Information Visualisation (7CS519)

Crime Data Analytics report

100653888

Table of Contents

1. Part 1 - Understanding of the Dataset **3**

1.1 Description of the dataset *3*

1.2 Distributional Analysis - Summary Statistics *3*

1.3 Basic Visualisations – Histogram *4*

1.3 Basic Visualisations – Charts *5*
Crime Distribution per Location – Pie Chart 5
Crime Distribution per Location – Stacked Bar Chart 5
Correlation Plot 5

Figure 5. Crime Distribution by Type **6**

2. Part 2 – Exploring Relationships in the Dataset **6**

2.1 Simple Linear regression *6*
2.1.1 Assumption model 6
2.1.2 Constant Variance 7
2.1.3 Normality 7
2.1.4 Independence Test 8

2.2 Relationship between Residuals *8*

2.3 Hierarchical Clustering *9*

3. Part 3 – Independent Evaluation and Commentary **10**

3.1 *10*
3.1.2 Data pre-processing 11

3.2. Ethical Considerations *12*

5. References **12**

1. Part 1 - Understanding of the Dataset

1.1 Description of the dataset

The table of crime statistics provides a detailed overview of the crime rates in the Derbyshire area, England. The data table includes different variables including…

i) LSOAs (Lower-layer Super Output Areas) information: The crime data provides the rate of crime in each of 642 LSOAs in the Derbyshire area.
Each LSOAs has been assigned different information variables including..
- LSOAs name
- LSOAs identifier code
- LSOAs population
- LSOAs land area

ii) Crimes: The dataset also includes the rate of occurrences of 14 kinds of crimes reported in each LSOA, including anti-social behaviour, burglary, robbery, vehicle crimes, violent crimes, shoplifting, criminal damage and arson, other theft, drugs, other.
- Also, we might be wrong but it might be useful to point out that from the crime distribution we have 2 sub-categories of crime viz.

Overall, the table provides valuable insight into the crime rates in different areas of Derby. The data can be useful for law enforcement agencies in the Derbyshire area to better understand the patterns of crime in the different LSOAs and devise strategies to reduce the incidence of crime. Additionally, the information can be used by policymakers to allocate resources and develop programs aimed at improving public safety.

1.2 Distributional Analysis - Summary Statistics

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
Population	993	1411	1584	1657	1815	3948	374.1384
Land Area in Hectares	12.81	37.26	71.56	408.84	209.55	16227.09	1115.232
Anti-Social Behaviour	3	22	36	52.35	58	1359	73.9463
Burglary	1	5	8	10.25	13	158	9.075658
Robbery	1	1	1	2.16	2	79	3.944852
Vehicle Crimes	1	5	8	9.34	12	60	6.564563
Violent Crimes	3	22	35	49.89	58	1280	63.81191
Shoplifting	1	1	1	9.73	5	613	32.70965
Criminal Damage & Arson	1	7	12	14.86	19	196	13.66408
Other Theft	1	5	8	11.91	13	382	19.13246
Drugs	1	2	3	4.67	5	150	9.166201
Other Crimes	1	2	3	3.83	5	46	3.652869
Bike Theft	1	1	1	2.55	2	170	7.321158
Possession of Weapons	1	1	2	2.22	3	60	2.885617
Public Order	1	4	7	10.57	11.75	404	20.16094
Theft From the Person	1	1	1	2.22	2	202	8.572847

Fig 1: Table of Summary Statisitics

The summary statistics above provide us with valuable insights into the distribution of the dataset, the skewness and the presence, if any of extreme values.
From the summary, we see that the population of the LSOAs as well as the Land Area variables show a high standard deviation which tells us there is a considerable spread in the population size and land area of the different LSOAs.
The crime variables exhibit a minimum value of 1 for most of the crime categories, except for Anti-Social Behaviour and Violent Crimes with a minimum value of 3. The maximum values for all variables are considerably higher than the third quartiles, suggesting the presence of a few areas with a significantly higher number of crime incidents. The means of all variables are greater than their respective medians, indicating positive skewness in the dataset.

