Faculty of Natural and Mathematical Sciences
Department of Informatics

King's College London
Strand Campus, London,
United Kingdom

7CCSMPRJ

Individual Project Submission 2023/24

Name: Alexander Smerdon

Student Number: 23031306

Degree Programme: Data Science

Project Title: Urban Planning, Infrastructure and Health Inequalities in North East London

Supervisor: Grigorios Loukides

Word Count: 11984

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

☑ I **agree** to the release of my project

□ I **do not** agree to the release of my project

Signature: A.SMERDON **Date:** August 2, 2024

Department of Informatics
King's College London
United Kingdom

7CCSMPRJ Individual Project

Urban Planning, Infrastructure and Health Inequalities in North East London

Name: **Alexander Smerdon**
Student Number: 23031306
Course: Data Science

Supervisor: Grigorios Loukides

This dissertation is submitted for the degree of MSc in Data Science.

Abstract

Urban inequality can be defined as the disparities in access to services, resources, and opportunities among different social groups, manifesting in areas such as income, education, housing, and healthcare. Whilst previous research has addressed these disparities through real-estate data and socio-economic indicators, there has been limited focus on North East (NE) London.

This project aims to enhance understanding of inequalities in NE London by employing and evaluating various statistical and machine learning methods across multiple datasets. Specifically, regression techniques (including Ridge, Lasso, and spatial regression) are utilised to build models predicting life expectancy and income. Additionally, clustering techniques are applied to healthcare infrastructure data to categorise neighbourhoods and highlight areas being under-served by current provisions.

By comparing NE London with broader data from London and the UK, this research highlights the unique challenges faced by NE London boroughs. The primary goal is to evaluate the effectiveness of these models in capturing underlying inequalities and to showcase a comprehensive understanding of statistical and computational methods. The findings contribute to a deeper comprehension of urban inequalities in NE London and could inform future research and policy development.

Acknowledgements

I would like to thank my supervisor, Grigorios Loukides, for his guidance and feedback throughout this project. I also appreciate the resources and support provided by the Department of Informatics at King's College London, with special thanks to my personal tutor Letizia Gionfrida. Lastly, thank you to my family and friends for their support.

Contents

List of Figures

Listings

Nomenclature

NHS (National Health Service):

The publicly funded healthcare system of the United Kingdom, providing a range of health services including hospitals and general practitioners (GPs).

TfL (Transport for London):

The government body responsible for the transport system in Greater London, including the Underground, buses, and other public transport services.

ONS (Office for National Statistics):

The executive office of the UK Statistics Authority, responsible for collecting and publishing statistics related to the economy, population, and society at national, regional, and local levels.

London Datastore:

An online data-sharing portal where users can access datasets related to London, including economic, demographic, and infrastructure related data.

Regression Techniques:

Statistical methods used to model the relationship between a dependent variable and one or more independent variables. Examples include Ridge, Lasso, and spatial regression.

Clustering Techniques:

Methods used to group a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Examples include K-means, fuzzy K-means, DBSCAN, and leader-follower clustering.

Geospatial Data:

Data that is associated with a specific location. In this context, it refers to the geographical location of hospitals, GPs, and other urban infrastructure.

Life Expectancy:

The average period that a person is expected to live, used as a key indicator in assessing health outcomes across different regions.

Income:

The monetary earnings of individuals or households, used as a measure of economic status.

Chapter 1

Introduction

Inequality in an urban context can be defined as the differences in access to services, resources, and opportunities among various social groups. These differences often manifest as disparities in income, education, housing, and healthcare. Previous work has focused on modelling these disparities using real-estate data and other socio-economic indicators, but there has been limited emphasis on the specific context of North East (NE) London.

This project aims to use and evaluate a variety of statistical and machine learning methods on different datasets to better understand the inequalities facing NE London, with the objective of improving outcomes. Specifically, it will evaluate regression techniques such as Ridge, Lasso, and spatial regression to build models predicting life expectancy and income. Clustering techniques will also be employed to group neighbourhoods based on healthcare infrastructure.

The choice of these techniques is motivated by their ability to handle complex datasets and capture various dimensions of inequality. Ridge and Lasso regressions are chosen for their effectiveness in handling multicollinearity and feature selection, respectively. Spatial regression is included to account for spatial dependencies in the data. Clustering methods are used to identify patterns and group similar neighbourhoods, which can inform targeted interventions. The primary results show that the regression models explain a significant portion of the variance in life expectancy and income, though with varying degrees of accuracy. The clustering analysis - depending on the method used - reveals distinct groups of neighbourhoods with similar healthcare infrastructure, highlighting areas that may require additional resources.

The remainder of this report is structured as follows: The Background section provides context to the struggles facing North East London, and the motivation for using machine learning methods to improve outcomes. The Literature Review section discusses the work of previous papers in this field, and how this dissertation attempts to develop these ideas. This section also provides contextual information for the statistical methods being used. The Methodology section details the data sources used and the implementation of the statistical/machine learning techniques. The Results section presents the findings from the regression and clustering analyses. There is then a discussion on the Legal, Social, Ethical and Professional Issues, and then the Conclusion summarises the key findings, discusses limitations, and suggests directions for future research. Although attempts have been made to help with reader accessibility, this paper assumes some background knowledge in statistical and machine learning techniques.

Chapter 2

Background

This project is motivated by the current rise in inequality that the UK is experiencing. Exacerbated by the Covid-19 pandemic and the recent cost-of-living crisis, the UK has seen an unprecedented rise in the amount of people living in both relative and absolute poverty. According to research done by the Child Poverty Action Group, this includes 4.2 million children in the year 2022, an increase of 800,000 since 2010-11 [1]. As one of the poorest areas of the country [2], North East London has been on the forefront of these developments.

Using the area covered by NHS North East London [3], the boroughs this project will focus on are: Barking and Dagenham, City of London, Hackney, Havering, Newham, Redbridge, Tower Hamlets and Waltham Forest. This is an area with a population of roughly two million, and is growing year on year [4]. This area will be compared to the rest of London.

Areas of NE London have the lowest life expectancy in London, as well as some of the lowest across the whole UK. As one of the most deprived areas of the country, the London borough of Barking and Dagenham is emblematic of the struggles currently facing many areas of North East London. In the previous two years (post-pandemic), life expectancy has fallen to the lowest since 2007 with average male life expectancy in Barking and Dagenham being 76.26 years according to research done by the Standard [5], compared with the UK average of 79.0 [6]. The borough also recorded some of the lowest income rates in the country, with the mean average salary in 2022 being £31,800 - lower than the UK average of £39,400 and far lower than the London average of £58,300 [7]. These stark differences highlight the urgent need to improve outcomes for the area, and this project aims to find the interventions that would have the most positive impact.

Chapter 3

Literature Review

3.1 Previous Research

Research into urban inequality is a highly active field due to its real-world significance, with numerous studies investigating its causes and impacts across cities worldwide. This has led to many papers using different machine learning techniques to both understand inequality and model outcomes. Some of these, and ways in which my dissertation attempts to improve on them, will be discussed below.

Research into urban health inequalities has shown significant differences in access to healthcare and quality of life across Socio-Economic groups. Bennett et al. (2023)[8] provide a perspective on the relationship between life expectancy and house prices in London from 2002 to 2019. Using geo-spatial data, Bayesian hierarchical models were employed to estimate death rates and house prices at a detailed level. Their analysis found that life expectancy increases were correlated with higher initial house prices and significant house price growth, particularly in already expensive areas. In this project, housing data will be used in the regression modelling.

Inequality affects not only direct health outcomes but also how people perceive environmental stressors like noise pollution. For example, Tong's (2022)[9] PhD thesis demonstrates that while actual noise levels do impact health, noise perception is heavily influenced by urban planning and Socio-Economic conditions. Using geo-spatial data from London and New York City, Tong employed ridge regression and Bayesian models to study noise complaints and urban morphology at the city level, Socio-Economic factors at the regional level, and traffic noise impacts on sleep and mental health at the national level. The key finding is that urban planning significantly affects noise-induced health issues, with factors like building density, street layout, and Socio-Economic conditions playing a more significant role in noise perception than the actual noise levels. This highlights the need for addressing

Socio-Economic disparities through effective urban planning and noise management to promote healthier cities. With this in mind, this project will consider transport infrastructure when modelling inequality.

Clustering algorithms have also been used to identify patterns and map inequalities in urban settings. For example, in Kadeem Khan's (2019)[10] thesis on Nairobi, K-Means clustering was applied to geo-spatial data, including satellite imagery and census information, to create residential typologies. This analysis revealed significant disparities in access to services and infrastructure across different neighbourhoods, highlighting areas in need of investment. The study demonstrated that clustering can effectively uncover underlying patterns in urban data, providing a detailed understanding of spatial inequalities. However, these techniques also have limitations. They depend heavily on the quality and granularity of available data, which can vary significantly across different contexts. Additionally, clustering algorithms may produce results that are difficult to interpret without validation and domain-specific knowledge .

In a similar vein, K-means clustering has also been explored for segmenting urban data and identifying patterns. This literature review paper of unsupervised ML techniques by Bochra Hadj Kilani (2023)[11] discusses its widespread use in urban design analysis, highlighting its effectiveness in providing insights into urban planning. However, the study highlights several limitations of K-means, such as sensitivity to initial centroid values, challenges in determining the optimal number of clusters, and scalability issues with large datasets. To address these constraints, the paper suggests integrating K-means with other algorithms, like DBSCAN, for more effective handling of complex urban data.

In this research, both regression and clustering techniques are employed to analyse inequalities in North East London. By integrating various data sources, such as socio-economic, environmental, and urban planning data, the aim is to map out disparities and discover targeted interventions to improve life expectancy and income distribution in the region. The regression analysis will predict life expectancy and income. Clustering techniques will be utilised to understand health inequalities and identify specific areas in need of investment. Building on the insights from the study by Bochra Hadj Kilani (2023)[11], which highlights the limitations of K-means clustering in urban design analysis, this research will explore alternative algorithms to achieve more effective geospatial clustering. The results are intended to influence policy and planning decisions, ultimately improving outcomes in North East London.

3.2 Machine Learning Methods

This project will use a variety of statistical models to help understand the biggest causes of inequalities, with the main one being regression models. Two different models will be built. The first will attempt to predict life expectancy in London, and the second will attempt to predict income in London.

Multiple Linear Regression with Regularisation (Lasso and Ridge)

As one of the simplest forms of regression, multiple linear regression [12] allows for the relationship between multiple independent variables and a dependent variable to be modelled. For modelling life expectancy, datasets on healthcare access, childhood obesity rates, housing tenure, and ethnicity will be used. For modelling income, datasets on whether English is spoken at home, GCSE grades of pupils eligible for free school meals, ethnicity, accessibility to transport, and higher education rates will be used.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{3.1}$$

Linear Regression Formula

These datasets may cause instability in the estimation of regression coefficients however, as the independent variables involved may be collinear. For example, life expectancy and childhood obesity rates are both indicators of health outcomes, and so are likely to be correlated. To account for this, regularisation methods will be used to make the model more accurate.

Lasso (L1) [13] regularisation works similarly to ordinary multiple linear regression, where the objective is to minimise the Ordinary Least Squares (OLS) loss function. However, in addition to minimising prediction errors, Lasso introduces a penalty term that encourages sparsity in the model by penalising the absolute values of the regression coefficients. This penalty term has the effect of shrinking some coefficients towards zero and can lead to the removal of irrelevant predictors from the model. Specifically, when the penalty term exceeds a certain threshold for a coefficient, the corresponding predictor is excluded from the model as its coefficient is set to zero. This helps minimise collinearity within the model.

$$SSE_{L1} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.2}$$

Lasso (L1) Regression Formula

Ridge (L2) [14] regularisation also minimises the sum of squared errors in Ordinary Least Squares (OLS) regression while adding a penalty term proportional to the squared magnitudes of the regression coefficients. This penalty, controlled by a regularisation parameter λ, mitigates multicollinearity and overfitting by constraining coefficient magnitudes. In short, Ridge regression aims to optimise model fit while stabilising predictions against input data variations. Unlike L1 regression, however, variables are not removed from the model and are instead just moved towards zero.

$$SSE_{L2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.3}$$

Ridge (L2) Regression Formula

Spatial regression techniques

Spatial regression [15] extends traditional regression techniques by including information about the spatial structure of the data into the model. This accounts for any uncertainty within the model caused by unseen factors relating to spatial dependencies, thereby improving the model's predictive accuracy. One method of accomplishing this is by using spatial lag models. Spatial lag models are particularly well-suited for addressing spatial autocorrelation in the dependent variable. By acknowledging that observations in nearby locations may have similar values due to unobserved spatial factors, these models provide more reliable estimates of the relationships between predictors and life expectancy. This leads to a more accurate assessment of the spatial patterns and determinants of life expectancy within North East London. Other techniques for spatial regression exist, such as Spatial Error Models and Geographically Weighted Regression, however due to the time constraints of the project these were not explored.

$$y = \rho W y + X\beta + \epsilon \tag{3.4}$$

Spatial Lag Model Formula

3.2.1 Model performance metrics

To assess the quality of the models in this project, standard statistical modelling performance metrics will be used. These are discussed in detail below.

The R^2 score, or the coefficient of determination, indicates the proportion of variance in the dependent variable that is predictable from the independent variables [16]. An R^2 score of 1 means the model explains all the variance, while a score of 0 means it explains none. In the context of this project, an R^2 score closer to 1 would indicate a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

R^2 (Coefficient of Determination)

The Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values [16]. A lower MSE indicates better model performance. In this project, MSE will be used to evaluate how well the models predict life expectancy and income. It's important to note that the magnitude of the MSE is relative to the scale of the data; for instance, an MSE of 10,000 might be acceptable in a context where the data values are in the hundreds of thousands.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.6}$$

Mean Squared Error (MSE)

Cross-validation [16] is a technique used to assess how a model generalises. It involves partitioning the data into multiple subsets, training the model on some subsets and validating it on the remaining ones. This process is repeated several times (folds) to ensure that the model's performance is consistent across different subsets of data. Cross-validation helps in detecting overfitting and provides a more robust estimate of model performance.

The coefficients in a regression model represent the relationship between each independent variable and the dependent variable [16]. For example, in the income model, a coefficient of 227.23 for "English Spoken at Home" indicates that a 1% increase in the proportion of residents who speak English at home is associated with an approximate £227.23 increase in average income, holding all other variables constant. Understanding these coefficients helps in interpreting the impact of each predictor on the outcome variable. This in turn will help identify the factors that most impact outcomes.

A Q-Q plot (quantile-quantile plot) compares the distribution of residuals to a normal distribution [16]. A residual is the difference between an actual value and a value predicted by the model. If the residuals follow a normal distribution, the points on the Q-Q plot will lie approximately along a straight line. Deviations from this line indicate departures from normality, which can suggest issues such as skewness or kurtosis in the residuals. This in turn means a weaker model. A good

Q-Q plot would show points closely aligned to the line, indicating that the residuals are normally distributed.

Residual plots display the residuals on the vertical axis and the predicted values on the horizontal axis (MSE) [16]. Ideally, the residuals should be randomly scattered around zero with no apparent pattern. This randomness indicates that the model's assumption of normality are valid and that the model captures the relationship between the variables well. Patterns or trends in the residual plot can indicate issues such as non-linearity, heteroscedasticity, or the presence of outliers.

Predicted vs actual plots compare the model's predicted values against the actual values (MSE) [16]. A perfect model would have all points lying on the 45-degree line (where predicted values equal actual values). Deviations from this line indicate prediction errors. A good plot would show points closely clustered around the line, suggesting that the model accurately predicts the outcomes. The effectiveness of these metrics should be interpreted in the context of the data and the specific objectives of the project. For instance, while a high R^2 score and a low MSE are desirable, the real-world applicability and interpretability of the model are equally important (as previously mentioned with MSE).

3.2.2 Clustering Methods

When it comes to using clustering algorithms to understand inequalities in North East London, this project will compare several approaches to see which works best. These clustering techniques include: K-means, fuzzy K-means, DBSCAN, and leader-follower clustering. These algorithms will use geospatial data about London NHS hospitals and GPs, to try and find trends in the data and see the areas / neighbourhoods that are being under-served by current infrastructure provisions. With the populations of each London borough predicted to rise, this analysis will also help with identifying areas most in need of future health infrastructure. Efficiency of the algorithms will also be discussed, however this is not particularly relevant to this specific project as the Hospital dataset being used is relatively small.

K-means

The K-means algorithm [17] works by first specifying the number of clusters, then iteratively assigning data points to the nearest cluster centroid. The centroids are then updated based on the mean of the points assigned to each cluster. This process continues until convergence, resulting in clusters that minimise the within-cluster sum of squares. In this context, K-means clustering will be applied to geospatial data on London NHS hospitals and GPs. By analysing the distribution of healthcare facilities and identifying clusters with similar spatial characteristics, the aim is to uncover areas or neighbourhoods facing the longest wait times for healthcare services. K-means is computationally efficient with a time complexity of O(knT), where k is the number of clusters, n is the number of data points, and T is the number of iterations until convergence, making it suitable for geospatial data.

$$\min_{C} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 \tag{3.7}$$

K-means Clustering Formula

Fuzzy K-means

The fuzzy K-means algorithm [17] is an extension of the basic K-means algorithm, allowing data points to be assigned to multiple clusters with varying degrees of membership. Unlike traditional K-means, where each point belongs exclusively to one cluster, fuzzy K-means assigns membership values to each point indicating the degree of belongingness to each cluster. This flexibility provides a more nuanced representation of spatial patterns, particularly in areas where hospitals/GPs may serve multiple neighbourhoods or boroughs. The time complexity is similar to K-means, at O(knT), with the added complexity of calculating membership values.

$$\min_{C,U} \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}^{m} \|x_i - \mu_k\|^2 \tag{3.8}$$

Fuzzy K-means Clustering (Fuzzy C-means) Formula

DBSCAN

The DBSCAN algorithm [18] works by identifying dense regions of points separated by areas of lower density in the data space. It defines clusters as continuous regions of high density, allowing for the detection of arbitrary-shaped clusters and handling outliers effectively. DBSCAN requires two parameters: epsilon (ϵ), which defines the maximum distance between two points to be considered neighbours, and minPts, the minimum number of points required to form a dense region. In the context of this project, DBSCAN will also be applied to the geospatial data on London NHS hospitals and GPs to identify clusters of facilities serving densely populated areas with potential disparities in healthcare access. The time complexity of DBSCAN is $O(n \log n)$, making it relatively efficient for large datasets, though it can become computationally demanding with particularly massive datasets.

