Developing a Predictive Information System for Healthcare Financial Risk Analysis: A Case Study on Public Health Funding Models in the UK

By
Camaren Rogers
Students Number M01012741
Master of Science Degree
at
MIDDLESEX UNIVERSITY
October 2025
Supervisor: Dr Stylianos Kapetanakis

Abstract

This dissertation aims to offer regional insights on NHS funding distribution patterns, uncover financial risk trends in healthcare outcomes, and support better-informed resource allocation decisions in the UK publicly funded healthcare systems. Method: By developing a predictive information system that uses SQL-based tools and business Intelligence (BI) techniques to analyse and visualise healthcare financial risks the NHS faces within the inequalities of its system. Research questions: Which regions in the UK demonstrate the highest variability in healthcare spending vs. outcomes, and how can a SQL-based information system cluster these regions based on financial risk levels? What trends and anomalies can be detected in NHS funding distribution over the last 10 years, and how can these be visualized using a SQL-integrated BI system?

Acknowledgement

I extend my heartfelt gratitude to Middlesex University, London, and my course Business Information Systems Management. A special thanks to my supervisor, Dr. Stylianos Kapetanakis for his invaluable guidance, wisdom, motivation, and unwavering support throughout this journey. I would also like to thank Dr. Carlisle George for his helpful insight into digital and data healthcare, for introducing me to other technical innovations and information in the UK healthcare sector and being such a pivotal support doing my research. I am deeply grateful to my Mom and Dad for their financial support, love, and care, which carried me through to this point and for continuing to support me throughout this journey. I also want to thank my friends and course mates for the roles they played in this journey, always giving me a reason to push myself further.

Table of Contents

Introduction

Problem Description

Healthcare data reveal social, regional, and ethnic inequalities in the United Kingdom (Bailoni, 2011). The United Kingdom is strong-armed to meet current challenges such as rising cost of care and treatment, limiting public spending, and remedying defective and failing information systems. The issue of health inequalities is particularly sensitive in the United Kingdom as it reinforces other political issues and thus feeds into certain representations of identity and geopolitical grievances (Bailoni, 2011). Because resources in the United Kingdom are not being critically allocated according to needs, the efforts to improve equity are also being undermined. Many argue that the current UK capitation formula that was introduced over 20 years ago is outdated and does not sufficiently account for clinical needs.

Objective and Scope

To maximize primary care's ability to reduce health inequalities a key step is to design a health system that efficiently ranges funding for primary care services. By making better use of routinely available information and big data resources more fair resource allocation could be achieved. I have two main objectives in preparing this dissertation. The first is to review the UK's formula-based system for the equitable allocation of health care funding across regions paying particular attention to England. The focus of this dissertation is the design of architecture and functionality of predictive information systems for clustering, trend detection, and anomaly identification focusing on areas like regional disparities in funding efficiency, policy implications of financial risk clusters, effectiveness of SQL-based systems in public health analytics within the NHS systems. I want to look closely at what is realistically achievable in terms of attainable equity and what the implications are for the emphasis that should be placed on other objectives of health policy such as effectiveness, efficacy, and operational flexibility. It is important to provide people with the greatest health needs, receive the same level and quality of care. Spending the same amount everywhere doesn't guarantee that, because how efficiently regions use their funding can vary. The financial resources required by a regional health authority will depend on the population of the region. But people of different ages and genders have different health care needs, so this allocation should be changed to reflect different demographic and gender patterns that may exist between regions. Although, even given age and gender, people can differ in health care needs due to various morbidity and socioeconomic factors (Bond and Conniffe, 2001). Real life situations are much more complicated with more regions, many age categories, plus gender and other factors, as well as various classifications of health care expenditure being allocated separately. This dissertation addresses this problem of regional inequalities pertaining to allocation of financial funding and systemic inefficiency by analysing available NHS data in the goal of answering the following questions:

Which regions in the UK demonstrate the highest variability in healthcare spending vs. outcomes, and how can a SQL-based information system cluster these regions based on financial risk levels?

What trends and anomalies can be detected in NHS funding distribution over the last 10 years, and how can these be visualized using a SQL-integrated BI system?

This systematic literature review (SLR) conducted an in-depth analysis into the available literature surrounding this topic, aggregated the findings and discussed the overarching arguments and evidence. The search terms used were NHS Resource Allocation, Health Inequalities, NHS funds, Geographical Inequalities, UK Public Health, Regional Inequalities, Public Expenditure, Systematic Review, Data-Driven Decision-Making (DDDM).

Literature review

Overview of United Kingdom National Health Service

The UK has a government-sponsored universal healthcare system the National Health Service (NHS). The NHS consists of a series of publicly funded healthcare systems in the UK. It includes the National Health Services (England), NHS Scotland, NHS Wales, and Social Care in Northern Ireland. Citizens are entitled to healthcare under this system but have the option to buy private health insurance as well. The NHS Plan promises more power and information for patients, more hospitals and beds, more doctors and nurses, significantly shorter waiting times for appointments, improved healthcare for older patients, and tougher standards for NHS organizations (Chang et al., n.d.). In England, the health service is subdivided into regional agencies, the ten Strategic Health Authorities (SHA), and local agencies, the 152 Primary Care Trusts (PCT), which all come under the authority of the Ministry of Health (Bailoni, 2011). The UK government introduced a new goal for the allocation of resources in the NHS in England in 1999: “to contribute to the reduction in avoidable health inequalities.” (Department of Health, 2000) To better achieve this aim a health inequalities part was introduced into the allocation formula in 2002, which targeted more resources at deprived areas (Department of Health 2008). Consequently, increases in allocations since that time have tended to favor more deprived areas. (Barr et al., 2014). Demand for the NHS over the years continues to increase, and poor decisions on spending and financial funding have hindered the system's stability.

NHS Funding

The sustainability of NHS funding has been a huge concern for UK citizens for a long time. NHS spending represented 3.2% of gross domestic product (GDP) only five years after the NHS went public. More than seven decades later, spending has exceeded the growth in GDP so that by 2022 we spent around 9.3% of GDP on the NHS. More spending has been a result of pressure contributed by higher price inflation compared to the economy rather than the growth in the country’s economic wealth. Budgets have not increased well from year to year, while annual spending on the NHS has increased on average by around 3.4% in real terms (figure 1). Over the decade, from 2000 to 2010 annual increases averaged 6.2%. This stands for a significant financial gap in funding that has had a direct impact on the performance of the NHS and the quality of care it has been able to deliver. Prospects for the financial year 2024-2025 don’t look good, with the Institute for Fiscal Studies analysis of

planned day-to-day spending in England suggesting a real cut in funding of 1.2%, equivalent to 2 billion pounds plus a 0.75% real cut for the Scottish NHS and only a modest real rise in Wales of 0.7% (Appleby, Leng and Marshall, 2024). Three main factors drive the pressure to spend more on the NHS: population changes, income effects, and other cost pressures. Previously, the NHS has been funded from a combination of general taxation (income tax, VAT, and other duties and taxes), National Insurance contributions, and charges to patients. In 2024, funding from general taxation contributed around 80% of the NHS budget, National Insurance contributions will be around 18%, and charges to patients 2%. Although the absolute amount of privately funded healthcare has increased over time, so too have public sources, leaving the proportion of healthcare funded privately in the UK stable at approximately 20% (Appleby, Leng, and Marshall, 2024). The current approach to setting up public spending involves a spending review led by the HM Treasury and, through various public service agreements, defines the key improvements that the public can expect from these resources. Spending reviews usually take place every two to four years. Throughout my research four distinct recommendations have been suggested about NHS funding for a secure future to ensure stable spending: funding model, performance monitoring, strategic planning, and cash injection. It is suggested that following these recommendations would ensure the stability of the NHS funding model, with ongoing independent analyses of population health outcomes and the healthcare system, and a five-year strategic plan that would increase government accountability and help healthcare planning.

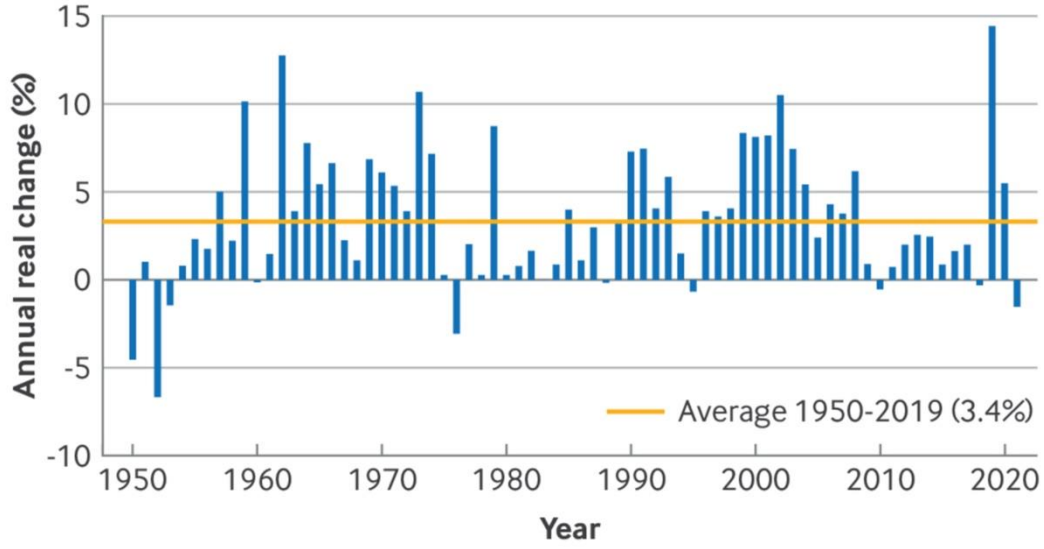


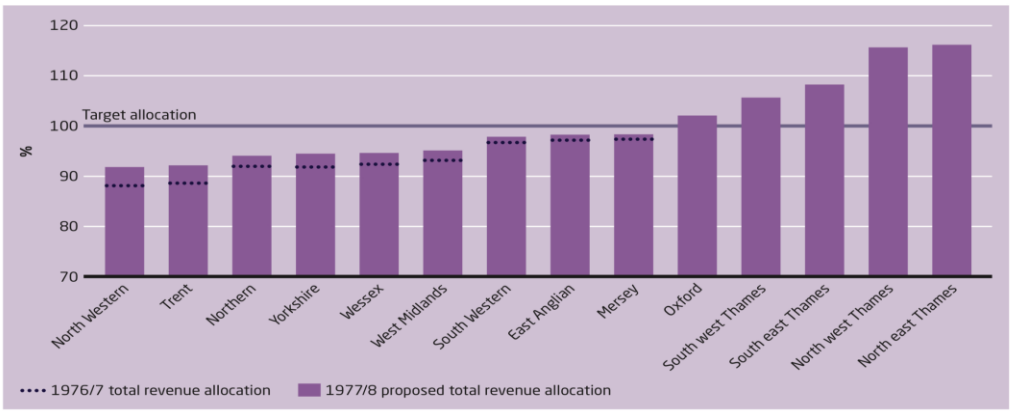
Fig 1 | Annual real changes in UK NHS spending 1950-51 to 2022-23. Data source: 1950-2018, British Social Attitudes survey[2]; 2019-22, authors' estimates based on UK health departments' annual accounts

Allocation of Health Resources

In the United Kingdom, resource allocation takes place at two levels: funds are first allocated between the constituent countries of the United Kingdom, each devolved nation has the authority to make its own decisions about how health services are organized, funded, and delivered, this is known as the Barnett formula. Second, resources are allocated with England to local purchasers. The Department of Health decides how its annual budget gets divided between commissioners responsible for buying services in different areas of the country. The Department of Health gets a fixed amount of money called the Departmental Expenditure Limit, and it is not allowed to

spend more than that. The department then must make two decisions: how much to allocate to NHS commissioners and other bodies; and then how much to allocate to each commissioner. The responsibility for these decisions currently lies with the Secretary of State for Health. The mission to improve healthcare regional inequality has been a problem for years. In the 2000s, the labour government continued to try and tackle avoidable health inequalities, aiming to reduce disparities in health outcomes, not just access. This led to further adjustments, directing more funds to areas with unmet or unexpressed health needs. Changes were made to fix unfair differences between regions by giving more NHS money to places that needed it most, however an analysis of the Resource Allocation Working Party (RAWP) reveals stark disparities. The RAWP was a group set up in 1975 within England's National Health Service to tackle how to fairly distribute NHS funding across different regions based on actual health needs, a principle that stays in place today. The results of this process showed a significant difference between the amount of resources areas were receiving and what they should be receiving based on RAWP's assessment of differences in need, each area's 'target funding'. In particular, the Thames regions (encompassing London) and Oxford were over-funded compared with the rest of England, which was underfunded, see figure 2 below (Buck, D. and Dixon, A., 2013). In the most recent data, which is not very recent, each primary care trust (PCT) makes its own decisions about how much to spend on public health, out of their overall allocation. Under the reforms, a range of public health functions transferred to local authorities, and a new executive agency of the Department of Health, Public Health England (PHE) has been set up. The secretary of State will decide how much to spend on public health overall, and with the advice of PHE, distribute it to distinct functions. This includes a new ring-fenced grant to each local authority to fulfil their responsibilities (Buck, D. and Dixon, A., 2013). However, in my research several pointers have been suggested to improve the current approach which I agree in especially pertaining to the lack of available data, first being greater transparency in the work of the Advisory Committee on Resource Allocation (ACRA), with greater opportunities to consult on proposals for future revision to formulas.

