project

Crime Rate Prediction

Team Members

Muhammad Elsyed Muhammad Ahmed	**2305273**
Yasser Ali Muhammad Muhammad	**2305268**
Muhammad Ashraf Fathy Ahmed	**2305181**
Yakot Shaker Naseem Shaker	**2305557**
Muhammad Ahmed Fayek	**2305298**
Muhammad Hazem Hafez Mustafa	**2305534**

SCAN ME

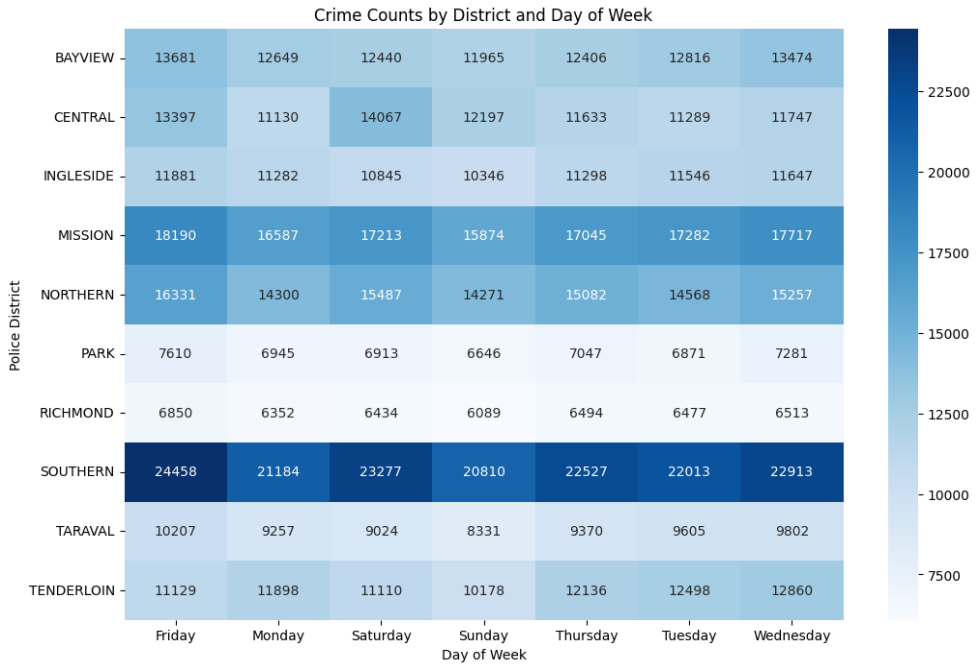
Crime Rate Prediction

Project Overview:

The crime rate is increasing now-a-days in many countries. In today's world with such a higher crime rate and brutal crime happening, there must be some protection against this crime. Here we introduced a system by which the crime rate can be reduced. Crime data must be fed into the system. We introduced data mining algorithms to predict crime. Crime data is analyzed which is stored in the database. Data mining algorithms will extract information and patterns from database. The system will commit group crime. Clustering will be done based on places where crime occurred, gangs who were involved in crime and the timing crime took place. This will help to predict crime which will occur in future. Admin will enter crime details into the system which is required for prediction. Admin can view criminal historical data. Crime incident prediction depends mainly on the historical crime record and various geospatial and demographic information.

