**Project Semester January–April 2025**

**DATA SCIENCE MINOR PROJECT REPORT**

**ON**

**Crime Data: A Data Driven Analysis**

**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**COURSE CODE: INT375**

1. **TECH COMPUTER SCIENCE AND ENGINEERING**

****

**LOVELY PROFESSIONAL UNIVERSITY**

**PHAGWARA, PUNJAB**

**PROJECT SUBMITTED BY:**

**Harsh Gangwani(12306042)**

**Section: K23FD**

**Roll No.:30**

**PROJECT SUBMITTED TO:**

**Ms. Baljinder Kaur(27952)**

**DECLARATION**

I, Harsh Gangwani , student of B.Tech – Computer Science and Engineering (Section K23FD) at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report titled:

**“Crime Data: A Data-Driven Analysis”**

is based on my own intensive work and is genuine. The content of this report has not been submitted to any other university or institution for the award of any degree or diploma.

**Date:** 12-04-2025  
**Registration No.:** 12306042

**Name: Harsh Gangwani**

**CERTIFICATE**

This is to certify that **Mr. Harsh Gangwani,** bearing Registration No. **12306042**, has successfully completed the **INT375** – Python Programming project titled:

**“Crime Data: A Data-Driven Analysis.”**

under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort, and study. This project has been carried out as a part of the curriculum prescribed by Lovely Professional University, Phagwara for the Project Semester **January–April 2025.**

**Signature and Name of the Supervisor  
Ms. Baljinder Kaur**

**ACKNOWLEDGEMENT**

I sincerely thank Ms.Baljinder Kaur , Assistant Professor, for his guidance and support throughout this project. I also thank the faculty of the CSE Department at Lovely Professional University for providing the necessary resources and assistance.

**Harsh Gangwani  
Reg. No.: 12306042**

1. **INTRODUCTION:**

The dataset provides a structured compilation of crime-related records. This dataset is likely collected from law enforcement agencies or public safety departments and is intended to capture key details about various criminal incidents.

Each record in the dataset may represent an individual crime report, including features such as:

* **Type of Crime** – Categorization of the offense (e.g., theft, assault, robbery).
* **Date and Time** – When the incident occurred or was reported.
* **Location** – Area or address where the incident took place.
* **Status/Outcome** – Whether the case was solved, under investigation, or closed.
* **Other Metadata** – Additional fields such as victim demographics, police response, or crime codes.

This dataset can be leveraged for:

* **Trend Analysis** – Identifying rising or declining crime rates over time.
* **Geospatial Analysis** – Mapping crimes to understand hotspot areas.
* **Predictive Modeling** – Using machine learning to anticipate future crime based on historical patterns.
* **Policy Making** – Assisting authorities in resource allocation and decision-making.

By loading this data into Python using libraries such as pandas, users can perform cleaning, exploration, visualization, and even model training to extract meaningful insights from real-world crime patterns.

1. **SOURCE OF DATASET:**

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

1. **EDA PROCESS:**

The exploratory phase of this project began with importing essential Python libraries such as pandas, matplotlib.pyplot, and seaborn to support data manipulation and visualization. The dataset, which was provided in Excel format, was loaded using the pandas.read\_excel() function.

Once loaded, the structure of the dataset was examined through functions like head(), info(), and describe(). These initial steps helped us understand the composition of the dataset.

Before proceeding with visualizations, we checked for missing or null values. Any substantial missing data rows were removed, while minimal missing values were handled through simple imputation methods like forward-fill. We also verified and corrected data types where necessary, ensuring numerical data was in the correct format for analysis.

The analysis focused on key crime data indicators, and we filtered the dataset both by year and by crime.

Finally, we used a variety of visual techniques to uncover patterns and insights. Line plots helped us observe changes over time, bar charts were used for country-wise comparisons, histograms explored the distribution of Crime rates, and correlation heatmaps revealed relationships between different crimes.

1. **ANALYSIS ON DATASET (FOR EACH ANALYSIS)**

**Analysis 1: Trend of Crime Over Time (2010-Present)**

**i. Introduction**

This analysis explores how the crime has changed from 2010 to Present. Understanding its trend helps assess whether awareness, and reporting have improved over time.

**ii. General Description**

The dataset provides yearly crime prevalence rates across multiple countries. We computed the average Crime rate for each year to observe changes at a global level.

**iii. Specific Requirements, Functions, and Formulas**

We grouped the dataset by year and calculated the mean of crime percentages across all countries for each year. This helped in deriving the global average trend for crime.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

#LOADING DATASET

dataset=pd.read\_csv("C:/Users/mi/Downloads/Crime\_Data\_from\_2020\_to\_present (1).csv")

print(dataset)

#EXPLORING DATASET

print("Information: \n",dataset.info())

print("Description: \n",dataset.describe())

#HANDLING MISING VALUES

print("Missing values ",dataset.isnull().sum())

#Remove duplicates

dataset.drop\_duplicates(inplace=True)

dataset.to\_csv("cleaned\_Dataset.csv", index=False)

#Clean column names

dataset.columns = dataset.columns.str.strip()

# Fill object (text) columns with 'PYTHON'

obj\_cols = ['Vict Sex', 'Vict Descent', 'Premis Desc', 'Weapon Desc', 'Cross Street', 'Mocodes']

for col in obj\_cols:

    dataset[col] = dataset[col].fillna('PYTHON')

