Analysing Crime Data from Greater Manchester Area: Applications of GLM

Jacob Mortimer

May 7, 2025

Contents

List of Figures

Abstract

This project investigates the relationship between key demographic and socioeconomic variables and crime rates across Lower Layer Super Output Area (LSOA's) in Greater Manchester. Using crime data from December 2018 and 2021 Census data, a generalized linear model framework was applied to model crime counts for each area. After exploratory data analysis and extensive variable assessment, LASSO regularization and bivaraite interaction analysis were used to produce a parsimonious, interpretable negative binomial model that balances statistical strength with simplicity. This final model explained approximately 41.6% of the variation in crime, identifying key predictors such as recent immigration, tenure type, and household deprivation as having the strongest associations. Fixed effects for districts were included but regional geography itself wasn't a significant predictor once demographics were accounted for. Overall, these findings not only reinforce key criminological ideas, but quantify their influence relative to each other in Greater Manchester.

Acknowledgment

I would like to thank my supervisor Madhuchhanda Bhattacharjee for her guidance and feedback throughout the project and also the Department of Mathematics at The University of Manchester for providing resources to make this project possible. Also, thanks to the work done by several researchers before me, I was able to identify crime predictors to watch out for and include in my own research because of them. Finally, I acknowledge the use of publicly available census and crime data from the Office for National Statistics and police.uk which formed the foundation of my dataset.

1 Background

1.1 Motivation

As a student living in Fallowfield Manchester, I constantly have a persistent unease about the safety of my possessions within my student house because, upon moving into my student home in September, two home invasions occurred within the space of a week on my road. One of which was directly next door. This does not come as a surprise to me as I have been aware of the problems with crime in Manchester, especially historically in areas such as Moss Side (famously nicknamed 'Gunchester') which have a negative reputation for being dangerous [2] . However, what came as a shock to me is that, through investigation of crime statistics due to curiosity about my recent experiences, I found that Greater Manchester had a staggering crime rate of 117.7 per 1000 people (the third highest county rate in the UK as of 2024).This is 12.6 higher than London, which I thought was one of the most dangerous areas in the country [1] . This inspired me to look at what causes crime variation over different areas.

1.2 Objective

A rather unsettling reality is that Greater Manchester is a hot spot for criminal activity, but since I have already taken up study here (like thousands of others) due to the pedigree of the university and the plethora of attractions the city of Manchester provides, I thought the purpose of this project would be to discover why area dependent crime patterns emerge within Greater Manchester rather than just relying on purely historical crime statistics – which only offer a static view of past events and fails to account for evolving factors. I want to determine the area-level factors that are most strongly associated with an increase in the crime rate over a fixed point in time in hopes of providing a broader understanding of how community-level characteristics in Greater Manchester influence public safety. A key analytical goal to achieve this is to construct a parsimonious statistical model that balances simplicity and interpretability with explanatory power, identifying a minimal set of strong predictors. Hopefully a clear, interpretable model can contribute to informing students which areas are the safest to live or operate within as well as inform long-term crime prevention discussions.

1.3 Chapter summary

The key sections of my project and the role they serve in satisfying my objective are as follows:

1.3.0.1 Research: In this section, I will discover what qualities of a specific area could potentially influence its respective criminal activity based on various research papers that have been conducted around the world at various points in time. I will also detail ways to quantify these predictive qualities an area possesses in the form of a variable such as 'Number of people aged 18-24'.

1.3.0.2 Data: This section will detail the process of obtaining a dataset which contains observations of all of the potential variables (which were previously flagged to have some influence on an areas crime rate) for every location in Greater Manchester at a fixed point in time. Additionally it will provide details about said data.

1.3.0.3 Modeling: This section outlines the statistical methods used to construct a model that explains the relationship between area-level explanatory variables and crime count. It focuses on the mathematical framework of Generalized Linear Models,

1.3.0.4 Data Analysis: This section will identify which GLM best fits our data. It will also establish which LSOA-level census variables are most strongly associated with crime using said GLM. This will be achieved using a combination of multicollinearity checks, univariate performance (Nagelkerke R^2), Wald tests, bivariate interactions, grouped models, AIC-based stepwise selection, and LASSO regularization. This analysis will result in the formulation of a parsimonious, interpretable model that balances statistical strength with simplicity (Note that analysis in this project is cross-sectional, not predictive over time, due to the fixed-point nature of the available data).

1.3.0.5 Conclusions: This section will interpret the results of the final model, discuss limitations of the data and model, and suggest areas for further research.

2 Background

2.1 Demographics Of An Area

What motivates a person to commit a crime in a specific area? In the study [3], the concept of a persons awareness space and activity space is mentioned when anticipating a criminals pattern for crime location. An awareness space is every location an individual is aware of, whether that be through first hand experience or word of mouth. An activity space however, are locations a person is well acquainted with such as their local area around their home, their workplace and recreational areas they frequent like pubs. A criminal is more likely to act within his/her activity space because of the decreased risk (as they are familiarized with escape routes and the local's routines) and because if they spend more time in certain areas, opportunities for criminal activity present themselves more frequently.

2.1.0.1 Accepting that many crimes are committed by local residents supports the use of area demographics to infer both individual motivations for offending and the broader social dynamics that shape community-level crime patterns making them valuable inputs to statistical models of crime. The following sections break down key demographic and contextual variables that I believe influence crime rates by area based on previous studies.

2.2 Crime and Poverty

Financial instability can induce crime as everyone has a survival instinct and needs money to live (also areas that are in poverty have more opportunities for crime due to lack of surveillance and urban decay). Research has shown that neighborhood poverty and associated structural factors continue to predict multiple crime related outcomes [15] [16]. Additionally, areas which have a wealth disparity could be subjected to more crime [17]. The cause of this disparity could be down to economic factors such as unemployment rates or job income inequality in the area which stems from a persons education level. Furthermore, a potential cause for areas of poverty could be systematic barriers such as housing discrimination and social exclusion against immigrant and ethnic minorities [6]. Also, urban decay (such as broken windows and graffiti) could be an indicator of the economic success of the people in the area [23]. This is why I think it will be constructive to find the following quantitative variables which capture poverty and affluence: employment rates, education level, indicators of wealth, and signs of urban decay. Additionally, incorporation of religion and ethnicity should be considered here too as this may also contribute to an areas affluence indirectly.

2.2.0.1 Quantifying Variables: Levels of Employment, Level of Education, Number of Rooms, Number of Reported Incidents of Urban Decay, Religion, Ethnicity.

2.3 Crime and Age

In study [4], Casey believes that a cornerstone of cognitive development is the ability to suppress inappropriate thoughts and actions in favor of goal-directed ones, especially in the presence of compelling incentives. A number of classic developmental studies have shown that this ability develops throughout childhood and adolescence and hence age will have an impact on the likelihood of a crime being committed, especially when the child is in a financially disadvantaged environment as the 'compelling incentive' would be to steal or otherwise to get money and resources. In addition, young people are more likely to take risks and be influenced by peer pressure [5]. This means that the percentage population of different age groups in an area could be significant to the overall crime rate so age range variables should be considered.

2.3.0.1 Quantifying Age Variable: Age Brackets of the population.

2.4 Crime and Opportunity

Urban decay, as well as being a sign of an areas wealth, also increases opportunity for crime. This is intuitive as a broken street light reduces visibility so an opportunity for crime without consequence of detection will present itself more frequently.

2.4.0.1 Additionally, surveillance has a profound impact on crime rates in the UK as shown in [11] where it's stated that 'actively monitored systems (e.g., real-time surveillance by operators) are more effective, showing approximately a 15 percent crime reduction, whereas passive systems often yield negligible effects.' This means that areas with more actively monitored systems could experience less crime as criminals will not have a risk-free opportunity to offend. The deterrent of surveillance extends to police presence as in [12] a study showed that visible marked police cars in crime hot spots reduced the crime rate by 31 percent.

2.4.0.2 High population density can influence crime rates in multiple ways. With more people, there are simply more potential targets and interactions, which increases the opportunity for criminal acts. Additionally, larger populations may include a greater number of individuals on the extreme ends of behavior (statistical outliers), potentially leading to a higher number of offenders. Densely populated areas may provide a sense of anonymity, which can encourage criminal behavior [7]. Population size can also indirectly affect conviction rates. As population (and crime) increases, local criminal justice systems may become overburdened, reducing court efficiency. A strained system—combined with limited police resources and possibly lower public willingness to report crime—can lead to lower conviction rates. If offenders perceive that the likelihood of being caught and convicted is low, this may further encourage criminal activity.

2.4.0.3 Road infrastructure affects crime rate as a location which is more accessible is more easily burgled and more likely to be in a criminals 'activity space'. This means that the location and frequency of crime could depend on the layout of the road in an area. Consequently, quantitative variables such as population size, court conviction rate, police funding/police activity in an area, number of surveillance systems, and quality of transport infrastructure could be insightful variables that correlate to crime differences area to area.

2.4.0.4 Quantifying Opportunity Variables: Population size, Conviction Rate, Number of Active Patrol Units, Number of Working Surveillance Systems, Metres of Road.

2.5 Crime and Social factors

There has been research done to suggest that the amount of pubs could be influential to crime rate as people are more likely to display disorderly conduct or be violent when under the influence of alcohol[18]. Interestingly, partly due to an increase in pub attendance but mainly due to more people being outside, the hot weather has also been shown to increase crime rates[19]. More generally, drug abuse and poor mental health could contribute to crime rate so areas with more resources in mental healthcare, substance abuse interventions and social services could see less crime. Additionally, children in unstable home environments as a result of divorce, domestic conflict or neglect among other things are more likely to offend [20].

Another factor to consider is areas with known gang presence and consequently an increased availability of illegal drugs and weapons. This will significantly contribute to an increase in criminal offenses relating to drugs, violence and property crimes. [21]

2.5.0.1 Social Cohesion: Finally, tenure of property impacts crime rate as areas with a higher ownership will tend to have more social cohesion making crime less likely [22]. Also, recency of residency and year of arrival into the country could play a part in overall social cohesion as well as its previously mentioned connection to poverty. The Mitigation Observatory at the University of Oxford notes that, while immigration and diversity can enrich communities, rapid demographic changes can challenge social cohesion when integration mechanisms (like employment services for migrants) aren't in place. Consequently, the study finds that areas with higher proportions of migrants/recent arrivals may experience weaker community networks which inhibits civic engagement and consequently deteriorates the social cohesion and trust that plays a role in crime prevention. [24]

2.5.0.2 The idea of gentrification (whereby an area of poverty is changed by wealthy people moving in) also changes crime rate in unique ways. On one hand, areas with high gentrification would experience a decrease in crime due to improved affluence but, much like tenure, this comes at a cost of social cohesion. This study shows gentrification decreases crime overall as the positive changes outweigh any social cohesion factors [25].

2.5.0.3 Here, to quantify mental healthcare, accessibility to appropriate services would need to be considered. To quantify family stability, divorce rates would be representative. Additionally, by considering the effects of social cohesion with relation to crime, it becomes important to observe tenure, residence year of arrival into the Uk and again ethnicity and religion (which we have already considered when thinking about area wealth). To measure gentrification, I would look at Residency duration. (Note that a more comprehensive statistic has been developed here named the gentrification index which also factors in house price and wage [26])

2.5.0.4 Quantifying Social Variables: Tenure, Year of Arrival to the UK, Residency Duration, Number of Mental Health Centres, Divorce Rate, Number of Gang Related Crimes Reported, Number of Pubs, Temperature.

3 Data

3.1 Crime Data

Police.uk [9]. Each police force generates a Crime and ASB (antisocial behaviour) file which contain reported crimes for that month. They also are required to release a police outcomes file which updates previous months cases with their outcome (eg Investigation complete no subject identified). The data goes through a quality assurance process involving format validation, automated testing and manual verification. This ensures the data is formatted correctly with valid dates and locations as well as having no required fields blank.

3.1.0.1 Anonymity Because this data must protect the privacy of the victims of crime, there is an anonymity process which adequately minimizes privacy risks whilst maximizing the data's transparency to the public. This is done by truncating date of crime to just the month and year, encrypting offense reference, categorizing crimes into generalized groups eg 'Violence and Sexual Offenses' and only revealing a local approximation of crimes place of occurrence. (Usually assigned to center point of a street or a public point of interest such as a pub/airport/shopping centre).

3.1.0.2 Known Issues with Crime data There is some location inaccuracy as police cannot be fully confident where the crime exactly occurred whether that be because the crime wasn't committed in the force gazetteer system or the victim couldn't report the specific location. Estimates of geocoding accuracy in different forces range from 60-97 percent. Court result matching is difficult as there is no unique identifier for crimes that runs from the police service to the CPS (crown prosecution service) so the police use 'fuzzy matchmaking' with a success rate between 19-97 percent. Finally, data from a certain month can be subject to change as the case progresses, (for example if it turns out to be more than just assault and it occurred in a different location)

3.1.1 Response Variable from police.uk

The crime data identifies crime location by LSOA (Lower Layer Super Output Area) and so from now on, all data gathered in this project must be separated into these spatial areas to match with the crime count. In Greater Manchester, there are around 1700 distinct LSOA's which are accounted for in the crime data and consequently we must match every observation with their own explanatory variables so we can build a consistent dataset that allows for valid comparisons and statistical modelling. This alignment ensures that each unit of analysis, the LSOA has both a crime count and a complete set of predictor variables, enabling the fitting of an appropriate model to assess which factors are most strongly associated with crime levels.

