DERBYSHIRE CRIME DATA VISUALISATION

Contents

1.Part 1 – Understand the Data Set and Descriptive Visualisation

(a)Dataset

The dataset used for this study is data on reported cumulative crime incidents in areas under the Derbyshire and Derby City during the year 2019. These are reported crime statistics for each lower layer super output area (LSOA) covering Derbyshire. Thus, the data covers the eight administrative districts: Amber Valley, Bolsover, Derbyshire Dales, North East Derbyshire, South Derbyshire, and the boroughs of Chesterfield, Erewash, and High Peak, as well as Derby City.

The data is categorized into 315 regions covering Derby city and Derbyshire, 69 regions in Chesterfield, 78 regions in Amber Valley, 73 regions in Erewash, 48 regions in Bolsover, and 59 regions in High Peak. It consists of 14 crime indicators that are matched to population and land area (hectares) statistics for each of the 642 regions.

The data contains 18 columns (variables) and 642 rows. The first and second columns contain unique identifiers for each region. All data points have been incremented by 1 to remove instances where a region reported 0, which could impact potential logarithmic transformations of the data.
The variables contained in the dataset are shown in table 1 below:

Table 1: List of Variables in Dataset

List of Variables in Crime Dataset									
S/N	Variable name	S/N	Variable name	S/N	Variable name	S/N	Variable name	S/N	Variable name
1	LSOA	5	Anti-Social Behaviour	9	Violent Crimes	13	Drugs	17	Public Order
2	Name	6	Burglary	10	Shoplifting	14	Other Crimes		
3	Population	7	Robbery	11	Criminal Damage & Arson	15	Bike Theft		
4	Land Area in Hectares	8	Vehicle Crimes	12	Other Theft	16	Possession of Weapons	18	Theft From the Person

Source: Compiled by author

This dataset could be used as follows:

i. Assess the trend in crime rate across Derbyshire in 2019
ii. Determine the crime rate per capita in each area
iii. Determine the impact of population densities on crime type
iv. Make prediction regarding the different types of crimes in each district
v. Ascertain the concentration of crime type in different areas.

(b)Data Mining with R programming

Data was read and stored as Data1 and first 6 rows were inspected as shown below:

```
#Loading of dataset and view the first 6 rows
Data1<-read.csv('AssessmentCrimeData.csv')
head(Data1)
```

```
##                 LSOA              Name  Population  Land.Area.in.Hectares
Anti.Social.Behaviour
##  1  E01013453  Derby  013A            2497                              47.79
205
```

```
##  2  E01013454  Derby  013B           1418                              21.89
53
## 3 E01013455 Derby 017A       1855                  63.71
```

Missing Values

Dataset was also checked for missing values and the result showed there are no missing values.

Summary Statistics

Summary statistics was also done to ascertain the statistical properties of the numerical variables. The results showed statistics like 1st quartile, median, mean, etc. The standard deviation was obtained separately with the relevant R code and all the results are shown in table 2 below.

Table 2: Summary Statistics

Table of Summary Statistics								
S/N	Variable Name	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
1	Population	993.00	1,411.00	1,584.00	1,657.00	1,815.00	3,948.00	371.13
2	Land Area in Hectares	12.81	37.26	71.56	408.84	209.55	16,227.00	1115.23
3	Anti-Social Behaviour	3.00	22.00	36.00	52.00	58.00	1,359.00	73.95
4	Burglary	1.00	5.00	8.00	10.25	13.00	158.00	9.08
5	Robbery	1.00	1.00	1.00	2.16	2.00	79.00	3.94
6	Vehicle Crimes	1.00	5.00	8.00	9.34	12.00	60.00	6.56
7	Violent Crimes	3.00	22.00	35.00	49.90	58.00	1,280.00	63.81
8	Shoplifting	1.00	1.00	1.00	9.73	5.00	613.00	32.71
9	Criminal Damage & Arson	1.00	7.00	12.00	14.86	19.00	196.00	13.66
10	Other Theft	1.00	5.00	8.00	11.91	13.00	382.00	19.13
11	Drugs	1.00	2.00	3.00	4.67	5.00	150.00	9.17
12	Other Crimes	1.00	2.00	3.00	3.83	5.00	46.00	3.65
13	Bike Theft	1.00	1.00	1.00	2.55	2.00	170.00	7.32
14	Possession of Weapons	1.00	1.00	2.00	2.22	3.00	60.00	2.89
15	Public Order	1.00	4.00	7.00	10.57	11.75	404.00	20.16
16	Theft From the Person	1.00	1.00	1.00	2.22	2.00	202.00	8.57

