M3 Challenge 2024:

“A Tale of Two Crises:
The Housing Shortage and Homelessness”

Team #17580
01 March 2024

Executive Summary

The housing crisis and homelessness are not merely statistics or policy issues; they are profound reflections of our societal values and priorities. Behind every statistic lies a human story of struggle, resilience, and hope. Our team aims to fabricate a long-term plan to support major cities in addressing the homelessness crisis by assisting the councils in deploying an adaptable model for which unforeseen circumstances of the nature of natural disaster, economic recessions and the current migration status can be considered in a Poisson regression model.

We attempted to forecast the quantity of dwelling units in Manchester and Brighton & Hove over the next 50 years by using a multilinear regression model. Our model used several indicators, such as the overall population and income levels, to estimate the growth in housing units in Manchester throughout the course of the projection period. By fitting the model to historical data and accounting for expected population and income growth rates, we projected an increase in housing units from 2023 to 2074, with an estimated total of **263,773** housing units in 2034, **292,389** housing units in 2044 and **418,699** units in 2074. Using a similar multivariate linear regression model, we also estimated that the number of housing units in Manchester and Brighton will rise to **139,227** units in 2034, **145,238** units in 2044 and **163,268** units in 2074.

As for the homeless population in Manchester and Brighton & Hove, we created an SEIR (Susceptible-Exposed-Infectious-Recovered) model for modelling change in homelessness population for which the choice of stems from its adaptability and effectiveness in capturing the complex dynamics of infectious diseases, which share similarities with the spread of homelessness within a population. The SEIR model has been successfully applied in epidemiology to understand the transmission dynamics of diseases, making it a suitable framework for analysing the progression of homelessness through different stages within a community hierarchy. By calculating the relevant coefficients, we were able to solve the differential equations and project that the homeless population in Manchester would reach **5,950**, **7,984** and **9,163** in 2034, 2044 and 2074 respectively. Likewise, Brighton & Hove's homeless population is projected to reach **1,394**, **2,430** and **4,556** in 2034, 2044 and 2074 respectively.

As a result of using a combination of the Poisson regression model and the multivariate regression model, this could exemplify the effects of the consequences of each individual factor, considering all socio-economic and economic factors on the effect of homelessness. Moreover, by using the statistical models, we can aim to bring light upon the harmful consequences that may impact the growth of the homeless population, creating awareness as a dire need to tackle this problem. By taking initiative to model a dangerous global challenge such as proverty, repositioning the homeless from its pivotal role in society exemplifies the ostracised effects of the efforts to benefit the homeless in society.

Contents Page

Executive summary **2**

Global assumptions for The Models **4**

Q1: It Was the Best of Times

1.1 Defining the problem **4**

1.2 Assumptions **5**

1.3 Variables **5**

1.4 The Model **7**

1.5 Results **8**

1.6 Sensitivity Analysis **9**

1.7 Strengths and Weaknesses of The Model **10**

Q2: It Was the Worst of Times

2.1 Defining the problem **11**

2.2 Assumptions **11**

2.3 Variables **12**

2.4 The Model **12**

2.5 Results **13**

2.6 Sensitivity Analysis **14**

2.7 Strengths and Weaknesses of The Model **15**

Q3: Rising from This Abyss

3.1 Defining the problem **15**

3.2 Assumptions **15**

3.3 Variables **16**

3.4 The Model **16**

3.5 Results **19**

3.6 Sensitivity Analysis **20**

3.7 Strengths and Weaknesses of The Model **20**

Conclusion **20**

References **21**

Code Appendix **22**

M3 Challenge 2024:

Global Assumptions

1. There will be no revolutionary change in the housing market within the next 50 years. Any "revolution" in the housing market yet to come will shift the housing market completely, hence making our model inaccurate and unreliable if not considered. This is practically impossible to consider.

2. Bungalows, Flats, Maisonettes, Terraced, Detached/Semi-Detached, Caravan, Houseboat and Mobile homes will all be classified as "housing". Under the definition for homelessness from the UK House of Commons [1]. this will qualify all the above to be categorically housing.

3. There will be no major shift in the economy of the UK. This will allow to simplify the models and for us to use historical growth data and growth rates to create a model which will allows us to bypass complex economic forecasting.

4. Assuming there is negligible percentage difference in the raw data values in the population of the unhoused and the real, defined approximation of the population in each year. This means that there is a limited number of variables to consider when modelling the multivariate regression model.

5. The availability of affordable housing remains inaccessible to the unhoused/ the cost of each affordable housing is not one that is available for the homeless. This will ensure that the government gives rise to the availability of affordable housing, the rate of the homeless population is not skewed.

*6. No revolutionary Technological Advancements and Economic Stability***.** We can assume that there is no major economic crisis such as large inflation rates, large tax inflation impacting the affordability of the homes, with significant fluctuations of GDP growth aligned with stark unemployment rates in the UK. In the case where UK's GDP fluctuates significantly below par, this may take on an effect on magnitude of the houses/infrastructure that can be built used to aid the homeless and thus, take an impact on the rate of homelessness. In light of this argument, where the UK's GDP may be skewed, the efficiency of technological aids in housing construction, may negatively constribute to the additional costs to build housing and improve quality of life, taking into account other social factors such as counselling.

Q1: It Was the Best of Times

1.1 Defining the Problem

In this endeavor, our objective is to predict the growth of housing units in the United States or the United Kingdom over the next several decades. For this purpose of the question, we have decided to formulate a linear regression model based on Manchester and Brighton and Hove. This regression model is created from multiple variables that are taken into account which may impact the effect of homelessness on the population and its future potential increase in magnitude of the population. This could mean that the predicted multivariate regression model may allign closely with true values, making our overall model to explain the effects of homelessness accurately holistic.

1.2 Assumptions

1. ***Stable Policy Environment.*** We assume no significant policy changes concerning housing regulations, construction standards, or government initiatives that could impact housing unit growth over the forecast period. We can also assume that there is a marginal impact of governmental power upon the availability of houses made accessible to the homeless. This means that there is no govermental intervention on the initiatives used to aid homelessness i.e. by infrastructure and repairment, which could effect the population of the homeless in future years.

2. ***Constant Housing Demand.*** We anticipate a relatively stable relationship between housing demand and economic factors, with no substantial shifts in consumer behavior or preferences that might significantly alter housing demand. By this statement, we can therefore assume that potential socio-economic factors such as employment, quality of life and affordability index, which represents the proportion of income used to allocate housing expenses, play a sublime impact on the effect of homelessness. Moreover, this means that our model strictly correlates the impact of homelessness as a proportional relationship to availability of housing and house demands.

