DATA ANALYSIS AND VISUALISATION PRINCIPLES ASSIGNMENT REPORT

JONES KWAME OSEI (S4112106)

2022-05-31

Table of Contents

[Introduction 4](#_Toc104931866)

[Overview 4](#_Toc104931867)

[General Hypothesis 4](#_Toc104931868)

[Data Cleaning 4](#_Toc104931869)

[Justification for Data Cleaning: 5](#_Toc104931870)

[Data Structure Before Cleaning 5](#_Toc104931871)

[Techniques for Cleaning 5](#_Toc104931872)

[1. removing of irrelevant entry: 5](#_Toc104931873)

[2. Rename County: 6](#_Toc104931874)

[Converting Data Type: 6](#_Toc104931875)

[Removing Percentage columns: 6](#_Toc104931876)

[Rename Columns with longer names: 6](#_Toc104931877)

[Standardise capitalisation and wording: 6](#_Toc104931878)

[Descriptive Analysis 6](#_Toc104931879)

[Basic Exploaratory Data Analysis 7](#_Toc104931880)

[The Head() and Tail() of our dataset 7](#_Toc104931881)

[Data Structure after Cleaning 7](#_Toc104931882)

[Dimension of Dataframe after cleaning 7](#_Toc104931883)

[let’s display the type and a preview of all columns as a row 7](#_Toc104931884)

[Summary of the dataset 8](#_Toc104931885)

[1. Descriptive statistics with the summary(). 8](#_Toc104931886)

[SUmmary statistics of a column 8](#_Toc104931887)

[2. Summary statistics using stat.desc() function: 8](#_Toc104931888)

[3. Descriptive statistics with describe() function 8](#_Toc104931889)

[Investigating the missing values. 8](#_Toc104931890)

[Visualise the Data Types 9](#_Toc104931891)

[Plotting the dataset structure 9](#_Toc104931892)

[Plotting the bar chart for each discrete feature 9](#_Toc104931893)

[The charts can also be grouped by a discrete variable, e.g. the presence of a Robbery 9](#_Toc104931894)

[Plotting quantile-quantile for each continuous feature 9](#_Toc104931895)

[Plots density estimates for each continues feature 10](#_Toc104931896)

[Visualise correlation heatmap 10](#_Toc104931897)

[Visualising the principal components 10](#_Toc104931898)

[Inspecting the dataset 10](#_Toc104931899)

[Advance Exploratory Data Analysis. 10](#_Toc104931900)

[Creating a variable called total\_crime to input all crimes together for each county. 10](#_Toc104931901)

[Visualising All crimes 10](#_Toc104931902)

[Visualising Successful crime conviction Distribution 10](#_Toc104931903)

[Successful crime conviction in 2015 10](#_Toc104931904)

[Successful crime conviction in 2015 10](#_Toc104931905)

[Successful crime conviction in 2016 10](#_Toc104931906)

[Successful crime conviction in 2017 10](#_Toc104931907)

[Visualising unsuccessful crime distribution 10](#_Toc104931908)

[unsuccessful crime distribution for 2014 10](#_Toc104931909)

[unsuccessful crime distribution for 2015 10](#_Toc104931910)

[unsuccessful crime distribution for 2016 10](#_Toc104931911)

[unsuccessful crime distribution for 2017 10](#_Toc104931912)

[Find top counties per total crime 10](#_Toc104931913)

[Top five Counties with Hihest Crime Cases 10](#_Toc104931914)

[Counties with most Crimes from 2014 to 2017 10](#_Toc104931915)

[Total Crimes from 2014 to 2017 with a voilin plot 10](#_Toc104931916)

[Average crime distribution for each year 10](#_Toc104931917)

[Average Robbery Within Each Year 10](#_Toc104931918)

[for 2014 10](#_Toc104931919)

[for 2015 11](#_Toc104931920)

[for 2016 11](#_Toc104931921)

[for 2017 11](#_Toc104931922)

[Average Crime Type Distribution 11](#_Toc104931923)

[Critical Review of the Visualisation Tools: 11](#_Toc104931924)

[Predictive Analysis: 11](#_Toc104931925)

[Covariance and Correlation Analysis: 11](#_Toc104931926)

[print the first 6 rows 12](#_Toc104931927)

[Compute covariance and correlation matrix 12](#_Toc104931928)

[Visualize correlation matrix 12](#_Toc104931929)

[Regression analysis 12](#_Toc104931930)

[Linear Regression: 12](#_Toc104931931)

[Simple Linear Regression. 12](#_Toc104931932)

[Simple Libear Model One 13](#_Toc104931933)

[Plotting our variables for Model One 13](#_Toc104931934)

[Conclusion of model1 14](#_Toc104931935)

[Compute p-value 14](#_Toc104931936)

[Prediction for Model One 14](#_Toc104931937)

[Plotting Simple Libear Model One Results 14](#_Toc104931938)

[Linear Model Two 14](#_Toc104931939)

[Plotting our variables for Model Two 14](#_Toc104931940)

[Regression Analysis and Summary for Model Two 14](#_Toc104931941)

[Conclusion of model2 15](#_Toc104931942)

