Data Analysis and Visualisation Principles

By

Adepoju, Akinlolu

May 2025.

Table of Contents

List of Abbreviations

CPS - Crown Prosecution Service

EDA - Exploratory Data Analysis

ANOVA - Analysis of Variance

RMSE - Root Mean Squared Error

MAE - Mean Absolute Error

R^2 - Coefficient of Determination

ROC-AUC - Receiver Operating Characteristic – Area Under the Curve

TP - True Positive

TN - True Negative

FP - False Positive

FN - False Negative

RF - Random Forest

DT - Decision Tree

GB - Gradient Boosting

XGBoost - Extreme Gradient Boosting

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

WCSS - Within-Cluster Sum of Squares

MinPts - Minimum Points (parameter in DBSCAN)

BI - Business Intelligence

API - Application Programming Interface

KDE - Kernel Density Estimation

CSV - Comma-Separated Values

1.0 Introduction

The Crown Prosecution Service (CPS) plays a pivotal role in delivering justice by prosecuting criminal cases in England and Wales. Understanding how prosecutions vary by region and offence type is critical to evaluating performance and identifying areas for systemic improvement. This report presents a comprehensive, quantitative analysis of CPS prosecution outcomes from 2014 to 2015, utilising an integrated dataset that encompasses a wide range of offence categories, including violence, theft, and drug-related crimes, across various regions. The dataset includes key performance indicators such as conviction counts, the number of unsuccessful cases, and offence rate percentages across various crime categories, providing valuable insights into the performance and operations of the justice system. This study aims to explore these patterns using predictive and unsupervised learning techniques to classify and cluster trends at the regional and offence levels. Additionally, this study seeks to optimise predictive models for forecasting high-conviction scenarios, thereby providing insights that can inform policy decisions and operational strategies within the CPS.

Research questions guiding this analysis include:

- Are there significant regional disparities in conviction and unsuccessful prosecution rates?
- Which offence types are most likely to result in unsuccessful outcomes?
- Can the likelihood of a given CPS case set resulting in a high conviction rate be accurately predicted using available performance indicators and case attributes?
- Do natural clusters exist among regions based on offence profiles, and can they inform resource allocation?

The methodologies employed include data preprocessing and exploratory analysis, utilising machine learning clustering and classification algorithms, along with appropriate statistical validation, as illustrated in Figure 1.1 below. Visualisations and metrics were used to effectively assess and convey insights, underpinning hypotheses with empirical evidence and contextual relevance.

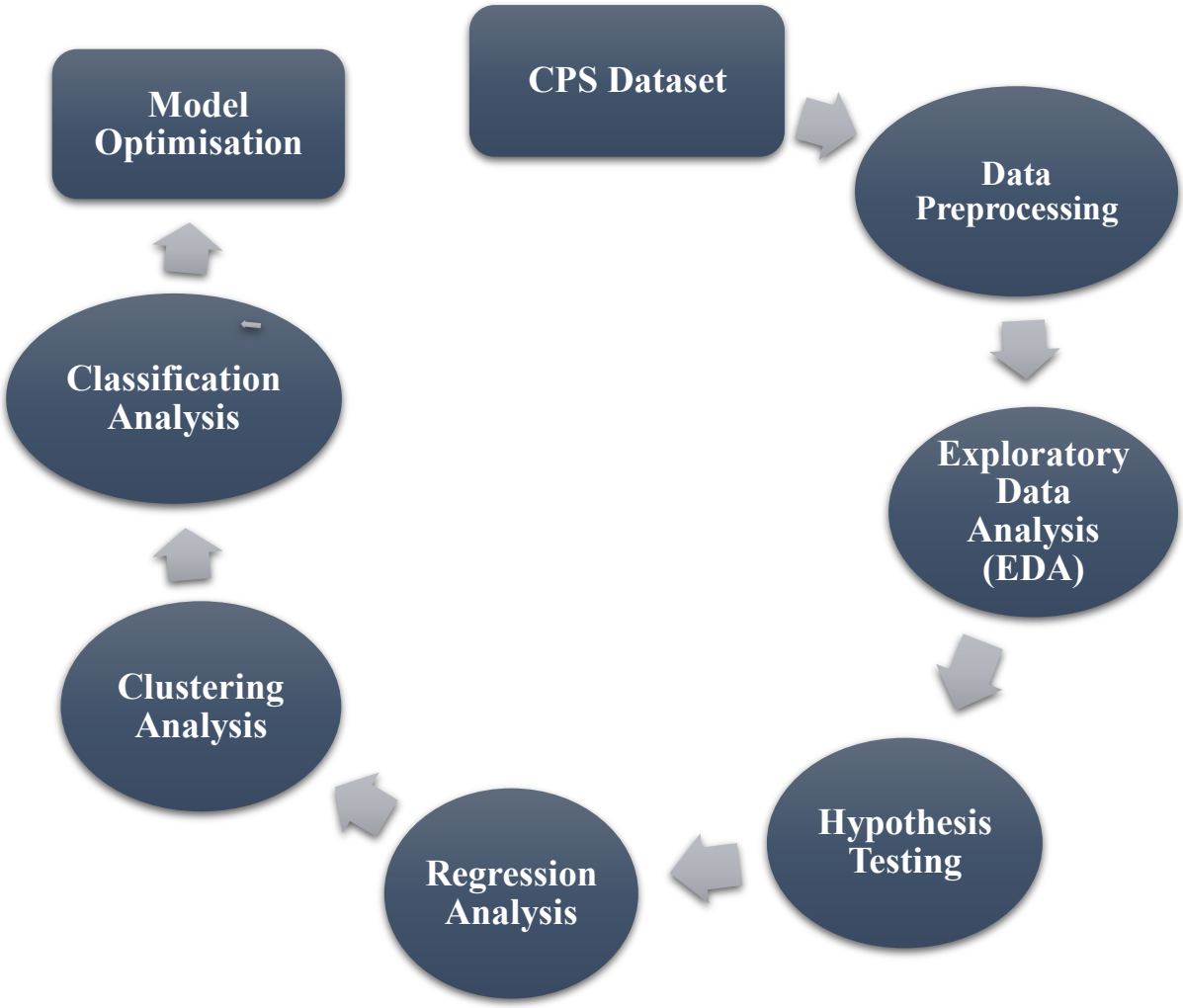

Figure 1.1: Data Analysis Workflow

In the subsequent sections, the report elaborates on data handling strategies, exploratory analysis, inferential tests, and machine learning implementations, concluding with a discussion of limitations and recommendations for operational or policy improvements within the CPS framework.

2.0 Data Cleaning and Preprocessing

This section presents an in-depth discussion of the procedures implemented to clean and preprocess the raw CPS dataset. The dataset used for this analysis comprises monthly CPS outcome reports from January 2014 to December 2015. These monthly reports were stored as individual CSV files in their respective yearly folders. The integration process was fully automated using R scripts, eliminating manual errors and ensuring reproducibility. Prior to processing the data, a comprehensive list of libraries was imported using an automated installation function to ensure all dependencies were available (Figure 2.1).

```r
## 1. INSTALL AND LOAD REQUIRED LIBRARIES
install_and_load_all <- function(packages) {
  for (pkg in packages) {
    if (!requireNamespace(pkg, quietly = TRUE)) {
      message(sprintf("Installing missing package: %s", pkg))
      install.packages(pkg, dependencies = TRUE)
    }
    suppressPackageStartupMessages(library(pkg, character.only = TRUE))
  }
}
required_packages <- c(
  # Core Data Handling
  "tidyverse", "fs", "lubridate", "janitor", "glue", "reshape2",

  # Tidymodels Core + Sub-packages
  "tidymodels", "recipes", "parsnip", "workflows", "rsample", "yardstick", "tune", "dials", "broom",

  # Modeling Engines
  "glmnet", "ranger", "xgboost", "rpart",

  # Tuning & Interpretation
  "vip", "doParallel",

  # Traditional ML & Evaluation
  "caret", "Metrics",

  # Clustering & Distance
  "cluster", "dbscan", "dendextend", "pracma",

  # Visualization Enhancements
  "ggrepel", "viridis", "scales", "GGally",

  # Statistical Analysis
  "rstatix", "FSA", "e1071", "inflection",

  # Plot Layouts and Tables
  "knitr", "gridExtra",

  # String Interpolation
  "glue",

  # Time Series Utilities
  "zoo"
)
install_and_load_all(required_packages)
```

Figure 2.1: Loading of Libraries in R Studio for analysis

These libraries were grouped by purpose:

- **Core Data Handling**:
  - tidyverse: An umbrella package containing essential packages like dplyr, ggplot2, readr, and tidyr for data manipulation, visualisation, and import.
  - lubridate: Facilitates the parsing and manipulation of date-time data.
  - janitor: Offers functions like clean_names() for standardizing variable names.
  - fs: Handles file system operations, such as listing and accessing directories and files.
  - glue: used for string interpolation.
  - Reshape2: used for reshaping dataframe.
- **Tidymodels Framework**:
  - Tidymodels,
    including recipes, parsnip, workflows, rsample, yardstick, tune, dials
    and broom: These packages support a modern, tidy approach to building machine learning workflows, including model specification, preprocessing, evaluation, and tuning.
- **Model Engines and Interpretation**:
  - glmnet, ranger, xgboost, rpart: Provide implementations of regression, random forest, boosting, and decision trees.
  - vip: Generates variable importance plots.
  - doParallel: Enable parallelised and efficient hyperparameter tuning.
- **Traditional Machine Learning**:
  - caret, Metrics: Provide traditional ML modelling and evaluation utilities such as RMSE and MAE.
- **Clustering and Distance Metrics**:
  - cluster, dbscan, dendextend, pracma: Support clustering analysis and visualisations, including silhouette scores and dendrogram enhancement.
- **Visualisation Enhancements**:
  - ggrepel, viridis, scales: Improve the aesthetics and readability of plots, such as handling overlapping text labels and ensuring colourblind-friendly palettes.
  - GGally: provides ggpairs for pairplot
- **Statistical Analysis**:
  - rstatix, FSA, e1071: Enable statistical testing, including ANOVA, Kruskal-Wallis, and Dunn's test, along with measures of distribution skewness.
  - inflection: used to detect inflection points in trend analysis
- **Formatting and Layout**:
  - knitr, gridExtra: Used to arrange multiple plots and format tables for presentation.
- **Time Series Utilities**:
  - zoo: Provides rolling average functions and smoothing for time series data (Sanderson, 2024).

The raw data consisted of monthly outcome reports in CSV format, spanning 2014 and 2015, stored in separate directories for each year. Using R, a reproducible and efficient preprocessing pipeline was developed to automate the integration of these datasets, ensuring high data fidelity, traceability, and operational efficiency—qualities essential when working with administrative records (Finak *et al.,* 2018). A custom function was employed to scan the target directories (CPS_Data/2014 and CPS_Data/2015), identify relevant CSV files using pattern recognition, and iterate over each file. For each dataset, the script attempted to extract the reporting month from the filename using lubridate::parse_date_time() and appended metadata columns including Year, Month, and a composite YearMonth.

A missing month (November) was observed in 2015. To address this, a copy of November from the 2016 folder was programmatically created and relabeled accordingly, as shown in Figure 2.2. This intervention ensured continuity in the time series and prevented downstream analytical gaps (Alwateer *et al.,* 2024). The individual dataframes were concatenated using bind_rows() and sorted chronologically. The resulting dataset (cps_2014_2015_merged.csv) served as the foundation for subsequent preprocessing, analysis, and modelling stages. Preliminary inspections using str(), dim(), colSums(is.na()), and duplicated() confirmed that the dataset encompassed 24 months of data (January 2014 – December 2015), comprising 1,032 rows and 54 columns, with no missing or duplicate entries—affirming its readiness for cleaning and transformation.

```r
# ==== Fill Missing November 2015 Using 2016 Data ====
november_2016_path <- NULL
year_2016_folder <- path(base_dir, "2016")

if (dir_exists(year_2016_folder)) {
  november_files <- dir_ls(year_2016_folder, regexp = "november.*\\.csv$", recurse = FALSE)
  if (length(november_files) > 0) {
    november_2016_path <- november_files[1]
  }
}

if (!is.null(november_2016_path)) {
  df_nov16 <- read_csv(november_2016_path, show_col_types = FALSE)
  if ("Date" %in% colnames(df_nov16)) {
    df_nov16 <- df_nov16 %>% mutate(Date = suppressWarnings(as.Date(Date)))
  }
  df_nov16 <- df_nov16 %>%
    mutate(Year = 2015,
           Month = "November")
  all_data[[length(all_data) + 1]] <- df_nov16
  cat("November 2016 data used to fill missing November 2015.\n")
} else {
  cat("No November 2016 data found to substitute for November 2015.\n")
}
# ==== Combine, Sort, and Format Merged Data ====
cps_data <- bind_rows(all_data)
# Create 'YearMonth' column and format it
cps_data <- cps_data %>%
  mutate(YearMonth = parse_date_time(paste(Month, Year), orders = "B Y", quiet = TRUE),
         YearMonth = format(YearMonth, "%Y-%m")) %>%
  arrange(YearMonth)
# Set month as ordered factor
month_order <- month.name
cps_data <- cps_data %>%
  mutate(Month = factor(Month, levels = month_order, ordered = TRUE))
# ==== Display Info and Save ====
print(cps_data %>% select(Year, Month, YearMonth) %>% distinct())
print(str(cps_data))
print(head(cps_data))
print(tail(cps_data))
# Dataset Inspection
# Check the structure of the dataset: variable names and types
str(cps_data)

# Check the dimensions: number of rows and columns
```

Figure 2.2: R code for filling in Missing Values in November 2015

As shown in Figure 2.3, following the successful merging and loading of the Crown Prosecution Service (CPS) dataset, column names were standardised using janitor::clean_names(), which reformatted all headers into lowercase snake_case for consistency and ease of manipulation. Unnamed indexing columns—typically imported as x1 or ...1—were identified and renamed to region, reflecting the fact that each row represented a distinct region within England and Wales. The region variable was cleaned by converting entries to lowercase, removing whitespace, and filtering out non-geographic labels, such as "national" rows, which represented aggregate totals and could distort regional comparisons. After filtering the national column, the dataset was reduced by 2.3% to a total of 1,008 rows.

```r
## 3. DATA CLEANING
# === Load and Clean Data ===
cps_df <- read_csv("cps_2014_2015_merged.csv", show_col_types = FALSE)

cps_df <- cps_df %>%
  rename_with(str_trim) %>%
  clean_names()

if ("x1" %in% names(cps_df)) {
  names(cps_df)[names(cps_df) == "x1"] <- "region"
} else if ("...1" %in% names(cps_df)) {
  names(cps_df)[names(cps_df) == "...1"] <- "region"
}

if (!"region" %in% colnames(cps_df)) {
  stop("ERROR: No 'region' column found in the dataset.")
}

# === Region and Date Cleaning ===
cps_df <- cps_df %>%
  mutate(region = str_to_lower(str_replace_all(as.character(region), " ", ""))) %>%
  filter(region != "national")

if ("year_month" %in% colnames(cps_df)) {
  cps_df <- cps_df %>%
    mutate(year_month = parse_date_time(year_month, orders = "ym"))
}


# Dataset Inspection
# Check the structure of the dataset: variable names and types
str(cps_df)

# Check the dimensions: number of rows and columns
dim(cps_df)

# Check for missing values in each column
colSums(is.na(cps_df))

# Check for duplicate rows in the dataset
sum(duplicated(cps_df))

# View the actual duplicate rows
cps_df[duplicated(cps_df), ]
```

Figure 2.3: Data Cleaning R-code

To provide an overview of the dataset, Table 1 summarises its key structural features, including the types and distribution of variables, the presence of missing values, and duplicates.

Table 2.1: Summary of Dataset Features

Aspect	Status
Number of rows	1,008
Number of columns	54
Character variables	27, including percentages (e.g., "78.6%"), month names, region names
Numeric variables	26, this includes number of each offences and year
Date/Time variables	1, year_month in POSIXct format
Missing values	No NAs detected, but possible non-standard placeholders like "-" exist
Duplicate entries	None

All percentage fields were stripped of % symbols and converted to numeric types. Embedded commas in numeric values (e.g., "1,000") were also removed to ensure correct parsing. Dashes (–) in the original CPS dataset, used to denote zero or absence of data, were replaced with numeric zeros in both percentage and count fields. Replacing these values with zero was based on a consistent pattern in which dashes appeared in percentage columns alongside corresponding raw counts, which were recorded as zero. Interpreting these as true zeros rather than missing values preserved data consistency and ensured the accuracy of aggregate statistics such as total convictions and unsuccessful outcomes, while preventing unnecessary data loss in subsequent analysis (Alwateer *et al.,* 2024).

For further analysis, additional engineered features were introduced. These included total_convictions and total_unsuccessful, which represent the row-wise summation of convictions and unsuccessful outcomes across all offence categories, respectively. Another derived metric, mean_conviction_percent, was computed as the average conviction rate across all offence types for each region-month entry. Additionally, to facilitate stratified analysis, regions were grouped into broader clusters—North, Midlands, South, South West, Wales, London, and East—based on predefined mappings using a region_group variable, as shown in Figure 2.4.

```r
# === Group Regions ===
region_groups <- list(
  North = c("greatermanchester", "lancashire", "cumbria", "merseyside", "northumbria", "durham", "northyorkshi
            "westyorkshire", "southyorkshire", "cleveland", "humberside", "cheshire"),
  Midlands = c("derbyshire", "nottinghamshire", "leicestershire", "lincolnshire", "westmidlands", "warwickshi
               "staffordshire", "northamptonshire", "westmercia"),
  South = c("essex", "kent", "hampshire", "surrey", "sussex", "thamesvalley", "cambridgeshire", "hertfordshir
            "bedfordshire", "wiltshire"),
  South_West = c("avonandsomerset", "dorset", "devonandcornwall", "gloucestershire"),
  Wales = c("southwales", "northwales", "dyfedpowys", "gwent"),
  London = c("metropolitanandcity"),
  East = c("norfolk", "suffolk")
)

region_map <- unlist(lapply(names(region_groups), function(group) {
  setNames(rep(group, length(region_groups[[group]])), region_groups[[group]])
}))
cps_df$region_group <- region_map[cps_df$region]

# Basic descriptive stats for all numeric columns
cps_df %>%
  select(where(is.numeric)) %>%
  summary()
```

Figure 2.4: R code for grouping Regions

The cleaned dataset was then evaluated through descriptive statistical analysis. Basic summaries (e.g., summary () function) highlighted considerable variation in prosecution outcomes across offence types. For instance, *offences against the person* recorded an average of 74 unsuccessful prosecutions per entry, whereas *homicide* had fewer than one on average. Similarly, conviction percentages ranged from approximately 51% (homicide) to over 94% (drug offences), underscoring the dataset's complexity and justifying the use of advanced modelling techniques for deeper insights. This approach was critical for further hypothesis testing, visualisation and modelling.

3.0 Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) phase served as a critical bridge between data cleaning and machine-learning modelling. Its primary objectives were to uncover underlying trends, validate assumptions, detect anomalies, and generate insights that would guide both hypothesis testing and predictive modelling, given the multidimensional nature of the CPS dataset, which spans multiple offence types, regions, and temporal periods (Dhummad, 2025). A comprehensive and varied set of visualisation techniques was employed. To begin, the dataset was assessed for distribution characteristics using skewness histograms for each numeric offence column, as shown in Figures 3.1 and 3.2. These visual summaries provided immediate insight into the shape and central tendency of conviction data. The presence of right-skewed distributions with a skewness value ranging from 3.0 to 6.0 in several offence categories highlights the importance of robust statistical techniques that can handle non-normality.

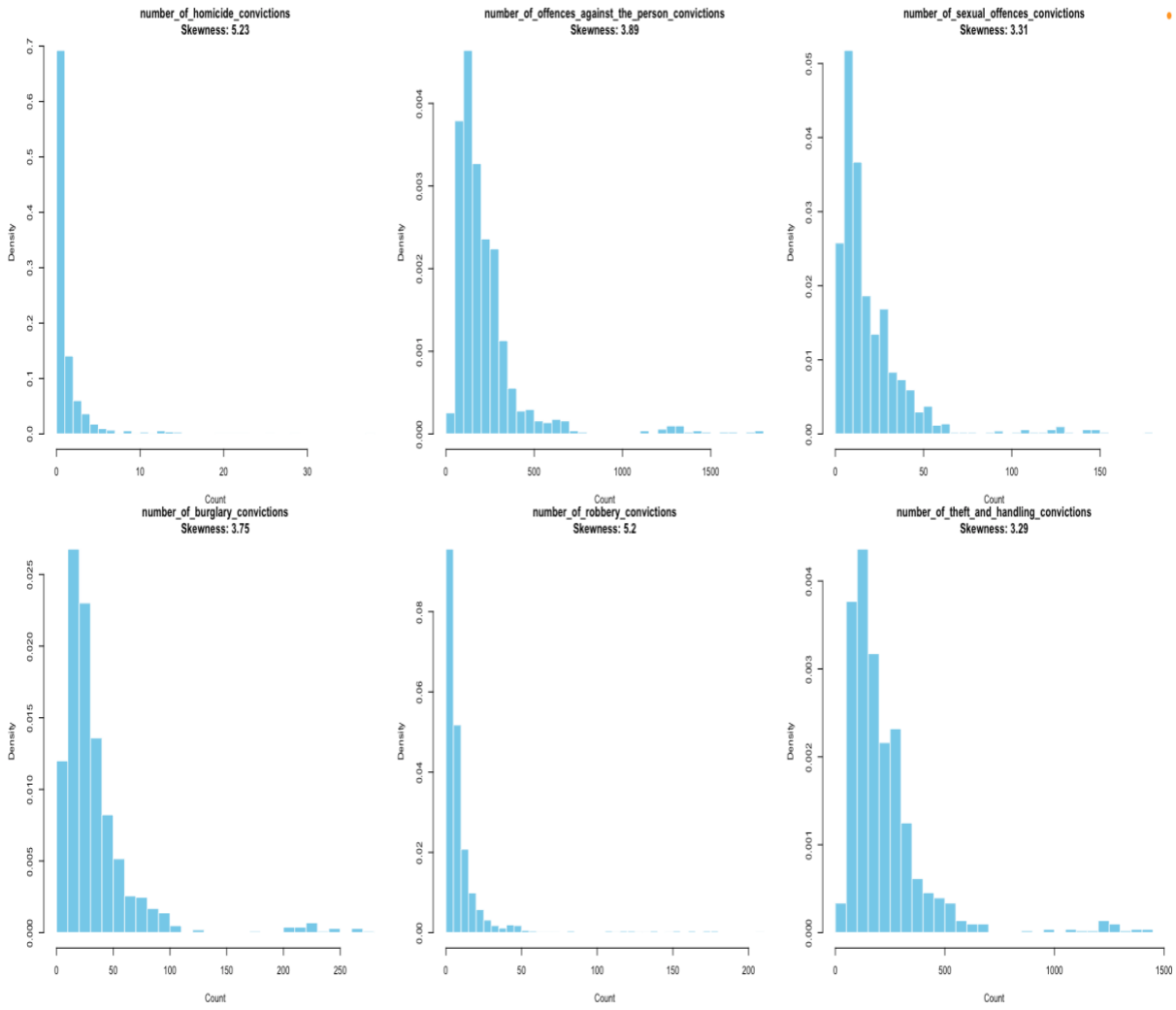

Figure 3.1: Skewness of Selected Number of Offences

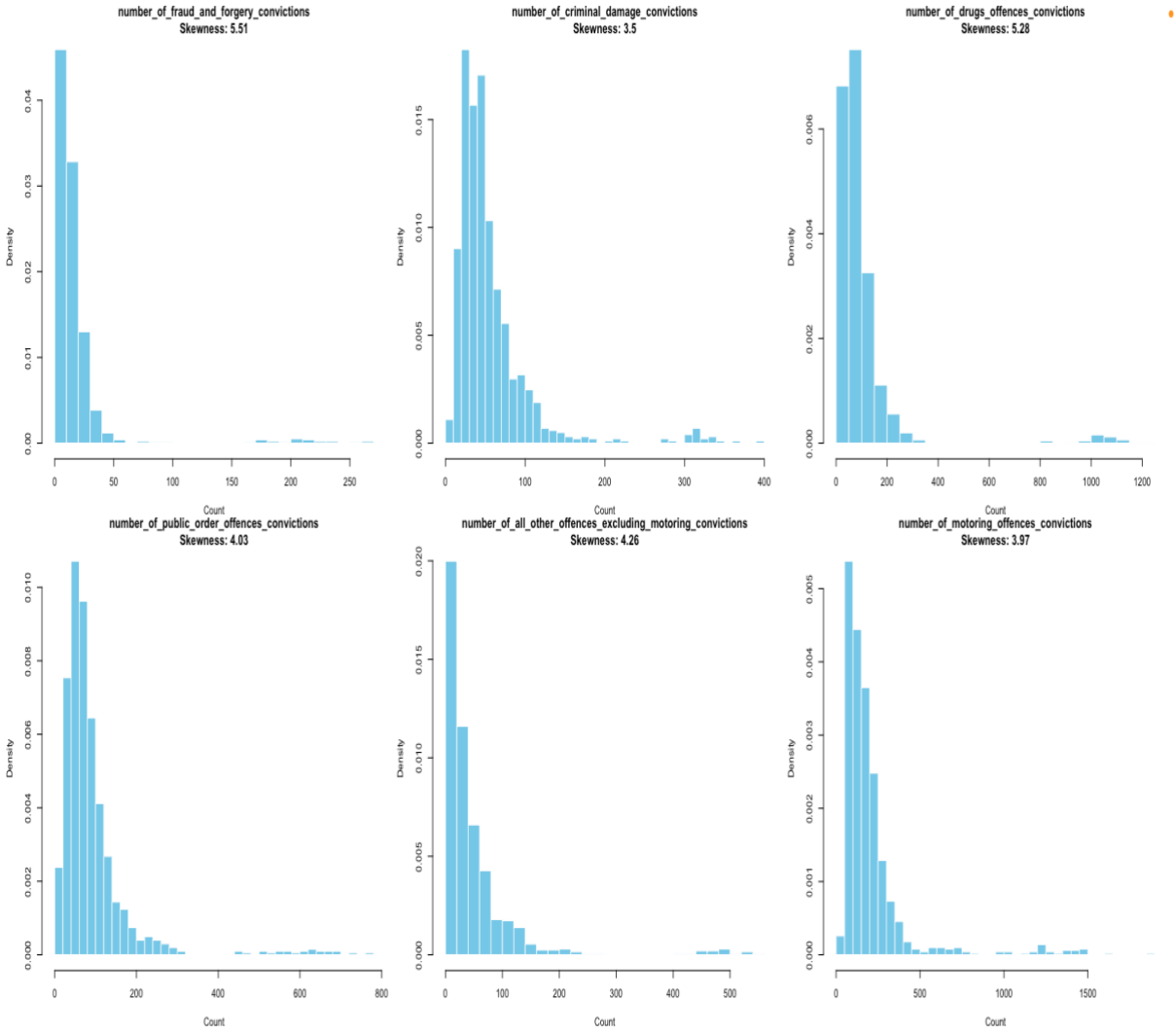

Figure 3.2: Skewness of Selected Number of Offences

To explore temporal dynamics and regional disparities in prosecution outcomes, line plots were employed because they enabled the clear identification of trends across the 24-month period spanning 2014 to 2015 (Pena-Araya, Pietriga and Bezerianos, 2019). A line plot depicting monthly total convictions by region group revealed that the North consistently recorded the highest number of convictions, with approximately 17,000 total convictions, followed by the South. In contrast, the East region displayed the lowest monthly conviction rate, with approximately 1,250 total convictions (Figure 3.3). This visualisation highlights persistent regional disparities in prosecution volumes, which may reflect underlying differences in crime rates, population density, or judicial efficiency.