These insights can guide further analysis and interpretation, such as identifying areas with high crime rates or exploring relationships between variables.

1.3 Basic Visualisations – Histogram

The first choice of plot was histogram to

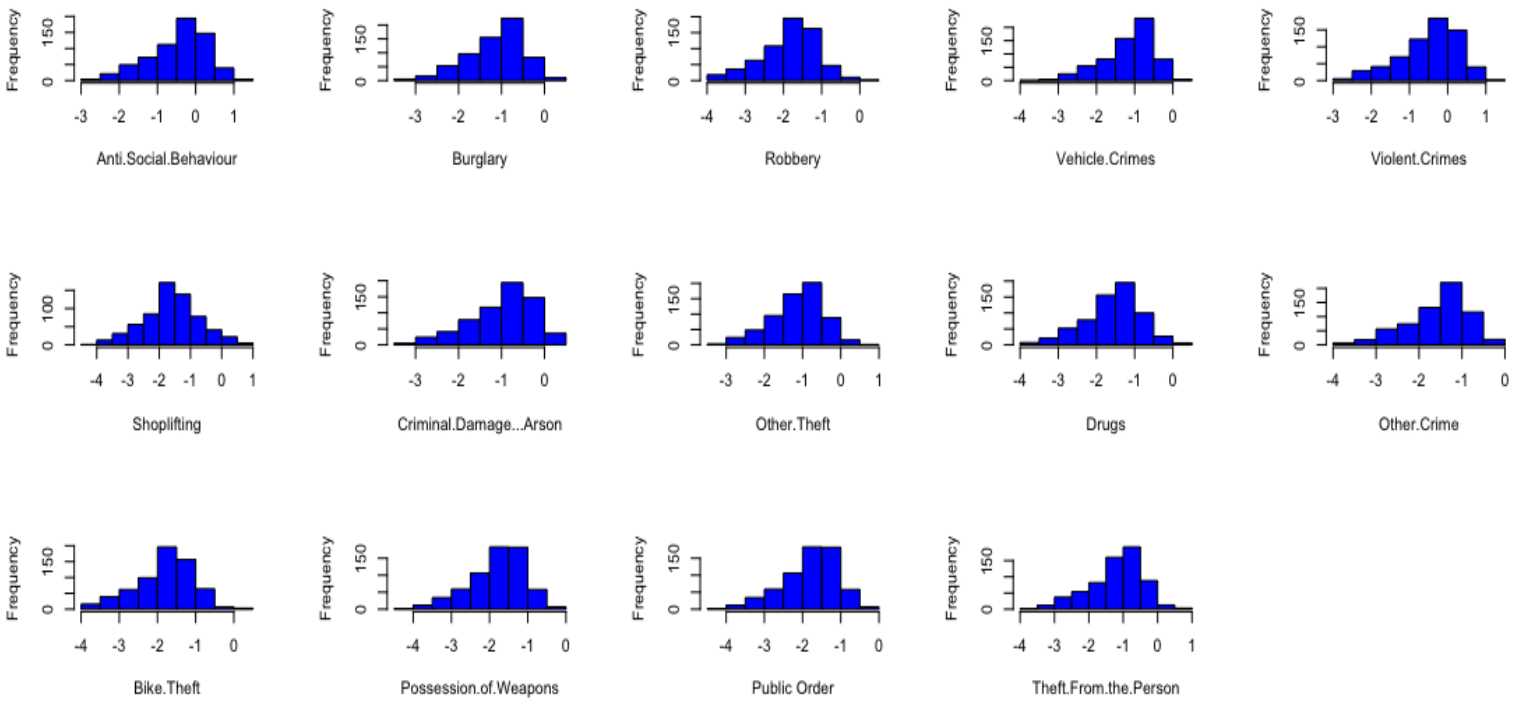

Figure 2. Log 10 Histogram of the independent crime Indicators

1.3 Basic Visualisations – Charts

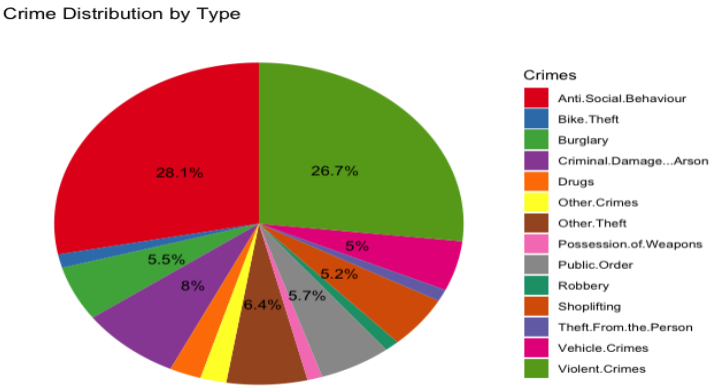

Figure 3. Crime Distribution by Type

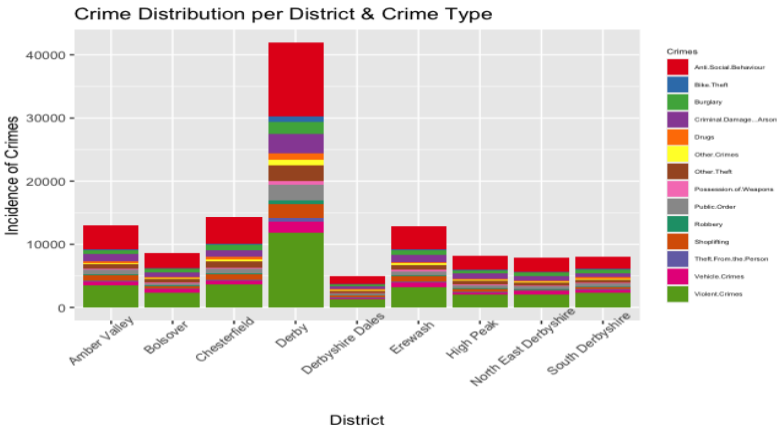

Figure 4. Crime Distribution by Type

Crime Distribution per Location – Pie Chart

Anti-social behaviour and violent crimes are the most common crime in Derbyshire, they cover more than the average of the total crimes in the County. Followed by Criminal Damage Arson and other theft which are 8% and 6.4% respectively. Bike Theft, theft from the person, Drugs, Other crimes, Possession of weapons, and Robbery are each less than 5% of the total crime.

Crime Distribution per Location – Stacked Bar Chart

The chart above shows the crime rate and type of crime across Derbyshire county towns, with each town name extracted from the Name column. It goes in line with the previous chart, as Anti-social behaviour and Violent Crimes seem to be most prominent in all areas. The crime rate at Derby City requires special attention as its bar goes higher compared to others. Although the crime rates are more than each other per city, they vary directly as well i.e. the higher the crime rate the higher each of the crime types, no type is left behind in any city.

Correlation Plot

The correlation plot helps to visualize the relationship between the crime types and the population density (Population divided by Land area) and it reveals that the crime types are highly correlated with each other but none is correlated with the population density as all population density aspects are completely gray out, this implies that the population of an area does not determine the crime type and rate in that area.

