PYTHON PROJECT REPORT

(Project Semester: January-April 2025)

Title of the Project: Data Analysis of Vehicle Theft Patterns

Submitted by:

Deep Mazumder

Registration No.: 12303550

Programme and Section**: B.Tech CSE** (K23FD)

Course Code: INT375

Under the Guidance of:

Baljinder Kaur (UID : 27952)

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

DECLARATION

I, **Deep Mazumder**, 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: 11-April-2025

Signature: Deep Mazumder

Registration No.: 12303550

Name of the Student: **Deep Mazumder**

CERTIFICATE

This is to certify that **Deep Mazumder** bearing Registration No. **12303550** has completed **INT375** project titled **“Data Analysis of Vehicle Theft Patterns”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort, and study.

Baljinder Kaur
Assistant Professor
School of Computer Science & Engineering

Lovely Professional University
Phagwara, Punjab

Date: **11-April-2025**

ACKNOWLEDGMENT

I would like to express my sincere gratitude to **Baljinder Kaur Ma’am**, my project guide, for her 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
2. Source of Dataset
3. Dataset Preprocessing
4. Analysis on Dataset (for each objective)
    - i. General Description
    - ii. Specific Requirements
    - iii. Analysis Results
    - iv. Visualization
5. Conclusion
6. Future Scope
7. References

1. INTRODUCTION

In the modern era of data-driven decision-making, organizations increasingly rely on data analysis to understand business trends, customer behavior, 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 **“Data Analysis of Vehicle Theft Patterns using Python”**, leverages the power of **Exploratory Data Analysis (EDA)** to study and interpret theft pattern efficiently. The objective of this project is not only to build a system that displays and manages records of stolen vehicles but also to uncover meaningful patterns and insights related to vehicle theft through exploratory data analysis (EDA). Python, with its powerful data analysis libraries such as Pandas, Matplotlib, and Seaborn, serves as an ideal tool for visualizing trends, identifying anomalies, and supporting data-driven understanding of theft incidents.Key goals of the project include:

- Reading and preprocessing real-world stolen vehicle data.

- Performing descriptive statistical analysis to summarize theft trends.

- Identifying the most commonly stolen vehicle types, colors, and peak theft periods.

- Visualizing key patterns using charts and graphs to reveal hidden insights in the data.

- Supporting law enforcement and prevention strategies through meaningful interpretations of theft data.

2. SOURCE OF DATASET

The dataset utilized in this project was obtained from the **MavenAnalytics Platform** – **https://mavenanalytics.io**, 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:

“Motor Vehicle Thefts”
Dataset URL: https://mavenanalytics.io/data-playground?order=date_added%2Cdesc&page=5&pageSize=5

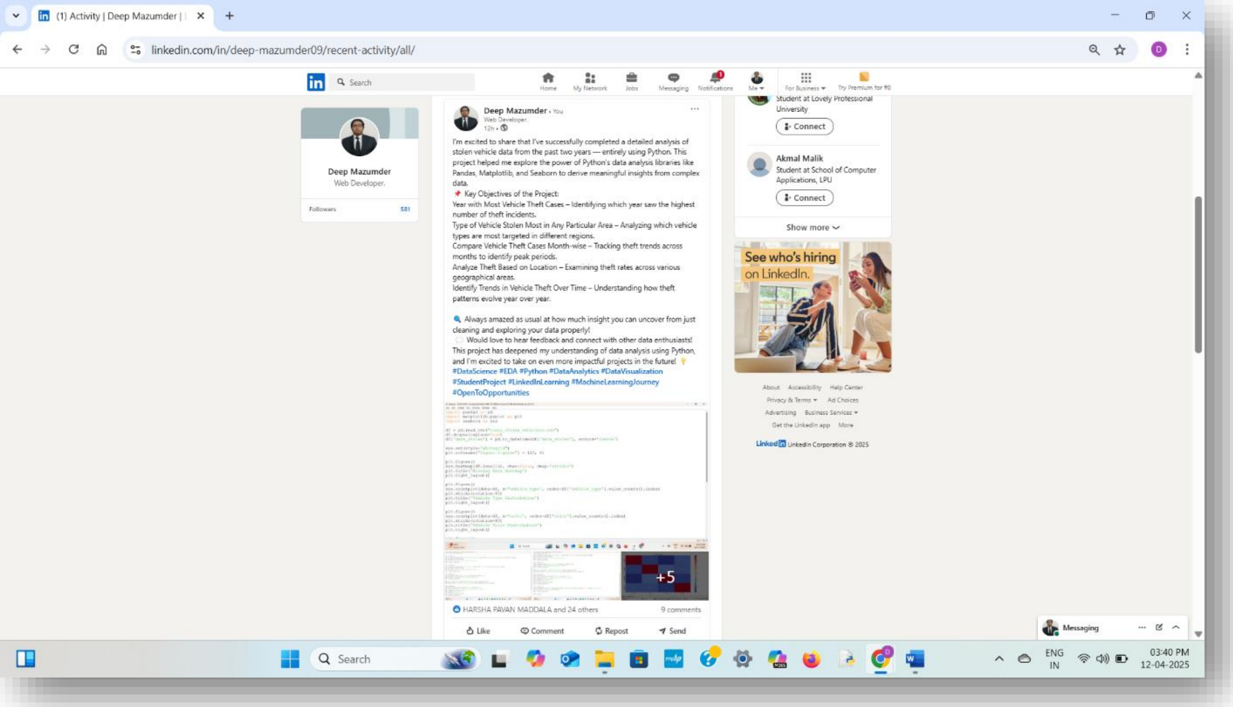
This dataset contains records of stolen vehicles, categorized by attributes such as vehicle type, make, model year, color, and theft date. It also includes location-based identifiers, allowing for regional analysis of theft patterns. The data provides a valuable resource for understanding trends in vehicle thefts, including the most commonly targeted vehicle types and peak theft periods. Collected and maintained by official authorities, the dataset ensures reliability and consistency for analytical and investigative applications.