1.3 Variables

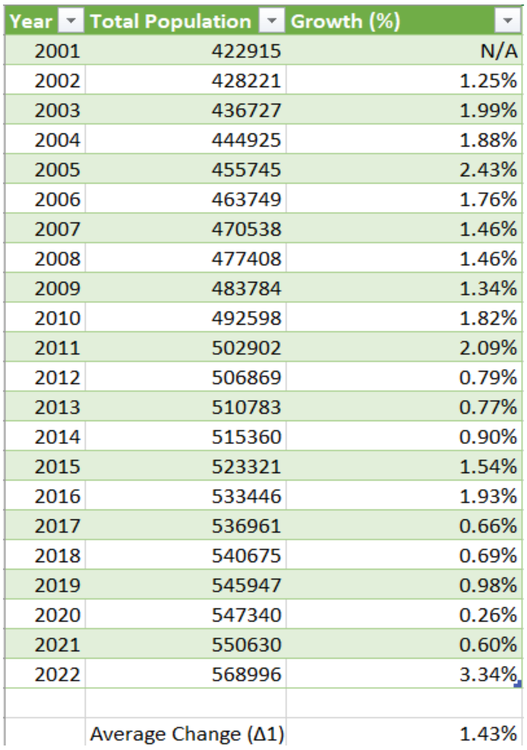
1.3.1 Manchester

Symbol	Definition	Units
Y	Total Housing Units	units
X_1	Total Population	units
Δ_1	Population Growth	Year^{-1}
X_2	Median Income	£
Δ_2	Median Income Growth	Year^{-1}
T	Years after 2019	units

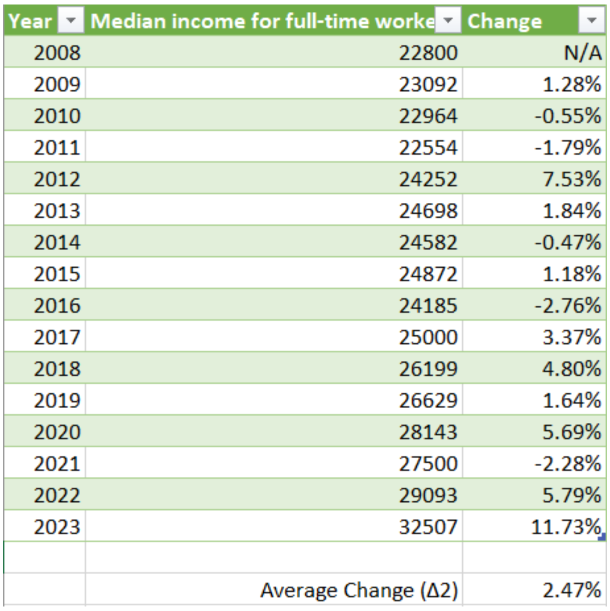
Table 1: Variable definitions for Problem 1

Total Population (X_1): Population growth is a significant factor that drives the growth of housing units. As the population grows, there is a corresponding increase in the number of households. Population growth can result from factors such as natural population growth and net migration. With more households being formed, there is a greater demand for housing units to accommodate these new households. We also calculated the percentage growth in population in Manchester annually and considered the average percentage growth in our model, giving us a value of 1.43% for Δ_1.

Median Income (X_2): People's purchasing power and, in turn, their capacity to afford housing are influenced by their income levels. Potential homeowners' affordability thresholds are frequently widened by rising salaries. This means that individuals who were previously unable to qualify for mortgages or afford down payments may meet the financial criteria necessary to purchase homes, thereby increasing the number of potential buyers in the housing market and therefore increasing the demand for housing units. We also considered the average income growth rate to estimate changes in affordability over time, giving us a value of 2.47% for Δ_2 as follows:



Year	Total Population	Growth (%)
2001	422915	N/A
2002	428221	1.25%
2003	436727	1.99%
2004	444925	1.88%
2005	455745	2.43%
2006	463749	1.76%
2007	470538	1.46%
2008	477408	1.46%
2009	483784	1.34%
2010	492598	1.82%
2011	502902	2.09%
2012	506869	0.79%
2013	510783	0.77%
2014	515360	0.90%
2015	523321	1.54%
2016	533446	1.93%
2017	536961	0.66%
2018	540675	0.69%
2019	545947	0.98%
2020	547340	0.26%
2021	550630	0.60%
2022	568996	3.34%
	Average Change (Δ1)	1.43%



Year	Median income for full-time worke	Change
2008	22800	N/A
2009	23092	1.28%
2010	22964	-0.55%
2011	22554	-1.79%
2012	24252	7.53%
2013	24698	1.84%
2014	24582	-0.47%
2015	24872	1.18%
2016	24185	-2.76%
2017	25000	3.37%
2018	26199	4.80%
2019	26629	1.64%
2020	28143	5.69%
2021	27500	-2.28%
2022	29093	5.79%
2023	32507	11.73%
	Average Change (Δ2)	2.47%

1.3.2 Brighton and Hove

Symbol	Definition	Units
Y	Total Housing Units	units
X	Current Year	units
M	Housing Growth	Year^{-1}
C	Initial Housing Units	units

Table 2: Variable definitions for Problem 1

1.4 Model Development

We decided to employ similar, yet different models for Manchester and Brighton & Hove due to their geographical differences and analysis of housing unit growth in both locations.

1.4.1 Regression Analysis (Manchester)

To predict the growth of housing units over time, we employed a multivariate regression model using the Ordinary Least Squares (OLS) method using MATLAB. This supervised machine learning approach utilises a linear equation to establish the relationship between the dependent variable, in this case, the total housing units, and multiple independent variables. We decided to use a multilinear regression model as data points for housing units in Manchester appear to be much more erratic and random compared to the data for Brighton & Hove. The multivariate regression model is represented by the following equation:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \mathcal{E}$$

For which:

- Y represents the dependent variable, which is the total number of housing units.
- X_1 and X_2 denote the independent variables: total population and median income respectively.
- β_0 is the intercept term, representing the value of Y when all independent variables are equal to zero.
- β_1 and β_2 are the coefficients that quantify the impact of the total population and median income respectively on the dependent variable.
- ε represents the error term, capturing the difference between the observed and predicted values of the dependent variable.

To predict future values of X_1 and X_2 for the model, we applied the following formulae to in a simple linear model using the following formulae:

$$X_1 = X_1 * \Delta_1^T$$

$$X_2 = X_2 * \Delta_1^T$$

1.4.2 Linear Regression (Brighton and Hove)

By representing the data graphically it was apparent the availibility of housing units has an extremely strong correlation with the elapsed time. After further research, this proved reasonable as Brighton and

Hove is a relatively small coastal town with plenty of space to expand outwards. A simple linear regression was the best way to utilise the this relationship.

$$Y = m \cdot X + c$$

Where:

- Y represent the dependant variable, which is the total number of housing units.
- X represent the current calendar year
- *m* represents the rate of increase of housing units
- *c* represents the y-intercept of the housing units at 0AD

This type of linear model will lead to negative values at for certain ranges of years outside its domain. As a result, it is only intended for use after 1993 and let:

$$m = 601.0246$$

$$c = -1.0833x10\text{^}6$$

1.5 Results

From the following regressions, the predicted housing units for Manchester in figure 1 and predicted housing units for Brighton and Hove in figure 2 can be seen as follows:

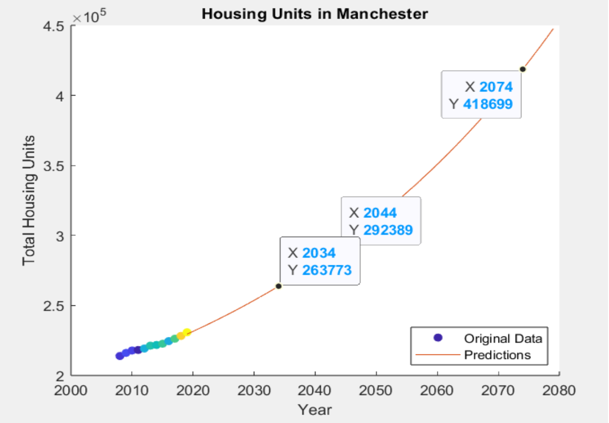


Figure 1

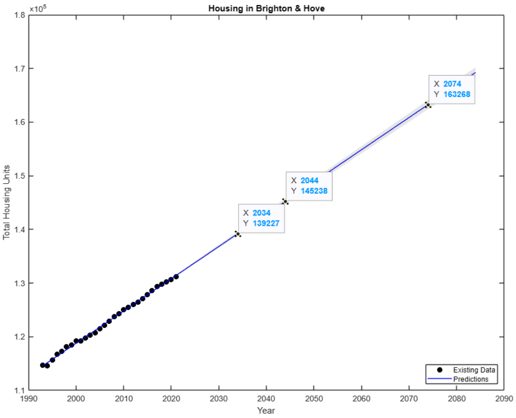


Figure 2

1.5.1 Manchester

In our model, we used the provided dataset containing information on housing units, median income and total population from the years 2008 to 2019. Following data preprocessing and feature selection, we fitted the multivariate regression model to the dataset using the least squares method to estimate the coefficients $\beta_{0.1,2}$. In our model, we obtained the following values for β:

- $B_0 = 110900$
- $B_1 = 0.163$
- $B_2 = 1.114$

From our model (Figure 1), we predict that the number of housing units will rise to **263773** units in 2034, **292389** units in 2044 and **418699** units in 2074 from its value of 230990 in 2019. This is shown in figure 2.

