**Crimes in Derbyshire: A Data Analysis Summary using R.**

**University of Derby, DE22 1GB, United Kingdom**

**100643247**

**100643247@unimail.derby.ac.uk.**

Table of Contents

[**Introduction** 3](#_Toc135586937)

[**Derby Crime Dataset Description & Summary Statistics** 3](#_Toc135586938)

[**Box Plot for Population against the 14 crime indicators** 4](#_Toc135586939)

[**Histogram Plot for Population against the 14 crime indicators** 5](#_Toc135586940)

[**Linear Regression** 6](#_Toc135586941)

[**Linearity Plots** 7](#_Toc135586942)

[**Histograms of Normality** 9](#_Toc135586943)

[**Scatter plot for residuals** 11](#_Toc135586944)

[**Using Shapefiles to visualise the data.** 14](#_Toc135586945)

[**Ethics of the Analysis** 18](#_Toc135586946)

[**References:** 19](#_Toc135586947)

# **Introduction**

Crime is a pervasive problem that affects every community, and decision-making about crime prevention and law enforcement can be aided by knowledge of patterns of criminal behaviour, places with the highest rates of crime, and the kinds of crimes that are most frequently committed there (Cozens et al., 2005). The Derby Crime dataset for the UK county of Derbyshire in 2019 is a comprehensive collection of crime information specific to the Derby area, encompassing a wide range of crime types, including Anti-Social Behaviour, Burglary, Robbery, Vehicle Crimes, Violent Crimes, Shoplifting, Criminal Damage & Arson, Other Theft, Drugs, Other Crimes, Bike Theft, Possession of Weapons, Public Order, and Theft from the Person. The use of visualization tools plays a vital role in analysing large and complex datasets like Derby Crime Data, enabling the exploration and discovery of critical information hidden within the dataset. Through visualizations such as tables, charts, graphs, and maps, complex information can be presented in a concise and easily interpretable manner, facilitating data-driven decision-making and effective communication of findings to various stakeholders (Kitchin, 2016).

The analysis of the Derby Crime dataset serves as a crucial tool for law enforcement agencies, policymakers, and community stakeholders. By examining the data and extracting meaningful insights, stakeholders can make informed decisions regarding resource allocation, crime prevention strategies, and community outreach initiatives. Additionally, understanding the nature and trends of crime in Derby can assist in identifying potential social and economic factors that contribute to criminal activities, allowing for targeted interventions and holistic approaches to crime reduction (Braga et al., 2019).

Furthermore, ethical considerations are of utmost importance when conducting any data analysis, including crime datasets. It is essential to approach the analysis and visualization of crime data with a strong ethical framework. This includes ensuring the privacy and confidentiality of individuals involved in the crime incidents, avoiding biases and discrimination, and presenting the data in a manner that accurately represents the realities of crime in Derby (Kitchin, 2016). By critically evaluating and applying ethical principles throughout the analysis process, we can ensure the integrity and reliability of our findings, avoiding misleading interpretations or stigmatization.

Finally, the analysis of the Derby Crime dataset offers a valuable opportunity to gain insights into the crime landscape in the Derby area. By leveraging visualization tools and techniques, we aim to extract meaningful and actionable information from the dataset to inform decision-making, develop evidence-based strategies, and improve public safety. The ethical considerations underlying the analysis and visualization process are crucial to ensure the accuracy, fairness, and integrity of the findings. Through this analysis, we hope to contribute to a safer and more secure community in Derby and provide valuable insights that can be applied to crime prevention efforts in other regions as well.

