SF Crime Forecasting

CS 133 - Data Visualization
Member: Tang, Ho Lam
Ruiyuan Li

1.Problem Definition & Inspiration

This project aims to analyze and predict crime trends in San Francisco using machine learning. By forecasting hourly crime counts and exploring spatial patterns, we provide insights to help improve public safety and resource planning.

Project Questions

Key Questions We Explore:

1.What are the spatial crime hotspots in San Francisco over the last 5 years?

2.What are the peak hours for different crime types?

3. Do neighborhoods show consistent patterns for specific crimes?

4. Are there seasonal patterns in crime trends?

5.Can we accurately predict the category of crime based on time, day of week, and police district?

2.Project Questions

Project Questions

Key Questions We Explore:

1.What are the spatial crime hotspots in San Francisco over the last 5 years?

2.What are the peak hours for different crime types?

3. Do neighborhoods show consistent patterns for specific crimes?

4. Are there seasonal patterns in crime trends?

5.Can we accurately predict the category of crime based on time, day of week, and police district?

3.Dataset Overview

Source: San Francisco Police Department Incident Reports (2018–Present)

Columns used:9 key column

incident_category, incident_datetime, analysis_neighborhood, police_district, incident_day_of_week, incident_time, incident_description, longitude, latitude

Total rows: 952,286

```
crime_data_path = "/content/drive/Shareddrives/SF_Crime_Forecasting/Police_Department_Incident_Reports__2018_to_Present_20
crime_df = pd.read_csv(crime_data_path)
crime_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 952287 entries, 0 to 952286
Data columns (total 35 columns):
 #   Column                                              Non-Null Count   Dtype
---  ------                                              --------------   -----
 0   Incident Datetime                                   952287 non-null  object
 1   Incident Date                                       952287 non-null  object
 2   Incident Time                                       952287 non-null  object
 3   Incident Year                                       952287 non-null  int64
 4   Incident Day of Week                                952287 non-null  object
 5   Report Datetime                                     952287 non-null  object
 6   Row ID                                              952287 non-null  int64
 7   Incident ID                                         952287 non-null  int64
 8   Incident Number                                     952287 non-null  int64
 9   CAD Number                                          741646 non-null  float64
 10  Report Type Code                                    952287 non-null  object
 11  Report Type Description                             952287 non-null  object
 12  Filed Online                                        185557 non-null  object
 13  Incident Code                                       952287 non-null  int64
 14  Incident Category                                   951133 non-null  object
 15  Incident Subcategory                                951133 non-null  object
 16  Incident Description                                952287 non-null  object
 17  Resolution                                          952287 non-null  object
 18  Intersection                                        900362 non-null  object
 19  CNN                                                 900362 non-null  float64
 20  Police District                                     952287 non-null  object
 21  Analysis Neighborhood                               900119 non-null  object
 22  Supervisor District                                 899893 non-null  float64
 23  Supervisor District 2012                            900228 non-null  float64
 24  Latitude                                            900362 non-null  float64
 25  Longitude                                           900362 non-null  float64
 26  Point                                               900362 non-null  object
 27  Neighborhoods                                       888210 non-null  float64
 28  ESNCAG - Boundary File                              10984 non-null   float64
 29  Central Market/Tenderloin Boundary Polygon - Updated 127671 non-null  float64
 30  Civic Center Harm Reduction Project Boundary        124884 non-null  float64
 31  HSOC Zones as of 2018-06-05                         200793 non-null  float64
 32  Invest In Neighborhoods (IIN) Areas                 0 non-null       float64
 33  Current Supervisor Districts                        900228 non-null  float64
 34  Current Police Districts                            899378 non-null  float64
dtypes: float64(14), int64(5), object(16)
memory usage: 254.3+ MB
```

