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

Interim Report

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

January 2023

UNIVERSITY OF
LINCOLN

School of Computer Science
University of Lincoln
United Kingdom

Submitted in partial fulfilment of the requirements for the Degree of BSc(Hons) Computer Science

Supervisor Dr. Kamaran Fathulla

Word Count: 1850

Contents

List of Figures

List of Tables

1 Literature Review

1.1 Background

Since the beginning of the COVID-19 pandemic in early 2020, the world has undergone a monumental change in how we handle and assess our healthcare data. The speed that coronavirus spread across the country was unprecedented and made this a necessity; in the first week of 2022, the number of new coronavirus cases recorded was approximately 1.25 million cases (Mathieu et al., 2020).

As we moves forward, we should look at the effectiveness of techniques that have been used and investigate if there are more efficient methods of data analysis and data collection. By ensuring the robustness of data methodology used for this pandemic, we can allow for more accurate and timely analysis.

The aim of this literature review is to explore techniques that have already been employed to analyse the data across the world to highlight high risk demographics, detailing the strengths and weaknesses of the current methods in order to gain an understanding of potential area of improvement for how we handled the data in the United Kingdom, and to identify possible areas of innovation and development.

1.2 Research Objectives

The questions that will guide the research for this project are:

- How has pre-existing data been collected?
- How has pre-existing data been stored and managed?
- What is the current role that data analytics occupies in modelling and tracking the pandemic?
- Have artificial intelligence and machine learning been utilised to analyse data, and how?

1.3 Previous Studies

1.3.1 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 to divide the inserted data into a set of Resilient Distributed Datasets, as seen in Figure 1. HDFS is a distributed file system designed to provide fault-tolerant data management. Spark is an analytics engine designed for fast data processing; it provides a programming interface to allow access to data parallelism.

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

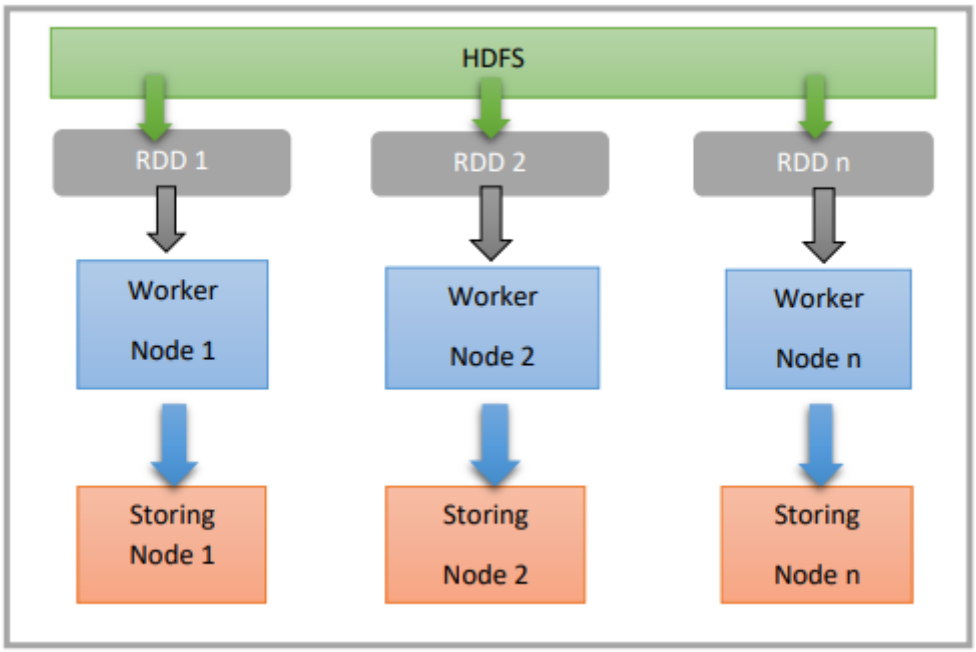

Figure 1: An implementation of an ASF framework (Elmeiligy et al., 2020).

1.3.2 Concerning the Analysis of the Data

There are a wide variety of methods that have been used to analyse data related to COVID-19; this will likely be due to the variety of the data.

In mid-2020, there was a meta-analysis into the potential of Big Data and Artificial Intelligence in managing the COVID-19 pandemic (Bragazzi et al., 2020), focusing on applications at differing time scales. They determined that for short-term applications, Big Data techniques could be used to monitor the progress of the pandemic outbreak in real-time. The meta-analysis noted a study by a team at the Fogarty International Center (Sun et al., 2020), where they monitored news media and social networks in China to reconstruct the progression of the outbreak in mainland China.

