DATA ANALYST - 2

DATA VISUALIZATION WITH R

Adedotun Teminiola Inaolaji

[Company Name]

Introduction

This analysis is to gain insights into the crime patterns and trends within Derby, this report presents an analysis of the crime data set obtained from Derbyshire, a city in the United Kingdom. The data set comprises information from 642 observations across 18 variables, providing a comprehensive overview of various crime-related aspects in different regions of Derby. The variables include socio-economic factors, population statistics, land area, and the occurrence of different types of crimes such as anti-social behavior, burglary, robbery, vehicle crimes, violent crimes, shoplifting, criminal damage, and more.

Objective

I will be visualizing to determine if and why certain areas were prone to crime. For example, if a particular area has a higher crime rate compared to another area with similar population and land area, further investigation can be done to determine the cause. The primary objective to visualize and provide information that can assist in the decision-making processes for law enforcement agencies, local authorities, and community organizations. By examining the relationships between population, land area, and different types of crimes. I aim to identify areas that may require targeted interventions, resource allocation, or policy adjustments to address and mitigate criminal activities effectively.

Additionally, this report explores the variations in crime rates across different regions within Derby. By analyzing the distribution of crimes across neighborhoods, we can identify areas with higher crime rates, understand the prevalent types of crimes in specific regions, and evaluate the effectiveness of existing crime prevention strategies.

DATASET INFORMATION

The structure of the dataset indicates that it is a data frame consisting of **642 observations (rows) and 18 variables (columns)**. Here is the interpretation of each variable:

LSOA (Character): This variable represents the Lower Layer Super Output Area (LSOA) code for each observation. LSOAs are geographic divisions used for statistical purposes in the UK.	**Population (Integer):** This variable denotes the population count for each LSOA, indicating the number of individuals residing in that particular region.
Name (Character): The Name variable contains the names or labels associated with each LSOA, representing the specific regions or areas within Derby.	**Land.Area.in.Hectares (Numeric):** The variable represents the land area in hectares associated with each LSOA, providing an indication of the geographical size or extent of the region.
Anti.Social.Behaviour(Integer): This variable represents the number of reported incidents related to anti-social behavior within each LSOA. An Acts that harass, upset, or cause pain to others	**Other.Theft (Integer):** The Other.Theft variable represents the count of reported incidents related to theft other than shoplifting within each LSOA
Burglary (Integer): The variable indicates the count of reported burglary incidents within each LSOA. The act of trespassing into buildings or other property with intent to commit a crime	**Drugs (Integer):** This variable indicates the count of reported incidents involve the use or supply of illegal drugs within each LSOA,
Robbery: (Integer): This variable represents the number of reported robbery incidents within each LSOA. An Act of taking someone else's property by force or threat. Take someone else's property by force or threat	**Other.Crimes (Integer):** The variable represents the count of reported incidents that do not fall into specific crime categories mentioned earlier within each LSOA.
Vehicle Crime (Integer): This variable denotes the count of reported vehicle-related crimes (such as theft or vandalism) within each LSOA. A crime that involving vehicles, such as theft or criminal damage	**Bike.Theft (Integer):** This variable denotes the count of reported incidents related to bike theft within each LSOA

Violent.Crimes (Integer): The variable indicates the number of reported incidents involving violence within each LSOA. Crimes that involve violence or threat of violence	**Possession.of.Weapons (Integer):** This variable represents the count of reported incidents involving the possession of weapons, such as guns or knives within each LSOA.
Shoplifting (Integer): This variable represents the count of reported incidents related to shoplifting within each LSOA. Theft of goods from shops and stores.	**Public.Order (Integer):** The Public.Order variable denotes the count of reported incidents related to public disorder within each LSOA
Criminal.Damage…Arson (Integer): This variable denotes the number of reported incidents (Property damage or vandalism) related to criminal damage or arson within each LSOA.	**Theft.From.the.Person (Integer):** This variable represents the count of reported incidents involving theft from a person within each LSOA.

This data set is very useful for analyzing crime patterns in different parts of Derbyshire. By matching crime indicators with population and land area statistics.

Summary Statistics For Crime-Data

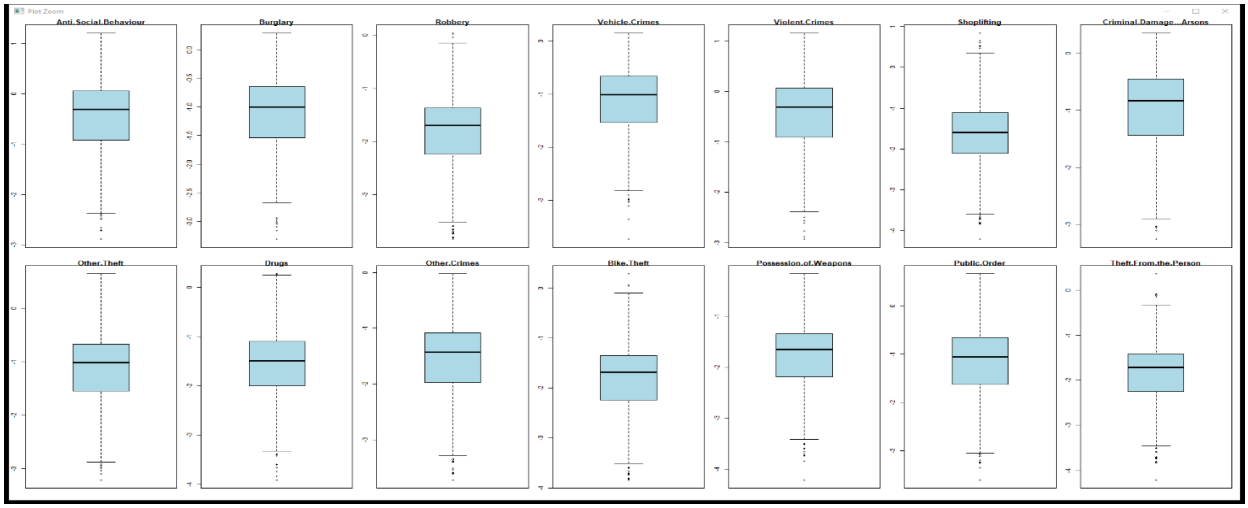
Crime Indicators	Mean	Median	Standard Deviation (SD)	Skew	Kurtosis	Range	Min	Max
Anti Social Behaviour	52.35	36.00	73.95	10.26	159.17	1356.00	3.00	1359.00
Burglary	10.25	8.00	9.08	7.55	109.30	157.00	1.00	158.00
Robbery	2.16	1.00	3.94	13.27	233.09	78.00	1.00	79.00
Vehicle Crime	9.34	8.00	6.56	2.42	10.87	59.00	1.00	60.00
Voilent Crimes	49.90	35.00	63.81	11.86	214.73	1277.00	3.00	1280.00
shoplifting	9.73	1.00	32.71	11.62	188.21	612.00	1.00	613.00
Criminal Damage & Arson	14.86	12.00	13.66	5.19	53.28	195.00	1	196.00
Other theft	11.91	8.00	19.13	12.44	221.45	381.00	1	382.00
Drugs	4.67	3.00	9.17	10.64	141.81	149.00	1	150.00
Other crimes	3.83	3.00	3.65	5.38	49.73	45.00	1	46.00
Bike theft	2.55	1.00	7.32	19.10	425.05	169.00	1	170.00
Possession of Weapon	2.22	2.00	2.89	13.29	250.67	59.00	1	60.00
Public order	10.57	7.00	20.16	13.44	236.79	403.00	1	404.00
Theft from the Person	2.22	1.00	8.57	20.52	462.30	201.00	1	202.00

