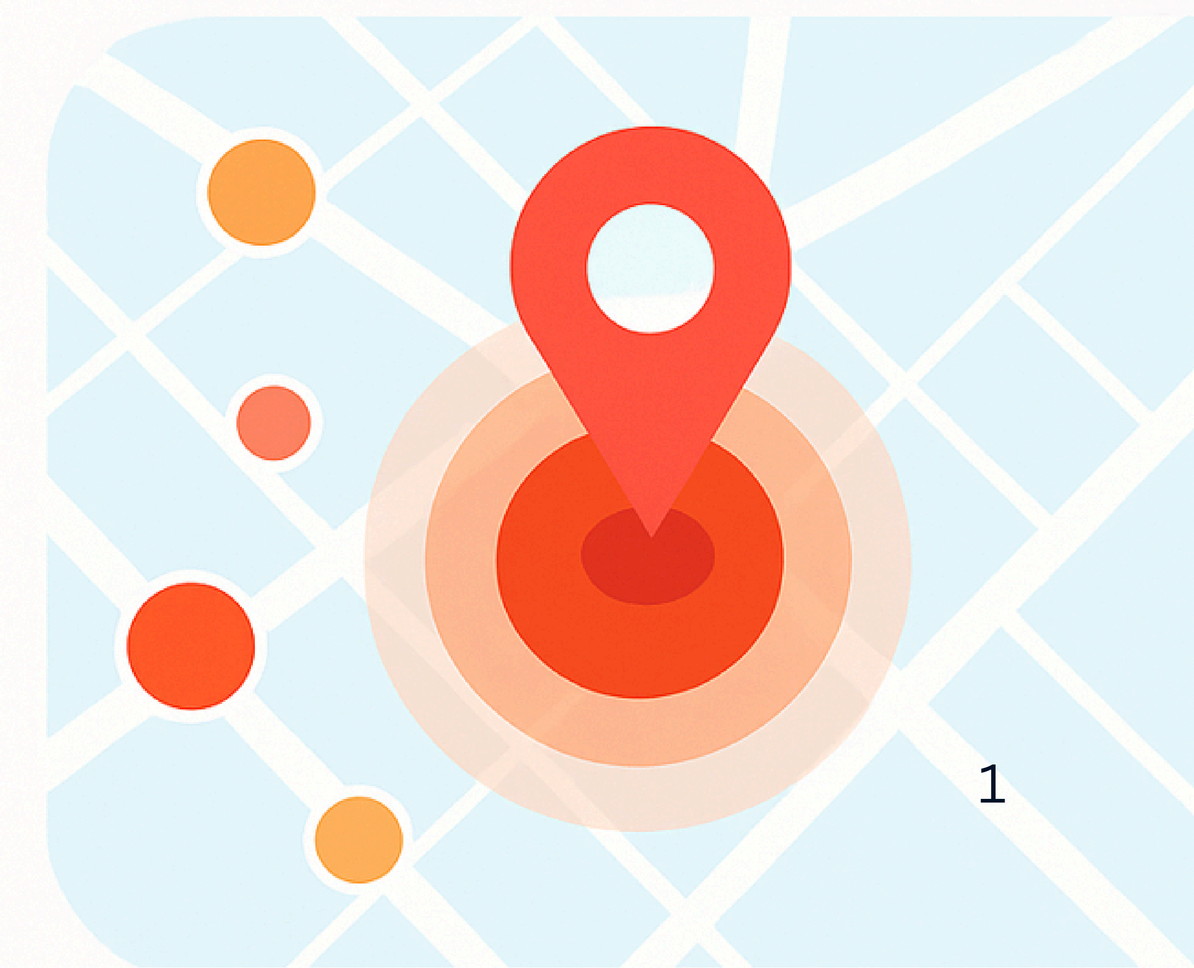
DBSCAN vs K-Means:

CRIME HOTSPOT DETECTION



Guide :

Asst. Prof. RESHMA SUDHAKARAN

Submitted By

RAJEEV R

ROLL NO. : 70

LNSS22CS072

Seminar Overview

01 **Introduction to Crime Hotspots:** Understanding the problem.

02 **Clustering Fundamentals:** The role of unsupervised learning.

03 **K-Means Clustering:** Algorithm, strengths, and limitations.

04 **DBSCAN Clustering:** Algorithm, strengths, and limitations.

05 **Comparative Analysis:** K-Means vs. DBSCAN for crime data.

06 **Conclusion & Q&A:** Choosing the right tool and future trends

What are Crime Hotspots?

Geographic areas with a statistically higher concentration of criminal incidents than surrounding areas.

Characteristics:

- High frequency of specific crime types.
- Often localized to specific streets, blocks, or intersections.
- Can be dynamic, shifting over time.

Importance: Identifying these areas is crucial for effective policing and resource allocation.

WHY DETECT CRIME HOTSPOTS?

- **Resource Optimization:** Directing limited police resources to areas where they are most needed.
- **Proactive Policing:** Shifting from reactive responses to proactive crime prevention strategies.
- **Targeted Interventions:** Implementing specific community programs or interventions in high-risk zones.
- **Understanding Crime Patterns:** Gaining insights into the spatial distribution and underlying causes of crime.
- **Public Safety:** Enhancing overall safety and security for citizens.

Traditional Methods vs. Data-Driven Approaches

Traditional Methods:

- Manual mapping (pin maps).
- Expert knowledge and anecdotal evidence.
- Simple aggregation (e.g., counting crimes per district).
- **Limitations**: Subjective, labor-intensive, may miss subtle patterns.

