

**Muhammad Umair Ashraf Khan**

**ID: 4215274**

**DATA ANALYSIS AND VISUALIZATION REPORT**

**Crown Prosecution Service Case Outcomes**

Contents

[Executive Summary 4](#_Toc138188882)

[Data Integration 5](#_Toc138188883)

[Load\_files\_from\_dataset 5](#_Toc138188884)

[Merge\_files 5](#_Toc138188885)

[Data Pre-processing 6](#_Toc138188886)

[Data Cleaning 6](#_Toc138188887)

[Data Description 9](#_Toc138188888)

[Data Explorer 9](#_Toc138188889)

[Missing Data Profile 10](#_Toc138188890)

[Data Distribution 10](#_Toc138188891)

[Summary of Columns of Interest 12](#_Toc138188892)

[Successful\_drugs 12](#_Toc138188893)

[County 12](#_Toc138188894)

[Year 12](#_Toc138188895)

[Month 12](#_Toc138188896)

[Region 13](#_Toc138188897)

[Correlation Matrix 13](#_Toc138188898)

[Predictive Analysis 15](#_Toc138188899)

[Regression 15](#_Toc138188900)

[Linear Regression 15](#_Toc138188901)

[Clustering 20](#_Toc138188902)

[K-Means Clustering 20](#_Toc138188903)

[Classification 22](#_Toc138188904)

[Binomial Classification 22](#_Toc138188905)

[References 24](#_Toc138188906)

# 

# Executive Summary

The data that was used for this analysis is the crown prosecution dataset. The dataset that was used contained data starting from 2014 up to 2018. Although it must be kept in consideration that the dataset did not contain records for each month on a yearly basis.

The dataset has been analyzed using various techniques such as data cleaning, data exploration, data description, data visualization, regression, and classification techniques. The report covers how the data was prepared before any further techniques were applied to the dataset.

After that correlation analysis was performed to know more the existing variables within the dataset and how they were linked with each other. The correlation matrix showed strong relationships between different variables such as Successful Fraud Convictions and Successful Forgery Convictions. Although strong correlation does not guarantee causation, it was intriguing to look at these variables in detail using regression, clustering, and classification techniques.

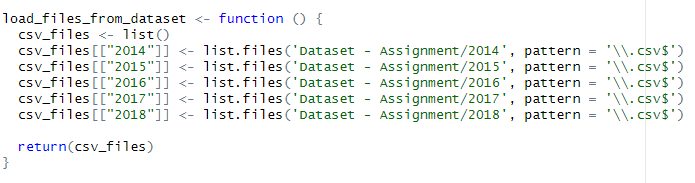
Some machine learning techniques were utilized for prediction and classification. The prepared and cleaned dataset was used and split into training and testing sets. For prediction linear regression was used to predict successful drugs convictions based on successful frauds conviction.

For classification, logistic regression was not suitable as the introduced region variable was categorical. Instead, multinomial logistic regression was used for data classification. K-means clustering algorithm was utilized to perform clustering.

# Data Integration

## Load\_files\_from\_dataset

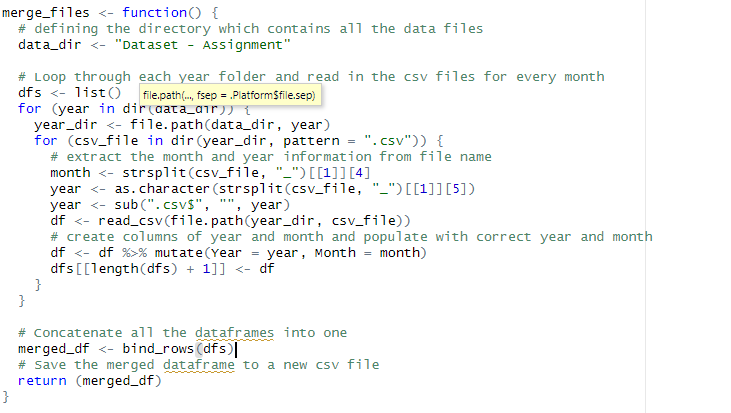
This function is used to fetch all the files within directories and returns a nested list. The first level of the list includes the year and for each year it contains a list containing all the files for that year.



## Merge\_files

This code starts reading files from the year directories, after that it loads all the files in the year directory one by one loading all the csv contents and pasting it in the master data frame called merged\_df.

