DATA SCIENCE TOOL BOX:PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

LOS ANGELES CRIME DATA

Submitted by : VISHNU

Registration No : 12307182

Section : KM005

Course Code : INT 375

Under the Guidance of

Maneet Kaur

UID : 15709

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Vishnu bearing Registration no. 12307182 has completed INT 375 project titled, **“Exploratory Data Analysis on Los Angeles Crime Data”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 12-04-2025

DECLARATION

I, Vishnu , student of Introduction To Data Management 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 : 12-04-2025	**Signature**
Registration No : 12307182	**VISHNU**

Table of Contents

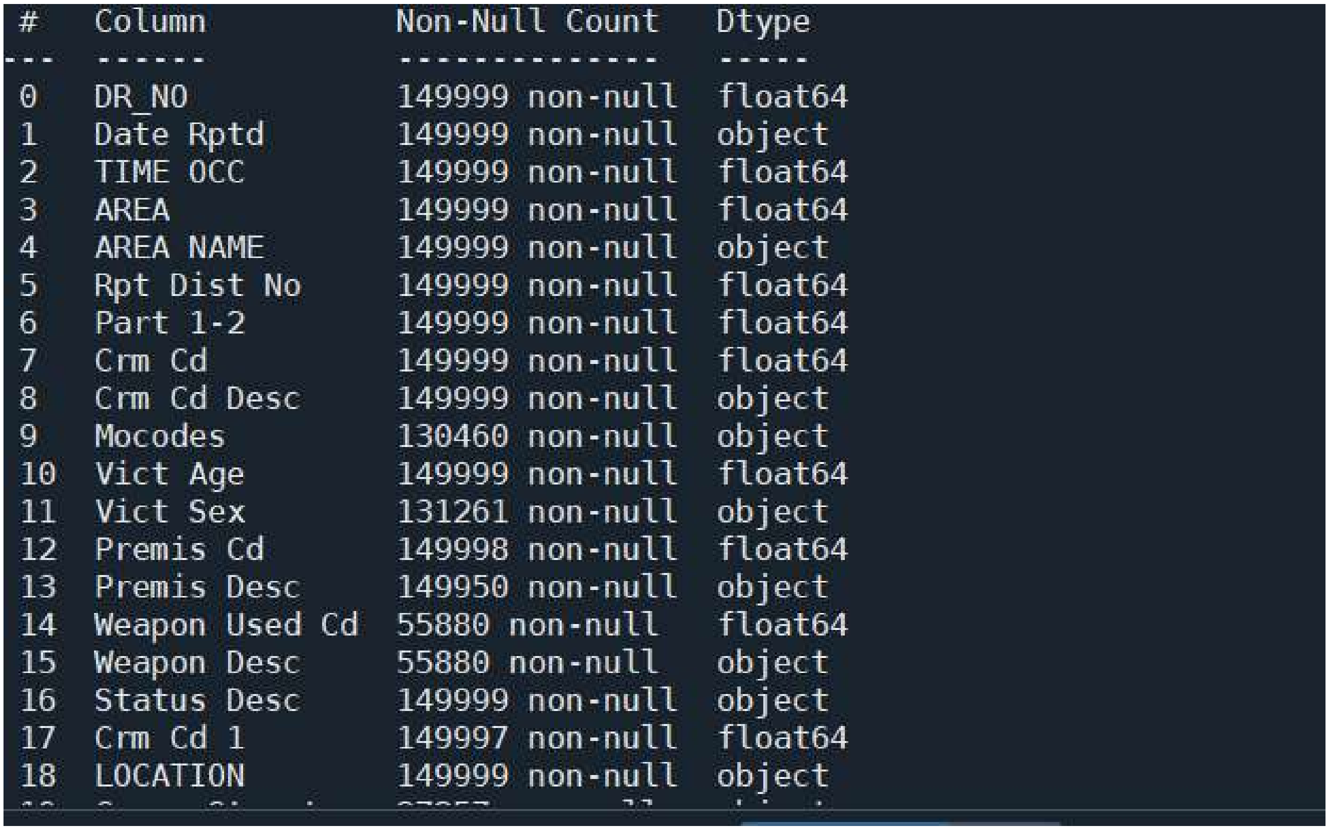
1. Introduction

In this project, I worked on analyzing a real-world crime dataset that contains incidents reported between 2020 and the present. The main goal was to explore the data, understand patterns related to crime types, victim information, and time-based trends, and finally, try predicting victim age using machine learning.