Source: Computed by Author

Analysis of summary statistics

i. Twelve variables have 1.00 as their minimum while Violent crimes and Anti-Social behavior have 3 as their minimum. Land Area in Hectares has 12.81 while the highest minimum value occurred in Population. This results shows that these twelve variables have instances where they reported no crimes (as 1 was added to such data points at the beginning).

ii. The closeness of media and mean shows how normally distributed a data is. It also shows the level of variability in the data. From the result above, Other Crimes has the highest closeness of median (3) and mean (3.83). Other variables have mixed differences with Land Area in Hectares having the highest difference of 337.28 (408.84-71.56). These variations suggest the data is not normally distributed. It also suggests high presence of outliers.

iii. Small SD signifies that results from data are very close the mean values. The larger the SD the higher the level of variance in the data. From the Standard deviation (SD), Land Area in Hectares with has the highest variability with SD value of 1115.23. This is followed by Population with

SD of 371.13 while Other Crimes has the lowest variability with SD of 3.65. These confirm high skewness of data and absence of normality.

iv. There are several outliers in the data as could be deduced from table 2. The wide differences between the minimum and maximum values, the mean and median, as well as the high SD values suggest presence of outliers.

Data transformation using Logarithm and Densities

Given the observed high variability in the data, logarithm transformation was employed to moderate the level of skewness in the data. Similarly, densities were used instead of the raw data to moderate the effect of concentration of people in certain cities vis- a-vis areas that are sparsely populated.

Relevant plots

Graphical presentation of the transformed data is shown in figures 1and 2.