The important aspect to note here is that file names contain the information of month and year. This information is not present as part of the dataset. This information is vital to differentiate between the observations. So, the code when loading the csv contents also reads that information from the filename, creates columns called year and month in the dataset and appends it correctly. As a result, each observation can be related to a year and month subsequently in the analysis.



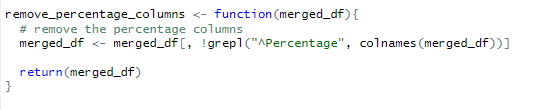
# Data Pre-processing

## Data Cleaning

For any data analysis to be performed data cleaning is one the most important steps. Data cleaning is an iterative process and may require multiple attempts depending on the quality of data maintained after each iteration. Domain knowledge and analysis goals are also important as they determine what steps are considered.

Data cleaning can also greatly increase the quality of the analysis. It also enhances reliability, improves accuracy and validity of the analysis results.

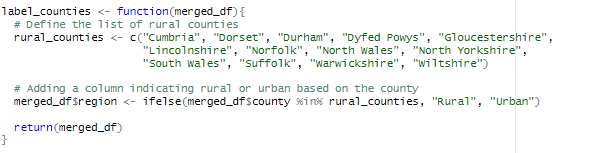
As the first step, data was viewed using the View() function. Apart from that summary was printed to get a glimpse of the dataset. The first step identified was the column names were long and had to be shortened. Then percentages columns needed removal as they could be calculated anytime using the existing columns.

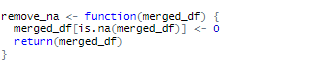


After removing the percentages column, we renamed the columns to make them more user friendly while we were coding the analysis. Names were shortened and spaces were replaced with “\_”. This improved the code readability as well as made the code writing easier. It also went along with the developer's style of writing the code.  


After that, the first column which contains information about the county names was not named. County as a column name was introduced.

In our analysis we needed to label the existing Counties in rural and Urban areas. We classified the County's into these two regions so that we can use this labelling in our clustering and classification analysis.



For values absent from the data columns, they were removed from the dataset, removing null values improved data accuracy further in our analysis. Although there are other methods of replacing null values like replacing by mean or median values of that variable, in our case it was simply easier to remove the null values as the count of null values was significantly low.  


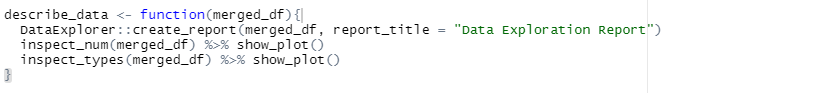
In the data observations we had a row that was called National. It contained the sum of all observations for all counties. Since we can calculate this anytime using the county's observations, we do not need to keep this in our dataset. We removed this observation from our data set.



