

CAMBRIGE'S CRIME ANALYSIS

Big Data Technology
Group 5

A dark, moody photograph of a man with short hair and a beard, wearing a light-colored suit jacket over a patterned shirt. He is looking out of the open rear window of a car. The interior of the car is visible, and the background is blurred.

Daffa Anis Fahrizi (2106731365)
Miranti Anggunsari (2106731472)
Lavly Rantissa Zunnuraina R. (2206830624)
Safia Amita Khoirunnisa (2206059420)
Annisa Ardelia Setiawan (2206059471)

LIST OF CONTENTS

- 01 PENDAHULUAN
- 02 TOOLS
- 03 IMPLEMENTASI
- 04 ANALISIS DATA
- 05 KESIMPULAN



A woman with short dark hair and bangs, wearing a black jacket with silver studs, is leaning into a car window. She is looking directly at a man whose profile is visible on the left. The scene is set at night or in low light, with the interior of the car visible.

PENDAHULUAN

Dalam beberapa dekade terakhir, urbanisasi yang cepat dan ketimpangan sosial telah menjadi pemicu utama peningkatan tingkat kriminalitas. Lingkungan perkotaan yang padat sering kali menjadi pusat berbagai jenis kejahatan. Menghadapi situasi ini, pendekatan tradisional dalam penanganan kejahatan sering kali kurang efektif tanpa adanya dukungan data yang akurat dan komprehensif. Oleh karena itu, analisis data kriminalitas menjadi sangat penting untuk memahami dinamika kejahatan di wilayah Cambrige.

TOOLS - HADOOP

Untuk melakukan analisa data, kami menggunakan tools Hadoop yang merupakan suatu framework open-source yang dapat digunakan untuk **mengolah** dan **menyimpan data** yang sangat besar. Penggunaan Hadoop ini memungkinkan kami untuk menganalisis dataset berukuran besar dengan membagi data menjadi bagian-bagian yang dapat diproses secara paralel.



HDFS

Sistem penyimpanan yang terdistribusi ini memungkinkan kami untuk menyimpan data dalam blok-blok kecil di dalam berbagai node pada kluster.



MapReduce

Model pemrograman ini digunakan untuk memetakan data mentah ke format yang dibutuhkan dan melakukan agregasi hasil untuk menghasilkan output.



YARN

Framework ini kami gunakan untuk manajemen sumber daya dalam kluster Hadoop.

TOOLS - SPARKSQL



SparkSQL memberikan pendekatan yang lebih terfokus pada pengolahan data terstruktur menggunakan SQL di dalam ekosistem Apache Spark. SparkSQL sendiri merupakan modul dalam Apache Spark yang memungkinkan pemrosesan data terstruktur dan semi-terstruktur menggunakan bahasa kueri SQL. Hal ini memungkinkan analisis yang lebih intuitif dan efisien ketika data yang digunakan memiliki struktur tabel, seperti data kriminal yang biasanya tersimpan dalam format CSV atau database SQL. Hasil dari kueri SparkSQL ini dapat langsung dianalisis lebih lanjut menggunakan visualisasi dengan Matplotlib.

