ANALYSIS OF THE CROWN PROSECUTION SERVICE CASES OUTCOMES DATA

by Zeba Khadhijah

I. Understanding the dataset

In data science, developing a basic understanding of the context of the data is the first step to make any meaningful analysis.

The Crown Prosecution Service (CPS) is a public agency that prosecutes criminal cases in England and Wales. While the cases may have been investigated by the police or other investigative organisations, the CPS carries out all prosecution related functions independently. Responsibilities of the CPS include deciding which cases to prosecute, what charges to press, presenting cases to the court and aiding victims and prosecution witnesses.[1]

Data used in this project is a monthly breakdown of criminal case outcomes by principal offence category and CPS Area. This dataset can be accessed freely from the data.gov.uk website. Data from the years 2014-2016 has been selected as the training data for this project, and data from 2017 and 2018 will be used as test data to validate our prediction models. Month wise data for 42 counties in England and Wales is available, with few months of missing data.

The dataset contains the following information for different crimes:

1. Number of convictions: when court rules in favour of the prosecution and the criminal is convicted and sentenced appropriately (i.e., a successful prosecution).

2. Conviction Rate: Percentage of successful prosecutions (calculated from the total number of prosecutions)
3. Number of unsuccessful convictions: When defendant is not convicted by the court (which may be for a variety of reasons).
4. Rate of unsuccessful convictions: Percentage of unsuccessful prosecutions calculated from the total number of cases prosecuted.
5. Number of admin finalised unsuccessful: Cases where prosecution cannot proceed due to variety of reasons (for e.g., insufficient evidence, suspect fled etc) and the proceedings are indefinitely adjourned.[2] CPS data does not identify the nature of crime for cases resulting in an administrative finalisation.

The above information is available for the 12 categories of offences:

1. Homicide: Killing of another person.
2. Offences against the person: range of crimes that harm another person such as assault and abuse[3].
3. Sexual offences: Molestation, rape, statutory rape etc.
4. Burglary: entering a property illegally usually with intent to steal.
5. Robbery: stealing, often by threatening or using physical aggression towards a person.
6. Theft & handling: stealing and selling stolen goods.
7. Fraud & forgery: cheating someone, creating fake legal materials with intent to defraud.
8. Criminal damage: destroy property/intent to destroy without lawful excuse[4].
9. Drugs offences: Possession, supply, production and importation/exportation[5].

10. Public order: Crimes that interfere with the regular functioning of people and society.
11. Motoring: careless or inconsiderate driving that breaks established road rules.
12. All other offences excluding motoring.

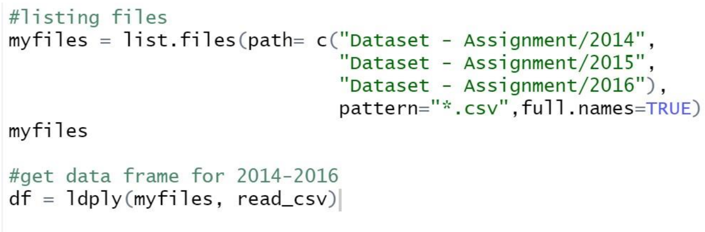
It is worth noting that number of convictions for a given crime can give us a picture of how widespread a particular crime may be. For example, a quick glance of the csv data files reveal that there are very few homicide cases in a given month compared to motoring offences, and that there are typically a greater number of convictions in counties with large cities like London. On the other hand, conviction rates for a specific crime shows how effective the prosecution is in getting a defendant convicted (for e.g., in April 2014, prosecution was successful in winning convictions for 94.2% of drug cases but only 72.2% of sex offences).

To further investigate the data we need to conduct exploratory data analyses (EDA) to identify trends and relationships between variables in our data. Data Manipulation will be required to perform EDA.

II. DATA PRE-PROCESSING

Reading the data

Our data set is split into numerous month-wise csv files. Reading the data and merging them into a single data frame in R makes analysis easier and code simpler.

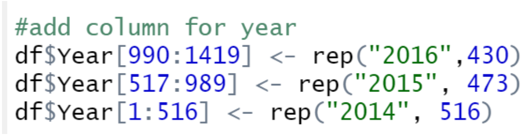


```
#listing files
myfiles = list.files(path= c("Dataset - Assignment/2014",
                             "Dataset - Assignment/2015",
                             "Dataset - Assignment/2016"),
                             pattern="*.csv",full.names=TRUE)
myfiles

#get data frame for 2014-2016
df = ldply(myfiles, read_csv)
```

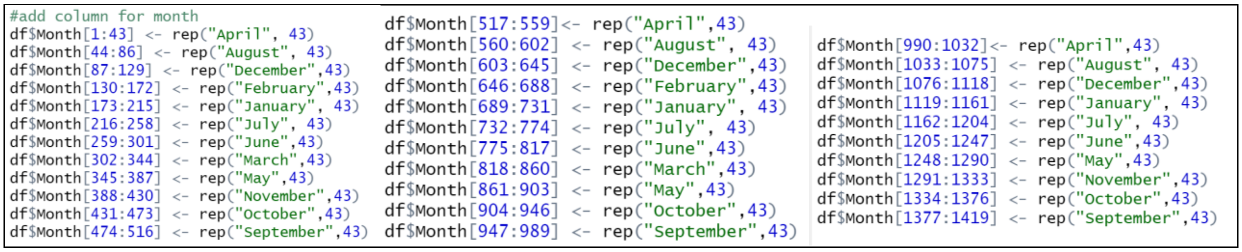
Fig 1. Reading all csv files into one data frame in R

Data Manipulation

On merging all the csv files, we do not have the core identifying information i.e., the month and years for each entry, as this information was only present in the title of the csv files. Adding a column for month and year will greatly help makes sense of our data. This can be done using the following code:



```
#add column for year
df$Year[990:1419] <- rep("2016",430)
df$Year[517:989] <- rep("2015", 473)
df$Year[1:516] <- rep("2014", 516)
```



```
#add column for month
df$Month[1:43] <- rep("April", 43)
df$Month[44:86] <- rep("August", 43)
df$Month[87:129] <- rep("December",43)
df$Month[130:172] <- rep("February",43)
df$Month[173:215] <- rep("January", 43)
df$Month[216:258] <- rep("July", 43)
df$Month[259:301] <- rep("June",43)
df$Month[302:344] <- rep("March",43)
df$Month[345:387] <- rep("May",43)
df$Month[388:430] <- rep("November",43)
df$Month[431:473] <- rep("October",43)
df$Month[474:516] <- rep("September",43)
```

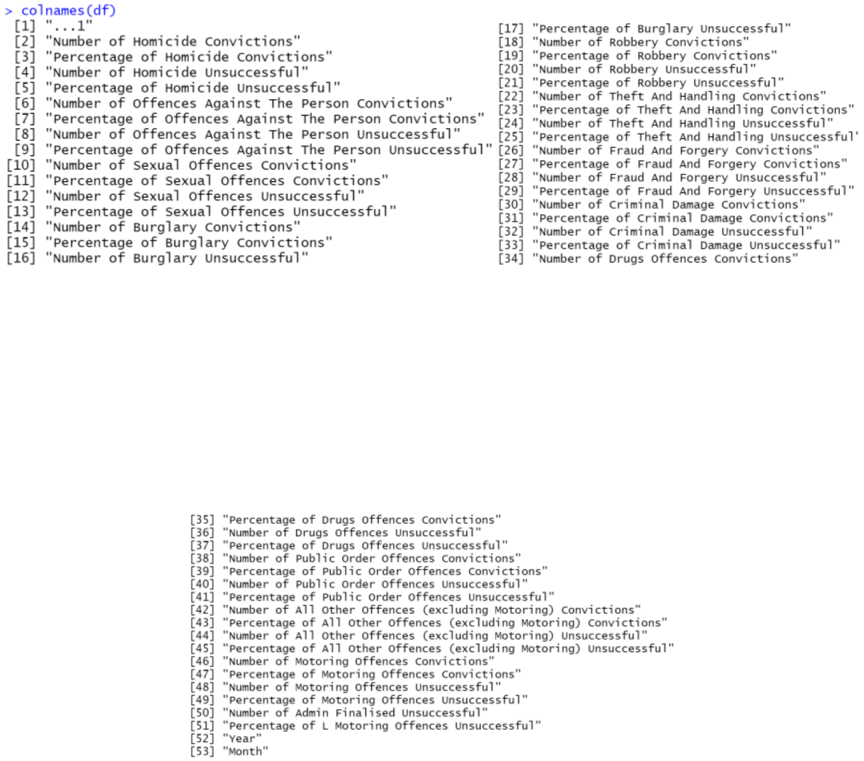
```
df$Month[517:559]<- rep("April",43)
df$Month[560:602] <- rep("August", 43)
df$Month[603:645] <- rep("December",43)
df$Month[646:688] <- rep("February",43)
df$Month[689:731] <- rep("January", 43)
df$Month[732:774] <- rep("July", 43)
df$Month[775:817] <- rep("June",43)
df$Month[818:860] <- rep("March",43)
df$Month[861:903] <- rep("May",43)
df$Month[904:946] <- rep("October",43)
df$Month[947:989] <- rep("September",43)
```

```
df$Month[990:1032]<- rep("April",43)
df$Month[1033:1075] <- rep("August", 43)
df$Month[1076:1118] <- rep("December",43)
df$Month[1119:1161] <- rep("January", 43)
df$Month[1162:1204] <- rep("July", 43)
df$Month[1205:1247] <- rep("June",43)
df$Month[1248:1290] <- rep("May",43)
df$Month[1291:1333] <- rep("November",43)
df$Month[1334:1376] <- rep("October",43)
df$Month[1377:1419] <- rep("September",43)
```

Fig 2. Adding columns for month and year in data frame

The above code the uses the *rep function* which repeats the same value for a specified number of times. Rows in the data frame related to each month and year are indexed and the rep functions with the appropriate arguments have been applied to get columns with the appropriate year and month.