3.1.1.1 Because police.uk (as of 6/12/24) don't have up to date crime data for Greater Manchester, I have used the most recent data available from the month of December 2018. The data I collected from police.uk was a list of all cases that occurred so I changed it to count data for each LSOA which will make it easier to handle statistically.

3.2 Census 2021

[13] The 2021 Census data collected by the Office for National Statistics (ONS) provides a detailed insight into the demographic and socioeconomic state of every output area (OA) in the UK and is consequently the main dataset I will be using to investigate which areas in greater Manchester are the safest. The census data was primarily collected via an online form sent to every household in the UK. To ensure maximized responses, paper forms were sent to houses that didn't complete the form online and follow up procedures were conducted by field officers sent to non-responding households. Additionally, there was local support centers to aid people who needed assistance. Accuracy of this data was ensured through collaborations with other data sources such as school enrollment records, tax records and health data. Also ONS used advanced imputation techniques to estimate missing data. The survey data included information on most of our explanatory variables listed above at every 'output area' in Greater Manchester. (For context, the smallest geographical unit for which data is grouped in the census is called an output area and usually consists of around 125 households or 300 people). The data is formatted is follows: for your desired location (whether that be as small as an OA or as large as a county) there are survey results totaled up for that area. For example, the survey asks for age of every household member so in the data you will see the number of people in every age bracket in that area.

3.3 Explanatory Variables from the Census

Firstly, the data is grouped into layer super output area (LSOA) to be consistent with crime data formatting. As speculated with evidence above, I have identified the following survey responses from the census 2021 to be relevant data for my investigation:

3.4 Data to measure age

Age is categorized in Age bands and reports the populations age from the day of the Census in 2021. This data is especially reliable as it's highly comparable to the census 10 years prior.

3.5 Data to measure affluence

3.5.1 Household Deprivation and Number of Rooms

Deprivation is identified through dimensions of deprivation which are used to categorize households (for example: Households with 1 dimension of deprivation). The 4 dimensions of deprivation are:

- Education: A household is deprived in the education dimension if no one has at least level 2 education and no one aged 16 to 18 is a full-time student

- Employment: A household is deprived in the employment dimension if any member excluding full-time students is either unemployed or economically inactive due to a long-term sickness

- Health: A household is deprived in the health dimension if any house member has general health that is very bad or is identified as disabled. (Identification for disability is in line with the Equality Act 2010 and states that it's a person who assessed that their day-to-day activities are limited by their long-term physical or mental health conditions/illnesses)

- Housing: A household is deprived in the housing dimension if the accommodation is either overcrowded, in a shared dwelling, or has no central heating.

3.5.2 Education Level

The census data provides information on all qualifications held by participants aged 16 and over. This may include foreign qualifications but this has been adjusted to the UK equivalent. Education level is distinguished by the following categories:

- Level 1 and entry level: 1 to 4 GCSEs grade A* to C , Any GCSEs at other grades, O levels or CSEs (any grades), 1 AS level, NVQ level 1, Foundation GNVQ, Basic or Essential Skills

- Level 2: 5 or more GCSEs (A* to C or 9 to 4), O levels (passes), CSEs (grade 1), School Certification, 1 A level, 2 to 3 AS levels, VCEs, Intermediate or Higher Diploma, Welsh Baccalaureate Intermediate Diploma, NVQ level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First or General Diploma, RSA Diploma

- Level 3: 2 or more A levels or VCEs, 4 or more AS levels, Higher School Certificate, Progression or Advanced Diploma, Welsh Baccalaureate Advance Diploma, NVQ level 3; Advanced GNVQ, City and

Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma

- Level 4 or above: degree (BA, BSc), higher degree (MA, PhD, PGCE), NVQ level 4 to 5, HNC, HND, RSA Higher Diploma, BTEC Higher level, professional qualifications (for example, teaching, nursing, accountancy)
- Other qualifications

3.5.3 Employment

Determined by whether or not a respondent aged 16 or over was economically active between 15 March and 21 March 2021. Being unemployed but looking for work that could start within two weeks, or waiting to start a job that had been offered and accepted counts as being economically active for this survey. Crucially, the category of economically inactive has been separated depending on circumstance. 'Economically Inactive Other' accounts for persons who are not in education, retired, looking after family/home or long-term sick/disabled.

3.6 Data to measure social factors

3.6.1 Ethnicity and Religion

The Census provides multiple categories for ethnicity and religion. I have included the 'not answered' category for religion.

3.6.2 Residency Duration and Tenure

Length of residence is applicable to those not born in the uk and shows how long they have lived in a given area. Tenure categorizes whether a home is owned (with or without a mortgage), has shared ownership, is socially or privately rented or is lived in rent free,

3.6.3 Year of Arrival in the UK

The year someone not born in the UK last arrived in the UK. This does not include returning from short visits away from the UK.

3.7 Variables not accounted for

3.7.1 Number of Reported Incidents of Urban Decay

The 'Fixmystreet' website is a reporting platform which allows the public to report local issues (including urban decay like broken windows, graffiti and broken street lights) to the local council and is therefore the best source for realizing data for an urban decay variable. Unfortunately there are no extensive historic reports by LSOA for this data and therefore it cannot be included in the dataset.

3.7.2 Excluded Social Variables

I could not include Number of Mental Health Centres because they are too sparsely distributed relative to the size of a LSOA. For example, a single centre might serve an entire city, meaning one LSOA could have one centre while dozens of surrounding LSOAs have none—despite residents in all areas having similar access. Divorce Rate was also an impractical statistic to implement as, due to its low counts in small population, the data it produces is highly variable so statistically unreliable.

3.7.2.1 Additionally, police.uk doesn't disclose whether or not a crime logged was gang related in the historical archives so we can't quantify this gang-crime correlation. Finally, data on number of pubs per LSOA wasn't readily available at the time of data collection and Temperature varies too much with time and location to be a meaningful factor as crime data and survey conduction happen over the course of multiple days.

3.7.3 Excluded Data That Could Influence Opportunity For Crime

Due to the need for data at a highly granular spatial level (specific to individual LSOAs) I was only successful in finding population using Census data. I was unsuccessful in finding detailed and consistent reports on 'Conviction Rate' and 'Number of Active Patrol Units'. Most available conviction or court efficiency data and police budget/resources are aggregated at the police force or local authority level, which lacks the spatial resolution required for this project. Similarly, the variable 'Metres of Road' and 'Number of Surveillance Systems' were excluded from the final model as suitable data at the LSOA level was not readily available.

3.8 Temporal Considerations and Summary

This section combines 2 datasets: Police.uk crime data from the month of December 2018 and Census 2021 data. Census data offers the most detailed demographic and socioeconomic information at the LSOA level but only comes out every 10 years. The closest and most up to date crime data is 3 years prior to Census data survey date. This temporal misalignment will be a small concern for the overall results of the project and efforts should be made at a future date to match the crime data dates. For now we can assume that key structural characteristics of Areas change slowly over time and therefore remain valid for comparing to historical crime data.

3.8.0.1 Also worth noting: the aim of this project is not to forecast future crime trends over time, but rather to investigate which community-level characteristics are most strongly associated with crime a given single point in time and understand fully why different areas fluctuate in crime. For this reason it doesn't matter that census data has low temporal frequency as time is not a variable we will be considering.

4 Modeling

This section provides a rigorous mathematical foundation for modeling count data whilst considering overdispersion using generalised linear modeling techniques. It draws from standard treatments of generalized linear models and statistical learning theory [27-32]

4.1 Classical Linear Model

To begin with we can consider a Classical Linear Model (CLM) which assumes that the response variable Y is continuous and errors ε_i are normally distributed

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In matrix form:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ minimizes the residual sum of squares:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2$$

Solving this yields the closed-form solution:

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$$

In R, we can fit a CLM using:

```
lm_model <- lm(Y ~ ., data = dataset)
summary(lm_model)
```

The validity of CLM relies on the assumptions that:

- There is a linear relationship between predictors and response
- Each observation is independent of one another
- The variance of the errors across all levels of predictors remain constant (homoscedasticity)
- The errors are normally distributed,
- There is no multicollinearity between predictors (predictors aren't linearly dependent).

While useful for continuous outcomes, the assumptions of normality and constant variance (homoscedasticity) do not hold for count data as the variance often increases with the mean. Furthermore, CLM predictions are unbounded and can produce negative fitted values, which are not meaningful in the context of non negative response variables. Adjusting this model by transitioning into a more flexible framework of generalised linear models will allow us to address these limitations.

4.2 Generalised Linear Models

Generalised linear models (GLMs) generalise classical linear models (CLMs) by relaxing the assumptions of normally distributed errors and constant variance. This becomes useful when the response variable is represented as count data or binary outcomes for example which are non-continuous. By linking the expected value of the response to a linear predictor through a link function, a wider class of response distributions can be accomodated for. A GLM is composed of three key components:

4.2.0.1 Random component: Each response Y_i follows a distribution from the exponential (dispersion) family. Its probability density function can be written as:

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right\}$$

where θ_i is the canonical parameter, ϕ is the dispersion parameter, and $a(\cdot), b(\cdot), c(\cdot)$ are known functions that would define a specific distribution.

4.2.0.2 Systematic component: A linear predictor is formed from covariates x_i:

$$\eta_i = x_i^\top \beta$$

Here x_i is the vector of covariates for observation i and β is

4.2.0.3 Link function: The mean value of the response $\mu_i = \mathbb{E}[Y_i]$ is connected to the linear predictor through a smooth, invertible link function $g(\cdot)$:

$$g(\mu_i) = \eta_i$$

The variance of Y_i depends on its mean via:

$$\text{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)\nu(\mu_i)$$

Here, $\nu(\mu_i)$ is called the *variance function*, and it differs depending on the chosen distribution.

4.2.1 Poisson GLM

We now specialise this framework to the Poisson distribution, commonly used for modelling count data.

4.2.1.1 Exponential Family Form: The Poisson distribution for $Y_i \in \mathbb{N}_0$ with mean μ_i has the probability mass function:

$$P(Y_i = y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

This can be written in exponential family form by noting:

$$\begin{aligned}
\theta_i &= \log(\mu_i), \quad \text{(canonical parameter)} \\
b(\theta_i) &= e^{\theta_i} = \mu_i \\
a(\phi) &= 1, \quad \text{(Poisson has fixed dispersion)} \\
c(y_i; \phi) &= -\log(y_i!)
\end{aligned}$$

Hence, the Poisson density becomes:

$$f(y_i|\theta_i) = \exp\left\{y_i\theta_i - e^{\theta_i} - \log(y_i!)\right\}$$

4.2.1.2 Mean and Variance Functions: From the general exponential family result:

$$\mathbb{E}[Y_i] = b'(\theta_i) = e^{\theta_i} = \mu_i, \quad \mathrm{Var}(Y_i) = b''(\theta_i) = e^{\theta_i} = \mu_i$$

4.2.1.3 Link Function: The canonical link function is the **log link**, mapping μ_i to η_i (ensuring the mean never goes below 0):

$$g(\mu_i) = \log(\mu_i) = \eta_i \Rightarrow \mu_i = e^{\eta_i}$$

4.2.1.4 Log-Likelihood Function: We assume the data $y_1, \dots, y_n$ is drawn independently from the poisson distribution and the likelihood function measures how likely it is that the model (with a given β) could have generated the observed data. For n independent observations, the log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \cdot x_i^\top \beta - e^{x_i^\top \beta} - \log(y_i!)\right]$$

4.2.1.5 Score Function: To estimate β, we take the derivative with respect to each β_j of our log-likelihood function to find a maximum.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij}$$

Setting this equal to zero gives the score equations:

$$X^\top (y - \mu) = 0$$

However, since $\mu_i = e^{x_i^\top \beta}$, this system is nonlinear and must be solved iteratively.

4.2.1.6 Estimation via IRLS (Iteratively Reweighted Least Squares): To solve $X^\top(y-\mu)=0$, GLMs use the IRLS algorithm. At iteration r, define:

$$\begin{aligned} \mu_i^{(r)} &= e^{x_i^\top \beta^{(r)}} \\ w_i^{(r)} &= \mu_i^{(r)} \\ z_i^{(r)} &= \eta_i^{(r)} + \frac{y_i - \mu_i^{(r)}}{\mu_i^{(r)}} \quad \text{(working response)} \end{aligned}$$

Then update:

$$\beta^{(r+1)} = (X^\top W X)^{-1} X^\top W z$$

where $W = \operatorname{diag}(w_1, \ldots, w_n)$ and z is the working response vector.

This algorithm continues until convergence where $\beta^{(\text{final})} = \hat{\beta}$ (typically when changes in β are below a specified tolerance). $\hat{\beta}$ is called the Maximum Likelihood Estimator (MLE).

4.2.1.7 Variance of Estimates: The variance-covariance matrix of the estimated coefficients will assess their uncertainty. We can find this matrix by calculating $\mathcal{I}(\beta)$ -the Fisher Information Matrix, defined as:

$$\mathcal{I}(\beta) = \mathbb{E}\left[-\frac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^\top}\right] = X^\top W X$$

where W is the diagonal matrix of weights, whose i-th diagonal element is:

$$w_i = \left(\frac{1}{\operatorname{Var}(Y_i)}\right)\left(\frac{d\mu_i}{d\eta_i}\right)^2$$

Which simplifies to $w_i = \mu_i$ in the Poisson case as $Var(Y_i) = \mu_i$ and $\frac{d\mu_i}{d\eta_i} = \mu_i$.