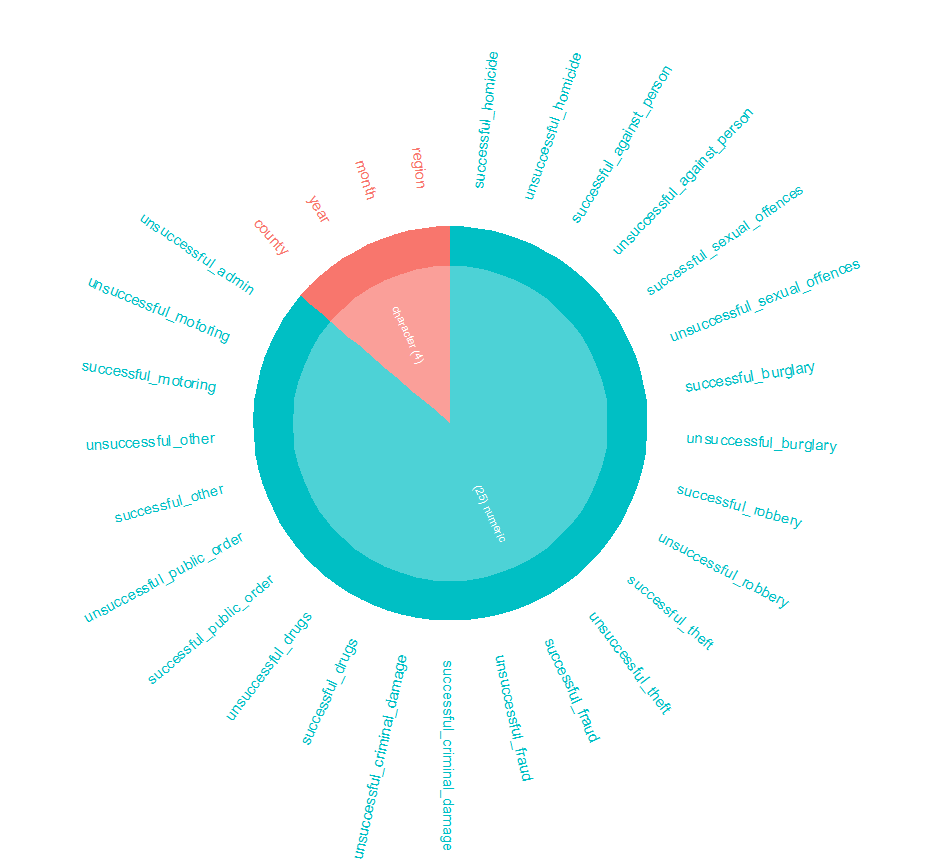
# Data Description

## Data Explorer

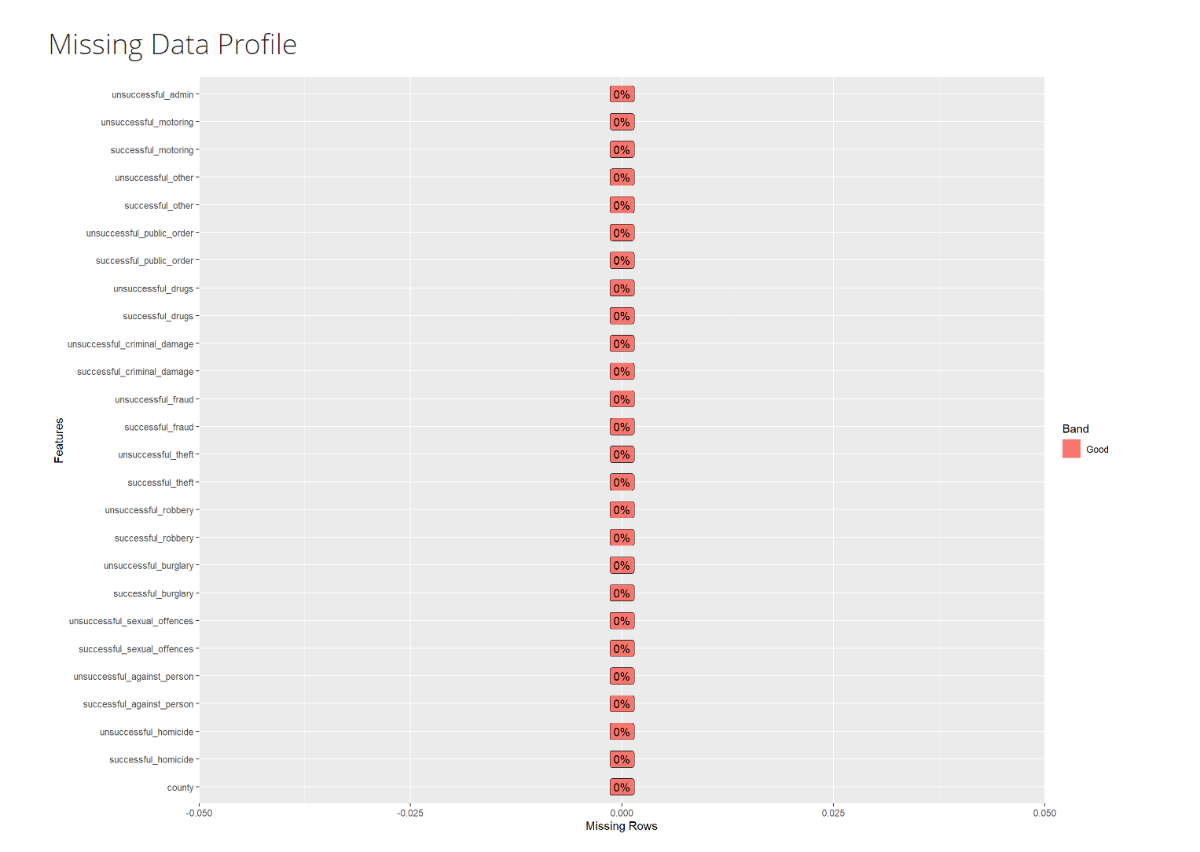
We used the data Explorer package. This package helped us to generate a detailed report about our data. The report contained numerous graphs and charts that helped us understand and describe our data in a better way.

