Durham
University

Analysis of Opioid Analgesic Prescriptions in England; Significant Features and Identification of Local Tier Authority Residual Outliers utilising Linear Regression

by

Samuel Mason

A master’s Dissertation for the degree of

Master of Data Science - Social Analytics

September 2024

Abstract

This study explores the key factors likely to be influencing the marked raise in opioid analgesic prescriptions in the UK; a trend exemplified by a 200% increase in prescriptions from 1998 to 2018 and a corresponding rise in opioid-related mortality rates (NHS, 2022; ONS, 2023). Through a comprehensive analysis of variables identified in literature, such as demographic factors (age and gender), socioeconomic factors (income, education, and employment), healthcare accessibility variables (number of GPs, pharmacists, and other healthcare professionals), and health-related factors (cancer registrations, smoking rates, and levels of sporting activity). This research utilises Stepwise Regression to identify the most significant variables and applies Ordinary Least Squares (OLS) regression to predict baseline opioid prescription rates at the Local Tier Authority level in England.

The analysis reveals that healthcare staff, particularly GPs in training and advanced nurse practitioners, significantly influence opioid prescription rates, more so than demographic or socioeconomic factors. Residual analysis further identifies Local Tier Authorities with the greatest deviations from predicted prescription rates, pinpointing areas of over- or under-prescription, and whether significant variables are driving these disparities.

This study fills a critical gap in the literature by focusing on England's opioid prescribing patterns, rather than the to-date predominantly US orientated research. By leveraging the NHS's English Prescribing Dataset (EPD), the study provides a unique perspective on how factors such as healthcare infrastructure, staff experience, deprivation statistics, and Local Tier Authority demographics impact opioid prescribing quantities within England's publicly funded healthcare system. Moreover, it confirms statistically significant trends in opioid prescribing, offering new insights into how these factors manifest in a universal healthcare context.

In conclusion, this research not only enhances the understanding of opioid prescribing practices in England but also contributes valuable comparative data to the global literature. The findings underscore the importance of targeted interventions in specific Local Tier Authorities to effectively address disparities in opioid prescription rates, informing both national and global opioid management strategies.

Table of contents

Abstract.......... 1

1. Introduction 5

1.1 The Opioid Crisis 5

1.2 Objective 7

1.3 Research Questions 7

1.4 Literature Review 8

1.4.1 Opioid Analgesics 8

1.4.2 Acute and Chronic Pain 9

1.4.3 Opioid Addiction and Withdrawal 9

1.4.4 Contributing Factors 9

2. Chapter One: Preprocessing 11

2.1 Data 11

2.1.1 Dependant Variable 12

2.1.2 Demographic Variables 12

2.1.3 Socioeconomic Variables 13

2.1.4 Economic Variable 13

2.1.5 Health Variables 13

2.1.6 Healthcare Variables 14

2.2 Ethics 14

2.2.1 Data Privacy 15

2.2.2 Informed Consent 15

2.2.3 Potential Biases 15

2.3 Methodology 15

2.3.1 Concatenation and Mapping 16

2.3.2 Data Cleaning 18

2.3.2.1 Random Forest Regression 19

2.3.2.2 Mean Squared Error: 21

2.3.2.3 Mean Absolute Error (MAE): 21

2.3.2.4 Coefficient of Determination (R^2): 21

2.3.2.5 Euclidean Distance 22

2.3.2.6 Feedforward Neural Network 22

2.3.3 Data Scaling and Normalisation 25
2.4 Results 26
2.4.1 Concatenation and Mapping 26
2.4.2 Data Cleaning 27
2.4.2.1 Random Forest 27
2.4.2.2 Euclidean Distance 27
2.4.2.3 Feedforward Neural Network 27
3. Chapter Two: Regression 28
3.1 Methodology 28
3.1.1 Dimensionality Reduction 28
3.1.2 Outliers 29
3.1.3 Stepwise Regression 30
3.1.4 Linear Regression 32
3.2 Results 34
3.2.1 Dimensionality Reduction and Outlier Handling 34
3.2.2 Model Evaluation 35
3.2.3 Regression Results 39
4. Chapter Three: Residual Analysis 42
4.1 Methodology 42
4.1.1 Identification 43
4.1.2 K-Means Clustering 43
4.1.3 Analysis of Variance 44
4.2 Results 45
4.2.1 Outlying Residuals 45
4.2.2 Clustering 45
4.2.3 Local Tier Authorities 46
4.2.4 Variable Influence 47
5. Discussion 50
5.1 Data Availability, Reliability and Quality 50
5.2 Methodology Considerations 51
5.3 Ethics 53
5.4 Results 54
6. Conclusion 58

7. Acknowledgments 59
8. Reference List 61

Table of figures

Figure 1: Age Standardised Mortality Rates 8
Figure 2: Random Forest Regression Tree 20
Figure 3: Neural Network Architecture 24
Figure 4: Residual vs Independent Variables (Pre-Transformation) 36
Figure 5: Residual vs Independent Variables (Post-Transformation)...............................37
Figure 6: QQ Plot and Histogram of Residuals..38
Figure 7: Linear Regression, Actual vs Predicted Plot...40
Figure 8: Variable Coefficients 40
Figure 9: Residuals with Highlighted top 10% 43
Figure 10: Largest Residuals Only 45
Figure 11: Clusters of Outlying Residuals 45
Figure 12: Outlying Residual Box Plots Over/Under Prediction Data...............................48

Table of tables

Table 1: Dataset, Variable, Outcome (District) 16
Table 2: Dataset, Variable, Authority Mapping 16
Table 3: Dataset, Variable, Outcome (NHS)...17
Table 4: Feature VIF Scores 39
Table 5: Feature Coefficients 41
Table 6: Local Tier Authorities, Clusters and Prediction 46
Table 7: ANOVA test variables F-statistics..49

1. Introduction

1.1 The Opioid Crisis

Opioid prescription abuse and addiction has become a major problem in public health worldwide, especially in the United States where there is compelling evidence that prescription opioids are being given extensively to patients driven by encouragement from pharmaceutical manufacturers, with little regard for opioid dependency issues (Celentano, 2020). The UK has been displaying similar increases in opioid prescribing, which has surpassed a 200% increase in the period 1998 to 2018 and has been referred to as the opioid epidemic for the UK (NHS, 2022). There has also been a concurrent rise in the age-standardised mortality rate for deaths related to drug poisoning, which has risen every year since 2012, with slightly under half of these deaths involving an opioid (ONS, 2023). The NHS (National Health Service) receives more pressure than ever to understand how many opioid analgesics should be prescribed in order to restrict this growing epidemic.

So, what makes opioid prescribing so difficult? Chronic primary pain of visceral origin (related to the internal organs) has no clear underlying condition and is a hard factor to analyse or quantify, this makes it challenging to gauge the severity of pain experienced by a patient (NICE, 2024 : WebMD, 2023). The WHO Analgesic Ladder for managing acute, chronic non-cancer, and chronic cancer related pain, illustrates this issue with its three staged approach. The first step is for mild pain, in which non-opioid analgesics such as nonsteroidal anti-inflammatory drugs are recommended; step two is for moderate pain in which weak opioids such as hydrocodone, codeine and tramadol are recommended; step three is for severe and persistent pain where potent opioids such as morphine or methadone are recommended (Anekar, Cascella and Hendrix, 2023). Each step's identification is solely dependent on how the patient portrays the level of pain to the prescriber, illustrating the vagueness in prescribing practice for opioids. This subjective nature of pain, coupled with the addictive nature of opioids and individual prescribing practices, can lead to dependency and misuse issues for patients, influencing higher rates of opioid prescriptions. The UK government has identified prescribing rates of opioids in the UK as a concern (MHRA, 2020).

This balance of opioid prescriptions and pain relief is further complicated by the fact that there are few other pharmaceutical alternatives that effectively block pain with convenient administration (Dey and Vrooman, 2022). Nonopioid and nonpharmacologic therapies are possible alternatives to opioids analgesics for treating non-malignant chronic pain, but their use is ultimately down to patient choice in consultation with their prescriber (Dey and Vrooman, 2022). Some of these alternative treatments include exercise, physical activity, physical therapy, transcutaneous electric nerve stimulation, medications, non-steroidal anti-inflammatory drugs, acetaminophen and antidepressants (Dey and Vrooman, 2022). These alternatives may lead to patients experiencing more breakthrough pain than if they were being medicated with opioids. Thus, it is important to consider the effect of under-prescribing opioids, prescribing centres that favour these treatments may potentially leave patients in pain.

The electronic health record service (the Spine) has entrenched overprescribing in surgical practice by setting higher default prescription quantities, even though most patients need fewer than 10 tablets or can manage with non-narcotic alternatives. A study from Makary, Overton and Wang highlighted that the Spine defaults to prescribe 30 tablets when a patient may not need as many, leading to overprescription (Makary, Overton and Wang, 2017). Additionally, too many patients leave hospital with unnecessary opioid prescriptions (Makary, Overton and Wang, 2017). For instance, after a standard elective laparoscopic cholecystectomy, some doctors judiciously prescribe non-opioid alternatives or up to five opioid tablets, whilst others routinely overprescribe, providing 30-60 tablets of oxycodone, leading to potential daily intakes that significantly increase overdose risks (Makary, Overton and Wang, 2017).

However, there are two sides of the coin with opioid prescriptions, a pro-opioid argument on the side of pain advocacy for anyone with persistent pain and no other way to mitigate this pain impacting their daily lives; and the other side of the coin, a clear reality that opioid prescribing causes death, addition and dependency. This study aims to generate the missing balance by looking at the nature of opioid reliance, exploring socio environmental, psychological, physical, economical, and geographic variables in order to generate a baseline of opioid prescribing rates. The hope is that findings from this study will lead policy developers to be able to understand which features are influencing opioid prescribing, in order to develop the policies behind healthcare practice to mitigate, limit or potentially increase prescribing rates. Furthermore, Local Tier Authorities in England experiencing the largest deviation in over or under prescribing will be highlighted.

1.2 Objective

This study aims to offer a baseline rate of opioid prescriptions for Local Tier Authorities and poinpoint which authorities are over or under prescribing, influenced by significant features initially identified in relevant literature, and subsequently refined by a Stepwise Regression. The study aims to answer a range of research questions by employing a Stepwise Regression and subsequent Linear Regression, harnessing variables and allowing for a detailed residual analysis. Data for this study was collected from various reputable sources, including NHS databases, Office for National Statistics, and other trustworthy bodies in order to provide reliable results.

The analysis of these results aims to contribute to existing literature by addressing the gap in opioid prescribing in the UK, as most current literature focuses on the US opioid epidemic. These results aim to provide statistically significant evidence on variables that are positively or negatively influencing England's opioid prescription rates by incorporating the publicly available data published by the English Prescribing Dataset (EPD) as our dependent variable. With the aim of the produced coefficients describing how factors such as healthcare infrastructure, staff experience, and location disparities specifically influence opioid prescription patterns in England. This work aims to not only broaden the understanding of opioid prescribing practices in England but also contribute valuable comparative data to the global literature by offering a basis for more targeted policy interventions utilising the findings of the output baseline prescription rate per Local Tier Authority, and the Local Tier Authorities identified which are leading the front of this opioid epidemic. Potentially, this study will provide a replicable methodology which other countries can adopt in order to understand their influential variables and highlight low level authority districts which are contributing to their epidemic.

1.3 Research Questions

- What is the baseline rate for opioid prescribing in Local Tier Authorities in England?
- What sociodemographic, demographic, socioeconomic, geographic, healthcare accessibility, and health factors influence opioid prescribing in England?
- Which Local Tier Authorities in England deviate the most from predicted opioid prescribing rates, indicating potential over or under-prescribing?

1.4 Literature Review

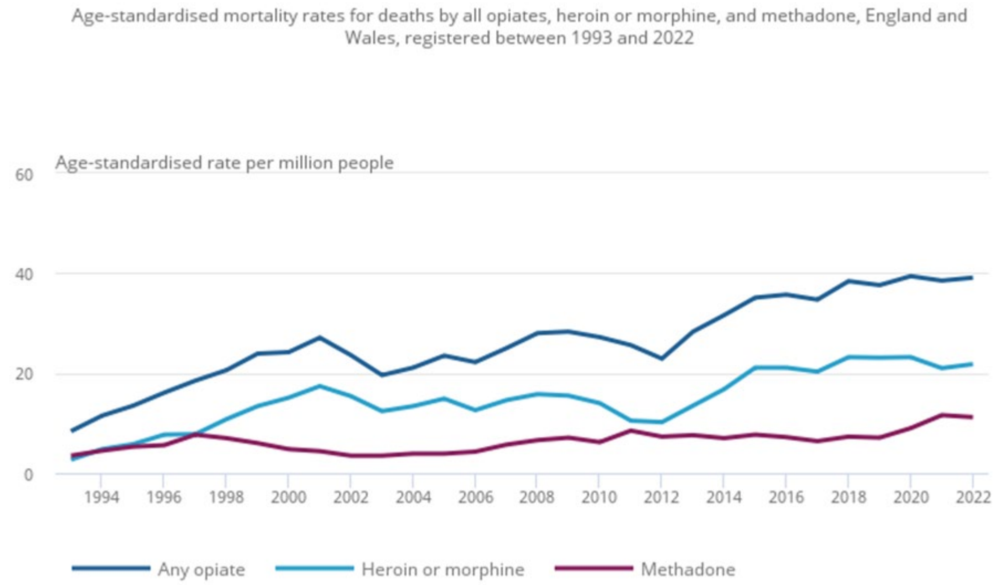

Figure 1: Age Standardised Mortality Rates

(ONS, 2021)

A clear upward trend for mortality rates by opioids from 2000 to 2012 is illustrated in '*Figure 1*', some fluctuations occurred but it was a fairly flat trend; however, in the ten-year span from 2012, mortality rates have increased by nearly 100%. 'Any Opioid' from '*Figure 1'* does not encompass only heroin and methadone, there are a range of opioid alkaloids and their semisynthetic and synthetic derivatives that are contributing to this influx of mortalities, particularly in the prescription industry in which prescribing opioids has become a norm regardless of their addictive and harmful tendencies.

1.4.1 Opioid Analgesics

Opioid analgesics are the classification of drugs for all opioid alkaloids and their semisynthetic derivatives, synthetic phenylpiperidines, and synthetic pseudo piperidines, all of which act on three major classes of receptors (Jamison and Mao, 2015). These opioids are used to relieve moderate to severe pain particularly of visceral origin; pain is classified into four categories: acute, chronic, primary or secondary (NICE, 2023). Opioids analgesics can be prescribed for any of these categories depending on the severity of the pain (NICE, 2023). The NHS state 'Opioids are very good analgesics for acute pain and pain at the end of life but there is little evidence that they are helpful for long-term pain' (NHS, 2022). Opioid analgesics are crucial for managing acute, cancer-related and end of life pain, additionally displaying significant effect in treating chronic non-cancer pain for specific populations (Wolfert et al., 2010).

1.4.2 Acute and Chronic Pain

Acute pain is of short duration and is sudden or urgent, mostly occurring after injuries or trauma, defined as pain linked with a cause that can be relieved with treatment (UPMC, 2024). Meanwhile chronic pain is pain that continues for longer than three months, either due to health conditions such as cancer, or due to no resolution through treatment (NHS inform, 2023). Chronic pain can be associated with some common conditions such as: back pain, neck pain, arthritis or joint pain, cancer pain near a tumour, testicular pain, headaches or migraines, lasting pain in scar tissue, muscle pain or neurogenic pain (Cleveland Clinic, 2021).

1.4.3 Opioid Addiction and Withdrawal

Opioid withdrawal is one of the most powerful factors in opioid dependence and addictive behaviours. Repeat exposure and escalating doses alters the patient's brain; opioid functionality decreases as the opioid receptors progressively become less responsive to the opioid stimulation (Kosten and George, 2002). This decline in functionality and increase in dosage required escalates opioid tolerance, leading to higher doses being necessary to achieve the same effect; dependence also initiates where the patient becomes susceptible to withdrawal symptoms (Kosten and George, 2002). Withdrawal symptoms stem from the changes in the locus coeruleus, as opioid molecules occupy mu (μ) receptors, suppressing release of noradrenaline, meaning when opioids are withdrawn, neurons release noradrenaline in excessive amounts, causing diarrhoea, muscle cramps, anxiety, and jitters (Kosten and George, 2002). This, coupled with the ‘high’ from opioids, causes opioids to be extremely addictive, and this gradual increase in tolerance results in patients requesting higher dosages in order to relieve their pain, building dependency. This illustrates why opioids should be administered for short term use only. This effect ratchets opioid overprescribing, and patients caught in this cycle should be referred to opioid withdrawal treatment.

1.4.4 Contributing Factors

To understand the sociodemographic, demographic, socioeconomic, geographic, healthcare accessibility, and health factors that influence opioid prescribing, a literature search highlighted key variables from these groups that would be incorporated in this study. Statistical models will be utilised to refine these identified variables, adjusting for factors that appropriately and significantly affect opioid prescription rates in England.

Most commonly, applicable literature for this study was collected and conducted in the US. Although the US differs from England, this literature identifies variables which impacted opioid prescribing for their epidemic in order to see if they have a similar relationship with opioid prescriptions in England. The difference between the US's healthcare system and England's is one of the most important features to note; the NHS provides healthcare for ordinary residents of the UK, whilst in the US healthcare is not free and often covered by insurance (GOV, 2023). This difference leads to socioeconomic factors being more pronounced in the US, where fewer people have access to healthcare and treatment, thereby reinforcing the stronger link between opioid use, crime, and deprivation compared to England.

Studies highlighted that chronic pain occurred most commonly in older people, occurring in 45%–85% of people in their later years, coupled with advancing age being the strongest risk factor for cancer, increasing exponentially in the final decades of life (Mikelyte, et al, 2020 : DePinho, 2000). Older women have also been shown to have the highest prevalence of long-term opioid use (Campbell et al, 2010). This literature provides evidence that the use of demographic data is important in our study due to the apparent trend that opioids are prescribed more frequently with increasing age and being female.

As demonstrated by Cremer et al, socioeconomic variables such as economic, housing, social environment, healthcare environment, and population characteristics all displayed significant predictor variables in morphine milligram equivalents dispensed per capita of opioids for America (Cremer et al., 2021). A further study by Martin Gulliford in Great Britain reinforced that opioid drug use was also strongly patterned by socioeconomic conditions, with income and education showing the strongest correlation to opioid use (Gulliford, 2020).

