ASSIGNMENT ON DATA ANALYSIS AND VISUALISATION PRINCIPLES

JOHNSON OHAKWE

4111678

CT7202

SCHOOL OF COMPUTING AND ENGINEERING

UNIVERSITY OF GLOUCESTERSHIRE, UNITED KINGDOM

MODULE TUTOR:

BHUPESH MISHRA

MAY, 2022

TABLE OF CONTENT

CHAPTER ONE

1.1 INTRODUCTION

Crime is an intentional demonstration that inflicts any kind of damage, harm to or loss of property, and is illegal. There are bunches of various sorts of crime and almost everybody will encounter a crime sooner or later in their lives. As distributed by Oxford University Press in 2016, Criminologists have long acknowledged that monetary conditions and social uniqueness expect a huge part in why explicit individuals become drawn to crime (Roger et al., 2017).

Homicide, drug abuse, fraud, forgery, motoring offence, burglary, robbery, and sexual abuse are all different types of crime in the UK (Home Office, 2013), and research has shown that there has been an increase in some of these crimes (National Crime Agency, 2022).

Lately, there is by all accounts an expansion in crime by and large and these violations carried out are most times not appropriately recorded, if appropriately recorded, they are not very much displayed either by utilizing a Linear model; for anticipating or estimating future event (future crime values), Classification model; to combine the volume of crime information so that likenesses and contrasts can be immediately perceived, and Clustering model; for gathering important crime information into bunches and picks fitting outcomes in view of various methods, which will be useful in making future practical and measurable strategies that will help in overseeing wrongdoing in the United Kingdom.

1.2 OBJECTIVES OF THE STUDY

The objectives of this study are:

- To clean the crime data sets for 2014, 2015, 2016, and 2017.
- To manipulate the cleaned crime data sets into our own choice for use.
- To perform descriptive analysis on the data sets.
- To explore and visualize the data sets.
- To create a new linear model function in R called the Akash model and compare it with the inbuilt linear model function in R (lm) using the data set.
- To build a linear model which will be used to predict the Number of Offences Against the Person Convictions using the Number of Sexual Offences Convictions and the Number of Motoring Offences Convictions as independent variables and the Number of Offences Against the Person Convictions as the dependent variable.
- To compare K-Nearest Neighbor (KNN) Algorithm and Decision tree algorithm while performing Classification.
- To study the K-Means Algorithm while performing Clustering.
- To test several hypotheses for the three models. (Linear model, Clustering, and Classification)

CHAPTER TWO

METHODOLOGY

This chapter talks about the methodology for statistical modelling of the crime data sets collected.

2.1 SOURCE OF DATA

The crime data sets used in this study are secondary data and it is sourced from Crown Prosecution Service Case Outcomes by Principal Offence Category. (link: https://data.gov.uk/dataset/89d0aef9-e2f9-4d1a-b779-5a33707c5f2c/crown prosecution-service-case-outcomes-by-principal-offence-category-data)

2.2 LINEAR MODEL

Linear regression discusses the relationship and the effect of independent variables on the dependent variable. Here we will discuss the assumptions of linear regression and the methods used to create the Akash model for regression analysis and the inbuilt R function for linear regression.

2.2.1 ASSUMPTIONS OF LINEAR REGRESSION

Assumptions of linear regression:

i. Normality.
ii. Linearity.
iii. Homoscedasticity.
iv. Multicollinearity.
v. Autocorrelation.

2.2.1.1 TEST FOR HOMOSCEDASTICITY

This assumption talks about equal variance across the error terms. The Breusch-Pagan test (Wikipedia, 2021) will be used to test for equal variance assumption. The following Lagrange multiplier (LM) yields the test statistic for the Breusch-Pagan test

The test statistic.

$$LM = \left(\frac{\partial \ell}{\partial \theta}\right)^T \left(-E\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta^t}\right]\right)^{-1} \left(\frac{\partial \ell}{\partial \theta}\right) \tag{2.1}$$

This test can be carried out by means of the accompanying three strategies:

Step 1: Apply OLS in the model

$$y_i = X_i\beta + \varepsilon_i, \quad i = 1, ..., \mathrm{n}$$

Step 2: Compute the regression residuals, ε_i, square them, and divide by the Maximum Likelihood estimate of the error variance from the Step 1 regression, to get what Breusch and Pagan call g_i:

$$g_i = \varepsilon_i^2 \Big/ \sigma^2 \quad , \qquad \sigma^2 = \sum \varepsilon_i^2 \Big/ n \tag{2.2}$$

Step 3: Estimate the auxiliary regression

$$g_i = \gamma_1 + \gamma_2 z_{2i} + \ldots + \gamma_p z_{pi} + \eta_i \tag{2.3}$$

Step 4: The LM test statistic is then half of the explained sum of squares from the auxiliary regression in Step 3:

$$LM = \tfrac{1}{2}\left(TSS - SSR\right) \tag{2.4}$$

where TSS is the sum of squared deviations of the g_i from their mean of 1, and SSR is the sum of squared residuals from the auxiliary regression.

The hypothesis for the test is stated below as:

H_0: $\sigma_1^2 = \sigma_2^2 =, \ldots, \sigma_n^2 = 0$ (Homoscedasticity)

H_1: $\sigma_1^2 \neq \sigma_2^2 \neq, \ldots, \sigma_n^2 \neq 0$ (Heteroscedasticity)

Decision rule:
If the p-value is greater than 0.05, we then accept H_0, but if the p-value is less than 0.05, we reject H_0 and accept H_1.

2.2.1.2 TEST FOR AUTOCORRELATION

We test for autocorrelation to find out if the error term (U_t) is correlated in the regression. We will use the Durbin-Watson [DW] statistics to test for the presence of autocorrelation.

Durbin-Watson statistic.

$$d^* = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{2.5}$$

Where.
e_t, t=1, 2, . . ., T

Hypothesis:

H_0: $\rho = 0$

H_1: $\rho > 0$

Decision rule:
If the p-value is greater than 0.05, we then accept H_0, but if the p-value is less than 0.05, we reject H_0 and accept H_1.

2.2.1.3 TEST FOR NORMALITY

The test will be conducted to find out if the error terms are normally distributed with zero mean and constant variance i.e $\mu \sim N(0, \sigma^2)$. We shall be using the Shapiro-Wilk's test (Wikipedia, 2022).

The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2.6}$$

where,

$x_{(i)}$ is the *i*th order statistic

$\bar{x} = \left(x_i, ..., x_n\right)/n$ is the sample mean.

The coefficients a_i, are given by.

$$\left(a_i, ..., a_n\right) = \frac{m^T V^{-1}}{C} \tag{2.7}$$

Where C is a vector norm.

$$C = \| V^{-1} m \| = \left(m^T V^{-1} V^{-1} m\right)^{1/2} \tag{2.8}$$

And m is a vector.

$$m = \left(m_1, ..., m_n\right)^T \tag{2.9}$$

is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.

The hypothesis is

H_0: $\mu_1 = 0$ Normally distributed

H_1: $\mu_1 \neq 0$ Not normally distributed

Decision rule:
If the p-value is greater than 0.05, we then accept H_0, but if the p-value is less than 0.05, we reject H_0 and accept H_1. The test statistic follows a chi-square distribution.

2.2.1.4 TEST FOR MULTICOLLINEARITY

Multicollinearity assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF).

Where,

$$VIF = \frac{1}{(1-R^2)} \tag{2.10}$$

If the value of VIF is <10, it is acceptable and can conclude that there is no multicollinearity among the explanatory variables. (Douglas et al. 2012).

2.2.2 AKASH LINEAR MODEL FUNCTION

This function was created using the matrix approach to linear regression analysis.

Considering a multiple linear regression straight line equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e \tag{2.11}$$

Where; β_0 = Intercept

β_1 = Coefficient of X_1

β_2 = Coefficient of X_2

β_k = Coefficient of X_k

e = Error term

Estimation of the model parameters using the matrix approach we have.

$$X^T X = \begin{pmatrix} n & \sum X_1 & \sum X_2 \ldots & \sum X_k \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \ldots & \sum X_1 X_k \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \ldots & \sum X_2 X_k \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \sum X_k & \sum X_1 X_k & \sum X_2 X_k \ldots & \sum X_k^2 \end{pmatrix} \quad (2.12)$$

$$X^T Y = \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ . \\ . \\ . \\ \sum X_k Y \end{pmatrix} \quad (2.13)$$

$$X^T Y = (X^T X)\beta \quad (2.14)$$

To solve for β, we have.

$$\beta = (X^T X)^{-1}(X^T Y) \quad (2.15)$$

Where:

$$(X^T X)^{-1} = \frac{1}{\det (X^T X)} \text{adj } (X^T X) \quad (2.16)$$

$$\beta = \begin{pmatrix} n & \sum X_1 & \sum X_2 \ldots & \sum X_k \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \ldots & \sum X_1 X_k \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \ldots & \sum X_2 X_k \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \sum X_k & \sum X_1 X_k & \sum X_2 X_k \ldots & \sum X_k^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ . \\ . \\ . \\ \sum X_k Y \end{pmatrix} \quad (2.17)$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} \tag{2.18}$$

2.2.3 INBUILT R LINEAR MODEL FUNCTION

The design was inspired by the S function of the same name described in Chambers (1992). The implementation of the model formula by Ross Ihaka was based on Wilkinson & Rogers (1973).

2.2.4 ROOT MEAN SQUARE ERROR (RMSE)

RMSE is regularly used as a measure to distinguish differences between values (sample or population values) predicted by a model or an estimator and the values observed (Wikipedia 2021).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - x_i\right)^2}{N}} \tag{2.19}$$

2.3 CLASSIFICATION

Classification is an arranged set of related categories used to group data according to its similarities. Here we shall discuss the methodology of the K-Nearest Neighbour Algorithm and Decision tree algorithm.

2.3.1 K-NEAREST NEIGHBOUR ALGORITHM (KNN)

KNN falls in the supervised learning algorithms. This implies that we have a dataset with titled training measurements (x, y) and would want to find the link between x and y. Our goal is to discover a function h:X→Y so that having an unknown observation x, h(x) can positively predict the identical output y. First, we will discuss the working of the KNN classification algorithm. In the classification problem, the K-nearest neighbour algorithm basically said that for a given value of K algorithm will find the K nearest neighbour of an unseen data point and then it will assign the class to the unseen data point by having the class which has the highest number of data points out of all classes of K neighbours. For distance metrics, we will use the Euclidean metric.

$$d\left(x, x^{\mathrm{I}}\right) = \sqrt{\left(x_1 - x_1^{\mathrm{I}}\right)^2 + \ldots + \left(x_n - x_n^{\mathrm{I}}\right)^2} \tag{2.20}$$

Finally, the input x gets assigned to the class with the largest probability.

$$p\left(y = j / X = x\right) = \frac{1}{K}\sum_{i \in \mathrm{A}} \mathrm{I}\left(y^{(i)} = j\right) \tag{2.21}$$

2.3.2 DECISION TREE ALGORITHM

The decision tree algorithm is one of the most well-known machine learning algorithms. It is a managed or supervised machine learning algorithm, used for both classification and regression tasks. In the decision tree algorithm, Entropy and Gini index is calculated to check the amount of information required in a sample and to measure inequality in the sample.

$$Entropy = -\sum_{i=1}^{n} p_i * \log\left(p_i\right) \tag{2.22}$$

$$Gininindex = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.23}$$

2.4 CLUSTERING

A process of organizing objects into groups such that data points in the same groups are like the data points in the same group. A cluster is a collection of objects where these objects are similar and dissimilar to the other cluster. Here we shall discuss the methodology of the K-Means Algorithm

2.4.1 K-MEANS ALGORITHM

K-Means clustering is a type of unsupervised learning. The main aim of this algorithm is to discover groups in data and the number of groups is addressed as K. This clustering algorithm separates data into the most appropriate group based on the information the algorithm already has. Data is separated in K different clusters, which are usually chosen to be far enough apart from each other spatially, in Euclidian_Distance, to be able to produce effective data mining results. Each cluster has a centre, called the centroid**,** and a data point is clustered into a certain cluster based on how close the features are to the centroid. K-means algorithm iteratively minimizes the distances between every data point and its centroid to find the most optimal solution for all the data points.

Given two points $A(x_2, x_1)$ and $\mathrm{B}(\mathrm{y}_2, \mathrm{y}_1)$ the Euclidean distance is estimated using

$$D(A, B) = \sqrt{\left(x_2 - x_1\right) + \left(y_2 - y_1\right)} \tag{2.24}$$

The centroid of each cluster v_j is estimated using

$$v_j = \frac{1}{m} \sum_{x \in v_j} X \tag{2.25}$$

The objective function for the K-means clustering algorithm is the squared error function:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(\| x_i - v_j \| \right)^2 = 1 \tag{2.26}$$

Where,

$\| x_i - v_j \|$ is the Euclidian distance between a point, x_i, and a centroid, v_j, iterated over all *k* points in the i^{th} cluster, for all n clusters. In simpler terms, the objective function attempts to pick centroids that minimize the distance to all points belonging to its respective cluster so that the centroids are more symbolic of the surrounding cluster of data points.

2.5 CONFUSION MATRIX

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix looks like the square matrix below:

$$\begin{array}{lcc} & \textit{Actually Positive (1)} & \textit{Actually Negative (2)} \\ \textit{Predicted Positive (1)} & \textit{True Positives (TPs)} & \textit{False Positives (FPs)} \\ \textit{Predicted Negative (2)} & \textit{False Negatives (FNs)} & \textit{True Negatives (TNs)} \end{array} \tag{2.27}$$

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN (2.28)

Where,

TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

CHAPTER THREE

DATA CLEANING AND MANIPULATION, VISUALISATION AND DESCRIPTIVE ANALYSIS

3.1 DATA CLEANING AND MANIPULATION

In this chapter, we shall import our data set, clean them, explore and give some visualisations of the data set and perform a descriptive analysis of the data set. We start by importing the monthly crime data set for 2014 from the 2014 folder.

```
$`./2014/principal_offence_category_april_2014.csv`
# A tibble: 43 x 51
   ...1    `Number of Hom~` `Percentage of~` `Number of Hom~` `Percentage of~` `Number of Off~`
   <chr>              <dbl> <chr>                       <dbl> <chr>                       <dbl>
 1 Nation~               81 85.3%                          14 14.7%                        7805
 2 Avon a~                1 100.0%                          0 0.0%                          167
 3 Bedfor~                0 -                               0 -                              69
 4 Cambri~                0 -                               0 -                              99
 5 Cheshi~                1 50.0%                           1 50.0%                        140
 6 Clevel~                0 -                               0 -                              85
 7 Cumbria                0 -                               0 -                              77
 8 Derbys~                0 -                               0 -                             151
 9 Devon ~                1 100.0%                          0 0.0%                          157
10 Dorset                 0 -                               0 -                              73
# ... with 33 more rows, and 45 more variables:
```

Let’s check the length of the months and the dimension of each month.

```
[1] 12
```

```
[1] 43 51
```

We can see that there are 12 months of data set which implies that 2014 has complete information, but there are missing months in the subsequent year and each month has 43 rows and 51 columns which means that each month has 2193 cells containing information, if we replace these missing months then our analysis will be very far from being accurate or true and we will have a high level of biasedness in our work. This is what led to the adding up of all available monthly data sets to get a year or annually data set.

As seen above in the imported data set, there are columns for percentages and they are rough and are not recognised as numeric by R, we also do not need them for our work (analysis), so we start data cleaning by removing columns with percentages. We went further to perform a data frame (or matrix) operation by adding the data for each month together to get an annually (yearly) data set, this was possible because the data for each month have the same number of rows and the same number of columns. The same thing was done for the 2015, 2016, and 2017 data sets. Below is a head view of the data for 2014.

	Number of Homicide Convictions	Number of Homicide Unsuccessful
National	731	188
Avon and Somerset	17	5
Bedfordshire	17	5
Cambridgeshire	3	0
Cheshire	11	2
Cleveland	11	3

	Number of Offences Against The Person Convictions
National	105123
Avon and Somerset	2706
Bedfordshire	906
Cambridgeshire	1188
Cheshire	2051
Cleveland	1272

Below is the monthly data set for 2015.

```
$`./2015/principal_offence_category_april_2015.csv`
# A tibble: 43 x 51
   ...1    `Number of Hom~` `Percentage of~` `Number of Hom~` `Percentage of~` `Number of Off~`
   <chr>              <dbl> <chr>                       <dbl> <chr>                       <dbl>
 1 Nation~               84 88.4%                          11 11.6%                        9554
 2 Avon a~                3 100.0%                         0 0.0%                          262
 3 Bedfor~                0 -                              0 -                              81
 4 Cambri~                0 -                              0 -                             115
 5 Cheshi~                1 100.0%                         0 0.0%                          177
 6 Clevel~                1 100.0%                         0 0.0%                          127
 7 Cumbria                0 -                              0 -                             100
 8 Derbys~                0 -                              0 -                             166
 9 Devon ~                0 -                              0 -                             169
10 Dorset                 0 -                              0 -                             101
# ... with 33 more rows, and 45 more variables:
```

Let’s check the length of the months and the dimension of each month.

```
[1] 11
```

```
[1] 43 51
```

We can observe that there is no complete information for 2015 as there is one missing month in the 2015 but each month has same number of rows and columns as the 2014 data set. We shall forge ahead to start data cleaning by removing the columns with percentages and adding up all the months together to get an annually (yearly) data set. Below is a head view of the 2015 data set.

