# Table of contents

## 1. Introduction

The analysis presented herein focuses on exploring the relationship between convictions and unsuccessful attempts in a crime dataset. The dataset encompasses various criminal incidents, with key variables of interest including the total number of convictions and unsuccessful attempts. The primary objectives of this analysis are to investigate correlations, identify patterns within distinct clusters, and develop a predictive model for administratively finalized unsuccessful attempts based on relevant factors.

## 2. Data Import and Exploration

The crime dataset used for this analysis contains information on a diverse range of criminal activities. It includes variables such as "Convictions_Total," representing the total number of convictions, and "Unsuccessful_Total," indicating the total number of unsuccessful attempts. The dataset also incorporates additional features for each criminal incident.

### 2.1 Importing the libraries & the Dataset

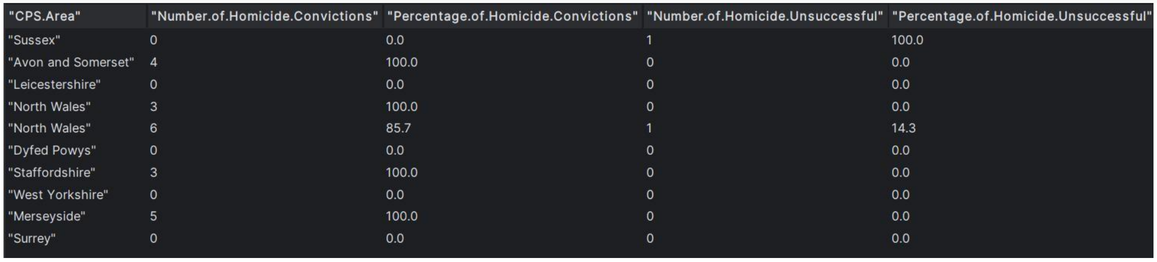
The dataset, named “Integrated_Data.csv,” is imported for exploration and analysis.

```
library('ggplot2') library('reshape2') library('GGally')

library('dplyr') library('lubridate') library('e1071')

library('factoextra') data <-

read.csv("../input/Integrated_Data.csv", header = TRUE)
```

## 2.2 Initial Data Exploration

A sample of 10 rows from the dataset is displayed to provide an overview.

```
data %>% sample_n(10) %>% print(.)
```

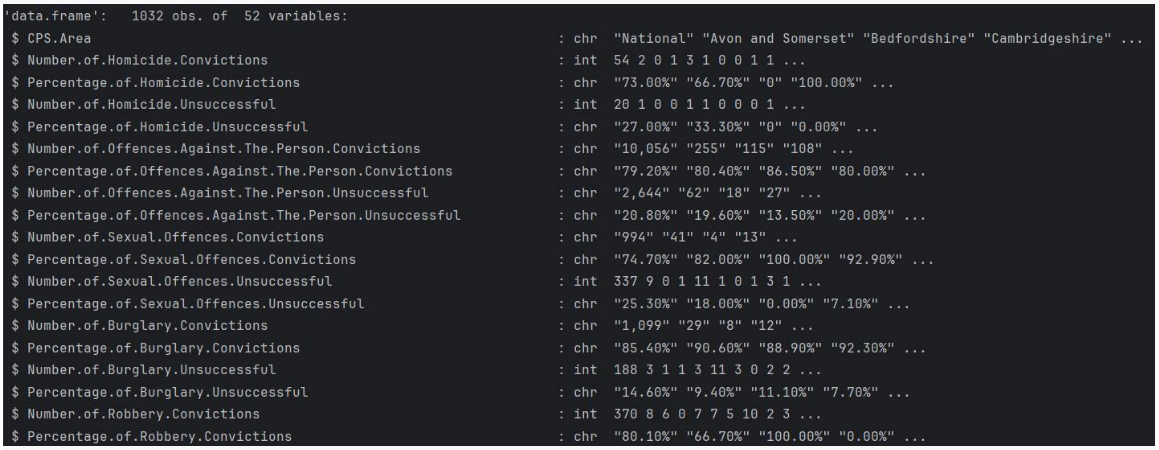


| "CPS.Area" | "Number.of.Homicide.Convictions" | "Percentage.of.Homicide.Convictions" | "Number.of.Homicide.Unsuccessful" | "Percentage.of.Homicide.Unsuccessful" |
|---|---|---|---|---|
| "Sussex" | 0 | 0.0 | 1 | 100.0 |
| "Avon and Somerset" | 4 | 100.0 | 0 | 0.0 |
| "Leicestershire" | 0 | 0.0 | 0 | 0.0 |
| "North Wales" | 3 | 100.0 | 0 | 0.0 |
| "North Wales" | 6 | 85.7 | 1 | 14.3 |
| "Dyfed Powys" | 0 | 0.0 | 0 | 0.0 |
| "Staffordshire" | 3 | 100.0 | 0 | 0.0 |
| "West Yorkshire" | 0 | 0.0 | 0 | 0.0 |
| "Merseyside" | 5 | 100.0 | 0 | 0.0 |
| "Surrey" | 0 | 0.0 | 0 | 0.0 |

### *2.2.1 Data Structure Examination*

The structure of the dataset is inspected using the str function to understand variable types and dimensions. The Statistical or 5-number summary of data

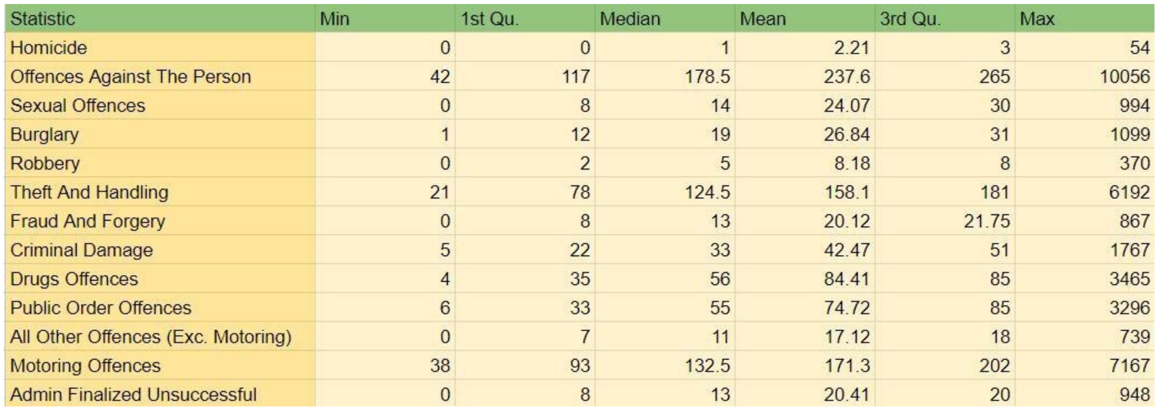
```
str(data)
```



```
'data.frame':   1032 obs. of  52 variables:
 $ CPS.Area                                          : chr  "National" "Avon and Somerset" "Bedfordshire" "Cambridgeshire" ...
 $ Number.of.Homicide.Convictions                    : int  54 2 0 1 3 1 0 0 1 1 ...
 $ Percentage.of.Homicide.Convictions                : chr  "73.00%" "66.70%" "0" "100.00%" ...
 $ Number.of.Homicide.Unsuccessful                   : int  20 1 0 0 1 1 0 0 0 1 ...
 $ Percentage.of.Homicide.Unsuccessful               : chr  "27.00%" "33.30%" "0" "0.00%" ...
 $ Number.of.Offences.Against.The.Person.Convictions : chr  "10,056" "255" "115" "108" ...
 $ Percentage.of.Offences.Against.The.Person.Convictions : chr  "79.20%" "80.40%" "86.50%" "80.00%" ...
 $ Number.of.Offences.Against.The.Person.Unsuccessful : chr  "2,644" "62" "18" "27" ...
 $ Percentage.of.Offences.Against.The.Person.Unsuccessful : chr  "20.80%" "19.60%" "13.50%" "20.00%" ...
 $ Number.of.Sexual.Offences.Convictions             : chr  "994" "41" "4" "13" ...
 $ Percentage.of.Sexual.Offences.Convictions         : chr  "74.70%" "82.00%" "100.00%" "92.90%" ...
 $ Number.of.Sexual.Offences.Unsuccessful            : int  337 9 0 1 11 1 0 1 3 1 ...
 $ Percentage.of.Sexual.Offences.Unsuccessful        : chr  "25.30%" "18.00%" "0.00%" "7.10%" ...
 $ Number.of.Burglary.Convictions                    : chr  "1,099" "29" "8" "12" ...
 $ Percentage.of.Burglary.Convictions                : chr  "85.40%" "90.60%" "88.90%" "92.30%" ...
 $ Number.of.Burglary.Unsuccessful                   : int  188 3 1 1 3 11 3 0 2 2 ...
 $ Percentage.of.Burglary.Unsuccessful               : chr  "14.60%" "9.40%" "11.10%" "7.70%" ...
 $ Number.of.Robbery.Convictions                     : int  370 8 6 0 7 7 5 10 2 3 ...
 $ Percentage.of.Robbery.Convictions                 : chr  "80.10%" "66.70%" "100.00%" "0.00%" ...
```

The figure represents some of the attributes

### *2.2.2 The Statistical or 5-number summary of data*



| Statistic | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Homicide | 0 | 0 | 1 | 2.21 | 3 | 54 |
| Offences Against The Person | 42 | 117 | 178.5 | 237.6 | 265 | 10056 |
| Sexual Offences | 0 | 8 | 14 | 24.07 | 30 | 994 |
| Burglary | 1 | 12 | 19 | 26.84 | 31 | 1099 |
| Robbery | 0 | 2 | 5 | 8.18 | 8 | 370 |
| Theft And Handling | 21 | 78 | 124.5 | 158.1 | 181 | 6192 |
| Fraud And Forgery | 0 | 8 | 13 | 20.12 | 21.75 | 867 |
| Criminal Damage | 5 | 22 | 33 | 42.47 | 51 | 1767 |
| Drugs Offences | 4 | 35 | 56 | 84.41 | 85 | 3465 |
| Public Order Offences | 6 | 33 | 55 | 74.72 | 85 | 3296 |
| All Other Offences (Exc. Motoring) | 0 | 7 | 11 | 17.12 | 18 | 739 |
| Motoring Offences | 38 | 93 | 132.5 | 171.3 | 202 | 7167 |
| Admin Finalized Unsuccessful | 0 | 8 | 13 | 20.41 | 20 | 948 |

# 3. Data Cleaning

## 3.1 Handling Missing Values

An initial check for missing values is performed, and subsequent data cleaning techniques are employed.

```
anyNA(data)


FALSE
```

## 3.2 Treatment of Percentage Columns

Percentage columns containing values represented as "-" are addressed by replacing them with 0 using Excel. This is deemed necessary as percentages cannot be averaged or estimated.

