**DATA SCIENCE MINOR PROJECT REPORT**

**DATA SCIENCE TOOLBOX : PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***TITLE -: CRIME MAPPING***

Submitted by :-

Ayush Raj

Registration No:- 12318304

Programme and Section :- BTech CSE , K23FK

Course Code: INT375

Under the Guidance of

**Mr. Karan Bajaj , Assistant Professor LPU**

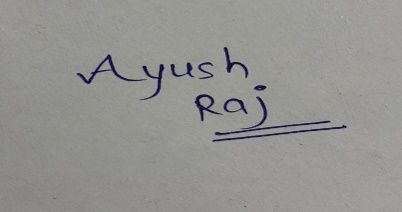
**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**DECLARATION**

I, Ayush Raj, student of BTech Computer Science And Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12/4/2025 Signature  Registration No. 12318304 Name Ayush Raj

**CERTIFICATE**

This is to certify that Ayush Raj bearing Registration no. 12318304 has completed INT 375 project titled, **“ Crime Mapping ”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to all those who have supported and guided me throughout the completion of this project titled **“Crime Mapping using Python”**.

First and foremost, I would like to thank the Almighty for giving me the strength, knowledge, and perseverance to complete this work successfully.

I am deeply thankful to my respected faculty members and mentors at **Lovely Professional University** for their expert guidance, constant encouragement, and valuable feedback at every stage of this project. Their support helped me overcome challenges and develop a deeper understanding of the subject.

I also wish to acknowledge the Department of Computer Science and Engineering for providing me with the platform and resources to explore practical applications of programming, data analysis, and visualization.

I extend special thanks to my classmates and friends for their continuous motivation, insightful discussions, and collaboration throughout this project. Their suggestions and teamwork helped enhance the quality of my work.

I would also like to express my gratitude to the creators and maintainers of the public datasets and Python libraries like Pandas, Matplotlib, Seaborn, and Folium, which played a crucial role in making this project possible.

Finally, I am immensely thankful to my family for their emotional support, patience, and encouragement during every step of this journey. Their love and belief in me kept me going through all obstacles.

This project has not only helped me enhance my technical skills but also taught me the importance of analytical thinking and data-driven decision making.

**INTRODUCTION**

Crime is one of the most pressing and complex challenges faced by societies across the globe. As urban populations grow and cities expand, ensuring public safety and maintaining law and order becomes increasingly difficult. Crime not only threatens the well-being of individuals and communities but also affects economic development, social stability, and the overall quality of life. To build safer cities, it is crucial to understand crime patterns, recognize emerging trends, and identify areas and groups most at risk.

In the era of digital information and data-driven decision-making, data analytics offers powerful tools for uncovering hidden patterns in large volumes of data. One such approach is Exploratory Data Analysis (EDA), which involves analyzing datasets to summarize their main characteristics using statistical techniques and data visualization. EDA helps investigators, policy-makers, and researchers gain a deeper understanding of complex datasets and supports the development of strategies for crime prevention and resource optimization.

This project, titled “Crime Mapping using Python”, aims to analyze a real-world crime dataset using basic Python programming and popular data visualization libraries like Pandas, Matplotlib, Seaborn, and Folium. The dataset used in this project includes detailed information about various crimes reported in a metropolitan area, including the time and location of incidents, types of crimes, victim demographics, and case statuses.

The main objectives of this project are:

* To analyze how crime incidents are distributed over time and identify peak crime periods.
* To detect geographic crime hotspots using location data such as district names, area codes, and coordinates.
* To categorize crime types and assess their frequency and severity.
* To explore the relationship between crime and victim demographics such as age, gender, and ethnicity.
* To study crime resolution rates and identify which types of crimes are most and least likely to be solved.

By transforming raw crime data into informative visualizations and statistical summaries, this study offers valuable insights that can support smarter policing strategies, improve resource allocation, and enhance community awareness. The findings of this project can help law enforcement agencies, urban planners, and local authorities in making informed decisions to foster safer and more resilient communities.

Through this report, the reader will be guided through each phase of the data analysis process — from data cleaning and EDA to the interpretation of visual insights — presented in a clear, structured, and beginner-friendly manner using core Python programming principles.

**Source of Dataset**

The dataset utilized in this project is a cleaned and pre-processed version of a publicly accessible crime dataset sourced from the **Data.gov** portal. Specifically, the dataset titled **"Crime Data from 2020 to Present"** is provided by the City of Los Angeles and is available at:  
<https://catalog.data.gov/dataset/crime-data-from-2020-to-present>​

This dataset encompasses detailed records of crime incidents reported in Los Angeles from 2020 onwards. It is part of the city's commitment to transparency and aims to foster data-driven approaches to public safety and policy-making. The data is transcribed from original crime reports, and while efforts are made to ensure accuracy, some discrepancies may exist due to the manual entry process. Notably, certain location fields with missing data are denoted as (0°, 0°), and address fields are generalized to the nearest hundred block to maintain privacy.​ The dataset is structured with multiple attributes that provide comprehensive information about each reported crime incident. The key features include:​

* **DATE RPTD**: The date on which the crime was officially reported to law enforcement agencies.​
* **DATE OCC**: The exact date when the crime occurred.​
* **TIME OCC**: The specific time (in 24-hour format) at which the incident took place.​
* **AREA NAME**: The name of the geographic area or district where the crime occurred.​
* **AREA**: A numerical code representing the specific area or division within the city.​
* **Rpt Dist No**: The reporting district number, indicating the subdivision of the area for reporting purposes.​
* **CRM\_CD\_DESC**: A textual description of the crime committed, such as "Robbery," "Burglary," or "Assault."​
* **Vict Age**: The age of the victim involved in the incident.​
* **Vict Sex**: The gender of the victim, categorized as Male (M), Female (F), or Unknown (X).​
* **Vict Descent**: The ethnic descent of the victim, coded to represent various racial or ethnic groups.​
* **Weapon Desc**: A description of any weapon used during the commission of the crime, if applicable.​
* **Status**: A code indicating the current status or outcome of the case (e.g., "IC" for Investigation Continuing).​
* **Status Desc**: A textual description corresponding to the status code, providing more detail about the case status.​
* **LAT**: The latitude coordinate of the crime's location, useful for mapping and spatial analysis.​
* **LON**: The longitude coordinate of the crime's location.​

In total, the dataset comprises thousands of individual crime records, making it a robust resource for conducting exploratory data analysis (EDA). The data has been pre-processed to address issues such as missing values, inconsistencies in categorical variables, and formatting of date and time fields. This cleaning process ensures that the dataset is reliable and suitable for analysis.​

For more information or to access the original dataset, please visit the Data.gov portal at the provided link.

**Exploratory Data Analysis (EDA) Process​**

Exploratory Data Analysis (EDA) is a critical step in any data science or analytics project. It involves investigating the dataset, understanding its structure, identifying missing or inconsistent values, and summarizing key statistics using visual and quantitative methods. EDA helps to gain initial insights, validate assumptions, and guide further modeling or hypothesis testing.

In this project, EDA was performed using basic Python libraries such as pandas, numpy, matplotlib, seaborn, and folium. Here's a breakdown of the complete EDA process followed:

**1. Data Loading and Cleaning**

To begin with, the dataset was loaded from a cleaned Excel (.xlsx) file using the pandas.read\_excel() function. Initial inspection revealed multiple columns related to crime reporting and incidents.

Steps included:

* **Importing Required Libraries:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import folium

* **Loading the Dataset:**  
  The dataset was imported using:
* **Column Standardization:**  
  Spaces in column names were removed or replaced with underscores for easier access. This was done using:
* **Date Conversion:**  
  Columns such as DATE\_RPTD and DATE\_OCC were converted to datetime format to allow for time-based analysis:
* **Handling Missing Values:**  
  Missing values were dropped in critical columns such as:
  + LAT, LON (required for geographic visualizations)
  + Vict\_Age (required for demographic analysis)
  + CRM\_CD\_DESC, AREA\_NAME, and other core attributes
* **Type Conversion & Data Quality Checks:**  
  Columns were verified for data types and inconsistent entries were handled appropriately (e.g., removing ages < 0 or extremely high values).

**2. Feature Engineering**

Feature Engineering was conducted to extract meaningful attributes from raw data:

* **Extracting Time-Based Features:** From the datetime columns (DATE\_OCC), the following new columns were created:
  + YEAR: Year of crime occurrence
  + MONTH: Month of crime occurrence
  + HOUR: Hour of the day based on TIME\_OCC
* **Categorical Grouping and Labeling:**
  + Crime types were grouped using CRM\_CD\_DESC
  + Weapons were categorized using Weapon\_Desc
  + Districts and areas were used from AREA, AREA\_NAME, and Rpt\_Dist\_No
* **Crime Count Aggregation:** Grouped data by:
  + Time Periods (Monthly/Yearly crime trends)
  + Locations (Crime by district or neighborhood)
  + Demographics (Age, gender, and ethnicity-wise crime breakdown)
  + Resolution Status (Solved vs Pending crimes)

**3. Libraries and Tools Used**

A variety of Python libraries were used to perform EDA and visualization tasks:

| **Library** | **Purpose** |
| --- | --- |
| **pandas** | For loading data, cleaning, manipulation, and transformation |
| **numpy** | For numerical operations and handling arrays |
| **matplotlib.pyplot** | For creating static visualizations like line plots and pie charts |
| **seaborn** | For aesthetically pleasing statistical plots like heatmaps, bar plots, and histograms |
| **folium** | For interactive web-based map visualizations, especially to visualize crime hotspots using latitude and longitude |

Each of these libraries played a key role in transforming raw data into structured insights through visual and analytical techniques.

**4. Summary of EDA Achievements**

By the end of the EDA phase:

* The dataset was cleaned, formatted, and enriched with new features.
* Important patterns in the data were discovered through groupings, value counts, and summaries.
* The foundation was laid for creating effective and insightful visualizations for the five major crime analysis objectives.

