**Introduction**

In the modern world, data is being produced at a rate that has never been seen before, making it more and more challenging to make sense of it all. Data visualisation can help with this. We may learn important lessons from data that could otherwise be concealed in rows and columns of raw data by presenting it in a visually appealing and simple-to-understand style.

In this project, I'll be utilising the R programming language to investigate various data visualisation strategies. By experimenting with various visualisation formats, such as charts, graphs, maps, and infographics, I hope to produce meaningful and effective visualisations that tell a narrative and convey key information.

I will have a firm grasp on how to efficiently visualise data and convey conclusions to others by the project's conclusion. So let's get started and explore the potential of data visualisation!

**Dataset Description**

Derbyshire is among the top 20 most dangerous counties in England, Wales, and Northern Ireland. The overall crime rate in Derbyshire in 2021 was 81 crimes per 1,000 people, and the most common crimes were violence and sexual offences, which happened to roughly every 39 out of 1,000 residents. This dataset reflects reported incidents of crime that occurred in Derbyshire and its environments covering 315 regions in the year 2019.

It is made up of 18 fields and 642 records while the first and second columns contain unique identifier for each region, comprises of 14 crime indicators matched to population and land area for each region.

**Metadata**

* **LSOA**: Lower Layer Super Output Areas are a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales.
* **Name:** Regions within Derbyshire associated with the crimes.
* **Population:** The number of people leaving in the region.
* **Land.Area.in.Hectares:** Primarily used in the measurement of the land area of the crime scene
* **Anti.Social.Behaviour:** personality disorder characterised by impulsive, irresponsible and often criminal behavior.
* **Burglary:** anillegal entry of a building with intent to commit a crime, especially theft.
* **Robbery:** is a crime of theft that involves force or intimidation against a person or property
* **Vehicle.Crimes:** refers to the**theft and trafficking of vehicles and the illicit trade in spare parts.**
* **Violent.Crimes: crime in which the offender uses or threatens to use violent force upon the victim**
* **Shoplifting:** the action of stealing goods from a shop while pretending to be a customer
* **Arson:** the criminal act of deliberately setting fire to property
* **Other.Theft:** Other forms of stealing.
* **Drugs:** these are crime involving illicit use or transfer of hard drug
* **Other.Crimes:** These are other recorded but unspecified crimes
* **Bike.Theft:** This is the crime of stealing bicycle
* **Possession.of.weapon:** This is Threatening with an offensive weapon in public.
* **Public.Order:** these crimes are acts that deviate from or conflict with society's general ideas of normal social behavior, moral values, and public opinion

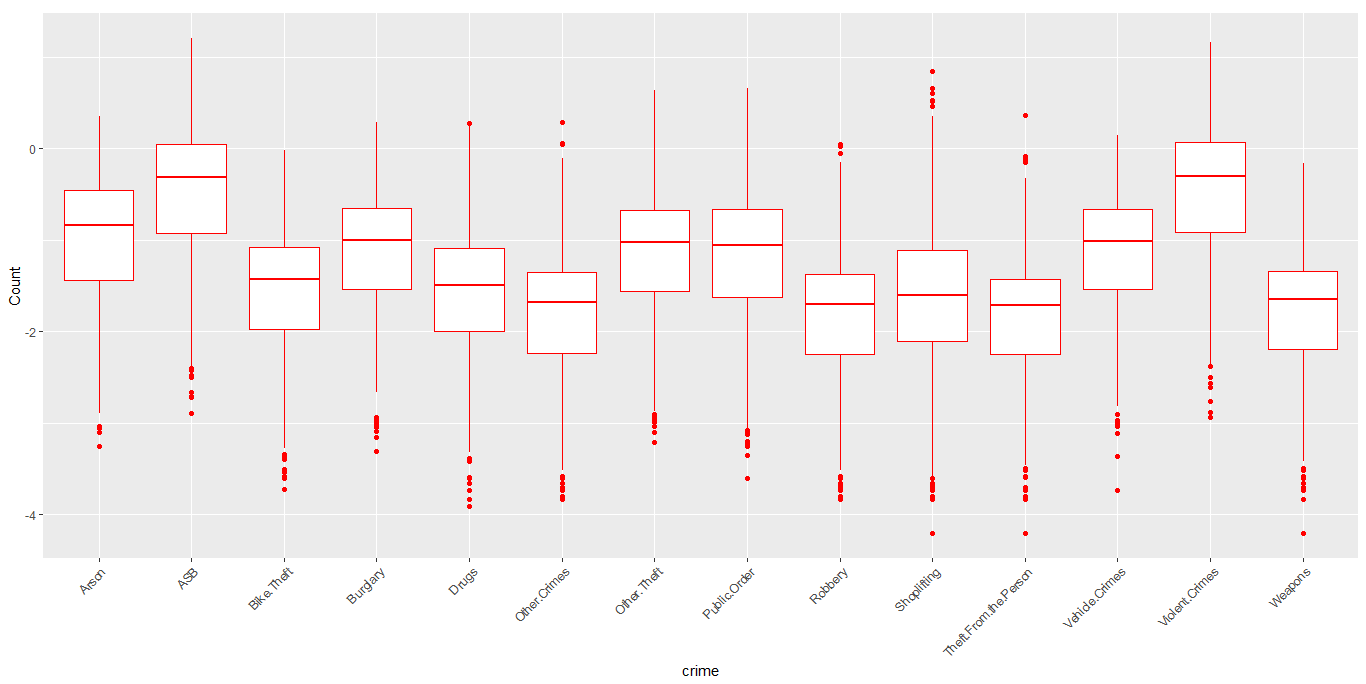
***Statistical Summary of Numerical Variables***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Mean | Median | SD | Skew | Min | Max | Range | 1st Quartile | 3rd Quartile |
| Population | 1657.32 | 1584 | 374.14 | 1.71 | 993 | 3948 | 2955 | 1411 | 1815 |
| Land.Area.in.Hectares | 408.84 | 71.56 | 1115.23 | 6.82 | 12.81 | 16227.09 | 16214.28 | 37.26 | 209.55 |
| Anti.Social.Behaviour | 52.35 | 36 | 73.95 | 10.26 | 3 | 1359 | 1356 | 22 | 58 |
| Burglary | 10.25 | 8 | 9.08 | 7.55 | 1 | 158 | 157 | 5 | 13 |
| Robbery | 2.16 | 1 | 3.94 | 13.27 | 1 | 79 | 78 | 1 | 2 |
| Vehicle.Crimes | 9.34 | 8 | 6.56 | 2.42 | 1 | 60 | 59 | 5 | 12 |
| Violent.Crimes | 49.9 | 35 | 63.81 | 11.86 | 3 | 1280 | 1277 | 22 | 58 |
| Shoplifting | 9.73 | 1 | 32.71 | 11.62 | 1 | 613 | 612 | 1 | 5 |
| CriminalArson | 14.86 | 12 | 13.66 | 5.19 | 1 | 196 | 195 | 7 | 19 |
| Other.Theft | 11.91 | 8 | 19.13 | 12.44 | 1 | 382 | 381 | 5 | 13 |
| Drugs | 4.67 | 3 | 9.17 | 10.64 | 1 | 150 | 149 | 2 | 5 |
| Other.Crimes | 3.83 | 3 | 3.65 | 5.38 | 1 | 46 | 45 | 2 | 5 |
| Bike.Theft | 2.55 | 1 | 7.32 | 19.1 | 1 | 170 | 169 | 1 | 2 |
| Possession.of.Weapon | 2.22 | 2 | 2.89 | 13.29 | 1 | 60 | 59 | 1 | 3 |
| Public.Order | 10.57 | 7 | 20.16 | 13.44 | 1 | 404 | 403 | 4 | 11.75 |
| Theft.From.the.Person | 2.22 | 1 | 8.57 | 20.52 | 1 | 202 | 201 | 1 | 2 |

