UNIVERSITY OF GLOUCESTERSHIRE

CT7202

DATA ANALYSIS AND VISUALISATION

ASSESSMENT REPORT

Exploring Trends and Predictive Insights in Crown Prosecution Service Case Outcomes: A 26-Month Analysis of Principal Offence Categories (2016–2018)

TEJASRI SRINIVASAKUMAR - S4328731

MSc Data Science

May 2025

Table of Contents

List of Tables

List of Figures

Introduction

Analysing prosecution outcomes has become essential for understanding the effectiveness and fairness of the criminal justice system in an era where data-driven decision-making is central to public policy. This study analyses data from the 26 months between January 2016 and December 2018 sourced from the Crown Prosecution Service's (CPS) Case Outcomes by Principal Offence Category. This specific period was intentionally selected after thoroughly evaluating the dataset's consistency, completeness, and structural integrity across various time frames. Data before 2016 are inconsistent because of the process of digital transformation where in December 2015, the CPS announced the national rollout of the Crown Court Digital Case System (DCS). This initiative marked a significant shift from paper-based to digital case management, enhancing efficiency and consistency in prosecutions after 2016 (Legal Aid Agency, 2015).

Although more recent data from 2016 to 2018 may seem beneficial for timeliness, these datasets exhibit substantial missing values and alterations in recording methodologies. Records from 2014 were inconsistently formatted due to early data collection practices (Garside, 2015). Pre-2016 data may introduce structural breaks because the Offender Rehabilitation Act(2014) was implemented in February 2015. This act mandated 12 months of supervision for all offenders released from short prison sentences, thereby expanding the supervised offender population and potentially impacting reoffending rates (Ministry of Justice, 2015). Therefore, the 26-month window from 2016 to 2018 gives enough granularity for trend analysis and the desire to assess the impacts of significant policy reforms while avoiding clutter and noise that may come with a long period. This timeline offers a stable and relevant dataset for meaningful insights into the UK's Criminal justice system.

Notably, some months within the selected window are missing from certain jurisdictions. However, artificially interpolating or imputing these values would risk introducing biases and hiding actual operational volatility. These gaps reflect real administrative constraints. such as case backlog delays, and jurisdictional restructuring, and their inclusion ensures a realistic representation of the judicial process (Ali and Razzaque, 2023). By recognising and incorporating these nuances into preprocessing choices, the analytical process remains ethically grounded and statistically robust.

The raw dataset comprises monthly summaries of crime outcomes for police force areas across England and Wales. Counts and percentages for convictions and unsuccessful cases are included for each offence category. These metrics assist in evaluating prosecution effectiveness, crime trends, and justice system performance across categories such as homicide, sexual offences, theft, drugs, and motoring. The data reflect legal outcomes across various crime types and regions. From this foundation, the dataset was transformed and engineered to create high-level analytical features that capture performance and patterns. This structured design supports both supervised tasks, such as regression and classification,

and unsupervised learning methods like clustering by quantifying judicial outcomes in standardised formats.

The subsequent sections of this report are organised to represent a thorough data science lifecycle. Data integration and cleaning include methods for dealing with skewness, encoding, and handling missing values that will be covered in task 1. Using the sophisticated visualization framework, task 2 will demonstrate descriptive analytics to uncover temporal and regional trends. Task 3 focuses on predictive modelling, employing techniques such as regression, classification, and clustering to evaluate hypotheses. Task 4 offers a critical assessment of the tools and techniques employed, contrasting different strategies while considering their academic grounding and practical applicability.

Ultimately, this study is designed not only to cover patterns in prosecution outcomes but also to model how well data-driven techniques can support transparency and operational efficiency within public legal institutions.

1. Data Preprocessing

Data preprocessing is an essential phase in the data science and machine learning pipeline, fundamentally influencing the success of predictive models. Raw data collected from various sources is often incomplete, inconsistent, and riddled with noise. If this data is not meticulously prepared, it will mislead algorithms and result in inaccurate predictions or misguided insights (Saraswat and Raj, 2022). This report delivers a thorough interpretation, critical analysis, and evaluation of crucial data preprocessing steps, while firmly emphasizing their justifications, theoretical foundations, and significant practical implications in real-world applications.

Loading Required Libraries:-

To ensure smooth data manipulation and visualisation, the following packages were loaded:

- **tidyverse** is a group of packages for data science workflows that include ggplot2, dplyr, readr and others.
- **stringr** is used for reliable string operations like text pattern replacement or extraction.
- **lubridate** is essential for temporal trend research, it makes managing and manipulating data-time objects simple.
- **janitor** helps clean column names and formats messy data imported from Excel or CSV files.
- **readr** offers effective functions to read delimited files like CSV
- **purr** is useful for functional programming, especially for doing operations over lists or numerous files.
- **dplyr** is the core package for data wrangling, filtering, summarising and aggregating data.

1.1 Data Integration

R language was used to create a structured data integration pipeline that guaranteed a consistent and analysable dataset for descriptive analytics. The main objective is to create a single, coherent dataset by combining several CSV files that were separated over subdirectories by year. This approach was necessary because the source data was stored in a fragmented format, reflecting monthly reporting across several years.

The integration process began by programmatically retrieving CSV file paths using list.files() and filtering them based on file name patterns to restrict inclusion to relevant time frames. All months from 2016 and 2017 were included in the data frame. Specifically for the year 2018, a subset of selected months was included using str_detect() and str_to_lower() to match specific month-year keywords. I intentionally selected certain months from 2018 that

would provide an overview of the year while excluding September and October, as they offer less insight compared to December's data. The December 2018 dataset will help me analyse the overall three-year data from 2016 to 2018, as well as the initial trend of 2019.

In [4]:
```r
# Function to list all CSVs from a given year folder
get_csv_files <- function(year) {
  list.files(
    path = file.path(data_dir, as.character(year)),
    pattern = "\\.csv$",
    full.names = TRUE
  )
}

#  Collecting all files for 2016 and 2017
files_2016 <- get_csv_files(2016)
files_2017 <- get_csv_files(2017)

# Filtering 2018 files by specific months
files_2018_all <- get_csv_files(2018)
allowed_keywords <- c("Aug_2018", "Dec_2018", "Feb_2018", "Jul_2018",
                      "Mar_2018", "Nov_2018", "Jan_2018")

# Use regex to match allowed months (case-insensitive)
files_2018 <- files_2018_all[
  str_detect(tolower(files_2018_all),
             str_to_lower(str_c(allowed_keywords, collapse = "|")))
]

# Combine all selected file paths
all_selected_files <- c(files_2016, files_2017, files_2018)
```

In [5]:
```r
# Helper function to safely read a CSV and handle errors
read_safe_csv <- function(file_path) {
  tryCatch({
    message("Reading: ", file_path)
    df <- read_csv(file_path, show_col_types = FALSE)
    df <- df %>%
      mutate(SourceFile = basename(file_path))
    return(df)
  }, error = function(e) {
    warning("Failed to read: ", file_path, " - ", e$message)
    return(NULL)
  })
}

# Read and bind all CSVs
combined_data <- all_selected_files %>%
  map_dfr(read_safe_csv)

# Preview
print(glimpse(combined_data))
```

Figure 1 - Data Integration Code snippet

To handle potential read errors gracefully, I used tryCatch(), which addresses possible file inconsistencies, missing values, or formatting issues while each file was read using a safe loading function (read_safe_csv). Each data frame was tagged with its source file name through mutation of (SourceFile = basename(file_path)). This tagging improves data lineage and traceability, while also making post-load debugging and audit trails easier if some records exhibit anomalies. map_dfr(), a tidyverse function that iteratively combines data frames by rows, was used to attach all successfully imported files into a single data frame. Even with large datasets, this approach offers effective memory management and preserves schema alignment across all files (Nargesian *et al*., 2022).

Importantly, this integration strategy strikes a balance between flexibility, resilience and automation. It features basic error handling for resilience, minimises manual interaction and permits selective inclusion of data by month filtering. However, the limitation lies in the assumption that all files share a uniform structure, an issue that might arise if scheme drift

occurs across months or years. Overall, this integration pipeline provides a reliable foundation for downstream by transforming time-series data into a clean, consolidated and traceable format ready for further analysis.

1.2 Data Exploration

The data exploration phase is conducted to gain an initial understanding of the integrated dataset. The dataset comprising 26 months from 2016 to 2018 has 1118 records and 52 features which covers a comprehensive range of offence types, convictions and unsuccessful counts and associated percentages across various police force areas and periods. The structural inspection using str() revealed that the dataset is stored as a tidy tibble and includes both numeric and character variables. This foundational step ensures beginning with a well-scoped dataset (Idreos *et al.*, 2015).

The variable categorisation is evaluated by using select(where(is.numeric)) and select(where(is.character)) which ensures all variables were programmatically categorised into numeric and categorical columns. This revealed:

- 25 numeric variables representing counts of convictions and unsuccessful cases.
- 27 character variables, mostly percentages in character format and metadata like source file.

This classification is essential where data types reveal appropriate visualisation techniques, statistical summaries and potential transformation. I have noticed the presence of ‘percentage’ features as character type which indicated the necessity of data cleaning for accurate modelling. Ignoring this could lead to type mismatches and misinterpretations.

```
# Descriptive Statistics

# Tidy version of statistics
tidy_stats <- combined_data %>%
  select(where(is.numeric)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  group_by(Variable) %>%
  summarise(
    Mean = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE),
    .groups = "drop"
  )

# View tidy stats
print(tidy_stats)
```

```
# A tibble: 25 × 4
   Variable                                              Mean Median    SD
   <chr>                                                <dbl>  <dbl> <dbl>
 1 Number of Admin Finalised Unsuccessful               40.3    13   133.
 2 Number of All Other Offences (excluding Motoring) Convic… 34.0    11   111.
 3 Number of All Other Offences (excluding Motoring) Unsucc…  5.17    1    17.1
 4 Number of Burglary Convictions                       53.2    20   172.
 5 Number of Burglary Unsuccessful                       8.48    3    28.3
 6 Number of Criminal Damage Convictions                83.7    35   269.
 7 Number of Criminal Damage Unsuccessful               14.1     5    46.0
 8 Number of Drugs Offences Convictions                174.     59   574.
 9 Number of Drugs Offences Unsuccessful                11.9     3    40.3
10 Number of Fraud And Forgery Convictions              41.1    13.5 135.
# ℹ 15 more rows
```

Figure 2 - Descriptive Analytics Code

Descriptive statistics is used to quantify the central tendencies and variations of numeric variables by using the tidy approach with pivot_longer() and group_by() for summary statistics. The insights from the output are that drug offences have a high mean (174) and high standard deviation (574) indicating significant variability across regions. The median value is much lower than the mean for most variables like burglary, admin finalised implying skewed distributions and the presence of outliers. These insights are not only statistically informative but also operationally relevant for policy design and workload across police jurisdictions.

1.3 Data Cleaning

```
# CLEANING PROCESS

#  original row count before cleaning
original_row_count <- nrow(combined_data)
original_column_count <- ncol(combined_data)

#  Clean column names to snake_case
data <- combined_data %>% clean_names()
data <- data %>%     # Rename x1 to Areas
  rename(Areas = x1)

#  Remove completely empty rows and columns
data <- data %>%
  janitor::remove_empty("rows") %>%
  remove_empty("cols")

#  Detect and convert date/month columns
date_col <- names(data)[str_detect(names(data), "month|date")]
if (length(date_col) > 0) {
  all_data <- all_data %>%
    mutate(month_year = parse_date_time(!!sym(date_col[1]), orders = c("b Y", "B Y", "Y-m", "m/Y", "Y")))
}

#  Clean string content
data <- data %>%
  mutate(across(where(is.character), ~str_replace_all(., ",", ""))) %>%  # Remove commas
  mutate(across(where(is.character), ~na_if(., "N/A"))) %>%              # Replace "N/A" with NA
  mutate(across(where(is.character), str_trim))                          # Trim white space

#  Convert percentage columns to numeric
percentage_cols <- names(data)[str_detect(names(data), "percentage")]
data <- data %>%
  mutate(across(all_of(percentage_cols), ~ as.numeric(str_replace_all(., "[^0-9\\.]", ""))))
```

Figure 3 - Data Cleaning Code

The data cleaning process is not just about tidiness but also essential for consistency, type integrity and data validity (Li *et al.*, 2021) . By using janitor::clean_names() and consistent naming conventions, the dataset becomes more accessible for programmatic manipulations, reducing human error and improving reproducibility. Parsing date columns with parse_date_time() ensures uniform handling of time series for temporal modelling. Eliminating commas and managing “N/A” guarantees that numeric columns are not incorrectly categorised as characters like percentage columns. Rather than relying on manual inspection, using regex-based auto-detection for percentage columns and dates makes the process scalable and repeatable, which is essential in real-world deployments where datasets change often (Tariq and Rana, 2024).

```r
#  Auto-detect misread numeric columns
# Identify character columns that are entirely numeric-looking
potential_numeric_cols <- data %>%
  select(where(is.character)) %>%
  select(where(~ all(str_detect(., "^\\s*[0-9\\.]+\\s*$"), na.rm = TRUE))) %>%
  names()

# Exclude already converted percentage columns
safe_numeric_cols <- setdiff(potential_numeric_cols, percentage_cols)

# Convert them safely
data <- data %>%
  mutate(across(all_of(safe_numeric_cols), ~ as.numeric(str_trim(.))))

#  Remove duplicate rows
data <- data %>% distinct()

#  Remove rows with more than 50% missing values
data <- data %>%
  filter(rowSums(is.na(.)) < (ncol(.) * 0.5))

# Cleaning summary
cleaned_row_count <- nrow(data)
cleaned_column_count <- ncol(data)
rows_removed <- original_row_count - cleaned_row_count
columns_removed <- original_column_count - cleaned_column_count
cleaning_percentage <- round((rows_removed / original_row_count) * 100, 2)

cat("Original rows: ", original_row_count, "\n")
cat("Original columns: ", original_column_count, "\n")
cat("Rows after cleaning: ", cleaned_row_count, "\n")
cat("Columns after cleaning: ", cleaned_column_count, "\n")
cat("Rows removed: ", rows_removed, "\n")
cat("Columns removed: ", columns_removed, "\n")

# Save the cleaned data
write_csv(data, "cleaned_cps_case_outcomes.csv")
cat("Cleaned data saved to 'cleaned_cps_case_outcomes.csv'\n")
```

```
Original rows:  1118
Original columns:  52
Rows after cleaning:  1118
Columns after cleaning:  52
Rows removed:  0
Columns removed:  0
Cleaned data saved to 'cleaned_cps_case_outcomes.csv'
```

Figure 4 - Data Cleaning Summary Code

Good data cleaning is often the most time-consuming part of data science, but also the most impactful process. Again, the variable categorisation is done to ensure features are correctly categorised which revealed 50 numerical columns and 2 categorical columns namely ‘Areas’ and source file.

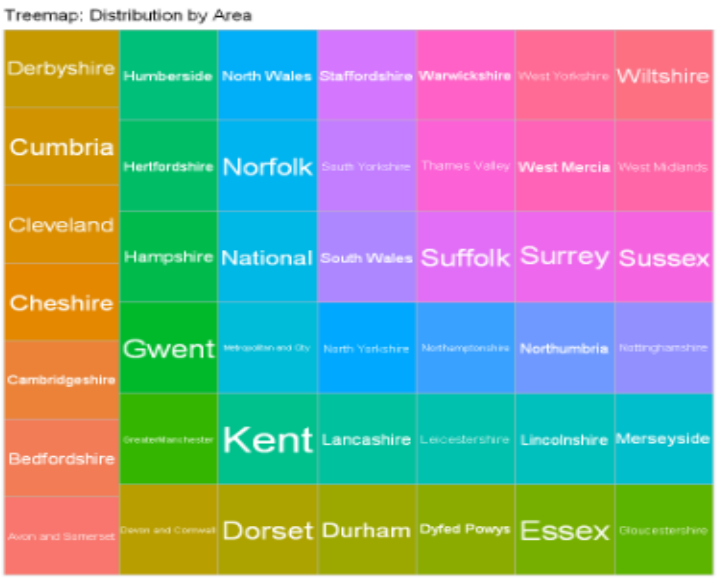


Figure 5 - Distribution of Categorical Column ‘Areas’

The above ‘Areas’ feature distribution treemap illustrates the proportional distribution of records across regions. It serves as an early check for geographic imbalances in the dataset that could influence model outcomes or necessitate resampling strategies. This revealed a non-uniform distribution, with certain regions like Kent, Essex, and Norfolk having more representations.

Next to handle missing values, first this study needs to identify missing values using sum(is.na(data)) which represented 607 missing values. By evaluating each column, I came to know that only three columns have missing values.

