Exploring the Role of Data Analytics in Combating Future Pandemics: Lessons Learned from the COVID-19 Pandemic in the United Kingdom

Stewart Charles Fisher II
Student ID: 25020928
QLS ID: FIS20772963
25020928@students.lincoln.ac.uk

April 2023

UNIVERSITY OF
LINCOLN

School of Computer Science
University of Lincoln
United Kingdom

Submitted in partial fulfilment of the requirements for the Degree of MComp Computer Science

Supervisor Dr. Kamaran Fathulla

Word Count: 11767

Acknowledgements

To my family, I would like to express the deepest gratitude, for your support over the past years. Through your efforts over my life, I am the first amongst us to be able to go to university. To know the humble beginnings and hard struggles of both of you, my parents, and those who came before, it is nothing short of inspiring for me. I only hope that I have made you proud.

To my friends, I cannot express enough how much I have valued the time we have spent together, no matter how little. It really is quite something to know that you have someone there to listen to odd ramblings in the quiet hours.

To the university, in particular Dr. Kamaran Fathulla, for giving me the privilege of an education, for the guidance in a field that was completely foreign to me prior to exploring, and for the advice and critique that I might not have liked at the time but was ultimately helpful in the end.

Abstract

Objective To examine the efficacy of applying Big Data techniques in modelling and visualising COVID-19 data from the United Kingdom, and to examine the impact of different regression analysis models in determining the factors of transmission cases and mortality.

Design A retrospective study, collecting COVID-19 data from government sources and analysing the data using Big Data techniques and statistical modelling.

Setting Publicly available data was collected from government sources in the United Kingdom. The data extracted was specific only to regions in England; regions from Scotland, Wales, and Northern Ireland could not be included. Analysis was conducted using a Jupyter Notebook file with the Python libraries Pandas, PySpark, NumPy, and scikit-learn. Data cleaning and pre-processing was carried out prior to the analysis.

Participants COVID-19 data records from individuals who have reported their COVID-19 infection and vaccinations statuses to government sources.

Conclusion Data visualisation and regression analysis are effective tools for analysing COVID-19 data and other pandemics, in order to identify patterns and relationships. Regression analysis in particular was shown to be successful in modelling factors in mortality. Data visualisation is a necessary supplement for the regression analysis, although the choice of the visualisations used should be tailored to the needs of the reader, as well as the applicability of the data.

Keywords COVID-19, Big Data, Data Visualisation, Machine Learning, Regression Analysis

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 The Project . . . 2
- 1.3 The Significance of the Project . . . 2
- 1.4 Aims and Objectives of the Project . . . 3
  - 1.4.1 Aims . . . 3
  - 1.4.2 Objectives . . . 3

2 Literature Review **4**
- 2.1 Previous Studies . . . 5
  - 2.1.1 Concerning the Handling of the Data . . . 5
  - 2.1.2 Concerning the Use and Impact of Data Visualisation . . . 6
  - 2.1.3 Concerning the Use and Impact of Regression Modelling . . . 7
  - 2.1.4 Concerning the Quality of Datasets for Use in Big Data . . . 8
- 2.2 Intended Contributions . . . 10

3 Requirements Analysis **12**
- 3.1 Stakeholders of the Project . . . 12
- 3.2 Requirements of the Data Collection . . . 12
  - 3.2.1 Source of the Data . . . 12
  - 3.2.2 Focus of the Data . . . 13
  - 3.2.3 Release Date of the Data . . . 13
- 3.3 Requirements of the Data Quality . . . 13
- 3.4 Requirements of the Data Processing . . . 13
  - 3.4.1 Python Libraries . . . 13
  - 3.4.2 Code Environment . . . 14
  - 3.4.3 Data Cleaning and Transformation . . . 14
  - 3.4.4 Data Storage . . . 14
- 3.5 Requirements of the Data Visualisation . . . 14
  - 3.5.1 Python Libraries . . . 14
  - 3.5.2 Communication of the Data Visualisations . . . 14
  - 3.5.3 Accessibility of the Data Visualisations . . . 15
- 3.6 Requirements of the Regression Modelling . . . 15
  - 3.6.1 Python Libraries . . . 15
  - 3.6.2 Metrics of the Regression Modelling . . . 15
  - 3.6.3 Accessibility of the Regression Modelling . . . 15
- 3.7 Requirements of the Conclusive Analysis . . . 15
  - 3.7.1 Project Diary . . . 15
  - 3.7.2 Regression Model Documentation . . . 15

4 Design and Methodology **16**
4.1 Project Management . . . 16
4.1.1 Task Scheduling . . . 16
4.1.2 Progress Reporting and Project Documentation . . . 17
4.2 Evolutionary Prototyping . . . 17
4.3 Software Development Tools . . . 18
4.3.1 Version Control . . . 18
4.3.2 Jupyter Notebook . . . 18
4.3.3 TensorFlow . . . 18
4.4 Graphical Implementations . . . 19
4.5 Regression Model Algorithms . . . 19
4.5.1 Linear Regression . . . 19
4.5.2 Ridge Regression . . . 19
4.5.3 Lasso Regression . . . 20
4.5.4 Elastic Net Regression . . . 20
4.5.5 Bayesian Ridge Regression . . . 20
4.5.6 Huber Regression . . . 20
4.5.7 Theil-Sen Regression . . . 20
4.5.8 Orthogonal Matching Pursuit . . . 21
4.5.9 Support Vector Regression . . . 21
4.5.10 Logistic Regression . . . 21
4.6 Dataset Acquisition . . . 22
4.7 Performance Metrics . . . 22
4.8 Risk Analysis . . . 22
4.8.1 Availability of the Data . . . 22
4.8.2 Data Privacy . . . 22
4.8.3 Technical Risk . . . 23

5 Implementation **24**
5.1 Data Preparation . . . 24
5.1.1 Data Ingestion . . . 24
5.1.2 Data Formatting . . . 24
5.2 Data Visualisation . . . 26
5.2.1 Graphic Selection . . . 26
5.2.2 Data Aggregation . . . 26
5.2.3 Data Enhancement . . . 26
5.2.4 Time-Series Plots . . . 26
5.2.5 Proportion Graphics . . . 29
5.3 Regression Model Analysis . . . 32
5.3.1 Model Selection . . . 32
5.3.2 Data Aggregation . . . 32
5.3.3 Model Creation . . . 33
5.3.4 Model Evaluation . . . 34
5.3.5 Cross-Model Evaluation . . . 36
5.4 Miscellaneous Issues . . . 37
5.4.1 Male and Female Deaths Datasets . . . 37
5.4.2 General Cases Dataset . . . 37
5.4.3 Incorrect Encoding of the Date . . . 37
5.4.4 Support Vector Regression . . . 37

6 Results and Discussion **38**
6.1 Data Visualisation 38
6.1.1 Findings on Potential Transmission Vectors 38
6.1.2 Findings on Variants 42
6.2 Regression Analysis 43
6.2.1 The Impact of Sex on COVID-19 Cases 43
6.2.2 The Impact of Variants on COVID-19 Cases and Deaths . 44
6.2.3 The Impact of Vaccinations on COVID-19 Cases and Deaths 46
6.3 Remarks on the Approaches 47

7 Conclusion **48**
7.1 Suitability of the Aims and Objectives 48
7.2 Improvements of Research 48
7.3 The Quality of the Data 48
7.4 Reflections on the Implementation 49
7.5 The Results of the Project 49

Bibliography **50**

A Obsidian Templates **54**

B Project Canvas **55**

C Model Prototype Documentation **56**

List of Figures

1.1 An example of regression modelling (Matulić, 2007). 2

2.1 An implementation of an ASF framework (Elmeiligy et al., 2020). 5
2.2 A treemap of the S&P 500 stock market index (FINVIZ, 2023). . 6
2.3 A combination of GIS regression and ML regression (Almalki et al., 2022). 7
2.4 The prediction results of modelling (Khan et al., 2021). 8
2.5 The proposed model of quality assessment by Cai et al. (Cai & Zhu, 2015). 10

4.1 The logistic sigmoid function (Richards, 2008). 21

5.1 A multi-line cumulative time-series plot. 27
5.2 A single-line discrete time-series plot. 28
5.3 A flawed multi-line cumulative time-series plot of once-vaccinated patients by region over time. 29
5.4 A pie chart. 31
5.5 A treemap. 32
5.6 A raw output of the `.coef_` array. 34
5.7 A clean output of the `.coef_` array. 35

6.1 A selection of pie charts from the artefact. 39
6.2 Sigmoidal characteristics displayed in younger age groups. 39
6.3 Linear characteristics displayed in older age groups. 40
6.4 Erroneous output from the General Cases dataset. 40
6.5 A comparison of male and female increase in cases by region over time. 41
6.6 A comparison of the magnitude of cases across age groups. 42
6.7 A plot of reinfections, likely influenced by the Omicron variant. . 42

B.1 An image of the project canvas on Obsidian. 55

List of Tables

4.1 The first implementation of the project Gantt chart (Fisher, 2022). 16
4.2 The second implementation of the project Gantt chart (Fisher, 2023). 17
4.3 The datasets collected from the UK Government Coronavirus Dashboard. 22

6.1 Cross-model evaluation for *The Impact of Sex on COVID-19 Cases*. 43
6.2 Coefficients for the *The Impact of Sex on COVID-19 Cases*, from COVID-RMA048. 44
6.3 Cross-model evaluation of *The Impact of Variants on COVID-19 Cases and Deaths*. 45
6.4 Coefficients for the *The Impact of Variants on COVID-19 Cases and Deaths*, from COVID-RMA049. 45
6.5 Cross-model evaluation of *The Impact of Vaccines on COVID-19 Cases and Deaths*. 46
6.6 Coefficients for the *The Impact of Vaccines on COVID-19 Cases and Deaths*, from COVID-RMA061. 47

Listings

5.1 The function of zero-padding. 25
5.2 An implementation of a multi-line cumulative time-series plot. . . 28
5.3 An implementation of the peak calculation. 30
5.4 An implementation of treemaps. 30
5.5 An implementation of data aggregation for the regression models. 33
5.6 An implementation of input data preparation for the regression models. 33
5.7 An implementation of the date correction for the regression model data. 34
5.8 An implementation of the coefficient table. 35
5.9 An implementation of the cross-model evaluation table. 36

Chapter 1

Introduction

1.1 Background

Beginning in late 2019, the COVID-19 pandemic had an unprecedented impact on the world and our healthcare institutions, with ∼760 million cases and ∼6.9 million cases having been reported globally (Organisation, 2023). The speed at which the pandemic progressed highlighted the critical importance of data and our ability to collect, process, and analyse it accurately. Governments, health organisations, and researchers have been closely working together to gain an understanding of the virus, in order to mitigate its spread with effective intervention methods. Due to the immense volume and complexity of the data that is continuously being generated by the pandemic, researchers have to employ Big Data techniques to be able to carry out their analyses.

Big Data is the use of advanced analytic techniques on quantities of data that surpass the capabilities of conventional databases to manage and process data in an efficient manner (IBM, 2021); it is crucial that the methods we use to study the data of the COVID-19 pandemic and future pandemics allow us to follow the progression as it unfolds so that public health policies influenced by the studies can be fruitful rather than outdated.

Data visualisation is a key technique in Big Data, whereby we create visual representations of the data to aid administrators to understand complex information in a quick and simple fashion; these visualisations can take many forms, such as graphs, maps, and infographics. By displaying data this way, an administrator or a general reader will be able to interpret trends and patterns that might have otherwise been unknowable when looking at the raw data. Notable uses of real-time data visualisation throughout the COVID-19 pandemic include hotspot identification and transmission rate identification (University, 2022).

While data visualisation is effective for delivering an interpretation of the data in a visual format, it does not allow the reader to gain a true understanding of the underlying relationships in the data. Large and complex datasets, such as those collected in the COVID-19 pandemic, can often contain multiple independent variables. To model the relationship between the independent variables and the dependent variables we are interested in, we can use regression analysis. Regression analysis is a method of statistical analysis that uses a best-fitting line to determine the relationships of the variables, see Figure 1.1. The objective of regression analysis is to minimise the distance between the best-fitting line, which represents the predicted values, and the observed values (Simplilearn, 2017); this distance is called the residual. A regression model will aim to identify the vari-

ables that have a larger or lesser impact on the outcome of the dependent variable. The strength of each independent variable is represented by a beta coefficient.

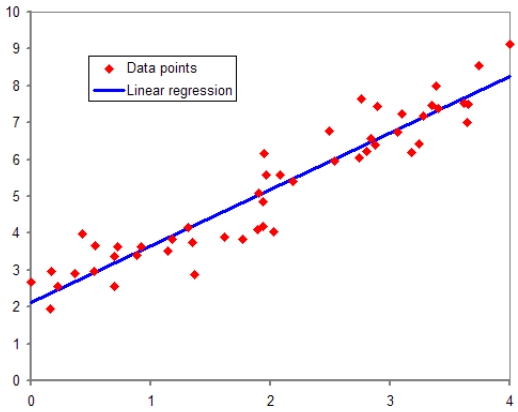

Figure 1.1: An example of regression modelling (Matulić, 2007).

1.2 The Project

In contrast with data visualisation, regression modelling is limited to modelling the relationship between individual variables and a dependent variable, and cannot provide a display of the complex patterns that exist within the data. By using a combination of the two techniques on multifaceted data harvested from the United Kingdom, the project was able to gain a more complete understanding of how the COVID-19 pandemic progressed in the United Kingdom, as well as a unique insight into the advantages and disadvantages of using particular models of regression analysis when applied to the harvested datasets, in order to determine the factors that lead to transmission or mortality.

1.3 The Significance of the Project

The project aims to explore the efficacy of data visualisation and regression analysis when applied to COVID-19 data from the United Kingdom, with a particular focus on examining risk factors that lead to transmission cases or deaths due to COVID-19. While the techniques used in this project are already established, there has been relatively little research conducted on how regression analysis can be applied to the particular datasets used in this project. By addressing this gap in research, this project has the potential to provide a valuable insight into the usage of Big Data techniques to combat a future pandemic.

Furthermore, this project aims to address the successes and shortcomings of the dataset; in order for Big Data research to be gainful, the data being used needs to be high quality and well structured, to allow for easy processing (Cai & Zhu, 2015). If data contains incomplete entries or inconsistencies, this can prevent effective research. Through examining the potential improvements to the data itself and how it is collected, this project could perhaps aid future research that uses the same source of data.

1.4 Aims and Objectives of the Project

1.4.1 Aims

- To explore the efficacy of applying Big Data techniques, particularly data visualisation and regression analysis, to COVID-19 data from the United Kingdom.
- To identify ways that the quality of the COVID-19 datasets from the United Kingdom, specifically governmental sources, can be improved.
- To contribute to the development of more effective analysis strategies for future pandemics, occurring either locally in the United Kingdom or internationally.

1.4.2 Objectives

- To collect and process COVID-19 data from official governmental sources in the United Kingdom.
- To develop and apply data visualisation techniques to explore and display multivariate patterns and complex trends that exist within the COVID-19 data.
- To apply regression analysis modelling to the COVID-19 data, to investigate the existence of relationships between key variables and outcomes, specifically transmission cases or mortality.
- To assess the the quality and consistency of the COVID-19 datasets acquired from the governmental sources in the United Kingdom.
- To develop methods to clean the harvested COVID-19 data, so that it is more suitable for the techniques that will be applied to it.
- To compare the results of the data visualisation and regression analysis, to find how the two different approaches compliment each other.

Chapter 2

Literature Review

This literature review aims to provide a comprehensive review of the existing literature concerning the use of Big Data techniques, such as data visualisation and regression analysis, to analyse COVID-19 data. Throughout the COVID-19 pandemic, the vast amounts of data that have been generated, as well as the plurality of available sources, have expedited the requirement for effective techniques to allow us to understand a progressing pandemic and how to guide public health policy decisions. Even three years on, it is still possible that we have not entirely perfected the methods we chose to combat this data.

The following literature review is focused on six main research objectives:

- What is the impact of using data visualisation and linear regression modelling in analysing COVID-19 data, concerning the provision of insight into its spread and the impact of the pandemic?
- How can the existing COVID-19 datasets be optimised and improved for usage with Big Data techniques, such as data visualisation and linear regression modelling?
- How do the results from using linear regression modelling compare to results acquired from alternative statistical analysis methods, such as non-linear regression modelling, in terms of accuracy and robustness?
- How can the insights and findings from the analysis of COVID-19 data using Big Data techniques be used to inform and guide public health policies in the United Kingdom?
- Are there any ethical considerations to be aware of when using Big Data techniques to analyse COVID-19 data, and how can these considerations be addressed and mitigated?
- What are the potential limitations and challenges of using Big Data techniques, such as data visualisation and linear regression modelling, to analyse COVID-19 data, and how can these be addressed and mitigated?

Through a critical evaluation of previous studies and existing literature, the project aims to demonstrate gaps in the current knowledge and highlight areas for further research, which the project hopes to provide. This literature review serves as the basis upon which the project conducted the subsequent analysis of the COVID-19 data for the United Kingdom, in which the project will use data visualisation and regression analysis techniques. This is an extended version of the literature review in the interim report (Fisher, 2023).

2.1 Previous Studies

2.1.1 Concerning the Handling of the Data

A study in 2020, *A Multi-Dimensional Big Data Storing System for Generated COVID-19 Large-Scale Data using Apache Spark*, outlined how Apache Software Foundation implementations, specifically Apache Spark could be integrated to handle the large quantities of data so that it could be processed efficiently.

Researchers from Mansoura University, Egypt (Elmeiligy et al., 2020) conducted research into the the value of using an ASF framework to analyse data; they used the Hadoop Distributed File System (HDFS) to divide the inserted data into a set of Resilient Distributed Datasets (RDDs), as seen in Figure 2.1. HDFS is a distributed file system designed to provide fault-tolerant data management (Borthakur, 2022). Spark is an analytics engine designed for fast data processing; it provides a programming interface to allow access to data parallelism (Foundation, n.d.-b).

