*Division of Computing Science and Mathematics*
*Faculty of Natural Sciences*
*University of Stirling*

# NHS Scotland Open Data: A data visualisation study about child health

**Natalie Polack**

**Dissertation submitted in partial fulfilment for the degree of Master of Science in Big Data**

**September 2019**

# Abstract

For the last five years, the Information Services Division (ISD), which is part of the Public Health and Intelligence strategic business unit within NHS National Services Scotland (NSS), has been using Tableau to generate data visualisations and data stories for their customers. Recently though, there has been a trend towards the use of the programming language R and the Shiny package for data analysis and the creation of visualisations. Benefits to ISD of moving from Tableau to R include potential savings on licence fees and the costs of IT support, and more flexibility with some functionality such as layouts of visualisations and dashboards.

During the same timeframe, the Scottish Government published their 'Open Data Strategy'. In line with this new strategic direction on the use of open data, NSS took the decision to establish the 'NSS Open Data Working Group'. In September 2017, the first dataset was released on the NHS Scotland Open Data portal and as of 1st September 2019, there are 38 open datasets now available. These open datasets are released under the UK Open Government Licence (OGL), which means they are free to use and re-use but ISD currently receive very little feedback from customers on how the data is being exploited.

This dissertation had three main objectives. Firstly, data on the NHS Scotland Open Data Portal was accessed via the API. This ensured that the option to query a dataset and select a subset of it, rather than always having to download the entire dataset, was available. Secondly, the programming language R and the R packages, ggplot2 and Shiny were evaluated for creating data visualisations. The NHS Scotland Open Data portal contains a group of datasets related to infants and children and these were used to create visualisations about Child Health in Scotland. Thirdly, ISD receives very few comments on the ease of use of the NHS Scotland Open Data portal or how the data is being used, so the 'User Experience' undergone during this dissertation was discussed in order to provide some feedback.

The programming language R was used to write scripts for making a set of calls to the API for the NHS Open Data portal to request resource ID numbers, the total number of records that matched each query and then the records themselves. The ggplot2 package was used to create visualisations in R, and the Shiny package was used to build an application about Child Health. Similarities and differences between the data tables available from the ISD website and the open data available on the NHS Open Data portal were assessed.

The first objective of this dissertation, to learn how to access the data via the API, was successfully achieved and a set of scripts have been written to query data in the 'Child Health' group of datasets. The methodology used is discussed in Section 4.1. The second objective was to evaluate R and ggplot2 for the creation of visualisations, and Shiny for the generation of a data story. Three datasets were investigated using R and a range of visualisations illustrating the effect of factors such as SIMD, maternal age, trends with time and by Health Board have been created using ggplot2. These can be seen in Section 4.2. A specification for a Shiny application about Child Health has been written and can be found in Section 4.3. A prototype application has been produced and suggestions for further work to develop this application are given in Section 5.3. The third objective was to provide some feedback on the 'User Experience' of working with the open data and this is given in Section 4.4.

# Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my University project except for the following cases:

- Much of the background information in Section 1.1 was obtained from a presentation entitled: 'Delivering a single open data portal for NHS in Scotland' given by Dr Stefan Teufl, Information Analyst for the Open Data Team in NHS National Services Scotland in June 2019.
- Use of the API for the NHS Open Data portal was guided by a useful blog post entitled: 'Accessing APIs from R (and a little R programming)' [1] a copy of which was supplied by Dr Stefan Teufl.
- In the R script entitled: P1andMatBMI, lines 331 – 336 were provided by Dr Stefan Teufl.

**Signature**: Dr Natalie Polack

**Date**: 6th September 2019

# Acknowledgements

The open data platform used in this dissertation is managed for NHS Scotland by NHS National Services Scotland. The data is released under the UK Open Government Licence (OGL) [2] which means it is free to use and re-use.

I would like to express my gratitude to Dr Stefan Teufl, Information Analyst for the Open Data Team in NHS National Services Scotland for all the guidance and support with which he has provided me as my industrial supervisor for this dissertation. His help has been invaluable.

I would also like to thank Jonathan Cameron and Dr Richmond Davies for their assistance with arranging my industrial placement with the NHS, Professor Gabriela Ochoa for agreeing to be my academic supervisor for this dissertation, and Dr David Cairns for providing me with a useful paper about the Shiny package for R [3] and for introducing me to the R cheat sheets on the RStudio website.

Marian and the team at the Data Lab have provided me with sponsorship plus many excellent opportunities over the last two years for meeting students, companies and organisations from all over Scotland who are involved in Data Science and for this I would like to thank them.

And finally, a big thank-you to Mike, Luke and Ross for all their support and encouragement over the last two years.

# Table of Contents

# List of Tables

# List of Figures

# 1 Introduction

## 1.1 Background and Context

NHS Scotland currently has about 140 000 employees who work within 14 Regional Health Boards, 7 Special Boards and one Public Health Body [4]. The Special Boards provide national services such as such as NHS24, Blood transfusion services and the Scottish Ambulance Service [5]. Another of these Special Boards is NHS National Services Scotland (NSS), which comprises 6 strategic business units. The Information Services Division (ISD) is part of the Public Health and Intelligence strategic business unit within NSS and the role of ISD as described on the division's website [6] is to provide:

> *'… health information, health intelligence, statistical services and advice that support the NHS in progressing quality improvement in health and care and facilitate robust planning and decision making.'*

NSS is authorised under The Official Statistics (Scotland) Order 2008 [7] to produce official statistics and works to the Code of Practice for Official Statistics, which is maintained by the UK Statistics Authority (UKSA) [8]. Each year, ISD releases over 230 publications via their website. A publication can consist of a summary and a full report in PDF format; supporting materials such as a Glossary, a technical report and concept guides in PDF format; data tables as Excel spreadsheets; and visualisations generated using Tableau [9], an analytics platform for producing data visualisations via a drag-and-drop methodology and the software language and environment, R [10].

In February 2015, the Scottish Government published their 'Open Data Strategy' [11]. In this document, the Scottish Government set out their ambition to:

> *'… create a Scotland where non-personal and non-commercially sensitive data from public services is recognised as a resource for wider societal use and as such is made open in an intelligent manner and available for re-use by others …'*

The Scottish Government's vision was to support the delivery of improved public services through public organisations making better use of data; to encourage innovative use of data leading to wider social and economic benefits; and to ensure accountability and transparency in the delivery of public services.

The publication of the 'Open Data Strategy' was quickly followed by the passing of 'The Re-use of Public Sector Information Regulations 2015' [12] [13] that came into force in July 2015. In line with this new strategic direction from the Scottish Government on the use of open data, NSS took the decision to establish the 'NSS Open Data Working Group'. This working group was tasked with delivering an options paper to debate the means by which NSS should release open data. Open Data is defined by the 'Open Definition' project from the Open Knowledge Foundation [14] as follows:

> *'Open data and content can be **freely used**, **modified**, and **shared** by **anyone** for **any purpose**'*

The goal for NSS was to develop a single open data portal that would encompass data from all 14 Regional Health Boards, 7 Special Boards and one Public Health Body within the NHS. Uniform technical standards were applied to ensure consistency of the data and that it was suitable for as wide a range of users as possible. In addition, the open data portal is covered by the UK Open Government Licence (OGL) [2], so data can be used freely, unlike data tables provided in a proprietary format.

The format that the open data should take had to be considered. Sir Tim Berners-Lee suggested a 5-star deployment system for open data, which classifies open data by the manner in which it is released. The 5 Star Open Data Model from the 5 ★ Open Data website [15] is shown in Figure 1 and the descriptions of 1 to 5 star data are given in Table 1.

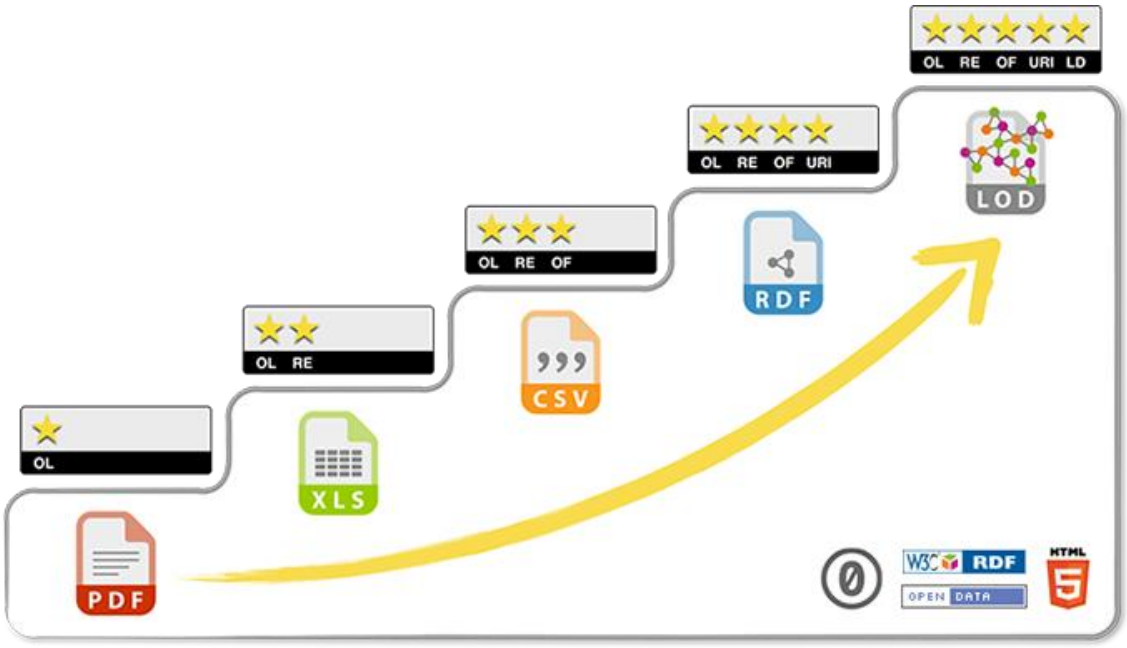


**Figure 1. The 5 ★ Data Model**

(Source: 5 ★ Open Data website [15]. Acronyms: OL = Open Licence, RE = Re-usable, CSV =Comma Separated Value, OF = Open Format, RDF = Resource Description Framework, URI = Uniform Resource Identifier, LD = Linked Data, LOD = Linked Open Data)

**Table 1. Descriptions of 1 to 5 Star Data**

| Number of stars | Description |
|---|---|
| ★ | Data is available on the Web under an open licence, in any format, for example as PDF document |
| ★★ | Data is available as structured data, for example as an Excel spreadsheet |
| ★★★ | Data is available in a structured and machine-readable open format that does not require any proprietary software, for example as a CSV file |
| ★★★★ | Data points can each have a Uniform Resource Identifier (URI) [16] and a Resource Description Framework or RDF graph [17] can be used to represent the data |
| ★★★★★ | Data points are linked to other data to give linked open data |

Most people are competent at using 1 – 3 star open data but specialist training and skills are often required for 4 to 5 star open data. 3 star open data is structured, machine readable and

available in a non-proprietary format such as CSV, which means that a wider range of users are able to work with it compared to 4 and 5 star open data. For these reasons, NSS chose to release NHS Scotland Open Data as 3 star open data.

The NSS Open Data Working Group then had to decide upon the platform to recommend, namely should open data be released as licensed tables on the existing ISD website or should a separate portal be used. The decision was taken to use CKAN, which is open source, free software for producing open data websites. CKAN stands for Comprehensive Knowledge Archive Network and users include the European Data Portal, the US, Canadian and Australian Governments, and many cities worldwide [18] [19].

In September 2017, the ''Prescriptions in the Community' dataset was the first to be released on the NHS Scotland Open Data portal [20]. Opinions were sought from a small group of initial users, who were especially interested in this prescriptions data.

Feedback from this soft launch identified a number of key points to address. These included technical features that were required from CKAN and customer requirements on how data should be presented. In addition, three key actions for the open data team were identified. Firstly, it was necessary to challenge the misconception that 'Open Data equates to All Data, including personal information'. This is not the case. Neither the data released as summary data tables on the ISD website nor the NHS Open Data published on the NHS Scotland Open Data Portal contain any patient-level information, so there are no information governance concerns when using data from either source.

Secondly, a cultural shift within NHS Scotland was required to facilitate the move to release more datasets as Open Data. The attitude had to be shifted from 'owning' the data towards enabling access of the data by anyone as the default position.

Thirdly, the need to create and maintain clear standards for Open Data was found to be essential. It was vital with data coming from many different health boards and special boards that the data was released in a consistent format with coding conventions to allow data to be linked easily and with good quality metadata to aid its re-use.

The NHS Scotland Open Data portal is having a positive impact. The Prescribing and Workforce teams have seen a decrease in routine information requests since the release of open data, as customers can often be directed to the portal to find the information they require themselves. New customers are starting to engage with the portal, for example academics and educators, who specifically look for open datasets, and having all the open data sets available on one portal has increased discoverability for customers by giving them a better overview of all data available.

Google Analytics data shows that over 1000 users visited the portal each month during the first quarter of 2019. In March 2019, there were 1377 visitors to the site with 67% of traffic coming from direct or organic searches and 29% of traffic being referred from the ISD website. Users of the portal include (1) The private sector, for example, pharmaceutical companies; (2) Academia using data for research projects; (3) Education sector for teaching purposes; (4) Analysts in the Public Sector such as Audit Scotland and the Office for National Statistics; and (5) Internal customers within the NHS using the data via automated API calls.

As of 1st September 2019, 38 datasets have been published on the NHS Scotland Open Data portal.

## 1.2 Scope and Objectives

The NHS Scotland Open Data Portal has two methods for interacting with data tables available via the portal. Firstly, it is possible to download Open Data tables in a CSV format. Secondly, the

Application Programming Interface (API) for the NHS Scotland Open Data Portal can be used and calls to the API can be integrated into various programming languages.

This dissertation provides an opportunity to engage with the NHS Scotland Open Data Portal in a number of ways.

Firstly, data on the NHS Scotland Open Data Portal will be accessed via the API. This will ensure that any application that is developed will interact with the most up-to-date version of the data and that submitting a query to the data store is possible. This allows the user to select a subset of data if they wish, rather than having to download the dataset in its entirety.

Secondly, the programming language R and the open source R package, Shiny [21] will be evaluated as an alternative to Tableau for creating interactive data visualisations. The NHS Scotland Open Data portal contains a group of datasets related to infants and children. These datasets will be used to create a data story with visualisations about Child Health in Scotland. The open datasets that will be used in this dissertation are:

- '**Births in Scottish Hospitals**' - last updated on 27th November 2018
- '**Infant Feeding**' - last updated on 30th October 2018
- '**Primary 1 Body Mass Index Statistics**' - last updated on 27th November 2018

The aim will be to create an output that highlights the key messages contained within the data, makes sense of the data for the end user and adds value to it by demonstrating that it is possible to link data from different datasets and visualise the outcomes to gain insights. Similarities and differences between the data tables available on the ISD website and the open data available via the portal will be reviewed, and any limitations of R and Shiny versus Tableau will also be discussed.

Thirdly, ISD receives very little feedback on the ease of use of the NHS Scotland Open Data portal or how the data is being used. The 'User Experience' undergone during this dissertation will be discussed in order to high-light the ease or difficulty of use of the open data portal. Recommendations will be made to encourage better engagement with the data, to create more value from the data and to suggest any improvements that might be made.

## 1.3 Achievements

In this dissertation, learning how to access the data using the API was required. This method of accessing the data allows for querying of the data, so that a subset can be downloaded rather than the entire dataset. This is particularly useful when dealing with large datasets and when considering that data products of the future must be mobile-friendly and therefore must allow the user to limit the amount of data downloaded. This objective was successfully achieved and a set of scripts have been written to query data in the 'Child Health' group of open datasets.

R and ggplot2 were evaluated for the creation of data visualisations and a range of visualisations were produced illustrating the effect of factors such as trends over time, variation by Health Board, SIMD quintile and maternal age using three different datasets: 'Births in Scottish Hospitals', 'Infant Feeding' and 'P1 Body Mass Index (BMI) Statistics'. A specification was written for a Shiny application about Child Health and a prototype product was produced.

It was successfully demonstrated that data from two different open data sets: 'Births in Scottish Hospitals' and 'P1 Body Mass Index (BMI) Statistics' can be joined together and that informative visualisations can be created from the open data, as well as from the patient-level data used by ISD. Finally, feedback on the 'User Experience' of working with the NHS Scotland Open Data portal was provided.

## 1.4 Overview of Dissertation

The subsequent chapters of this dissertation cover the following topics:

- Chapter 2 provides an overview of why data visualisation is important and some examples of data visualisation tools and software used with public health data.
- Chapter 3 gives an overview of the NHS Scotland Open Data portal, an outline of the datasets relating to Child Health that were used in this dissertation, the learning about R and Shiny that was required and an explanation of how the API was used.
- Chapter 4 explains how the API was used to access the data available on the NHS Scotland Open Data portal and the User Experience of working with the portal is discussed. Results of this dissertation are presented.
- Chapter 5 comprises a discussion of the results and suggestions for further work.

# 2 State-of-The-Art

## 2.1 Why we should use data visualisations

*'The world cannot be understood without numbers.*
*But the world cannot be understood with numbers alone'*
- Hans Rosling [22]

Data in today's society is generated in a whole myriad of ways ranging from video footage posted on YouTube, photographs on Facebook and Instagram, text in documents, e-mails and on Twitter, information from internet transactions and production processes, and sensors used for monitoring in cars, aircraft and personal wearable devices. The term 'Big Data' is widely used and this term defines data as having four dimensions: Volume – the magnitude of data being produced; Velocity – the speed at which data is being generated by streaming processes; Variety – the diversity of data types; and Veracity – the quality and accuracy of the data. More recently, Value has been mentioned as a fifth dimension and this refers to the value that can be obtained by gaining better insights through analysis [23] [24]. Data is being generated in such sheer volumes and at such a rapid rate that analysis and gaining insights and understanding in order to make informed decisions is often only possible with some kind of visual way of reviewing the data.

Berinato [25] discusses how to make visualisations that really work by the consideration of two key questions. Firstly, what is the information that you have – is it conceptual or is it data-driven? And secondly, do you want to communicate the information or explore it to find out more about its patterns, trends and anomalies? He states that:

*'Visualization is merely a process. What we actually do when we make a good chart is get at some truth and move people to feel it—to see what couldn't be seen before. To change minds. To cause action.'*

He discusses the use of four types of visualisation depending on the aim. For 'Idea Illustration' the focus should be on simple, clear communication of the logic and structure of the idea. 'Idea Generation' is explained as being like a 'brain-storming' of different visual approaches. 'Visual Discovery' can be either a visual confirmation of an answer to a question or a visual exploration of the data to look for patterns. The fourth type, 'Everyday Dataviz' is usually a simple visualisation, such as a bar chart or scatter graph that communicates a key point.

Grinstein at al [26] discuss a wide range of data visualisation types from line graphs, scatter plots and heat maps to parallel co-ordinates, polar charts and principal component analysis. The main focus of this piece of work is on illustrating which visualisations are most effective for high-dimensional data. The authors concluded that visualisations such as Pixel Displays, RadViz and PolyViz work well when a high number of dimensions need to be displayed. However, they also commented on screen resolution and colour perception having an impact on the effectiveness of a visualisation with multiple dimensions. They stated that in some cases it is better to have a series of linked visualisations or offer some level of interaction with the visualisation.

Berinato [25] also mentions that exploration of data lends itself to an interactive process. Sedig at al [27] describe interaction with data as being a complex phenomenon. The area of Public Health information is discussed and it is suggested that tool designers should consider user activity as consisting of several levels of interaction with the data: Complex cognitive activities; Tasks involving location and categorisation; Filtering and transforming the data; and User-interface events like a click or a swipe.

