**PYTHON PROJECT REPORT**

(Project Semester: January-April 2025)

**Title of the Project: EDA ON CRIME ANALYSIS USING PYTHON**

**Submitted by:**

**Sudeep Kumar Reddy Eaga  
Registration No.: 12318529  
Programme and Section: B-Tech CSE (K23KM)  
Course Code: INT375**

**Under the Guidance of:  
Anchal Kaundal (UID : 29612)**

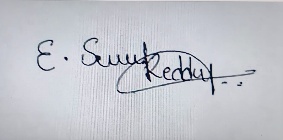
**Discipline of CSE/IT**  
**Lovely School of Computer Science & Engineering**  
**Lovely Professional University, Phagwara**

**DECLARATION**

I, **Sudeep Kumar Reddy Eaga**, student of **Bachelors of Technology (B.Tech)** 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: 03-April-2025

Signature:

  
Registration No.: 12318529  
Name of the Student: **Sudeep Kumar Reddy Eaga**

# ****CERTIFICATE****

This is to certify that **Sudeep Kumar Reddy Eaga** bearing Registration No. **12318529** has completed **INT375** project titled **“EDA ON CRIME ANALYSIS USING PYTHON”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort, and study.

**Anchal Kaundal**  
**Assistant Professor**  
**School of Computer Science & Engineering**

**Lovely Professional University**  
**Phagwara, Punjab**

Date: **08-April-2025**

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to **Anchal Kaundal Sir’s**, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to **Lovely Professional University** for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

# ****TABLE OF CONTENTS****

1. [Introduction](#_1._INTRODUCTION)
2. [Source of Dataset](#_2._SOURCE_OF)
3. [Dataset Preprocessing](#_3._DATASET_PREPROCESSING)
4. [Analysis on Dataset (for each objective)](#_4._ANALYSIS_ON)
5. [Conclusion](#_5._CONCLUSION)
6. [Future Scope](#_6._FUTURE_SCOPE)
7. [References](#_7.REFERENCES)

# ****1. INTRODUCTION****

In the modern era of data-driven decision-making, organizations increasingly rely on data analysis to understand business trends, customer behaviour, and sales performance. One of the most effective techniques for uncovering insights from raw datasets is **Exploratory Data Analysis (EDA)**. EDA plays a crucial role in identifying patterns, spotting anomalies, testing hypotheses, and checking assumptions through statistical summaries and visualizations.

This project, titled **“Crime Analytics and Prediction System”**, leverages the power of **Exploratory Data Analysis (EDA)** to study and interpret sales data efficiently. The objective of the project is not only to build a system that displays and manages sales records but also to provide valuable insights into sales performance through data exploration. Python, with its rich ecosystem of data analysis libraries such as **Pandas**, **Matplotlib**, and **Seaborn**, offers an ideal platform for implementing EDA-based solutions.

### Key goals of the project include:

* Reading and preprocessing crime data for accurate and structured analysis.
* Performing statistical and exploratory analysis to summarize crime trends.
* Developing machine learning models to predict crime types based on various features.
* Visualizing crime patterns geographically using interactive maps.
* Supporting law enforcement and public safety efforts through data-driven insights and dashboards.

Unlike traditional record-keeping systems that simply store and retrieve crime data, this project emphasizes the analysis and interpretation of data to generate actionable insights. Through exploratory data analysis (EDA) techniques, we move beyond basic listings to uncover relationships, identify hotspots, detect outliers, and understand crime behavior across various dimensions such as time, location, and crime category.

This system is particularly useful for:

* Law enforcement agencies aiming to make data-informed operational decisions.
* Analysts and students exploring the intersection of data science and public safety.
* Policy makers and municipal authorities seeking a visual summary of crime statistics and patterns.

Through this project, users not only enhance their Python programming skills but also build a strong foundation in interpreting real-world data. The integration of EDA, machine learning, and visualization transforms raw crime records into compelling stories and predictive insights that can aid in crime prevention and resource planning.