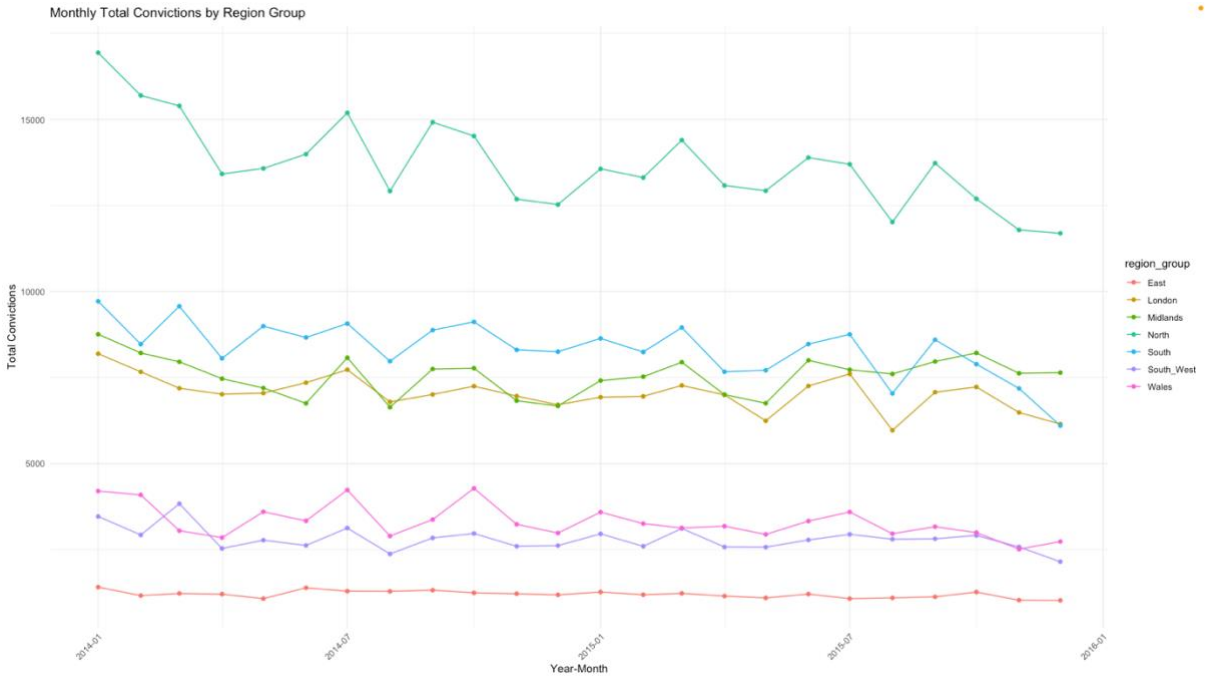

Figure 3.3: Monthly Total Convictions by Region Group

In terms of conviction quality, a line plot of the average conviction rate over time showed a similar shift in regional performance. The East region had the highest average conviction rate, above 87.5%, at the start of 2014, suggesting an initially higher prosecutorial success. However, at the end of 2015, the North had overtaken other regions, with a rate above 85%, indicating possible procedural improvements or shifts in case mix (Figure 3.4). This evolving pattern offers insight into how prosecution strategies or regional judicial dynamics may have changed over time. Furthermore, analysis of monthly unsuccessful outcomes revealed that the North recorded the highest number (more than 2,500 cases) at the beginning of 2014, but by December 2015, both London and the North with approximately 2000 records were leading in unsuccessful outcomes, with the East consistently maintaining the lowest levels (less than 500 cases) (Figure 3.5). This pattern suggests the potential for region-specific challenges in securing convictions, possibly related to caseload complexity, resource allocation, or the cooperation of witnesses. To enhance interpretability and reduce short-term volatility in the data, a national-level rolling average for conviction rates was introduced using a 3-month moving average smoothing technique. This clarified overarching trends and revealed a gradual upward trajectory in the national mean conviction rate, ranging between 80% and 84% across the 24-month period, underscoring an overall improvement in prosecution success (Figure 3.6).

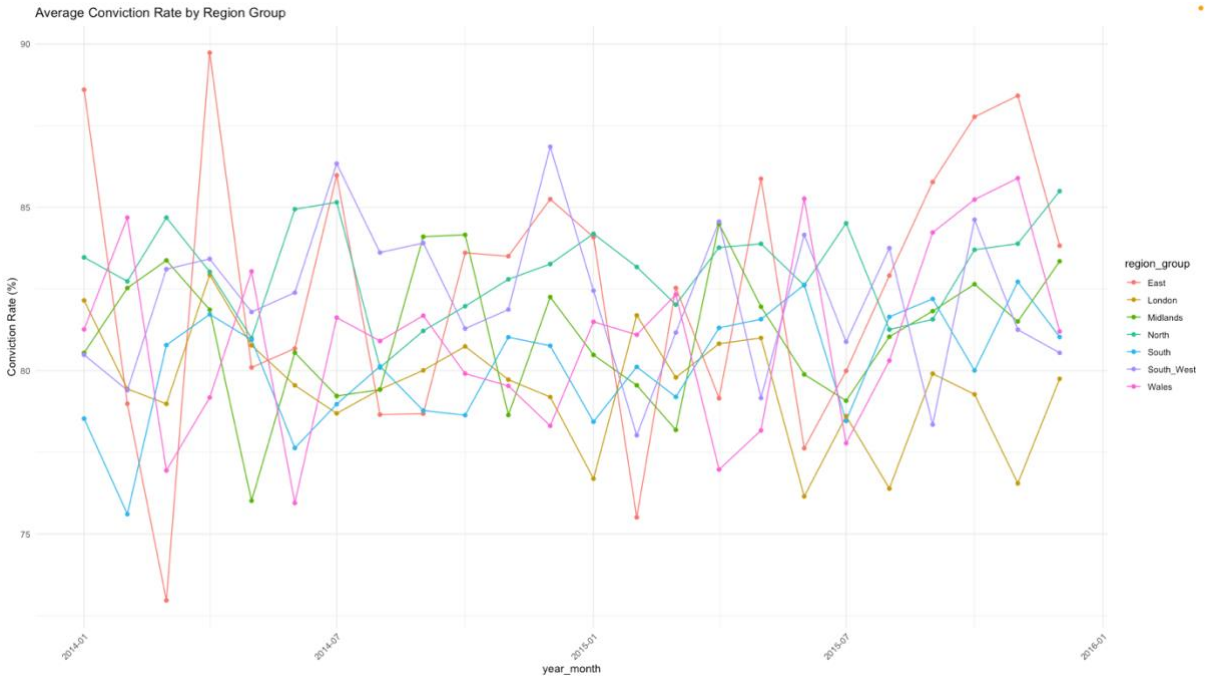

Figure 3.4: Average Conviction Rate by Region Group

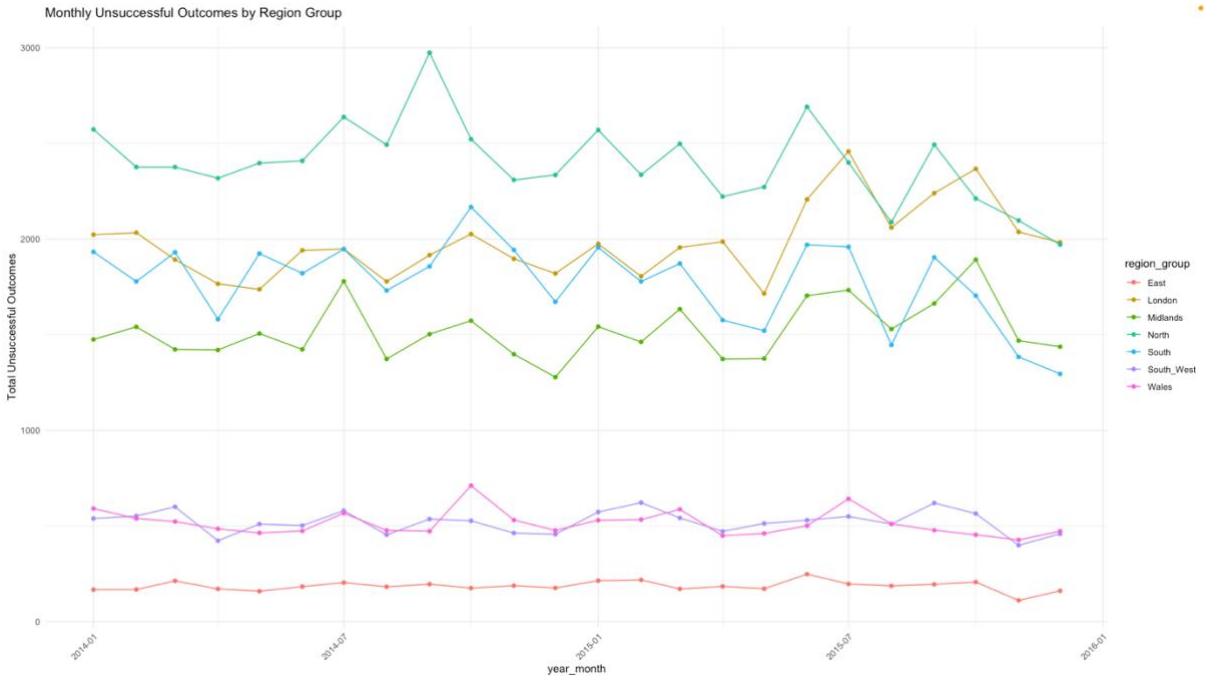

Figure 3.5: Monthly Unsuccessful Outcomes by Region Group

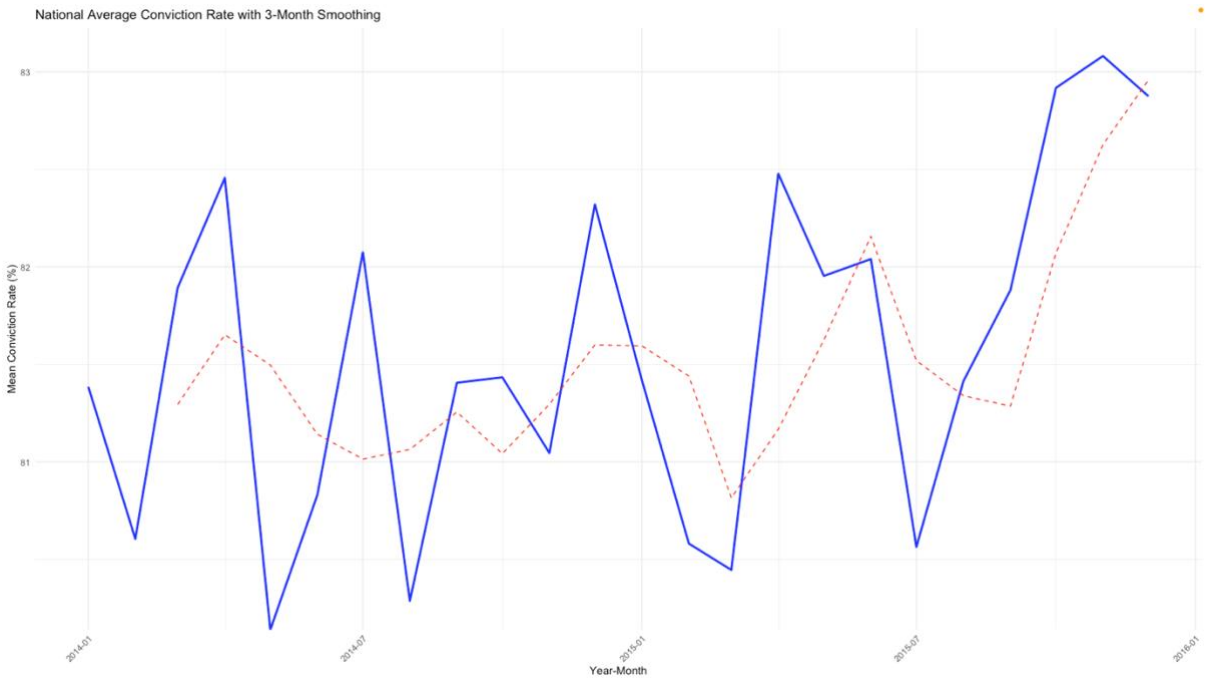

Figure 3.6: National Average Conviction Rate with 3-Month Smoothing

To evaluate regional disparities in the scale of prosecutions, Figure 3.7 uses bar charts to summarise the average total number of convictions across grouped regions (Setiawan and Suprihanto, 2021). The analysis revealed that London had the highest average number of convictions, with approximately 7000, followed by the North, Midlands, South, Wales, South West, and East. These findings suggest regional concentration of prosecutorial activity, which may be influenced by factors such as population density, regional crime rates, and judicial infrastructure. To provide a deeper understanding of prosecution efficiency, stacked bar plots were created to display the ratio of convictions to unsuccessful outcomes for each regional group (Figure 3.8). This visualisation showed systemic differences in case outcomes, with London, the South, and the Midlands exhibiting relatively higher proportions of unsuccessful prosecutions. In contrast, the North region demonstrated a higher share of successful convictions, suggesting potential differences in procedural effectiveness, resourcing, or prosecutorial practices across regions.

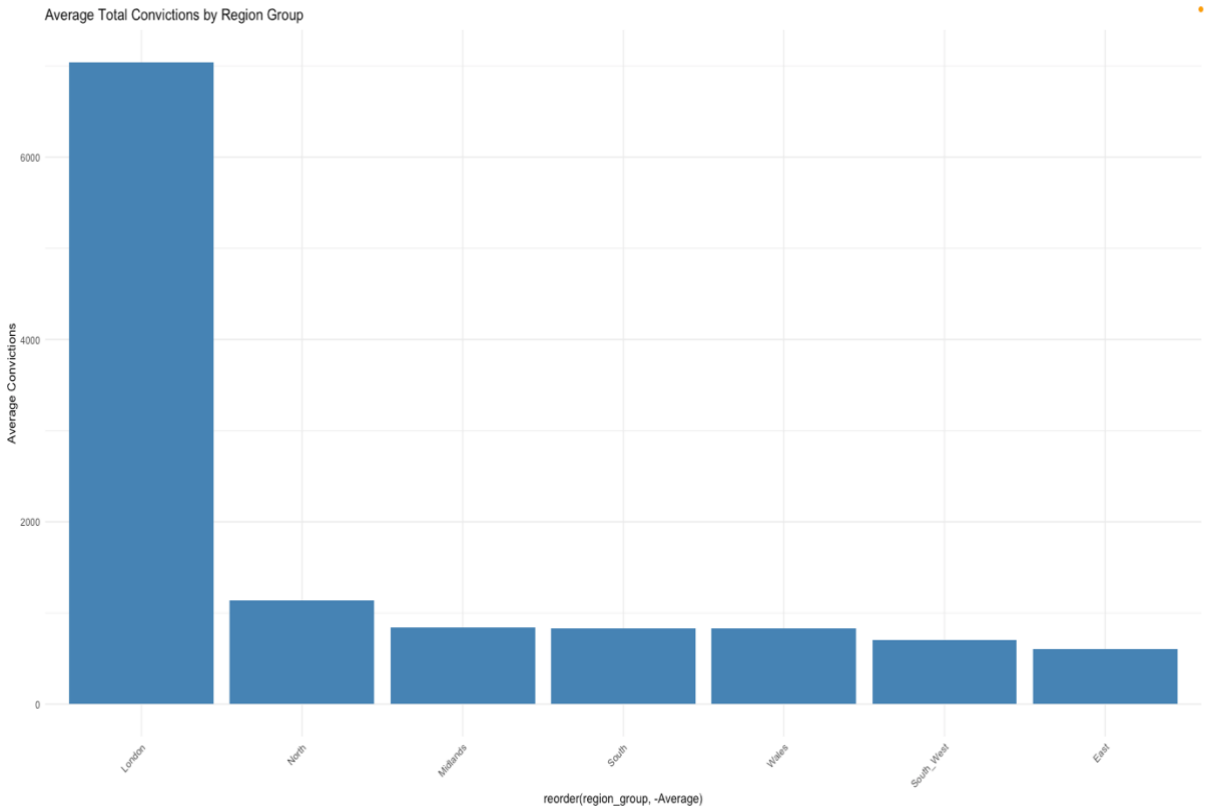

Figure 3.7: Average Total Convictions by Region Group

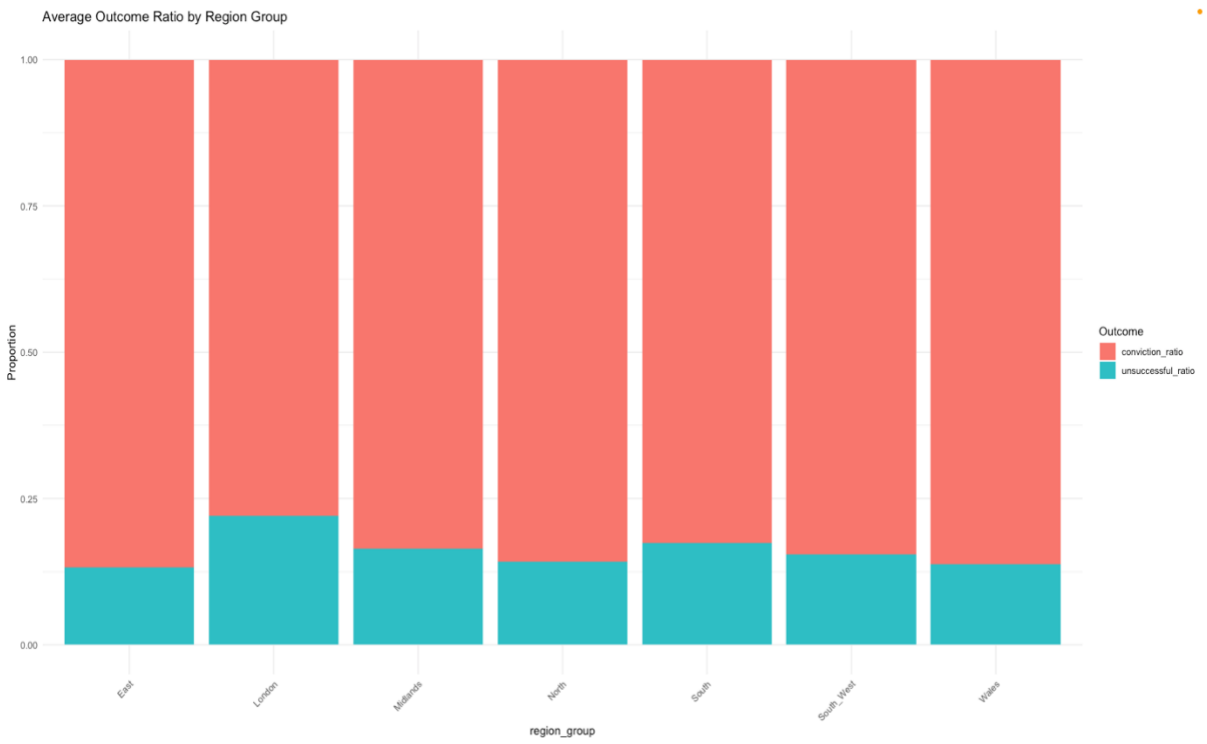

Figure 3.8: Average Outcomes Ratio by Region Group

Boxplots were also used to assess intra-regional variability and detect outliers, as shown in Figure 3.9 below (Setiawan and Suprihanto, 2021). The East, South, and South West regions displayed no outliers, indicating consistent conviction patterns. However, the North, Midlands, and Wales showed outliers above the upper quartile, while London exhibited both upper and lower outliers, highlighting greater variability. These findings suggest uneven performance and potential anomalies in conviction reporting or case characteristics, which may be a result of some offences potentially having a higher conviction rate (i.e., theft and handling, and drug offences) than others (i.e., sexual offences). To explore distributional characteristics, KDE-enhanced histograms were applied to assess the shape and skewness of total conviction counts (Figure 3.10). The histograms confirmed a right-skewed distribution, consistent with earlier visualisations, indicating that while most regions reported moderate conviction totals, a few recorded exceptionally high values. This suggests non-uniform caseloads or distinct prosecutorial trends across regions.

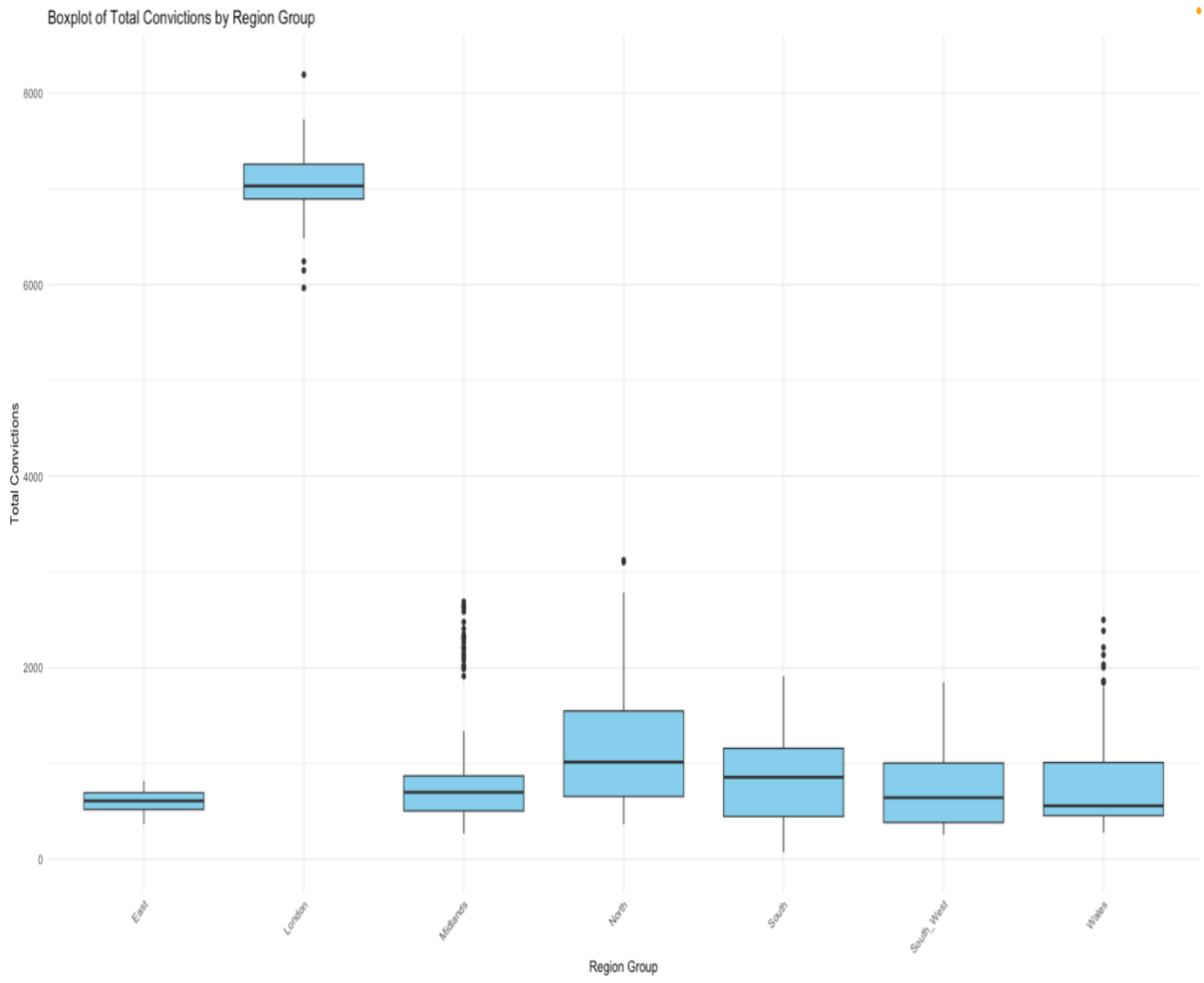

Figure 3.9: Boxplot of Total Convictions by Region Group

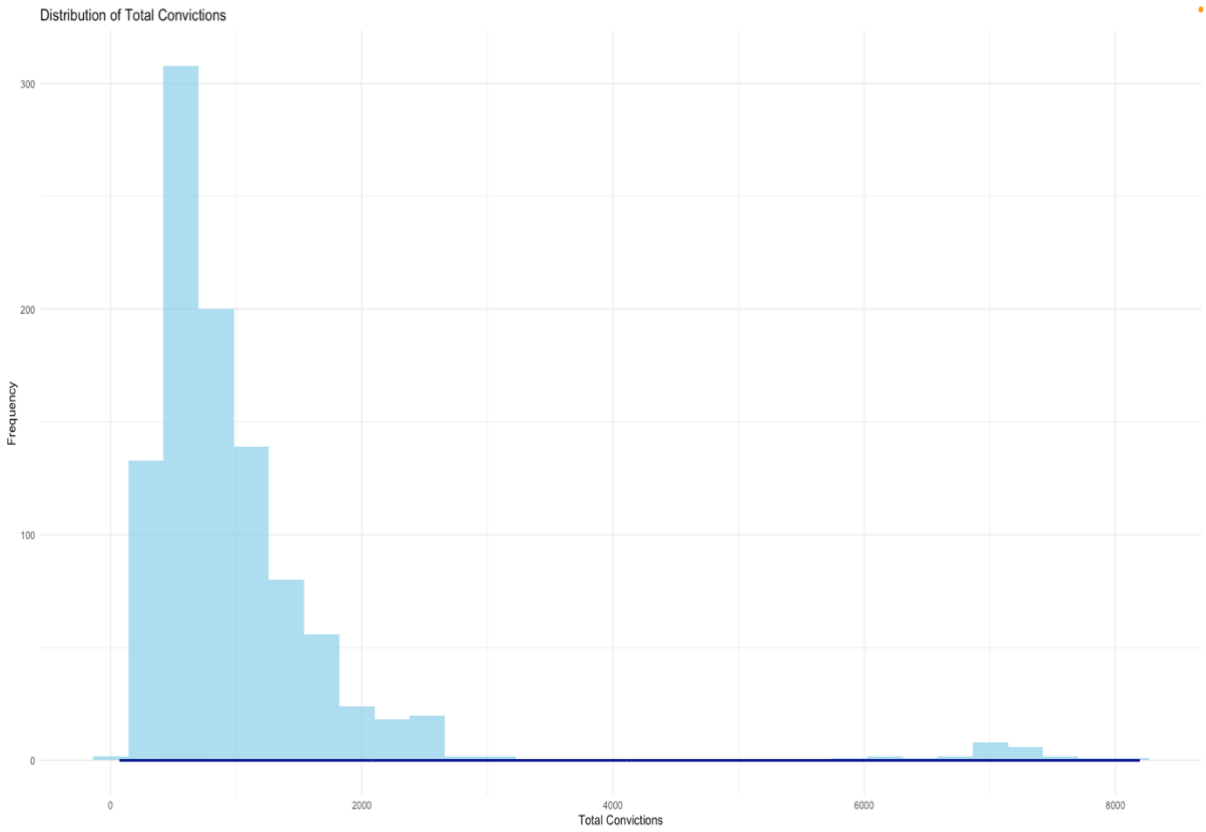

Figure 3.10: Distribution of Total Convictions

Seasonality was also investigated through a grouped bar chart (Setiawan and Suprihanto, 2021), comparing total convictions across the four meteorological seasons (Winter, Spring, Summer, Autumn), stratified by year. This visualisation was guided by the hypothesis that seasonal effects, such as public holidays, crime incidence patterns, or staffing levels, could impact CPS case volumes. Figure 3.11 highlights that, although the seasonal variations were subtle, a modest increase in convictions was observed during Autumn and Winter, particularly in 2014 and 2015 (e.g., Winter 2014: 26%; Autumn 2015: 29%). This was used to reflect end-of-year administrative pressures and an increase in case finalisation before the close of the judicial calendar.