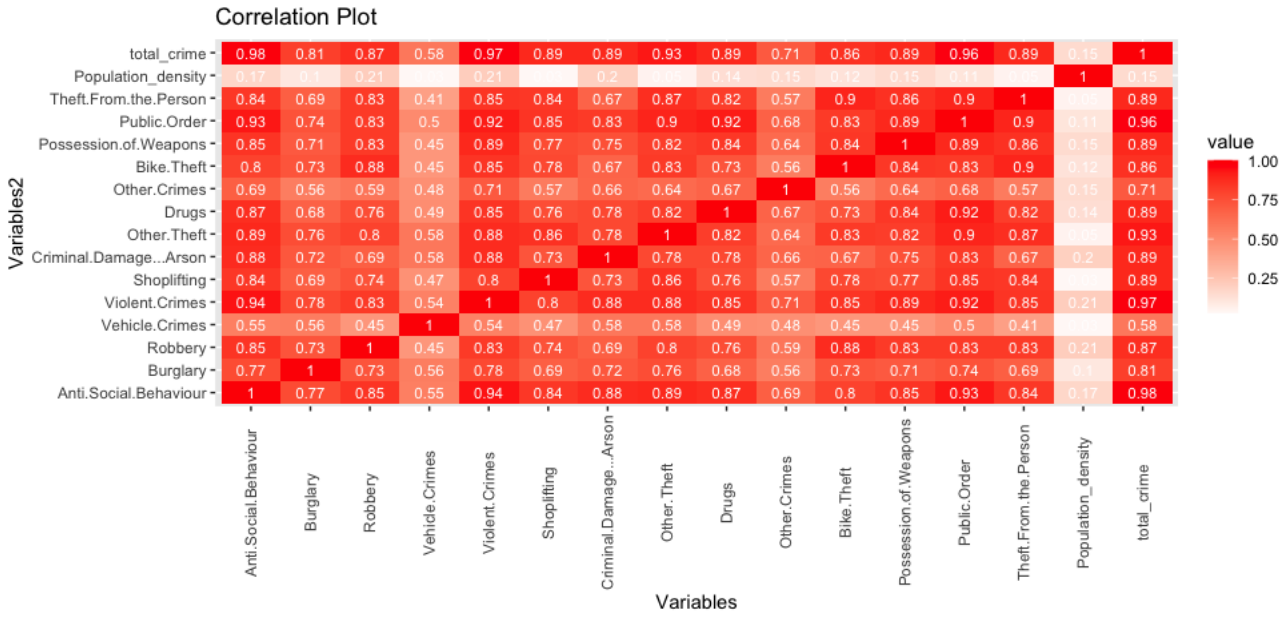

Figure 5. Crime Distribution by Type

2. Part 2 – Exploring Relationships in the Dataset

2.1 Simple Linear Regression

2.1.1 Assumption model

To explore the relationships within the dataset, we set out to define 8 models to be used to determine the extent of the relationships. The assumptions of the first model (were used because it contains the entire dataset. The population density is regressed on total crime. Population density talks about the population and land area and the total crime comprise all the crimes, checking this can serve as a basic assumption for others.

