Assignment 2

C. Twaddle, L. De Jong and D Grassman

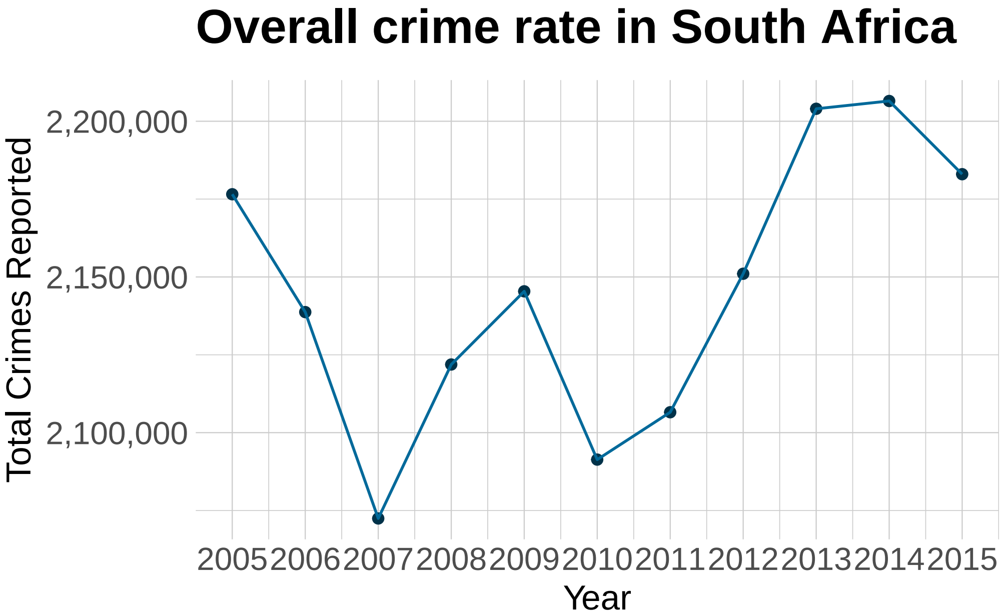
1 Data Wrangling

- External population data sourced from: https://data.openup.org.za/dataset/police-district-population-mwzd-ptz7

2 Background

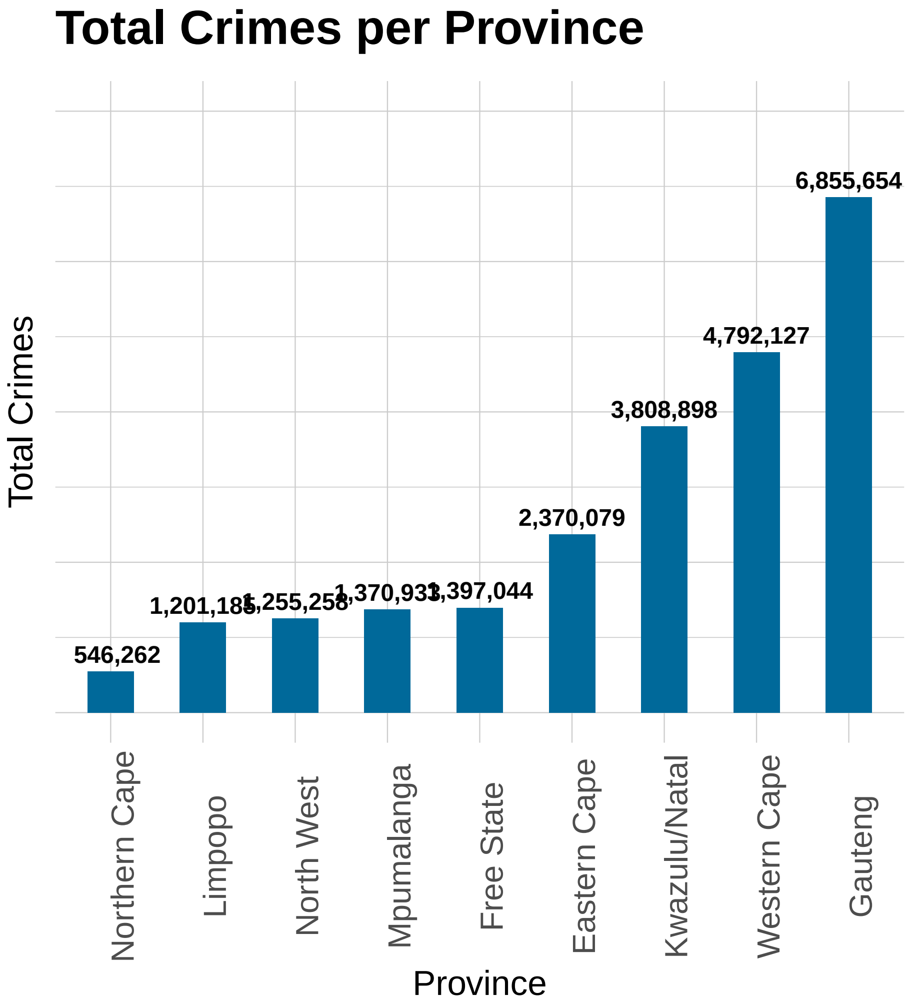
This section aims to provide insight into the state of crime in South African for a 10 year period (2005 - 2015). It will include various visualisations to illustrate the statistics of interest.

2.0.1 Total crimes per province



The graph above shows the overall crime rate for the whole country of South Africa over the 10 years. We can see that while the line does dip once or twice, the crime rate has increased over the period.

2.1 Total crimes per province



The above graph represents the total cumulative crime count per province. Gauteng has the highest crime rate, which will influence our choice of sampling the dataset in later models.