	Number of Homicide Convictions	Number of Homicide Unsuccessful
National	763	182
Avon and Somerset	30	2
Bedfordshire	8	0
Cambridgeshire	6	0
Cheshire	8	0
Cleveland	5	1

	Number of Offences Against The Person Convictions
National	111847
Avon and Somerset	2996
Bedfordshire	953
Cambridgeshire	1228
Cheshire	2405
Cleveland	1344

	Number of Offences Against The Person Unsuccessful

Below is the monthly data set for 2016.

```
$`./2016/principal_offence_category_april_2016.csv`
# A tibble: 43 x 51
   ...1    `Number of Hom~` `Percentage of~` `Number of Hom~` `Percentage of~` `Number of Off~`
   <chr>              <dbl> <chr>                       <dbl> <chr>                       <dbl>
 1 Nation~              125 83.3%                          25 16.7%                       11455
 2 Avon a~                2 66.7%                           1 33.3%                         275
 3 Bedfor~                0 -                               0 -                              89
 4 Cambri~                2 100.0%                          0 0.0%                         151
 5 Cheshi~                1 50.0%                           1 50.0%                        280
 6 Clevel~                1 100.0%                          0 0.0%                         121
 7 Cumbria                0 -                               0 -                            116
 8 Derbys~                0 -                               0 -                            242
 9 Devon ~                4 80.0%                           1 20.0%                        267
10 Dorset                 1 100.0%                          0 0.0%                         142
# ... with 33 more rows, and 45 more variables:
```

Let’s check the length of the months and the dimension of each month.

```
[1] 10
```

```
[1] 43 51
```

We can observe that there is no complete information for 2016 as there are two missing months in the 2016 but each month has same number of rows and columns as the 2014, 2015 data sets. We shall forge ahead to start data cleaning by removing the columns with percentages and adding up all the months together to get an annually (yearly) data set. Below is a head view of the 2016 data set.

```
                  Number of Homicide Convictions Number of Homicide Unsuccessful
National                                     999                              234
Avon and Somerset                             27                               11
Bedfordshire                                  14                                2
Cambridgeshire                                13                                3
Cheshire                                      15                                2
Cleveland                                     15                                0
                  Number of Offences Against The Person Convictions
National                                                     106141
Avon and Somerset                                              2939
Bedfordshire                                                    925
Cambridgeshire                                                 1300
Cheshire                                                       2652
Cleveland                                                      1317
                  Number of Offences Against The Person Unsuccessful
```

Below is the monthly data set for 2017.

```
$`./2017/Principal_Offence_Category_Aug.csv`
# A tibble: 43 x 51
   ...1    `Number of Hom~` `Percentage of~` `Number of Hom~` `Percentage of~` `Number of Off~`
   <chr>              <dbl> <chr>                      <dbl> <chr>                       <dbl>
 1 Nation~               54 73.0%                         20 27.0%                       10056
 2 Avon a~                2 66.7%                          1 33.3%                         255
 3 Bedfor~                0 -                              0 -                             115
 4 Cambri~                1 100.0%                         0 0.0%                          108
 5 Cheshi~                3 75.0%                          1 25.0%                         263
 6 Clevel~                1 50.0%                          1 50.0%                         118
 7 Cumbria                0 -                              0 -                             102
 8 Derbys~                0 -                              0 -                             173
 9 Devon ~                1 100.0%                         0 0.0%                          195
10 Dorset                 1 50.0%                          1 50.0%                         132
# ... with 33 more rows, and 45 more variables:
```

Let's check the length of the months and the dimension of each month.

```
[1] 9
```

```
[1] 43 51
```

We can observe that there is no complete information for 2017 as there are three missing months in the 2017 but each month has same number of rows and columns as the 2014, 2015, 2016 data sets. We shall forge ahead to start data cleaning by removing the columns with percentages and adding up all the months together to get an annually (yearly) data set. Below is a head view of the 2017 data set.

```
                 Number of Homicide Convictions Number of Homicide Unsuccessful
National                                    826                             202
Avon and Somerset                            18                               6
Bedfordshire                                  5                               0
Cambridgeshire                               14                               2
Cheshire                                     10                               5
Cleveland                                     8                               1
                 Number of Offences Against The Person Convictions
National                                                     89729
Avon and Somerset                                             2382
Bedfordshire                                                   791
Cambridgeshire                                                1086
Cheshire                                                      2241
Cleveland                                                     1073
                 Number of Offences Against The Person Unsuccessful
```

3.2 VISUALISATION

In this section, we shall look mainly at plotting graphs which will represent information on our data set. The two main types of graph we will be using here is Bar chart and Pie chart. We will use Bar chart to visualise the lowest to the highest Convictions in the crime data set and we will also use Bar chart to visualise the lowest to the highest Unsuccessful Convictions in the crime data set, for the Pie chart, we will use it to visualise the lowest to the highest percentage of Convictions and the lowest to the highest percentage of Unsuccessful Convictions in the crime data set. Both Bar and Pie charts are very important in representing information or figures in a better way that even a layman can see and understand without any explanation from a professional. One of the major deficiencies of the Bar chart is that it works mainly with vectors

and does not visualise other statistical information like the mean etc, like the Box plot which gives a better picture of a data set and also shows the mean of the information. The pie chart is also an important visualisation tool because it gives a circular picture of the percentage covered by each variable in a particular data set. As we know, percentages may be misunderstood by a layman who has no knowledge of what percentage represents.

For the 2014 data set.

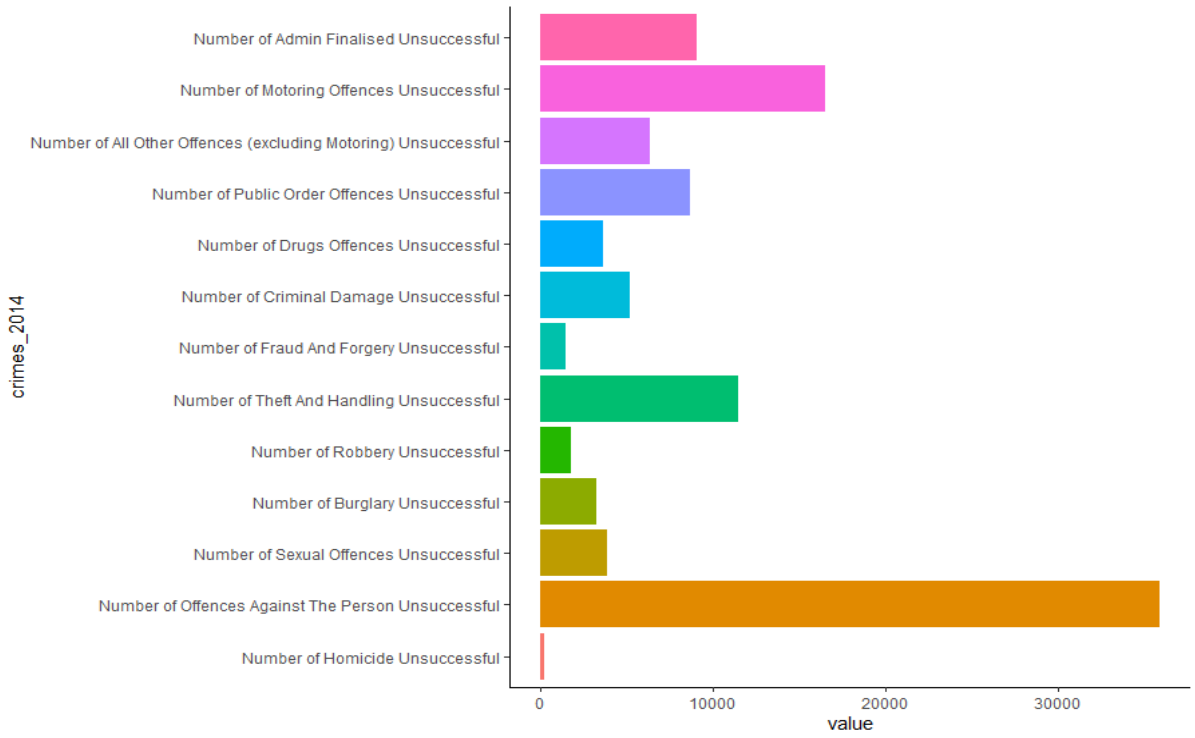

Figure 1: A Bar chart of Unsuccessful Convictions for 2014

From figure 1, we can see that the crime with the highest number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest number of unsuccessful convictions is the Number of homicides.

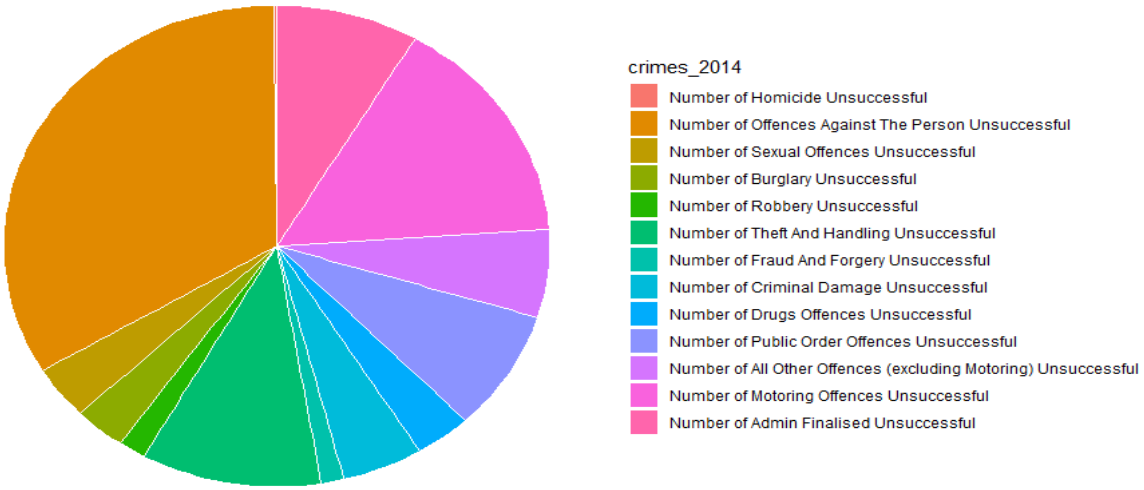

Figure 2: A Pie chart of Unsuccessful Convictions for 2014

From figure 2, we can see that the crime with the highest percentage number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest percentage number of unsuccessful convictions is the Number of homicides which is almost invisible in the plot.

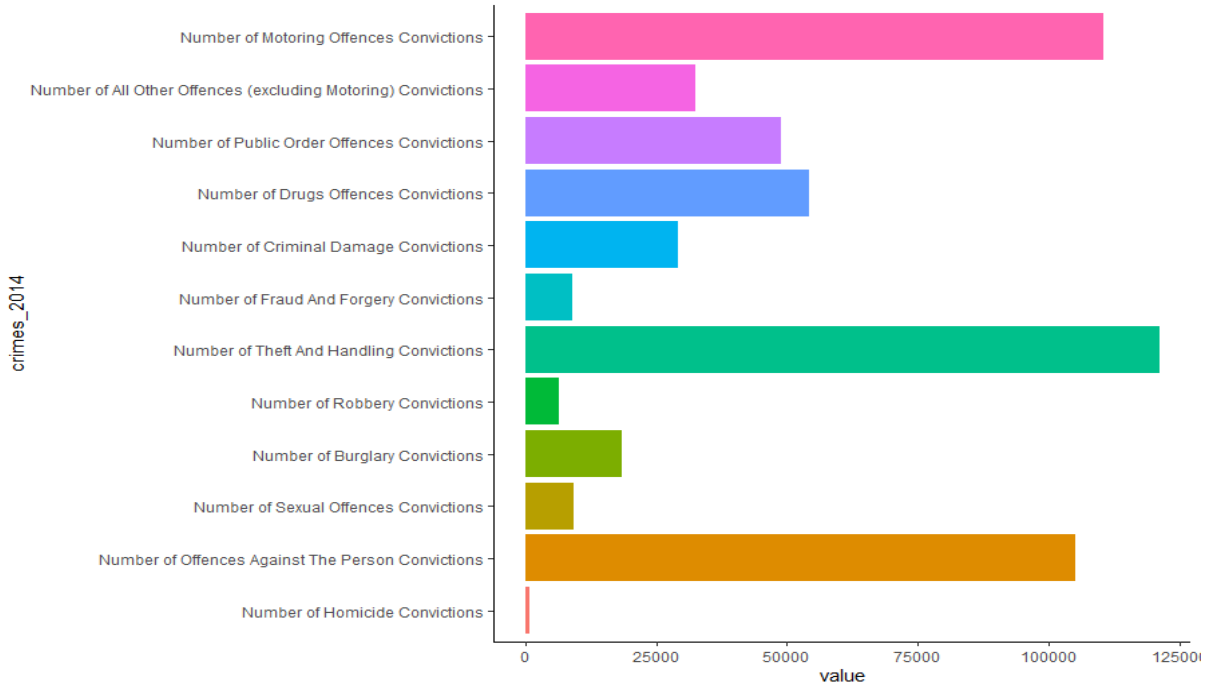

Figure 3: A Bar chart of Convictions for 2014

From figure 3, we can see that the crime with the highest number of convictions is the Number of theft and handling followed by the Number of motoring offences, the Number of offences against person, and the crime with the lowest number of convictions is the Number of homicides.

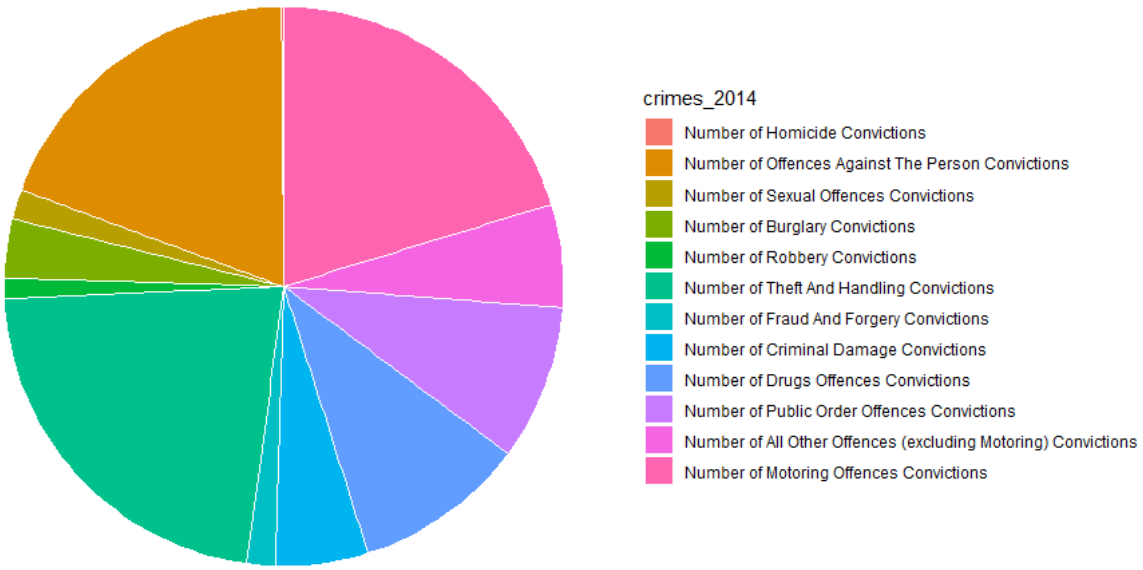

Figure 4: A Pie chart of Convictions for 2014

From figure 4, we can see that the crime with the highest percentage number of convictions is the Number of theft and handling followed by the Number of motoring offences, the Number of offences against person, and the crime with the lowest percentage number of convictions is the Number of homicides.

For 2015 data set.

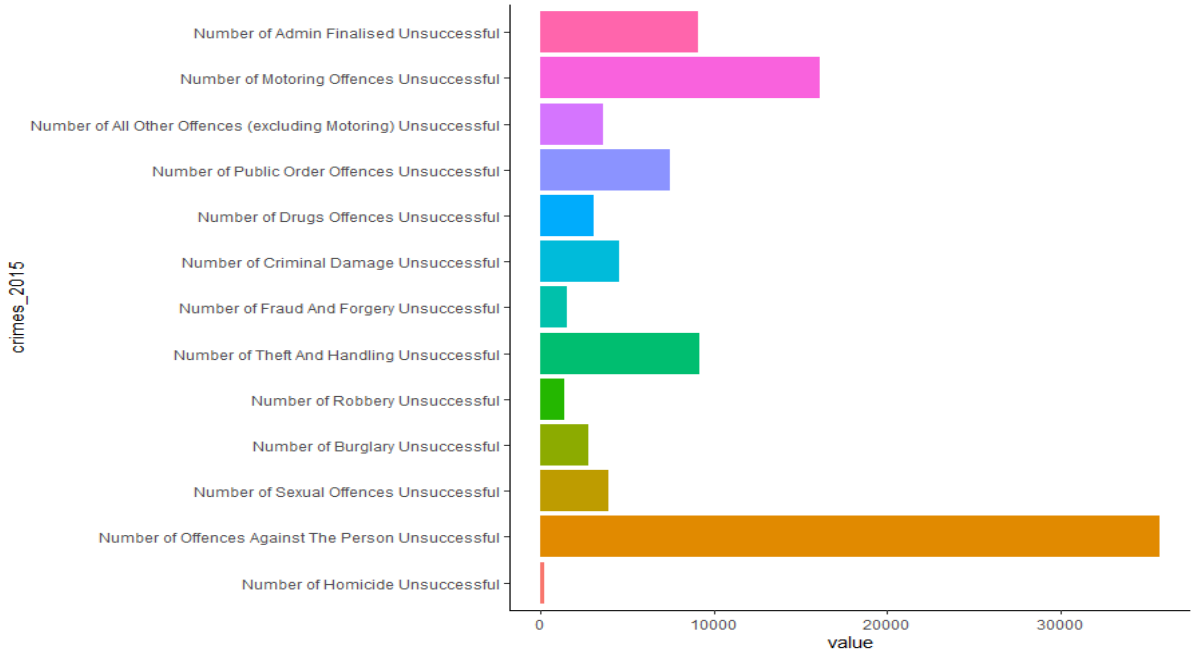

Figure 5: A Bar chart of Unsuccessful Convictions for 2015

From figure 5, we can see that the crime with the highest number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest number of unsuccessful convictions is the Number of homicides.

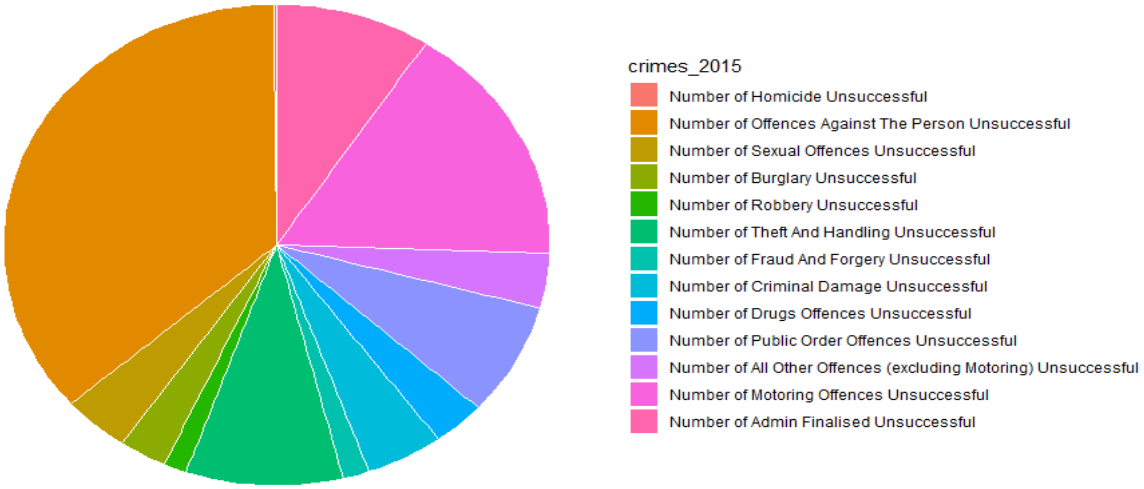

Figure 6: A Pie chart of Unsuccessful Convictions for 2015

From figure 5, we can see that the crime with the highest percentage number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest percentage number of unsuccessful convictions is the Number of homicides.

