**Table of Content**

1.Introduction……………………………………………………………………………………………………………………………5

1.1 Background………………………………………………………………………………………………………………….………5

1.2 Objectives…………………………………………………………………………………………………………………..………..5

2.Data Integration and Cleaning……………………………………………………………………………………………….…5

2.1 Data Download and Integration …………………………………………………………………………………………..5

2.2 Data Cleaning and Exploration………………………………………………………………………………………………7

2.2.1 Initial data exploration……………………………………………………………………………………………….………7

2.2.2 Changing data types…………………………………………………………………………………………………….…… 8

2.2.3 Changing the column names………………………………………………………………………………………………8

2.2.4 Inspect for mislabeled column names…………………………………………………………..……………………8

2.2.5 Fixing string inconsistencies ………………………………………………………………………………………………8

2.2.6 Missing Values……………………………………………………………………………………………………………………9

2.2.7 Re-organize Columns and Data Frames………………………………………………….…………………………13

3 Descriptive Analysis………………………………………………………………………………………………………………14

3.1 Summary Statistics………………………………………………………………………………………………………..……15

3.1.1 Summary Statistics based on Conviction type (Successful) ………………………………………………15

3.1.2 Summary Statistics based on Conviction type (Unsuccessful) …………………………….……………15

3.1.3 Summary Statistics based on Area (Successful) …………………………………………………………….…16

3.1.4 Summary Statistics based on Area (Unsuccessful) …………………………………………………..………17

3.2 Distribution of observations ………………………………………………………………………………………….……18

3.3 Overall Conviction Categories Across Years…………………………………………….……………………………19

3.4 Homicide ……………………………………………………………………………………………………………………………20

3.4.1 Distribution of Homicide Convictions vs Average all other offences over time…………..………20

3.4.2 Distribution of Homicide Convictions Across the Area…………………….………………….…………….22

3.5 Offences Against Person Convictions…………………………………………………………………………………..24

3.5.1 Distribution of Offences Against Person Convictions ………………………………………………..……..24

3.5.2 Distribution of Offences Against Person Across the Area………………………………………….………25

3.6 Sexual Offences …………………………………………………………………………………………………………….……27

3.6.1 Distribution of Sexual Offences vs Average all other offences over time…………………………..27

3.6.2 Distribution of Sexual Offences Across the Area…………………………………………………………..….28

3.7 Burglary………………………………………………………………………………………………………………………..……29

3.7.1 Distribution of Burglary vs Average all other offences over time…………………….……….….…….29

3.7.2 Distribution of Burglary Across the Area…………………………………………………………….…………….30

3.8 Robbery……………………………………………………………………………………………………………………………..32

3.8.1 Distribution of Robbery vs Average all other offences over time……………………….……………..32

3.8.2 Distribution of Robbery Across the Area…………………………………………………………………..………33

3.9 Theft And Handling……………………………………………………………………………………………………………..35

3.9.1 Distribution of Theft And Handling vs Average all other offences over time……………….……..35

3.9.2 Distribution of Theft And Handling Across the Area…………………………………………………….…..36

3.10 Fraud And Forgery…………………………………………………………………………………………………………….37

3.10.1 Distribution of Fraud And Forgery vs Average all other offences over time……….……………37

3.10.2 Distribution of Fraud And Forgery Across the Area……………………………..…………………………..39

3.11 Criminal Damage ……………………………………………………………………………………………………..………40

3.11.1 Distribution of Criminal Damage vs Average all other offences over time…………………..….40

3.11.2 Distribution of Criminal Damage Across the Area……………………………………………………….….41

3.12 Drugs Offences………………………………………………………………………………………………………….………42

3.12.1 Distribution of Drugs Offences vs Average all other offences over time……………….…………..42

3.12.2 Distribution of Drugs Offences Across the Area………………………………………………….………….44

3.13 Public Order Offences………………………………………………………………………………………………….……45

3.13.1 Distribution of Public Order Offences vs Average all other offences over time…..…………..45

3.13.2 Distribution of Public Order Offences Across the Area……………………………………….…………..46

3.14 Motoring Offences………………………………………………………………………………………………………..….48

3.14.1 Distribution of Motoring Offences vs Average all other offences over time………….………..48

3.14.2 Distribution of Motoring Offences Across the Area………………………………………………..………50

3.15 Total Criminal Offence Convictions by Area……………………………………………………..51

4 Hypothesis …………………………………………………………………………………………………………………..……….52

4.1 Hypothesis Development……………………………………………………………………………………………………52

4.2 Hypothesis Testing…………………………………………………………………………………………………….……….52

5 Predictive Modeling………………………………………………………………………………………………………………57

5.1 data preparation for model…………………………………………………………………………………………………57

5.2 Linear Regression………………………………………………………………………………………………………….……57

5.2.1 Assumptions Checking……………………………………………………………………………………………………..57

5.2.2 Regression Analysis………………………………………………………………………………………………………….59

5.3 Clustering………………………………………………………………………………………………………………….……….61

5.3.1 Determining Optimal Number of Clusters………………………………………………………………………..61

5.3.2 K-Means Clustering…………………………………………………………………………………………………….……61

5.3.3 Making Predictions Using a Clustering Model………………………………………………..………………..63

5.4 Classification……………………………………………………………………………………………………………….……..63

5.4.1 Data Preprocessing………………………………………………………………………………………………………….64

5.4.2 Model Training…………………………………………………………………………………………………………………64

5.4.3 Model Prediction and Evaluation…………………………………………………………………..…………………64

5.4.4 Summary…………………………………………………………………………………………………………………..…….66

6 Critical Reviews of The Data Analytics Tools and Techniques……………………………..………………….66

7 Critical Review of the Visualisation Tools Used………………………………………………………………………68

8 References…………………………………………………………………………………………………………………………….69

1.Introduction

1.1 Background

The dataset, published by the Crown Prosecution Service encapsulates a monthly breakdown of criminal case outcomes in the United Kingdom across various primary offence categories and CPS Areas. The Crown Prosecution Service (CPS) was established in 1986. Which is a key component of the criminal justice system in England and Wales and operates independently to make decisions about criminal prosecutions. CPS’s primary role is to review evidence gathered by the police and other investigative authorities and determine whether there is enough evidence to pursue criminal charges against an individual (The Crown Prosecution Service, n.d.)

The original dataset includes 51 variables related to various criminal case outcomes that happened over 43 geographical locations across England and Wales. These variables include numerous criminal activities such as homicide, fraud and forgery, criminal damage, drug offences, theft and handling, offences against the person, robbery, sexual offences, burglary, public order offences, motoring offences, and other offences excluding motoring. Moreover, the dataset reports the number of convictions and unsuccessful outcomes for each criminal category for each CPS Area, presenting a detailed examination of the legal outcomes across different locations.

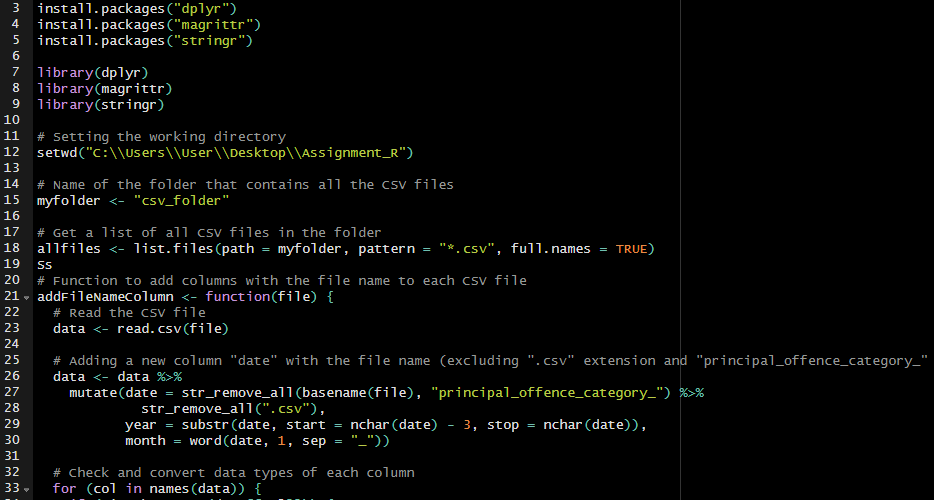
1.2 Objectives

The objective of this analysis is to provide valuable insights into the patterns, trends, and factors driving case outcomes inside the Crown Prosecution Service, Applying a combination of descriptive and predictive analytics tools. Such as the use of descriptive analytics to analyze the dataset, as well as the development of prediction models utilizing linear regression, clustering, and classification approaches.

2.Data Integration and Cleaning

2.1 Data Download and Integration

Data integration is the problem of merging data from several sources and delivering a consistent representation of these facts to the user (Lenzerin, 2002). The dataset download and integration process involved several key steps to assemble a comprehensive dataset for subsequent analysis. The dataset was sourced from the website called data.gov.uk and which covered the data on various principal offences, from the period of 2014 to 2018. The original dataset consists of five distinct folders for each year. In each folder, that represents a year, individual CSV files were allocated to represent monthly breakdowns of criminal case outcomes across various principal offence categories and CPS Areas in the United Kingdom

A subset of the dataset covering the years 2014 to 2016 was chosen, yielding 33 months of data, for this analysis. Remarkably, records for 2014 were available in full, for 11 months in 2015, and 10 months in 2016. To account for differences in data completeness throughout the designated years and cover a meaningful timeframe, this sample was carefully selected. After the selection procedure, the 33 separate files were combined into a special folder while keeping their original filenames.

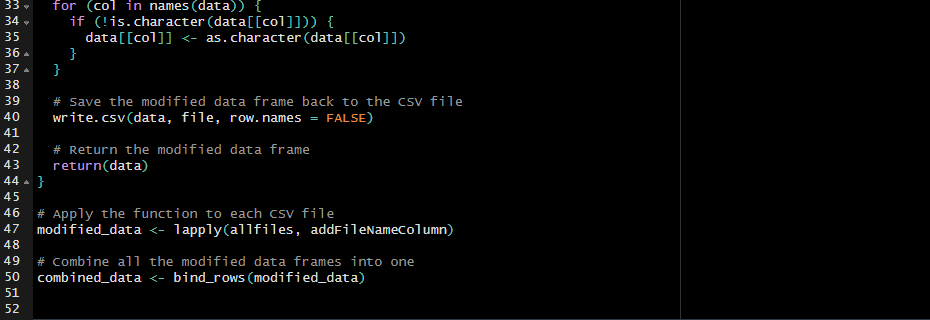


Figure 2.1 Data Download and Integration

The above code demonstrates a systematic approach to integrate and preprocess the Crown Prosecution Service (CPS) Case Outcomes dataset. After configuring the working directory to the relevant folder, using dplyr, magrittr and stringr packages, the following steps have been taken for further analysis. Firstly, a new column called "date” was added to every CSV file. The file name was used to create this new column by removing the ".csv" extension and the particular prefix "principal\_offence\_category\_." The code then continues to extract and add further temporal data, the year and month, from the newly formed "date" column. After that, each column's data type was converted to a common data type for further calculations. Moreover, the updated dataset is saved back into the original CSV files, safeguarding the changes for later use. Then using **lapply**, the script

iterates through each CSV file, applying the custom function to add date-related columns to the data frames. The script uses the bind\_rows function to combine the modified frames into a single dataset. The valuable information extracted from each monthly breakdown of criminal case outcomes across various principal offense categories and CPS Areas in the United Kingdom is combined in this final step to create a cohesive and enriched dataset.

2.2 Data Cleaning and Exploration

2.2.1 Initial data exploration

The process of finding and removing issues from a data warehouse is known as data cleaning. When gathering and merging data from multiple sources into a data warehouse, maintaining good data quality and consistency becomes critical (1Vaishali Chandrakant Wangikar and 2Ratnadeep R. Deshmukh, 2011). During initial data exploration, using tidyverse package, initial look of the data frame has been explored with the help of str() and summary() functions. With the help of Skimr package, skim() has been used to generate further details of summary statistics. Furthermore, using vis\_dat() function, visual representation of missing values, summary statistics and distribution and correlation plots of the data set has been generated.



Figure 2.2 Initial data exploration

Figure 2.2 show that the dataset has 76 626 of total observations with 54 columns and 1419 rows.

* + 1. Changing data types

The data type of “Year” has been converted to numeric. Also, from the 2nd variable to the 51st variable, the data type of every other variable was changed to as integer since the observations of particular variables consist of rounded values. Subsequently, from the 3rd variable to the 51st variable, the data type of every other variable was changed to numeric, since it contains percentage values

* + 1. Changing the column names

With the help of tidyverse and dplyr packages, I have renamed the first variable as “Area”. After that every other column from the 2nd variable to the 51st variable which had “Number.of” as a prefix was removed. Subsequently, every variable which had “Percentage.of” as a prefix was replaced with “%” symbol for easy reference.

* + 1. Inspect for mislabeled column names

After examining for mislabeled column names, it was discovered that there were no reports of columns with improper labels or mislabeling.

* + 1. Fixing string inconsistencies

Column names 52, 53, and 54 were changed to "Date," "Year," and "Month," accordingly. This ensures consistency and adheres to the naming practice in which variable names begin with capital letters. The names(mydata) function is used to display the modified column names. Moreover, using “gsub” dots (“.”) between column names were removed

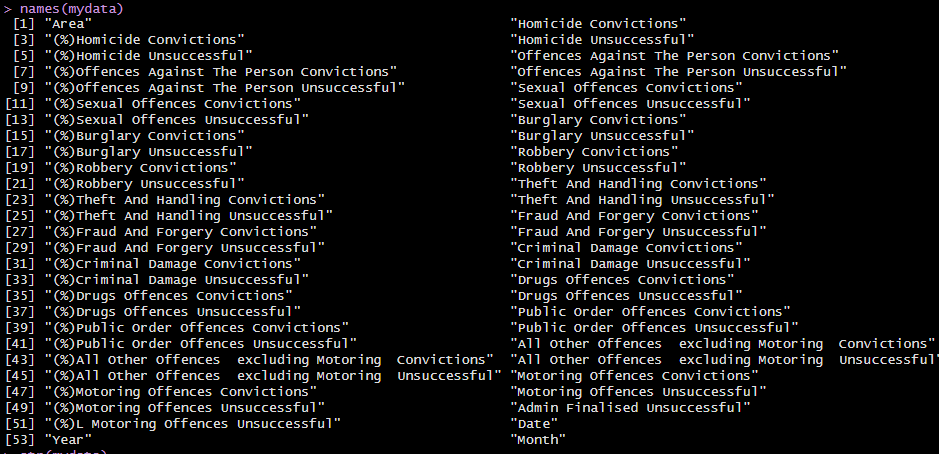
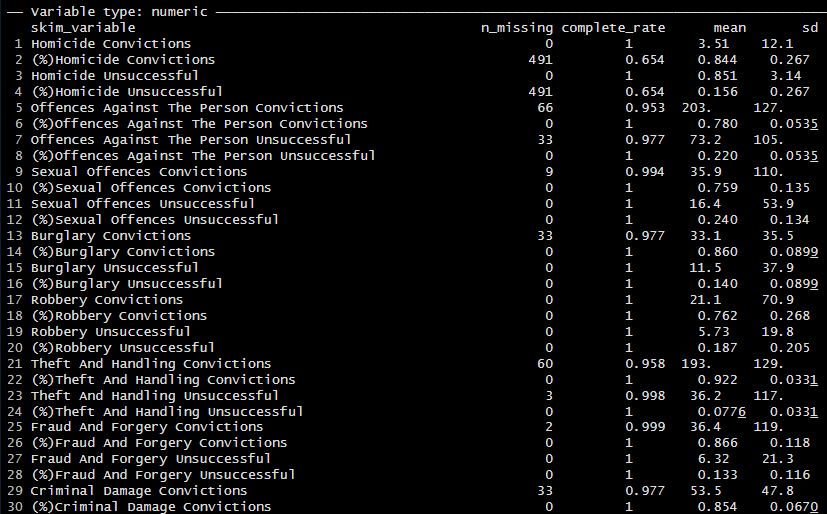
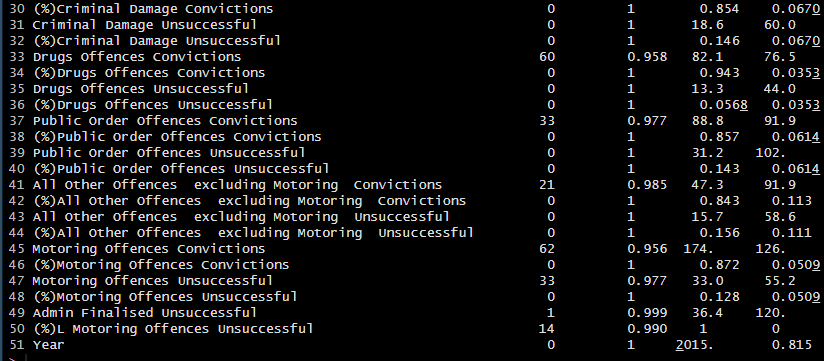


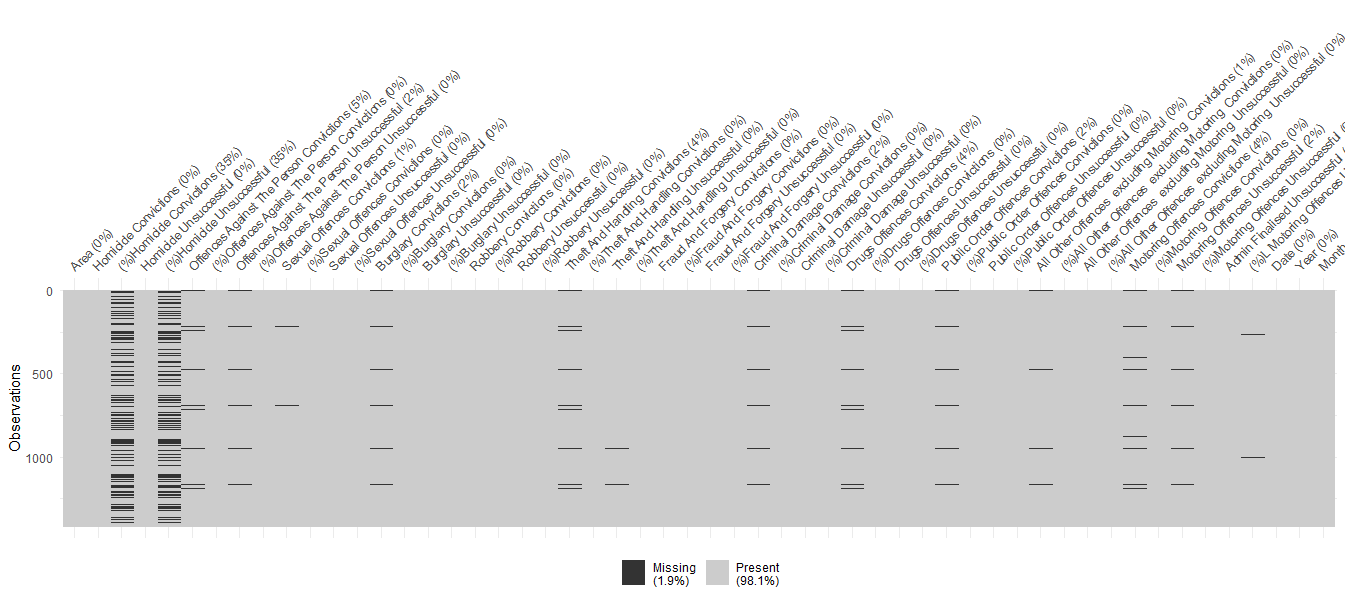
Figure 2.3 Variable Names

* + 1. Missing Values

Missing values can arise as a result of a variety of causes, including missing totally at random, missing at random, or missing not at random. All of this could be the consequence of a system failure during data collecting or human error during data pre-processing (Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago & Oteng Tabona , 2021). The missing values problem is common in all domains that deal with data and produces a variety of concerns such as performance degradation, data analysis issues, and biased results caused by disparities in missing and complete values (Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM, 2019). The package's skim\_without\_charts() function is then used to provide a summary of the data frame, including missing values, data kinds, and summary statistics. Figure 2.4 and Figure 2.5 shows the missing Values in the original data frame.



Figure 2.4 Missing Values

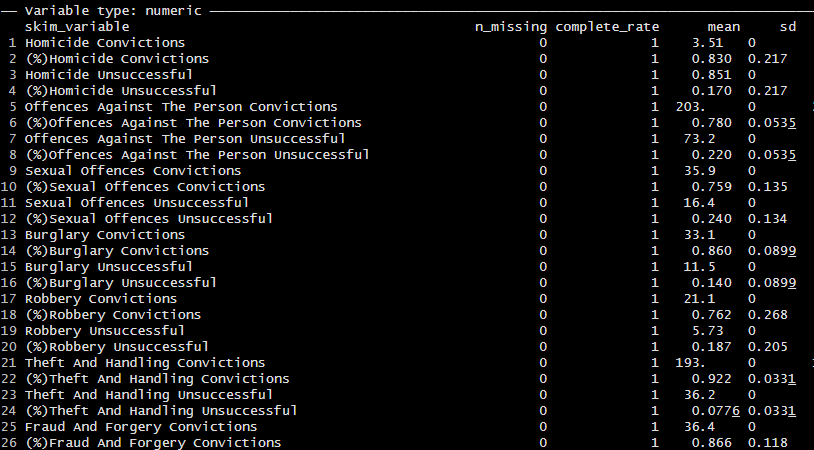
Figure 2.5 Missing Values

As the first step, the columns with integer data types were selected. After that missing values of each variable were replaced with the mean values of respective variables. When data for specific observations is missing, eliminating those rows may result in the loss of valuable information and a biased representation of the dataset. The report has maintained the original mean of the variable by imputing missing values with the mean, ensuring that the statistical summary of the dataset remains relatively unchanged. This method provides a decent estimate for missing values without significantly altering the overall properties of the data. Thereafter, missing values of percentage variables have been calculated by using the below formulas,

Ex: %Homicide Convictions (Success) = Homicide Convictions` / (`Homicide Convictions` + Homicide Unsuccessful)

Ex: %Homicide Convictions = `Homicide Unsuccessful` / (`Homicide Convictions` + `Homicide Unsuccessful`)

Below figure 2.6 and figure 2.7 shows summary of variables after removing missing values.



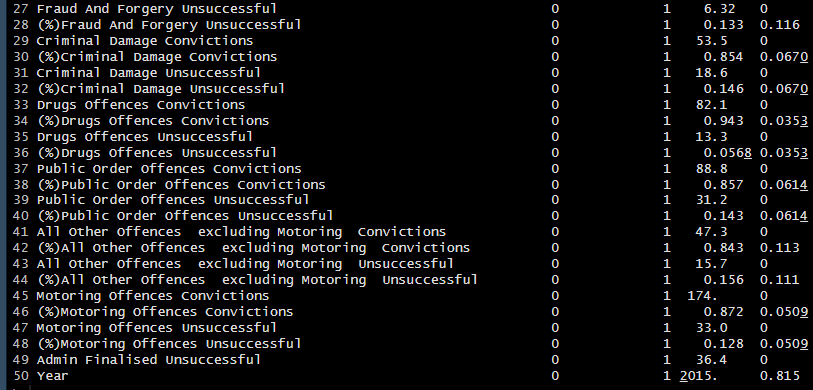


Figure 2.6 Missing Values

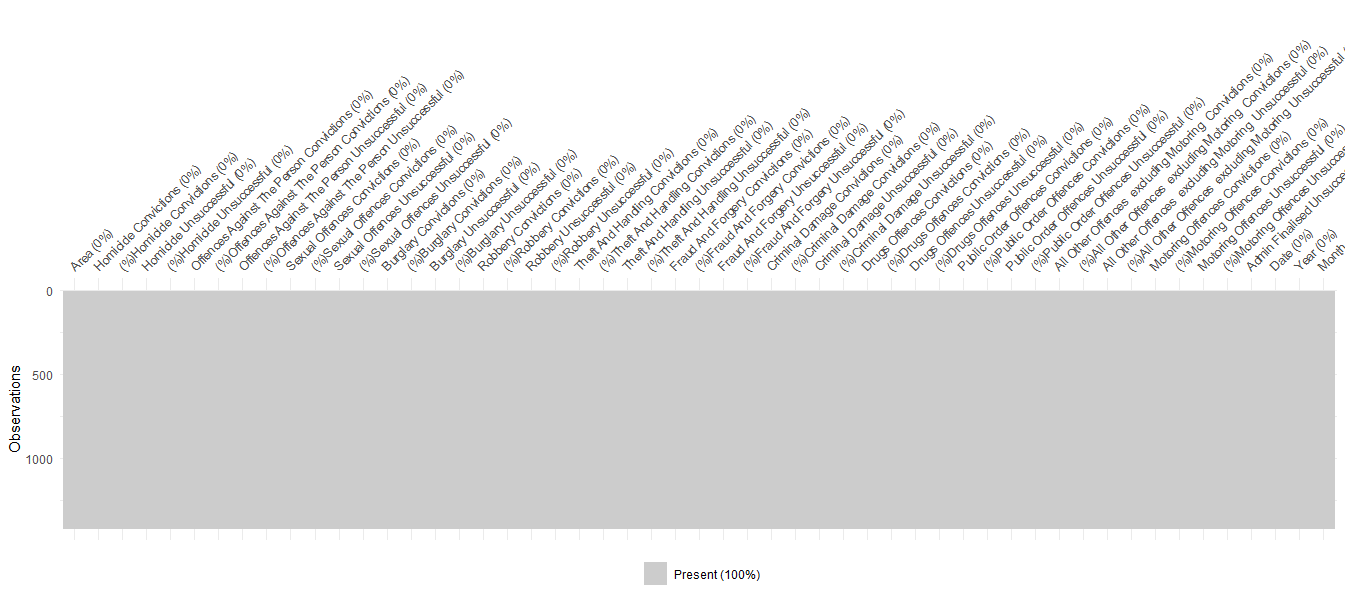


Figure 2.7 Missing Values

The figure show that there are no missing values in the data set anymore.