Data-Driven Approaches:

- Leveraging large datasets of crime incidents.
- Employing statistical and machine learning algorithms.
- **Advantages:** Objective, efficient, identifies complex patterns, predictive capabilities.

Introduction to Clustering

- **What is Clustering?** An unsupervised machine learning technique that groups similar data points together.
- **Goal:** To partition a dataset into subsets (clusters) such that data points within the same cluster are more similar to each other than to those in other clusters.
- **No Labels**: Unlike supervised learning, clustering does not require pre-labeled data.
- **Applications:** Customer segmentation, anomaly detection, document analysis, and crime hotspot detection.

Clustering in Crime Analysis

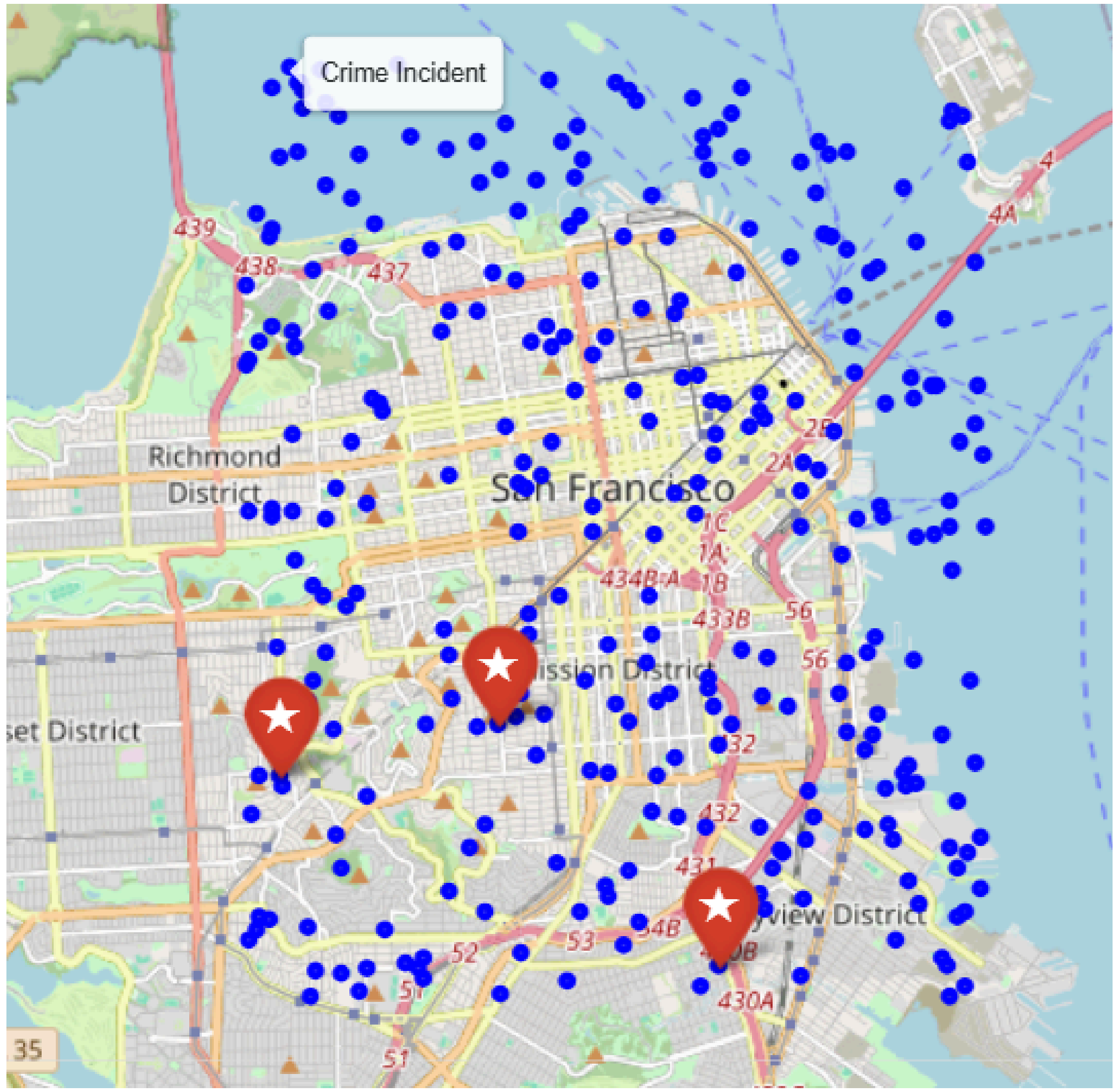
- **Application:** Grouping crime incidents based on their geographical coordinates (latitude and longitude).
- **Output:** Each cluster represents a potential crime hotspot.
- **Benefits:**
  - Automated identification of high-density crime areas.
  - Reveals spatial patterns that might not be obvious manually.
  - Provides a quantitative basis for resource deployment.
- **Key Challenge:** Choosing the right clustering algorithm and parameters.

Introduction to K-Means Clustering

- **Algorithm Type**: Centroid-based clustering algorithm.
- **Core Idea:** Partitions 'n' observations into 'k' clusters, where each observation belongs to the cluster with the nearest mean (centroid).
- **"K" Value**: The number of clusters (k) must be specified beforehand.
- **Iterative Process:** The algorithm iteratively assigns data points to clusters and updates cluster centroids.

