DATA SCIENCE TOOLBOX : PYTHON PROGRAMMING

PROJECT REPORT

Table of Content

- Cover page 2

- Certificate 3

- Declaration 4

- Acknowledgement 5

1. Introduction 6

2. Source of dataset 7

3. EDA process 8

4. Analysis on dataset (for each objective) 9

1. General Description
2. Specific Requirements
3. Analysis results
4. Visualization

5. Conclusion 27

6. Future scope 28

7. References 29

DATA SCIENCE TOOLBOX : PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Crime Insights: Data-Driven Crime Analysis

Submitted by

Himanshi Chaudhari

Registration No 12318164

Programme and Section K23GN

Course Code INT375

Under the Guidance of

Mrs. Aashima (UID : 28968)

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

**CERTIFICATE**

This is to certify that Himanshi Chaudhari bearing Registration no. 12318164 has completed

INT 375 project titled, **“**Crime Insights: Data-Driven Crime Analysis **”** 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.

**Mrs. Ashima**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Himanshi Chaudhari, student of Data Science 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: 11-04-2025 Signature

Registration No. 12318164 Himanshi Chaudhari

**Acknowledgement**

I would like to express my heartfelt gratitude to *Mr./Ms Ashima* , my project guide,

for their continuous guidance, helpful suggestions, and motivation throughout this project. Their support played a crucial role in the successful completion of this work.

I also thank *Lovely Professional University* and the Department of Computer Science for providing the required resources and an encouraging environment.

I’m grateful to my friends and classmates for their constant encouragement, helpful insights, and collaborative spirit during the development of this project.

Finally, I sincerely thank my family for being my constant source of motivation and for their unwavering support during this journey.

**INTRODUCTION**

Crime is an ever-growing concern in modern society, affecting

communities, governance, and public safety. In the era of data

driven decision-making, analysing crime data can offer critical

insights that help in understanding patterns, identifying hotspots,

and predicting future incidents. This project focuses on

exploring and analysing a real-world crime dataset using Python

and its data science libraries.

By applying techniques such as Exploratory Data Analysis (EDA),

data visualization, this project aims to uncover hidden trends,

frequency of various crime types, temporal patterns (like which

time of year/day crime peaks), and geographical distribution of

incidents.

The dataset used in this project was sourced from a public catalog

platform and contains detailed records of reported crimes, including

attributes like crime type, location, date, and more. Using tools

such as Pandas, Matplotlib, Seaborn, and Scikit-learn, this analysis

aims not only to understand the dataset, but also to provide

insights that could help policymakers, police departments,

and the general public.

The ultimate goal of this project is to bridge the gap between raw

crime data and actionable insights using the power of Python

based analytics

Crime Insights: Data-Driven Crime Analysis

(2020-2025)

* Name – Himanshi Chaudhari
* Reg – no – 12318164
* Roll-no- 58
* Section – K23GN
* In This project I have covered almost every point of python libraries including NumPy pandas matplotlib and seaborn
* The Website from which I have taken this dataset is -- [https://catalog.data.gov/dataset/crime-](https://catalog.data.gov/dataset/crime-data-from-2020-to-present) [data-from-2020-to-present](https://catalog.data.gov/dataset/crime-data-from-2020-to-present)
* This project is based on the crime dataset between the years 2020 to 2025

**EDA Process**

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, often using visual methods. In this project, EDA was performed to better understand the structure of the crime dataset, detect patterns, identify outliers, and discover relationships between variables.

**🔹 1. Dataset Overview**

Checked the number of rows and columns

Printed the first few rows using df.head()

Understood column names and data types using df.info() and df.describe().

**🔹 2. Missing Value Analysis**

Identified missing values using df.isnull().sum()

Decided how to handle them: dropped rows / filled with mean, mode, etc.

**🔹 3. Univariate Analysis**

Analyzed individual columns like crime\_type using value\_counts() and bar plots

Visualized with histograms, bar charts.

**🔹 4. Bivariate / Multivariate Analysis**

Looked at relationships between variables (e.g., crime type vs. location)

Used groupby, scatter plots, heatmaps.

