Project Proposal

BSCS *[6-A]*

Introduction to Data Science Lab

Analysis and Prediction of Crime Trends in Chicago (2022)

Group Members:

Zaynah Nadeem	03-134231-061
Natasha Fatima	03-134231-055
Hafsa Mueen	03-134231-017
Laveeza Azmat	03-134232-042

Submitted To: Mr. Muhammad Umar Tariq

Department of Computer Sciences
Bahria University Lahore Campus, Lahore

Abstract

Urban crime presents a persistent challenge for modern cities, particularly large metropolitan areas such as Chicago, where crime patterns vary significantly across time and location. The dynamic and complex nature of criminal activities makes it difficult for law enforcement agencies to continuously monitor trends, predict high-risk areas, and allocate resources effectively using traditional methods.

This project focuses on analyzing and predicting crime trends in Chicago using data science and machine learning techniques. The Chicago Crime Dataset for the year 2022 was utilized to study spatial and temporal crime patterns. After data preprocessing and exploratory analysis, machine learning models including Random Forest and Logistic Regression were trained to predict crime-prone areas, achieving an accuracy of approximately **91%**. A Python-based backend integrated with a SQL Server database was developed to store crime records and prediction results. These outcomes are visualized through an interactive Power BI dashboard that provides near real-time insights into crime trends and predicted hotspots. The system supports data-driven decision-making and demonstrates how analytics, machine learning, and visualization can be combined for proactive crime management.

Introduction

Crime in urban environments has a direct impact on public safety, economic stability, and the overall quality of life of citizens. Chicago has historically experienced high crime rates across various districts, making crime analysis and prevention a critical concern for law enforcement agencies. Conventional crime monitoring systems are often reactive, relying heavily on historical reports without offering predictive insights.

With the increasing availability of large-scale crime datasets, data science and machine learning techniques provide powerful tools to analyze historical data and uncover hidden patterns. These techniques enable authorities to move from reactive policing to predictive and preventive strategies. This project aims to analyze crime patterns in Chicago during 2022, identify crime hotspots, and develop a predictive system supported by a real-time visualization dashboard.

The scope of this project includes data preprocessing, exploratory data analysis, predictive modeling, backend and database integration, and dashboard visualization. The study does not involve personal identification of individuals, direct law enforcement intervention, or crime data beyond the year 2022.

Problem Statement

Urban crime data is large, complex, and continuously evolving. Law enforcement agencies often struggle to analyze this data efficiently and identify crime-prone areas in advance. The absence of predictive systems makes it difficult to allocate resources optimally and respond proactively to emerging crime trends. Therefore, there is a need for an integrated system that

can analyze historical crime data, predict potential hotspots, and present insights through an easy-to-understand visual interface.

Objectives

The main objectives of this project are:

1. To analyze crime patterns in Chicago during 2022 across different locations, crime categories, and time periods.
2. To develop a machine learning model capable of predicting high-risk crime areas based on historical data.
3. To design and implement a backend system connected to a database for storing crime data and model predictions.
4. To create an interactive Power BI dashboard for real-time visualization of crime trends and predictive insights.

Literature Review

Several studies have explored crime analysis and prediction using data science techniques:

- ***Predicting Crime Using Machine Learning*** employed Random Forest and Decision Tree models and concluded that spatial and temporal factors play a major role in predicting crime hotspots.
- ***Real-Time Crime Data Visualization Using BI Tools*** demonstrated that dashboards built using Power BI and live data connections enhance situational awareness and decision-making efficiency.
- ***Crime Forecasting in Chicago*** applied Logistic Regression and time-series analysis to forecast high-risk areas in advance.

This project builds upon existing research by integrating predictive modeling, backend systems, and real-time visualization into a single, unified framework.

Methodology

1. Data Collection

- Dataset: Chicago Crime Dataset (2022)
- Source: Chicago Data Portal / Kaggle
- Key attributes include ID, Date, Primary Type, Description, District, Beat, Ward, Latitude, Longitude, and Location Description.

2. Data Preprocessing

- Missing and inconsistent values were handled to ensure data quality.
- Categorical variables such as crime type and district were encoded.

- Feature engineering was performed by extracting hour, day of week, month, and day/night indicators to capture temporal patterns.

3. Exploratory Data Analysis

- Crime distribution was analyzed by type, district, and time period.
- Temporal trends were examined using hourly, daily, and monthly analyses.
- Geographic heatmaps were created to visually identify crime hotspots.

4. Model Training

- Algorithms used: Random Forest and Logistic Regression.
- Data was split into 80% training and 20% testing sets.
- Models were evaluated using accuracy, precision, recall, and F1-score.

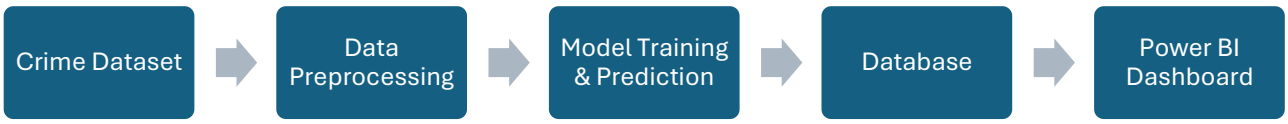
5. Backend and Database Integration

- A backend system was developed using Python.
- A database was created to store cleaned crime data and prediction outputs.
- The backend supports data insertion, prediction retrieval, and communication with the dashboard.

