R Assignment CPS Dataset

Derrick Owusu Ofori-s4113984

20/05/2022

Contents

1. Introduction

In this assignment, I will analyze the Crown Prosecution Service Case Outcomes by Principal Offence Category(CPS). The CPS is in charge of prosecuting criminal cases that have been investigated by the police service and other investigative units in England and Wales.

The data presented from the CPS has been segregated into months from 2014 to 2017. Each dataset contains the results of proceedings in magistrates' courts and the Crown Court. Convictions and unsuccessful outcomes are the two categories of outcomes. The number and percentage of defendants in each category are shown in the report. `Convictions` include guilty pleas, convictions following a trial, and cases proven in the defendant's absence. Discontinuances and withdrawals discharged committals, dismissals and acquittals, and administrative finalizations all fall under the category of `unsuccessful outcomes`. When a case cannot proceed because a warrant for the defendant's arrest has not been served, the defendant cannot be located by the police to serve a summons, or the defendant has died or been judged incompetent to plead, `administrative finalisations` are recorded.

Homicide, personal offences, sexual offences, burglary, robbery, theft and dealing, fraud and forgery, criminal damage, drugs offences, public order, motoring, and all other offences excluding motoring are all classified as offences in the CPS dataset.

2. Objectives

The objectives of this report are to consolidate the monthly data sets, perform some descriptive analysis, test four different hypotheses and implement predictive and clustering techniques. A critical review of the data analytics and visualization tools utilized was also undertaken.

2.1. Hypotheses

Four hypotheses were investigated and tested. These are ;

A. An increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions.

$H0$: **An increase in the number of homicide and robbery convictions will not result in an increase in sexual convictions.**

$H1$: **An increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions.**

$\alpha = 0.05$

B. Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.

$H0$: **Ridge regression will not result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

$H1$: **Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

C. metropolitan areas and cities have higher average convictions than other counties in the United Kingdom.

$H0$: **metropolitan areas and cities do not have higher average convictions than other counties in the United Kingdom.**

$H1$: **metropolitan areas and cities have higher average convictions than other counties in the United Kingdom.**

$\alpha = 0.05$

D. Random forest algorithm will result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.

$H0$: **Random forest algorithm will not result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.**

$H1$: **Random forest algorithm will result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.**

3. Data importation and preprocessing

3.1 Data Importation, Integration and feature engineering

To facilitate easy data wrangling, the monthly datasets for each year were extracted and consolidated into one single dataset. To identify observations based on year and month, date information was extracted from the file pathname, and feature engineered into the dataset as the variables, `year`, `month` and `date`.

```
##   year month       date
## 1 2014   Apr 2014-04-01
## 2 2014   Apr 2014-04-01
## 3 2014   Apr 2014-04-01
## 4 2014   Apr 2014-04-01
## 5 2014   Apr 2014-04-01
## 6 2014   Apr 2014-04-01
```

The data set ended up with 1806 observations and 54 variables.

```
## [1] 1806   54
```

3.2 Data Cleaning

The R programming language recognizes variables with the per cent sign (%), the minus sign signifying zero percentage (-), and commas(,) as characters, making numerical computations impossible. As such all instances of percentages and commas were removed and the minus sign replaced with 0.0. The respective variables were then made numerical. The variables with county renamed from `X` to `County.Name`

This output of the first 40 rows and 4 columns demonstrates the data cleaning applied.

County.Name	Number.of.Homicide.Convictions	Percentage.of.Homicide.Convictions	Number.of.Homicide.Unsuccessful
National	81	85.3	14
Avon and Somerset	1	100.0	0
Bedfordshire	0	0.0	0
Cambridgeshire	0	0.0	0
Cheshire	1	50.0	1
Cleveland	0	0.0	0
Cumbria	0	0.0	0
Derbyshire	0	0.0	0
Devon and Cornwall	1	100.0	0
Dorset	0	0.0	0
Durham	2	100.0	0
Dyfed Powys	0	0.0	0
Essex	1	100.0	0
Gloucestershire	0	0.0	0
GreaterManchester	1	100.0	0
Gwent	0	0.0	0
Hampshire	2	100.0	0
Hertfordshire	1	100.0	0
Humberside	0	0.0	0

County.Name	Number.of.Homicide.Convictions	Percentage.of.Homicide.Convictions	Number.of.Homicide.Unsuccessful
Kent	9	64.3	5
Lancashire	1	100.0	0
Leicestershire	4	80.0	1
Lincolnshire	0	0.0	0
Merseyside	5	100.0	0
Metropolitan and City	21	91.3	2
Norfolk	1	100.0	0
Northamptonshire	2	100.0	0
Northumbria	3	100.0	0
North Wales	1	100.0	0
North Yorkshire	1	50.0	1
Nottinghamshire	2	100.0	0
South Wales	0	0.0	0
South Yorkshire	2	100.0	0
Staffordshire	0	0.0	0
Suffolk	1	100.0	0
Surrey	0	0.0	0
Sussex	0	0.0	0
Thames Valley	9	100.0	0
Warwickshire	0	0.0	0
West Mercia	1	50.0	1

Additionally The sum of all missing data is also zero.

```
## [1] 0
```

4. Descriptive analysis

4.1 Summary statistics

To commence descriptive analysis, some summary statistics were assessed, namely; the mean, mode, maximum and minimum. The mean summarizes an entire dataset with a single number representing the data's centre point or typical value. The standard deviation on the other hand measures how spread out data is relative to its mean and is simply the square root of the variance. The maximum and minimum values show the most prominent and smallest values per offence category.

The number of offences again the person takes the highest mean value of about 234. This means that this type of offence has the most monthly occurrences on average. In comparison to other convictions, its standard deviation demonstrates that the data are considerably spaced out from the mean. For this offence, the maximum and minimum sentences were 1904 and 29, respectively.

Another measure of central tendency is the median. The strength of the median is that it is less affected by outliers and skewed data as compared to the mean. This is also a good measure of central tendency when the data distribution is not symmetrical. An advantage of the mean over the median is that since it is affected by every value in a data it can be said to truly reflect the central tendency of that data unlike the median.

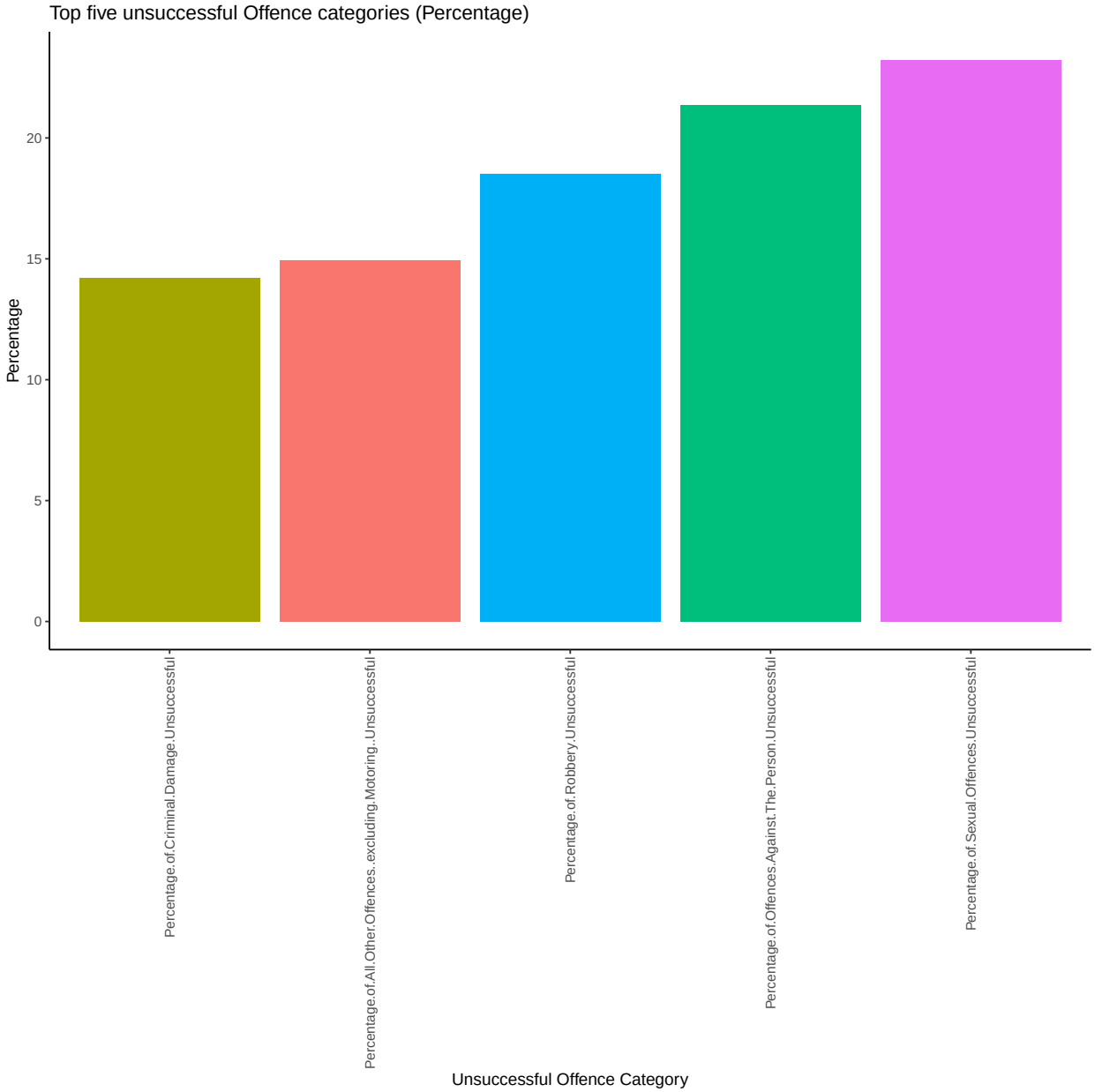
The variance is an alternative measure of spread. The merit of variance is that it treats all departures from the mean equally, regardless of direction. Because the squared deviations cannot equal zero, the data appears to have no variability. However, the weakness of variance is that it is more impacted by outliers within the data.

Offense	mean	SD	max	min
Number.of.Homicide.Convictions	1.88	3.22	38	0
Number.of.Homicide.Unsuccessful	0.46	1.30	14	0
Number.of.Offences.Against.The.Person.Convictions	234.04	234.20	1904	29
Number.of.Offences.Against.The.Person.Unsuccessful	71.22	102.81	862	5
Number.of.Sexual.Offences.Convictions	22.32	24.56	181	0
Number.of.Sexual.Offences.Unsuccessful	8.41	12.15	107	0
Number.of.Burglary.Convictions	31.81	34.13	278	1
Number.of.Burglary.Unsuccessful	5.49	8.64	83	0
Number.of.Robbery.Convictions	10.24	19.37	209	0
Number.of.Robbery.Unsuccessful	2.77	6.38	61	0
Number.of.Theft.And.Handling.Convictions	201.18	181.28	1426	13
Number.of.Theft.And.Handling.Unsuccessful	18.18	25.61	220	0
Number.of.Fraud.And.Forgery.Convictions	19.70	34.04	299	0
Number.of.Fraud.And.Forgery.Unsuccessful	3.23	6.54	58	0
Number.of.Criminal.Damage.Convictions	51.00	45.55	400	3
Number.of.Criminal.Damage.Unsuccessful	8.87	10.29	85	0
Number.of.Drugs.Offences.Convictions	98.40	155.14	1228	4
Number.of.Drugs.Offences.Unsuccessful	6.57	12.94	111	0
Number.of.Public.Order.Offences.Convictions	85.63	88.41	779	2
Number.of.Public.Order.Offences.Unsuccessful	15.20	22.85	194	0
Number.of.All.Other.Offences..excluding.Motoring..Convictions	36.26	61.28	551	0
Number.of.All.Other.Offences..excluding.Motoring..Unsuccessful	6.87	13.36	129	0
Number.of.Motoring.Offences.Convictions	191.28	194.78	1889	1
Number.of.Motoring.Offences.Unsuccessful	31.83	52.25	491	0
Number.of.Admin.Finalised.Unsuccessful	19.45	33.92	419	0

4.2 Data visualization

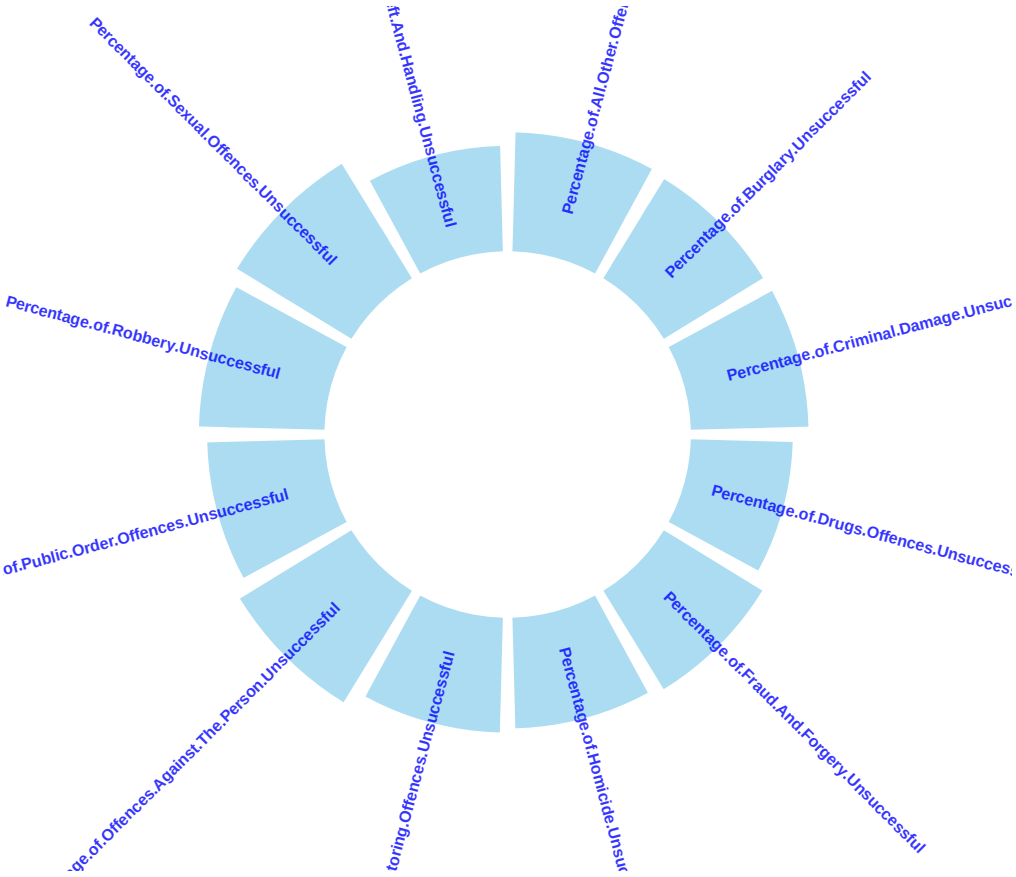
4.2.1 Average percentage unsuccessful Offence (Boxplot & Circular Barplot)

An area of interest was finding the offence categories the CPS was most unsuccessful in convicting. To find this, the average of all percentage unsuccessful convictions was extracted and the top five were visualized using a bar plot graph. A bar graph is a chart or graph that displays categorical data in the form of rectangular bars with weights proportional to the values they represent. From the plot, I see that sexual offence convictions had the highest unsuccessful percentage of about 22%. Sexual offences are indeed difficult to convict. With regards to rape for instance, In the year ending March 2020, 99% of rapes reported to police in England and Wales resulted in no legal proceedings against alleged attackers[1].