Studies by Drug Alcohol Depend illustrated that socioeconomic factors don't just contribute more to opioid prescribing but also to a significant association between at least one socioeconomic factor and overdose in the USA; opioid misuse is an important factor in the pipeline of opioid prescribing, addition, and overdosing, and will overlap with the variables that are incorporated in this study (van Draanen et al., 2020). Studies also found that smokers are 8.23 times more likely to have opioid use disorders and 2.51 times more likely to use opioids compared to nonsmokers, and opioids were prescribed more frequently to high-SES (socioeconomic status) patients than low-SES patients (Rajabi et al., 2019 : Joynt et al., 2013). Black patients also received fewer prescriptions in the US (Rajabi et al., 2019 : Joynt et al., 2013).

Giles and Malcolm highlighted that misuse is significantly associated with an increase in property crimes (Giles and Malcolm, 2021). This was prevalent among young adults, illustrating the need for targeted policy interventions to address both opioid misuse and its broader social impacts (Giles and Malcolm, 2021). In terms of geographical impact on opioid prescribing rates, in the US prescribing rates in rural areas are higher in comparison to urban areas because of closer patient-provider relationships, lower socioeconomic status, and income inequality having a higher impact (Yang et al, 2021). Additionally, in America paediatric athletes were found to seek pain relief for injuries and to be at high risk for opioid use, with 28% to 46% of high school athletes having used opioids and those participating in high school sports having a 30% higher chance of future opioid misuse (Benjamin et al., 2024).

2. Chapter One: Preprocessing

Preprocessing of data is a critical step in any data-oriented research study. Data in the real world is not clean and is often incomplete, noisy or inconsistent (Yang, 2018). This study aims to combine data from a range of different published datasets into a single, comprehensive data frame whilst ensuring the data remains contextual, representational, and accessible. This chapter will outline the data sources, ethical considerations, concatenation, cleaning, and transformations that were employed to preprocess the data in preparation for regression and residual analysis.

2.1 Data

In a perfect world, data from all the features outlined in the introduction would be available on a Local Tier Authority level and published annually. However, one of the most important features for opioid prescribing could not be sourced: quantitative data for pain. This missing variable is one of the largest limitations of this study and the hope is that quantitative data will be released for populations experiencing pain and the associated causative conditions, so that this could be added to the study at a later date. This study will utilise monthly prescription data across

England alongside other features identified to create an accurate representation of the landscape.

2021 was the year with the most accessible published datasets available across all sources. For data that did not have 2021 availability or was deemed to be inappropriate, alternative data was used from previous years, after being checked to ensure that the data did not vary enough to cause significant discrepancies, thereby maintaining the integrity and consistency of the analysis. To provide the model with the most accurate patterns, the ideal level of data would be the lowest administrative level possible. This study aimed to collect data at the Local Tier Authority level for general data sources, and at the Sub Integrated Care Board Location level (SIBC) for all healthcare-related data, ensuring detailed and precise relationships. However, this was not always possible and certain variables utilised higher administrative level data which will still provide valuable insight to the model.

2.1.1 Dependant Variable

The dataset used for the dependent variable was the English Prescribing Dataset (EPD). The EPD merges the detailed prescribing Information data from NHSBSA (National Health Service Business Services Authority), and PLP (Practice Level Prescribing) in England from NHS Digital (BSA, 2024). The dataset contains all prescriptions issued and dispensed in England to provide an accessible source of prescribing information available to the public. For this study the monthly datasets published across 2021 were concatenated to provide us with an annual dataset (English Prescribing Dataset: Jan 2021 - Dec 2021).

2.1.2 Demographic Variables

The Office of National Statistics (ONS) publishes census data estimating population density, median age, percentage broad age bands and gender for Local Tier authorities across England, Scotland and Wales. For this study the ‘’Sex’’ dataset was employed for providing Census 2021 estimates that classify residents in England and Wales by gender per Local Tier Authority (ONS, 2022). Furthermore, Census 2021 ‘Population estimates for the UK, England, Wales, Scotland, and Northern Ireland: mid-2022’ from the ONS, provided the ‘population density’ dataset for the population per square kilometre and the median age mid-2022 (ONS, 2024). Additionally, providing data for percentage distribution of mid-2022 UK population estimates by broad age bands. These variables do not vary greatly year to year, and the impact of this on the study will be minimal from using data from different time series.

2.1.3 Socioeconomic Variables

To obtain deprivation data, the source was the UK government's release of the Indices of Deprivation (IoD), which provides statistics on deprivation in small areas across England (Local Authorities and lower-layer super output areas (LSOA)), containing 7 domains of deprivation which are combined to create the Index of multiple deprivation (IMD) (MHCLG, 2019). The 7 domains and their weightings consist of: income (22.5%), employment (22.5%), health deprivation and disability (13.5%), education and skills training (13.5%), crime (9.3%), barriers to housing and services (9.3%), and living environment (9.3%) (MHCLG, 2019). The IoD provides the average rank and score for Local Authority districts for each subdomain and the IMD. The average score is utilised for this study and is calculated by averaging the LSOA scores for each larger area with population weighting, providing the relevant measurement for our study (MHCLG, 2019). Furthermore, the most recent release of the IoD was 2019, and as these variables do not vary dramatically year to year, this dataset was employed, allowing us to still capture socioeconomic variables (MHCLG, 2019).

2.1.4 Economic Variable

Economic data was captured by regional gross domestic product (GDP) for Local Tier Authority districts; the ONS provides annual releases of GDP from 1998 to 2018-2022 for Local Authority districts in pounds, millions (ONS, 2024). 2018 data was used for this study because between April and June 2020, GDP fell by a record 19.4%, as this was during the height of the first national lockdown (ONS, 2021). To avoid this skew in data, the 2018 published dataset provides a pre-COVID representation of GDP, ensuring that COVID-19 implications are not considered (ONS, 2024).

2.1.5 Health Variables

Cancer, sport and smoking datasets were collected from NHS Digital, Sport England and the ONS. The cancer registrations statistics report counts newly diagnosed cancers registered in England for the year 2021, with the diagnostic labelled as an international classification of disease 10th revision 3-digit code (NHS Digital, 2023). Cancer data was only obtainable at regional and sub-Integrated care board geography; regional data was incorporated for this study as sub integrated care boards overlap with Local Tier authorities. Furthermore, the ONS publishes ‘The proportion of current smokers by Local Authority of the UK’, which consists of estimated percentage of smokers above the age of 18 by Local Tier Authorities’ populations from 2015 to 2021 (ONS, 2022). Finally, our sports health variable comes from Sport England; May 2018 - May 2019, with the data being obtained from surveys within this period (Sport England, 2023). The rationale behind selecting data from this period is to ensure that the rates are not affected by covid lockdowns or respiratory issues caused by covid. However, the dataset May 2019 - May 2020 was used accidentally, leading to a limitation that Covid could still produce some alteration in the data, although daily exercise was encouraged throughout lockdowns. The dataset contains percentages of active (150+ minutes), fairly active (30-149 minutes) or inactive (<30 minutes) levels of weekly activity by Local Tier Authority (Sport England, 2023).

2.1.6 Healthcare Variables

To provide numerical representation of healthcare availability, the ‘General Practice Workforce’ and ‘Patients Registered at a GP Practice’ datasets from NHS Digital were utilised (NHS Digital, 2024). The patients registered dataset provided the number of patients registered at a GP practice on the first day of each month, whilst the general practice workforce documented full-time equivalent (FTE) headcount figures by gender, role, age band, and work commitment of England's primary care general practice workforce (NHS Digital, 2024). For the number of patients dataset, April 2023 was employed as post April 2023 sub integrated and integrated care board locations (SICB (Sub Integrated Care Board), ICB (Integrated Care Board)) changed mapping codes. To ensure that mapping stayed consistent April 2023 was also utilised for the general practice workforce dataset for these same reasons. These variables’ data do not vary dramatically over the year, so for this study, April was used to represent the annual data.

2.2 Ethics

This study has taken several ethical considerations into account in order to ensure that the data is used in a responsible and respectful manner throughout the research process. The primary ethical concerns addressed in this study include data privacy, informed consent, data accuracy, and potential biases.

2.2.1 Data Privacy

The datasets employed in this study are sourced from publicly available repositories provided by reputable organisations such as NHS Digital, the Office for National Statistics (ONS), and the UK Government. These datasets do not include individual-level data, ensuring that personal privacy is maintained from the outset. For instance, the EPD ensures patient anonymity by omitting personally identifiable information. The credibility of these sources means they are required to adhere to stringent data collection and privacy laws, including the General Data Protection Regulation (GDPR) and compliance with the Information Commissioner's Office (ICO) guidelines.

2.2.2 Informed Consent

As this study relies on secondary data sources, the issue of informed consent is primarily the responsibility of the original data collectors. Organisations such as NHS Digital and the ONS have protocols in place to ensure that data is collected ethically, with appropriate consent obtained from participants where necessary.

2.2.3 Potential Biases

Recognising and mitigating biases is crucial for ethical research. This study acknowledges the limitations and potential biases in the datasets used, such as the reliance on regional Cancer data, GDP data from 2018 to avoid the skew of COVID-19 impacts, datasets from years other than 2021, and data which may not pass confidence intervals representing statistically insignificant representations. By transparently discussing these limitations the study aims to provide a balanced interpretation of the findings, ensuring that conclusions are drawn with appropriate context.

2.3 Methodology

The methodology section of this chapter provides an overview of the processes and techniques employed to prepare and preprocess our data. It covers the steps involved in concatenating datasets, mapping variables, cleaning data, and scaling variables to ensure comparability. This systematic approach is crucial for ensuring that the data is robust, reliable, and suitable for subsequent analysis in order to deliver effective and meaningful results.

2.3.1 Concatenation and Mapping

To concatenate and map our data, variables were extracted from various datasets and placed into a mapping data frame; our mapping data frame consists of district and healthcare authority identifiers. These identifiers are used to merge variable data at correct district or healthcare levels, *'Table 1'* highlights the methodology for creating our mapping data frame.

Table 1: Dataset, Variable, Outcome (District)

Dataset	Variable(s)	Outcome
'Lower Tier to Upper Tier Authority' (ONS, 2024 : Open Geography Portal, 2022).	Lower Tier and Upper Tier Code Identifiers and Names.	Upper Tier Code -> Upper Tier Name -> Local Tier Code -> Local Tier Name
'Local Tier Authority to regional location' (ONS, 2024 : Open Geography Portal, 2022).	Regional Code and Regional Name.	Region Code -> Region Name -> Upper Tier Code -> Upper Tier Name -> Local Tier Code -> Local Tier Name

Only English locations were incorporated by filtering Local Tiers Codes to exclude any not starting with 'E' ('E' represents England). In order to build on our mapping data frame, variables are added from respective datasets at their level of authority. *'Table 2'* highlights: the dataset, variable(s) included, and the level of authority to be merged upon. Each dataset and variable are covered in detail in the 'Data' section earlier in chapter one. Any variables labelled as 'Pivoted' means that the variable was changed into a longitudinal format to match the rest of the data.

Table 2: Dataset, Variable, Authority Mapping

Dataset	Variable(s)	Level of District Authority Mapping
'Population estimates for the UK, England, Wales, Scotland, and Northern Ireland: mid-2022' (ONS, 2024).	Population density per square kilometre, median age mid-2022.	Local Tier Authority
'Percentage distribution of mid-2022 UK population estimates, by broad age bands and local authority' (ONS, 2024).	Broad age bands: Ages 0 to 15, Ages 16 to 64, Ages 65+, Ages 85+.	Local Tier Authority
'Sex' (ONS, 2022).	Sex (2 categories), Observation. (Pivoted on Local Tier Code).	Local Tier Authority
'Cancer Registration Statistics, England 2021' (Table 1) (NHS Digital, 2023).	ICD10 code, Site description, Geography code, Gender. (Pivoted on Site Description).	Regional
'The English Indices of Deprivation 2019 (IoD2019)' (File 10) (Sheets: 'IMD', 'Income', 'Employment', 'Education', 'Health', 'Crime', 'Barriers', 'Living').	IMD - Average score, Income - Average score, Employment - Average score, Education, Skills and Training - Average score, Health Deprivation and Disability - Average score, Crime - Average	Local Tier Authority

	score, Barriers to Housing and Services - Average score, Living Environment - Average score.	
‘The proportion of current smokers by Local Authority of the UK’ (ONS, 2022).	Estimated proportion of current smokers.	Local Tier Authority
‘Regional gross domestic product: local authorities’ (ONS, 2024) (Table 5).	Gross Domestic Product (GDP) at current market prices, pounds million: 2018.	Local Tier Authority
‘Active Lives data tables’ (May 2019 - May 2020) (Sport England, 2023).	Active Rate, Fairly Active Rate, Inactive Rate.	Local Tier Authority
N/A	Population totals derived from summed totals of male and female observations.	Local Tier Authority

To concatenate and map our data for NHS variables, the same process was applied for mapping and variable extraction. However, NHS mapping requires Sub Integrated Care Board (SICB), Integrated Care Board (ICB) locations, and SIBC code hierarchy identifiers for identification and merging. SIBC locations are the lowest level available and overlap with Local Tier authorities, thus healthcare identifiers will be mapped onto a Local Tier Authority level. *‘Table 3’* highlights the methodology for incorporating NHS identifiers to our mapping data frame:

Table 3: Dataset, Variable, Outcome (NHS)

Dataset	Variable(s)	Outcome
‘LSOA (2021) to SICBL to ICB to LAD (April 2023) Lookup in EN’ (LSOA, 2021).	SIBC Code, SIBC Hierarchy Code, ICB Code.	Local Tier Authority -> ICB Code -> SIBC Code -> SIBC Hierarchy Code

Manipulations were required for the NHS data before merging to achieve variable totals at a SICB level. To apply the ‘Patients Registered at a GP Practice’ dataset from NHS Digital, the SICB codes and number of patients were extracted, grouped by SICB, summed, and merged on SICB code to our data frame (NHS Digital, 2024). For the ‘General Practice Workforce, England, Bulletin Tables, September 2015 - April 2023’ dataset, worksheet ‘6b’, the headcounts for full time equivalents by SIBC code hierarchy, the following staff were unable to prescribe opioids and were removed: apprentices, therapists, physiotherapists, phlebotomists, general practice assistants, dispensers, dieticians, care coordinators, healthcare assistants, physician associates, podiatrists, advanced therapist practitioners, advanced podiatrist practitioners, advanced physiotherapist practitioners, nurse specialists, extended role practice nurses, practice nurses, trainee nurses, nursing partners, other nurses, all direct patient care and all nurses (NHS Digital, 2024). After removal, remaining columns of relevant staff data were merged onto our mapping data frame at an SIBC level. Utilising these staff counts, for each SIBC hierarchy code, the columns were summed and a new column for ‘Total Prescribers’ was added containing each row’s totals.

The EPD January - December 2021 datasets were concatenated and filtered for BNF codes starting with ‘040702’ (opioid analgesic BNF codes start with ‘040702’) (British National Formulary) (BSA, 2024 : OpenPrescribing, 2024). All residual columns except the postcode and total quantity were dropped, followed by grouping postcodes and summing. Utilising the ‘Patients Registered at a GP Practice’ dataset, the grouped postcodes and total prescription quantities were matched to SICB codes, grouped, and summed; these total opioid analgesic prescriptions per SICB location were then merged onto our mapping data frame.

2.3.2 Data Cleaning

Data cleaning is a critical step in the preprocessing phase, ensuring that the dataset is accurate, consistent and ready for analysis. Here, we address the methodology behind filling missing data for region names and codes, population, GDP, and other key variables.

In the concatenation process, 14 Local Tier authorities failed to match with regional or Upper Tier Authority codes due to being ‘Inactive’; these Local Authorities were missing codes and all regional and Local Tier level data. However, it was vital to fill these data points rather than excluding rows with Null values to retain the relationship to the dependent variable. These ‘Inactive’ Local Authorities did contain SIBC level data and their Local Tier code identifiers, so in order to fill the missing data points, Local Tier codes were utilised and inputted into resources such as 'Find that postcode' and the ONS to acquire their respective region and Upper Tier Authority (FindThatPostcode, 2024 : ONS, 2024).

Populating Local Tier level data for these missing rows required available data to be re-entered. With lookup codes in place, the missing age distribution, population density, median age, GDP and cancer data was inputted into the missing rows. However, Gender observations in ‘Inactive’ Local Tier authorities was not available in original datasets. To circumvent this, the ‘Population estimates mid 2022’ dataset from the UK Government for Local Tier districts was employed, unfortunately this dataset has since been made inaccessible. The estimated population totals from this dataset were then split according to Gender using the 2021 Census, which indicated that women and girls constituted 51% of the population of England and Wales, while men and boys made up 49% (ONS, 2023).

Deprivation, smoking and sport activity were also missing variables in the Local Tier Authority level data. One of the missing rows 'Isles of Scilly' had all its data available in the original datasets, allowing this row to be filled with its corresponding data.

2.3.2.1 Random Forest Regression

At a Local Tier level, deprivation, smoking and sport activity each had 13 rows with missing data, meaning estimates had to be created to ensure any meaningful patterns could still be observed. A Random Forest Regression was chosen to predict the missing data points because our data is continuous, and this model effectively captures non-linear relationships between the dependent and independent variables (AnalytixLabs, 2023). Additionally, Random Forest Regression is less prone to overfitting, which makes it well-suited for handling high dimensional data like ours (AnalytixLabs, 2023).

Random Forest Regression is a supervised learning algorithm that uses an ensemble of decision trees for regression tasks, combining the predictions from multiple trees to make more accurate predictions than a single model (Bakshi, 2020). A decision tree consists of a root node, branches, internal nodes and leaf nodes; where the root node represents the entire dataset, branches represent decisions or tests on features, internal nodes represent decision points, and leaf nodes represent the final output or prediction (IBM, 2023). The Random Forest works by constructing several decision trees during training and outputs the average of their predictions. The process involves (Bakshi, 2020):

1. Randomly selecting k data points from the training set.
2. Building a decision tree based on these k points.
3. Repeating steps 1 and 2 to create N trees.
4. For a new data point, averaging the predictions from all N trees to get the final result.