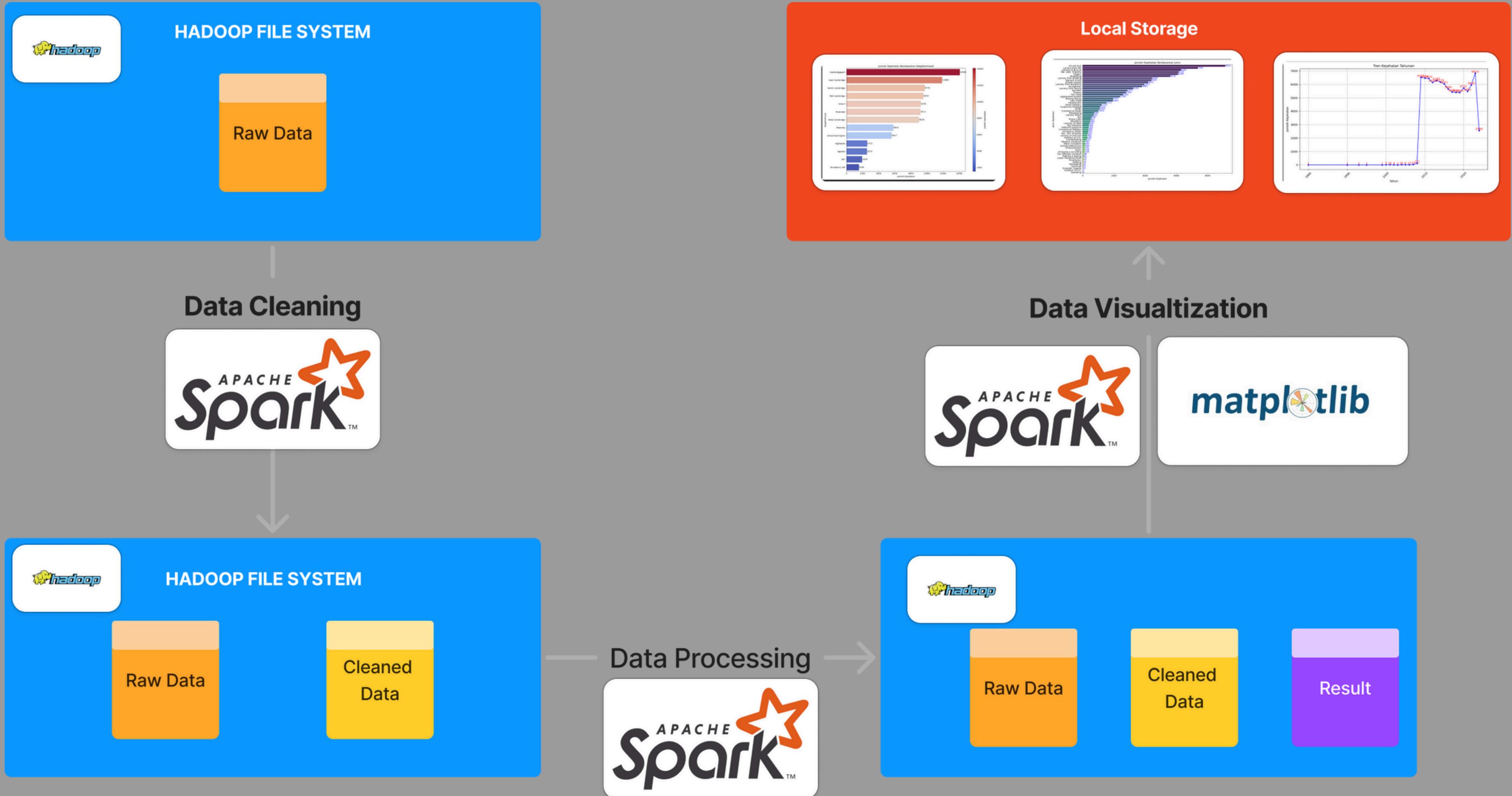
TOOLS - PYTHON



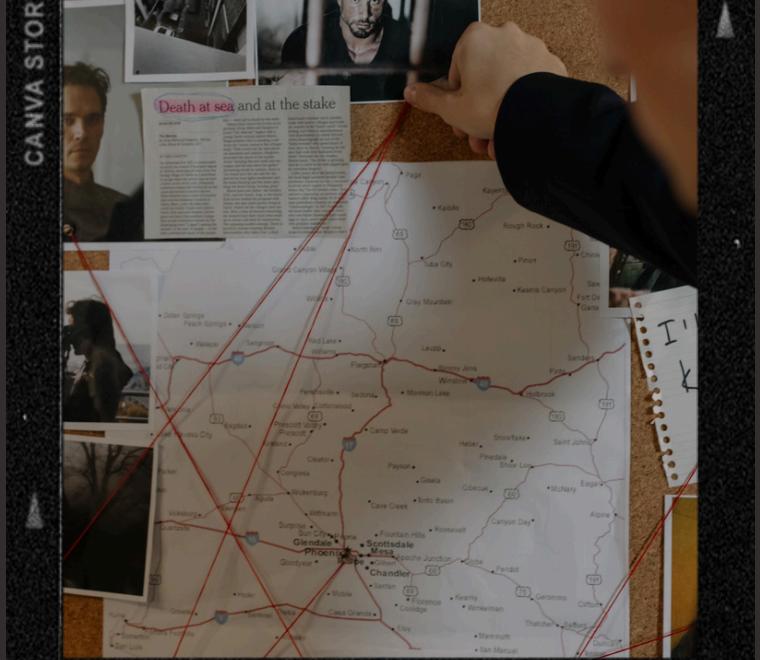
Dalam proyek ini, beberapa library dari Python, seperti Matplotlib, Pandas, dan Seaborn digunakan untuk membuat visualisasi distribusi kejahatan berdasarkan tahun, jenis kejahatan, dan wilayah (neighborhood). Dengan library-library ini, data dari Hadoop dan Spark yang telah diproses dapat divisualisasikan menjadi grafik garis dan batang yang memberikan wawasan lebih dalam. Fitur anotasi dan kustomisasi Matplotlib membantu menyoroti poin-poin penting, seperti jumlah kejahatan di setiap tahun atau perbedaan mencolok antar jenis kejahatan.







IMPLEMENTASI



DATA YANG DIPAKAI



Cambridge Crime Data

Reported crime in Cambridge (2009-2024)

Last Updated: 5 months ago (Version 1)

About this Dataset

Overview:

This dataset comprises crime incidents reported in the City of Cambridge, as featured in the Cambridge Police Department's Annual Crime Reports, spanning from 2009 to 2024. The data provides detailed information about various crime types and their occurrences across different neighborhoods in Cambridge.

Dataset Details:

- **File Number:** A unique identifier for each crime report (Text).
- **Date of Report:** The date when the crime incident was reported (Floating Timestamp).
- **Crime Data Time:** The specific time when the crime incident occurred (Floating Timestamp).
- **Crime:** The type of crime committed (Text).
- **Reporting Area:** A numerical identifier for the community area where the crime occurred (Number).
- **Neighborhood:** The name of the neighborhood where the crime was reported (Text).
- **Location:** The street information indicating the approximate location of the crime (Text).

- **Last Updated:** June 15, 2024.

- **Revisions:** The dataset will be periodically updated to ensure compliance with local, state, and federal privacy rights and legal requirements.

- **Volume:** Dataset ini memiliki 95923 row data yang mencakup data kejahatan selama periode 15 tahun dengan berbagai variabel.
- **Variety:** Dataset ini mengandung berbagai jenis data, seperti: Data teks (jenis kejahatan, nama area), Data temporal (waktu kejadian, waktu laporan), Data numerik (kode wilayah).
- **Velocity:** Meskipun dataset ini tidak real-time, data ini mendukung analisis tren dan memungkinkan pembaruan secara berkala untuk menangkap pola terkini dalam waktu tertentu.