Their research found that by implementing the framework to segment the data, they were able to increase the system performance in their modelling.

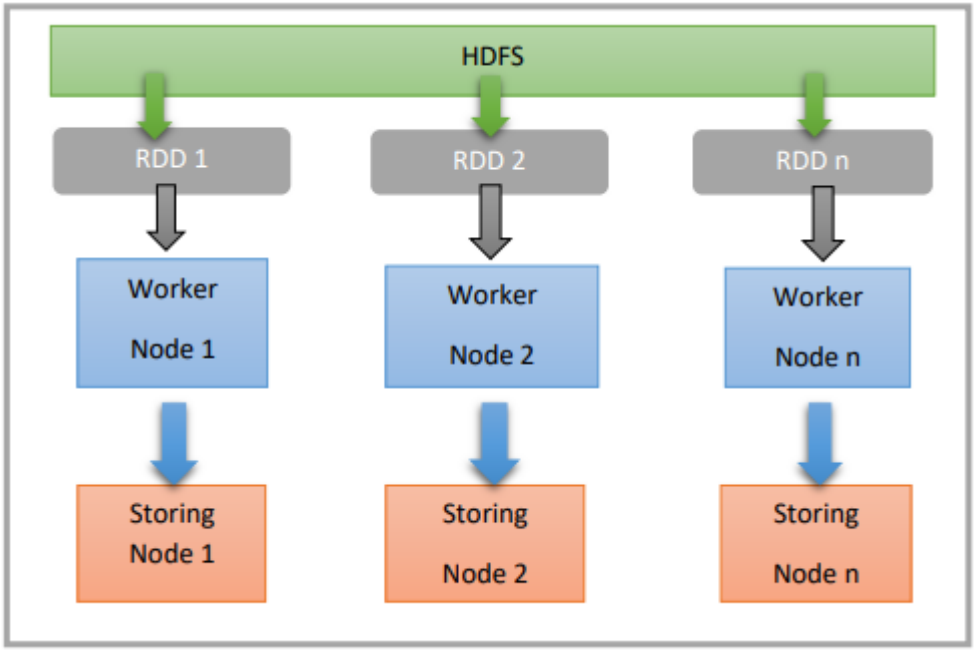

Figure 2.1: An implementation of an ASF framework (Elmeiligy et al., 2020).

However, while PySpark, the Python API for Spark, may be more efficient for handling larger datasets, there are scenarios where pandas may be a more efficient choice. Pandas is a Python library, created by Wes McKinney, designed for data manipulation and analysis, which provides an easier user experience and efficient data structures similar to those provided by PySpark, such as the DataFrame (McKinney, 2011). While pandas can offer higher efficiency than PySpark, it does not offer the ability to use RDDs across a cluster of machines, and so a user is limited by the amount of memory on a single machine.

2.1.2 Concerning the Use and Impact of Data Visualisation

The early progressions in modern data visualisation began with the advent of computer graphics, notably in a study in 1989; DeFanti et al. determined that as the definition of a supercomputer changed, going from 0.1 to 1.0 gigaFLOPS[1] to 1.0 to 10 gigaFLOPS, computer scientists would require a solution to handle the barrage of data in order to conduct research (DeFanti et al., 1989).

By the mid-1990s, Ben Shneiderman et al. of the University of Maryland Human-Computer Interaction Lab had invented treemapping (Shneiderman, 1992). Treemapping is a method of data visualisation that can be used to display hierarchical data in nested rectangles (Jadeja & Shah, 2015). Each rectangle in the structure represents a node of the tree structure, with the size of each rectangle representing the proportional value of the respective node; these rectangles are generated with tiling algorithms (Vernier et al., 2020). The colour coding of the squares can also be used to represent additional variables or categories. Treemapping has notably been used for financial analysis of stocks markets, see Figure 2.2.

Figure 2.2: A treemap of the S&P 500 stock market index (FINVIZ, 2023).

While treemaps can visualise data with an efficient use of space, they can suffer if the data that is examined is too large or too complex; vast proportional differences between nodes, or large quantities of nodes, can make a treemap almost unreadable. To mitigate this, a user must either be selective with the data they put into the model or choose an alternative.

The context of the data can affect the adequacy of different data visualisation formats; a study in 2018 by Saket et al. examined how tasks would determine the efficiency of basic visualisation models, such as tables, line charts, bar charts, scatter plots, and pie charts (Saket et al., 2019). The authors found that users prefer pie charts and bar charts for identifying clusters, and line charts and scatter plots for identifying correlations. It was also noted that line charts should be avoided in cases were precises value identification is necessary, due to the uni-

[1] 1 gigaFLOPS is equal to one billion floating point operations per second. (Base, 2020)

formity of the axes values, and that pie charts should be avoided for correlation tasks.

2.1.3 Concerning the Use and Impact of Regression Modelling

Regression analysis is an area of data analysis that has already shown some promise; in early 2022, there was a study into the usage of regression analysis for COVID-19 infections and deaths due to issues with access to food, and health (Almalki et al., 2022). The study used a combination of a machine learning (ML) regression model and a geographic information system[2] (GIS) regression model, as seen in Figure 2.3; the authors used scikit-learn software for the ML regression and ArcGIS-ArcMap software for the GIS analysis. The cases and deaths were taken as the dependent variables and the remaining factors as independent variables. The results showed that while there were instances of strong correlation between the independent variables, there was only weak correlations being presented between the dependent and independent variables, across both methods.

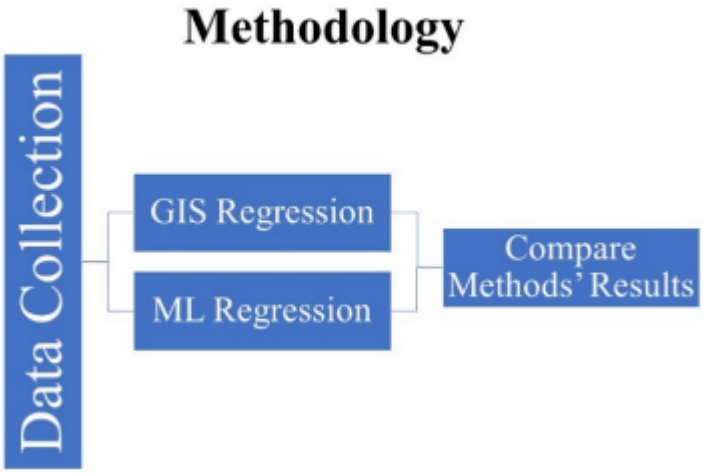

Figure 2.3: A combination of GIS regression and ML regression (Almalki et al., 2022).

Other forms of regression analysis, beyond linear regression have been found to produce compelling results when identifying relationships within data; in 2020, a study at Yildiz Technical University, Turkey (Qudrat-Ullah & Tsasis, 2017) implemented a negative binomial regression analysis to determine the impact of human mobility on the causality of COVID-19 cases. Unlike the prior study, this study displayed relationships with the dependent variables, showing that there are correlations between COVID-19 cases and the volume of airline traffic, and also the number of airports.

Alongside causality analysis, regression modelling has been used to predict outcomes. Researchers from the College of Computing and Information Technology at the University of Bisha in Saudi Arabia, the ABES Institute of Technology in India, and the Department of Computer Science and Engineering at Graphic

[2]A geographic information system pairs geographic data with software tools to manage and analyse them. (Chang, 2019)

Era Hill University in India conducted a joint study to evaluate the usage of five different regression models in predicting cases, deaths, recoveries internationally (Khan et al., 2021). Their analysis found that polynomial ridge regression was the best at predicting the number of cases, the linear regression model was the best at predicting the number of recoveries, see Figure 2.4, and the support vector machine regression model was the best at predicting the number of death cases.

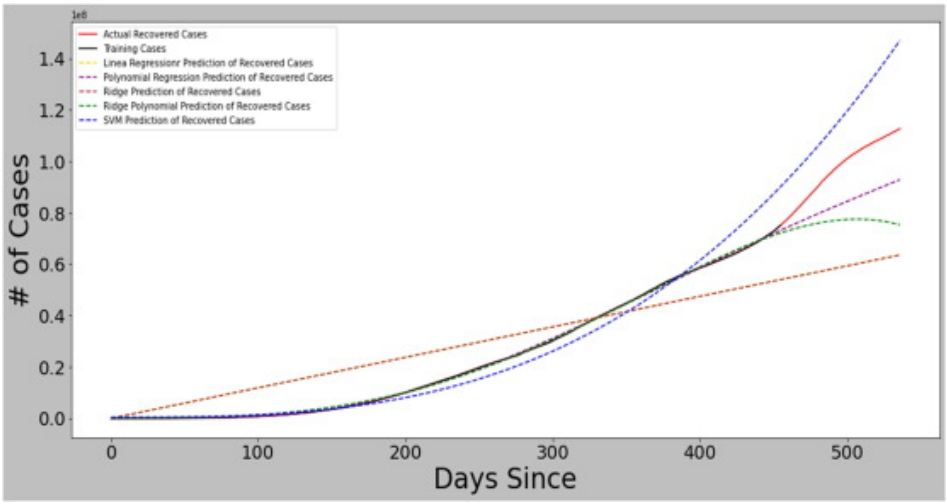

Figure 2.4: The prediction results of modelling (Khan et al., 2021).

2.1.4 Concerning the Quality of Datasets for Use in Big Data

The quality of the data that is being used in data analysis can affect the results produced and, consequently, the conclusions that can be drawn. In 2015, at study was conducted into the contemporary processes that were being employed in the field of Big Data (Juddoo, 2015). The study found that traditional processes, such as conditional functional dependencies (CFD) and denial constraints (DC) to ensure quality, there were still improvements that needed to be made. Furthermore, the latest methods required more development in order to be able to address the challenges posed by the volume, variety, and velocity of Big Data; this study doesn't address ways in which these processes can be improved.

Also in 2015, a MITRE research team sponsored by the United States Air Force conducted four separate studies into the applications of Big Data and how the issues with data quality differed from issues in traditionally-sized data collection (Becker et al., 2015). The four studies were:

- *Hyperspectral Imaging (HSI) Using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)*
  - An exploration of the use of HSI technology, with a particular focus on the data that is generated by the AVIRIS.
- *The Whole [Human] Genome Sequencing Project (WGS)*
  - An examination of how advances in computing and data storage technologies have enabled significant progress in gene sequencing research,

considering the quality of the data and contextual information of bioinformatics[3] and computational genomics.

- *U.S. Army Medical Command (MEDCOM) Medical Operational Data System (MODS)*
  - A discussion of the MODS data warehouse project, that aims to improve medical readiness and eventually progress onto improving other health-related areas.
- *Federal Aviation Administration (FAA)/Center for Advanced Aviation System Development (CAASD): Threaded Track and Leviathan Projects*
  - An evaluation of the use of platforms, such as the Hadoop MapReduce (Foundation, n.d.-a) stack, to gain insight into data processing and workflow variations across the analytic pipeline.

The meta-study highlighted several key findings related to the quality of Big Data. A notable discovery was that despite the volume of the data increasing by large orders of magnitude, the quality of the data doesn't exhibit a change; the number of issues in the data is proportional to the increase in the volume. Additionally, data errors that have been manually generated can be difficult to find and are typically semantic in nature, making these errors undetectable to common mechanisms of error detection. Meanwhile, Big Data that has been acquired from an automated source can exhibit signs of quality increase, thereby improving the ease of use of the data and removing the potential for human error. Other findings of the study suggested that errors that appear in Big Data can be treated as noise, and that Big Data algorithms can introduce unintended errors into the data; in the context of data, noise refers to any unwanted variation in the data that can disrupt or distort the perception of a particular pattern or relationship. Despite this, the authors stated a sentiment similar to the prior study; improvements could be made to mitigate these issues. Two noteworthy recommendations of this study were to utilise a Big Data quality framework and to minimise the chance of errors in the harvested data by minimising the number of sources.

Cai et al. examined the challenges produced by data quality in the Big Data Era, to establish their own model on how to assess and ensure data quality (Cai & Zhu, 2015). The authors determined that in order for a researcher to ensure the quality of their data, a research should evaluate the data they plan to collect according to five meta-categories:

- Availability
- Presentation Quality
- Relevance
- Reliability

[3]Bioinformatics is the application of computing and statistical analysis to analyse biological data and processes, such as genetics, to identify patterns in the large data that is generated. Bioinformatics plays a large role in drug discovery and the personalisation of medicine (Bayat, 2002).

- Usability

The model the authors designed, see Figure 2.5, is designed to address all five of the meta-categories. By employing a data quality framework with a hierarchical structure, with a dynamic quality assessment, the authors believe this model can ensure data quality by combining external, contextual knowledge of the data with an internal requirement satisfaction analysis.

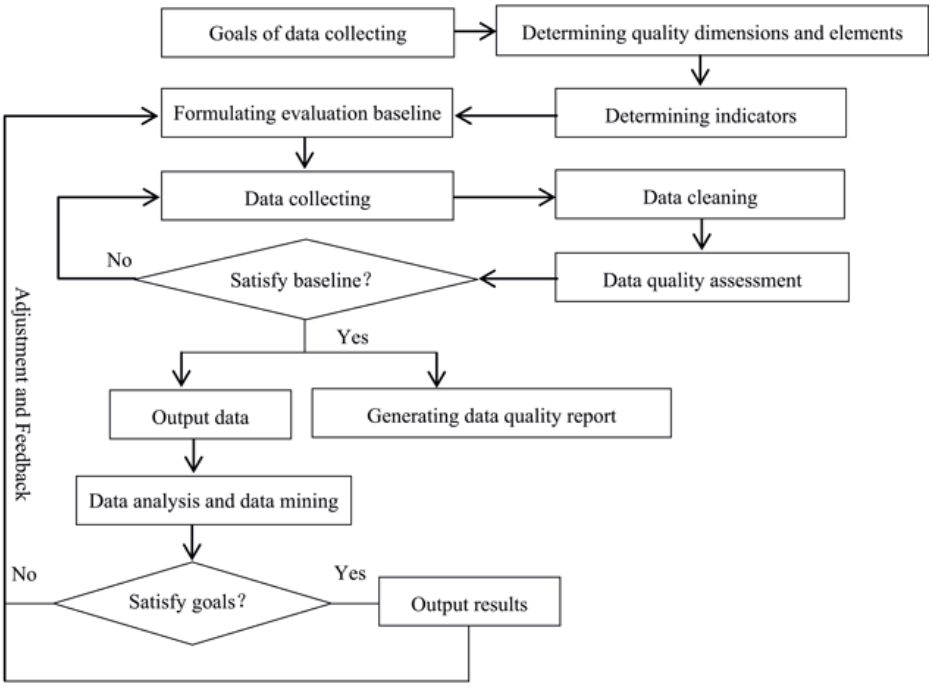

Figure 2.5: The proposed model of quality assessment by Cai et al. (Cai & Zhu, 2015).

2.2 Intended Contributions

Upon final summary of the literature, it is clear that there is potential for Big Data techniques, especially data visualisation and regression analysis, to be integrated into our health crisis management protocols, to observe and track progressions.

Regarding the management of the data, it is evident that both pandas and PySpark display advantages to the user. Neither tool can replace the other as they are designed for different environments, with pandas being a data manipulation library and PySpark being a distributed computing framework; a combination of both tools is possible and can enable a researcher to effectively manage and analyse data. Therefore the project shall primarily employ the use of the pandas library, and will also explore how the PySpark framework can be implemented into the artefact if the volume of the data requires me to do so.

Moreover, data visualisation is an essential aspect of a data analysis project, capable of providing vital communication of findings to stakeholders, so that they can identify the complex patterns and trends that appear within the data. In the artefact, the project will focus on the usage of line graphs, for their ability to display correlative relationships in the data. The project will also focus on the usage of treemapping and pie charts, for their ability to display clusters in the data in a proportional fashion.

In addition to data visualisation, regression analysis has also shown its potential to be a crucial statistical analysis tool, to analyse the relationship between complex variables in the data. In the artefact, the project will focus on using a variety of linear-based regression models, as non-linear-based regression models have a higher complexity in the coefficients that can be obtained; linear-based regression models assume a linear relationship and so there is only one beta coefficient between each independent variable and dependent variable relationship. Despite this, linear-based regression models have shown promise, as seen in the study by Khan et al. Therefore, the models that the project will look towards are:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Bayesian Ridge Regression
- Theil-Sen Regression
- Huber Regression
- Orthogonal Matching Pursuit

Lastly, the quality of the data that is used in the artefact is paramount for producing accurate, complete, and consistent results. The data quality assessment framework that was outlined by Cai et al. will enable me to identify issues within the data and implement a process of measures to proactively mitigate them. By doing this, there can be confidence that the results that are gained from the research is truthful.

In conclusion, with this research and the artefact, the aim is to improve upon prior studies by evaluating how established data visualisation and regression modelling techniques can be applied to government sourced data on COVID-19 from the United Kingdom. By doing so, this project aims to provide a valuable insight into how administrators can address the current pandemic and future outbreaks. There is also an aim to provide insight into the relationships between variables, to discover potential factors in transmission cases and mortality.

Chapter 3

Requirements Analysis

3.1 Stakeholders of the Project

A stakeholder is any individual, group, or organisation that has a vested interest or concern in the the outcome and the results of this project. By understanding who the stakeholders of this project are, the project can be tailored towards satisfying their demands. The five key stakeholders identified in this project are:

Researchers The project should provide insight into the efficacy of Big Data techniques in analysing the COVID-19 pandemic.

Data Analysts The project should contribute to the advancement of Big Data techniques for future analyses.

Public Health Administrators The project should provide insight into the factors that contribute to the spread and impact of COVID-19.

Government Representatives The project should provide information capable of informing policy decisions for outbreak mitigation and resource allocation.

General Public The project should provide an understanding of the spread of the COVID-19 virus, so individuals or businesses can implement private mitigation strategies.