The team cross-referenced reports from DXY.cn[1] with international data sources via media sources (Kyodo News, The Strait Times, and CNN), governments, and official health authorities. They found that while the model worked early on, the overwhelming of healthcare systems and reporting fatigue meant that the data captured from non-traditional sources was becoming diminished over time and couldn't keep up.

Regression analysis is an area of data analysis that has 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 regression model and a GIS regression model[2], as seen in Figure 2; 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.

[1] DXY.cn is an online health-focused community for physicians and healthcare professionals in China.
[2] A geographic information system pairs geographic data with software tools to manage and analyse them. (Chang, 2019)

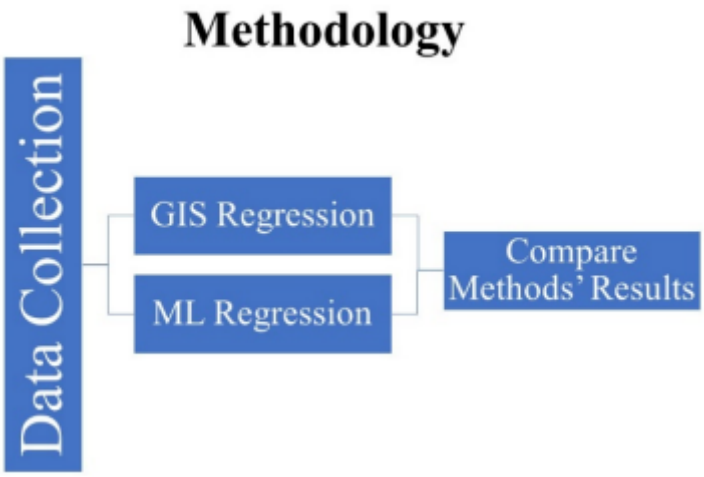

Figure 2: A combination of GIS regression and ML regression (Almalki et al., 2022).

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.

1.4 Conclusion and Future Directions

Upon final summary, it appears that there is a clear potential for Big Data techniques to be integrated into our health crisis management protocols, to observe and track progressions. Regarding the collection of health data, it appears that there could be an expansion to the harvesting pool beyond official sources, although I should be wary of their limitations. So far, regression models have returned varying qualities in their outputs and so I should aim to compare their efficacy, and aim to observe their performance on different types of data.

The project aims to examine whether an alternative understanding of potential underlying trends and patterns in the large quantities of data can be found and exploited by using statistical and regression modelling to discover them. If there are in fact more underlying trends, both risk assessment and healthcare forecasting could be improved; the results of the project could potentially guide future recommendations for mitigation strategies and resource allocation.

The proposed contributions of the project are:

- The project will aim to build my own model for healthcare data originating from the United Kingdom, using ASF software to manage the data.
- The project will aim to explore the usage of different regression analysis models on the harvested healthcare data and, if successful, compare their performance.
- The project will aim to explore more approaches as to how healthcare data could be collected.

2 Progress Update

2.1 Status of Objectives

2.1.1 Week 1 - 5/10/22

The original project plan was to conduct a machine learning project to use regression modelling to highlight high-risk demographics in London for COVID-19. The project blueprint was scheduled to begin on 5th December 2022 and conclude 16 weeks later on the week beginning 20th March 2023, with six weeks allotted for surplus in the event of any unforeseen delays; the Gantt chart for this plan can be seen in Table 1.

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 1: The original Gantt chart (Fisher, 2022).

The first stage of the project was to collect data from the sources outlined in the project proposal (Fisher, 2022). These sources were:

- The London Datastore (Authority, 2022)
- The UK Government Coronavirus Data Hub[3] (Government, 2022)
- The NHS Digital Hub[4] (NHS, 2019)

Four datasets were sourced from the Government Data Hub and two from the London Datastore; the data on the NHS Hub was deemed no longer necessary and so data has not been sourced, as of the time of writing. What I did not realise when I sourced the data is that I failed to properly document the parameters selected and the date I collected the data, making replication of the initial batch of data unfeasible. The datasets[5] that I collected at the time can be seen in Table 2.

