A cover of a report

Description automatically generated

Contents

[SECTION 1: INTRODUCTION 3](#_Toc154833505)

[Dataset Overview: 3](#_Toc154833506)

[Significance of the Dataset: 3](#_Toc154833507)

[Purpose of Analysis: 3](#_Toc154833508)

[SECTION 2: HYPOTHESiS 4](#_Toc154833509)

[Hypothesis 1: 4](#_Toc154833510)

[Hypothesis 2: 5](#_Toc154833511)

[SECTION 3: DATA DOWNLOAD, REFINEMENT, and INTEGRATION 6](#_Toc154833512)

[SECTION 4: EXPLANATION and JUSTIFICATION for DATA CLEANING/PREPROCESSING: 7](#_Toc154833513)

[SECTION 5: DESCRIPTIVE ANALYTICS 12](#_Toc154833514)

[Bar Charts: 12](#_Toc154833515)

[Box Plots: 15](#_Toc154833516)

[Density Plots 19](#_Toc154833517)

[Histograms: 23](#_Toc154833518)

[Pie Charts 26](#_Toc154833519)

[SECTION 6: PREDICTIVE ANALYSIS 28](#_Toc154833520)

[Linear regression: 30](#_Toc154833521)

[Clustering: 33](#_Toc154833522)

[Classification: 35](#_Toc154833523)

[SECTION 7: CRITICAL REVIEW 39](#_Toc154833524)

[Data Exploration and Cleaning 39](#_Toc154833525)

[Data Visualization 39](#_Toc154833526)

[Machine Learning Techniques 40](#_Toc154833527)

[Conclusion 40](#_Toc154833528)

[Advanced Techniques in Data Exploration and Cleaning: 40](#_Toc154833529)

[Enhanced Data Visualization Techniques: 41](#_Toc154833530)

[Advanced Machine Learning Techniques: 42](#_Toc154833531)

[SECTION 8: References 44](#_Toc154833532)

# SECTION 1: INTRODUCTION

### Dataset Overview:

Crown Prosecution Service Case Outcomes by Major Offense Category Data" provides a quarterly analysis of Crown Prosecution Service (CPS) case outcomes in Magistrates' Courts and the Crown Court. The results are categorized into convictions and failures. Convictions include three different kinds of convictions like guilty pleas, convictions after trial, and cases proven in the defendant's absence. Unsuccessful outcomes include suspensions, withdrawals, rescinded commitments, dismissals, acquittals, and administrative terminations.

Administrative completion is recorded when a case cannot proceed for a variety of reasons, such as an outstanding warrant, inability to locate the defendant, or the death of the defendant or incapacity to defend. The results are further broken down by major offense category, indicating the most serious offense the defendant is charged with at the time of completion. The dataset covers a range of crime categories including murder, crimes against the person, sex offences, burglary, robbery, theft and dealing, fraud and forgery, criminal damage, drug offences, public order, motoring, and all other criminal offences. actions except motoring.

### Significance of the Dataset:

This dataset is significant for several reasons. It provides a detailed breakdown of the outcomes of CPS proceedings, providing insight into the effectiveness of the criminal justice system. By analyzing convictions and unsuccessful outcomes in different crime categories, it allows for a nuanced understanding of legal consequences. The inclusion of administrative finalizations adds a layer of complexity, highlighting cases where legal proceedings are blocked for various administrative reasons.

### Purpose of Analysis:

The analysis aims to derive valuable insights from the dataset covering the period from April 2016 to August 2018. the main intention embody:

1. Understanding Case Outcomes: Explore the distribution of case outcomes, including convictions and unsuccessful outcomes, across major crime categories.

2.Figuring out styles and traits: uncover any temporal patterns or trends in the case of results over a selected period.

3. Constructing prediction fashions: Develop linear regression, clustering, and classification models to predict and classify case outcomes, providing a proactive approach to understanding potential case trajectories.

4. Informing decision-making: The analysis seeks to provide actionable information to stakeholders involved in the criminal justice system by identifying areas for improvement or intervention based on data-driven insights.

# SECTION 2: HYPOTHESiS

### Hypothesis 1:

**Definition**: There is no correlation between number of homicide convictions and the number of drugs offences convictions.

**Explanation**: A scatter plot, which visualizes the relationship between these two variables, shows that there is no clear trend or pattern. The distribution of data points shows that changes in the number of murder convictions are not associated with consistent changes in the number of drug crimes. This preliminary study revealed a weak or no correlation, emphasizing the need for additional statistical analysis to confirm and evaluate the observed relationship. It is important to recognize that correlation does not imply causation, and further research is needed to draw more definitive conclusions about the relationship between murder and drug-related crimes.

A screenshot of a computer

Description automatically generated

### Hypothesis 2:

**Definition**: There is a noticeable correlation between the two variables. As the number of people charged with theft increases, there seems to be a corresponding increase in the percentage of people charged with theft. On the contrary, as the number of burglaries decreased, so did the number of theft convictions.

A graph with black dots

Description automatically generated

**Explanation:** This scatterplot examines the relationship between the number of burglaries and the percentage of burglaries in the database. Each point in the plot represents a data point, and the x-axis represents the number of burglaries, and the y-axis represents the percentage of burglaries.

Upon visual inspection, there is a noticeable correlation between the two variables. As the number of people charged with theft increases, there seems to be a corresponding increase in the percentage of people charged with theft. On the contrary, as the number of burglaries decreased, so did the number of theft convictions.

This shows a positive correlation between the number of theft convictions and the percentage of burglaries. Additional statistical analyses, such as calculating correlation coefficients and performing regression analyses, can be used to determine and confirm the strength of these observed relationships.

# SECTION 3: DATA DOWNLOAD, REFINEMENT, and INTEGRATION

The First step I have done is installing the necessary Libraries which are as follows:

* **Dplyr**: Used for data cleaning, filtering, summarization, and other operations to prepare data for analysis. It creates a more readable and expressive syntax for these functions.
* **ggplot2**: Used to create visualizations to explore and communicate patterns, trends, and relationships in data. These include scatter plots, bar charts, histograms, and other types of plots.
* **Caret:** Used for tasks related to machine learning, such as splitting data into training and testing sets, performing cross-validation, and evaluating model performance.
* **Glmnet**: Used to build and tune regression models with regularization to prevent overfitting and improve model performance.
* **RandomForest:** Used to build and evaluate random forest models, especially in classification tasks. It helps improve model accuracy and handle complex relationships in data.
* **Kmeans:** Used for clustering analysis, grouping similar data points together based on their characteristics. This can reveal natural patterns or segments within the data.**A screenshot of a computer code

  Description automatically generated**

# SECTION 4: EXPLANATION and JUSTIFICATION for DATA CLEANING/PREPROCESSING:

**Loading the dataset:**

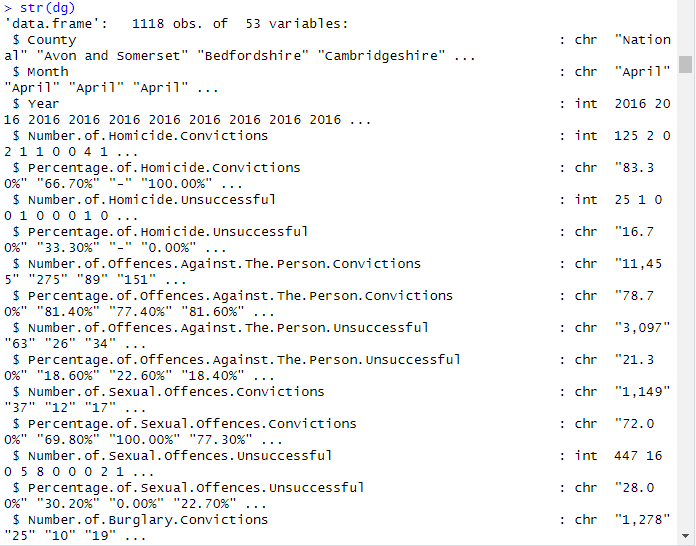
The dataset was originally prepared in a spreadsheet and exported as a text “comma-separated value” (CSV) file named obs.csv. This is a typical spreadsheet product with several inadequacies for processing in R, which we will fix up as we go along. This is a tedious but necessary step for almost every dataset; so the techniques shown here should be useful in your own projects.