In conclusion, the Crime Analytics and Prediction System demonstrates how data science methodologies can enhance public safety strategies by turning unstructured crime data into valuable knowledge for smarter governance.

**2. SOURCE OF DATASET**

The dataset utilized in this project was obtained from the **U.S. Government’s Open Data Platform** – [**catalog.data.gov**](https://catalog.data.gov), which serves as a comprehensive repository of datasets across various sectors including business, economics, health, and environment. The specific dataset used in this project is titled:

**“Retail Sales of Goods by Kind of Business: Total and E-commerce”**  
**Dataset URL:** <https://opendata.dc.gov/api/download/v1/items/c5a9f33ffca546babbd91de1969e742d/csv?layers=6>

The crime dataset used in this project is sourced from Kaggle and government portals like data.gov.in. It contains comprehensive records of crime incidents reported across various districts. Each entry includes fields such as date, time, district, location coordinates, crime type, and status, which are vital for geospatial and temporal analysis.

Rationale for Choosing This Dataset: This dataset was selected due to its suitability in analyzing:

* Trends in crime over time (daily, monthly, seasonal).
* High-crime zones and district-wise crime comparison.
* The relationship between time of day and crime occurrence.
* The spatial distribution of different types of crimes.

Preprocessing and Enrichment: To prepare the dataset for analysis:

* Data Cleaning: Missing values and inconsistencies were handled using Pandas.
* Date-Time Formatting: Timestamps were converted to datetime format for trend analysis.
* Feature Engineering: Extracted features such as hour, weekday, and month to enhance temporal analysis.
* Categorical Encoding: Standardized district names and encoded crime categories for machine learning.
* Geo-Spatial Preparation: Latitude and longitude fields were cleaned for accurate map visualizations.

Benefits of This Dataset for Crime Analysis: This dataset offers valuable opportunities for exploratory data analysis and predictive modeling. It supports:

* Visualization of crime trends using line graphs, bar plots, and heatmaps.
* Identification of crime hotspots through interactive maps.
* Comparative study of crime rates across different districts.
* Detection of outliers and anomalies in crime reports.

By applying EDA and machine learning techniques, the project derives meaningful insights from crime data, enabling better understanding and strategic planning for public safety.

**3. DATASET PREPROCESSING**

To ensure the dataset was suitable for analysis, a systematic data preprocessing phase was carried out. The raw dataset, sourced from a government crime portal, contained monthly crime incident figures across various neighbourhoods and crime categories. Upon loading the dataset, an initial review was conducted to understand the structure, format, and completeness of the data. This review revealed several inconsistencies and missing values that needed to be addressed.

The first step in preprocessing involved handling missing data. A detailed check was performed to identify any null or incomplete entries. Depending on the nature and significance of the missing values, different imputation techniques were used. For example, in time-series columns (such as the date of occurrence), missing entries were filled based on previously observed values (forward fill) to maintain trend continuity. For numerical columns (like crime counts or incident rates), mean or median imputation was applied to preserve the dataset's statistical balance. If certain rows or columns contained excessive missing data and did not contribute meaningfully to the analysis, they were removed.

Next, data cleaning was conducted. Redundant columns that did not offer analytical value (such as unnecessary ID fields or irrelevant location details) were dropped. Column names were reformatted for consistency—removing special characters, converting to lowercase, and making names more readable. In cases where categorical data entries showed inconsistencies (e.g., varied naming for crime types), standardization was applied to unify them. This helped avoid duplication and ensured that grouping and filtering operations would yield accurate results.

Data type validation and conversion were essential parts of the preprocessing phase. Date fields were converted into a standard datetime format to support chronological sorting and time-based analysis. Numeric fields, such as crime counts and incident rates, were checked to ensure all values were in the correct format and free of unexpected characters or text, which could interfere with computations. Ensuring correct data types allowed for smooth statistical operations and reliable visual representations.

To further enrich the dataset, feature engineering techniques were applied. New columns were created to support deeper analysis. For example, month and year were extracted from date entries to allow for monthly trend analysis in crime incidents. Additionally, changes in crime rates over months, crime type contributions, and cumulative crime figures were computed to highlight hidden patterns in the data. These derived features made it easier to compare and contrast crime activity across different neighborhoods and time periods.