Ola and Sedig [28] discuss how many visualisation tools limit users' interaction with the data by only allowing them to perform simple manipulations. They describe visualisation tools as having three aspects: cognitive tasks such as generating a hypothesis, visual tasks that are carried out by the visuoperceptual system when the user looks at a data visualisation and interactive tasks where the user manipulates the data. In a user study with public health data, they demonstrate how interactive visualisations can play a role in assisting with data-orientated health tasks. The term 'Interactivity' is used to refer to the quality of the user's interaction with the data visualisation tool, i.e. whether the user can adjust properties within the tool's interface, explore relationships between attributes of their choice and look for links between different data. Ola and Sedig suggest that users need to be able to interact fully with public health data in order to gain maximum insight from the data but that designing interaction is a non-trivial issue. They state that there is a need for conceptual structures to assist with the design process of new data visualisation tools.

## 2.2 Use of Public Health data

Ola and Sedig [28] discuss how, in the past, the health community has been slow to use data to maximise insights and they believe this is partly due to the complex nature of the data. Aung et al [29] carried out a study of data literacy and statistical skills with junior, mid-level and senior decision makers working for the Taiwanese Government. These employees were responsible for reproductive, maternal, new born, child health and nutrition policy decisions. It was found that the majority of these staff had not had any data literacy or visualisation training since leaving university and it was suggested that basic data usage training for policy makers at all levels would be beneficial.

Ola and Sedig [30] describe two types of tools that are used for evaluating public health data: data analytics tools and interactive visualisation tools. They use the term 'Visual Analytics' tools to define a new class of tool that combines the strengths of both. Applications of visual analytic tools include the early detection and monitoring of epidemics, the exploration of complex interactions of different health policy options in order to aid policy-making decisions, and health assurance by making the public aware of preventative measures and ensuring access to and effectiveness of health care services. Commercial visual analytics tools such as Tableau [9] and Spotfire [31] are referred to as a means of creating interactive dashboards that are useful for tools for public health stakeholders and provide a means of managing and dispensing health resources efficiently.

In a more recent paper, Ola and Sedig [32] continue to discuss how interactive data visualisations can play a key role in utilising Public Health data gathered from hospitals, sensors, and social media in a variety of formats. They discuss how many health visualisation tools use simple charts that only illustrate one or two attributes of the data. They believe that there is a need for more sophisticated data visualisations that allow the user to distinguish patterns and develop insights by encoding many data attributes simultaneously. The idea of conceptual frameworks is mooted to help tool designers create more advanced visualisation tools with interactivity as part of the design thinking. Examples of complex visualisations are shown using global data concerning causes and risk factors that are associated with mortality. Visualisations are shown that display data by geography, chronology and demography.

In 2018, a Whitepaper entitled: 'Making NHS data work for everyone' was published by Harwich and Lasko-Skinner from Reform, a leading Westminster think tank for public service reform [33]. This paper focuses on the use of data from NHS England and NHS Trusts by private sector companies for development of commercial products and services. Harwich and Lasko-Skinner explain that:

*'Everyone and everything within the NHS generate data daily,*
*from patients to doctors and nurses, MRI scanners to appointment booking systems'*

The primary purpose of this data is direct patient care and the data is used when making medical decisions about the type of care a patient should receive. Secondary use of the data requires permission and could include analysis, generation of statistics, machine learning, and informing the development of new products and services.

Harwich and Lasko-Skinner suggest that NHS England and NHS Trusts should have a national strategy for value exchange between the NHS, patients and industry, and for commercial agreements. They believe that a clear understanding by the NHS of their value proposition, i.e. the benefit the NHS provides, to whom and the unique way in which this is done, is central to any mutually beneficial value exchange. The value of healthcare data is shown in Figure 2. They go on to discuss how the Department of Health and Social Care, Caldicott Guardians, the NHS and industry should ensure that there is a dialogue with the public to discuss different commercial models such as Data sharing agreements, Grant-funded collaboration, Licencing Cost-Recovery, and Commercial arrangements.

The paper then goes on to discuss concerns about ease of access of NHS data and describes the data as being held in 'silos of healthcare activity' with no easy way to communicate between different departments. The use of Open Data during the early stages of a commercial agreement is discussed and NHS Digital [34] has a number of open datasets that are available via their website and via the UK government's portal [35]. An Open Access model, where the results of the research would be made available is debated but is thought to be of limited use, as private sector companies are likely to want any findings to be proprietary. A recommendation is made that NHS organisations should offer synthetic datasets, which could be shared with private sector companies for research or for product and service development at the early stages of a project.

## 2.3 Data Visualisation Tools being used with Public Health Data

### 2.3.1 Tableau

In Section 1.1, the current use of Tableau by NSS and ISD for the production of data visualisations and infographics was discussed. In a Whitepaper entitled: 'Data Storytelling: Using visualisation to share the human impact of numbers', Kosara and Wallace discuss the process of storytelling and how the Tableau software can aid this process [36].

Kosara and Wallace explain how stories build connections and context around facts thus making them more memorable. They describe how stories do not just have '*a beginning, a middle and an end*' but how a good story also has a '*story arc'* consisting of increasing action or conflict presented in a logical order and leading to a conclusion.

In Section 2.1, the importance of user interaction with a data visualisation in order to gain more value from the data was discussed. Tableau offers not only the ability to create an interactive dashboard to monitor, explore and analyse the data but also the Story Points feature that provides a framework for arranging a set of interactive data visualisations with annotations in a sequence in order to tell the reader a story. This is believed to be important as stories can motivate action. Kosara and Wallace state that:

*'Dashboards tell you what's happening, but stories explain why.'*

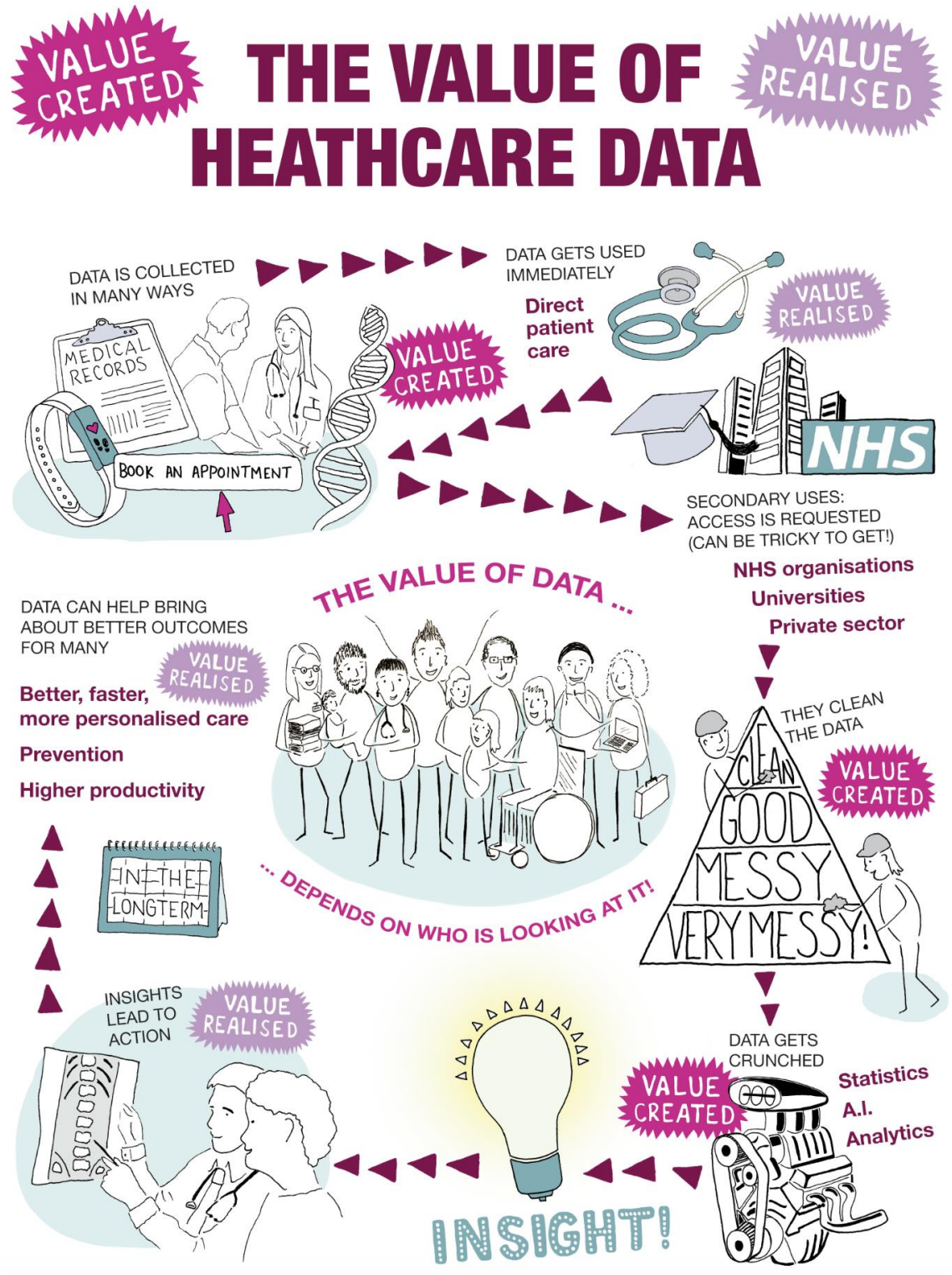


**Figure 2. The Value of NHS data**
(Source: 'Making NHS data work for everyone', Whitepaper by Reform, London)

Tableau offers a range of products including Tableau Desktop that can be connected to a data source and used for analytics and visualisation, Tableau Server for publishing dashboards and sharing them using a web-based server and Tableau Online, which is a hosted version of Tableau Server. Additional services include Tableau Prep for combining, shaping and cleaning

data, Tableau Data Management for managing data in a Tableau environment and Tableau Mobile, which can connect to an organisation's Tableau Server or Tableau Online site. Tableau also offers a variety of pricing packages for individuals and teams with different access levels available. A price of $70 per user per month is quoted on the website for Tableau Desktop, Tableau Prep Builder, and one Creator license of Tableau Server.

Tableau is widely used for analytics and visualisation within ISD and NSS. The NHS pay for a number of licences and these are accessed by the teams who use Tableau for generating interactive visualisations. A sum of £350k would be a fair estimate for the total investment in Tableau over the last few years (see Appendix 1), so potential cost-reduction is a key driver in the move to look at alternative software and tools for data analysis and generation of visualisations.

### 2.3.2 D3.js

D3.js is an open source JavaScript library that can be used to connect data to a Document Object Model (DOM) and then apply transformations to the document. For example, an HTML table can be created from an array of numbers. D3 stands for Data-Driven Documents and the D3.js website [40] and Smith [41] discusses its key features.

D3.js works with existing web standards such as HTML (Hypertext Markup Language), SVG (Scalable Vector Graphics) and CSS (Cascading Style Sheets) and can be easily implemented across platforms without the need for proprietary software or plug-ins. D3.js is data-driven. It works with static data or data fetched via an API from a remote server and it is compatible with a variety of data formats including CSV (Comma-Separated Values), JSON (JavaScript Object Notation) and XML (eXtensible Markup Language). D3.js is also very flexible and combines a wide range of visualisation components with a data-driven method to manipulating document object models. It assists the user in the creation of any type of visualisation, whilst still allowing the user to control visualisation features.

Gratzl et al [37] discuss the idea of Vistories, where a visual story is produced that is based on the history of the data exploration process. D3.js is used to create the visualisations. They describe a model they call CLUE (Capture, Label, Understand and Explain) that they use to explore public health data. The user can switch between exploration, authoring and presentation modes and create a Vistory of their chosen path through the exploration process. The term provenance graph is used to describe this history of the exploration process and the user can enrich their vistory with annotations and high-light key points. The integration of the provenance graph into the presentation allows the user to switch back to exploration mode at any point.

Hadjar et al [38] investigate the use of virtual reality (VR) to aid understanding of multi-dimensional health data collected from open data portals. D3.js is one of the libraries they use with their Web Virtual Reality based system. Reyes and Labelle [39] wrote a summary article about tools, practices and transformations. They mention the D3.js implementation of the Web VR framework of Hadjar et al [38].

### 2.3.3 Python

Python is an object-oriented, general-purpose programming language. It is often referred to as a scripting language, as it was originally intended to be used for small projects but as the popularity of Python has grown, so has its use. Many platforms and web applications now use Python including Google, YouTube and the New York Stock Exchange [42].

Python has many libraries available for creating graphics. Firstly, there is Matplotlib, which is a low level library that is good for creating simple charts such as line charts, bar charts, histograms and scatter plots. Secondly, there is Pandas Visualisation, which is a good tool for

creating graphics when working with a Pandas data frame. It can be used to create line charts, bar charts, histograms and scatter plots just like Matplotlib. Then there is Seaborn, which can also be used with a Pandas data frame. It has a higher-level interface, which provides extra features such as the option to overlay different types of plot and to use colour to highlight different classes. In addition, it is easier to use Seaborn than Matplotlib when using features such as annotations and faceting, and to produce more detailed visualisations such as box plots, heat maps and pairs plots [43].

Animated visualisations are also possible with Matplotlib and Celluloid. Both involve plotting a number of graphs and then creating a GIF (Graphics Interchange Format) file using these graphs to create the animation [44].

Interactive visualisations can be generated with the plotly package, a library built on plotly.js, which is built on D3.js. A wrapper called cufflinks, which is designed to work with Pandas data frames can be used on plotly. This combines the ability to code in Python with the graphics capability of D3.js. Plotly is a graphics company and the plotly library for Python is one of the open-source tools it offers to use in either off-line or on-line mode. Using plotly and cufflinks, it is possible to make interactive histograms and box plots, to overlay histograms, to plot an interactive time series of a scatter plot, to introduce colour to differentiate classes, and to generate a scatterplot matrix, annotated heat maps and 3D graphs. Bokeh is another library available in Python for making interactive graphs, which are constructed by stacking layers on top of one another. A figure is created and then elements called glyphs are added. Both plotly and Bokeh can be used to create dashboards [44] [45].

### 2.3.4 R

R is a software language and environment that can be used for data manipulation, statistical calculations and graphics where the term ‘environment’ is used to describe a planned and coherent system. RStudio is an integrated development environment (IDE) for R. The Open Source edition of RStudio Desktop is free and there are a variety of other options and add-ons with different price points. These include RStudio Server, RStudio Server Pro, RStudio Connect and RStudio Package Manager. The RStudio Team bundle comprises RStudio Server Pro for analysing data and creating data products, RStudio Package Manager for ensuring that all users in the organisation have access to the same R packages and package versions, and RStudio Connect for publishing data products. This bundle offers businesses and organisations a more secure environment, an agreed number of licences, access to a dedicated server and more computing power for large pieces of work and the option of running multiple sessions for one user when analytical jobs need to be run in parallel[46].

R can very easily be extended by packages and one popular package is ggplot2, which was created by Hadley Wickham for data visualisation [10] . This package is based upon the ‘Grammar of Graphics’, which is the idea that any graph can be constructed from the same components: a dataset, a co-ordinate system and a set of visual markers that are called geoms in ggplot2 [47].

Blevins et al [48] use R to present an analysis of patient-level data. CD4 is a glycoprotein that is found on the surface of some cells in the immune system. ‘CD4+ T-helper’ cells are a type of white blood cell that expresses the protein and they form part of the human immune system [49]. Blevins et al looked at trends in CD4 cell count and AIDS when antiretroviral therapy is started, CD4 cell curves after therapy initiation, and mortality. R is used to generate three different types of plots. Longitudinal plots are used to examine changes in measurements versus outcomes. Bubble plots illustrate changes in key indicators across various groups over time. Heat maps provide spatial and temporal data for the group of patients. There is mention

of including a graphical user interface (GUI) in future work that could display statistics and provide some level of interaction.

Chein et al [50] used the National Health Insurance Research Database for their analysis. The database was constructed by the Taiwan National Health Insurance Administration, Ministry of Health and Welfare and it contains data about the health of all individuals. They discuss a visualisation model for analysing disease trajectory. It is proposed that the results could be used to predict the risk of certain diseases, inform national health policy and assist with health education programmes. Chien et al analysed data such as medication, surgery, length of stay, cost and the number of patients who suffered from another disease during the subsequent two years. The Shiny package was used to provide an input panel for the user to add information such as gender, age, the date and the disease and an output panel showing a Sankey diagram of the disease trajectory.

As already mentioned in Section 2.3.2, Hadjar et al [38] used virtual reality to look at multi-dimensional health data from open data portals. As well as D3.js, they also used R, Shiny and ggplot2 and found ggplot2 to be fast at plotting large networks when looking at network visualisation. Shiny also offered the benefit of being able to deploy the Shiny application on the Shinyapps.io server [51]. In their summary article, Reyes and Labelle [39] also mention the Shiny and ggplot2 implementation of the Web VR framework of Hadjar et al [38].

Konkoľová and Paralič [52] discuss the use of R, RStudio and Shiny to support active learning of data science during the module Knowledge Discovery (Data Mining) at the Technical University in Košice.

## 2.4 Data Visualisation in NHS Scotland

In Section 2.3.1, it was mentioned that Tableau has been the software tool of choice for ISD and NSS for the last five years (see Appendix 1). Recently though, there has been a move towards the use of R, and the R package Shiny has been used to produce some visualisations and storyboards.

The Acute Hospital Activity & NHS Beds Data Release dated 26th June 2018 contains charts which have been produced using R and a JavaScript library D3 [53].

The Data Lab blog contains an article about the Accelerator programme run by The Data Lab during 2018 in which Constantinescu of the Data Lab describes how she mentored Thrower, Senior Analyst from ISD and supported her in learning R and Shiny in order to create visualisations about the Burden of disease in Scotland [54].

The Drug-Related Hospital Statistics, Drug and Alcohol Misuse visualisation based on the data release dated 28th May 2019 [55] is available on the shinyapps.io website, as is the Prescribing and Medicines, National Therapeutic Indicators visualisation based on the data release dated 16th July 2019 [56].

The Drug-Related Hospital Statistics, Drug and Alcohol Misuse visualisation offers different views of the data including time trend data where two geographical locations or two drug types can be compared, plus time trend data by level of deprivation or age and gender. Interaction with the data is available via selections made from drop-down menus [55].

The National Therapeutic Indicators visualisation uses prescription data to compare prescribing activity across NHS Boards, Health and Social Care Partnerships (HSCPs), GP clusters and GP practices in specified therapeutic areas. Box plots are available to look at a specific point in time; a line chart is used to look at trends over time; and time trend data can be examined simultaneously across three points in time with pop-up annotations to provide details [56].

It is believed that these Shiny visualisations are generated from data tables that are specifically generated and formatted to feed the visualisations. The data tables sit on a server and can be downloaded through the same interface in the data explorer.

# 3 Methodology

## 3.1 The NHS Scotland Open Data portal

In September 2017, the first dataset was published on the NHS Scotland Open Data portal (see Section 1.1). As of 1st September 2019, nearly two years later, the number of datasets available has risen to 38 according to the landing page of the portal which is shown in Figure 3 .