We can use the *view()* function to see the data frame with the newly added columns, or the *colnames()* function to get a list of column headers as seen below:



```
> colnames(df)
 [1] "...1"
 [2] "Number of Homicide Convictions"
 [3] "Percentage of Homicide Convictions"
 [4] "Number of Homicide Unsuccessful"
 [5] "Percentage of Homicide Unsuccessful"
 [6] "Number of Offences Against The Person Convictions"
 [7] "Percentage of Offences Against The Person Convictions"
 [8] "Number of Offences Against The Person Unsuccessful"
 [9] "Percentage of Offences Against The Person Unsuccessful"
[10] "Number of Sexual Offences Convictions"
[11] "Percentage of Sexual Offences Convictions"
[12] "Number of Sexual Offences Unsuccessful"
[13] "Percentage of Sexual Offences Unsuccessful"
[14] "Number of Burglary Convictions"
[15] "Percentage of Burglary Convictions"
[16] "Number of Burglary Unsuccessful"
[17] "Percentage of Burglary Unsuccessful"
[18] "Number of Robbery Convictions"
[19] "Percentage of Robbery Convictions"
[20] "Number of Robbery Unsuccessful"
[21] "Percentage of Robbery Unsuccessful"
[22] "Number of Theft And Handling Convictions"
[23] "Percentage of Theft And Handling Convictions"
[24] "Number of Theft And Handling Unsuccessful"
[25] "Percentage of Theft And Handling Unsuccessful'
[26] "Number of Fraud And Forgery Convictions"
[27] "Percentage of Fraud And Forgery Convictions"
[28] "Number of Fraud And Forgery Unsuccessful"
[29] "Percentage of Fraud And Forgery Unsuccessful"
[30] "Number of Criminal Damage Convictions"
[31] "Percentage of Criminal Damage Convictions"
[32] "Number of Criminal Damage Unsuccessful"
[33] "Percentage of Criminal Damage Unsuccessful"
[34] "Number of Drugs Offences Convictions"
[35] "Percentage of Drugs Offences Convictions"
[36] "Number of Drugs Offences Unsuccessful"
[37] "Percentage of Drugs Offences Unsuccessful"
[38] "Number of Public Order Offences Convictions"
[39] "Percentage of Public Order Offences Convictions"
[40] "Number of Public Order Offences Unsuccessful"
[41] "Percentage of Public Order Offences Unsuccessful"
[42] "Number of All Other Offences (excluding Motoring) Convictions"
[43] "Percentage of All Other Offences (excluding Motoring) Convictions"
[44] "Number of All Other Offences (excluding Motoring) Unsuccessful"
[45] "Percentage of All Other Offences (excluding Motoring) Unsuccessful"
[46] "Number of Motoring Offences Convictions"
[47] "Percentage of Motoring Offences Convictions"
[48] "Number of Motoring Offences Unsuccessful"
[49] "Percentage of Motoring Offences Unsuccessful"
[50] "Number of Admin Finalised Unsuccessful"
[51] "Percentage of L Motoring Offences Unsuccessful"
[52] "Year"
[53] "Month"
```

Fig 3. Output for colnames() function

As seen in the above output, the new columns have been added to the data frame.

The next step in preparing our data for analysis is to select the required columns and rows, and changing the column names.

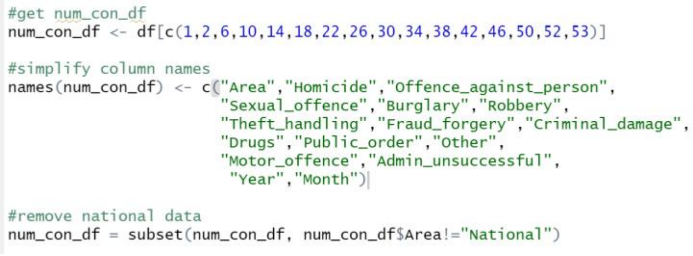
The two main areas of interest in the present data are the number of convictions and the percentage of convictions (i.e., conviction rates). These will be split into two data frames (resulting in not using the columns with information about unsuccessful convictions and their percentage).

In terms of selecting rows, it is important to remove rows relating to National data. This is because, national data is not an independent observation on its own, but is computed by adding up the numbers for the 42 counties in the data set for each month. Generating descriptive statistics, visualisations or models will be meaningless if all rows are not independent observations.

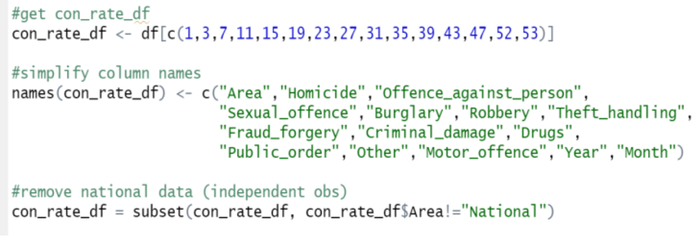
Changing column names serve the following purpose:

(i) R code will be cleaner and more efficient if column names do not have spaces in them.

(ii) Visualisations will be neater when names are shorter, especially when there is a large amount of data to plot.



```
#get num_con_df
num_con_df <- df[c(1,2,6,10,14,18,22,26,30,34,38,42,46,50,52,53)]

#simplify column names
names(num_con_df) <- c("Area","Homicide","Offence_against_person",
                       "Sexual_offence","Burglary","Robbery",
                       "Theft_handling","Fraud_forgery","Criminal_damage",
                       "Drugs","Public_order","Other",
                       "Motor_offence","Admin_unsuccessful",
                        "Year","Month")

#remove national data
num_con_df = subset(num_con_df, num_con_df$Area!="National")
```



```
#get con_rate_df
con_rate_df <- df[c(1,3,7,11,15,19,23,27,31,35,39,43,47,52,53)]

#simplify column names
names(con_rate_df) <- c("Area","Homicide","Offence_against_person",
                        "Sexual_offence","Burglary","Robbery","Theft_handling",
                        "Fraud_forgery","Criminal_damage","Drugs",
                        "Public_order","Other","Motor_offence","Year","Month")

#remove national data (independent obs)
con_rate_df = subset(con_rate_df, con_rate_df$Area!="National")
```

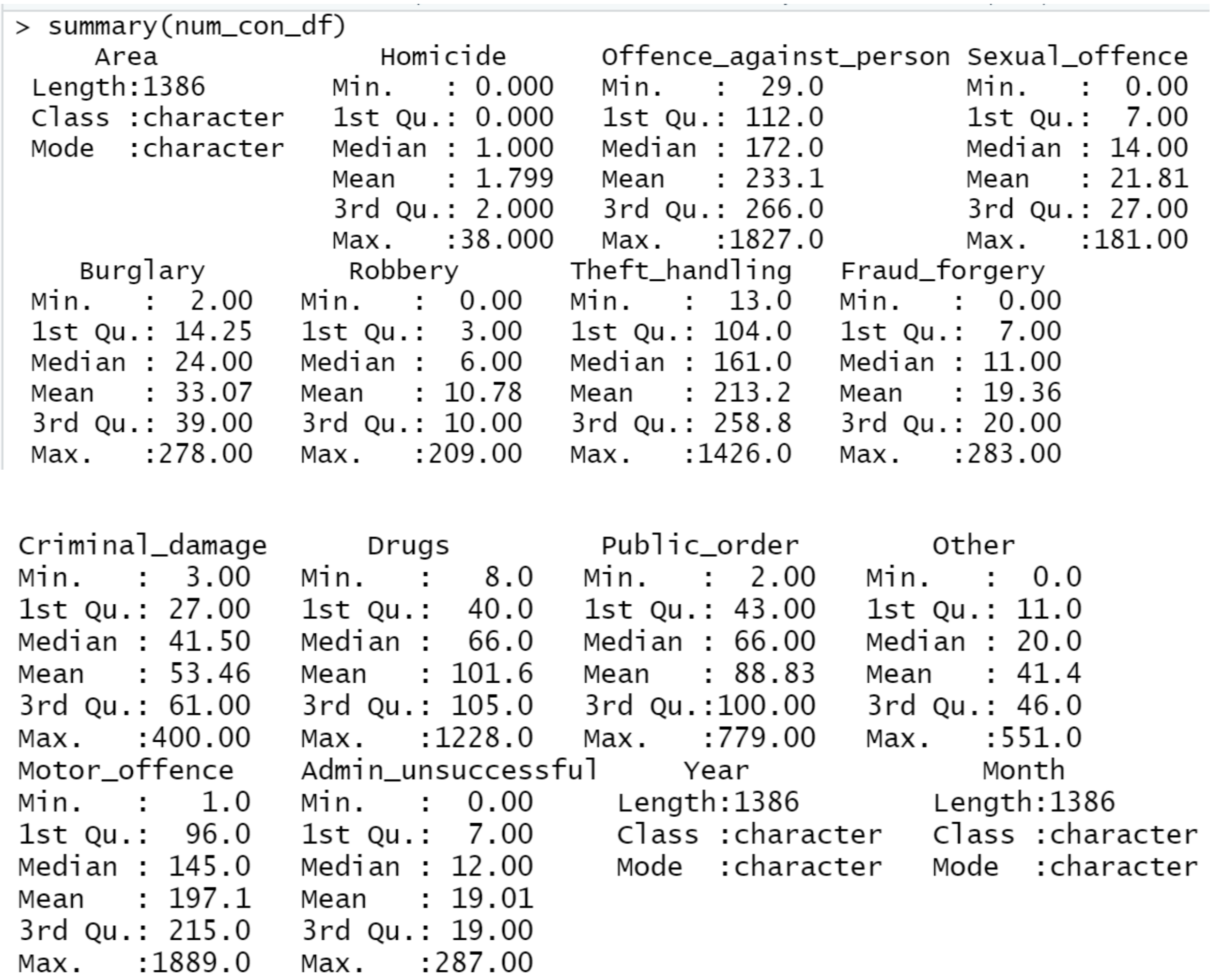
Fig 4. Modifying data frames: Selecting columns, rows and renaming columns

III. EXPLORATORY DATA ANALYSES (EDA)

EDA involves generating summary descriptive statistics and various graphical visualisations to understand the data better[6].