[Plotting Simple Linear Model Two Results 15](#_Toc104931943)

[Critcal Review of Model1 and Model2 of our linear regression. 15](#_Toc104931944)

[Multiple Linear Regression: 15](#_Toc104931945)

[Regression Analysis and Summary for Mutltiple Regression 16](#_Toc104931946)

[Conclusion of Multiple Regression 16](#_Toc104931947)

[Critcal Review of Simple Linear Regression and Multiple Linear Regression. 16](#_Toc104931948)

[Clustering: 16](#_Toc104931949)

[K- Means Clustering 17](#_Toc104931950)

[Hierarchical clustering: 18](#_Toc104931951)

[Agglomeration Clustering (AGNES): 18](#_Toc104931952)

[Divisive Hierarchical Clustering: Top down approach: 18](#_Toc104931953)

[Distance Used: 19](#_Toc104931954)

[Critcal Review Of The K-Means and Hierarchical clustering: 19](#_Toc104931955)

[Classification: 20](#_Toc104931956)

[data partition (2 clusters) 20](#_Toc104931957)

[Decision Tree 20](#_Toc104931958)

[data partition (For 3 clusters) 20](#_Toc104931959)

[Decision Tree 20](#_Toc104931960)

[General Review of the Analytical Tools 20](#_Toc104931961)

[Comparing Land Size of Metropolitan and city to Gloucester 20](#_Toc104931962)

[Map of Meropolitan and City (London) 20](#_Toc104931963)

[Map of Gloucestershire 20](#_Toc104931964)

# Introduction

The purpose of this assignment project is to download, reshape and combine two years of *Crown Prosecution Service Case Outcomes* by *Principal Offence Category*. We will use various data analytic tools and techniques for descriptive and predictive analysis by using different techniques. Varying analyses and interpretations of various outcomes will be made through descriptive analytics.

The analytical tools used were also critically analysed in this report as well as the effectiveness of the techniques employed. We will have a careful look at the visualisation tools utilised in the assignment.

## Overview

### General Hypothesis

Definitions of the hypotheses based on the current datasets are that there exists a correlation between specific variables. The more specific hypothesis is that the number of robbery offenses successful can be predicted using the remaining numerical variables.

# Data Cleaning

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data to improve the quality of data(Chai 2020). This should be the first data analyst should begin with. Without acknowledging errors in data makes the analysis vulnerable and susceptible to failure.

Detecting and repairing dirty data is one of the perennial challenges in data analytics, and failure to do so can result in inaccurate analytics and unreliable decisions(Natarajan, Li et al. 2010). Today’s business and services decision-making hugely depend on storing and obtaining data big amounts of data. These datasets can support and improve management decisions as well as exponentially optimise services.

Notwithstanding this importance, the quality of data continues to be a greater concern for individuals, service providers, and private and government institutions. The quality of data can positively or negatively affect the decision and make analysis unrealistic. Thus, dirty or uncleaned data can lead to an inaccurate decisions. To avoid this, we have a duty as data scientists to effectively detect the very error in our data and subsequently repair or correct those errors. Some of these errors are a result of data entry especially if those making the entry are untrained to carry out that task. Some common mistakes like typo errors, adding commas to figures, incorrect date format, duplicate entries and many can render a data uncleaned.

Due to the forgone, there is the need to clean our data to have something more meaningful to make tangible inferences from it. The conversion of raw data into a form that will make it less difficult to understand and interpret. This can be achieved by rearranging, ordering, and manipulating data to provide more insightful information about our data.

## Justification for Data Cleaning:

We are cleaning our dataset to have cleaned and good data to derive a realistic analysis to obtain a result that can effectively support a decision. Generally, data cleaning reduces errors and improves data quality(Natarajan, Li et al. 2010). Some of the data cleaning techniques adopted in this project include removing an irrelevant entry, handling missing values, converting data type, standardising capitalization and wording, renaming, clear formatting, and fixing errors and duplication.

The justification for the integration and data cleaning is that the data is provided in parts, and it is desirable to implement the analysis every year or for the entire period also it is necessary to clean the data from na values.

First, we have to have a look at the structure of our data to reshape it to a finished dataset that can be used for our various analysis

### Data Structure Before Cleaning

The dataset before cleaning composes of characters and integers with characters being dominant. We can find 1806 observations of 54 variables in the All\_Crimes data frame. In addition, there are no missing values in the dataset. The missing values will be further investigated later.

We will now have to know the type of data we have uploaded.

### Techniques for Cleaning

Real Data are never perfect because errors are inevitable and may occur creatively and unexpectedly (Natarajan, Li et al. 2010). So we have to painstakingly first pinpoint issues in the data rendering it bad. The techniques for cleaning our data include the following:

#### 1. removing irrelevant entries:

Entries that are not important to our analysis will only make it slow and might even confuse the analysis we want to do. Ensuring only what is relevant and necessary before we begin our analysis is key to success. In our dataset, we could see that there are percentage columns that are not needed in the analysis we want to do. But we decided to deal with the percentage symbol first leaving the number. This was done to prove the way only the symbol is cleaned if we had wanted to use that column. Also, in our dataset, there are dashes (-), commas (,), and empty spaces that if not removed, can cause errors. The empty spaces are replied to with zero (0).