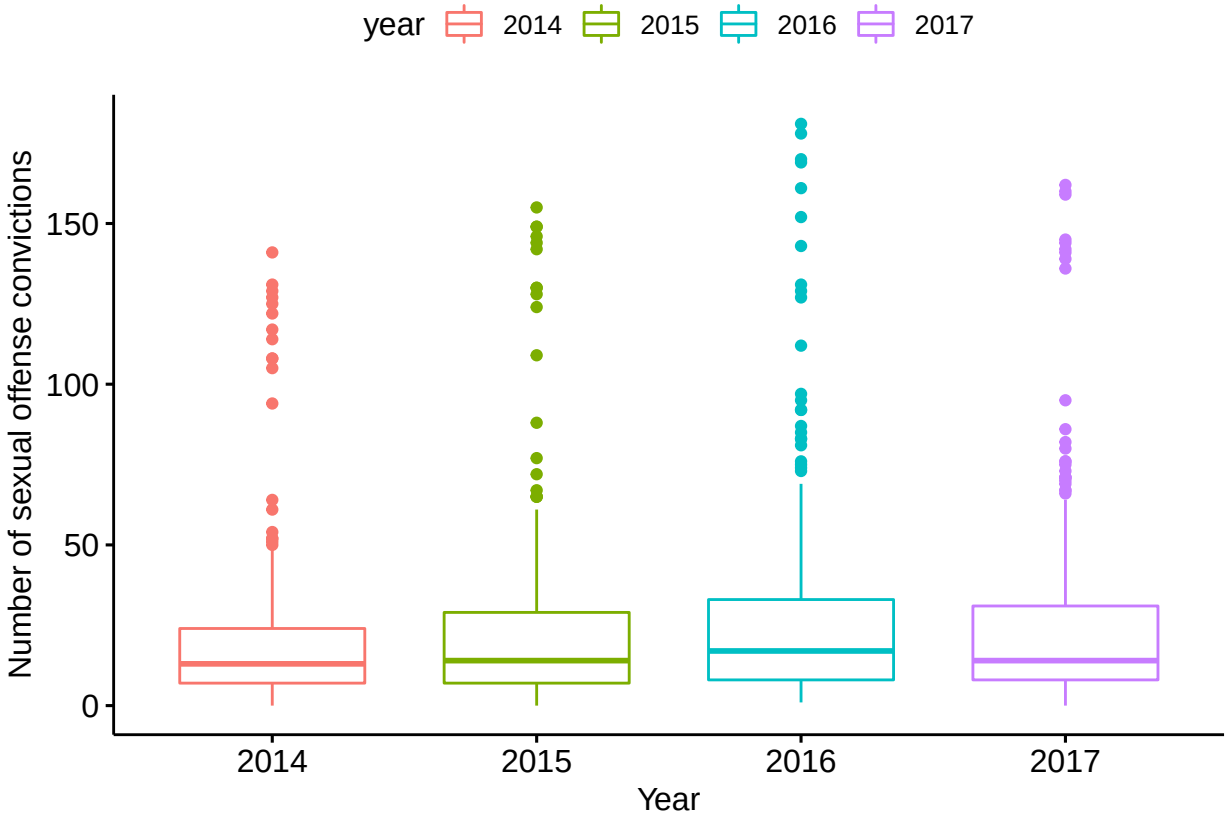
Following sexual harassment, other crimes with a high unsuccessful conviction rate are offences against the

person, robbery, and all Other Offences excluding motoring and criminal damage. But what about the other offence categories? A limitation of bar plots is that when visualizing multiple variables, stacking them all on the X-axis can make the graph seem overpopulated, and as such necessitate the visualization of a few variables. This however can be resolved with a circular bar plot. I can visualize not just the five but all the offence categories at the same time on a 360-degree axis. I can see the hierarchy of unsuccessful convictions by comparing the respective heights of the bars. The weakness of the circular bar plot however is that is no y-axis to infer a corresponding value.

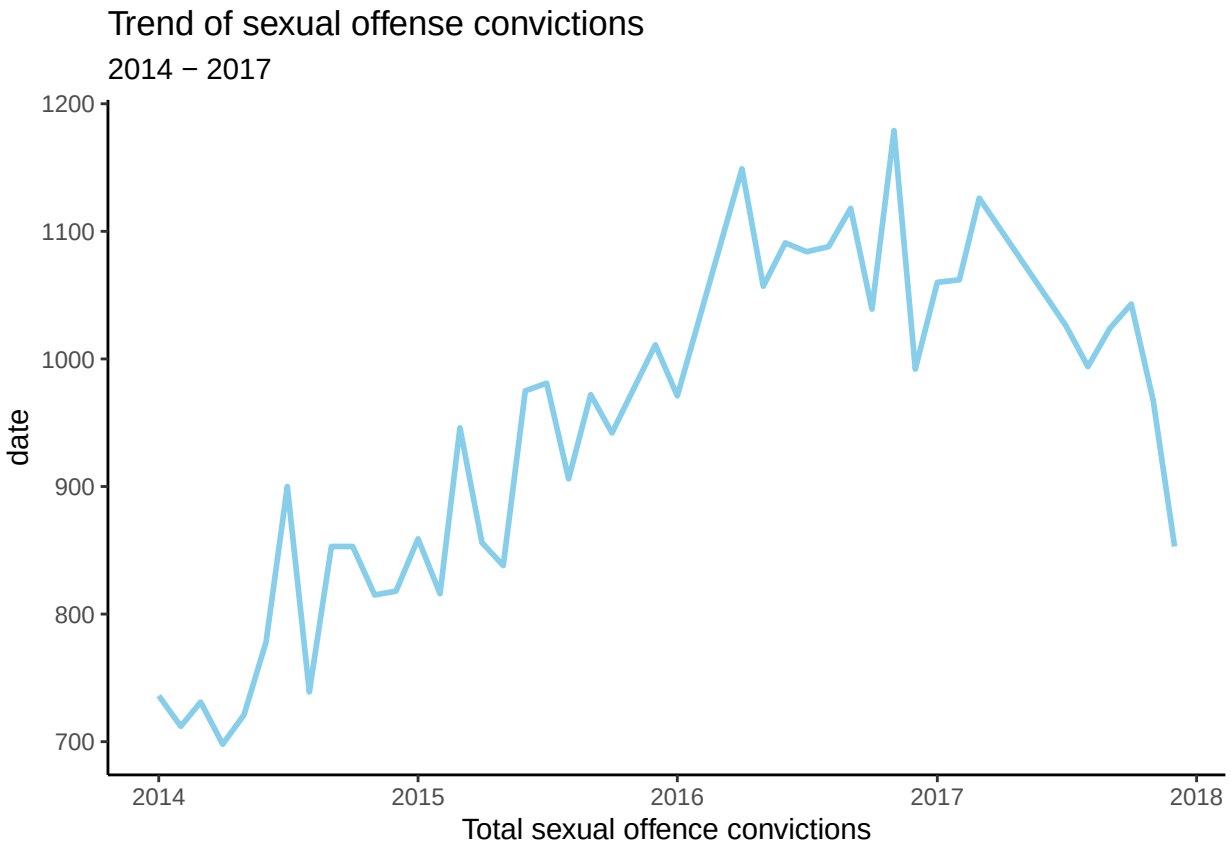


4.2.2 Distribution and trend of yearly sexual offence convictions. (Boxplot, Line graph & Line graph with Vline)

To gather further insight into sexual offences, a yearly boxplot was constructed to visualize the distribution of the number of sexual offence convictions. Box plots use quartiles (or percentiles) and averages to visually depict the distribution of numerical data and skewness. The minimum value, first (lower) quartile, median, third (upper) quartile and maximum value are all shown in box plots. The yearly sexual offence was on the rise since 2014 and in 2016 had a median value of about 20 a lower quartile of about 15 and an upper quartile of about 27. However, in 2017 I see that the median sexual conviction dropped to about 17, with other summary statistics seeing a decline. To understand why I must overcome the limitations of the box plot.

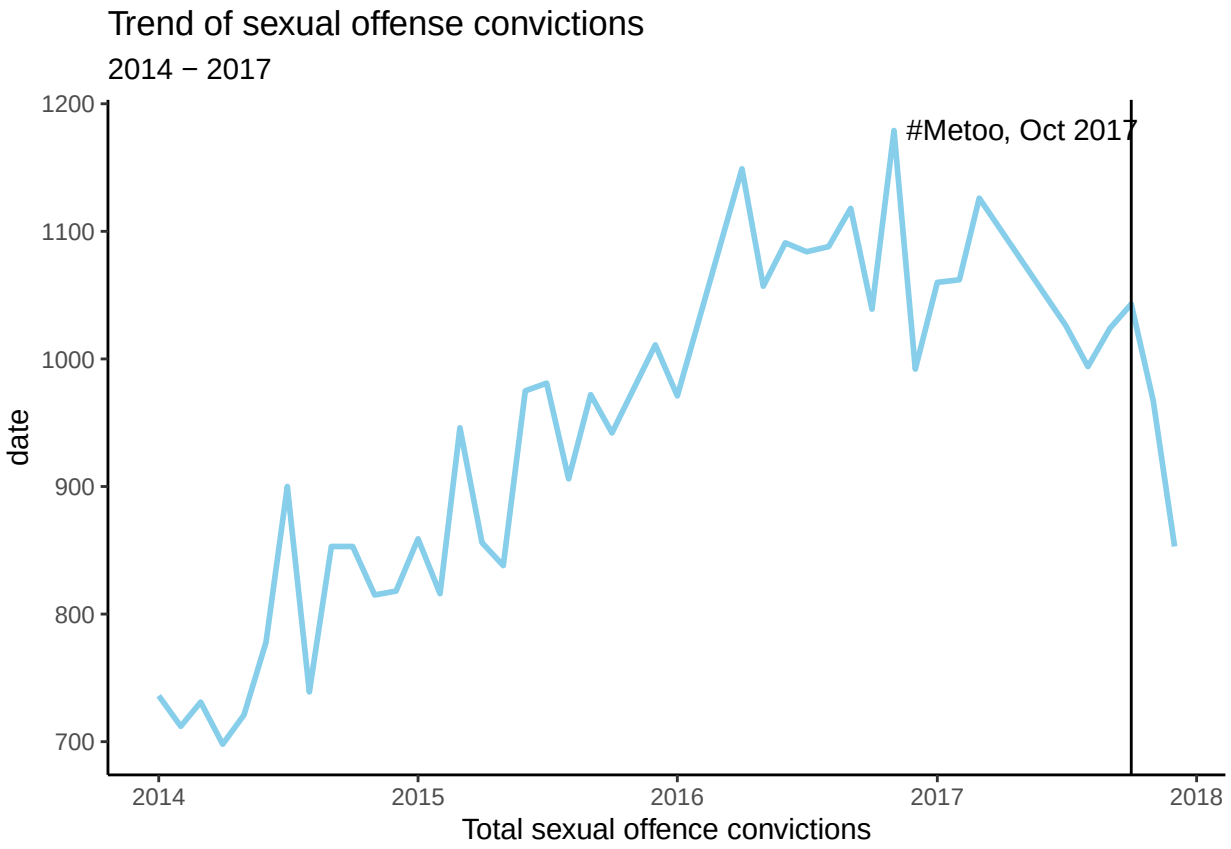


Although the box plot is good at portraying several statistical values at once, it doesn't truly capture the dynamic changes in sexual conviction over time. The line graph on the other hand is designed to fulfil that purpose. A line graph displays quantitative values over a specified time interval. Using the line graph I can truly capture the trend in sexual convictions and see its peaks and slumps.



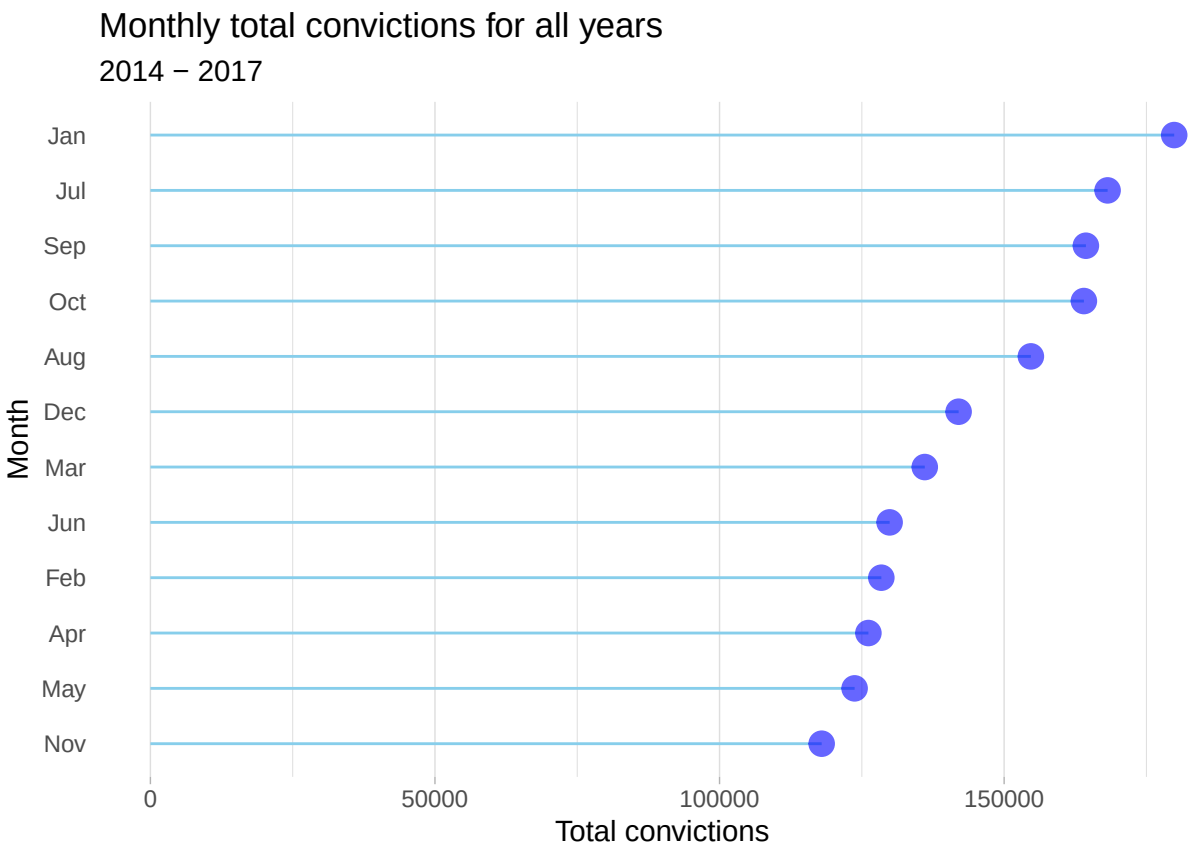
Although the line graph is limited to showing a single value across time, it has brought forth an interesting insight. Form the line graph to see that although sexual convictions rise and fall over the years. However, I witnessed a sudden fall around the latter part of 2017.