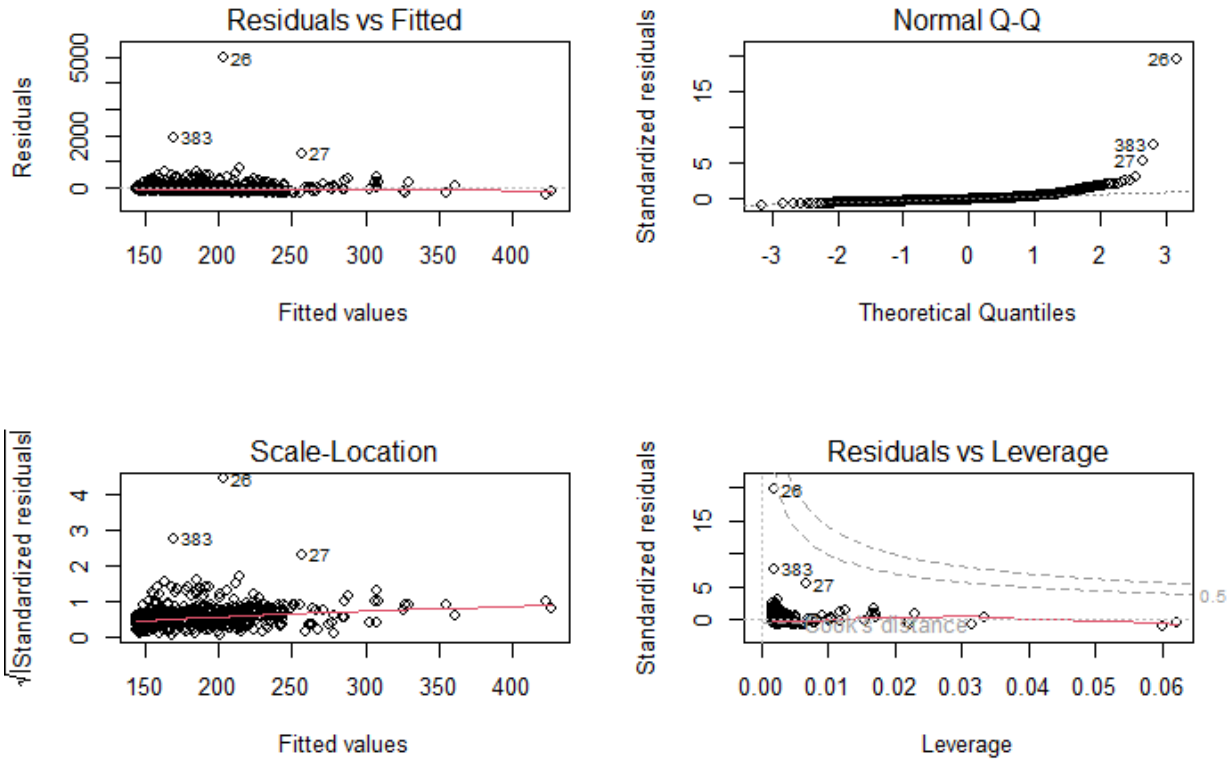

Figure 6. Test of Assumptions

2.1.2 Constant Variance

The **Residuals vs Fitted Values Plot** supported by the **Scale-Location Plot** is used to check for constant variance. With all values randomly scattered around the horizontal line at y = 0, except three observations which can be considered outliers and can be dealt with and with no sign of heteroscedasticity, then we can say the constant variance assumption is met.

2.1.3 Normality

The **Normal Q-Q Plot** is used to verify if the Normality assumption is met, and with all the observations except the three identified outliers on the straight line then we can conclude that the Normality assumption is approximately met.

2.1.4 Independence Test

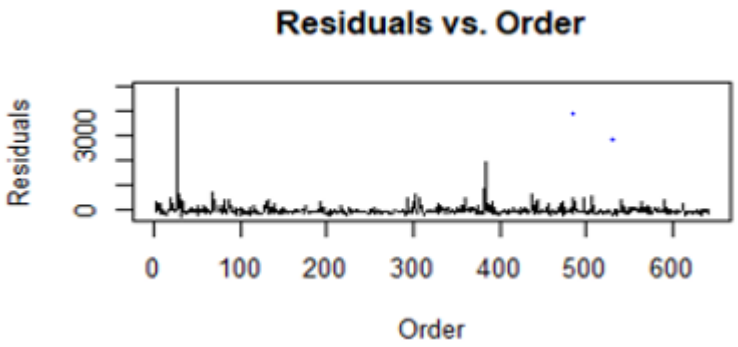

Figure 7. Independent Test

The line plot above is the plot of residuals against their order in the data set. It is plotted to verify the independence assumption of the model. Since the plot shows no clear pattern or trend then it indicates that the independence assumption is not violated.