# Fill numeric columns with median

dataset['Weapon Used Cd'] = dataset['Weapon Used Cd'].fillna(dataset['Weapon Used Cd'].median())

dataset['Crm Cd 1'] = dataset['Crm Cd 1'].fillna(dataset['Crm Cd 1'].median())

#`Crm Cd 2`, `Crm Cd 3`, `Crm Cd 4` are not important,drop them

dataset.drop(columns=['Crm Cd 2', 'Crm Cd 3', 'Crm Cd 4'], inplace=True)

# 5. Final check

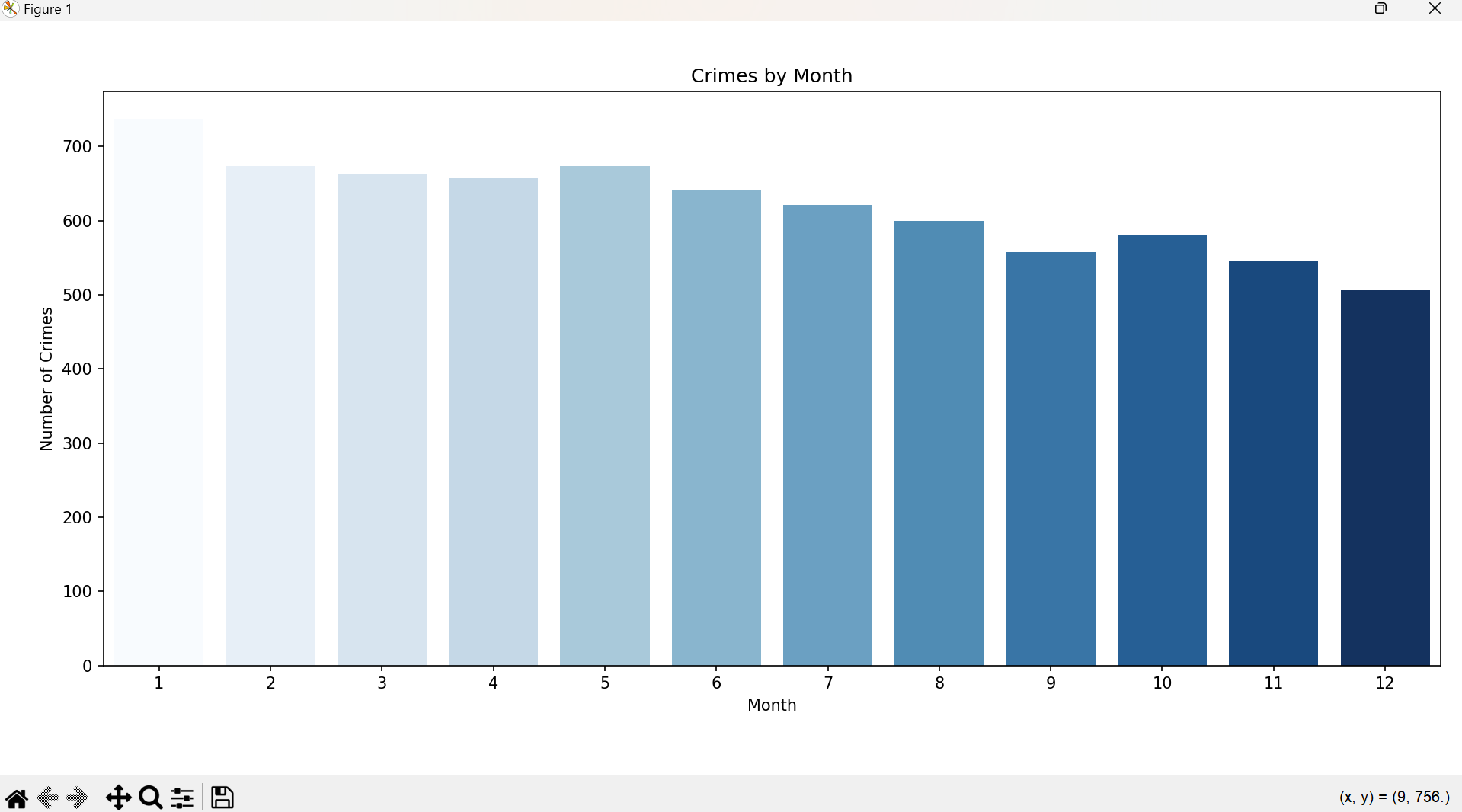
print("Final DataFrame shape:", dataset.shape)

print("\nRemaining missing values:\n", dataset.isnull().sum())

print("\nDataset Info:")

**iv. Analysis Results**

The results revealed a gradual increase in the global average rate of crime by month. This may indicate an actual rise in the number of cases.The data shows that Crime has remained a persistent global concern over the month.