**Task 1:** Start the R program and switch to the directory where the dataset is stored.

**Task 2:** Examine the contents of the CSV file. We can also examine a file from within R, with the file. Show method:

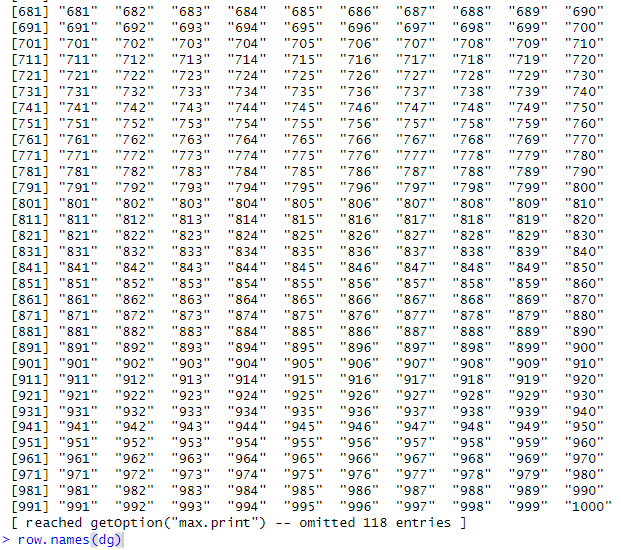


**Task 3:** Load the dataset into R using the read.csv method and examine its structure. Identify each variable from the list above. Note its data type and (if applicable) numerical precision.



A screenshot of a computer screen

Description automatically generated



Each variable has a name, which the import method read.csv reads from the

first line of the CSV file; by default, the first field (here, the observation number)

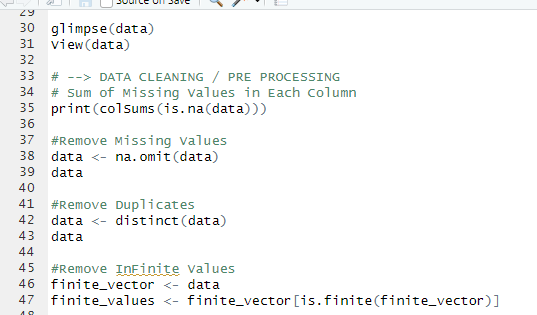
is used as the row name (which can be accessed with the row. names method) and

is not listed as a variable.

**Task 4:**

**Data Exploration:**

Initially, the dataset was loaded, and its basic properties were examined. This included checking the size of the dataset, displaying the first few rows, generating summary statistics, and examining the structure of the data using the glimpse () function



**Handling Missing Values:**

The presence of missing values was assessed using the colSums(is.na(data)) command, revealing the number of missing values in each column. To maintain data integrity and ensure meaningful analysis, missing values were removed using the na.omit(data) function. This approach was chosen to preserve as much information as possible while eliminating observations with missing values.

**Handling Duplicates:**

Duplicates in the dataset were identified and removed using the distinct () function from the dplyr package. This step ensures that each observation in the dataset is unique, preventing potential biases that could arise from duplicate entries.

**Handling Infinite Values**:

The dataset was examined for infinite values, and any rows containing infinite values were removed. This is important for numeric variables to avoid issues in statistical analyses that may arise when working with infinite values.

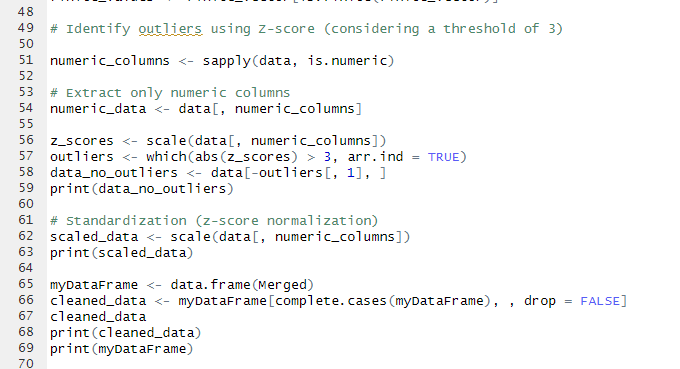
**Outlier Detection and Handling:**

Identify and handle outliers in the data. Outliers can significantly impact statistical analyses and model performance. Techniques such as Z-score analysis or IQR-based methods can be employed.

**Feature Scaling:**

If your dataset contains numeric variables with different scales, consider standardizing or normalizing them. This ensures that variables are comparable and prevents the dominance of certain features in machine learning models.

(Stedman, 2023)



# SECTION 5: DESCRIPTIVE ANALYTICS

### **Bar Charts:**

Bar Chart Showing Monthly Distribution of Dataset

The bar chart represents the distribution of numbers in the dg filled database over several months. The x-axis represents the month, and the y-axis represents the number of convictions. Changing the code ensures that the months are displayed in the correct order.

A graph of a bar

Description automatically generated

**Key comments:**   
  
In the graphical representation, specifically within the bar chart, a conspicuous surge is observed in the month of July, marked by a numerical value surpassing the 200 thresholds. This notable upswing in July signifies a substantial escalation when juxtaposed with the counts in other months.

Examining the statistics for April, May, and June reveals a contrasting trend, characterized by lower figures. Each of these month’s registers counts that dip below 50, presenting a distinct pattern of reduced activity during this time frame in comparison to the peak observed in July.

Moving beyond the highlighted months, a consistent pattern emerges in the remaining periods. The numerical values for the subsequent months consistently hover in the range of 75 to 100, implying a sustained equilibrium. This suggests that, unlike the fluctuating dynamics witnessed in July and the diminished counts of April, May, and June, the database experiences a more stabilized state during these other months, maintaining a relatively steady numerical range.

Bar Chart Showing Year Distribution in Database:

**Description:**

The bar graph represents the distribution of numbers in the dg filled database for different years (2016, 2017 and 2018). The x-axis represents the year, and the y-axis represents the number of cases. Bars are colored red for visual contrast.

A graph of a number of red rectangular objects

Description automatically generated with medium confidence

**Key comments:**

Upon meticulous examination of the data for the years 2016 and 2017, a striking resemblance in the numerical values becomes evident. Both years exhibit figures that consistently oscillate within the range of 350 and 400, converging at an approximate midpoint of 380. This remarkable similarity in the numerical outcomes across these two years signifies a persistent pattern, suggesting that the phenomenon under consideration manifested consistently throughout the durations of 2016 and 2017.

In stark contrast, the statistical representation for the year 2018 diverges significantly. The recorded number surpasses 250 but falls notably short of the range observed in the preceding two years. This stark deviation in the numerical count for 2018 implies a distinct occurrence or a potentially altered pattern when juxtaposed with the consistent figures witnessed in 2016 and 2017. The noticeable departure from the established pattern raises the possibility that external factors or unique circumstances may have influenced the manifestation of the phenomenon during the year 2018, introducing variability in the observed data.