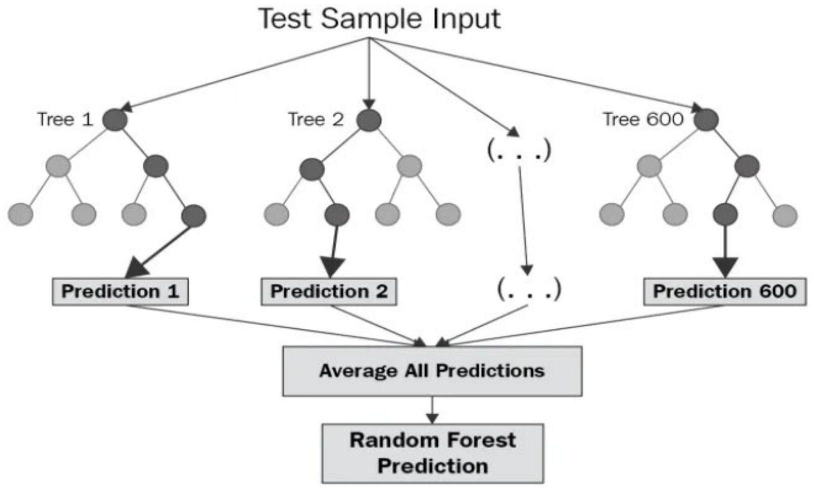

Figure 2: Random Forest Regression Tree

(Bakshi, 2020)

The model is built using training and testing data, where it learns patterns from the training data and then makes predictions on the unseen testing data. The predicted results are then compared against the actual values to assess the model's performance. The model generates subsets of the training data, known as bootstrap samples which have the same number of data points as the original dataset but with some duplicates. For each bootstrap sample, a decision tree is constructed by the model, growing by recursively splitting the data at each node. The trained Random Forest Regression model then predicts the target values for the validation sets of the dependant variable. Each decision tree in the forest makes a prediction for each data point, and the final prediction is the average of all the tree predictions (Bakshi, 2020).

Our Random Forest Regression iterated each column with missing data for our deprivation, smoking and sport variables. For each iteration, complete and incomplete rows of our geographic data were separated; complete rows counted as our training (80% of data) and testing (20% of data) data with a random state of 42. The model was then used to predict the incomplete rows of each column per iteration, allowing us to focus on one dependant variable for prediction at a time.

To assess the performance of our Random Forest Regression, three performance metrics are utilised that will be used throughout this study: Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2).

2.3.2.2 Mean Squared Error:

MSE in regression represents the average squared residual and measures the amount of error in the average squared difference between observed and predicted values; the lower the MSE, the less error found in the model (Frost, 2021). The statistical formula for MSE is:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

(Frost, 2021)

- y_i is the i^{th} observed value.
- $\hat{y}_i$ is the corresponding predicted value.
- n = the number of observations.
- ∑ is the summation symbol, indicating summation of all squared observed minus predicted values.

2.3.2.3 Mean Absolute Error (MAE):

MAE calculates the average magnitude of the errors utilising the absolute difference between predicted and actual values, to measure the accuracy of a regression model (Schneider, 2022). The goal of MAE is to evaluate the quality of predictions with their absolute magnitude rather than relative magnitude (Arize AI, 2023). The lower the MAE, the better the model's performance. The statistical formula for MSE is:

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

(Sharma, 2022)

- $\hat{y}_i$ is the predicted value for the i^{th} observation.
- y_i is the actual (true) value for the i^{th} observation.
- n is the total number of observations.
- ∑ is the summation symbol, indicating that you sum over all observations from 1 to n.

2.3.2.4 Coefficient of Determination (R^2):

Coefficient of determination (R^2) produces a number between 0 and 1 which indicates how well the independent variables of our model explain the variation in the dependent variable, 1 indicates a perfect fit of data, whilst 0 represents data that does not fit at all (Fernando, 2024). The statistical formula for the coefficient of determination is:

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

(Kumar, 2022)

- ȳ is the mean of the actual values.
- $\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ represents the total sum of squares (TSS) which measures the total variance in the actual data.
- $\sum_{i=1}^{n} (y_i - \bar{y}_i)^2$ represents the residual sum of squares (RSS) which measures the variance in the data that the model does not explain.

2.3.2.5 Euclidean Distance

For the number of patients registered at a GP, 18 data points were missing from our data frame. Utilising deprivation, population, and our other geographical level data, the Euclidean distance was employed to match any incomplete rows to the most similar complete row. Euclidean distance helped to ensure that imputed approximations were closely aligned with similar observations based on our relevant variables, thereby maintaining the integrity and distribution of our dataset.

The Euclidean distance works by measuring the distance between two points in an n-dimensional plane called Euclidean space (GeeksforGeeks, 2024). In one dimensional space, the Euclidean distance between two cells would be the simple arithmetic difference, in 2-D space Pythagoras Theorem calculates the distance, whilst in N-Dimensional space the number of variables (‘n’) represents the dimensions and the distance is calculated with the following statistical formula (Stanford.edu, 2024 : Stanford.edu, 2021):

$$D_{ij}^2 = \sum_{v=1}^{n} (X_{vi} - X_{vj})^2$$

(Stanford.edu, 2024 : Stanford.edu, 2021)

- Where D_{ij} is the distance ('D') between data points i and j.
- Where ‘X’ represents each variable in data points i and j.

2.3.2.6 Feedforward Neural Network

Our dependent variable (opioid analgesic prescriptions) contained 18 missing data points. The method deemed most applicable for estimating missing opioid prescription counts was a non-linear, deep feedforward neural network utilising Rectified Linear Unit (ReLU) activation functions to capture non-linear patterns in our data.

A feed forward neural network feeds the information in a forward direction, through input nodes and hidden nodes, to an output node(s); there are no cycles or loops between nodes (respected nodes are part of input, hidden and output layers) (DeepAI, 2019). The input layer consists of neurons that receive data and pass them to the hidden layers, with the neurons in the input layer being determined by the dimensionality of the data (DeepAI, 2019). The hidden layers are the computational engine of the neural network, the neurons of each hidden layer take the weighted sum of the output from the previous layer, apply an activation function (ReLU in our case) and pass the resulting output to the next layer (DeepAI, 2019). The ReLU activation function is:

$$f(X) = max(0, X)$$

(Dremio, 2024)

This function introduces non-linearity, allowing the model to learn complex patterns and relationships and overcome the vanishing gradient problem (Dremio, 2024). It works by applying the max function to the inputted value; if the value is greater than or equal to zero, it returns the input value, whilst if the input is negative it returns zero (Dremio, 2024). Finally, the last hidden layer passes the weighted sum to the output layer which produces output based on the number of neurons in the output layer; the number of neurons for a regression task is one, the predicted value (Dremio, 2024).

To ensure we had complete data to train the model, any rows with missing prescription data were dropped, as well as any non-numeric, or location identifying variables. In order to optimise predictions and their accuracy, we scaled and normalised our features (see ‘Data Scaling and Normalisation’ below). The data frame was then split into X and Y training, and X and Y testing data using a random seed of 42 and 20% of the data for testing (X = Independent Variables, Y = Dependent Variable). Our dense input layer consisted of 128 neurons and the dimensionality was equal to the shape of our X training data (geographic and SIBC level data). The hidden layers consisted of ten dense layers with ReLU activation functions with four layers containing 64 neurons, and six layers containing 32 neurons, and since this was a regression our output layer was a dense layer containing 1 neuron (our predicted value) (*‘Figure 3’*).

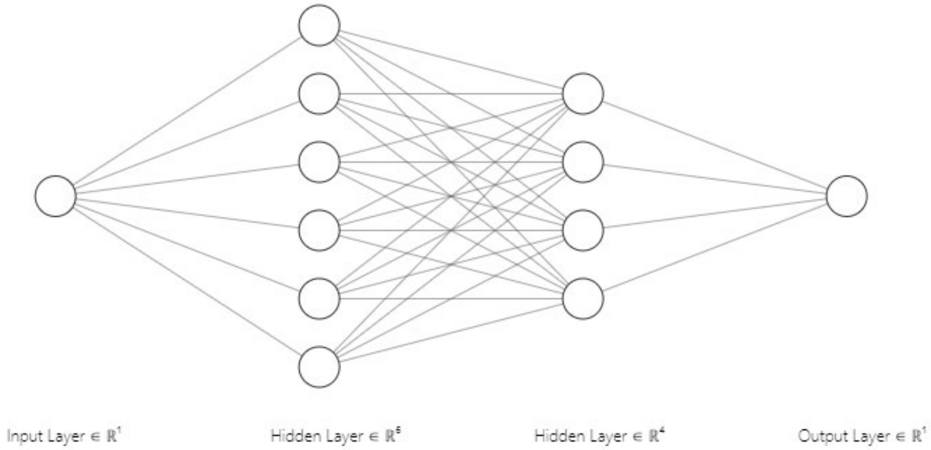

Figure 3: Neural Network Architecture

(Alexlenail.me, 2024)

The model trained on the training data and validated on the X and Y testing data. This meant that the model could predict the dependent variable (Y) from what it had learnt from the training data to predict values for unseen data and calculate the MAE. An Adam optimiser with a learning rate of 0.01 was employed, this learning rate was the rate in which the model overwrote old information with new information in the learning process (ScienceDirect, 2024). The loss metric was set to MAE to focus the model on improving predicted values by reducing the absolute difference between actual and predicted values. With an 'EarlyStopping' callback, the model terminates when the monitored metric (validation loss) had stopped improving; with a patience of twenty, the model waits for twenty iterations without improvement before calling 'EarlyStopping' (TensorFlow, 2024). 'RestoreBestWeights' was also enabled, meaning after termination, the iteration with the lowest validation loss would be used rather than the final iteration (TensorFlow, 2024). These features helped to combat any overfitting by preventing the model from being trained on seen data too many times. Additionally, the 'ReduceLROnPlateau' callback reduced the learning rate by a factor of 0.5 when validation loss had stopped improving, meaning for each reduction the learning rate would be multiplied by 0.5, to a minimum of 1e-5; coupled with a set patience of ten, the model would wait ten iterations of no improvement before callback (TensorFlow, 2024). The model was then compiled and run for 400 epochs with a batch size of 64.

2.3.3 Data Scaling and Normalisation

Data scaling and normalisation is essential for preprocessing the data to ensure consistent and meaningful analysis. In our context, scaling involved dividing all quantity-related data points by their respective Local Tier Authority population to achieve a uniform scale (per capita). This approach scales the data, making it easier to compare across different populations. Additionally, normalisation was performed to create a normalised range of data between 0 to 1. This process adjusts the values based on the minimum and maximum values for each feature, bringing together features with different units and magnitudes to a consistent scale. This section will describe the normalisation and scaling techniques applied to our dataset.

Before normalisation our quantity related data points were scaled by Local Authority populations to be per-person in order to remove any potential bias from varying population sizes. This simple process divided the following variables: dependent variable, paramedics, pharmacists, advanced pharmacist practitioners, nurse dispensers, advanced nurse practitioners, GP regular locums, advanced nurse practitioners, GP regular locums, GP retainers, GPs in training grade, salaried GPs, GP partners, all GPs, total opioid prescribers, cancer data, and GDP Pounds, millions by the total Local Tier population for each row elementwise. Once complete, population totals were dropped as the data was now scaled by population. This produced a scaled dataset with quantity variables scaled per capita on the Local Tier Authority level, providing an unbiased dataset regardless of population.

In order to prepare the dataset for linear regression, in the now complete data frame, any names of locations, non-numerical or identifying codes (other than Local Tier Authority as this is the lowest level we are analysing) were dropped; all data was now compiled to each Lower Tier Authority code. To transform the Lower Tier Codes into numerical representation, the Local Tier Code column was converted into dummy variables, pivoting the data so that each Local Tier Code has its own column, and each row represents a Logical Boolean variable of 1 if the row is in Local Tier Code, or 0 if not, providing a numerical identifier for each Local Tier Code to extract for outlying residuals in the residual analysis.

With all data in a numerical format, each variable's minimum and maximum values were obtained for numerical features across the dataset (apart from Local Tier Codes or our dependent variable). The minimum value of each feature was transformed into a 0 and the maximum value was transformed into a 1, every other value was transformed into a float between 0 and 1 according to its hierarchy (CodeAcademy, 2024). The following statistical formula is how this is achieved:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

(Abhinav Bandaru, 2022)

- x = The value to be Normalised.
- x' = The Normalised value of X.
- Min(x) = The Minimum value of the variable which X belongs to.
- Max(x) = The Maximum value of the variable which X belongs to.

2.4 Results

2.4.1 Concatenation and Mapping

The resulting mapping data frame incorporated data from various sources at different authority levels across England. At a Local Tier Authority level, the data frame included:

- Population Data: Population density per square kilometre, median age as of mid-2022, age distribution (ages 0-15, 16-64, 65+, and 85+), male and female population counts, and total population.
- Health and Lifestyle Data: Percentage of smokers, and sport activity rates (active, fairly active, and inactive).
- Economic Data: Gross Domestic Product (GDP) at current market prices (pounds million for 2018).
- Socioeconomic Data: Average scores for various domains, including overall Index of Multiple Deprivation (IMD), income, employment, education, health, crime, barriers to housing, and living environment.

At a regional level the data frame included:

- Cancer Data: Cancer types, ICD-10 codes, and cancer registrations by gender.

At the Sub-Integrated Care Board (SICB) Level, the data frame included:

- Healthcare Data: The number of patients registered at GP practices, headcounts for full-time equivalent NHS workforce with opioid analgesics prescribing rights, total prescribers, and total opioid analgesic prescriptions.

This resulting data frame provided a robust foundation for our analysis, presenting important potential influential variables at their respective levels.

2.4.2 Data Cleaning

2.4.2.1 Random Forest

The averaged MSE across all missing variables from our Random Forest model equated to 6.19; the average standard deviation across all predicted variables was 3.07 meaning that the MSE is approximately 2 times the average standard deviation of the predicted variables. This indicates that the prediction errors are about twice the average variability of the predictions, suggesting that while the model's performance is reasonable, there is still a notable level of prediction error relative to the inherent variability in the data. Therefore, while the MSE level is acceptable for proceeding with estimates to fill in missing data points, it should be used with caution due to this notable level of prediction variability.

The averaged MAE for our Random Forest model for all predicted variables equated to 1.18, indicating that, on average, the predictions deviated from the actual values by approximately 1.18. The average range of our deprivation, smoking and sport rate data was 16, meaning in the context of our data, these estimates can be used as approximations without varying enough to cause significant impact as predictions varied by under 10% of our averaged range. The averaged R^2 for our Random Forest model across all predicted dependent variables was 0.79, indicating that the model explains 79% of the variance in the predicted data. This suggests that the model's estimates are appropriate and fit the data well.

2.4.2.2 Euclidean Distance

For the 18 most similar data points fulfilled by the Euclidean distance, the average Euclidean distance was 3,854 (this measures the dissimilarity between imputed and actual rows). This average distance is relatively small in comparison to the overall range of the dependent variable of 2,717,235, indicating that the imputed values were reasonably close to the existing data points and provides evidence that the estimates are reasonably effective and introduce minimal distortion relative to the natural variability of the dependent variable.

2.4.2.3 Feedforward Neural Network

Our feed forward neural networks model's evaluation was conducted using the MAE metric. The resulting validation MAE was 103.70, which equates to a difference of less than 1% of the scaled range of our dependent variable (scaled dependent variable range: 11865.16, and unscaled dependent variable range: 88256470). This suggests that the model was effective and accurate in estimating the missing data points.

Overall, chapter one has produced a dataset that has no Null values, is all numerical, and is scaled and normalised. The next chapter will highlight the dimensionality reduction process, identification and handling of outliers, and the regression's procedures. By addressing these aspects, we aim to improve the model's accuracy and reliability, ensuring that our final analysis is based on the most pertinent and clean data possible before it is inputted into our Linear Regression.

3. Chapter Two: Regression

This chapter outlines the methodology and results of our Stepwise and Linear Regressions, alongside dimensionality reductions and outlier handling. The Stepwise Regression in this section identifies the most influential variables in our dataset with respect to our dependent variable. Subsequently, the Linear Regression employs these influential variables to determine the baseline rate of opioid analgesic prescriptions (regression line of fit) and each feature's relationship in the form of a positive or negative coefficient.

3.1 Methodology

3.1.1 Dimensionality Reduction

Dimensionality reduction was a critical first step in order to address the issue of overfitting. Overfitting means our regression model learns the noise in our training data and redundant variables instead of vital underlying patterns (IBM, 2024). The aim of dimensionality reduction is to reduce the number of features whilst still capturing important information; this section will outline the methodology applied to reduce the dimensionality and minimise overfitting (IBM, 2024).

Variables which present a near perfect correlation coefficient when analysed for correlation indicate that the model is being fed nearly identical data. To counteract this, our study employed a correlation matrix utilising Pearson's correlation coefficient to identify pairs of features that exhibited high correlation, comparing attributes and calculating a score (correlation coefficient) ranging from -1 to +1. A score close to positive one represents linear similarity and a score close to 0 represents almost no linear similarity (Nettleton, 2014). The correlation coefficient (‘r’) is calculated with the following statistical formula (Pearson's correlation coefficient formula):

$$r = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

(Turney, 2022)

- Where ‘r’ is Pearson's correlation coefficient.
- Where ‘n’ is the number of data points (x, y pairs).
- Where ‘x’ is the value of variable one.
- Where ‘y’ is the value of variable two.

Utilising Pearson’s correlation coefficient, a correlation matrix was calculated for all variables (other than the dependent variable and Local Tier Codes) to calculate the correlation coefficient for every independent variable against each other. The threshold for correlation coefficients was set for 0.95 to highlight any features with a near perfect linear correlation.

3.1.2 Outliers

Points that significantly differ from the majority of data points are referred to as outliers. Outliers are objectively anomalous and can occur from variability in the data, measurement errors, or other factors (Neural Data Science, 2023). This anomalous data can cause issues as it has disproportionate leverage on statistical properties such as the mean and variance of the data and can violate their assumptions; removing outliers can improve the statistical significance of results (Neural Data Science, 2023 : Frost, 2019). Outliers can be identified through a range of methods; for this study, the chosen method employed was the standard score (Z-Score) which measures how many standard deviations above or below the population mean a data point is, when the sample size is larger than thirty (Glen, 2023). The statistical formula for the Z-Score is:

$$Z = \frac{(x - \mu)}{\sigma}$$

(National Library of Medicine, 2024)

To identify outliers, we calculated Z-scores for each variable, setting the threshold to three standard deviations from the mean. Using the empirical rule that meant that roughly 99.7% of the data fell within three standard deviations, and any data points over the threshold (positively or negatively) were extreme outliers (Frost, 2021).