***Breakdown***

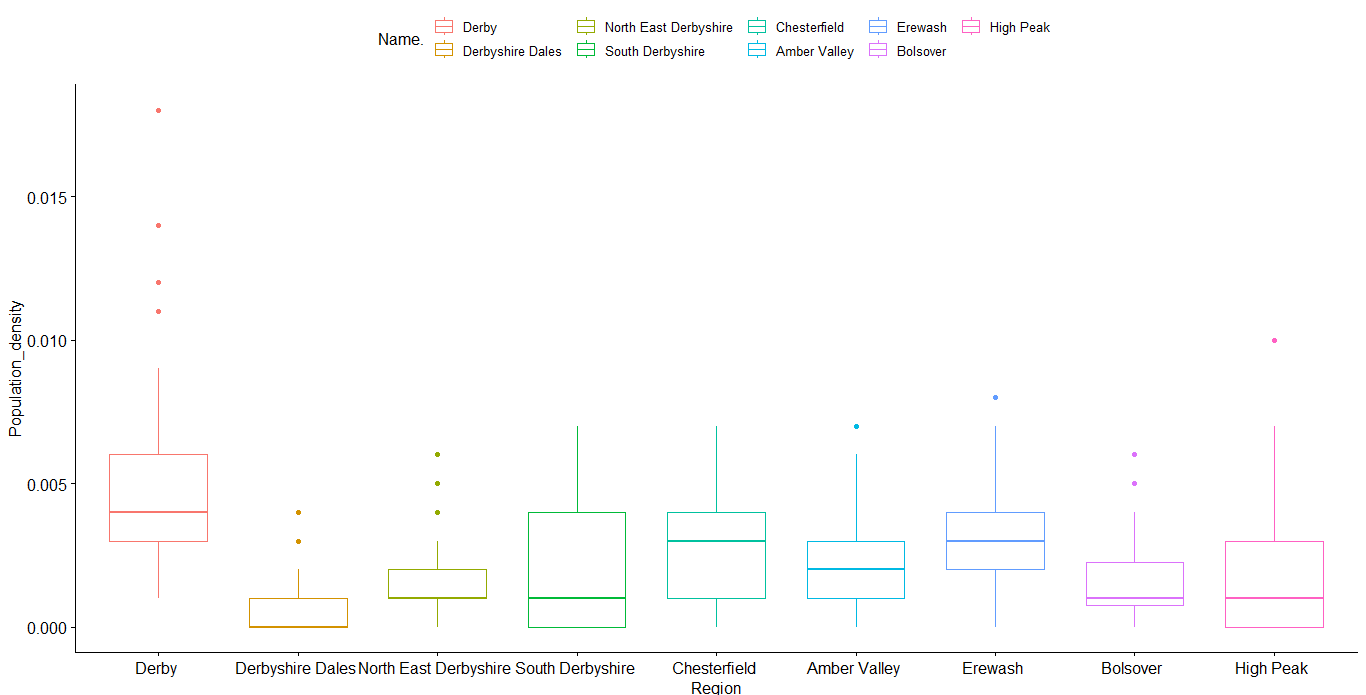
* The LSOA has a lot of variation in population size, land area, and antisocial behaviour, suggesting there may be inequalities.
* The variables "Anti.Social.Behaviour", "Robbery", "Shoplifting", "Drugs", "Bike.Theft", "Possession.Weapons", and "Theft.From.the.Person" have positive skewness values.
* Land.Area.Hectares and Public.Order have positive skewness values of 6.82 and 13.44, suggesting larger land areas and higher public order crimes.
* The variable "Other.Crimes" has a positive skewness value of 5.38, indicating a potential longer tail on the right side of the distribution, suggesting potential areas with higher occurrences of other types of crimes.
* Land areas, incidents of anti-social behaviour, violent crimes, and public order crimes have high standard deviations, suggesting heterogeneity or diversity.
* Theft has lower variability than other variables, suggesting homogeneity or consistency across the LSOA.

***Boxplot of Crimes***



By visually inspecting the dataset, a combined boxplot is created, which simultaneously indicates the 1st quantile, 3rd quantile, interquartile range, the mean, median and the outliers contained in the variable of the individual crimes.

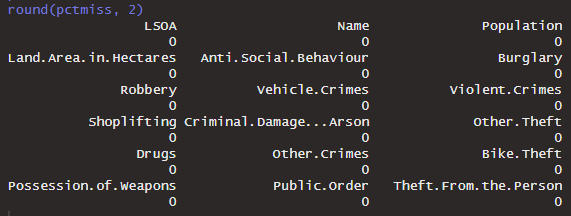
***Boxplot of Regions by Population Density***



The population data for the dataset's nine separate areas is shown in the boxplot above. The population is depicted on the vertical axis, while the regions are represented on the horizontal axis.

***Manipulations made on Dataset***

1. A new column created called **Population\_Density\_Level** from the original **Population** and **Land\_Area** column, then partitioned into two observarions using the mutate function: a region with population density above average is termed **High,** otherwise, **Low.**
2. For ease of exploration, in the **Name** categorical variable, the region name was seperatated from the entire observation which is the combination of both the redion name and the postcode, which is called **New\_name.**
3. Created a new column, named **Crime\_Rate,** which estimates the rate of crime in each region under consideartion.
4. The name variable was later renamed to **LSOA11CD** in order to be used as a point of merging with the shapelife provided.
5. all the numerical variables, aside the **Land Area were** converted to density form for a better analysis and, eg **Population** was converted to **Population Density** By using the formula **(Population/(Land Area \* 10000))**
6. All numerical variables were normalised before analysis in order to minimize redundancy and makes it easy for a model to learn and understand the problem.
7. I visually inspected the presence of zero values and NAs, the table below confirms none.



Linear Regression

Linear regression is a statistical method used to study the relationship between two or more variables, with the aim of predicting the value of one variable based on the values of the other variables. In the context of crime, linear regression was used to model the relationship between crime rates and Population density withing Derbyshire

The disparities between the actual values of the dependent variable and the values predicted by the regression equation are known as residuals in linear regression. The residual for the nth observation is determined mathematically by**: residual = observed - predicted**

where **observed** is the actual value of the dependent variable for the nth observation, and **predicted** is the value of the dependent variable predicted by the regression equation for the same observation.

The method we employ to quantify the connection between one or more predictor variables and a response variable is called linear regression.

***Model estimate table***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Crime* | *1st Coeff* | *2nd Coeff* | *Crime* | *1st Coeff* | *2nd Coeff* |
| Anti Social | -1.82146 | 1.14829 | **Other theft** | -2.30958 | 1.00335 |
| Burglary | -2.30167 | 1.00289 | **Drugs** | -2.84592 | 1.08812 |
| Robbery | -3.18957 | 1.14904 | **Other Crimes** | -3.1588 | 1.14374 |
| Vehicle Crimes | -2.35954 | 1.02457 | **Bike theft** | -2.83232 | 1.07552 |
| Violent Crimes | -1.8708 | 1.18624 | **Possession** | -3.08451 | 1.09586 |
| Shoplifting | -2.95277 | 1.14475 | **Public Order** | -2.53849 | 1.14124 |
| CDA | -2.30629 | 1.12458 | **Theft from per** | -3.13702 | 1.07455 |

It's interesting to see the coefficients of different types of crimes in this table. The negative coefficients for most types of crimes suggest that there is a negative correlation between the occurrence of those crimes and the predictor variable. In other words, as the predictor variable increases, the incidence of those crimes decreases.

***Crimes = 1st Coeff \* Quantity + 2nd Coeff \* Discount***

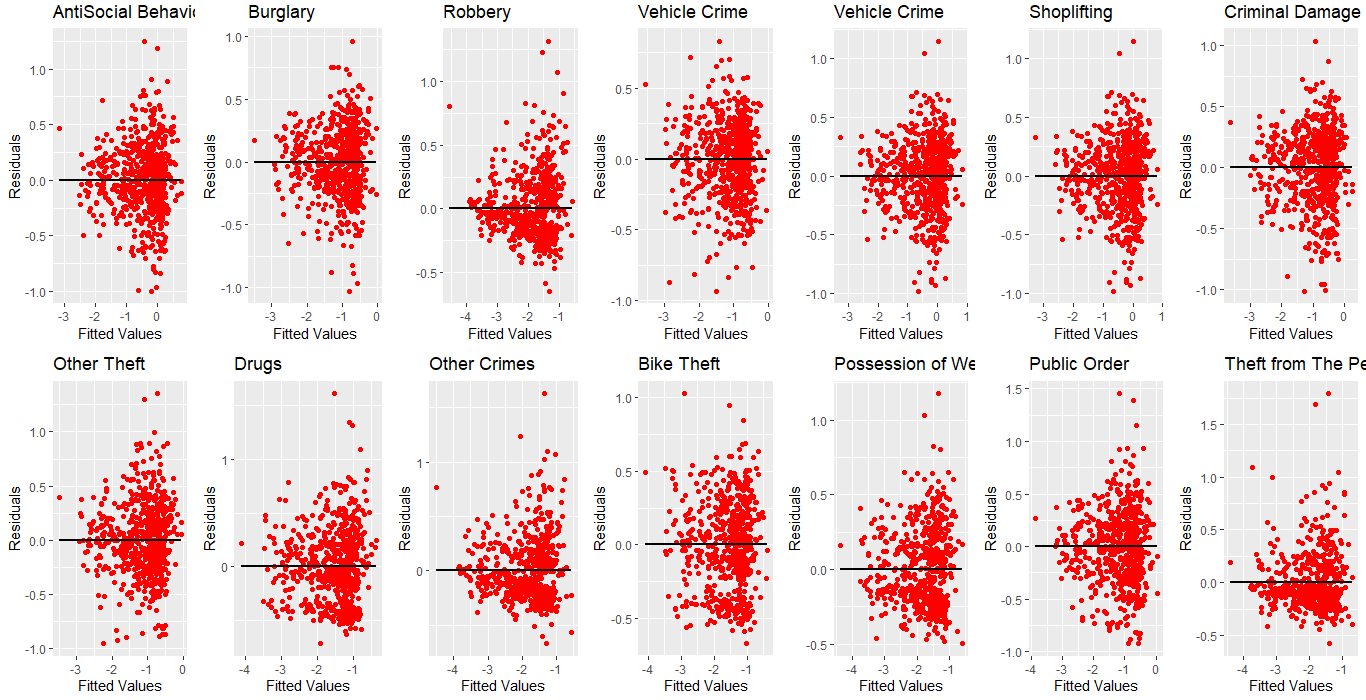
Based on the values of Quantity and Discount, this equation may be used to forecast the value of the crime variables.