CrimeAnalysis_FeatureEngineering

Updated 5mo ago

6 comments · Cambridge Crime Data

▲ 34

Bronze ...

DATA CLEANING

- Menginisialisasi spark session
- Memuat dataset dari hdfs
- Menghapus baris dengan nilai admin error
- Mengganti nilai kosong dengan “NA”
- Mengekstrasi tahun dari kolom crime data time
- Menyimpan dataset bersih ke HDFS
- Menutup spark session

Data awal berisi 95924 => menjadi 92755

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, regexp_extract, col

# 1. Inisialisasi Spark Session
spark = SparkSession.builder \
    .appName("Crime Data Cleaning") \
    .config("spark.executor.memory", "2g") \
    .config("spark.driver.memory", "2g") \
    .getOrCreate()

# 2. Memuat Dataset dari HDFS
hdfs_path = "hdfs:///data/crime_reports/crime_reports.csv" # Path dataset asli
df = spark.read.csv(hdfs_path, header=True, inferSchema=True)

# Data sebelum pembersihan
print(f"Jumlah baris sebelum pembersihan: {df.count()}")

# 3. Menghapus baris dengan nilai "Admin Error" di kolom 'Crime', 'Neighborhood', atau 'Location'
df_clean = df.filter(~(
    (df['Crime'] == 'Admin Error') |
    (df['Neighborhood'] == 'Admin Error') |
    (df['Location'] == 'Admin Error')
))

# 4. Mengganti nilai kosong (None) dengan "NA" di semua kolom
df_clean = df_clean.fillna("NA")

# 5. Ekstrak tahun dari kolom 'Crime Date Time'
# Menggunakan regex untuk menangkap tahun pertama dari kolom
df_clean = df_clean.withColumn(
    "Crime Year",
    regexp_extract(col("Crime Date Time"), r"\d{4}", 1) # Ambil tahun pertama yang ditemukan
)

# Verifikasi data setelah pembersihan
print(f"Jumlah baris setelah pembersihan: {df_clean.count()}")
df_clean.show(5)

# 6. Simpan dataset bersih ke HDFS dalam satu file
output_clean_path = "hdfs:///data/cleaned_crime_reports2.csv" # Path dataset bersih
df_clean.coalesce(1).write.csv(output_clean_path, header=True)

# 7. Menutup Spark session
spark.stop()
```

OUTPUT

File Number	Date of Report	Crime Date Time	Crime Reporting
2009-01323 02/21/2009 09:53:... 02/21/2009 09:20 ... Threats			
2009-01324 02/21/2009 09:59:... 02/20/2009 22:30 ... Auto Theft			
2009-01327 02/21/2009 12:32:... 02/19/2009 21:00 ... Hit and Run			
2009-01331 02/21/2009 03:05:... 02/21/2009 15:00 ... Larceny (Misc)			
2009-01346 02/22/2009 05:02:... 02/22/2009 05:02 OUI			

only showing top 5 rows

DATA KOTOR

/data									Gol				
Show 25 entries									Search:				
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 20:58	0	0 B	cleaned_crime_reports1.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 21:04	0	0 B	cleaned_crime_reports2.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 17:54	0	0 B	crime_reports					

Showing 1 to 3 of 3 entries

Previous 1 Next

DATA BERSIH

/data									Gol				
Show 25 entries									Search:				
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 20:58	0	0 B	cleaned_crime_reports1.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 21:04	0	0 B	cleaned_crime_reports2.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 17:54	0	0 B	crime_reports					

Showing 1 to 3 of 3 entries

Previous 1 Next

DATA PROCESSING

- Menginisialisasi spark session
- Memuat dataset yang sudah dibersihkan
- menghitung jumlah kejahatan per tahun
- menghitung jumlah kejahatan per neighbourhood
- menghitung jumlah kejahatan per crime type
- menampilkan hasil
- menyimpan hasil ke HDFS
- menutup spark session

```
# 7. Simpan hasil pengolahan ke HDFS
output_base_path = "hdfs:///output/Crime11"
annual_crimes.write.csv(f"{output_base_path}/annual_crimes.csv",
header=True)
neighborhood_crimes.write.csv(f"{output_base_path}/neighborhood_crimes.csv",
, header=True)
crime_type_crimes.write.csv(f"{output_base_path}/crime_type_crimes.csv",
header=True)

# 8. Menutup Spark session
spark.stop()
```

IMPLEMENTASI

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# 1. Inisialisasi Spark Session
spark = SparkSession.builder \
    .appName("Analyze Crime Data by Year") \
    .config("spark.executor.memory", "2g") \
    .config("spark.driver.memory", "2g") \
    .getOrCreate()

# 2. Memuat Dataset yang Sudah Dibersihkan
cleaned_path = "hdfs://data/cleaned_crime_reports2.csv"
df_cleaned = spark.read.csv(cleaned_path, header=True, inferSchema=True)

# 3. Hitung jumlah kejahatan per tahun
annual_crimes = df_cleaned.groupBy("Crime Year").count().orderBy("Crime Year")

# 4. Hitung jumlah kejahatan per Neighborhood
neighborhood_crimes
df_cleaned.groupBy("Neighborhood").count().orderBy("count",
ascending=False)

# 5. Hitung jumlah kejahatan per Crime Type
crime_type_crimes = df_cleaned.groupBy("Crime").count().orderBy("count",
ascending=False)

# 6. Tampilkan hasil
print("Jumlah Kejahatan per Tahun:")
annual_crimes.show()

print("Jumlah Kejahatan per Neighborhood:")
neighborhood_crimes.show()

print("Jumlah Kejahatan per Crime Type:")
crime_type_crimes.show()
```

