CT7202 DATA ANALYSIS AND VISUALISATION PRINCIPLES (SEM2 - 2022/23)

DATA ANALYSIS AND VISUALISATION PRINCIPLES ASSESSMENT

CRITICAL DATA VIUSALIZATION ANALYSIS OF CROWN PROSECUTION SERVICE CASE OUTCOMES DATA SET GROUPED BY PRINCIPAL OFFENCE CATEGORIES

BY

OLUWATOYIN ODENIYI (s4115181)

TABLE OF CONTENT

INTRODUCTION

The data set is to be analysed is from a twenty-four-month period is the number of offences which were convicted and some of which were unsuccessful (no convictions). The 25 offences range from burglaries, thefts, sexual offences in 44 Counties in the UK for specific months in the years 2014,2015,2016,2017. Observing the data set's numeric values, the number of unsuccessful convictions of offences reduce as the number of successful convictions of the same convictions increased in the UK Counties. The first step in this assessment which is also the foundation of all data analysis is to decide on the hypotheses which will serve as a guide for this assessment. Observing the data through a thorough study of the numeric values(observations) for the variables, it was possible to tentatively decide which variables will be used for data analysis, considering the data set is a large one.

RESEARCH QUESTION AND HYPOTHESES

The main research question is: How effective is the successful convictions of offences by the Crown Prosecution in reducing the occurrence of the same offences in the Counties?

The Null Hypothesis (H_0): The number of unsuccessful offences were not influenced by the number of convictions of the same offences (There is no relationship between the two specific variables).

The Alternative Hypothesis (H_a): The number of unsuccessful offences reduced as more of the same offences were convicted (The frequency of unsuccessful offences (no convictions) in the Counties reduced as the convictions increased).

Hypotheses testing using the Probability value (P value) method:

From the calculation shown in the R studio workspace and shown below, calculating the Confidence Interval (CI) with a random sample mean size of 26 and proportion size of 0.49. The Null Hypothesis is rejected as using the CI (0.2928, 0.6821), the value of alpha was 0.05 which was less than the Probability value (P value) of 1. The interpretation of this is that the Alternative Hypothesis is accepted. **This will be the first working hypotheses for the analysis of the data set which is that: The frequency of unsuccessful offences (no convictions) in the Counties reduced as the convictions for the same offences increased.**

Based on the numeric values of the data set, another observation for a second hypothesis is: The higher the number of theft and handling convictions, the lesser the number of homicide convictions in all Counties.

FIRST HYPOTHESIS

TESTING THE FIRST HYPOTHESIS

```
# Using the Two Tailed Test:
```

Picking two variables and assigning random mean size and proportion size

```
'NSexualOffencesUnsuccessful' = 26

'NSexualOffencesConvictions' = 0.49


# Find the margin of error

alpha = 0.05      # replace with your chosen alpha (it is a 95% confidence level)

moe = qnorm(1-alpha/2, 0, 1) * sqrt(NSexualOffencesConvictions*(1-
NSexualOffencesConvictions)/NSexualOffencesUnsuccessful)


# Find the confidence interval
```

```
upper_bound <- NSexualOffencesConvictions + moe

lower_bound <- NSexualOffencesConvictions - moe

lower_bound

upper_bound

#95% Confidence Interval is [0.2928, 0.6821]
```

Note that the:

```
#P-value is the Probability value and ranges between 0 and 1.

#P<alpha ; reject the null hypothesis and using the Confidence Interval; alpha is greater
hence the null hypothesis is rejected.

#P>alpha ; accept the null hypothesis.
```

The number of Unsuccessful Offences reduced as there was an increase in the number of the offences successfully convicted in the 26 UK Counties

The data set, https://data.gov.uk/dataset/89d0aef9-e2f9-4d1a-b779-5a33707c5f2c/crown-prosecution-service-case-outcomes-by-principal-offence-category-data).%C2%A0, was imported and all the excel files of the specific years(2014,2015,2016,2017) were combined so that all the data can be observed and analysed easily. The data set (final_combination) has 26 columns and 1806 entries (26 variables and 1806 observations). In this report, the hypotheses that the higher the number of convictions by the Crown Prosecution Service, in the UK Counties, the less the number of unsuccessful convictions.

DATA WRANGLING

Importing and cleaning the data set

The data set was imported and merged so that all the files were all under one file. This made the data from all the years to be combined as one data set. This was easily verified to be successful after using the necessary function, as all the files had the same column names. The number of variables (column names) remained the same after the merger, however the observations(rows) increased.

The cleaning and tidying of the data set was important as raw data might not have consistent data needed for analysis. It was therefore a requirement to rename the variables, X which represented Counties in the UK and all the number of offences convicted and unsuccessful were renamed so it could be read easily by the programming language R. In R, all missing values must be removed or dropped as R can only work with values it can read. Therefore, the data set was checked for any missing values, and it did not have any. Also values with commas, like thousands, had to be worked on by removing the commas using the function: read_csv. This was done as R can easily read numeric data with no commas separating the figures. Also, there was a check for structural errors, correcting misspelled column names and changing the column names to a shorter version was done (in R, long column names throw errors as it will not be recognised as a variable). R Packages like readr and tidyverse were installed for cleaning the data set. Also checked for duplicate values as this could affect the results in R and there were no duplicate values in the data set. Columns which had the percentage values were dropped as they were unneeded since they were calculated adding two value sets which are already available. Overall, the cleaning and integration made the data set easier to understand and interpret by observation, visualization and modelling.

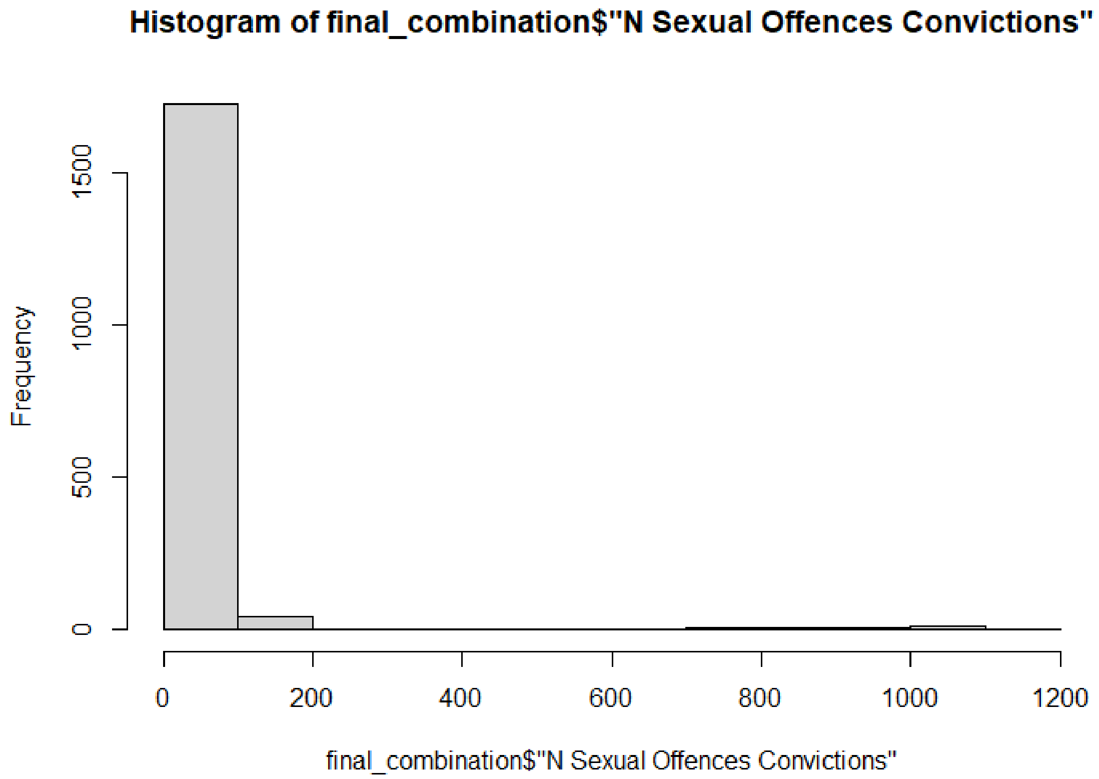
Analysis of the data set and the interpretation of the analysis

This is important in observing the numbers of offences (unsuccessful ones and the convicted ones) recorded for the Counties and collated by the Crown Prosecution for the specified years. To determine if the data set had strings or characters, to carry out some statistical analysis of the data set (mean, median, mode, minimum, maximum, percentiles). Using descriptive analytical models like histograms, pie charts, bar plots to present a clear visualization of the observations in the data set. From the statistical analysis, the mean of the offences successfully convicted was higher than the mean of the offences not successfully convicted. For offences in columns 4 and 5, the mean for Number of offences against the persons convictions was 457.2 while the mean for Number of offences against the persons unsuccessful was 139.1. The mean for the Number of sexual offences convictions is 43.61 while the mean for the unsuccessful convictions of this same offence is 16.43. After observing the table using the ‘glimpse’ function, it was clear enough to see which offences had the highest frequency of successful convictions and which had low frequency of unsuccessful convictions. Number of theft and handling convictions and Number of drug convictions were the highest, while the Number of unsuccessful convictions of Homicide and that of Burglary were the lowest. These variables had high relevance of occurrence in the data set and could be used as focal points in visualization and modelling.

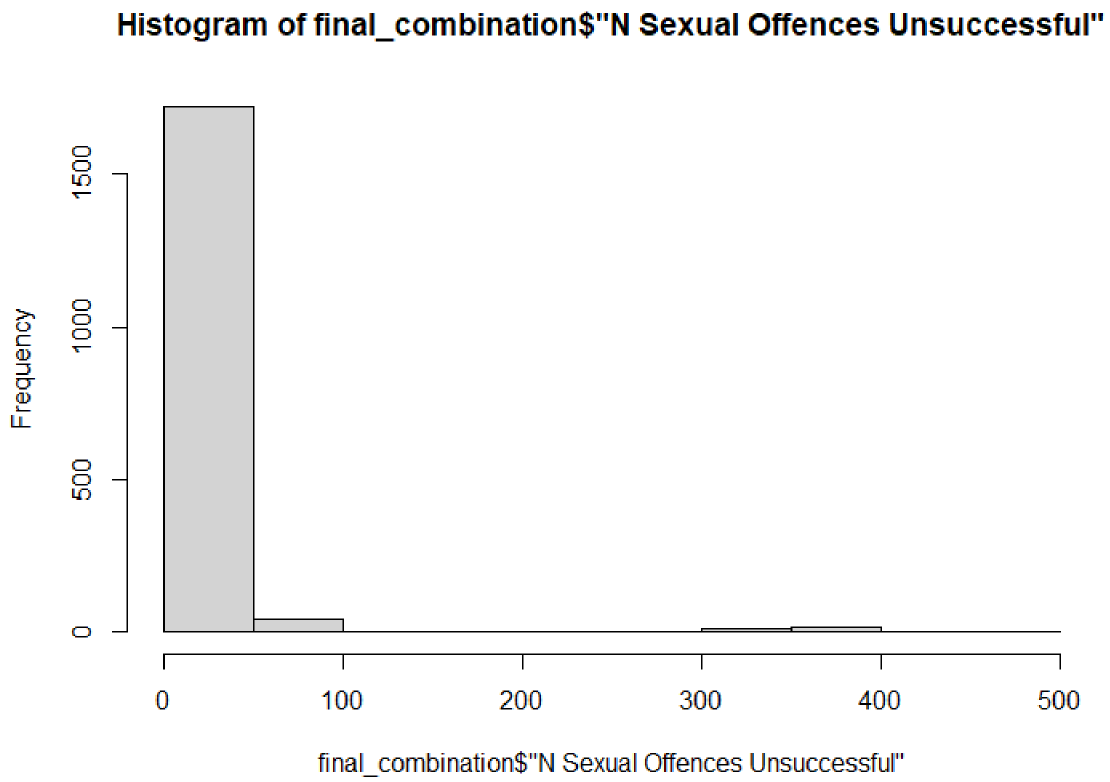
A frequency table showed how high and how low the successful and unsuccessful convictions of the offences were in the Counties. Data visualization using histogram was a good graphical representation of the frequency of the successful and unsuccessful convictions of the offences as observed in frequency table. A box plot was also a good representation for the visualization of the data set as it illustrated the distribution of the data.

Visualization of the data set using specific variables

Histogram of specific variables in the data set



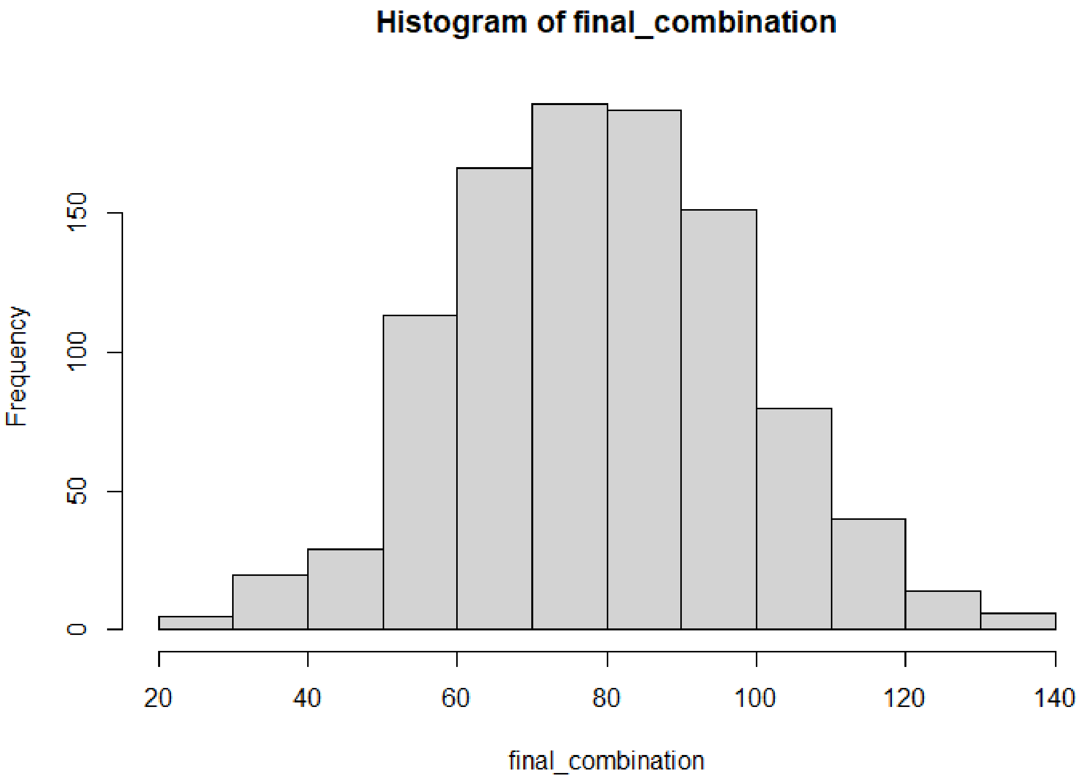
The frequency of Sexual offences convictions with numbers as high as 100 was above 1500 times in the counties for the combined years in the data set. For numbers as high as 800-1000, they occurred only singly or just a bit higher than once.