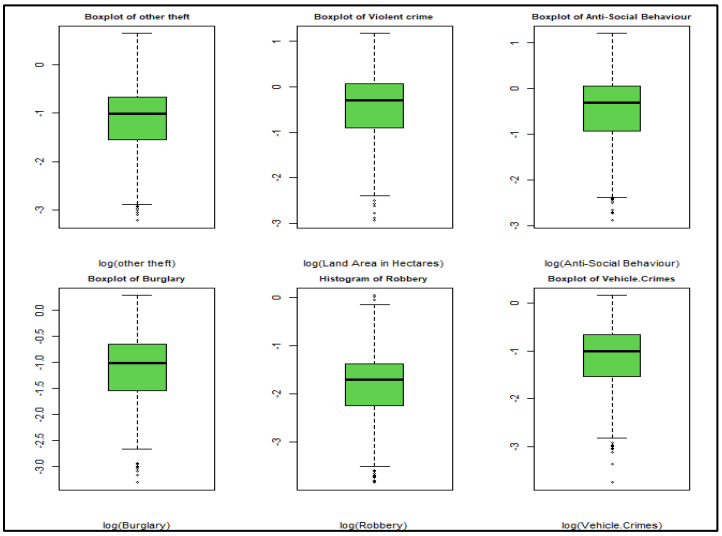

Figure 1: Boxplot of selected variables

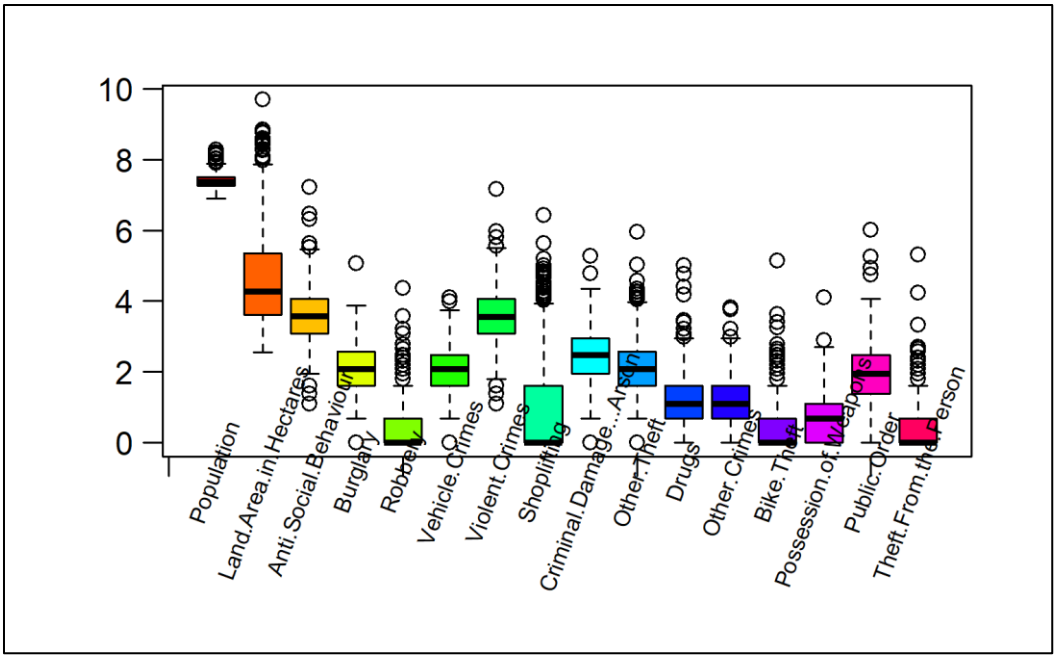

Figure 2: Boxplot of all the numerical variables

Analysis of plots

1. There appears to be some convergence to normality after data transformation as could be seen in the boxplot (fig1 and fig 2). In crimes such as Other theft, Burglary, Vehicle crime, Drugs, etc, the median is in the middle of the box, suggesting symmetric distribution (normality). Thus the transformation improved the data quality.
2. From fig 2, the median is closer to the bottom of the box in crimes such as Robbery, shoplifting and Theft from the person. This suggests their distributions are positively skewed. Furthermore, some outliers however, are still present as depicted in figure 2. High number of outliers is present in variables such as population, Robbery, violent crime and shoplifting.

Part 2 – Exploration of the relationships in data set

(2a) Linear Regression

Simple linear regression model was used to examine the relationships in the model. The general form of the model is given by Y=B0 + B1X + e.

Where: Y = the dependent or response variable
X= independent or explanatory variable
B0 and B1 = parameters of the model to be estimated
e= random error term

Assumptions of a linear regression model

Some of the basic assumptions of a linear regression model are: 1.There is linear relationship between the dependent variable and the independent variable(s) 2. The error term is normally distributed 3. Independence of the successive values of the error term. 4. Homoscedasticity or constant variance of the error term.

Estimation of regression model

Fourteen models of simple linear regression were estimated for each of the 14 crimes in the dataset. The data was transformed such that the log of density of each crime is the response variable while log of population density is the explanatory variable. Thus one explanatory variable "log10 Population density" was used for all the 14 models, while the response variables differ according to the various crimes (log of densities of each crime). The estimated model is presented in table 2 below.

Table 2: Regression results

REGRESSION RESULTS								
MODEL/ PARAMETER	**DEPENDENT VARIABLE**	**B0**	**B1**	**P-VALUE(B0)**	**P-VALUE(B1)**	**F-STATISTIC**	**P-VALUE (F-STAT)**	**ADJUSTED R^2**
MODEL1	log (Anti-Social Behaviour density)	-1.82146	1.14829	<2e-16	<2e-16	3063	<2e-16	0.8269
MODEL2	log(Burglary density)	-2.30167	1.00289	<2e-17	<2e-17	3082	<2e-17	0.8278
MODEL3	log(Robbery density)	-3.18957	1.14904	<2e-18	<2e-18	4579	<2e-18	0.8772
MODEL4	log(Vehicle Crimes density)	-2.35954	1.02457	<2e-19	<2e-19	3367	<2e-19	0.84
MODEL5	log10(Violent Crimes density)	-1.8708	1.18624	<2e-20	<2e-20	3721	<2e-20	0.853
MODEL6	log10(Shoplifting density)	-2.95277	1.14475	<2e-21	<2e-21	951.3	<2e-21	0.5972
MODEL7	log10(Criminal Damage & Arson density)	-2.30629	1.12458	<2e-22	<2e-22	3175	<2e-22	0.832
MODEL8	log10(Other Theft densty)	-2.30958	1.00335	<2e-23	<2e-23	2018	<2e-23	0.7588
MODEL9	log10(Drugs density)	-2.84592	1.08812	<2e-24	<2e-24	2301	<2e-24	0.782
MODEL10	log10(Other Crimes density)	-2.83232	1.07552	<2e-25	<2e-25	3128	<2e-25	0.8299
MODEL11	log10(Bike Theft density)	-3.1588	1.14374	<2e-26	<2e-26	3631	<2e-26	0.8499
MODEL12	log10(Possession of Weapons density)	-3.08451	1.09586	<2e-27	<2e-27	4228	<2e-27	0.8683
MODEL13	log10 (Public Order density)	-2.53849	1.14124	<2e-28	<2e-28	2555	<2e-28	0.7994
MODEL14	log10 (Theft From the Person density)	-3.13702	1.07455	<2e-29	<2e-29	3823	<2e-29	0.8564

Source: Computed by author

Interpretation of estimated models

From the results, the models have negative intercept and are statistically significant. This implies that the regression lines pass below the origin (zero). Similarly, all the models have positive slopes which are also statistically significant at the 5% level of significance (α=0.05). The estimated parameters (bo and b1) are statistically significant because their respective p-values are individually less than 0.05 (that is 2e-29 < 0.05) or (0.00000000000000000000000000002 < 0.05) respectively.

From Model1, the result implies that one unit change in population density will result in 1.14829 units increase in Anti-Social Behaviour density, holding all other factors constant. Similarly, from model 10 one unit change in population density will result in 1.07552 units increase in Bike Theft density, holding all other factors constant.

