DEPARTMENT OF ENGINEERING MATHEMATICS

Reasons for Vaccine Hesitancy in England

Joshua Bibby

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Sunday 12th September, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Joshua Bibby, Sunday 12th September, 2021

Contents

List of Figures

List of Tables

List of Listings

Abstract

Hesitancy to get a vaccine has become an even more pressing issue during the COVID-19 pandemic. Understanding what kind of people are unwilling to get vaccinated and their reasons behind it is a key issue to resolve in order for us to exit this current crisis.

The main aim of this project was to explore vaccine hesitancy in England. Exploring what factors most greatly influence vaccine hesitancy to be able to better target vaccine sceptics and ultimately improve vaccine coverage. We will explore vaccine uptake throughout England in 2021 and compare them to factors such as age, ethnicity and deprivation metrics.

Our data is taken mainly from the Public Health England's (PHE) daily and weekly COVID-19 vaccination reports, as well as from the National Immunisation Management Service (NIMS), LG Inform and the Office for National Statistics (ONS).

> My research hypothesis is that (when using the PHE's COVID-19 vaccinations data set) suitable analysis will yield results on the most important factors that influence whether a person will be unwilling to take a vaccine.

Main points of achievement:

- We spent many hours collating and cleaning vaccine data in preparation for further analysis.
- We wrote a total of 3000 lines of *Python* source code.
- We defined a new hesitancy measure for computing at which date and value does vaccine uptake start to asymptote in an area.
- We implemented algorithms to explore this hesitancy measure in comparison to multiple demographic factors.

Supporting Technologies

- I used the *Matplotlib*,*Pandas* and *Seaborn* public-domian Python Libraries.
- I used the *Skilearn* Python Library to perform my data analysis.
- I used LaTeX to format my thesis, via the online service *Overleaf*.

Notation and Acronyms

PHE	:	Public Health England
NIMS	:	National Immunisation Management Service
ONS	:	Office of National Statistics
IMD	:	Index of Multiple Deprivation
LTLA	:	Lower Tier Local Authority
LSOA	:	Lower Layer Super Output Area
NHS	:	National Health Service
WHO	:	World Health Organisation
SVM	:	Support Vector Machine
RNN	:	Recurrent Neural Network
LSTM	:	Long Short-Term Memory
PPE	:	Personal Protective Equipment

Acknowledgements

I would like to acknowledge the endless support and patience of my friends and family as well as the insightful inputs of my supervisor, Dr. Leon Danon.

Chapter 1

Introduction

This paper contains the following chapters:

- We start with the **Introduction** chapter where we will talk about what we are going to do, gain familiarity with the topic, and talk about the challenges we hope to tackle.
- We'll then move to the **Background** chapter where we will look in more detail at the various issues we are dealing with and the methods/tools we may use to tackle them.
- After this we'll move on to the **Execution** chapter where we will present the analysis we performed and any results we created.
- Next we have the **Critical Evaluation** chapter where we will evaluate what we did; discussing the limitations of, decisions in, and potential improvements to, our analysis process.
- And finally comes the **Conclusion** chapter where we will discuss the work we have presented in this paper as well as any recommendations for future work that could be conducted on this topic.

Please visit our GitHub page to see our code and additional materials [19].

1.1 COVID-19

The COVID-19 pandemic has had a devastating effect on the world. At the time of writing (August 2021) there have reportedly been over 200 million cases and 4.5 million deaths worldwide [33]. COVID-19 has had a massive impact on lives, crippled healthcare systems and has reportedly been the causes of over 500 excess deaths per 100k people in some countries [35]. There is no doubt that COVID-19 is one of the major crises of our time.

A crisis of this size required a equally sized response. Many systems and policies have been put in place across the globe in an attempt to lessen the effects of this pandemic. Lock downs, social distancing, working from home, face masks, personal protective equipment (PPE), self-isolation, vaccines and many more. All these protective measures have had knock on effects on not only people's physical and mental health [38, 44], but also on the health of economies across the globe. We need to find solutions to this crisis, a way to stop the pain, suffering and death of our populations, and vaccines may be one of these solutions.

1.2 Vaccines

"Vaccines save lives" [47]. Ever since their inception over 200 years ago, vaccines have helped see the eradication or control of multiple once terrible and deadly diseases, Smallpox, Polio, Tetanus, just to name a few [27]. According to the WHO, vaccines currently prevent 4-5 million deaths every year and, "is one of the most successful and cost-effective public health interventions." Not only that, but they say an additional 1.5 million deaths could be avoided, if global vaccination coverage improves [18].

There have been many successful global vaccination efforts happening now and in the past few decades [13], making the reasons why certain populations still have low vaccination coverage ever more poignant. Some of the reasons for low vaccine coverage include; low vaccine availability, high vaccine dose costs, poor

vaccine initiatives/roll-outs by governments, people's unwillingness to receive a vaccine. These reasons and more are often intertwined, and make the issue of low vaccine coverage a highly convoluted issue that will be complex to address. In the next section we will discuss vaccine hesitancy; what it is, current research on the topic, and how we can tackle it.

1.3 Vaccine Hesitancy

People's unwillingness to receive a vaccine, more commonly called *Vaccine Hesitancy* is a very complex issue as well as a very important issue during this current pandemic. *Vaccine Hesitancy* is described by the WHO as, "the reluctance or refusal to vaccinate despite the availability of vaccines" [14]. Pre-COVID studies have been conducted that form their analysis around a 5C model for vaccine hesitancy [39, 40]:

- Confidence (trust in the vaccines)
- Complacency (not perceiving diseases as high risk)
- Constraints (structural and psychological barriers)
- Calculation (engagement in extensive information searching)
- Collective responsibility (willingness to protect others)

These (and similar [45, 42, 50]) studies have usually been theoretical or empirical in nature, gathering feedback from questionnaires or surveys from various populations. This 5C model is suitable for analysing a person's individual opinion or reasons for their vaccine hesitancy but in this paper we will explore the potential deprivation factors that lie behind a person's vaccine hesitancy. Does a person's education level play an important role in determining their vaccine hesitancy level? Or maybe the levels of crime or poverty in the area where they live? These are the kinds of factors will we be exploring here.

Vaccine hesitancy is a very complex issues that has to be tackled with great care and consideration, no matter what angle it is being approached from. The reasons behind someone's hesitancy is likely to be multifaceted and can't necessarily be boiled down to one key issue. When discussing, exploring and analysing people who are vaccine hesitant, it is important to do it with no judgement, the upmost respect and the upmost sensitivity for the person, their beliefs and the data surrounding them.

1.4 Challenges

All that we have discussed above relates just in the same way to the COVID-19 vaccines. The speed at which each of these vaccines were created and the attention they have received in the media has lead to much debate over the "ethical, legal and practical considerations" of COVID-19 vaccines [11]. This is meant there has been high individual interest around the topic of vaccinations which has led to the spread of lots of information on the topic as well as a spread of a lot of misinformation on the topic.

Misinformation in our current information age is an area of growing concern, especially when looking at research said to be from within the academic community. It has lead to a growing mistrust in science and scientific research and more specifically, a growing mistrust of vaccines and vaccine research. All of this and more has meant vaccine hesitancy has become a very important issue as of late, an issue that needs to be tackled quickly, effectively and with much collaborative effort if we are going to be able to see benefits during this current pandemic.

1.5 Goals of this paper

In this paper we will be specifically looking at England. Data describing COVID-19 and deprivation measures released by the UK government concerning the areas of England. This analysis could obviously be expanded to other areas or countries in a project with larger scope. We would like to thank the UK government and the associated organisations for public access to the data we use in this paper, without which ours and many similar analysis projects would not be possible.

We will be using data surrounding COVID-19 which is gathered by the UK government agencies on a daily and weekly basis [7]. As well as data from the English Indices of Deprivation report which was last updated in 2019 and covers various measures of deprivation for areas within England.

Below we list research questions and objectives of this paper:

Aims:

1. Do factors such as age or ethnicity play a large role in predicting an areas' overall vaccine hesitancy?

2. What are the most influential measures of deprivation when predicting an areas' overall vaccine hesitancy?

Objectives:

1. We will gather publicly available data on COVID-19 cases, deaths and vaccinations for the local authorities of England, as well as data that describe the various levels of deprivation (and other demographic factors) in areas of England.

2. Preprocess and plot our data to get an initial idea of how our data looks.

3. Apply machine learning model to help us figure out/predict the most important indicators of different levels of vaccine hesitancy.

Chapter 2

Background

In this chapter we will explain the key concepts and rationale that lead us to analysing this particular topic. We will explain what challenges we are tackling, in what area we are working, the relevance of our work and the methods we will apply to tackle these challenges. We will also define any concepts or techniques that we will use in this paper to ensure the paper is as self-contained as possible.

2.1 COVID-19

2.1.1 What is COVID-19?

The Coronavirus disease is an infectious disease cause by the SARS-CoV-2 virus, which arose in late 2019 in Wuhan, China [46]. The Coronavirus disease has gone through many naming conventions, initially called *the 2019 novel coronavirus* and more recently as *the virus responsible for COVID-19*, *the COVID-19 virus* and now most commonly just, *COVID-19* [23].

2.1.2 The personal impacts of COVID-19

COVID-19 manifests as mainly a respiratory illness but its symptoms can vary widely in severity and longevity in each case. COVID-19 has been reported to affect older generations more severely but also people with underlying respiratory conditions.

The National Health Service (NHS) lists the mains symptoms of COVID-19 as [22]:

- a high temperature.
- a new, continuous cough.
- a loss or change to your sense of smell or taste.

The World Health Organisation (WHO) lists the mains symptoms of COVID-19 as [3]:

- Fever.
- Cough.
- Tiredness.
- Loss of taste or smell.

It's important when performing our data analysis that we remember the extremely devastating effects COVID-19 can and has had on people and their families since its initial discovery. And we do our analysis to help understand past impacts and lessen future impacts of COVID-19 and similar outbreaks.

Since its initial discovery, the disease has spread and turned into a global pandemic. In the next section we will discuss the global impact of COVID-19.

2.1.3 The global impacts of COVID-19

At the time of writing (August 2021), the coronavirus has reportedly caused 133000 deaths in the UK [10] and over 4.5 million deaths worldwide [6]. COVID-19 has had huge health, economic and societal impacts on people all across the globe. Countries are continually experiencing new waves of COVID-19 cases, especially with new variants emerging. These large impacts have caused major shifts in economic and societal norms in many countries; what people buy and how they spend their time has changed significantly.

In January 2020 the WHO declared the novel coronavirus outbreak a, "public health emergency of international concern (PHEIC), WHO's highest level of alarm" [29]. In the next section we will discuss how people, countries and governments have responded to the COVID-19 pandemic.

2.1.4 The response to COVID-19

The monumental impact of COVID-19 has also caused a monumental response from the global community. If anything good can be taken from this pandemic, its the timely and sizable response people have had to this pandemic. COVID-19 has had a horrible effect on the world and it's through the efforts of people, communities, governments and other organisations (scientific or otherwise) that we are starting to see the light at the end of the tunnel.

People and communities have put in a tremendous amount of effort into helping the most vulnerable, be it offering to buy them food, be their taxi or simply to provide company. People are at the root of the positive response to this pandemic and they should be at the forefront of anyone's mind when discussing the response to this pandemic.

Governments have had to take quick and drastic action to lessen the impact of this pandemic on people's lives and their economies. Social distancing rules were enforced as well as lockdowns during periods in which cases numbers were rising sharply. All this limiting of people's actions predictably had major impacts on global economies. It was predicted that the coronavirus would cost the UK government nearly £400 billion in 2020/21 alone [5]. These extreme measures were in an effort to not only curtail the effects on COVID-19 but to also keep the economy afloat during the periods of harsh restrictions.

It is clear that COVID-19 has had a major effect on people's health as well as the health of global economies. There has been a concerted effort to get back to an old norm and consequently improve people's and economy's health, and this effort was aided in no small effort by the development and roll-out of vaccines. In the next section we will discuss vaccines, their roll-out during the coronavirus pandemic and why people might be unwilling or hesitant to receive a vaccine.

2.2 Vaccines and Vaccine Hesitancy

In this section we will discuss vaccines, their history, their development for COVID-19 and why people may be hesitant or unwilling to receive a vaccine.

2.2.1 What is a vaccine?

The WHO describes a vaccine as, "a simple, safe, and effective way of protecting you against harmful diseases" [32]. Vaccines are essentially a way to train your immune system against a disease by offering it a killed or weakened version of the virus and allowing your body to develop antibodies against this disease without having to contract the disease itself.

Vaccines have had a major impact on global health and people's fight against deadly diseases, namely the development of a smallpox vaccine lead to the eradication of the disease. In the next section we will review a brief history of vaccines and the major impacts they have had on us all.

2.2.2 A brief history of vaccines

The first evidence of vaccines date back to as early as 1000 CE, when smallpox material was used to aid inoculation against the disease. More commonly known is the breakthrough of Edward Jenner in 1796 when he showed how an injection with cowpox (an illness among cattle) could help protect a person against smallpox, further developments and changes to this vaccine eventually lead to the eradication of smallpox. The next major impact in the history of vaccine came when Louis Pasteur developed a vaccine for rabies in 1885.

By the end of the 19th century, with the greater ability to grow virus cultures in a laboratory, many more vaccines and vaccine research has been developed to the point where an extremely wide variety of illnesses and conditions are being researched [15]. The WHO currently states that vaccines exist that help protect against many diseases, including [31]:

- Cervical cancer
- Cholera
- COVID-19
- Diphtheria
- Ebola virus disease
- Hepatitis B
- Influenza
- Japanese encephalitis
- Measles
- Meningitis
- Mumps
- Pertussis
- Pneumonia
- Polio
- Rabies
- Rotavirus
- Rubella
- Tetanus
- Typhoid
- Varicella
- Yellow fever

2.2.3 A brief history on the development of COVID-19 Vaccines

The rapid expansion of vaccine research over the past 200 years has meant vaccine developers have not come completely unprepared for the arrival of COVID-19.

The way out of the COVID-19 pandemic was through the development of a COVID-19 vaccine. The genetic sequence of the coronavirus was first release in January 2020 [53]. With this knowledge there was concerted effort and collaboration amongst different health authorities to quickly and safely develop COVID-19 vaccines. As of December 2020, the UK medicine regulator had authorised use of the Oxford/AstraZenica and the Pfizer–BioNTech COVID-19 vaccines, as well as the Moderna vaccine [24].

With the development and approval of these vaccines in the UK, next came their distribution throughout the population. In the section we will briefly discuss how the COVID-19 vaccines are being rolled out in the UK.

2.2.4 The roll-out of COVID-19 vaccines

The UK's mass COVID-19 vaccination programme began in December 2020, when 90 year old Margaret Keenan received her first vaccine. Since then a schedule has been set up to decide who will receive their vaccines first. The priority list was developed by the government's Joint Committee on Vaccination and Immunisation. this list started with health workers and care home residents, then clinically extremely vulnerable and older people. As the months went on progressively younger age groups were encourage to come forward and receive their vaccines [36]. In the UK, at the time of writing (August 2021), 88.5% of the population have received their first doses of a COVID-19 vaccine and 78.5% have received their second [30].

These data do seem to be asymptoting however and brings up the question of why someone might be unwilling to receive a vaccine. In the next section we will discuss vaccine hesitancy, the circumstances around why someone may be unwilling to receive a vaccine.

2.2.5 Vaccine Hesitancy

The Sage working group on vaccine hesitancy defines it as follows [25]:

> ***Definition: Vaccine Hesitancy***
>
> *Vaccine hesitancy refers to delay in acceptance or refusal of vaccines despite availability of vaccine services. Vaccine hesitancy is complex and context specific, varying across time, place and vaccines. It is influenced by factors such as complacency, convenience and confidence.*

Vaccine hesitancy is a very complex and case specific topic. Vaccine hesitancy is often intertwined with vaccine availability. Vaccine hesitancy may be present in an area even if it's not the driving forcing behind low uptake, such as in the presence of low vaccine available.

Vaccine hesitancy is often spoken of as a continuum. A person won't simply definitely or definitely not get a vaccine, there are often a multitude of reasons someone may have for not getting a vaccine. Ignoring the factors of eligibility and availability, vaccine hesitancy often comes down to a combinations of confidence, convenience and complacency; not believing the vaccine will work, getting the vaccine will cause too much disruption in their life, believe they don't need the vaccine as it won't help them personally [34].

Understanding why someone might choose not to receive a vaccine is a complicated issue. Understanding the type of people who may choose not to receive a vaccine is an equally complex issue and is something we hope to shed some light on in this paper. In the next section we discuss the English Indices of Deprivation and how we might use it to help decide the kinds of areas that suffer the most from vaccine hesitancy.

2.3 English Indices of Deprivation

All information in this section is taken from, "The English Indices of Deprivation 2019 (IoD2019)" [12]. The English Indices of Deprivation looks at various deprivation measures across small areas of England. Local measures of Deprivation in England have been collected since the 1970s and new reports exploring local area's levels of deprivation are usually released every 3-5 years (2019, 2015, 2010, 2007, 2004). Most recently, the Indices of Deprivation (IoD) has been made up of seven major areas:

- Income
- Employment
- Health Deprivation and Disability
- Education, Skills Training
- Crime
- Barriers to Housing and Services
- Living Environment

There are many new acronyms and definitions that come with the English Indices of Deprivation and we will be using a lot in this paper. In the next sections we will discuss; what local areas are used and how they are defined, how the deprivation measures are calculated, a weighted summary measure known as the Index of multiple deprivation, and the measures we will use in this paper. Please note that throughout this paper we will use IMD (Index of multiple deprivation) to refer to any measure we use from the English Indices of Deprivation.

2.3.1 Local areas in England

The deprivation measures presented in the English Indices of Deprivation are calculated at a small neighbourhood level known as a Lower-Layer Super Output Area (LSOA). The Ministry of Housing, Communities & Local Government defines an LSOA as follows:

> "LSOAs are small areas designed to be of a similar population size, with an average of approximately 1,500 residents or 650 households. There are 32,844 LSOAs in England. They are a standard statistical geography and were produced by the Office for National Statistics for the reporting of small area statistics."

To aid using the English Indices of Deprivation for further analysis; aggregate and summary measures are often calculated for the local authority level. There are over 300 of these areas in England which are most commonly known as Lower Tier Local Authorities (LTLAs). These LTLAs are a higher level partition of the areas of England compared to LSOAs.