3.2 Requirements of the Data Collection

3.2.1 Source of the Data

The data that is collected must be obtained from governmental sources in the United Kingdom. Governmental sources are considered to be reliable and authoritative sources of data, and so the COVID-19 data that can be obtained from these sources is likely to be accurate, and consistent.

The consistency of the data will allow me to compare different regions from within the United Kingdom. There are 219 NHS trusts in the United Kingdom, covering 1257 hospitals (Raw, 2022); by using government-sourced data, there can be confidence that the data that has been collected is consistent, so that the analysis can truthfully identify patterns, trends, and relationships.

The data that is sourced from the government will be publicly available, not requiring payment or special access, meaning that other researchers can easily

access the data to continue with their own research, or verify the findings of this project.

As the focus of this project is on the United Kingdom, the data that can be collected from governmental sources is guaranteed to be relevant to the focus of the project; the government will collect a wider range of datasets than a project with an objective. Therefore, the project is more likely to be able to collect all of the information necessary for the analysis, rather than a fraction of it.

Moreover, by using government data, the information obtained in this project could be compared to information obtained from the analyses of data sourced from government sources of other countries, allowing researchers to identify differences in patterns and trends on an international scale, and allowing for additional policy-making to mitigate the international impact.

3.2.2 Focus of the Data

The data that is collected should be limited to the focus of the project; the data should include information such as:

- Cases
- Deaths
- Vaccinations
- Reinfections
- Variants

3.2.3 Release Date of the Data

The data that is collected must have a consistent date of sourcing. During this project, it is highly likely that there will have to be a combination of multiple datasets in order to perform the analysis. As it is possible for datasets to be amended over time, there must be maintenance of datasets that were collected with consistent timing, in case one dataset was to be updated while another was not; if this were to occur and there were not measures in place to anticipate for this, the results of the data analysis could be untruthful.

3.3 Requirements of the Data Quality

Throughout the project, there must be maintenance of the aforementioned framework to ensure that the quality of the data and the subsequent results are of a suitable standard. This framework should be used on a frequent basis, to regularly monitor the project along each step, to ensure the accuracy of the results.

3.4 Requirements of the Data Processing

3.4.1 Python Libraries

The libraries that are selected alongside PySpark and pandas should allow for data manipulation. Libraries capable of fast numerical operations are necessary

to perform calculations on large scale datasets. The libraries used should also allow for complex mathematical and statistical functions, to reduce manual implementation. The chosen libraries should have active communities, with large amounts of documentation, so that issues found during production are easier to solve.

3.4.2 Code Environment

The environment used to produce the project artefact should also be suitable for data processing. As such, it should allow for interactive computing, so that the implemented code can be segmented in blocks, and executed independently from the rest of the code; this will allow immediate results without having to execute the entire code, which could be large.

Since Python is the code language that will be used to produce the project artefact, the environment should support the language, and similar to the code libraries, the environment should also have an active community with a large amount of documentation.

3.4.3 Data Cleaning and Transformation

Data cleaning and transformation techniques should ensure that the data is accurate, complete, and consistent prior to being used for data visualisation and regression analysis, including but not limited to the removal of duplicates, and the standardisation of data formats.

3.4.4 Data Storage

The data should be stored in a format that is suitable for both the data visualisation and regression analysis sections of the artefact. The storage format should be robust and scalable, and it must allow for efficient manipulation of data via the necessary libraries.

3.5 Requirements of the Data Visualisation

3.5.1 Python Libraries

The libraries that are selected for the data visualisation implementations should be capable of handling the scalability that comes along with Big Data. Alongside this, the libraries should be able to be customised, so that the output can be tailored to the needs and requirements of the end user. Lastly, the libraries should have efficient performance; a large number of data visualisations could need to be produced and optimising the time expenditure is crucial to the timeline of the project.

3.5.2 Communication of the Data Visualisations

The project should produce visualisations that are capable of displaying the complex patterns and trends that exist within the COVID-19 data.

3.5.3 Accessibility of the Data Visualisations

The project should ensure that the visualisations that are produced are easily understood, not just by researchers but also by non-technical recipients, most notably public health administrators and government representatives.

3.6 Requirements of the Regression Modelling

3.6.1 Python Libraries

Similar to the data visualisation, the libraries that are selected for the regression analysis implementations should be capable of handling the scalability of Big Data, and should also have a large active communities. Additionally, the libraries should provide a wide range of regression models, for instance, linear regression and ridge regression.

3.6.2 Metrics of the Regression Modelling

Since this project aims to discover the efficacy of using regression modelling on the data, the project must collect the necessary metrics to assess the performance of each model, for instance, the coefficient of determination, the mean squared error and root mean squared error, and the mean average error.

3.6.3 Accessibility of the Regression Modelling

The project should ensure that both the beta coefficients and the visualisations of the predictions are interpretable; unlike the data visualisation section, these outputs are more tailored towards the researchers and the data analysts, who are responsible for interpreting these results and reporting them in a manner that is understandable for non-technical recipients.

3.7 Requirements of the Conclusive Analysis

3.7.1 Project Diary

The project should be accompanied by a project diary, to record any alterations and issue fixes that happen throughout the process; changes that are made can have a severe impact on the outcome of the results and so the project must be able to account for these changes.

3.7.2 Regression Model Documentation

As part of maintaining a recorded history of the project, there must be documentation for every single regression model that is implemented, taking care to document the performance metrics and variables used. Since the artefact won't keep the code for each model that is built, provided the documentation is maintained, the project can easily rebuild any model as necessary to validate the harvested results.

Chapter 4

Design and Methodology

4.1 Project Management

To ensure that the project ran smoothly, it was essential that a set of project management tools was prepared; this need is escalated according to the multiplicity of the project tasks. The tools that were employed in this project involved task scheduling, progress reporting, and project documentation.

4.1.1 Task Scheduling

Being able to properly schedule the tasks of a project is crucial to its success. Scheduling allows a project leader to make sure that the project is completed on time, and to break down the project into smaller tasks that are easier to comprehend; resources can be allocated efficiently, avoiding unnecessary burnout or delay. A popular method of task scheduling is to use a Gantt chart.

A Gantt chart is a visual illustration of a project schedule, first designed by Henry Gantt (Gantt, 1974). Similar in structure to a bar chart, the tasks of the project are listed along the vertical axis while the time intervals are listed along the horizontal axis (Richman, 2002); the granularity of the time intervals is customisable by the user.

During the pre-production phase of the project, it was decided that the production phase project would begin on the 5th of December 2022 and end on the week beginning the 13th of March 2023, lasting ∼16 weeks. The original Gantt chart that outlined the steps and time allocations for the production phase can be seen in Table 4.1; due to the length of the project, weekly intervals were used for the time allocations.

Weeks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Collect data	■	■														
Clean the data			■													
Integrate the data together				■												
Research regression analysis models					■	■	■	■								
Build the regression models									■	■	■	■				
Process the data													■	■		
Document findings															■	■

Table 4.1: The first implementation of the project Gantt chart (Fisher, 2022).

During Week 8 of the project, an update was made to the Gantt chart, as there were adjustments in the time expenditure expectations; the second edition of the Gantt chart can be seen in Table 4.2.

Weeks	9	10	11	12	13	14	15	16
Week Beginning	30/1	6/2	13/2	20/2	27/2	6/3	13/3	20/3
Build and test the statistical modelling for the data.	■	■						
Build and test the regression modelling for the data.			■	■	■			
Analyse the results of the modelling.						■	■	
Document findings.								■

Table 4.2: The second implementation of the project Gantt chart (Fisher, 2023).

4.1.2 Progress Reporting and Project Documentation

As mentioned before, it is crucial to keep a log of the changes and alterations that are made, in order to account for any consequential changes in the results. A diary of the project was kept, detailing every accomplishment, issue, or change during each session of work throughout the production phase and beyond.

It was decided that the project diary would be hosted on Obsidian, a personal knowledge base and note-taking platform that uses Markdown files (Obsidian, n.d.-a). The use of Markdown allows the individual files, known as notes, to be interconnected with each other. Applications similar to Obsidian are Notion and Roam Research; the reasons Obsidian was chosen over the other two options are:

Familiarity There is prior experience using Obsidian, having used the application throughout my tenure at university.

Customisation Obsidian allows for customisable note templates; by using templates, there can insurance that my documentation is consistent. Alongside this, Obsidian allows third-party plugins for extended user experience.

Price Roam Research does not have a free option of usage, unlike Obsidian and Notion; Obsidian is the only option that can be used entirely for free, albeit through the use of plug-ins and GitHub.

Accessibility Everything done through the Obsidian application is saved natively to your local computer, with the option to synchronise the data to a GitHub repository or a proprietary service called Obsidian Sync (Obsidian, n.d.-b). Notion is only hosted on an online server, meaning that if there was to be a server downtime, there would be an inability to access the documentation, thereby hindering the progress on the project.

The templates that were used for the project diary entries and the model prototype documentation can be found in Appendix A.

4.2 Evolutionary Prototyping

In every project that involves software development, there must be a software development methodology. A software development methodology is a structured approach concerning the design, build, and deployment of software, that outlines the processes and practices to follow, in a unique but iterative fashion; no software development approach is the best, and the suitability of an approach varies depending on the individual or team.

For this project, it was decided that the evolutionary prototyping approach would be adopted. Evolutionary prototyping is an approach that relies heavily upon incremental iterations to a codebase, slowly refining it over time (O'Reilly, n.d.). The benefit of using evolutionary prototyping is that allows for fast development of systems that need particular focus. Alongside this, it is the approach that there is the most comfort with, having used it all throughout the previous projects that have undertaken.

4.3 Software Development Tools

4.3.1 Version Control

Git is a version control system designed to allow a developer to track the changes made to a codebase over time, with different versions of the code being stored in branches (Torvalds & Google, 2007). GitHub is one of the complimentary systems that allow a Git repository to be hosted online (GitHub, n.d.); in the event that my system was to fail, my project would be stored in a secondary backup and the event would not be a critical failure for the project.

4.3.2 Jupyter Notebook

Jupyter Notebook is an open source application designed for data scientists, to provide an interactive environment for data analysis. Multiple models can be built and executed within a single document due to the code cell implementation (Team, 2015). By using Jupyter Notebook, a developer can reduce a large amount of time expenditure, when compared to using a traditional code environment. The reason Jupyter Notebook was chosen over similar alternatives was purely a matter of comfort and aesthetic; as with most tools used, there is prior experience, and a preference for the workspace it provides.

Python

Jupyter Notebook offers support for three languages; Python, R, and Julia. Early on in the planning stage of the project, it was decided that the artefact was going to be built in Python. The factors that decided this were:

Confidence Out of the three available options, Python is the only language with prior experience. Beyond this, Python has a reputation for its simplicity to use.

Community Python has one of the largest and most active communities in software development, giving access to a vast wealth of information if necessary.

Libraries As mentioned before, the analysis and visualisation libraries that are available with Python are immensely powerful, and are backed by a large amount of documentation.

4.3.3 TensorFlow

TensorFlow is a library, developed by Google DeepMind, that is designed for machine learning. This library is being implemented to allow for GPU utilisation,

to improve execution times when training the regression models (Google, 2022). While the TensorFlow library does provide regression models via the Keras library, these models will not be used in the artefact, in favour of the utility benefits of using scikit-learn models.

4.4 Graphical Implementations

Line Graphs To implement line graphs, the artefact used the `plot` function from the `pyplot` module in Matplotlib (Hunter, 2007).

Pie Charts To implement pie charts, the artefact used the `pie` function from the `pyplot` module in Matplotlib (Hunter, 2007).

Treemaps To implement treemaps, the artefact used the Squarify library (Laserson, 2023).

4.5 Regression Model Algorithms

The scikit-learn regression models (Pedregosa et al., 2011) used in this project are:

- `LinearRegression`
- `Ridge`
- `Lasso`
- `ElasticNet`
- `BayesianRidge`
- `HuberRegressor`
- `TheilSenRegressor`
- `OrthogonalMatchingPursuit`

4.5.1 Linear Regression

Linear regression is a format of regression analysis, where the model assumes the relationship between a dependent variable and one or more independent variables is linear, and thus can be explained by a linear expression. The aim of linear regression is to find a line of best fit that minimises the sum of squared errors between the predicted values and the actual values for the dependent variable.

4.5.2 Ridge Regression

Ridge regression is a variation of linear regression that is used when the data that is used contains multicollinearity. Multicollinearity is a phenomenon where two or more independent variables in a regression model are highly correlated with each other, thereby making it harder for the model to determine the effects of each independent variable on the dependent variable (Franke, 2010). To counteract this, ridge regression adds a penalty term to the sum of squared errors, in

order to prevent the beta coefficients from being too large, using a regularisation parameter to control the penalty strength; this is called the L2 penalty.

4.5.3 Lasso Regression

Lasso regression is another variation of linear regression that is intended to address multicollinearity in the data. The difference between lasso and ridge regression is that the penalty term used by lasso regression is the absolute value of the coefficients instead of the squared value; this is called the L1 penalty. By doing this, some coefficients are set to zero, effectively performing feature selection. Feature selection is the process of selecting a subset of the relevant features and predictors, in order to remove irrelevant features and thereby increase the performance of the model.

4.5.4 Elastic Net Regression

Elastic net regression combines the capabilities of ridge regression and lasso regression, by using a penalty term that is a combination of the L1 and L2 penalties. Elastic net regression can provide a more flexible approach to regularised regression analysis than lasso regression or ridge regression alone, although this requires specific tuning of the penalty term.

4.5.5 Bayesian Ridge Regression

Bayesian ridge regression is a probabilistic approach to regression analysis that uses Bayesian inference[1] to estimate the coefficients. The model assumes a prior distribution for the coefficients and updates this assumed distribution based on the data. This method allows prior information to be included in the modelling, to acquire updated knowledge.

4.5.6 Huber Regression

Huber regression is a robust variation on linear regression that is prepared to handle outliers that might exist within the data. Huber regression does this by using the Huber loss function, which combines the squared loss of linear regression and the absolute loss of robust regression models; when the residual is small, the loss function tends towards the squared loss, and while the residual is large, it will tend towards the absolute loss.

4.5.7 Theil-Sen Regression

Theil-Sen regression is another robust regression variation, which computes the slopes between all pairs of points and then computes a median. This method is useful for analysing datasets that contain outliers. Theil-Sen does not rely upon assumptions of the data distribution.

[1]Bayesian inference is a method of statistical inference that takes into account prior knowledge, in accordance with Bayes' theorem (Joyce, 2016).

4.5.8 Orthogonal Matching Pursuit

Orthogonal matching pursuit is a sparse regression model that aims to find a subset of independent variables that are the most relevant to the dependent variable; it is commonly employed when the number of independent variables is larger than the number of dependent variables. The model will iteratively select the independent variable that appears most relevant and add it to the model, until a specific limit of variables has been hit or the model can no longer iterate.

4.5.9 Support Vector Regression

Support vector regression is a method of regression analysis that aims to find the hyperplane in the feature space that maximises the margin between the closest points of different classes. A hyperplane is a surface that separates the data into two groups, based upon their predicted values. When a linear kernel is used for support vector regression, the resulting hyperplane is a linear function of the independent variables.

4.5.10 Logistic Regression

Logistic regression is a regression analysis method that models the relationship between a binary dependent variable and one or more independent variables. Unlike linear regression, logistic regression predicts the probability of the dependent variable being in a particular class. The sigmoid function, see Figure 4.1, is used to transform the linear regression equation into the probability scale. Even though logistic regression is not itself a linear-based regression model in the strictest sense, it is a type of generalised linear model that is capable of being interpreted in a similar fashion to typical linear-based regression models. The coefficients that can be calculated from logistic regression represent the strength and direction of the relationship between the predictor variables and the probability of the outcome variable.

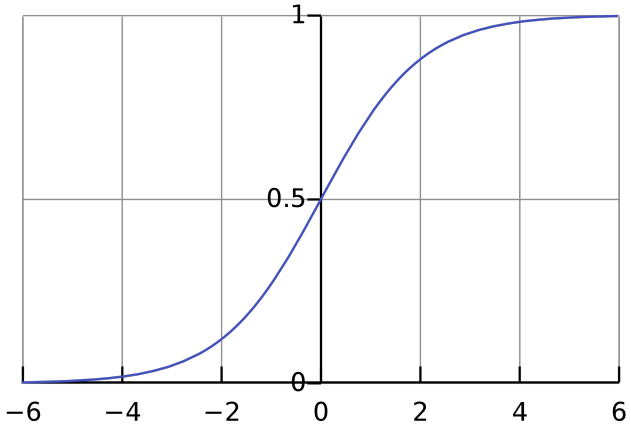

Figure 4.1: The logistic sigmoid function (Richards, 2008).

This model was erroneously implemented into my artefact, as the data does not display a binary outcome; as the model is included within the Linear Models module, it was falsely believed to be a linear regression model. Despite this, the results of this model can be seen in Appendix C but will not be included in the final evaluation.

4.6 Dataset Acquisition

The data that is being used in this project is sourced from the UK Government Coronavirus Dashboard (Government, 2022a); the dashboard is updated every Thursday at 4pm. The dashboard allows researchers to collect data from the dashboard via an API. The datasets that were collected for use in this project were all released on the 22nd of December 2022, are specific to the region of reporting, and can be seen in Table 4.3.

Dataset	**Description**	**Source**
Male Cases	A dataset of the cumulative male cases, by age and region.	Government, 2022f
Female Cases	A dataset of the cumulative female cases, by age and region.	Government, 2022b
Male Deaths	A dataset of the cumulative male deaths, by age and region.	Government, 2022g
Female Deaths	A dataset of the cumulative female deaths, by age and region.	Government, 2022c
General Cases	A dataset of the cumulative cases, by region.	Government, 2022d
General Deaths	A dataset of the cumulative deaths, by region.	Government, 2022e
Vaccinations	A dataset of the cumulative populations of vaccinated patients, by the number of vaccinations received, age, and region.	Government, 2022i
Reinfections	A dataset of the cumulative reinfection cases, by region and age.	Government, 2022h
Variants	A dataset of the cumulative variants sequenced, by region.	Government, 2022j

Table 4.3: The datasets collected from the UK Government Coronavirus Dashboard.