Thirteen of the models have high explanatory powers, with R-squared ranging from 75.88% to 87.72%. Only Model6 has a moderate explanatory power with R-squared of 59.72%. For instance, the results means that 82.69% of variations in Anti-Social Behaviour density is explained by Model1, implying a high explanatory ability. The other models could be explained in the same manner.

The F-statistic results for all the models show p-values less than the significance level of 5% (that is 2e-29 < 0.05) respectively. This means that overall, all the models are statistically significant.

Checking for the assumptions of the regression model

i. **Linearity:** The response variables were plotted against the explanatory variable from all the 14 models to check for linearity. The results show linearity in all the models except Model 13 (public order). In Pubic order, the deviations of the points from the lines are wide. For the other models, the points cluster more on the line.

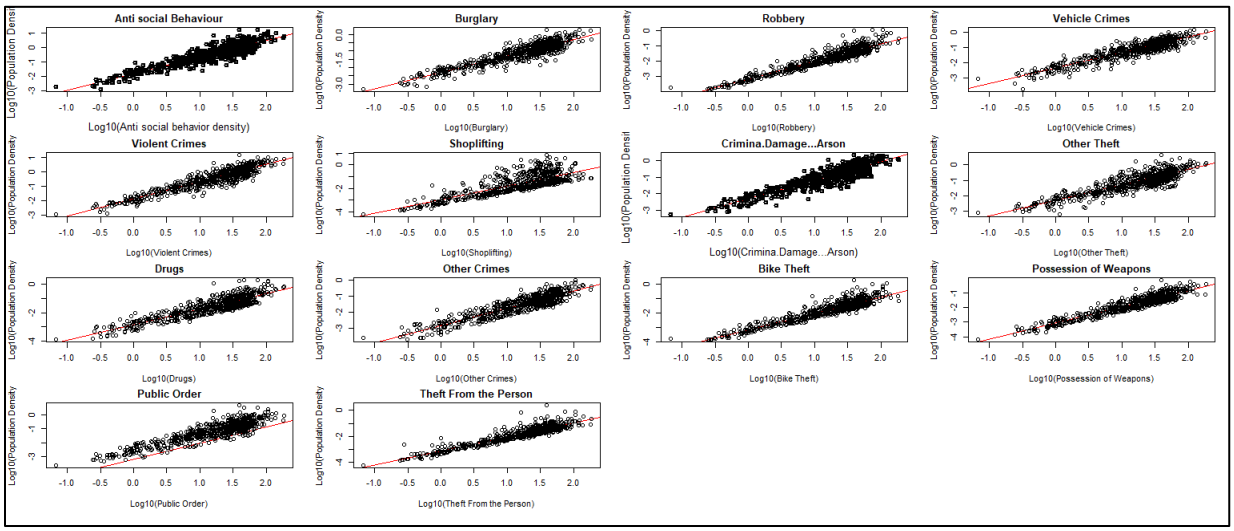

Figure 4: Scatter plots of the response variables and the explanatory variables

ii. **Normality:** The residuals were plotted in a histogram to check for normality. The results show that Anti-social behaviour, Burglary, vehicle crime, violent crimes, Criminal Damage & Arson, Other theft and Public Order have bell-shaped histograms suggesting normal distribution. However the other 7 variables are not normally distributed. This is shown in figure 5 below.

