A close-up of a logo

Description automatically generated

**ANALYSIS OF PRINCIPAL CRIMINAL OFFENCES IN THE UNITED KINGDOM**

**ABDURAHEEM ABDUSALAM TEMITOPE – S4210507**

MODULE TITLE: CT7202 DATA ANALYSIS AND VISUALISATION PRINCIPLES

MODULE TUTOR: DR. BHUPESH MISHRA

SCHOOL OF COMPUTING AND TECHNOLOGY

UNIVERSITY OF GLOUCESTERSHIRE

Table of Contents

[1.1 Principal Criminal Offenses 3](#_Toc156383190)

[1.3 Offenses in The UK 4](#_Toc156383191)

[1. Dataset 5](#_Toc156383192)

[2. Data Exploration 7](#_Toc156383193)

[3. Hypotheses Definition 10](#_Toc156383194)

[4. Data Cleaning and Integration Techniques Used 11](#_Toc156383195)

[5. Descriptive Statistics 12](#_Toc156383196)

[6. Analysis and Interpretation of Results 22](#_Toc156383197)

[Exploratory Analysis 22](#_Toc156383198)

[Correlation and Regression Analysis 36](#_Toc156383199)

[Clustering Analysis 39](#_Toc156383200)

[Hypothesis Test Results 41](#_Toc156383201)

[Hypothesis 1 Conclusion 42](#_Toc156383202)

[Hypothesis 2 Conclusion 42](#_Toc156383203)

[7. Critical Review of Data Analytics Tools and Techniques Used 43](#_Toc156383204)

[Regression Analysis and Correlation Analysis 43](#_Toc156383205)

[Exploratory and Clustering Analysis 43](#_Toc156383206)

[Data Manipulation Tools 43](#_Toc156383207)

[Predictive Modeling Tools 43](#_Toc156383208)

[8. Critical Review of Data Visualization Tools and Techniques Used 45](#_Toc156383209)

**References**

1. Introduction

## 1.1 Principal Criminal Offenses

Criminal offenses form an extensive and intricate landscape within the realm of legality, encapsulating a wide spectrum of actions that contravene established laws and are subject to punitive measures. This expansive domain encompasses a diverse array of transgressions, each varying in severity, societal impact, and legal repercussions. Ranging from minor infractions like petty theft, where the value of the stolen property is minimal, to heinous felonies like homicide, the tapestry of criminal offenses traverses a complex labyrinth that demands a nuanced understanding of its intricacies.

At the core of comprehending the multifaceted nature of criminal offenses lies the exploration and analysis of their categories, elements, and consequential legal outcomes. These offenses are not merely isolated actions but are embedded within a matrix of legal frameworks, societal norms, and individual responsibilities. To unravel this intricate fabric, it becomes imperative to delve deep into their categorizations, dissect the constituent elements that define them, and comprehend the extensive legal ramifications that ensue from their commission.

The legal consequences tied to criminal offenses exhibit a vast spectrum, heavily contingent on their severity and the jurisprudence prevailing in a given jurisdiction. Penalties range from nominal fines, probation, or community service for minor infractions or misdemeanors, aiming at restorative justice or correction, to the most severe forms of punishment such as imprisonment or even capital punishment for grave felonies like murder. Beyond these direct penalties, collateral consequences such as the loss of certain civil rights or enduring social stigma often accompany a criminal conviction, impacting an individual's life beyond the judicial sentence.

However, these offenses, with their varied legal implications, extend beyond mere classification and repercussions. They embody a complex interplay of legal tenets, societal norms, and individual responsibility. Understanding these offenses in their entirety necessitates a comprehensive exploration that goes beyond surface-level categorizations. It requires a meticulous dissection of their constituent elements and an astute comprehension of the profound legal, social, and ethical ramifications they entail.

This comprehensive understanding forms the bedrock for legal practitioners, policymakers, and society as a whole in their collective endeavor to uphold justice, maintain societal order, and safeguard the rights of individuals within the purview of a robust and fair legal framework. Such insights derived from a deep exploration of criminal offenses empower stakeholders to make informed decisions, shape policies, and foster a more just and equitable society.

## 1.3 Offenses in The UK

In the United Kingdom, principal criminal offenses form the cornerstone of legal actions pursued by law enforcement agencies and the Crown Prosecution Service (CPS). Central to the UK's criminal justice system, the CPS shoulders the responsibility of prosecuting criminal cases investigated by various law enforcement entities. Their role extends from reviewing cases submitted by the police to making informed decisions regarding whether a case should proceed to court based on available evidence and the public interest.

The recorded data on principal criminal offenses, meticulously cataloged by the CPS, presents a comprehensive monthly breakdown of case outcomes. This repository of information categorizes outcomes based on the principal offense category, further segmented by the CPS Area. This granular classification offers a detailed panorama of how criminal cases progress and conclude in various geographic regions across the UK.

Within this monthly breakdown lie invaluable insights into the nature of offenses prosecuted, the diverse outcomes stemming from these cases, and the geographic distribution of legal proceedings. Such comprehensive data serves as a reservoir for understanding the complex dynamics of criminal prosecutions, showcasing patterns, trends, and potential disparities across different offense categories and regions.

For policymakers, this data acts as a compass, guiding the formulation of strategies and policies aimed at tackling specific crime patterns prevalent in different areas. By identifying trends in prosecution outcomes or the prevalence of certain offenses in specific regions, policymakers can tailor interventions and allocate resources more effectively to address underlying societal issues contributing to criminal activities.

Similarly, researchers find immense value in this trove of data. Analyzing long-term trends or conducting comparative studies across regions aids in unveiling underlying sociodemographic or economic factors influencing crime rates. Such analyses facilitate evidence-based research that delves into the effectiveness of different legal approaches or social interventions in curbing specific types of criminal activities.

Moreover, for law enforcement agencies, this data serves as a vital tool for resource allocation, strategic planning, and understanding localized crime dynamics. It aids in optimizing efforts, directing resources to areas experiencing surges in specific offenses, or refining investigative approaches based on observed case outcomes.

In essence, the meticulous recording and categorization of principal criminal offenses by the CPS serve as a critical cornerstone in understanding the multifaceted landscape of criminal prosecutions in the UK. This data-driven approach offers insights that not only inform legal proceedings but also empower stakeholders across various domains to make informed decisions, ultimately contributing to the pursuit of a more just and secure society.

# Dataset

Data from 24 months (2016-2018) were downloaded and merged. After the data had been merged from the downloaded files, data exploration was conducted to help clarify what data were in the final merged file and how they would be used in the analysis. From the visualization, there are some missing data represented using -. Also, columns with percentages in the name tend to have a % sign in the value. This could hinder the ability to perform analysis and is addressed in the data-cleaning phase of this report.

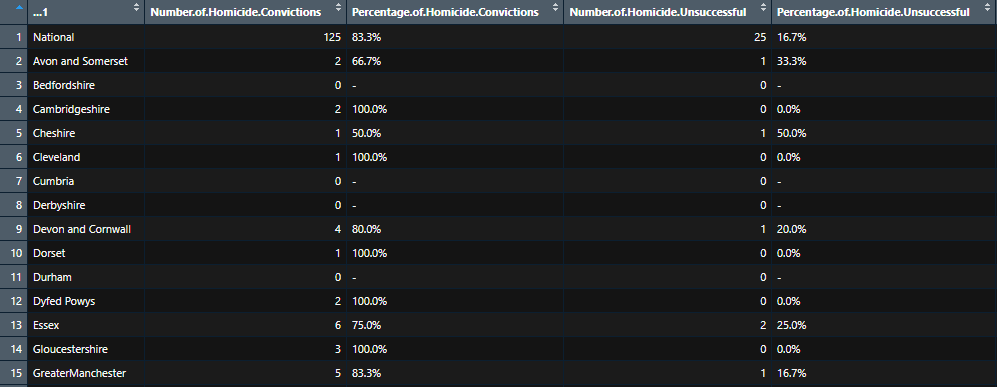


Figure 1. Dataset overview

The dataset had 1204 rows and 51 columns. This was shown through using the glimpse() function and str() function.

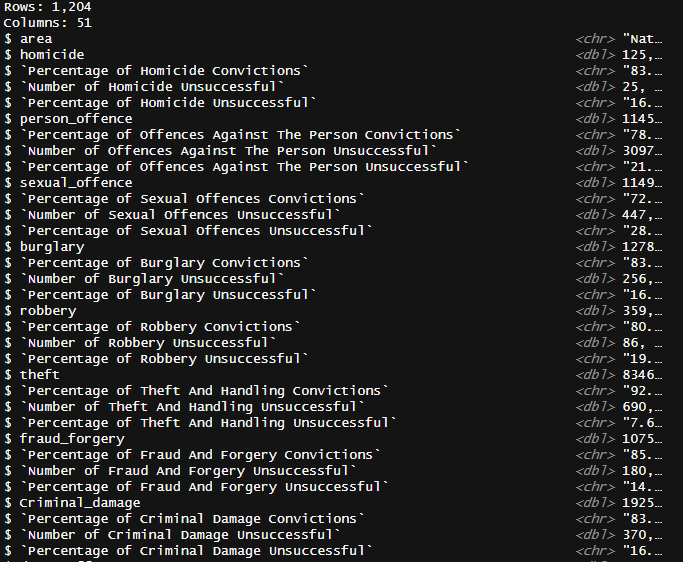


Figure 2. Column names and types

# Data Exploration

Using sum(is.na()), it was shown that there were 2193 missing data in all. The visualization (figure 3) showed where these missing data were mostly located. In all missing data formed 1.9% of the total dataset.