One global phenomenon that took place at the same time this fall is the `#Metoo Movement`. In October 2017 the `#Metoo movement` took various social media platforms by storm[2]. Women from all works of life voiced out about the frequency with which they experience sexual assault and harassment. This drew worldwide attention and had rippling effects across several industries. The vertical line on the graph shows when the movement commences, and since this time point, sexual offence convictions saw a sharp fall. In the USA for instance sexual harassment at work saw a decrease since the inception of the `#Metoo` movement[3]. Similar research in the United Kingdom needs to be conducted before I can attribute this fall in sexual offence convictions to the `Metoo` movement, however, it paints an optimistic outlook for protection again sexual offences.

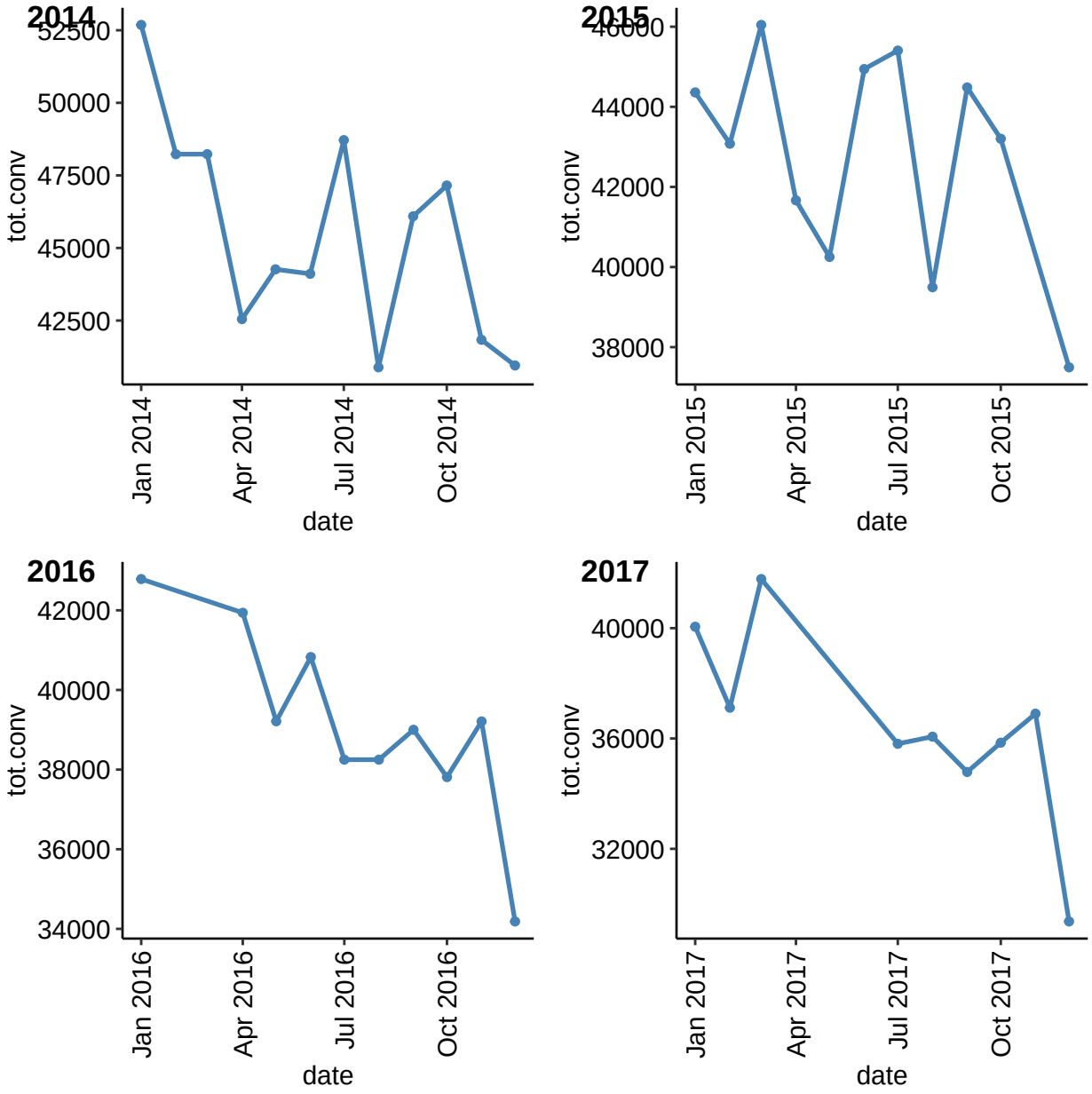


4.2.3 Monthly and yearly total conviction pattern (Lollipop chart, Grouped line graph & Stacked area chart)

Another interesting insight I can draw is finding out which months had to most total convictions. I will do this by utilizing a lollipop plot which is essentially a bar plot with the bar converted into a line and a dot. It depicts the link between two numerical and categorical variables. I realize that January seems to lead with an excess of 150,000 total convictions. This was tailed by July and September. But why does January seem to lead all other months in terms of convictions?

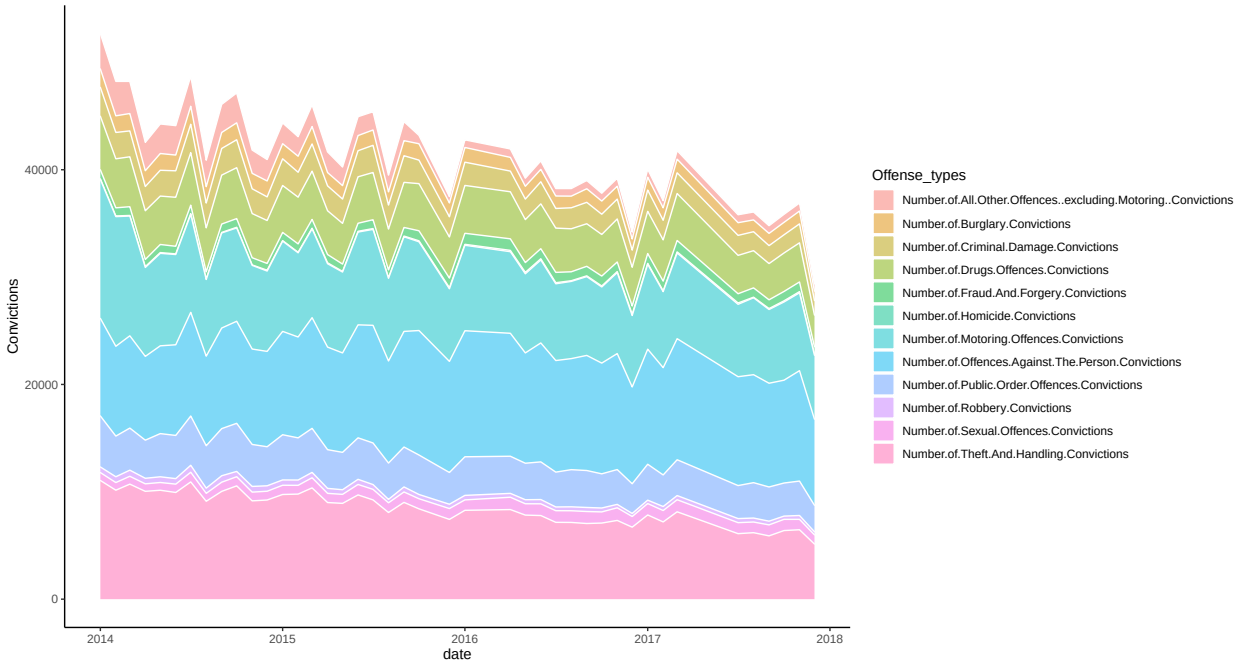


The lollipop plot like the bar plot can only show the length of a bar and its corresponding values, depriving further insight. Once again an alternative visualization tool, the grouped line graph can be utilized. Just like a regular line plot, I will assess a quantitative metric across time, and in this case, I will take it on a year by year bases visualizing trends across each month of each year.



In 2014 the total conviction was 52,000. Total convictions trended downwards after this to a low of about 41,000 at the end of the year. A similar pattern seemed to occur each year, where total convictions start high but then after some undulations reach a low at the end of the year. This indicates that January simply has a high number of total convictions because it is simply carrying over the previous year's lows to represent its high. To get a more holistic picture an alternative visualization may prove useful. The line graphs presented so far show the quantitative changes in a single variable. However, a stacked area chart shows all relationships change over time. The stacked area chart however can become quite overwhelming to read when the variables under consideration are large.

From the stacked area chart an interesting observation can be drawn. Every single conviction seems to follow a downward trend over time. In essence, total convictions by the CPS are taking are reducing overtime. People are simply committing fewer crimes over time. Factors contributing to this might include education, labour market policies and increased spending on police resources[4].


Convictions
40000
20000
0
2014
2015
2016
2017
2018
date
Offense_types
Number.of.All.Other.Offences..excluding.Motoring..Convictions
Number.of.Burglary.Convictions
Number.of.Criminal.Damage.Convictions
Number.of.Drugs.Offences.Convictions
Number.of.Fraud.And.Forgery.Convictions
Number.of.Homicide.Convictions
Number.of.Motoring.Offences.Convictions
Number.of.Offences.Against.The.Person.Convictions
Number.of.Public.Order.Offences.Convictions
Number.of.Robbery.Convictions
Number.of.Sexual.Offences.Convictions
Number.of.Theft.And.Handling.Convictions

5. Hypothesis testing

5.1 An increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions.

`H0:` **An increase in the number of homicide and robbery convictions will not result in an increase in sexual convictions.**

`H1:` **An increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions.**

$\alpha = 0.05$

To investigate this hypothesis, I will utilize multiple linear regression. This kind of regression takes multiple features and inputs and tries to predict the dependent varibale. It follow the same assumptions as the linear regression and seeks to determine a postive relationship between the predictor and predicted variables in the data.

$yi = \beta 0 + \beta 1x1 + \beta 2x2... + \beta pxip + e$

where $i = n$ observations: yi = the dependent variable $x1$ = the explanatory variable $\beta 0$ =y - intercept (constant term) βp = the slope coefficients for each independent variable e = the model error term or residuals.

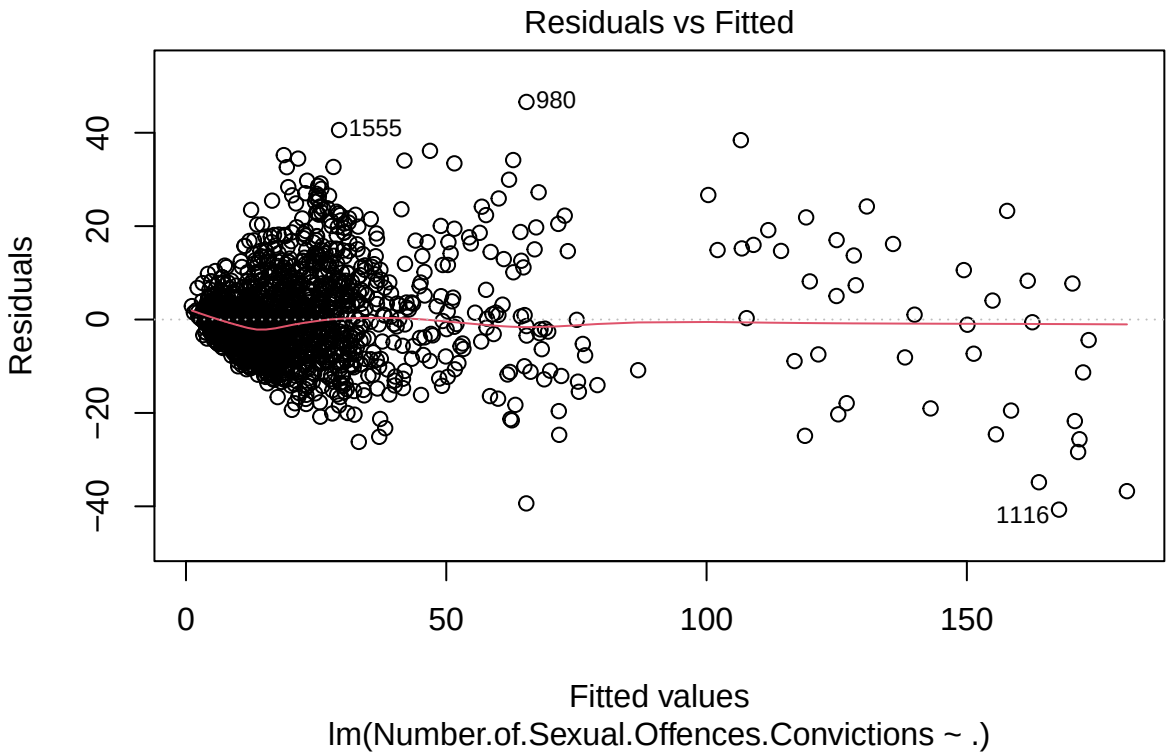
By utilizing multiple linear regression I can indeed find out if there is a positive relationship between homicide and robbery and consequently sexual convictions. However before this method can be applied, certain assumptions must be checked. These are:

1. Linear relationship: Each predictor variable and the responder variable have a linear relationship.
2. Multivariate Normality: The model's residuals have a normal distribution.
3. Homoscedasticity: At every point in the linear model, the residuals have the same variance.
4. Autocorrelation: Residuals' error terms are independent.
5. No multicollinearity: Independent variables are not highly correlated with each other.

To check these assumptions, a multiple linear regression model was constructed using sexual convictions as the dependent variable and another number of sexual convictions as the independent variable. Several diagnostic plots were then piloted from the model and used to test the assumptions.

5.1.1 Linear relationship

The dependent variable and the independent factors must have a linear relationship. To check this the residuals vs fitted graphic can be used to verify the linearity assumption.

