**Assignment Report**

CT7202 Data Analysis and Visualisation Principles

Chanya Subasingha Arachchige-S4107143

Abstract

Court case results of England and Wales from June 2015 to March 2018 are analyzed to identify trends and patterns in the data. Several trends and patterns are outlined in this report for theft and handling, robbery, burglary, drug offences and motoring offences.

Table of Contents

[1 Introduction 2](#_Toc104241983)

[2 Hypothesis 2](#_Toc104241984)

[2.1 The number of total court cases for theft and handling, burglary and robbery in England and Wales decreased from July 2015 to March 2018. 2](#_Toc104241985)

[2.2 Total drug offences court cases depend on the month and the court area from July 2015 to March 2018 in England and Wales. 2](#_Toc104241986)

[2.3 The percentage of successful convictions for motoring offences is dependent on the court area from July 2015 to March 2018 in England and Wales. 2](#_Toc104241987)

[2.4 The percentage of successful convictions for motoring offences is dependent on the total number of court cases from July 2015 to March 2018 in England and Wales. 2](#_Toc104241988)

[2.5 The percentage of successful convictions for motoring offences is dependent on the month from July 2015 to March 2018 in England and Wales. 2](#_Toc104241989)

[3 Data preparation 3](#_Toc104241990)

[3.1 Data Cleaning 3](#_Toc104241991)

[3.2 Feature engineering 4](#_Toc104241992)

[3.3 Missing value imputation 5](#_Toc104241993)

[4 Descriptive analysis 5](#_Toc104241994)

[4.1 Analysis of theft and handling, burglary, and robbery national court cases 6](#_Toc104241995)

[4.2 Analysis of drug offence court cases 10](#_Toc104241996)

[4.3 Analysis of motoring cases 14](#_Toc104241997)

[5 Linear regression 16](#_Toc104241998)

[6 Clustering 21](#_Toc104241999)

[6.1 Clustering drug offence cases finalised with the K-Means algorithm. 21](#_Toc104242000)

[6.2 Fuzzy clustering 26](#_Toc104242001)

[7 Classification 27](#_Toc104242002)

[7.1 Decision trees classifier 27](#_Toc104242003)

[8 Conclusion 30](#_Toc104242004)

[9 References 31](#_Toc104242005)

[10 Appendix 33](#_Toc104242006)

# Introduction

The Crown Prosecution Service (CPS) publishes a monthly report on the outcome of CPS proceedings in magistrates' courts and the crown court by principal offence category. The data set contains twenty-seven monthly court outcome reports for the thirty-three-month period from July 2015 to March 2018. This is the period after the new data assurance regime was introduced [1]. Each report contains the number and percentage of convictions and unsuccessful convictions by defendant [2].

One court case can represent several observations in the data set as there can be several defendants in one court case [1]. However, in this report, a court case per defendant is called a court case for simplicity. The principal offence is decided at the time of case finalisation which is the most serious charge against the person. They are categorised as homicide, offences against the person, sexual offences, burglary, robbery, theft and handling, fraud and forgery, criminal damage, drugs offences, public order, motoring, or other offences excluding motoring. The data set also contains a column for administrative finalisations. These are the cases which could not proceed due to an administrative issue such as an unexecuted warrant for the arrest of the defendant or summons that have not been served by the police because they were unable to trace the defendant, or the defendant has died or is unfit to plead. These cases are summarised without categorising into principal offences [1].

# Hypothesis

There are five hypotheses defined based on the data set. Descriptive analysis and machine learning modelling have been used to analyse data in order to accept or reject the null hypothesis.

## The number of total court cases for theft and handling, burglary and robbery in England and Wales decreased from July 2015 to March 2018.

## Total drug offences court cases depend on the month and the court area from July 2015 to March 2018 in England and Wales.

## The percentage of successful convictions for motoring offences is dependent on the court area from July 2015 to March 2018 in England and Wales.

## The percentage of successful convictions for motoring offences is dependent on the total number of court cases from July 2015 to March 2018 in England and Wales.

## The percentage of successful convictions for motoring offences is dependent on the month from July 2015 to March 2018 in England and Wales.

# Data preparation

All the data files have the same columns. Each data file has 51 variables holding the number of court outcomes and percentage of court outcomes, and 43 observations representing each crown court and magistrate court and national values.

First, I noticed that the column names had long names with spaces, making it hard to work within R functions as they are not compatible with R variable names. It is possible to remove spaces with the below code.

colnames(dataframe)<-str\_replace\_all(colnames(dataframe), c(" " = "" ))

However, the data column names are too long even without the spaces, and the code lines tend to get longer with the long column names. Additionally, it causes overlaps in the graph axis. Therefore, all the column names are renamed to their abbreviations. This also makes the code more user-friendly and readable. I used capital letters for the column names to make it clear that they are abbreviations. Three court areas with long names changed to shorter strings to be able to show in the graphs properly.

Then a new data frame was created, merging all the data sets. As a preparation to combine datasets, a new variable *Date* was introduced to all the data sets using the *lubridate* library [3]. Even though the data set only has a month and a year, the date was added as the first day of the respective month for simplicity. Merging data sets reduce code complexity and processing time as one data set can be processed easily in R, with better performance than looping through all the data sets each time.

## Data Cleaning

The dataset was examined for missing values, but there are no missing values in the data file for a particular month. Some observations are recorded with a dash in the percentage of homicide offence convictions and rate of homicide successful percentages columns; these are not missing values; instead resulted from zero number of court cases for the respective month. However, some months were missing when considering the whole period as a time series, as shown in Figure 3.1.

Then data types of the dataset were checked. Columns with the number of court cases are numeric, and the columns with percentages of court outcomes have character type values. These percentages columns were converted to numeric because numerical values can carry more information, and more insight can be drawn from them. I wrote a code to remove the percentage sign first and then convert values to numeric.