Number of Convictions Data

The easiest way to get descriptive statistics and summary information for the entire data is to use the *summary()* function in R. The below figure shows its output:



```
> summary(num_con_df)
     Area              Homicide      Offence_against_person Sexual_offence
 Length:1386        Min.   : 0.000   Min.   :  29.0         Min.   :  0.00
 Class :character   1st Qu.: 0.000   1st Qu.: 112.0         1st Qu.:  7.00
 Mode  :character   Median : 1.000   Median : 172.0         Median : 14.00
                    Mean   : 1.799   Mean   : 233.1         Mean   : 21.81
                    3rd Qu.: 2.000   3rd Qu.: 266.0         3rd Qu.: 27.00
                    Max.   :38.000   Max.   :1827.0         Max.   :181.00
    Burglary         Robbery       Theft_handling   Fraud_forgery
 Min.   :  2.00   Min.   :  0.00   Min.   :  13.0   Min.   :  0.00
 1st Qu.: 14.25   1st Qu.:  3.00   1st Qu.: 104.0   1st Qu.:  7.00
 Median : 24.00   Median :  6.00   Median : 161.0   Median : 11.00
 Mean   : 33.07   Mean   : 10.78   Mean   : 213.2   Mean   : 19.36
 3rd Qu.: 39.00   3rd Qu.: 10.00   3rd Qu.: 258.8   3rd Qu.: 20.00
 Max.   :278.00   Max.   :209.00   Max.   :1426.0   Max.   :283.00

Criminal_damage      Drugs         Public_order       Other
Min.   :  3.00   Min.   :   8.0   Min.   :  2.00   Min.   :  0.0
1st Qu.: 27.00   1st Qu.:  40.0   1st Qu.: 43.00   1st Qu.: 11.0
Median : 41.50   Median :  66.0   Median : 66.00   Median : 20.0
Mean   : 53.46   Mean   : 101.6   Mean   : 88.83   Mean   : 41.4
3rd Qu.: 61.00   3rd Qu.: 105.0   3rd Qu.:100.00   3rd Qu.: 46.0
Max.   :400.00   Max.   :1228.0   Max.   :779.00   Max.   :551.0
Motor_offence    Admin_unsuccessful     Year               Month
Min.   :   1.0   Min.   :  0.00     Length:1386        Length:1386
1st Qu.:  96.0   1st Qu.:  7.00     Class :character   Class :character
Median : 145.0   Median : 12.00     Mode  :character   Mode  :character
Mean   : 197.1   Mean   : 19.01
3rd Qu.: 215.0   3rd Qu.: 19.00
Max.   :1889.0   Max.   :287.00
```

Fig 5. Descriptive statistics of number of convictions for each offence

Measures of central tendency are statistics that describe the Central or typical value of a dataset. In the above output we have two common measures of central tendency - the mean and median. The mean is the sum of all observations divided by the number of observations and the median is the middle value of a dataset when it is arranged in ascending order. While the mean is affected by presence of outliers, the median is not. In the case of a perfectly normal distribution, the value of the mean and the median would be same. We can see above that there is a fairly large difference between the mean and median value in many of the offence categories – indicating that the distribution of the data is skewed.

The minimum and maximum values as the name suggests identifies the highest and lowest value in the distribution. These values help us understand the range of our data. Quartiles are values that divide a dataset into four equal parts, or quarters. The first quartile (Q1) or the lower quartile, is the value that separates the lowest 25% of the observations from the rest of the dataset. The third quartile (Q3), also known as the upper quartile, is the value that separates the lowest 75% of the observations from the highest 25% of the observations. In our data, we can see that the 3rd quartile value and the max value are fairly different for many of the offence categories, indicating the presence of outliers and skewed data distribution.

Another measure of descriptive statistics that has not been included in the above output is the standard deviation. Calculated by computing the square root of the variance, it is a measure of how dispersed that data is. We can use the code '*lapply(df, sd)*'to produce the standard deviations of all the columns in our data. A low standard deviation means that most of the data points are closer to the mean, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The below table shows the mean and standard deviations of the different offence categories per month in the years 2014-16:

Table 1. Descriptive statistics (mean and standard deviation)

Offence Category	Mean	Standard Deviation
Homicide	1.79	3.2
Offence against the person	233.1	233.42
Sexual Offence	21.81	23.91
Burglary	33.07	35.49
Robbery	10.78	20.55
Theft and Handling	213.2	190.26
Fraud and Forgery	19.36	33.99
Criminal Damage	53.46	47.77
Drugs	101.6	158.08
Public Order	88.83	91.9
Other	41.4	66.88
Motor Offences	197.1	205.64
Admin Unsuccessful	19.01	32.85

Except for the case of homicide, standard deviation is rather high, inidcating that the data is widely dispersed.

By examining the means and max values, we can conclude that the top five offences with the largest number of convictions are:

1. Offences against the person

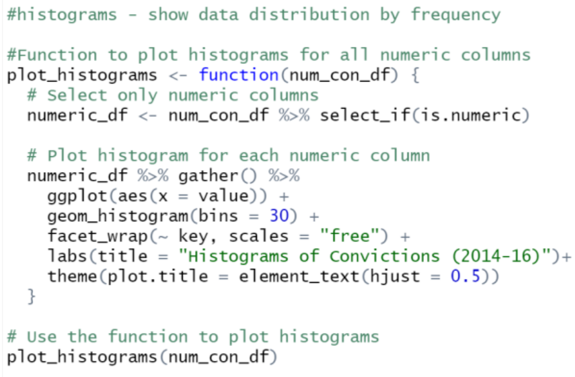
2. Motor offences

3. Theft and handling

4. Drug offences

5. Public order offences.

To visualise the frequency distribution of data for each offence, histograms can be used.



```
#histograms - show data distribution by frequency

#Function to plot histograms for all numeric columns
plot_histograms <- function(num_con_df) {
  # Select only numeric columns
  numeric_df <- num_con_df %>% select_if(is.numeric)

  # Plot histogram for each numeric column
  numeric_df %>% gather() %>%
    ggplot(aes(x = value)) +
    geom_histogram(bins = 30) +
    facet_wrap(~ key, scales = "free") +
    labs(title = "Histograms of Convictions (2014-16)")+
    theme(plot.title = element_text(hjust = 0.5))
  }

# Use the function to plot histograms
plot_histograms(num_con_df)
```

Fig 6. Code to plot histograms of all numerical columns in the data frame

In the above code, bins refer to the bars in the histogram, where each bar indicates a specific interval of data points[7]. Number of bins to use is a subjective choice, but since our data is large, 30 is a fair number of bins to specify. Facet wraps has been used to view individual variables in their own histograms. By specifying scales as free, the X and Y axis of histograms can use different scale ranges that are best suited for each variable.

While one can write code to generate individual histograms for each offence category which may provide a better visualisation, the advantage of the facet wrap method is that we can visualise all the columns at once. At this initial stage of data exploration, this might be more useful and comprehensive.

The visual output of the above code is shown below:

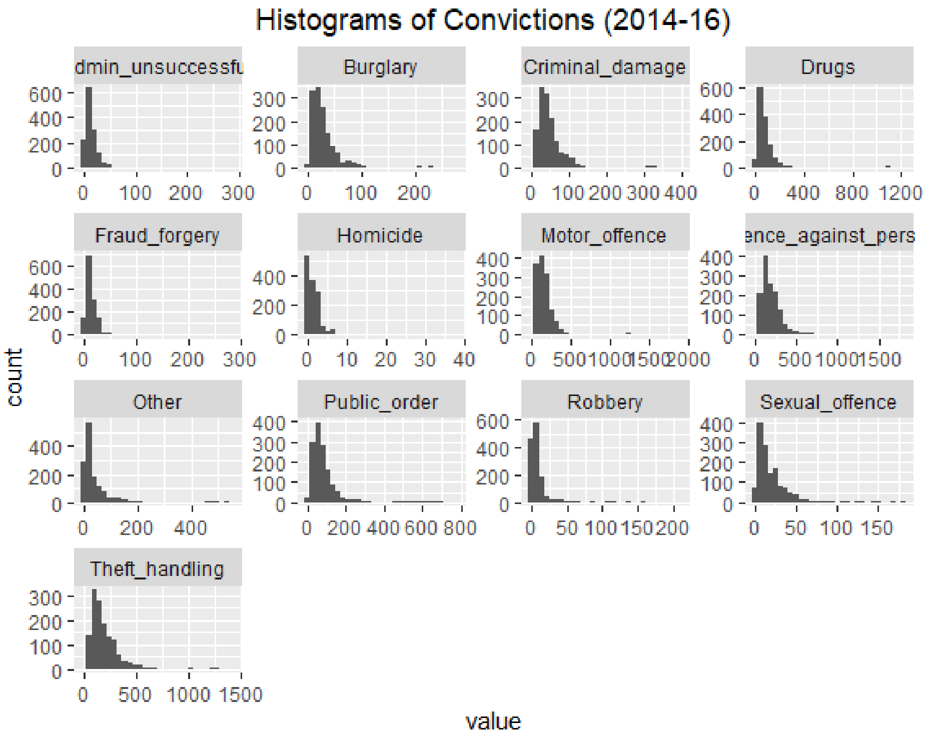
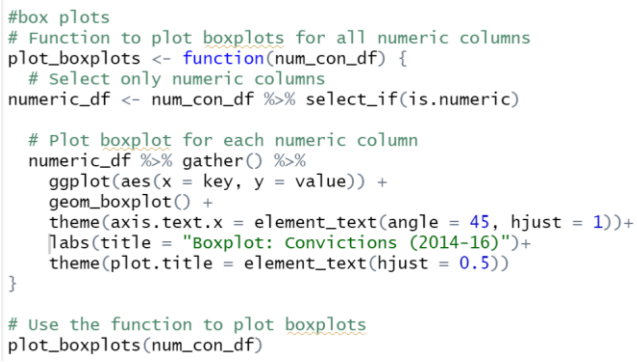


Fig 7. Frequency distribution of the data

We can see a mix of skewed and roughly normal distributions with some outliers. Comparing the Y-axis scale shows that offences like homicide are often under 10 convictions a month, while offences against the person, theft and handling cases have under 500 convictions a month.

Boxplots are also a very useful form of visualising data. The following code was used to make boxplots of the 12 offences:



```
#box plots
# Function to plot boxplots for all numeric columns
plot_boxplots <- function(num_con_df) {
  # Select only numeric columns
numeric_df <- num_con_df %>% select_if(is.numeric)

  # Plot boxplot for each numeric column
  numeric_df %>% gather() %>%
    ggplot(aes(x = key, y = value)) +
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    labs(title = "Boxplot: Convictions (2014-16)")+
    theme(plot.title = element_text(hjust = 0.5))
}

# Use the function to plot boxplots
plot_boxplots(num_con_df)
```