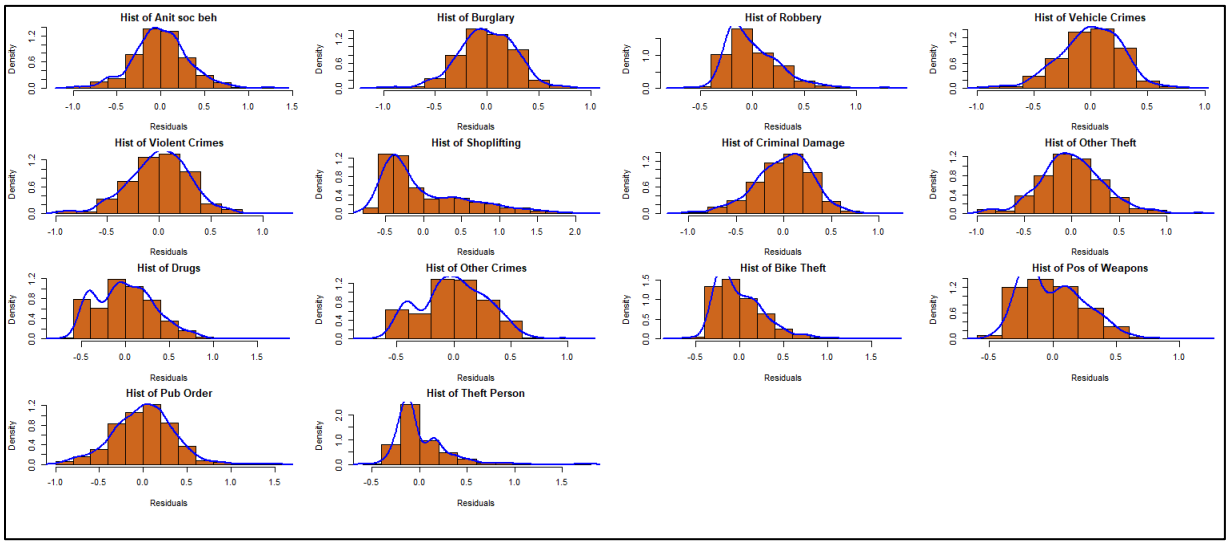

Figure 5: Histogram of residuals

iii. **Constant Variance (Homoscedasticity):** The residuals were plotted against Fitted values to check for Homoscedasticity. The results suggest the models for Anti-social behaviour, Burglary, vehicle crime, violent crimes, Criminal Damage & Arson, Other theft, Other Crime and Public Order have constant variance in the error term. This is because the points are fairly equally spread around the red fairly horizontal lines which pass through the origin (zero). This is shown in figure 6 below.

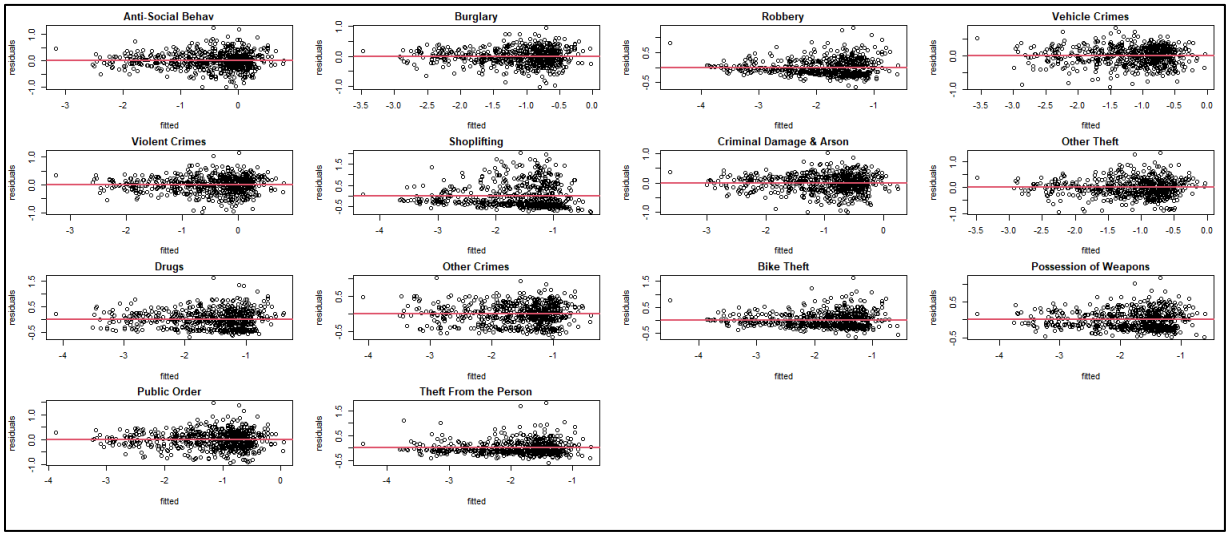

Figure 6: Residuals Vs Fitted plot

iv. **Independence:** The residuals were plotted with lagged length values and epsilon to check for independence. From the results in Figure 7 below the points are fairly scattered in model1 to Model 13, thus suggesting independence. Only model14 (Theft from the person) appear to lack independence as the points are somewhat clustered together.