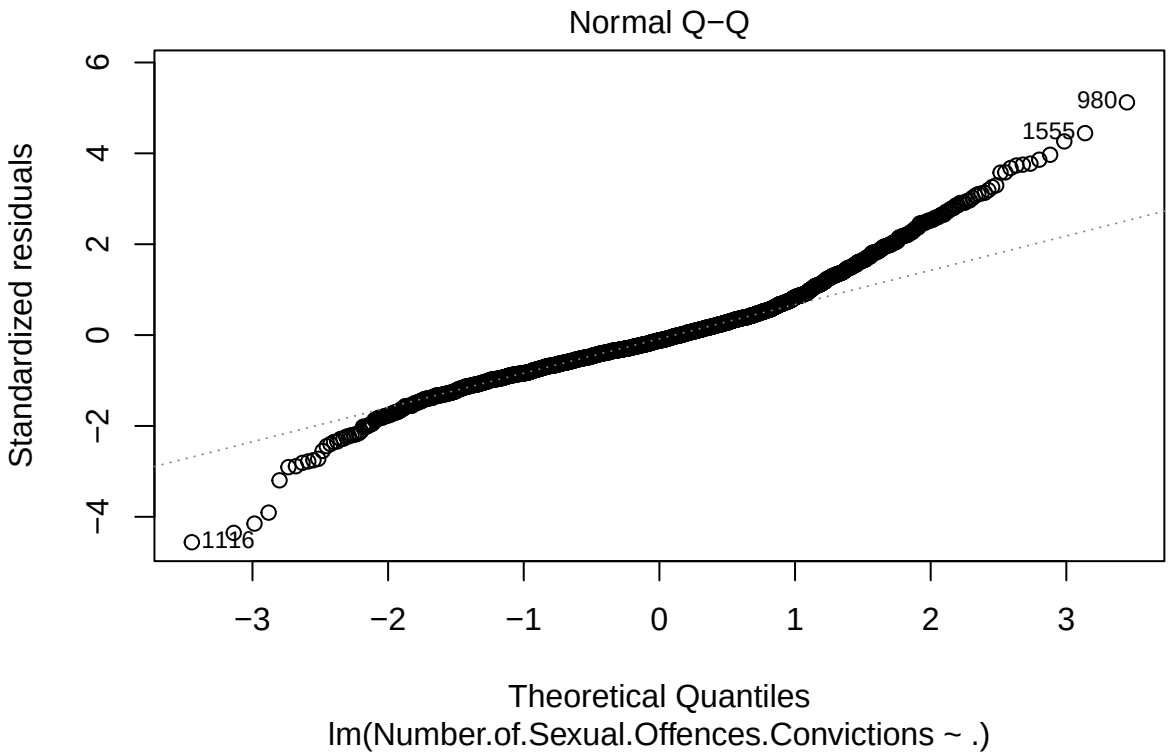


The residual plot should, in theory, exhibit no fitted pattern. That is, at 0, the red line should be nearly horizontal. The appearance of a pattern could suggest a fault with the linear model in some way. Our plot on the other had shown no pattern and as such has met the first assumption.

5.1.2 Multivariate Normality

To visually check the normality assumption, I utilized the QQ plot of residuals. The residual normal probability plot should roughly follow a straight line.

Because all of the points in our case don't fall roughly along this reference line, I cannot infer normalcy. This will be corrected with transformation later on.



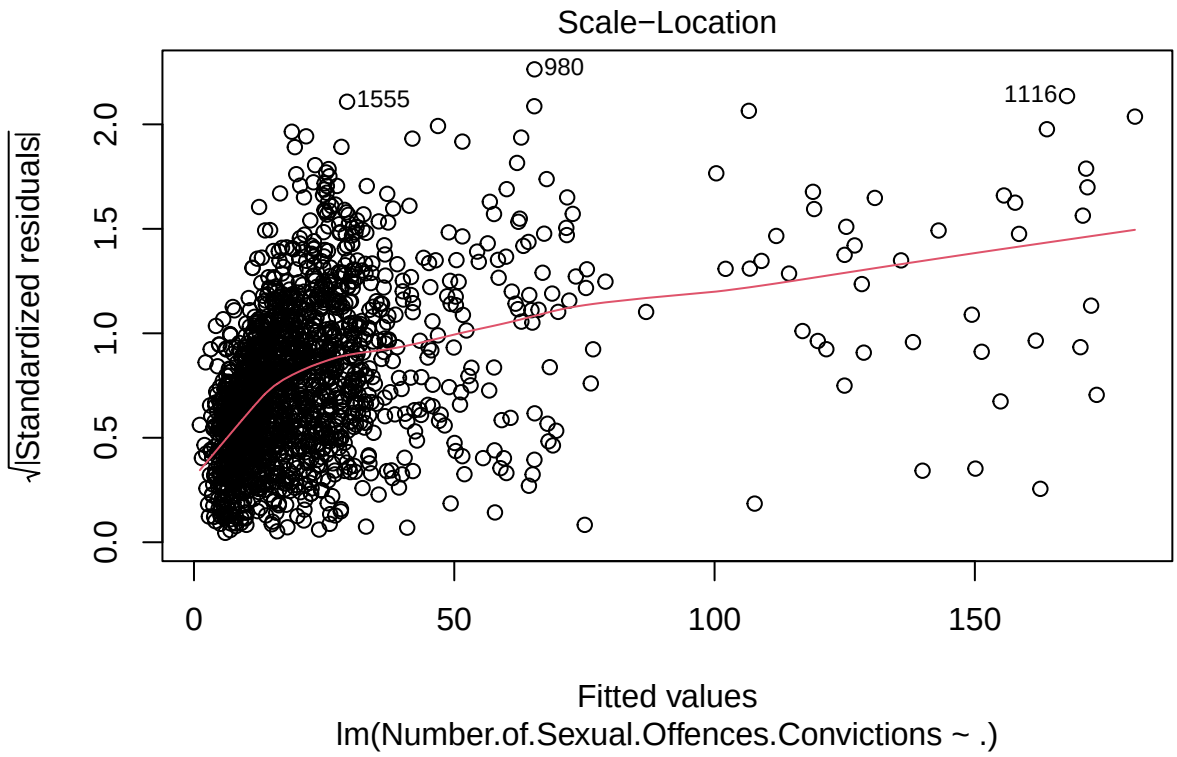
5.1.3. Homoscedasticity

To test this assumption, utilise the scale-location plot, also known as the spread-location plot.

When residuals are dispersed evenly across predictor ranges, this graph depicts what happens. It's a good indicator if you see a horizontal line with equally spaced points. In our case, however, this is not the case.

The residual points' variability (variances) increases as the fitted outcome variable's value increases, meaning that the residual errors have non-constant variances (or heteroscedasticity).

To reduce heteroscedasticity, the data could be transformed into a log or square root format.



5.1.4. Autocorrelation

Multiple linear regression assumes that each observation in the data set is independent. This can be determined with a Durbin-Watson test, indicates if residuals or variables show some form of autocorrelation.

this test uses the following hypotheses:

$H0$: There is no autocorrelation among the residuals.

$H1$: The residuals are autocorrelated.

$\alpha = 0.05$

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.008479216      2.013045   0.756
##  Alternative hypothesis: rho != 0
```

From the output, I can see that the test statistic is 2.013045 and the corresponding p-value is 0.782. Since this p-value is more than 0.05, I can fail to reject the null hypothesis and conclude that the residuals in this regression model are not autocorrelated.

5.1.5. No multicollinearity assumption.

Multicollinearity occurs when the independent variables are highly linked, making it impossible to isolate their influence on the outcome variable. To put it another way, one of the predictor variables can almost completely predict another predictor variable. This assumption can be tested using the Variance Inflation Factor(VIF). When the VIF is above 5 then there is multicollinearity.

```
##                                                  Variables       VIF
## 1                                Number.of.Homicide.Convictions  2.743416
## 2             Number.of.Offences.Against.The.Person.Convictions 20.166259
## 3                                Number.of.Burglary.Convictions 14.422758
## 4                                 Number.of.Robbery.Convictions  8.914715
## 5                       Number.of.Theft.And.Handling.Convictions 14.160046
## 6                        Number.of.Fraud.And.Forgery.Convictions 17.069539
## 7                          Number.of.Criminal.Damage.Convictions 19.236453
## 8                           Number.of.Drugs.Offences.Convictions 22.222686
## 9                    Number.of.Public.Order.Offences.Convictions 14.450624
## 10 Number.of.All.Other.Offences..excluding.Motoring..Convictions  6.398277
## 11                       Number.of.Motoring.Offences.Convictions  6.503109
```

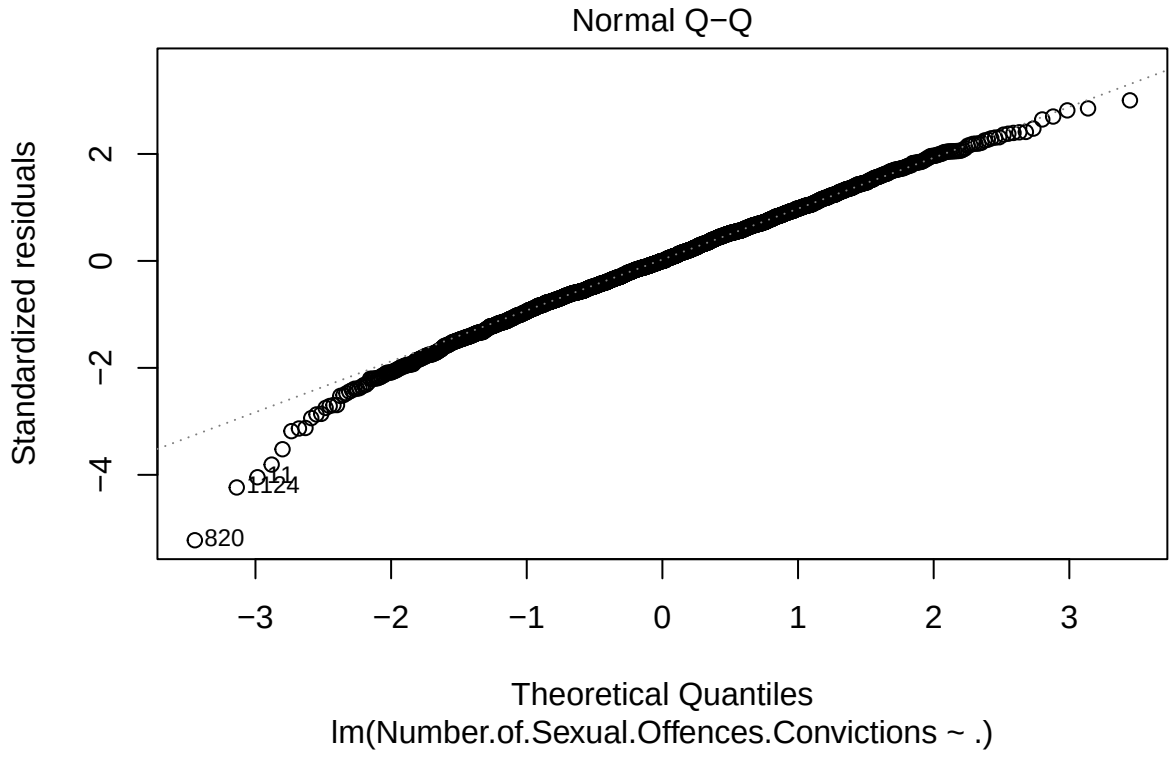
Most of the variables have a VIF of more than 5. Transformation may be applied to remedy this.

5.1.6 Cube root transformation to meet linear assumptions

To meet the assumption requirements, data transformation is very useful. For this case we will use a cube root transformation. Transformation enables non-linear data to be made more linear. The cube root transformation was tried and make the data more linear and a new multiple linear regression was constructed.

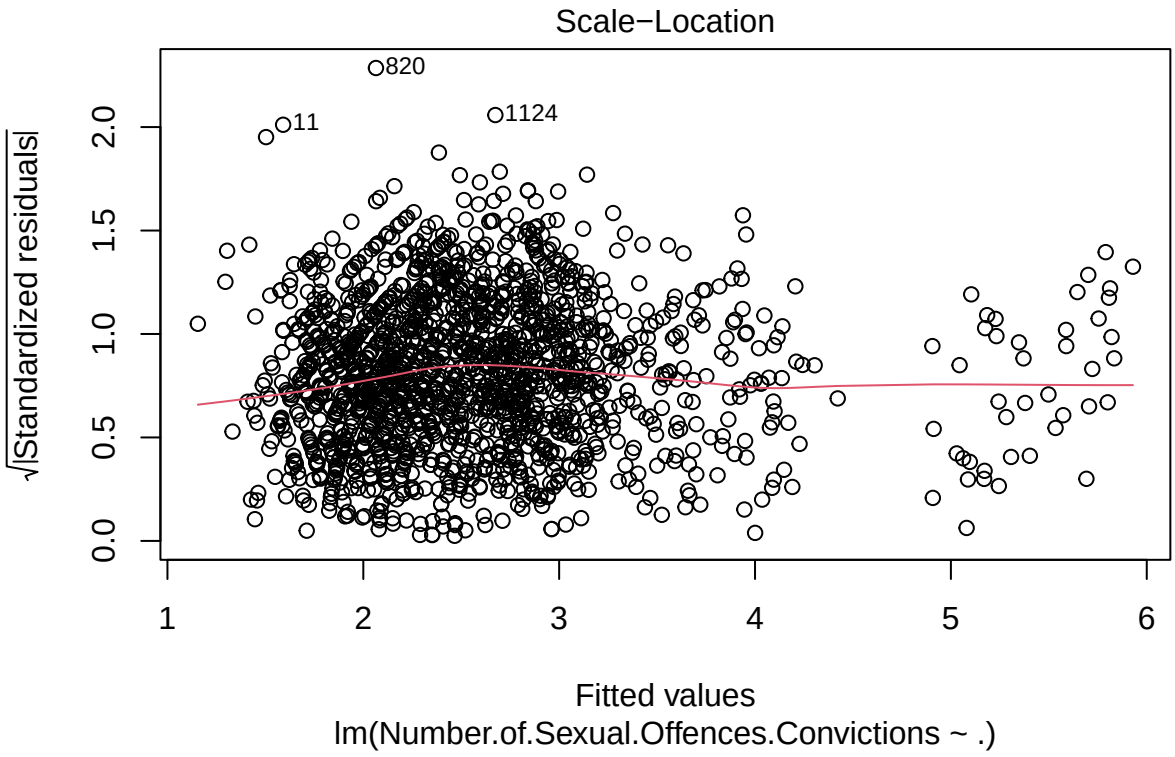
5.1.7 Multivariate normality assumption met

After applying the cube root transformation, the normality of the residuals has greatly improved as seen by its alignment on the diagonal line.



5.1.8 Homoscedasticity assumption met

The residual points' variability (variances) have levelled off and are nearly constant with the value of the fitted outcome variable, meaning that the residual errors have a constant variance. (Homoscedastic)



5.1.9 Multicollinearity rectified

The VIF for all variables shrunk after variables transformation. However, some of them are still larger than 5. These variables will be removed to get rid of multicollinearity.