2.2 Relationship between Residuals

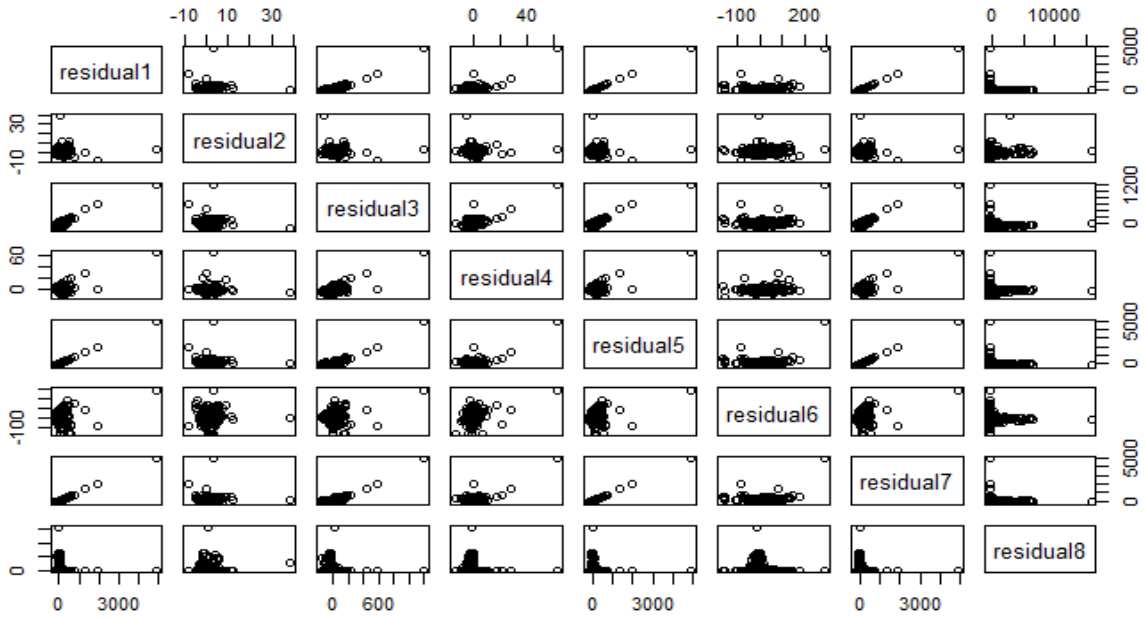

Figure 8. Scatterplot of Residuals

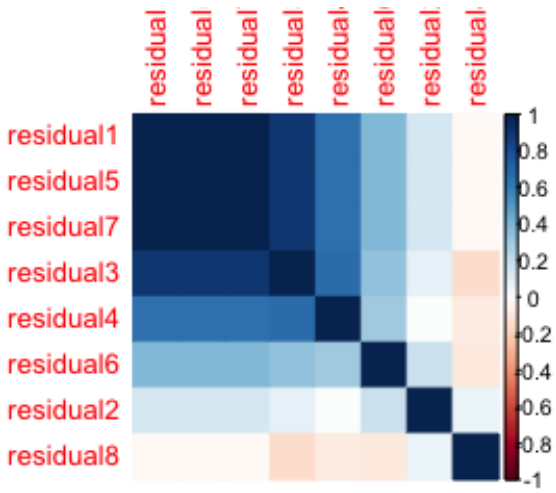

Figure 9. Heatmap of Residuals

The Scatterplot and heatmap are used to check the relationship between the residuals, and these charts reveal that residuals 3, 7, 5, and 1 are highly correlated with one another with residual 4 being partially correlated with them.

2.3 Hierarchical Clustering

Figure 9. Scatterplot of Residuals

100 observations were sampled to organize and group the residuals, this was done to produce a clearer visualization as all observations failed to show clearly. The chart does not only help to show grouping but also figure out the outliers in the data set. Observations 26 and 27 with the higher height indicate outliers.