Chart, bar chart, histogram

Description automatically generated

Figure 3.1. Number of theft and handling convictions from July 2015 to March 2018

Then after a round of data visualisation with box plots and bar charts, noticed that the data set contains national values of the principal offences court outcomes that have significantly larger values than other values. As shown in Figure 4.1, national values that are outliers and much more significant than the rest hinder data visualisation for identifying patterns and prominent features. Hence, national data are separated from the data set, copied to another data frame, and analysed separately.

Column *percentage of L motoring offences unsuccessful* has been removed from the data set as this column has values only 100 and NA. The column has no data variation; therefore, it adds no value for pattern recognition or prediction analysis.

## Feature engineering

The total number of finalised court cases for each offence category was calculated using the summation of the number of convictions and unsuccessful cases. Because both convictions and unsuccessful convictions followed a similar trend. The new variables *total\_theft\_handling, total\_burglary, total\_robbery, total\_cases* (for drug offence) and *total\_motoring* were created for theft and handling, burglary, robbery, drug offence and motoring offence, respectively.

The date column was used to extract the month [3] and create another column; then, the month column was used to introduce the season column using a conditional statement. Month and season columns are used in the analysis to find seasonal patterns. The variable *motor\_cat* was introduced to store levels of success rates for motoring offences. The percentage of successful convictions is divided into three categories, average, high and very high, depending on the value. The court names are string values they cannot use in classification. Therefore, they are converted to numeric values and stored in the new column *court\_num*.

## Missing value imputation

A column *no\_months* (number of months) was created using the date column of the data set. Then using a vector containing the complete value set for the number of months, identified missing values, and imputed the missing data point of each data column.

Linear interpolation [4] [5] [6] is used to fill missing values for columns *'N\_BC', 'N\_BU', 'N\_RC', ' N\_RU', 'N\_THC'* and *'N\_THU'* as the column values has a downward trend. Using simple imputations such as mean, or median will produce values disrupting the trend. Hence, the linear imputation, which uses values closer to missing data points to impute missing data, was a more appropriate choice to preserve the trend. Missing values of the *Date* column (timestamp) were also filled with linear interpolation method *approx* [7],which uses the midpoint of neighbouring values to impute missing values.

# Descriptive analysis

Analysis was carried out by plotting bar graphs against court names and timestamps separately to identify trends or seasonality in the data. Once such a pattern is identified, those columns are analysed further. Histograms and box plots are used to visualise the data distribution and spot irregularities and outliers [8]. Histograms are better to understand the frequency of the values, while boxplot enable clearly identify outliers and the distribution of the data.

Initial descriptive analysis using a box plot shows that there are outliers present in the dataset; the box plot is used here as it enables clearly identify outliers better than any other visualisation method such as bar plots or scatter plots. National data is recognised as an extreme outlier, as shown in Figure 4.1. It is, therefore, separated from the dataset as described in section 3.1 (Data cleaning). Because having an outlier far apart from all the other values hinders the analysis by suppressing the features in the rest of the data when visualising data. National data being an outlier is an obvious result as the national column contains the summation of all the other observations in the same column for the respective month.

The Metropolitan and city court area is also an outlier in all the categories as shown in Figure 4.7. These data are far apart from the other courts’ number of finalised court cases. The Metropolitan and city area has the highest court cases among all courts in England and Wales. This is justifiable as Metropolitan and City court areas include the most populous city London in the UK [9]. There are other outliers identified, but they vary with the type of offence.

Chart, scatter chart

Description automatically generated

Figure 4.1. Number of robbery convictions in July 2015

## Analysis of theft and handling, burglary, and robbery national court cases

The national total number of finalised theft and handling court cases has a minimum of 5559 maximum of 10114. The national total number of finalised burglary and robbery court cases has a maximum of 1775 and 608, respectively. The number of national theft and handling cases is much higher when compared to robbery cases and burglary cases, as shown in table 4.1.

The distribution of theft and handling cases is a multimodal and does not show skewness, according to Figure 4.2. The histograms also show that the distribution of the total national robbery court cases is skewed to left, indicating that there are only a smaller number of months with a higher number than the median total robbery cases finalised in courts. Burglary cases follow a distribution close to a normal distribution. Both distributions for burglary and robbery are unimodal while the histogram of national theft and handling cases shows the multimodal distribution.

Table 4.1. Summary of the basic statistics for theft and handling, robbery and burglary total national court cases from July 2015 to March 2018

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | Range |
| Theft and handling court cases | 5559 | 6981 | 7793 | 7911 | 8783 | 10114 | 4555 |
| Burglary court cases | 982 | 1317 | 1395 | 1414 | 1516 | 1775 | 793 |
| Robbery court cases | 373 | 443 | 464 | 464 | 474 | 608 | 235 |

Chart

Description automatically generated

Figure 4.2. Number of robbery convictions in June 2015

It is clear from Figure 4.2 that the national data for theft and handling outcomes are following a clear downward trend. There is a spike in the first three months of 2017 for the national number of theft and handling convictions, other than that only slight deviations from the downward trend is observed. The number of convictions seems to be higher in the first four-month each year as seen in Figure 3.1. Even though, the number of unsuccessful convictions is much smaller than convictions it also follows a downward trend. Bar charts are better to visualise data in order to identify any trends. Separate graphs for convictions and unsuccessful convictions are included in the Appendix Figure 1 and Figure 2.

Chart, bar chart, histogram

Description automatically generated

Figure 4.3. The national number of theft and handling court cases from July 2015 to March 2018

Figure 4.4 shows that the national data for burglary outcomes are following a downward trend with some variations. Unsuccessful convictions are also following a downward trend over the period. The minimum number of both convictions and unsuccessful convictions is reported in December 2017 which is towards the end of the period under this analysis. The maximum number of burglary convictions are reported in October 2015 with over 1500 cases.