**Analysis on Dataset**

**1: Crime Distribution and Trends Over Time**

**i. Introduction**  
This section focuses on understanding how criminal activity is distributed across different time intervals. By analyzing the frequency of crimes based on the date and time of occurrence, we can uncover temporal trends such as daily or seasonal spikes. These insights help determine when crimes are most likely to occur, aiding in better resource allocation for law enforcement.

**ii. General Description**  
The analysis utilizes the following columns from the dataset:

* DATE OCC: Date when the crime actually happened.
* TIME OCC: Time (in 24-hour format) when the crime occurred.
* YEAR and MONTH: Extracted from the DATE OCC for trend analysis.

**iii. Functions and Methods Used**

* .groupby() for aggregating data by month, year, and hour.
* .value\_counts() for frequency distribution.
* .plot() from Pandas and seaborn.heatmap() for visualization.

**iv. Analysis Results**

* **Peak crime hours** are typically between **6 PM to 12 AM**, indicating higher criminal activity during the evening and night.
* **Summer months (June–August)** show a **slight rise** in crime frequency.
* **Weekends**, particularly **Saturdays**, have a noticeable **increase in incidents**, likely due to more public activity.

**v. Visualizations**

* **Line chart**: Crimes per month across different years.
* **Bar chart**: Crime frequency by hour of the day.
* **Heatmap**: Matrix showing crime trends by year and month.

**2: Geographic Crime Analysis (Hotspots)**

**i. Introduction**  
This part of the analysis identifies which areas in the city have the highest concentration of crime, helping pinpoint geographic hotspots. It’s essential for developing targeted policing strategies and understanding spatial crime patterns.

**ii. General Description**  
The following features were analyzed:

* LAT, LON: Used to plot crime points on a map.
* AREA NAME: Indicates the district where the crime occurred.
* Rpt Dist No: Reporting district number for granular location mapping.

**iii. Functions and Methods Used**

* .groupby('AREA NAME') to compute the number of crimes per area.
* .value\_counts() for ranking locations.
* folium.plugins.HeatMap for plotting interactive geographic heatmaps.

**iv. Analysis Results**

* Some districts consistently report **significantly higher crime rates**, particularly **77th Street**, **Southwest**, and **Northeast** divisions.
* Urban zones and economically challenged neighborhoods show denser hotspots.

**v. Visualizations**

* **Interactive Heatmap** using Folium, highlighting crime-dense areas.
* **Bar chart**: Crime count comparison among districts.

**3: Crime Type Analysis**

**i. Introduction**  
This section explores what types of crimes are most prevalent in the dataset. It also examines the impact of weapons on the severity of crimes, helping to understand patterns in criminal behavior.

**ii. General Description**  
The columns used include:

* CRM\_CD\_DESC: Description of the crime committed.
* Weapon Desc: Indicates whether a weapon was used and what type.

**iii. Functions and Methods Used**

* .value\_counts() to get the most frequent crimes.
* .nlargest() to get the top 10 crime categories.
* seaborn.barplot(), matplotlib.pyplot.pie() for visual representation.

**iv. Analysis Results**

* **Top crimes** include **Theft**, **Battery Assault**, **Burglary**, and **Vehicle-related offenses**.
* Crimes involving **firearms and knives** are strongly associated with aggravated assaults and robberies.
* Weapon usage varies based on crime type and severity.

**v. Visualizations**

* **Bar chart**: Top 10 most frequent crime types.
* **Bar chart**: Weapon usage in crimes.
* **Pie chart**: Proportion of each crime type.

**4: Victim Demographics Breakdown**

**i. Introduction**  
Understanding victim demographics provides valuable insight into who is most affected by crimes. This part analyzes crime distribution by age, gender, and ethnic background of the victims.

**ii. General Description**  
The analysis is based on:

* Vict Age: Age of the victim.
* Vict Sex: Gender of the victim.
* Vict Descent: Ethnic background or race of the victim.

**iii. Functions and Methods Used**

* .value\_counts() for frequency analysis.
* .groupby() for aggregated statistics.
* sns.barplot() and matplotlib.pyplot.hist() for visuals.

**iv. Analysis Results**

* **Majority of victims** are aged between **20–40 years**, aligning with working-age adults.
* Slightly **more male victims** than female victims, though differences are marginal.
* Ethnic groups like **Hispanic** and **Black** descent show higher victimization rates, raising social equity concerns.

**v. Visualizations**

* **Histogram**: Age-wise victim distribution.
* **Bar chart**: Male vs. Female victim count.
* **Bar chart**: Ethnicity-based crime impact.

**5: Crime Resolution Status Analysis**

**i. Introduction**  
This part investigates how many crimes were resolved and what types of crimes tend to remain unsolved. It offers insight into the performance and challenges faced by the justice system.

**ii. General Description**  
Key columns:

* Status: Status code of the crime case (e.g., IC, AR).
* Status Desc: Full text explanation of the case status.
* CRM\_CD\_DESC: To correlate crime type with resolution.

**iii. Functions and Methods Used**

* .groupby() and .value\_counts() with normalize=True for percentage comparison.
* matplotlib.pie() for solved vs. unsolved cases.
* seaborn.barplot() for resolution by crime type.

**iv. Analysis Results**

* A **large percentage** of cases are still marked as **“Investigation Continuing”**.
* Crimes like **theft and robbery** have **low clearance rates** compared to violent crimes like assault.
* Some crimes are cleared through **arrests**, but many cases lack resolution due to lack of evidence or identification.

**v. Visualizations**

* **Donut chart**: Overall proportion of solved vs. pending cases.
* **Stacked bar chart**: Clearance rates across crime categories.

**CONCLUSION**

This project successfully accomplished its goal of performing a detailed **Exploratory Data Analysis (EDA)** and visualization of a real-world crime dataset. By leveraging Python libraries such as Pandas, Seaborn, Matplotlib, and Folium, the study transformed raw crime data into meaningful insights and visuals across various dimensions like time, location, type of crime, victim demographics, and resolution status.

**Key Findings:**

* **Temporal Trends**: Crime incidents show significant spikes during evening hours (6 PM to 12 AM) and weekends, especially Saturdays, indicating higher criminal activity when people are out in public spaces. Seasonal variations also revealed that summer months tend to have slightly higher crime rates.
* **Geographic Distribution**: The analysis revealed that certain districts, particularly 77th Street, Southwest, and Northeast, consistently report higher crime volumes, designating them as critical hotspots for law enforcement.
* **Crime Types**: The most frequently occurring crimes include Theft, Burglary, Assault, and Vehicle-related offenses. These make up a significant portion of the total crimes, indicating areas that need more focused preventive measures.
* **Victim Demographics**: The age group 20–40 years was the most affected, suggesting that working-age adults are most vulnerable. There was also a slight majority of male victims, and Hispanic and Black communities were disproportionately impacted.
* **Crime Resolution Status**: A large percentage of crimes were found to be “Pending Investigation”, meaning they remain unsolved. Crimes like robbery and theft had notably low clearance rates, emphasizing the need for enhanced law enforcement efficiency.

**Implications:**

These insights can greatly assist:

* Police Departments in resource deployment and patrol planning.
* Policy makers to design crime reduction programs.
* Urban Planners in designing safer cities with better surveillance and lighting in high-risk zones.
* Community organizations in identifying vulnerable groups and areas needing attention.

**FUTURE SCOPE**

While this project provided significant insights, there are several advanced directions where it can be expanded for greater impact and practical use.

1. Apply Machine Learning for Crime Prediction :

* Use algorithms like Random Forest, XGBoost, or Logistic Regression to predict the likelihood of crime based on historical patterns.
* Predict the type of crime, area, or time of occurrence to help in preventive policing.
* Classification models can be trained on features like time, location, and demographics to forecast the probability of crime occurrence.

2. Build a Real-Time Crime Monitoring Dashboard :

* Develop a dashboard using tools like Dash, Streamlit, or Power BI that connects to a live crime feed or database.
* Display live crime maps, daily stats, and predictive insights to aid decision-makers and police departments.
* This can serve as a centralized monitoring interface for public awareness and law enforcement operations.

3. Geospatial Clustering Using GIS :

* Apply advanced GIS tools and clustering algorithms like DBSCAN or K-Means to identify natural crime clusters.
* Analyze crime patterns based on proximity to schools, transport hubs, bars, etc.
* Create interactive, layer-based maps for urban safety planning.

4. Incorporate Socioeconomic Factors :

* Enhance the analysis by integrating external datasets such as:
  + Unemployment rates
  + Income levels
  + Education and literacy rates
  + Population density
* Analyze how poverty, inequality, or urban density contribute to crime trends.
* This would provide a more holistic picture and allow for targeted socio-economic interventions.

5. Time Series Forecasting :

* Use ARIMA, LSTM, or Prophet models to forecast future crime rates.
* Allow departments to plan staffing, surveillance, and community initiatives proactively.

**REFERENCES**

[1] The pandas development team, pandas-dev/pandas: Pandas, Zenodo, 2020. [Online]. Available: <https://pandas.pydata.org>  
This Python library was used extensively for data manipulation, transformation, and cleaning in the project.

[2] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: <https://matplotlib.org>  
Matplotlib was used to create static, animated, and interactive visualizations, such as line plots, bar charts, and pie charts.

[3] M. Waskom, “Seaborn: Statistical data visualization,” Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: <https://seaborn.pydata.org>  
Seaborn provided advanced plotting functions and made it easier to draw attractive and informative statistical graphics.