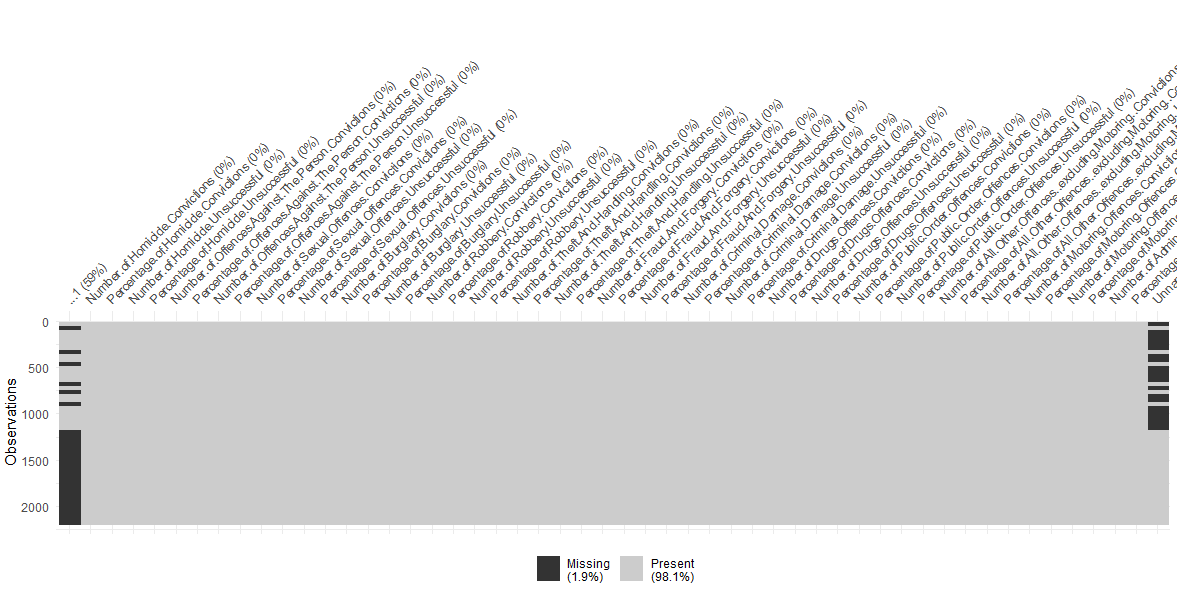


Figure 3. Missing data

The missing data also tend to appear under columns which had character values.

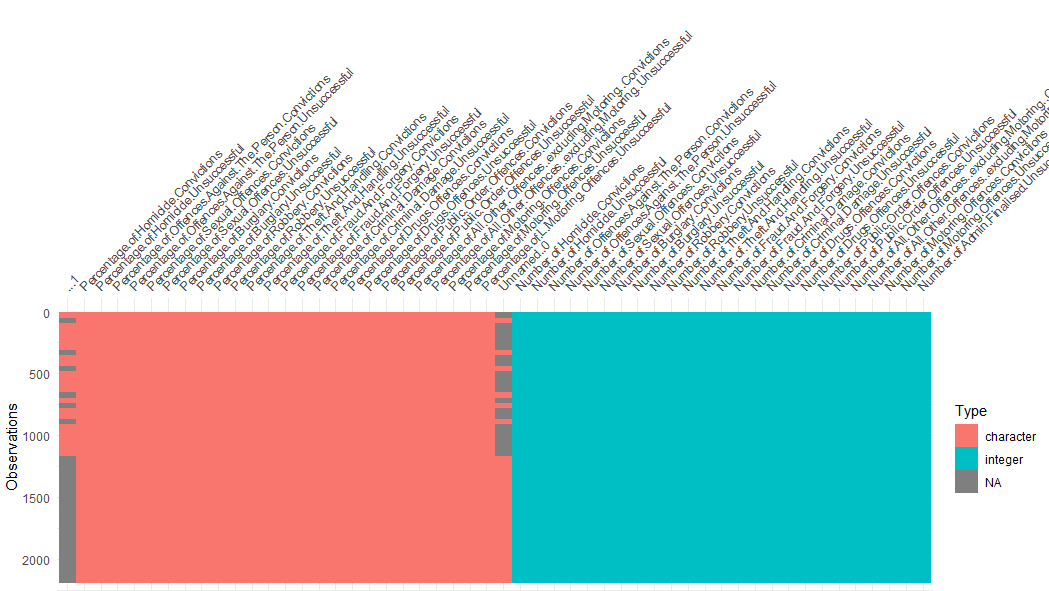


Figure 4. visualizing missing data

The monthly breakdown of criminal case outcomes, meticulously cataloged by the Crown Prosecution Service (CPS), stands as an invaluable repository of data encapsulating the multifaceted dimensions of the UK's criminal justice landscape. This detailed breakdown isn't merely a compendium of legal proceedings; rather, it serves as a comprehensive chronicle, capturing the ebbs and flows of criminal prosecutions across diverse offense categories and geographical regions.

Each entry within this meticulously crafted dataset offers a panoramic view of the intricate legal journey undertaken by criminal cases. It delineates the outcomes arising from legal proceedings, categorizing them based on the principal offense category and further segmenting them by the CPS Area. This granular classification doesn't merely signify statistical data; rather, it forms a tapestry of insights into the prosecution and resolution of myriad criminal cases within the UK's diverse sociocultural and geographic landscapes.

This data isn't just a numerical representation; it's a reflection of the multifarious legal deliberations, the intricate interplay of evidentiary aspects, and the pursuit of justice in its varied forms. It encapsulates the triumphs and challenges faced by the judicial system, portraying the intricate complexities inherent in prosecuting offenses of distinct natures, magnitudes, and societal impacts.

At its core, this data repository isn't a static collection of figures but a dynamic canvas that unravels patterns, trends, and potential disparities existing within the realm of criminal prosecutions. It's an evolving narrative, offering invaluable insights into the nuances of the legal process, the societal dynamics influencing criminal proceedings, and the regional variations shaping prosecutorial outcomes.

Moreover, beyond its statistical significance, this data serves as a compass guiding policymakers, researchers, and legal practitioners alike. It empowers stakeholders to discern patterns, identify systemic issues, and forge evidence-based strategies aimed at enhancing the efficacy and fairness of the criminal justice system. By analyzing the patterns emerging from this dataset, policymakers gain insights crucial for informed decision-making, researchers unearth trends pivotal for evidence-based studies, and legal practitioners derive nuanced understandings vital for improving legal strategies and approaches.

In essence, this monthly breakdown of criminal case outcomes, meticulously dissected by principal offense categories and CPS Areas, doesn't merely represent statistics; it embodies a narrative of justice, societal intricacies, and the tireless pursuit of fairness within the contours of the UK's criminal justice system.

# Hypotheses Definition

# 

Hypothesis 1

Null Hypothesis. In this case, there won’t be any connection between being convicted of robbery and being convicted of murder.

Alternative Hypothesis. In this case, there will be a correlation between robbery convictions and homicide convictions. Decision Rule.

Do not accept the hypothesis if the p value is less than α.

Do not reject the hypothesis if the p-value is greater than α.

Hypothesis 2

Null Hypothesis. In this case, there will be a correlation between being convicted of an offense and being convicted of homicide.

Alternate Hypothesis. Convictions for homicide and sexual offenses will show a correlation.

Conclusion

Reject the hypothesis if the p value is less than α.

Do not reject the hypothesis if the p value is greater, than α.

# Data Cleaning and Integration Techniques Used

The approach is done in two phases, integration and cleaning. First was the integration of the individual monthly datasets and joining to become one, and the cleaning is removal of unwanted characters in the dataset.

Data integration is a crucial step in the data preparation process, especially when dealing with datasets from multiple sources or time periods. Using R programming to read datasets directly from different sources and joining them together is a robust technique that ensures seamless integration. This approach is supported by studies emphasizing the significance of effective data integration in enhancing the quality and usability of datasets.

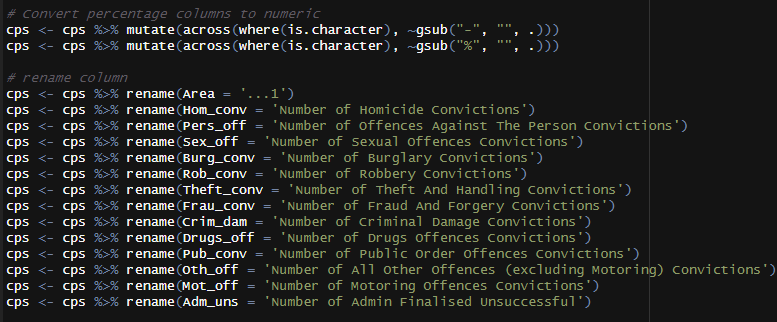


Figure 5. Data cleaning

Data cleaning involves the identification and correction of errors or inconsistencies in a dataset. In the case of removing the '%' sign and the '-' sign from values, this process is crucial for preparing data for analysis. Employing R programming for cleaning is a common practice, and the specific techniques used align with established principles in data cleaning.

# Descriptive Statistics

**Homicide Conviction**

Using base R, the distribution of homicide (figure 6) was gave a summarise picture of the homicide conviction data.

