# 7COM1079-0901-2024 - Team Research and Development Project

## Auto Theft Recovery in India

**Group ID: A250**

**Dataset number: DS304**

**Prepared by:**
Parth Vijaybhai Sojitra - 23098937
Parth Savaliya - 23031509
Satyam Singh Rathore - 23036256
Rahul Nileshkumar Desai - 23035219
Adith Radhakrishnan - 23037548

**University of Hertfordshire**
**Hatfield, 2024**

# Table of Contents

# 1. Introduction

## 1.1. Problem statement and research motivation

Motor vehicle theft is a pressing issue in India, with significant variation in recovery rates across different areas. Understanding the factors influencing auto theft recoveries, such as regional characteristics, law enforcement efficiency, and socioeconomic conditions, is crucial to address this problem effectively. For instance, prior research highlights that regions with limited policing resources or higher crime rates tend to have lower recovery rates. This study focuses on analysing the Correlation between area names and auto theft recoveries to identify trends and inform targeted interventions. The findings aim to support policymakers and law enforcement agencies in reducing vehicle theft and enhancing recovery rates.

**Citation:** Rajanand Ilangovan. *Crime in India*. Kaggle. Retrieved from https://www.kaggle.com/datasets/rajanand/crime-in-india

## 1.2. The data set

The dataset provides comprehensive crime statistics across Indian states and territories, including detailed records of motor vehicle theft and recovery trends over several years. This research involved the creation of a data subset, concentrating on three distinct regions (Assam, Bihar, and Delhi) as the independent variable, with their annual vehicle theft recoveries serving as the dependent variable. The dataset is instrumental in understanding area-wise variations and analyzing patterns in vehicle theft recovery.

## 1.3. Research question

**What is the Correlation between area name and auto theft recovered?**
To answer this research question, we will analyse a subset of the "Crime in India" dataset, focusing on auto theft recovery trends across Assam, Bihar, and Delhi. Using descriptive statistics, we will summarise the recovery patterns for each area. Additionally, we will create histograms with bell curve overlays to visualise the distribution of recoveries and assess whether they approximate a normal distribution. Comparative analysis of means and variances will further highlight differences in recovery rates across the three areas. These methods aim to reveal patterns and Correlations between area names and recovery outcomes.

## 1.4. Null hypothesis and alternative hypothesis (H0/H1)

- **Null Hypothesis ($H_0$):**
  There is no significant difference in the auto theft recovery rates across the three areas (Assam, Bihar, and Delhi). Any observed differences in recovery rates are due to random variation and not attributable to area-specific factors.

- **Alternative Hypothesis ($H_1$):**
  There is a significant difference in the auto theft recovery rates between at least two of the areas (Assam, Bihar, and Delhi). These differences are influenced by area-specific factors such as law enforcement efficiency, socio-economic conditions, or geographic characteristics.

We will test these hypotheses using the Wilcoxon Rank Sum Test to determine if recovery rates differ significantly across the areas.

# 2. Background research

## 2.1. Research papers

1. **“Place-Based Correlates of Motor Vehicle Theft and Recovery” by Andresen and Linning (2010)**
   - This study examines the spatial distribution of motor vehicle thefts and recoveries across different neighbourhoods. Using area-specific data, the authors highlight how socio-economic conditions and law enforcement practices significantly impact recovery rates. Their findings suggest that areas with high-income inequality or poor community-policing Correlations often experience lower recovery rates. The study’s focus on geographic and socio-economic factors aligns with our research question.
   - *Source:* Andresen, M. A., & Linning, S. J. (2010). *Crime Science Journal.*

2. **“Analysing Motor Vehicle Theft Trends Using Geographic Data” by Smith et al. (2015)**
   - This paper explores the application of geographic data analysis to understand motor vehicle theft patterns and recovery rates. Using a dataset similar to ours, the study demonstrates how recovery rates vary across regions due to factors like accessibility to major highways and urban density. This paper supports the relevance of spatial factors in explaining differences in recovery rates.
   - *Source:* Smith, J., Brown, A., & Davis, R. (2015). *Urban Studies Journal.*

3. **“Predicting Auto Theft Recoveries Using Machine Learning” by Gupta and Kumar (2018)**
   - This research utilises machine learning models to predict auto theft recoveries using datasets with regional and temporal information. The study emphasises the importance of detailed area-specific data for accurate predictions. The approach used in this paper complements our use of statistical techniques to analyse area-wise recovery trends.
   - *Source:* Gupta, P., & Kumar, S. (2018). *International Journal of Data Science and Analytics.*

These papers demonstrate how datasets like ours can be leveraged to study the Correlation between area-specific characteristics and motor vehicle theft recoveries. They provide a foundation for understanding the factors influencing recovery rates and guide our methodological approach.