DATA PROCESSING

Jumlah kejahatan per tahun

```
24/12/04 15:26:58 INFO CodeGenerator: Code ge
+-----+-----+
|Crime Year|count|
+-----+-----+
| 1980| 2|
| 1990| 1|
| 1993| 2|
| 1995| 1|
| 1999| 1|
| 2000| 5|
| 2001| 14|
| 2002| 4|
| 2003| 2|
| 2004| 16|
| 2005| 8|
| 2006| 17|
| 2007| 25|
| 2008| 121|
| 2009| 6515|
| 2010| 6474|
| 2011| 6433|
| 2012| 6144|
| 2013| 6285|
| 2014| 6179|
+-----+-----+
only showing top 20 rows
```

Jumlah kejahatan berdasarkan jenis

```
+-----+-----+
|          Crime|count|
+-----+-----+
| Hit and Run| 9121|
| Larceny from MV| 7369|
| Larceny of Bicycle| 6292|
| Mal. Dest. Property| 6130|
| Forgery| 6109|
| Shoplifting| 5610|
| Larceny from Buil...| 4464|
| Warrant Arrest| 4405|
| Simple Assault| 4201|
| Larceny from Resi...| 3953|
| Housebreak| 3804|
| Larceny from Person| 3228|
| Accident| 2871|
| Threats| 2766|
| Flim Flam| 2583|
| Aggravated Assault| 2399|
| Missing Person| 1953|
| Auto Theft| 1941|
| Harassment| 1521|
| Street Robbery| 1223|
+-----+-----+
only showing top 20 rows
```

IMPLEMENTASI

Jumlah kejahatan per wilayah

```
+-----+-----+
| Neighborhood|count|
+-----+-----+
| Cambridgeport| 14080|
| East Cambridge| 11898|
| North Cambridge| 9733|
| Mid-Cambridge| 9549|
| Area 4| 9236|
| Riverside| 9211|
| West Cambridge| 9028|
| Peabody| 5824|
| Inman/Harrington| 5617|
| Highlands| 2571|
| Agassiz| 2533|
| MIT| 1939|
| Strawberry Hill| 1536|
+-----+-----+
Jumlah Kejahatan per Crime Type:
```

HASIL DATA PROCESSING

Browse Directory													
									Go!				
Show 25 entries									Search:				
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 21:04	0	0 B	data					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 22:27	0	0 B	output					

Showing 1 to 2 of 2 entries

Previous 1 Next

Browse Directory													
									Go!				
/output/Crime11									Search:				
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 22:27	0	0 B	annual_crimes.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 22:27	0	0 B	crime_type_crimes.csv					
<input type="checkbox"/>	drwxr-xr-x	daffafahrizi	supergroup	0 B	Dec 04 22:27	0	0 B	neighborhood_crimes.csv					

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2024.

DATA VISUALIZATION

- Mengimpor library yang dibutuhkan
- Menginisialisasi spark session
- Membaca data dari HDFS
- Mengkonversi data ke panda dataframe

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 from pyspark.sql import SparkSession
5
6 # Inisialisasi Spark session
7 spark = SparkSession.builder.appName('CrimeDataAnalysis').getOrCreate()
8
9 # Membaca data dari Hadoop/HDFS
10 annual_crimes = spark.read.csv('hdfs:///output/Crime11/annual_crimes.csv/part-00000.csv', header=True, inferSchema=True)
11 crime_type_crimes = spark.read.csv('hdfs:///output/Crime11/crime_type_crimes.csv/part-00000.csv', header=True, inferSchema=True)
12 wilayah_crimes = spark.read.csv('hdfs:///output/Crime11/wilayah_crimes.csv/part-00000.csv', header=True, inferSchema=True)
13
14 # Convert DataFrame to pandas DataFrame untuk visualisasi
15 annual_crimes_df = annual_crimes.toPandas()
16 crime_type_crimes_df = crime_type_crimes.toPandas()
17 wilayah_crimes_df = wilayah_crimes.toPandas()
```

IMPLEMENTASI

IMPLEMENTASI

DATA VISUALIZATION

```
# 1. Visualisasi Data Kejahatan Tahunan (Annual Crimes)
plt.figure(figsize=(10,6))
sns.lineplot(x='Crime Year', y='count', data=annual_crimes_df, marker='o', color='b')
plt.title('Tren Kejahatan Tahunan')
plt.xlabel('Tahun')
plt.ylabel('Jumlah Kejahatan')
plt.grid(True)
plt.xticks(rotation=45)
# Menambahkan anotasi data di titiknya
for i in range(len(annual_crimes_df)):
    plt.text(annual_crimes_df['Crime Year'][i], annual_crimes_df['count'][i],
             str(annual_crimes_df['count'][i]), color='red', ha="center", va="bottom", fontsize=10)
plt.tight_layout()
plt.savefig('annual_crimes_trend.png') # Menyimpan grafik
plt.close() # Menutup figure agar tidak ditampilkan