4.Data Cleaning & Preparation

1.Handling missing values

2.Simplifying categories

3.Feature creation

4.Before/after table or bullet list

	Incident Datetime	Incident Date	Incident Time	Incident Year	Incident Day of Week	Report Datetime	Row ID	Incident ID	Incident Number	CAD Number	...	Longitude	Point	Neighborhoods	ESNCAG - Boundary File	Central Market/Tenderloin Boundary Polygon - Updated	Civic Center Harm Reduction Project Boundary	HSOC Zones as of 2018-06-05	Invest In Neighborhoods (IIN) Areas	Current Supervisor Districts	Current Police Districts
0	2023/03/01 05:02:00 AM	2023/03/01	05:02	2023	Wednesday	2023/03/11 03:40:00 PM	125379506374	1253795	236046151	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2023/03/14 06:44:00 PM	2023/03/14	18:44	2023	Tuesday	2023/03/14 06:45:00 PM	125402407041	1254024	230176728	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2023/02/15 03:00:00 AM	2023/02/15	03:00	2023	Wednesday	2023/03/11 04:55:00 PM	125378606372	1253786	236046123	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2023/03/11 03:00:00 PM	2023/03/11	15:00	2023	Saturday	2023/03/13 08:29:00 AM	125420606244	1254206	236045937	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2023/03/13 07:30:00 AM	2023/03/13	07:30	2023	Monday	2023/03/14 07:11:00 AM	125412306244	1254123	236047096	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	2023/03/16 09:26:00 AM	2023/03/16	09:26	2023	Thursday	2023/03/16 09:26:00 AM	125467916780	1254679	230185672	230750962.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	2023/03/16 05:30:00 PM	2023/03/16	17:30	2023	Thursday	2023/03/16 06:02:00 PM	125482604134	1254826	230187101	230752550.0	...	-122.401324	POINT (-122.40132418490647 37.76228996810526)	54.0	NaN	NaN	NaN	NaN	NaN	9.0	2.0
7	2023/03/16 01:49:00 PM	2023/03/16	13:49	2023	Thursday	2023/03/16 01:49:00 PM	125473107041	1254731	230178047	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	2023/03/16 10:15:00 PM	2023/03/16	22:15	2023	Thursday	2023/03/17 12:03:00 PM	125561906374	1255619	236049456	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	2023/02/11 02:00:00 PM	2023/02/11	14:00	2023	Saturday	2023/03/18 01:20:00 PM	125564606244	1255646	236050049	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5.Feature Engineering

Parsed **incident_datetime** into: **hour, day, month, year, date_of_week**

Cleaned column names for consistency

Aggregated incidents by hour to create the target: crime_count_hourly

Label-encoded categorical columns:**incident_category, analysis_neighborhood, police_district**

	hour	day	month	year	date_of_week	analysis_neighborhood_label	police_district_label	incident_category_label	crime_count_hourly
0	0	1	1	2018	0	18	3	17	121
1	0	1	1	2018	0	0	0	19	121
2	0	1	1	2018	0	19	8	12	121
3	0	1	1	2018	0	25	2	22	121
4	0	1	1	2018	0	15	4	21	121

6.Data Cleaning Confirmation

Confirm data cleanliness and structure.

No missing (null) values in any of the 12 columns — that means the dataset is fully cleaned.