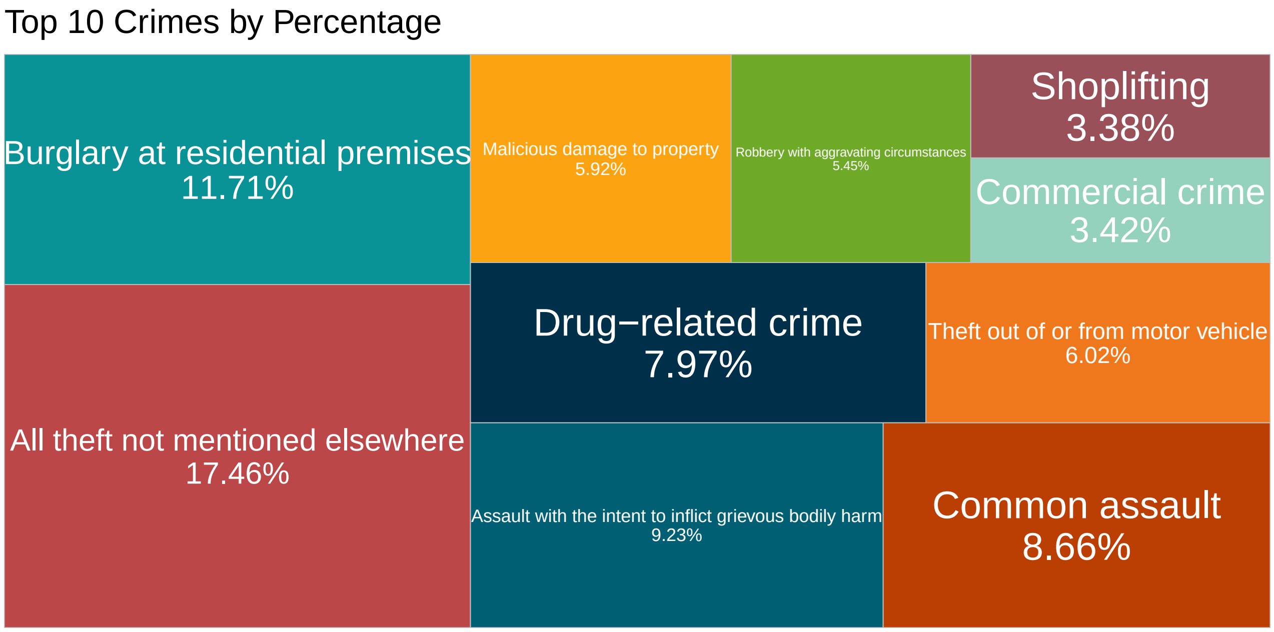
2.1.1 Areas that experience higher crime per province

```
# A tibble: 90 x 3
# Groups:   Province [9]
   Province     Station     total
   <chr>        <chr>       <int>
 1 Eastern Cape East London 92360
 2 Eastern Cape Mount Road  85777
 3 Eastern Cape Mthatha     76451
 4 Eastern Cape Humewood    71392
 5 Eastern Cape Cambridge   67733
 6 Eastern Cape Bethelsdorp 62241
 7 Eastern Cape Kwazakele   59962
 8 Eastern Cape Grahamstown 58862
 9 Eastern Cape Gelvandale  57233
10 Eastern Cape Walmer      53781
# i 80 more rows
```

This table shows the top 10 stations with the highest cumulative crime count per province. ## #Most common crimes in South Africa

```
# A tibble: 27 x 2
   Category                                             total
   <chr>                                                <int>
 1 All theft not mentioned elsewhere                  4120351
 2 Burglary at residential premises                   2763950
 3 Assault with the intent to inflict grievous bodily harm 2179207
 4 Common assault                                     2043267
 5 Drug-related crime                                 1879871
 6 Theft out of or from motor vehicle                 1420789
 7 Malicious damage to property                       1397845
 8 Robbery with aggravating circumstances             1284991
 9 Commercial crime                                    807206
10 Shoplifting                                         798079
# i 17 more rows
```

This table ranks the crime types from the highest cumulative count to the lowest.



The above tree map displays the most often occurring crimes by percentage.

3 Regression

Our goal with our regression model is to investigate whether there is a strong correlation between the number of males in the population and the amount of sexual offences (so) that are reported. We aim to test if the model could predict the number of so crimes that will be reported. Please note that the population estimates were sourced from the 2011 South African Census, and the regression model will therefore be limited to the year 2011.

3.1 EDA

Going forward our hypotheses are as follows: H0: The number of males in a population has no influence on the number of sexual offences reported. HA: The number of males in a population has an influence on the number of sexual offences reported.

Our first step into exploring the data is to test for correlation.

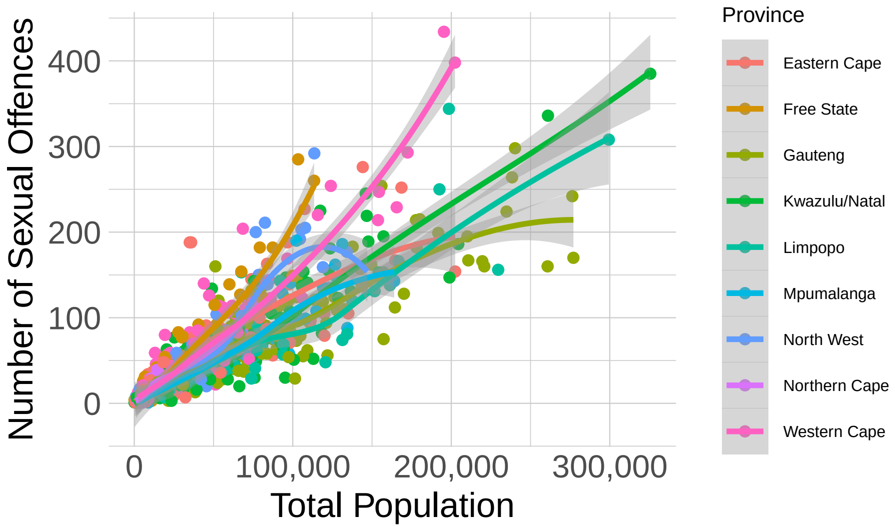
```
[1] 0.7865257
```

The above result shows that the two variables of interest, number of males in population and number of reported sexual offences, is highly positively correlated. The following function gives more insight into this correlation.

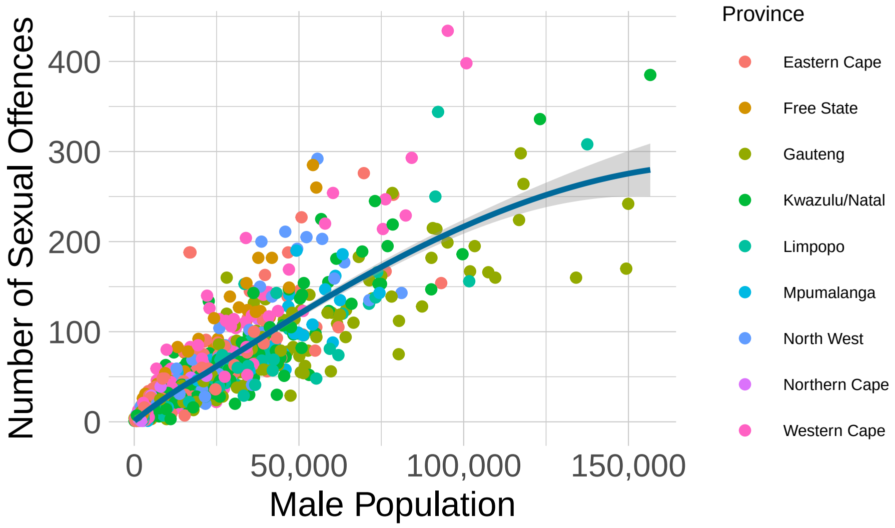
```

    Pearson's product-moment correlation

data:  so_data$sexual_offences and so_data$estimated_male
t = 130.62, df = 10518, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7791266 0.7937057
sample estimates:
      cor 
0.7865257 
```

In the following output shows a t-value of 130,62. This value indicates how far the correlation coefficients are from zero in standard errors. In our case, the t-value shows that the correlation coefficients are quite far from zero and a larger t-value provides strong evidence to support the alternative hypothesis. The p-value is extremely small, much less than 0,05, which once again supports the alternative hypothesis.