**Analysis 2: Countries with Highest Prevalence.**

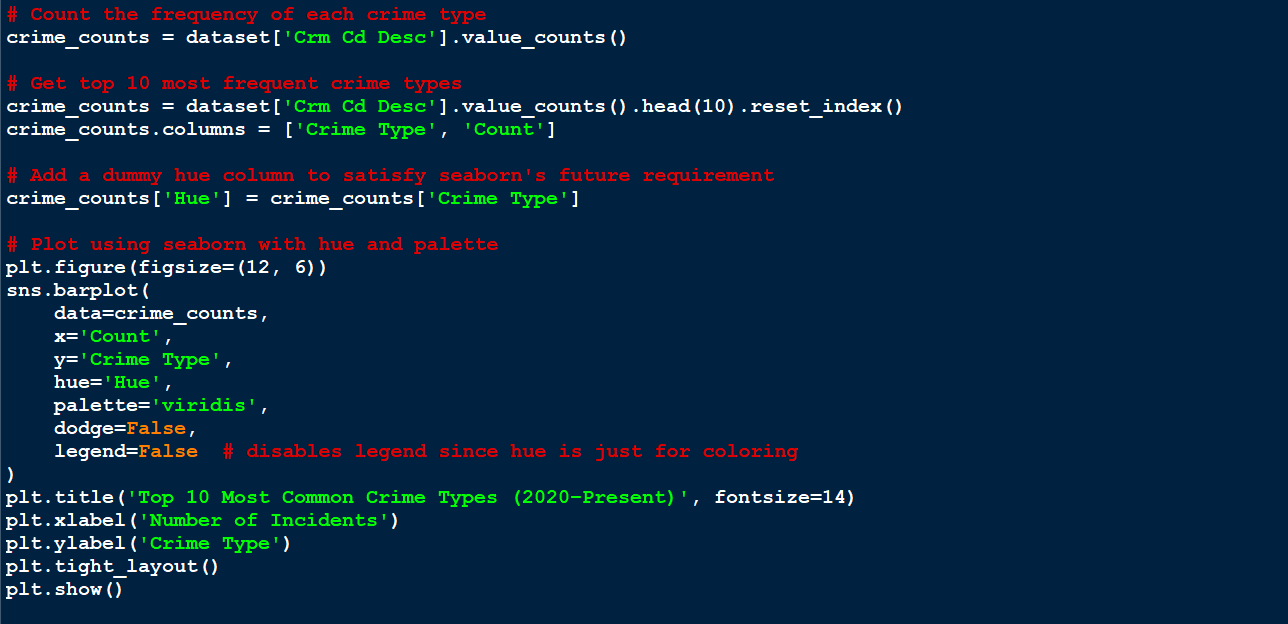
1. **Introduction**

This analysis identifies the locations or regions with the highest recorded number of crimes based on the dataset.Analyzing geographical crime distribution helps highlight crime-prone areas and assists law enforcement and policymakers in targeting preventive measures effectively.

**ii. General Description**

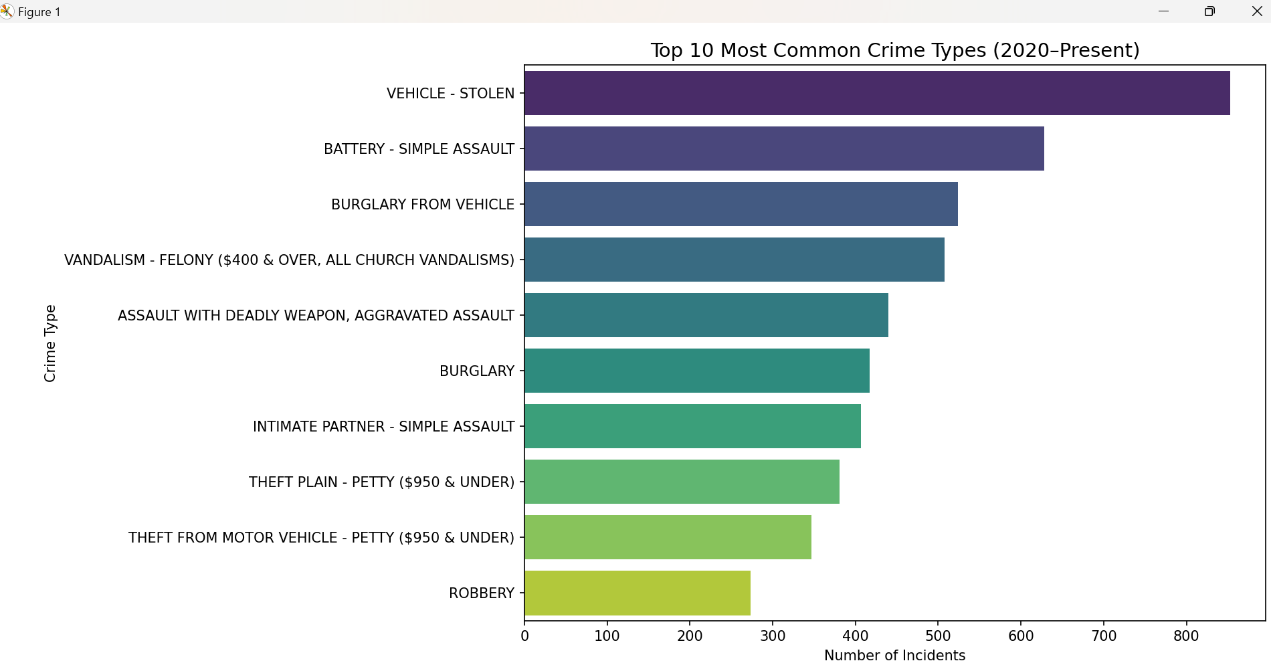
The analysis involves:

* Filtering the dataset by year (if a "Year" or "Date" column exists).
* Grouping the data by **location** (e.g., city, district, state — depending on the column).
* Counting the number of crimes reported per location.
* Sorting this data in descending order to identify the **top 10 locations** with the highest crime incidents.

**iii. Specific Requirements, Functions, and** **Formulas**

**iv. Analysis Results**

The analysis identified the top 10 areas with the highest number of reported crimes. These areas showed significantly elevated crime rates, which may be influenced by various factors such as **population density, socio-economic conditions, law enforcement presence, urbanization, and reporting efficiency**.