1.5.2 Brighton and Hove

In this model, we used the provided dataset containing information on housing units and their corrresponding years, from 1993 to 2021. Figure 2 contains prediction for the existing housing units 10, 20 and 50 years into the future. The grey area represents the range of housing units where we are 95% confident the true value will lie.

We predict that the number of housing units will rise to **139227** units in 2034, **145238** units in 2044 and **163268** units in 2074 from its value of 131240 in 2021.

1.6 Sensitivity Testing

We conducted sensitivity testing by % error for our values using the following formula:

$$\%\,Error\ =\ |\frac{E-T}{T}|\cdot 100$$

Where E is the predicted value and T is the true value, and calculated the R^2 by using the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SS_{res} is the seridual sum of squares and SS_{tot} is the total sum of squares.

1.6.1 Manchester

Using the estimated value of 233591 and the real value of 238800, we can calculate the percentage error to be **2.18%**

1.6.2 Brighton and Hove

Using the estimated value for 2021 and the real value for 2021 gave a percentage error of **0.13%.** Furthermore an R^2 value of **0.99725** shows this model is quite accurate.

1.7 Strengths and Weaknesses

The regression models offer a quantitative framework for understanding the drivers of housing unit growth and making informed projections. They enable policymakers and stakeholders to assess the potential impacts of different scenarios and interventions. Weaknesses: While regression analysis provides valuable insights, it is subject to certain limitations. The models rely on historical data and assumptions about future trends, which may not fully capture unforeseen developments or structural shifts in housing markets.

For the model for Manchester, we obtained a ordinary R^2 value of **0.9755** and an adjusted R^2 value of **0.9701**. A high R-squared value indicates that the model fits the observed data points well. The adjusted R^2 value is particularly useful when comparing models with different numbers of predictors. It penalizes the addition of unnecessary predictors that do not improve the model's performance. In this case, the adjusted R-squared value of **0.9701** suggests that the model for Manchester is reliable and provides a good balance between explanatory power and model simplicity.

Q2: It Was the Worst of Times.

2.1 Defining the Problem

The second problem asks us to predict the changes in the homeless population in the next few decades. We chose to analyse the two areas in the UK, this being Manchester and Brighton and Hove.

2.2 Assumptions

1. ***Homelessness can be modelled as an infectious disease.*** The dynamic nature of homelessness, where individuals may move from being at risk of homelessness to actually experiencing it, and potentially to becoming housed, much like the transitions between susceptibility, exposure, infection, and recovery in disease spread. While homelessness is not a disease and does not spread through biological contagion, the SEIR model offers a structured way to quantitatively represent the flow of people between different regionds due to external socioeconomic pressures and policy interventions, which can influence these transitions.

2. ***Homelessness defines those who do not have access to accommodation for his/her occupation.*** The UK House of Commons [1] defines homelessness as such; thus, we will recognise the following also.

3. Homogeneous population. Individuals experiencing homelessness in Brighton and Hove, and Manchester are considered as one group, regardless of specific demographic characteristics (age, gender, etc.). This allows us to eliminate any variables which influence the likelihood of homelessness based on the demographic characteristics of an individual and what that may entail.

4. Transmission Dynamics are reflected like that of a disease. We can model the spread of "homelessness" as a disease based on certain transmission dynamics. For example, individuals who are already homeless (Infectious) can transmit homelessness to susceptible individuals through various channels such as exposure to poverty, unemployment, lack of affordable housing, mental health issues, addiction, etc, through the influence of a linear regression model.

5. The infection will be contained within the city and not be allowed to spread beyond to the point at which it affects Brighton and Hove, and vice versa, over the next 50 years. Assume a closed population, meaning there are no new individuals entering or leaving the homeless population during the modelling period, except through transitions between SEIR states.

2.3 Variables

2.3.1 Manchester/ Brighton and Hove

Symbol	Definition	Units
α	The transmission rate from At-Risk to Homeless	year^{-1}
β	The transmission rate from Homeless to Housed	year^{-1}
γ	The recovery rate of At-Risk individuals	year^{-1}
δ	The relapse rate from Housed to At-risk	year^{-1}

Table 3: Variable definitions for Problem 2

2.4 The Model

2.4.1 Developing the Model

We chose a Susceptible-Exposed-Infectious-Removed (SEIR) model to predict the "spread" of homelessness in the cities of Brighton and Hove, and Manchester. Because each proportion of those homeless are represented as percentages and therefore probabilities, we can employ an infection model. Moreover, the SEIR model is the benchmark used in mathematical to describe the diffusion of an epidemic disease. Since we are analyzing the data for the UK, this provides an accurate prediction after testing and adjusting. The SEIR model consists of a system of ordinary differential equations (ODEs) that describe the flow of individuals between the compartments. The method formulas are as follows:

$$\frac{dS}{dt} = -\alpha \cdot \frac{I \cdot S}{N}$$

$$\frac{dE}{dt} = \alpha \cdot \frac{I \cdot S}{N} - \delta \cdot E$$

$$\frac{dI}{dt} = \delta \cdot E - \beta \cdot I$$

$$\frac{dR}{dt} = \gamma \cdot I$$

For which:

- S (Susceptible): In the context of homelessness, 'Susceptible' individuals are those at risk of becoming homeless.
- E (Exposed): It will represent individuals who are experiencing conditions that could imminently lead to homelessness, such as those with eviction notices.
- I (Infectious): For homelessness, 'Infectious' individuals would be those who are currently experiencing homelessness.
- R (Recovered): In our adaptation, 'Recovered' individuals are those who have secured housing and are no longer homeless.
- N (Total Population): This represents the total population of the area.