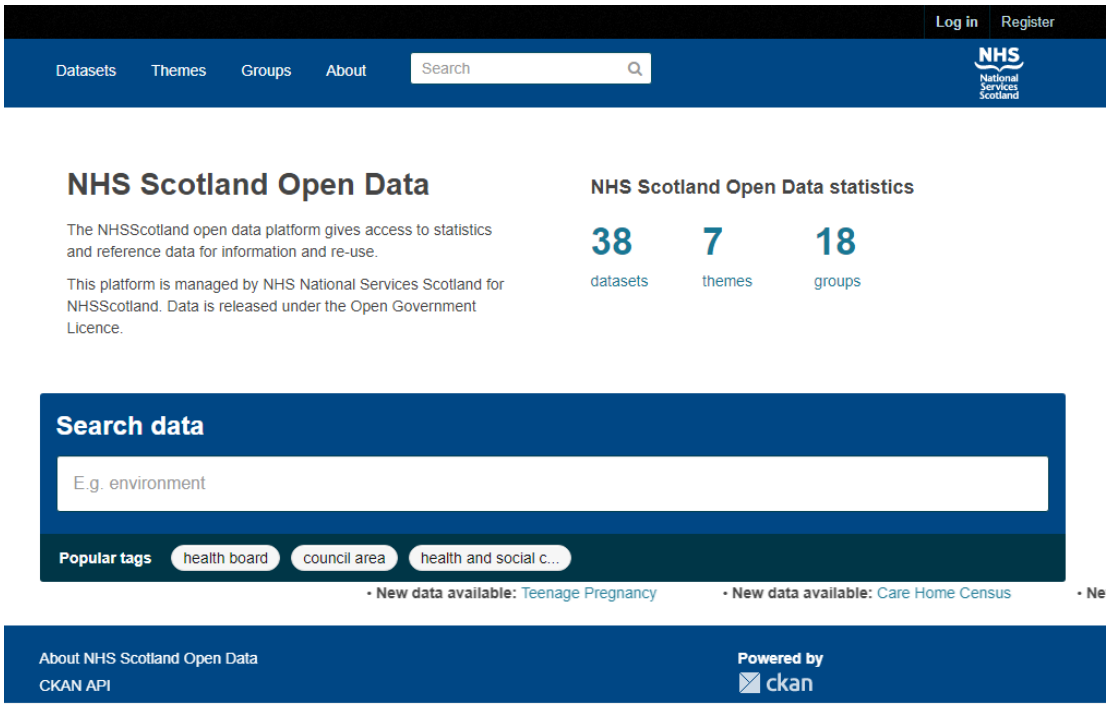


**Figure 3. The NHS Scotland Open Data portal**

Data on the portal can be explored in different ways. For example, the full list of datasets can be viewed by clicking on 'Datasets' on the Home page. Alternatively, data on the portal can be explored by group including topics such as Child Health, Waiting Times and Mental Health. The page displaying the group of Child Health datasets is shown in Figure 4.

The page for a particular dataset can be selected from the list and this page contains a description of the dataset giving background information, the metadata and a list of available resources. A resource is the CKAN term for a data file. The resources for the 'Infant Feeding' dataset are shown in Figure 5.

Each resource has a preview window to show the user the data's attributes and underneath the preview window is a Data Dictionary that lists the attributes and provides links to lookup tables for any codes that are used within the resource and a table containing additional information specific to the data resource. The preview window for the Infant Feeding by Maternal Age resource is shown in Figure 6.

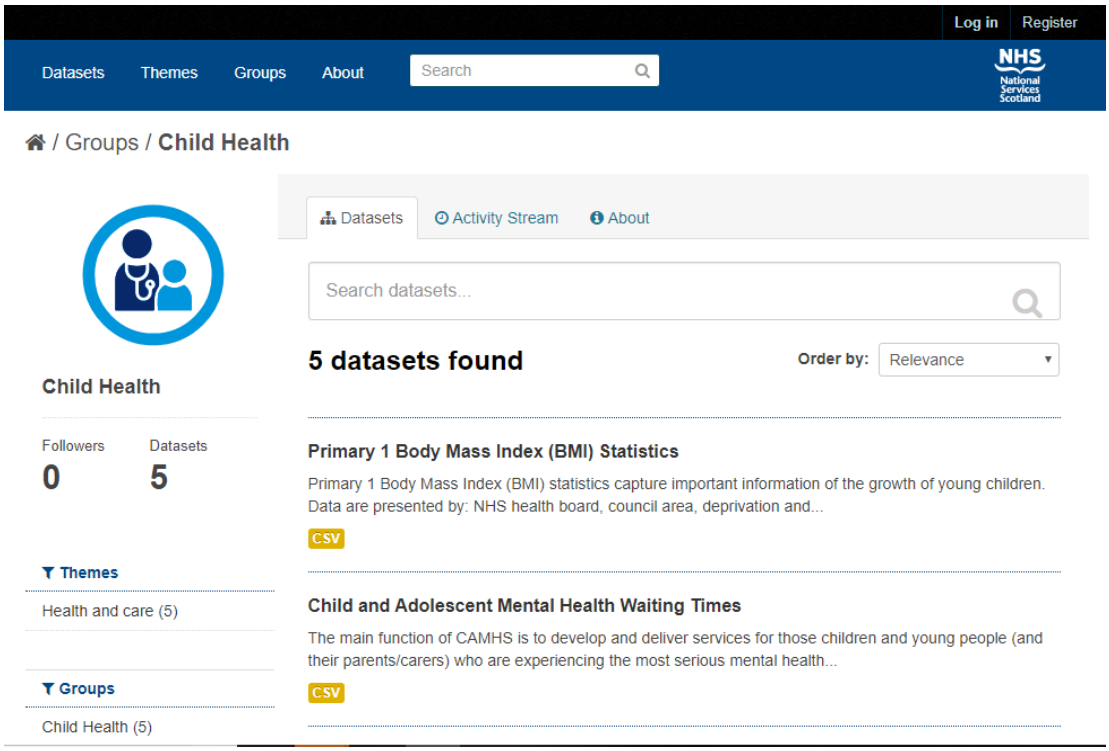


**Figure 4. The Group of Child Health datasets**

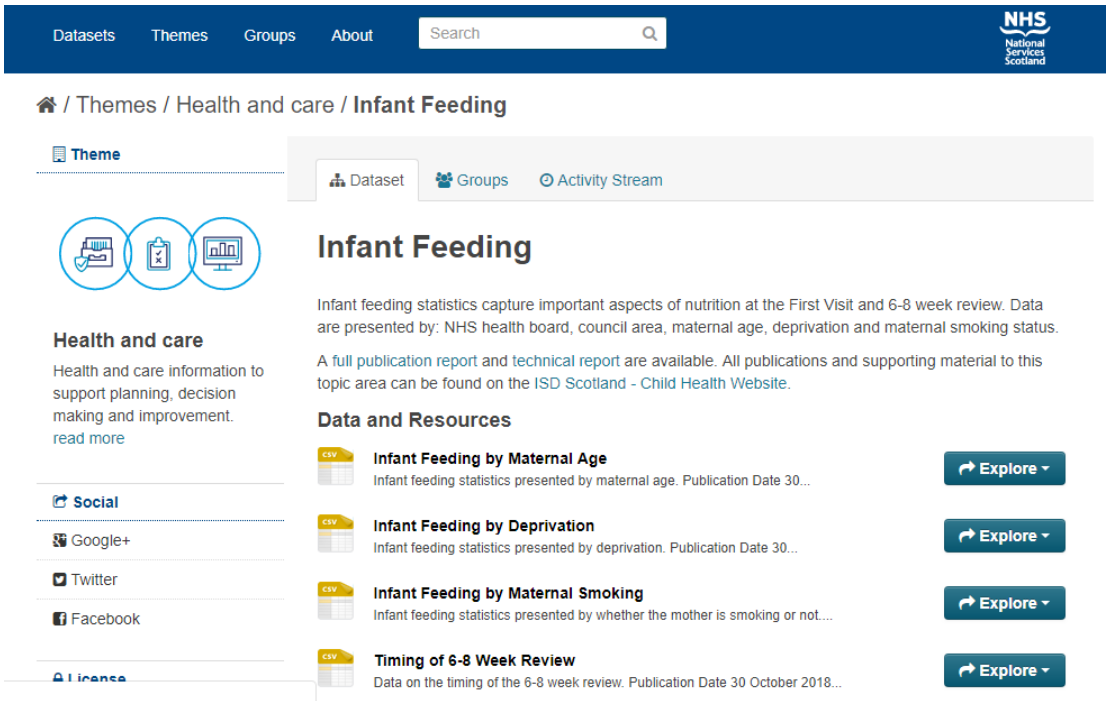


**Figure 5. The Resources for the 'Infant Feeding' dataset**

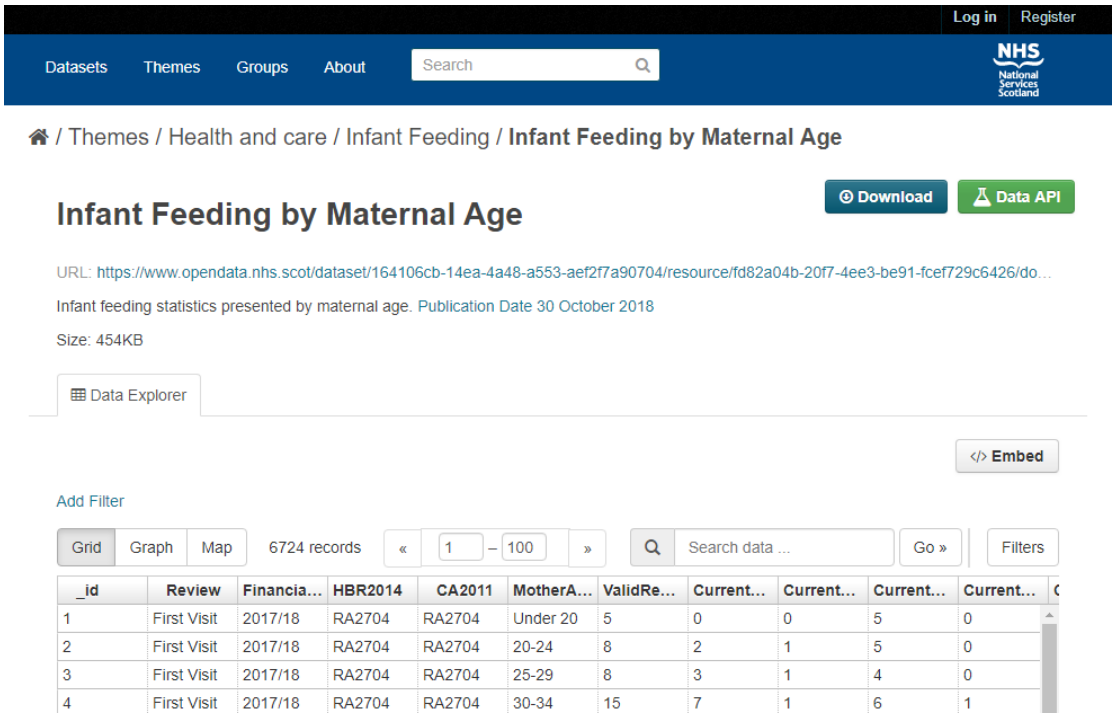


**Figure 6. The preview window for the Infant Feeding by Maternal Age resource**

## 3.2 Datasets about Child Health

The group of datasets relating to Child Health is shown in Figure 4 and the datasets used in this dissertation are:

- '**Births in Scottish Hospitals**' - last updated on 27th November 2018
- '**Infant Feeding**' - last updated on 30th October 2018
- '**Primary 1 Body Mass Index Statistics**' - last updated on 27th November 2018

A publication report, a technical report and a data visualisation produced in Tableau are available for each of the datasets listed above, apart from 'Births in Scottish Hospitals' for which no visualisation is available. The official statistics used in these reports and visualisations are generated by ISD from patient-level data. However, the information has been aggregated and summarised in such a way that it is not possible to identify any individual patient in the reports.

Data tables for these datasets are available on both the ISD website and as NHS Open Data on the NHS Scotland Open Data Portal. Data and visualisations are assessed for statistical disclosure before they are released [57]. This means that some numbers are rounded or suppressed to preserve confidentiality. For example, aggregated figures might be used for areas of low population density to prevent identification of an individual child. Data tables do not contain any information about individual patients and so there are no information governance concerns when using data tables from either source.

Links to the PDF copies of the publication reports and the Tableau visualisations for the four datasets used can be found on the Publications page of the ISD website after selecting 'Maternity and Births' and 'Child Health' from the drop-down list of Health Topics[58]. Links to the technical reports can be found towards the end of each publication report. Links to each publication reports technical report and Tableau visualisation are also available:

- '**Births in Scottish Hospitals**' - [59][60]
- '**Infant Feeding**' - [61][62][63]
- '**Primary 1 Body Mass Index Statistics**' - [64][65][66]

It should be noted that the publication and technical reports are all classified as National Statistics publications for Scotland. This means that they only contain statistics relating to the financial year that is the subject of the report and any trends observed compared to the last few years. No weighting or emphasis is allowed to be placed upon any particular findings. Possible relationships between, for example, maternal health during pregnancy and child health during the first few years of life cannot be explored. Opinions cannot be expressed nor conclusions drawn by the analyst writing the report. These reports are required to be strictly factual and to report the statistics and nothing further [67].

This dissertation however is not bound by these constraints and possible associations between the datasets will be explored.

## 3.3 Learning about R and Shiny that was required

It was decided to use the programming language R and the open source R packages, ggplot2 and Shiny [21] for creation of data visualisations for this dissertation. Key considerations in the decision to use R and Shiny were the free availability of the shinyapps.io [51] website for possible publication of an application in the future and a shift towards using R within the NHS.

It was necessary to undertake a course of study on R and Shiny before commencing any analysis and the following on-line courses were studied on the DataCamp website:

- Introduction to R [68]
- Importing Data in R (Part 1) [69]
- Importing Data in R (Part 2) [70]
- Working with the RStudio IDE (Part 1) [71]
- Working with the RStudio IDE (Part 2) [72]
- Intermediate R [73]
- Data Visualization with ggplot2 (Part 1) [74]

A tutorial about Shiny available on the RStudio website was also used:

- How to Start Shiny tutorial [75]

## 3.4 Using the API

### 3.4.1 Submitting a query to the API

In Section 1.1, it was mentioned that there are two methods for accessing data available via the portal. Firstly, a resource can be downloaded in a CSV format by using the URL underneath the title on the page for each individual resource or by using the green download button (see Figure 6). Alternatively, the API for the NHS Scotland Open Data Portal can be used. Use of the API allows interaction with the website and the data in the data store, and interrogation of the data. This means that the user can select what they wish to download, rather than always having to down-load the whole dataset.

## 3.5 Development of a Shiny App

The website about Shiny from RStudio describes Shiny as being an R package that can be used to build interactive web applications [21] and the shinyapps.io website provides a means of deploying a Shiny application. The free version of a shinyapps.io account can be used to deploy up to five applications for up to 25 active hours per month [51]. The website provides both video and written tutorials that give an introduction to building a Shiny application [75].

The tutorial explains that a Shiny application consists of a single script called app.R that has its own directory. This script consists of three components: a user interface object, a server function and a call to the shinyApp function. In some cases, the directory may also contain a second script called helpers.R and a folder containing any images or data required by the application. Older versions of Shiny did not support single file applications, so it used to be the case that separate scripts were required for the user interface object (ui.R) and the server function (server.R). Currently, both single file applications and those with separate ui.R and server.R scripts are supported.

The basic structure of the code for a single script Shiny application is shown in Figure 7 and this script can be opened in RStudio and run by clicking 'Run App' that appears with a green arrow in the top left hand corner of the console. The drop-down menu beside 'Run App' gives the options to run the application in the viewer pane of RStudio (bottom right) or in a pop up window. When opened in a pop-up window, the additional option to run the app in a new browser tab is available.

```
1  #Load the Shiny package
2  library(shiny)
3
4  #Define the user interface
5  ui <- fluidPage("Hello")
6
7  #Define the server logic
8  server <- function(input, output){}
9
10 #Make a call to shinyApp
11 shinyApp(ui = ui, server = server)
```

**Figure 7. Code snippet showing the structure of a Shiny application**

# 4 Results

## 4.1 Using the API for the NHS Scotland Open Data portal

### 4.1.1 Obtaining all the resource ID numbers for a package

The first aim of this dissertation was to access the data on the NHS Scotland Open Data Portal via the API. The blog post entitled: 'Accessing APIs from R (and a little R programming)' [1] provides an excellent introduction to using an API as well as a worked example of how to use the API for the EU's EUR-Lex database. The worked example from this blog post was replicated in R and then adapted to produce an initial script that would work with the NHS Scotland Open Data portal.

An example of a call to the API to get the resource ID's for the 'Infant Feeding' package is shown in the code snippet in Figure 8. The term 'package' is used by CKAN to refer to a 'dataset'.

In this initial script, the httr and jsonlite R packages were used to assist with calls to the API and converting JSON objects into R objects. After loading these packages, the next step was to add the web address for the NHS Scotland Open Data portal and to create the path to be used, which includes the call to the API and the package name (see lines 11-12 in Figure 8).

```
1  #Get resource id's for the Infant Feeding package
2
3  #Call packages
4  library(httr) #Version 1.4.0
5  library(jsonlite) #Version 1.6
6  library(ggplot2) #Version 3.1.1
7  library(tidyverse) #Version 1.2.1
8  library(reshape2) #Version 1.4.3
9
10 #First call to API to get all resource ID's for the infant
   feeding package
11 url  <- "https://www.opendata.nhs.scot"
12 path <- "/api/3/action/package_show?id=infant-feeding"
13 raw.result <- GET(url = url, path = path)
14
15 #A result of 200 means the server has received the request
16 raw.result$status_code
17
18 #Translate it into text
19 this.raw.content <- rawToChar(raw.result$content)
20
21 #Parse character string containing JSON file into something R can
   work with
22 this.content <- fromJSON(this.raw.content)
23
24 #Create a variable to store the list of resource id's for the
   package
25 if_id_list <- this.content[3]$result$resources$id
26
27 #Print the list of resource id's
28 if_id_list
```

**Figure 8. Code snippet for an API call to get resource ID's for the 'Infant Feeding' package**

If the call to the API has been successful, the status code will return a result of 200, so it is useful to check this (see line 16 in Figure 8). The response to a call to the CKAN API for 'package_show' always has the same format and this is a JSON dictionary with three keys: 'help, 'success' and 'result'. The list of resource ID's for a package can be accessed via index numbers and the $ nomenclature as shown in line 25 in Figure 8.

### 4.1.2 Using multiple calls to the API

In subsequent scripts, multiple calls to the API were used to obtain: (1) the ID numbers for the resources available for a package, (2) the number of records for a particular query, and then (3) the records themselves. The code snippet shown in Figure 9 is a continuation of the script shown in Figure 8.

```
29 #This is the id for Infant Feeding by Maternal Age
30 if_age <- if_id_list[1]
31
32 #Second call to API to using id number and limit=1 to get total
   number of records
33 url  <- "https://www.opendata.nhs.scot"
34 path <- file.path("/api/3/action/datastore_search?id=", if_age,
   "&limit=1", fsep="")
35 raw.result2 <- GET(url = url, path = path)
36
37 #Translates it into text
38 this.raw.content2 <- rawToChar(raw.result2$content)
39
40 #Parse character string containing JSON file into something R can
   work with
41 this.content2 <- fromJSON(this.raw.content2)
42
43 #The third element in the list contains the data and notes
44 #and the $total field contains the number of records
45 if_age_total <- this.content2[[3]]$total
46 if_age_total
47
48 #Third call to API using id and total number of records
49 url  <- "https://www.opendata.nhs.scot"
50
51 #Uses calls 1 and 2 to API
52 path <- file.path("/api/3/action/datastore_search?id=", if_age,
   "&limit=", if_age_total, fsep="")
53
54 raw.result3 <- GET(url = url, path = path)
55
56 #Translates it into text
57 this.raw.content3 <- rawToChar(raw.result3$content)
58
59 #Parse character string containing JSON file into something R can
   work with
60 this.content3 <- fromJSON(this.raw.content3)
61
62 #The third element contains the data and notes
63 #The $records field contains the individual records
64 this.content3.df <- data.frame(this.content3[[3]]$records)
```

**Figure 9. Code snippet to get the number of records for the resource and the records themselves for the Infant Feeding by Maternal Age resource**

Once a particular resource has been selected - Infant Feeding by Maternal Age in the example shown in Figure 9 - the path for the second call to the API needs to be created. In line 34 of Figure 9, it can be seen that 'datastore_search' is now used and the ID is set by using either the resource ID or in this example, the variable in which the resource ID obtained from the first call to the API has been stored. The ID for an individual resource can also be found on the NHS Scotland Open Data portal by scrolling to the bottom of the web page for an individual resource and looking for 'id' in the Additional Information table.