3.1.3 Stepwise Regression

The "Curse of Dimensionality" is the problem of additional variables in regression causing the data space to expand exponentially (DataCamp, 2023). This exponential increase in space causes data sparsity (where most of the high-dimensional space is empty), increased computation, overfitting, and a negative impact on model performance (DataCamp, 2023). With our data frame containing 34 variables, fitting a linear regression now would likely exacerbate these issues making it difficult for our model to perform satisfactorily. To address this issue, a Stepwise Regression was employed to extract the most statistically significant variables from our data frame for regression.

A Stepwise Regression iteratively examines the statistical significance of independent variables using Ordinary Least Squares (OLS) regressions (see ‘Linear Regression’ below for OLS regression explanation), selecting the most significant variables (Hayes, 2022). It starts with an empty model working with forward and backward steps. The forward steps add the most significant variable per iteration based on a variable's coefficient falling under a p-value threshold (Hayes, 2022). The p-value measures the probability that any observed difference is due to chance, whilst the coefficient is the mathematical relationship between the independent and dependent variable (Dahiru, 2008 : Frost, 2017). Since a p-value is a probability, it has a range between 0 and 1; the lower the value the less probable that the result was due to chance, determining how statistically significant a result is (Dahiru, 2008). To test the regression coefficients a two-tailed hypothesis test is conducted to obtain the p-value for the level of significance (Watts, 2022). The null hypothesis states that there is no linear relationship between the independent and dependent variable, whilst the alternative hypothesis states that there is a significant linear relationship (Watts, 2022).

$$H_0 : \beta_i = 0 | H_1 : \beta_i \neq 0$$

(Watts, 2022)

The standard error (SE) then measures how accurately the model estimates the coefficient (β), a smaller standard error describes less error in the estimate (Minitab, 2024). The statistical formula for the SE is:

$$\sqrt{\frac{\sum(\hat{y}-y)^2/(n-k)}{\sum(x_i-\bar{x_i})^2}}$$

(Watts, 2022)

- Where $\sum(\hat{y}-y)^2$ is the Residual Sum of Squares (RSS).
- Where *n* - *k* is the degrees of freedom, *n* is the sample size and *k* are the number of predictors.
- Where $\sum(x_i-x_i)^2$ is the sum of squares of the differences between the independent variable and its mean.

For each coefficient (β_i) the t-test is computed to measure how far the estimated coefficient is from zero (Zach, 2021) using the statistical formula:

$$t=\frac{\beta_i}{SE(\beta_i)}$$

(Zach, 2021)

The p-value is then derived from the t-statistic by comparing it to the t-distribution with its respective degrees of freedom (Omnicalculator, 2024). For our two tailed test, the following statistical formula calculates the p-value:

$$p=2\times P(T>|t|)$$

(Omnicalculator, 2024)

- Where $P(T>|t|)$ represents the probability that the t-statistic is as extreme or more extreme than the observed value under the null hypothesis.

With the derived p-value we can now reject or fail to reject the null hypothesis. If the p-value is less than the significance level of our threshold (0.05), we reject the null hypothesis determining statistical significance, if it is greater, we fail to reject the null hypothesis, determining no statistical significance.

After each addition, the model repeats the OLS regression for all included variables to see if the new addition decreases variables' coefficient's p-values, if it exceeds the threshold, it is removed (this is the backwards step of the model) (Hayes, 2022). Alternating iterations of forward and backward steps are performed until no more changes can be made, ensuring that the resulting variables are statistically significant and contribute to the model's predictive power, balancing the inclusion of relevant predictors to avoid overfitting (Hayes, 2022).

A p-value threshold of 0.05 and bidirectional elimination of forward and backward steps for our model was run until termination, instead of a capped limit on variables.

3.1.4 Linear Regression

In order to predict the values of opioid analgesic prescriptions using the independent variables identified through our Stepwise Regression, a multiple linear Ordinary Least Squares regression was employed to estimate the coefficients of the linear equation (IBM, 2024). A regular linear regression models the relationship between two variables by fitting a line of best fit to the data, one variable is the explanatory variable, and the other is the dependent variable (Yale University, 2019). The equation for a linear line of best fit is:

$$Y = \alpha + \beta x$$

(Hutcheson, 2011)

- Where x is the independent variable.
- Where Y is the dependent variable.
- Where β is the slope of the line (regression coefficient).
- Where α is the intercept (the value of Y when X = 0).

In order to understand how well a regression performed, OLS compares the observed values of the dependent variable Y with the predicted values from the regression equation (residuals) (Hutcheson, 2011). Residuals indicate the difference between the observed values and the values predicted by the regression line, showing how well the model predicts each data point. By summing the squared residuals (removing negative deviations), the Residual Sum of Squares (RSS) is obtained (Hutcheson, 2011). The RSS is the measure of model-fit for an OLS regression; a large RSS represents a poor fit whilst a perfect fit would be zero (Hutcheson, 2011).

By utilising OLS Linear Regression, our machine learning model aimed to minimise the Residual Sum of Squares (RSS) as the target metric. The model iteratively fits the regression by adjusting the coefficients until the RSS is minimised as much as possible (scikit-learn developers, 2019). In order to incorporate all significant independent variables in our study, the OLS regression model was extended to include all our explanatory variables (Multiple Linear Regression), by simply adding additional variables into the equation (Hutcheson, 2011). The regression equation is the same as the single response variable equation, but this time was predicted by all our explanatory variables (X_1 to X_{10}).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots + \beta_{10} X_{10}$$

(Hutcheson, 2011)

After running our OLS regression, in order to check for linearity, homoscedasticity and independence, each independent variable's residual plot was generated to understand the fit of the variable. Each of the independent variables from the regression were analysed for random patterning, as a random pattern for the residual plot represents a good fit of data (Stattrek, 2022). If any variables exhibited non-linearity or homoscedasticity, a range of methods were incorporated to improve the fit of the data and tested by re-running the OLS regression. Firstly, principal component analysis (PCA) was used to linearly transform the data onto a new coordinate system, capturing the largest variation and reducing dimensionality (scikit-learn, n.d.). However, this resulted in a reduction of normality significance, a heavily tailed QQ plot, and no improvement in the residual plots, so this was not incorporated. Secondly, square rooting the independent variables which exhibited non-linearity or homoscedasticity resulted in the residual plots showing no significant improvement, and once again further compromised the claims of normality, so this was also not employed. Finally, logarithmically transforming the highlighted independent variables resulted in slightly improved residual plots. The logarithmic transformation returned the natural logarithm plus one of each variable array (np.log1p) elementwise (Numpy, 2024). Since our data was normalised, this logarithmic transformation was chosen as it works where values can be zero or very close to zero (Numpy, 2024). With this improvement in residual plots, this iteration of the regression was used for analysis.

After rerunning the regression, the Shapiro-Wilk test was employed in order to test for normality. The Shapiro-Wilk test uses a right tailed hypothesis test with a null hypothesis of the sample being generated from a normal distribution, to determine whether the data deviates from a normal distribution (Malato, 2023).

Finally, to detect multicollinearity and certify the statistical significance of the independent variables, Variance Inflation Factor (VIF) was utilised to check for correlation between multiple independent variables in our multiple regression model, estimating the amount of variance inflation in the coefficients (Potters, 2019). The formula for VIF is:

$$VIF_i = \frac{1}{1 - R_i^2}$$

(Potters, 2019)

- Where R_i^2 is our coefficient of determination for regressing the i^{th} independent variable against the remaining.

A VIF of 1 indicates no correlation between the variable in question and the others, if the VIF falls between 1 and 5 it suggests moderate correlation, whilst a VIF is greater than 5 it suggests high correlation among variables (Potters, 2019). When VIF exceeds 10 it signals significant multicollinearity (Potters, 2019).

3.2 Results

3.2.1 Dimensionality Reduction and Outlier Handling

62/165 features were shown to exhibit a correlation coefficient over the threshold following dimensionality reduction. This illustrates that a significant quantity of the data was redundant. Upon analysis of these highly correlated pairs, cancer data was found to account for the majority of the variables; for each cancer type male and female registrations were nearly identical. In order to resolve this issue, only C00-C97 malignant cancers were included, using the total count column instead of each registration type of malignant cancers ('C00 to C97 excl. C44_All malignant cancers excluding non-melanoma skin cancer (NMSC)'). The rationale behind this, is that ICD codes C00-C97 represent all cases of tumours that invade into surrounding tissues, whilst D00-D48 are benign tumours which are non-invasive and have not yet spread from the surface layer of cells in an organ or other tissue and are thus presumed to have no relation to chronic cancer-pain (NHS Digital, 2023 : Healthline, 2022). Additionally, non-melanoma skin cancer symptoms do not include chronic pain, so the total malignant cancer registration variable excluding non-melanoma skin cancer was incorporated to represent only malignant cancers that can cause chronic cancer-pain (NHS, 2023). With this reduction of variables, two total registration columns for malignant cancers remained: male and female observation counts. The average was taken elementwise for male and female registrations in order to obtain an averaged registration count, reducing cancer data to one singular averaged representative variable.

Male and female observations exhibited a near perfect correlation in the correlation matrix after re-running the matrix, which deemed that gender was redundant and thus removed from the data frame. 5 highly correlated pairs remained, median age, ages 65+, all GP's per capita, total prescribers per capita, and salaried GPs per capita. These important variables were chosen to be incorporated regardless of their correlation to capture demographic and healthcare practice data. The removal of these variables reduced our number of features from 165 to 34, a dimensionality reduction of almost 80%.

A total of 70 out of 11445 data points in our data frame were highlighted as extreme outliers. In order to maintain central tendency, stabilise model performance and mitigate the impact of these outliers, every outlier identified was replaced with the mean of its respective column. The decision to replace outliers with the mean was chosen as it provides a straightforward approach to maintaining the integrity of the dataset by reducing the disproportionate influence of extreme values, ensuring that the central tendency and overall structure of the data remain intact, whilst balancing the trade-offs between model stability and the potential of introduced bias.

Further reducing the dimensionality, Stepwise Regression deemed the following variables statistically significant: number of patients per capita, advanced nurse practitioners per capita, GP regular locums per capita, GPs in training grade per capita, total prescribers per capita, ages 65+, pharmacists per capita, median age, crime - average score , and IMD - average score. With these features extracted, the dimensionality of our Linear Regression is reduced to 10 statistically significant variables, leaving us in a position to conduct our linear regression.

3.2.2 Model Evaluation

After running our first iteration of our OLS regression: number of patients per person, total prescribers per person, GPs in training grade per person, GP regular locums per person, and pharmacists per person residual plots (See *'Figure 4'*) all exhibited non-linearity or homoscedasticity.

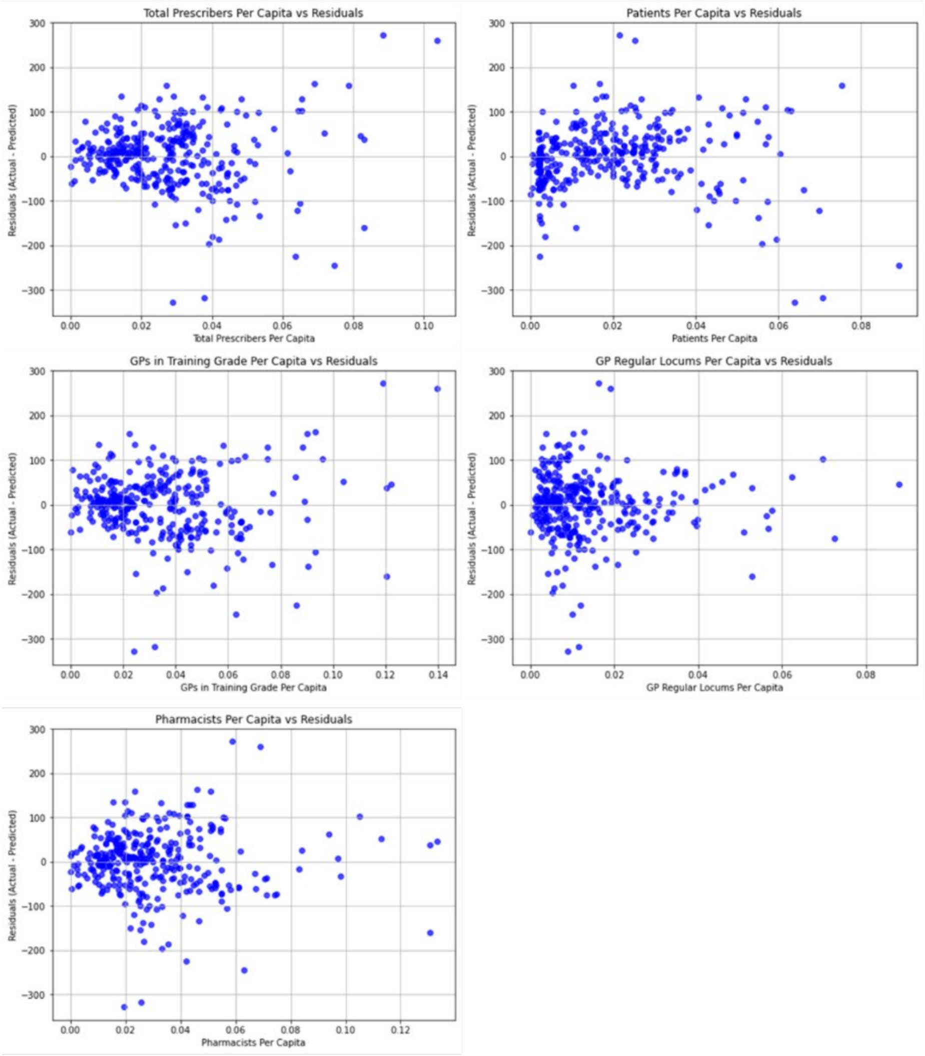

Figure 4: Residual vs Independent Variables (Pre-Transformation)

The logarithmic transformation applied to these independent variables helped to improve the non-linearity or homoscedasticity as shown in *'Figure 5'* from rerunning the OLS regression with the transformed variables.

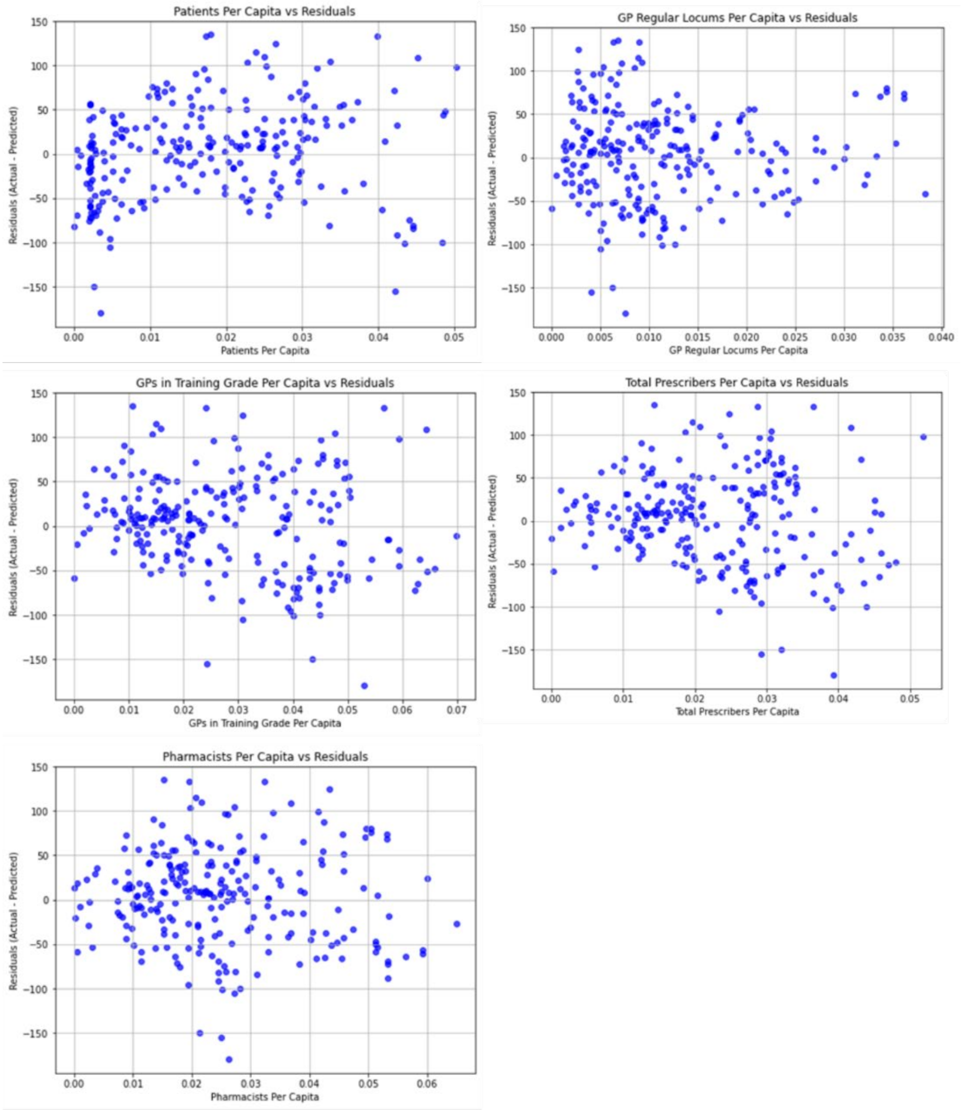

Figure 5: Residual vs Independent Variables (Post-Transformation)

The residual plots for the independent variables now indicated that the assumptions of linearity and homoscedasticity were reasonably met. This suggests that the linear regression was well-specified and that the relationships between the independent variables and the dependent variable were appropriately captured by the model. To further understand if the model performed adequately the model's metrics were analysed, which indicated that the Linear Regression performed well. The model's R^2 of 0.853 meant that 85.3% of the variance in the dependent variable was explained by our selected independent variables, providing further evidence that the model fitted the data well. Furthermore, the F-statistic of 183.4 and p-value of 4.04e-125 illustrated that the model was statistically significant, meaning at least one of the predictor variables is meaningfully associated with opioid prescription quantities per person.

In addition to these metrics, we examined the normality of the residuals to ensure the reliability of our model's parameter estimates (University of Wisconsin, 2024). The Quantile-Quantile (QQ) plot and histogram of residuals suggested that the residuals are approximately normally distributed (See *‘Figure 6’*). However, deviation in the tails of the QQ plot, particularly at the extremes, suggested that while the residuals are approximately normally distributed there may be some outliers or skewness in the data (the red diagonal line represents the theoretical quantiles of a normal distribution). Meanwhile, the histogram of residuals presented a roughly bell-shaped distribution with some skewness present, as seen in the slight asymmetry and the presence of outliers on both sides of the distribution.