Verifying that the dataset is clean and ready.

```
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

dtype: int64

7.DATA TRANSFORMATION

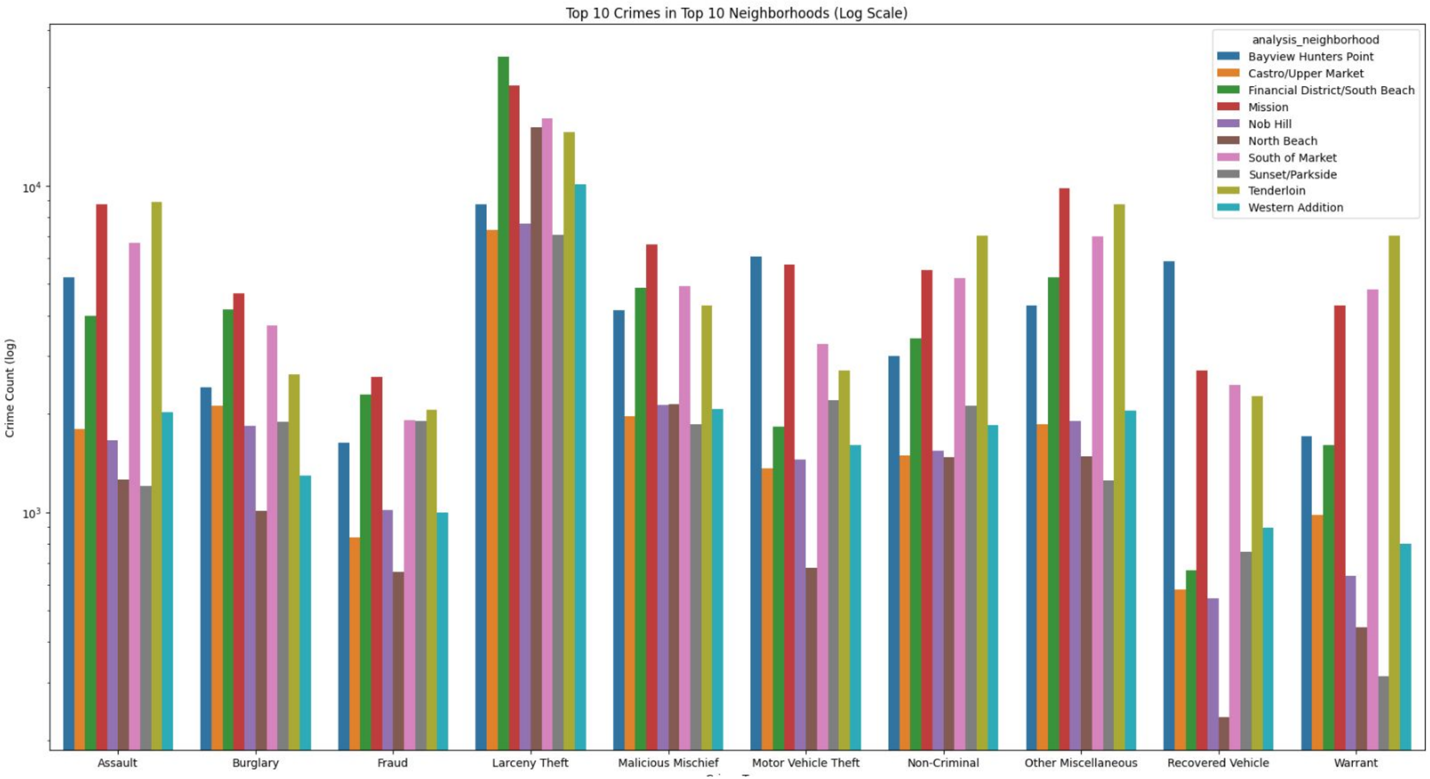
1.This table shows that our dataset is complete.

2.many different crime types

	0
incident_category	44
analysis_neighborhood	42
police_district	11
incident_date	2675
incident_day_of_week	7
longitude	12049
latitude	12538
hour	24
day	31
month	12
year	8
date_of_week	7
crime_count_hourly	81
crime_count_neighborhood	31
crime_count_police_district	33
crime_count_incident_category	39

dtype: int64

8.Data Exploration

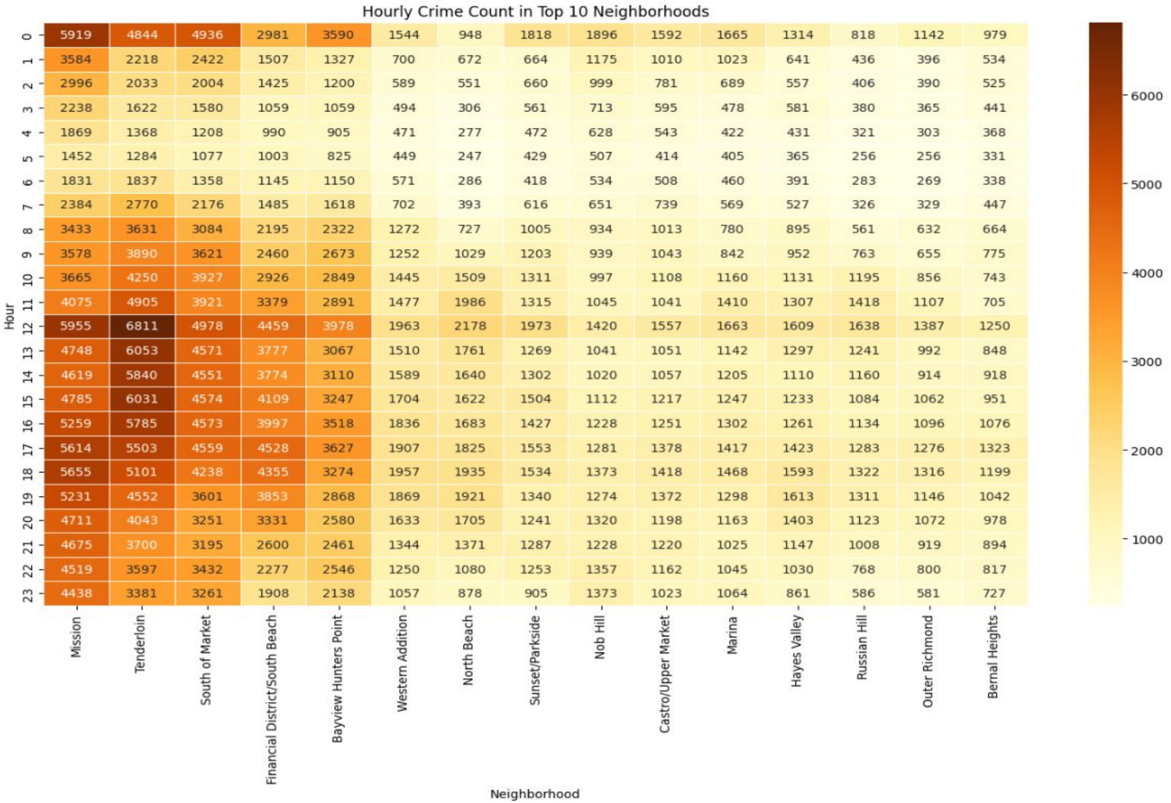


1.Shows top 10 crime types in 10 high-crime neighborhoods.

Larceny Theft is the most common crime.