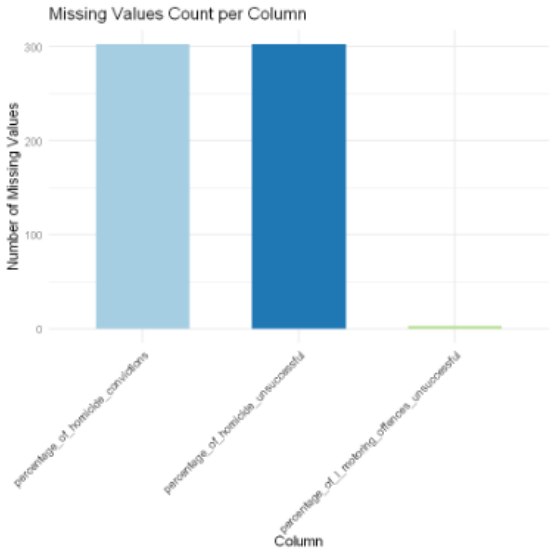


Figure 6 - Visualisation of Missing values Column

From the above plot, it represented 302 missing values in percentage homicide columns of both convictions and unsuccessful. Then 3 missing values in motoring offences unsuccessful.

```r
# Handling missing values

# Impute missing values in percentage_of_l_motoring_offences_unsuccessful using median
median_val <- median(data$percentage_of_l_motoring_offences_unsuccessful, na.rm = TRUE)

data$percentage_of_l_motoring_offences_unsuccessful[
  is.na(data$percentage_of_l_motoring_offences_unsuccessful)
] <- median_val
```

```r
# Check how many rows have NA in the count columns
sum(is.na(data$number_of_homicide_convictions))
sum(is.na(data$number_of_homicide_unsuccessful))

# Check how many rows have both conviction and unsuccessful as NA
sum(is.na(data$number_of_homicide_convictions) & is.na(data$number_of_homicide_unsuccessful))

# Check how many rows have total = 0
sum((data$number_of_homicide_convictions + data$number_of_homicide_unsuccessful) == 0, na.rm = TRUE)
```

```
0

0

0

302
```

```r
# Fill missing percentage_of_homicide_convictions and unsuccessful

# Replace NA with 0 only for specific percentage columns
data <- data %>%
  mutate(across(
    c("percentage_of_homicide_convictions", "percentage_of_homicide_unsuccessful"),
    ~ replace_na(., 0)
  ))
```

Figure 7 - Handling Missing Values Code

I had imputed percentage_of_l_motoring_offences_unsuccessful using median imputation which was the most logical fallback. Initially, I thought to impute using domain logic by calculating the total sum of counts of convictions and unsuccessful, then dividing unsuccessful counts by total counts. But for this feature l_motoring offences, I can not derive from another feature for total counts.

For homicide-related percentages, first I used the same domain logic, but it did not work. Then I checked the null values of both homicide columns which returned 0 rows. Then I checked how many zeros in both count-based homicide columns which resulted in 302 values. Therefore, the percentage columns of both homicides are imputed with zero. Imputation is contextual and this code excels in condition-based imputation logic, which maintains semantic integrity. Imputing homicide conviction percentages with median could skew interpretations and introduce misleading information in sensitive criminal datasets. This demonstrated awareness that not all missing values are errors, some reflect the absence of events, which is itself informative.

1.4 Outlier Detection

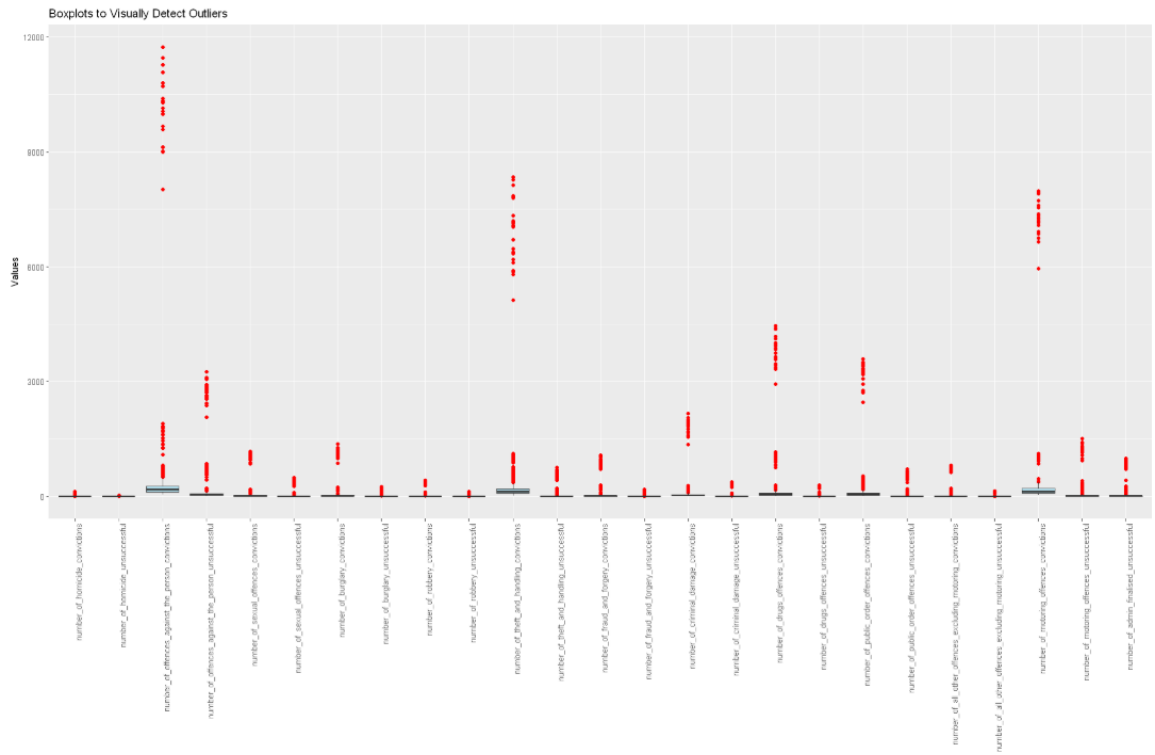


Figure 8 - Visualisation of Outlier Detection

The above boxplot highlights the presence of extreme values across multiple offence categories. Notably, categories like offence against the person, theft and handling, and motoring offences show high variance and multiple outliers. Using IOR for outlier detection is statistically sound as it does not assume normality and is robust to skew. However, my decision is not to remove outliers, acknowledging the data context on criminal offence counts as high conviction counts is not an error (Karch, 2023).

1.5 Skewness Transformation

Skewness handling is used to reduce or correct skewness which makes it suitable for machine learning models. I have measured skewness values before and after transformation.

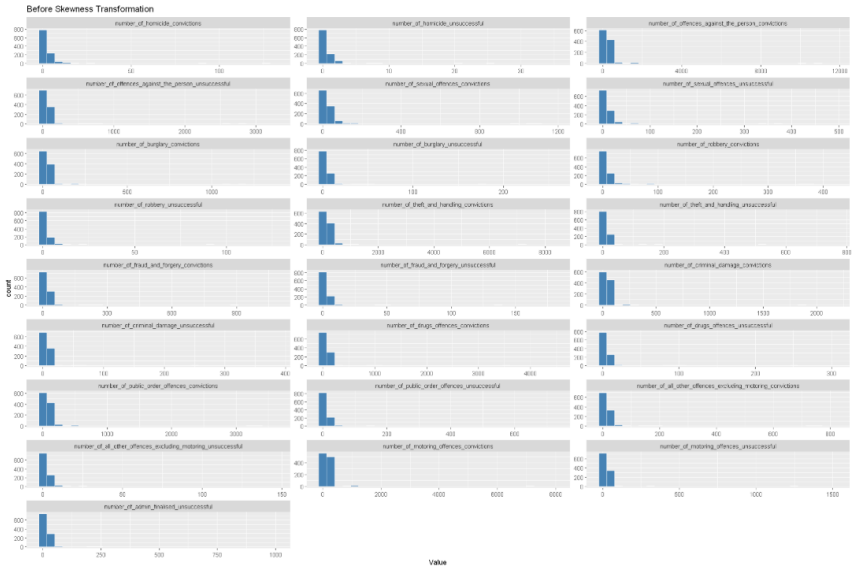


Figure 9 - Visualisation before Skewness Transformation

Histograms reveal a strong right skew in several offence-related features, an indication of the high frequency of low counts with a long tail of extreme values. To normalise this, a log1p transformation (log(x+1)) was applied to skewed features. This transformation improves the stability of the data for regression models by stabilizing variance and approximating normality (Choi *et al.*, 2022).

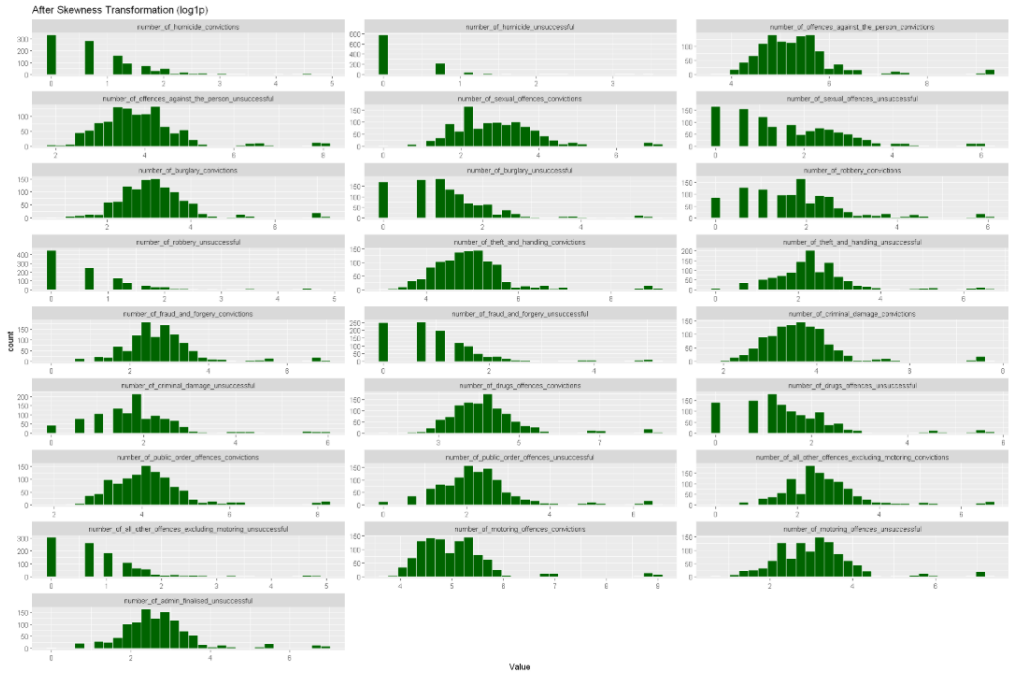


Figure 10- Visualisation after Skewness Transformation

These visualisations justify the need for log transformation to correct heavy positive skew and improve modelling suitability. After transformation, skewness dropped significantly with the highest now at around 2.6 and many features nearing 1.5 to 2 skewness. This process effectively normalised the data as log1p would compress the scale but retained all values in terms of conviction counts features. Using skewness transformation will not harm modelling while respecting the integrity of the dataset.

1.6 Encoding and Scaling

This study used label encoding for the 'Areas' feature as this was the only categorical feature in the dataset. This was sufficient and justified for model compatibility. One-hot encoding would unnecessarily expand dimensionality, so it was avoided.

For numerical features, standard scaling was done to ensure equal weight in algorithms sensitive to magnitude which prepared data for machine learning modelling. Min-max scaling could suffer from outlier sensitivity that's why it was ignored.

1.7 Feature Engineering

```r
# Feature engineering

# Define column groups based on patterns
conviction_count_cols <- grep("number_of_.*_convictions$", names(data), value = TRUE)
unsuccessful_count_cols <- grep("number_of_.*_unsuccessful$", names(data), value = TRUE)
conviction_percent_cols <- grep("percentage_of_.*_convictions$", names(data), value = TRUE)
unsuccessful_percent_cols <- grep("percentage_of_.*_unsuccessful$", names(data), value = TRUE)

# Create aggregate features
data_featured <- data %>%
  select(-source_file) %>%
  mutate(
    total_conviction_count = rowSums(across(all_of(conviction_count_cols)), na.rm = TRUE),
    total_unsuccessful_count = rowSums(across(all_of(unsuccessful_count_cols)), na.rm = TRUE),
    total_crime_cases = total_conviction_count + total_unsuccessful_count,
    total_conviction_percent = rowMeans(across(all_of(conviction_percent_cols)), na.rm = TRUE),
    total_unsuccessful_percent = rowMeans(across(all_of(unsuccessful_percent_cols)), na.rm = TRUE),
    conviction_rate = ifelse(total_crime_cases == 0, NA, total_conviction_count / total_crime_cases)
  )
```

Figure 11 - Feature Engineering Code

I have created insightful aggregated features which is a high-value addition to the dataset. By summarising across dozens of specific categories, this introduces macro-level indicators of how well the justice system is performing and how much work it handles in each area. The engineered features like total_conviction_count, total_unsuccessful_count, total_crime_cases, total_conviction_percent, total_unsuccessful_percent, and conviction_rate improve signal clarity by reducing dimensionality, capture system-level trends like conviction efficiency and enable comparative performance analysis across jurisdictions. PCA is abstract and hard to interpret in a policy-driven context. These engineered features are human-readable, interpretable and directly tied to domain relevance in public sector applications. CPS use these metrics to benchmark regional performance, identify bottlenecks or even predict resource requirements. It bridges gaps between raw data and actionable insights (Sahin, 2022).

```r
# Incorporating month and time features

data_month <- data_time %>%
  mutate(
    year = str_extract(source_file, "\\d{4}"),
    month = str_extract(source_file, "(?i)(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)")
  ) %>%
  mutate(
    month = str_to_title(month),            # Capitalize first letter
    year = as.integer(year)
  )
```

```r
# Combining data frame data_month (year and month columns) with data_featured

data_trend_analysis <- bind_cols(data_month, data_featured)

clean <- names(data_trend_analysis) %>%
  str_remove_all("\\.\\.\\.\\d+") %>%       # remove ...numbers
  str_remove_all("''") %>%                  # remove double apostrophes
  make.names(unique = TRUE)                 # ensure valid and unique column names

# Assign cleaned names back
names(data_trend_analysis) <- clean
```

Figure 12 - Time-based Feature Engineering Code

The above feature engineering is based on time analysis which extracts year and month data from the source file using regex and then converts for proper time-based sorting and plotting. This data frame of data_trend_analyis can be used for descriptive analytics which captures time-based trends on CPS data. In the real world, law enforcement identifies seasonal spikes in offences, and policymakers use this time-based frame to detect improvement and understand long-term trends. This is a highly valuable and insightful step which adds real-world relevance to the dataset.