Finally England is commonly divided into nine larger geographical regions (also called *standard regions*), these include:

- North East
- North West
- Yorkshire
- East Midlands
- West Midlands
- South East
- South West
- East of England
- London

2.3.2 How deprivation measures are calculated

For every deprivation metric, the measure is calculated as a ranking of all the 32844 LSOAs. So, the LSOA in 1st position is the most deprived for the metric, the one in 32844th place is the least. Importantly this means that we can't compare the rankings mathematically. That is to say the LSOA in 10th place is more deprived than an LSOA in 20th place, but is not necessarily twice as deprived. To hammer it home, it terms of deprivation level, 1st place could be closer to 32000th place than 32000th place is to 32844th place. We can only know an area's deprivation level relative to other areas.

2.3.3 Index of multiple deprivation

The English Indices of Deprivation contains over 38 different measures used to describe an area's deprivation, the measures are categorised into seven major themes (mentioned earlier) which are then summarised into a weighted measure called the Index of Multiple Deprivation (IMD). The idea behind this measure is to try and get an overall level of deprivation for a particular area.

The Ministry of Housing, Communities & Local Government defines the Index of Multiple Deprivation as follows [12]:

> "The Index of Multiple Deprivation 2019 combines information from the seven domains to produce an overall relative measure of deprivation. The domains are combined using the following weights: Income Deprivation (22.5%), Employment Deprivation (22.5%), Education, Skills and Training Deprivation (13.5%), Health Deprivation and Disability (13.5%), Crime (9.3%), Barriers to Housing and Services (9.3%), Living Environment Deprivation (9.3%). The weights have been derived from consideration of the academic literature on poverty and deprivation, as well as consideration of the levels of robustness of the indicators."

2.3.4 Local authority level

As mentioned before, aggregation measures are often created to help people understand and analysis the LSOA-level metrics but at a higher level, such as LTLA or geographical region. Now because the English Indices of Deprivation calculates their measures in a ranking system rather than using an absolute scale, we can't simply find the average of the LSOA's measure within an LTLA as this would create inaccurate results. Instead, we have to create metric that allows us to better understand how deprived a LTLA might be whilst also using the ranking system created in the English Indices of Deprivation 2019. In our paper we will make use of the following aggregated metrics [17]:

> **IMD: Overall - Extent (%)** "is a local authority level measure which represents the proportion of an authority's population living in the most deprived LSOAs in the country. This is a weighted measure of the population in the most deprived 30 per cent of all areas: the population living in the most deprived 10 per cent of LSOAs in England receive a 'weight' of 1.0; the population living in the most deprived 11 to 30 per cent of LSOAs receive a sliding weight, ranging from 0.95 for those in the eleventh percentile, to 0.05 for those in the thirtieth percentile."

So, the higher the IMD: Overall - Extent (%) score is the more deprived an area is. Additionally we will make use of the following aggregated metric (note that here we use IMD: Crime as an example but this aggregation measure can be applied to any of the available metrics) [16]:

> **IMD: Crime** "is a local authority level measure which represents the proportion of an authority's LSOAs that fall in the most deprived 10% of LSOAs nationally."

So again, the larger the IMD: Crime measure is the more deprived an area is for that category. We now have an understanding of and access to COVID-19 data for England and its local authorities as well as an understanding of and access to deprivation measures for their local authorities. The next stage is to explore the methods and tools we can employ to help us analysis these data and produce some meaningful results. In the following sections we will look at what methods and tools have been applied in similar scenarios to ours and how we might apply these during our analysis process.

2.4 Data Analysis

The technological age we live has meant this pandemic can be tackled in a way that a pandemic never has been before. The instant communication and information sharing potential of our age has meant we can confront this problem with much more pace, communication and collaboration. This together with the ever-expanding field of data science has meant predictions, and therefore policy decisions, can be made in a very timely fashion.

A brief history of data analysis

One could argue that data analysis has been around for as long as statistics itself, dating back to ancient Egypt. As the technological era arose, the speed at which data analysis could be performed increased rapidly. The US census of 1880 took seven years to complete and it wasn't until the invention of computers that data analysis really flourished [1].

The analysis of data has been around for centuries but it wasn't until the 1970s, when the term "data science" was first coined, that it really started to show its full potential [41]. Now as the speed and availability of computers has rapidly increased, so has the capabilities of data science.

Data analysis and COVID-19

The potential of data science has really shone during this devastating pandemic. With the increased power and speed of computers, data scientists were able to create data sets and results at much greater speeds. Not only would this data science ability increase the speed at which, say, vaccines were created but also the speed at which governments were able to see the potential risk and predictions of the pandemic in the near future. Not only the risks but also the possible decisions that can be made to lessen the effect of COVID-19 on their population, potential saving many lives. As horrible as this pandemic is, the power of data scientists has really shone through in a way that has never been seen before. The work of people and organisations, such as the Turing Institute, in producing collaborative and timely results during this pandemic has been exemplary [9].

That being said, many issues (both good and bad) have been brought to the forefront in the field of data science during this pandemic. Firstly, the decisions in response to this pandemic need to be made quickly, and in order to do that we need accurate and timely data. Accuracy and speed are often the opposite sides of a coin, the quicker something is released often means the less accurate it is. Governments have made efforts to make important data available quickly during this pandemic, such as the *Control of patient information (COPI) notice* introduced in the UK by the secretary for health and social care to allow certain organisations access to confidential patient information to help in the response to COVID-19 [2]. This unprecedented access to vital data has lead to many important and timely results such as the creation of data analysis platforms such as OpenSafely [26], but there have still been reports of problems around, "data availability, access and standardisation" [9].

Secondly comes the issue of communication. Creating predictive results in such a timely fashion inevitably creates uncertainties. Communicating these uncertainties to policy makers and the general public whilst keeping their confidence in the results can be a very difficult issue. And then comes the issue around communication between researchers. The ability to perform effective 'Open Science' increases the speed at which research is conducted, especially important in the response to the COVID-19 pandemic [54].

Thirdly comes the issue of inequality. As with many of things, the pressures of the pandemic has brought the issue of inequality in data science to the forefront. The lack of proper representation of minorities in data sets (and therefore biases in the results and decisions made), the inequalities that already exist within the data science community, and the reported inequalities around who could gain access to important data sets for their research [9].

There are clearly many issues that have risen up in the field of data science during this pandemic and these are issues that we need to be acutely aware of when conducting our own research. Similar issues around inequality and interpretability arise when discussing our next topic, machine learning. In this next section we will discuss what machine learning is and how we can utilise it when performing data analysis.

2.5 Machine Learning Models

2.5.1 What is machine learning?

Machine learning is essentially teaching a computer to do a task. With the advent of machine learning, computers have been able to perform a wide variety of tasks in many different fields both quicker and more accurately than humans. Not everything can be performed by a computer at this stage, but machine learning has shown great potential in the fields of voice and image recognition that has wide ranging practicalities. As the development of machine learning continues, computers will be able to perform more complex and diverse problems. The Royal Society describes machine learning as follows [21]:

> "Machine learning is a form of artificial intelligence that allows computer systems to learn from examples, data, and experience. Through enabling computers to perform specific tasks intelligently, machine learning systems can carry out complex processes by learning from data, rather than following pre-programmed rules."

2.5.2 A brief history of machine learning

Machine learning is largely based on statistical models. So one could argue that the beginnings of machine learning began with the introduction of Bayes' Theorem in 1763 [37]. Continuing from this came the

introduction of Least Square in 1805 and Markov Chains in 1913, both methods still widely used in modern machine learning [28].

As with many of the histories we have looked at, large progress came to the field of machine learning during the 20th century. In the 1950s, Alan Turing first propose his *Learning Machine*, that decade also saw the introduction of concepts such as artificial neurons and neural networks, now widely used in modern machine learning. More recent major announcements include, recurrent neural networks (RNNs), reinforcement learning, random forest, support-vector machines (SVMs), long short-term memory RNNs (LSTM) [28]. All of which are used greatly in the field of machine learning to date, with great potential for further development.

2.5.3 Machine learning and health

With the ever increasingly speed and performance of machine learning models, their use in more and more fields has continue to grow. One of those fields being Health. With the ever increasing amount of data being gathered across many fields, including health, the potential of machine learning's application has really shone.

Recently the benefits of machine learning in health vary from things like, aiding radiologists in the task of MRI and x-ray image recognition [51] and predicting breast cancer risk using personal health data [52]. As well as in many other aspects of predicting people's health, although there is still room to really capture the potential of big data. The increased size of and access to health data sets has really captured the potential of machine learning in the world of health data [49]. And the COVID-19 crisis has no shortage of health data.

2.5.4 Machine learning and COVID-19

As we have already explained greatly; data science, machine learning and health data have combined excellently in helping in the response to the COVID-19 pandemic. The amount of papers published and academic research being conducted around COVID-19 has increased greatly over the past two years and will likely keep increasing in the years to come [43]. Machine learning models have been developed to predict numbers of infections, hospitalisation and deaths from COVID-19 [48], and organisations such as the Turing Institute having invested considerable effort and resources into, "tackling the spread and effects of coronavirus" [29, 4].

Clearly, data scientists can help greatly during this crisis and what we learn from this pandemic will hopefully prepare us and future generations in tackling similar crises, and ultimately protect and save people's lives, which is what we do all this for. In the next section we will outline what we hope to achieve from our analysis.

2.6 Outcomes

In this section we will discuss the kinds of results we want to produce and the kinds of outcomes we might expect from our analysis.

Our analysis will involve exploring the relationships between COVID-19 and deprivation data in England. What we want to discover is what kinds of people or factors effect vaccine hesitancy the most. We will first aim to create a robust metric for defining vaccine hesitancy that isn't affected by issues such as ineligibility or vaccine availability.

We will then, after pre-processing, apply suitable machine learning models to our data in an effort to discover the most important factors when predicting vaccine hesitancy. An exploration of multiple different types of model will likely be required to complete this process adequately and robustly.

An example result might be,

> 'through the application of our machine learning models to our COVID-19 and IMD data we found that Education, Skills Training and Crime were the most important metrics for determining vaccine hesitancy. Access to education appeared as the biggest influencer in the majority of our models. The less educated someone is, the more hesitant they are to receive a vaccine. This analysis was performed at a local authority level.'

We do anticipate education to be a large factor in determining vaccine hesitancy but also understand that this is a very complex topic to tackle and there may be many hidden co-factors influencing our results. We now have a thorough understanding of the areas we're working in and what analysis we want

to perform. In the next chapter we will present the analysis we perform and offer some interpretations of the results.

Chapter 3

Execution

In this section we will look at the various stages of our data analysis process. We analysed COVID-19 vaccine data within the different regions and local authorities of England and compared them to demographic factors and various Index of Multiple Deprivation (IMD) metrics. We first collated and visualised COVID-19 vaccine data in England using data from LG inform and then went to to implement more complex analysis to explore the 'importance' of each factor in determining vaccine hesitancy.

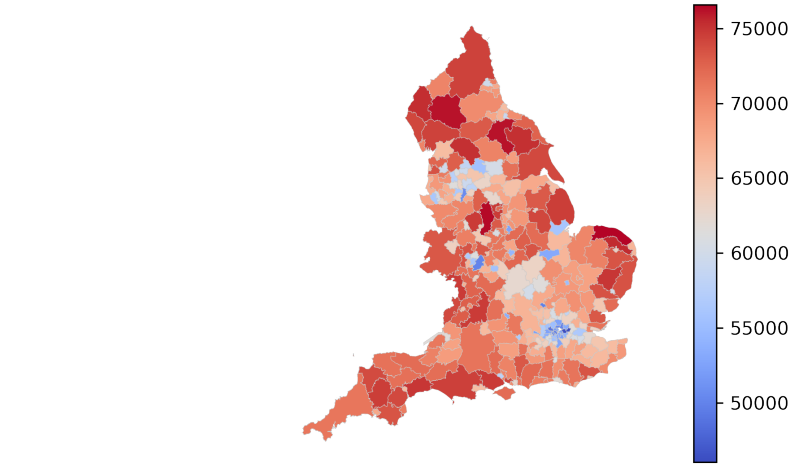
3.1 COVID-19 Data: LG Inform

In this first section we did some initial exploration of COVID-19 data accessed from the Local Government Association's website, LG Inform [20]. The idea behind this initial exploration was to get an idea of how COVID-19 data is distributed throughout England and to practice handling COVID-19 and geographical data.

We accessed data on:

- cumulative first dose numbers
- cumulative second dose numbers
- cumulative confirmed case numbers
- IMD: Overall - extent (%)

3.1.1 First Doses



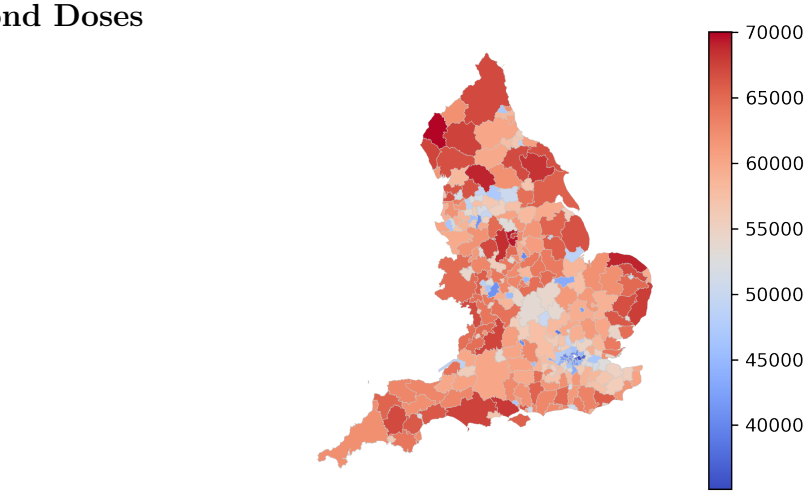
Source: LG Inform

Figure 3.1: Cumulative first dose numbers per 100k population stratified by local authority.

We mapped first dose COVID-19 data onto the local authorities of England. The colour and numbers show the number of people who have received their first dose of a COVID-19 vaccine per 100000 population within each area.

It can be seen that a lower numbers of first vaccine have been administered in the London area as well as in parts of the North West to a lesser extent. It does appear from this that there is a discrepancy between different areas and why this may be is something we will investigate later. There could be multiple reasons for this discrepancy at this stage; people's hesitancy to received the vaccine, people's ineligibility or inability to receive the vaccine, or possibly the local government's speed and spread of their vaccine roll out.

3.1.2 Second Doses



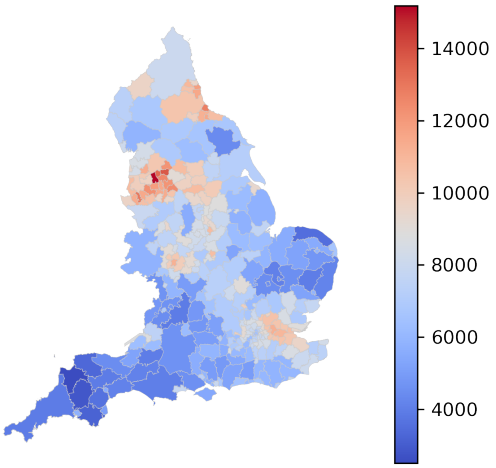
Source: LG Inform

Figure 3.2: Cumulative second dose numbers per 100k population stratified by local authority.

Here we mapped second dose COVID-19 data onto the local authorities of England. The colour and numbers show the number of people who have received their second dose of a COVID-19 vaccine per 100000 population within each area.

This map shows a very similar pattern to that of first doses although obviously at a lower rate. This is understandable as; people's hesitancy to received the vaccine, people's ineligibility or inability to receive the vaccine, or the local government's speed and spread of their vaccine roll out would extend to the second dose as well.

3.1.3 Confirmed Cases



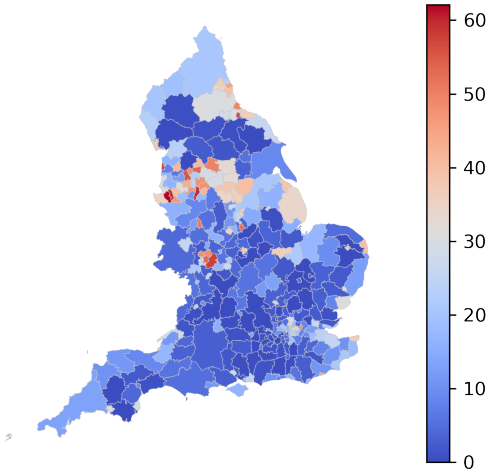
Source: LG Inform

Figure 3.3: Cumulative confirmed case numbers per 100k population stratified by local authority.

Here we mapped data on the number of cumulative confirmed cases of COVID-19 onto the local authorities of England. The colour and numbers show the number of people who have received a positive result of a COVID-19 test per 100000 population within each area.

This map differs from that of the first and second doses maps as it shows a greater number of confirmed cases in parts of the North West but not so much in the London area. The similarities of these maps could lead to questions about the relationship between vaccine and case numbers.

3.1.4 Index of Multiple Deprivation



Source: LG Inform

Figure 3.4: IMD: Overall - Extent (%) stratified by local authority.

Here we mapped data on IMD: Overall - Extent (%) onto the local authorities of England. The colour and numbers show the percentage of Lower Layer Super Output Areas (LSOAs) within each area that are some of the most deprived areas nationally.

This maps shows higher percentages in parts of the North West and less so in the London area. Worse numbers in parts of the North West is a recurring theme in this, the vaccine and the cases maps and leads to questions about the relationship between an area's deprivation and it's vaccine uptake.