Mission & Tenderloin have high counts across many crime types.

Data Exploration



2.Hourly Crime Count in Top 10 Neighborhoods

Crime peaks in late afternoon and evening.

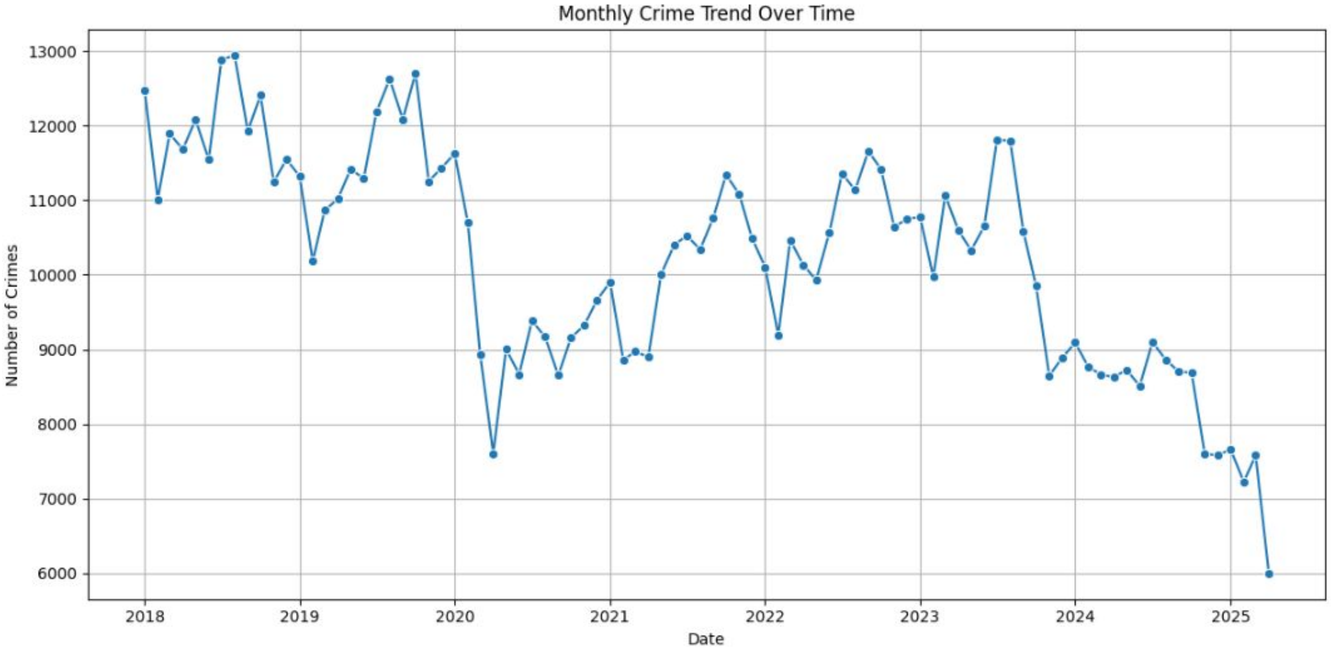
Mission and Tenderloin stay high all day

Data Exploration

3.Line Chart: Monthly Crime Trend Over Time

Crime dropped sharply around 2020.

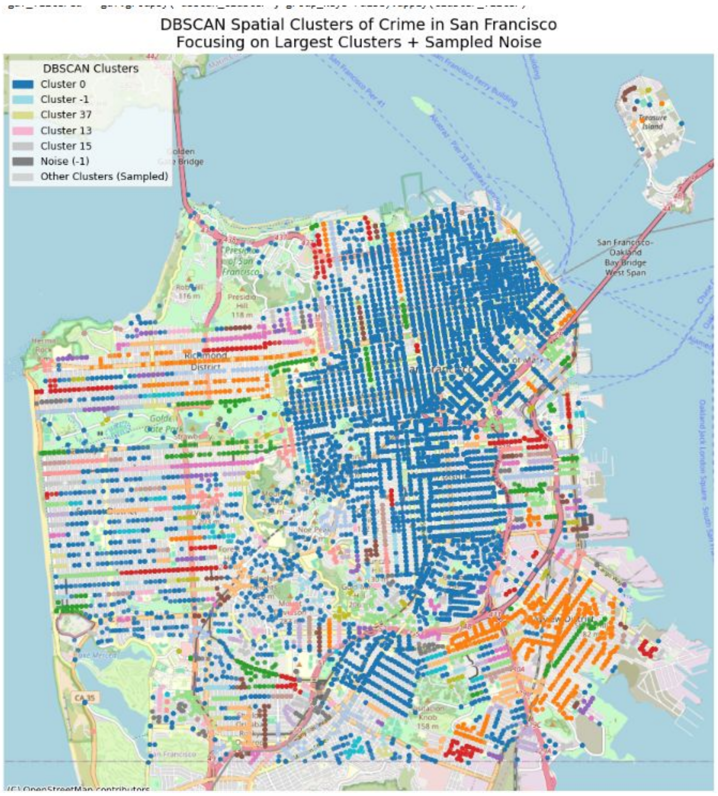
Seasonal ups & downs are visible each year.



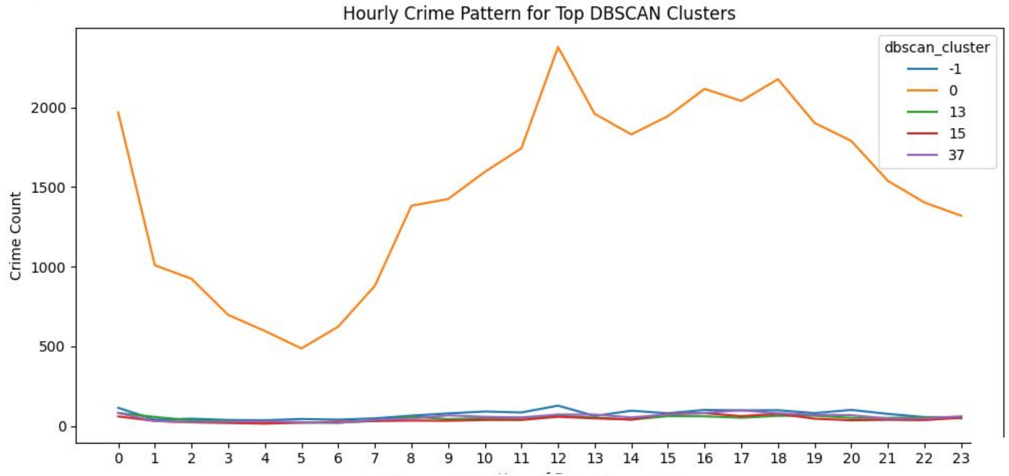
9.CLUSTERING

The main goal for this part is to find out two different pattern

1. Using DBSCAN Clustering to Visualize and understand crime hotspots
2. Find out the crime density or frequency context Include features that indicate how active or dangerous a time/place is