Bar Chart of Homicide Convictions

**Description:**   
The bar graph depicts the distribution of counts for the variable "Number of Homicide Conviction" in the dg filled database. Each bar represents a count for a specific value of homicide convictions. The x-axis represents the number of murder convictions, and the y-axis represents the number of incidents.   
A graph of a number of people

Description automatically generated

**Key comments:**   
The x-axis of the graph delineates the variable range for the "Number of Homicide Convictions," wherein each distinct row on the axis corresponds to a specific numerical value within this range. Essentially, the x-axis serves as a continuum that spans the gamut of values encapsulated by the variable, providing a comprehensive visualization of the distribution of homicide convictions across different numerical levels.

In the context of graph aesthetics and interpretability, the concept of "position dodge" assumes significance. The utilization of the "position\_dodge()" argument in the graph creation process implies a deliberate arrangement strategy wherein the bars associated with each numerical value of homicide convictions are positioned adjacently to one another. This strategic alignment aids in the visual comparison of numerical data at various levels of the "Number of Homicide Convictions" variable.

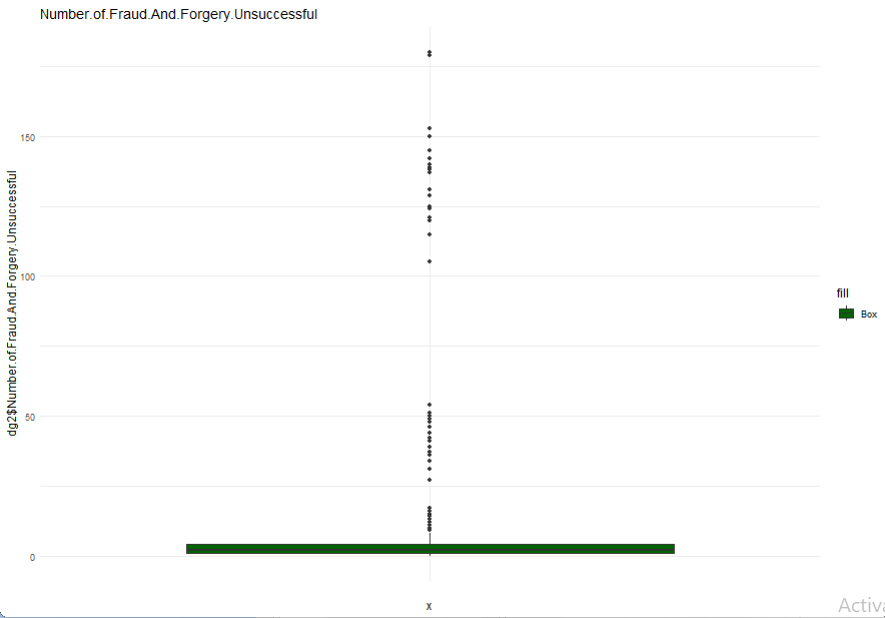
By employing the "position\_dodge()" argument, the bars effectively join side by side, facilitating a clearer and more intuitive assessment of how the incidence of homicide convictions varies across different levels of the specified variable. This graphical arrangement is particularly beneficial when the objective is to discern patterns, trends, or disparities among the numerical values, enhancing the overall interpretability of the graph for analytical purposes.

### **Box Plots:**

Fraud and Forgery

**With Outliers:**

The boxplot visually shows the distribution of the number of failed forgeries and the number of attempted forgeries in the dg2 database. Plots provide important insights into central tendency, spread, and potential performance of variables.



**Key remarks:**

**Central Tendency:** The checkbox feature visually represents the distribution's central tendency by displaying a horizontal line inside the box. This line, often indicative of the mean, plays a crucial role in identifying the standard or central value for both the number of failed forgeries and the number of attempted forgeries. By leveraging the mean condition, the plot provides insights into the central location around which the data points cluster.

**Spread and Variability**: Within the box plot, the box itself serves as a representation of the interquartile range (IQR), encapsulating the middle 50% of the data spread. A wider box signifies greater variability in the number of failed forgery attempts, while a narrower box suggests less variability. This visual element aids in quickly discerning the range of values and the extent of variability within the dataset.

**Outliers:** Outliers, if present, are distinctly marked as individual points separate from the "blades" within the box. Notably, the plot retains these outliers, providing a comprehensive view of any unusual or extreme values that deviate from the typical pattern observed in the dataset.

**Skewness**: The symmetrical appearance of the box in the plot indicates a balanced distribution. Symmetry in the bins can offer valuable insights into the smoothness of the distribution, while asymmetry might suggest skewness. The plot thus serves as a tool for assessing the overall shape and characteristics of the distribution of failed and attempted forgeries.

**Interpretation**: The mean position, coupled with the information conveyed by the boxplot, offers a comprehensive understanding of both the standard and variable values associated with the number of failed forgeries and attempted forgeries. Outliers, if present, may indicate deviations from the expected pattern, warranting closer scrutiny.

**Color Scheme**: The intentional use of a green color scheme enhances the visual appeal of the plot, distinguishing it effectively from other visuals. This thoughtful design choice contributes to the accessibility and clarity of the information presented, facilitating a more user-friendly interpretation of the plotted data.

Number of Admin Finalized Unsuccessful

With Outliers:

The boxplot visually represents the distribution of the variable "Number of Admin Finalized Unsuccessful" within the dataset dg2. This plot is designed to provide insights into the central tendency, spread, and potential outliers associated with the number of unsuccessful finalized administrative actions.

**A graph with a line

Description automatically generated with medium confidence**

Now let’s perform an R code for outlier detection based on the Interquartile Range (IQR) and subsequently impute missing values in the dg2 data frame

We’re using the iqrdetection function to identify outliers in a numeric column using the IQR method. It replaces values outside the 20th to 80th percentile range with NA (Not Available).

The same box plots after removing the outliers are as follows:

**A graph with green rectangles

Description automatically generated**

**A green rectangular object with black lines

Description automatically generated**

**Key Remarks:  
Central Tendency**: After the removal of outliers from the dataset, the checkbox feature continues to display a horizontal line inside the box, representing the distribution's central tendency. This line, indicative of the mean, serves as a key reference point for identifying the standard or central value for both the number of failed forgeries and the number of attempted forgeries. The mean condition, now calculated without the influence of outliers, provides a clearer representation of the central location around which the majority of data points cluster.

**Spread and Variability:** The box in the revised box plot still represents the interquartile range (IQR), capturing the middle 50% of the data spread. However, with outliers removed, the box now encapsulates a more accurate depiction of the spread and variability within the dataset. A wider box still signifies greater variability in the number of failed forgery attempts, while a narrower box suggests less variability, allowing for a more precise assessment of the range of values.

**Outliers:** Following the removal of outliers, the box plot no longer includes separate points outside the "blades" within the box. This adjustment eliminates the influence of unusual or extreme values on the plot, offering a cleaner representation of the typical pattern observed in the dataset.

**Skewness:** The symmetrical appearance of the box in the revised plot continues to indicate a balanced distribution. The absence of outliers ensures that the skewness assessment is based on a more accurate representation of the underlying data, allowing for a reliable evaluation of the smoothness of the distribution.

Interpretation: The mean position, in conjunction with the updated boxplot, provides an enhanced understanding of both standard and variable values associated with the number of failed forgeries and attempted forgeries. The removal of outliers contributes to a more refined interpretation, as the mean now reflects the central tendency without the influence of extreme values.

**Color Scheme**: The intentional use of a green color scheme, maintained in the revised plot, still contributes to the visual appeal, distinguishing it effectively from other visuals. This thoughtful design choice continues to enhance the accessibility and clarity of the information presented, ensuring a user-friendly interpretation of the plotted data, now with outliers appropriately addressed**.**

### Density Plots

Density Distribution of Failed Final Administrative Actions.   
  