* + 1. Re-organize Columns and Data Frames

Several actions have been taken to the original data frame (mydata) for future reference. Initially, a lookup table for month names is built, and data frame was connected with this database to replace month names with corresponding numeric values. The "Month" column is then changed to a numeric type. A new data frame (mydata2) is formed by picking specified variables from the original one while omitting those containing "(%)" and "unsuccessful" in their names. "Mydata2"'s "Month" column is likewise transformed to numeric, and a new "Time" column is created by combining "Year" and "Month." Following that, a third dataframe "mydata3" is formed by selecting important columns from "mydata" such as "Area," "Year," "Month," "Time," columns containing "unsuccessful," and columns excluding "(%)." "mydata3"'s "Month" column is transformed to numeric, and a new "Time" column is created in the same manner as "mydata2." The end result is a structured dataframe "mydata3" that is ready for further analysis, complete with proper data types and time-related columns. In summary, successful and unsuccessful principal components were separated from the original data frame.

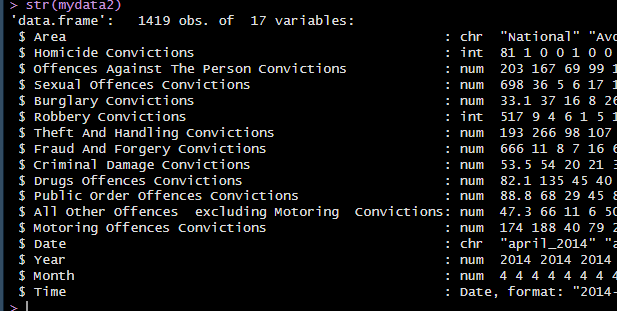


Figure 2.8 Criminal convictions Successful

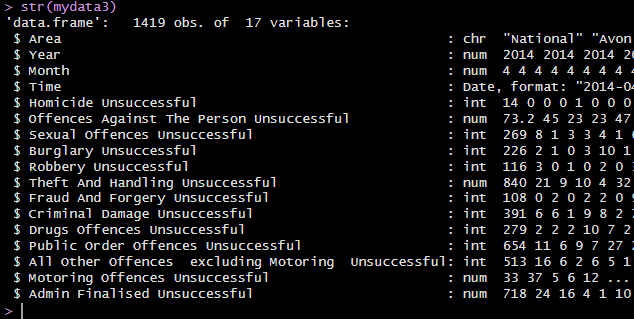


Figure 2.9 Criminal Convictions Unsuccessful

Figure 2.8 shows criminal convictions successful and Figure 2.9 shows successful criminal convictions

3 Descriptive Analysis

Descriptive analysis, known as descriptive statistics, is the process of describing or summarizing a set of data using statistical techniques. As one of the major types of data analysis, descriptive analysis is known for producing actionable insights from otherwise uninterpreted data (BUSH, 2020). Statistics can be classified into two categories which are descriptive statistics and inferential statistics. Descriptive analysis serves as a basis for more complex statistical analyses and assists researchers, analysts, and decision-makers in understanding the data's properties. It is an important phase in the whole data analysis process, assisting in the identification of patterns, trends, outliers, and other traits that may be valuable for further examination

3.1 Summary Statistics

3.1.1 Summary Statistics based on Conviction type (Successful)

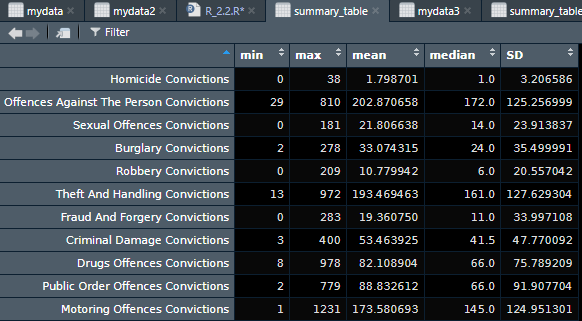


Figure 3.1 Summary Statistics Conviction Successful

Figure 3.1 indicate the summary statistics of various primary offense conviction categories in England and Wales from 2014 to 2016.Motoring Offence shows the highest number among all the convictions (1231). Theft And Handling (193.42) present the highest variation in the number of convictions while Homicide (8.32) has the lowest variation in the number of convictions. Offence against the person convictions shows the highest mean (202.87), while homicide shows the lowest.

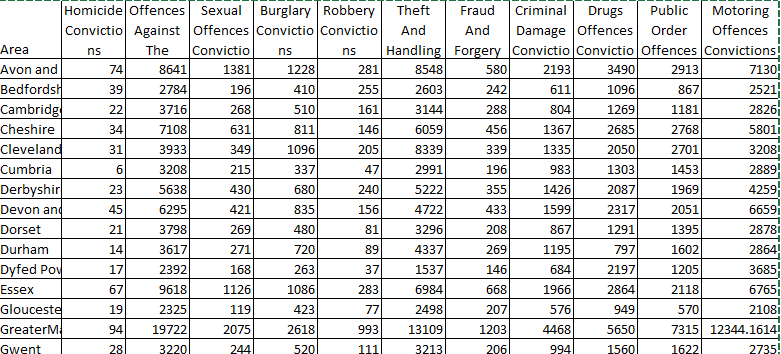
3.1.2 Summary Statistics based on Conviction type (Unsuccessful)

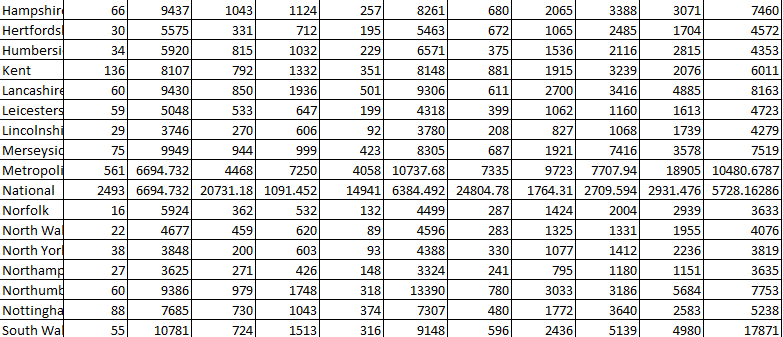


Figure 3.2 Summary Statistics Conviction Unsuccessful

Figure 3.2 indicate the summary statistics of various primary offense conviction categories Unsuccessful in England and Wales from 2014 to 2016. Offence against the person unsuccessful shows the highest number among all the convictions (862) as well as highest variation (104.6) . while Homicide (1.26) has the lowest variation in the number of convictions. Offence against the person unsuccessful shows the highest mean (73.2), while homicide shows the lowest (0.43).

3.1.3 Summary Statistics based on Area (Successful)





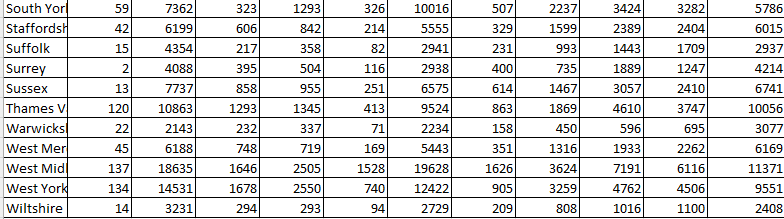
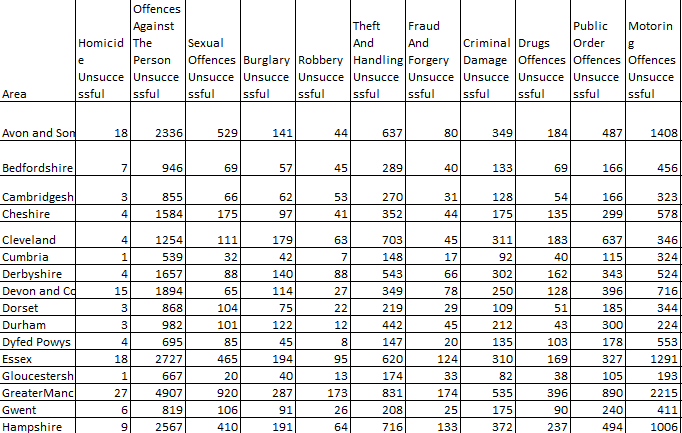
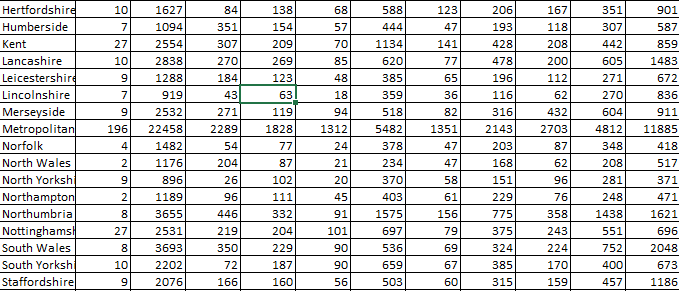


Figure 3.3 area summary successful

3.1.4 Summary Statistics based on Area (Unsuccessful)





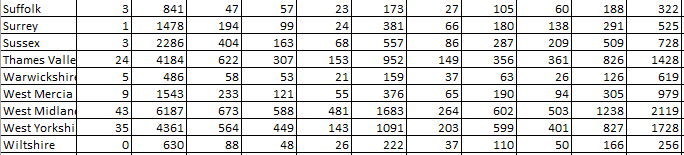


Figure 3.4 area summary unsuccessful

3.2 Distribution of observations

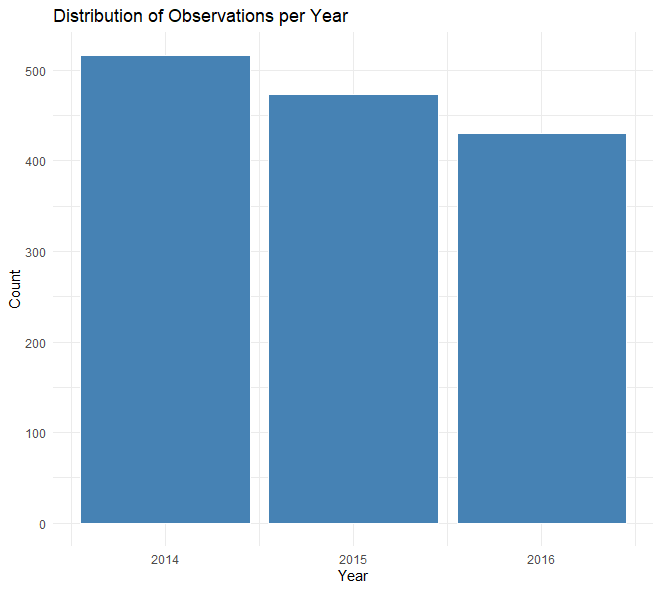


Figure 3.5 Distribution of observations

Figure 3.5 indicates the distribution of observations throughout the years. From 2014 to 2016 the highest number of observations can be seen in 2014 and 2016 is the lowest.

3.3 Overall Conviction Categories Across Years

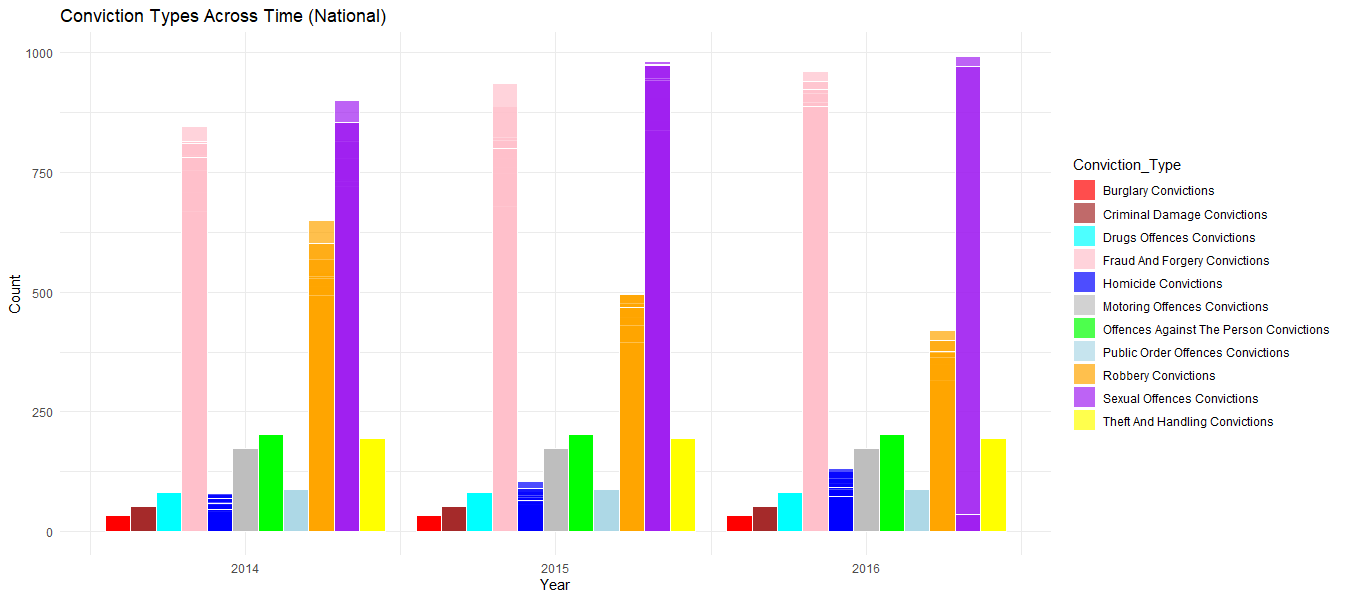


Figure 3.6 Conviction Categories Across Time

Figure 3.6 indicates overall principal offence category types, regardless of area from 2014 to 2016. Which shows sexual offence convictions as highest number of conviction category in all three years and the Burglary as the lowest type of convictions. sexual offence has increased over the time. There can be seen a gradual decrease in Robbery convictions

3.4 Homicide

3.4.1 Distribution of Homicide Convictions vs Average all other offences over time

* Homicide Successful

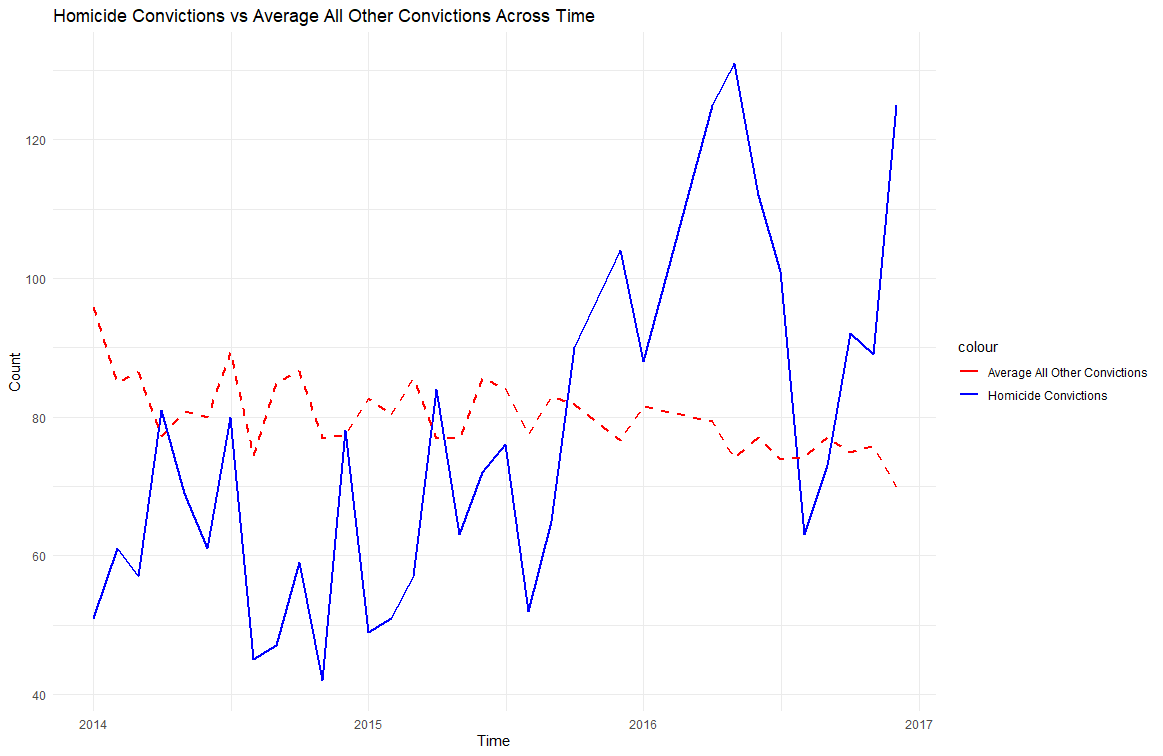


Figure 3.7 Homicide Successful

The figure 3.7 shows the number of Homicide Convictions successful vs average criminal offence of all the criminal offence categories in England and Wales, from 2014 to 2016. There can be seen an increased trend of homicide Convictions while the average number of all other convictions shows a slightly decreasing trend. It hit its lowest point in the last quarter of 2014 (43) and the first half of 2016 was when it hit its highest point (over 130). Also, there was a serious drop of Homicide Convictions successful right after it hit the highest.

* Homicide Unsuccessful

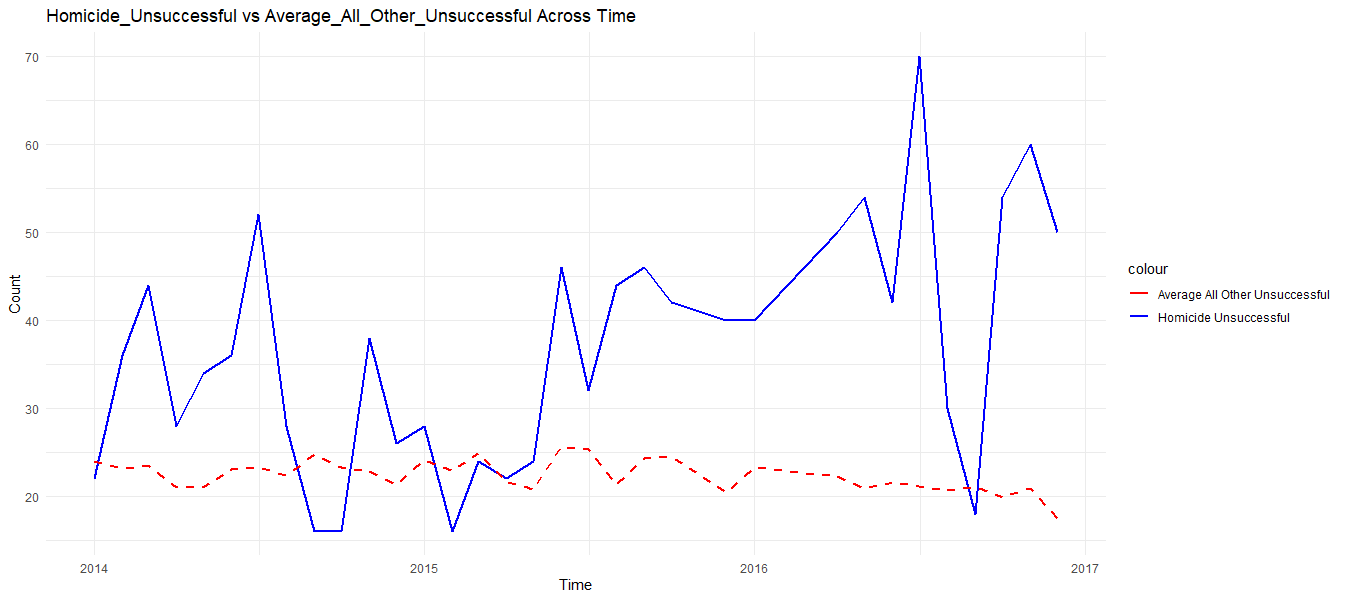


Figure 3.8 Homicide Unsuccessful

The figure shows the distribution of Homicide Convictions unsuccessful vs average criminal offence unsuccessful of all the criminal offence categories from 2014 to 2016. There can be seen an overall increasing trend of homicide Convictions while the average number of all other convictions shows a slightly decreasing trend. Homicide unsuccessful has hit its lowest point in the 3rd quarter of 2014 (below 10) and the beginning of the 2nd half of 2016 was when it hit its highest point (70). Also, there was a significant drop in Homicide Convictions unsuccessful right after it hit the highest.

3.4.2 Distribution of Homicide Convictions Across the Area

* Homicide Successful

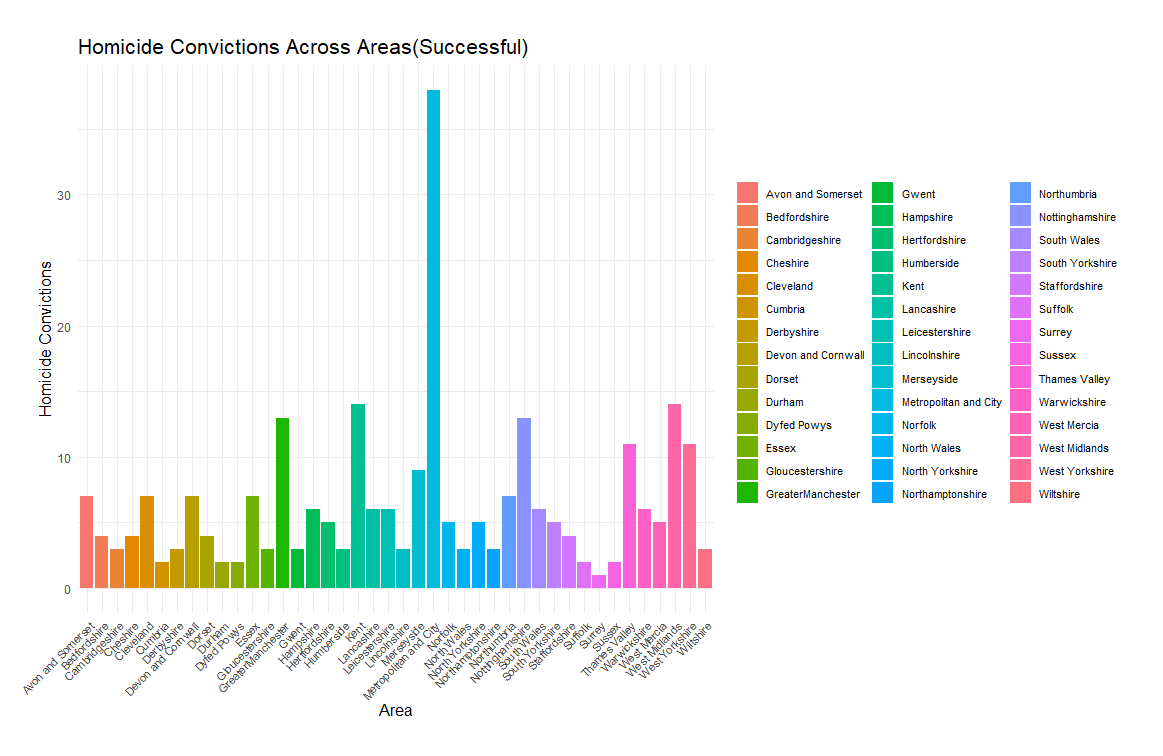


Figure 3.9

Figure 3.9 shows the number of Homicide convictions successful across different areas in England and Wales, from 2014 to 2016. Metropolitan and City shows the highest rate of Homicide convictions successful (over 35), which is more than two times higher compared to all other areas and the Surey shows the lowest number

* Homicide Successful

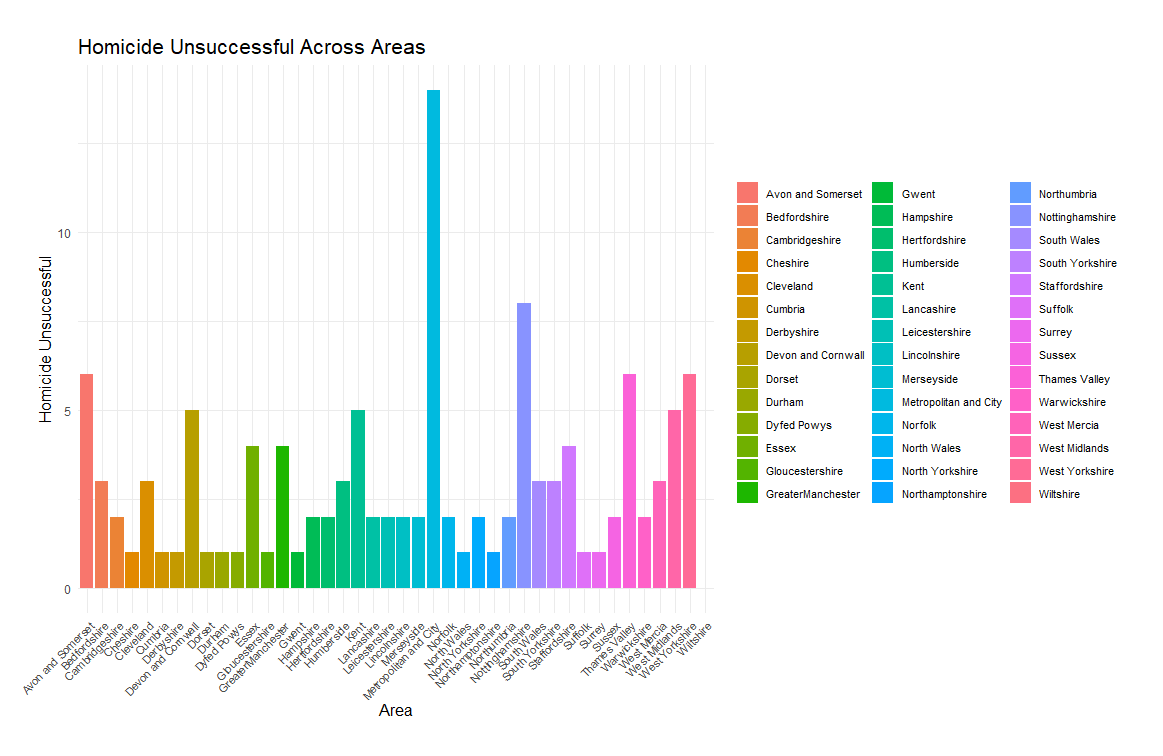


Figure 3.10

Figure 3.10 shows the number of Homicide convictions unsuccessful across different areas in England and Wales, from 2014 to 2016. Metropolitan and City show the highest rate of Homicide convictions unsuccessful (12.5). Overall all the counties have less than 10 unsuccessful crime numbers except Metropolitan