**🔹 5. Outlier Detection**

Used boxplots and standard deviation methods to spot any outliers

**🔹 6. Trend Analysis (if time series is involved)**

Analyzed how crimes changed over months or years

Used line plots for time-based insights

**Visualizations Used:**

Mention a few examples:

Bar charts to show crime counts

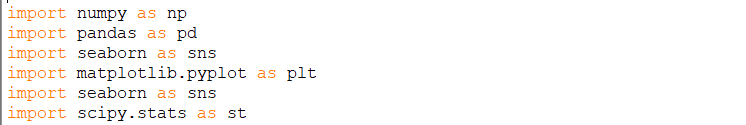
Heatmaps for correlation

Line charts for time-based trends

Boxplots for outlier detection

EDA helped in gaining a clear understanding of the dataset and prepared the foundation for further modeling and analysis. These insights also guided the feature selection and helped in identifying potential areas for deeper investigation

* Importing the warnings and python libraries in idle python --



* 2. Importing the data set



* 3. Overview of the data set

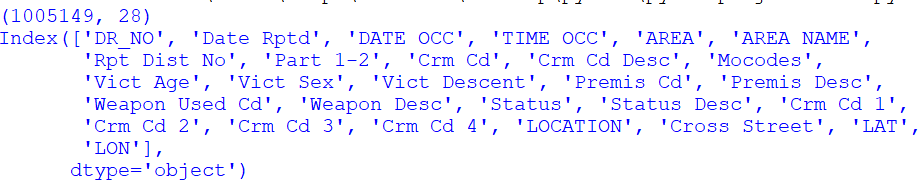
. check the dimension of the data set for that we have use shape attribute

Ans – The output of the code is the



check the columns of the dataset for that I used attribute

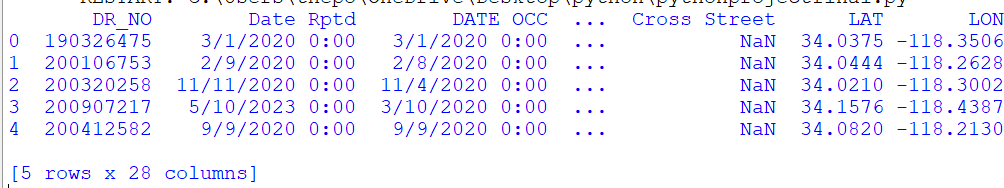




. check the top 5 row of the dataset

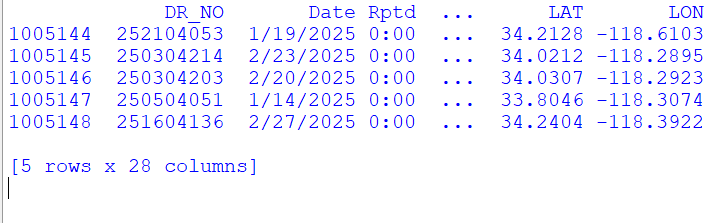


Ans – The output of the code is



. check the list 5 rows of the dataset

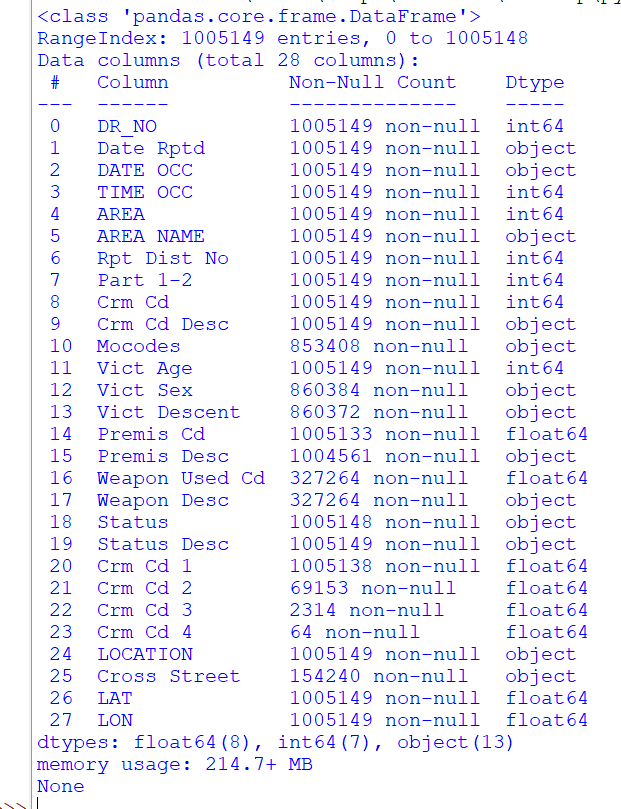


Ans – The output of the code is

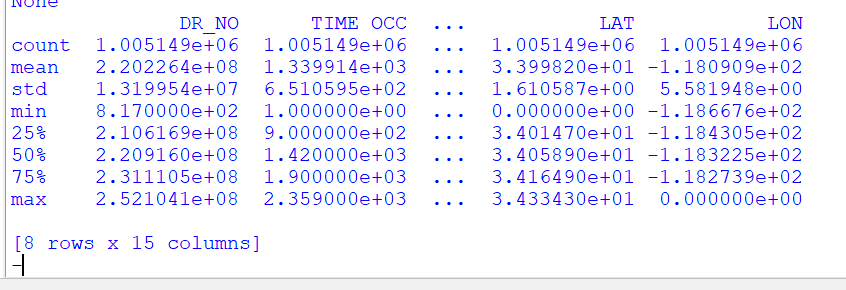
. checking all the information of the dataset and details then we use info function