```
##                                                  Variables       VIF
## 1                                Number.of.Homicide.Convictions  1.517829
## 2             Number.of.Offences.Against.The.Person.Convictions 12.338253
## 3                                Number.of.Burglary.Convictions  6.736244
## 4                                 Number.of.Robbery.Convictions  3.020976
## 5                      Number.of.Theft.And.Handling.Convictions 10.645811
## 6                       Number.of.Fraud.And.Forgery.Convictions  4.949779
## 7                         Number.of.Criminal.Damage.Convictions 10.903673
## 8                          Number.of.Drugs.Offences.Convictions  7.755476
## 9                  Number.of.Public.Order.Offences.Convictions  8.038214
## 10 Number.of.All.Other.Offences..excluding.Motoring..Convictions  3.418363
## 11                     Number.of.Motoring.Offences.Convictions  5.428450
```

5.1.10 Hypothesis testing results

In our final model, four variables have a significant linear relationship with the number of sexual convictions; `Number of Homicide Convictions`, `Number of Robbery Convictions`, `Number of Fraud And Forgery Convictions` and `Number of All Other Offences excluding Motoring Convictions`.

The variables under consideration for our hypothesis however were `Number of Homicide Convictions`, and `Number of Robbery Convictions`. The coefficient of the number of homicide convictions is 0.17709, this means that holding all other variables constant, a unit increase in homicide convictions will lead to sexual convictions increasing by 0.17709. For the number of robbery convictions, holding all other variables constant, a unit increase in homicide convictions will lead to sexual convictions increasing by 0.17318.

With both p-values of the number of homicide and robbery convictions being far below $\alpha = 0.05$,

I reject the null hypothesis

$H0$: **An increase in the number of homicide and robbery convictions will not result in an increase in sexual convictions.**

and accept the alternative hypothesis

$H1$: **An increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions.**

```
##
## Call:
## lm(formula = Number.of.Sexual.Offences.Convictions ~ ., data = number_sex_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3397 -0.3191 -0.0012  0.3157  1.5010
##
## Coefficients:
##                                                          Estimate
## (Intercept)                                               0.70421
## Number.of.Homicide.Convictions                            0.17709
## Number.of.Robbery.Convictions                             0.17318
## Number.of.Fraud.And.Forgery.Convictions                   0.51736
## Number.of.All.Other.Offences..excluding.Motoring..Convictions  0.05499
##                                                          Std. Error
## (Intercept)                                                 0.03885
## Number.of.Homicide.Convictions                              0.01931
## Number.of.Robbery.Convictions                               0.02121
## Number.of.Fraud.And.Forgery.Convictions                     0.02299
## Number.of.All.Other.Offences..excluding.Motoring..Convictions   0.01394
##                                                          t value Pr(>|t|)
## (Intercept)                                               18.124  < 2e-16
## Number.of.Homicide.Convictions                             9.172  < 2e-16
## Number.of.Robbery.Convictions                              8.165 6.08e-16
## Number.of.Fraud.And.Forgery.Convictions                   22.505  < 2e-16
## Number.of.All.Other.Offences..excluding.Motoring..Convictions   3.945 8.29e-05
##
## (Intercept)                                              ***
## Number.of.Homicide.Convictions                           ***
## Number.of.Robbery.Convictions                            ***
## Number.of.Fraud.And.Forgery.Convictions                  ***
## Number.of.All.Other.Offences..excluding.Motoring..Convictions ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4865 on 1759 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6433
## F-statistic: 795.8 on 4 and 1759 DF,  p-value: < 2.2e-16
```

5.2 Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.

$H0$: **Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

$H1$: **Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

For this hypothesis, I seek to find the best regression model to predict sexual convictions. Since I have already constructed a linear model, I will expatiate on the ridge regression.

As I show in the previous hypothesis, our data suffered from multicollinearity. A powerful solution to this issue is the Ridge Regression. Ridge Regression is a technique for studying multicollinear data in multiple regression models. When there is multicollinearity, least squares estimates are unbiased, but their variances are huge, thus they could be far off the true value. Ridge regression reduces standard errors by adding a degree of bias to the regression estimates[5]. In ridge regression, I introduce a parameter lambda. The model coefficient changes as the tuning parameter lambda changes, this results in reducing the impact of multicollinearity within the variables.

To evaluate the prediction ability of both the ridge and linear regression I will use the root mean squared error (RMSE) as a tool to determine the model with the least error rate. RMSE uses Euclidean distance to demonstrate how far predictions differ from observed true values, and in this case will be applied to both the predictions of the ridge and linear regression, the closer the RMSE number is to zero, the more precise the sample estimate.

I will split the data into a train and test set with a ratio of 70%:30%. The train set will be used to train both models and then predictions tested with the dependent variable in the test set. The data will be transformed to meet the assumptions of linear regression as discussed in the previous hypothesis.

5.2.1 Linear regression

For linear regression, I obtained an R- squared of 0.63. This implies that about 63% of the variations in the dependent variable are explained by the linear model.

```
##
## Call:
## lm(formula = Number.of.Sexual.Offences.Convictions ~ ., data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35203 -0.31939 -0.00121  0.31500  1.48765
##
## Coefficients:
##                                                          Estimate
## (Intercept)                                               0.70776
## Number.of.Homicide.Convictions                            0.17470
## Number.of.Robbery.Convictions                             0.16521
## Number.of.Fraud.And.Forgery.Convictions                   0.51602
## Number.of.All.Other.Offences..excluding.Motoring..Convictions  0.06553
##                                                          Std. Error
## (Intercept)                                                 0.04750
## Number.of.Homicide.Convictions                              0.02317
## Number.of.Robbery.Convictions                               0.02552
## Number.of.Fraud.And.Forgery.Convictions                     0.02850
```

```
## Number.of.All.Other.Offences..excluding.Motoring..Convictions    0.01686
##                                                                t value Pr(>|t|)
## (Intercept)                                                     14.899  < 2e-16
## Number.of.Homicide.Convictions                                   7.539 9.16e-14
## Number.of.Robbery.Convictions                                    6.473 1.38e-10
## Number.of.Fraud.And.Forgery.Convictions                         18.108  < 2e-16
## Number.of.All.Other.Offences..excluding.Motoring..Convictions   3.887 0.000107
##
## (Intercept)                                                    ***
## Number.of.Homicide.Convictions                                 ***
## Number.of.Robbery.Convictions                                  ***
## Number.of.Fraud.And.Forgery.Convictions                        ***
## Number.of.All.Other.Offences..excluding.Motoring..Convictions ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4867 on 1227 degrees of freedom
## Multiple R-squared:  0.6321, Adjusted R-squared:  0.6309
## F-statistic: 527.1 on 4 and 1227 DF,  p-value: < 2.2e-16
```

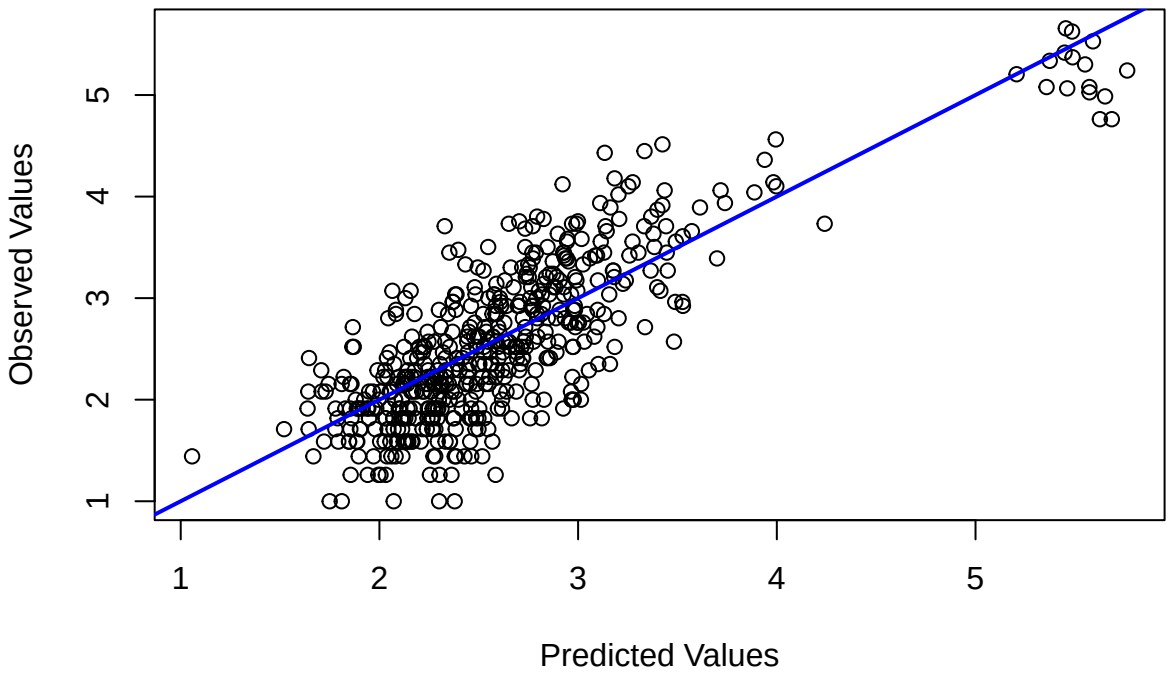
I can see an overview of the actual vs predicted outputs for the linear regression below.

```
##      actual  pred_lm
## 4  2.571282 2.589355
## 5  2.223980 2.126912
## 15 1.817121 2.113066
## 21 1.817121 2.527550
## 27 2.154435 3.015154
## 28 1.912931 2.461552
```

I was able to obtain an RMSE of 0.487053, which is quite close to 0 and stands as a good result.

```
## [1] 0.487053
```

Finally, by plotting the predicted vs the actual I can see a good linear relationship, indicating that the model is

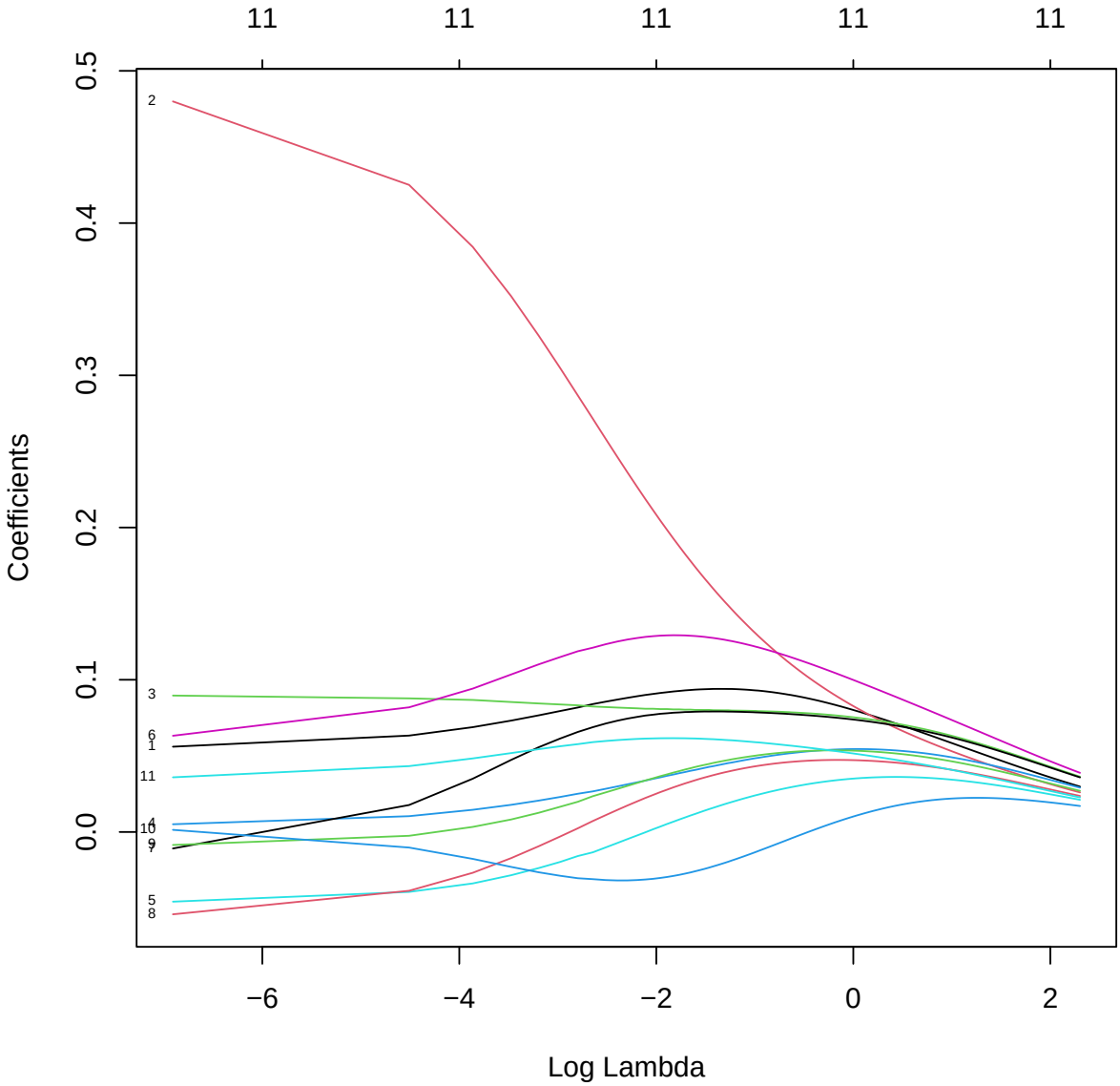


a good one.

5.2.2 Ridge regression

Next, I use ridge regression. The lambdas utilized range from 0.001 to 10 by a 0.001 step. I will first test the median lambda value and then find the optimal using cross-validation.

A plot of the model coefficients shows that I am working with a total of 11 variables as can be seen at the top of the plot. I also see that some variables shrink as lambda increases. Variables that don't shrink as quickly like variables 3 and 5 are more relevant to the model. Some of our least significant like 11 and 7 go towards 0 much faster than the other variables.



I will first use the median lambda value and later find lambda with cross-validation. Using the median lambda I obtained an R squared of 0.58. This implies that about 58% of the variations in the dependent variable are explained by the ridge regression model.

```
## [1] 0.5839801
```

The RMSE for the un-optimized ridge regression was about 0.544, which is worse than the linear regression.