The second call to the API is used to find the total number of records for the chosen resource or if a filter is used, the total number of records that match the specific query. This is required because the default setting for the number of records that will be downloaded is 100 and many datasets and responses to queries will be larger than this. The response to a call that includes 'datastore_search' always has the same format and this is a JSON dictionary with three keys: 'help', 'success', and 'result'. The total number of records matching the API call can be accessed via index numbers and the $ nomenclature as shown in line 45 in Figure 9 and this total is stored as a variable for use in the third call to the API.

The path for the third call to the API to obtain the records themselves uses the variables created from the first two API calls (see line 52 in Figure 9). The records are accessed via index numbers and the $ nomenclature and converted into a data frame as shown in line 64 of Figure 9. Information about a resource such as the data dictionary can be accessed via $result, as well as via the web page for a resource on the NHS Scotland Open Data portal.

### 4.1.3 Defining the fields or rows of a resource to be returned by the API call

The API calls shown in Figure 8 and Figure 9 return all records for the Infant Feeding by Maternal Age resource. However, it may be the case that only some fields are required for a piece of analysis or perhaps, only records for a certain year or health board are needed. In these cases, filters can be built into the API call, so that only the records of interest are downloaded. This is particularly useful when the dataset being used is very large and downloading the whole dataset would be time-consuming or in some cases not possible. Some examples of filters that can be added to API calls are shown in Table 2.

**Table 2. Examples of filters that can be added to API calls**

| Filter | Result |
|---|---|
| & limit=1 | Limits the number of records downloaded to one |
| &q=2017/18 | Limits the records to those from the year 2017/18 |
| &fields=MotherAgegroup, EverBreastfed | Limits the response to just these fields |

## 4.2 Analysis and visualisations in R

The second aim of this dissertation was to evaluate the programming language R as an alternative to Tableau for creating data visualisations. Three datasets related to Child Health were studied and visualisations were created using the ggplot2 package.

### 4.2.1 'Births in Scottish Hospitals' by SIMD and Maternal Age

The first set of data to be investigated was the 'Parity' resource of the 'Births in Scottish Hospitals' dataset. The number of maternities in 2017/18 was analysed by the level of deprivation of the maternal area of residence and by maternal age group. The level of deprivation was assessed using the Scottish Index of Multiple Deprivation quintiles (SIMD 1 – 5), where 1 refers to the most deprived areas and 5 refers to the least deprived areas. The SIMD quintile data is updated every three to four years and 'SIMD2016' was the version used with the 2017/18 data in the 'Parity' resource.

Figure 10 shows the percentage of births in Scottish hospitals by SIMD quintile during 2017/18. It can be seen that births are not distributed equally across the five SIMD quintiles. A higher than average percentage of the births occurred in the most deprived areas SIMD 1 (25%) and SIMD 2 (21%) and that there is a downward trend in the percentage of births per quintile when moving from SIMD 1 (25%) to SIMD 5 (17%).

In Figure 11, the data is split by SIMD quintile and Maternal Age Group. In the 'Under 25' age group, a linear trend in the percentage of births can be seen with level of deprivation. The percentage of births decreases from 7% in SIMD 1 to 1% in SIMD 5. A trend in the same direction is seen with the '25 – 34' age group but it is less marked than with the younger age group. The percentage of births decreases from 14% in SIMD 1 to 10% in SIMD 5. By contrast, the trend observed with mothers aged '35 and over' is for an increasing percentage of births in the least deprived areas with 4% in SIMD 1 increasing to 6% in SIMD 5.

Figure 12 shows the same data as Figure 11 but with the data grouped initially by Maternal Age Group and then split by SIMD quintile. The same linear trends observed in Figure 11 can be clearly seen and the distribution of births between the age groups is 18% for 'Under 25', 60% for '25 – 34' and 22% for '35 and over'.

### 4.2.2 'Infant Feeding'

#### 4.2.2.1 Trends over time

The second set of data to be investigated was the 'Infant Feeding' dataset. In Figure 13, the trends in the four types of infant feeding for the whole of Scotland are shown from 2000/01 to 2017/18.

Over this time period, the percentage of infants being exclusively breastfed has remained at a similar level of 39% in 2000/01 to 36% in 2017/18. However, a decrease in the percentage of infants being exclusively formula fed can be seen, together with a corresponding increase in the percentage of infants receiving a mixture of breast milk and formula. Formula feeding has decreased from 57% in 2000/01 to 49% in 2017/18. Mixed feeding has increased from 4% in 2000/2001 to 15% in 2017/18. So, over the time period the percentage of infants receiving some breastmilk, either via being exclusively breastfed or via mixed feeding has increased from 43% in 2000/01 to 51% in 2017/18.

#### 4.2.2.2 Effect of location

The effect of four different factors on infant feeding was studied using the data from 2017/18. The first of these was location, which was investigated by looking at the percentage of infants who had been breastfed at some point in time by Health Board. Figure 14 shows the average

for Scotland as an orange line and it can be seen that the Health Boards with the highest percentages of infants who have been breastfed at some point are NHS Orkney, NHS Lothian and NHS Shetland. The Health Boards with the lowest percentages of infants who have been breastfed at some point are NHS Lanarkshire, NHS Ayrshire and Arran, and NHS Forth Valley.

#### 4.2.2.3 Effect of SIMD

The effect of SIMD quintile on the likelihood of starting and continuing to breastfeed in 2017⁄18 was investigated and the results can be seen in Figure 15. A linear relationship between SIMD quintile and breastfeeding can be seen with the highest rates of breastfeeding being observed in the least deprived areas.

In SIMD 1, 48% of infants are initially breastfed. This falls to 37% by the time of the health visitor's first visit at 10 – 14 days after birth, and then to 28% by the 6 – 8 week review. In SIMD 5, 81% of infants are initially breastfed. This falls to 70% by the time of the health visitor's first visit at 10 – 14 days after birth, and then to 60% by the 6 – 8 week review. Across SIMD quintiles, the drop-off in breastfeeding is slightly higher between initially starting and the 10 – 14 day review (a drop of 11 – 13 percentage points) than between the 10 – 14 day review and the 6 – 8 week review (a drop of 9 – 11 percentage points).

#### 4.2.2.4 Effect of Maternal Age

In Figure 16, the likelihood of starting and continuing to breastfeed in 2017⁄18 is shown by maternal age. Again, a relationship can be between maternal age and breastfeeding with the highest rates of breastfeeding being observed with older mothers. The relationship is linear for the youngest four age bands: 'Under 20', '21 – 24', '25 – 29', and '30 – 34' years and then starts to level out for the oldest two age bands: '35 – 39' and '40 and over' years.

In the 'Under 20' age group, 37% of infants are initially breastfed. This falls to 21% by the time of the health visitor's first visit at 10 – 14 days after birth, and then to 13% by the 6 – 8 week review. In the 'Over 40' age group, 76% of infants are initially breastfed. This falls to 66% by the time of the health visitor's first visit at 10 – 14 days after birth, and then to 56% by the 6 – 8 week review. Across the age groups, the drop-off in breastfeeding is slightly higher between initially starting and the 10 – 14 day review (a drop of 10 – 15 percentage points) than between the 10 – 14 day review and the 6 – 8 week review (a drop of 8 – 10 percentage points).

#### 4.2.2.5 Effect of Maternal Smoking Status

The effect of maternal smoking status on the type of infant feeding in 2017⁄18 is shown in Figure 17 and Figure 18. In Figure 17, numbers of infants and the type of feeding at the Health Visitor's first visit 10 – 14 days after birth are displayed by maternal smoking status: 'Non-smoker', 'Smoker' or 'Unknown/Invalid'. For infants with a mother who is a non-smoker, 56% are fed some breast milk, whether this is via being exclusively breastfed or via mixed feeding. For infants with a mother who is a smoker, only 28% are fed some breast milk.

In Figure 18, percentages of infants, the type of feeding and the drop-off in breast and mixed feeding between the first visit at 10 – 14 days and the 6 – 8 week review are shown by maternal smoking status. For infants with a mother who is a non-smoker, the number fed some breast milk drops from 56% to 46% by the time of the 6 – 8 week review. The drop-off in breastfeeding percentage points for infants with a mother who is a smoker is similar, as the number fed some breast milk drops from 28% at 10 – 14 days to 17% by the time of the 6 – 8 week review. The main difference appears to be that mothers who are smokers appear to be less likely to be breastfeeding by the time of the first visit at 10 - 14 days. This may be due to mothers who are smokers being less likely to try breastfeeding initially.

Some records were marked with a smoking status of 'Unknown/Invalid'. It can be seen that the pattern of feeding types for these records is very similar to those where the maternal smoking status was recorded as 'Non-smoker', so these records may belong to mothers who are non-smokers. Overall, these records account for less than 1% of the records for 2017/18.

### 4.2.3 'P1 Body Mass Index Statistics' by SIMD

The third dataset to be used was the 'P1 Body Mass Index (BMI) Statistics' dataset. The data from 2017/18 was studied by SIMD quintile and Figure 19 was generated using the 'Epidemiological BMI by Deprivation at Council Area level' resource with data grouped by SIMD quintile and BMI group: 'Healthy weight', 'Overweight and Obese' and 'Underweight'. Figure 19 shows that BMI in P1 children is related to SIMD quintile. In 2017/18, 73% of P1 children living in SIMD 1 had a healthy body weight. For P1 children living in SIMD 5, this figure rose to 82%.

### 4.2.4 Joining data from different datasets

One of the aims of this dissertation mentioned in Section 1.2 was to demonstrate that it is possible to join data from different NHS Scotland open datasets in order to gain new insights. It was decided to investigate whether there might be a relationship between the body mass index of mothers and children. Maternal BMI data recorded in 2011/12 at the booking appointment, which is a pregnant woman's first appointment with the midwife to register her pregnancy, was joined with data on BMI for P1 children from 2017/18. These two open datasets do not contain information at the patient level, so joining of records for mothers with those of their own children is not possible. However, by selecting women who were pregnant in 2011/12 and children who were in P1 in 2017/18, a link can be made because women who were pregnant in 2011/12 would have had children of P1 age in 2017/18.

In Figure 20, data from joining the 'Epidemiological BMI by Deprivation at Council Area level' resource of the 'Primary 1 Body Mass Index (BMI) Statistics' dataset for 2017/18 with the 'Maternal Body Mass Index (BMI)' resource of the 'Births in Scottish Hospitals' dataset for 2011/12 are shown. Data for Maternal BMI in 2011/12 shows that 44% of pregnant women living in SIMD 1 had a healthy body weight at the time of their first appointment with the midwife in the early stages of pregnancy. This figure rises to 56% for those living in SIMD 5. As discussed in Section 4.2.3, 73% of P1 children living in SIMD 1 in 2017/18 had a healthy body weight and this figure rises to 82% for those living in SIMD 5.

The results displayed in Figure 20, show that both Maternal BMI and P1 children's BMI have a linear relationship with SIMD quintile and the highest percentage of healthy BMI results are seen for those living in SIMD 5 for both datasets.

Joining of two data frames was also carried out in Section 4.2.2.2, where the Infant Feeding by SIMD resource of the 'Infant Feeding' dataset was joined with the lookup table containing the Health Board names in order to display the Health Board names in Figure 14. This is discussed further in Section 4.4.3.

Other factors that were considered for investigation by joining datasets were Maternal Age and Ethnicity. The possibility of joining the 'Births in Scottish Hospitals' dataset with the 'Infant Feeding' dataset by maternal age was contemplated. However, different groupings have been used for maternal age in the two datasets. This difference is discussed further in Section 4.4.3. The 'Infant Feeding' and '27 – 30 Month Review Statistics' publication reports both discuss the role of ethnicity, so this was another option for joining of datasets. However, the 'Infant Feeding' dataset on the NHS Scotland Open Data portal does not contain any data about ethnicity, so it was not possible to investigate this factor by joining the two datasets. This difference between the open data and the publication report is discussed further in Section 4.4.2.