## 3.3 Extracting and Converting Percentage Columns

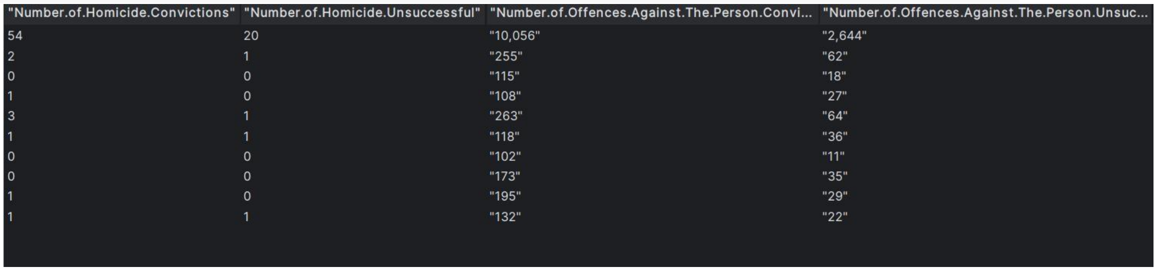
Columns containing percentage values are extracted and converted to numeric format for further analysis.

```
percentage_columns <- grep("Percentage.of.", names(data)) data[,
percentage_columns] <- apply(data[, percentage_columns], 2, func
tion(x) as.numeric(sub("%", "", x)))
typeof(data$Percentage.of.Burglary.Convictions)

"double"
```

### 3.4 Numeric Columns Extraction

Numeric columns are extracted for analysis, and any issues related to commas causing coercion are addressed using Excel.

```r
count_cols <- data[!grepl("Percentage", names(data))]

count_cols <- count_cols[-c(1, length(count_cols))]

head(count_cols, 10)
```



| "Number.of.Homicide.Convictions" | "Number.of.Homicide.Unsuccessful" | "Number.of.Offences.Against.The.Person.Convi... | "Number.of.Offences.Against.The.Person.Unsuc... |
|---|---|---|---|
| 54 | 20 | "10,056" | "2,644" |
| 2 | 1 | "255" | "62" |
| 0 | 0 | "115" | "18" |
| 1 | 0 | "108" | "27" |
| 3 | 1 | "263" | "64" |
| 1 | 1 | "118" | "36" |
| 0 | 0 | "102" | "11" |
| 0 | 0 | "173" | "35" |
| 1 | 0 | "195" | "29" |
| 1 | 1 | "132" | "22" |

**NOTE**: During the analysis, It was noted there were many String Numeric columns. Some values had commas in them like "1,0000" which caused coercion which ultimately caused Null values to generate. To avoid this, Excel was used to remove the extra commas. This file was saved separately in "output" folder.

### 3.5 Converting Numeric Strings to Integers

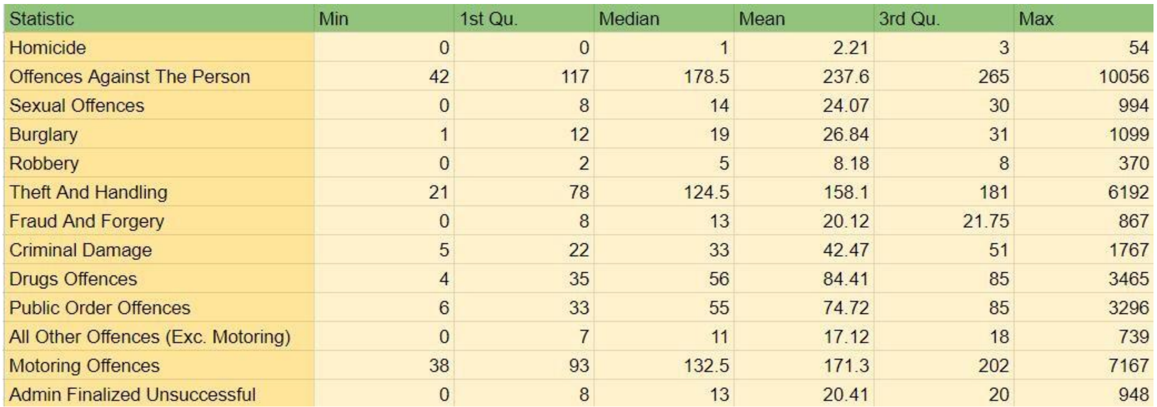
Numeric columns initially in string format are loaded and converted to integers.

```r
count_cols <- read.csv("../output/numeric_cols.csv", header = TRUE)
count_cols <- as.data.frame(lapply(count_cols, function(x) if(is.charac
ter(x)) as.integer(x) else x))

count_cols$period <- data$period
count_cols$CPS.Area <- data$CPS.Area

count_cols <- na.omit(count_cols)
anyNA(count_cols)

FALSE
```

## 4. Descriptive Analytics

### 4.1 Data Distribution

A summary of the distribution of numeric columns is presented.



| Statistic | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Homicide | 0 | 0 | 1 | 2.21 | 3 | 54 |
| Offences Against The Person | 42 | 117 | 178.5 | 237.6 | 265 | 10056 |
| Sexual Offences | 0 | 8 | 14 | 24.07 | 30 | 994 |
| Burglary | 1 | 12 | 19 | 26.84 | 31 | 1099 |
| Robbery | 0 | 2 | 5 | 8.18 | 8 | 370 |
| Theft And Handling | 21 | 78 | 124.5 | 158.1 | 181 | 6192 |
| Fraud And Forgery | 0 | 8 | 13 | 20.12 | 21.75 | 867 |
| Criminal Damage | 5 | 22 | 33 | 42.47 | 51 | 1767 |
| Drugs Offences | 4 | 35 | 56 | 84.41 | 85 | 3465 |
| Public Order Offences | 6 | 33 | 55 | 74.72 | 85 | 3296 |
| All Other Offences (Exc. Motoring) | 0 | 7 | 11 | 17.12 | 18 | 739 |
| Motoring Offences | 38 | 93 | 132.5 | 171.3 | 202 | 7167 |
| Admin Finalized Unsuccessful | 0 | 8 | 13 | 20.41 | 20 | 948 |

## 4.2 Outlier Detection

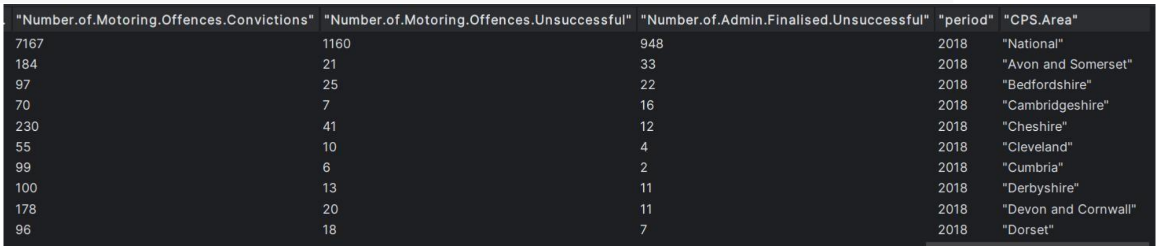
When examining percentage columns, avoid traditional **Boxplot** or outlier detection methods for unbounded data. Here's why:

**Bounded Nature**: Percentages range from 0% to 100%, causing skewed distributions with "extreme" values near boundaries. Using symmetric-distribution methods may lead to false positives.

**Interpretation**: Percentage outliers often hold meaningful interpretations, like a high conversion rate in marketing data being a positive outlier, not an error. Context is key for accurate analysis.