Chart, bar chart, histogram

Description automatically generated

Figure 4.4. The national number of burglary court cases from July 2015 to March 2018

Total national finalised robbery cases have the highest values, around 600, at the beginning of the period in July 2015 and September 2015. Then the values slowly decreased with fluctuations to 375 in December 2016. Then shows a slight increase until March 2017, then dropped again and reached its lowest value of 373 in December 2017. Overall, the values fall over the period with fluctuations. Robbery convictions follow the same pattern as total court outcomes. However, the unsuccessful cases mostly stay between 75 and 100. Relatively higher values are observed at the beginning of the period in July and September 2015, and January 2017, exceeding 125 cases.

Chart, bar chart

Description automatically generated

Figure 4.5. The national number of robbery court cases from July 2015 to March 2018

When compared with the number-of-month variable (representing the timestamp), all three variables are negatively correlated. The total theft and handling and total burglary court cases have a strong negative correlation with the number-of-month variable, while total robbery court cases do not show such a strong relationship with time.

Chart, scatter chart

Description automatically generated

Figure 4.6. Scatter plot matrix for the national number of theft and handling court cases from July 2015 to March 2018

Correlation calculation of the national number of convictions and unsuccessful cases of burglary, robbery and theft and handling with the date was done using Pearson correlation to measure linear relationship [10]. National burglary case outcomes and theft and handling case outcomes have a strong negative correlation with the date column, as shown in Table 4.2, which suggests that both those offence categories have a good linear relationship with time. National robbery convictions also have a negative correlation. However, that is not strong as the other two types of case outcomes.

There are outliers present in robbery convictions, national robbery unsuccessful and burglary convictions. Since Pearson correlation is highly sensitive to outliers [10], Spearman Correlation analysis is also carried out to identify monotonic relationship [10]. Both correlation analysis shows closer values. Spearman correlation results show that there is a decline in the number of national court cases for burglary, theft and handling and robbery court cases over time as both conviction and unsuccessful convictions are decreasing. Especially, burglary and theft and handling court cases show a clear decline over time with correlation coefficients closer to -1.

Table 4.2. Correlation of the number of case outcomes and date

|  |  |  |
| --- | --- | --- |
| Number of national case outcome | Pearson correlation with date | Spearman correlation with date |
| Theft And Handling Convictions | -0.883 | -0.861 |
| Theft And Handling Unsuccessful | -0.913 | -0.919 |
| Burglary Convictions | -0.761 | -0.761 |
| Burglary Unsuccessful | -0.815 | -0.851 |
| Robbery Convictions | -0.494 | -0.478 |
| Robbery Unsuccessful | -0.582 | -0.479 |

## Analysis of drug offence court cases

Total drug offence court cases are visualised against the court in Figure 4.7. It is clear that the Metropolitan and city court has the highest number of drug offences in England and Wales for the period June 2015 to March 2018. West Midlands and Merseyside have the next higher values however, they are around one-fifth of the highest value. Warwickshire and Durham have the lowest drug-related court cases. There is a variation in the number of drug offence court cases finalised among the courts. Therefore, the court is a good variable to consider for classify the total drug offence court cases data into clusters.

Timeline

Description automatically generated with medium confidence

Figure 4.7. Total drug offence cases vs court they finalised from July 2015 to March 2018

Table 4.3 shows the summary of the drug offence court cases in England and Wales. Minimum value, the first quartile, median, mean and the third quartile lie closer when compared to the maximum. The values clustered towards the higher end of the scale have resulted from the Metropolitan and city court area when cross-referencing Table 4.3 and Figure 4.7. The distribution has a high standard deviation which can be a result of higher values from the Metropolitan and City area.

Table 4.3. Summary of the basic statistics for drug offence total court cases from July 2015 to March 2018

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Standard Deviation |
| 4.00 | 38.00 | 63.00 | 98.36 | 97.00 | 1281 | 164.5618 |

When looking at the histogram in Figure 4.8, it is evident that the distribution of drug offence court cases is highly skewed to the left. Furthermore, in the histogram, there are no observations recorded after 400 until 800 which shows the separation of the Metropolitan and city values and other values. It is also clear how far apart the values of the Metropolitan and city court area and the values of other courts lie. One could expect that when the clustering algorithm is used to identify clusters there will be a cluster consisting of Metropolitan and city court area.

Chart, histogram

Description automatically generated

Figure 4.8. Histogram for total drug offence cases

Total drug offence cases show a decline for the most part of the period from July 2015 to March 2018, as visualised in Figure 4.9. The highest values were recorded in Jan 2016, and the lowest values were recorded in December 2017. There is a spike in March 2017, which goes against the downward trend beginning in June 2016. There is no seasonality visible in the graph. However, there are several patterns observed such as December has low court cases compared to the following January. Due to the data variation shown in Figure 4.9, the month is also considered a good variable to classify total drug offence cases.

Chart, bar chart

Description automatically generated

Figure 4.9. Number of drug offence court cases from July 2015 to March 2018

Mean drug Offence court cases are highest in April, according to Figure 4.10, and lowest in December. There is a slight variation in each month. Here the mean number of cases per month is used to visualise the number of court cases each month rather than the total number of cases, as the number of data points for each month is different. When considered Figure 4.9 and 4.10, while there is data variation in the mean number of cases per month, and number of cases each month, no clear seasonality or trend is identified.

Chart, bar chart

Description automatically generated

Figure 4.10. Mean drug offence court cases per month from July 2015 to March 2018

Diagram

Description automatically generated with medium confidence

Figure 4.11. Scatter plot matrix for total drug offence cases finalised from July 2015 to March 2018

There are two clear clusters in the month, and total\_cases scatter plot, months with a low number of drug offence court cases and a high number of drug offences court cases, as shown in Figure 4.11. A small cluster of data points with a higher number of cases can be identified as the data representing Metropolitan and city when cross-referenced with Figure 4.7 given that there are twenty-seven data points in the small cluster representing each month of the data under analysis. Similarly, in the court\_num and tatal\_cases scatter plot, the cluster of data points separated from the other data can be identified as Metropolitan and city.