- The summary statistics show us that difference in Anti Social Behaviour mean 52.35 and medium (36.00) is high, it means that the data is skewed or not evenly distributed. its positively skewed distribution
- Theft from the person mean (2.22) and medium (1.00) show the data is highly skewed to the right with a skewness value of 20.52, indicating the presence of some extremely high values

- The differences in the mean and median values shows that each crime type may be skewed in different directions.

Visual Inspection of Crime Data for the Year 2019 Using a Box plot

Boxplot visually shows the distribution of numeric data and skewness by showing the quartiles (or percentiles) and means of the data. Boxplots are useful for visualizing because they show outliers in your dataset.

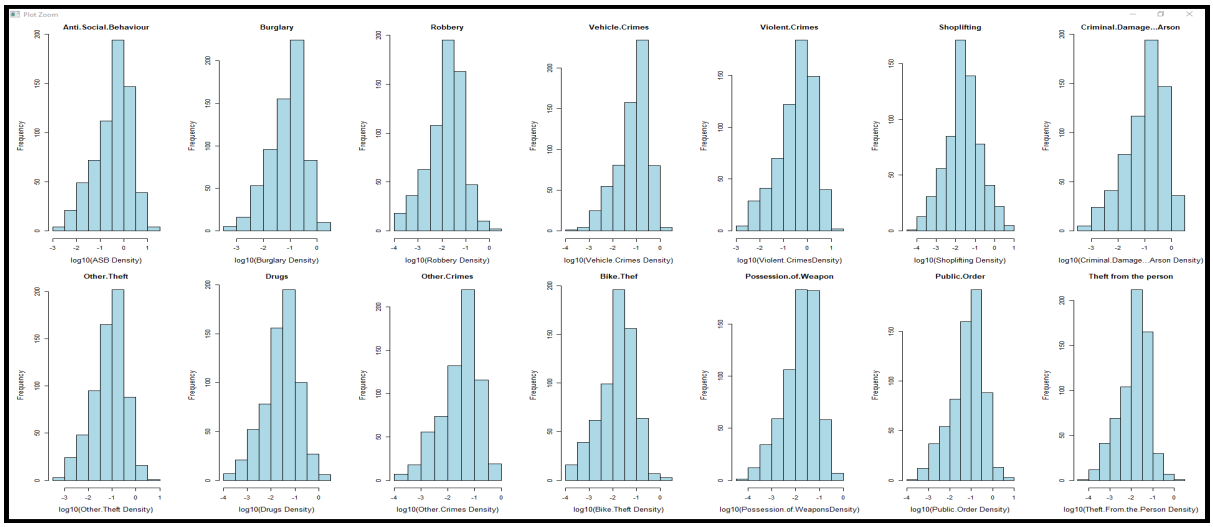


Comment: the boxplots above shows that data is not normally distributed around the mean and is skewed towards one side. Most of the boxplots above has a longer whisker and outliers on one side that shows a skewed distribution expect for Shoplifting.

- Shoplifting's plot shows a roughly normal distribution as both whiskers are of similar length and the median is in the centre of the box. However, all other plots are left-skewed.

Visual Inspection of Crime Data for the Year 2019 Using a Histogram

The shape of the histogram provides us a valuable insight into data distribution. For example, a histogram with a bell-shaped curve indicates a normal distribution, and a skewed histogram indicates a non-normal distribution.



Comment: the histogram plot shows that the data set are not normally distributed, they are mostly skewed to the left the distribution has a longer tail on the left-hand side, most data value are not on the right-hand side.

- The shoplifting data appear normally distributed, as evidenced by the symmetric appearance of its boxplot above and confirmed by the relatively normal shape of its histogram. Non-normal distribution of the data can be due to outliers, sampling errors, or measurement errors.

Histograms and boxplots provide important insight into the distribution of your data and help you identify patterns and trends in your data.

Data Transformation

(a) Log10 transformation technique was applied for scaling and normalizing the Crime dataset to address issues with the wide range of values and skewed distribution of the original data. This helped to normalize the distribution and make it more symmetrical, improving the accuracy of statistical analysis and making it easier to compare and interpret the data.
(b) The data was pre-process, whereby we have to remove the columns 'LSOA' and 'Name', as they are not relevant for clustering.
(c) A new column was created “regio” by separating the region and the post code from the name column.
(d) In creating the geo-plot, I renamed the LSOA11CD on the crime data set in order to merge it with the shapefile.

Part 2:

Simple Linear Regression

Simple Linear Regression: Conducting linear regression to make sure that the four assumptions are met. Based on Linear relationship, Independence, Homoscedasticity (constant variance), Normality. Basically, to understand the relationship between two variables, where population and land are independent variable while each crime indicators is the dependent variable.