### Grouping data based on year and area for a more informative analysis

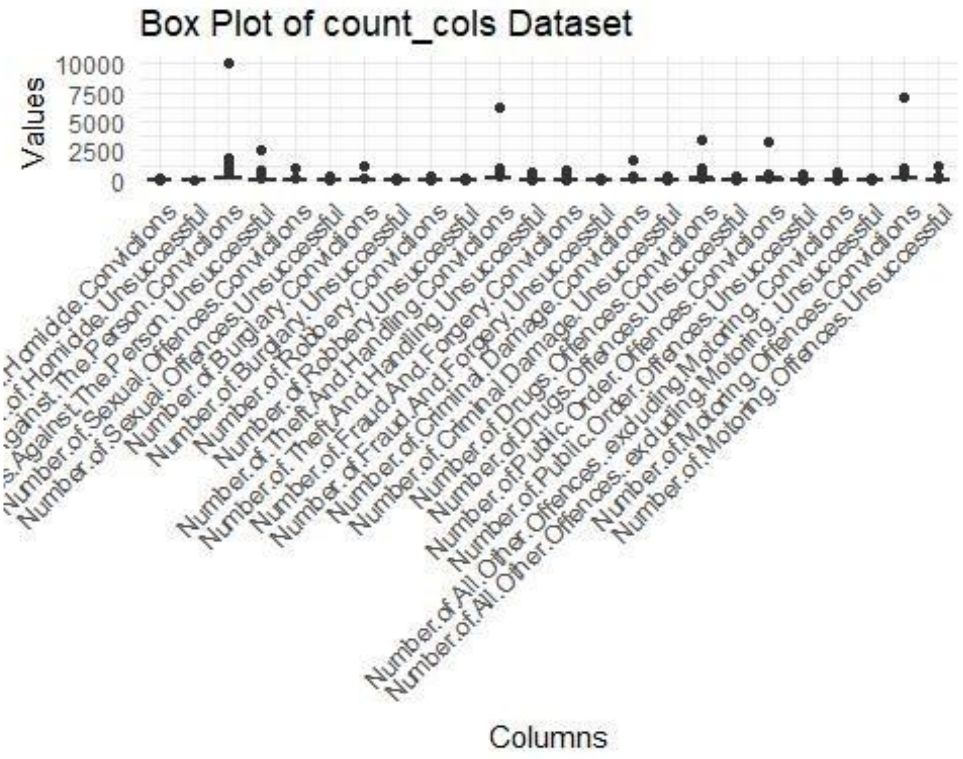
```
grouped_data <- count_cols %>%
  group_by(CPS.Area, period) head(grouped_data, 10)
```



| "Number.of.Motoring.Offences.Convictions" | "Number.of.Motoring.Offences.Unsuccessful" | "Number.of.Admin.Finalised.Unsuccessful" | "period" | "CPS.Area" |
|---|---|---|---|---|
| 7167 | 1160 | 948 | 2018 | "National" |
| 184 | 21 | 33 | 2018 | "Avon and Somerset" |
| 97 | 25 | 22 | 2018 | "Bedfordshire" |
| 70 | 7 | 16 | 2018 | "Cambridgeshire" |
| 230 | 41 | 12 | 2018 | "Cheshire" |
| 55 | 10 | 4 | 2018 | "Cleveland" |
| 99 | 6 | 2 | 2018 | "Cumbria" |
| 100 | 13 | 11 | 2018 | "Derbyshire" |
| 178 | 20 | 11 | 2018 | "Devon and Cornwall" |
| 96 | 18 | 7 | 2018 | "Dorset" |

### Boxplot

```
theme_set(theme_minimal()) melted_data <-
melt(grouped_data[, 1:24]) ## No id variables;
using all as measure variables

box_plot <- ggplot(melted_data, aes(x = variable, y = value)) +
  geom_boxplot() +
  labs(title = "Box Plot of count_cols Dataset",
       x = "Columns", y = "Values") +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) print(box_plot)
```



### Limitations

- Focus on quartiles: Boxplot prioritize quantifying the distribution for the middle 50% of the data, which might not be as relevant for extreme values in percentages.
- Misinterpretation of box size: The size of the box in a boxplot represents the inter-quartile range (IQR), which doesn't directly translate to variability in percentages.

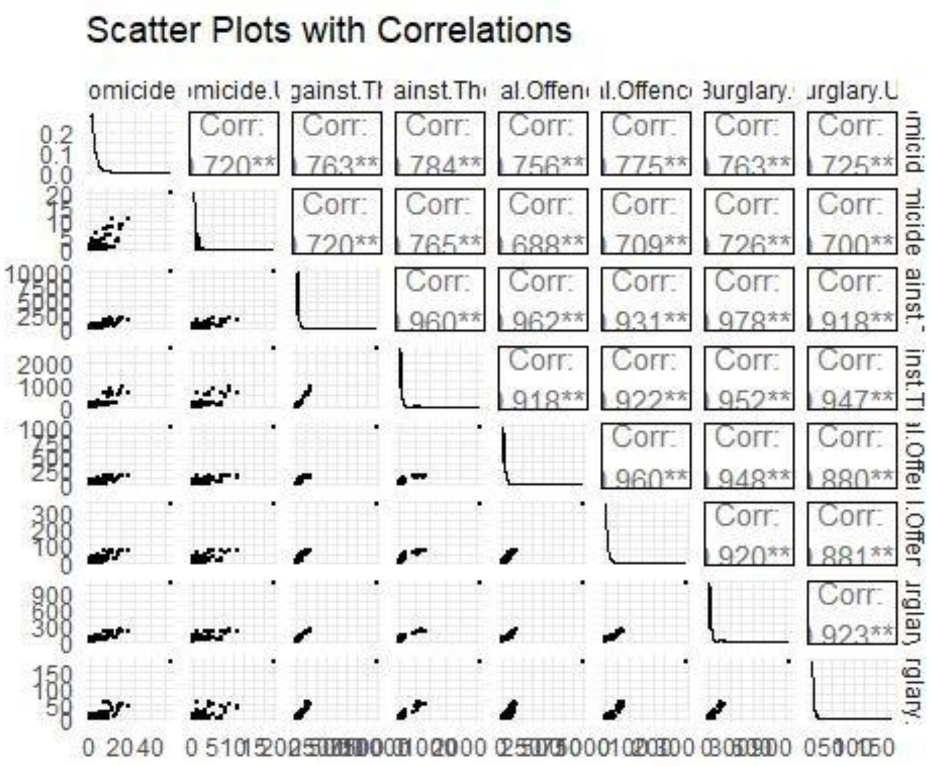
So based on these things, we are not going to remove the outliers here.

### 4.3 Visual Correlation Analysis

Visualizations, including scatter plots with correlations, are employed to analyze how variables are correlated.

Since the frame cannot be fitted into screen hence using only first 8 columns

```
ggpairs(grouped_data[1:8], columns = 1:8, lower = list(continuous = wrap
("points", size = 0.5))) + labs(title = "Scatter
  Plots with Correlations")
```

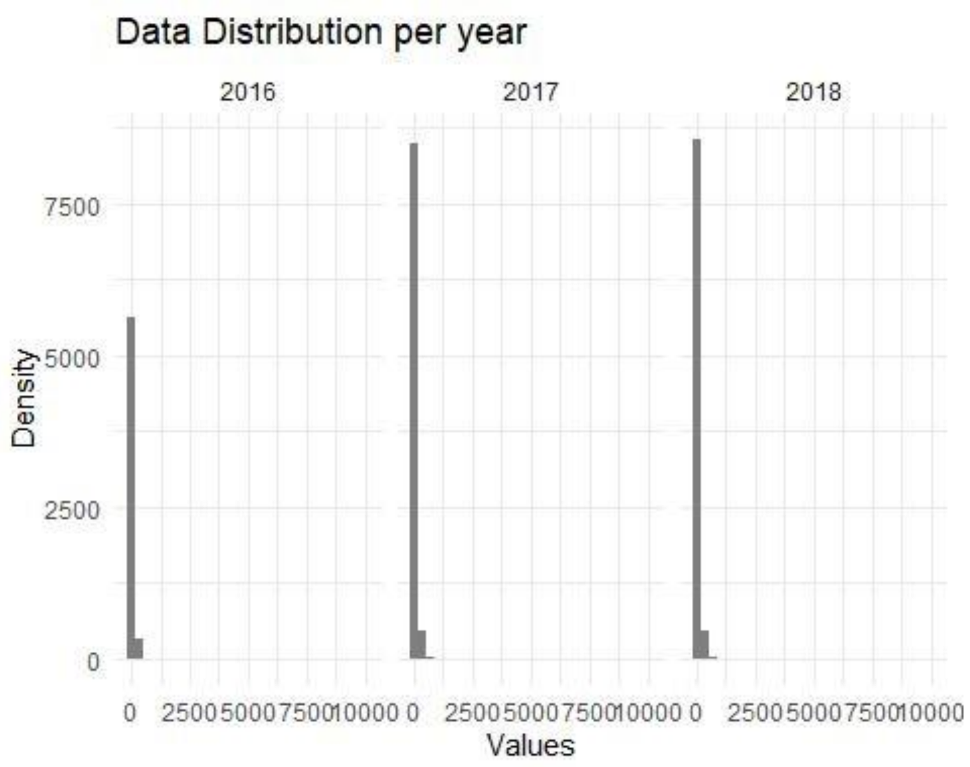


We can see the data is highly correlated with most of the columns having correlation above 90.

### 4.4 Data Distribution Visualization

Histograms are used to visualize data distribution per year.

```
melted_data <- melt(grouped_data[, 1:24]) ## No

id variables; using all as measure variables

melted_data$period <- grouped_data$period

hist_plot <- ggplot(melted_data[1:3], aes(x = value, fill = variable))
  + geom_histogram() + facet_wrap(~period) +
  labs(title = "Data Distribution per year",
      = "Values", y =
      "Density") +
  theme_minimal() + scale_fill_manual(values = c("Convictions" =
"skyblue", "Unsuccessful" = "orange")) print(hist_plot)
```



### *Data Skewness and Transformation*

The data exhibits significant right skewness, commonly observed in count data. However, the presence of both ratio and raw count attributes, coupled with poor visualizations, hampers efficient analysis.