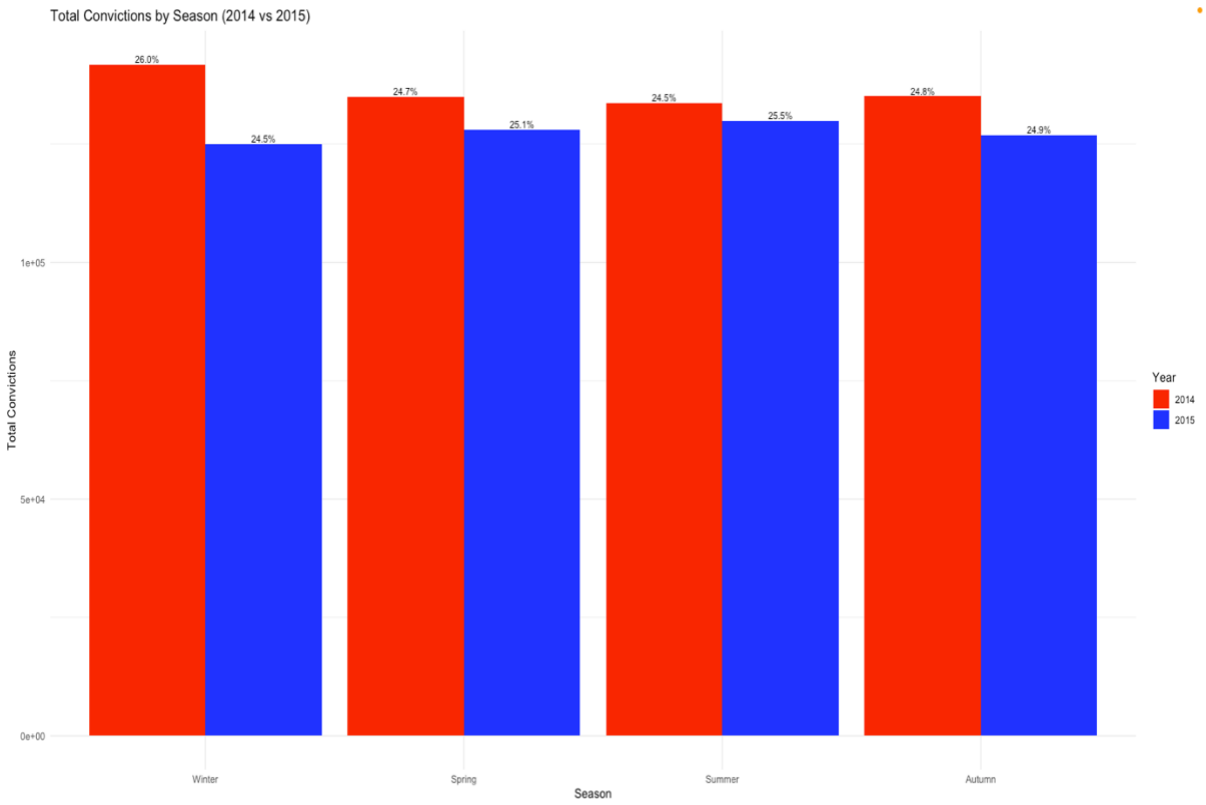

Figure 3.11: Total Convictions by Season (2014 / 2015)

To identify patterns in offence prevalence across regions, the top five offences by conviction count were plotted for each region using the code shown in Figure 3.12. Consistently, 'theft and handling', 'offences against the person', 'motoring offences', 'drug offences', and 'public order offences' featured as the most frequent convictions, although in varying order. This suggests that these offence categories represent a significant portion of the CPS's prosecutorial workload nationwide. To further examine spatial disparities, a comparative bar chart was generated for three of the most common offences—theft and handling, motoring offences, and public order offences—across regional groups (Figure 3.13). The North exhibited the highest conviction volumes for all three offence types, with theft and handling particularly dominant. In contrast, the East recorded the lowest conviction totals, suggesting either a lower incidence of these crimes or reduced prosecutorial throughput in that region. These findings can inform regional resource allocation and effective strategic planning.

```r
# ==== Top 5 Offences by Region ====
# Extract conviction columns
conviction_cols <- names(cps_df)[str_detect(names(cps_df), "number_of_.*_convictions")]

# Prepare data
top5_df <- cps_df %>%
  group_by(region) %>%
  summarise(across(all_of(conviction_cols), sum, na.rm = TRUE), .groups = "drop") %>%
  pivot_longer(-region, names_to = "offence", values_to = "convictions") %>%
  mutate(
    offence = offence %>%
      str_replace_all("number_of_|_convictions", "") %>%
      str_replace_all("_", " ") %>%
      str_to_title()
  ) %>%
  group_by(region) %>%
  slice_max(order_by = convictions, n = 5, with_ties = FALSE) %>%
  ungroup()

# Plot individually by region
unique_regions <- unique(top5_df$region)

for (reg in unique_regions) {
  region_data <- top5_df %>% filter(region == reg)

  p <- ggplot(region_data, aes(x = convictions, y = fct_reorder(offence, convictions))) +
    geom_col(fill = "#0073C2FF") +
    labs(
      title = paste("Top 5 Offences in", reg),
      x = "Total Convictions",
      y = "Offence Type"
    ) +
    theme_minimal(base_size = 14) +
    theme(
      plot.title = element_text(face = "bold", hjust = 0.5, size = 16),
      axis.text.y = element_text(size = 11),
      axis.text.x = element_text(size = 11)
    )

  print(p)
}
```

Figure 3.12: R code for analysing the Top 5 Offences by Region

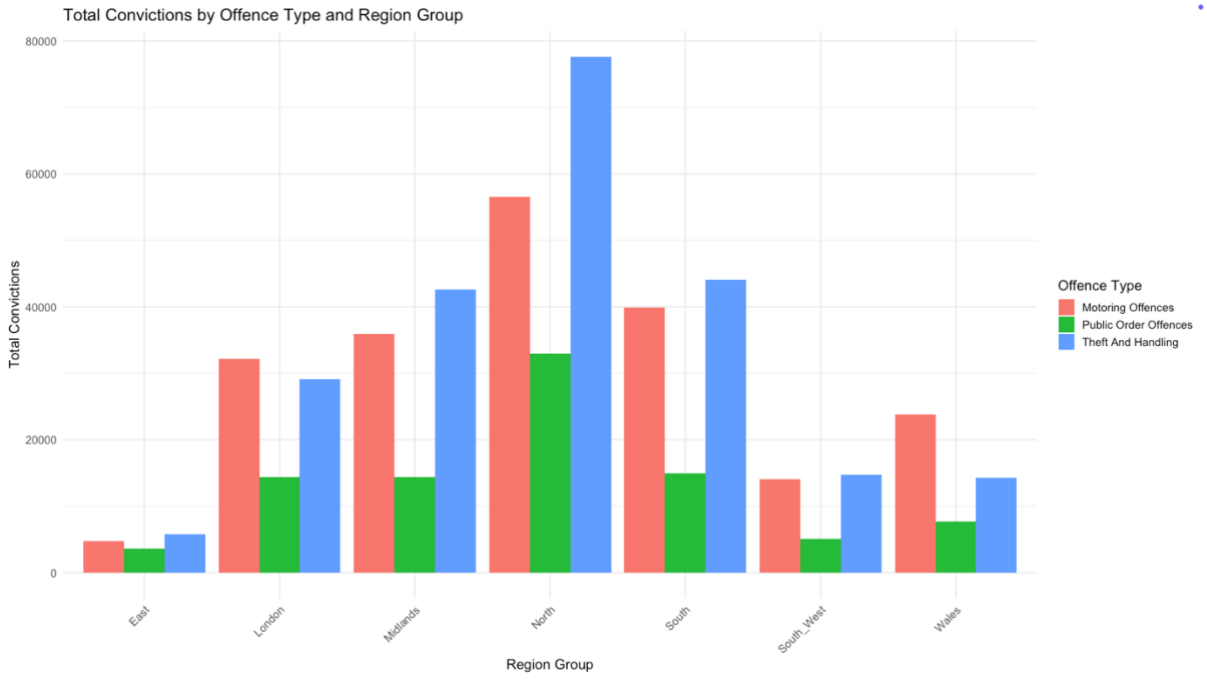

Figure 3.13: Total Convictions by Offence Type and Region Group

Conviction efficiency was then explored via a bar chart of total conviction success rates across all offence types over the 24-month period (Figure 3.14). This analysis revealed that drug offences, theft and handling, and motoring offences had the highest conviction success rates, which were above 75%, suggesting streamlined prosecution or strong evidentiary support in these categories. Conversely, sexual offences exhibited the lowest success rate below 75%, which may reflect greater evidentiary challenges, lower victim cooperation, or more complex legal proceedings.

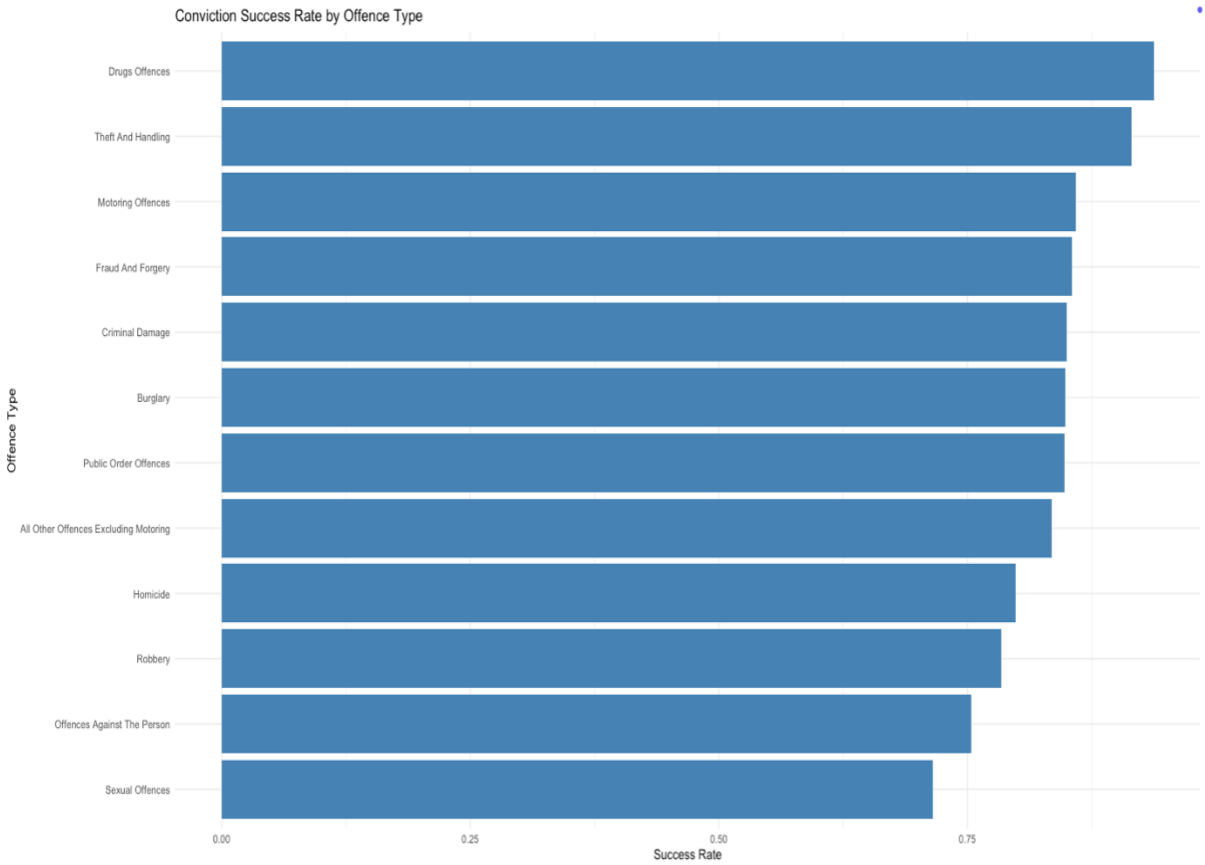

Figure 3.14: Conviction Success Rate by Offence Type

To capture national-level prosecution outcomes, a pie chart was constructed to represent the proportion of cases resulting in convictions, unsuccessful outcomes, and administrative finalisations (Figure 3.15). Administrative finalisations were separated due to their non-conviction nature, though conceptually they could be grouped under unsuccessful outcomes. The visualisation showed that approximately 81.9% of cases resulted in convictions, with 18.1% accounted for by unsuccessful outcomes and administrative finalisations combined. This suggests a relatively high overall success rate for CPS prosecutions during the analysed period.

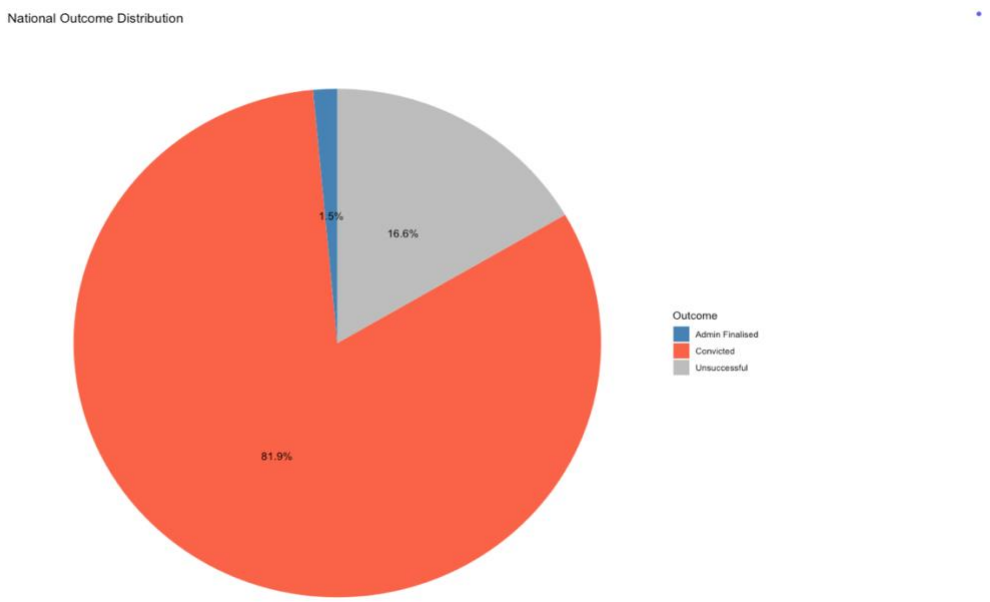

Figure 3.15: National Outcome Distribution

According to Zappia *et al.* (2025), to support machine learning processes and reduce redundancy among predictors, correlation heatmaps were generated for conviction counts, unsuccessful outcomes, and total case volumes, each segmented by offence type (Figures 3.16, 3.17, and 3.18). The heatmaps revealed strong correlations ($r > 0.9$) between 'offences against the person' and several other offence types, including 'sexual offences', 'burglary', 'theft and handling', 'fraud and forgery', 'criminal damage', 'drug offences', and 'public order offences. These high correlations were mirrored in the corresponding unsuccessful outcome variables, indicating shared structural patterns across offence categories. Notably, total conviction volumes for 'theft and handling' were also strongly correlated with 'burglary' and 'criminal damage', while 'drug offences' correlated highly with 'fraud and forgery' and 'robbery'.

Furthermore, none of the variable pairs in the three correlation heatmaps—covering conviction counts, unsuccessful outcomes, and total offence volumes—exhibited correlation coefficients below 0.5. These relationships informed feature selection strategies and highlighted the potential risk of multicollinearity in predictive models. Overall, this multi-layered EDA approach provided a comprehensive structural understanding of offence patterns, regional disparities, conviction efficiencies, and variable interdependencies in the CPS dataset. The insights directly informed feature engineering, supported hypothesis development, and laid a robust foundation for the subsequent application of statistical and machine learning techniques. This analysis also provides valuable evidence for policymakers seeking to improve prosecutorial equity and efficiency throughout the justice system.

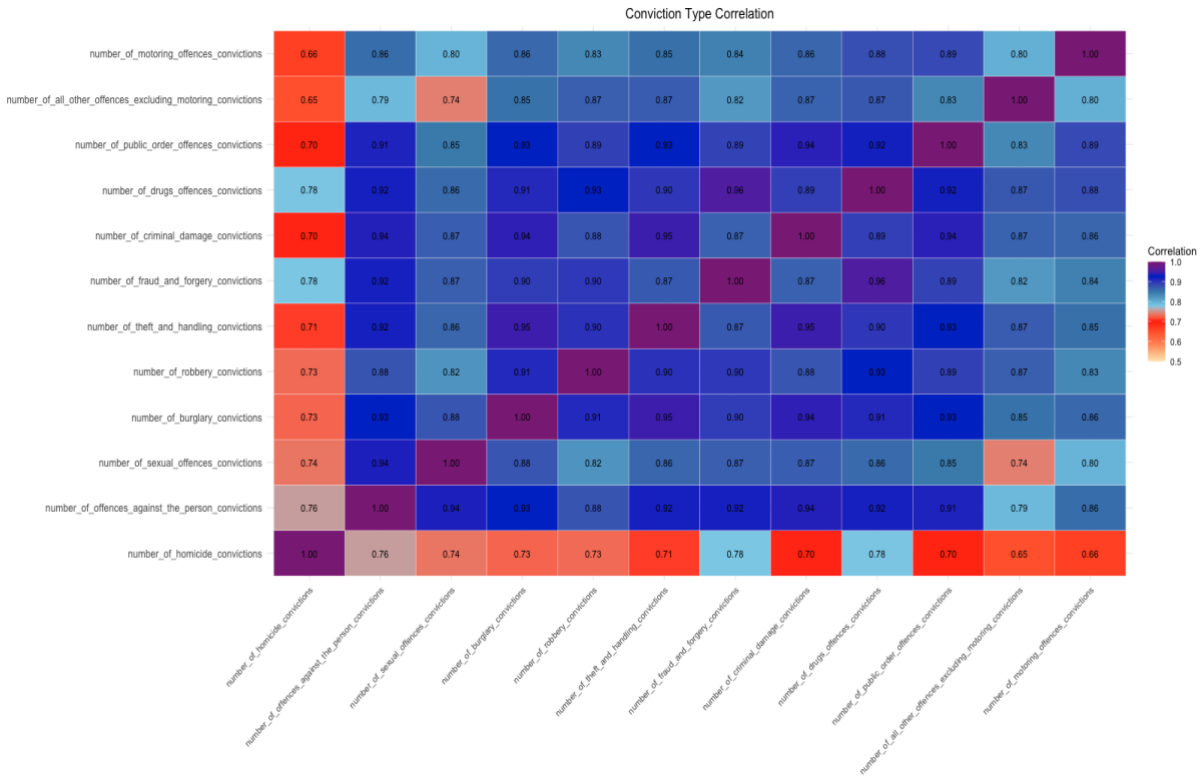

Figure 3.16: Conviction Type Correlation

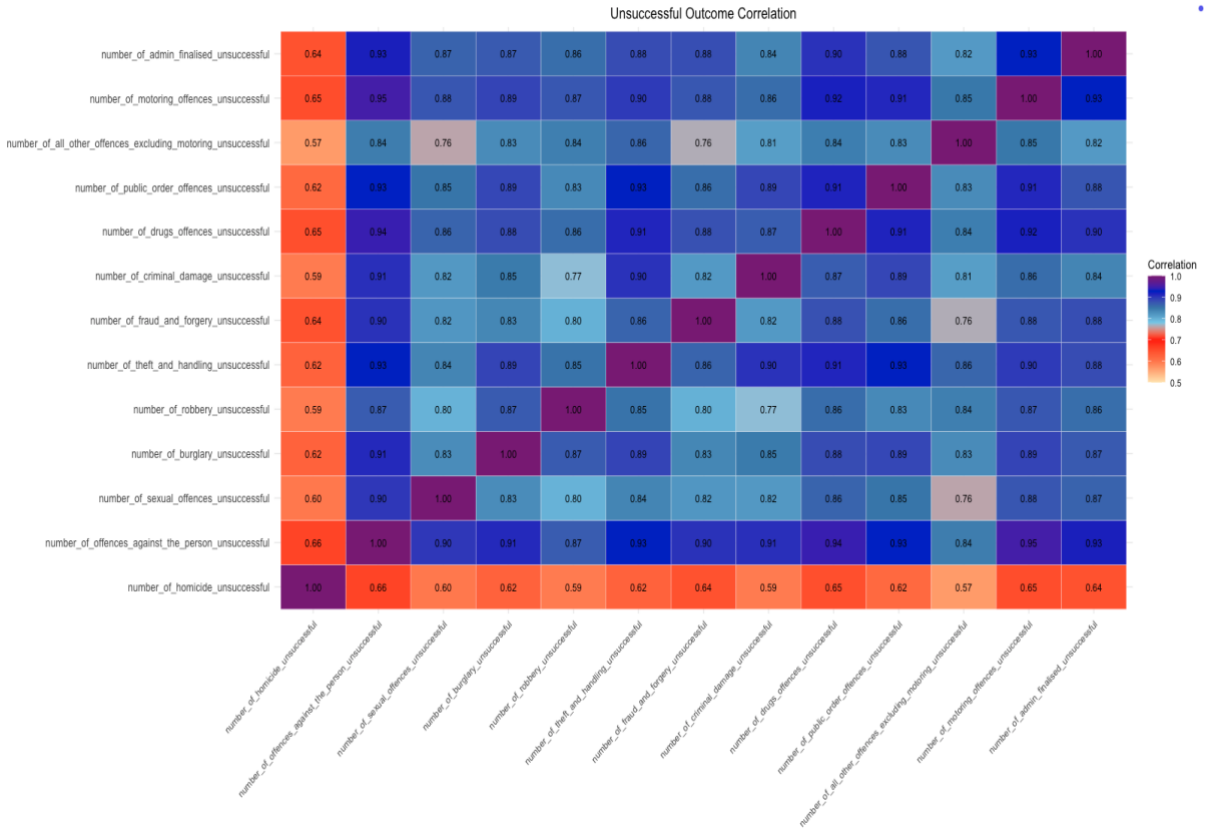

Figure 3.17: Unsuccessful Outcome Correlation

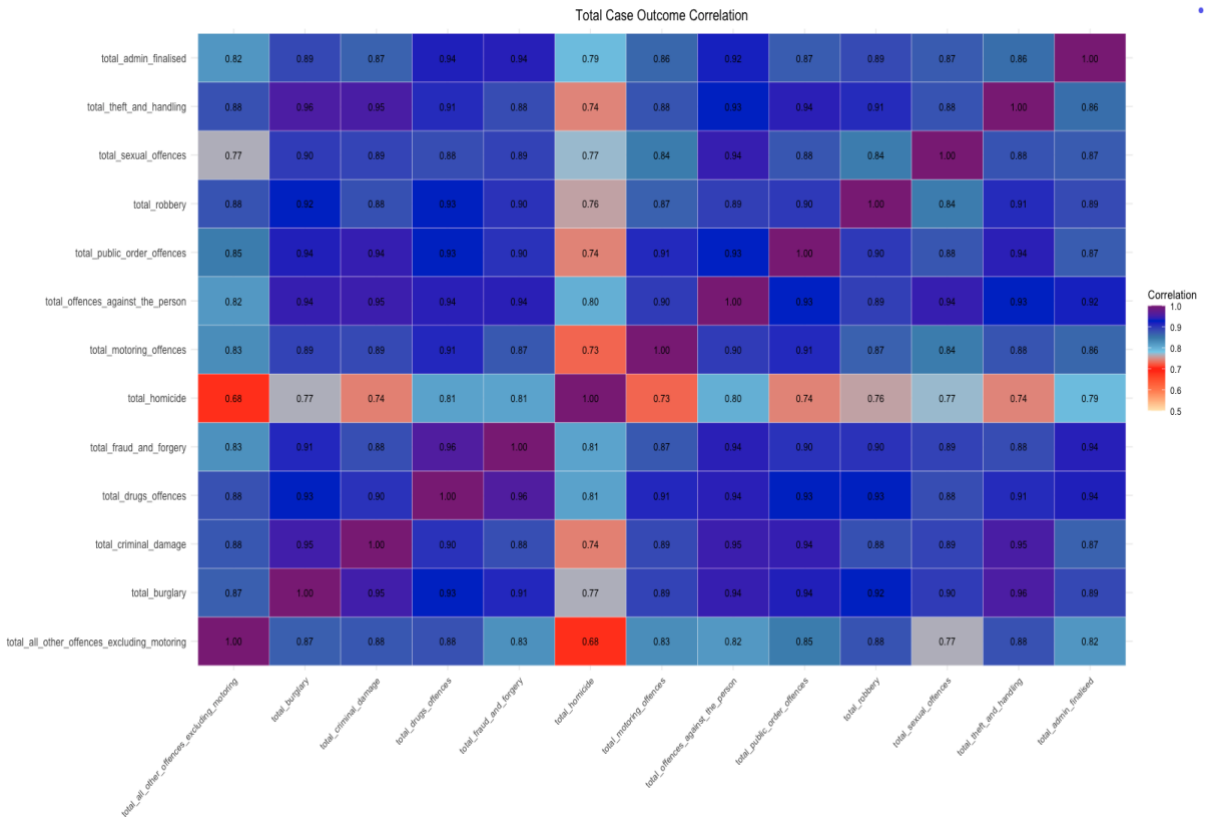

Figure 3.18: Total Case Outcome Correlation

According to Erdely and Rubio-Sánchez (2025), pair plots can be constructed to visually assess bivariate relationships among offence types, as shown in Figure 3.19. The plot revealed a general trend of positive correlation between most offence categories, suggesting that regions with higher conviction counts in one offence type (e.g., theft and handling) were also likely to report elevated counts in related categories (e.g., burglary, public order offences). This clustering of offences reflected underlying socio-criminal dynamics and shared regional reporting practices. The pair plot served as a useful exploratory tool to validate trends observed in the correlation heatmap and to inform further feature selection by identifying potentially redundant variables due to collinearity.

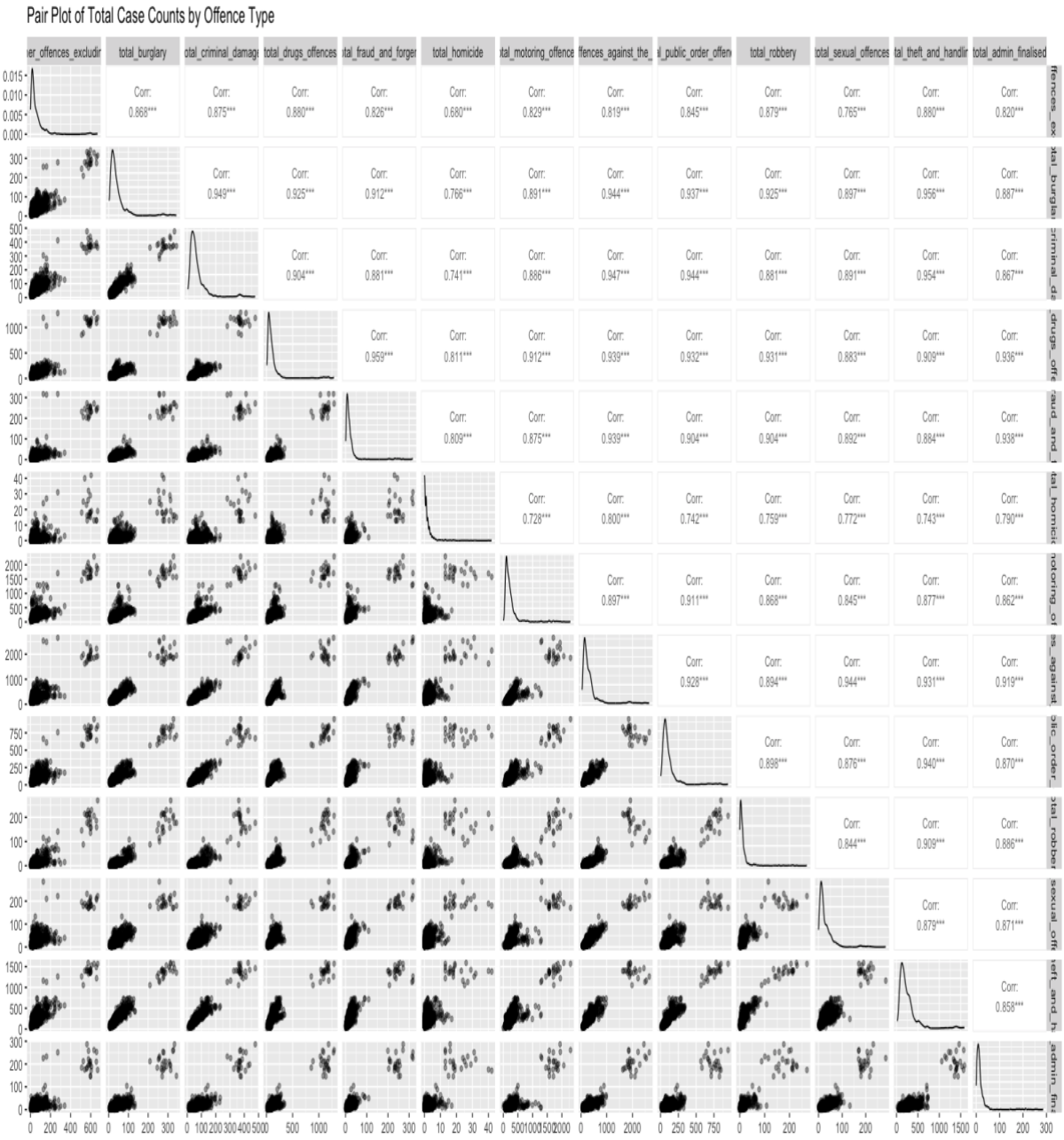

Figure 3.19: Pair Plot of Total Case Counts by Offence Type

4.0 Statistical Hypothesis Testing

Following exploratory data analysis (EDA), which revealed substantial disparities in the number of unsuccessful case outcomes across different offence types, as shown in Figure 4.1, formal statistical hypothesis testing was conducted to assess whether these observed differences were statistically significant or occurred by chance.