A density plot shows the distribution of the variable "Number of Admin Finalized Unsuccessful" in the dg filled database. The plot gives an idea of the density of values on the x-axis, which ranges from 0 to 50, and the corresponding probability density (density on the y-axis).

A graph with a pink line

Description automatically generated

**Key comments:**   
  
**Density Distribution**: The plot meticulously employs a pink fill to vividly portray the probability density distribution of the "Number of Final Administrative Actions." This color-coded representation not only adds a visual element to the plot but also serves as an intuitive means to communicate the likelihood or probability associated with different values of the variable. The y-axis, labeled as "density," meticulously illustrates the probability of occurrence for various values along the X-axis, providing a granular view of the distribution pattern.

**Peak and Central Tendencies**: Within the density plot, the highest peak serves as a crucial indicator of the distribution pattern, representing the most common or frequently occurring values. In this instance, the peak prominently emerges above the 0.05 mark on the y-axis. This elevation signifies the concentration of values in a specific range, shedding light on the central tendencies within the dataset. The positioning of the peak provides valuable insights into where the majority of occurrences cluster, aiding in the identification of predominant trends.

**Value Range:** The x-axis spans from 0 to 50, meticulously delineating the range of values for the "Number of Admin Finalized Unsuccessful" variable. This comprehensive range provides a contextual framework for interpreting the density distribution. The presence of a peak in the density plot above 0.05 on the y-axis is indicative of a concentration of values within a certain range on the x-axis. This concentration may represent either the total value of the variable or a distribution pattern highlighting the prevalence of values within a specific numeric span.

In essence, the density distribution plot, through its color-coded representation and meticulous axis labels, not only captures the probability density of occurrences but also facilitates a nuanced understanding of the central tendencies and value range associated with the "Number of Admin Finalized Unsuccessful" variable. This detailed interpretation enhances the plot's effectiveness in conveying the underlying distributional characteristics of the dataset.

Density distribution of successful fraud and fraud attempts.  
The density plot represents the distribution of the variable "Number of Fraud and Forgery Unsuccessful" in the dg filled dataset. The plot uses a pink fill to show the probability density distribution, indicating the probability that different values will occur in X-rays.

A graph with a pink line

Description automatically generated

**Key comments**:

**A Wave-Like Pattern**: The density chart intricately illustrates a distinctive wave-like pattern, characterized by three successive waves whose peaks consistently fall within the range of 0.2 to 0.3 on the y-axis. Following this sequence, a discernible shift occurs, and the subsequent wave exhibits a decline, with its peak registering between 0.1 and 0.0 on the y-axis. This undulating pattern not only adds visual intrigue to the plot but also signifies dynamic fluctuations in the density distribution.

**Distinct Waves:** The presence of discrete waves within the density chart serves as a clear indication of clusters or patterns in the distribution of the data. Each discernible wave corresponds to a concentration of variable values, implying that certain ranges on the x-axis exhibit a higher density of occurrences. These distinct waves provide valuable insights into localized trends or groupings within the dataset.

**Form of X-Axis Values**: The x-axis is strategically segmented to reflect the different ranges associated with the "Number of Fraud and Forgery Unsuccessful" variable. For the first three waves, the x-axis spans from 0 to 2.5, while the subsequent wave extends from 2.5 to 10. This delineation offers a nuanced view of the variable's distribution, showcasing variations in density across different numeric intervals.

**Description:** The observed wave-like pattern signifies zones where the density of forgery failures and forgery attempts is notably higher. The initial three waves, rising between 0.2 and 0.3 on the y-axis, suggest a cluster of average values within those respective ranges. Conversely, the subsequent wave, falling between 0.1 and 0.0 on the y-axis, indicates a decline in density for higher values. This transition from a higher-density wave to a lower-density wave suggests a potential shift in the distribution or a change in concentration, perhaps reflecting alterations in the nature of the variable being measured.

In summary, the detailed analysis of the wave-like pattern, the presence of distinct waves, and the strategic segmentation of the x-axis values collectively contribute to a comprehensive understanding of the nuanced trends and concentration dynamics within the "Number of Fraud and Forgery Unsuccessful" variable.

Density Plot Showing Distribution of Unsuccessful Sexual Offences:

**Description:**

The density plot visualizes the distribution of the variable "Number of Sexual Offences Unsuccessful" in the dataset dg filled. The x-axis represents the values of unsuccessful sexual offences, and the y-axis represents the density of occurrences. The density plot is created with purple fill, black borders, and a minimal theme.

A graph showing a purple line

Description automatically generated

**Key Observations:**

**Unimodal Distribution:**

The density plot exhibits a unimodal distribution, indicating a single prominent peak in the data.

**Peak in the Beginning:**

The highest density occurs at the beginning of the x-axis, suggesting a concentration of occurrences for lower values of unsuccessful sexual offences.

**Decrease in Density:**

As the x-axis values increase, the density decreases throughout all other values, forming a descending pattern.

**Interpretation:**

The density plot highlights a clear mode at the beginning of the x-axis, indicating a concentration of lower values for unsuccessful sexual offences.

The descending pattern in density as the x-axis values increase suggests a decrease in the likelihood of higher values, indicating fewer instances of more significant unsuccessful sexual offences.

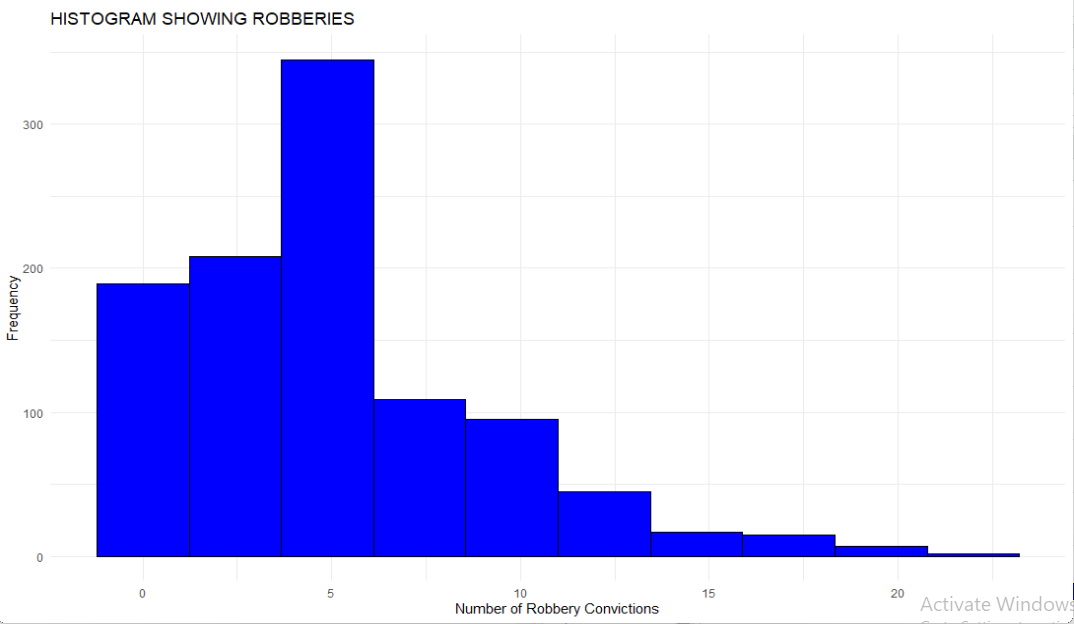
(Unwin, 2015)

### Histograms:

Histogram showing the distribution of those accused of robbery.