2.4.2 Executing the Model

In order to leverage the available data, we began by connecting those who were **priority need** Homeless and **not priority need** Homeless. We used the statistics for the average household size[3] to correct the households which are homeless to the homeless population, by using a multipier of 2.36 to adjust the data as such. To calculate the variables for our SEIR model, we coded a set of equations:

Variables	Values calculated
α	0.13040167618273908
β	0.12936643303985582
γ	0.5213003472739447
δ	1.4210526315789473

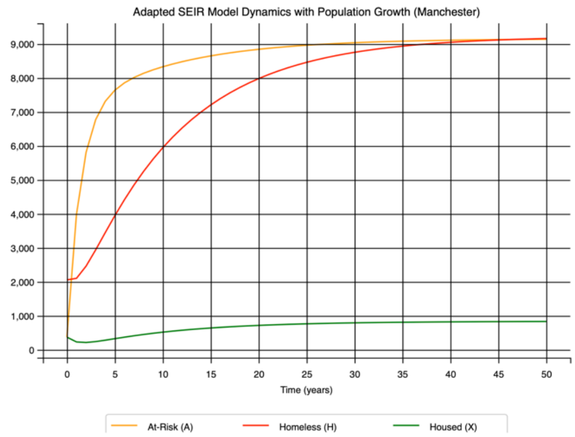
Table 4: Variable values for Manchester

Variables	Values calculated
α	0.015496127233052227
β	0.18429413982985354
γ	0.03619756660739407
δ	0.14592595905172612

Table 5: Variable values for Brighton and Hove

2.5 Results

From the code for the SEIR model using the variable values calculated above,we were able to produce a model which outlines the homelessness in the future through the infectious disease model, for which we obtained the following relations:



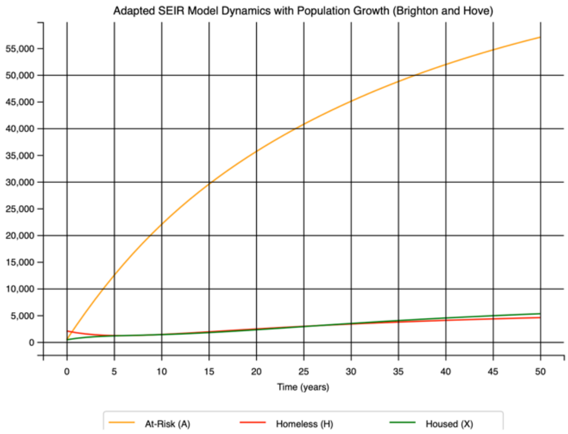


Figure 3*

Figure 4*

*Y axis represents the population on both Figure 3 and Figure 4

2.5.1 Manchester

The SEIR model predicts that in 2034 there will be **5950** homeless, in 2044 there will be **7984** homeless and in 2074 there will be **9163** homeless.

2.5.2 Brighton and Hove

The SEIR model predicts that in 2034 there will be **1394** homeless, in 2044 there will be **2430** homeless and in 2074 there will be **4556** homeless.

2.6 Sensitivity analysis

We tested the true values of homelessness with considering the multiplier of 2.36 units per household against the values from the model we built as follows:

Year	**Manchester**	**Manchester Predicted**	**Brighton and Hove**	**Brighton And Hove predicted**
2015	2692.76	n/a	1125.72	1138.40
2016	4915.88	4238.34	1151.68	1209.50

2017	5253.49	4846.24	1295.64	1253.62
2021	6549.00	6403.41	1123.36	1293.21

Table 5: Comparison of model to real data values

We conducted sensitivity testing for our values using the following formula:

$$\% \, Error \; = \; |\frac{E-T}{T}| \cdot 100$$

Where E is the predicted value and T is the true value. This gave us an average error of **1.13%** for Brighton and Hove, and an average percentage error of **7.75%** for Manchester, which shows that our model is fairly accurate with respect to the true data values given to us in the data sheet.

2.7 Strengths and Weaknesses

Even though our model seems to fit the data provided quite accurately, there is some major flaws which, if given more time, we could eradicate by not simplifying assumptions. The model relies on simplifying assumptions about the complex factors contributing to homelessness, which may oversimplify the real-world dynamics and overlook certain nuances. Moreover, the accuracy of the model depends on the availability and quality of data used for calibration, which was limited, leading to uncertainties in the model outcomes. Homelessness is a dynamic and multifaceted issue influenced by evolving socioeconomic conditions, policy changes, and individual circumstances. The model may not fully capture the dynamic nature of homelessness and its interactions with other social phenomena. Overall, while the SEIR model offers valuable insights into homelessness dynamics and interventions, we should be mindful of its limitations and uncertainties, using it as one of several tools in understanding and accurately addressing the magnitude of the effects of homelessness effectively in the future.

Q3: Rising from the Abyss:

3.1 Defining the Problem

As part of question 3, our aim is to exemplify the potential strategies to diminish the magnitude of homelessness, considering all possible economical, socio-economic factors (and political factors), we have decided to use a Poisson model by placing the measure of homelessness as a binary outcome and a dependent variable, whilst putting all other factors such as employment and housing demands as independent variables in order to illustrate the effects of each possible factor on the effects of homelessness. By illustrating this idea of modeling each factor's influence on the magnitude of homelessness, the aim is to provide council bodies with the information needed to make informed choices as to where they should focus their efforts to tackle homelessness.

3.2 Assumptions

1. There is only a sublime effect on the consequence of homelessness because of other factors such as employment rates and housing demands. It is for that reason; we can assume that the homeless population is susceptible to death and severe consequences. In the case where there is a natural hazard, i.e. an earthquake, it can be presumed that the unhoused are unlikely to be protected from the dangers of the consequences, so far so there is a risk of death. In this assumption, we can therefore assume that the severity of the consequences is such that there is no impact on the consequences that causes the population of the homeless to be damaged in the process.

2. There is economic stability in the UK, the UK GDP does not change, thus it does not have any impact on the population of the homeless by economic growth or loss. This means that there is no economic or governmental influence on the position of the homeless population because of the aid and infrastructure to improve the quality of life. Furthermore, this means that although the population rate may not remain constant, we can assume that the population of the homeless increases at a greater rate in proportion to the population's decrease.

3.3 Variables

Symbols	**Definitions**	**Units**
Y	Total Homeless population	units
y	Total Housing units	units
μ	Expected count predicted	units
α	Dispersion parameter	dimensionless
Γ	Gamma functions	dimensionless

Table 6: Variables definition for Problem 3

3.4 The model

3.4.1 Developing the model

To determine the significance of varied factors on homelessness, our approach centres around a regression model designed to handle count data. Here is an overview of why we made this choice and how the chosen regression model works:

- **Count Data and Overdispersion**: When modelling count data such as the number of homeless individuals, it is common to encounter "overdispersion"—where the variance of the data is greater than the mean. Traditional Poisson regression assumes that the mean and variance are equal, which often is not the case with real-world data. When overdispersion is present, as indicated by preliminary data analysis, using a Poisson regression can lead to biased standard errors, which influences confidence intervals and significance tests. We ran a test to determine if our data was suitable and found that the variance was in fact greater than the mean.
- **Negative Binomial Regression Model**: This model extends the Poisson regression by adding a parameter to account for the overdispersion. Our program implements a Negative Binomial regression to analyse the influence of factors like total housing units, median sales price, total population, and median income on the counts of homeless individuals.
- **Data Merging and Preparation**: Our data comes from various sources, each focusing on a specific aspect like housing or income. We merge these datasets using common identifiers such as the 'Year' column to create a single, coherent dataset for analysis. The merger ensures that each observation contains all the variables required to assess the impact on homelessness for that particular year. To provide an accurate estimation, we collated 50 predicted data points for each of the variables from our previous models.

Negative Binomial Regression is a statistical method used to model the relationship between a set of independent variables (like housing units, income, and population) and a dependent variable (homeless count).

In a Negative Binomial Regression, the count of homelessness is connected to the independent variables via a log-link function, which ensures the model's predictions are always positive. The formula for predicting homelessness counts (`μ`) is given by ` $\mu = e^{X\beta}$`, where `X` represents our independent variables, and `β` are the coefficients that the model estimates. If we exponentiate a coefficient from the model, we get what is called an Incidence Rate Ratio (IRR). An IRR higher than 1 suggests that increasing that independent variable is associated with higher counts of homelessness, whereas an IRR below 1 suggests the opposite.