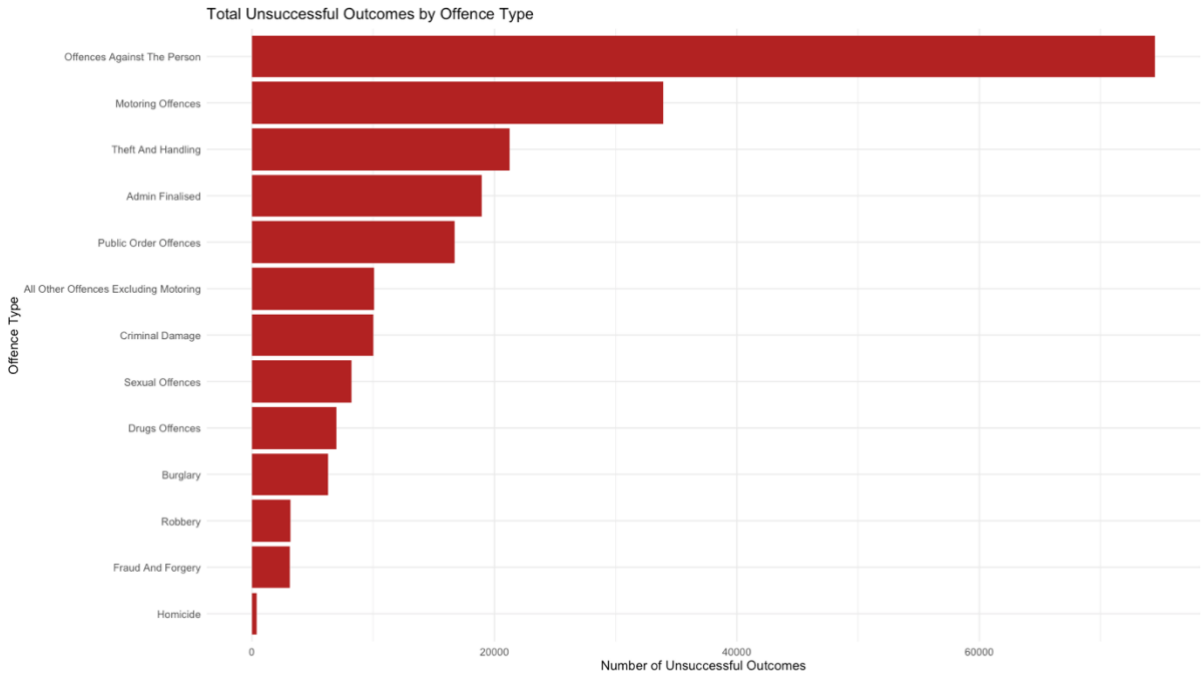

Figure 4.1: Total Unsuccessful Outcomes by Offence Type

4.1 Hypothesis Formulation

Null Hypothesis (H_0): There is no significant difference in the average number of unsuccessful prosecutions across offence types.

Alternative Hypothesis (H_1): There is a significant difference in the average number of unsuccessful prosecutions across offence types.

This hypothesis was based on visual trends that suggested varying prosecution success rates across offences such as sexual offences, motoring offences, and offences against the person. These inconsistencies highlighted the need for further testing to support data-driven conclusions for stakeholders in the criminal justice system.

4.2 ANOVA and Kruskal-Wallis

Using the code as shown below (Figure 4.2), the dataset was first reshaped, isolating 13 distinct offence types (e.g., Homicide, Sexual offences, Theft and Handling), all of which relate to unsuccessful prosecutions. Numeric cleaning and filtering ensured comparability across all rows. As the dependent variable, number of unsuccessful prosecutions, was not guaranteed to follow a normal distribution across all offence groups, both parametric and non-parametric tests were employed to validate robustness:

One-Way ANOVA (parametric): Assumes homogeneity of variance and normality of residuals (Kim and Cribbie, 2017).

Kruskal-Wallis Test (non-parametric): A rank-based test used to validate significance where ANOVA assumptions may not hold (Ostertagová, Ostertag and Kováč, 2014).

```r
## 5. HYPOTHESIS TESTING USING ANOVA
# === Null Hypothesis (H₀):
# There is no significant difference in the average number of unsuccessful case outcomes
# across offence types (e.g., sexual offences, drug offences, fraud).

# Alternative Hypothesis (H₁):
# There is a significant difference in the average number of unsuccessful case outcomes
# across offence types.

# Filter out unwanted columns
exclude_keywords <- c("percentage", "total", "unsuccessful_ratio")

unsuccessful_cols <- names(cps_df)[
  str_detect(names(cps_df), regex("unsuccessful", ignore_case = TRUE)) &
    !str_detect(names(cps_df), regex(paste(exclude_keywords, collapse = "|"), ignore_case = TRUE)) &
    sapply(cps_df, is.numeric)
]

# Reshape the data to long format
df_melted <- cps_df %>%
  select(all_of(unsuccessful_cols)) %>%
  pivot_longer(cols = everything(), names_to = "Offence", values_to = "Unsuccessful") %>%
  mutate(
    Offence = str_to_title(
      str_replace_all(Offence, c("number_of_" = "", "_unsuccessful" = "", "_" = " "))
    )
  ) %>%
  drop_na()

# View included offence types
cat("Final offence types included:\n")
unique_offences <- unique(df_melted$Offence)
for (i in seq_along(unique_offences)) {
  cat(paste0(i, ". ", unique_offences[i], "\n"))
}

# Run ANOVA and Kruskal-Wallis tests
anova_result <- df_melted %>%
  anova_test(Unsuccessful ~ Offence)

kruskal_result <- df_melted %>%
  kruskal_test(Unsuccessful ~ Offence)

cat("\nANOVA Results:\n")
print(anova_result)
```

Figure 4.2: R code for Anova and Kruskal-Wallis Testing

Results

Degrees of freedom (df) = 12, p = p-value (probability value), χ^2 = (Chi-square statistic) and η^2 = effect size.

The ANOVA test reported a statistically significant difference in mean unsuccessful prosecutions across offence types:

$$F(12, 13091) = 306.52,\ p < 0.001,\ \eta^2 = 0.219.$$

This indicates a large effect size, with approximately 22% of the variation in unsuccessful outcomes attributable to offence type.

The Kruskal-Wallis test confirmed these findings non-parametrically:

$$\chi^2(12) = 7516,\ p < 0.001.$$

This result indicates that the differences between the groups are statistically significant. The p-value < 0.001 suggests that the probability of these differences occurring due to random chance is extremely low. Because the Kruskal-Wallis test does not assume normality, it provides a robust confirmation of the ANOVA findings, strengthening the conclusion that offence type significantly influences the rate of unsuccessful prosecutions.

4.3 Post-Hoc Analysis

To determine which specific offence categories differed significantly in terms of unsuccessful prosecution outcomes, Dunn's test with Bonferroni correction was conducted following a statistically significant Kruskal-Wallis result (Dinno, 2015). This post-hoc analysis revealed several notable patterns:

Sexual offences were statistically more likely to result in unsuccessful outcomes than many other offence types. Significant differences were observed when compared to offences against the person (Z = 36.378, p.adj < 0.001), motoring offences (Z = 25.613, p.adj < 0.001), theft and handling (Z = -19.849, p.adj < 0.001), robbery (Z = -15.284, p.adj < 0.001), fraud and forgery (Z = -14.050, p.adj < 0.001), and public order offences (Z = 14.996, p.adj < 0.001). Although sexual offences also differed from drug offences (Z = -2.602, p.unadj = 0.009), this was not statistically significant after adjustment (p.adj = 0.723). These results suggest that sexual offences had consistently poorer prosecution outcomes than most other major categories.

Motoring offences and homicide were associated with substantially lower numbers of unsuccessful prosecutions, reflecting potentially higher conviction efficiency. Key comparisons included homicide versus offences against the person ($Z = -62.862$, $p.adj < 0.001$), homicide versus motoring offences ($Z = -52.097$, $p.adj < 0.001$), and homicide versus theft and handling ($Z = -46.333$, $p.adj < 0.001$). Additionally, motoring offences differed significantly from offences against the person ($Z = -10.765$, $p.adj < 0.001$) and burglary ($Z = -28.997$, $p.adj < 0.001$). Several other offence types were frequently associated with higher rates of unsuccessful outcomes. For example, fraud and forgery showed substantial differences when compared to motoring offences ($Z = -39.663$, $p.adj < 0.001$), sexual offences ($Z = -14.050$, $p.adj < 0.001$), and theft and handling ($Z = -33.899$, $p.adj < 0.001$). Likewise, offences against the person significantly differed from robbery ($Z = 51.662$, $p.adj < 0.001$) and theft and handling ($Z = 16.529$, $p.adj < 0.001$).

The largest observed absolute Z-score was found between homicide and offences against the person ($Z = -62.862$, $p.adj < 0.001$), indicating a significant disparity in unsuccessful outcomes between two high-severity categories. Another substantial result involved admin finalised and homicide ($Z = 41.487$, $p.adj < 0.001$), though this likely reflects a procedural classification difference rather than an actual case disposition. Out of the 78 pairwise comparisons, the overwhelming majority yielded adjusted p-values below 0.001, confirming that the rates of unsuccessful prosecution differ significantly across offence types. These disparities reinforce the importance of offence-specific interventions and evidence-based reforms within the criminal justice system. Table 4.1 presents summaries of the results from selected pairs in this analysis.

Table 4.1: Selected Post-Hoc Dunn's Test Results for Unsuccessful Outcomes pairs by Offence Type

Comparison	Z	P.unadj	P.adj
Homicide – Offences Against The Person	-62.862	0.000	0.000
Homicide – Motoring Offences	-52.097	0.000	0.000
Offences Against The Person – Sexual Offences	36.378	0.000	0.000
Motoring Offences – Sexual Offences	25.613	0.000	0.000

Theft and Handling – Sexual Offences	-19.849	0.000	0.000
Fraud and Forgery – Motoring Offences	-39.663	0.000	0.000
Fraud and Forgery – Sexual Offences	-14.050	0.000	0.000
Offences Against The Person – Robbery	51.662	0.000	0.000
Admin Finalised – Homicide	41.487	0.000	0.000
Motoring Offences – Offences Against The Person	-10.765	0.000	0.000
Robbery – Sexual Offences	-15.284	0.000	0.000
Homicide – Sexual Offences	-26.484	0.000	0.000
Public Order Offences – Sexual Offences	14.996	0.000	0.000
Drugs Offences – Sexual Offences	-2.602	0.009	0.723
Burglary – Motoring Offences	-28.997	0.000	0.000
Motoring Offences – Robbery	40.896	0.000	0.000
Homicide – Theft and Handling	-46.333	0.000	0.000
Offences Against The Person – Theft and Handling	16.529	0.000	0.000
Drugs Offences – Theft and Handling	-22.451	0.000	0.000
Fraud and Forgery – Theft and Handling	-33.899	0.000	0.000
Public Order Offences – Theft and Handling	-4.854	0.000	0.000
Burglary – Sexual Offences	-3.385	0.001	0.056

4.4 Interpretation

The statistical tests provide strong evidence to reject the null hypothesis (H_0), which posits that there is no significant difference in the average number of unsuccessful prosecutions across offence types. Both the one-way ANOVA ($F(12, 13091) = 306.52$, $p < 0.001$, $\eta^2 = 0.219$) and the Kruskal-Wallis test ($\chi^2(12) = 7516$, $p < 0.001$) revealed statistically significant differences in prosecution outcomes across the 13 offence categories. The substantial effect size suggests that approximately 22% of the variation in unsuccessful prosecutions can be attributed to offence type, confirming that the observed disparities are not due to random variation but reflect structural or procedural differences in how different offences are handled. The results of this statistical analysis also highlight a clear need for offence-specific prosecution strategies and differentiated resource allocation across the criminal justice system.

Notably, sexual offence cases—despite their severity—were consistently associated with higher rates of unsuccessful prosecution outcomes. This trend may reflect structural challenges such as limited physical evidence, victim non-engagement, or complex legal thresholds. These findings point to the urgent need for specialised support services, enhanced victim protection, tailored investigative protocols, and targeted legal reforms to improve outcomes in these cases. Conversely, the consistently high conviction success observed in motoring and drug-related offences suggests these cases benefit from clearer evidentiary standards, streamlined procedures, or more readily available forensic data. These characteristics could serve as benchmarks for improving case management practices in more complex offence types.

These findings support the stratification of offences by type in predictive model design. Since outcome variability appears strongly offence-dependent, stratification can help reduce bias and improve predictive accuracy by accounting for structural differences between case types. For policy-makers, legal practitioners, and justice system stakeholders, this evidence provides a quantitative foundation for prioritising prosecutorial resources and designing interventions in areas with persistently high rates of unsuccessful prosecutions. Such targeted reforms may help advance the goals of procedural fairness, resource efficiency, and equitable access to justice nationwide.

5.0 Predictive Modelling: Regression

In this research, regression analysis is employed as a supervised machine learning technique to model the relationship between a dependent variable (total unsuccessful outcomes) and several independent variables (conviction counts for violent offences).

5.1 Hypothesis

H_0 (Null Hypothesis): Violent offence conviction counts do not significantly predict the total number of unsuccessful outcomes.

H_1 (Alternative Hypothesis): Violent offence conviction counts significantly predict the total number of unsuccessful outcomes.

Figure 5.1 illustrates the code used to prepare the dataset for regression analysis, selecting four variables related to violent offences as predictors and the total number of unsuccessful outcomes as the target variable. It creates a copy of the dataset, extracts only the necessary columns, and verifies that any rows with missing values are removed to ensure the model is trained on complete and relevant data.

```r
## 6. REGRESSION TECHNIQUE
# Predict the total number of unsuccessful outcomes using only violent offence conviction counts as predictors.

# === Prepare the data ===
cps_lr <- cps_df

# Select predictors and target
predictors <- c(
  "number_of_homicide_convictions",
  "number_of_offences_against_the_person_convictions",
  "number_of_sexual_offences_convictions",
  "number_of_robbery_convictions"
)
target <- "total_unsuccessful"

# Filter complete cases
data <- cps_lr %>% select(all_of(c(predictors, target))) %>% drop_na()

# Split into train/test (70/30)
set.seed(42)
train_index <- createDataPartition(data[[target]], p = 0.7, list = FALSE)
train_data <- data[train_index, ]
test_data  <- data[-train_index, ]

X_train <- as.matrix(train_data[, predictors])
y_train <- train_data[[target]]
X_test  <- as.matrix(test_data[, predictors])
y_test  <- test_data[[target]]

# === Train models ===
lm_model <- lm(total_unsuccessful ~ ., data = train_data)
ridge_model <- cv.glmnet(X_train, y_train, alpha = 0)
lasso_model <- cv.glmnet(X_train, y_train, alpha = 1)

# === Predict & Evaluate ===
predict_and_evaluate <- function(model, model_type) {
  if (model_type == "Linear") {
    preds <- predict(model, newdata = test_data)
  } else {
    preds <- predict(model, newx = X_test, s = "lambda.min")
  }
  rmse_val <- rmse(y_test, preds)
  mae_val  <- mae(y_test, preds)
  r2_val   <- caret::R2(preds, y_test)
```

Figure 5.1: R code for Regression Analysis

The primary aim is to evaluate how effectively violent crime convictions can predict prosecution inefficiency, operationalised as the number of unsuccessful case outcomes. This section contributes to the overarching research objective of applying interpretable machine learning to identify systemic inefficiencies within the Crown Prosecution Service (CPS). The model estimates the total number of unsuccessful prosecutions based on four key offence types, including the number of homicide convictions, the number of sexual offence convictions, the number of robbery convictions, and the number of offences against the person. These categories were selected based on both theoretical relevance and preliminary correlation analysis. All pairwise correlations between predictors were below 0.9, confirming the absence of multicollinearity and supporting both model stability and interpretability (Desboulets, 2018).

Consequently, the modelling approach provides valuable insights into which specific types of violent offences are most strongly associated with prosecution failure, thereby informing performance evaluation and evidence-based strategic planning within the Crown Prosecution Service (CPS). In addition to quantifying the influence of violent offence conviction rates on CPS prosecution inefficiency—given the complexity, resource demands, and high-stakes nature of such cases—the analysis also aimed to inform operational decision-making by identifying areas where targeted interventions, such as enhanced training, procedural improvements, or strategic resource allocation, could strengthen conviction outcomes. By centring the analysis on offences known for their legal complexity and societal impact, the study maintains a balance between methodological rigour and practical relevance.

5.2 Data Preparation

The dataset was filtered to retain only the selected variables. A 70/30 train-test split was applied, and predictor matrices were formatted appropriately for penalised regression algorithms (Vrigazova, 2021).

5.3 Model Development

Three regression models were trained and compared:

Linear Regression – to serve as a baseline (Maulud and Abdulazeez, 2020).

Ridge Regression – incorporates L2 regularisation to manage multicollinearity (Magklaras *et al.,* 2024).

Lasso Regression – uses L1 regularisation to enhance interpretability and perform feature selection (Bhattacharyya, 2025).

5.4 Model Evaluation

The model's performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R-squared, R^2), as highlighted by Chicco, Warrens, and Jurman (2021). The results are summarised in Table 5.1.

Table 5.1: Regression Performance Table

Model	**RMSE**	**MAE**	**R^2**
Linear	79.23	48.68	0.920
Ridge	84.50	49.49	0.913
Lasso	79.46	48.53	0.920

As shown above, Linear and Lasso regression achieved the highest R^2 (0.920), indicating that the selected predictors explain 92.0% of the variance in unsuccessful outcomes. Ridge regression yielded slightly lower accuracy (91.3%). Also, to assess model generalisability, 5-fold cross-validation was applied. Results are shown in Table 5.2.

Table 5.2: Cross-Validation Results

Model	**Mean R^2**	**SD R^2**
Linear	0.936	0.021
Ridge	0.940	0.034
Lasso	0.939	0.020

All three models demonstrated high consistency across folds, confirming the robustness of the model. Ridge had the highest average R^2 value of 0.94 among the three models, and it also exhibited a slightly greater variance.

5.5 Discussion of Results

The model coefficients, summarised in Table 5.3, provide insight into each predictor's contribution and feature importance.

Table 5.3: Model Coefficients Across Techniques

Feature	**Linear**	**Ridge**	**Lasso**
Homicide convictions	9.834	12.859	9.539
Offences against the person convictions	0.746	0.502	0.750
Sexual offences convictions	0.285	2.227	0.230
Robbery convictions	4.467	4.242	4.435

Homicide convictions consistently emerged as the strongest predictor, likely due to the inherent complexity and risk of trial failure in such cases. Robbery convictions also carried significant weight, while sexual offences contributed less, particularly in the Lasso model, which reduced its coefficient to close to zero, indicating potential redundancy or weaker predictive power relative to other variables. The actual vs. predicted plots showed that the predicted values clustered closely around the diagonal reference line, indicating a strong linear relationship and high predictive accuracy (Figure 5.2). This alignment suggests that the models were effective in capturing the underlying patterns in the data. Additionally, the residual plots exhibited a random scatter of residuals around the horizontal line at zero, with no discernible patterns or funnel shapes (Figure 5.3). This supports the assumptions of homoscedasticity (constant variance of errors) and normality of residuals, reinforcing the validity of the linear regression approach and confirming that model errors were randomly distributed.

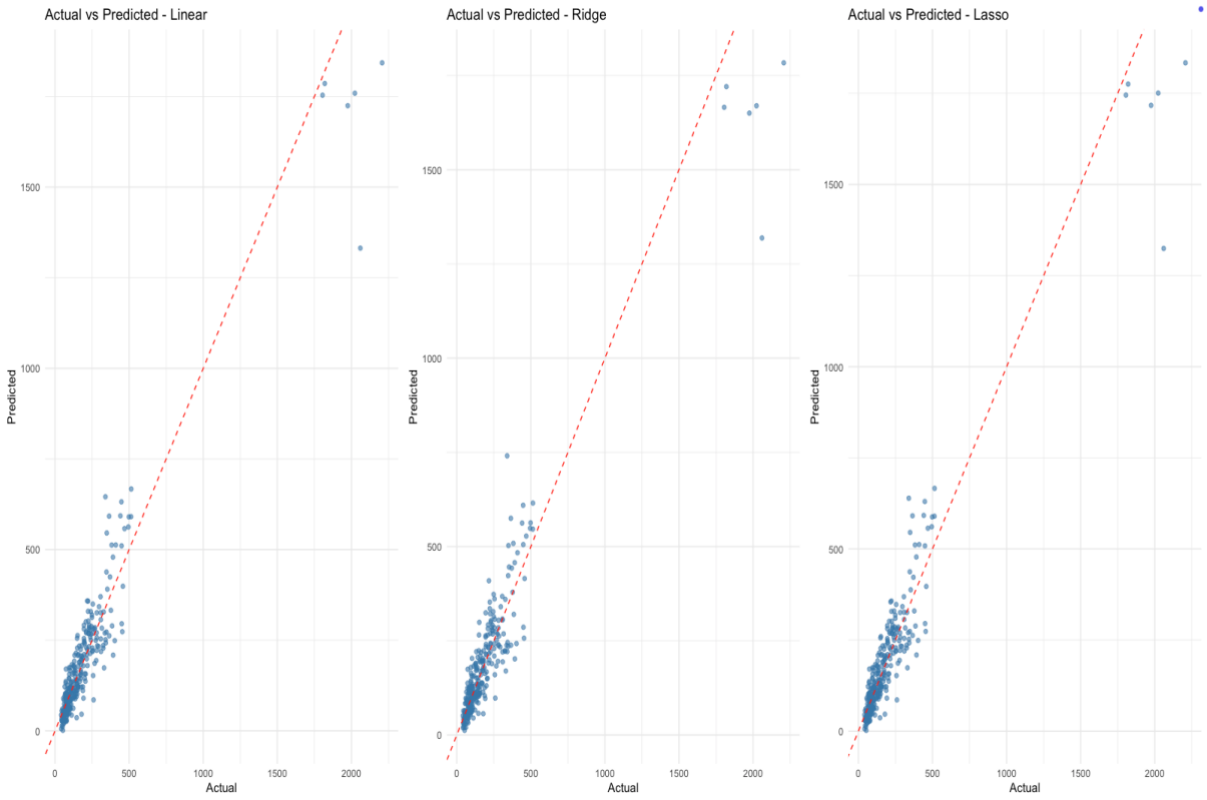

Figure 5.2: Plots of Actual against Predicted Values across all Regression Models

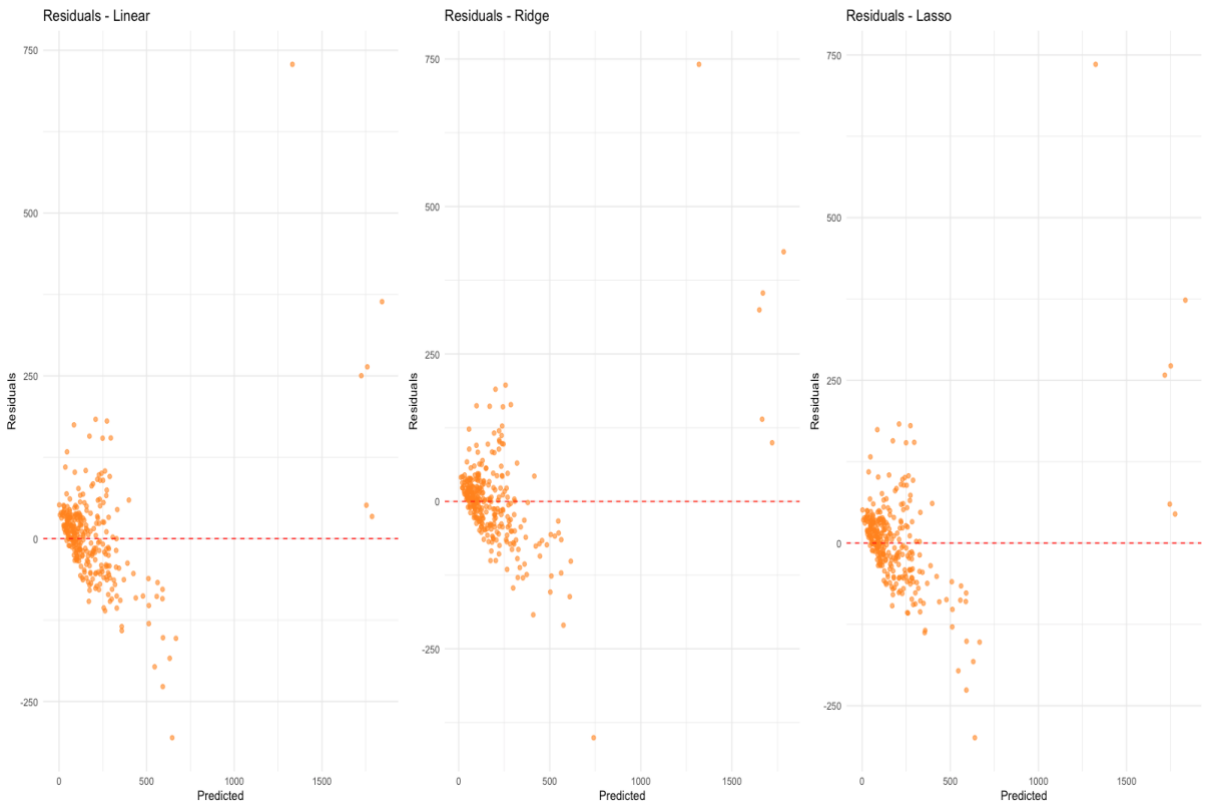

Figure 5.3: Plots of Residuals across all Regression Models

5.6 Interpretations

The regression results provide strong empirical support for rejecting the null hypothesis (H_0), affirming that conviction counts for violent offences, particularly homicide and robbery, significantly predict the total number of unsuccessful prosecution outcomes. This finding supports the alternative hypothesis (H_1) and suggests that specific types of violent offences are associated with higher prosecution failure rates. Among the predictors, homicide convictions consistently showed the highest positive association with unsuccessful outcomes across all models. This may reflect the legal and evidentiary complexity inherent in such cases, which often require intricate forensic analysis and involve high-risk trial dynamics. Consequently, these cases warrant enhanced procedural oversight, possibly through targeted reviews, forensic quality audits, or improved trial preparation protocols.

Robbery convictions also emerged as a strong predictor, indicating that this offence category may be vulnerable to similar inefficiencies or systemic pressures. Addressing these issues through specialised legal training, effective evidentiary guidelines, or case screening procedures could help reduce prosecutorial failure rates in these areas (Rossmo and Pollock, 2019). However, sexual offence convictions had a lesser predictive impact, particularly in the Lasso regression model, where its coefficient was significantly reduced. This suggests that while relevant, its influence may be comparatively lower when controlling for other types of violent offences, or that more dominant predictors minimise its effect.

The models' high R-squared values (up to 0.94 in cross-validation) and the random distribution of residuals further validate the reliability and generalisability of the findings. The close alignment of predicted and actual values suggests that the models effectively capture the underlying patterns, offering a high degree of explanatory power without overfitting. Overall, this analysis demonstrates how interpretable machine learning techniques can provide actionable insights into the structural inefficiencies of public legal systems. By linking specific offence conviction patterns to prosecution outcomes, the models support evidence-based decision-making in areas such as workload forecasting, strategic resource planning, and targeted procedural reforms, contributing to a more effective and equitable Crown Prosecution Service (CPS).

6.0 Clustering

In addition to regression analysis, clustering offers a powerful unsupervised learning approach for uncovering hidden patterns and segmenting data based on feature similarity. It identifies natural groupings by measuring the closeness of data points to one another, without relying on predefined labels (Naeem *et al.,* 2023). In this study, clustering was applied to analyse CPS regions based on offence and outcome profiles, enabling the discovery of underlying structural similarities. This technique is particularly valuable for exploratory analysis, as it supports regional benchmarking, resource allocation, and strategic planning by revealing latent similarities or outliers in prosecutorial workloads. Ultimately, clustering enhances the interpretive depth of the analysis and delivers actionable insights for policy and operational decisions within the criminal justice system (Naeem *et al.,* 2023).

6.1 Hypothesis

Null Hypothesis (H_0): There is no meaningful regional grouping based on offence types.

Alternative Hypothesis (H_1): Regions can be grouped into meaningful clusters based on offence types, reflecting systematic similarities or differences.

To determine whether meaningful groupings exist among CPS regions based on offence type patterns, clustering analysis was conducted using three unsupervised learning techniques: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. The hypothesis guiding this analysis posited that regional offence profiles may contain systematic similarities, which, if successfully identified, could inform CPS strategy, resourcing, and further statistical modelling.