## **Derby Crime Dataset Description & Summary Statistics**

We are going to describe the dataset to help in understanding the characteristics of the data, such as the type of variables, the range of values, and the distribution of the data. This is an essential step in the data analysis process, enabling to gain a deeper understanding of the data and extract meaningful insights.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **SD** | **Median** | **Trimmed** | **MAD** | **Min** | **Max** | **Range** | **Skew** | **Kurtosis** | **SE** |
| LSOA\* | 321.5 | 185.47 | 321.5 | 321.5 | 237.96 | 1 | 642 | 641 | 0 | -1.21 | 7.32 |
| Name\* | 321.5 | 185.47 | 321.5 | 321.5 | 237.96 | 1 | 642 | 641 | 0 | -1.21 | 7.32 |
| Population | 1657.32 | 374.14 | 1584 | 1610.96 | 284.66 | 993 | 3948 | 2955 | 1.71 | 5.06 | 14.77 |
| Land.Area.in.Hectares | 408.84 | 1115.23 | 71.56 | 149.04 | 63.48 | 12.81 | 16227.09 | 16214.3 | 6.82 | 70.73 | 44.01 |
| Anti.Social.Behaviour | 52.35 | 73.95 | 36 | 40.46 | 23.72 | 3 | 1359 | 1356 | 10.26 | 159.17 | 2.92 |
| Burglary | 10.25 | 9.08 | 8 | 9.05 | 4.45 | 1 | 158 | 157 | 7.55 | 109.3 | 0.36 |
| Robbery | 2.16 | 3.94 | 1 | 1.53 | 0 | 1 | 79 | 78 | 13.27 | 233.09 | 0.16 |
| Vehicle.Crimes | 9.34 | 6.56 | 8 | 8.44 | 4.45 | 1 | 60 | 59 | 2.42 | 10.87 | 0.26 |
| Violent.Crimes | 49.9 | 63.81 | 35 | 40.75 | 25.2 | 3 | 1280 | 1277 | 11.86 | 214.73 | 2.52 |
| Shoplifting | 9.73 | 32.71 | 1 | 3.54 | 0 | 1 | 613 | 612 | 11.62 | 188.21 | 1.29 |
| Criminal.Damage...Arson | 14.86 | 13.66 | 12 | 12.74 | 8.9 | 1 | 196 | 195 | 5.19 | 53.28 | 0.54 |
| Other.Theft | 11.91 | 19.13 | 8 | 9.02 | 5.93 | 1 | 382 | 381 | 12.44 | 221.45 | 0.76 |
| Drugs | 4.67 | 9.17 | 3 | 3.27 | 2.97 | 1 | 150 | 149 | 10.64 | 141.81 | 0.36 |
| Other.Crimes | 3.83 | 3.65 | 3 | 3.27 | 1.48 | 1 | 46 | 45 | 5.38 | 49.73 | 0.14 |
| Bike.Theft | 2.55 | 7.32 | 1 | 1.63 | 0 | 1 | 170 | 169 | 19.1 | 425.05 | 0.29 |
| Possession.of.Weapons | 2.22 | 2.89 | 2 | 1.79 | 1.48 | 1 | 60 | 59 | 13.29 | 250.67 | 0.11 |
| Public.Order | 10.57 | 20.16 | 7 | 7.82 | 4.45 | 1 | 404 | 403 | 13.44 | 236.79 | 0.8 |
| Theft.From.the.Person | 2.22 | 8.57 | 1 | 1.38 | 0 | 1 | 202 | 201 | 20.52 | 462.3 | 0.34 |

The population ranges from 993 to 3948, with a mean of 1657, according to the summary. With a mean of 408, the land area in hectares ranged from 12.81 to 16227.09. The number of offences also varies greatly, with some having minimum counts of 1 and maximum counts of 1359. Another evidence that the data is positively biased comes from the fact that the mean and median of the crime counts are typically higher than their minimum and 1st quartile values. In other words, cities with low crime rates are more prevalent than cities with high crime rates. A few cities may have very high crime rates that distort the general distribution of crime counts towards higher values.

**Chart, box and whisker chart

Description automatically generated****Box Plot for Population against the 14 crime indicators**

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Box plots are an effective graphic tool for summarising and illustrating the distribution of a dataset, therefore we used them. They are especially helpful when comparing many groups or variables since they make it simple to examine each group or variable's central tendency, variability, skewness, and outliers.

From the plots above, most cities have very low crime rates This is a common pattern in crime data, we also noticed that Certain Variables, including Land Area in Hectares, Population, Anti-Social Behaviour, Burglary, Robbery, Vehicle Crimes, Shoplifting, Criminal Damage & Arson, Other Theft, Drugs, Bike Theft, Possession of Weapons, Public Order, and Theft from the Person, have different means and medians. The mean and median are identical for other variables like Other Crimes and Violent Crimes.