To fine-tune this model, we estimate a parameter called `alpha`. This adjusts for the level of overdispersion in our count data, providing a better fit than the simpler Poisson regression, which does not account for overdispersion. We can refine our estimate of `alpha` to improve the model's performance by checking how well it fits the data and adjusting accordingly.

.By using this model with our estimated `alpha`, we can analyse how varied factors may influence the number of homeless individuals. After fitting the model and reviewing its results, we come to understand that none of the considered predictors, total housing units, median sales prices, and median income have a statistically significant influence on homelessness, indicating the complexity of the issue and the possibility that other factors not included in the model might be at play.

3.4.2 Executing the Model

To execute the Negative Binomial Regression Model and find the coefficients for each independent variable while considering the overdispersion and the impact of total housing units, median sales price, total population, and median income on homelessness, the following steps are taken:

1. Prepare Data for Modelling:

- Compile the actual observed values and the 50 predicted data points for each of the variables into a single dataset. Ensure that each column is appropriately named and corresponds to the variables of interest.

2. Estimate Preliminary Poisson Model:

- Fit a Poisson regression model to the observed data. This model serves as a baseline to measure overdispersion in the count of homeless individuals.

3. Calculate Overdispersion Parameter (α):

- Using the results from the Poisson model, especially the Pearson chi-squared statistic, estimate the degree of overdispersion to inform the parameter for the Negative Binomial model.

4. Fit Negative Binomial Regression Model:

- With the overdispersion parameter estimated, fit the Negative Binomial regression model using the combined observed and predicted data. This will yield coefficients for each variable that quantify their impact on homelessness.

5. Interpret the Results:

- Examine the coefficients, their associated z-values, p-values, and confidence intervals. Pay particular attention to the sign and magnitude of the coefficients and their statistical significance, which will indicate the impact of each factor on homelessness.

6. Assess Model Fit:

- Evaluate the model's fit using pseudo-R^2 and information criteria such as AIC or BIC. These metrics will help determine the explanatory power of the model and the robustness of the coefficients.

3.5 Results

Our model shows us that the number of housing units has the most significant impact on homelessness, followed by the median sales price of houses, as shown in Figure 5. Median income has a marginal effect compared to the other two factors. The total population has the opposite effect, because the population increases at a faster rate than homelessness and so the proportion of homeless people decreases. The P value greater than 0.05 shows that the variables we have taken to affect the homelessness in the cities are insignificant in the projected homelessness count. Thus, we can show that in the event of any natural disasters or radical changes to the population, the models will remain robust and valid in the projected homeless count.

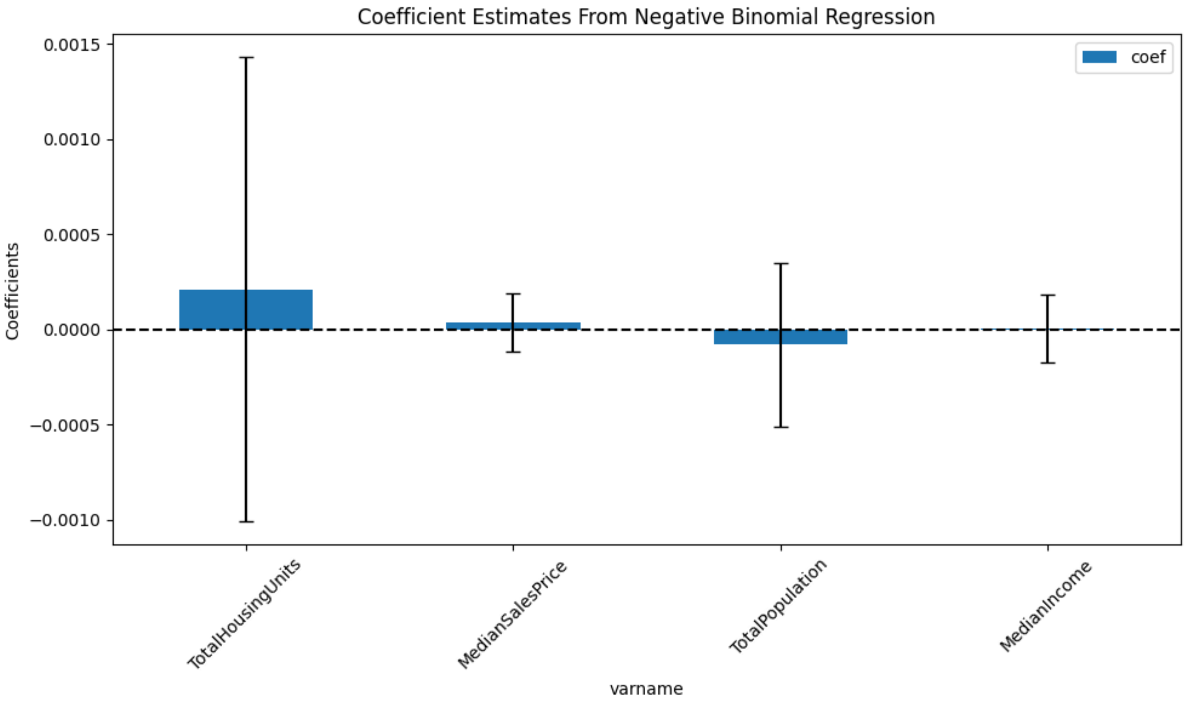
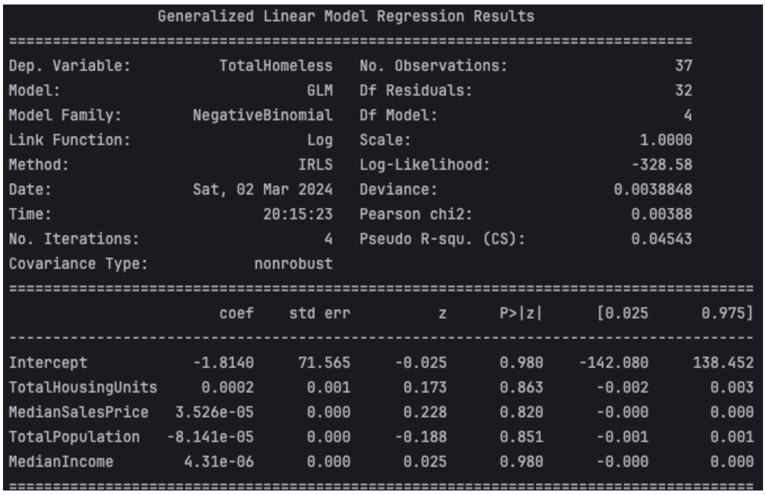