## Analysis of motoring cases

Metropolitan and City and Bedfordshire have reported more average success of convictions for motoring offences. Wiltshire, Suffolk, Norfolk and Durham have a very high success rate for motoring court offences. There is no identifiable seasonality in the success rate of convictions for motoring offences, as shown in Figure 4.12. Data variation can be observed in the Figure 4.12; therefore, court is considered to include in the classification of motoring cases success rate.

Table

Description automatically generated with medium confidence

Figure 4.12. Motoring court case success rate from July 2015 to March 2018

Table 4.4. Summary of the basic statistics for the success rate of motoring offences court cases from July 2015 to March 2018

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Standard Deviation |
| 61.5 | 84.20 | 87.9 | 87.28 | 91.1 | 100.00 | 5.3 |

The distribution of motoring case success rates is right-skewed because the values are generally high, most values lie over 70%. There are no observations recorded below 61.5% for this period as mentioned in Table 4.4. The distribution in Figure 4.13 shows that the frequency of the values less than 70 is very low and the highest frequency is recorded in the range of 85% to 90%.

Chart, histogram

Description automatically generated

Figure 4.13. Histogram of motoring court case success rate from July 2015 to March 2018

# Linear regression

A simple linear regression model is used here because the data shows a downward trend over the period. A linear model uses the relationship between dependent and independent variables to fit a straight line to the data points. The line results in minimum residual error are selected as the best fit for the given data set.

Three separate simple linear regression models were fitted to the total national number of cases for theft and handling, burglary, and robbery offences. A simple linear model is selected here as the number of total cases and time is used to analyse the trend. Residuals are summarised in Table 5.1. Residuals of the total theft and handling cases and burglary cases are negatively skewed. However, robbery cases have almost normal distributed residual distribution.

Table 5.1. Regression model summary of residuals

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | 1Q | Median | 3Q | Max |
| Total theft and handling cases | -1037.29 | -427.36 | 35.03 | 315.00 | 1276.55 |
| Total burglary cases | -250.66 | -45.87 | 22.92 | 60.26 | 180.39 |
| Total robbery cases | -86.191 | -18.020 | -1.118 | 24.613 | 102.176 |

Then the residual plots [11] are analysed to see the spread of the residuals. Figure 5.1 and Figure 5.2 shows that the errors are randomly distributed and there is no clear pattern to the residuals of the models for theft and handling and burglary. In Figure 5.1, the residuals of most points do not lie much further from the regression line suggesting the line is a good fit.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated |  |

Figure 5.1. Regression model for total theft and handling court cases and residual plot

Figure 5.2 shows that there are two points which lie relatively further away from the regression line. Other than that, most of the points lie closer to the regression line. In both Figure 5.1 and 5.2 the decline over the time is clearly visible with the angle of the line.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated |  |

Figure 5.2. Regression model for total burglary court cases and residual plot

The linear regression model for total robbery cases has resulted in a higher variance in residuals at the beginning of the period as shown in Figure 5.3. There is a higher fluctuation in data for the first three months of the period which is hard to fit into a linear model. After that, while one point towards the middle of the period shows higher residual other points lie relatively closer to the regression line. When looking at the minimum and maximum errors from the Table 5.1 and cross reference to the Figure 5.3, it can be seen that there are only few data points have higher errors. Furthermore, when looking at the magnitude of the residuals, the model for total robbery cases has much higher magnitude error given that the total robbery cases have maximum value of 608.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated |  |

Figure 5.3. Regression model for total robbery court cases and residual plot

According to Table 5.2, the linear model shows that each month finalized theft and handling court cases are reduced by 101.129. The linear model for total burglary cases shows that total cases finalized for burglary offences are reduced by 13.921 each month. Similarly, the model for robbery cases shows that there is a 2.9755 finalized case reduction each month. This result shows that all three categories of total court cases are in decline for the period from July 2015 to March 2018 in England and Wales.

Table 5.2. Regression model summary of coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Estimate | Std. Error | t value | Pr(>|t|) |
| Total theft and handling cases | Intercept | 9529.039 | 176.578 | 53.97 | < 2e-16 \*\*\* |
| no\_months | -101.129 | 9.484 | -10.66 | 6.76e-12 \*\*\* |
| Total burglary cases | Intercept | 1636.370 | 35.241 | 46.434 | < 2e-16 \*\*\* |
| no\_months | -13.921 | 1.893 | -7.355 | 2.8e-08 \*\*\* |
| Total robbery cases | Intercept | 511.7745 | 13.6157 | 37.587 | < 2e-16 \*\*\* |
| no\_months | -2.9755 | 0.7313 | -4.069 | 0.000302 \*\*\* |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |

The standard deviation of the total theft and handling values is 1103.147. The residual standard error is 518.8, as mentioned in Table 5.3. As the residual standard error is much smaller than the standard deviation of the total theft and handling data, the fit is considered a good fit [12]. Additionally, R squared is 0.78. This means explanatory variable can explain 77% of the variation in total theft and handling variable. R squared has a value closer to one also supports that the regression line is a good fit for the variable total theft and handling [12].

Similarly, the regression line for total burglary cases can be accepted as a good fit as it has a 0.63 R squared value. On the other hand, the regression line for the total robbery cases does not have a good fit when considering the 0.34 R squared value. The standard deviation of the total robbery cases is 48.763. When comparing the residual error of 40 for total robbery cases from Table 5.3, it is clear that the regression line is not a good fit for total robbery cases. Because the standard deviation of the variable and residual standard error is much closer, suggesting that using linear regression only a slightly better than using the mean to predict the total robbery cases [12]. Nevertheless, all three models have p values less than 0.001 as shown in Figure 5.3. Therefore, I can conclude that there is very strong evidence against the null hypothesis [13] [14]

Table 5.3. Regression model summary of coefficients

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Residual standard error  on 31 degrees of freedom | Multiple R-squared: | F-statistic: on 1 and 31 DF | p-value |
| Total theft and handling cases | 518.8 | 0.7858 | 113.7 | 6.761e-12 |
| Total burglary cases | 103.5 | 0.6357 | 54.1 | 2.796e-08 |
| Total robbery cases | 40 | 0.3481 | 16.56 | 0.0003015 |

The linear model for the robbery court cases has a higher error rate. Therefore, a different prediction method could be used for robbery cases to obtain a better prediction, such as a nonlinear model. A polynomial regression model was used to fit a regression line for total robbery cases.