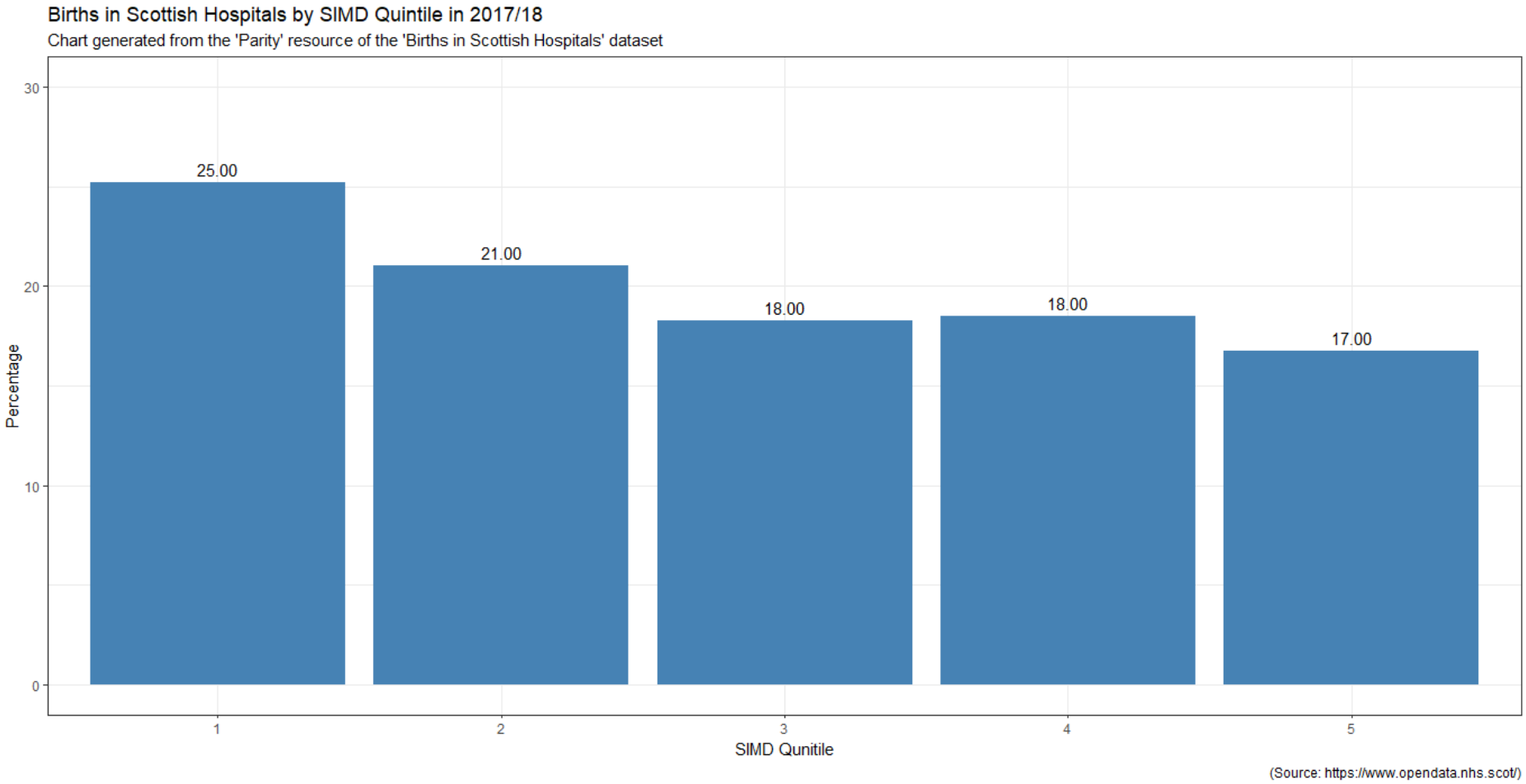


**Figure 10. Births in Scottish Hospitals by SIMD Quintile in 2017/18**

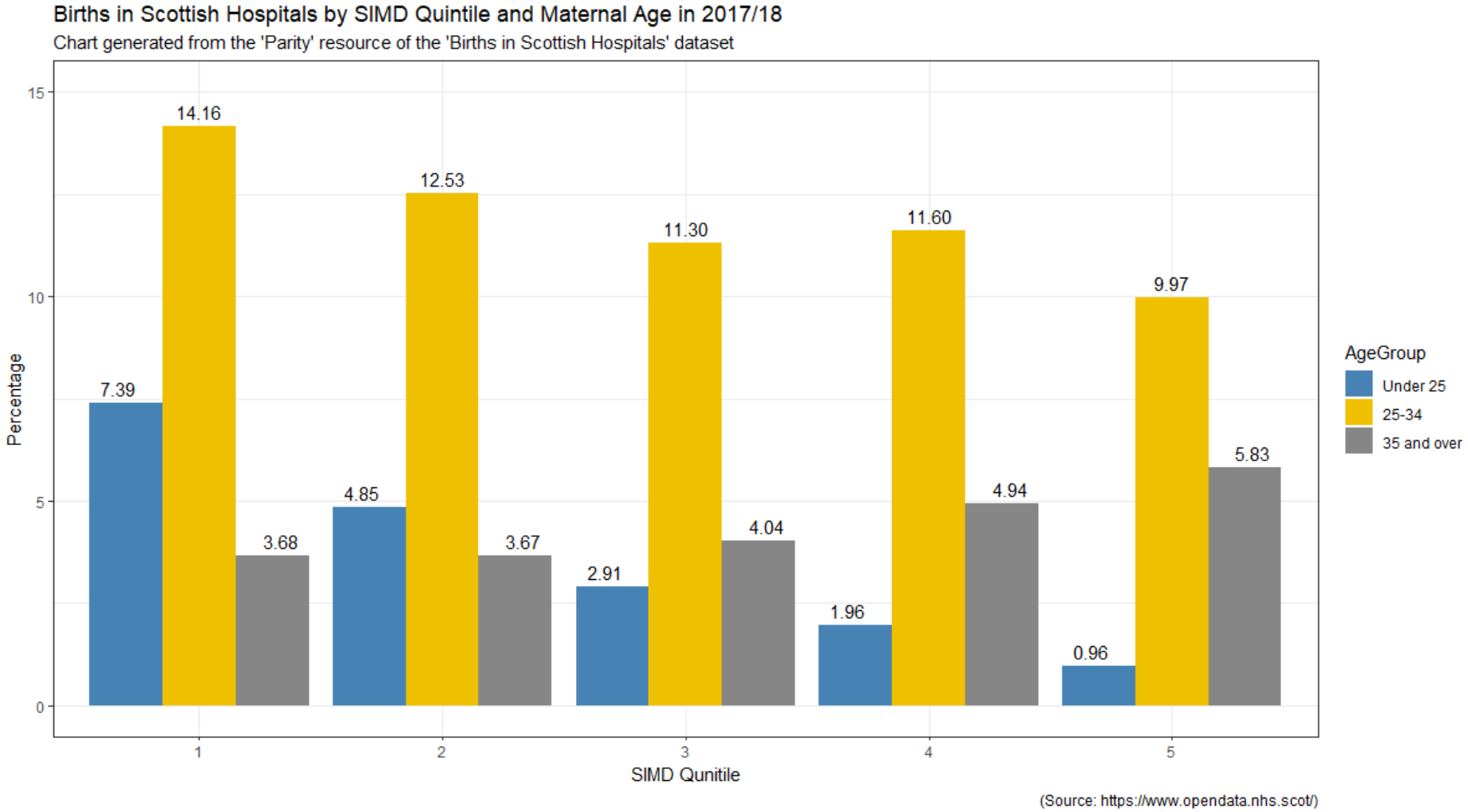


**Figure 11. Births in Scottish Hospitals by SIMD Quintile and Maternal Age in 2017/18**

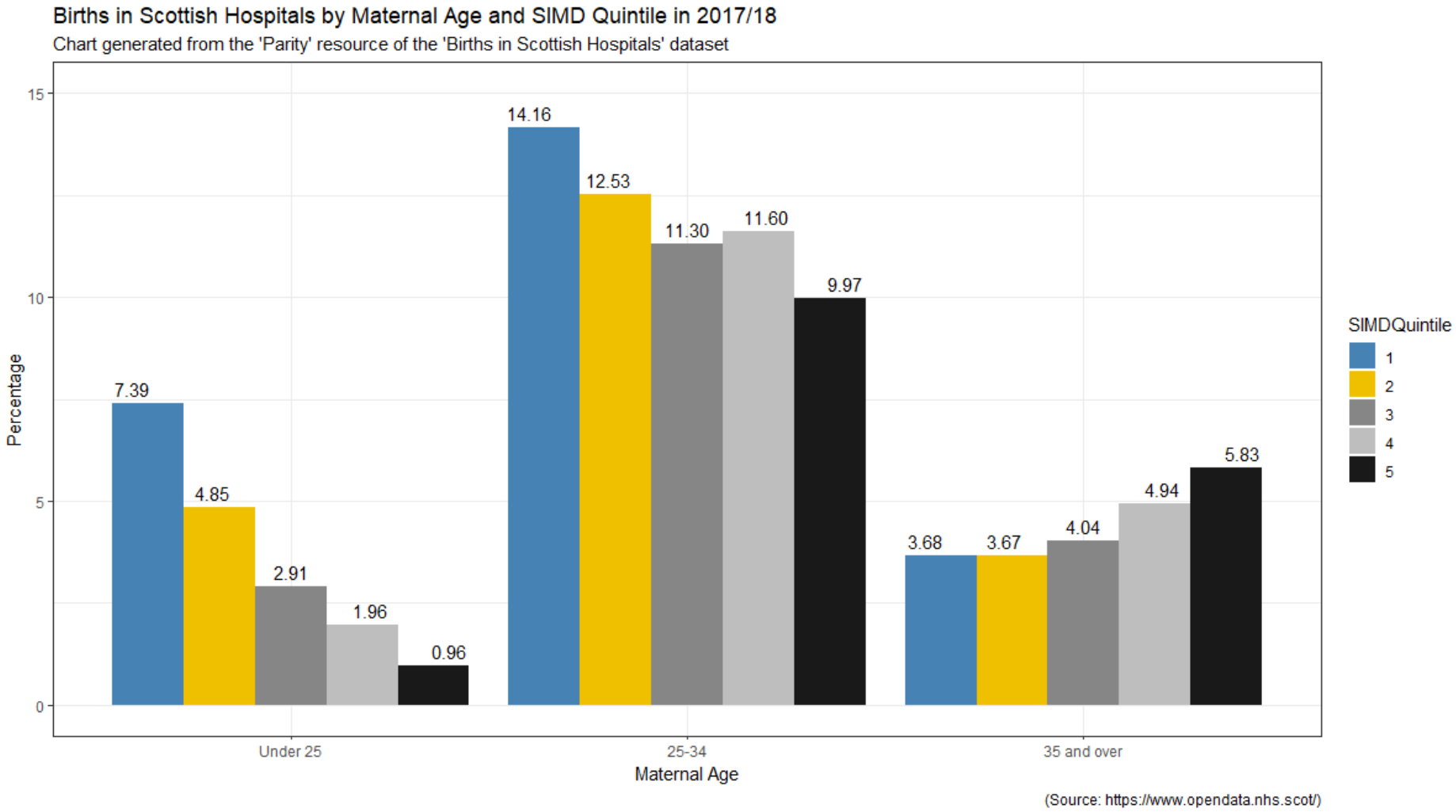


**Figure 12. Births in Scottish Hospitals by Maternal Age and SIMD Quintile in 2017/18**

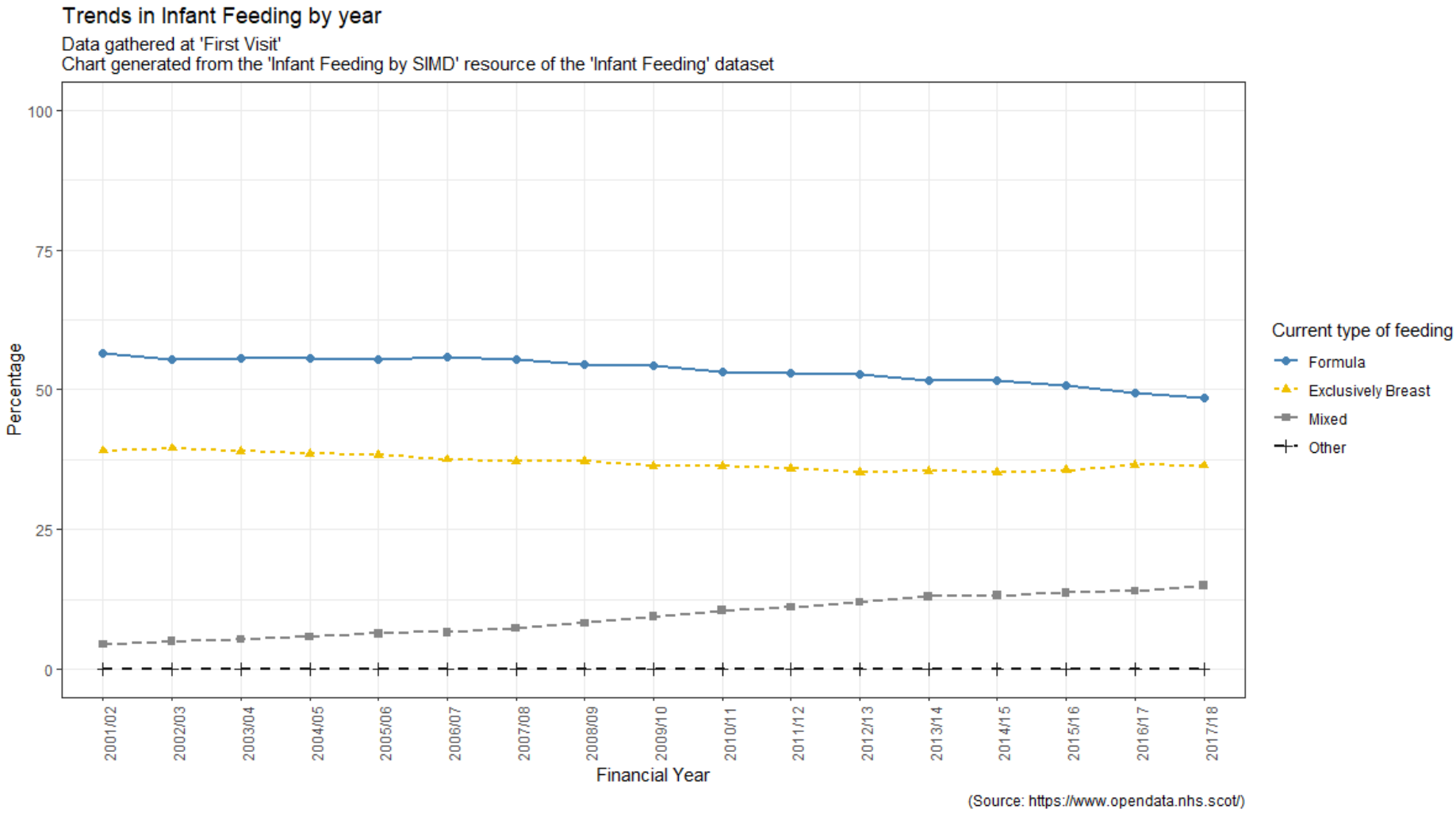


**Figure 13.Trends in Infant Feeding by Year**

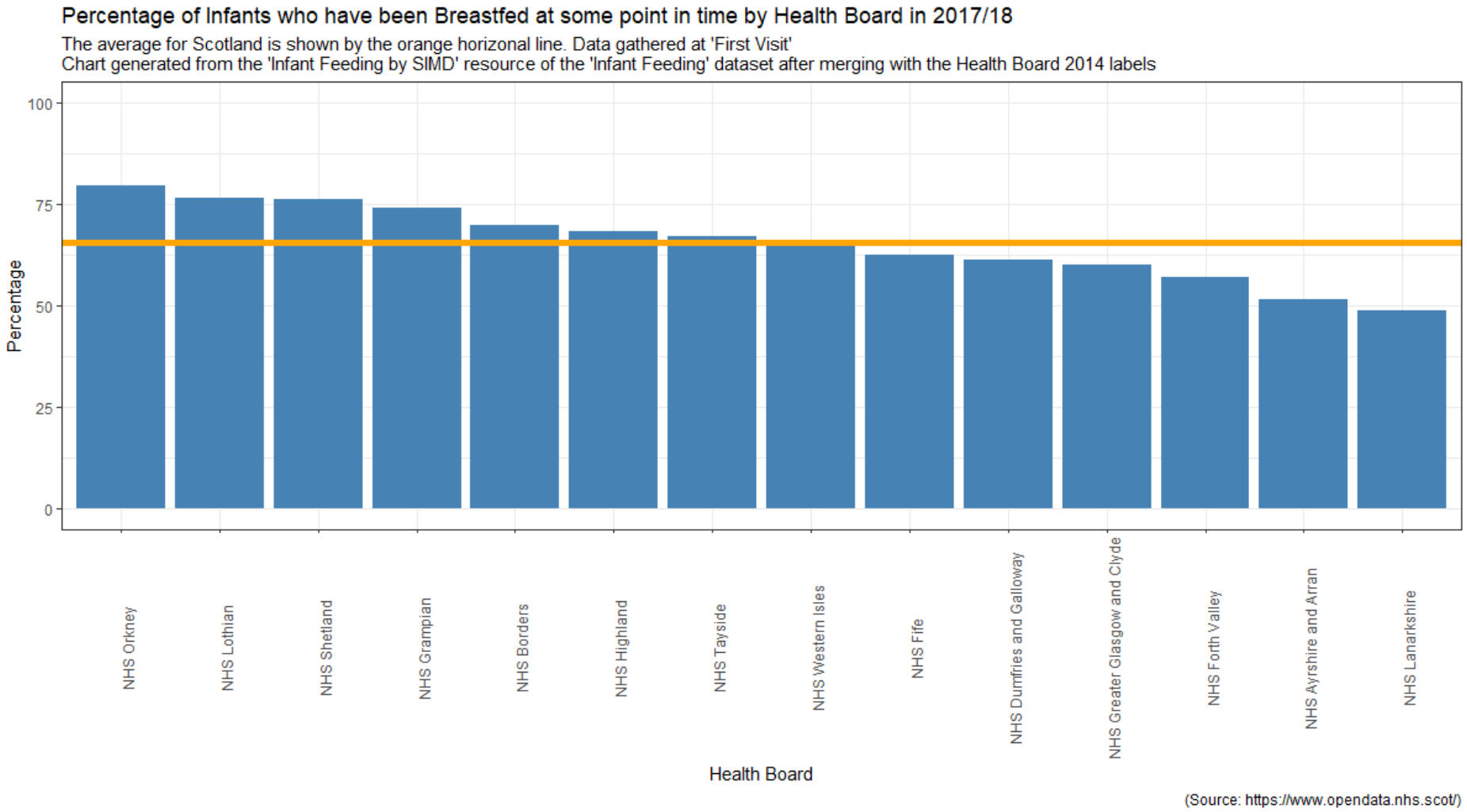


**Figure 14. Percentage of Infants who have been Breastfed at some point in time by Health Board in 2017/18**

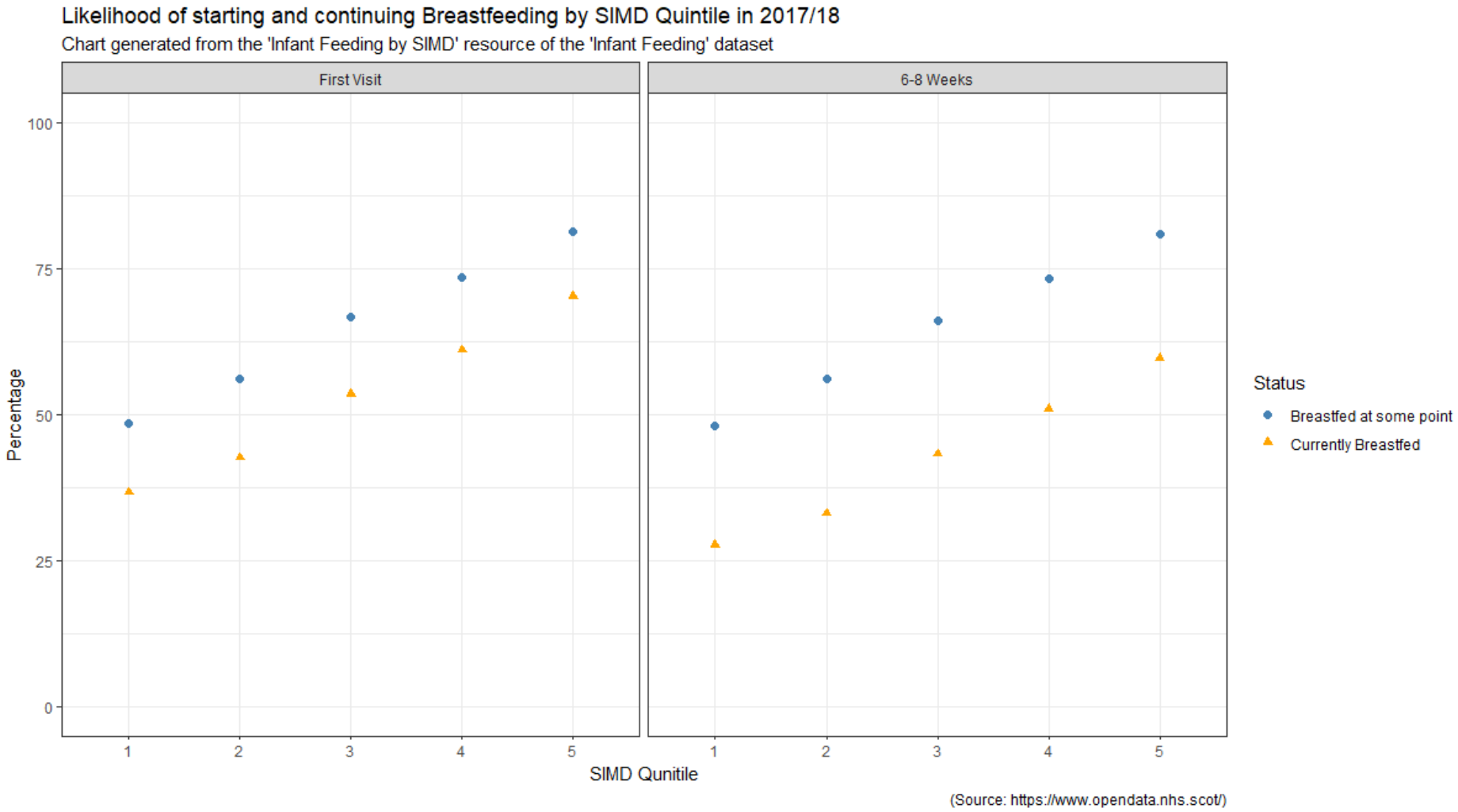


**Figure 15.Likelihood of starting and continuing Breastfeeding by SIMD Quintile 2017/18**

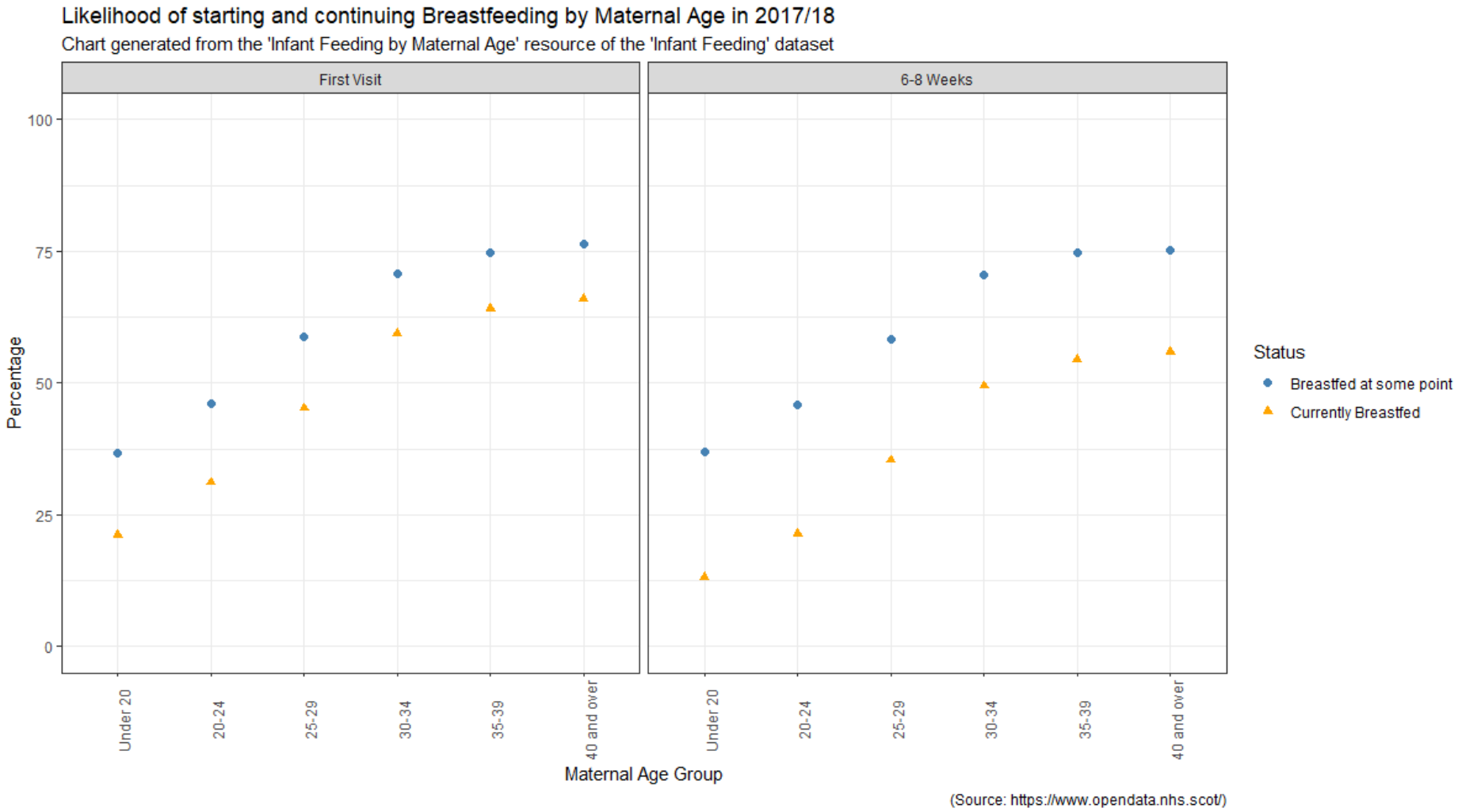


**Figure 16.Likelihood of starting and continuing Breastfeeding by Maternal Age 2017/18**

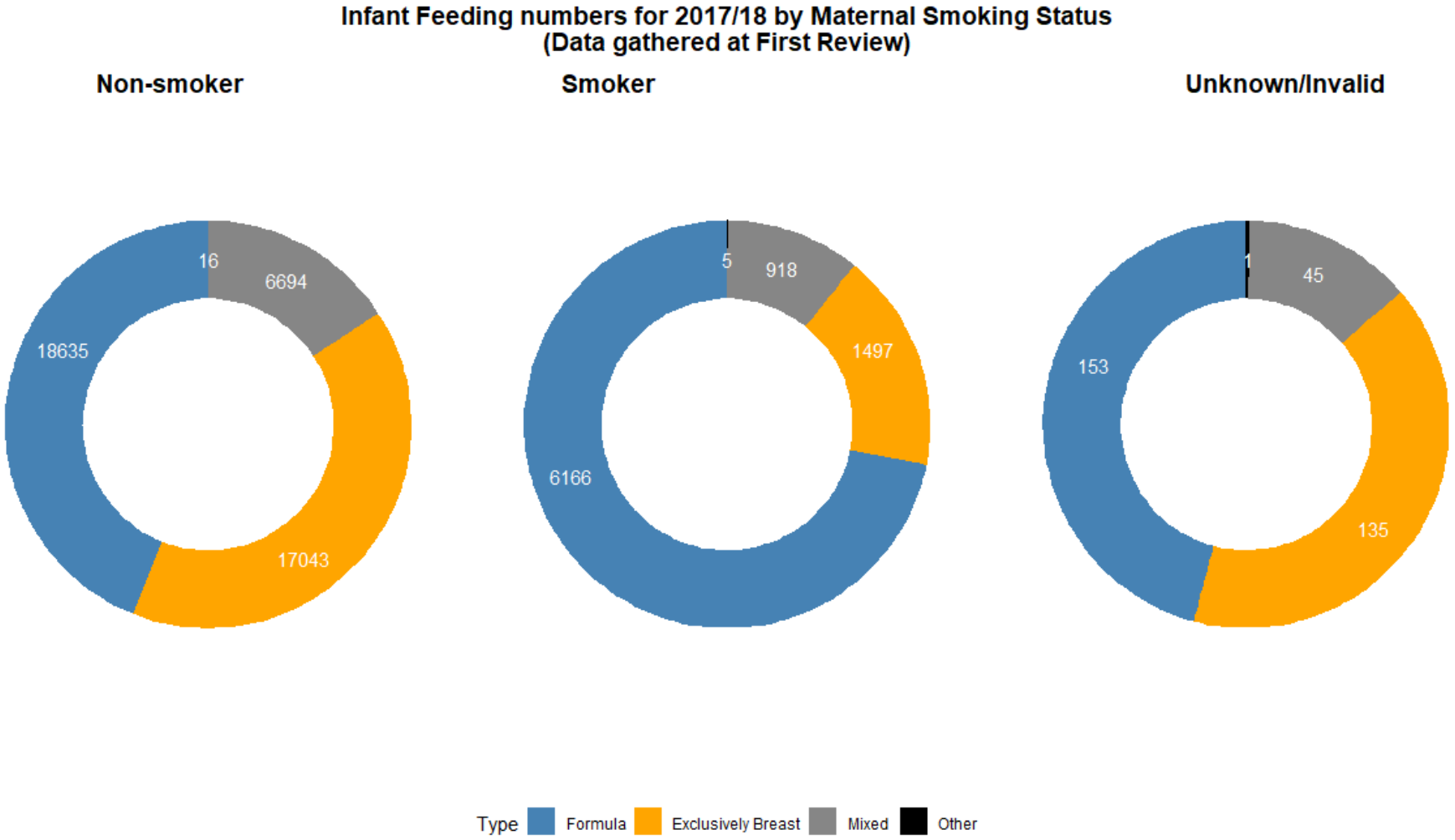


**Figure 17. Infant Feeding numbers at the First Review by Maternal Smoking Status in 2017/18**

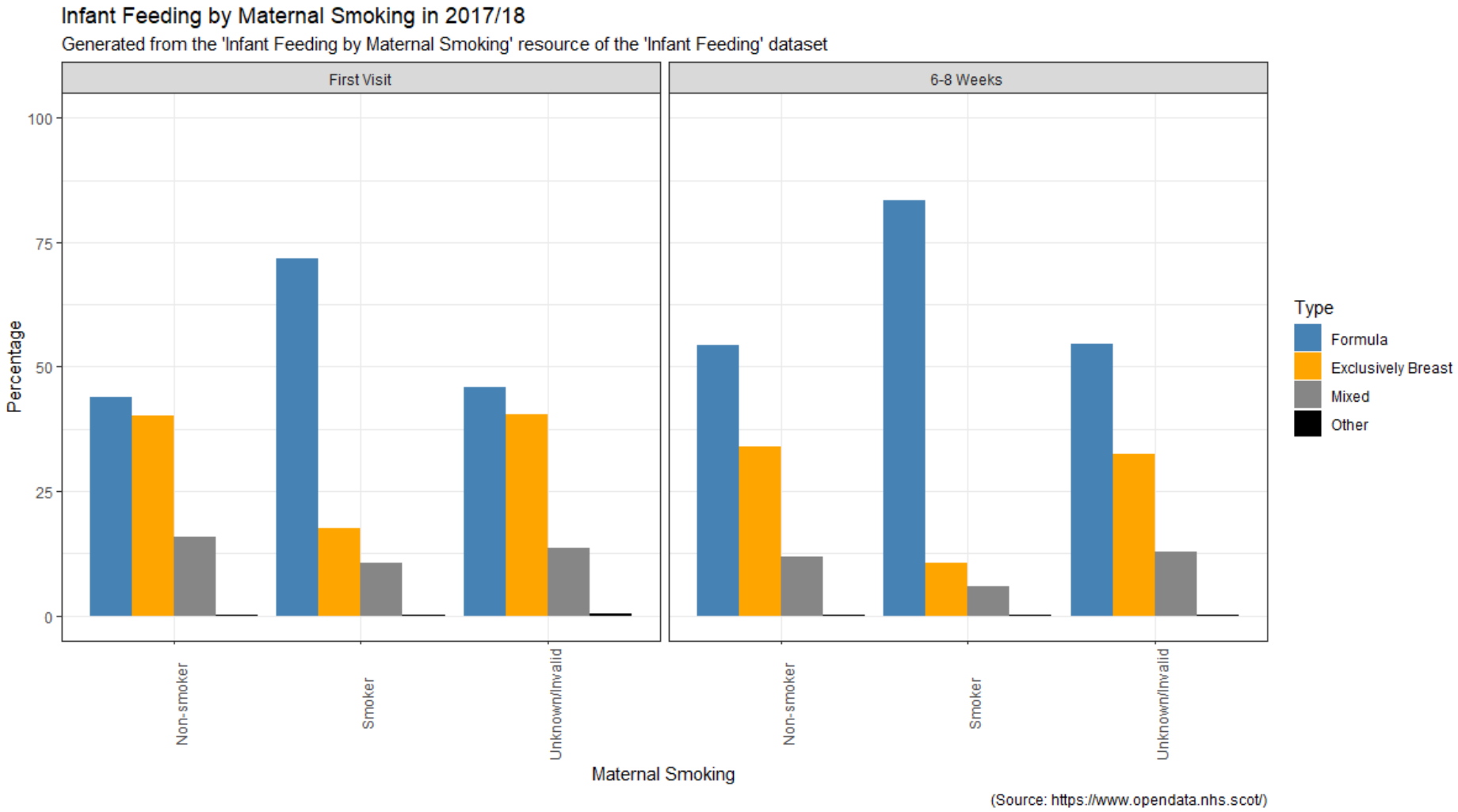


**Figure 18.Infant Feeding by Maternal Smoking Status in 2017/18**

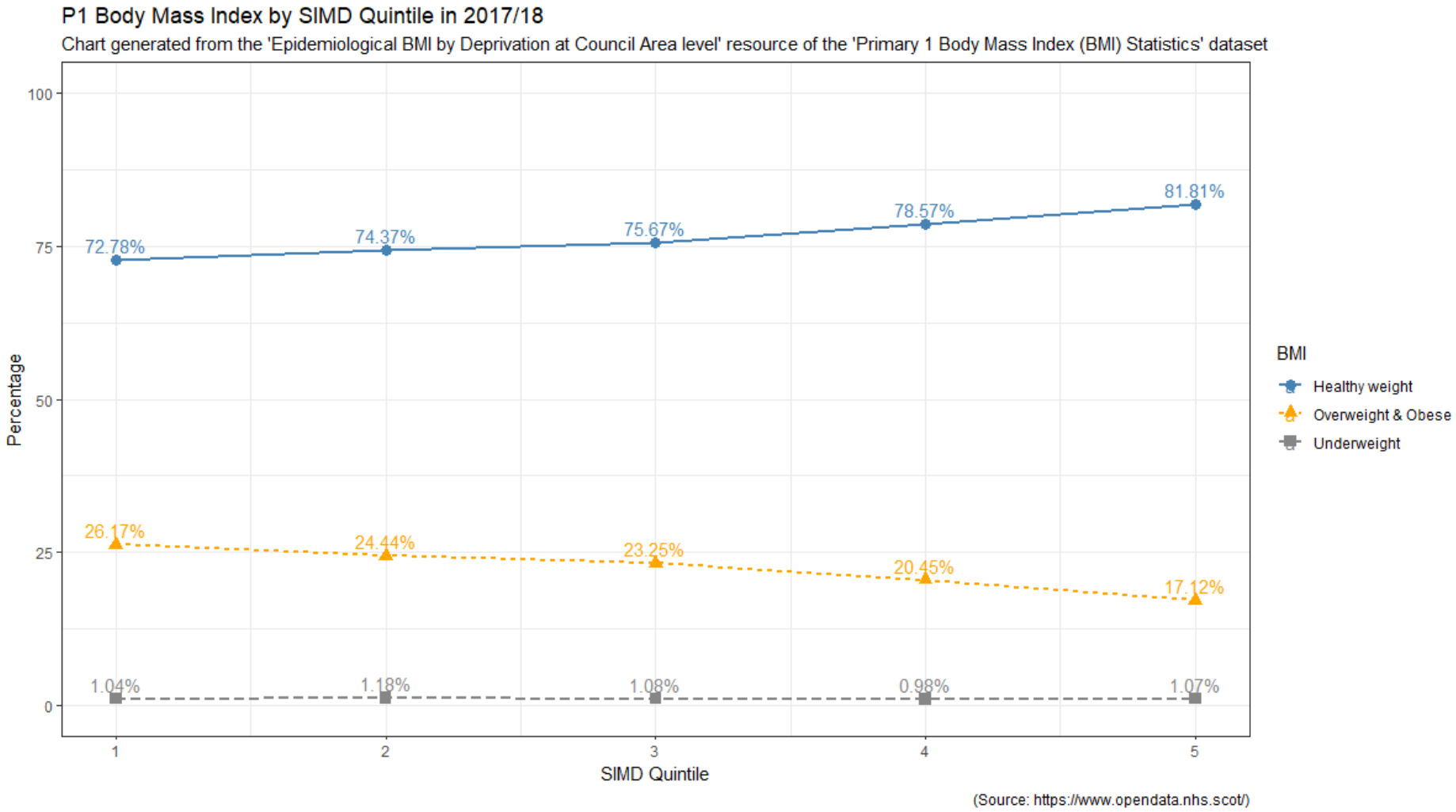


**Figure 19. Primary 1 Body Mass Index by SIMD Quintile in 2017/18 generated from the Epidemiological BMI data**

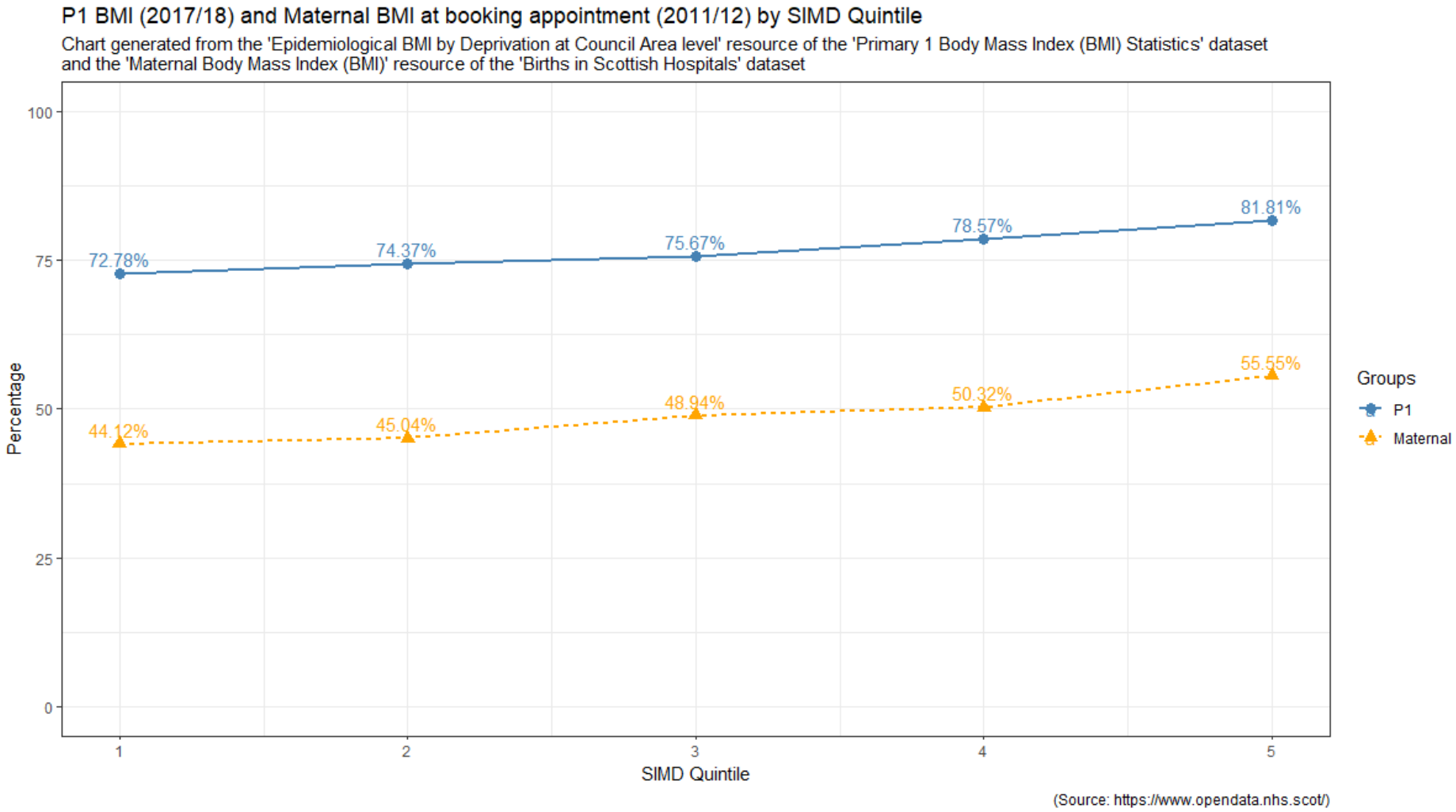


**Figure 20.P1 Body Mass Index (2017/18) and Maternal Body Mass Index at booking appointment (2011/12) by SIMD Quintile**

## 4.3 Specification for a Shiny app for 'Child Health in Scotland'

As well as the evaluation of R for producing data visualisations, the appraisal of the R package, Shiny [21] for creating an interactive data story was also part of the second aim of this dissertation. Before creating any data product, though, it is important to consider who the users of the product might be and what requirements they might have.

Potential users of a Shiny application about Child Health in Scotland could include NHS Boards and other NHS organisations, the Scottish Government, partner organisations, commercial organisations, research bodies and academics, the media and the public.

Important features for a specification for these user groups to include are:

- A navigation bar or introductory page that lists the datasets in the 'Child Health' group and allows the user to select which dataset they would like to explore: (1) Births in Scottish Hospitals, (2) Infant Feeding, (3) 27 – 30 Month Review Statistics, or (4) P1 Body Mass Index (BMI) Statistics
- Once a dataset has been selected, a menu that lists the resources available for that dataset. For example, for 'Infant Feeding' there are five resources available
- Once a resource has been selected, a set of menus to allow the user to filter the data or select a subset. For 'Infant Feeding by Deprivation', the user might wish to select data from a particular year or just certain fields in the resource
- The menu selections would need to feed into the path for the call to the API in the server function of the application
- The ability to download and save the data selection used to create a plot via a 'Download the data' button
- The option to include annotations to provide additional information about some plots
- The ability to export and save a plot via a 'Download the plot' button
- The option to display data from different datasets on the same plot or in two plots side by side
- An interactive map of Scotland with the option to display the boundaries of the NHS Health Boards or the local authorities and additional information appearing on hovering or clicking the cursor over areas of the map
- Data Dictionaries and Metadata would need to be available as tables in the application, perhaps with a download option
- Links to 'Other Useful Information' related to Child Health such as the publication reports from ISD or reports from Scottish Government

## 4.4 Feedback on the 'User Experience' of the NHS Scotland Open Data portal

### 4.4.1 Ease of use of the open data portal

Currently, ISD does not receive much feedback on how their open data is used or the ease of use of the NHS Scotland Open Data portal. In this section, positive feedback from the 'User Experience' undergone during this dissertation will be given and some suggestions for possible future improvements will be provided.

In Section 3.1, an introduction to the portal was given. Figure 3 shows the landing page, Figure 4 displays the group of Child Health datasets, Figure 5 lists the resources for the 'Infant Feeding' dataset, and Figure 6 shows the page with a preview window for the Infant Feeding by

Maternal Age resource. The website has a clear and consistent layout. Navigation is straightforward and it is easy to locate specific groups, packages and their resources. The page for a specific dataset includes metadata in a Table entitled 'Additional Information' and links to published reports and the relevant page on the ISD website. The page for a specific resource includes a 'Data Dictionary' which describes the fields in the resource and a table of 'Additional Information'.