This kind of analysis helps give a clearer picture of crime patterns and can be a great starting point for developing smarter safety strategies in the future.

2. Source of Dataset

The dataset used in this project is titled **"Crime_Data_from_2020_to_Present.csv"**. It includes detailed records like report dates, times of occurrence, types of crimes, location codes, victim demographics, and more. It's a solid dataset for doing exploratory analysis



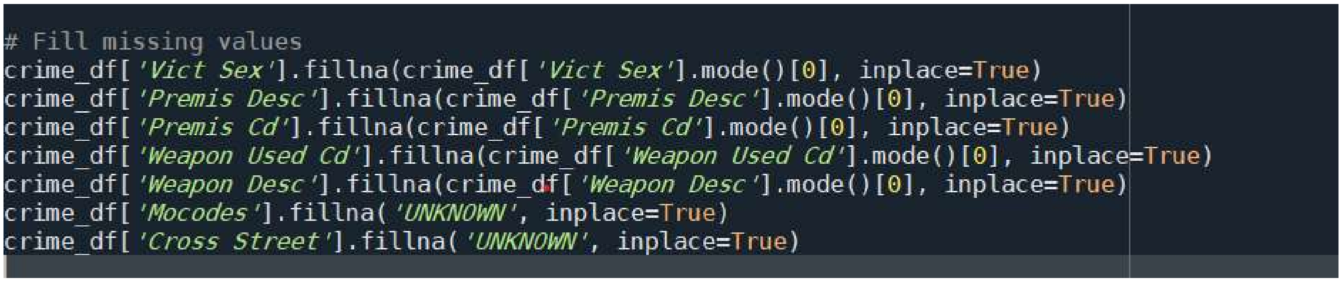
```
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   DR_NO           149999 non-null  float64
 1   Date Rptd       149999 non-null  object
 2   TIME OCC        149999 non-null  float64
 3   AREA            149999 non-null  float64
 4   AREA NAME       149999 non-null  object
 5   Rpt Dist No     149999 non-null  float64
 6   Part 1-2        149999 non-null  float64
 7   Crm Cd          149999 non-null  float64
 8   Crm Cd Desc     149999 non-null  object
 9   Mocodes         130460 non-null  object
 10  Vict Age        149999 non-null  float64
 11  Vict Sex        131261 non-null  object
 12  Premis Cd       149998 non-null  float64
 13  Premis Desc     149950 non-null  object
 14  Weapon Used Cd  55880 non-null   float64
 15  Weapon Desc     55880 non-null   object
 16  Status Desc     149999 non-null  object
 17  Crm Cd 1        149997 non-null  float64
 18  LOCATION        149999 non-null  object
```

3. Data Cleaning & Preprocessing

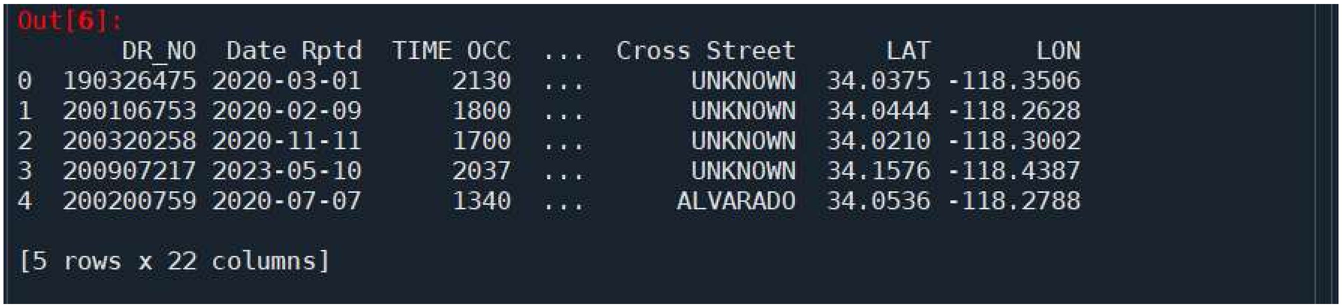
Before jumping into the actual analysis, a good amount of data cleaning was done. Here's what I focused on:

- Removed rows where key info (like crime date, time, or victim age) was missing.
- Filled in missing values for things like victim gender, premise code/description, and weapon info using the most frequent values (mode).
- Converted date columns to proper datetime format and made sure numeric columns were in the right type.
- Added a new column to extract the year from the report date so we could analyze trends over time.

After all this, the data was ready for proper analysis and modeling.



```python
# Fill missing values
crime_df['Vict Sex'].fillna(crime_df['Vict Sex'].mode()[0], inplace=True)
crime_df['Premis Desc'].fillna(crime_df['Premis Desc'].mode()[0], inplace=True)
crime_df['Premis Cd'].fillna(crime_df['Premis Cd'].mode()[0], inplace=True)
crime_df['Weapon Used Cd'].fillna(crime_df['Weapon Used Cd'].mode()[0], inplace=True)
crime_df['Weapon Desc'].fillna(crime_df['Weapon Desc'].mode()[0], inplace=True)
crime_df['Mocodes'].fillna('UNKNOWN', inplace=True)
crime_df['Cross Street'].fillna('UNKNOWN', inplace=True)
```



```
Out[6]:
       DR_NO  Date Rptd  TIME OCC  ...  Cross Street      LAT       LON
0  190326475 2020-03-01      2130  ...       UNKNOWN  34.0375 -118.3506
1  200106753 2020-02-09      1800  ...       UNKNOWN  34.0444 -118.2628
2  200320258 2020-11-11      1700  ...       UNKNOWN  34.0210 -118.3002
3  200907217 2023-05-10      2037  ...       UNKNOWN  34.1576 -118.4387
4  200200759 2020-07-07      1340  ...      ALVARADO  34.0536 -118.2788

[5 rows x 22 columns]
```

4. Analysis

Here are the key points I explored:

- Found the 10 most common types of crimes.
- Looked at how victim age and gender were distributed.
- Tracked crime numbers by year to check for patterns.
- Used a boxplot to spot outliers in victim age.
- Created a correlation map to see how numeric features relate.

I also trained a **Linear Regression model** to predict a victim's age based on:

- Area
- Time of Occurrence
- Crime Code
- Premise Code

Here’s how the model performed (values shown are from testing):

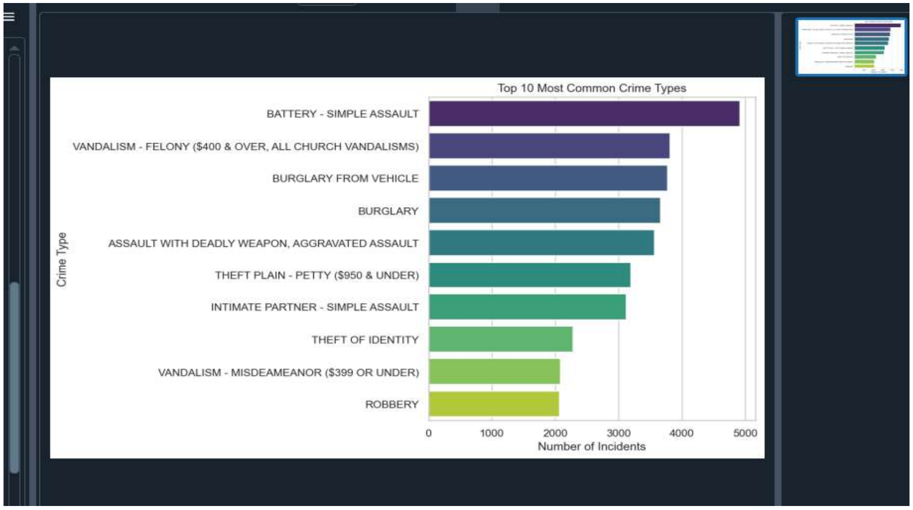
- **MAE (Mean Absolute Error)**: 15.569934296765167
- **RMSE (Root Mean Squared Error)**: 19.717333218495796
- **R² Score**: 0.002696484740044358