Data Types

For data description we inspect the data types of our present variables. As you can see in the picture below in the variables called year, month, region and county are of type character and the rest of the variables are of type numerical



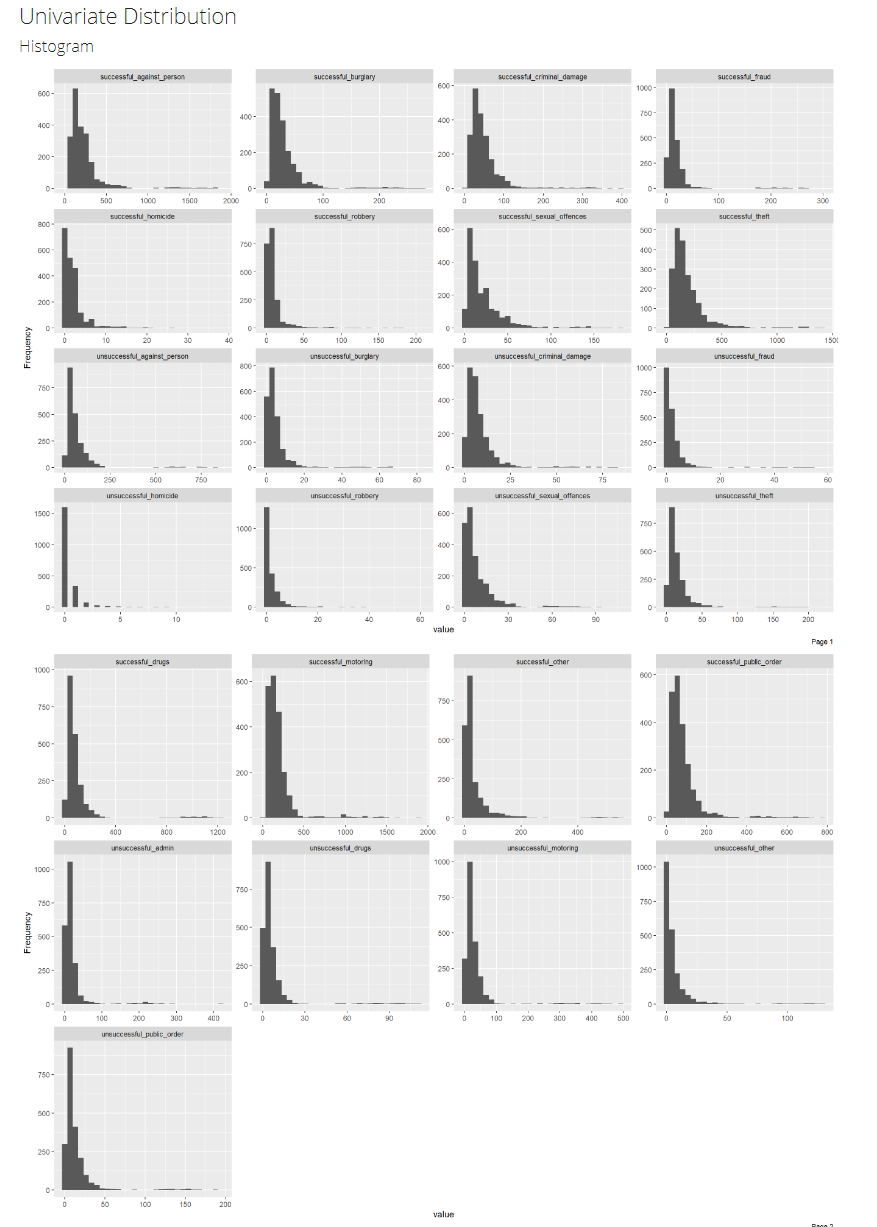
### Missing Data Profile



### Data Distribution

We used univariate distribution (Flury, 1986) to analyze all the variables in our data set. Univariate distribution elaborates a single variable or attribute in our data set. It analyses how the data is distributed for a single attribute without the consideration of relationships with other variables.