As a new user of the website, it was not immediately obvious how to locate the ID number for a resource, which is required when building the path for an API call. This ID number is located in the 'Additional Information' table and is labelled as 'id'. However, this table is at the very bottom of the page underneath the preview window and Data Dictionary, and only the first few rows of the table are visible when the page is initially viewed. The resource ID is included in the link for downloading the data as a CSV file but this is only clear to the user once they know the resource ID. In future, it might be useful to list the resource ID at the top of the page underneath the title and the link for downloading the data as a CSV file.

There is a link to the API Guide in the CKAN documentation on the left hand side of the footer of the website [76]. This link gives some examples of how to structure a call for a package or resource matching a query but the link currently points to version 2.7.3 of the documentation instead of the latest version of the documentation 2.9.0a.

A green 'Data API' button can be seen in Figure 6, which shows the page for the Infant Feeding by Maternal Age resource. This button opens a pop-up window that provides useful information such as links to the CKAN documentation for the DataStore extension, which is part of the Maintainer's Guide [77], examples of different actions that can be carried out such as 'datastore_search' and examples of different queries. This section of the documentation provides more useful information than the link in the footer, so perhaps the link in the footer should be updated. The pop-up window also contains two examples of requests to the API written in JavaScript and Python. The analysis carried out during this dissertation has been done using R. So, for other users of R, perhaps it would be useful to add an example of a request to the API written in R to this pop-up window.

### 4.4.2 Differences between ISD data tables and the open data resources

Some differences were noted when comparing the ISD data tables that are released alongside the annual publication reports with the open data available via the portal. Sometimes, not all of the data used in an annual publication report is released as open data. For example, the Infant Feeding publication available on the ISD website contains some interesting data which shows how breastfeeding rates vary with ethnicity. However, this information is not available as open data on the portal. Investigation of this discrepancy revealed that the role of ethnicity in breastfeeding rates was added as a new experimental category in the recent publication report. This was after the open data had been generated and loaded onto the portal, which is why the data on ethnicity is not available as open data. Experimental categories such as this one are carefully monitored and if the numbers in certain groups within the category become low enough for potential identification of individuals to become an issue, then the category will be removed from future reports.

The other main difference between the ISD data tables and the open data is format. The data tables on the ISD website are Excel files. Sometimes, a dataset is presented as one file with a separate tab for each data table. In other cases, a separate Excel file is provided for each data table. On the open data portal however, the format is consistent. Each package is split into a number of resources and each resource is available to download as a separate CSV file or via the API. In Section 1.1, the decision to provide NHS Scotland open data as 3 star data in a non-proprietary format covered by the UK Open Government Licence (OGL) [2] was discussed. The

data tables on the ISD website available in Excel format are not covered by this licence and do not use non-proprietary software but they are currently released alongside each publication report, so for now, the decision has been taken to continue to make these tables available on the ISD website with the publication reports.

### 4.4.3 Consistency of variable and band names

The decision was taken by the open data team to use codes for some of the variables in the open data resources. One example of this is in the Infant Feeding by Maternal Age resource where the 'HBR2014' variable is used for health board with 14 codes agreed and 'CA2011' is used for council area with 32 codes agreed. The reason for using codes instead of names was to prevent multiple different data entries for the same health board or council area name. For example, 'Argyll and Bute', 'Argyll and bute', 'ArgyllandBute', 'Argyll & Bute', 'Argyll&Bute' are just a few of the possible variations of one council area name that might occur without pre-agreed codes. The names of the health boards and council areas are provided in lookup tables in the Data Dictionary for the resource.

