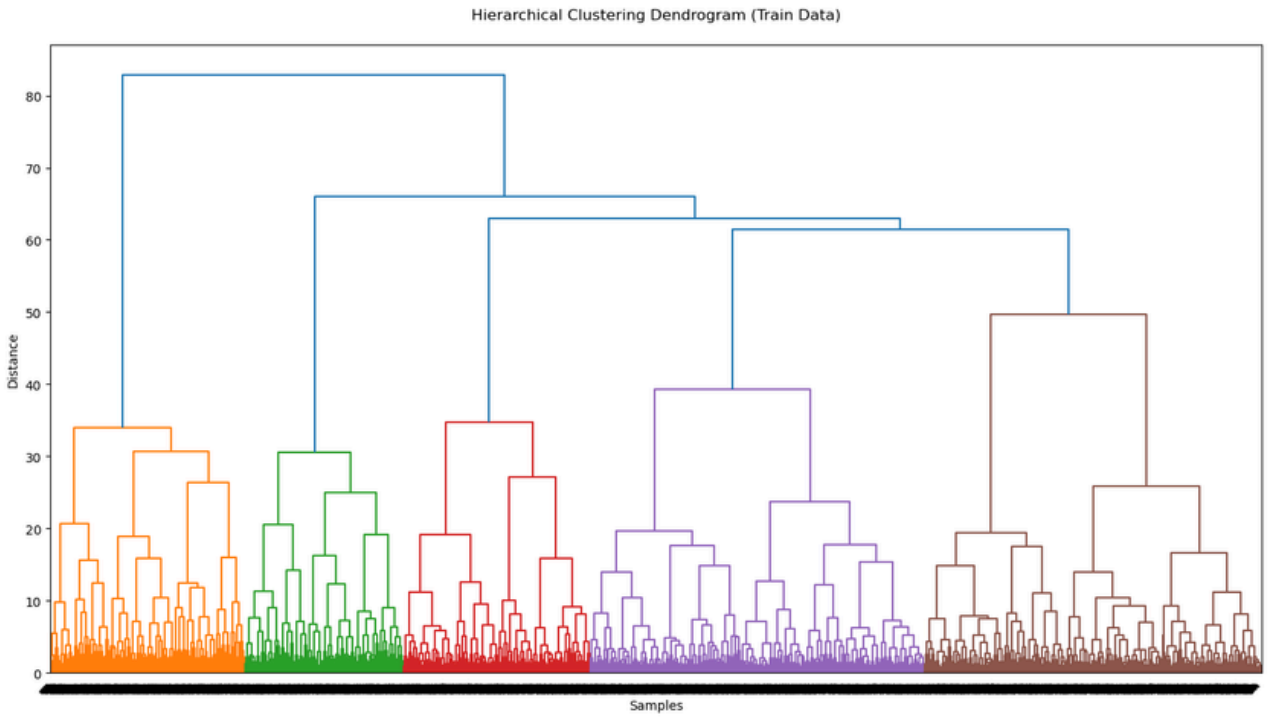
We started by selecting **five main features** that represent the dimensions of time and space

Geographically: **X, Y**
Temporally: **Hour, Day, Month**
We applied a **StandardScaler** to standardize the values, as each feature has a different unit of measurement, which can affect the clustering results

Next, we **randomly sampled 5,000** records from the training data to reduce computational overhead and improve performance

We then applied the **linkage** function using the **Ward** method to construct a hierarchical tree representing the similarity of the data
We plotted a dendrogram on the selected data to determine the appropriate number of clusters