We used the Log10 transformations to make skewed distributions more symmetrical, stabilize variance in the statistical models, transform multiplicative relationships into additive ones, and make the data more interpretable. Also, with the aim of unlocking insights that may have been hidden in the raw data. Furthermore, we would also visualise using histogram plots below.

## **Histogram Plot for Population against the 14 crime indicators**

**Chart, histogram

Description automatically generated**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# **Linear Regression**

We will use linear regression produce appropriate visualisations that allows us to check for independence, normality, linearity & homoscedasticity. The plot for independence is below:

A diagram of burglary

Description automatically generated with low confidence

A picture containing text, diagram, screenshot, line

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

For the majority of crime categories, there is no discernible pattern or association between the residuals at various time intervals. This suggests that for these categories of crime, the assumption of independence is true.

## **Linearity Plots**

We are going to carry out a Linear regression as a statistical technique to model the relationship between the dependent variable and one or more independent variables in our dataset. In our crime data analysis, linear regression will be used to investigate the relationship between crime rates and various factors (Kutner et al., 2005). By using linear regression, we can identify the factors that have the strongest impact on crime rates and develop evidence-based interventions that target these factors. The linearity plots are below:

A picture containing text, diagram

Description automatically generated

A picture containing text, diagram, sketch, drawing

Description automatically generated

A picture containing text, diagram

Description automatically generated

Anti.Social.Behaviour: The plot suggests a potential deviation from linearity. The skewness and kurtosis indicate a significant departure from a normal distribution, further suggesting non-linearity.

Burglary: This indicates a relatively linear relationship. The skewness and kurtosis indicate a departure from a normal distribution but not as pronounced as in the previous variable.

Robbery: Indicates a non-linear relationship. The skewness kurtosis indicates a significant departure from normality, reinforcing the non-linearity.

Vehicle.Crimes: Indicates a relatively linear relationship. The skewness and kurtosis indicate a slight departure from a normal distribution.

Violent.Crimes: Indicates a potential deviation from linearity. The skewness kurtosis indicates a significant departure from normality, suggesting non-linearity.

Shoplifting: Indicates a non-linear relationship. The skewness and kurtosis indicate a departure from normality.

Criminal.Damage...Arson, Other.Theft, Drugs, Other.Crimes, Bike.Theft, Possession.of.Weapons, Public.Order, and Theft.From.the.Person: These variables have relatively linear relationship. The skewness and kurtosis indicate departures from normality, but not as pronounced as in some previous variables.

## A picture containing cone, diagram Description automatically generated**Histograms of Normality**

A picture containing diagram, cone

Description automatically generated



These variables plotted above have values that are close, suggesting approximately symmetric distributions.

The skewness values for some variables (e.g., Violent.Crimes, Shoplifting) indicate moderate to high positive skewness, suggesting right-skewed distributions.

The kurtosis values for some variables (e.g., Violent.Crimes, Shoplifting) indicate moderately high kurtosis, suggesting distributions with heavy tails.

**Homoscedasticity Plots**

A picture containing text, sketch, diagram, drawing

Description automatically generatedPlots for homoscedasticity are important for checking the assumptions of linear regression in our crime data analysis, ensuring that the results are valid and reliable.

A picture containing sketch, text, drawing, diagram

Description automatically generated

A picture containing text, diagram, sketch, design

Description automatically generated

we can examine that from the above plots that:

* Anti.Social.Behaviour, Shoplifting, Bike.Theft, Possession.of.Weapons, Public.Order, and Theft.From.the.Person: These variables exhibit relatively high standard deviation values compared to their mean, indicating heteroscedasticity. The range values also vary significantly for these variables.
* Burglary, Vehicle.Crimes, Criminal.Damage...Arson, Other.Theft, Drugs, Other.Crimes: These variables indicate potential homoscedasticity.
* Robbery and Violent.Crimes: These variables indicate potential homoscedasticity. However, the range values are relatively larger, suggesting some variation in the spread of the data points.

## **Scatter plot for residuals**

We will create a scatter plot to examine the residuals between the 14 crimes indicators. Additionally, we will generate a heatmap and perform hierarchical clustering analysis to further explore the data (refer to the plots below).