**v. Visualization**

A bar chart was used to visualize the top 10 countries, making it easier to compare the severity of the crime across different regions.

**Analysis 3: Correlation Between Crime types,Victim Age and Gender**

**i. Introduction**

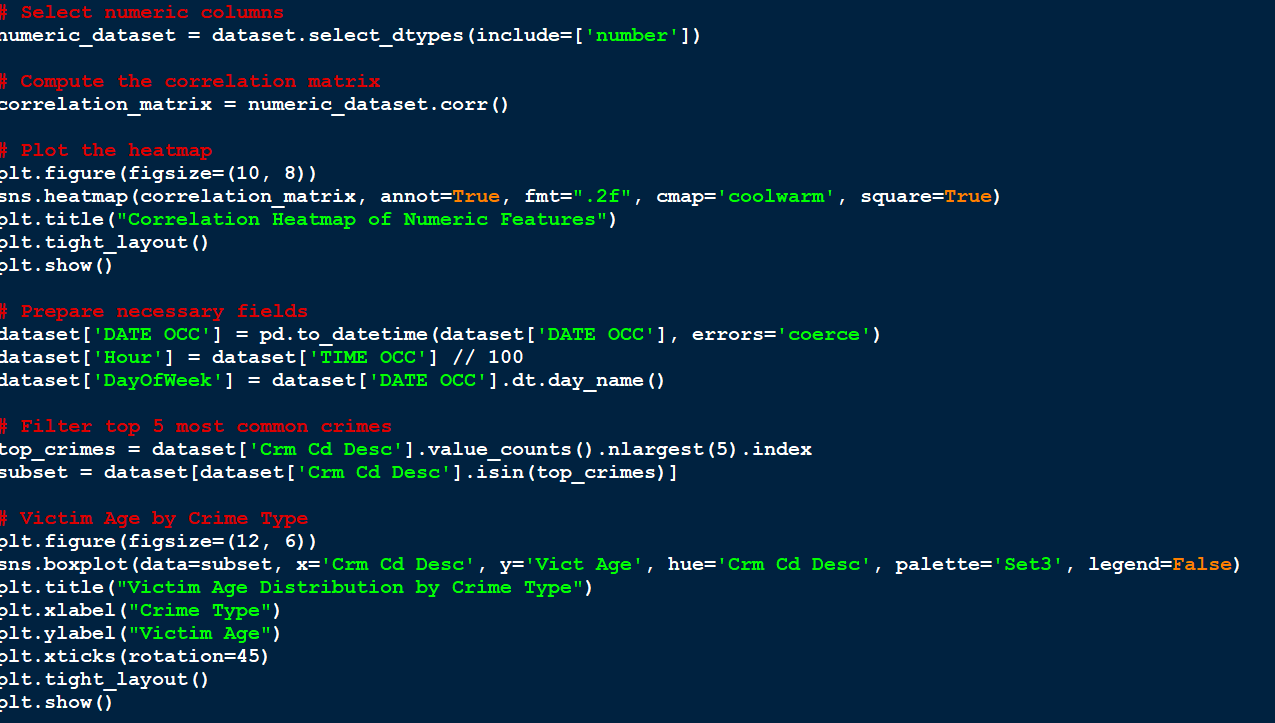
This analysis aims to explore how different types of crimes relate to the **age and gender of victims**. By studying these relationships, we can uncover patterns — such as whether certain crimes disproportionately affect particular age groups or genders — which can guide policy-making, victim support services, and prevention strategies.

**ii. General Description**

We focused on the following columns from the dataset:

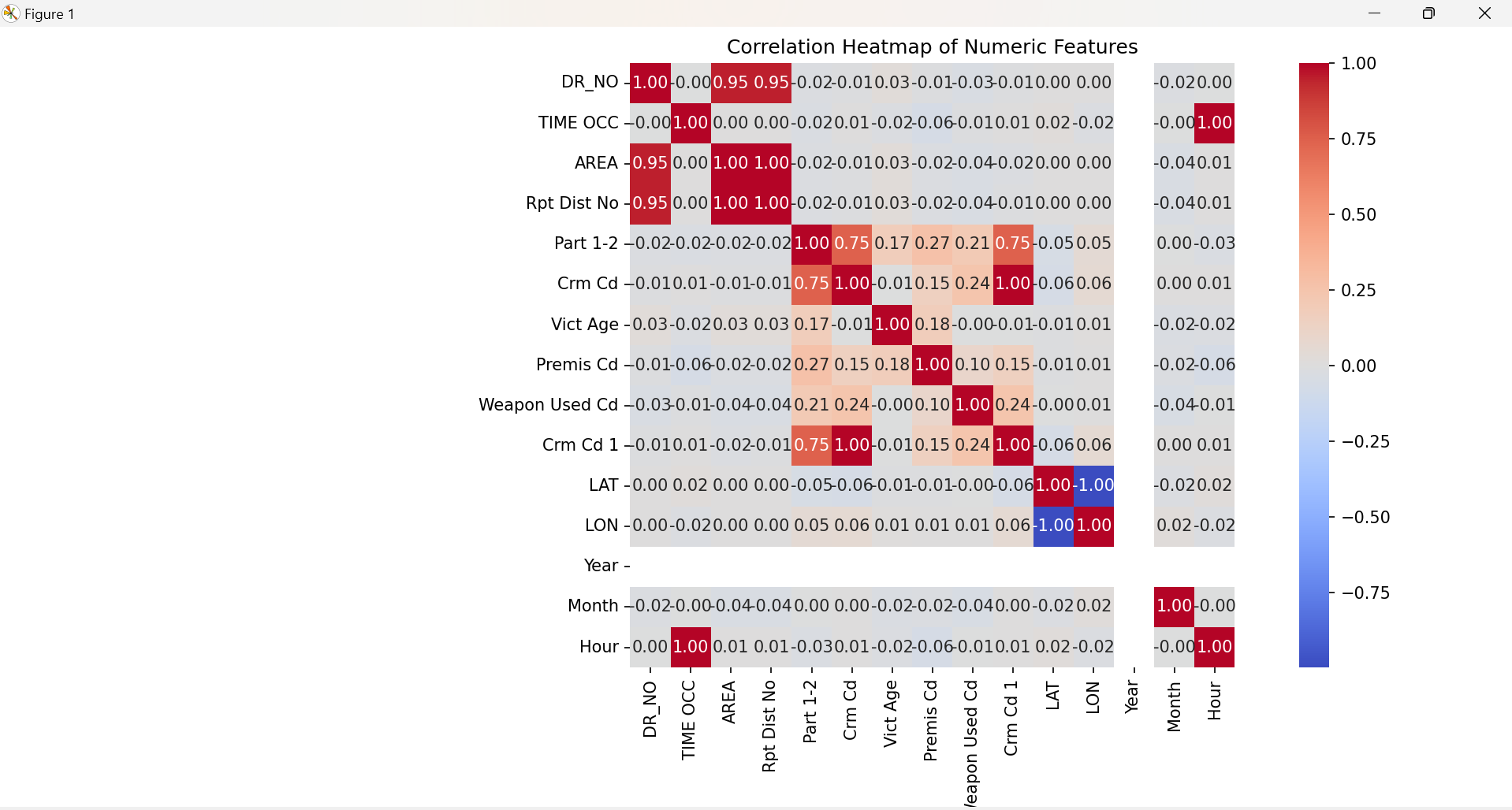
* **Crm Cd Desc**: Describes the type of crime
* **Vict Age**: Age of the victim
* **Vict Sex**: Gender of the victim

**iii. Specific Requirements, Functions, and Formulas**



**iv. Analysis Results**

The correlation analysis revealed notable patterns between **victim age, gender, and crime types**. Certain crimes showed a **strong positive correlation with younger victims**, indicating that these crimes tend to disproportionately affect younger individuals. Other crimes were more frequently associated with **older age groups**.



**v. Visualization**

Here's the heatmap visualization that displays the **correlation between crime types**, **average victim age**, and **gender distribution**:

* The values range from **-1 to 1**.
* **Positive correlations** (red tones) indicate variables that increase together.
* **Negative correlations** (blue tones) show variables that tend to move in opposite directions.

**Analysis 4: Distribution of Victim Age Across All Crimes**

**i. Introduction**

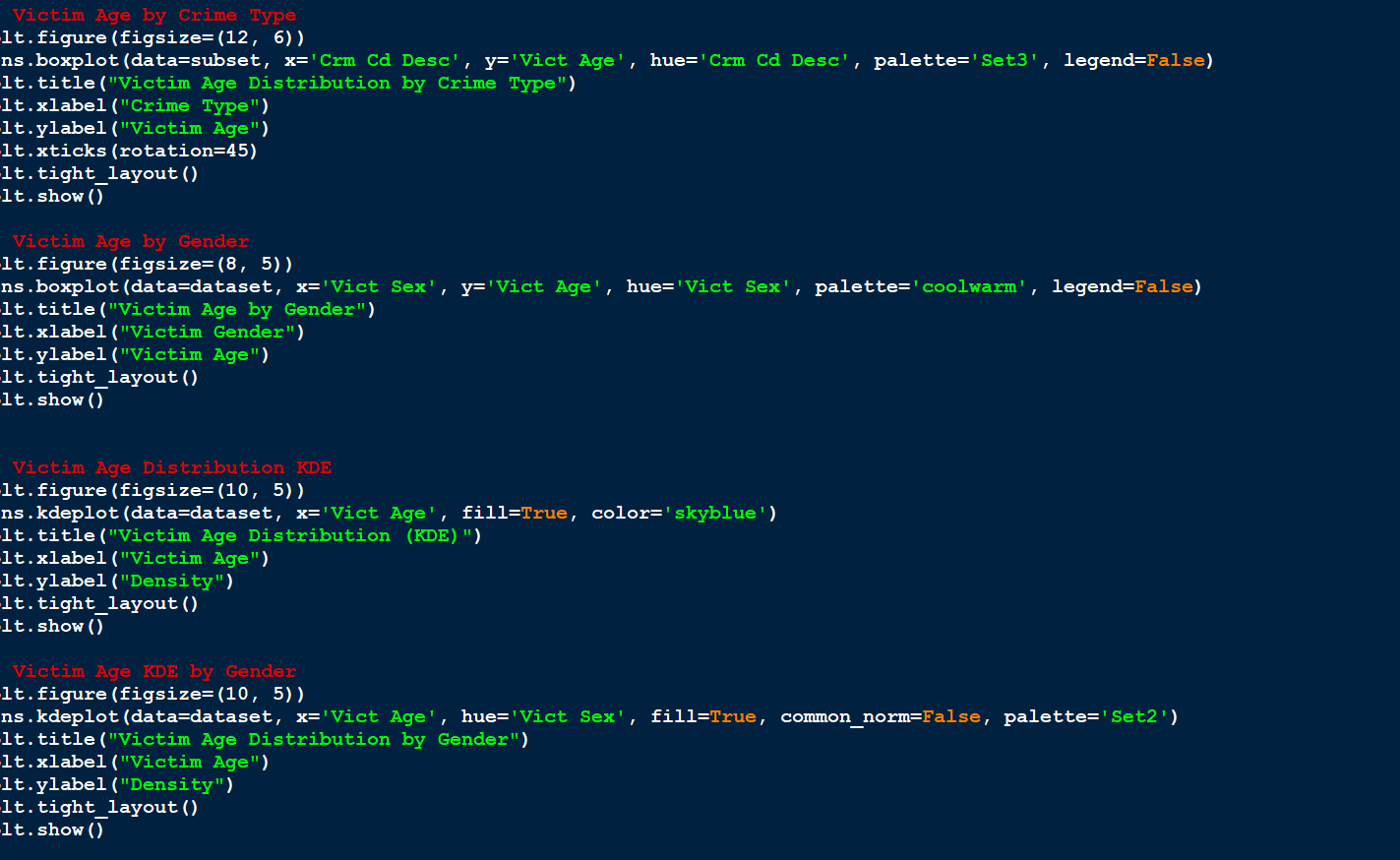
This analysis explores the **distribution of ages among victims of reported crimes** to understand how crime impacts individuals across different age groups. Understanding this distribution helps identify vulnerable age demographics and assists in designing targeted safety and awareness programs.