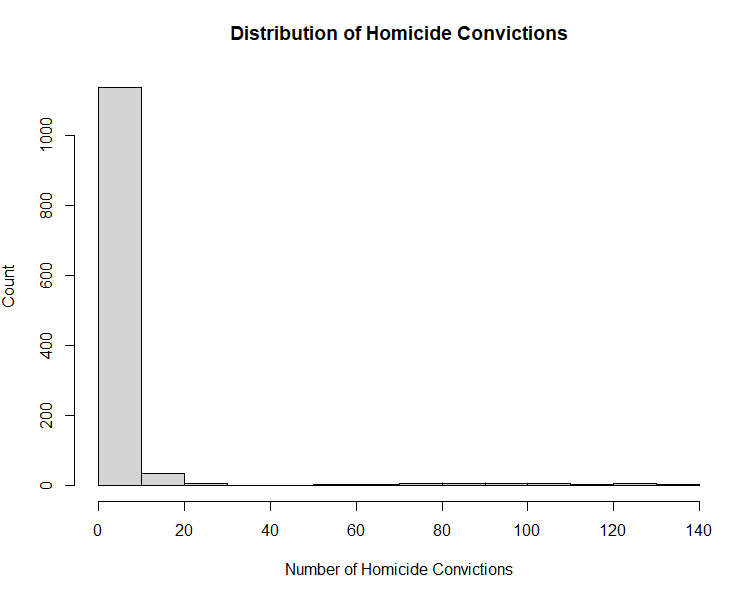


Figure 6. Homicide Distribution Plot

Figure 6 presents a clear visual representation of the frequency distribution of homicide convictions across different jurisdictions. The most striking feature of this histogram is its pronounced right skew, indicating that a large number of areas have a low number of convictions, while a small number of areas have high convictions.

Along the right of the x-axis, the bars rapidly diminish in height. This shows that areas with a moderate to high number of convictions are substantially less common. The absence of bars past a certain point on the x-axis indicates there's an upper limit to the number of convictions reported, suggesting a ceiling effect or a maximum threshold for the dataset.

This distribution raises several points for consideration. The concentration of areas with lower conviction counts could point to widespread low crime rates or could reflect variances in law enforcement effectiveness, reporting practices, or judicial processes across areas. The few areas with higher conviction rates may warrant additional investigation to understand the contributing factors, such as socio-economic conditions, law enforcement policies, or population density.

**Offences against the person**

A graph of a number of persons offense cases

Description automatically generated

Fig 7. Distribution of offences against the person Convictions

Figure 7 outlines the frequency distribution of convictions for offenses against individuals across various regions or entities in the UK. It provides a visual representation of the data pertaining to personal offence convictions, such as assault or battery. At the x-axis, the number of personal offense convictions are represented, ranging from 0 to 12,000 with the y-axis denoting the count of the frequency of these conviction numbers across the dataset.

The most prominent feature of this histogram is the large initial bar, which indicates that a vast majority of the entities accounted for in the dataset have a relatively low number of convictions, near zero. The frequency quickly tapers off as the number of convictions rises, as evidenced by the shorter bars proceeding to the right.

This distribution is heavily right-skewed, showing that while most entities have a small number of convictions, there are a few entities with very high numbers of convictions. This skewness could indicate that personal offences are concentrated in a small number of regions, or perhaps that the reporting or enforcement intensity varies greatly between different areas.

This visualization signals areas of interest for deeper investigation. Regions corresponding to the high-conviction bars may be experiencing higher crime rates, or they might have more robust law enforcement and reporting mechanisms, leading to a higher number of recorded convictions.

Further analysis might involve cross-referencing these regions with socio-economic data, population density, law enforcement policies, or other relevant datasets to discern the underlying factors contributing to the skewed distribution. This data could also be valuable for allocating resources for crime prevention or judicial processes, as it highlights areas with potentially higher needs for intervention.

**Sexual Offences Conviction**

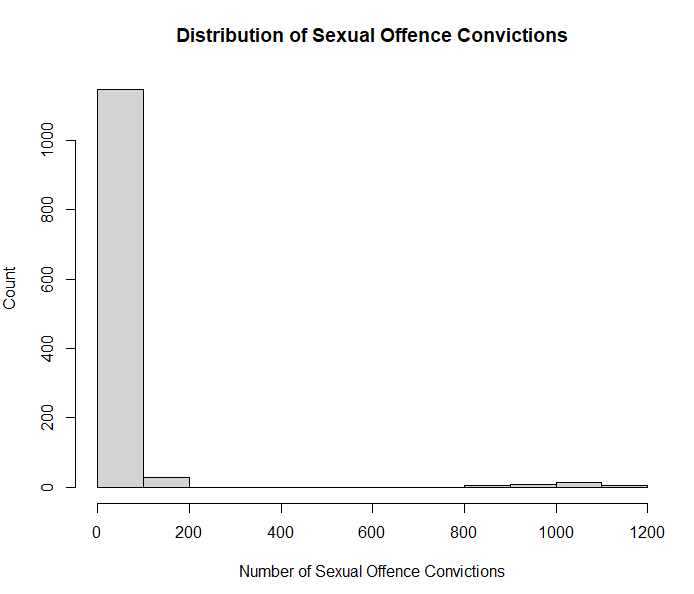


Figure 8. Sexual offences distribution

Figure 8 shows the distribution of sexual offence convictions across various jurisdictions. The x-axis of the histogram shows the range of sexual offence convictions, starting at 0 and extending past 1000. The y-axis displays the count or frequency of entities falling within each bin or range of convictions.

From the chart, the following can be deduced:

1. Skewness: The distribution is right-skewed with a very tall first bar and progressively shorter bars as the number of convictions increases. This suggests that the vast majority of entities have a low number of sexual offence convictions, while a small number have higher counts.

2. Frequency: The height of the first bar indicates that the highest frequency of entities has between 0 to 200 sexual offense convictions. This bar's prominence implies that such convictions are relatively uncommon occurrences in most entities.

3. Declining Numbers: The number of entities with higher counts of convictions drops sharply, as seen by the rapid decline in bar height moving rightward along the x-axis. This is indicative of fewer entities with a substantial number of sexual offense convictions.

4. Outliers: The presence of bars at higher conviction numbers, although significantly lower in height, suggests that there are outliers with unusually high numbers of convictions. These could be regions with higher population densities or those with more active or effective reporting and judicial processes.

The skewness is indicative of disparities in the distribution of sexual offense convictions across entities. Such disparities could be due to a variety of factors including, but not limited to, regional population differences, varying law enforcement effectiveness, differing societal attitudes towards reporting such crimes, or possibly inconsistencies in data collection methods.

This visualization serves as extra support for the analysis of the distribution for resource allocation and the need for targeted intervention in areas with higher conviction rates, and the potential for underreporting in regions with suspiciously low rates of convictions. It is also possible to explore correlations with other datasets to gain a deeper understanding of the context for these convictions, such as socioeconomic factors, education levels, and regional support services for victims.

A graph of a number of criminals

Description automatically generated

Figure 9. Robbery convictions distribution

Figure 9 displays the frequency of robbery convictions across various entities, which could be interpreted as jurisdictions, court districts, or perhaps police precincts.

Key observations from the histogram are:

1. Concentration of Low Conviction Counts: The x-axis, which details the number of robbery convictions, shows that the overwhelming majority of entities report a very low number of robbery convictions, within the 0 to 50 range, as depicted by the first and tallest bar.

2. Right-Skewed Distribution: The histogram is right-skewed, meaning there are fewer entities with a higher number of robbery convictions. This is evident from the rapidly decreasing height of the bars as we move right on the x-axis.

3. Frequency Decline: The frequency of entities with higher conviction counts declines steeply after the first interval. The bars representing the number of entities with 50 to 100, 100 to 150, and so on, robbery convictions, are much shorter, indicating these are less common.

4. Sparse High Conviction Counts: The flat, elongated tail of the histogram indicates that while entities with high conviction counts (e.g., over 200) exist, they are rare.

The following can be deduced from this chart:

Commonality of Few Convictions: The data suggests that most entities experience few robbery convictions, which might reflect effective crime prevention and law enforcement strategies, or it might indicate a lower propensity for such crimes in those areas.

Analysis of Outliers: The entities corresponding to the bars on the right might require further analysis to understand the reasons behind the higher number of convictions. Factors could include socioeconomic conditions, urbanization, policing strategies, or reporting practices.

Policy Implications: The distribution could inform policymakers about where to allocate resources for crime prevention, law enforcement, and community support, focusing on the areas with higher robbery rates.

Further Contextual Analysis: For a comprehensive understanding, this data should be analyzed in the context of other socioeconomic variables. It might also be useful to compare it with population data to assess if the rate of convictions correlates with population density.

The skewed nature of the data suggests that robbery, much like other specific crime categories, is not uniformly distributed across the dataset's entities. This could be valuable information for targeting interventions and understanding crime patterns.

**A graph of a number of criminal damage cases

Description automatically generated**

Figure 10. Criminal Damages Convictions Distribution

Figure 10 depicts the distribution of criminal damage convictions. It is a right-skewed distribution; the modal class is the interval closest to zero, indicating a high frequency of entities with minimal recorded convictions. A closer look at the x-axis reveals a range that extends past 2000 convictions, yet there's a pronounced scarcity of entities with counts exceeding even the lower hundreds. This kind of distribution is characteristic of datasets where a particular event - criminal damage, in this case - is infrequent or its detection and subsequent conviction is inconsistently applied across the sampling frame.

Notably, the y-axis, which tops out at 1200, suggests a significant number of entities with zero to negligible convictions, after which the frequency drops. This precipitous drop-off warrants a discussion on the underlying causes - whether it's due to a variance in enforcement efficacy, reporting practices, or indeed a reflection of the actual incidence rates of such crimes.

Given the nature of criminal damage as a crime category, which typically has a lower incidence rate compared to petty theft or vandalism, the distribution may not be unexpected. However, the absence of entities with high conviction counts could also suggest a ceiling effect where, beyond a certain point, additional convictions are not being captured or reported in the data.

The implications of such a distribution are varied. It is essential to consider the robustness of the data collection mechanisms and the uniformity of legal processes across the surveyed areas. Furthermore, the data could be cross-referenced with socio-economic factors to determine if there's a correlation that might explain the lower incidence in most entities and the outlier cases.