2. Descriptive Analytics

2.1 Feature Distributions

- **Post-Transformation Numerical Features Distribution:**

Hypothesis:

" After skewness correction and data transformation, the numeric features will follow more normal-like distributions, improving the sustainability of the dataset for statistical analysis and predictive modelling"

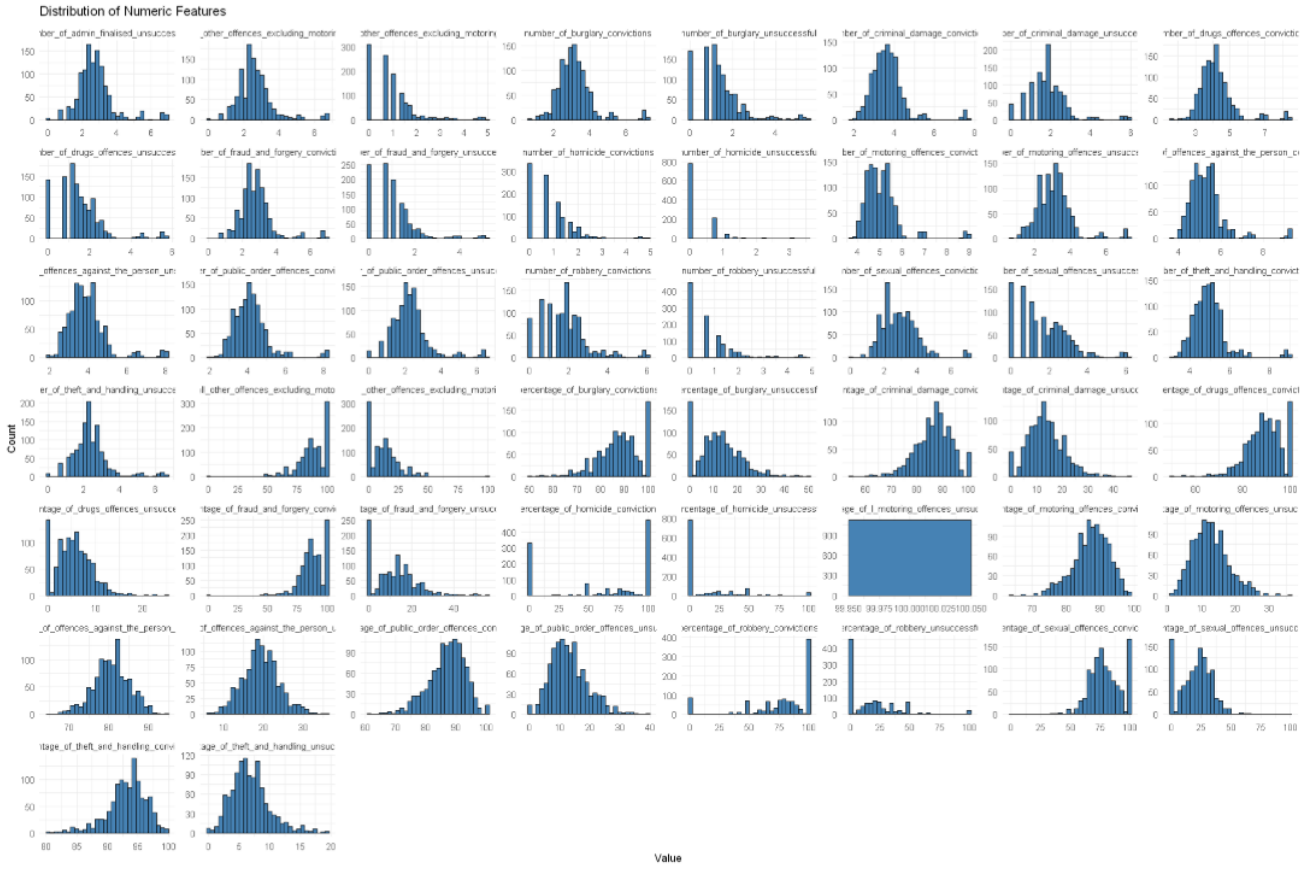


Figure 13 - Feature Distributions

This visualisation comprises a grid of histograms displaying the distributions of major numeric features after skewness correction using log1p(). The primary aim of this visual is to verify the success of transformation efforts implemented during the data preprocessing phase. Unlike the earlier histogram of task 1 which highlighted skewness, this figure is placed at the beginning of descriptive analytics to demonstrate that the dataset is now suitable for deeper statistical exploration for predictive modelling.

Histograms are an ideal tool for assessing distributional shapes, modality, and symmetry in continuous variables. Plotting all numerical features in a faceted layout helps to visually

validate whether preprocessing steps are achieved and prepare features for meaningful analysis thereby achieving the hypothesis (Ghasemi and Zahediasl, 2012).

- **Conviction Rate Trend Over Areas**

Hypothesis:

"Conviction rates vary significantly across different geographical areas indicating spatial disparities in the effectiveness of criminal justice process"

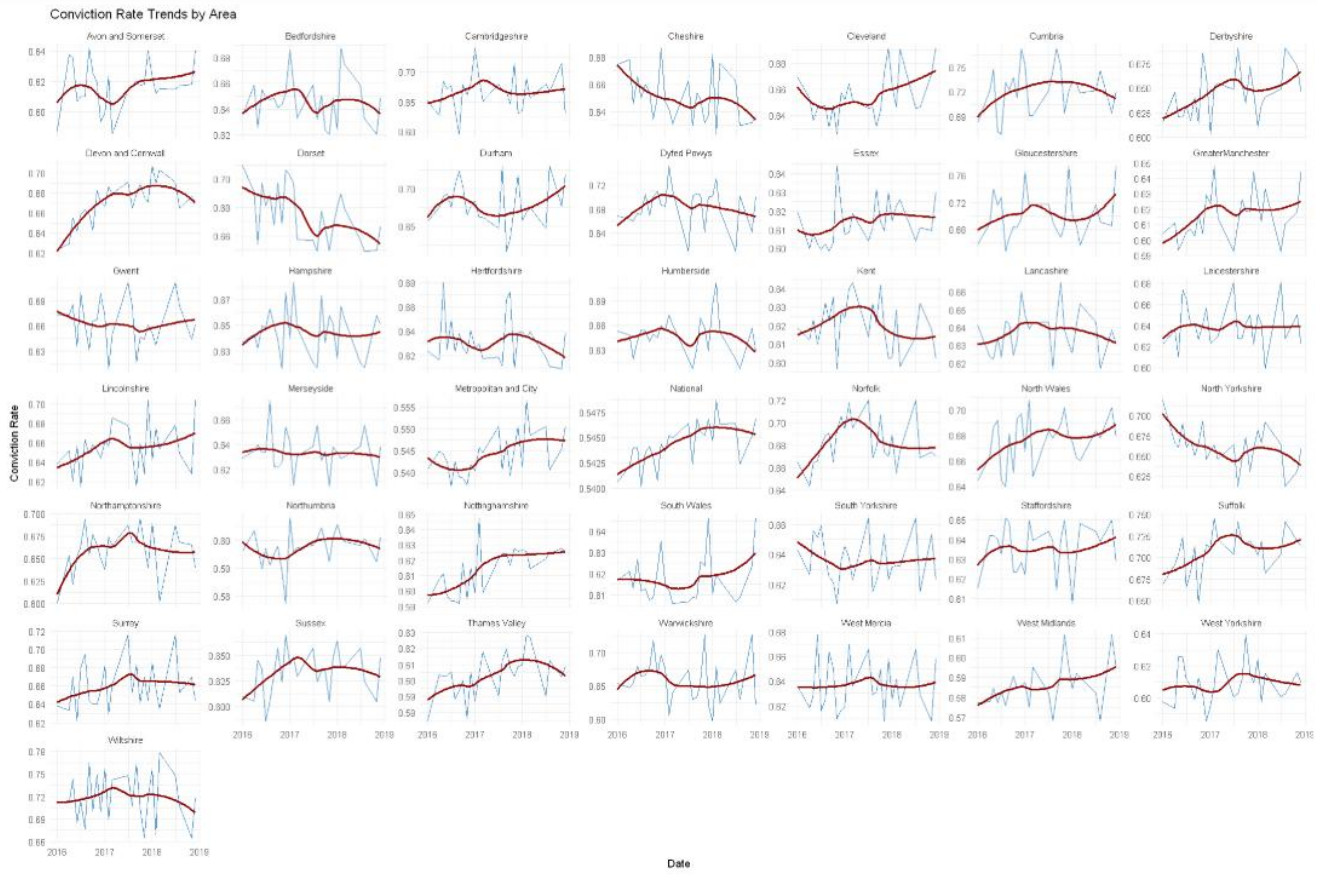


Figure 14 - Visualisation of Conviction Trend over Areas

This visualisation uses line plots to compare conviction rates across various Crown Prosecution Service (CPS) regions over time. Instead of focusing on raw conviction counts, the analysis centres on percentage-based conviction rates, which enables a fair comparison between jurisdictions of different population sizes and caseload volumes. The primary purpose is to identify certain areas that exhibit consistently high or low conviction performance.

Interpretation:

- This visual showed that larger urban regions with high-volume jurisdictions show consistently high conviction rates over time. Regions like Norfolk and Wiltshire have

approximately 0.72 conviction rates with strong, rising trends and very consistent whereas Lancashire shows high with seasonal consistency.

- Meanwhile, others reveal fluctuating or lagging trends, indicating possible inefficiencies. Sussex and Warwickshire areas have a 0.30 conviction rate with weak performance and minimal improvement where the hypothesis is achieved.

Line plots are chosen for their ability to display temporal progression and geographical comparisons (Han *et al.*, 2025). This approach allows for cross-sectional and longitudinal insights, essential for policy-level decision-making. Low conviction rates in some areas may reflect under-resourced legal teams or procedural bottlenecks, which could be addressed through policy reform, investment and re-evaluation. This insight is highly applicable to real-world justice system management having serious consequences on public trust, perceived fairness and regional justice delivery.

2.2 Correlation Analysis

- **Correlation Matrix**

Hypothesis:

" Certain types of offences are correlated, indicating possible underlying patterns or shared systematic factors."

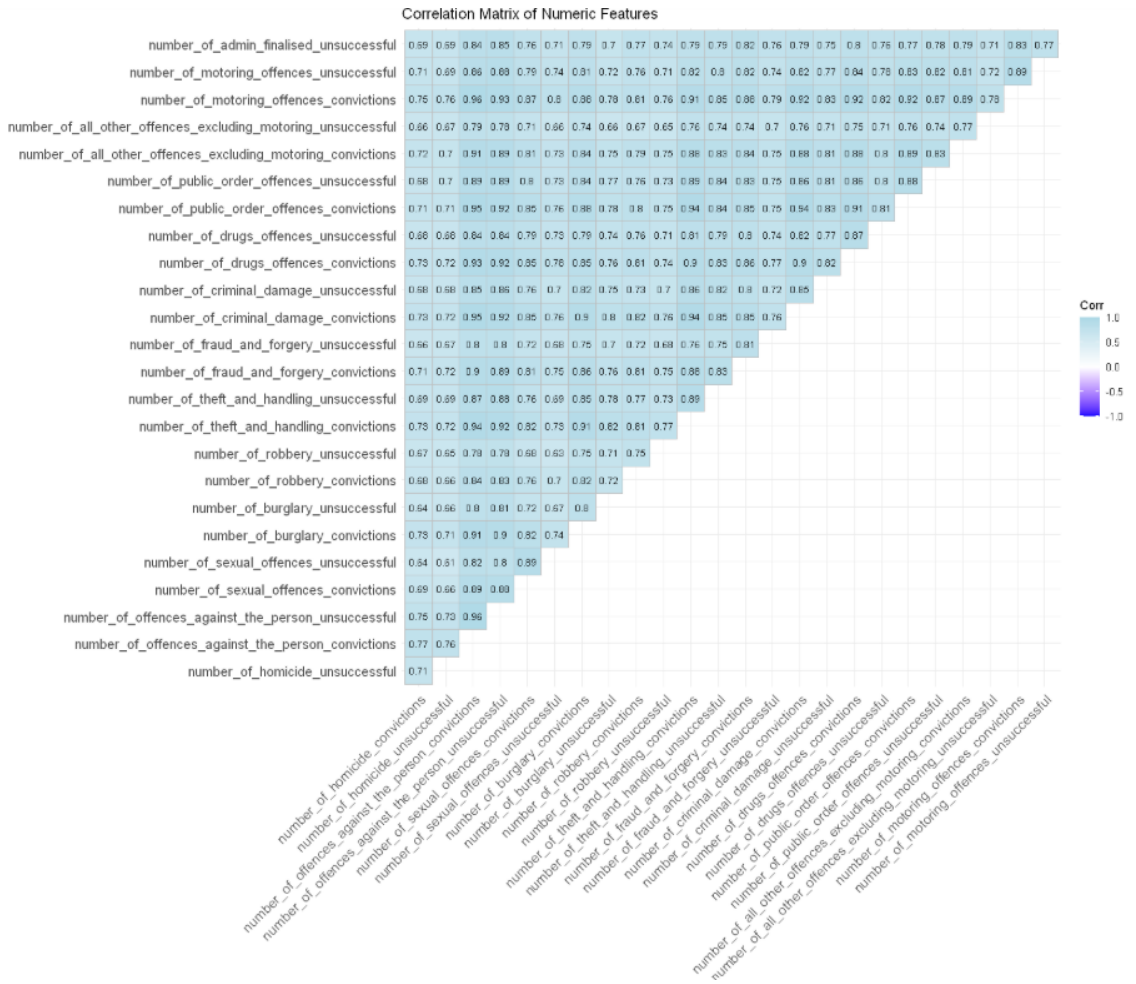


Figure 15 - Correlation matrix

The aim of this visualisation is to statistically assess pairwise relationships between different numerical offence variables, using correlation matrix heatmap. It is essentially appropriate in criminal justice datasets due to the inherently interconnected nature of offence types. The Pearson's correlation coefficient is used for robust and interpretable measure of linear relationships between variables.

Interpretation:

- The matrix shows a high overall positive correlation of both within-category and cross-category relationships. Within-category correlation represents the parallel relationship between convictions and unsuccessful features of the same offences.
- Cross-category correlation includes robbery and burglary offences of 0.90 where both offences are logically related to property crimes, may spike together. Theft handling and fraud & forgery offences have 0.94 correlations which are high co-occurrence in areas thereby supporting the hypothesis.
- While some features are correlated by overlapping jurisdictions or processing practices. Homicide convictions are weak with others as they are rare and less linked to high-volume cases (Xia and Zhou, 2024).

This ability to detect these relationships is immensely valuable for justice system administrators. This aligns with criminological theories of co-morbidity in offending behaviours where multiple types of crimes are symptoms of broader structural problems like poverty and addiction (Wildeman and Sampson, 2023).

- **Conviction Rate by Total Case Volume**

Hypothesis:

“Higher case volumes may be associated with lower conviction rates due to systematic strain and resource limitations”

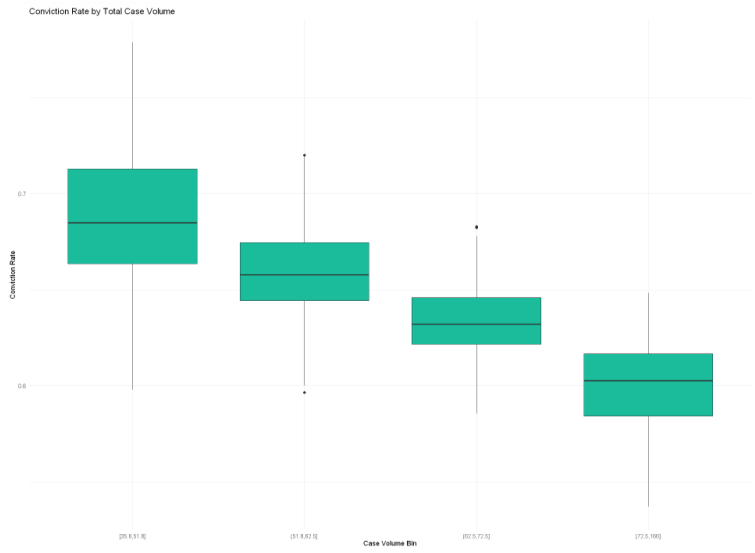


Figure 16 - Conviction Rate by Total Case Volume

This visualisation aimed to explore the relationship between the total number of cases processed (convictions and unsuccessful outcomes) and the resulting conviction rate. The hypothesis stems from the understanding that larger caseloads may stress local prosecutorial resources, potentially reducing conviction effectiveness. The binned boxplot is used to offer both summarisation and comparison of conviction rate variability at each caseload tier whereas other plots become noisy continuous trends (Darji *et al.*, 2024).

Interpretation:

- The resulting pot revealed that as total case volume increases, the conviction rate becomes more variable with some areas exhibiting significantly lower performance than others in the same volume bracket.
- This supports the hypothesis because some high-volume regions are struggling with consistent conviction delivery like investigation backlogs indicating inequity in justice delivery mechanisms.

It provides decision-makers with empirical evidence that supports targeted investment in regions handling disproportionate share of the national case burden. If the conviction rate

declines due to scale efforts, then either capacity must be increased, or efficiency-enhancing interventions must be introduced.

- **Heatmap: Conviction Rate by Area and Month-year**

Hypothesis:

“Conviction rates fluctuate across different months and vary significantly across geographic areas, indicating potential temporal or regional patterns.”