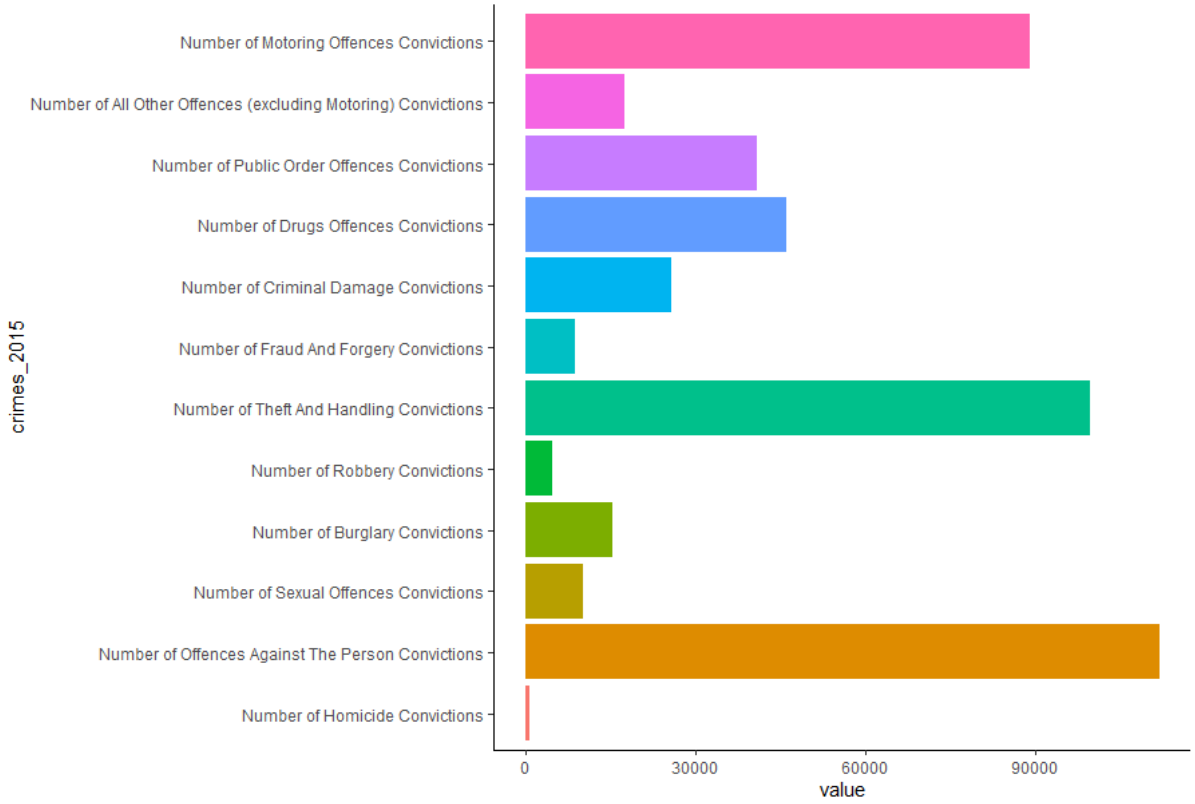

Figure 7: A Bar chart of Convictions for 2015

From figure 7, we can see that the crime with the highest number of convictions is the Number of offences against person followed by the Number of theft and handling, the Number of motoring offences, and the crime with the lowest number of convictions is the Number of homicides.

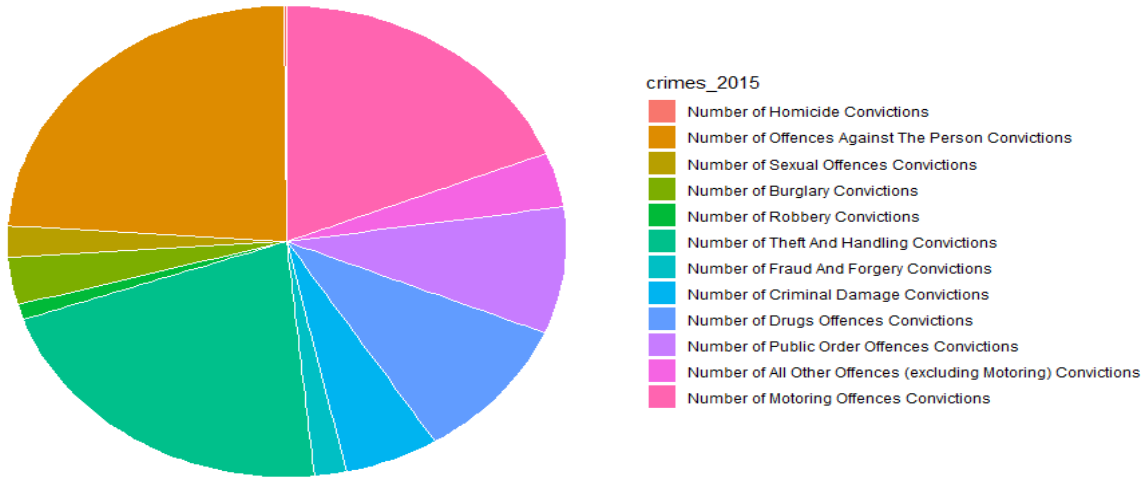

Figure 8: A Pie chart of Convictions for 2015

From figure 8, we can see that the crime with the highest percentage number of convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest percentage number of convictions is the Number of homicides.

For 2016 data set

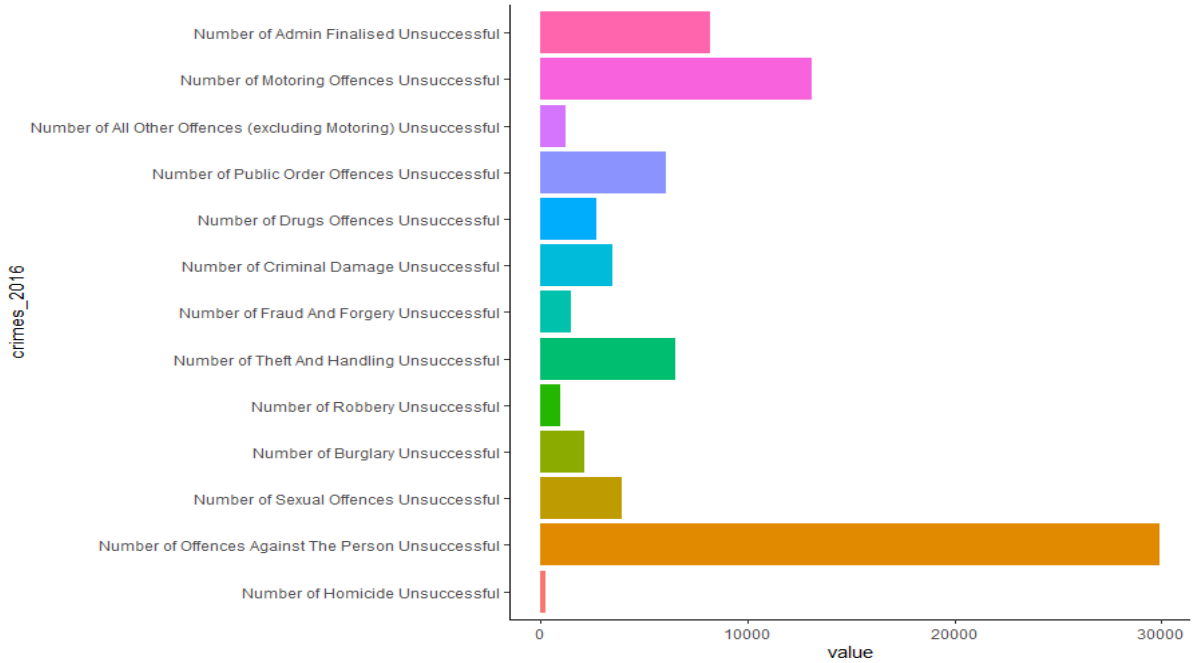

Figure 9: A Bar chart of Unsuccessful Convictions for 2016

From figure 9, we can see that the crime with the highest number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of admin finalised, and the crime with the lowest number of unsuccessful convictions is the Number of homicides.

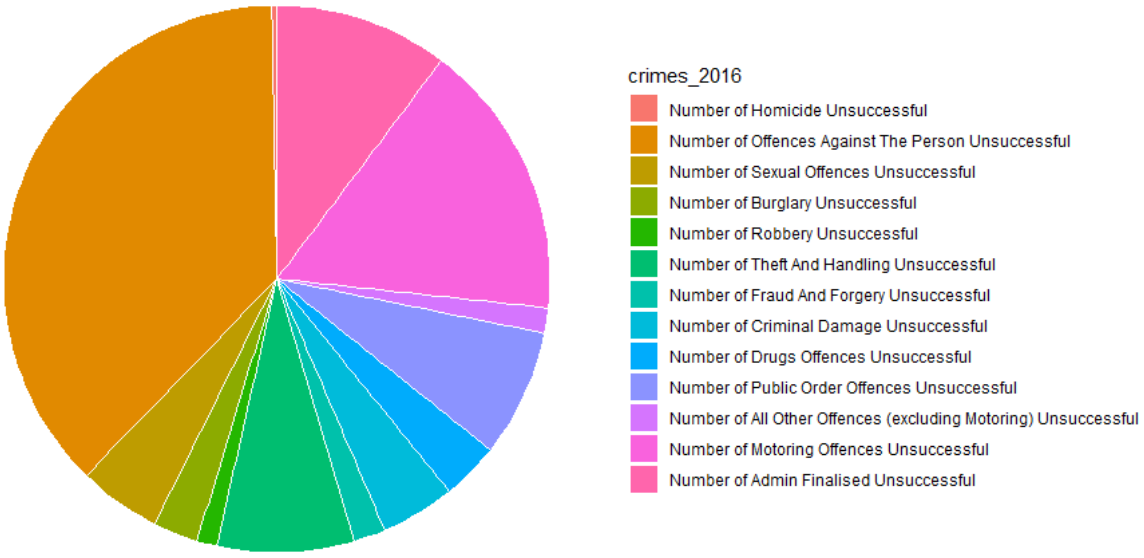

Figure 10: A Pie chart of Unsuccessful Convictions for 2016

From figure 10, we can see that the crime with the highest percentage number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of admin finalised, and the crime with the lowest percentage number of unsuccessful convictions is the Number of homicides.

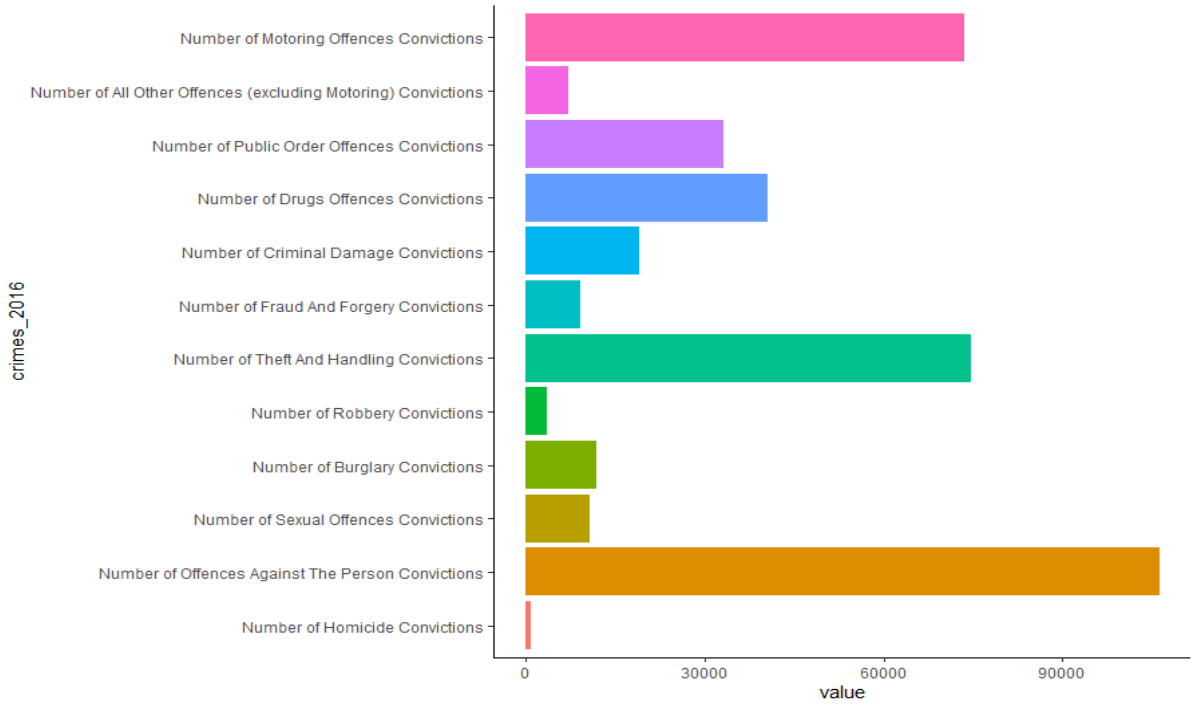

Figure 11: A Bar chart of Convictions for 2016

From figure 11, we can see that the crime with the highest number of convictions is the Number of offences against person followed by the Number of theft and handling, the Number of motoring offences, and the crime with the lowest number of convictions is the Number of homicides.

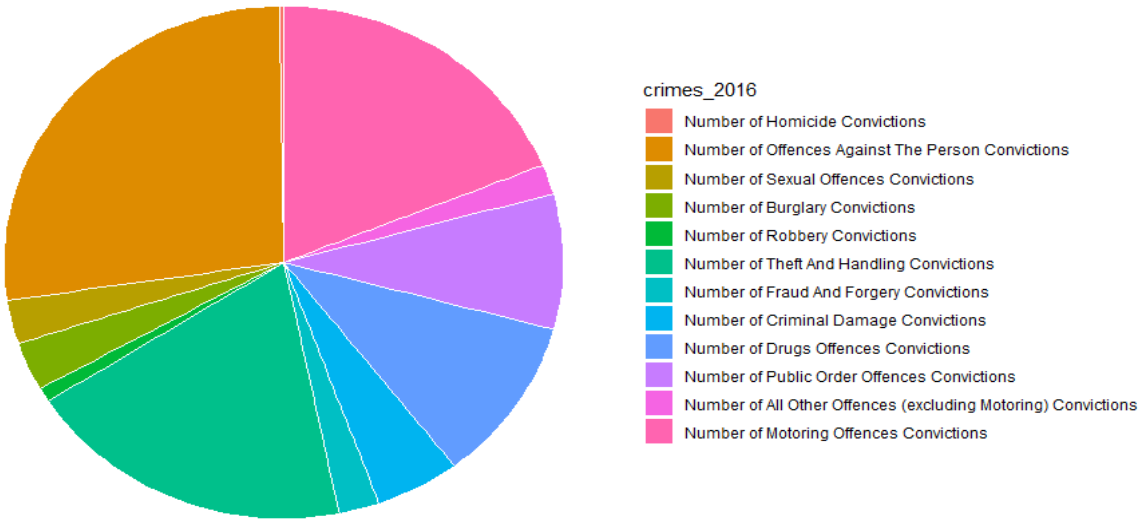

Figure 12: A Pie chart of Convictions for 2016

From figure 12, we can see that the crime with the highest percentage number of convictions is the Number of offences against person followed by the Number of theft and handling, the Number of motoring offences, and the crime with the lowest percentage number of convictions is the Number of homicides.

For 2017 data set.

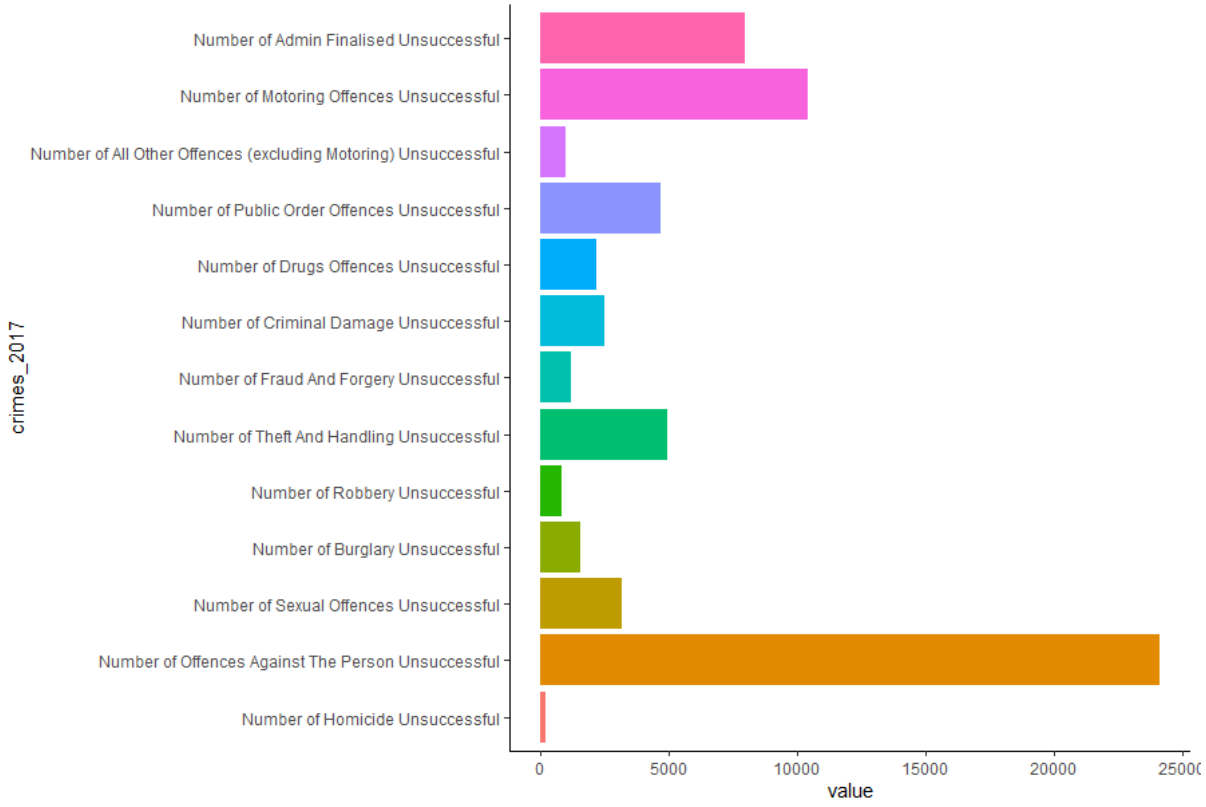

Figure 13: A Bar chart of Unsuccessful Convictions for 2017

From figure 13, we can see that the crime with the highest number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of admin finalised, and the crime with the lowest number of unsuccessful convictions is the Number of homicides.

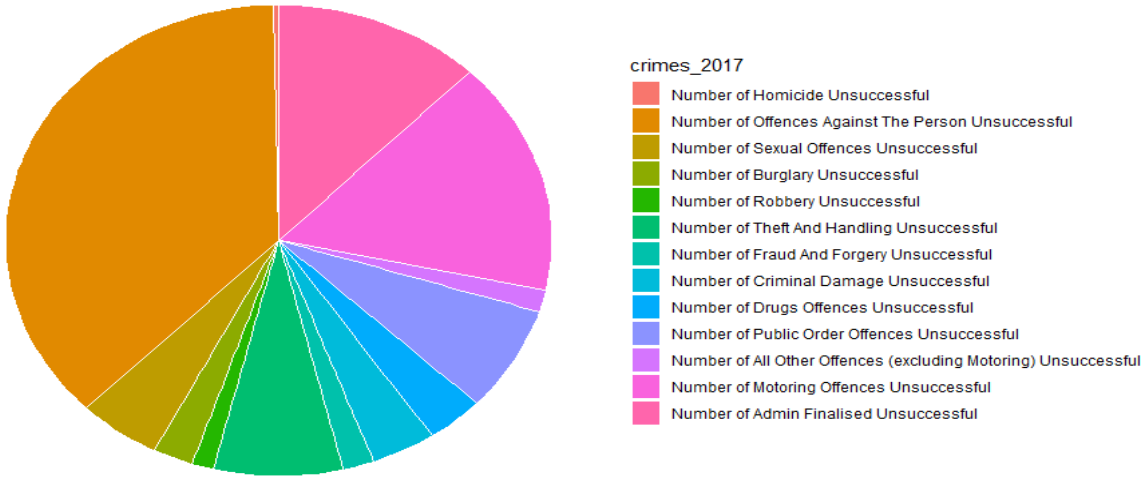

Figure 14: A Pie chart of Unsuccessful Convictions for 2017

From figure 14, we can see that the crime with the highest percentage number of unsuccessful convictions is the Number of offences against person followed by the Number of motoring offences, the Number of admin finalised, and the crime with the lowest percentage number of unsuccessful convictions is the Number of homicides.