The top five offences identified earlier in Figure 3.12, including theft and handling, offences against the person, motoring offences, drug offences, and public order offences, were selected for clustering analysis. For each category, total offence counts were computed by aggregating convictions and unsuccessful outcomes to capture the full prosecutorial workload. These aggregated figures were then averaged by region to account for differences in frequency and population, and subsequently standardised using z-score scaling to ensure comparability across features and prevent dominance by variables with larger magnitudes (Figure 6.1) (Wongoutong, 2024). After this process, various clustering models were performed on the dataframe.

```r
## 7. CLUSTERING
# Null Hypothesis (H₀): There is no meaningful regional grouping based on offence types
# Alternative Hypothesis (H₁): Regions can be grouped into meaningful clusters based on offence types, which reflect systematic similarities or differences.

#  K-MEans Clustering
# === Create total offence columns ===
cps_ml <- cps_df %>%
  mutate(
    total_theft_offences = number_of_theft_and_handling_convictions + number_of_theft_and_handling_unsuccessful,
    total_violence_offences = number_of_offences_against_the_person_convictions + number_of_offences_against_the_person_unsuccessful,
    total_motoring_offences = number_of_motoring_offences_convictions + number_of_motoring_offences_unsuccessful,
    total_drugs_offences = number_of_drugs_offences_convictions + number_of_drugs_offences_unsuccessful,
    total_public_order_offences = number_of_public_order_offences_convictions + number_of_public_order_offences_unsuccessful
  )

# === Aggregate by region ===
region_cluster_df <- cps_ml %>%
  group_by(region) %>%
  summarise(
    across(starts_with("total_"), mean, na.rm = TRUE),
    .groups = "drop"
  )

# === Scale features ===
offence_columns <- c(
  "total_theft_offences",
  "total_violence_offences",
  "total_motoring_offences",
  "total_drugs_offences",
  "total_public_order_offences"
)
scaled_matrix <- scale(region_cluster_df[, offence_columns])

# === Elbow and Silhouette Methods ===
wss <- vector()
silhouette_scores <- vector()

for (k in 2:10) {
  kmeans_model <- kmeans(scaled_matrix, centers = k, nstart = 25)
  wss[k] <- kmeans_model$tot.withinss
  sil <- silhouette(kmeans_model$cluster, dist(scaled_matrix))
  silhouette_scores[k] <- mean(sil[, 3])
}

# Plot Elbow Method
```

Figure 6.1: R code for Clustering Analysis

6.2 K-Means Clustering

K-Means is a partition-based clustering algorithm that groups observations into k non-overlapping clusters by minimising the within-cluster sum of squares (WCSS). The algorithm operates iteratively: initial centroids are assigned, data points are grouped based on Euclidean distance, and centroids are recalculated until convergence (Ikotun *et al.,* 2022). When applied to the scaled CPS regional offence dataset, the optimal number of clusters was determined to be k = 3. This decision was supported by both the Elbow Method, which showed a visible inflection at k = 3, and the Silhouette Analysis, which yielded an average silhouette score of 0.5434, indicating moderate intra-cluster cohesion and inter-cluster separation (Figures 6.2 and 6.3). A pairwise scatter plot matrix was constructed to visualise the clustering structure across combinations of total offence types. The plots revealed that variables such as total theft offences and drug offences, as well as violence offences and public order offences, were positively correlated (Figures 6.4 and 6.5).

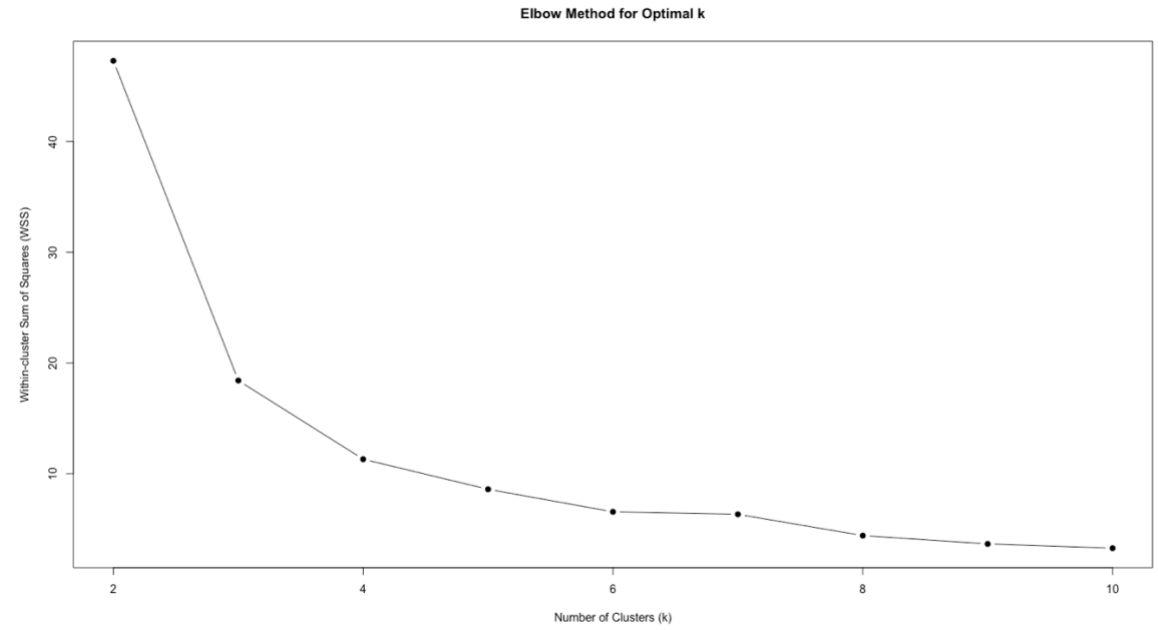

Figure 6.2: Elbow Method for Optimal k

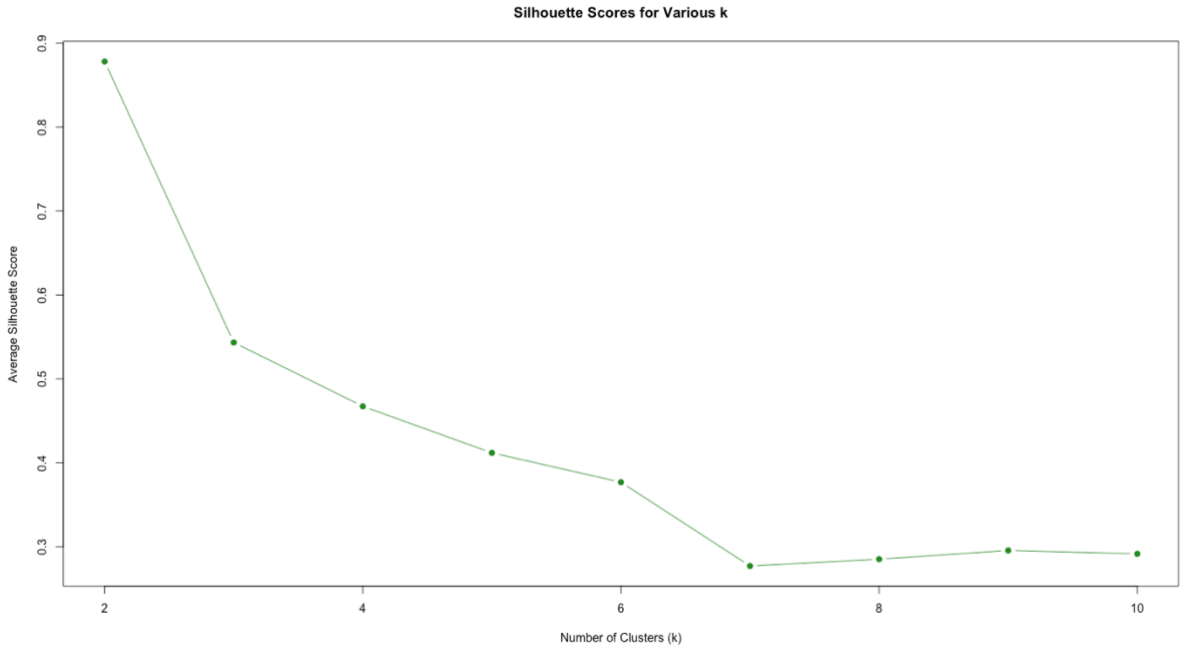

Figure 6.3: Silhouette Scores for Various k (K-means)

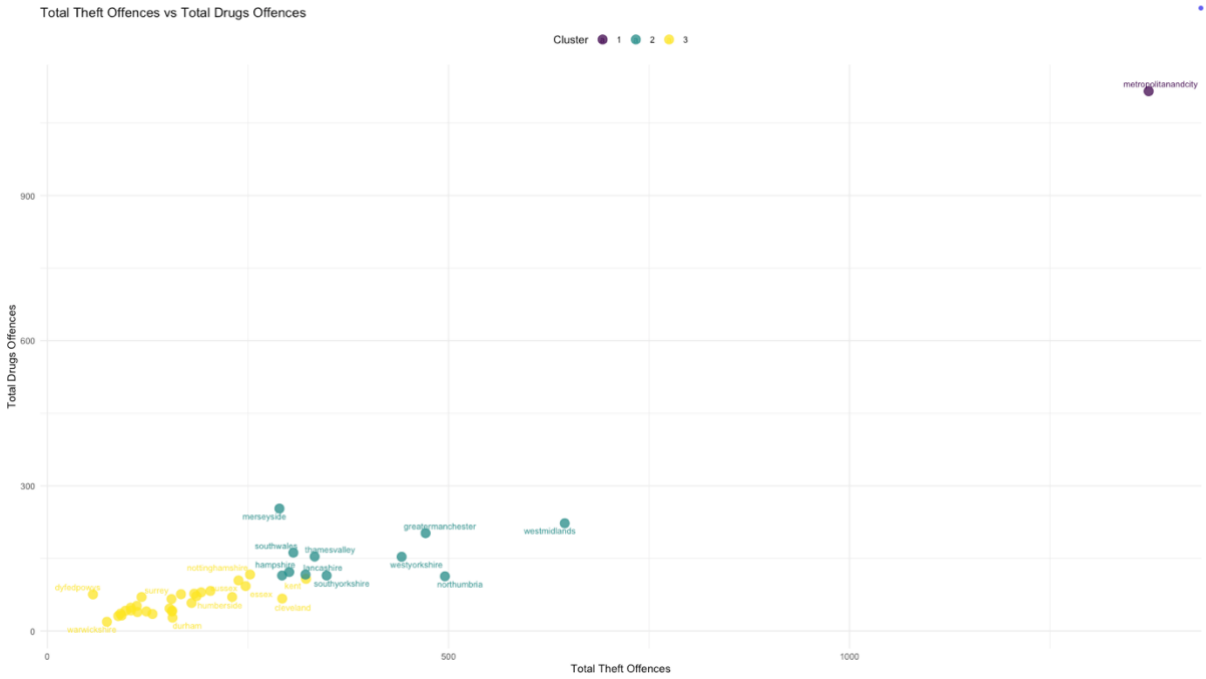

Figure 6.4: Total Theft Offences Against Total Drug Offences (K-means)

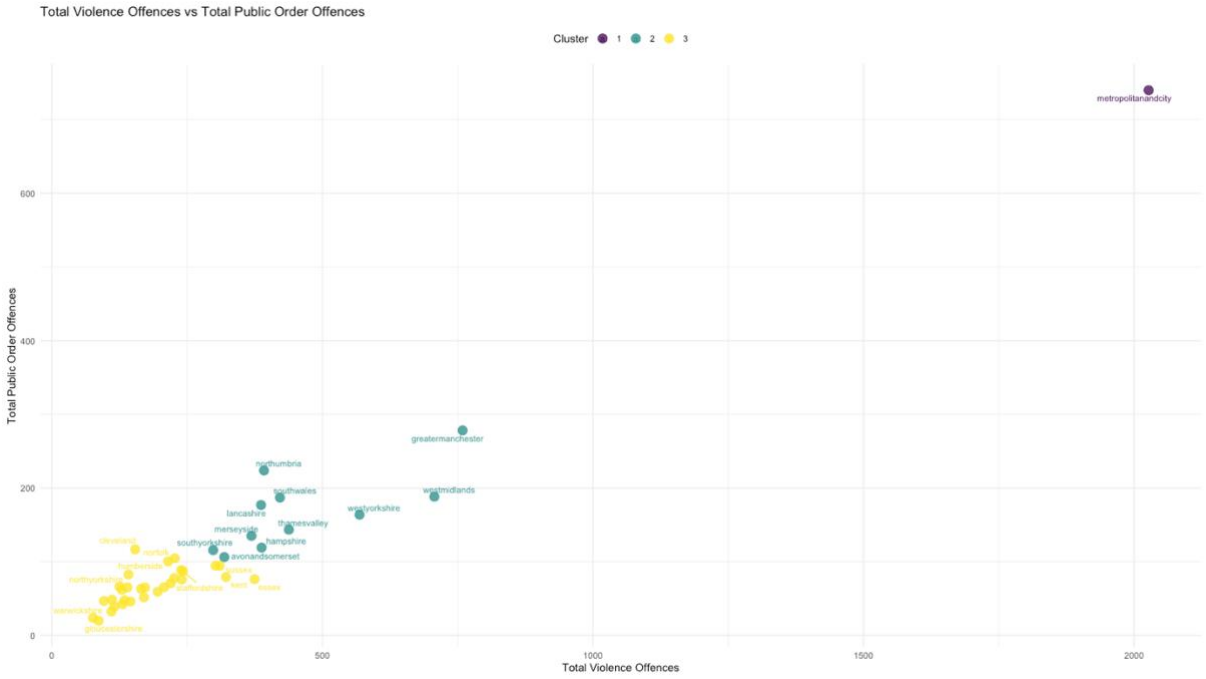

Figure 6.5: Total Violence Offences Against Total Public Order Offences (K-means)

These relationships contributed to the formation of three distinguishable clusters: High-crime urban centres, where offence counts, particularly theft and motoring, were significantly elevated (Cluster 1). Moderate-activity regions, with balanced offence volumes (Cluster 2). Low-volume rural regions are generally characterised by lower offence totals across all categories (Cluster 3). However, upon visual inspection, it became clear that the 'metropolitanandcity' region was a substantial outlier, forming a cluster of its own due to its disproportionately high offence volumes. Its presence skewed the clustering structure and obscured interpretability for the remaining regions. To enhance visual and analytical clarity, 'metropolitanandcity' was excluded from the analysis. After re-running the clustering on the reduced dataset, k = 3 remained the optimal choice, although the silhouette score dropped to 0.4367, reflecting an 11% reduction in overall cluster separation (Figures 6.6 and 6.7). Nevertheless, this score still signified an acceptable structure.

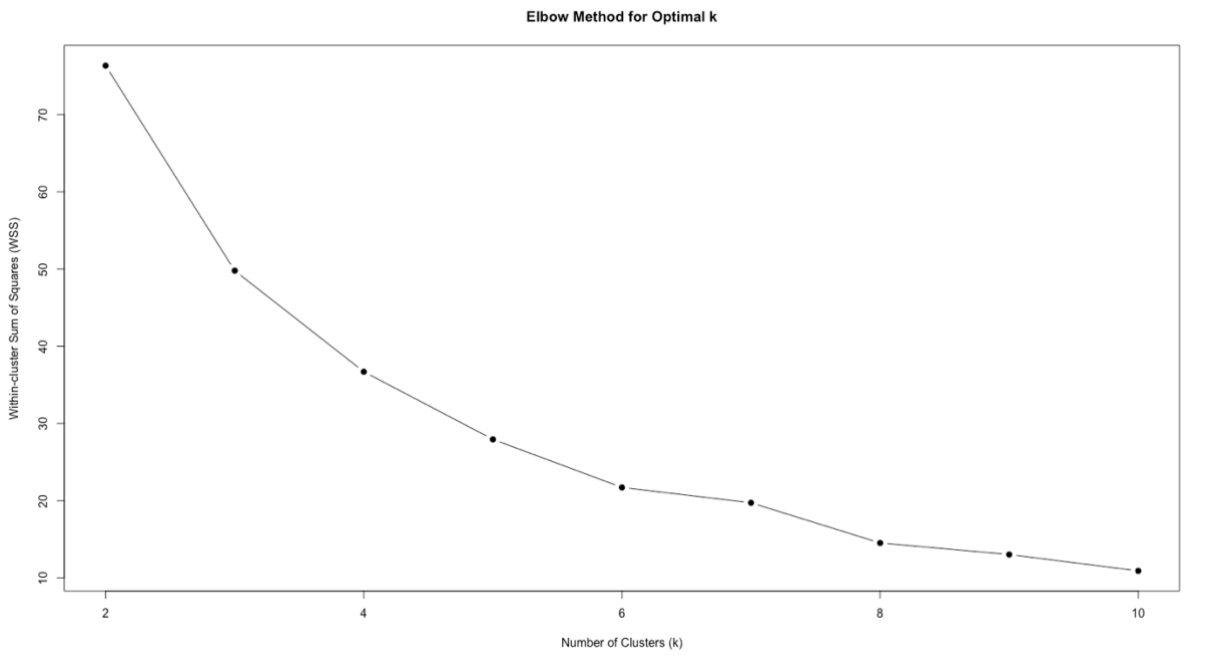

Figure 6.6: Elbow Method for Optimal k (Metropolitan and City Excluded)

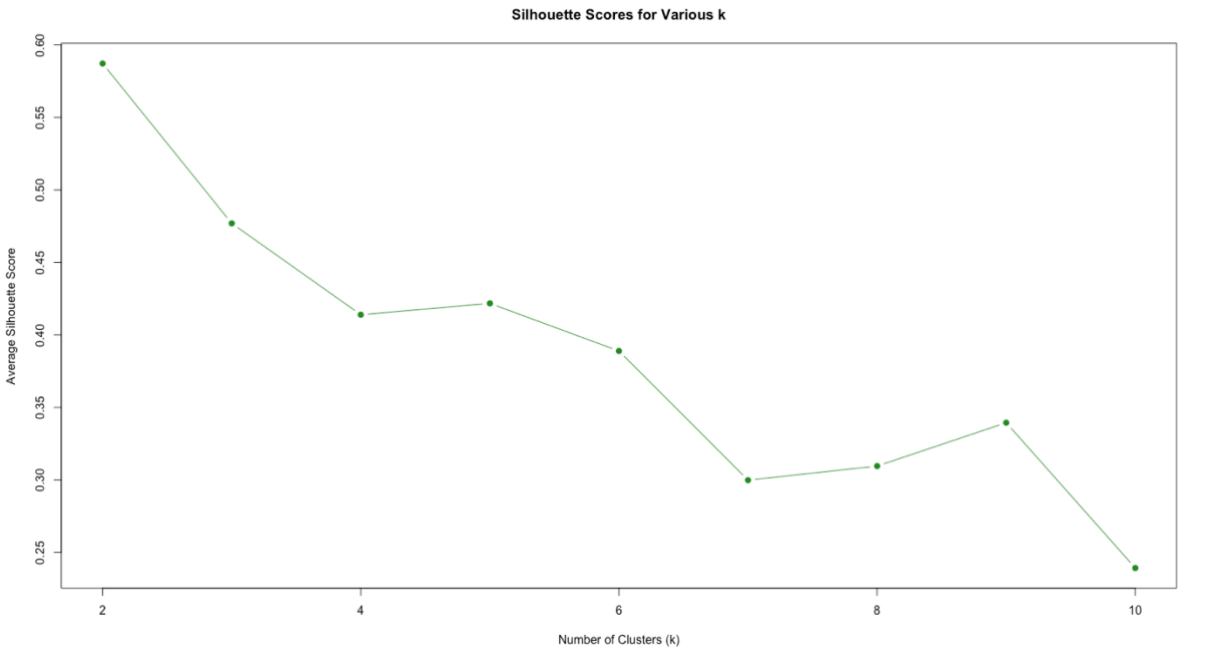

Figure 6.7: Silhouette Scores for Various k (K-means) Metropolitan and City Excluded

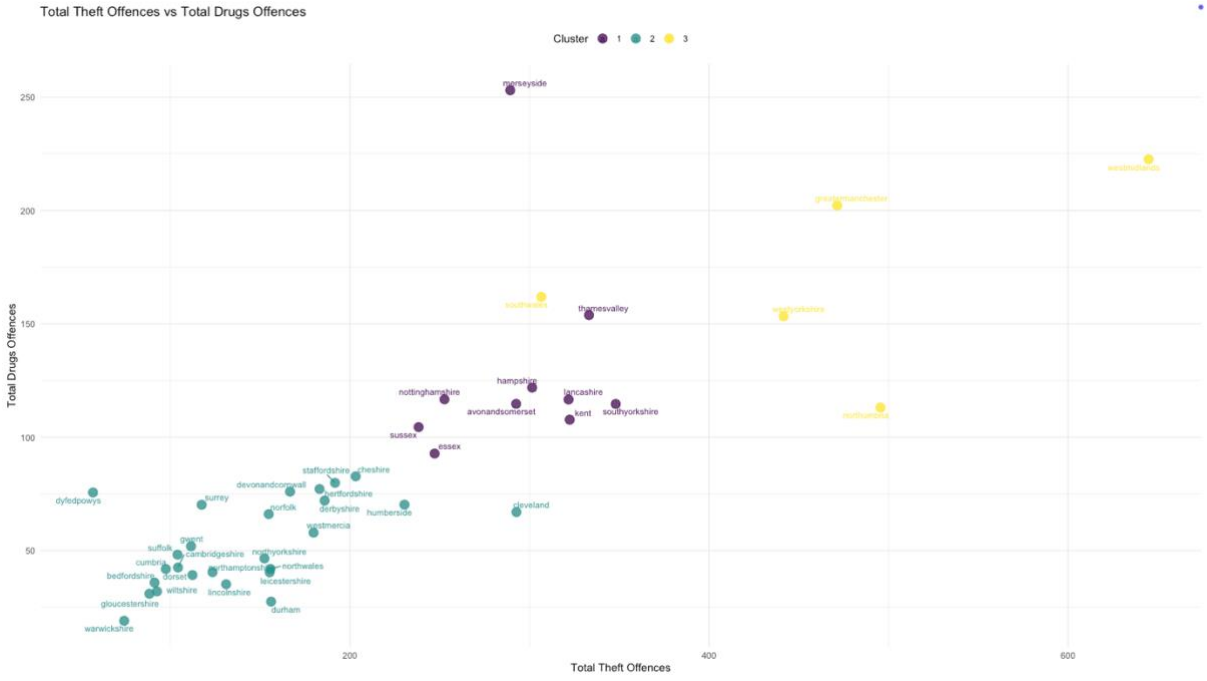

Figure 6.8: Total Theft Offences Against Total Drug Offences (K-means)

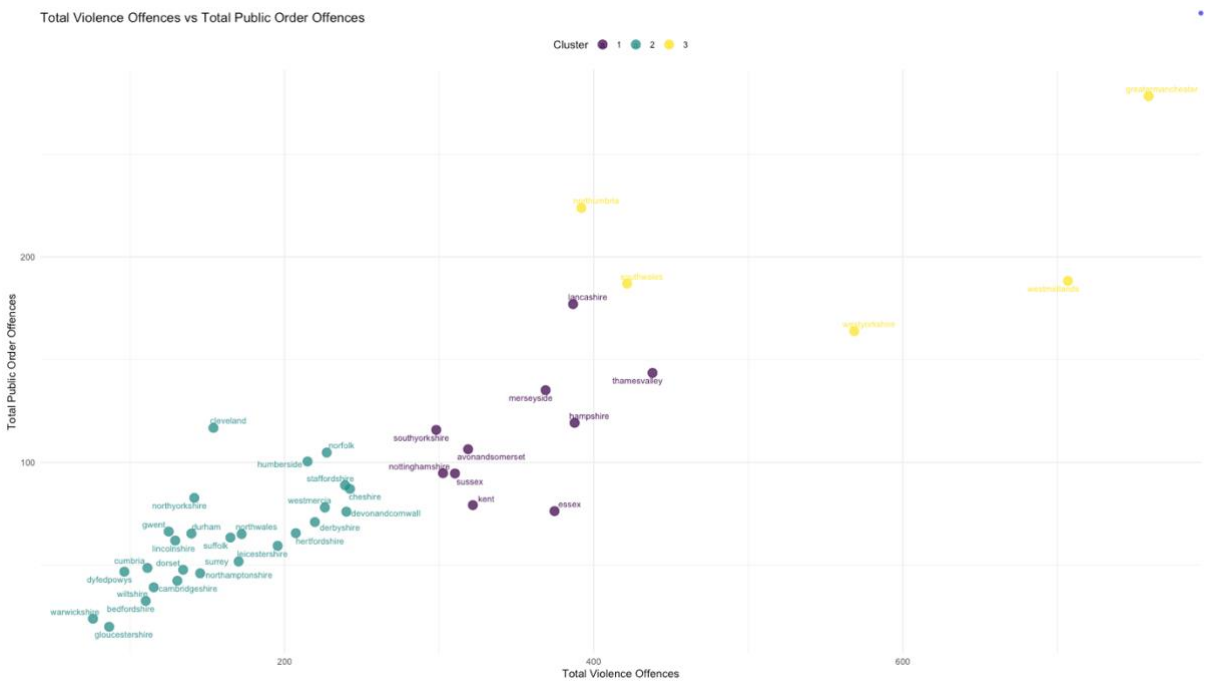

Figure 6.9: Total Violence Offences Against Total Public Order Offences (K-means)

As shown in Figures 6.8 and 6.9, the revised clustering output retained its interpretive value. The regions were grouped into:

Cluster 2: Low-volume rural areas (e.g., Warwickshire, Gloucestershire, Durham)

Cluster 1: Moderate-offence regions (e.g., Kent, Nottinghamshire, Avon and Somerset)

Cluster 3: High-crime urban centres (e.g., Greater Manchester, West Midlands, West Yorkshire)

This outcome reinforces the hypothesis (H_1) that CPS regions can be meaningfully grouped based on offence patterns. The clustering reveals underlying socio-geographic structures in prosecution workload and can support evidence-based planning, such as targeted resourcing, workload balancing, and comparative regional performance analysis.

6.3 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering, a bottom-up unsupervised learning method, was applied to explore potential hierarchical relationships among CPS regions based on offence patterns. This approach iteratively merges regions based on similarity, beginning with each region as an individual cluster and gradually combining them. Ward's linkage criterion was employed, which minimises the total within-cluster variance at each step, making it well-suited for detecting compact, spherical clusters (Murtagh and Legendre, 2011). The resulting dendrogram offered a clear visualisation of the clustering process and revealed meaningful nested structures among regions. A horizontal cut at a dendrogram height of approximately 5.0 yielded an interpretable three-cluster solution, aligning with the refined K-Means outcome (Figure 6.10). With three clusters identified, the silhouette score of approximately 0.44 supported the validity of this clustering structure, indicating moderate cohesion within and separation between regional groups (Figure 6.11).

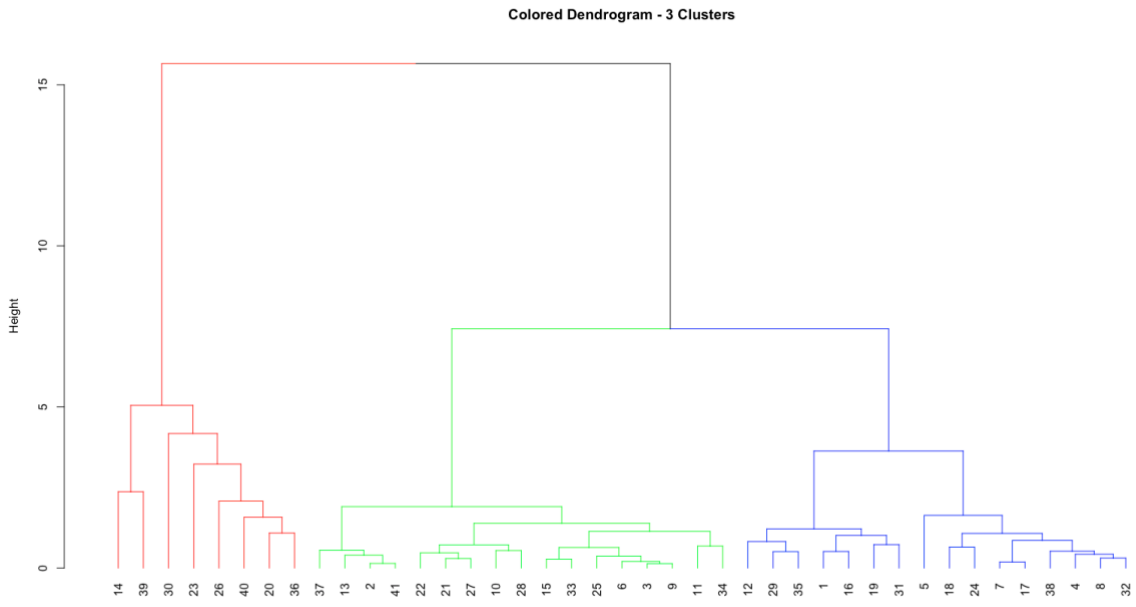

Figure 6.10: Dendrogram for Agglomerative Clustering

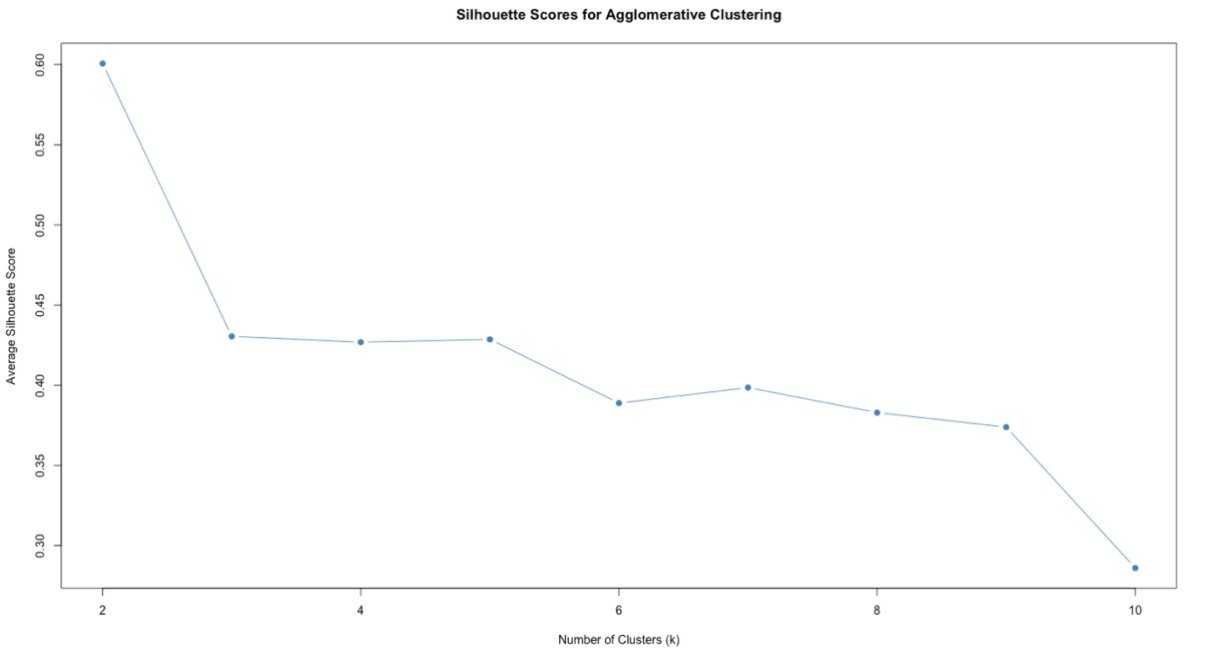

Figure 6.11: Silhouette Scores for Agglomerative Clustering

A significant advantage of hierarchical clustering lies in its ability to reveal graduated and nested relationships (Shetty and Singh, 2021). For example, rural regions such as Staffordshire, Cheshire, and West Mercia were initially clustered together (cluster 2), suggesting high similarity in offence distribution. These regions were first merged with mid-level regions, such as Gwent and North Wales (cluster 1), and were later grouped into larger clusters with regions like Greater Manchester, the West Midlands, and South Wales (cluster 3), indicating a tiered similarity in prosecutorial profiles (Figure 6.12). These clusters were visualised for offence pairs using scatterplots to show their distributions after cutting the dendrogram (Figures 6.13 and 6.14). This layered structure is informative for policymakers and regional planners seeking to understand the gradual transitions in offence dynamics across geographic areas. Overall, the agglomerative approach complemented the K-Means results, enhancing interpretability and providing a richer understanding of regional interdependence and cluster composition in the context of offence trends.