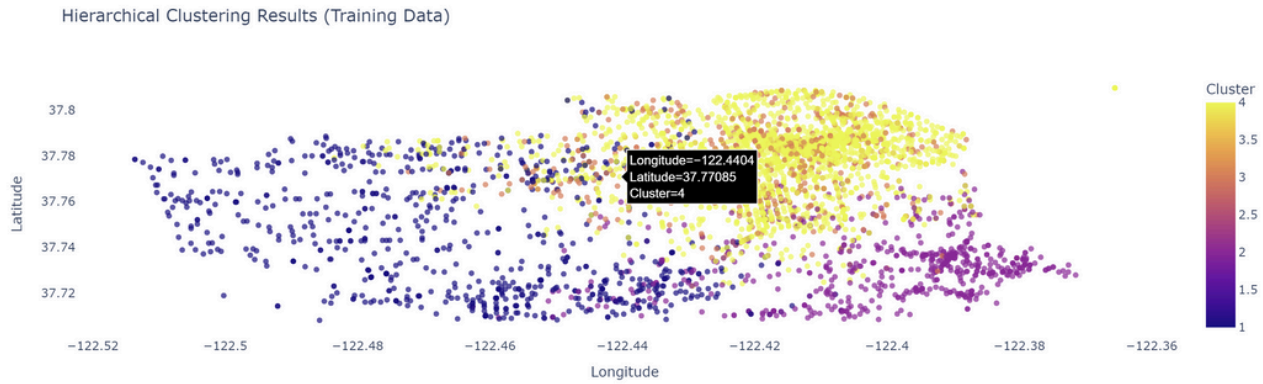
We found significant spacing between some branches, which helped us decide on **four clusters** because they strike a balance between accuracy and simplicity



Using the **fcluster function**, we divided the data into **four clusters**.
We then used **plotly.express** to create an **interactive Scatter Plot** with different colors for each cluster

The results revealed **distinct geographic clusters**:

One cluster **centered around the city cente**r, which could be an indicator of high crime density areas
Other clusters were **scattered along the outskirts**, which could indicate a different temporal distribution or type of crime

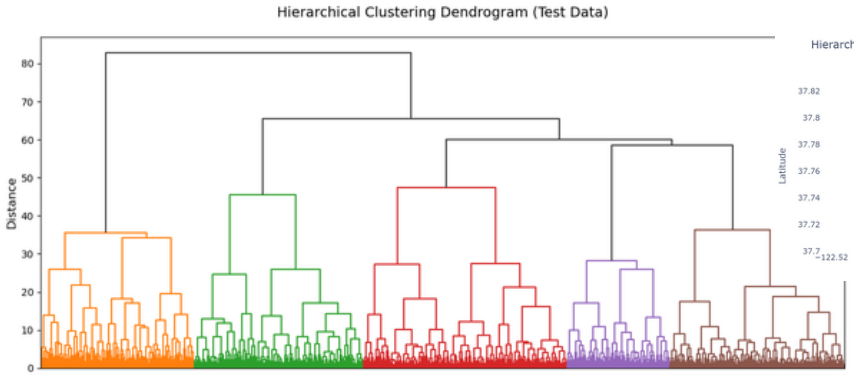


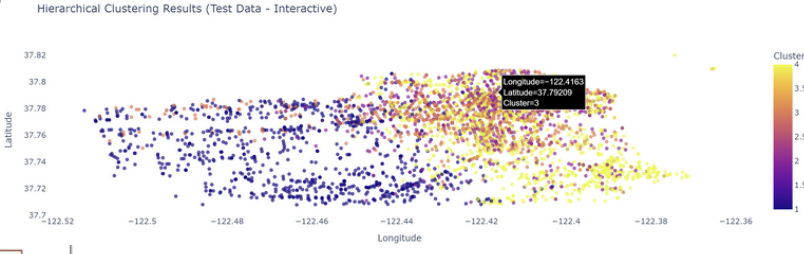
Training Data: Preparation & Training and Visualitaion

We repeated exactly the same preparation steps on the test data:
The same five attributes **(X, Y, Hour, Day, Month)**
The same standardization method using **StandardScaler**
And the same **sample size** (**5,000 records**)

The goal here was to **verify** that the **clustering structure** we obtained from the **training data** was **not just a coincidence**, and that it actually reflects a consistent pattern.
We drew a **dendrogram** for the test data using the same method (**Ward linkage**). Indeed, the tree structure was very similar to the one we obtained in the training data, supporting the fact that the distribution of crimes follows a recurring pattern.

We applied the same number of clusters **(4 clusters)** to the **test data**, and used the same **Plotly** visualization method. The result was: The clusters appeared in roughly the same locations as those seen in the training data. **The central regions remained distinct**, and the outer regions also maintained their distribution. This symmetry in the partitioning gives us confidence that the model can generalize well and is not affected by specific data.