Figure 5: The Coefficient Estimates from Negative Binomial Regression



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          TotalHomeless   No. Observations:                   37
Model:                            GLM   Df Residuals:                       32
Model Family:        NegativeBinomial   Df Model:                            4
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -328.58
Date:                Sat, 02 Mar 2024   Deviance:                    0.0038848
Time:                        20:15:23   Pearson chi2:                  0.00388
No. Iterations:                     4   Pseudo R-squ. (CS):            0.04543
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            -1.8140     71.565     -0.025      0.980    -142.080     138.452
TotalHousingUnits     0.0002      0.001      0.173      0.863      -0.002       0.003
MedianSalesPrice   3.526e-05      0.000      0.228      0.820      -0.000       0.000
TotalPopulation   -8.141e-05      0.000     -0.188      0.851      -0.001       0.001
MedianIncome        4.31e-06      0.000      0.025      0.980      -0.000       0.000
=====================================================================================
```

Figure 6: The results of the Linear Model Regression

3.6 Sensitivity Analysis

A McFadden's pseudo R^2 ranging from 0.2 to 0.4 indicates very good model fit. As such, the model mentioned above with a McFadden's pseudo R^2 of 0.04 is likely not a terrible model, at least by this metric, but it isn't particularly strong either.

3.7 Strengths and Weaknesses of the Model

The negative binomial regression is well-suited for modelling count data. As all variables which we are assessing are expressed as counts, the negative binomial regression is particularly effective for analysing such data. The model also allows for overdispersion, which is common in such data due to numerous factors contributing to homelessness.

However, coefficients obtained from the negative binomial regression can be challenging, especially when dealing with multiple predictors and interaction terms. The inclusion of the dispersion parameter adds complexity to the model, making it more challenging to specify and interpret compared to simpler models like Poisson regression.

Conclusion

By considering the trends in socio-economic factors we have implemented models to predict future trends and deduce relationships between these variables and their impact on homelessness. In the concluding remarks of our analysis, we encapsulate not only the quantitative findings but also the profound insights gleaned from our exploration of housing and homelessness in the UK. Through the lens of our regression models, we have unveiled a trajectory of housing unit growth that, while promising, underscores a pressing need for more robust solutions to housing shortages. Our projections indicate an increase to 418,699 housing units in Manchester and 163,268 units in Brighton & Hove by 2074. Yet, these figures juxtapose starkly against the backdrop of a rising homelessness crisis, with predictions reaching up to 9,163 homeless individuals in Manchester and 4,556 in Brighton & Hove by the same year. From our Negative Binomial Regression Model, we can identify housing unit growth as the most significant cause of homelessness in Manchester and hence should be a key focus for the government within the coming years.

Evaluation

Housing data, often derived from valuations and comparable sales, may require a data smoothing process to enhance accuracy, particularly in scenarios involving inflation. This could influence the R^2 value of our analysis. For this study, we adopt a simplified approach to assess growth trends, fully acknowledging that this method does not account for all economic variables. While further research, adjustments, and hypothesis testing could refine our understanding, these steps are considered beyond the current scope, aiming to maintain focus within the limitations of our simplified model. Throughout the challenge, we have utilised historical data from a range of different years. As a result, there may be some challenges that could undermine the validity of our conclusions. Data from different years reflect different social, environmental, and political conditions. Changes in these factors over time could lead to inconsistencies in the relationship between predictor variables and the outcomes we extrapolate from our models. Ultimately because of the unpredictability of the factors affecting homelessness and the nature of the statistical models, it may be impossible to statistically identify and quantify the homeless as simply numerical units of data to compare against economic factors such as median housing value.

References

1. [1] House of Commons ODPM: Housing, Planning, Local Government, and the Regions Committee HC 61-I.(n.d.).

Available at: https://publications.parliament.uk/pa/cm200405/cmselect/cmodpm/61/61i.pdf.

2. Ministry of Housing, Communities & Local Government (2012). *Live tables on homelessness*. [online] GOV.UK.
Available at: https://www.gov.uk/government/statistical-data-sets/live-tables-on-homelessness.

3. Statista. (n.d.). *UK average household size 2022*. [online]
Available at: https://www.statista.com/statistics/295551/average-household-size-in-the-uk
[Accessed 2 Mar. 2024].

4. Manchester Homelessness Partnership. (n.d.). *Homelessness in Manchester*. [online]

Available at: https://mhp.org.uk/homelessness-in-manchester/.

5. GOV.UK. (n.d.). *Homelessness statistics*. [online]

Available at: https://www.gov.uk/government/collections/homelessness-statistics#statutory-homelessness.

6. www.manchester.gov.uk. (n.d.). *Intelligence Hub | Intelligence Hub | Manchester City Council*. [online]

Available at:
https://www.manchester.gov.uk/info/200088/statistics_and_intelligence/7611/intelligence_hub.

7. GOV.UK. (n.d.). *Council Tax: stock of properties, 2021*. [online]

Available at: https://www.gov.uk/government/statistics/council-tax-stock-of-properties-2021.

8. Cross Validated. (n.d.). *regression - McFadden's Pseudo-R^2 Interpretation*. [online] Available at: https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation#:~:text=A%20rule%20of%20thumb%20that.

7. Office for National Statistics (2023). *Families and households in the UK - Office for National Statistics*. [online]www.ons.gov.uk.
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2022.

Code Appendix

```
data = X; %load data from spreadsheet
yearsToPredict = [10, 20, 50]; % Example years
scatter(data.Year, data.TotalHousingUnits, [], data.Income, 'filled'); % Plot the original data with color representing income
hold on;
% Fit a multiple linear regression model
mdl = fitlm(data, 'TotalHousingUnits ~ TotalPopulation + Income');
% Predict for future years
startYear = max(data.Year); % Start predicting from observed year
endYear = startYear + 60; % Predict up to 60 years from the start year
yearsForPrediction = (startYear:endYear)';
predictTable = table(yearsForPrediction, 'VariableNames', {'Year'});
% Assuming linear growth rates for income and population
income_growth_rate = 0.0247; % Specify the income growth rate (e.g., 0.02 for 2% annual growth)
population_growth_rate = 0.0143; % Specify the population growth rate (e.g., 0.01 for 1% annual growth)
% Calculate projected income and population for future years starting from one year after the last observed year
initial_income = data.Income(end); % Initial income value (last observed value)
initial_population = data.TotalPopulation(end); % Initial population value (last observed value)
projected_income = initial_income * (1 + income_growth_rate) .^ (yearsForPrediction - startYear);
projected_population = initial_population * (1 + population_growth_rate) .^ (yearsForPrediction - startYear);
% Populate predictTable with projected Income and Population data
predictTable.Income = projected_income;
predictTable.TotalPopulation = projected_population;
predictedHousing = predict(mdl, predictTable);
% Plot the predicted data
plot(predictTable.Year, predictedHousing, '-');
xlabel('Year');
ylabel('Total Housing Units');
title('Housing Units in Manchester');
legend('Original Data', 'Predictions', 'Location', 'southeast');
% Highlight the predicted years and indicate them
for i = 1:length(yearsToPredict)
    yearToPredict = yearsToPredict(i);
    predictedIndex = find(predictTable.Year == yearToPredict);
    if ~isempty(predictedIndex)
        predictedValue = predictedHousing(predictedIndex);

        % Plot red crosses at predicted years
        plot(yearToPredict, predictedValue, 'rx', 'MarkerSize', 10);

        % Add text annotation indicating the predicted year
        text(yearToPredict, predictedValue, num2str(yearToPredict), ...
            'HorizontalAlignment', 'center', 'VerticalAlignment', 'bottom');
    end
end
predictedYears = (startYear:endYear)';
predictedHousingUnits = zeros(length(predictedYears), 1);


for i = 1:length(predictedYears)
   yearToPredict = predictedYears(i);
   predictedIndex = find(predictTable.Year == yearToPredict);


   if ~isempty(predictedIndex)
        predictedValue = predictedHousing(predictedIndex);
        predictedHousingUnits(i) = predictedValue;
   end
end
```

^^<u>A script that predicts the number the growth of housing units using a multi-linear regression model in Manchester.</u>

A script that calculates the recovery rate of at-risk individuals (Gamma) and the relapse rate from housed to at-risk (Delta)

```
import pandas as pd
homelessness_data = pd.read_csv('homelessness_data_bnh.csv')

# Prepare lists to hold calculated values for gamma and delta
gammas = [None]  # Initial None for the first delta value
deltas = [None]  # Initial None for the first gamma value