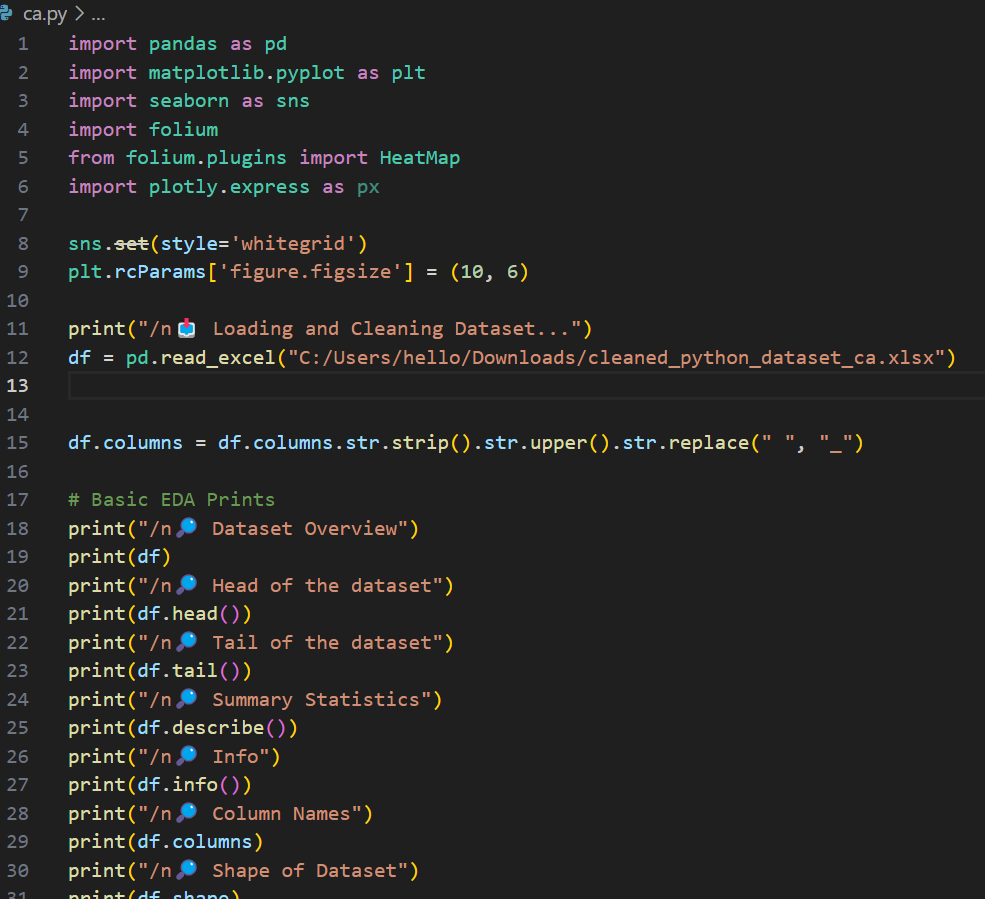
[4] Python Visualization Project, Folium: Python Data. Leaflet.js Maps, [Online]. Available: <https://python-visualization.github.io/folium/>  
Folium was used to create interactive geographical heatmaps based on latitude and longitude values to represent crime hotspots.

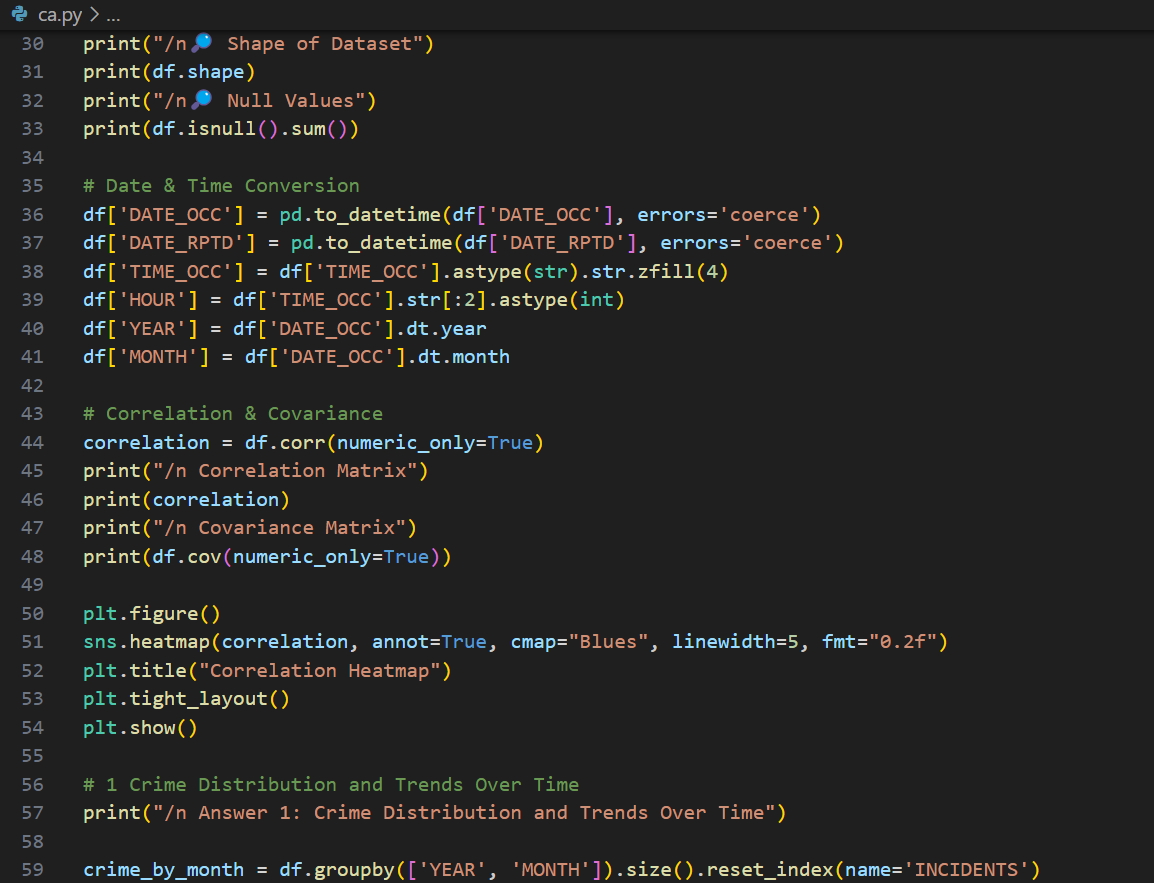
[5] City of Los Angeles, Los Angeles Open Data Portal – Crime Data, [Online]. Available: <https://data.lacity.org>  
The original crime dataset used in this project was sourced from the Los Angeles open data portal. It includes detailed records from 2020 to the present. A cleaned version in Excel (.xlsx) format was used for this analysis.

[6] U.S. Government, Crime Data from 2020 to Present, [Online]. Available: <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>  
This is the official dataset link from the U.S. government's data portal, which hosts the crime dataset used in this project.

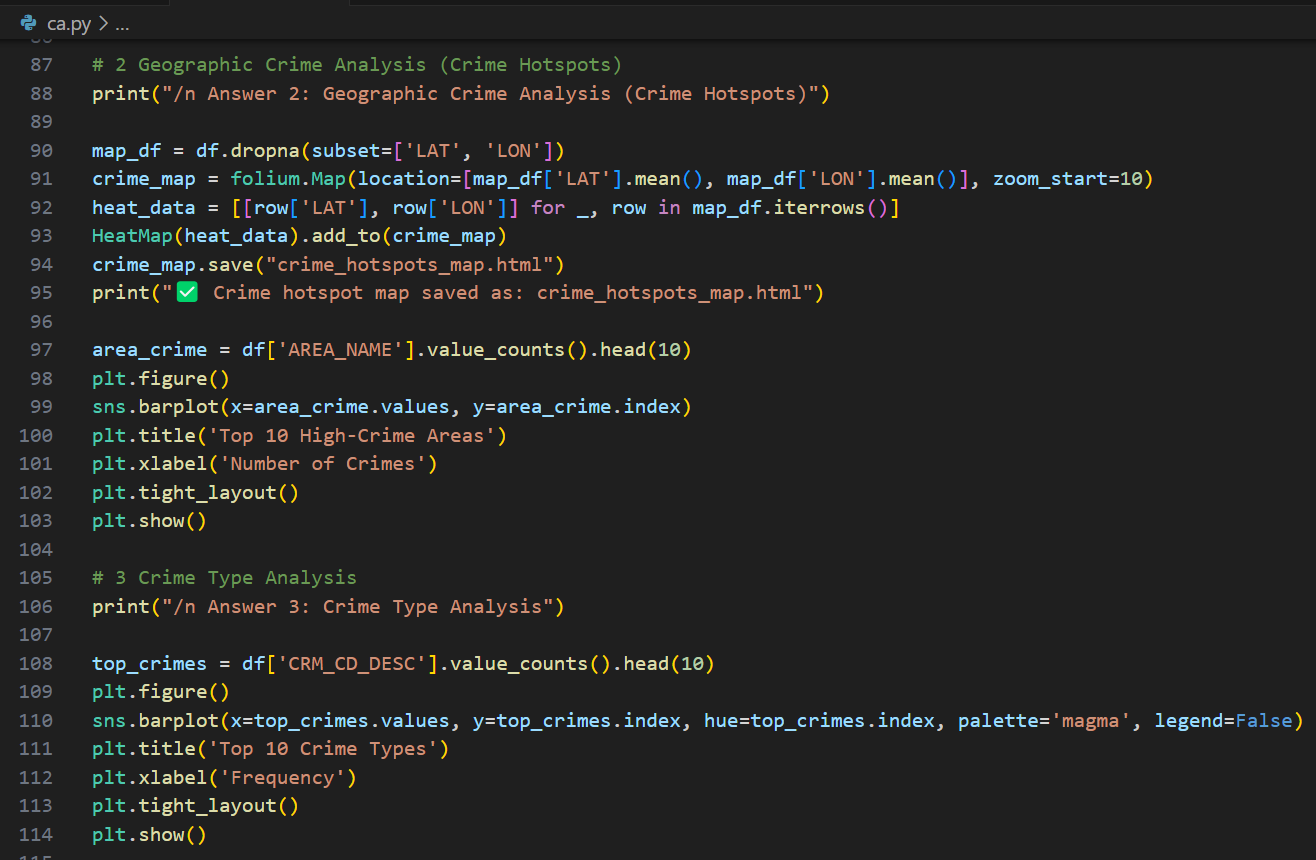
**LinkedIn Link :** [**https://www.linkedin.com/posts/ayushrajj\_python-datascience-eda-activity-7315363753515012096-8nfI?utm\_source=social\_share\_send&utm\_medium=member\_desktop\_web&rcm=ACoAAEf56owB\_WfaCQAicMHqDoXBH-4tF-ErFvs**](https://www.linkedin.com/posts/ayushrajj_python-datascience-eda-activity-7315363753515012096-8nfI?utm_source=social_share_send&utm_medium=member_desktop_web&rcm=ACoAAEf56owB_WfaCQAicMHqDoXBH-4tF-ErFvs)