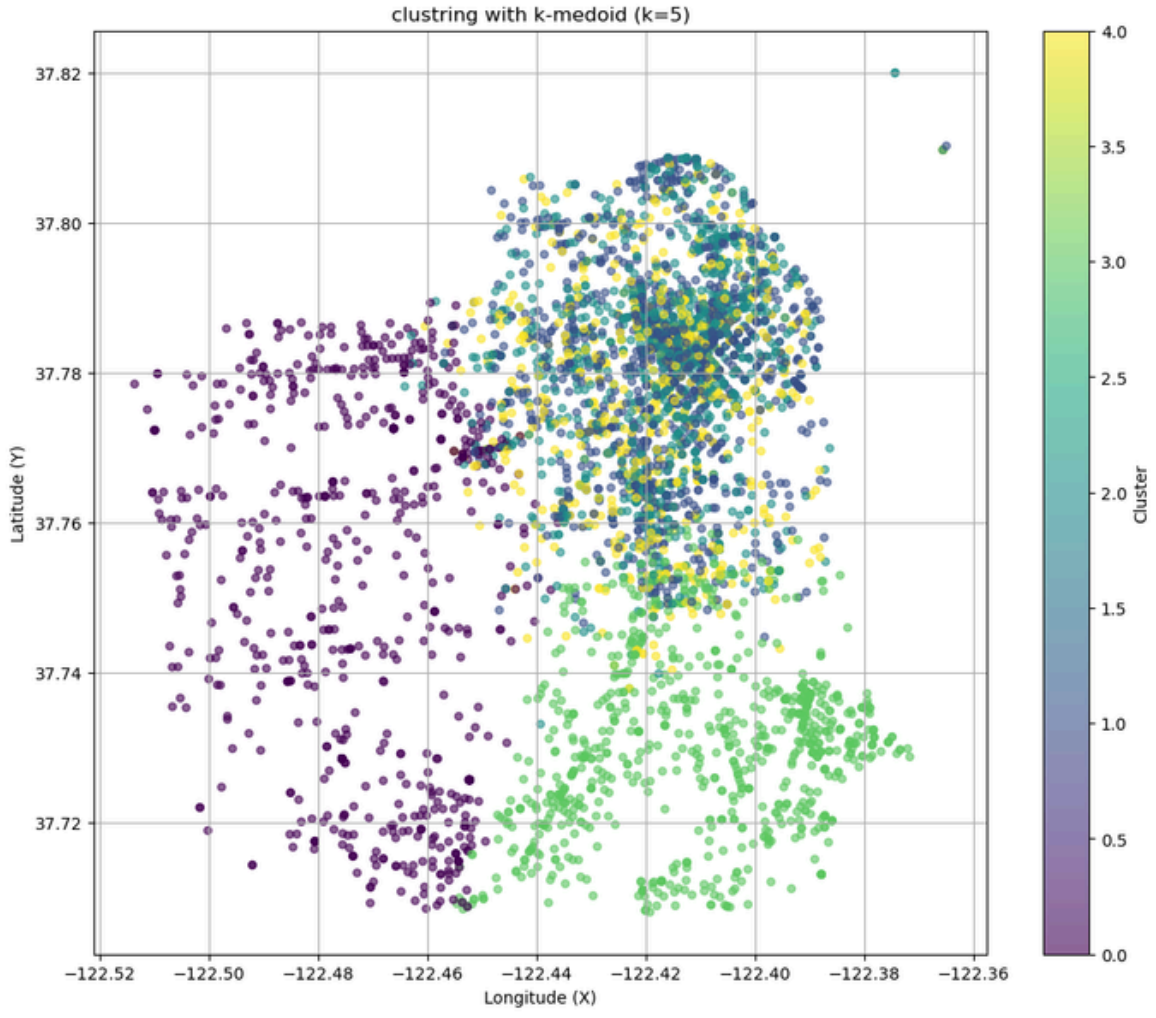


K-Medoids Clustering

The objective of this task is to determine **the optimal number of clusters (k)** for applying the **K-Medoids clustering algorithm** on the dataset. To improve computational efficiency and scalability, the process is initiated with a smaller random sample of the dataset and the sample size is increased gradually to verify the consistency of the chosen k

Features used for clustering: **X, Y, Hour, and Category_Encoded**
Standardization is applied using **StandardScaler** to normalize the feature values
Clustering Procedure