**A graph of a distribution of theft cases

Description automatically generated**

Figure 11. Theft convictions distribution

Figure 11 visualizes theft conviction frequencies across different regions. The initial observation reveals a heavy concentration of units with low conviction counts, as evidenced by the predominant initial bin. This suggests that, while theft may be pervasive as a crime category, it typically manifests at relatively lower frequencies within the vast majority of these units.

The distribution's rightward skew is indicative of the rapid decline in the number of units with increasing conviction counts. Units with higher numbers of theft convictions, particularly those exceeding the median range of the x-axis, are markedly less common, highlighting a significant variability in theft incidences across the dataset.

For a data scientist analyzing patterns of crime within the UK, such a distribution could be telling. It implies a potential focus for resource allocation: areas with higher conviction counts might benefit from enhanced crime prevention and law enforcement efforts. However, it is also critical to consider that the low conviction numbers in some units could be reflective of underreporting.

The implications of this distribution are substantial, necessitating further research into the underlying factors that contribute to this skewness. Such factors may include, but are not limited to, socioeconomic variables, population densities, and the effectiveness of local law enforcement. Moreover, a longitudinal analysis could provide insights into whether these are enduring trends or if there are developing pockets of increased theft convictions.

In summary, this histogram acts as a macro-level indicator, prompting a more nuanced statistical exploration to guide policy and strategic decision-making within the realm of crime prevention and law enforcement.

# Analysis and Interpretation of Results

The datasets, after collection and integration were prepared by cleaning the data and replacing unnecessary characters with empty string values. This cleaning was done on the entire dataset, and was done in two phases

1. Removal of the % sign and the - sign
2. The removal of commas from numbers going into the thousands

These cleanings were done to ensure the data was in the right format for analysis.

Descriptive analysis was carried out on the dataset to explore the dataset and extract insights that are useful for analysis. The analysis performed was implemented using the R programming language, and the different analysis techniques used were

* Exploratory Analysis
* Correlation and Regression Analysis
* Clustering Analysis

These analyses will be discussed and the interpretation of the results will be provided including code sources to back them up.

## Exploratory Analysis

The crime rates across different areas were examined using exploratory analysis. This is visually represented in the charts below.



Figure 12. Homicide conviction across areas

Figure 12 presents a distribution of conviction counts across various units of analysis, which represent distinct geographical regions. The most striking feature of this distribution is the prominent peak at the lower end of the conviction count spectrum, which rapidly tapers off as conviction counts increase. This suggests that incidents leading to criminal damage convictions are predominantly low across the majority of the units surveyed, with high conviction counts being an exception rather than the norm.

More so, the distribution exhibits a pronounced right skew, with the tail extending towards the higher end of the conviction count axis. Such a skew indicates that while most units have fewer convictions, a small number have disproportionately more, although they are still relatively rare. This pattern could be reflective of varying enforcement intensities, reporting rates, or actual incidence rates of criminal damage across different areas.

This visualization serves as a preliminary step in identifying areas with outlier conviction rates that may warrant further investigation. It also prompts questions regarding the underlying factors contributing to the observed distribution, such as socio-economic disparities or differences in judicial processing. These insights are foundational for allocating resources effectively and tailoring intervention strategies to address criminal damage within the context of a broader criminological framework.

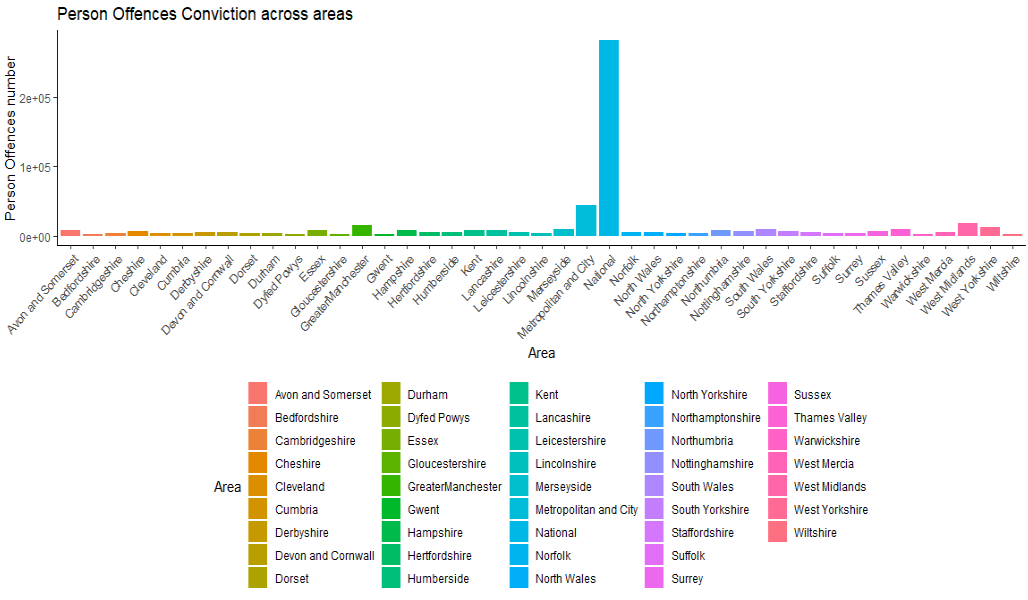


Figure 13. Person offense convictions across areas

Figure 13 presents a comparative analysis of personal offense convictions across different geographical or administrative regions. A notable observation from the chart is the considerable variability in conviction numbers across the depicted areas. A majority of the areas are characterized by lower conviction rates, yet an outlier with a significantly elevated conviction count is evident. The utilization of a logarithmic scale on the y-axis suggests a broad variance in conviction numbers, with the potential for these numbers to span multiple orders of magnitude.

The outliers highlighted in the chart – National, Metropolitan, and City – have a disproportionately high number of convictions for personal offenses. This is most likely due to the way the data was collected and organized. Thus, national and citywide data tend to have higher values compared to other cells.

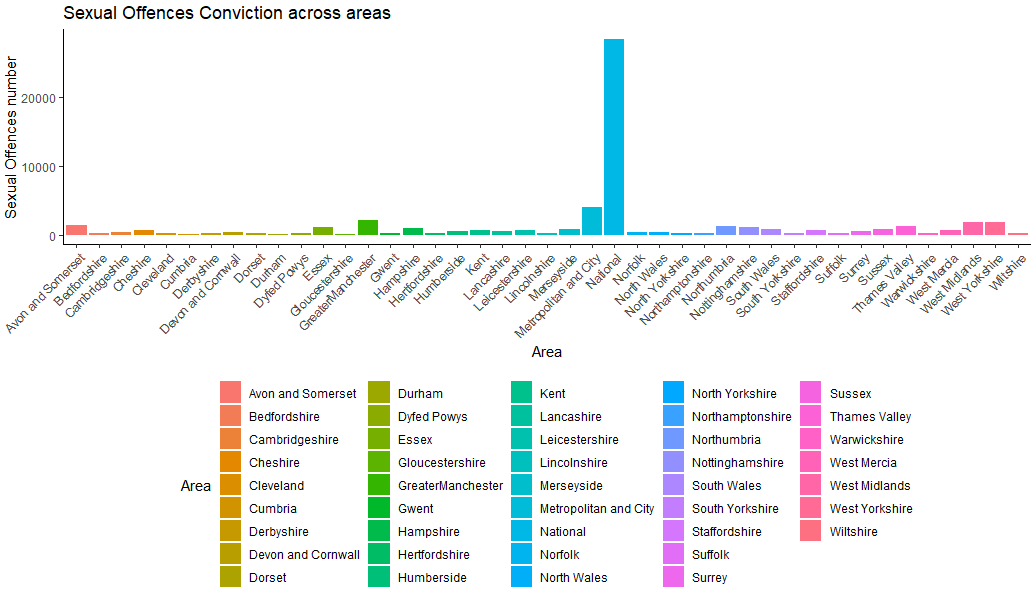


Figure 14. Sexual offenses conviction across areas

Figure 14 displays the number of sexual offense convictions in various areas.

Figure 14 depicts the distribution of sexual offense convictions across various regions. The distribution is unbalanced, with the majority of regions exhibiting relatively minimal conviction counts, while a single region is characterized by a disproportionately high number of convictions, as visually represented by a bar that significantly outstrips the others. This outlier within the data set is the national date.

A critical analysis from a data science perspective would encompass several dimensions:

A comprehensive contextual analysis is warranted to unravel the reasons behind the pronounced number of convictions in the outlier region. This could include an assessment of population dynamics, offence rates relative to the population, and the efficiency of the judicial process.

The integrity of the data must be scrutinized, particularly in light of the outlier, to confirm its validity and to rule out any potential for error or misinterpretation.

From a policy standpoint, these findings could be instrumental in directing focused interventions within the outlier region, potentially addressing the need for enhanced victim support services or law enforcement training. A comparative analysis may reveal whether the highlighted region also experiences elevated conviction rates across other crime categories, suggesting the possibility of broader systemic challenges. Additionally, the low conviction rates in numerous regions raise the question of potential underreporting, which could skew the true representation of sexual offenses within these areas.