The simple linear regression model:

$$Crime = b0 + b1*Population$$

B0 and B1 represent the intercept and slope of the regression line,
e is the residual standard error of the model

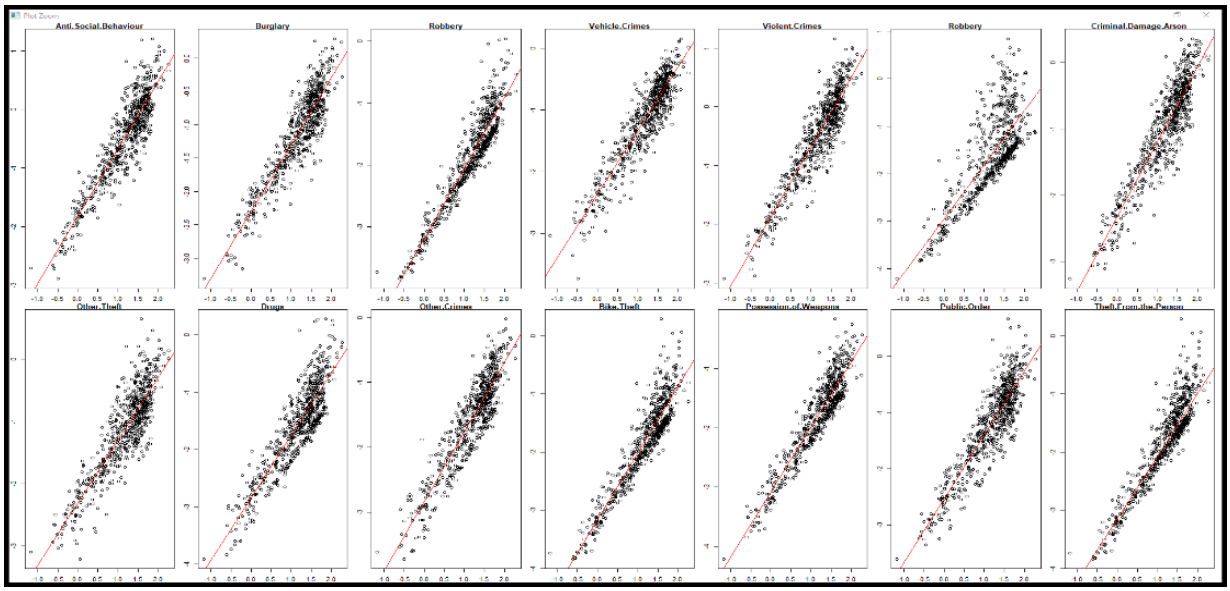
Th simple linear regression models predict the frequency of different types of crime based on population density. Each of the model uses the log-transformed crime rate per unit of land area as the response variable and the log-transformed population density per unit of land area as the predictor variable.

Model	Dependent variable	Independent variable	B0	B1	e
1	Anti.Social.Behaviour/Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-1.82146	1.14829	0.313
2	Burglary/Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.30167	1.00289	0.2725
3	Robbery/Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-3.18957	1.14904	0.2561
4	Vehicle.Crimes/Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.35954	1.02457	0.223395
5	Violent.Crimes/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-1.87080	1.18624	0.2933
6	Shoplifting/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.95277	1.14475	0.5598
7	Criminal.D.Arson/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.30629	1.12458	0.3011
8	Other.Theft/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.30958	1.00335	0.3369
9	Drugs/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.84592	1.08812	0.3422
10	Other.Crimes/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.83232	1.07552	0.2901

11	Bike.Theft/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-3.15880	1.14374	0.2863
12	Possession.of.Weapons/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-3.08451	1.09586	0.2542
13	Public.Order/ Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-2.53849	1.14124	0.3406
14	Theft.From.the.Person// Land.Area.in.Hectares	Population/Land.Area.in.Hectares	-3.13702	1.07455	0.2621

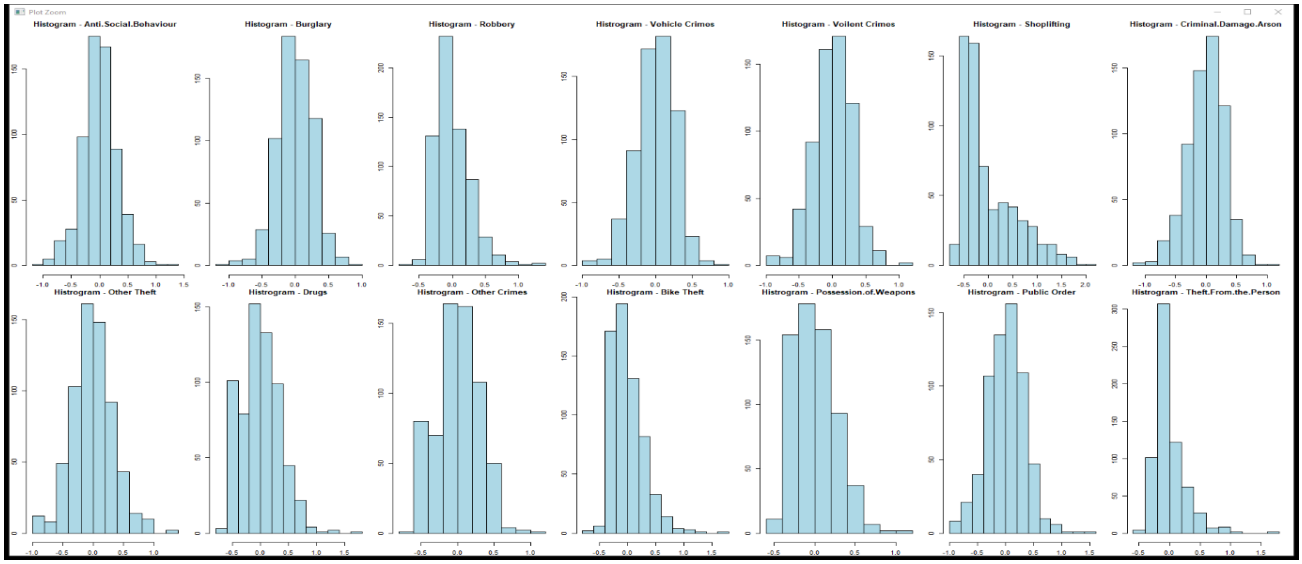
Comment:
Using the first four models above correspond to the different types of crime. The four models show a strong positive relationship between population density and crime rate, as evidenced by the estimated coefficients of the predictor variables being significantly greater than zero (p-value < 2.2e-16).
The R-squared value of the model fit is also high, shows that the predictor variables explain much of the variability in the response variable.

Assumption 1: linearity:
Linearity is an important step in many statistical analyses, especially regression analyses. Linearity means that the relationship between two variables is linear. In other words, the relationship can be represented by a straight line.