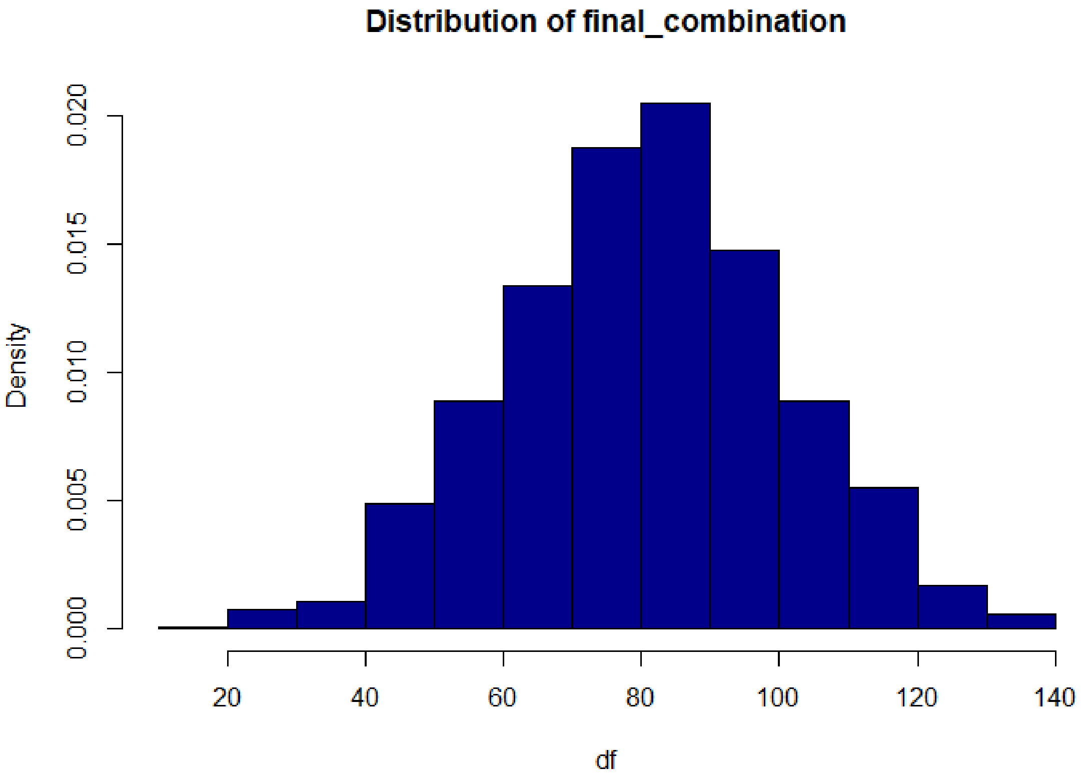


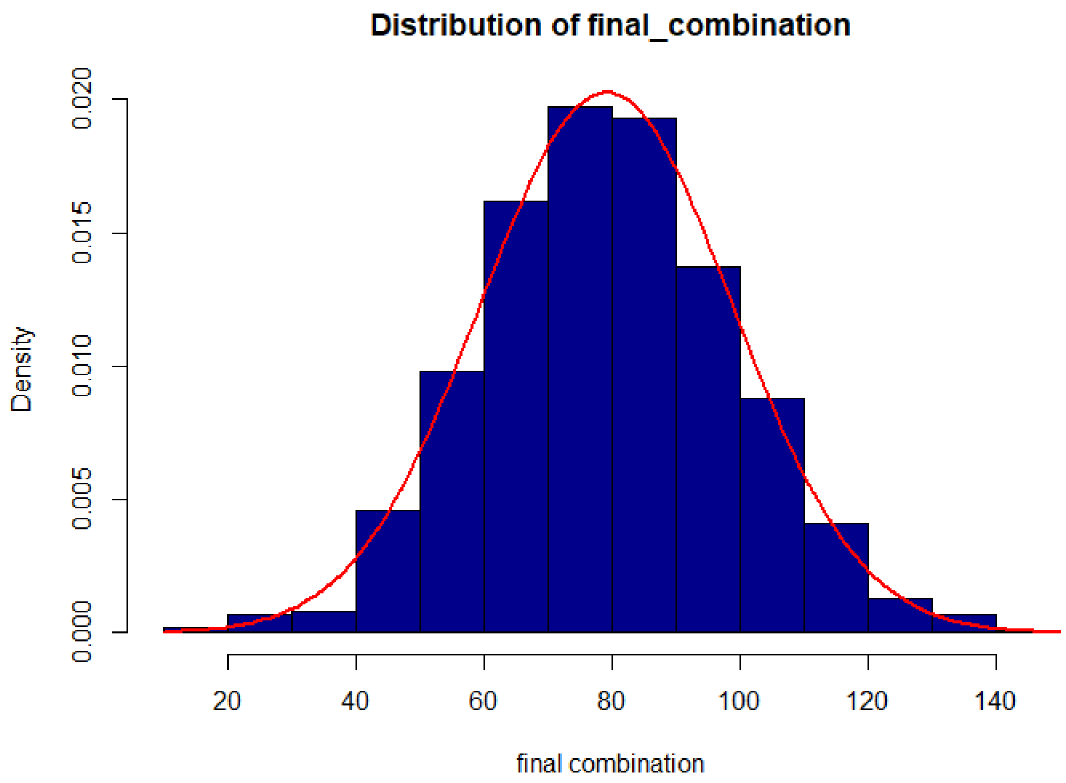
The frequency of Sexual offences unsuccessful with numbers as high as 100 was above 1500 times in the counties for the years in the data set. For numbers as high as 300-400, they occurred only singly or just a bit higher than once.

Comparing the histograms above, it shows clearly that the frequency of convicted sexual offences was higher than the unsuccessful ones of the same offence. This can be observed to be the same for all other offences in the final_combination data set.

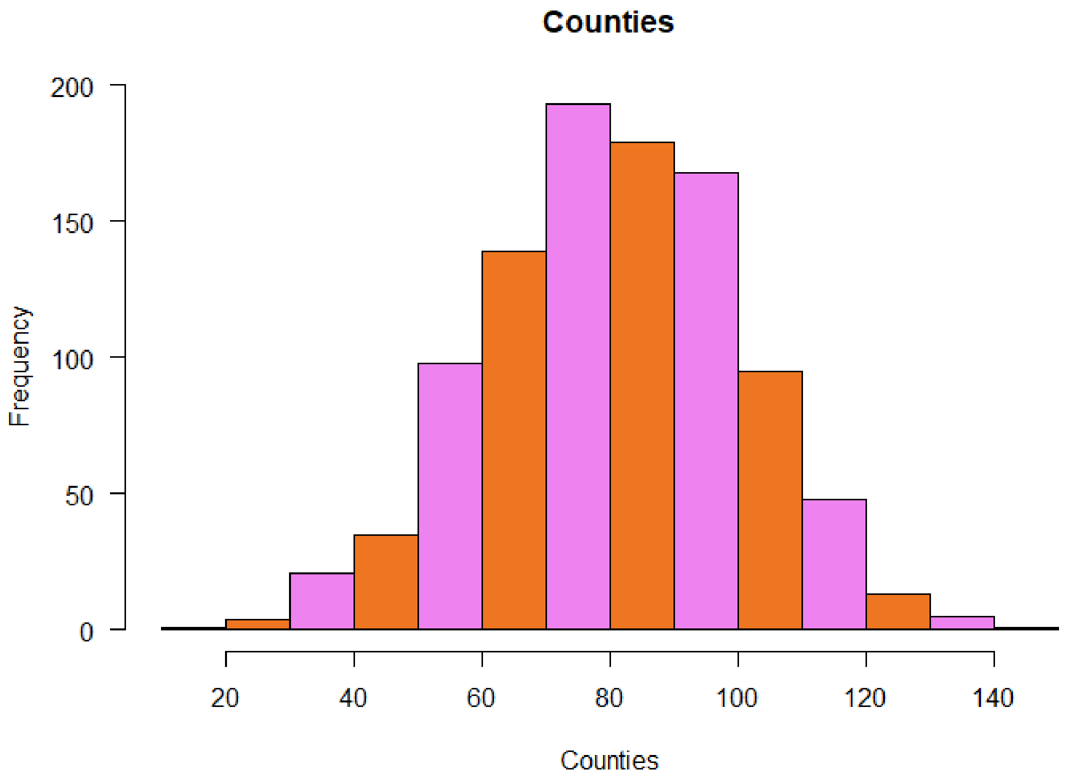
Histogram of the entire data set


Histogram of final_combination
Frequency
0
50
100
150
20
40
60
80
100
120
140
final_combination


Distribution of final_combination
Density
0.000
0.005
0.010
0.015
0.020
20
40
60
80
100
120
140
df



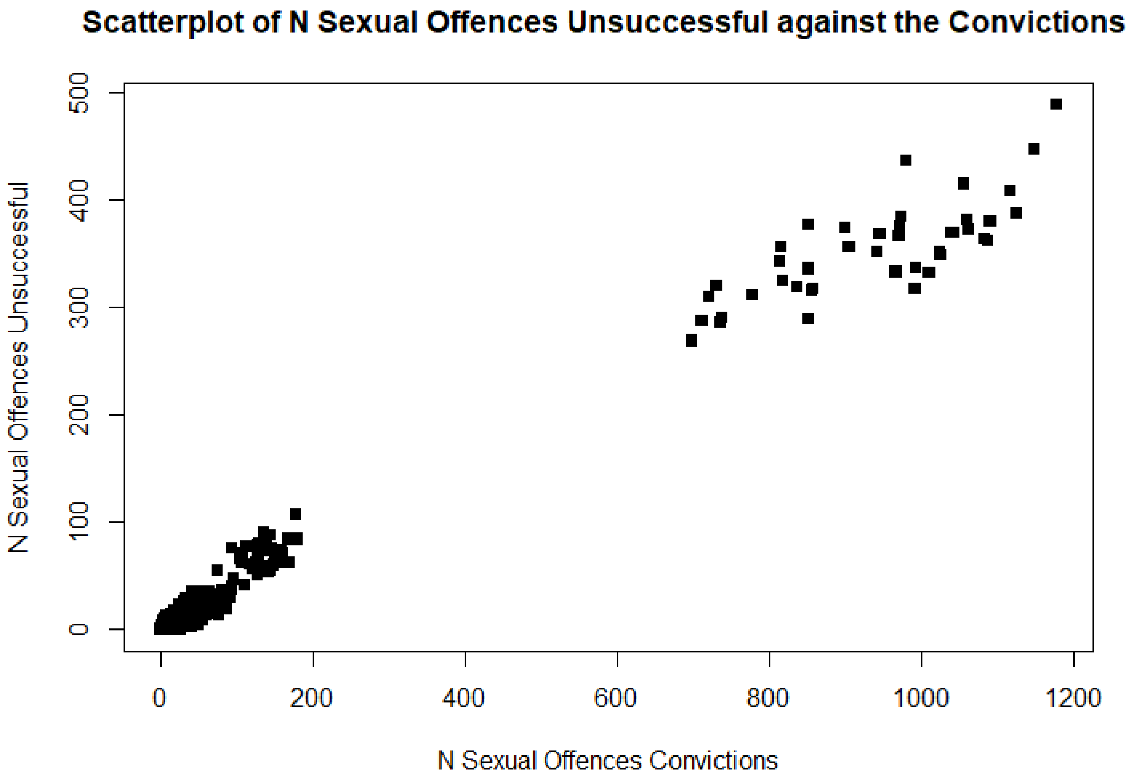
When the density of the distribution was 0.0210, the middle point(median) and mean of the distribution of the data set was around 80. The rest of the data is evenly distributed on either side of the median. Therefore, the data fit the distribution to a good degree as seen with the red line on the histogram.



Histogram showing the distribution of the Counties in the data set. Counties with frequency of offences just below 200 are in the middle (median) of the distribution.

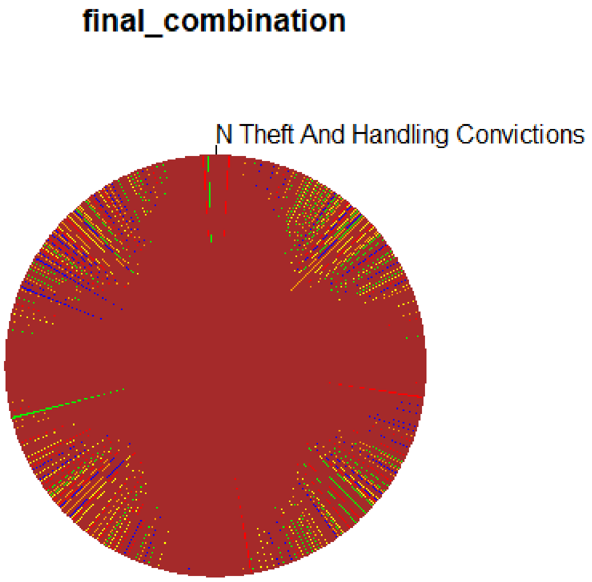
Scatterplot showing the relationship between the Number of Sexual Offences Unsuccessful and Number of Sexual Offences Convictions

The visualization below is an indication of first, a relationship and second the type of relationship between the two specific variables. The number of successful convictions of this offence has a close relationship to the number of its unsuccessful ones. For 200 successful convictions, there were about 100 unsuccessful offences. For 1200 successful convictions, there were about 500 unsuccessful offences. There is a dense number of points closer to the 0 range and its dispersed from the 800 range on the x-axis of the plot.

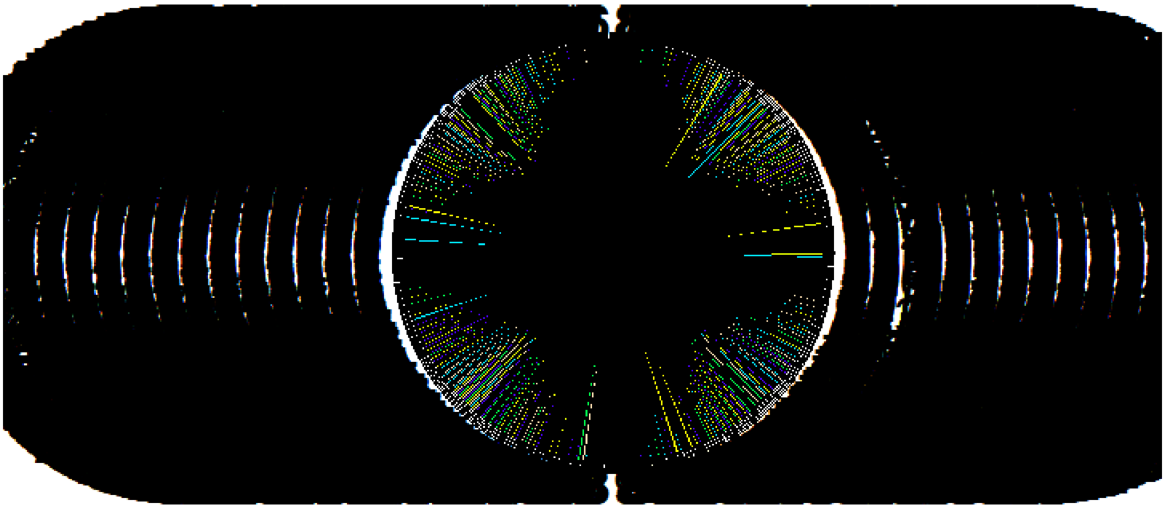


<u>Pie chart</u>

However, a pie chart was not a good option for this same data visualization as pie charts work best for small-size vector variables.