Then because the asymptotic distribution of the MLE is:

$$\hat{\beta} \overset{approx}{\sim} \mathcal{N}(\beta,\ \mathcal{I}(\beta)^{-1})$$

we get:

$$\operatorname{Var}(\hat{\beta}) = (X^\top W X)^{-1}$$

This can be used to construct confidence intervals and test statistics for the regression coefficients.

4.2.1.8 Overdispersion: The Poisson GLM assumes that the mean and variance of the response are equal $\operatorname{Var}(Y_i) = \mu_i$ however real world count data like in our case may violate this assumption as most likely $\operatorname{Var}(Y_i) > \mu_i$ which would imply there is overdispersion. This leads to incorrect analysis of data so must be addressed with a new model. A diagnostic measure for overdispersion is the dispersion parameter:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where n is the number of observations, p is the number of parameters (including the intercept), y_i is the observed response, and $\hat{\mu}_i$ is the fitted mean under the Poisson model. We use this pearson estimate for our overdispersion test (score test) which uses the test statistic:

$$Z = \frac{\hat{\phi} - 1}{\sqrt{2/n}} \sim \mathcal{N}(0,1)$$

Our hypothesis:

$$H_0 : \phi = 1 \quad \text{vs} \quad H_1 : \phi > 1$$

In R this is calcualted in the following code:

```
library(AER)
m2 = glm(y ~ x, family = poisson)
dispersiontest(m2)
```

4.2.2 Quasi-Poisson

The Quasi-Poisson model extends the Poisson GLM to account for overdispersion by relaxing the equidispersion assumption $\text{Var}(Y_i) = \mu_i$. Instead, the Quasi-Poisson assumes $\text{Var}(Y_i) = \phi\mu_i$ where $\phi > 1$ is an unknown dispersion parameter estimated from the data (the pearson estimate from above). ϕ isn't estimated via maximum likelihood since the quasi-likelihood does not correspond to a fully specified probability distribution but the parameter estimates $\hat{\beta}$ are obtained using the same IRLS method and link function as Poisson. Additionally, ϕ cancels from the score equations leaving the same $\hat{\beta}$ estimate as in Poisson. Crucially, the standard errors of $\hat{\beta}$ are scaled by ϕ:

$$\text{Var}(\hat{\beta}) = \phi(X^\top W X)^{-1}.$$

In R, the dispersion parameter $\hat{\phi}$ is returned via:

```
m3 = glm(y ~ x, family = quasipoisson)
summary(m3)$dispersion
```

This value is the Pearson estimate:$\hat{\phi}$ which directly influences the reported standard errors, confidence intervals, and test statistics in the Quasi-Poisson summary output. It provides a data-driven correction for overdispersion, even though no full likelihood is specified in the quasi-likelihood framework.

4.2.3 Negative Binomial

The Negative Binomial model deals with overdispersion by assuming the variance of the response variable increases quadratically with the mean.

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

Here $\theta > 0$ is the dispersion parameter which controls amount of overdispersion. As this parameter tends to infinity the Negative Binomial model converges to the Poisson Model so a low $\theta >$ would imply greater dispersion

4.2.3.1 Exponential Family Form: The probability mass function for $Y_i \sim \text{NB}(\mu_i, \theta)$ is:

$$P(Y_i = y_i) = \frac{\Gamma(y_i + \theta)}{y_i!\,\Gamma(\theta)} \left(\frac{\mu_i}{\mu_i + \theta}\right)^{y_i} \left(\frac{\theta}{\mu_i + \theta}\right)^{\theta}$$

4.2.3.2 Canonical Link: The NB model assumes the same canonical link as the Poisson:

$$\eta_i = \log(\mu_i) = x_i^\top \beta \quad \Rightarrow \quad \mu_i = \exp(x_i^\top \beta)$$

4.2.3.3 Log-Likelihood Function: The log-likelihood for n independent observations is:

$$\ell(\beta, \theta) = \sum_{i=1}^{n} \left[\log\Gamma(Y_i + \theta) - \log\Gamma(\theta) - \log(Y_i!) + \theta \log\left(\frac{\theta}{\theta + \mu_i}\right) + Y_i \log\left(\frac{\mu_i}{\theta + \mu_i}\right)\right]$$

4.2.3.4 Score Function for β: Taking derivatives with respect to β_j, and using the chain rule:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (Y_i - \mu_i^*) x_{ij}$$

where $\mu_i^* = \mu_i \left(\frac{Y_i + \theta}{\mu_i + \theta}\right)$. In practice, this simplifies to a quasi-score:

$$\frac{\partial \ell}{\partial \beta} = X^\top (y - \mu)$$

just as in Poisson GLM, but μ_i is estimated via maximum likelihood considering the NB variance.

4.2.3.5 Estimation via IRLS: The NB model can be fitted using a variant of the IRLS algorithm by modifying the weight matrix to reflect the NB variance:

$$w_i^{(r)} = \left(\frac{1}{\mathrm{Var}(Y_i)}\right)\left(\frac{d\mu_i}{d\eta_i}\right)^2 = \left(\frac{1}{\mu_i^{(r)} + \frac{(\mu_i^{(r)})^2}{\theta}}\right) \cdot (\mu_i^{(r)})^2$$

Using this, define the working response:

$$z_i^{(r)} = \eta_i^{(r)} + \frac{y_i - \mu_i^{(r)}}{g'(\mu_i^{(r)})} = \eta_i^{(r)} + \frac{y_i - \mu_i^{(r)}}{\mu_i^{(r)}}$$

as in Poisson, since $g(\mu_i) = \log(\mu_i)$.

Then, update:

$$\beta^{(r+1)} = (X^\top W^{(r)} X)^{-1} X^\top W^{(r)} z^{(r)}$$

4.2.3.6 Variance of Estimates: As in the Poisson case, the variance-covariance matrix of the coefficient estimates $\hat{\beta}$ in the Negative Binomial model is derived from the observed information matrix. However, the variance structure differs due to the overdispersion parameter θ. The asymptotic variance of $\hat{\beta}$ is given by:

$$\mathrm{Var}(\hat{\beta}) = \left(X^\top W X\right)^{-1}$$

where the weight matrix W (from paragraph above) has diagonal entries:

$$w_i = \frac{\mu_i}{1 + \mu_i/\theta}$$

4.3 Model Evaluation

Once a Poisson or Negative Binomial GLM is fitted, its performance can be evaluated using the log-likelihood at the maximum likelihood estimate (MLE), the Akaike Information Criterion (AIC), and Nagelkerke's pseudo-R^2.

4.3.0.1 Akaike Information Criterion (AIC): AIC evaluates model fit while penalizing model complexity. It is defined as:

$$\mathrm{AIC} = -2\ell(\hat{\beta}) + 2k,$$

Where $\ell(\hat{\beta})$ is the loglikelihood for a Poisson or Negative Binomially distributed response at the MLE $\hat{\beta}$ (that was estimated using IRLS) and k is the number of parameters (Note: there will always be one more parameter in the Negative Binomial case to factor in the dispersion parameter θ). A smaller AIC indicates a better model (fit vs. complexity trade-off).

4.3.0.2 Nagelkerke's R^2: Nagelkerke's pseudo-R^2 is the proportion of variance explained by the model so can be used as a measure of fit (similarly to R^2 in linear regression but adapted to models which don't have normally distributed errors) . It is computed from the likelihoods of the null model ($\mathcal{L}_0$) and the fitted model ($\mathcal{L}_1$):

$$R^2 = \frac{1-\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right)^{2/n}}{1-\mathcal{L}_0^{2/n}} = \frac{1-\exp\left(\frac{2}{n}(\ell_0-\ell_1)\right)}{1-\exp\left(\frac{2}{n}\ell_0\right)}.$$

Here ℓ_0 is the log-likelihood of the null model (only includes the intercept) and ℓ_1 is the log-likelihood for the fitted. The numerator computes how much better the full model fits relative to the null model. This statistic ranges from 0 to 1 (as the denominator serves to normalize the result) and offers an interpretable measure of model fit where 1 would suggest the model perfectly explains the response variable. In R , AIC and Nagelkerke's R^2 for Negative Binomial is returned via:

```
m4 = glm.nb(y ~ x)
out4a = summary(m4)$aic
out4b = nagelkerke(m4)
```

5 Data Analysis

The aim of this section is to firstly establish which GLM to use, appropriately deal with the Area grouping variable, reduce variables to find a parsimonious model, and finally optimize the model.

5.1 Model Assessment

After developing a rigorous theoretical foundation for Poisson, Quasi-Poisson, and Negative Binomial models, we now apply these frameworks to our real crime dataset for Greater Manchester. With the response variable of `Crime`, and the census-derived predictor variables, we fit the following four models:

5.1.0.1 CLM(4.1): Classical linear model via OLS (ordinary least squares).

5.1.0.2 Poisson GLM(4.2.1): Assumes $\mathrm{Var}(Y_i) = \mu_i$.

5.1.0.3 Quasi-Poisson GLM(4.2.2): Assumes $\mathrm{Var}(Y_i) = \phi\mu_i$, correcting for overdispersion.

5.1.0.4 Negative Binomial GLM(4.2.3): Allows $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2/\theta$, estimating θ from the data.

5.1.0.5 The table below summarizes key model evaluation statistics extracted from the R output.

Model	**AIC**	**Nagelkerke R^2**	**Dispersion**	**Theta (NB)**
CLM (OLS)	—	0.308	—	—
Poisson GLM	18374	0.99999	7.144	—
Quasi-Poisson GLM	—	0.99999	7.858	—
Neg. Binomial GLM	12181	0.5189	—	2.6487

Table 1: Model diagnostic statistics for OLS, Poisson, Quasi-Poisson, and Negative Binomial regressions.

5.1.1 Results of The Overdispersion Test (Score test)

The test run on our poisson model showed $Z = 8.50$, and $p < 2.2 \times 10^{-16}$ which is a low enough probability to reject $H_0 : \phi = 1$ and conclude that there is significant overdispersion. A model which doesn't rely on $\mathrm{Var}(Y_i) = \mu_i$ is therefore required.

5.1.2 Interpretation and Model Choice

The Poisson model yields a dispersion estimate far greater than 1 ($\hat{\phi} = 7.14$), indicating substantial overdispersion and violating the equidispersion assumption $\mathrm{Var}(Y_i) = \mu_i$ which aligns with the results of the score test. The Quasi-Poisson model appropriately corrects for this, inflating the standard errors using a Pearson-based dispersion estimate $\hat{\phi} = 7.86$, but lacks a true likelihood and therefore cannot be compared using AIC.

The Negative Binomial model accommodates overdispersion structurally by introducing an additional parameter θ, and outperforms other models in terms of AIC (12181, lowest among valid models). Although the Nagelkerke R^2 is slightly lower than the Poisson model, the latter's near-perfect R^2 is misleading due to poor variance assumptions.

5.1.2.1 Conclusion: Based on overdispersion correction, information criteria, and model interpretability, the **Negative Binomial GLM** is the most appropriate choice for modelling crime counts in Greater Manchester. Considering real world context, If we looked at a high average crime rate urban area, we'd observe more variance in crime count month to month due to the unpredictable nature of densely populated environments. In contrast, a rural area would see a lower mean crime rate but more consistent monthly counts. This disparity reflects overdispersion (where the variance exceeds the mean), a feature that the Negative Binomial model is specifically designed to handle.

5.2 Exploratory Data Analysis

I performed exploratory data analysis to gain a visual and statistical intuition of key relationships within the raw data before proceeding with detailed variable analysis.

5.2.1 Response Variable

In the previous section we concluded what model was the best fit. This can also be visualised in a Histogram of our raw data with an overlay of intercept only poisson and negative binomial model probability mass functions where we see that the green negative binomial line more accurately emulates the shape of the crime count histogram. Note, for readability, Deansgate and other disproportionally high crime areas were excluded from this histogram. (see Appendix 6.1.0.11 for code).

5.2.1.1

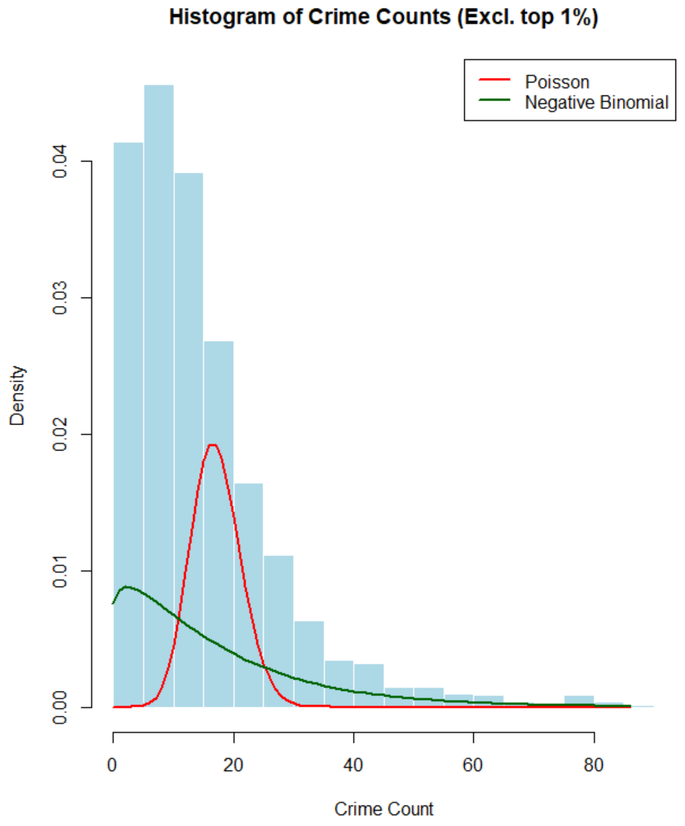

Figure 1: Histogram of Crime Counts

5.2.2 Area

Boxplots revealed moderate differences in average crime rates across districts in Greater Manchester (see Appendix 6.1.0.11 for code). This could suggest that districts have a structural role in shaping crime patterns. Potentially this variation reflects latent regional factors that weren't accounted for in the data such as police funding or access to mental healthcare services. To account for district level influences I considered a categorical fixed effect variable for Area when fitting my models.