To address this, I am aggregating all conviction columns into a single total column and applying log transformation. Why log transformation?

- **Stabilizing Variances:** Log transformation enhances data for modeling by stabilizing variances, benefiting models assuming constant variance.

- **Linearizing Relationships:** It linearizes relationships, particularly aiding linear models in capturing complex patterns.

In terms of visualization: - Log transformation normalizes skewed distributions, improving plot clarity. - It mitigates the impact of outliers, contributing to more insightful visualizations.

### *Next Steps*

1. **Exclusion of Ratio Columns:** Removing ratio columns, as they don’t significantly contribute to our analysis.

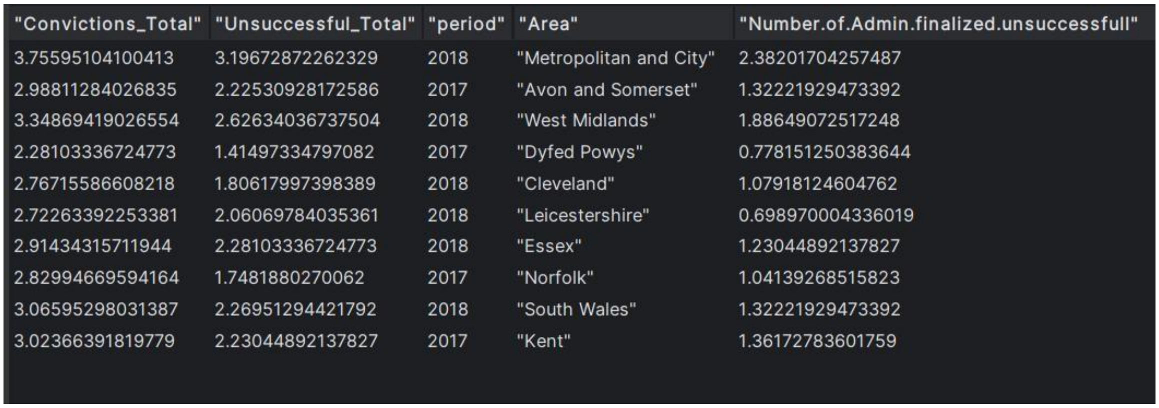
2. **Aggregation and Transformation:** Aggregating data by summing up conviction columns and applying log transformation for a more meaningful representation.

## 4.5 Log Transformation for Data Representation

To provide a more meaningful representation of the data, log transformation is applied after aggregating and excluding ratio columns.

```
aggregated_data <- grouped_data[, 1:24] %>%
  mutate(
    Convictions_Total  =  rowSums(select(.,  ends_with(".Convictions")),
rm = TRUE),
    Unsuccessful_Total = rowSums(select(., ends_with(".Unsuccessful")),
na.rm = TRUE) ) %>%
  select(-ends_with(".Convictions"), -ends_with(".Unsuccessful"))

transformed_data <- log10(aggregated_data[, 1:2] + 1) # Adding 1 to han
dle zeros transformed_data$period <- grouped_data$period
transformed_data$Area <- grouped_data$CPS.Area
transformed_data$Number.of.Admin.finalized.unsuccessfull <- grouped_dat
a$Number.of.Admin.Finalised.Unsuccessful
transformed_data$Number.of.Admin.finalized.unsuccessfull <- log10(trans
formed_data$Number.of.Admin.finalized.unsuccessfull + 1)
transformed_data %>% sample_n(10) %>% print(.)
```

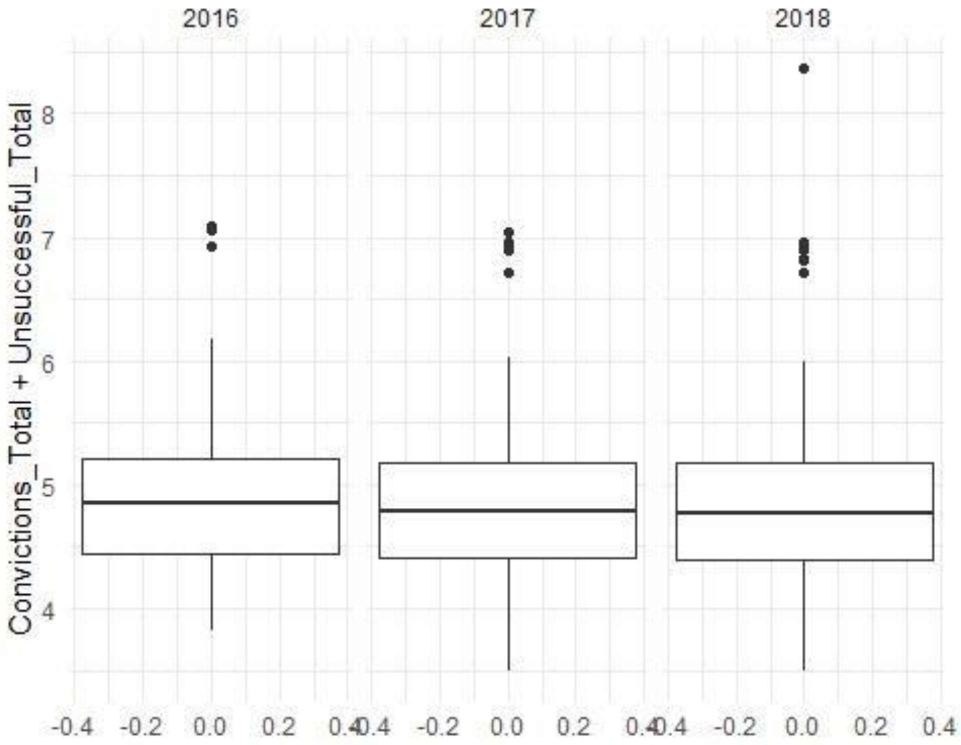
na.



| "Convictions_Total" | "Unsuccessful_Total" | "period" | "Area" | "Number.of.Admin.finalized.unsuccessfull" |
|---|---|---|---|---|
| 3.75595104100413 | 3.19672872262329 | 2018 | "Metropolitan and City" | 2.38201704257487 |
| 2.98811284026835 | 2.22530928172586 | 2017 | "Avon and Somerset" | 1.32221929473392 |
| 3.34869419026554 | 2.62634036737504 | 2018 | "West Midlands" | 1.88649072517248 |
| 2.28103336724773 | 1.41497334797082 | 2017 | "Dyfed Powys" | 0.778151250383644 |
| 2.76715586608218 | 1.80617997398389 | 2018 | "Cleveland" | 1.07918124604762 |
| 2.72263392253381 | 2.06069784035361 | 2018 | "Leicestershire" | 0.698970004336019 |
| 2.91434315711944 | 2.28103336724773 | 2018 | "Essex" | 1.23044892137827 |
| 2.82994669594164 | 1.7481880270062 | 2017 | "Norfolk" | 1.04139268515823 |
| 3.06595298031387 | 2.26951294421792 | 2018 | "South Wales" | 1.32221929473392 |
| 3.02366391819779 | 2.23044892137827 | 2017 | "Kent" | 1.36172783601759 |

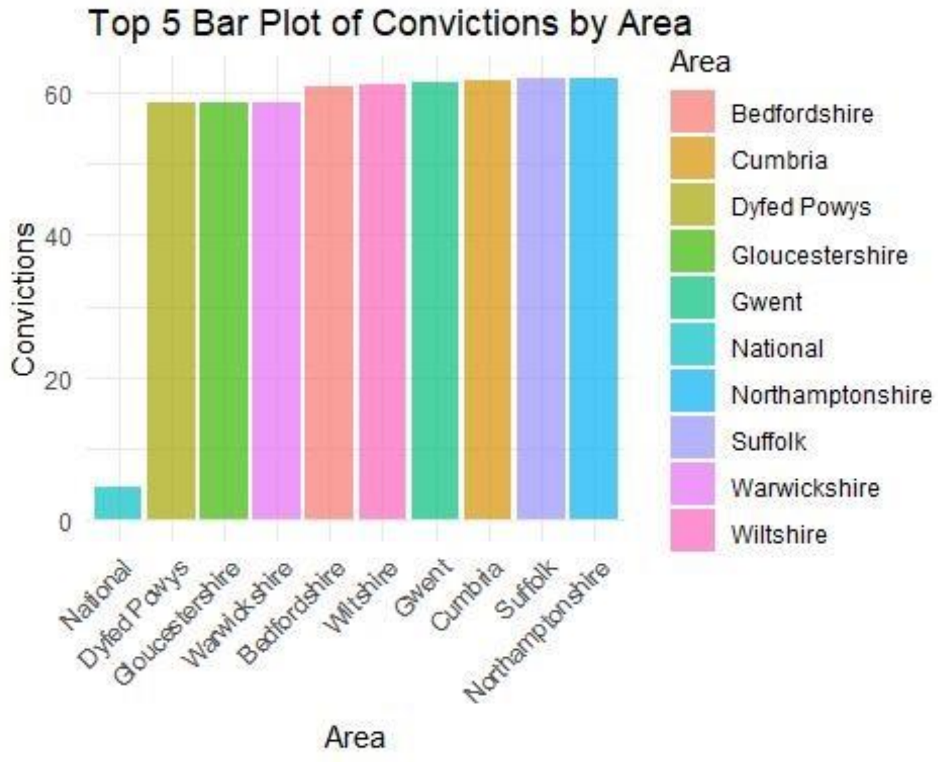
Now let us visualize this transformed data

```
ggplot(transformed_data, aes(x = Convictions_Total + Unsuccessful_Total)) +
  geom_boxplot() +
  facet_wrap(~period) +
  coord_flip()
```