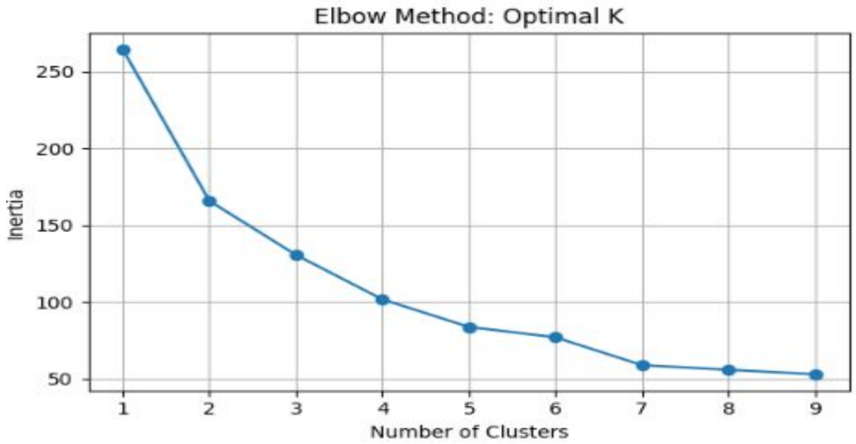


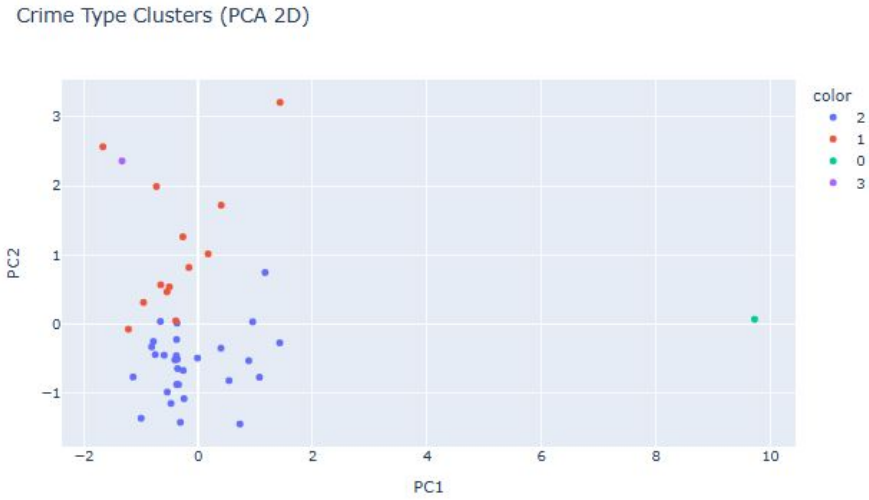
10.CLUSTERING



crime_type_group	incident_category	count
0	Human Trafficking - Involuntary Servitude	3
1	Other Miscellaneous	64765
	Warrant	28770
	Drug Offense	25798
	Weapons Offense	12008
	Traffic Violation Arrest	8599
2	Malicious Mischief	61539
	Assault	59704
	Non-Criminal	54216
	Burglary	52287
	Motor Vehicle Theft	51536
3	Larceny Theft	252675

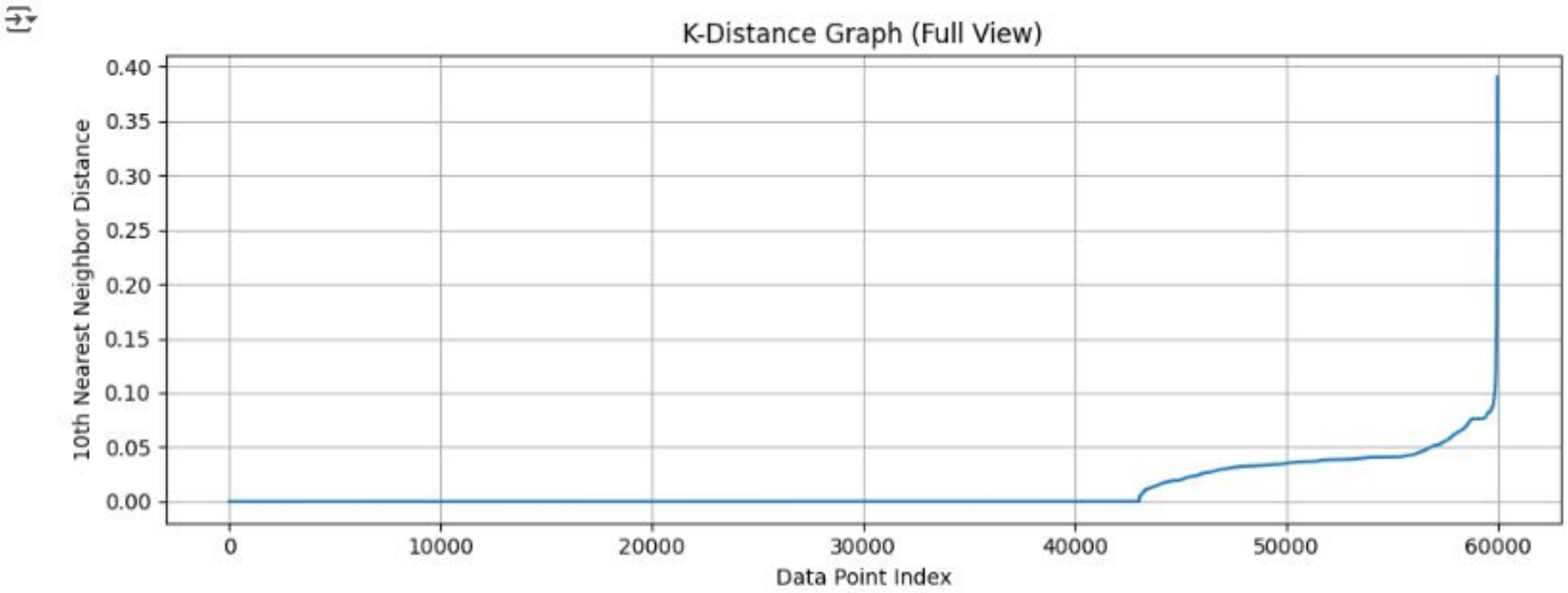
dtype: int64





K-Distance

We used this K-Distance Graph to pick the eps value for DBSCAN. The elbow point (around 0.05) shows the best distance threshold for detecting crime clusters.

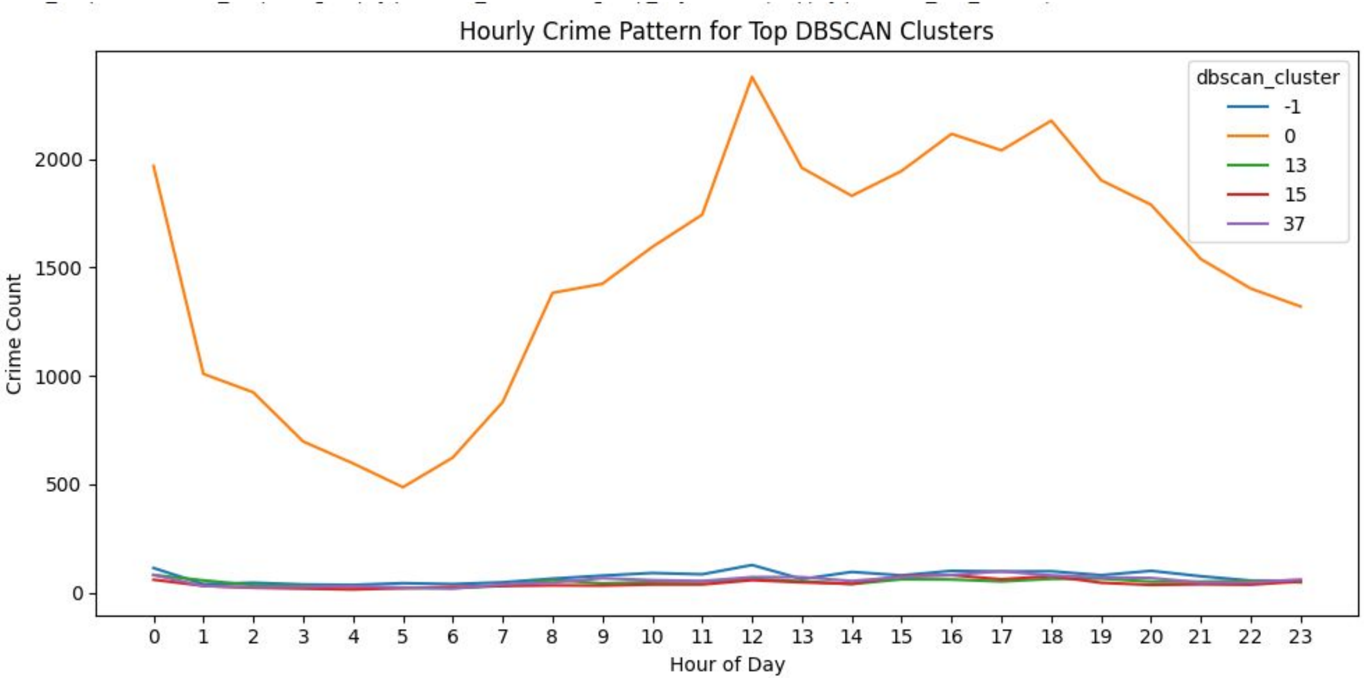


Hourly Crime Pattern for Top DBSCAN Clusters