When doing univariate distribution, we are interested in how data is spread, where its mean median and mode are lying. We are also interested in how data is skewed. As you can see in the figure below data is cured on the left side more than it is on the right side for some of the variables.



## Summary of Columns of Interest

### Successful\_drugs

It represents that the minimum value of the variable is 4, 25% of the values are below 38, the median value is 62, the mean value is 95.72, 75% of the values are below 97 and the maximum value is 1228. It also indicates that the variable has some outliers as the max value is quite high as compared to the mean and the median.

  
Successful\_fraud

It represents that the minimum value of the variable is 0, 25% of the values are below 7, the median value is 12, the mean value is 19.71, 75% of the values are below 20 and the maximum value is 299. It also indicates that the variable has some outliers as the max value is quite high as compared to the mean and the median.



### County

This column is of type character and contains all the possible county values from England

### Year

This column has values from 2014 up to 2018. The dataset contained data for 5 years.

### Month

This column is of type character and values range from january to december.

### Region

This contains only two values i.e. rural and urban.

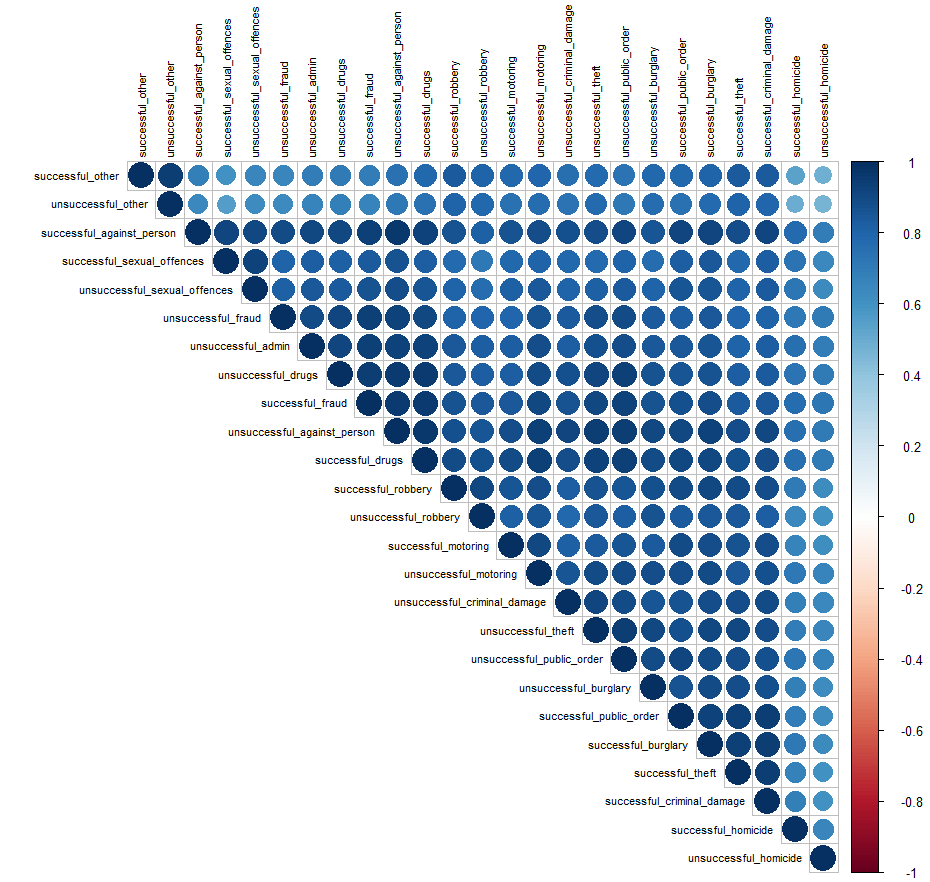
## Correlation Matrix

The correlation matrix (Dziuban, 1974) shows how variables are dependent or affected by the other variables within the dataset. It provides valuable insights of the dataset variables showing their dependencies and relationships.

The correlation coefficient ranges from -1 to 1. Negative correlation means that increase in value of one variable causes the values of other variables to decrease. Positive correlation on the other hand means that an increase in the values of one variable causes an increase in the values of the other variable. 0 means that variables are not dependent on each other and variation in one variable has no impact on the other variable.