Note that all of the datasets used are specific to England; the API provided by dashboard does not provide regional data for Scotland, Wales, or Northern Ireland.

4.7 Performance Metrics

The performance metrics from the regression model prototypes are saved in both the prototype documentation, and in a `csv` file. The metrics are observed using the Metrics module from the scikit-learn library.

4.8 Risk Analysis

4.8.1 Availability of the Data

There was a risk that the data that was supposed to be collected could have been unavailable. As the aim of this project is specific to the use of governmental sources, this issue could have been catastrophic to the success of this project.

4.8.2 Data Privacy

The data that was collected for this project had to be ethically sourced and handled; the Coronavirus Dashboard collects and presents all of their statistics in line with the Code of Practice for Statistics, which affirms that researchers can have 'confidence that published government statistics have public value, are high quality, and are produced by people and organisations that are trustworthy' (for Statistics Regulation, 2022).

4.8.3 Technical Risk

The computing power required to process the implementations of the project could exceed that available to me on my home device. In order to mitigate this issue, care was taken to use a system equivalent to those available in the university laboratories. Furthermore, the artefact could be processed via Google Colab, which has a wider availability of processing power.

Chapter 5

Implementation

5.1 Data Preparation

5.1.1 Data Ingestion

The data from each dataset was loaded into the Jupyter Notebook environment by reading the respective `csv` files into pandas DataFrames. Pandas is able to infer the tabular structure of the data, where each column represents a feature and each row represents an observation (pandas, n.d.). Alongside this, pandas assumes that the first row of the imported data contains the header data for each column and will assume the datatype of each column. Despite this, it is still possible for the inference to be incorrect and so there had to be insurance that the datatypes were correct; in each DataFrame, the only column to not format correctly was the column containing the date.

5.1.2 Data Formatting

Normalising the DataFrames

The columns in the DataFrames that would not be necessary for either the data visualisation or the regression analysis had to be removed, in order to preserve the completeness of the data; this is called normalisation. The column that had to be removed from each DataFrame was the `areaType` column that appeared in every dataset; this column was presumably an artefact from the data being derived from a larger, original dataset, and so it would only hold the value of 'region'. For this reason, the column was never used. Alongside the removal of the `areaType` column, it was essential to remove any rows that contained any null values, as they can cause discrepancies in analyses.

None of the datasets contained any rows that had null values, although the Reinfections dataset did contain rows that had redundant rows for the age column. These rows were subsequently removed from the DataFrame, reducing the number of rows from 205810 to 196455, a removal of 9355 rows.

Renaming Columns

It was crucial that the interpretation of the data in each DataFrame was correct, so that the artefact could work with them properly. To do this, the columns in each DataFrame were renamed to improve clarity and consistency; many datasets had the column name 'value', which can be ambiguous. To rectify this, column

names were changed to be more specific to the data it represents, such as changing 'value' to 'Deaths' in datasets about deaths.

Datatype Conversion

To convert the date column from a standard categorical column to a `datetime` column, a built-in function with pandas DataFrames was used to apply it to the entire column for each DataFrame.

Syntax Correction

The datasets that contained columns for age formatted the age such that spaces were represented by underscores, and single digit numbers were not zero-padded. Zero-padding is the addition of a zero to a numerical value to satisfy a particular format, such as formatting the number 5 as '05' to satisfy a two-digit format.

To implement zero-padding into the DataFrames, a function was written, see Listing 5.1, that could detect the format of the input string and the value of the number, and apply zero-padding if necessary. A second implementation was then applied to replace underscores in the string with a space character.

```
def zeroPrefix(x):
    # If the string contains 'to' a delimiter
    if "_to_" in x:
        # Split the string into a start and an end range
        x_range = x.split("_to_")

        # Convert the start and the end ranges to integers
        x_start = int(x_range[0])
        x_end = int(x_range[1])

        # If the start or end range is less than 10, add a leading zero
        x_start = "0" + str(x_start) if x_start < 10 else str(x_start)
        x_end = "0" + str(x_end) if x_end < 10 else str(x_end)
        return x_start + "_to_" + x_end

    # If the string contains '+' as a delimiter
    elif "+" in x:
        # Split the string into an integer and a '+'
        x = int(x.split("+")[0])

        # If the integer is less than 10, add a leading zero
        x = "0" + str(x) if x < 10 else str(x)
        return x + "+"

    # If the string is just a single integer
    else:
        # Convert the integer to an integer
        x = int(x)

        # If the integer is less than 10, add a leading zero
        x = "0" + str(x) if x < 10 else str(x)
        return x
```

Listing 5.1: The function of zero-padding.

5.2 Data Visualisation

5.2.1 Graphic Selection

For all of the available DataFrames, the graphic models that were selected were:

Region Displaying the distributions and proportions of each dataset, according to each region.

Age Displaying the distributions and proportions of each dataset, according to each age group.

Variant Displaying the distribution and proportions of each dataset, according to each variant. Note that this is only applied to the Variants dataset, which does not display graphics according to region or age group.

As region, age, and variant are the only categorical variables that appear in the DataFrames, they were natural choices.

5.2.2 Data Aggregation

Data aggregation is the grouping of multiple data points, based upon a shared characteristic. In order to produce visualisations of the data according to the chosen categories, the artefact would have to employ aggregation. For certain DataFrames, each region could be subdivided by the age column, and vice versa. Where necessary, a DataFrame was grouped by the region and date, and reduce[1] the values of their respective numerical column, for instance, the Cases column. The indices were then reset, creating a new sequential order for the DataFrame, to allow further operations on the DataFrame.

5.2.3 Data Enhancement

Data enhancement is the addition of new information or features to an existing dataset, in order to improve the quality or usefulness. This was applied to each DataFrame, by calculating the difference in the numerical column, between each date interval, comparing the current value to the prior value, and adding it into a new column to store the increase; the DataFrames were already sorted in date order. Any rows in this new column that had a null value, due to being the latest date entry, were replaced with a value of 0. Using this method is what allowed me to plot the additional visualisations for the discrete time-series plots.

5.2.4 Time-Series Plots

Cumulative and Discrete Time-Series Plots

Cumulative time-series plots are used in the artefact to visualise the progression of the COVID-19 pandemic, by region, age, or variant, over a continuous period of time. A cumulative formatting allows a reader to gain an insight of the total impact of the pandemic, and how effective measures have been in mitigating the

[1]Reduction is the process of reducing the amount of data in a dataset while maintaining the relevant information; it is also known as a fold operation ("Fold in Functional Programming", 2016).

impact, as seen by the gradient of the curve. An example of the cumulative time-series plots can be seen in Figure 5.1.

On the other hand, discrete time-series plots are effective for identifying trends and patterns at specific points in time, for instance, periods of time that exhibit in transmission cases will be clearly displayed, see Figure 5.2. Correlations between national or regional events and spikes on discrete time-series graphs can help determine future mitigation strategies.

Multi-Line and Single-Line Time-Series Plots

Multi-line time-series plots, see Figure 5.1, are used to allow the reader to compare the trends between the variables, to display how certain policies may be effective in some areas but not in others; it is often that policies have to be tailored to fit specific regions.

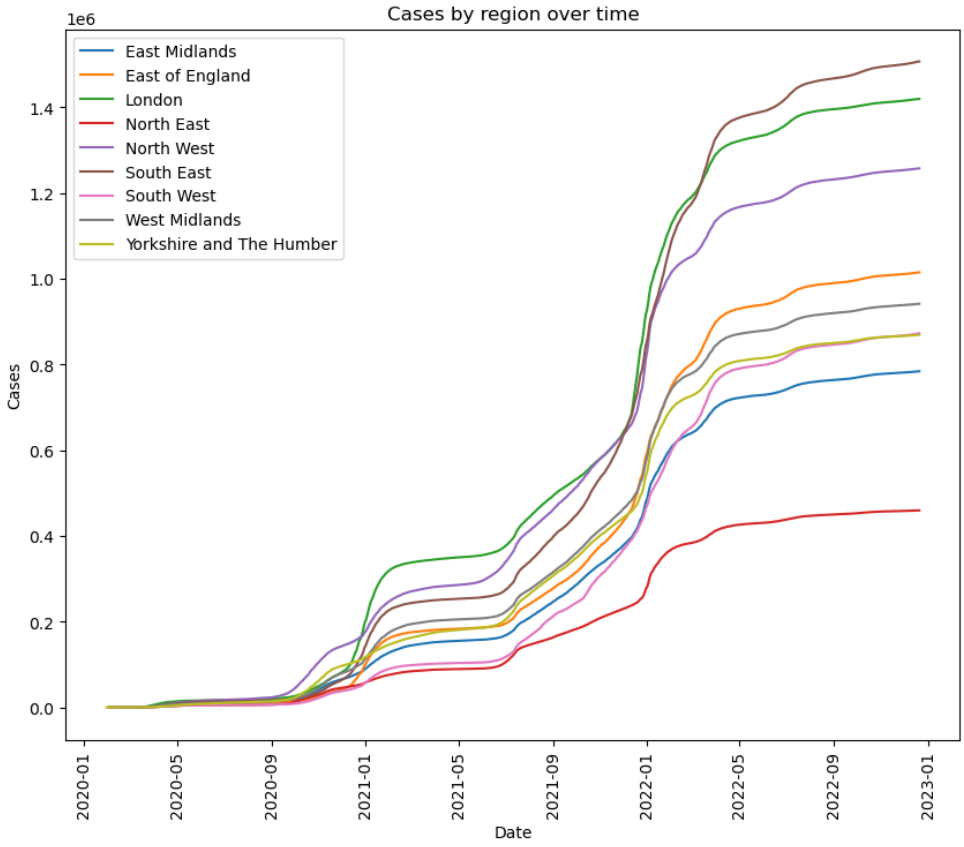

Figure 5.1: A multi-line cumulative time-series plot.

Single-line time-series plots, see Figure 5.2, were an addition to allow the reader to gain a clear insight for particular areas, as multi-line graphs can become crowded as the number of variables increases.

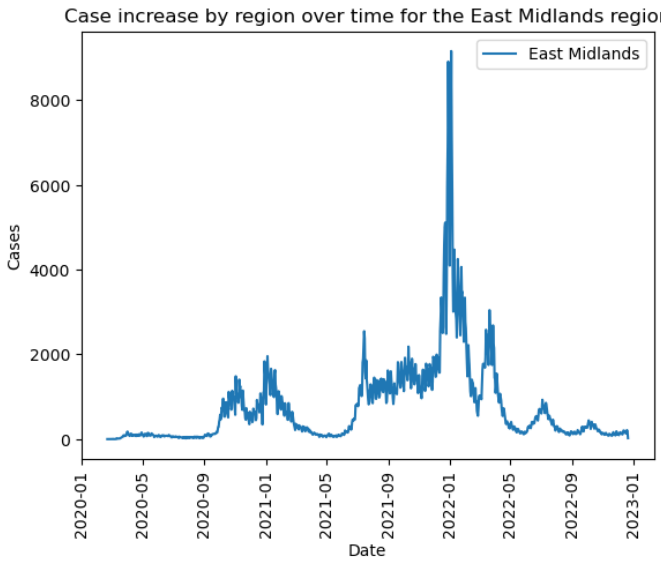

Figure 5.2: A single-line discrete time-series plot.

Implementation of the Time-Series Plots

To plot the multi-line time-series plots, the artefact looped through each categorical variable, region, age group, or variant, and plotted the numerical value along the Y-axis and the date along the X-axis. An example can be seen in Listing 5.2.

The implementation of the single-line time-series plots followed the same loop and plot logic, except for the fact that the functions responsible for the configuration and display of the plots were contained within the loop, unlike the multi-line plots.

```
# Create a figure with size (10, 8)
plt.figure(figsize=(10,8))

# Loop through each region
for maleCasesRegionName in maleCasesRegion["Region"].unique():
    # Filter for the current region
    maleCasesRegionData = maleCasesRegion[maleCasesRegion["Region"] ==
        maleCasesRegionName]

    # Plot the cases against date for the current region
    plt.plot(maleCasesRegionData["Date"], maleCasesRegionData["Cases"],
        label=maleCasesRegionName)

# Configure the layout of the plot
plt.xlabel("Date")
plt.ylabel("Cases")
plt.legend()
plt.xticks(rotation=90)
plt.title("Cases by region over time")

# Show the plot
plt.show()
```

Listing 5.2: An implementation of a multi-line cumulative time-series plot.

Correcting the Vaccinations Dataset

When originally plotting the cumulative time-series plots for the Vaccinations dataset, the plot that was produced displayed oscillations, each approximately separated by one week on the date axis. Naturally, a cumulative sum should not ever display a decrease in the subsequent values. This error was potentially being caused due to how the data was being collected; according to the API documentation for this dataset, 'it is possible that the number of people vaccinated in surveillance figures may reduce over time, due to people dying or moving out of a resident population' (Government, 2023).

To rectify this issue, a reduction algorithm was used to create new values for the three cumulative value columns, replacing the original, incorrect values in place.

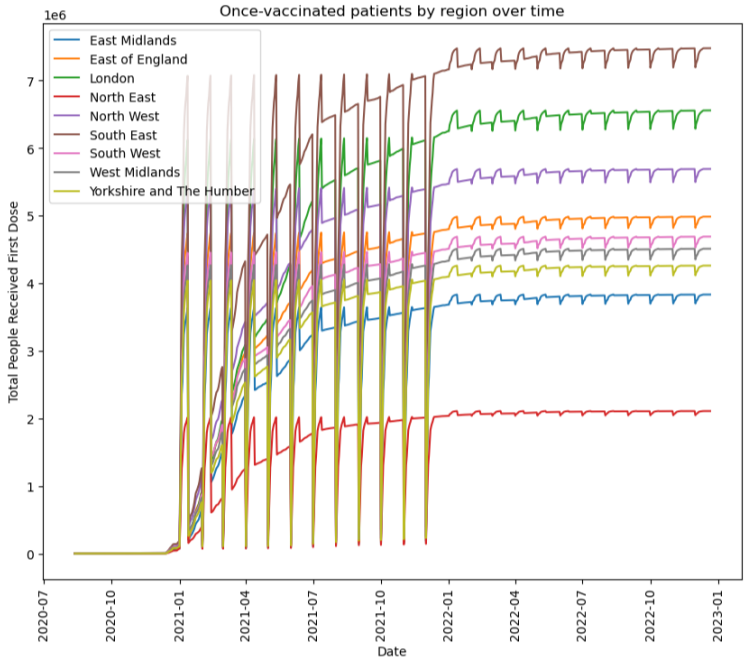

Figure 5.3: A flawed multi-line cumulative time-series plot of once-vaccinated patients by region over time.

5.2.5 Proportion Graphics

Peak Identification

Plotting pie charts and treemaps for each date in the data would be a fruitless and inefficient endeavour. Rather than doing that, proportion graphics were plotted, for each respective categorical variable, for each dataset, on the five highest peak dates.

The algorithm that was used to calculate the five highest peaks of the subdivided data can be seen in Listing 5.3. The algorithm smooths the data using a seven day rolling mean of the increase column, reducing the noise of the fluctuating values and thus highlighting the underlying trend. After this, the algorithm groups the newly-smoothed data by the month of the entry and finds the highest increase value within each month. Lastly, the values are sorted in descending order and the five highest rows are taken to be plotted for the proportion graphics.

When the proportion graphs were originally being plotted, the dates that it

selected for the peaks were within days of each other. This hindered the artefact's capabilities, as the data that was displayed showed minimal to no difference. As opposed to showing multiple days within one peak, the algorithm allows me to display five separate peaks.

```
# Calculate the rolling mean of case increase and add to a new column
maleCasesRegion["Rolling Mean"] = maleCasesRegion["Cases Increase"].
    rolling(7).mean()

# Group the data by month and find the peak number of case increase for
     each month
maxCasesPerMonth = maleCasesRegion.groupby(maleCasesRegion["Date"].dt.
    strftime("%Y-%m"))["Cases Increase"].max()

# Sort the values in descending order
maxCasesPerMonth = maxCasesPerMonth.sort_values(ascending=False)

# Get the five highest months
top5Months = maxCasesPerMonth.head(5).index.tolist()
```

Listing 5.3: An implementation of the peak calculation.

Pie Charts and Treemaps

Both the pie charts and treemaps were created using a similar process; for each of the top five months, the data for that month was isolated, to calculate the percentage that each category represents of the total increase.

For the pie charts, the percentage for each category is plotted as a slice of the pie, with the category value as the label. The treemaps uses this percentage to calculate the size of the rectangle in the treemap. Examples of the pie charts and treemaps can be seen in Figures 5.4 and 5.5 respectively.

For the treemaps, the process is mostly similar to that of the pie charts, and can be seen in Listing 5.4.

```
# Loop through each month
for month in top5Months:
    # Create a copy of the data for that month
    month2 = maleCasesRegion[maleCasesRegion["Date"].dt.strftime("%Y-%m"
        ) == month].copy()

    # Calculate the percentage increase in cases and add to a new column
    month2.loc[:, "Percentage Increase"] = month2["Cases Increase"] /
        month2["Cases Increase"].sum()

    # Find the peak date and add to a new column
    month2.loc[:, "Peak"] = (month2["Rolling Mean"].diff(1) > 0) & (
        month2["Rolling Mean"].diff(-1) > 0)

    # Get the peak date
    peakDate = month2[month2["Peak"]].sort_values("Cases Increase",
        ascending=False)["Date"].iloc[0]

    # Calculate the sizes of the rectangles as a percentage of the total
```

```
    volume
sizes = month2[month2["Date"] == peakDate]["Cases Increase"].values
total_size = sizes.sum()
sizes_percent = sizes / total_size * 100

# Create a dictionary of sizes and regions for the peak date
values_dict = dict(zip(month2[month2["Date"] == peakDate]["Region"],
    sizes_percent))

# Create the label for each region with the percentage value
labels = [f"{k}: {v:.1f}%" for k, v in values_dict.items()]

# Create a tree map with the dictionary values and labels
sq.plot(sizes=sizes_percent, label=labels, alpha=.8, text_kwargs={'
    rotation': 15})

# Configure the layout of the plot
plt.title(str(peakDate))

# Show the plot
plt.show()
```

Listing 5.4: An implementation of treemaps.