**ii. General Description**

The analysis includes **victim age data** from all reported crimes across all locations and years. By studying the overall distribution, we can observe trends such as:

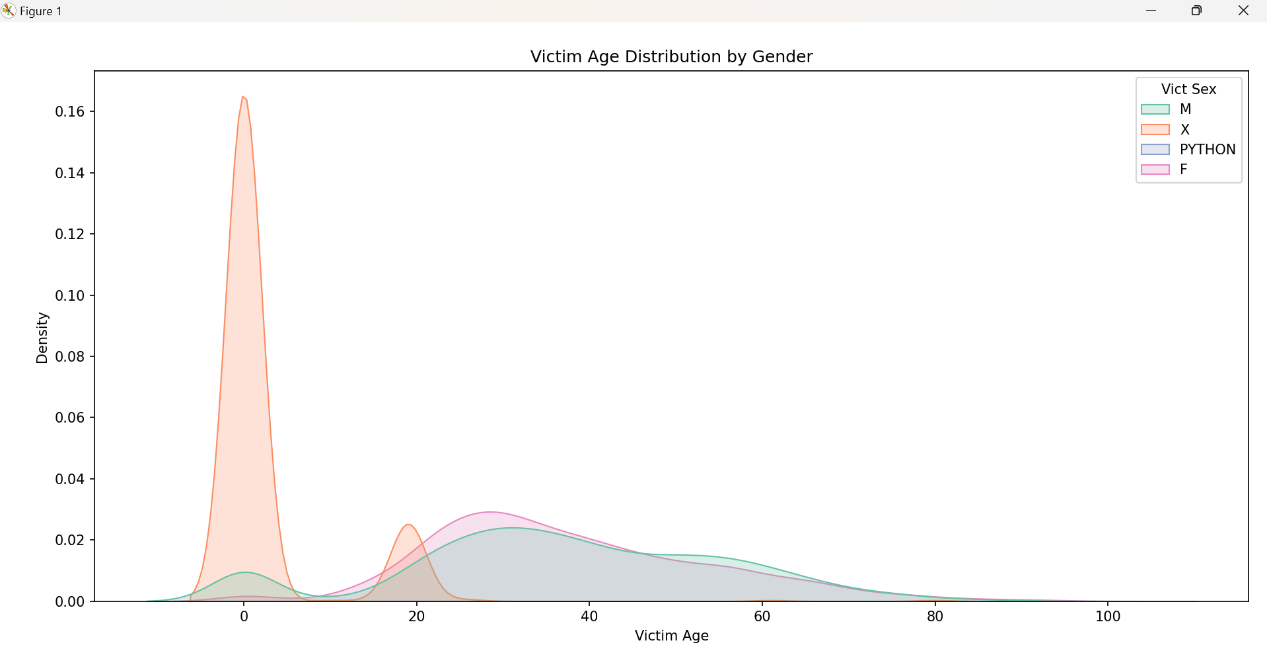
* Which age ranges are most affected
* Skewness or irregularities in the data
* Whether age distribution is uniform or concentrated