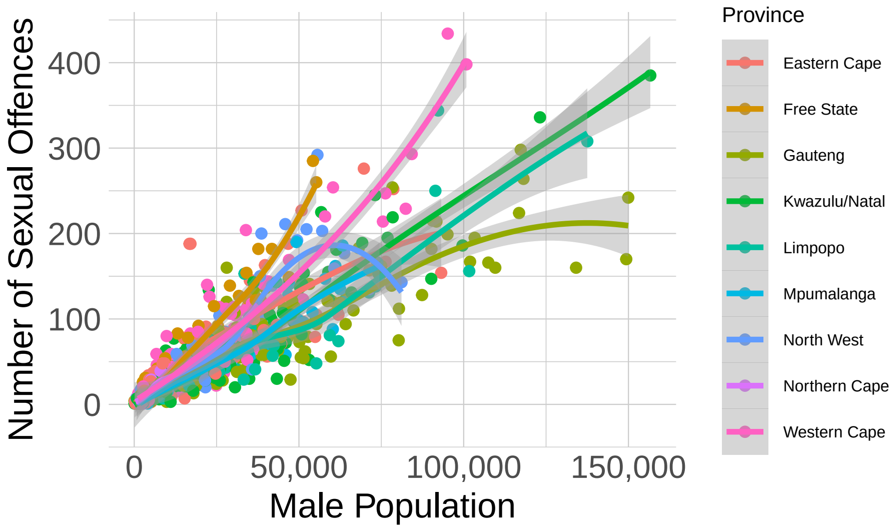
Below are some graphs to visualise these findings.



The above graph shows the correlation between the number of sexual offences reported against the total population, split by province. We can see a common upward trend in most provinces.



This graph indicates the correlation between sexual offences and the male population for the country as a whole. This graph supports the statistics that were explored above.



This graph expands on both previous graphs by exploring the relationships among the all the provinces to see if this is a consistent trend. We can see that most provinces follow a general upward trajectory of sexual offences as male population increases, except for Mpumalanga. This variance will be investigated further.

3.2 Models

The first model created is a simple, two variable linear regression.

```
Call:
lm(formula = sexual_offences ~ estimated_male, data = so_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-208.46  -12.93   -4.54    6.56 1301.86 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.106e+00  5.593e-01   10.92   <2e-16 ***
estimated_male 2.246e-03  1.719e-05  130.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.31 on 10518 degrees of freedom
Multiple R-squared:  0.6186,	Adjusted R-squared:  0.6186 
F-statistic: 1.706e+04 on 1 and 10518 DF,  p-value: < 2.2e-16
```

This summary shows us that the models sees these variables as also having a strong positive correlation as the p-value is less than 0,05. The residual standard error (RSE) suggest that on average the predicted values differ from the observed values at a rate of 41,36 units. Multiple R-squared shows that 61,86% of the predicted variables are explained by the predictors of the model. The F-statistics is the result of an f-test, and shows the overall significance of the model. It tries to explain the variability of the model. Our f-statistic is quite large and therefore the model can handle big variability.

```
                 Df   Sum Sq  Mean Sq F value Pr(>F)    
estimated_male     1 29110500 29110500   17061 <2e-16 ***
Residuals      10518 17946453     1706                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ANOVA Table (type II tests)

          Effect DFn   DFd        F p p<.05   ges
1 estimated_male   1 10518 17060.99 0     * 0.619
```

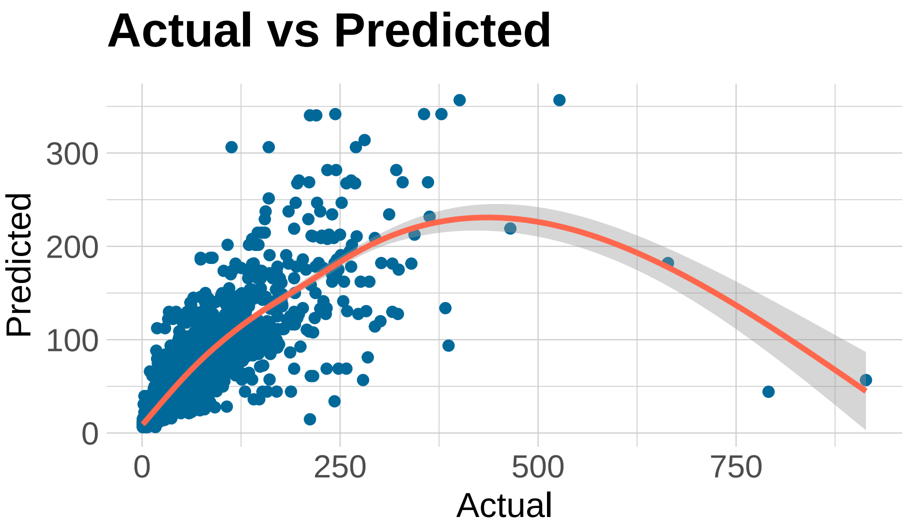
The above model is a simplified analysis of variance, which also uses a linear method, and confirms the linear regressions findings.

```
                          Df   Sum Sq  Mean Sq  F value Pr(>F)    
estimated_male             1 29110500 29110500 19336.38 <2e-16 ***
Province                   8   812103   101513    67.43 <2e-16 ***
estimated_male:Province    8  1323816   165477   109.92 <2e-16 ***
Residuals              10502 15810535     1505                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above model has been expanded to investigate whether the province has any significance in our investigation. The model indicates that province as a factor may be of importance.

3.2.1 Machine learning model

The following is a linear regression machine learning model that was trained on a subset of data and then tested on the rest.



The above graph plots the predicted outcomes of the model versus the actual values from the training data. We can see from this graph that the model is relatively accurate the smaller the variables, but gets less accurate as the variables increase. This could be explained by the fact that most of the data lies closer to the origin so training for larger values is difficult.

```
Call:
lm(formula = sexual_offences ~ estimated_male, data = training)