A picture containing text, black and white, pattern, symmetry

Description automatically generated

A close-up of a color chart

Description automatically generated with low confidence**Heatmap of Derby Crimes**

A picture containing text, diagram, sketch, technical drawing

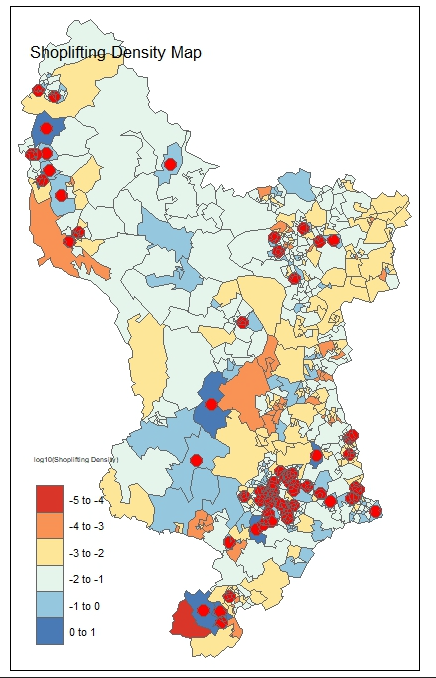
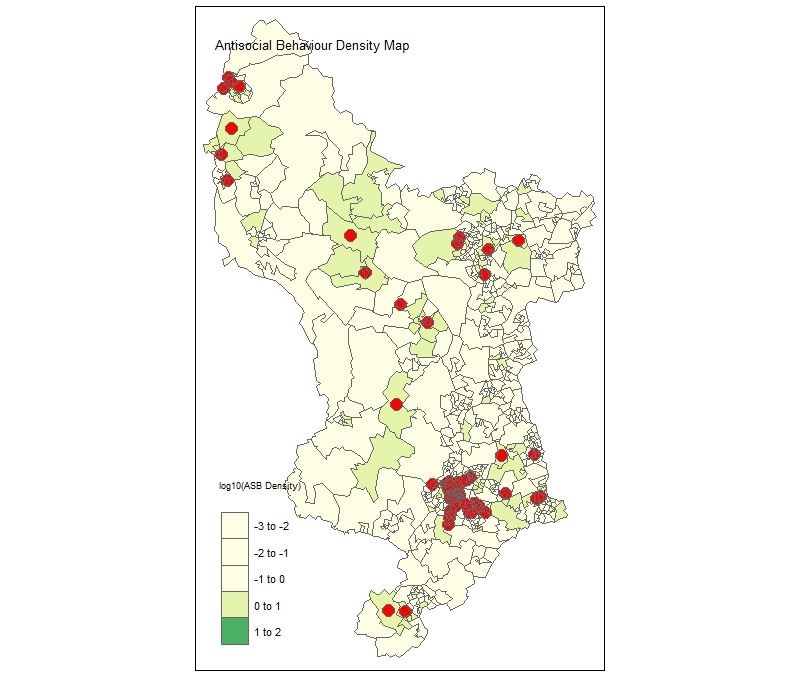
Description automatically generated

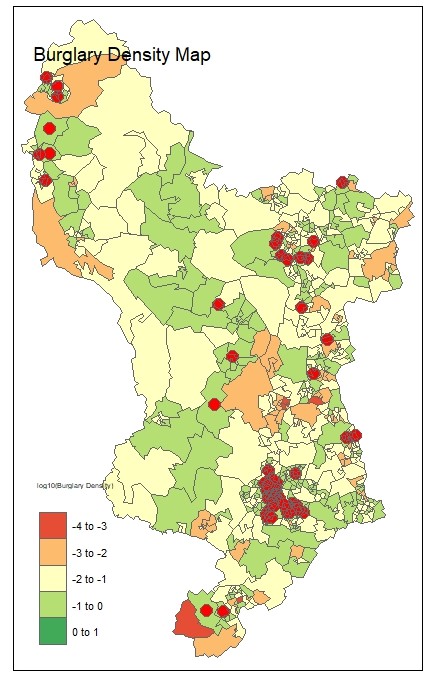
# **Using Shapefiles to visualise the data.**