Finally, the dataset was sorted and filtered to facilitate focused analysis. Locations were grouped based on crime rate, and the dataset was rearranged to highlight areas with the highest crime incidents. Outlier detection was also performed to identify unusual spikes or drops in crime incidents, which were examined further to understand their impact. Once the dataset was fully prepared, it was stored in a structured format, ready for visualization and exploratory data analysis.

This preprocessing phase was vital in transforming the raw dataset into a high-quality, analysis-ready format. It ensured that the data was not only accurate and complete but also tailored for meaningful insights and decision-making. Proper preprocessing greatly enhanced the reliability of the results obtained during subsequent stages of the project.

**4. ANALYSIS ON DATASET**

**Objective 1: Total Crime Incidents by Crime Type**

**i. General Description**  
This objective aims to calculate the total number of crime incidents for each crime type across the entire dataset. Each entry in the dataset includes a crime type and its corresponding count, allowing for aggregation and comparison. This analysis is fundamental in understanding which crime types contribute most to overall crime activity and helps law enforcement or government agencies identify high-crime areas and take necessary actions.

**ii. Specific Requirements**The specific goal here is to:  
• Group the dataset by crime type.  
• Calculate the sum of incidents for each crime type.  
• Sort the crime types based on total incidents to identify the most frequent and least frequent crimes.  
• Compare the crime figures among all crime types in a visually understandable format.

This analysis helps law enforcement focus on high-priority crime types and allocate resources more effectively.

**iii. Analysis Results**The aggregated crime analysis showed that certain crime types, such as Burglary, Theft, and Assault, contributed to the highest number of incidents over the recorded period. These crimes made up a significant portion of the overall crime activity. On the other hand, crimes like Vandalism and Fraud showed comparatively lower incident numbers, which could suggest either lower occurrence rates or underreporting.

This breakdown provides a clear picture of how different crime types are distributed across the dataset and sets the stage for deeper analysis, such as identifying trends over time or pinpointing crime hotspots.

**iv. Visualization**To support this analysis, the following visualizations were created:  
• Bar Chart: Displayed the total crime incidents for each crime type, making it easy to compare their frequency at a glance.

**Objective 2: Identifying High-Crime Areas**

**i. General Description**This objective focuses on identifying the areas with the highest crime rates based on total incidents reported. Understanding which neighbourhoods or districts experience the most crime can help law enforcement prioritize patrols and implement targeted interventions.

**ii. Specific Requirements**This analysis required:  
• Group the dataset by location (e.g., neighbourhood, district, or precinct).  
• Calculate the total number of crime incidents for each location.  
• Rank the locations based on total incidents to find the areas with the highest crime rates.  
• Create a visualization to highlight these high-crime locations.

This insight is valuable for deploying resources and planning interventions to reduce crime in high-risk areas.

**iii. Analysis Results**From the analysis, it was observed that certain locations, such as Downtown and the Industrial District, had significantly higher crime rates compared to others. These areas might require more law enforcement attention or targeted crime prevention strategies. In contrast, residential areas showed relatively lower crime rates, suggesting a need for different approaches in crime prevention.

**iv. Visualization**To visualize the crime hotspots:  
• A bar chart was created showing the total number of incidents per location.  
• The chart clearly highlighted the top-performing areas (i.e., the areas with the highest crime rates).  
• Color-coding and proper labeling enhanced clarity and presentation**.**

**Objective 3: Gender Breakdown of Crime Suspects**

**i. General Description**This objective aims to understand the gender distribution of crime suspects and compare how male and female suspects contribute to total crime incidents. This analysis will provide insights into potential demographic patterns in criminal activity.

**ii. Specific Requirements**For this analysis, the following steps were performed:  
• Group the dataset based on gender of the suspect.  
• Count the total number of unique suspect IDs for each gender.  
• Analyze the number of crimes committed by male and female suspects.  
• Compare gender distribution visually using pie charts for both count of suspects and crime incidents.