Residuals:
    Min      1Q  Median      3Q     Max 
-207.36  -12.86   -4.51    6.55 1302.04 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.139379   0.616837   9.953   <2e-16 ***
estimated_male 0.002238   0.000019 117.772   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.73 on 8416 degrees of freedom
Multiple R-squared:  0.6224,	Adjusted R-squared:  0.6223 
F-statistic: 1.387e+04 on 1 and 8416 DF,  p-value: < 2.2e-16
```

The summary of the model and the graph of the results show conflicting results. The table shows that the model is fairly accurate and the variables are strongly correlated, however the graph shows that the outcomes of the model get less accurate. This difference in evaluation of the model could be explained by coincidental correlation. As population in general increases it is expected that crime in general will also increase. Since an increase in males in a population still increases the population overall, the result of increased sexual offences is expected since overall crime will increase. However this exploration was still insightful into whether there is an effect at all.

4 Clustering

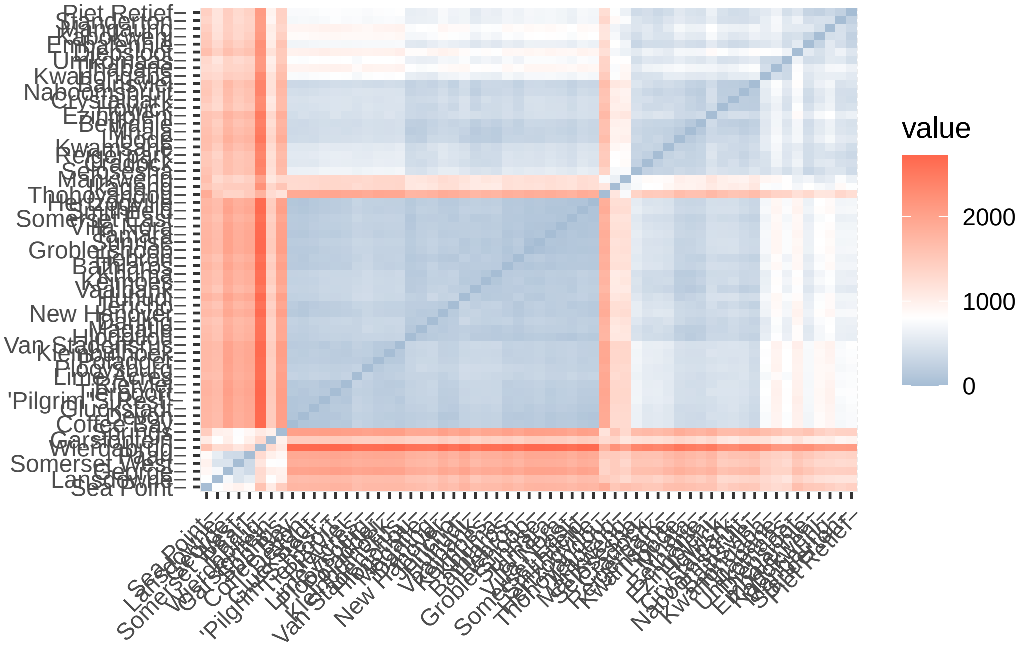
Since clustering is an unsupervised machine learning technique, we do not have any expectations for the outcome of our exploration.

4.1 EDA

We gave the model a subset of data that included station names, categories of crime, and the number of reports per crime time. Our aim was for the model to cluster stations based on either crime type or crime frequency.

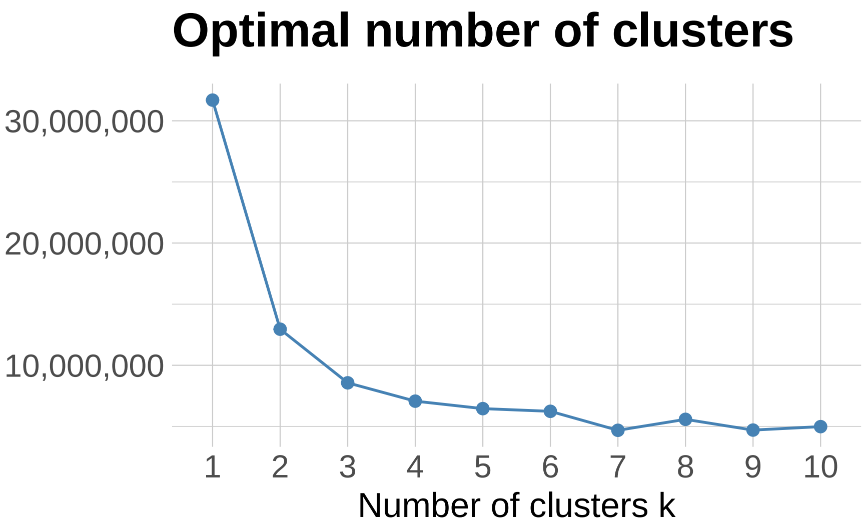
Our first step into exploring the data is to clean and cut the data. We chose to select a random sample of stations from each province to limit the messiness of out EDA and to test if the model could create cohesive clusters. We once again reduced the data set to 2011. We did this to avoid problems in the modelling and chose 2011 to be consistent with our regression model.

Our next step is to create and visualise graphs to help build the clustering model.

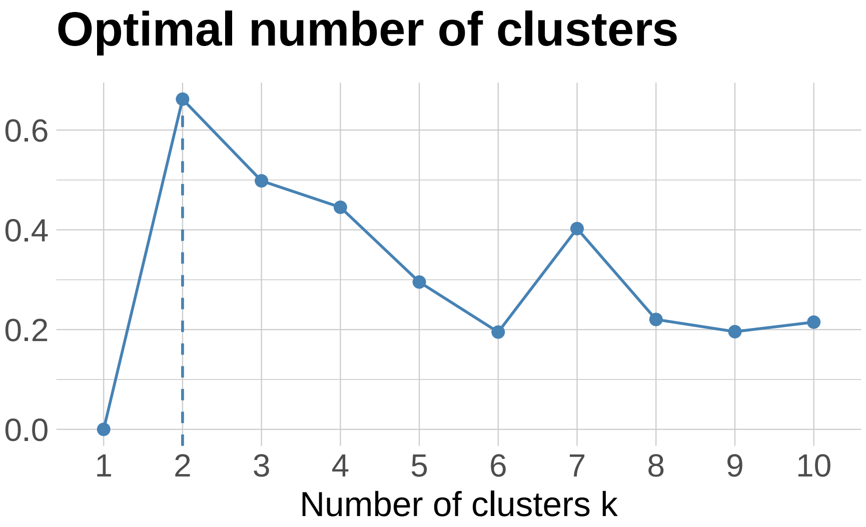


The above matrix plots the different station against each other. The darker colour represents stations with higher correlations. This could indicate that these stations have similar crime patterns.

We ran multiple tests to determine the optimal number of clusters.



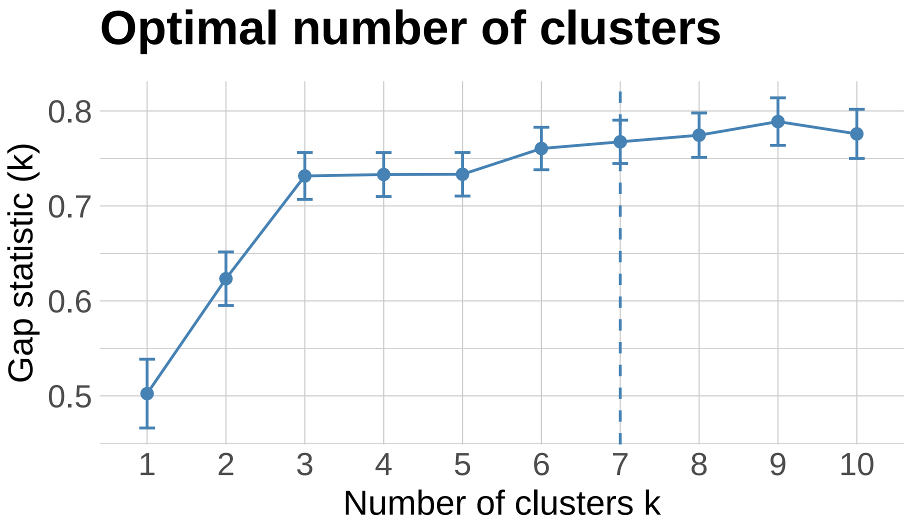
The above graph uses the data given and the ‘elbow’ method to guess the optimal number of clusters. This graph seems to indicate that the optimal clusters lies at either 2 or 3.



The above graph uses the data given and the silhouette method to guess the optimal number of clusters. This graph indicates that the optimal clusters is 2.