```
## [1] 0.5447794
```

5.2.3 Ridge regression with cross validation

Cross-validation allows you to estimate model performance using data that was not used during training. I will now find the optimal lambda through cross-validation.

Cross-validation is a statistical method for evaluating and comparing learning algorithms that divide data into two segments: one for learning or training a model and the other for validating it. The training and validation sets must cross over in successive rounds in traditional cross-validation so that each data point gets a chance to be validated against. An examples of cross - validations is the k-fold cross validation technique and repeating k-fold cross validation.

The data is first partitioned into k equally (or nearly equally) sized segments or folds in k-fold cross-validation. Following that, k iterations of training and validation are carried out, with each iteration holding out a different fold of the data for validation.

In this application I will use k = 10 folds, that is 10 fold cross-validation to find the optimal lambda values. Through this application, I arrived at a lambda value of 0.001

```
## [1] 0.001
```

This overview shows the first 6 actual and predicted values(s1)

```
##      actual       s1
## 4  2.571282 2.288942
## 5  2.223980 1.768066
## 15 1.817121 1.734452
## 21 1.817121 2.367503
## 27 2.154435 2.723510
## 28 1.912931 2.041169
```

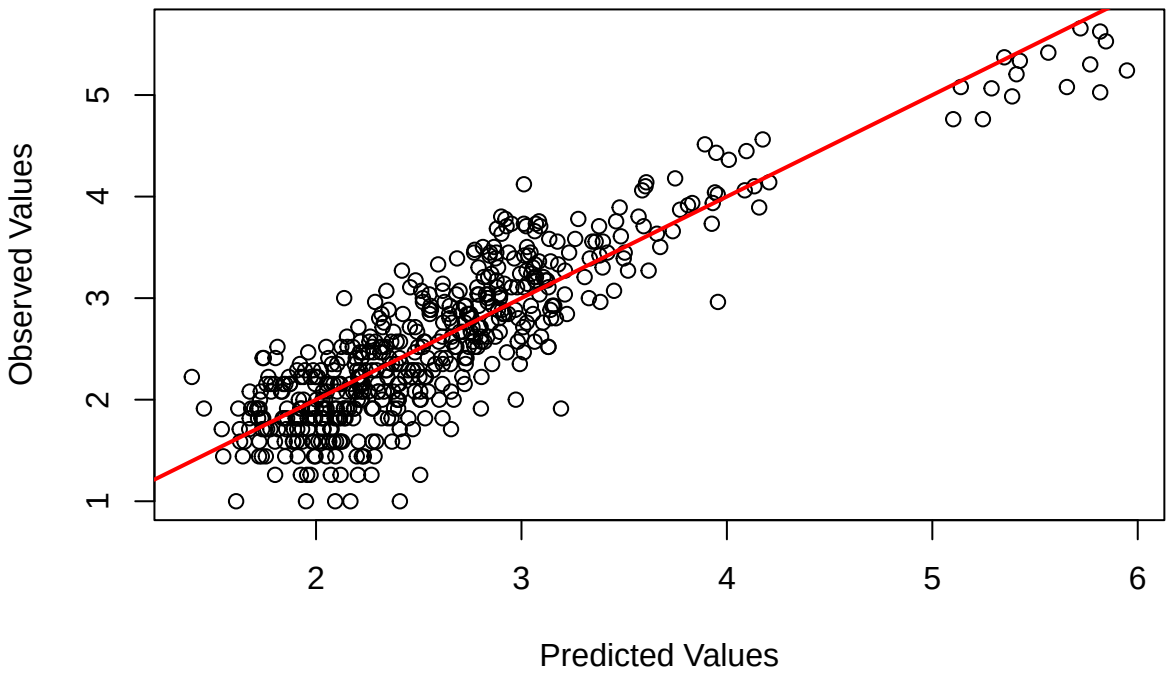
After tunning the model I was able to obtain a higher R - squared value of about 0.78. This means that about 78% of the variations in sexual convictions are explained by the model. This is a significant improvement.

```
## [1] 0.7821065
```

Finally, I obtained an RMSE of around 0.394.

```
## [1] 0.3942627
```

Finally, by plotting the predicted vs the actual I can see a better linear relationship as compared to the linear regression.


Observed Values
1
2
3
4
5
2
3
4
5
6
Predicted Values

5.2.4 Hypothesis testing results

From our tests the ridge regression had a lower RMSE.

```
##               Model      RMSE
## 1  Ridge Regression 0.3942627
## 2 Linear Regression 0.5447794
```

As such **I reject the null hypothesis**

$H0$: **Ridge regression will not result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

and accept the alternative hypothesis

$H1$: **Ridge regression will result in a lower root mean squared error than linear regression in the prediction of sexual conviction.**

5.3 Metropolitan areas and cities have higher average convictions than other counties in the United Kingdom.

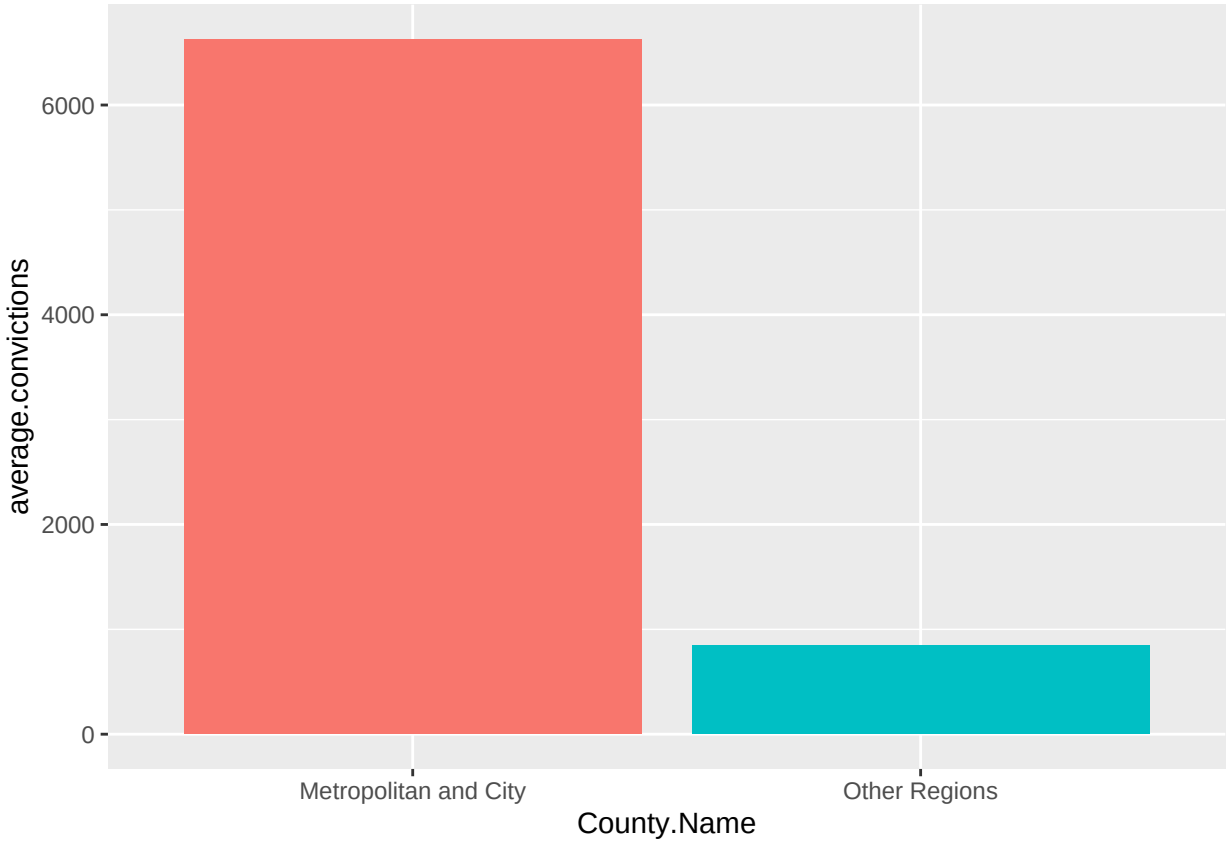
$H0$: **metropolitan areas and cities do not have a higher average conviction than other counties in the United Kingdom.**

$H1$: **metropolitan areas and cities have higher average convictions than other counties in the United Kingdom.**

For this hypothesis, I will find out if the `metropolitan area and city` county has more mean convictions compared to all other counties. I will test this hypothesis with a bar plot, the K-means clustering and the Wilcoxon signed-rank test.

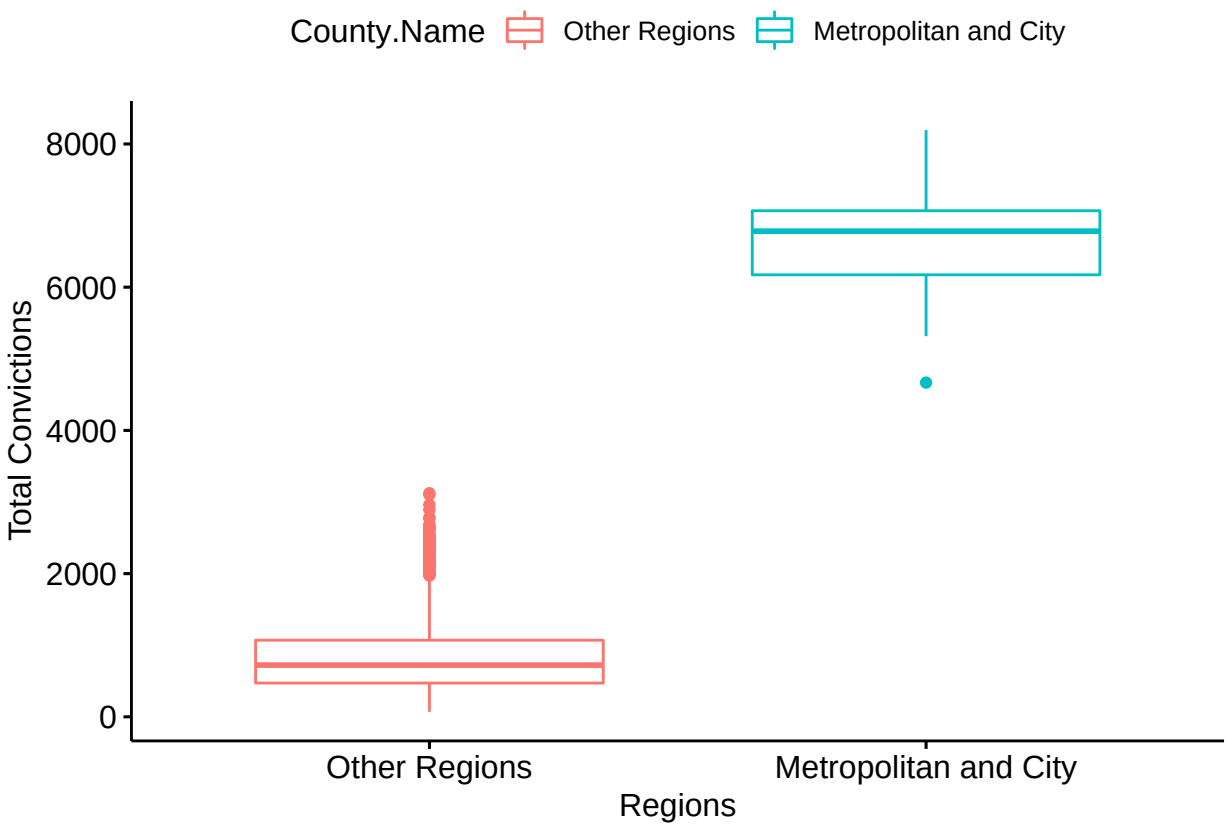
5.3.1 Metropolitan area and city mean conviction vs other counties (Barplot and Boxplot)

To find the mean difference in convictions all other counties were labelled as 'Other Regions'. The average for the total of all convictions was extracted and visualized on the bar plot below. I can immediately see a huge difference. The average conviction for Metropolitan and City stands around 6,600 and that for other regions stands around 840. This difference could be attributed to some factors such as population, however, the combined population of people in other regions far exceed that of the Metropolitan and City[6]. The bar plot fails to provide insight into the distribution of total convictions, as such I will utilize the box plot.



From the box plot, I can see the clear differences between these two regions. The Metropolitan and City have a higher median of about 6600 as compared to that of other regions which stands at around 900. Other summary statistics also point to this disparity. The reason for this disparity may not lie in the population itself but the population density. Metropolitan and cities are characterized by densely populated communities, as well as increased levels of population turnover, this subsequently facilitates high levels of

crime and consequently total convictions[7]. Again I would like to determine if the data itself represents this pattern. This will be done with the K-Means Clustering



5.3.2 K-Means clustering

Clustering occurs when a values in a dataset are grouped based on some similarities and unique characteristic dissimilar to data points in other groups. It is essentially a collection of objects based on their similarity and dissimilarity.

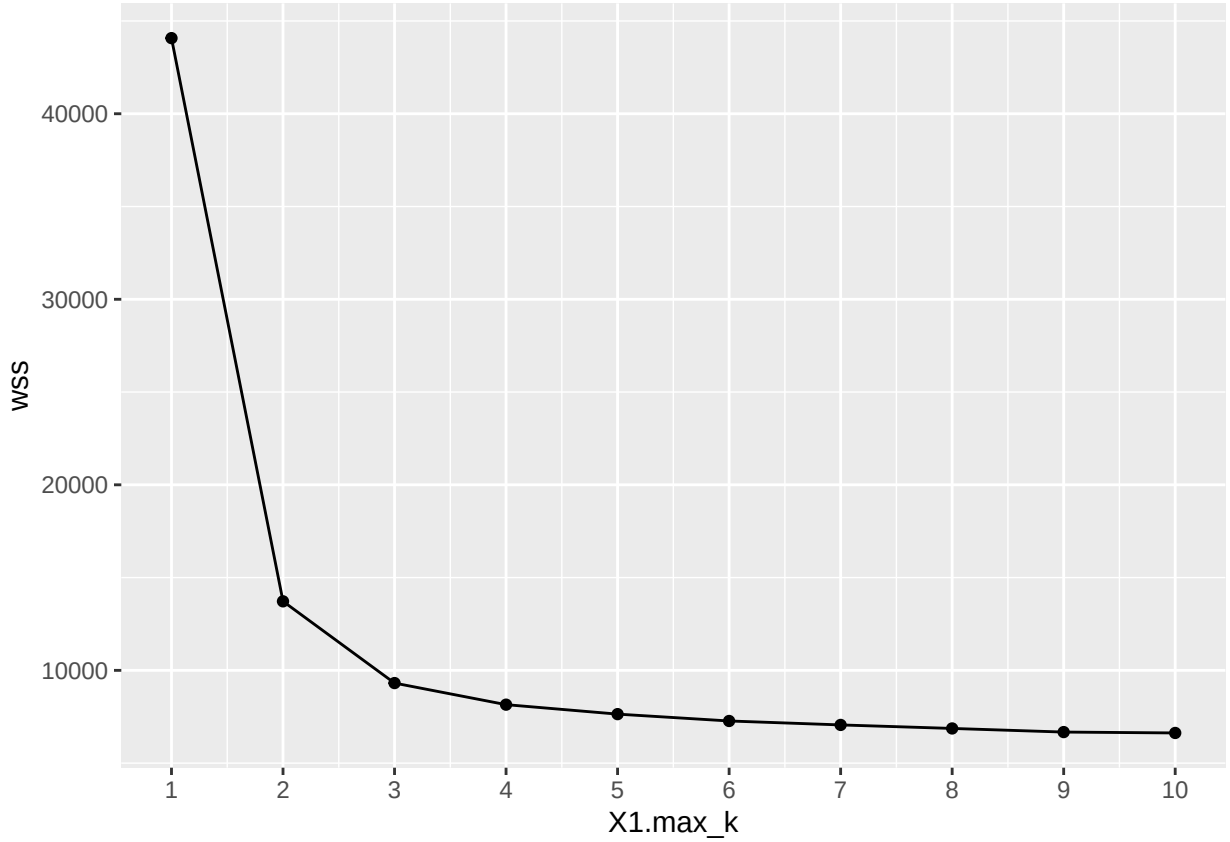
K-means clustering is a simple unsupervised learning approach. It uses a straightforward method of classifying a given data set into a series of clusters, each specified by the letter "k," which is predetermined. The clusters are then positioned as points, and all observations or data points are associated with the closest cluster, computed, and adjusted, before repeating the procedure with the new adjustments until the desired result is achieved. This method of clustering works by;

1. The initial group of clusters is represented by K points in the object data space.
2. The closest k is allocated to each item or data point.
3. The positions of the k clusters are recalculated when all objects have been allocated.
4. Steps 2 and 3 are performed until the cluster positions do not change.

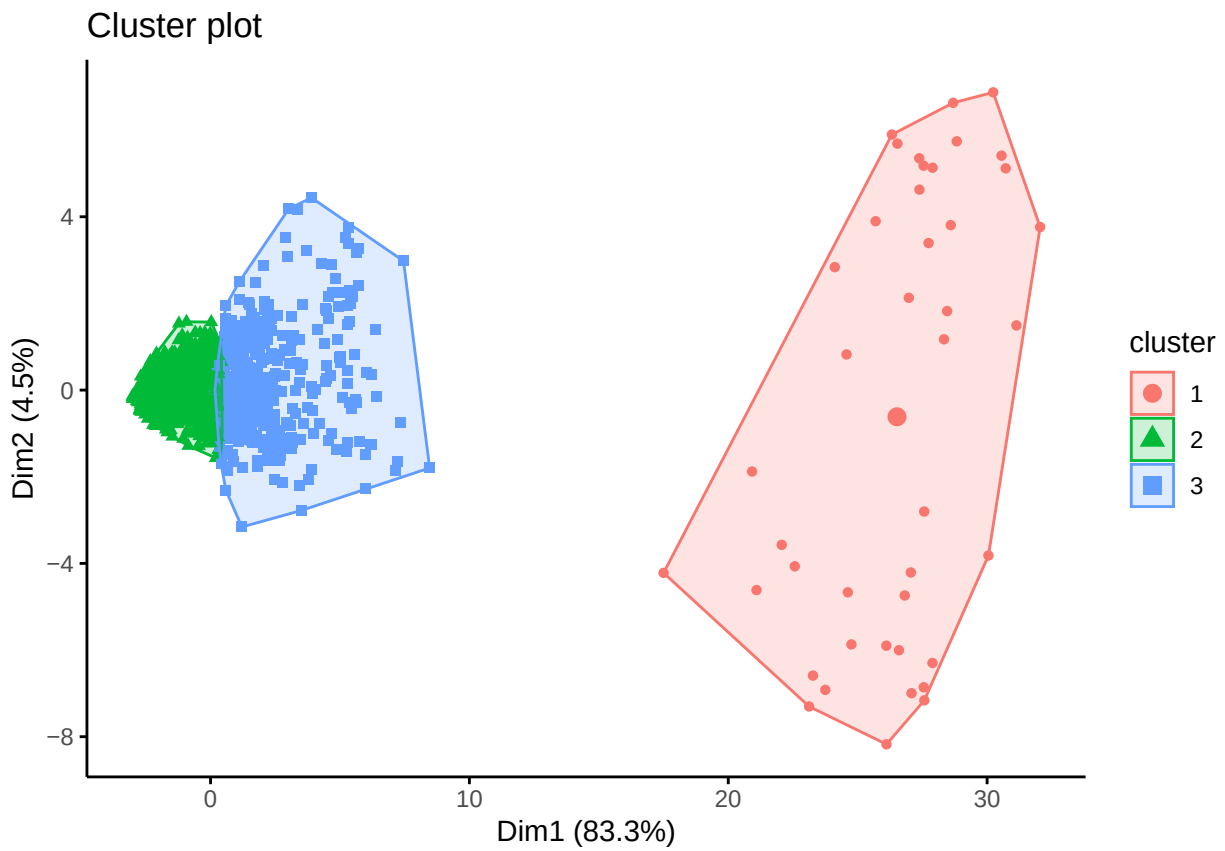
One way to find the number of clusters to use is the elbow method will be utilized. The elbow method uses a range of values for k and selects the best one. When k is 1, the sum of the squares within the cluster will be large. To determine the k value, I will plot a graph between k-values and the within-cluster sum of the square. Our graph will dramatically start experiencing some diminishing returns at a particular k value. The value of k will be assigned to that point.

From the elbow plot, I see that diminishing returns sets in at cluster 3. So k = 3 will be out of the cluster.

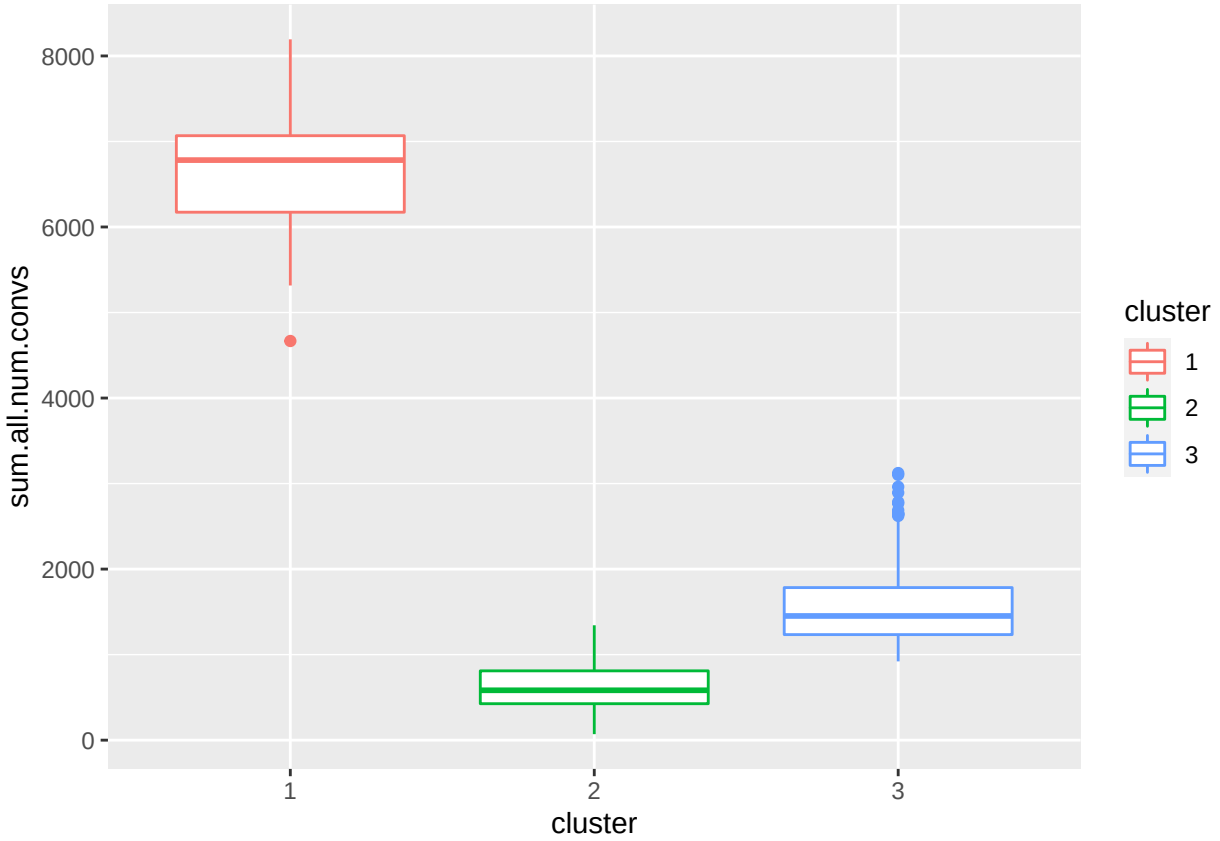
However, before I commence clustering, I will utilize only the numerical convictions and scale the data to enable the optimal calculation of the distance. Data scaling guarantees that the clustering process weights each variable in your data equally. Otherwise, clustering results will be influenced by features with a significantly broader range of values than other features.



Interestingly i see that clusters 2 and 3 are similar. However cluster 1 is completely different.



To gain some additional insight a box plot of the clusters was visualized based on total convictions. Interestingly cluster 1's median approximates that of the Metropolitan and City as seen in the previous figure. The other clusters again had relatively low but similar medians.



To validate my assertion, the means total conviction for both cluster 1 and Metropolitan and City were calculated. Both produced the same mean of 6,625.81 indicating that cluster 1 signifies the distinctively high conviction levels of Metropolitan and City.