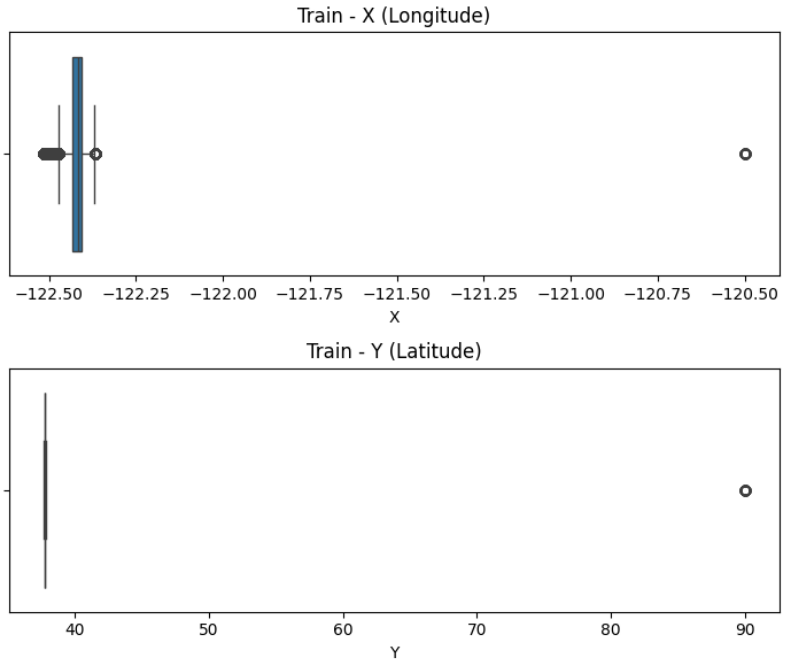
Understanding the Project and Setting Objectives:

The dataset is sourced from the Kaggle SF Crime Rate Prediction competition. It contains 878,049 crime incidents in San Francisco with the following features: Dates: Timestamp of the crime. Category: Type of crime (target variable). Descript: Detailed description of the crime. DayOfWeek: Day of the week. PdDistrict: Police district. Resolution: Outcome of the crime. Address: Location of the crime. X, Y: Longitude and latitude coordinates.



Data Preprocessing: (Cleaning Data, Normalization)

- **Nulls: there is no nulls in the data.**
- **outliers:**



- **duplicated: there are 2323 rows duplicated.**
- **Cleaning Done.**

Data transformation:

- **data scaling using standardization (StandardScaler from scikit-learn)**
- **data encoding using LabelEncoder from scikit-learn**

Clustering methods:

- **K-medoids:**

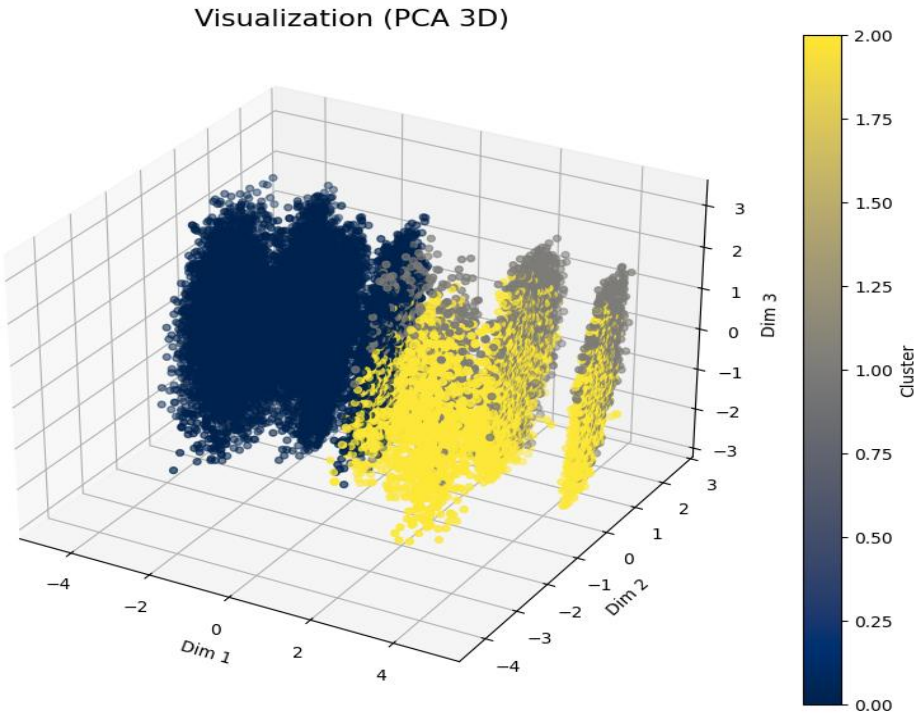
In [6]:
```
kmedoids_range = range(2, 15)
silhouette_scores = []          #For Score Calculations

for k in kmedoids_range:
    kmedoids = KMedoids(n_clusters=k, random_state=0)
    kmedoids.fit(X)
    score = silhouette_score(X, kmedoids.labels_)
    silhouette_scores.append(score)

best_k = kmedoids_range[silhouette_scores.index(max(silhouette_scores))]
```