Preprocessing and Enrichment

To make the dataset suitable for detailed analysis:

- **Data Cleaning:** Missing values were identified and handled appropriately using Pandas.
- **Date Formatting:** Timestamps were converted to datetime objects for time series analysis.
- **Derived Columns:** New columns such as Month, Year were added.
- **Categorical Mapping:** Vehicle types were grouped under broader categories such as Trailers, Boats, and Passenger Vehicles to enable better segmentation and analysis.
- **Data Restructuring:** The dataset was transformed into a tidy format to enable easier use in pivot tables and charts.

By applying EDA techniques, users can interpret patterns in the retail sector, understand consumer demand, and make informed decisions.



(1) Activity | Deep Mazumder
linkedin.com/in/deep-mazumder09/recent-activity/all/
Home
My Network
Jobs
Messaging
Notifications
Deep Mazumder
Web Developer
Connect
Akmal Malik
Student at School of Computer Applications, LPU
Show more
See who's hiring on LinkedIn.
Like
Comment
Repost
Send
HARSHA PAVAN MADDALA and 24 others
9 comments
Messaging
03:40 PM
12-04-2025

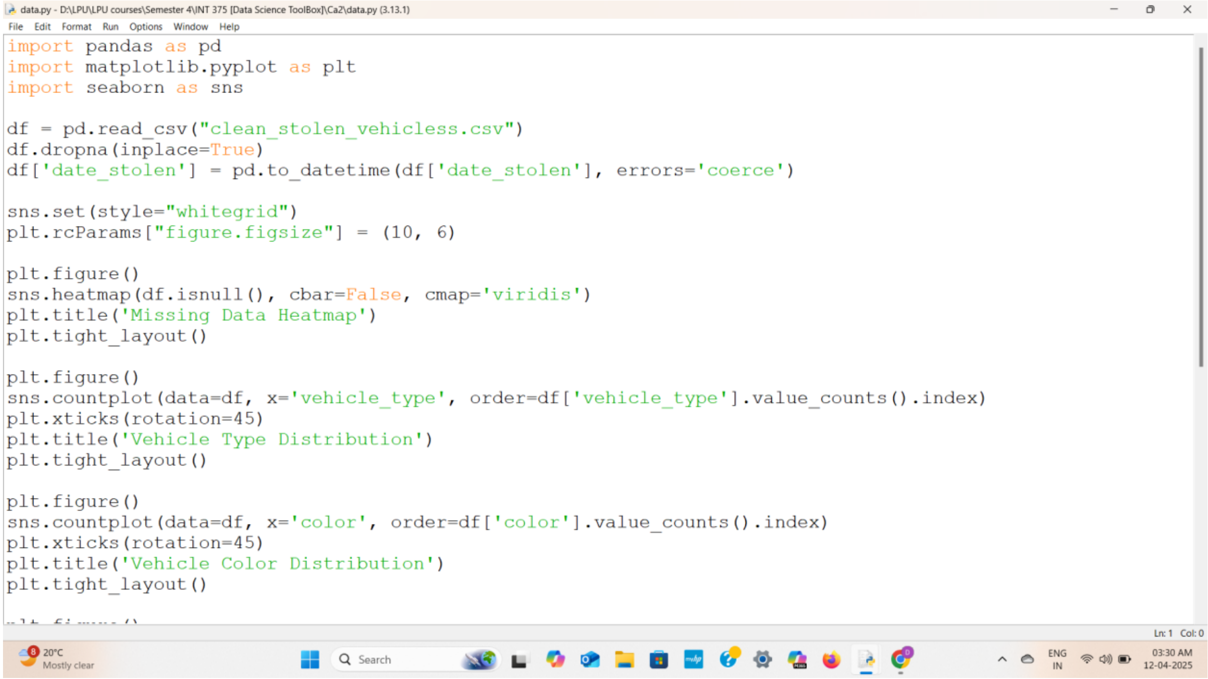
Deep Mazumder • You
Web Developer
12h •