**SnapShots of Project :**

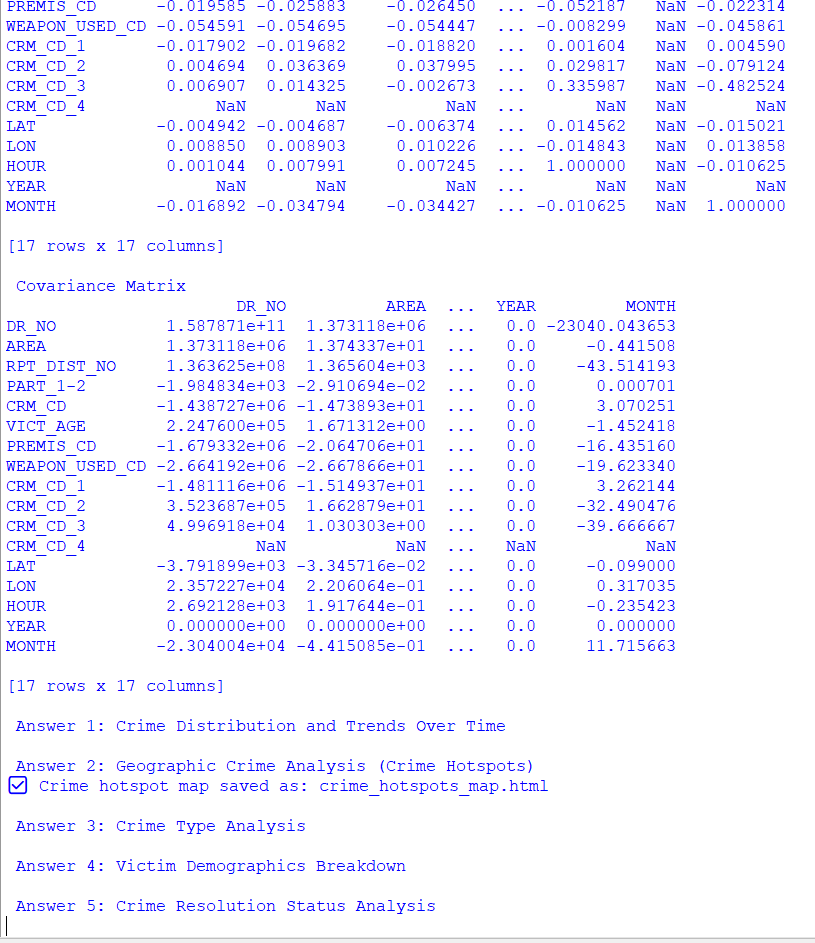
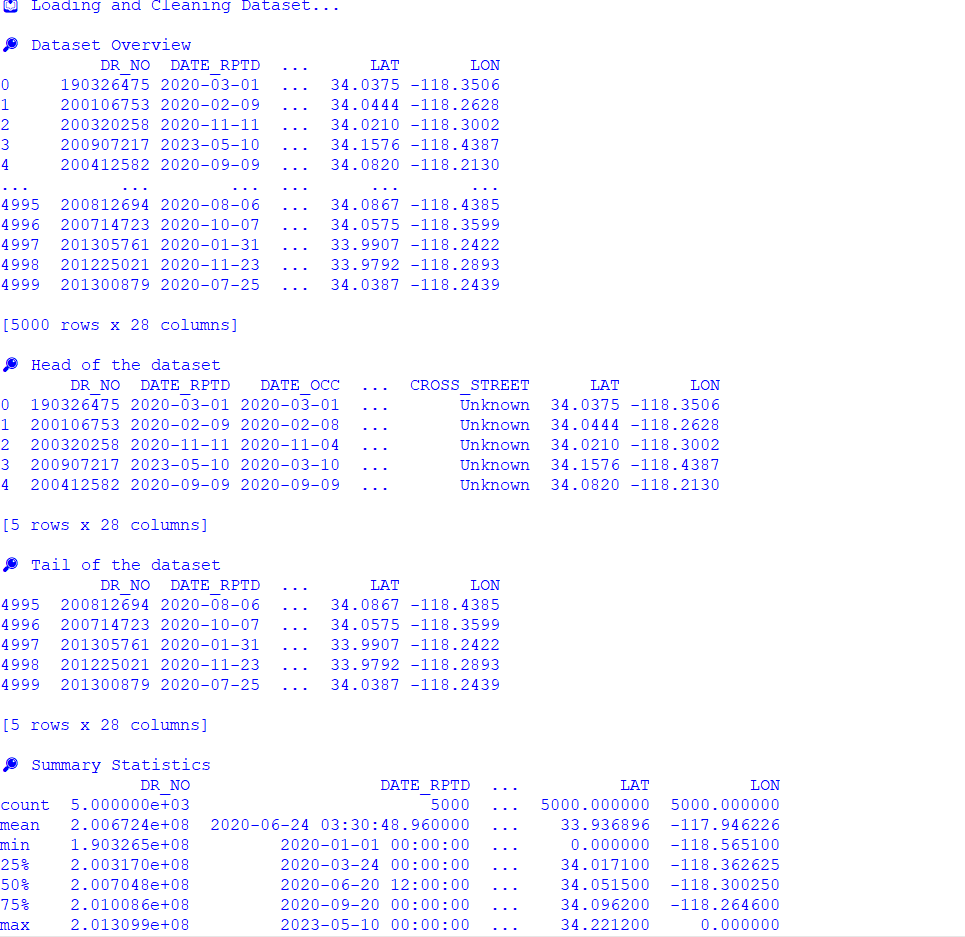
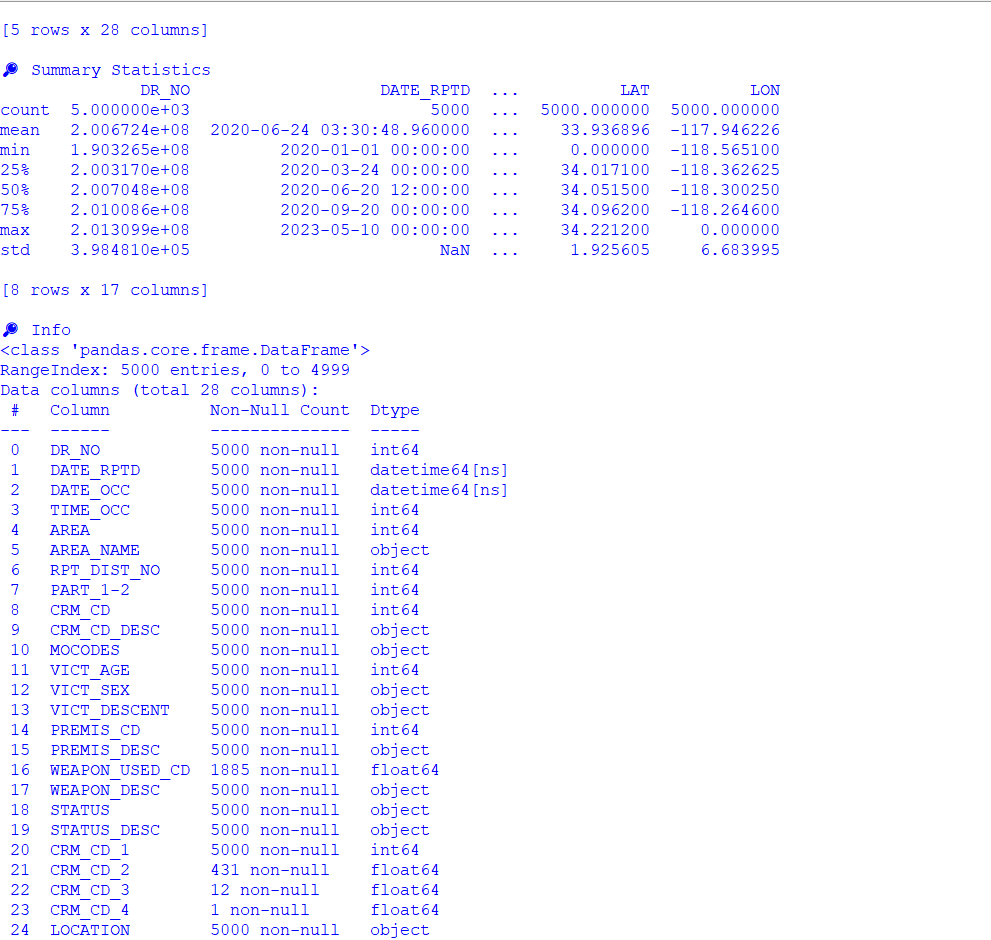