From the Government Data Hub		**From the London Datastore**	
Dataset	Description	Dataset	Description
General Cases	A dataset of the cumulative cases by date of publishing.	Cases	A second dataset of the cumulative cases by date of publishing.
Male Cases	A dataset of the cumulative male cases by date of publishing.	Admissions	A dataset of the cumulative number of admissions by date of publishing.
Female Cases	A dataset of the cumulative female cases by date of publishing.		
Vaccinations	A dataset of the cumulative number of vaccinated patients.		

Table 2: The original collection of datasets.

2.1.2 Week 3 - 19/12/22

The second stage of project began on 19th December 2022, although the data collection did not take as long as previously believed. Around this time, I had been examining the data and began to have doubts about the possible relationships that could be discovered from it. I contacted my supervisor to query if I could make alterations to the scope of my project; I received confirmation that the scope could be altered around technical issues and so I shifted my project towards a Big Data style focus. I decided that I would have the predictive model as a more aspirational target rather than a solidified target, due to my lack of understanding on how successful a regression model might be.

While cleaning the data, I discovered that some of the data had corrupted. The Government Data Hub offers the user the ability to combine up to five separate datasets into a single dataset, however it doesn't guarantee that the purity of the resulting file. Due to the corruption of the files, I decided that I would have to implement RDDs to configure the data to the desired output.

I decided to break down the datasets according to age and location where appropriate, as it would allow me to get a more granular look at the data when it came time to analyse them.

[3]The UK Government Coronavirus Data Hub shall be referred to as the 'Government Data Hub' for the rest of the report.
[4]The NHS Digital Hub shall be referred to as the 'NHS Hub' for the rest of the report.
[5]Most of these datasets contained categories for age and the boroughs of London.

2.1.3 Week 4 - 26/12/22

By this stage in the project, I had decided to expand the harvesting pool for the data that I would be using; if I wanted this project to be an analysis of how Big Data could be used, I would ideally be using as large a pool of data as possible. Subsequently, I discarded the currently collected data and collected data covering the entirety of England. This time, I made sure to properly document the exact parameters[6] used and the date of collection, via a naming scheme for the `.csv` files. The new datasets, still in use as of the time of writing, were all collected from the Government Data Hub and can be seen in Table 3.

Dataset	**Description**	**Date of Release**
Male Cases	A dataset of the cumulative male cases by date of publishing.	22-12-2022
Female Cases	A dataset of the cumulative female cases by date of publishing.	22-12-2022
Male Deaths	A dataset of the cumulative male deaths by date of publishing.	22-12-2022
Female Deaths	A dataset of the cumulative female deaths by date of publishing.	22-12-2022
General Cases	A dataset of the cumulative cases by date of publishing.	22-12-2022
General Deaths	A dataset of the cumulative deaths by date of publishing.	22-12-2022
Vaccination Demographics Dataset	A dataset of the cumulative populations of vaccinated patients, by the number of vaccinations received by date of publishing.	22-12-2022
Reinfections Dataset	A dataset of the cumulative number of reinfections by date of publishing.	22-12-2022
Variants	A dataset of the cumulative number of COVID-19 variant cases by date of publishing.	22-12-2022

Table 3: The current collection of datasets.

Cleaning the new data was a relatively simple task as the formatting of the data is mostly identical, and I only had to make minor adjustments to the code within the span of some hours.

2.1.4 Week 5 - 2/1/23

By the fifth week, I was still on track with my targets and so I began my research into previous studies. I had purposely given myself an extended period of time of research to account for the wave of external assignments that I had due in the month of January. This time management was successful and I was able to easily balance external work with my project.

2.1.5 Week 8 - 23/1/23

This week, I concluded my research. As mentioned before, the research has guided me towards a more adjusted path and so I felt it necessary to update my time plan to go along with that; the new Gantt chart can be seen in Table 4.

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: The updated Gantt chart.

2.2 Updated Aims

- Build a statistical model for the selected healthcare data covering England.
- Evaluate the efficacy of regression modelling on the selected healthcare data.
- Evaluate where the healthcare data could be improved to allow for better modelling.

2.3 Preliminary Results and Deliverables

Currently, the implementation has rudimentary completions of the statistical models, as seen in Figure 3, but there not currently any key results.

[6]The parameters are stored within a saved hyperlink.