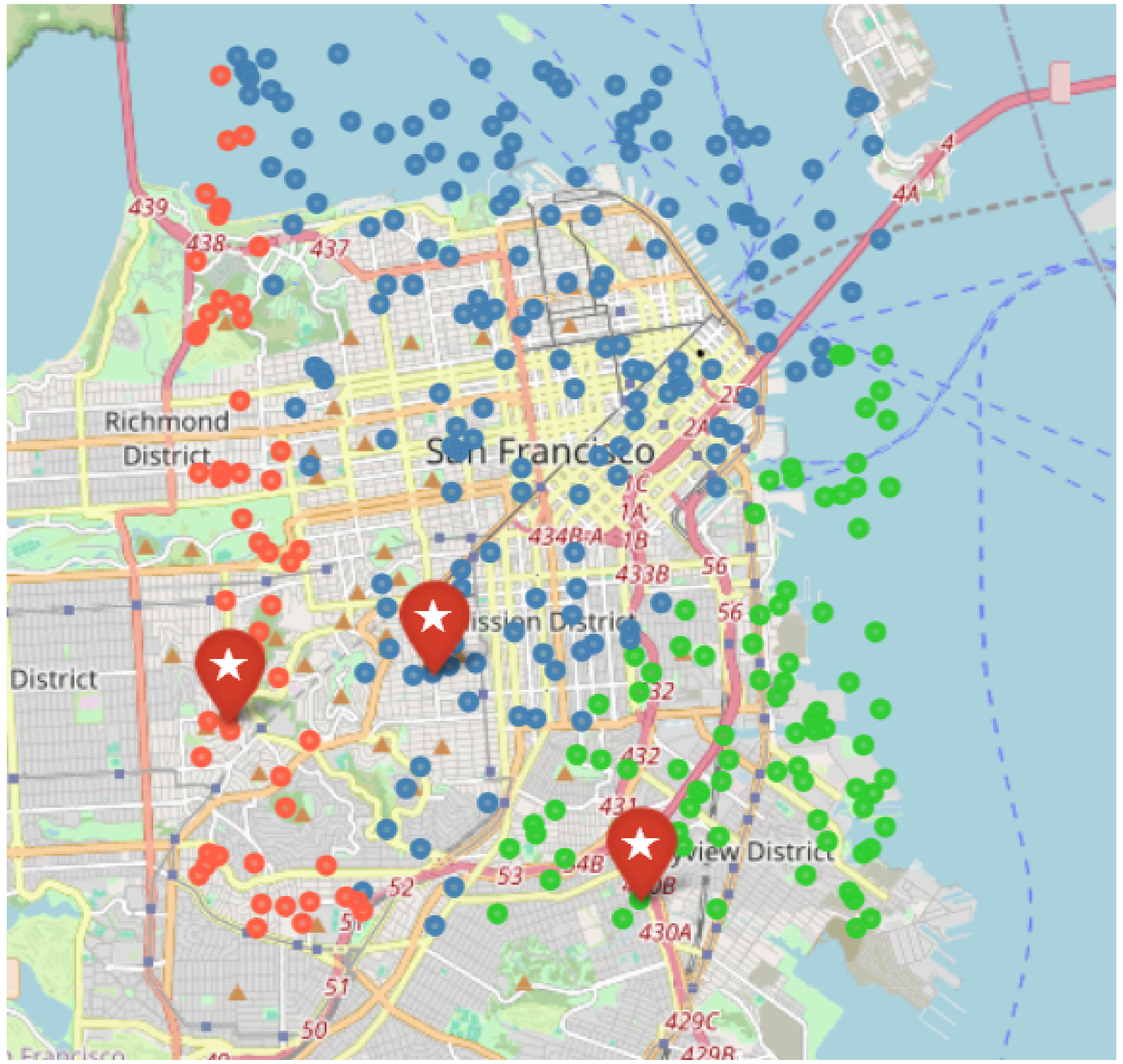
How K-Means Works: The Algorithm (Step 1)

- **Step 1: Initialization**
    - Randomly select k data points from the dataset to serve as initial cluster centroids.



How K-Means Works: The Algorithm (Step 2)

- **Step 2: Assignment Step**
  - Each data point (crime incident) is assigned to the cluster whose centroid is closest to it.
  - Distance is typically measured using Euclidean distance.
  - Formula: $d((x_1,y_1),(x_2 ,y_2))=(x_2-x_1)^2+(y_2-y_1)^2$



How K-Means Works: The Algorithm (Step 3)

- **Step 3: Update Step**
  - Recalculate the centroids for each cluster by taking the mean of all data points assigned to that cluster.
  - **Formula**: New Centroid $(C_x, C_y) = (\frac{1}{n} \sum x_i, \frac{1}{n} \sum y_i)$ where n is the number of points in the cluster.
- **Iteration:** Repeat Steps 2 and 3 until the centroids no longer move significantly or a maximum number of iterations is reached.
- **Convergence:** The algorithm converges when cluster assignments no longer change.