6. Power BI Dashboard

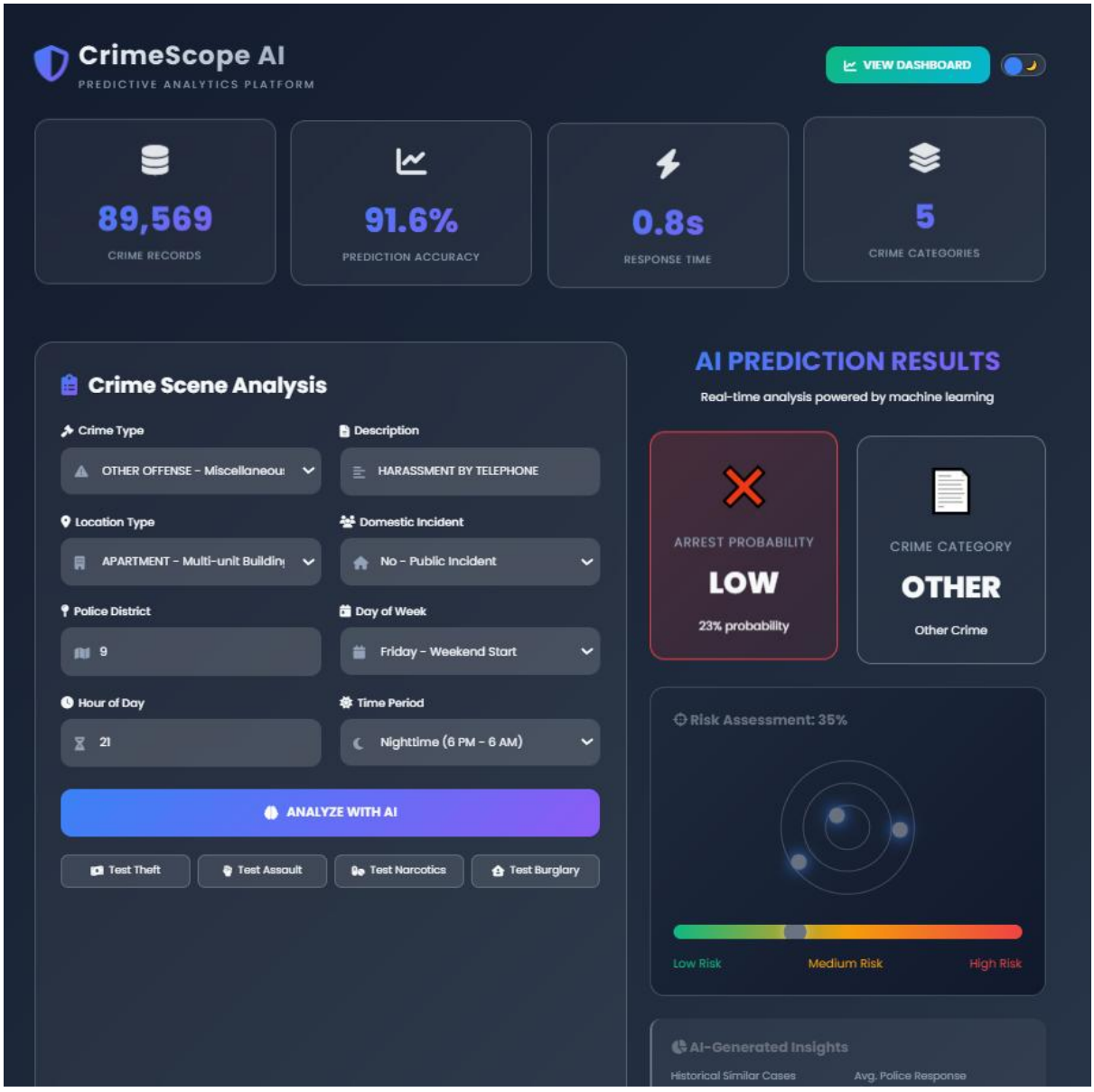
- The database was connected live to Power BI
- Interactive filters were provided for district, crime type, and time.
- Visualizations include charts, maps, and predictive risk indicators.