However, the accuracy of the model did not increase significantly more than the linear model. For higher power polynomials p-values of the coefficients get higher (Refer to appendix Figure 13). When polynomial with power two was used to fit a model, the residual standard error was 39.07. Since the linear model has a residual standard error of 40 the polynomial model with power two does not show much improvement as shown in Figure 5.5. Additionally, Figure 5.4 shows the residuals of the points with higher residuals have not changed significantly.

Chart, line chart

Description automatically generated

Figure 5.4. Polynomial models of power two for predict total robbery court cases

Text, letter

Description automatically generated

Figure 5.5. Summary of the polynomial models of power two for predicting total robbery court cases

When the power of the polynomial function increased, the p-value increased significantly, and the residual standard error stayed around 39; refer to appendix Figure 13. Therefore, these polynomial plots did not result in better models worth considering.

# Clustering

Clustering is a technique used to divide data into groups based on the homogeneity of the observation. It is unsupervised data classification method and is also a good method for data exploration. Finding the correct variables, the right metrics, and the right number of clusters is challenging and crucial in creating good data clusters.

## Clustering drug offence cases finalised with the K-Means algorithm.

The number of finalised drug offence cases is clustered with the K-Means [15] algorithm. K-Means clustering algorithm is used here as it is computationally efficient and scalable [16]. One of the drawbacks is that number of clusters k must define prior to using the algorithm. Choosing an inappropriate k can result in poor clustering. Elbow method, Average Silhouette Method [17] [16] was used in this analysis to evaluate the quality of clustering to choose an appropriate value for the number of clusters.

The variables total drug offence cases, month, and numeric transformation of the court variable are used for clustering. The scaled values of the variables are used here as the K-Means algorithm uses Euclidean distance to create clusters and the total court cases, which have values much more significant than other attributes with a maximum of 1281 (refer to Table 4.3), which can dominate the distance calculation with the month and court variable with a maximum of 12 and 42 respectively.

Chart

Description automatically generated with medium confidence

Figure 6.1. Within-cluster sum of squares for total Drug offence cases finalised data.

Within-cluster sum of squares (Within\_ss) which gives a variance within the group, was calculated for the different numbers of clusters, and it is used to identify the best value for the number of clusters. The lower variance within the cluster results in better clustering. Figure 6.1 suggests that the elbow of the Within\_ss curve is at five clusters [17], while four clusters and six clusters have closer values to the recognised elbow value.

The average Silhouette Method was used to measure the quality of clustering. It determines how well each observation lies within the cluster [16]. A Silhouette score closer to 1 means the data lie closer to the cluster centre while a zero score means data lie closer to the boundary. Maximum average Silhouette score is produced when using five clusters confirming the selection of the number of clusters earlier with the Within\_ss curve. However, four clusters also stand out as a good option in Figure 6.2.

Chart, line chart

Description automatically generated

Figure 6.2. Within-cluster sum of squares for total Drug offence cases finalised data.

K-Means clustering for four and five clusters is analysed as the number of attributes is relatively small, and the Silhouette score and Within\_ss are fairly close [17].

Those clusters are visualised in Figure 6.3 and Figure 6.4. The function fviz\_cluster used to draw clusters transforms the initial set of variables into a new set through the principal component analysis (PCA) [18]. When dimension reduction was used to plot the clusters, there were overlapping cluster borders. However, in K-Means clustering, there are no overlapping points between clusters.

Chart, radar chart

Description automatically generated

Figure 6.3. Cluster plot for total drug offence cases finalised data with four clusters.

Chart

Description automatically generated

Figure 6.4. Cluster plot for total drug offence cases finalised data with five clusters

Cluster centres are summarised in Table 6.1 and Table 6.2. Cluster number one in the four clusters model can be identified as the cluster containing Metropolitan and City data. It has an extreme z value for the total number of cases. This can be verified by de-normalizing the z value for the court number using the mean and standard deviation of the court number column [19]. The de-normalized values are used to identify the court name of the cluster centres. Similarly, cluster number three can be identified as observations representing Metropolitan & City in the five clusters model.

Both models have a cluster centre around September and February when considering the month. And court variable has cluster centres Metropolitan & City, Dyfed Powys, Staffordshire and Leicestershire. Single cluster based on Metropolitan & City interpreted as the result of the higher number of total cases reported in the court area. Other cluster centres should be analysed further to get more insight into the data by finding similarities in the formed clusters.