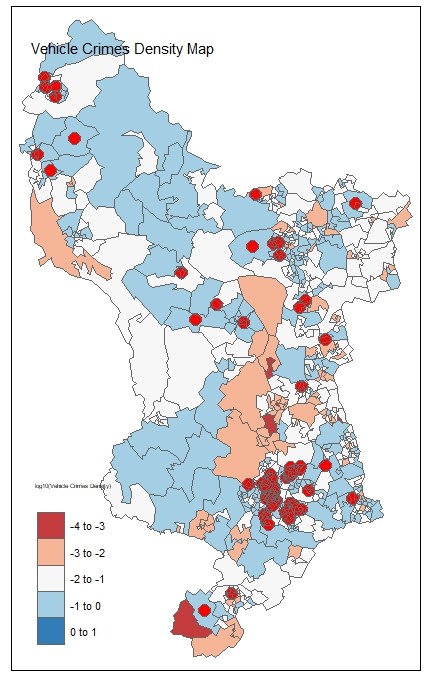
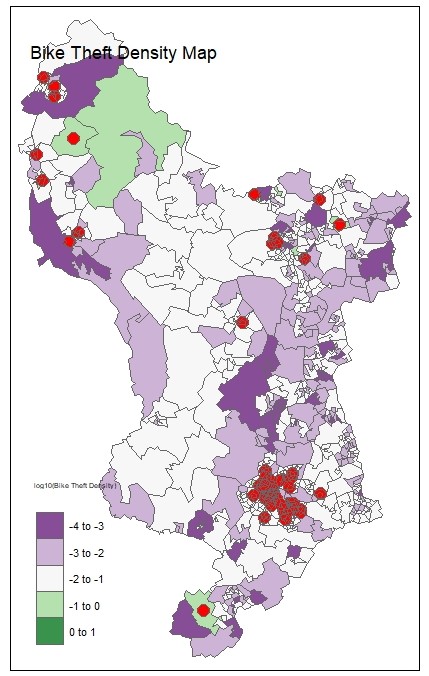
Shapefiles will be used to visualise the data by overlaying it on a map (Kwan, 2004). The purpose of using shapefiles to visualize the crime data are as follows:

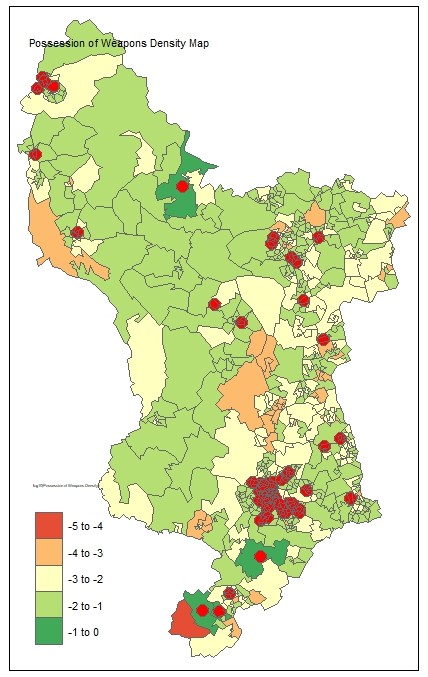
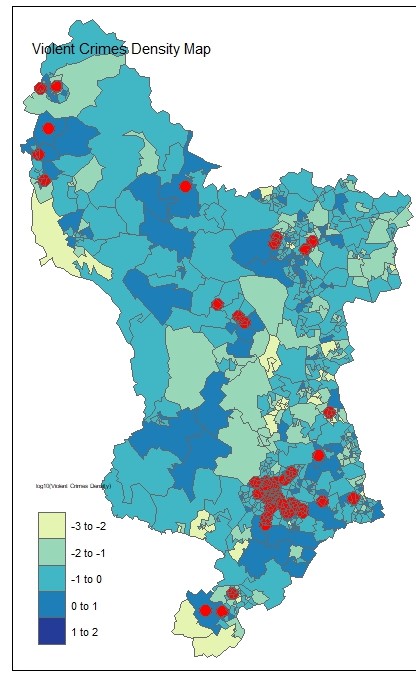
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJECTID** | **LSOA11CD** | **LSOA11NM** | **LSOA11NMW** | **BNG\_E** | **BNG\_N** | **LONG** | **LAT** | **Shape\_\_Are** | **Shape\_\_Len** |
| 13049 | E01013453 | Derby 013A | Derby 013A | 433917 | 335978 | -1.49699 | 52.92017 | 518889.6 | 3480.642 |
| 13050 | E01013454 | Derby 013B | Derby 013B | 434028 | 335510 | -1.49539 | 52.91595 | 227687.6 | 2111.753 |
| 13051 | E01013455 | Derby 017A | Derby 017A | 433477 | 335509 | -1.50358 | 52.91598 | 730268.7 | 3950.947 |
| 13052 | E01013456 | Derby 013C | Derby 013C | 434579 | 335647 | -1.48718 | 52.91715 | 217176.2 | 2092.651 |
| 13053 | E01013457 | Derby 013D | Derby 013D | 434297 | 335174 | -1.49142 | 52.91292 | 242066.8 | 1968.101 |
| 13054 | E01013458 | Derby 017B | Derby 017B | 433913 | 334625 | -1.49719 | 52.90801 | 440384.5 | 3322.047 |