#### 2. Rename County:

The county column was initially named “X” and needed to be renamed.

#### Converting Data Type:

All variables were read by R as characters. We needed to change from range two to 51 (2:51) from characters to numbers so that R can read them as numbers instead of characters. This is because R initially imputed the numbers as texts. However, to be able to process them, they need to be changed to numerals. Our analysis algorithms will not compute any mathematical input because they are classed as strong characters as they appear in texts. Same as the dates as they are stored as texts. Hence, we also convert them to numerical.

#### Removing Percentage columns:

Since our analysis will not be based on percentages, we decided to remove them from the dataset.

#### Rename Columns with longer names:

We could observe that the column names of our data frame were very long and could be difficult to call in code. For this reason, they are renamed for a shorter name so they can be called easily.

#### Standardize capitalization and wording:

Within this dataset, we have to ensure that the texts are congruent with each other. Any mixture or consistencies of wording and capitalisation can easily cause errors in categories. Due to this, all dates and months were given capital letters in their initials. This will avoid any issues when we need to translate before processing as capitalisation can cause R to misinterpret.

**Remove National from the county row**. The national column sums all the criminal cases for each case which we chose not to include in our analysis, hence, removed it from the data frame.

# Descriptive Analysis

*Descriptive Analysis* is a way to show, describe and summarise data in a more constructive way so that the patterns that emerge will complement every condition of the data(Abt 1987).

All over the world, *big data and data science* have been in *ascendancy*. They help support organizations and cooperate to maximize profits and services. However, to get the most out of every piece of data, there has to be extensive research into it as this makes the data to be processed and studied with a higher level of scrutiny.

The data should be analyzed to produce an in-depth insight and influential trends which allows those that follow to be made to gain public approval. Descriptive Data Analysis is one of the most relevant procedures when making a statistical analysis of our dataset. It assists us with a conclusion of the distribution of our dataset. Also, it enables us to detect outliers and helps us to identify similarities within our variables that will put us in a position for making further statistical analysis.

In this section, we will further break our descriptive analysis into two parts. That is the Basic Data Exploratory Analysis and Advanced Exploratory analysis.

## Basic Exploratory Data Analysis

Exploratory Data Analysis (EDA) is usually intended to generate hypotheses and not to lead to conclusions based on the results of the study (Shinde, Majumder et al. 2022). This is done to know the dataset we will be working with. It is a good step even before data cleaning and after data cleaning before any data analysis is done.

Data analysis is the process of extracting insights from data. Data is heterogeneous in all ways, and processing such data is a challenge. Before applying any machine learning model to any dataset, it is necessary to understand the problem, deal with the missing values and noise, visualize the dataset, and select the machine learning model to analyze the data(Shinde, Majumder et al. 2022).

### The Head() and Tail() of our dataset

Since it will be hard to digest a huge number of rows of our data which has twenty columns and thousands of rows, we will be looking at the first ten and last rows of our dataset.

Table

Description automatically generated

The above displays the first ten (10) rows present in the input data frame.

A screenshot of a computer

Description automatically generated with medium confidence

The above displays the last ten (10) rows present in the input data frame.

### Data Structure after Cleaning

A picture containing text

Description automatically generated

From the above, the dataset after cleaning composes of characters and numbers with numbers being the dominant as compared to the structure before cleaning. Also, We can see the number of observations after cleaning is 1764 obs. of 29 variables instead of the 1806 observations of 54 variables in the All\_Crimes data frame before cleaning. Again, no missing values were identified in the dataset. The missing values will look into the missing values later.

### Dimension of Dataframe after cleaning

We will have a look at the dimension of our data

Text

Description automatically generated

### let’s display the type and a preview of all columns as a row

Background pattern

Description automatically generated

We have seen a preview of all the columns as well as rows and the type of data. We can see every column in a data frame showing details of 1,764 rows belonging to 29 columns.

### Summary of the dataset

Descriptive statistics of our data frame can be computed by three methods. These are:

#### Descriptive statistics with the summary().

A picture containing background pattern

Description automatically generated

Using the summary() function is automatically applied to each column and the resulting format relies on the type of data in each column. Thus:

-A column with numeric variables returns mean, median, minimum (min), maximum (max), and quantile.

- Columns with factor variables also return the number of observations in each group. Hence, descriptive Statistics simple calculates:

- minimum value of each column

- maximum value of each column

- mean value of each column

- median value of each column

- 1st quantile of each column (25th percentile)

- 3rd quantile of each column (75th percentile)

* Example:

#### Summary statistics of a column

Graphical user interface, text, application

Description automatically generated

A close-up of a document

Description automatically generated with medium confidence

#### 2. Summary statistics using stat.desc() function:

Table

Description automatically generated

This does a bit more than the simple describe() function. It also calculates the number of missing values and null of each column the number of non-missing values of each column sum, range, variance, and standard deviation for each column too.

#### 3. Descriptive statistics with describe() function

Here, the *Hmsic* package is used and it calculates the distinct values of each column, frequency of each value, and proportion of that values in the column as indicated below.