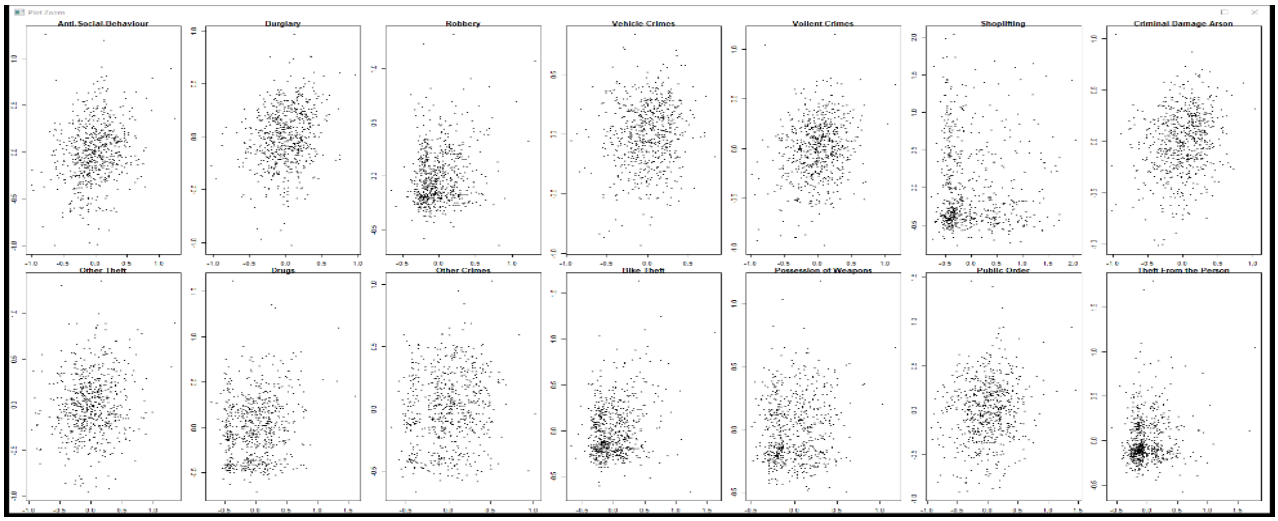
Comment: Most of the points in the plots fall along the straight line to some extent, not of all of them. It shows that there is some linear relationship between variables.

Normality Distributions
A normal distribution plot provides important information about the distribution of residuals in a linear regression model and helps you better understand the underlying patterns and potential limitations of your model.



Comment: The normality distribution plot of the linear regression shows that the residuals are not normally distributed. However, it is worth noting that antisocial behaviour, burglary, and motor vehicle crime variables show slightly symmetrical distributions.
that tell us, there may be a linear relationship between these variables, but the non-normality of the overall distribution could be due to other factors such as outliers, sampling bias, or measurement error.

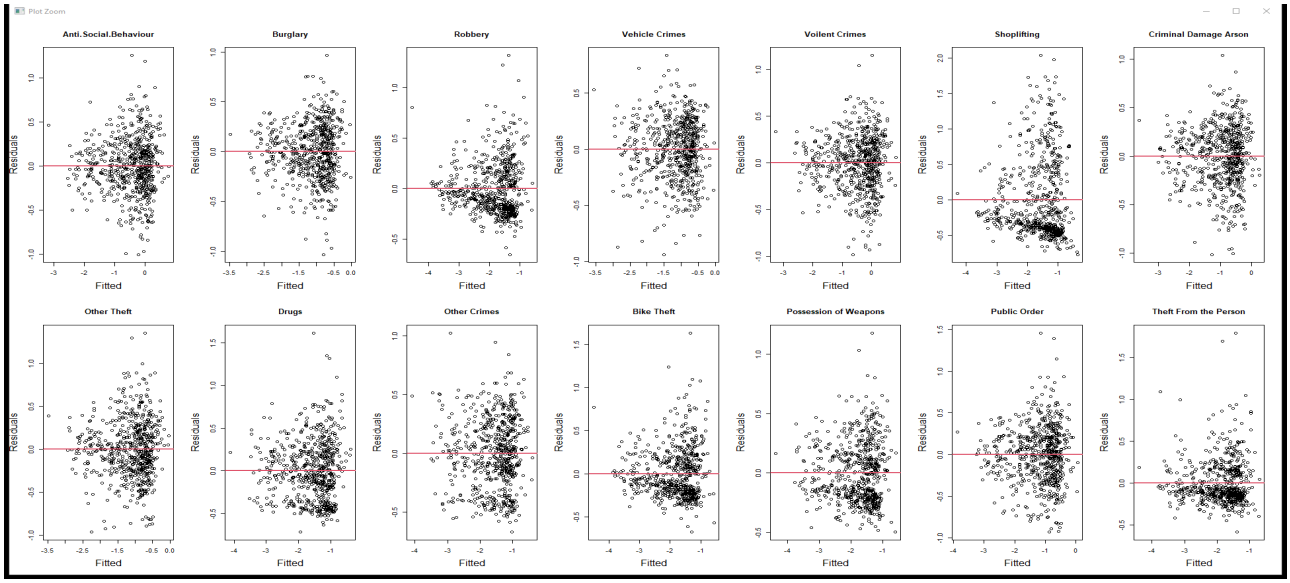
Independence.
Checking of independence is an important assumption in statistical analysis and helps ensure that results are accurate and reliable. To confirm that the observations are not related to each other in a systematic way. If there is a relationship or dependency between observations



Comment: The independence plot shows that the residuals are randomly distributed around the zero line, indicating that the residuals have no pattern, and the independence assumption is met. However, bike theft and shoplifting violate the error of independence assumption.

Constant Variance

Homoscedasticity is referring to the assumption that the variance of the errors in the regression model is constant at all levels of the predictor variable, which means that the spread of the residuals is the same for all levels of the predictor variable.