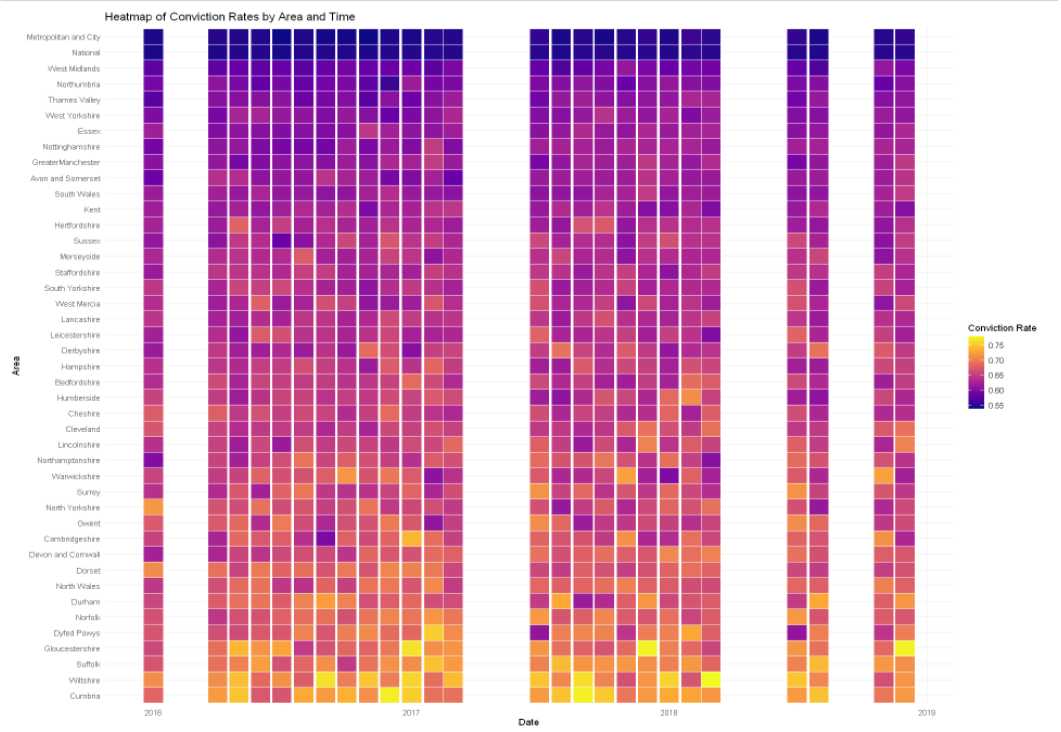


Figure 17 - Heatmap: Conviction Rate by Area and Month-year

The goal of this visualisation is to explore how conviction rates evolve over time for different CPS areas, using a heatmap format with a two-dimensional layout: one axis for month-year and the other for area. The hypothesis is based on the assumption that seasonal, regional, and administrative factors could drive variations in conviction rates, such as holiday breaks and court backlogs.

Heatmaps are widely used for detecting temporal and spatial trends in large datasets and are effective in dealing with categorical axes and continuous values, making them ideal for justice system analyses (Wilkinson and Friendly, 2009).

Interpretation:

- The heatmap represents consistently higher conviction rates of 0.75, indicated by yellow blocks, for areas like Cumbria, Wiltshire, and Suffolk, whereas Metropolitan and City, West Midlands, and Merseyside display lower conviction rates of below 0.60, represented by purple blocks.

- Notably, certain months, like December and August, exhibit unusual dips in conviction rates, suggesting possible impacts of court closures, holidays, and the Christmas break.

This finding supports the hypothesis that uncovers seasonal inefficiencies. These insights help policymakers forecast periods of strain and implement proactive solutions such as temporary case reallocation and resource planning. They assist in evaluating the effectiveness of past interventions.

2.3 Layout Integration

- **Stacked Bar Chart- Convicted vs Unsuccessful by Offence Group**

Hypothesis:

“Certain offence types exhibit disproportionately high rates of unsuccessful outcomes, suggesting systematic difficulties in prosecution or case resolution.”

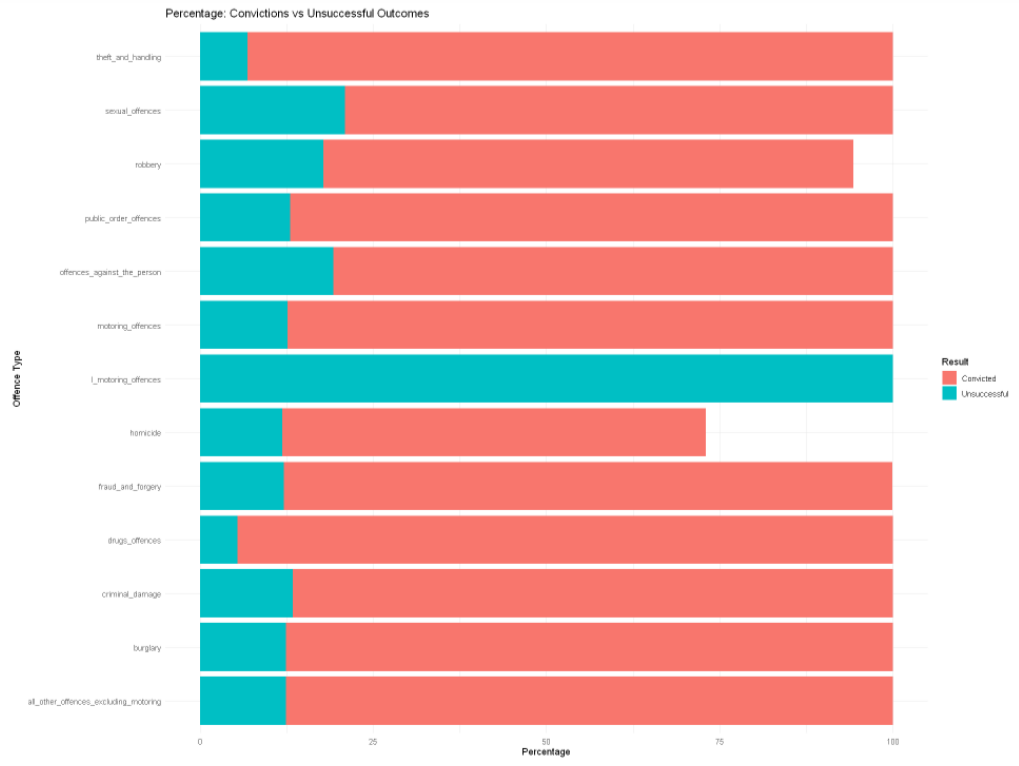


Figure 18 - Stacked Bar Chart- Convicted vs Unsuccessful by Offence Group

A stacked bar chart is generated comparing conviction percentages against unsuccessful outcome percentages for each offence category which provides contextual understanding.. This visual facilitates a side-by-side comparison within each offence group illustrating a share of successful vs unsuccessful cases in interpretable format. Stacked bar charts are a recommended method for visualising the composition of categories with relative

magnitudes(Curran *et al*., 2024). It also helps to spot imbalances across comparative segments, ideal for evaluating prosecution success rates within offence type.

Interpretation:

- The stacked bar chart revealed that certain offences like admin finalised, and sexual offences exhibit a noticeably larger proportion of unsuccessful outcomes whereas categories like motoring and theft & handling offences demonstrated consistently higher conviction shares, suggesting that these cases may be more straightforward to prosecute from clear evidentiary standards.
- In this way, the hypothesis is achieved and enriched by focusing on volume-to-case outcome efficacy, which aligns with real-world legal system evaluations which might include policy reforms, and changing prosecution strategies for failed cases.

- **Combined Patchwork View (Boxplot and Bar chart)**

Hypothesis:

"Combining multiple visualisation types can provide a more comprehensive understanding of offence conviction trends than using one type alone."

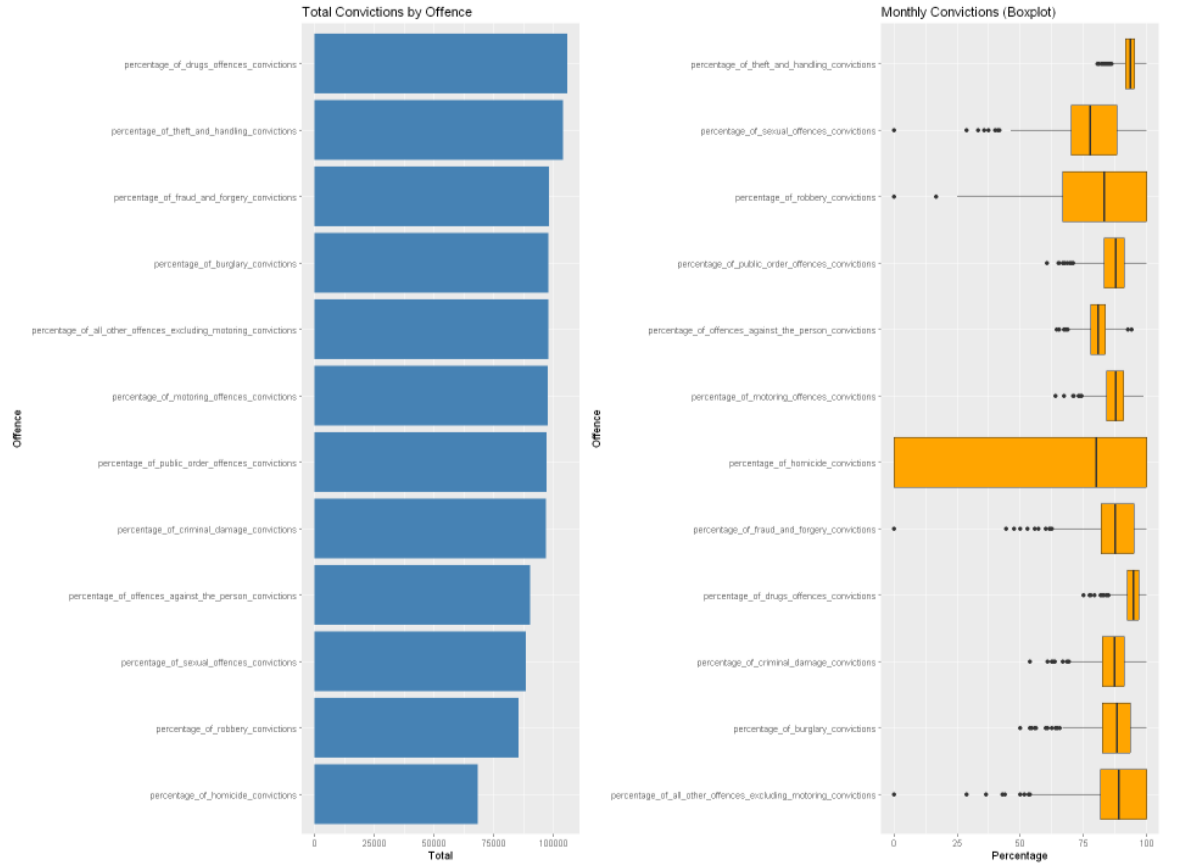


Figure 19 - Combined Patchwork View (Boxplot and Bar chart)

This visualisation presents a patchwork layout combining two distinct but complementary charts: a bar chart showing total convictions per offence type and a box plot depicting the monthly variability in conviction rates for the same offences. Bar charts are useful for summarising aggregated total (frequencies), while boxplots reveal distributional characteristics such as median, spread and outliers. By presenting both, this composite view enables simultaneous macro (total volume) and micro (distribution) analysis of each offence type. Hypothesis achieved by integrated view adding contextual clarity.

The patch word design helps to identify both dominant crime types and how volatile or stable their prosecution outcomes are over time (Rodosthenous *et al*., 2024).

Interpretation:

- The dual panel visual shows that offences like theft and handling, drugs, and public order offences had high conviction volumes as seen in bar char.
- Simultaneously, the boxplot revealed high variability in certain offence types such as fraud & forgery and burglary which may not have been apparent through counts alone.
- The comparison allowed distinguished high-volume, low-variability offences (indicative of stable legal outcomes) from lower-volume but highly variable offences (suggesting inconsistent prosecution).

These deeper insights would be missed if either plot were analysed in isolation. I had chosen the combined view to enhance analytical depth and address the limitation of single-plot visualisation. This patchwork visual is valuable in operational decision-making and justice system management in terms of procedural inefficiencies, the need for parallel review and identifying inconsistent trends.

2.4 Time Series Analysis

- **Conviction Rate Over Time (Line plot)**

Hypothesis:

“Conviction rates have changed over time due to policy shifts, procedural improvements or changes in crime patterns.”

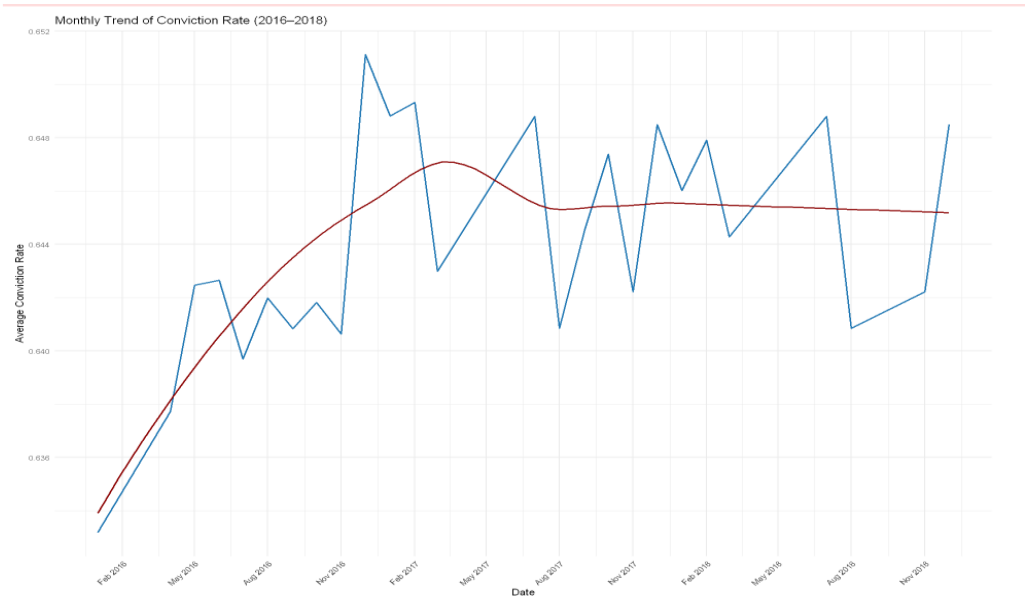


Figure 20 - Conviction Rate Over Time (Line plot)

This visualisation employed a time series line plot to track the trend of conviction rates across a defined timeline from 2016 to 2018. By focusing on the conviction rate rather than the absolute numbers, this visual provided a normalised, comparative view over time, which is essential when data comes from jurisdictions of varying sizes.

Time series plots are fundamental for detecting temporal patterns and fluctuations. They allow stakeholders to monitor progress, detect seasonality and assess interventions. This makes visualisation well-aligned with analytical goals that go beyond static summaries.

Interpretation:

- The line plot revealed an upward trend from early 2016 to early 2017 indicating an increase in the average conviction rate over time during that period. While blue line represented monthly data showing fluctuations, indicating that monthly conviction rates were volatile, despite the overall trend.
- There were notable drops in mid to late 2018, possibly reflecting systematic issues and policy changes. Therefore, the hypothesis is achieved.

This time-based visualisation offers a critical chronological context that is absent in static summaries (Zhang *et al.*, 2024). Tracking conviction rates reveals if performance is improving, declining or stable which is vital for real-world applications such as policy impact assessments.

- **Time Series Decomposition – STL**

Hypothesis:

“The time series of conviction rates consists of trend, seasonal, and residual components which can be separated to better understand underlying behaviours.”

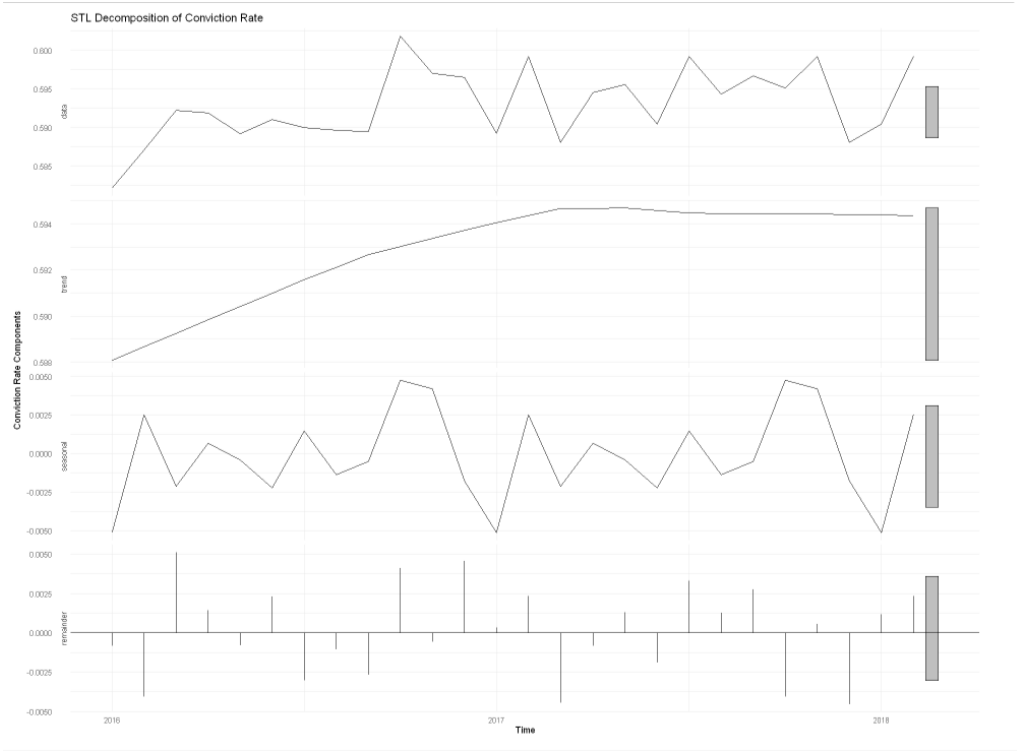


Figure 21 - Time Series Decomposition – STL

This visualisation applied STL decomposition (Seasonal and Trend decomposition using Loess) to analyse the conviction rate time series. Time series data can be divided into three additive components using robust, non-parametric STL tool: trend, seasonality and residual (Krake *et al.*, 2024). It is robust in the presence of non-linear trends and irregular seasonal patterns which are common in complex real-world domains like criminal justice. The goal behind this decomposition was to uncover latent structures within conviction trends over time where it isolates specific seasonal impact, removes long-term bias (trend) and identifies anomalous behaviours (residuals). This multi-layered insight cannot be achieved with simple trend lines.