These insights can help law enforcement target appropriate outreach programs or public awareness campaigns.

**iii. Analysis Results**The gender-wise breakdown revealed that males represented the majority of crime suspects in the dataset, contributing to a significantly higher number of total incidents. Females, though fewer in number, still contributed to a considerable portion of criminal activity. These findings suggest that crime prevention programs may need to be more tailored towards the demographics with the higher incidence of crimes.

**iv. Visualization**To illustrate this analysis:  
• A pie chart was used to show the proportion of male and female suspects.  
• A second pie chart represented the gender breakdown of total crime incidents.  
• Colours were used to differentiate the gender segments clearly, aiding in quick interpretation.

**Objective 4: Crime Trends Over Time**

**i. General Description**This objective explores crime patterns over time to identify trends, seasonal spikes, or drops in criminal activity. Understanding crime trends is essential for resource allocation, emergency response planning, and preventive measures.

**ii. Specific Requirements**To carry out this objective, the following steps were taken:  
• Group the dataset by the date of the crime incident.  
• Calculate the total number of incidents for each date to reflect daily crime rates.  
• Generate a line graph to visualize the crime trend over time.

This time-series analysis provides insights into patterns of criminal activity, including peak crime periods and potential causes behind fluctuations in crime rates.

**iii. Analysis Results**The analysis revealed fluctuating crime rates over time, with significant spikes on certain days, possibly linked to local events or holidays. The trendline displayed a rising pattern over time, suggesting that crime rates increased gradually, which could be due to factors like increased population density or reduced law enforcement presence.

**iv. Visualization**To visually represent the crime trends:  
• A line chart was used, with the x-axis representing dates and the y-axis showing the total number of incidents.  
• The chart clearly illustrated periods of high and low crime activity, making it easy to understand crime fluctuations over time.

**Objective 5: Crime Trend by Category Over Time**

**i. General Description**This objective analyse how different types of crime perform each month, identifying seasonal patterns or spikes in specific crime types, which could be crucial for planning interventions and resource distribution.

**ii. Specific Requirements**The following steps were undertaken to perform this analysis:  
• Extract month and year from the crime date to create a chronological sequence.  
• Group the dataset by both month and crime category, calculating the number of incidents for each category per month.  
• Use a line graph to track monthly crime trends by category.

**iii. Analysis Results**Certain crime types, such as Theft and Assault, showed higher rates during specific months, possibly coinciding with holidays or local events. Other crime types, like Vandalism, remained fairly stable year-round. This helps prioritize areas where crime prevention campaigns may be necessary, particularly during peak months for certain crime types.

**iv. Visualization**To visualize these trends:  
• A multi-line chart was created, with each line representing a different crime type over the months.  
• Markers and colour coding enhanced the clarity of the trends, making it easy to differentiate between categories.

**Objective 6: Relationship Between Crime Rate and Location**

**i. General Description**This objective explores how the crime rate correlates with different types of locations (e.g., residential, commercial, public areas) and whether crime rates vary based on the location.

**ii. Specific Requirements**For this objective, two analyses were performed:

1. Scatter plot with regression line to observe if there is a correlation between crime rate and location type.
2. Heatmap of crime density by location to visually represent areas with high crime rates.

**iii. Analysis Results**The scatter plot showed a modest positive correlation between certain location types (e.g., commercial areas) and crime rates, suggesting that more public or commercial spaces might attract higher crime rates. The heatmap clearly highlighted areas with high crime concentrations, offering insights for focused law enforcement intervention.

**iv. Visualization**

1. Crime Rate vs Location:  
   • A scatter plot was created with a regression line to observe the relationship.  
   • Crime rate was plotted on the y-axis and location type on the x-axis.
2. Heatmap of Crime Density:  
   • A heatmap was generated to show crime density by location.  
   • Brighter colours indicated higher crime rates, highlighting the most affected areas.