To visualize the correlation matrix heat map is a trendy way. Each correlation coefficient is represented by a color gradient. It aids in visual identification of strong and weak correlations.

Correlation can only determine the variable relationships and does not mean a cause-and-effect relationship.



# Predictive Analysis

Predictive analysis is a method where we use machine learning techniques and statistical algorithms. The aim of predictive analysis is to predict future outcomes based on the historical data. It helps us to understand the data and to make models that can predict patterns and identify relationships. Some techniques that we covered for our predictive analysis include

* Regression
* Clustering
* Classification

## Regression

Regression (Freund, 2006) is a statistical approach used to examine how a dependent variable (also known as the response or outcome variable) is related to one or more independent variables (also known as predictor or explanatory variables). The primary objective of regression is to find the optimal line or curve that describes the relationship between the dependent variable and the independent variables. This line or curve is referred to as the regression line or model. Several types of regression techniques are available, including linear regression, multiple regression, logistic regression, and polynomial regression. For our analysis, we will be focusing on linear regression.

### Linear Regression

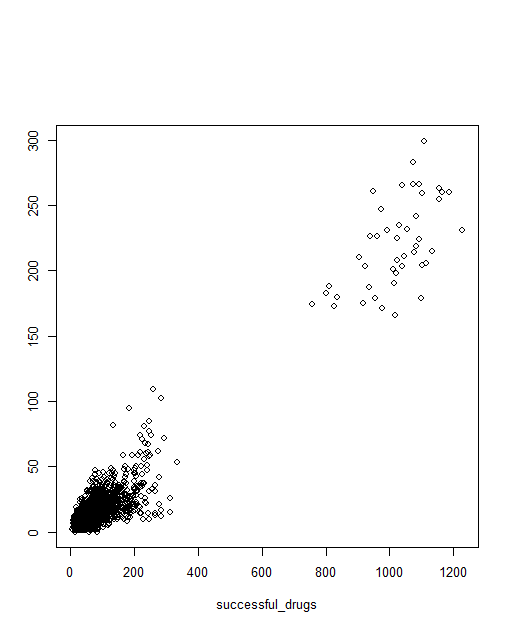
Linear regression (Montgomery, 2021) is a statistical method that is used to model the relationship between a dependent variable and an independent variable. As is evident by its name, it assumes linear relationship.

#### Hypothesis 1:

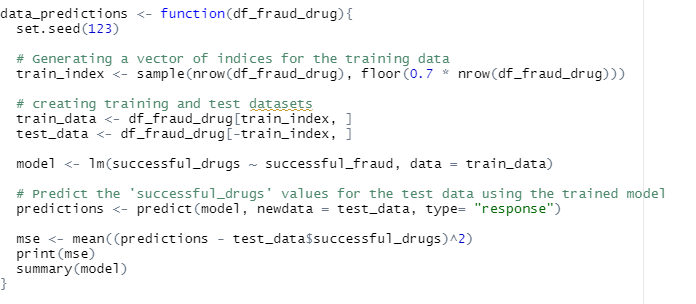
To be able to predict successful drug convictions based on successful fraud convictions. The value of coefficient of correlation between the two variables is 0.957. This value is remarkably close to a perfect positive correlation i.e. 1.



After plotting scatter plots between these two variables the result is as follows

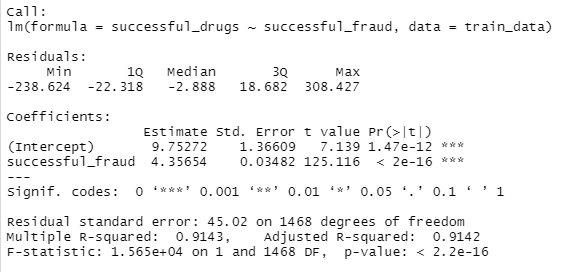


The code for producing data predictions after splitting the data into training dataset and testing dataset.



The summary of the model is as follows.