5.2.2.1 Mathematically, using a fixed effect variable for Area would assign each district its own intercept so the model can estimate baseline crime levels that are exclusive to each district. For observation i, if we let A_i be the district it belongs to then the linear predictor becomes:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \gamma_{A_i}$$

Here γ_{A_i} is a fixed effect term which captures how the intercept shifts for that district. Controlling for unobserved district characteristics that influence crime rate by isolating baseline differences avoids attributing district level variation to census variables. This improves the validity of the estimated effects for the explanatory variables and so from this point forwards, I will treat the grouping variable 'Area' as a fixed effect in every model I fit.

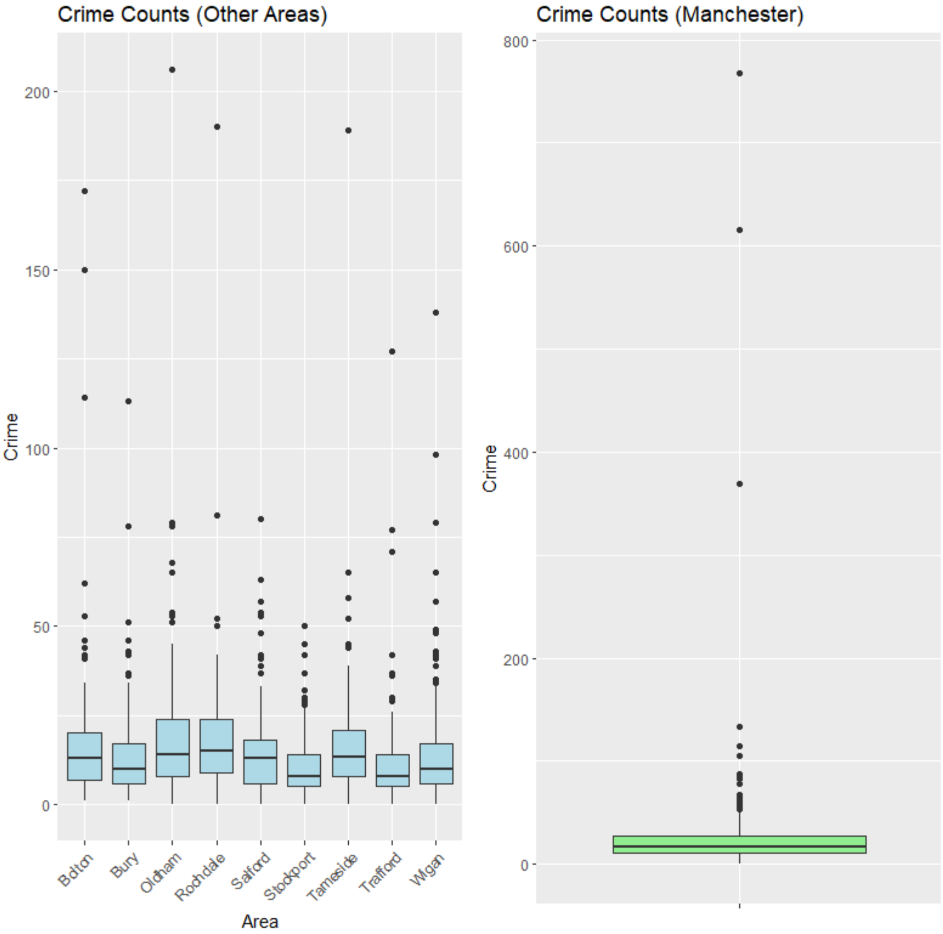

Figure 2: Boxplot of crime count per district

5.2.2.2

5.2.2.3 Here is a spatial heat map of observed crime in Greater Manchester, showing areas with a higher crime rate in a darker shade (see Appendix 6.1.0.11 for code):

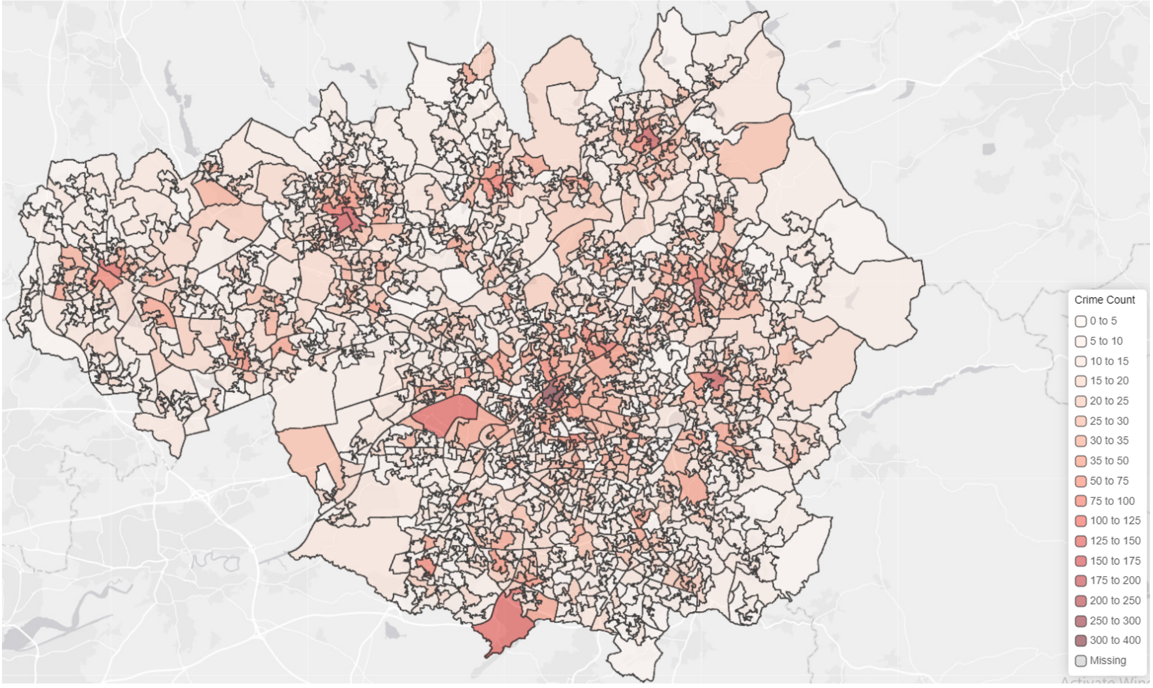

Figure 3: Crime distribution map of Greater Manchester

5.2.3 Predictors

I examined pairwise correlations among the predictor variables to gain an insight into potential multicollinearity issues that will be addressed later on. Upon computing a correlation matrix for every explanatory variable (taken as a percentage of the population as opposed to raw data), I flagged pairs of variables with strong correlations $|\rho| > 0.9$. In R:

```
cor_matrix <- cor(DS2_prcnt, use = "complete.obs")
```

Variable 1	**Variable 2**	**Correlation**
Household.not.deprived	No.qualifications	-0.9186
Born.in.the.UK...137	Born.in.the.UK...153	0.99996
Household.not.deprived	Household.is.deprived.in.two.dimensions	-0.9640
No.qualifications	Household.is.deprived.in.two.dimensions	0.9021
Economically.(excluding students)	Economically.inactive	-0.9643
Asian..Asian.British.or.Asian.Welsh	White	-0.9129
Born.in.the.UK...137	White	0.9160
Born.in.the.UK...153	White	0.9160
Asian..Asian.British.or.Asian.Welsh	Muslim	0.9699
White	Muslim	-0.9372
White	10.years.or.more	-0.9490
Born.in.the.UK...137	5.years.or.more..but.less.than.10.years	-0.9030
Born.in.the.UK...153	5.years.or.more..but.less.than.10.years	-0.9028
White	Arrived.1991.to.2000	-0.9015
10.years.or.more	Arrived.1991.to.2000	0.9241
5.years.or.more..but.less.than.10.years	Arrived.2011.to.2013	0.9512
5.years.or.more..but.less.than.10.years	Arrived.2014.to.2016	0.9614
2.years.or.more..but.less.than.5.years	Arrived.2014.to.2016	0.9098
2.years.or.more..but.less.than.5.years	Arrived.2017.to.2019	0.9699
Less.than.2.years	Arrived.2020.to.2021	0.9802
Private.rented.or.lives.rent.free	Private.rented	0.9997

Table 2: Highly correlated variable pairs ($|correlation| > 0.9$)

Also, 4 variables which, based on my logical reasoning, would be good potential predictors for crime are: 16–19-year-olds, the percentage of economically inactive individuals (other), households with only 2 rooms, and households that are privately rented or lived in rent-free. We can see that these variables aren't linearlly correlated with this interaction model (see Appendix 6.1.0.11 for code):

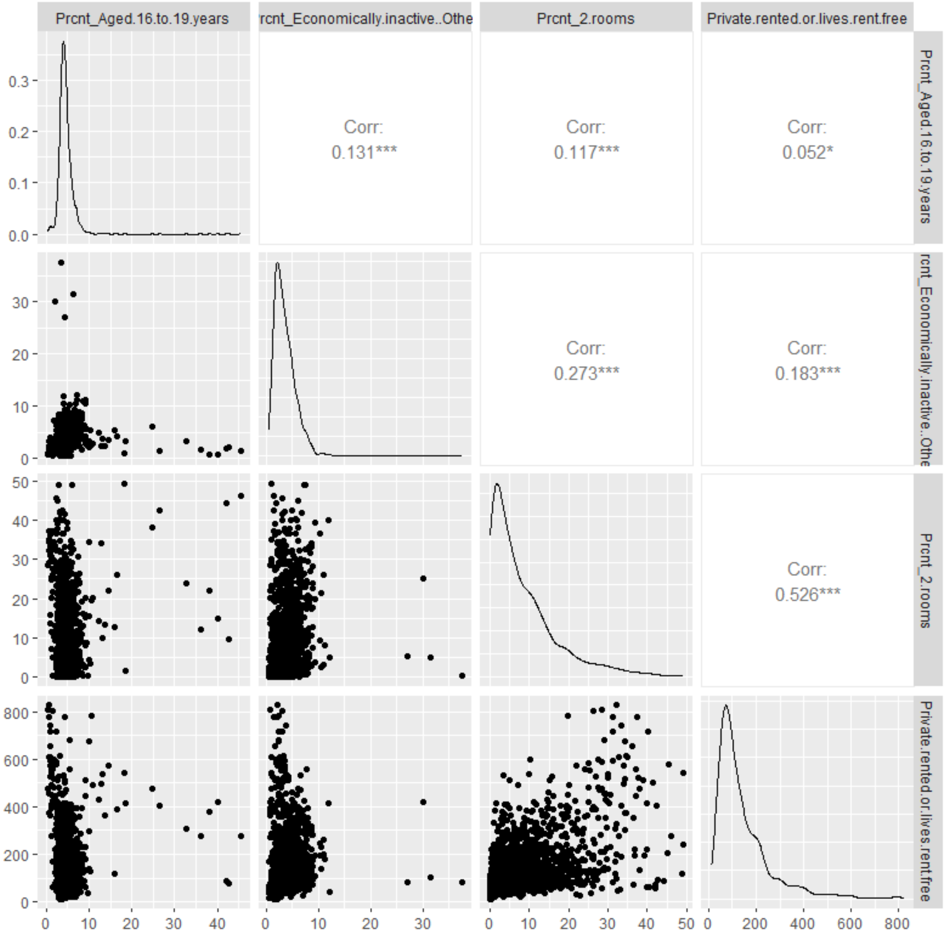

Figure 4: scatterplot matrix

5.2.3.1

5.3 Variable Assessment

5.3.1 Wald Test

Given our selection of the appropriate GLM - Negative Binomial, a good place to start assessing our covariates would be to see which variables contribute significantly to explaining the response in our full model. In R, I used the function:

```
summary(m4)$coefficients
```

which performs the Wald test on each estimated coefficient β_j. The Wald test determines if the deviation of each regression coefficient β_j from 0 is significant. If it is, then the corresponding covariate contributes in some way to explaining the response variable. If not, then the variable has no effect ($\beta_j = 0$).

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

5.3.1.1 Test Statistic:

The Wald test statistic is given by:

$$Z_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

where $\text{SE}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)}$ is the standard error of β_j

5.3.1.2 p-value:

If we assume asymptotic normality of $\hat{\beta}$, the test statistic under H_0 approximately follows the standard Normal distribution $z_j \sim \mathcal{N}(0, 1)$ we can calculate the p-value as:

$$p_j = 2 \cdot P(Z > |z_j|) = 2 \cdot (1 - \Phi(|z_j|))$$

Which tells us how likely we'd observe a coefficient this far from 0 (in either direction) assuming the Null hypothesis were true and this variable had no real effect. (ϕ is the cumulative distribution function of the standard normal distribution).