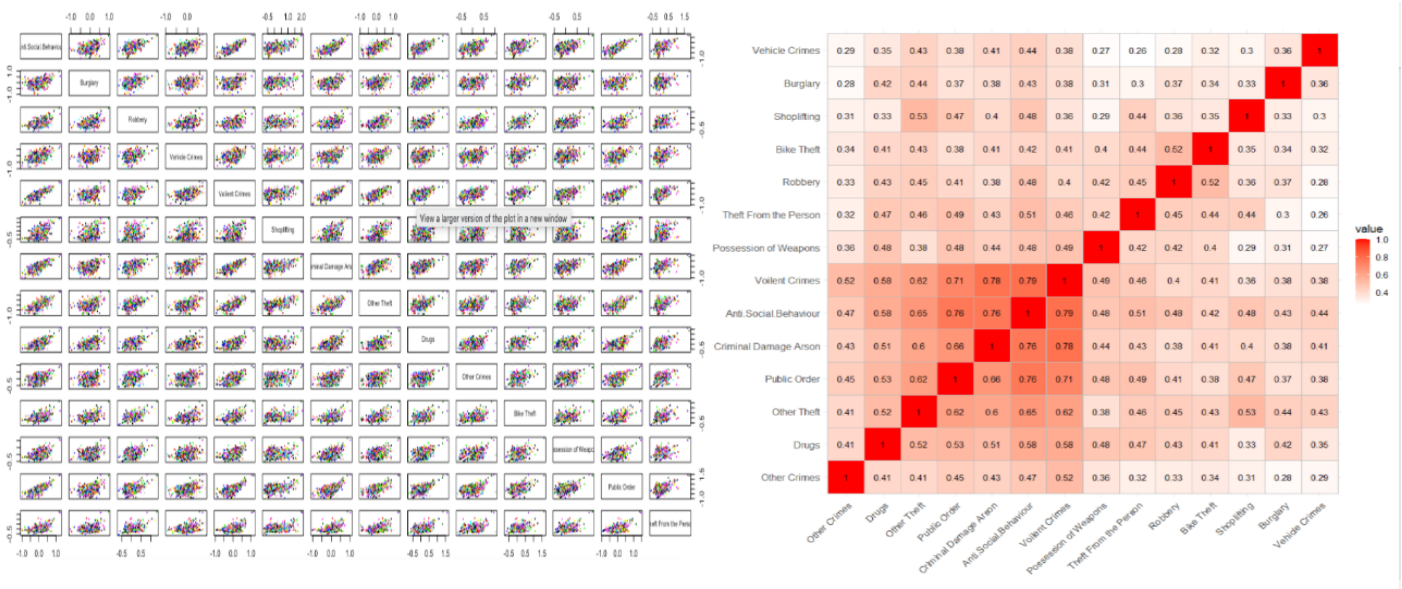
Comment:

Looking at the scatterplot above, it shows the fitted value of the model vs the residuals of those fitted value, It shows that most plots are homoscedasticity due to the random pattern with no significant trends. There are few plots like (Anti. Social. Behaviour) that indicate heteroscedasticity is present, due the the way the residual spread out increasing and decreasing as the predictor variable increase, The cone/funnel shape is a sign of heteroscedasticity.

Part 2b.

Residual Data frame: All the residual scatterplots are compiled in a Dataframe "crimedataresidual" that contains the residuals (the differences between the predicted values and the actual values) of 14 linear regression models. Each column in the data frame represents a different crime indicator, and the rows represent the residuals for each observation in the dataset. Separate colors were assigned to each column in the data frame, as shown below.

<u>Correlation Plots: Residual Data frame/Heat Map</u>

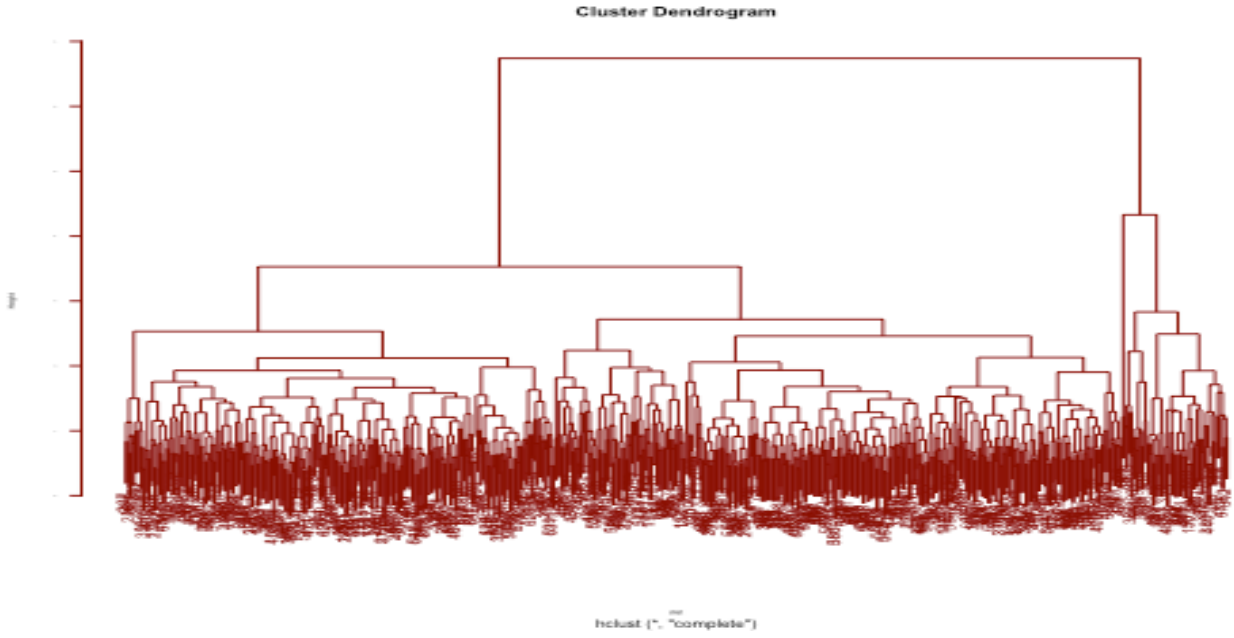


Comment: The scatter plots and correlation plot show the relationships between the residuals of the variables in the dataset. The correlation coefficients indicate the strength and direction of these relationships. there seems to be a strong correlation between the predicted value and the actual value for the following types of crimes:
Anti-social crime, theft, robbery, vehicle crime, violent crime, shoplifting, vandalism and arson, other theft, drugs, other crime, burglary bicycles, possession of weapons, public order and theft.

- Violent crime and anti.social.behaviour appear to have much stronger positive correlation.