***Linear Model Assumptions***

This model comes with some general assumptions, in which three of them will be examined in this work, the assumptions are explained thus:

***Constant Variance***

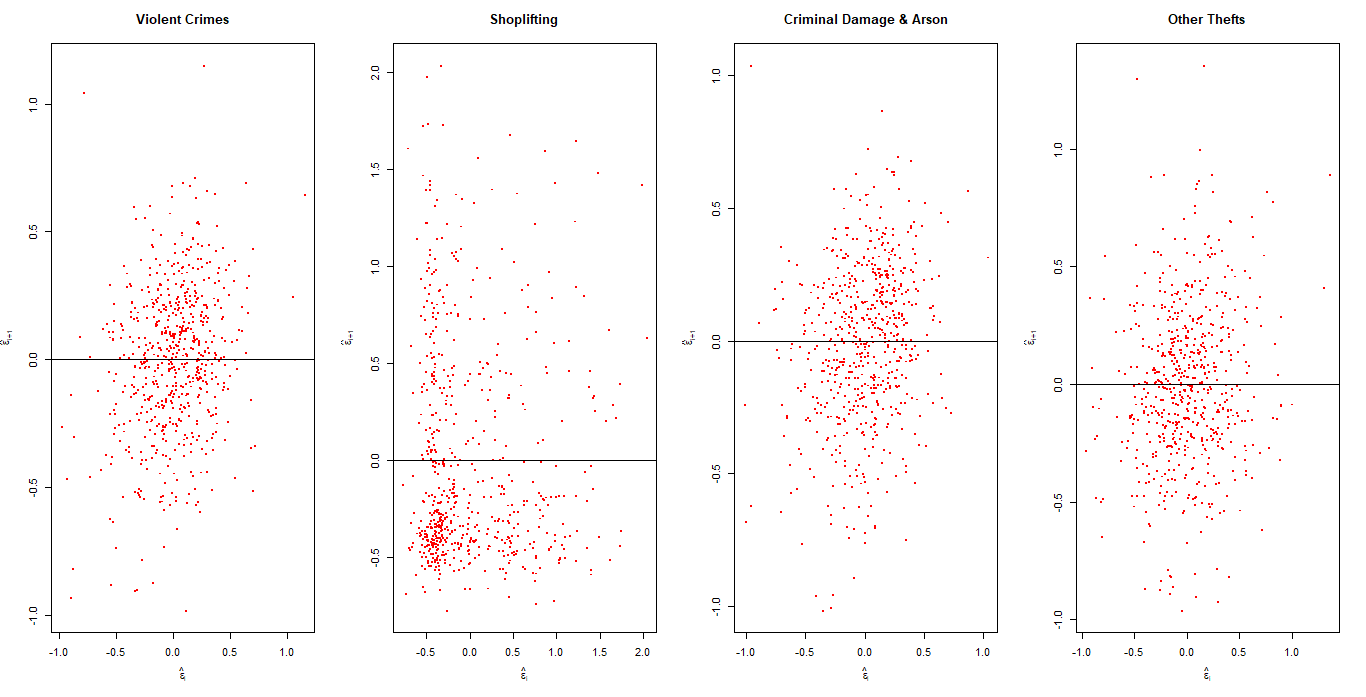
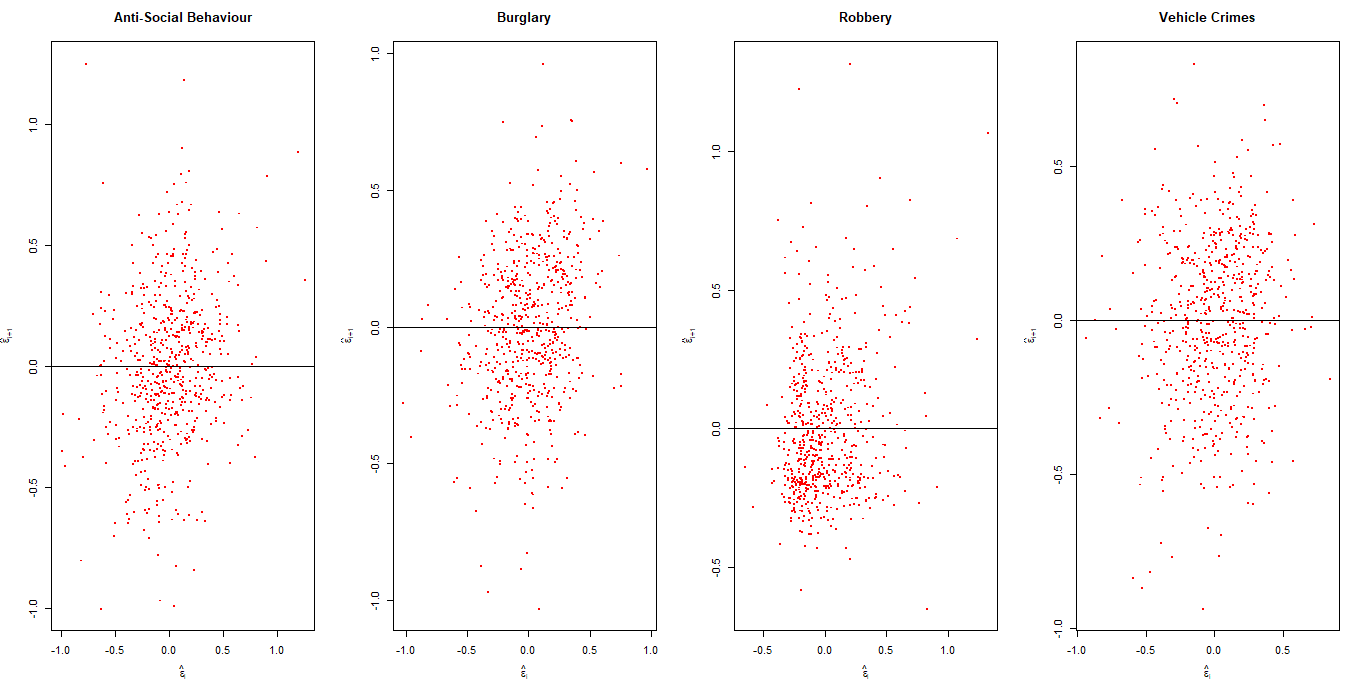
At every level of the predictor variable, it is expected for linear regression that the residuals have a constant variance. The quickest way to find this is to plot the **fitted** value against the **residual**. The plot is as seen below:

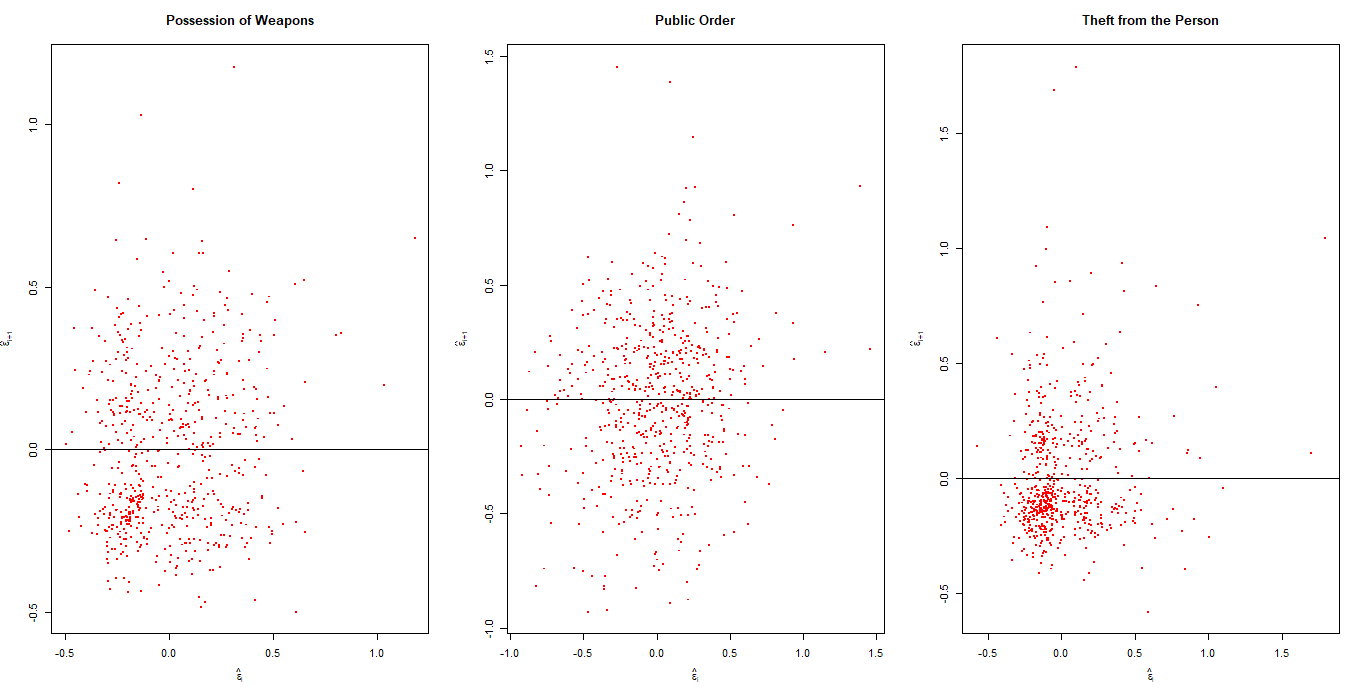
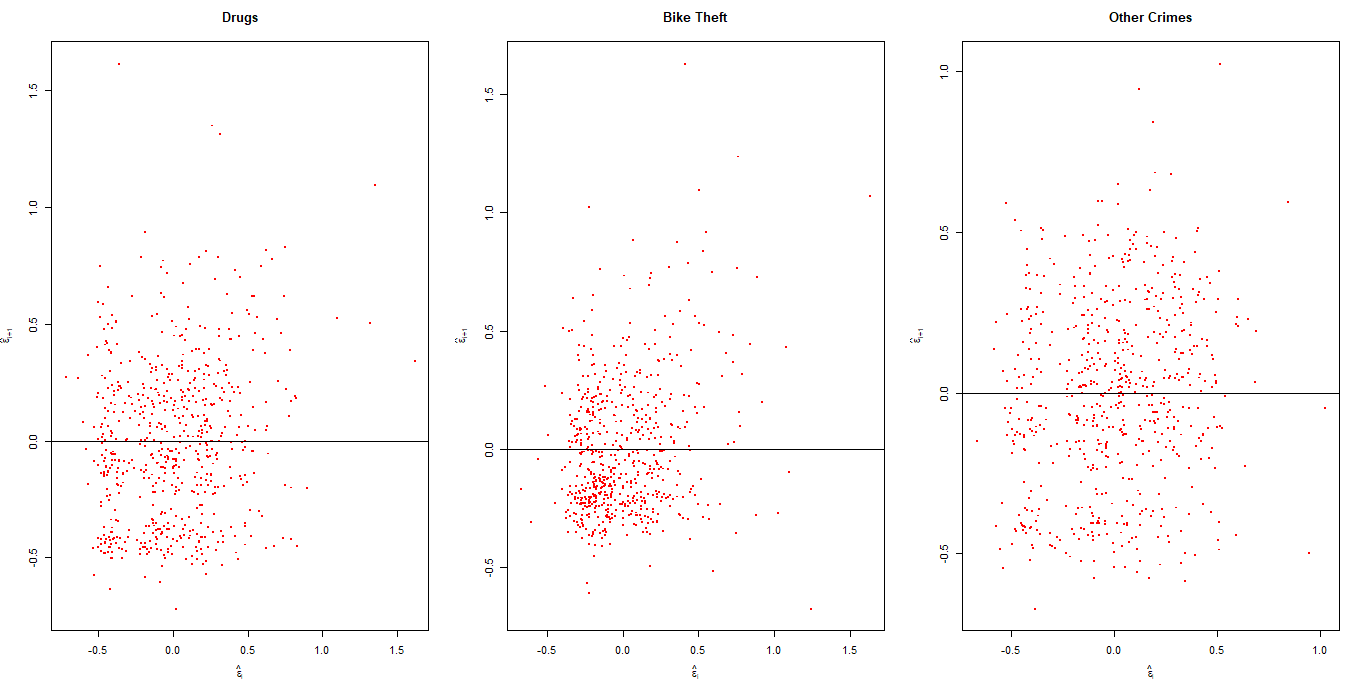


The group of regression models are seen to violate the constant variance assumption such that they exhibit heteroscedasticity, consequently, at all levels of the independent variable, the variance of the errors is not constant. As a result of the uneven weighting of data and potentially skewed estimations of model parameters caused by these models' residuals, which are not evenly distributed around the regression line, the models exhibit this characteristic.

***Independence***

This refers to the circumstance in which there is no correlation between the errors or residuals of the dependent variable. In other words, there shouldn't be any additional unobserved factors influencing the dependent variable; rather, the variation in the dependent variable should only be explained by the independent variables included in the model. The plots are as follow:

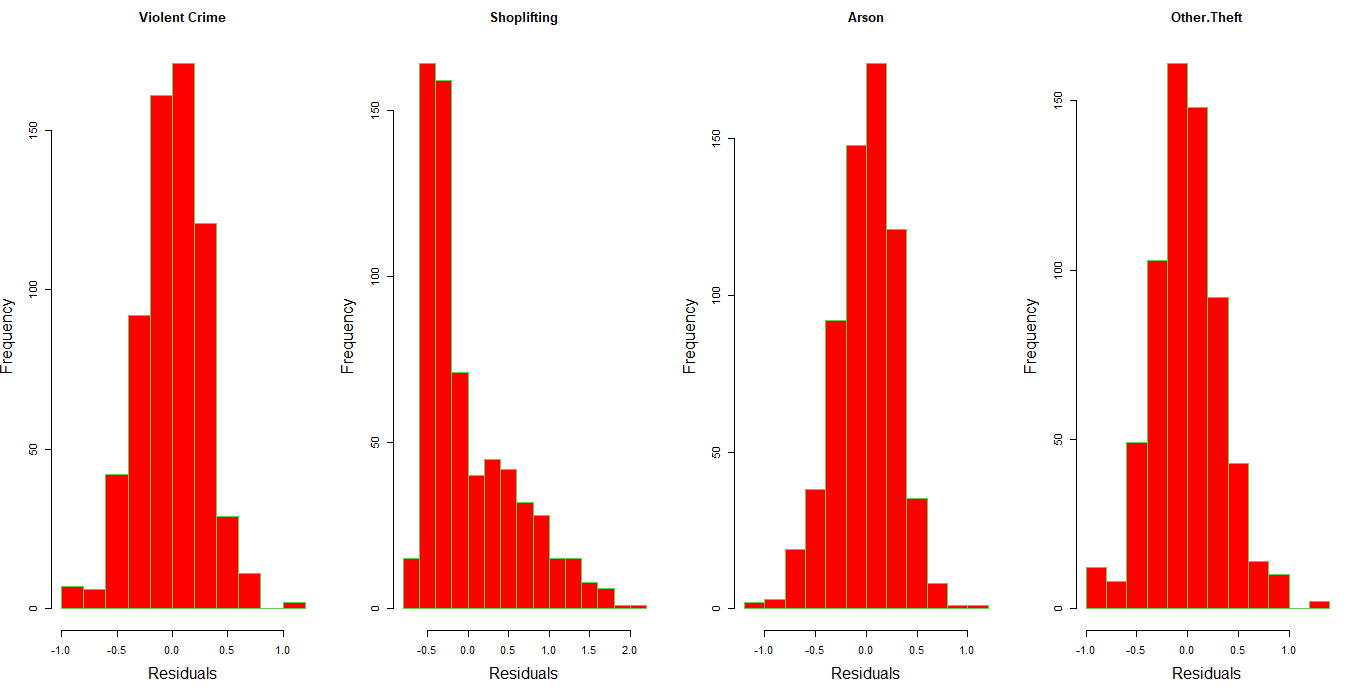
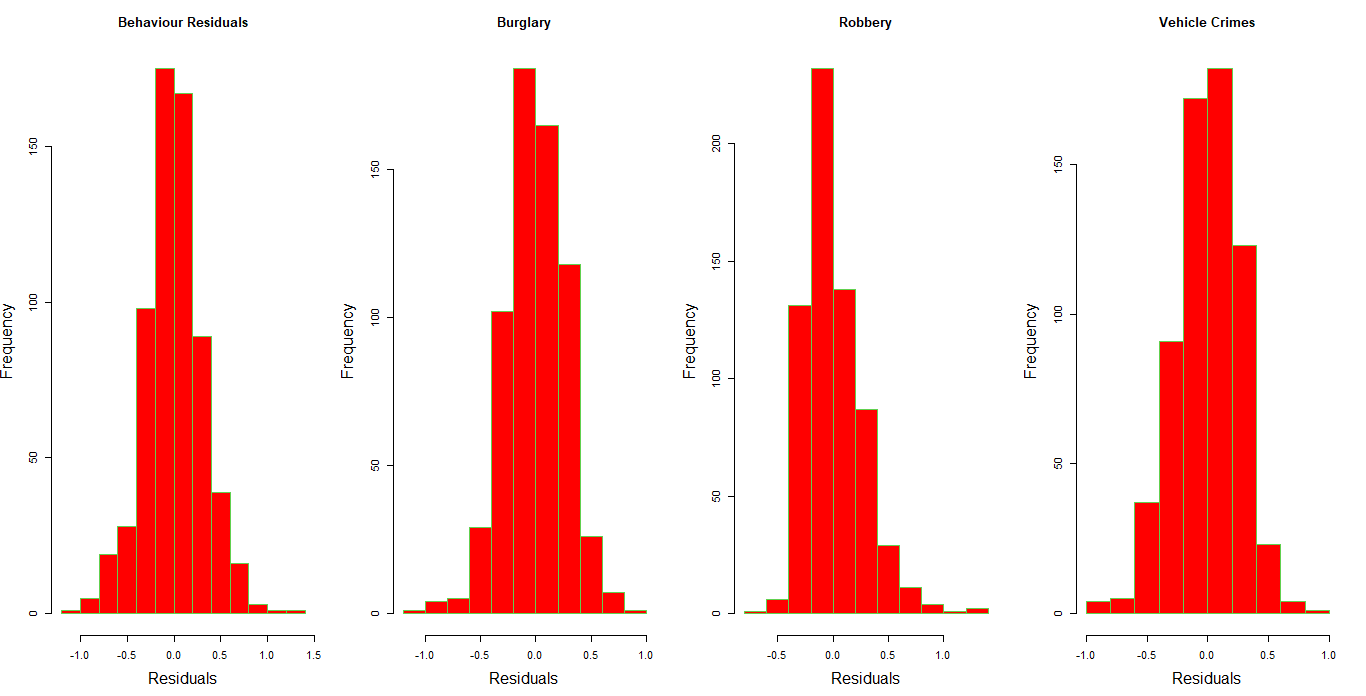