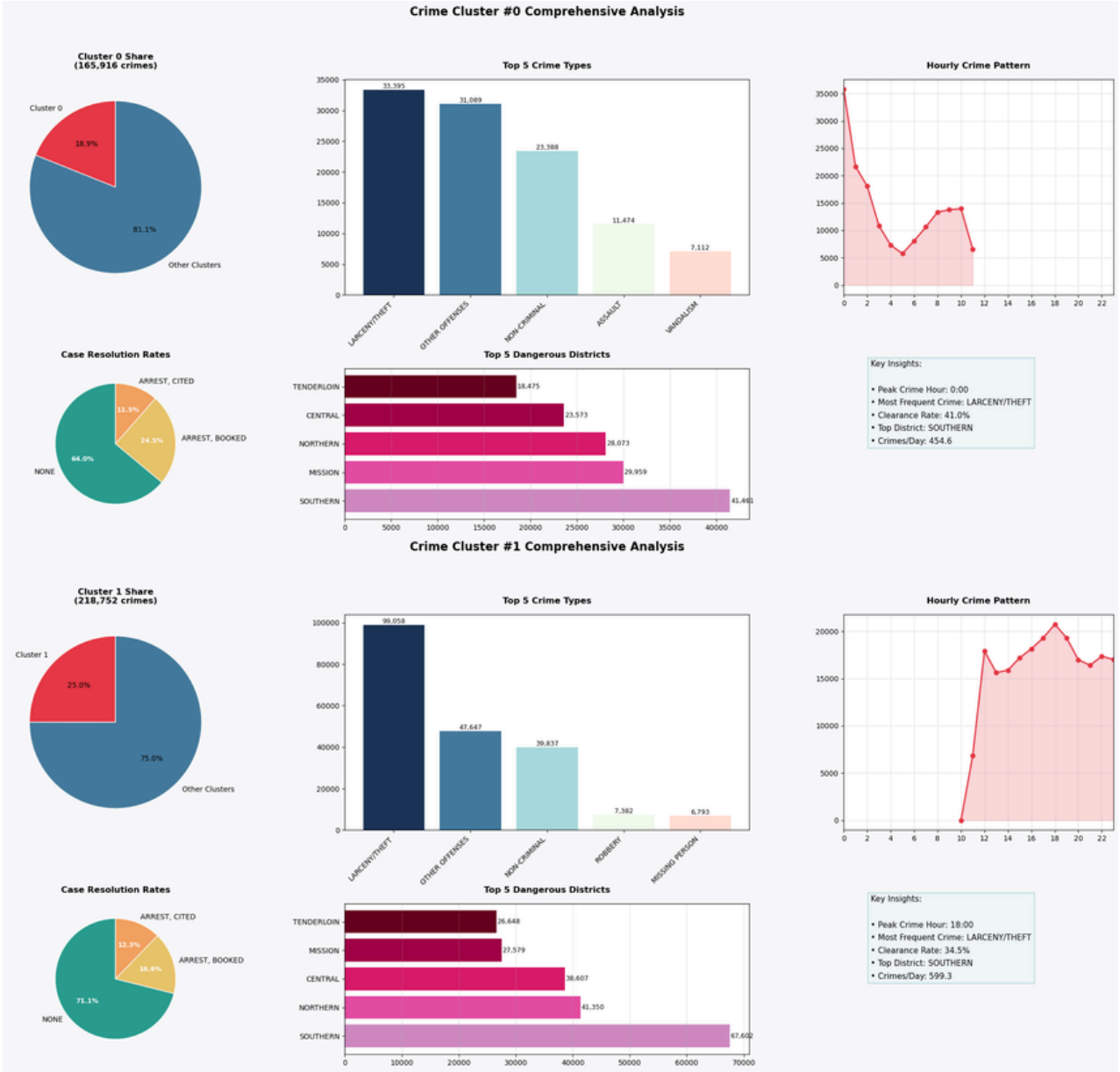
For each iteration:
1,000-record sample:
Best **K=5** (silhouette index: **0.2590**)
Performance improved from **K=2 (0.2035) to K=5 (0.2590)**, then decreased.
5,000-record sample:
Best **K=5** (silhouette index: **0.2360**).
Confirmation that **K=5** remains optimal with more data.
7,000-record sample:
Best **K=6** (silhouette index: **0.2230**).
Slight change due to **increased** data diversity, but **K=5 remained a strong choice**.
Final Decision
(We chose **K=5** because it achieved a balance between accuracy and clarity



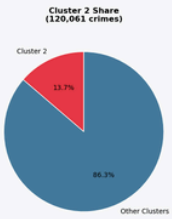
After setting **K=5**, we applied the model to the entire dataset **(~800,000 records)**. Cluster **1** was the **largest**, indicating an area with a **high crime density**.