Leader-follower clustering

The Leader-follower algorithm [17] works by iteratively partitioning the dataset into leader points, which act as centroids for their respective clusters, and follower points, which are assigned to the nearest leader point - but only if they fall within a defined distance threshold of that leader point. If a follower is too far away, it becomes a the new leader cluster centroid. By iteratively updating the leaders and followers, the algorithm converges to a partitioning of the data into cohesive clusters. Leader-follower clustering offers advantages in terms of scalability and sensitivity to noise, making it suitable for large datasets with complex spatial distributions. As with the other clustering algorithms, the aim is to understand areas that may be currently under-served by current healthcare infrastructure. The time complexity is O(n) for each iteration, which means it is not so efficient for larger datasets.

3.2.3 Excluded Clustering Methods

Hierarchical Clustering

Hierarchical clustering [19] was not chosen for this project primarily due to its less flexible approach for geospatial clustering. Hierarchical clustering lacks the ability to reassign points once merged or split, which is a limitation for geospatial data where boundaries of clusters might need adjustment. This rigidity can lead to less accurate representations of the data's natural groupings. While hierarchical clustering can reveal nested data structures, it may not perform well for the primary objective of identifying clusters based on spatial proximity and density.

Gaussian Mixture Models (GMM)

GMMs [20] were considered but ultimately not chosen for this project. A GMM assumes that data points are generated from a mixture of several Gaussian distributions, which might not align well with the irregular shapes of geospatial clusters. The computational complexity of GMMs, especially for large datasets, and their sensitivity to the initial parameter settings can make them less practical. In the context of geospatial clustering, where clusters can have irregular and non-Gaussian shapes, GMMs may not provide the most accurate results. Additionally, the convergence of GMMs can be slow and require significant computational resources, further limiting their suitability for this project.

By focusing on algorithms better suited for the nature of geospatial data, such as K-means, Fuzzy K-means, DBSCAN, and Leader-Follower, the analysis ensured that the chosen methods were both computationally feasible and effective in capturing spatial relationships.

Chapter 4

Approach

4.1 Datasets

For this project, a variety of online datasets were used. The primary source of data was Trust for London, a charitable organisation focused on tackling poverty and inequality [1]. Trust for London provided datasets that offered insights into social and economic disparities across different boroughs (such as housing tenure) [2] [3]. These datasets helped in understanding the Socio-Economic factors affecting income and life expectancy.

The London Datastore, an initiative by the Greater London Authority, served as another source of data [4]. This government website provides historic datasets covering numerous areas (such as demographics, economics, education, health, and transport). The datasets from the London Datastore were valuable for obtaining borough-level information (such as population statistics [5], economic indicators, and social metrics) [6].

Kaggle, a platform known for its repository of datasets shared by the data science community, provided additional demographic and Socio-Economic datasets [7]. Kaggle is where the dataset of London Hospitals was sourced [8].

The Office of National Statistics (ONS) was utilised for acquiring data related to health, life expectancy, and educational attainment [9]. The ONS datasets are useful for understanding broader trends and contextualising the findings within UK [10] [11] [12].

OpenStreetMap (OSM) was used to gather geospatial data (such as the locations of hospitals and London Underground stations) [13]. OSM's open data was used for the geospatial analysis component of the models, particularly for calculating distances and accessibility scores.

Wikipedia served as a supplementary source, particularly for historical and contextual information about London boroughs, as well as source for London Underground station location data [14].

Finally, the Transport for London (TfL) website provided journey data between London Underground stations, used in the Income Model [15]. This dataset was used for constructing the transport accessibility model, as it offered insights into commuting patterns and the connectivity of different areas within London.

Combining these datsets allowed for the creation of predictive models that integrate Socio-Economic and geospatial factors to predict income and life expectancy across London boroughs.

4.2 Software and Tools

The main software used in this project is Python, specifically the interactive computing platform Jupyter Notebook. Within the Jupyter Notebook, several libraries are used for data analysis, modelling, and visualisation:

- **Pandas** [21]: Used for data manipulation and analysis, particularly for reading data from various file formats, cleaning data, and performing exploratory data analysis. For example, Pandas is used to load and preprocess the income and accessibility datasets, merge them, and prepare them for modelling.

- **NumPy** [22]: In this project, NumPy is used for numerical operations, such as normalising the features of the datasets and handling numerical calculations.

- **Scikit-learn** [23]: A machine learning library providing tools for data mining and analysis. It is extensively used for building and evaluating the regression models:

  - `train_test_split` and `cross_val_score` from `sklearn.model_selection` are used for splitting datasets into training and testing sets and performing cross-validation to ensure the models' generalisability.

  - `KFold` for k-fold cross-validation to assess the stability and performance of the models.

  - `Ridge` from `sklearn.linear_model` for implementing Ridge regression models to handle multicollinearity and overfitting.

  - `Lasso` for implementing Lasso regression models to perform feature selection and regularisation.

  - Metrics such as `mean_squared_error`, `r2_score`, and `mean_absolute_error` from `sklearn.metrics` for evaluating model performance and ensuring the accuracy of predictions.

- **Matplotlib** [24]: Used for creating static, animated, and interactive visualisations. In this project, Matplotlib is used to create various plots such as scatter plots, line plots, and residual plots to visualise the results of the regression models and the distribution of residuals.

- **Seaborn** [25]: Built on top of Matplotlib, Seaborn provides a high-level interface for drawing statistical graphics. It is used for creating complex visualisations such as heatmaps and pair plots, which help in understanding the relationships between different variables in the dataset.

- **Scipy** [26]: Used for scientific and technical computing. In this project, Scipy's stats module is used for generating Q-Q plots to assess the normality of residuals in the regression models.

For clustering analysis, additional libraries and algorithms are used:

- **Folium** [27]: Used for creating interactive maps to visualise clustering results and geospatial data, providing a spatial context to the clusters formed. Uses Leaflet for map markers [28], whilst OpenStreetMap provided the background map [29].

- **skfuzzy** [30]: Provides tools for fuzzy clustering, such as the fuzzy k-means algorithm, which allows for overlapping clusters and a more nuanced analysis of the data.

- **Scikit-learn Clustering Algorithms** [23]:

  - `k-means` and `DBSCAN` for traditional clustering analysis to identify distinct groups within the data.

  - `pairwise_distances` for implementing the Leader-Follower clustering algorithm, which helps in identifying clusters based on distance metrics.

Finally, QGIS is used to compare cluster results to a map of life expectancy in each borough, providing a visual analysis of the spatial distribution of life expectancy across different clusters [31].

4.3 The Life Expectancy Model

4.3.1 Socio-economic model

The process of identifying relevant datasets for the Life Expectancy model involved extensive trial and error. Models built using individual independent variables were tested to identify datasets that would provide stable and meaningful predictions. The following independent variables were chosen for the Socio-Economic model:

- Percentage of rented households from Local Authorities or Housing Associations.
- Ethnic makeup of a borough (percentage who are different ethnic minorities).
- Child obesity rate in a borough (measured in Year 6 of Primary School).
- Percentage of people who speak English at home.

Forward sequential selection, where variables are added only if they have an impact on the model, was used to select the datasets.

The Socio-Economic data was loaded and pre-processed, with all relevant columns converted to numeric format and missing values handled appropriately. The dataset was then split into training and testing subsets (using an 80-20 split) with a fixed random state to ensure reproducibility.

```python
# Splits data into training and testing sets
X = df[['Rented_from_LA_or_HA', 'Minority', 'Obesity', 'Speak_English_at_Home']]
y = df['Life_Expectancy']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Listing 4.1: Life Expectancy Dataset Training and Testing

The Ridge regression model was built to allow for comparison. The training phase involved fitting the model to the training data and performing 5-fold cross-validation. This approach tried to ensure that the model was stable and provided meaningful predictions.

```python
from sklearn.linear_model import Ridge
from sklearn.model_selection import cross_val_score, KFold

# Initialises and trains Ridge regression model
ridge_model = Ridge(alpha=1.0)
ridge_model.fit(X_train, y_train)
```

```python

# Performs k-fold cross-validation for Ridge regression model
kf = KFold(n_splits=5, shuffle=True, random_state=42)
cv_mse_scores_ridge = cross_val_score(ridge_model, X, y, cv=kf,
    scoring='neg_mean_squared_error')
cv_r2_scores_ridge = cross_val_score(ridge_model, X, y, cv=kf,
    scoring='r2')
```

Listing 4.2: Building the Ridge Regression Model (Life Expectancy)

Aforementioned performance metrics, including Mean Squared Error (MSE) and R-squared (R^2), were calculated to evaluate the model. This will be discussed in Chapter 5.

4.3.2 Geospatial model

In addition to Socio-Economic factors, geographical data was also incorporated into the Life Expectancy model. This involved creating a dataset that included the number of hospitals in each borough and the average distance to the nearest hospital.

The geospatial data was loaded and preprocessed, including converting the coordinates of hospital locations to a consistent coordinate system and calculating the centroid of each borough. The distance between each borough's centroid and the nearest hospital was calculated to create a new feature: the average distance to the nearest hospital.

```python
def nearest_hospital_distance(row, hospitals_gdf):
    centroid = row['centroid']
    nearest_geom = hospitals_gdf.geometry.distance(centroid).min()
    return nearest_geom
```

Listing 4.3: Average Distance to Nearest Hospital

The data for the geospatial model was then prepared, with the hospital count and distance features used as independent variables and the life expectancy as the dependent variable. To account for spatial dependencies between boroughs, a Maximum Likelihood Lag (ML_Lag) regression model was used. As mentioned in Chapter 3, this model considers the influence of neighbouring regions on each other, which is particularly important in geographical data where the characteristics of one region can affect its neighbours. The spatial weights matrix was constructed using the K-Nearest Neighbours (KNN) method (with k=5), which defines the five neighbouring boroughs based on their geographical proximity.

```python
boroughs_train = boroughs
w_train = KNN.from_dataframe(boroughs_train, k=5)
```

```
model_geo = ML_Lag(y_geo, X_geo, w=w_train, name_y='life_expectancy
    ', name_x=['constant', 'hospital_count', '
    nearest_hospital_distance'])
```

Listing 4.4: Training the Geospatial Model

The ML_Lag model was trained using the prepared geospatial data, and the predictions from this model were obtained. This model helps to capture the spatial autocorrelation present in the data, providing a more accurate understanding of how geographical factors influence life expectancy. The accuracy of this model will be discussed in Chapter 5.

4.3.3 Combination of the two models

The predictions from both the Socio-Economic and geospatial models were combined to form a meta-model. This involved using the predictions from the two individual models as inputs for a final regression model. The combined model was then trained and evaluated using the same approach, allowing for a prediction of life expectancy across London boroughs. Both Ridge and Lasso models were built.

```
# Predictions from the geospatial model
y_pred_geo = model_geo.predy
# Combines predictions from both models
ridge_predictions = ridge_model.predict(X)
X_combined = np.column_stack((ridge_predictions, y_pred_geo))
# Trains meta-model
meta_model = Ridge(alpha=1.0)
X_train_combined, X_test_combined, y_train_combined,
    y_test_combined = train_test_split(X_combined, y, test_size=0.2,
     random_state=42)
meta_model.fit(X_train_combined, y_train_combined)
y_pred_meta = meta_model.predict(X_test_combined)
# Evaluates the combined model
mse_meta = mean_squared_error(y_test_combined, y_pred_meta)
r2_meta = r2_score(y_test_combined, y_pred_meta)
# Performs k-fold cross-validation for the combined model
cv_mse_scores_combined = cross_val_score(meta_model, X_combined, y,
     cv=kf, scoring='neg_mean_squared_error')
cv_r2_scores_combined = cross_val_score(meta_model, X_combined, y,
    cv=kf, scoring='r2')
```

Listing 4.5: Building the Combined Meta-Model (Life Expectancy)

The final results demonstrated the effectiveness of combining Socio-Economic factors with geospatial data to predict life expectancy distribution more accurately.

4.4 The Income Model

Similar to the Life Expectancy model, the Income model can be broken down into two distinct parts: the Socio-Economic regression model (which uses datasets about the Socio-Economic composition of London Boroughs), and the accessibility model (which uses London Underground journey data released by TfL). These two models are combined together for the final results.

4.4.1 Socio-economic model

As with the Life Expectancy model, the process of finding which datasets to include in the model was one of continual trial and error, where models built using individual independent variables were tested to find sets of data that would give stability and predictions that made sense. In the end, the following independent variables were chosen for this Socio-Economic model:

- Percentage of residents in London boroughs with a degree.
- Percentage of residents in London boroughs that spoke English as a first language at home.
- Ethnic makeup of a borough (percentage who are an ethnic minority).
- Percentage of students who achieved an average of level 8 at GCSEs and were accessing free school meals during secondary school.

Again, forward sequential selection was used to select the datasets.

The Socio-Economic data was loaded and preprocessed, converting all relevant columns to numeric format and handling missing values. The dataset was split into training and testing subsets (using an 80-20 split) with a fixed random state to ensure reproducibility. The data was also normalised.

```
# Splits the normalised data
X_ridge_train, X_ridge_test, y_ridge_train, y_ridge_test =
    train_test_split(X_ridge_normalized, y_ridge, test_size=0.2,
    random_state=42)
X_ridge2_train, X_ridge2_test, y_ridge2_train, y_ridge2_test =
    train_test_split(X_ridge2_normalized, y_ridge2, test_size=0.2,
    random_state=42)
# Initialises and trains the first Ridge regression model
ridge_model = Ridge(alpha=1.0)
ridge_model.fit(X_ridge_train, y_ridge_train)
```

Listing 4.6: Income Dataset Training and Testing

Both Ridge and Lasso regression models were built so that the difference could be compared. The training phase involved fitting both models to the training data and performing 5-fold cross-validation. This ensured the models were stable and provided meaningful predictions.

```
# Performs k-fold cross-validation for the first Ridge regression model
kf = KFold(n_splits=5, shuffle=True, random_state=42)
cv_mse_scores_ridge = cross_val_score(ridge_model, X_ridge_normalized, y_ridge, cv=kf, scoring='neg_mean_squared_error')
cv_r2_scores_ridge = cross_val_score(ridge_model, X_ridge_normalized, y_ridge, cv=kf, scoring='r2')
# Fits the first Ridge model and make predictions
ridge_model.fit(X_ridge_train, y_ridge_train)
y_pred_ridge = ridge_model.predict(X_ridge_test)
```

Listing 4.7: Building the Ridge Regression Model (Income)

Like before, performance metrics, including Mean Squared Error (MSE) and R-squared (R^2), were calculated to evaluate the models.

4.4.2 Accessibility model

As previously mentioned, data about London's transport infrastructure has been included. Using a dataset released by Transport for London that shows every journey made between two London Underground stations in 2021, an "accessibility score" for each station has been created that reflects the ease of which a passenger is able to get to central London.

The first step in this was to separate out each journey to allow for easier parsing of the data. Then, a list of central London stations was manually defined. To fit this criteria, a station had to be within the historic City of London / Square Mile (or near enough that it is frequently a commuter destination), or a central hub frequently used by commuters (for example, London Bridge station). These stations can be seen in Listing 4.8.

```
central_london_stations = [
    "Oxford Circus", "Piccadilly Circus", "Leicester Square",
    "Charing Cross", "Waterloo", "Bank", "Liverpool Street",
    "Victoria", "Green Park", "Bond Street", "London Bridge",
    "Kings Cross", "St Pancras", "Canary Wharf", "Heron Quays",
    "West India Quay", "Monument", "Moorgate", "Tower Hill",
    "Mansion House", "Blackfriars", "Barbican", "Shoreditch High Street", "Aldgate", "Farringdon", "Temple", "Chancery Lane", "
```

```
   Holborn", "Covent Garden", "Leicester Square",
    "Charing Cross", "Embankment", "Piccadilly Circus",
    "Goodge Street", "Euston Square", "Euston", "Old Street",
    "Green Park", "Marble Arch", "Hyde Park Corner",
    "Knightsbridge", "Victoria", "St James's Park",
    "Westminster", "Angel"
]
```

Listing 4.8: Defined Central London Stations

This then allowed for an initial accessibility score to be calculated. Working on the assumption that a higher number of journeys from a station reflects that station's accessibility, the total number of journeys between every station, and the list of central London stations, was summed and normalised to give an "accessibility" score.

This first attempt revealed many outliers in the dataset that needed sorting. For example, there were far more journeys made between North Greenwich and central London stations when compared to other similar outer-borough Underground stations. This is likely due to the fact many people travel to North Greenwich to visit the O2 arena to attend concerts and other events. A decision was made to remove these kinds of outliers, as it negatively impacted the accuracy of the model. Other stations removed include Finsbury Park, Stratford, and Brixton. Heathrow Airport terminal stations were also omitted.

After sorting these outliers, the model needed further refining to better reflect station accessibility. To do this, the coordinates of every station were gathered, and the distance (in kilometres) between each station and all the Central London stations were averaged (as seen in Listing 4.9).

```
def calculate_average_distance(lat, lon):
    distances = [geodesic((lat, lon), coords).kilometers for coords
     in central_london_stations.values()]
    return sum(distances) / len(distances)
```

Listing 4.9: Code to calculate the average distance between a station and all Central London stations.

A station's initial (non-normalised) accessibility score was then weighted inversely to this new calculated average (as seen in Listing 4.10). This had the impact of improving the scores of stations close to Central London (that in some cases may have had relatively low ridership numbers), whilst dampening the scores of stations which are further out (but may have high ridership numbers). These scores were finally normalised to between 1 and 100.