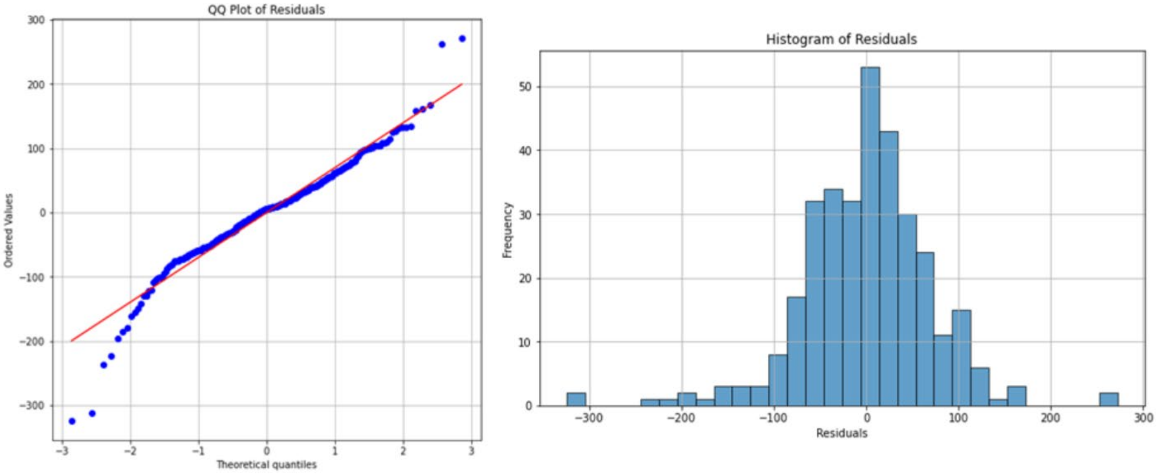

Figure 6: QQ Plot and Histogram of Residuals

The Shapiro-Wilk test statistic of 0.956, with a p-value of 2.37e-08, suggested that we reject the null hypothesis of perfectly normal residuals, indicating that there is some statistically significant deviation from normality.

Despite these deviations, the residuals do not show strong non-linear patterns or heteroscedasticity, as evidenced by the plots (See *‘Figure 6’*). This indicates that the assumptions of linearity and homoscedasticity are reasonably met, ensuring that the model’s estimates are still reliable. The slight non-normality may have some impact on the accuracy of standard errors and confidence intervals, but it is not severe enough to undermine the overall validity of the model.

Finally, to detect any multicollinearity, variance inflation factor (VIF) estimated the variance of the coefficients and their inflation (Potters, 2019). *'Table 4'* displays each feature and their VIF score:

Table 4: Feature VIF Scores

Feature	VIF
Patients Per Capita	1.833798
Advanced Nurse Practitioners Per Capita	4.829971
GP Regular Locums Per Capita	2.399359
GPs in Training Grade Per Capita	11.625991
Total Prescribers Per Capita	15.617462
Ages 65+	21.378165
Pharmacists Per Capita	4.525025
Median Age	21.755956
Crime - Average score	3.848805
IMD - Average score	2.766592

The VIF values revealed that the features: GPs in Training Grade per capita, Total prescribers per capita, Ages 65+, and Median age, exhibit high multicollinearity, well above the commonly accepted threshold of 10. This suggested that these features are highly correlated with others in the model, affecting the stability and interpretability of the regression coefficients. When analysing these coefficients from the results of the model, these variables featuring high multicollinearity results should be taken with precaution as they could be inflated.

3.2.3 Regression Results

Analysis of the model's performance deemed a well-performing and well-fitting enough model to proceed with analysis of the regression's output. The line of best fit, fitted by the OLS regression (Dotted line in *'Figure 7'*), represents a predicted quantity of opioid analgesic prescriptions prescribed per Local Tier authority (our baseline rate of predicted prescription quantities).

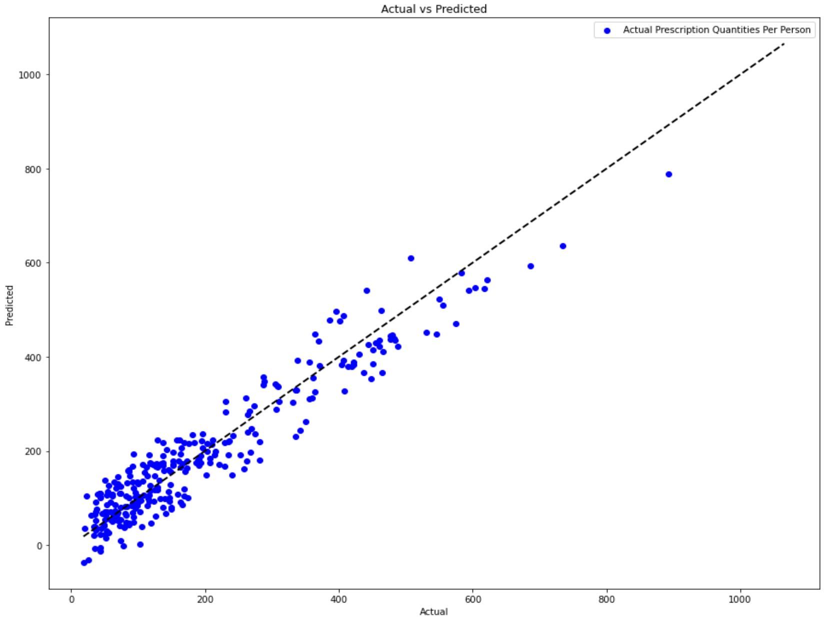

Figure 7: Linear Regression, Actual vs Predicted Plot

Furthermore, the feature importance of our OLS regression model tells us which variable coefficients have the strongest linear positive or negative influence upon opioid analgesic prescriptions. *'Figure 8'* visually represents the feature importance of our independent variables, whilst the coefficients in *'Table 5'* reveal the impact of each variable on the quantity of opioid analgesic prescriptions.

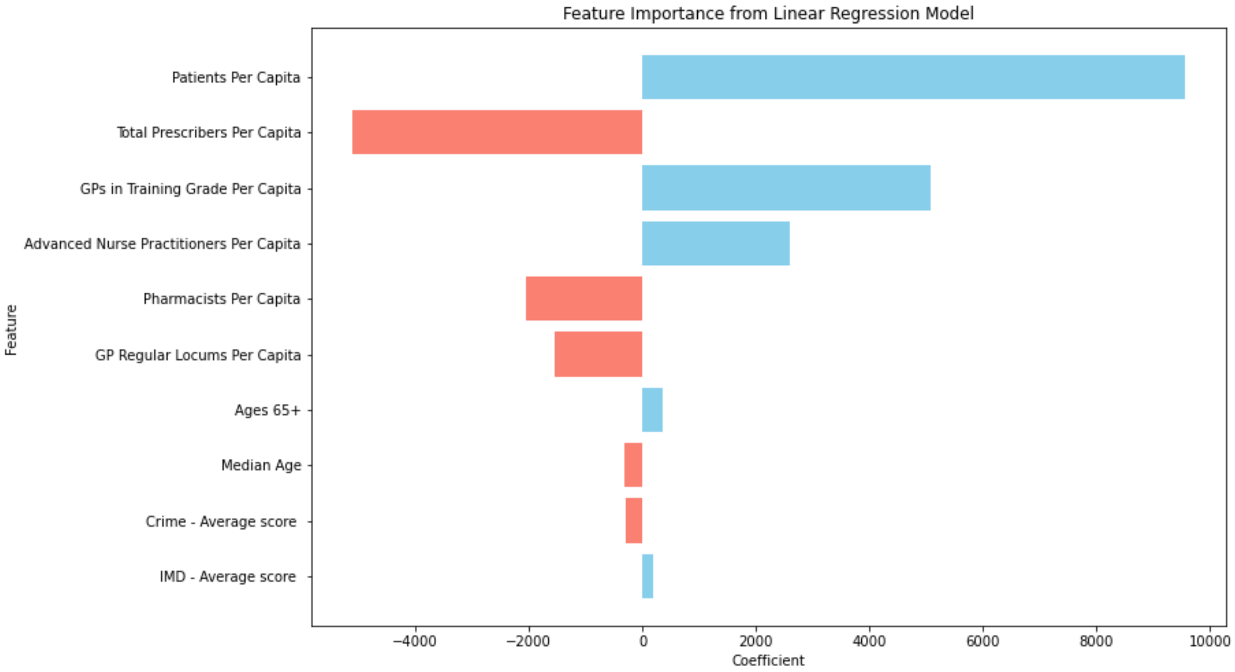

Figure 8: Variable Coefficients

Table 5: Feature Coefficients

Independent Variable	Feature Coefficient	Positive/Negative
Patients Per Capita	9551.46165288098	Positive
Total Prescribers Per Capita	-5109.724942864283	Negative
GPs in Training Grade Per Capita	5083.34746405798	Positive
Advanced Nurse Practitioners Per Capita	2592.3970397892435	Positive
Pharmacists Per Capita	-2053.387625096992	Negative
GP Regular Locums Per Capita	-1533.3815291704288	Negative
Ages 65+	367.20434081405125	Positive
Median Age	-314.3327827748178	Negative
Crime - Average score	-291.3144972082217	Negative
IMD - Average score	190.26410768205005	Positive

The variable "Patients Per Capita" measured the ratio of registered patients at GP practices to the total population within a Local Tier Authority. This variable has a coefficient of 9551.46, indicating the strongest relationship with the quantity of opioid analgesic prescriptions. Specifically, this coefficient suggests that for each additional unit increase in the number of patients per capita, the predicted quantity of opioid prescriptions increases by approximately 9551.46 units.

A more interesting coefficient is the ‘Total Prescribers Per Capita’, this variable represents the number of staff with authorised opioid prescribing abilities at a Local Tier Authority per capita. Displaying the second strongest coefficient, this indicates a potential inverse relationship between the availability of prescribers and the quantity of opioids prescribed.

However, ‘GPs in Training Grade Per Capita’ and ‘Advanced Nurse Practitioners Per Capita’ exhibited strong positive coefficients, indicating the positive linear relationship between these prescribing staff and total opioid prescription quantities. Whilst GPs in training grades VIF score was relatively high so the coefficient should be taken with caution, advanced nurse practitioners displayed a relatively low VIF indicating that it can be deemed an accurate observed relationship. Either way, both variables still indicate a positive relationship.

‘Pharmacists Per Capita’ and ‘GP Regular Locums Per Capita’ both exhibited an inverse relationship to opioid prescription quantities. Additionally, both variables displayed a relatively low VIF score deeming their coefficients to be accurately observed without multicollinearity.

Demographic variables displayed an obscure linear relation, with ‘Ages 65+’ positively influencing opioid prescriptions whilst ‘Median Age’ inversely related, negatively influencing opioids prescription quantities. Both of these variables exhibited the largest VIF scores indicating that their coefficients should be taken with precaution and are most likely inflated.

Deprivation was the least influential of the predictor variables, with ‘Crime’ and ‘IMD’ average scores both ranking lowest out of our coefficient features. Both of these variables scored relatively low VIF’s, deeming their observed relationships as trustworthy and not overly inflated.

These results should be taken with precaution as the VIF analysis highlighted that GPs in training grade per capita, total prescribers per capita, ages 65+, and median age variables displayed high multicollinearity. This implies that these coefficients may be inflated, potentially distorting their true impact. Therefore, while interpreting the coefficients it is important to consider the possibility that these values could be exaggerated due to multicollinearity.

4. Chapter Three: Residual Analysis

This chapter focuses on pinpointing Local Tier Authorities that are either over-prescribing or under-prescribing opioid analgesics by identifying the largest residuals deviating from our line of best fit. By analysing these outliers, we aim to uncover relationships between the variables and identify any underlying trends or patterns that may be causing certain areas to deviate significantly from the expected prescription rates. This analysis will provide insight into potential factors driving these deviations and help understand the dynamics behind opioid prescription practices that differ from ‘normal’ prescription rates.

4.1 Methodology

4.1.1 Identification

In order to identify the Local Tier Authorities that deviate the most, the top 10% of residuals furthest from the line of best fit of our OLS regression were identified (See *'Figure 9'*). This 10% threshold was chosen to identify the most significant outliers while managing the balance between sensitivity and specificity. The selection of this threshold aims to extract extreme cases that will provide meaningful insights into Local Tier Authorities with unusually high or low opioid prescription rates.

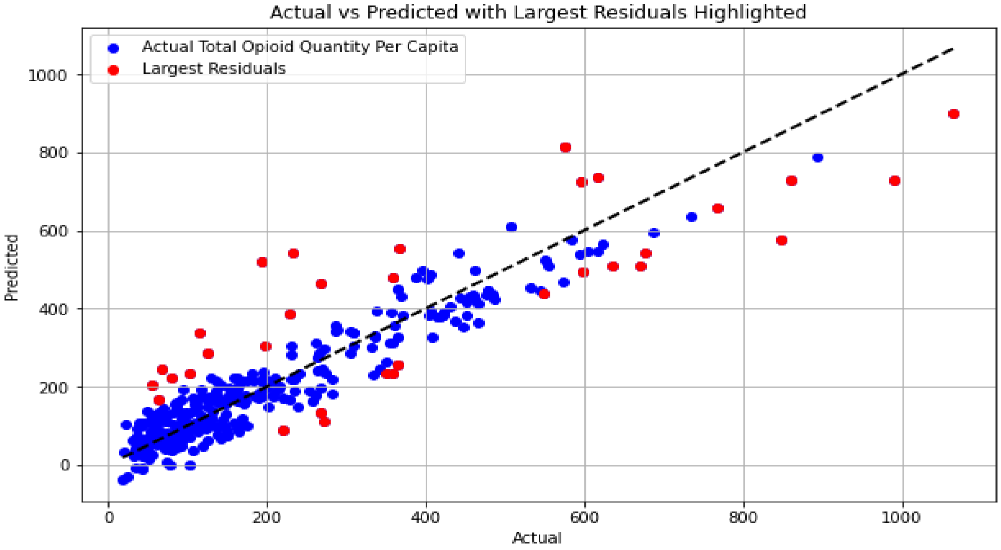

Figure 9: Residuals with Highlighted top 10%

4.1.2 K-Means Clustering

The resultant outlying residuals were then grouped into clusters based on their proximity to the nearest outlying residuals utilising K-means clustering. K-means clustering is an unsupervised machine learning algorithm using *k* number of centroids (a centroid represents the centre of a cluster) to allocate every point in a dataset to the nearest cluster, whilst keeping the centroids as small as possible (Garbade, 2018). To assign data points to their closest centroid, 'k' number of centroids are initialised by randomly selecting data points and defining them as the centroids for each cluster, the model then iteratively calculates the distance between each data point and each cluster using the Euclidean Distance, selecting the smallest distance between a datapoint and centroid (Sharma, 2021). After selecting all data points nearest centroid, the centroids are re-initialised by calculating the average of all data points in a cluster using the following formula:

$$Ci = \frac{1}{Ni} \sum xi$$

(Sharma, 2021)

The model will iteratively assign data points and re-initialise clusters to optimise the positions of the centroids for the clusters, halting when the centroids are stabilised (no improvement upon Euclidean distance) or when the defined number of iterations have been achieved (Garbade, 2018).

Clusters were then calculated and labelled as either over or under predicting according to the difference between predicted and actual values. This provides a data frame of outlying residuals, their original data, their corresponding cluster, and if they are over or under predicted. Each variable's original data was then plotted in a two-figure plot comprised of two box plots, one for under-prediction's residual data and the other for over-prediction's residual data. Thus, allowing for comparative analysis to understand variables influencing over or under prediction in comparison to the residual's pre-processed data.

4.1.3 Analysis of Variance

In addition to visual analysis, Analysis of Variance (ANOVA) testing on each clustered group highlighted features with significantly different values across clusters. ANOVA testing measures the level of variance between more than two groups in which samples are taken from (Qualtrics, 2024). A One-Way ANOVA test compares the means from independent groups to determine if there is a statistically significant difference between them, comparing numeric features across different clusters (our independent groups) (Qualtrics, 2024). High F-statistics and low p-values for variables across our independent groups show a statistically substantial difference between the means of a feature across clusters and vice versa for low F-statistics and high p-values. Highlighted features with significant variance are compared in a graph consisting of box plots of data for each cluster relevant to the cluster's significant variable.

4.2 Results

4.2.1 Outlying Residuals

Eight groups of clusters were identified in our top 10% of residuals (33 residuals), displayed in *‘Figure 10’*. This clustering highlighted eight distinct groups of residuals, resulting in our K-means ‘k’ value being set as eight centroids for the model to group by.

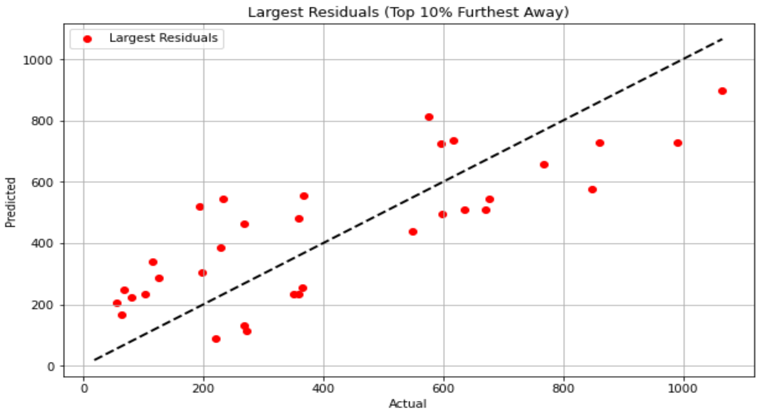

Figure 10: Largest Residuals Only

4.2.2 Clustering

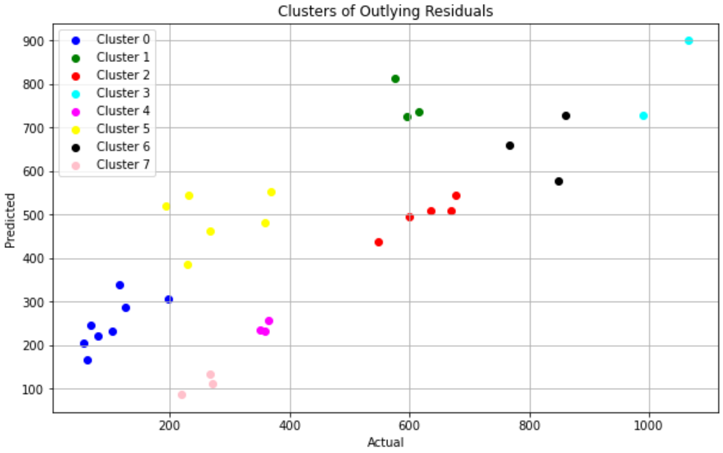

Figure 11: Clusters of Outlying Residuals

‘Figure 11’ illustrates the resulting clusters from our K-means algorithm. After clustering, calculating each residual in each cluster as an over-or under prediction resulted in eight

clusters. Five clusters of under-predicting and three clusters over-predicting, in each cluster there was no individual residual which opposed the overall prediction label.