Figure 2 Funding relative to need in the NHS by Region – the RAWP formula



SQL-based Information Systems

Significant advancements in the healthcare sector have involved the implementation and development of SQL-based information systems. Hospital information systems were designed primarily for administrative purposes to ensure that all charges were billed and collected correctly. Today the application of information technology in healthcare domain emerged as a new multimillion dollar industry. Many government and private organizations are making huge investments to produce health related tools (Abdullah, Sawar & Ahmed, 2009). Structured Query Language (SQL) is one of the most pivotal tools for managing and analyzing data stored in relational databases. SQL allows analysts and data professionals to perform complex queries, join multiple data sources, aggregate and filter large datasets, and automate repetitive tasks thereby significantly enhancing the efficiency and accuracy of data analysis. The most frequent SQL queries used in applications are those that retrieve information from one or more tables in the database. The select clause determines which fields (columns) constitute the query output, the from clause determines which tables are used, and the join determines the criterion for joining rows from different tables. Then where clause filters the row based on some other criteria. The group by cause indicates how to combine the selected rows, and the having clause performs a final filter based on other conditions. Additionally, the order by clause determines how to order the resulting set of data (Tuya, Suárez-Cabal and de la Riva, 2006). The following simple query shows these clauses in SQL previously mentioned:

```
SELECT hospital_name, COUNT(*) AS  patient_count
FROM patient_records
WHERE diagnosis = 'Type 2 Diabetes'
GROUP BY hospital_name
ORDER BY patient_count DESC;
```

In the query above, the FROM patient_records pulls data from a table of patient records, then uses the WHERE diagnosis = 'Type 2 Diabetes' filters for patients diagnosed with the specific type of diagnosis. To group the data by hospital to see how many cases each one has, the GROUP BY hospital_name query is used. SELECT hospital_name, COUNT(*) is used to count the number of patients per hospital and lastly the ORDER BY patient_count DESC sorts the results so hospitals with the most cases appear first. Below shows a sample output table based on the SQL query which counts patients diagnosed with Type 2 Diabetes per hospital:

hospital_name	patient_count
St. Mary's Hospital	245
Royal London Hospital	198
Queen Elizabeth Hospital	174
Guy's and St Thomas'	162
Homerton University Hospital	139

The query ranks hospitals by the number of patients diagnosed with Type 2 Diabetes, helping analysts spot where the condition is most prevalent. This kind of data could inform targeted outreach, resource allocation, or further investigation into regional health disparities. As the healthcare sector increasingly depends on large-scale data infrastructures, SQL has appeared as a vital competency for deriving insight from data lakes and warehouses. In the realm of business intelligence, SQL acts as a foundational layer for integrating disparate data sources and supporting advanced analytics platforms such as Power BI, Tableau, and Looker. These tools rely heavily on SQL queries for data extraction, transformation, and loading (ETL) processes. According to Stack Overflow, SQL is still one of the most widely used programming languages globally, emphasizing its universality and relevance across roles. It is adaptability to cloud-based systems like Amazon Redshift, Google BigQuery, and Microsoft Azure

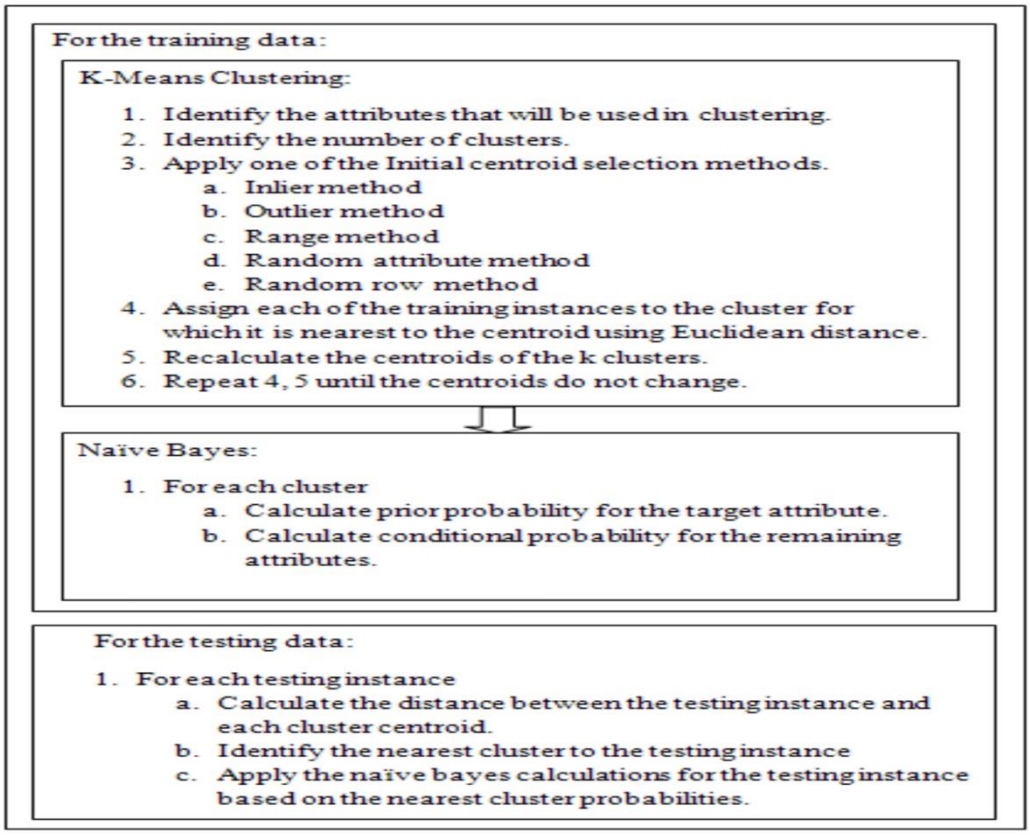
SQL Database further proves its enduring significance in modern business ecosystems (Maruf et al., 2022). Data-driven decision-making (DDDM) is a systematic approach wherein business choices and organizational strategies are informed through rigorous analysis and interpretation of data. The adoption of DDDM practices spans across industries, healthcare uses data to improve patient outcomes, retail predicts inventory needs, and financial institutions detect fraud using real-time analytics (Maruf et al., 2022). The advent of healthcare information management systems is now generating huge volumes of patient-centered, granular-level healthcare data. The high velocity of this data influences the relationship of hospitals and clinics with their patients and needs the use of analytics to tap into the needs, attitudes, preferences, and characteristics of clinical entities such as patients and practitioners. Healthcare information management systems are required to implement different data deployment, management and analytics strategies with the usage of big data tools, techniques and technology to utilize and healed the transformation of the heterogeneous healthcare data into valuable and useful insights (Imran et al., 2021). The integration of data mining algorithms with a relational Data Base Management System (DBMS) is an important and challenging problem. Carlos Ordonez, researcher on database technology for Teradata (NCR), introduces three SQL implementations of the popular K-means clustering algorithm to incorporate it with a relational DBMS in his research paper, "Integrating K-means Clustering with a Relational DBMS using SQL". During my research, similar to Ordonez's research, I also looked at data and query model from ESQP: An Efficient SQL Query Processing for Cloud Data Management by Jing Zhao, Xiangmei Hu and Xiaofend Meng from School of Information, Renmin University of China, where they discuss providing low latency for SQL query, including generalized selection, projection, aggregation and join. Low latency means that the first record of query results should be returned as soon as possible to prevent the client's longtime waiting. They explain a table is a collection of records, each of which is identified by a unique key, and each table is divided into n parts, each part replicated k times and are stored in different nodes in cluster. k is usually much smaller than the number of nodes in cluster while k = 2 for some SQL-based systems, mostly cloud computing, but it's a common starting point for examples (Zhao, Hu and Meng, 2010). In the next section I will explore more on clustering techniques and algorithms.

Clustering Techniques and Algorithms

Healthcare is one of the areas that significantly benefit from the development of machine learning techniques. Machine learning has the potential to help both patients and providers in terms of better care and lower costs. Developers at Stanford Center for Biomedical Informatics Research developed a machine learning model for predicting the diagnosis of depression up to one year in advance. Dr. Parag C. Pendharkar, Professor of Information Systems in the School of Business Administration at Penn State University, and Dr. Hitesh Khurana, Professor in the Department of Psychiatry at Pt. B.D. Sharma Post Graduate Institute of Medical Sciences, have also compared three different machine learning prediction methods for predicting patient's length of stay in Pennsylvania Federal and Specialty hospitals. Machine learning is a collection of data analytical techniques programmed to learn patterns from data sets. Using mathematical rules and statistical assumptions, machine learning techniques: supervised learning and unsupervised learning. Clustering methods are the most common unsupervised learning methods. Through the trading in steps of the machine learning methods, the optimal model's parameters are found by calculating the errors and evaluating the model's performance through some back-and-forth steps. Then, using the

optimal parameters, the model can be used for any new data set (Yousefi et al., 2020). There also have been several studies on the classification and clustering of patients based on specific diseases. Researchers at the School of Engineering and Information Technology at University of New South Wales at the Australian Defense Force Academy integrated decision tree and K-means clustering to predict heart disease on Cleveland Clinic Foundation Heart disease data set. Clustering patients into different priority classes immediately after their arrival could help the healthcare facility schedules patients in a better way (Yousefi al., 2020). K-means clustering is one of the most popular and well-known clustering techniques because of its simplicity and good behavior in many applications (Shouman, Turner & Stocker, 2012). The steps used in k-means clustering are shown in Figure 3.