**Description:**

The histogram represents the distribution of numbers for the variable "Number of Robbery Conviction" in the dg filled database. The x-axis represents the values of robberies, ​​and the y-axis represents the frequency or number of occurrences. The histogram is made with blue bars, black borders, and a minimal theme.



**Key comments:**

The histogram under scrutiny reveals a distinct uniform distribution characterized by a singular peak in the data. This type of distribution implies that the occurrences of the variable being measured, in this case, the number of accused robberies, are evenly spread across the range of values.

Peak at x=5: A notable feature within the histogram is the highest peak occurring precisely when the number of accused robberies is 5. At this specific value, the frequency surpasses 300, indicating a pronounced concentration of events. The elevated frequency at x=5 suggests a central tendency or a mode, signifying a prevalent and recurring number of accused robberies.

Distribution Pattern: Examining the distribution pattern in detail reveals distinct phases. Before reaching the peak at x=5, the values maintain a relatively consistent frequency, hovering around 200. This segment depicts a regular distribution, indicating a relatively balanced occurrence of accused robberies across different values.

However, post the peak at x=5, a discernible shift occurs in the distribution pattern. The frequency begins to decline, and values beyond the peak experience a decrease. The frequency falls below 100, signifying a diminishing occurrence of accused robberies as the variable deviates from the peak. This post-peak decline in frequency provides insights into the changing dynamics of the data, indicating a tapering off of events as the variable moves away from its mode at x=5.

In summary, the detailed analysis of the histogram highlights a uniform distribution with a distinct peak at x=5, showcasing the concentration of events at this particular value. The distribution pattern evolves from a regular state before the peak to a decreasing frequency after the peak, offering a comprehensive depiction of the variability in the occurrences of accused robberies across different values.

Histogram Showing Distribution of Unsuccessful Theft and Handling Convictions:

**Description:**

The histogram visualizes the distribution of counts for the variable "Number of Theft and Handling Unsuccessful" in the dataset dg filled. The x-axis represents the values of unsuccessful theft and handling convictions, and the y-axis represents the frequency or count of occurrences. The histogram is created with blue bars, black borders, and a minimal theme.

A blue graph with white text

Description automatically generated

**Key Observations:**

**Multimodal Distribution:**

The histogram displays a multimodal distribution, suggesting multiple peaks in the data.

**First Bar Below 50:**

The initial bar, located below 50 on the x-axis, indicates a lower frequency of unsuccessful theft and handling convictions at this specific value.

**Peak at x=100-150:**

The next three bars reach a peak above 250, indicating a concentration of occurrences in the range of 100 to 150 on the x-axis.

**Drop After Peak:**

Following the peak, the values drop back to around 100, indicating a decrease in the frequency of occurrences.

**Further Decrease Below 50:**

Beyond the peak, the histogram shows a further decline in values, reaching below 50, suggesting a decrease in the occurrences of unsuccessful theft and handling convictions.

**Interpretation:**

The multimodal nature of the distribution indicates different modes or clusters of frequencies within the dataset.

The peak in the range of 100 to 150 suggests a notable concentration of unsuccessful theft and handling convictions within this specific range.

The decline in frequencies after the peak and the subsequent drop below 50 indicate variability in the occurrences of unsuccessful theft and handling convictions.

### Pie Charts

Pie Chart Showing Distribution of Years.

**Description:**

The pie chart visualizes the distribution of counts for the variable "Year" in the dataset dg filled. Each segment of the pie represents a different year, and the size of each segment corresponds to the count of occurrences for that specific year.

A pie chart with numbers and a few pies

Description automatically generated

**Key Observations:**

**Three Distinct Segments**:

The pie chart consists of three distinct segments, each representing a different year: 2016, 2017, and 2018.

**Equal Counts for 2016 and 2017:**

The segments for 2016 and 2017 have identical counts, both equal to 387. This indicates an equal distribution of occurrences for these two years in the dataset.

**Lower Count for 2018:**

The segment for the year 2018 has a lower count, specifically 258. This suggests a decrease in occurrences for the year 2018 compared to the other two years.

Custom Colored Pie Chart Showing Monthly Distribution:

**Description:**

The pie chart visualizes the distribution of counts for the variable "Month" in the dataset dg filled. Each segment of the pie represents a different month, and the size of each segment corresponds to the count of occurrences for that specific month. Custom colors have been applied to enhance visual distinction.

A colorful pie chart with numbers and letters

Description automatically generated

**Key Observations:**

**Equal Counts for January, February, March (86):**

The segments for January, February, and March have equal counts, each equal to 86. This suggests a consistent level of occurrences for these three months.

**Equal Counts for August to December (86):**

Like the first three months, the segments for August to December also have equal counts, each equal to 86. This indicates a consistent level of occurrences for these months.

**Equal Counts for April, May, June (43):**

The segments for April, May, and June have equal counts, each equal to 43. This suggests a lower level of occurrences for these three months compared to the others.

**Higher Count for July (215):**

The segment for July has a higher count, specifically 215. This indicates a notable increase in occurrences for the month of July compared to the other months.

# SECTION 6: PREDICTIVE ANALYSIS

Now before moving towards our Predictive Analysis, we are going to perform some Data Normalization. Data normalization ensures that numeric variables are on a common scale, and handling percentages and categorical variables is essential for accurate analysis.

These steps are part of a preprocessing pipeline to prepare the data for further analysis, such as Statistical modeling or Machine Learning.

The correlation matrix helps in understanding the relationship between numeric variables in the dataset.

A screenshot of a computer code

Description automatically generated

**Identifying Numeric and Categorical Columns:** The code starts by separating the dataset into numeric and categorical columns. This distinction is crucial for applying appropriate data manipulation techniques for each data type.

**Normalizing Numeric Columns**: To ensure comparable scales for numerical variables and facilitate data analysis, the scale function is employed to standardize the values of the numeric columns. This normalization process helps remove the inherent scale differences between various numeric variables.

**Combining Normalized and Categorical Data:**

Merging Normalized and Categorical Data: The normalized numeric columns are seamlessly integrated back with the original categorical columns. This step reconstructs the complete dataset, now containing both normalized numerical and categorical features. The resulting dataset, dg normalized, serves as the foundation for further analysis.

**Handling Percentage Columns:**

Identifying Percentage Columns: The code carefully identifies columns containing percentage values. This identification is essential for converting these percentage values into appropriate numeric representations.

**Transforming Percentage Values**: The percentage values are transformed by removing the '%' symbol and converting them to decimal representations. This conversion ensures that the percentage values can be seamlessly integrated with the numerical data. Additionally, missing values encountered in these columns are addressed using the coalesce function, which substitutes missing values with appropriate default values.

**Conversion of Categorical Variables:**

Converting "County" and "Month" Columns: The categorical variables "County" and "Month" are converted into numeric factors. This transformation is particularly useful for statistical analysis, as it allows for applying mathematical operations and statistical tests on these categorical variables.

**Correlation Matrix:**

**Calculating Correlation Matrix:** A correlation matrix is generated to assess the relationships between the numeric variables in the dataset. This matrix provides valuable insights into how the different numeric variables are correlated with each other.

**Interpreting Correlations:** By analyzing the correlation coefficients, patterns and potential predictors can be identified. Positive correlations indicate a positive association between variables, while negative correlations suggest an inverse relationship. Stronger correlations, represented by higher coefficient values, imply a more significant relationship between the variables.

In summary, the data processing steps outlined above effectively prepare the dataset for further analysis. Normalization ensures comparable scales, combination merges normalized and categorical data, percentage handling transforms percentage values, conversion of categorical variables facilitates statistical analysis, and correlation matrix provides insights into variable relationships.