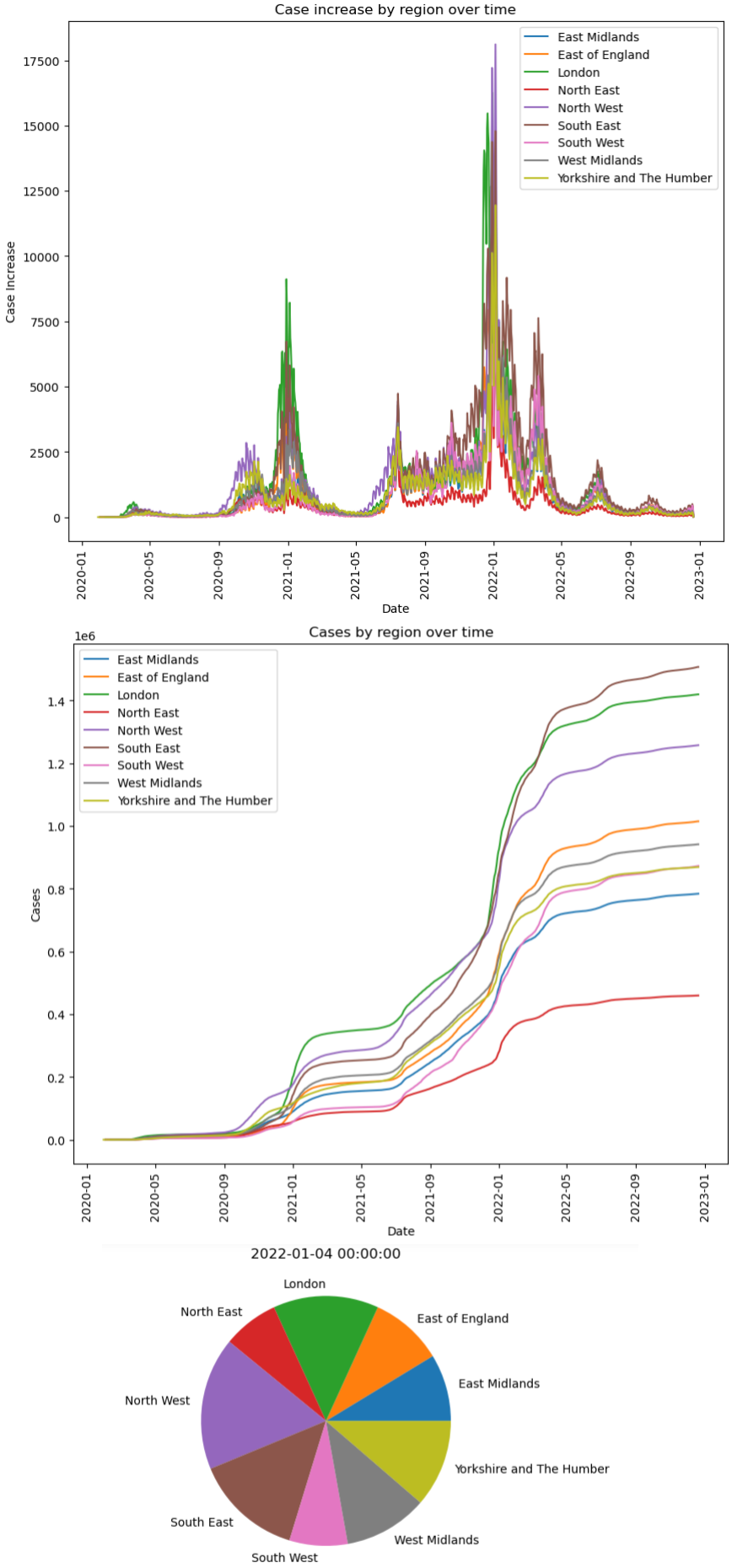

Figure 3: The initial implementation of statistical modelling.

2.4 Supervisory Engagement and Project Management

Over the time span of the project, I have had two meetings with my supervisor, however I have kept regular contact where necessary. I have not asked for more frequent meetings simply due to not needing a full face-to-face meeting.

In the meantime, I have kept a project diary that is updated whenever any action is completed, so that I have a fully detailed record of what I have done, but also to aid in keeping track of what I need to do next. I plan to publish this diary as an appendix item in the final dissertation report. Alongside the diary, I have a supplementary map of project ideas and importances, as seen in Figure 4.