```
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = crime_data_clust, FUNcluster = kmeans, K.max = 10, B = 50, nstart = 25)
B=50 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstmax'): 9
```

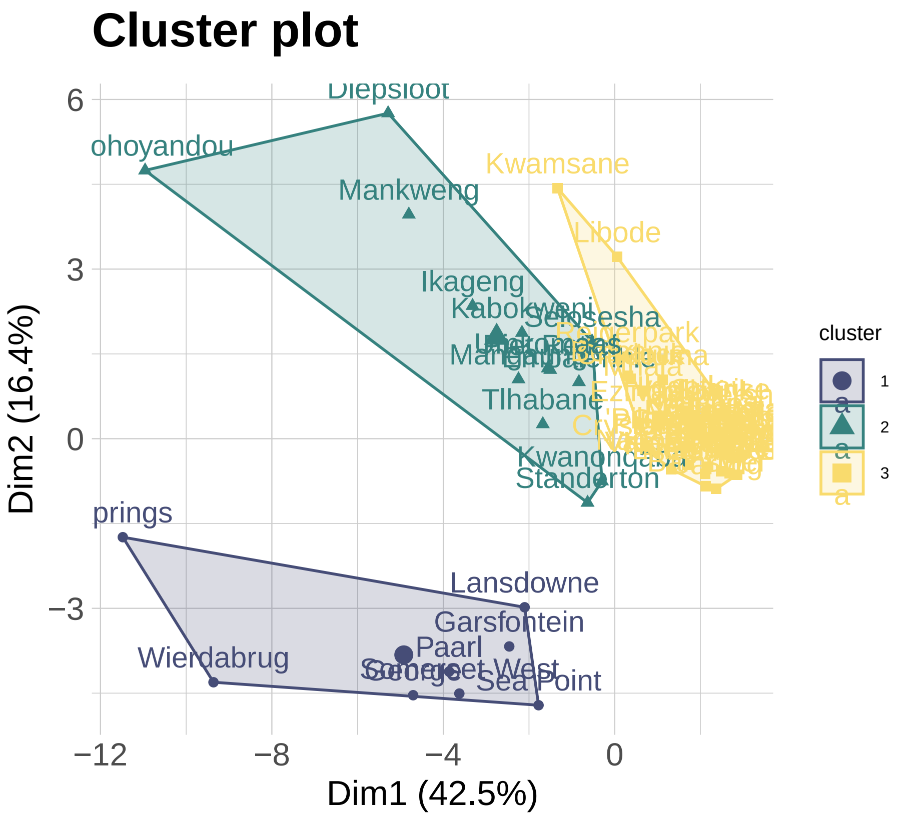
```
          logW   E.logW       gap     SE.sim
 [1,] 9.410123 9.912490 0.5023669 0.03619965
 [2,] 9.018380 9.641755 0.6233752 0.02820841
 [3,] 8.784668 9.516230 0.7315620 0.02469070
 [4,] 8.685019 9.418095 0.7330760 0.02314152
 [5,] 8.608601 9.341930 0.7333287 0.02293710
 [6,] 8.519590 9.280061 0.7604704 0.02234593
 [7,] 8.457481 9.224988 0.7675069 0.02282229
 [8,] 8.400713 9.175191 0.7744779 0.02338886
 [9,] 8.339008 9.127830 0.7888219 0.02504521
[10,] 8.308014 9.083893 0.7758788 0.02590177
```



The above table and graph uses the gap statistic method to guess the optimal number of clusters. These outputs indicate that the optimal clusters is 7.

4.2 Model

The following code creates and visualises the clustering model using k-means methods. We decided to set 3 clusters based on the on the previous cluster graphs.



The graph above shows the outcome of the k-means clustering model. We can see three clearly defined clusters. Each cluster groups a number of stations together.

The following table aims to provide insight into the clusters created from the sample data and why certain stations may have been grouped together. It shows the average of each crime time for each cluster.