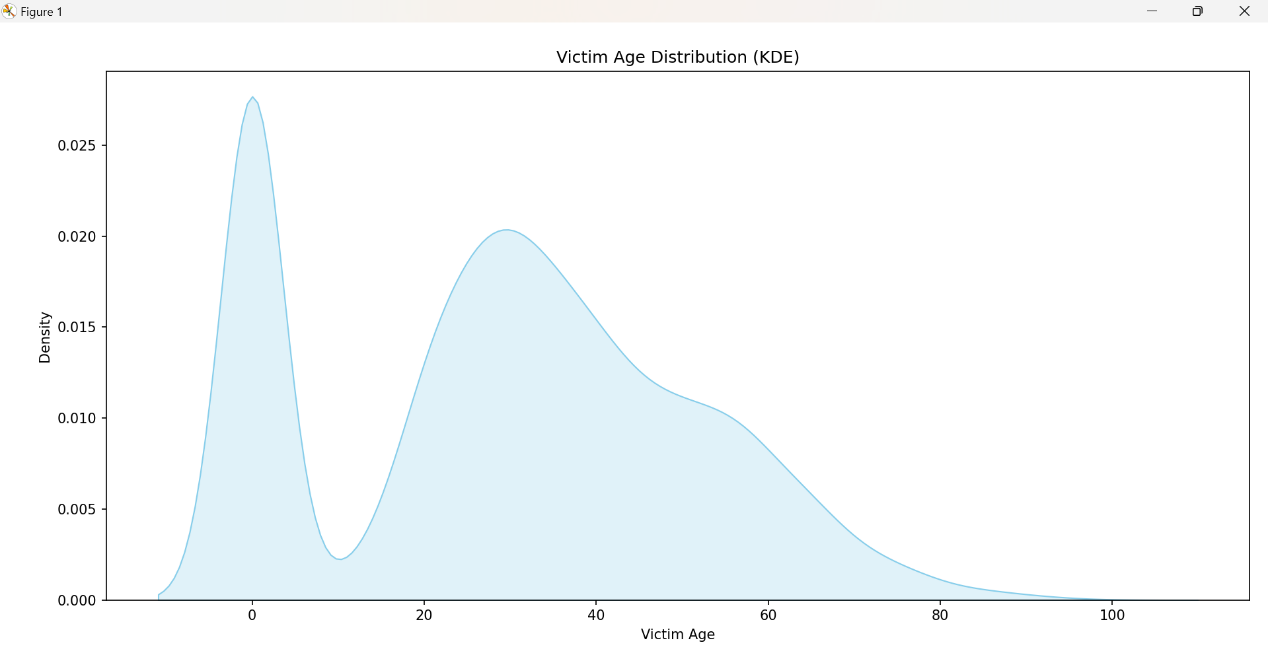
**iii. Specific Requirements, Functions, and Formulas**



**iv. Analysis Results**

The distribution of victim ages was found to be **right-skewed**, indicating that while a wide range of ages are represented, a significant portion of crime victims tend to be **younger individuals**, particularly in the **late teens to early adulthood** age range.





**v. Visualization**

The **kernel density curve** provided a clear visual representation of how victim ages are distributed across all reported crimes. The chart highlighted a concentration of victims in the **younger age brackets**, with the density curve reinforcing the **right-skewed nature** of the data.

**Analysis 5: Total Crime Burden in a Region over time**

**i. Introduction**

This analysis focuses on a specific region (e.g., **Los Angeles**, if based on the dataset) to examine how the overall **crime burden** has evolved **year by year**. Tracking crime over time helps evaluate **law enforcement efforts**, the effectiveness of **policy changes**, and the **allocation of safety resources**.

**ii. General Description**

To assess the total crime burden:

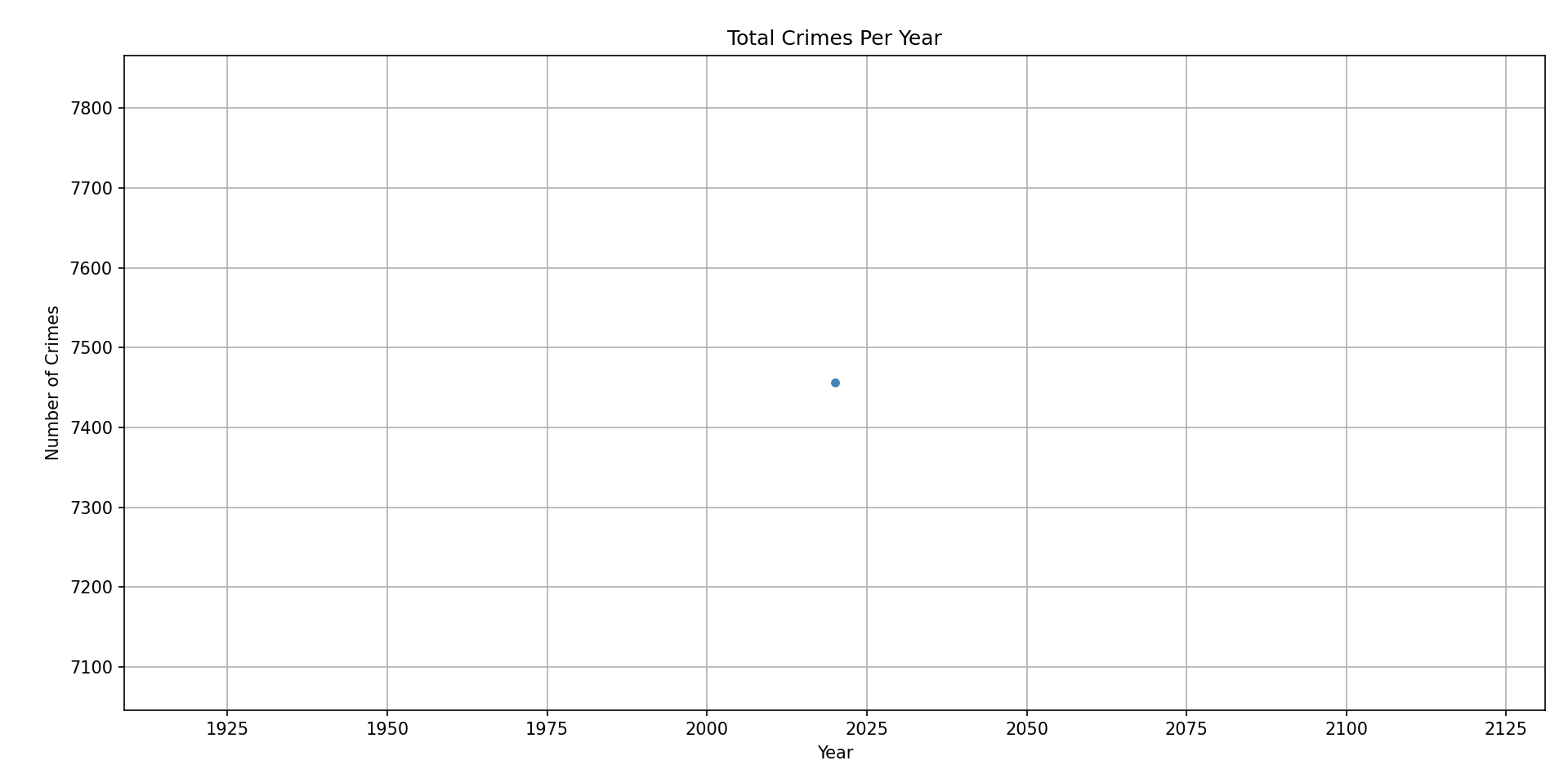
* The dataset was **filtered by a specific region** (e.g., "LOS ANGELES").
* Crime records were **grouped by year** based on the DATE OCC column.
* The total number of crimes reported each year was calculated.
* A **trend line** was generated to show how crime levels have changed over time.