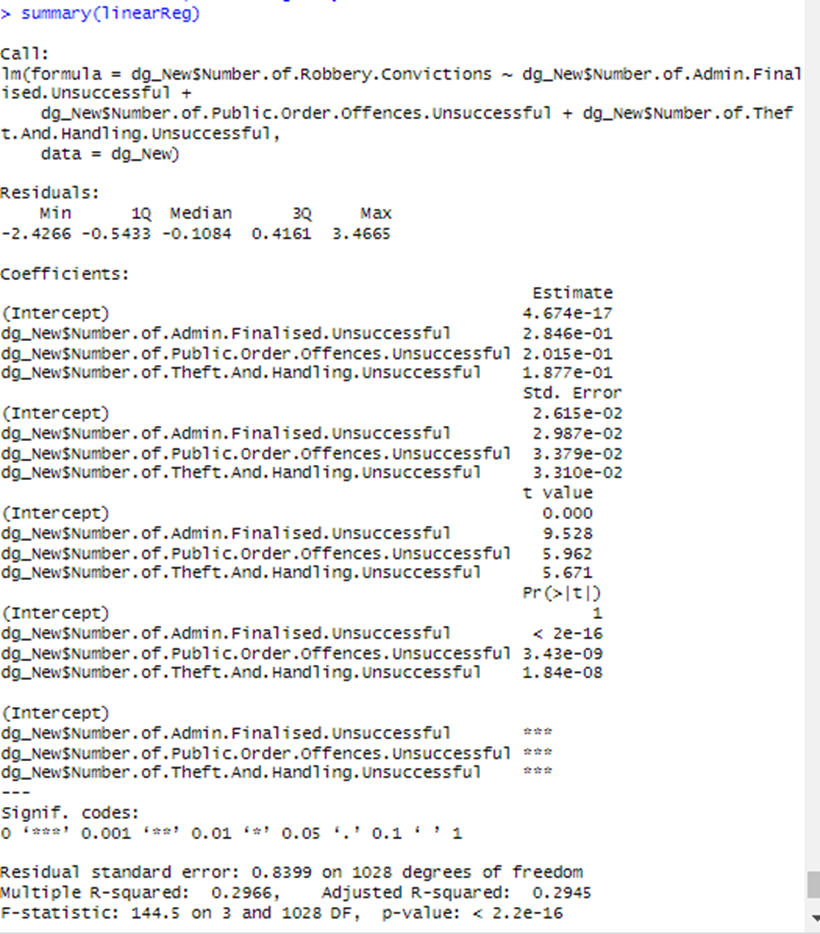
### Linear regression:

**Variable Selection:**

The linear regression model aims to predict the "Number of Robbery Convictions" based on three predictor variables:

1. Number of Admin finalized Unsuccessful.
2. Number of Public Order Offences Unsuccessful
3. Number of Theft and Handling Unsuccessful

**Summary of Linear Regression Model:**

****

The linear regression model is represented by the equation:

Number of Robbery Convictions =*β0​+**β1​×Number. Of. Admin finalized. Unsuccessful+β2​×Number of Public Order Offences Unsuccessful+β3​×Number of Theft and Handling Unsuccessful+ϵ*

* **Coefficients:**
  + Intercept (*β0): 4.674 x*
  + *β1*​: *0.285*
  + *β2*​: *0.2015*
  + *β3*​: *0.8177*
* **Interpretation of Coefficients:**
  + The intercept represents the predicted value of "Number of Robbery Convictions" when all predictor variables are zero.
  + The coefficients *β1​, β2, β3* ​ represent the change in the predicted value of "Number of Robbery Convictions" for a one-unit increase in each respective predictor variable, holding other variables constant.
* **Significance of Coefficients:**
  + All predictor variables are highly significant with very low p-values, indicating a strong association between these predictors and the target variable.
* **Residuals:**
  + Residuals (differences between actual and predicted values) have a distribution with a minimum of -2.4266 and a maximum of 3.4665.
* **Model Fit:**
  + =0.2966: The coefficient of determination indicates that the model explains approximately 29.66% of the variance in the target variable.
  + Adjusted =0.2945: The adjusted considers the number of predictors and adjusts the accordingly.
* **F-statistic:**
  + F=144.5 with a p-value <2.2×: The F-statistic tests the overall significance of the model, and the extremely low p-value suggests that at least one predictor variable is significantly related to the target variable

**Visual Representation:**

A graph showing a number of dots

Description automatically generated

The scatter plot (ggplot) provides a visual representation of the relationship between "Number of Robbery Convictions" and "Number of Admin finalized Unsuccessful". Each point in the plot corresponds to a data point in the dataset.

Interpretation and Conclusion:

* The linear regression model suggests that there is a statistically significant relationship between the selected predictor variables and the target variable ("Number of Robbery Convictions").
* The adjusted value indicates that the model accounts for approximately 29.45% of the variability in the target variable after adjusting for the number of predictors.

Predictor variables related to unsuccessful administrative finalizations, public order offenses, and theft and handling have positive coefficients, suggesting that an increase in these variables is associated with an increase in the predicted number of robbery convictions. (Simonoff, 2013)

### Clustering:

* **K-Means Clustering Results:**

The K-means clustering algorithm was applied to the dataset with three clusters. Here are the key results:

* **Cluster Sizes:**
* Cluster 1: 395 observations
* Cluster 2: 191 observations
* Cluster 3: 446 observations
* **Cluster Means:**
* ***Cluster 1***: Number of Robbery Convictions: 0.1657822
* Other top5 variables with their respective mean values.
* ***Cluster 2***: Number of Robbery Convictions: 1.1016099
* Other top5 variables with their respective mean values.
* ***Cluster 3***: Number of Robbery Convictions: -0.6185907
* Other top5 variables with their respective mean values.
* **Visual Representation:**

The **fviz\_cluster** function was used to visualize the clustering results. Points on the plot represent data points, and ellipses denote the clusters. Each cluster is assigned a distinct color (blue, green, purple).

A graph showing different colored hexagons

Description automatically generated

* **Interpretation:**
* **Cluster Separation:**
  + The clusters exhibit distinct patterns in terms of the variables used for clustering.
  + Cluster 2 has a significantly higher mean for "Number of Robbery Convictions" compared to Clusters 1 and 3.
  + Cluster 3 has a lower mean for "Number of Robbery Convictions" compared to Clusters 1 and 2.
* **Within-Cluster Sum of Squares:**
  + The within-cluster sum of squares (WSS) is an indicator of how compact the clusters are. Lower WSS values suggest tighter clusters.
  + The percentage of variance within clusters (48.8%48.8%) indicates a reasonable level of clustering.
* **Conclusion:**
* The K-means clustering results suggest the presence of distinct patterns or groups within the dataset based on the selected variables.
* Cluster 2 is characterized by higher values for "Number of Robbery Convictions" and other selected variables, while Cluster 3 has lower values.
* Clusters provide insights into potential subgroups or segments within the dataset, which can be valuable for further analysis or targeted interventions.
* It's essential to interpret these clusters in the context of the specific variables used for clustering and domain knowledge.