Performing a join in R to link the resource with the lookup table, so that health boards or council area names can be used in plots generated during analysis is straightforward. The only point to note is that slightly different field names were used for the health board codes in the resource (HBR2014) and in the lookup table (HB2014). The same code was used for council area code in both the resource and the lookup table (CA2011).

Variation in naming was found between different datasets. This is believed to be due to each publication being produced by a separate team, each of whom may have their own naming conventions. Also, some differences between datasets were observed in the way that data was grouped, which made it difficult to compare data with other datasets. For example, the Parity resource of the 'Births in Scottish Hospitals' dataset uses three bands for Maternal Age: 'Under 25, '25 – 34' and '35 and over'. The Infant Feeding by Maternal Age resource of the 'Infant Feeding' dataset however, splits the data into six bands for Maternal Age: 'Under 20', '20 – 24', '25 – 29', '30 – 34', '35 – 39' and '40 and over'. It is likely that the broader age bands in the Parity resource of the 'Births in Scottish Hospitals' dataset have been used to prevent the numbers in a particular age band from becoming small enough to make statistical disclosure an issue. However, this lack of consistency makes it more difficult to join datasets.

# 5 Conclusion

## 5.1 Summary

Learning how to access the NHS Scotland open data via the API was the first task in this dissertation. This objective was successfully achieved and a set of scripts were written to query the data from three datasets related to 'Child Health' which were Births in Scottish Hospitals', 'Infant Feeding' and 'P1 Body Mass Index (BMI) Statistics'.

The second objective was to evaluate R, ggplot2 ad Shiny for the creation of data visualisations. The NHS has been using Tableau for data analysis and visualisation for the last five years but recently, there has been a trend within the NHS towards the use of R as an alternative to Tableau. A range of visualisations were produced using R and these illustrated the effect of factors such as trends over time, variation by Health Board, SIMD quintile and maternal age. A draft specification for a Shiny application about Child Health in Scotland was written bearing in mind potential user groups. These users could include NHS Boards and other NHS organisations, the Scottish Government, partner organisations, commercial organisations, research bodies and academics, the media and the public. A prototype of a Shiny application was produced.

Obtaining feedback on how NHS Scotland open data is being used and the ease or difficulty of use of the NHS Scotland Open Data portal is of interest to ISD. So, the third objective of this dissertation was to document the 'User Experience'. It was shown that visualisations can be created from the open data, as well as from the patient-level data that is currently used by ISD. It was also demonstrated that two datasets can be joined to create an informative visualisation and gain new insights.

## 5.2 Evaluation

Accessing the NHS Scotland open data via the API was successfully achieved and in Section 4.1, Figure 7 and Figure 8 , the method that was used for making calls to the API can be seen. Calls to the API were used to obtain: (1) the ID numbers for the resources available for a package, (2) the number of records that matched a particular query, and then (3) the records themselves. This method of using three calls to the API was developed with a view to using drop-down menus to select a resource or a subset of the data in a future version of the Shiny application about 'Child Health'.

In the prototype application, an introductory page describes the datasets and invites the user to select one of the datasets about 'Child Health' from the navigation bar. Scripts containing the API calls have been saved as functions which are called in the server function. In a future version of the application, a set of drop-down menus could be linked to conditional functions to select a resource for a package and the subset of data that the user wishes to view, such as data for a particular year or for a certain Health Board.

Evaluation of R and the packages ggplot2 and Shiny for creating data visualisations and a data story as an alternative to Tableau was the second objective of this dissertation. Three data sets from the 'Child Health' group of datasets on the NHS Scotland Open Data portal were studied: 'Births in Scottish Hospitals', 'Infant Feeding' and Primary 1 Body Mass Index Statistics'.

Analysis of the 'Parity' resource for the 'Births in Scottish Hospitals' dataset for 2017/18 by SIMD quintile revealed a higher than average percentage of births occurred in the most deprived areas. Analysis by SIMD quintile and maternal age showed this same trend in the 'Under 25' and '25 – 34' age groups with a higher percentage of births occurring in the most deprived areas. The trend is less marked for the '25 – 34' age group. By contrast, the trend

observed for the '35 and over' age group is for an increasing percentage of births in the least deprived areas. Results can be seen in Figure 10, Figure 11 and Figure 12.

The publication reports for both 'Births in Scottish Hospitals' and 'Infant Feeding' discuss maternal age in terms of six age groups and it had been hoped to join these two datasets by maternal age. However, this was not possible as it was discovered that in the open data, only three age groups are used for reporting maternal age in the 'Births in Scottish Hospitals' dataset. It would be necessary to group data to reduce the number of factors in the 'Maternal Age' variable of the 'Infant Feeding' dataset from six age groups to three before these two datasets could be joined.

There is plenty of scope for further work using the 'Births in Scottish Hospitals' dataset. There are sixteen resources available but only two were used in this dissertation to assess the effects of SIMD quintile, maternal age and maternal BMI. Further work to investigate the effects of smoking behaviour, alcohol consumption and drug misuse during pregnancy over time and by Health Board would be an interesting area to pursue.

The 'Infant Feeding' dataset was studied in more depth with trends over time, variation by Health Board, and the effects of SIMD quintile, maternal age and maternal smoking status all being investigated.

Trends in the four types of infant feeding for the whole of Scotland are shown from 2000/01 to 2017/18 in Figure 13. Over the time period, the percentage of infants receiving some breastmilk, either via being exclusively breastfed or via mixed feeding has increased.

Data on infant feeding from 2017/18 was assessed for the effect of four different factors. First, the effect of location was investigated by looking at the percentage of infants who had been breastfed at some point in time by Health Board. Figure 14 shows the average for Scotland as an orange line. It can be seen that the Health Boards with the highest percentages of infants who have been breastfed at some point are, in general, located in the north and east of Scotland, whilst those with the lowest percentages of infants who have been breastfed at some point are located in the west of Scotland. A detailed study of particular Health Boards was not carried out in this dissertation but it would be interesting to examine this finding in more detail by comparing data for two Health Boards with very different breastfeeding rates such as NHS Lothian and NHS Lanarkshire.

When the effect of SIMD quintile on the likelihood of starting and continuing to breastfeed in 2017/18 was investigated, the results indicate that the highest rates of breastfeeding are observed in the least deprived areas. A linear relationship between SIMD quintile and breastfeeding can be seen in Figure 15. It appears that mothers living in the most deprived areas are less likely to try breastfeeding, so more education about the benefits of breastfeeding or different ways of delivering this education would be beneficial for pregnant women and their families who are living in the most deprived areas.

The drop-off in breastfeeding is higher between initially starting and the 10 – 14 day visit, than between the 10 – 14 day visit and the 6 – 8 week review. This drop-off also appears to occur across all SIMD quintiles. Support for mothers from midwives, health visitors, breastfeeding groups, and mother and baby groups to encourage mothers to persevere with breastfeeding would be beneficial across all SIMD quintiles.

A relationship can be seen between maternal age and breastfeeding with the highest rates of breastfeeding being observed with older mothers. In Figure 16, the relationship is linear for the youngest four age groups and then appears to level out for mothers aged 35 years and over. The results suggest that it might be beneficial to target education about the benefits of breastfeeding at women under the age of 30.

As observed with the analysis by SIMD quintile, the drop-off in breastfeeding is higher between initially starting and the 10 – 14 day visit, than between the 10 – 14 day visit and the 6 – 8 week review. The drop-off is observed across all age groups. This suggests that mothers of all age groups might benefit from additional support from midwives, health visitors, breastfeeding groups, and mother and baby groups in the first 10 – 14 days after birth to encourage them to continue with breastfeeding.

The effect of maternal smoking status on the type of infant feeding in 2017/18 is shown in Figure 17 and Figure 18. Infants with a mother who is a non-smoker are twice as likely to receive some breast milk whether this is via being exclusively breastfed or via mixed feeding. This suggests mothers who are smokers may be less likely to try breastfeeding. There is no resource in the open data to allow possible correlations between maternal smoking and SIMD, or between maternal smoking and maternal age to be studied but these would both be interesting areas to investigate should the information be made available in future releases of the open data.

One of the aims of this dissertation was to demonstrate that it is possible to join data from different NHS Scotland open datasets in order to gain new insights. A possible relationship between maternal BMI and P1 children's BMI was investigated. The 'Maternal Body Mass Index (BMI)' resource of the 'Births in Scottish Hospitals' dataset for 2011/12 was joined with the 'Epidemiological BMI by Deprivation at Council Area level' resource of the 'Primary 1 Body Mass Index (BMI) Statistics' dataset for 2017/18. It was interesting to note that on average, less than 50% of women had a healthy BMI at the time of booking their first appointment with the midwife. The results displayed in Figure 20, show that both Maternal BMI and P1 children's BMI have a linear relationship with SIMD quintile and the highest percentage of healthy BMI results are seen for those living in SIMD 5 for both datasets.

There is a fourth dataset in the 'Child Health' group on the NHS Scotland Open Data portal that was not used during this dissertation. The '27 - 30 Month Review Statistics' publication report contains some interesting data on ethnicity [78] [79] [80]. Initially, it was hoped to try and join this dataset with the 'Infant Feeding' dataset to investigate whether there were any patterns related to ethnicity, breastfeeding rates and the likelihood of any developmental concerns being noted during the 27 – 30 month review. However, the open data for 'Infant Feeding' does not contain the information on ethnicity that is mentioned in the publication report, so this piece of analysis was not possible.

The appraisal of the R package, Shiny for creating an interactive data story was also part of the second aim of this dissertation. In Section 4.3, the specification for a Shiny application about 'Child Health' in Scotland has been considered and a prototype product has been produced. There is scope for developing this product further to include information from the resources that were not investigated during this dissertation.

The third objective of this dissertation was to provide ISD with some feedback on the 'User Experience' of working with the NHS Scotland Open Data portal. It was successfully demonstrated that informative visualisations can be created from the open data, as well as from patient-level data which is what is used by ISD when producing publication reports. It was also demonstrated that two datasets can be joined to create an informative visualisation and gain new insights. Data from the 'Maternal Body Mass Index (BMI)' resource of the 'Births in Scottish Hospitals' dataset was joined with data from the 'Primary 1 Body Mass Index (BMI) Statistics' dataset and both sets of data were found to have a linear relationship with SIMD quintile. Other suggestions for improvements were given in Section 4.4.

## 5.3 Future Work

### 5.3.1 Joining datasets

In Section 4.2.4, Figure 20 shows that both Maternal BMI from the 2011/12 data and P1 children's BMI from data collected six years later in 2017/18have a linear relationship with SIMD quintile. The highest percentages of healthy BMI results are seen for those living in SIMD 5 for both datasets. As a possible area for future work, linear regression could be used to study the relationships the lines show in more detail and compare them.

A similar comparison could be carried out between breastfeeding and P1 children's BMI. The 'Infant Feeding by deprivation' resource of the 'Infant Feeding' dataset could be linked with the 'P1 Body Mass Index Statistics' dataset. Data from 2012/13 on infants who were breastfed at some point could be linked with P1 children's BMI from data collected five years later in 2017/18 as there would be a five year gap between an infant being breastfed and reaching P1 age. Again, linear regression could be used to compare the relationships shown. Alternatively, links between maternal BMI and breastfeeding could be investigated, as there is some evidence that women who are obese stop breastfeeding sooner that those who have a healthy BMI [81] [82]. It would also be of interest to investigate the role of ethnicity in infant feeding, developmental concerns noted during the 27 – 30 month reviews and P1 BMI. Recent research has indicated that standard BMI tests for children over estimate body fat in UK children of black African descent and under estimate body fat of UK children of South Asian descent. A method of adjusting BMI measurements for ethnicity has been developed and it would be interesting to evaluate some open data for P1 BMI adjusted for ethnicity in the future [83] [84].

Joining datasets by maternal age and by ethnicity were both considered but these ideas were not pursued. Different groupings were used for maternal age in the 'Births by Scottish Hospitals' and 'Infant Feeding' datasets in the open data. Ethnicity data for Scotland as a whole is available for the last five years in the '27 – 30 Month Review Statistics' but not for the 'Infant Feeding' dataset. In order to determine what other options are possible for joining open datasets, all the resources available in the 'Child Health' group of datasets would need to be examined, as some differences between the open data and the ISD data tables issued with the publication reports do exist.

The idea of linking the open data with data from other sources was also considered. There appears to be a correlation between the local authorities selected by the Scottish Government as Challenge Authorities and the health boards with the lowest rates of breastfeeding. The Challenge Authorities have the highest concentration of school children who are living in the 20% most deprived parts of Scotland [85]. Looking at Figure 14, it is interesting to note that eight out of nine of these Challenge Authorities are in the Health Boards with the lowest rates of breastfeeding. In addition, there may be a correlation between the Health Boards with the highest rates of development concerns observed during the 27 – 30 Month reviews and the Challenge Authorities. The Scottish Government consultation paper entitled: 'Measuring the attainment gap: consultation' published in 2017 refers to the 27 – 30 Month Review Statistics and discusses an attainment gap of almost 17% between children from the most and least deprived areas. The percentage of children receiving a report of 'No development concerns' is used as a measure of attainment in this report [86].

However, it was discovered that the NHS uses a population weighting of SIMD, so that SIMD 1 – 5 referred to in ISD publications each contain 20% of the population of Scotland. This is different to SIMD used by the Scottish Government, so this would need to be taken into consideration if NHS open data was to be joined with Scottish Government data by SIMD quintile.

### 5.3.2 Case study

The majority of the analysis in this dissertation has focused on the effects of SIMD quintile and maternal age with only a limited amount of work carried out on the trends over time and the variation by health board for infant feeding. It would be interesting to select two Health Boards with very different rates of breastfeeding and carry out a more detailed case study of the these two areas by factors such as population size, demographics, teenage pregnancy rates, percentages of children with development concerns at 27 – 30 months, P1 body mass index, as well as by SIMD quintile and maternal age.

### 5.3.3 Further development of the Shiny application on 'Child Health'

There is plenty of scope for further improvements to be made to the prototype Shiny application. The '27 – 30 Month Review Statistics' dataset was not investigated during this dissertation, so analysis and visualisations to identify factors that might be involved in development issues in pre-school children would be of interest. There is also a new open dataset entitled: 'Teenage Pregnancy' that was released in July 2019 and it would be interesting to look for ways to analyse and link this data with the other datasets in the 'Child Health' group.

A draft specification for a Shiny application was discussed in Section 4.3 and a prototype product has been produced. The next step would be to include interactive features such as a map with additional information available on hovering or clicking areas of the map and drop-down menus to select which data to use. Feedback from potential users would be necessary to make further improvements.

### 5.3.4 Alternative software for data analysis and visualisation

This dissertation has focused on the use of R, ggplot2 and Shiny as alternatives to Tableau and this is in-line with a trend within the NHS towards the use of R. Some work has been carried out within ISD on evaluating D3 as part of a proof on concept study but no visualisations using D3 have been published. It would also be worth evaluating Python as an alternative tool for data analysis. There are a number of libraries such as Matplotlib, Pandas Visualisation, Seaborn, plotly and Bokeh available for producing visualisations and these are outlined in Section 2.3.3. It might also be worth evaluating Data Studio from Google, which does not have such a steep learning curve as R and Shiny but has less data wrangling capability.

### 5.3.5 Encourage engagement with the data

More could be done to encourage use of the NHS Scotland Open Data portal. As mentioned in Section 4.4.1, it would be useful to include an example of a request to the API written in R in the pop-up window that appears when the green 'Data API' button on the page for a resource is selected (see Figure 6).

The website could be marketed to universities by offering to give lectures or computer lab sessions on its use to students. Alternatively, the NHS could run a workshop for students at the Data Talent event (March), submit a challenge based on open data for the Data Lab's Innovation week (June), or sponsor a competition for teams of students to enter with a prize for the most innovative uses of the NHS Scotland open data.

# References


[1] Accessing APIs from R (and a little R programming): https://www.r-bloggers.com/accessing-apis-from-r-and-a-little-r-programming/ [Accessed: 24th May 2019]

[2] Open Government Licence for public sector information: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ [Accessed: 24th August 2019]

[3] Doi, J, Potter, G, Wong, J, Alcaraz, I, & Chi, P, Web Application Teaching Tools for Statistics Using R and Shiny, *Technology Innovations in Statistics Education*, 9(1), 2016

[4] NHS Scotland: https://www.scot.nhs.uk/about-nhs-scotland/ [Accessed: 14th July 2019]

[5] NHS Scotland: Organisations: https://www.scot.nhs.uk/organisations/ [Accessed: 13th July 2019]

[6] Information Services Division, ISD Scotland is part of NHS National Services Scotland: https://www.isdscotland.org/ [Accessed: 10th July 2019]

[7] The Official Statistics (Scotland) Order 2008: http://www.legislation.gov.uk/ssi/2008/131/contents/made [Accessed: 10th July 2019]

[8] Code of Practice for Statistics, UK Statistics Authority: https://www.statisticsauthority.gov.uk/code-of-practice/ [Accessed: 10th July 2019]

[9] Tableau platform for data analytics: https://www.tableau.com/ [Accessed 10th July 2019]

[10] R - What is R? https://www.r-project.org/about.html [Accessed: 27th July 2019]

[11] Scottish Government Open Data Strategy: https://www.gov.scot/publications/open-data-strategy/ [Accessed: 13th July 2019]

[12] The Re-use of Public Sector Information Regulations 2015: http://www.legislation.gov.uk/uksi/2015/1415/introduction/made [Accessed: 13th July 2019]

[13] Information Commissioner's Office: What is re-use of public sector information? https://ico.org.uk/for-organisations/guide-to-rpsi/what-is-rpsi/ [Accessed: 13th July 2019]

[14] Open Definition, a project of the Open Knowledge Foundation: http://opendefinition.org/ [Accessed: 13th July 2019]

[15] 5 Star Open Data: https://5stardata.info/en/ [Accessed: 13th July 2019]

[16] The Difference Between URLs, URIs, and URNs:
https://danielmiessler.com/study/url-uri/ [Accessed: 13th July 2019]

[17] Data modelling with RDF(S): What is RDF?
http://graphdb.ontotext.com/free/devhub/rdfs.html [Accessed: 13th July 2019]

[18] The DataStore API:
https://docs.ckan.org/en/2.8/maintaining/datastore.html#the-datastore-api
[Accessed: 29th July 2019]

[19] CKAN API Guide:
https://docs.ckan.org/en/latest/api/index.html [Accessed: 23rd August 2019]

[20] NHS Scotland Open Data platform:
https://www.opendata.nhs.scot/ [Accessed: 13th July 2019]

[21] RStudio: Shiny:
https://www.rstudio.com/products/shiny/ [Accessed 12th July 2019]

[22] Hans Rosling, Ola Rosling, Anna Rosling Rönnlund, *Factfulness: Ten Reasons We're Wrong About the World – and Why Things Are Better Than You Think*, Flatiron Books, 2018

[23] The Four V's of Big Data:
https://www.ibmbigdatahub.com/infographic/four-vs-big-data [Accessed 24th July 2019]

[24] Extracting business value from the 4 V's of big data:
https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data
[Accessed 24th July 2019]

[25] Berinato, S, *Visualizations That Really Work*, Harvard Business Review, 92–100, June 2016

[26] Grinstein, G and Trutschl, M, High-Dimensional Visualizations, *KDD 2001*: San Francisco, California, 2001

[27] Sedig, K, Parsons, P, Dittmer, M, Ola, O, Beyond information access: Support for complex cognitive activities in public health informatics tools, *Online Journal of Public Health Informatics*, 4(3):e2, 2012

[28] Ola, O and Sedig, K, Discourse with Visual Health Data: Design of Human-Data Interaction, *Multimodal Technologies and Interact*, 2(10), 2018