Figure 3: Intergrating K-means Clustering and Naïve Bayes



Naïve Bayes is one of the data mining techniques that show considerable success in classification problems and specially in diagnosing heart disease patients. Naïve Bayes is based on probability theory to find the most likely possible classifications. It is based on prior probability of the target attribute and the conditional probability of the remaining attributes. For the training data the prior and conditional probability are calculated for each cluster. For each testing instance in the testing dataset, the probability is calculated with each of the target attribute values and the target attribute value with the largest probability is then selected. The probability of the testing instance for the target attribute value is calculated using the following formula:

$$P(v=ci) = P(ci) \times \quad (12)$$

Where v is the testing instance, ci is the target attribute value, aj is a data attribute, and vj is its value (Shouman, Turner & Stockers, 2012). Many existing efficient clustering algorithms are hard to implement inside a RDBMS where the programmer needs to worry about storage management, concurrent access, memory leaks, fault tolerance, security, and so on. Using SQL has not been considered an efficient and feasible way to implement data mining algorithms. SQL does not work with data mining, machine learning, or statistical algorithms. During my research I've found that it is possible to get an efficient SQL implementation of the popular K-means clustering algorithm that can work on top of a relational DBMS. As I mentioned earlier research done by Carlos Ordonez, from a performance point of view, explains how to cluster large data sets defining and indexing tables to store and retrieve intermediate and results, optimizing and avoiding joins, improving and simplifying clustering aggregations, and taking advantage of sufficient statistics. The basic input for K-means is a dataset Y containing n points in d dimensions, Y = {y1, y2,..., yn}, and k, the desired number of clusters. The output is three matrices W, C, R, containing k weights, k means and k variances respectively corresponding to each cluster and a partition of Y into k subsets. Matrices C and R are d x k and W is k x 1. Throughout this work three subscripts are used to index matrices: i = 1,..., n, j = 1,..., k, l = 1,...,d. To refer to one column of C or R we use the j subscript (e.g. Cj, Rj); Cj can be understood as a d-dimensional vector containing the centroid of the jth cluster having the respective squared radiuses per dimension given by Rj. For transposition we will use the T superscript. For instance, Cj refers to the jth centroid in column form and CTj is the jth centroid in row form. Let's say we've split the dataset Y into k groups, called X_1, X_2, ..., X_k, based on clustering. Each group is separate, with no overlap between them. K-means works by measuring how close each data point y_i is to the center of a cluster (C_j). It uses Euclidean distance, which is basically the straight-line distance between two points. The formula for this distance is:

$$d(y_i, C_j) = \sqrt{[(y_{i1} - C_{j1})^2 + (y_{i2} - C_{j2})^2 + ... + (y_{i\_d} - C_{j\_d})^2]}$$

In other words, it adds up the squared differences between each feature of the point and the centroid, then takes the square root (Ordonez, 2004). Centroids Cj are generally initialized with k points randomly selected from Y for an approximation when there is an idea about potential clusters. The algorithm iterates through executing the E and the M steps starting from some initial solution until cluster centroids become stable. The E step determines the nearest cluster for each point and adds the point to it. That is, the E step determines cluster membership and partitions Y into k subsets. The M step updates all centroids Cj by averaging points belonging to the same cluster. Then the k cluster weights Wj and the k diagonal variance matrices Rj are updated based on the new Cj centroids. The quality of a clustering solution is measured by the average quantization error q(C). K-means try to make each group as tight and compact as possible. It does this by finding the best set of centroids (group centers) that minimize the average distance between each data point and the center of the group it belongs to. It calculates the average of all the distances between each point y_i and its assigned centroid C_j and tries to make that number called q(C), as small as possible. This quantity measures the average squared distance from each point to the cluster where it was assigned according to the partition into k subsets. K-means finishes when the value it's trying to minimize, q(C), changes by only a tiny amount between steps, meaning the clusters have stabilized. K-means is theoretically guaranteed to converge decreasing q(C) at each iteration, but it is common to set a maximum number of iterations to avoid long runs (Ordonez, 2004). There are two main schemes presented in Ordonez's research. The first one presents a simple implementation of K-

means explaining how to program each computation in SQL. He refers to this scheme as the Standard K-means implementation. The second scheme presents a more complex K-means implementation incorporating several optimizations that dramatically improve performance. Ordonez's calls this scheme the Optimized K-means implementation.

Basic Framework

The basic scheme to implement K-means in SQL, having Y and k as input follows these steps:

1. Setup. Create, index and populate working tables.
2. Initialization. Initialize C.
3. E step. Compute k distances per point yi.
4. E step. Find closest centroid Cj to each point yi.
5. M step. Update W, C and R.
6. M step. Update table to track K-means progress.

Steps 3-6 are repeated until K-means converges.

Most of the SQL code to implement K-means involves Data Manipulation Language (DML) statements. The columns making up the primary key of a table are underlined. Tables are indexed on their primary key for efficient join access. Subscripts i, j, l are defined as integer columns and the d numerical dimensions of points of Y, distances, and matrix entries of W, C, R are defined as FLOAT columns in SQL. Before each INSERT statement it is assumed there is a "DELETE FROM ... ALL;" statement that leaves the table empty before insertion. Like mentioned the input data set has d dimensions which means there exists a table Y with several numerical columns out of which d columns are picked for clustering analysis. In practice the input table may have many more than d columns but to simplify exposition we will assume its definition is Y (Y1, Y2,..., Yd). So the SQL implementation needs to build its own reduced version projecting the desired d columns (Ordonez, 2004). This motivates defining the following "horizontal" table with d + 1 columns: Y H (i, Y1, Y2,..., Yd) having i as primary key. The first column is the i subscript for each point and then Y H has the list of d dimensions. This table saves Input/Output access (I/O) since it may have fewer columns than Y and it is scanned several times during the algorithm progress. In general, it is not guaranteed i (point id) exists because the primary key of Y may consist of more than one column, or it may not exist at all because Y is the result of some aggregations. It is necessary to automatically create i guaranteeing a unique identifier for each point yi. The following statement goes through Y once, adds up values step by step, and keeps only the d columns wanting (Ordonez, 2004).

```
INSERT INTO YH
SELECT sum(1) over(rows unbounded preceding) AS i
,Y1, Y2,..., Yd
FROM Y;
```

The point identifier can be generated with some other SQL function than returns a unique identifier for each point. Getting a unique identifier using a random number is not a good idea because it may get repeated, especially for very large data sets. Clustering results are stored in matrices W, C, R, motivating having one table for each of them storing one matrix entry per row to allow queries access each matrix entry by subscripts j and l. So the tables are as follows: W(j, w), C(l, j, val), R(i, j, val), having

k, dk and dk rows respectively. The table YH defined above is useful to seed K-means, but it is not adequate to compute distances using the SQL "sum()" aggregate function. So it has to be transformed into a "vertical" table having d rows for each input point, with one row per dimension (Ordonez, 2004). This leads to table YV with definition YV (i,l, val). Then table YV is populated with d statements as follows:

INSERT INTO YV SELECT i, 1, Y1 FROM YH;
...
INSERT INTO YV SELECT i,d,Yd FROM YH;

Ordonez defines a table to store several useful numbers to track K-means progress. Table model serves this purpose: model(d, k, n, iteration, avg_q, di f f_avg_q). Standard K-means presents scalability problems with increasing number of clusters or number of points. Its performance graphs exhibit nonlinear behavior. Optimized K-means is significantly faster and exhibits linear scalability (Ordonez, 2004). To conclude, my research shows that regional differences in healthcare funding and outcomes continue to be a major issue. Traditional evaluation methods often fall short in capturing the full complexity of these disparities, especially when it comes to linking financial inputs with health outcomes. This dissertation builds on newer approaches, like SQL-based clustering and interactive BI tools, to offer a clearer, data-driven view of where gaps exist and how inequalities in healthcare can be identified.

Methodology

Overview of Research Design

Industries have adopted data-centric decision frameworks because of how substantially more productive and profitable they perform compared to the result of their counterparts. Data-driven cultures have been linked to improvements in cost reduction and innovation. The adoption of DDDM practices spans across industries, healthcare use data to improve patient outcomes, retail predicts inventory needs, and financial institutions detect fraud using real-time analytics (Maruf et al., 2022). SQL is one of the most pivotal tools for managing and analyzing data stored in relational databases. SQL allows analysts and data professionals to perform complex queries, join multiple data sources, aggregate, and filter large datasets, and automate repetitive tasks thereby significantly enhancing the efficiency and accuracy of data analysis (Maruf et al., 2022). The use of data analytics, and particularly SQL, offers a powerful remedy for addressing the shortcomings of inefficiencies, real-time decision making, and data silos within healthcare information systems. SQL-based analytics enables significantly more efficient heart workflows because we can see how this helps with data analytics to drive operational efficiency. SQL systems support real-time decision making and resource optimisation (Chanda, 2024). To develop a predictive information system for healthcare financial risk analysis I aimed to design a database schema compiled of tables for regions, funding, outcomes, and risk scores. Using SQL queries for clustering, trend detection and anomaly identification I aim to highlight results applying BI integration, dashboards displaying regional clusters, funding vs. outcomes over time, and anomalies and outliers. The specific goal is to analyze available data and design a streamlined workflow and system to solve the mismatch between policy rhetoric and funding decisions. The multiple organizations and actors in national and local public health are often collectively defined as a complex system (Evans, 2020). For my research design I used a quantitative as well as a case study approach. My data sources were from NHS Digital, ONS, and Public Health England. I chose to analyze 10-year

historical data on funding, outcomes including mortality, and multidimensional poverty. I also used to normalize financial and outcome data for comparability. For analytical techniques I chose using a clustering algorithm, k-means, and hierarchical clustering to group regions by financial risk. To analyze trends, I used time-series analysis to detect funding anomalies. To display visualization, I used SQL-integrated BI tools to create dashboards. Historic records of quantitative measures were sourced from the Office for National Statistics which provide a comprehensive, internationally comparable time series of UK healthcare spending, disaggregated by key dimensions (finance, function, provider) since 2013–14 to 2021–22. The most recent data shows a slight overall rise in nominal healthcare spending, but a first-ever decline in real-terms government spending reflecting a pandemic aftermath shift and notable growth in private expenditure. My primary unit of analysis was UK Health Accounts data collected from the ONS, along with data from Index of Multiple Deprivation (IMD), National Institute for Health and Care Research (NIHR), Institute for Fiscal Studies (IFS) etc.

Data Preparation

Once all the necessary datasets are gathered, I started applying data cleaning and transformation techniques to identify and correct errors, inconsistencies, or inaccuracies in the datasets. This ensures my data is trustworthy before my modeling and analyzing. My cleaning tasks included handling missing values, either filling, removing, or flagging them, removing duplicates to avoid skewed results, correcting inaccuracies like typos or misclassifications, standardizing formats, such as date formats and capitalization, and managing outliers to fit my analytical goals. Data transformation was a necessary step to reshape my cleaned data into formats that suited my analytical goals, making data usable and compatible across my systems and models. This included aggregating, joining datasets, normalizing or scaling, encoding categorical variables and converting formats. Excel was my primary program for my initial data cleaning process. For advanced cleaning I mostly leveraged power query in excel for removing rows, filter, transform types, and automate workflows. Problems that I ran into with cleaning in excel were formatting data correctly to be able to come up with a cohesive and legible visual analysis. I was looking at the enormous amount of data and the number of unfamiliar Heathcare jargon and variables I needed to assess and prepared data in excel no longer seemed to be the most efficient way to avoid errors. Ultimately my goal is to present the results of my analysis visually using a SQL-integrated BI system. Therefore, given my issue with Excel, I moved to data preparation, processing, and analysis to Microsoft Power BI. Analyzing datasets directly in Microsoft Power BI fundamentally is faster, more dynamic, and is built for visual insight. Microsoft Power BI also has its version of Power Query; I believe it was more tailor-made for the analysis I aim to conduct. Power Query is Power BI's built-in ETL engine (Extract, Transform, Load). In Power Query you can connect to data sources, clean and reshape data, transform formats, and automate workflows. I found Power BI's Power Query more scalable than in Excel, especially features like its column distribution, which gives you a quick snapshot of nulls, distinct values, and potential outliers, perfect for clustering prep, create calculated columns, for example adding logic like:

EquityScoreCategory = if [EquityScore] > 80 then "High" else if [EquityScore] > 50 then "Medium" else "Low"

A notable feature during my processing is Merge Queries to combine input and output growth tables by name or variable. This shift from Excel to Power BI marked a turning point in my workflow, streamlining data preparation while unlocking more powerful, scalable tools for analysis and visualization.

Power BI Integration for Visualization

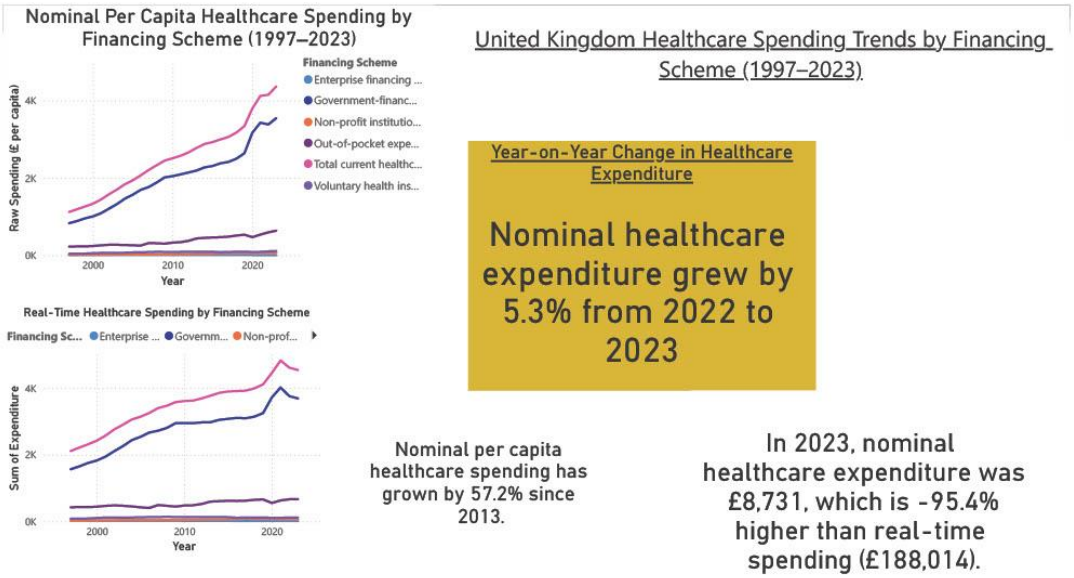
The first dataset I looked at in Power BI was a dataset from the Office for National Statistics (ONS) and accompanies the UK Health Accounts: 2023 and 2024 release. I chose it because it provides detailed reference tables on healthcare expenditure in the UK, segmented by financing scheme, healthcare function, and provider type. The data spans from 1997 to 2024, with 2024 figures marked as provisional. The Healthcare Financing Schemes break down UK healthcare spending into five key schemes, Government-financed, Voluntary health insurance, Non-profit institutions, Enterprise financing, Out-of-pocket. The key metrics tracked are total and per capita expenditure, annual growth rates, share of GDP, revenue sources for each scheme, healthcare spending by function, and provider types.