3.1.5 Case Rates by Dates

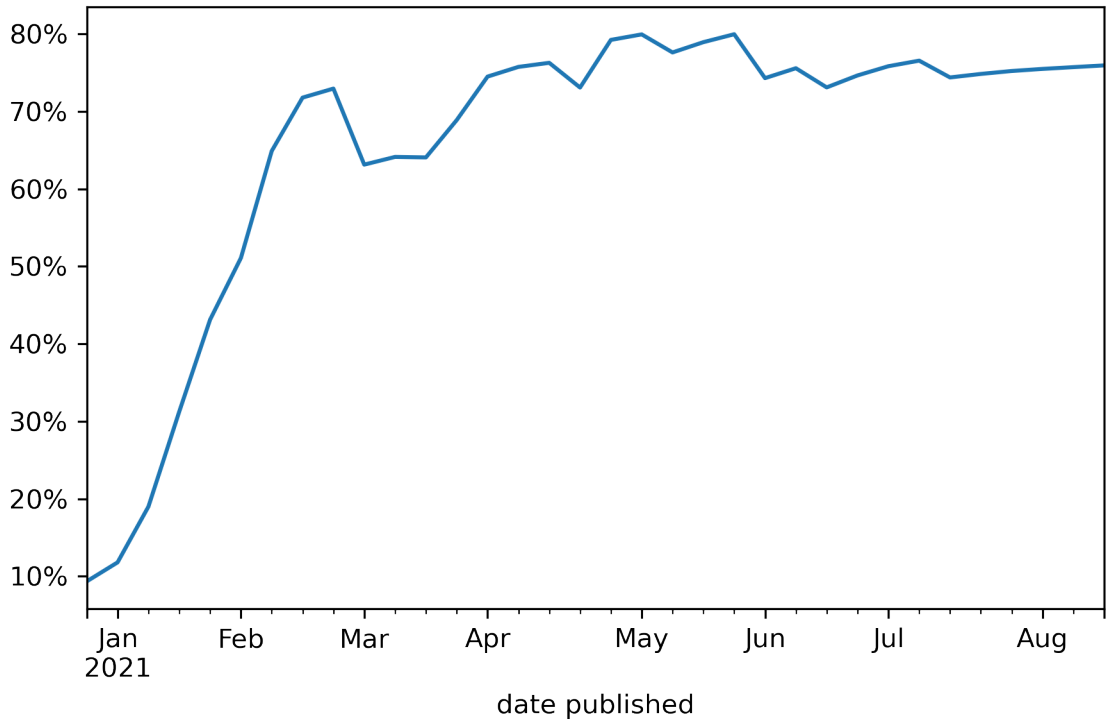


Figure 3.5: Cumulative 1st dose vaccine rate by date.

Here we plotted the percentage of people in England who have received at least one dose of a COVID-19 vaccine. As you can see the plot looks widely incorrect for a cumulative graph. This is because we used ONS bi-yearly population estimates with our weekly data. As the weeks and months went on, additional age ranges were tracked by our data source and, because the population estimates were static, therefore the total percentage of the available data was lowered whenever a new age range (and population estimate for that age range) was tracked and added. For example, in the earlier weeks, only the 80+ age group was tracked, additional age groups were added periodically and by the end multiple age groups were part of the dataset from the 18-24 to the 80+ age groups.

These misleading results and multiple other limitations (such as lack of data from earlier weeks and mismatching vaccine and population data) lead to us looking for more comprehensive data sources. This lead to us continuing our analysis using Public Health England (PHE) vaccine data.

3.2 COVID-19 Data: Public Health England

After out initial exploration of COVID-19 data from LG Inform we look for more comprehensive and complete data sets.

We continue our analysis using data from [8, 12]:

- Public Health England's (PHE) weekly COVID-19 vaccinations
- National Immunisation Management Service's weekly population estimates
- Ministry of Housing, Communities & Local Government's English Indices of Deprivation 2019

3.2.1 PHE: Case Rates by Dates

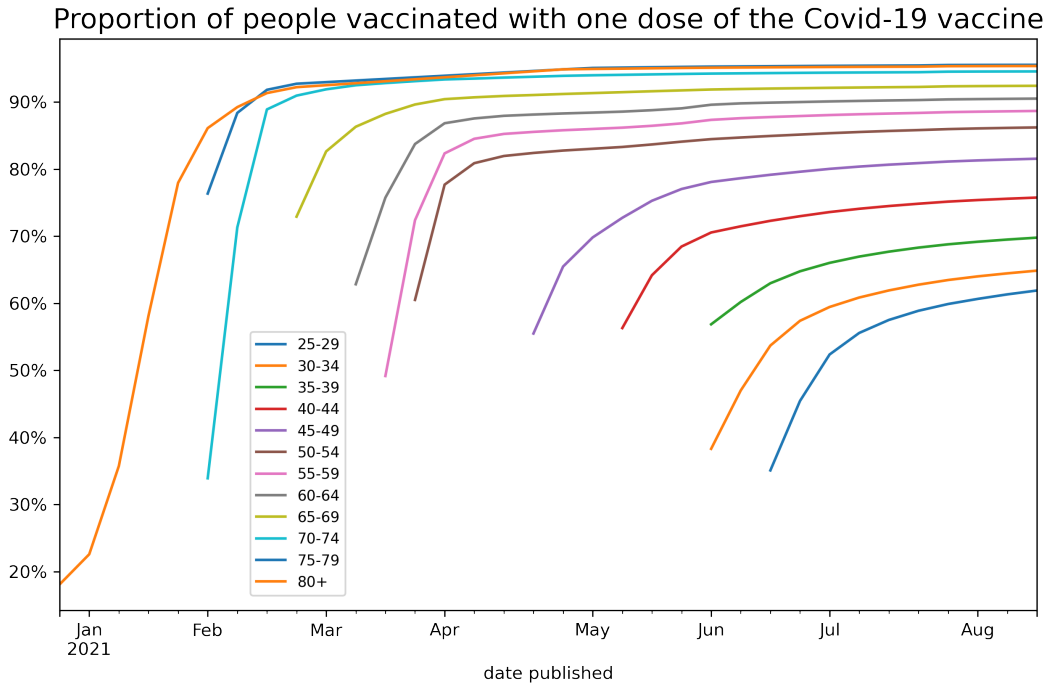


Figure 3.6: Proportion of people vaccinated with one dose of the Covid-19 vaccine.

Here we have plotted the percentage of the English population who have received a COVID-19 vaccination stratified by age group over time. As you can see, different age groups start at different weeks and this is because different age groups' vaccination rates started being recorded by PHE at different times.

As you can see every age groups' percentage starts to asymptote at a certain date and percentage. Eventually the weekly increase in vaccination rates "tails off" to a much slower rate that it never then increases from. This is likely because the vaccination of the population is reaching a natural saturation. That is, it has likely been offered to everyone and the only people who remain unvaccinated are the people who are unable or unwilling to receive the vaccine. The reasons why people would be unwilling to receive a vaccine, more commonly referred to as vaccine hesitancy, is something we will explore in the remainder of this chapter. First we need to find a way to quantify and analyse the point at which the vaccine rate starts to asymptote (or "tail off"/slow down) for a population.

3.2.2 Hesitancy Measure

In this paper we will define this *Hesitancy Measure* (the point at which the data starts to asymptote) as the weekly date and percentage at which the weekly vaccine rate increase drops and stays below 1%. More concretely, for a certain population with data in the form of a *column* of weekly vaccine rates and a value which we decide as the value that it starts to asymptote (the *asymptote_value*), we have the following,

```
import Pandas as pd

def hesitancy_measure(column, asymptote_value):

    reverse_column = list(column[::-1])
    N = -1

    for i,val in enumerate(reverse_column):

        weekly_difference = reverse_column[i]-reverse_column[i+1]

        if weekly_difference < asymptote_value:
            N = i
        else:
            break

    if N == -1:
        return False
    else:
        return column.index[-(N+2)], column[-(N+2)]
```

Listing 3.1: Code for calculating the hesitancy measure.

This code snippet shows how we calculate the hesitancy measure for some temporal data. Essentially we look at the weekly vaccine rate increases starting from the most recent date and working backwards. Once we find a weekly increase that is equal to or above the *asymptote_value* we stop and return the date and value of that week. By doing it this way we find the earliest point at which the weekly vaccine rate increase drops and stays below our *asymptote_value* up to our most recent data.

Our current choice of the *asymptote_value* for our hesitancy measure is 1%. This of course can be changed if more recent data are added or if the method is applied to different data sets. The choice of the *asymptote_value* here could affect later results and therefore further sensitivity analysis could be conducted to find the best choice of *asymptote_value*. In the next sections we will explore some initial results of applying this hesitancy measure.

3.2.3 When do vaccine rates tail off

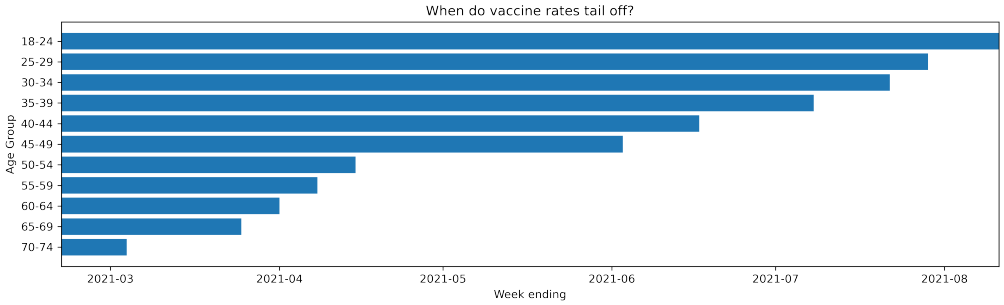


Figure 3.7: Dates at which different age groups' vaccine rates start to "tail off".

Here we have plotted the dates at which the vaccine rates of different age groups asymptote for the entire English population. We can see that older age groups have asymptoted earlier and this is most likely due to the government's initial vaccine roll-out targeting older, more vulnerable, age groups before moving on to younger ones.

3.2.4 At what value do weekly vaccination rates tail off

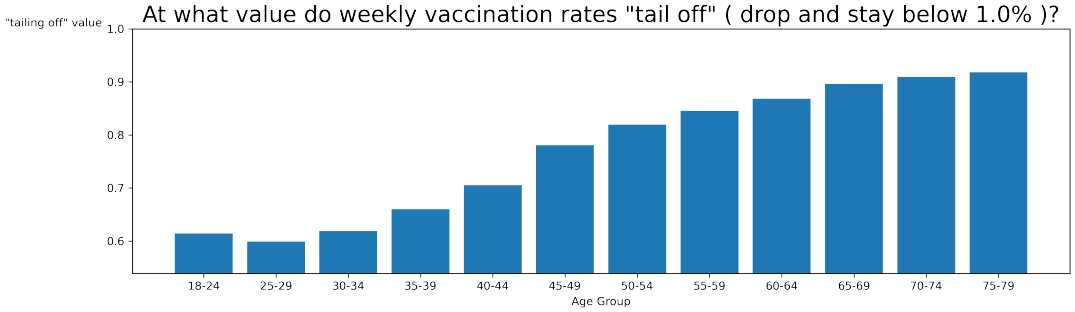


Figure 3.8: Proportion of population at which different age groups' vaccine rates start to "tail off".

Here we have plotted the values at which the vaccine rates of different age groups asymptote for the entire English population. We can see that older age groups seem to asymptote at a higher value. This could be because there was higher urgency to vaccinate older age groups and they have therefore reached actual saturation earlier, whereas there has been less urgency to vaccinate younger age groups. The weekly increase for younger age groups may therefore have started to "tail off" but may be a long way from its "true" saturation. Further analysis of the *asymptote_value* used in this paper could help answer some of these questions brought up by this plot.

However if there is a true discrepancy of the asymptote value for different age groups then that is something worth investigating. What makes people more hesitant? Is it to do with their economic status? Their education level? Access to healthcare? Is it to do with where they live? What kinds of media they are consuming? How they perceive the risks to themselves compared to others? In the next few sections we will look at what kinds of factors are most 'important' when determining an area's vaccine hesitancy.

3.3 COVID-19 Vaccine and the English Indices of Deprivation

In this section we look at the relationship between COVID-19 vaccine data and the English Indices of Deprivation 2019.

3.3.1 IMD metrics vs Cumulative Vaccine rate per 100000 population

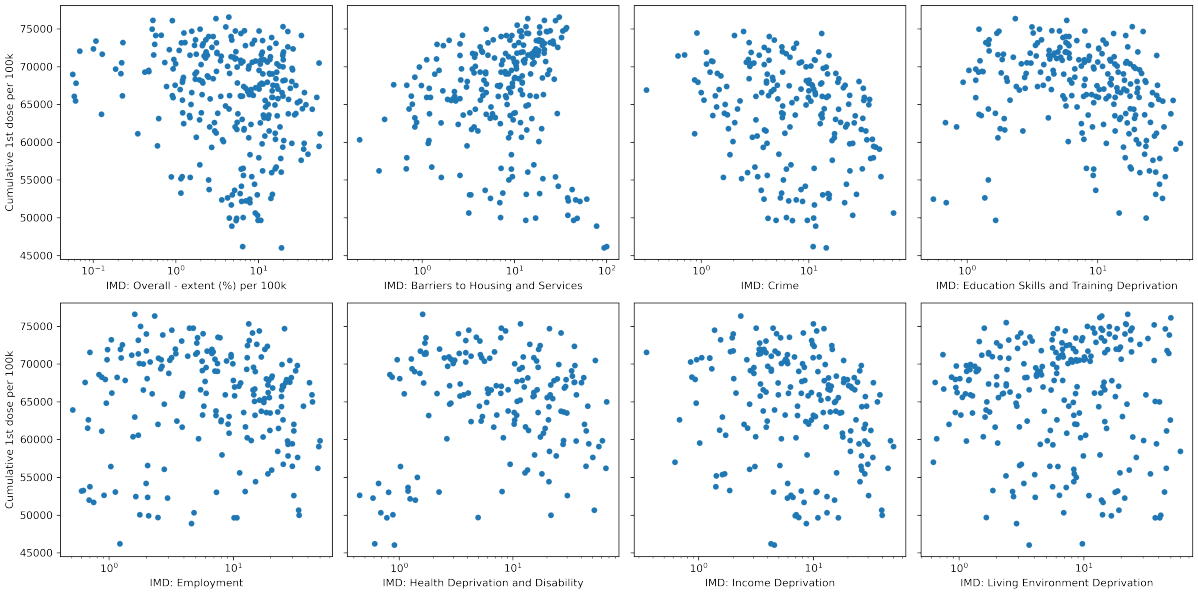


Figure 3.9: Plots of IMD metrics vs vaccination rates per 100k in local authorities of England.

In our first exploration in this section we plot various IMD metrics vs vaccination rates for 100000 people in England stratified by local authority. A small reminder that IMD: Overall - Extent (%) is a local authority level measure that shows the percent of LSOAs in the area that lie among the most deprived LSOAs in England. For the individual IMD metrics the measure is calculated as, the percent of LSOAs in the area that are in the 10% most deprived LSOAs nationally.

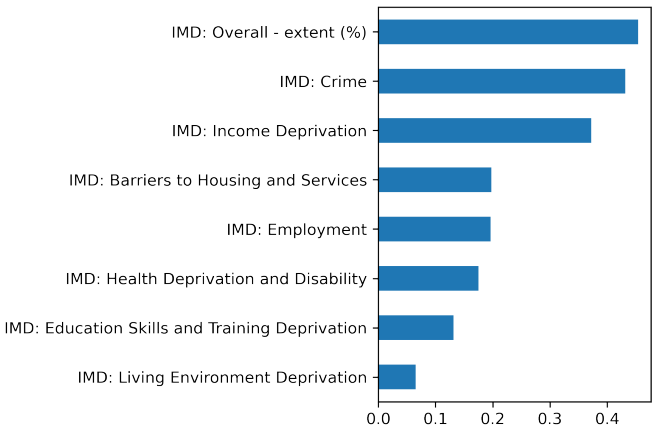


Figure 3.10: Pearsons Correlation coefficient for Cumulative 1st dose per 100k compared to IMD metrics.

We can see from the plots of IMD metrics vs. vaccination rates that there may relationships between vaccine rate and certain IMD metrics. We will also compare the Pearson's correlation coefficient of each IMD metric with the vaccine rate per 100000 people. From this it appears that *IMD: Overall - extent (%)*, *IMD: Crime* and *IMD: Income Deprivation* are the most correlated, with the other metrics lagging behind. This is contrary to our initial belief that *IMD: Education Skills and Training Deprivation* would be ammong the most related metrics. This is our initial exploration but let's also explore our vaccine measure that we defined earlier.

3.3.2 IMD metrics vs Vaccine Hesitancy

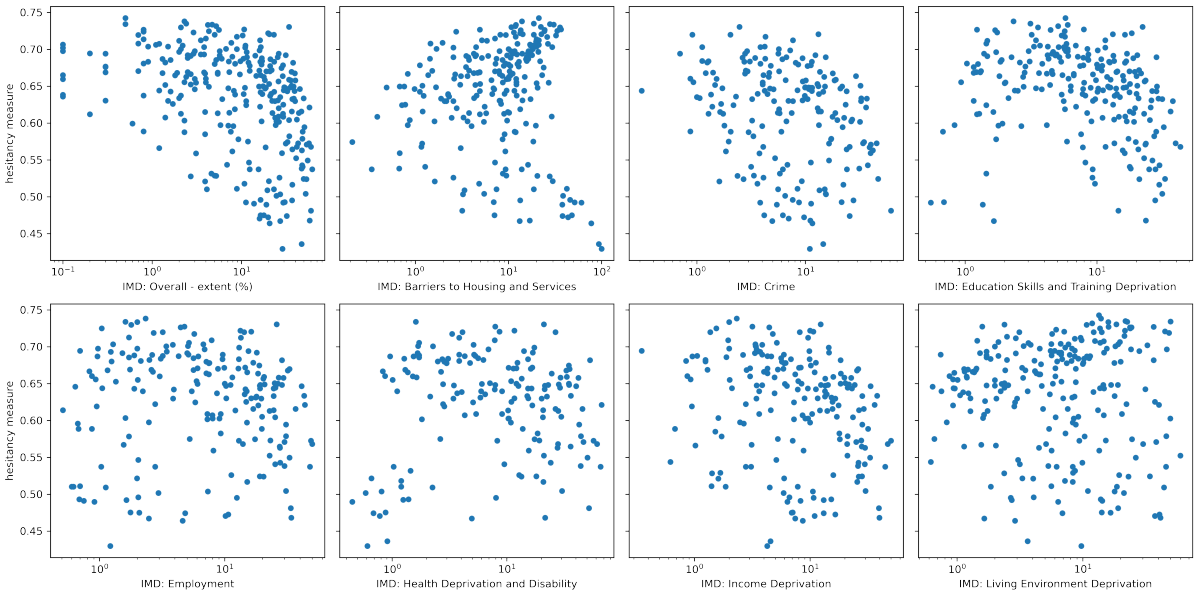


Figure 3.11: Plots of IMD metrics vs our hesitancy measure.