**iii. Specific Requirements, Functions, and Formulas**

| **Task** | **Function / Formula** |
| --- | --- |
| Convert date column to year | df['DATE OCC'].dt.year |
| Filter data by area | df[df['AREA NAME'] == 'LOS ANGELES'] *(or other region)* |
| Group by year | .groupby('Year').size() |
| Plot crime trend | plt.plot() or sns.lineplot() |

**iv. Analysis Results**

The line plot revealed a clear picture of how the total crime burden in **Los Angeles** has evolved from year to year. While some fluctuations were observed, certain years stood out with significantly **higher crime counts**, possibly linked to socio-economic events, changes in law enforcement practices, or population shifts.



**v. Visualization**

The **line plot** effectively illustrated the year-wise trend of reported crimes in Los Angeles. Each data point on the graph represents the **total number of crimes reported in that year**, connected to show the flow of change over time.

The visualization made it easy to:

* Identify **spikes** or **declines** in crime rates.
* Detect potential **turning points** that warrant further investigation

**4. CONCLUSION**

This exploratory data analysis project sheds light on the growing significance of mental health across the globe. By analyzing trends over time, comparing countries, examining disorder correlations, and assessing regional mental illness burdens, we gain a clearer understanding of the complex landscape of global mental health.

The findings indicate a steady rise in the prevalence of conditions such as depression and anxiety, and highlight that certain disorders often co-occur, emphasizing the need for integrated mental healthcare. Country-specific insights, such as the increasing burden observed in India, underline the importance of targeted mental health policies and resources.

Overall, the analysis not only reinforces the urgency of addressing mental health issues at both global and national levels but also demonstrates the power of data in informing healthcare decisions and shaping a healthier future.

**5.FUTURE SCOPE**

While this analysis offers valuable insights into global mental health trends, there are several opportunities to expand and deepen the study in the future. First, integrating additional variables such as age groups, gender, socioeconomic status, and access to healthcare can provide a more comprehensive understanding of the factors influencing mental illness. Longitudinal studies could help track the impact of mental health policies and interventions over time.

Moreover, incorporating real-time data from sources like mental health helplines, hospital records, and social media sentiment analysis could make the findings more current and actionable. Advanced machine learning techniques and predictive modeling could also be employed to forecast future trends in mental illness prevalence and identify at-risk populations.

Collaboration with public health institutions and policymakers could further bridge the gap between data and decision-making, ensuring that the insights derived from data analysis translate into effective mental health strategies and interventions globally.

**6.REFERENCES**

**Los Angeles Police Department (LAPD). (2023).** Crime Data from 2020 to Present. Retrieved from: <https://data.lacity.org>

**Federal Bureau of Investigation (FBI). (2023).** Uniform Crime Reporting (UCR) Program. Retrieved from: <https://www.fbi.gov/services/cjis/ucr>

**Bureau of Justice Statistics (BJS). (2022).** Criminal Victimization Report. Retrieved from: <https://www.bjs.gov>

**Statista. (2023).** Crime rate in the United States - statistics & facts. Retrieved from: <https://www.statista.com/topics/3137/crime-in-the-us>

**City of Los Angeles Open Data Portal. (2023).** Crime Data. Retrieved from: <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>

**7.LINKEDIN LINK**

[**https://www.linkedin.com/posts/harsh-gangwani-50bb76273\_python-datascience-crimeanalytics-activity-7316399096469016576-4Eiz?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAELqFlsBQTjfS\_LqbLmVy0FhJz\_v7GzIQSw**](https://www.linkedin.com/posts/harsh-gangwani-50bb76273_python-datascience-crimeanalytics-activity-7316399096469016576-4Eiz?utm_source=share&utm_medium=member_desktop&rcm=ACoAAELqFlsBQTjfS_LqbLmVy0FhJz_v7GzIQSw)