We will deeply investigate to know more about our dataset by using the skim(). It presents most of the numerical attributes from the summary. Again, it displays missing values, more quantile details, and an inline histogram for each variable.

From the output, we can see the data summary of our dataset that has *three column type frequencies* including:

|  |  |
| --- | --- |
| column Type | Frequency |
| Character | 3 |
| Date | 1 |
| Numeric | 25 |

The data summary also indicates there are no missing values found in the 29 columns and 1764 rows and has no group variables.

## Investigating the missing values.

We are now delving into our dataset to ascertain if there are no missing values in our data set. We can do this by testing the dataset or visualizing it. - test:

This returned a number zero in each element location indicating there are no missing values in our dataset.

Alternatively, we can do this by visualizing it:

The graphical representation above indicates that we have 100% of the values in our data, hence, there are no missing values.

### Visualize the Data Types

Graphical user interface

Description automatically generated with medium confidence

Shape

Description automatically generated with medium confidence

## Plotting quantile-quantile for each continuous feature

Graphical user interface, application

Description automatically generatedFrom quantile plots, we can observe that the charts showing the data are not normally distributed because the plots are skewed from straight lines. It can be inferred that there are outliers in the dataset. Hence, the distributions are skewed. Plots density estimates for each continuous feature

## Inspecting the dataset

Graphical user interface, text, application, email, Teams

Description automatically generated

Chart, sunburst chart

Description automatically generated

# Table Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

# Advance Exploratory Data Analysis.

## Visualising Successful crime conviction Distribution

Chart, bar chart

Description automatically generated

## Successful crime conviction in 2014

Chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

## Successful crime conviction in 2015

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## Successful crime conviction in 2016

Chart, bar chart

Description automatically generated

## Successful crime conviction in 2017

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## Visualising unsuccessful crime distribution

Chart, bar chart

Description automatically generated

### unsuccessful crime distribution for 2014

Chart, bar chart

Description automatically generated

### unsuccessful crime distribution for 2015

Chart

Description automatically generated

### unsuccessful crime distribution for 2016

Chart, bar chart

Description automatically generated

### unsuccessful crime distribution for 2017

Chart, bar chart

Description automatically generated

#### Top five Counties with Highest Crime Cases

A screenshot of a computer

Description automatically generated with medium confidence

### Counties with the most Crimes from 2014 to 2017

Chart, bar chart

Description automatically generated

Plotted are five counties with the most crimes from 2014 to 2017 in which Metropolitan and city topped the group.

### Total Crimes from 2014 to 2017 with a Violin plot

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

It is no doubt that 2014 is seen to record the higher number of crimes whilst 2017 had the lowest. It is also seen that there was a consistent decrease from 2014 to 2016.

### Average Robbery Within Each Year

#### for 2014

Chart, bar chart

Description automatically generated

The average crime distribution for only Robbery in 2014. It can be observed that January recorded the most robbery cases in 2014 with about 1450 as compared to December which had 1200.

### for 2015

Chart, bar chart

Description automatically generated

In 2015, March and July recorded the highest robbery cases with an estimated 1500 to December which recorded the lowest with cases of robbery of around 1000.

#### for 2016

Chart, bar chart

Description automatically generated

Once again, January recorded the highest robbery crime cases. This time, April was next with an estimated 1220 cases of robbery compared to about 1240 in January.

#### for 2017

Chart, bar chart

Description automatically generated

March 2017 topped with about 1350 robbery cases. Next to it is January which recorded 1300 to December which recorded the lowest with cases of robbery around 1000.

It can be concluded that, from 2014 to 2017, January had the highest number of robbery cases whilst December recorded the least consistently.

# Critical Review of the Visualisation Tools:

The advanced exploratory data analysis was done by using various visualisation tools. However, some were more effective than others in terms of appearance and readability. Some too could give an aesthetic look and could make decision-making easy by making it less difficult for everyone to comprehend the.

For instance, figures which have fills, labels, and themes give more accurate and vivid information than those which did not.

Chart, bar chart

Description automatically generated

Again, those which have no label fills, and themes gave no information and were ambiguous to read. Just Looking at them gave no adequate information as compared to the former. Chart, bar chart, histogram

Description automatically generated

# Predictive Analysis:

Predictive analysis is simply forecasting the future using data combined with statistical modeling through machine learning. The data in this case can be historical, that is, having a look at past events and real-time data, events happening currently around us. The data collected is of no use unless some useful information is derived from it. Therefore, it is essential to think of some predictive analysis for analyzing data and to get meaningful information(Dash, Panigrahi et al. 2015).

In this session, we will make predictions using regression, clustering, and classification.

## Covariance and Correlation Analysis:

There is a need to do correlation analysis before the regression. Correlation simply is when a change in a variable affects or causes a significant change in another. Vice versa is the case of uncorrelated variables. Thus, correlation measures the relationship between variables. In testing whether robbery conviction was dependent on burglary, if there is an increase in burglary causes a corresponding increase in robbery, then we can statistically assume that there is a correlation between the two. Hence, the need for correlation analysis.