Ans – The output of the code is



. checking for the describe method it will give you the summary of the invention

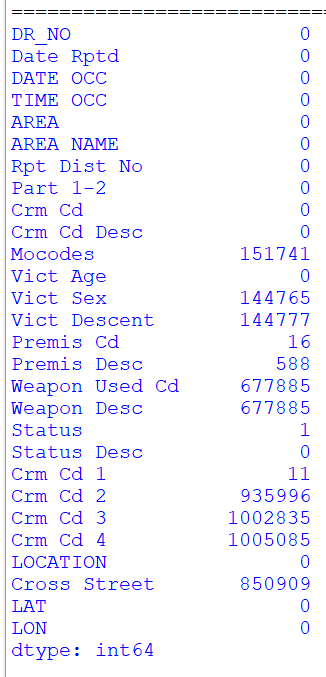
Ans -

* 4. **Check for anomalies in the dataset**

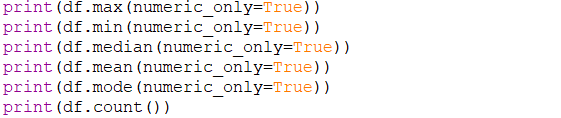
. check for missing numeric values Check for the missing number in the dataset and their sum



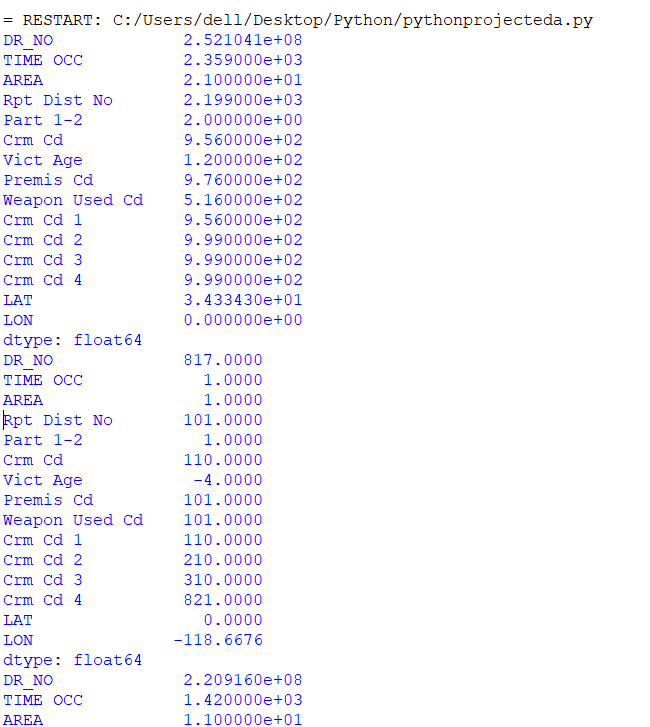
Ans – The output is

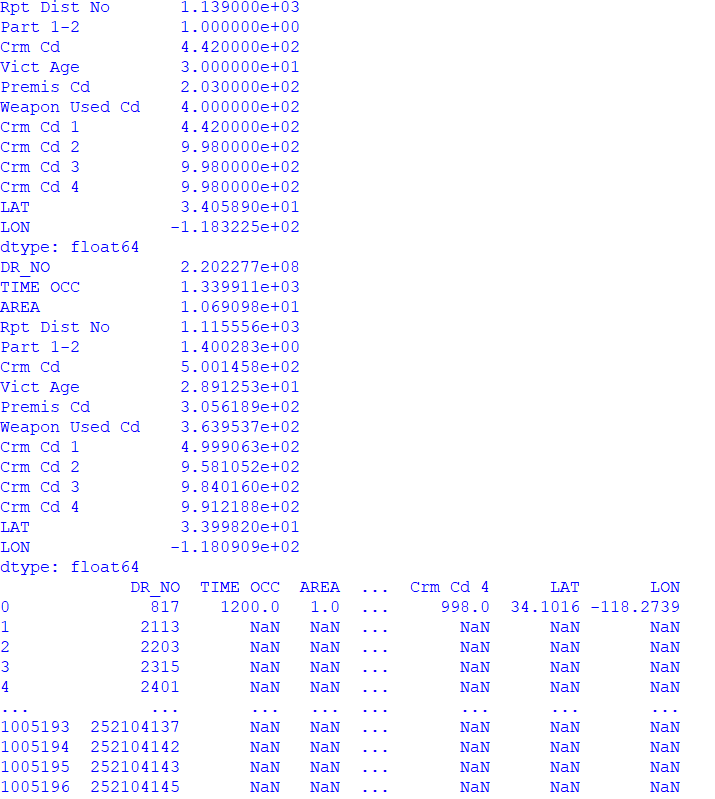


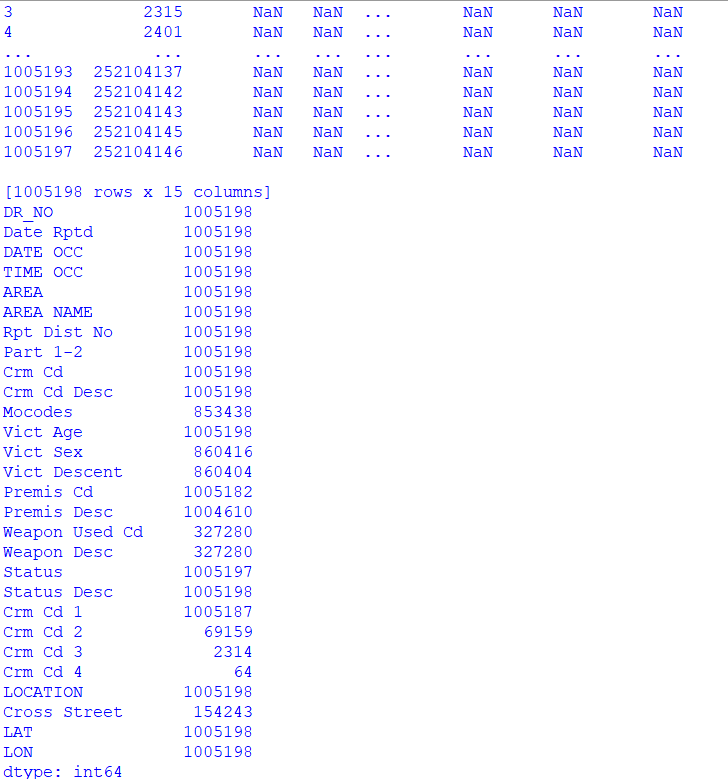