Fig 8. Function to plot boxplots of numeric columns in data

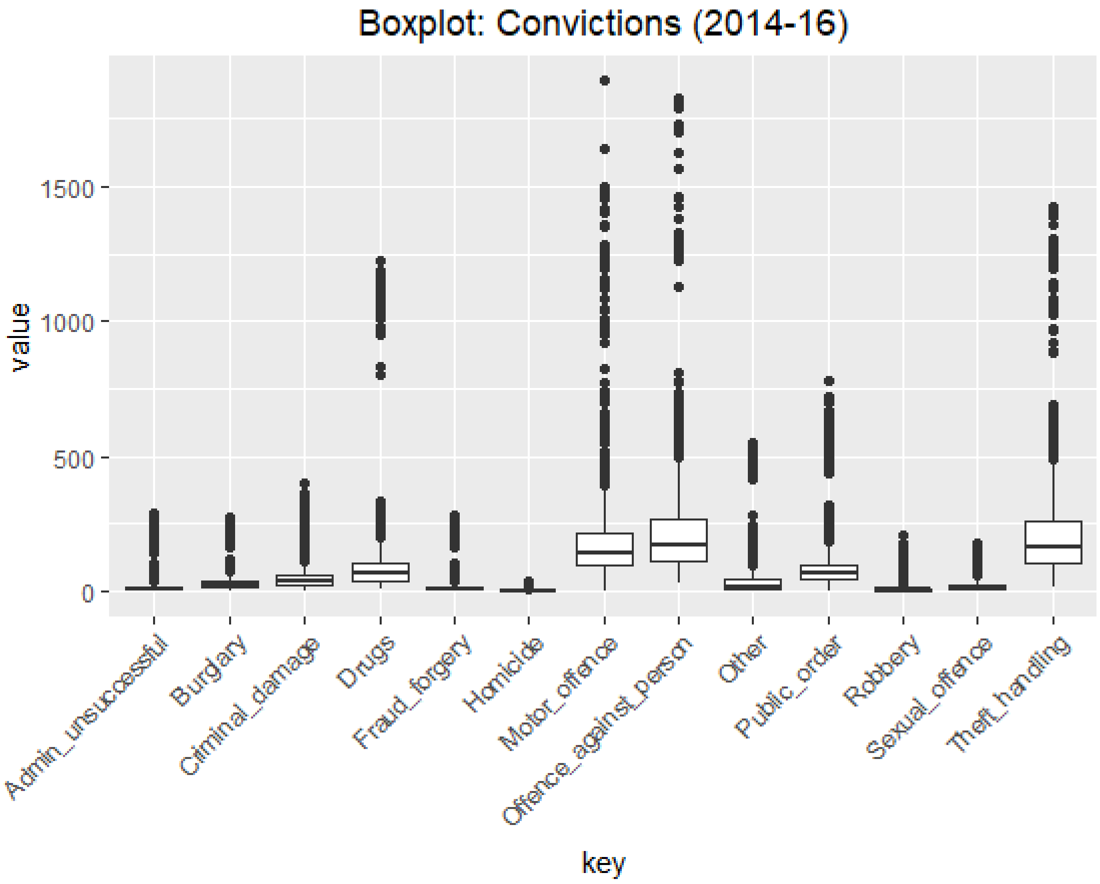


Fig 9. Boxplot of convictions of different offences and admin finalised unsuccessful cases

Box plots are very informative and visualise various descriptive statistics parameters, namely, the minimum score, first (lower) quartile, median, third (upper) quartile, maximum score and outliers[8]. The below image depicts how to interpret a box plot:

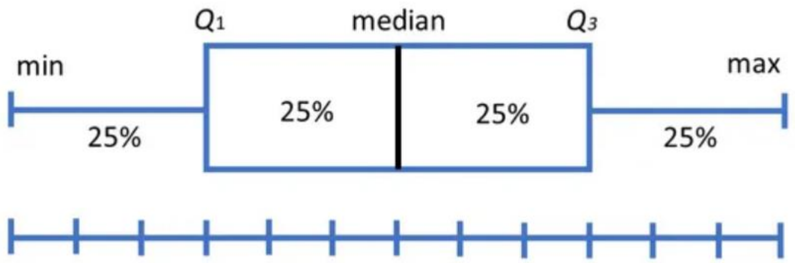


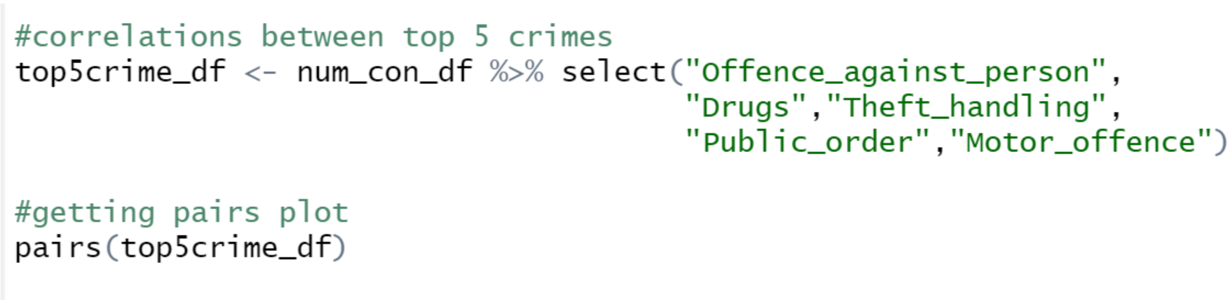
Fig 10. Interpreting a box plot

The 'whiskers', i.e., the lines at the ends of the boxes represent the highest and lowest halves of the data while the 'box' represents the middle half. The line shows the median – which gives us an idea of how skewed the data is (in the above figure the data is perfectly normal – which is rarely the case in real world data).

On observing fig 9, we can visually confirm what we saw in the summary descriptive statistics, i.e., categories of offences with larger number of convictions include offences against the person, motor offences, and theft and handling offences.

Next, it may be of interest to examine relationships between variables – is convictions in one type of offence related to another offence?

A scatterplot between variables will demonstrate this. To simplify visualisation, let's take into consideration the correlations between the top 5 crimes alone.



```
#correlations between top 5 crimes
top5crime_df <- num_con_df %>% select("Offence_against_person",
                                      "Drugs","Theft_handling",
                                      "Public_order","Motor_offence")

#getting pairs plot
pairs(top5crime_df)
```

Fig 11. Code to select specific offence categories and generate their scatterplots

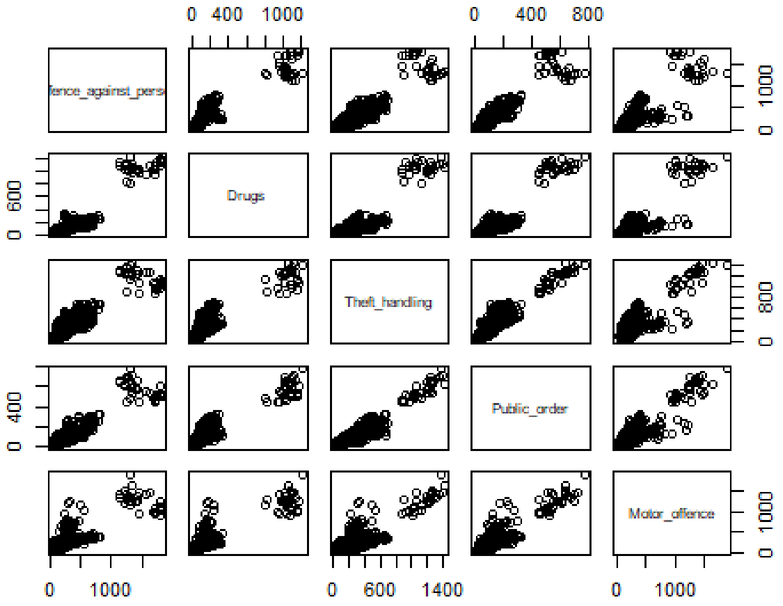
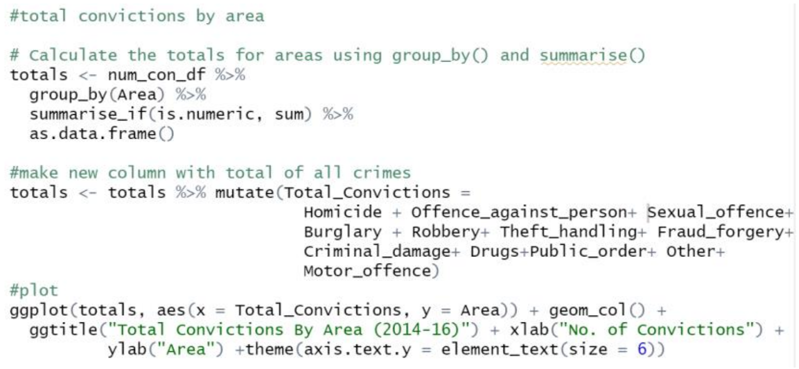


Fig 1. Pair plots between the top 5 offences with highest convictions

All the plots show a roughly straight line, indication a somewhat linear relationship between the variables. The data appearing as two clusters in the graph may possibly indicate that there are outliers in all the variables which are also correlating with each other.

Other relevant information present in the data include the areas or counties. It is of interest to us to know which areas have higher number of convictions.



```
#total convictions by area

# Calculate the totals for areas using group_by() and summarise()
totals <- num_con_df %>%
  group_by(Area) %>%
  summarise_if(is.numeric, sum) %>%
  as.data.frame()

#make new column with total of all crimes
totals <- totals %>% mutate(Total_Convictions =
                              Homicide + Offence_against_person+ Sexual_offence+
                              Burglary + Robbery+ Theft_handling+ Fraud_forgery+
                              Criminal_damage+ Drugs+Public_order+ Other+
                              Motor_offence)
#plot
ggplot(totals, aes(x = Total_Convictions, y = Area)) + geom_col() +
  ggtitle("Total Convictions By Area (2014-16)") + xlab("No. of Convictions") +
          ylab("Area") +theme(axis.text.y = element_text(size = 6))
```

Fig 13. Computing total convictions by area and creating its visualisation

The above code groups the data by area, and then adds up the number of convictions in each area. Therefore, in the new totals data, each row represents the total number of convictions in each county in the UK from 2014-2016 (excluding 3 months of missing data). We then add a column totalling the number of convictions for all the offence categories and plot those values for each county using the *ggplot* package.