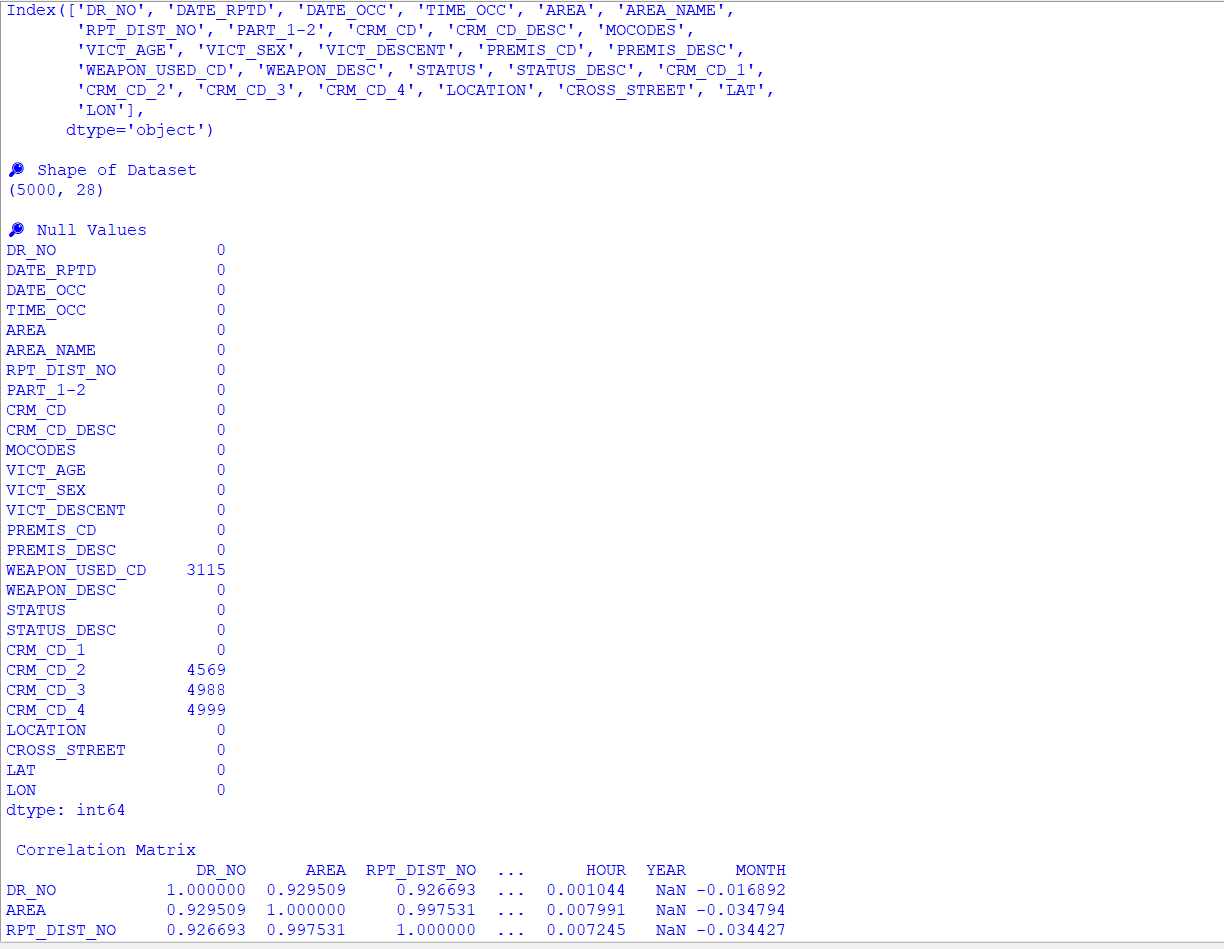
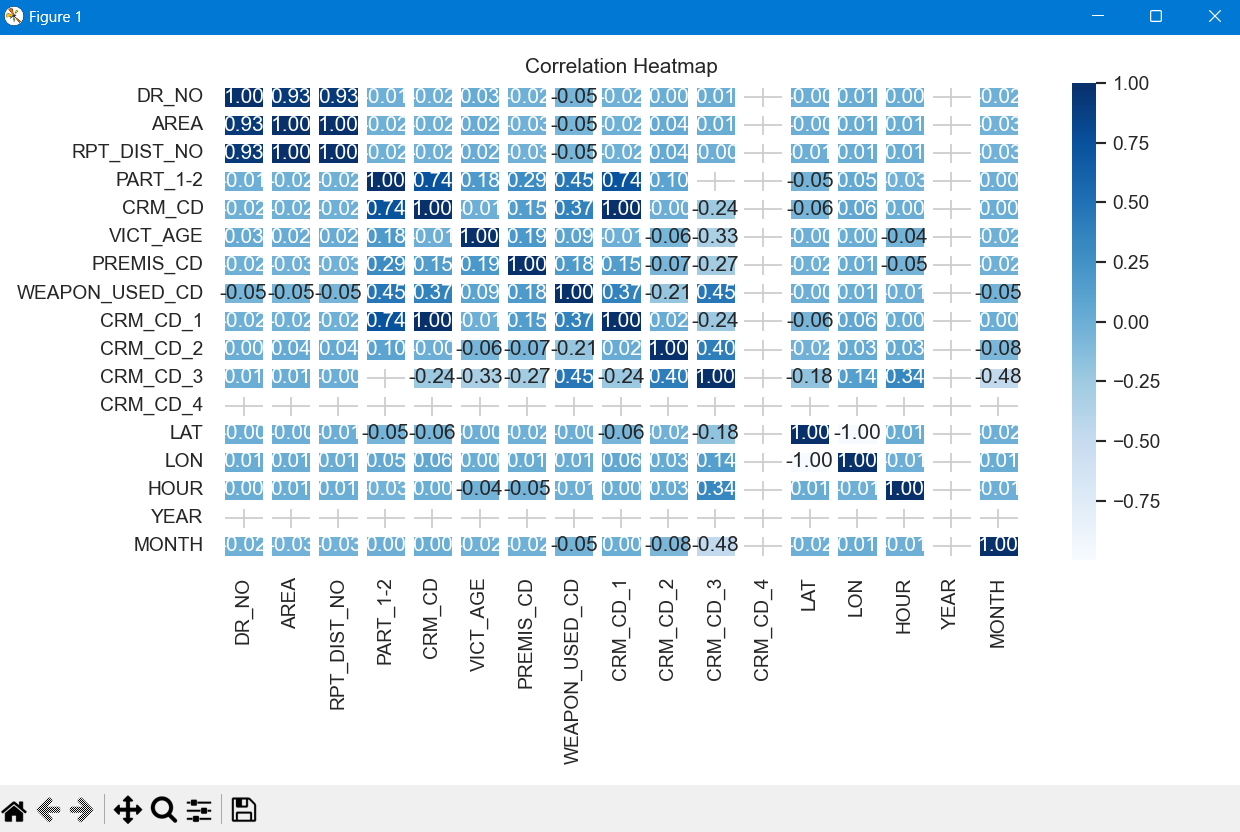
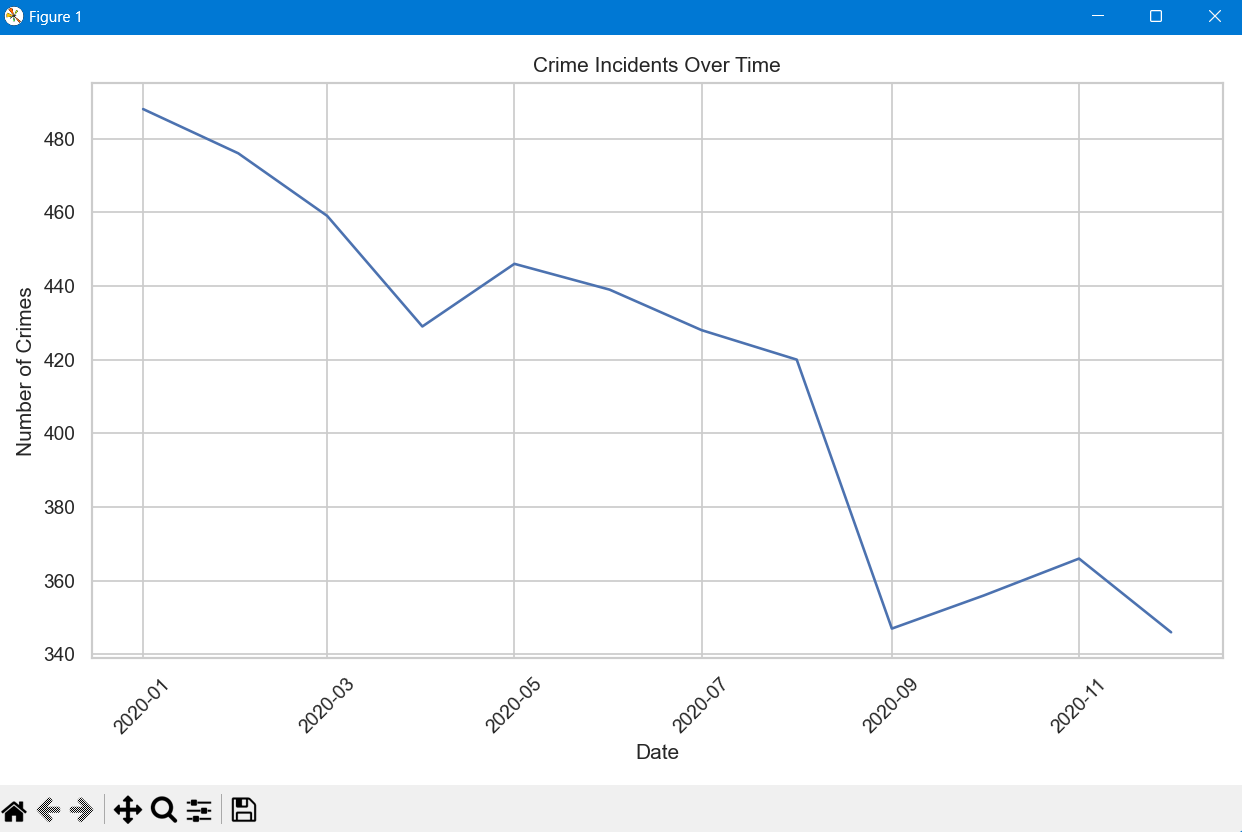