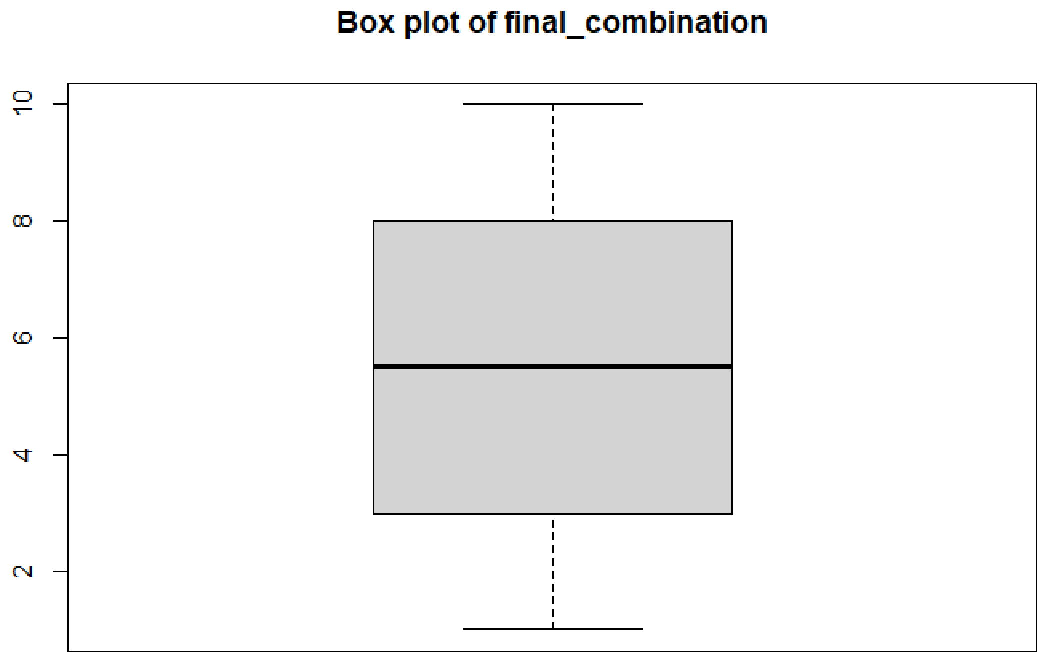


The above pie chart is not a clear visualization for the number of the specified offence in the Counties.



The pie chart is a poor visualization tool for the data set as there is no clear representation of the variables or their values.

Box Plot

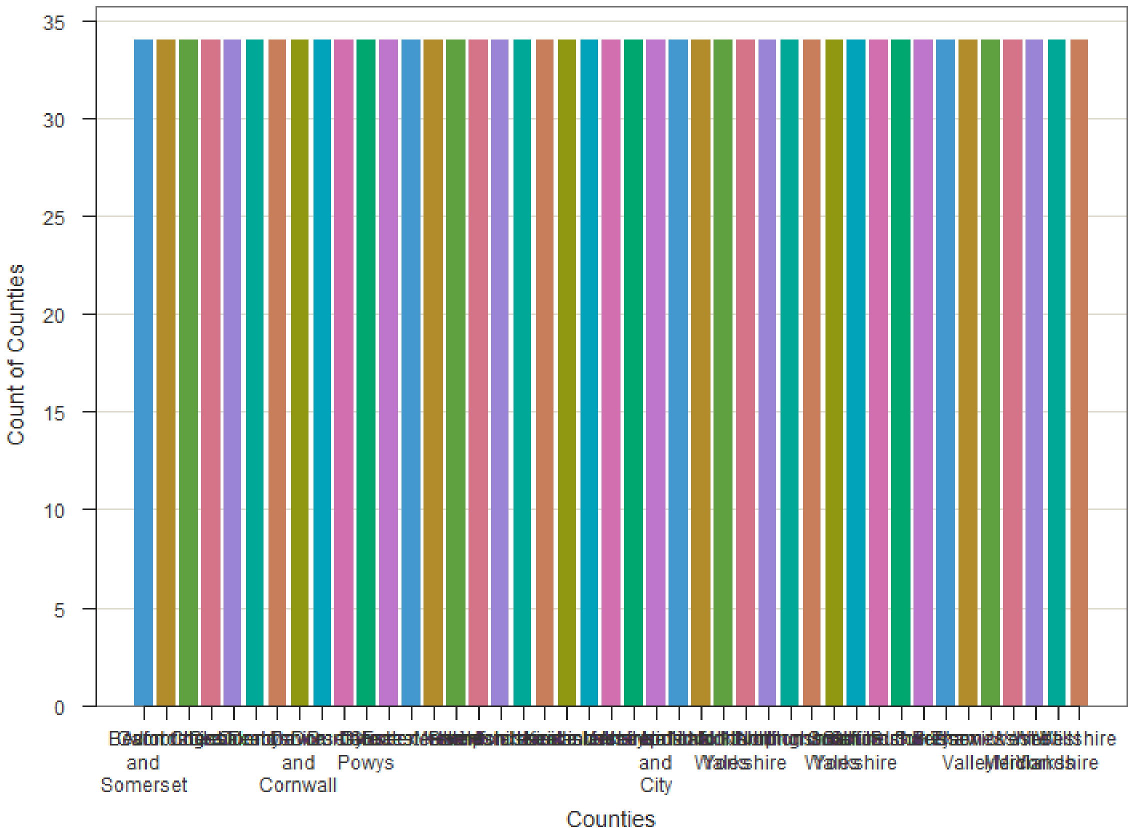


A Box plot represents 50% of the data set, with the median and starts with 1st quartile and second quartile. Here, the median of the data set is about 4.8 and there are no outliers/values varying very differently from majority of the values. The first quartile (25%) starts at 3 and the third quartile (75%) is at 8.

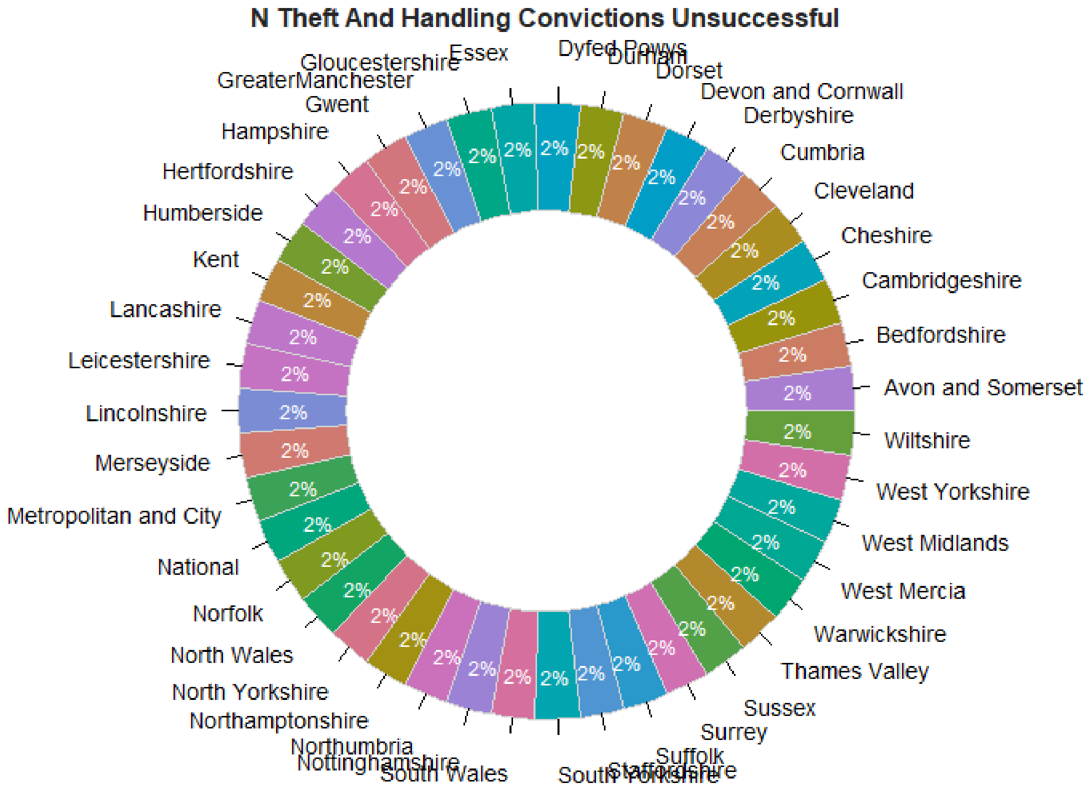
IMPLEMENTATION OF LINEAR REGRESSION, CLUSTERING AND CLASSIFICATION TECHNIQUES TO BUILD ANALYTICAL MODELS OF THE DATA

Classification technique

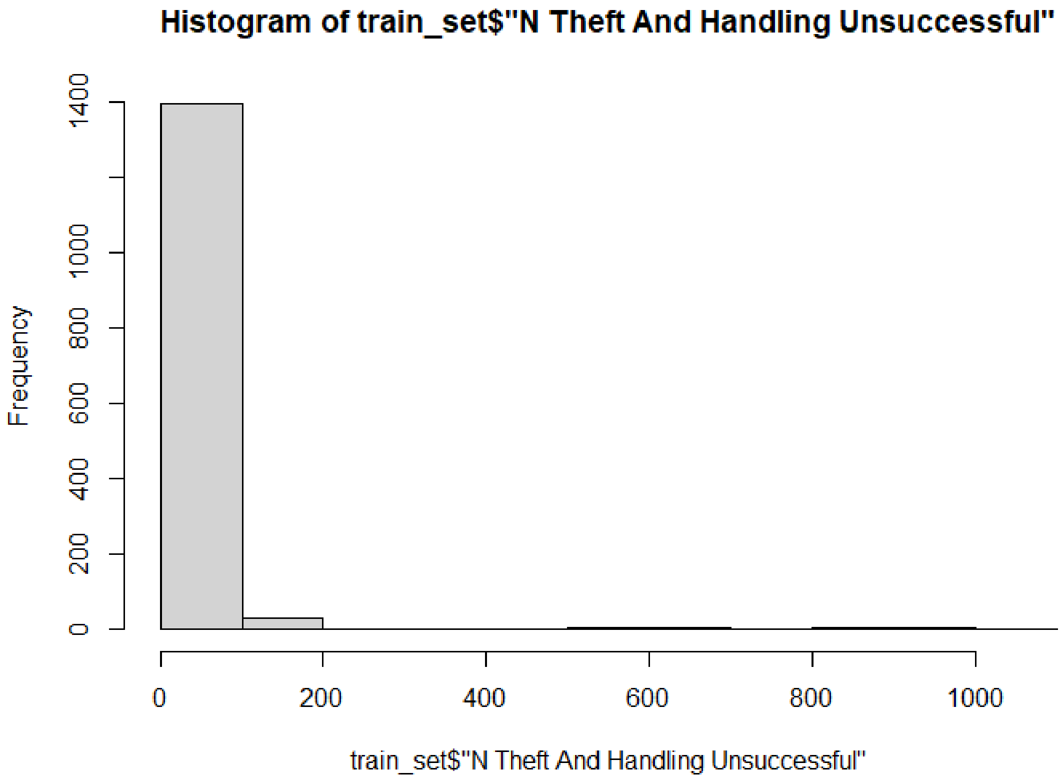
First splitting the data set into a train and test set (80% and 20% respectively). Then visualization of the train data set was carried out using bar chart, piechart and histogram.



Bar chart shows an equal distribution of the variable in the train data set.



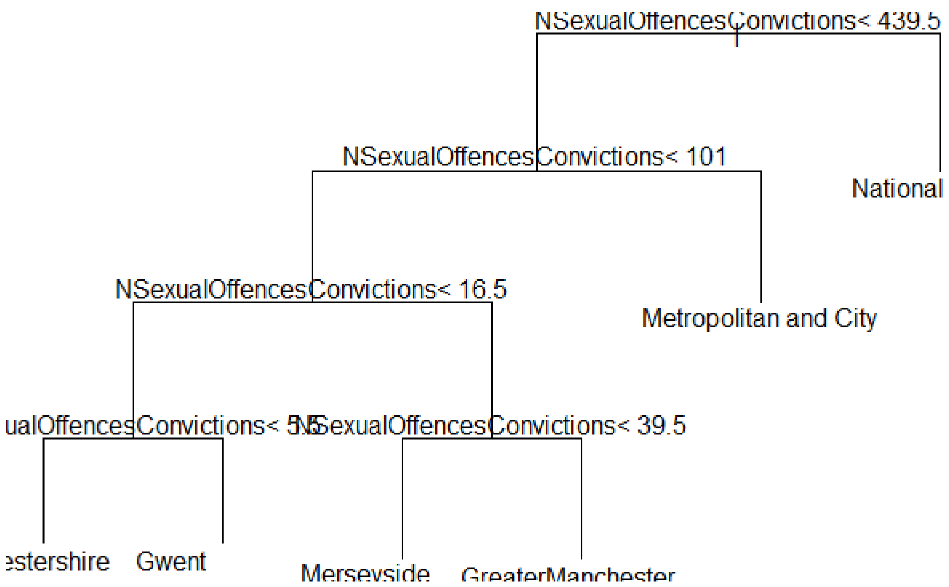
Pie chart shows an equal distribution of the specific offence in all the Counties in the train data set.

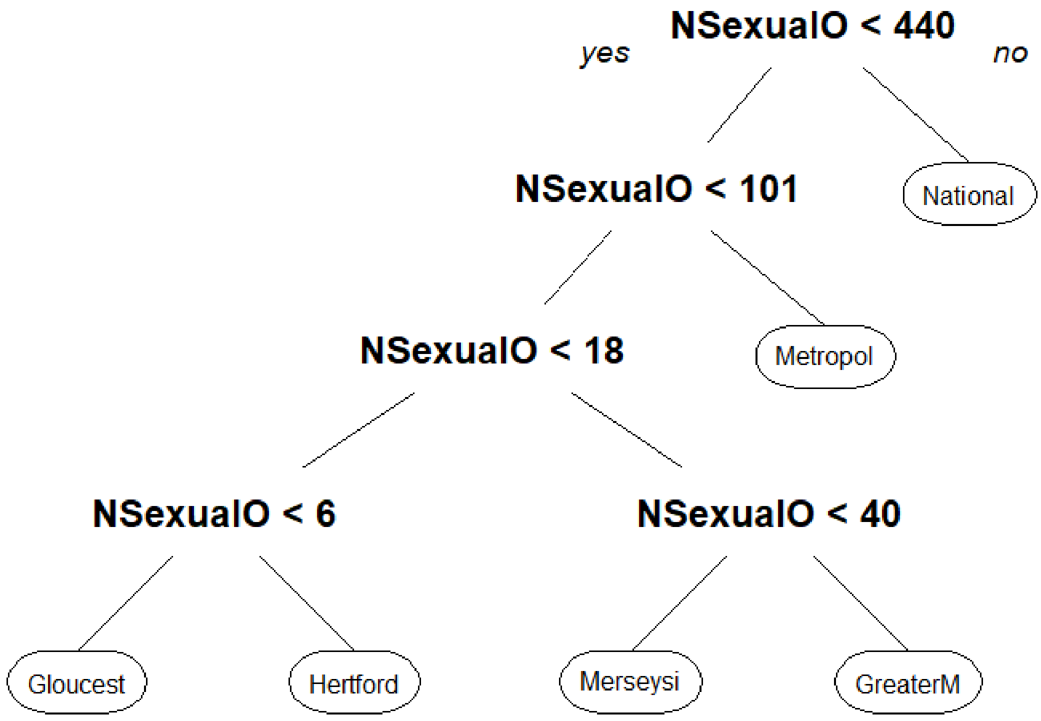


Histogram shows the frequency of this offence was1400 for about 100 of this offence and reduced from between 100 and 200, to about 50.