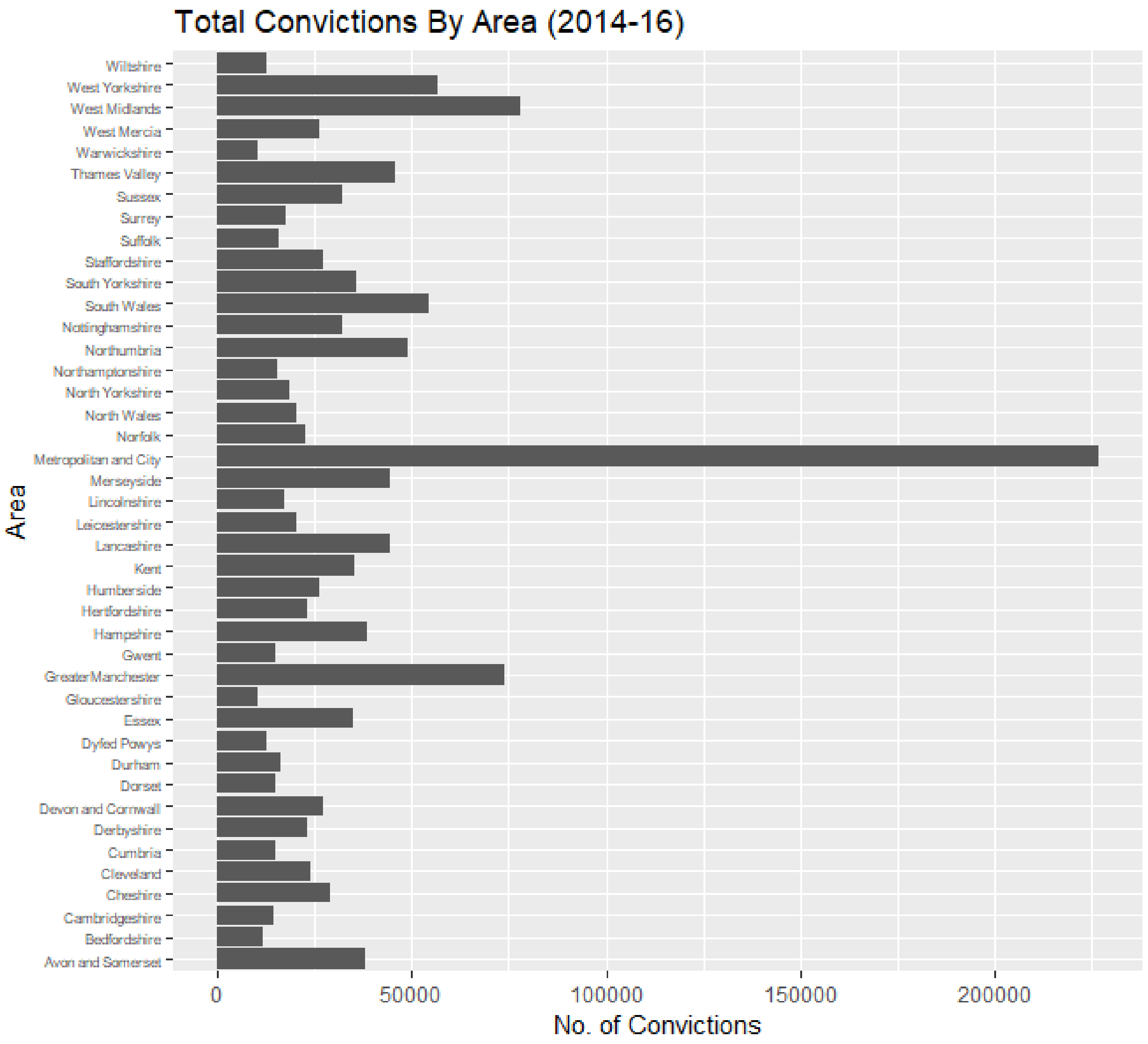


Fig 14. Total Number of Convictions by Area

This graph reveals some important information about the data. The number of convictions in Metropolitan and City area appears to be significantly higher than the other counties. Much of the outliers we see in our boxplots and histograms may actually belong to this region too.

Based on the above graph, the top 5 ceremonial counties with the maximum number of convictions are:

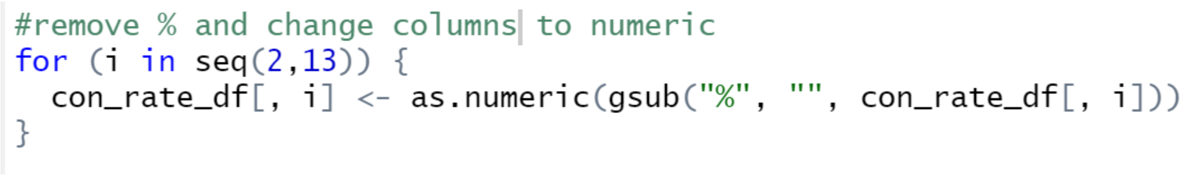
1. Metropolitan and City
2. West Midlands
3. Greater Manchester
4. West Yorkshire
5. South Wales

Higher number of convictions indicate higher number of crimes taking place, so we may conclude that these might be the least safe counties in the UK.

Conviction rate data

Now let's look at the conviction rates data. How might it be different from the number of convictions data?

The original conviction rate data is not numeric, but in character form, and every entry has the special character "%". In order to perform analyses, we need to format the data first:

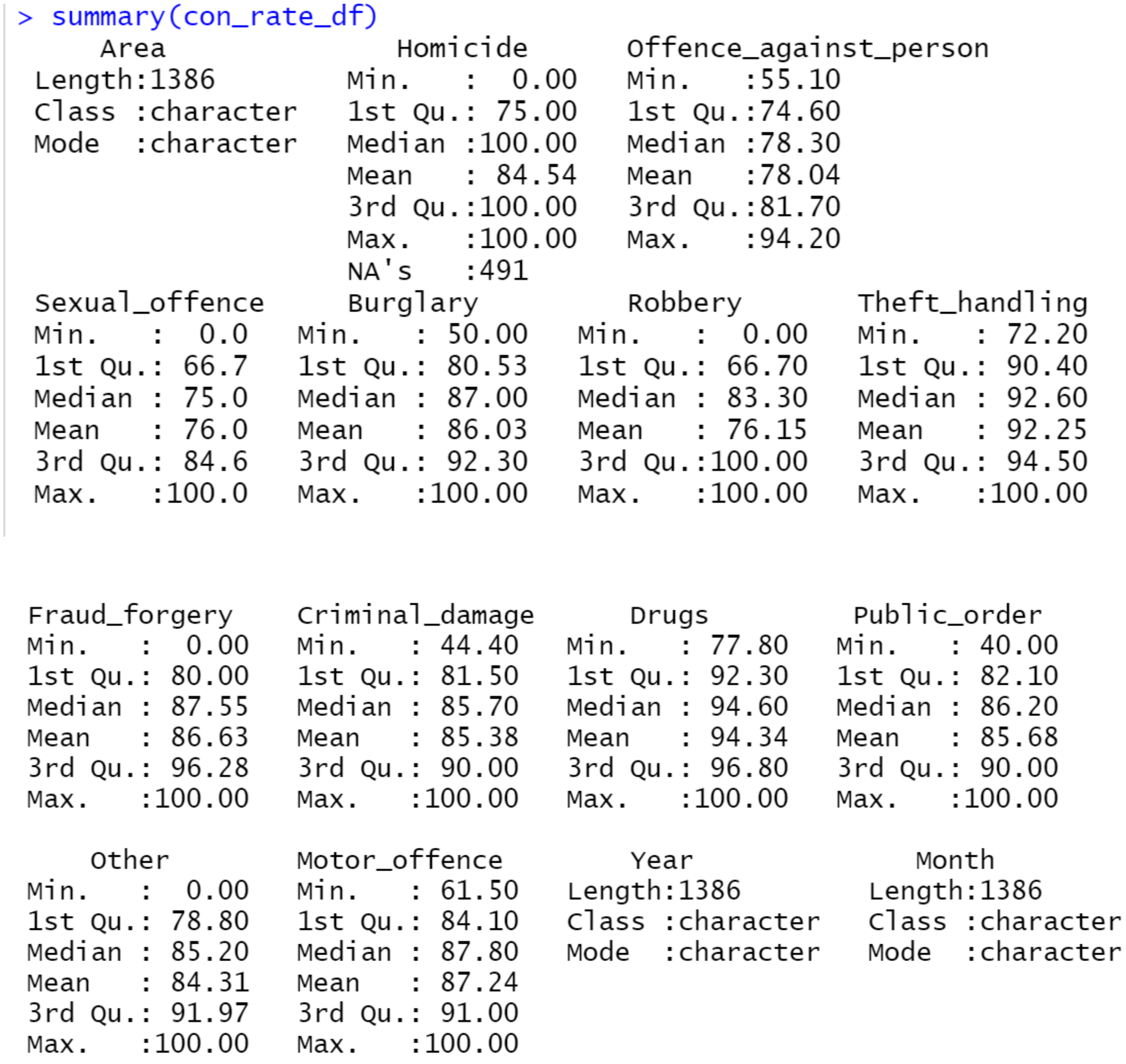


```
#remove % and change columns to numeric
for (i in seq(2,13)) {
  con_rate_df[, i] <- as.numeric(gsub("%", "", con_rate_df[, i]))
}
```

Fig 15. Data Formatting: Removing "%" and changing data type

In the above code, a for loop has been used to iterate the code through columns 2 to 13 in the data frame.

Using code similar to what we previously used for the number of convictions data, we can get descriptive statistics, histograms and boxplots for our data.



```
> summary(con_rate_df)
     Area              Homicide       Offence_against_person
 Length:1386        Min.   :  0.00   Min.   :55.10         
 Class :character   1st Qu.: 75.00   1st Qu.:74.60         
 Mode  :character   Median :100.00   Median :78.30         
                    Mean   : 84.54   Mean   :78.04         
                    3rd Qu.:100.00   3rd Qu.:81.70         
                    Max.   :100.00   Max.   :94.20         
                    NA's   :491                            
 Sexual_offence     Burglary         Robbery        Theft_handling 
 Min.   :  0.0   Min.   : 50.00   Min.   :  0.00   Min.   : 72.20  
 1st Qu.: 66.7   1st Qu.: 80.53   1st Qu.: 66.70   1st Qu.: 90.40  
 Median : 75.0   Median : 87.00   Median : 83.30   Median : 92.60  
 Mean   : 76.0   Mean   : 86.03   Mean   : 76.15   Mean   : 92.25  
 3rd Qu.: 84.6   3rd Qu.: 92.30   3rd Qu.:100.00   3rd Qu.: 94.50  
 Max.   :100.0   Max.   :100.00   Max.   :100.00   Max.   :100.00  

 Fraud_forgery    Criminal_damage      Drugs         Public_order   
 Min.   :  0.00   Min.   : 44.40   Min.   : 77.80   Min.   : 40.00  
 1st Qu.: 80.00   1st Qu.: 81.50   1st Qu.: 92.30   1st Qu.: 82.10  
 Median : 87.55   Median : 85.70   Median : 94.60   Median : 86.20  
 Mean   : 86.63   Mean   : 85.38   Mean   : 94.34   Mean   : 85.68  
 3rd Qu.: 96.28   3rd Qu.: 90.00   3rd Qu.: 96.80   3rd Qu.: 90.00  
 Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :100.00  

     Other        Motor_offence        Year              Month          
 Min.   :  0.00   Min.   : 61.50   Length:1386        Length:1386       
 1st Qu.: 78.80   1st Qu.: 84.10   Class :character   Class :character  
 Median : 85.20   Median : 87.80   Mode  :character   Mode  :character  
 Mean   : 84.31   Mean   : 87.24                                        
 3rd Qu.: 91.97   3rd Qu.: 91.00                                        
 Max.   :100.00   Max.   :100.00                                        
```

Fig 16. Descriptive Statistics: Conviction rates (2014-16)

Examining the minimum and mean values, we can identify that the top 3 offences with the highest conviction rates are:

1. Theft and handling offences
2. Drug offences
3. Motor offences

The crimes with the lowest conviction rates are:

1. Sexual offences
2. Robbery
3. Offences against the person

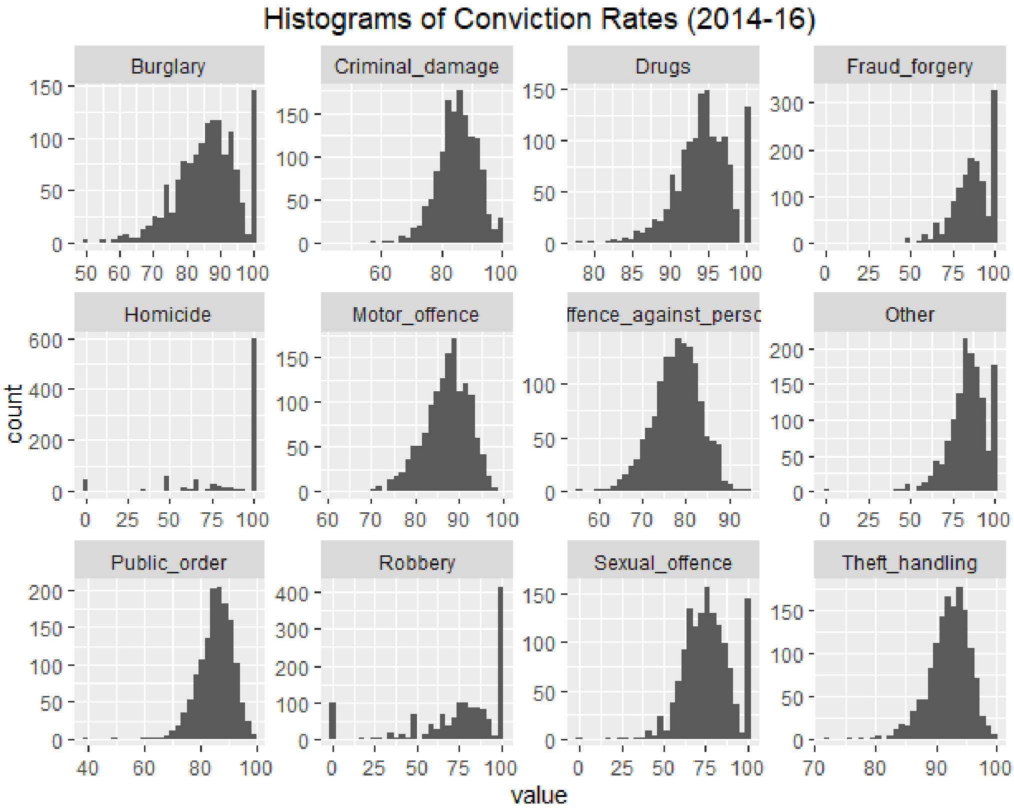


Fig 17. Frequency distribution of the data

We can see that most histograms skewed towards the right -indicating that conviction rates of the CPS are typically on the higher side.

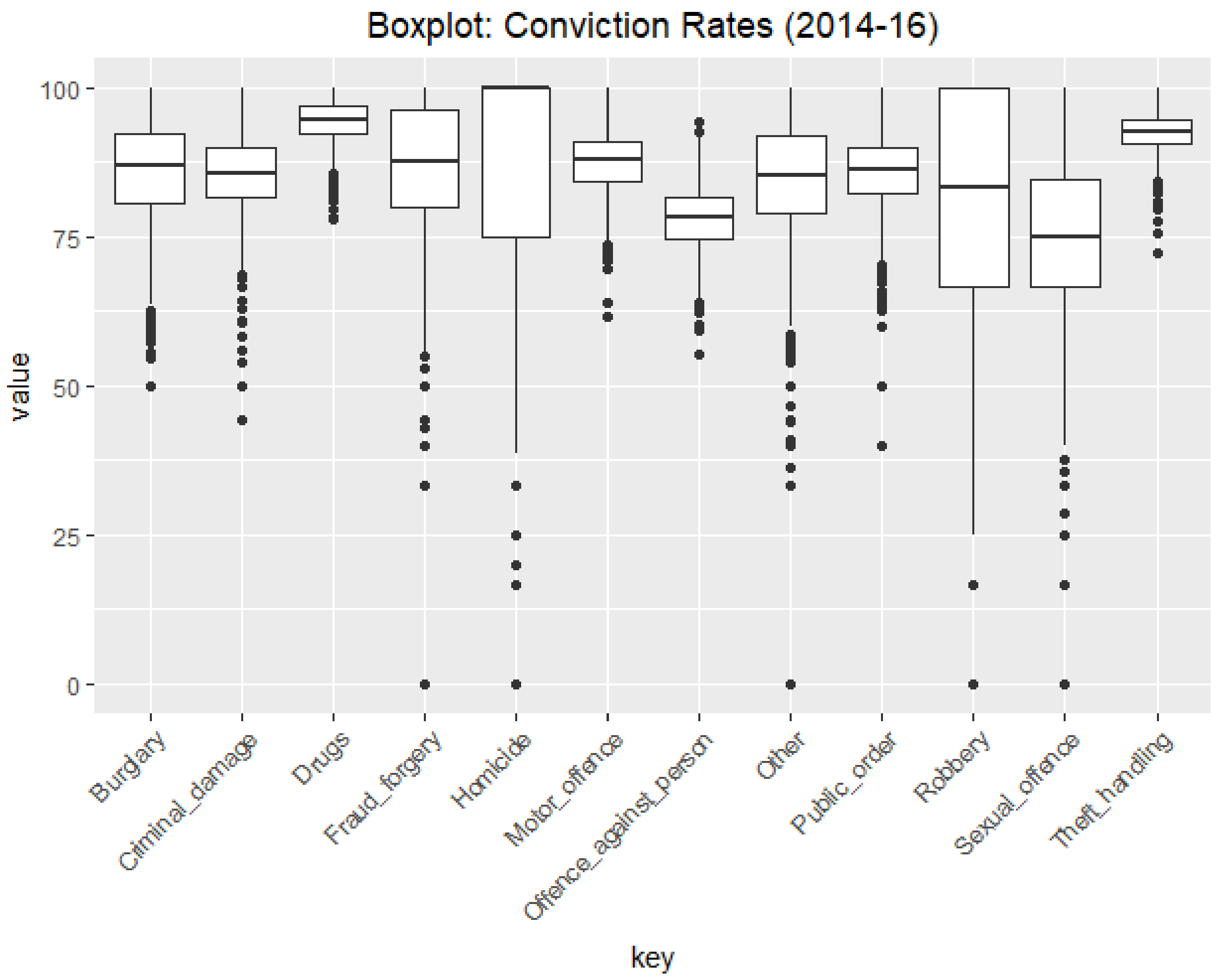
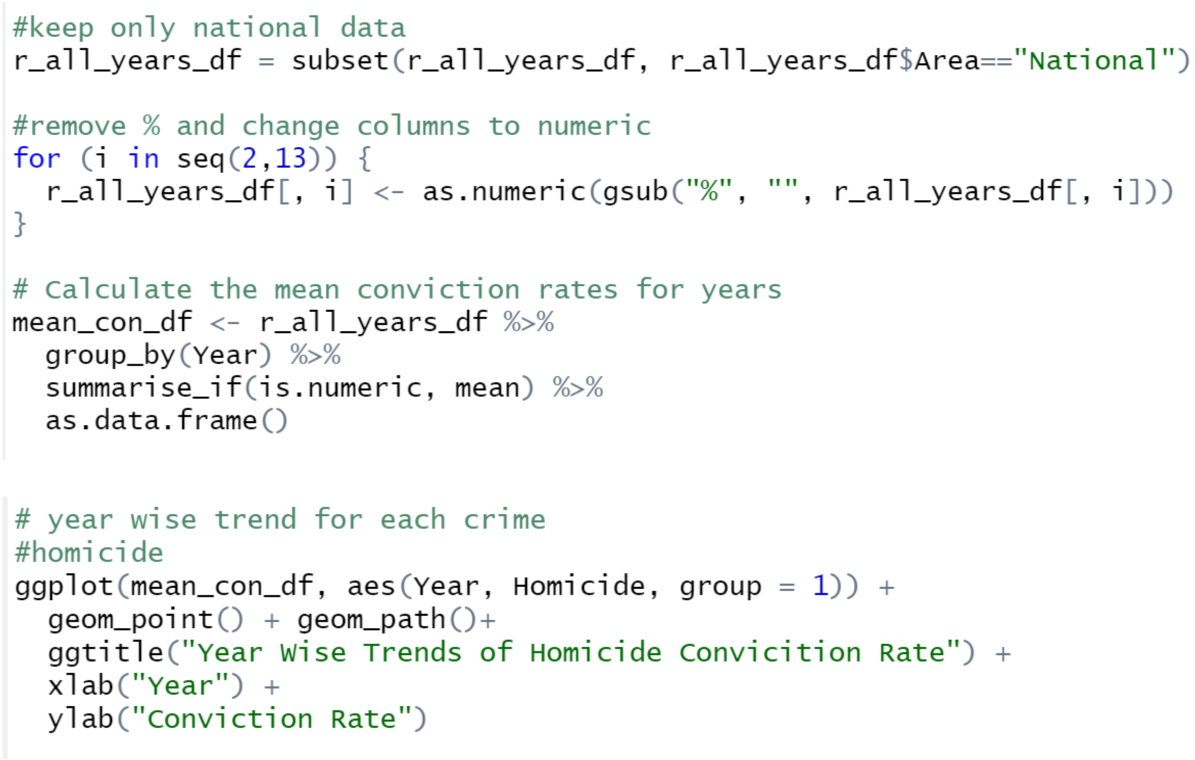


Fig 18: Boxplot of conviction rates of different offence categories

The above graph validates our conclusions from the descriptive statistics that sexual offences, robbery and offences against the person have some of the least conviction rates, while motor offences, drug related offences and theft and handling offences have highest conviction rates.

Another question we may want to answer is whether the conviction rates for each offences have improved or declined over the years?

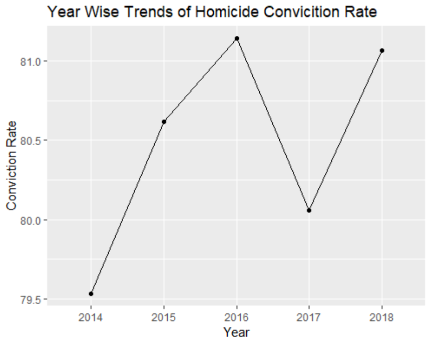
The below code was used to extract national conviction rates, calculate their year wise means and plot the year wise trends (for sake of brevity, only code for one column is displayed below).