The model gave okay results — not perfect, but a good baseline.

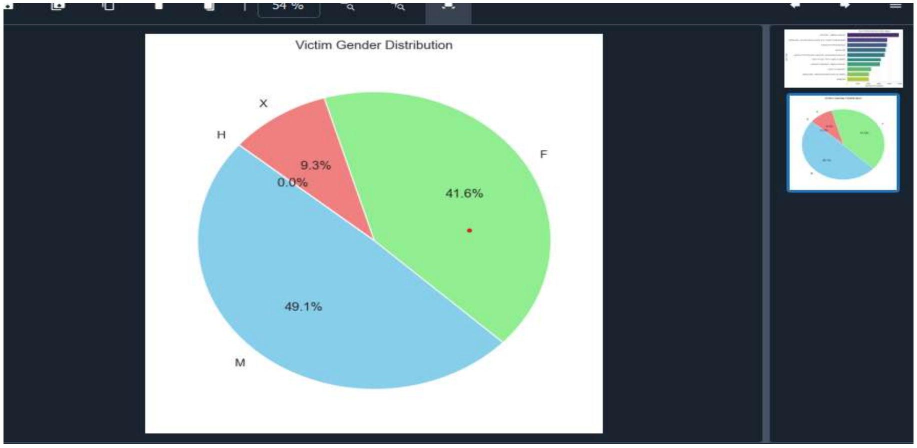
5. Visualizations

To make sense of everything visually, I created several plots:

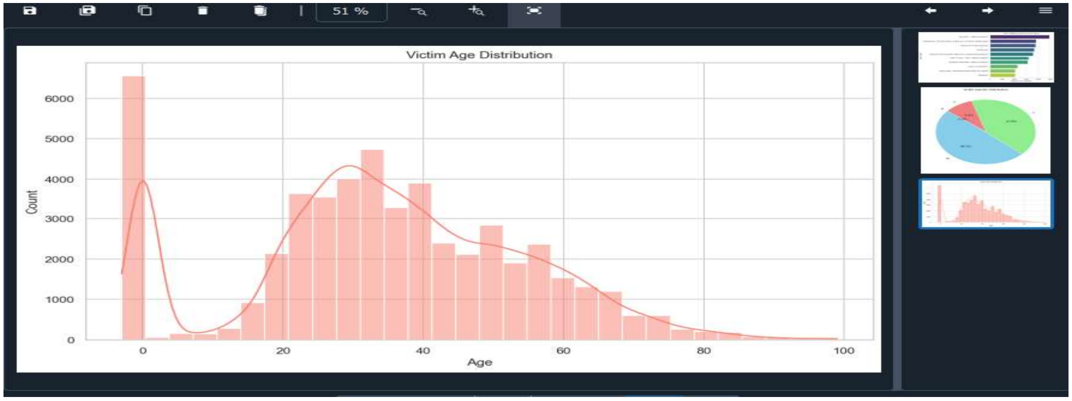
1. **Bar chart** showing the top 10 crime types.