K-Means: Strengths for Crime Hotspot Detection

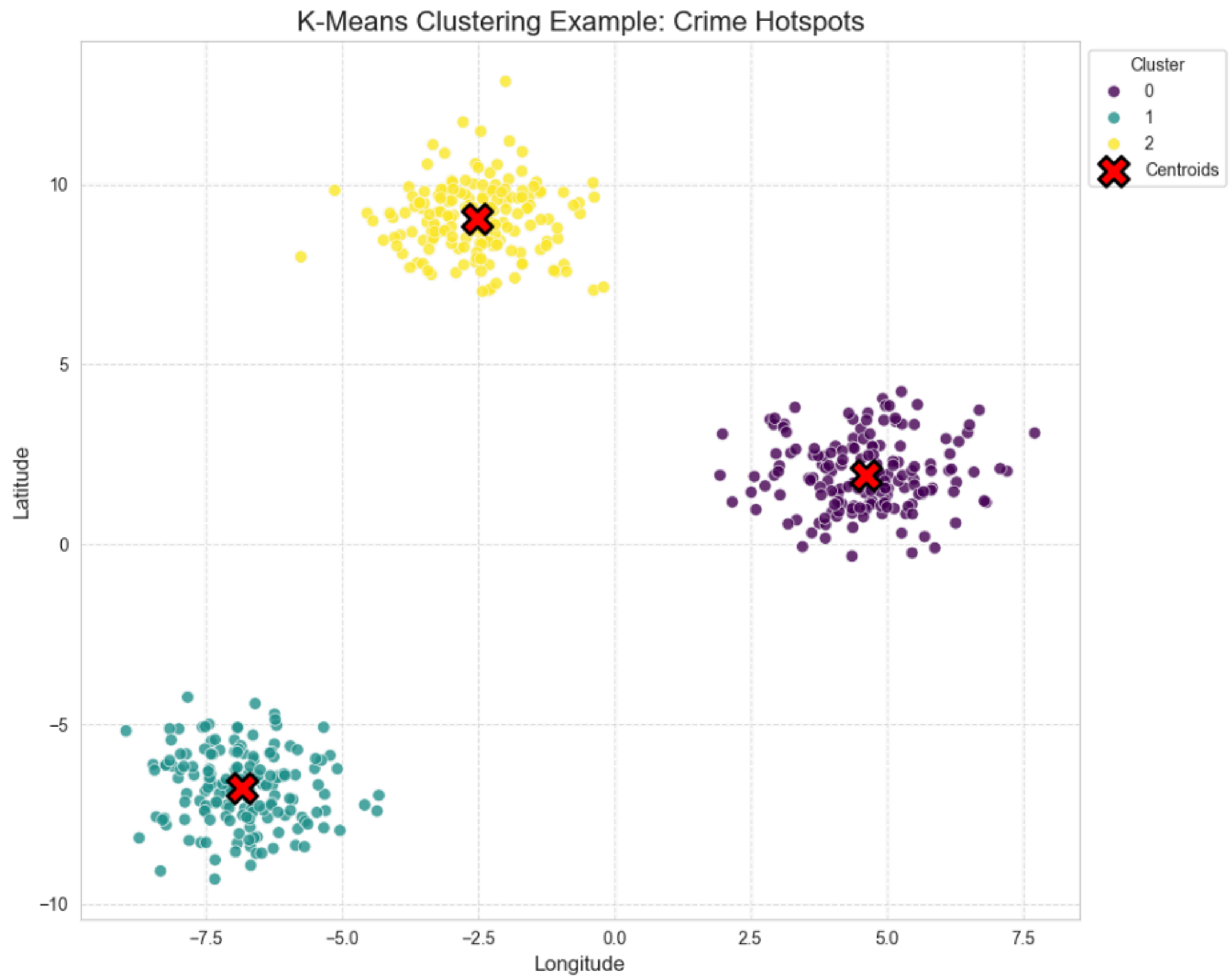
- **Simplicity & Speed:** Relatively simple to understand and computationally efficient, especially for large datasets.
- **Scalability:** Can handle a large number of crime incidents.
- **Guaranteed Convergence:** The algorithm is guaranteed to converge to a solution.
- **Interpretability:** Centroids provide a clear "center" for each hotspot.
- **Widely Used:** A well-established and understood algorithm.

K-Means: Limitations for Crime Hotspot Detection

- **Requires k:** Must specify the number of clusters (k) beforehand, which is often unknown for crime hotspots.
- **Spherical Clusters:** Assumes clusters are spherical and equally sized, which is rarely true for real-world crime patterns.
- **Sensitivity to Outliers:** Outliers (isolated crime incidents) can significantly affect centroid positions.
- **Handles Noise Poorly:** Treats all points as belonging to a cluster, even isolated noise.
- **Initial Centroid Sensitivity:** Results can vary based on the initial random placement of centroids.

K-Means: Summary

- **Observation:** K-Means identifies distinct, compact groups of crime incidents.
- **Challenge:** What if hotspots are irregularly shaped or vary in density?