* 5. Checking for the max value min values median mode count and sum in one pic



Ans - The output of the code is





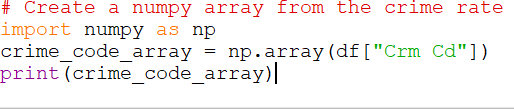


* A close up of a word  AI-generated content may be incorrect.6. Checking for the cleaning of the dataset

Ans – The output of the code is the

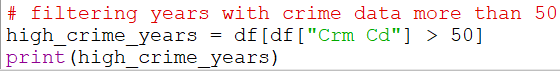


# CREATION OF NUMPY ARRAY 1.



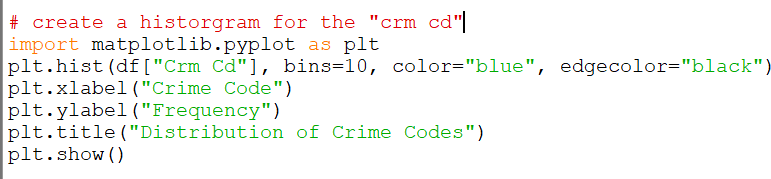
Ans – The output of the code is



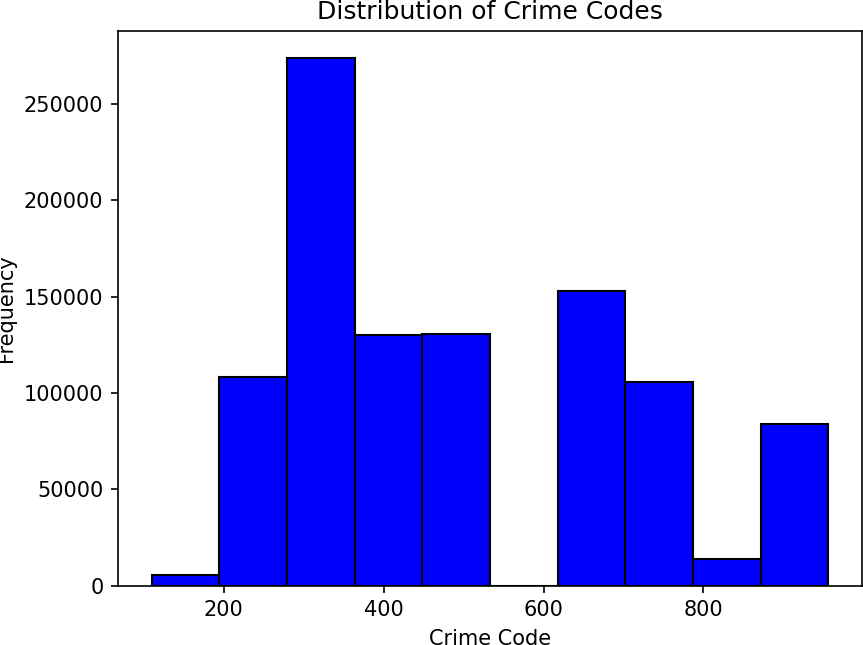
2. filtering years with crime data more than 50