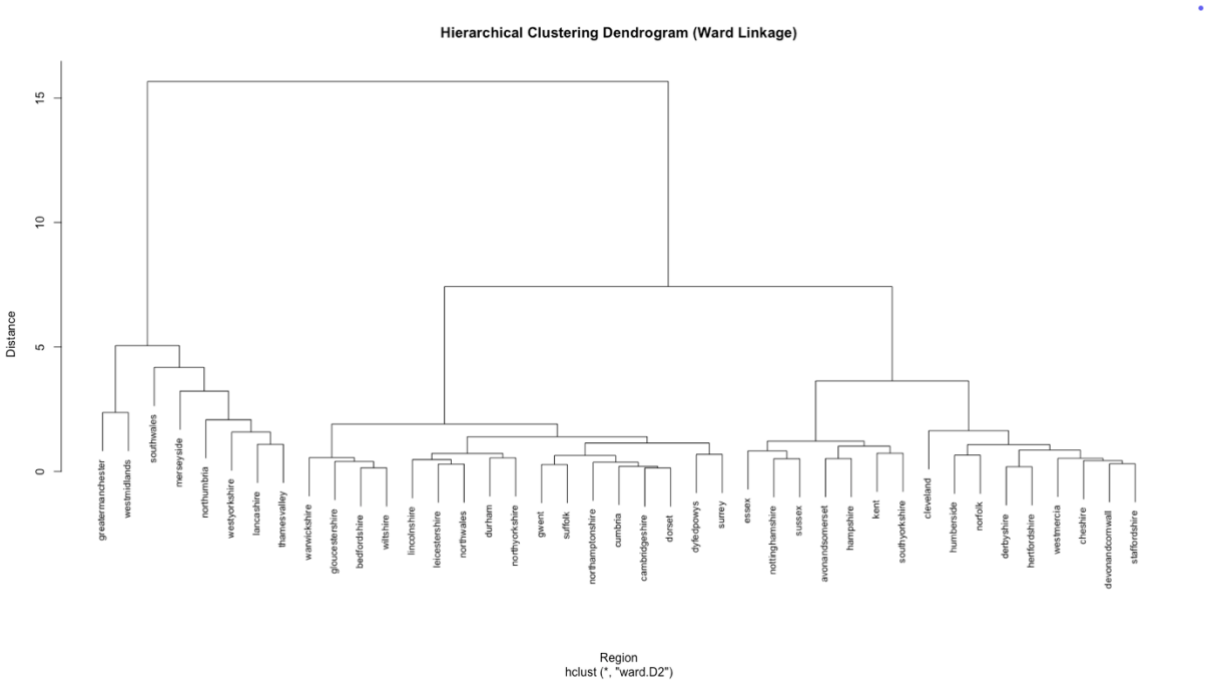

Figure 6.12: Hierarchical Clustering Dendrogram (Ward Linkage)

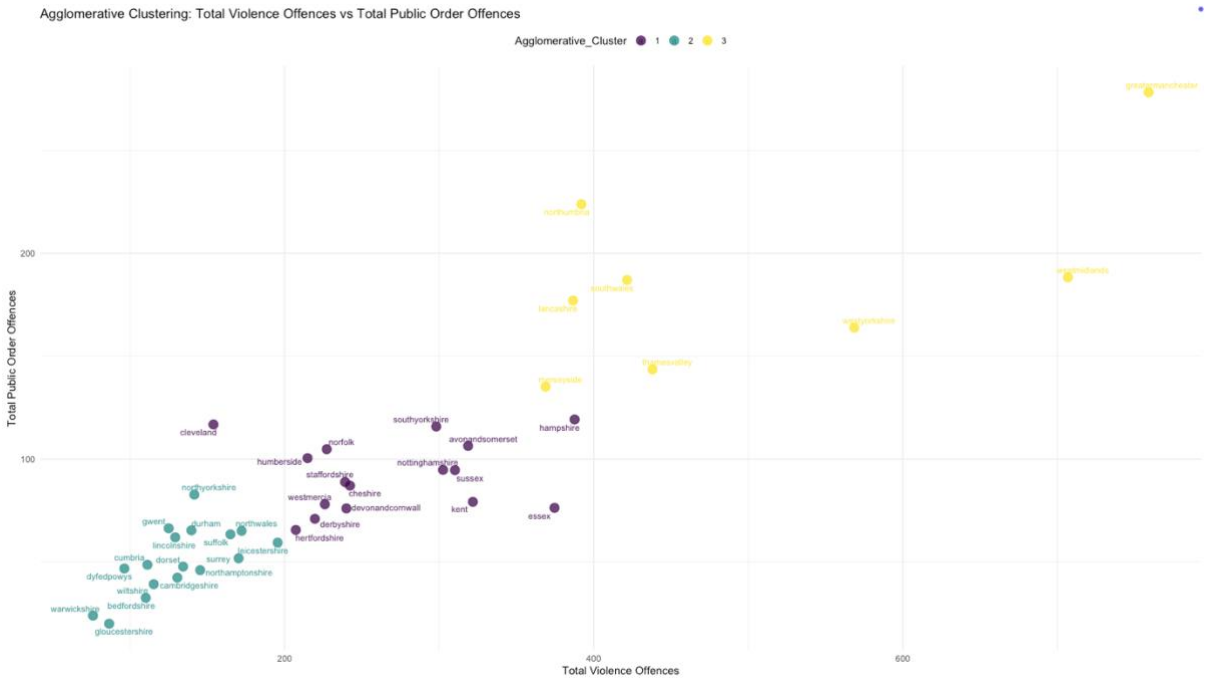

Figure 6.13: Total Violence Offences Against Total Public Order Offences (Agglomerative Clustering)

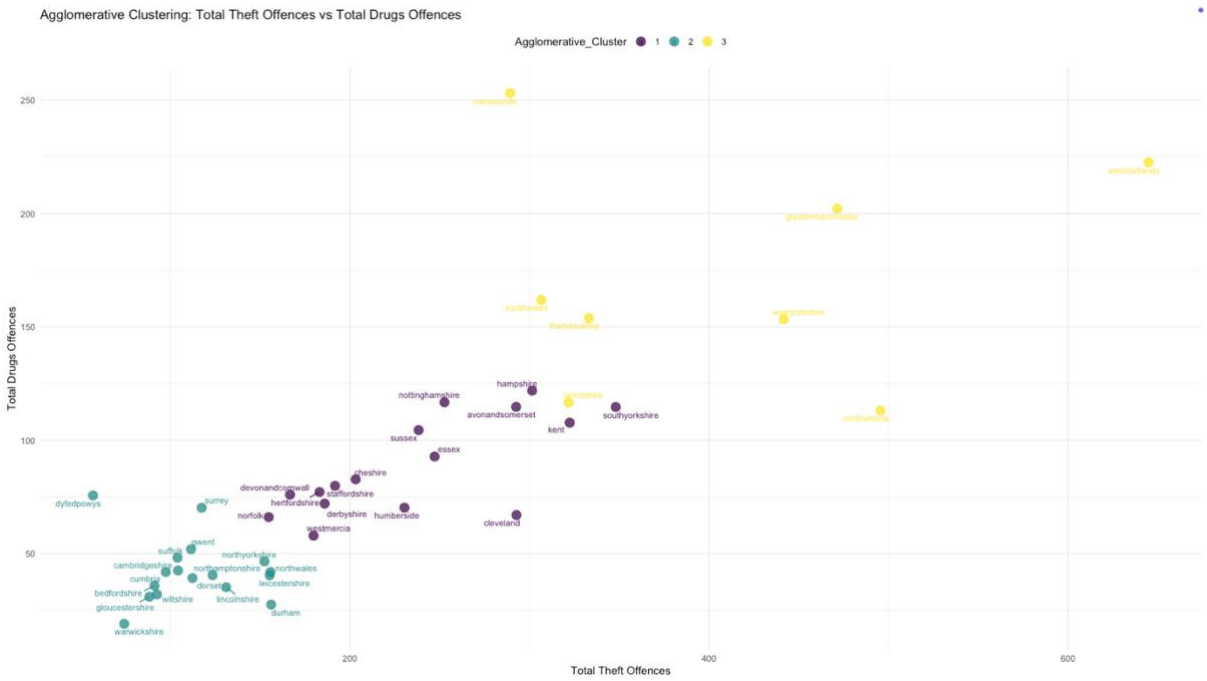

Figure 6.14: Total Theft Offences Against Total Drug Offences (Agglomerative Clustering)

6.4 DBSCAN Clustering: Density-Based Structure and Outlier Detection

To further interrogate the structure of regional offence data and identify potential outliers, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was implemented. Unlike centroid-based methods, such as K-Means, or linkage-based methods like Agglomerative Clustering, DBSCAN does not require the prior specification of the number of clusters. Instead, it groups regions based on local density, enabling the detection of arbitrarily shaped clusters and noisy or anomalous regions (Sánchez-Vinces *et al.,* 2025). DBSCAN identifies core points as those that have a minimum number of neighbours (MinPts) within a specified radius (ε). Following best practice, MinPts was set to 2 × number of features (5 offence types), yielding a value of 10 (Chauhan, 2022). The k-distance plot was used to estimate an appropriate ε, with the optimal value determined to be approximately 1.498 using the second-derivative (elbow) method as shown in Figure 6.15.

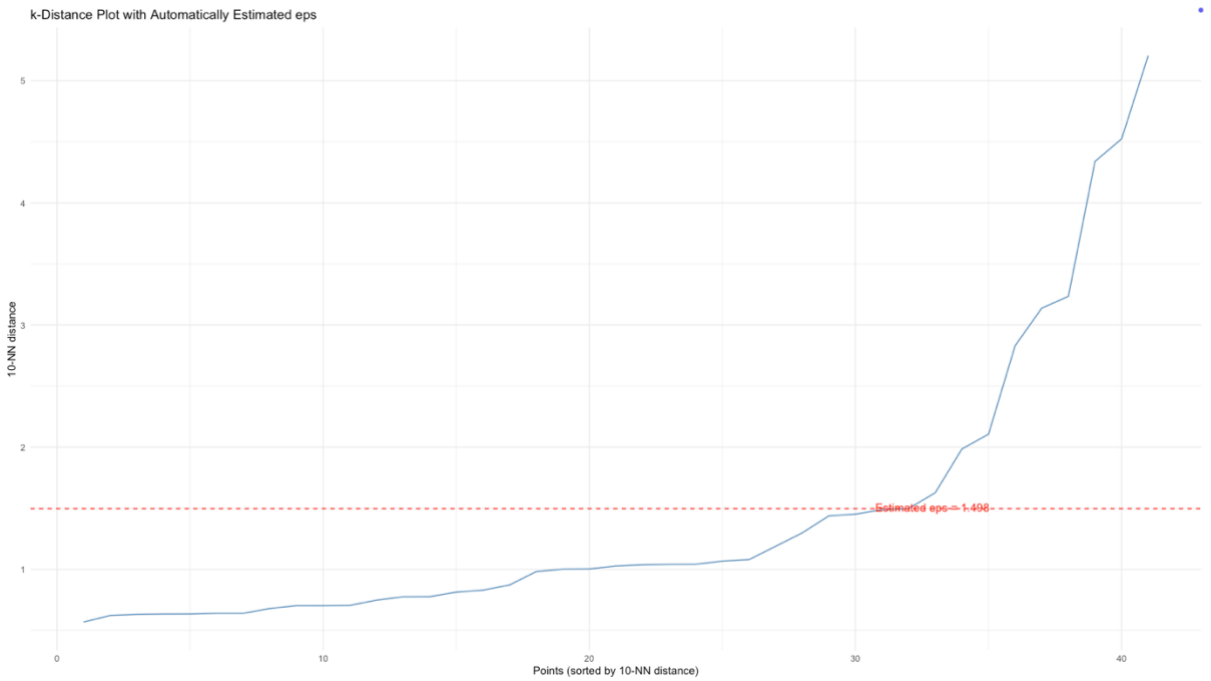

Figure 6.15: k-Distance Plot (DBSCAN)

The DBSCAN algorithm identified only one valid cluster (cluster 0) and flagged six regions as noise (cluster -1), specifically: Greater Manchester, Merseyside, Northumbria, South Wales, West Midlands, and West Yorkshire. Due to DBSCAN detecting fewer than two clusters, the silhouette score could not be computed. This outcome suggests that, based on DBSCAN's density-based criteria, the majority of CPS regions formed a single, broadly homogeneous group. The lack of sufficient density variation prevented the identification of multiple distinct clusters, indicating that regional offence patterns may not exhibit strong localised groupings under DBSCAN's assumptions. The flagged noise points correspond to high-volume or high-variance regions, as shown in Figures 6.16 and 6.17 below. Although DBSCAN did not yield multiple meaningful clusters, it still provided valuable insights. Specifically, the detection of outlier regions provides a data-driven basis for focused scrutiny. These regions may represent areas with unique socio-criminal dynamics, operational constraints, or inconsistencies in case classification, and may therefore require tailored interventions or separate modelling strategies to avoid skewing broader conclusions.

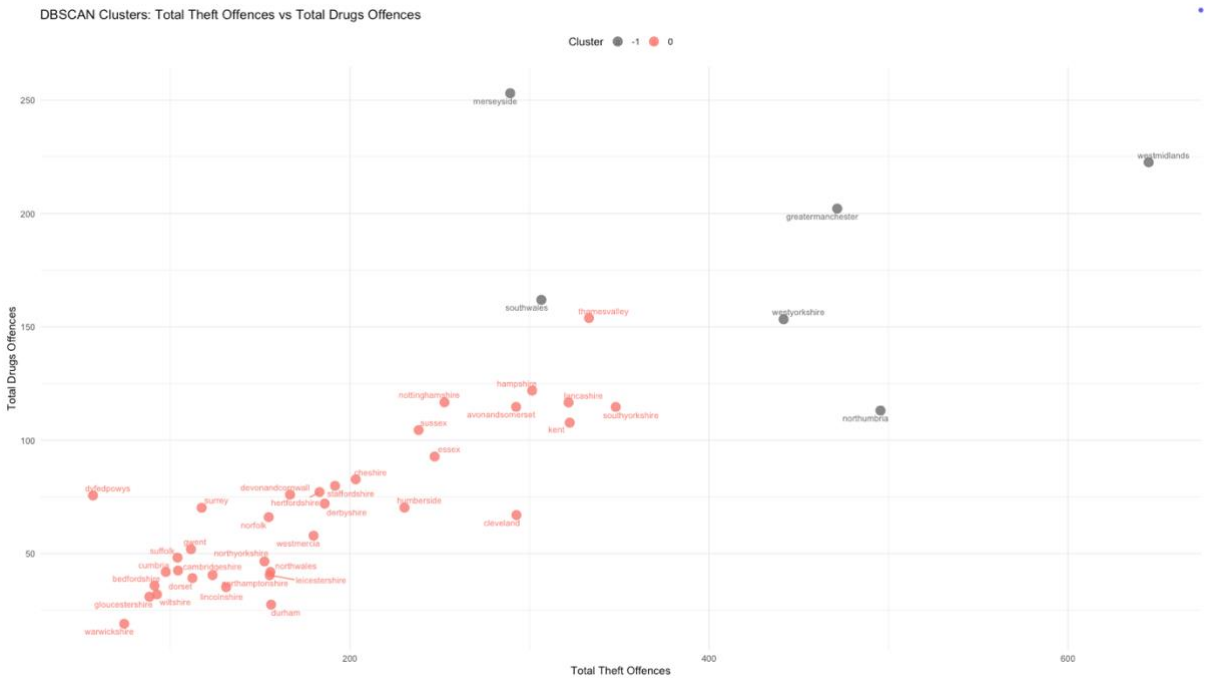

Figure 6.16: Total Theft Offences Against Total Drug Offences (DBSCAN)

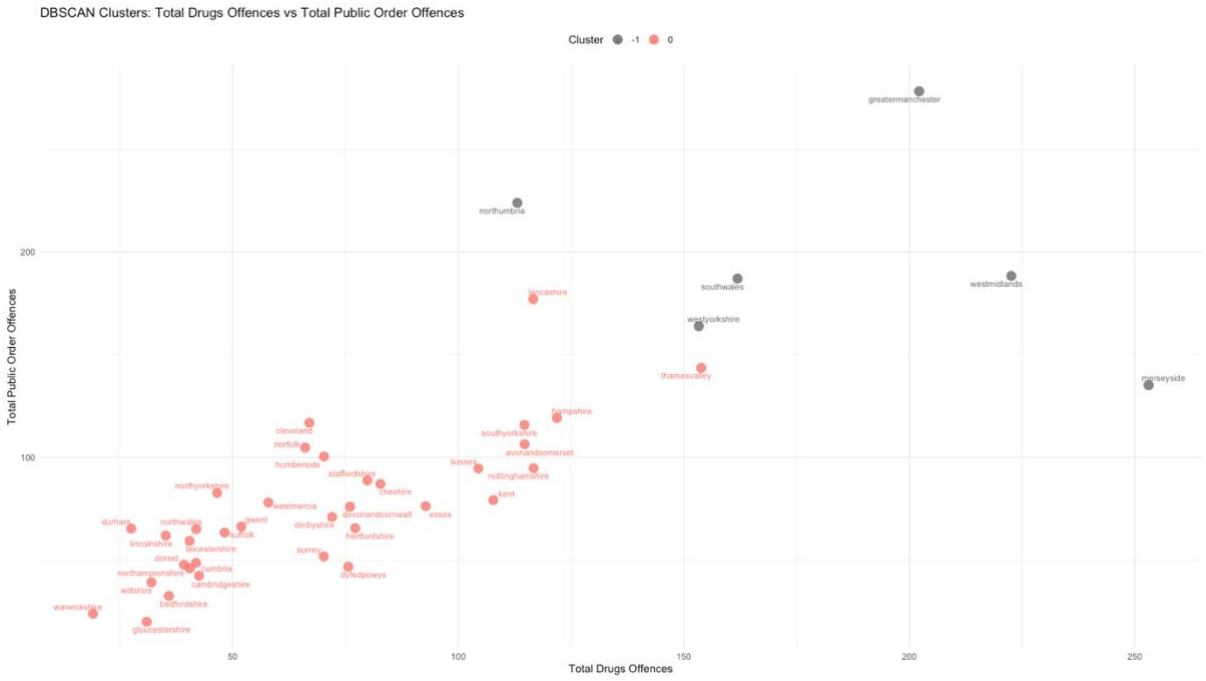

Figure 6.17: Total Drugs Offences Against Total Public Order Offences (DBSCAN)

6.5 Comparison and Interpretation of Clustering Results

In comparison, both K-Means and Agglomerative Hierarchical Clustering produced interpretable and statistically robust groupings of CPS regions, with silhouette scores of 0.5434 and 0.4367, respectively. These scores reflect moderate intra-cluster cohesion and inter-cluster separation, validating the presence of latent regional structure in offence and outcome profiles. K-Means offered clearer segmentations in flat cluster form, ideal for practical policy applications, while Agglomerative Clustering added hierarchical depth, useful for exploring graduated similarities and sub-regional affinities. These findings hold practical significance for stakeholders involved in criminal justice planning and reform. Regions identified within high-offence clusters may warrant increased prosecutorial resources, staffing, or targeted intervention funding to address elevated workloads. Conversely, consistently low-offence regions could be prioritised for workflow audits, digital transformation, or procedural streamlining to enhance operational efficiency. Furthermore, the clustering results provide a foundation for stratified or cluster-aware predictive modelling, ensuring that regional variation is appropriately captured and accounted for in subsequent analytical stages and decision-making processes.

By contrast, DBSCAN was less effective in identifying distinct regional groupings under its density-based assumptions. It identified only one cluster and flagged six regions as noise. Although DBSCAN's silhouette score could not be computed due to the insufficient number of clusters, its utility lay in flagging high-volume or structurally dissimilar regions that may merit separate analysis or modelling treatment. These outlier regions may reflect unique socio-economic contexts, operational pressures, or inconsistencies in data recording. Despite the differences in clustering outcomes, the collective results affirm the alternative hypothesis (H_1)—that CPS regional offence patterns are not random, but instead form structured groupings across the regions.

7.0 Classification

Classification is a supervised machine learning technique used to assign input data to predefined categories based on labelled training examples (Loog, 2017). In this research, classification models were employed to determine whether a regional crime profile would result in a high or low conviction success rate, thereby framing the problem as a binary classification task. Each case profile was labelled as either ‘1’ (high conviction outcome) or ‘0’ (low conviction outcome) based on historical Crown Prosecution Service (CPS) data. The aim was to provide operational insights that enable the CPS to identify underperforming regions and anticipate risks associated with specific offence patterns.

7.1 Hypothesis

H_0 (Null Hypothesis): Crime type distributions, temporal variables (month, year, season), and regional groupings do not significantly influence conviction success.

H_1 (Alternative Hypothesis): Crime type distributions, temporal variables (month, year, season), and regional groupings significantly influence conviction success.

By leveraging these features, the classification phase supports evidence-based decision-making and targeted strategic planning.

7.2 Data Preparation and Feature Engineering

Figure 7.1 below illustrates how the R code was utilised to preprocess the dataset, supporting binary classification by engineering new features and ensuring the data was effective with the models. For each offence type, total case counts were calculated by summing convictions and unsuccessful outcomes, providing a comprehensive representation of caseload distribution. Administrative finalisations were excluded from these totals, as they are not considered part of conviction metrics and could otherwise distort predictive patterns. To define the binary classification target, a new variable, high_conviction, was created. This was based on the median conviction rate across the dataset: regions with a conviction percentage above the median were labelled '1' (high conviction outcome), while those below the median were labelled '0' (low conviction outcome). This approach would also help to ensure the absence of imbalance in the target variable.

Categorical variables—including region group, season, month, and year—were converted into factors. These variables were preprocessed for modelling using one-hot encoding, and zero-variance filtering was applied to remove any non-informative predictors. The preprocessed data were subsequently split into training (80%) and test (20%) subsets, with stratification on the target variable to preserve class distribution and ensure a balanced evaluation across outcome categories.

```r
## 8. CLASSIFICATION
# ==== HYPOTHESIS: Predicting High Conviction Outcomes Using Crime Type and Regional Attributes: A Binary Classific
# Create a copy of the DataFrame
cps_cl <- cps_df

# ==== Define Metric-set and Confusion-Matrix Plot ====
# ==== Metric set ====
metrics_all <- metric_set(
  yardstick::accuracy,
  yardstick::precision,
  yardstick::recall,
  yardstick::f_meas,
  yardstick::roc_auc
)

plot_conf_mat_heatmap <- function(preds, title) {
  cm <- conf_mat(preds, truth = high_conviction, estimate = .pred_class)
  autoplot(cm, type = "heatmap") +
    scale_fill_gradient(low = "#D6EAF8", high = "#2E86C1") +  # Optional: change colors
    labs(
      title = title,
      x = "Actual",
      y = "Predicted",
      fill = "Count"
    ) +
    theme_minimal(base_size = 14) +
    theme(
      plot.title = element_text(face = "bold", hjust = 0.5),
      legend.position = "right"
    )
}

# ==== Prepare Dataset ====
offence_types <- c(
  "homicide", "offences_against_the_person", "sexual_offences",
  "burglary", "robbery", "theft_and_handling", "fraud_and_forgery",
  "criminal_damage", "drugs_offences", "public_order_offences",
  "all_other_offences_excluding_motoring", "motoring_offences"
)

for (offence in offence_types) {
  conviction_col <- paste0("number_of_", offence, "_convictions")
  unsuccessful_col <- paste0("number_of_", offence, "_unsuccessful")
  total_col      <- paste0("total_",         offence, "_cases")
```

Figure 7.1: R code for Classification Analysis

7.3 Models

Three classification algorithms were employed to predict conviction success outcomes, selected for their effective predictive power, interpretability, and computational efficiency:

Random Forest (RF): An ensemble learning method that builds multiple decision trees using randomly selected subsets of data and features. The final prediction is made by aggregating the outputs of individual trees through a majority voting process (Mienye and Sun, 2022). This approach enhances generalisation and mitigates overfitting, making it well-suited for high-dimensional and non-linear classification problems.

Decision Tree (DT): A rule-based model that partitions the dataset into subsets by applying decision rules based on feature values. It is interpretable and helpful in understanding data-driven decision logic. However, without pruning or regularisation, it can overfit to training data (Balcan and Sharma, 2024; Mankar *et al.*, 2024).

Gradient Boosting (GB) – XGBoost Implementation: A sequential ensemble method where each tree is trained to correct the residual errors of its predecessor. The XGBoost implementation is optimised for speed and performance, often achieving high accuracy on structured, tabular datasets. It is particularly effective in capturing complex patterns in the data (Bentéjac, Csörgő and Martínez-Muñoz, 2019).

These models were selected to provide a robust and interpretable comparative framework for evaluating classification performance on regional conviction outcomes.

7.3.1 Evaluation Metrics and Results

To comprehensively evaluate model performance, a range of classification metrics was applied. Battineni, Chintalapudi and Amenta (2020) highlighted that these metrics align with the confusion matrix, which compares predicted outcomes with actual outcomes and consists of the following components:

- **True Positives (TP):** Cases correctly predicted as high conviction outcomes
- **True Negatives (TN):** Cases correctly predicted as low conviction outcomes
- **False Positives (FP):** Cases incorrectly predicted as high conviction outcomes
- **False Negatives (FN):** Cases incorrectly predicted as low conviction outcomes

According to Battineni, Chintalapudi and Amenta (2020), the following metrics were used:

- **Accuracy:** Accuracy = (TP+TN) / (TP+TN+FP+FN)
  This measures the overall correctness of the model by capturing the proportion of total correct predictions.
- **Precision (Positive Predictive Value):** Precision = TP / (TP + FP)
  Precision refers to the accuracy of cases predicted as high-conviction outcomes, which helps reduce the number of false positives.
- **Recall (Sensitivity or True Positive Rate):** Recall = TP / (TP + FN)
  Recall measures the model's ability to correctly identify all actual high-conviction outcomes, thereby reducing the number of false negatives.

- **F1-Score:** This is the harmonic mean of precision and recall, providing a balanced metric especially useful when the class distribution is imbalanced.
- **ROC-AUC (Receiver Operating Characteristic – Area Under the Curve):** This metric evaluates the model's ability to discriminate between the two classes across all threshold values. A higher AUC indicates better performance in distinguishing high vs. low conviction outcomes, regardless of the classification threshold.

These metrics provide a detailed understanding of model behaviour across different dimensions—accuracy, risk of false alerts, and sensitivity to true outcomes—enabling a balanced and rigorous model comparison.

The table below provides a summary of the model results.