So, the purpose of correlation analysis is to determine the absence (dissociation) or presence (association) of a relationship between two variables. If they are uncorrelated, we will be able to measure the weakness of their association. Also, if they are correlated it is easy to measure their strength of association. Thus, we can find the numerical values that show the degree of relationship between the two variables. This clearly and conclusively summarizes the relationship between the two variables that we will find in our regression analysis.

## 

## Regression analysis

Regression analysis is a conceptually simple method for investigating the functional relationship among variables (Freund, Wilson et al. 2006). For example, we want to examine whether offenders who were convicted of robbery are related to those convicted of burglary or drugs. This relationship can be expressed in the form of an equation or a model connecting the response or dependent variables and one or more explanatory or predictive variables. In this example, the response variable is robbery conviction (measured by the number of people that were successfully convicted in the various counties per year) and the explanatory or predictor variable is either burglary convictions or drug convictions.

This means that regression analysis is mostly concerned with specifying the relationship between a single numeric dependent variable (The variable to be predicted) and one or more numeric independent variables (the predictors)(Sykes 1993). The type of regression we will use in the analysis is Linear Regression.

### Linear Regression:

This is a regression model that uses a straight line to show the relationship between the dependent variables and the predictors.

It is known that linear regression provides the simplest model form to model the regression function as a linear combination of predictors. It is popular in applications, and several reasons account for its popularity given below. Because of the linear form, the model parameters are easily interpretable. In addition, linear model theories are well established with mathematical elegance. Moreover, linear regression is the building block for many modern modeling tools. In particular, when the sample size is small, or the signal is relatively weak, linear regression often provides a satisfactory approximation to the underlying regression function(Freund, Wilson et al. 2006).

#### Simple Linear Regression.

Simple regression is the study of the dependence of response on one predictor, usually by assuming that the mean of the response as a function of the predictor is a straight line(2005).

Null Hypothesis: There is no linear association between robbery and burglary conviction, in other words, the increase in a robbery is not dependent on a burglary conviction.

Alternative hypothesis: There is a statistical linear association between robbery and burglary, in other words, the increase in robbery comes as a result of an increase in burglary convictions.

##### Simple Linear Model One

##### Plotting our variables for Model One

Chart, scatter chart

Description automatically generated

Our guess about the statistically significant of this plot is we want to see if Burglary Cases and for that matter Burglary Convictions are a significant linear predictor of Robbery conviction. Based on the plot, it seems reasonable to guess that we will reject the null hypothesis here.

The residuals are the difference between observed and predicted values. It can be calculated as Residual = Observed value – Predicted value.

| min | 1Q | Median | 3Q | Max |
| --- | --- | --- | --- | --- |
| 33.219 | -3.805 | 0.289 | 3.883 | 81.574 |

*The Min*: It represents the data point below the regression line

*1Q*: Is the 1st quantile which means 25% of the residual of the model1 is less than -3.805.

*Median*: The median of the residual of mode1 is 0.289. The median residual should be as close to zero (0) as possible

*3Q*: The 3rd quantile where 25% of the residual of model1 is greater than 3.883 From the summary of our model1, The Coefficients part presents:

*max*: It represents the data point the spread above the regression line

The *mean* is not seen because it is always zero in linear regression since thats what they are optimized for.

1. The Estimate for the model parameters – the value of the y-intercept to be -6.043195
   * the estimated effect of Burglary\_convict on Robbery Conviction cases is 0.509424
2. The standard error of the estimated values (Std. Error) is 0.006453
3. The test statistic (t value, in this case, the t-statistic) is 78.95
4. The p-value ( Pr(>| t | ) ), aka the probability of finding the given t-statistic if the null hypothesis of no relationship were true.
5. The final three lines are model diagnostics – the most important thing to note is the value (here it is 2.2e-16, or almost zero), which will indicate whether the model fits the data well.
6. From these results, we can say that there is a significant positive relationship between Burglary conviction and Robbery conviction (p-value < 0.001), with a 0.509-unit (+/- 0.01) increase in Robbery conviction for every unit increase in Burglary conviction.

In other words, we can see that the b1, obs value is around 0.509 which corresponds to an obs value of 78.95. We can interpret the meaning of the b1, obs in the context that, for everyone one case increase in Burglary conviction, we can expect Robbery conviction to increase by 0.509 conviction.

##### Conclusion of model1

From the foregoing, we have sufficient primary evidence that the null hypothesis can be rejected. The initial guess that Burglary conviction is a significant linear predictor of Robbery conviction has supporting evidence.

##### Prediction for Model One

Graphical user interface, text, application

Description automatically generated

##### Plotting Simple Linear Model One Results

Chart, scatter chart

Description automatically generated

##### Linear Model Two

##### Plotting our variables for Model Two

Chart, scatter chart

Description automatically generated

The residuals are the difference between observed and predicted values. It can be calculated as Residual = Observed value – Predicted value.

| min | 1Q | Median | 3Q | Max |
| --- | --- | --- | --- | --- |
| -45.919 | -3.323 | -0.763 | 1.964 | 57.745 |

*The Min*: It represents the data point below the regression line

*1Q*: This is the 1st quantile which means 25% of the residual of model1 is less than -3.323.