Table 6.1. Resulted cluster centres of K-Means clustering with four clusters for total drug offence cases finalised

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster number | Total cases | De-normalized total cases | Month | De-normalized month(rounded) | Court number | Court name |
| 1 | 6.0459 | 1093.3 | 1.6447e-17 | 7 | 0.2061 | Metropolitan & City |
| 2 | -0.0531 | 89.6 | 0.5771 | 9 | 0.9655 | Staffordshire |
| 3 | -0.1399 | 75.3 | -1.2836 | 2 | -0.0117 | Leicestershire |
| 4 | -0.2365 | 59.4 | 0. 5938 | 9 | -0.85 | Dyfed Powys |

Table 6.2. Resulted cluster centres of K-Means clustering with five clusters for total drug offence cases finalized

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster number | Total cases | De-normalized total cases | Month | De-normalized month(rounded) | Court number | Court name |
| 1 | -0.234 | 59.7 | 0.624 | 9 | -0.825 | Dyfed Powys |
| 2 | -0.22 | 62.1 | -1.249 | 2 | -0.825 | Dyfed Powys |
| 3 | 6.04 | 1093.2 | 1.64e-17 | 7 | 0.206 | Metropolitan & City |
| 4 | -0.0571 | 88.9 | 0.624 | 9 | 0.944 | Staffordshire |
| 5 | -0.042 | 91.4 | -1.249 | 2 | 0.944 | Staffordshire |

Table 6.3. Results of K-Means clustering for total drug offence cases finalised.

|  |  |  |
| --- | --- | --- |
| Number of clusters | Total Within\_ss | Between\_ss |
| 4 clusters | 923.9763 | 2475.024 |
| 5 clusters | 653.266 | 2745.734 |

When four clusters are used Within\_ss is higher and Between\_ss is lower. While five clusters have lower Within\_ss and higher Between\_ss. In the cluster plot, Figure 6.3 and Figure 6.4 can also refer to see the difference in Betweenness. The five-cluster model has produced a clearer separation between clusters than the four-cluster model and lower variance within the cluster. Therefore, a model with five clusters fits better to classify the data [20]. Average silhouette plots are analysed to get a further understanding of the separation distance of clusters [20].

Chart, histogram

Description automatically generated with medium confidence

Figure 6.5. Silhouette plot for four clusters with K-Means Classifier

The silhouette plot in Figure 6.5 shows that all the clusters have silhouette scores over the average silhouette width line, which means most values are closer to their cluster centres than other clusters. There is one cluster with a small number of observations. The other three clusters are approximately equal in size. Most of the clusters have positive Silhouette scores, while cluster two has a small number of observations with negative Silhouette scores. This means that cluster two has few observations closer to other clusters than its cluster centre [16].

Chart, histogram

Description automatically generated

Figure 6.6. Silhouette plot for five clusters with K-Means Classifier

In the case of the five clusters model, the Silhouette plot in Figure 6.6 also shows one cluster with a small number of observations. This is the cluster with the Metropolitan and City court area data points as identified analysing cluster centre. The figure also shows that there are no negative Silhouette scores, meaning that there is no probability of data assigned to the wrong cluster. Furthermore, most indexes in the five-cluster model are higher than the values of four cluster model. Based on these observations model with five clusters seems a better option to classify data even though the cluster sizes vary in size [21]. The results show that both court and month can be used to classify total finalised drug offence cases in England and Wales for the period from July 2015 to March.

## Fuzzy clustering

Fuzzy clustering(Fuzzy C-Means) [22] [23]with five clusters was used to cluster data and compare it with the results of the K-Means Classifier, as the best clustering was observed in K-Means with five clusters. The one small cluster with Metropolitan and City data is not observed here. The cluster with Metropolitan and City data has expanded to include other data points.

Chart, diagram

Description automatically generated

Figure 6.7. Fuzzy clusters model with five clusters

All the clusters have approximately the same width when using the Fuzzy clustering. According to Figure 6.8, the silhouette plot shows that all of the clusters have points closer to other clusters than their clusters because all clusters have negative silhouette scores. Especially cluster number five has more negative indexes than positive indexes meaning that possibility of those points belonging to another cluster is high.

Chart, funnel chart

Description automatically generated

Figure 6.8. Silhouette plot for five clusters with Fuzzy clustering

The K-Means clustering algorithm with five clusters results in the best clustering for total finalised drug offence cases in England and Wales for the period from July 2015 to March using month and court variables.

# Classification

A classification model can be trained to classify data when the class labels of the data are known. However, when a data set is used to train a model, it is possible that the model works well for the data used to train the model. Therefore, to use the model to classify unknown data, the accuracy of the model should be tested with known data. Using a randomly selected sample of the data set to train a model and then using the rest of the data to validate the model is a common practice [24]. This method is called data partitioning. Here, the train data set is used to train the model and then the test data set is used to evaluate the performance of the model using a confusion matrix.

## Decision trees classifier

A decision tree [25] [26]classifier is used here as it is easy to identify important variables used in the classification and interpret how the analysis has happened. Furthermore, decision trees produce fast classifier models.

The percentage of convicted court cases for motoring offences was categorised as average, high and very high. Then three dependent variables, court, month, and total motoring court cases, were used to build a decision tree classifier.

Timeline

Description automatically generated

Figure 7.1. Decision tree classifier for the percentage of motoring offence convictions

When looking at the right side of the tree, it is evident that when total court cases are low, the court plays the main role in classifying the success rate of the motoring offence convictions. Therefore, it is possible to say that there are areas with a lower success rate than other areas. Namely, when looking at the tree in Figure 7.1 and cross-reference to the Appendix table 1, Bedfordshire and Avon and Somerset have higher average success rates. It is possible to investigate causes and factors affecting the success rate of these court areas to get a better understanding.

Table 7.2. Confusion matrix of the decision tree classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | average | high | very high |
| average | 8 | 18 | 0 |
| high | 0 | 144 | 15 |
| very high | 1 | 31 | 40 |

The accuracy of the model is 71.36, as calculated using the equation below. The function calculate accuracy based on the confusion matrix.

accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) \* 100}

The decision tree was pruned [27] using a minimum split, maximum depth and minbucket methods. The best results for pruning methods were resulted from minimum split method. The minimum split is set to 10% of the train data set to avoid overfitting as the data set can be considered large with 907 observations in the training set [28].

Timeline

Description automatically generated

Figure 7.2. Decision tree classifier for the percentage of motoring offence convictions with a minimum split of 90

The decision tree uses the total motoring offence and court to classify the success rate of the motoring offences. When the total court cases for motoring offences is larger than or equal to 136, it does not result in a very high success rate. Furthermore, the success rate has become average when the total motoring court cases go above 604. I can suggest looking into this factor further to improve the success rate of convictions.