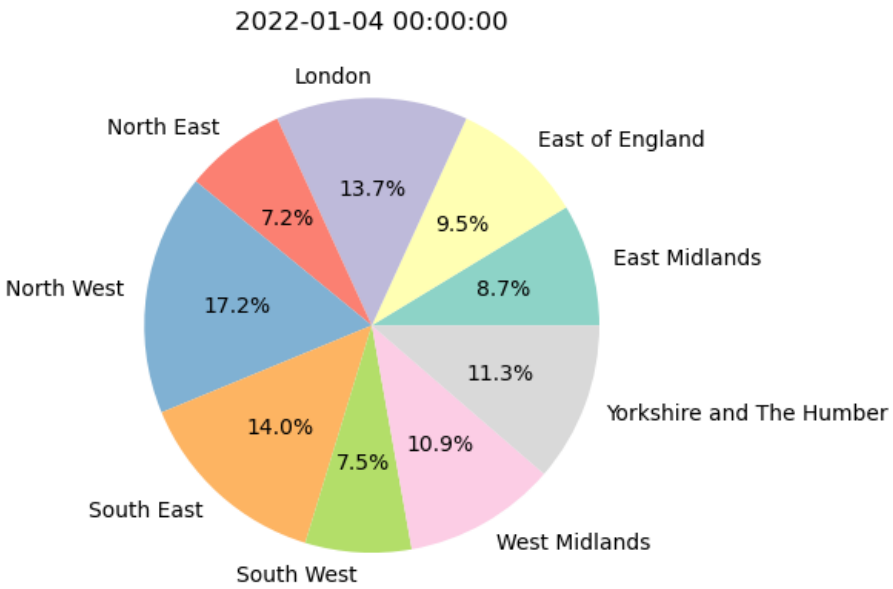

Figure 5.4: A pie chart.

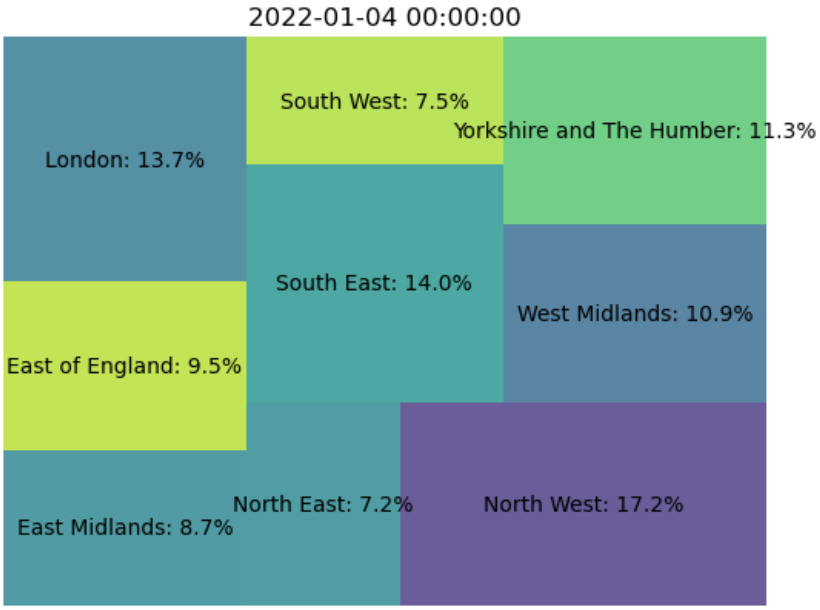

Figure 5.5: A treemap.

5.3 Regression Model Analysis

5.3.1 Model Selection

Three categories of regression models were built for the artefact, with multiple models for each. The full results for every model can be found in the documentation, as seen in Appendix C.

The Impact of Sex on COVID-19 Cases

A series of models aiming to examine the impact of sex, region, age, and date.

The Impact of Variants on COVID-19 Cases and Deaths

A series of models aiming to examine the impact of the variant, region, sequences[2] of the variant, and date.

The Impact of Vaccinations on COVID-19 Cases and Deaths

A series of models aiming to examine the impact of region, the number of people vaccinated, and date. The number of people vaccinated is separated into the total number of people to receive one, two, or three doses.

5.3.2 Data Aggregation

To build the DataFrame for *The Impact of Sex on COVID-19 Cases* models, there had to be a combination of the Male Cases and Female Cases DataFrames. Since both of these DataFrames already had the same structure, a sex column was added to each and put 'Male' and 'Female' as the value for the respective

[2]A sequence is an incidence of identification for each variant.

DataFrames, which were then vertically concatenated together and the index was reset.

For *The Impact of Variants on COVID-19 Cases and Deaths* and *The Impact of Vaccinations on COVID-19 Cases and Deaths* models, the necessary DataFrames did not have the same structures and so they could not be concatenated to one another. Instead they had to be merged according to their shared characteristics; for both models, this was the 'Region Code', 'Region', and 'Date' columns. After merging, all of the columns that could be used for the subsequent modelling were selected. An example of the aggregation can be seen in Listing 5.5.

```
# Join variants1 and generalCases1 on the common columns
variantImpact = pd.merge(variants1, newGeneralCases, on=['Region', '
    Date'])

# Join the result above and generalDeaths1 on the common columns
variantImpact = pd.merge(variantImpact, generalDeaths1, on=['Region
    Code', 'Region', 'Date'])

# Select the columns you need
variantImpact = variantImpact[['Region Code', 'Region', 'Date', '
    Variant', 'Sequenced', 'Percentage', 'Cases', 'Deaths']]
```

Listing 5.5: An implementation of data aggregation for the regression models.

5.3.3 Model Creation

To make the input arrays for the models, a subset of the newly made DataFrames with the required columns is copied to a separate DataFrame. This DataFrame is then separated into the independent and dependent variables, X and Y respectively, and stored in arrays.

To make the X array, one-hot encoding is applied to the categorical variables, which converts each category into a binary column to indicate presence or absence. This method does not imply to a model any ranking or ordering, making it the best choice to use. The new encoded columns are concatenated back into the array before usage. This can be seen in Listing 5.6. The encoder used to do this is from the Preprocessing and Normalisation module of scikit-learn (Pedregosa et al., 2011).

```
X = variantImpact1[["Variant", "Region", "Sequenced", "Days"]].values
y = variantImpact1[["Cases", "Deaths"]].values

catCols = [0, 1]

encodeX = encoder.fit_transform(X[:, catCols])
X = np.concatenate([encodeX, X[:, 2:]], axis = 1)
```

Listing 5.6: An implementation of input data preparation for the regression models.

Next, the X and Y variables are split into the train and test splits, of a 4:1 ratio; the 4:1 ratio was chosen as it is the most commonly chosen ratio, however it is possible that the ratio could have a hindrance on the produced results. The

seed used for the randomisation was set as 42, and was used in every model made, to ensure that all models are able to be reproduced.

Correcting for the Date Variable

In the original model implementations, for models *COVID-RMA001* to *COVID-RMA008*, the date was directly passed onto the model. This resulted in poor results; *COVID-RMA004* and *COVID-RMA008* had a coefficient of determination value of 0, meaning that they could not explain any variation of the data. This occurred because the date itself from each entry does not convey meaningful information to the models.

To fix this issue, a new column was added to the DataFrames that contained the number of days passed since the earliest date in the data, as seen in Listing 5.7. After implementing this column and passing it to the model, the coefficient of determination of the subsequent models improved significantly.

```
firstDate = variantImpact["Date"].min()

variantImpact["Days"] = (variantImpact["Date"] - firstDate).dt.days
```

Listing 5.7: An implementation of the date correction for the regression model data.

5.3.4 Model Evaluation

Beta Coefficients

When linear-based models are used in scikit-learn, the beta coefficients from the model can be collected using `.coef_`. The raw output from this command is returned as an array of values, see Figure 5.6, which can be difficult to interpret as it doesn't clearly display which variable each beta coefficient relates to.

```
Coefficients:  [[-3.38916530e+04  1.24646085e+04 -7.20606799e+03  5.34851717e+02
   1.84241780e+04  2.16433163e+04 -5.44706763e+03  5.50481963e+03
   1.96126945e+03 -1.39902550e+04 -4.59775986e+04 -5.31937511e+03
   9.57325667e+04 -1.15563989e+05  5.64518387e+04  9.07332264e+04
  -4.32698841e+04 -9.31767099e+03 -2.34661141e+04  7.27008077e-02
   2.94463135e+03]
 [-2.98005068e+02  1.07329252e+02 -6.23818324e+01 -3.53500580e+00
   1.75036070e+02  1.77302379e+02 -3.87928259e+01  3.71300864e+01
   2.57239675e+01 -1.19807022e+02 -5.02981747e+02  9.30282044e+01
   7.80043458e+02 -1.46372321e+03  1.15548262e+03  1.15838383e+03
  -1.18821900e+03  2.43955326e+02 -2.76969480e+02  6.40153007e-04
   3.19844846e+00]]
```

Figure 5.6: A raw output of the `.coef_` array.

To display the beta coefficients in a readable fashion, it was decided to put them into a DataFrame, which displayed the column the variable originated from, the variable itself, and it's corresponding beta coefficients. To implement this, the beta coefficients are extracted into an array and check the shape of the array. If there is only one column, the array is reshaped to a 2-dimensional array. The names of the variables are collected and appended to the output DataFrame in a for loop. This implementation can be seen in Listing 5.8, and the output in Figure 5.7.

```
# Get the coefficients of the model
coefficients = model2.coef_

# If coefficients has only one column, reshape it to a 2D array
if coefficients.ndim == 1:
    coefficients = coefficients.reshape(1, -1)

# Get the names of the encoded categorical variables
encodedCatNames = encoder.get_feature_names_out(["Variant", "Region"])

# Concatenate the names of the encoded categorical variables with the
    numerical column names
columnNames = np.concatenate([encodedCatNames, ["Sequenced", "Days"]])

# Create a dataframe with the variable names and their coefficients
variantImpactCoefficient = pd.DataFrame({'Variable': columnNames})

for i, coefficient in enumerate(coefficients):
    coefficientName = "Coefficient " + str(i + 1)
    variantImpactCoefficient[coefficientName] = coefficient

# Replace underscores with colons in the variable names
variantImpactCoefficient["Variable"] = variantImpactCoefficient["
    Variable"].str.replace("_", ": ")
```

Listing 5.8: An implementation of the coefficient table.

	Variable	Coefficient 1	Coefficient 2
0	Variant: Other	-33891.652962	-298.005068
1	Variant: V-20DEC-01 (Alpha)	12464.608463	107.329252
2	Variant: V-21APR-02 (Delta B.1.617.2)	-7206.067988	-62.381832
3	Variant: V-21OCT-01 (Delta AY 4.2)	534.851717	-3.535006
4	Variant: V-22JUL-01 (Omicron BA.2.75)	18424.177994	175.036070
5	Variant: V-22OCT-01 (Omicron BQ.1)	21643.316285	177.302379
6	Variant: VOC-21NOV-01 (Omicron BA.1)	-5447.067627	-38.792826
7	Variant: VOC-22APR-03 (Omicron BA.4)	5504.819627	37.130086
8	Variant: VOC-22APR-04 (Omicron BA.5)	1961.269447	25.723968
9	Variant: VOC-22JAN-01 (Omicron BA.2)	-13990.254955	-119.807022
10	Region: East Midlands	-45977.598590	-502.981747
11	Region: East of England	-5319.375112	93.028204
12	Region: London	95732.566740	780.043458
13	Region: North East	-115563.988960	-1463.723213
14	Region: North West	56451.838707	1155.482622
15	Region: South East	90733.226371	1158.383834
16	Region: South West	-43269.884097	-1188.219003
17	Region: West Midlands	-9317.670989	243.955326
18	Region: Yorkshire and The Humber	-23466.114070	-276.969480
19	Sequenced	0.072701	0.000640
20	Days	2944.631355	3.198448

Figure 5.7: A clean output of the `.coef_` array.

The performance metrics are returned using scikit-learn functions, for example, `.score` and `mean_squared_error`. Additionally, the option to plot residual graphs, to display the line of best fit of the predicted values against the actual values, was implemented to be used for the best performing models.

5.3.5 Cross-Model Evaluation

After each model was trained and tested, the captured performance metrics were added to a `csv` file along with the analysis series, the model name, and the type of regression model. This `csv` file is read into a DataFrame and split into the three analysis series. From there, each model is assigned a rank based upon the coefficient of determination, mean squared error, and mean average error, which is then assigned to a rank column. Lastly, the DataFrames are sorted by the value in the rank column in descending order to determine the best performing models. The implementation can be seen in Listing 5.9.

```
# Select the rows applicable to this analysis series
sexImpactPrototypes = modelPrototypes.loc[modelPrototypes["Analysis
    Series"] == "Sex Impact"]

# Calculate the rank of each row based on the selected columns
sexImpactPrototypes.loc[:, "Rank Value"] = 0
sexImpactPrototypes.loc[:, "Rank Value"] += sexImpactPrototypes.loc[:,
    "R2"].rank(ascending = False)
sexImpactPrototypes.loc[:, "Rank Value"] += sexImpactPrototypes.loc[:,
    "MSE"].rank()
sexImpactPrototypes.loc[:, "Rank Value"] += sexImpactPrototypes.loc[:,
    "MAE"].rank()

# Sort the DataFrame by the rank values
sexImpactPrototypes = sexImpactPrototypes.sort_values(by="Rank Value",
    ascending=True)

# Add a new column with proper ranks that handle joint places
sexImpactPrototypes["Rank"] = sexImpactPrototypes["Rank Value"].rank(
    method="min", ascending=True)

# Drop the "Analysis Series" column from the output
sexImpactPrototypes = sexImpactPrototypes.drop(columns=["Analysis
    Series", "Variables"])

# Output DataFrame
sexImpactPrototypes
```

Listing 5.9: An implementation of the cross-model evaluation table.

The reasoning behind this is that the values for the mean squared error and mean average error can be so large that manually selecting the best would be a futile effort.

5.4 Miscellaneous Issues

5.4.1 Male and Female Deaths Datasets

On the 2nd of February 2023, it became apparent that the Male and Female Deaths dataset would be unusable due to the fact that they only covered a span of the prior 28 days; there was simply no time-series that existed within the data. Even if there was, the small number of entries in the datasets, 171 to be exact, meant that if the data was used, the project would have been knowingly producing false results for the project and it would have been unethical.

5.4.2 General Cases Dataset

On the 24th of April 2023, an issue with the General Cases dataset was spotted via the data visualisation, where there was a series of peaks in the cases that did not align with the information displayed by the Male and Female Cases datasets, most notably on the 31st of January; when compared to non-governmental data, this peak appears to be erroneous (Organisation, 2023). Due to the fact that the data visualisation was the key in identifying the issue, the dataset is still being included in the final evaluation, although it is in a reduced manner.

To allow for more truthful regression modelling, a composite DataFrame was created by a combination of the Male and Female Cases datasets and used for the modelling. Note that the documentation written for the models that used the General Cases dataset can still be found in Appendix C. It should also be noted that despite a change in the datasets being used, there was absolutely no difference in the performance of any of the models in either series, *The Impact of Variants on COVID-19 Cases and Deaths* or *The Impact of Vaccines on COVID-19 Cases and Deaths*.

5.4.3 Incorrect Encoding of the Date

The regression models that were produced between models *COVID-RMA001* and *COVID-RMA017* were programmed such that the date or days variables were being one-hot encoded along with the categorical variables; this was an oversight during production and was not an intentional feature of the models. After rectifying this mistake, the models which used days passed, *COVID-RMA009* to *COVID-RMA017*, all saw a perceived reduction in performance when remade with the same input data. Despite the lesser performance, all models that suffered this error were removed from the final evaluation, as they produced false results and to include them would be unethical for the project. As noted before, these models can still be found in Appendix C.

5.4.4 Support Vector Regression

The complexity of the data meant that the expected run-time of the model was unfeasible for proper usage. A single model was made using this algorithm, and the documentation can be seen in Appendix C, but is not included in the final evaluation.

Chapter 6

Results and Discussion

6.1 Data Visualisation

6.1.1 Findings on Potential Transmission Vectors

The data visualisation produced by the artefact proved to be successful in uncovering underlying patterns related to the spread of COVID-19, specifically yielding findings on the dynamics of the transmission of COVID-19, as well as the impact of particular factors.

The peak dates for transmission cases that were identified across all of the proportion graphics, as seen in Figure 6.1, were:

- 27th of December 2020
- 4th of January 2021
- 19th of July 2021
- 27th of September 2021
- 18th of October 2021
- 29th of December 2021
- 4th of January 2022
- 22nd of March 2022

All of these dates, except for the 18th of October 2021 and the 22nd of March 2022, align with common school holiday periods and events, for instance, the Christmas break, New Years Day, the summer break, and the return to school. These peaks suggest that schoolchildren could be a major vector for the virus, although it does not suggest that transmission is directly linked to age.