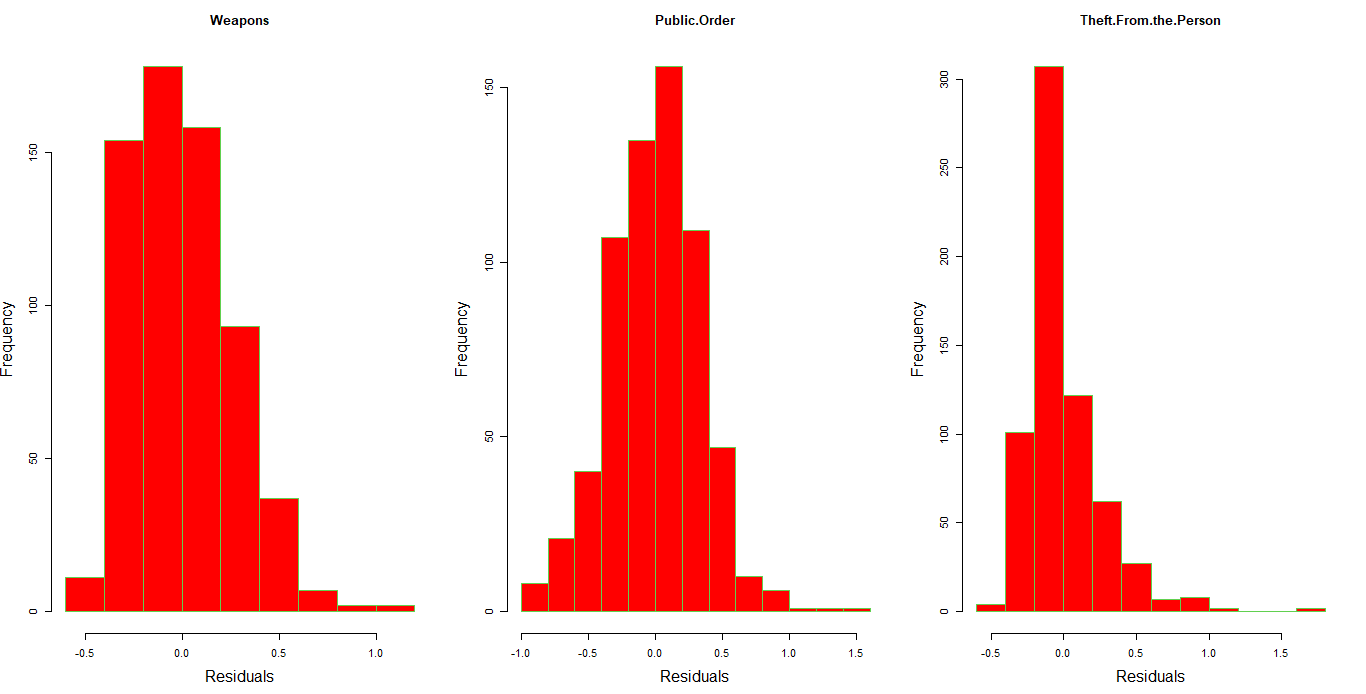
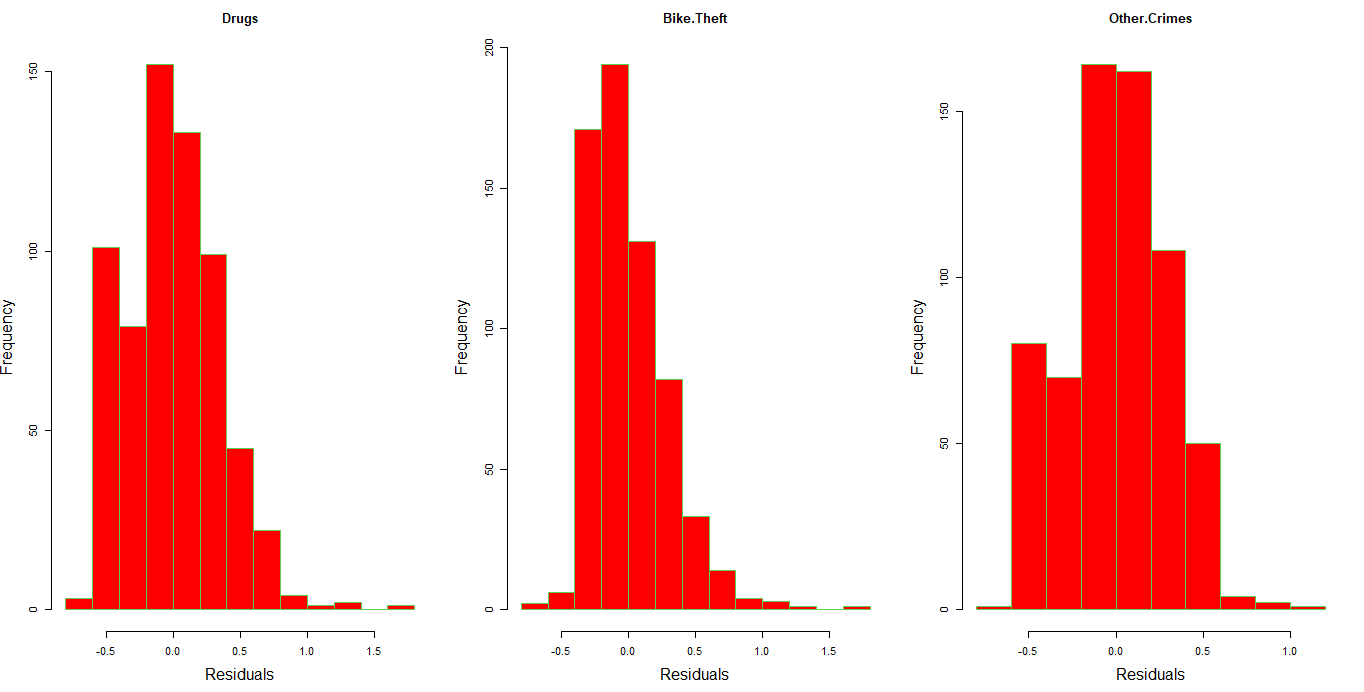


Since the errors (or residuals) of the dependent variable are not correlated with one another, the group of regression models mentioned above **abide** by the independence assumption because the variation in the dependent variable is solely explained by the independent variables included in the model and is not affected by any other unobserved factors. For accurate statistical inference and unbiased estimate of the regression coefficients, this condition is required.

***Normality***

The notion that the residuals (or errors) of the dependent variable would have a normal (Gaussian) distribution is known as the normality assumption in a model. This presumption is crucial because many statistical techniques used in regression analysis, such as confidence intervals and hypothesis testing, depend on the validity of the assumption of normalcy. The plots below are used to check for the models normality.