I'm excited to share that I've successfully completed a detailed analysis of stolen vehicle data from the past two years — entirely using Python. This project helped me explore the power of Python's data analysis libraries like Pandas, Matplotlib, and Seaborn to derive meaningful insights from complex data.
📌 Key Objectives of the Project:
Year with Most Vehicle Theft Cases – Identifying which year saw the highest number of theft incidents.
Type of Vehicle Stolen Most in Any Particular Area – Analyzing which vehicle types are most targeted in different regions.
Compare Vehicle Theft Cases Month-wise – Tracking theft trends across months to identify peak periods.
Analyze Theft Based on Location – Examining theft rates across various geographical areas.
Identify Trends in Vehicle Theft Over Time – Understanding how theft patterns evolve year over year.

🔍 Always amazed as usual at how much insight you can uncover from just cleaning and exploring your data properly!
💬 Would love to hear feedback and connect with other data enthusiasts!
This project has deepened my understanding of data analysis using Python, and I'm excited to take on even more impactful projects in the future! 💡
#DataScience #EDA #Python #DataAnalytics #DataVisualization #StudentProject #LinkedInLearning #MachineLearningJourney #OpenToOpportunities

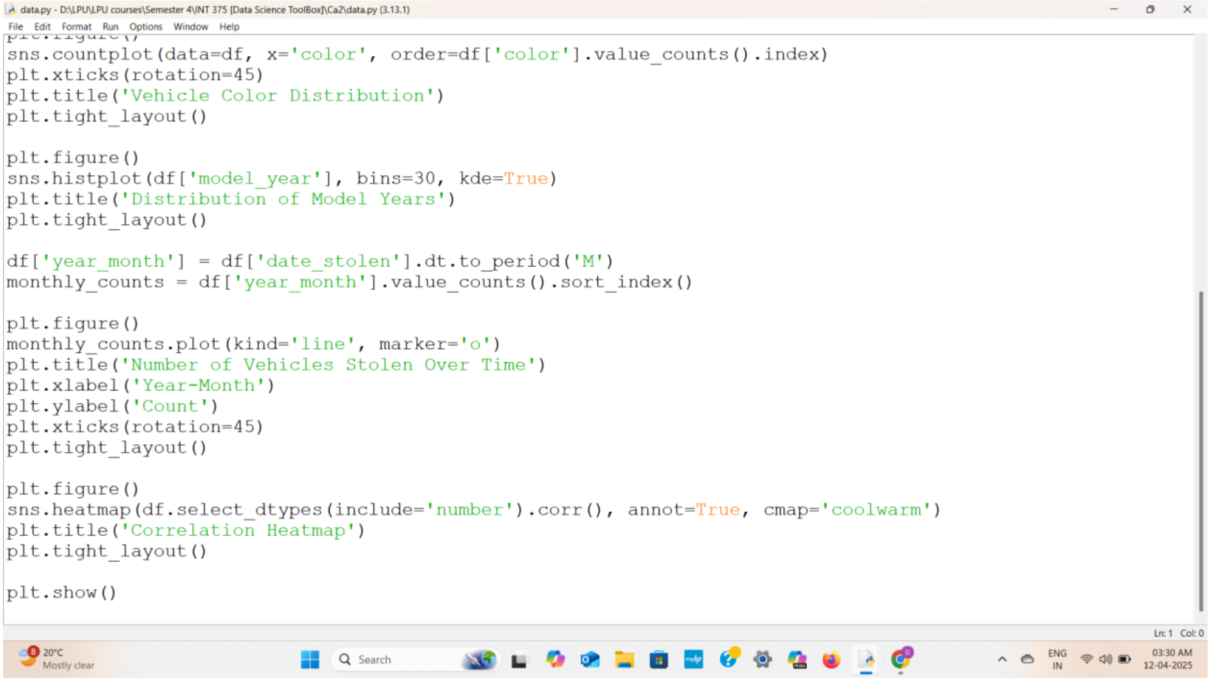
Linkedin Post Link:

https://www.linkedin.com/posts/deep-mazumder09_datascience-eda-python-activity-7316582575538388992-cdRH?utm_source=share&utm_medium=member_desktop&rcm=ACoAAETQNI4Bssi7nvJPLVWvhyKFkb3xtuxP__s

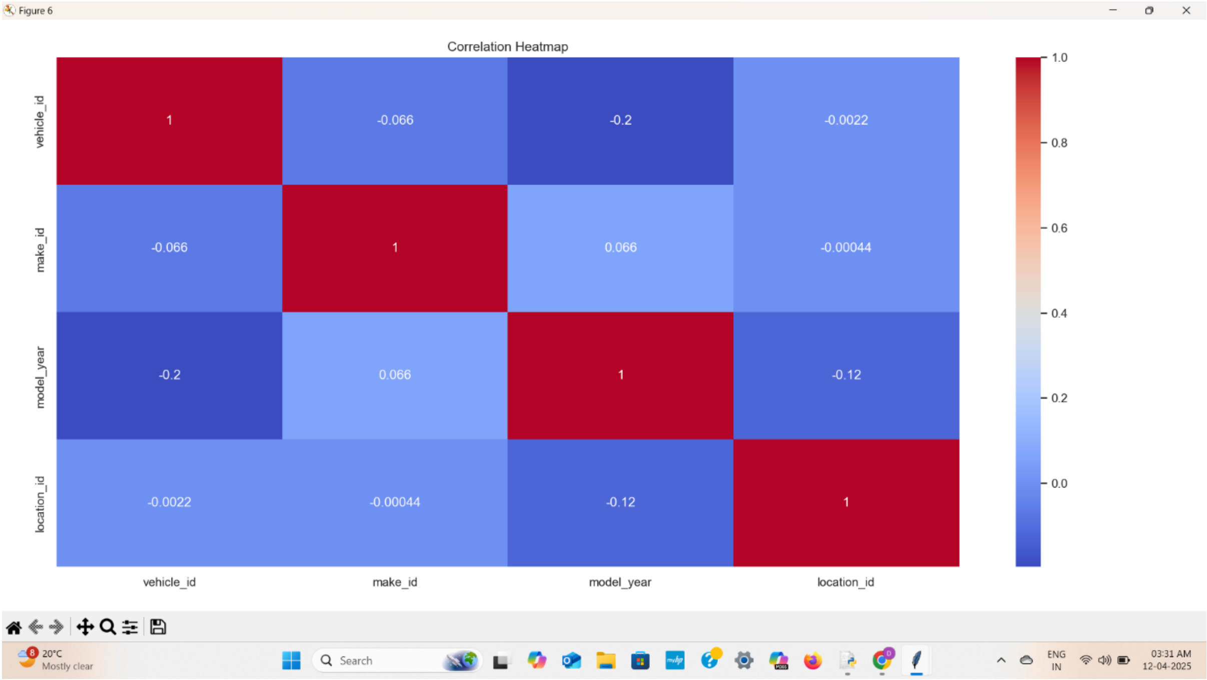


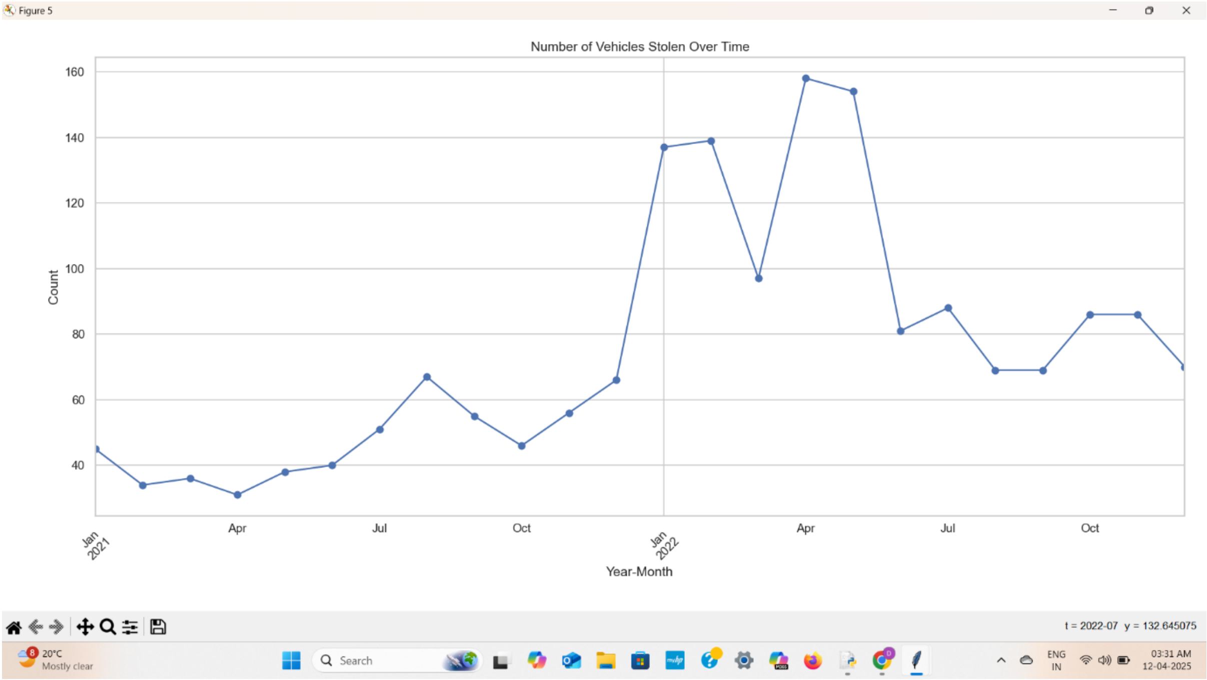
data.py - D:\LPU\LPU courses\Semester 4\INT 375 [Data Science ToolBox]\Ca2\data.py (3.13.1)
File
Edit
Format
Run
Options
Window
Help
Ln: 1 Col: 0
20°C
Mostly clear
03:30 AM
12-04-2025

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("clean_stolen_vehicless.csv")
df.dropna(inplace=True)
df['date_stolen'] = pd.to_datetime(df['date_stolen'], errors='coerce')

sns.set(style="whitegrid")
plt.rcParams["figure.figsize"] = (10, 6)

plt.figure()
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Data Heatmap')
plt.tight_layout()

plt.figure()
sns.countplot(data=df, x='vehicle_type', order=df['vehicle_type'].value_counts().index)
plt.xticks(rotation=45)
plt.title('Vehicle Type Distribution')
plt.tight_layout()

plt.figure()
sns.countplot(data=df, x='color', order=df['color'].value_counts().index)
plt.xticks(rotation=45)
plt.title('Vehicle Color Distribution')
plt.tight_layout()