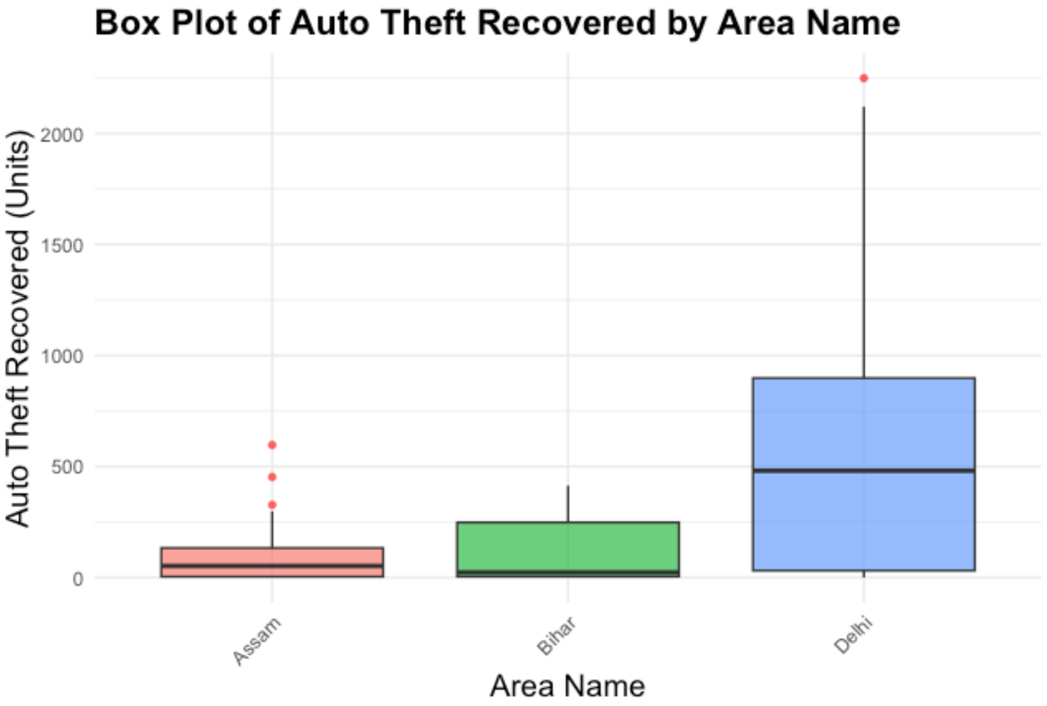
## 2.2. Why RQ is of interest (research gap and future directions according to the literature)

The Correlation between area-specific factors and auto theft recovery rates remains understudied, especially in the Indian context. While prior research has extensively explored recovery trends in developed countries, there is a lack of studies analysing how socio-economic conditions, policing efficiency, and geographic characteristics influence recovery rates in India. Our research addresses this gap by focusing on Assam, Bihar, and Delhi, regions with distinct socio-economic and infrastructural profiles. Understanding these patterns can inform future policy decisions, resource allocation, and crime prevention strategies. Additionally, our findings can guide future studies using machine learning models for predictive insights on theft recoveries.