This chart shows the hourly crime counts for the top DBSCAN clusters.

Cluster 0 (orange line) has the highest crime activity, peaking around noon and early evening.

Other clusters (13, 15, 37) have much lower crime counts and flatter patterns.



Preparation For Training

This table shows the crime categories grouped into 4 crime type groups.

- **Group 3 (Larceny Theft)** is the largest, with over 250,000 cases.
- **Group 2** includes common crimes like Assault, Burglary, and Motor Vehicle Theft.
- **Group 1** covers smaller categories such as Warrant and Drug Offense.
- **Group 0 (Human Trafficking)** is rare, with only 3 cases.

crime_type_group	incident_category	count
0	Human Trafficking - Involuntary Servitude	3
1	Other Miscellaneous	64765
	Warrant	28770
	Drug Offense	25798
	Weapons Offense	12008
	Traffic Violation Arrest	8599
2	Malicious Mischief	61539
	Assault	59704
	Non-Criminal	54216
	Burglary	52287
	Motor Vehicle Theft	51536
3	Larceny Theft	252675

dtype: int64

Model Training

DecisionTree

	precision	recall	f1-score	support
1	0.46	0.53	0.49	30104
2	0.78	0.75	0.77	99206
3	0.81	0.80	0.80	50535
accuracy			0.73	179845
macro avg	0.68	0.69	0.69	179845
weighted avg	0.74	0.73	0.73	179845

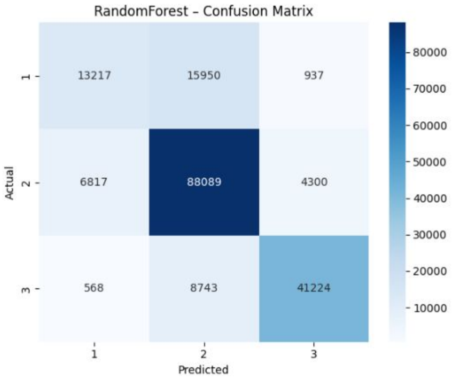
LogisticRegression

	precision	recall	f1-score	support
1	0.36	0.55	0.43	30104
2	0.78	0.63	0.70	99206
3	0.78	0.83	0.81	50535
accuracy			0.67	179845
macro avg	0.64	0.67	0.65	179845
weighted avg	0.71	0.67	0.69	179845

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