Here we plot various IMD metrics vs our hesitancy measure in England stratified by local authority. A small reminder that IMD: Overall - Extent (%) is a local authority level measure that shows the percent of LSOAs in the area that lie among the most deprived LSOAs in England. For the individual IMD metrics the measure is calculated as, the percent of LSOAs in the area that are in the 10% most deprived LSOAs nationally. These plots look very similar to that of vaccine rate per 100k and that observation is shown further in the similarities of both of their Pearsons Correlation coefficient charts and the plot comparing our hesitancy measure with the cumulative 1st doses per 100000 people. We will perform our analysis using our hesitancy measure for the remainder of this paper.

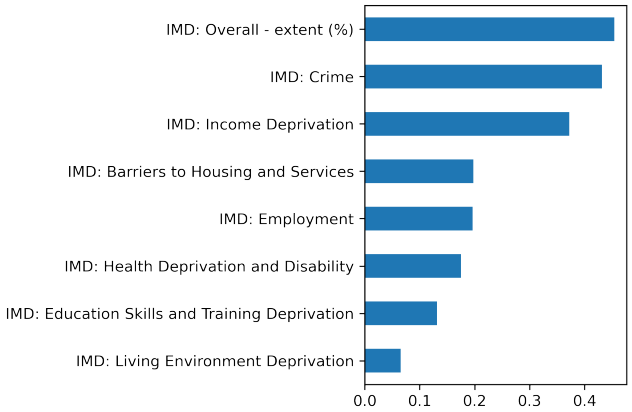


Figure 3.12: Pearsons Correlation coefficient for our hesitancy measure compared to IMD metrics.

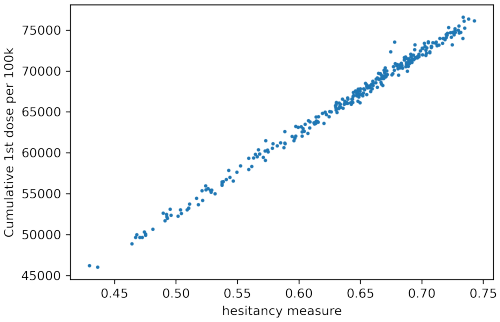


Figure 3.13: Hesitancy measure vs Cumulative 1st dose per 100k

3.3.3 Are there significant differences between the IMD Metrics?

Our goal here is to compare the IMD metrics to decide which one is the most 'important' when predicting an area's vaccine hesitancy. To be able to continue our comparison of the seven IMD metrics we first need to make sure they are significantly different. As if there are not any significant differences then there is not much point comparing their differences.

So, let's first look at how each IMD metric is distributed. A small reminder that our IMD metrics here measure the percentage of a local authorities' LSOAs that are in the 10% most deprived nationally.

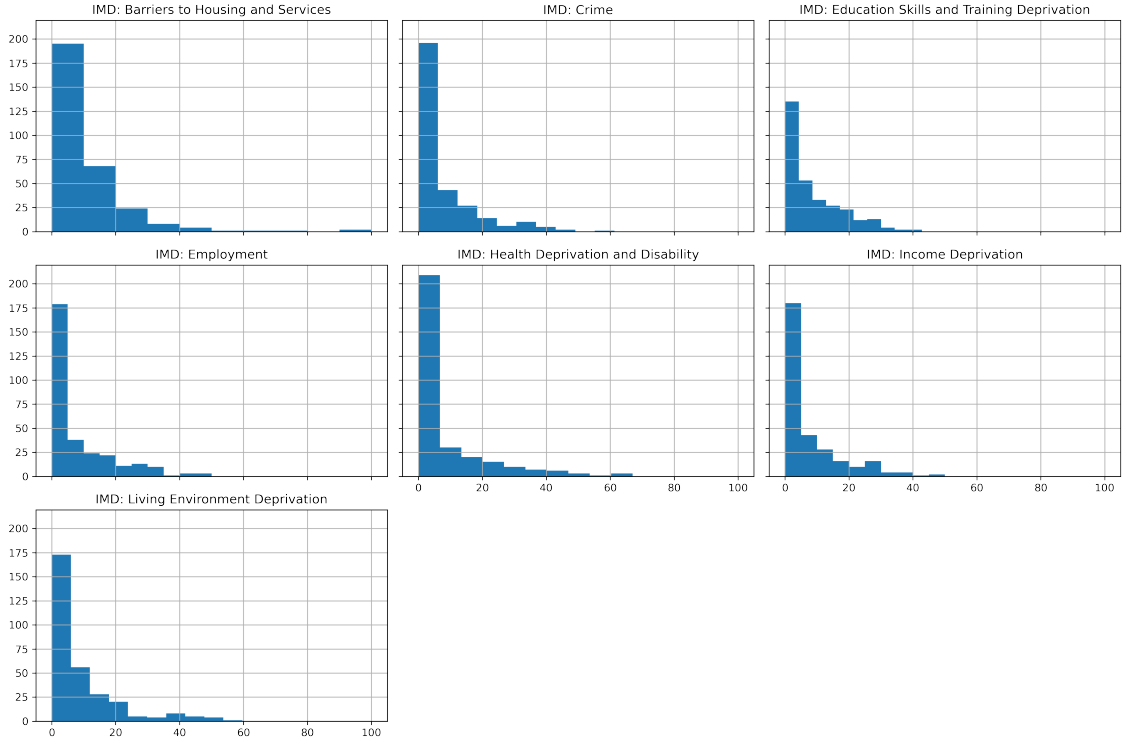


Figure 3.14: Histograms of IMD metrics stratified by local authority.

These histograms show that most local authorities are below 10% with differences between the metrics when we look at the higher percentages. Let us also compare the variances within the groups. As we can see there does appear to be a difference between the variances within the groups.

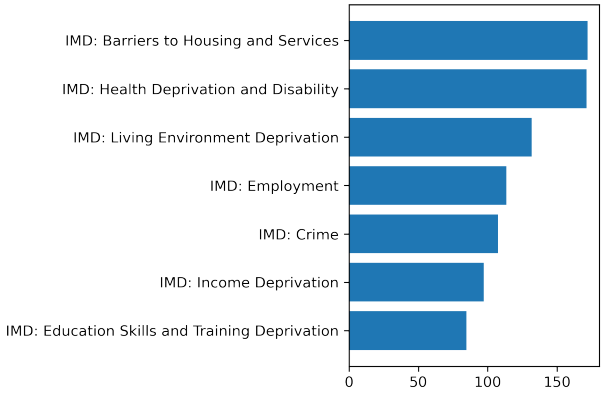


Figure 3.15: The variances of IMD metrics between local authorities.

As a final step we performed a one-way ANOVA test. The computed F statistic of our test came out to be approximately 1.9, this larger value means that the groups probably do have different means, as well as the variation between the IMD metrics is larger than the variation within them. Additionally, the associated p-value from this F distribution is approximately 0.077 which shows taht this result is statistically significant.

All of this analysis shows that there likely is a significant difference between the IMD metrics and therefore the differences between them is worth analysing.

3.3.4 Regression models

In this section we will apply various regression models to our first dose vaccine data to be able to figure out which of the seven IMD metrics is most 'important' when predicting an area's vaccine hesitancy, that is to say a unit change in which metric causes the biggest change in the vaccine hesitancy measure.

Simple Linear Regression: Number of first doses per 100000 population

We will start by creating individual linear regression models for each IMD metric and plot the results.

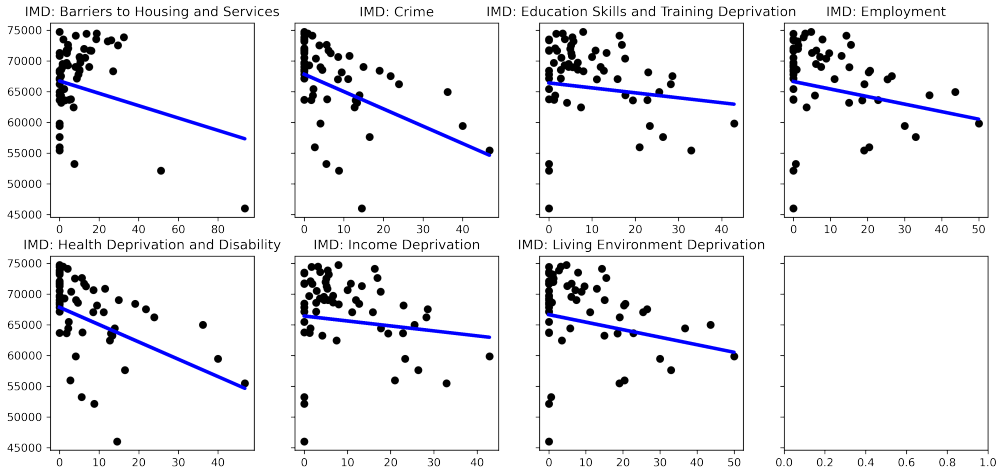


Figure 3.16: Linear Regression of Vaccine Rates per 100k and IMD Metrics.

From these models we can see that every metric has a negative coefficient, which is understandable as an area that has a higher IMD metric is likely more deprived and therefore likely has a worse vaccine rate. However linear regression models have many limitations that might mean they are not appropriate to be used in this case, as can be seen in the poor mean-squared and R-squared scores.

IMD Metric	Mean squared error	R-squared scores
IMD: Barriers to Housing and Services	35548940	−0.00
IMD: Crime	29456670	0.17
IMD: Education Skills and Training Deprivation	36470471	−0.03
IMD: Employment	35716791	−0.01
IMD: Health Deprivation and Disability	29456670	0.17
IMD: Income Deprivation	36470471	−0.03
IMD: Living Environment Deprivation	35716791	−0.01

Table 3.1: Linear regression of vaccine rates per 100k and IMD metrics: accuracy scores.

Simple Linear Regression: Hesitancy Measure

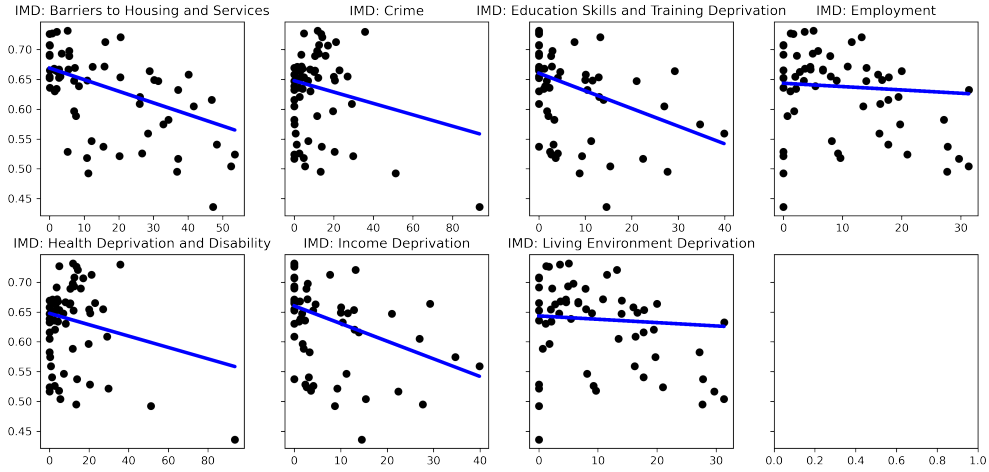


Figure 3.17: Linear regression of the hesitancy measure and IMD metrics.

Replicating the last section, we will create individual linear regression models for each IMD metric with our hesitancy measure, and plot the results. Results and accuracy scores are very similar to that of the vaccine rate per 100k section above. To note, the mean-squared error scores are not appropriate here but the R-squared scores are slightly higher for our hesitancy measure compared to the first dose vaccine rate per 100k.

IMD Metric	Mean squared error	R-squared scores
IMD: Barriers to Housing and Services	0.0035	0.30
IMD: Crime	0.0048	0.04
IMD: Education Skills and Training Deprivation	0.0044	0.12
IMD: Employment	0.0049	0.02
IMD: Health Deprivation and Disability	0.0048	0.04
IMD: Income Deprivation	0.0044	0.12
IMD: Living Environment Deprivation	0.0049	0.02

Table 3.2: Linear regression of our hesitancy measure and IMD metrics: accuracy scores.

Multiple Linear Regression: Hesitancy Measure

In this section we applied a multiple regression model to our IMD metrics. In our simple linear regression models we compared one independent variable (the one IMD metric) to one dependent variable (our hesitancy measure). In our multiple regression model we compare all seven IMD metrics as our independent variables to one dependent variable (our hesitancy measure).

From this model we will get coefficients for each of our IMD metrics. To be able to appropriately compare these coefficients we need to standardise our data. So before we apply our model we applied the standard (or Z) score to each of our data points. Here, x is our data point, μ is the mean of the IMD metric data and α is one standard deviation of the IMD metric data:

$$Z = \frac{x - \mu}{\alpha} \tag{3.1}$$

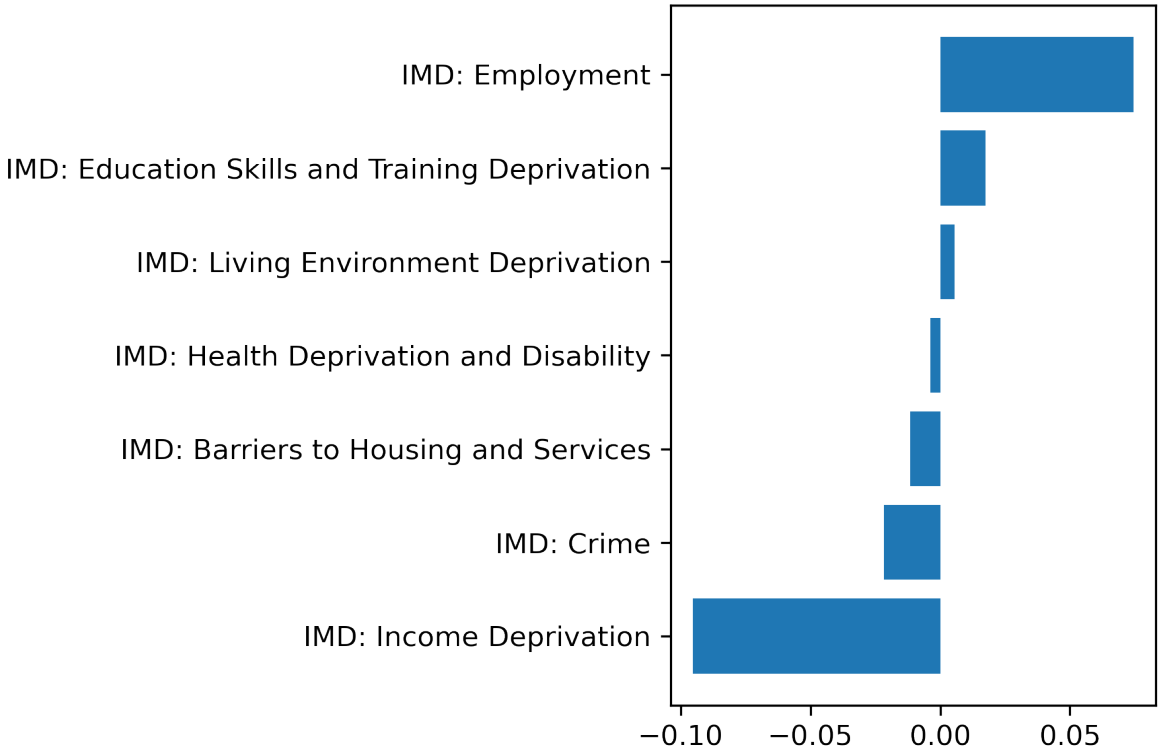


Figure 3.18: Multiple linear regression model. How 'important' is each IMD metric.

We then applied our multiple regression model to our standardised data to get seven coefficients, one for each of our IMD metrics, that represent the predicted change in our hesitancy measure for the unit change in the metric. The idea here is that the relative size of each of the coefficients represents the 'importance' of that metric (compared to the other metrics in the model) for determining the value of our hesitancy measure. Here we show the values of each metric's coefficient from our multiple linear regression model. From this model we see that, $IMD : Employment$ and $IMD : IncomeDeprivation$ seem to be the most influential metrics in determining our hesitancy measure. The R-squared accuracy score here is higher than those of our simple linear models but still lower than we would like.

Model	R-squared scores
Multiple Linear Regression	0.638

Table 3.3: Multiple linear regression of our hesitancy measure and IMD metrics: accuracy scores.

Generalised Linear Regression: Hesitancy Measure

In this section we applied a generalised regression model to our IMD metrics. We used Sklearn's TweedieRegressor that can be used to model different GLMs depending on the power, p, parameter, which determines the underlying distribution as well as a penalty term, α, that determines the regularisation strength. We initialised our model with parameters $p = 26$ and $\alpha = 0.5$. In our simple and multiple linear regression models we compared independent and dependent variables assuming our results could be predicted by linear functions. In our generalised regression model we compare all seven IMD metrics as our independent variables to one dependent variable, with the addition of a $linkfunction$ that allows the data to have arbitrary distributions whilst still being able to be predicted by a linear model.

From this model we will get coefficients for each of our IMD metrics. To be able to appropriately compare these coefficients we again used the standardise data.