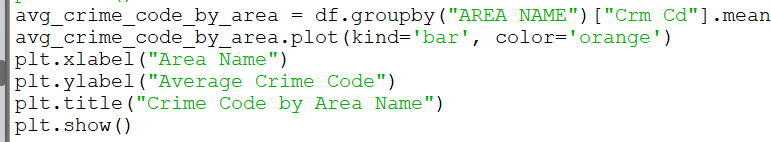
# HISTOGRAM

1. creating a histogram based on crime data

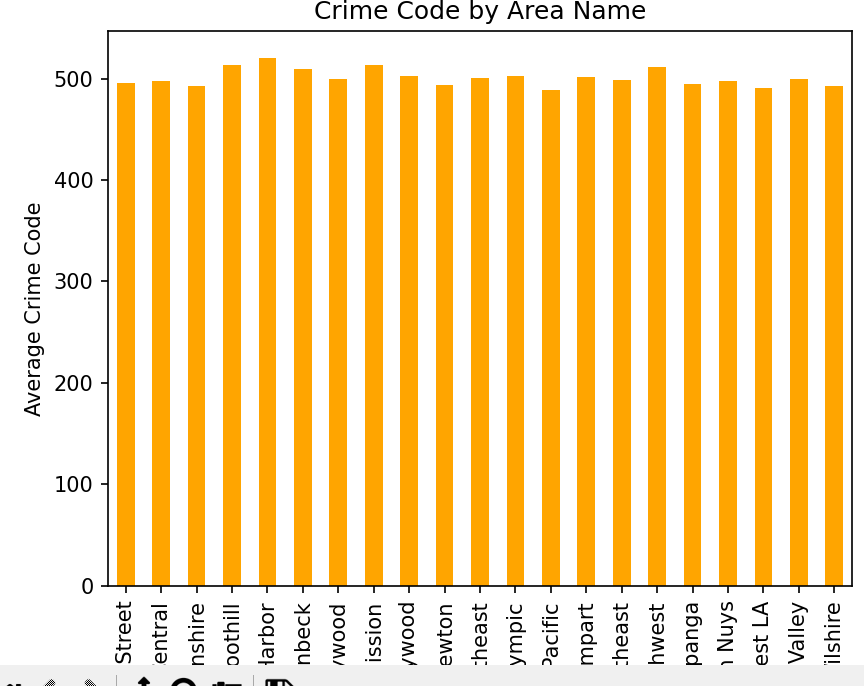


Ans – The output of the code is

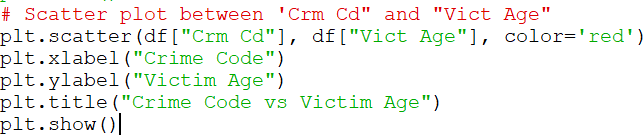


1. creating a Bar chart in which Bar chart comparing “ Crm Cd” across “Area Name”