Table 7.3. Confusion matrix of the decision tree classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | average | high | very high |
| average | 4 | 15 | 7 |
| high | 1 | 109 | 19 |
| very high | 0 | 35 | 37 |

The accuracy of the new model is 66.08. The accuracy is reduced after changing the minimum split parameter. Therefore, the better results are observed for the decision tree without minimum split.

According to the resulted decision trees, it is clear that the month is not a good branching factor for the percentage of successful convictions. Therefore, I can conclude that the rate of successful convictions for motoring offences does not depend on the month for the period from July 2015 to March 2018 in England and Wales. The total number of court cases and the court is suitable parameters to predict the successful conviction rate for motoring offences for this period. Since the accuracy of the model is not very high, it is good to look for more independent variables to classify success rates using decision trees. Once the factors contributing to the success rate are identified, using a Random Forest for predictions is a better option as this will improve the accuracy of the classification. Because it is a classifier using a group of Decision Trees and averaging them to produce a result.

# Conclusion

According to the descriptive analysis and the result of the linear regression models, it is clear that there is enough evidence to reject the null hypothesis for hypothesis 2.1. Because all three regression lines have p-values less than 0.001 and, when analysing R-squared values, it also suggested the lines are a good fit for the data set. Therefore, I can conclude that total court cases for theft and handling, burglary and robbery in England and Wales decreased from July 2015 to March 2018.

Total drug offences court cases from July 2015 to March 2018 in England and Wales can be clustered using the court and month variables. Clusters show good separation and low variance when using the K-Means clustering algorithm with five clusters. Therefore, for hypothesis 2.2, the null hypothesis can be rejected.

Based on the classification results with the decision tree classifier, I can conclude that the percentage of successful convictions for motoring offences depends on the court area and the total number of court cases for the period from July 2015 to March 2018 in England and Wales. A model has classified the data with 71% accuracy, which suggests that there is evidence that the percentage of successful convictions for motoring offences depends on the court area and the total number of court cases. Therefore, for hypotheses 2.3 and 2.4 null hypothesis can be rejected.

The decision tree did not use the variable month for classifying the percentage of successful convictions for motoring offences. Therefore, there is no evidence against null hypnosis 2.5. The analysis did not find that the rate of successful convictions for motoring offences is dependent on the month for the period from July 2015 to March 2018 in England and Wales. According to this analysis, null hypothesis cannot be rejected.

# References

|  |  |
| --- | --- |
| [1] | The Crown Prosecution Service , “CPS case outcomes by principal offence,” 13 March 2022. [Online]. Available: https://www.cps.gov.uk/publication/cps-case-outcomes-principal-offence. |
| [2] | Crown Prosecution Service, “Crown Prosecution Service Case Outcomes by Principal Offence Category Data,” 05 October 2018. [Online]. Available: https://data.gov.uk/dataset/89d0aef9-e2f9-4d1a-b779-5a33707c5f2c/crown-prosecution-service-case-outcomes-by-principal-offence-category-data. [Accessed 10 February 2022]. |
| [3] | W. Yarberry, “Lubridate: Date and Time Processing,” in *CRAN Recipes: DPLYR, Stringr, Lubridate, and RegEx in R*, New York, Apress, 2021. |
| [4] | RDocumentation, “na\_interpolation: Missing Value Imputation by Interpolation,” version 3.2. [Online]. Available: https://www.rdocumentation.org/packages/imputeTS/versions/3.2/topics/na\_interpolation. [Accessed 14 April 2022]. |
| [5] | V. a. T. Ashikaga., “An examination of the nearest neighbor rule for imputing missing values.,” *Proc. Statist. Computing Sect., Amer. Statist. Ass,* p. 326–331, 1980. |
| [6] | S. Moritz and T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” *The R Journal,* vol. 9, pp. 207-218, 2017. |
| [7] | RDocumentation, “approxfun: Interpolation Functions,” 1 Jan 2019. [Online]. Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/approxfun. [Accessed 10 May 2022]. |
| [8] | C. D. B. M. Ç.-R. David M. Diez, “Examining Numerical Data,” in *OpenIntro Statistics*, OpenIntro, 2019, pp. 41-60. |
| [9] | Parliment of the United Kindom, “REORGANISATION OF CPS TO MATCH POLICE FORCE AREAS,” 19 July 2000. [Online]. Available: https://publications.parliament.uk/pa/cm199900/cmselect/cmhaff/476/0050913.htm. [Accessed 20 May 2022]. |
| [10] | R. K. Mariette Awad, “Machine Learning: Similarity Matrix,” in *Efficient Learning Machines*, Berkeley, CA, Apress, 2015, pp. 35-37. |
| [11] | M. C.-R. C. B. David Diez, “Introduction to linear regression,” in *OpenIntro Statistics.*, OpenIntro, 2019, pp. 304-340. |
| [12] | W. L. W. S Christian Albright, “Regression Analysis: Estimating Relationships,” in *Business analytics : data analysis and decision making*, Boston, MA, Cengage, 2019, pp. 412-471. |
| [13] | P. Singh, “P Value, Statistical Significance and Clinical Significance,” *Journal of Clinical and Preventive Cardiology,* vol. 2, no. 4, pp. 202-204, Oct 2013. |
| [14] | D. Dietrich, R. Heller and B. Yang, “Linear Regression,” in *Data Science and Big Data Analytics*, Indianpolis, IN, John Wiley & Sons,, 2015, pp. 189-207. |
| [15] | M. Awad and R. Khanna, “Popular Machine Learning Algorithms,” in *Efficient learning machines : theories, concepts, and applications for engineers and system Designers*, New York, Apress Open, 2015, pp. 10-15. |
| [16] | A. Géron, “Clustering,” in *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*, Sebastopol, CA, O'Reilly Media, Incorporated, 2019, pp. 236 - 259. |
| [17] | D. R. H. B. Y. Dietrich, “Advanced Analytical Theory and Methods: Clustering,” in *Data Science and Big Data Analytics* , Indianpolis, John Wiley & Sons. , 2015, pp. 143-162. |
| [18] | Airforce Institute Of Technology, “AFIT Data Science Lab R Programming Guide,” Airforce Institute Of Technology, [Online]. Available: https://afit-r.github.io/kmeans\_clustering. [Accessed 22 April 2022]. |
| [19] | RDocumentation Search all packages and functions, “scale: Scaling and Centering of Matrix-like Objects,” 12 12 2019. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale. [Accessed 1 5 2022]. |
| [20] | F. Z. T. Iglesias and A. Zimek, “Clustering refinement,” *International Journal of Data Science and Analytics,* no. 12, p. 333–353, 2021. |
| [21] | P. J.Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987. |
| [22] | E. Cox, “Fuzzy Clustering,” in *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*, Morgan Kaufmann, 2005, pp. 207-264. |
| [23] | I. B. Türksen, “Fuzze clustering method,” in *An Ontological and Epistemological Perspective of Fuzzy Set Theory*, Elsevier Science, 2005. |
| [24] | W. L. W. S. Christian Albright, “Classification methods,” in *Business Analytics*, Boston, MA, Cengage, 2019, pp. 841-859. |
| [25] | P. N. Barry de Ville, “Chapter 1: Decision Trees—What Are They?,” in *Decision Trees for Analytics Using SAS Enterprise Miner*, Cary:NC, SAS Institute, 2013. |
| [26] | S. Saxena, “Decision Trees: Strengths, Weaknesses, and Uses,” in *Tree-Based Machine Learning Methods in SAS Viya*, SAS Institute, 2022. |
| [27] | S. Saxena, “Tuning a Decision Tree,” in *Tree-Based Machine Learning Methods in SAS Viya*, SAS Institute, 2022. |
| [28] | D. R. H. B. Y. Dietrich, “Discovering, Analyzing, Visualizing and Presenting Data,” in *Data Science and Big Data Analytics*, Indianpolis, John Wiley & Sons., 2015, pp. 237-240. |
| [29] | R. Yucel, “State of the Multiple Imputation Software,” *Journal of statistical software,* vol. 45, no. 1, p. 1–7, 2011. |
| [30] | The Crown Prosecution Service, “Theft Act Offences,” 2017. [Online]. Available: https://www.cps.gov.uk/legal-guidance/theft-act-offences#:~:text=%22A%20person%20is%20guilty%20of,and%20there%20subjected%20to%20force.%22. [Accessed 20 April 2022]. |