In [7]:
```
best_k
```

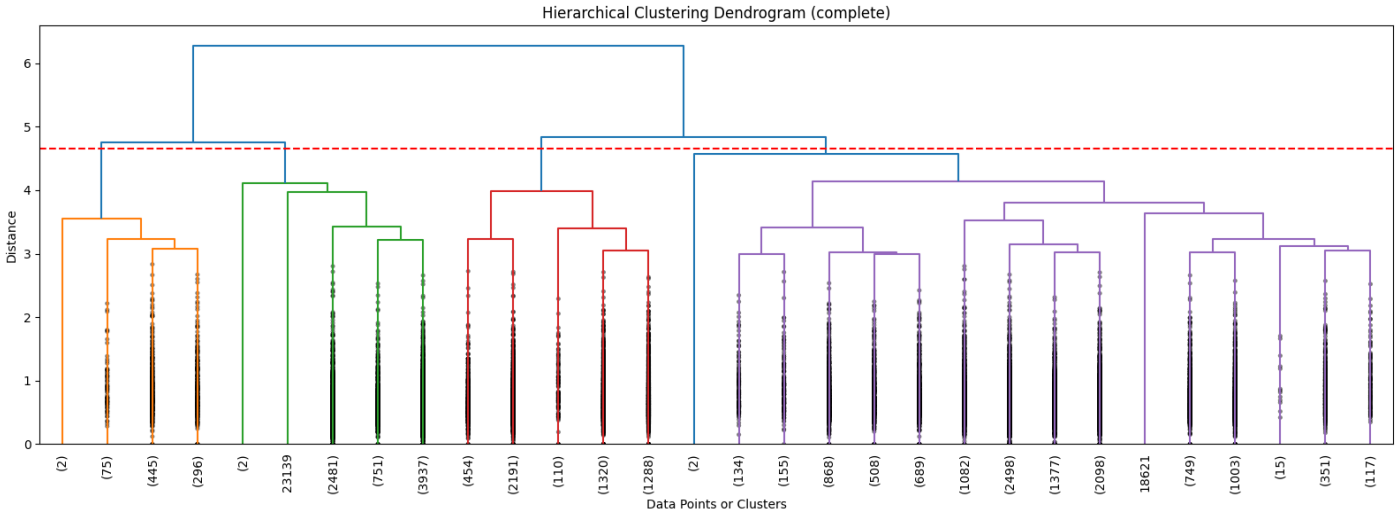
Out[7]: 3

  - **silhouette:** 0.23923224890959097
  - **calinski_harabasz:** 10546.71502162948
  - **davies_bouldin:** 2.2726734680236276

- **Dimensions reduction**:



- **Hierarchical**:



  - **silhouette:** 0.19457079219017495
  - **calinski_harabasz:** 6334.564726658944
  - **davies_bouldin:** 1.3180004238816498