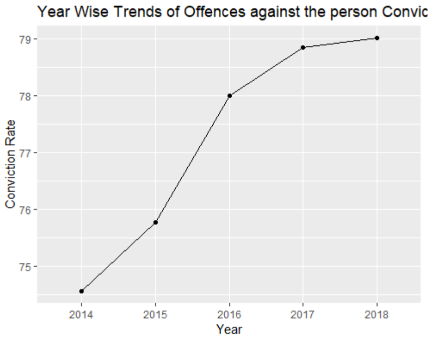


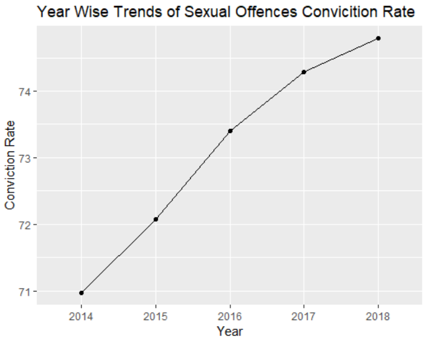
```
#keep only national data
r_all_years_df = subset(r_all_years_df, r_all_years_df$Area=="National")

#remove % and change columns to numeric
for (i in seq(2,13)) {
  r_all_years_df[, i] <- as.numeric(gsub("%", "", r_all_years_df[, i]))
}

# Calculate the mean conviction rates for years
mean_con_df <- r_all_years_df %>%
  group_by(Year) %>%
  summarise_if(is.numeric, mean) %>%
  as.data.frame()
```

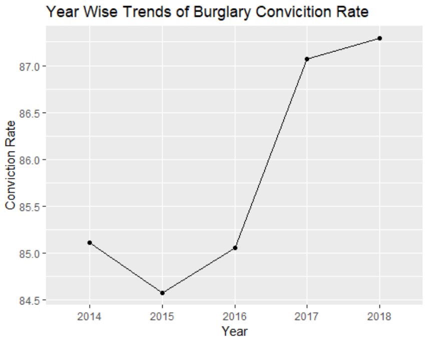
```
# year wise trend for each crime
#homicide
ggplot(mean_con_df, aes(Year, Homicide, group = 1)) +
  geom_point() + geom_path()+
  ggtitle("Year Wise Trends of Homicide Convicition Rate") +
  xlab("Year") +
  ylab("Conviction Rate")
```

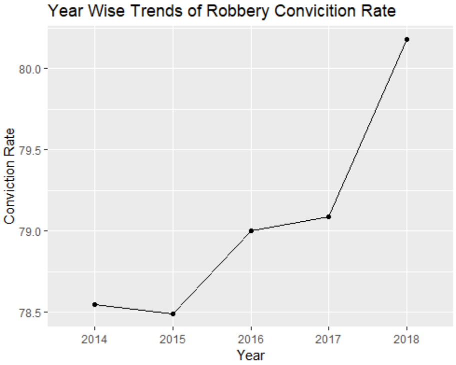
Fig 19. Sample code for line plots to visualise year wise trends

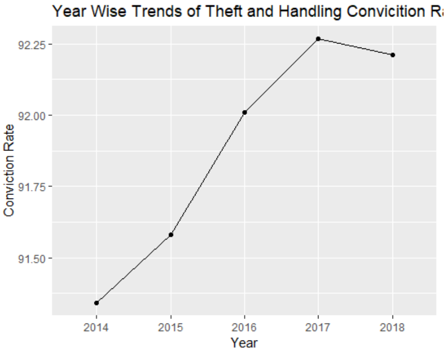

Year Wise Trends of Homicide Convicition Rate
81.0
80.5
80.0
79.5
Conviction Rate
2014
2015
2016
2017
2018
Year

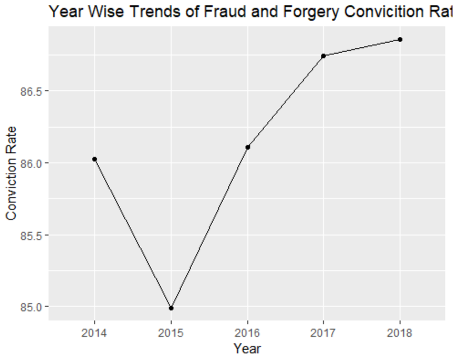

Year Wise Trends of Offences against the person Convic
79
78
77
76
75
Conviction Rate
2014
2015
2016
2017
2018
Year

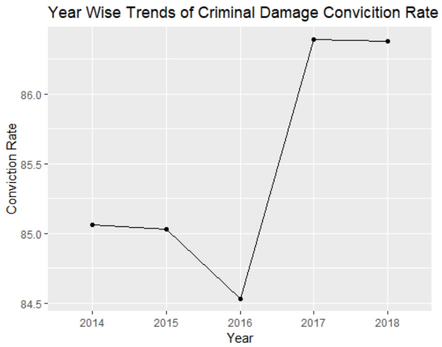

Year Wise Trends of Sexual Offences Convicition Rate
74
73
72
71
Conviction Rate
2014
2015
2016
2017
2018
Year


Year Wise Trends of Burglary Convicition Rate
87.0
86.5
86.0
85.5
85.0
84.5
Conviction Rate
2014
2015
2016
2017
2018
Year


Year Wise Trends of Robbery Convicition Rate
80.0
79.5
79.0
78.5
Conviction Rate
2014
2015
2016
2017
2018
Year


Year Wise Trends of Theft and Handling Convicition R
92.25
92.00
91.75
91.50
Conviction Rate
2014
2015
2016
2017
2018
Year


Year Wise Trends of Fraud and Forgery Convicition Rat
86.5
86.0
85.5
85.0
Conviction Rate
2014
2015
2016
2017
2018
Year


Year Wise Trends of Criminal Damage Convicition Rate
86.0
85.5
85.0
84.5
Conviction Rate
2014
2015
2016
2017
2018
Year

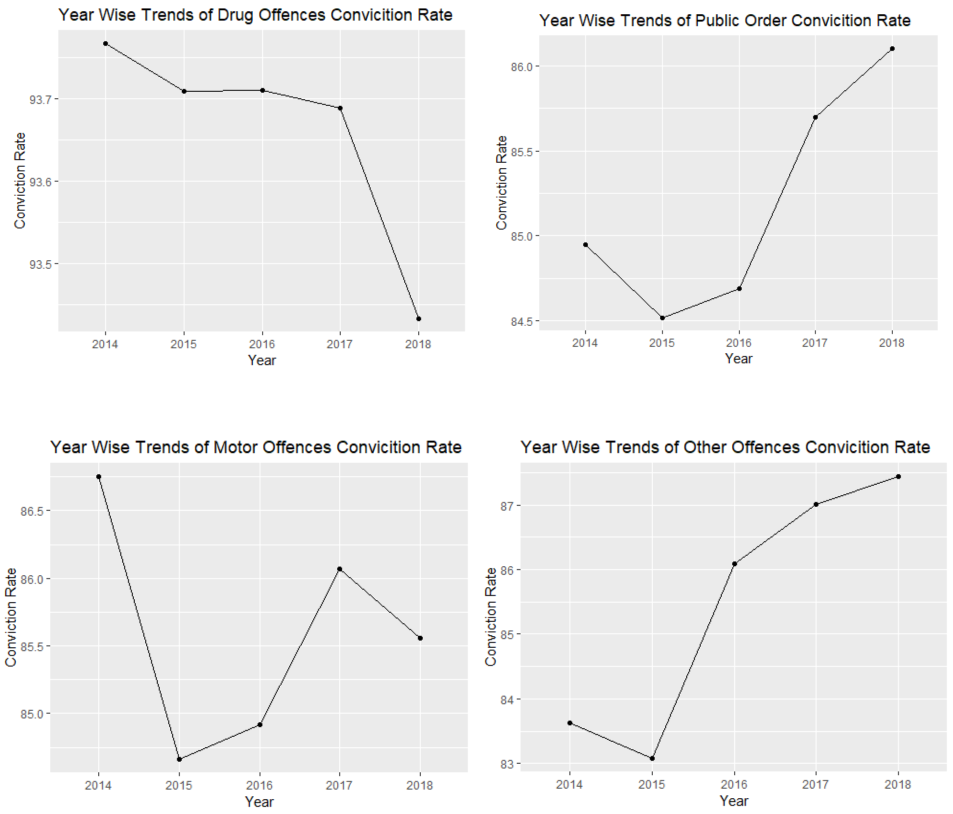


Fig 20. Year wise conviction rates of 12 offence categories

The above graphs show that only sexual offences and offences against the person have had a consistent upward trend in conviction rates over the years. Some offences like robbery have an overall upward trend, and drug offences have an overall downward trend. The other offence categories have ups and downs across the years as seen in the images.

IV. ARE THERE ANY PATTERNS AMONG THE NUMBER OF CONVICTIONS ACROSS DIFFERENT OFFENCE CATEGORIES IN OUR DATA?

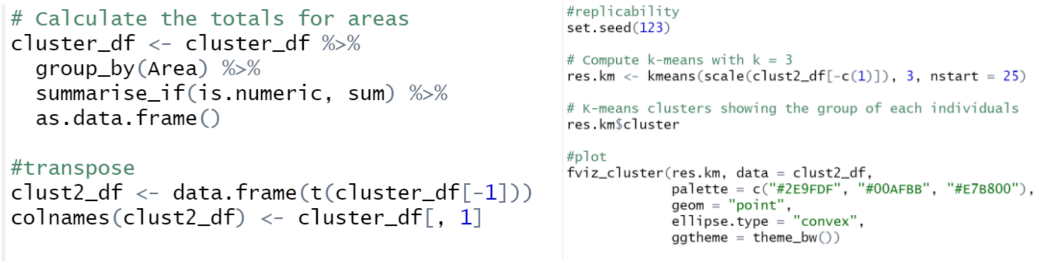
Clustering is a data science technique where the goal is to divide data points into groups or clusters such that the data within each cluster is similar to each other[11]. Clustering is an unsupervised learning technique. This means that it does not require labelled data, i.e., data that already has the answer or the category membership of the data points. Therefore, it is the technique to use when we are looking to identify trends without any expectations or priori knowledge of what the patterns may be.

The K-means clustering model is one of the most common clustering algorithms, that works by dividing the data into *k* clusters, where *k* is the number of clusters specified by the analyst. The algorithm choses *k* points to serve as cluster centers and then iteratively assigns each data point to the cluster with the closest centroid, which is the mean of all the data points currently in the cluster[11].

Data manipulation – pre-processing for cluster analysis

In order to find clusters among the number of convictions across different offence categories, we need to first manipulate our data and make it suitable for clustering analysis. As we have seen in the exploratory analysis section, our initial data frame simply merges the data of each month available, resulting in multiple rows having the same area or county name. To make our analysis comprehensible, we will group by area and calculate the total convictions of each offence category for all the months in the data – which will be saved as a data frame. Since our hypothesis is to identify patters within convictions in different offence categories, we transpose the rows and columns such that each offence category becomes a row, since clusters are identified between rows in the data frame, not columns.