```
# A tibble: 3 x 27
  Cluster Burglary at residential prem~1 `Common assault` All theft not mentio~2
    <int>                          <dbl>            <dbl>                  <dbl>
1       1                          788              328.                   1369.
2       2                          512.             347.                    418.
3       3                           95.2             78.6                   113.
# i abbreviated names: 1: `Burglary at residential premises`,
#   2: `All theft not mentioned elsewhere`
# i 23 more variables: `Theft out of or from motor vehicle` <dbl>,
#   `Assault with the intent to inflict grievous bodily harm` <dbl>,
#   `Theft of motor vehicle and motorcycle` <dbl>,
#   `Malicious damage to property` <dbl>, `Common robbery` <dbl>,
#   `Drug-related crime` <dbl>, ...
```

Cluster 1: This cluster has high averages of the following crimes, All theft not mentioned elsewhere, Burglary at residential premises and Theft out of or from motor vehicle. This seems to shows stations in this cluster experience high levels of different types of theft and burglary.

Cluster 2: This seems to experience most crimes at a median average compared to the other clusters.

Cluster 3: This cluster has the lowest overall average crime across categories.

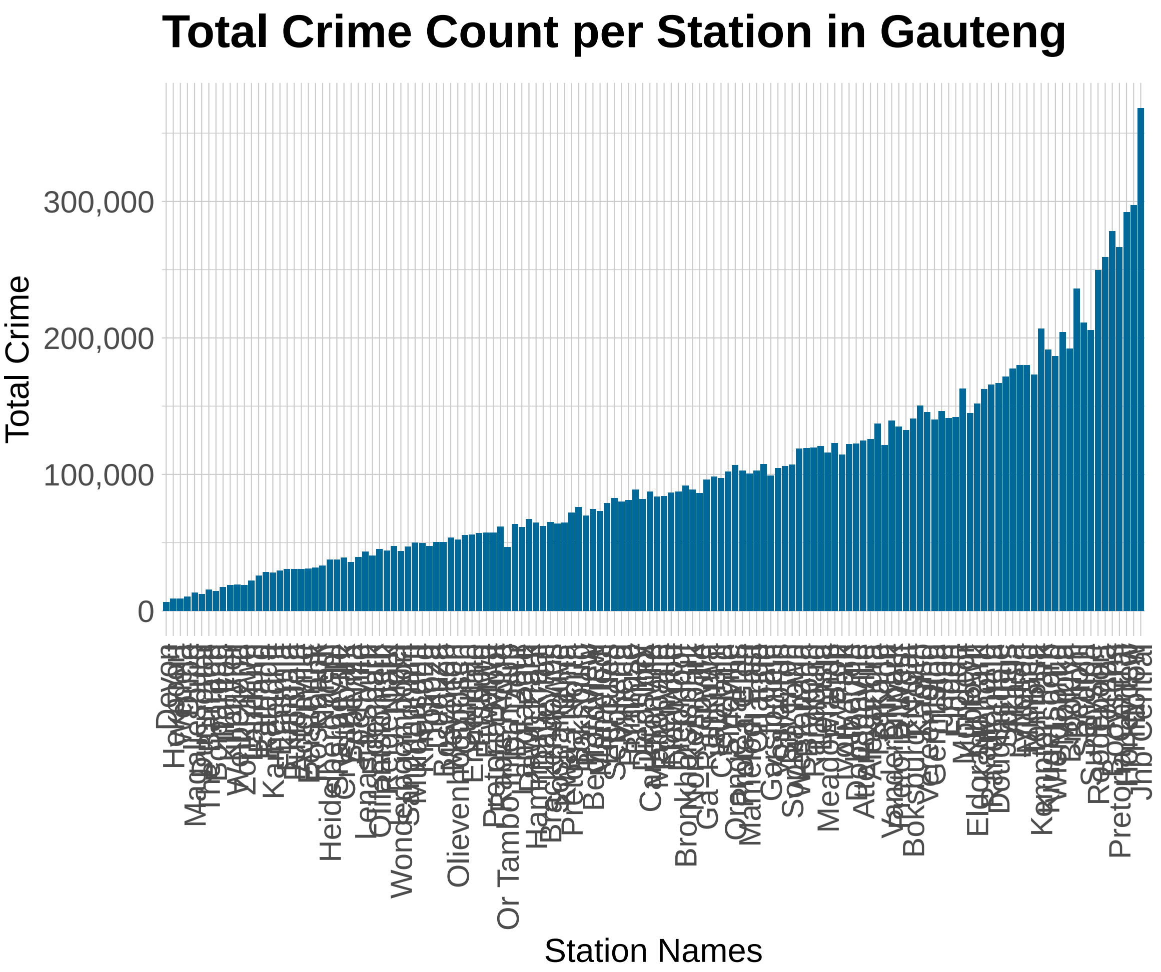
5 Classification

Our goal with our classification model is to investigate whether the model could correctly classify each station into a certain risk category. With this model, we wanted the type of crime to contribute to the crime level, so that an area with 5 murders would not be classified as "less dangerous" than an area with 10 shoplifting reports. In order to take this into account, we consulted SAPS website and classified the crimes into broader categories and assigned each broad category a weight. The process of adding this to our dataset is as follows:

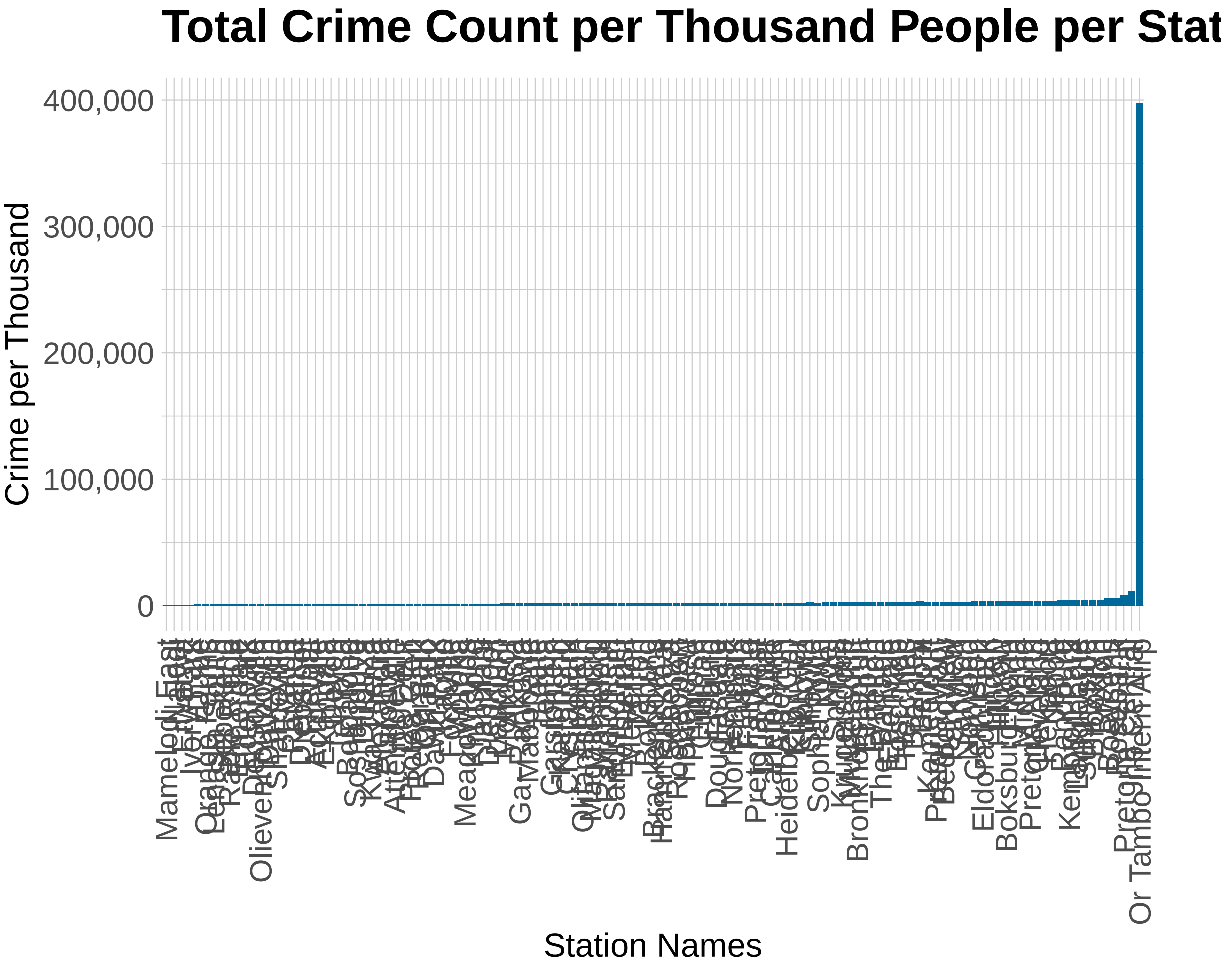
5.1 EDA

In order to visual the changes we made in a easy to read graph, we decided to cut our dataset down to the province of Gauteng in 2011. This made the data small enough to read the graph, but still include stations in diverse areas. In this chunk we also computed the weighted count of the crimes.