Table 1. Presents a subset of data from the UK Health Accounts: 2023 and 2024 release, showing nominal per capita healthcare expenditure from 1997 to 2023, disaggregated by financing scheme. Contains 162 rows and 128 distinct values.

Financing Scheme	Year	Expenditure
Total current healthcare expenditure	2023	4365
Total current healthcare expenditure	2022	4147
Total current healthcare expenditure	2021	4120
Total current healthcare expenditure	2020	3795
Government-financed expenditure	2023	3546
Government-financed expenditure	2021	3428
Government-financed expenditure	2022	3381
Total current healthcare expenditure	2019	3338
Total current healthcare expenditure	2018	3173
Government-financed expenditure	2020	3169
Total current healthcare expenditure	2017	3060
Total current healthcare expenditure	2016	2995
Total current healthcare expenditure	2015	2926
Total current healthcare expenditure	2014	2877
Total current healthcare expenditure	2013	2776
Total current healthcare expenditure	2012	2661
Government-financed expenditure	2019	2638
Total current healthcare expenditure	2011	2574
Total current healthcare expenditure	2010	2509
Government-financed expenditure	2018	2498
Total current healthcare expenditure	2009	2448
Government-financed expenditure	2017	2414
Government-financed expenditure	2016	2379
Total current healthcare expenditure	2008	2328
Government-financed expenditure	2015	2309
Government-financed expenditure	2014	2273
Total current healthcare expenditure	2007	2207
Government-financed expenditure	2013	2191

After importing and cleaning the dataset in Power BI Power Query I started to create narrative visuals to understand the dataset more and derive an analysis from it. I first looked at the nominal healthcare expenditure dataset and compared it with the real-term healthcare expenditure data, created line charts using both information. From there compiled a dashboard I named “United Kingdom Healthcare Spending Trends by Financing Scheme (1997-2023)”. Figure 4 shows the Power BI dashboard, which highlight year-to-year charge in healthcare expenditure and other key point indicators (KPIs).

Figure 4: Nominal Per Capita Healthcare Spending by Financing Scheme (1997-2023)

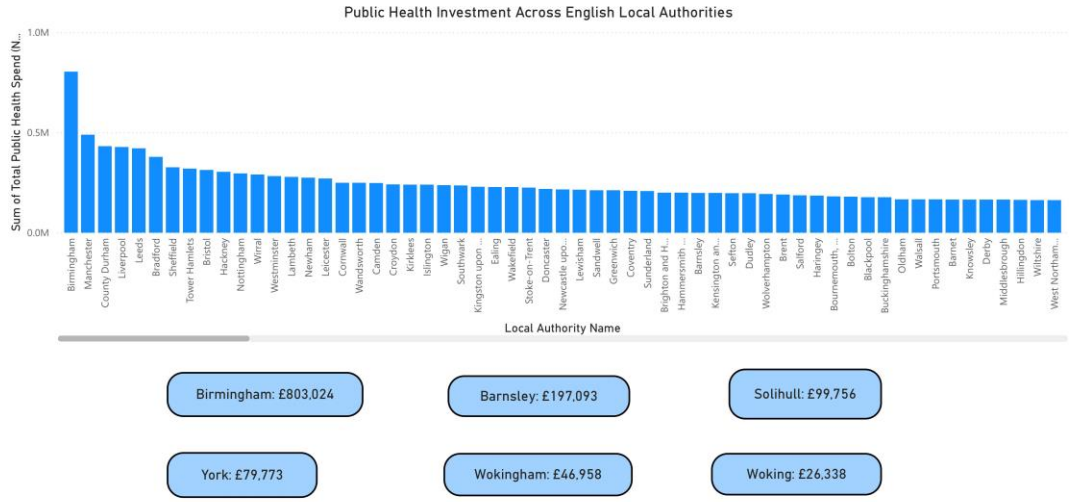


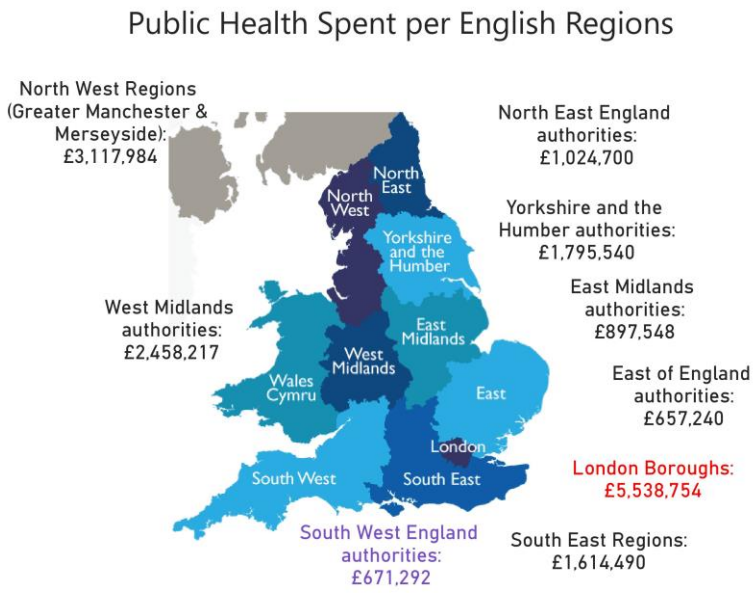
Since 1997, the UK's healthcare landscape has undergone dramatic shifts, not just in how much is spent, but in how that spending is financed. This dashboard explores the evolution of per capita healthcare expenditure across different financing schemes, revealing both long-term growth and recent acceleration. The line charts show a steady upward trajectory in healthcare spending per person, with notable divergence between financing schemes. Public schemes dominate in scale, but private and out-of-pocket spending have grown significantly, especially post-2010. This suggests increasing pressure on individuals and private insurers, possibly reflecting policy shifts or service gaps. In 2023, nominal expenditure was £8.731B, but real-terms spending was £188.014M, a staggering 95.4% inflation adjustment. This dataset and dashboard set the precedence and gave me a general knowledge of UK Healthcare spending, the government-financed scheme consistently accounts for the majority of healthcare expenditure, reinforcing the central role of the NHS and public health bodies. Out-of-pocket costs are rising, and voluntary and non-profit schemes remain small, which enterprise financing is minimal.

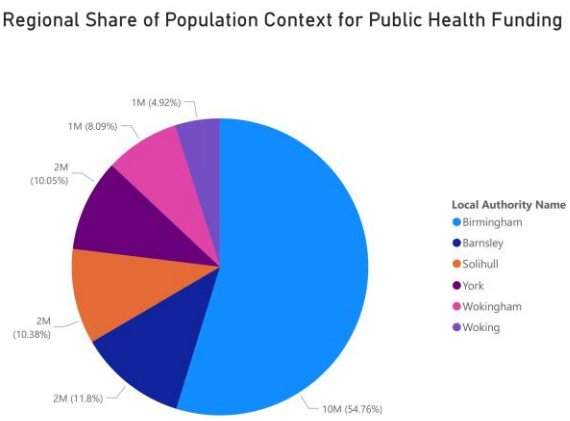
Initially my aim for this dissertation was to evaluate the whole of the United Kingdom's healthcare system in terms of spending and outcomes, however during my data preparation, collecting datasets from all four countries within the United Kingdom proved to be too difficult a task. Therefore, my focus was designated to England. The first England spending data I analyzed was a public health investment across English local authorities' dataset published by Public Health England, now part of the UK Health Security Agency and the Office for Health Improvement and Disparities. It's part of the UK government's commitment to transparency in local health spending and is typically released alongside the Public Health Grant allocations and Local Authority Health Profiles. The dataset is a detailed financial and programmatic snapshot of how public health funding is distributed across England's local councils. It's designed to support transparency, equity analysis, and strategic planning by showing not just how much is spent, but where, on what, and relative to population needs. The key features of the dataset are total net public health spend, which shows annual public health investment per local authority, population context, programmatic breakdown, and regional aggregates, summarizes total public health spend by region. Figure 5

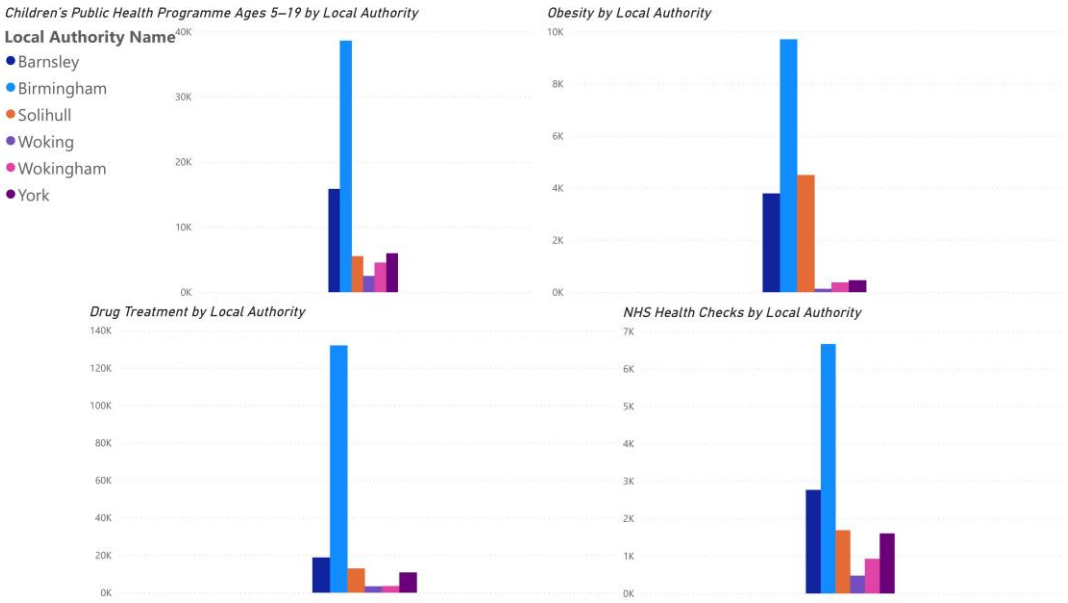
shows the Power BI dashboard containing the bar graph showing public health investment across English local authorities, as well as regional share of population context for public health funding, and public health spent per English regions, as well as certain program investment of specific local authority names.