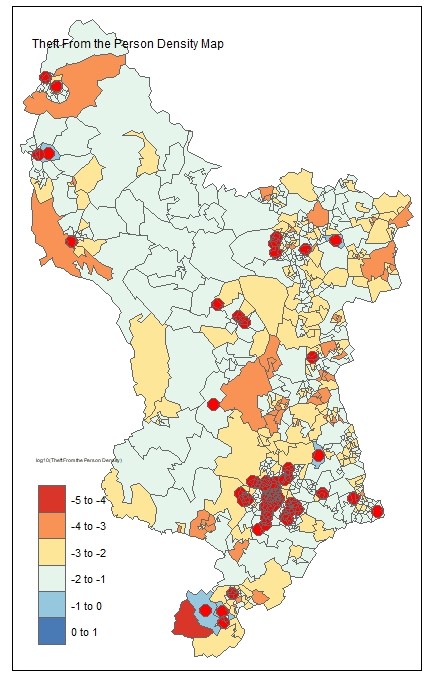
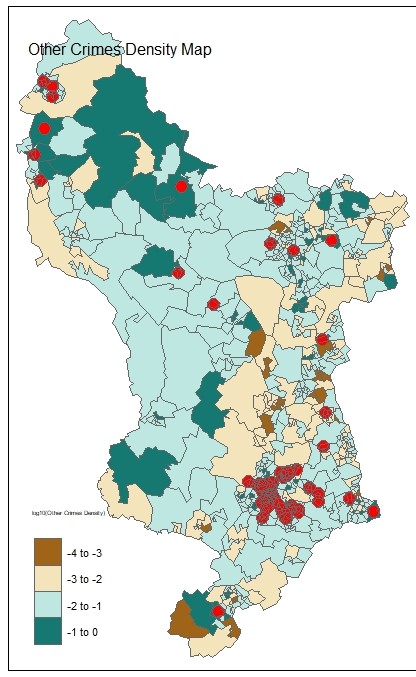
To provide a spatial context for the data, allowing users to see how the data is distributed across a geographic area. This can help in identifying patterns and trends that may not be immediately visible from the data alone (Zhang, Sun and Li, 2022). For example, a shapefile overlaying the Derby crime data on a map can help identify crime hotspots and areas of high risk, providing valuable information for crime prevention strategies and resource allocation decisions. This precision can be useful for analysing the Derby crime data, as it can help in identifying crime hotspots and areas of high risk with greater accuracy (Zhang, Sun and Li, 2022).