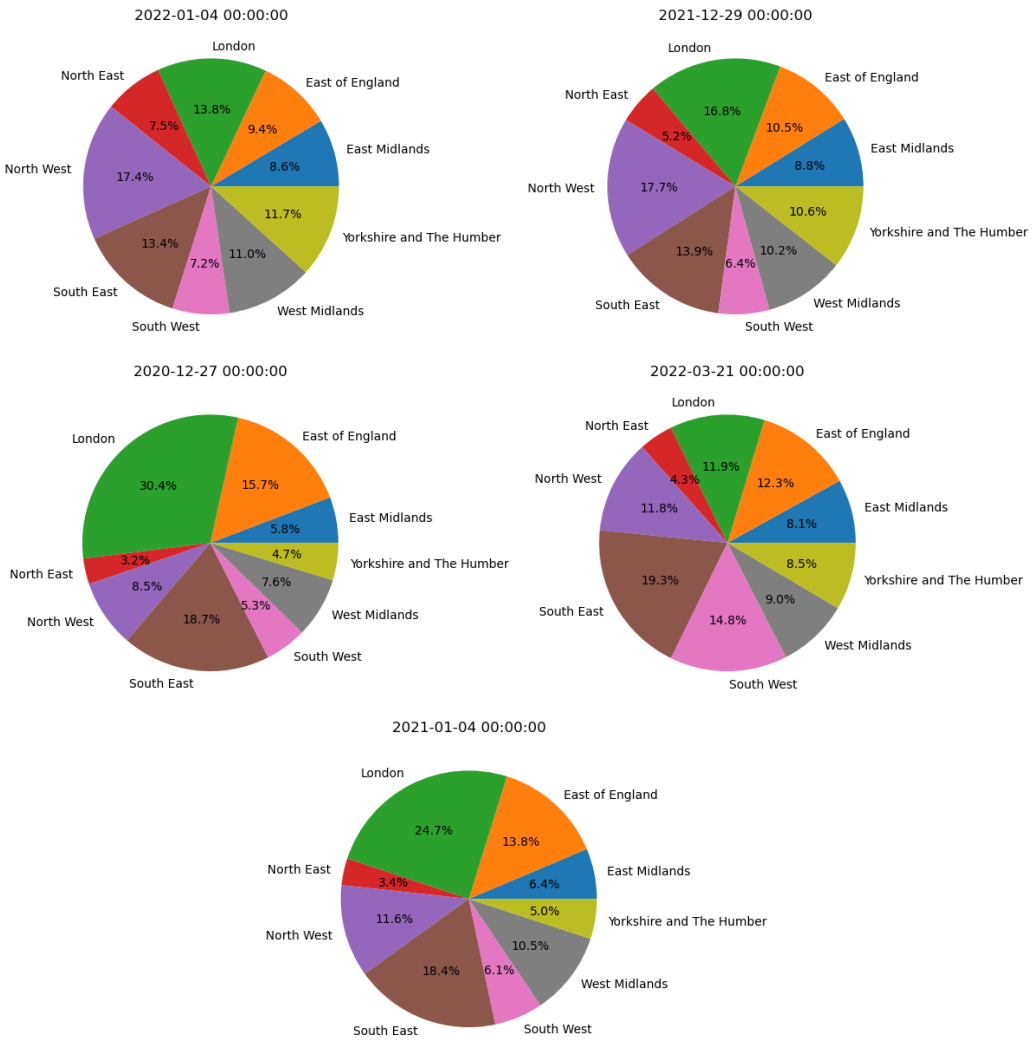

Figure 6.1: A selection of pie charts from the artefact.

Alongside the identified peaks in the proportion graphics, the cumulative time-series graphs of transmission cases exhibited more sigmoidal characteristics in the age groups of schoolchildren, compared to significantly older age groups who displayed more linear characteristics; this can be seen in Figures 6.2 and 6.3.

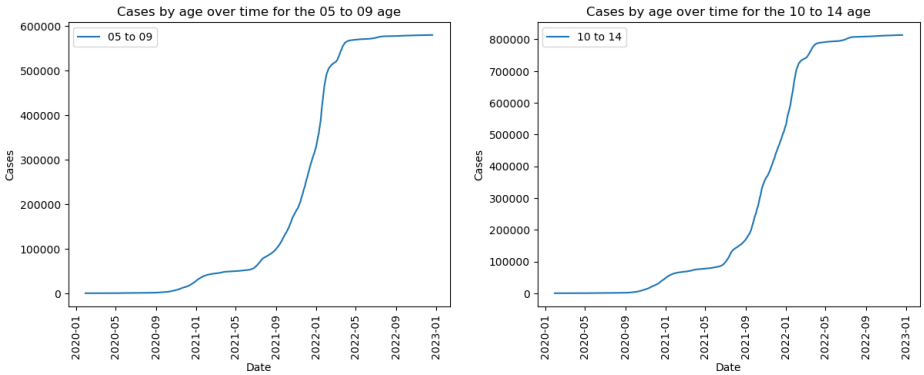

Figure 6.2: Sigmoidal characteristics displayed in younger age groups.

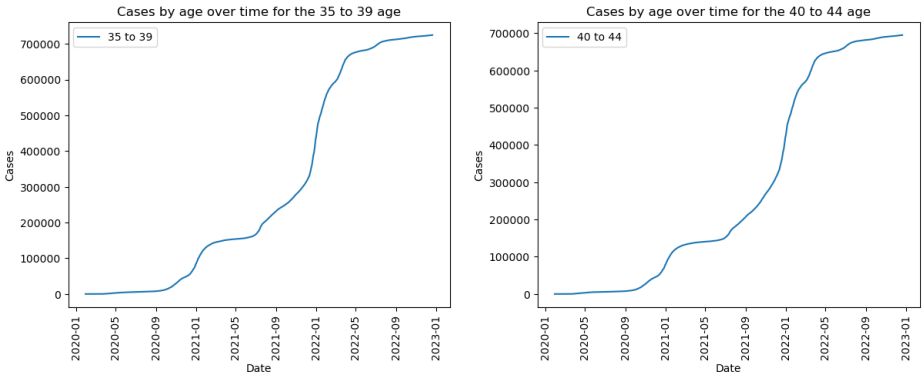

Figure 6.3: Linear characteristics displayed in older age groups.

The General Cases dataset displayed different characteristics to the Male Cases and Female Cases datasets when plotted, as seen in Figure 6.4; peak dates that were identified in this dataset were:

- 23rd of December 2021
- 31st of January 2022
- 28th of March 2022
- 19th of April 2022
- 6th of July 2022

The lack of correlation with the other graphs indicated that the dataset contained faulty data.

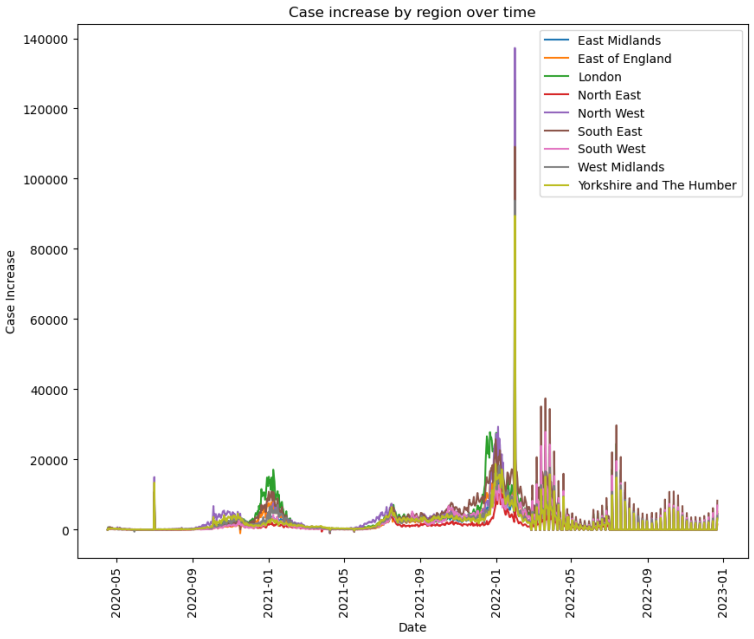

Figure 6.4: Erroneous output from the General Cases dataset.

The comparison of the graphs produced from the Male and Female Cases datasets, as seen in Figure 6.5, showed negligible differences, implying that sex has little to no impact on the transmission of COVID-19.

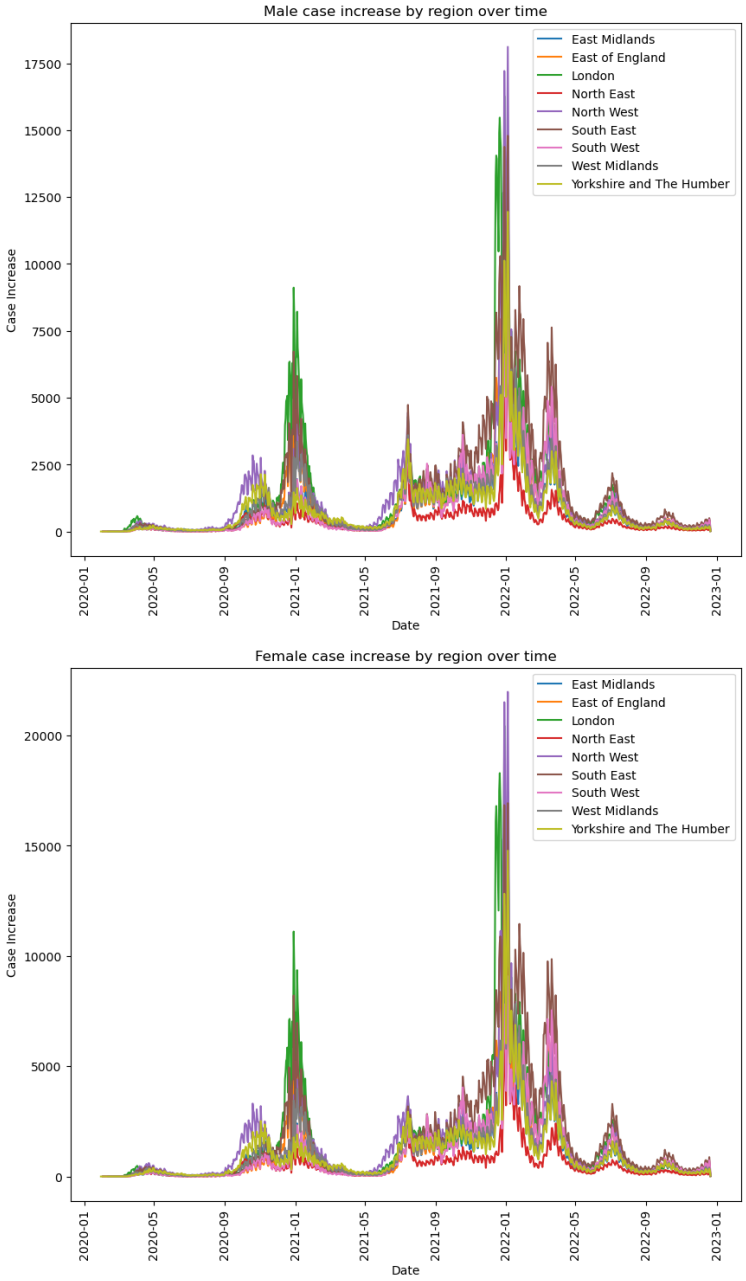

Figure 6.5: A comparison of male and female increase in cases by region over time.

When comparing the impact across age groups, it appears that senior age groups faced relatively higher impacts for transmission cases, implying that older people face a higher risk of catching the virus. This can be seen in Figure 6.6.

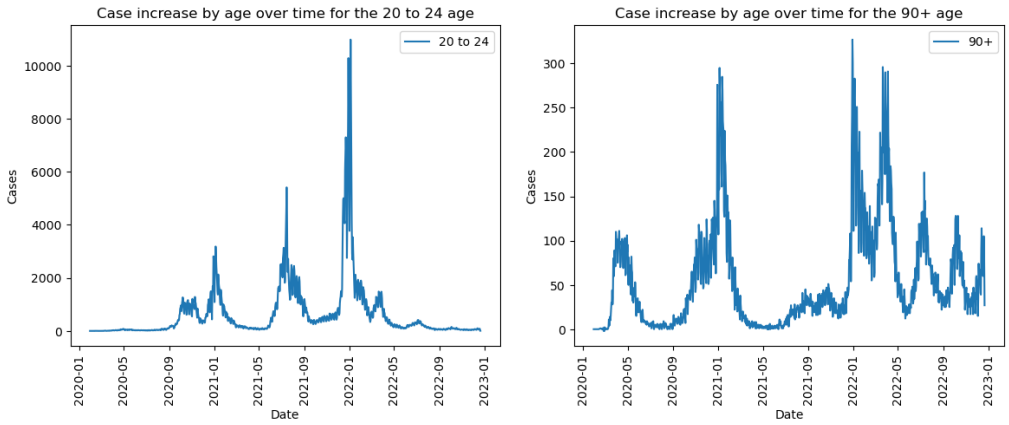

Figure 6.6: A comparison of the magnitude of cases across age groups.

The results for each region displayed minor differences in pattern, trending between a linear or sigmoidal pattern, but no observation could be made from them.

6.1.2 Findings on Variants

The 18th of October 2021 and the 22nd of March 2022 were the two peak dates for the transmission cases that did not align with school events and instead aligned with peaks in variant sequencing; the former aligned with the emergences of Omicron BA.1 and Omicron BA.2, and the latter with Omicron BA.4 and Omicron BA.5.

The reinfections saw significant rises during the onset of Omicron, notably witnessing peaks on the 4th of January 2022 and the 22nd of March 2022. Given the negligible levels of reinfections prior to New Years of 2022, it can be inferred that Omicron played a large factor in reported reinfections. This can be seen in Figure 6.7.

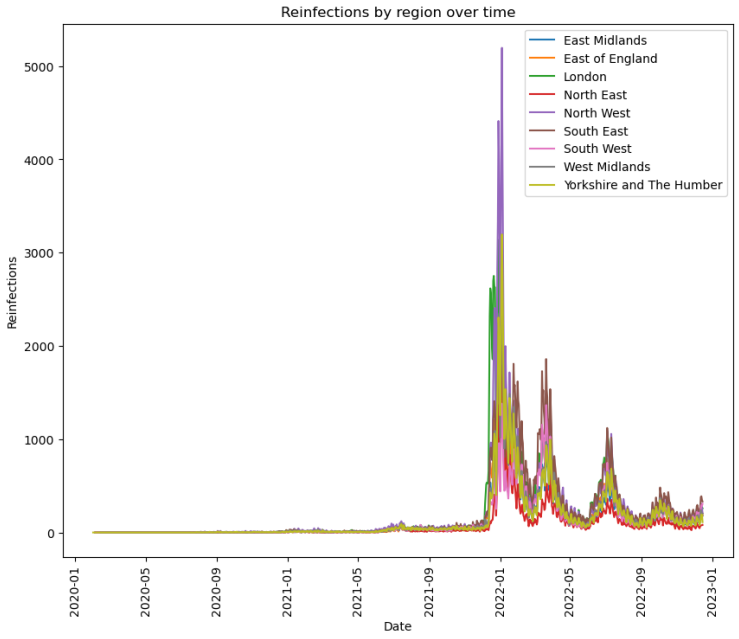

Figure 6.7: A plot of reinfections, likely influenced by the Omicron variant.

The peaks that were identified for the deaths did not strongly correlate with the peaks in the transmission cases. These peaks were:

- 21st of March 2020
- 9th of April 2020
- 31st of December 2020
- 20th of January 2021
- 1st of February 2021

The earliest date available in the Variants dataset was the 14th of February 2021, while the earliest date available in the General Deaths dataset was the 11th of March 2020. Due to this, a direct comparison of the peaks in deaths and variants is not possible. Despite this, if we were to extrapolate the existing information of graphs, it would suggest that there is a significant correlation between the earlier dates and the Alpha variant. On the other hand, the other variants, Delta and Omicron, do not appear to be as significant in impacting deaths.

6.2 Regression Analysis

6.2.1 The Impact of Sex on COVID-19 Cases

In this series of models, see Table 6.1, all of the models, except for the orthogonal matching pursuit, had an R^2 above 0.5, showing a reasonable understanding of the data, however the variance of the other metrics suggest that some of the models may be best suited to predicting specific outcomes. The lasso regression model performed the best, suggesting that the dataset contains redundant features.

Model Name	Type	R2	MSE	MAE	Rank Value	Rank
COVID-RMA048	Lasso Regression	0.68	325809013.1	13645.82	6.5	1.0
COVID-RMA073	Bayesian Ridge Regression	0.68	325810485.0	13647.62	8.5	2.0
COVID-RMA019	Ridge Regression	0.68	325810538.4	13647.74	10.5	3.0
COVID-RMA083	Theil Sen Regression	0.67	338318818.8	13308.00	12.0	4.0
COVID-RMA020	Linear Regression	0.68	325810547.9	13647.85	12.5	5.0
COVID-RMA078	Huber Regression	0.64	370881703.1	12713.05	13.0	6.0
COVID-RMA047	Elastic Net Regression	0.50	511139607.2	15293.80	21.0	7.0
COVID-RMA088	Orthogonal Matching Pursuit	0.49	521059149.2	15710.41	24.0	8.0

Table 6.1: Cross-model evaluation for *The Impact of Sex on COVID-19 Cases*.

From the coefficients in Table 6.2, the subsequent findings are:

- Females are more likely to have a higher number of cases compared to males.
- Living in either London, the North West or the South East is more likely to be associated with a higher number of cases, while living in either the North East.
- Age groups above 60 are less likely to have a high number of cases compared to younger age groups.

- There is a positive correlation between the days passed and the number of cases.

Variable	Coefficient: Cases
Sex: Female	3.773444e+03
Sex: Male	-1.893338e-10
Region: East Midlands	-4.100118e+03
Region: East of England	3.948921e+02
Region: London	1.188107e+04
Region: North East	-1.175962e+04
Region: North West	8.853731e+03
Region: South East	1.147036e+04
Region: South West	-3.809978e+03
Region: West Midlands	-0.000000e+00
Region: Yorkshire and The Humber	-1.248485e+03
Age: 00 to 04	-1.457593e+04
Age: 05 to 09	-2.663467e+02
Age: 10 to 14	1.173843e+04
Age: 15 to 19	9.271373e+03
Age: 20 to 24	1.354463e+04
Age: 25 to 29	1.592084e+04
Age: 30 to 34	1.615347e+04
Age: 35 to 39	1.371517e+04
Age: 40 to 44	1.163838e+04
Age: 45 to 49	8.343782e+03
Age: 50 to 54	7.092231e+03
Age: 55 to 59	2.534611e+03
Age: 60 to 64	-5.111944e+03
Age: 65 to 69	-1.221220e+04
Age: 70 to 74	-1.447157e+04
Age: 75 to 79	-1.730995e+04
Age: 80 to 84	-1.904625e+04
Age: 85 to 89	-2.001698e+04
Age: 90+	-2.059648e+04
Days	7.195987e+01

Table 6.2: Coefficients for the *The Impact of Sex on COVID-19 Cases*, from COVID-RMA048.