Figure 5: Health Investment Dashboard England's Local Authorities





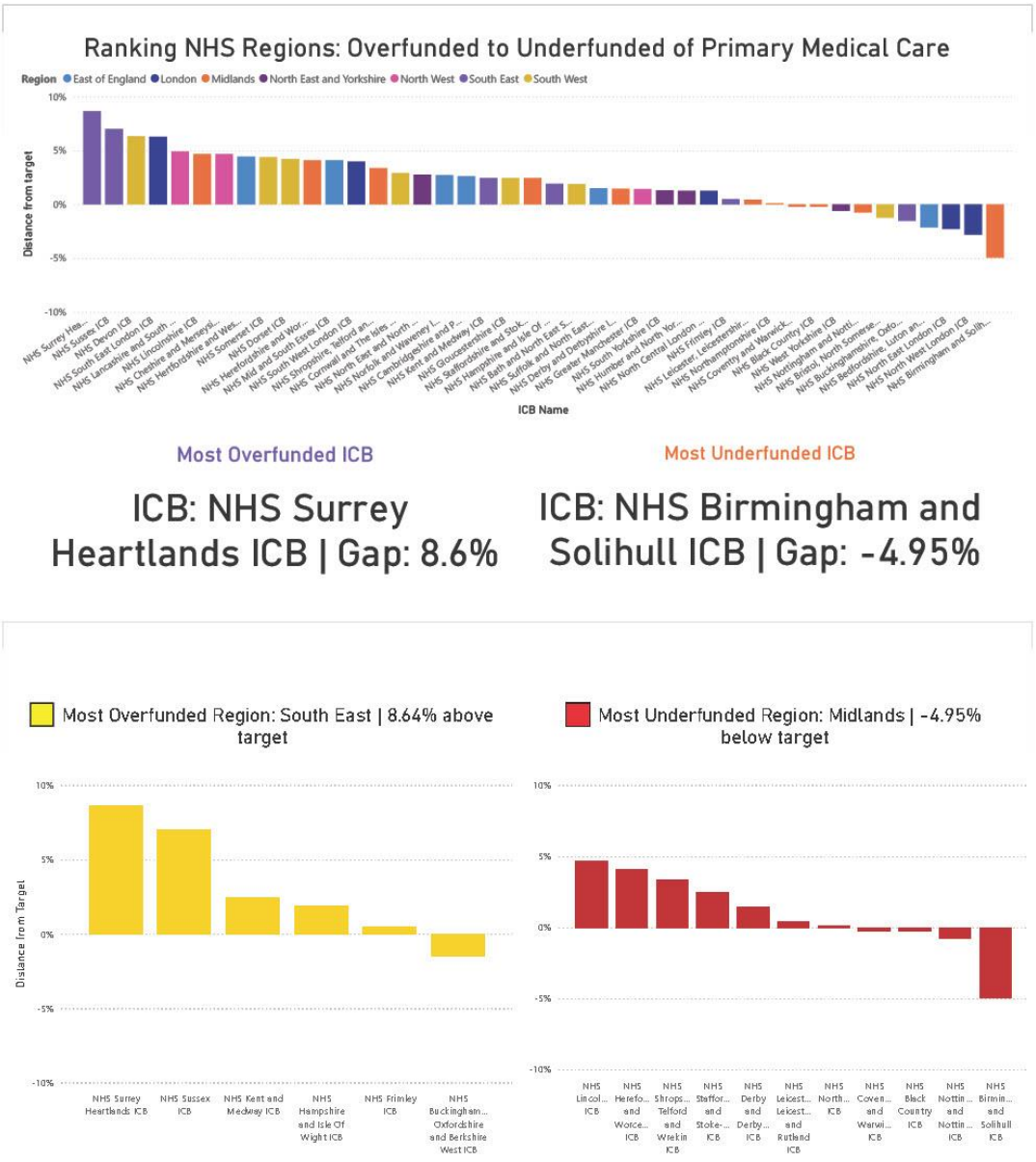




Birmingham receives the highest amount of public health funding per year among all local authorities in England, reflecting both the scale of its population and the complexity of its health needs. Birmingham is the largest local authority in the UK, with over 1.1 million residents. More people need more demand for public health services like vaccinations, mental health support, and sexual health clinics. The city has areas with significant socioeconomic challenges, including poverty, unemployment, and housing insecurity. These factors are strongly linked to poorer health outcomes, which increases the need for targeted public health interventions. Birmingham faces wide disparities in life expectancy, chronic disease prevalence, and access to care across different neighborhoods. Public health investment aims to reduce these inequalities through prevention and outreach. This makes Birmingham a prime case study for analyzing how public health investment correlates with outcomes. The dashboard also highlights Barnsley, Solihull, York, Wokingham, and Woking. These outliers were added to the dashboard after analyzing other data, making them a point of interest during my overall analysis.

The focus of my objective is assessing regional spending, so next I looked at a comprehensive financial dataset published by NHS England. The spreadsheet outlines funding distributions for the Integrated Care Boards (ICBs) across England for the fiscal years 2024/25 and 2025/26, with granular breakdowns by service type, region, and adjustment category. The dataset originates from NHS England's ICB Allocations and Planning Guidance documents, which are publicly released to support transparency and strategic planning across the health system. Each row in the spreadsheet corresponds to an ICB, showing region and ICB name, 2024/25 baseline funding, adjustments for pay deals, service transfers, and capacity, 2025/26 growth rates and convergence metrics, and final recurrent allocations per capita. The dataset covers multiple funding categories including core services, specialized services, primary medical care, POD services, and running costs. To assess unequal spending in healthcare I used this data to evaluate regional overfunding and underfunding of primary medical care by creating a chart that shows distance from target per region. "Distance from target" is a funding metric used by NHS England to measure how far an Integrated Care Board (ICB) is from its fair share allocation, the amount of funding it should receive based on factors like population size, health needs, deprivation levels, and service demand. Positive distance, for example, +8.6% means the ICB is overfunded, receiving more than its calculated target. Negative distance, for example -4.95%, means the ICB is underfunded, receiving less than its fair share. Figure 6 shows the Power BI dashboard containing the bar graph showing ranked NHS regions based of overfunded to underfunded of primary medical care received, as well as most overfunded region and most underfunded region.

Figure 6: Integrated Care Board Funding Analysis



Ranking NHS ICBs by Deviation from Target Funding for Primary Medical Care. Surrey Heartlands ICB is overfunded by 8.6%, while Birmingham and Solihull ICB are underfunded by 4.95%. These disparities inform predictive modeling and highlight regions requiring policy attention.

My analysis of ICB funding reveals a shocking contrast between the most overfunded and most underfunded ICB, while Surrey Heartlands enjoys an 8.6% surplus, Birmingham and Solihull face a 4.95% shortfall. The chart shows a spectrum of funding deviation, with some ICBs receiving significantly more than their target allocation, while others fall short, potentially impacting service delivery and population health. Underfunded regions often coincide with higher deprivation, greater disease burden, and elevated avoidable mortality. It reveals systemic imbalances that may correlate with regional health outcomes, particularly avoidable mortality.

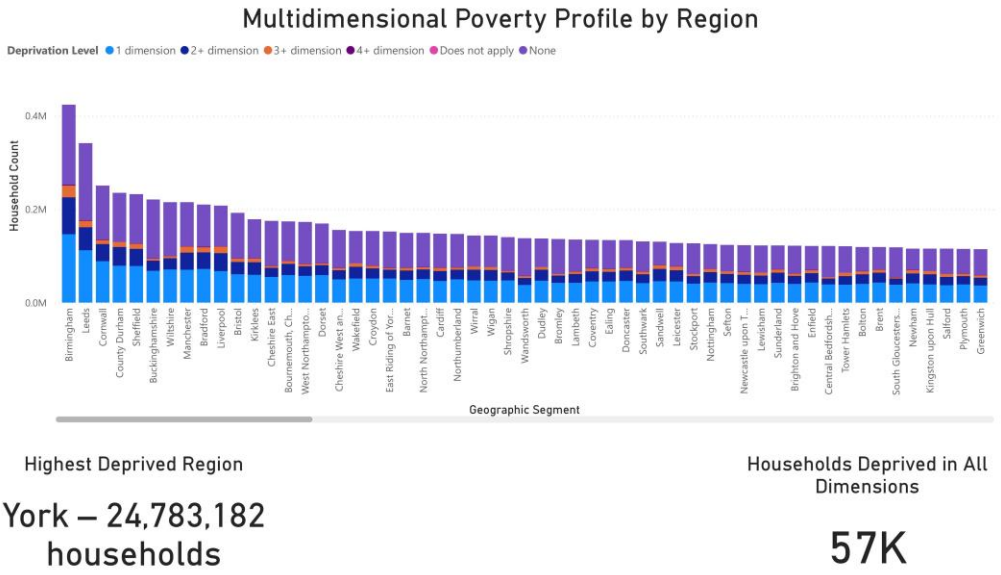
To understand regional vulnerability in health-related dimensions I next looked at households by deprivation dimensions. This information was published by the ONS using Census 2021 data. Households are classified into six categories based on four deprivation dimensions. Employment, unemployment or economic inactivity due to long-term illness/ disability, education, no one with Level 2 qualifications and no full-time students aged 16-18, health and disability, presence of bad/very bad health or disability, housing, overcrowding, shared dwellings, or lack of central heating. Each household is assigned a deprivation score from 0 (none) to 4+ (severely deprived across multiple dimensions). Each row represents a count of households in a specific Lower Tier Local Authority with a given deprivation score and includes over 300 local authorities, with granular counts for each deprivation level.

Table 2. This table presents household counts across varying levels of multidimensional deprivation ranging from one to four or more dimensions for specific UK local authorities, offering a snapshot of socio-economic vulnerability in regions like York and Wyre Forest.

Authority Code	Authority Name	Household Count	Deprivation Level
E06000014	York	*1918*	3+ dimension
E06000014	York	*27910*	1 dimension
E06000014	York	*9518*	2+ dimension
E06000014	York	*46025*	None
E06000014	York	*88*	4+ dimension
E06000014	York	*0*	Does not apply
E07000239	Wyre Forest	*20798*	None
E07000239	Wyre Forest	*15760*	1 dimension
E07000239	Wyre Forest	*6928*	2+ dimension
E07000239	Wyre Forest	*1730*	3+ dimension
E07000239	Wyre Forest	*106*	4+ dimension
E07000239	Wyre Forest	*0*	Does not apply
E07000128	Wyre	*1813*	3+ dimension
E07000128	Wyre	*23366*	None
E07000128	Wyre	*17707*	1 dimension
E07000128	Wyre	*7793*	2+ dimension
E07000128	Wyre	*88*	4+ dimension

This dataset serves as a highly valuable resource for identifying regional patterns of socio-economic vulnerability, particularly in the context of analyzing disparities in healthcare outcomes across England. One of the four deprivation dimensions is explicitly health-related, it flags households where someone reports bad or very bad health, or where there's a long-term disability. That gives you a direct signal of self-reported health burden at the household level, something that's often missing from spending datasets. Figure 7 shows the Power BI dashboard containing a bar graph showing Multidimensional Poverty Profile by Region.

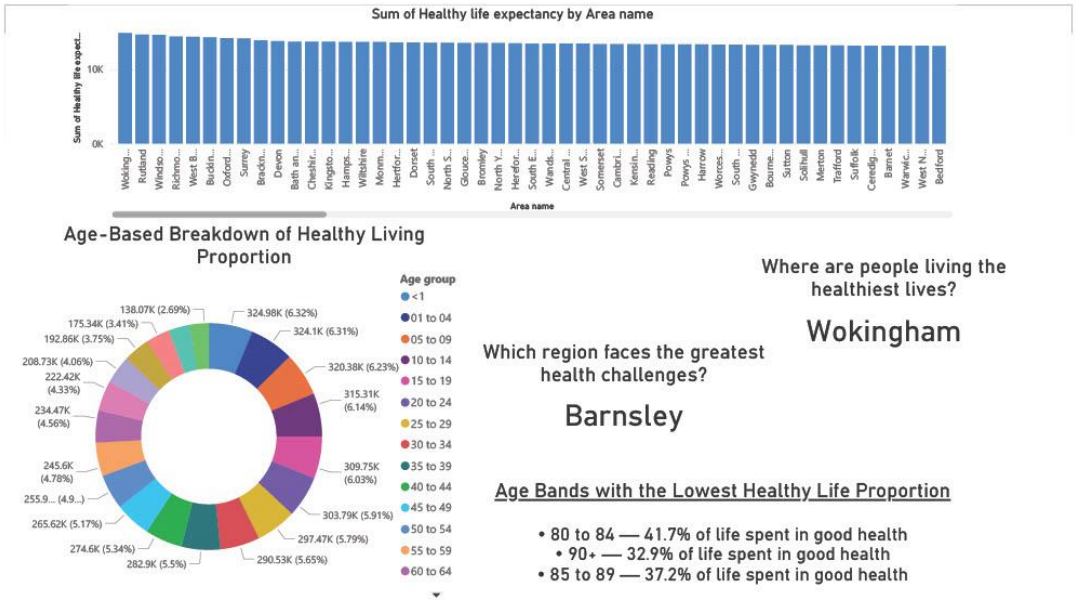
Figure 7: Multidimensional Poverty Profile by Region



By visualizing household counts across regions and deprivation levels, it challenges simplistic views of poverty and highlights where support is most urgently needed. Some regions may have fewer households in total poverty, but a higher proportion experiencing multiple, compounding disadvantages. The Highest Deprived Region is York, which tops the chart with 24,789 households experiencing deprivation, an unexpected finding that may reflect hidden structural challenges in a region often perceived as affluent. Severe Multidimensional Poverty: 57,000 households are deprived in all measured dimensions, representing the most vulnerable segment of the population. These households likely face intersecting barriers to health, education, employment, and housing. When combined with healthcare and financial risk models, this data can help forecast where deprivation may lead to higher public service costs, primary medical care is considered a public service cost in the UK and other publicly funded health systems, or poorer health outcomes.