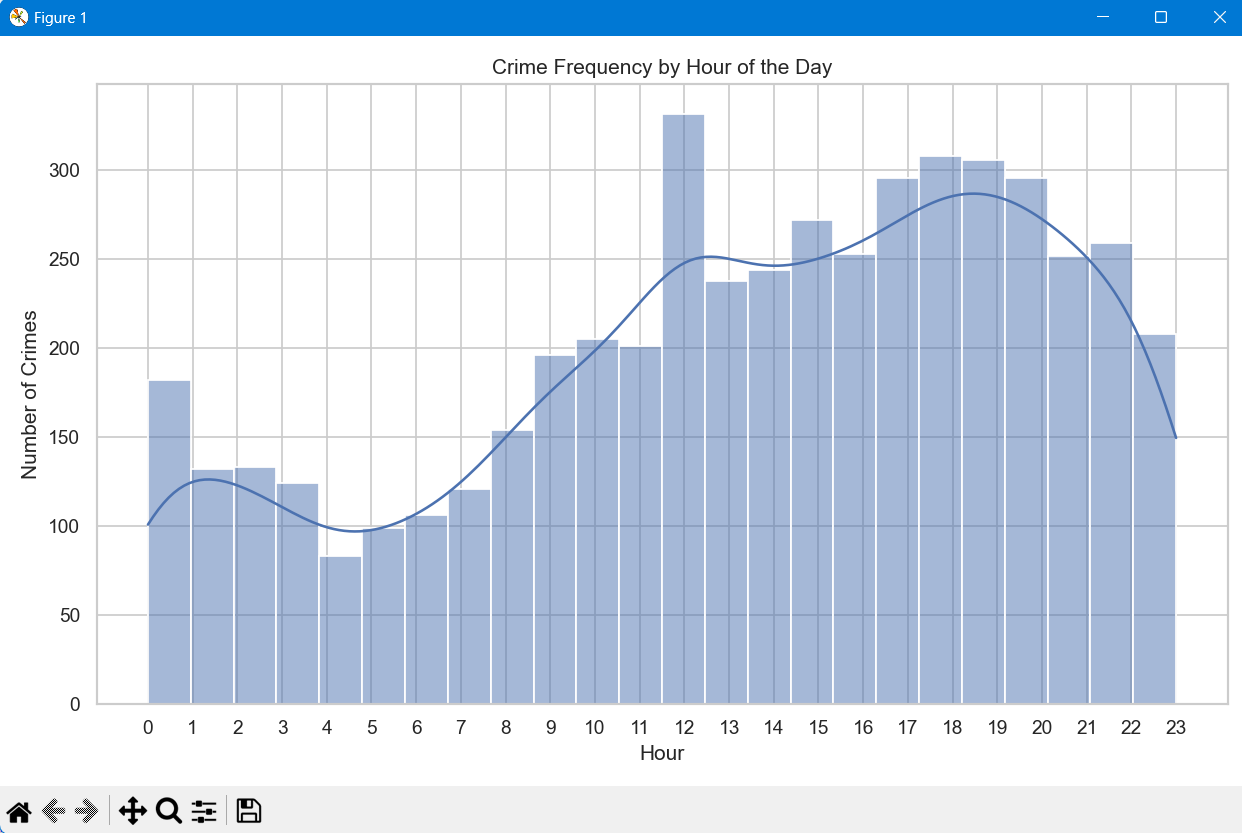
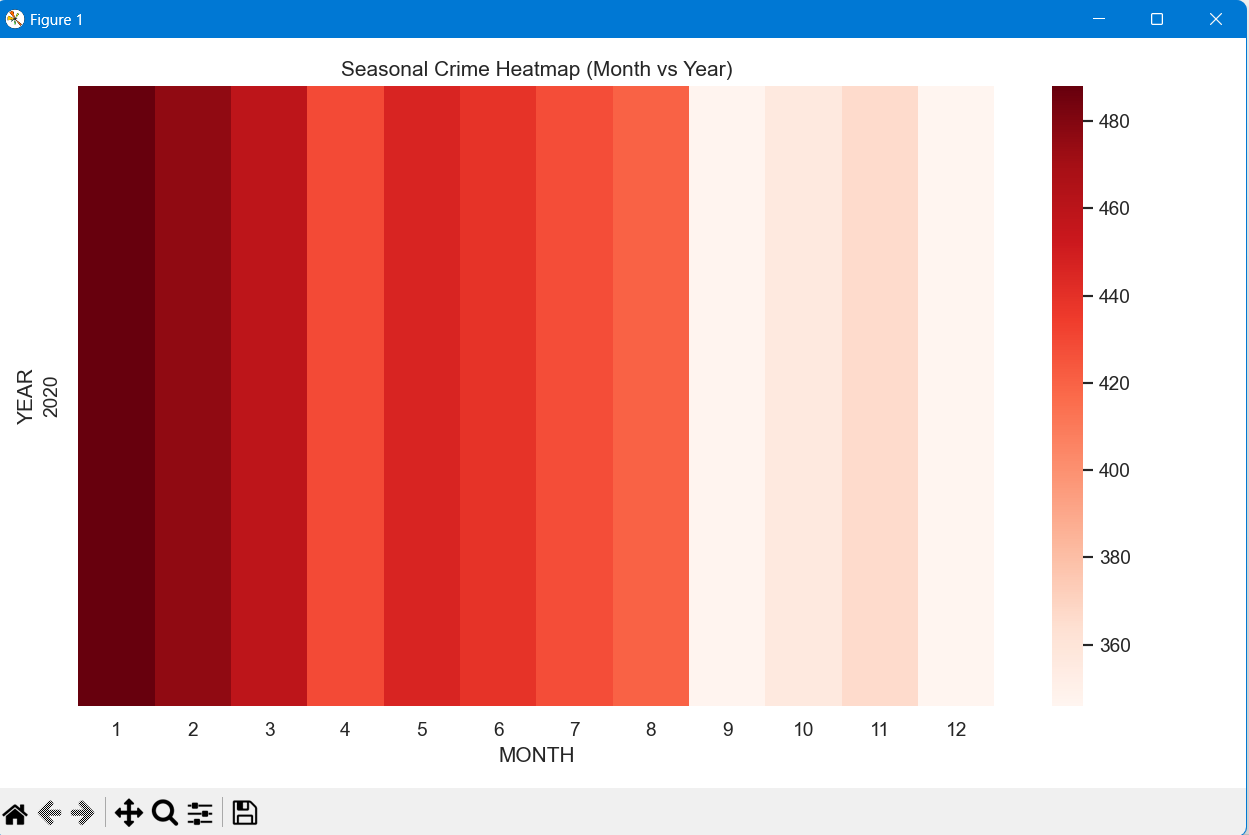
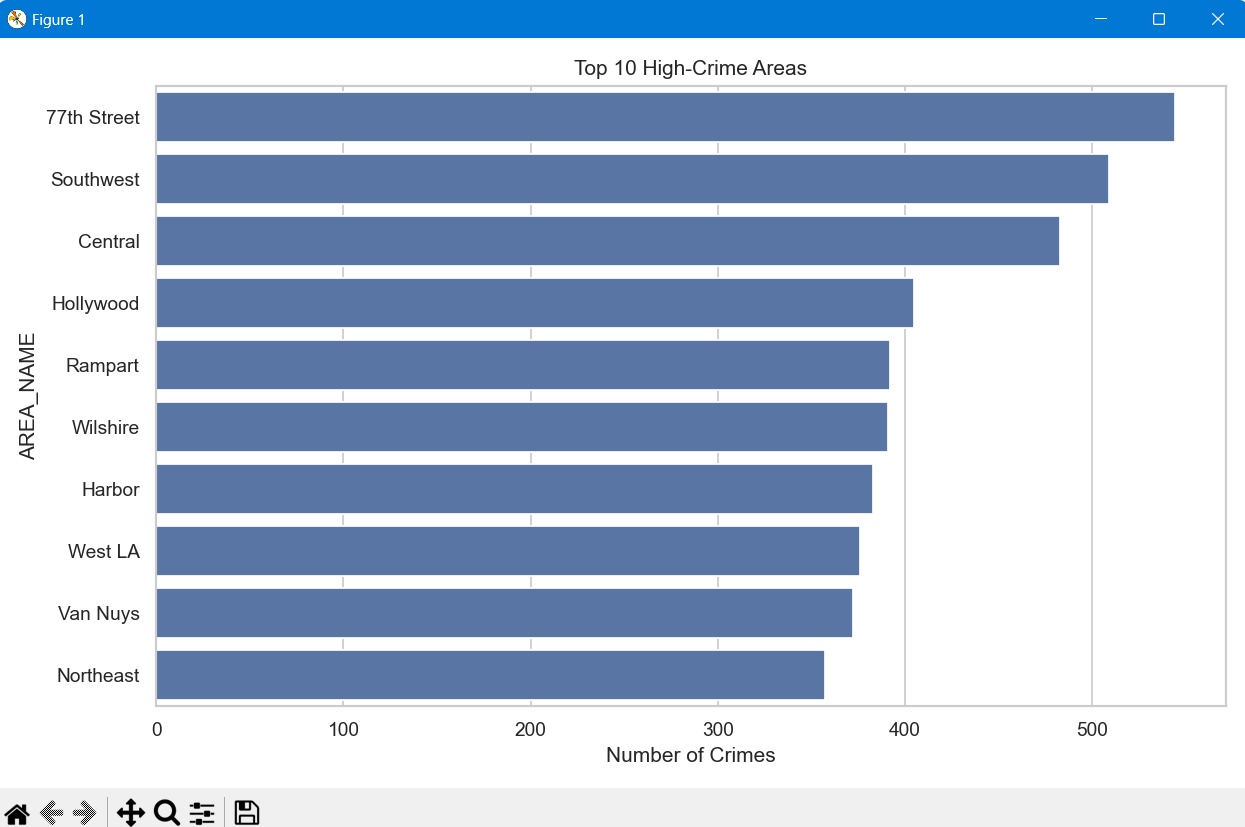