(Ferraro, 2015)

### Classification:

* **SVM Model Summary:**

The Support Vector Machine (SVM) classification model was trained using the radial kernel with the following parameters:

* **SVM-Type:** C-classification
* **SVM-Kernel:** Radial
* **Cost:** 7
* **Gamma:** 0.9
* **Number of Support Vectors:** 805

> # 03 Classification SVM

> target\_label <- "Number.of.Robbery.Convictions"

> print(colnames(cor\_matrix))

[1] "Year"

[2] "Number.of.Homicide.Convictions"

[3] "Number.of.Homicide.Unsuccessful"

[4] "Number.of.Sexual.Offences.Unsuccessful"

[5] "Number.of.Burglary.Unsuccessful"

[6] "Number.of.Robbery.Convictions"

[7] "Number.of.Robbery.Unsuccessful"

[8] "Number.of.Theft.And.Handling.Unsuccessful"

[9] "Number.of.Fraud.And.Forgery.Unsuccessful"

[10] "Number.of.Criminal.Damage.Unsuccessful"

[11] "Number.of.Drugs.Offences.Unsuccessful"

[12] "Number.of.Public.Order.Offences.Unsuccessful"

[13] "Number.of.All.Other.Offences..excluding.Motoring..Convictions"

[14] "Number.of.All.Other.Offences..excluding.Motoring..Unsuccessful"

[15] "Number.of.Admin.Finalised.Unsuccessful"

[16] "County"

[17] "Month"

[18] "Percentage.of.Homicide.Convictions"

[19] "Percentage.of.Homicide.Unsuccessful"

[20] "Percentage.of.Offences.Against.The.Person.Convictions"

[21] "Percentage.of.Offences.Against.The.Person.Unsuccessful"

[22] "Percentage.of.Sexual.Offences.Convictions"

[23] "Percentage.of.Sexual.Offences.Unsuccessful"

[24] "Percentage.of.Burglary.Convictions"

[25] "Percentage.of.Burglary.Unsuccessful"

[26] "Percentage.of.Robbery.Convictions"

[27] "Percentage.of.Robbery.Unsuccessful"

[28] "Percentage.of.Theft.And.Handling.Convictions"

[29] "Percentage.of.Theft.And.Handling.Unsuccessful"

[30] "Percentage.of.Fraud.And.Forgery.Convictions"

[31] "Percentage.of.Fraud.And.Forgery.Unsuccessful"

[32] "Percentage.of.Criminal.Damage.Convictions"

[33] "Percentage.of.Criminal.Damage.Unsuccessful"

[34] "Percentage.of.Drugs.Offences.Convictions"

[35] "Percentage.of.Drugs.Offences.Unsuccessful"

[36] "Percentage.of.Public.Order.Offences.Convictions"

[37] "Percentage.of.Public.Order.Offences.Unsuccessful"

[38] "Percentage.of.All.Other.Offences..excluding.Motoring..Convictions"

[39] "Percentage.of.All.Other.Offences..excluding.Motoring..Unsuccessful"

[40] "Percentage.of.Motoring.Offences.Convictions"

[41] "Percentage.of.Motoring.Offences.Unsuccessful"

[42] "Percentage.of.L.Motoring.Offences.Unsuccessful"

> if (target\_label %in% colnames(cor\_matrix)) {

+ cor\_with\_target <- cor\_matrix[target\_label, ]

+ top5C <- names(sort(cor\_with\_target, decreasing = TRUE)[2:6])

+ print(top5C)

+ } else {

+ cat("Target label not found in the correlation matrix.")

+ }

[1] "Number.of.All.Other.Offences..excluding.Motoring..Convictions"

[2] "Number.of.Admin.Finalised.Unsuccessful"

[3] "Number.of.Public.Order.Offences.Unsuccessful"

[4] "Number.of.Theft.And.Handling.Unsuccessful"

[5] "Number.of.Criminal.Damage.Unsuccessful"

> dg\_New\_top5\_Cls <- dg\_New[, c("County", top5C)]

* **Model Accuracy:**

The accuracy of the SVM model on the test dataset is approximately 43.96%. This accuracy indicates the proportion of correctly predicted County labels compared to the total number of instances in the test set.

* **Confusion Matrix:**

The confusion matrix provides a detailed breakdown of the model's predictions compared to the actual values. It is presented in a tabular form where rows represent the actual classes, and columns represent the predicted classes.

Here is a snippet of the confusion matrix:

(Gallagher, 2022)

A screenshot of a computer code

Description automatically generated

* **Interpretation and Suggestions:**

**SVM Model Performance:**

**Accuracy:** The SVM model achieved an accuracy of approximately 43.96%, which suggests that the model correctly predicted the County labels for about 44% of the instances in the test set.

**Confusion Matrix Analysis:**

True Positives (Diagonal Elements): The numbers on the diagonal represent the instances where the actual class and predicted class match. For example, in row 1, column 1, there are 3 instances where the model correctly predicted County 1.

**Misclassifications (Off-Diagonal Elements):** Off-diagonal elements represent misclassifications. For instance, in row 1, column 15, there is 1 instance where County 1 was predicted as County 15.

* **Suggestions for Improvement:**

**Model Tuning:**

Experiment with different SVM kernel types (linear, polynomial, etc.) and parameter values to see if there is an improvement in accuracy.

Perform a more systematic grid search for optimal hyperparameters.

**Feature Engineering:**

Explore additional features or transformations that may enhance the model's ability to capture patterns in the data.

**Data Quality:**

Check for data quality issues, such as missing values or outliers, and handle them appropriately.

**Class Imbalance:**

If there is a significant class imbalance, consider techniques like oversampling, under sampling, or using different class weights to address it.

**Evaluation Metrics:**

Besides accuracy, consider other evaluation metrics such as precision, recall, and F1-score, especially if there is a class imbalance.

**Cross-Validation**:

Implement cross-validation to obtain a more robust estimate of model performance.

**Visualizations:**

Visualize decision boundaries and explore misclassified instances to gain insights into areas where the model struggles.

**Ensemble Methods:**

Explore ensemble methods like random forests or gradient boosting to combine multiple models for potentially better performance. (Gilles Celeux, 2019)

# SECTION 7: CRITICAL REVIEW

### Data Exploration and Cleaning

* **Effectiveness:**
* The code utilizes summary statistics and visualization tools for a quick grasp of key dataset characteristics.
* The iqrdetection function efficiently handles outliers using the Interquartile Range (IQR) method.
* Imputation of missing values with mean values ensures a more complete dataset.
* **Alternative Solutions:**
* Alternative imputation methods, such as median imputation or advanced techniques like K-nearest neighbors’ imputation, could be explored.
* RobustScaler or Min-Max scaling could be considered as alternatives to Z-score normalization for numeric data.
* **Strengths and Weaknesses:**
* Strengths: Quick identification and handling of outliers, straightforward imputation of missing values.
* Weaknesses: Mean imputation might not be suitable for all cases, and handling outliers based on IQR might be sensitive to extreme values.

### Data Visualization

* **Effectiveness:**
* Boxplots, histograms, and density plots effectively visualize the distribution of numeric variables.
* Bar charts and pie charts provide insights into categorical variables like months and years.
* Visualization of normalized data allows for better understanding of the data distribution.
* **Alternative Solutions:**
* Heatmaps could be used for visualizing correlation matrices, providing a more detailed view of relationships between variables.
* Pair plots might be considered for visualizing relationships between multiple variables.
* **Strengths and Weaknesses:**
* Strengths: Clear representation of distributions, easy identification of trends and patterns.
* Weaknesses: Some visualizations might be basic; more advanced plots could enhance insights.