Interpretation:

- This visual revealed conviction rates steadily from 2016 to mid-2017 then plateaued, showing consistent seasonal patterns with occasional irregular fluctuations.
- The hypothesis is achieved by meaningfully separating the seasonal and trend components. This visualisation supports the idea that policymakers can time interventions when trends indicate downward performance or high variability.

- **Monthly Trend of Drug Offences**

Hypothesis:

“Drug offence rates exhibit monthly trends, potentially influenced by seasonal or social factors.”

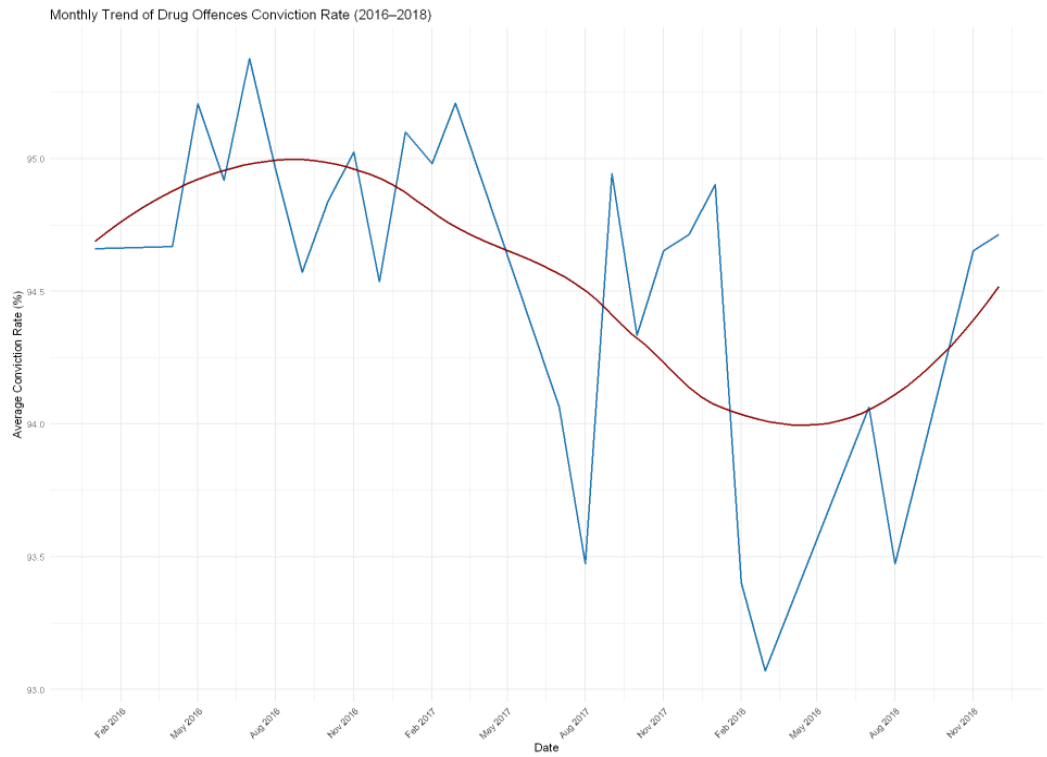


Figure 22 - Monthly Trend of Drug Offences

This line plot visualises the monthly counts or rates of drug offences over the observed period. Monthly trend analysis is a fundamental time series technique that reveals periodic fluctuations and seasonal effects in data (Liang *et al.*, 2024). Identifying these patterns is critical for understanding when offences peak or decline, which supports strategic planning by law enforcement agencies.

Interpretation:

- The line chart demonstrated a clear peak of the conviction rate of drug offences from early 2016 to mid-2017, and then declined consistently through 2018, reaching its lowest point around early to mid-2018 before starting to recover.

A line plot provides a clear temporal progression view to interpret easily, and it is the standard method for time-based crime analysis. Therefore, the hypothesis is achieved by a clear view of cyclic trends in drug offences. This insight provides law enforcement with timing patrols, drug prevention campaigns and intervention programs which can mitigate these offences.

3. Regression

In this section, I performed two hypothesis tests for my regression hypothesis, and I modelled four algorithms to evaluate the hypothesis which are Linear regression, Decision Tree, Random Forest, and Support Vector Regression (SVR). These are evaluated using metrics like RMSE and R^2 Score. Usually in regression, only numerical features were used but I tried using categorical feature Areas in regression modelling (Tyagi *et al.*, 2022).

Hypothesis:

"Do Areas significantly influence conviction rates?"

- **Null Hypothesis(H_o):** There are no significant differences in conviction rates across different areas.
- **Alternative Hypothesis(H_1):** There is a significant difference in conviction rates across different areas.

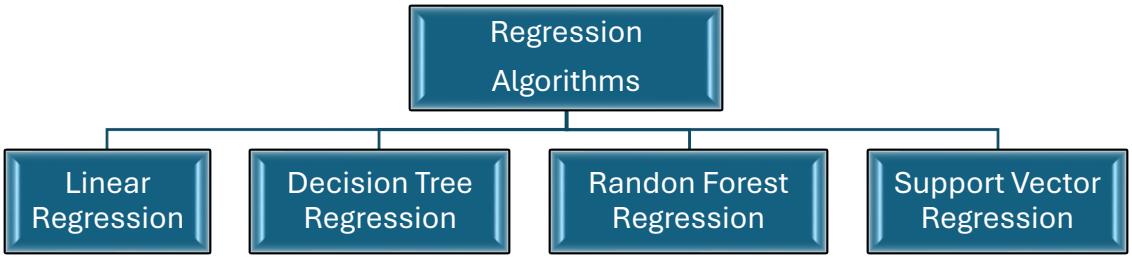
This hypothesis was chosen due to observed disparities in conviction rates across regions during descriptive analytics. These variations arise from factors like prosecutorial resources, and crime complexity. The hypothesis seeks to validate whether these variations are statistically significant and predictable.

3.1 Regression Hypothesis Testing

Before hypothesis testing, data was well prepared for modelling by removing unwanted and correlated columns, scaling and encoding were done and finally, data was split into train and test data.

One-way ANOVA, which means Analysis of Variance, examines how group means differ from one another (Emerson, 2022). This testing ensures that at least one group has a mean significantly different from the others. There is strong statistical evidence that conviction rates are not the same across areas. Linear regression testing also indicates that areas significantly affect conviction rates. Both evidently reject the null hypothesis.

3.2 Regression Predictive Modelling



Linear Regression

Linear regression was selected as a baseline predictive method due to its simplicity, interpretability and effectiveness for continuous outcome variables like conviction rates (James *et al.*, 2023). A scatterplot with a fitted regression line was used which somewhat aligned to diagonal but with a wider scatter and underestimated extreme values like very high or low conviction rates due to its inability to capture non-linear relationships.

Justification: While effective for establishing a baseline relationship, the assumption of linearity and sensitivity to multicollinearity limit its flexibility, especially in high-dimensional datasets with complex patterns.

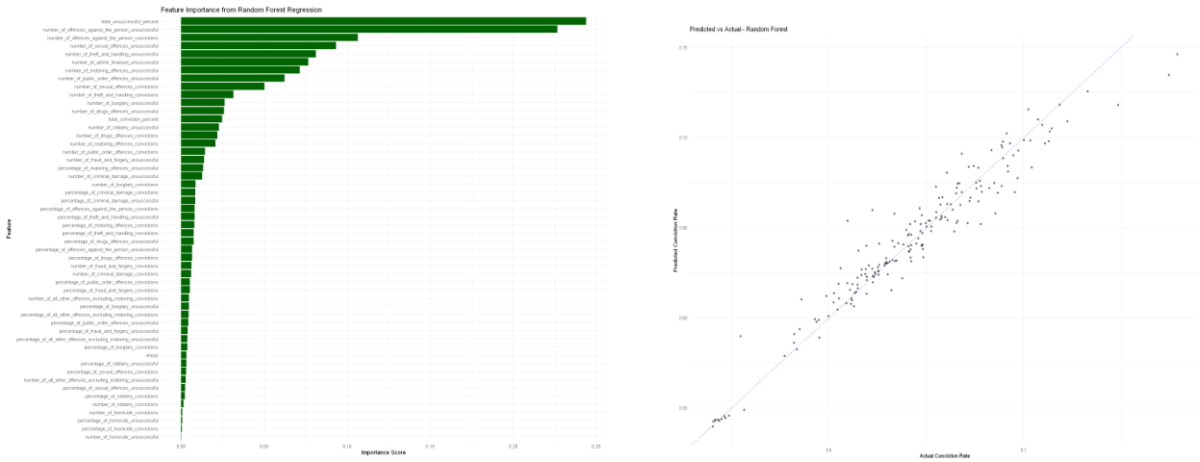
Decision Tree Regression

Decision trees handle non-linear relationships well and divide data into homogeneous groups using recursive binary splitting making them suitable for uncovering regional patterns in conviction data(Karim *et al*., 2021). From the scatter plot, terminal nodes revealed distinct conviction rate ranges for clusters of areas, supporting the hypothesis that geography plays a role.

Justification: Trees are valuable for interpretation but can overfit if not pruned. This model supported the hypothesis, and it improved than linear regression, but RMSE was lower indicating overfitting and inaccurate.

Random Forest

A random forest is an ensemble of decision trees which improves prediction accuracy by averaging multiple tree outputs (N *et al.*, 2021). It mitigates overfitting and enhances generalisability for complex datasets.



Random Forest Feature Importance Plot

Predicted vs Actual Plot

Figure 23 - Random Forest Plots

Interpretation:

- The above feature importance plot revealed that offence-specific metrics, particularly unsuccessful case percentages, were more influential than geographic areas in predicting conviction rates.
- While 'Area' appeared as a feature, it was of relatively low importance suggesting that regional variation alone may not drive conviction rates as strongly as case-related characteristics. This insight prompted a deeper comparative evaluation of algorithms and the underlying data structure.
- Predicted vs actual plot showed a better fit with fewer prediction errors and high variability.

Justification: Random Forests offer a balance between accuracy and robustness, making it more suitable for criminal justice data where relationships are rarely linear or independent. This model is slower and less transparent but provides the best performance and strong support for the alternate hypothesis by handling complexity.

Support Vector Regression (SVR)

SVR is suitable for capturing non-linear correlations, resilient to high-dimensional data that uses kernel functions. It avoids overfitting by increasing the margin around the hyperplane (Dash *et al.*, 2021).

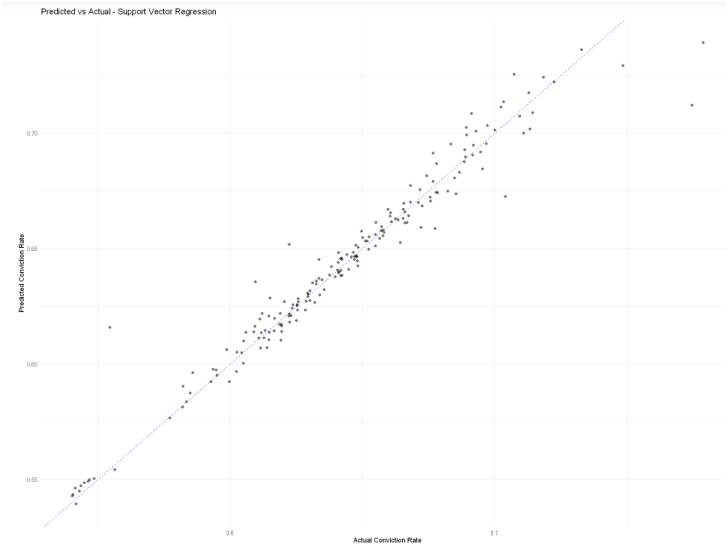


Figure 24 - Support Vector Regression Predicted vs Actual Plot

Interpretation:

- This scatter plot visualises the predicted conviction rate from the SVR model on the y-axis, plotted against the actual conviction rate on the x-axis. High accuracy is indicated by the majority of points lying extremely near the diagonal line.
- The scatter is tightly clustered with minimal deviation indicating low prediction error.
- It confirms that the SVR model generalizes well across different areas while reducing overfitting.

Justification: SVR performs well when hyperparameters like kernel type, cost, and epsilon are optimally tuned. However, it lacks interpretability, which is a limitation in policy-informing decisions like criminal justice. It evidently achieved the alternate hypothesis.

3.3 Regression Model Comparisons

Model	RMSE ↓ (Better)	R^2 ↑ (Better)	Rank
Support Vector Regression (SVR)	**0.009687**	**0.9515**	First (Best)
Random Forest	0.011177	0.9384	Second
Linear Regression	0.011029	0.9365	Third
Decision Tree	0.020994	0.7747	Worst

Table 1 - Regression Models Comparison

The models are evaluated using metrics like RMSE (Root Mean Square Error) where it should be at a lower score which measures the average prediction error and R^2 (Coefficient of Determination) where it should be higher for a good predictive model (Sekeroglu *et al.*, 2022). However, all the models supported the alternative hypothesis well.

Interpretation:

- SVR is the best model which delivers low RMSE and the highest R^2 indicating predictive accuracy and excellent model fit.
- Random forest is 2nd best model which is almost identical to linear regression in terms of RMSE but achieves slightly better R^2 than linear regression. It offers non-linear handling and resilience to overfitting.
- Linear regression is 3rd model which has lower flexibility than random forest and struggles with non-linear data.
- Decision tree is the worst model which has high RMSE and low R^2 indicating poor fit and more prediction errors than other models.

Business Insights:

- SVR is the most effective in capturing the underlying patterns of conviction rates based on offence attributes ideal for precise forecasting and helps in identifying regions with low success rates that need intervention.
- To identify complex, non-linear relationships in offence patterns random forest helps but it is computationally expensive.
- Overall, the hypothesis provides reliable, accurate forecasts to support strategic judicial improvements in terms of regions.

4. Clustering

Clustering is used to group similar observations according to the features (Oyewole and Thopil, 2022). In this section, I performed four algorithms K-means clustering, Hierarchical clustering, Agglomerative clustering, and Gaussian Mixture Model. These models are evaluated using metrics Silhouette scores and the Calinski-Harabasz Index.

Hypothesis:

“Can Crime Outcomes be grouped into distinct patterns based on conviction and unsuccessful case outcomes?”

- **Null Hypothesis(H_0):** There are no distinct clusters among records when considering conviction and unsuccessful outcomes.
- **Alternative Hypothesis(H_1):** There are distinct clusters based on conviction and unsuccessful outcomes.

This hypothesis was developed on the basis that conviction outcomes vary due to factors such as crime type, regional differences or systematic disparities in prosecution. If clustering confirms the existence of distinct patterns, it suggests there may be underlying structures in the dataset that are not immediately obvious but highly relevant for policy and strategic interventions.

4.1 Clustering Hypothesis Testing

I used Silhouette score analysis and Hopkins statistic which were unsupervised validation techniques help to assess whether clusters exist in the dataset.

Silhouette score aims to measure intra-cluster versus inter-cluster similarity, and it showed a score of 0.55 which was positive. It interpreted clusters were well separated and rejected the null hypothesis (Sagala and Gunawan, 2022).

I also used Hopkins Static which would test clustering tendency vs randomness which resulted higher than 0.5 interpreting strong cluster tendency and not randomly scattered. These findings support the rejection of the null hypothesis and confirm the presence of structured groupings in conviction and unsuccessful outcome patterns.