```
## # A tibble: 1 x 2
##   cluster average
##   <fct>     <dbl>
## 1 1         6626.
```

```
## # A tibble: 1 x 2
##   County.Name           average
##   <chr>                   <dbl>
## 1 Metropolitan and City   6626.
```

Since the clusters represent the level of conviction, the variable `conviction. level` was introduced with a `High conviction level` for cluster 1, a `Medium conviction level` for cluster 2 and a `Low conviction level` for cluster 3.

5.3.3 Wilcoxon signed-rank test and hypothesis testing results

Finally, to statistically test for the difference between the Metropolitan and City Region again in other counties I will use the Wilcoxon signed-rank test. This is a statistical test that compares two paired groups. The Wilcoxon signed-rank test is nonparametric implying that the distribution of the variables does not affect the outcome of the test.

this test uses the following hypotheses:

$H0$: metropolitan areas and cities do not have a higher average conviction than other counties in the United Kingdom.

$H1$: metropolitan areas and cities have higher average conviction than other counties in the United Kingdom.

$\alpha = 0.05$

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sum.all.num.convs by County.Name
## W = 72324, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

Since the p - value 2.2e-16 is less than $\alpha = 0.05$:

As such **I reject the null hypothesis**

$H0$: **metropolitan areas and cities do not have a higher average conviction than other counties in the United Kingdom.**

and accept the alternative hypothesis

$H1$: **metropolitan areas and cities have higher average convictions than other counties in the United Kingdom.**

5.4 Random forest algorithm will result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.

$H0$: **Random forest algorithm will not result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.**

$H1$: **Random forest algorithm will result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.**

For this hypothesis, I will evaluate the performance of the random forest algorithm by determining its accuracy in the prediction of conviction level. I will compare this to the gradient boosting machine and decision tree algorithm to test the hypothesis. The metric for evaluation, accuracy, is used for determining which model is the best at recognizing relationships and patterns between variables in a data set based on the input, or training, data. The models will first be used in their base state and results will be evaluated, afterwards, hyperparameter tuning and cross-validation will be performed to make the final evaluation and test model performance.

The first model used was the Gradient Boosting Machine algorithm. Gradient Boosting is an iterative functional gradient approach that minimizes a loss function by choosing a function that points towards the negative gradient iteratively; a weak hypothesis. This model is made up of;

1. The loss function's job is to estimate how good the model is at making predictions using the data it's given.
2. Weak Learner - A weak learner attempts to classify our data but fails miserably, maybe no better than guessing at random. It has a high mistake rate, in other words.
3. This is the iterative and sequential strategy of gradually increasing the trees (weak learners). I need to get closer to our final model with each cycle. In other words, the value of our loss function should decrease with each repetition[8].

Next, I used the Decision tree algorithm. Decision trees are created using an algorithm that finds different ways to segment a data set based on certain conditions. It is one of the most popular and practical supervised learning algorithms. Decision Trees are a non-parametric supervised learning method that can be used for classification and regression. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable[9].

Finally, I used the Random Forest algorithm. Random Forest is a powerful machine learning technique that may be used for regression and classification. It's an ensemble method, which means a random forest model is made up of many little decision trees called estimators, each of which makes its predictions. The random forest model combines the estimators' predictions to get a more precise prediction. Simply put, it lowers instability by averaging many decision trees, or a forest of trees built with randomness[10].

5.4.1 Data partitioning

To commence the data was partitioned into a train and test set with a 75% to 25% split for the train and test set. The only number of convictions and unsuccessful were selected for this model. The year, date, month and county variables were all removed as are of no use to our model. The dependent variable will be the conviction level.

5.4.2 Results with no hyperparameter tuning and cross-validation

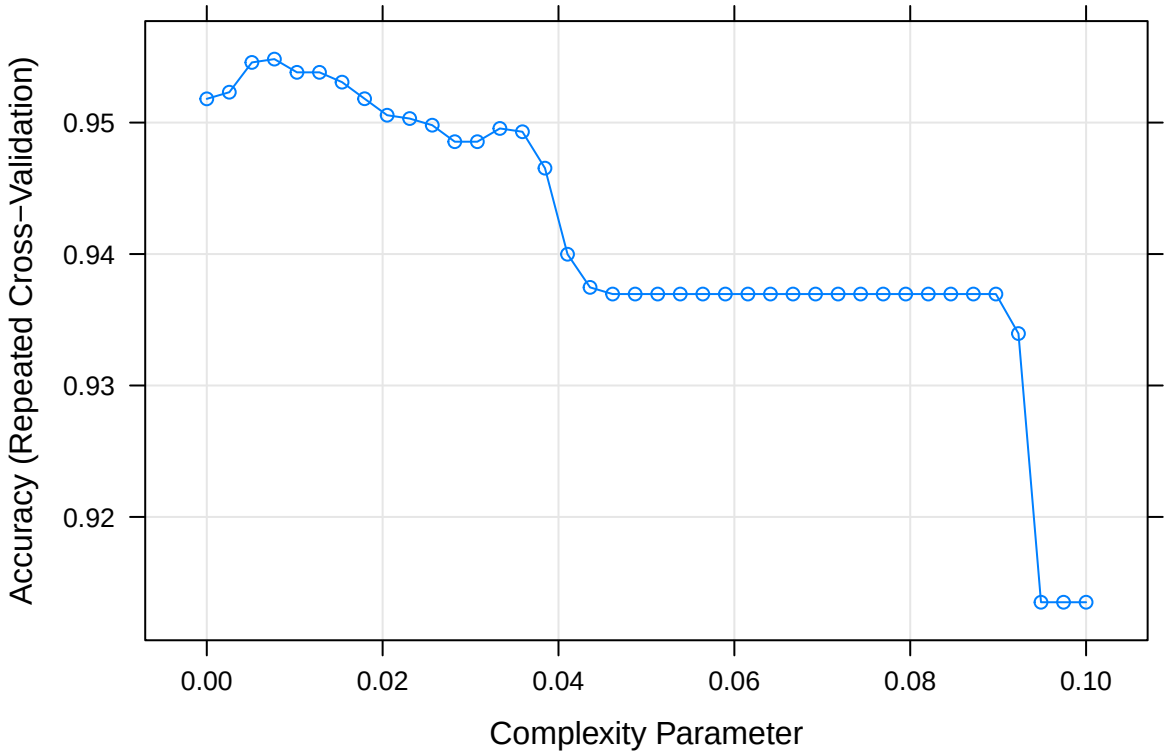
First I run the Gradient Boosting Machine(GBM), Decision trees and Random Forest algorithms without any hyperparameter tuning on the train set. The results indicate that the was the best Random Forest algorithms performer with an accuracy of around 96.8%.