[29] Aung , T, Niyeha, D, Shagihilu, S, Mpembeni, R, Kaganda, J, Sheffel, A and Heidkamp, R, Optimizing data visualization for reproductive, maternal, newborn, child health, and nutrition (RMNCH&N) policymaking: data visualization preferences and interpretation capacity among decision-makers in Tanzania, *Global Health Research and Policy*, 4:4, 2019

[30] Ola, O, Sedig, K, The Challenge of Big Data in Public Health: An Opportunity for Visual Analytics, *Online Journal of Public Health Informatics*, 5(3):e223, 2014

[31] Spotfire:
https://www.tibco.com/products/tibco-spotfire [Accessed: 25th July 2019]

[32] Ola, O, Sedig, K, Beyond simple charts: Design of visualizations for big health data, *Online Journal of Public Health Informatics*, 8(3):e195, 2016

[33] Harwich, E and Lasko-Skinner, R , *Making NHS data work for everyone*, Whitepaper by Reform who is established as the leading Westminster think tank for public service reform, Published by Reform, London, December 2018

[34] NHS Digital:
https://digital.nhs.uk/ [Accessed: 26th July 2019]

[35] UK Government Open Data portal:
https://data.gov.uk/ [Accessed: 26th July 2019]

[36] Mackinlay, J, Kosara, R, Wallace, M, *Data Storytelling: Using visualization to share the human impact of numbers*, Whitepaper about tableau software, 2016

[37] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit, From Visual Exploration to Storytelling and Back Again, *Eurographics Conference on Visualization (EuroVis)*, 35(3), 2016

[38] Hadjar, H, Meziane, A, Gherbi, R, Setitra, I and Aouaa, N, WebVR based Interactive Visualization of Open Health Data, Web Studies, 2, Paris, France. ACM, New York, NY, USA, October 2018

[39] Reyes, E and Labelle, S, Seeing Through the Web: Tools, Practices, Transformations, Web Studies, 2, Paris, France. ACM, New York, NY, USA, October 2018

[40] Data Driven Documents:
https://d3js.org/ [Accessed: 26th July 2019]

[41] Interactive Data Visualization with D3.js:
https://towardsdatascience.com/interactive-data-visualization-with-d3-js-43fc3428a27e
[Accessed: 26th July 2019]

[42] What is Python?
https://www.pythonforbeginners.com/learn-python/what-is-python/
[Accessed: 28th July 2019]

[43] Introduction to Data Visualization in Python:
https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed [Accessed: 27th July 2019]

[44] Interactive Data Visualization:
https://towardsdatascience.com/interactive-data-visualization-167ae26016e8
[Accessed: 27th July 2019]

[45] The Next Level of Data Visualization in Python:
https://towardsdatascience.com/the-next-level-of-data-visualization-in-python-dd6e99039d5e [Accessed: 27th July 2019]

[46] RStudio pricing options: https://www.rstudio.com/pricing/ [Accessed: 23rd August 2019]

[47] RStudio - Data Visualization with ggplot2 Cheat Sheet: https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf [Accessed: 27th July 2019]

[48] Blevins M, Wehbe FH, Rebeiro PF, Caro-Vega Y, McGowan CC, Shepherd BE, et al., Interactive Data Visualization for HIV Cohorts: Leveraging Data Exchange Standards to Share and Reuse Research Tools. *PLoS ONE* , 11(3), Editor: Scarlett L. Bellamy, University of Pennsylvania School of Medicine, UNITED STATES, 2016

[49] Definition of CD4: https://en.wikipedia.org/wiki/CD4 [Accessed: 23rd August 2019]

[50] Ting-Ying Chien, Chong-Yi Chen, Guo-Lun Jin, Hsien-Wei Ting, Disease Trajectory Visualization System Based on Big Data Analytics, *Association for Computing Machinery (ACM)*, ICMHI, Tsukuba, Japan, June 2018

[51] shinyapps.io by RStudio: https://www.shinyapps.io/ [Accessed: 27th July 2019]

[52] Konkoľová, V and Paralič, J, Active learning in data science education, *16th IEEE International Conference on Emerging eLearning Technologies and Applications ICETA 2018*, Starý Smokovec, Slovakia, November 2018

[53] Acute Hospital Activity & NHS Beds Data Release – dated 26th June 2018: https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/2018-06-26/Acute-Hospital-Publication/ [Accessed: 28th July 2019]

[54] Using Shiny for interactive displays of health data: The Scottish Burden of Diseases: https://www.thedatalab.com/tech_blog/using-shiny-for-interactive-displays-of-health-data-the-scottish-burden-of-diseases/ [Accessed: 28th July 2019]

[55] Drug-Related Hospital Statistics, Drug and Alcohol Misuse: https://scotland.shinyapps.io/nhs-drhs-data-explorer/ [Accessed: 28th July 2019]

[56] Prescribing and Medicines, National Therapeutic Indicators: https://scotland.shinyapps.io/nhs-prescribing-nti/ [Accessed: 27th July 2019]

[57] NHS National Services Scotland - Statistical Disclosure Control Protocol - Version 3.0: https://www.isdscotland.org/About-ISD/Confidentiality/disclosure_protocol_v3.pdf [Accessed: 23rd August 2019]

[58] Information Services Division - Publications: https://www.isdscotland.org/Publications/index.asp [Accessed: 29th July 2019]

[59] Publication report - Births in Scottish Hospitals Year ending 31 March 2018. Publication date - 27 November 2018: https://www.isdscotland.org/Health-Topics/Maternity-and-Births/Publications/2018-11-27/2018-11-27-Births-Report.pdf [Accessed: 29th July 2019]

[60] Technical report - Births in Scottish Hospitals. Publication date - 27 November 2018. https://www.isdscotland.org/Health-Topics/Maternity-and-Births/Publications/2018-11-27/2018-11-27-Births-Technical.pdf [Accessed: 29th July 2019]

[61] Publication report - Infant Feeding Statistics Scotland, Financial Year of Birth 2017/18. Publication date - 30 October 2018: https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2018-10-30/2018-10-30-Infant-Feeding-Report.pdf [Accessed: 29th July 2019]

[62] Technical Report - Infant Feeding Statistics Scotland, Financial Year of Birth 2017/18. Publication date - 30 October 2018: https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2018-10-30/2018-10-30-Infant-Feeding-Technical-Report.pdf [Accessed: 29th July 2019]

[63] Infant Feeding in Scotland Tableau Dashboard: https://www.isdscotland.org/health-topics/Child-Health/publications/2018-10-30/visualisation.asp [Accessed: 29th July 2019]

[64] Publication report - Body Mass Index of Primary 1 Children in Scotland School Year 2017/18. Publication date - 11 December 2018: https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2018-12-11/2018-12-11-P1-BMI-Statistics-Publication-Report.pdf [Accessed: 29th July 2019]

[65] Technical Report - Body Mass Index of Primary 1 Children in Scotland, School Year 2017/18. Publication date - 11 December 2018: https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2018-12-11/2018-12-11-P1-BMI-Statistics-Technical-Report.pdf [Accessed: 29th July 2019]

[66] Primary 1 BMI Tableau Dashboard: https://www.isdscotland.org/health-topics/Child-Health/publications/2018-12-11/visualisation.asp [Accessed: 29th July 2019]

[67] UK Statistics Authority - Code of Practice for Statistics: https://www.statisticsauthority.gov.uk/code-of-practice/ [Accessed: 23rd August 2019]

[68] Introduction to R: https://www.datacamp.com/courses/free-introduction-to-r [Accessed: 29th July 2019]

[69] Importing Data in R (Part 1): https://www.datacamp.com/courses/importing-data-in-r-part-1 [Accessed: 29th July 2019]

[70] Importing Data in R (Part 2): https://www.datacamp.com/courses/importing-data-in-r-part-2 [Accessed: 29th July 2019]

[71] Working with the RStudio IDE (Part 1): https://www.datacamp.com/courses/working-with-the-rstudio-ide-part-1 [No longer available to new subscribers]

[72] Working with the RStudio IDE (Part 2): https://www.datacamp.com/courses/working-with-the-rstudio-ide-part-2 [No longer available to new subscribers]

[73] Intermediate R:
https://www.datacamp.com/courses/intermediate-r [Accessed: 29th July 2019]

[74] Data Visualization with ggplot2 (Part 1):
https://www.datacamp.com/courses/data-visualization-with-ggplot2-1
[Accessed: 29th July 2019]

[75] How to Start Shiny tutorial:
https://shiny.rstudio.com/tutorial/ [Accessed: 29th July 2019]

[76] CKAN API Guide: https://docs.ckan.org/en/ckan-2.7.3/api/index.html
[Accessed: 24th August 2019]

[77] CKAN DataStore extension: https://docs.ckan.org/en/latest/maintaining/datastore.html
[Accessed: 24th August 2019]

[78] Publication report - Child Health 27-30 Month Review Statistics, Scotland 2017/18.
Publication date - 9 April 2019:
https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2019-04-09/2019-04-09-Child-Health-27m-review-Report.pdf [Accessed: 29th July 2019]

[79] Technical report - Child Health 27-30 Month Review Statistics, Scotland 2017/18.
Publication date - 9 April 2019:
https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2019-04-09/2019-04-09-Child-Health-27m-review-Technical-Report.pdf [Accessed: 29th July 2019]

[80] Child Health 27-30 Month Review Tableau Dashboard:
https://www.isdscotland.org/health-topics/Child-Health/publications/2019-04-09/visualisation.asp [Accessed: 29th July 2019]

[81] Research reveals why obese mothers less likely to breastfeed:
https://medicalxpress.com/news/2018-03-reveals-obese-mothers-breastfeed.html
[Accessed: 24th August 2019]

[82] Overweight mothers are more likely to stop breastfeeding:
https://medicalxpress.com/news/2018-09-overweight-mothers-breastfeeding.html
[Accessed: 24th August 2019]

[83] New BMI readings for children of different ethnicities:
https://www.sgul.ac.uk/news/news-archive/bmi-ethnicity-children
[Accessed: 24th August 2019]

[84] M T Hudda, C M Nightingale, A S Donin, M S Fewtrell, D Haroun, S Lum, J E Williams, C G Owen, A R Rudnicka, J C K Wells, D G Cook & P H Whincup, Body mass index adjustments to increase the validity of body fatness assessment in UK Black African and South Asian children, *International Journal of Obesity*, 41, 1048–1055 (2017).
https://www.nature.com/articles/ijo201775

[85] The Scottish Attainment Challenge: Equality Impact Assessment results: https://www.gov.scot/publications/equality-impact-assessement-eqia-results-scottish-attainment-challenge/ [Accessed: 2nd September 2019]

[86] Measuring the attainment gap: consultation: https://www.gov.scot/publications/consultation-measuring-attainment-gap-milestones-towards-closing/#t1 [Accessed: 24th August 2019]

[87] Comparison of Tableau, R's Shiny and Google's Data Studio: https://www.homeagency.co.uk/hub/battle-of-the-dashboards/ [Accessed: 24th August 2019]

[88] Supermetrics for Data Studio Pricing: https://supermetrics.com/pricing/data-studio [Accessed: 2nd September 2019]

# Appendix 1

**Email 1**

Copy of an e-mail from Jonathan Cameron, Head of Service - Strategic Development, Public Health & Intelligence (PHI), NHS National Services Scotland, to Natalie Polack on 17th July 2019. The e-mail address of the sender has been removed.

CAMERON, Jonathan (NHS NATIONAL SERVICES SCOTLAND)

Wed 17/07/2019 15:58

Hi Natalie

It is a little difficult to be precise on this as we also host the system for other organisations such as Scottish Government but could easily say that an investment has been made of £350k in Tableau by NSS across multiple years.

No problem to include this figure in any write up.

Kind regards

Jonathan

**Email 2**

Copy of an e-mail from Jonathan Cameron, Head of Service - Strategic Development, Public Health & Intelligence (PHI), NHS National Services Scotland, to Natalie Polack on 26th August 2019. The e-mail address of the sender has been removed.

CAMERON, Jonathan (NHS NATIONAL SERVICES SCOTLAND)
Mon 26/08/2019 09:45

Hi Natalie

NHS have been using Tableau for 5 years now.

Not sure if we have published any D3 visualisations outside of the organisation – I think we have only done this internally as early proof of concept. R Shiny is being preferred as it is easier to use and links in with the new R platform that we have.

Hope to see you soon.

Kind regards

Jonathan

# Appendix 2

This page contains a list of web pages, blog posts, and questions and answers from forums that were useful when writing code for this dissertation.


[1] DataCamp tutorial: How to install R on Windows, Mac OS X and Ubuntu that was used for downloading R, RStudio and installing the Tidyverse: https://www.datacamp.com/community/tutorials/installing-R-windows-mac-ubuntu#comments [Accessed: 24th August 2019]

[2] Explains how to use Source to run a script: http://www.rexamples.com/8/How%20to%20run%20the%20code [Accessed: 24th August 2019]

[3] jsonlite and httr packages for using an API: https://www.programmableweb.com/news/how-to-access-any-restful-api-using-r-language/how-to/2017/07/21 [Accessed: 24th August 2019]

[4] Introduction to R Shiny: https://medium.com/@ODSC/introduction-to-r-shiny-b6acdf17c963 [Accessed: 24th August 2019]

[5] Downloadable ggplots in shiny: https://ildiczeller.com/2018/02/11/downloadable-ggplots-in-shiny/ [Accessed: 24th August 2019]

[6] Run a plotting function and save the output as a PNG: https://shiny.rstudio.com/reference/shiny/0.11/plotPNG.html [Accessed: 24th August 2019]

[7] Join in R: How to join (merge) data frames (inner, outer, left, right) in R: http://www.datasciencemadesimple.com/join-in-r-merge-in-r/ [Accessed: 24th August 2019]

[8] Use file.path: https://stackoverflow.com/questions/12650260/using-paste-to-construct-windows-path-in-r [Accessed: 24th August 2019]

[9] Bar and line graphs (ggplot2): http://www.cookbook.com/Graphs/Bar_and_line_graphs_(ggplot2)/ [Accessed: 24th August 2019]

[10] R ggplot - Error stat_bin requires continuous x variable: https://stackoverflow.com/questions/34428440/r-ggplot-error-stat-bin-requires-continuous-x-variable [Accessed: 24th August 2019]

[11] Position = 'dodge' and stat = 'identity': https://github.com/tidyverse/ggplot2/issues/2229 [Accessed: 24th August 2019]

[12] ggplot2: Changing the Default Order of Legend Labels and Stacking of Data: https://learnr.wordpress.com/2010/03/23/ggplot2-changing-the-default-order-of-legend-labels-and-stacking-of-data/ [Accessed: 24th August 2019]

[13] Drop bars from a plot using filter: https://stackoverflow.com/questions/44054876/selectively-drop-bars-from-bar-plot-without-changing-formatting [Accessed: 24th August 2019]

[14] Data tidying - how to use gather: https://tidyr.tidyverse.org/articles/tidy-data.html [Accessed: 24th August 2019]

[15] Using filter with two != conditions: https://stackoverflow.com/questions/47191727/r-filtering-by-two-columns-using-is-not-equal-operator-dplyr-subset [Accessed: 24th August 2019]

[16] Using mutate: https://dplyr.tidyverse.org/reference/mutate.html [Accessed: 24th August 2019]

[17] Joining 2 R data sets with different column names: https://www.r-bloggers.com/joining-2-r-data-sets-with-different-column-names/ [Accessed: 24th August 2019]

[18] Renaming levels of a factor: http://www.cookbook-r.com/Manipulating_data/Renaming_levels_of_a_factor/ [Accessed: 24th August 2019]

[19] What exactly IS an API? https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f [Accessed: 24th August 2019]

[20] Top 50 ggplot2 Visualizations - The Master List (With Full R Code) http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html [Accessed: 24th August 2019]

[21] NHS R Resources: https://scotland.shinyapps.io/nhs-r-resources/ [Accessed: 24th August 2019]

[22] R style guide for PHI analysts: https://github.com/Health-SocialCare-Scotland/R-Resources/blob/master/PHI%20R%20style%20guide.md [Accessed: 24th August 2019]

[23] Example of Shiny app: https://github.com/TheDataLabScotland/Public_ScotGovAccelerator_2018/blob/master/BurdenOfDiseasesShiny/app.R [Accessed: 24th August 2019]

[24] Plotting the same output in two tabPanels in shiny: https://stackoverflow.com/questions/44205137/plotting-the-same-output-in-two-tabpanels-in-shiny?rq=1 [Accessed: 24th August 2019]

[25] Shiny: Re-using the same plot in multiple tabs is not working: https://stackoverflow.com/questions/52968020/shiny-re-using-the-same-plot-in-multiple-tabs-is-not-working [Accessed: 24th August 2019]

[26] How To Create A Pie Chart In R Using ggplot2: https://www.datanovia.com/en/blog/how-to-create-a-pie-chart-in-r-using-ggplot2/ [Accessed: 24th August 2019]

[27] Joining 2 R data sets with different column names: https://www.r-bloggers.com/joining-2-r-data-sets-with-different-column-names/ [Accessed: 24th August 2019]

[28] ggplot Axis Ticks: Set And Rotate Text Labels: https://www.datanovia.com/en/blog/ggplot-axis-ticks-set-and-rotate-text-labels/ http://www.rpubs.com/dvdunne/reorder_ggplot_barchart_axis [Accessed: 24th August 2019]

[29] Pie Charts in ggplot2: https://www.r-bloggers.com/pie-charts-in-ggplot2/ [Accessed: 24th August 2019]

[30] R - ggplot pie chart with facet_wrap: https://stackoverflow.com/questions/40088819/r-ggplot-pie-chart-with-facet-wrap [Accessed: 24th August 2019]

[31] ggplot2 and Cowplot - Easy way to mix multiple graphs on the same page: http://www.sthda.com/english/wiki/wiki.php?id_contents=7930 [Accessed: 24th August 2019]

[32] Plot two lines on one graph in ggplot and R: https://rpubs.com/euclid/343644 [Accessed: 24th August 2019]

[33] Shapes and line types: http://www.cookbook-r.com/Graphs/Shapes_and_line_types/ [Accessed: 24th August 2019]

[34] ggplot2 barplots : Quick start guide - R software and data visualization: http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization#change-fill-colors [Accessed: 24th August 2019]

[35] Change color of only one bar in ggplot: https://stackoverflow.com/questions/22894102/change-color-of-only-one-bar-in-ggplot [Accessed: 24th August 2019]

[36] ggplot legends - change labels, order and title: https://stackoverflow.com/questions/12075037/ggplot-legends-change-labels-order-and-title [Accessed: 24th August 2019]

[37] Colors (ggplot2): http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/ [Accessed: 24th August 2019]

[38] Cowplot: Shared legends: https://wilkelab.org/cowplot/articles/shared_legends.html [Accessed: 24th August 2019]

[39] Plotting an average line in ggplot2: https://stackoverflow.com/questions/48013633/plotting-an-average-line-in-ggplot?rq=1 [Accessed: 24th August 2019]

[40] How to put exact number of decimal places on label ggplot bar chart: Put % above bars: https://stackoverflow.com/questions/38369855/how-to-put-exact-number-of-decimal-places-on-label-ggplot-bar-chart [Accessed: 25th August 2019]

[41] Showing data values on stacked bar chart in ggplot2: https://stackoverflow.com/questions/6644997/showing-data-values-on-stacked-bar-chart-in-ggplot2 [Accessed: 25th August 2019]

[42] How to write the first for loop in R: https://www.r-bloggers.com/how-to-write-the-first-for-loop-in-r/ [Accessed: 25th August 2019]

[43] Call R script from Shiny app: https://stackoverflow.com/questions/44524619/call-r-script-from-shiny-app [Accessed: 3rd September 2019]

[44] "source" can't read the other file: https://community.rstudio.com/t/source-cant-read-the-other-file/18570 [Accessed: 3rd September 2019]

[45] Error: unexpected datatables/htmlwidget output: https://community.rstudio.com/t/error-unexpected-datatables-htmlwidget-output/16521/2 [Accessed: 3rd September 2019]