*Median*: The median of the residual of mode1 is -0.763. The median residual should be as close to zero (0) as possible

*3Q*: The 3rd quantile where 25% of the residual of model1 is greater than 1.964 From the summary of our model1, The Coefficients part presents:

*max*: It represents the data point the spread above the regression line

The *mean* is not seen because it is always zero in linear regression since thats what they are optimised for.

1. The Estimate for the model parameters – the value of the y-intercept to be -0.635
   * the estimated effect of Burglary\_convict on Robbery Conviction cases is 0.1099
2. The standard error of the estimated values (Std. Error) is 0.001412
3. The test statistic (t value, in this case, the t-statistic) is 77.896
4. The p-value ( Pr(>| t | ) ), aka the probability of finding the given t-statistic if the null hypothesis of no relationship were true.
5. The final three lines are model diagnostics – the most important thing to note is the value (here it is 2.2e-16, or almost zero), which will indicate whether the model fits the data well.
6. From these results, we can say that there is a significant positive relationship between Burglary conviction and Robbery conviction (p-value < 0.001), with a 0.509-unit (+/- 0.01) increase in Robbery conviction for every unit increase in Burglary conviction.

In other words, we can see that the b1, obs value is around 0.109952 which corresponds to an obs value of 77.896. We can interpret the meaning of the b1, obs in the context that, for everyone one case increase in Drug conviction, we can expect Robbery conviction to increase by 0.109952 conviction.

##### Conclusion of model2

From the foregoing, we have sufficient primary evidence that the null hypothesis can be rejected. The initial guess that Drug conviction is a significant linear predictor of Robbery conviction has supporting evidence.

#####Prediction for Model Two

##### Plotting Simple Linear Model Two Results

#### Critical Review of Model1 and Model2 of our linear regression.

| Model | Prediction MSE |
| --- | --- |
| model1 | 62.51 |
| model2 | 69.94 |

In comparing the two prediction results above, model1 had the mean standard error closest to zero, 62.51 as compared to 69.94 for the model2. Hence, model1 has a better accuracy since it had a lower value.

#### Multiple Linear Regression:

Multiple regression analysis (MR) is a highly flexible system for examining the relationship of a collection of independent variables (or predictors) to a single dependent variable (or criterion). The independent variables may be quantitative or categorical [mul reg]

Null Hypothesis: There is no linear association between robbery and burglary and drug conviction, in other words, the increase in the robbery is not dependent on burglary and drug conviction.

Alternative hypothesis: There is a statistical linear association between robbery and burglary and drug conviction. In other words, the increase in robbery comes as a result of an increase in burglary and drug convictions.

##### Regression Analysis and Summary for Multiple Regression

1. The estimated effect of a Drug conviction on a Robbery conviction is 0.058, while the estimated effect of Burglary\_convict is 0.278.
2. This means that for every 1% increase in Drug conviction, there is a correlated 0.058% increase in the conviction of robbery, and for every 1% increase in Burglary\_convict, there is a 0.278% increase in Robbery conviction.
3. The standard errors for these regression coefficients are very small, and the t-statistics are comparatively small (21.89 and 23.25, respectively). The p-values reflect these small errors and t-statistics. For both parameters, there is almost zero probability that this effect is due to chance.

##### Conclusion of Multiple Regression

From the foregoing, we have sufficient primary evidence that the null hypothesis can be rejected. There is a shred of statistical evidence that Burglary and Drug convictions are significant linear predictors of the Robbery conviction.

**Plotting our Multiple Regression results**

**Chart, scatter chart

Description automatically generated**

#### Critcal Review of Simple Linear Regression and Multiple Linear Regression.

| Linear Regression | Prediction MSE |
| --- | --- |
| Simple model1 | 62.51 |
| Simple model2 | 69.06 |
| multiple Reg | 64.64 |

In comparing the prediction results above, simple regression model1 had the lower mean standard error, 64.64 as compared to 69.06 and 64.64 of model2 and multiple regression respectively. Hence, multiple regression has a better accuracy since it had a lower value close to zero.

# Clustering:

Cluster analysis is one of the Pattern Recognition techniques and should be appreciated as such. It may be characterized by the use of resemblance or dissemblance measures between the objects to be identified(Diday and Simon 1976).

When data points are broken into several groups the data points belonging to the same groups are more identical to the other data points in the same group than those belonging to different groups. In other words, it is to separate groups with the same traits and locate them into clusters(Lee 1981).

In database design, clustering analysis can be used to group similar records into one bucket to shorten retrieval time. In the program restructuring problem, clustering analysis can be used to group programs frequently calling one another. This grouping reduces page faults. Clustering analysis can also be used to discover possible errors in a set of data and thus enhance data integrity(Dash, Panigrahi et al. 2015).

We going to make predictions using two clustering techniques namely, K-Means and Hierarchical Clustering.

## K- Means Clustering

This is a type of unsupervised learning technique that is used to group observations together. K-Means is relatively easy to comprehend and can be used almost and readily applied to every type of statistical problem. Here, we used several clusters also known as a centroid to group observations based on similarities. The centroid is a pivotal point or the middle with a huge hyperspace. So we are doing this for K-Clusters and essentially calculate the distance between the centroids.