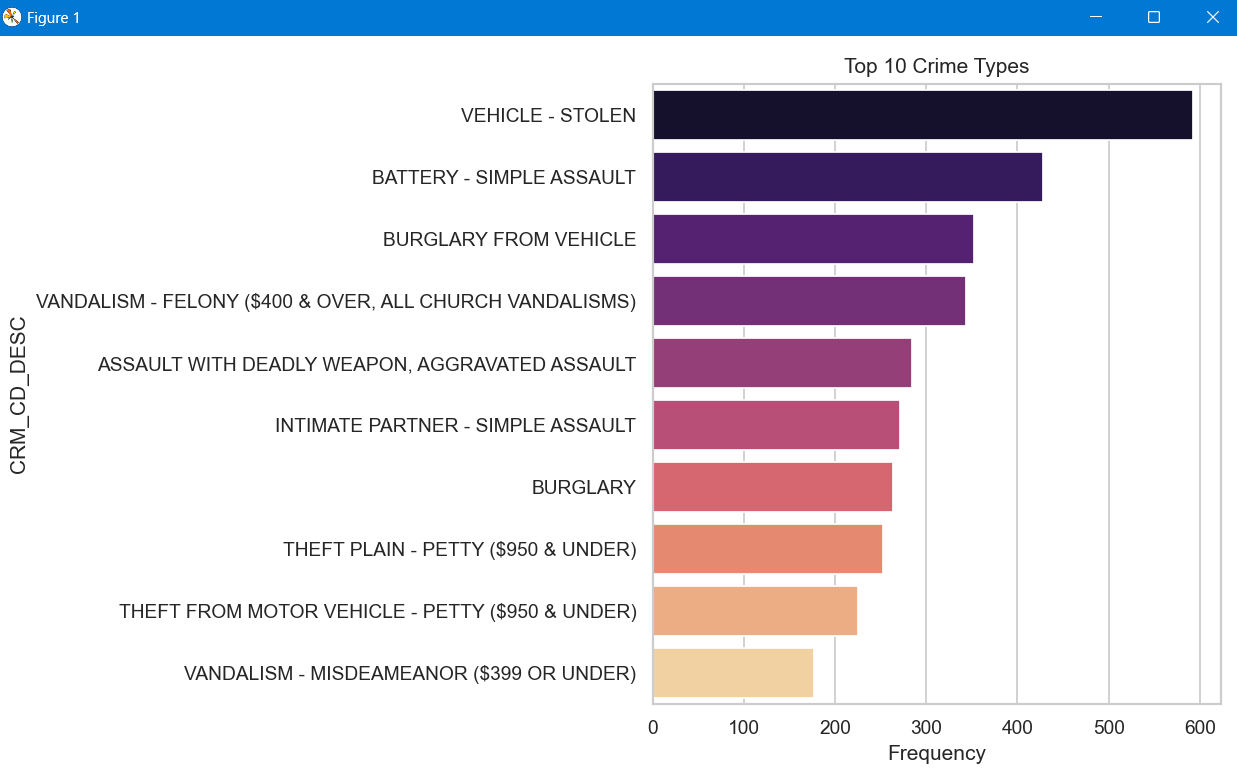
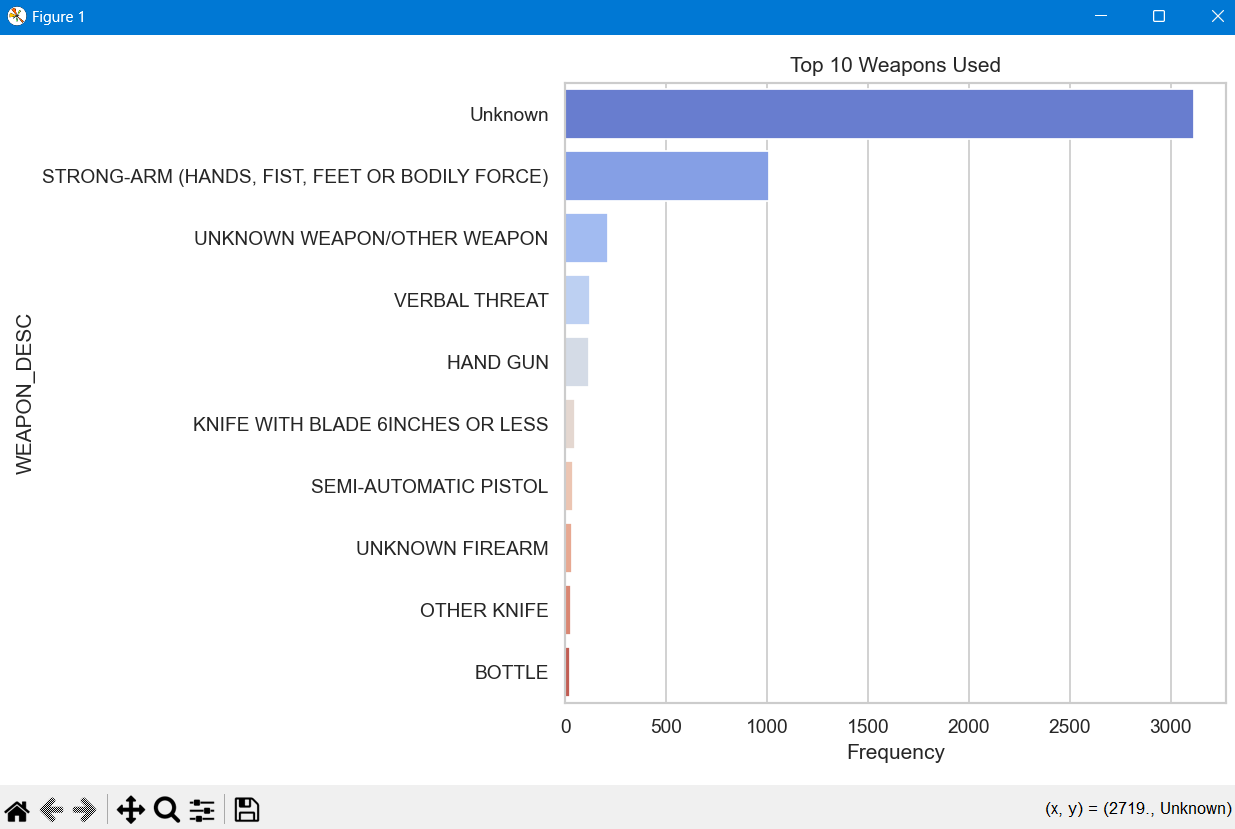
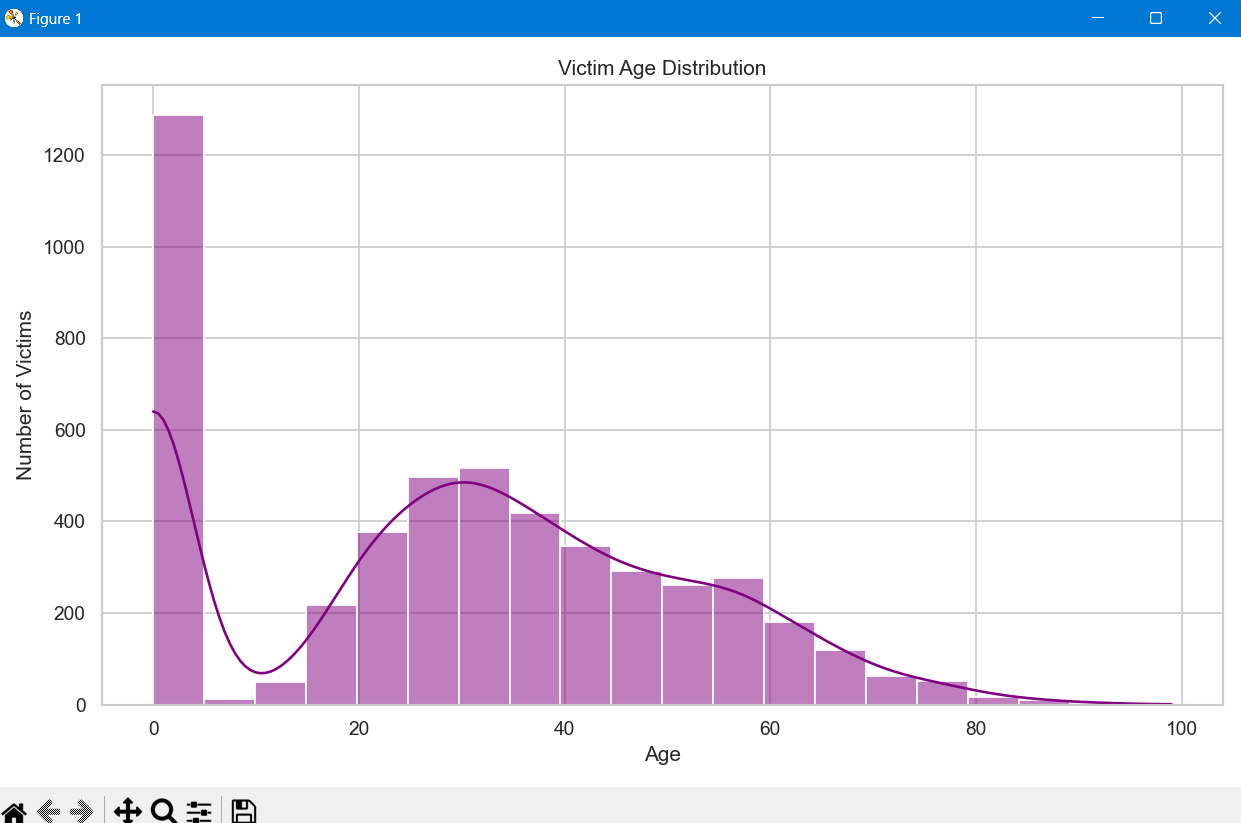
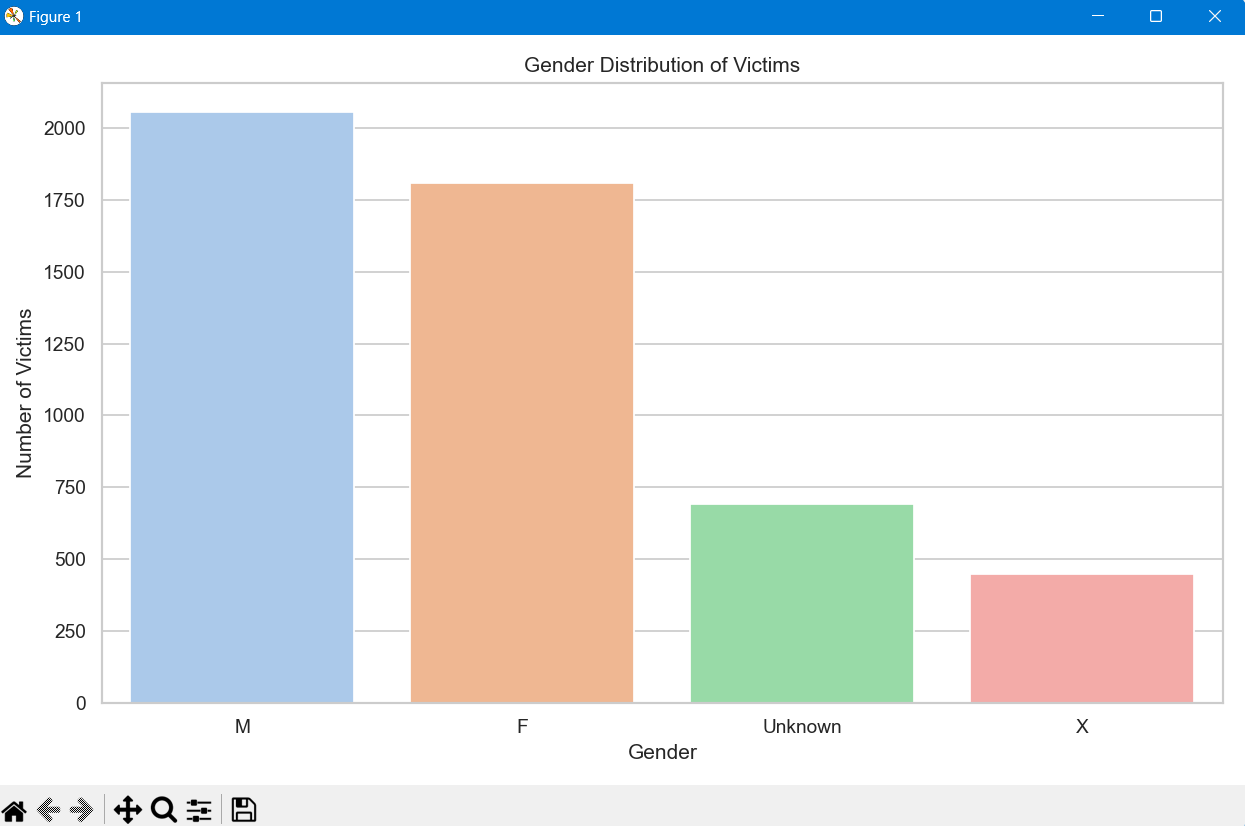