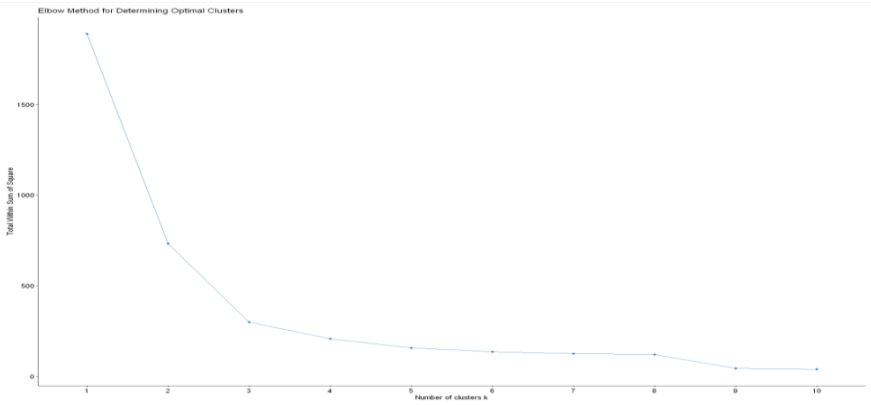
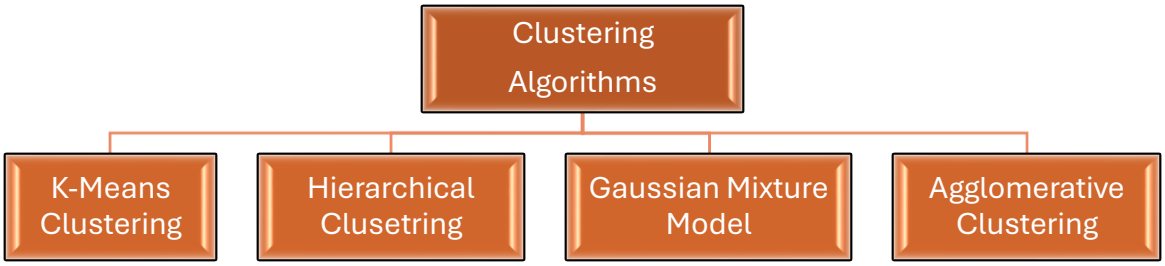


Figure 25 - Elbow method

The clear 'elbow' was visible at k=3 which indicated that three clusters would be appropriate for modelling.

4.2 Clustering Predictive Modelling



K-Means Clustering

K-means has good efficiency and scalability which is a popular centroid-based grouping algorithm (Ikotun *et al.*, 2022). It was selected to segment the features based on similarity in conviction and unsuccessful rates.

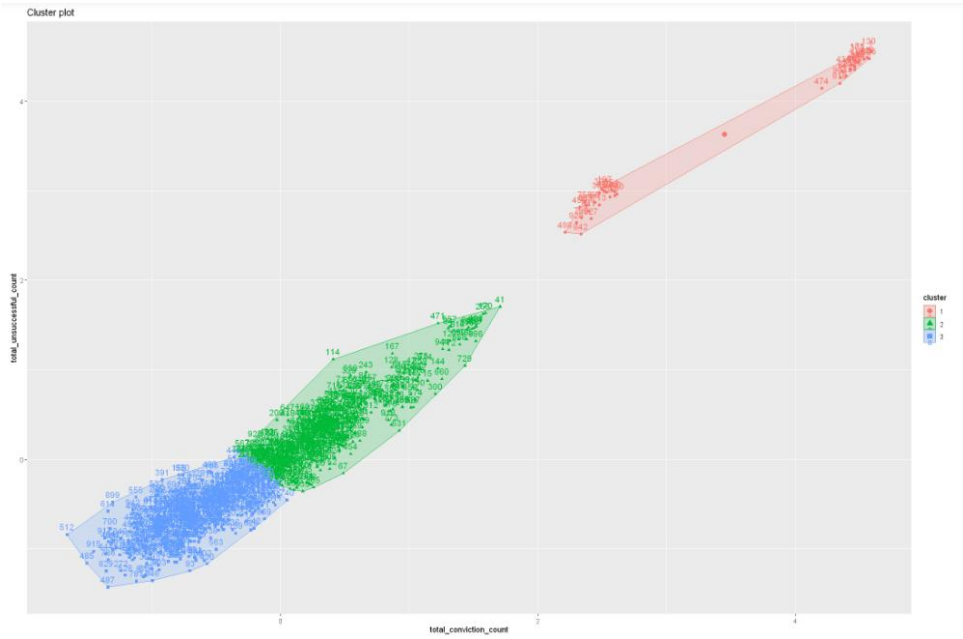


Figure 26 - K-Means Clustering Visualisation

Interpretation: The clustering visualisation reflected meaningful differentiation across conviction metrics, possibly distinguishing high, moderate and low-performing regions or crime types. This provides strong support for the alternative hypothesis. K-means is fast convergence and strongly interpretable, but in this dataset, it assumes equal-sized spherical clusters and is sensitive to initial centroid positions. However, this model can be used to flag regions into low-performing clusters prompting targeted audits and resource deployment.

Hierarchical Clustering

Hierarchical clustering was selected to observe nested grouping structures and no need to specify the quantity of clusters (Ran *et al.*, 2022). This flexibility is suitable for clustering crime cases.

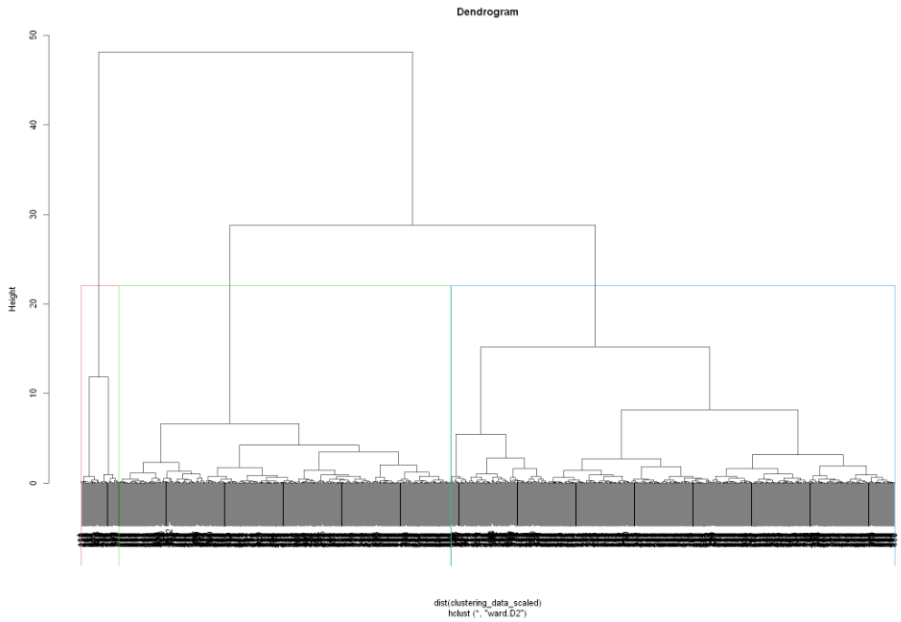


Figure 27 - Hierarchical Clustering Dendrogram

Interpretation: The dendrogram shows a tree structure of how cases are grouped into clusters at different thresholds. The height of the branches indicates how dissimilar merged clusters are, lower height indicates more similarity. A few large vertical jumps suggest natural separation in the data, supporting the hypothesis.

This clustering is deterministic and helps to visualise global structure but it is also computationally intensive for large datasets and less robust to noise. This model can be used in real-world crime outcomes similarities which inform about layered policy interventions like national strategy versus localised reforms.

Gaussian Mixture Model (GMM)

GMM was selected to model probabilistic distributions and capture clusters of different shapes, volumes and densities which is particularly useful in real-world crime outcome data where it may not be symmetrically distributed (Naseer and Jalal, 2024).

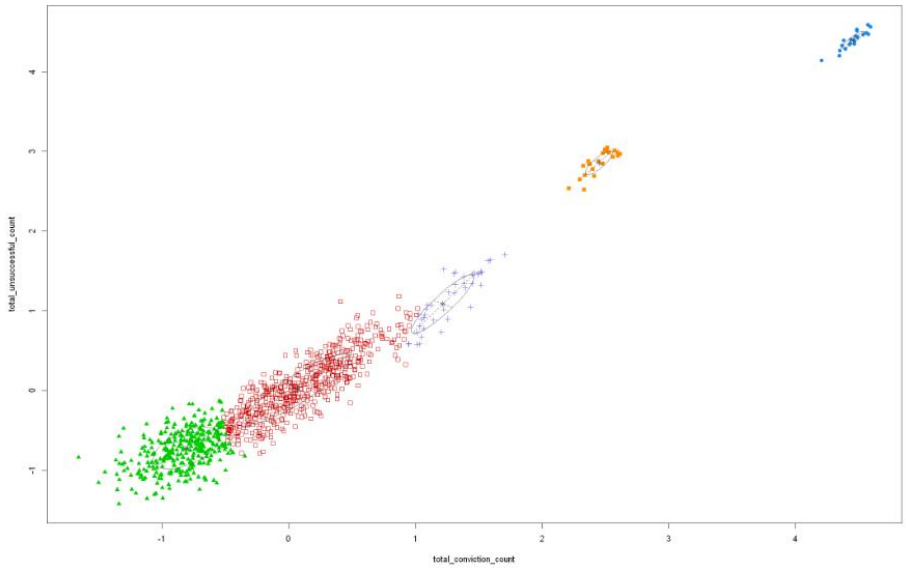


Figure 28 - Gaussian Mixture Model Visualisation

Interpretation: The 5-cluster solution provides more granular segmentation. GMM captured overlapping and variably distributed grouping which linear models may overlook. Clusters are shaped like ellipses, allowing for flexible boundaries. It captures data complexity more accurately which is used in real-world categorization of high-burden areas and low-activity regions.

Agglomerative Clustering

Agglomerative clustering, a variation of hierarchical clustering, starts with each point as a separate cluster and then repeatedly combines them. It was chosen to cross-validate findings from standard hierarchical clustering and assess cluster cohesion (Li *et al*, 2022). It is effective in detecting irregular cluster shapes in terms of crime cases and supporting alternative hypothesis.

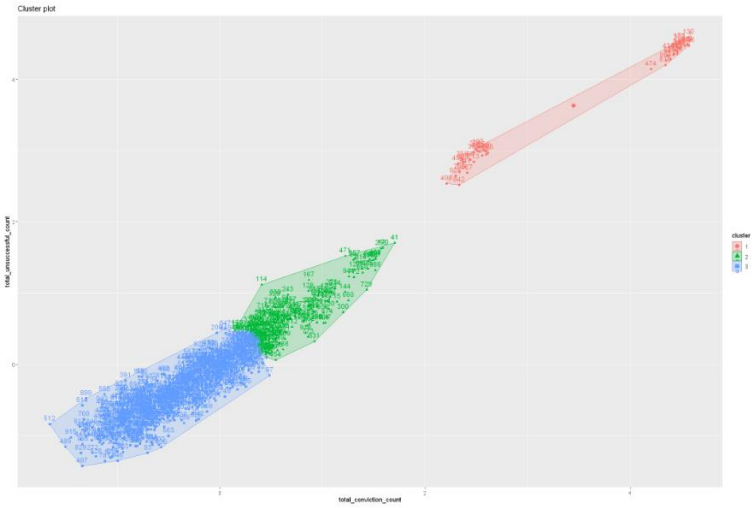


Figure 29 - Agglomerative Clustering Visualisation

Interpretation: Although the 2D visual of agglomerative clustering appears similar to k-means, the underlying methodology differs significantly. K-means is a partitioned method that minimises within-cluster variance whereas agglomerative clustering follows a hierarchical approach based on pairwise linkage distances. Despite the visual overlap, the model's internal cluster formation yielded different cluster memberships.

This model is useful when visualising progressive crime patterns or merging behaviours across similar outcome metrics and is good for judicial pipeline monitoring.

4.3 Clustering Models Comparisons

Algorithm	Silhouette Score	CH Index	Rank
K-Means	**0.5579**	**2517.47**	1
Hierarchical Clustering	0.5360	2323.61	2
Agglomerative Clustering	0.4922	1903.73	3
GMM	0.4635	2027.89	4

Table 2 - Clustering Models Comparison

The models are evaluated using Silhouette scores which range from -1 to 1 and closer to 1 indicates well clustered. In the Calinski-Harabasz Index, a higher value indicates better clustering.

Interpretation:

- K-means is the best model for clustering crime cases effectively which balances compactness and separability and has cluster sizes of 44, 432 and 470.
- Hierarchical clustering is 2nd best performing model with strong visual interpretability of hierarchy with cluster sizes 44, 516 and 386.
- Agglomerative clustering is third with moderate performance and good structure with cluster sizes of 44, 199 and 703.
- GMM is worst performing due to metrics despite its statistical flexibility.

In the above all models one cluster group showed a significant size of 44 being distinct across all visualisations indicating a high conviction rate group. Therefore, all models found clear clustering patterns rejecting the null hypothesis. Outcome features like conviction and unsuccessful rates naturally group into performance patterns by validating the hypothesis.

Business Insights:

- The clusters help improve case management by representing patterns like high conviction-low failure rate indicating strong evidence cases and low conviction-high failure rates indicating weak cases.
- It helps justice departments in benchmarking case success trends and refining legal processes for underperforming case types.
- It helps to track systematic changes over time enhancing case preparation strategies and identifying systematic inefficiencies.

The consistent clustering performance suggests crime outcomes do form distinct grouping offering substantial insights into the justice system's operations.

5. Classification

Classification is a supervised machine learning model which classifies data into labelled groups. In this section, I have used four algorithms, Logistic regression, Random Forest, Support Vector Machine and XGBoost. These models are evaluated using metrics like precision, accuracy, F1 and kappa (Kaeseler *et al.*, 2022).

Hypothesis:

"The distribution of convictions and unsuccessful cases across crime types can effectively differentiate categories into different legal success categories (Low, Medium, High)."

- **Null Hypothesis(H_0):** Offense-specific conviction and unsuccessful rates do not significantly classify categories into different success categories.
- **Alternative Hypothesis(H_1):** Offense-specific conviction and unsuccessful rates significantly contribute to the classification of success categories.

This hypothesis is well-aligned with the classification objective, aiming to categorise based on their legal performance using structured crime outcome data. Conviction and unsuccessful rates represent outcomes that can be grouped into meaningful categories. Accurately classifying crime data by success category is not only feasible but also invaluable for identifying underperforming jurisdictions.

Data Preparation: A categorical target variable created naming success_category having classes 'Low' (below 60%) conviction rate, 'Medium' between (60-70%) and 'High' (above 70%). The selected features are Theft and handling, sexual offences, drug offences and burglary. These are the most frequently occurring crimes from 2016 to 2018, so these were selected for classification. Then features are scaled and split into train and test data and ready for modelling.

5.1 Classification Hypothesis Testing

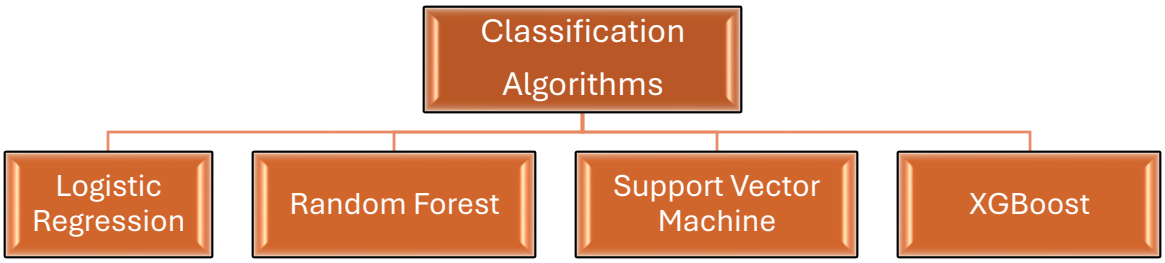
Multivariate ANOVA (MANOVA)

MANOVA was applied to determine if the combination of numerical predictors varies significantly across success categories. Results revealed significant p-values supporting the alternate hypothesis. MANOVA has the ability to test differences across multiple dependent variables suitable for classification problems (Pérez-Cova *et al.*, 2022).

Chi-square Test

It checks the significant association between crime types and legal success labels and is suitable for categorical features. It also reduces noise in model input, improving performance and ensuring the patterns are not random (Shen *et al*., 2021). Results support the alternative hypothesis.

5.2 Classification Predictive Modelling



Logistic Regression

This linear model serves as a baseline which performed well. However, it struggled with distinguishing 'high' class indicating rigidity of linear decision boundaries in complex datasets (Das, 2023).