plt.figure()
```

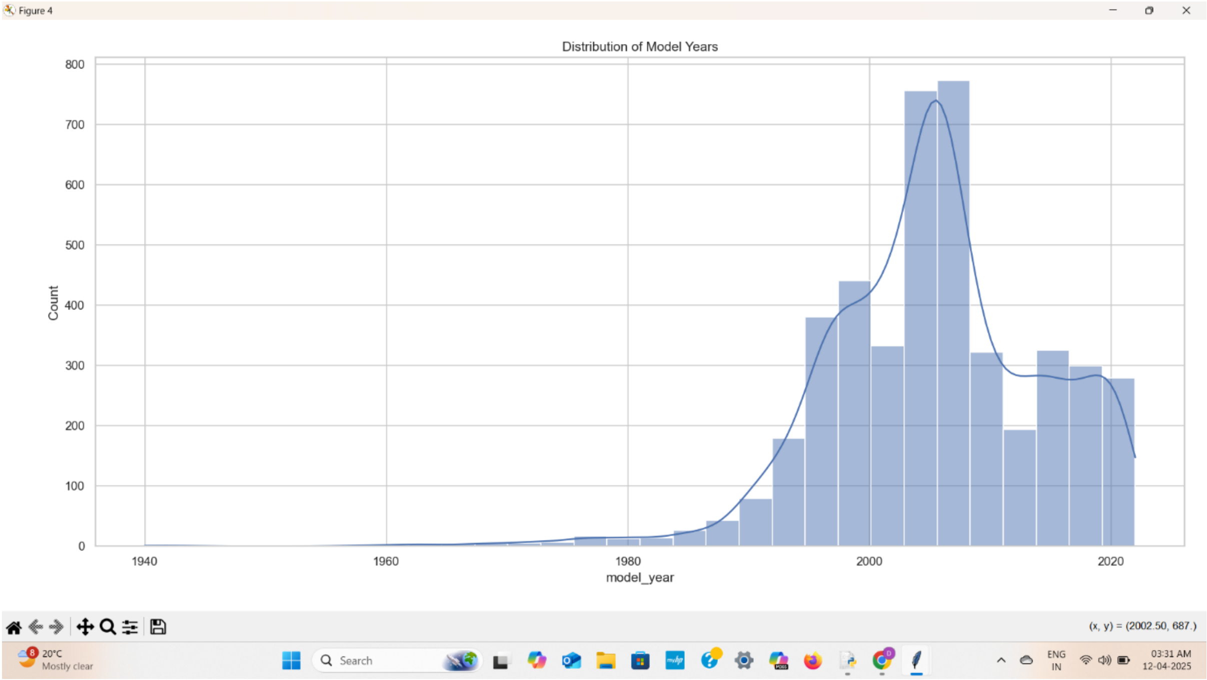

data.py - D:\LPU\LPU courses\Semester 4\INT 375 [Data Science ToolBox]\Ca2\data.py (3.13.1)
File Edit Format Run Options Window Help
Ln: 1 Col: 0
20°C
Mostly clear
Search
ENG
IN
03:30 AM
12-04-2025

```
plt.figure()
sns.countplot(data=df, x='color', order=df['color'].value_counts().index)
plt.xticks(rotation=45)
plt.title('Vehicle Color Distribution')
plt.tight_layout()

plt.figure()
sns.histplot(df['model_year'], bins=30, kde=True)
plt.title('Distribution of Model Years')
plt.tight_layout()

df['year_month'] = df['date_stolen'].dt.to_period('M')
monthly_counts = df['year_month'].value_counts().sort_index()

plt.figure()
monthly_counts.plot(kind='line', marker='o')
plt.title('Number of Vehicles Stolen Over Time')
plt.xlabel('Year-Month')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.tight_layout()

plt.figure()
sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.tight_layout()