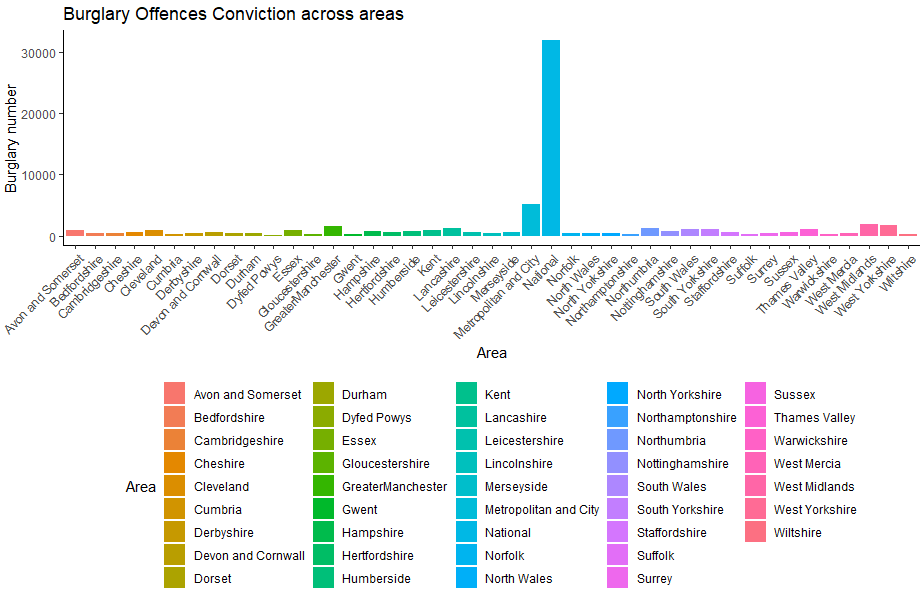


Figure 15. Burglary Offences

The bar chart (figure 15) provides a visual representation of burglary conviction counts across various regions. A clear disparity is evident from the chart, where a predominant number of regions report low conviction counts, contrasted sharply by an outlier with a significantly elevated conviction count. This outlier, marked by the tallest bar on the chart, suggests several possibilities, such as a heightened incidence of burglary, a larger population, or more effective law enforcement activity in that particular region.

The y-axis employs a linear scale that ascends to 30,000, accentuating the stark contrast between the outlier and other regions. Such a scale choice simplifies the interpretation of the data, making the vast differences in conviction counts across regions more pronounced.

Each region is denoted by a distinct color, as indicated by the chart's legend, simplifying the task of correlating the bars with their respective regions. This color coding is particularly helpful when navigating the complexity of data involving numerous categories.

From a data science standpoint, the chart invites a deeper examination of why such disparity exists, particularly in the outlier region. This investigation would encompass analyzing demographic and socioeconomic factors, law enforcement efficiency, and judicial processes. Confirming the accuracy of the data is paramount, especially considering the anomaly presented by the outlier.

Furthermore, the data's implications for policy-making are significant. If validated, it suggests a need for focused resource allocation to enhance burglary prevention and community support in the outlier region. A per capita conviction rate analysis would provide additional clarity, offering a standardized metric for comparison across regions with diverse population sizes.

Incorporating this data with other crime statistics and demographic insights could yield a more comprehensive understanding of crime patterns and underlying causes. Additionally, a temporal analysis would reveal whether the outlier region's high conviction count is an emerging trend or an established pattern.

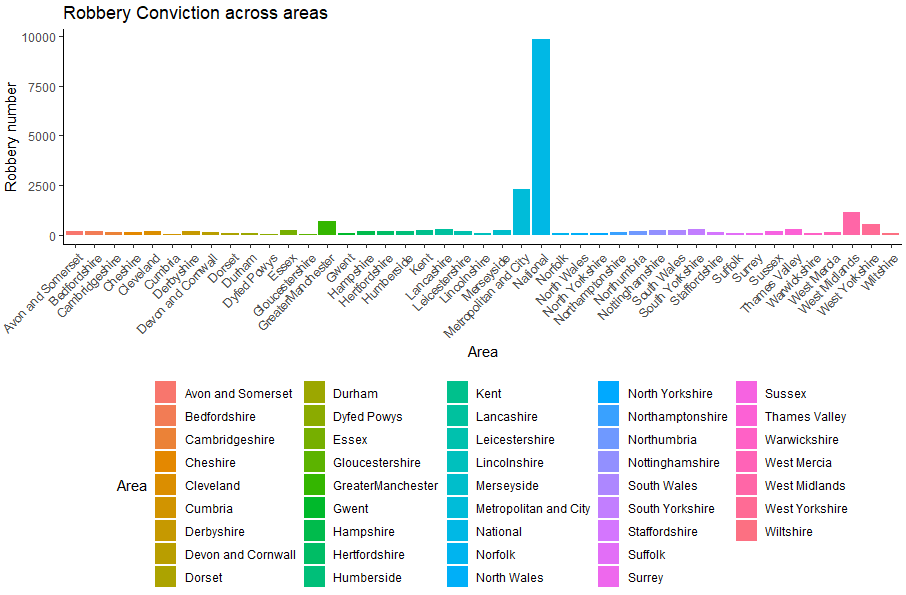


Figure 16. Robbery Conviction across areas

Figure 16 quantifies the incidence of robbery convictions across a spectrum of units, delineating distinct regional boundaries. A cursory analysis reveals a substantial uniformity in the number of convictions for a majority of these units, with the prevalence of robbery offenses generally low, as indicated by the near-baseline bars. However, an anomaly is identified - a singular unit exhibits a conviction count markedly higher than its counterparts. This spike may suggest an elevated occurrence of robberies, potentially attributable to dense population clusters or a more robust crime detection and prosecution framework within that locale.

The linear scaling of the y-axis up to 10,000 convictions enables a straightforward comparative assessment of the data across units. Accompanying this is a color-coding scheme within the chart's legend, which provides a visual shorthand for differentiating between the units in question.

From the vantage point of data science, several interpretative and analytical steps emerge:

- Data Integrity: Initial steps would include a thorough validation of the outlier data to preclude any statistical anomalies or collection errors.

- Contextual Exploration: A holistic understanding of the outlier necessitates an exploration of demographical and socio-economical variables, along with an appraisal of the law enforcement efficacy within the area.

- Strategic Policy Development: The insights garnered here could be pivotal in shaping targeted crime prevention strategies and guiding the allocation of law enforcement resources or community support services.

- Normalization Through Per Capita Analysis: Adjusting the conviction figures to a per capita basis would provide a more nuanced perspective, allowing for equitable comparisons across units with diverse population sizes.

- Inter-Crime Correlation Analysis: Integrating this data with broader crime statistics could unveil underlying trends or systemic issues, facilitating a comprehensive view of the crime dynamics within these units.

- Longitudinal Analysis: Evaluating the data temporally could discern whether the identified peak in convictions is an emergent pattern or an entrenched historical trend.

In essence, the chart serves as an impetus for an advanced analytical discourse, potentially driving a data-informed approach to understanding and mitigating the incidence of robberies within the highlighted area.

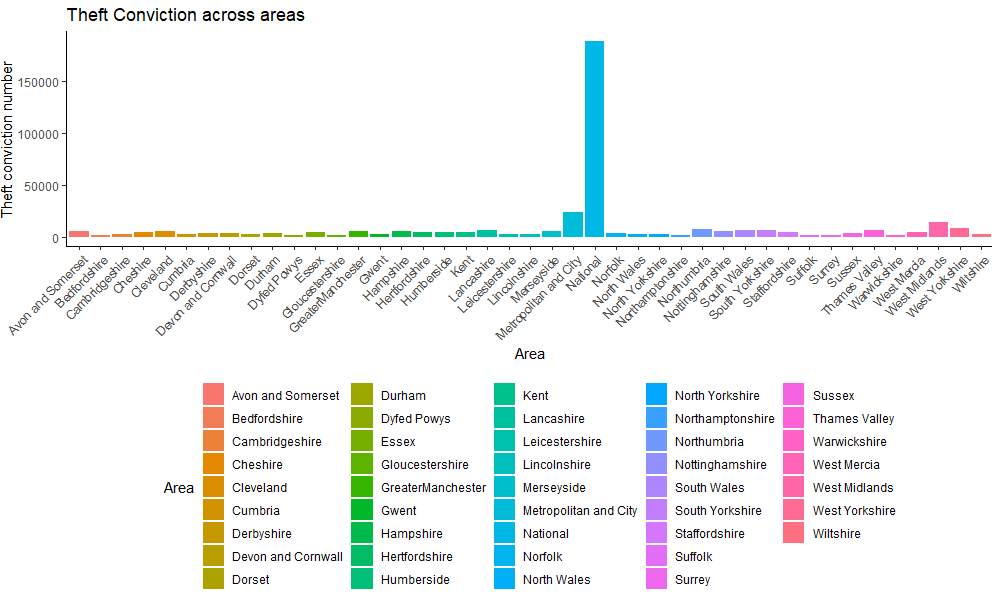


Figure 17. Theft Conviction across Areas

Figure 17 provides a detailed visual representation of theft convictions across regions. The chart reveals a significant disparity in conviction rates, with the majority of units reporting low numbers of theft convictions. However, one unit stands out prominently with a considerably higher conviction rate, as denoted by the chart's tallest bar. This notable deviation is due to the larger population base (national) of the data. The y-axis, employing a linear scale that extends up to 1.5 million, accentuates the stark contrast between the outlier and the other units. The chart further incorporates a color-coding system, facilitating easy tracking and comparison of data across the units.

From a data science perspective, this chart prompts a multi-faceted analysis:

- Data Accuracy: The extremity of the outlier underscores the need for meticulous data verification to rule out errors or anomalies.

- Contextual Exploration: Assuming data accuracy, it becomes crucial to understand the dynamics driving the outlier’s high conviction rate, including factors like population density, economic conditions, regional reporting practices, and law enforcement efficiency.

- Policy Decision-Making: The data, particularly if reflective of a genuine trend, can significantly influence policy decisions and resource allocation, targeting regions with elevated theft rates.

- Per Capita Analysis: To gain a more comprehensive understanding, examining theft conviction rates on a per capita basis would offer a balanced view, especially for units with diverse population sizes.

- Cross-Statistical Analysis: A comparative analysis with other socio-economic and law enforcement data could shed light on underlying causal factors.

- Temporal Analysis: Evaluating the trend over time would help ascertain whether the high conviction rate in the outlier unit is a recent development or a longstanding issue.