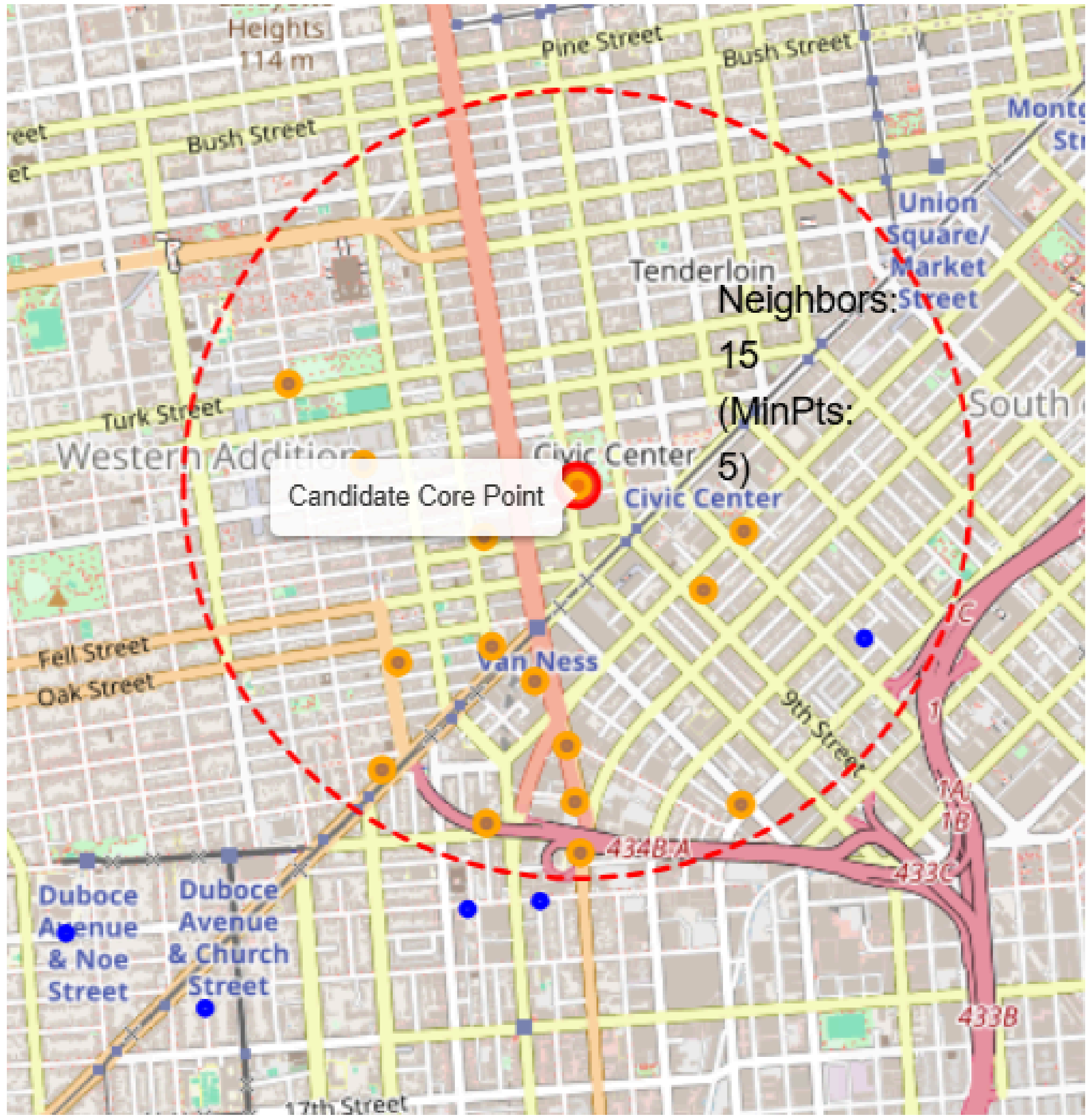


Introduction to DBSCAN

- **Algorithm Type:** Density-based spatial clustering of applications with noise.
- **Core Idea:** Groups together points that are closely packed together, marking as outliers points that lie alone in low-density regions.
- **No Pre-defined k:** Does not require specifying the number of clusters beforehand.
- **Handles Noise:** Explicitly identifies and labels noise points.
- **Discovers Arbitrary Shapes:** Can find clusters of arbitrary shapes, unlike K-Means.