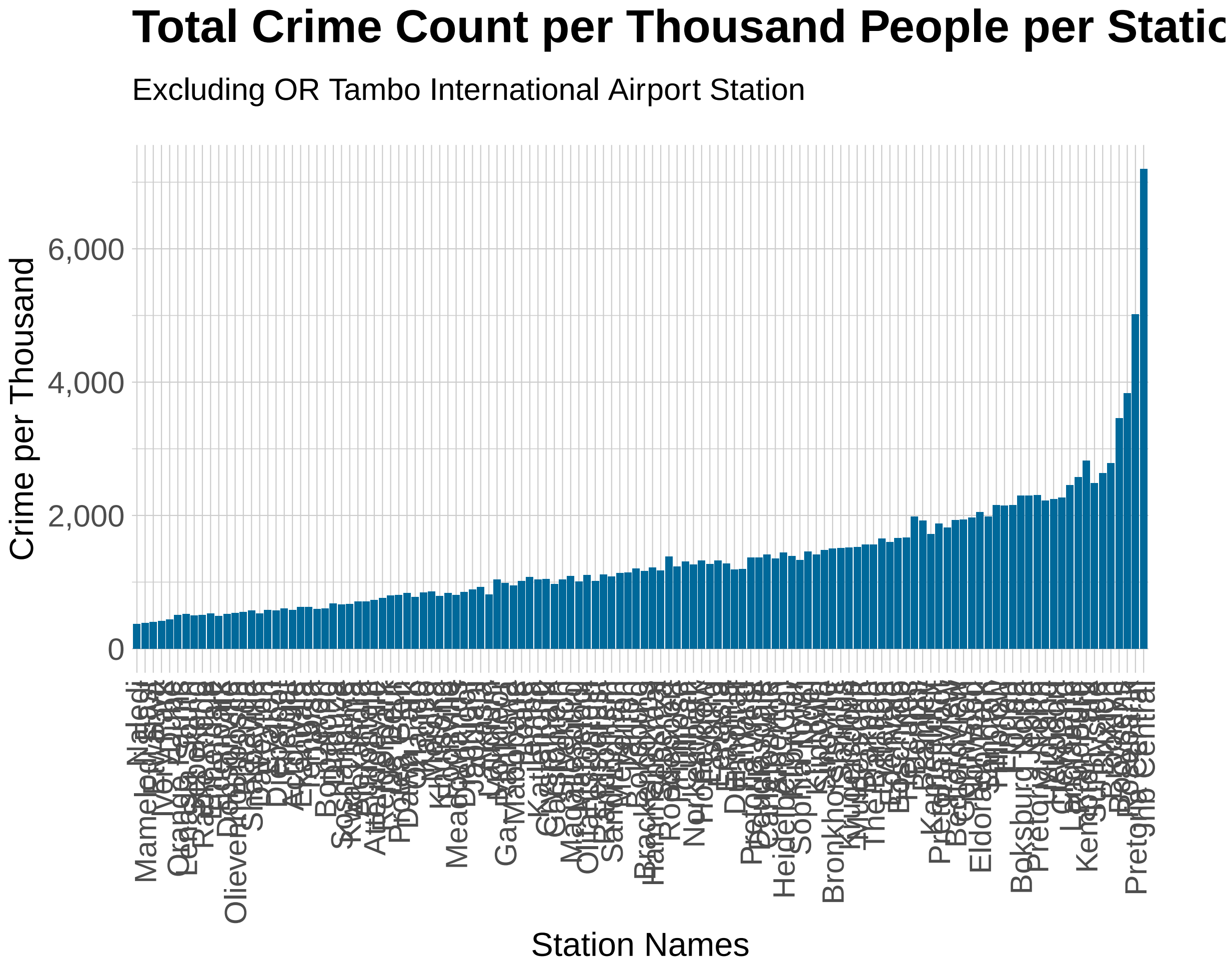
When we first explored this idea, we wanted to take district population into account. We computed extra columns calculating the crime per thousand people, as well as the weighted crime per thousand people.



The graph above shows the total crime count per district in Gauteng. The graph is ordered from lowest crime areas to highest crime areas. This shows that Gauteng has areas of diverse crime levels.

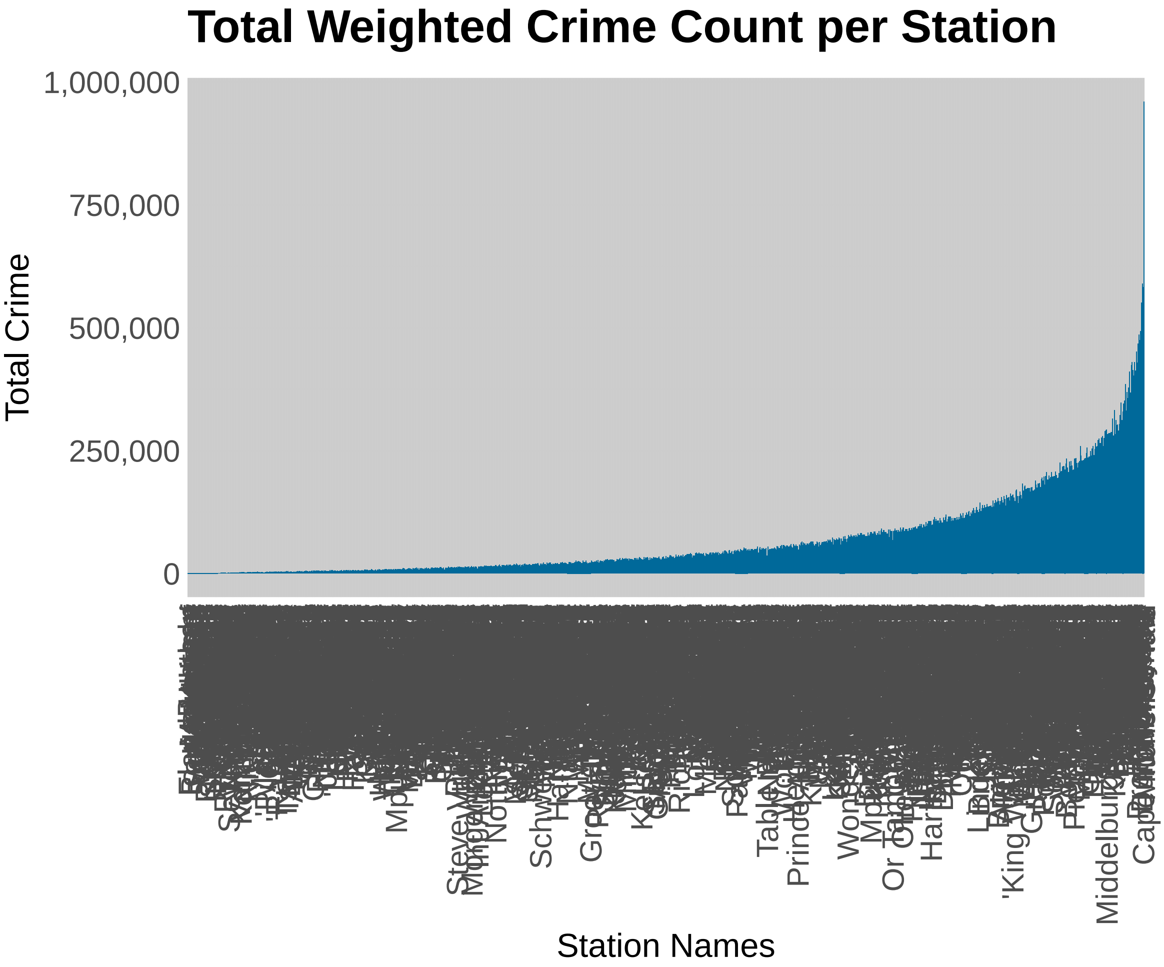


The graph above plots the total weighted crime per thousand people per district. As we can see, there is one station that is heavily skewing the data. This is OR Tambo International Airport police station. This high weighted crime rate is due to the fact that many crimes are reported to that station, despite the fact that the actual district population is very low (Around 170 as of the 2011 Census).



For the graph above, we removed OR Tambo Airport station to show just how skewed the data had become. We also noticed that overall, the effect that population was having was not reflecting the true status of each district, and therefore have not taken it into account further.

Below we increased the dataset to include every province in the year 2011, and this is the data we used to train our model.



The above graph shows the total weighted crime for all stations in the year 2011. This graph helps to visualise the areas that would be high risk, medium risk and low risk.

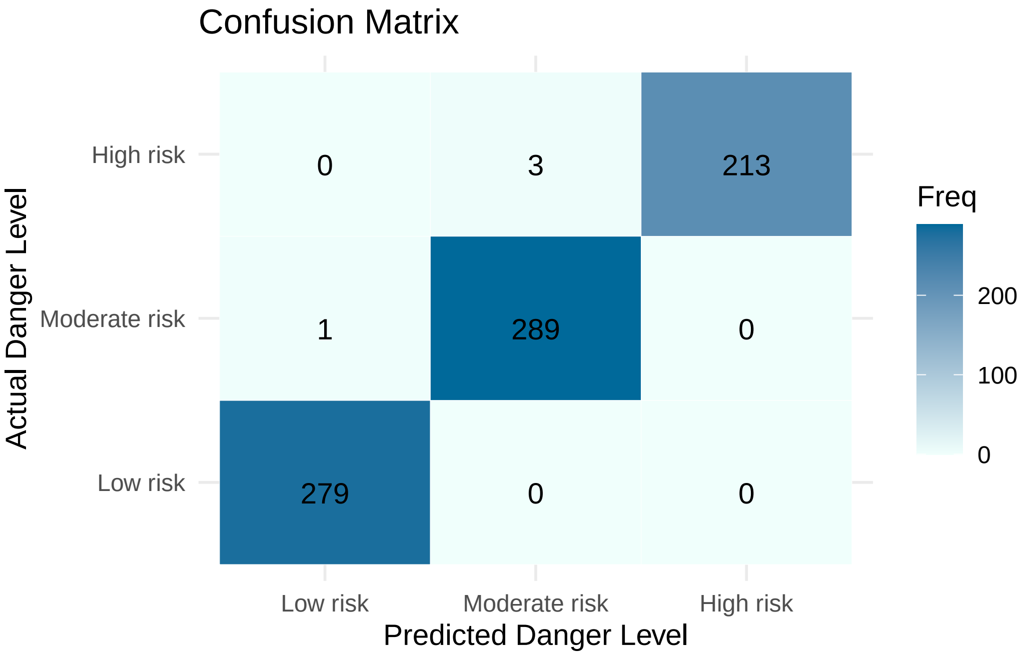
We decided to use the interquartile range to classify the risk levels of each station. Stations with a weighted crime count of greater than the 75% quartile were classified as high risk. Stations with a weighted crime count of less than the 25% quartile were classified as low risk. Every other station was classified as moderate risk.

The dataset is split into 70% train data, and 30% testing data. We used a randome forest classification model since we have 3 categories for it to classify stations into.

We then give the model the testing data for it to make predictions.