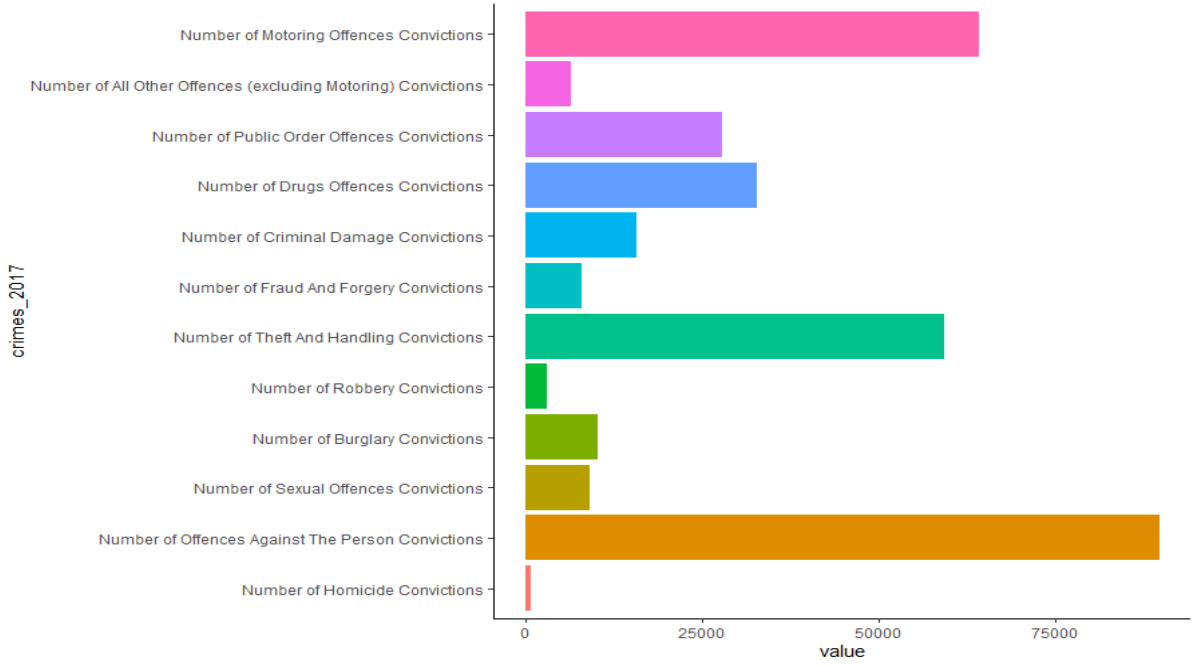

Figure 15: A Bar chart of Convictions for 2017

From figure 15, we can see that the crime with the highest number of convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest number of convictions is the Number of homicides.

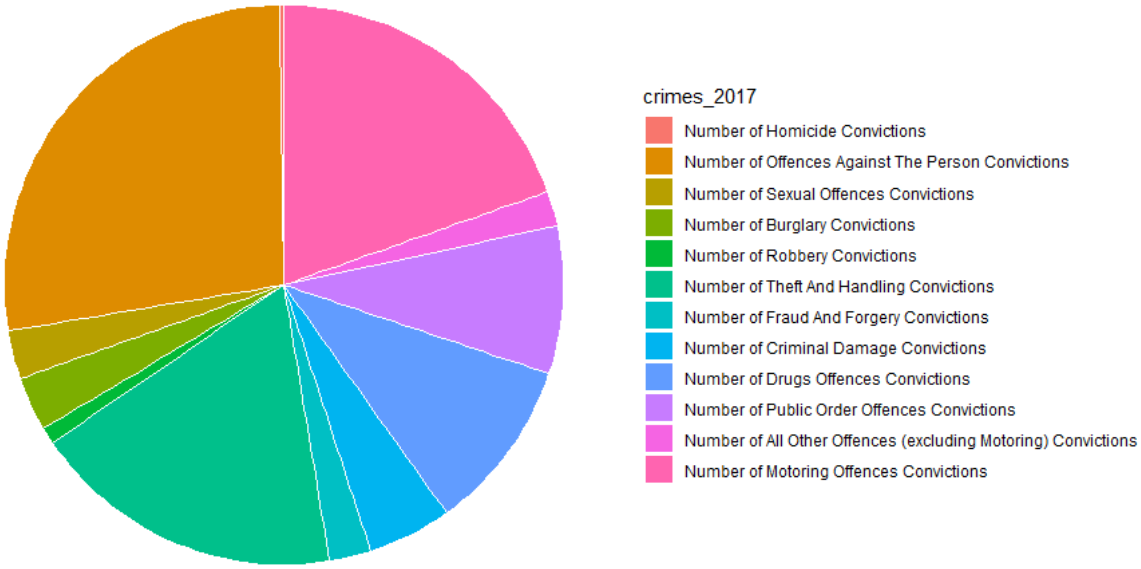

Figure 16: A Pie chart of Convictions for 2017

From figure 16, we can see that the crime with the highest percentage number of convictions is the Number of offences against person followed by the Number of motoring offences, the Number of theft and handling, and the crime with the lowest percentage number of convictions is the Number of homicides.

3.2.1 REMARKS

We observed that the Number of offences against person and the Number of homicides had the highest and lowest number respectively in both Convictions and Unsuccessful Convictions for all the years, except for Convictions in 2014 where both the Number of theft and handling and the Number of motoring offences were bigger than the Number of offences against the person. From our visualisations, we can deduce that the most kinds or highest number of crimes committed in the UK are the crime of offences against the person, crime of motoring offences, and the crime of theft and handling, while the lowest kind of crime in the UK is the crime of Homicides.

We went ahead to combine the four years data sets together to get a general or combine information of all the years and did some visualisations on it.

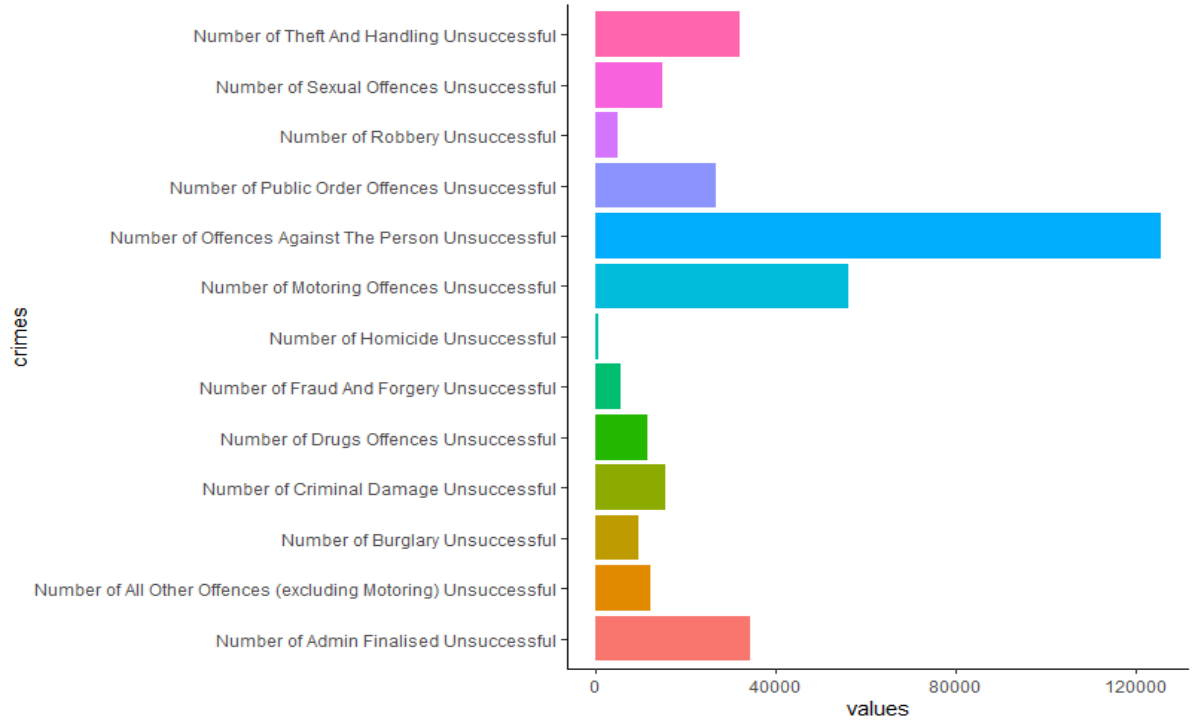

Figure 17: A Bar chart of Unsuccessful Convictions for the four years (2014, 2015, 2016, and 2017).

From figure 17, we observed that the crime with the highest number of unsuccessful convictions for the four years is the number of offences against the Person followed by the number of motoring offences, the number of admin finalised, and the crime with the lowest number of unsuccessful convictions for the four years is the number of homicides.

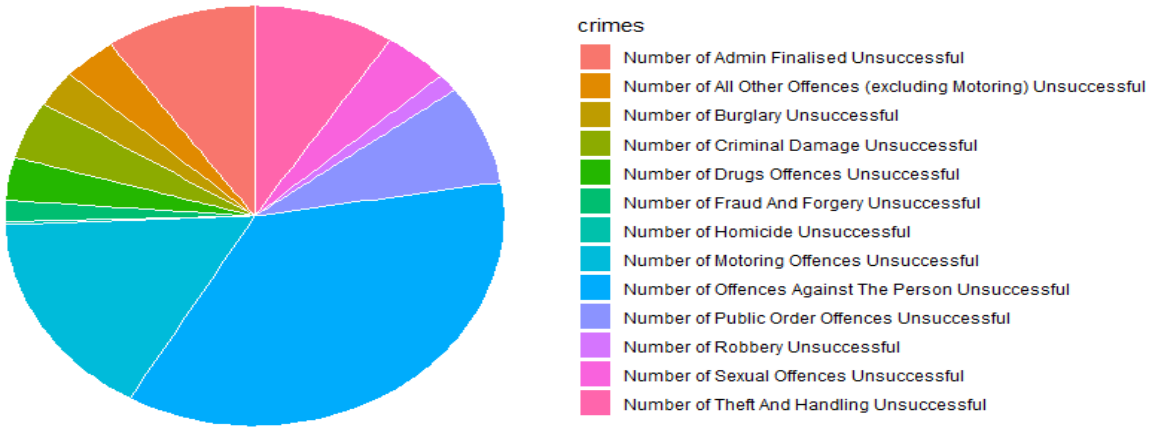

Figure 18: A Pie chart of Unsuccessful Convictions for the four years (2014, 2015, 2016, and 2017).

From figure 18, we observed that the crime with the highest percentage number of unsuccessful convictions for the four years is the number of offences against the Person followed by the number of motoring offences, the number of admin finalised, and the crime with the lowest percentage number of unsuccessful convictions for the four years is the number of homicides.

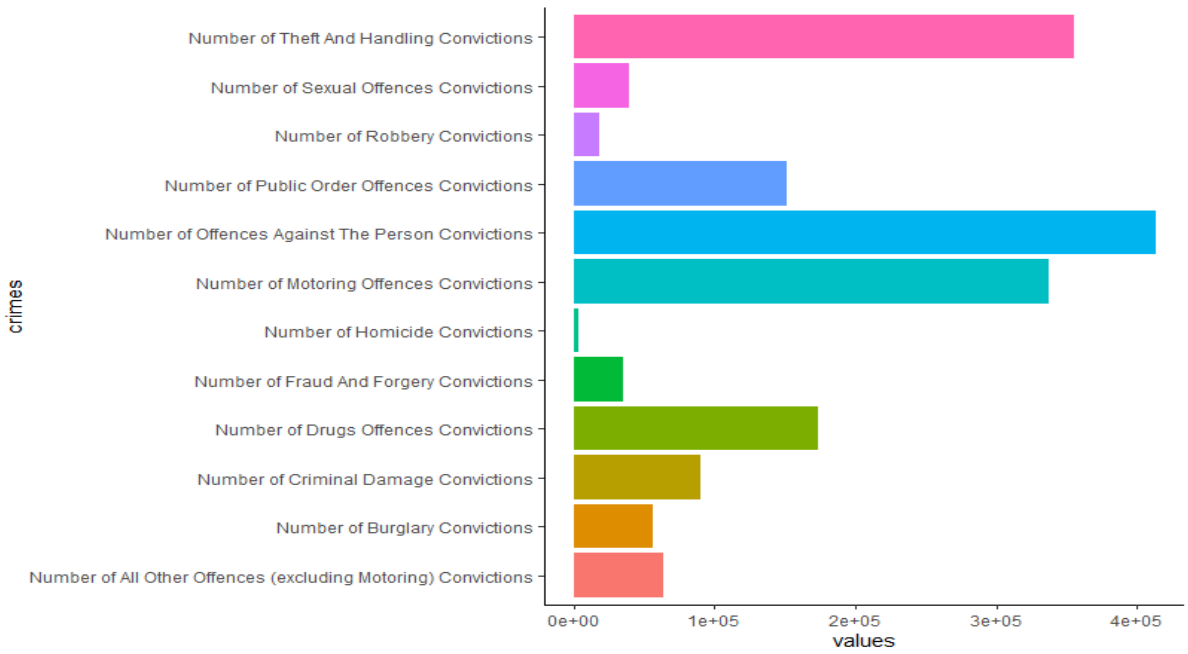

Figure 19: A Bar chart of Convictions for the four years (2014, 2015, 2016, and 2017).

From figure 19, we observed that the crime with the highest number of unsuccessful convictions for the four years is the number of offences against the Person followed by the number of theft and handling, the number of motoring offences, and the crime with the lowest number of unsuccessful convictions for the four years is the number of homicides.

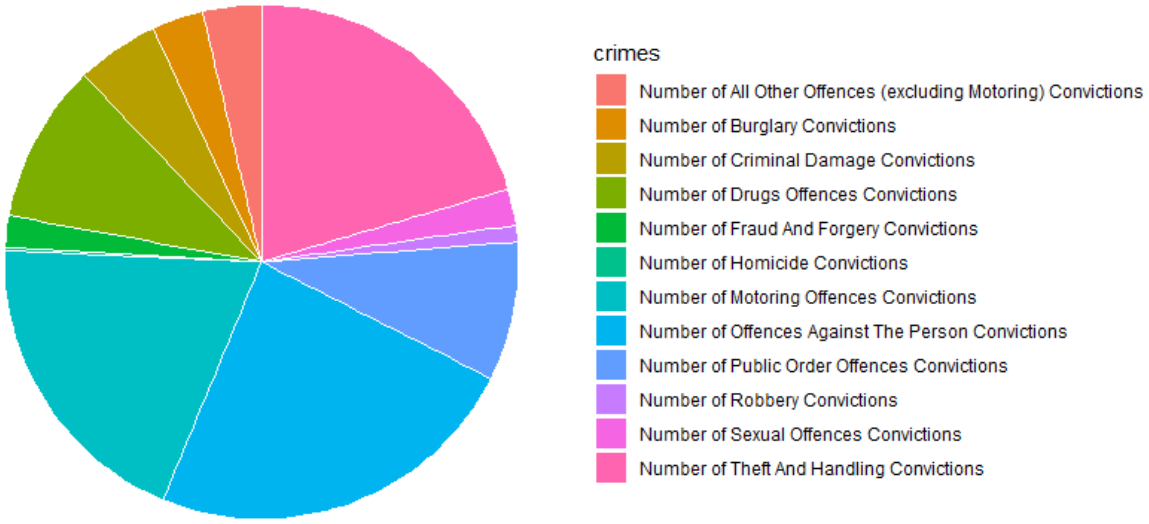

Figure 20: A Pie chart of Convictions for the four years (2014, 2015, 2016, and 2017).

From figure 20, we observed that the crime with the highest percentage number of unsuccessful convictions for the four years is the number of offences against the Person followed by the number of theft and handling, the number of motoring offences, and the crime with the lowest percentage number of unsuccessful convictions for the four years is the number of homicides.

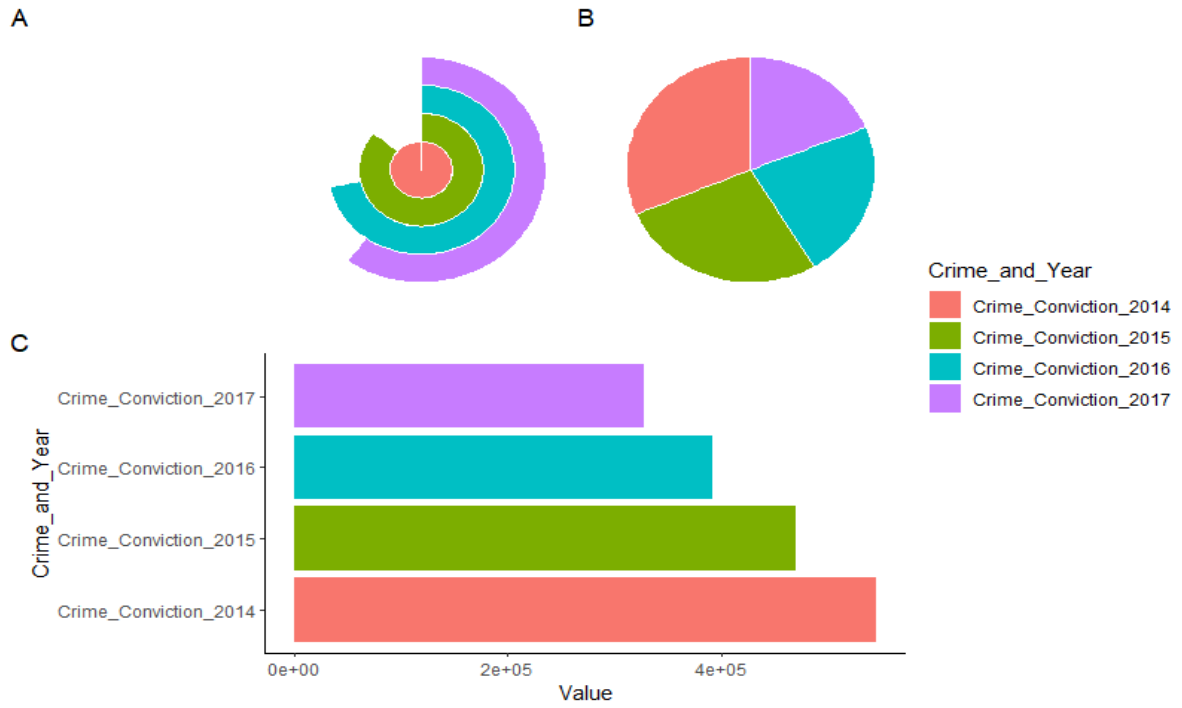

Figure 21: A combine plot containing Pie chart, Bar and Circle Bar chart for Convictions.

Here in figure 21, we observed that 2014 had the highest number of crime convictions and 2017 had the lowest crime conviction.