We can then run the clustering algorithm and plot its graph.



```
# Calculate the totals for areas
cluster_df <- cluster_df %>%
  group_by(Area) %>%
  summarise_if(is.numeric, sum) %>%
  as.data.frame()

#transpose
clust2_df <- data.frame(t(cluster_df[-1]))
colnames(clust2_df) <- cluster_df[, 1]
```

```
#replicability
set.seed(123)

# Compute k-means with k = 3
res.km <- kmeans(scale(clust2_df[-c(1)]), 3, nstart = 25)

# K-means clusters showing the group of each individuals
res.km$cluster

#plot
fviz_cluster(res.km, data = clust2_df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw())
```

Fig 25. R code for K means clustering

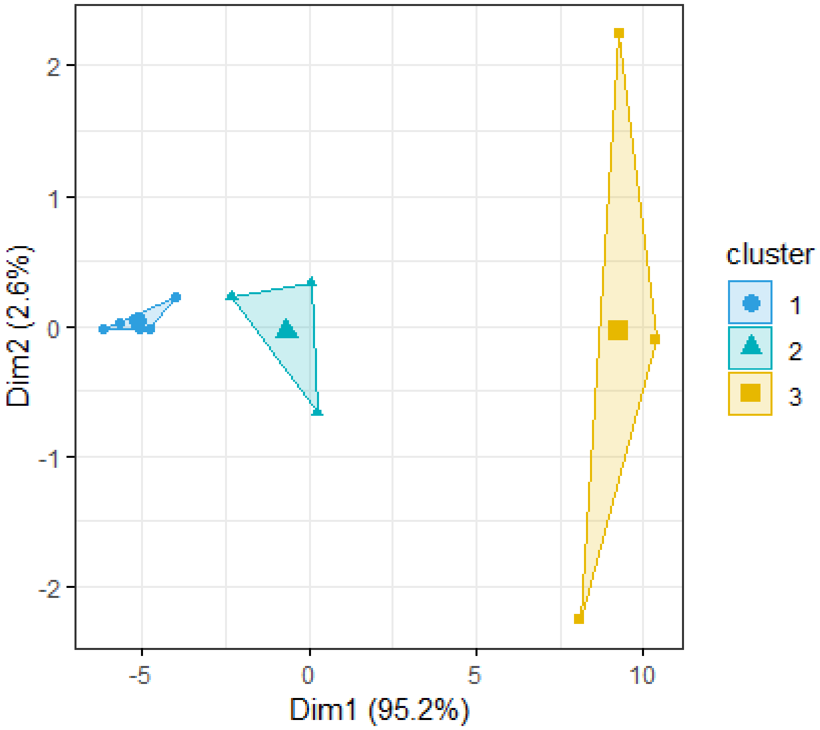


Fig 26. K-means cluster plot

Table 2. Clusters of Number of convictions of offence categories (from k-means model)

Cluster 1	Cluster 2	Cluster 3
Homicide	Public Order	Offences against the person
Burglary	Drug related offences	Motor Offences
Fraud and Forgery	Criminal Damage	Theft and Handling
Robbery		
Sexual Offences		

Now we interpret the clusters with domain knowledge. While a cursory look may not reveal any obvious similarities within elements of a cluster, a deeper gauging reveals a possible commonality. In the offences in cluster one, either the likelihood of someone reporting the crime is less (for example, sexual offences and fraud cases are often less reported due to many reasons including social stigma[18,19]), or the likelihood of an eyewitness being able to identify the perpetuator is less (Burglary, Homicide, Robbery) possibly because the perpetuator is masked or because of fear. It is relatively easier to identify the perpetuator in the offences in cluster two, and even easier in cluster 3 (you often know your assailant, the vehicle number plate, or if you identify that a product is stolen it is easy to investigate the person selling the stolen goods).

Therefore, if we had to name the clusters it could be:

1. Offences with most unidentifiable perpetuators
2. Offences with moderately identifiable perpetuators
3. Offences with easily identifiable perpetuators

Logically this would make sense since our data represents the number of convictions in each category, and it is likely that convictions are more successful in crimes where the perpetuator is easily identifiable. However, it is worth verifying with industry experts and other resources to have stronger confidence in the accuracy of the clusters and their interpretation.

V. CRITICAL REVIEW OF CLUSTERING MODELS

The main strength of K-means clustering is that it is easy to understand and implement, and is fairly efficient in its performance[20]. Another strength of K-means is that it is able to find globular or spherical clusters[20], which can be useful in cases where the clusters have a clear center, and the data points within each cluster are similar to each other.

However, K-means has several weaknesses as well. One issue with the K-means algorithm is that it is sensitive to the initialization of the cluster centroids[21]. This means that the algorithm can result in different solutions depending on the starting point of the cluster centroids. An alternative algorithm to overcome this issue is discussed below. Additionally, the k-means model requires a reasonable guess as to how many clusters naturally exist in the data which might not necessarily be accurate[21].

Another criticism of the model we have used is that our data had a large number of dimensions – 42 columns to be exact. High dimensional data may bear the "curse of dimensionality". This refers to the fact that as the dimensionality of the data increases, the volume of the space increases vastly, and the data points become sparse[22]. This makes can cause the clusters to overlap with each other, rather than being distinct. Moreover, higher the number of dimensions, higher the number of outliers, which can affect the clusters. One way to deal with this is reducing the dimensions of our data using principal component analysis. In our data set, an option would have been to simply use the national data instead of the data for 42 counties, or add up the number of convictions for different regions in the UK (midlands, south east etc.) to reduce the number of dimensions.

Moreover, K-means is sensitive to the presence of outliers[23]. This is because the value of means is affected by outliers. A recent study examined the robustness of a variant - k-medoids clustering to overcome this limitation by using medians[23]. A direction for future analyses may be to examine whether the rates of conviction data form similar clusters as the number of convictions data.

VI. KEY FINDINGS AND CONCLUSIONS FROM THE DATA

- Offences with the highest conviction rates are Theft and handling offences, Drug offences and Motor offences.
- Offences with the lowest conviction rates are: Sexual offences, Robbery, Offences against the person
- The conviction rate for sexual offences and offences against the person have had a consistent upward trend over the years.
- Offences that are less likely to be reported and where perpetuators are less likely to be identified have the least convictions (e.g. Sexual offences, homicide) while offences where it is easy to identify the perpetator (e.g. motor offences) have the highest number of convictions.

References

[1]. The Crown Prosecution Service (2022). About CPS. Available at: https://www.cps.gov.uk/ (Accessed: 5 January 2023).

[2]. The Crown Prosecution Service (2022). CPS data summary Quarter 1 2022-2023. Available at: https://www.cps.gov.uk/publication/cps-data-summary-quarter-1-2022-2023 (Accessed: 5 January 2023).

[3]. The Crown Prosecution Service (2022). Offences against the Person, incorporating the Charging Standard. Available at: https://www.cps.gov.uk/legal-guidance/offences-against-person-incorporating-charging-standard (Accessed: 5 January 2023).

[4]. The Crown Prosecution Service (2022). Criminal Damage. Available at: https://www.cps.gov.uk/legal-guidance/criminal-damage (Accessed: 5 January 2023).

[5]. Sentencing Council (n.d.). Drug Offences. Available at: https://www.sentencingcouncil.org.uk/outlines/drug-offences/ (Accessed: 5 January 2023).

[6]. Lam, K.S. (2022). Exploratory Data Analysis Project Using Python. Available at: https://medium.com/@lamsampathkumar0/eda-exploratory-data-analysis-project-using-python-de90cbf4e128 (Accessed: 5 January 2023).

[7]. Stats4Stem. R- Histogram. Available at: https://www.stats4stem.org/r-histogram (Accessed: 6 January 2023).

[8]. McLeod, S. (2019). What does a boxplot tell you? Available at: https://www.simplypsychology.org/boxplots.html. (Accessed: 6 January 2023).

[9]. Gov.UK. (2022). From harm to hope: A 10-year drugs plan to cut crime and save lives. Available at: https://www.gov.uk/government/publications/from-harm-to-hope-a-10-year-drugs-plan-to-cut-crime-and-save-lives/from-harm-to-hope-a-10-year-drugs-plan-to-cut-crime-and-save-lives (Accessed: 6 January 2023).

[10]. Nolan III, J.J., 2004. Establishing the statistical relationship between population size and UCR crime rate: Its impact and implications. *Journal of Criminal Justice*, *32*(6), pp.547-555.

[11]. Lantz, B., 2019. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

[12]. Jaccard, J., Wan, C.K. and Turrisi, R., 1990. The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate behavioral research*, *25*(4), pp.467-478.

[13]. Open Genus IQ: Computing Expertise and Legacy. (n.d). Advantages and Disadvantages of Linear Regression. Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/. (Accessed 10 January 2023)

[14]. Schmidt, A.F. and Finan, C., 2018. Linear regression and the normality assumption. *Journal of clinical epidemiology*, *98*, pp.146-151

[15]. John, G.H., 1995, August. Robust Decision Trees: Removing Outliers from Databases. In *KDD* (Vol. 95, pp. 174-179).

[16]. S. Chowdhury, Y. Lin, B. Liaw and L. Kerby, "Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance," 2022

International Conference on Intelligent Data Science Technologies and Applications (IDSTA), San Antonio, TX, USA, 2022, pp. 17-25, doi: 10.1109/IDSTA55301.2022.9923169.

[17]. Frost, J. (n.d.). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ (Accessed 10 January 2023).

[18]. Kimble, C. & Chettiar, I. (2018). Sexual Assault Remains Dramatically Underreported. Available at: https://www.brennancenter.org/our-work/analysis-opinion/sexual-assault-remains-dramatically-underreported. (Accessed 11 January 2023).

[19]. SWNS & Ballinger, A. (2017). Fraud Victims are too embarrassed to tell their partners, research has found. Available at: https://www.bristolpost.co.uk/news/uk-world-news/fraud-victims-embarrassed-tell-partners-687340 (Accessed 11 January 2023).

[20]. Kaushik, M. and Mathur, B., 2014. Comparative study of K-means and hierarchical clustering techniques. *International Journal of Software & Hardware Research in Engineering*, *2*(6), pp.93-98.

[21]. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J., 2022. K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*.

[22]. Pantola, P. (2018). The curse of dimensionality. Available at: https://medium.com/@paritosh_30025/curse-of-dimensionality-f4edb3efa6ec. Accessed (13 January 2023).

[23]. Lin, X., Han, J. (2011). *K*-Medoids Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_426.