Flow Chart

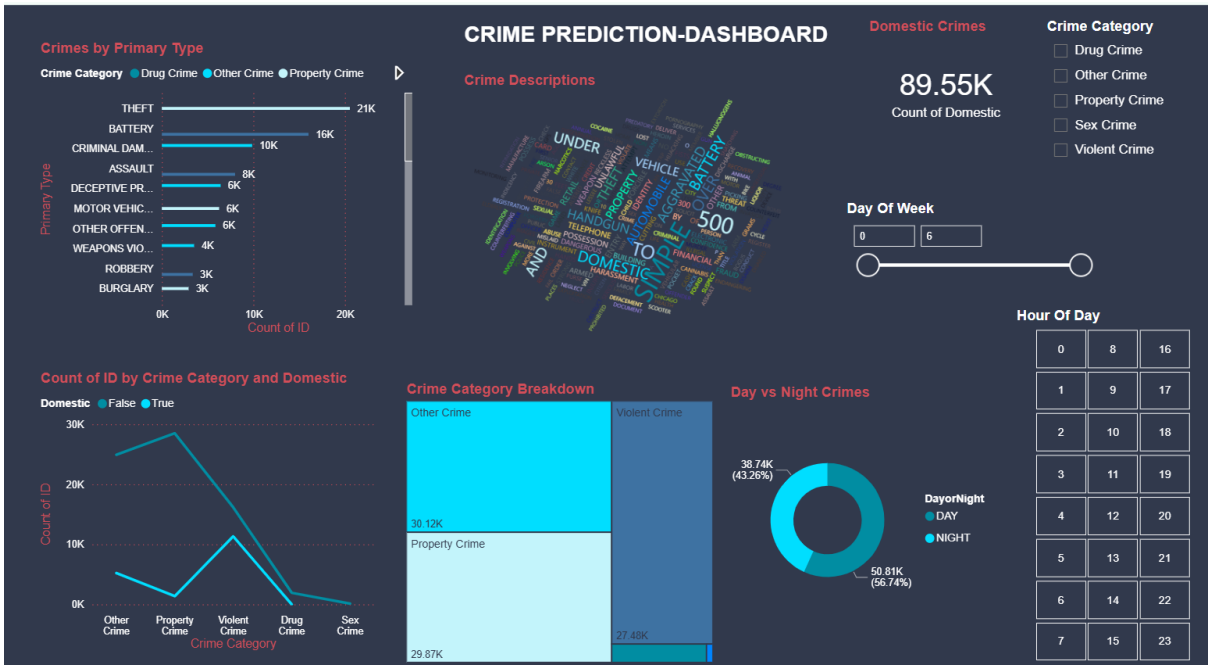


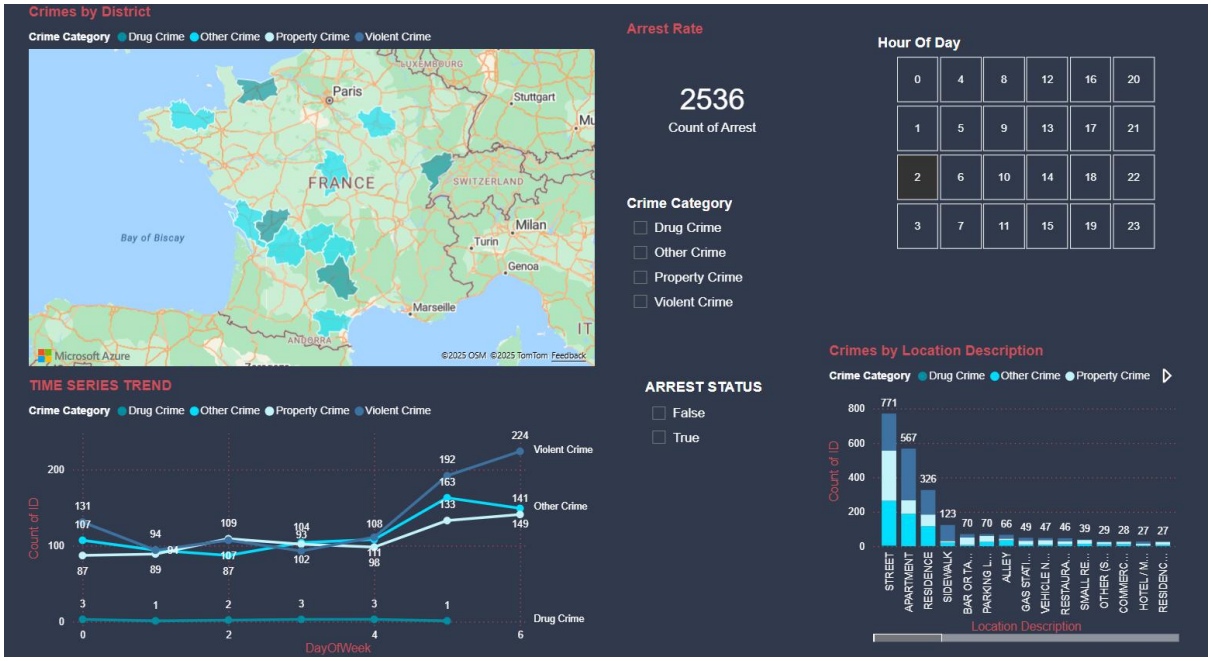
Results

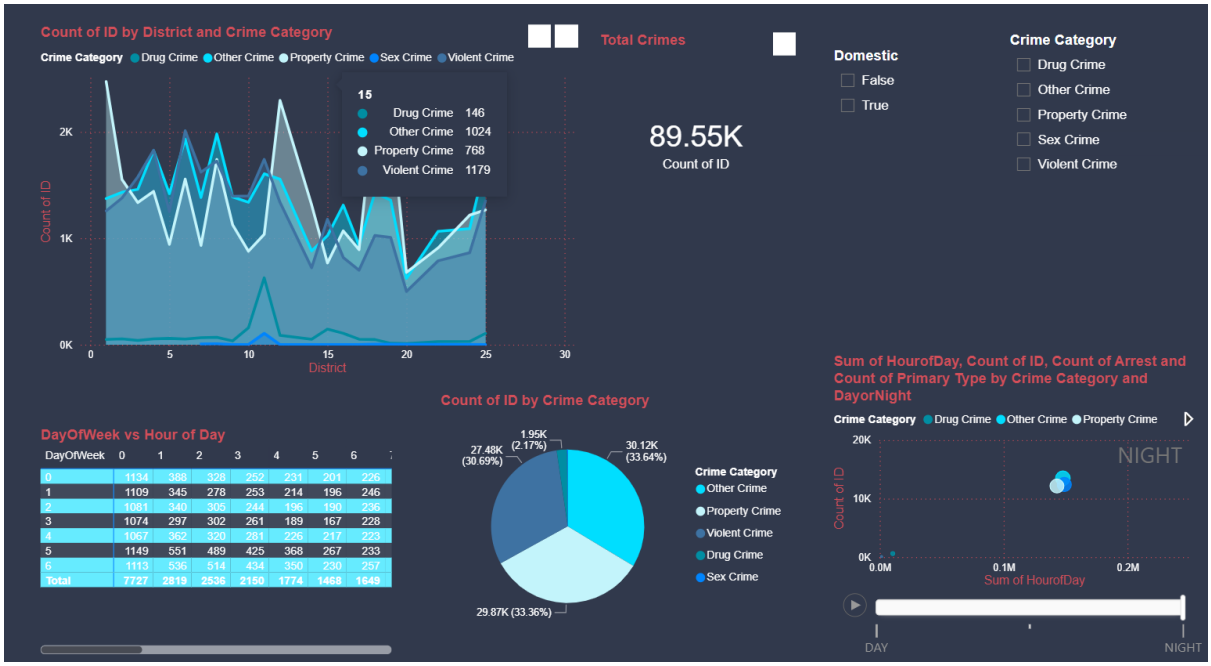
- The model was able to predict the category of crime (such as property crime, violent crime, or other crime types) based on historical patterns.
- Property crimes were observed to be more frequent compared to violent crimes. The Random Forest model achieved approximately **91%** accuracy in predicting high-risk areas.