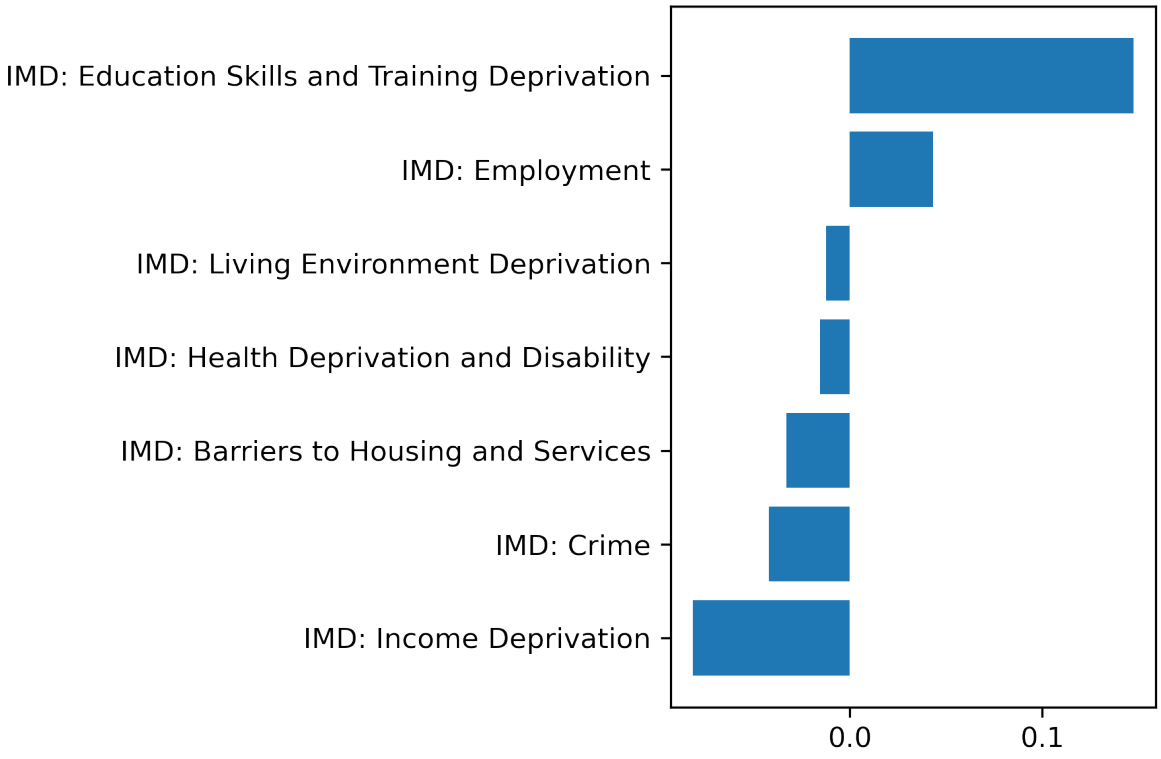


Figure 3.19: Generalised linear regression model. How 'important' is each IMD metric.

Again we get seven coefficients, one for each of our IMD metrics, that represent the predicted change in our hesitancy measure for the unit change in the metric. Here we show the values of each metric's coefficient from our generalised linear regression model. From this model we see that, *IMD: Education Skills and Training Deprivation* and *IMD: Income Deprivation* seem to be the most influential metrics in determining our hesitancy measure. The R-squared accuracy score however is negative which indicates the model is worse. So to help with this we need to tune our hyperparameters p and α.

Model	R-squared scores
Generalised Linear Regression	-1.028

Table 3.4: Generalised linear regression of our hesitancy measure and IMD metrics: accuracy scores.

Improved Generalised Linear Regression: Hesitancy Measure

In this section we performed some hyperparameter tuning for the power value, p, and penalty value, α, in our TweedieRegressor.

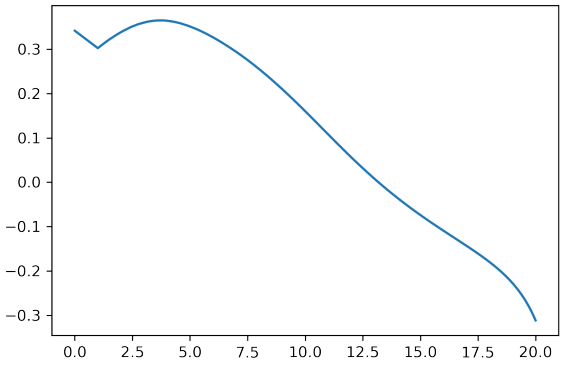


Figure 3.20: Power value parameter tuning using our validation set. (0-20)

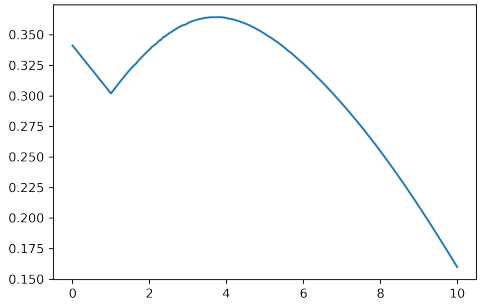


Figure 3.21: Power value parameter tuning using our validation set. (0-10)

parameter tuning range	power value	R-squared scores
$0-20$	3.8	0.364374
$0-10$	3.78	0.364377

Table 3.5: Parameter values that maximise the R-squared score on the validation set.

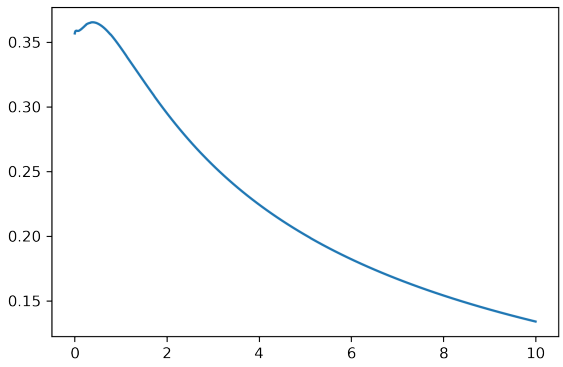


Figure 3.22: Alpha value parameter tuning using our validation set. (0-10)

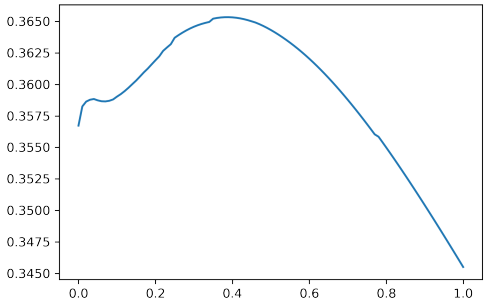


Figure 3.23: Alpha value parameter tuning using our validation set. (0-1)

parameter tuning range	alpha value	R-squared scores
$0-10$	0.39	0.365
$0-1$	0.39	0.365

Table 3.6: Parameter values that maximise the R-squared score on the validation set.

From this hyperparameter tuning we get that having the power value, $p = 3.78$, and the alpha value, $\alpha = 0.39$, give the highest R-squared accuracy score on the validation set. After this hyperparameter tuning we get the following accuracy score on the test set.

Model	power value	alpha value	R-squared scores
Generalised Linear Regression	3.78	0.39	0.211

Table 3.7: Improved generalised linear regression of our hesitancy measure and IMD metrics: accuracy scores.

Although this R-squared score is an improvement on the one before the tuning, it is lower than that of the multiple linear regression model. Again we get seven coefficients, one for each of our IMD metrics, that represent the predicted change in our hesitancy measure for the unit change in the metric. Here we show the values of each metric's coefficient from our improved generalised linear regression model. From this model we see that, *IMD: Crime* and *IMD: Income Deprivation* seem to be the most influential metrics in determining our hesitancy measure. The R-squared accuracy score however is still too low to use this model with confidence. So to help with this we need to add more input variables.

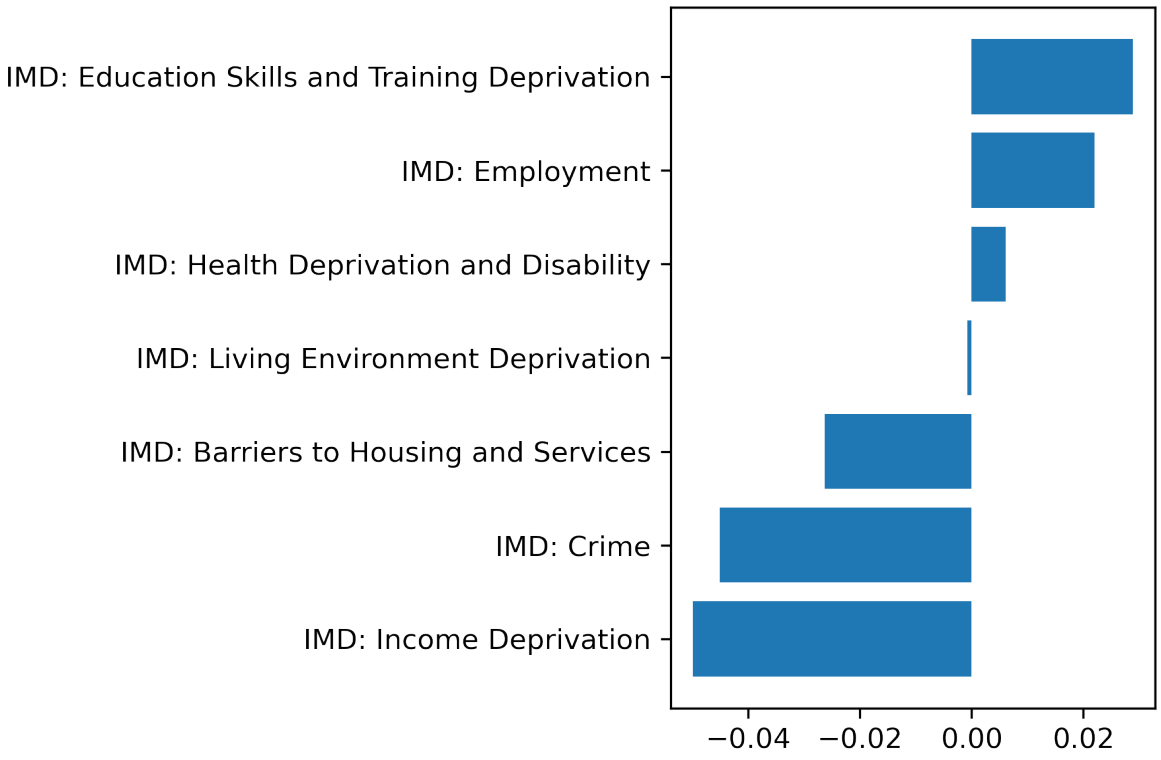


Figure 3.24: Generalised linear regression model after hyperparameter tuning. How 'important' is each IMD metric.

3.3.5 Regression models: additional input variables

In this section we will repeat our previous steps of model development and application in the hopes that doing so will increase the accuracy of our models and give us a better understanding of which IMD metric is most influential when trying to predict an area's vaccine hesitancy. Please note that we are performing these processes with less details than before as to not repeat ourselves too much.

The additional variables for the given area we have added are:

- Mean Age
- Median Age
- The proportion of people with a white British ethnicity
- The proportion of people with a multiple ethnic group ethnicity
- The proportion of people with a black British ethnicity
- The proportion of people with an other ethnic group ethnicity
- The proportion of people with an Asian British ethnicity

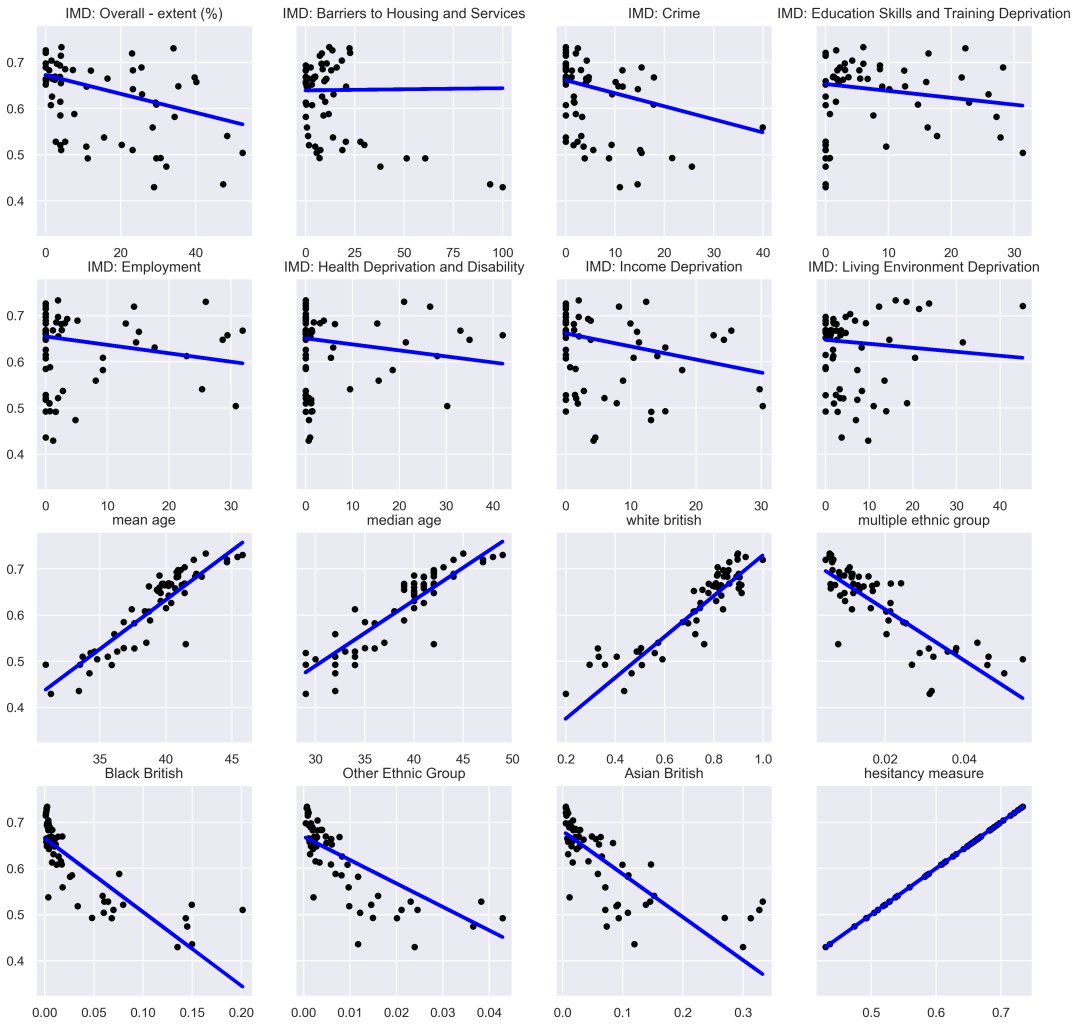


Figure 3.25: Linear regression of the hesitancy measure ,IMD metrics and additional variables.

IMD Metric	Mean squared error	R-squared scores
IMD: Overall - extent (%)	0.0060	0.10
IMD: Barriers to Housing and Services	0.0071	−0.08
IMD: Crime	0.0063	0.05
IMD: Education Skills and Training Deprivation	0.0075	−0.13
IMD: Employment	0.0077	−0.17
IMD: Health Deprivation and Disability	0.0076	−0.15
IMD: Income Deprivation	0.0070	−0.05
IMD: Living Environment Deprivation	0.0074	−0.11
Mean Age	0.0013	0.81
Median Age	0.0013	0.81
White British	0.0013	0.81
Multiple Ethnic Group	0.0020	0.69
Black British	0.0024	0.64
Other Ethnic Group	0.0028	0.57
Asian British	0.0034	0.49

Table 3.8: Linear regression of our hesitancy measure, IMD metrics and additional variables: accuracy scores.

Multiple Linear Regression: Hesitancy Measure

In this section we applied a multiple regression model to our IMD metrics and additional variables. Recall that all the data have been standardised using the z-score formula.

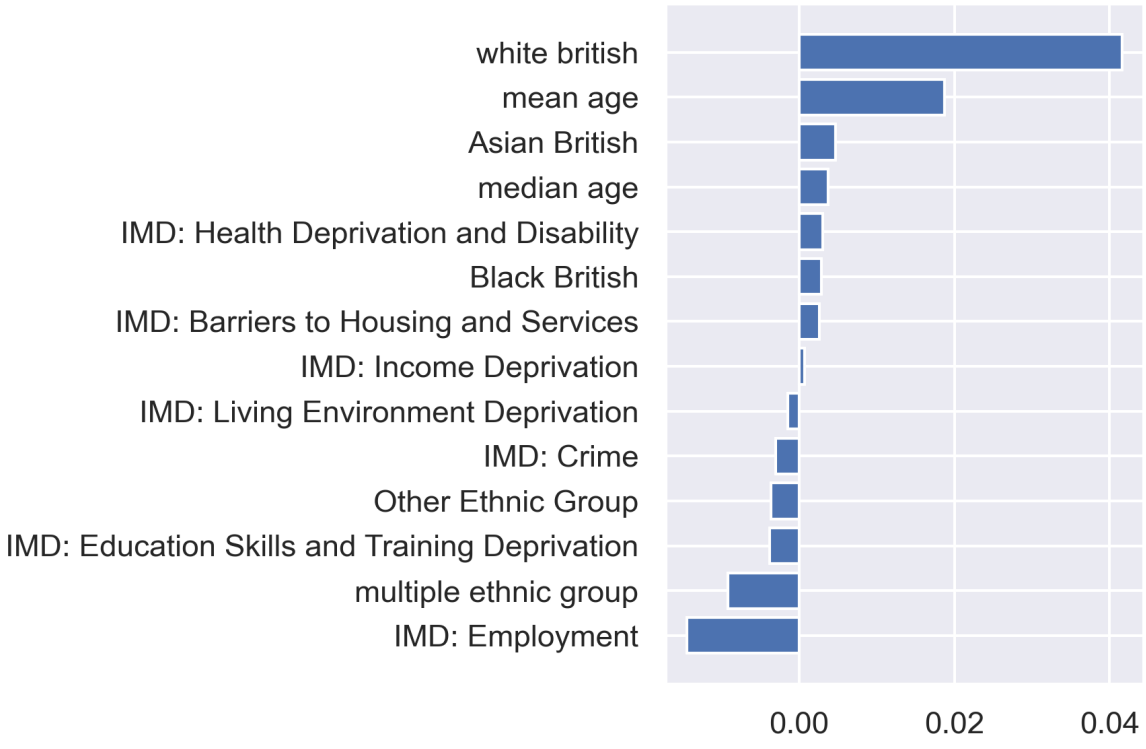


Figure 3.26: Multiple linear regression model. How 'important' is each variable.

We then applied our multiple regression model to our standardised data to get seven coefficients. Here we show the values of each metric's coefficient from our multiple linear regression model. From this model we see that, *White British*, *Mean Age* and *IMD: Employment* seem to be the most influential metrics in determining our hesitancy measure.

Isolating only the IMD metrics we see that *IMD: Employment* far outweighs the other metrics as to how influential it is in predicting vaccine hesitancy.