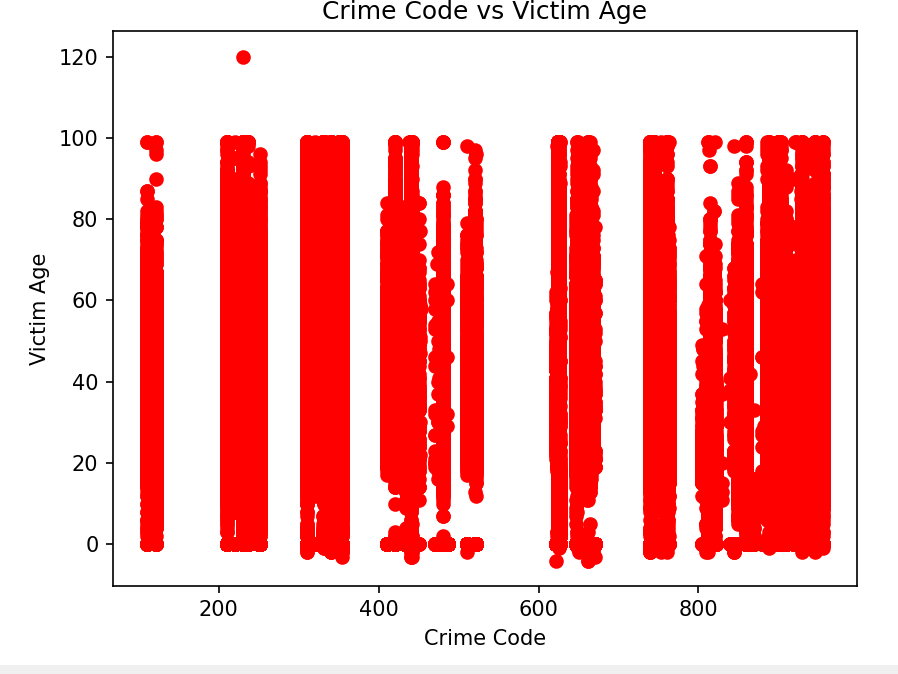
Ans – The output of the code is



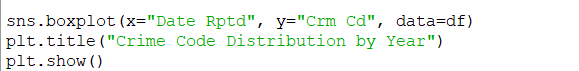
1. creating a scatter plot between “CRM Cd” and “Vict Age”:\