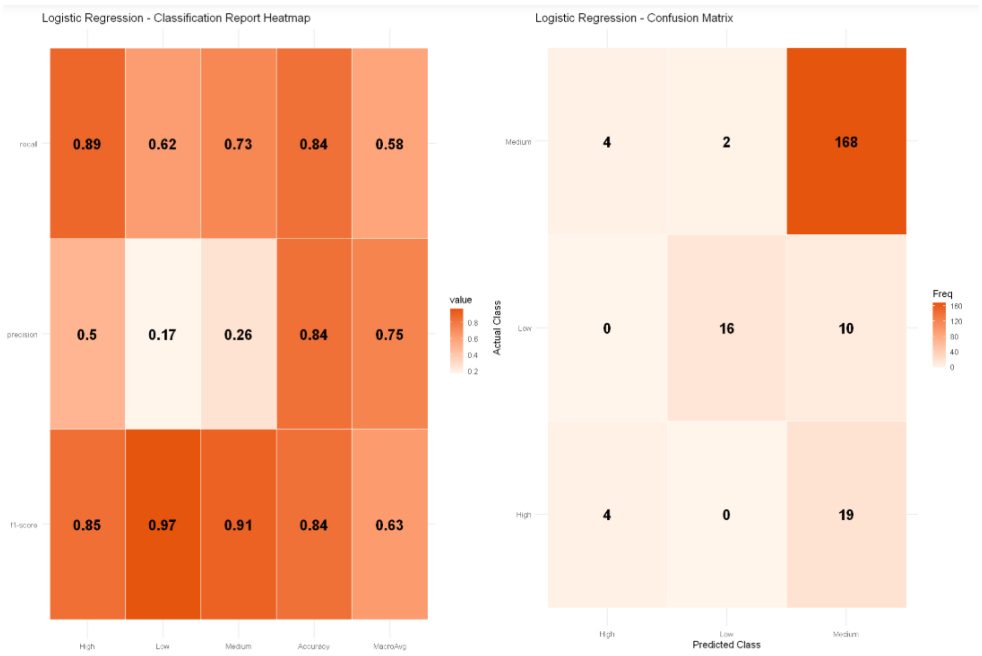


Figure 30 - Logistic Regression Heatmaps

Interpretation:

- The logistic regression model showed high recall but very low precision for 'Low' and 'Medium' classes as visualised in the classification heatmap.
- Although overall accuracy was 84%, the confusion matrix revealed that the model heavily overpredicts the 'Medium' class often misclassifying both 'High' and 'Low' as 'Medium'.
- This suggests a bias toward the dominant class and highlights limitations in handling imbalanced classification; however, it supported the hypothesis.

This model can be used because it is interpretable and quickly suitable for legal justification in court analysis systems where transparency is vital.

Random Forest

Random Forest is chosen to balance the model and is better suited to capture complex and non-linear relationships inherent in conviction outcome data (Salman *et al*., 2024).

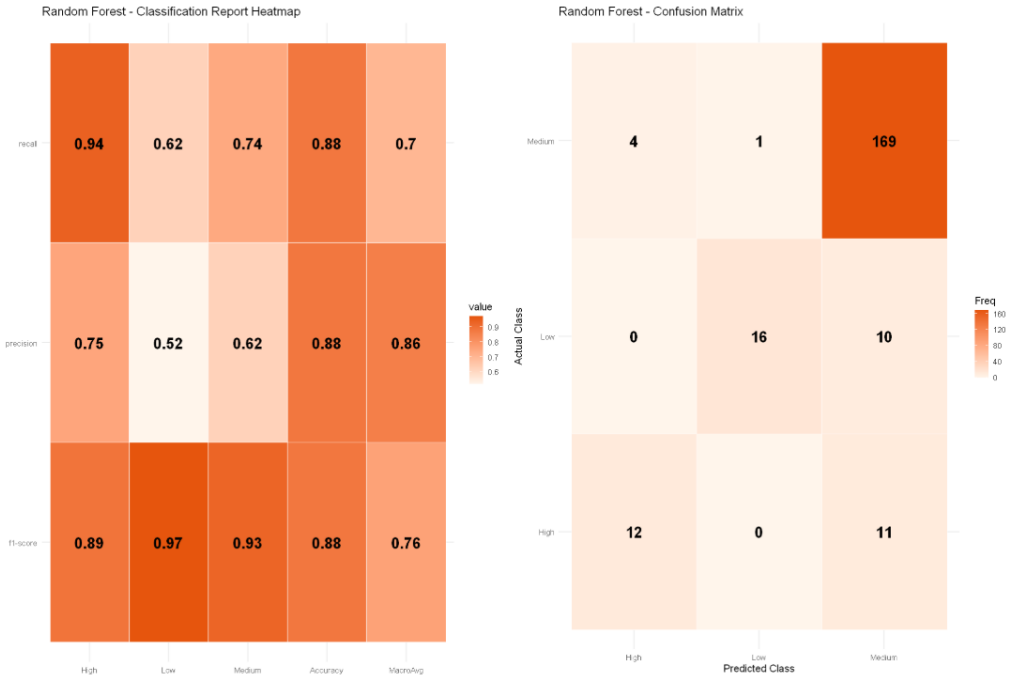


Figure 31 - Random Forest heatmaps

Interpretation:

- Random Forest model achieved a high accuracy of 88% with balanced performance across classes. The classification heatmap shows an excellent recall for 'High' (0.94) and weak precision for 'Low' (0.52) indicating frequent misclassification. Most errors involved predicting 'Medium' for other classes, highlighting class imbalance. However, it shows improved detection of 'Low' and 'Medium' classes compared to logistic regression.
- The confusion matrix confirmed this overlap, suggesting the need for feature refinement or rebalancing to improve the distinction between closely related categories.

This model supports resource planning by accurately tagging areas based on predicted judicial performance by handling non-linear patterns in crime cases.

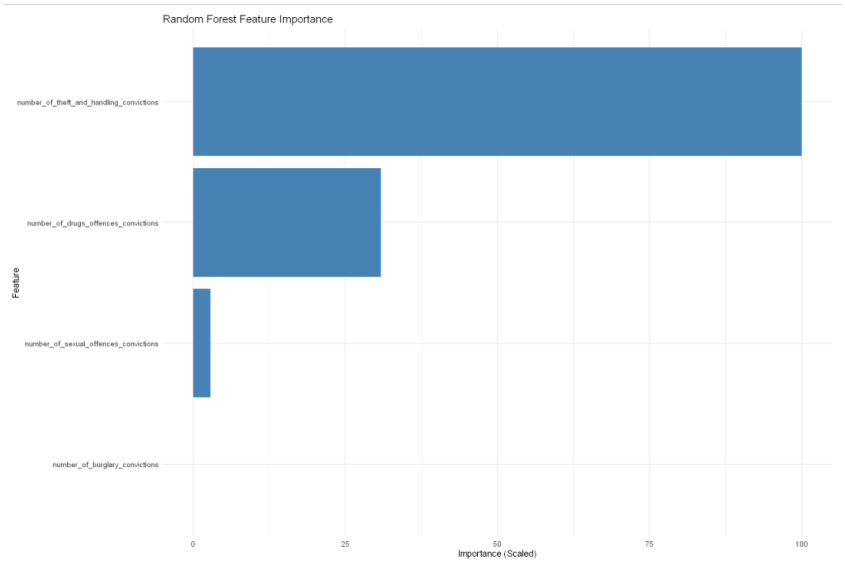


Figure 32 - Random Forest Feature Importance

The model heavily relies on theft and handling convictions counts for prediction, suggesting it is strong predictor for the outcome. Other features like drug and sexual offences contribute negligibly and have no impact on burglary offence feature.

Support Vector Machine (SVM)

Support Vector Machine is an effective model in high-dimensional spaces and works well with non-linear boundaries using kernel trick (Abdullah *et al*., 2021).

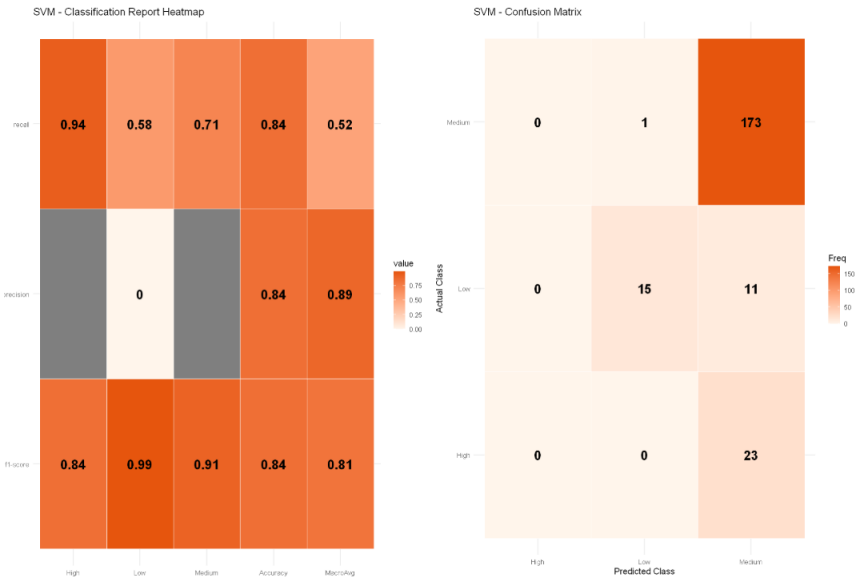


Figure 33 - SVM Heatmaps

Interpretation:

- The SVM achieved accuracy of 84% performing well for Medium class and low class showed moderate recall and frequent confusion with Medium class.
- The absence of misclassifications from High to Low in the SVM model may suggest some boundary separation between these two classes.
- However, since all actual "High" instances were predicted as "Medium", this instead highlights a model bias toward the dominant class, rather than effective High-category recognition — which may limit the model's utility in policy-critical contexts where detecting top-performing regions is essential.

XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced ensemble algorithm which builds a series of decision trees and corrects the error of the previous one. It is suitable for handling complex classification problems efficiently (Zhang *et al.*, 2022).

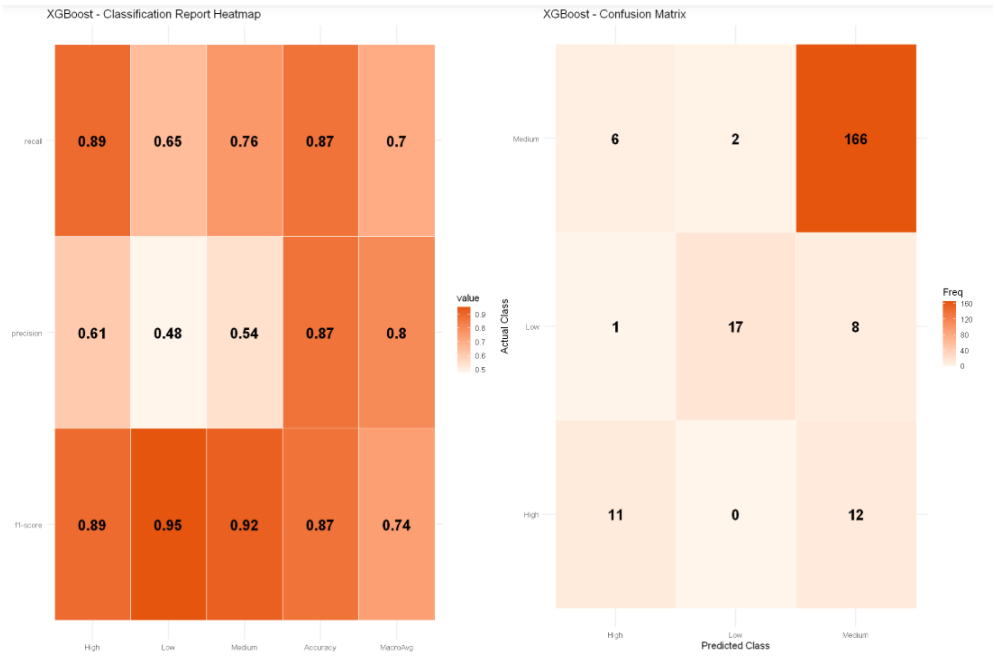


Figure 34 - XGBoost Heatmaps

Interpretation:

- XGBoost model performs well for ‘Medium’ and ‘High’ classes with high precision and recall. However, ‘Low’ class suffers from precision (0.48) and is often misclassified as Medium.
- This model struggles to distinguish ‘Low’ class from others. However, supporting the alternate hypothesis.

XGBoost is highly efficient for deployment in judicial dashboards or alert systems due to its scalability and tunability.

5.3 Classification Model Comparison

Rank	Model	Accuracy	Kappa	Sensitivity (High)	Precision (High)	Balanced Accuracy
1	**Random Forest**	88.3%	0.63	0.52	0.75	0.77
2	**XGBoost**	87%	0.60	0.47	0.61	0.78
3	**Logistic Regression**	84.3%	0.47	0.17	0.50	0.69
4	**SVM**	84.3%	0.41	0.00	NA	0.64

Table 3 - Classification Models Comparison

Interpretations:

- Random Forest is the best performer in both accuracy (88%) and class-wise balance. It handled non-linear patterns making it relevant for complex interactions between various conviction categories.
- XGBoost offered slightly lower accuracy 87% than random forest and excellent precision for high class. It is more tuneable and balance classes using boosting techniques.
- Logistic Regression remains standard for interpretability and ease of deployments with fair performance. However, it lacks flexibility to capture interactions in multi-dimensional data.
- SVM was the least effective model which completely failed to classify 'High' success category with poor kernel selection.

Overall, models supported the hypothesis and stated that it can significantly specify categories into low, medium and high classes enabling policymakers for resource allocation planning accordingly.

Business Insights:

The classification models using the hypothesis offer practical and impactful insights into legal system performance at a national or regional level. They can be deployed in:

- Public sector analytics dashboards, automatically flagging areas with poor excellent conviction outcomes.
- Risk-based policy making where governments can identify at-risk or underperforming jurisdictions and intervene early.

By incorporating data-driven modelling into legal performance measurement, my classification system transforms static records into actionable intelligence. This directly supports public accountability and justice reforms.

6. Critical Evaluation of Tools and Techniques

6.1 Critical Review of Regression Analysis

Effectiveness:

Regression Modelling was used to assess how offence-related features especially geography influenced conviction rates. Among the four models, Linear Regression, Decision Tree, Random Forest and Support Vector Regression – SVR demonstrated superior performance with R^2 of 0.95 and the lowest RMSE (0.009) indicating predictive accuracy. This suggests a strong fit for capturing non-linear and complex trends within conviction data. The model's robustness to high-dimensional features (Bansal *et al*., 2022) proves valuable in criminal justice datasets where multiple offence factors interrelate.

Limitations:

Linear regression, while interpretable and fast, failed to capture the non-linear nuances present in the dataset and inability to model outliers or skewed distributions. Similarly, decision tree models performed poorly due to their overfitting and high variance. These models also offered limited generalisability (Davies, 2021).

Alternative Solutions:

Although SVR and Random Forest performed well, the report could benefit from experimenting with Gradient Boosting Regression or ElasticNet Regression, which can provide better bias-variance trade-offs and are known for handling multicollinearity and high-dimensionality (Zou and Hastie, 2005). Moreover, cross-validation techniques such as k-fold validation could have been used to validate model robustness rather than a single train-test split.

Visual Tools:

Scatter plots were well applied for regression model interpretation. However, additional residual vs fitted plots or quantile-quantile (Q-Q) plots could have strengthened the diagnostic evaluation of assumptions (Paria *et al*., 2022).

Relevance to Dataset:

The regression models captured region-specific variations in legal outcomes, allowing the Crown Prosecution Service to benchmark judicial performance. However, more advanced interpretable models such as SHAP values could improve explainability in policy contexts.

6.2 Critical Review of Clustering Analysis

Effectiveness:

Clustering was used to identify patterns in offence outcomes without predefined labels. The hypothesis testing via Hopkins and Silhouette score indicated natural groupings in the data, affirming the modelling direction. K-means was the top-performing algorithm with a Calinski-Harabasz Index of 2517.47. The elbow method confirmed k=3 as optimal, showcasing solid cluster separation and compactness (Fraihat *et al.*, 2022).

Limitations:

Clustering results are sensitive to feature scaling and selection. Only two aggregated features total conviction and unsuccessful counts were used, which simplifies the model and may have missed finer granularity. GMM, though flexible in handling irregular clusters, performed the worst in silhouette scores (Hansen and Lee, 2021). It struggled due to overlapping distributions and potential overfitting in sparse clusters.

Alternative solutions:

Future studies could explore DBSCAN or spectral clustering, which are more resilient to noise and do not require predefined cluster members (Tang *et al.*, 2022). These are suitable for identifying anomalies such as outlier jurisdictions in legal systems.

Visual Tools:

The 2D scatter plots used for cluster visualisation were appropriate but limiting. Techniques like PCA or t-SNE could have been implemented for dimensionality reduction and better visual representation when clustering is applied to higher-dimensional data (Roy, 2024).

Relevance to Dataset:

Clustering revealed groupings such as high-performing versus low-performing crime cases. These are useful in segmenting jurisdictions for policy intervention. It also emphasizes system profiling or burden detection. However, the lack of temporal clustering like changes over months, limited time-based operational insights.