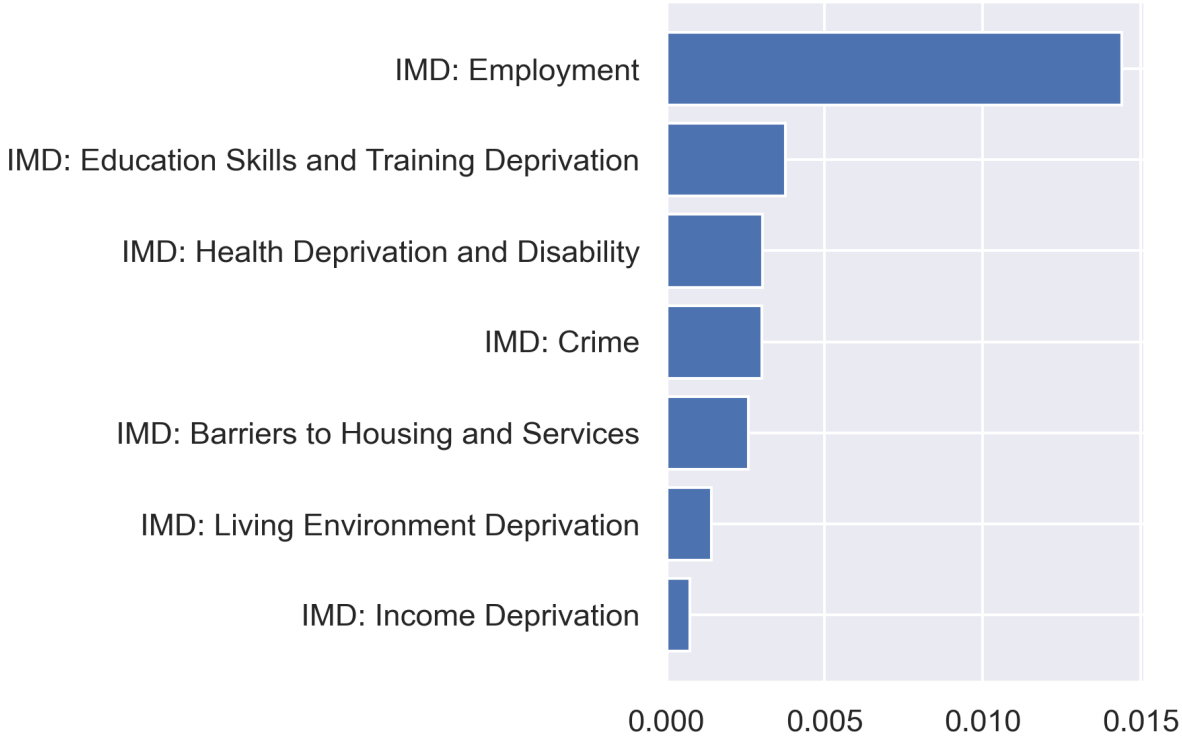


Figure 3.27: Multiple linear regression model. How 'important' is each IMD metric.

Model	R-squared scores
Multiple Linear Regression	0.9031

Table 3.9: Multiple linear regression of our hesitancy measure, IMD metrics and additional variables: accuracy scores.

Generalised Linear Regression: Hesitancy Measure

From this model we will get coefficients for each input variable. To be able to appropriately compare these coefficients we again used the standardise data.

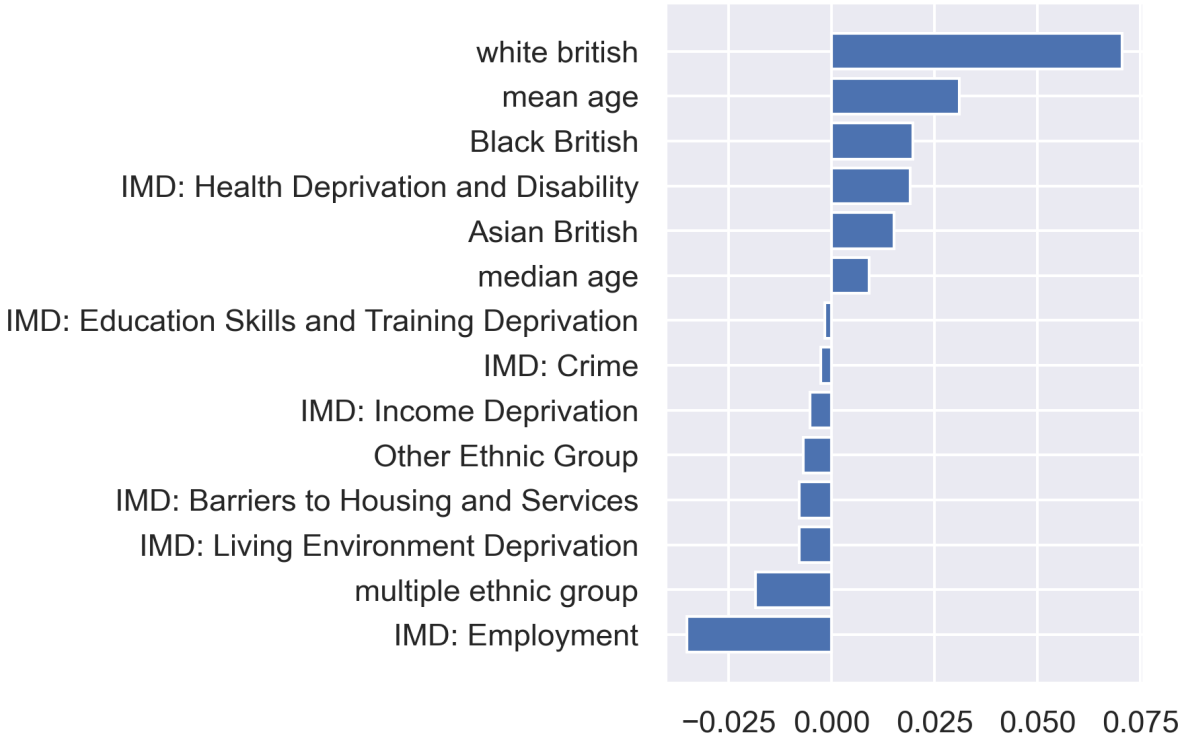


Figure 3.28: Generalised linear regression model. How 'important' is each variable.

Again we get seven coefficients, one for each of our IMD metrics, that represent the predicted change in our hesitancy measure for the unit change in the input variable. Here we show the value of each variables' coefficient from our generalised linear regression model. From this model we see that, *White British,Mean Age* and *IMD: Employment* seem to be the most influential variable in determining our hesitancy measure.

Isolating only the IMD metrics we see that *IMD: Employment* outweighs the other metrics with *IMD: Health Deprivation and Disability* as the second most influential.

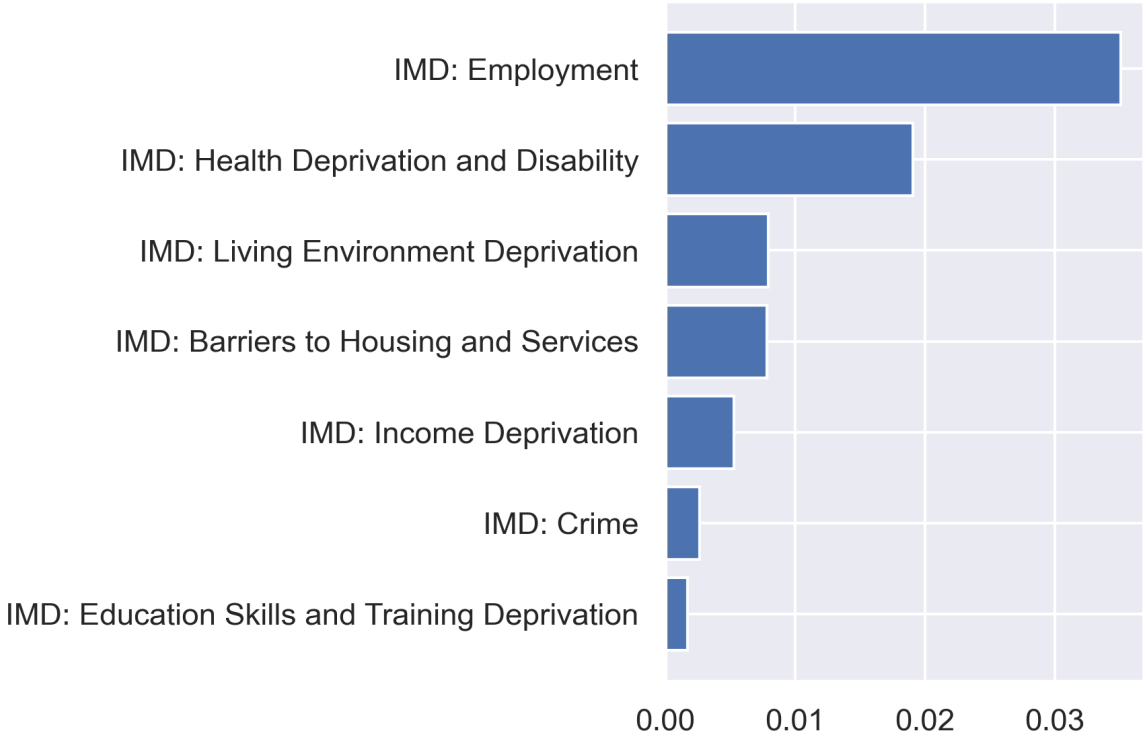


Figure 3.29: Generalised linear regression model. How 'important' is each IMD metric.

Model	R-squared scores
Generalised Linear Regression	0.9029

Table 3.10: Generalised linear regression of our hesitancy measure,IMD metrics and additional variables: accuracy scores.

Improved Generalised Linear Regression: Hesitancy Measure

In this section we performed some hyperparameter tuning for the power value, p, and penalty value, α, in our TweedieRegressor.

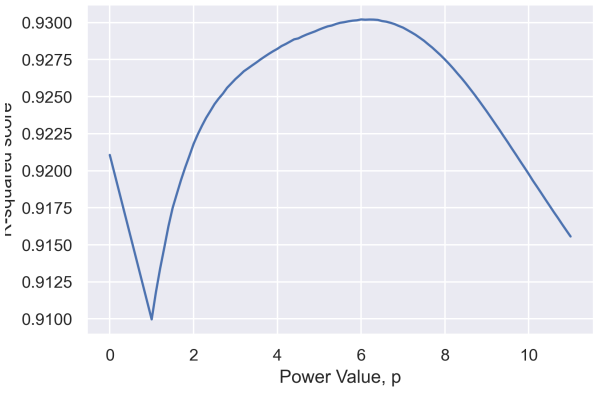


Figure 3.30: Power value parameter tuning using our validation set. (0-11)

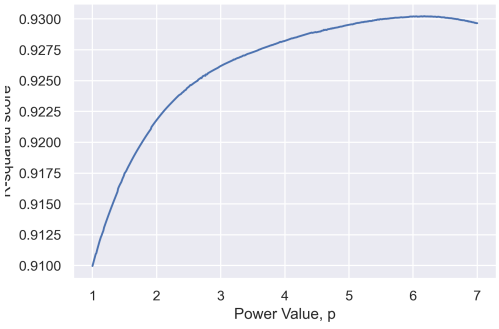


Figure 3.31: Power value parameter tuning using our validation set. (1-7)

parameter tuning range	power value	R-squared scores
$0-11$	6.0	0.930211
$1-7$	6.18	0.930229

Table 3.11: Parameter values that maximise the R-squared score on the validation set.

Figure 3.32: Alpha value parameter tuning using our validation set. (0-10)

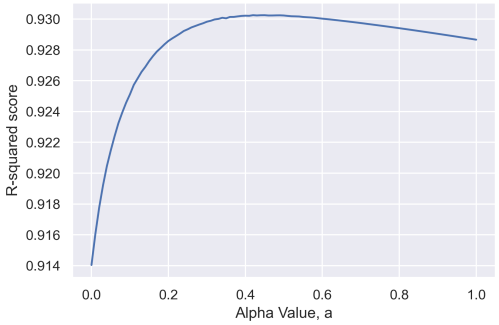


Figure 3.33: Alpha value parameter tuning using our validation set. (0-1)

parameter tuning range	alpha value	R-squared scores
$0-10$	0.5	0.930229
$0-1$	0.42	0.930250

Table 3.12: Parameter values that maximise the R-squared score on the validation set.

From this hyperparameter tuning we get that having the power value, $p = 6.18$, and the alpha value, $\alpha = 0.42$, give the highest R-squared accuracy score on the validation set. After this hyperparameter tuning we get the following accuracy score on the test set.

Model	power value	alpha value	R-squared scores
Generalised Linear Regression	6.18	0.42	0.902978

Table 3.13: Improved generalised linear regression of our hesitancy measure, IMD metrics and additional variables: accuracy scores.

Although this R-squared score is an improvement on the one before the tuning, it is lower than that of the multiple linear regression model. Again we get seven coefficients, one for each of our IMD metrics. From this model we see that, *white British* seems to be the most influential metric in determining our hesitancy measure.

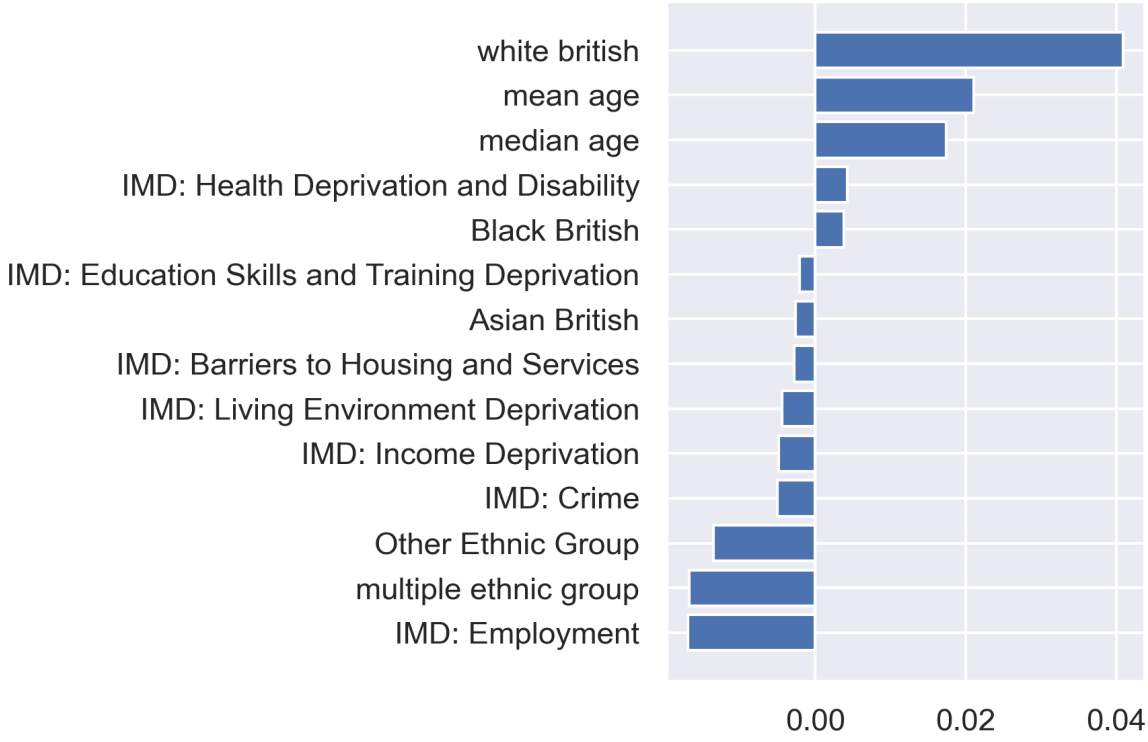


Figure 3.34: Generalised linear regression model after hyperparameter tuning. How 'important' is each variable.

Isolating only the IMD metrics we see that *IMD: Employment* still outweighs the other metrics as the most influential.

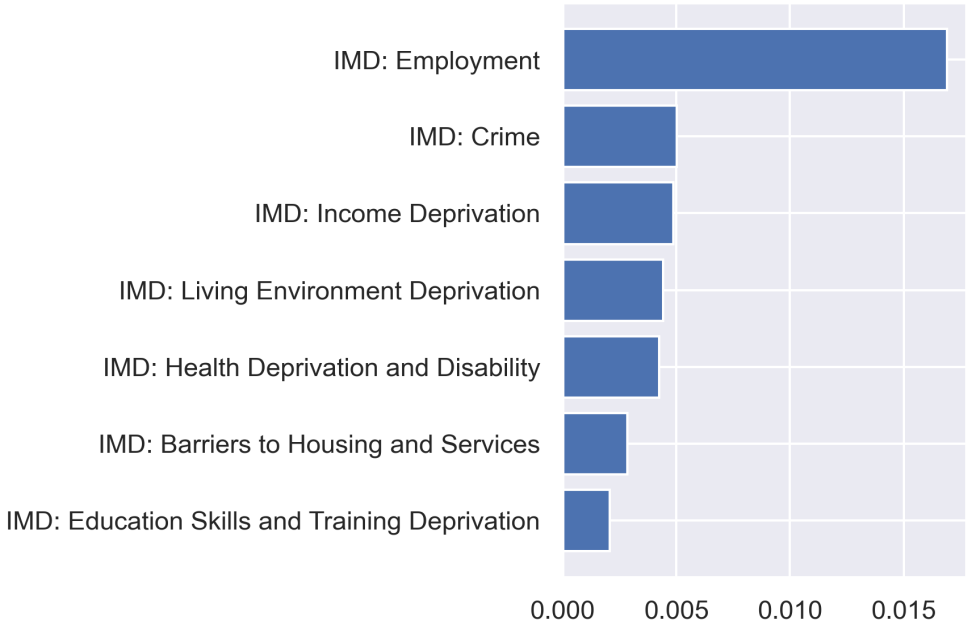


Figure 3.35: Generalised linear regression model. How 'important' is each IMD metric.

3.3.6 Closing Remarks

We have applied many models here so below we list each model, their resulting R-squared score (an average is taken if there are multiple models from one section) and the IMD metric that was shown to be the most influential IMD metric in determining vaccine hesitancy.

Model	R-squared score	Most Influential IMD Metric
MLM: Hesitancy Measure	0.638	IMD: Income Deprivation
Improved GLM: Hesitancy Measure	0.211	IMD: Income Deprivation
SLM: Hesitancy Measure	0.094	N/A
SLM: Number of first doses per 100k	0.037	N/A
GLM: Hesitancy Measure	-1.028	IMD: Education Skills and Training Deprivation

Table 3.14: Summary of models.

