Chanya Deshani Subasingha Arachchige

[Email address]

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

[Document title]

[Document subtitle]

Table of Contents

[1 Introduction 1](#_Toc100352310)

[2 Hypothesis 1](#_Toc100352311)

[2.1 Court cases related to stealing are decreasing from 2015 June to 2018 March. 1](#_Toc100352312)

[3 Data preparation 2](#_Toc100352313)

[3.1 Data Cleaning 2](#_Toc100352314)

[3.2 Feature engineering 2](#_Toc100352315)

[4 Descriptive analysis 3](#_Toc100352316)

[5 linear regression 6](#_Toc100352317)

# Introduction

The Crown Prosecution Service (CPS) publishes a report monthly on outcome of CPS proceedings in magistrates' courts and in the crown court by principal offence category. The data set contains 27 monthly analysis reports for the thirty-three months period from July 2015 to March 2018, the period after new the new data assurance regime was introduced [1]. Each report contains number and percentage of convictions and unsuccessful conviction by defendant basis.

One court case can represent several observations in the data set as there can be number of defendants in one court case [1]. Principal offence is deciding at the time of case finalization. They are categorized 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 column for administrative finalisations, the cases which could not proceed due to an administrative issue such as unexecuted warrant for the arrest of the defendant, or summons have not 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 summarized without categorising by into principal offences [1].

# Hypothesis

## Number of national court cases in the United Kingdom for theft and handling, burglary and robbery are decreasing from 2015 June to 2018 March.

# Data preparation

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

First, I noticed that the column names had long names with spaces which make it hard to work with in R functions. Therefore, all the long column names renamed to their abbreviations. This also make the code more user-friendly and readable. I used capital letters for the column names to make it clear that it is abbreviations. Then a new data frame was created merging all the data sets. As a preparation to combine datasets, new variable *Date* was introduced to all the data sets, even the data set only had the month and year, date was added as first day of the respective month for the simplicity. This reduce code complexity and processing time as one data set can be processed with better performance than looping through all the data sets each time.

## Data Cleaning

The dataset examined for missing values but there are no missing values in the data file for month. There are some observations recorded with “-“ in the percentage column in homicide percentages columns, these are not missing values but resulted by zero number of court cases for respective month. However, some months were missing when considering whole period.

Then data types of the dataset were checked. Columns with number of court cases are numeric, the columns with percentages of court outcomes have character type values. 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 first, remove percentage sign, and then convert values to numeric values.

Then after round of data visualizing with box plots and bar graphs, noticed that the data set containing national values of the principal offences court outcomes have extremely lager values. As shown in Figure 4.1, the national values which are outliers and much larger than all the other values, hinders the data visualizing and identifying patterns and prominent features. Hence the national data has separated from the data set and copied to another data frame. Which can analyse separately.

Column *percentage of L motoring offences unsuccessful* has removed from the data set as this column has values only 100 and NA. Hence the column has no data variation resulting it not being valuable to analysis for pattern recognition or prediction.

## Feature engineering

New columns, N\_SC (for stealing convictions) and N\_SU (for stealing unsuccessful) was created using the values of the four columns that contained number of court outcomes related to burglary and theft and handling. Because the initial exploration showed that they have followed a downward trend and correlation with the number of months are close. Furthermore, the type of crime is same manner. Similarly, created another dataset with combining robbery values.

As a preparation to combine the columns, which has different ranges, I first scaled the columns using *scale* function of the R which returns the z-score of the values in one column.

## Missing value imputation

A column no\_months (number of months) created using the date column. Then using the complete value set for number of months, filled the missing value.

Linear interpolation [2] [3] [4]is used to fill missing values for columns 'N\_BC', 'N\_BU', 'N\_RC', ' N\_RU', 'N\_THC', 'N\_THU' as the column values has a downward trend. Using simple imputations such as mean, or median will produce values that go against the trend. Hence, the linear imputation, which uses values closer to missing data point to impute missing data was a more appropriate choice to preserve the trend.

# Descriptive analysis

Initial descriptive analysis using box plot showed that there are outliers present in the dataset, the box plot is used here as it enables clearly identify outliers. National data is identified as an extreme outlier as shown in figure 4.1 and separated from the dataset as described in section 3.1(Data cleaning). National data being an outlier is an obvious result as the national column contains the summation over all the other observation in the same column for the respective month.

Chart, histogram, scatter chart

Description automatically generated

Figure 4.1 Number of theft and handling conviction in June 2015

It is clear from the figure 4.2 that the national data for theft and handling outcomes are following a clear downward trend. Even though, the unsuccessful convictions numbers are much smaller than convictions it also follows a downward trend. The correlation analysis is in Table 4.1 showing the trend properly.

Chart, bar chart, histogram

Description automatically generated

Figure 4.2 National number of theft and handling court cases from July 2015 to March 2018

Total case numbers of robbery, burglary and theft and handling have a positive correlation as shown in the scatter plot matrix in the Figure 4.3. Even though the robbery case count does not follow a linear relationship with theft and handling total cases, it shows a positive relationship. When compared with number of month, all three variables shows negative correlation.

Chart, scatter chart

Description automatically generated

Figure 4.3 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 was done using spearman correlation. Burglary case outcomes and theft and handling case outcomes have a strong negative correlation with the date column as shown in Table 4.1. Robbery convictions also have a negative corelation, however, that is not strong as other two type of case outcomes. Correlation analysis results shows that there is a decline in the number court cases for burglary, theft and handling and robbery court cases over the time as both conviction and unsuccessful convictions are decreasing. Specially, burglary and theft and handling court cases show clear decline over the time.

Table 4.1 Correlation of the number of case outcome and date

|  |  |
| --- | --- |
| Number of national case outcome | Spearman correlation with date |
| Theft And Handling Convictions | -0.861 |
| Theft And Handling Unsuccessful | -0.919 |
| Burglary Convictions | -0.761 |
| Burglary Unsuccessful | -0.850 |
| Robbery Convictions | -0.478 |
| Robbery Unsuccessful | -0.479 |

# linear regression

Linear model was fitted to a new variable created using theft and handling and burglary scaled values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total theft and handling cases | Total burglary cases | Total robbery cases |
| Residuals: | Min 1Q Median 3Q Max  -1037.29 -427.36 35.03 315.00 1276.55 | Min 1Q Median 3Q Max  -250.66 -45.87 22.92 60.26 180.39 | Min 1Q Median 3Q Max  -86.191 -18.020 -1.118 24.613 102.176 |
| Coefficients: | Estimate Std. Error t value Pr(>|t|)  (Intercept) 9529.039 176.578 53.97 < 2e-16 \*\*\*  no\_months -101.129 9.484 -10.66 6.76e-12 \*\*\* | Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1636.370 35.241 46.434 < 2e-16 \*\*\*  no\_months -13.921 1.893 -7.355 2.8e-08 \*\*\* | Estimate Std. Error t value Pr(>|t|)  (Intercept) 511.7745 13.6157 37.587 < 2e-16 \*\*\*  no\_months -2.9755 0.7313 -4.069 0.000302 \*\*\* |
| Residual standard error: | 518.8 on 31 DF | 103.5 on 31 DF | 40 on 31 DF |
| Multiple R-squared: | 0.7858, | 0.6357, | 0.3481 |
| Adjusted R-squared: | 0.7789 | 0.6239 | 0.3271 |
| F-statistic: | 113.7 on 1 and 31 DF, | 54.1 on 1 and 31 DF, | 16.56 on 1 and 31 DF |
| p-value: | 6.761e-12 | 2.796e-08 | 0.0003015 |