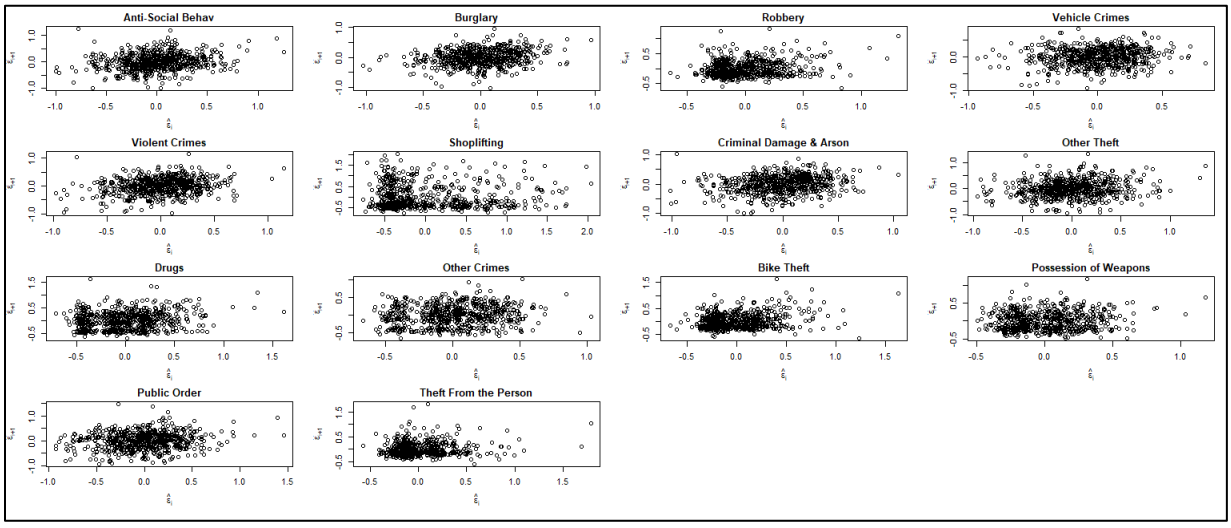

Figure 7: Residuals Vs lagged Values

2b: Correlation Matrix and Heatmap of the residuals

The Correlation matrix and heatmap of residuals are shown in figures 8 and 9 respectively. From the correlation matrix, there is strong evidence of correlation among the residuals as shown by the scatter plots in each pair. Most of the scatter points flow upwards from left to right, suggesting positive correlation. Similarly, from the heatmap in figure 9, there is strong positive correlation between Anti Soc Behaviour and Public order; Violent crimes, and Criminal Damage & Arson, respectively. This is shown by the pink colour corresponding to 0.7 correlation from the colour key. On the other hand, low positive correlation is observed between shop lifting and Theft From the Person; Burglary and Drugs; Robbery and Shoplifting, etc as shown by the corresponding colour pallet.

There are three main clusters: (1) Shop lifting, vehicle crime and burglary. (2) other crimes, possession of weapons, Theft From the Person, bike theft and Robbery (3) Drugs, other theft, Criminal Damage & Arson, Violent crime, Public Order and Anti Social Behaviour.

Outliers: The outliers are Shop lifting, vehicle crime and burglary as shown in figure 9.

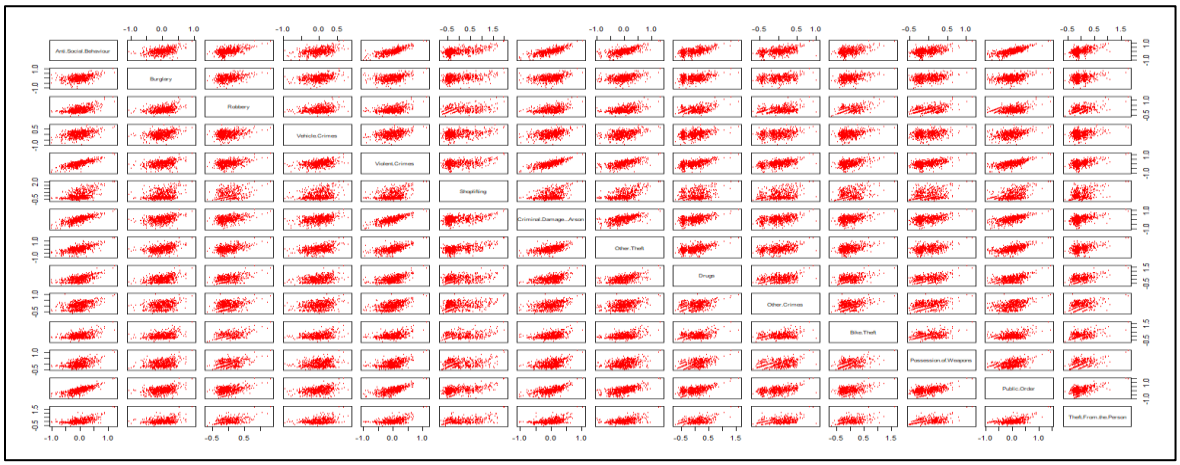

Fig8: correlation matrix of the residuals.