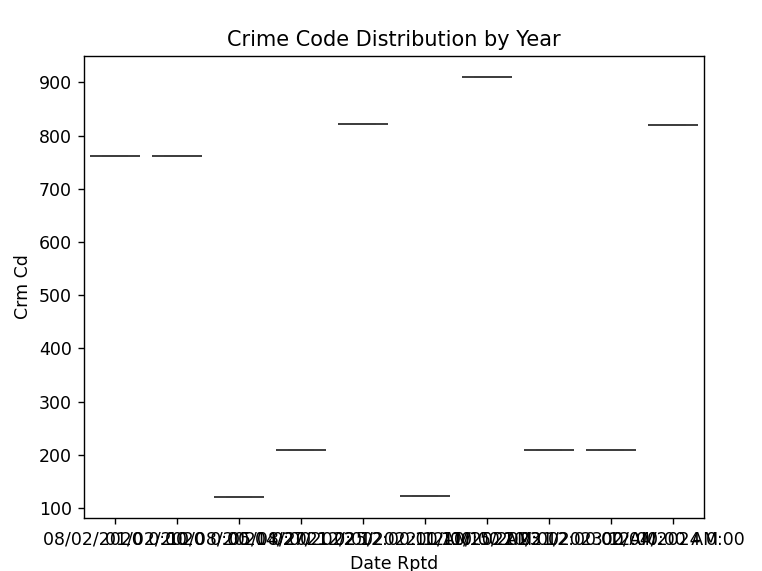
Ans – The output of the code is



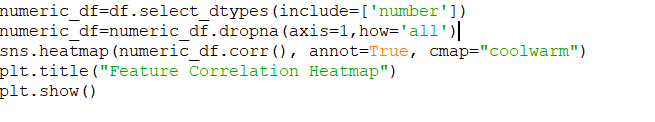
4.Crate boxplot Date Rptd and Crm cd



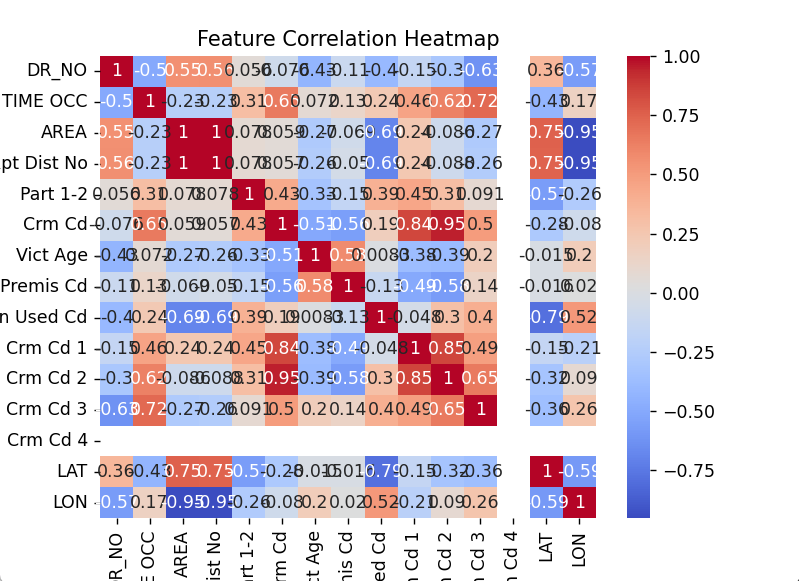
Ans - Output of the code



5-Creating a heatmap to visualize the correlation



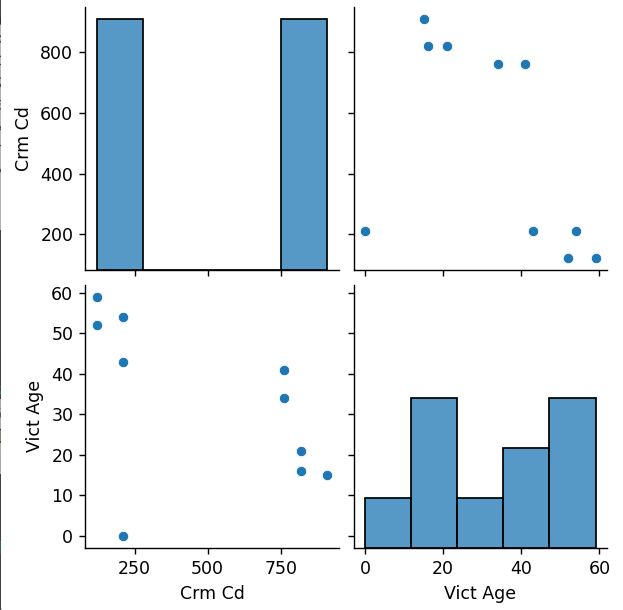
Output of the program-



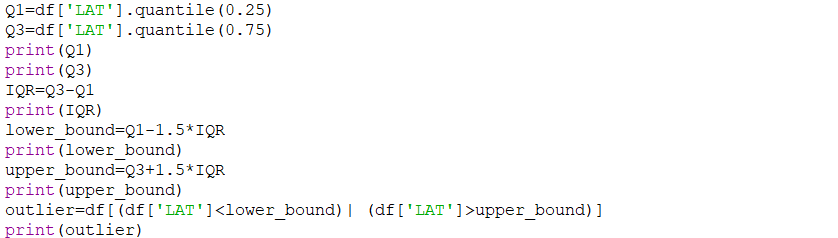
Create Pair plot for selected columns



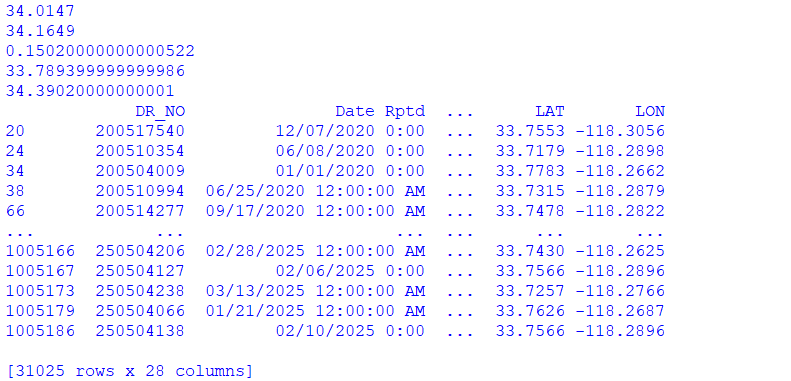
Output of the code is-



Finding outlier



Output-



BELOW I HAVE WRITTEN ALL THE CODE WHICH I HAVE WRITE IN IDLE PYTHON

import numpy as np import pandas as pd import seaborn as sns

import matplotlib.pyplot as plt import scipy.stats as st

# Importing the dataset

df = pd.read\_csv("crime.csv")

# Checking the dimensions of the dataset # print(df.shape)

# Listing the columns of the dataset

# print(df.columns)

# Viewing the first five rows of the dataset # print(df.head())

# Viewing the last five rows of the dataset # print(df.tail())

# Viewing all the information about the dataset # print(df.info())

# Descriptive statistics of the dataset # print(df.describe())

# Checking for missing values in the dataset and their total count # print(df.isnull().sum())

# Viewing the maximum values in the dataset # print(df.max(numeric\_only=True))

# Viewing the minimum values in the dataset # print(df.min(numeric\_only=True))

# Viewing the median values in the dataset # print(df.median(numeric\_only=True))

# Viewing the mean values in the dataset # print(df.mean(numeric\_only=True ))

# Viewing the mode values in the dataset # print(df.mode(numeric\_only=True))

# Counting non-null values in each column # print(df.count())

# Cleaning the dataset by dropping rows with missing values # print(df.dropna(inplace=True))

# Creating a numpy array from the crime rate # crime\_code\_array = np.array(df["Crm Cd"])

# print(crime\_code\_array)

# Filtering years with crime data greater than 50 high\_crime\_years = df[df["Crm Cd"] > 50]

# print(high\_crime\_years)

# Creating a histogram for the "Crm Cd" column

# plt.hist(df["Crm Cd"], bins=10, color="blue", edgecolor="black") # plt.xlabel("Crime Code")

# plt.ylabel("Frequency")

# plt.title("Distribution of Crime Codes") # plt.show()

# Creating a bar chart to show the average crime code by area

# avg\_crime\_code\_by\_area = df.groupby("AREA NAME")["Crm Cd"].mean() # avg\_crime\_code\_by\_area.plot(kind='bar', color='orange')