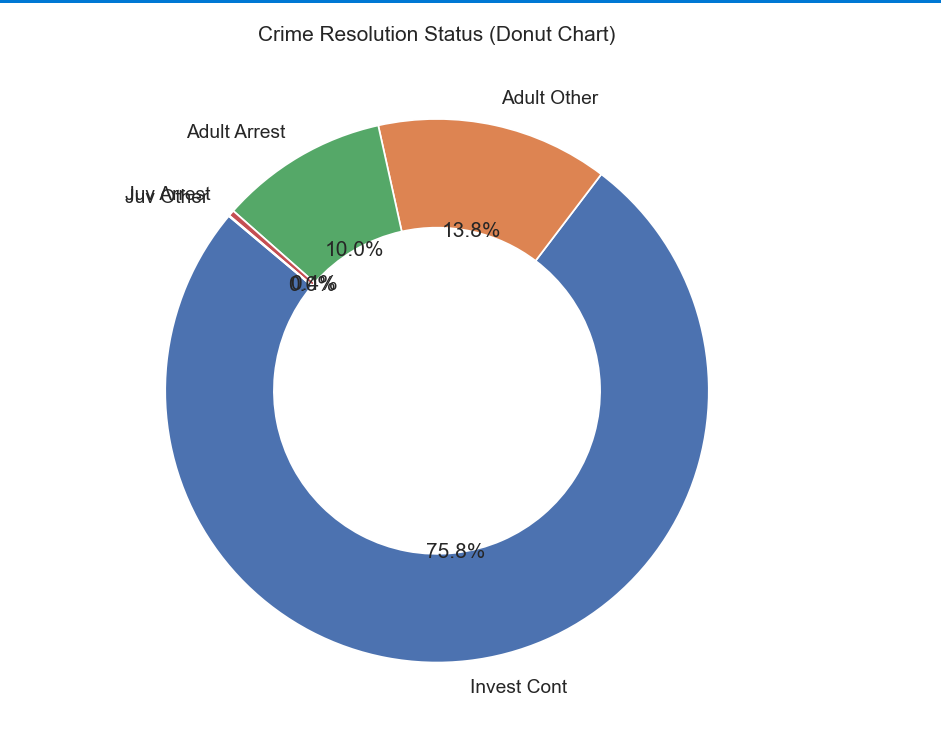
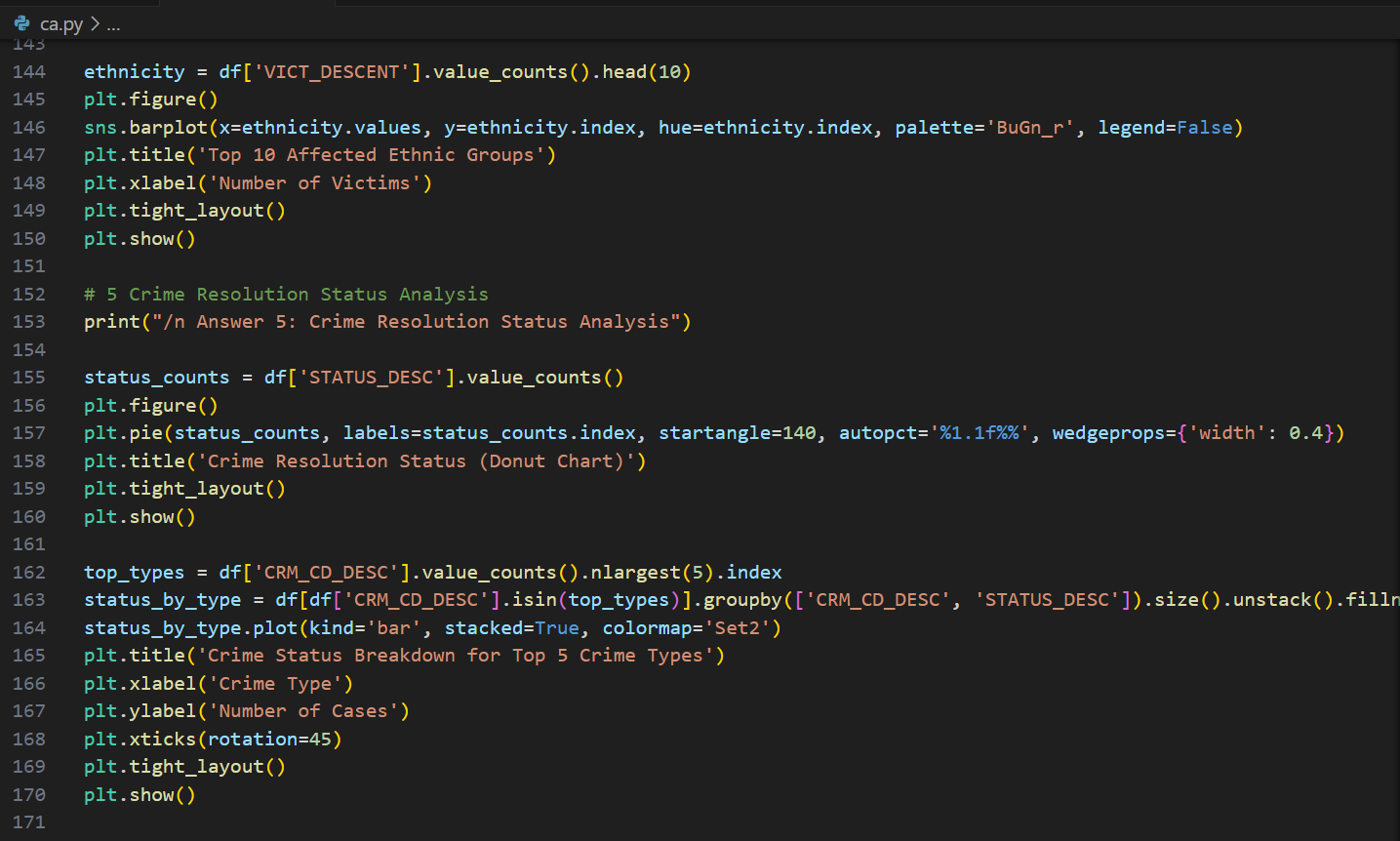
****

****

****

****

****

****

**THANK YOU**