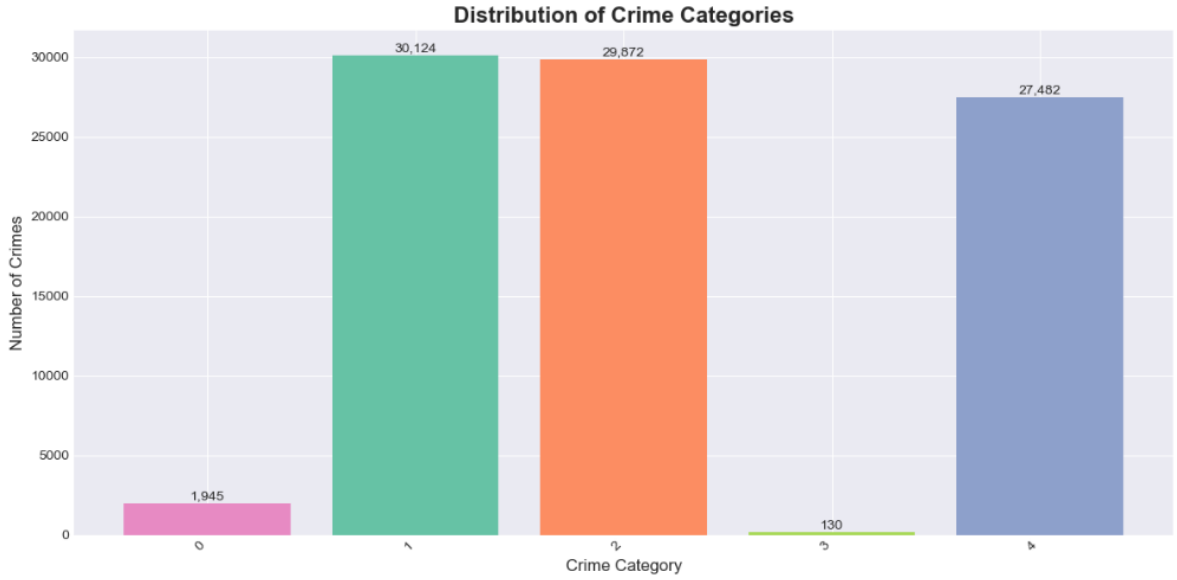
- The Power BI dashboard provides real-time visualization of crime trends, predicted risk levels, and predicted crime categories, enabling more informed decision-making.

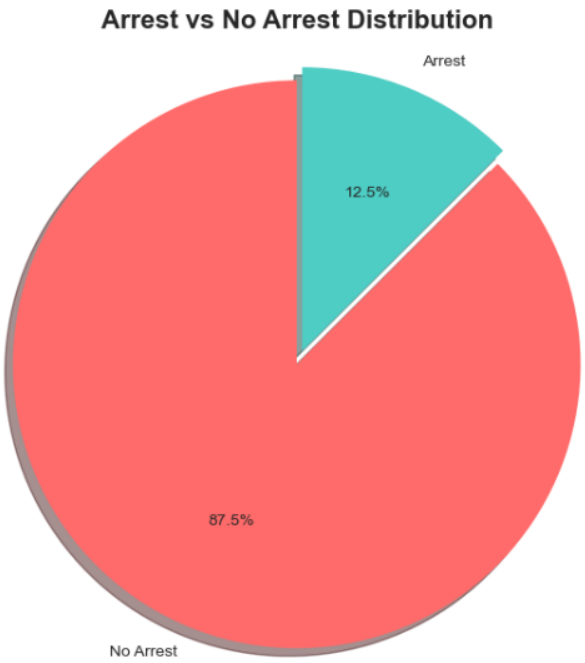

CRIME PREDICTION-DASHBOARD
Crimes by Primary Type
Crime Category Drug Crime Other Crime Property Crime
THEFT 21K
BATTERY 16K
CRIMINAL DAM... 10K
ASSAULT 8K
DECEPTIVE PR... 6K
MOTOR VEHIC... 6K
OTHER OFFEN... 6K
WEAPONS VIO... 4K
ROBBERY 3K
BURGLARY 3K
Primary Type
0K 10K 20K
Count of ID
Crime Descriptions
Domestic Crimes
89.55K
Count of Domestic
Crime Category
Drug Crime
Other Crime
Property Crime
Sex Crime
Violent Crime
Day Of Week
0 6
Hour Of Day
0 8 16
1 9 17
2 10 18
3 11 19
4 12 20
5 13 21
6 14 22
7 15 23
Count of ID by Crime Category and Domestic
Domestic False True
30K
20K
10K
0K
Count of ID
Other Crime
Property Crime
Violent Crime
Drug Crime
Sex Crime
Crime Category
Crime Category Breakdown
Other Crime
30.12K
Violent Crime
27.48K
Property Crime
29.87K
Day vs Night Crimes
38.74K
(43.26%)
50.81K
(56.74%)
DayorNight
DAY
NIGHT