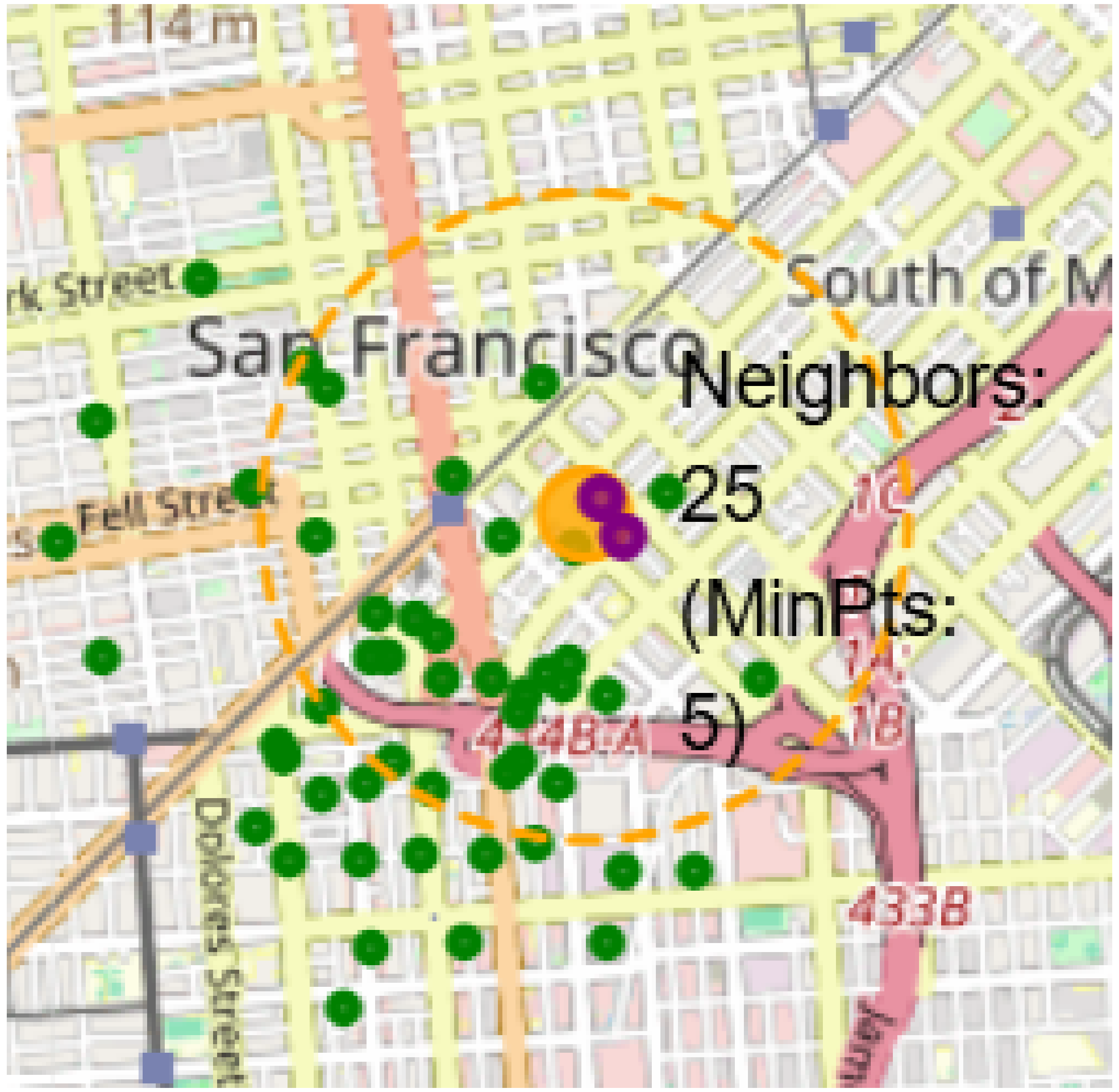
How DBSCAN Works: Core point

- **Definition:** A point is a core point if there are at least MinPts (minimum number of points) within a distance of ϵ (epsilon) from it.
- **Role:** Core points form the "dense" parts of clusters.