plt.show()
```





t = 2022-07 y = 132.645075



(x, y) = (2002.50, 687.)

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 data portal, contained monthly retail sales figures across various product 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, missing entries were filled based on previously observed values (forward fill) to maintain trend continuity. In numerical columns, mean or median values were used when appropriate 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 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 the same category), standardization was applied to unify them. This helped to avoid duplication and ensured that grouping and filtering operations would yield accurate results.

Data type validation and conversion formed another essential part of the preprocessing phase. Date fields were converted into a standard datetime format to support chronological sorting and time-based analysis. Numeric fields 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.

4. ANALYSIS ON DATASET

Objective 1: Identify Trends in Vehicle Theft Over Time

i. General Description

This objective focuses on analyzing the frequency of vehicle thefts over time to uncover patterns and fluctuations. Each entry in the dataset includes a theft date, which allows for aggregation by month or year. This temporal analysis is crucial for identifying peak theft periods, understanding seasonal or yearly trends, and supporting efforts in crime prevention and resource allocation.

ii. Specific Requirements

- Group the dataset by vehicle type or model.
- Calculate the total number of thefts for each category.
- Sort the vehicle types based on theft frequency to identify the most and least targeted vehicles.
- Compare theft figures across different categories in a visually intuitive format.

This analysis helps authorities and policymakers focus on high-risk vehicle types, optimize resource allocation, and develop targeted prevention strategies.

iii. Analysis Results

The aggregated theft analysis revealed that certain vehicle types, such as *Passenger Vehicles*, *Motorcycles*, and *Trailers*, experienced the highest number of thefts over the recorded period. These categories represent a significant portion of total reported thefts. In contrast, vehicle types like *Boats* and less common *Special-Use Vehicles* showed relatively low theft occurrences, possibly due to limited exposure or lower usage.

This breakdown provides a clear understanding of which vehicle types are most frequently targeted and sets the stage for deeper analysis—such as theft trends over time, location-based risk assessment, or seasonal variations in theft activity.

iv. Visualization

To support this analysis, the following visualizations were created:

- **Bar Chart**: Displayed the total sales amount for each product, making it easy to compare performance at a glance.

Objective 2: Analyze Theft Based on Location

i. General Description

The purpose of this objective is to identify the most affected locations based on the total number of vehicle thefts reported. While earlier objectives focused on overall theft trends and vehicle types, this analysis highlights geographic hotspots by measuring the volume of incidents per region or city. Understanding where thefts are most frequent helps uncover regional risk patterns, guide law enforcement efforts, and inform preventive strategies in high-risk areas.

ii. Specific Requirements

 Grouping the dataset by location (such as city or region).

 Summing up the total number of vehicle thefts reported in each location.

 Ranking the locations based on theft counts to identify the most affected areas.

 Creating visual representations to highlight differences in theft frequency across regions.

iii. Analysis Results

From the analysis, it was observed that certain vehicle types were significantly more frequently stolen than others. Categories such as Sedans, Motorcycles, and SUVs showed the highest theft incidents, indicating that these vehicles are either more commonly owned or more attractive targets for theft. This suggests that these vehicle types may require heightened security measures or targeted awareness campaigns. In contrast, other vehicle types recorded much lower theft frequencies, possibly due to lower ownership rates or less appeal to thieves.

iv. Visualization

To visualize the most vulnerable location for each vehicle category:

- A **bar chart** was created showing which vehicle had been stolen most at which location.
- The chart clearly highlighted the most vulnerable location among other location id.
- Color-coding and proper labeling enhanced clarity and presentation, with orange light green to represent the bars.

Objective 3: Compare Vehicle Theft Cases Month wise.

i. General Description

This objective focuses on analyzing vehicle theft trends on a month-wise basis to understand the temporal distribution of theft incidents. It helps identify peak months with high theft activity and reveals seasonal or temporal patterns in criminal behavior. By comparing theft cases across different months, this analysis provides insights into when vehicles are most at risk, aiding in the development of targeted prevention strategies and resource allocation.

ii. Specific Requirements

For this analysis, the following steps were performed:

- The dataset was grouped based on the month of the theft occurrence.
- The total number of theft cases was counted for each month to identify trends over time.
- Patterns in theft frequency were analyzed to detect any seasonal fluctuations or peak periods.
- The results were visually represented using bar charts and line graphs to provide a clear month-wise comparison.

These insights can help law enforcement and city planners optimize patrol schedules, allocate resources more effectively, and implement timely preventive measures.

iii. Analysis Results

The month-wise breakdown revealed a noticeable increase in vehicle theft incidents during the year-end months, particularly in November and December. This trend suggests that theft activity tends to rise during this period, possibly due to increased vehicle usage, holiday travel, or reduced vigilance. Understanding this seasonal spike is crucial, as it can help authorities and communities implement targeted prevention strategies, enhance surveillance efforts, and raise public awareness during high-risk months.

Such time-based insights are valuable for optimizing resource deployment and reducing theft incidents during peak periods.

iv. Visualization

To illustrate this analysis:

- A **line chart** was used to show the increasing vehicle theft cases month wise.

Objective 4: Type of Vehicle Stolen Most in any Particular Area.

i. General Description