```python
def adjust_accessibility_score(row):
    original_score = row['Accessibility_Score']
    avg_distance = row['Average_Distance_to_Central_London']
    adjusted_score = original_score / (1 + avg_distance)  # adjusts by inversely proportional factor
    return adjusted_score
```

Listing 4.10: Code to calculate a new accessibility score weighted on the average distance of a station to Central London.

Finally, this dataset (which includes the borough of each Underground station) is merged with the previously mentioned income dataset. The accessibility data was used to prepare a second Ridge and Lasso regression model, with the borough-level accessibility scores serving as the independent variable and the borough-level average income as the dependent variable. Similar to the Socio-Economic model, the dataset was split into training and testing sets, and the models were trained and evaluated using cross-validation and performance metrics.

```python
# Initialises and trains the second Ridge regression model (transport model)
ridge_model2 = Ridge(alpha=1.0)
ridge_model2.fit(X_ridge2_train, y_ridge2_train)
# Performs k-fold cross-validation for the second Ridge regression model (transport model)
cv_mse_scores_ridge2 = cross_val_score(ridge_model2, X_ridge2_normalized, y_ridge2, cv=kf, scoring='neg_mean_squared_error')
cv_r2_scores_ridge2 = cross_val_score(ridge_model2, X_ridge2_normalized, y_ridge2, cv=kf, scoring='r2')
# Fits the second Ridge model (transport model) and make predictions
y_pred_ridge2 = ridge_model2.predict(X_ridge2_test)
# Evaluates the second Ridge model (transport model)
mse_ridge2 = mean_squared_error(y_ridge2_test, y_pred_ridge2)
r2_ridge2 = r2_score(y_ridge2_test, y_pred_ridge2)
```

Listing 4.11: Building the Accessibility/Transport Ridge Model.

4.4.3 Combination of the two models

The predictions from both the Socio-Economic and accessibility models were combined to form a meta-model. This involved using the predictions from the two individual models as inputs for a final regression model. The combined model was then trained and evaluated using the same approach, ensuring a prediction of income across London boroughs.

```python
# Evaluates the second Ridge model (transport model)
mse_ridge2 = mean_squared_error(y_ridge2_test, y_pred_ridge2)
r2_ridge2 = r2_score(y_ridge2_test, y_pred_ridge2)
# Combines predictions from both models
ridge_predictions = ridge_model.predict(X_ridge_normalized)
ridge2_predictions = ridge_model2.predict(X_ridge2_normalized)
X_combined = np.column_stack((ridge_predictions, ridge2_predictions))
# Trains meta-model
meta_model = Ridge(alpha=1.0)
X_train_combined, X_test_combined, y_train_combined, y_test_combined = train_test_split(X_combined, y_ridge, test_size=0.2, random_state=42)
meta_model.fit(X_train_combined, y_train_combined)
y_pred_meta = meta_model.predict(X_test_combined)
# Evaluates the combined model
mse_meta = mean_squared_error(y_test_combined, y_pred_meta)
r2_meta = r2_score(y_test_combined, y_pred_meta)
```

Listing 4.12: Building the Combined Meta-Model (Income)

The final results showed the effectiveness of combining Socio-Economic factors with accessibility data to predict income distribution more accurately. These results will be discussed in further detail in Chapter 5.

4.5 Clustering

The clustering analysis aims to group hospitals in London based on their geographical coordinates to identify patterns and insights into healthcare accessibility across the city. Various clustering algorithms were used, including K-means, Fuzzy K-means, DBSCAN, and Leader-Follower clustering. The dataset used is the UK hospital locations dataset previously used in the spatial clustering for the Life Expectancy model. Various hyper-parameters for each clustering algorithm were experimented with to improve results. All clustering results were overlaid onto a GeoJSON file of the 32 London Boroughs [16], as well as a map from OpenStreetMaps (using Folium).

4.5.1 K-means Clustering

The hospital dataset was filtered to include only London hospitals, ensuring the relevant columns were numeric and free of missing values. The K-means algorithm was then applied, assigning each hospital to one of five clusters based on their geographical coordinates. The results were visualised using Folium, with markers colour-coded by cluster and borough boundaries for context (this can be seen in Chapter 5).

```
# Prepares data for clustering
X = valid_coords_london[[’Latitude’, ’Longitude’]].values
# K-means clustering
kmeans = KMeans(n_clusters=20, random_state=42)
valid_coords_london[’KMeans_Cluster’] = kmeans.fit_predict(X)
```

Listing 4.13: Generating the K-means results.

4.5.2 Fuzzy K-means Clustering

The Fuzzy K-means algorithm allowed for each hospital to belong to multiple clusters with varying degrees of membership.

```
# Prepares data for clustering
X = valid_coords_london[[’Latitude’, ’Longitude’]].values
# Fuzzy C-means clustering
cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(X.T, c=10, m=2,
    error=0.005, maxiter=1000, init=None)
# Assigns cluster with highest membership value to each point
valid_coords_london[’FuzzyKMeans_Cluster’] = np.argmax(u, axis=0)
```

Listing 4.14: Generating the Fuzzy K-means results.

Two visualisation methods were used to represent these results. In the first method, the colour intensity of each marker reflected the degree of membership to the nearest cluster centre.

In the second method, the colours were blended based on the memberships to different clusters, providing a more nuanced visual representation of the fuzzy memberships. These visualisations helped in understanding the extent to which each hospital belonged to multiple clusters.

```
# Function to blend colours based on membership values
def blend_colors(memberships, colors):
    blended_color = np.zeros(3)
    for i, color in enumerate(colors):
        blended_color += memberships[i] * np.array(mcolors.to_rgb(color))
    return to_hex(blended_color)
```

Listing 4.15: Blending the colours of the Fuzzy K-means results.

4.5.3 DBSCAN Clustering

The DBSCAN algorithm was applied to identify clusters of hospitals in densely populated areas and isolate those in more remote locations. The results were visualised on a map similar to the previous clustering methods.

```
# Prepares data for clustering
X = valid_coords_london[['Latitude', 'Longitude']].values
# DBSCAN clustering with eps-0.05
dbscan = DBSCAN(eps=0.05, min_samples=5)
valid_coords_london['DBSCAN_Cluster'] = dbscan.fit_predict(X)
```

Listing 4.16: Generating the DBSCAN results.

4.5.4 Leader-Follower Clustering

The Leader-Follower clustering algorithm was implemented using the previously mentioned pairwise_distances function. This method provided a flexible approach to clustering, accommodating various distance metrics to suit the specific needs of the analysis.

```
# Leader-Follower clustering with threshold=0.05
threshold_distance = 0.05
leaders = []
followers = []
labels = -1 * np.ones(len(X))
for i, point in enumerate(X):
    if not leaders:
        leaders.append(point)
        labels[i] = len(leaders) - 1
    else:
        distances = pairwise_distances([point], leaders)
        min_distance = np.min(distances)
        if min_distance < threshold_distance:
            labels[i] = np.argmin(distances)
        else:
            leaders.append(point)
            labels[i] = len(leaders) - 1
```

Listing 4.17: Generating the Leader-follower clustering results.

4.5.5 Visualisation and Results

The results of each clustering algorithm were visualised using Folium maps. Each map included markers for the hospitals, colour-coded according to their respective clusters, with borough boundaries added for context. Additionally, the clustering results were exported to a CSV file and loaded into QGIS. The results were then compared to a population density map of London. This comparison aimed to highlight under-served areas and assess the alignment of hospital clusters with population needs. Further discussion on the insights from this comparison will be provided in Chapter 5.

Chapter 5

Results

5.1 Life Expectancy Model Results

For the results of the Life Expectancy (LE) model, the results of the first two smaller models will be presented, and then the meta-model (which is a combination of the two models) will be presented.

5.1.1 Socio-Economic Model

In this model, data on borough ethnicity was combined into one variable (minority). The other two independent variables are child obesity rates and social housing rates. The result of this regression model can be seen on Figure 5.1.

Metric	**Value**
Cross-Validation Mean Squared Error	1.0819
Cross-Validation R^2 Score	0.2459
Ridge Regression - Mean Squared Error	0.4121
Ridge Regression - R^2 Score	0.6668
Ridge Regression - Coefficients	$[-0.0789,$ $-0.000008819,$ $-0.1993]$
Ridge Regression - Intercept	81.7648

Figure 5.1: Ridge Regression Model Metrics

Model interpretation:

- **Ridge Regression Mean Squared Error (MSE):** The MSE for the test set is 0.4121, showing how well the model fits the training data. A lower MSE indicates better model performance, suggesting that the model predictions are close to the actual life expectancy values in the test set.

- **Ridge Regression R^2 Score:** The R^2 score for the test set is 0.667, indicating that approximately 66.7% of the variance in life expectancy can be explained by the model. This is a reasonably good fit, showing that the model captures the relationship between the predictors and life expectancy to a significant extent.

- **Ridge Regression Intercept:** The intercept of 81.765 suggests that when all predictors are zero, the expected life expectancy is approximately 81.765 years. This value serves as the baseline prediction for life expectancy when the predictor values are at their minimum.

- **Cross-Validation Mean Squared Error:** An MSE of 1.0819 translates to a Root MSE (translated back into years) of about 1.04 years, indicating that the model's life expectancy predictions are off by about 1 year on average.

- **Cross-Validation R^2 Score:** The cross-validation R^2 score is 0.246, suggesting that only about 24.6% of the variance in life expectancy is explained by the model during cross-validation. This relatively low value indicates that the model might not generalise well to new data, pointing to potential issues such as overfitting or insufficient feature representation.

In terms of the coefficients, they can be interpreted as:

- **For renting from a local housing authority** a coefficient of -0.0789 indicates living in social housing has a negative impact on life expectancy, however the value is so close to zero that it is very marginal.

- **A value of -0.000008819 for the percentage of ethnic minorities in a borough** is so small that it has zero impact on the results of the model.

- **Obesity** has the biggest impact in this model on life expectancy with a coefficient value of -0.1993, which indicates higher obesity rates are associated with a lower life expectancy.

As can be seen in Figure 5.2, the model predicts life expectancy reasonably well within the mid-range of values. However, the deviations at the extremes (both lower and higher life expectancy) suggest that the model struggles with accurately predicting the most extreme values. One significant limitation visible from this plot is the small number of data points. With only seven samples, it is challenging to draw definitive conclusions about the model's performance. The small sample size can lead to overfitting, where the model performs well on the training data but poorly on unseen data.

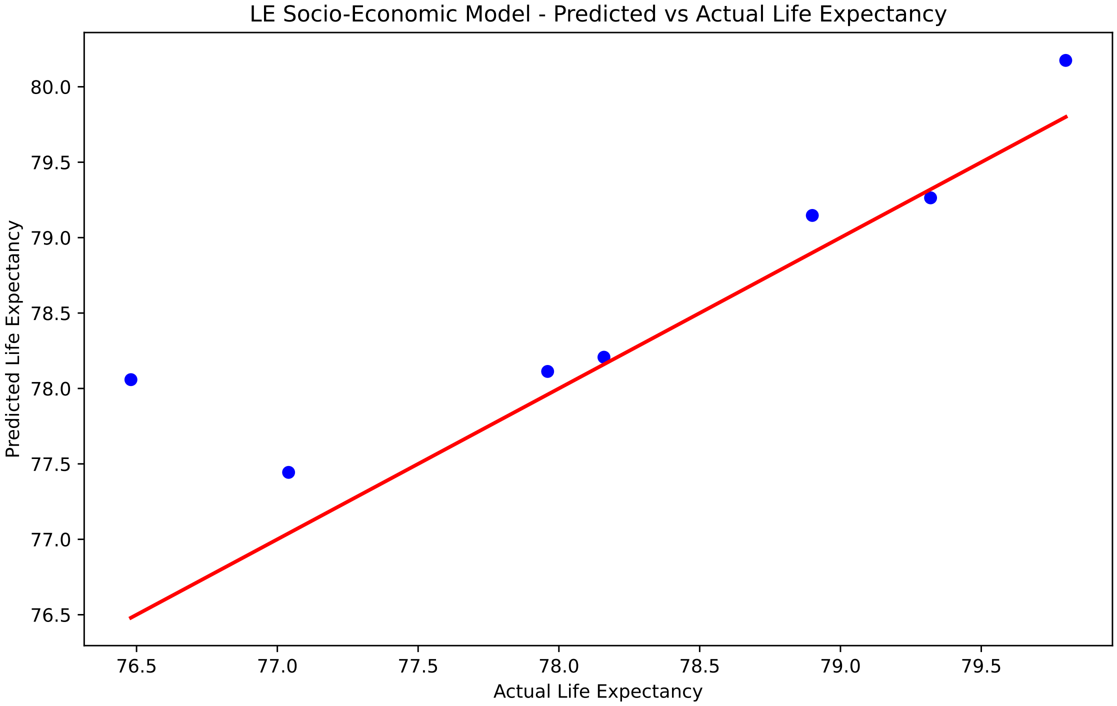

Figure 5.2: Predicted vs Actual Life Expectancy for the Socio-Economic LE Model

Increasing the sample size would likely improve the model's performance and provide a more reliable evaluation of its predictive power. Incorporating additional relevant features might help in capturing more variability in the life expectancy data, potentially leading to better predictions.

As shown in Figure 5.3, the Q-Q plot suggests that while the residuals of the model somewhat follow a normal distribution, there are significant deviations at the tails. This indicates that the model's residuals are not perfectly normally distributed. In particular, the residuals appear to be more dispersed at the lower end, which can affect the model's performance and reliability. The deviations at the tails may indicate the presence of outliers or other non-normal characteristics in the data.

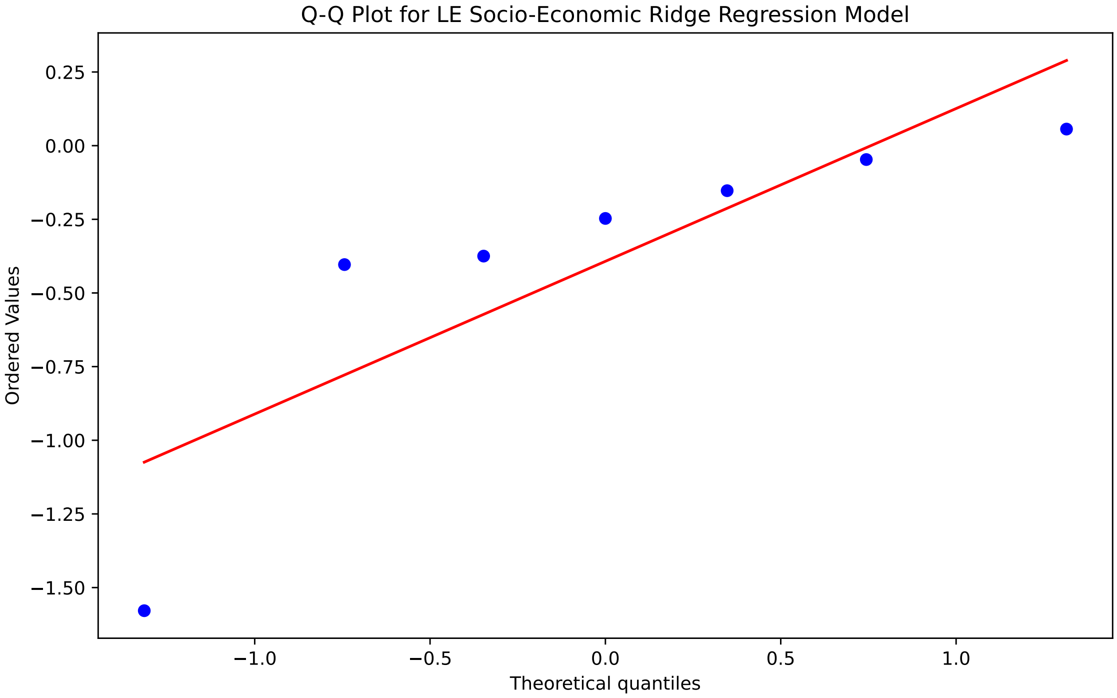

Figure 5.3: Q-Q Plot for the Socio-Economic Life Expectancy Model

The “Renting from Housing Authorities” residual plot (Figure 5.4) suggests that while the residuals are somewhat randomly distributed, there are areas where the residuals appear more spread out, particularly at higher values of the percentage renting from housing authorities. This indication of heteroscedasticity suggests that the model might not be fully capturing the complexities in the relationship between renting from housing authorities and life expectancy. Specifically, the spread of residuals at higher values suggests that the variance of the errors increases with higher percentages of a Borough’s population renting from housing authorities, which the current linear model does not adequately address.

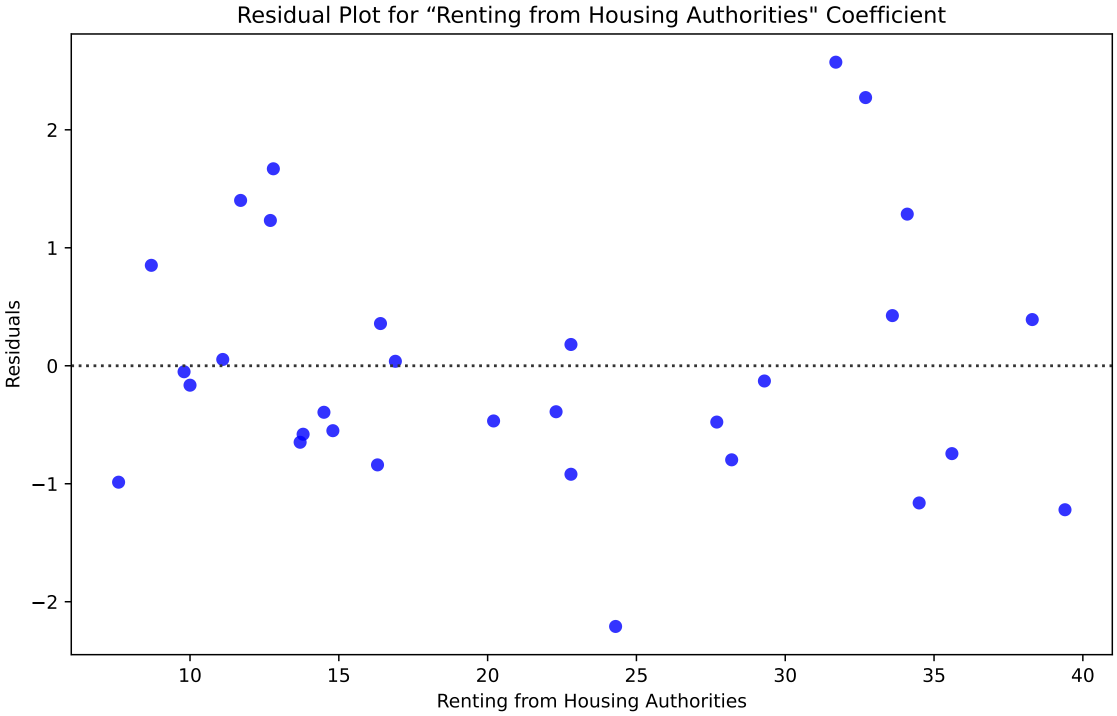

Figure 5.4: Residual Plot for “Renting from Housing Authorities” Coefficient

Again, as Figure 5.5 shows, there is noticeable heteroscedasticity in the residual plot for the childhood obesity rates. The spread of residuals varies, especially at mid-range obesity rates, indicating that the model is not fully capturing the relationship between obesity rates and life expectancy.

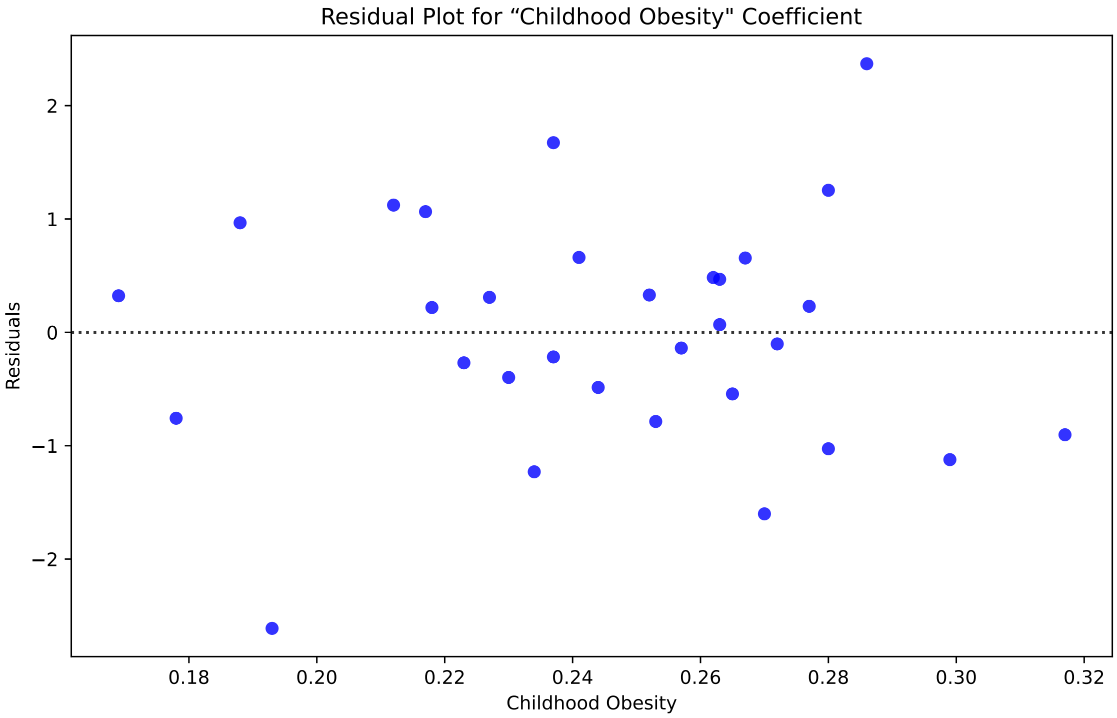

Figure 5.5: Residual Plot for Chilhood Obesity rates

The presence of clusters and variability in the spread of residuals across different values of the independent variable suggests that the model might benefit from including additional variables or exploring non-linear relationships.

5.1.2 Spatial Regression Model

The spatial regression model is used to account for spatial dependencies in the data. In this case, it is applied to the life expectancy data to see how spatial factors, like the number of hospitals and the distance to the nearest hospital, influence life expectancy across different boroughs. The results of this model can be observed in Figure 5.6:

```
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :life_expectancy             Number of Observations:          32
Mean dependent var  :     79.0172                Number of Variables   :           4
S.D. dependent var  :      1.4441                Degrees of Freedom    :          28
Pseudo R-squared    :      0.0932
Spatial Pseudo R-squared:  0.0935
Log likelihood      :    -55.0931
Sigma-square ML     :      1.8321                Akaike info criterion :     118.186
S.E of regression   :      1.3535                Schwarz criterion     :     124.049

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT        78.10600         1.06247        73.51329         0.00000
      hospital_count         0.15388         0.08893         1.73032         0.08357
nearest_hospital_distance    0.00013         0.00036         0.36158         0.71767
   W_life_expectancy        -0.00015         0.00253        -0.06018         0.95201
```

Figure 5.6: Summary of Output for Spatial Lag Model

Model Interpretation:

- **Number of Observations**: The model is based on 32 observations, which corresponds to the boroughs in London.

- **Mean Dependent Variable (Life Expectancy)**: The average life expectancy across these boroughs is 79.0172 years.

- **Pseudo R-squared**: This is a measure of the model's explanatory power. A value of 0.0932 indicates that the model explains about 9.32% of the variance in life expectancy. This is relatively low, suggesting that the spatial factors included in the model don't account for much of the variability in life expectancy.

- **Spatial Pseudo R-squared**: Similarly, this is the R-squared value considering spatial dependencies, which is also 0.0935.

- **Log Likelihood**: The log-likelihood value of -55.0931 is a measure of model fit, with higher values indicating better fit. This will be compared to the meta-model.

In terms of the coefficients, they can be interpreted as:

- **The constant term**: (78.10600) is the base life expectancy when all other variables are zero. This is statistically significant with a very low p-value (¡0.00000).

- **hospital_count**: The coefficient (0.15388) represents the change in life expectancy for each additional hospital. This positive coefficient suggests that having more hospitals is associated with higher life expectancy. The p-value (0.08357) is marginally significant, indicating a weak but not conclusive relationship.

- **nearest_hospital_distance**: The coefficient (0.00013) indicates the effect of the distance to the nearest hospital on life expectancy. This very small positive value suggests a negligible impact. The high p-value (0.71767) indicates this is not statistically significant.

- **W_life_expectancy**: This spatial lag term (coefficient of -0.00015) represents the influence of neighbouring boroughs' life expectancy on a borough's life expectancy. The very small negative coefficient and high p-value (0.95201) suggest that there is no significant spatial dependency captured by this model.

The spatial regression model's low pseudo R-squared values indicate that the included spatial factors (hospital count and distance to nearest hospital) unfortunately explain only a small portion of the variance in life expectancy across boroughs.

5.1.3 Combined Model

The combined model integrates both the Socio-Economic and spatial regression models to try and predict life expectancy more accurately. Both Ridge regression and Lasso regression models were built. The results of **the Ridge regression model** can be seen in Figure 5.7:

Metric	**Value**
Cross-Validation Mean Squared Error (Combined)	0.8850
Cross-Validation R^2 Score (Combined)	0.3768
Combined Model - Mean Squared Error	0.2662
Combined Model - R^2 Score	0.7848
Combined Model - Coefficients	[0.9657, -0.2478]
Combined Model - Intercept	22.2521

Figure 5.7: Combined Life Expectancy Model Metrics (Ridge)

Model interpretation:

- **Mean Squared Error (MSE)**: The MSE of the combined model is 0.2662. This indicates that, on average, the squared difference between the actual and predicted life expectancy values is around 0.2662 years. This relatively low MSE suggests that the combined model provides a good fit to the data.

- **R-squared (R^2) Score**: The R^2 score of 0.7848 signifies that approximately 78.48% of the variability in life expectancy is explained by the combined model. This is a significant improvement over the individual models, indicating that the integration of Socio-Economic and spatial factors enhances the predictive power of the model.

- **Cross-Validation Results**: The cross-validation MSE of 0.8850 and R^2 score of 0.3768 indicate the model's performance on unseen data. The lower cross-validation R^2 score compared to the training R^2 score suggests that the model may be overfitting to some extent. However, it still captures a reasonable amount of variance in the life expectancy data when cross-validated.

In terms of the coefficients, they can be interpreted as:

- **The coefficient for the Socio-Economic model prediction** is 0.9657, indicating a strong positive impact on life expectancy.

- **The coefficient for the spatial model prediction** is -0.2478, suggesting a small negative impact on life expectancy. However, the strength of this coefficient is less than the positive Socio-Economic coefficient.

- **The intercept** of 22.2521 is the baseline life expectancy when all predictors are zero, which in the context of this combined model, serves as an adjustment factor for the predictions.

The negative coefficient for the spatial model prediction in the combined model could be explained by the interaction dynamics between the Socio-Economic and spatial models. While the individual spatial model shows a positive effect of hospital count on life expectancy, the combined model adjusts for overlaps or redundancies between the two sets of variables. Since Socio-Economic factors already capture a part of the variability in life expectancy, the spatial model's contribution is adjusted negatively to fine-tune the overall prediction. This does not imply that better spatial attributes reduce life expectancy; rather, it reflects the complex interplay between the two models. The results highlight the complexity of modelling life expectancy, showing the need for further refinement (and larger, more granular datasets) to improve model quality and interpretability.

As shown in Figure 5.8, the Ridge meta-model predicts life expectancy reasonably well, with a few outliers at the tails of the line. Much like the individual Socio-Economic model, this indicates there could be an issue with the underlying data, or that more data could be required to make a more accurate assessment.

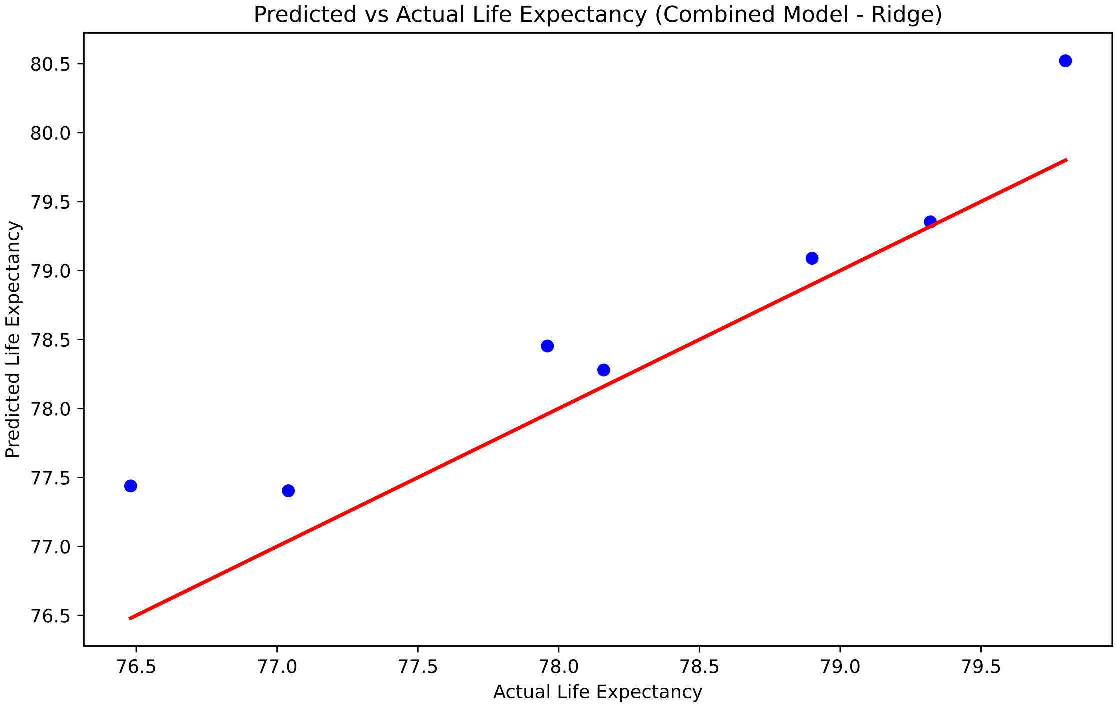

Figure 5.8: Predicted vs Actual Life Expectancy (Combined Model - Ridge)

The Q-Q plot, as shown in Figure 5.9, suggests that the model follows a relatively normal distribution, however it is a large improvement over the Q-Q plot for the initial Socio-Economic Model seen in Figure 5.3. There still remain outliers at both ends of the plot however, which implies some non-normal characteristics of the model.

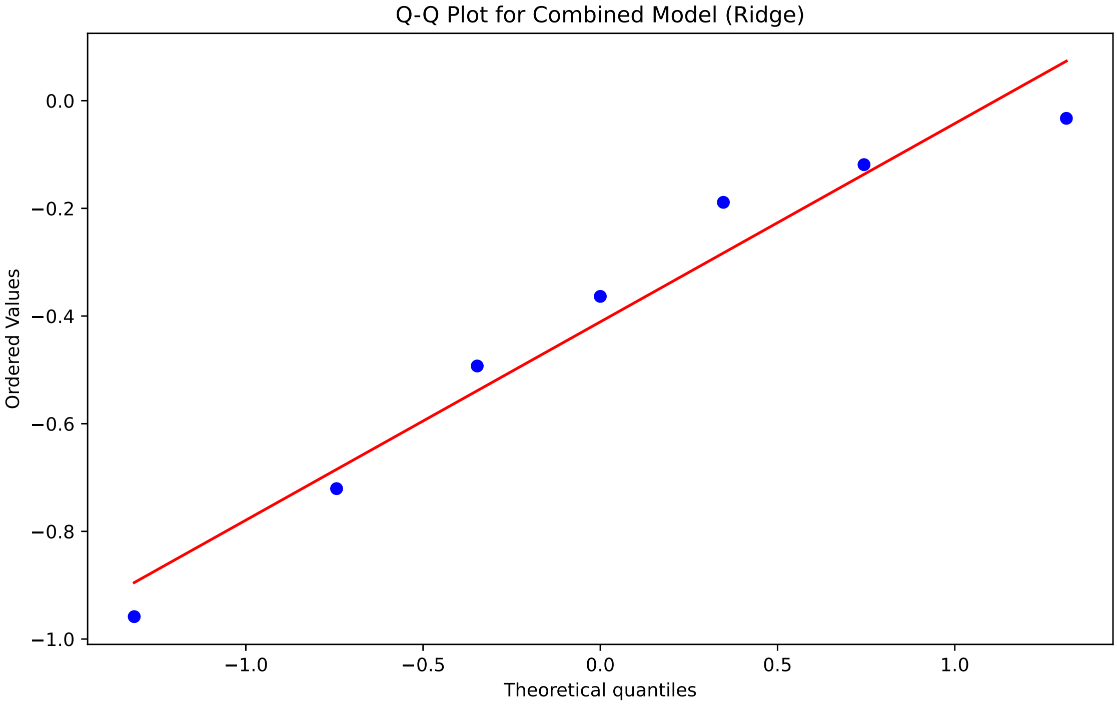

Figure 5.9: Q-Q Plot for Combined Model (Ridge)

Finally, the map of predicted Life Expectancy by Borough for the Ridge regression model can be seen in Figure 5.10. Generally for North East London, it predicts a lower average life expectancy when compared to the rest of London. The notable outlier here is its prediction for Newham however. Here, this model predicts an average life expectancy of around 80.5 years. This is significantly higher than the 76.48 years in the dataset. A discrepancy this large indicates that there certainly is an issue with the dataset causing inaccuracies with the model.

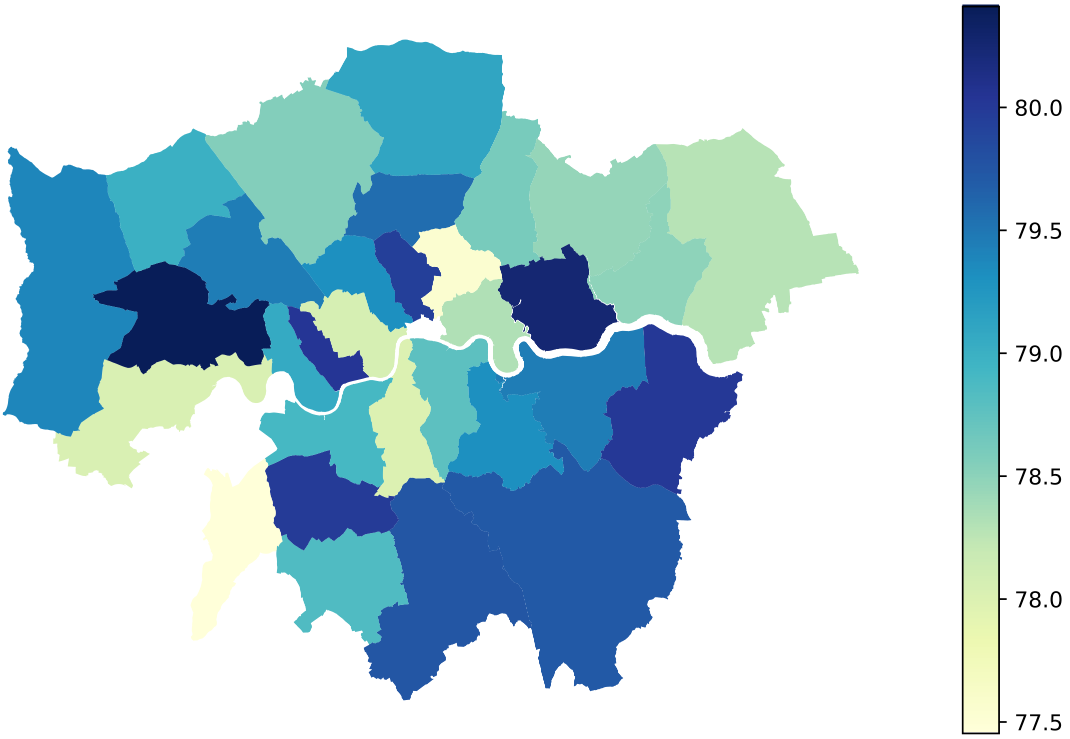

Figure 5.10: Predicted Life Expectancy by Borough (Ridge)

The results of **the Lasso regression model** can be seen in Figure n:

Metric	**Value**
Cross-Validation Mean Squared Error (Combined)	0.9450
Cross-Validation R^2 Score (Combined)	0.3323
Combined Model - Mean Squared Error	0.4805
Combined Model - R^2 Score	0.6115
Combined Model - Coefficients	[0.8733, -0.0000]
Combined Model - Intercept	10.0255

Figure 5.11: Lasso Combined Life Expectancy Model Metrics

The Lasso regression results indicate that the second coefficient (associated with the spatial regression model) has been shrunk to zero. This implies that Lasso has effectively excluded the spatial model variable, suggesting that it does not add significant predictive power to the model. As shown in Figure 5.12 and in Figure 5.13 however, this has led to less accurate predictive results when compared to the Ridge model (seen in Figure 5.8).

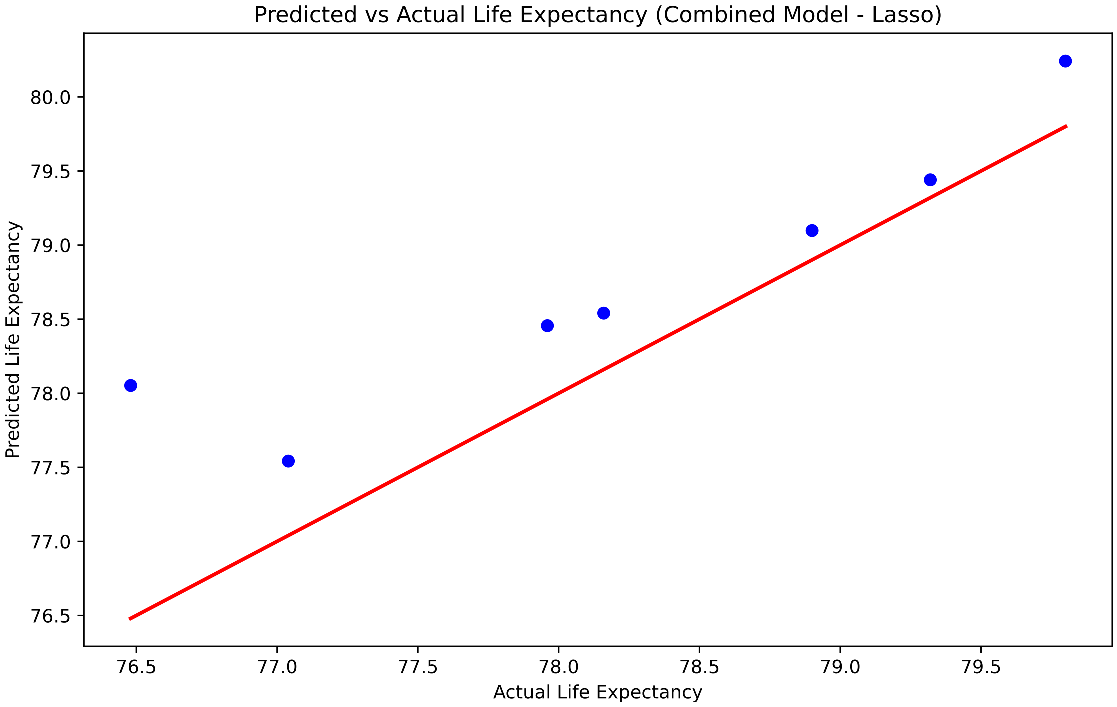

Figure 5.12: Predicted vs Actual Life Expectancy (Combined Model - Lasso)

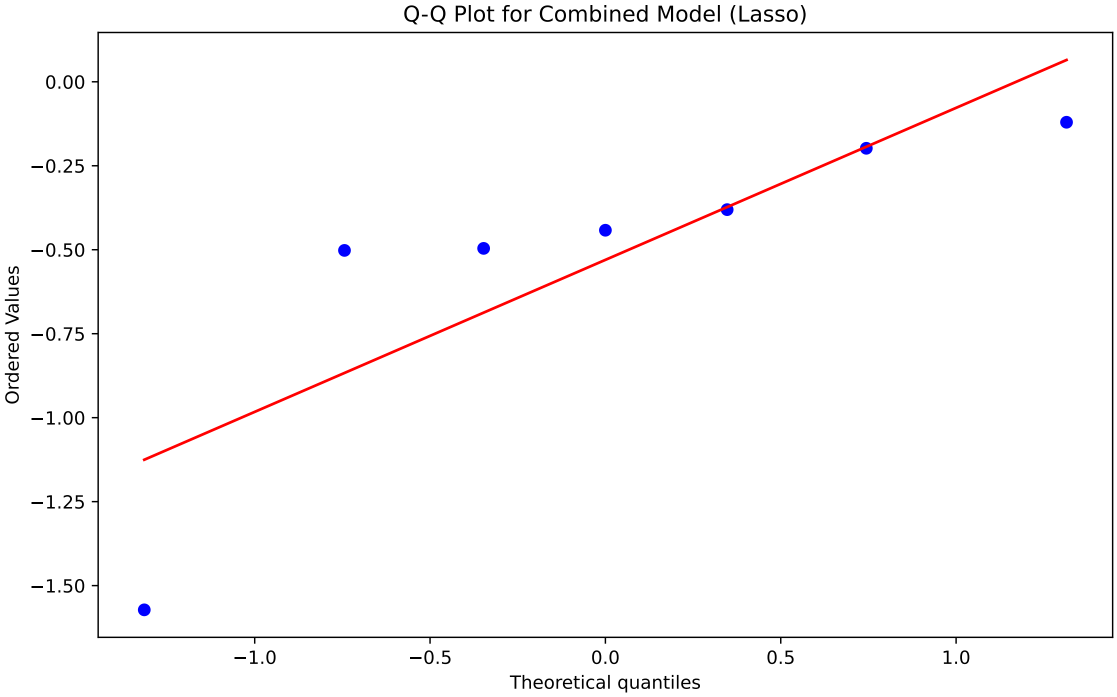

Figure 5.13: Q-Q Plot for Combined Model (Lasso)

The map of predicted Life Expectancy by Borough for the Lasso regression model can be seen in Figure 5.14. Much like before with the Ridge model, Newham stands out with a very high life expectancy. Again, this highlights the importance of dataset accuracy, and is an area where improvement could be made.