	Estimate	**Std. Error**	**z value**	**Pr**$(> \|\mathbf{z}\|)$
Arrived.1981.to.1990	-0.05442	0.01860	-2.92567	0.00344
Arrived.2011.to.2013	-0.05245	0.01810	-2.89727	0.00376
% Arrived.1981.to.1990	0.80354	0.29189	2.75291	0.00591
Economically.inactive..Other	0.01319	0.00488	2.70155	0.00690

Table 3: Wald Test Output for the Covariates with lowest p-value

5.3.2 Univariate

Looking at a variable in isolation allows us to test its marginal explanatory power (its individual ability to explain crime variation when considered alone). Importantly, doing so will alleviate the effects of multicollinerity among covariates so hidden significant variables can be identified. An efficient method of testing every variable is with a univariate model loop in R which fits every explanatory variable individually with every model mentioned above and can be found in the appendix. Appendix 6.1.0.11

5.3.2.1 Mathematically, each univariate model still follows the generalised linear model structure but we are now looking at one singular predictor in isolation so the linear predictor η_i will only contain a singular explanatory variable x_i

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i$$

5.3.2.2 Like mentioned previously, to compare which univariate models performed the best, I examined the Nagelkerke R^2 statistic from the negative binomial generalised linear models. Ordering every model from largest R^2 to smallest offers insight into a variables stand alone power of explaining the variance in the response variable (crime), with the best predictors having the highest statistic as shown in this table:

	Linear Model	Poisson		Quasi Poisson	Negative Binomial		
Variable	Adj R^2	AIC	Psdo R^2	Disp.	AIC	Psdo R^2	Disp.
% Owned	0.0920	29429.61	0.9921	27.5535	12443.42	0.3345	1.8334
2 rooms	0.1290	29278.73	0.9928	22.6234	12610.64	0.2658	1.6604
% Owns with a mortgage or loan or shared ownership	0.0772	30833.06	0.9821	31.3862	12629.21	0.2577	1.6356
% 2 rooms	0.1028	30271.70	0.9871	22.5065	12656.62	0.2456	1.6110
Private rented	0.0960	31259.16	0.9770	23.88705	12693.19	0.2292	1.5717
3 rooms	0.0753	31606.31	0.9718	32.1897	12694.88	0.2285	1.5680
Aged 25 to 34 years	0.0705	32712.54	0.9459	30.4732	12702.64	0.2249	1.5547
% Arrived 2017-2019	0.0759	32330.10	0.9568	26.8734	12737.67	0.2088	1.5266
Arrived 2017-2019	0.0729	32690.09	0.9466	32.8171	12742.84	0.2064	1.52047

Table 4: Comparison of model metrics across LM, Poisson, Quasi-Poisson, and Negative Binomial GLMs for selected individual variables. Note that Adj R^2 refers to Adjusted R^2, "Psdo R^2" refers to Nagelkerke pseudo R^2 and "Disp." refers to (estimated) dispersion parameter.

5.3.3 Multivariate

While univariate models are useful, they overlook the real possibility that many explanatory variables can be statistically related. To better understand how groups of related predictors contribute to explaining our crime count, I decided to fit census categories together in a multivariate model. The following table shows the highest performing categories based on the Nagelkerke R^2 under the negative binomial framework. The table shows that tenure and residency duration are contributing more significantly (relative to other groups) to explaining crime variation.

	Linear Model	Poisson		Quasi Poisson	Negative Binomial		
Variable	Adj R^2	AIC	Psdo R^2	Disp.	AIC	Psdo R^2	Disp.
Tnr_cnt	0.1383	27123.93	0.9980	22.2108	12310.92	0.3880	2.0266
Res_cnt	0.1569	26443.06	0.9987	17.1664	12383.26	0.3636	1.9467
Tnr_prcnt	0.1045	28727.10	0.9948	22.4996	12391.62	0.3575	1.9084
Res_prcnt	0.1264	27670.36	0.9972	16.9472	12444.53	0.3387	1.8594
Dep_cnt	0.0915	30334.29	0.9867	49.9134	12624.51	0.2632	1.6482
Age_cnt	0.0744	31564.35	0.9728	28.8985	12656.94	0.2543	1.6203
Arvl_prcnt	0.0919	31324.68	0.9764	24.0273	12680.94	0.2437	1.6015
Durn_cnt	0.0850	32004.03	0.9645	32.7037	12674.08	0.2414	1.5939
Durn_prcnt	0.0922	31475.61	0.9739	24.7798	12679.63	0.2381	1.5893

Table 5: Comparison of model metrics across LM, Poisson, Quasi-Poisson, and Negative Binomial GLMs for selected variable groups. cnt = raw count prcnt = %

5.3.3.1 Individual Variable performance within the Group Modelling Within each group model, Wald tests were also performed on each variable to identify which predictors were responsible for the groups overall performance. We can see here that within the economic activity group, 'Economically inactive other' is the most statistically significant.

Variable	Estimate	Std. Error	z-value	p-value	Wald Statistic NB
(Intercept)	1.7368	0.1306	13.2993	2.34E-40	176.87
Economically inactive Other	0.0128	0.0006	20.0991	7.52E-90	403.97
Economically active and a full time student	0.0047	0.0005	8.6856	3.77E-18	75.43
Economically active excluding full time students.	0.0006	0.0001	5.0919	3.54E-07	25.93
Economically inactive	-0.0005	0.0002	-2.5859	0.0097	6.68

Table 6: Wald test results for employment-related covariates in the `Emp_cnt` group model.

5.4 Variable Selection

Given that the current dataset has over 100 covariates, it is essential to identify which predictors meaningfully contribute to explaining the variation in crime rates across Greater Manchester. Including all variables in the model may result in several issues:

5.4.0.1 Multicollinearity: Occurs when 2 or more explanatory variables in the regression model are highly linearly correlated which results in difficulties when isolating individual effects of each variable. This can contribute to unstable coefficient estimates meaning small changes in data could lead to a large difference in the estimated coefficients ($\hat{\beta}$). Furthermore, multicollinearity increases standard errors so variables appear to be statistically insignificant despite actually being relevant to the model.

5.4.0.2 Overfitting: A model with too many variables will pick up random noise in the data (like a fluke variation unrelated to crime rate) rather than the underlying relationship. This results in a model that despite fitting the dataset well, will generalize poorly to new data and consequently have no predictive power.

5.4.0.3 Interpretability and Computational Complexity: When dealing with a large number of variables, the models produced will be challenging to understand and interpret from a human perspective and more intensive to run in R.

5.4.0.4 Consequently, the goal of this section is to identify a parsimonious model that balances explanatory power and simplicity.

5.4.1 Stepwise Model Selection

One way to reduce the amount of variables that are included in the model is to apply stepwise model selection using the Akaike Information Criterion (AIC). In R, this is performed using:

```
m5_step = stepAIC(m5, direction = "both", trace =
   TRUE)
```

This function begins with the full model `m5` (which is the negative binomial GLM fitted with every variable) and iteratively adds or removes variables to minimize the AIC. The AIC balances goodness-of-fit with model complexity so models with lower AIC values are preferred as they indicate better fit with fewer parameters (see above section on AIC). In my code, $direction = both$ allows both forward selection (adding variables) and backward elimination (removing variables) at each step. The procedure stops when no further addition or removal leads to a lower AIC. The result is the most parsimonious model (i.e., simplest) that still achieves a strong fit according to the AIC criterion.

5.4.1.1 Applying Stepwise: Applying Stepwise to the full model reduced the amount of variables in the model by 7. This is promising as it shows that our research section was actually successful in identifying variables which do in fact correlate with crime in some way. As a variable selection technique however, stepwise is not effective in finding a reduced model so the next section will detail my logical manual selection of variable process using my own analysis of variables conducted earlier.

5.4.2 Manual Variable Selection

In order to select candidate variables to include in my final model, I created a spreadsheet which included the following information about each variable:

Univariate: From the univariate model fittings, I decided to choose the Nagelkerke R^2 statistic for the negative binomial model as my first benchmark as it showed how well the individual predictor variable explained the variance in crime. If the statistic was > 0.10 for any univariate model, I considered its predictor variable to have a strong explanatory power in this context and marked the 'univariate strength' column with a 'YES' for that variable. I decided to give a 'MAYBE' to any statistic between $0.05 < R^2 \leq 0.1$ and a 'NO' otherwise.

Wald: I chose to assess the $p-value$ for every variable in my full model first. A low $p-value$ indicates that the observed effect of the variable ($\hat{\beta}_j$) is unlikely given that the null hypothesis is true. Furthermore, variables with low $p-value's$ are more likely to contribute meaningfully to explaining the response variable so if $p < 0.01$, my 'Wald Significance' column would be marked with 'YES'. If $0.01 \leq p < 0.05$ then I would assign my 'MAYBE' but everything above got a 'NO'.

5.4.2.1 After the full model variables had been assessed using $p-value$, I did the same for each of my group models (corresponding to census data categories) so I could again see which variables contributed within their category to explaining crime. I used the same thresholds for p as above to assign 'YES', 'MAYBE' or 'NO' where appropriate.

5.4.2.2 Finally, for each 'MAYBE' or 'YES' a variable received in the statistics from the 3 models, it was awarded 1 point. Then I moved all variables with at least 2 points to a reduced model containing only candidate variables. I decided to run another stepwise model selection on these selected variables and no variables were reduced implying that no further removals in variables contributed to a the smaller Akaike Information Criterion (AIC) fit.

5.4.2.3 Additionally, I used the correlation matrix to remove 1 variable out of the pair of variables which were flagged to have high linear correlation if they both appeared in the candidate list. See appendix for candidate list of variables Appendix 6.1.0.11 .

5.4.3 Lasso

The Least Absolute Shrinkage and Selection Operator (LASSO) automatically performs variable selection with our candidate variables by shrinking some coefficients to exactly zero and consequently removing their contribution in the linear predictor of the model.

5.4.3.1 Penalized log-likelihood: As previously described in the GLM section, we aim to maximise the log-likelihood function $\ell(\boldsymbol{\beta})$ to estimate our model. In Lasso, we choose to instead maximise a penalized log-likelihood:

$$\ell^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - s(\boldsymbol{\beta})$$

$s(\boldsymbol{\beta})$ is a penalty function based on the ℓ_1-norm of the coefficients so a more complex model (which will have a larger ℓ_1-norm) will receive a larger penalty:

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j| = \lambda ||\boldsymbol{\beta}||_1$$

Here, $\lambda \geq 0$ is a tuning parameter that controls the strength of the penalty. The larger the value of λ, the greater the shrinkage applied to the coefficients.

So finally, our LASSO estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\max_{\boldsymbol{\beta}} \{\ell(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_1\}$$

Which can be equivalently expressed as a minimization problem by convention:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Essentially LASSO selects variables that increase the log-likelihood (meaning making the observed data most likely and improving model fit) whilst justifying their penalty cost $\lambda|\beta_j|$. Clearly, if the penalty is larger than the log likelihood, the variable in question doesn't serve to minimize $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ so will be set to 0. This process can be visualized in the LASSO cross-validation plot which can be coded in R as:

```
library(glmnet)
cv_lasso = cv.glmnet(X, y, family = "poisson",
   alpha = 1, nfolds = 10)
plot(cv_lasso)
```

5.4.3.2 The *glmnet* package in R doesn't support the Negative Binomial model so here I temporarily fitted a Poisson GLM during the LASSO procedure.

5.4.3.3 The y axis here is 'Poisson Deviance' which is what the *glmnet* package uses to quantify model error. Since Poisson deviance is proportional to the negative log-likelihood (up to a constant), lower deviance values indicate better model fit under the Poisson assumption. The x axis is $log(\lambda)$ therefore, as we move from right to left, lambda gets smaller and less penalty is given to the model so more variables are kept in. The first vertical line in the plot marks the value of λ that minimizes deviance (known as λ_{min}) . The corresponding model produced by λ_{min} is expected to generalize best to new data, balancing fit and regularization.

5.4.3.4 Notice that as more and more variables get removed from models with $\lambda > \lambda_{min}$, the deviance starts to increase and Nagelkerke R^2 would decrease as the model becomes too simple to capture key relationships and maintain explanatory power. This illustrates the utility of LASSO as it automatically selects a parsimonious model that retains the most useful predictors whilst keeping a strong predictive performance.

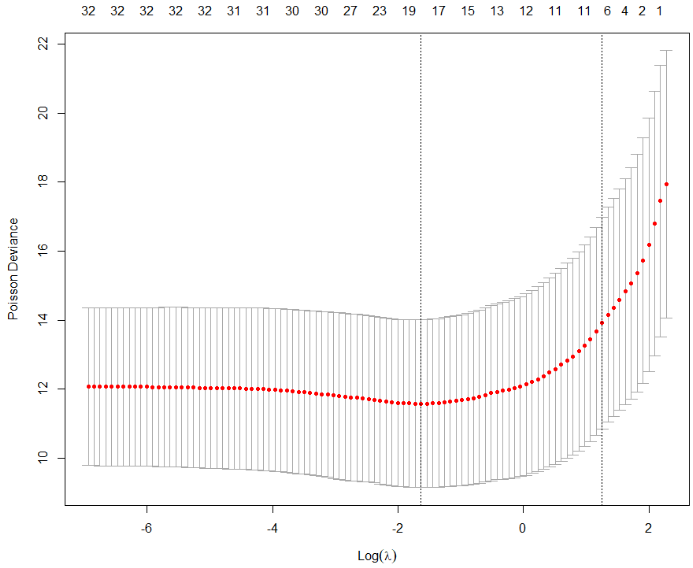

Figure 5: LASSO cross-validation plot used to select λ_{min}

5.4.3.5 Applying LASSO: After applying the LASSO algorithm to the shortlisted candidate variables, the model produced using λ_{min} determined the final set of predictors to include. These selected variables were then refitted in a Negative Binomial model to evaluate their performance. The following tables present the Wald test results for each variable and the overall fit statistics of the reduced model.