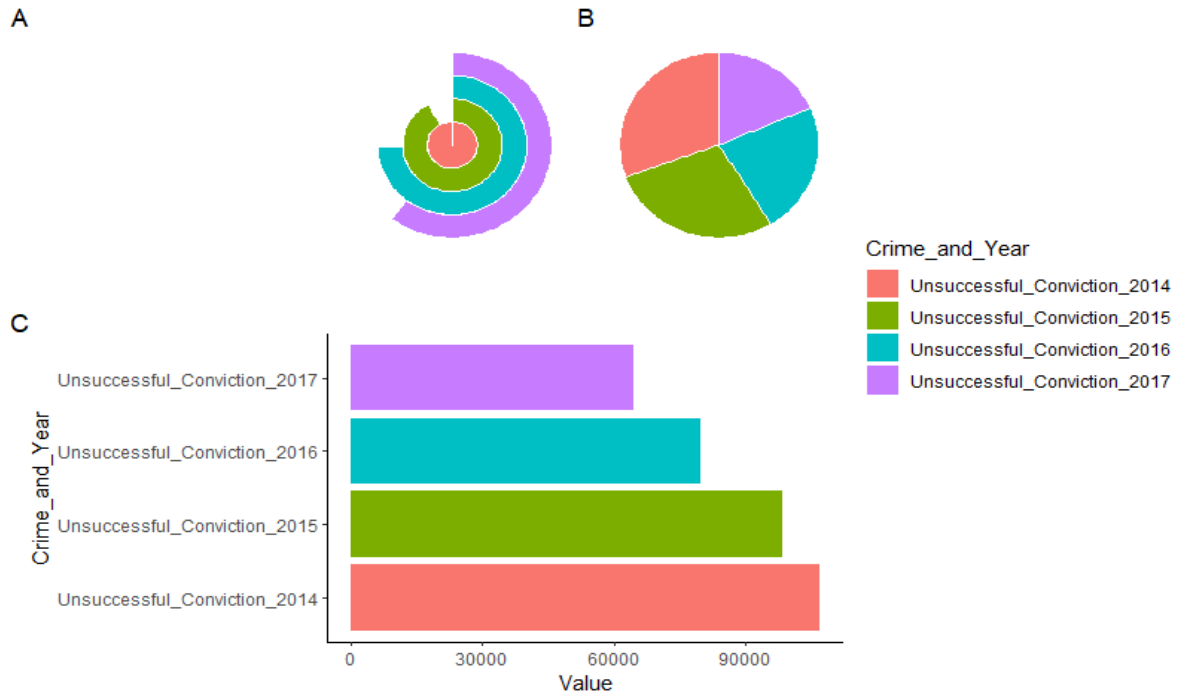

Figure 22: A combine plot containing Pie chart, Bar and Circle Bar chart for Unsuccessful Convictions.

Here in figure 22, we observed that 2014 had the highest number of crime unsuccessful convictions and 2017 had the lowest crime unsuccessful conviction.

Figure 23: A combine plot containing Group Bar Chart and Stacked Bar chart for Convictions and Unsuccessful Convictions.

From figure 22, we observed that 2014 had the highest number of crime convictions and unsuccessful convictions and 2017 had the lowest crime convictions and unsuccessful conviction.

3.2.2 REMARKS

From our visualisations, we observed that 2014 had the highest number of convictions and unsuccessful convictions and 2017 had the lowest number of convictions and unsuccessful convictions. A very important trend we observed from our visualisations is that there is a decrease in both convictions and unsuccessful convictions from 2014 to 2017, which also implies a decrease in crime in the UK from 2014 to 2017. We cannot totally conclude that there is a decrease in crime in the UK from 2014 to 2017 as we know that there were missing months for 2015, 2016, and 2017.

3.3 DESCRIPTIVE ANALYSIS

Here we shall be looking at some descriptive analysis parameters such as minimum, maximum, mean, and median for 2014, 2015, 2016, and 2017 data sets. The mean will tell us about the most common value in the data set, the median will give an idea of the centre value in the data set, the minimum value will give us an idea of the smallest value in the data set, and the maximum value will show us the largest value in the data set. Descriptive analysis is very important because it gives us some very important statistical insight into our data set before going for prescriptive analysis.

For the 2014 data set.

```
Number of Homicide Convictions Number of Homicide Unsuccessful
Min.   :  0.00                 Min.   : 0.000
1st Qu.:  5.25                 1st Qu.: 1.000
Median : 11.00                 Median : 2.000
Mean   : 17.40                 Mean   : 4.476
3rd Qu.: 16.75                 3rd Qu.: 4.750
Max.   :185.00                 Max.   :63.000
Number of Offences Against The Person Convictions
Min.   :  625
1st Qu.: 1189
Median : 1964
Mean   : 2503
3rd Qu.: 2998
Max.   :15048
Number of Offences Against The Person Unsuccessful Number of Sexual Offences Convictions
Min.   : 150.0                                     Min.   :  40.0
```

We can see that the number of motoring offences convictions has both the highest maximum and highest minimum values of 17062 and 810 respectively. We can also see that the number of theft and handling convictions has the highest value for both mean and median at 2883 and 2126 respectively. We observed that the number of homicide unsuccessful convictions has the lowest minimum value, maximum value, mean value, and median value as 0, 63, 4.476, and 2 respectively for the 2014 data set.

For the 2015 data set.

```
Number of Homicide Convictions Number of Homicide Unsuccessful
Min.   :  1.00                 Min.   : 0.000
1st Qu.:  5.25                 1st Qu.: 0.250
Median : 11.00                 Median : 2.000
Mean   : 18.17                 Mean   : 4.333
3rd Qu.: 22.50                 3rd Qu.: 4.000
Max.   :191.00                 Max.   :61.000
Number of Offences Against The Person Convictions
Min.   :  763
1st Qu.: 1302
Median : 2038
Mean   : 2663
3rd Qu.: 3108
Max.   :16247
Number of Offences Against The Person Unsuccessful Number of Sexual Offences Convictions
Min.   : 179.0                                     Min.   :  45.0
```

We can see that the number of offences against the person conviction has the highest value of minimum value, maximum value, mean value, and median value as 763, 16247, 2663, and 2038 respectively. We also observed that the number of homicide unsuccessful convictions has

the lowest value of minimum value, maximum value, mean value, and median value as 0, 61, 4.333, and 2 respectively in the 2015 data set.

For the 2016 data set.

```
Number of Homicide Convictions Number of Homicide Unsuccessful
Min.   :  1.00                 Min.   : 0.000
1st Qu.:  9.00                 1st Qu.: 1.250
Median : 14.50                 Median : 2.500
Mean   : 23.79                 Mean   : 5.571
3rd Qu.: 28.50                 3rd Qu.: 5.000
Max.   :185.00                 Max.   :72.000
Number of Offences Against The Person Convictions
Min.   :  725
1st Qu.: 1246
Median : 1882
Mean   : 2527
3rd Qu.: 2645
Max.   :17332
Number of Offences Against The Person Unsuccessful Number of Sexual Offences Convictions
Min.   : 143.0                                     Min.   :  34.0
```

We observed that the number of offences against the person convictions has the highest value of minimum value, maximum value, mean value, and median value as 725, 17332, 2527, and 1882 respectively. We also observed that the number of homicide unsuccessful conviction has the lowest value of minimum value, maximum value, mean value, and median value as 0, 72, 5.571, and 2.5 respectively in the 2016 data set.

For the 2017 data set.

```
Number of Homicide Convictions Number of Homicide Unsuccessful
Min.   :  2.00                 Min.   : 0.00
1st Qu.:  7.25                 1st Qu.: 1.00
Median : 14.00                 Median : 3.00
Mean   : 19.67                 Mean   : 4.81
3rd Qu.: 19.75                 3rd Qu.: 5.00
Max.   :141.00                 Max.   :69.00
Number of Offences Against The Person Convictions
Min.   :  631
1st Qu.: 1076
Median : 1628
Mean   : 2136
3rd Qu.: 2345
Max.   :13661
Number of Offences Against The Person Unsuccessful Number of Sexual Offences Convictions
Min.   : 130.0                                     Min.   :  34.0
```

We observed that the number of offences against the person convictions has the highest value of minimum value, maximum value, mean value, and median value as 631, 13661, 2136, and 1628 respectively. We also observed that the number of homicides unsuccessful convictions has the lowest value of minimum value, maximum value, mean value, and median value as 0, 69, 4.81, and 3 respectively in the 2017 data set.

CHAPTER FOUR

LINEAR MODEL, CLUSTERING, AND CLASSIFICATION

4.1 LINEAR MODEL

We will be performing regression analysis here using the number of offences against the person convictions as our dependent variable, and the number of sexual offences convictions and the number of motoring offences convictions as our independent variables. The data set we will be using here is a combined data for 2014, 2015, 2016, and 2017 containing just the listed variables. Below is the head view of our data set.

	Number of Homicide Convictions	Number of Homicide Unsuccessful
National	3319	806
Avon and Somerset	92	24
Bedfordshire	44	7
Cambridgeshire	36	5
Cheshire	44	9
Cleveland	39	5
	Number of Offences Against The Person Convictions	
National	412840	
Avon and Somerset	11023	
Bedfordshire	3575	
Cambridgeshire	4802	
Cheshire	9349	
Cleveland	5006	
	Number of Offences Against The Person Unsuccessful	
National	125626	
Avon and Somerset	2945	

We will start by comparing my linear model function (Akash Linear Model) and the inbuilt R linear model function (lm). First, we will state our hypothesis that Akash linear model is better than the R inbuilt linear model based on which functions run faster and their outputs.

Statement of hypothesis:

H_0: Akash Linear Model function is better than the R inbuilt Linear Model function.

Against

H_1: R inbuilt Linear Model function is better than the Akash Linear Model function.

As we have stated our hypothesis, we begin by performing regression analysis using both functions.

4.1.1 AKASH LINEAR MODEL FUNCTION (OUTPUT)

	[,1]
Intercept	67.9030
Number of Sexual Offences Convictions	5.0979
Number of Motoring Offences Convictions	0.6201

Summary of the linear model using Akash function

```
       V1
 Min.   : 0.6201
 1st Qu.: 2.8590
 Median : 5.0979
 Mean   :24.5403
 3rd Qu.:36.5005
 Max.   :67.9030
```

4.1.2 R INBUILT LINEAR MODEL FUNCTION (OUTPUT)

```
Call:
lm(formula = `Number of Offences Against The Person Convictions` ~
    ., data = conv.data1)

Coefficients:
                                (Intercept)    `Number of Sexual Offences Convictions`
                                    67.9030                                     5.0979
`Number of Motoring Offences Convictions`
                                     0.6201
```

Summary of the linear model using R function

```
Call:
lm(formula = `Number of Offences Against The Person Convictions` ~
    ., data = conv.data1)

Residuals:
    Min      1Q  Median      3Q     Max
-3722.2  -835.7   172.3   625.8  4359.3

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             67.90303  344.18564   0.197    0.845
`Number of Sexual Offences Convictions`  5.09794    0.65018   7.841 1.55e-09 ***
`Number of Motoring Offences Convictions`  0.62012    0.08386   7.395 6.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1527 on 39 degrees of freedom
Multiple R-squared:  0.9765,    Adjusted R-squared:  0.9753
F-statistic:   810 on 2 and 39 DF,  p-value: < 2.2e-16
```

4.1.3 REMARKS

We observed that both Akash function and R function gave same output of the estimates for the intercept (67.9030) and coefficients of the number of sexual offences convictions (5.0979) and the number of sexual offences convictions (0.6201), but the R function performed better in the summary aspect as it went further to calculation other estimates such as standard error, residual standard error, R-Square and Adjusted R-Square, F-statistics, etc, which are very important in Statistics in explaining the effectiveness of the model and the explanatory variables on the response variable, on like the Akash function which only gave the mean, median, minimum, maximum, 1st Quarter, and 3rd Quarter values of the model, which has

little or no importance in regression analysis. From our findings, we can conclude and say that the R function is better in the output aspect.

4.1.4 FOR AKASH LINEAR MODEL FUNCTION (TIME IT TAKES TO RUN)

```
Time difference of 0.0009999275 secs
```

4.1.5 FOR R INBUILT LINEAR MODEL FUNCTION (TIME IT TAKES TO RUN)

```
Time difference of 0.003000975 secs
```

4.1.6 REMARKS

We can see that the Akash function runs faster than the R function with 0.0009999275 secs. There is an issue with the whole time checking because each time we run this operation in R, we get a different time, so the stated time above is the result from the last time we performed this operation but nevertheless each time we perform this operation, we will always get a smaller time for the Akash function which implies that the Akash function runs fastest than the R function. But if we are to be realistic, we will see that the time difference between Akash and R functions is not significant owning to fact that the R function does a lot of things than just providing the regression estimates just like the Akash function, so we can conclude that the R function performed better also here.

4.1.7 CONCLUSION

Since the R function performed better based on output and time taken to run, we then reject H_0, and accept H_1, hence, we conclude that the R inbuilt Linear Model function is better than the Akash Linear Model function.

We move further to perform machine learning using our data, but before that, we first need to check the correlation (relationship or linearity) between our variables and then test for the linear regression assumption.

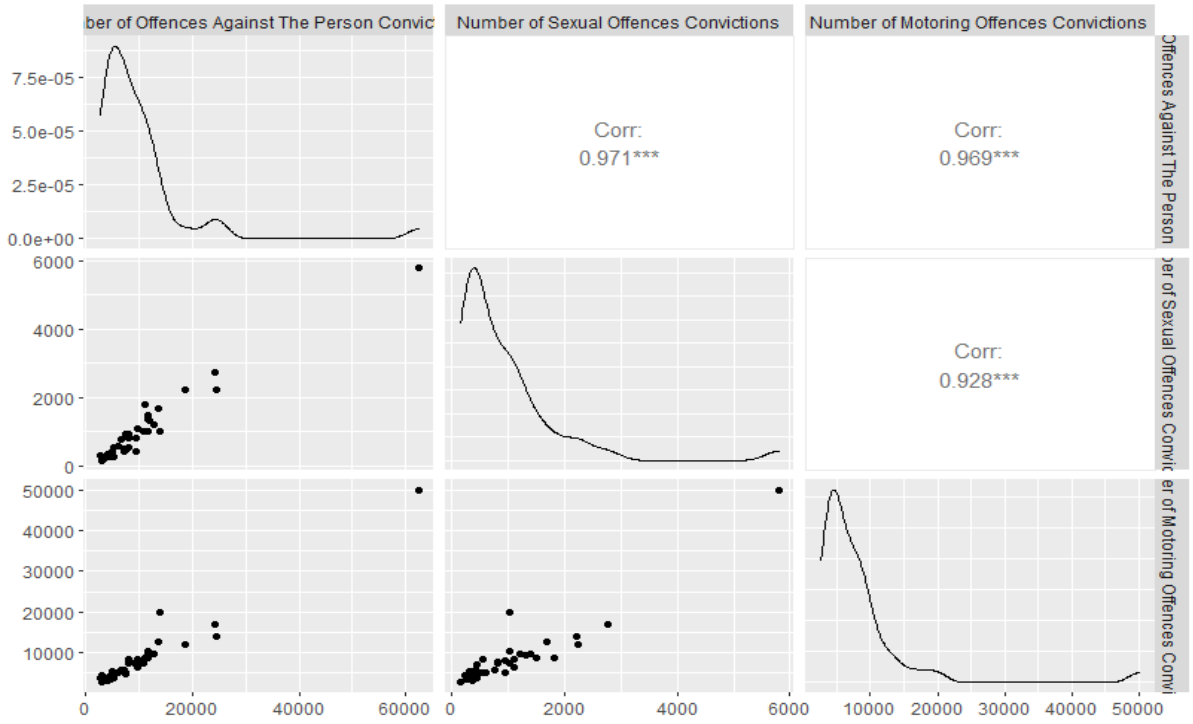

Figure 24: A combine correlation plot showing scatter plots and correlation values.

From figure 24, we can see that there is high correlation between our variables and the scatter plots shows a straight line which is good, so we can go on with the test for assumptions of linear regression.

4.1.8 NORMALITY ASSUMPTION

We are using the Shapiro-Wilk's Test for Normality and the Normal Q-Q plot.

Statement of hypothesis:

H_0: $\mu_1 = 0$ Normally distributed

H_1: $\mu_1 \neq 0$ Not normally distributed

```
        Shapiro-Wilk normality test

data:  lm1$residuals
W = 0.94507, p-value = 0.04311
```

From our result, we conclude that there is no normality since our p-value (0.04311) is less than 0.005, so our data set failed the normality assumption.

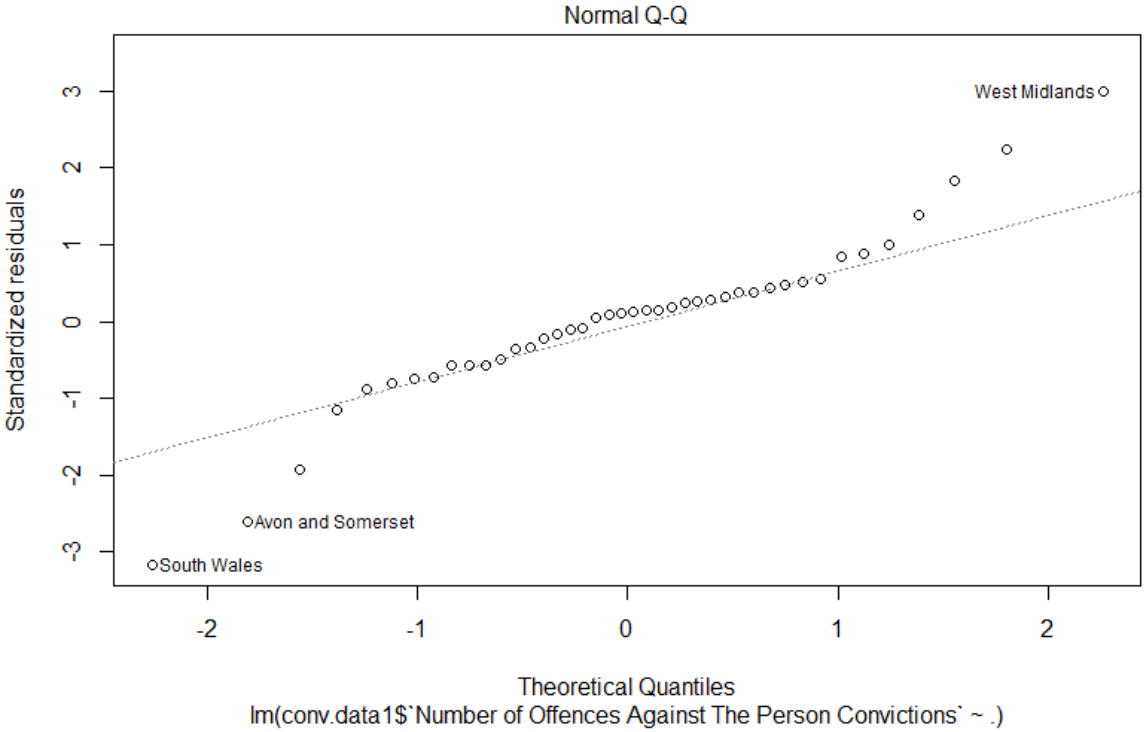

Figure 25: A Normal Q-Q plot used for testing normality assumption.

Figure 25 shows a failed normality assumption.

4.1.9 HOMOSCEDASTICITY ASSUMPTION

We are using the Breusch-Pagan test for homoscedasticity and the Q-Q plot of residual against the fitted.

Statement of hypothesis:

H_0: $\sigma_1{}^2 = \sigma_2{}^2 =, \ldots, \sigma_n{}^2 = 0$ (Homoscedasticity)

H_1: $\sigma_1{}^2 \neq \sigma_2{}^2 \neq, \ldots, \sigma_n{}^2 \neq 0$ (Heteroscedasticity)

```
        studentized Breusch-Pagan test

data:  lm1
BP = 3.6089, df = 2, p-value = 0.1646
```

From our result, we accept H_0, and hence, we conclude that there is equal variance since our p-value (0.1646) is greater than 0.05, so our data set passed the homoscedasticity assumption.