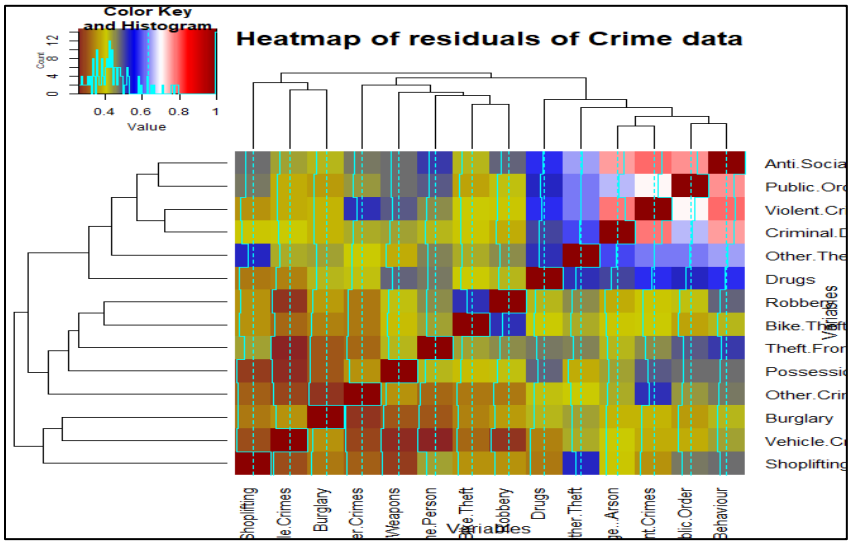

Fig9: Heatmap of residuals

2c: Hierarchical clustering

Clustering analysis was performed on the residuals using Dendograms (figure 10). The Elbow analysis was also done to determine the optimal number of clusters (figure11). Through the elbow analysis, the optimal number of clusters was determined to be 3. Figure 12 shows thee clusters as determined by the elbow analysis and the outliers could be identified at the extreme left corner.

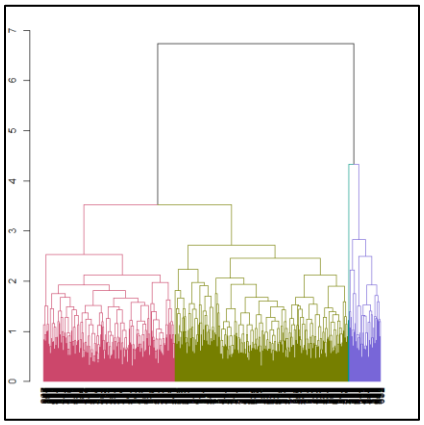

Fig 10: Elbow analysis

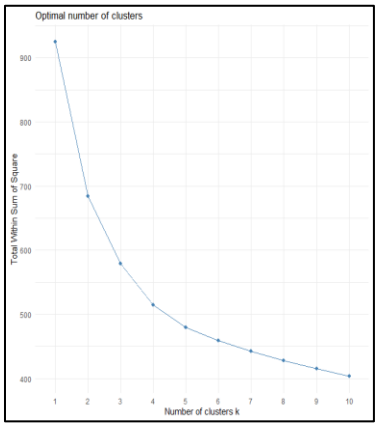

Fig 11 Dendogram of residuals

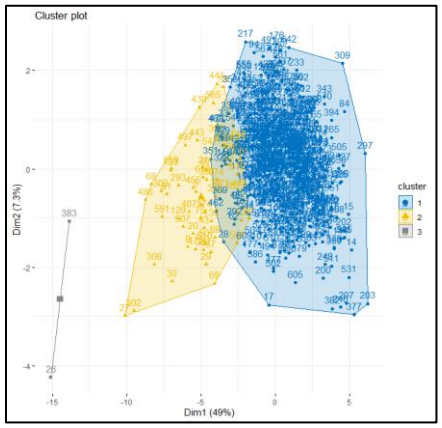

Fig. 12 Clusters of residuals.