# ****5. CONCLUSION****

This project focused on conducting a comprehensive exploratory data analysis (EDA) on a crime analytics dataset using Python. The goal was to derive meaningful insights from the dataset, visualize trends, and inform strategic decisions related to crime prevention and resource allocation. The analysis was carried out using Python libraries like Pandas, Matplotlib, and Seaborn.

**Key Findings:**

1. **Crime Performance by Category:** The analysis highlighted the crime categories with the highest incident counts. Categories such as **violent crimes** and **property crimes** showed high frequencies, helping law enforcement prioritize areas with the most significant public safety risks.
2. **Crime Trends Over Time:** The temporal analysis revealed fluctuations in crime activity over different months. Peaks in crime were identified, possibly linked to specific events or seasons. These trends help agencies plan for periods of increased crime and allocate resources accordingly.
3. **Crime Severity by Location:** The analysis of crime incidents by location and type of crime allowed for identifying high-risk zones. This information can guide policy-making, resource allocation, and targeted crime reduction efforts in specific neighbourhoods or districts.
4. **Gender-Based Insights in Crime Victimization:** Gender-based insights revealed patterns of victimization, showing differences in crime types experienced by males and females. This helps shape awareness campaigns and initiatives designed to protect vulnerable groups.
5. **Outlier Detection in Crime Data:** Through outlier detection, the analysis identified irregular spikes in crime, possibly linked to anomalies like special events or data reporting errors. Addressing these outliers is essential for ensuring accurate trend forecasting.
6. **Correlation Between Crime Types:** A mild correlation between certain crime types was observed, suggesting patterns of co-occurrence. For example, property crimes and violent crimes might often happen together, which can inform joint intervention strategies.

**Overall Impact:**

By performing EDA on this crime dataset, this project was able to transform raw crime data into actionable insights for law enforcement and policymakers. The visualizations created helped communicate findings clearly and facilitated understanding of patterns across various crime categories and locations.

The findings are crucial for:

* **Resource allocation** for crime prevention in high-risk areas.
* **Crime reduction strategies** based on patterns of behaviour.
* **Targeting specific crime types** with tailored approaches.
* **Public awareness campaigns** to educate and protect vulnerable demographics.
* **Data-driven decision-making** for law enforcement agencies.

**Final Thoughts:**

This project underscores the importance of data analytics in criminal justice and public safety. By using Python and its data science ecosystem, crime data can be effectively analyzed to uncover actionable insights. The project also demonstrates the value of real-world datasets in making informed decisions that can improve societal safety.

Going forward, the analysis can be extended by incorporating predictive models, such as crime forecasting, and applying machine learning techniques for more advanced insights into crime patterns and prevention strategies.

**6. FUTURE SCOPE**

The current project has successfully demonstrated the potential of Exploratory Data Analysis (EDA) in extracting meaningful insights from crime analytics data. However, there are several opportunities to expand and deepen this analysis in future work. These enhancements would not only add more value to the current findings but also provide more advanced decision-support tools for law enforcement and policymakers.

**1. Integration of Machine Learning Models**

In the future, this project can be extended by incorporating predictive analytics using machine learning algorithms. Models such as:

* **Classification models** (e.g., Decision Trees, Random Forest) to predict crime likelihood based on location, time, and type.
* **Clustering techniques** (like K-Means) to identify crime hotspots and trends within specific geographic regions.
* **Anomaly detection models** to flag unusual crime patterns or spikes that may warrant immediate attention.

These techniques can help predict crime patterns, improve resource allocation, and enhance crime prevention strategies.

**2. Real-Time Data Integration**

Currently, the project uses static historical data. Future versions can be improved by integrating real-time crime data streams, which will allow:

* **Live crime tracking** and updates.
* **Instant alerts** for spikes in specific crime types or areas.
* **Real-time feedback analysis** on law enforcement actions and public safety measures.

This would enable quicker responses to emerging crime patterns and more dynamic decision-making.