3.5 Offences Against Person Convictions

3.5.1 Distribution of Offences Against Person Convictions

* Offences Against Person Convictions Successful

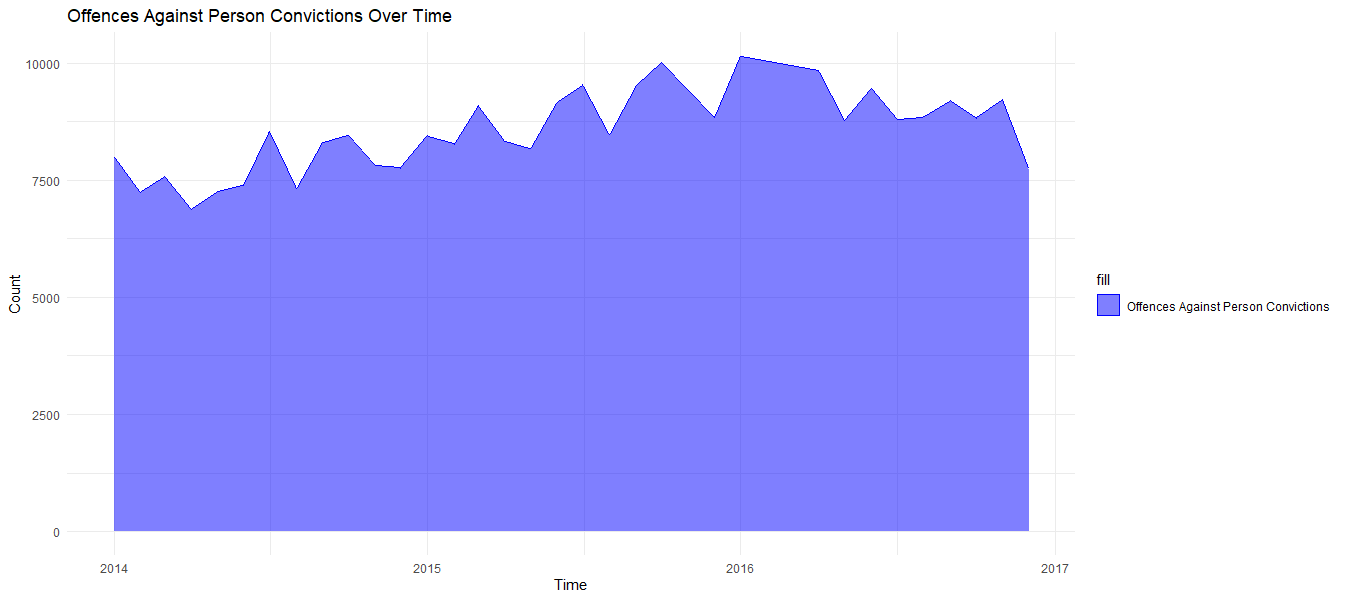
Figure 3.11

Figure 3.11 shows the distribution of Offences Against Person Convictions Successful from 2014 to 2016, in England and Wales. There can be seen a slightly increasing trend throughout years with some seasonal fluctuations. In the beginning of 2016 number of Offences Against Person Convictions have reached the maximum level which is 10000, after that it has started to reduce gradually. The minimum number of convictions can be seen around the first quarter of 2014.

* Offences Against Person Unsuccessful

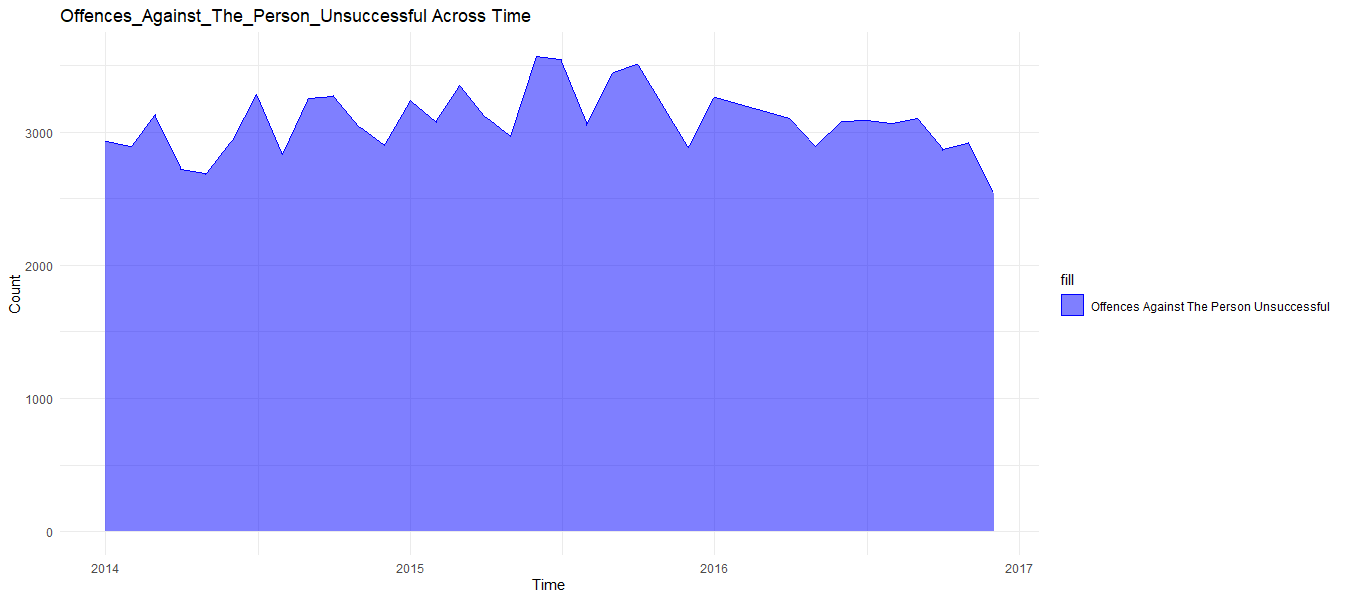


Figure 3.12

Figure 3.12 shows the distribution of Offences Against Person Convictions unsuccessful from 2014 to 2016, in England and Wales. Apart from seasonal fluctuations, there cannot be seen a significant trend throughout years. In beginning of May 2015, the distribution has reached to the maximum number of unsuccessful convictions which is about 3500 in numbers.

3.5.2 Distribution of Offences Against Person Across the Area

* Offences Against Person Successful

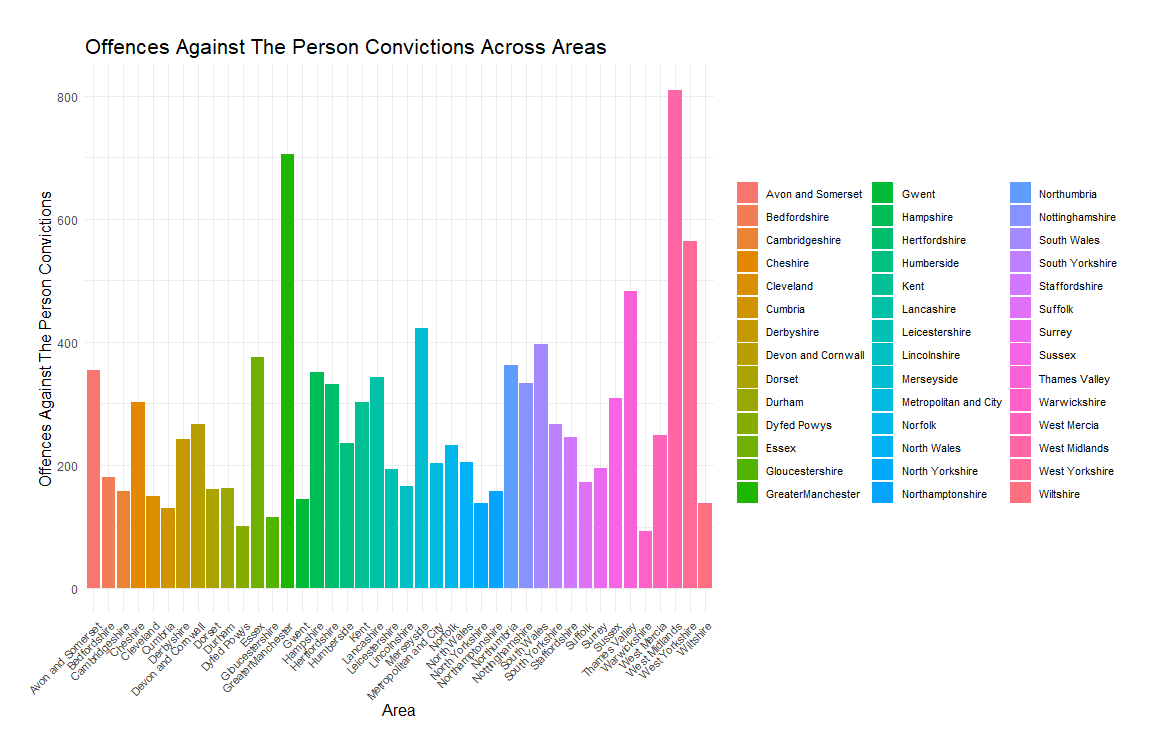


Figure 3.13

Figure 3.13 shows the number of offences Against Person successful across different areas in England and Wales, from 2014 to 2016. West Midlands shows the highest number of offences Against Person (over 800) while Warwickshire shows the lowest (100). Greater Manchester has the second highest convictions number which is 700.

* Offences Against Person Unsuccessful

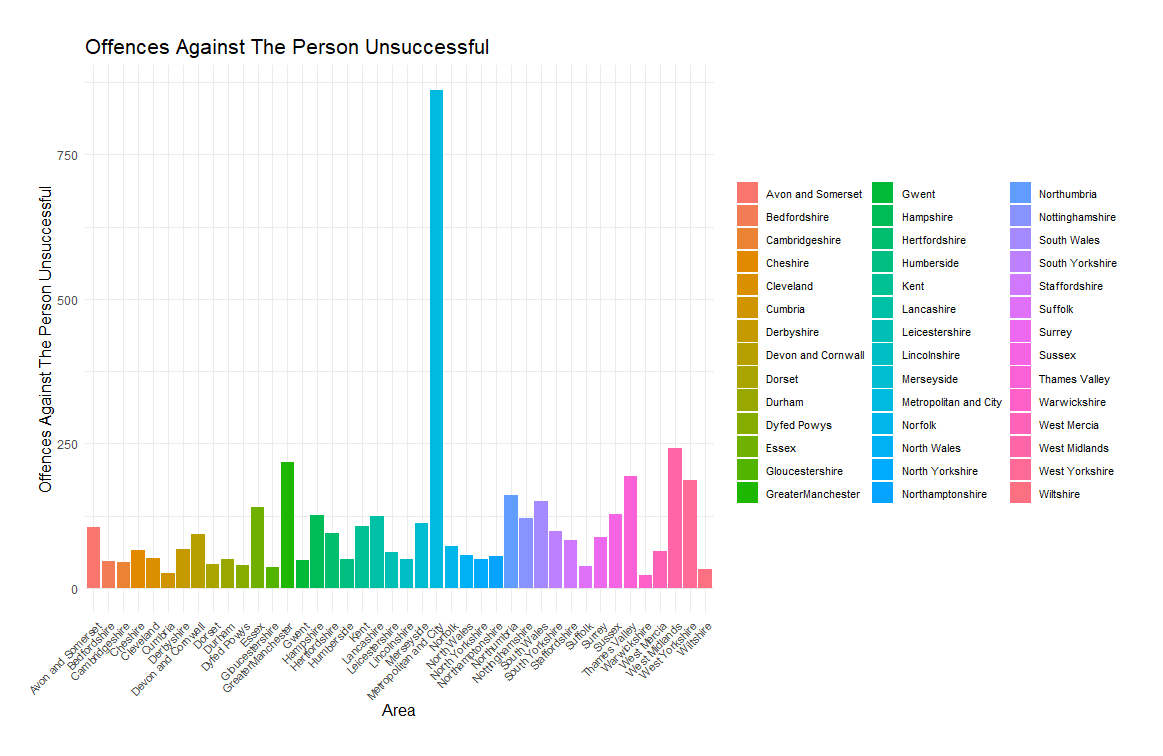


Figure 3.14

Figure 3.14 shows the number of offences Against Person convictions unsuccessful across different counties in England and Wales, from 2014 to 2016. Northamptonshire shows the highest number of unsuccessful rates which is about 800 and more than two times higher than other areas. Overall, all the counties represent below 250 number of unsuccessful offences Against Person convictions except Northamptonshire

3.6 Sexual Offences

3.6.1 Distribution of Sexual Offences vs Average all other offences over time

* Sexual Offences Successful

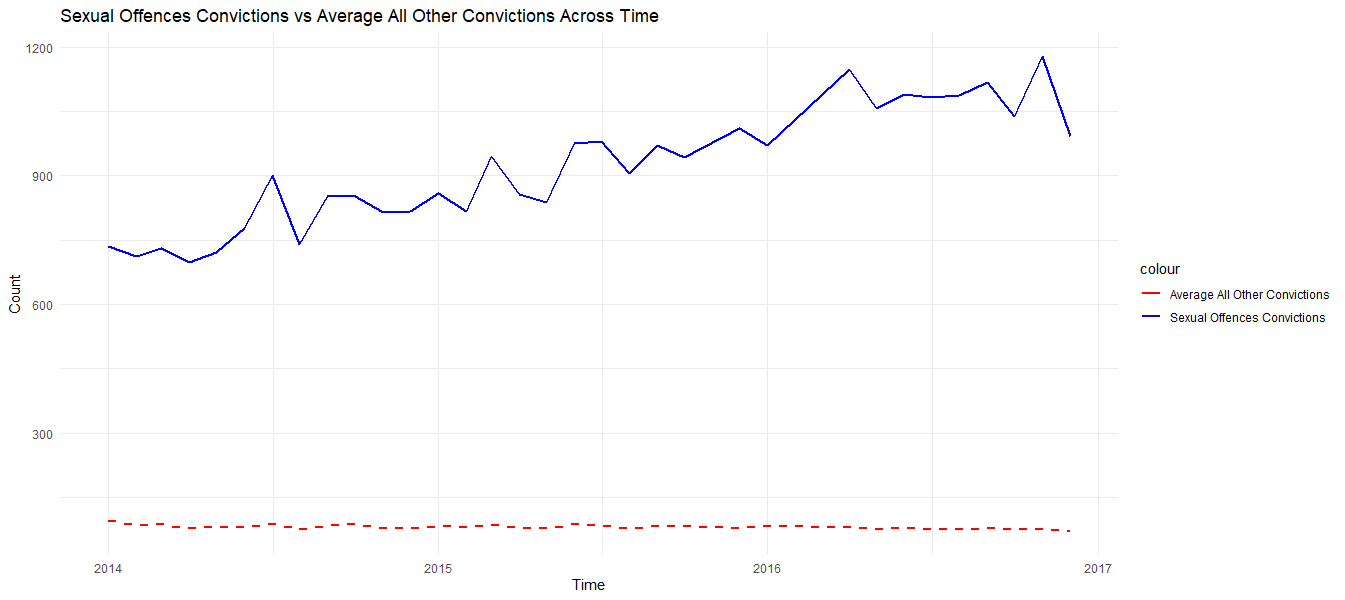


Figure 3.15

Figure 3.15 shows the number of Sexual Offences Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen an increase trend of Sexual Offences while the average number of all other convictions shows constant trend line. Sexual Offences has hit the lowest point in the end of first quarter of 2014 (700) and the beginning of 2nd quarter, 2015 was where it hit the highest point (1100)

* Sexual Offences Unsuccessful

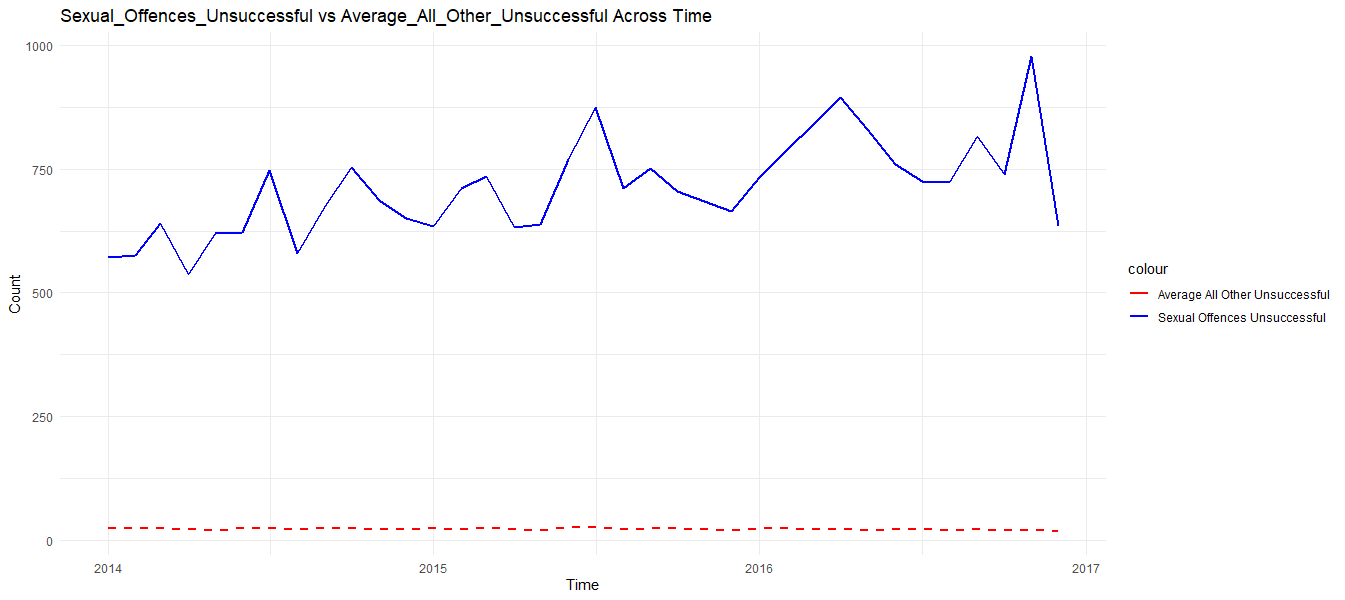


Figure 3.16

Figure 3.16 shows the number of Sexual Offences Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. The graph shows slightly increasing trend throughout years. The lowest number can be identified in first half of 2014 while the highest Sexual Offences happened in the end of 2016 which is 1000 in numbers. After the line hit highest there can be seen a significant drop in unsuccessful Sexual Offences.

3.6.2 Distribution of Sexual Offences Across the Area

* Sexual Offences Successful

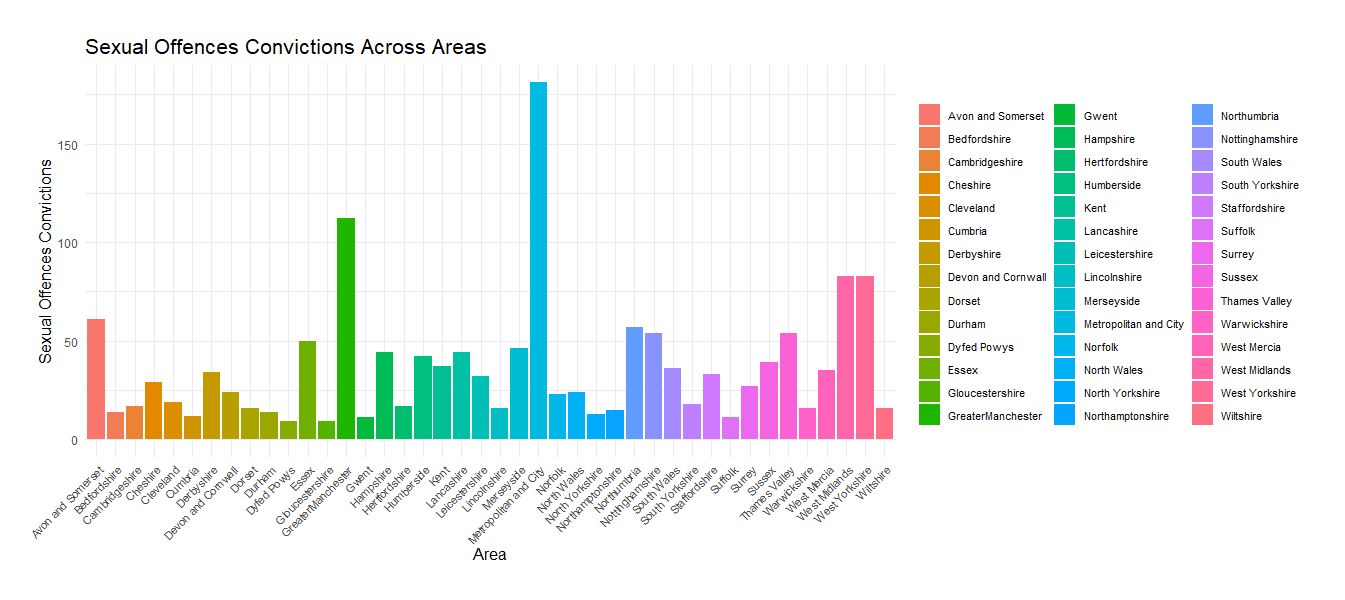


Figure 3.17

Figure 3.17 shows the number of Sexual offences convictions across different counties in England and Wales, from 2014 to 2016. Metropolitan and city shows the highest number of sexual convictions among counties, which is more than 175. The second highest represent Greater Manchester. Overall, all the counties have less than 100 sexual convictions except Metropolitan and city and Greater Manchester

* Sexual Offences Unsuccessful

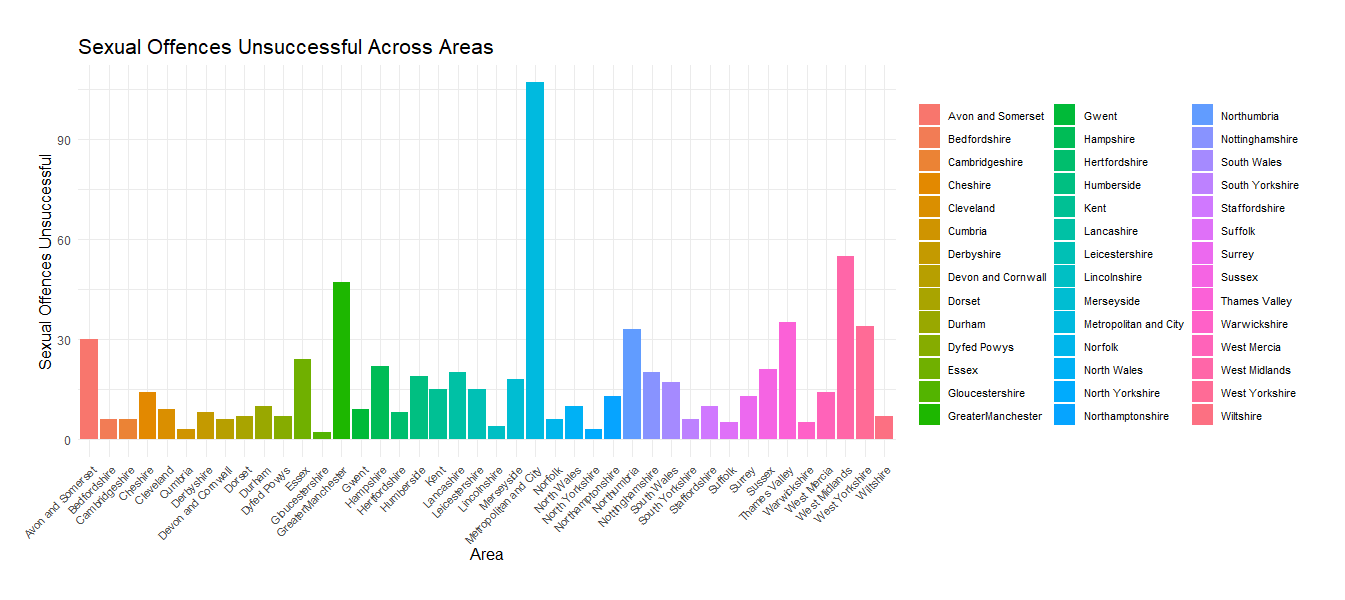


Figure 3.18

Figure 3.18 shows the number of Sexual offences convictions across different counties in England and Wales, from 2014 to 2016. Metropolitan and city shows the highest number of sexual convictions unsuccessful which is more than 100 as well as two times higher than other areas. Overall, all the areas has less has 60 unsuccessful sexual convictions except Metropolitan and city.