This chart, therefore, serves as a critical starting point for an in-depth exploration into the factors influencing the distribution of theft convictions, particularly in the significantly divergent unit.

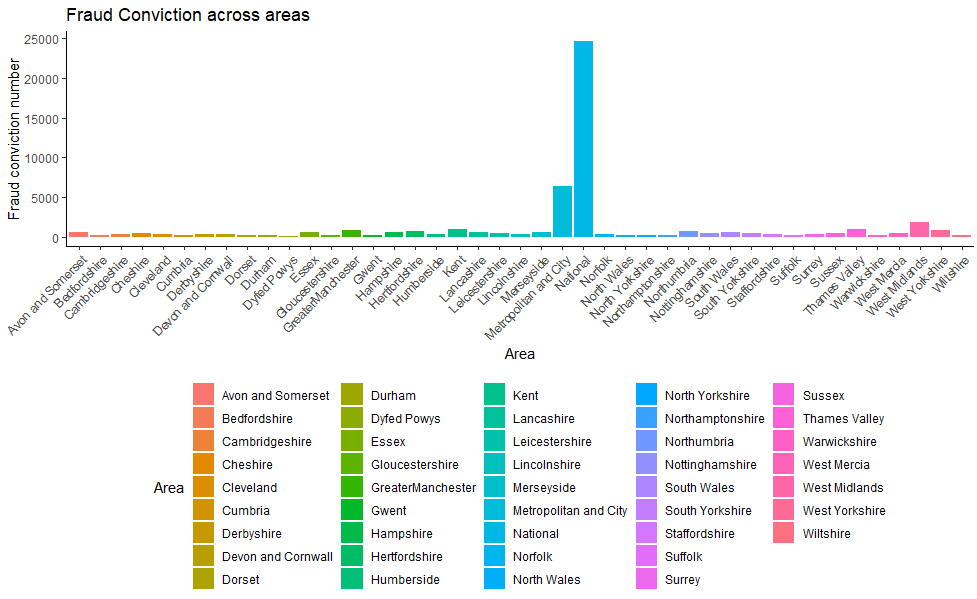


Figure 18. Fraud conviction

Figure 18 provides a visualization of the distribution of fraud convictions among a variety of defined areas. This visualization reveals several critical insights from a data science perspective:

Firstly, the distribution of fraud convictions across these areas is notably skewed. The majority of areas are characterized by relatively modest conviction counts, a trend discernible from the prevalence of shorter bars across the breadth of the chart. However, a singular area emerges as a significant outlier, exhibiting a markedly higher count of fraud convictions. This anomaly may suggest several underlying factors: a denser population, heightened economic activity that might inherently increase the propensity for fraud, or perhaps a more robust mechanism for fraud detection and legal prosecution in that specific area.

The y-axis, quantifying the number of convictions and extending to a ceiling of 25,000, employs a linear scale that markedly accentuates the visual prominence of the outlier. Furthermore, the use of a color-coded legend facilitates an intuitive and efficient comparative analysis across the areas represented in the chart.

Interpreting this data, several actions and considerations are imperative:

- Data Verification: A primary step involves ensuring the integrity and accuracy of the data, particularly for the area identified as an outlier, to exclude any possibilities of statistical anomalies or reporting errors.

- Contextual Analysis: Delving into the factors contributing to the elevated fraud convictions in the outlier area is crucial. This exploration might encompass economic dynamics, population density considerations, and the prevalence of financial institutions or districts within the area.

- Policy Implications and Resource Allocation: The insights derived from this data are pivotal in guiding policy decisions. Strategic resource allocation for fraud prevention initiatives and legal enforcement could be particularly targeted towards the outlier area.

- Per Capita Analysis: To achieve a balanced understanding, evaluating the rate of fraud convictions on a per capita basis is advisable, especially if the outlier area is characterized by a significantly large population.

- Comparative Analysis: Correlating these findings with other crime statistics could uncover broader trends or systemic issues, contributing to a more holistic understanding of the region’s crime profile.

- Temporal Analysis: Investigating whether the observed pattern represents an emerging trend or a consistent historical phenomenon is essential in understanding the evolution of fraud-related activities in these areas.

In summation, this chart not only underscores a marked disparity in fraud convictions across the surveyed areas but also serves as a catalyst for a thorough investigation into the factors influencing the outlier area's disproportionate conviction rate.

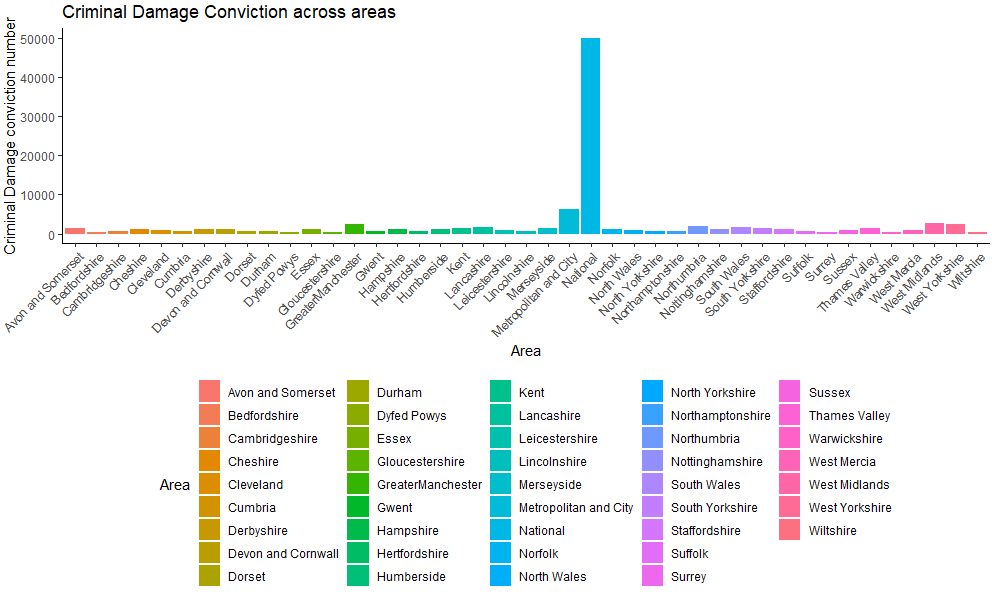


Figure 19. Criminal Damage Conviction across areas

Figure 19 illustrates the distribution of criminal damage convictions among various areas within the UK. A preliminary analysis reveals a highly skewed distribution of convictions. A majority of the areas exhibit relatively low conviction numbers, as evidenced by the cluster of bars near the baseline of the chart. This could suggest that criminal damage, as a category of crime, is generally less prevalent or less frequently results in convictions across most of these regions.

However, one area starkly contrasts with the rest, displaying an anomalously high number of convictions. This outlier dominates the chart and indicates an area where either the incidence of criminal damage is substantially higher, or the rate of reporting and successful conviction is significantly more pronounced. As a data scientist, one must consider various factors that could contribute to this observation. This includes the area's population density, socio-economic conditions, effectiveness of law enforcement, or even the possibility of a data recording anomaly.

To draw more meaningful conclusions, further investigation would be required. This would involve verifying the data's accuracy, examining the specific legal framework for criminal damage in the outlier area, and comparing per capita conviction rates rather than absolute numbers to account for population differences. Additionally, correlating these findings with other crime statistics could help identify whether this area has a generally high crime rate or if this trend is specific to criminal damage.

In writing a report on crime in the UK, it would be crucial to contextualize these findings within the broader societal landscape, taking into account the potential impact of recent events, policy changes, and local crime prevention strategies. The extreme value highlighted in the chart would be emphasized as a point of interest for policymakers and law enforcement, possibly denoting a need for targeted interventions or resource allocation.

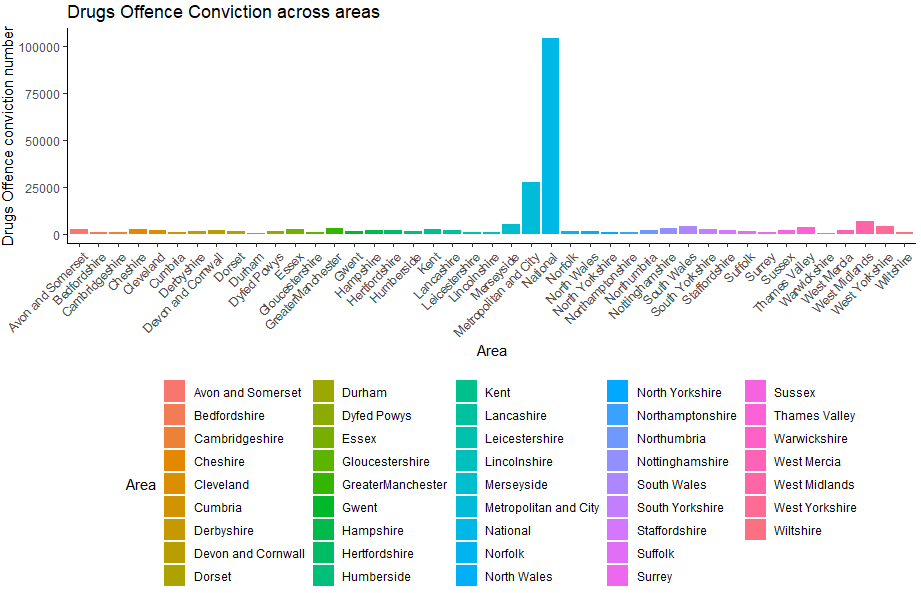


Figure 20. Drugs Offence across areas

The bar chart represents the number of drug offense convictions in various areas across the UK. The y-axis is on a logarithmic scale, which is evident from the exponential notation (1e+05, which equates to 100,000), suggesting a wide range of conviction numbers. The data exhibits a highly skewed distribution, with the vast majority of areas having relatively low numbers of drug offense convictions. This is illustrated by the cluster of bars near the base of the chart.