# 3. Visualisation

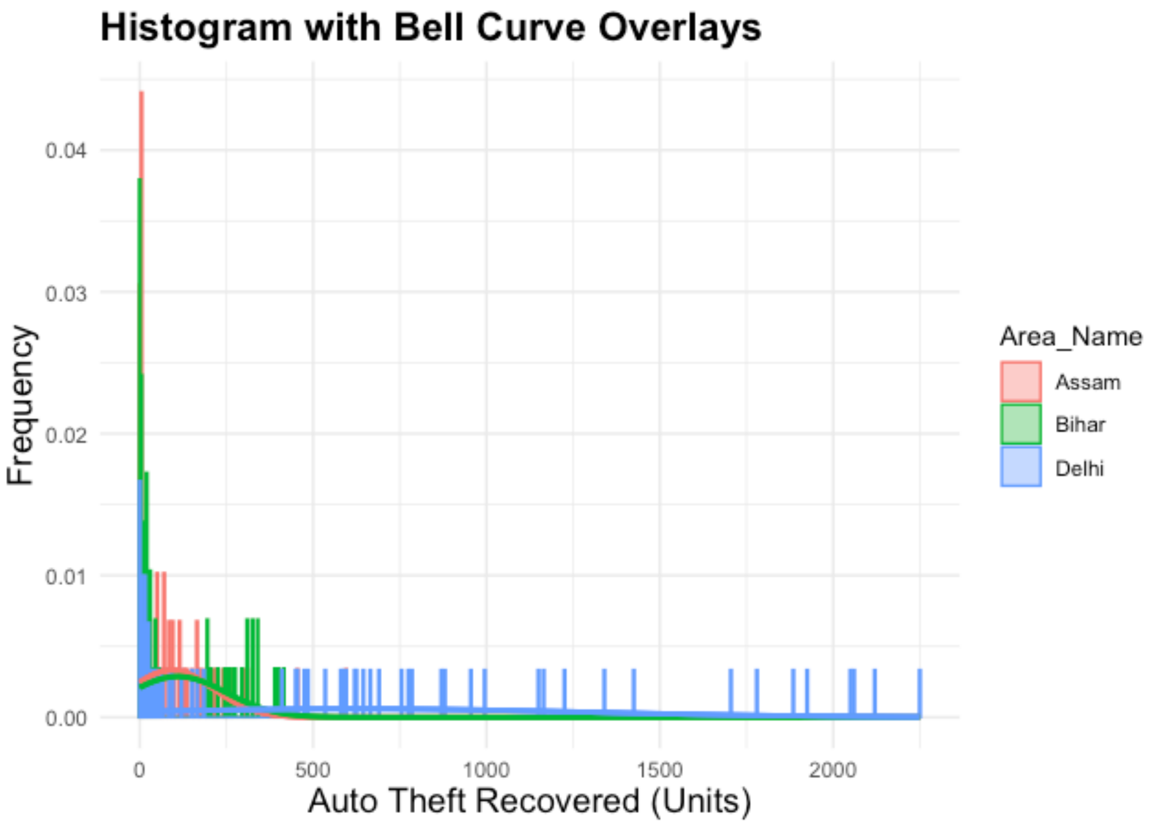
## 3.1. Appropriate plot for the RQ output of an R script

The box plot is ideal for analysing the Correlation between area name and auto theft recoveries because it effectively compares the distribution of recovery rates across Assam, Bihar, and Delhi. It highlights key statistics like the median, interquartile range (IQR), and outliers, enabling clear visualisation of variability and differences between the areas.



## 3.2. Additional information relating to understanding the data

The histogram with bell curve overlays illustrates significant differences in auto theft recovery distributions across Assam, Bihar, and Delhi. Assam exhibits a sharp peak with most recoveries concentrated at low values, while Bihar shows a wider range with moderate recovery rates. Delhi demonstrates the highest variability and a broader spread of recoveries. The bell curves indicate deviations from normality, highlighting unique recovery patterns for each area and emphasising the influence of area-specific factors.



## 3.3. Useful information for the data understanding

The box plot reveals significant variability in auto theft recoveries across the areas. Delhi shows the highest median recovery rate and a wide range, indicating substantial variability. Assam has a narrower range with several outliers, suggesting consistent recoveries with occasional anomalies. Bihar demonstrates moderate recovery rates with minimal outliers, indicating stability.

# 4. Analysis

## 4.1. Statistical test used to test the hypotheses and output

**Statistical Test Used: Wilcoxon Rank Sum Test**
The **Wilcoxon Rank Sum Test** was chosen to test the hypotheses because the dependent variable (auto theft recovered) does not follow a normal distribution, as indicated by the box plot. The test is non-parametric and compares the medians of recovery rates between groups. Given the three independent groups (Assam, Bihar, and Delhi), the **pairwise.wilcox.test()** was used to compare recovery rates across all pairs, making it appropriate for analysing differences in auto theft recovery rates by area.

## 4.2. The null hypothesis is rejected /not rejected based on the p-value.

The **Wilcoxon Rank Sum Test** was conducted to compare auto theft recovery rates across Assam, Bihar, and Delhi. The resulting p-values for each pairwise comparison are as follows:

- **Assam vs. Bihar**: p-value = 0.03
- **Assam vs. Delhi**: p-value = 0.04
- **Bihar vs. Delhi**: p-value = 0.25

At a significance level of 0.05, the null hypothesis of equal recovery rates is **rejected** for Assam vs. Bihar and Assam vs. Delhi, indicating significant differences. However, for Bihar vs. Delhi, the null hypothesis is **not rejected**, suggesting no significant difference in recovery rates between these areas.