3.7 Burglary

3.7.1 Distribution of Burglary vs Average all other offences over time

* Burglary Successful

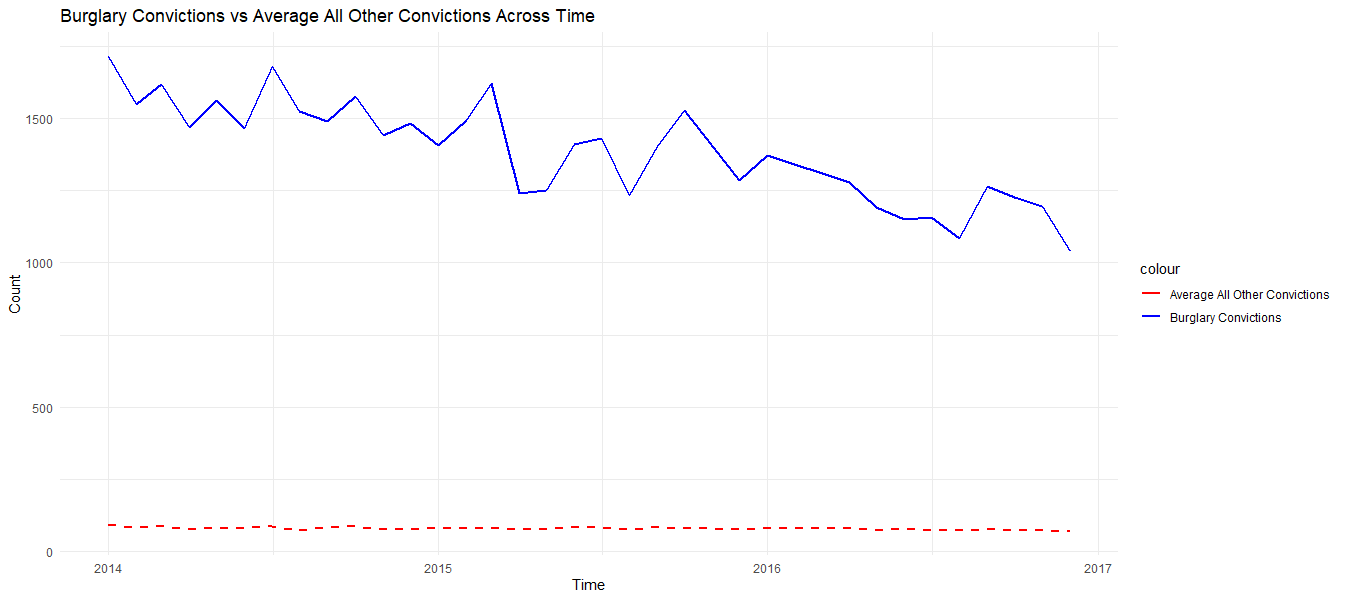


Figure 3.19

Figure 3.19 shows the number of Burglary Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decrease trend of Burglary Offences while the average number of all other convictions shows constant trend line. From the beginning of 2014 it shows approximately 1750 of successful rate but at the end of 2016 which shows 1000 nearly.

* Burglary Unsuccessful

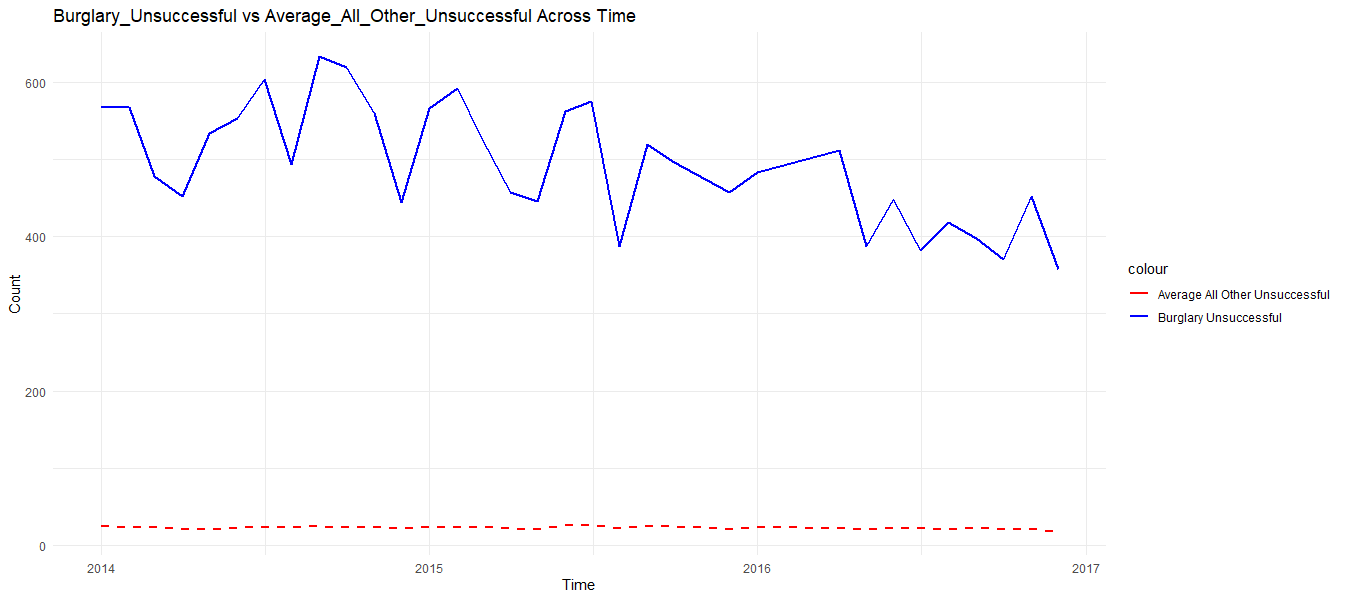


Figure 3.20

Figure 3.20 shows the number of Burglary Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. The graph shows the decreasing trend from 2014 to 2016. The distribution reached to the highest point in the second half of 2014 which indicates approximately 650 number of unsuccessful burglaries.

3.7.2 Distribution of Burglary Across the Area

* Burglary Successful

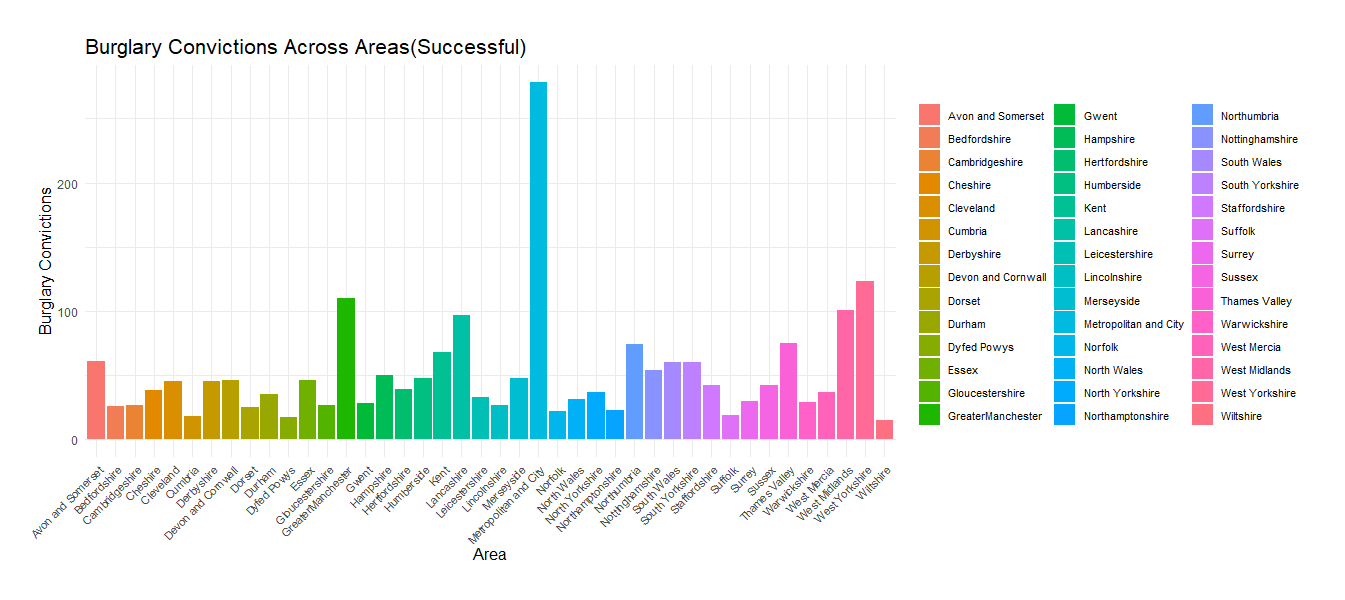


Figure 3.21

Figure 3.21 indicates the number of Burglary convictions (successful) across different counties in England and Wales, from 2014 to 2016. According to the statistics the highest number of Burglary convictions has happened in Metropolitan City (270 approximately). Which is two times more than second highest county Great Manchester.

* Burglary Unsuccessful

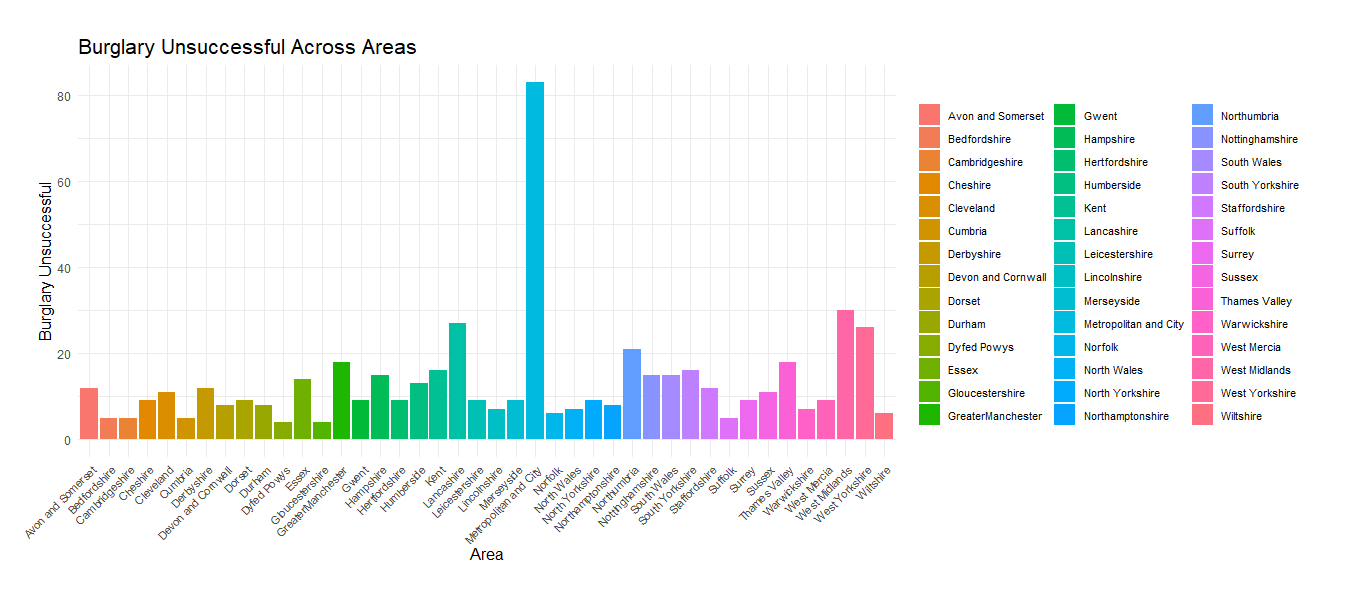


Figure 3.22

Figure 3.22 indicates the number of Burglary convictions (Unsuccessful) across different counties in England and Wales, from 2014 to 2016. The highest number of Unsuccessful Burglary convictions has happened in Metropolitan City, which is slightly more than 80 in numbers. Overall, all other counties have less than 30 number of unsuccessful Burglary convictions.

3.8 Robbery

3.8.1 Distribution of Robbery vs Average all other offences over time

* Robbery Successful

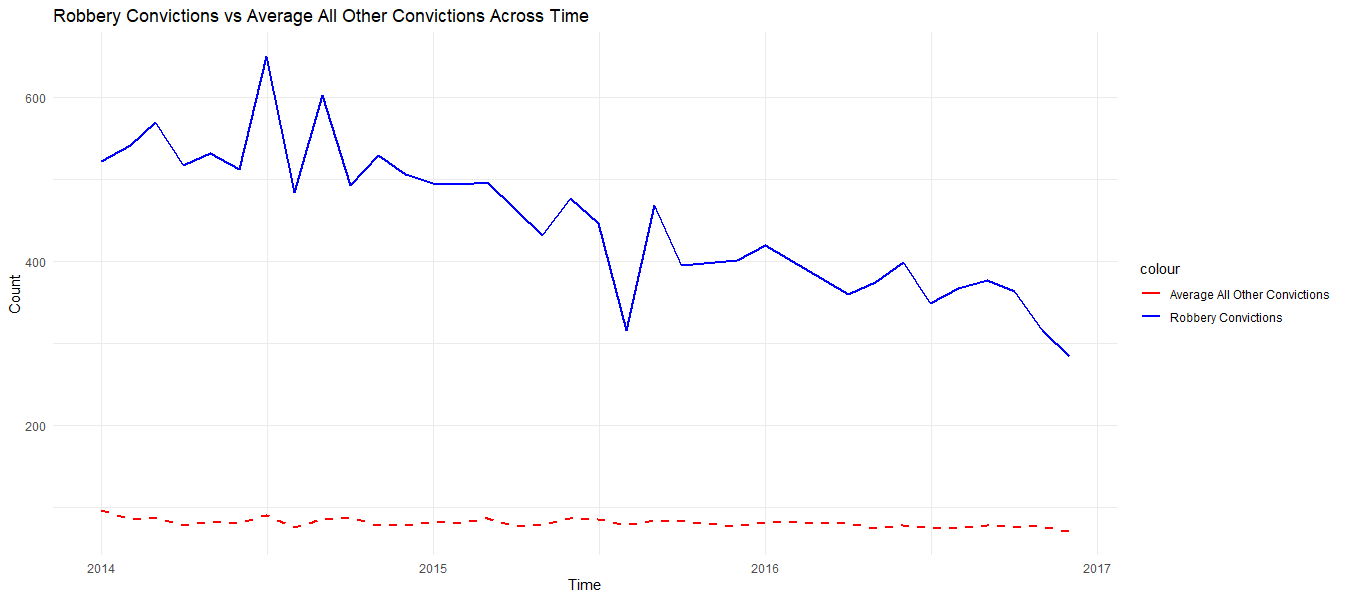


Figure 3.23

Figure 3.23 shows the number of Robbery Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decrease trend of Robbery offences while the average number of all other convictions shows constant trend line. From the beginning of 2014 it shows approximately 500 of successful Robbery rate but at the end of 2016 which has reduced to 300.

* Robbery Unsuccessful

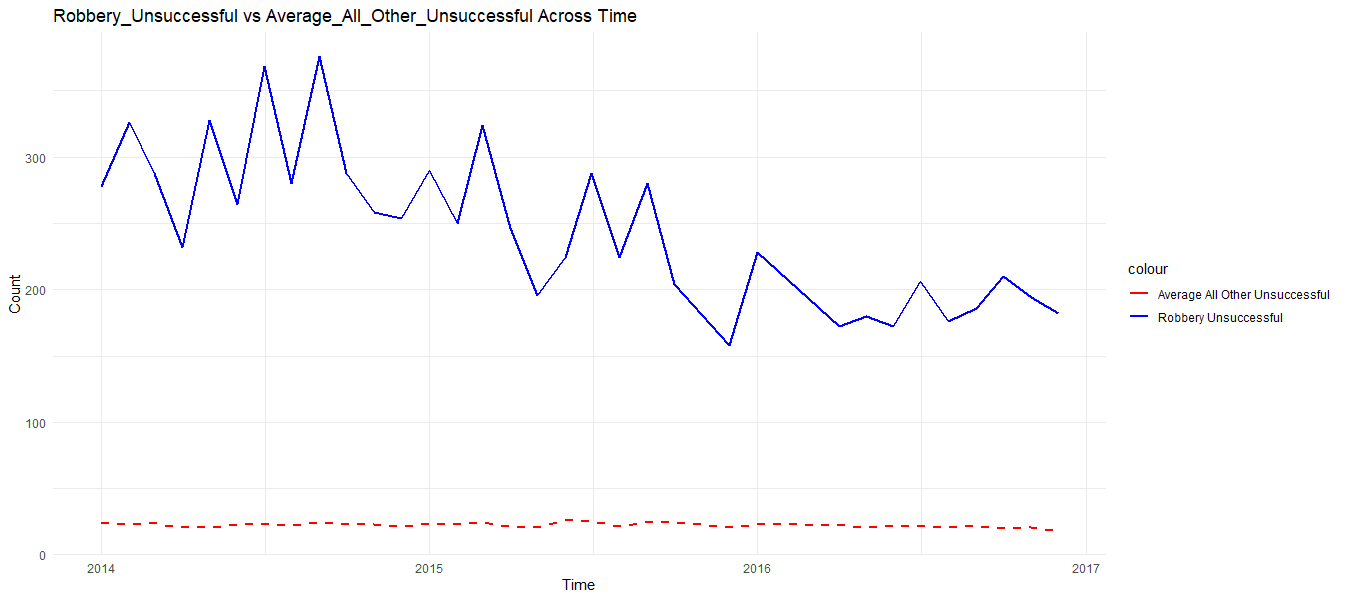


Figure 3.24

Figure 3.24 shows the number of Robbery Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a slightly decreasing trend with significant seasonal fluctuations.

3.8.2 Distribution of Robbery Across the Area

* Robbery Successful

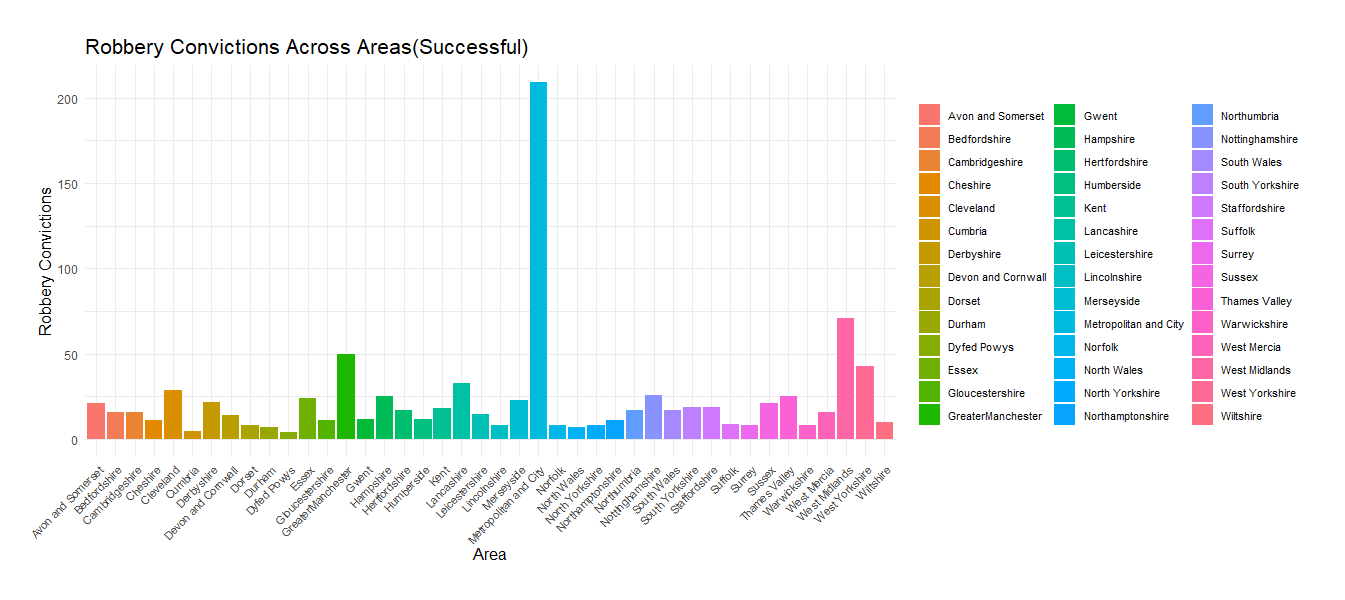


Figure 3.25

Figure 3.25 indicates the number of Robbery convictions (Successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan City shows the highest number of successful robbery (210) convictions among the counties. Which is three times higher than other counties.

* Robbery Unsuccessful

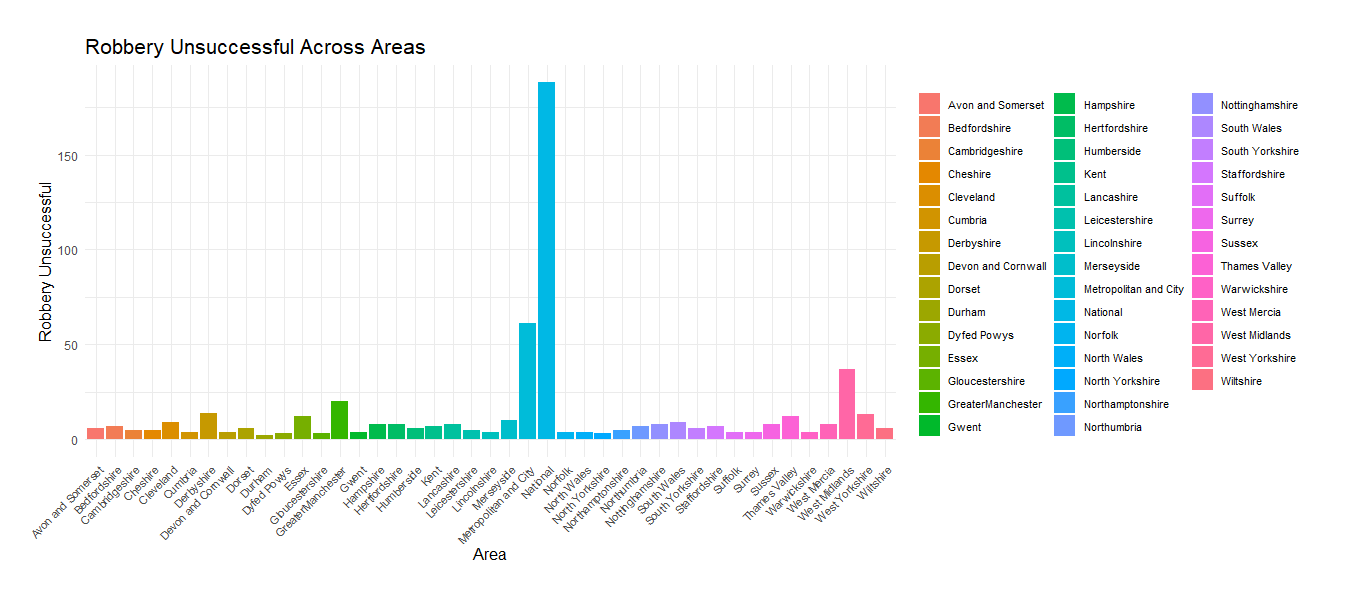


Figure 3.26

Figure 3.26 indicates the number of Robbery convictions (successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan City shows the highest number of successful robbery (over 125) convictions among the counties.

3.9 Theft And Handling

3.9.1 Distribution of Theft And Handling vs Average all other offences over time

* Theft And Handling Successful

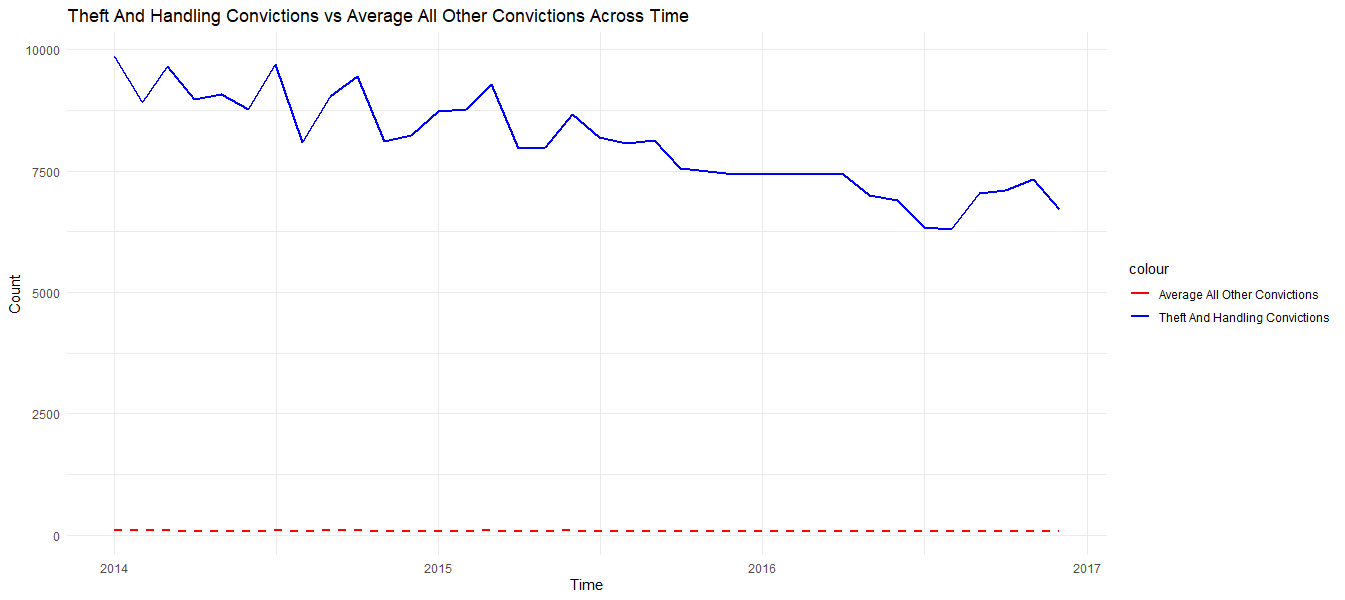


Figure 3.27

The figure shows the number of Theft And Handling Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decrease trend while the average number of all other convictions shows constant trend line. It indicates 10 000 of Theft And Handling in the begning of 2014 and it has reduced to approximately 7000.