* The estimated intercept value is 9.75 with a standard error of 1.37
* The estimated coefficient for the variable successful\_fraud is 4.35 with standard error of 0.034
* The significance of the estimated coefficient also known as t-value is valued at 7.13 for the intercept and 125.116 for successful\_fraud. Both t-values are highly significant.
* P-value which is indicative of the significance of the coefficients also suggest that both variables have strong relationship and is contrary to the null hypothesis (no relation between the dependent variable and independent variable)

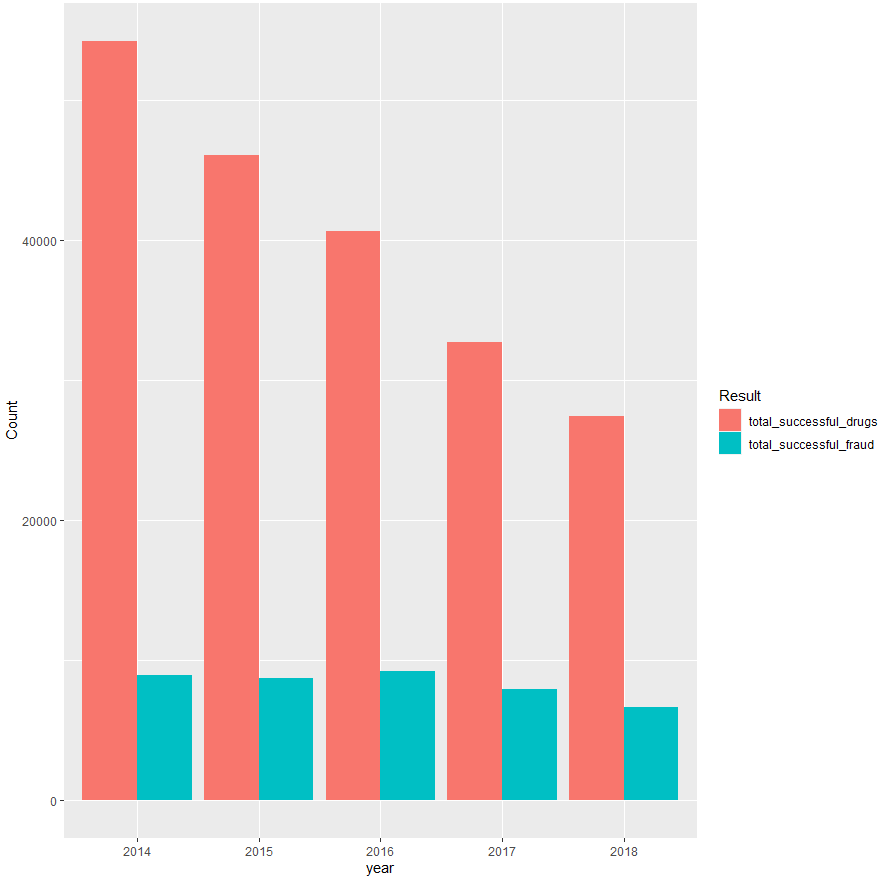


The mean squared error between predicted and actual values came out to be 1680.622. Mean square value is the absolute distance between the actual point in the dataset from the predicted point in the dataset. The smaller the error means the better the result.

Year wise Increase in Successful convictions To look at how successful convictions for drug and fraud-related crimes behaved over the years. To do this we calculated total\_successful\_drugs and total\_successful\_fraud and then plotted the reading in a bar graph. This is how it was coded.



Following was the outcome of the running the above code.



#### Hypothesis 2:

After demonstrating the significance of successful Frauds convictions and successful Drugs convictions, we divide the dataset into rural and urban areas for both convictions.

Here is the code that divided the dataset into rural and urban regions. After that, we calculated the mean values for both conviction types for urban and rural areas.

Mean rural Drug convictions is 45.66 whereas mean urban successful drugs convictions is 118.16. This clearly indicates that successful drug-related convictions in drug-related offences are higher than urban regions.

Similarly, rural fraud convictions is 8.09 whereas mean urban successful fraud convictions is 24.92. This clearly indicates that successful fraud related convictions in rural areas are relatively lower than urban regions.   