# 2. Visualisasi Jenis Kejahatan (Crime Types)
plt.figure(figsize=(10,6))
crime_type_crimes_sorted = crime_type_crimes_df.sort_values(by='count', ascending=False)
sns.barplot(x='count', y='Crime', data=crime_type_crimes_sorted, palette='viridis')
plt.title('Jumlah Kejahatan Berdasarkan Jenis')
plt.xlabel('Jumlah Kejahatan')
plt.ylabel('Jenis Kejahatan')
# Menambahkan anotasi pada setiap bar
for i in range(len(crime_type_crimes_sorted)):
    plt.text(crime_type_crimes_sorted['count'].iloc[i] + 5, i,
             str(crime_type_crimes_sorted['count'].iloc[i]), color='blue', ha="left", va="center", fontsize=10)
plt.tight_layout()
plt.savefig('crime_type_distribution.png') # Menyimpan grafik
plt.close() # Menutup figure agar tidak ditampilkan

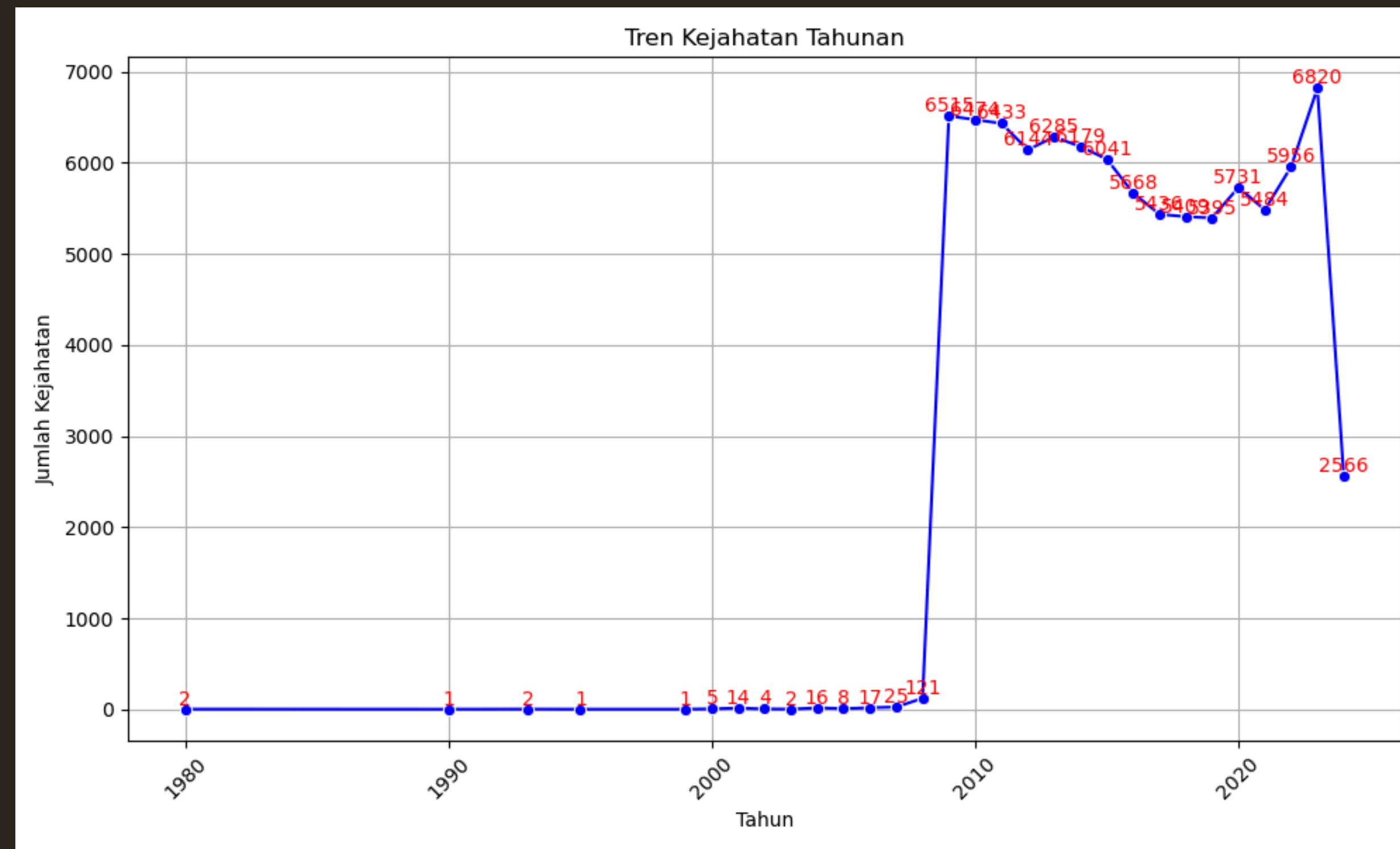
# 3. Visualisasi Kejahatan per Neighborhood (Neighborhood Crimes)
plt.figure(figsize=(10,6))
neighborhood_crimes_sorted = neighborhood_crimes_df.sort_values(by='count', ascending=False)
sns.barplot(x='count', y='Neighborhood', data=neighborhood_crimes_sorted, palette='coolwarm')
plt.title('Jumlah Kejahatan Berdasarkan Neighborhood')
plt.xlabel('Jumlah Kejahatan')
plt.ylabel('Neighborhood')
# Menambahkan anotasi pada setiap bar
for i in range(len(neighborhood_crimes_sorted)):
    plt.text(neighborhood_crimes_sorted['count'].iloc[i] + 5, i,