What’s this ? Outliers ? Yes. Even after the transformation there are still some outliers. Let us analyze which area is this.

```
transformed_data %>%
  select(Convictions_Total, Area) %>%
  group_by(Area) %>%
  summarise(Convict = sum(Convictions_Total)) %>%
  arrange(desc(Convict)) %>% top_n(10, -Convict) %>% ggplot(aes(x =
  reorder(Area, Convict), y = Convict, fill = Area)) + geom_bar(stat
  = "identity", position = "dodge", alpha = 0.7) + labs(title = "Top
  5 Bar Plot of Convictions by Area", x = "Area", y = "Convictions")
  +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



**Interpretation:** Consider the national area, where the smallest bar might mislead. The log transformation compresses larger values and spreads out smaller ones. Despite the seemingly diminutive National bar, it likely had substantial original values. Thus, the current outliers may originate from this area, given the comparatively larger actual counts.

However, we opt not to “handle” these outliers. Despite transformation, they represent actual counts. Intuitively, addressing such outliers may not align with the inherent nature of the data.

# 5. Hypothesis Testing

## 5.1 Hypothesis 1: Correlation Between Convictions and Unsuccessful Attempts

### *5.1.1 Context*

We aim to explore the relationship between the number of convictions and unsuccessful attempts. The hypothesis stems from the intuitive notion that a higher number of convictions might correlate with a higher frequency of unsuccessful attempts, indicating a potential pattern in criminal behavior.

Hence,

**Null Hypothesis (H0):** There is no correlation between the number of convictions and unsuccessful attempts.

**Alternative Hypothesis (H1):** There is a significant positive or negative correlation between the number of convictions and unsuccessful attempts.

### *5.1.2 Methodology*

To test this hypothesis, we employed the Pearson correlation coefficient, a widely used statistical measure to gauge the strength and direction of a linear relationship between two variables. The null hypothesis (H0) posits no correlation, while the alternative hypothesis (H1) suggests a significant positive or negative correlation.

Applying Pearson correlation coefficient test

```
cor_test_result <- cor.test(transformed_data$Convictions_Total, transfo
rmed_data$Unsuccessful_Total) cor_test_result
```

```
                    Pearson's product-moment correlation

t = 91.126, df = 1000, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0 95
percent confidence interval: 0.9376623 0.9510209 sample
estimates: cor 0.9447324
```

### *5.1.3 Results*

The results of the Pearson's product-moment correlation test unveiled a robust and statistically significant correlation between the number of convictions and unsuccessful attempts (r = 0.9447, p < 2.2e-16). The 95% confidence interval for the correlation coefficient ranged from 0.9377 to 0.9510. Given the very low p-value, we reject the null hypothesis, providing substantial evidence for a significant correlation.

This outcome suggests that individuals with a higher number of convictions tend to have a higher frequency of unsuccessful attempts, hinting at a potential link between criminal history and failed endeavors.

## 5.2 Hypothesis 2: Clusters Show Different Patterns

### *5.2.1 Background and Context*

Moving beyond individual correlations, we sought to investigate whether distinct clusters within our dataset exhibited different patterns in the relationship between convictions and unsuccessful attempts. This hypothesis is grounded in the idea that various subgroups might engage in criminal behavior differently, leading to unique patterns in their unsuccessful attempts.

Therefore, the hypothesis shall be;

**Null Hypothesis (H0):** The clusters do not show different patterns in the relationship between convictions and unsuccessful attempts.

**Alternative Hypothesis (H1):** The clusters exhibitdistinct patterns in the relationship between convictions and unsuccessful attempts.

### *5.2.2 Methodology*

To address this hypothesis, we employed k-means clustering, a partitioning method that identifies natural groupings within the data. Subsequently, an Analysis of Variance (ANOVA) test was conducted to determine if these clusters demonstrated significant differences in the relationship between convictions and unsuccessful attempts.

#### 5.2.2.1 K-Means Clustering

**Why K-Means Clustering?**

K-means clustering was chosen as the method for grouping data points into clusters. This algorithm is particularly well-suited for this task due to its simplicity and efficiency in identifying natural groupings within the data. K-means assigns each data point to the cluster whose mean is nearest, resulting in clusters that capture the inherent structure of the data.

Moreover, the interpretability of k-means clusters makes it advantageous in our context. By visualizing the clusters through scatter plots with ellipses, we can easily discern the different patterns in the relationship between convictions and unsuccessful attempts, offering a valuable exploratory tool in criminological analysis.

#### 5.2.2.2 ANOVA Test

**Why ANOVA Test?**

The Analysis of Variance (ANOVA) test is employed to assess whether there are statistically significant differences in the means of multiple groups. In our case, it helps determine if the clusters exhibit distinct patterns in the relationship between convictions and unsuccessful attempts.

ANOVA is preferred in this scenario because it allows us to compare means across more than two groups simultaneously. Given that we have identified multiple clusters through k-means, ANOVA is well-suited to evaluate whether these clusters significantly differ in their criminal behavior patterns.

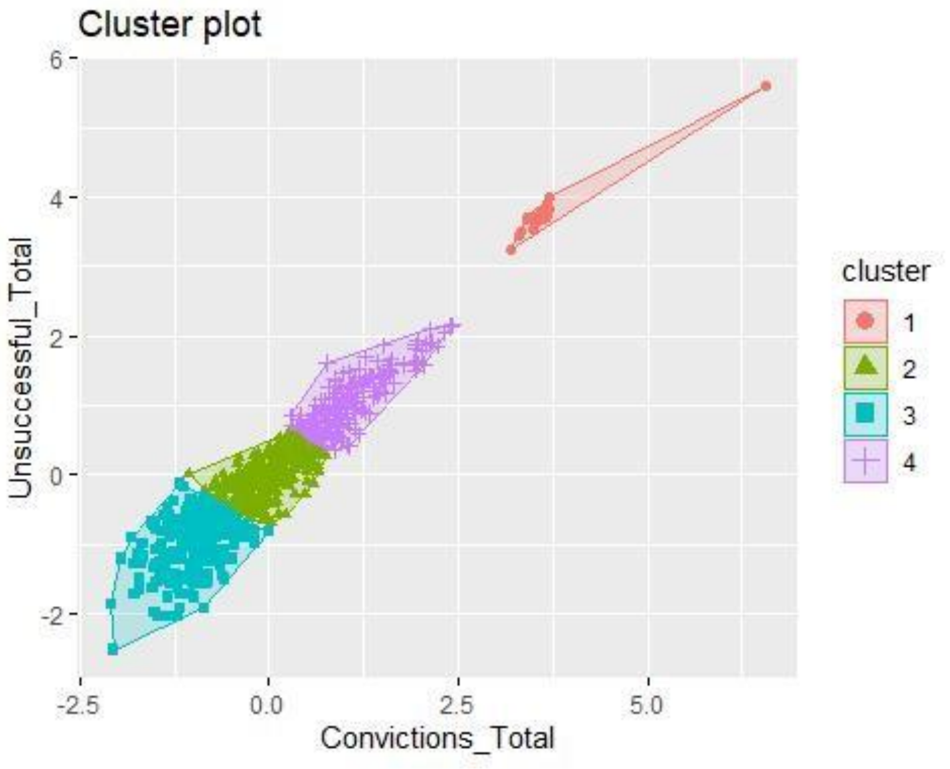
### *5.2.3 Performing k-means clustering*

```
k <- 4 set.seed(123)
kmeans_result   <-   kmeans(transformed_data[,
1:2], centers = k) transformed_data$Cluster <-
kmeans_result$cluster
```

***Visualizing Clusters***

Visual representation of clusters through scatter plots with ellipses is provided.