4.2.3 Local Tier Authorities

Any residuals identified as over-predicting actually represent an under prescription of opioids, whilst any under-predicted residuals represent overprescription, this is because any residuals that are under predicted highlight that actual quantity is greater than expected, whilst being inverted for over-predicted residuals. Each rows Local Tier Authority Code was extracted from residuals to display Local Tier Authorities which were under or overprescribing, as shown in *'Table 6'*:

Table 6: Local Tier Authorities, Clusters and Prediction

Local Tier Code	Local Tier Name	Over/Under Predicted	Cluster
E06000003	Redcar and Cleveland	Over	2
E06000002	Middlesbrough	Over	2
E07000037	High Peak	Over	2
E07000107	Dartford	Over	2
E07000109	Gravesham	Over	2
Identifier Lost	Identifier Lost	Over	3
E07000047	West Devon	Over	3
E06000005	Darlington	Over	6
E07000033	Bolsover	Over	6
E07000035	Derbyshire Dales	Over	6
E06000014	York	Under	0
E06000063	Cumberland	Under	0
E06000064	Westmorland and Furness	Under	0
E07000124	Ribble Valley	Under	0
E08000011	Knowsley	Under	0

E08000022	North Tyneside	Under	0
E08000023	South Tyneside	Under	0
E08000037	Gateshead	Under	0
E06000046	Isle of Wight	Under	5
E07000208	Epsom and Ewell	Under	5
E07000217	Woking	Under	5
E07000226	Crawley	Under	5
E07000228	Mid Sussex	Under	5
E07000229	Worthing	Under	5
E07000079	Cotswold	Over	4
E07000080	Forest of Dean	Over	4
E07000234	Bromsgrove	Over	4
E07000078	Cheltenham	Over	7
E07000213	Spelthorne	Over	7
E07000224	Arun	Over	7
E07000218	North Warwickshire	Under	1
E07000223	Adur	Under	1
E07000088	Gosport	Under	1

4.2.4 Variable Influence

Each variable’s actual data was plotted using box plots to display actual residual data with respect to if the residual was under-predicted or over-predicted (*‘Figure 12’*; only variables with a clear difference are displayed), showing a variable’s actual data distribution, central tendency and variability to compare between over- or under-predictions.

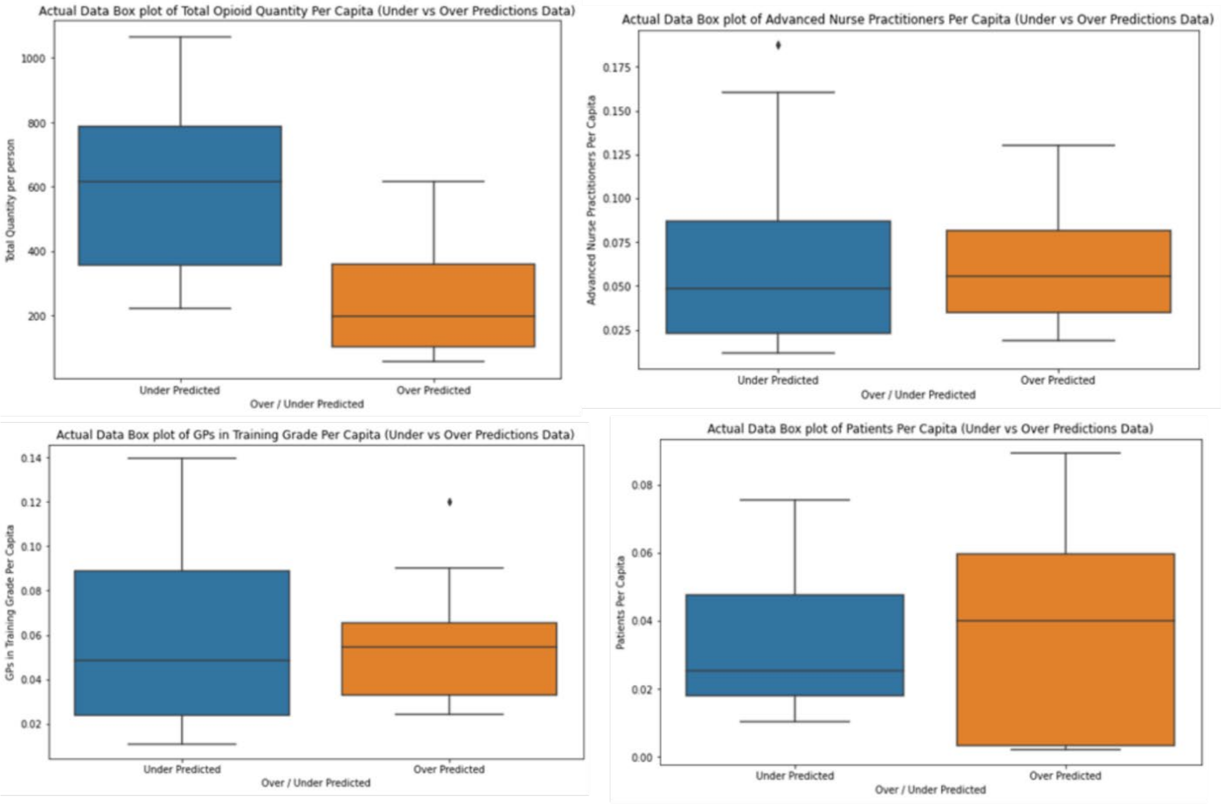

Figure 12: Outlying Residual Box Plots Over/Under Prediction Data

Total opioid quantities per capita displayed the largest difference between medians for over or under predicted data. Specifically, in cases where the model under-predicted opioid prescriptions, the actual amounts prescribed were higher, as shown by the larger median and fourth quartile. This suggests that areas where the model underestimated opioid quantities were, in reality, over-prescribing opioids, and vice versa for over-predictions. This pattern indicates that the discrepancies were not due to random influential variables but rather reflect systematic over- or under-prescribing practices by the Local Tier Authorities involved.

Additionally, the number of patients per capita demonstrated the second largest difference in medians between over-predictions and under-predictions. Over-predictions contained a higher median and a broader distribution of patients per capita, suggesting that the model may have overestimated the influence of this variable when predicting for these outlying residuals. As a result, outlying locations with higher patient numbers in comparison to the population which were over-predicted, actually under-prescribed from the predicted baseline. This could highlight that in these outlying residuals, when a higher percentage of the Local Tier Authorities population are registered to GP practices, GPs may exercise more caution in prescribing opioids, leading to lower prescription rates than anticipated by the model.

Furthermore, the median number of GPs in training grades was higher in areas where opioid prescriptions were over-predicted, suggesting that outlying Local Tier Authorities with more GPs in training had lower-than-expected opioid prescription rates. This may indicate that in areas with under-prescribing, GPs in training grades were not contributing to the actual number of opioids being prescribed as the model predicted. When this is considered alongside the similar results for advanced nurse practitioners, it suggests that these types of staff were not acting in the same way as the majority of Local Tier Authorities from our OLS, potentially contributing to evidence as why they are outlying residuals or of potential confounding variables.

To further understand numerically, instead of visual analysis, ANOVA testing highlighted: number of patients per capita, total prescribers per capita, GPs in training grade per capita and advanced nurse practitioners per capita, as showing significant variation between over and under prediction. Each of these variables passed the significance test with p-values close to zero and F-statistics highlighting the level of variance against the other variables. *‘Table 7’* displays each variable and their corresponding F-Statistic:

Table 7: ANOVA test variables F-statistics

Variable	F-Statistic
Number of patients per capita	11
GPs in training grade per capita	7
Total prescribers per capita	6
Advanced nurse practitioners per capita	4

These results highlight that these specific healthcare-related variables significantly affect the model’s accuracy through variation. The number of patients per capita had the highest F-statistic, indicating that it contributes the most variance between over- and under-predictions. Followed by GPs in training grade per capita also showing a substantial F-statistic (second largest), indicating the strong variation between over- and under-predictions, leading to a strong relation between the number of GPs in training and the opioid analgesic prescription rates. Total prescribers per capita and advanced nurse practitioners per capita also showed notable F-statistics.

These findings suggest that variation in these variables was most pronounced in our outlying residuals, providing evidence that not only were these locations over- or under-prescribing relative to the baseline rate, but also that the influence of these variables diverges from the model's expected coefficients, indicating that in these Local Tier Authorities different prescribing practices might be in place, leading to the under-prescription rates compared to non-outlying residuals. Additionally, these results suggest the potential presence of confounding variables affecting these coefficients, causing prescription rate divergence in these areas from the majority of the model.

5. Discussion

5.1 Data Availability, Reliability and Quality

Preprocessing created a comprehensive dataset, aligning data to the appropriate authority level, further enabling meaningful analysis. However, a major limitation was the exclusion of some variables; despite a thorough literature review, unmeasured confounders influencing opioid prescribing rates may have been overlooked. Furthermore, using data from different years and geographic levels introduced constraints in consistency, comparability, and spatial coverage, potentially leading to biases by the over- or under-representation of certain trends. The absence of data for 'inactive' Local Tier Authorities and missing variables necessitated imputation through use of statistical estimates, which may have introduced inaccuracies and failed to capture true underlying patterns. To improve validity and reduce bias, utilising data of the same year, at the Local Authority level would mitigate these temporal and spatial limitations. Additionally, the lack of chronic pain data was a significant constraint, and incorporating this in future research could enhance the model's robustness.

To ensure data validity, data was collected from reputable sources such as the Office of National Statistics, NHS digital and the UK Government in order to achieve reliability and trustworthiness in our results. The credibility of ONS data is well-established as the ONS is the largest independent producer of official statistics in the UK and operates as a non-ministerial government department (ONS, 2023). Meanwhile, the UK Government was deemed as a credible source of data collection as it is an authoritative and official governmental institution, ensuring rigorous data collection, comprehensive coverage, and adherence to strict quality standards. Furthermore, using data from National Health Service (NHS Digital), ensured accuracy and reliability as the credibility of the EPD is supported by its comprehensive collection and integration of data. Our dependent variable is particularly important and is the backbone of this study, thus utilising the EPD for our opioid prescription quantities was vital. However, the EPD is not without its limitations as prescriptions dispensed outside England, unfulfilled prescriptions, and certain institutions are omitted, introducing considerations for the data's scope and potential biases.

Some limitations were identified in our dataset, suggesting the need for improvements or alternative data sources. For instance, the use of 2018 GDP data to mitigate COVID-19 impacts may not accurately reflect current economic conditions or post-COVID changes, potentially biasing our regression analysis. Additionally, our cancer data was only available at the regional level, limiting the model's ability to capture chronic cancer pain at the Local Tier Authority level. Furthermore, smoking data lacked confidence intervals, making it less reliable, and only covered individuals over 18, missing underage smokers; it is well documented that there are many underage smokers in the UK (ASH, 2024). Finally, our sports data from Sports England, while less reliable due to its survey-based collection method, included confidence intervals, but better statistical significance could have been achieved by excluding estimates that didn't pass these intervals.

5.2 Methodology Considerations

The methods employed to fill estimates for missing data successfully generated the required data by leveraging various techniques, each utilising existing data. The Random Forest Regression leveraged our data frame to predict missing deprivation, smoking and sport's data, whilst Euclidean Distance filled missing 'number of patients registered at a GP' with their most similar Local Tier Authorities. The Feed Forward Neural Network with ReLU activation, effectively handled non-linear relationships and used the estimates of the Random Forest and Euclidean Distance to utilise the entire filled dataset. Additionally, implementing early stopping callbacks minimised the risk of overfitting and scaled data was utilised. Each model performed adequately, providing successful estimations for the missing data. However, several caveats must be considered. The Random Forest and Euclidean Distance methods were executed before the scaling and normalisation process, potentially affecting their performance. Additionally, the Random Forest model used independent variables not highlighted by literature as being strongly linked to the predicted dependent variables for the regression. Furthermore, the Euclidean Distance approach assumes that the dependent variable is similar to its most similar row, overlooking potential outliers or data discrepancies caused by external phenomena. To improve these models, incorporating variables known to be linked to the dependent variables in the Random Forest, as well as normalising and scaling data for both Random Forest and Euclidean Distance would enhance their performance. Further tuning, such as adjusting the train-test split ratio or conducting a grid search for Random Forest and Neural Networks could also improve predictive power. For the Neural Network, experimenting with different activation functions or layer configurations might yield better results (Baheti, 2022). Lastly, replacing the Euclidean Distance formula with a more sophisticated model could help uncover underlying patterns and account for contextual factors, whilst adopting normalisation during scaling would ensure each feature contributes equally to the model, particularly when features have varying distributions.

The methods employed to handle outliers and reduce dimensionality were effective in refining the dataset. Extreme outliers were successfully identified utilising a Z-score threshold of three standard deviations, capturing any data outside of the majority (99.7%) and replacing with the mean (Kenton, 2021). Dimensionality reduction ensured correlated pairs highlighted through Pearson's correlation matrix were removed (except 5 which caused VIF inflation in the analysis), and Stepwise Regression highlighted significant variables influencing opioid prescribing rates. The Stepwise Regression process was particularly effective in reducing dimensionality whilst ensuring statistical significance between variables (McNeese, 2015). However, these methods had limitations. The Z-Score threshold of three standard deviations might have been too lenient or too strict, potentially missing important outliers, and replacing outliers with the mean, whilst convenient, could mask rare but significant events and reduce the dataset's variability, skewing results. Adjusting the Z-Score threshold through trial and error could improve outlier detection. Additionally, Stepwise Regression, while useful for dimensionality reduction, may have overlooked complex interactions between variables. Moreover, leaving some highly correlated variables in the analysis proved problematic, as evidenced by issues with VIF during the analysis (*'Table 4'*). Re-running the correlation matrix and removing highly correlated variables, especially those not identified as significant in the Stepwise Regression or appearing redundant, would help reduce variance inflation.

The Linear Regression and residual analysis employed in this study effectively identified relationships between dependent and independent variables, providing clear interpretations of coefficients (IBM, 2023). Linear regression was particularly suitable for understanding the positive or negative relationships between influential variables using these coefficients to opioid prescription quantities. Residual analysis further supported this by identifying the top 10% of residuals as outliers, offering insights into how outlying residual's actual data compared to the estimated coefficients of the overall model, identifying trends in these outlying residuals; this analysis also provided evidence of confounding variables. However, some limitations need to be addressed. Linear Regression only captures linear relationships, potentially overlooking more complex interactions. The residual analysis focusing on the top 10% of residuals does not capture the overall distribution, risking the false categorisation of some authorities as outliers, when they may simply be within a normal range of variability. Additionally, examining residuals only for overall variability of over- or under-prediction's data does not reveal the nuances between different individual clusters. To improve these methods, considering multicollinearity in the model, employing a ridge regression would add a penalty term to the OLS, reducing the impact of highly correlated variables and addressing issues with variance inflation (Murel, 2023). Incorporating the Interquartile Range (IQR) into the residual analysis could offer a more statistically sound method for identifying true outliers (Third Space Learning, 2024). Furthermore, analysing individual clusters within the residuals could provide deeper insights into how variables change with over or under-prediction, capturing the nuances that the overall residual analysis might miss.

5.3 Ethics

This study was conducted with a strong emphasis on ethical considerations. Ethical and privacy was paramount given the sensitivity of the results and healthcare, prescription and deprivation data. All datasets utilised contained no individual level information and the data sources utilised followed relevant data protection guidelines such as GDPR (GDPR, 2018). Throughout the methodology of the study, the critical focus was to maintain statistical integrity to ensure results were statistically significant particularly in identifying specific Local Tier Authorities. Inaccurate findings could potentially affect public health policy or misinform stakeholders, leading to ethical implications. The baseline rates generated through Linear Regression, while useful for analysis, may not fully capture the complexity of the environments in which opioid prescribing occurs. This limitation was carefully considered to avoid ethical pitfalls, such as misrepresenting the prescribing behaviours of healthcare providers or oversimplifying the factors that influence opioid use. Thus, the study's approach to identifying Local Tier Authorities that were under- or over-prescribing opioids was handled with sensitivity. Highlighting these areas carried ethical considerations, as it could lead to stigmatisation or unfair scrutiny. The results of this research are produced with the intent to inform and support improvements in public health, rather than to cast blame or create negative perceptions of specific Local Tier Authorities.

5.4 Results

This study successfully reduced the dimensionality of variables identified in literature to incorporate statistically significant features to England's opioid epidemic. Estimated coefficients from the OLS regression's independent variables provided information on whether significant variables positively or negatively influenced opioid prescription rates. Local Tier Authorities that deviated most in their prescription rates from the model's baseline rate were identified and whether they were over- or under-prescribing. These Local Tier Authorities original data was analysed to gain a deeper understanding into what was influencing these outlying residual's deviations.

The results of the Stepwise Regression will be a valuable addition to current literature, as they provide statistically significant variables relevant to England's opioid prescription quantities. These results highlight that the number of patients registered at a GP, ages 65+, median age, crime, and overall deprivation statistics are all statistically significant in the landscape of opioid prescribing in England. Additionally, the Stepwise Regression identified cohorts of staff who had a significant relation to England's opioid epidemic; advanced nurse practitioners, GPs in training, pharmacists, and also that the total number of prescribing staff all statistically influence opioid prescription rates.

Building on these identified features, the OLS regression coefficient analysis provided deeper understanding of the influence of these variables. The largest coefficient produced was 'Patients Per Capita', displaying the strongest positive coefficient. This variable represents the number of patients registered at a GP, scaled by the Local Tier Authorities population, providing the percentage of the population that are registered to surgeries within a Local Tier Authority. As the percentage of a population registered at a GP increases, so do opioid prescription quantities for a Local Tier Authority. This coefficient may imply that better GP and healthcare services in a Local Tier Authority enables a higher percentage of the population to be registered, resulting in more frequent medical consultations and treatments such as opioids. Opposingly, this could also imply that areas with more patients registered out of the population reduces the amount of time available to doctors, causing prescription practices to be more lenient. Due to this, this coefficient cannot be interpreted with any rationale as to why it causes the largest increase due to an abundance of potential factors, the most logical argument would be that when the number of patients registered increases, there are more patients that will require opioids, thus, while interesting, this coefficient does not contribute to the understanding of the overall landscape of opioid prescribing in England.