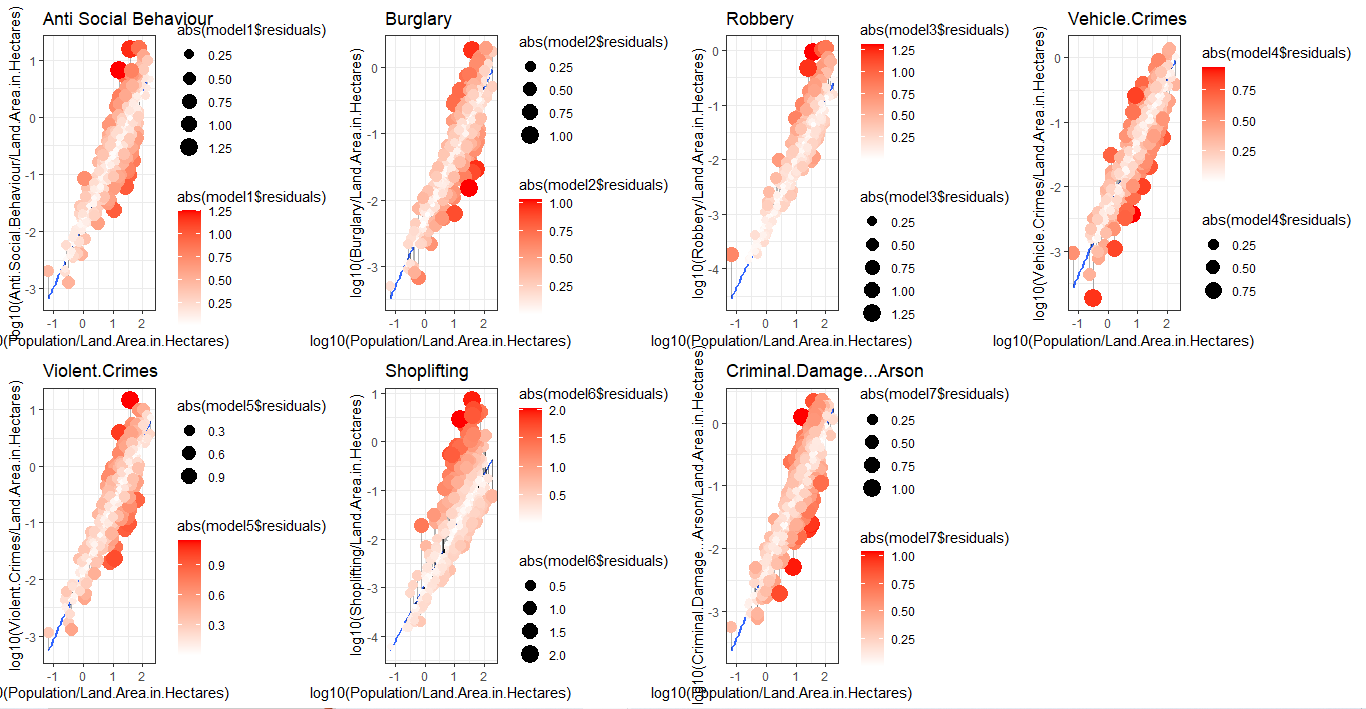




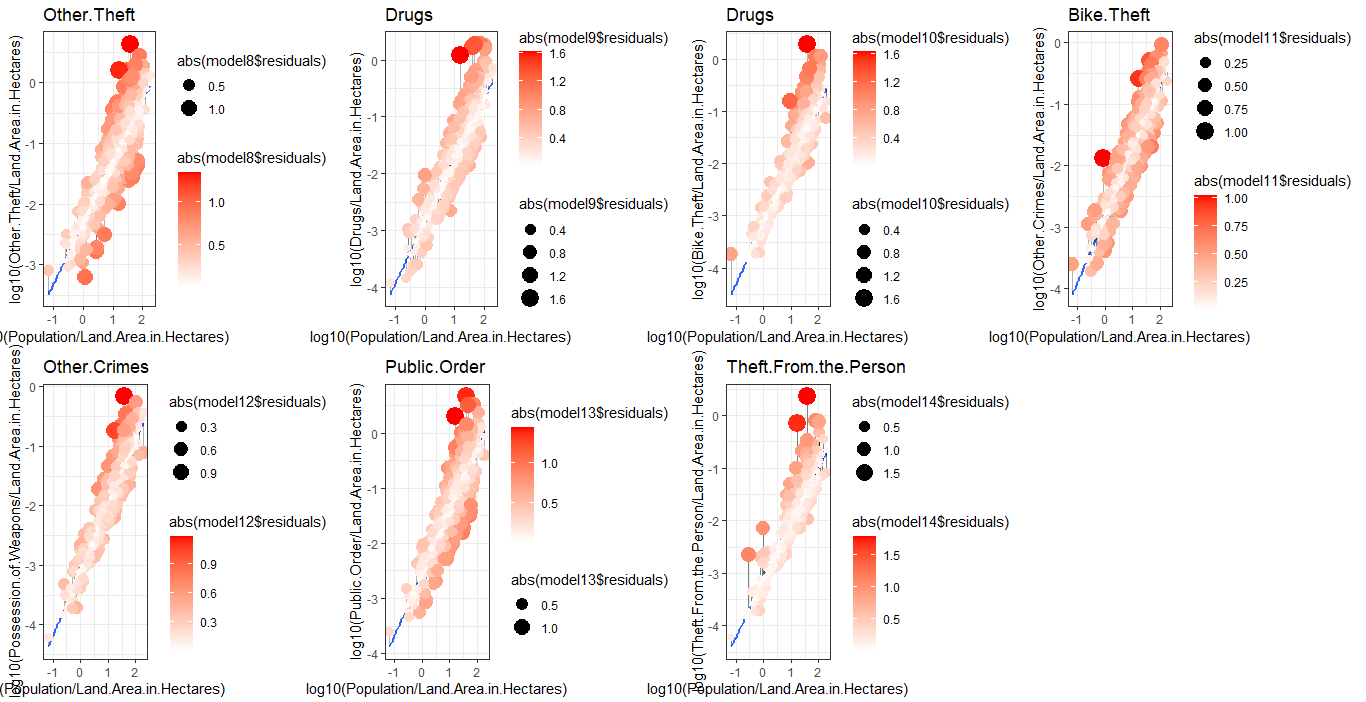
A number of the models do not obey the normality assumptions, while Anti Social Behaviour, Burglary, Other Crimes can be fairly accepted to be normally distributed, meaning they obey the normality assumption.

***The residual plot***

*T*his is a scatterplot of the residuals against the predictor variable, is employed to assess the following linear regression model assumptions:

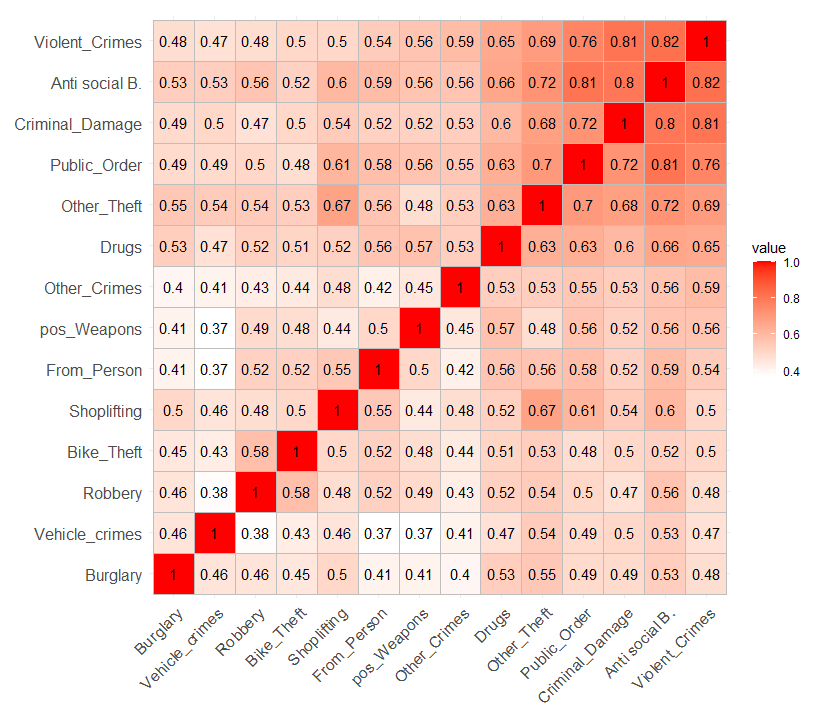


* Linearity: The plot should display a haphazard distribution of dots around the zero line. A nonlinear regression model could be more suitable if there is a clear pattern, such as a curved or U-shaped pattern, which indicates that the connection between the predictor and response variable is not linear.
* Homoscedasticity: The graphic should demonstrate that the residuals' variance is constant for all values of the predictor variable. An alternative error structure, such as weighted least squares regression, may be more appropriate if the variance of the residuals rises or decreases when the predictor variable changes, indicating that the variance is not constant.
* Independence: The residuals should be randomly distributed on the plot, with no patterns indicating that they are connected with, reliant on, or independent of the predictor variable or one another.



If these assumptions are met, the residual plot should show a random scatter of points around the zero line, with no discernible patterns or trends. However, **as observed in majority of the plots seen above**, **assumptions are not met,** hence the residual plot can help to identify the problems and guide further analyses.

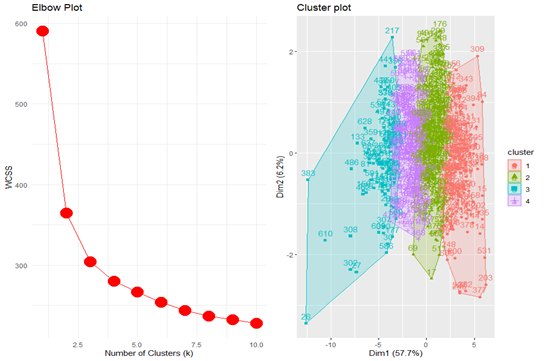
***The Heatmap.***



With the use of heatmaps, which graphically describe data by colouring values, it is simple to see and quickly comprehend complicated data. The residuals with the highest correlations are shown with rich colour, and the colour fades out as the correlation decreases, as we have in this example of regression model residuals.