In a comprehensive report, these data would be contextualized against socioeconomic indicators, population data, and perhaps the availability and accessibility of substances within these areas. A deeper dive into the data might include a time-series analysis to track conviction trends over the years, an examination of the types of drugs involved, and a correlation with other forms of crime.

## Correlation and Regression Analysis

After the exploration, a correlation test was done to see if there was a relationship between robbery offense convictions and homicide convictions. To achieve that, a scatterplot was used to show the linearity of the data values, and then a Pearson correlation test was done on the data

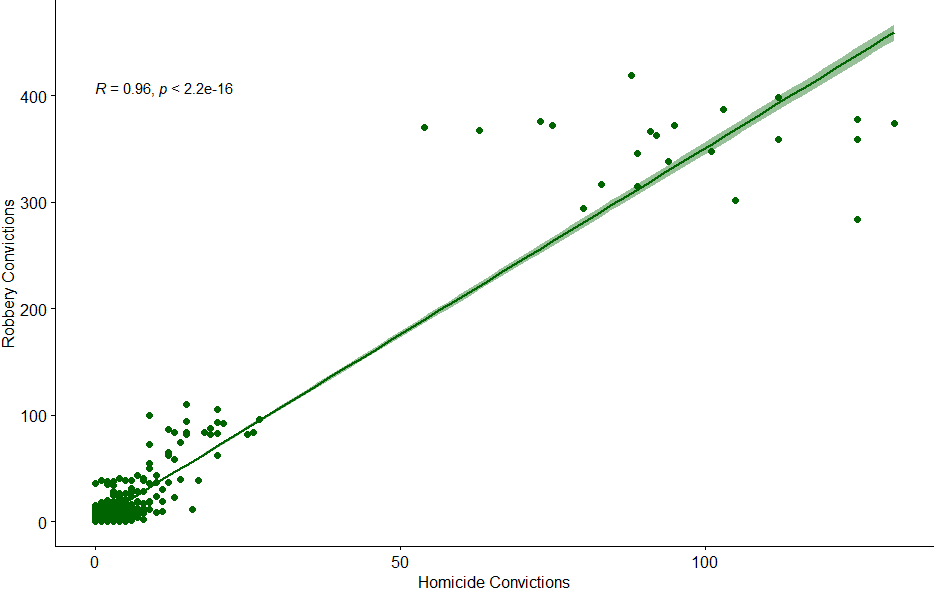


Fig 21. Correlation Plot

From the plot, there is a linear relationship between the two columns, which shows that somehow sexual offenses are related to homicides. To gain deeper insight into this chart, a correlation analysis was conducted (fig 22).

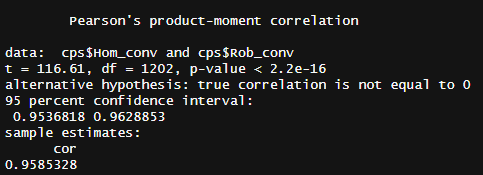


Fig 22. Correlation Result

Figure 22 presents the results of Pearson's product-moment correlation test between two variables: the number of homicide convictions (`Hom\_conv`) and the number of robbery convictions (`Rob\_conv`). The Pearson correlation coefficient (r) is 0.9585328, indicating a very strong positive linear relationship between the two variables. This means that as the number of homicide convictions increases, the number of robbery convictions tends to increase as well, and vice versa.

The t-value of 116.61 with 1202 degrees of freedom is extremely high, which supports the strong correlation. The p-value is less than 2.2e-16, which is essentially zero. This extremely low p-value indicates that the probability of observing such a high correlation by chance is virtually nonexistent, assuming that there's no true correlation between the variables. This provides very strong evidence against the null hypothesis, which is that there is no correlation between the two types of convictions.

The 95% confidence interval for the correlation coefficient ranges from about 0.9537 to 0.9629, indicating that we can be 95% confident that the true correlation coefficient lies within this range. This narrow interval further supports the strength and precision of the estimated correlation.

There is compelling statistical evidence to suggest that the number of homicide convictions is highly correlated with the number of robbery convictions in the data provided.

Regression Analysis performed was done, by analyzing robbery convictions and sexual offense as predictor variables for homicide convictions. Seeing that most of the cases could be related, it is right to use them as determinants in predicting the value.

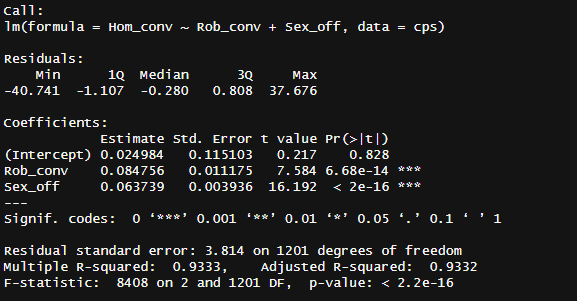


Fig 23. Regression Analysis

This regression analysis output encapsulates the relationship between homicide convictions and two predictor variables: robbery and sexual offense convictions within the `cps` dataset. The model's residuals span a substantial range, indicating variability in the model's predictive accuracy across the data. The intercept, while quantitatively minimal, is not statistically significant, suggesting that the baseline level of homicide convictions when both predictors are zero is not distinguishable from zero within the confidence of this model.

The coefficients for robbery and sexual offense convictions both exhibit strong positive correlations with homicide convictions, as indicated by their respective estimates and the negligible p-values, which are well below the conventional alpha levels for statistical significance. These findings suggest that increases in robbery and sexual offense convictions are associated with increases in homicide convictions.

The model's diagnostic metrics reveal a compelling narrative. With a Multiple R-squared of over 93%, the model accounts for a vast majority of the variance in homicide convictions. The Adjusted R-squared corroborates the model's efficacy, adjusting for the number of predictors and affirming that the explanatory variables robustly account for the outcome variable's variability. The overwhelming F-statistic and its associated p-value further cement the model's validity, confirming that the predictors' collective contribution is statistically significant. The analysis delineates a strong statistical relationship between the predictors and the response variable. It's a testament to the model's strength in explaining the variability in homicide convictions and underscores the robustness of the predictors.

## Clustering Analysis

The dataset was grouped into 2 clusters using the columns of the dataset. The clustering was performed using K-Means clustering. The result of the clustering is shown below:

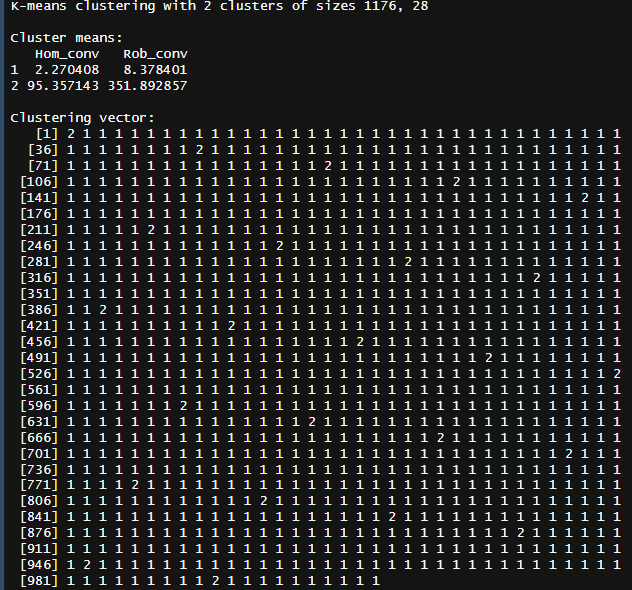


Fig 24. Clustering Analysis

After the clusters were obtained, they were visualized on the dataset to show how the clustering model performed the clustering. The plot below shows that

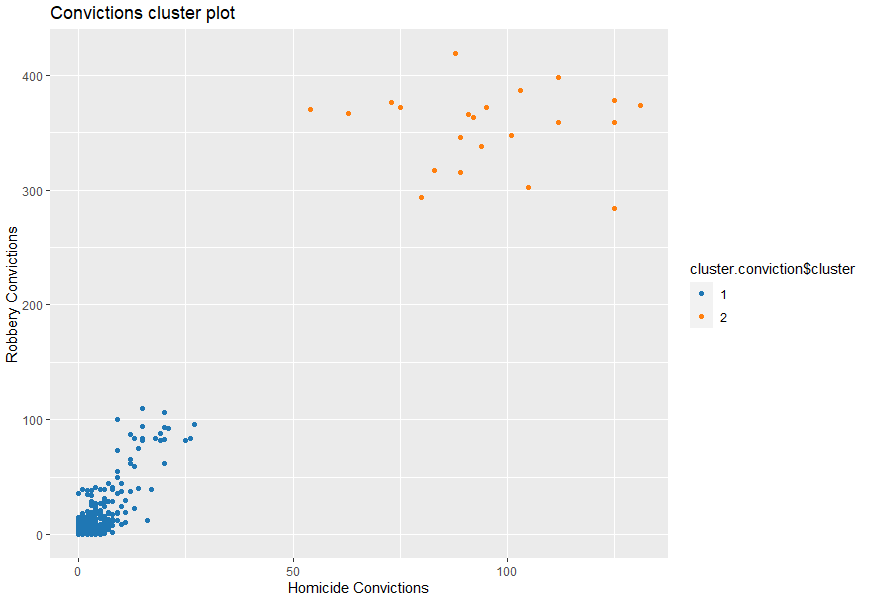


Fig 25. Clusters

The clusters showed the K-Means model was able to cluster the data very well without any problem.

## Hypothesis Test Results

This regression analysis is a statistical assessment of the relationship between the number of homicide convictions (Hom\_conv) and the number of robbery convictions (Rob\_conv), as well as sexual offence convictions (Sex\_off), using data from the `cps` dataset.