To look at direct outcome metrics that can be compared against NHS spending data to see where investment does or doesn't translate into better health, I analyzed healthy life expectancy (HLE) compiled from ONS Healthy Life Expectancy estimates and regional demographic data derived from Census 2021, Public Health England, and NHS Digital sources. HLE measures the number of years individuals are expected to live in good health as well as total healthy years by region aggregated HLE values across local authorities. This dataset is a highly sensitive outcome metric, making it ideal for evaluating how well healthcare investment translates into real-world health. Figure 8 shows the Power BI dashboard containing a bar graph showing sum of life expectancy by area name, as well as an age-based breakdown of healthy living proportion.

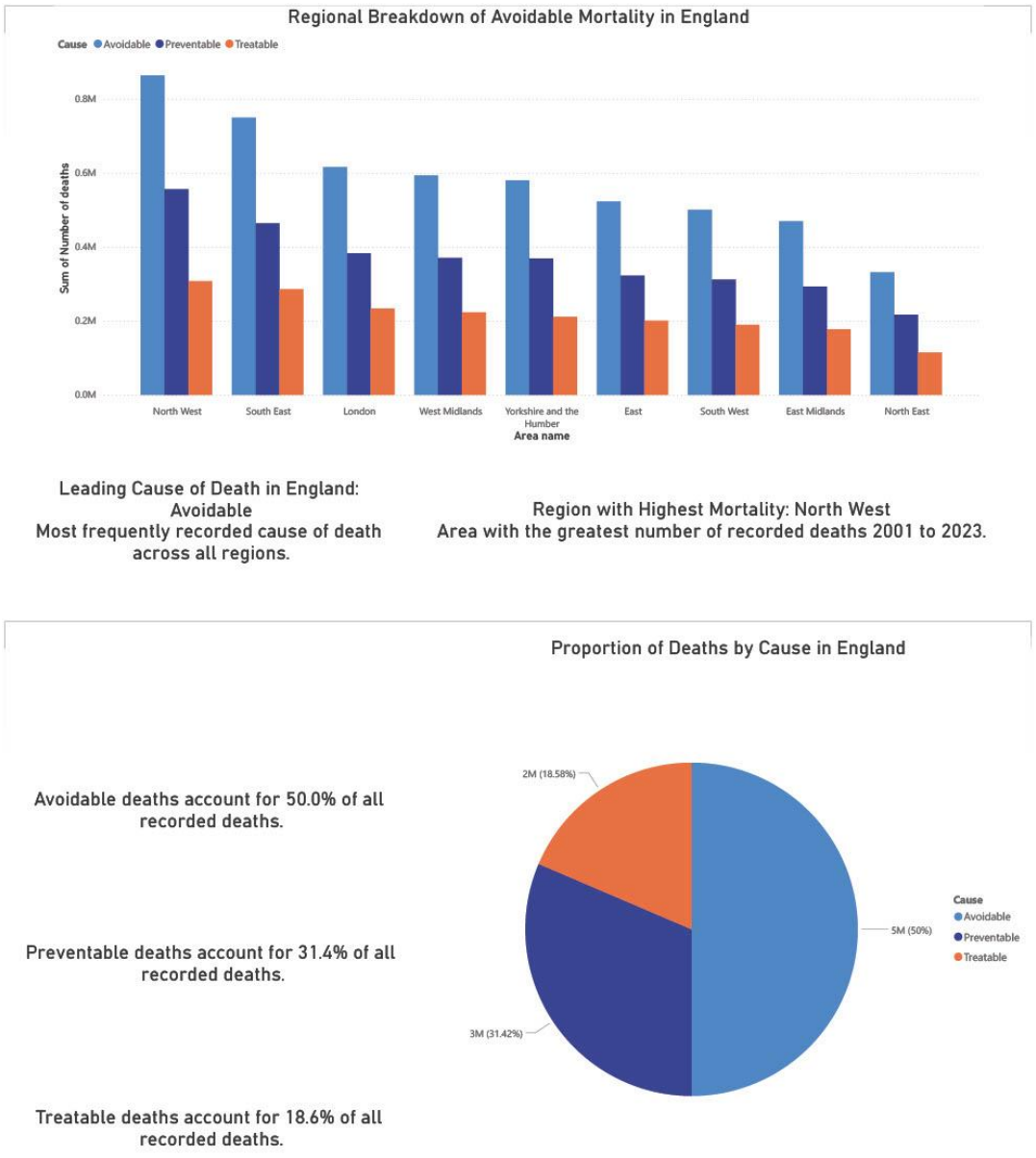
Figure 8: Healthspan Inequities Across Age and Region



The bar chart ranks regions by total years spent in good health. Wokingham stands out as the healthiest region, suggesting strong public health infrastructure, socioeconomic stability, or lifestyle factors. Regions like Wokingham may serve as benchmarks for best practice, while others may require targeted health interventions. Wokingham leads to a healthy life expectancy, indicating a population that not only lives longer but lives better. Barnsley faces the lowest healthy life expectancy, highlighting a potential concentration of chronic illness, poor access to care, or socioeconomic disadvantage. Regions like Barnsley may benefit from targeted health funding, preventive care programs, and infrastructure investment. Healthy life expectancy is a clean, interpretable outcome, it reflects not just survival, but quality of life, which is central to equity-focused healthcare analysis.

My second direct outcome measure to analyse and compare spending data to identify outliers is avoidable mortality statistics for deaths registered in 2001 to 2023. The data is published by Office for National Statistics and NHS Digitals, likely based on death registrations, cause-specific mortality, and age standardized death rates and covers regional breakdowns across England. The dataset measures all-cause mortality rates, cause-specific deaths, age-standardized mortality rates, and avoidable, preventable, and treatable mortality. Figure 9 shows the Power BI dashboard containing a bar graph showing regional breakdown of avoidable mortality in England, as well as a pie chart showing proportion of deaths by cause in England.

Figure 9: England Mortality Dashboard Avoidable, Preventable, and Treatable Deaths (2001 –2023)



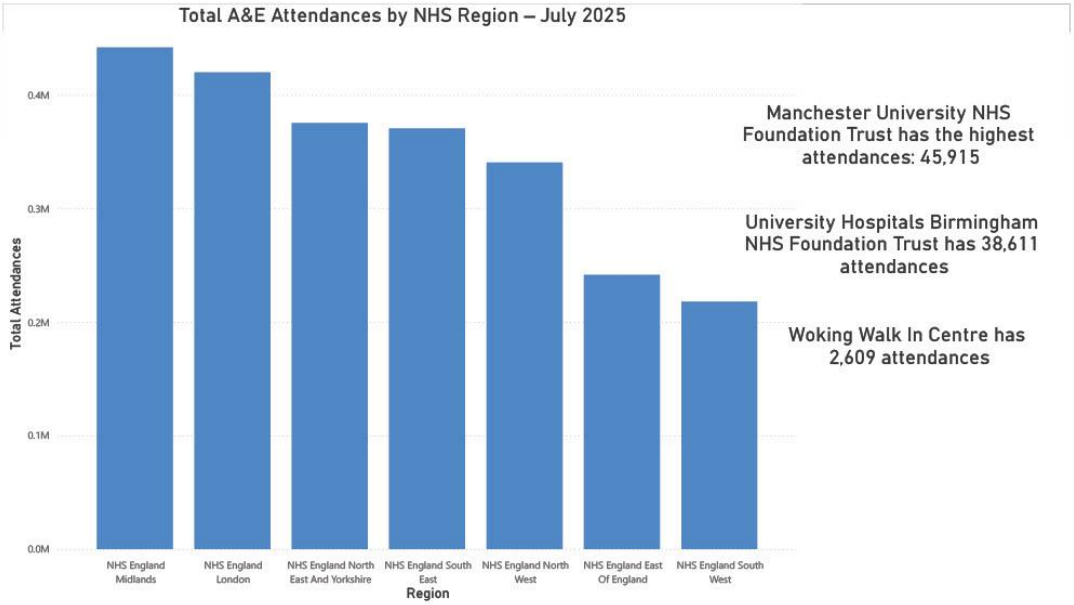
The Northwest region exhibits the highest cumulative avoidable deaths, indicating elevated public health and financial risk. Stratified bars suggest internal disparities by gender or cause, informing targeted predictive modeling. Avoidable deaths constitute 50% of all recorded mortality, with preventable and treatable deaths accounting for 31.4% and 18.6%, respectively. These categories inform the predictive system's risk stratification logic and highlight areas for targeted intervention.

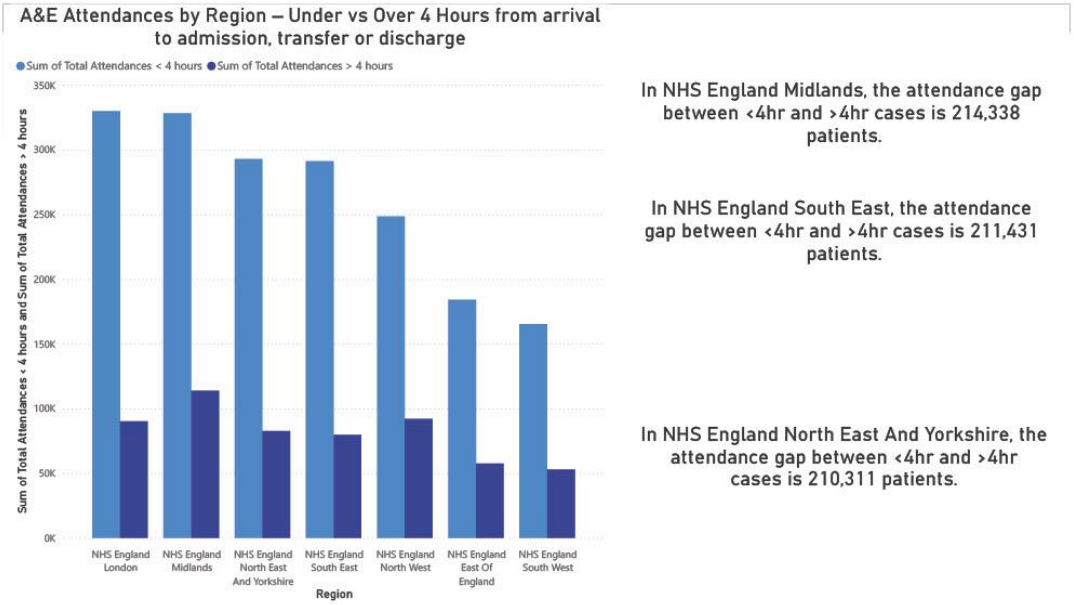
The visualizes longitudinal patterns in avoidable mortality across English regions, offering both descriptive insight and a foundation for predictive modeling. To identify regional disparities in avoidable mortality and correlate these with potential financial risk indicators, such as healthcare demand, resource allocation inefficiencies, and population health vulnerabilities. The Northwest recorded the highest number of