* Theft And Handling Unsuccessful

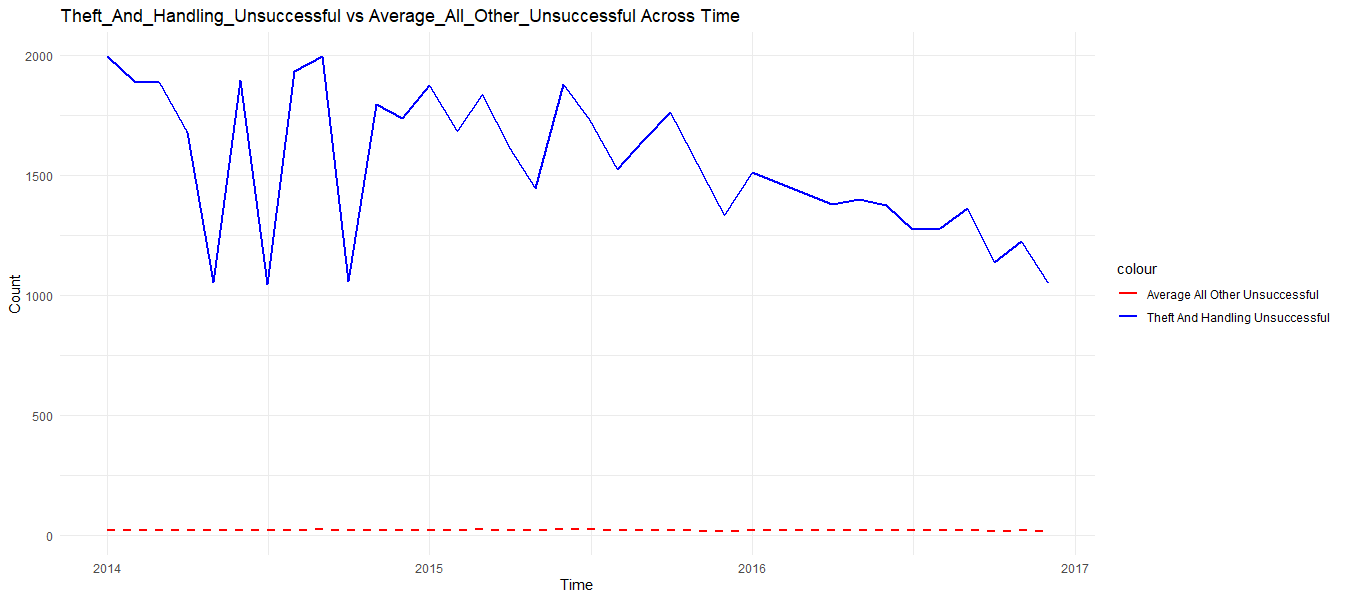


Figure 3.28

Figure 3.28 shows the number of Theft And Handling Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. In 2014 the graph shows some significant fluctuations of numbers between 1000 to 2000 and by the end o the time period the numbers have been decreased.

3.9.2 Distribution of Theft And Handling Across the Area

* Theft And Handling Successful

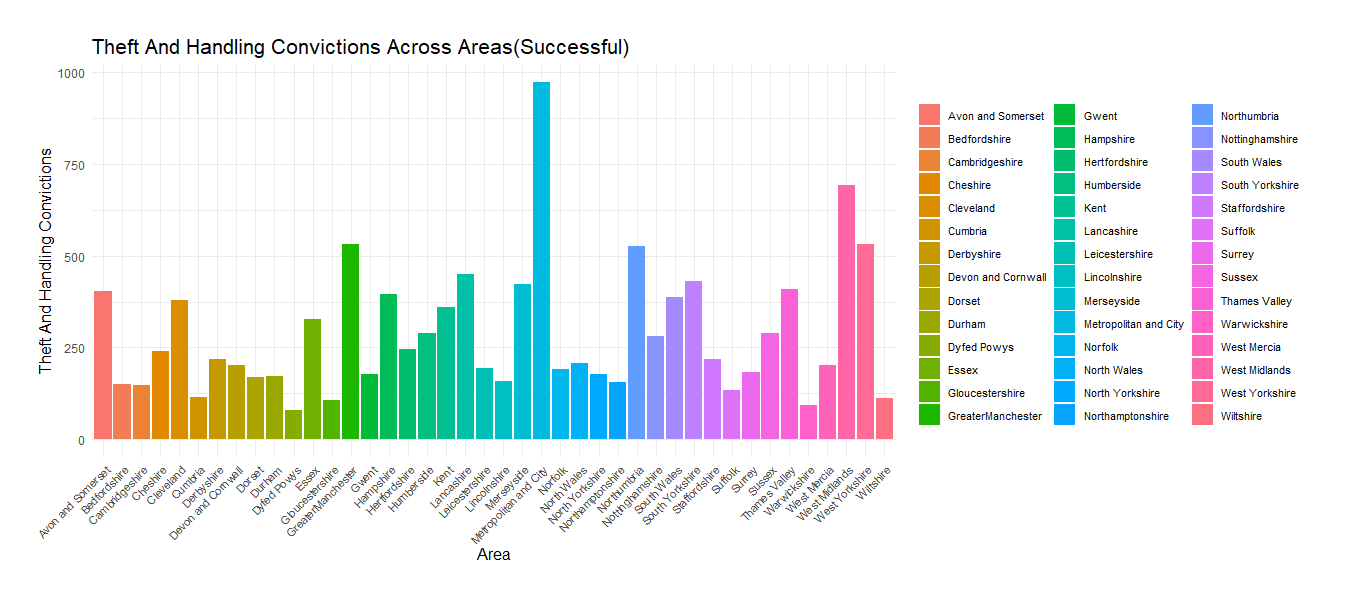


Figure 3.29

Figure indicates the number of Theft And Handling convictions (Successful) across different counties in England and Wales, from 2014 to 2016. The highest number of Theft And Handling convictions can be seen in the Metropolitan and City which is approximately 1000. And the second highest shows in West Midlands.

* Theft And Handling Unsuccessful

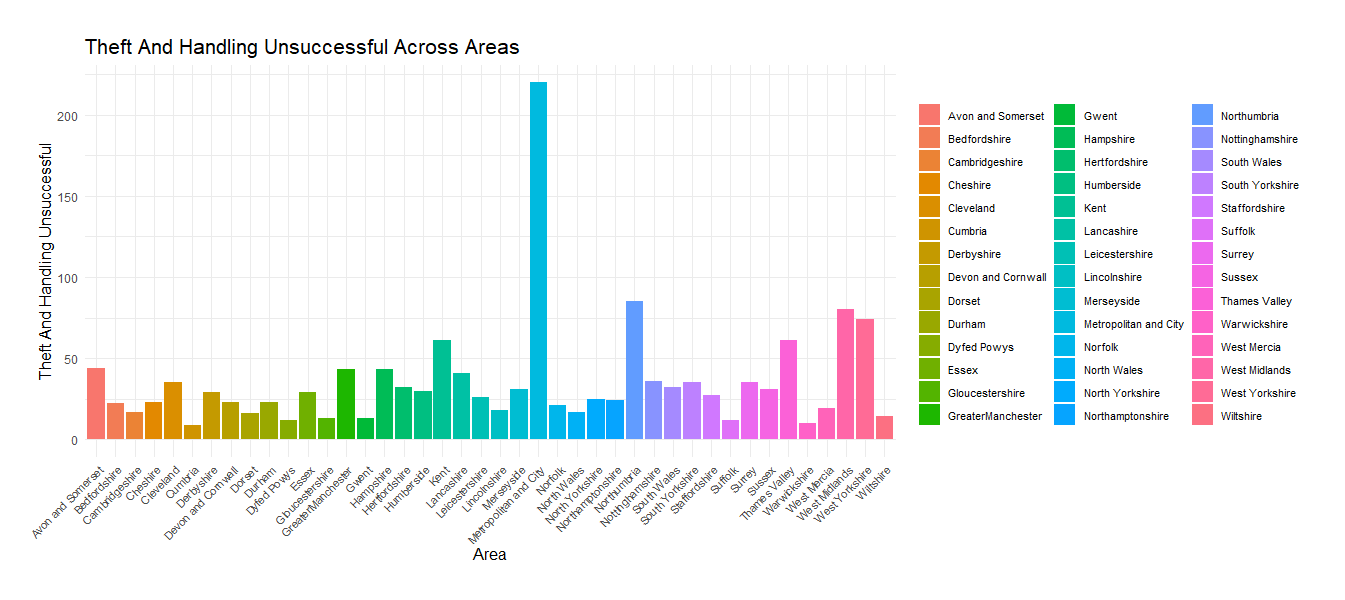


Figure 3.30

Figure 3.30 indicates the number of Theft And Handling convictions (Unsuccessful) across different counties in England and Wales, from 2014 to 2016. The highest number of Theft And Handling convictions can be seen in the Metropolitan and City, which is more than 200 in numbers and more than two times higher than other areas.

3.10 Fraud And Forgery

3.10.1 Distribution of Fraud And Forgery vs Average all other offences over time

* Fraud And ForgerySuccessful

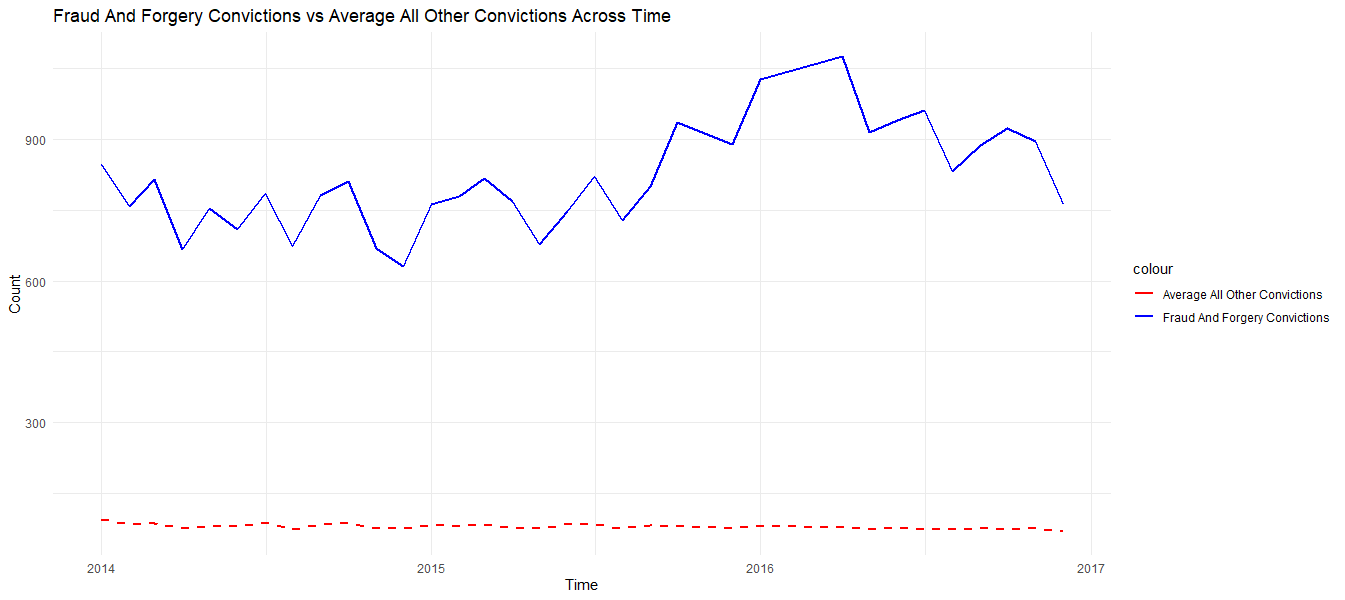


Figure 3.31

This data shows a distribution fraud and forgery convictions successful from 2014 to 2016, compared to average gneral criminal convictions. There can be seen an increase trend in fraud and forgery convictions throughout the years with some seasonal fluctuations. At the end of 2014 the particular conviction type has hit the lowest point while in the first quarter of 2016 hitting the highest rate.

* Fraud And Forgery Unsuccessful

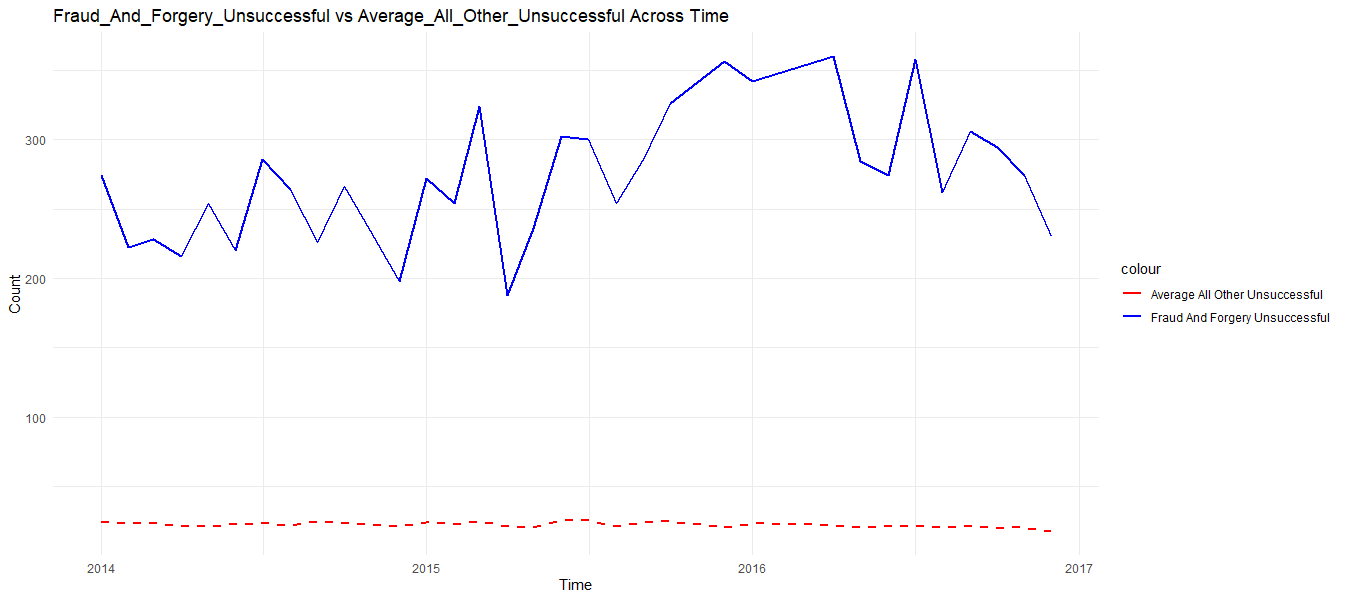


Figure 3.32

This data shows a distribution fraud and forgery convictions Unsuccessful from 2014 to 2016, compared to average gneral criminal convictions. The numbers have reached to the highest in the begning of 2016 which is approximately 350.

3.10.2 Distribution of Fraud And Forgery Across the Area

* Fraud And ForgerySuccessful

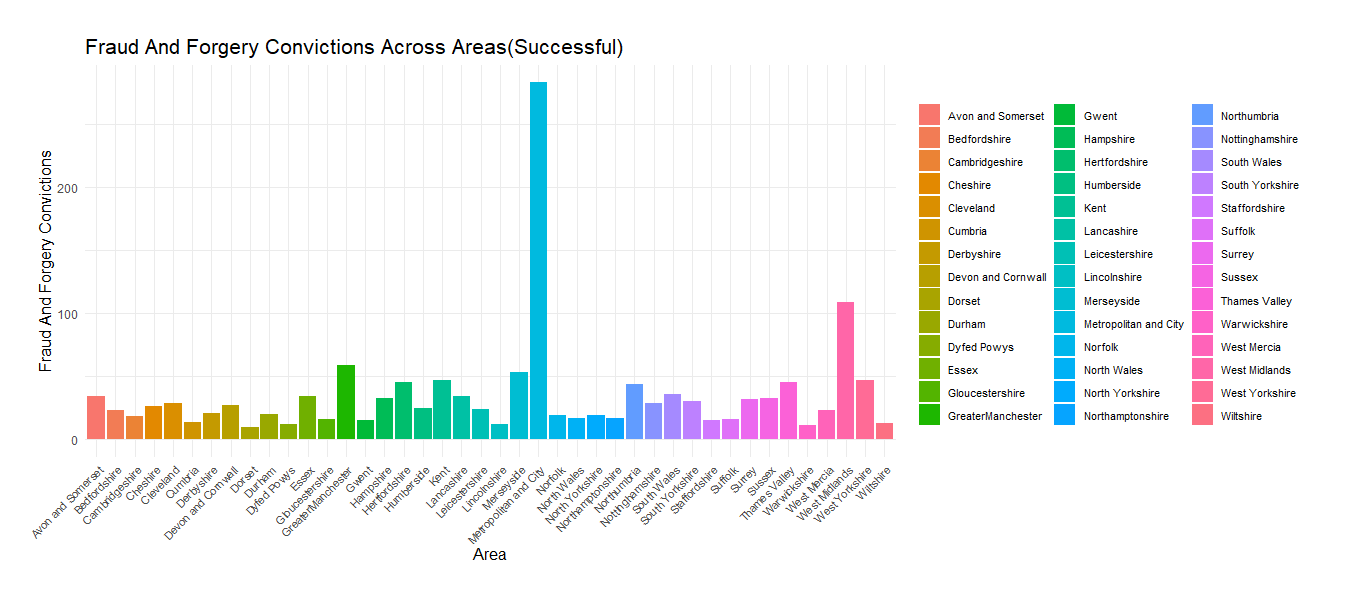


Figure 3.33

Figure 3.33 indicates the number of Fraud And Forgery convictions (Successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City shows the highest numbers in here as well. Which is approximately 300 and three times higher than other counties.

* Fraud And ForgeryUnsuccessful

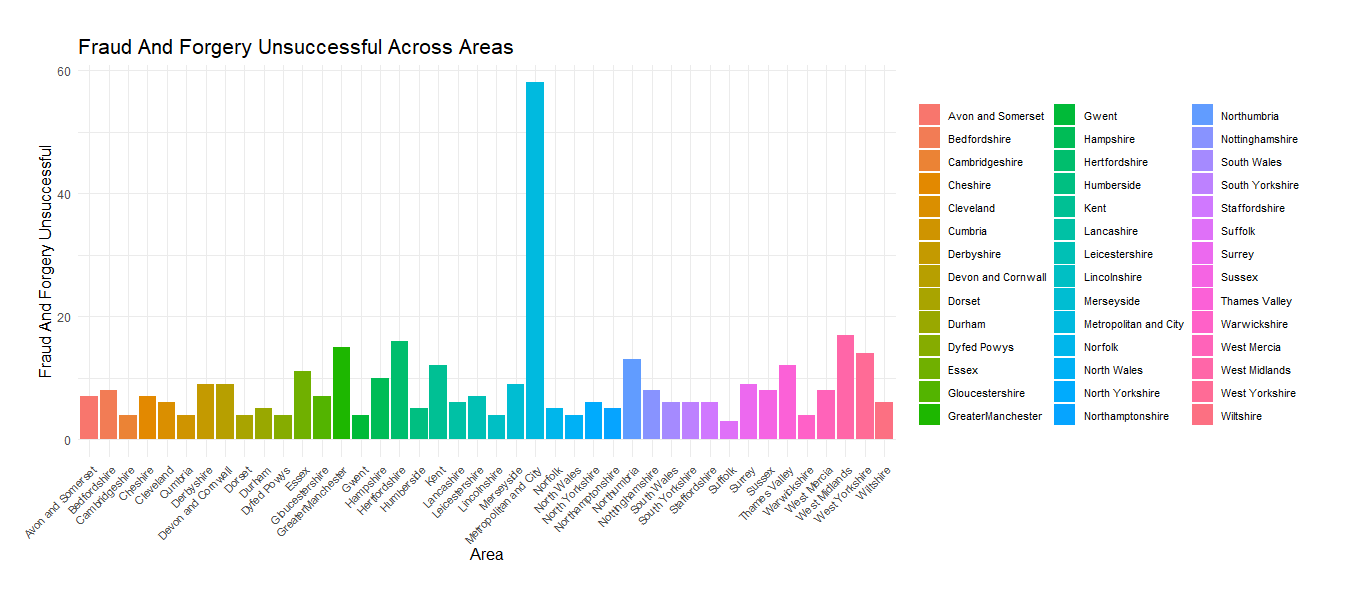


Figure 3.34

Figure indicates the number of Fraud And Forgery convictions (Unuccessful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City shows the highest numbers.

3.11 Criminal Damage

3.11.1 Distribution of Criminal Damage vs Average all other offences over time

* Criminal Damage Successful

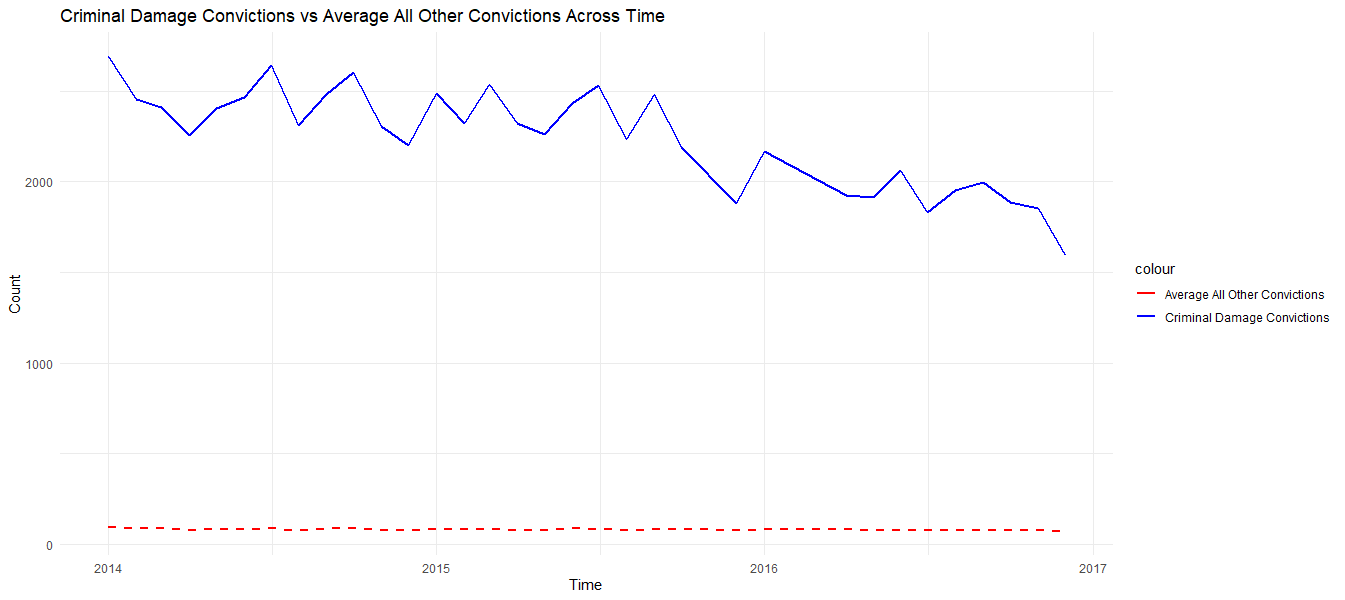


Figure 3.35

The figure shows the number of Criminal Damage Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decrease trend of Criminal Damage offences while the average number of all other convictions shows constant trend line. It indicates more than 2500 cases of Criminal Damage in the begning of 2014 and it has reduced to 1500 nearly

* Criminal Damage Unsuccessful

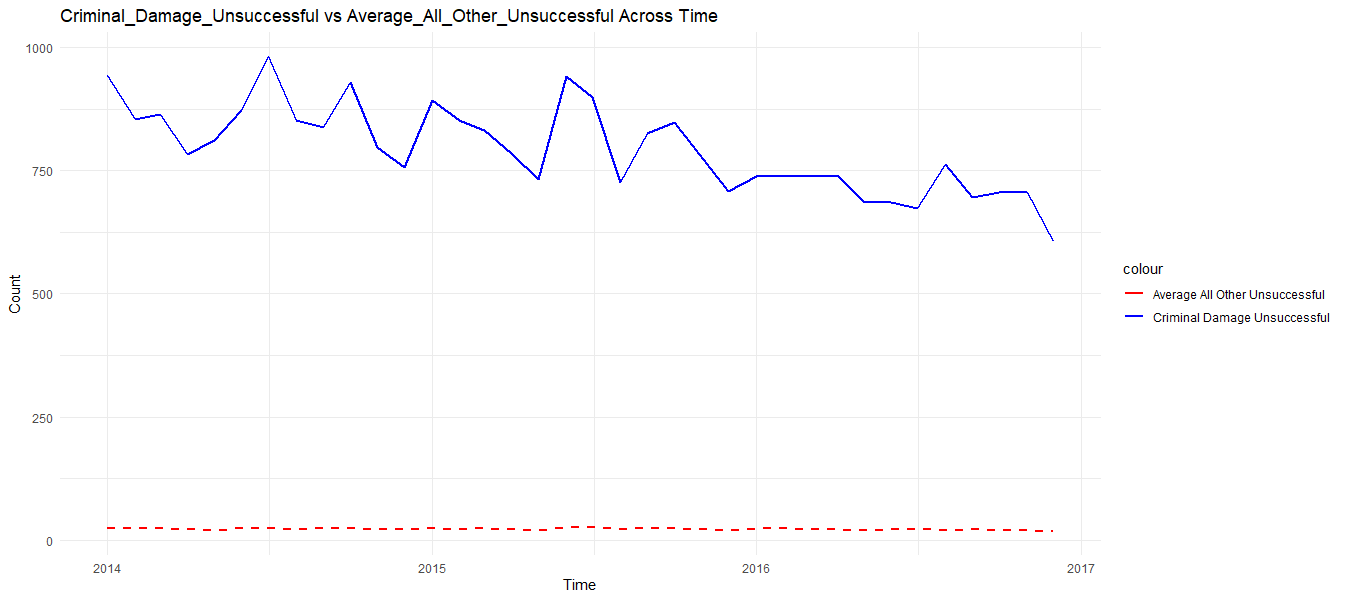


Figure 3.36

The figure shows the number of Criminal Damage Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. The numbers fluctuate between 500 and 1000 throughout years.