Model (with addition variables)	R-squared score	Most Influential IMD Metric
MLM: Hesitancy Measure	0.9031	IMD: Employment
Improved GLM: Hesitancy Measure	0.9030	IMD: Employment
GLM: Hesitancy Measure	0.9029	IMD: Employment
SLM: Hesitancy Measure	0.285	N/A

Table 3.15: Summary of models with additional variables.

Chapter 4

Critical Evaluation

In this chapter we will evaluate the analysis we have performed. We will present some failure cases, what hurdles/limitations we came up against, what parts of our analyses went well and what parts can be improved upon. For each situation we will attempt to suggest improvement, what could have been done differently or what might have benefited from a different approach. This chapter isn't solely to critique our own work but to also add a stepping stone for any future work that is performed. From this evaluation future data scientists may understand what will work in their analysis or will figure out what not to do in the future.

4.1 Working with LG Inform Data

We started our analysis looking at COVID-19 vaccine data from the UK Government funded website, Local Government Inform (LG Inform). This data allowed us to map COVID-19 data at a local authority level and identify areas of the country that seem to be deviating from the norm and could be potential areas to look at in more detail in the future

Now this process did come with several limitations. Firstly when connecting our COVID-19 data with our mapping tools we came up against a lot of naming inconsistencies that slowed the process down considerably. Similarly when mapping the *IMD: Overall - Extent (%)* we ran into a problem of analysing data that had been gathered at different times. The IMD data came from the UK Governement's 'English indices of deprivation 2019' report whereas our COVID-19 and geographical data came from 2021. This meant that inconsistencies arose, mainly for example *North Northamptonshire* and *West Northamptonshire* didn't existed until 2021 and were therefore not in the 2019 IMD report. We resolved this issue by calculating the *IMD: Overall - Extent (%)* measure ourselves from the individual LSOA IMD scores. This was an adequate solution for this case but, in the future, creating a general function for calculating this measure ourselves may help circumvent this issue in the future.

Additionally, we came across an issue when analysing and plotting data per 100000 population. To create this measure we naturally needed to use population data which is most commonly a population estimate created by organisations such as the ONS. From LG Inform we were able to access COVID-19 data gathered by NIMS and population estimate data gathered by the ONS. Using these data sets from two different sources led to inaccurate or impossible results, although the relative values between the local authorities would likely remain the same. Another limitation we came across is that ONS population estimates are released bi-annually whereas a lot of the COVID-19 data we could access are released weekly which meant no weekly proportional data analysis could be performed. These limitations or inaccuracies led us to move away from LG Inform as we continued our analyses.

4.2 Working with Public Health England Data

In this section of our analysis we wanted to get an idea of what the underlying causes of vaccine hesitancy might be. Transitioning to using Public Health England (PHE) data sets turned out to be a very good choice as these data sets were much more comprehensive and complete. These data sets added many more variables to analyse with our weekly COVID-19 data as well as having population estimate from both ONS and NIMS.

4.2.1 Hesitancy Measure

We did however still come across some limitations in this stage of our analysis. Firstly, our hesitancy measure was created to try and understand vaccine hesitancy, to try and put a value with the number of people who are unwilling or hesitant to receive a COVID-19 vaccine. However, this train of thought doesn't account for how the local government's vaccine roll-out may be affecting weekly vaccine rates or how weekly vaccine rates may have slowed down but haven't actually started to asymptote. Similarly, vaccine data for different age groups was made available at different times meaning we couldn't meaningfully compare them in the earlier weeks of our data set. This also meant that, for some stratifications of the data, the weekly vaccine increase was already below the hesitancy threshold in its first week and the hesitancy measure would therefore pick that week as the 'tailing off' point rather than whatever the true date may be. Our hesitancy measure did however allow us to quickly identify a big discrepancy in the values at which different age groups 'tail off' (younger age groups seeming to 'tail off' at lower values, i.e. potentially more hesitant).

4.2.2 English Indices of Deprivation

We decided we wanted to compare our COVID-19 data with the English Indices of Deprivation as we wanted to try and get an idea of the kinds of people (or areas in which people live) that are most hesitant when it comes to getting a vaccine. With the ultimate goal being increasing vaccine coverage by gaining the ability to better understand of these people and target them with appropriate information campaigns.

A similar limitation as before arises where IMD data comes from the UK Government's 'English indices of deprivation 2019' report whereas our weekly COVID-19 data is much more recent. If we were to perform this or similar analysis again then we would try to find more recent deprivation metrics because area's rankings may have changed between 2019 and August 2021 (the time of writing).

The issue of interpretability arises when handling IMD metrics at a local authority level. As the IMD metrics are recorded at an LSOA level we have had to resort to looking at the proportion of LSOAs in a local authority. If we were to perform this analysis again we may looking into better ways to either keep the results interpretable or perform the analysis entirely at the LSOAs to potentially improve accuracy.

Additionally further analysis of the significant differences between the IMD metrics would allow for greater confidence in the results.

4.3 Regression Models

In this section we applied linear regression models to our COVID-19 and IMD data to try and understand what metrics most greatly influence vaccine hesitancy.

Simple Linear Regression

When performing this analysis we did not present any evidence to show that our data met the assumptions of a regression model. This is something we would analyse if we were to perform these processes again. The use of the mean-squared error for our simple linear regression model definitely lead to some misleading results. Using mean-squared error for errors less than one causes small errors to be ignored whereas using mean-squared error for values greater than one causes outliers to have a much greater weighting. This highlights the possible limitations of using the mean-squared error measure and is the reason we stepped away from it in the later parts of our analysis. R-squared scores seemed to be a more reliable accuracy measure as we talk about later on.

4.3.1 Multiple Linear Regression

As with the simple linear regression, we neglected the key assumptions of linear regression, namely the presence of a linear relationship, normal distribution and no co-linearity. Co-linearity is something we touched on but not to the extent it should have been. If we were to perform this analysis again we would make sure to be more thorough when laying the foundations for our analysis.

4.3.2 Generalised Linear Regression

The initial application of the generalised linear model return a negative R-squared score which meant the model really didn't fit the data at all. This is mainly because we instantiated our hyperparameters

randomly. This was worthwhile keeping in the analysis process as to show the base line from which we can build upon.

We then performed some hyperparameter tuning and acquire more appropriate values of the power value (which describes the distribution of the data) and the alpha value (which determines the strength of the regularisation of the model). This improved our R-squared accuracy score by a reasonable amount but didn't come out as high as our accuracy scores from our multiple regression model. If we were given more time then we would thoroughly compare these two models to better understand why one might be more or less appropriate for modelling our data.

4.3.3 Adding Additional Variables

Adding additional variables to our model understandably increased its accuracy but what we neglected was the facts that it made the assumptions of our linear models much harder to meet. Many of the additional metrics had strong co-linearity (possibly affecting results) and so this is definitely something we would look into more thoroughly given more time.

That being said, the Public Health England data we were using had a plethora of variables to choose from and we could have definitely experimented with adding even more/more appropriate variables to our models give more time.

4.3.4 Closing Remarks

There is a lot that can be improved in this analysis process. More thorough analysis of the assumptions required for our models, more careful selection and use of our additional variables (to maintain the assumptions of our models) and better analysis of what conclusions we can make/further analysis we can perform once the process is complete. Furthermore, there is definite room for expansion until the wider UK counties and other countries to help tackle the same problem.

Chapter 5

Conclusion

In this paper we set out to better understand what factors influence levels of vaccine hesitancy the most. We wanted to find out whether demographic factors, such as age or ethnicity, played an important role. And whether we could predict the level of an area's vaccine hesitancy based on various measures of deprivation.

5.1 Key points of our work

In this section we present a recap of our analysis process and key findings:

1. During the execution of our work we explored data from the website, LG Inform, and were able to map out data on COVID-19 vaccine doses and cases, as well as IMD metrics.

2. Once this was complete we moved on to analysing from Public Health England's weekly data releases. We presented our definition of a vaccine hesitancy measure and then went on to explore how this hesitancy measure relates to age, ethnicity and IMD metrics at a local authority level. We found that younger age groups in England have a lower hesitancy measure which could indicate higher levels of vaccine hesitancy.

3. We then went to on to exploring the IMD Metrics themselves and whether they have significant differences to make them worth comparing.

4. Once this was complete we started applying our machine learning models to our data. We started out with simple linear regression models, comparing each of the seven main IMD domain metrics to our hesitancy measure, to get an initial understanding of how they might relate to one another.

5. We then standardised our data to be able to properly compare the IMD metrics using a multiple linear regression model. This model showed that *IMD: Income Deprivation* seemed to be the largest influencer of vaccine hesitancy, followed by *IMD: Employment*.

6. We then went on to explore additional models and applied a generalised linear model to our data and, after some hyperparameter tuning, the model showed that *IMD: Income Deprivation* again seemed to be the largest influencer of vaccine hesitancy, followed by *IMD: Crime* this time.

7. Wanting to go further, we then added additional variables to our input data, namely age and ethnicity. We replicated previous steps now with this larger input data. We compared our standardised data to our hesitancy measure using a multiple linear regression model and it showed that *IMD: Employment* seemed to overwhelmingly be the largest influencer of vaccine hesitancy.

8. We then went on to repeat our use of a generalised linear regression model which showed, after some hyperparameter tuning, that *IMD: Employment* again seemed to overwhelmingly be the largest influencer of vaccine hesitancy.

We created applied many models during our analysis so for clarity and understanding we will reproduce here our model summary tables that were presented at the end of the Execution chapter as well as a bar chart that represents these results.

Model	R-squared score	Most Influential IMD Metric
MLM: Hesitancy Measure	0.638	IMD: Income Deprivation
Improved GLM: Hesitancy Measure	0.211	IMD: Income Deprivation
SLM: Hesitancy Measure	0.094	N/A
SLM: Number of first doses per 100k	0.037	N/A
GLM: Hesitancy Measure	−1.028	IMD: Education Skills and Training Deprivation

Table 5.1: Summary of models.

Model (with addition variables)	R-squared score	Most Influential IMD Metric
MLM: Hesitancy Measure	0.9031	IMD: Employment
Improved GLM: Hesitancy Measure	0.9030	IMD: Employment
GLM: Hesitancy Measure	0.9029	IMD: Employment
SLM: Hesitancy Measure	0.285	N/A

Table 5.2: Summary of models with additional variables.

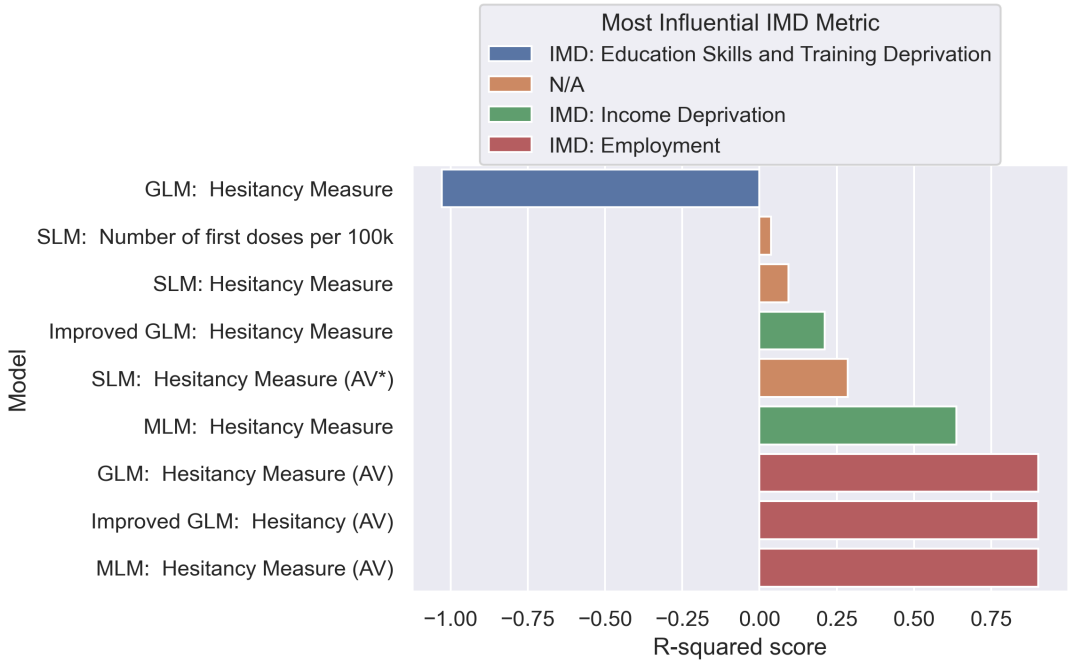


Figure 5.1: Summary of models (*AV = Additional Variables)

5.2 Current project status

The work presented in this paper provides a solid introduction and exploration of how data science and machine learning can be applied to the task of understanding vaccine hesitancy. Currently our models that solely explore IMD metrics and our hesitancy measure show that *IMD: Income Deprivation* is the most influential metric, these models do however come with relatively low accuracy scores. Our models that contain additional input variables show that *IMD: Employment* is the most influential metric, these models come with relative high accuracy scores. We are currently using a weekly 1% growth rate as the cut off point for are hesitancy measure, this value can be adjust as more data becomes available and growth rates continue to shrink.

Aim 1. which we talked about in our Introduction chapter was to answer the question, "Do factors such as age or ethnicity play a large role in predicting an areas' overall vaccine hesitancy?". We performed analysis of ages to show that there is a discrepancy in our hesitancy measure between ages groups. We didn't explore ethnicity as much although that, and age, did play a larger role in the later sections of our execution chapter.

Aim 2. which we talked about in our Introduction chapter was to answer the question, "What are the most influential measures of deprivation when predicting an areas' overall vaccine hesitancy?". We applied multiple linear regression models to answer this exact question which we have outlined above.

Objective 1. which we talked about in our Introduction chapter was about gathering data. Our initial data sets were accessed through the website, LG Inform, before we moved on to using data from Public Health England, Gov.uk, ONS and NIMS.

Objective 2. which we talked about in our Introduction chapter was about cleaning our data and creating some initial exploratory plots. We complete this objective and this is what took up most of our time. Processing the data and getting it into a suitable format for analysis as well as our initial attempts at creating geographical plots.

Objective 3. which we talked about in our Introduction chapter was about applying machine learning models to our data to find the indicators of vaccine hesitancy. For this we did apply multiple models to complete this tasks but did not extend this into how we can use IMD metrics to predict hesitancy measure values.

5.3 Future work

This paper is simply a first foray into using machine learning to predict vaccine hesitancy. Our initial assumptions were that *IMD: Education* would be the (or one of the) most influential factors, but our models showed that this was not the case at all. This could be because it truly isn't as strong a factor as is popularly believed or that its hidden colinearity with other input variables had an effect on our results that we did not pick up on. Either of these would be a good basis for further analysis.

There is no shortage of variables about people that we can access in the current information age in which we live. Our initial analysis around adding additional variables to our models showed promising results, and so further analysis around adding even more variables could lead to interesting results. So long as the assumptions of our models can still be adhered to.

The potential scope for a project such as this is worldwide. This paper only deals with data in England but there is no scientific reason why it couldn't be extend into other countries in the UK and around the world. Doing so would increase the size of the data sets being used, therefore increasing the accuracy of our models and ultimately people's confidence in the results.

5.4 Closing remarks

The relevance of vaccine hesitancy research is constantly growing. To be able to get out of this pandemic we need to have effective vaccination programs, to do that we need to address the reasons behind low vaccine uptake in a robust and timely fashion, and to do that we need equally as robust and timely research into vaccine hesitancy. The worse our populations' vaccine hesitancy is, the longer this and future pandemic will last.

To properly tackle the issues around vaccine hesitancy we need to understand what kinds of people might be hesitant, the reasons why they might be hesitant and the best methods, tools, and information needed to lessen this hesitancy, improve vaccine coverage and ultimately save lives.

This paper takes a look at what factors might influence current vaccine hesitancy. This is just an initial exploration of a highly complex topic, one that needs to be tackled in a collaborative effort by multiple scientific fields. Will this kind of research be able to help during this current pandemic? Will this problem still be around during future pandemics, and if so, will it be able to help? Will the reasons behind vaccine hesitancy during future pandemics be the same as they are now?

Bibliography

[1] A brief history of data analysis. https://www.flydata.com/blog/a-brief-history-of-data-analysis/. Accessed: 2021-09-11.

[2] Control of patient information (copi) notice. https://www.opensafely.org/. Accessed: 2021-09-11.

[3] Coronavirus. https://www.who.int/health-topics/coronavirus#tab=tab_3. Accessed: 2021-09-11.

[4] Coronavirus disease (covid-19) pandemic. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov#:~:text=On%2031%20December%202019%2C,2019%2DnCoV%E2%80%9D. Accessed: 2021-09-11.

[5] The cost of coronavirus. https://www.instituteforgovernment.org.uk/explainers/cost-coronavirus. Accessed: 2021-09-11.

[6] Covid-19 coronavirus pandemic. https://www.worldometers.info/coronavirus/. Accessed: 2021-09-11.

[7] Covid-19 vaccinations. https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-vaccinations/. Accessed: 2021-09-11.

[8] Covid-19 vaccinations. https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-vaccinations/. Accessed: 2021-09-11.

[9] Data science and ai in the age of covid-19 – report. https://www.turing.ac.uk/research/publications/data-science-and-ai-age-covid-19-report. Accessed: 2021-09-11.

[10] Deaths in united kingdom. https://coronavirus.data.gov.uk/details/deaths. Accessed: 2021-09-11.

[11] Debate on the report "covid-19 vaccines: ethical, legal and practical considerations". https://www.who.int/director-general/speeches/detail/debate-on-the-report-covid-19-vaccines-ethical-legal-and-practical-considerations. Accessed: 2021-09-11.

[12] The english indices of deprivation 2019 (iod2019). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf. Accessed: 2021-09-11.

[13] Global vaccine action plan. https://www.who.int/teams/immunization-vaccines-and-biologicals/strategies/global-vaccine-action-plan. Accessed: 2021-09-11.

[14] Global vaccine action plan. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019. Accessed: 2021-09-11.

[15] The history of vaccines. https://www.historyofvaccines.org/timeline/all. Accessed: 2021-09-11.