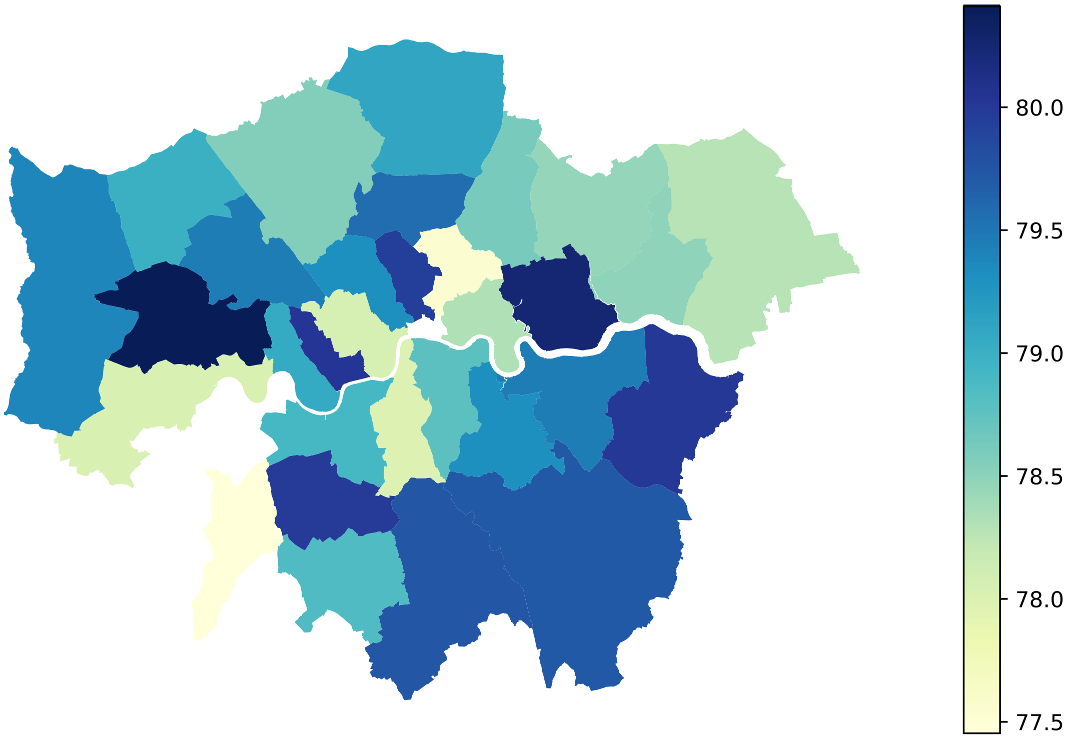

Figure 5.14: Predicted Life Expectancy by Borough (Lasso)

The comparison between the Lasso and Ridge regression models also highlights the impact of different regularisation techniques on the predictive models. In the Lasso results, the second coefficient (for the spatial regression model) is zero, indicating that Lasso has excluded this variable as it does not add predictive power. This simplification suggests that Lasso is effective in eliminating less significant predictors, focusing on the most impactful variables.

Conversely, the Ridge regression retains both coefficients, with the spatial model having a smaller, non-zero impact. This implies that Ridge regression maintains all features (suggesting some value in the spatial data), even if their contribution is minimal. Ridge regression's approach helps in stabilising predictions by not completely discarding any predictors, which can be useful when all variables have some degree of relevance.

Overall, Ridge regression maintains a more complex model by retaining all variables (and gives a more accrate prediction), whilst Lasso simplifies the model by excluding less significant predictors. This difference highlights the importance of choosing the appropriate regularisation technique based on the data's characteristics and the modelling objectives.

5.2 Income Model

For the results of the Income model, the results of the first two smaller models will be presented, followed by the results of the meta-model.

5.2.1 Socio-Economic Model

When building the Socio-Economic Model, both non-normalised and normalised data were experimented with. The results of the non-normalised Socio-Economic Model can be seen in Figure 5.15:

Metric	**Value**
Cross-Validation Mean Squared Error (Ridge)	347323001.6146
Cross-Validation R^2 Score (Ridge)	0.5149
Ridge Regression - Mean Squared Error	279429217.0469
Ridge Regression - R^2 Score	0.4580
Ridge Regression - Coefficients	[227.2322, -0.2415, 2316.1488, 64835.8129]
Ridge Regression - Intercept	-54748.5169

Figure 5.15: Ridge Regression Model Metrics (Socio-Economic) - Non-Normalised

The results of the normalised Socio-Economic model can be seen in Figure 5.16:

Metric	**Value**
Cross-Validation Mean Squared Error (Ridge)	315887897.7507
Cross-Validation R^2 Score (Ridge)	0.5630
Ridge Regression - Mean Squared Error	239500007.2105
Ridge Regression - R^2 Score	0.5354
Ridge Regression - Coefficients	[-1112.7315, -10270.3448, 4891.0477, 13981.8203]
Ridge Regression - Intercept	61676.7544

Figure 5.16: Ridge Regression Model Metrics (Socio-Economic) - Normalised

The results of the Socio-Economic Model (both non-normalised and normalised) indicate a moderate ability to explain the variance in income within the dataset. The cross-validation R^2 score of 0.5630 and the test set R^2 score of 0.5354 (for the normalised results) suggest the model captures around 56.30% and 53.54% of the variance in income, respectively. Despite these reasonable R^2 scores, the Mean Squared Error (MSE) values are quite high (315,887,897.75 for cross-validation and 239,500,007.21 for the test set), indicating substantial average squared differences between predicted and actual income values. For the rest of this section the normalised results will be referred to.

The coefficients offer further insights:

- **English Spoken at Home (-1112.73)**: This negative coefficient suggests that a 1% increase in residents who speak English at home corresponds to a £1112.73 decrease in average income (holding other variables constant).

- **Minority (-10270.34)**: This significant negative coefficient indicates that a 1% increase in minority populations correlates with a £10270.34 decrease in average income (holding other variables constant), highlighting socioeconomic challenges faced by minority communities.

- **GCSE FSM (4891.05)**: For each 1% increase in the proportion of students who achieved high GCSE scores while on free school meals, the average income in the borough increases by approximately £4891.05, holding all other variables constant. This suggests that educational attainment among disadvantaged students is a strong predictor of higher income.

- **Percentage with Degree (13981.82)**: For each 1% increase in the proportion of residents with a degree, the average income in the borough increases by approximately £13981.82, holding all other variables constant. This large positive coefficient shows the substantial impact of higher education on income levels.

The prediction plot (Figure 5.17) shows significant deviations between predicted and actual values (particularly at higher income levels), indicating issues with model accuracy. The high MSE values further support this, suggesting a need for model refinement and more granular data.

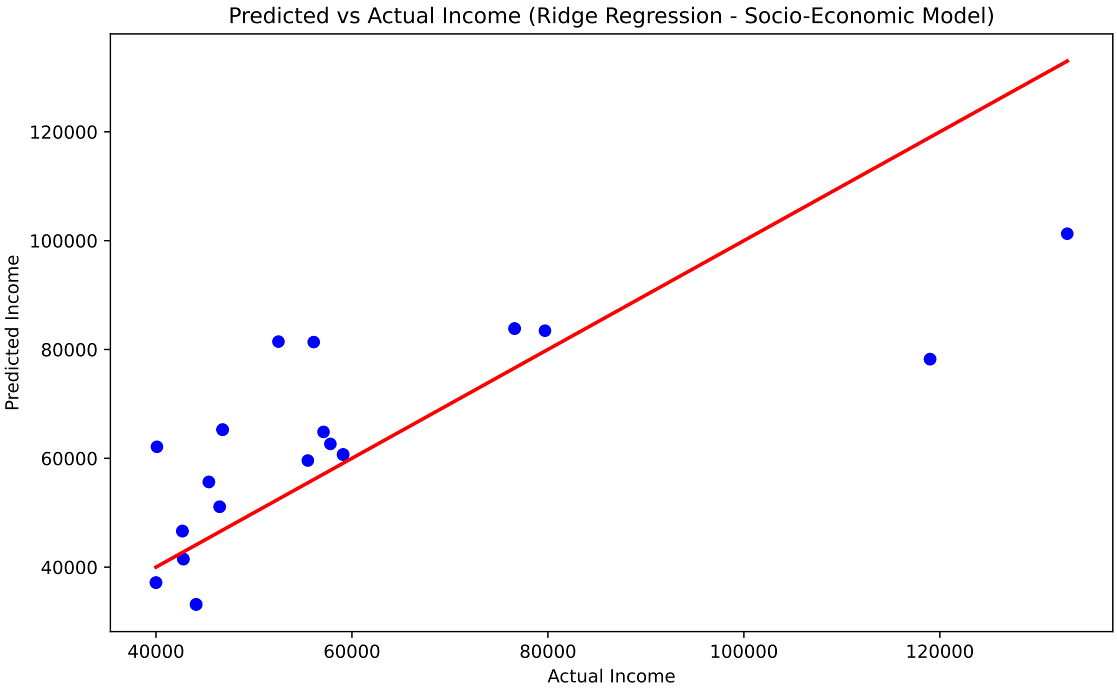

Figure 5.17: Predicted vs Actual Income (Ridge Regression - Socio-Economic Model

The results of the Socio-Economic Model highlight the complex interplay between language proficiency, minority status, and income in London boroughs. The model indicates that higher proportions of residents speaking English at home correlate with lower average incomes, which could be explained by the poor educational performance of white working-class pupils in the UK (who are most likely to speak English at home). This aligns with broader evidence showing that white working-class children tend to perform less well educationally when compared to other demographics, impacting their economic prospects.[32]

Conversely, the significant negative coefficient for minority status highlights systemic socioeconomic disparities faced by immigrant and minority communities in London. Despite broader evidence suggesting that immigrants often achieve better educational outcomes and integrate well in Anglophone countries[32], the model indicates substantial income disparities for minority populations. This highlights the need for targeted interventions to address these disparities and support minority communities.

The model's substantial positive coefficients for educational attainment variables (GCSE FSM and Percentage with Degree) reinforce the importance of education in predicting higher income. Improving access to education for minority groups could mitigate income disparities and enhance economic prospects. However, the predicted vs actual plot shows significant deviations, particularly at higher income levels, suggesting issues with model accuracy and again highlighting the need for more granular data and refined models to capture the complexities of socioeconomic factors influencing income.

Overall, whilst this Socio-Economic model provides some insights, the contradictions and limitations in the data highlight the need for further investigation and targeted policies to address systemic disparities and improve educational and economic outcomes for all demographic groups. This is particularly relevant for North East London, which has a high proportion of minority residents. The region's demographic profile shows the importance of addressing educational and socioeconomic disparities to improve income levels and overall quality of life in these boroughs.

5.2.2 Accessibility Model

When it came to the accessibility model, several iterations were considered to try and improve the accessibility score results. As discussed in Chapter 4, the initial accessibility results only considered the total journeys between a station and the list of central London stations. A sample of these results can be seen in Figure 5.18:

Metric	**Value**
Minimum Accessibility Score	1.0 (Station: Roding Valley)
Maximum Accessibility Score	100.0 (Station: Stratford)
Average Accessibility Score	16.27

Origin	**Accessibility Score**	**Normalised Accessibility Score**
Acton Town	364660.0	14.582749
Aldgate East	623051.0	24.629150
Alperton	91249.0	3.952361
Amersham	123804.0	5.218119
Archway	722703.0	28.503681

Figure 5.18: Accessibility Scores for Selected Stations

As previously discussed in Chapter 4, there were several outliers that were removed in an attempt to get more useful results. These were Stratford, Brixton, Finsbury Park and North Greenwich. The results (the top 5 % of stations) after removing these outliers can be seen in Figure 5.19:

Origin	**Accessibility Score**	**Normalised Accessibility Score**
Bethnal Green	1261121.0	71.950571
Camden Town	1486866.0	84.852467
Canada Water	1264317.0	72.133230
Canning Town	1710837.0	97.652975
Earls Court	1285685.0	73.354465
Elephant & Castle	1261279.0	71.959601
Highbury & Islington	1175660.0	67.066259
Mile End	1231281.0	70.245139
Seven Sisters	1662534.0	94.892336
Tottenham Court Road	1529898.0	87.311854
Vauxhall	1751903.0	100.000000
Wembley Park	1336925.0	76.282960

Figure 5.19: Stations in the Top 5% of Accessibility Scores

To more accurately represent the proximity of stations to central London, the accessibility scores were adjusted by inversely weighting them based on the average distance of each station to the central London stations. The results of this can be seen in Figure 5.20:

Borough	**Station**	**Average Distance to Central London**	**Normalised Adjusted Accessibility Score**
Lambeth	Vauxhall	3.619477	100.000000
Camden	Camden Town	3.687142	83.730441
Southwark	Elephant & Castle	3.073537	81.738129
Islington	Highbury & Islington	3.890811	63.573099
Westminster	Warren Street	2.447946	58.288607

Figure 5.20: Average Distance to Central London and Normalised Weighted Accessibility Scores for Selected Stations

This final weighting of stations can be seen on the heatmap in Figure 5.21, and in the individual stations in Figure 5.22. As expected, stations nearest to the central London stations had the highest accessibility score, whilst stations further out (including stations in North East London) had a lower accessibility score.

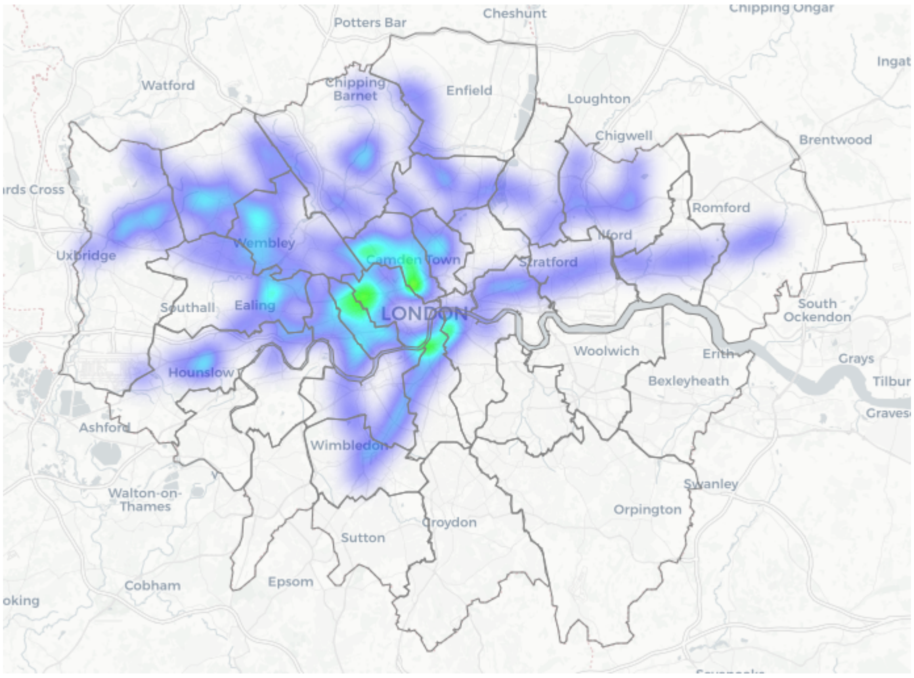

Figure 5.21: Heatmap of London Underground station accessibility

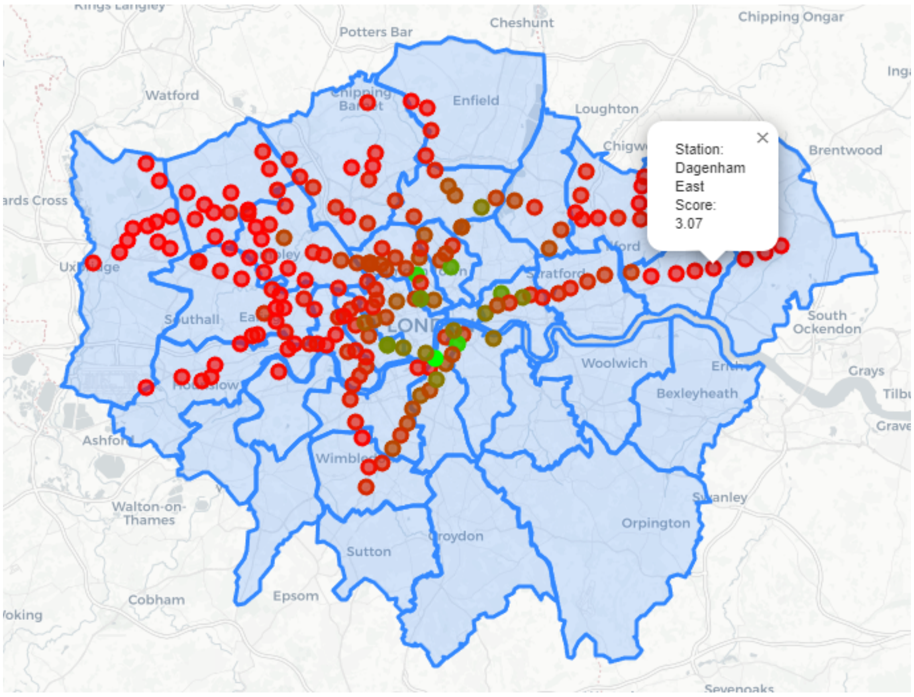

Figure 5.22: Individual London Underground stations based on their accessibility score. Dagenham East station is highlighted.

These station scores were then used as the basis of the Accessibility Model. The results of this model can be seen in in Figure 5.23:

Metric	**Value**
Cross-Validation Mean Squared Error (Ridge2)	624557455.0365
Cross-Validation R^2 Score (Ridge2)	0.1502
Ridge Regression - Mean Squared Error (Ridge2)	375408532.3828
Ridge Regression - R^2 Score (Ridge2)	0.2718
Ridge Regression - Coefficients (Ridge2)	[10851.5337]
Ridge Regression - Intercept (Ridge2)	62344.3116

Figure 5.23: Ridge Regression Model Metrics (Transport/Accessability)

The results of this Ridge Regression Model unfortunately indicate a limited ability to explain income variance within the dataset. The cross-validation R^2 score of 0.1502 and the test set R^2 score of 0.2718 suggest the model captures some variance (around 15.02% and 27.18%, respectively), but not a substantial amount. The high Mean Squared Error (MSE) values (624,557,455.04 for cross-validation and 375,408,532.38 for the test set) indicate significant average squared differences between predicted and actual income values. The coefficient of 10851.53 for the Normalised Adjusted Accessibility Score suggests that increased accessibility correlates with higher average income, holding other variables constant. However, the high MSE values and low R^2 scores highlight significant issues with model accuracy, suggesting the need for more refined models and better data to better capture the relationship between transport accessibility and income. One significant way to improve the model would be to find a dataset that includes London Overground and DLR journeys, as a lot of North East London is serviced by these networks. This data would need to be provided by TfL however.

The prediction plot (Figure 5.24) also shows that the Accessibility / Transport model led to quite inaccurate predictions for income. To further improve the model, an attempt was made to account for the influence of having proximity to a major interchange station (such as London Bridge or Paddington), however this unfortunately turned out to be beyond the scope of this project.

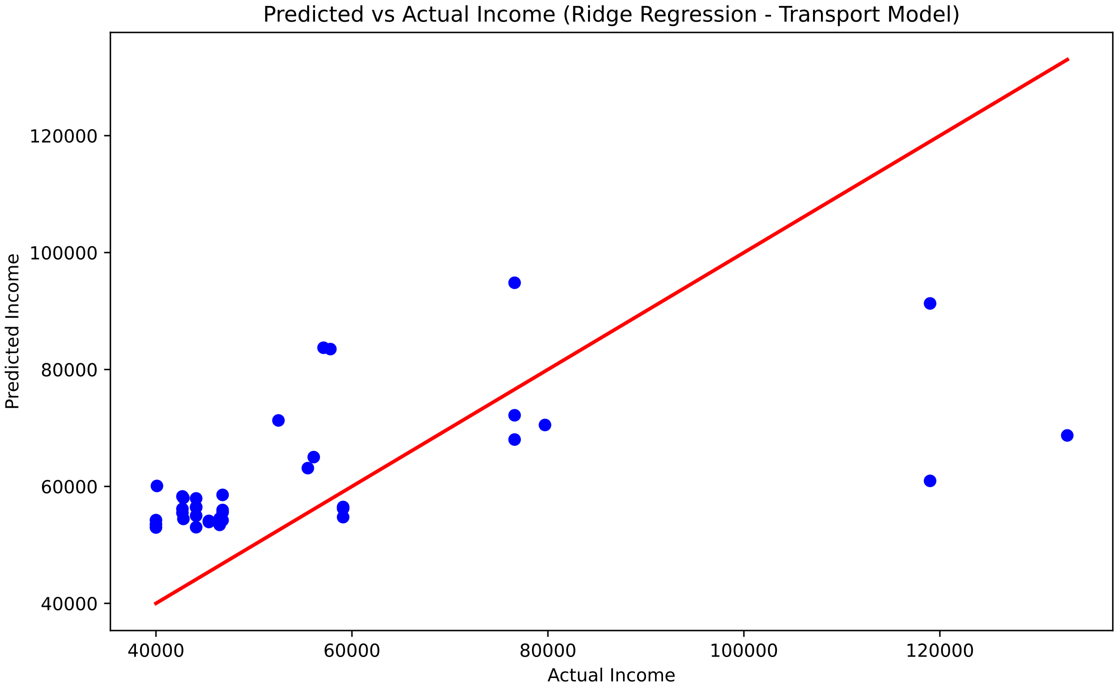

Figure 5.24: Predicted vs Actual Income (Ridge Regression - Transport Model)

5.2.3 Combined Model

Finally, the results of the two models were combined. For comparison, versions of the models were built using Ridge regression and Lasso regression. The results of the Ridge model can be seen in Figure 5.25, whilst the results of the Lasso model can be seen in Figure 5.26.

Metric	**Value**
Combined Model - Mean Squared Error	232777365.1731
Combined Model - R^2 Score	0.5485
Combined Model - Coefficients	[0.9486, 0.2991]
Combined Model - Intercept	-15483.5137

Figure 5.25: Combined Model Metrics (Ridge)

Metric	**Value**
Combined Model - Mean Squared Error	232439650.2098
Combined Model - R^2 Score	0.5491
Combined Model - Coefficients	[0.9439, 0.2983]
Combined Model - Intercept	-15140.2702

Figure 5.26: Combined Model Metrics (Lasso)

The difference between the two models is incredibly minimal, with both models providing a MSE of 232 million, and a R2 score of 0.54. This indicates that these models explain roughly 54% of Income in London boroughs. Unlike with the combined Life Expectancy model, the Lasso regression in this model has not reduced any of the coefficients towards zero. For the rest of this section, the Ridge model will be referred to.

The predicted vs actual plot (Figure 5.27) for the combined model indicates that combining the two models did not significantly improve predictions. Whilst there is an upward trend (as expected), the plot reveals several clusters of points towards the bottom and outliers, highlighting issues with the model's accuracy. Specifically, the clustering of points around lower income levels and the scattered outliers (particularly at higher income levels) suggest the model struggles to capture the full range of income variability accurately. These deviations imply potential flaws in the data or the need for further refinement and additional relevant features to enhance predictive performance.

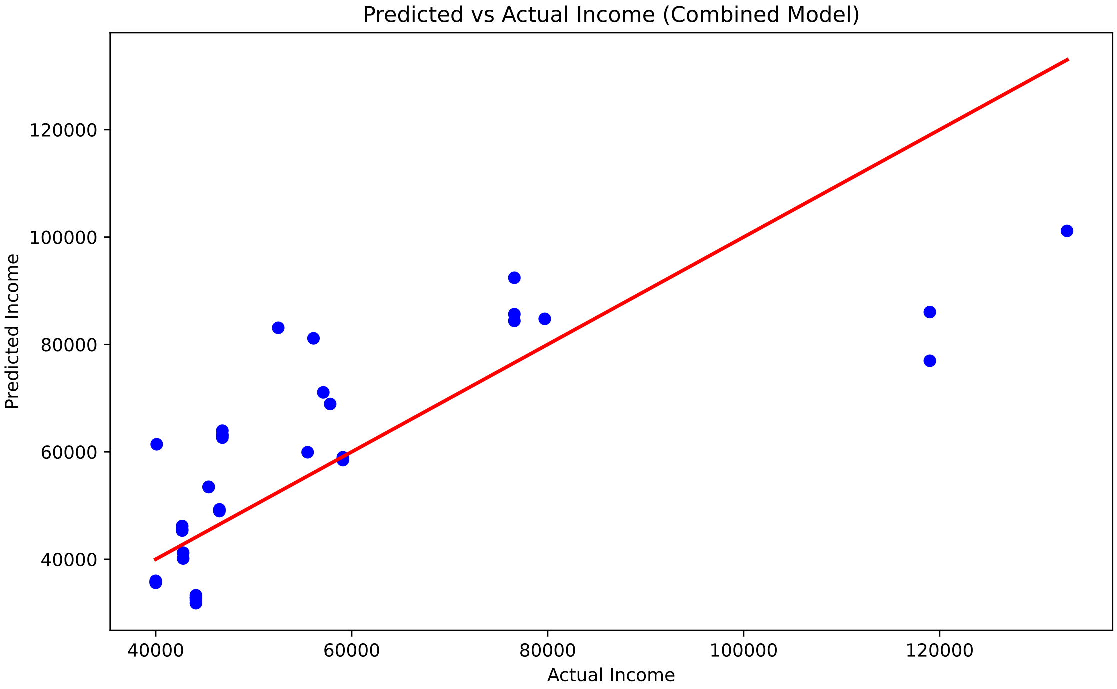

Figure 5.27: Predicted vs Actual Income - Combined Model (Ridge)

This issue is apparent in the residuals plot (Figure 5.28). The plot shows a clear pattern rather than a random distribution, indicating that the model is not capturing all underlying relationships in the data. The residuals exhibit a pronounced curve, particularly at higher income levels, where the residuals increase sharply. This suggests heteroscedasticity and non-linearity, where the model's errors are not evenly distributed across all income levels. This pattern highlights the need for further model refinement and the inclusion of more relevant variables to better capture the complexities of income distribution in London boroughs.

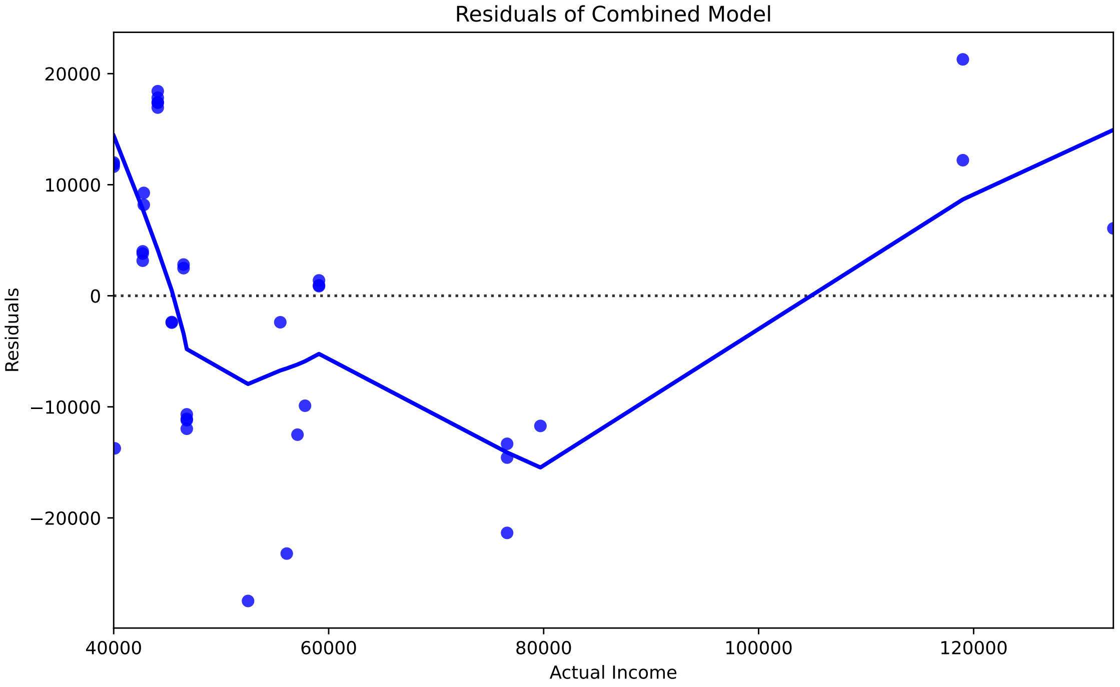

Figure 5.28: Residual Plot for the Combined Income Model (Ridge)

The Q-Q plot for the combined model (Figure 5.29) further supports these observations. The plot shows deviations from the reference line, particularly at the tails, indicating that the residuals do not follow a normal distribution. This non-normality is marked by clear patterns and outliers, suggesting the model's residuals are not purely random and that there may be underlying structural issues or omitted variables affecting performance. These deviations highlight the need for further investigation and potential model adjustments to better capture the underlying data distribution.

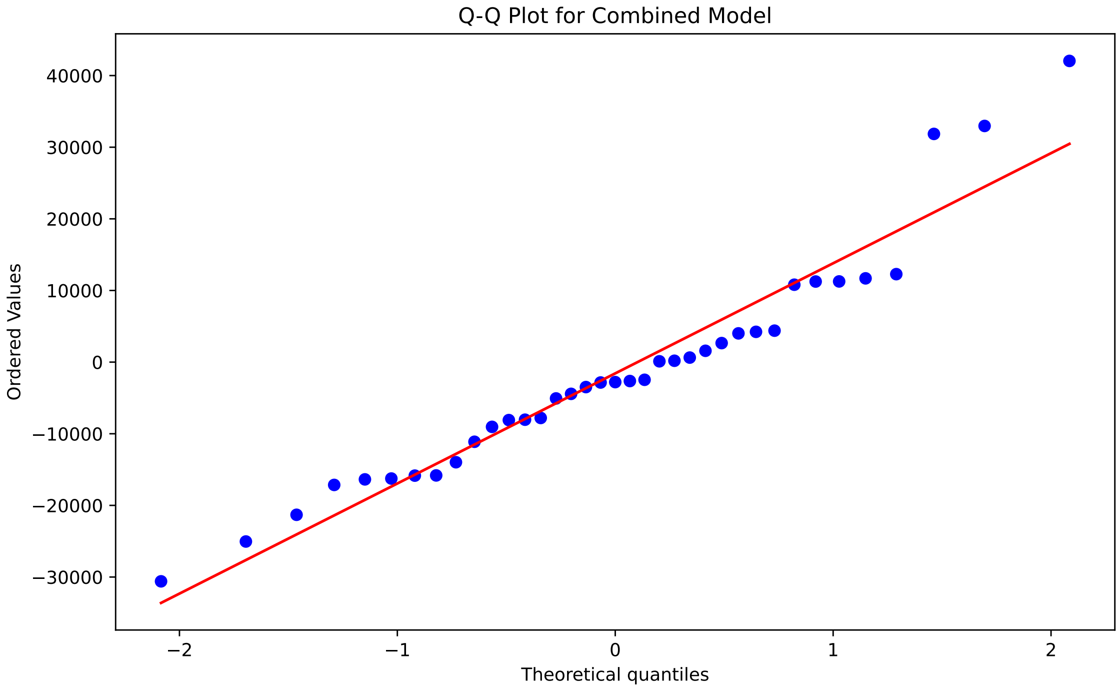

Figure 5.29: Q-Q Plot for the Combined Income Model (Ridge)

5.3 Clustering results

As mentioned, the dataset used in clustering was the UK Hospital dataset, which was refined to only include hospitals with a Greater London Postcode. The locations of every hospital can be seen in Figure 5.30:

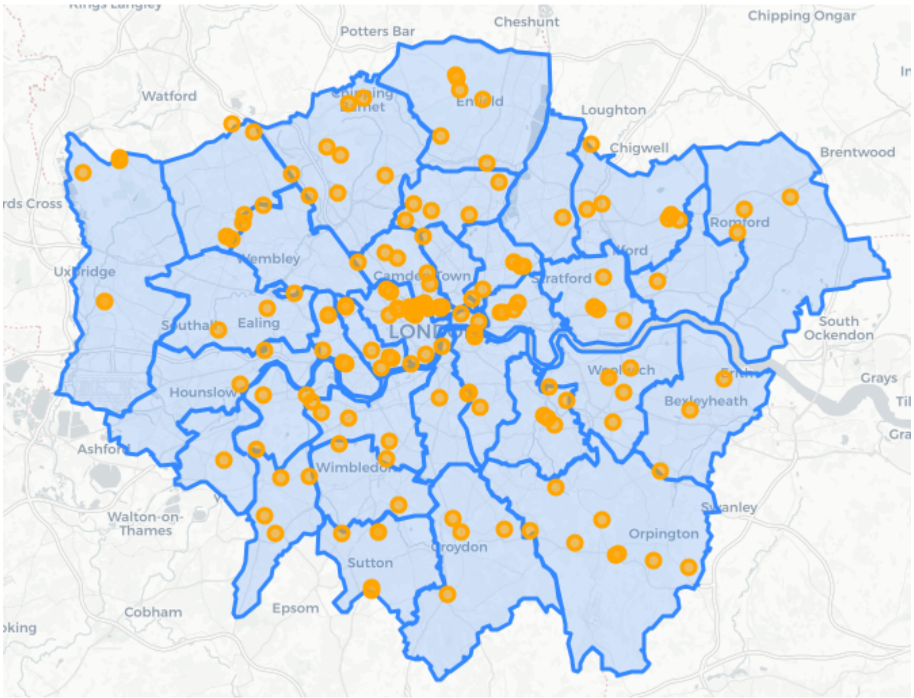

Figure 5.30: UK Hospitals Dataset refined to show all London Hospitals

5.3.1 K-means Clustering

The results of using K-means clustering on the hospital dataset can be seen in Figure 5.31 (k=5) and Figure 5.32 (k=10). It should be noted that for all results, the colours for each cluster region are randomly assigned. Depending on the initial starting parameters (namely the number of cluster centres (k) defined), the level of detail for different regions can vary. However, a consistent observation across different outputs is that the K-means algorithm often groups hospitals in North East London with those in South East London (specifically Woolwich and Bexleyheath), despite being separated by the River Thames. This clustering does not accurately represent the real-life geographical layout of London or the typical movement patterns of residents in North East London. Due to poor accessibility and limited crossing points, it is unlikely that people living in North East London would travel across the river to access healthcare services in South East London.

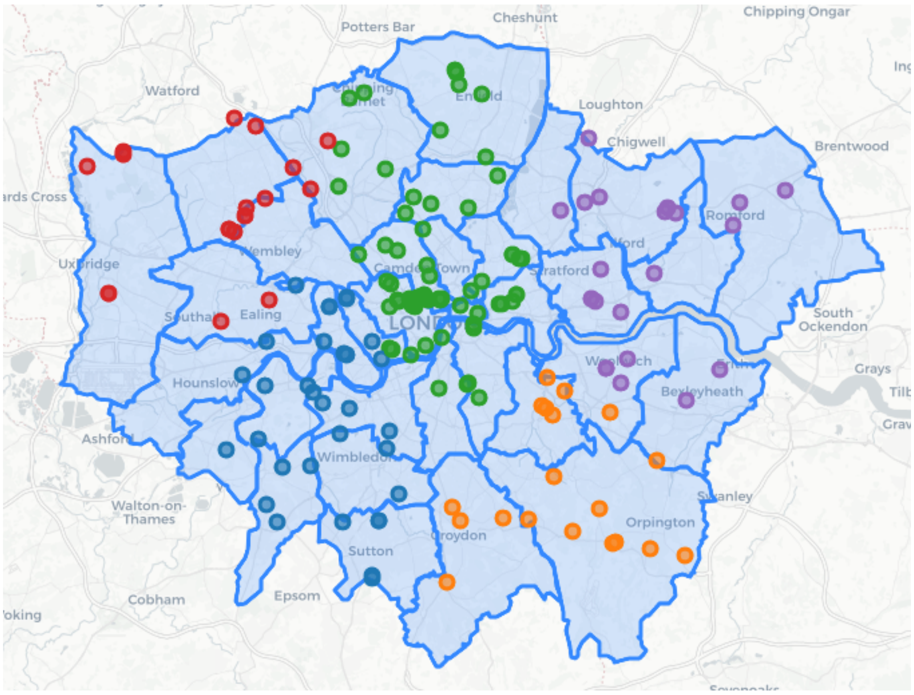

Figure 5.31: K-means clustering results with initial parameter set to k=5

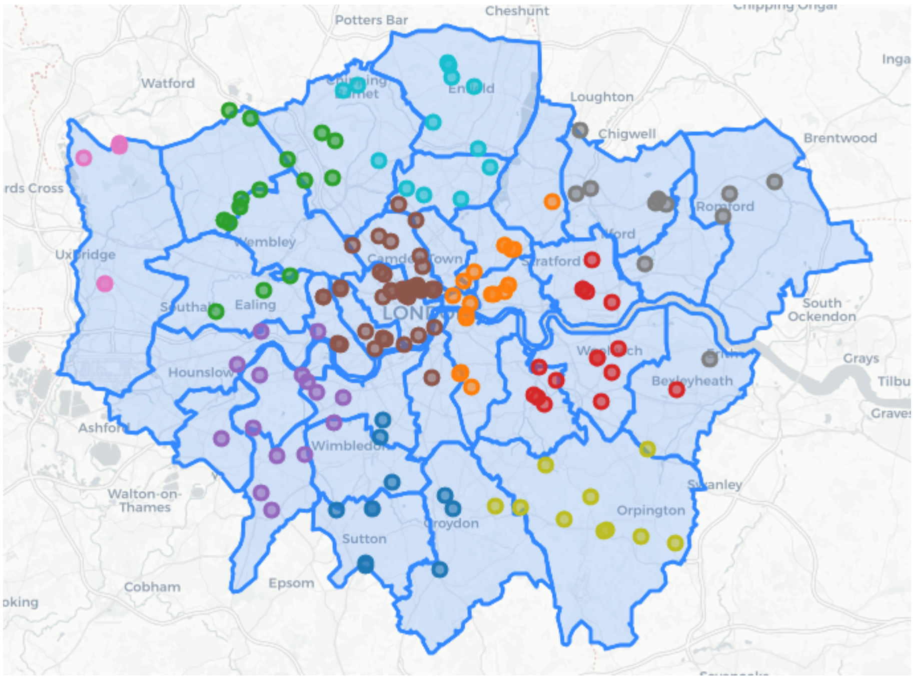

Figure 5.32: K-means clustering results with initial parameter set to k=10

K-means clustering, while efficient and straightforward, has notable limitations. The algorithm's reliance on initial cluster centres can lead to different final clusters depending on the starting conditions. K-means also assumes clusters to be spherical and evenly sized, which may not align with the actual distribution of hospitals.

The tendency of K-means to create clusters based purely on proximity, without accounting for geographical barriers (like the River Thames), can result in impractical groupings. This highlights a significant limitation of the algorithm in urban studies where natural and man-made barriers significantly impact accessibility and movement patterns. Despite these limitations, K-means provides a useful starting point for analysing the distribution of healthcare facilities, helping to identify areas with varying densities of hospitals, which can be further refined with additional geographical or demographic data.