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(> |z|)**
(Intercept)	2.119743	0.234156	9.053	<2e-16 ***
% Private rented or lives rent free	0.012655	0.002920	4.334	1.46e-05 ***
% Aged 20 to 24 years	-0.011063	0.005479	-2.019	0.04350 *
% Household is deprived in two dimensions	0.017000	0.006900	2.464	0.01375 *
% Household is deprived in three dimensions	0.035318	0.013554	2.606	0.00917 **
% Household is deprived in four dimensions	0.087740	0.070122	1.251	0.21084
% Apprenticeship	0.008589	0.014965	0.574	0.56601
% Economically inactive Other	0.014235	0.010769	1.322	0.18623
% Other ethnic group	-0.018878	0.011627	-1.624	0.10445
% Buddhist	0.122964	0.065170	1.887	0.05919 .
% Other religion	0.172118	0.065811	2.615	0.00891 **
% 5 years or more but less than 10 years	0.017324	0.012921	1.341	0.18001
% 2 rooms	0.009657	0.003410	2.832	0.00463 **
% 4 rooms	-0.006301	0.002281	-2.762	0.00575 **
% 5 rooms	-0.005669	0.002141	-2.648	0.00809 **
% 6 rooms	-0.013096	0.005095	-2.570	0.01017 *
% 7 rooms	-0.016611	0.010310	-1.611	0.10715
% Arrived 2020 to 2021	0.038451	0.012715	3.024	0.00249 **
Area Bury	-0.068363	0.090804	-0.753	0.45153
Area Manchester	0.066960	0.084190	0.795	0.42641
Area Oldham	0.074229	0.084881	0.875	0.38184
Area Rochdale	0.113194	0.085356	1.326	0.18480
Area Salford	-0.161441	0.084504	-1.910	0.05607 .
Area Stockport	-0.134311	0.083960	-1.600	0.10966
Area Tameside	0.002045	0.084875	0.024	0.98078
Area Trafford	0.108745	0.093916	1.158	0.24691
Area Wigan	-0.001136	0.080710	-0.014	0.98877

Table 7: Regression Coefficients from Negative Binomial Model

Metric	Value
Adjusted R^2 (Linear Model)	0.1718
AIC (Poisson Model)	24357.55
Nagelkerke R^2 (Poisson Model)	0.9996
Dispersion (Quasi-Poisson Model)	13.8555
AIC (Negative Binomial Model)	12291.24
Nagelkerke R^2 (Negative Binomial Model)	0.4091
Theta (Negative Binomial Model)	2.1176

Table 8: Model Performance Metrics

5.5 Optimization/Refinement

Here we motivate and add second order terms a long with exploring random effects for the categorical variable to reach our final model.

5.5.1 Bivariate

To further investigate high performing variables in the reduced model, I fitted second order (bivariate) models to asses how pairs of these variables contribute to explaining crime. The advantage of fitting a bivariate model is that it reveals if a variables significance is conditional on the inclusion of a different predictor. This scenario could be common in census data for variables that quantify the same phenomena. For example, both "number of rooms" and "household deprivation" reflect aspects of regional affluence and their individual effects may be masked or inflated due to collinearity. Including them in a bivariate model helps clarify whether they jointly improve explanatory power beyond what each offers alone.

In R:

```r
m4a = tryCatch({summary(glm.nb(y ~ x*z))$aic}, error
  =function(e) NA)
```

Mathematically, we can model this by extending the linear predictor in our generalised linear model to include an interaction term between the 2 covariates:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \cdot x_{2i})$$

η_i is the linear predictor for observation i as seen in the formulation of GLM chapter above, x_{1i} and x_{2i} are the two explanatory variables used in the bivariate model, and $\beta_3(x_{1i} \cdot x_{2i})$ is the interaction term which allows the effect of 1 variable to vary depending on the value of another. This is shown by the marginal effect of x_1 varying with x_2 :

$$\frac{\partial \eta_i}{\partial x_{1i}} = \beta_1 + \beta_3 x_{2i}$$

Assessing each bivariate model based on Nagelkerke R^2, much like in the univariate case, will show us how much the pairs explain crime variance. These are the top pairs of reduced model variables:

Variable 1	**Variable 2**	**Nagelkerke NB**
% Private rented or lives rent free	% Household is deprived in three dimensions	0.3557
% Private rented or lives rent free	% Household is deprived in two dimensions	0.3412
% Household is deprived in three dimensions	% 2 rooms	0.3311
% Household is deprived in three dimensions	% Arrived 2020 to 2021	0.3299
% Household is deprived in two dimensions	% 2 rooms	0.3286

Table 9: Nagelkerke R^2 from Negative Binomial Models for Selected Interaction Terms

5.5.2 Raising Nagelkerke R^2

To further improve the explanatory power of the reduced model, I included the 5 top pairs (as found above) which had high Nagelkerke R^2 values indicating strong joint explanatory ability for crime count variation. Including these interaction terms captures conditional relationships that may be masked in additive-only models. This will enhance model fit without drastically increasing complexity.

Metric	Value
Adjusted R^2 (Linear Model)	0.1804102
AIC (Poisson Model)	24195.11
Nagelkerke R^2 (Poisson Model)	0.99965
Dispersion (Quasi-Poisson Model)	14.07090
AIC (Negative Binomial Model)	12282.00
Nagelkerke R^2 (Negative Binomial Model)	0.415739
Theta (Negative Binomial Model)	2.146191

Table 10: Model performance metrics across linear, Poisson, Quasi-Poisson, and Negative Binomial models

5.5.2.1

5.5.2.2 In this refined model, the Nagelkerke R^2 for the negative binomial model has increased to 0.416 from 0.409 which, though not significant, is an improvement on the models explanatory power. In R, the negative binomial glm with interaction terms that makes up our final parsimonious model and is coded as follows:

```
m_nb <- glm.nb(Crime ~ var1 + var2 + var3 + var1:var
    2 + var2:var3 + Area, data = DS2)
```

Here varx refers to variables which survived the LASSO algorithm found in Table 7, varx:vary refers to selected interaction terms given in Table 9

5.5.3 Testing Random Effects for Area: Final Model Refinement

This section explores an alternative way of treating the grouping variable 'Area' which specifies districts in Greater Manchester. Instead of using fixed effects to estimate an intercept for each district as I have done up to this point, random effects treat district specific intercepts as random deviations from a global mean. This approach assumes observed areas like 'Manchester' are a sample from a larger population of districts, and models their effects as being drawn from a probability distribution. The linear predictor becomes:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + b_{A_i}$$

Here, $b_{A_i} \sim \mathcal{N}(0, \sigma^2_{\text{Area}})$ is the random intercept for the district that observation i belongs to. This term now accounts for unobserved differences between regions like police activity as discussed in the previous fixed case.

5.5.3.1 Assuming district-specific deviations come from a common distribution allows for partial pooling of information across areas (so for example districts with less data borrow strength form the broader dataset).

Additionally, random effects only requires the addition of one variance parameter σ^2_{Area} so there wouldn't have to be an estimate for every district coefficient intercept like in fixed. This increases the reduced models parsimony and allows it to be more generalized.

5.5.3.2 To investigate if district-level variation in crime could be more appropriately modeled through random effects, I considered the use of a generalized linear mixed model (GLMM) for our reduced Model. In R:

```
formula_str_re <- paste("Crime ~", paste(all_terms,
   collapse = " + "), "+ (1 | Area)")
model_formula_re <- as.formula(formula_str_re)
    library(lme4)
m4_re <- glmer.nb(model_formula_re, data = DS2,
   control = glmerControl(optimizer = "bobyqa"))
summary(m4_re)
```

The output revealed that the estimated variance of the random intercept was extremely small $9.841e-12$ - indicating that once the fixed effects were included, there was negligible residual variation across districts. Additionally, AIC for the random effects model was slightly higher at 12284 compared to the fixed models 12282 suggesting no improvement on model fit.

5.5.3.3 Given the negligible random intercept variance and slightly poorer performance based on AIC, the fixed effects model is more appropriate for this final model. Fixed effects better capture the structured differences between districts, without sacrificing interpretability or performance.

6 Discussion/Conclusions

This project set out to identify the most impactful explanatory variables associated with crime variation across Greater Manchester by using a parsimonious statistical modeling framework. I have successfully managed to find a reduced model and several clear conclusions have emerged.

6.0.0.1 Limitations Firstly, as mentioned in the Data section, there is a temporal mismatch between census and crime data which serves to limit causal interpretation as we don't know how close Area predictors were in 2021 to their value in December 2018. A fundamental extension would be to source crime data from March 2021 to match the Census time period. Additionally, some collinearity between variables may influence effect size interpretation a long with the temporal mismatch.

6.1 Variables In Order Of Least To Most Positive Association

6.1.0.1 Interpretation Of The Results Table: The following table shows covariates from the reduced model (without bivariate terms for clarity) in order of estimate size $\hat{\beta}$. The larger the estimated effect size, the stronger the association the variable has to crime assuming all other variables are held constant. This association is positive when $\hat{\beta} > 0$ and negative when $\hat{\beta} < 0$. The p value provided is the metric I previously used to assess statistical significance which shows how likely said estimator is to be observed this far from the value $\hat{\beta} = 0$ assuming that the variable has no real effect. A higher p-value would imply that a variables observed effect on crime rate could have occurred by chance (random variation) hence isn't statistically significant.

Variable	Estimate	p-value
% Other.religion	0.17212	0.00891
% Buddhist	0.12296	0.05919
Area Rochdale	0.11319	0.18480
Area Trafford	0.10874	0.24691
% Household is deprived in four dimensions	0.08774	0.21084
Area Oldham	0.07423	0.38184
Area Manchester	0.06696	0.42641
% Arrived 2020 to 2021	0.03845	0.00249
% Household is deprived in three dimensions	0.03532	0.00917
% 5 years or more but less than 10.years	0.01732	0.18001
% Household is deprived in two dimensions	0.01700	0.01375
% Economically inactive Other	0.01423	0.18623
% Private rented or lives rent free	0.01265	0.00001
% 2 rooms	0.00966	0.00463
% Apprenticeship	0.00859	0.56601
% Tameside	0.00204	0.98078
Area Wigan	-0.00114	0.98877
% 5 rooms	-0.00567	0.00809
% 4 rooms	-0.00630	0.00575
% Aged 20 to 24 years	-0.01106	0.04350
% 6 rooms	-0.01310	0.01017
% 7 rooms	-0.01661	0.10715
% Other ethnic group	-0.01888	0.10445
Area Bury	-0.06836	0.45153
Area Stockport	-0.13431	0.10966
Area Salford	-0.16144	0.05607

Table 11: Model Coefficients Ranked by Strength of Association

- **Age, $\hat{\beta} = -0.011$:** A higher percentage of individuals aged 20 to 24 was negatively associated with crime. This contradicts evidence presented by Casey [4] which suggested the opposite. A possible explanation for this could be the large student population in Greater Manchester which, despite potentially being easy targets for crime, may not actually contribute to crime rates themselves due to their education level.

- **Tenure, $\hat{\beta} = 0.01265$:** The percentage of people whose tenure was categorized as rented or lives rent free was found to be very strongly statistically significant and slightly positive in correlation with crime counts. This confirms the research conducted by Ferrazzi [22] and goes to show that lower tenure security and weaker social cohesion as a result of renting property increases crime in Greater Manchester.

- **Immigration, $\hat{\beta} = 0.038$:** Percentage of people who arrived from 2020 to 2021 has a significant and positive correlation with crime. This confirms evidence brought forward by The Mitigation Observatory at the University of Oxford [24] . Possible reasons for this outcome, as discussed previously, could be due to new arrival's economic situation or their impact on social cohesion.

- **Deprivation, $\hat{\beta} = 0.08774$:** We can clearly see from the table that the more dimensions of deprivation households satisfy, the more positive the correlation with crime rate the variable becomes. 2 dimensions deprived has an estimate of 0.017, 3 dimensions has an estimate of 0.035, and 4 dimensions has an estimate of 0.088 (Note here that 4 dimensions of deprivation is not as statistically significant as the other 2 variables). Additionally we can see that if a house has 4 rooms or above, it is negatively correlated to crime but if the house has 2 rooms, there is a positive correlation. (All results here are statistically significant)

  6.1.0.2 Additionally, we can see a significant positive correlation between Economically inactive (Other) and crime ($\hat{\beta} = 0.01423$) which suggests areas where more people are out of work experience more crime.

  6.1.0.3 These results reinforces the well established link between economic hardship and increased crime, confirming the work of Graif et al. (2014) [15] also applies to us in Greater Manchester.

- **Religion, $\hat{\beta} = 0.17212$:** The percentage of people identifying with 'Other religion' had a significant and the most positive correlation with crime at an estimate of 0.17. This was followed by Buddhist which was 0.12. Before definitive interpretation of these results can be made, further investigation must be conducted to avoid misattribution but for our model, these 2 variables showed the strongest explanatory power. Note that 'Other Religion' is very broad so includes individuals with vastly different beliefs making generalizations about this variable's explanatory power unreliable.

6.1.0.4 Interpretation of Fixed Effects for Area: In the final negative binomial model, the categorical variable Area was included using fixed effects instead of random. The model estimated a separate intercept adjustment for each district. Among the area coefficients, no area variable was statistically significant at $p - value < 0.05$ which suggests that regional differences in crime rate were not significant on their own once other demographic and contextual variables were accounted for. This essentially means that geographic districts didn't explain additional variation so district-specific latent regional factors that we couldn't quantify in the Data section like police activity didn't appear to explain additional variation in crime.