```
Confusion Matrix and Statistics

               Reference
Prediction      Low risk Moderate risk High risk
  Low risk           279             1         0
  Moderate risk        0           289         3
  High risk            0             0       213

Overall Statistics

               Accuracy : 0.9949
                 95% CI : (0.987, 0.9986)
    No Information Rate : 0.3694
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9923

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Low risk Class: Moderate risk Class: High risk
Sensitivity                   1.0000               0.9966           0.9861
Specificity                   0.9980               0.9939           1.0000
Pos Pred Value                0.9964               0.9897           1.0000
Neg Pred Value                1.0000               0.9980           0.9948
Prevalence                    0.3554               0.3694           0.2752
Detection Rate                0.3554               0.3682           0.2713
Detection Prevalence          0.3567               0.3720           0.2713
Balanced Accuracy             0.9990               0.9952           0.9931
```

The above output shows the important information about our model. It is 99.62% accurate and has a high confidence interval.

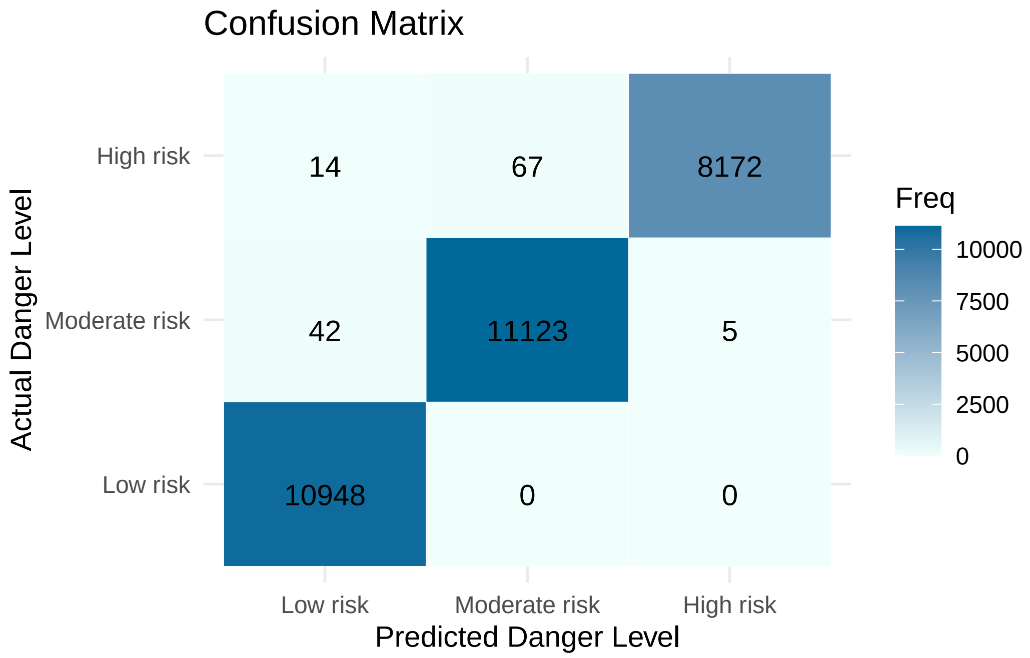


The above plot is a visual representation of our confusion matrix.

5.2 Extra test

In order to see how well our model handles a bigger dataset, we tested it on every year except for 2011 (which it was partially trained on). The increase in data only reduced the model accuracy by 0.6%.

```
Confusion Matrix and Statistics

               Reference
Prediction      Low risk Moderate risk High risk
  Low risk         10948            42        14
  Moderate risk        0         11123        67
  High risk            0             5      8172

Overall Statistics

               Accuracy : 0.9958
                 95% CI : (0.995, 0.9965)
    No Information Rate : 0.3678
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9936

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Low risk Class: Moderate risk Class: High risk
Sensitivity                   1.0000               0.9958           0.9902
Specificity                   0.9971               0.9965           0.9998
Pos Pred Value                0.9949               0.9940           0.9994
Neg Pred Value                1.0000               0.9975           0.9964
Prevalence                    0.3605               0.3678           0.2717
Detection Rate                0.3605               0.3662           0.2691
Detection Prevalence          0.3623               0.3684           0.2692
Balanced Accuracy             0.9986               0.9962           0.9950
```



6 Recommendations

Based on our background and findings we present the following recommendations.

Despite the mixed results of our regression model, gender based violence is a prevalent issue in South Africa. Initiatives should be started to help victims and make them feel safe to come forward for help. Every police station should have a representative able to deal with gender based violence and sexual offences, and collaborate with NGOs that aim to support victims.

Each station should be aware of what type of crime is most prevalent in their district and develop strategies for managing those crimes. For example, areas that our model assigned to cluster 1 should be aware of the heightened prevalence of theft and implement a prevention strategy, such as more frequent patrols around residential areas.

Our classification model is an example of something that be implemented by SAPS to keep track of which areas are at what risk level. If SAPS implements a better, automated system it could automatically assign, weight and track all the reports that are made at a station. It could then update the level of that areas based on current statistics in real time.

Better records of the results of cases should be kept. It was not possible to find data on the resolution of reports that were made, and that type of insight could be useful in identifying areas where the SAPS system is lacking.