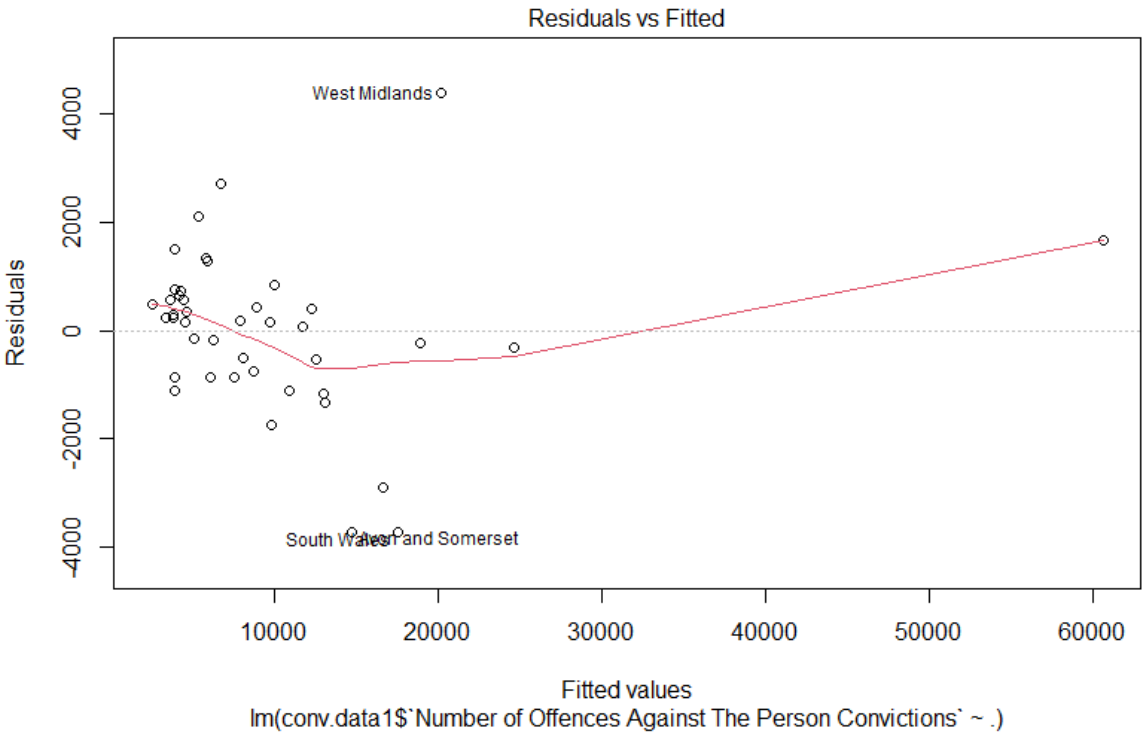

Figure 26: A plot of the residual against the fitted.

Figure 26 shows a passed homoscedasticity assumption.

4.1.10 AUTOCORRELATION ASSUMPTION

We will be using the Durbin-Watson test for autocorrelation

Statement of hypothesis:

H_0: $\rho = 0$

H_1: $\rho > 0$

```
        Durbin-Watson test

data:  lm1
DW = 2.0812, p-value = 0.6043
alternative hypothesis: true autocorrelation is greater than 0
```

From our result, we can conclude that there is no autocorrelation since our p-value (0.6043) is greater than 0.05, so our data set passed the autocorrelation assumption.

4.1.11 MULTICOLLINEARITY ASSUMPTION

We will be using the variance inflation factor test for multicollinearity

```
`Number of Sexual Offences Convictions` `Number of Motoring Offences Convictions`
                              7.238567                                 7.238567
```

From our result, we can conclude that our independent variables are not highly correlated since our values (7.238567 and 7.238567) is less than 10, so our data set passed the multicollinearity assumption.

Most of the assumptions were met but the normality assumption was not met, so we need to transform our data. We will transform the data using Cube Root Transformation. Below is a head view of the transformed data set.

```
                 Number of Offences Against The Person Convictions
Avon and Somerset                                          22.25529
Bedfordshire                                               15.29063
Cambridgeshire                                             16.87100
Cheshire                                                   21.06631
Cleveland                                                  17.10660
Cumbria                                                    16.10350
                 Number of Sexual Offences Convictions
Avon and Somerset                             12.177905
Bedfordshire                                   6.240251
Cambridgeshire                                 7.166096
Cheshire                                       9.302477
Cleveland                                      7.547842
Cumbria                                        6.299605
                 Number of Motoring Offences Convictions
Avon and Somerset                                20.63465
Bedfordshire                                     14.89106
Cambridgeshire                                   15.28064
Cheshire                                         19.72113
Cleveland                                        15.67848
Cumbria                                          15.40665
```

Since we have transformed our data, we again need to test for linear regression assumptions

4.1.12 NORMALITY ASSUMPTION

```
        Shapiro-Wilk normality test

data:  lm2$residuals
W = 0.9774, p-value = 0.563
```

From our result, we can conclude that the normality assumption is now met since our p-value (0.563) is greater than 0.05.

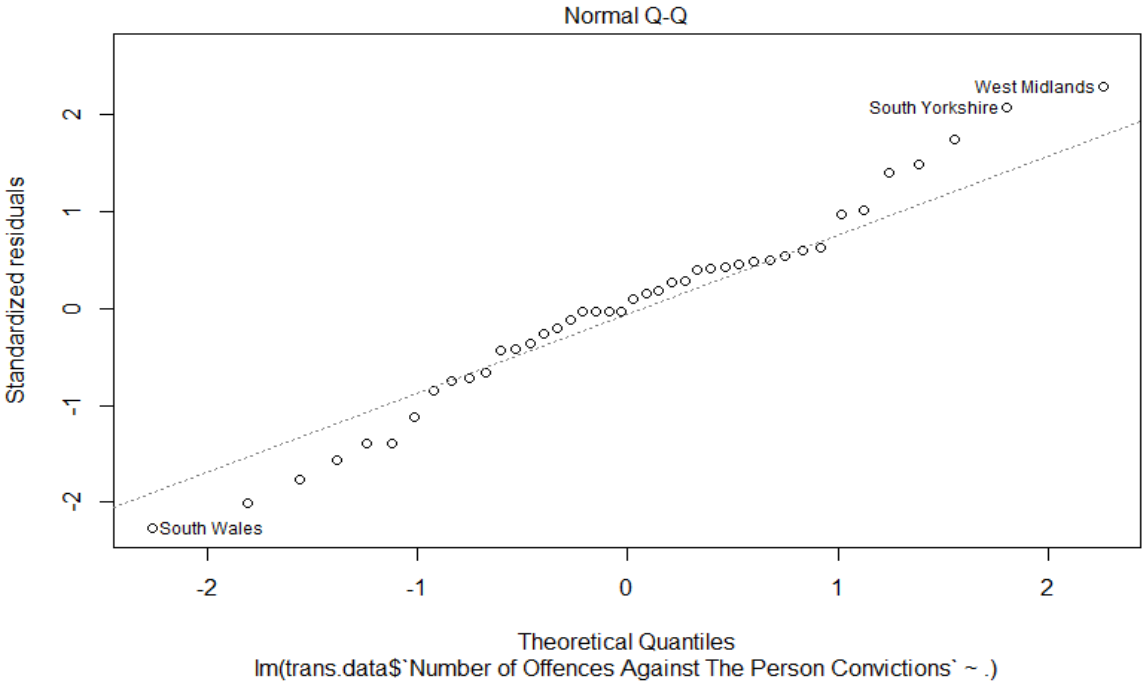

Figure 27: A Q-Q Normal plot used for testing the normality assumption for the transformed data.

Figure 27 shows a passed normality assumption.

4.1.13 HOMOSCEDASTICITY ASSUMPTION

```
        studentized Breusch-Pagan test

data:  lm2
BP = 3.2105, df = 2, p-value = 0.2008
```

From our result, we can conclude that the homoscedasticity assumption was met since our p-value (0.2008) is greater than 0.05.

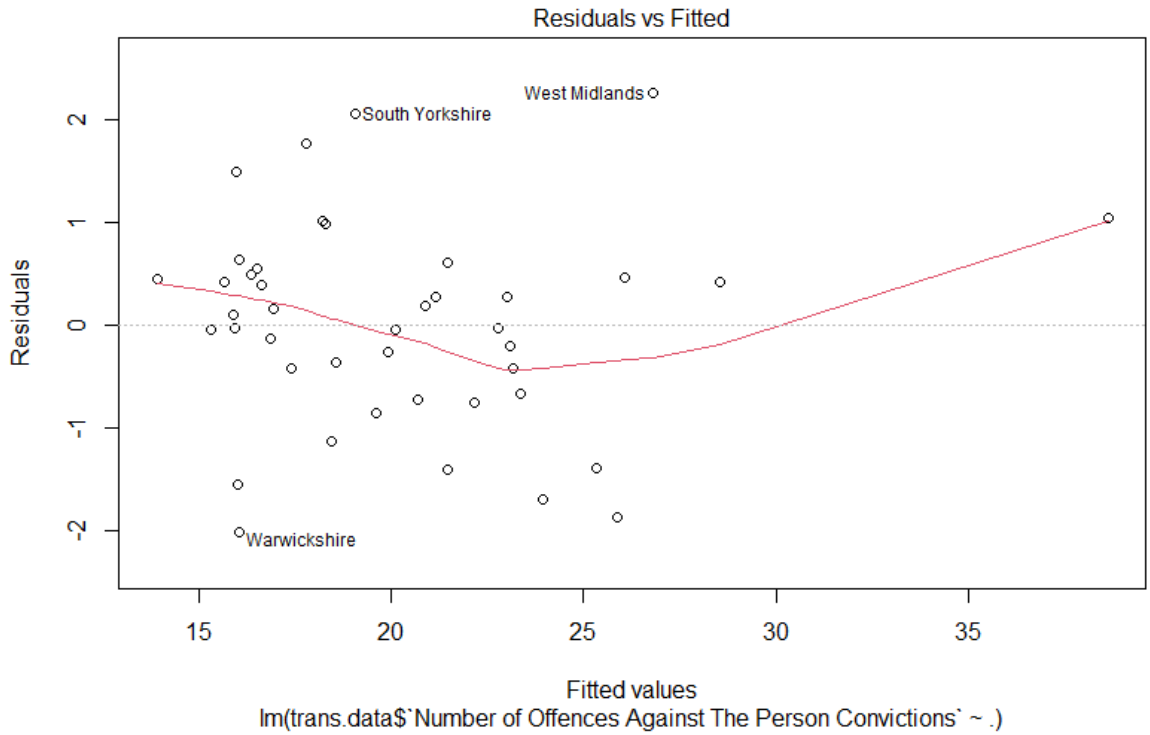

Figure 28: A plot of the residual against the fitted.

Figure 28 shows a passed homoscedasticity assumption.

4.1.14 AUTOCORRELATION ASSUMPTION

```
        Durbin-Watson test

data:  lm2
DW = 1.9631, p-value = 0.462
alternative hypothesis: true autocorrelation is greater than 0
```

From our result, we can conclude that the autocorrelation assumption was met since our p-value (0.462) is greater than 0.05.

4.1.15 MULTICOLLINEARITY ASSUMPTION

```
`Number of Sexual Offences Convictions` `Number of Motoring Offences Convictions`
                              5.491488                                 5.491488
```

From our result, we can conclude that the multicollinearity assumption was met as our values (5.491488 and 5.491488) are lesser than 10.

Since all assumptions have been met, we move to split the data into train and test for machine learning and prediction purpose. We will split 70% of the data set for train and 30% for test. Below is a head view of the data sets for train and testing.

Train data set

	Number of Offences Against The Person Convictions		
Gloucestershire	14.37587		
Staffordshire	19.97831		
Hertfordshire	19.24251		
Surrey	17.33033		
Sussex	21.38883		
Bedfordshire	15.29063		
	Number of Sexual Offences Convictions	Number of Motoring Offences Convictions	
Gloucestershire	5.348481	13.84871	
Staffordshire	9.325532	19.39169	
Hertfordshire	7.500741	17.92043	
Surrey	8.183269	17.30143	
Sussex	10.307137	20.37220	
Bedfordshire	6.240251	14.89106	

Test data set.

	Number of Offences Against The Person Convictions
Avon and Somerset	22.25529
Cheshire	21.06631
Cumbria	16.10350
Durham	16.68988
Dyfed Powys	14.45926
GreaterManchester	28.96707
	Number of Sexual Offences Convictions
Avon and Somerset	12.177905
Cheshire	9.302477
Cumbria	6.299605
Durham	6.753313
Dyfed Powys	6.091199
GreaterManchester	14.013592
	Number of Motoring Offences Convictions
Avon and Somerset	20.63465
Cheshire	19.72113
Cumbria	15.40665
Durham	15.35593
Dyfed Powys	16.26259
GreaterManchester	25.69768

We use the train data set to build our regression model and test the hypothesis that an increase in the number of sexual offences convictions and the number of motoring offences convictions will lead to an increase in the number of offences against the person convictions.

Statement of hypothesis:

H_0: An increase in the number of sexual offences convictions and the number of motoring offences convictions will lead to an increase in the number of offences against the person convictions.

Against

H_1: An increase in the number of sexual offences convictions and the number of motoring offences convictions will not lead to an increase in the number of offences against the person convictions.

```
Call:
lm(formula = train.data$`Number of Offences Against The Person Convictions` ~
    ., data = train.data)

Coefficients:
                                (Intercept)    `Number of Sexual Offences Convictions`
                                     2.3416                                     0.9966
`Number of Motoring Offences Convictions`
                                     0.4688
```

We will use the summary function to see more information in our model before we conclude our hypothesis.

```
Call:
lm(formula = train.data$`Number of Offences Against The Person Convictions` ~
    ., data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1945 -0.7482 -0.1075  0.4304  2.3706

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2.3416     1.2282   1.907 0.067689 .
`Number of Sexual Offences Convictions`  0.9966     0.1946   5.120 2.45e-05 ***
`Number of Motoring Offences Convictions`  0.4688     0.1231   3.809 0.000768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.099 on 26 degrees of freedom
Multiple R-squared:  0.9154,    Adjusted R-squared:  0.9089
F-statistic: 140.6 on 2 and 26 DF,  p-value: 1.142e-14
```

We get a regression straight line model of $y = 2.3416 + 0.9966x_1 + 0.4688x_2$

Where x_1 is number of sexual offences convictions and x_2 is number of motoring offences convictions.

4.1.16 CONCLUSION.

Our model implies that both number of sexual offences convictions and number of motoring offences convictions are significant in explaining the response or dependent variable and our model implies that an increase in both number of sexual offences convictions and number of motoring offences convictions will lead to an increase in the number of offences against the person convictions, so we accept H_0 and hence we conclude that an increase in both the number of sexual offences convictions and number of motoring offences convictions will lead to an increase in the number of offences against the person convictions

Our model seems to be particularly good because the independent variables were significant in explaining the dependent variable, so we will perform an analysis of variance test using an ANOVA table to see if our independent variables are the same or if there is a significant difference between them.

Statement of hypothesis:

H_0: The explanatory or independent variables are the same.

Against

H_1: There is a significant difference in the explanatory or independent variables.

4.1.17 ANOVA TABLE

```
Analysis of Variance Table

Response: train.data$`Number of Offences Against The Person Convictions`
                                     Df Sum Sq Mean Sq F value    Pr(>F)
`Number of Sexual Offences Convictions`    1 322.21  322.21 266.719 3.465e-15 ***
`Number of Motoring Offences Convictions`  1  17.53   17.53  14.508 0.0007681 ***
Residuals                                 26  31.41    1.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4.1.18 CONCLUSION.

From our result, we reject H_0, and accept H_1, hence, since our p-values (3.465e-15 and 0.0007681) are smaller than 0.05, so we conclude that there is a significant difference in the explanatory or independent variables.

We then forge ahead to use the test data set to test our model and make some predictions. Below is a table of the predicted values and the actual values.

	Predicted	Actual
Avon and Somerset	14088.273	11023
Cheshire	9074.416	9349
Cumbria	3976.328	4176
Durham	4307.695	4649
Dyfed Powys	4123.870	3023
GreaterManchester	22797.699	24306
Humberside	8300.589	7611
Metropolitan and City	52798.073	62288
Northamptonshire	4872.773	4686
Northumbria	12793.807	11712
North Wales	6605.157	6058
North Yorkshire	4617.703	4948
Nottinghamshire	9779.025	9852

We can see that there is not much difference between the predicted values and the actual or observed values, which implies that our model fits well, but this can be determined by calculating error. We shall be using the Root mean square error (RMSE) Test to show this.

Statement of hypothesis:

H_0: The model fit well.

Against

H_1: The model does not fit well.

4.1.19 RMSE

```
[1] 0.9821679
```

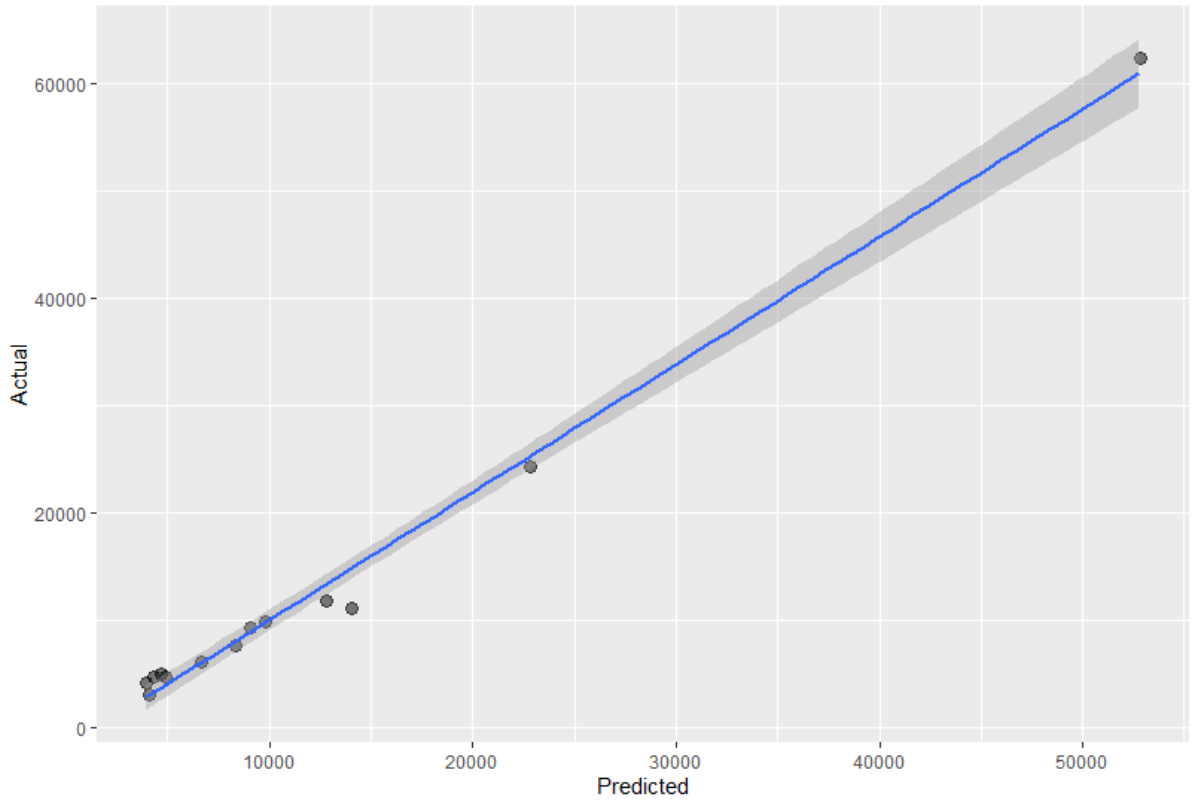

Figure 29: A scatter plot of the actual values against the predicted values.

From figure 29, we can see our model fit well by showing a straight line.

4.1.20 CONCLUSION.

Since our RMSE value (0.9821679) is less than 1 and our plot in figure 29, we accept H_0 and conclude that our model fits the data set well and it shows that our model predicted values are close to the actual or observed data.

4.2 CLUSTERING

Here we shall look at the K-Means Clustering Algorithm and we will test the hypothesis that the Metropolitan and City have the highest crime rate (convictions).

Statement of hypothesis:

H_0: Metropolitan and City have the highest average crime rate (convictions) compared to other counties in the UK.

Against

H_1: Metropolitan and City do not have the highest average crime rate (convictions) compared to other counties in the UK.

4.2.1 K-MEANS ALGORITHM

Here we row bind 2014, 2015, 2016, and 2017 data sets picking only the columns with convictions together to get a combined data set for our K-means clustering. Below is a head view of the data set.