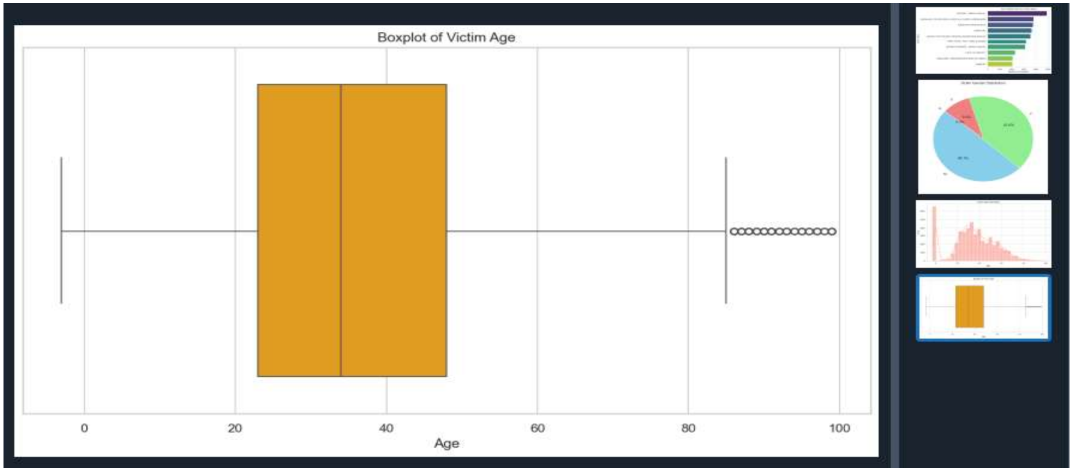
2. **Pie chart** showing victim gender distribution.



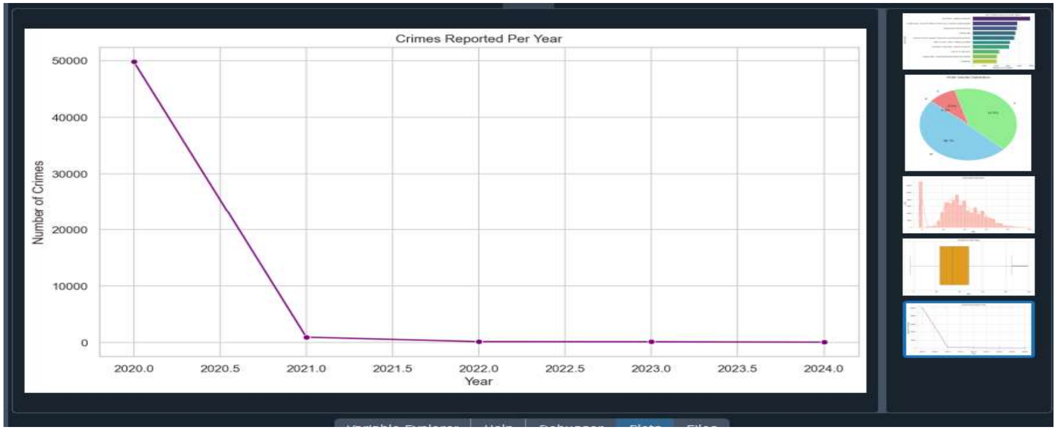
3. **Histogram** for victim age distribution.