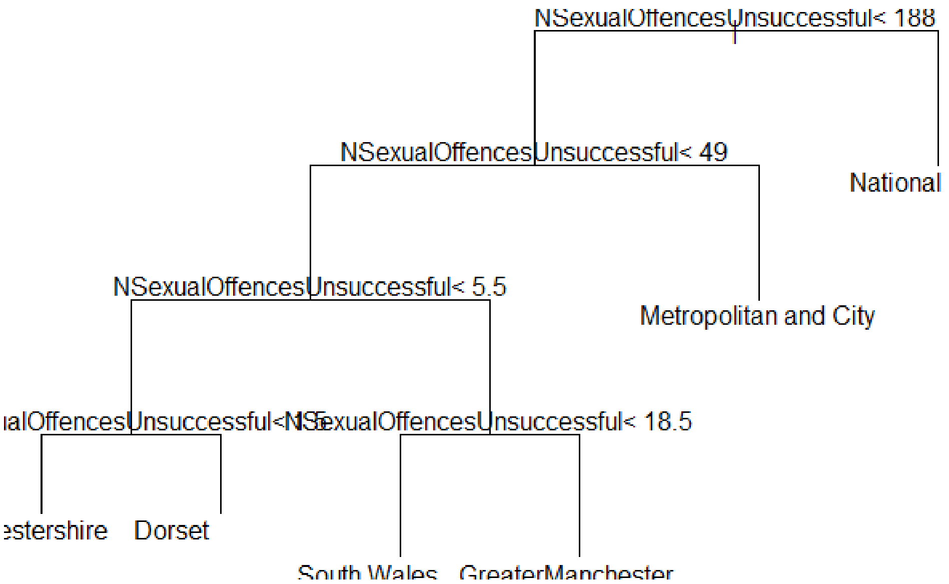
Classification with Decision Tree

With a decision tree, to specifically classify Number of sexual offences unsuccessful in the Counties and Number of sexual offences convictions in relation to the Counties:


NSexualOffencesConvictions< 439.5
NSexualOffencesConvictions< 101
National
NSexualOffencesConvictions< 16.5
Metropolitan and City
ualOffencesConvictions< 5.5
NSexualOffencesConvictions< 39.5
estershire
Gwent
Merseyside
GreaterManchester


NSexualO < 440
yes
no
NSexualO < 101
National
NSexualO < 18
Metropol
NSexualO < 6
NSexualO < 40
Gloucest
Hertford
Merseysi
GreaterM

From the two decision trees above, the most important County based on the high frequency of convictions in all the Counties under National (which is at the root of the tree) and has more than 440 number of Sexual offences convictions. In Metropolitan and City, it is more than 101. It is 117. In Gloucestershire, number of this offence is less than 6 as it is 4.


NSexualOffencesUnsuccessful< 188
NSexualOffencesUnsuccessful< 49
National
NSexualOffencesUnsuccessful< 5.5
Metropolitan and City
ualOffencesUnsuccessful< 5
NSexualOffencesUnsuccessful< 18.5
estershire
Dorset
South Wales
GreaterManchester

Here, in all the Counties under National in relation to the number of sexual offences unsuccessful, the number of these offences is greater than 188; it is 269 in number. So, this is accurate. Also, this frequency of this offence is more than 49; from the train data set it is 56. It can be deduced from these models that are a good representation of the train data set.

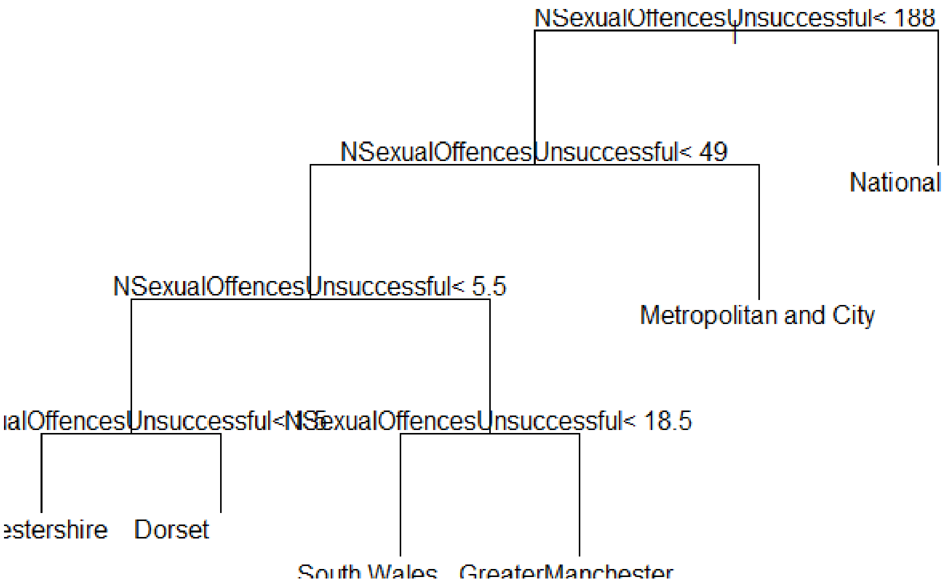
Probability of Predictions of the decision tree

```
predict(Convictions_tree, train_set, type='prob')

p1 = predict(Convictions_tree, train_set)

p1     # 23rd observation of the train data set is classified as belonging to the highest
probability which is 0.04948


#Calculate how well the decision tree model performed using the test set
```


NSexualOffencesUnsuccessful< 188
NSexualOffencesUnsuccessful< 49
National
NSexualOffencesUnsuccessful< 5.5
Metropolitan and City
ialOffencesUnsuccessful< 1.5
NSexualOffencesUnsuccessful< 18.5
estershire
Dorset
South Wales
GreaterManchester

The decision tree above shows that the number of sexual offences unsuccessful is more than 188 total in all the Counties. Observing the remaining nodes, the numbers are accurate representations of the observations of the variables in the train data set.

Linear regression:

```
linear_regression <- lm(NSexualOffencesConvictions ~ NSexualOffencesUnsuccessful, data = train_set)

print(linear_regression) #Coefficients for Intercept is 0.84 and for NSexualOffencesUnsuccessful it is 2.59

linear_regression <- lm(NSexualOffencesUnsuccessful ~ NSexualOffencesConvictions, data = train_set)

print(linear_regression)

plot(linear_regression)

 ## #Coefficients for Intercept is 20.28 for NSexualOffencesConvictions and for NSexualOffencesUnsuccessful it is -0.25.
```

The above results shows that estimated effect of ‘N Sexual Offences Unsuccessful’ on ‘N Sexual Offences Convictions’ is 2.59 and in the second code, for the other way round, it was 0.038. This shows that the number of Unsuccessful sexual offences is strongly related to the number of Sexual offences convictions in the Counties.

From the summary of the linear regression model:

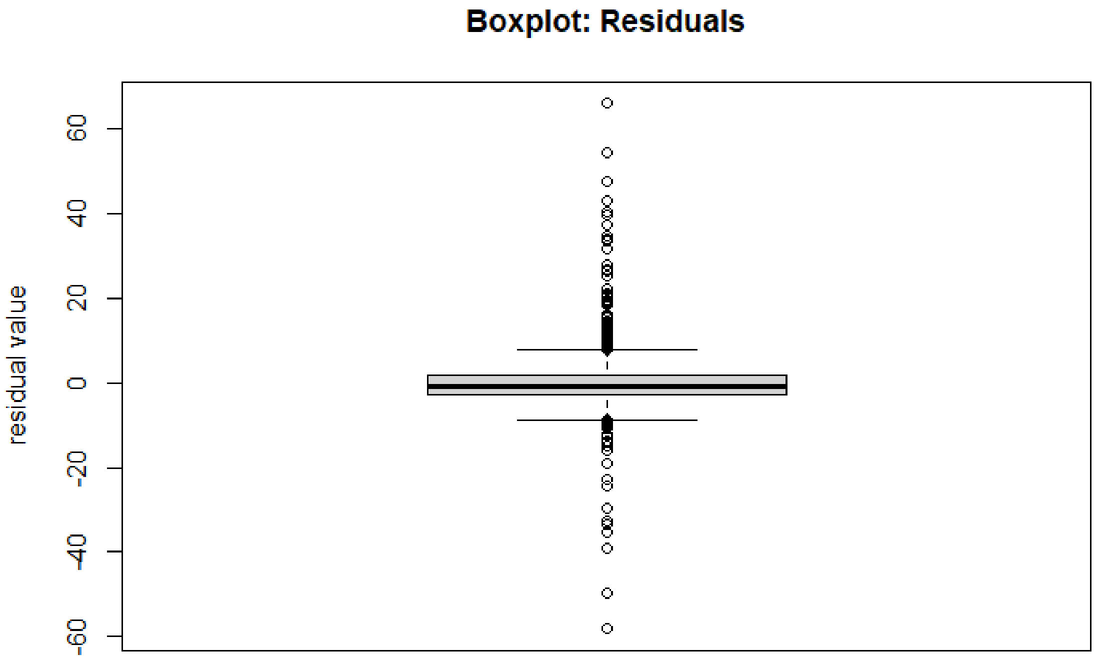
#R squared has a high value at 0.983 which means the model is good; the higher the R squared value, the better the model fits the data. In this instance, it also shows the proportion of change in the dependent variable (N Sexual Offences Unsuccessful) caused by the independent variable ('N Sexual Offences Convictions').

#Adjusted R squared is 98% of the variance in the data which explains the goodness of the model.

#The residual summary statistics provides information about the residual distribution.

#The min and max are close to each other in magnitude which shows symmetry of the distribution.

To investigate this model's residuals further, if they are normally distributed using a boxplot:



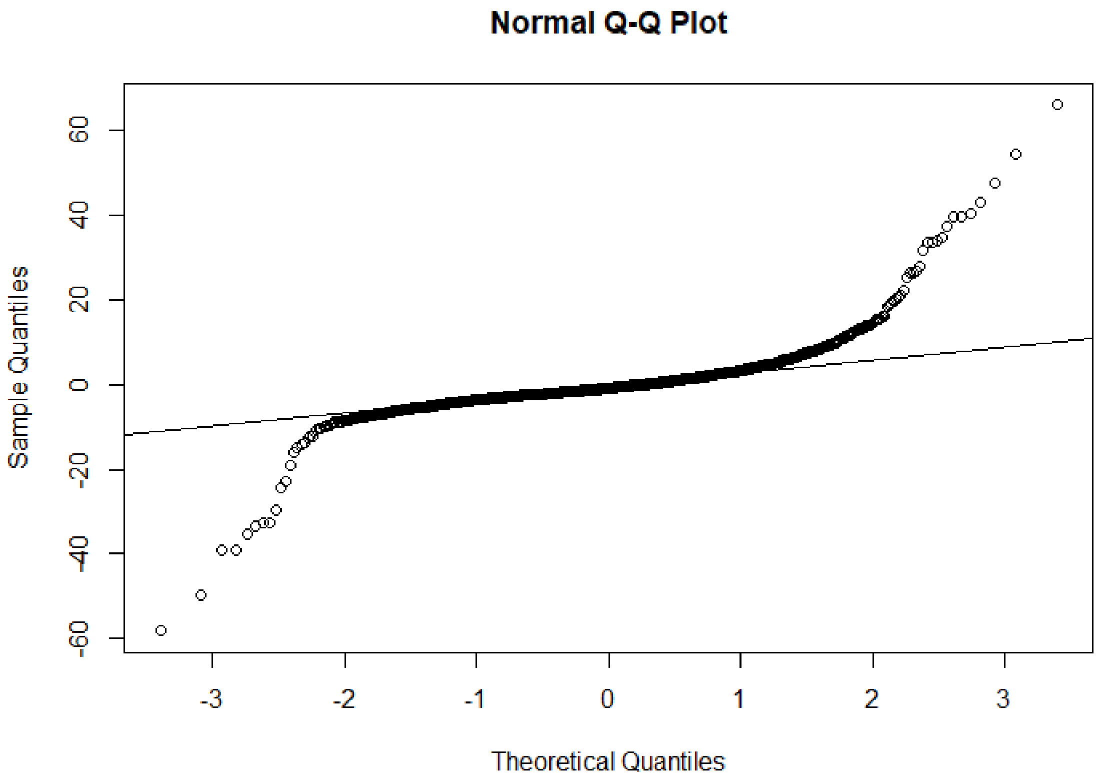
boxplot(linear_regression[['residuals']],main='Boxplot: Residuals',ylab='residual value')

#The median is close to zero and the non-outlier min and max look about the same distance from 0, which is good as it suggests correct model specification

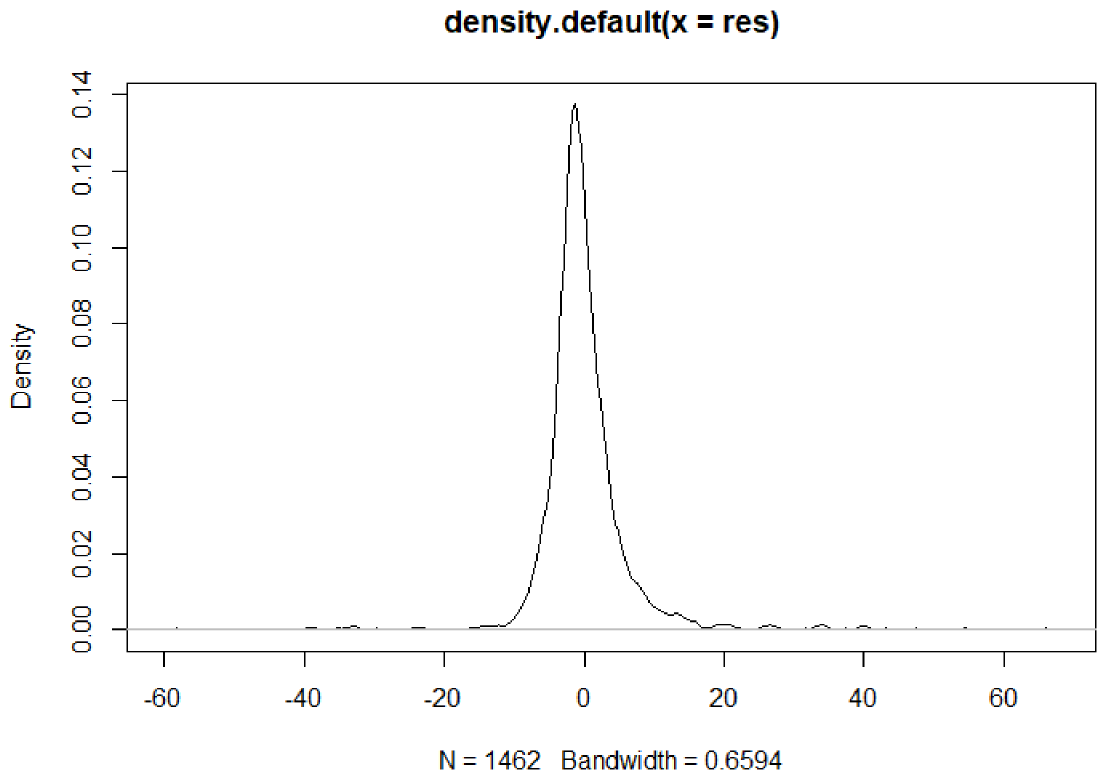
#The confidence interval and assume the estimates of the Gaussian distribution:

	2.5 %	97.5 %
Intercept	-0.4746575	0.2668450
N Sexual Offences Convictions	0.3757984	0.3807696

#Get a list of residuals and then create a Q-Q plot of the residuals



The residuals might not be normally distributed as they stray from the tails but to be sure, using a density plot for further analysis as if it is bell shaped it would show a normal distribution of residuals:

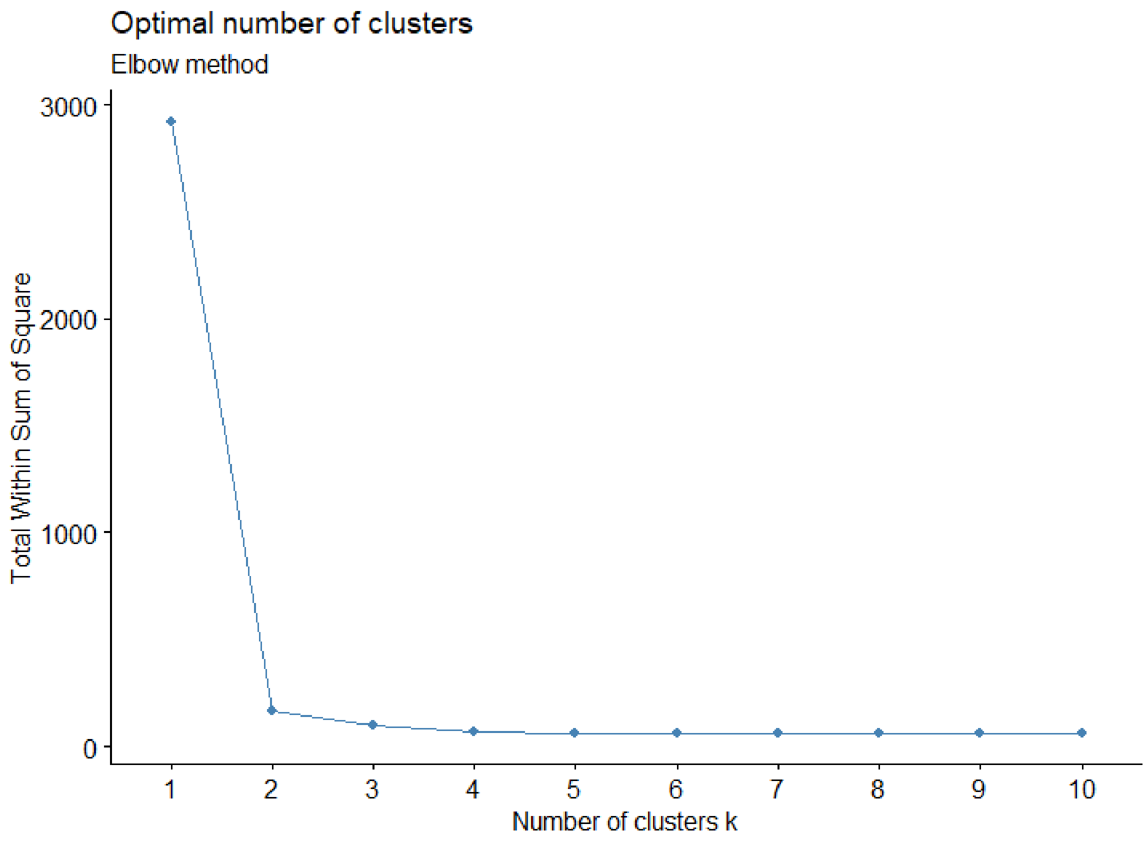


The plot is showing normal distribution as it shows no skewness to the right or left. The highest point is from 0 on the x axis.

Clustering technique

K means Clustering

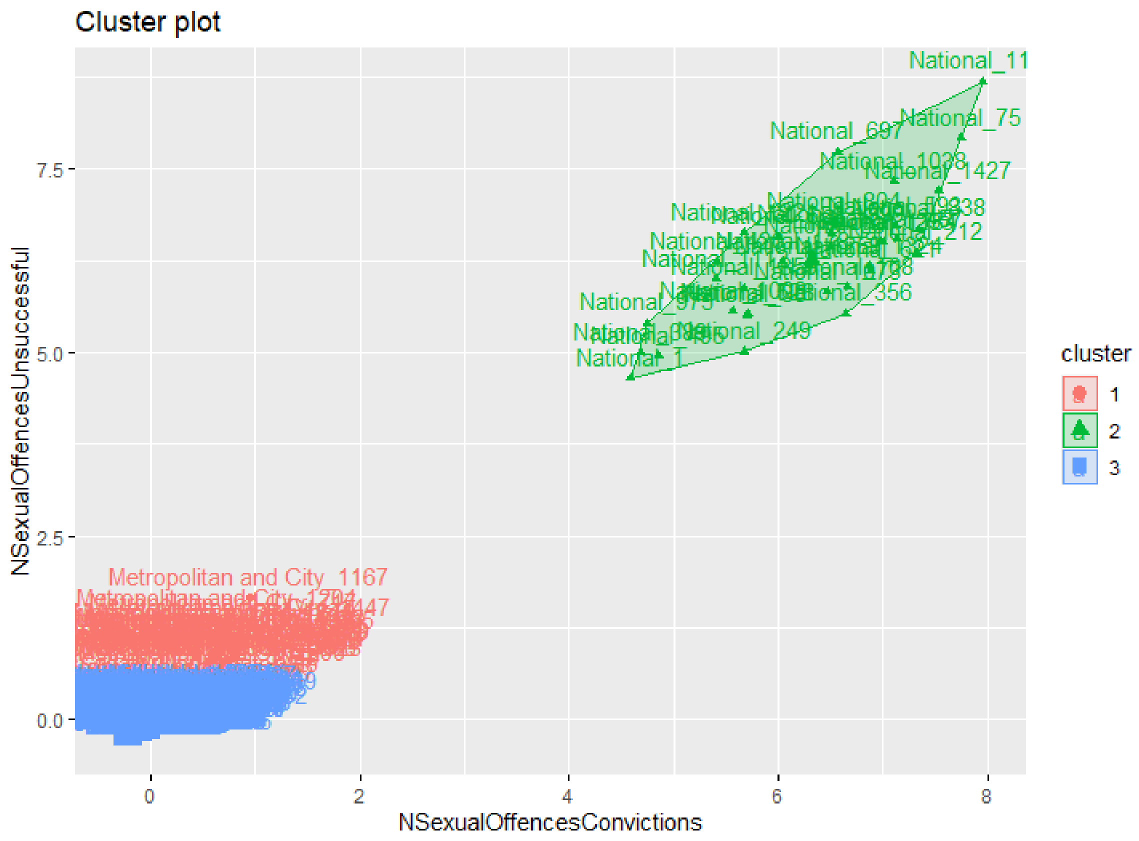
Focusing on the two variables (Columns 6 and 7 of the data set) to prove the hypotheses. First scaling the data set to make the distance matrix unweighted and then finding the Euclidean distance, the number of clusters needed for modelling is calculated using the within sum squares(wss) method:



Focusing on the slope of the magnitude closer to 0, the number of clusters for modelling will be 2 or 3.

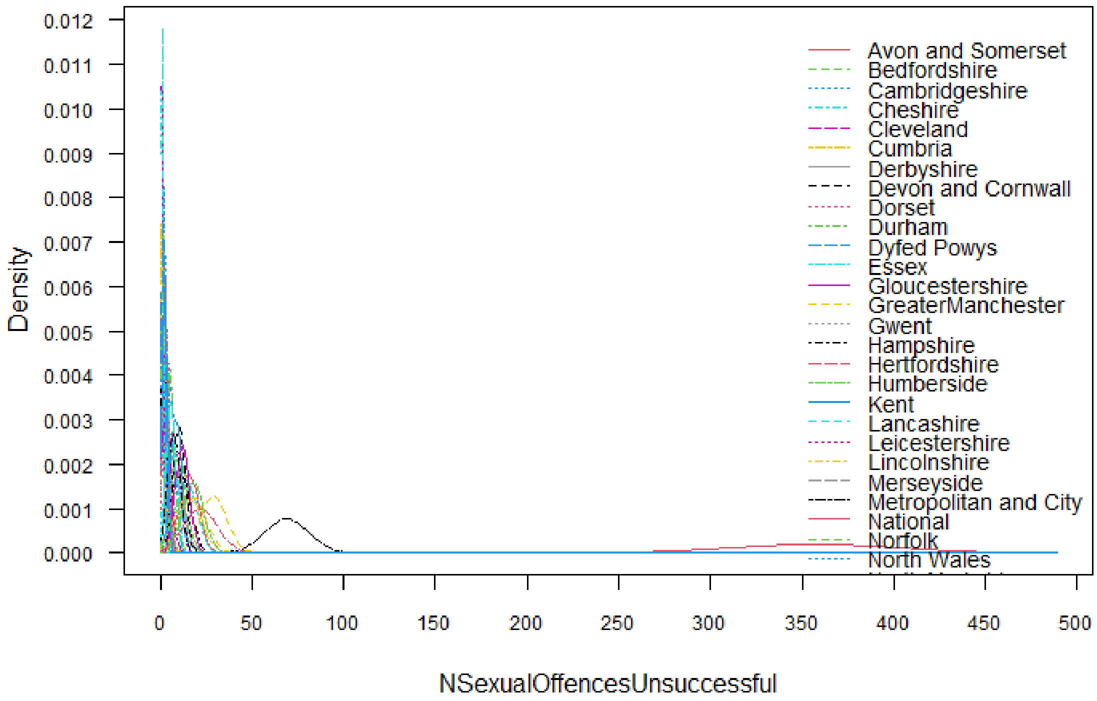
```
Within cluster sum of squares by cluster:
[1]  4.564204 51.870639 40.875214
 (between_SS / total_SS =  96.7 %)
```

#Using 3 clusters for modelling:



#The Counties under National is the biggest cluster in green as it is a total of all the Counties and the Metropolitan and City is in orange while in blue is Midlands. These clusters correspond to the frequency of these offences in these Counties as observed in the data set. Therefore, K means clustering is a good machine model for the data set with the focus on the hypotheses.

<u>Naive Bayes Classification and Normalization in cluster analysis</u>



```
#Prediction model of train data set

#Confusion matrix train data set


p1 <- predict(model, train_set)

(tab1 <- table(p1, train_set$Counties))

p1


1 - sum(diag(tab1)) / sum(tab1)  #Training model accuracy is 80%


#Prediction model of test data set

#Confusion matrix test data set


p2 <- predict(model, test_set)

(tab2 <- table(p2, test_set$Counties))

p2
```

1 - sum(diag(tab2)) / sum(tab2) #Test model accuracy is 97%

Therefore, this machine model is a good representation of the data set focusing on the hypotheses.

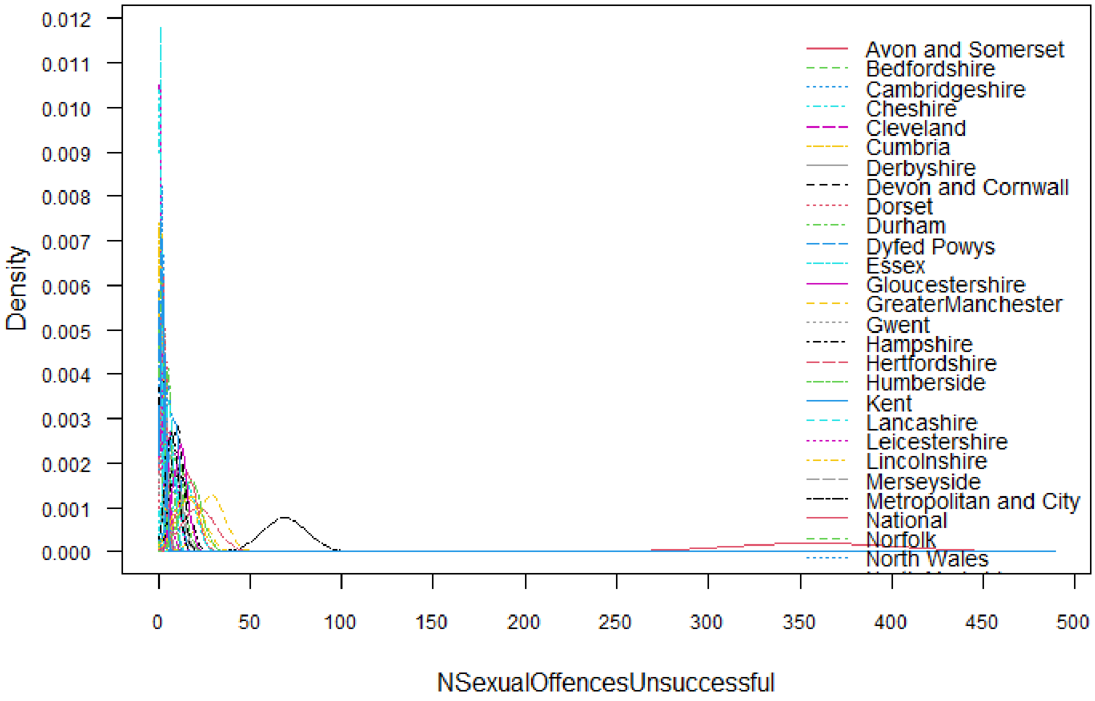
<u>Hierarchical Agglomerative Clustering (All observations are grouped into clusters, from the bottom to top)</u>