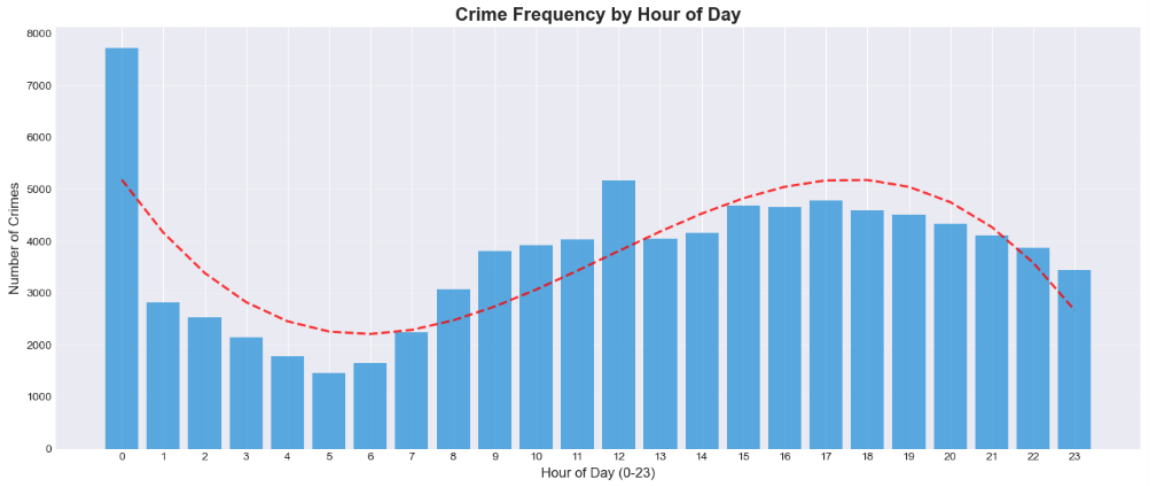

Crimes by District
Crime Category Drug Crime Other Crime Property Crime Violent Crime
LUXEMBOURG
Paris
Stuttgart
FRANCE
SWITZERLAND
Bay of Biscay
Milan
Turin
Genoa
Marseille
ANDORRA
Microsoft Azure
©2025 OSM ©2025 TomTom Feedback
Arrest Rate
2536
Count of Arrest
Hour Of Day
0 4 8 12 16 20
1 5 9 13 17 21
2 6 10 14 18 22
3 7 11 15 19 23
Crime Category
Drug Crime
Other Crime
Property Crime
Violent Crime
ARREST STATUS
False
True
TIME SERIES TREND
Crime Category Drug Crime Other Crime Property Crime Violent Crime
Count of ID
Violent Crime
Other Crime
Drug Crime
DayOfWeek
Crimes by Location Description
Crime Category Drug Crime Other Crime Property Crime
771
567
326
123
70 70 66 49 47 46 39 29 28 27 27
STREET
APARTMENT
RESIDENCE
SIDEWALK
BAR OR TA...
PARKING L...
ALLEY
GAS STATI...
VEHICLE N...
RESTAURA...
SMALL RE...
OTHER (S...
COMMERC...
HOTEL / M...
RESIDENC...
Location Description