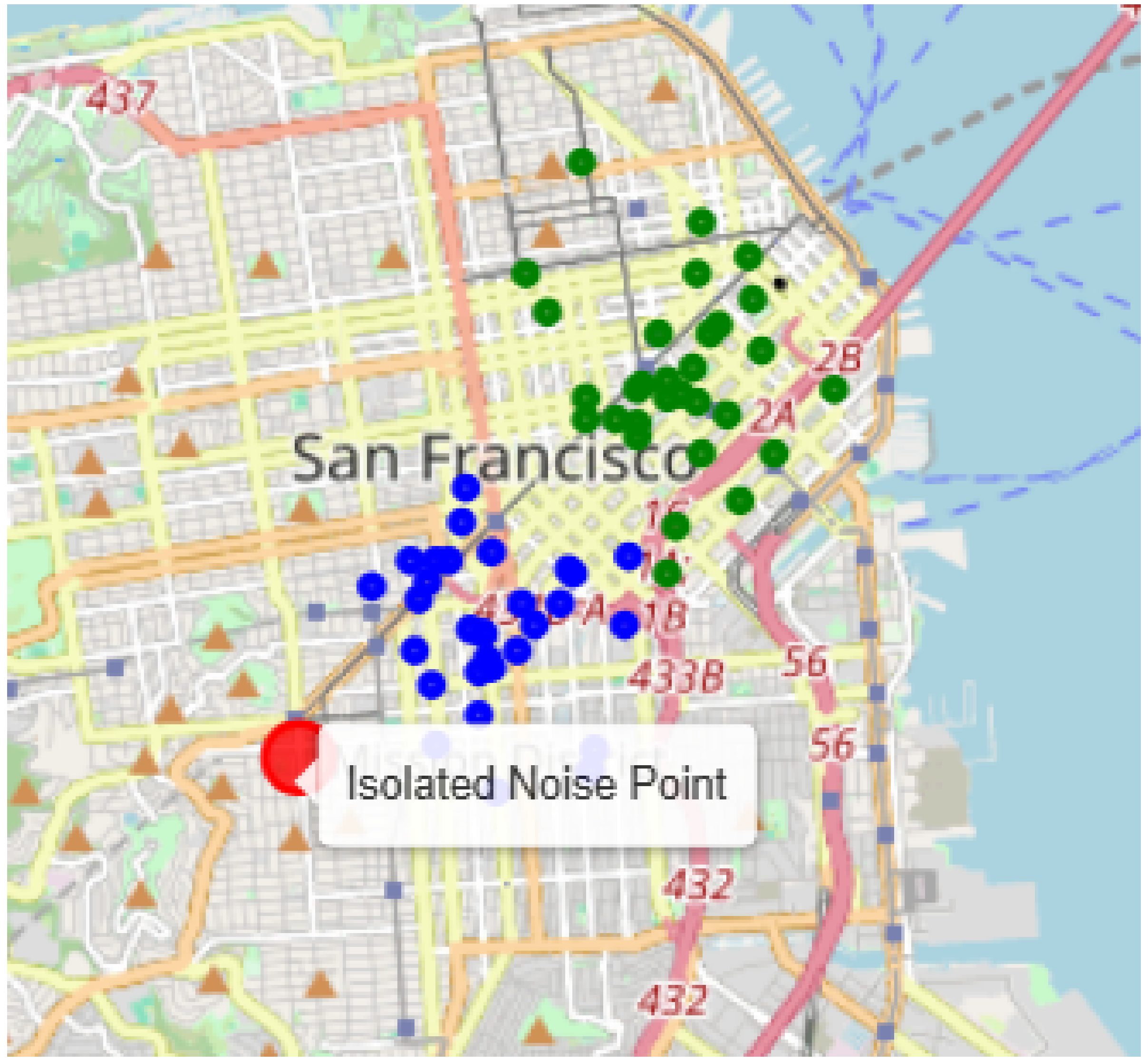
How DBSCAN Works: Border Points

- **Definition:** A point is a border point if it is within a distance of ϵ from a core point, but it is not a core point itself (i.e., it doesn't have MinPts within its own ϵ-neighborhood).
- **Role:** Border points are on the "edge" of a cluster.



How DBSCAN Works: Noise Points

- **Definition:** A point is a noise point (or outlier) if it is neither a core point nor a border point.
- **Role:** These are isolated crime incidents that do not belong to any dense cluster.
- **Benefit:** DBSCAN explicitly identifies these, which can be useful for understanding sporadic crime.



DBSCAN: Key Parameters: Epsilon (ϵ)

- **Definition:** The maximum distance between two samples for one to be considered as in the neighborhood of the other.
- **Impact:** Defines the radius of the neighborhood to search for points.
- **Too Small ϵ:** Many points might be labeled as noise, and clusters might be fragmented.
- **Too Large ϵ:** Different clusters might merge into a single large cluster.
- **Selection:** Often determined by domain knowledge or by analyzing the k-distance graph.

DBSCAN: Key Parameters: MinPts

- **Definition:** The minimum number of points required to form a dense region (i.e., the minimum number of points in an ϵ-neighborhood for a point to be considered a core point).
- **Impact:** Influences the density required to form a cluster.
- **Too Small MinPts:** Can lead to noisy clusters, as even sparse regions might be considered dense.
- **Too Large MinPts:** May cause sparse clusters to be labeled as noise.
- **General Rule:** A common heuristic is MinPts≥D+1, where D is the dimensionality of the data .

DBSCAN: Strengths for Crime Hotspot Detection

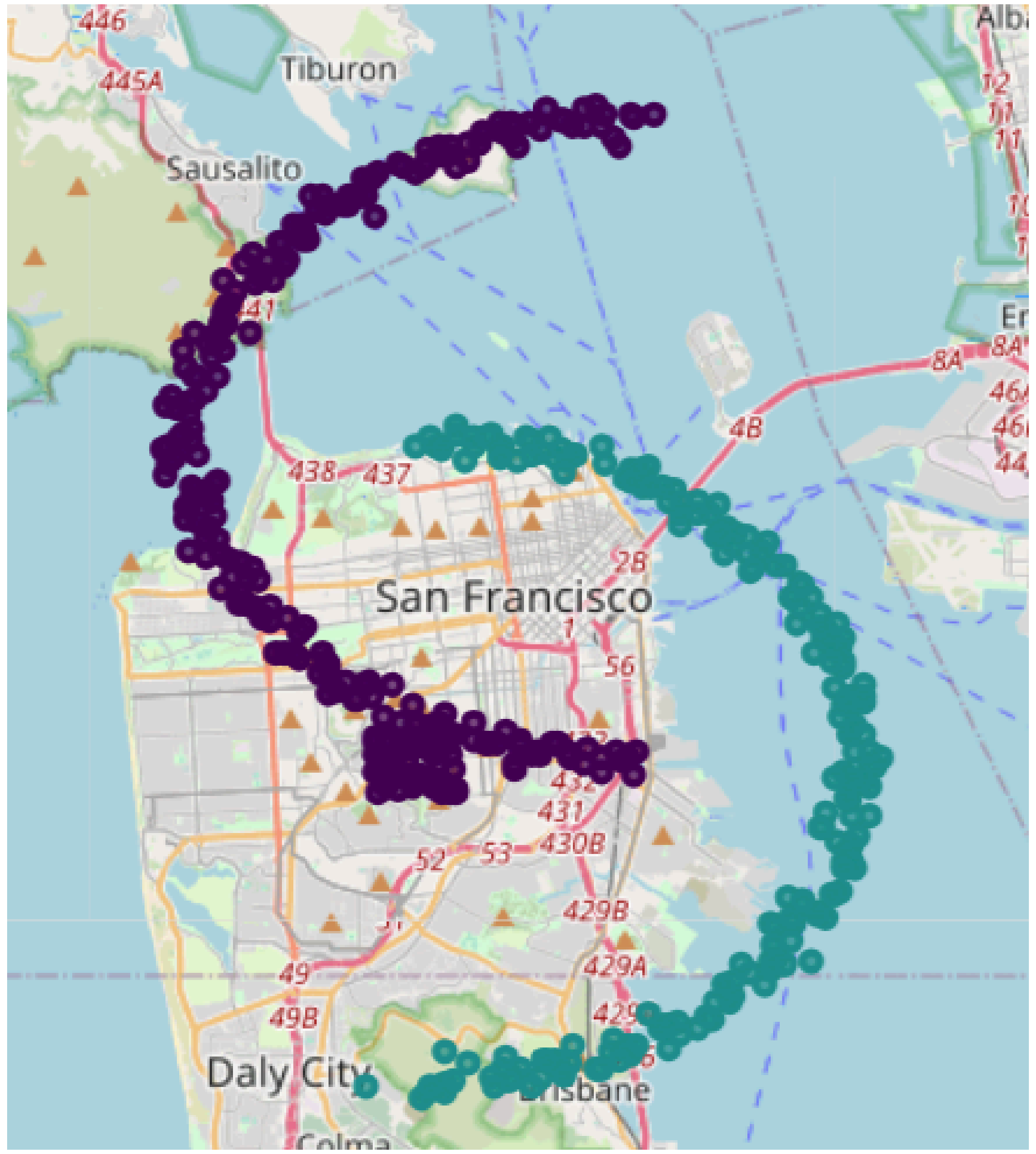
- **Arbitrary Cluster Shapes:** Can discover clusters of complex, non-spherical shapes, which is common for crime hotspots.
- **Handles Noise Naturally:** Explicitly identifies outliers/noise points, which is valuable for real-world crime data.
- **No Pre-defined k:** Does not require the user to specify the number of clusters beforehand.
- **Robust to Outliers:** Less sensitive to individual outliers compared to K-Means.
- **Density-Based:** Aligns well with the concept of "hotspots" as areas of high crime density.