6.1.0.5 Summary of Variable Findings and Extensions: In conclusion, this project has found 'Other Religion' to have the most positive correlation with crime rate in Greater Manchester. This variable requires further investigation to understand why it has shown up to be this significant in order to be acted upon. The variable with the highest statistical significance whilst also showing a strong positive correlation was '% Arrived between 2020 and 2021'. Understanding the reasons behind this correlation (such as inadequate integration mechanisms challenging social cohesion) and ways to prevent this number rising should be prioritized by the government in order to decrease crime rates in Greater Manchester. For students, living in areas with a high % of rented housing, signs of deprivation, and a higher population of recent arrivals will be subjected to a higher crime rate so keep this in mind when moving in to your student house.

6.1.0.6 As mentioned in the Data section, ways to enhance this research would be to find ways to incorporate variables that weren't accounted for in the dataset. The easiest of which would be to include the gentrification index (constructed by Flinders [26]) which would more extensively cover gentrification beyond just using Residency Duration. Also gaining access to historical reports on 'FixMyStreet' could help give insights to urban decay and its relationship with crime.

6.1.0.7 Model The final model is a Negative Binomial generalized linear model enhanced by LASSO selection and bivariate interaction analysis that achieved a strong balance of simplicity and explanatory power - with a Nagelkerke R^2 of 0.416. This means that the final model contributes to explaining 41% of crime variation which is reasonably strong for a demographic model that has been reduced to this extent to avoid the possibility overfitting and have a better out-of-sample generalization.

6.1.0.8 The 'Predicted Map' below is a visualization of the final model (which was trained on Crime data from 2018) predicting crime by fitting with the reduced selection of Census 2021 variables. This is a purely data-driven prediction of crime using area characteristics alone. Comparing this with a raw observed response, we can see that our model still can pick up key trends in crime variation despite having reduced complexity. This validates internal consistency of our model, showing it can reproduce the response that it already knows, but doesn't show true model validation.

```r
Dataset2$Predicted_Crime <- predict(model_reduced,
   type = "response")
```

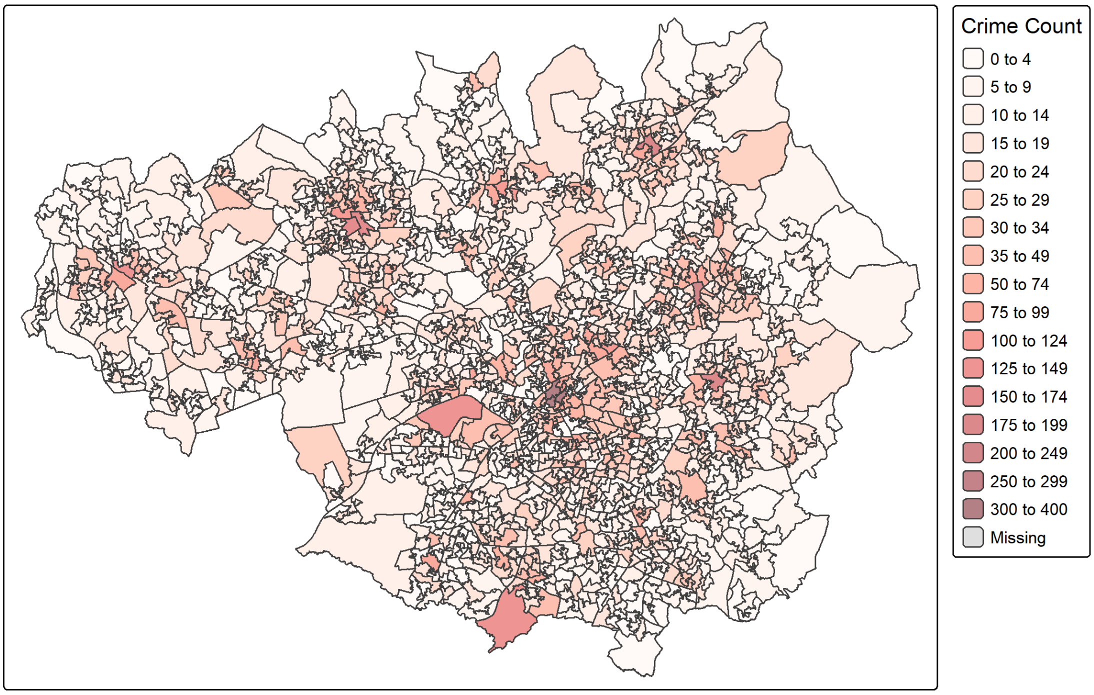

Figure 6: Crime Map

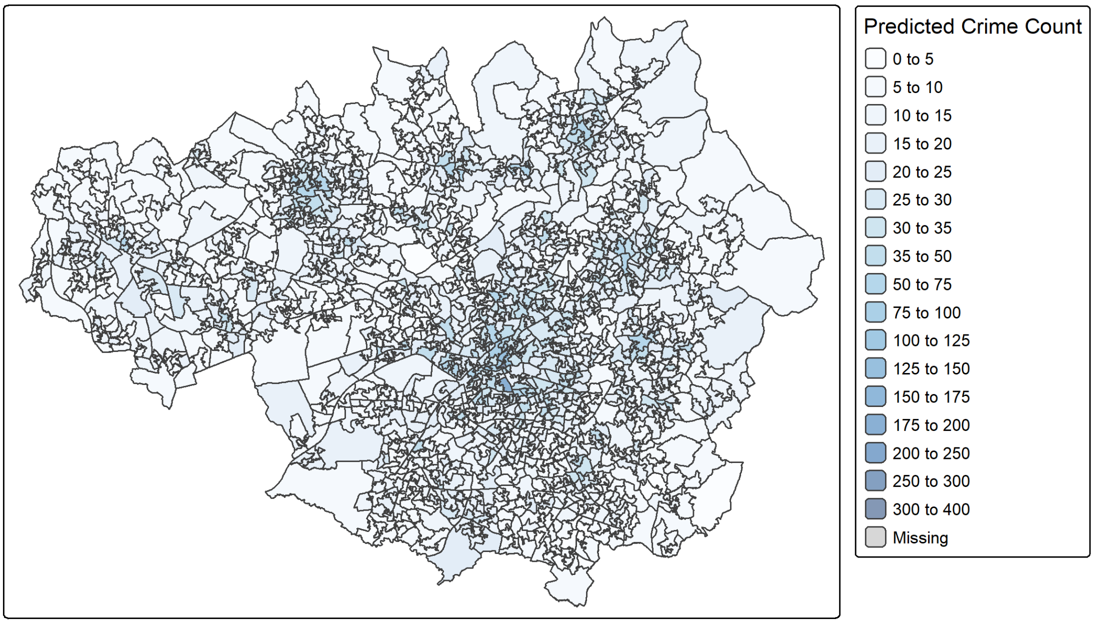

Figure 7: Predicted Map

6.1.0.9 Further Model Development: To test whether the model's relationships generalize beyond Greater Manchester, a non-temporal out of sample validation must be conducted. This would involve gathering another dataset with the same variables from the same 2021 Census survey but in a different city/location and testing the model to see how accurate it is in replicating crime count of that location. If this turns out to be somewhat accurate, it would support spatial generalizability meaning the identified demographic and socioeconomic drivers of crime in Greater Manchester are applicable more broadly across the UK.

6.1.0.10 Further valuable extensions of this model would involve the incorporation of temporal data to evaluate how well the model performs over time. This would assess the models ability to capture evolving crime patterns and allow for time series forecasting which essential for informing proactive crime prevention strategies.

6.1.0.11 Ultimately this project demonstrates the potential of demographic and socioeconomic modeling to explain crime variation, offering a foundation for further spatial and temporal validation.

Appendix

Methods used for plotting histogram

```
library(MASS)


DS2 <- read.csv("~/GLM/OALS/Meeting/DS2_MB.csv",
   header = TRUE)
dim(DS2)  # [1] 1702 175
y <- DS2$Crime
pois_model <- glm(y ~ 1, family = poisson)
nb_model <- glm.nb(y ~ 1)
q99 <- quantile(y, 0.99)
y_trimmed <- y[y <= q99]
hist_data <- hist(y_trimmed, breaks = 30, freq =
   FALSE,
                  main = "Histogram of Crime Counts
                     (Excl. top 1%)",
                  xlab = "Crime Count", col = "
                     lightblue", border = "white")

x_vals <- 0:max(y_trimmed)
bin_width <- hist_data$breaks[2] - hist_data$breaks
   [1]
lambda <- mean(y)
mu <- mean(y)
theta <- nb_model$theta

pois_probs <- dpois(x_vals, lambda) / bin_width
nb_probs   <- dnbinom(x_vals, mu = mu, size = theta)
    / bin_width

lines(x_vals, pois_probs, col = "red", lwd = 2)
lines(x_vals, nb_probs, col = "darkgreen", lwd = 2)

legend("topright", legend = c("Poisson", "Negative
   Binomial"),
       col = c("red", "darkgreen"), lwd = 2)
```

Methods used for Boxplots

```
 Filter the two datasets
mcr_data <- DS2 %>% filter(Area == "Manchester")
rest_data <- DS2 %>% filter(Area != "Manchester")

# Boxplot for rest
p1 <- ggplot(rest_data, aes(x = Area, y = Crime)) +
  geom_boxplot(fill = "lightblue") +
  theme(axis.text.x = element_text(angle = 45, hjust
      = 1)) +
  labs(title = "Crime Counts (Other Areas)", x = "
     Area", y = "Crime")

# Boxplot for Manchester only
p2 <- ggplot(mcr_data, aes(x = Area, y = Crime)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Crime Counts (Manchester)", x = "",
     y = "Crime") +
  theme(axis.text.x = element_blank())

# Combine them
grid.arrange(p1, p2, ncol = 2)
```

Methods for creating spatial images from shapefiles

```
library(readxl)
library(tidyverse)
library(sf); library(spdep); library(tmap)
library(MASS); library(lmtest); library(pscl)


# Read dataset
#Dataset <- read_excel("~/GLM/OALS/Dataset.xlsx")
Dataset = read.csv("~/GLM/OALS/Meeting/DS2_MB.csv",
   header=T); dim(DS2)
# Read LSOA shapefile
shapefile_path <- "~/GLM/OALS/Lower_layer_Super_
   Output_Areas_(December_2021)_Boundaries_EW_BFC_(V
   10).shp"
lsoa_shapefile <- st_read(shapefile_path)

# Filter for Greater Manchester areas
gm_keywords <- c("Wigan", "Rochdale", "Manchester",
   "Salford", "Oldham",
                 "Bury", "Bolton", "Tameside", "
                    Trafford", "Stockport")

greater_manchester_lsoa <- lsoa_shapefile %>%
  filter(str_detect(LSOA21NM, paste(gm_keywords,
     collapse = "|")))

# Merge shapefile with dataset
Dataset2 <- left_join(greater_manchester_lsoa,
   Dataset, by = c("LSOA21CD" = "LSOA21CD"))

# Plot base map of Greater Manchester LSOAs
plot(st_geometry(greater_manchester_lsoa))

# Check for missing values in crime data
sum(is.na(Dataset2$All))
view(Dataset2)
# Visualize crime distribution
crime_breaks <- c(0, 5, 10, 15, 20, 25, 30, 35, 50,
   75, 100, 125, 150, 175, 200, 250, 300, 400)
crime_map <- tm_shape(Dataset2) +
  tm_polygons("Crime", palette = "Reds", alpha =
     0.5, breaks = crime_breaks, title = "Crime
```

```
      Count") +
  tm_layout(legend.outside = TRUE)


# Compute Geometric Centroids for Each LSOA
geometric_centroids <- greater_manchester_lsoa %>%
  mutate(geometric_centroid = st_centroid(geometry))
       %>%
  st_as_sf(coords = c("geometry"), crs = st_crs(
     greater_manchester_lsoa))
# Read the already population-weighted centroids
centroid_shapefile_path <- "~/GLM/OALS/LSOA_
   PopCentroids_EW_2021_V3.shp"
cs <- st_read(centroid_shapefile_path)

# Merge with dataset (to add crime data if needed)
merged <- inner_join(cs, Dataset, by = c("LSOA21CD"
   = "LSOA21CD"))

# Plot using ggplot
ggplot() +
  geom_sf(data = merged, aes(geometry = geometry),
     fill = "green", color = "black", alpha = 0.5) +
  geom_sf(data = cs, color = "red", size = 2, shape
     = 4) +  # Using pre-existing centroids
  ggtitle("Existing Population-Weighted Centroids (
     Red)")

# Overlay centroids on crime map
centroid_map <- crime_map +
  tm_shape(geometric_centroids) +
  tm_symbols(size = 0.3, col = "green", shape = 4,
     title = "Geometric Centroids") +
  tm_shape(cs) +
  tm_symbols(size = 0.3, col = "red", shape = 4,
     title = "Population-Weighted Centroids")  # Use
      ‘cs‘ directly

# Enable interactive mode and display final map
tmap_mode("view")
crime_map
centroid_map
```

Methods used for plotting scatterplot matrix

```
library(readxl)
library(dplyr)
library(GGally)

DS2 <- read_excel("~/GLM/OALS/DS2.xlsx")

# View the column names for safety
colnames(DS2)

# Now select the correct columns
dt <- dplyr::select(DS2, ‘%_Owned‘, ‘2 rooms‘, ‘%_
   Owns with a mortgage or loan or shared ownership
   ‘, ‘Private rented or lives rent free‘)

# Base R scatterplot matrix
pairs(dt, main = "R base package", upper.panel =
   NULL)

# GGally scatterplot matrix
ggpairs(dt)
```