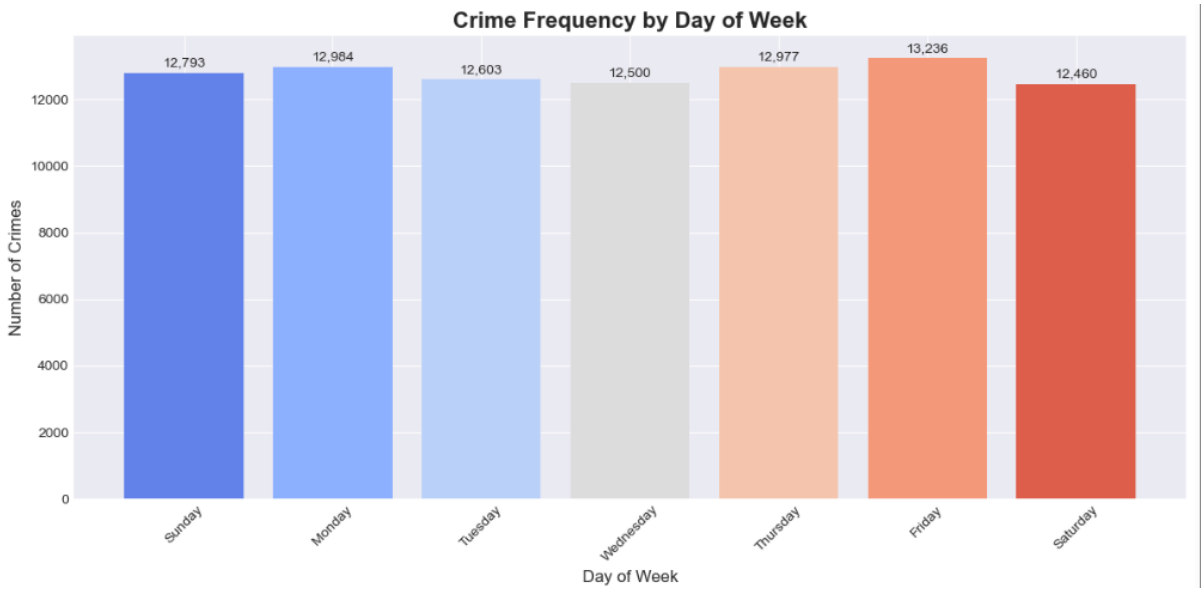


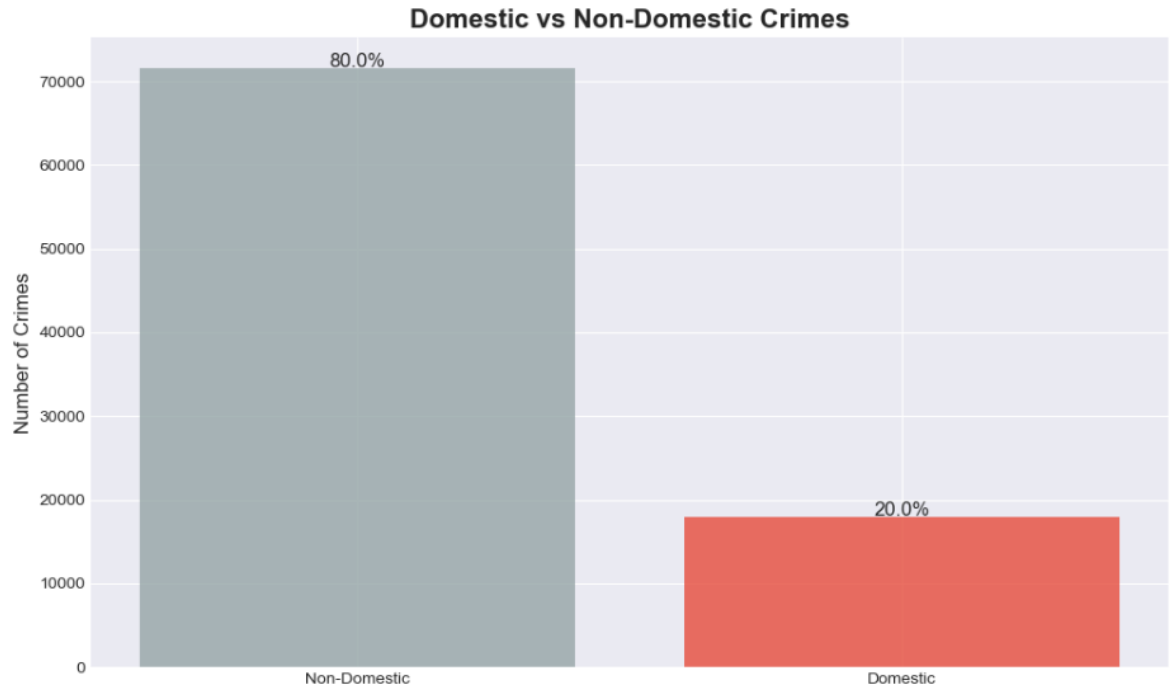
Data Visualization and Insights

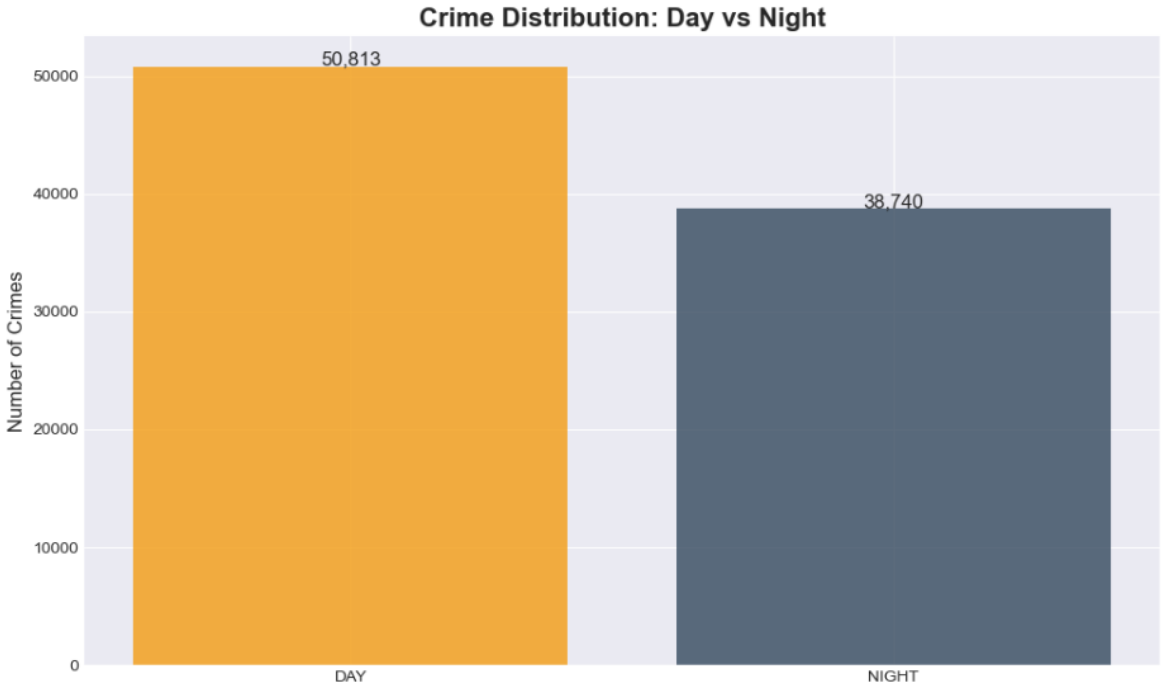


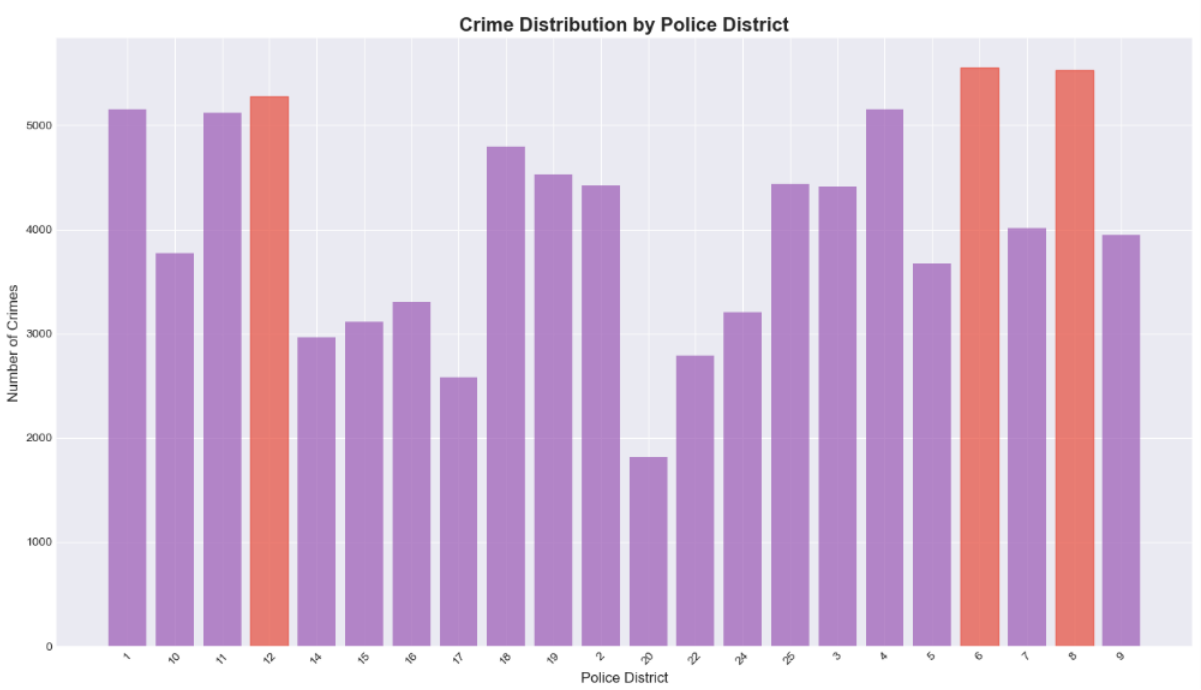