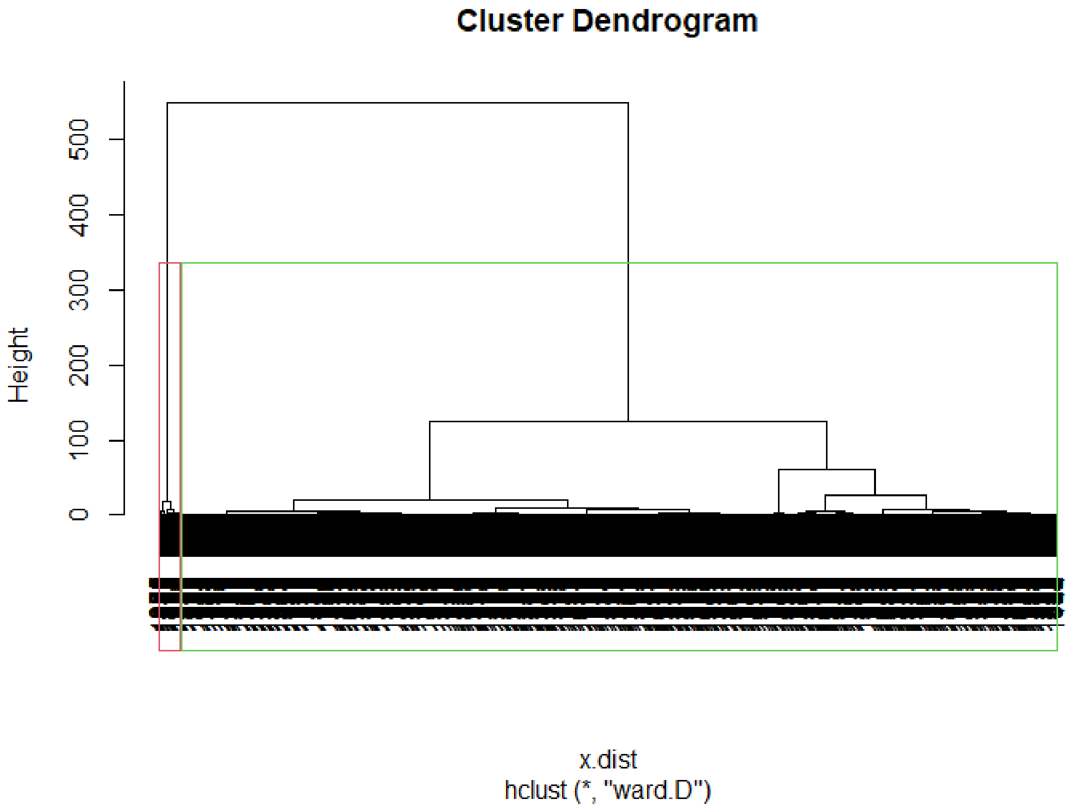
For 'NSexualOffencesConvictions' and 'NSexualOffencesUnsuccessful' in columns 6 and 7 of the train data set as an example to prove the hypotheses

#Hierarchical Agglomerative Algorithim

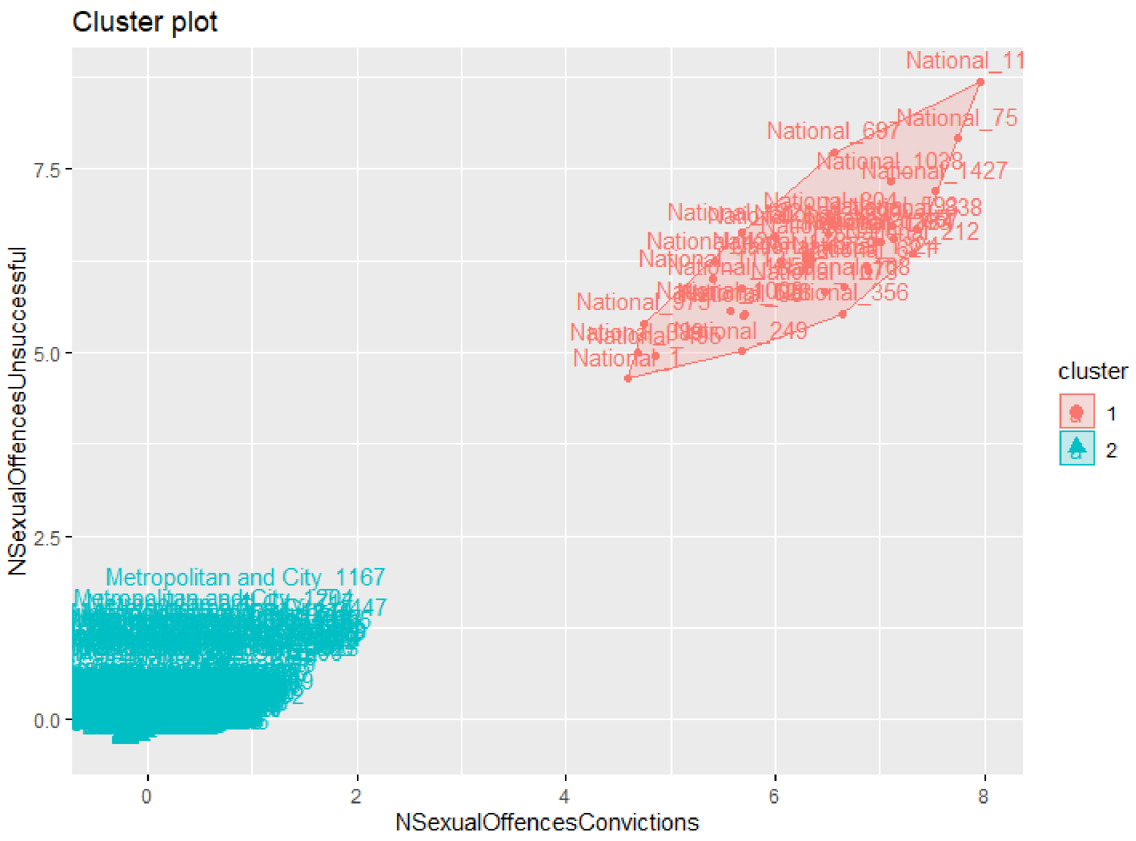
hc.out_x = hclust(x.dist, method = 'ward.D') # Using the ward.D method of clustering

hc.out_x #Number of objects is the same number of observations of the train data set which is 1462, same as in the data set.





Which shows two clusters in coloured rectangles. To show the specific groups the clusters belong to:



The cluster (1) in orange which is where all the Counties fall under, shows that the Number of Sexual Offences convictions increased in all the counties as the number of unsuccessful convictions for the same offence reduced. The cluster (2) in green showing Metropolitan and City shows that the two offences had a lower occurrence recorded.

VALIDITY OF THE HYPOTHESIS

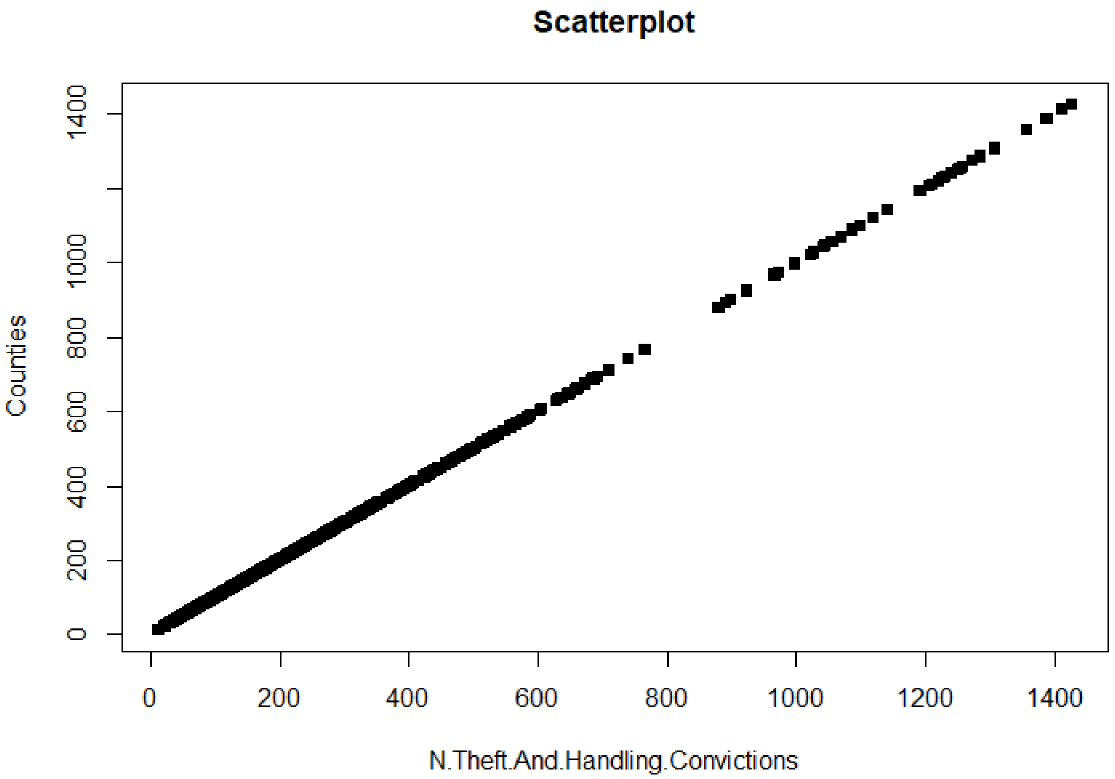
Interpreting the results of the analysis performed by descriptive and predictive tools, the validity of the Alternative Hypothesis holds that as **the frequency of unsuccessful offences (no convictions) in the Counties reduced as the convictions for the same offences increased.**

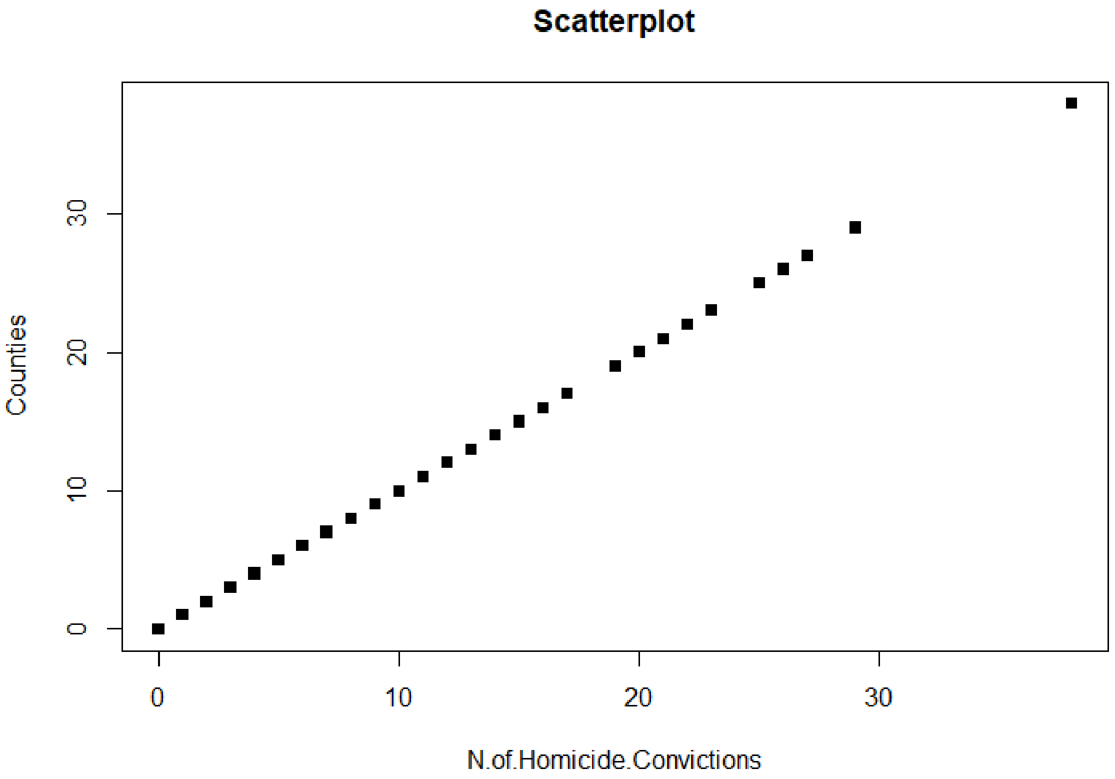
SECOND HYPOTHESIS

The higher the **number of theft and handling convictions, the lesser the number of homicide convictions in all Counties**

DATA ANALYSIS WITH VISUALIZATION TOOLS

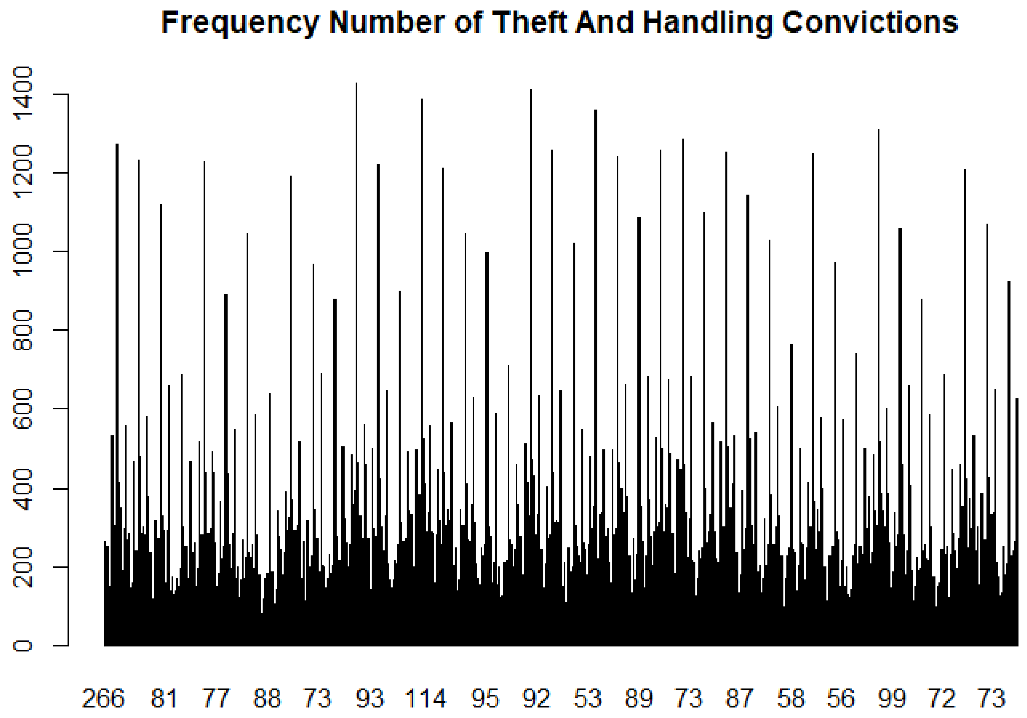
IMPLEMENTATION OF SCATTERPLOT WITH THE TWO SPECIFIC VARIABLES IN THE DATA SET FOR MODELLING

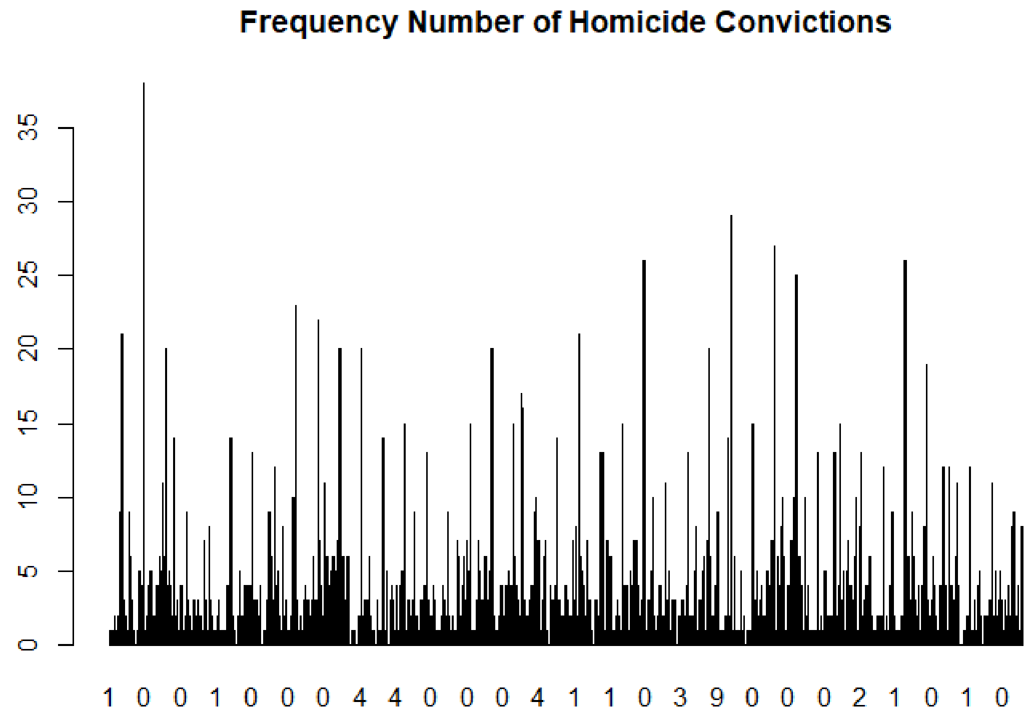

Scatterplot
Counties
0
200
400
600
800
1000
1400
0
200
400
600
800
1000
1200
1400
N.Theft.And.Handling.Convictions


Scatterplot
Counties
0
10
20
30
0
10
20
30
N.of.Homicide.Convictions

From the scatterplots, the Number of Theft and Handling Convictions increased in the Counties, more than 1400 in frequency while the Number of Homicide convictions were less in frequency across the counties.