The objective of this analysis is to examine theft patterns across different locations and identify which types of vehicles are most frequently stolen in specific areas. Understanding these localized trends is crucial for recognizing high-risk zones and vehicle types, enabling law enforcement and communities to take targeted preventive measures. Insights from this analysis can support strategic decisions related to patrol planning, public awareness campaigns, and vehicle security recommendations.

ii. Specific Requirements

To carry out this objective, the following steps were taken:

- The dataset was grouped based on the area or location of each theft incident.
- Within each area, the most frequently stolen vehicle types were identified by counting the number of theft cases per vehicle type.
- Visualizations such as bar charts were created to highlight the most targeted vehicle types in each region.

This location-based analysis provides valuable insights into regional theft patterns, helps identify high-risk vehicle categories, and supports the development of area-specific security strategies.

iii. Analysis Results

From the analysis, it was observed that the frequency of vehicle theft varied across different types. All-Terrain Vehicles recorded the highest number of thefts, followed by Trucks and then Cabs. This indicates that All-Terrain Vehicles are particularly vulnerable and may be more attractive targets for theft, possibly due to their versatility or higher resale value.

The trend highlights a clear preference among thieves, suggesting targeted theft behavior rather than random selection.

These insights are crucial for law enforcement and vehicle owners, as they help in identifying which vehicle types require more robust security measures, better tracking systems, and focused public awareness in high-risk areas.

iv. Visualization

- A **Pie chart** was used to clearly visualize the difference between choice of vehicle.

Objective 5: Year with Most Vehicle Theft Cases.

i. General Description

This objective focuses on analyzing vehicle theft trends over the years to identify which year recorded the highest number of theft incidents. Understanding year-wise fluctuations in theft cases helps in recognizing long-term trends, assessing the effectiveness of past security measures, and anticipating future risks. These insights are critical for law enforcement to adjust prevention strategies, allocate resources effectively, and plan targeted awareness campaigns during peak theft years.

ii. Specific Requirements

The following steps were undertaken to perform this analysis:

- A new column was derived to represent the year of each vehicle theft incident, ensuring the years were ordered chronologically (e.g., 2021, 2022).
- The dataset was grouped by year, and the total number of theft cases for each year was calculated.
- A bar chart was used to visualize the number of vehicle thefts in each year.

iii. Analysis Results

The results indicated that 2022 recorded the highest number of theft cases when compared to 2021, highlighting a significant increase in vehicle theft incidents. This comparison helps in understanding theft trends over time and supports the allocation of resources to areas with rising theft cases.

iv. Visualization

To visually interpret the monthly sales patterns:

- A bar chart was created, with each bar representing the total number of theft cases per year.
- The y-axis represented the years (2021 and 2022), and the x-axis represented the total number of vehicle thefts.
- Distinct colors for each year helped differentiate the theft counts clearly.
- Grid lines and well-labeled axes enhanced the readability and interpretation of the data, making it easier to compare theft trends between the two years.

5. CONCLUSION

This project focused on performing an in-depth exploratory data analysis (EDA) on a Vehicle Stolen dataset using Python. The goal was to derive meaningful insights from raw data, visualize trends, and support strategic decision-making through data-driven interpretations. The entire analysis was conducted using Python libraries such as **Pandas**, **Matplotlib**, and **Seaborn**.

6. FUTURE SCOPE

The current project has successfully demonstrated the potential of Exploratory Data Analysis (EDA) in extracting meaningful insights from retail sales 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 businesses.

1. Integration of Machine Learning Models

In the future, this project can be extended by incorporating predictive analytics to better understand and prevent vehicle theft. Machine learning models such as:

- **Linear Regression** or **XGBoost** for predicting theft trends based on historical data
- **Classification models** (e.g., **Decision Trees**, **Random Forest**) to identify factors contributing to theft and predict high-risk theft periods
- **Clustering techniques** (like **K-Means**) for identifying patterns in theft incidents across different regions and vehicle types

These techniques can help law enforcement and urban planners anticipate theft trends, enhance resource allocation, and implement proactive security measures.

2) **Dashboard Development**

Creating interactive dashboards using tools like Tableau, Power BI, or Plotly Dash can enhance the user experience by allowing non-technical stakeholders to explore the data visually and interactively.

3. Location based Theft Analysis

If location-based data is available, incorporating geospatial analysis could reveal:

- Regional differences in vehicle theft rates
- High-risk areas for targeted prevention measures
- Area-wise vehicle type popularity and theft patterns

This can help law enforcement and city planners optimize resource deployment, implement area-specific security initiatives, and tailor awareness campaigns based on regional theft trends.

4. Multivariate Analysis

Applying multivariate statistical methods can help analyze the combined effect of multiple variables like price, category, customer demographics, and time. This can uncover deeper patterns and interactions that are not evident in univariate or bivariate analysis.

Conclusion of Future Scope

In summary, this project provides a strong foundation for data-driven decision-making in vehicle theft prevention. By leveraging advanced analytics, predictive models, and geospatial tools, the project can evolve into a comprehensive security solution. These future enhancements can enable more accurate risk predictions, optimize resource allocation, and improve targeted prevention strategies, ultimately leading to a reduction in vehicle theft incidents.

7.REFERENCES

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 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 Retail Sales," *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*, Packt Publishing, 2017.

[7] A. Anuradha and D. N. Kumar, "Customer Segmentation and Sales Prediction in Retail 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.