# Calculate gamma (Improvement rate) and delta (Relapse rate)
for i in range(1, len(homelessness_data)):
    # Calculate gamma
    change_eligible_not_homeless = (homelessness_data.loc[i, 'Eligible_Not_Homeless'] - homelessness_data.loc[i - 1, 'Eligible_Not_Homeless'])
    gamma = -change_eligible_not_homeless / homelessness_data.loc[i - 1, 'Eligible_Not_Homeless']
    gammas.append(gamma)

    # Calculate delta - Note we use the assumption that an increase in priority need might indicate relapse
    change_priority_need = (homelessness_data.loc[i, 'Priority_Need'] -
                            homelessness_data.loc[i - 1, 'Priority_Need'])
    if change_priority_need > 0 and change_eligible_not_homeless < 0:
        delta = change_priority_need / (homelessness_data.loc[i - 1, 'Priority_Need'] -
                                        homelessness_data.loc[i - 1, 'Eligible_Not_Homeless'])
        deltas.append(delta)
    else:
        deltas.append(None)  # Cannot calculate delta if we don't have a decrease in 'Eligible_Not_Homeless'

# Add the calculated values back to the DataFrame
homelessness_data['Gamma'] = gammas
homelessness_data['Delta'] = deltas

homelessness_data.fillna(value={'Gamma': 0, 'Delta': 0}, inplace=True)

# Print the DataFrame with the calculated values
print(homelessness_data)

# Calculate the average Gamma and Delta, ignoring 'None' values and preventing division by zero
average_gamma = sum(g for g in gammas if g is not None) / (len([g for g in gammas if g is not None]) or 1)
filtered_deltas = [d for d in deltas if d is not None]
average_delta = sum(filtered_deltas) / (len(filtered_deltas) or 1)

#Display average Gamma and Delta values
print(f"Average Gamma: {average_gamma}")
print(f"Average Delta: {average_delta}")
```

A script that calculates the rate of moving from being at-risk to homeless (Alpha) and rate of moving from homeless to housed (Beta)

```
import pandas as pd
available_houses_data = pd.read_csv('available_houses_bnh.csv')
homelessness_data = pd.read_csv('homelessness_data_bnh.csv')
# Initialize lists for alpha and beta calculations
alpha_list = [None]  # No alpha for the first year
beta_list = [None]  # No beta for the first year
# Calculate alpha and beta together since they depend on consecutive years of data
for i in range(1, len(homelessness_data)):
    # Alpha Calculation: Year over year change in total eligible households for homelessness
    increase_in_eligible_households = (homelessness_data.loc[i, 'Total_Eligible_Households'] -
                                       homelessness_data.loc[i-1, 'Total_Eligible_Households'])
    # Avoid division by zero by ensuring the previous value is at least 1
    alpha = abs(increase_in_eligible_households) / max(homelessness_data.loc[i-1, 'Total_Eligible_Households'], 1)
    alpha_list.append(alpha)
    # Beta Calculation: Change in the number of available houses year-to-year
    # Assuming available housing units data aligns with total eligible households data
    if i < len(available_houses_data):
        # Absolute value ensures that both increases and decreases in availability affect beta
        change_in_houses_available = (available_houses_data.loc[i, 'Houses_Available'] -
                                      available_houses_data.loc[i - 1, 'Houses_Available'])
        beta = abs(change_in_houses_available) / max(available_houses_data.loc[i - 1, 'Houses_Available'], 1)
        beta_list.append(beta)
    else:
        # If there isn't corresponding data in housing units, append None
        beta_list.append(None)
# Update the homelessness DataFrame with the calculated alpha and beta values
homelessness_data['Alpha'] = alpha_list
# Align the length of beta_list with homelessness_data in case of length mismatch
# This can happen if the housing units data has fewer records
while len(beta_list) < len(homelessness_data):
    beta_list.append(None)
homelessness_data['Beta'] = beta_list
# Calculate and print the average alpha and beta, excluding 'None' values
average_alpha = sum(filter(None, alpha_list)) / len([a for a in alpha_list if a is not None])
average_beta = sum(filter(None, beta_list)) / len([b for b in beta_list if b is not None])
# Display the DataFrame with the calculated Alpha and Beta values
print(homelessness_data[['Year', 'Alpha', 'Beta']])
#Display average Alpha and Beta values
print(f"Average Alpha: {average_alpha}")
print(f"Average Beta: {average_beta}")
```

A script that predicts the number of homeless people using the SEIR Model. This represents the script with the variables for Manchester and the collection of the 50 data points for the number of homeless people over the next 50 years

```python
import numpy as np
from scipy.integrate import odeint
import pandas as pd

# Defined parameters and initial conditions
annual_growth_rate = 0.01
alpha = 0.13040167618273908
beta = 0.12936643303985582
gamma = 0.5213003472739447
delta = 1.4210526315789473
A0 = 372.88
H0 = 2057
X0 = 373.2
N = 477408

# SEIR model differential equations including population growth
def d_states_dt(states, t, N, alpha, beta, gamma, delta, growth_rate):
    A, H, X = states
    new_population = growth_rate * N
    dA_dt = -alpha * A + delta * X - gamma * A + new_population
    dH_dt = alpha * A - beta * H
    dX_dt = beta * H - delta * X
    return [dA_dt, max(dH_dt, 0), dX_dt]  # Ensure never negative

# Time settings for the simulation over 50 years
years_to_simulate = 50
t = np.linspace(0, years_to_simulate, int(years_to_simulate) + 1)

# Solve the differential equations
states0 = [A0, H0, X0]
states = odeint(d_states_dt, states0, t, args=(N, alpha, beta, gamma, delta, annual_growth_rate))

# Extract the Homeless results
Homeless = states[:, 1]

# Prepare the data for the DataFrame
data = {
    'Year': np.arange(2008, 2008+years_to_simulate+1).tolist(),
    'Homeless': Homeless.tolist()
}

# Create the DataFrame
homeless_df = pd.DataFrame(data)

# Save the DataFrame to a CSV file
homeless_df.to_csv('homeless_projection.csv', index=False)

# Print success message
print('Homeless data for the next 50 years has been saved to homeless_projection.csv')
```

A script that predicts the number of homeless people using the SEIR Model. This is the script used for making an estimation for the homeless in Brighton and Hove. The same implementation was used for Manchester with different variables.

```python
import numpy as np
from scipy.integrate import odeint
import matplotlib.pyplot as plt
from scipy.stats import linregress

# Annual population growth rate for Manchester
annual_growth_rate = 0.01

# Use the calculated average values for transition rates
alpha = 0.13040167618273908  # Transmission rate from At-Risk to Homeless
beta = 0.12936643303985582  # Transmission rate from Homeless to Housed
gamma = 0.5213003472739447  # Recovery rate of At-Risk individuals
delta = 1.4210526315789473  # Relapse rate from Housed to At-Risk

# Initial populations in each compartment
A0 = 372.88  # Initial At-Risk population
H0 = 2057    # Initial Homeless population
X0 = 373.2   # Initial Housed (formerly homeless) population
N = 477408   # Total population

# SEIR model differential equations including population growth
def d_states_dt(states, t, N, alpha, beta, gamma, delta, growth_rate):
    A, H, X = states
    new_population = growth_rate * N  # New people added to the population annually
    dA_dt = -alpha * A + delta * X - gamma * A + new_population  # New at-risk
    dH_dt = alpha * A - beta * H  # New homeless
    dX_dt = beta * H - delta * X  # New housed