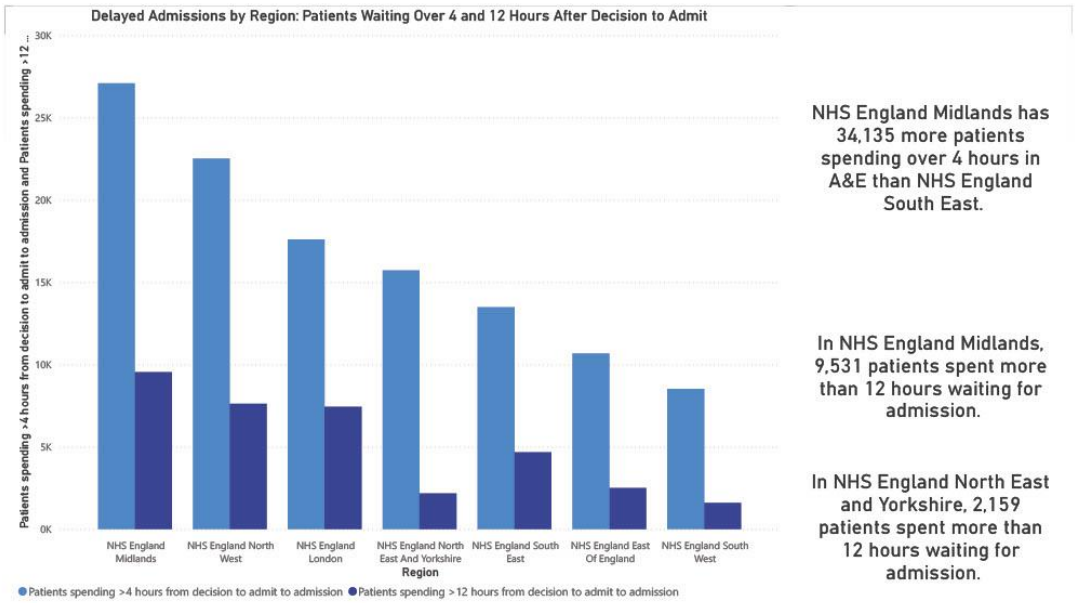
avoidable deaths over the 22-year period, positioning it as a region of elevated financial and public health risk. This trend may reflect compounded factors, higher deprivation indices, chronic disease prevalence, and systemic underinvestment. This region becomes a focal point for predictive modeling, with clustering algorithms, K-means or DBSCAN, used to identify high-risk local authorities within it. By visualizing historical mortality patterns, it enables the identification of high-risk clusters and informs the design of predictive algorithms that anticipate future financial strain on healthcare systems. The pie chart reveals that half of all recorded deaths in England are classified as avoidable, underscoring the systemic opportunity for intervention. Notably, 31.4% are preventable, suggesting that upstream public health strategies, such as smoking cessation, pollution control, and dietary education, could significantly reduce mortality. Meanwhile, 18.6% are treatable, pointing to potential inefficiencies or inequities in clinical care delivery." Avoidable deaths form the primary input for risk scoring, as they signal both clinical and systemic vulnerabilities. Preventable deaths may correlate with socioeconomic indicators, IMD scores, guiding resource allocation models. Treatable deaths highlight areas for investment in service delivery, staffing, and infrastructure, whether through increased demand for acute services or long-term investment in public health

The last direct outcome measurement I analyzed specifically to identify pressure points in the healthcare system was total attendances at Accident & Emergency departments. The dataset was published by NHS England and based on monthly and annual A&E activity reports. The data covers NHS regions and measures type of attendance, major A&E, minor injury units, walk-in centers, as well as arrival mode, ambulance vs. Self-presenting, and age group, time of day, or admission outcome. Figure 10 shows the Power BI dashboard containing a bar graph showing total A&E attendance by NHS region in July 2025, as well as A&E attendances by region under vs over 4 hours from arrival to admission, transfer or discharge, and delayed admissions by regions patients waiting over 4 and 12 hours after decision to admit.

Figure 10: Regional Pressures on A&E Services – July 2025


Total A&E Attendances by NHS Region – July 2025
Total Attendances
0.4M
0.3M
0.2M
0.1M
0.0M
NHS England Midlands
NHS England London
NHS England North East And Yorkshire
NHS England South East
NHS England North West
NHS England East Of England
NHS England South West
Region
Manchester University NHS Foundation Trust has the highest attendances: 45,915
University Hospitals Birmingham NHS Foundation Trust has 38,611 attendances
Woking Walk In Centre has 2,609 attendances


A&E Attendances by Region – Under vs Over 4 Hours from arrival to admission, transfer or discharge
Sum of Total Attendances < 4 hours
Sum of Total Attendances > 4 hours
Sum of Total Attendances < 4 hours and Sum of Total Attendances > 4 hours
350K
300K
250K
200K
150K
100K
50K
0K
NHS England London
NHS England Midlands
NHS England North East And Yorkshire
NHS England South East
NHS England North West
NHS England East Of England
NHS England South West
Region
In NHS England Midlands, the attendance gap between <4hr and >4hr cases is 214,338 patients.
In NHS England South East, the attendance gap between <4hr and >4hr cases is 211,431 patients.
In NHS England North East And Yorkshire, the attendance gap between <4hr and >4hr cases is 210,311 patients.



The Northwest and Midlands regions report the highest emergency care activity, with Manchester and Birmingham trusts leading in attendances. The Northeast and Yorkshire region reports the highest breach gap, with over 21,000 patients waiting beyond the 4-hour target. The Midlands region reports the highest number of patients waiting over 4 and 12 hours after the decision to admit.

This dashboard visualizes total A&E attendances by NHS region for July 2025, highlighting variations in emergency care demand across England. It serves as a proxy for system pressure, unmet primary care needs, and population health vulnerabilities. Some regional trends include Northwest and Midlands regions show consistently higher A&E activity, and Southeast and East of England report lower volumes, possibly reflecting better primary care access or demographic differences. The Northwest and Midlands regions lead in A&E attendances, with Manchester and Birmingham trusts absorbing thousands of cases in a single month. In contrast, lower attendances in regions like the Southeast may suggest stronger primary care infrastructure, better chronic disease management, or demographic advantages. The breach of the four-hour A&E target represents more than a performance metric; it reflects underlying patient distress, workforce strain, and systemic inefficiencies within urgent care delivery. In the Northeast and Yorkshire, over 21,000 patients exceeded the recommended waiting threshold. This figure highlights not only elevated demand but also challenges in case throughput, pointing to potential resource misallocation or operational bottlenecks. Comparable patterns emerge in the Midlands and Southeast, where even moderate attendance volumes result in thousands of delayed cases. Such delays can trigger extended hospital stays, deteriorating health outcomes, and escalating costs. By quantifying breach volumes across regions, this dashboard offers a critical lens for identifying where financial exposure intersects with operational inefficiency, an essential input for predictive modeling and strategic resource planning. The Midlands region emerges as the epicenter of delay, with over 34,000 patients spending more than 4 hours in limbo after being deemed in need of admission. The Northwest and Northeast regions echo this pattern, with thousands of patients breaching the 12-hour threshold. These figures suggest that even when emergency care is delivered, the hospital system

lacks the elasticity to absorb patients promptly. NHS England Midlands, 9,531 patients spent more than 12 hours waiting for admission. NHS England Northeast and Yorkshire, 2,159 patients spent more than 12 hours waiting for admission. The next section advances this analysis by theorizing an implementation of K-means clustering using SQL-native logic, allowing for the segmentation of NHS regions based on multidimensional indicators such as A&E breach rates, deprivation scores, and healthspan metrics.

K-Means Clustering Implementation: SQL-Native

My objective for this section is to theorize implementing a scalable, SQL-native K-means clustering algorithm, inspired by Carlos Ordonez's SQL-based K-means model, that groups England's regions based on financial risk indicators, enabling data-driven regional policy insights. In my literature review Ordonez's work on "Programming the K-means Clustering Algorithm in SQL" was a pivotal and influential concept that has anchored the framework of my dissertation. Ordonez introduces two implementations of K-means in SQL in his research. The proposed implementations allow clustering large data sets in a relational DBMS eliminating the need to export or access data outside the DBMS. Only standard SQL was used; no special data mining extensions for SQL were needed. This work concentrated on defining suitable tables, indexing them and writing efficient queries for clustering purposes. The first implementation, and the one my model of a SQL-based clustering of England regions by financial risk is guided by, is a naive translation of K-means computations into SQL and serves as a framework to introduce an optimized version with superior performance. It is called Standard K-means.

Setup

Following the basic framework to implement K-means in SQL, having Y and k as input, articulated in Ordonez's research, step one is setting up a scheme, creating, indexing and populating working tables. My tables support multidimensional analysis across funding, need, and outcomes that analyze regional equality in healthcare financing based on the datasets I collected from ONS, UK Gov open data, and local authority datasets.

Table 3. This table presents a comprehensive set of indicators designed to assess regional disparities in healthcare funding, access, and outcomes across England. Each column captures a distinct dimension of equality, enabling multidimensional analysis for policy evaluation, clustering, or dashboarding.

Column Name	Description
Region Name	Name of the geographic region or ICB area.
Total ICB Allocation (£)	Total annual funding allocated to the Integrated Care Board.
Per Capita ICB Funding (£)	ICB funding divided by regional population—used to compare funding equity.
Public Health Investment (£)	Total investment in public health initiatives by local authorities.
Nominal Healthcare Spending (£)	Overall healthcare spending in nominal terms, across all schemes.

Spending by Financing Scheme	Breakdown of spending by source: NHS, private, charitable, etc.
Change in Funding (YoY %)	Year-over-year percentage change in total healthcare funding.
Capital vs Revenue Split	Ratio or percentage of capital investment vs operational (revenue) spending.
Multidimensional Poverty Index	Composite index capturing deprivation across income, housing, education, etc.
Income Poverty Rate (%)	Percentage of population living below the income poverty threshold.
Housing Deprivation Score	Index score reflecting housing quality, overcrowding, and affordability.
Education Access Index	Measure of access to quality education and attainment levels.
Health Access Score	Composite score reflecting access to healthcare services and facilities.
Unemployment Rate (%)	Percentage of working-age population currently unemployed.
A&E Attendance Rate (per 1,000)	Number of Accident & Emergency visits per 1,000 residents.
Mortality Rate (Standardized)	Age-standardized mortality rate for comparability across regions.
Healthspan Estimate (Years)	Estimated years of life spent in good health.
Preventable Mortality Rate	Deaths per 100,000 that could be avoided through timely and effective care.
Primary Care Access Score	Index measuring ease of access to GPs and primary care services.
Mental Health Service Coverage (%)	Percentage of population with access to mental health services.
Funding-to-Need Ratio	Ratio comparing allocated funding to measured regional need.
Spending Efficiency Index	Metric assessing how effectively funds translate into improved outcomes.
Health Outcome Disparity Score	Measure of inequality in health outcomes across demographic or regional lines.
Regional Equity Index	Composite score reflecting overall equity in funding, access, and outcomes.

This schema is designed to support SQL-based modelling, geospatial mapping, and narrative-driven policy analysis. It can be adapted for use in dashboards, regression models, or clustering algorithm to uncover hidden patterns in regional health equality.

Initialization

Step two in the basic framework is initialization, where each data point, region, is assigned to a starting cluster before any calculations begin. K-Means Initialization is a crucial step in the K-Means clustering algorithm. This process involves selecting the initial centroids for the clusters before the iterative optimization begins (Statisticseasily.com, n.d.). Ordonez's SQL-native approach uses random initialization for its simplicity and scalability. Randomly assigning each region to a cluster is the starting point for iterative refinement. It creates an initial partition of the data so that centroids

can be calculated in the next step. Without this, the algorithm has no basis for comparison.

```
UPDATE regional_metrics
SET cluster_id = FLOOR(RAND() * K);  -- K is the number of clusters
```

This line uses the RAND() function to generate a random float between 0 and 1, multiplies it by K (the number of clusters), and then floors it to get an integer between 0 and K–1. Each region is thus randomly assigned to one of the K clusters. To avoid procedure loops, instead of looping through rows, Ordonez uses set-based operations like UPDATE and SELECT to handle entire tables at once. In this step I also compute the centroid of each cluster using standardized indicators. In K-means clustering, a centroid is the average position of all data points in a cluster across multiple dimensions. Ordonez uses GROUP BY cluster_id and AVG() functions to compute centroids.

```
SELECT cluster_id,
        AVG(z_unemployment_rate) AS centroid_unemployment,
        AVG(z_income_poverty_rate) AS centroid_income_poverty, AVG(z_housing_deprivation_score) AS centroid_housing,
        AVG(z_education_access_index) AS centroid_education,
        AVG(z_health_access_score) AS centroid_health_access
FROM regional_metrics
GROUP BY cluster_id;
```

Ordonez introduces the idea of sufficient statistics, to improve performance, precomputed aggregates like sum and count that allow you to calculate means without scanning the full dataset each time. This is especially useful when working with large datasets.