3.11.2 Criminal Damage Across the Area

* Criminal Damage Successful

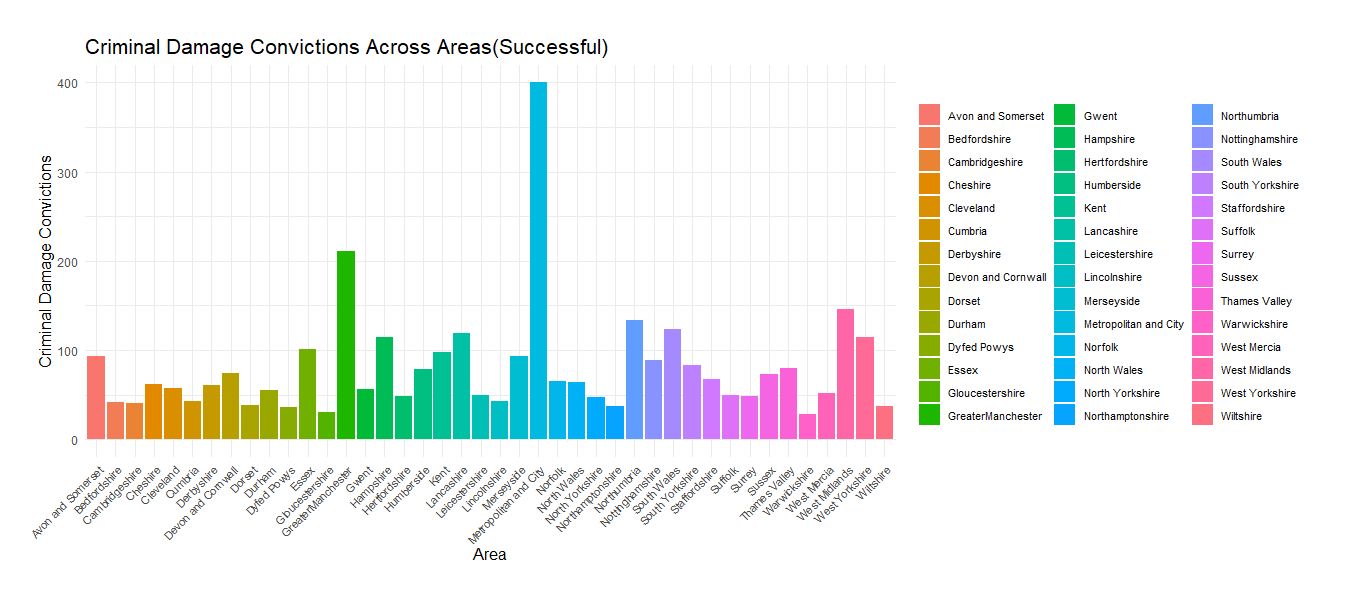


Figure 3.37

Figure indicates the number of Criminal Damage convictions (Successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City shows the highest numbers and the second highest is Greater Manchester. (400 and 200 ).

* Criminal Damage Unsuccessful

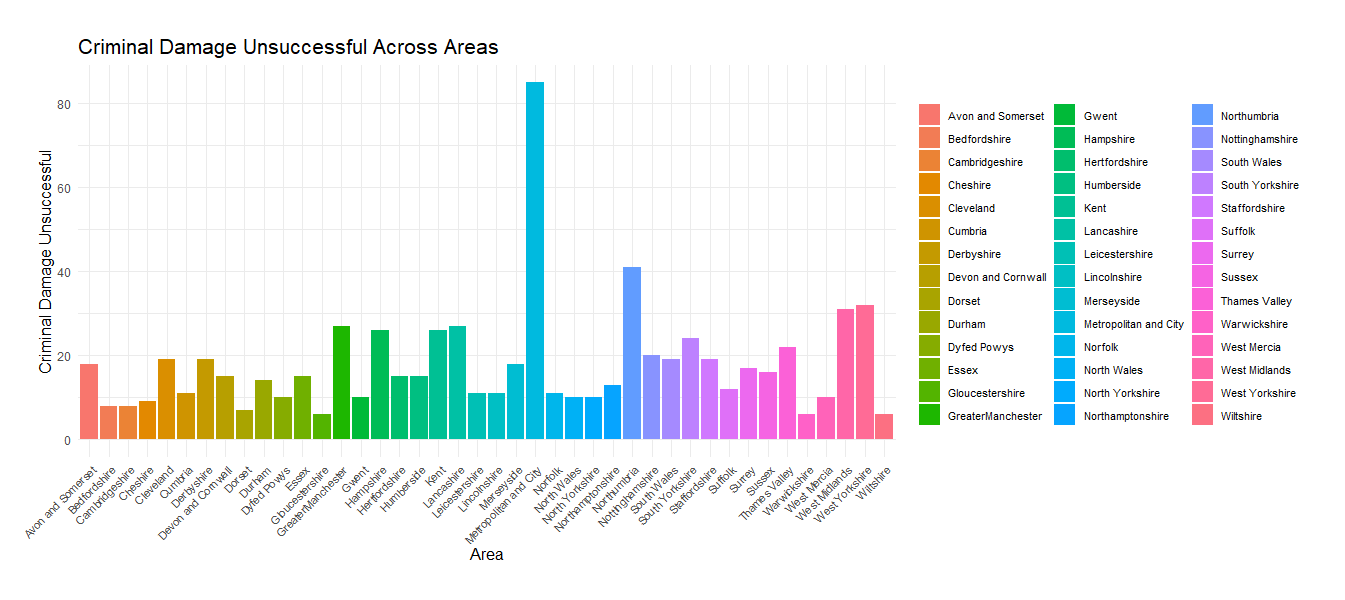


Figure 3.38

Figure indicates the number of Criminal Damage convictions (Unuccessful) across different counties in England and Wales, from 2014 to 2016. The highest number of Theft And Handling convictions can be seen in the Metropolitan and City, which is more than 80 in numbers and more than two times higher than other areas.

3.12 Drugs Offences

3.12.1 Distribution of Drugs Offences vs Average all other offences over time

* Drugs Offences Successful

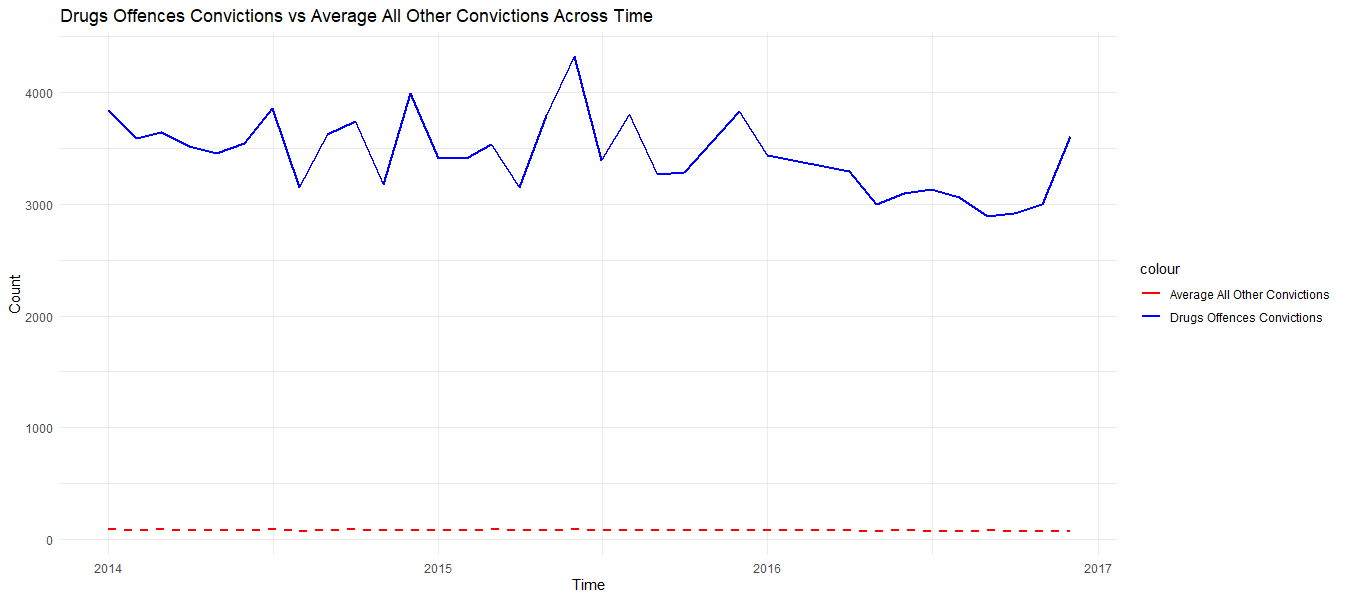


Figure 3.39

The figure shows the number of Drugs Offences Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There is no significant fluctuation can be seen in the category of Drugs Offences throughout years. The line fluctuates between 2700 and 4500.

* Drugs Offences Unsuccessful

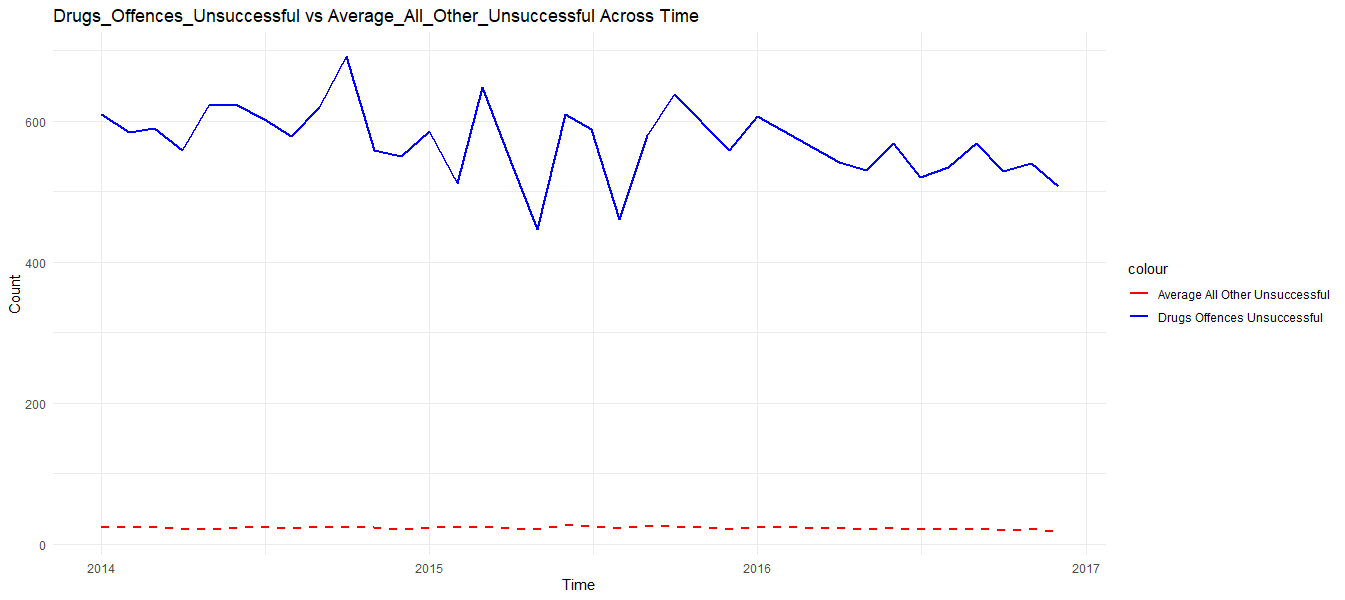


Figure 3.40

The figure shows the number of Drugs Offences Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a stable trend throughout years with some significant fluctuations around 2015.

3.12.2 Distribution of Drugs Offences Across the Area

* Drugs Offences Successful

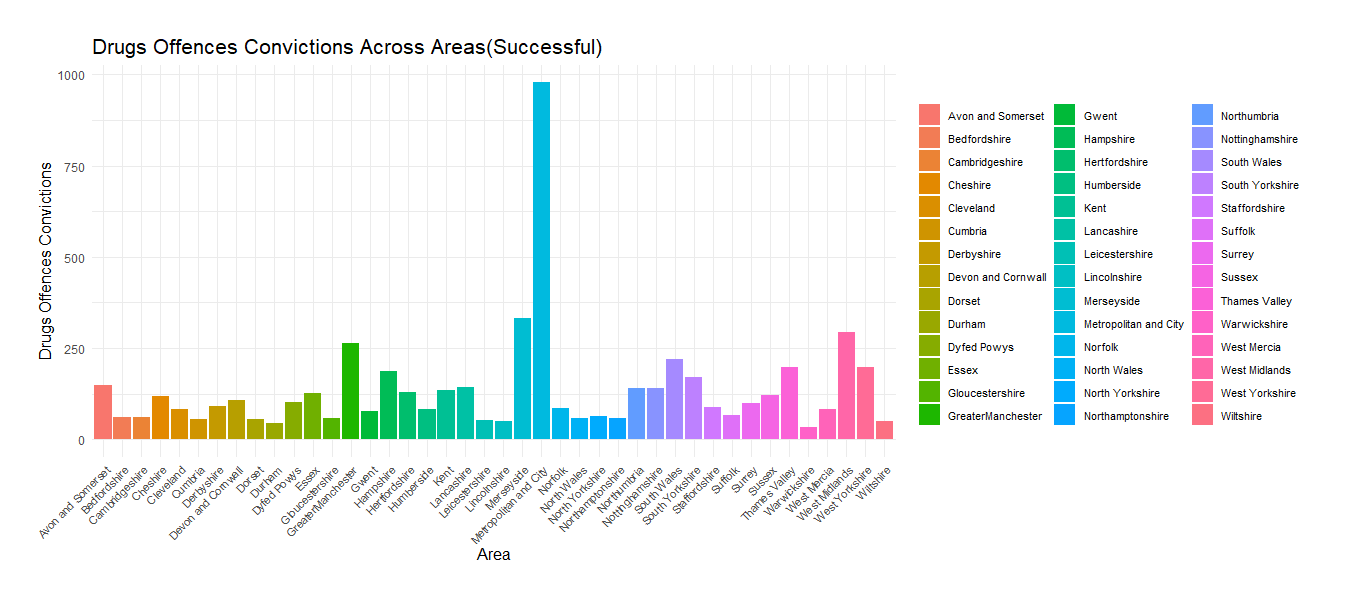


Figure 3.41

Figure indicates the number of Drugs Offences convictions (Successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City has the highest number of Drugs Offences convictions and which is around 1000 in numbers.

* Drugs Offences Unsuccessful

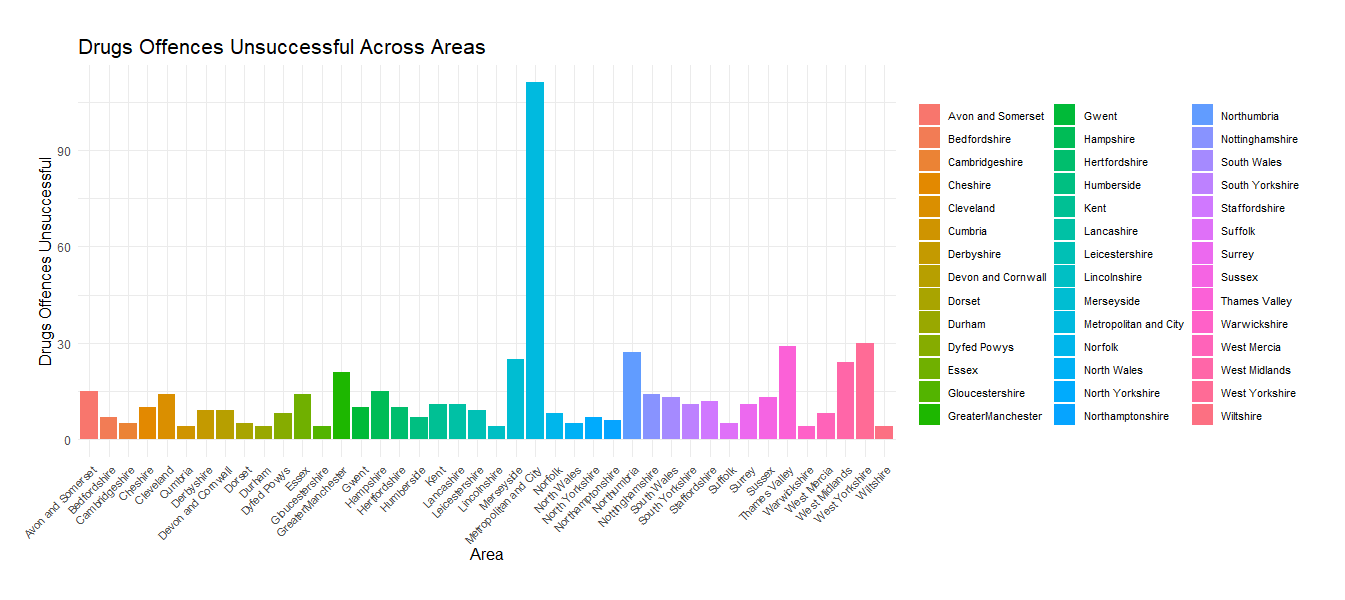


Figure 3.42

Figure indicates the number of Drugs Offences convictions (Unsuccessful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City has the highest number of Drugs Offences convictions and which is around 100 in numbers and three times higher than other counties

3.13 Public Order Offences

3.13.1 Distribution of Public Order Offences vs Average all other offences over time

* Public Order Offences Successful

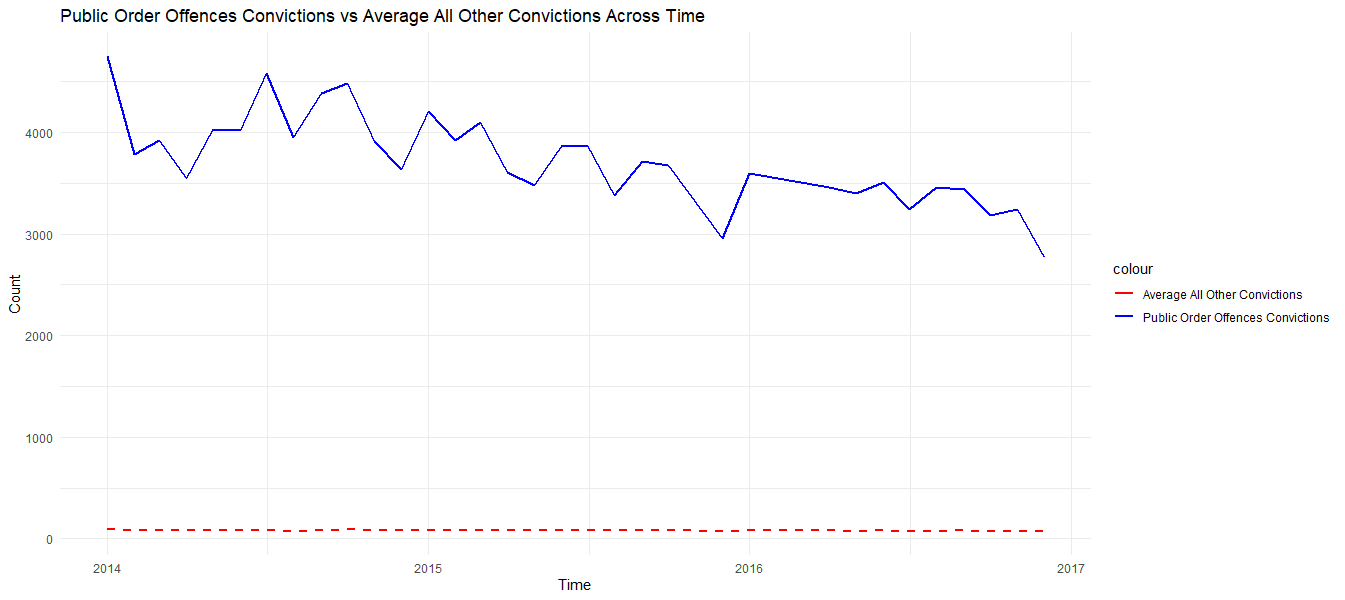


Figure 3.43

The figure shows the number of Public Order Offences Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decrease trend of Public Order Offences while the average number of all other convictions remains constant trend line. The highest point indicates more than 4500 cases of Public Order Offences in the begning of 2014 and it has reduced to 2500 at the end of the 2016.

* Public Order Offences Unsuccessful

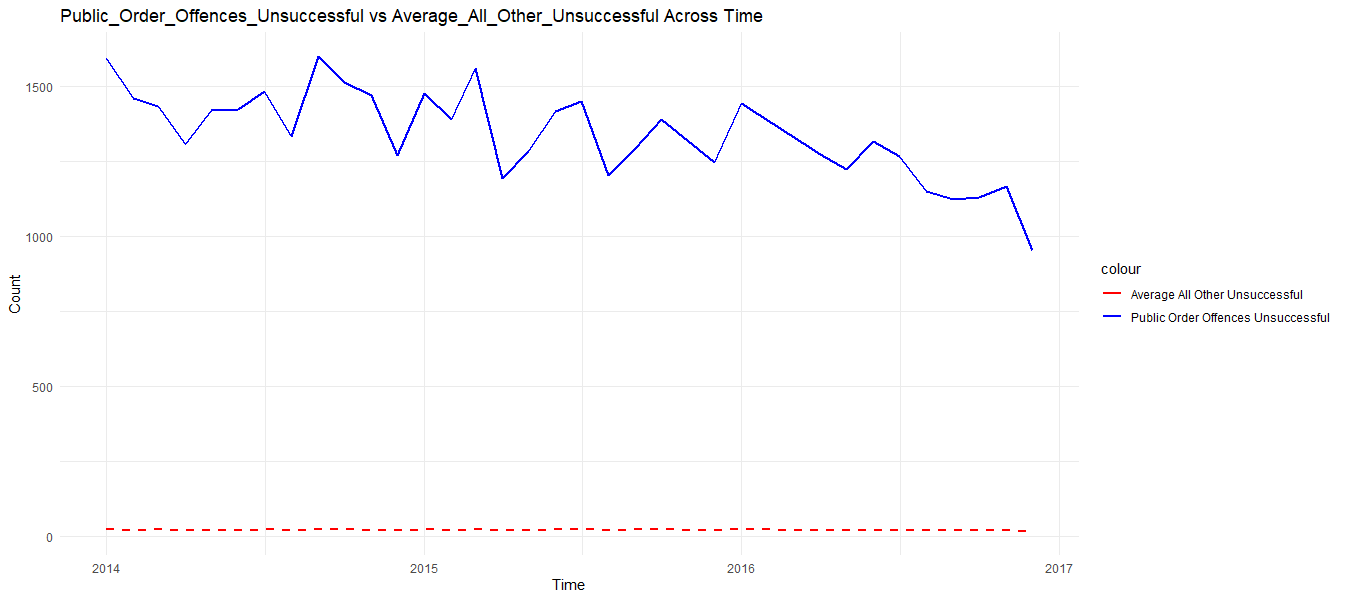


Figure 3.44

The figure shows the number of Public Order Offences Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. A decreasing trend can be identified from the above graph and the number of public order offences varies between 1000 and 1600 approximately.

3.13.2 Distribution of Public Order Offences Across the Area

* Public Order Offences Successful

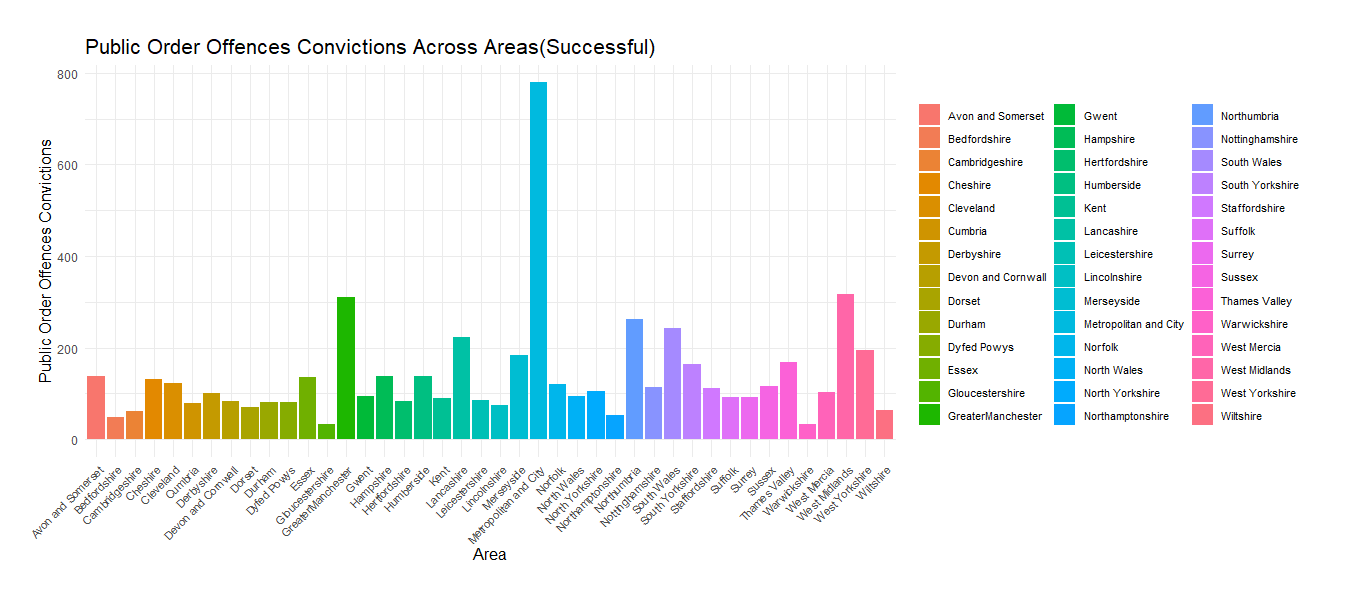


Figure 3.45

Figure indicates the number of Public Order Offences (successful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City has the highest number of Public Order Offences while Grater Manchester and West midlands shows approximately equal number of successful public order offence.