3. Part 3 – Independent Evaluation and Commentary

3.1

Location by Total Crime

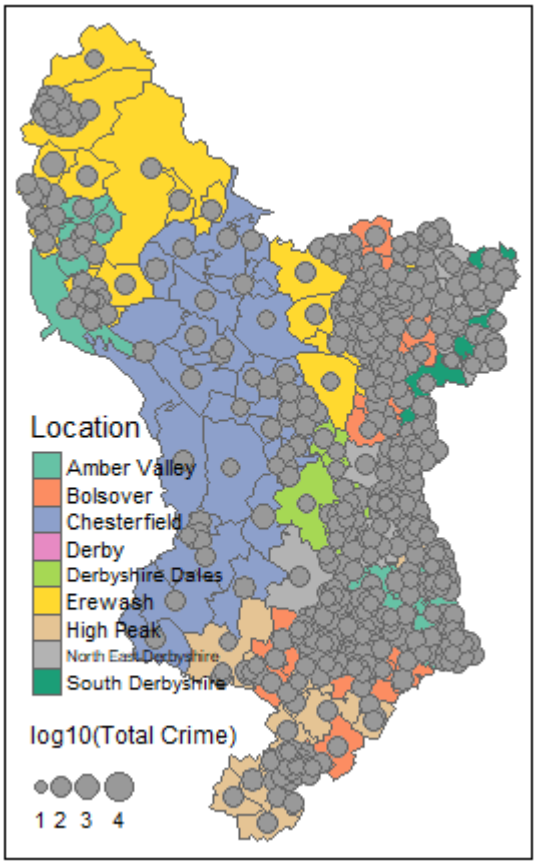

Figure 3.1

Population Density by Anti-Social Behaviour

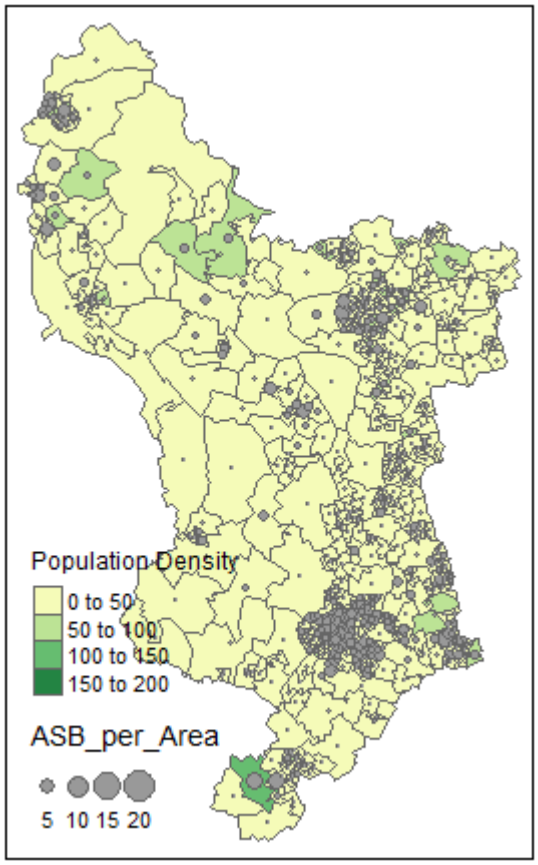

Figure 3.2

Figure 3.1 helps in revealing where each location falls on the map as well as the area where crimes are committed the most, the colours indicate the towns and the ball sizes represent the transformed total size. The chart shows that the east sides, both the northern and the southern parts of the east are the prominent location for crime. Although, this is influenced by the excessive crime rate in Derby town. The total crime is transformed to logarithm Values to help lower the influence of the outliers, if not done, the classification of the ball sizes will be variant.

Figure 3.2. ASB (Anti-social Behaviour) is the top crime in Derbyshire, the chart above shows where this crime occurs and as well reveals that it does not correlate with the population density. As some greener places have fewer dots while some faded-green parts have many dots. It was placed alongside figure 3.1 to gain adequate information on the name of the concerned Locations. The ASB_per Area was used to indicate the rate per land area and to lower the power of the influential points in the dataset.