## Clustering

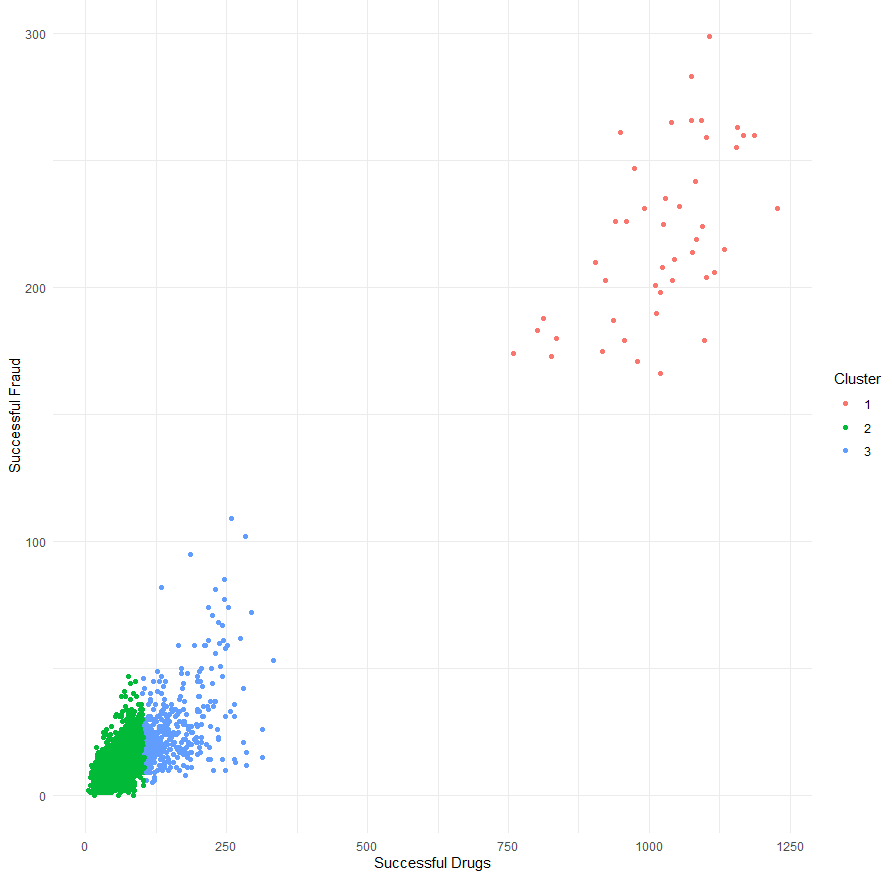
Clustering (Fraley, 1998) is used in unsupervised machine learning to identify and group similar data points together. Clustering algorithms group the dataset into a predefined number of clusters based on the similarity between the data points. This allows us to identify patterns, relationships, and subgroups within a dataset. We will look at k-means clustering

### K-Means Clustering

K-means (Hartigan, 1979) is a famous clustering algorithm that is used to divide or partition data into clusters. It is a type of unsupervised machine learning algorithm used to divide data into different clusters. It is an iterative method where centroids are initialized randomly then repeatedly assign each point to the nearest centroid. It repeats until the centroids no longer move or the maximum iterations are reached.

Here the function only fetches two columns successful fraud convictions and successful drugs convictions from the complete dataset. The number of specified clusters in 3. Each data point is assigned a cluster and then it is plotted to see the results.

 Following is the result of running code.



## Classification

Classification is a technique in data science that involves categorizing data into different classes based on their features or attributes. It is a supervised learning technique that requires labelling the training data so that it can learn and build a classification model. The goal of this exercise is to build a model that can accurately classify unseen and new data instances into its classes based on the data attributes that it had learned while training.

### Binomial Classification

Binomial Classification (Rayan, 2019) is a technique also known as binary classification. In this technique we classify data into two distinct classes or groups. The response variable is binary meaning that it can only take two possible outcomes. The outcome is based on the attributes of the input data.

In the code below we take columns of our interest i.e successful\_fraud and successful\_drugs. The target variable is set as a region. The data is split into a training and testing set using 70% - 30% split. The sample function randomly selects indices for the training set and the remaining are used for testing the dataset.



The accuracy of the model is reported as 79.37%. This accuracy can vary as the training set and test set are randomly selected each time the code is executed.

# References

Dziuban, C. D. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological bulletin, 81(6)*, 358.

Flury, B. K. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician, 40(3)*, 249-251.

Fraley, C. &. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal, 41(8)*, 578-588.

Freund, R. J. (2006). *Regression analysis. .* Elsevier.

Hartigan, J. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics), 28(1)*, 100-108.

Montgomery, D. C. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Rayan, J. C. (2019). Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. *Radiology: Artificial Intelligence, 1(1)*, e180015.