Table 7.1: Performance Metrics of Classification Models for Predicting Conviction Success

Metric	Random Forest	Decision Tree	Gradient Boosting
Accuracy	79.2%	76.2%	77.2%
Precision	75.2%	71.9%	74.8%
Recall	87.1%	86.1%	82.2%
F1-Score	80.7%	78.4%	78.3%
ROC-AUC	82.5%	78.1%	81.9%

Among the three models, Random Forest exhibited the highest overall performance, achieving the best accuracy (79.2%), recall (87.1%), and F1-score (80.7%), as well as the highest ROC-AUC score (82.5%). This suggests that the model was especially effective at correctly identifying regions with high conviction outcomes, minimising false negatives—an essential consideration for public interest and judicial accountability. The model's ensemble nature likely contributed to its generalisation strength, as it captures complex feature interactions without overfitting, making it particularly suitable for this high-dimensional dataset.

The Decision Tree model showed the highest recall after Random Forest (86.1%), indicating that it was also sensitive to high-conviction outcomes. However, it had the lowest ROC-AUC (78.1%) and precision (71.9%), indicating a greater tendency toward false positives—predicting a high conviction when it was not actually a high conviction. While its

interpretability makes it useful for rule-based insights, the model's simplicity may limit its ability to capture complex decision boundaries.

Gradient Boosting performed competitively, with an accuracy of 77.2% and ROC-AUC of 81.9%. Although its recall (82.2%) was lower than that of the other models, it had a slightly higher precision (74.8%) than the Decision Tree. This balance reflects its strength in handling nuanced relationships within the data, particularly in cases where marginal values or outliers exist. However, the lower recall compared to Random Forest implies it may miss some high-conviction cases, which could be critical depending on the operational goals of the CPS. Figure 7.2 presents a bar plot that compares these metrics across all models, while Figure 7.3 visualises the comparison of the ROC-AUC curves across all models.

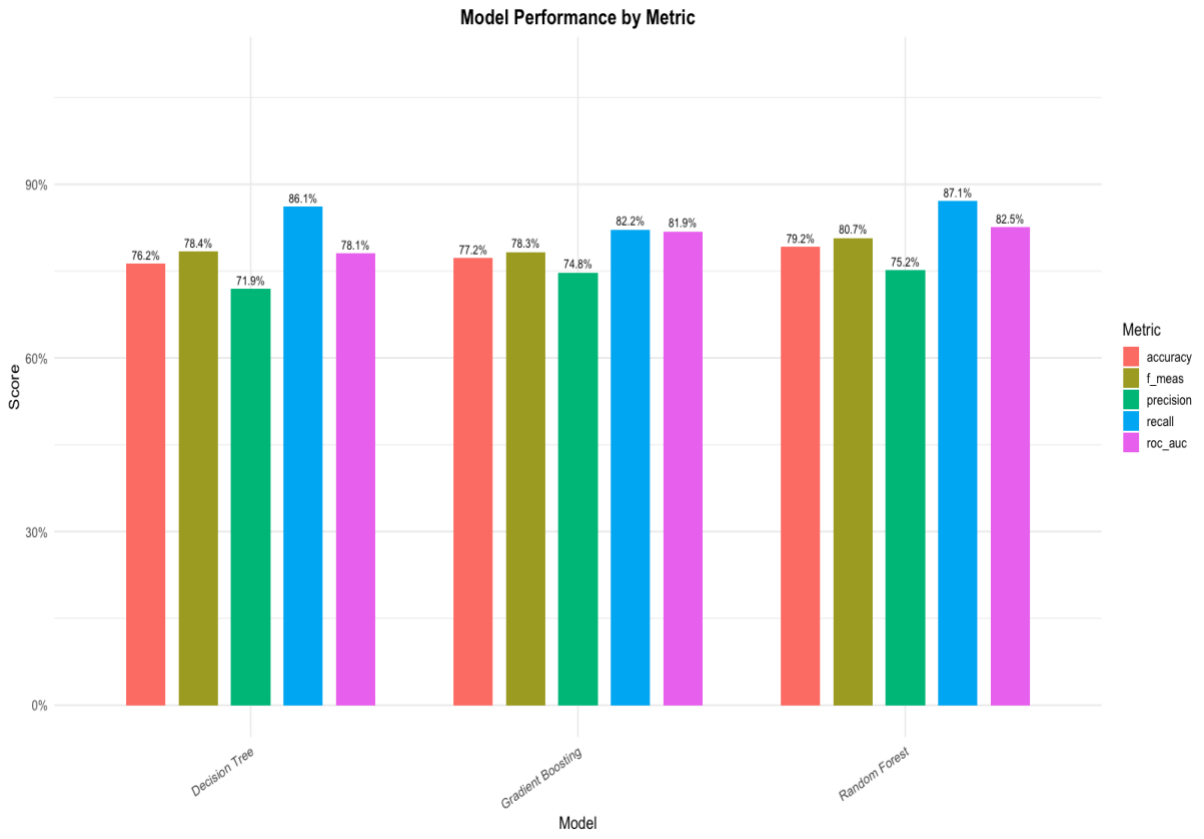

Figure 7.2: Model Performance by Metric

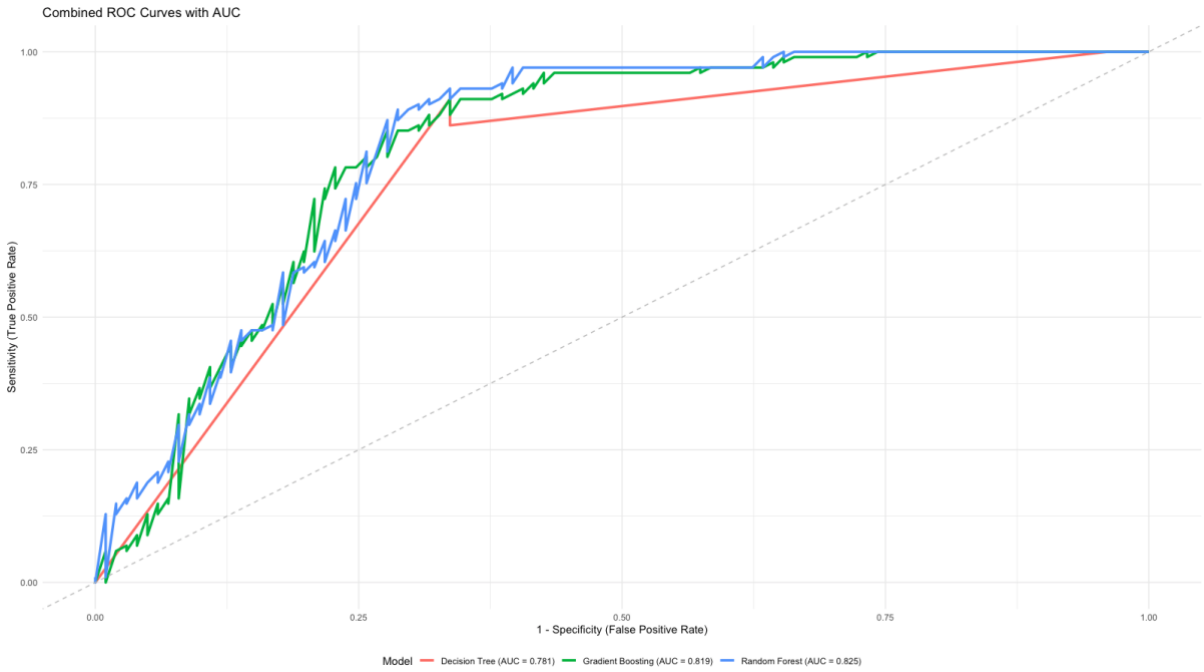

Figure 7.3: Combined ROC Curves with AUC

7.4 Confusion Matrix and Feature Importance

Confusion Matrix

To supplement the quantitative evaluation metrics, confusion matrices were analysed for each model to assess the distribution of true and false predictions across the binary outcome classes—high conviction (1) and low conviction (0). This further inspection reveals strengths and weaknesses in model behaviour that aggregate metrics alone may obscure (Battineni, Chintalapudi and Amenta, 2020).

The Random Forest model demonstrated a strong ability to accurately classify both classes, yielding 88 true positives and 72 true negatives. However, it misclassified 29 high conviction outcomes as low (false negatives) and 13 low outcomes as high (false positives). The low number of false positives supports its high precision, while an overall recall of 87.1% balanced its moderate false negatives (Figure 7.4).

The Decision Tree model correctly predicted 87 high-conviction outcomes but had the highest number of false negatives (34) among the three models. This suggests that while it maintained a high recall (86.1%), it occasionally failed to generalise from the training data, a standard limitation of unpruned decision trees. The slightly higher number of false positives (14) also reflects its lower precision (71.9%) (Figure 7.5).

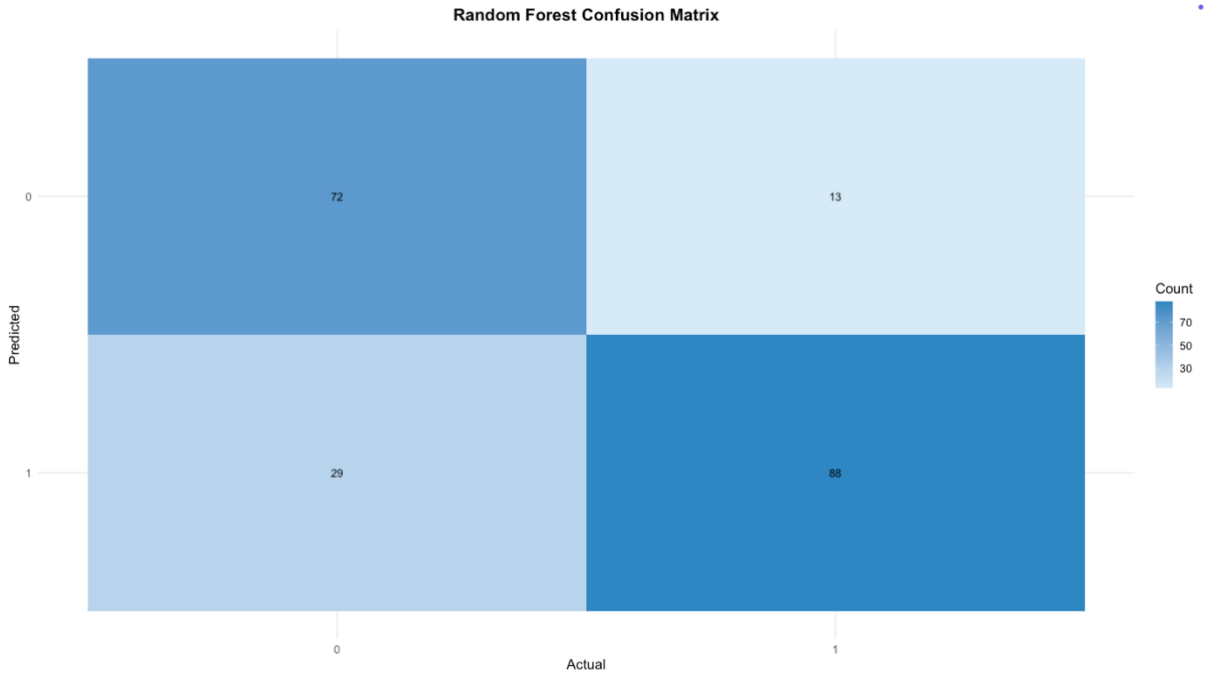

Figure 7.4: Random Forest Confusion Matrix

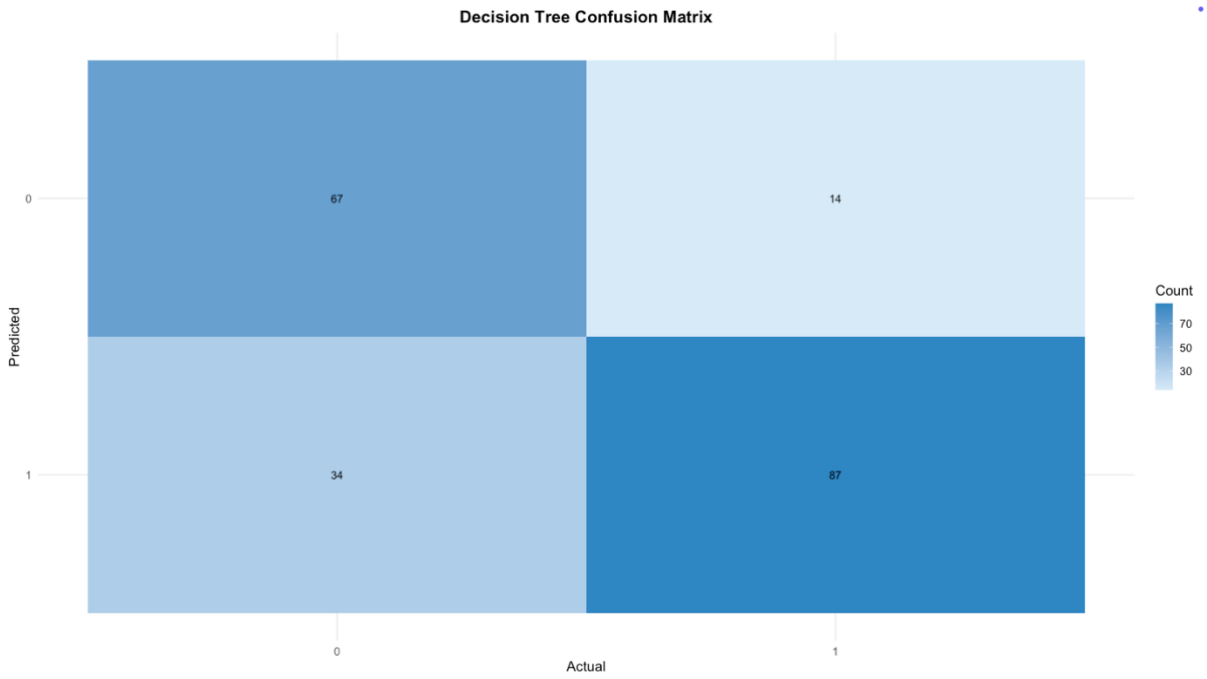

Figure 7.5: Decision Tree Confusion Matrix

The Gradient Boosting model demonstrated balanced performance, correctly identifying 83 high-conviction outcomes and 73 low-conviction cases. However, it produced the highest number of false positives (18), which likely contributed to its slightly reduced precision (74.8%). Also, the number of false negatives (28) was lower than that of the Decision Tree, explaining its higher F1-score (Figure 7.6).

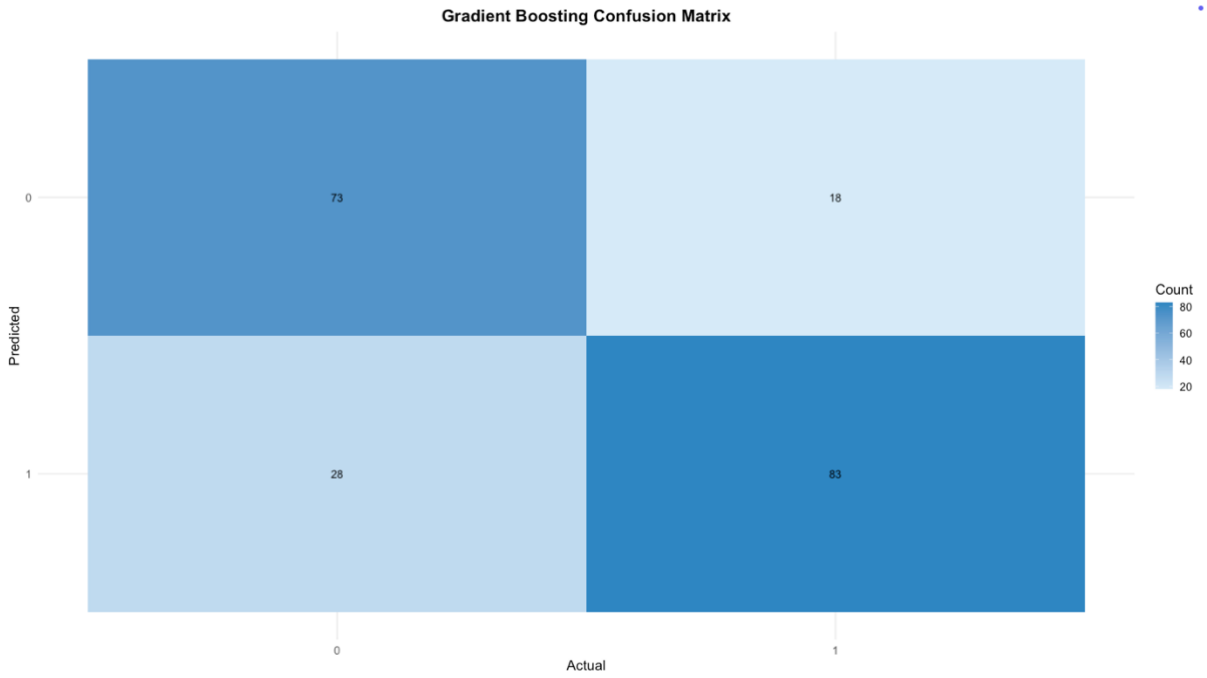

Figure 7.6: Gradient Boosting Confusion Matrix

From an operational perspective within the CPS, false negatives—where high-conviction regions are misclassified as low-conviction—can lead to an underestimation of performance, resulting in unwarranted scrutiny or the misallocation of resources. Conversely, false positives—where low-conviction regions are misclassified as high—may mask underperformance, reducing accountability and delaying necessary interventions. Among the models tested, performance metrics highlight the Random Forest model as the most effective, balancing sensitivity to high-performing regions (recall) with a lower incidence of false positives (precision). This balance makes it particularly well-suited for deployment in high-stakes contexts such as regional performance monitoring or case triaging. The model's strong performance supports fairer, more accurate, and data-driven decision-making, enabling the CPS to allocate resources more effectively and avoid reinforcing systemic biases through misclassification.

Feature Importance

To interpret the internal decision logic of each model and enhance transparency, feature importance rankings were extracted from the trained classifiers. This analysis offers critical insight into which offence categories most strongly influenced predictions of high or low conviction outcomes across regional profiles.

Random Forest

The top five predictors in the Random Forest model were: homicide cases, offences against the person, burglary cases, motoring offences, and theft and handling (Figure 7.7). This suggests that serious violent offences (e.g., homicide, offences against the person) and property-related crimes(e.g., burglary, theft) were the most influential in shaping the conviction outcome classification. The inclusion of motoring offences—typically considered less severe—may indicate high variance in conviction rates across regions, making it a strong signal for class separation. The diversity of these top predictors reinforces Random Forest's strength in capturing both high-impact and high-frequency offence types, contributing to its effective overall performance.

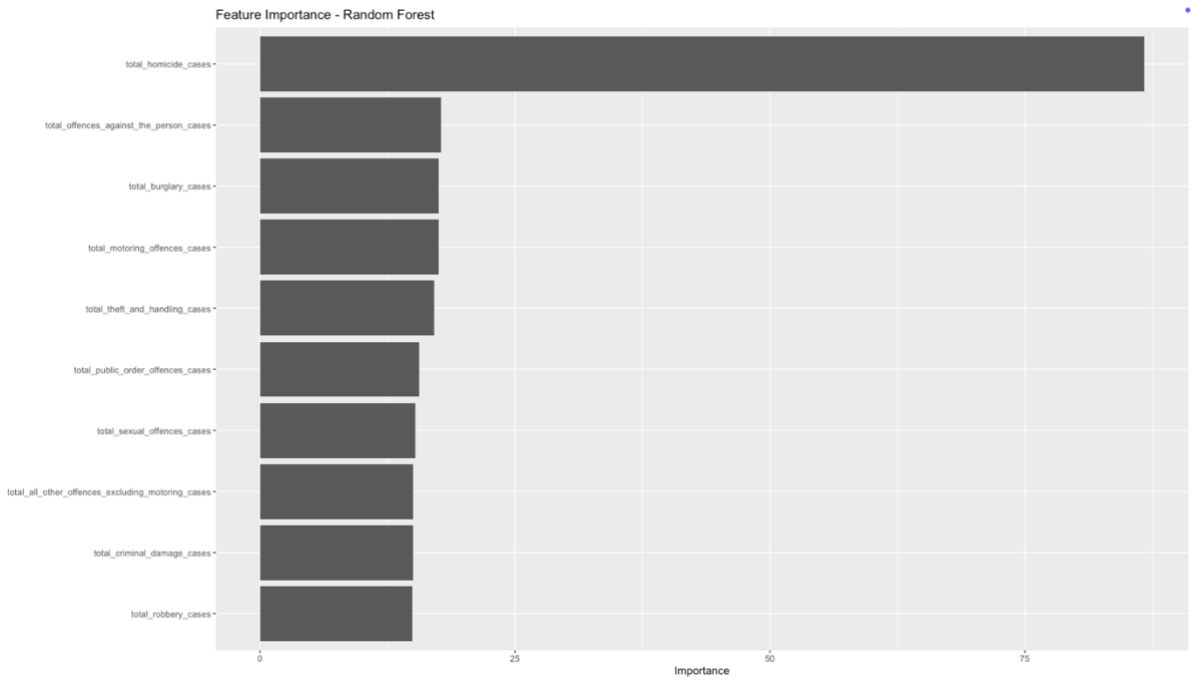

Figure 7.7: Feature Importance – Random Forest

Decision Tree

For the Decision Tree model, the top five features were: homicide cases, theft and handling offences, offences against the person, criminal damage, and burglary cases (Figure 7.8). This model placed more emphasis on homicide cases and theft and handling, possibly due to their prevalence in the dataset and their association with variable regional conviction rates. While Decision Trees offer transparent logic, they are also susceptible to overfitting to dominant or highly variable features, which may explain the slight reduction in generalisation performance compared to Random Forest.

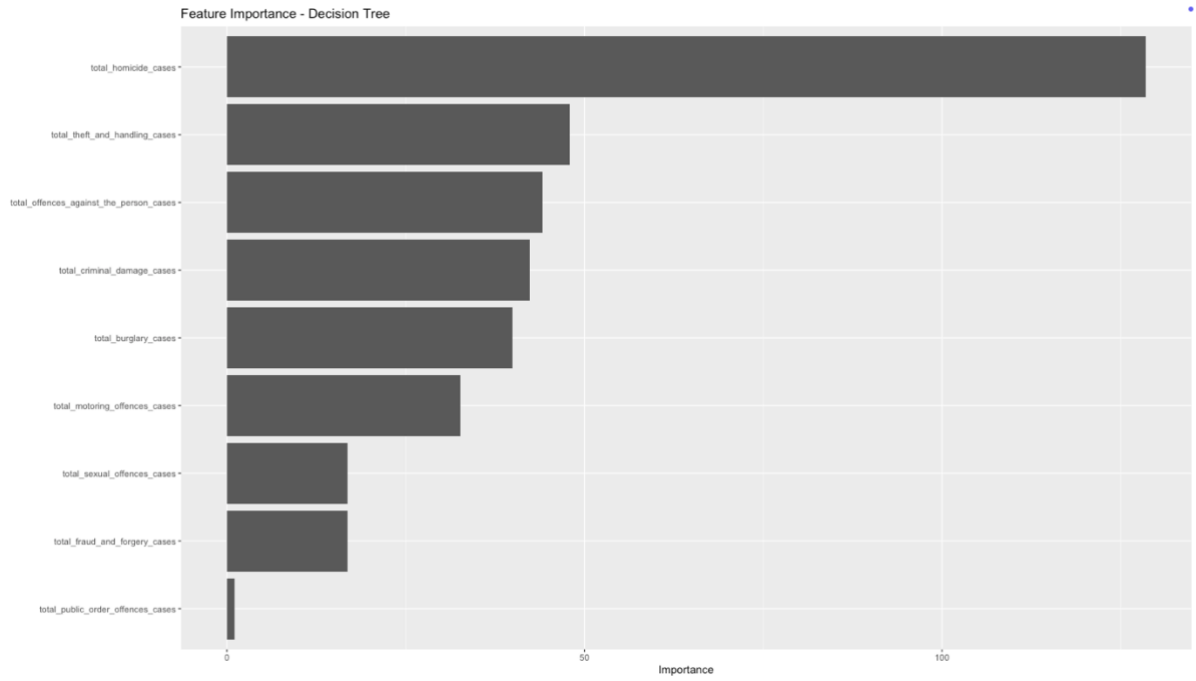

Figure 7.8: Feature Importance – Decision Tree

Gradient Boosting

The Gradient Boosting model identified the following as its most important features: homicide cases, burglary cases, motoring offences, robbery, fraud and forgery (Figure 7.9). This ranking indicates that violent and acquisitive crimes played a central role in the model's prediction logic. The inclusion of robbery, fraud and forgery highlights the model's sensitivity to offences that may involve complex investigations or evidentiary challenges, affecting conviction outcomes across regions. Gradient Boosting's ability to prioritise nuanced feature interactions may account for its competitive, though slightly less stable, performance.

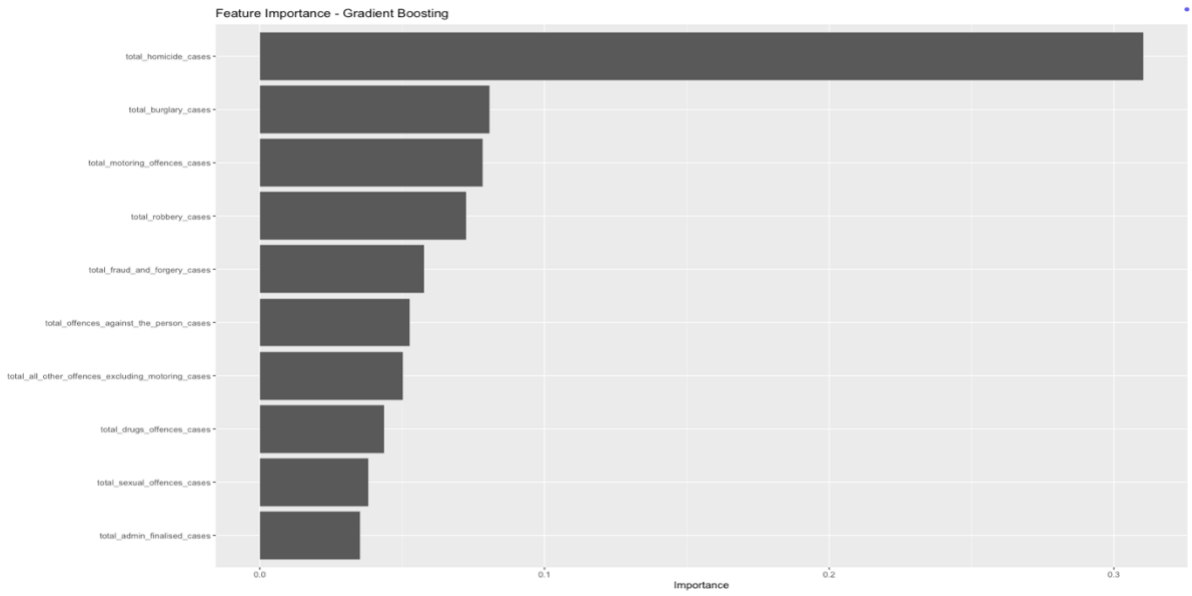

Figure 7.9: Feature Importance – Gradient Boosting

7.5 Interpretation

This classification section demonstrated that offence composition, temporal trends, and regional context are significant predictors of prosecutorial success. Across all models, Random Forest emerged as the most effective classifier, offering the best balance between precision and recall. Its superior performance and interpretability make it suitable for real-world deployment, where it could assist CPS analysts in forecasting conviction outcomes, identifying regional underperformance, and guiding strategic planning. By highlighting offence categories such as homicide, burglary, and offences against the person as key predictors, the model offers actionable intelligence. This can inform operational improvements, resource reallocation, and the prioritisation of high-risk or complex cases.

The consistently high recall values, particularly in Random Forest, support rejection of the null hypothesis (H_0), providing empirical evidence for the alternative (H_1): that regional groupings, offence types, and time-based factors significantly influence conviction success. From a policy and practice standpoint, these insights can help the CPS identify systemic disparities and optimise workflows. For example, frequent false positives in specific regions may signal a need for procedural review or training, while offences with persistently low conviction success could be subject to evidentiary or legislative reassessment. Understanding which features drive predictive decisions also enables the CPS to develop data-informed performance dashboards, monitor conviction trends over time, and ensure equitable outcomes across jurisdictions.

8.0 Classification Model Optimisation

To enhance predictive performance and reduce model-specific bias, hyperparameter optimisation was conducted **for** three classification algorithms: Random Forest (RF), Decision Tree (DT), and Gradient Boosting (GB) (Hamed, Sadek and El-Hafeez, 2023). The primary objective was to enhance model generalisation on unseen data and ensure a robust, balanced evaluation across multiple metrics, particularly in the context of predicting regional conviction success.

8.1 Multi-Metric Evaluation Framework

A comprehensive metric set was defined using the yardstick package to evaluate models across **ROC-AUC, accuracy, precision, recall, and F1-score**. These metrics reflect distinct performance dimensions and are critical for ensuring the reliability of classification results. Recall and F1-score were of particular importance, given their sensitivity to false negatives and balanced consideration of predictive errors, especially relevant in legal and policy contexts where failing to recognise high-conviction patterns can have real-world implications.