             str(neighborhood_crimes_sorted['count'].iloc[i]), color='green', ha="left", va="center", fontsize=10)
plt.tight_layout()
plt.savefig('neighborhood_crimes_distribution.png') # Menyimpan grafik
plt.close() # Menutup figure agar tidak ditampilkan
```

Memvisualisasi data kejahatan tahunan

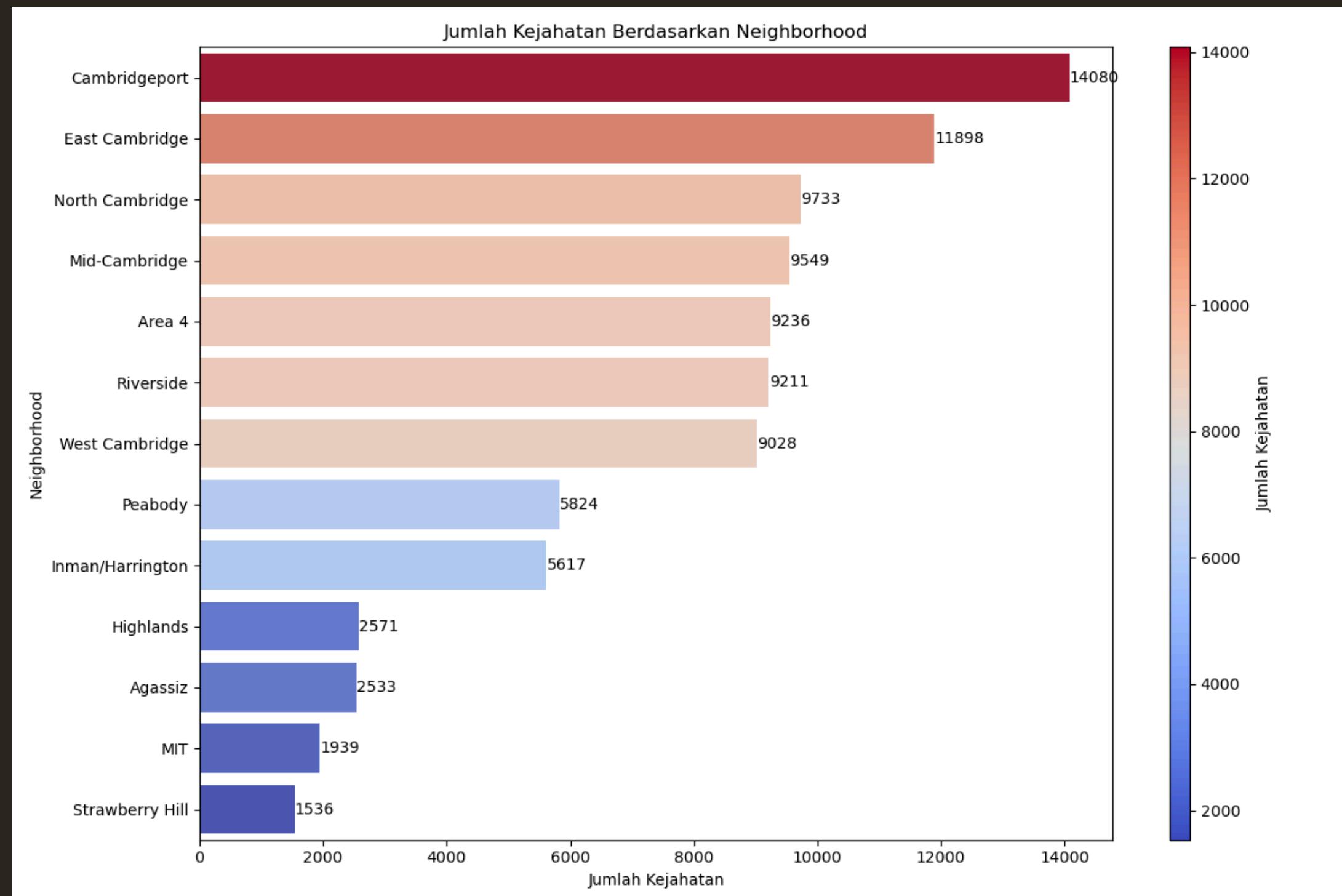
Memvisualisasi jenis kejahatan

Memvisualisasi kejahatan per neighbourhood

ANALISIS DATA



ANALISIS DATA



ANALISIS DATA

Berdasarkan pengolahan data, distribusi kejahatan di berbagai wilayah menunjukkan pola yang unik, dengan beberapa wilayah memiliki tingkat risiko kejahatan yang berbeda.

Potensi Kejahatan Tinggi

Wilayah seperti Cambridgeport dan East Cambridge tercatat memiliki angka kriminalitas yang sangat tinggi. Salah satu faktor utamanya adalah kepadatan penduduk yang tinggi, sehingga meningkatkan peluang konflik atau kejahatan

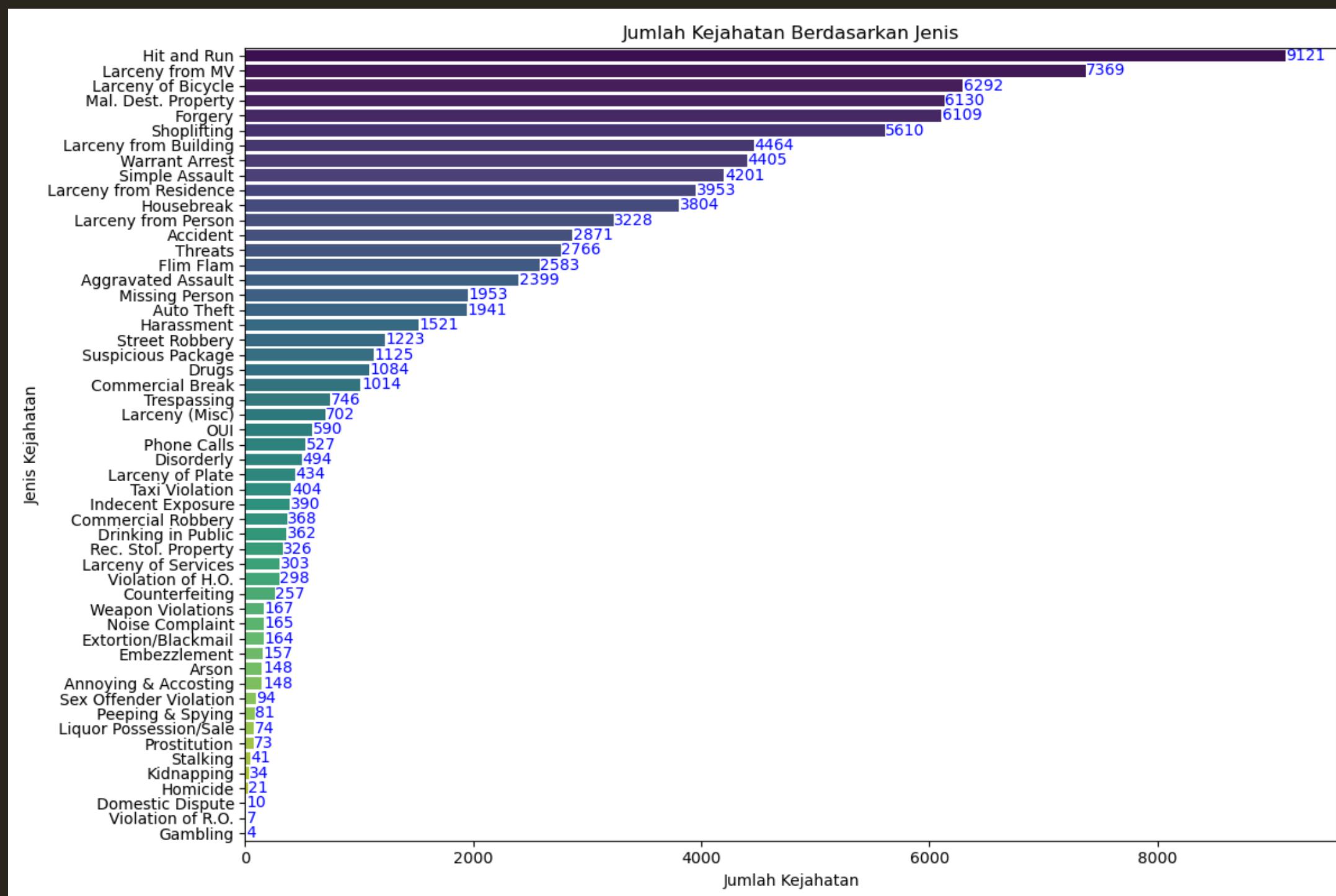
Potensi Kejahatan Rendah

Di lain sisi, MIT dan Strawberry Hill memiliki tingkat kejahatan yang rendah karena di MIT yang merupakan area akademik, sistem keamanannya cukup baik dan rendahnya populasi penduduk di wilayah Strawberry Hill menjadi faktor utama rendahnya angka kriminalitas.

ANALISIS DATA