The K-means will then merge similarities between the two clusters. Then whichever clusters are the same or minimal in distance will be merged. If there are no similarities then they will not be merged but will keep on iterating through as many iterations as there are. K-Means uses euclidean Distance to compare different clusters together to determine whether merging them will be possible or not. This algorithm randomly picks a point to calculate the centroid and based on that point it will look for the best position within each of the observations that have been given to calculate the centroid from them.

in other words, this algorithm randomly configures K centroids in the Data Space based on the position of each observation where the algorithm will optimize for the best position of the centroid. The K-Means algorithm only stops when no change in centroid values occurred. Again, it only stops when the number of iterations has been reached.

The k-means algorithm finds locally optimal solutions concerning the clustering error. It is a fast iterative algorithm that has been used in many clustering applications. It is a point-based clustering method that starts with the cluster centers initially placed at arbitrary positions and proceeds by moving each step to the cluster centers to minimize the clustering error. The main disadvantage of the method lies in its sensitivity to the initial positions of the cluster centers. Therefore, to obtain near-optimal solutions using the k-means algorithm several runs must be scheduled differing in the initial positions of the cluster centers(Likas, Vlassis et al. 2003).

*Clustering Problem Statement:* The overall goal is to do a hierarchical clustering on observations in our dataset, All\_crimes [4:28] (explain the range) and we will essentially compare unsuccessful crime cases (Unsuc\_crimes) to the type of cluster we have generated in the to ascertain why there are more unsuccessful crimes cases in some counties than other counties.

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Hierarchical clustering:

In data mining, hierarchical clustering works by grouping data objects into a tree of the cluster(Murtagh and Contreras 2012). It also groups the observations but does it at a very higher level. It will typically formulate the number of clusters from either descending order or ascending other. This algorithm does not require a pre-specified number of clusters metric to use compared to the K-Means where every distance is a euclidean distance. It gives us room to specify the distance or different distances for this particular method.

Hierarchical methods suffer from the fact that once we have performed either merge or split step, it can never be undone. This inflexibility is useful in that it leads to smaller computation costs by not having to worry about a combination number of different choices. However, such techniques cannot correct mistaken decisions that once have been taken. Two approaches can help in improving the quality of hierarchical clustering: (1) Firstly to perform a careful analysis of object linkages at each hierarchical partitioning or (2) By integrating hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters using another clustering method such as iterative relocation(Rani¹ and Rohil 2013).

### Agglomeration Clustering (AGNES):

This is a bottom-up approach in which each observation starts as its cluster and merges with other observations that are most similar It groups these observations into huge numbers. Agglomerative hierarchical clustering algorithms can be characterized as greedy, in the algorithmic sense. A sequence of irreversible algorithm steps is used to construct the desired data structure. Assume that a pair of clusters, including possible singletons, is merged or agglomerated at each step of the algorithm. Then the following are equivalent views of the same output structure constructed on n objects: a set of *n − 1* partition, starting with the fine partition consisting of *n* classes and ending with the trivial partition consisting of just one class, the entire object set; a binary tree (one or two child nodes at each nonterminal node) commonly referred to as a dendrogram; a partially ordered set (poset) which is a subset of the power set of the n objects; and an ultrametric topology on then objects(Murtagh and Contreras 2012).

### Divisive Hierarchical Clustering: Top-down approach:

All observations are in one group. They start to split the more different (heterogeneous) they are. This keeps on iterating through until each observation is its cluster. In doer words, everything starts with one cluster of all observations then it begins to branch off in a tree-like shape with two or more branches as the number of observations decreases. It puts observations in larger groups. It splits individual observations into clusters producing a set of cluster groups. From this larger set of groups, it can be further clustered into smaller groups. We can keep on iterating it until you get one cluster.

### Distance Used:

This algorithm dynamically uses different forms of distance being Euclidean Dist, Manhattan dist, Maximum Dist, and many more. Distances affect the shape of the cluster

The hierarchical algorithm is characterised by its different forms of ‘linking’ on how to combine different clusters

* 1. Complete linkage clustering: It obtains the distance between ALL the observations in the cluster
  2. 1 and cluster 2 and merge the cluster if the distances between them are a minimum. In other words, it combines them if they are very close. Close to each other than the other clusters.

1. Average Linkage Clustering: This form of clustering link computes the distance between ALL the observations in two clusters cluster 1 and cluster 2 and averages the values.
2. Centroid Linkage Clustering: This link calculates the centroid within each cluster and uses the distance between each centroid to determine the merge.

The overall goal is to do a hierarchical clustering on observations in our dataset, All\_crimes[4:28] (explain the range) and we will essentially compare unsuccessful crime cases (Unsuc\_crimes) to the type of cluster we have generated the to ascertain why there are more unsuccessful crimes cases in some counties than other counties.

***Conclusion****:* In comparing our observation to the cluster we have; we could deduce that Cluster 1 had the highest Unsuccessful Crime convictions and this cluster belongs to Metropolitan and City with an average of 8571.69 unsuccessful crime convictions. 1002.07 average crime convictions were realised from the other counties all grouped in Cluster 2.