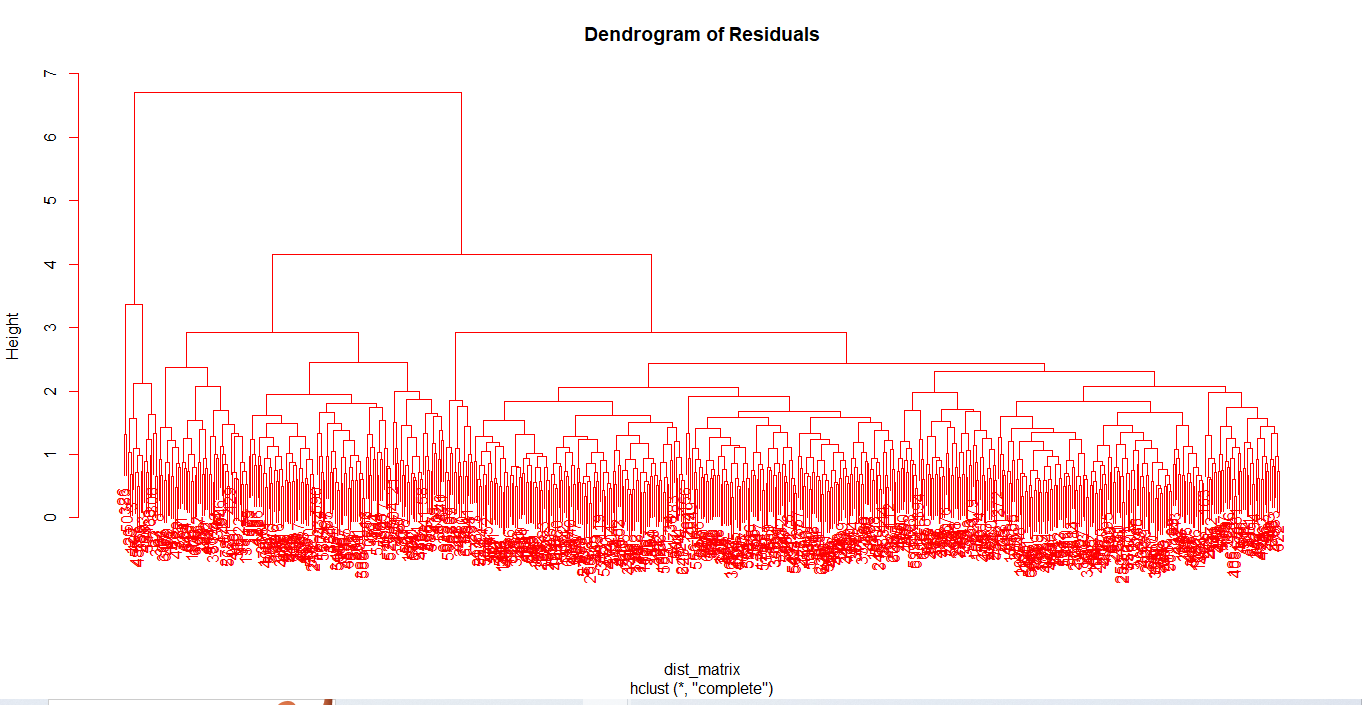
***Heirachical Clustering***

An obvious "elbow point" at **k=4** when the WCSS values start to level out after a sharp decline can be seen on the elbow plot for a k-means clustering analysis with four categories. This shows that clustering the data into 4 clusters would be the best option. In order to evaluate the clustering's quality visually, a cluster plot is also made, where data points are coloured or marked in accordance with the given cluster membership. The cluster plot displays discrete clusters with well separated data points inside each cluster, demonstrating that the data were reasonably grouped into **4 clusters**.



***The Dendogram***

As seen in the dendogram of residuals below, the groupings are closely packed towards the base of the tree due to high number of observations, but has the tree narrows towards the top, the groupings become clearer.



Has being observed from previous plots, the observations have quite a number of outliers, same applies to this dendogram of residuals as 10-15 main outliers can be observed physically, while 42 was gotten when derived using code. This conclusion is reached due to the distance they maintain away from the rest of the tree.

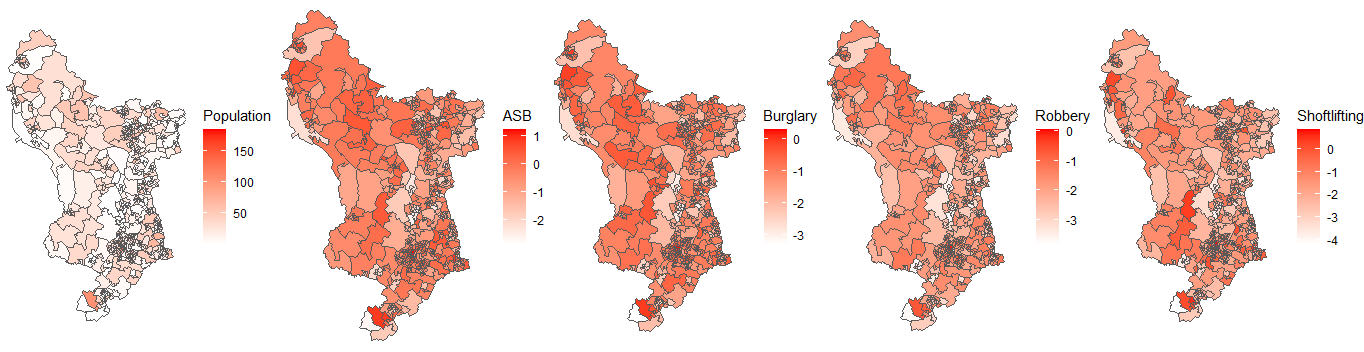
The dendrogram shows the hierarchical clustering of the data points, with each branch representing the similarity between the data points. The outliers can be identified as the individual data points that are farthest away from the rest of the data on the dendrogram. They may be shown as separate branches or as isolated points in the diagram.

Further Investigation On The Dataset

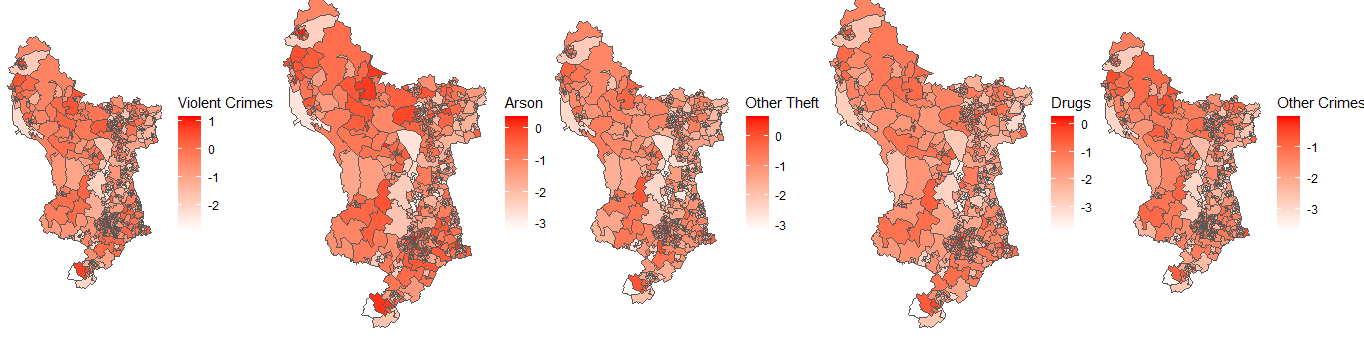
|  |  |  |  |
| --- | --- | --- | --- |
| Plot |  |  |  |
| Description | As seen above, the Population of the entire reqion is a bit right-skewed. With concetration between 1000 and 2500 and outliers around 4000. | As evident in the Treemap, Derby has the highest criminal cases, followed by Amber Valley and Derbyshire Dales comes last. | As seen in the barplot, most of the regions under consideration fall under Low population with a figure of 308 while only 334 regions fall under High population density. |

***The Geomap***

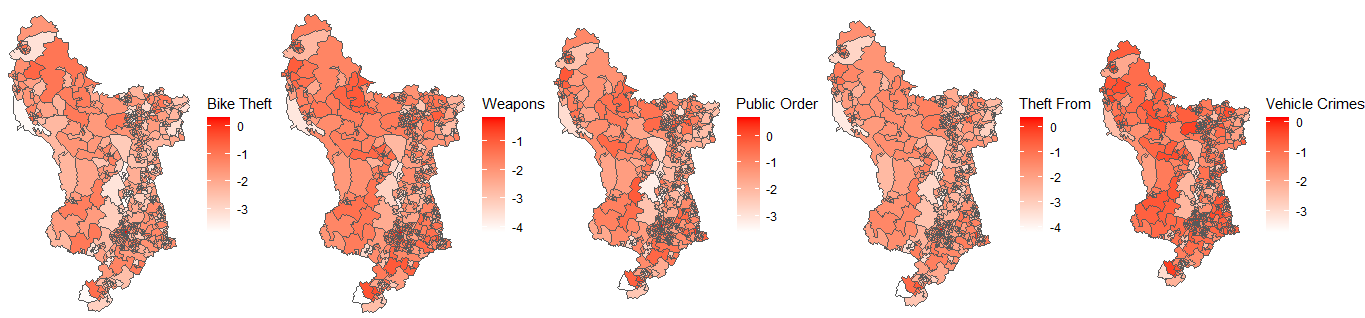
It is obvious that some crimes are more common in some places based on the map of crime density in Derbyshire, UK.