**3. Dashboard Development**

Creating interactive dashboards using tools like **Tableau**, **Power BI**, or **Plotly Dash** can enhance the user experience by allowing law enforcement agencies and policymakers to explore crime data visually and interactively. Dashboards can:

* Filter data by time, location, crime type, etc.
* Visualize key performance indicators (KPIs) such as crime reduction rates, arrest success, or crime recurrence.
* Provide **drill-down** analysis features for deeper insights into specific regions or time periods.

**4. Geographical Crime Analysis**

If location-based data is available, adding **geospatial analysis** could highlight:

* **Regional differences** in crime performance.
* **High-risk areas** for targeted policing or community engagement.
* **Area-wise crime patterns** to support planning and intervention.

This can assist in resource planning, pinpointing crime hotspots, and formulating targeted community safety programs.

**5. Enhanced Crime Analysis**

Future developments can focus more deeply on crime patterns and behaviour, such as:

* **Crime recidivism** analysis to understand repeat offenders.
* **Offender profiling** based on crime type, location, and time.
* **Predicting crime spikes** based on historical data and emerging patterns.

These insights can assist law enforcement agencies in improving crime prevention strategies and community outreach programs.

**6. Time Series and Seasonal Trend Modelling**

While the current analysis provides insights into general crime trends, more advanced **time series forecasting models** (like **ARIMA**, **Prophet**) can be used to:

* Predict future crime rates and trends, factoring in seasonality and long-term patterns.
* Plan for resource allocation during high-crime periods (e.g., holidays, events).
* Anticipate emerging crime trends based on historical data.

**7. Recommendation Systems for Law Enforcement**

Based on historical crime data, location, and offender behaviour, recommendation systems could:

* **Suggest patrol routes** and resource deployment strategies.
* **Personalize policing tactics** to match emerging crime patterns.
* Enhance **community policing efforts** by targeting areas with higher chances of crime incidents.

**8. Multivariate Analysis**

Applying multivariate statistical methods can help analyze the combined effect of multiple variables like location, time of day, type of crime, and offender characteristics. This can uncover deeper patterns and interactions, such as:

* How certain variables influence crime rates in different neighbourhoods.
* Which factors are predictive of high-crime areas or times.

**9. Incorporation of External Factors**

External datasets such as **weather patterns**, **economic indicators**, or **social events** can be integrated to provide a more holistic view of crime performance. This would make the analysis more robust and context-aware, allowing law enforcement to better anticipate crime trends based on these external factors.

**10. Automation of Reporting**

Finally, automating the EDA and reporting process using Python scripts or workflow tools like **Apache Airflow** can:

* Save time in routine crime data analysis.
* Ensure **consistency** and **reproducibility** in crime reports.
* Enable **scheduled updates** and notifications, providing timely insights to authorities.

**Conclusion of Future Scope**

In summary, this project lays a solid foundation for data-driven decision-making in crime prevention and law enforcement. By leveraging advanced analytics, real-time systems, and interactive dashboards, the project can evolve into a comprehensive tool for improving public safety. These future enhancements can make a significant impact on crime reduction, resource management, and community safety efforts.

By incorporating predictive analytics, geographic insights, and real-time data, the project can provide authorities with better tools for proactive crime prevention and faster responses to emerging crime patterns.

**7.REFERENCES**

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Python*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.

[2] M. Waskom et al., “Seaborn: Statistical data visualization,” *Journal of Open-Source Software*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: <https://doi.org/10.21105/joss.03021>

[3] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[4] R. G. Brown and R. F. Meyer, “Time Series Analysis and Forecasting of Crime Incidents,” *International Journal of Forecasting*, vol. 1, no. 1, pp. 25–38, 1985.  
[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016, pp. 785–794.

[6] P. Raj and A. C. Pattabhi, *Data Analytics for Public Safety*, Packt Publishing, 2017.

[7] A. Anuradha and D. N. Kumar, “Crime Pattern Recognition and Predictive Analytics in Law Enforcement using Machine Learning,” *2021 3rd Int. Conf. on Signal Processing and Communication (ICPSC)*, Coimbatore, India, 2021, pp. 188–192.

[8] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2014.

**8.Snap Shots of the project**

**LOAD AND READING DATA**