Part 3: Independent Evaluation

(a)Geomapping

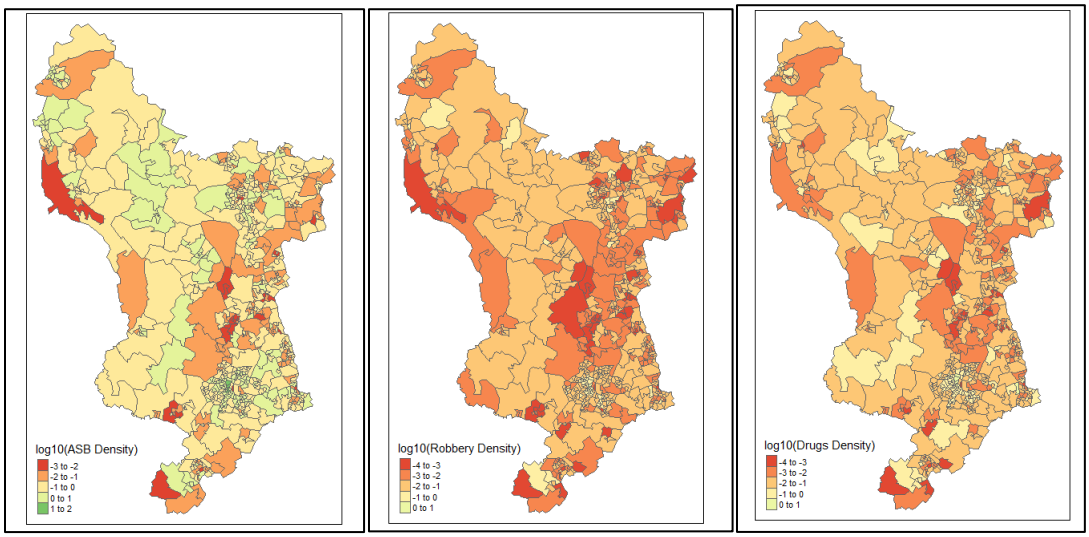

Fig 13: Maps of ASB, Robbery and Drugs densities

Geomapping and spatial analysis were carried out on the data to establish further insights. All the crime variables were analysed however, 3 crimes are presented in Fig 13 above. From the results, ASB density of -1 to 0 dominate the map while ASB density of 1 to 2 only appear in a tiny area on the southern part of Derbyshire. Density -3 to -2 is concentrated in the north eastern part. Similarly, in the Robbery density map, the density value of -2 to -1 dominate, while Drugs

density -2 to -1 also dominate the map. In general therefore, the result shows that these crimes are widely spread in Derbyshire but at very low rates.

(b)Ethical Issues in Information Visualization

Ethics are moral principles that influence personal behaviour. It is concerned with doing what is right in every endeavor. Hence, ethics in visualization seeks to ensure that analysis and visualization are done in the right way, devoid of misrepresentation of fact. In other words, ethical principles seeks to ensure that visualization avoids any form of harm to users (Hagendorff , 2020).

Visualization extracts insights and involves interpretation of data usually designed to tell a story so as to promote good decisions making. Hence, peoples' views could be manipulated through unethical visualization. Producing data visualization that is confusing, manipulation of facts or misleading will result in harming the users. This is of serious ethical concern. It could lead to people making the wrong decisions or choices. In general, visual representations of data should be honest, responsible and embodied in fairness.

Factors that will prevent misleading analysis and visualization include:

1. Data privacy rights of data subjects and users should be respected and should be enforceable. The General Data Protection Regulation (GDPR) should be adhered to.
2. Honest data presentation: – Manipulating or distorting data to mislead or misinform viewers must be avoided. This involves choosing appropriate chart types, scales, and data transformations that accurately show the underlying data.
3. Eliminating Bias through Fairness and objectivity. Visualization should be presented objectively without introducing personal bias. Transparency about data sources, methodology, and limitations can help create suitable design to users need embodied fairness and credibility.
4. Visualization should not be ambiguous but show clarity, simplicity and objectivity
5. Enforceable regulatory frameworks should be enshrined to forestall misrepresentation of facts.

Conclusion

This study employed a dataset to investigate fourteen types of crime in the various cities of Derbyshire. The evaluation of the statistical properties of the data revealed high level skewness and divergence from normality. Therefore, the data was transformed into logarithmic state, to reduce skewness and encourage normality. Given that the various cities differ in size, the effect of varying population and land mass was diluted by converting the variables to densities.

Fourteen models of simple regression analysis were used to ascertain the impact of population density on the densities of the various crimes. The results showed that population density is a statistically significant factor affecting the various crimes in Derbyshire. The models also show high explanatory power evidenced in high Adjusted R^2. The models largely conform to the assumptions of linear regression as shown by the various tests of normality, linearity, independence and homoscedasticity carried out on the model residuals.

A cluster analysis done on the dataset revealed 3 clusters with Shop lifting, vehicle crime and burglary identified as outliers. Derbyshire maps were also created and the prevalence of each crime was visualize on the maps. The comparism of Antisocial behaviour, Robbery and Drugs result shows that these crimes are widely spread in Derbyshire but at very low rates.
Finally, it was suggested that that adhering to GDPR, honesty, objectivity, fairness are some of the ways to imbibe ethical principles in data visualization.

Bibliography

Bisoux T. (2019). The Ethics of Data Visualization. AACSB. Accessed 18 May 2023. https://www.aacsb.edu/insights/articles/2019/12/the-ethics-of-data-visualization

Breheny, P and Burchett W (2017). Visualization of Regression Models Using visreg. The R Journal Vol. 9/2, December 2017

Hagendorff Thilo (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines (2020) 30:99–120

Siddiqui, A. T (2021) Data Visualization: A Study of Tools and Challenges. *Asian Journal of Technology & Management Research (AJTMR) ISSN: 2249 –0892 Vol11 Issue–01, Jun -202118*

Zicari R. V, & Zwitter, A. (2016) Data for Humanity: An Open Letter. Goethe University, Frankfurt. Retrieved from: http://www.bigdata.uni-frankfurt.de/dataforhumanity/