# plt.xlabel("Area Name")

# plt.ylabel("Average Crime Code")

# plt.title("Crime Code by Area Name") # plt.show()

# Creating a line graph to show the trend of crime code across dates # plt.plot(df["DATE OCC"], df["Crm Cd"], marker='o')

# plt.xlabel("Date of Occurrence") # plt.ylabel("Crime Code")

# plt.title("Trend of Crime Code Across Dates") # plt.show()

# Scatter plot between 'Crm Cd' and 'Vict Age'

# plt.scatter(df["Crm Cd"], df["Vict Age"], color='red') # plt.xlabel("Crime Code")

# plt.ylabel("Victim Age")

# plt.title("Crime Code vs Victim Age") # plt.show()

# Boxplot for "Crm Cd" distribution by year # import seaborn as sns

# sns.boxplot(x="Date Rptd", y="Crm Cd", data=df) # plt.title("Crime Code Distribution by Year") # plt.show()

# Creating a heatmap to visualize the correlation between features # sns.heatmap(df.corr(), annot=True, cmap="coolwarm")

# plt.title("Feature Correlation Heatmap") # plt.show()

# Pair plot for selected columns

# sns.pairplot(df[["Crm Cd", "Vict Age", "Year"]]) # plt.show()

#creating outlier

Q1=df['LAT'].quantile(0.25)

Q3=df['LAT'].quantile(0.75)

print(Q1)

print(Q3)

IQR=Q3-Q1

print(IQR)

lower\_bound=Q1-1.5\*IQR

print(lower\_bound)

upper\_bound=Q3+1.5\*IQR

print(upper\_bound)

outlier=df[(df['LAT']<lower\_bound)| (df['LAT']>upper\_bound)]

print(outlier)

Conclusion :

Working on this crime dataset project has been such a wild and eye-opening journey! As a student still learning data science, I wasn’t sure where this would go at first—but with the help of Python and its amazing libraries like **Pandas**, **Matplotlib**, **Seaborn**, and **numpy**, I was able to explore real-world crime data and turn raw numbers into meaningful insights.

Through the process, I learned how to clean and organize the dataset, visualize it using various types of plots. It was super interesting to observe trends like which crimes are most common, which areas experience more criminal activity, and how certain patterns change depending on the time or location. It felt like doing detective work, but with code instead of a magnifying glass .

What really stood out to me was how much you can discover just by visualizing the data. Heatmaps, bar graphs, and scatter plots helped me see trends I never would have noticed otherwise. Knowing how data could be used to predict or classify crime types—opening the door to the idea that tech can actually help solve real-life problems.

As a student, this project pushed me out of my comfort zone. I had to debug, rethink my approach, and learn new concepts on the go—but that’s what made it so rewarding. It helped me see how powerful Python can be in the world of data analysis and how meaningful stories can be hidden in numbers.

In the end, this wasn’t just about writing code—it was about learning to *ask better questions*, and using data to try and answer them. I’ve still got a lot to learn, but this project gave me a solid foundation and a real sense of how data science can make a difference.

Future Scope :

While this project gave me a strong foundation in data analysis and visualization using Python, there’s still so much more that can be explored. In the future, I’d love to expand this project in a few exciting directions:

**Geospatial Analysis:** Integrating location-based libraries like **Folium** or **GeoPandas** could help visualize crime data on interactive maps, making it easier to identify crime hotspots and patterns geographically.

**Time Series Forecasting:** By focusing more on the time aspect of crime data, I could try building time series models (using libraries like **statsmodels** or **Prophet**) to forecast crime trends in the future. That could actually be useful for law enforcement planning!

**Bigger, Cleaner Datasets:** In the future, I’d love to work with larger and more recent datasets that include additional features like demographic data, weather info, or socio-economic factors. This could help find deeper insights and stronger correlations.

Machine Learning Models and NLP : can be used to find more accurate pattern in crimead eventually will give good results.

**Real-time Dashboards:** It’d be super cool to build a live dashboard using **Dash** so that others can interact with the visualizations and get insights in real-time.

Overall, this project is just a starting point. As I continue learning and building my skills, I hope to make this kind of data analysis more detailed, more useful, and maybe even a bit more impactful.

**References**

**Catalog, “Crime data from 20220 to present ”, catalog.data.gov,2020. [Online].**

**Available:** [**https://catalog.data.gov/dataset/crime-%20data-from-2020-to-present**](https://catalog.data.gov/dataset/crime-%20data-from-2020-to-present)

**[Accessed: Apr. 2- 2025].**