# Appendix

Table 1. Court names and numeric transformed numbers for machine modelling

|  |  |
| --- | --- |
| Court name | Representing number |
| Avon & Somerset | 1 |
| Bedfordshire | 2 |
| Cambridgeshire | 3 |
| Cheshire | 4 |
| Cleveland | 5 |
| Cumbria | 6 |
| Derbyshire | 7 |
| Devon & Cornwall | 8 |
| Dorset | 9 |
| Durham | 10 |
| Dyfed Powys | 11 |
| Essex | 12 |
| Gloucestershire | 13 |
| GreaterManchester | 14 |
| Gwent | 15 |
| Hampshire | 16 |
| Hertfordshire | 17 |
| Humberside | 18 |
| Kent | 19 |
| Lancashire | 20 |
| Leicestershire | 21 |
| Lincolnshire | 22 |
| Merseyside | 23 |
| Metropolitan & City | 24 |
| Norfolk | 25 |
| North Wales | 26 |
| North Yorkshire | 27 |
| Northamptonshire | 28 |
| Northumbria | 29 |
| Nottinghamshire | 30 |
| South Wales | 31 |
| South Yorkshire | 32 |
| Staffordshire | 33 |
| Suffolk | 34 |
| Surrey | 35 |
| Sussex | 36 |
| Thames Valley | 37 |
| Warwickshire | 38 |
| West Mercia | 39 |
| West Midlands | 40 |
| West Yorkshire | 41 |
| Wiltshire | 42 |

Chart, bar chart, histogram

Description automatically generated

Figure 1. The national number of theft and handling convictions from July 2015 to March 2018

Chart, bar chart, histogram

Description automatically generated

Figure 2. The national number of theft and handling unsuccessful convictions from July 2015 to March 2018

Chart, bar chart, histogram

Description automatically generated

Figure 3. The national number of burglary convictions from July 2015 to March 2018

Chart, bar chart, histogram

Description automatically generated

Figure 4. The national number of unsuccessful burglary convictions from July 2015 to March 2018

Chart, bar chart

Description automatically generated

Figure 5. The national number of robbery convictions from July 2015 to March 2018

Chart, bar chart

Description automatically generated

Figure 6. The national number of robbery unsuccessful convictions from July 2015 to March 2018

Chart, bar chart

Description automatically generated

Figure 7. The national number of drug offence convictions from July 2015 to March 2018

Chart, bar chart

Description automatically generated

Figure 8. The national number of drug offence unsuccessful convictions from July 2015 to March 2018

Chart, bar chart, histogram

Description automatically generated  
Figure 9. The national percentage of motoring offences from July 2015 to March 2018

Chart, scatter chart

Description automatically generated

Figure 10. Residual plot for linear regression model of total theft and handling cases

Chart, scatter chart

Description automatically generated

Figure 11. Residual plot for linear regression model of total burglary cases

Chart, scatter chart

Description automatically generated

Figure 12. Residual plot for linear regression model of total robbery cases

Text

Description automatically generated

Figure 13. Polynomial models with the power of three and four for total robbery court cases

Chart, line chart

Description automatically generated

Figure 14. Polynomial model with power three for total robbery court cases

Chart, line chart

Description automatically generated

Figure 15. Polynomial model with power four for total robbery court cases