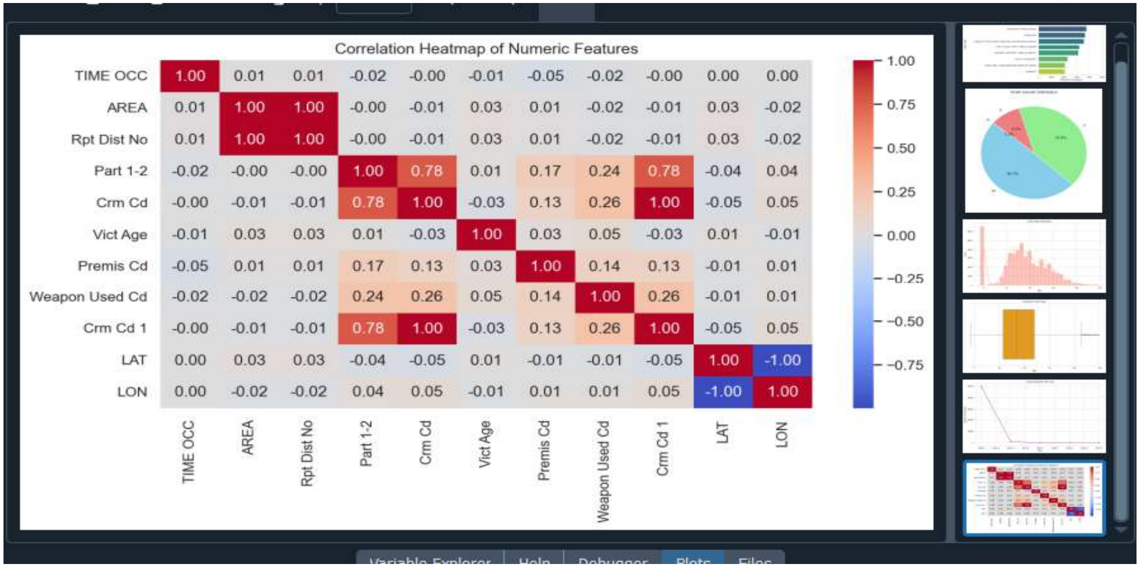
4. **Boxplot** to detect any outliers in victim ages.



5. **Line plot** to show number of crimes reported each year.



6. **Correlation heatmap** to understand numeric relationships.



	TIME OCC	AREA	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Premis Cd	Weapon Used Cd	Crm Cd 1	LAT	LON
TIME OCC	1.00	0.01	0.01	-0.02	-0.00	-0.01	-0.05	-0.02	-0.00	0.00	0.00
AREA	0.01	1.00	1.00	-0.00	-0.01	0.03	0.01	-0.02	-0.01	0.03	-0.02
Rpt Dist No	0.01	1.00	1.00	-0.00	-0.01	0.03	0.01	-0.02	-0.01	0.03	-0.02
Part 1-2	-0.02	-0.00	-0.00	1.00	0.78	0.01	0.17	0.24	0.78	-0.04	0.04
Crm Cd	-0.00	-0.01	-0.01	0.78	1.00	-0.03	0.13	0.26	1.00	-0.05	0.05
Vict Age	-0.01	0.03	0.03	0.01	-0.03	1.00	0.03	0.05	-0.03	0.01	-0.01
Premis Cd	-0.05	0.01	0.01	0.17	0.13	0.03	1.00	0.14	0.13	-0.01	0.01
Weapon Used Cd	-0.02	-0.02	-0.02	0.24	0.26	0.05	0.14	1.00	0.26	-0.01	0.01
Crm Cd 1	-0.00	-0.01	-0.01	0.78	1.00	-0.03	0.13	0.26	1.00	-0.05	0.05
LAT	0.00	0.03	0.03	-0.04	-0.05	0.01	-0.01	-0.01	-0.05	1.00	-1.00
LON	0.00	-0.02	-0.02	0.04	0.05	-0.01	0.01	0.01	0.05	-1.00	1.00

These visualizations made it a lot easier to see what's really going on in the data.

6. Conclusion

From this analysis, I got a clearer picture of:

- Which crimes are most common.
- What kind of victims are most affected.
- How crime patterns change over time.

Even though the regression model wasn't highly accurate, it did show that there's *some* relationship between crime features and victim age. There's definitely potential for improvement using better models.

7. Future Scope

Here's how this project could be taken further:

- Analyze where crimes happen geographically using maps.
- Bring in more datasets — like economic data or weather — to see if they impact crime.
- Build predictive tools for future crime trends.

8. References

- Dataset: *Crime_Data_from_2020_to_Present.csv*
- Libraries used: *pandas, numpy, seaborn, matplotlib, sklearn*
- Subject: *INT375 - Python Programming*
- *Dataset Link :* https://catalog.data.gov/dataset/crime-data-from-2020-to-present