IMPLEMENTATION OF BAR PLOT WITH THE TWO SPECIFIC VARIABLES IN THE DATA SET FOR MODELLING

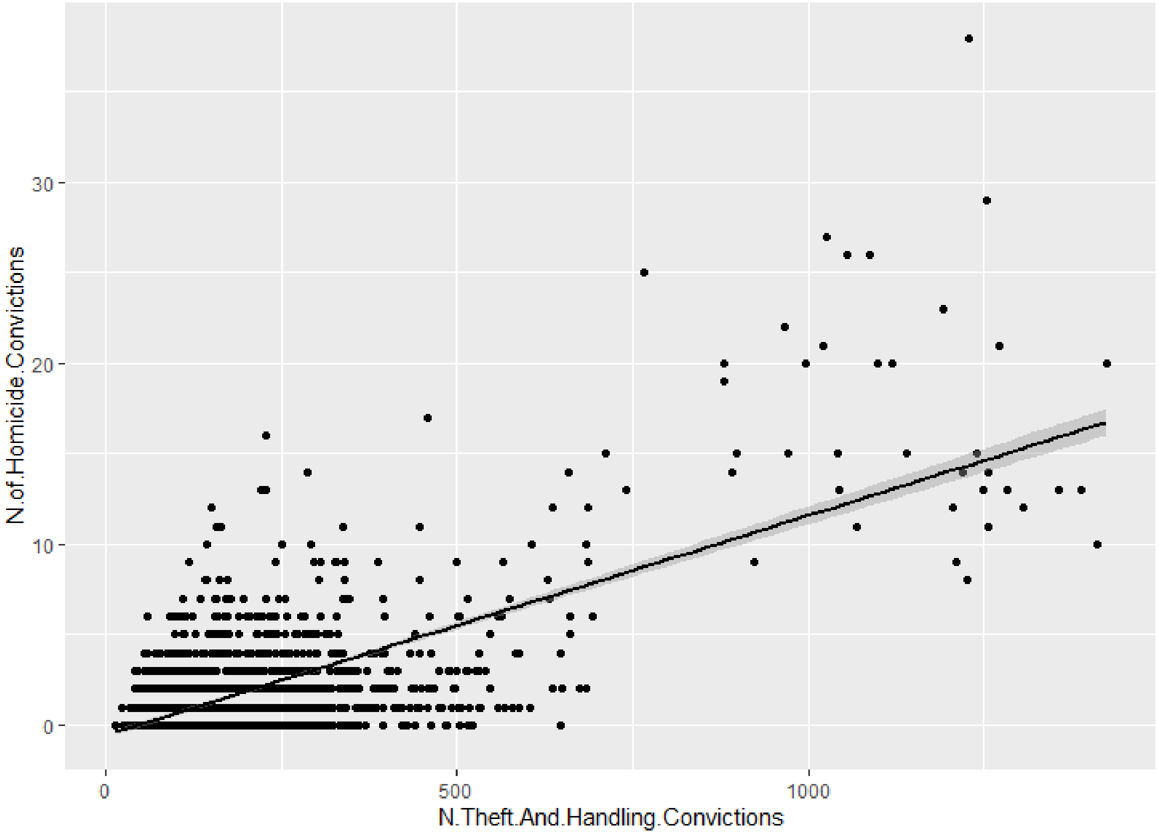




The bar plots show that the Number of Theft and Handling Convictions were high in all the Counties, however across the same counties the Number of Homicide Convictions reduced.

LINEAR REGRESSION, CLUSTERING AND CLASSIFICATION TECHNIQUES TO BUILD ANALYTICAL MODELS OF THE DATA

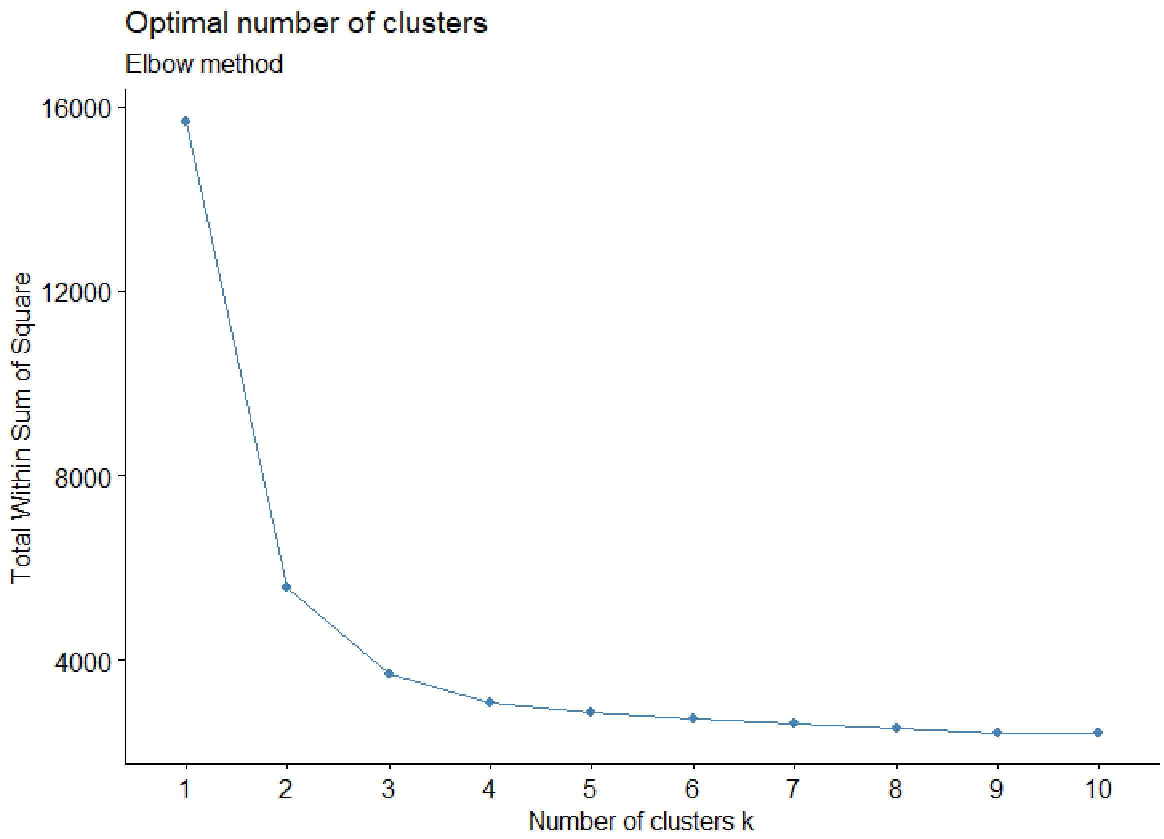
IMPLEMENTATION OF LINEAR REGRESSION WITH THE TWO SPECIFIC VARIABLES IN THE DATA SET FOR MODELLING



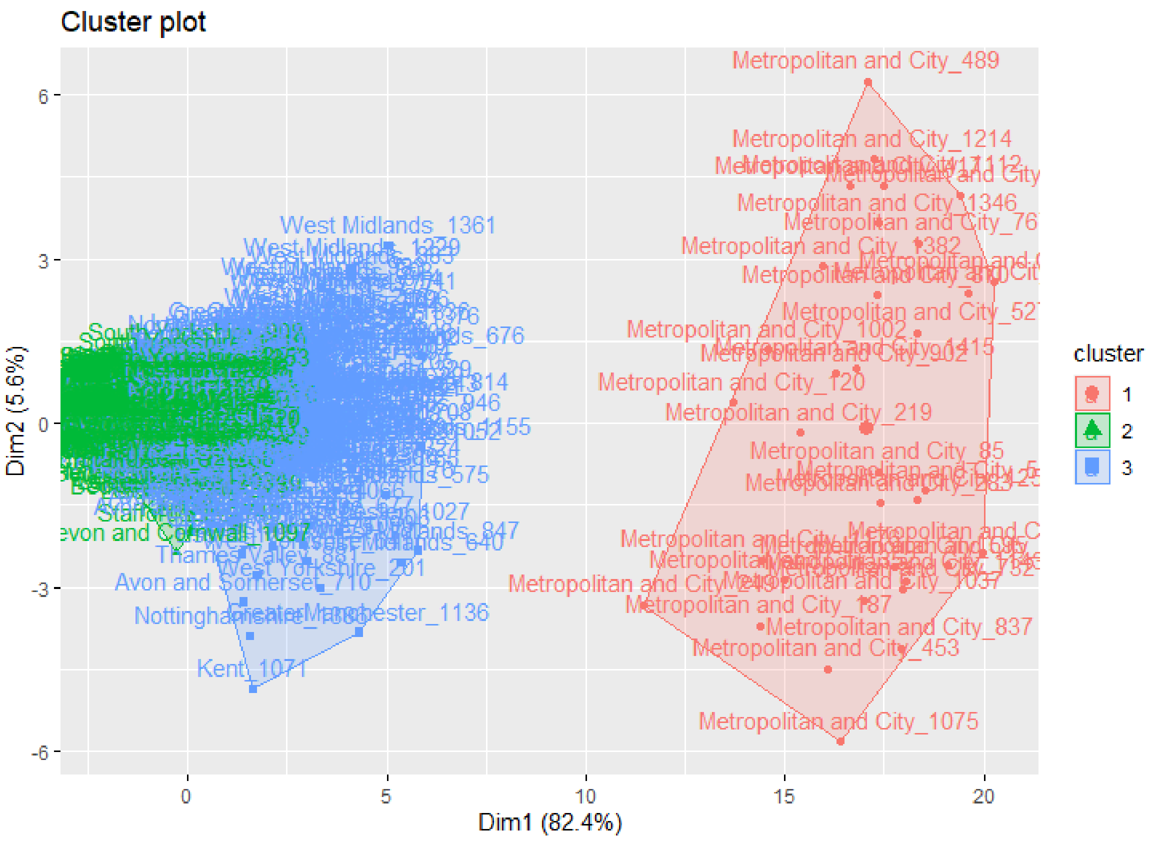
There is a positive linear relationship between the two variables. In this model it can be observed that when the Number of Theft and Handling convictions was close to 1500, the Number of Homicide Convictions was less than 40. The denser points at the start of the regression line shows that there is a stronger relationship between the two variables between numbers 0 and 500 on the x axis (independent variable) and numbers 0 and 8 on the y axis (dependent variable). This is a clear model of how the higher the convictions of theft and handling in the counties, the less the number of homicide convictions (offences which lead to death of individuals).

IMPLEMENTATION OF CLUSTERING WITH THE TWO SPECIFIC VARIABLES IN THE DATA SET FOR MODELLING

K Means Clustering for 'N.Theft.And.Handling.Convictions' and 'N.of.Homicide.Convictions' in columns 6 and 7 of the train data set

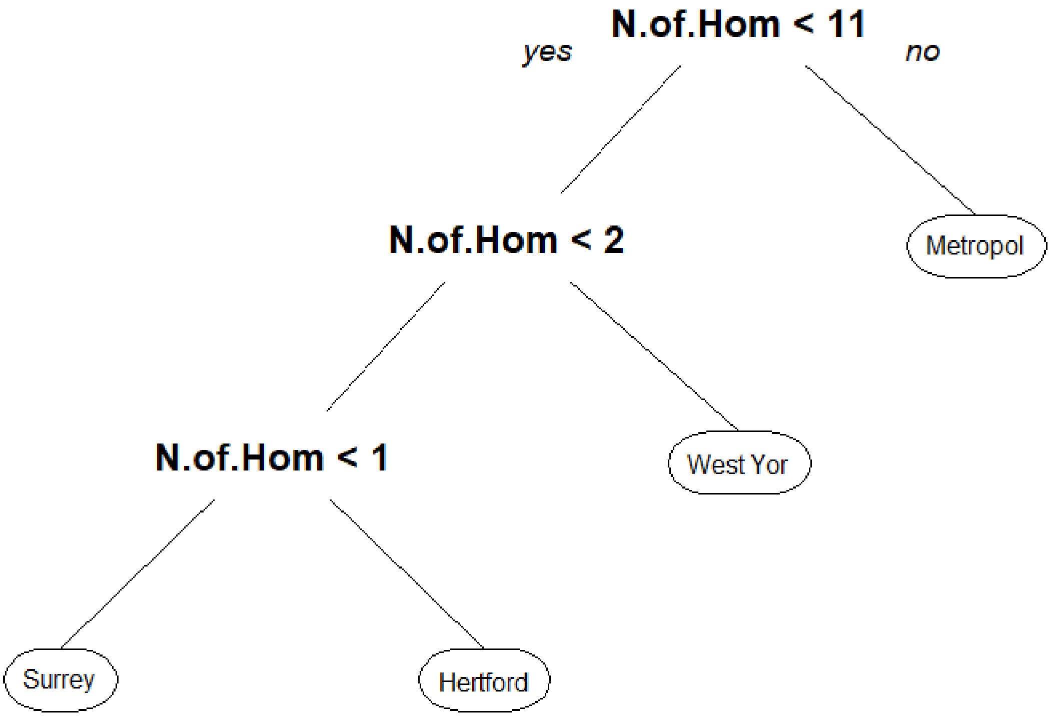


With the aid of the elbow method above, the number of clusters for the Kmeans clustering is 3 or 4.