Arrest vs No Arrest Distribution
Arrest
12.5%
87.5%
No Arrest

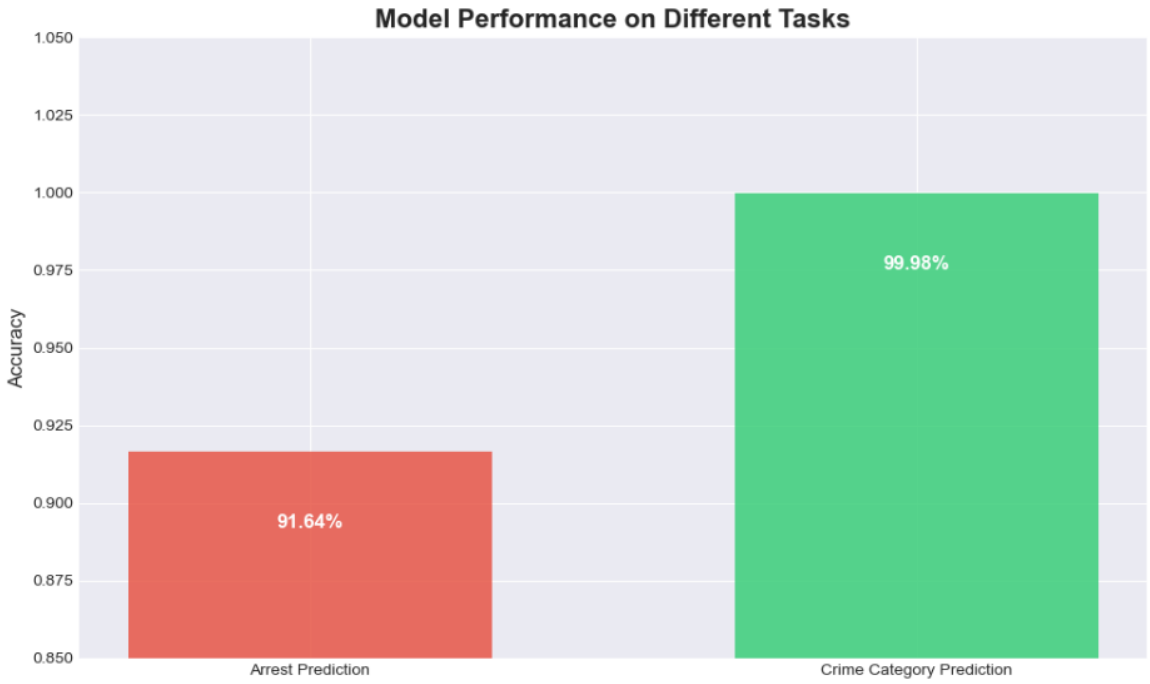

Crime Frequency by Hour of Day
Number of Crimes
Hour of Day (0-23)