* Public Order Offences Unsuccessful

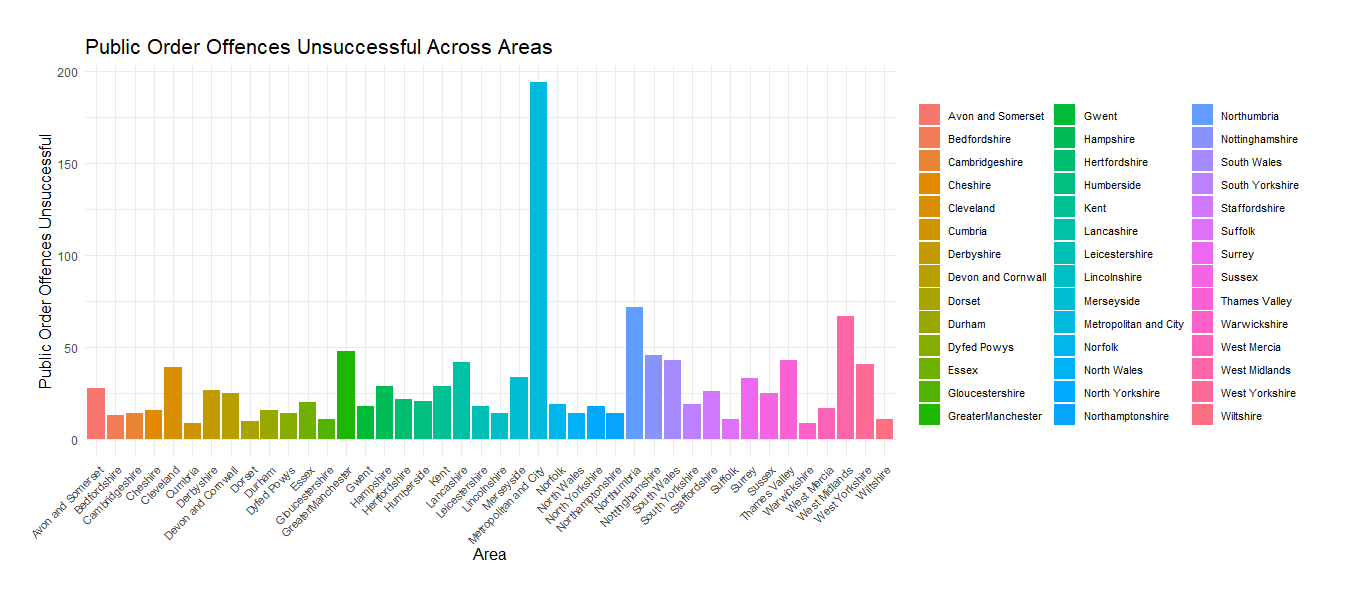


Figure 3.46

Figure indicates the number of Public Order Offences (Unsuccessful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City has the highest number of Unsuccessful Public Order Offences, which is approximately 200 in numbers.

3.14 Motoring Offences

3.14.1 Distribution of Motoring Offences vs Average all other offences over time

* Motoring Offences Successful

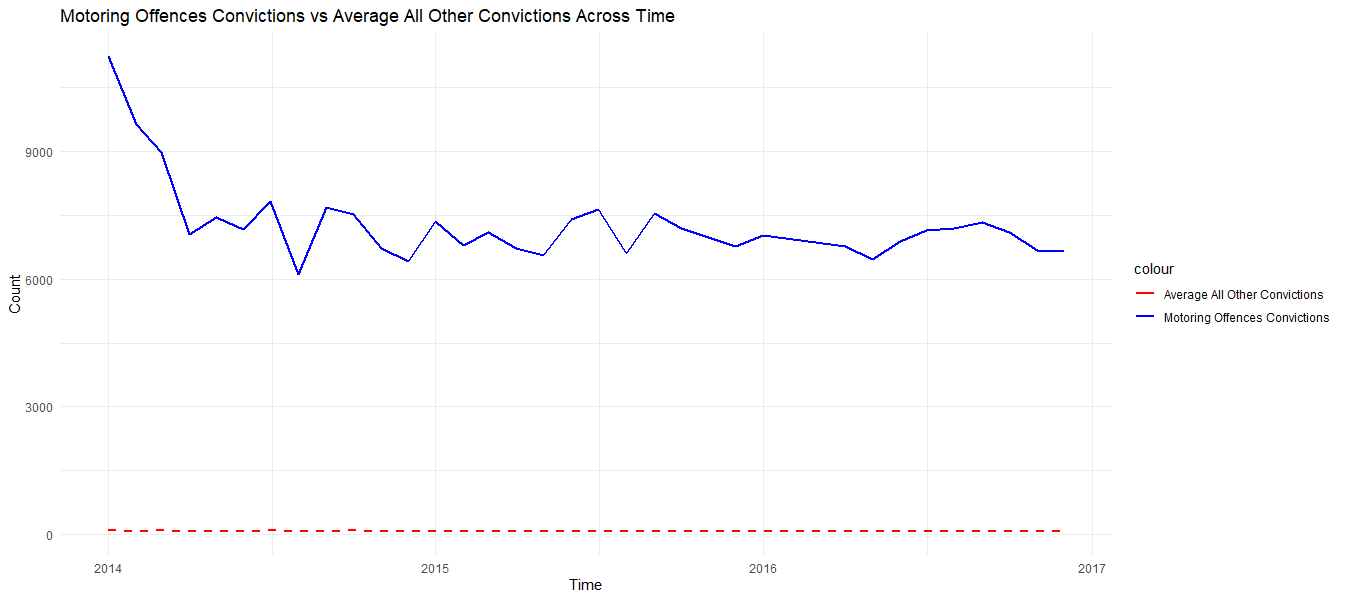


Figure 3.47

The figure shows the number of Motoring Offences Successful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a significant decrease in trend of Motoring Offences in the first quarter of 2014 after that it shows a steady trend until the end of the 2016 and it fluctuates between 6000 and 7500 approximately.

* Motoring Offences Unsuccessful

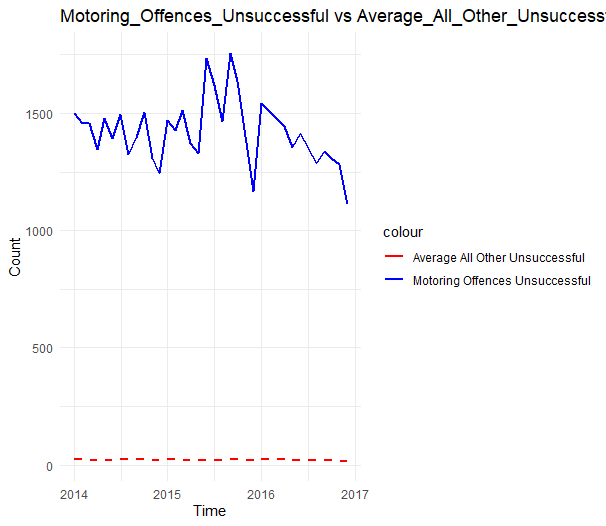


Figure 3.48

The figure shows the number of Motoring Offences Unsuccessful vs the average number of criminal offences of all other criminal offence categories in England and Wales, from 2014 to 2016. There can be seen a decreasing trend with some significant fluctuations in 2015. The numbers have reached to the highest in 2015 which is 1750 approximately.

3.14.2 Distribution of Motoring Offences Across the Area

* Motoring Offences Successful

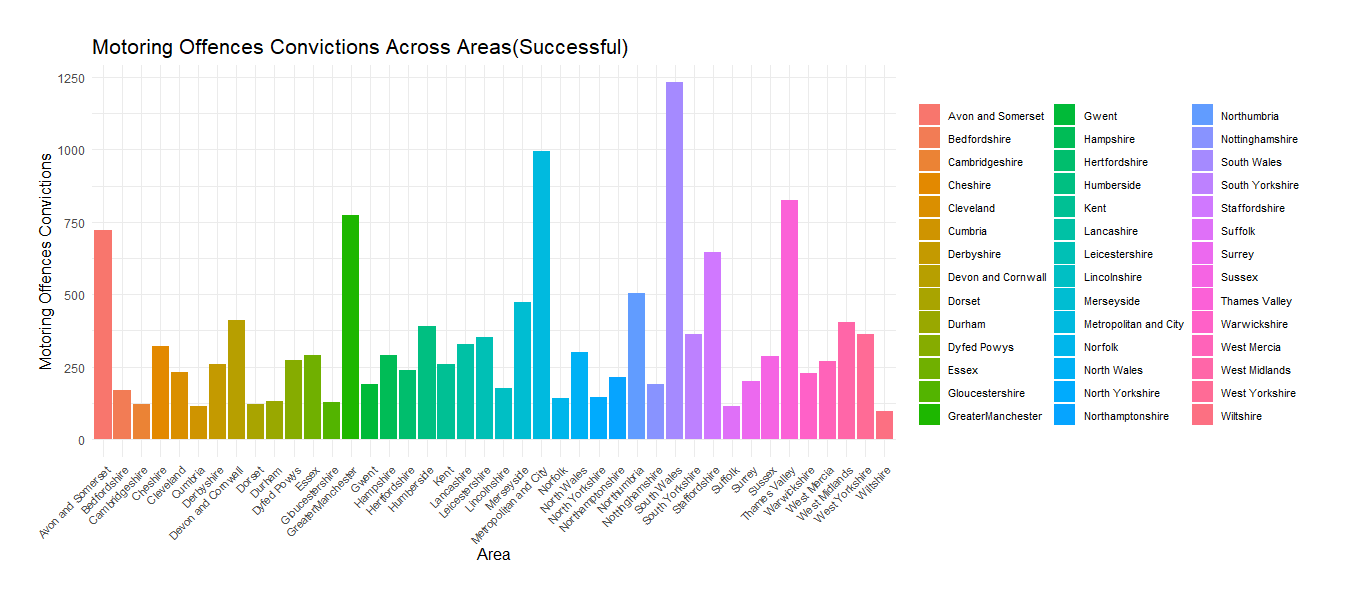


Figure 3.49

Figure 3.49 indicates the number of Motoring Offences (successful) across different counties in England and Wales, from 2014 to 2016. South Wales shows the highest number of successful Motoring Offences which is about 1250 in numbers and the second highest represent Metropolitan and City.

* Motoring Offences Unsuccessful

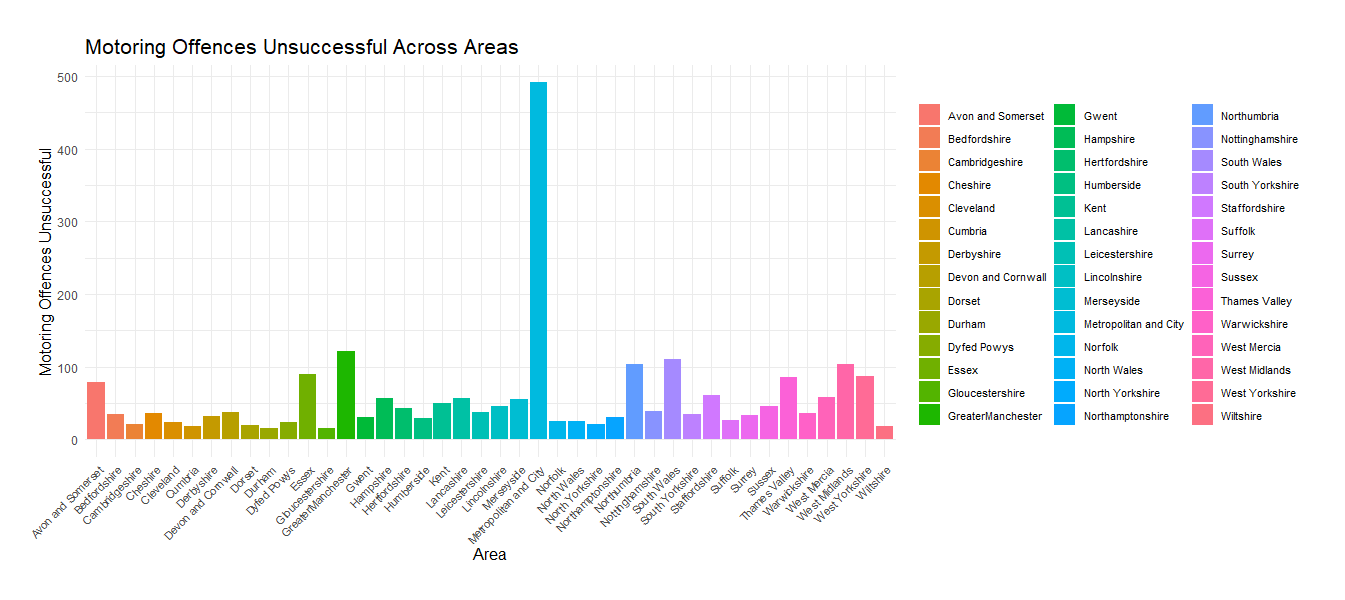


Figure 3.50

Figure indicates the number of Motoring Offences (Unsuccessful) across different counties in England and Wales, from 2014 to 2016. Metropolitan and City shows the highest number of unsuccessful Motoring Offences (500).

3.15 Total Criminal Offence Convictions by Area

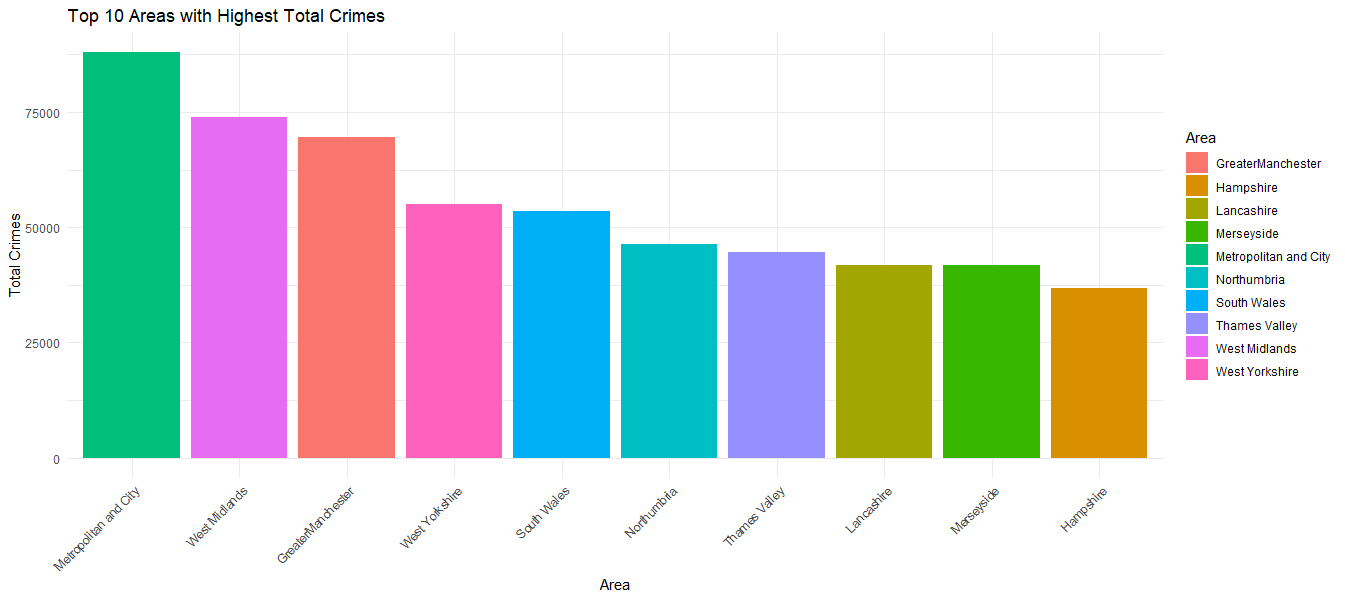


Figure 3.51

Figure 3.51 indicates the top 10 countries which has highest number of overall criminal convictions. Metropolitan and City has the highest numbers while West Midlands shows the lowest.

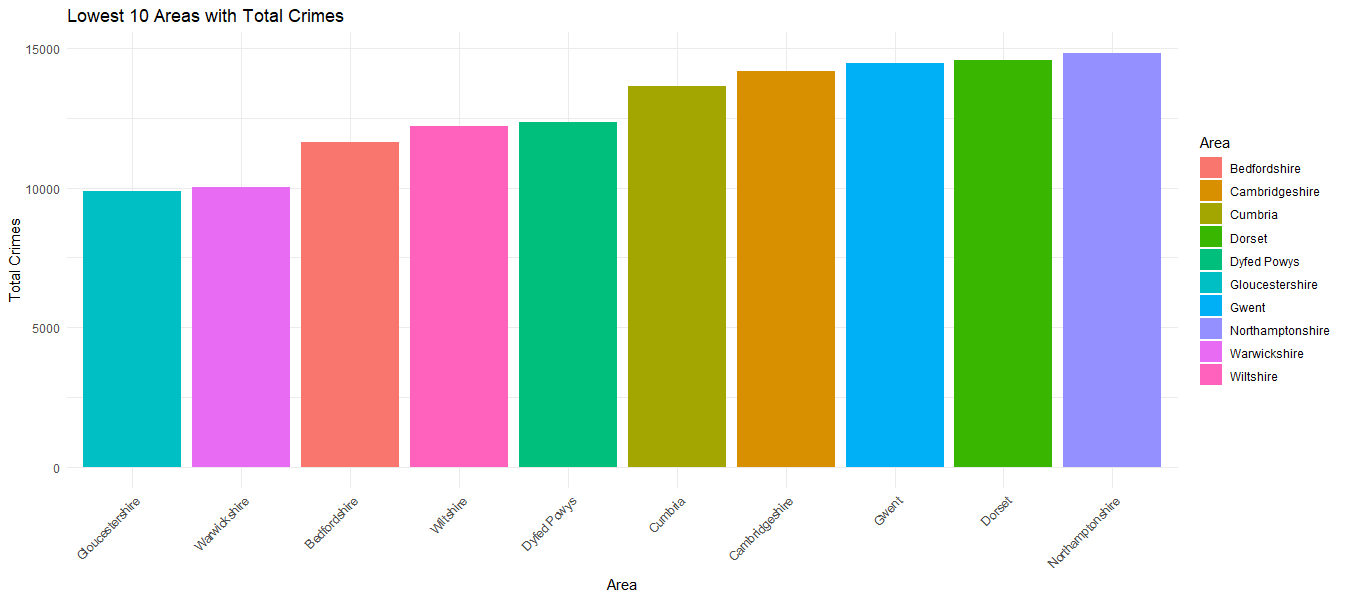


Figure 3.52

Figure 3.52 indicates the top 10 countries which has lowest number of overall criminal convictions throughout the years. Gloucestershire and Warwickshire show the lowest numbers which is 10 000 in numbers.

4 Hypothesis

4.1 Hypothesis Development

A hypothesis is a statement that expresses the researcher's expectation or prediction regarding the relationship between study variables. The hypothesis is where the research process begins and finishes. It is crucial to the entire procedure because it is at the heart of it. The heart of the investigation is the hypothesis. Without a theory, research cannot provide adequate service (Dayanand, 2018)

This report indicates the following hypothesis for further analysis

**H0: There is no significant relationship between Time and Criminal Offences (Success)**

H1: There is a significant relationship between Time and Homicide Convictions

H2: There is a significant relationship between Time and Offences Against the Person Convictions

H3: There is a significant relationship between Time and Sexual Offences Convictions

H4: There is a significant relationship between Time and Burglary Convictions

H5: There is a significant relationship between Time and Robbery Convictions

H6: There is a significant relationship between Time and Theft and Handling Convictions

H7: There is a significant relationship between Time and Fraud and Forgery Convictions

H8: There is a significant relationship between Time and Criminal Damage Convictions

H9: There is a significant relationship between Time and Drugs Offences Convictions

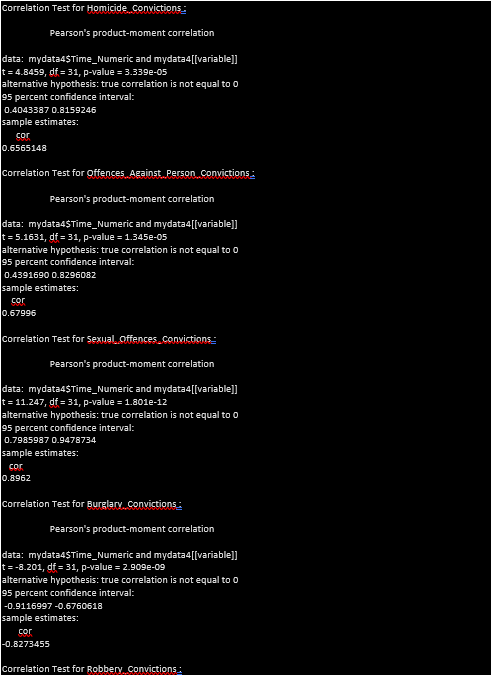
H10: There is a significant relationship between Time and Public Order Offences Convictions

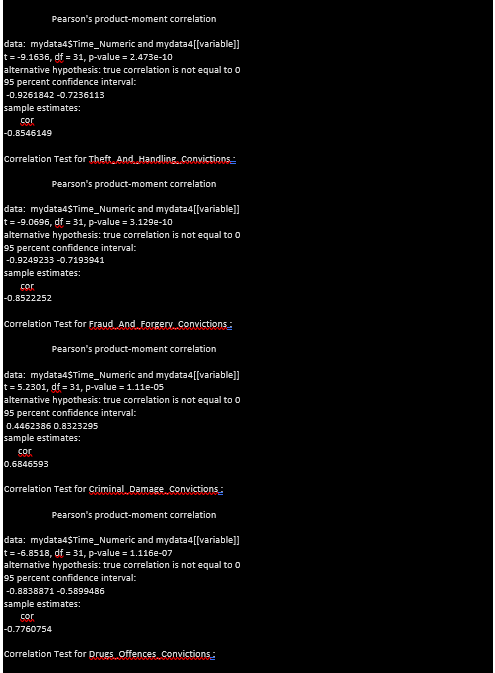
H11: There is a significant relationship between Time and Motoring Offences Convictions

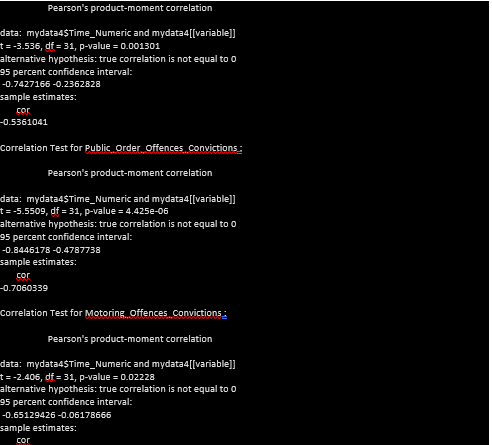
4.2 Hypothesis Testing

**• Pearson Correlation Analysis**

The Pearson correlation coefficient, with a value ranging from - 1 to 1, is used to calculate the linear correlation between two variables. The numerical value is 1, suggesting that the straight-line equation can represent two positively related variables. The coefficient value of -1 implies that the two variables are linearly stated, yet they are negatively related. (e, 1948). Pearson's correlation coefficient is a test statistic for determining the statistical relationship (or association) between two continuous variables. Because it is based on the covariance approach. It reveals the degree and direction of the association, or correlation. The following hypotheses were utilized to determine the relationship, as well as the decision rule that was used to make the final decisions.







|  |  |  |
| --- | --- | --- |
| Variable | P Value | Conclusion |
| Homicide Convictions | 3.339e-05 | There is a significant positive correlation between Time and Homicide Convictions |
| Offences Against the Person Convictions | 1.345e-05 | There is a significant positive correlation between Time and Offences Against the Person |
| Sexual Offences Convictions | 1.801e-12 | There is a significant positive correlation between Time and Sexual Offences |
| Burglary Convictions | 2.909e-09 | There is a significant negative correlation between Time and Burglary Offences |
| Robbery Convictions | 2.473e-10 | There is a significant negative correlation between Time and Robbery |
| Theft and Handling Convictions | 3.129e-10 | There is a significant negative correlation between Time and Theft and Handling |
| Fraud and Forgery Convictions | 1.11e-05 | There is a significant positive correlation between Time and Fraud and Forgery |
| Criminal Damage Convictions | 1.116e-07 | There is a significant negative correlation between Time and Criminal Damage |
| Drugs Offences Convictions | 0.001301 | There is a significant negative correlation between Time and Drugs Offences Convictions |
| Public Order Offences Convictions | 4.425e-06 | There is a significant negative correlation between Time and Public Order Offences Convictions |
| Motoring Offences Convictions | 0.02228 | There is a significant negative correlation between Time and Motoring Offences |

➢ Decision rule

If the P value is less than alpha level (95%), then we have enough evidence to reject null hypothesis. (P<0.05)

➢ Decision

The P values of all the variables are less than 0.05, which indicates that all the variables are significant. Homicide Convictions, Offences Against the Person Convictions, and Sexual Offences Convictions are positively correlated with time. This suggests an increasing trend over time for these types of offenses. Variables like Burglary Convictions, Robbery Convictions, Theft and Handling Convictions, Drugs Offences Convictions, Public Order Offences Convictions, and Motoring Offences Convictions show a negative correlation with time. This implies a decreasing trend over time for these types of offenses.