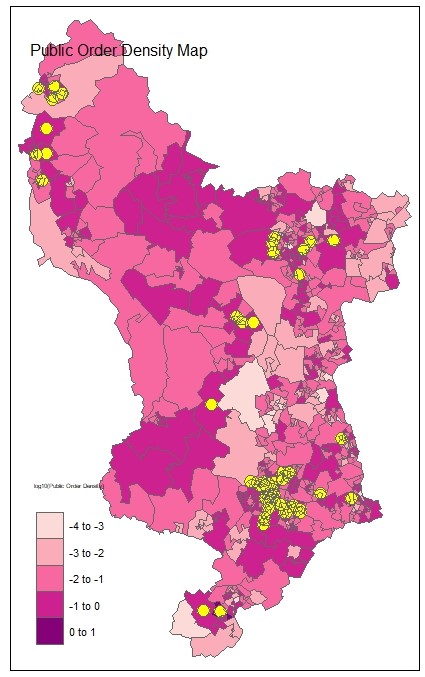
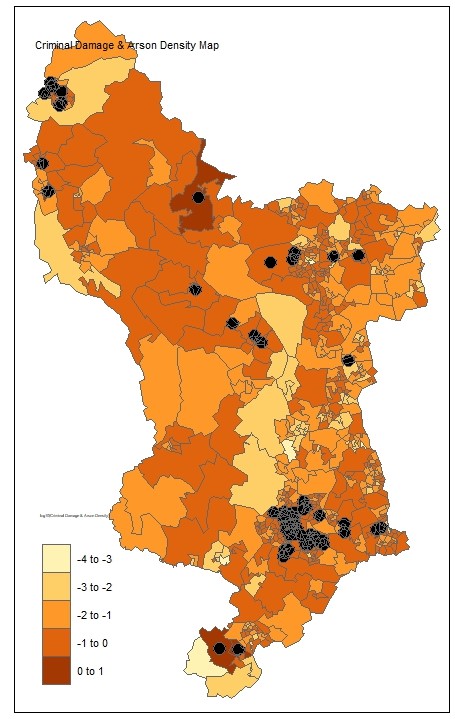


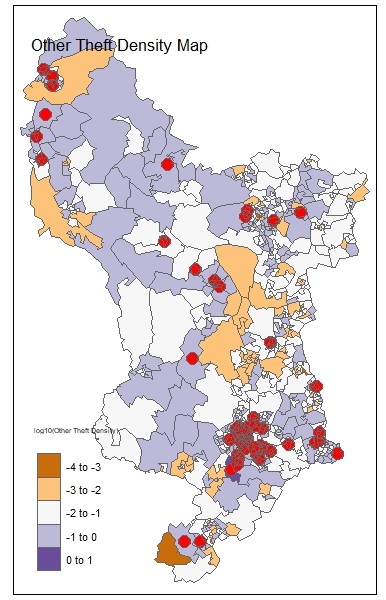
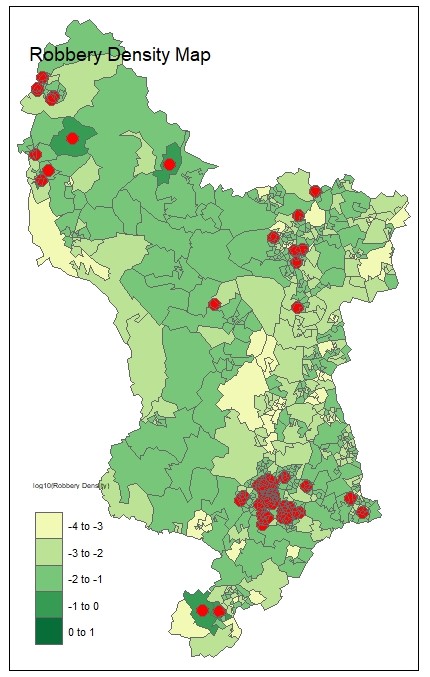












The purpose of the above plots is to visualise the density of various variables across geographic areas using a shapefile. It achieves this by creating density maps for each variable of interest. The plot identifies hotspots, which are areas with values above the 90th percentile for each variable and highlights them on the map. This allows users to easily identify areas with high density values compared to the rest of the dataset.

To represent the hotspots, symbols (dots) are added to the map at the centroid of each hotspot. The symbols are displayed next to each symbol, indicating the identifier of the corresponding geographic area. The main shapefile serves as the base map onto which the density maps and hotspots are overlaid. This provides spatial context and allows for comparison between different variables within the same geographic regions.

Overall, the plots using shapefiles helps visualise the density distribution of various variables and highlights the areas with high values, allowing for quick identification of hotspots and understanding of the geographic patterns associated with each variable enhancing the overall visualisation and making it easier to interpret the map.