Ditemukan juga pola unik kejahatan yang terjadi berdasarkan wilayah. Misalnya, pada wilayah Wiverside dan Mid-Cambridge, tingkat kejahatan yang relatif tinggi ini mencakup vandalisme atau pencurian properti kecil, karena adanya aktivitas sosial yang cukup tinggi. Sedangkan, pada wilayah Peabody dan Inman/Harrington, sering terjadi kejahatan bertenagat sedang. Hal ini mungkin saja disebabkan oleh adanya kombinasi karakteristik residensial dan komersial, sehingga menciptakan peluang kejahatan, meskipun tidak seintensif pada wilayah Cambridgeport.

ANALISIS DATA



TREN SOSIAL DAN KRIMINALITAS

Rendahnya Pelaporan Sebelum 2009

- Kurangnya kepercayaan pada sistem hukum mengurangi pelaporan insiden.
- Sistem pelaporan digital dan edukasi meningkatkan akurasi data.

Krisis Ekonomi & Lonjakan Kejahatan (2009)

- Tekanan finansial akibat krisis global memicu peningkatan kejahatan properti.
- Stabilitas ekonomi penting untuk mencegah kriminalitas di masa depan.

Teknologi Keamanan Canggih (2024)

- AI, kamera pengenal wajah, dan big data sukses menurunkan angka kejahatan.
- Deteksi dan pencegahan lebih efektif, menciptakan lingkungan yang lebih aman.

SOLUSI DAN REKOMENDASI

Pencegahan Kejahatan Oportunistik

- Memasang lampu terang di area publik dan lengkapi dengan kamera pengawas berbasis AI.
- Memberikan edukasi kepada masyarakat untuk lebih peduli menjaga keamanan diri dan barang pribadi.

Kebijakan Kejahatan Khusus

- Menerapkan hukuman tegas untuk pelanggaran seperti tabrak lari dan vandalisme.