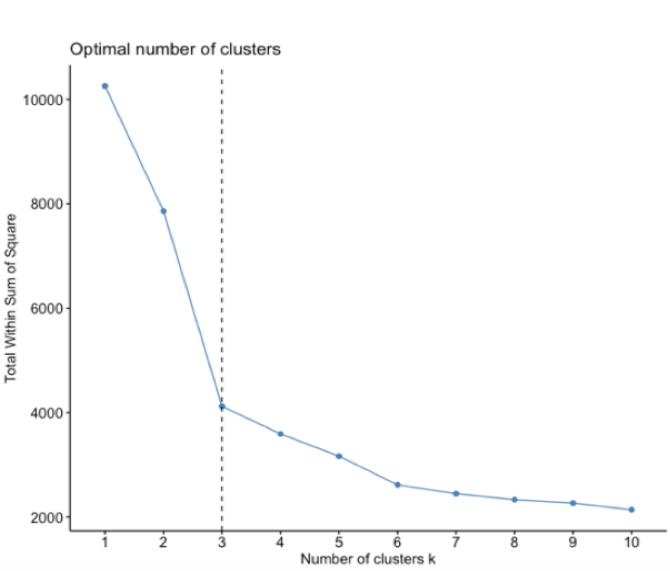
A correlation value of 1 indicates a perfect positive correlation between the predicted value and the actual value. This means that as the predicted values increase, the actual values also increase proportionally. A correlation of 1 is a desirable outcome of the predictive model because the strong correlations shown in the heatmap are a positive indicator of the accuracy of the model's predictions for these crime categories. While 0.4 indicate the weak correlation.

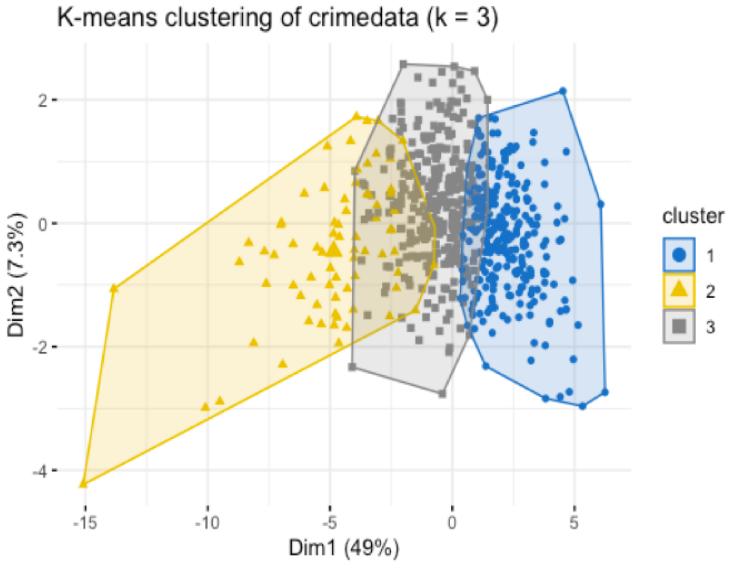
Hierarchical Clustering



Based on the visualize inspection of dendrogram above, it shows that the dendrogram has quite number of outlier, about 10 outliers, which means they are distant away from the rest of the data point.
A dendrogram is structured as a tree-like diagram, with data points represented by leaves and branches that show the joining of similar clusters.

K. Means Clustering:
To perform k-means clustering on this dataset, first had to pre-process the data, Remove the columns 'LSOA' and 'Name', as they are not relevant for clustering. To normalize the remaining columns using the scale() function.





The K-Means clustering algorithm identified **three (3)** distinct clusters**:**

Elbow plot - Comment:
Elbow Plot shows that there is a sharp bend after the **third cluster**. This indicates that the fourth cluster does not add much information to the clustering, and that three clusters is the optimal number.

The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

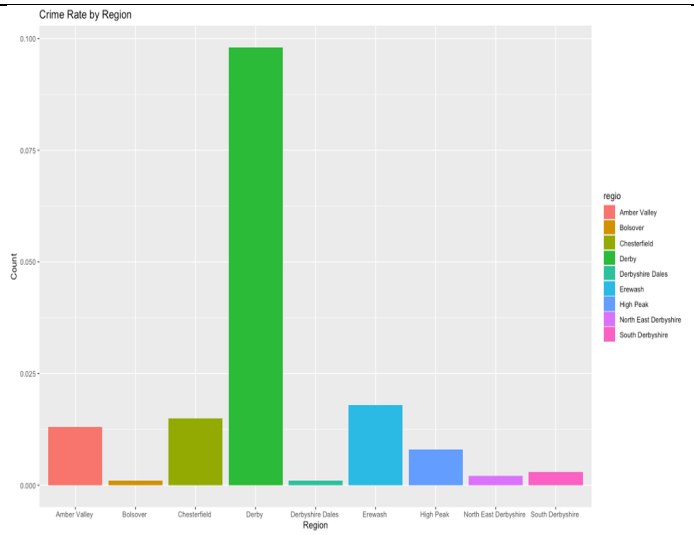
Cluster Analysis:
Cluster analysis is a type of unsupervised learning that groups data points together based on their similarities.
The optimal number of clusters for a data set is typically determined using a combination of methods, such as the elbow method and the silhouette coefficient.

The K-Means clustering algorithm identified three (3) distinct clusters:

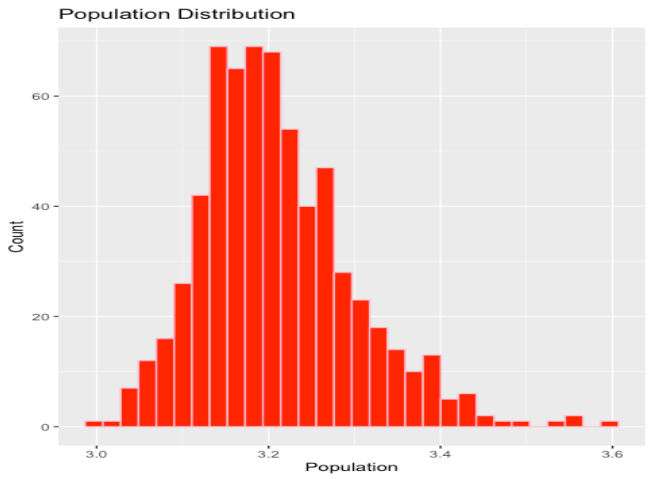
- Examining the clusters reveals that there are similarities between all the clusters, suggesting that they share certain traits or attributes. The clusters graph show us that the data points within each cluster tend to be more similar to each other than to those in other clusters.