```
fviz_cluster(kmeans_result, data = transformed_data[, 1:2], geom =
"poin t", ellipse = TRUE)
```



### *5.2.4 Results and Intelligent Interpretation*

The application of k-means clustering resulted in the identification of four distinct clusters within the dataset. A detailed examination of the clusters reveals intriguing patterns that merit intelligent interpretation.

## Cluster Analysis:

1. **Clusters Proximity:**

- Three of the clusters exhibit a closer proximity to each other, indicating similarities in the relationship between convictions and unsuccessful attempts.

This suggests that individuals within these clusters share common characteristics or engage in criminal behavior in a comparable manner.

- The fourth cluster, however, stands out by being more distant from the others. This indicates a unique pattern, suggesting a subgroup of individuals with markedly different criminal behavior tendencies.

2. **Sparse Points in the Fourth Cluster:**

- The fourth cluster, with fewer data points, raises interesting questions. The scarcity of points might signify a distinct and less prevalent type of criminal behavior within the dataset.
- This scarcity doesn't diminish the significance of the cluster; rather, it emphasizes its uniqueness. It might represent a specialized category of individuals with a specific modus operandi or background that sets them apart from the larger groups.

## Implications and Considerations:

1. **Diversity in Criminal Behavior:**

- The proximity of three clusters suggests a shared pattern in criminal behavior, possibly indicating common socio-economic factors, motives, or demographic characteristics.
- The divergence of the fourth cluster highlights the heterogeneous nature of criminal behavior, emphasizing the need to acknowledge and analyze diverse patterns within the dataset.

2. **Outliers and Uncommon Patterns:**

- The sparsity of points in the fourth cluster makes it an outlier, drawing attention to a potential rare or unusual category of criminal behavior.
- Investigating the characteristics of this cluster could unveil unique insights into less prevalent but distinctive criminal patterns, contributing to a more comprehensive understanding of criminal behavior diversity.

## Recommendations for Further Analysis:

1. **Individual Cluster Profiling:**

- Conduct in-depth profiling of each cluster, particularly focusing on the fourth cluster, to identify key characteristics that distinguish it from the others.

- Explore demographic, geographic, or behavioral factors that might contribute to the distinct criminal patterns observed in each cluster.

2. **External Validation:**

- Validate the findings with external datasets or domain experts to ensure the robustness of the identified clusters and the reliability of the observed patterns.

3. **Policy and Intervention Strategies:**

- Tailor intervention and policy strategies based on the identified patterns within each cluster. Understanding the diversity in criminal behavior can inform more targeted and effective approaches to crime prevention and law enforcement.

In conclusion, the k-means clustering results not only reveal the existence of clusters but also provide a nuanced understanding of the diverse patterns in criminal behavior. The intelligent interpretation considers both the similarities among clusters and the distinctiveness of the fourth cluster, setting the stage for further detailed analyses and informed decision-making.

### *5.2.4 Applying the Statistical Test: ANOVA*

Analysis of Variance (ANOVA) test is applied to compare means or medians of convictions and unsuccessful attempts across clusters.

```
anova_result <- aov(Convictions_Total + Unsuccessful_Total ~ Cluster,
da ta = transformed_data) summary(anova_result)

              Df Sum Sq Mean Sq F value Pr(>F)
Cluster        1     8.0 8.034 24.83 7.38e-07 ***
Residuals 1000 323.6 0.324
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### *5.2.3 ANOVA Results Interpretation*

The Analysis of Variance (ANOVA) test was conducted to compare means or medians of convictions and unsuccessful attempts across the identified clusters.

Here’s the interpretation of the above results;

1. **Significant Difference Among Clusters:**

- The ANOVA test yielded a highly significant F value (F = 24.83, p < 0.001), indicating that there is a substantial difference among the means of convictions and unsuccessful attempts across the identified clusters.

2. **Cluster Influence on Convictions and Unsuccessful Attempts:**

- The 'Cluster' factor significantly influences the variability in both convictions and unsuccessful attempts. The low p-value (p < 0.001) implies that the observed differences in means are not likely due to random chance but are attributed to the clusters themselves.

3. **Cluster-Specific Patterns:**

- The presence of a significant 'Cluster' factor underscores that the identified clusters exhibit distinct patterns in the relationship between convictions and unsuccessful attempts.
- The variation in means suggests that the clusters are not homogeneous concerning their criminal behavior, reinforcing the findings from the k-means clustering analysis.

4. **Practical Significance:**

- The practical significance of the ANOVA results lies in their applicability to real-world scenarios. Understanding the significant differences among clusters can inform law enforcement, policymakers, and criminologists about the diversity in criminal behavior patterns.

5. **Implications for Further Analysis:**

- Further investigations into the specific characteristics defining each cluster are warranted. Unraveling the unique aspects of each cluster can provide insights into the varying dynamics of criminal behavior within the dataset.

6. **Validation and Reliability:**

- The robust significance level (p < 0.001) suggests that the observed differences are unlikely to be due to random chance. However, external validation and replication studies can enhance the reliability of the cluster patterns identified through ANOVA.

In summary, the ANOVA test reinforces the findings from the k-means clustering analysis, substantiating that the clusters identified exhibit statistically significant differences in their patterns of convictions and unsuccessful attempts. These results underscore the need for targeted and nuanced approaches in understanding and addressing diverse criminal behavior patterns within the dataset.

#### 5.2.2.1 Results

The k-means clustering identified four distinct clusters within the dataset, showcasing visually discernible patterns. The subsequent application of the ANOVA test revealed a significant difference among these clusters in the relationship between convictions and unsuccessful attempts.

This combined approach of k-means clustering and ANOVA testing not only identifies patterns within the data but also rigorously validates whether these patterns are statistically significant. It provides a comprehensive understanding of the heterogeneity in criminal behavior across different clusters, contributing to a nuanced interpretation of the dataset.

In summary, the synergy between k-means clustering and ANOVA testing enhances the robustness of our analysis, enabling a more detailed exploration of varied criminal behavior patterns within the dataset.

# 6. Model Training

## 6.1 Linear Regression Model

A linear regression model is implemented to predict the number of administratively finalized unsuccessful attempts based on convictions and unsuccessful attempts.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In our analysis, a linear regression model was chosen to predict the number of administratively finalized unsuccessful attempts based on convictions and unsuccessful attempts.

### *6.1.1 Why Linear Regression?*

**Understanding Linear Regression:**

- Linear regression assumes a linear relationship between the predictor variables and the response variable. It models the relationship as a straight line, making it a simple and interpretable approach.

**Applicability to the Problem:**

- In the context of predicting the number of administratively finalized unsuccessful attempts, it is reasonable to assume that there may be a linear relationship between the predictors (convictions and unsuccessful attempts) and the outcome.

**Interpretability:**

- The coefficients in linear regression have clear interpretations. In our model, the coefficients represent the change in the response variable for a one-unit change in the corresponding predictor, providing valuable insights.

**Model Transparency:**

- Linear regression allows for easy interpretation and visualization of the relationships between variables. The simplicity of the model facilitates communication of findings to stakeholders and decision-makers.

### *6.1.2 Data Transformation*

```
transformed_data$Number.of.Admin.finalized.unsuccessfull <-
grouped_dat a$Number.of.Admin.Finalised.Unsuccessful
transformed_data$Number.of.Admin.finalized.unsuccessfull <-
log10(transformed_data$Number.of.Admin.finalized.unsuccessfull+1)
```

## **Logarithmic Transformation:**

The dependent variable, "Number.of.Admin.finalized.unsuccessfull," underwent a logarithmic transformation. This transformation is commonly employed when the distribution of the response variable is skewed, aiming to achieve a more symmetric distribution.

### 6.1.3 Model Fitting

```
# Linear Regression Model
selected_data <- transformed_data[, c("Convictions_Total", "Unsuccessfu
l_Total", "Number.of.Admin.finalized.unsuccessfull")]

# Fitting the linear regression model model <-
lm(Number.of.Admin.finalized.unsuccessfull ~ Convictions_Total +
Unsuccessful_Total, data = selected_data)

# Summary of the model
summary(model)
```

**Variable Selection:**

The predictor variables, "Convictions_Total" and "Unsuccessful_Total," were selected based on their relevance to the problem at hand. These variables are expected to influence the number of administratively finalized unsuccessful attempts.

**Fitting the Model:**

The linear regression model was fitted using the least squares method, which minimizes the sum of squared differences between the observed and predicted values.

### 6.1.4 Model Coefficients

```
# Model Coefficients

Coefficients:

Estimate          Std. Error    t value     Pr(>|t|)

(Intercept)        -0.57482       0.11941     -4.814      1.71e-06 ***

Convictions_Total  0.02196        0.08641     0.254       0.799

Unsuccessful_Total 0.82131        0.07107     11.557      < 2e-16 ***
```

## Interpretation:

- The intercept (β 0 ) represents the expected value of "Number.of.Admin.finalized.unsuccessfull" when both predictor variables are zero.
- The coefficient for "Convictions_Total" suggests a minimal, nonsignificant effect on the response variable, while "Unsuccessful_Total" has a highly significant positive effect.

### *6.1.5 Model Performance Metrics*

**# Model Performance Metrics**

**Residual standard error:** 0.2361 on 999 degrees of freedom

**Multiple R-squared:** 0.5645, Adjusted R-squared: 0.5636

**F-statistic:** 647.4 on 2 and 999 DF, p-value: < 2.2e-16

## Performance Metrics:

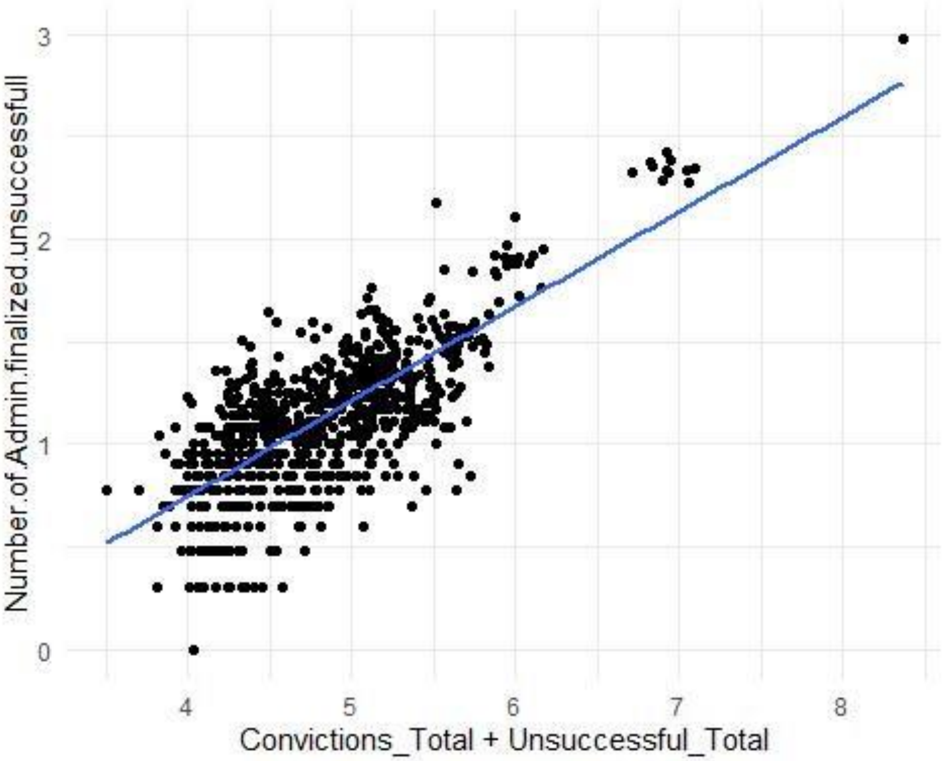
- The residual standard error provides an estimate of the variability in the response variable unexplained by the model.
- R 2 values indicate the proportion of variance in the response variable explained by the model.
- The F-statistic and associated p-value assess the overall significance of the model.