[16] Imd - crime - proportion of lsoas in most deprived 10 https://lginform.local.gov.uk/reports/lgastandard?mod-metric=8368&mod-area=E92000001&mod-group=AllRegions_England&mod-type=namedComparisonGroup. Accessed: 2021-09-11.

[17] Imd - overall - extent (%) in england. https://lginform.local.gov.uk/reports/lgastandard?mod-metric=8985&mod-area=E92000001&mod-group=AllRegions_England&mod-type=namedComparisonGroup. Accessed: 2021-09-11.

[18] Immunization. https://www.who.int/news-room/facts-in-pictures/detail/immunization. Accessed: 2021-09-11.

[19] is20347-vaccine-hesitancy-in-england-2021. https://github.com/JoshBibby/is20347-Vaccine-Hesitancy-in-England-2021. Accessed: 2021-09-12.

[20] Lg inform. https://lginform.local.gov.uk/. Accessed: 2021-09-11.

[21] Machine learning. https://royalsociety.org/topics-policy/projects/machine-learning/. Accessed: 2021-09-11.

[22] Main symptoms of coronavirus (covid-19). https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/main-symptoms/. Accessed: 2021-09-11.

[23] Naming the coronavirus disease (covid-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed: 2021-09-11.

[24] Oxford university/astrazeneca vaccine authorised by uk medicines regulator. https://www.bbc.com/future/article/20210720-the-complexities-of-vaccine-hesitancy. Accessed: 2021-09-11.

[25] Report of the sage working group on vaccine hesitancy. https://www.who.int/immunization/sage/meetings/2014/october/1_Report_WORKING_GROUP_vaccine_hesitancy_final.pdf. Accessed: 2021-09-11.

[26] Secure analytics platform for nhs electronic health records. https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice. Accessed: 2021-09-11.

[27] Ten threats to global health in 2019. https://www.cdc.gov/vaccines/parents/diseases/forgot-14-diseases.html. Accessed: 2021-09-11.

[28] Timeline of machine learning. https://en.wikipedia.org/wiki/Timeline_of_machine_learning. Accessed: 2021-09-11.

[29] Timeline: Who's covid-19 response. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#! Accessed: 2021-09-11.

[30] Vaccinations in united kingdom. https://coronavirus.data.gov.uk/details/vaccinations. Accessed: 2021-09-11.

[31] Vaccines and immunization. https://www.who.int/health-topics/vaccines-and-immunization#tab=tab_2. Accessed: 2021-09-11.

[32] Vaccines and immunization: What is vaccination? https://www.who.int/news-room/q-a-detail/vaccines-and-immunization-what-is-vaccination. Accessed: 2021-09-11.

[33] Who coronavirus (covid-19) dashboard. https://covid19.who.int/. Accessed: 2021-09-11.

[34] Why some people don't want a covid-19 vaccine. https://www.gov.uk/government/news/oxford-universityastrazeneca-vaccine-authorised-by-uk-medicines-regulator. Accessed: 2021-09-11.

[35] Tracking covid-19 excess deaths across countries. *The Economist*, 2021. Accessed: 2021-09-11.

[36] Chris Baraniuk. Covid-19: How the uk vaccine rollout delivered success, so far. *The BMJ*, Feb 2021.

[37] Thomas Bayes and null Price. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[38] Torsten Bell. The pandemic has affected poor people's mental health the most. *The Guardian*, 2021.

[39] Cornelia Betsch, Philipp Schmid, Dorothee Heinemeier, Lars Korn, Cindy Holtmann, and Robert Böhm. Beyond confidence: Development of a measure assessing the 5c psychological antecedents of vaccination. *PLOS ONE*, 2018.

[40] Noel T. Brewer, Gretchen B. Chapman, Alexander J. Rothman, Julie Leask, and Allison Kempe. Increasing vaccination: Putting psychological science into action. *Psychological Science in the Public Interest*, 18(3):149–207, 2017. PMID: 29611455.

[41] Longbing Cao. Data science: A comprehensive overview. *ACM Computing Surveys (CSUR)*, Oct 2017.

[42] Robert Böhm Cornelia Betsch. Using behavioral insights to increase vaccination policy effectiveness. *SAGE Journals*.

[43] Holly Else. How a torrent of covid science changed research publishing - in seven charts, Dec 2020.

[44] PHOSP-COVID Collaborative Group, Rachael Andrea Evans, Hamish McAuley, Ewen M Harrison, and et al. Physical, cognitive and mental health impacts of covid-19 following hospitalisation – a multi-centre prospective cohort study. *medRxiv*, 2021.

[45] Era Dabla-Norris Hibah Khan. Who doesn't want to be vaccinated? determinants of vaccine hesitancy during covid-19. *IMF*.

[46] Chaolin Huang, Yeming Wang, Xingwang Li, and et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 395(10223):497–506, 2020.

[47] Beate Kampmann and Sandra Mounier Jack. Covid-19 vaccines save lives. *BMJ*, 373, 2021.

[48] Ajinkya Kunjir, Dishant Joshi, Ritika Chadha, Tejas Wadiwala, and Vikas Trikha. A comparative study of predictive machine learning algorithms for covid-19 trends and analysis. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3407–3412, 2020.

[49] Jason Denzil Morgenstern, Emmalin Buajitti, Meghan O'Neill, Thomas Piggott, Vivek Goel, Daniel Fridman, Kathy Kornas, and Laura C Rosella. Predicting population health with machine learning: a scoping review. *BMJ Open*, 10(10), 2020.

[50] Gretchen B. Chapman Noel T. Brewer. Increasing vaccination: Putting psychological science into action. *SAGE Journals*.

[51] Sara Reardon. Rise of robot radiologists. *Nature News*, Dec 2019.

[52] Gigi F. Stark, Gregory R. Hart, Bradley J. Nartowt, and Jun Deng. Predicting breast cancer risk using personal health data and machine learning models. *PLOS ONE*.

[53] Tung Thanh Le, Zacharias Andreadakis, Arun Kumar, Raúl Gómez Román, Stig Tollefsen, Melanie Saville, and Stephen Mayhew. The covid-19 vaccine development landscape. *Nature News*, Apr 2020.

[54] Michael Woelfle, Piero Olliaro, and Matthew H. Todd. Open science is a research accelerator. *Nature News*, Sep 2011.

Appendix A

Appendix

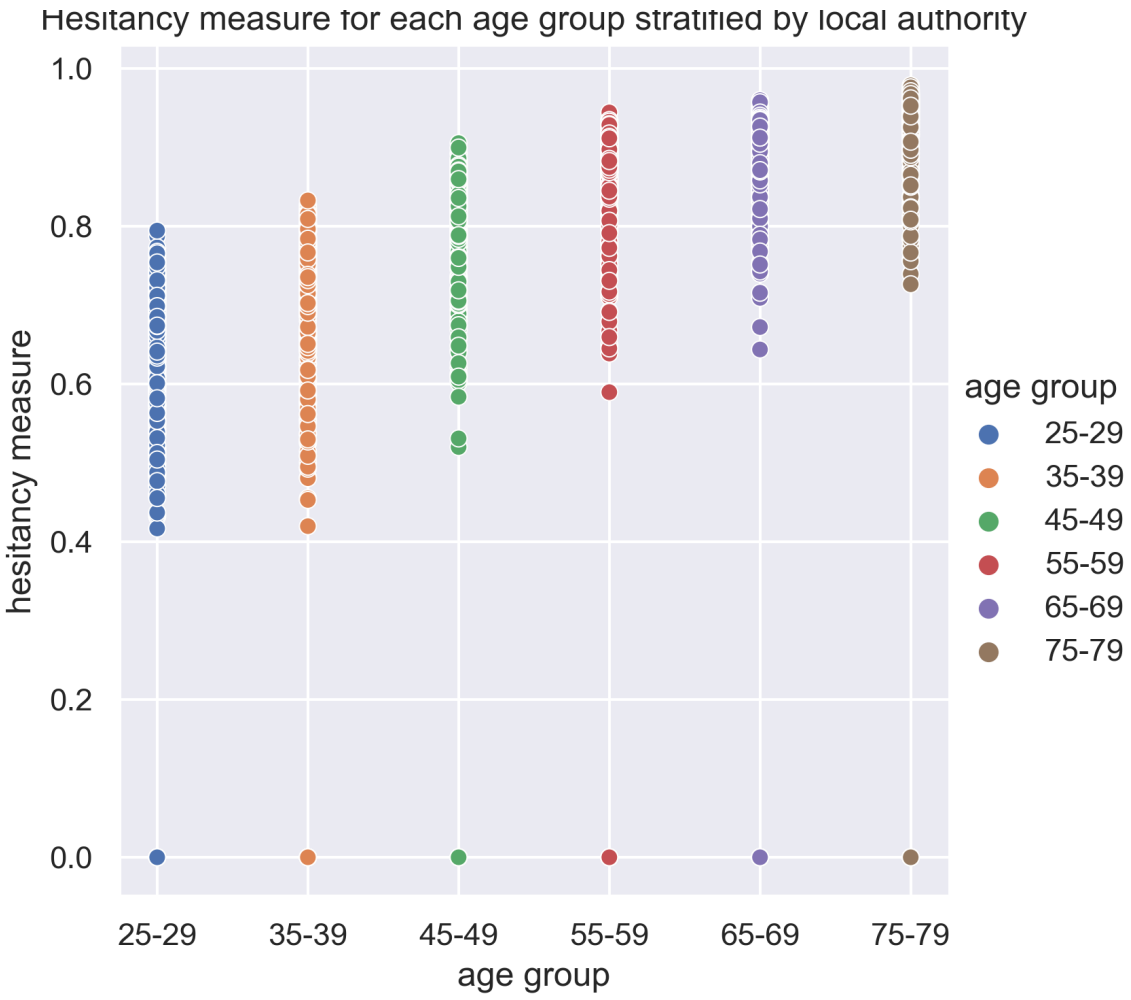


Figure A.1: Hesitancy measure for each age group stratified by local authority.

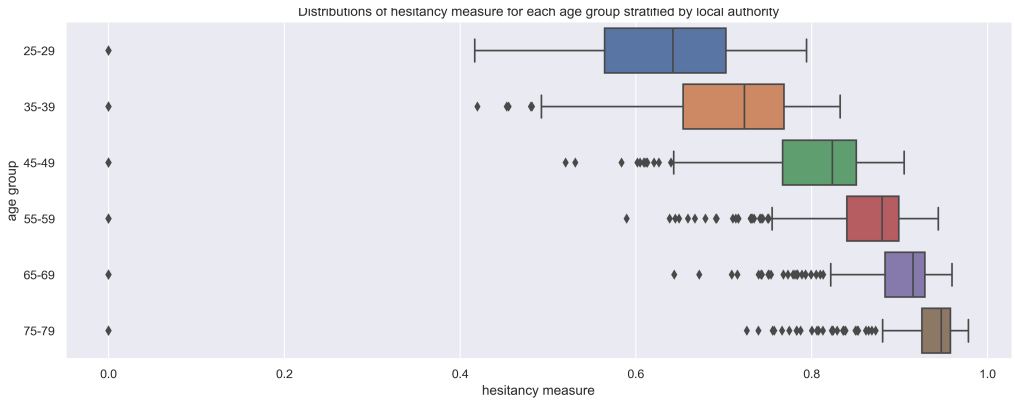


Figure A.2: Distributions of hesitancy measure for each age group stratified by local authority.

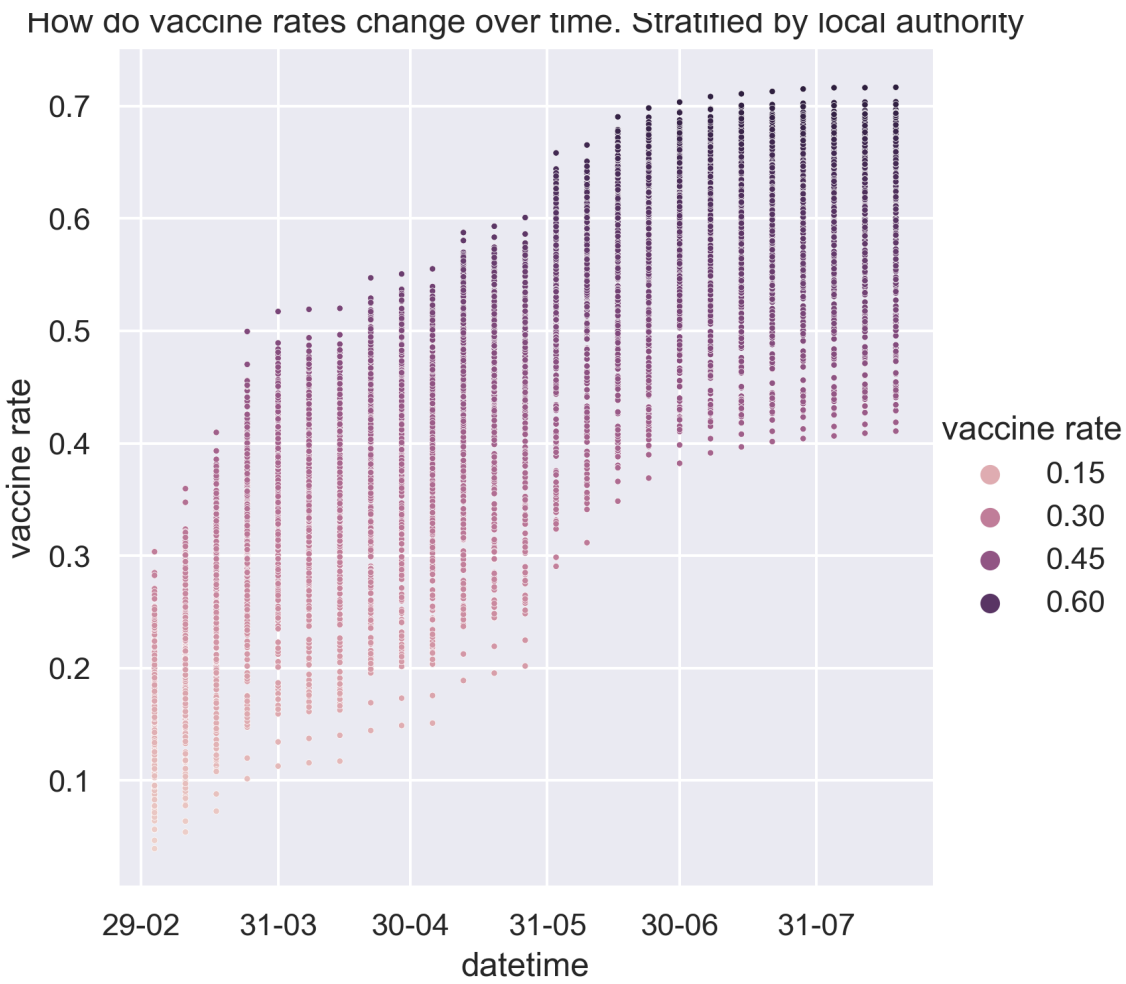


Figure A.3: How do vaccine rates change over time. Stratified by local authority.

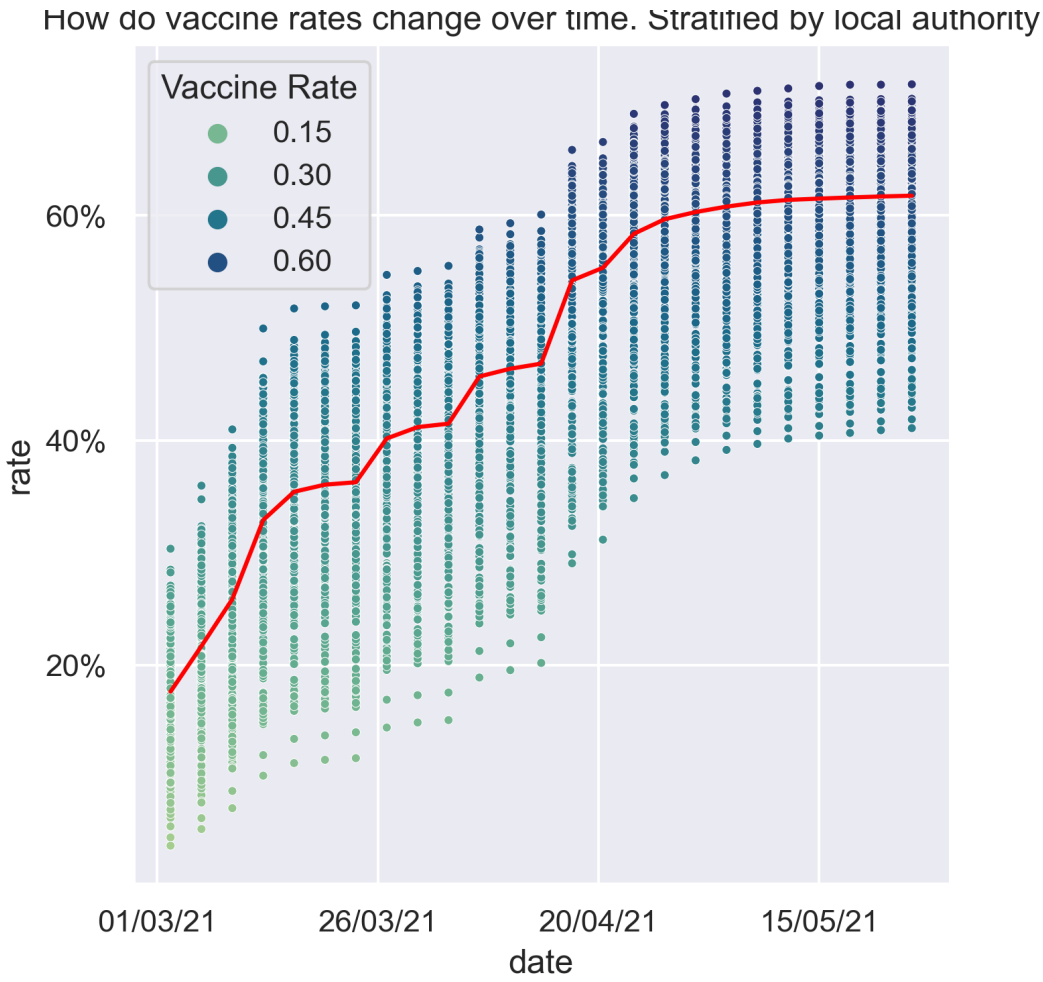


Figure A.4: How do vaccine rates change over time. Stratified by local authority.