Data Interrogation and Insightful Visualizations

(a) The visual interrogation and analysis of crime data using crime density and a bar plot provide actionable insights for addressing crime effectively.

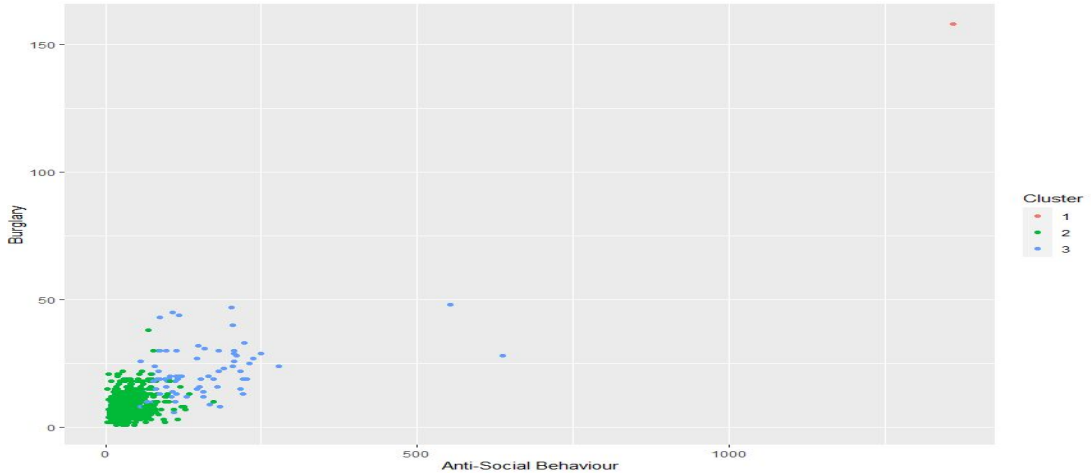


Bar plot above; represent the crime density in the region under consideration and it appears that derby has highest crime rate. By calculating crime density and creating a bar plot, it becomes possible to identify regions with higher crime rates and prioritize them for intervention.



Here is the distribution of population, which appears to be Right-
Skewed, the majority of the population found to between 3.0 to 3.4, with some outliers.

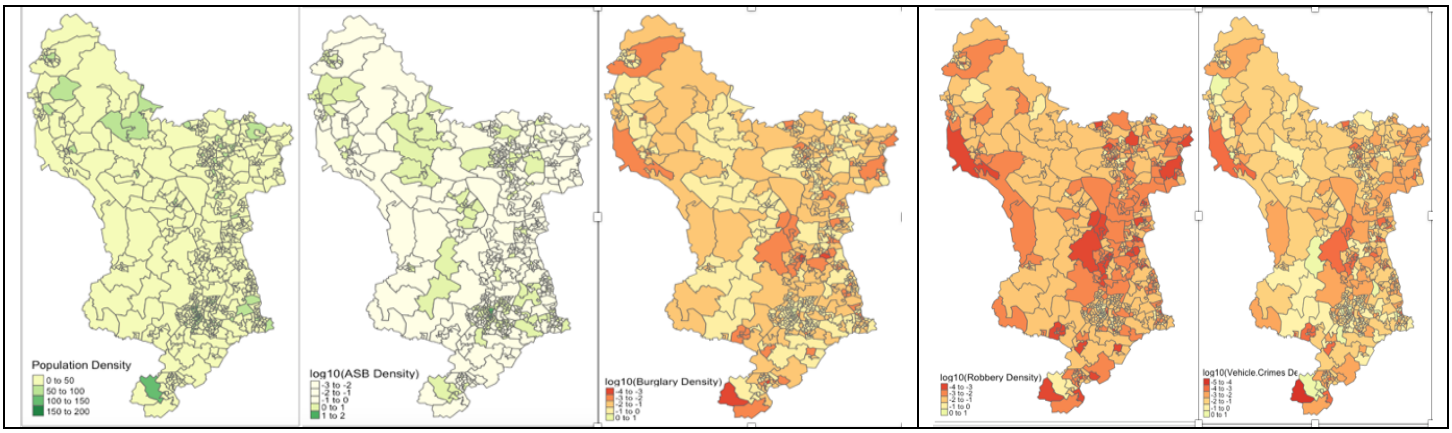
(b) Visualize the clustering results using a scatter plot



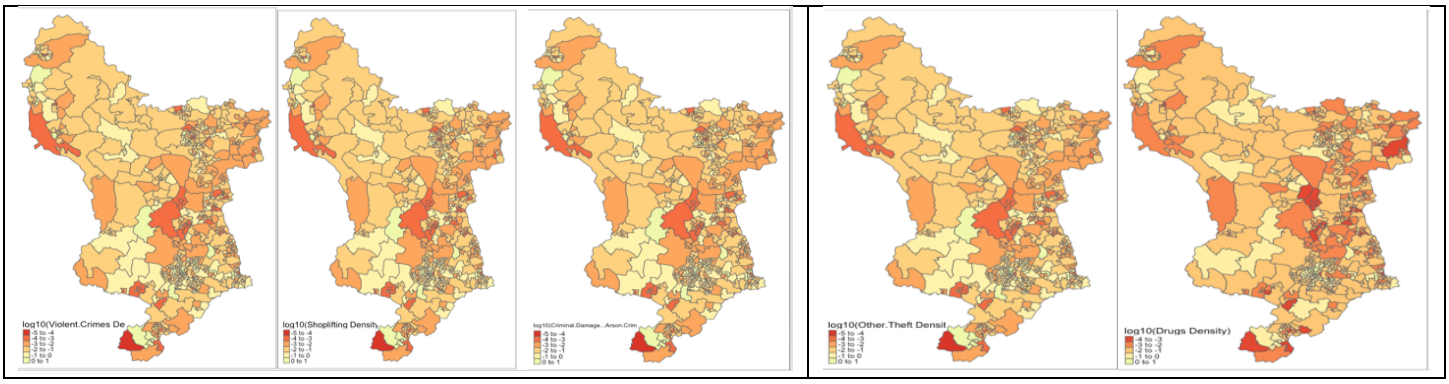
- The antisocial behavior and robbery scatterplots, with each point color-coded according to cluster affiliation, show that the clusters are not perfectly separated. The presence of points of overlap suggests that there is some similarity or overlap between antisocial behavior and robbery clusters.