- **Compare Hierarchical and KMedoids:**

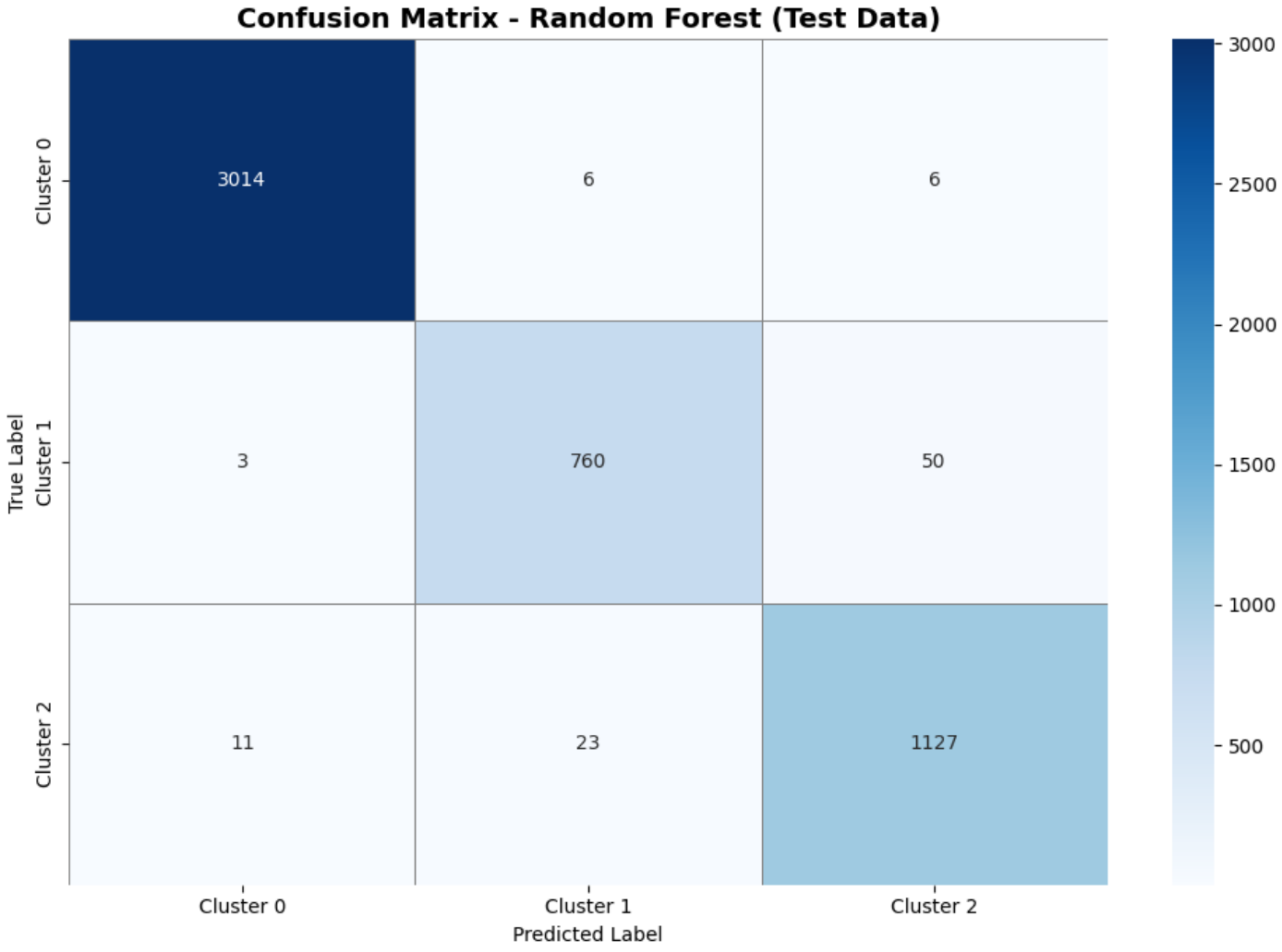
	Metric	KMedoids	Hierarchical
0	Silhouette Score	0.239232	0.194571
1	Calinski-Harabasz Index	10546.715022	6334.564727
2	Davies-Bouldin Index	2.272673	1.318000

Results:

- **Based on the analysis, it can be concluded that KMedoids has a more significant impact on our decision-making process.**

after using classification methods:

- **Decision tree's accuracy = 0.966**
- **Random Forest's accuracy = 0.9802**
    - ❖ ***Random forest has accuracy better than decision tree.***



Source code:
https://github.com/MohamedElsyed2005/crime_rate_prediction_project.git