5 Predictive Modeling

The process of constructing and applying mathematical models or algorithms to make predictions or forecasts about future events or outcomes based on historical data is known as predictive modeling. The main goal is to discover patterns, relationships, or trends in data that can be used to forecast or estimate the outcome of interest for new or unknown data. Understanding the mechanics of predictive modeling is beneficial for diagnosing and improving performance (Lawton, n.d.).

5.1 data preparation for model

In order to run predictive model, one variable called “Sum Convictions” was added to the

data frame by getting the summation of all the criminal offence categories per each month.

5.2 Linear Regression

5.2.1 Assumptions Checking

* Linearity

In linear regression it assumes there is a linear relationship between independent and dependent variable. Scatter plot can be used to determine this relationship. The primary assumption of the linear regression model, as the name suggests, is that the dependent and independent variables have a linear relationship. Only the parameters are linear in this case. Surprisingly, there are no such constraints on the extent or shape of the explanatory variables themselves ( Vivek Krishnamoorthy and Udisha Alok, 2022 ).

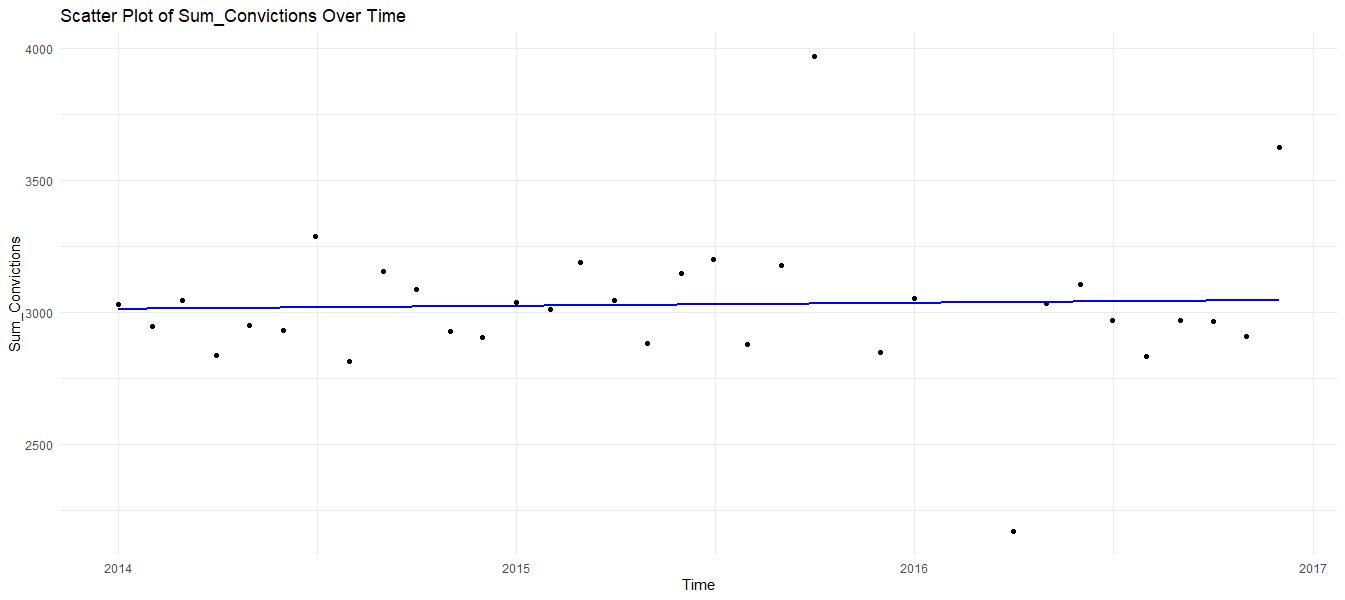


Figure 5.1

The above figure indicates that there is an approximate linear relationship between Total criminal convictions and time.

* Normality of Residuals

Under linear regression its assumes that residuals are normally distributed. histogram provides a visual representation of the distribution of residuals.

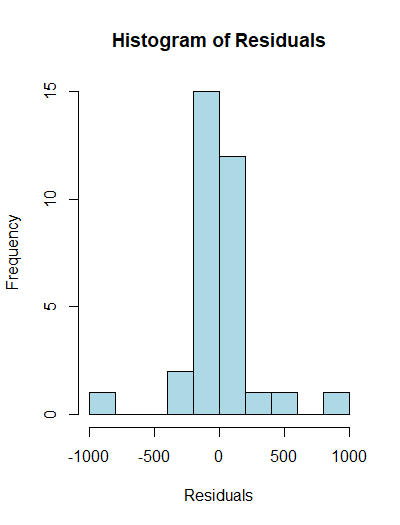


Figure 5.2

The figure indicates that residuals are normally distributed.

5.2.2 Regression Analysis

With the use of regression analysis and coefficient value analysis, a function with their respective beta values can be developed. The following is an illustration of the function:

Y0 = β0 + β1X

The part of the equation is indicated by the letters β0 + βXi. The independent variable is denoted by the letter X. The intercept is defined by β0 , or the value that Y takes when X is zero, is represented by the symbol β0. The slope of a line is represented by the symbol β. It indicates how many units Y changes when X changes one unit

Residuals:

Min 1Q Median 3Q Max

-868.16 -116.12 -12.04 66.76 937.20

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.516e+03 2.500e+03 1.006 0.322

Time 3.094e-02 1.507e-01 0.205 0.839

Residual standard error: 280.1 on 31 degrees of freedom

Multiple R-squared: 0.001358, Adjusted R-squared: -0.03086

F-statistic: 0.04215 on 1 and 31 DF, p-value: 0.8387

Based on above data the suggested regression model is

Total Criminal Convictions = 2516 + 0.031Time + ϵ

The estimated relationships between the independent and the dependent variables are represented by coefficients of a linear regression model. The intercept (β0) in a regression model represents the value of the response variable when all of the predictor variables in the model are equal to zero. Here it's the predicted value of 'Total Convictions' at the starting point of the period. when the independent variable equals zero Total Convictions is 2516.

The coefficient of independent variable (Time) indicates the estimated change in the dependent variable ('Sum\_Convictions') in response to a one-unit change in the independent variable ('Time'). In this situation, it implies that total Convictions increases by 0.031 units on average for each additional unit of time. The p-value (0.839) associated with the calculated coefficient for 'Time' indicates whether it is statistically significant. The high p-value indicates that the coefficient is not substantially different from zero in this example, showing that 'Time' is not a significant predictor of Total Convictions. Furthermore, The F-ratio yield efficient model has a value larger than 1. The value in the table above is 0.04215. A low F-statistic and a high p-value indicate that the model may be statistically significant for predictions.

The value of R, the correlation coefficient, can be thought of as a metric for the accuracy of the dependent variable's prediction; in this case, a value of Multiple R-squared is 0.001358, which means the proportion of the variance in the total convictions or dependent variable explained by the time or independent variable. Here, the model explains only a very small fraction of the variance. Also, the adjusted R square is -0.03086.

In summary, the linear regression model does not provide strong evidence for a significant relationship between independent ('Time') and dependent variables ('Sum\_Convictions') The model's low explanatory power and non-significant coefficients indicate that the chosen variables may not be effective predictors of total criminal convictions.

5.3 Clustering

Clustering is defined as the grouping of items in which little or no knowledge about the object relationships in the given data exists (Gbeminiyi John Oyewole & George Alex Thopil, 2022)

5.3.1 Determining Optimal Number of Clusters

The Elbow Method was utilized to determine the optimal number of clusters for the analysis. According to this method, within-cluster sum of squares reduces as the number of clusters grows. An examination of the Elbow Method plot reveals that the best number of clusters are four.

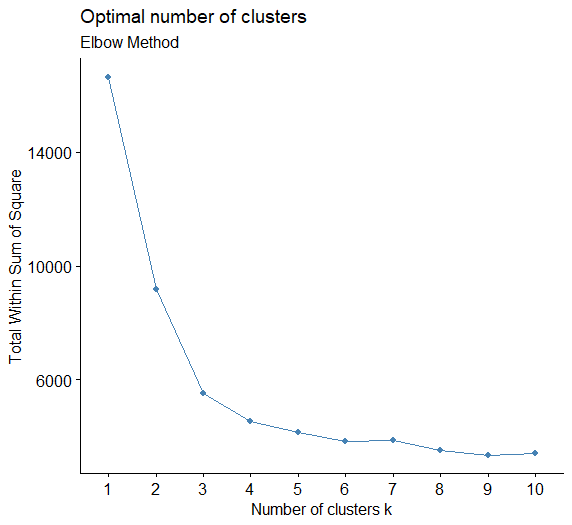
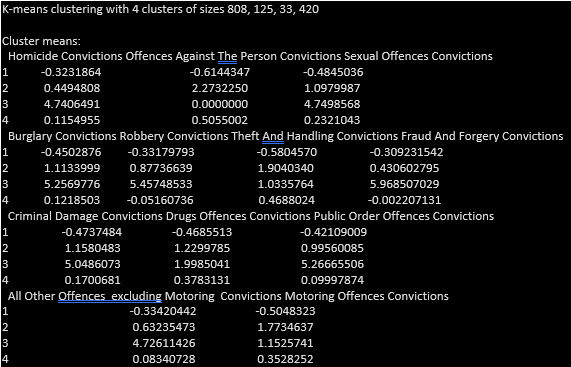


Figure 5.3

5.3.2 K-Means Clustering

The k-means clustering technique discovered four unique clusters. In terms of crime-related characteristics, each cluster differs. The cluster centers represent the average values for each variable inside the relevant clusters.



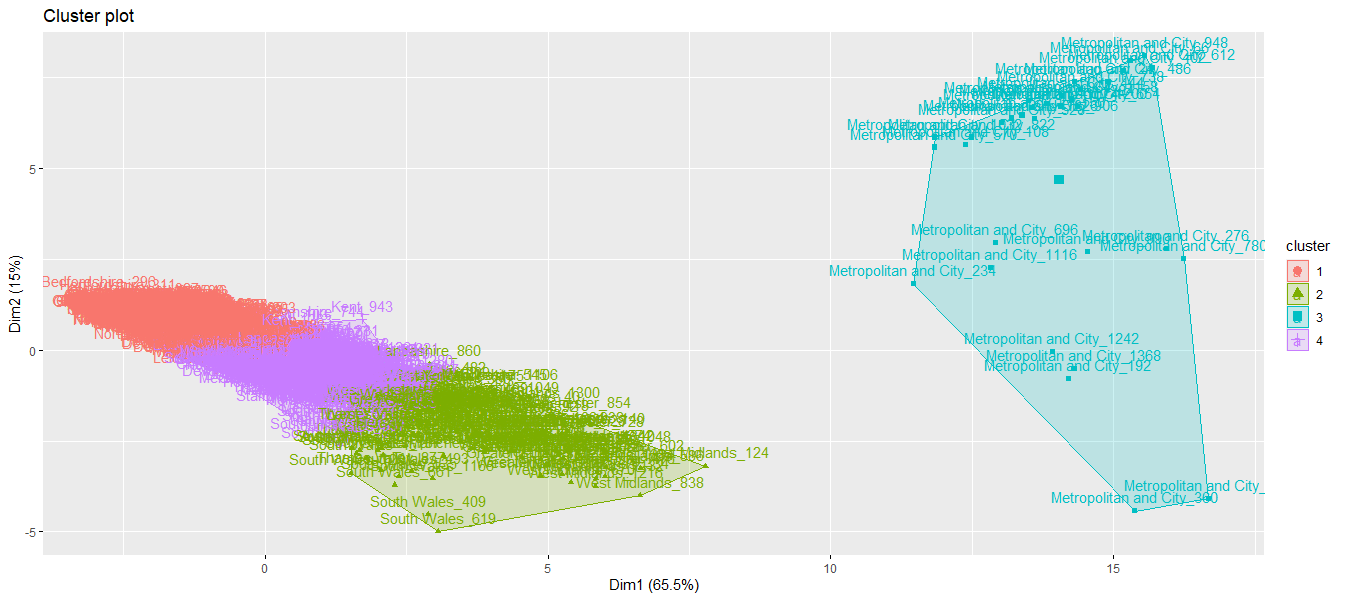


Figure 5.4

According to the above figure, data has been divided into four clusters: 808 (biggest), 125, 33 (smallest), and 420. Across crime kinds, each cluster reveals diverse trends. Cluster 1 is identified as relatively lower rates across all crime types, except motoring offences and Southwest Regions (e.g., Avon and Somerset, Devon and Cornwall): These regions are primarily found in Cluster 1, implying that crime patterns are similar. Cluster 2 has higher rates of Offenses Against the Person, Sexual Offenses, Robbery, Theft and Handling, and Motoring Offenses, indicating regions that may need focused interventions. Cluster 3 stands out for having very high rates of homicide, sexual offences, burglary, robbery, fraud and forgery, and public order offenses, areas like Metropolitan and City are uniquely represented in Cluster 3, showing specific crime characteristics in this location. potentially indicating regions that require immediate action. Cluster 4 depicts locations with moderate rates of crime, which are generally lower than Cluster 2 but higher than Cluster 1. Also, Greater Manchester is spread across Clusters 1 and 2, showcasing heterogeneity in crime profiles within the region.

5.3.3 Making Predictions Using a Clustering Model

When a clustering model is built on a dataset, it can be used to predict additional, previously unknown data. Based on its similarity to the existing cluster center points, each new data point is assigned to one of the current clusters. This includes preparing and preprocessing new data, predicting cluster assignments using the trained model, and analyzing the results. The clustering model predicts cluster assignments for each data point in the new dataset. The new data clustering can optionally be seen to see how it is distributed throughout the specified clusters.

5.4 Classification

Classification is a data mining technique that categorizes or classifies elements in a collection. The purpose of classification is to anticipate the correct target class for each case in the data. (V. Krishnaiah1\*, Dr.G.Narsimha 2, 2014). Through the use of an algorithm, classification analysis can be used to ask questions, make decisions, or forecast behavior. It works by generating a set of training data that includes a certain set of attributes as well as the expected outcome. The classification algorithm's task is to figure out how that set of qualities gets to its conclusion (Indicative, 2022). The goal of classification is to create a model that can accurately predict the class of unseen new instances.

5.4.1 Data Preprocessing

Initially, a new data frame was created by selecting specific variables from an original data frame. These variables indicate types of criminal convictions in various counties. The "Area" column is then turned into a factor, which is a categorical variable in R. In addition, the column "Robbery Convictions" is transformed into the numeric format. The column names of the new data frame are changed by replacing spaces with underscores, making them more useful for subsequent analysis. Thereafter The dataset was divided into two sections: train\_data, which contains the first 1000 rows, and test\_data, which contains the remaining rows. Training the model on one subset and evaluating its performance on another is a frequent method in machine learning.

5.4.2 Model Training

Random Forest is a collective learning approach that, during training, builds a large number of decision trees and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Here randomForest function is used to train a random forest model in the analysis. The area variable was considered as the target variable, and the model should be predicted using all other variables in the dataset. ntree = 500 indicates that there are 500 trees in the forest.

5.4.3 Model Prediction and Evaluation

Using the trained random forest method, the model has made predictions on the test set. After that the accuracy of the model has been calculated by comparing the predicted values with the actual values.

Accuracy : 0.6945

95% CI : (0.648, 0.7383)

No Information Rate : 0.0239

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6872

Mcnemar's Test P-Value : NA

With a Kappa statistic of 0.6872, the random forest model attained an accuracy of 69.45%, showing impressive agreement beyond chance. The accuracy 95% confidence interval ranged from 0.648 to 0.7383. The No Information Rate (NIR) was calculated to be 2.39%, which is the accuracy that would be obtained by always guessing the most common class. The p-value (2.2e-16), which compares the model's accuracy to the No Information Rate, was found to be significantly low, implying that the model's accuracy is significantly better than chance.

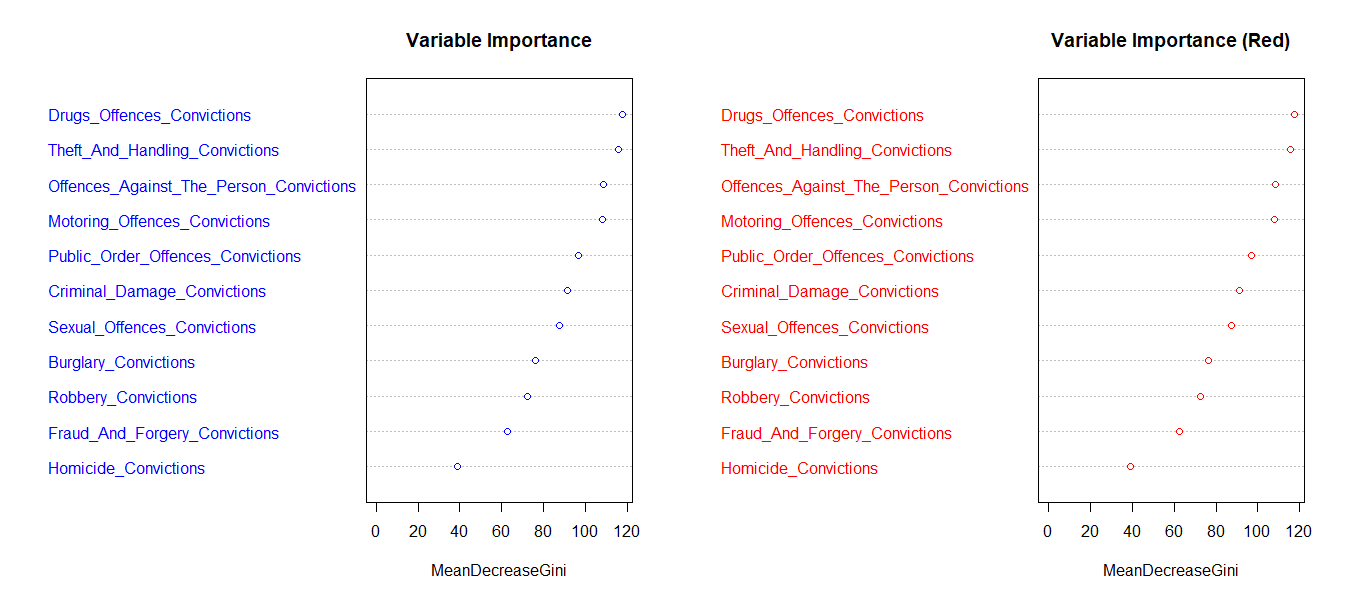


Figure 5.5

Based on the above figure, The variable importance for "Drugs Offences Convictions" is shown on the left side of the graph, while the variable importance for "Theft and Handling Convictions" is shown on the right side. Each variable's importance is displayed by a blue bar, and the red line depicts the mean drop in Gini for each variable.

Gini impurity is a measure of how well a collection of data items is classified. A lower Gini impurity suggests that the data points are more likely to be grouped. The average drop in Gini impurity that happens when a certain variable is used to separate the data is referred to as the mean decrease in Gini. The graph demonstrates that the most relevant variables for "Drug offences convictions" and "Theft and Handling Convictions" are the same. These categories include "Offences Against the Person Convictions," "Motoring Offences Convictions," as well as "Public Order Offences Convictions." This implies that these criteria are the most relevant in determining whether someone will be convicted of drug offences, theft, or handling offences.

Also, Statistics were provided by region, including sensitivity, specificity, positive predictive value, negative predictive value, prevalence, detection rate, detection prevalence, and balanced accuracy. These metrics provide a complete review of the model's performance across different classes, providing insights into both positive and bad examples.

5.4.4 Summary

Finally, the random forest model, trained with 500 trees and utilizing the "Area" variable as the target, displayed a considerable accuracy of 69.45% and a robust Kappa statistic of 0.6872, suggesting significant agreement beyond what would be predicted by chance. The model's accuracy exceeded the No Information Rate substantially, as demonstrated by a p-value of 2.2e-16. These findings highlight the random forest approach's effectiveness in forecasting the "Area" variable based on the given dataset.

6 Critical Reviews of The Data Analytics Tools and Techniques

* Descriptive Analysis

Descriptive Analysis provides a basis for data exploration and initial insights. It lacks the ability to forecast or determine relationships. Under descriptive analysis, the report includes mean, standard deviation, minimum and maximum values as summary statistics. The mean identified as the average or most common value in a set of numbers (Taylor, 2015). Provides a measure of central tendency, it is sensitive to extreme values, making it vulnerable to outliers. A standard deviation is a statistic that measures a dataset's dispersion relative to its mean (HARGRAVE, 2023). This is useful for understanding the variability within the dataset. Outliers or extreme values can have a major impact on the standard deviation. As alternative tools median and mode can be use and which are not affected by extreme values as well as useful when using categorical data.

* Linear Regression

Linear regression can be used to determine the relationship between one or more independent variables and a dependent variable. The size and direction of the link between two variables are represented by the coefficients. However, it presumes a linear relationship, which may not be valid in all cases. Outliers and assumptions that are violated can have an impact on accuracy. Linear regression can only be utilized with continuous data, so tests such as Ordinal logistic regression can be used to develop prediction models.

* Correlation

The degree and direction of a linear relationship between two variables are measured by correlation. Correlation is scale-independent, which means it is unaffected by measurement units. Aids in recognizing patterns and trends in data. It can disclose whether variables flow in the same direction or opposite directions, But the correlation is sensitive to outliers, A single outlier can have a considerable impact on the correlation coefficient, which can lead to incorrect assumptions. Assumes linear relationship between variables and can only be used to analyze quantitative data

* Cluster Analysis

It groups similar data points together to reveal underlying patterns and natural divisions in the collection. Its strength resides in its adaptability. Cluster analysis aids in identifying distinct subgroups or categories within the data. However, its effectiveness can be affected by variables such as the distance measure used and the clustering technique used, potentially resulting in different outcomes. Understanding the significance and characteristics of each cluster requires significant consideration and domain expertise, which can make interpretation difficult.

* Classification

Classification is a useful data analysis approach that predicts the categorical class or category of an observation based on its characteristics. It can handle binary and multiclass classification issues with ease, making it suitable for a wide range of situations. Classification methods can be affected by imbalanced datasets, in which one class outnumbers the others, potentially resulting in biased predictions. Furthermore, noise and outliers in the data may have an impact on the model's accuracy.

7 Critical Review of the Visualisation Tools Used

ggplot2 in R is a powerful visualization tool noted for its remarkable customizability and flexibility, allowing users to construct a wide range of charts. ggplot2 supports faceting, allowing the construction of numerous plots based on different factors. However, compared to other visualization tools, it has a higher learning curve, especially for beginners, and its syntax might be lengthier for simpler displays.

Plotly is a versatile and dynamic tool for creating interactive visualizations in R. Its benefits are interactivity, chart type variety, support for web-based dashboards, and cross-language compatibility. However, when considering whether Plotly is the ideal solution for their individual data visualization needs, users should evaluate the learning curve and potential complications in customization.

# References

Vivek Krishnamoorthy and Udisha Alok. (2022 ). *Linear Regression: Assumptions and Limitations*. Retrieved from https://blog.quantinsti.com/linear-regression-assumptions-limitations/

1Vaishali Chandrakant Wangikar and 2Ratnadeep R. Deshmukh. (2011). Data Cleaning: Current Approachesand Issues.

Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. (2019). Impact of missing data on bias and precision when estimating change in patient. *reported outcomes from a clinical registry. Health Qual Life Outcomes.*

BUSH, T. (2020). *escriptive Analysis: How-To, Types, Examples*. Retrieved from PESTLE ANALYSIS: https://pestleanalysis.com/descriptive-analysis/

e, Q. Y. (1948). Linear correlation analysis of ammunition storage environment based on Pearson correlation. *Journal of Physics: Conference Series*.

Gbeminiyi John Oyewole & George Alex Thopil. (2022). Data clustering: application and trends.

HARGRAVE, M. (2023). *Standard Deviation Formula and Uses vs. Variance*. Retrieved from Investopidea: https://www.investopedia.com/terms/s/standarddeviation.asp

Indicative. (2022). *What Is Classification Analysis?* Retrieved from Indicative: https://www.indicative.com/resource/classification-analysis/#:~:text=Classification%20analysis%20is%20a%20data,programming%2C%20neural%20network%20and%20statistics.

Lawton, G. (n.d.). *predictive modeling*. Retrieved from Techtarget: https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling

Lenzerin, M. (2002). Data Integration: A Theoretical Perspective. *Data Integration: A Theoretical Perspective*.

Taylor, S. (2015). *Mean*. Retrieved from corporatefinanceinstitute.com: https://corporatefinanceinstitute.com/resources/data-science/mean/

*The Crown Prosecution Service*. (n.d.). Retrieved from The Crown Prosecution Service: https://www.cps.gov.uk/

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago & Oteng Tabona . (2021). *A survey on missing data in machine learning*.

V. Krishnaiah1\*, Dr.G.Narsimha 2. (2014). Survey of Classification Techniques in Data Mining. *International Journal of Computer Science and Engineering*, 10.

*What is Data Integration?* (n.d.). Retrieved from Tibco: https://www.tibco.com/reference-center/what-is-data-integration#:~:text=Data%20integration%20is%20the%20process,process%20by%20systems%20and%20users.