6.2.2 The Impact of Variants on COVID-19 Cases and Deaths

Among these models, see Table 6.3, the models that are purely predicting deaths perform much better, suggesting that cases are harder to predict than cases. The repeated preferences for ridge regression suggests there were not any redundant features. Moreover, ridge, lasso, linear, and Theil-Sen all had near-perfect R^2 values and superb error metrics when predicting for deaths.

Model Name	Type	R2	MSE	MAE	Variables	Rank Value	Rank
COVID-RMA057	Ridge Regression	0.99	2.636936e+05	399.20	Deaths	4.5	1.0
COVID-RMA059	Linear Regression	0.99	2.640566e+05	399.30	Deaths	7.5	2.0
COVID-RMA058	Lasso Regression	0.99	2.638117e+05	399.36	Deaths	7.5	2.0
COVID-RMA085	Theil Sen Regression	0.99	3.982558e+05	479.46	Deaths	10.5	4.0
COVID-RMA049	Ridge Regression	0.94	4.028838e+10	115580.51	Cases and Deaths	24.0	5.0
COVID-RMA050	Lasso Regression	0.94	4.030008e+10	115607.64	Cases and Deaths	26.0	6.0
COVID-RMA091	Orthogonal Matching Pursuit	0.52	1.554359e+07	3153.84	Deaths	27.0	7.0
COVID-RMA051	Linear Regression	0.94	4.030026e+10	115608.07	Cases and Deaths	28.0	8.0
COVID-RMA060	Elastic Net Regression	0.46	1.749133e+07	3472.73	Deaths	31.0	9.0
COVID-RMA075	Bayesian Ridge Regression	0.22	2.552737e+07	4199.26	Deaths	34.0	10.0
COVID-RMA080	Huber Regression	-1.26	7.344673e+07	7055.83	Deaths	37.0	11.0
COVID-RMA053	Ridge Regression	0.89	8.057650e+10	230761.82	Cases	37.5	12.0
COVID-RMA054	Lasso Regression	0.89	8.059989e+10	230815.93	Cases	39.5	13.0
COVID-RMA084	Theil Sen Regression	0.89	8.109005e+10	227322.56	Cases	39.5	13.0
COVID-RMA052	Elastic Net Regression	0.62	8.051237e+10	147168.31	Cases and Deaths	40.0	15.0
COVID-RMA055	Linear Regression	0.89	8.060025e+10	230816.83	Cases	41.5	16.0
COVID-RMA089	Orthogonal Matching Pursuit	0.64	9.078060e+10	165864.17	Cases and Deaths	45.0	17.0
COVID-RMA056	Elastic Net Regression	0.79	1.610000e+11	290863.88	Cases	48.0	18.0
COVID-RMA090	Orthogonal Matching Pursuit	0.76	1.820000e+11	328574.50	Cases	51.0	19.0
COVID-RMA079	Huber Regression	0.71	2.150000e+11	343032.82	Cases	54.0	20.0
COVID-RMA074	Bayesian Ridge Regression	0.48	3.900000e+11	496982.70	Cases	60.0	21.0

Table 6.3: Cross-model evaluation of *The Impact of Variants on COVID-19 Cases and Deaths*.

From the coefficients in Table 6.4, the subsequent findings are:

- The Delta, Omicron BA.1, and Omicron BA.2 variants show the lowest correlation with cases and deaths.
- The number of times a variant has been sequenced has no correlation with cases or deaths, as to be expected.
- Correlations are much stronger with cases than deaths; this could be due to the lesser performance of the models on the cases variable.

Variable	Coefficient: Cases	Coefficient: Deaths
Variant: Other	-104721.175668	-241.912642
Variant: V-20DEC-01 (Alpha)	38287.025360	96.250182
Variant: V-21APR-02 (Delta B.1.617.2)	-23140.008148	-65.977353
Variant: V-21OCT-01 (Delta AY 4.2)	5744.891966	9.683242
Variant: V-22JUL-01 (Omicron BA.2.75)	55387.040216	138.237112
Variant: V-22OCT-01 (Omicron BQ.1)	71252.535371	173.977558
Variant: VOC-21NOV-01 (Omicron BA.1)	-18786.844854	-41.199808
Variant: VOC-22APR-03 (Omicron BA.4)	18067.080851	31.097133
Variant: VOC-22APR-04 (Omicron BA.5)	2513.357018	9.358690
Variant: VOC-22JAN-01 (Omicron BA.2)	-44603.902116	-109.514114
Region: East Midlands	-275942.072393	-3111.398230
Region: East of England	-35498.295711	431.539401
Region: London	554363.935821	4734.060984
Region: North East	-674019.940874	-8630.772389
Region: North West	353545.246040	6919.419214
Region: South East	531411.946449	6820.252064
Region: South West	-254172.981208	-6775.091372
Region: West Midlands	-57112.580585	1170.148377
Region: Yorkshire and The Humber	-142575.257569	-1558.158049
Sequenced	0.038172	0.000093
Days	3294.027441	9.225898

Table 6.4: Coefficients for the *The Impact of Variants on COVID-19 Cases and Deaths*, from COVID-RMA049.

6.2.3 The Impact of Vaccinations on COVID-19 Cases and Deaths

Within these models, see Table 6.5, there appears to be a difference in the performance of the models; in the second model series, both cases and deaths performed best with ridge regression but in this series, cases and deaths perform better with ridge and linear respectively.

Model Name	Type	R2	MSE	MAE	Variables	Rank Value	Rank
COVID-RMA071	Linear Regression	0.96	1.497265e+06	749.52	Deaths	7.5	1.0
COVID-RMA077	Bayesian Ridge Regression	0.96	1.497263e+06	749.53	Deaths	7.5	1.0
COVID-RMA070	Lasso Regression	0.96	1.496989e+06	751.33	Deaths	7.5	1.0
COVID-RMA069	Ridge Regression	0.96	1.497240e+06	749.63	Deaths	7.5	1.0
COVID-RMA087	Theil Sen Regression	0.93	2.325238e+06	1159.33	Deaths	16.5	5.0
COVID-RMA061	Ridge Regression	0.93	3.697456e+10	107367.59	Cases and Deaths	24.5	6.0
COVID-RMA063	Linear Regression	0.93	3.697481e+10	107368.84	Cases and Deaths	27.5	7.0
COVID-RMA062	Lasso Regression	0.93	3.697477e+10	107369.41	Cases and Deaths	27.5	7.0
COVID-RMA072	Elastic Net Regression	0.52	1.680340e+07	3363.21	Deaths	30.0	9.0
COVID-RMA094	Orthogonal Matching Pursuit	0.27	2.583256e+07	4149.24	Deaths	34.0	10.0
COVID-RMA065	Ridge Regression	0.90	7.394763e+10	213985.54	Cases	36.5	11.0
COVID-RMA082	Huber Regression	0.07	3.300246e+07	4664.53	Deaths	37.0	12.0
COVID-RMA076	Bayesian Ridge Regression	0.90	7.394777e+10	213986.31	Cases	38.5	13.0
COVID-RMA066	Lasso Regression	0.90	7.394803e+10	213987.49	Cases	40.5	14.0
COVID-RMA067	Linear Regression	0.90	7.394811e+10	213988.16	Cases	42.5	15.0
COVID-RMA064	Elastic Net Regression	0.66	7.761766e+10	149968.42	Cases and Deaths	45.0	16.0
COVID-RMA086	Theil Sen Regression	0.89	8.686361e+10	225822.16	Cases	48.0	17.0
COVID-RMA092	Orthogonal Matching Pursuit	0.48	1.180000e+11	183646.95	Cases and Deaths	50.0	18.0
COVID-RMA068	Elastic Net Regression	0.80	1.550000e+11	296573.64	Cases	52.0	19.0
COVID-RMA081	Huber Regression	0.72	2.180000e+11	332441.99	Cases	55.0	20.0
COVID-RMA093	Orthogonal Matching Pursuit	0.69	2.350000e+11	363144.66	Cases	58.0	21.0

Table 6.5: Cross-model evaluation of *The Impact of Vaccines on COVID-19 Cases and Deaths*.

From the coefficients in Table 6.6, the subsequent findings are:

- There is a negative correlation with having one or two doses of the vaccine, and cases; this was to be expected.
- There is a positive correlation with having one dose of the vaccine, but a negative correlation with having two doses of the vaccine. Despite this, the magnitude of the coefficients makes them almost negligible.
- The models found a positive correlation with having three doses for both cases and deaths; this could be caused by the population of people who have received three doses being insignificant when compared to the population of once-vaccinated patients, or even the national population.

Variable	Coefficient: Cases	Coefficient: Deaths
Region: East Midlands	-271543.335823	-2959.231445
Region: East of England	-48806.047141	300.433182
Region: London	565220.326294	4503.898652
Region: North East	-640232.611200	-8094.775733
Region: North West	370778.886247	6771.652792
Region: South East	483774.563660	6337.537852
Region: South West	-268450.584024	-6611.888202
Region: West Midlands	-60397.181010	1136.132304
Region: Yorkshire and The Humber	-130344.027454	-1383.759601
Total People Received First Dose	-0.016705	0.008278
Total People Received Second Dose	-0.955376	-0.007186
Total People Received Third Dose	2.028702	0.000777
Days	3103.685391	12.236531

Table 6.6: Coefficients for the *The Impact of Vaccines on COVID-19 Cases and Deaths*, from COVID-RMA061.

6.3 Remarks on the Approaches

On the whole, it can be determined that neither approach is sufficient without the other. Certain findings in the regression analysis, such as the positive correlations for thrice-vaccinated patients, lack the context that can be provided in the data visualisation. Conversely, the data visualisations cannot conclusively determine relationships like the regression analysis can, provided the data is truthful.

With that in mind, the outputs of the pie charts are inferior to those of the treemaps; the utility of both of these models was found to be mostly equal, however as the number of variables increased, the interpretation of the pie charts suffered due to the limited volume of a 360 degree circle. It should be noted that as the variables continue to increase in number, the same could eventually happen to a treemap, although at a much later stage.

Chapter 7

Conclusion

7.1 Suitability of the Aims and Objectives

Upon a final summary, the project was highly successful in achieving the aims and objectives that were laid out. A key achievement of the project was the investigation into underlying relationships using regression analysis; the regression models performed much better than expected, especially for modelling the relationships to deaths.

Another achievement was the creation of data visualisation graphics, which not only were able to display trends that could not be seen in the regression analysis, but also identified issues within particular datasets. Moreover, the suitability and application of particular visual models has been outlined.

Furthermore, the project was successful in addressing the quality and cleanliness of the data, and implementing methods to ensure the continuation of the project, notably in handling the General Cases dataset.

7.2 Improvements of Research

An area of research that could be notably improved concerns the research of particular regression models. A sufficient understanding of linear regression, ridge regression, and lasso regression was acquired prior to production however certain misunderstandings about models such as support vector regression lead to initially pursuing overly-ambitious model prospects. Alongside this, a non-linear-based regression model was included due to this same error of research. While these events did not hinder the project's progress, they certainly highlighted an area for improvement.

7.3 The Quality of the Data

While the overall quality of the datasets used in the artefact was generally good, there were some issues that may or may not need to be addressed. One such issue was the inconsistency that appeared in later entries in particular datasets, for instance, the General Cases dataset changed the update schedule from daily to weekly in 2022. This had an effect on the outcome of the time-series plots, and added to the untruthfulness of the data. Alongside this, the synchronicity of the data entries caused the large outlier spike on the 31st of January 2022, as it appears a backlog of cases were uploaded within that single day and the dataset

was not adjusted to account for it. Maintaining a protocol for updating national datasets could be the solution to solving this issue.

7.4 Reflections on the Implementation

Particular implementations in the code to produce the visualisations, while successfully implemented, could possibly be reduced into functions in a future update to the artefact, post-project. While this was originally believed to be unfeasible at the beginning of the development, due to the variety of the datasets, it later become apparent that there is a chance to explore the possibility of reducing the code down within functions, to reduce code redundancy. Due to the time constraints of the project and the current scale of the artefact, this is no longer possible. Retrospectively, this approach could have also improved time spent testing the artefact.

As discussed before, GitHub was a tool that was identified for usage throughout the project, to maintain integrity of the codebase in the event of a system error or an internal error. While GitHub was used to store the codebase, the frequency with which updates of the artefact were pushed to the remote repository were lacking. Typically updates were pushed after large quantities of work were completed, rather than incrementally. Not only did this increase the risk on the safety of the artefact but also impeded the ability to follow the chosen software development methodology. In future endeavours, this behaviour should be addressed.

7.5 The Results of the Project

Considering all this, it can be said that data visualisation and regression analysis are suitable tools for analysing COVID-19 data, as well as data created from future pandemics, to determine patterns, trends, and relationships. More specifically, it should be noted that regression analysis has shown exemplary success in modelling and understanding the factors that lead to mortality cases; there is still some improvement to be made in modelling transmission cases.

In addition to this, data visualisation has proven itself to not only be advantageous in determining the progression of the pandemic but also crucial in ensuring the quality of the data being used, which is paramount depending on what could be inspired by these results.

In conclusion, this project has displayed the ability and efficacy for data visualisation and regression analysis, amongst other Big Data techniques, to be employed to tackle large, fast developing emergencies, in order to ensure the national security of public health.

Bibliography

Almalki, A., Gokaraju, B., Acquaah, Y., & Turlapaty, A. (2022). Regression analysis for covid-19 infections and deaths based on food access and health issues. *Healthcare*, *10*, 324. https://doi.org/10.3390/healthcare10020324

Base, I. U. K. (2020). Understand measures of supercomputer performance and storage system capacity. *University Information Technology Services*. Retrieved April 6, 2023, from https://kb.iu.edu/d/apeq

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ*, *324*, 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Becker, D., King, T. D., & McMullen, B. (2015). *Big data, big data quality problem*. IEEE. https://doi.org/10.1109/BigData.2015.7364064

Borthakur, D. (2022). Hdfs architecture guide. *Apache Hadoop*. Retrieved April 6, 2023, from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*, 2. https://doi.org/10.5334/dsj-2015-002

Chang, K.-T. (2019). *Introduction to geographic information systems*. Mcgraw-Hill Education.

DeFanti, T. A., Brown, M. D., & McCormick, B. H. (1989). Visualization: Expanding scientific and engineering research opportunities. *Computer*, *22*, 12–16. https://doi.org/10.1109/2.35195

Elmeiligy, M. A., Desouky, A. I. E., & Elghamrawy, S. M. (2020). A multi-dimensional big data storing system for generated covid-19 large-scale data using apache spark. *arXiv*. Retrieved January 23, 2023, from https://arxiv.org/abs/2005.05036

FINVIZ. (2023). Sp 500 map. *FINVIZ*. Retrieved April 7, 2023, from https://finviz.com/map.ashx

Fisher, S. C. (2022). Forecasting the high-risk demographics for covid-19 in london using deep learning.

Fisher, S. C. (2023). Exploring the role of data analytics in combating future pandemics: Lessons learned from the covid-19 pandemic in the united kingdom: Interim report.

Fold in functional programming. (2016). *Tcler's Wiki*. Retrieved April 18, 2023, from https://wiki.tcl-lang.org/page/Fold+in+functional+programming

for Statistics Regulation, O. (2022). Code of practice. *Code of Practice for Statistics*. Retrieved April 13, 2023, from https://code.statisticsauthority.gov.uk/

Foundation, A. S. (n.d.-a). Apache hadoop 3.3.1 – mapreduce tutorial. *Apache Hadoop*. Retrieved April 11, 2023, from https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Foundation, A. S. (n.d.-b). Overview - spark 3.3.2 documentation. *Apache Spark*. Retrieved April 6, 2023, from https://spark.apache.org/docs/3.3.2/

Franke, G. R. (2010). Multicollinearity. *Wiley International Encyclopedia of Marketing*. https://doi.org/10.1002/9781444316568.wiem02066

Gantt, H. L. (1974). *Work, wages, and profits.* Easton [Pa.] Hive Pub. Co.