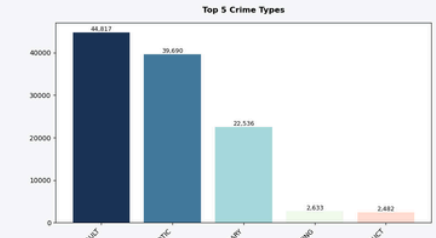
Cluster	Number of crimes	Ratio
1	218,752	~24.5%
0	165,916	~18.6%
5	136,193	~15.3%
2	120,061	~13.5%
4	117,800	~13.2%
3	116,937	~13.1%

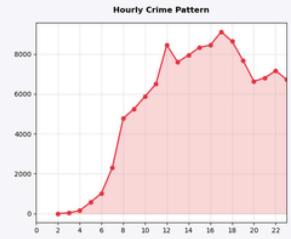
An integrated function was developed to **visualize the characteristics of each identified crime cluster**. This function displays a comprehensive analytical dashboard containing six key elements for each cluster:

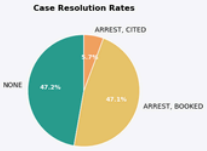


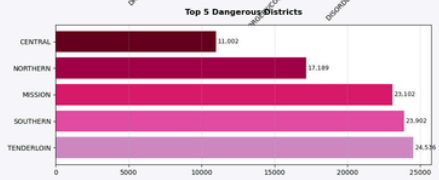
Crime Cluster #2 Comprehensive Analysis







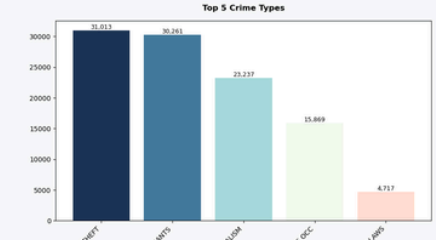


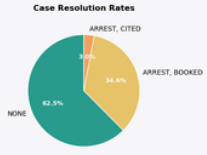


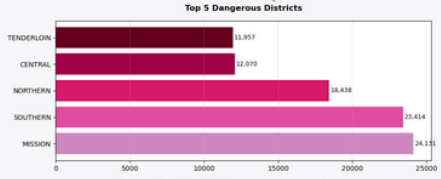
Key Insights:

- Peak Crime Hour: 17:00
- Most Frequent Crime: ASSAULT
- Clearance Rate: 54.8%
- Top District: TENDERLOIN
- Crimes/Day: 328.9

Crime Cluster #3 Comprehensive Analysis



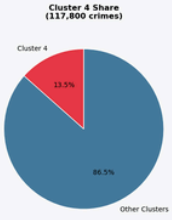


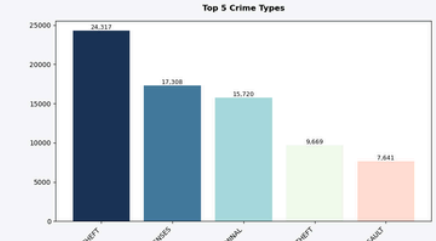


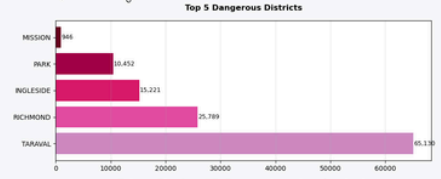
Key Insights:

- Peak Crime Hour: 18:00
- Most Frequent Crime: VEHICLE THEFT
- Clearance Rate: 40.5%
- Top District: MISSION
- Crimes/Day: 320.4

Crime Cluster #4 Comprehensive Analysis



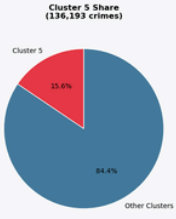


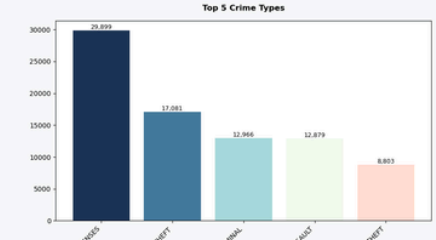


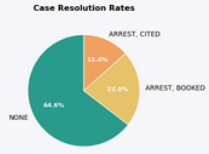
Key Insights:

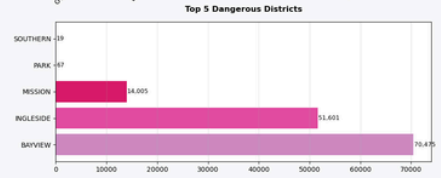
- Peak Crime Hour: 12:00
- Most Frequent Crime: LARCENY/THEFT
- Clearance Rate: 30.7%
- Top District: TARAVAL
- Crimes/Day: 322.7

Crime Cluster #5 Comprehensive Analysis









Key Insights:

- Peak Crime Hour: 12:00
- Most Frequent Crime: OTHER OFFENSES
- Clearance Rate: 42.4%
- Top District: BAYVIEW
- Crimes/Day: 373.1

Crime Type Prediction using Decision Tree

A decision tree is an algorithm similar to the game **"Guess Who?"** , Imagine you have a **series of questions** that gradually lead you to determine the type of crime.

Each question divides the data into **smaller**, **more homogeneous groups** until we arrive at the final classification. In our case, the questions revolve around:

Where? (In which **neighborhood** did the crime occur)
When? (On **which day** of the week and at **what time**)
What? (**What type** of crime is expected)

How does a tree build its decisions?

The algorithm follows these steps:

It starts at the root: It asks the most effective question that divides the data into two as distinct groups as possible. For example: **"Did the crime occur after 8 p.m.?" Yes or No**

It continues branching: At each subsequent level, it chooses the best question to separate the data again. It might ask: **"Is the area the northern region?" then "Is today a Friday?"**

It stops when: either **all the elements in the group are of the same type**, it reaches a certain depth, or the partitioning becomes ineffective.

Why are decision trees suitable for crime analysis?

They handle categorical data: such as **neighborhood names** and **days of the week** easily.

They produce understandable rules: police can easily understand their logic.

They don't require complex processing: they handle missing values and nonlinear relationships.

They reveal important relationships: such as that **car thefts increase on Fridays** in a **business district**.

Evaluation Metrics

For the **Decision Tree**

a set of metrics were used to evaluate the model's performance, such as **accuracy**, weighted **precision**, weighted **recall**, and weighted **F1 score**. These metrics help assess the model's ability to correctly predict the appropriate crime type in the test data. The results of these metrics reflect the tree's effectiveness in handling data that includes various characteristics such as geographic location and time.

Accuracy: 22.2% - indicates the percentage of correct classifications
Precision: 15.8% - reflects the accuracy of positive predictions
Recall: 22.2% - demonstrates the model's ability to detect actual cases
F1 score: 14.4% - average of accuracy and recall

For **hierarchical clustering**

the **Silhouette**, **Calinski-Harabasz**, and **Davies-Bouldin** metrics were used to evaluate the quality of the clusters

Silhouette coefficient: 0.146 (training) and 0.111 (test)
-A **positive** silhouette indicates **some structure** in the data.

Kalinsky-Harabaz index: 727 (training) and 697 (test)
The **difference between training and testing** is acceptable and **does not indicate overfitting.**

Davies-Bolden index: 1.789 (training) and 2.001 (test)
A moderate Davies-Bolden index **confirms the presence of distinct clusters**. There are some **rare crimes**.

For **K-medoids**

,the model was evaluated using the same metrics: **Silhouette**, **Calinski-Harabasz** and **Davies-Bouldin**

High Silhouette Coefficient: 0.254
Indicates that clustering **succeeded in forming cohesive and well-defined clusters**