6.3 Critical Review of Classification Analysis

Effectiveness:

The classification hypothesis focused on categorising legal outcomes into ‘High’, ’Medium’ and ‘Low’ based on conviction rates and offence types. Models testing included Logistic Regression, Random Forest, SVM and XGBoost. Random Forest is the best model having an accuracy of 88.3% and kappa of 0.63 displaying strong class-wise balance and high sensitivity for all three categories(Ayala-Berdon *et al.*, 2020).

Limitations:

Even though Random Forest and XGBoost had high accuracy, it struggled to balance classes. SVM failed to detect any 'High' class instances, indicating poor handling of minority classes and inadequate kernel tuning. Logistic regression, though easy to interpret, lacked the flexibility to model complex boundaries leading to misclassification.

Furthermore, while some models performed well on accuracy, they lacked interpretability like SVM and XGBoost, despite strong results (Pande *et al.*, 2023). In jurisdiction systems, transparency is critical. The absence of interpretability tools like LIME or SHAP was a missed opportunity for in-depth model introspection.

Alternate solutions:

Applying ensemble voting classifiers, Naïve Bayes for categorical dominance or LightGBM could offer performance improvements while optimising speed and memory usage. Also, SMOTE could be used to tackle class imbalance, particularly by improving the detection of 'High' success cases (Chawla *et al.*, 2002).

Visual Tools:

Confusion matrices and classification heatmaps were effective and sufficiently interpretable. However, ROC-AUC curves and precision-recall plots would offer further performance insights, especially for imbalanced class scenarios (Miao and Zhu, 2021). Precision- recall is particularly relevant for justice systems where false positives like tagging areas vs low-performing can have severe consequences.

Relevance to Dataset:

The classification model's ability to predict legal success categories empowers CPS and policymakers to strategically allocate resources and attention. Early identification of 'Low' performance areas enables prompt policy intervention and resource allocation. However, greater emphasis on class balance and calibration like Platt scaling would have improved fairness in predictions.

6.4 Overall Evaluation of Tools and Techniques

Effectiveness of Tools:

Using R was an appropriate choice for data preprocessing and modelling due to its statistical depth, visualisation power via ggplot2, patchwork and machine learning packages like caret and xgboost (Giorgi *et al*., 2022). The use of standardisation, label encoding, log transformation and feature engineering was theoretically justified and applied effectively across all models.

Visualisation Techniques:

A strong variety of visuals was used across the study including line charts, heatmaps, patchwork views, correlation matrices, histograms, bar plots, scatter plots, decomposition plots and confusion matrices (Bhatia *et al.*, 2025). However, some visualisations like overlapping clustering plots could have been enhanced with dimensionality reduction techniques for better readability.

Alternative Considerations:

Python-based libraries like scikit-learn, seaborn or SHAP could have been considered for broader tool access and scalable deployment pipelines (Géron, 2022). Furthermore, AutoML platforms offer faster tuning and ensemble selection useful for complex systems like judicial analytics.

Critical Reflection:

While the model selection was diverse and justified, model explainability is vital in justice datasets. Incorporating ethical considerations such as fairness in model outcomes (Selbst *et al.*, 2019) and ensuring non-discriminatory patterns like regional bias would strengthen the analytical rigour. Additionally, validating models with cross-validation or external datasets could improve generalisability.

Conclusion

This project critically explored the application of data analytical techniques across descriptive, predictive and evaluation phases using criminal case outcome data. Through rigorous data integration, cleaning, feature engineering and advanced modelling, which extracted meaningful trends and relationships, notably in conviction rates across areas and offence types. Predictive modelling including regression, clustering and classification was evaluated using multiple algorithms, enhancing model comparison and validity. Tools such as Random Forest, XGBoost, SVM and K-means proved particularly effective supported by metrics like accuracy, RMSE and silhouette scores. Hypotheses were tested statistically, and models were interpreted both visually and contextually. Furthermore, ethical considerations, generalisability and model fairness were critically reviewed. This comprehensive pipeline demonstrates how data-driven insights can inform resource allocation, policy decisions and performance monitoring in the justice system. Overall, this report not only fulfils academic and technical expectations but also promotes responsible real-world applications of data science in sensitive domains.

References

Abdullah, D.M. and Abdulazeez, A.M. (2021) 'Machine Learning Applications based on SVM Classification A Review,' *Qubahan Academic Journal*, 1(2), pp. 81–90. https://doi.org/10.48161/qaj.v1n2a50.

Ali, M. and Razzaque, N. (2023) 'The Key Problems Facing Civil Justice Today Are Cost, Delay & Complexity: A Critical Review,' *Scholars International Journal of Law Crime and Justice*, 6(08), pp. 438–446. https://doi.org/10.36348/sijlcj.2023.v06i08.007.

Ayala-Berdon, J. *et al.* (2020) 'Random forest is the best species predictor for a community of insectivorous bats inhabiting a mountain ecosystem of central Mexico,' *Bioacoustics*, 30(5), pp. 608–628. https://doi.org/10.1080/09524622.2020.1835539.

Bansal, M., Goyal, A. and Choudhary, A. (2022) 'A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,' *Decision Analytics Journal*, 3, p. 100071. https://doi.org/10.1016/j.dajour.2022.100071.

Bhatia, N. et al. (2025) 'Advanced techniques for fusion data visualisation,' Frontiers in Physics, 13. https://doi.org/10.3389/fphy.2025.1569248.

Chawla, N.V. *et al.* (2002) 'SMOTE: Synthetic Minority Over-sampling technique,' *Journal of Artificial Intelligence Research*, 16, pp. 321–357. https://doi.org/10.1613/jair.953.

Choi, G. *et al.* (2022) 'Log-transformation of independent variables: Must we?,' *Epidemiology*, 33(6), pp. 843–853. https://doi.org/10.1097/ede.0000000000001534.

Curran, F.C., Carlo, S. and Harris-Walls, K. (2024) 'Making the Data Visible: A Systematic Review of Systems-Level Data Dashboards for leadership and Policy in Education,' *Review of Educational Research* [Preprint]. https://doi.org/10.3102/00346543241288249.

Darji, J. *et al.* (2024) 'Efficient use of binned data for imputing univariate time series data,' *Frontiers in Big Data*, 7. https://doi.org/10.3389/fdata.2024.1422650.

Das, A. (2023) 'Logistic regression,' in *Springer eBooks*, pp. 3985–3986. https://doi.org/10.1007/978-3-031-17299-1_1689.

Dash, R.K. *et al.* (2021) 'Fine-tuned support vector regression model for stock predictions,' *Neural Computing and Applications*, 35(32), pp. 23295–23309. https://doi.org/10.1007/s00521-021-05842-w.

Davies, L. (2021) 'Linear regression, covariate selection and the failure of modelling,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2112.08738.

Emerson, R.W. (2022) 'ANOVA assumptions,' *Journal of Visual Impairment & Blindness*, 116(4), pp. 585–586. https://doi.org/10.1177/0145482x221124187.

Fraihat, M. *et al.* (2022) 'An efficient enhanced k-means clustering algorithm for best offer prediction in telecom,' *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, 12(3), p. 2931. https://doi.org/10.11591/ijece.v12i3.pp2931-2943.

Garside, R. (2015) 'Criminal justice since 2010. What happened, and why?,' *Criminal Justice Matters*, 100(1), pp. 4–8. https://doi.org/10.1080/0268117x.2015.1061331.

Géron, A. (2022) *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/.

Ghasemi, A. and Zahediasl, S. (2012) 'Normality Tests for Statistical Analysis: A Guide for Non-Statisticians,' *International Journal of Endocrinology and Metabolism*, 10(2), pp. 486–489. https://doi.org/10.5812/ijem.3505.

Giorgi, F.M., Ceraolo, C. and Mercatelli, D. (2022) 'The R language: an engine for bioinformatics and data science,' Life, 12(5), p. 648. https://doi.org/10.3390/life12050648.

Han, X. *et al.* (2025) 'Multimodal Spatio-Temporal Data Visualization Technologies for Contemporary Urban Landscape Architecture: A Review and Prospect in the context of Smart Cities,' *Land*, 14(5), p. 1069. https://doi.org/10.3390/land14051069.

Hansen, B.E. and Lee, S. (2021) 'Inference for iterated GMM under misspecification,' *Econometrica*, 89(3), pp. 1419–1447. https://doi.org/10.3982/ecta16274.

Idreos, S., Papaemmanouil, O. and Chaudhuri, S. (2015) 'Overview of Data Exploration Techniques,' *Google Scholar*, pp. 277–281. https://doi.org/10.1145/2723372.2731084.

Ikotun, A.M. *et al.* (2022) 'K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,' *Information Sciences*, 622, pp. 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

James, G. *et al.* (2023) 'Linear regression,' in *Springer texts in statistics*, pp. 69–134. https://doi.org/10.1007/978-3-031-38747-0_3.

Kaeseler, R.L. et al. (2022) 'Feature and Classification Analysis for Detection and Classification of tongue movements from Single-Trial Pre-Movement EEG,' IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, pp. 678–687. https://doi.org/10.1109/tnsre.2022.3157959.

Karch, J.D. (2023) 'Outliers may not be automatically removed.,' *Journal of Experimental Psychology General*, 152(6), pp. 1735–1753. https://doi.org/10.1037/xge0001357.

Karim, R., Alam, M.K. and Hossain, M.R. (2021) 'Stock Market Analysis Using Linear Regression and Decision Tree Regression,' *IEEE Xplore* [Preprint]. https://doi.org/10.1109/esmarta52612.2021.9515762.

Krake, T. *et al.* (2024) 'Uncertainty-Aware Seasonal-Trend decomposition based on loess,' *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16. https://doi.org/10.1109/tvcg.2024.3364388.

Legal Aid Agency (2015) 'Crime news: national rollout for Crown Court Digital Case System,' *GOV.UK*, 3 December. https://www.gov.uk/government/news/crime-news-national-rollout-for-crown-court-digital-case-system?utm_source=chatgpt.com.

Li, P. *et al.* (2021) 'CleanML: A study for evaluating the impact of data cleaning on ML classification tasks,' *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 13–24. https://doi.org/10.1109/icde51399.2021.00009.

Li, T., Rezaeipanah, A. and Din, E.M.T.E. (2022) 'An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement,' *Journal of King Saud University - Computer and Information Sciences*, 34(6), pp. 3828–3842. https://doi.org/10.1016/j.jksuci.2022.04.010.

Liang, Y. *et al.* (2024) 'Foundation Models for Time Series Analysis: A tutorial and survey,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.1145/3637528.3671451.

Miao, J. and Zhu, W. (2021) 'Precision–recall curve (PRC) classification trees,' *Evolutionary Intelligence*, 15(3), pp. 1545–1569. https://doi.org/10.1007/s12065-021-00565-2.

Ministry of Justice (2015) 'Criminal Justice Management 2015,' *GOV.UK*, 23 September. https://www.gov.uk/government/speeches/criminal-justice-management-2015?utm_source=chatgpt.com.

N, G. *et al.* (2021) 'Random Forest Regression-Based Machine Learning model for accurate estimation of fluid flow in curved pipes,' *Processes*, 9(11), p. 2095. https://doi.org/10.3390/pr9112095.

Nargesian, F., Asudeh, A. and Jagadish, H.V. (2022) 'Responsible Data Integration: next-generation challenges,' *Proceedings of the 2022 International Conference on Management of Data*, pp. 2458–2464. https://doi.org/10.1145/3514221.3522567.

Naseer, A. and Jalal, A. (2024) 'Multimodal Objects Categorization by Fusing GMM and Multi-layer Perceptron,' *IEEE Xplore*, pp. 1–7. https://doi.org/10.1109/icacs60934.2024.10473242.

Oyewole, G.J. and Thopil, G.A. (2022) 'Data clustering: application and trends,' *Artificial Intelligence Review*, 56(7), pp. 6439–6475. https://doi.org/10.1007/s10462-022-10325-y.

Pande, C.B. *et al.* (2023) 'Comparative Assessment of Improved SVM Method under Different Kernel Functions for Predicting Multi-scale Drought Index,' *Water Resources Management*, 37(3), pp. 1367–1399. https://doi.org/10.1007/s11269-023-03440-0. R libraries

Paria, S.S., Rahman, S.R. and Adhikari, K. (2022) 'fastman: A fast algorithm for visualizing GWAS results using Manhattan and Q-Q plots,' *bioRxiv (Cold Spring Harbor Laboratory)* [Preprint]. https://doi.org/10.1101/2022.04.19.488738.

Pérez-Cova, M. *et al.* (2022) 'Comparison of multivariate ANOVA-Based approaches for the determination of relevant variables in experimentally designed metabolomic studies,' *Molecules*, 27(10), p. 3304. https://doi.org/10.3390/molecules27103304.

Ran, X. *et al.* (2022) 'Comprehensive survey on hierarchical clustering algorithms and the recent developments,' *Artificial Intelligence Review*, 56(8), pp. 8219–8264. https://doi.org/10.1007/s10462-022-10366-3.

Rodosthenous, T., Shahrezaei, V. and Evangelou, M. (2024) 'Multi-view data visualisation via manifold learning,' *PeerJ Computer Science*, 10, p. e1993. https://doi.org/10.7717/peerj-cs.1993.

Roy, A. (2024) 'Data visualisation to understand how data is structured using K-Means and hierarchical cluster analyses with interactive graphics,' *IOSR Journal of Applied Geology and Geophysic*, 12(6), pp. 01–05. https://doi.org/10.9790/0990-1206010105.

Sagala, N.T.M. and Gunawan, A.A.S. (2022) 'Discovering the optimal number of crime cluster using elbow, silhouette, gap statistics, and NBClust methods,' *ComTech Computer Mathematics and Engineering Applications*, 13(1), pp. 1–10. https://doi.org/10.21512/comtech.v13i1.7270.

Sahin, E.K. (2022) 'Implementation of free and open-source semi-automatic feature engineering tool in landslide susceptibility mapping using the machine-learning algorithms RF, SVM, and XGBoost,' *Stochastic Environmental Research and Risk Assessment*, 37(3), pp. 1067–1092. https://doi.org/10.1007/s00477-022-02330-y.

Salman, H.A., Kalakech, A. and Steiti, A. (2024) 'Random Forest algorithm Overview,' *Deleted Journal*, 2024, pp. 69–79. https://doi.org/10.58496/bjml/2024/007.

Saraswat, P. and Raj, S. (2022) 'DATA PRE-PROCESSING TECHNIQUES IN DATA MINING: a REVIEW,' *International Journal of Innovative Research in Computer Science & Technology*, pp. 122–125. https://doi.org/10.55524/ijircst.2022.10.1.22.

Sekeroglu, B. *et al.* (2022) 'Comparative evaluation and comprehensive analysis of machine learning models for regression problems,' *Data Intelligence*, 4(3), pp. 620–652. https://doi.org/10.1162/dint_a_00155.

Selbst, A.D. *et al.* (2019) 'Fairness and Abstraction in Sociotechnical Systems,' *ACM DL*, pp. 59–68. https://doi.org/10.1145/3287560.3287598.

Shen, C., Panda, S. and Vogelstein, J.T. (2021) 'The Chi-Square test of distance correlation,' *Journal of Computational and Graphical Statistics*, 31(1), pp. 254–262. https://doi.org/10.1080/10618600.2021.1938585.

Tang, C. *et al.* (2022) 'Unified One-Step Multi-View spectral clustering,' *IEEE Transactions on Knowledge and Data Engineering*, 35(6), pp. 6449–6460. https://doi.org/10.1109/tkde.2022.3172687.

Tariq, S. and Rana, T.A. (2024) 'Automatic regex synthesis methods for english: a comparative analysis,' *Knowledge and Information Systems* [Preprint]. https://doi.org/10.1007/s10115-024-02232-1.

Tyagi, K. *et al.* (2022) 'Regression analysis,' in *Elsevier eBooks*, pp. 53–63. https://doi.org/10.1016/b978-0-12-824054-0.00007-1.

Wildeman, C. and Sampson, R.J. (2023) 'Desistance as an intergenerational process,' *Annual Review of Criminology*, 7(1), pp. 85–104. https://doi.org/10.1146/annurev-criminol-022422-015936.

Wilkinson, L. and Friendly, M. (2009) 'The history of the Cluster Heat Map,' *The American Statistician*, 63(2), pp. 179–184. https://doi.org/10.1198/tas.2009.0033.

Xia, Y. and Zhou, X. (2024) 'Improving the use of parallel analysis by accounting for sampling variability of the observed correlation matrix,' *Educational and Psychological Measurement*, 85(1), pp. 114–133. https://doi.org/10.1177/00131644241268073.

Zhang, K. *et al.* (2024) 'Self-Supervised Learning for Time Series Analysis: Taxonomy, progress, and Prospects,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10), pp. 6775–6794. https://doi.org/10.1109/tpami.2024.3387317.

Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the elastic net,' *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2), pp. 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.