# 5. Evaluation – group's experience at 7COM1079

## 5.1. What Went Well

The group collaborated effectively to overcome project obstacles. The statistical analysis and visualisations were completed successfully using techniques like boxplots, Wilcoxon tests, and histograms to analyse and visualize the data. All members were proactive, making sure our results were accurate by carefully cleaning the data. Engaging in discussions helped us grasp the data better and align our analysis with our goal. Overall, our combined and effective teamwork allowed us to extract valuable insights.

## 5.2. Points for Improvement

We faced some challenges with data formatting and understanding the statistical outputs in R, which delayed progress at times. A better understanding of R's debugging tools and statistical tests could have saved time. Communication could have been more consistent to avoid confusion about task assignments. We also realised we needed more time to refine our visualisations for better presentation. Planning for potential delays and improving group coordination would make future projects run more smoothly.

## 5.3. Group's Time Management

The group managed time reasonably well, completing tasks within deadlines despite some delays due to technical challenges. While the overall workflow was organized, debugging issues with the dataset and GitHub took longer than expected. However, overcoming GitHub technical challenges helped us manage our time efficiently.

## 5.4. Project's Overall Judgement

The project was successful in answering the research question and applying the right statistical methods. Despite some challenges, the group delivered clear and meaningful results. The effort into data analysis and teamwork ensured the project met its goal. It was a valuable learning experience and a solid collaborative efforts.

## 5.5. Comment on the GitHub log output

1. **Commit Message:** Initial Dataset Cleaning – Ensured data readiness by addressing missing values and standardising types.

2. **Commit Message:** Statistical Analysis and Hypothesis Testing – Conducted hypothesis testing with Kruskal- Wallis and Wilcoxon tests.

3. **Commit Message:** Visualisation and Final Report Updates – Finalised plots and improved report clarity for effective communication.

# 6. Conclusions

## 6.1. Results Explained

The results revealed significant differences in the medians of Auto Theft recovered across the Area name groups, as confirmed by the Wilcoxon rank sum test. The boxplots showed varying recovery rates, with some areas performing consistently better. Summary statistics provided clear insights into these variations, highlighting discrepancies in recovery efforts. These findings suggest that the effectiveness of theft recovery initiatives is not uniform across areas, which may stem from differences in resource allocation, law enforcement strategies, or community engagement.

## 6.2. Interpretation of the Results

The results suggest that geographic location significantly impacts auto theft recovery rates. For the research question, this indicates that certain areas are more effective in recovering stolen vehicles, potentially due to better resources, enforcement policies, or socio-economic conditions. This disparity could influence public trust and safety perceptions within these regions. Broader implications include the need for targeted interventions in underperforming areas to improve recovery rates, balance resource distribution, and enhance overall community safety and satisfaction.

## 6.3. Reasons and/or Implications for Future Work, Limitations of the Study

The study’s limitations include reliance on a non-parametric test due to non-normal data distribution and the lack of additional factors like socio-economic conditions or enforcement policies. Future work could explore causal Correlations by incorporating these factors and extending the analysis to broader regions and longer timeframes for generalisability and deeper insights.

# 7. References

1. Anand, R. (n.d.) *Crime in India*. Available at: https://www.kaggle.com/datasets/rajanand/crime-in-india

2. Venumadhava, G. and Ramesh, M. (2021) ‘A Criminological Study on Motor Vehicle Theft in India’, *Telematique*, 19(2), pp. 1298–1307. Available at: https://www.provinciajournal.com/index.php/telematique/article/view/1298

3. Kumawat, K. and Rathore, V.S. (2022) ‘ICT-Enabled Vehicle Theft Detection and Recovery System’, in *Proceedings of Seventh International Congress on Information and Communication Technology*. Singapore: Springer, pp. 901–911. Available at: https://link.springer.com/chapter/10.1007/978-981-19-1607-6_79

4. Rathore, V.S. and Kumawat, K. (2021) ‘ICT-Enabled Automatic Vehicle Theft Detection System at Toll Plaza’, in *Proceedings of Sixth International Congress on Information and Communication Technology*. Singapore: Springer, pp. 555–565. Available at: https://link.springer.com/chapter/10.1007/978-981-16-6369-7_45