```
             County Number of Homicide Convictions
1 Avon and Somerset                             17
2      Bedfordshire                             17
3    Cambridgeshire                              3
4          Cheshire                             11
5         Cleveland                             11
6           Cumbria                              1
  Number of Offences Against The Person Convictions Number of Sexual Offences Convictions
1                                              2706                                   429
2                                               906                                    63
3                                              1188                                    84
4                                              2051                                   185
5                                              1272                                   144
6                                              1126                                    86
  Number of Burglary Convictions Number of Robbery Convictions
1                            520                            98
2                            156                           109
3                            202                            56
```

To perform K-means clustering, we first standardise our data set because some of the columns contain exceptionally large values (outliers). Below is a head view of the standardised data set.

```
     Number of Homicide Convictions Number of Offences Against The Person Convictions
[1,]                     -0.09784606                                        0.1022027
[2,]                     -0.09784606                                       -0.6377442
[3,]                     -0.59489556                                       -0.5218192
[4,]                     -0.31086727                                       -0.1670557
[5,]                     -0.31086727                                       -0.4872883
[6,]                     -0.66590264                                       -0.5473062
     Number of Sexual Offences Convictions Number of Burglary Convictions
[1,]                             0.7828069                      0.5145287
[2,]                            -0.6894725                     -0.4923340
[3,]                            -0.6049975                     -0.3650931
[4,]                            -0.1987127                     -0.0110315
[5,]                            -0.3636402                      0.1908943
[6,]                            -0.5969522                     -0.5421239
     Number of Robbery Convictions Number of Theft And Handling Convictions
[1,]                  -0.046388001                                 0.7312466
```

Since we have standardised our data set, we can now perform the k-means clustering, but before that, we need to choose a value for K. We will use the Elbow test to choose our K value. Below is an Elbow test plot for selecting our K value.

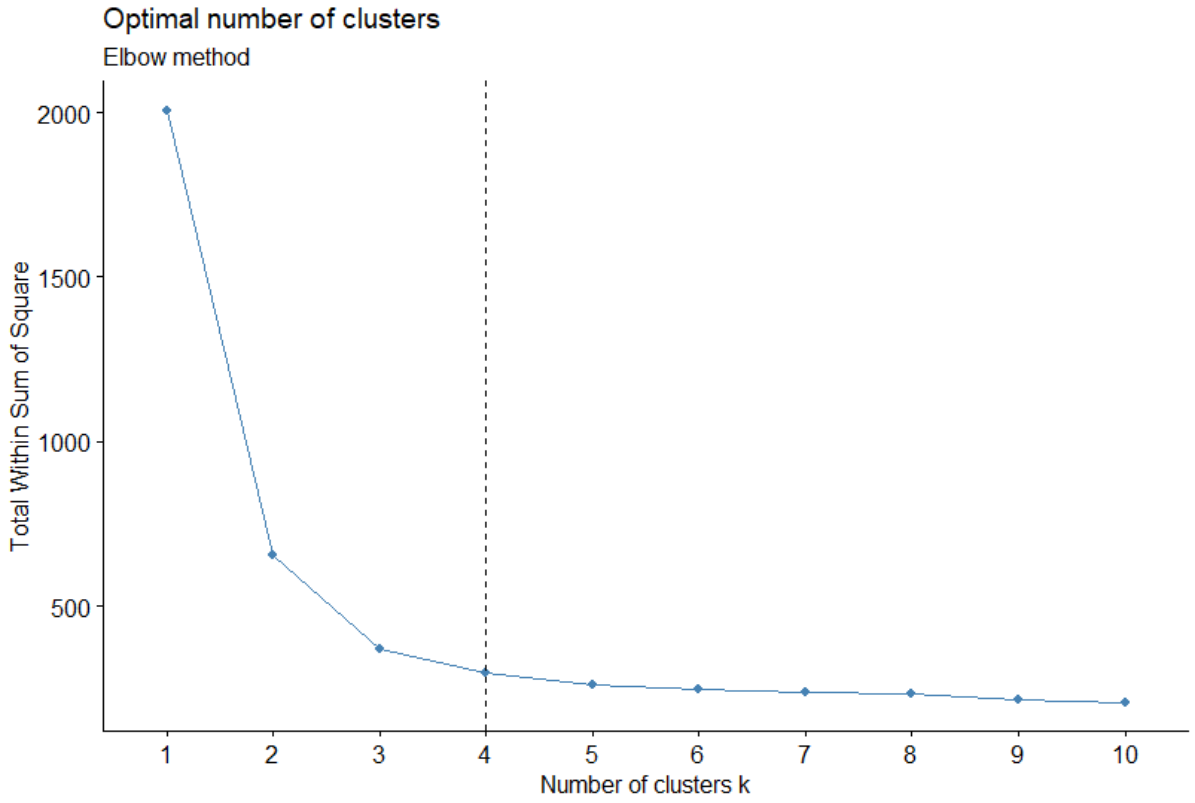

Figure 30: An Elbow method plot for selecting K value.

4 is the best k value for our data set just as seen in figure 30. We can now go ahead to perform the K-means clustering. Below is our result from the k-means clustering.

```
K-means clustering with 4 clusters of sizes 4, 12, 104, 48

Cluster means:
  Number of Homicide Convictions Number of Offences Against The Person Convictions
1                      5.5294644                                          5.3911787
2                      0.5737625                                          1.1941382
3                     -0.3692433                                         -0.4309938
4                      0.1957978                                          0.1860205
  Number of Sexual Offences Convictions Number of Burglary Convictions
1                             4.8858808                      5.1927061
2                             1.1367976                      1.3160080
3                            -0.4771245                     -0.4340596
4                             0.3424137                      0.1787349
  Number of Robbery Convictions Number of Theft And Handling Convictions
1                   5.313250386                                 4.6906199
2                   1.062348932                                 1.4450172
3                  -0.323486479                                -0.4648907
4                  -0.007470727                                 0.2551240
```

```
  Number of Fraud And Forgery Convictions Number of Criminal Damage Convictions
1                              6.07942646                              4.8801772
2                              0.50144558                              1.3903433
3                             -0.30265225                             -0.4537559
4                              0.02376628                              0.2288706
  Number of Drugs Offences Convictions Number of Public Order Offences Convictions
1                           5.96153667                                   5.1172230
2                           0.51755687                                   1.2270978
3                          -0.33030338                                  -0.4160983
4                           0.08947339                                   0.1683367
  Number of All Other Offences (excluding Motoring) Convictions
1                                                    4.45275084
2                                                    1.00263301
3                                                   -0.33193877
4                                                    0.09747984
  Number of Motoring Offences Convictions
1                               5.1277506
2                               1.0423456
3                              -0.4049012
4                               0.1893869

Clustering vector:
  [1] 4 3 3 4 4 3 3 4 3 3 3 4 3 2 3 4 3 4 4 2 3 3 4 1 3 3 2 3 3 4 2 4 4 3 3 4 4 3 4 2 2 3 4 3 3
 [46] 3 3 3 3 3 3 3 3 4 3 2 3 4 3 3 4 4 3 3 4 1 3 3 4 3 3 4 4 4 3 3 3 4 4 3 3 2 2 3 4 3 3 3 3 3
 [91] 3 3 3 3 3 4 3 2 3 4 3 3 4 4 3 3 4 1 3 3 4 3 3 4 4 3 3 3 3 3 4 3 3 2 4 3 4 3 3 3 3 3 3 3 3
[136] 3 3 3 3 4 3 3 3 3 4 4 3 3 4 1 3 3 4 3 3 4 4 3 3 3 3 3 4 3 3 2 4 3

Within cluster sum of squares by cluster:
[1] 111.00741  55.25415  53.20446  79.47871
 (between_SS / total_SS =  85.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Our K-means cluster have 4 clusters with sizes 4, 12, 104, 48 in each cluster. We could see that the number of fraud and forgery convictions had the highest mean for the first cluster (6.07942646), the number of theft and handling convictions had the highest mean in the second cluster (1.4450172), the number of fraud and forgery convictions had the highest mean in the third cluster (-0.30265225) and the number of sexual offences convictions has the highest mean in the fourth cluster (0.3424137). We shall look at the plot of the clusters to have a clearer picture of the clusters.

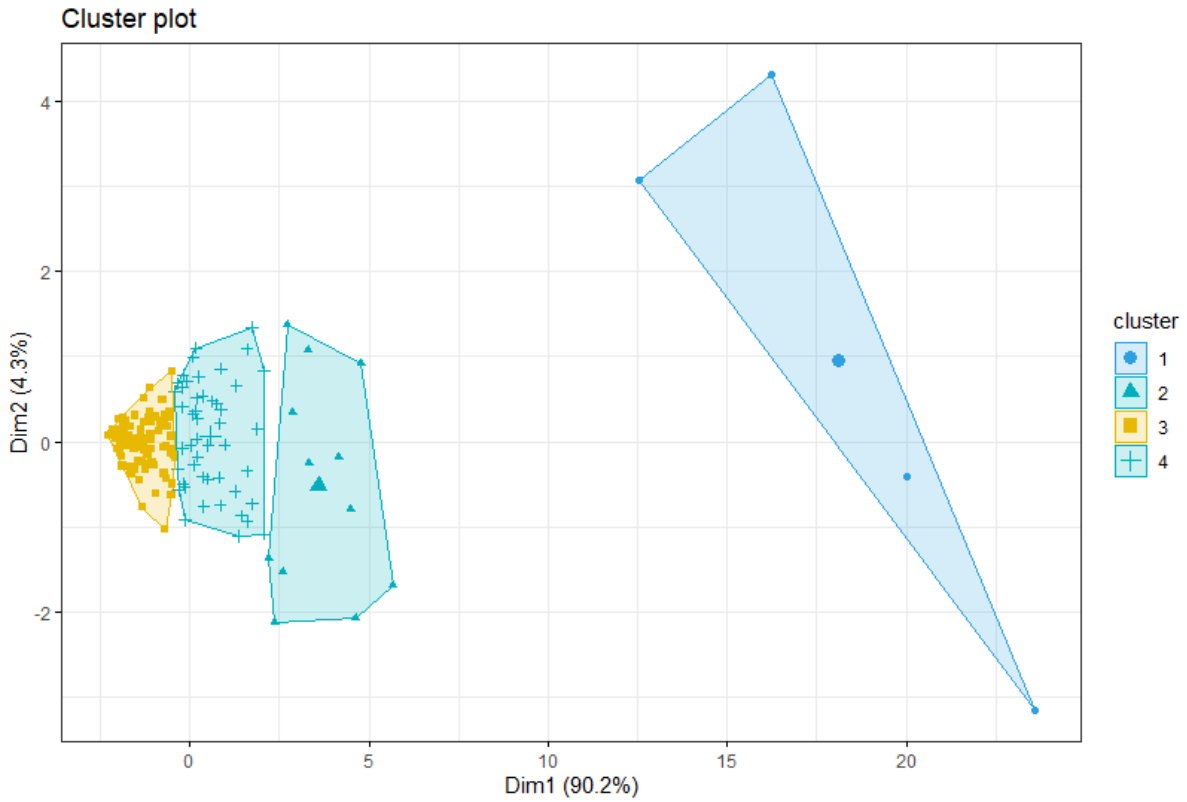

Figure 31: A cluster plot

From figure 31, we can see that we have four clusters and we observed that cluster three has more numbers.

We went further to add the total convictions column and cluster (class) column to our data so that we can check the counties with the highest convictions (crimes). Below is a head view of the data.

```
  Class Total.Conviction            County Number of Homicide Convictions
1     4            14764 Avon and Somerset                             17
2     3             4408      Bedfordshire                             17
3     3             5480    Cambridgeshire                              3
4     4            10327          Cheshire                             11
5     4             9581         Cleveland                             11
6     3             5897           Cumbria                              1
  Number of Offences Against The Person Convictions Number of Sexual Offences Convictions
1                                              2706                                   429
2                                               906                                    63
3                                              1188                                    84
4                                              2051                                   185
5                                              1272                                   144
6                                              1126                                    86
  Number of Burglary Convictions Number of Robbery Convictions
1                            520                            98
2                            156                           109
3                            202                            56
4                            330                            54
5                            403                            73
6                            138                            20
  Number of Theft And Handling Convictions Number of Fraud And Forgery Convictions
```

We can now check for the cluster with the highest average number of total convictions.

	Class	Average
	<fct>	<dbl>
1	1	69572.
2	2	22725.
3	3	5704.
4	4	12315.

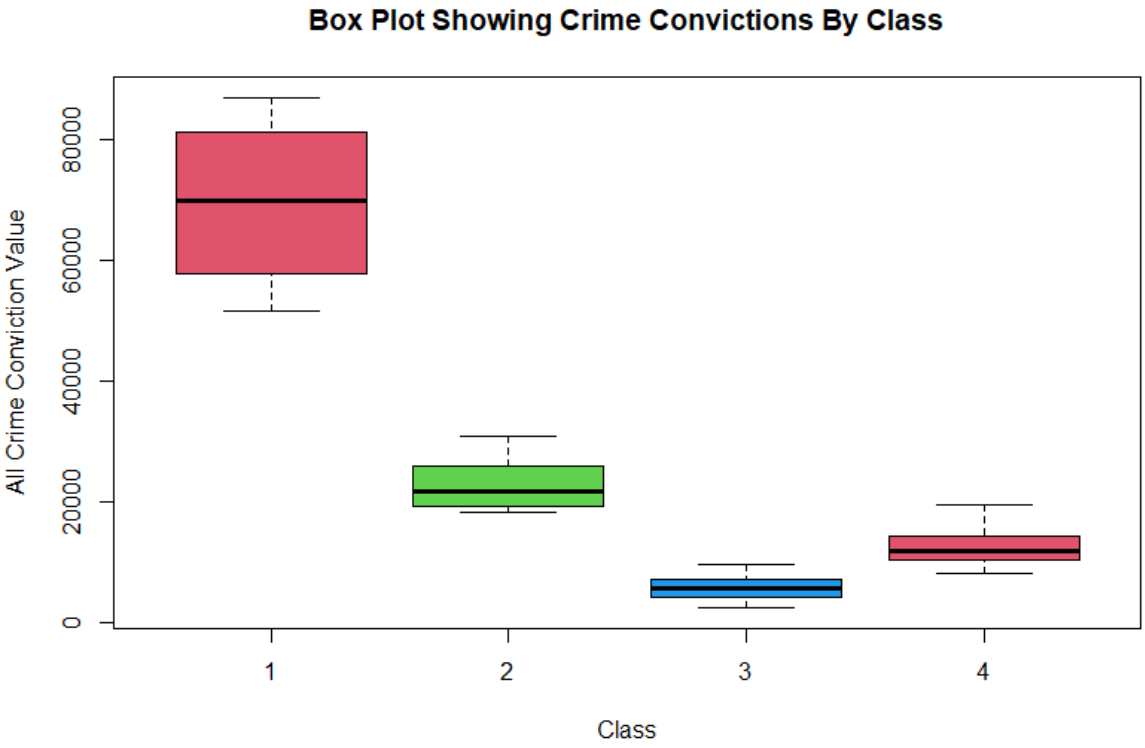

Figure 32: A Box plot

From the above result and figure 32, we can see that cluster 1 has the highest average value of total convictions as 69572. We need to further investigate to know which counties are in cluster 1 (the first cluster).

	County
1	Metropolitan and City
2	Metropolitan and City
3	Metropolitan and City
4	Metropolitan and City

Our investigation shows that there are four counties in the first cluster, and they are all Metropolitan and City.

4.2.2 CONCLUSION.

Since the Metropolitan and City have higher average value of total convictions (69572) than other counties, we accept H_0, and hence, we conclude that using K-Means clustering,

Metropolitan and City have the highest average crime rate (convictions) compared to other counties in the UK.

4.3 CLASSIFICATION

We will study how better the K-nearest neighbour classification algorithm model our data set using its accuracy level to conclude. We will be testing the hypothesis that the K-nearest neighbour classification algorithm model our data better.

Statement of hypothesis:

H_0: K-nearest neighbour classification algorithm is better than the Decision tree classification algorithm for our data set.

Against

H_1: Decision tree classification algorithm is better than the K-nearest neighbour classification algorithm for our data set.

4.3.1 K-NEAREST NEIGHBOUR ALGORITHM

Here we row bind all the yearly data sets of 2014, 2015, 2016, and 2017 together to get combined information for the four years. Below is a head view of the data.

	Class	Number of Homicide Convictions	Number of Offences Against The Person Convictions
1	4	17	2706
2	3	17	906
3	3	3	1188
4	4	11	2051
5	4	11	1272
6	3	1	1126

	Number of Sexual Offences Convictions	Number of Burglary Convictions	Number of Robbery Convictions
1	429	520	98
2	63	156	109
3	84	202	56
4	185	330	54
5	144	403	73
6	86	138	20

	Number of Theft And Handling Convictions	Number of Fraud And Forgery Convictions
1	3550	205
2	1071	91
3	1317	118
4	2434	159
5	3251	104
6	1176	52

	Number of Criminal Damage Convictions	Number of Drugs Offences Convictions
1	854	1453
2	219	442
3	297	493
4	475	852
5	506	736
6	434	501

	Number of Public Order Offences Convictions	Number of All Other Offences (excluding Motoring) Convictions
1	1152	927
2	314	132
3	524	223
4	952	595
5	1223	305
6	556	726

We will normalise our data set and split them into train and test. Below is a head view of the normalised data set.

	Number.of.Homicide.Convictions	Number.of.Offences.Against.The.Person.Convictions
1	0.089005236	0.12455857
2	0.089005236	0.01681930
3	0.015706806	0.03369845
4	0.057591623	0.08535344
5	0.057591623	0.03872628
6	0.005235602	0.02998743

	Number.of.Sexual.Offences.Convictions	Number.of.Burglary.Convictions	Number.of.Robbery.Convictions
1	0.26211015	0.16952242	0.049129990
2	0.01924353	0.03682100	0.054759468
3	0.03317850	0.05359096	0.027635619
4	0.10019907	0.10025520	0.026612078
5	0.07299270	0.12686839	0.036335722
6	0.03450564	0.03025884	0.009211873

	Number.of.Theft.And.Handling.Convictions	Number.of.Fraud.And.Forgery.Convictions
1	0.21113887	0.069858713
2	0.04813256	0.025117739
3	0.06430826	0.035714286

We go further to splitting the data, 70% to train and 30% to test. Below is a head view of both the train and test data.

Train data set.

	Number.of.Homicide.Convictions	Number.of.Offences.Against.The.Person.Convictions
20	0.10471204	0.18195966
104	0.12041885	0.11941103
102	0.06806283	0.07015024
103	0.34031414	0.11959059
142	0.07329843	0.11498174
105	0.15706806	0.05315137

	Number.of.Sexual.Offences.Convictions	Number.of.Burglary.Convictions	Number.of.Robbery.Convictions
20	0.2149967	0.28180824	0.10951894
104	0.1108162	0.14363835	0.04708291
102	0.1426676	0.08421436	0.04196520
103	0.1267419	0.08931826	0.05322416
142	0.1579297	0.07437113	0.02661208
105	0.1373590	0.04046664	0.02098260

	Number.of.Theft.And.Handling.Convictions	Number.of.Fraud.And.Forgery.Convictions
20	0.23586270	0.06789639
104	0.13111520	0.07339089
102	0.07976065	0.02864992

Test data set.