DBSCAN: Limitations for Crime Hotspot Detection

- **Parameter Sensitivity:** Highly sensitive to the choice of ϵ and MinPts. Incorrect parameters can lead to poor results.
- **Varying Densities:** Struggles with clusters of widely varying densities. A single pair of (ϵ,MinPts) values might not work for all clusters.
- **Border Points:** Border points can be part of multiple clusters, leading to ambiguity.
- **High Dimensionality:** Performance can degrade in very high-dimensional data (though less of an issue for 2D spatial data).
- **Computational Cost:** Can be slower than K-Means for very large datasets, especially with inefficient spatial indexing.

DBSCAN: Summary

- **Observation**: DBSCAN successfully identifies clusters that are not necessarily circular and effectively separates noise.
- **Benefit**: More accurately reflects the organic shapes of crime hotspots.



Parameter Sensitivity: K-Means vs. DBSCAN

- **K-Means:**
  - **k:** The most critical parameter. Incorrect k leads to over- or under-segmentation.
  - **Initialization:** Can get stuck in local optima.
- **DBSCAN:**
  - **ϵ & MinPts:** Highly impactful. Small changes can drastically alter results.
  - **Challenge:** Finding optimal values often requires experimentation and domain knowledge.

Handling Noise: K-Means vs. DBSCAN

- **K-Means:**
  - Every data point is forced into a cluster.
  - Outliers can distort cluster centroids, leading to less accurate hotspot definitions.
  - Requires pre-processing to remove noise if desired.
- **DBSCAN:**
  - A significant advantage: inherently identifies noise points.
  - This is highly beneficial for crime data, where isolated incidents (noise) should not be considered part of a dense hotspot.
  - Provides a cleaner representation of true dense areas.

Cluster Shape: K-Means vs. DBSCAN

- **K-Means:**
  - Designed for isotropic (spherically shaped) clusters.
  - Struggles with elongated, crescent-shaped, or irregularly shaped hotspots.
  - May split a single, large, irregular hotspot into multiple smaller, spherical ones.
- **DBSCAN:**
  - Excels at discovering clusters of arbitrary shapes.
  - Can accurately delineate hotspots that follow street networks or geographical features.
  - More suitable for real-world crime patterns that are rarely perfectly circular.

Scalability: K-Means vs. DBSCAN

- **K-Means**:
  - Generally faster for very large datasets, especially with optimizations (e.g., mini-batch K-Means).
  - Time complexity: O(nkdI), where n is data points, k is clusters, d is dimensions, I is iterations.
- **DBSCAN:**
  - Can be slower for very large datasets without spatial indexing.
  - Time complexity: O(nlogn) or O(n^2) depending on implementation (e.g., using k-d trees vs. brute-force distance calculations).

Use Cases in Crime Hotspot Detection

- **When to Use K-Means:**
  - When the number of hotspots (k) is known or can be reasonably estimated.
  - When hotspots are expected to be roughly circular and distinct.
  - For quick, preliminary analysis on very large datasets.
- **When to Use DBSCAN:**
  - When the number of hotspots is unknown.
  - When hotspots are expected to have irregular shapes.
  - When identifying and separating noise (isolated incidents) is crucial.
  - For more precise and realistic hotspot delineation

Summary

- Crime hotspot detection is vital for effective policing.
- **K-Means** is simple, fast, and good for spherical clusters, but requires pre-defined k and handles noise poorly.
- **DBSCAN** can find arbitrary shapes and handles noise well, but is sensitive to parameters and can be slower.
- The choice depends on data characteristics, desired output, and problem context.
- Visualization and domain knowledge are crucial for validation.

References

[1] Yiqun Xie, Shashi Shekhar, and Yan Li. 2022. Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey. ACM Comput. Surv. 55, 2, Article 36 (January 2022). https://doi.org/10.1145/3487893

[2] J. Bonam, L. R. Burra, G. S. V. N. S. Susheel, K. Narendra, M. Sandeep and G. Nagamani, "Crime Hotspot Detection using Optimized K-means Clustering and Machine Learning Techniques," 2023. https://doi.org/10.1109/ICESC57686.2023.10193563

Any questions?

Thank you