The higher magnitude of unsuccessful convictions in Metropolitan and City, which is London and its immediate environs could be the overpopulation in the area as compared to a county like Gloucestershire. Due to this, the authorities should ensure that enough staff is provided in that area to work properly to achieve the required results to promote Justice.

Graphical user interface, chart

Description automatically generated

Timeline

Description automatically generated

**Conclusion:** In comparing our observation to the cluster we have; we could deduce that Cluster 1 had the highest Unsuccessful Crime convictions and this cluster belongs to Metropolitan and City with an average of 8571.69 unsuccessful crime convictions. 1002.07 average crime convictions were released from the other counties all grouped in Cluster 2.

The higher magnitude of unsuccessful convictions in Metropolitan and City, which is London, and its immediate environs could be the overpopulation in the area as compared to a county like Gloucestershire. Due to this, the authorities should ensure that enough staff is provided in that area to work properly to achieve the required results to promote Justice.

### Critical Review of The K-Means and Hierarchical clustering:

Having used both techniques, K-Means was the most effective due to its flexibility as mistakes can be corrected as compared to the robust hierarchical technique. The K-Means technique is also best for a large amount of data as the hierarchical technique suffers to compute a dendrogram if there is a huge amount of data. Notwithstanding these, the K-means has a limitation where the number of clusters, K, must always be determined beforehand as compared to the hierarchical format. Another weakness of the K-means is that it can only handle numerical data but Hierarchical techniques can handle different data types.

## Classification:

Classification and regression trees are machine-learning methods for constructing prediction models from data(Loh 2011). The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error

### Decision Tree

Timeline

Description automatically generated

# General Review of the Analytical Tools

The data analytics tools used are linear regression, k-means, hierarchical clustering, and classification. The least effective is linear regression given that it assumes that the relationship is linear. Hierarchical clustering and k-means do not assume a linear relationship. Between the k-means and the clustering method, the k-means seems to be the more effective method. Notwithstanding, although the K-Means gave a perfect classification when K=2, the accuracy for the overall accuracy confusion metric was 1 for the classification until K=2 was adjusted to k=3.

# Comparing Land Size of Metropolitan and city to Gloucester

The population of Metropolitan and cities affected has an association with the unsuccessful convictions. As compared to those with a lesser population.

## Map of Metropolitan and City (London)

Map

Description automatically generated

#### 

#### Map of Gloucestershire

Map

Description automatically generated

References:

(2005). "Simple Linear Regression." 19-46.

Simple regression is the study of the dependence of a response on one predictor, usually by assuming that the mean of the response as a function of the predictor is a straight line. This chapter presents all the details of the simple linear regression model based on fitting using ordinary least squares. Estimates are derived, and attention is paid to testing, confidence procedures, analysis of variance for simple linear regression, prediction and prediction errors, and graphical methods.

Abt, K. (1987). "Descriptive data analysis: a concept between confirmatory and exploratory data analysis." Methods of information in medicine **26**(02): 77-88.

Chai, C. P. (2020). "The importance of data cleaning: Three visualization examples." Chance **33**(1): 4-9.

Dash, S. S., et al. (2015). Artificial Intelligence and Evolutionary Algorithms in Engineering Systems : Proceedings of ICAEES 2014, Volume 1. Advances in Intelligent Systems and Computing,. New Delhi, Springer India : Imprint: Springer,**:** 1 online resource (XVII, 862 pages 376 illustrations).

The book is a collection of high-quality peer-reviewed research papers presented in Proceedings of International Conference on Artificial Intelligence and Evolutionary Algorithms in Engineering Systems (ICAEES 2014) held at Noorul Islam Centre for Higher Education, Kumaracoil, India. These research papers provide the latest developments in the broad area of use of artificial intelligence and evolutionary algorithms in engineering systems. The book discusses wide variety of industrial, engineering and scientific applications of the emerging techniques. It presents invited papers from the inventors/originators of new applications and advanced technologies.

Diday, E. and J. Simon (1976). Clustering analysis. Digital pattern recognition, Springer**:** 47-94.

Freund, R. J., et al. (2006). Regression analysis, Elsevier.

Lee, R. C. (1981). Clustering analysis and its applications. Advances in information systems science, Springer**:** 169-292.

Likas, A., et al. (2003). "The global k-means clustering algorithm." Pattern recognition **36**(2): 451-461.

Loh, W. Y. (2011). "Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(1): 14-23.

Murtagh, F. and P. Contreras (2012). "Algorithms for hierarchical clustering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(1): 86-97.

Natarajan, K., et al. (2010). Data mining techniques for data cleaning. Engineering Asset Lifecycle Management, Springer**:** 796-804.

Rani¹, Y. and H. Rohil (2013). "A study of hierarchical clustering algorithm." ter S & on Te SIT **2**: 113.

Shinde, G. R., et al. (2022). Exploratory Data Analysis. Quality of Work-Life During Pandemic, Springer**:** 97-105.

Sykes, A. O. (1993). "An introduction to regression analysis."