A more interesting coefficient is the ‘Total prescribers per capita’, representing the number of staff with authorised opioid prescribing abilities per capita at a Local Tier Authority. Displaying the second strongest coefficient (negative), this indicates a potential inverse relationship between the availability of prescribers and the quantity of opioids prescribed. This apparently contradictory relationship could be due to a range of factors, such as more available time for prescribing staff to ensure patients only receive opioids when necessary, more effective peer to peer review, improved patient management, or the adoption of alternative treatment options. However, this variable was highlighted by our VIF testing to be influenced by multicollinearity meaning this could be inflated from the actual perceived influence, thus this finding should be taken with precaution.

On the contrary to ‘Total prescribers per capita’ inverse relationship, ‘GPs in training grade per capita’ and ‘Advanced nurse practitioners per capita’ exhibited strong positive coefficients, indicating the positive linear relationship between these prescribing staff and total opioid prescription quantities. However, GPs in training grade’s VIF score was relatively high so the coefficient should be taken with caution, and advanced nurse practitioners displayed a relatively low VIF, indicating that it can be deemed an accurate observed relationship. Either way, both variables still indicate a positive relationship, suggesting that an increase in the number of these staff per capita is associated with higher opioid prescription quantities. This reflects a trend where undertrained GP’s or advanced nurse practitioners with less training and education than qualified GPs, contribute to an increased overall prescribing rate; potentially due to their learning phase involving more hands-on prescribing practices, a tendency to prescribe opioids more leniently, or different prescribing behaviours compared to more experienced practitioners. A study by Levy et al in the US, investigated the trends in opioid prescribing rates by speciality and highlighted the rates by practice specialty, indicating that there is a difference between types of practice speciality and their prescription rates, but little evidence in literature could be found for individual staff levels and prescribing rates (Levy et al., 2015).

To provide further evidence of this trend, coefficient analysis of the more experienced members of staff such as ‘Pharmacists per capita’ and ‘GP regular locums per capita’ both exhibited an inverse relationship to opioid prescription quantities. Additionally, both variables displayed a relatively low VIF score deeming their coefficients to be accurately observed without multicollinearity. This inverse relationship could be due to the level of experience from these members of staff, refraining them from over prescribing opioids. Coupled with the coefficients of advanced nurse practitioners and GPs in training grade, one clear factor differentiating the two is the level of experience, or the level of education; indicating these factors could be influencing the opposing prescription rates. This further compounds the trend between experienced members of staff and lesser experienced or under-educated staff and opioid prescription rates.

Whilst demographic variable’s coefficients were less than those of healthcare accessibility or healthcare providers, the demographic variables still helped to understand the landscape of opioid prescribing across demographics in England. For instance, ‘Ages 65+’ coefficient was positive, indicating as the quantity of this older demographic increases in a Local Tier Authority, so do the rates of opioids per capita. This trend was expected as it was highlighted by Mikelyte, et al, in their study that concluded that chronic pain occurred most commonly in older people being found in 45%–85% of people in their later years (Mikelyte, et al, 2020). However, this trend was mainly down to cancer, as cancer risks exponentially increased in the latter years of life, illustrated by DePinho in their study (DePinho, 2000). This posed the question as to why cancer data was deemed not significant by our Stepwise Regression; this may have been because cancer data was only available on a regional level.

Additionally, the demographic variable ‘Median age’ was inversely related to opioid prescription rates, suggesting that as the median age of a Local Tier Authority increases, opioid prescription rates decrease. This trend potentially indicates that as a population ages, healthcare services in the area may become overwhelmed due to the higher demand for medical care, which is typically associated with older populations (Keehan et al., 2004). Consequently, this strain on healthcare resources might lead to a reduction in the resources allocated to pain management, including opioid prescriptions. Although, both demographic variables exhibited the largest VIF scores, indicating that their coefficients should be taken with precaution and are most likely inflated.

Our least influential type of predictor variables was Deprivation; with ‘Crime’ and ‘IMD’ average scores both ranking last out of our coefficients (*‘Table 5’*). Deprivation data scoring the lowest coefficient was an interesting result, as studies from Cremer et al, Martin Gulliford, van Draanen et al, Rajabi et al, and Joynt et al, all highlighted in our researched literature the important link between socioeconomic variables and opioid use, prescriptions, or abuse (Cremer et al., 2021 : Gulliford, 2020 : van Draanen et al., 2020 : Rajabi et al., 2019 : Joynt et al., 2013). Furthermore, deprivation and socioeconomic variables were the most prevalent in literature, so deprivation data scoring the lowest coefficients indicates that this is not likely to be the root cause of the epidemic. Furthermore, both of these variables scored relatively low VIF’s, deeming their relationships observed as trustable and not overly inflated, further solidifying this finding. To further justify this relationship, deprivation variables not included in this study should be incorporated to see if they have a larger influence on opioid prescription quantities in England. For example, our literature review included a study from Rajabi et al, which indicated the link between black patients receiving fewer prescriptions in the US; incorporating this variable might help to see if this relationship also exists in England. However, from this study we can conclude that the deprivation and socioeconomic variables did not cause as much influence as expected from our literature analysis.

Overall, our linear regression results demonstrate the link between the level of staff and experience of opioid practitioners and the rate at which they prescribe. Healthcare staff variables illustrated a more significant influence on opioid prescription staff than demographic variables, or deprivation variables. This suggests that the epidemic is being fuelled by the staff prescribing the opioids rather than the patients receiving them, counteracting the original assumption that patients abuse the fact that pain is unquantifiable. These results help to build on current literature, not only by providing statistical evidence between the variables included in our Linear Regression in England, but also by filling in the missing gaps in understanding what is fuelling this epidemic.

To further understand the epidemic, residual analysis highlighted Local Tier Authorities deviating from the model’s baseline prediction rate. However, the resulting analysis of these residuals illustrated that advanced nurse practitioners and GPs in training grade prescribed less than predicted, as the residuals data was disparate from the coefficients of the regression. This suggested that in these areas prescribing mitigation may be in place, and healthcare providers should take a deeper look into how these members of staff prescribe in the Local Tier authorities highlighted. This provided further evidence of confounding variables in deviating Local Tier Authorities, meaning that this study should be reconducted, incorporating more healthcare variables to further understand where this unforeseen prescription quantity disparities are coming from. An individual cluster analysis was originally conducted to further understand these deviations, determining which clusters had confounding variables swaying the expected influence of opioid prescriber variables. However, due to spatial limitations, this analysis has not been incorporated into this study, replicators may wish to build on this study by including this in their analysis. Furthermore, a geographical analysis was conducted looking at deviating Local Tier Authorities geographically, this did not highlight an obvious link between geographic locations between these Local Tier Authorities, this also was not included due to spatial constraints.

6. Conclusion

This study has illustrated the significant influence of prescribing staff on England’s opioid epidemic. While previous literature has focused on demographic and socioeconomic factors, our findings shift the spotlight to the critical role of healthcare providers, particularly GPs in training and advanced nurse practitioners; these less experienced practitioners are associated with higher opioid prescription rates, suggesting that the epidemic is being fuelled more by prescriber behaviour than by patient demand.

Policymakers should take these results as compelling evidence to guide resource allocation and interventions aimed at mitigating this public health crisis. A targeted approach might involve limiting the prescribing authority of GPs in training and advanced nurse practitioners, such as requiring all opioid prescriptions by these practitioners to be approved by an experienced GP before dispensing or mandating specialised training in opioid prescribing for advanced nurse practitioners which could further reduce redundant opioid prescribing. Furthermore, the implications of these findings extend beyond policy to clinical practices, emphasising the need for enhanced oversight and strict regulation in prescribing. Implementing these suggestions could play a crucial role in understanding differences between prescribing practices, preventing England’s opioid epidemic from mirroring the epidemic observed in America.

While this study provides valuable insights, it also highlights the complexity of opioid prescribing practices and the need for further research. Future studies should incorporate additional variables, such as specific healthcare provider behaviours and patient outcomes, to better understand the causal pathways and confounding factors influencing prescribing rates. This will be essential for developing more targeted and effective interventions.

Finally, it is important to acknowledge that the data used in this study, from 2021 and earlier, may not fully reflect the current landscape. As the healthcare environment continues to evolve, ongoing research will be crucial to ensure that policies remain effective and responsive to new trends. This study lays a strong foundation for such future efforts, offering a robust framework for further exploration of this critical public health issue. Given the significant public health implications, it is imperative that policymakers and healthcare leaders act swiftly to address these findings, ensuring that England's opioid prescribing practices do not spiral out of its currently somewhat refined remits.

7. Acknowledgments

My journey through education has been an interesting one. I never would have thought I would be studying Data Science for a master's degree at Durham; If I had told that to my younger self he would have laughed. Nevertheless, this has only been achievable with the help and support of some very important people to me, so I wish to acknowledge and thank the following people.

I first would like to thank my mum, dad and sister. Each one of you has provided unwavering support and belief in me, stuck with me through the hard times, and never stopped motivating me, you are all greatly appreciated. I would also like to acknowledge my extended family, my late Grandfather, Norman Mayes, who inspired me throughout my childhood and continues to guide me throughout my life. Finally, Maureen Mayes, my grandmother, who has always provided me with unconditional love and support that I will always cherish. Thank you all for being my foundation.

I would also like to thank Zuzer Musaji (Z), although many years ago, Z gave me the cultivating environment to fuel the development of my work ethic, even throughout many trials and tribulations, Z stuck with me and helped sow the seeds for my drive that I hold today. I hope your business continues to succeed, and that your hard work continues paying off.

Additionally, I would like to thank Sophie King. Sophie has been one of my best friends since the start of university; she has been my bundle of positivity; always being an optimist, always

looking for an adventure, and always ensuring that I have worked to the best of my ability. Have fun in Australia Sophie, look forward to visiting you.

Finally, I would like to thank Hana Hussian, your support, kindness, and encouragement through my masters meant a lot to me. I’m confident you’ll excel—study hard, pursue your passion, and success will follow. Thank you for believing in me and for all the ways you have contributed to my journey, I am truly grateful.

8. Reference List

1. Abhinav Bandaru (2022). *Min-Max Normalization*. Medium. [online]. Available at: https://medium.com/@abhi1achiever/min-max-normalization-db1f515b08b4 [Accessed 05/08/2024]
2. alexlenail.me. (2024). *NN SVG*. [online] Available at: https://alexlenail.me/NN-SVG/index.html. [Accessed 05/08/2024]
3. Alto, V. (2023). *Understanding Ordinary Least Squares (OLS) Regression*. Built In. [online] builtin.com. Available at: https://builtin.com/data-science/ols-regression. [Accessed 13/08/2024]
4. AnalytixLabs (2023). *Random Forest Regression — How it Helps in Predictive Analytics?* [online] Medium. Available at: https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4#:~:text=Moreover%2C%20it%20is%20less%20prone. [Accessed 17/08/2024]
5. Anekar, A.A., Cascella, M. and Hendrix, J.M. (2023). *WHO analgesic ladder*. [online] National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/books/NBK554435/. [Accessed 18/08/2024]
6. Arize AI. (2023). *Mean Absolute Error In Machine Learning: What You Need To Know*. [online] Available at: https://arize.com/blog-course/mean-absolute-error-in-machine-learning-what-you-need-to-know/. [Accessed 04/08/2024]
7. ASH (2024). *Young people and smoking*. [online] ASH. Available at: https://ash.org.uk/resources/view/young-people-and-smoking#:~:text=About%20400%2C000%2011%2D%20to%2015. [Accessed 02/08/2024]
8. Baheti, P. (2022). *12 Types of Neural Networks Activation Functions: How to Choose?* [online] www.v7labs.com. Available at: https://www.v7labs.com/blog/neural-networks-activation-functions. [Accessed 23/08/2024]
9. Bakshi, C. (2020). *Random Forest Regression*. [online] Medium. Available at: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84. [Accessed 03/08/2024]
10. Ballantyne, J.C. and LaForge, S.K. (2007). *Opioid dependence and addiction during opioid treatment of chronic pain*. Pain, 129(3), pp.235–255. doi:https://doi.org/10.1016/j.pain.2007.03.028.

11. Benjamin, H.J., Perri, M.M., Leemputte, J., Lewallen, L. and DeVries, C. (2024). *Opioids and Youth Athletes*. Sports Health. doi: https://doi.org/10.1177/19417381241228629.
12. BSA Business Services Authority. (2024). *'English Prescribing Dataset (EPD)'*. Available at: https://opendata.nhsbsa.net/dataset/english-prescribing-data-epd [Accessed 25/07/2024]
13. Campbell, C.I., Weisner, C., LeResche, L., Ray, G.T., Saunders, K., Sullivan, M.D., Banta-Green, C.J., Merrill, J.O., Silverberg, M.J., Boudreau, D., Satre, D.D. and Von Korff, M. (2010). *Age and Gender Trends in Long-Term Opioid Analgesic Use for Noncancer Pain.* American Journal of Public Health, [online] 100(12), pp.2541–2547. doi: https://doi.org/10.2105/AJPH.2009.180646.
14. Celentano, D. (2020). *The Worldwide Opioid Pandemic: Epidemiologic Perspectives*. Epidemiologic Reviews, 42(1), pp.1–3. doi: https://doi.org/10.1093/epirev/mxaa012.
15. Cleveland Clinic (2021). *Chronic Pain*. [online] Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/diseases/4798-chronic-pain. [Accessed 21/07/2024]
16. Codeacademy. (2024). *Normalization*. [online] Codecademy. Available at: https://www.codecademy.com/article/normalization. [Accessed 05/08/2024]
17. Cremer, L.J., Underwood, N., Robinson, A., Guy, G.P. and Rooks-Peck, C.R. (2021). *Association between county-level sociodemographic characteristics and county-level differences in opioid dispensing.* Preventive Medicine Reports, 24, p.101612. doi: https://doi.org/10.1016/j.pmedr.2021.101612.
18. Dahiru, T. (2008). *P-Value, a true test of statistical significance? A cautionary note.* Annals of Ibadan Postgraduate Medicine, [online] 6(1), pp.21–26. doi: https://doi.org/10.4314/aipm.v6i1.64038. [Accessed 11/08/2024]
19. DataCamp. (2023). *The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions.* [online] Available at: https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning [Accessed 10/08/2024]
20. DeepAI (2019). *Feed Forward Neural Network.* [online] DeepAI. Available at: https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network. [Accessed 04/08/2024]
21. DePinho, R.A. (2000). *The age of cancer. Nature*, [online] 408(6809), pp.248–254. doi: https://doi.org/10.1038/35041694.
22. Dey, S. and Vrooman, B.M. (2022). *Alternatives to Opioids for Managing Pain*. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK574543/. [Accessed 18/08/2024]

23. Dremio. (2024). *ReLU Activation Function*. Dremio. [online] Available at: https://www.dremio.com/wiki/relu-activation-function/. [Accessed 04/08/2024]
24. Fernando, J. (2024). *R-Squared: Definition, Calculation Formula, Uses, and Limitations.* [online] Investopedia. Available at: https://www.investopedia.com/terms/r/r-squared.asp. [Accessed 04/08/2024]
25. FindThatPostcode. (2024). *Find that Postcode*. [online] Available at: https://findthatpostcode.uk/. [Accessed 03/08/2024].
26. FPM Faculty of Pain Medicine (2020). *Opioids Aware: A resource for patients and healthcare professionals to support prescribing of opioid medicines for pain.* [online] Faculty of Pain Medicine. Available at: https://fpm.ac.uk/opioids-aware. [Accessed 18/07/2024]
27. Frost, J. (2017). *How to Interpret P-values and Coefficients in Regression Analysis.* [online] Statistics By Jim. Available at: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/. [Accessed 11/08/2024]
28. Frost, J. (2019). *Guidelines for Removing and Handling Outliers in Data.* [online] Statistics By Jim. Available at: https://statisticsbyjim.com/basics/remove-outliers/. [Accessed 09/08/2024]
29. Frost, J. (2021). *Empirical Rule: Definition, Formula, and Uses.* [online] Statistics By Jim. Available at: https://statisticsbyjim.com/probability/empirical-rule/. [Accessed 10/08/2024]
30. Frost, J. (2021). *Mean Squared Error (MSE).* [online] Statistics By Jim. Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/. [Accessed 04/08/2024]
31. Garbade, M. (2018). *Understanding K-means Clustering in Machine Learning.* [online] Towards Data Science. Available at: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed 14/08/2024]
32. GDPR (2018). *General data protection regulation (GDPR)*. [online] General Data Protection Regulation (GDPR). Available at: https://gdpr-info.eu/. [Accessed 23/08/2024]
33. GeeksforGeeks. (2024). *Euclidean Distance | Formula, Derivation & Solved Examples.* [online] Available at: https://www.geeksforgeeks.org/Euclidean -distance/. [Accessed 04/08/2024]
34. Giles, M. and Malcolm, M. (2021). *Prescription Opioid Misuse and Property Crime.* Social Science Quarterly, 102(2), pp.663–682. doi: https://doi.org/10.1111/ssqu.12945.
35. GOV (2023). *NHS entitlements: migrant health guide.* [online] Available at: https://www.gov.uk/guidance/nhs-entitlements-migrant-health-guide#:~:text=Hospital%20treatment%20is%20free%20of. [Accessed 24/07/2024]

36. GOV.UK (2020). *Opioid medicines and the risk of addiction.* [online] GOV.UK. Available at: https://www.gov.uk/guidance/opioid-medicines-and-the-risk-of-addiction. [Accessed 19/07/2024]
37. Gulliford, M. (2020). *Opioid use, chronic pain and deprivation.* EClinicalMedicine, 21, p.100341. doi: https://doi.org/10.1016/j.eclinm.2020.100341.
38. Hammersley, R., Forsyte, A., Morrison, V. and Davies, J.B. (1989). *The Relationship between Crime and Opioid Use.* Addiction. [online] 84(9), pp.1029–1043. doi: https://doi.org/10.1111/j.1360-0443.1989.tb00786.x.
39. Healthline. (2022). *Benign Tumors: Causes, Types, Symptoms, Diagnosis, Treatment.* [online] Available at: https://www.healthline.com/health/benign#vs-malignant. [Accessed 05/08/2024]
40. Hutcheson, G. D. (2011). *Ordinary Least-Squares Regression*. In L. Moutinho and D. Hutcheson, The SAGE Dictionary of Quantitative Management Research. Pages 224-228. [online] Available at :https://datajobs.com/data-science-repo/OLS-Regression-[GD-Hutcheson].pdf [Accessed 13/08/2024]
41. IBM (2023). *What is a Decision Tree*. IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a. [Accessed 17/08/2024]
42. IBM (2024). *What Is Linear Regression?*. IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/linear-regression#:~:text=IBM-. [Accessed 12/08/2024]
43. IBM. (2024). *What is Dimensionality Reduction?*. IBM. [online] Available at: https://www.ibm.com/topics/dimensionality-reduction#:~:text=Dimensionality%20reduction%20is%20a%20method. [Accessed 05/08/2024]
44. Jani, M. and Dixon, W.G. (2017). *Opioids are not just an American problem*. BMJ, p.j5514. doi: https://doi.org/10.1136/bmj.j5514.
45. Joynt, M., Train, M.K., Robbins, B.W., Halterman, J.S., Caiola, E. and Fortuna, R.J. (2013). *The Impact of Neighborhood Socioeconomic Status and Race on the Prescribing of Opioids in Emergency Departments Throughout the United States.* Journal of General Internal Medicine, 28(12), pp.1604–1610. doi: https://doi.org/10.1007/s11606-013-2516-z.
46. Keehan, S.P., Lazenby, H.C., Zezza, M.A. and Catlin, A.C. (2004). *Age Estimates in the National Health Accounts.* Health care financing review, [online] 26(2), pp.1–16.

Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194874/. [Accessed 23/08/2024]

47. Kenton, W. (2021). *Empirical Rule*. [online] Investopedia. Available at: https://www.investopedia.com/terms/e/empirical-rule.asp. [Accessed 21/08/2024]
48. Kosten, T. and George, T. (2002). *The Neurobiology of Opioid Dependence: Implications for Treatment.* Science & Practice Perspectives, [online] 1(1), pp.13–20. doi: https://doi.org/10.1151/spp021113.
49. Kumar, A. (2022). *R-squared, R2 in Linear Regression: Concepts, Examples*. [online] Data Analytics. Available at: https://vitalflux.com/r-squared-explained-machine-learning/. [Accessed 04/08/2024]
50. Levy, B., Paulozzi, L., Mack, K.A. and Jones, C.M. (2015). *Trends in Opioid Analgesic–Prescribing Rates by Specialty, U.S., 2007–2012.* American Journal of Preventive Medicine, 49(3), pp.409–413. doi: https://doi.org/10.1016/j.amepre.2015.02.020. [Accessed 23/08/2024]
51. Makary, M.A., Overton, H.N. and Wang, P. (2017). *Overprescribing is major contributor to opioid crisis.* BMJ, 359, p.j4792. doi: https://doi.org/10.1136/bmj.j4792.
52. Malato, G. (2023). *An Introduction to the Shapiro-Wilk Test for Normality*. Built In. [online] builtin.com. Available at: https://builtin.com/data-science/shapiro-wilk-test. [Accessed 14/08/2024]
53. McNeese, B. (2015). *Stepwise Regression. SPC for Excel*. Available at: https://www.spcforexcel.com/knowledge/root-cause-analysis/stepwise-regression/ [Accessed 23/08/2024]
54. MHCLG (2019). *File 10: Local Authority district summaries*. [online] GOV.UK. Available at: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.
55. MHCLG (2019). *The English Indices of Deprivation 2019*. [online] Available at: https://assets.publishing.service.gov.uk/media/5d8e26f6ed915d5570c6cc55/IoD2019_Statistical_Release.pdf. [Accessed 30/07/2024]
56. MHRA (2020). *Opioids: Risk of Dependence and Addiction*. [online] GOV.UK. Available at: https://www.gov.uk/drug-safety-update/opioids-risk-of-dependence-and-addiction. [Accessed 21/07/2024]
57. Mikelyte, R. et al. (2020) *'Factors influencing trends in opioid prescribing for older people: a scoping review'*, Primary Health Care Research & Development, 21, p. e36. doi:10.1017/S1463423620000365.

58. Minitab. (2024). *What is the standard error of the coefficient?* [online] Available at: https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/regression/supporting-topics/regression-models/what-is-the-standard-error-of-the-coefficient/#:~:text=The%20standard%20error%20of%20the%20coefficient%20measures%20how%20precisely%20the. [Accessed 12/08/2024]
59. Murel, J. (2023). *What is ridge regression?* | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/ridge-regression. [Accessed 23/08/2024]
60. National Library of Medicine. (2024*). Finding and Using Health Statistics*. [online] Available at: https://www.nlm.nih.gov/oet/ed/stats/02-910.html. [Accessed 10/08/2024]
61. Nettleton, D. (2014). *Pearson Correlation - an overview*. [online] www.sciencedirect.com. Available at: https://www.sciencedirect.com/topics/computer-science/pearson-correlation. [Accessed 05/08/2024]
62. Neural Data Science. (2023). *Data Cleaning - Dealing with Outliers — Data Science for Psychology and Neuroscience — in Python.* [online] Available at: https://neuraldatascience.io/5-eda/data_cleaning.html. [Accessed 09/08/2024]
63. NHS (2022). *Opioid prescribing for chronic pain*. [online] www.england.nhs.uk. Available at: https://www.england.nhs.uk/south/info-professional/safe-use-of-controlled-drugs/opioids/. [Accessed 18/07/2024]
64. NHS Digital. (2023). *Cancer Registrations Statistics, England 2021- First release, counts only.* [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/cancer-registration-statistics/england-2021---summary-counts-only. [Accessed 31/07/2024]
65. NHS Digital. (2023). *Counts of cancer diagnoses tables.* [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/cancer-registration-statistics/england-2021---summary-counts-only. [Accessed 31/07/2024]
66. NHS Digital. (2024). *General Practice Workforce, 30 April 2023*. [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services. [Accessed 31/07/2024]
67. NHS Digital. (2024). *General Practice Workforce*. [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services. [Accessed 31/07/2024]

68. NHS Digital. (2024). *Patients Registered at a GP Practice, April 2023.* [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice. [Accessed 31/07/2024]
69. NHS Digital. (2024*). Patients Registered at a GP Practice.* [online] Available at: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice. [Accessed 31/07/2024]

70. NHS England, (2023). *Optimising personalised care for adults prescribed medicines associated with dependence or withdrawal symptoms: Framework for action for integrated care boards (ICBs) and primary care*. [online] Available at: https://www.england.nhs.uk/long-read/optimising-personalised-care-for-adults-prescribed-medicines-associated-with-dependence-or-withdrawal-symptoms/#:~:text=Prescribing%20of%20low%20dose%20opioids. [Accessed 16/07/2024]
71. NHS inform. (2023). *What is chronic pain?* [online] Available at: https://www.nhsinform.scot/illnesses-and-conditions/brain-nerves-and-spinal-cord/chronic-pain/what-is-chronic-pain/. [Accessed 21/07/2024]
72. NHS. (2023). *Symptoms of non-melanoma skin cancer.* [online] Available at: https://www.nhs.uk/conditions/non-melanoma-skin-cancer/symptoms/. [Accessed 05/08/2024]
73. NICE (2023). *Pain, chronic*. [online] National Institute for Health and Care Excellence. Available at: https://bnf.nice.org.uk/treatment-summaries/pain-chronic/. [Accessed 21/07/2024]
74. NICE (National Institute for Health and Care Excellence), (2024). *Analgesics*. [online] Available at: https://bnf.nice.org.uk/treatment-summaries/analgesics/#:~:text=Opioid%20analgesics%20are%20usually%20used,pain%20particularly%20of%20visceral%20origin. [Accessed 18/07/2024]
75. Numpy. (2024). *numpy.log1p — NumPy v1.26 Manual.* [online] Available at: https://numpy.org/doc/stable/reference/generated/numpy.log1p.html. [Accessed 13/08/2024]
76. Office for National Statistics (ONS) (2023). *Deaths related to drug poisoning in England and Wales - Office for National Statistics*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deat

hs/bulletins/deathsrelatedtodrugpoisoninginenglandandwales/2022registrations. [Accessed 16/07/2024]

77. Omnicalculator. (2024). *t-test Calculator | Formula | p-value*. [online] Available at: https://www.omnicalculator.com/statistics/t-test. [Accessed 12/08/2024]
78. ONS (2021). *1998 to 2022 edition of this dataset*. [online] www.ons.gov.uk. Available at: https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/gdpandeventsinhistoryhowthecovid19pandemicshockedtheukeconomy/2022-05-24#:~:text=Between%20April%20and%20June%202020. [Accessed 31/07/2024]
79. ONS (2021). *GDP and events in history: how the COVID-19 pandemic shocked the UK economy*. Office for National Statistics. [online] www.ons.gov.uk. Available at: https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/gdpandeventsinhistoryhowthecovid19pandemicshockedtheukeconomy/2022-05-24#:~:text=Between%20April%20and%20June%202020. [Accessed 31/07/2024]
80. ONS (2022). *Adult smoking habits in the UK.* Office for National Statistics. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2021#:~:text=In%202021%2C%20England%20had%20the. [Accessed 31/07/2024]
81. ONS (2022). *Figure 3: The proportion of current smokers by Local Authority of the UK.* [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2021#:~:text=In%202021%2C%20England%20had%20the. [Accessed 31/07/2024]
82. ONS (2022). *Sex - Office for National Statistics.* [online] Available at: https://www.ons.gov.uk/datasets/TS008/editions/2021/versions/4. [Accessed 25/07/2024]
83. ONS (2023). *How we collect and use data at the ONS*. Office for National Statistics. [online] Available at: https://www.ons.gov.uk/aboutus/usingpublicdatatoproducestatistics/howwecollectandusedataattheons#:~:text=We%20are%20the%20largest%20independent. [Accessed 25/07/2024]
84. ONS (2023). *Male and female populations*. [online] www.ethnicity-facts-figures.service.gov.uk. Available at: https://www.ethnicity-facts-figures.service.gov.uk/uk-

population-by-ethnicity/demographics/male-and-female-populations/latest/. [Accessed 03/08/2024]

85. ONS (2024). *Figure 2: Percentage distribution of mid-2022 UK population estimates, by broad age bands and Local Authority* [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2022#:~:text=The%20median%20age%20of%20the%20population%20in%20the%20UK%20was. [Accessed 25/07/2024]
86. ONS (2024). *Figure 3: Population density (people per sq. kilometre) and median age estimates for mid-year 2022 and mid-year 2011* [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2022#:~:text=The%20median%20age%20of%20the%20population%20in%20the%20UK%20was. [Accessed 25/07/2024]
87. ONS. (2024). *Home - Office for National Statistics*. [online]. Available at: https://www.ons.gov.uk/ [Accessed 03/08/2024]
88. ONS. (2024). *Lower Tier Local Authority to Upper Tier Local Authority (December 2022) Lookup in EW.* [online] www.data.gov.uk. Available at: https://www.data.gov.uk/dataset/d7d58c6a-96e5-4af6-aea5-cfdff2e67794/lower-tier-local-authority-to-upper-tier-local-authority-december-2022-lookup-in-ew [Accessed 02/08/2024].
89. ONS. (2024). *LSOA (2021) to SICBL to ICB to LAD (April 2023) Lookup in EN*. [online] www.data.gov.uk. Available at: https://www.data.gov.uk/dataset/c90a701f-e240-4633-a6ff-f677746a6d3f/lsoa-2021-to-sicbl-to-icb-to-lad-april-2023-lookup-in-en [Accessed 02/08/2024].
90. ONS. (2024). *Regional gross domestic product Local Authorities*. Office for National Statistics. [online] Available at: https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductlocalauthorities. [Accessed 31/07/2024]
91. Open Geography Portal. (2022). *Local Authority District to Region (December 2022) Lookup in EN.* [online] Available at: https://geoportal.statistics.gov.uk/datasets/78b348cd8fb04037ada3c862aa054428/explore. [Accessed 02/08/2024].

92. OpenPrescribing. (2024). *BNF 4.7.2: Opioid analgesics*. OpenPrescribing. [online] Available at: https://openprescribing.net/bnf/040702/. [Accessed 02/08/2024].

93. Potters, C. (2019). *Variance Inflation Factor Definition*. [online] Investopedia. Available at: https://www.investopedia.com/terms/v/variance-inflation-factor.asp. [Accessed 14/08/2024]

94. Qualtrics. (2024). *What is ANOVA (Analysis Of Variance)*. [online] Available at: https://www.qualtrics.com/en-gb/experience-management/research/anova/?rid=ip&prevsite=en&newsite=uk&geo=GB&geomatch=uk. [Accessed 16/08/2024]

95. Rajabi, A., Dehghani, M., Shojaei, A., Farjam, M. and Motevalian, S.A. (2019). *Association between tobacco smoking and opioid use: A meta-analysis*. Addictive Behaviors, 92, pp.225–235. doi: https://doi.org/10.1016/j.addbeh.2018.11.043.

96. Schneider, P. (2022). *Mean Absolute Error - an overview*. ScienceDirect Topics. [online] www.sciencedirect.com. Available at: https://www.sciencedirect.com/topics/engineering/mean-absolute-error. [Accessed 04/08/2024]

97. ScienceDirect. (2024). *Learning Rate - an overview*. ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/computer-science/learning-rate. [Accessed 05/08/2024]

98. scikit-learn developers (2019). *sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. [Accessed 14/08/2024]

99. scikit-learn. (n.d.). *sklearn.decomposition.PCA*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#:~:text=Principal%20component%20analysis%20(PCA). [Accessed 13/08/2024]

100. Sharma, N. (2021). *K-Means Clustering Explained*. [online] neptune.ai. Available at: https://neptune.ai/blog/k-means-clustering. [Accessed 15/08/2024]

101. Sharma, P. (2022). *Basic Introduction to Loss Functions*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/08/basic-introduction-to-loss-functions/. [Accessed 04/08/2024]

102. Sport England. (2023). *Active Lives Adult Survey May 2019-20 Report. Tables 1-3 Levels of Activity* [online] Available at: https://www.sportengland.org/research-and-data/data/active-lives/active-lives-data-tables. [Accessed 31/07/2024]

103. Sport England. (2023). *Active Lives data tables*. [online] Available at: https://www.sportengland.org/research-and-data/data/active-lives/active-lives-data-tables. [Accessed 31/07/2024]

104. Stanford.edu. (2021). *'n'-Dimensional Euclidean Distance*. [online] Available at: https://hlab.stanford.edu/brian/euclidean_distance_in.html. [Accessed 04/08/2024]

105. Stanford.edu. (2024). *Measuring Dis/Similarities*. [online] Available at: https://hlab.stanford.edu/brian/making_measurements.html [Accessed 04/08/2024].

106. Stannard, C. (2013). *Opioids in the UK: what's the problem?* BMJ, 347(aug15 1), pp.f5108–f5108. doi: https://doi.org/10.1136/bmj.f5108

107. Stattrek.com. (2022). *Residual Analysis in Regression*. [online] Available at: https://stattrek.com/regression/residual-analysis. [Accessed 13/08/2024]

108. TensorFlow. (2024*). tf.keras.callbacks.EarlyStopping.* TensorFlow Core v2.4.0. [online] Available at: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping. [Accessed 05/08/2024]

109. TensorFlow. (2024). *tf.keras.callbacks.ReduceLROnPlateau*. TensorFlow Core v2.1.0. [online] Available at: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/ReduceLROnPlateau. [Accessed 05/08/2024]

110. Third Space Learning. (2024). *Interquartile range*. [online] Available at: https://thirdspacelearning.com/gcse-maths/statistics/interquartile-range/. [Accessed 23/08/2024]

111. Turney, S. (2022). *Pearson Correlation Coefficient (r).* [online] Scribbr. Available at: https://www.scribbr.com/statistics/pearson-correlation-coefficient/. [Accessed 05/08/2024]

112. University of Wisconsin. (2024). *Normality.* [online] sscc.wisc.edu. Available at: https://sscc.wisc.edu/sscc/pubs/RegDiag-R/normality.html. [Accessed 14/08/2024]

113. UPMC | Life Changing Medicine. (2024). *Acute Pain Causes, Symptoms, and Treatments* | *UPMC.* [online] Available at: https://www.upmc.com/services/pain-management/conditions/acute-pain#:~:text=Acute%20pain%20is%20sudden%20or. [Accessed 21/07/2024]

114. van Draanen, J., Tsang, C., Mitra, S., Karamouzian, M. and Richardson, L. (2020). *Socioeconomic marginalization and opioid-related overdose: A systematic review.* Drug and Alcohol Dependence, 214, p.108127. doi: https://doi.org/10.1016/j.drugalcdep.2020.108127.

115. Watts, V. (2022). *13.6 Testing the Regression Coefficients.* ecampusontario.pressbooks.pub. [online] Available at: https://ecampusontario.pressbooks.pub/introstats/chapter/13-6-testing-the-regression-coefficients/. [Accessed 12/08/2024]

116. WebMD. (2023). *What Is Visceral Pain?* [online] Available at: https://www.webmd.com/pain-management/what-is-visceral-pain. [Accessed 17/08/2024]

117. WHO (2023). *Opioid overdose*. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/opioid-overdose#:~:text=Among%20them%2C%20about%2060%20million. [Accessed 21/07/2024]

118. Wolfert, M.Z., Gilson, A.M., Dahl, J.L. and Cleary, J.F. (2010). *Opioid Analgesics for Pain Control: Wisconsin Physicians' Knowledge, Beliefs, Attitudes, and Prescribing Practices. Pain Medicine*, 11(3), pp.425–434. doi: https://doi.org/10.1111/j.1526-4637.2009.00761.x.

119. Yale University (2019). *Linear Regression*. [online] Yale.edu. Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm. [Accessed 12/08/2024]

120. Yang, H., 2018. *Data preprocessing*. [online] Pennsylvania State Univ. Citeseer. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=53fef985237ae14efddeaf202d44c35ce714d8e2 [Accessed 23/07/2024]

121. Yang, T.-C., Kim, S. and Shoff, C. (2021), *Income Inequality and Opioid Prescribing Rates: Exploring Rural/Urban Differences in Pathways via Residential Stability and Social Isolation.* Rural Sociology, 86: 26-49. https://doi.org/10.1111/ruso.12338

122. Zach (2021). *Understanding the t-Test in Linear Regression*. [online] Statology. Available at: https://www.statology.org/t-test-linear-regression/. [Accessed 12/08/2024]