5.3.2 Fuzzy K-means Clustering

The results of applying Fuzzy K-means clustering to the hospital dataset are displayed in Figure 5.33 and Figure 5.34. Similar to traditional K-means, Fuzzy K-means does not account for geographical boundaries such as the River Thames. However, this method provides additional insights into how people in North East London might access healthcare. Figure 5.33 highlights the proximity of a hospital to its cluster center, whilst Figure 5.34 (which blends colours between clusters) illustrates potential overlaps in healthcare service areas, suggesting that residents of North East London might opt for treatment at hospitals in neighbouring boroughs that are geographically closer.

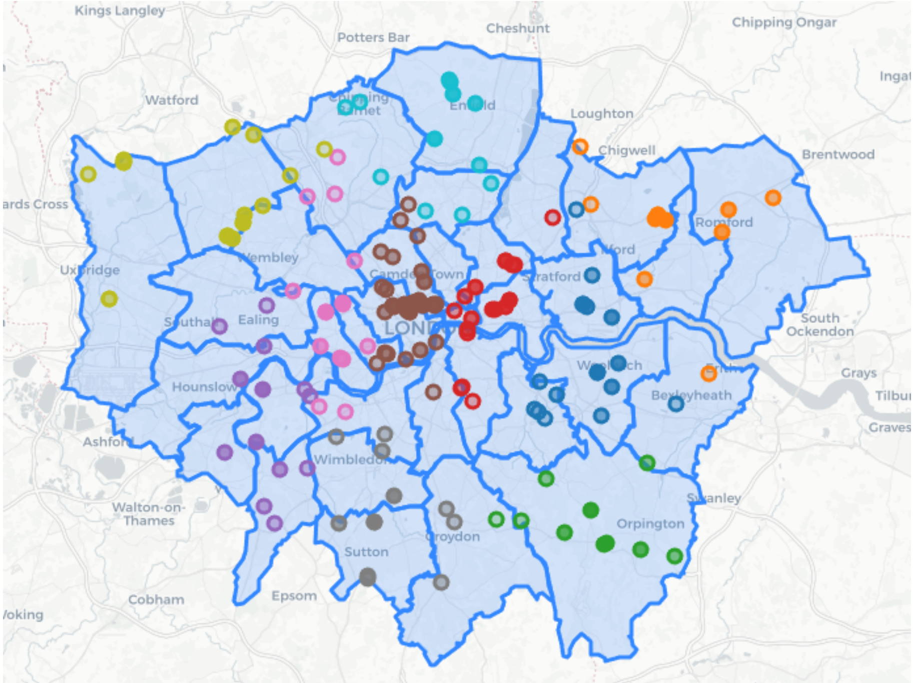

Figure 5.33: Fuzzy K-means clustering results with initial parameter set to k=5

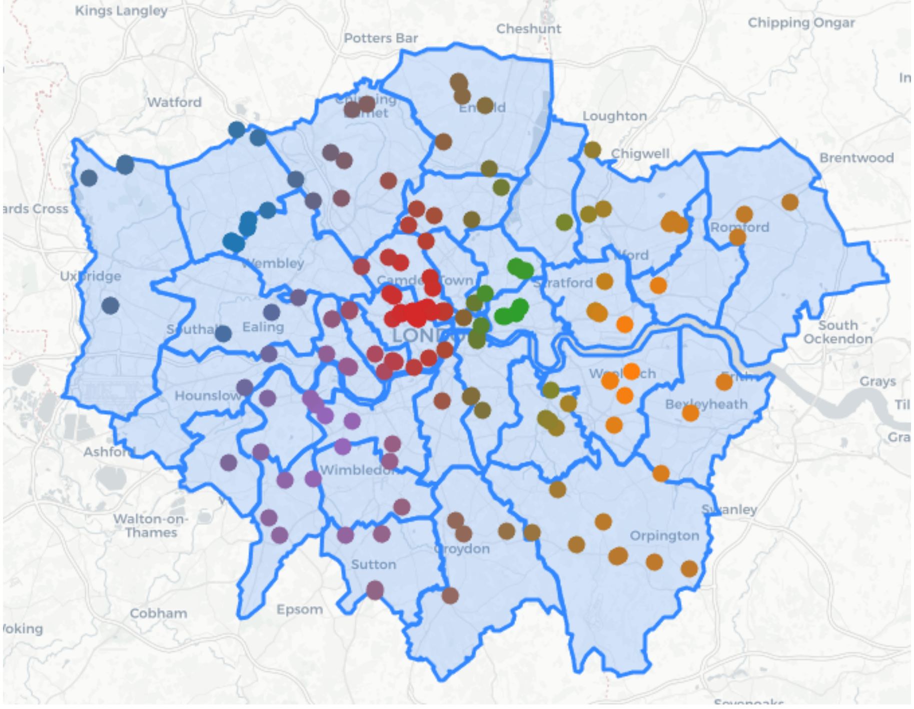

Figure 5.34: Fuzzy K-means clustering results with blended clusters (k=5)

Fuzzy K-means clustering assigns membership levels to each data point, indicating the degree to which each hospital belongs to different clusters. This approach highlights the flexibility in hospital choice for residents, reflecting real-world scenarios where proximity can outweigh administrative boundaries. The blended colours in the second figure effectively convey this overlapping membership, showing that some hospitals serve multiple communities. For example, residents of Tower Hamlets might prefer hospitals in Hackney or Islington due to their closer proximity, despite administrative boundaries suggesting otherwise.

5.3.3 DBSCAN Clustering

The results of applying DBSCAN to the hospital dataset are displayed in Figure 5.35. DBSCAN is known for its ability to identify clusters based on the density of data points and to handle noise effectively. However, this algorithm requires a certain level of density to function optimally.

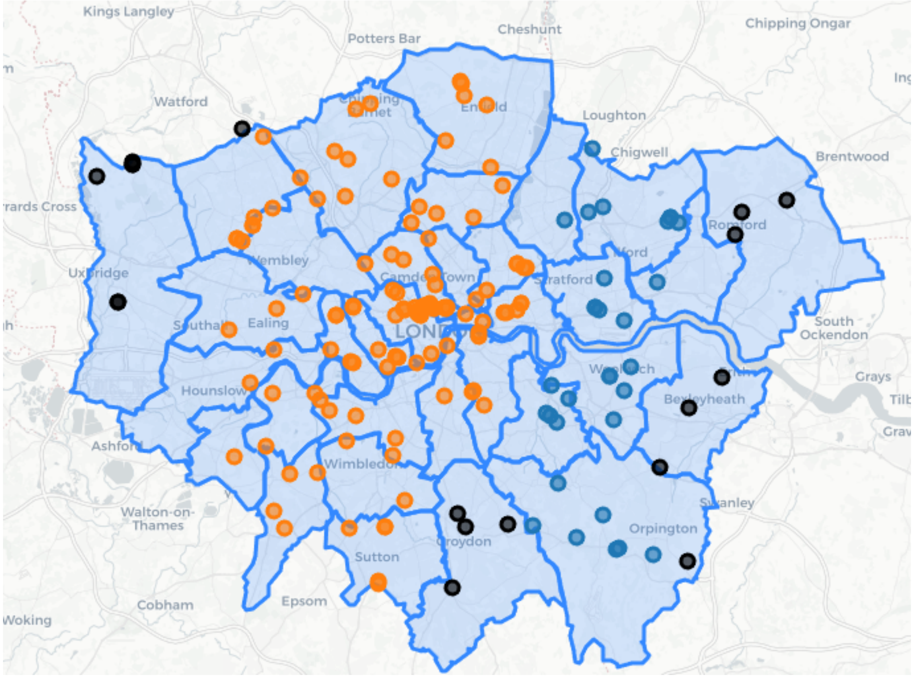

Figure 5.35: DBSCAN clustering results

In this case, the hospital dataset appears too sparse for DBSCAN to identify meaningful clusters. The resulting clusters are scattered and do not provide substantial insights into the distribution of hospitals across London. Adjusting the parameters, such as epsilon (the maximum distance between two points to be considered as in the same neighbourhood) and min_samples (the minimum number of points to form a dense region), did not significantly improve the clustering results.

The poor performance of DBSCAN in this scenario highlights the importance of data density for this algorithm. The sparsity of the dataset means that DBSCAN cannot effectively differentiate between densely packed regions and noise, leading to suboptimal clustering outcomes. This suggests that DBSCAN may not be the most suitable clustering method for this particular dataset, especially when compared to methods like K-means or Fuzzy K-means, which have shown more informative results.

5.3.4 Leader-Follower Clustering

The Leader-Follower clustering algorithm produced impressive results, as demonstrated in Figure 5.36. This algorithm managed to accurately group various regions together, accounting for geographic factors effectively. An intriguing outcome of this method was the formation of 33 different clusters, which is only one more than the total number of boroughs in London, suggesting a high level of precision.

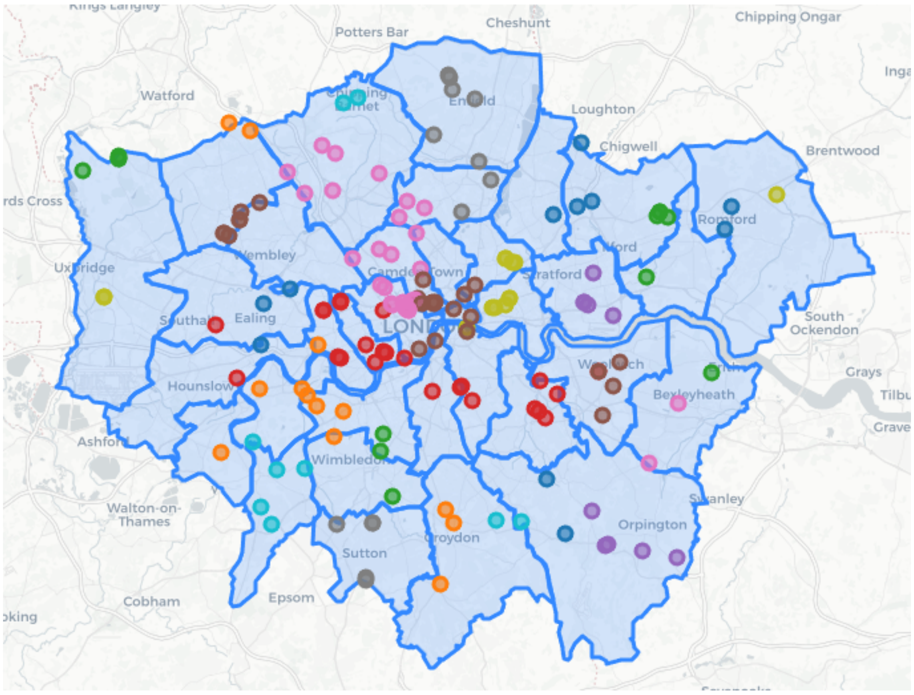

Figure 5.36: Leader-Follower clustering results

Additionally, the algorithm displayed a nuanced understanding of the area's geography. For instance, it adeptly clustered neighbouring boroughs which are separated by the River Thames, reflecting realistic movement and accessibility patterns. There was minimal overlap in the clustering, only visible in the central areas - namely the City of London / Tower Hamlets (see the zoomed-in view in Figure 5.37). However, this overlap is minor and does not significantly detract from the overall accuracy of the clusters.

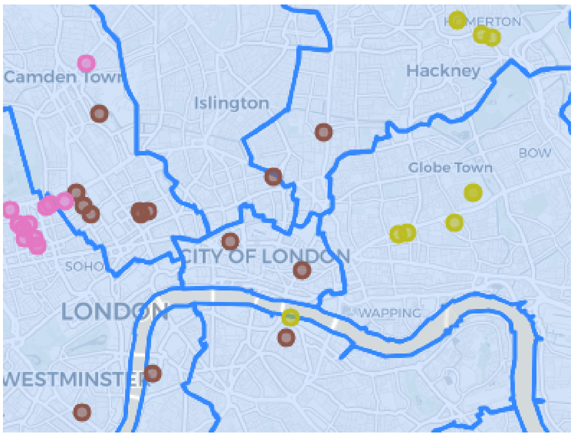

Figure 5.37: Leader-Follower results overlap

Overall, the Leader-Follower algorithm stands out when compared with the rest of the clustering algorithms for its ability to produce meaningful and geographically coherent clusters, aligning closely with the actual boroughs of London.

5.3.5 Comparison with population density

The most revealing result of this work is how it highlights the way in which the boroughs of North East London are under-served by current public health infrastructure. As shown in Figure 5.38, the number of hospitals in these boroughs is noticeably lower than in other areas of London (namely West London), particularly in Waltham Forest and Barking and Dagenham (where there is only one hospital). This scarcity of healthcare facilities is concerning as it can lead to longer travel times for residents seeking medical attention, increased pressure on existing services, and poorer health outcomes overall.

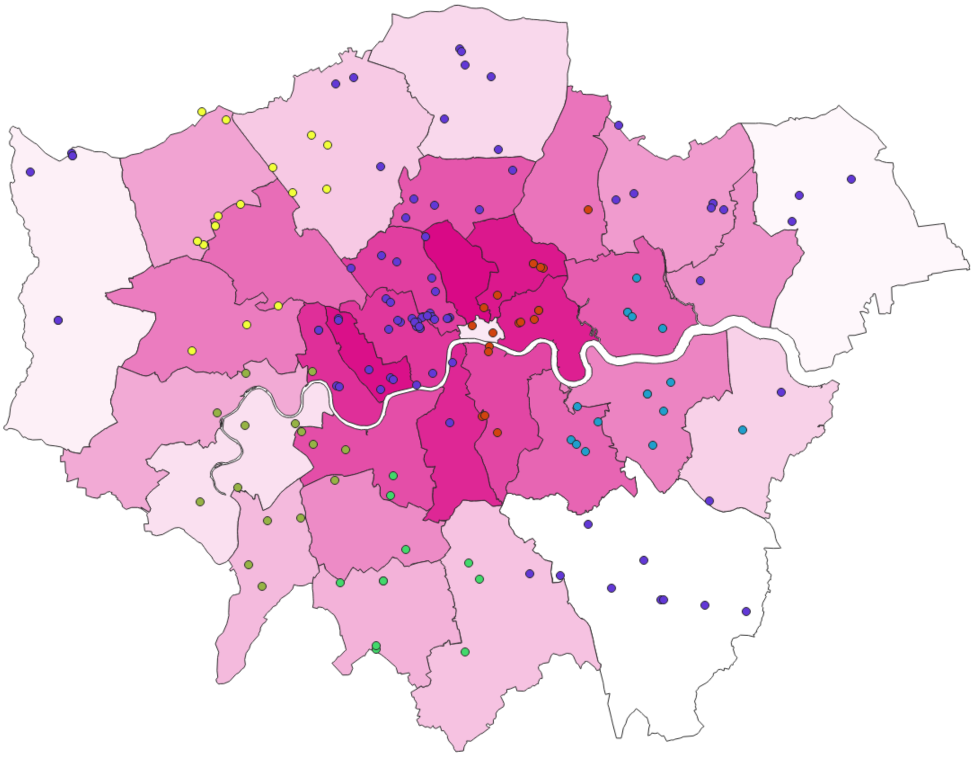

Figure 5.38: Map of London Borough average population density overlaid with K-means clustering results (darker pink indicates higher density)

To improve these outcomes, there should be a concerted effort to invest in additional healthcare infrastructure in these underserved areas. This could include building new hospitals or expanding existing ones, as well as enhancing public transport links to ensure easier access to healthcare facilities.

Chapter 6

Legal, Social, Ethical and Professional Issues

The models and analyses developed in this research could potentially inform policy and urban planning strategies within North East London (and possibly beyond). To ensure compliance with relevant ethical standards and professional guidelines, several key considerations must be addressed.

6.1 Data Privacy and Security

The datasets used in this project include sensitive information related to socio-economic factors, health outcomes, and urban infrastructure. It is important to handle this data responsibly, ensuring compliance with data protection regulations such as the General Data Protection Regulation (GDPR). Measures should be taken to anonymise the data where possible and secure it against unauthorised access.

6.2 Ethical Use of Data

Using data to inform policy decisions has ethical implications. The results of this research could impact various socio-economic groups, potentially influencing resource allocation and public health interventions. Therefore, it is essential to ensure that the analysis and resulting policies do not disadvantage any group. Ethical considerations include avoiding biases in data interpretation and ensuring transparency in the decision-making process.

6.3 Professional Conduct

Adhering to the British Computer Society's Code of Conduct [33], this research prioritises the integrity and reliability of the methods and models used. All findings are verified and validated through testing and cross-validation procedures.

6.4 Transparency and Accountability

Transparency is important in this research. All data sources are clearly documented, and credit is given to the original authors and creators of the datasets and technologies used. The results and insights derived from this research should be shared with stakeholders, including local authorities and the public, in an accessible and understandable format.

6.5 Impact on Public Policy

The insights gained from this research aim to influence public policy and urban planning decisions. It is essential to ensure that these decisions are based on comprehensive analysis. Therefore, any policy recommendations derived from this research should be reviewed and supported by additional studies and expert opinions to ensure their validity and effectiveness.

6.6 Social Implications

This research aims to address health and income inequalities in North East London. The goal is to identify under-served areas and provide actionable insights to improve public health and economic outcomes. It is important to engage with community stakeholders and consider their perspectives in the analysis and recommendations.

By addressing these legal, social, ethical, and professional issues, this research aims to contribute responsibly and positively to the field of urban planning and public policy, ensuring that the findings are used ethically and effectively to benefit the community.

Chapter 7

Conclusion

7.1 Summary of Key Findings

The study aimed to predict life expectancy and income in North East London using a variety of regression and clustering techniques. The Life Expectancy model revealed that obesity rates had the most significant negative impact on life expectancy, highlighting the need for targeted health interventions. The Income model identified education, particularly the percentage of residents with degrees, as a significant factor influencing income levels. These findings suggest a strong link between educational attainment and economic outcomes, highlighting the importance of investment in education in the region.