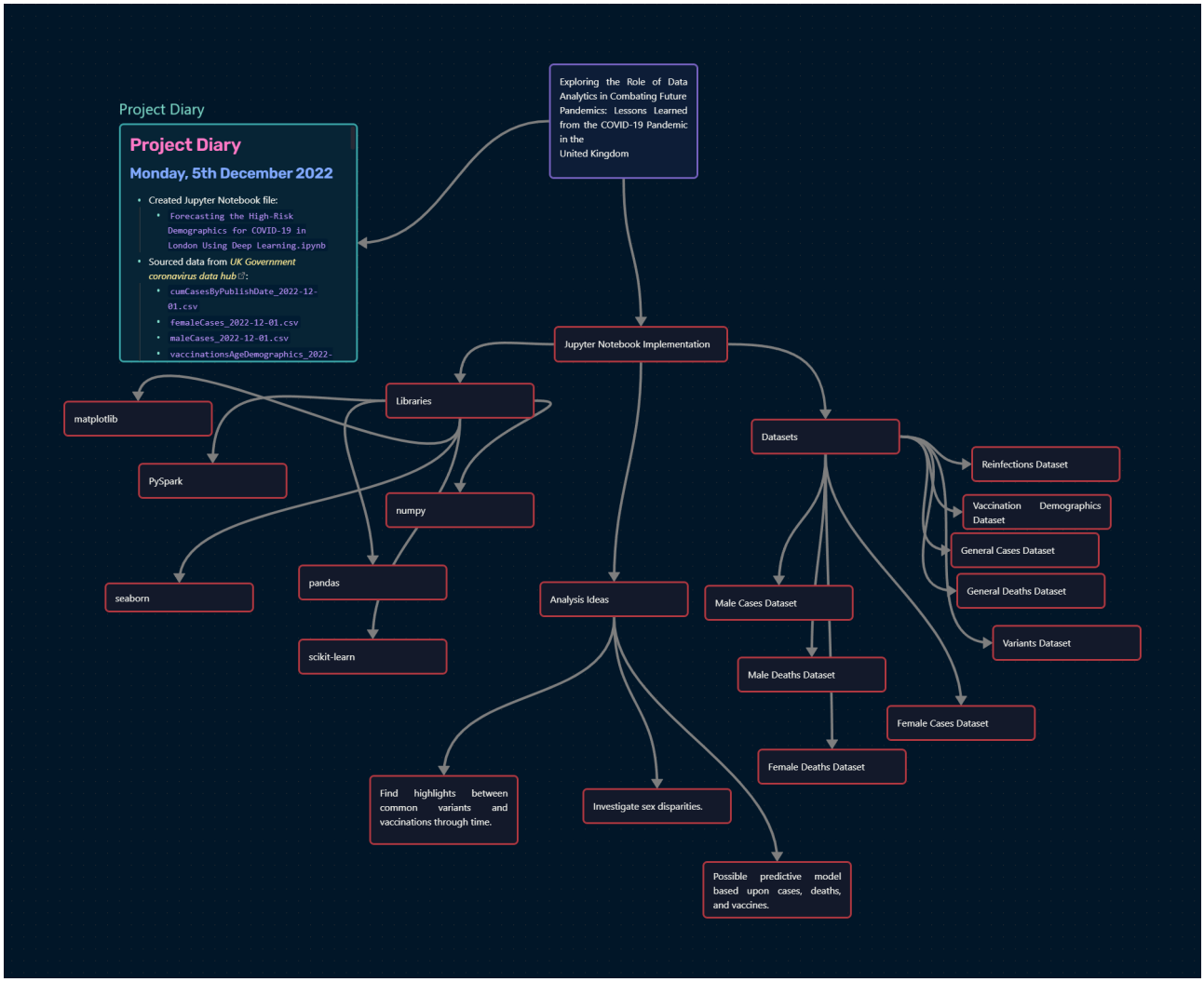

Figure 4: The project map.

3 References

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

Authority, G. L. (2022). Coronavirus (covid-19) cases – london datastore. *London Datastore*. Retrieved October 17, 2022, from https://data.london.gov.uk/dataset/coronavirus--covid-19--cases

Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., & Wu, J. (2020). How big data and artificial intelligence can help better manage the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, *17*, 3176. https://doi.org/10.3390/ijerph17093176

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

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:2005.05036 [cs]*. Retrieved January 23, 2023, from https://arxiv.org/abs/2005.05036

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

Government, U. (2022). Coronavirus (covid-19) cases in the uk. *coronavirus.data.gov.uk*. Retrieved October 24, 2022, from https://coronavirus.data.gov.uk/

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus pandemic (covid-19). *Our World in Data*.

NHS. (2019). Home - nhs digital. *NHS Digital*. Retrieved October 25, 2022, from https://digital.nhs.uk/

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

Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *The Lancet Digital Health*, *2*, e201–e208. https://doi.org/10.1016/s2589-7500(20)30026-1