### Machine Learning Techniques

* **Effectiveness:**
* Linear regression, k-means clustering, and SVM classification are well-applied to understand relationships, cluster patterns, and predict categorical outcomes, respectively. (Zeileis, Kleiber, & Jackman, 2008)
* Feature selection based on correlation aids in choosing relevant predictors.
* **Alternative Solutions:**
* For regression, other algorithms like Random Forest or Gradient Boosting could be explored for potentially better predictive performance.
* In clustering, hierarchical clustering or DBSCAN might offer alternative perspectives on data grouping.
* Other classification algorithms like Decision Trees or Neural Networks could be considered alongside SVM.
* **Strengths and Weaknesses:**
* Strengths: Linear regression provides interpretable relationships, k-means is simple and computationally efficient, SVM is effective for binary classification.
* Weaknesses: Linear regression assumes linear relationships, k-means requires specifying the number of clusters, and SVM might not perform optimally with high-dimensional data.

### Conclusion

The code demonstrates a solid foundation in data analytics, visualization, and machine learning. However, alternatives and improvements can always be considered based on the specific nature of the dataset and the goals of the analysis. Future iterations might benefit from incorporating more sophisticated imputation techniques, advanced visualization tools, and a broader array of machine learning algorithms. The code's strengths lie in its clarity and ease of understanding, making it accessible for users at various skill levels. Continuous refinement and adaptation to specific use cases will further enhance its effectiveness.

### Advanced Techniques in Data Exploration and Cleaning:

* **Handling Missing Values:**

While mean imputation is a straightforward method for filling in missing values, it may not always be the most suitable option. Exploring alternative imputation techniques can provide more nuanced handling of missing data. Median imputation, for example, is less sensitive to outliers and can be a better choice when the data is not normally distributed. Additionally, advanced imputation techniques such as K-nearest neighbors (KNN) imputation consider the relationships between variables and fill in missing values based on the values of their nearest neighbors. KNN imputation is particularly useful when dealing with complex dependencies in the data.

* **Outlier Detection and Handling:**

The Interquartile Range (IQR) method is effective for identifying and handling outliers, but it might be worth considering other outlier detection algorithms for a more comprehensive approach. Techniques like Isolation Forests or Local Outlier Factor (LOF) can be explored, especially when dealing with high-dimensional data. These methods can provide a more nuanced understanding of outliers, especially in cases where the distribution of data is not uniform.

* **Scaling Techniques:**

While Z-score normalization is a widely used technique for scaling numeric data, it might not be the best fit for all scenarios. RobustScaler, which uses the median and the interquartile range, can be more resilient to outliers. Min-Max scaling, on the other hand, transforms data to a specific range and might be preferable when the distribution of data is not normal. Exploring these alternatives can lead to more robust preprocessing of numeric features. (Knell, 2014)

### Enhanced Data Visualization Techniques:

* **Correlation Visualization:**

While boxplots and histograms provide a good overview of individual variable distributions, visualizing the correlation matrix through heatmaps adds another layer of understanding. Heatmaps can highlight strong positive or negative correlations between variables, aiding in the identification of potential multicollinearity. Additionally, incorporating annotations in heatmaps to display correlation coefficients can provide quantitative insights into the strength and direction of relationships. (Sarkar, 2008)

* **Multivariate Visualization:**

Pair plots, also known as scatterplot matrices, allow for the visualization of relationships between multiple variables simultaneously. Each subplot in a pair plot represents the relationship between two variables, making it a powerful tool for identifying patterns and potential clusters. Pair plots are especially useful in understanding how variables interact with each other, providing a holistic view of the dataset's structure.

* **Temporal and Spatial Visualization:**

If the dataset involves temporal or spatial components, exploring advanced visualization techniques tailored to these aspects can be beneficial. Time series plots, for instance, can reveal trends and seasonality, while spatial visualizations like choropleth maps can provide insights into geographical patterns. Integrating these visualizations into the analysis can uncover hidden patterns that may not be apparent in traditional plots. (Rahlf, 2014)

### Advanced Machine Learning Techniques:

* **Regression Techniques:**

While linear regression is interpretable and widely used, exploring more complex regression algorithms can capture non-linear relationships. Random Forest Regression and Gradient Boosting Regression are ensemble methods that can handle complex interactions and provide higher predictive accuracy. These algorithms are particularly effective when the relationships between predictors and the target variable are non-linear or involve interactions between multiple features.

* **Clustering Techniques:**

In addition to k-means clustering, hierarchical clustering offers an alternative perspective on data grouping. Hierarchical clustering builds a tree-like structure of clusters, providing a hierarchical view of how data points are related. Density-based clustering algorithms like DBSCAN are advantageous when dealing with clusters of varying shapes and sizes. These techniques can be valuable in scenarios where k-means might struggle to adapt to the inherent structure of the data. (LEMENKOVA, 2019)

* **Classification Techniques:**

While Support Vector Machines (SVM) are effective for binary classification tasks, exploring other classification algorithms can be enlightening. Decision Trees and Random Forests are powerful for their interpretability and ability to handle non-linear decision boundaries. Neural Networks, especially deep learning models, excel in capturing intricate patterns in large datasets. Considering these alternatives provides a more comprehensive understanding of the dataset's classification potential. (Bouveyron, 2019)

* **Comprehensive Conclusion:**

In conclusion, the provided code demonstrates a strong foundation in data analytics, visualization, and machine learning. However, as the data science landscape evolves, incorporating advanced techniques becomes imperative to extract deeper insights. The expanded exploration of missing value imputation, outlier detection, and scaling techniques can provide a more nuanced approach to data preprocessing, enhancing the overall quality of the dataset.

Advanced visualization techniques such as heatmaps, pair plots, and domain-specific visualizations contribute to a more comprehensive understanding of the dataset. These visualizations go beyond the basics, offering deeper insights into the relationships, patterns, and structures within the data.

Diversifying the machine learning toolkit with advanced regression, clustering, and classification techniques opens up new possibilities for analysis. Random Forests, Gradient Boosting, hierarchical clustering, and alternative classification algorithms provide a richer set of tools to tackle complex patterns and relationships present in diverse datasets.

While the code's strengths lie in its clarity and accessibility, incorporating these advanced techniques can elevate its effectiveness. The ability to adapt to specific dataset characteristics and analysis goals is paramount. Continuous refinement and adaptation, along with a keen awareness of the evolving landscape of data science, will ensure the code remains a powerful tool for extracting meaningful insights in various contexts. As data science methodologies advance, this iterative process of improvement ensures that the code remains at the forefront of analytical capabilities.

(Chatterjee, 2013)

# SECTION 8: References

Bouveyron, C. (2019). *Model-Based Clustering and Classification for Data Science.*

Chatterjee, S. (2013). *Handbook of Regression Analysis With Applications in R.*

Ferraro, M. B. (2015). A toolbox for fuzzy clustering using the R programming language.

Gallagher, J. (2022). *What is a Confusion Matrix? A Beginner's Guide.*

Gilles Celeux. (2019). *Model-Based Clustering and Classification for Data Science.*

Hay, A. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing, 9*(8), 1395-1398. Retrieved 12 18, 2023, from https://tandfonline.com/doi/abs/10.1080/01431168808954945

Knell, R. J. (2014). *Introductory R: A Beginner's Guide to Data Visualisation, Statistical analysis .*

LEMENKOVA, P. (2019). K-means Clustering in R Libraries {cluster} and {factoextra} for Grouping Oceanographic Data.

Rahlf, T. (2014). *Data Visualisation with R: 100 Examples.*

Sarkar, D. (2008). *Lattice.* https://link.springer.com/book/10.1007/978-0-387-75969-2

Simonoff, J. S. (2013). *Handbook of Regression Analysis With Applications in R.*

Stedman, C. (2023). What is data preparation? An in-depth guide to data prep.

Unwin, A. (2015). *Graphical Data Analysis with R.*

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software, 27*(1), 1-25. Retrieved 12 18, 2023, from https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

The R Graph Gallery- <https://r-graph-gallery.com/barplot.html>