For the clustering analysis, the Leader-follower algorithm demonstrated superior performance compared to other methods like K-means, fuzzy K-means, and DBSCAN. This algorithm very effectively grouped healthcare facilities, providing a clearer picture of under-served areas. This result builds on Bochra Hadj Kilani's (2023)[11] work, suggesting improvements in clustering techniques for urban data analysis. However, the limitations of the dataset, which only included major hospitals and excluded smaller health facilities like GPs must be considered when drawing conclusions in how to improve the region. The results do however suggest that North East London is likely under-served by current healthcare infrastructure.

7.2 Limitations

The primary issue encountered during the project was the limited availability and granularity of data. The small sample sizes, particularly for the life expectancy model, led to overfitting and flawed predictions. Including more diverse and comprehensive datasets could enhance model accuracy and allowed them to better generalise.

Additionally, the exclusion of smaller healthcare facilities from the clustering analysis limited the understanding of healthcare accessibility. Incorporating data on GPs and other local health services would provide a more accurate representation of healthcare infrastructure distribution in the region.

The issues of this project may have been avoided if perhaps there was a clearer focus when it came to the scope. Perhaps if more time had been dedicated to say just one model, instead of both Life Expectancy and Income, more insightful results could have been gathered. Likewise when it comes to the clustering results; if perhaps a variety of different geospatial datasets could have been experimented with, more insightful results for how to improve infrastructure in North East London could have been gathered.

7.3 Future Work

Future research should focus on obtaining more comprehensive datasets, including a wider range of socio-economic and health-related variables. This could improve the quality and accuracy of the predictive models. Exploring other advanced regression techniques, such as spatial error models and geographically weighted regression, could also provide deeper insights into spatial dependencies. One notable limitation of the study was the lack of consideration for temporal data. Incorporating data collected over different time periods could potentially enhance the accuracy and quality of the models by capturing trends and changes over time. Future research should aim to include such temporal data to better understand the dynamics of socio-economic and health variables.

With the Transport/Accessibility Model specifically, the previously mentioned work on accounting for proximity to large interchange stations could also be expanded on to provide a more accurate model of accessibility to central London.

For clustering, integrating additional data sources, such as transport accessibility and real-time service availability, could also refine the analysis of healthcare accessibility. Further development of hybrid clustering algorithms that combine the strengths of different methods might also yield better results. To further test the strength of the Leader-follower clustering algorithm (and to see if the results of this paper are accurate), similar spatial datasets should be experimented with.

7.4 Final thoughts

This research has highlighted significant socio-economic and health inequalities in NE London. While the models developed offer valuable insights, the limitations in data quality and availability have constrained their predictive power. Nonetheless, the findings emphasise the critical role of education in economic outcomes and the pressing need for improved healthcare infrastructure in the region. By addressing these challenges in future work, more effective and targeted interventions can be developed to mitigate inequalities and enhance the quality of life for residents in North East London.

Bibliography

[1] Child Poverty Action Group. Official Child Poverty Statistics: 350,000 More Children in Poverty and Numbers Will Rise. Accessed: April 2024.

[2] GOV.UK. English Indices of Deprivation 2019. Accessed: April 2024.

[3] NHS North East London. About NHS North East London. Accessed: April 2024.

[4] Trust for London. Population Over Time. Accessed: April 2024.

[5] The Standard. Life Expectancy Falls in Two Years in London's Poorest Areas. Accessed: April 2023.

[6] Office for National Statistics (ONS). National Life Tables: United Kingdom. Accessed: April 2024.

[7] London Datastore. Average Income of Tax Payers by Borough. Accessed: April 2024.

[8] Bennett, J., et al. *The Relationship Between Life Expectancy and House Prices in London: 2002-2019.* 2023. Accessed: March 2024.

[9] Tong, Y. *Noise Perception and Urban Planning: A Comparative Study of London and New York City.* PhD thesis, 2022. Accessed: March 2024.

[10] Khan, K. *Decoding Urban Inequality: The Applications of Machine Learning for Mapping Inequality in Cities of the Global South.* Master's thesis, 2019. Accessed: March 2024.

[11] Bochra Hadj Kilani. *K-Means Clustering algorithms in Urban studies: A Review of Unsupervised Machine Learning techniques.* University of Carthage, Lab GADEV, December 2023. Accessed: March 2024.

[12] University of Manchester. *Multiple Linear Regression.* CMIST, 2020. Accessed: March 2024.

[13] Ranstam, J., and Cook, J. *LASSO Regression.* British Journal of Surgery, Volume 105, Issue 10, Page 1348, 2018. Accessed: March 2024.

[14] Hoerl, A. E., and Kennard, R. W. *Ridge Regression: Applications to Non-Orthogonal Problems.* Wiley Interdisciplinary Reviews: Computational Statistics, Volume 2, Issue 3, Pages 352-358, 2010. Accessed: March 2024.

[15] Rogerson, P. A., and Fotheringham, A. S. *The SAGE Handbook of Spatial Analysis.* Sage Publications, 2008. Accessed: March 2024.

[16] Lewis-Beck, C., and Lewis-Beck, M. S. *Applied Regression: An Introduction.* SAGE Publications, 2015. Accessed: March 2024.

[17] Xu, R., and Wunsch, D. *Survey of Clustering Algorithms.* IEEE Transactions on Neural Networks, Volume 16, Issue 3, Pages 645-678, 2005. Accessed: March 2024.

[18] Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., and Xu, X. *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.* ACM Transactions on Database Systems (TODS), Volume 42, Issue 3, Article 19, 2017. Accessed: March 2024.

[19] Rokach, L., and Maimon, O. *Clustering Methods.* In: Maimon, O., and Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, 2005. Accessed: March 2024.

[20] Reynolds, D. A. *Gaussian Mixture Models.* Tutorial, 2009. Accessed: March 2024.

[21] Pandas Development Team. pandas: Powerful Python data analysis toolkit. Accessed May 2024.

[22] NumPy Developers. NumPy: The fundamental package for scientific computing with Python. Accessed May 2024.

[23] Scikit-learn Developers. Scikit-learn: Machine Learning in Python. Accessed May 2024.

[24] Matplotlib Development Team. Matplotlib: Visualization with Python. Accessed May 2024.

[25] Seaborn Development Team. Seaborn: Statistical data visualization. Accessed May 2024.

[26] SciPy Developers. SciPy: Open source scientific tools for Python. Accessed May 2024.

[27] Folium Developers. Folium: Python Data. Leaflet.js Maps. Accessed June 2024.

[28] Leaflet.js Developers. Leaflet: An Open-Source JavaScript Library for Interactive Maps. Accessed July 2024.

[29] OpenStreetMap Contributors. OpenStreetMap: Open Data Commons Open Database License (ODbL). Accessed July 2024.

[30] Scikit-Fuzzy Developers. Scikit-Fuzzy: Fuzzy Logic Toolbox for Python. Accessed May 2024.

[31] QGIS Development Team. QGIS: A Free and Open Source Geographic Information System. Accessed June 2024.

[32] Burn-Murdoch, J. *The Anglosphere has an advantage on immigration.* Financial Times, April 26, 2024. Accessed: July 2024.

[33] BCS, The Chartered Institute for IT. *BCS Code of Conduct.* Accessed: July 2024.

Data Sources

[1] Trust for London. About Trust for London. Accessed April 2024.

[2] Trust for London. Qualifications Obtained by Borough. Accessed April 2024.

[3] Trust for London. Child Obesity Statistics. Accessed April 2024.

[4] London Datastore. Home Page. Accessed April 2024.

[5] London Datastore. Land Area and Population Density by Ward and Borough. Accessed April 2024.

[6] Trust for London. Life Expectancy by Borough. Accessed April 2024.

[7] Kaggle. Home Page. Accessed April 2024.

[8] Kaggle. UK Hospitals Dataset. Accessed April 2024.

[9] Office for National Statistics. About Us. Accessed April 2024.

[10] Office for National Statistics. Life Expectancy for Local Areas of the UK. Accessed April 2024.

[11] Office for National Statistics. Licences. Accessed April 2024.
Source: Office for National Statistics licensed under the Open Government Licence v.3.0

[12] Office for National Statistics. Life Expectancy for Local Areas of the UK (2001-2019). Accessed April 2024.

[13] OpenStreetMap. List of London Underground Stations. Accessed June 2024.

[14] Wikipedia. Tube Stations in London by Borough. Accessed June 2024.

[15] Transport for London. FOI Request Detail: Reference ID FOI-0759-2425. Accessed June 2024.

[16] Cartography Vectors. London Boroughs GeoJSON. Accessed June 2024.