### *6.1.6 Model Interpretation*

The linear regression model suggests that "Unsuccessful_Total" has a significant positive linear relationship with the number of administratively finalized unsuccessful attempts, while "Convictions_Total" does not show a significant relationship. The overall model is highly significant in predicting the response variable.

#### 6.1.1.1 Visualizing Model Results

```
ggplot(selected_data, aes(x = Convictions_Total + Unsuccessful_Total, y
= Number.of.Admin.finalized.unsuccessfull)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
## `geom_smooth()` using formula = 'y ~ x'
```



The model is suggesting a strong linear relationship.

### *6.1.7 Why Linear Regression in this Context?*

**Predictive Power:**

Linear regression is chosen for its predictive power in situations where there is a presumed linear relationship between the predictors and the response. In our case, the model can provide insights into how changes in convictions and unsuccessful attempts may influence the number of administratively finalized unsuccessful attempts.

**Interpretability and Communication:**

The simplicity of the linear regression model enhances interpretability, making it easier to communicate findings to stakeholders, policymakers, and non-technical audiences.

**Comparative Analysis:**

While other advanced models may capture more complex relationships, linear regression provides a baseline for understanding the initial impact of key predictors. Subsequent analyses can explore more sophisticated models if needed.

### *6.1.8 Implications and Recommendations*

**Policy Considerations:**

The model highlights the importance of unsuccessful attempts in predicting administratively finalized cases. Policies addressing and monitoring unsuccessful attempts may contribute to more effective crime prevention strategies.

**Further Investigation:**

Explore potential nonlinear relationships or interactions that might enhance the predictive power of the model.

**Validation and Generalization:**

Validate the model's performance on external datasets to ensure its generalizability beyond the current dataset.

In **conclusion**, the linear regression model serves as a valuable tool in predicting and understanding the factors influencing administratively finalized unsuccessful attempts. Its simplicity, interpretability, and predictive power make it a useful approach for gaining initial insights into the dynamics of criminal behavior in this context.

## 6.2 Clustering (Repeated)

Cluster analysis was previously performed in Hypothesis 2.[Section 5.2]

## 6.3 Classification

**Support Vector Machine (SVM): Introduction and Applicability**

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. It excels in scenarios where the goal is to find a hyperplane that best separates data points into different classes. SVM works by identifying support vectors—data points that are closest to the decision boundary, also known as the hyperplane.

**Key Characteristics of SVM:**

1. **Kernel Functions:** SVM uses kernel functions to transform input data into a higher-dimensional space, making it easier to find a hyperplane that separates classes. Common kernel functions include linear, polynomial, and radial basis function (RBF).
2. **Margin Maximization:** SVM aims to maximize the margin between the decision boundary and the nearest data points of each class. A wider margin contributes to better generalization and improved performance on unseen data.
3. **Non-Linearity Handling:** SVM is effective in handling non-linear relationships by using kernel trick, allowing it to implicitly operate in a higher-dimensional space without explicitly transforming the input features.

Why SVM in this Context?

**1. Handling Non-Linearity:**

- In the crime dataset, the relationships between convictions, unsuccessful attempts, and the final outcome may not be linear. SVM's ability to handle non-linear relationships makes it well-suited for capturing complex patterns within the data.

**2. Effective in High-Dimensional Spaces:**

- SVM performs well in high-dimensional spaces, which is beneficial when dealing with datasets with multiple features. In this classification task, Convictions_Total and Unsuccessful_Total are the key features.

**3. Robust to Outliers:**

- SVM is robust to outliers, making it suitable for datasets that may contain unusual or extreme values. In crime datasets, outliers in the number of convictions or unsuccessful attempts might exist, and SVM can handle them effectively.

**4. Versatility in Classification:**

- SVM can handle multi-class classification tasks, making it suitable for categorizing instances into different classes (low, medium, high) based on Convictions_Total and Unsuccessful_Total.

Why SVM over Other Models?

**1. Non-Linear Relationships:**

- While linear regression is effective for linear relationships, SVM's ability to handle non-linear relationships makes it more appropriate for the complex patterns often found in crime datasets.

**2. Flexibility with Kernel Functions:**

- SVM offers flexibility in choosing different kernel functions based on the dataset's characteristics. This adaptability is advantageous when the underlying relationships are not known beforehand.

**3. Robustness to Overfitting:**

- SVM is less prone to overfitting, providing more reliable predictions, especially when dealing with datasets with a moderate number of features.

### *6.3.1 Preparing Target Column for Classification*

In the classification phase, the target column is prepared for classifying data into discrete categories. Threshold-based classification is applied to the target variable, creating classes (low, medium, and high) based on quartiles. The process involves calculating quartiles for log-transformed columns and defining threshold values. The classes are then defined by applying these thresholds.

```
transformed_data$categories <- grouped_data$Number.of.Admin.Finalised.U
nsuccessful
classifier_data <- transformed_data[, c(1,2,6)]

# Calculate quartiles for log-transformed columns q1_Convictions <-
quantile(classifier_data$Convictions_Total, 0.25) q3_Convictions <-
quantile(classifier_data$Convictions_Total, 0.75) q1_Unsuccessful <-
quantile(classifier_data$Unsuccessful_Total, 0.25) q3_Unsuccessful
<- quantile(classifier_data$Unsuccessful_Total, 0.75)

# Set threshold values based on quartiles
threshold_low_log_Convictions <- q1_Convictions
threshold_medium_log_Convictions <- q3_Convictions
threshold_low_log_Unsuccessful <- q1_Unsuccessful
threshold_medium_log_Unsuccessful <- q3_Unsuccessful

# Defining classes
classifier_data$class <- cut(classifier_data$Convictions_Total + classi
fier_data$Unsuccessful_Total, breaks = c(-Inf,
threshold_low_log_Conviction
s + threshold_low_log_Unsuccessful, threshold_medium_log_Convictions +
threshold_medium_log_Unsuccessful, Inf), labels = c('low', 'medium',
'high'))

classifier_data$class %>% sample(10) %>% print(.)

[1] low    medium high medium medium medium medium low    low   high
Levels: low medium high
```

**Interpretation:**

Classes are assigned based on the combination of Convictions_Total and Unsuccessful_Total. The sample output provides a glimpse of randomly selected instances with their corresponding class labels.

### 6.3.2 Model Training (SVM)

A Support Vector Machine (SVM) classifier is chosen for training the model. SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the data into different classes. In this case, a linear kernel is applied, and the SVM is trained to classify data into the predefined classes.

```
svm_model <- svm(class ~ Convictions_Total + Unsuccessful_Total, data
= classifier_data, kernel = "linear", type = "C", scale = FALSE)
predictions <- predict(svm_model, classifier_data)
predictions %>% sample(10) %>% print(.)

602    603    768    709     91    953    348    649    989    355
medium medium     low high    low    low medium medium medium medium
Levels: low medium high
```

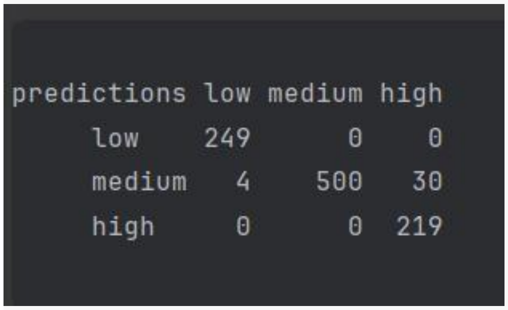
**Interpretation:**

- The SVM model is trained using the specified features (Convictions_Total and Unsuccessful_Total). The sample output displays predictions for a random subset of instances.