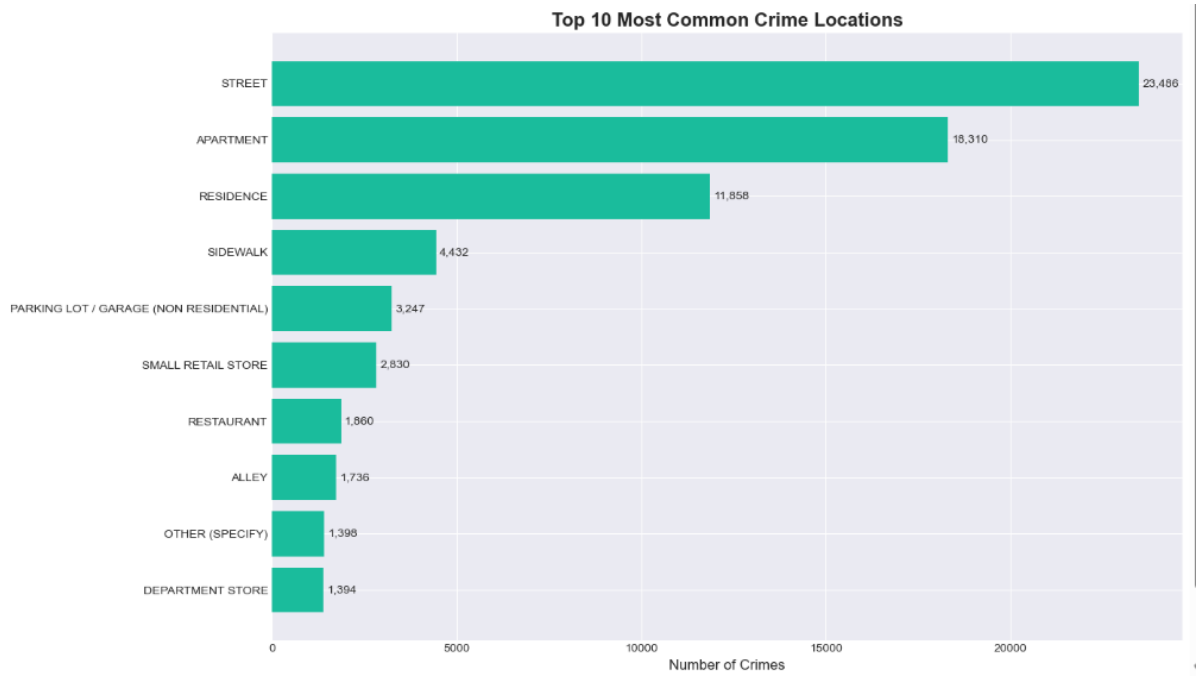

Crime Frequency by Day of Week
12,793
12,984
12,603
12,500
12,977
13,236
12,460
Number of Crimes
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week


Domestic vs Non-Domestic Crimes
80.0%
20.0%
Number of Crimes
70000
60000
50000
40000
30000
20000
10000
0
Non-Domestic
Domestic


Crime Distribution: Day vs Night
50,813
38,740
Number of Crimes
50000
40000
30000
20000
10000
0
DAY
NIGHT


Crime Distribution by Police District
Number of Crimes
0
1000
2000
3000
4000
5000
1
10
11
12
14
15
16
17
18
19
2
20
22
24
25
3
4
5
6
7
8
9
Police District


Model Performance on Different Tasks
Accuracy
1.050
1.025
1.000
0.975
0.950
0.925
0.900
0.875
0.850
91.64%
99.98%
Arrest Prediction
Crime Category Prediction



Evaluation:

```
Logistic Regression Accuracy: 0.8753965348356373
              precision    recall  f1-score   support

N              0.89      0.81      0.84       98
Y              0.82      0.90      0.86       97

accuracy                           0.99       195
macro avg      0.85      0.85      0.85       195
weighted avg   0.85      0.85      0.85       195

 Random Forest Accuracy: 0.9163642454357658
              precision    recall  f1-score   support

N              0.93      0.97      0.95     15686
Y              0.74      0.51      0.60      2225

accuracy                           0.92     17911
macro avg      0.84      0.74      0.78     17911
weighted avg   0.91      0.92      0.91     17911
```

Discussion

The results indicate that historical crime data contains meaningful spatial and temporal patterns that can be leveraged for crime prediction. The predictive model successfully identifies high-risk areas, while the dashboard allows users to explore crime trends interactively.

These findings align with previous research, confirming that location and time are strong predictors of crime activity. The integration of real-time visualization significantly improves the usability and practical value of the system for decision-makers.

Limitations

- The analysis is limited to data from a single year (2022).

- Socioeconomic and demographic variables were not included.
- Predictions are probabilistic and should complement, not replace, human judgment.
- Model accuracy depends on data quality and timely updates.

Conclusion and Future Work

This project demonstrates the effective application of data science, machine learning, and visualization techniques for analyzing and predicting crime trends. The system provides actionable insights that support proactive crime prevention and efficient resource allocation.

Future improvements include:

- Extending the dataset to multiple years
- Incorporating socioeconomic indicators
- Exploring advanced machine learning and deep learning models
- Implementing continuous model retraining with new data

References

1. Chicago Police Department, “Crime Data 2022,” Chicago Data Portal. https://www.kaggle.com/datasets/georgehanyfouad/crime-prediction-in-chicago-in-2022