	Number.of.Homicide.Convictions	Number.of.Offences.Against.The.Person.Convictions
1	0.08900524	0.12455857
4	0.05759162	0.08535344
15	0.04188482	0.02208655
16	0.06282723	0.17160472
23	0.13612565	0.14652541
28	0.03141361	0.04201832

	Number.of.Sexual.Offences.Convictions	Number.of.Burglary.Convictions	Number.of.Robbery.Convictions
1	0.26211015	0.16952242	0.04912999
4	0.10019907	0.10025520	0.02661208
15	0.02853351	0.06635071	0.02098260
16	0.16721964	0.13525337	0.05066530
23	0.18380889	0.13889902	0.10081883
28	0.05441274	0.06562158	0.01944729

	Number.of.Theft.And.Handling.Convictions	Number.of.Fraud.And.Forgery.Convictions
1	0.21113887	0.06985871
4	0.13775644	0.05180534
15	0.06503156	0.01766091
16	0.21902946	0.07653061
23	0.21455813	0.09850863
28	0.10284061	0.02040816

Next is to perform the k-nearest neighbour classification. We need to select our k value, so we went ahead to perform a test for k value selection. In this test, we see several k values and the accuracy level they will give. Below is the result from our test.

```
1 = 86.27451
2 = 86.27451
3 = 86.27451
4 = 86.27451
5 = 82.35294
6 = 82.35294
7 = 82.35294
8 = 84.31373
9 = 84.31373
10 = 86.27451
11 = 84.31373
12 = 84.31373
```

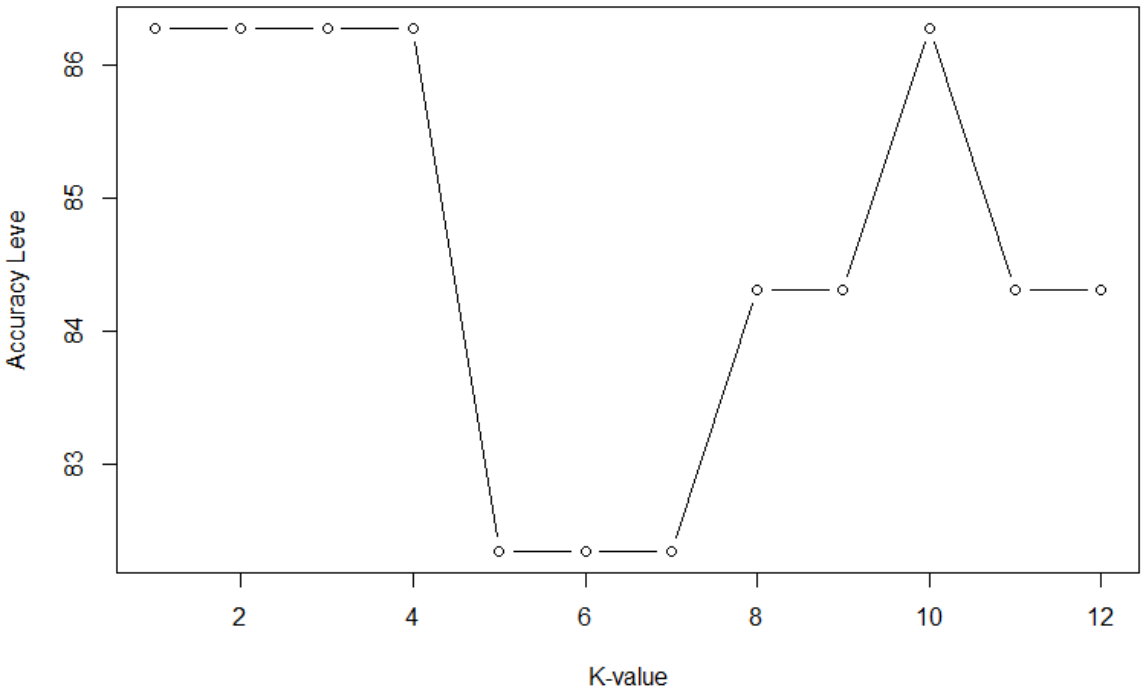

Figure 33: A line plot of accuracy levels against k-values.

From our result and figure 32, we notice that k values for 1, 2, 3, 4, and 10 will give the best accuracy level, so we will choose 4 as our k value and use it to perform the k-nearest neighbour classification. Below is the output of our k-nearest neighbour classification.

```
 [1] 4 3 3 4 4 3 4 3 3 4 3 3 3 3 4 4 3 4 1 3 3 3 4 3 4 3 3 3 3 3 4 3 4 4 3 3 3 2 4 3 3 4 4 3 3
[46] 4 3 3 3 3 4
Levels: 1 2 3 4
```

We can see that the k-nearest neighbour algorithm has classify the class into 4.

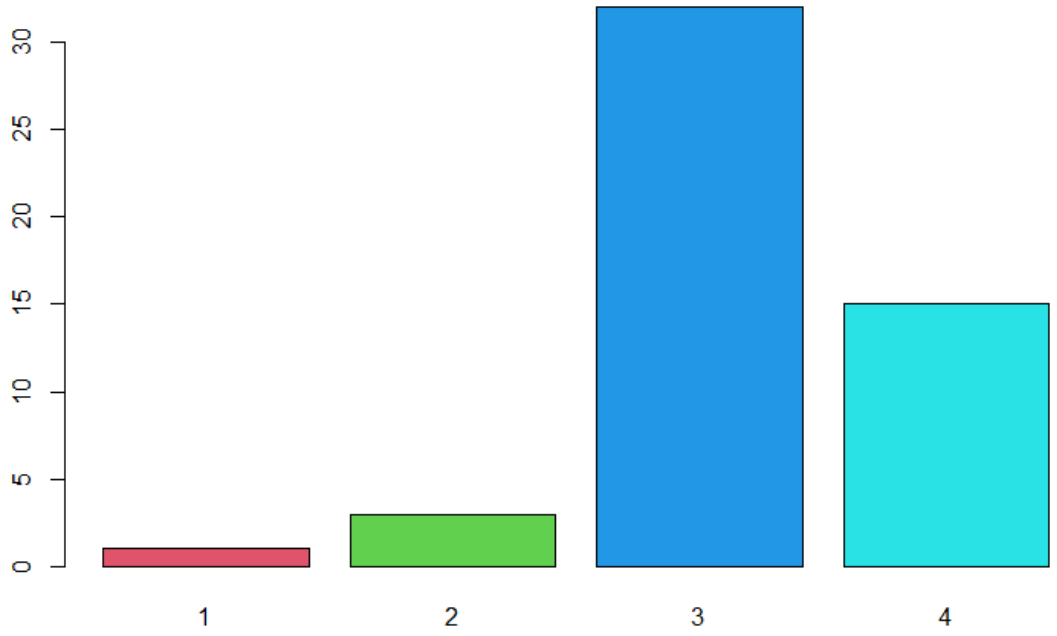

Figure 34: A Bar chart of class

Figure 34 give a clearer picture of the classes, and we can see that the class that have the highest number of members or neighbours is class 3 followed by class 4. We forge ahead to calculate the table of the predicted and the test data. Below is the table of the k-nearest neighbour algorithm.

```
    vn1_test_class
pred  1  2  3  4
   1  1  0  0  0
   2  0  1  0  0
   3  0  0 27  4
   4  0  1  1 16
```

From the table, we can see that classes one and two were well classified, in class three, we can see that 4 neighbours were wrongly classified, and class four had 2 neighbours wrongly classified. we need to perform the confusion matrix on the table to be able to check its accuracy.

```
Confusion Matrix and Statistics

    vn1_test_class
pred  1  2  3  4
   1  1  0  0  0
   2  0  1  0  0
   3  0  0 27  4
   4  0  1  1 16

Overall Statistics

               Accuracy : 0.8824
                 95% CI : (0.7613, 0.9556)
    No Information Rate : 0.549
    P-Value [Acc > NIR] : 3.423e-07

                  Kappa : 0.7766

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           1.00000  0.50000   0.9643   0.8000
Specificity           1.00000  1.00000   0.8261   0.9355
Pos Pred Value        1.00000  1.00000   0.8710   0.8889
Neg Pred Value        1.00000  0.98000   0.9500   0.8788
Prevalence            0.01961  0.03922   0.5490   0.3922
Detection Rate        0.01961  0.01961   0.5294   0.3137
Detection Prevalence  0.01961  0.01961   0.6078   0.3529
Balanced Accuracy     1.00000  0.75000   0.8952   0.8677
```

From the above result, we can see that the K-nearest neighbour algorithm have 88.24% Accuracy level.

4.3.2 DECISION TREE ALGORITHM

For the decision tree algorithm, we will be using the same data set used for the k-nearest neighbour algorithm. To perform decision tree classification, firstly, we will split our data set into train and test. Below is a head view of both the train and test data set.

For the train data set.

```
    Class Number of Homicide Convictions Number of Offences Against The Person Convictions
20      2                             20                                              3665
104     4                             23                                              2620
102     3                             13                                              1797
103     4                             65                                              2623
142     3                             14                                              2546
105     3                             30                                              1513
    Number of Sexual Offences Convictions Number of Burglary Convictions
20                                    358                             828
104                                   201                             449
102                                   249                             286
103                                   225                             300
142                                   272                             259
105                                   241                             166
    Number of Robbery Convictions Number of Theft And Handling Convictions
20                            216                                     3926
```

For the test data set.

	Class	Number of Homicide Convictions	Number of Offences Against The Person Convictions	Number of Sexual Offences Convictions	Number of Burglary Convictions	Number of Robbery Convictions	Number of Theft And Handling Convictions
1	4	17	2706	429	520	98	3550
4	4	11	2051	185	330		
15	3	8	994	77	237		
16	4	12	3492	286	426		
23	4	26	3073	311	436		
28	3	6	1327	116	235		

We can now build our decision tree model with the train data and use it (the model) for prediction. Below is a plot of our decision tree model.

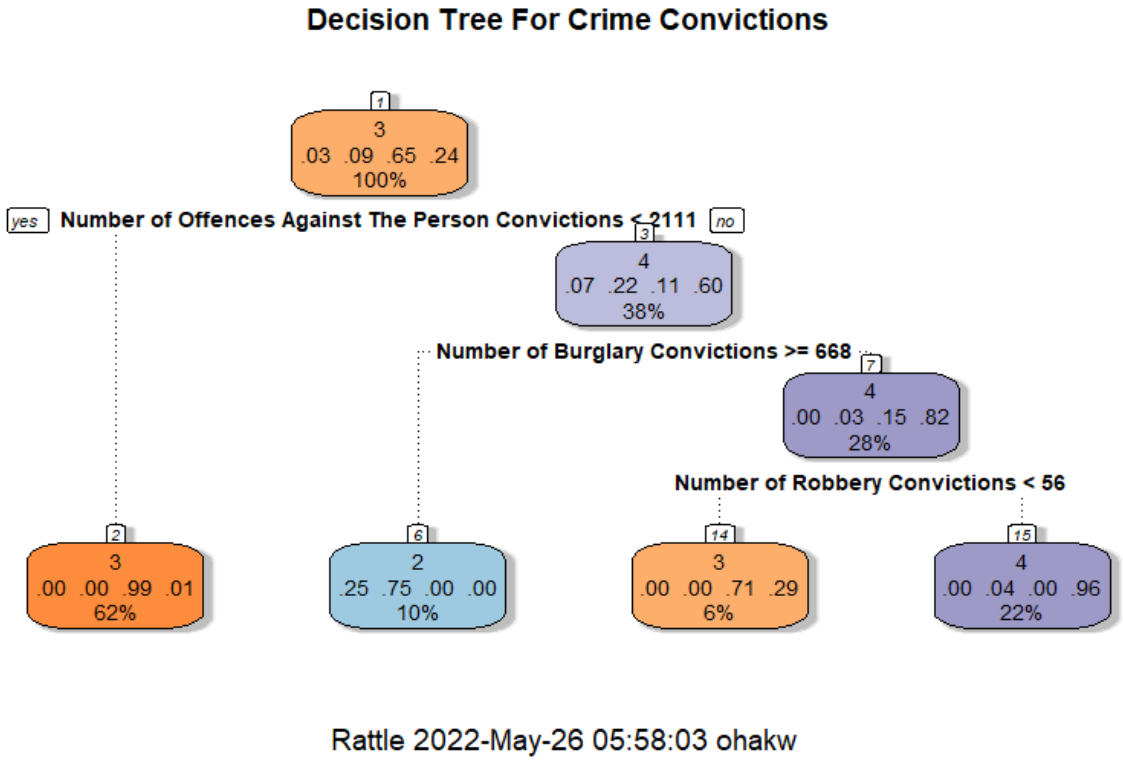

Figure 35: A Decision Tree plot.

From figure 35, the dependent variable of this decision tree is class which has four classes, 1, 2, 3, and 4. the root of the tree (number of offences against person convictions) have 2111 observations from our data set.

We will forge ahead to prediction using our decision tree model and calculate the table of the predicted and the test data. Below is the table of the decision tree algorithm.

```
   pred1
     1  2  3  4
1  0  1  0  0
2  0  1  0  1
3  0  0 26  2
4  0  0  3 17
```

Our findings show that class 1 had 1 member misclassified, class 2 had 1 member that was misclassified, class 3 had 2 members misclassified, and class 4 had 3 members misclassified. We go ahead to calculate the accuracy level using the confusion matrix and below is our result.

```
Confusion Matrix and Statistics

   pred1
     1  2  3  4
  1  0  1  0  0
  2  0  1  0  1
  3  0  0 26  2
  4  0  0  3 17

Overall Statistics

               Accuracy : 0.8627
                 95% CI : (0.7374, 0.943)
    No Information Rate : 0.5686
    P-Value [Acc > NIR] : 6.538e-06

                  Kappa : 0.7422

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity               NA  0.50000   0.8966   0.8500
Specificity           0.98039  0.97959   0.9091   0.9032
Pos Pred Value            NA  0.50000   0.9286   0.8500
Neg Pred Value            NA  0.97959   0.8696   0.9032
Prevalence            0.00000  0.03922   0.5686   0.3922
Detection Rate        0.00000  0.01961   0.5098   0.3333
Detection Prevalence  0.01961  0.03922   0.5490   0.3922
Balanced Accuracy         NA  0.73980   0.9028   0.8766
```

From our result above, we can see that the accuracy level for our decision tree model is 86.27%.

4.3.3 CONCLUSION.

Since the K-nearest neighbour classification algorithm percentage value of accuracy level is greater than the Decision tree classification algorithm with 88.24%, we accept H_0 and hence, we conclude that the K-nearest neighbour classification algorithm is better than the Decision tree classification algorithm for our data set.

4.3.4 MODEL IMPROVEMENT

Note that we can further improve the better classification model. This can be done by cross validation, which means we are going to perform the K-nearest neighbour classification for several values of k and choose the best model with accuracy level, below are the results from our test.

```
k-Nearest Neighbors

168 samples
 12 predictor
  4 classes: '1', '2', '3', '4'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 135, 134, 134, 134, 135
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9046346  0.8112374
   2  0.9162210  0.8362311
   3  0.8926916  0.7881682
   4  0.9222816  0.8468223
   5  0.9105169  0.8237546
   6  0.9105169  0.8237546
   7  0.9169340  0.8370831
   8  0.8928699  0.7880759
   9  0.8868093  0.7730335
  10  0.8746881  0.7445285

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 4.
```

Result from our cross validation (for model improvement), suggest if the model is further improved, we will have the model at an optimal level when k is equal to 4 with 92.23% as accuracy level.

CHAPTER FIVE

GENERAL CONCLUSION AND RECOMMENDATIONS

5.1 GENERAL CONCLUSION

This work consists of data cleaning, manipulation, visualization, descriptive analysis, and modelling using machine learning. In the data cleaning, we imported all the monthly data sets for 2014, 2015, 2016, and 2017 and we cleaned them by removing the columns with percentages, and for data manipulation, all the data sets for all months in each year were added together using data frame (or matrix) operation to get an annually (yearly) data sets for 2014, 2015, 2016, and 2017. For visualization, we used both Bar charts (including grouped and stacked Bar charts) and Pie charts to visualise or represent our data sets for each year, although, there were other visualisations we explored while modelling our data, such as Box plots, scatter plot, dendrogram, cluster plot, decision tree plot, line plot, etc. We did a descriptive analysis to see the mean, median, minimum, and maximum values for each year's data set. We used the Linear model, Clustering, and Classification for modelling. In linear modelling, we studied the Akash linear model function (built by us) and the R inbuilt linear model function and from our investigations, we found out that the R inbuilt function is better when compared to our new function (Akash linear model function). We went further to build a linear model for machine learning purposes where we predicted the number of offences against the person convictions using the number of sexual offences convictions and the number of motoring offences convictions as our explanatory variables, our study shows that an increase in the number of sexual offences convictions and the number of motoring offences convictions will lead to an increase in the number of offences against the person convictions. For the Clustering, we study both the K-Means clustering algorithm and we found out that the Metropolitan and City had more criminal convictions compared to other counties. In the Classification model, we compared the K-Nearest Neighbor classification algorithm and the Decision Tree classification algorithm using our data set and our investigations show that the K-Nearest Neighbor classification algorithm was a better model for our data seeing its accuracy level, and we did cross validation for model improvement. Generally, we can conclude from our research that the Metropolitan and City have the highest total and an average number of criminal convictions compared to other counties in the UK, and we also conclude from our research that there are more crime offences against person than other crimes in the UK.

5.2 RECOMMENDATIONS

Here are some recommendations based on this research.

- From our research, we recommend that the UK Government should improve security in the Metropolitan and City (such as London and the rest) to leverage or reduce the high rate of crime in these areas.
- We recommend that the citizen of the UK should be more helpful in providing the crime authorities (like the police) with vital crime information at their disposal to aid the society with more security.

REFERENCE

[1] Douglas, C.M, Elizabeth, A.P, Vining, G.G (2012), Introduction to Linear Regression Analysis.

[2] Rose Ihaka, Wilkinson, and Rogers, (1973), Lm (R inbuilt Linear Function).

[3] Matthew Heeks, Sasha Reed, Mariam Tafsiri and Stuart Prince, (2018), The economic and social costs of crime. *Home Office.* Research Report 99.

[4] Oxford University Press, (2016), Social disadvantage, crime, and punishment. *London School of Economics and Political Science Research Online.* pp. 322-340.

[5] Grahame Allen and Yago Zayed, (2021), Homicide Statistics. *House of Commons Library.* Number 8224.

[6] Nick Morgan, Oliver Shaw, Jennifer Mailley, and Rebecca Channing, (2020), Trends and drivers of homicide. *Home Office.* Research Report 113.

[7] Grahame Allen, Richard Tunnicliffe, (2021), Drug Crime: Statistics for England and Wales. *House of Commons Library.* Number 9029.

[8] Wikipedia, (2021), Breusch-Pagan Test for Homoscedasticity. *www.wikipedia.com*.

[9] Wikipedia, (2022), Shapiro-Wilk Test for Normality. *www.wikipedia.com*.

[10] Wikipedia, (2022), Variance Inflation Factor Test for Multicollinearity. *www.wikipedia.com*.

[11] Wikipedia, (2021), Durbin-Watson Test for Autocorrelation. *www.wikipedia.com*.

[12] Wikipedia, (2021), Root Mean Square Error. *www.wikipedia.com*.

Appendix