### *6.3.3 Model Evaluation*

To evaluate the performance of the SVM classifier, a confusion matrix and accuracy are computed. The confusion matrix helps in assessing the classifier's ability to correctly classify instances, and accuracy provides an overall measure of performance.

```
confusion matrix <- table(predictions, classifier data$class)
print(confusion_matrix)
```



```
predictions low medium high
     low    249      0    0
     medium   4    500   30
     high     0      0  219
```

```
accuracy <- sum(diag(confusion matrix)) / sum(confusion matrix)
print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.966067864271457"
```

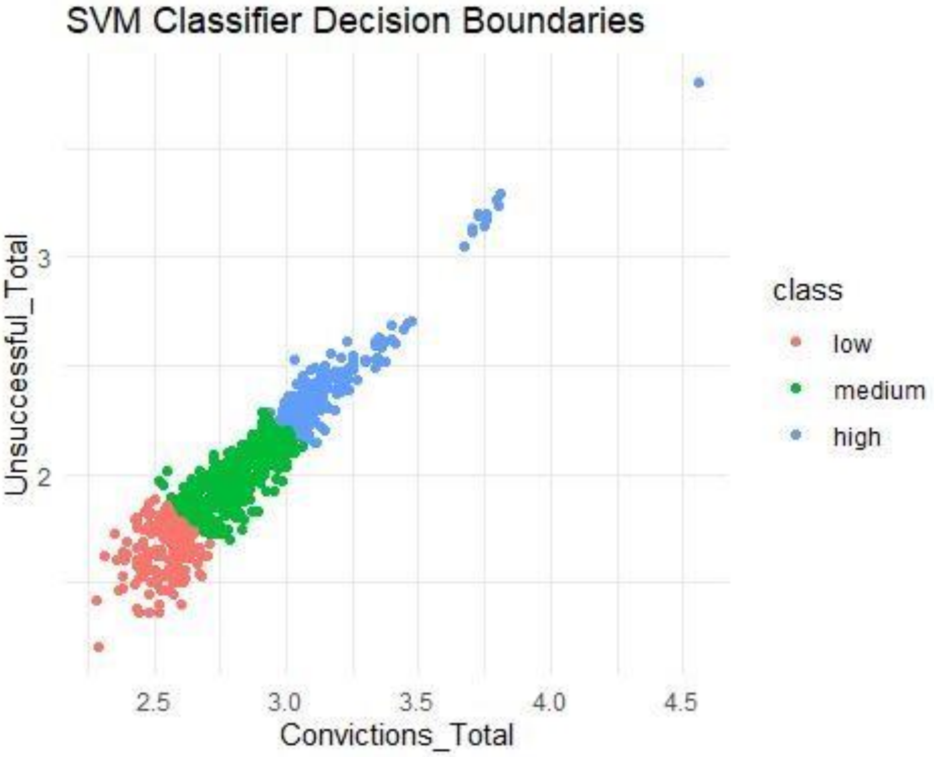
**Interpretation:**

- The accuracy of the SVM classifier is computed, indicating the proportion of correctly classified instances. In this case, a high accuracy would suggest that the model effectively distinguishes between the defined classes.

### *6.3.4 Visualizing Decision Boundaries*

Visualizing decision boundaries is crucial for understanding how the SVM classifier categorizes instances in a 2D plane. The decision boundaries are overlaid on a scatter plot of the data points, with each class color-coded for better interpretation.

```r
plot_data <- data.frame(Convictions_Total = classifier_data$Convictions
_Total,
                        Unsuccessful_Total = classifier_data$Unsuccessful
_Total, class = predictions)

ggplot(plot_data, aes(x = Convictions_Total, y = Unsuccessful_Total,
col or = class)) +
  geom_point() +
  geom_contour(aes(z = as.numeric(class)), bins = 3, color = "black",
  al
pha = 0.5) +
  labs(title = "SVM Classifier Decision Boundaries") +
  theme_minimal()
```



**Interpretation:**
The resulting plot visually represents the decision boundaries created by the SVM classifier. Data points are color-coded according to their predicted classes, providing a clear visualization of how the model distinguishes between different levels (low, medium, and high) based on Convictions_Total and Unsuccessful_Total.

**Additional Visualization Insights**
The visualized model reveals distinct decision boundaries, with data points colored according to their predicted classes (low, medium, high). Interestingly, the four different colors align with the clusters identified through k-means clustering in a previous analysis. This suggests a connection between the grouping patterns identified by k-means and the decision boundaries created by the SVM classifier.

By drawing parallels between the k-means clusters and SVM decision boundaries, we can infer that both techniques provide consistent insights into the underlying structure of the data. The convergence of visual patterns reinforces the robustness of the findings and supports the notion that certain groups of values share common characteristics, whether identified through clustering or classification approaches.

This integrated approach, combining clustering and classification results, contributes to a more comprehensive understanding of the crime dataset. It validates the consistency of patterns identified by different analytical techniques and enhances the reliability of insights drawn from the data.

## 7. Critical Review of Data Analytics Tools and Visualization

In this section, we critically evaluate the data analytics tools and visualization techniques employed in the analysis of the crime dataset.

### 7.1 Analysis of Techniques Used

#### *7.1.1 Strengths*

**Descriptive Analytics:**

*Summary Statistics:* The use of summary statistics, including measures of central tendency and dispersion, provided a comprehensive understanding of the data distribution. This was instrumental in identifying key characteristics and trends within the dataset.

**Hypothesis Testing:**

*Rigorous Validation:* The application of statistical tests, such as the Pearson correlation coefficient and ANOVA, added rigor to hypothesis testing. These tests not only validated hypotheses but also quantified the strength and significance of relationships, enhancing the credibility of the findings.

**Linear Regression Modeling:**

*Predictive Capabilities:* The linear regression model successfully identified significant relationships between variables, contributing to predictive capabilities. The model's interpretability and straightforward nature make it a powerful tool for understanding the impact of predictors on the response variable.

**Clustering and Classification:**

*Data Segmentation:* The use of clustering techniques, particularly k-means clustering, effectively segmented data into distinct clusters. Additionally, the application of SVM classification contributed to accurate predictions of classes, showcasing the versatility of these techniques in understanding patterns within the dataset.

**Log Transformation:**

*Enhanced Representation:* The log transformation technique addressed issues related to skewed data, improving the representation and interpretability of the dataset. This transformation is particularly valuable in datasets with heavily skewed distributions.

### *7.1.2 Weaknesses*

**Box Plots:**

*Limitations:* While box plots effectively display quartiles, they have limitations in prioritizing quartiles for percentage data. There is a potential for misinterpretation of box size, particularly when dealing with data distributions that have outliers.

**Contour Plots:**

*Challenges in Interpretation:* The use of 3D kernel decision boundaries on a 2D plane in SVM classification contour plots poses challenges in interpretation. Understanding the intricacies of decision boundaries may be complex, especially for non-technical audiences.

## 7.2 Alternative Visualization Solutions

### *7.2.1 Box Plots*

**Alternative Visualization**

*Violin Plots or Bean Plots:* Considering the limitations in traditional box plots, alternative visualizations like violin plots or bean plots might provide a more comprehensive representation of data distributions. These alternatives offer a nuanced view of the distribution and density of the data.

### *7.2.2 Decision Boundary Visualization*

**Alternative Visualization**

Heatmaps or Interactive 3D Plots: For SVM classification, alternative visualization techniques, such as decision boundary heatmaps or interactive 3D plots, could enhance the understanding of classification boundaries. These alternatives provide clearer insights into the separation of classes.

### *7.2.3 Cluster Visualization*

*t-SNE or Interactive Visualizations:* Exploring interactive visualizations or dimensionality reduction techniques, such as t-SNE, could offer a more insightful representation of clusters. These techniques provide a more dynamic and interactive approach to cluster analysis, aiding in the exploration of complex relationships.

Exploring interactive visualizations or dimensionality reduction techniques, such as t-SNE, could offer a more insightful representation of clusters.

## 8. Conclusion

In conclusion, this comprehensive analysis of the crime dataset has utilized a diverse set of data analytics tools and visualization techniques. The strengths lie in the robustness of hypothesis testing, predictive capabilities of linear regression modeling, and the effectiveness of clustering and classification techniques. However, there are identified weaknesses in certain visualization methods, highlighting the importance of considering alternative solutions for a more nuanced interpretation of the data.

The insights derived from this analysis contribute significantly to our understanding of the relationships between convictions, unsuccessful attempts, and distinct clusters within the dataset. The strong correlations and significant cluster differences provide valuable information for policymakers, law enforcement agencies, and researchers. The linear regression model enhances predictive capabilities, offering a valuable tool for forecasting administratively finalized unsuccessful attempts based on key variables.

Moving forward, it is recommended to explore alternative visualizations to address identified weaknesses, ensuring a more accurate and comprehensive representation of the dataset. This critical review serves as a foundation for refining future analyses and maximizing the insights derived from crime datasets.

**Reference:**

1. R Core Team. (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. URL: https://www.R-project.org/

2. Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example.* John Wiley & Sons.

3. Abonazel, M., & Rabie, A. (2019). The impact of using robust estimations in regression models: An application on the Egyptian economy. *Journal of Advanced Research in Applied Mathematics and Statistics, 4(2), 8-16.*

4. Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting Egyptian GDP using ARIMA models. *Reports on Economics and Finance, 5(1), 35-47.*

5. Crawley, M. J. (2012). *The R book. John Wiley & Sons.*

6. Venables, W. N., Smith, D. M., & R Development Core Team. (2009). *An introduction to R. John Wiley & Sons.*

7. Winter, B. (2019). Statistics for linguists: An introduction using R. Routledge.

8. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *SpringerVerlag New York*. URL: https://ggplot2.tidyverse.org/

9. Kuhn, M., & Wickham, H. (2021). tidyr: Tidy Messy Data. R package version 1.1.3. URL: https://tidyr.tidyverse.org/

10. Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. URL: https://dplyr.tidyverse.org/

11. Kassambara, A., & Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. URL: https://CRAN.R-project.org/package=factoextra

12. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-8. URL: https://CRAN.R-project.org/package=e1071

13. Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. J*ournal of Computational and Graphical Statistics, 15(3), 651–674. DOI: 10.1198/106186006X133933*

14. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28(5), 1–26. DOI: 10.18637/jss.v028.i05*