```
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 180s 20ms/step - accuracy: 0.5820 - loss: 0.9284 - val_accuracy: 0.5804 - val_loss: 0.9322
Epoch 12/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 204s 20ms/step - accuracy: 0.5831 - loss: 0.9266 - val_accuracy: 0.5806 - val_loss: 0.9316
Epoch 13/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 200s 20ms/step - accuracy: 0.5836 - loss: 0.9256 - val_accuracy: 0.5794 - val_loss: 0.9324
Epoch 14/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 182s 20ms/step - accuracy: 0.5821 - loss: 0.9272 - val_accuracy: 0.5802 - val_loss: 0.9330
Epoch 15/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 203s 20ms/step - accuracy: 0.5827 - loss: 0.9267 - val_accuracy: 0.5809 - val_loss: 0.9314
Epoch 16/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 183s 20ms/step - accuracy: 0.5828 - loss: 0.9264 - val_accuracy: 0.5804 - val_loss: 0.9313
Epoch 17/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 216s 22ms/step - accuracy: 0.5834 - loss: 0.9256 - val_accuracy: 0.5801 - val_loss: 0.9324
Epoch 18/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 185s 20ms/step - accuracy: 0.5836 - loss: 0.9252 - val_accuracy: 0.5816 - val_loss: 0.9309
Epoch 19/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 182s 20ms/step - accuracy: 0.5851 - loss: 0.9239 - val_accuracy: 0.5807 - val_loss: 0.9319
Epoch 20/20
8992/8992 ━━━━━━━━━━━━━━━━━━━━ 201s 20ms/step - accuracy: 0.5842 - loss: 0.9246 - val_accuracy: 0.5820 - val_loss: 0.9295
<keras.src.callbacks.history.History at 0x7d9cdaa58a10>
```



Based on accuracy and F1-scores, RandomForest is the best-performing model, achieving 79% accuracy and the highest scores across all classes.

11.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.