# **Ethics of the Analysis**

When analysing and visualizing the Derby crime data, we considered ethical considerations to ensure the accuracy and integrity of the findings. Here are some steps that was taken:

* **Privacy and Confidentiality:** Respecting the privacy and confidentiality of individuals involved in crime incidents is crucial. Personally identifiable information (PII) should be anonymized or aggregated to protect the identities of victims and offenders. Only aggregate statistics was used for analysis and visualization.
* **Data Bias and Discrimination:** Care was taken to avoid bias and discrimination in the analysis and visualisation of crime data. Unfair or prejudiced interpretations can lead to stigmatization of certain communities (D'Ignazio et al., 2020).
* **Accurate Representation:** We ensured that the Data visualizations accurately represents the realities of crime in Derby. Manipulating visualizations or using misleading scales can distort the perception of crime rates and patterns. Clear labelling, appropriate scaling, and accurate data representation should be employed to present the information transparently (Few, 2012).
* **Contextualization and Interpretation:** Proper contextualization is necessary to understand the underlying factors contributing to crime. It is important to interpret the findings in conjunction with socio-economic, cultural, and historical factors that may influence crime patterns. This will help stakeholders develop comprehensive strategies for crime prevention and community safety (Barlow and Decker, 2010).
* **Stakeholder Engagement:** Engaging stakeholders, including community members, law enforcement agencies, and policymakers, is essential throughout the analysis and visualization process. Their input and perspectives can provide valuable insights and ensure the analysis addresses the specific needs and concerns of the community (Bannister & Kearns, 2013).

# **References:**

1. Bannister, J. and Kearns, A. (2013). Understanding Community Safety: Maps and Analysis. Policy Press.
2. Barlow, H.D. and Decker, S.H. (2010). Criminology and public policy: putting theory to work. Philadelphia, Pa.: Temple University Press.
3. Braga, A.A., Weisburd, D. and Turchan, B. (2019). Focused deterrence strategies effects on crime: A systematic review. Campbell Systematic Reviews, [online] 15(3). Available at: https://prohic.nl/wp-content/uploads/2020/11/2020-03-31-FocussedDeterrenceBraga.September2019.pdf [Accessed 15 May 2023].
4. Cozens, P., Saville, G. and Hillier, D. (2005). (PDF) Crime Prevention through Environmental Design (CPTED): A Review and Modern Bibliography. [online] ResearchGate. Available at: https://www.researchgate.net/publication/239746349\_Crime\_Prevention\_through\_Environmental\_Design\_CPTED\_A\_Review\_and\_Modern\_Bibliography [Accessed 17 May 2023].
5. Few, S. (2012). Show Me the Numbers: Designing Tables and Graphs to Enlighten. [online] Google Books. Analytics Press. Available at: https://books.google.co.uk/books/about/Show\_Me\_the\_Numbers.html?id=1xiHLgEACAAJ&redir\_esc=y [Accessed 18 May 2023].
6. Kitchin, R. (2016). The ethics of smart cities and urban science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), p.20160115. doi:https://royalsocietypublishing.org/doi/10.1098/rsta.2016.0115.
7. Klein, L.F. and D’Ignazio, C. (2020). Data feminism. [online] Cambridge, Massachusetts: The Mit Press. Available at: https://data-feminism.mitpress.mit.edu/ [Accessed 17 May 2023].
8. Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2005). Applied Linear Statistical Models. [online] McGraw-Hill/Irwin. Available at: https://users.stat.ufl.edu/~winner/sta4211/ALSM\_5Ed\_Kutner.pdf [Accessed 17 May 2023].
9. Research Geografiska Annaler ; Kwan, G.M.I.T.-G. (2004). GIS METHODS IN TIME-GEOGRAPHIC RESEARCH: GEOCOMPUTATION AND GEOVISUALIZATION OF HUMAN ACTIVITY PATTERNS. Geogr. Ann, [online] 86, pp.86–90. Available at: http://meipokwan.org/Paper/GeogAnnaler\_2004.pdf [Accessed 18 May 2023].
10. Tilley, N. and Sidebottom, A. (2017). Handbook of Crime Prevention and Community Safety. Taylor & Francis.
11. Zhang, K., Sun, H. and Li, X. (2022). Aging Population Spatial Distribution Discrepancy and Impacting Factor. Sustainability, [online] 14(15), p.9528. doi:https://doi.org/10.3390/su14159528.