GitHub. (n.d.). Hello world. *GitHub Docs*. Retrieved April 13, 2023, from https://docs.github.com/en/get-started/quickstart/hello-world

Google. (2022). Use a gpu — tensorflow core. *TensorFlow*. Retrieved April 26, 2023, from https://www.tensorflow.org/guide/gpu

Government, U. (2022a). Coronavirus (covid-19) cases in the uk. *Coronavirus Dashboard*. Retrieved April 13, 2023, from https://coronavirus.data.gov.uk/

Government, U. (2022b). $\text{Femalecases}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=femaleCases&format=csv&release=2022-12-22

Government, U. (2022c). $\text{Femaledeaths}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=femaleDeaths28Days&format=csv&release=2022-12-22

Government, U. (2022d). $\text{Generalcases}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=cumCasesByPublishDate&metric=newCasesBySpecimenDateRollingRate&format=csv&release=2022-12-22

Government, U. (2022e). $\text{Generaldeaths}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=cumDailyNsoDeathsByDeathDate&format=csv&release=2022-12-22

Government, U. (2022f). $\text{Malecases}_2022 - 12 - 22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=maleCases&format=csv&release=2022-12-22

Government, U. (2022g). $\text{Maledeaths}_2022 - 12 - 22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=maleDeaths28Days&format=csv&release=2022-12-22

Government, U. (2022h). $\text{Reinfections}_2022 - 12 - 22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=newReinfectionsBySpecimenDateAgeDemographics&format=csv&release=2022-12-22

Government, U. (2022i). $\text{Vaccinationdemographics}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=vaccinationsAgeDemographics&format=csv&release=2022-12-22

Government, U. (2022j). $\text{Variants}_2022-12-22.csv$. *Coronavirus Dashboard*. https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=variants&format=csv&release=2022-12-22

Government, U. (2023). Vaccinations age demographics breakdown. *Coronavirus Dashboard*. Retrieved April 19, 2023, from https://coronavirus.data.gov.uk/metrics/doc/vaccinationsAgeDemographics

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, *9*, 90–95. https://doi.org/10.1109/mcse.2007.55

IBM. (2021). Big data analytics — ibm. *IBM*. Retrieved April 3, 2023, from https://www.ibm.com/analytics/big-data-analytics

Jadeja, M., & Shah, K. (2015). *Tree-map: A visualization tool for large data.*

Joyce, J. (2016). Bayes' theorem. *Stanford Encyclopedia of Philosophy*. Retrieved April 17, 2023, from https://plato.stanford.edu/entries/bayes-theorem/

Juddoo, S. (2015). *Overview of data quality challenges in the context of big data*. IEEE. https://doi.org/10.1109/cccs.2015.7374131

Khan, M. A., Khan, R., Algarni, F., Kumar, I., Choudhary, A., & Srivastava, A. (2021). Performance evaluation of regression models for covid-19: A statistical and predictive perspective. *Ain Shams Engineering Journal*, *13*. https://doi.org/10.1016/j.asej.2021.08.016

Laserson, U. (2023). Squarify. *GitHub*. Retrieved April 17, 2023, from https://github.com/laserson/squarify

Matulić, A. (2007). Normdist$_r$*egression.png*. *Wikimedia Commons*. Retrieved April 3, 2023, from https://commons.wikimedia.org/wiki/File:Normdist_regression.png

McKinney, W. (2011). Pandas: A foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*, 1–9. Retrieved April 6, 2023, from https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf

Obsidian. (n.d.-a). Obsidian: A knowledge base that works on local markdown files. *Obsidian*. Retrieved April 13, 2023, from https://obsidian.md/

Obsidian. (n.d.-b). Sync. *Obsidian*. Retrieved April 13, 2023, from https://obsidian.md/sync

O'Reilly. (n.d.). 21. evolutionary prototyping - rapid development: Taming wild software schedules [book]. *O'Reilly*. Retrieved April 13, 2023, from https://www.oreilly.com/library/view/rapid-development-taming/9780735634725/ch21.html

Organisation, W. H. (2023). Who covid-19 dashboard. *World Health Organisation*. Retrieved April 3, 2023, from https://covid19.who.int/

pandas. (n.d.). Pandas.read$_c$*sv*|*pandas*1.2.4*documentation*. *pandas*. Retrieved April 18, 2023, from https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qudrat-Ullah, H., & Tsasis, P. (2017). *Innovative healthcare systems for the 21st century*. Cham Springer International Publishing.

Raw, G. (2022). How many nhs hospitals are there in the uk? *Sanctuary Personnel*. Retrieved April 12, 2023, from https://www.sanctuarypersonnel.com/blog/2020/10/how-many-nhs-hospitals-are-there-in-the-uk?source=google.com

Richards, G. (2008). The logistic sigmoid function. *Wikimedia Commons*. Retrieved April 17, 2023, from https://commons.wikimedia.org/wiki/File:Logistic-curve.svg

Richman, L. (2002). *Project management step-by-step [electronic resource]*. Amacom.

Saket, B., Endert, A., & Demiralp, Ç. (2019). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*, 2505–2512. https://doi.org/10.1109/TVCG.2018.2829750

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, *11*, 92–99. https://doi.org/10.1145/102377.115768

Simplilearn. (2017). Regression analysis — data science tutorial — simplilearn - youtube. *YouTube*. Retrieved April 3, 2023, from https://www.youtube.com/watch?v=DtOYBxi4AIE

Team, J. (2015). The jupyter notebook — jupyter notebook 6.5.4 documentation. *The Jupyter Notebook*. Retrieved April 13, 2023, from https://jupyter-notebook.readthedocs.io/en/stable/notebook.html#code-cells

Torvalds, L., & Google. (2007). Tech talk: Linus torvalds on git. *YouTube*. Retrieved April 13, 2023, from https://www.youtube.com/watch?v=4XpnKHJAok8&t=90s

University, J. H. (2022). Johns hopkins coronavirus resource center. *Johns Hopkins Coronavirus Resource Center*. Retrieved April 3, 2023, from https://coronavirus.jhu.edu/map.html

Vernier, E., Sondag, M., Comba, J., Speckmann, B., Telea, A., & Verbeek, K. (2020). Quantitative comparison of time-dependent treemaps. *Computer Graphics Forum*, *39*, 393–404. https://doi.org/10.1111/cgf.13989

Appendix A

Obsidian Templates

Prototype Documentation Template

```
Created {{date}} | {{time}}
# {{title}}


---
Type:
---
Dependent variables:
-

Independent variables:
-
---
Coefficient of determination/R-Squared:
---
MSE:
MAE:
```

Project Diary Template

```
## {{date}}


---
```

Appendix B

Project Canvas

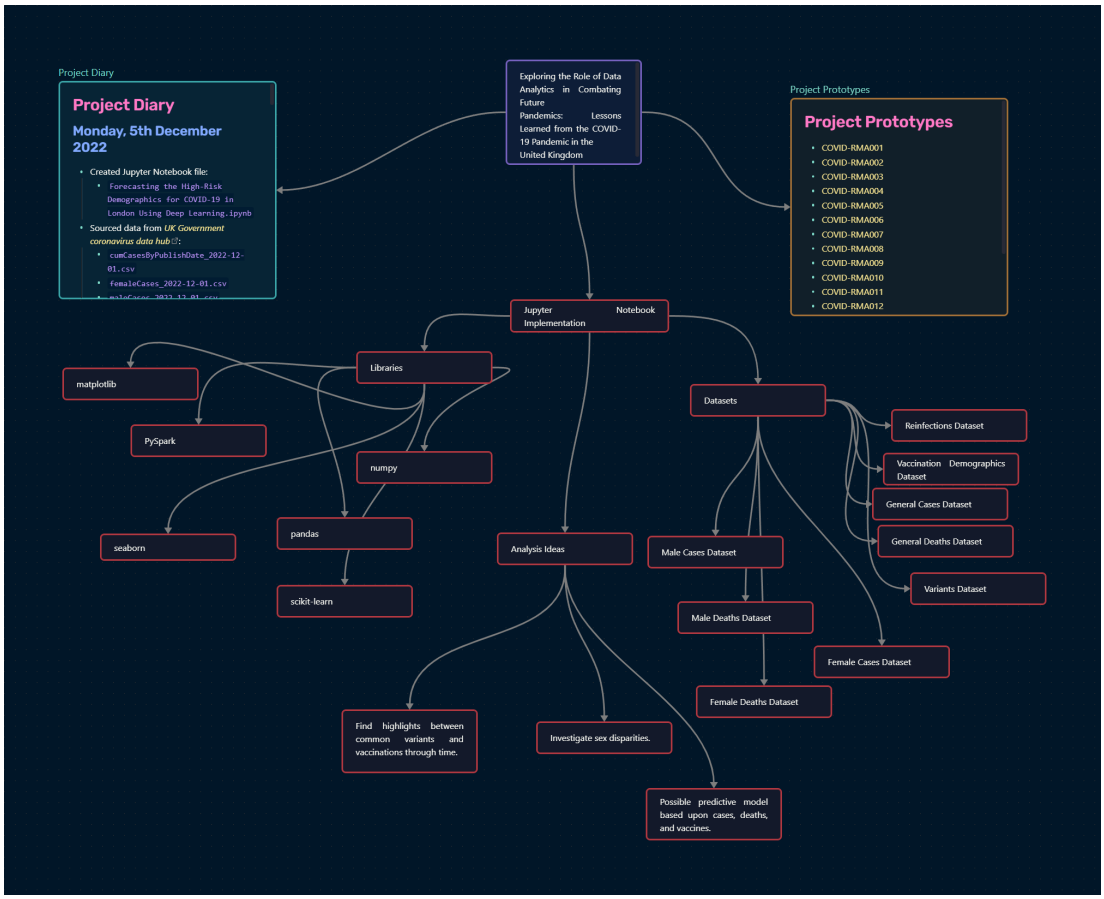

Figure B.1: An image of the project canvas on Obsidian.

Appendix C

Model Prototype Documentation

COVID-RMA001

A model aimed to examine the impact of sex, region, and age, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Region
- Age

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.23

MSE: 789001641.45 MAE: 21189.57

COVID-RMA002

A model aimed to examine the impact of sex and age, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Age

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.18

MSE: 843555118.56 MAE: 20991.71

COVID-RMA003

A model aimed to examine the impact of sex and region, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Region

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.06

MSE: 971845853.32 MAE: 23720.35

COVID-RMA004

A model aimed to examine the impact of sex, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.00

MSE: 1025673656.29 MAE: 24191.30

COVID-RMA005

A model aimed to examine the impact of sex, region, and age, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Region
- Age

Independent variables:

- Rate

Coefficient of determination/R-Squared: 0.13

MSE: 192669864.45 MAE: 11549.41

COVID-RMA006

A model aimed to examine the impact of sex and region, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Region

Independent variables:

- Rate

Coefficient of determination/R-Squared: 0.01

MSE: 218468738.78 MAE: 12099.62

COVID-RMA007

A model aimed to examine the impact of sex and age, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex
- Age

Independent variables:

- Rate

Coefficient of determination/R-Squared: 0.12

MSE: 193781104.32 MAE: 11590.38

COVID-RMA008

A model aimed to examine the impact of sex, using a combination of the male cases and female cases datasets.

Type: Linear Regression

Dependent variables:

- Sex

Independent variables:

- Rate

Coefficient of determination/R-Squared: 0.00

MSE: 219519568.00 MAE: 12140.02

COVID-RMA009

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 285544521.62 MAE: 12507.01

COVID-RMA010

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 285568354.80 MAE: 12504.66

COVID-RMA011

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. This model used a 1

Type: Logistic Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.01

MSE: 1581811440.23 MAE: 24436.90

COVID-RMA012

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The alpha constant was set to 0.1.

Type: Lasso Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 285578027.07 MAE: 12504.84

COVID-RMA013

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The alpha constant was set to 1.0.

Type: Lasso Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 286874914.45 MAE: 12517.37

COVID-RMA014

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. This model used a 1

Type: Support Vector Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: -0.18

MSE: 1103690743.36 MAE: 19838.70

COVID-RMA015

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. This model used a 1

Type: Elastic Net Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.05

MSE: 889099287.44 MAE: 22951.40

COVID-RMA016

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.05

MSE: 973540895.92 MAE: 23527.81

COVID-RMA017

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 285558920.82 MAE: 12505.42

COVID-RMA018

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The model has been adjusted for prior errors in computation, due to encoding.

Type: Bayesian Ridge Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.68

MSE: 325810485.03 MAE: 13647.62

COVID-RMA019

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The model has been adjusted for prior errors in computation, due to encoding.

Type: Ridge Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.68

MSE: 325810538.37 MAE: 13647.74

COVID-RMA020

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The model has been adjusted for prior errors in computation, due to encoding.

Type: Linear Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.68

MSE: 325810547.88 MAE: 13647.85

COVID-RMA021

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 81537942784.51 MAE: 231023.47

COVID-RMA022

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 81514486074.53 MAE: 230971.80

COVID-RMA023

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.48

MSE: 390362179465.29 MAE: 496343.60

COVID-RMA024

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 81537592909.86 MAE: 231022.54

COVID-RMA025

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 263693.62 MAE: 399.20

COVID-RMA026

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 263811.70 MAE: 399.36

COVID-RMA027

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 264056.58 MAE: 399.30

COVID-RMA028

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.22

MSE: 25527367.18 MAE: 4199.26

COVID-RMA029

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40757374884.08 MAE: 115685.50

COVID-RMA030

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40768928360.78 MAE: 115710.95

COVID-RMA031

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40769103420.55 MAE: 115711.39

COVID-RMA032

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.63

MSE: 80063235873.42 MAE: 146610.18

COVID-RMA033

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 76501919948.48 MAE: 217752.55

COVID-RMA034

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 76502325987.07 MAE: 217754.49

COVID-RMA035

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 76502405037.38 MAE: 217755.25

COVID-RMA036

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 76502045354.62 MAE: 217753.25

COVID-RMA037

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.80

MSE: 156654225005.24 MAE: 297726.27

COVID-RMA038

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497240.39 MAE: 749.63

COVID-RMA039

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1496988.71 MAE: 751.33

COVID-RMA040

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497265.17 MAE: 749.52

COVID-RMA041

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497263.24 MAE: 749.53

COVID-RMA042

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.52

MSE: 16803400.32 MAE: 3363.21

COVID-RMA043

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 38251708594.44 MAE: 109251.09

COVID-RMA044

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 38251911487.89 MAE: 109252.91

COVID-RMA045

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 38251951151.29 MAE: 109252.39

COVID-RMA046

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.66

MSE: 78335514202.78 MAE: 150544.74

COVID-RMA047

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.50

MSE: 511139607.21 MAE: 15293.80

COVID-RMA048

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.68

MSE: 325809013.06 MAE: 13645.82

COVID-RMA049

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40288379625.42 MAE: 115580.51

COVID-RMA050

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40300076715.88 MAE: 115607.64

COVID-RMA051

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.94

MSE: 40300255778.24 MAE: 115608.07

COVID-RMA052

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.62

MSE: 80512372372.91 MAE: 147168.31

COVID-RMA053

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 80576495557.23 MAE: 230761.82

COVID-RMA054

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 80599889620.06 MAE: 230815.93

COVID-RMA055

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 80600247499.90 MAE: 230816.83

COVID-RMA056

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.79

MSE: 161007253416.23 MAE: 290863.88

COVID-RMA057

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 263693.62 MAE: 399.20

COVID-RMA058

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 263811.70 MAE: 399.36

COVID-RMA059

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 264056.58 MAE: 399.30

COVID-RMA060

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.46

MSE: 17491329.58 MAE: 3472.73

COVID-RMA061

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 36974561498.10 MAE: 107367.59

COVID-RMA062

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 36974765824.48 MAE: 107369.41

COVID-RMA063

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 36974805283.60 MAE: 107368.84

COVID-RMA064

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.66

MSE: 77617660395.90 MAE: 149968.42

COVID-RMA065

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 73947625755.82 MAE: 213985.54

COVID-RMA066

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 73948034660.25 MAE: 213987.49

COVID-RMA067

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 73948113302.04 MAE: 213988.16

COVID-RMA068

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.80

MSE: 155218517391.47 MAE: 296573.64

COVID-RMA069

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497240.39 MAE: 749.63

COVID-RMA070

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Lasso Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1496988.71 MAE: 751.33

COVID-RMA071

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Linear Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497265.17 MAE: 749.52

COVID-RMA072

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Elastic Net Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.52

MSE: 16803400.32 MAE: 3363.21

COVID-RMA073

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.68

MSE: 325810485.03 MAE: 13647.62

COVID-RMA074

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.48

MSE: 389972501439.44 MAE: 496982.70

COVID-RMA075

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.22

MSE: 25527367.18 MAE: 4199.26

COVID-RMA076

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.90

MSE: 73947767847.63 MAE: 213986.31

COVID-RMA077

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Bayesian Ridge Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.96

MSE: 1497263.24 MAE: 749.53

COVID-RMA078

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Huber Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.64

MSE: 370881703.08 MAE: 12713.05

COVID-RMA079

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Huber Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.71

MSE: 214761280522.90 MAE: 343032.82

COVID-RMA080

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Huber Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: -1.26

MSE: 73446731.95 MAE: 7055.83

COVID-RMA081

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Huber Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.72

MSE: 217720937314.27 MAE: 332441.99

COVID-RMA082

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Huber Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.07

MSE: 33002461.98 MAE: 4664.53

COVID-RMA083

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date. The random state is set to 42.

Type: Theil Sen Regression

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.67

MSE: 338318818.81 MAE: 13308.00

COVID-RMA084

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date. The random state is set to 42.

Type: Theil Sen Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 81090054698.60 MAE: 227322.56

COVID-RMA085

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date. The random state is set to 42.

Type: Theil Sen Regression

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.99

MSE: 398255.82 MAE: 479.46

COVID-RMA086

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date. The random state is set to 42.

Type: Theil Sen Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.89

MSE: 86863613347.34 MAE: 225822.16

COVID-RMA087

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date. The random state is set to 42.

Type: Theil Sen Regression

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.93

MSE: 2325238.06 MAE: 1159.33

COVID-RMA088

A model aimed to examine the impact of sex, region, age, and date, using a combination of the male cases and female cases datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Sex
- Region
- Age
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.49

MSE: 521059149.16 MAE: 15710.41

COVID-RMA089

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.64

MSE: 90780602754.99 MAE: 165864.17

COVID-RMA090

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.76

MSE: 181545661918.13 MAE: 328574.50

COVID-RMA091

A model aimed to examine the impact of variant, region, sequences of variants, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Variant
- Region
- Sequenced
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.52

MSE: 15543591.84 MAE: 3153.84

COVID-RMA092

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases
- Deaths

Coefficient of determination/R-Squared: 0.48

MSE: 117732716359.42 MAE: 183646.95

COVID-RMA093

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Cases

Coefficient of determination/R-Squared: 0.69

MSE: 235439600156.29 MAE: 363144.66

COVID-RMA094

A model aimed to examine the impact of vaccines, region, and date, using a combination of the general cases and general deaths datasets. The date is formatted as days since the earliest date.

Type: Orthogonal Matching Pursuit

Dependent variables:

- Vaccinations
  - Separated by one, two, or three doses received
- Region
- Date

Independent variables:

- Deaths

Coefficient of determination/R-Squared: 0.27

MSE: 25832562.55 MAE: 4149.24