Methods for fitting GLM and for obtaining overdispersion parameter estimates

```
m1 = lm(y ~ x+DS2$Area)
m2 = glm(y ~ x+DS2$Area, family = poisson)
m3 = glm(y ~ x+DS2$Area, family = quasipoisson)
m4 = glm.nb(y ~ x+DS2$Area)
out1  = summary(m1)$adj.r.squared
out2a = summary(m2)$aic
out2b = nagelkerke(m2)[[2]][3]
out3  = summary(m3)$dispersion
out4a = summary(m4)$aic
out4b = nagelkerke(m4)[[2]][3]
out4c = summary(m4)$theta
tmp = c(out1, out2a, out2b, out3, out4a, out4b, out4
   c)
tmp
disp_test = dispersiontest(m2)
disp_test
```

Example of methods used for univariate analyses

```
for ( i in 28:165) {
x = DS 2[ , i ]
m 1 = summary ( lm ( y ~ x ) ) $ adj . r . squared

m 2 a = summary ( glm ( y ~ x , family = poisson ) )
    $ aic
m 2 b = nagelkerke ( glm ( y ~ x , family = poisson
   ) )
[[2]][3]
m 3 = summary ( glm ( y ~ x , family = quasipoisson
   ) ) $
dispersion
m 4 a = summary ( glm . nb ( y ~ x ) ) $ aic
m 4 b = nagelkerke ( glm . nb ( y ~ x ) ) [[2]][3]
m 4 c = summary ( glm . nb ( y ~ x ) ) $ theta
tmp = c ( cl . nm [ i ] , " NA " , m 1 , m 2a , m 2b
     , m 3 , m 4a , m 4b
, m 4 c )
out = rbind ( out , tmp )
```

Output of first order GLM with all variables

Variable	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	4.6565	2.8459	1.6362	0.1018
% Private.rented.or.lives.rent.free	-0.1386	0.2392	-0.5797	0.5621
Aged.25.to.34.years	-0.0014	0.0005	-2.7464	0.0060
Aged.75.to.84.years	0.0025	0.0035	0.6997	0.4841
Household.is.deprived.in.one.dimension	0.0009	0.0010	0.9126	0.3614
Household.is.deprived.in.three.dimensions	0.0018	0.0077	0.2277	0.8199
Household.is.deprived.in.four.dimensions	-0.0138	0.0486	-0.2836	0.7767
No.qualifications	0.0026	0.0007	3.8105	0.0001
Economically.active.and.a.full.time.student	0.0002	0.0023	0.0693	0.9447
Economically.inactive	0.0016	0.0006	2.7616	0.0058
Economically.inactive..Other	0.0060	0.0038	1.5742	0.1154
Asian..Asian.British.or.Asian.Welsh	0.0007	0.0003	2.4741	0.0134
Mixed.or.Multiple.ethnic.groups	0.0056	0.0043	1.2828	0.1996
Other.ethnic.group	-0.0123	0.0033	-3.6708	0.0002
No.religion	0.0006	0.0003	1.9908	0.0465
Buddhist	0.0213	0.0135	1.5769	0.1148
Other.religion	0.0514	0.0172	2.9839	0.0028
Born.in.the.UK...53	-0.0013	0.0003	-3.9550	7.65×10^{-5}
10.years.or.more	0.0075	0.0026	2.8868	0.0039
5.years.or.more..but.less.than.10.years	-0.0187	0.0119	-1.5699	0.1164
Less.than.2.years	0.0175	0.0081	2.1455	0.0319
1.room	-0.0040	0.0038	-1.0453	0.2959
2.rooms	-0.0011	0.0019	-0.5723	0.5671
3.rooms	-0.0004	0.0013	-0.3008	0.7636
6.rooms	0.0072	0.0037	1.9391	0.0525
7.rooms	-0.0110	0.0087	-1.2635	0.2064
9.or.more.rooms	-0.0018	0.0071	-0.2467	0.8051
Arrived.1991.to.2000	-0.0259	0.0097	-2.6745	0.0075
Arrived.2011.to.2013	0.0113	0.0132	0.8552	0.3925
Arrived.2014.to.2016	0.0038	0.0095	0.3966	0.6917
Arrived.2020.to.2021	-0.0370	0.0112	-3.2993	0.0010
Social.rented	0.0008	0.0016	0.4773	0.6331
Private.rented	0.0049	0.0017	2.9843	0.0028
Lives.rent.free	0.0341	0.0375	0.9104	0.3626
% Aged.20.to.24.years	-0.0340	0.0096	-3.5351	0.0004
% Aged.50.to.64.years	0.0014	0.0105	0.1363	0.8916
% Aged.65.to.74.years	-0.0345	0.0151	-2.2815	0.0225
% Aged.75.to.84.years	-0.0961	0.0533	-1.8023	0.0715
% Household.is.deprived.in.two.dimensions	0.0086	0.0101	0.8506	0.3950
% Household.is.deprived.in.three.dimensions	-0.0070	0.0566	-0.1233	0.9019
% Household.is.deprived.in.four.dimensions	0.0654	0.3352	0.1952	0.8453
% Apprenticeship	0.0248	0.0199	1.2458	0.2128
% Economically.active.and.a.full.time.student	0.0313	0.0436	0.7184	0.4725
% Economically.inactive..Other	-0.0873	0.0592	-1.4745	0.1403
% Mixed.or.Multiple.ethnic.groups	-0.0733	0.0773	-0.9487	0.3428
% White	0.0293	0.0068	4.3353	1.46×10^{-5}
% Other.ethnic.group	0.2419	0.0625	3.8712	0.0001

Table 12: Full regression output

Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>|z|$)**
% Christian	-0.0048	0.0045	-1.0695	0.2849
% Buddhist	-0.4227	0.2465	-1.7147	0.0864
% Other religion	-0.8100	0.2895	-2.7983	0.0051
% 10 years or more	-0.1294	0.0479	-2.6992	0.0070
% 5 years or more but less than 10 years	0.5021	0.2248	2.2336	0.0255
% Less than 2 years	-0.3513	0.1501	-2.3400	0.0193
% 2.rooms	-0.0228	0.0312	-0.7295	0.4657
% 3.rooms	-0.0337	0.0279	-1.2096	0.2264
% 4.rooms	-0.0398	0.0265	-1.5055	0.1322
% 5.rooms	-0.0384	0.0264	-1.4575	0.1450
% 6.rooms	-0.0878	0.0356	-2.4695	0.0135
% 7.rooms	0.0028	0.0682	0.0417	0.9668
% Arrived.1991.to.2000	0.5198	0.1770	2.9372	0.0033
% Arrived.2011.to.2013	-0.3452	0.2487	-1.3879	0.1652
% Arrived.2014.to.2016	-0.2133	0.1827	-1.1675	0.2430
% Arrived.2020.to.2021	0.7741	0.2092	3.7001	0.0002
% Social.rented	-0.0122	0.0116	-1.0530	0.2923
% Private.rented	0.1095	0.2398	0.4568	0.6478
DS2$Area Bury	-0.1275	0.0904	-1.4100	0.1585
DS2$Area Manchester	0.0121	0.0979	0.1238	0.9015
DS2$Area Oldham	0.0815	0.0842	0.9682	0.3330
DS2$Area Rochdale	0.0735	0.0829	0.8865	0.3754
DS2$Area Salford	-0.2974	0.0932	-3.1902	0.0014
DS2$Area Stockport	-0.2397	0.0927	-2.5854	0.0097
DS2$Area Tameside	-0.1135	0.0873	-1.2999	0.1936
DS2$Area Trafford	-0.0234	0.1022	-0.2285	0.8192
DS2$Area Wigan	-0.1902	0.0881	-2.1585	0.0309

Table 13: Full regression output

References

[1] Statista Inc., USA. (2024) "Crime rate per 1,000 population in England and Wales in 2023/24, by police force area". `https://www.statista.com/statistics/866788/crime-rate-england-and-wales-by-region/`

[2] ITV News: Insight (2024) "Moss Side: How a history of violence still affects people today", 13-Sep-2024. *https://www.itv.com/news/granada/2024-09-13/moss-side-how-a-history-of-violence-still-affects-people-today*

[3] Curtis-Ham, S., Bernasco, W., Medvedev, O.N. et al. (2020) "A framework for estimating crime location choice based on awareness space". Crime Sci 9, 23 (2020). *https://doi.org/10.1186/s40163-020-00132-7.*

[4] Casey BJ, Jones RM, Hare TA. (2008) "The adolescent brain". Ann N Y Acad Sci. Mar;1124:111-26. *doi: 10.1196/annals.1440.010.*

[5] Steinberg L. (2008) "A Social Neuroscience Perspective on Adolescent Risk-Taking". Dev Rev. 28(1):78-106. *doi: 10.1016/j.dr.2007.08.002.*

[6] Massey, D. S., & Denton, N. A. (1993) "American Apartheid: Segregation and the Making of the Underclass". *Harvard University Press.*

[7] Glaeser, E. L., & Sacerdote, B. (1999). "Why is There More Crime in Cities?" Journal of Political Economy, 107(S6), S225–S258. *https://doi.org/10.1086/250109*

[8] FixMyStreet (2024) Url: *https://www.fixmystreet.com/*

[9] Police.uk (2024). Mayor's Office for Policing and Crime (MOPAC). *https://www.police.uk/*

[10] Nomis - Office for National Statistics (2024) "Census 2021" *https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?theme=93*

[11] Piza, Eric L. (2024) "CCTV Video Surveillance and Crime Control: The Current Evidence and Important Next Steps". in Brandon C. Welsh, Steven N. Zane, and Daniel P. Mears (eds), The Oxford Handbook of Evidence-Based Crime and Justice Policy, Oxford Handbooks. *https://doi.org/10.1093/oxfordhb/9780197618110.013.14.*

[12] College of Policing (2021) "The effectiveness of visible police patrol". *https://www.college.police.uk/research/what-works-policing-reduce-crime/visible-police-patrol*

[13] Compton, G. (2015) "An Overview of the 2021 Census Design". UKDS Conference. https://dam.ukdataservice.ac.uk/media/455472/comptonblake.pdf

[14] Office for National Statistics (2023) "Measures showing the quality of Census 2021 estimates" *https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/measuresshowingthequalityofcensus2021estimates*

[15] Graif, C., Gladfelter, A. S., & Matthews, S. A. (2014). "Urban Poverty and Neighborhood Effects on Crime: Incorporating Spatial and Network Perspectives". *Sociological Compass*, 8(9), 1140–1155. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4928692/

[16] Western, B. & Pettit, B. (2010). "Incarceration and Social Inequality". *Daedalus*, 139(3), 8–19. https://doi.org/10.1162/DAED_a_00019

[17] Itskovich, E. (2024). "Economic Inequality, Relative Deprivation, and Crime: An Individual-Level Examination." *Justice Quarterly. https://doi.org/10.1080/07418825.2024.2435859*

[18] Northumbria University. (2023). "Research investigating links between pubs and crime rates offers insights into better policing" *Northumbria University News.* Retrieved from https://www.northumbria.ac.uk/about-us/news-events/news/pubs-and-crime/

[19] Choi, H. M., et al. (2024). "Temperature, Crime, and Violence: A Systematic Review and Meta-Analysis". *Environmental Health Perspectives*, 132(10), 106001. https://doi.org/10.1289/EHP14300

[20] Bosick, S., & Fomby, P. (2019). Family instability in childhood and criminal offending during the transition into adulthood. *American Behavioral Scientist.* https://pmc.ncbi.nlm.nih.gov/articles/PMC6889959/

[21] National Gang Center. (n.d.). *National Youth Gang Survey Analysis: Gang-Related Offenses.* Retrieved May 5, 2025, from https://nationalgangcenter.ojp.gov/survey-analysis/gang-related-offenses

[22] Ferrazzi, D. (2022). *Social Cohesion and Tenure in Urban Communities: A Study of Housing Stability and Community Engagement.* PhD thesis, University of Leeds. Available at: https://etheses.whiterose.ac.uk/id/eprint/32478/1/Ferrazzi%2C%20Dario%2C%20reg.170151355_Post-viva%20edits.pdf. *page 45*

[23] Wilson, J. Q., & Kelling, G. L. (1982). "Broken Windows: The police and neighborhood safety". *The Atlantic Monthly*, 249(3), 29–38. `https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/`

[24] Demireva, N. (2019) "Immigration, Diversity and Social Cohesion". *The Migration Observatory, University of Oxford.* Available at: `https://migrationobservatory.ox.ac.uk/resources/briefings/immigration-diversity-and-social-cohesion/`

[25] Palmer, C., Pathak, P., & Autor, D. (2017). "Does gentrification reduce crime?" *VoxEU.org*, 16 November 2017. *https://cepr.org/voxeu/columns/does-gentrification-reduce-crime*

[26] Flinders, S., & Almeida, A. (2025). "The Greater Manchester Gentrification Index". *Common Wealth.* Retrieved from `https://www.common-wealth.org/interactive/the-greater-manchester-gentrification-index`

[27] Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley.

[28] Dunn, P.K., & Smyth, G.K. (2018). *Generalized Linear Models with Examples in R.* Springer.

[29] Faraway, J.J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* CRC Press.

[30] Dey, D., Ghosh, M., & Mallick, B. (Eds.). (2000). *Generalized Linear Models: A Bayesian Perspective.* Marcel Dekker, Inc.

[31] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer Texts in Statistics.

[32] McCullagh, P., & Nelder, J.A. (1983). *Generalized Linear Models.* Chapman and Hall.