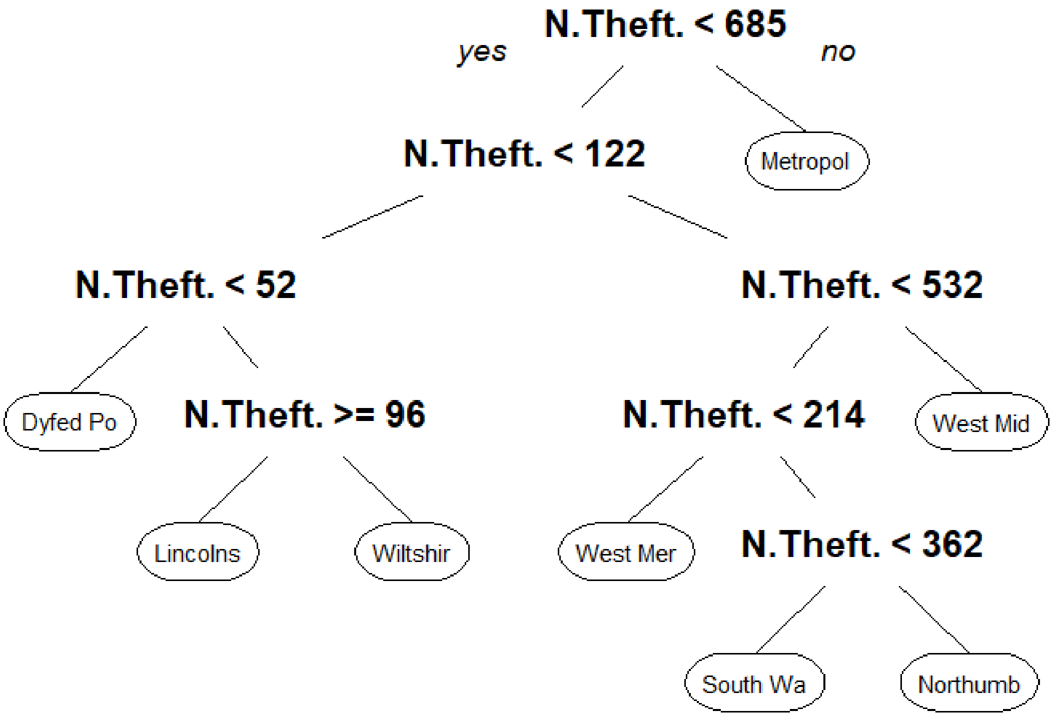


Metropolitan and City has the biggest cluster (1) in orange, followed by West Midlands, Kent, Avon and some other Counties in the blue cluster (2) while the green cluster (3) has Devon and some other counties. This can also be seen as being like the observations from the train set of the data.

Using scatterplot as a visualization tool the focus variables of the data set

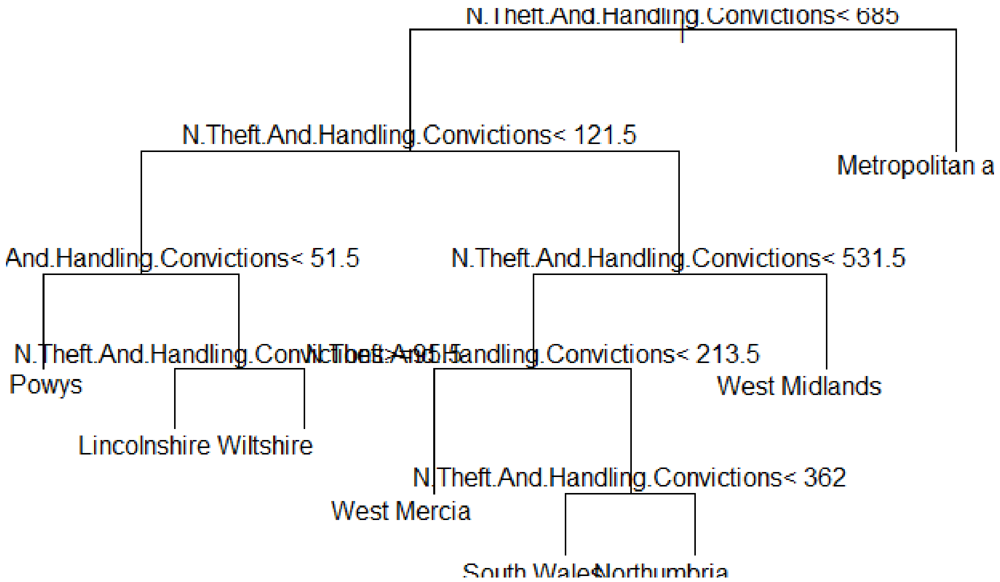


Taking Hertfordshire and Surrey as an example, the number of Homicide convictions in the County is just above 1, same as in the train data set. In Metropolitan and city, this offence is a little above 11. Overall, the number of homicides convictions is few in these counties.



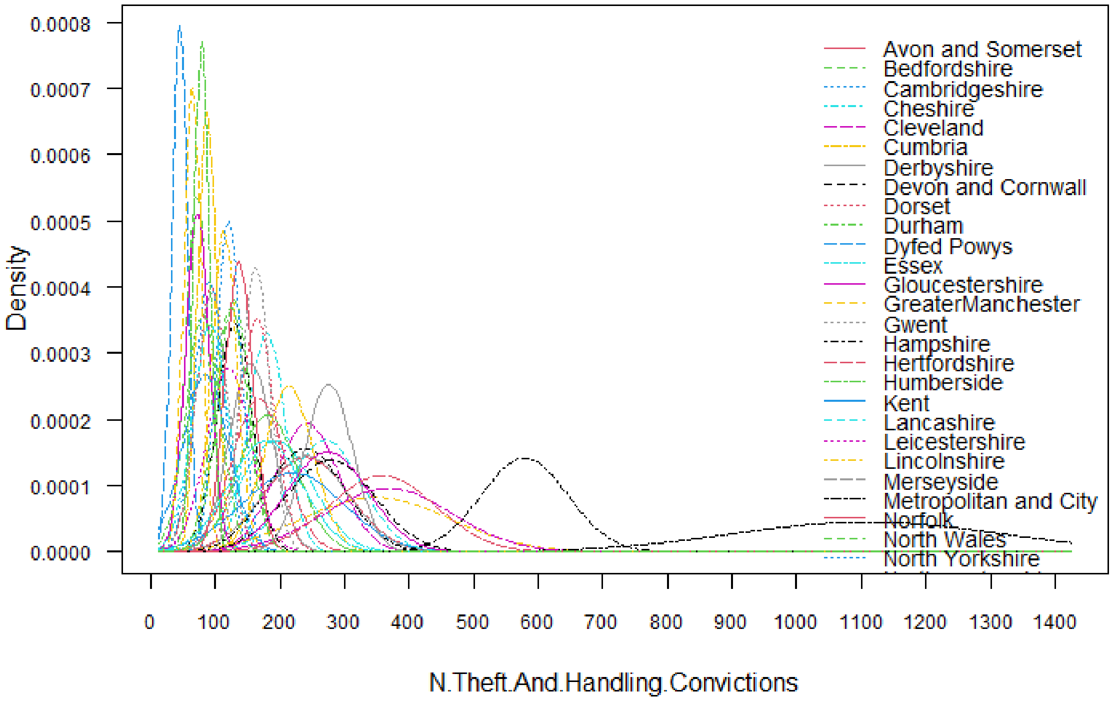
In Metropolitan and city, the Number of Theft and Handling is higher than 685 and this can be observed from the data set as 1388. Overall, this offence has a high frequency in all the Counties in the data set.

Studying the two decision trees, my observation is that the more convictions of Theft and Handling, the less the convictions of Homicide in the Counties.


N.Theft.And.Handling.Convictions< 685
N.Theft.And.Handling.Convictions< 121.5
Metropolitan a
And.Handling.Convictions< 51.5
N.Theft.And.Handling.Convictions< 531.5
N.Theft.And.Handling.Conv
N.Theft.And.Handling.Convictions< 213.5
Powys
West Midlands
Lincolnshire
Wiltshire
N.Theft.And.Handling.Convictions< 362
West Mercia
South Wales
Northumbria

A prediction model of the decision tree with the variable Number of Theft and Handling Convictions.

IMPLEMENTATION OF NAÏVE BAYES CLASSIFICATION WITH THE TWO SPECIFIC VARIABLES IN THE DATA SET FOR MODELLING



Observing the County Cleveland, when the Number of Theft and Handling Convictions was 254, the offence was closer to 0 for the Number of Homicide Convictions as seen above. In Dyfed Powys, when the Number of Theft and Handling Convictions was around 56, there were no convictions for Homicide in the County.

```
#Prediction model of train data set

#Confusion matrix train data set


p1 <- predict(model, train_set)

(tab1 <- table(p1, train_set$Counties))

p1


1 - sum(diag(tab1)) / sum(tab1)  #Training model accuracy is 81% which is a good accuracy for this model.




#Prediction model of test data set

#Confusion matrix test data set
```

```
p2 <- predict(model, test_set)

(tab2 <- table(p2, test_set$Counties))

p2
```

```
1 - sum(diag(tab2)) / sum(tab2) #Test model accuracy is 83%.
```

VALIDITY OF THE SECOND HYPOTHESIS

Interpreting the results of the analysis performed by descriptive and predictive tools, the validity of the Hypothesis holds that as **the higher the number of Theft and Handling offences convictions in the Counties , the lesser the convictions for homicide offences.**

CRITICAL REVIEW OF THE VISUALIZATION TOOLS AND THEIR EFFECTIVENESS AND ALTERNATE SOLUTIONS

Visualizing the data set with a histogram is the best method to me as it clearly showed the variables of the data set and the frequency of occurrence of the offences in the Counties. Pie charts are best to visualize small data sets. As seen in the two pie charts above, the data set visualization was not clear or easy to understand the variables in relation to the occurrence of offences in the Counties. The scatter plot is nearly as good as the histogram as regards visualization as it showed the distribution of variables in the data set and included the statistical measurement, variance which was seen as the mean values. The box plot clearly showed the median of the values of the data set, and this was also a good visualization. The scatterplot was a good representation of the relationship between the dependent and independent variable which is the basis of the hypotheses for this data analysis. My conclusion is that this visualization(scatterplot) is on the same level of effectiveness as a visualization tool, like the histogram. The bar plot was also an effective visualization tool as it was clear how the two variables for the second hypothesis were of high and low frequencies in all the Counties.

An alternate and more effective solution to using these tools could start by using a small sample size of the data set to apply pie chart as a tool for visualization. County per highest offence convictions or County per lowest offence unsuccessful would be easier, faster to plot and visualize. This is a proffered solution as pie charts are very clear to understand. However, choosing a random sample size might not be the best representation of the whole data set since it covers four years of recorded offences and 43 Counties.

WEAKNESSES AND STRENGTHS OF THE MACHINE MODELS USED IN DATA ANALYSIS AND MODEL PREDICTIONS

#1. Hierarchical Agglomerative clustering is not the best for very large data set (with numerous observations) as it has to do with grouping clusters from bottom top. This is a weakness of this model. K Means is a good option for large data sets, which is a strength of k means clustering. I tend to prefer this as a machine model technique as the clusters were easy to identify, differentiate and study, in this case, showing the clusters of the Counties in relation to the specific offences and it is a fast algorithm. Naive Bayes Classification for cluster analysis was also good for modelling, though it is usually recommended for categorical variables, it can work with numerical variables as seen in this analysis. However, this model quickly generated output unlike the decision trees. Naïve Bayes Classification showed that the variables were not independent of each other; it also had a fast output and good accuracy of 80% for the train data set and 97% for the test data set for this first hypothesis and for the second hypothesis it was 81% and 83% accuracies for the train set and test set respectively. Overall, I prefer K means clustering for machine modelling because of its speed, high accuracy and because it does not require the prediction of a response variable.

#2. Decision trees can scale to big data sets but might not best for such as it can have many nodes and lead to overfitting. They also have the disadvantage of having moderate to high variance. However, decision trees are easy to interpret; easy to visualize. They showed the distribution of the specified offences per County which was clear to observe in this data analysis (which is good as decision trees can handle both categorical and numerical data). However, I did not much prefer it for modelling the data as the branches could be confusing if not shortened. This means important data needed for the model might be cut off in the bid to achieve this aim.

#3. Linear regression is easy to understand, however cannot be used to address non-linear relationship of variables. However, I prefer the linear regression model as it was accurate to a very high level and good as a model to illustrate the relationship between the specific variables which were the dependent and independent variables in the data set. R squared of the linear regression has a high value at 0.983 which means this machine model is good. It showed the proportion of change in the specified dependent variable (‘NSexualOffencesUnsuccessful’) caused by the independent variable(‘NSexualOffencesConvictions’). Adjusted R squared is 98% of the variance in the data and explains the goodness of the model.

CONCLUSION

It would be worthwhile to see how using machine models like K means clustering and Naïve Bayes for modelling data sets of convicted offences in not only Counties but sampled populations of Countries; since they produce accurate models and are fast in their outputs. This means they are time-saving tools for machine modelling of data sets. Also comparing the model’s outputs with that of unsuccessful offences would be interesting. Seeing as in this data visualization and machine modelling, the recorded frequency of convicted offences had a noticeable impact on the occurrence of the unsuccessful convictions of the same offences; meaning there was more justice for victims as the offenders were convicted. As a means of prescriptive analysis, it could help security operatives and the justice system to have a better idea of where they need to focus their strategies to reduce offences or crimes in communities or Countries for the future. Bigger picture is that all this can save lives and make the society safer for everyone.

The second hypothesis points to the fact that less people died, that is less homicide convictions when the Number of theft and Handling Convictions in the Counties went higher. It can be inferred from this that there is a strong relationship between the two offences and when more people were caught and convicted of theft and handling, less people died or there were less homicides. This could be a good focus for the people in charge of security in the Counties to focus on to reduce the number of homicides and make people safer.

REFERENCES AND CITATIONS

1. https://www.researchgate.net/figure/Strengths-and-weaknesses-of-common-machine-learning-models_tbl1_341573546

2. https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

3. http://www.physics.csbsju.edu/stats/box2.html

4. What is Prescriptive Analytics? How does it work? Examples & Benefits (valamis.com)

5. A Comprehensive Guide to Data Visualisation in R for Beginners | by Parul Pandey | Towards Data Science

6. A Step-by-Step Guide to the Data Analysis Process [2022] (careerfoundry.com)

7. R for Data Science by Hadley Wickham and Garret Grolemund (Chapter 5)

8. https://www.dataquest.io/blog/load-clean-data-r-tidyverse/

9. Classification in R Programming: The all in one tutorial to master the concept! - DataFlair (data-flair.training)

10. Classification in R Programming: The all in one tutorial to master the concept! - DataFlair (data-flair.training)

11. In-text-citation: Introduction to Multivariate Regression Analysis Alexopoulos EC Department of Public Health, Medical School, University of Patras, Rio Patras, Greece. Corresponding author: Evangelos Alexopoulos, Department of Public Health, Medical School, University of Patras, 26500 Rio Patras, Greece (Page 26)