The crimes mentioned, including stealing, arson, violent crimes, car crimes, and more, all contribute to the total crime rate in the region. Law enforcement and other organisations can use the map as a useful tool to concentrate their efforts and resources in the most crime-prone regions.



However, there can be serious risks to people and communities if the crimes are not stopped quickly. For instance, burglaries may result in physical harm to property, the loss of priceless objects, and psychological grief for the victims. Robberies and other violent crimes can result in bodily harm or even death, as well as instilling fear and instability in the communities they influence. Vehicle crimes, such as theft and vandalism, can interrupt everyday life and lead to financial loss for the victims. In addition, if allowed unchecked, crimes like arson and possession of firearms can have major negative effects on public safety.



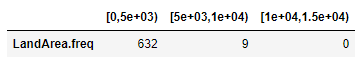
In order to lower crime rates and maintain community safety, it is imperative that law enforcement and other organisations collaborate. Increased patrols in high-crime areas, community outreach and education initiatives, and proactive steps to stop crimes before they start can all be part of this. By taking steps to reduce crime, we can build communities that are safer and more secure for everyone.

Investigating the Regional Land Area

There is need to investigate the Land Area of the Regions under consideration, this would further solidify the reasons for any actionable decision to be made from the previous insights generated and to also create more understanding of the dataset.

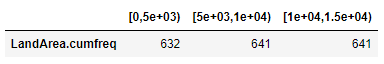
***Frequency Distribution of Land Area***

As seen in the table below, 632 Land Areas fall within 0 to 5000 hectares, 9 falls within 5000 to 10000.



***Frequency Distribution of Land Area***

The cumulated value of the entire observation can be reconfirmed below to be 641.

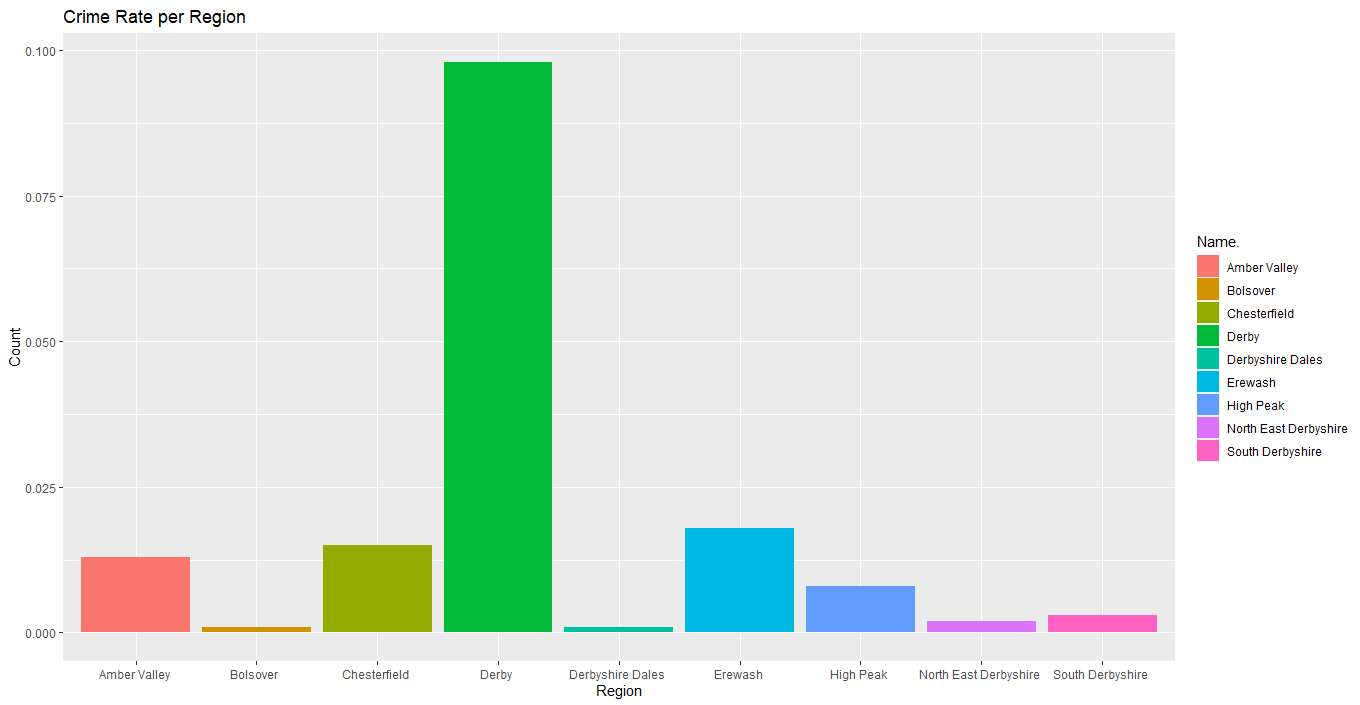


***Mean Regional Land Area***

The average land area in hectares according to the regions under observation can be seen in the table below, with Derbyshire Dales coming with the highest of value of 1842.91hectares, followed by High Peak with 913.81hectares of Land, in the third place we have South Derbyshire with 582.97hecatres of land. Coming last on the list, we have Derby with 51.68hectares of land.



Checking for the Crime Rate by Region



The plot above clearly illustrates that in the given dataset, Derby has the highest crime rate among the mentioned regions, followed by Erewash, and then Amber Valley, while Bolsover and Derbyshire Dales come last with close figures.

**Ethics of Analysis**

The moral standards that direct the way data is gathered, analysed, and published are referred to as ethics in analysis. It's crucial to think about the ethical ramifications of data analysis as the practise spreads and more data is gathered.

Privacy is one of the most important ethical issues in analysis. It is crucial to make sure that any personal information is kept private and that the necessary precautions are taken to prevent unwanted access to it when analysts work with data. Prior to gathering and evaluating the data from participants, it is also crucial to have their informed permission.

Bias is a further ethical issue to be addressed in analysis. Regardless of their own opinions or preferences, analysts must make sure that their analysis is impartial and objective. This entails being conscious of potential bias sources and taking action to reduce them.

Another crucial ethical factor in analysis is transparency. It is crucial for analysts to be open and honest about their processes, data sources, and any analysis's constraints. This enables stakeholders to make defensible decisions based on the analysis and helps to establish confidence and trust with them.

Last but not least, ethical analysis also takes into account how the analysis will affect society as a whole. Analysts must think about the possible outcomes of their study, such as how it could harm certain groups of people or maintain current inequities. In order to achieve social justice and equality, data must be used in a way that does not reinforce current power dynamics or marginalise particular groups.

In conclusion, ethics in analysis is a crucial factor that should direct the way data is gathered, examined, and published. Protecting privacy, reducing prejudice, being open and honest about their processes and constraints, and taking into account how their research may affect society are all important considerations for analysts.

**To avoid misleading analysis and visualization, the following steps should be taken**:

1. Define the scope and objectives of the analysis: Before starting the analysis, it is important to clearly define the scope and objectives of the project. This will help to ensure that the analysis is relevant and aligned with the needs of the stakeholders.
2. Use reliable and accurate data: The quality of the analysis is highly dependent on the quality of the data used. It is important to ensure that the data is reliable and accurate, and that any errors or biases are identified and corrected.
3. Avoid cherry-picking: Cherry-picking refers to the practice of selecting only the data that supports a particular argument or conclusion. This can be misleading and unethical. To avoid cherry-picking, it is important to consider all relevant data and to be transparent about any exclusions.
4. Be transparent about methodology: The methodology used in the analysis should be transparent and clearly documented. This will help to ensure that the analysis can be replicated and that any limitations or assumptions are understood.
5. Avoid over-simplification: Visualisation can be a powerful tool for communicating complex data, but it can also be misleading if the data is over-simplified or misrepresented. It is important to ensure that the visualisations accurately represent the data and that any simplifications are clearly explained.
6. Consider the impact of the analysis: The impact of the analysis should be considered, both in terms of the potential benefits and risks. It is important to ensure that the analysis is used ethically and responsibly, and that any potential harms are identified and mitigated.

Overall, the ethics of analysis requires a commitment to transparency, accuracy, and accountability. By following these steps, analysts can help to ensure that their work is reliable, relevant, and trustworthy.

**Conclusion**

In conclusion, the information visualisation project has given me insightful knowledge about the data I've been examining. I was able to spot patterns and trends utilising a number of visualisation approaches that would have been challenging to spot using conventional data analysis tools.

My visualisations not only helped me understand the data more deeply, but also provided a way to communicate my insights to others but also able to create clear and compelling visualisations that conveyed complex information in a simple and intuitive way, critically evaluate and apply ethical principles and standards to the whole of the visualisation process.

**Reference**

*13 powerful ways to visualize your data (with examples)* (2022) *Sisense*. Available at: https://www.sisense.com/blog/10-useful-ways-visualize-data-examples/ (Accessed: April 23, 2023).

*Data ethics* (no date) *www.cognizant.com*. Available at: https://www.cognizant.com/us/en/glossary/data-ethics (Accessed: April 23, 2023).

*R tutorial: Learn R programming language tutorial - javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/r-tutorial (Accessed: April 23, 2023).