Here are the key points from the analysis:

1. Model Formula: The linear model predicts homicide convictions based on the number of robbery and sexual offence convictions.

2. Residuals: The residuals, which are the differences between the observed values and the values predicted by the model, seem to be relatively small for the majority of the data, with the median at -0.280. There are some outliers, as indicated by the minimum and maximum values (-40.741 and 37.676, respectively).

3. Coefficients:

- The intercept coefficient is 0.024984, but it is not statistically significant (p-value: 0.828), suggesting that when the number of robbery and sexual offence convictions is zero, the baseline number of homicide convictions is not significantly different from zero.

- The coefficient for Rob\_conv is 0.084756, which is statistically significant (p-value: 6.68e-14). This implies a positive relationship where for each additional robbery conviction, the number of homicide convictions increases by approximately 0.084756, all else being equal.

- The coefficient for Sex\_off is 0.063739, also statistically significant (p-value: < 2e-16). This indicates that for each additional sexual offense conviction, the number of homicide convictions increases by approximately 0.063739, holding all else constant.

4. Statistical Significance: The t-values for both predictors are high and the p-values are very low, indicating strong statistical evidence to reject the null hypothesis that the respective coefficients are zero. This means both robbery and sexual offense convictions are statistically significant predictors of homicide convictions in this model.

5. Fit of the Model:

- The Residual Standard Error (RSE) is 3.814, which gives a sense of the typical size of the residuals.

- The Multiple R-squared value is 0.9333, which means that approximately 93.33% of the variability in homicide convictions can be explained by the model. This is a very high R-squared value, indicating a good fit to the data.

- The Adjusted R-squared is also high at 0.9332, which takes into account the number of predictors in the model and is more appropriate for comparing models with a different number of predictors.

- The F-statistic is very large (8408) and the associated p-value is extremely small (< 2.2e-16), indicating that the model is a significantly better fit than a model with no predictors.

The regression analysis suggests that both robbery and sexual offense convictions are significant predictors of homicide convictions, and the model explains a very high proportion of the variance in homicide convictions.

Based on the hypotheses defined earlier, the test was conducted and the results of the tests are stated below

### Hypothesis 1 Conclusion

The coefficient for robbery conviction is 0.084756, which is statistically significant (p-value: 6.68e-14). This implies a positive relationship where for each additional robbery conviction, the number of homicide convictions increases by approximately 0.084756, all else being equal. Based on this the alternative hypothesis was accepted.

### Hypothesis 2 Conclusion

The coefficient for sexual offense is 0.063739, also statistically significant (p-value: < 2e-16). This indicates that for each additional sexual offense conviction, the number of homicide convictions increases by approximately 0.063739, holding all else constant.

# 

# 7. Critical Review of Data Analytics Tools and Techniques Used

### Regression Analysis and Correlation Analysis

Regression analysis, as expounded by Fox and Weisberg (2018), is fundamental for modeling relationships between variables. Correlation analysis, discussed by Muenchen (2017), complements regression by quantifying the strength and direction of associations.

### Exploratory and Clustering Analysis

Exploratory analysis, according to Tukey (1977), aids in uncovering patterns and trends. For clustering analysis, James et al. (2013) provide insights into the application of clustering techniques for pattern recognition and classification.

### Data Manipulation Tools

**tidyverse, dplyr, readr:**

These tools streamline data manipulation tasks, ensuring data is in a suitable format for analysis. Wickham (2017) introduces the tidyverse philosophy, emphasizing consistency and ease of use. The dplyr package, as discussed by Wickham et al. (2017), provides a set of intuitive functions for efficient data manipulation, while the readr package facilitates data importation.

### Predictive Modeling Tools

**tidymodels**

This suite of tools provides a consistent framework for modeling. Kuhn et al. (2021) present the tidymodels approach, highlighting its modularity and integration with tidyverse principles. The framework facilitates the application of regression analysis, ensuring a coherent and reproducible modeling pipeline.

**Strengths:**

* Consistent interface with tidyverse principles (Kuhn et al., 2021).
* Integrates seamlessly with other tidy tools.

**Weaknesses:**

* Smaller community compared to Scikit-learn.

Alternative to tidymodels is Scikit-Learn, which is available in Python

Scikit-learn (Python)

**Strengths:**

* Broad range of machine learning algorithms (Pedregosa et al., 2011).
* Extensive documentation and community support.

**Weaknesses:**

* Learning curve, especially for beginners.

# 

# 8. Critical Review of Data Visualization Tools and Techniques Used

**ggplot and base R**

These tools play a crucial role in data visualization, providing interactive and aesthetically pleasing plots. Wickham (2009) emphasizes ggplot's grammar of graphics for creating sophisticated visualizations, while Sievert (2020) underscores plotly's dynamic capabilities. The integration of ggplot and plotly in R offers a powerful combination for exploratory data analysis and presentation.

In the realm of data visualization within the R ecosystem, the traditional base R graphics and the more contemporary ggplot2 library present divergent paradigms. Base R graphics, inherent to the R language, offer a foundational plotting system that excels in straightforward plot creation and provides a granular level of customization. Its intrinsic presence within R means that users can leverage its capabilities without supplementary packages. However, this granularity comes at a cost: the syntax can be verbose, and achieving nuanced customizations necessitates a deep dive into intricate coding, often with a lack of consistency across various plotting functions.

On the contrary, ggplot2, a brainchild of the tidyverse collection, encapsulates a coherent and unified grammar of graphics. This framework allows for the construction of complex and layered visual representations through a consistent and structured syntax (Wickham, 2009). Its philosophy fosters an intuitive layering of graphical components, facilitating the creation of aesthetically pleasing and information-rich visualizations. The modern defaults of ggplot2 ensure that the resultant graphics align with contemporary design standards, enhancing interpretability and visual appeal. Nonetheless, ggplot2's abstraction and reliance on the tidyverse can introduce a steep learning curve, particularly for those unacquainted with its grammatical structure. Additionally, it may exhibit performance bottlenecks with large datasets and sometimes demands workarounds for highly specialized customizations.

The dichotomy between base R graphics and ggplot2 is emblematic of the balance between simplicity and abstraction in visualization tools. The choice between them is often governed by the user’s proficiency, the complexity of the visualization task at hand, and the requisite level of detail. While base R graphics provide a solid foundation, ggplot2 offers a progressive system that caters to the nuanced demands of modern data analysis and visualization.

Alternatives to base R graphics and ggplot2 for data visualization in R are several with distinct advantages and drawbacks.

Lattice:

Lattice is a powerful visualization package in R that is based on Trellis graphics, providing a framework for data visualization with an emphasis on multivariate data.

Strengths:

- Lattice excels in creating conditioned plots (panel plots) allowing for the exploration of data across several variables and is adept at handling complex multivariate data.

- It integrates deeply with the R environment and maintains performance efficiency, particularly with large datasets.

Weaknesses:

- Lattice plots can be less intuitive to customize due to its unique syntax, and it often requires more verbose code to achieve the same results as ggplot2.

- The package lacks the extensive community support that ggplot2 benefits from, meaning users may find fewer external resources and extensions.

Plotly:

Plotly for R is an interface to the Plotly JavaScript graphing library. It enables the creation of interactive, web-based graphs that are easily shared and embedded in web pages.

Strengths:

- Interactivity is Plotly’s standout feature, allowing end-users to zoom, pan, and hover over elements in the plot for more information.

- The ability to integrate with web applications through frameworks like Shiny makes Plotly an ideal choice for creating interactive web visualizations.

Weaknesses:

- Plotly's interactivity comes with increased complexity, potentially requiring users to have some familiarity with web technologies.

- While the syntax is designed to be similar to ggplot2, there can be a learning curve for users to fully exploit its interactive capabilities.

Highcharter:

Highcharter is another R package that provides bindings for the Highcharts JavaScript library, known for its rich interactive charts.

Strengths:

- Similar to Plotly, Highcharter shines with its interactive and aesthetically pleasing chart options, along with a high degree of customization.

- It offers a wide variety of chart types, including some that are not easily achievable with ggplot2.

Weaknesses:

- Highcharter’s dependency on Highcharts means that it is not entirely open-source, which could limit its use in commercial environments without proper licensing.

- The syntax may not be as familiar to R users who have not worked with Highcharts previously.

Leaflet:

Leaflet for R is a package that allows for the creation of interactive maps. It is particularly well-suited for spatial data visualization.

Strengths:

- Leaflet is the go-to choice for spatial data due to its ease of use in creating interactive maps and integrating geographic data.

- The package supports a wide range of features like markers, shapes, and layers, which can be easily customized and extended.

Weaknesses:

- Leaflet is specialized for mapping and is not a general-purpose visualization tool, limiting its use to spatial analysis.

- Users looking for more advanced GIS capabilities may need to integrate with other spatial packages or move outside the R ecosystem.

Each of these tools offers a unique set of features and design philosophies. The selection among them is contingent upon the specific needs of the visualization task, the data's nature, and the target audience's requirements. The decision matrix for a data scientist not only involves the technical merits of these tools but also the broader context of the project's goals, the desired user experience, and the scalability of the visualization solution.

**References**

Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

Muenchen, R. A. (2017). Introduction to R for Business Intelligence.

Redman, T. C., & Blanton, A. (1997). Data quality for the information age. Artech House. Inc., Norwood, MA, USA.

Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. CRC Press.

Wickham, H. (2017). Tidyverse: Easily install and load the ‘tidyverse’. R package version, 1(1), 2017.

Wickham, H., & Wickham, H. (2016). Data analysis (pp. 189-201). Springer International Publishing.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.

Wickham, H., François, R., Henry, L., Müller, K., & Wickham, M. H. (2019). Package ‘dplyr’. A Grammar of Data Manipulation. R package version, 8.