- Mengadakan kampanye anti-pencurian bagi remaja, serta ajak bisnis dan pihak berwenang bekerja sama menggunakan teknologi anti-pencurian.

Strategi Berbasis Wilayah

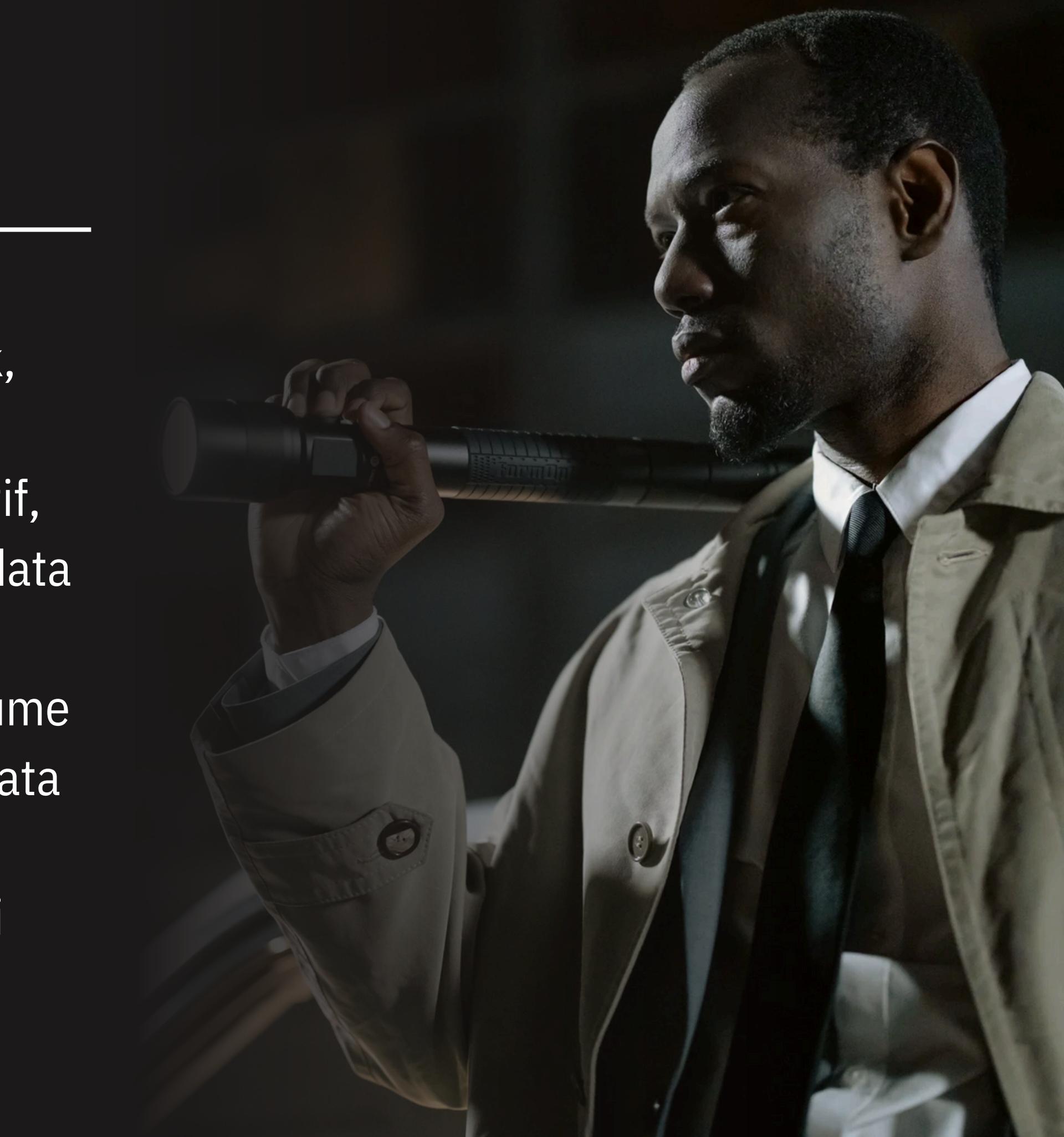
- Meningkatkan patroli polisi dan buat zona aman di daerah rawan, seperti Cambridgeport and East Cambridge.
- Melibatkan warga melalui program Neighborhood Watch untuk laporan cepat dan responsif.

PELUANG ANALISIS LANJUTAN

- Clustering Wilayah: Menggunakan algoritma seperti K-Means untuk mengelompokkan wilayah berdasarkan jenis kejahatan dan fokuskan kebijakan sesuai karakteristiknya.
- Prediksi Tren: Menganalisis data historis dengan ARIMA atau LSTM untuk memprediksi pola kejahatan dan alokasikan sumber daya secara efisien.
- Korelasi Sosio-Ekonomi: Menggabungkan data sosial dengan kriminalitas untuk mengidentifikasi akar masalah dan rancang program yang tepat sasaran.

KESIMPULAN

Dengan memanfaatkan Hadoop, Apache Spark, dan Matplotlib, analisis tingkat kriminalitas di Cambridge dapat dilakukan dengan lebih efektif, terutama dalam mengelola dan menganalisis data yang besar dan kompleks. Hadoop berperan penting dalam menyimpan dan mengelola volume data yang sangat besar. Serta dengan Spark, data ini dapat diproses dengan cepat untuk mengekstrak informasi yang relevan mengenai tren kejahatan berdasarkan lokasi dan waktu.





THANK YOU