Geo mapping

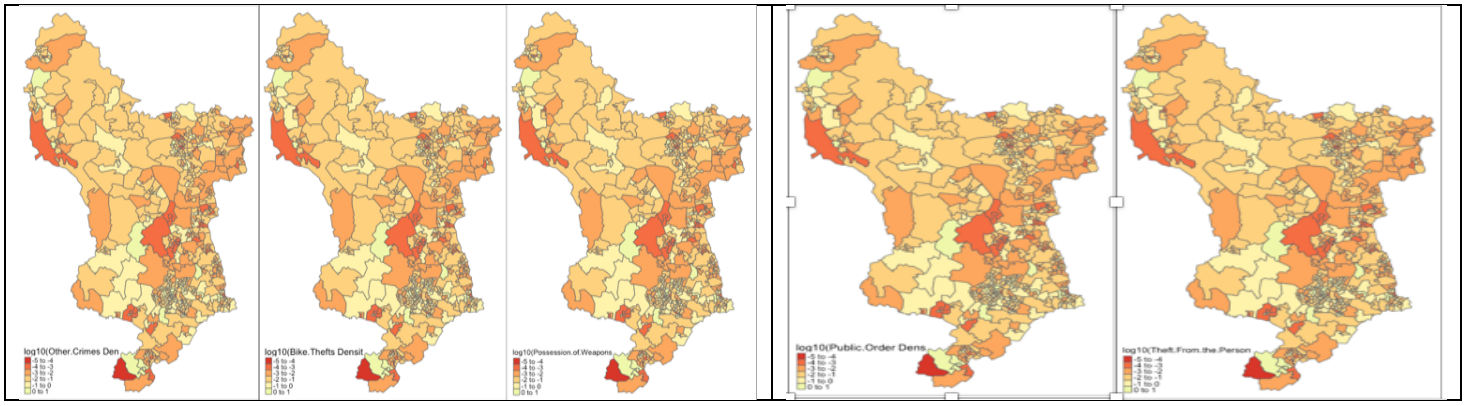
Geo-mapping gives us valuable insights about the distribution and patterns of reported crimes in Derbyshire.



After analyzing the map below, it shows that there are noticeable patterns and similarities among the crimes reported in Derbyshire. The crimes include Anti-Social Behaviour, Burglary, Robbery, Vehicle Crimes, Violent Crimes, Shoplifting, Criminal Damage and Arson, Other Theft, Drugs, Other Crimes, Bike Theft, Possession of Weapons, Public Order, and Theft From the Person



By utilizing geo mapping, we gain a comprehensive understanding of crime dynamics in Derbyshire. This enables us to identify high-risk areas, strategically allocate resources, and adjust prevention strategies accordingly. The insights gained from analyzing crime patterns support collaboration among law enforcement agencies, policy makers and community organizations to drive effective crime reduction responses. With geo mapping, we can proactively address crime issues, enhance public safety, and create a safer environment for the residents of Derbyshire.



Additionally, geo-mapping allows us to monitor the effectiveness of implemented interventions by tracking changes in crime patterns over time. By harnessing the power of geo mapping, we can work towards creating a safer and more secure community in Derbyshire.

3b

Ethics of Analysis

Ethical issues related to analysis encompass a wide range of considerations that arise throughout the process of conducting data analysis. These ethical issues can occur at various stages, from data collection and storage to the interpretation and communication of results. Here are some comprehensive ethical issues related to analysis:

- Obtaining informed consent from participants or data subjects is crucial, especially when their data is collected for analysis. Participants should be fully informed about the purpose, risks, and potential uses of their data and provide voluntary consent. Transparency and clarity in explaining how their data will be used are essential.

- Privacy and Confidentiality: Safeguarding the privacy and confidentiality of individuals and organizations is of utmost importance. Data should be collected, stored, and shared securely, ensuring that personal identifiers are removed or appropriately anonymized. Adequate measures should be in place to protect sensitive or confidential information.

- Data Integrity and Quality: Maintaining data integrity and ensuring the accuracy and reliability of the analysis are ethical responsibilities. Analyses should be conducted using sound statistical methods and best practices. Proper data validation, verification, and cleaning procedures should be implemented to minimize errors and biases.

- Avoiding Bias and Discrimination: Ethical analysis requires addressing potential biases and avoiding discrimination. Analysts should be aware of biases that may exist in data collection or analysis techniques and take steps to mitigate their impact. Analysis should be conducted in a fair, objective, and impartial manner, without favoring any particular group or perpetuating stereotypes or discrimination.

- Responsible Use of Results: Ethical analysis involves considering the potential impacts and consequences of the results. Analysts should use the results responsibly, considering the broader social, economic, and ethical implications. Ensuring that the analysis contributes positively to decision-making, policy formulation, or societal understanding is crucial.

- Transparency and Reproducibility: Transparency in analysis involves clearly documenting and explaining the methods, assumptions, and steps involved in the analysis process. Providing access to the data, analysis code, and documentation allows for scrutiny, reproducibility, and verification by peers. Transparency promotes trust and accountability in the analysis.

- Reporting and Communication: Ethical analysis requires responsible reporting and communication of results. Findings should be presented accurately and without misrepresentation. Providing appropriate context, acknowledging limitations and uncertainties, and avoiding sensationalism or exaggeration are essential to ensure the proper understanding and interpretation of the results.

Addressing all ethical issues promotes the responsible, transparent, and unbiased use of data analysis. By upholding ethical standards, analysts can contribute to the trustworthiness, reliability, and positive impact of their work on individuals, communities, and society as a whole.

Conclusion

In conclusion, the visualization of derby crime data set has yielded significant insights into crime patterns, enabling targeted actions to address and reduce crime effectively. Through visual interrogation techniques, I gained a comprehensive understanding of crime distribution and identified regions with higher crime rates.

The insights derived from the analysis have tangible implications. It will enable Law enforcement agencies to prioritize their efforts by focusing on high-crime areas, leading to more efficient patrols and investigations. Policymakers can develop targeted policies and initiatives to tackle the root causes of crime, promoting long-term solutions.

Utilizing the power of data visualization, this analysis has contributed to evidence-based decision-making, it will provide the community better understanstanding of the crime in the entire city Derbyshire.