E step

After centroids are computed, each region, data point, needs to be reassigned to the nearest cluster. This is done by calculating the Euclidean distance between the region's standardized indicators and each cluster's centroid. The following statement computes a cross join between all regions and all centroids, calculating the distance for every possible region cluster pairing.

```
INSERT INTO region_distances

SELECT r.region_name, c.cluster_id, c.cluster_label,

SQRT(

        POWER(r.z_unemployment_rate - c.centroid_unemployment, 2) +

        POWER(r.z_income_poverty_rate - c.centroid_income_poverty, 2) +

        POWER(r.z_housing_deprivation_score - c.centroid_housing, 2) +

        POWER(r.z_education_access_index - c.centroid_education, 2) +
```

```
    POWER(r.z_health_access_score - c.centroid_health_access, 2)
) AS distance
FROM regional_metrics r
JOIN centroids c ON 1=1;
```

The purpose of this SQL query is to insert new rows into the region_distances table, which will store the calculated distances between each region and every cluster centroid. The query selects the region name (r.region_name), the cluster ID (c.cluster_id), and the cluster label (c.cluster_label). The region name, such as "West Midlands," is sourced from the regional_metrics table, while the cluster ID (e.g., 0, 1, 2) and descriptive label (e.g., "High Risk") are retrieved from the centroids table. To compute the distance, the query uses the Euclidean formula: the square root of the sum of squared differences between each region's standardized indicators and the corresponding centroid values. Specifically, it calculates the squared differences for unemployment rate, income poverty rate, housing deprivation score, education access index, and health access score. Each POWER(x, 2) operation squares the difference between a region's value and the centroid's value for a given indicator, and the SQRT(...) function aggregates these into a single distance metric. This result represents how far a region is from a cluster centroid in multidimensional space. The data is pulled from the regional_metrics table and joined with the centroids table using a cross join (JOIN centroids c ON 1=1), ensuring that every region is paired with every cluster. This setup is essential for calculating all possible region-to-centroid distances. The final output of the query will include the region name, cluster ID and label, and the calculated distance. This lays the groundwork for the next analytical step: assigning each region to the nearest cluster based on the smallest computed distance.

Table 4. Sample Table: Region Distances to Cluster Centroids

Region Name	Cluster ID	Cluster Label	Distance
West Midlands	0	High Risk	0.872
West Midlands	1	Emerging Risk	1.456
West Midlands	2	Stable	2.013
Yorkshire and Humber	0	High Risk	1.234
Yorkshire and Humber	1	Emerging Risk	0.998
Yorkshire and Humber	2	Stable	1.876
Southeast	0	High Risk	2.102
Southeast	1	Emerging Risk	1.321
Southeast	2	Stable	0.654

This stage marks the analytical core of the clustering process, where meaningful groupings begin to emerge. By computing Euclidean distances across standardized indicators such as poverty, educational access, and healthcare availability, regions are systematically assigned to clusters that reflect multidimensional vulnerability profiles. The SQL-native methodology proposed by Ordonez enables this process to be executed with high efficiency and scalability, making it particularly well-suited for large, policy-relevant datasets in the public sector. Once cluster assignments is stabilized, each region was mapped to its corresponding cluster using geographic identifiers.

Cluster profiles were developed by aggregating key indicators, such as funding-to-need ratios and health outcome disparities, allowing for intuitive labelling of clusters as “High Risk,” “Emerging Risk,” or “Stable.” This interpretive layer transforms raw clustering output into policy-relevant insights, enabling targeted interventions and resource reallocation. In Southeast, the cities with the greatest variation between healthcare expenditure and resulting health outcomes are Wokingham and Woking.

Results

The regions in the UK with the highest variability in healthcare are West Midlands, Yorkshire and The Humber, and Southeast. The cities with the biggest differences between how much money is spent on healthcare and the results people get from it in the West Midlands is Birmingham and Solihull. Barnsley and York are the cities in Yorkshire and the Humber that see the greatest disparity between healthcare spending and health outcomes. From my analysis, the trends and anomalies can be detected in NHS funding distribution over the last 10 years do support my regional outliers that I identified. One of the most prominent has be unequal allocation of Public Health investments. From 2013/14 to 2021/22, Public Health investment was disproportionately distributed across local authorities. 20% of authorities receive the bulk of investment, Northern cities like Barnsley and York rank high, while southern or suburban areas like Wokingham m and Woking receive significantly less. See Figure 5. A cluster of authorities sits between £50,000–£70,000, Solihull receives nearly £100k despite not being among the most deprived areas nationally. Wokingham and Woking are among the lowest funded, Wokingham is often classified as a wealthy area, which means it receives less government funding for services like healthcare and social care. But certain neighborhoods or populations, such as people with disabilities, mental health challenges, or low-income families, still face serious unmet needs (Wokingham Borough Council, 2024). Another important anomaly I derived from my data analysis was misleading healthcare spending per person. While nominal healthcare spending per person has surged by 57.2% since 2013, the real value and impact of that growth depend on inflation, regional disparities, and outcome improvements. In 2023, the reported (nominal) healthcare spending per person was £8,731. But when you adjust for inflation, real spending is much lower, about 95% less than the nominal figure. See figure 4. One of the most significant data I’ve detected to analyze inequality in NHS funding distribution patterns is primary medical care overfunding and underfunding. Primary medical care serves as the foundational entry point to England’s healthcare system, playing a critical role in facilitating day-to-day access to NHS services (Key Health, 2024). From my data analysis I identified distinctly regional bias. See figure 6. Southern ICBs, Surrey, Sussex, and Hampshire, dominate the overfunded end, while urban and Midlands/Northern ICBs, Birmingham, Greater Manchester, and West Yorkshire appear underfunded. I gained a clear understanding of systemic regional imbalance from this data analysis.

Conclusion

In this dissertation I researched the UK healthcare system, breaking down the structure of the NHS funding, allocation, and formula. I also studied SQL-based information systems, focusing on IS advancements in the healthcare sector. My dissertation’s core concept was exploring clustering techniques and algorithms, Carlos Ordonez’s research on “Programming the K-means Clustering Algorithm in SQL” and

“Integrating K-means Clustering with a Relational DBMS using SQL” was the breakthrough when I was finding how to design a predictive information system for healthcare financial risk analysis. Power BI was the SQL-integrated BI system tool I used to analyse my data, while I didn’t use SQL in practice during my data analysis, I theorized implementing a scalable, SQL-native K-means clustering algorithm, inspired by Carlos Ordonez’s SQL-based K-means model, that groups England’s regions based on financial risk indicators, enabling data-driven regional policy insights. My aim for this dissertation was to uncover a new and unique way to analyze healthcare data using SQL but not using the traditional techniques. I believe that my theorized model has the potential to introduce and advance the way SQL is seen and used to be more than just a domain-specific language but could possibly be a clustering algorithmic programming language.

References

Abdullah, U., Sawar, M.J. & Ahmed, A., 2009. Design of a rule-based system using Structured Query Language. Proceedings of the Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp.223–228. IEEE. Available at: https://doi.org/10.1109/DASC.2009.78

Al Maruf, A., Paul, R., Imam, M.H. and Babar, Z., 2022. A Systematic Review of The Role Of SQL And Excel In Data-Driven Business Decision-Making For Aspiring Analysts. American Journal of Scholarly Research and Innovation, 1(01), pp.249-269.

Bailoni, M., 2011. Regional inequalities and political challenges for healthcare in the United Kingdom. Hérodote, 143(4), pp.162-183.

Barr, B., Bambra, C. and Whitehead, M., 2014. The impact of NHS resource allocation policy on health inequalities in England 2001-11: longitudinal ecological study. Bmj, 348.

Bond, D., 2001. Cross-regional equity in health care funding (129 KB). -2216.

Buck, D. and Dixon, A., 2013. Improving the allocation of health resources in England: How to decide who gets what. London: The King's Fund.

Chanda, D., 2024. Improving healthcare efficiency with data analytics: A case study on SQL-based patient information systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(4), p.3.

Chang, J., Peysakhovich, F., Wang, W. and Zhu, J., 2011. The UK health care system. United Kingdom, 30, p.2019.

Department of Health and Social Care. (n.d.) *Public health investment across English local authorities*. [online] Available at: https://www.gov.uk/government/collections/local-authority-revenue-expenditure-and-financing.

Evans, D., 2021. What price public health? Funding the local public health system in England post-2013. Critical Public Health, 31(4), pp.429-440.

Huang, S.H., LePendu, P., Iyer, S.V., Tai-Seale, M., Carrell, D. & Shah, N.H. (2014). Toward personalizing treatment for depression: predicting diagnosis and severity. Journal of the American Medical Informatics Association, 21(6), 1069–1075. https://doi.org/10.1136/amiajnl-2014-002733

Key Health. (2024). *The Role of Primary Care in the UK Healthcare System*

Maruf, A.A., Paul, R., Imam, M.H. & Babar, Z., 2022. A systematic review of the role of SQL and Excel in data-driven business decision-making for aspiring analysts. American Journal of Scholarly Research and Innovation, 1(1), pp.249–269. Available at: https://doi.org/10.63125/n142cg62

NHS England. (n.d.) *A&E attendances and emergency admissions*. [online] Available at: https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/.

NHS England. (n.d.) *Integrated Care Board (ICB) funding allocations*. [online] Available at: https://www.england.nhs.uk/allocations/.

OECD. (n.d.) *Health spending (indicator)*. [online] Available at: https://www.oecd.org/en/data/indicators/health-spending.html.

Office for Health Improvement and Disparities. (n.d.) *Healthspan inequities across age and region*. [online] Available at: https://fingertips.phe.org.uk/.

Office for National Statistics. (n.d.) *Multidimensional poverty profile by region* [online] Available at: https://www.ons.gov.uk/census.

Office for National Statistics. (n.d.) *Mortality statistics for England*. [online] Available at: https://www.ons.gov.uk/census.

Ordonez, C., 2004. Programming the K-means Clustering Algorithm in SQL. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), Seattle, WA, USA, 22–25 August 2004. ACM, pp. 823–828. Available at: https://dl.acm.org/doi/10.1145/1014052.1016921.

Ordonez, C., 2006. Integrating K-means Clustering with a Relational DBMS using SQL. IEEE Transactions on Knowledge and Data Engineering, 18(2), pp.188–201. Available at: https://ieeexplore.ieee.org/document/1563982.

Pendharkar, P.C. & Khurana, H. (2014). Machine learning techniques for predicting hospital length of stay in Pennsylvania federal and specialty hospitals. International Journal of Computer Science and Applications, 11(3), pp.45–56

Shouman, M., Turner, T. & Stocker, R. (2012). Integrating Naïve Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Computer Science & Information Technology (CS & IT), pp.125–137. doi:10.5121/csit.2012.2511

Statisticseasily.com.(n.d). What is: K-Means Initialization Explained. Retrieved September 2, 2025, from https://statisticseasily.com/glossario/what-is-k-means-initialization/#google_vignette.

Sohail Imran, Tariq Mahmood, Ahsan Morshed and Timos Sellis, "Big Data Analytics in Healthcare — A Systematic Literature Review and Roadmap for Practical Implementation," IEEE/CAA J. Autom. Sinica , vol. 8, no. 1, pp. 1-22, Jan. 2021. doi: 10.1109/JAS.2020.1003384 (http://dx.doi.org/10.1109/JAS.2020.1003384)

Tuya, J., Suárez-Cabal, M.J. and de la Riva, C., 2006. SQLMutation: A tool to generate mutants of SQL database queries. University of Oviedo, Spain.

Wokingham Borough Council. (2024). *Vital care for those most in need under threat from funding shortfall*. Available at: https://www.wokingham.gov.uk/news/2024/vital-care-those-most-need-under-threat-funding-shortfall

Yousefi, N., Hasankhani, F., Kiani, M. & Yousefi, N. (2020). Appointment scheduling model in healthcare using clustering algorithms. International Journal of Optimization and Control: Theories and Applications (IJOCTA), 10(3), pp.1–13. arXiv:1905.03083 [cs.SY]

Zhao, J., Hu, X. and Meng, X., 2010. ESQP: An Efficient SQL Query Processing for Cloud Data Management