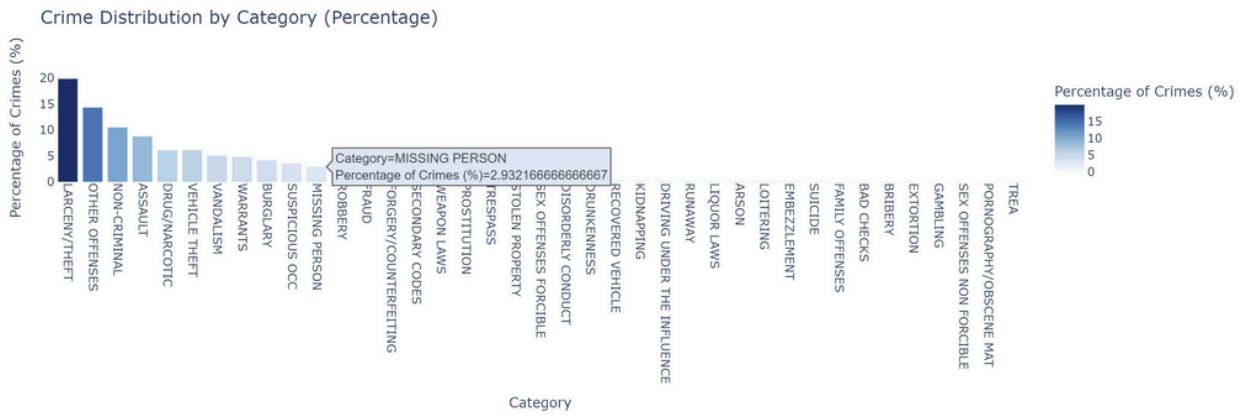
High Kalinsky-Harabaz Index: 1987.1
This very high index indicates that clustering produced dense and widely separated clusters. This is considered excellent and demonstrates that the model **succeeded in clearly and effectively distinguishing between different crime patterns**

Low Davies-Bolden Index: 1.179
This is considered very good and confirms that **the model produced a clear and distinct clustering of crime patterns**

Visualization

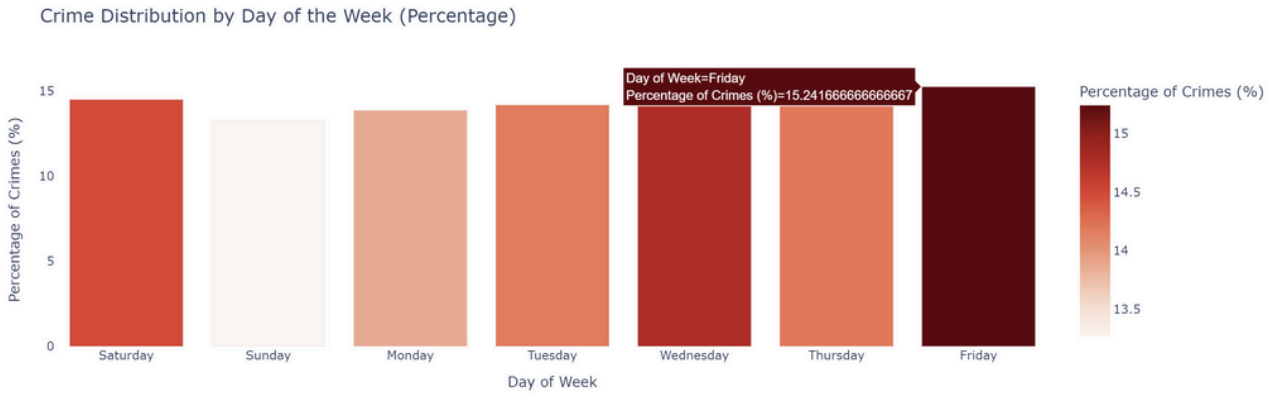
Crime Distribution by Category (Percentage)

Purpose: Understand the frequency of different crime types.
Insights:
The most frequent crime category is LARCENY/THEFT, dominating the distribution · WARRANTS and OTHER OFFENSES follow as significant categories ·
Less frequent crimes (e.g., "EMBEZZLEMENT") appear as smaller bars, indicating · rare occurrences
Implication: Law enforcement could prioritize theft prevention and warrant enforcement due to their prevalence