    # Prevent negative populations
    dA_dt = max(dA_dt, -A)
    dH_dt = max(dH_dt, -H)
    dX_dt = max(dX_dt, -X)

    return [dA_dt, dH_dt, dX_dt]

# Time settings for the simulation over 50 years
years_to_simulate = 50
t = np.linspace(0, years_to_simulate, years_to_simulate + 1)  # One entry per year

# Solve the differential equations
states0 = [A0, H0, X0]
states = odeint(d_states_dt, states0, t, args=(N, alpha, beta, gamma, delta, an-
nual_growth_rate))

# Extract the results
At_Risk, Homeless, Housed = states.T

# Specific years we want to extract data for
years_of_interest = [1, 2, 8, 10, 20, 30, 50]
data_points = {year: {} for year in years_of_interest}

# Extract data points for specific years
for year in years_of_interest:
    index = int(year)
    data_points[year]["At_Risk"] = At_Risk[index]
    data_points[year]["Homeless"] = Homeless[index]
    data_points[year]["Housed"] = Housed[index]

# Print the extracted data
for year, data in data_points.items():
    print(f"Year {year}: {data}")

# Plot the results
plt.figure(figsize=(12, 8))
plt.plot(t, At_Risk, label='At-Risk (A)', color='orange')
plt.plot(t, Homeless, label='Homeless (H)', color='red')
plt.plot(t, Housed, label='Housed (X)', color='green')
plt.title('Adapted SEIR Model Dynamics with Population Growth (Manchester)')
plt.xlabel('Time (years)')
plt.ylabel('Population')
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.15), shadow=True, ncol=3)

plt.tight_layout()  # Adjust the padding to make room for the legend
plt.grid(True)
plt.show()
```

A portion of the script containing the sensitivity analysis

```
# Hypothetical observed data points for the Homeless population (for specific years)
observed_H_data = np.array([2430, 2572, 5360])
# Corresponding prediction years in the simulation
prediction_years = [0, 1, 7]  # These are indexes in your simulation that correspond to the
observed data points
# Make sure this matches the actual observed data

# Extract the predicted Homeless values for the same specific years:
seir_H_predictions = states[:, 1]  # Assuming column 1 corresponds to the Homeless compartment
'H'
predicted_H_data = np.array([seir_H_predictions[year] for year in prediction_years])

# Verify observed and predicted arrays have the same length
assert len(observed_H_data) == len(predicted_H_data), "The length of observed and predicted
data arrays must match."

# Calculate R-squared using linear regression from SciPy:
slope, intercept, r_value, p_value, std_err = linregress(observed_H_data, predicted_H_data)
r_squared = r_value**2

print(f"The R-squared value for the Homeless compartment is: {r_squared}")
```

A script which creates and plots a linear regression model in addition to a confidence interval. This was used in estimating available housing units for Brighton and Hove.

```
% Load data used in the plot
BH.special = [10, 20, 50];
BH.data = readtable("stats.xlsx", Sheet="Brighton and Hove", ...
    Range="A3:B31");

BH.data.Properties.VariableNames = ["Year", "Units"];

% Create a linear model and predict future data
BH.yrs = (min(BH.data.Year):year(datetime("now"))+max(BH.special)+10)';
BH.mdl = fitlm(BH.data);

BH.pred = table(BH.yrs, VariableNames={'Year'});
[BH.pred.Units, BH.pred.conf] = predict(BH.mdl, BH.pred);

% Plot the data
plot(BH.data.Year, BH.data.Units, "o", Color="Black", MarkerSize=20, ...
    Marker=".");
hold on;
plot(BH.pred.Year, BH.pred.Units, "-", Color="Blue");

% Plot uncertainty (95% confidence)
fill([BH.pred.Year; flipud(BH.pred.Year)], ...
    [BH.pred.conf(:,1); flipud(BH.pred.conf(:,2))], ...
    "k", FaceAlpha=0.1, EdgeColor="none");

% Output R^2 value (5dp)
sprintf("R^2: %.5f", BH.mdl.Rsquared.Ordinary)

% Mark predicted years
for BH_i = 1:length(BH.special)
    BH_i_year = year(datetime("now")) + BH.special(BH_i);
    BH_i_units = BH.pred.Units(BH.pred.Year == BH_i_year);

    plot(BH_i_year, BH_i_units, "rx", LineWidth=2, MarkerSize=10, ...
        Color="Black");

    clearvars("BH_i", "BH_i_year", "BH_i_units");
end
hold off;

% Annotate
xlabel("Year");
ylabel("Total Housing Units");
title("Housing in Brighton & Hove");
legend("Original Data", "Linear Prediction");
legend("Existing Data", "Predictions", "Location", "southeast");

% Remove all locals from workspace
clear BH;
```

A script used to determine if the data is over dispersed or not in order to decide which linear regression model is suitable to use.

```
#This will determine whether we use a poisson model or a negative binomial regression
import pandas as pd

# Load your dataset
df = pd.read_csv('dataset.csv')

# Calculate the observed mean and variance of the count data
mean_homeless = df['TotalHomeless'].mean()
variance_homeless = df['TotalHomeless'].var()

# Calculate the dispersion statistic (variance-to-mean ratio)
dispersion_statistic = variance_homeless / mean_homeless
print(f"Dispersion Statistic: {dispersion_statistic}")

# Check for overdispersion
if dispersion_statistic > 1:
    print("The count data is overdispersed.")
else:
    print("The count data is not overdispersed.")
```

A script showing the implementation of the negative binomial regression model along with the calculation of the necessary variables.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import glm
from statsmodels.genmod.families import NegativeBinomial
from functools import reduce
import matplotlib.pyplot as plt

# Load data from separate CSV files; ensure they all have a 'Year' column and the same length
housing_units_df = pd.read_csv('housing_units_data.csv')
sales_price_df = pd.read_csv('sales_price_data.csv')
population_df = pd.read_csv('population_data.csv')
income_df = pd.read_csv('income_data.csv')
homelessness_df = pd.read_csv('homeless_projection.csv')

# Merge all dataframes on the 'Year' column
df_merged = reduce(
    lambda left, right: pd.merge(left, right, on=['Year']),
    [housing_units_df, sales_price_df, population_df, income_df, homelessness_df])

# Handle any missing values if necessary
df_merged = df_merged.dropna()

# Fit a Negative Binomial regression model; estimate alpha using a preliminary model or specify it directly
# Define the regression formula, ensure column names match those in the merged dataframe exactly
formula = 'TotalHomeless ~ TotalHousingUnits + MedianSalesPrice + TotalPopulation + MedianIncome'

# Initial Poisson model to estimate alpha
poisson_model = glm(formula, data=df_merged, family=sm.families.Poisson()).fit()
alpha_est = poisson_model.pearson_chi2 / poisson_model.df_resid

# Negative Binomial model
nb_model = glm(formula, data=df_merged, family=NegativeBinomial(alpha=alpha_est)).fit()

# Print the model summary
print(nb_model.summary())

# Visualization of coefficients
coefs = pd.DataFrame({
    'coef': nb_model.params.values[1:],  # Excludes Intercept
    'err': nb_model.bse.values[1:],  # Excludes Intercept
    'varname': nb_model.params.index[1:]  # Excludes Intercept
})

fig, ax = plt.subplots(figsize=(10, 6))
coefs.plot(x='varname', y='coef', kind='bar', ax=ax, yerr='err', capsize=4)
plt.title('Coefficient Estimates From Negative Binomial Regression')
plt.ylabel('Coefficients')
plt.axhline(y=0, color='black', linestyle='--')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```