```
## Using 100 trees...
```

Model	Accuracy
Gradient boosting machine	0.9660633
Random Forest	0.9683258
Decision Tree	0.9479638

5.4.3 Decision tree with hyperparameter tunning of complexity parameter

To improve model performance I will tune the complexity parameter which is used to control the size of the decision tree and facilitate the selection of the optimal tree size. Values from 0 to 40 with an increment of 0.01 were used to tune this parameter. Additionally, a 10 fold cross-validation with three repetitions was applied. 10% of the data was used for validation and 90% for training.

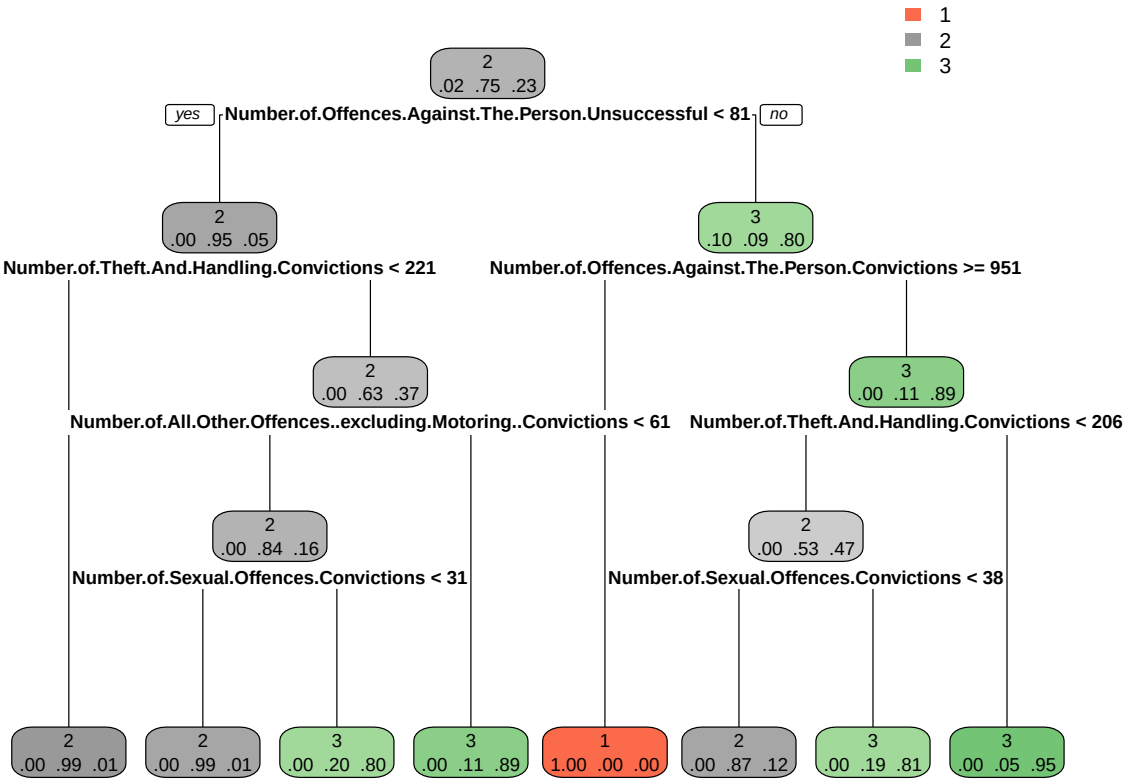


From the plot above I realize the highest accuracy was obtained with a complexity parameter of around 0.0077, as such this was the best tune.

```
##           cp
## 4 0.007692308
```

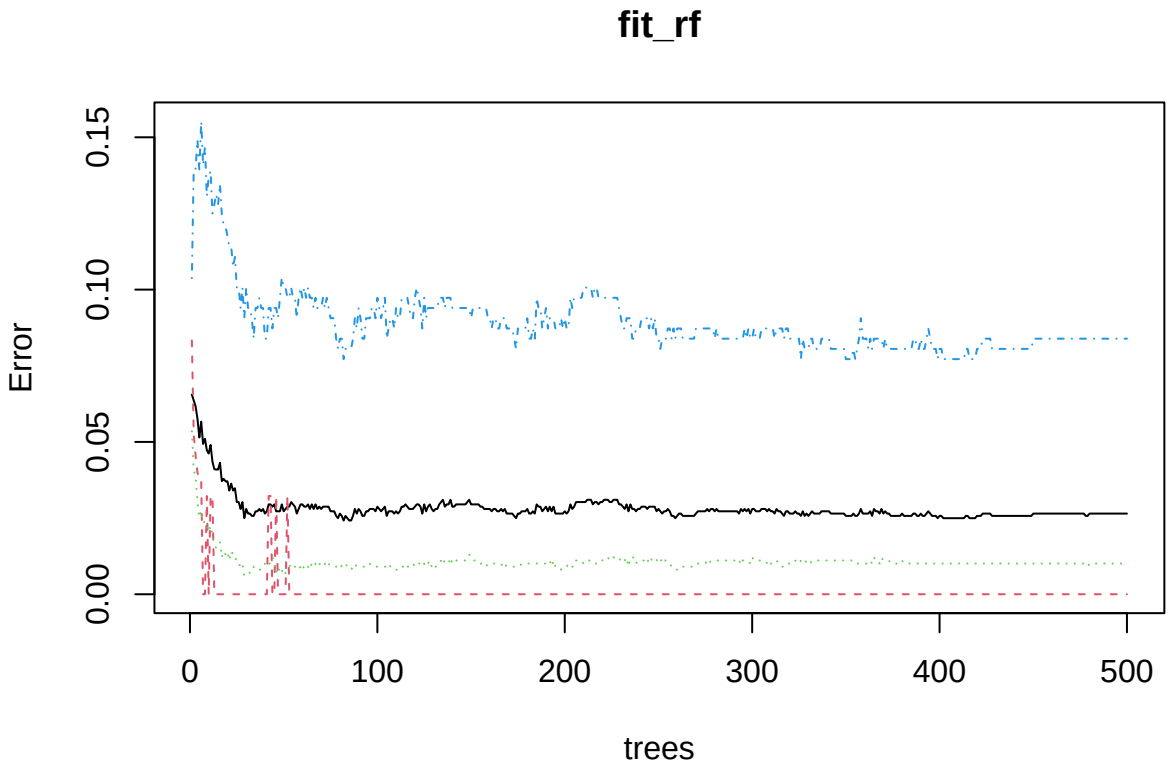
Below I can see the decision tree plot, this diagram shows the various outcomes of a set of related options. In determining the conviction level, the most important variables at the top are `Number of offence against the person unsuccessful`, `Number of theft and handling convictions` and `Number of offence against the person convictions`. If `Number of offences against the person unsuccessful` is less

than 81 then class 2 will be selected and moved to the next variable `Number of theft and handling convictions` for the next decision making. If `Number of offence against the person unsuccessful` is more than 81, class 3 will be selected and moved to the next variable `Number of offence against the person convictions` for the next decision making



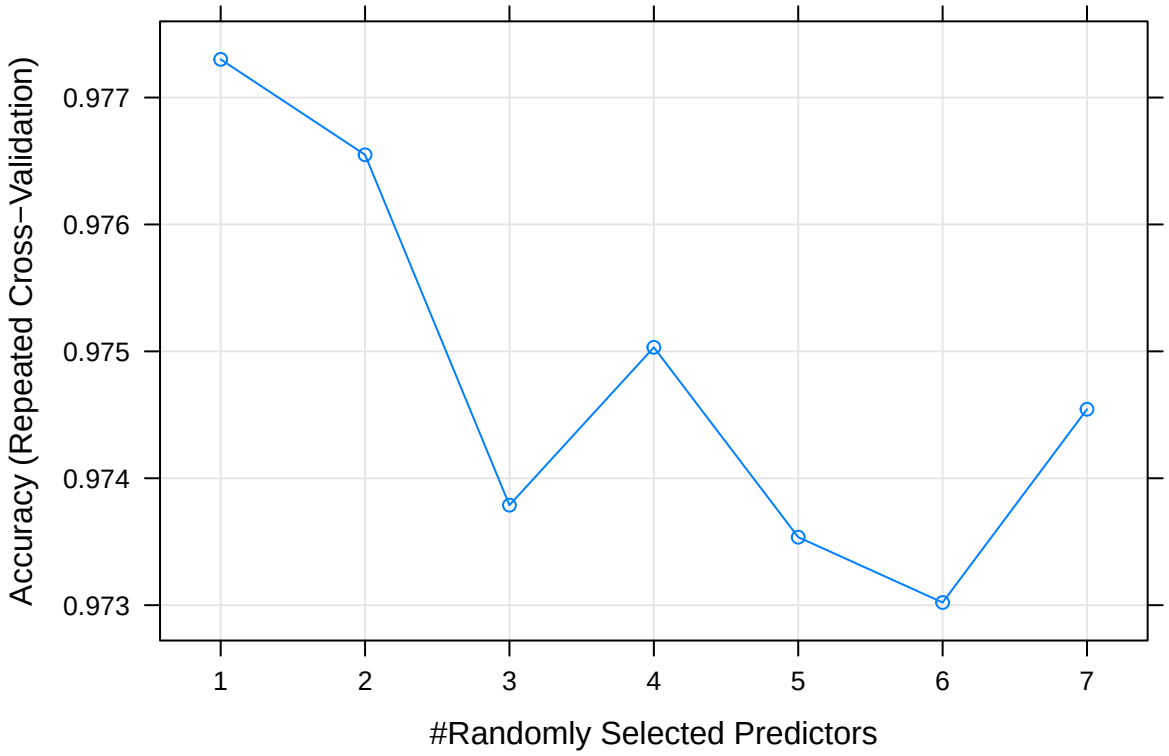
5.4.4 Random forest with hyperparameter tuning and cross-validation

To tweak the random forest, I'll first plot the base random forest model to determine the ideal number of trees to employ. The error rates for conviction level 1, conviction level 2, total error rate, and conviction level 3 are represented by the blue, black, green, and red lines. On the graphic below, the error rare is on the y axis, with the number of trees on the x-axis. When looking at the total error rate (green), it appears that the mistake rate is at its lowest at 70 trees. As a result, I'll set 'ntrees = 70' in our model.



10-fold cross-validation with a 10%:90% ratio of validation and training data was used. The cross-validation was also repeated three times. The tuning parameter value `mtry`(Number of variables randomly sampled as candidates at each split) was assessed with a range of values from 1 to 7.

Plotting the tunned model shows a y-axis which illustrates the accuracy with an x-axis showing the randomly selected predictors. The best `mtry` value was 1.




```
##   mtry
## 1    1
```

5.4.5 GBM Tree with Preprocessing and repeated cross-validation

To improve performance with GBM I will preprocess through scaling and centre, The average of a variable is subtracted from the data when it is centring. When data is scaled, the standard deviation of a variable is subtracted from the total.10 fold cross-validation was undertaken with three repetitions.

In an interesting turn of events, the GBM algorithm surpassed the Random forest algorithm after a series of hyperparameter tuning and cross-validation. The GBM, Random forest and Decision tree algorithms had accuracies of about 98%, 96% and 95% respectively

5.4.6 Model evaluation and hypothesis testing results

Model	Accuracy
Gradient boosting machine	0.9796380
Random Forest	0.9615385
Decision Tree	0.9479638

The confusion matrix of the tuned GMB model produced an accuracy of about 98%. The ability of the algorithm/model to predict a true negative of each available category is known as specificity. It's also known as the actual negative rate in literature. The statistic used to evaluate a model's ability to predict the true

positives of each accessible category is known as sensitivity in Machine Learning. For class 1 both algorithms had a perfect sensitivity and specificity. class 2 and 3 had sensitivity and specificity values of about 99% and 95% and then 94% and 99%.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1  11   0   0
##          2   0 328   6
##          3   0   3  94
##
## Overall Statistics
##
##                Accuracy : 0.9796
##                  95% CI : (0.9617, 0.9906)
##     No Information Rate : 0.7489
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.947
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity           1.00000   0.9909   0.9400
## Specificity           1.00000   0.9459   0.9912
## Pos Pred Value        1.00000   0.9820   0.9691
## Neg Pred Value        1.00000   0.9722   0.9826
## Prevalence            0.02489   0.7489   0.2262
## Detection Rate        0.02489   0.7421   0.2127
## Detection Prevalence  0.02489   0.7557   0.2195
## Balanced Accuracy     1.00000   0.9684   0.9656
```

Since the Gradient boost machine algorithm had a higher accuracy than the random forest($0.9796380 > 0.9615385$).

As such **I fail to reject the null hypothesis**

$H0$: **Random forest algorithm will not result in better accuracy than gradient boosting machine and decision trees algorithm in the prediction of conviction level.**

6. Conclusion

6.1 Summary

Through data integration and cleaning a consolidated data set was created facilitating the creation of four hypotheses, I had sufficient evidence to prove that an increase in the number of homicide and robbery convictions will result in an increase in the number of sexual offence convictions. The power of the ridge regression was demonstrated by producing a lower RMSE score than linear regression while handling multicollinearity and high dimensionality. Sufficient evidence was also found supporting the claim that the Metropolitan area and city had a higher conviction rate than any other county. Finally, the random forest algorithm was outperformed by the Gradient Boosting Machine algorithm in the prediction of conviction levels after hyperparameter tuning and cross-validation, disproving my initial hypothesis. Additionally, through data visualization, I showed that the total number of convictions overall had been on the decline and that the trend of sexual offence convictions was starting to take a more favourable downward turn.

6.2 Limitation

A major limitation was the absence of some months from the four-year dataset. These months could have facilitated further insight into the true nature of convictions in the UK. Additionally, some advanced classification models like the `XGboost` could not be implemented due to their long-run time from computational requirements.

6.3 Future work

As discussed earlier, research on the impact of the `#Metoo Movement` in workplaces showed a decline in sexual harassment. I would like to conduct similar research in the UK to ascertain if this movement played a part in reducing sexual offence convictions in the UK. Additionally, I would like to find out the reasons why Metropolitan and city has a disproportionately high conviction rate compared to other regions in the UK.

7 References

[1] Liv Moloney, "Is rape the perfect crime." news.sky.com. https://news.sky.com/story/99-of-rapes-reported-to-police-in-england-and-wales-do-not-result-in-legal-proceedings-why-12104130#:~:text=In%20order%20to%20fir

[2] ELENA NICOLAOU, COURTNEY E. SMITH, "A #MeToo Timeline To Show How Far I've Come — & How Far I Need To Go" www.refinery29.com https://www.refinery29.com/en-gb/2018/10/213189/me-too-movement-history-timeline-year-weinstein

[3] Johnson, S.K., Keplinger, K., Kirk, J.F. and Barnes, L., 2019. Has sexual harassment at work decreased since# metoo. Harvard Business Review.

[4] DRACA, M., 2016. It's Prices, Stupid: Explaining falling crime in the UK.

[5] McDonald, G.C., 2009. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), pp.93-100.

[6] 'Population of England in 2020, by ceremonial county' https://www.statista.com/statistics/971694/county-population-england/

[7] Craglia, M., Haining, R. and Signoretta, P., 2001. Modelling high-intensity crime areas in English cities. Urban Studies, 38(11), pp.1921-1941.

[8] Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, p.21.

[9] Quinlan, J.R., 1996. Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), pp.71-72.

[10] Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25(2), pp.197-227.

8 Abbreviation

GBM: Gradient Boosting Machine CPS: Crown Prosecution Service RMSE: Root Mean Squared Error