3.1.2 Data pre-processing

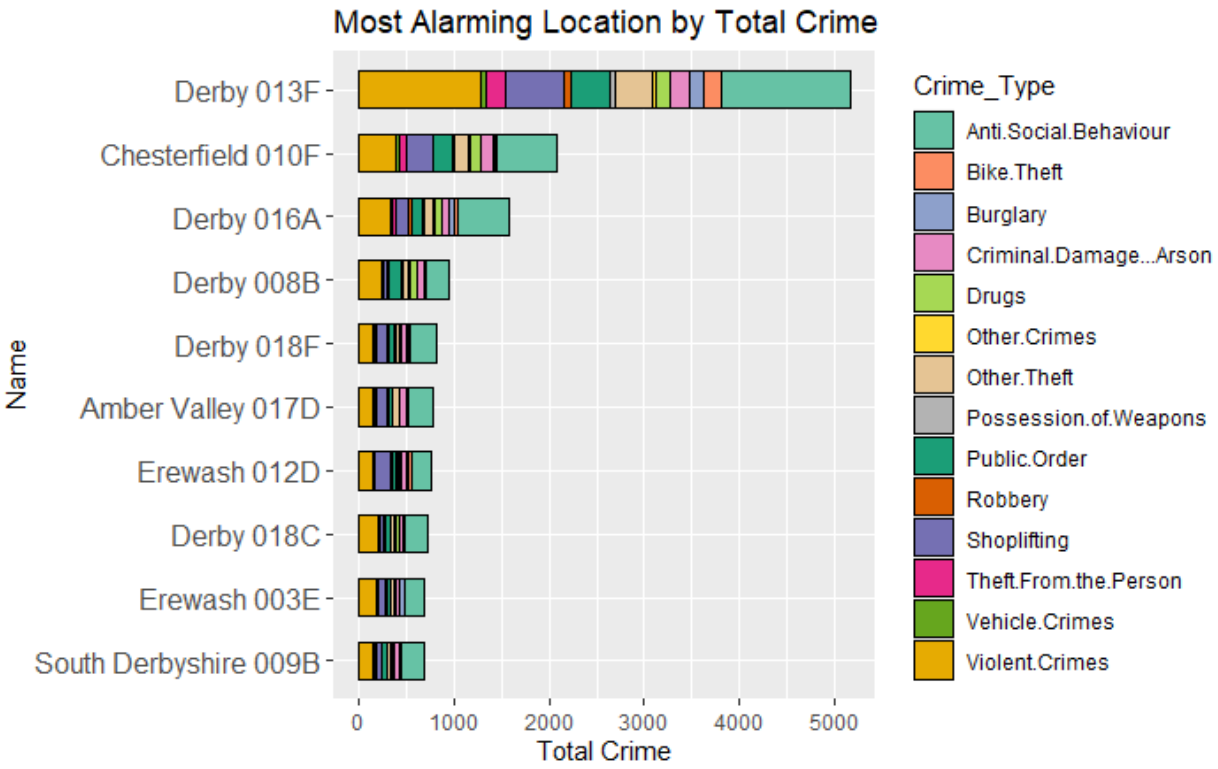

Figure 3.3: Most Alarming Location

Sequel to the analysis in part one, and the outliers detected in part two. This chart was plotted to reveal those regions that were suspicious in the analysis and as well need special attention in order to curb the crime rate on time. It comprises the Name of the top ten regions with high crime rates, coloured by the crime types. The first three bars represent observations 26, 383, and 27 that are pointed out as influential points in the data set.

3.2. Ethical Considerations

For this analysis, there were ethical considerations applied to ensure that the analytics provided were conducted and presented in a morally and responsible fashion. This was considered to ensure that the report did not constitute a misinformation to influence misguided decisions, and other undue harm to potential users of this report. Some of the steps taken include…

Data quality: The analysis was conducted using verified data sources with a high degree of accuracy and reliability

Transparency: the report was documented to ensure that all the analytical steps, assumptions were clearly stated. Limitations encountered were also documented to help the avoidance of misleading analyses.

Contextual Interpretation: The report also tries to provide proper context and interpretation of the analysis results whilst also avoiding the use of generalisations and unnecessary conclusions.

Use of ideal analytical tools: The analysis also employed the use of relevant analytical tools to provide the most ideal results and interpretations.

5. References

Donoho, D.L. (2017) '50 years of data science', Journal of Computational and Graphical Statistics, 26(4), pp. 745-766. Available at: https://doi.org/10.1080/10618600.2017.1384734 (Accessed: May 2023).

Dwork, C., Hardt, M., Pitassi, T., et al. (2012) 'Fairness through awareness', in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12), pp. 214-226. Available at: https://doi.org/10.1145/2090236.2090255 (Accessed: May 2023).

https://observatory.derbyshire.gov.uk/wp-content/uploads/reports/documents/deprivation/ID_2019_Report.pdf