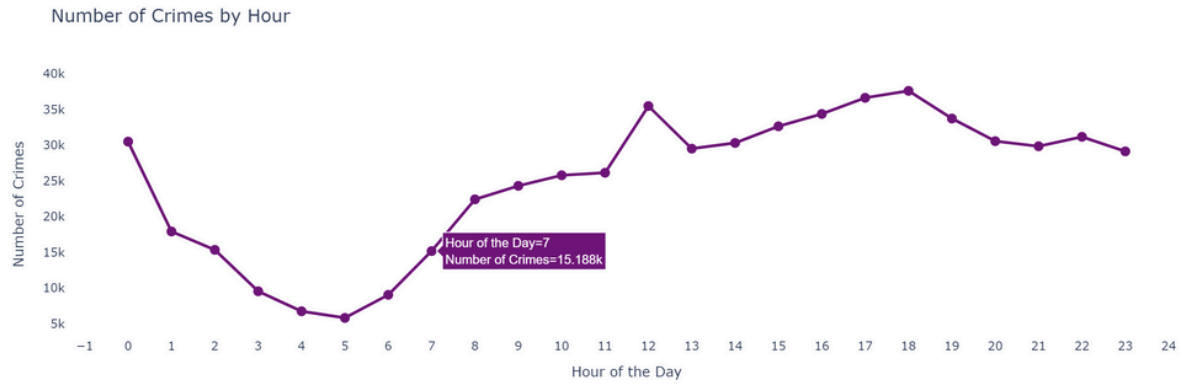
Crime Distribution by Day of the Week (Percentage)

Purpose: Analyze weekly crime patterns.
Insights:
Crimes are evenly distributed across days, with slight peaks on Friday and Saturday (each 15%~)
Sunday shows the lowest crime rate (~13.5%), possibly due to reduced activity
Weekdays (Monday–Thursday) hover around 14%, indicating consistent crime rates
Implication: Weekend policing might require additional resources due to higher activity



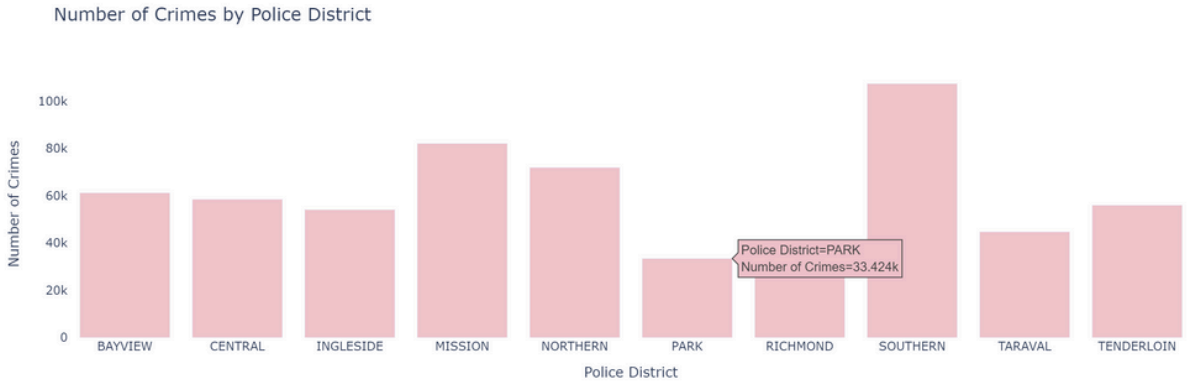
Number of Crimes by Hour

Purpose: Examine hourly crime trends.
Insights:
Peak Hours: Crimes spike during late afternoon/evening (12 PM–8 PM), with the · highest count at 6 PM
Lowest Activity: Early morning hours (2 AM–5 AM) show the fewest crimes ·
Pattern: Crimes rise steadily from 6 AM, peak at 6 PM, and decline afterward
Implication: Patrols could be intensified during peak hours, especially around commute times



Number of Crimes by Police District

Purpose: Identify high-crime districts.
Insights:
SOUTHERN and MISSION districts have the highest crime counts, exceeding 120,000 · incidents
NORTHERN and CENTRAL follow as moderate-risk areas ·
TARAVAL and PARK show the lowest crime rates ·
Implication: Resource allocation (e.g., patrols, community programs) should prioritize high-crime districts like SOUTHERN and MISSION



Crime Heatmap (Hexbin)

Purpose: Visualize geographical crime density.
Insights:
Hotspots: The densest areas (red/yellow) are near downtown (e.g., Tenderloin Mission District), aligning with the district analysis
Spread: Crimes radiate outward from the city center, with lower density in · peripheral neighborhoods
Waterfront: Low activity near the coastline (e.g., Presidio) ·
Implication: Targeted interventions (e.g., lighting, surveillance) could focus on central hotspots