8.2 Model Configuration and Grid Search

Three models were tuned using the tune() function with specific hyperparameters:

- **Random Forest**: mtry, min_n, trees
- **Decision Tree**: tree_depth, min_n, cost_complexity
- **Gradient Boosting (XGBoost)**: trees, learn_rate, tree_depth

Parameter grids were created using grid_regular(), and 5-fold stratified cross-validation was applied to preserve class distribution and ensure reliable selection.

8.3 Performance Comparison: Baseline vs. Optimised Models

Following grid search, the best-performing hyperparameters were used to retrain each model on the training data. The final optimised models were evaluated on the test set, and their performance was compared to the corresponding baseline models.

Table 8.1: ROC-AUC Scores – Baseline vs. Optimised

Model	Baseline ROC-AUC	Optimised ROC-AUC
Random Forest	0.8254	0.8261
Gradient Boosting	0.8187	0.7995
Decision Tree	0.7806	0.7673

The Random Forest model achieved the highest ROC-AUC, with a marginal improvement following optimisation. However, Gradient Boosting and Decision Tree models experienced a decline in ROC-AUC, potentially due to overfitting during the tuning phase or suboptimal parameter boundaries. This outcome underscores the importance of validating hyperparameter tuning results on independent test sets.

8.4 Accuracy Comparison

Table 8.2: Accuracy – Baseline vs. Optimised

Model	Baseline Accuracy	Optimised Accuracy
Random Forest	79.2%	80.7%
Gradient Boosting	77.2%	78.2%
Decision Tree	76.2%	76.7%

Random Forest improved by 1.5 per cent, reinforcing its suitability for this classification task. The Decision Tree showed minimal improvement, while Gradient Boosting gained a modest 1% increase in accuracy, despite a drop in ROC-AUC, suggesting a trade-off between threshold-based and overall discrimination performance.

8.5 Confusion Matrix Comparison

As shown in Figures 8.1, 8.2 and 8.3, the optimised Random Forest reduced false negatives and improved true positives, contributing to stronger recall and F1-score. Decision Tree optimisation decreased false positives but had a marginal impact on precision and F1 score, indicating over-sensitivity to high-conviction outcomes. Gradient Boosting showed slightly more false positives after optimisation, suggesting an increased risk of false alarms, which may not be desirable in operational environments requiring high certainty.

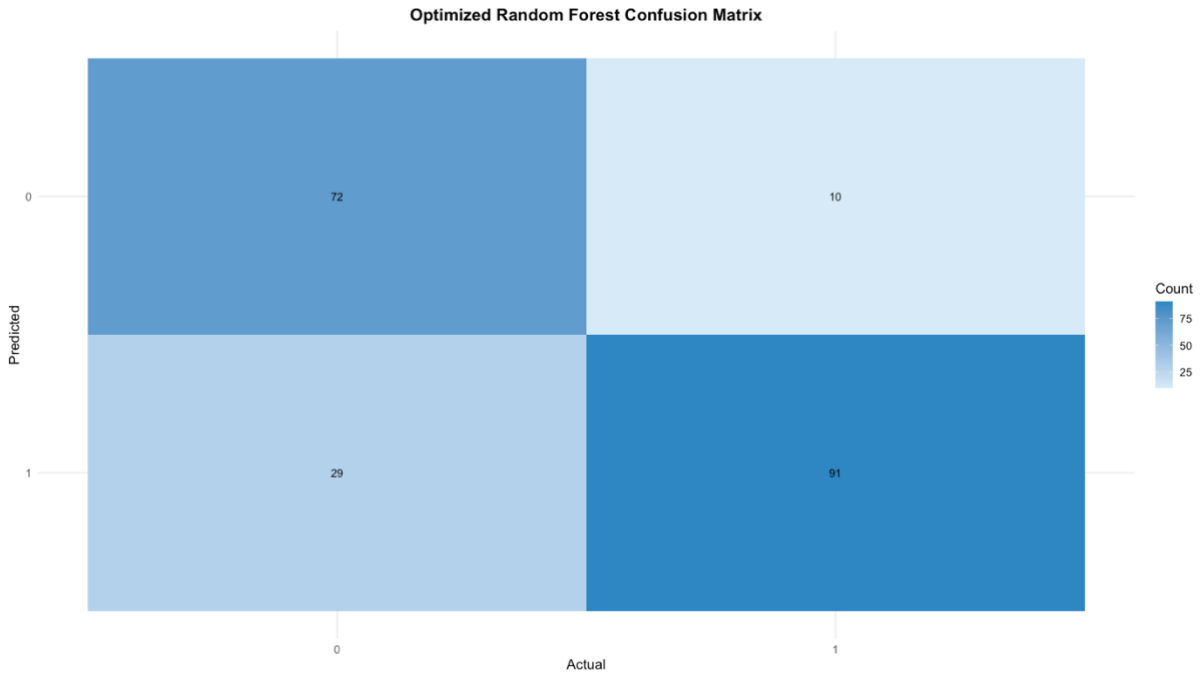

Figure 8.1: Optimised Random Forest Confusion Matrix

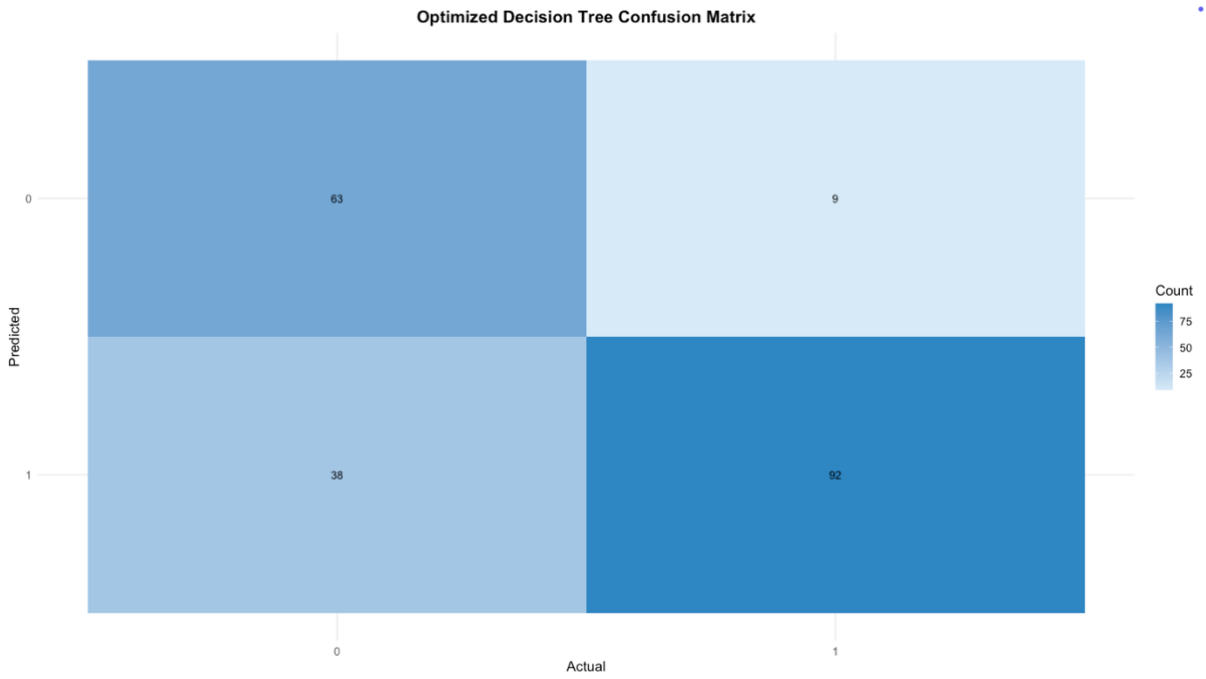

Figure 8.2: Optimised Decision Tree Confusion Matrix

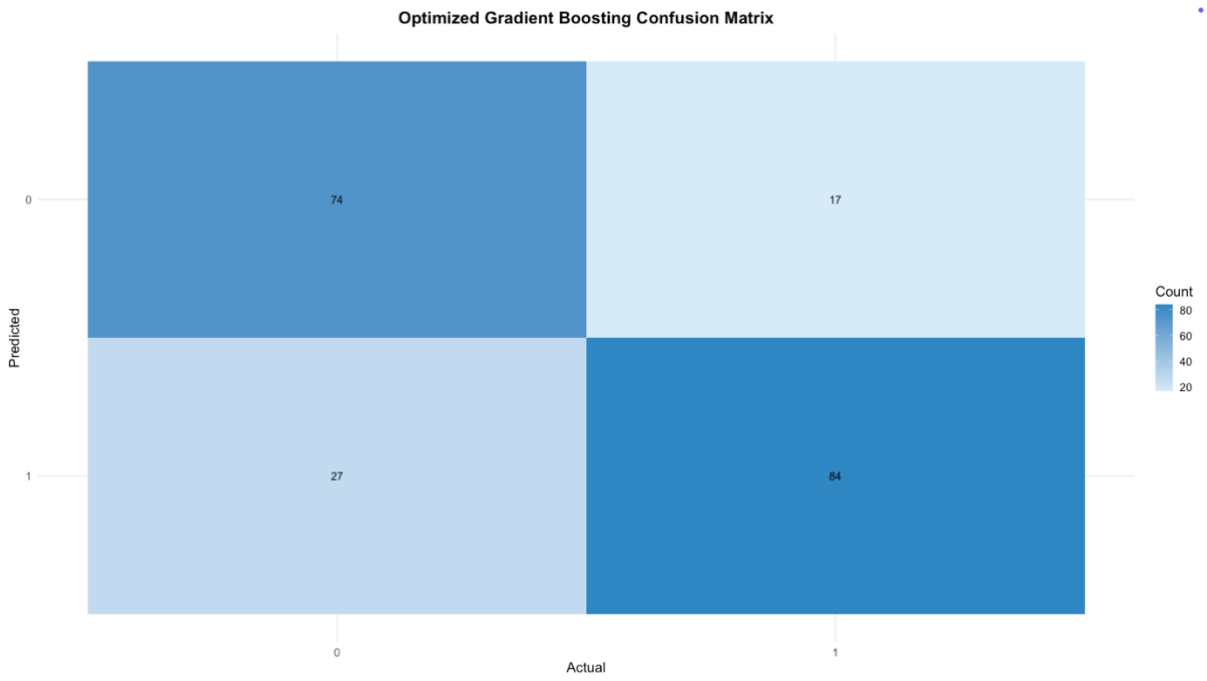

Figure 8.3: Optimised Gradient Boosting Confusion Matrix

8.6 Summary

Among the models, Random Forest demonstrated the best overall trade-off in terms of accuracy, recall, precision, and AUC, making it the most suitable model for predicting conviction success based on regional crime patterns. Its robust recall and F1-score are significant for the CPS in minimising missed high-conviction cases, which could inform resource allocation, training needs, or case prioritisation. The findings also highlight the critical role of validation and metric diversity in model selection.

9.0 Further Evaluation of Analytical Methods, Tools, and Techniques

This section further evaluates the analytical techniques, statistical tests, machine learning models, and data visualisation tools employed throughout this analysis. Each method is examined in terms of appropriateness, effectiveness, and limitations, with comparisons to alternative approaches.

9.1 Statistical Hypothesis Testing

Tools Employed:

- rstatix, FSA, and e1071 in R for ANOVA, Kruskal–Wallis, Dunn's test, and skewness calculations.

Effectiveness:

- Enabled robust hypothesis testing of disparities in prosecution outcomes across offence types (Ogle, 2025).
- Integrated parametric and non-parametric analysis to enhance the reliability of results in non-normal distributions.
- Effect sizes and adjusted p-values facilitated nuanced interpretation, promoting transparency and fairness (Erceg-Hurn and Mirosevich, 2008).

Alternatives:

- Python equivalents (e.g., SciPy, statsmodels) offer similar tests but lack seamless integration with machine learning (ML) workflows (Virtanen *et al.,* 2020).
- SPSS and SAS offer GUI-driven simplicity, but are less flexible for automated and reproducible pipelines (Rahman and Muktadir, 2021).

9.2 Regression Modelling

Tools Employed:

- tidymodels framework, including recipes, parsnip, and workflows for model specification (Tidymodels-org, 2019).
- glmnet for Lasso and Ridge regression; caret for resampling and error estimation.

Effectiveness:

- High R^2 values (up to 0.94) validated model strength in predicting unsuccessful outcomes.
- Penalised regressions reduced overfitting and enabled variable selection, particularly in high-dimensional offence data (Lima *et al.,* 2020).
- Cross-validation reinforced the model's generalisability and reduced variance in performance metrics.

Alternatives:

- Python's scikit-learn can offer performance advantages for larger datasets, but at the cost of increased integration complexity (Virtanen *et al.,* 2020).
- Bayesian regression (e.g., brms) can capture posterior uncertainty, but it adds computational overhead and interpretive complexity (Bürkner, 2017).

9.3 Clustering Analysis

Tools Employed:

- Cluster, dendextend, and dbscan for K-Means, hierarchical clustering, and density-based clustering (STHDA, 2017).
- factoextra and pracma for determining optimal cluster counts and evaluating silhouette scores (STHDA, 2017).

Effectiveness:

- K-Means and Hierarchical Clustering (silhouette ≈ 0.54) revealed natural groupings of CPS regions.
- DBSCAN highlighted urban outliers not detected by centroid-based methods.
- Triangulation of clustering methods added robustness and supported strategic CPS resource segmentation.

Alternatives:

- HDBSCAN (Python) can automatically handle varying densities but requires Python integration (Malzer and Baum, 2019).
- Gaussian Mixture Models or Self-Organising Maps offer alternative paradigms but are less accessible to non-specialist audiences (Hunt and Reffert, 2021).

9.4 Classification Modelling

Tools Employed:

- ranger for Random Forest, rpart for Decision Tree, and xgboost for Gradient Boosting.
- yardstick for evaluation metrics (accuracy, precision, recall, F1-score, ROC-AUC).
- vip for variable importance plots.

Effectiveness:

- Random Forest achieved top performance (ROC-AUC ≈ 0.83, recall = 87%), crucial for identifying high-conviction regions.
- Decision Trees offered interpretability, while XGBoost captured nuanced patterns.
- Evaluation across multiple metrics ensured balanced model comparison.

Alternatives:

- LightGBM and CatBoost, implemented in Python, may offer speed advantages, particularly with large datasets (Ke *et al.,* 2017).
- Neural networks can uncover deeper, non-linear relationships; however, they lack transparency, which is inconsistent with the justice system's requirements (Rawat and Wang, 2017).

9.5 Integrated Toolchain and Visualisation Environment

Tools Employed:

- tidyverse for data wrangling (dplyr, janitor, lubridate, glue, reshape2)
- ggplot2 and extensions (ggrepel, viridis, scales, GGally) for data visualisation (Sanderson, 2024).

Effectiveness:

- Delivered a unified workflow from ingestion to interpretation, ensuring consistency and clarity.
- Visuals, including line plots, kernel density estimates (KDEs), boxplots, and receiver operating characteristic (ROC) curves, effectively conveyed trends and outliers, supporting hypothesis generation and model evaluation.

Limitations:

- Static outputs limited stakeholder interactivity.
- Complex plotting requires verbose code, which is not ideal for operational deployment without additional tooling.

Alternatives:

- Tableau and Power BI are ideal for creating interactive dashboards, particularly for managerial stakeholders, but they lack full scriptability (Parthe, 2023).
- Python's Plotly offers greater interactivity but requires integration effort (Van-Der-Donckt *et al.,* 2022).

Conclusion

This study adopted a comprehensive, multi-method approach to investigate prosecution outcomes within the Crown Prosecution Service (CPS) for the years 2014 and 2015. By combining data preprocessing, exploratory data visualisation, statistical hypothesis testing, and both supervised (regression and classification) and unsupervised (clustering) machine learning techniques, the research aimed to uncover structural patterns in conviction and unsuccessful prosecution outcomes across regions and offence types. Leveraging the R programming language and the tidymodels framework, the analysis was designed to be reproducible and interpretable.

The study's results revealed significant regional disparities in prosecution performance. The North and London consistently reported the highest volumes of convictions and unsuccessful outcomes, whereas the East region recorded the lowest, suggesting substantial differences in regional caseloads, resource distribution, and procedural efficiency. When analysed by offence category, sexual offences emerged as the type most prone to unsuccessful prosecution. This may reflect a complex interplay of evidentiary limitations, low victim engagement, and higher legal thresholds. In contrast, motoring offences, homicide, and robbery exhibited relatively higher conviction success rates, reinforcing the necessity for offence-specific strategies in both policy and predictive model design.

The regression modelling component of the project produced evidence that CPS case outcomes can be predicted with a high degree of accuracy. Linear, Ridge, and Lasso regression models—especially those incorporating violent crime counts—explained up to 94% of the variance in total unsuccessful prosecutions, as reflected by R^2 values. These findings highlight the utility of statistical modelling in identifying systemic inefficiencies. Similarly, the classification phase demonstrated the predictive capability of ensemble learning methods. Among the models tested, Random Forest yielded the best overall performance, achieving an ROC-AUC of approximately 0.83 and a recall of 87%. These results underscore the potential of interpretable machine learning for developing early warning systems and risk stratification tools within the CPS.

Unsupervised clustering added further insight into structural regional differences. Both K-Means and Hierarchical Clustering techniques revealed three natural region types—high-crime urban centres, mid-range regions, and low-volume rural areas—while DBSCAN highlighted specific urban hubs, such as Greater Manchester and West Midlands, as statistical outliers. These findings support the argument that prosecution patterns are not random but rather shaped by underlying geographic and operational factors, thus offering a data-driven framework for targeted resource allocation and policy intervention.

Despite these contributions, the research encountered several limitations. Firstly, the temporal scope was limited to two years, which restricted the ability to detect long-term trends or the effects of recent reforms. Secondly, the imputation of missing November 2015 data using records from 2016 may introduce a degree of temporal distortion. Thirdly, replacing dashes ("–") with zeros could underestimate actual missingness or introduce bias. Finally, while the exclusive use of R ensured a reproducible and academic-standard workflow, it may limit broader operational adoption by stakeholders more familiar with business intelligence platforms, such as Power BI and Tableau.

In response to these insights, several recommendations are proposed, including future studies that extend the analysis to include data from years beyond 2015, which would allow for the identification of longer-term policy effects and cyclical trends. Additionally, model performance could be enhanced by incorporating a broader range of variables, including socioeconomic indicators (i.e. unemployment rate), case complexity (i.e. case duration), and victim cooperation metrics. Also, the development of interactive dashboards using platforms such as Power BI or R Shiny is encouraged to facilitate real-time, stakeholder-accessible decision-making. Implementing advanced time-series forecasting techniques such as ARIMA would also enable better modelling of seasonal and cyclical variations. Lastly, targeted interventions should be prioritised for high-risk offence types and urban regions consistently associated with elevated rates of prosecution failure, particularly those identified as outliers via DBSCAN clustering.

In conclusion, this study addressed the objectives, including confirming the existence of regional and offence-type disparities in prosecution outcomes, demonstrating the feasibility of reliable predictive modelling, and revealing the presence of meaningful regional clusters. Collectively, these insights provide a robust foundation for evidence-based policy design and operational improvement across the criminal justice system.

References

Alwateer, M., Atlam, E.-S., Abd, M., Ghoneim, O.A. and Gad, I. (2024). Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications*, [online] 12(11), pp.53–75. doi:https://doi.org/10.4236/jcc.2024.1211004.

Balcan, M.F. and Sharma, D. (2024). *Learning Accurate and Interpretable Decision Trees*. [online] Openreview.net. Available at: https://openreview.net/forum?id=skdlnUYRzQ [Accessed 15 May 2025].

Battineni, G., Chintalapudi, N. and Amenta, F. (2020). Performance analysis of different machine learning algorithms in breast cancer predictions. *EAI Endorsed Transactions on Pervasive Health and Technology*, 6(23), p.166010. doi:https://doi.org/10.4108/eai.28-5-2020.166010.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2019). A Comparative Analysis of XGBoost. *ResearchGate*. [online] doi:https://doi.org/10.48550/arXiv.1911.01914.

Bhattacharyya, J. (2025). LASSO Regression -A Procedural Improvement. [online] *ResearchGate*. doi:https://doi.org/10.13140/RG.2.2.17732.13444.

Bürkner, P.-C. (2017). Advanced Bayesian Multilevel Modeling with the R Package brms. [online] *arXiv.org*. Available at: https://arxiv.org/abs/1705.11123 [Accessed 25 May 2025].

Chauhan, N.S. (2022). *DBSCAN Clustering Algorithm in Machine Learning*. [online] KDnuggets. Available at: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html. [Accessed 15 May 2025].

Chicco, D., Warrens, M.J. and Jurman, G. (2021). The Coefficient of Determination R-squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science*, [online] 7(5), p.e623. doi:https://doi.org/10.7717/peerj-cs.623.

Desboulets, L. (2018). A Review on Variable Selection in Regression Analysis. *Econometrics*, 6(4), p.45. doi:https://doi.org/10.3390/econometrics6040045.

Dhummad, S. (2025). The imperative of exploratory data analysis in machine learning. *Scholars Journal of Engineering and Technology*, 13(1), pp.27–31. Available at: https://doi.org/10.36347/sjet.2025.v13i01.005

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal: Promoting communications on statistics and Stata*, [online] 15(1), pp.292–300. doi:https://doi.org/10.1177/1536867x1501500117.

Erceg-Hurn, D.M. and Mirosevich, V.M. (2008). Modern robust statistical methods: An easy way to maximise the accuracy and power of your research. *American Psychologist*, [online] 63(7), pp.591–601. doi:https://doi.org/10.1037/0003-066x.63.7.591.

Erdely, A. and Rubio-Sánchez, M. (2025). D-plots: Visualisations for Analysis of Bivariate Dependence Between Continuous Random Variables. *Stats*, [online] 8(2), p.43. doi:https://doi.org/10.3390/stats8020043.

Finak, G., Mayer, B., Fulp, W., Obrecht, P., Sato, A., Chung, E., Holman, D. and Gottardo, R. (2018). DataPackageR: Reproducible data preprocessing, standardisation and sharing using R/Bioconductor for collaborative data analysis. *Gates Open Research*, 2, p.31. doi:https://doi.org/10.12688/gatesopenres.12832.1.

Hamed, B.A., Sadek, A. and El-Hafeez, T.A. (2023). Optimising classification efficiency with machine learning techniques for pattern matching. *Journal of Big Data*, 10(1). doi:https://doi.org/10.1186/s40537-023-00804-6.

Hunt, E.L. and Reffert, S. (2021). Improving the open cluster census. *Astronomy & Astrophysics*, 646, p.A104. doi:https://doi.org/10.1051/0004-6361/202039341.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J. (2022). K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*, 622(622). doi:https://doi.org/10.1016/j.ins.2022.11.139.

Ke, G., Meng, Q., Finley, T., Wang, T. and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. [online] Available at: https://www.researchgate.net/publication/378480234_LightGBM_A_Highly_Efficient_Gradient_Boosting_Decision_Tree.

Kim, Y.J. and Cribbie, R.A. (2017). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology*, 71(1), pp.1–12. doi:https://doi.org/10.1111/bmsp.12103.

Lima, E., Davies, P., Kaler, J., Lovatt, F. and Green, M. (2020). Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. *Scientific Reports*, 10(1). doi:https://doi.org/10.1038/s41598-020-64829-0.

Loog, M. (2017). *Supervised Classification: Quite a Brief Overview*. [online] arXiv.org. Available at: https://arxiv.org/abs/1710.09230.

Magklaras, A., Gogos, C., Alefragis, P. and Birbas, A. (2024). Enhancing Parameters Tuning of Overlay Models with Ridge Regression: Addressing Multicollinearity in High-Dimensional Data. *Mathematics*, 12(20), p.3179. doi:https://doi.org/10.3390/math12203179.

Malzer, C. and Baum, M. (2019). *A Hybrid Approach To Hierarchical Density-based Cluster Selection*. [online] ResearchGate. doi:https://doi.org/10.48550/arXiv.1911.02282.

Mankar, A., Bhoite, S., Kharade, K. and Raskar, K.A. (2024). *Metaanalysis of Overfitting of Decision Trees*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/389499622_METAANALYSIS_OF_OVERFITTING_OF_DECISION_TREES.

Maulud, D. and Abdulazeez, A.M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, [online] 1(4), pp.140–147. doi:https://doi.org/10.38094/jastt1457.

Mienye, I.D. and Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, [online] 10, pp.99129–99149. Available at: https://ieeexplore.ieee.org/abstract/document/9893798.

Murtagh, F. and Legendre, P. (2011). *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/51962445_Ward%27s_Hierarchical_Clustering_Method_Clustering_Criterion_andAgglomerative_Algorithm.

Naeem, S., Ali, A., Anam, S. and Ahmed, M.M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, 13(1), pp.911–921. doi:https://doi.org/10.12785/ijcds/130172.

Ogle, D.H. (2025). *Dunn's Kruskal-Wallis Multiple Comparisons.* [online] Rdrr.io. Available at: https://rdrr.io/cran/FSA/man/dunnTest.html [Accessed 15 May 2025].

Ostertagová, E., Ostertag, O. and Kováč, J. (2014). *Methodology and Application of the Kruskal-Wallis Test*. [online] Applied Mechanics and Materials. Available at: https://www.scientific.net/AMM.611.115.

Parthe, R.M. (2023). Comparative Analysis of Data Visualisation Tools: Power BI and Tableau. *Indian Scientific Journal Of Research In Engineering And Management*, [online] 07(10), pp.1–11. doi:https://doi.org/10.55041/ijsrem26272.

Pena-Araya, V., Pietriga, E. and Bezerianos, A. (2019). A Comparison of Visualisations for Identifying Correlation over Space and Time. *IEEE Transactions on Visualisation and Computer Graphics*, 1, pp.1–1. doi:https://doi.org/10.1109/tvcg.2019.2934807.

Rahman, A. and Muktadir, G. (2021). *SPSS: An Imperative Quantitative Data Analysis Tool for Social Science Research*. [online] ResearchGate. doi:https://doi.org/10.47772/IJRISS.2021.51012.

Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9), pp.2352–2449. doi:https://doi.org/10.1162/neco_a_00990.

Rossmo, K. and Pollock, J. (2019). Confirmation Bias and Other Systemic Causes of Wrongful Convictions: A Sentinel Events Perspective. *SSRN Electronic Journal*, 11(2). doi:https://doi.org/10.2139/ssrn.3413922.

Sánchez-Vinces, B.V., Schubert, E., Zimek, A. and Cordeiro, R.L.F. (2025). A comparative evaluation of clustering-based outlier detection. *Data Mining and Knowledge Discovery*, 39(2). doi:https://doi.org/10.1007/s10618-024-01086-z.

Sanderson, S.P. (2024). *Unveiling the Smooth Operator: Rolling Averages in R – Steve's Data Tips and Tricks*. [online] Steve's Data Tips and Tricks. Available at: https://www.spsanderson.com/steveondata/posts/2024-01-05/index.html [Accessed 15 May 2025].

Setiawan, I. and Suprihanto, S. (2021). Exploratory data analysis of crime report. *Matrix : Jurnal Manajemen Teknologi dan Informatika*, 11(2), pp.71–80. doi:https://doi.org/10.31940/matrix.v11i2.2449.

Shetty, P. and Singh, S. (2021). Hierarchical Clustering: A Survey. *International Journal of Applied Research*, 7(4), pp.178–181. doi:https://doi.org/10.22271/allresearch.2021.v7.i4c.8484.

STHDA (2017). *Determining The Optimal Number Of Clusters: 3 Must Know Methods - Articles - STHDA*. [online] Sthda.com. Available at: https://www.sthda.com/english/articles/index.php?url=%2F29-cluster-validation-essentials%2F96-determining-the-optimal- [Accessed 15 May 2025].

Tidymodels-org (2019). *A predictive modeling case study – tidymodels*. [online] Tidymodels.org. Available at: https://www.tidymodels.org/start/case-study [Accessed 15 May 2025].

Van-Der-Donckt, J., Van-Der-Donckt, J., Deprost, E. and Van-Hoecke, S. (2022). *Plotly-Resampler: Effective Visual Analytics for Large Time Series*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2206.08703.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E. and Carey, C.J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), pp.261–272.

Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research Journal*, 12(1), pp.228–242. doi:https://doi.org/10.2478/bsrj-2021-0015.

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLOS ONE*, 19(12), p.e0310839. doi:https://doi.org/10.1371/journal.pone.0310839.

Zappia, L., Richter, S., Ramírez-Suástegui, C., Kfuri-Rubens, R., Vornholz, L., Wang, W., Dietrich, O., Frishberg, A., Luecken, M.D. and Theis, F.J. (2025). Feature selection methods affect the performance of scRNA-seq data integration and querying. *Nature Methods*, [online] 22(4), pp.834–844. doi:https://doi.org/10.1038/s41592-025-02624-3.