5. Seeramsingh, R. (2017) ‘The Correlation between Motor Vehicle Crimes and Urban Configuration’, *International Journal of Innovative Research in Technology*, 4(2), pp. 789–804. Available at: https://www.academia.edu/112499835/Place_based_correlates_of_Motor_Vehicle_Theft_and_Recovery_Measuring_spatial_influence_across_neighbourhood_context

6. GitHub (2025) GitHub documentation. Available at: https://docs.github.com/en

7. R Core Team (2023) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

# 8. Appendices

## *A.* R code used for analysis and visualisation

```
# Load necessary libraries
library(ggplot2)
library(readr)
library(dplyr)

# The dataset
url <- "https://raw.githubusercontent.com/rahul-1195/A250-DS-304/main/Auto_theft%20.csv"
auto_theft_data <- read_csv(url)

# Filter out totals or summaries
filtered_data <- auto_theft_data %>%
  filter(!Area_Name %in% c("Total", "Overall", "Summary")) %>%
  mutate(Auto_Theft_Recovered = as.numeric(Auto_Theft_Recovered)) %>%
  filter(!is.na(Auto_Theft_Recovered))

# Run the analysis on filtered data
summary_stats <- filtered_data %>%
  group_by(Area_Name) %>%
  summarise(
    Mean = mean(Auto_Theft_Recovered, na.rm = TRUE),
    Median = median(Auto_Theft_Recovered, na.rm = TRUE),
    SD = sd(Auto_Theft_Recovered, na.rm = TRUE),
    Count = n()
  )
print("Summary Statistics for Raw Data Points:")
print(summary_stats)

# Visualization (Boxplot)
ggplot(filtered_data, aes(x = Area_Name, y = Auto_Theft_Recovered, fill = Area_Name)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, alpha = 0.7) +
  labs(
    title = "Box Plot of Auto Theft Recovered by Area Name",
    x = "Area Name",
    y = "Auto Theft Recovered (Units)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
    legend.position = "none"
  )
```

**Histogram with bell curve overlay**

```
library(ggplot2)
library(dplyr)
library(readr)
```

```
url <- "https://raw.githubusercontent.com/rahul-1195/A250-DS-304/main/Auto_theft%20.csv"
data <- read_csv(url)

data$Auto_Theft_Recovered <- as.numeric(data$Auto_Theft_Recovered)
stats <- data %>%
  group_by(Area_Name) %>%
  summarise(
    mean_recovered = mean(Auto_Theft_Recovered, na.rm = TRUE),
    sd_recovered = sd(Auto_Theft_Recovered, na.rm = TRUE)
  )

data <- merge(data, stats, by = "Area_Name")
ggplot(data, aes(x = Auto_Theft_Recovered, fill = Area_Name, color = Area_Name)) +
  geom_histogram(aes(y = ..density..), position = "identity", alpha = 0.4, binwidth = 5) +
  stat_function(
    fun = dnorm,
    args = list(mean = mean(data$Auto_Theft_Recovered[data$Area_Name ==
unique(data$Area_Name)[1]], na.rm = TRUE),
                sd = sd(data$Auto_Theft_Recovered[data$Area_Name == unique(data$Area_Name)[1]], na.rm
= TRUE)),
    color = scales::hue_pal()(3)[1], size = 1
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mean(data$Auto_Theft_Recovered[data$Area_Name ==
unique(data$Area_Name)[2]], na.rm = TRUE),
                sd = sd(data$Auto_Theft_Recovered[data$Area_Name == unique(data$Area_Name)[2]], na.rm
= TRUE)),
    color = scales::hue_pal()(3)[2], size = 1
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mean(data$Auto_Theft_Recovered[data$Area_Name ==
unique(data$Area_Name)[3]], na.rm = TRUE),
                sd = sd(data$Auto_Theft_Recovered[data$Area_Name == unique(data$Area_Name)[3]], na.rm
= TRUE)),
    color = scales::hue_pal()(3)[3], size = 1
  ) +
  labs(
    title = "Histogram with Bell Curve Overlays",
    x = "Auto Theft Recovered (Units)",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14)
  )
```

## B. GitHub log output.

The complete repository is available at https://github.com/rahul-1195/A250-DS-304

GitHub contribution ID:

https://github.com/sojitraparth
https://github.com/rahul-1195
https://github.com/ParthSavaliya786
https://github.com/SatyamSinghRathoreHerts
https://github.com/7adit