



UNIVERSITY OF
LEICESTER

Department of Informatics University of Leicester

CO7201 Individual Project

Mine, Analyse and Visualise Healthcare Data

Shashidar Ette

se146@student.le.ac.uk

169050065

MSc Advanced Software Engineering

Final Report

Project Supervisor: Prof Reiko Heckel

Second Marker: Dr Rayna Dimitrova

Submitted: 18th May 2018

Word count: 9956

DECLARATION

All sentences or passages quoted in this report, or computer code of any form whatsoever used and/or submitted at any stages, which are taken from other people's work have been specifically acknowledged by clear citation of the source, specifying author, work, date and page(s). Any part of my own written work, or software coding, which is substantially based upon other people's work, is duly accompanied by a clear citation of the source, specifying author, work, date and page(s). I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this module and the degree examination as a whole.

Name: Shashidar Ette

Date: 18th May 2018

Abstract

In the modern world, one of the main challenges for organizations and communities is the huge volumes of data available. In past decade, with data being made publicly available via open platforms, the challenge has increased manifold. With the advent of technology, governments are working towards digitization and providing effective healthcare to comply with set processes and policies. With their implementation, vast amounts of data are captured at several levels. This data can contain vital insights and clues for the decisions to be made for a better future. However, the systems available can only be used view the data in isolation and there is a need to develop a methodology, models and systems to explore the data in broader perspective. There is a dire necessity of systems which can help key personnel and stakeholders within an organization to explore the data effectively, independently to assist in quick decision making and help them to induce an amendment or propose new policies.

This project dwells into this challenge of high volumes of data for healthcare domain. Specifically, the focus is on practice level prescription data available from NHS. With the help of various data mining and exploration techniques, it investigates various healthcare datasets. The initial focus is to discover correlations between them, followed by in-depth analysis to explore the hidden insights in a broader spectrum. The report also details the data model for per-patient cost and the challenges. It is followed by the details of the web application prototype developed, its design and architecture based on the domain knowledge accumulated as part of the project. The report concludes with the learnings, outcomes and details the future developments.

Acknowledgements

I want to express my sincere gratitude and thanks to Prof. Reiko Heckel for his guidance, encouragement and constant supervision. I would like to make a special mention of the open discussions we had about the problem domain and project in general. Your constant feedback helped me to deliver better on the project.

I sincerely thank Dr. Rayna Dimitrova, for her feedback on the project and guidance which helped to improve the prototype.

My thanks and appreciation also go to Prof. Umesh Kadam, College of Life Sciences, Diabetes Research Centre, for providing me with the domain knowledge and help me to get into the field of prescriptions and medicines.

Your support to get to meet the team in East Leicester CCG to portray my work and share the knowledge is inspirational.

To acknowledge the sense of achievement and satisfaction, Thank You Prof. Reiko Heckel and Prof. Umesh Kadam for giving an opportunity to work on this project. It helped me gain knowledge and experience in the field of Data Science. I hope the outcome met the expectations and I truly believe it can be further developed to a stage which we envisioned at its inception.

Table of Contents

Introduction	7
Aims and Objectives	7
Project Overview	9
Background.....	9
Healthcare Data.....	9
Data Sources	11
Related Work.....	15
Data Analytics.....	16
Anaconda Data Science Platform.....	16
Python	17
SciPy	17
Pandas.....	18
Numpy	18
Jupyter notebooks	18
Orange	18
Machine Learning	19
Scikit.....	20
Orange	20
Data Visualization	21
Plotly	24
2D Visualization.....	24
3D Visualization.....	25
Web Technologies.....	25
Wireframes using Draw.io	25
HTML	26
Javascript.....	26
jQuery	26
Bootstrap	27
JSON	27
Ajax	28

Azure	28
Visual Studio	28
Learning	29
Requirements	30
Data Mining	30
Software.....	31
Approach	32
Data Exploration	33
Datasets linking.....	33
GP Profile data	34
Prescription costs data.....	36
Public Health Indicators & Index of Deprivation Measures.....	37
Combining Datasets	38
Jupyter notebooks.....	45
Data Model.....	45
Technical Specifications	48
Application prototype.....	48
Wireframes.....	49
Final Prototype	51
Conclusion	53
Project	53
Challenges.....	54
Future Developments.....	54
Glossary	55
References	56

Introduction

This document is a final report generated as part of MSc individual project work for the CO7201 module. The document details the project's main aim and objectives. It then discusses the background and work done related to the topic under consideration. As part of the background, it covers several aspects such as published articles related to the healthcare data & visualization, existing systems, technologies or frameworks available for data analysis and visualization. Later different project and software requirements are discussed. It also details the status of the requirements and challenges faced to deliver them.

The primary focus is on practice level prescription data available from NHS digital. Next section, gives a detailed account of the approaches considered for data exploration, analysis and visualization. Further to the approach, the next part details the design and architecture of the web application. It is followed by the brief conclusion and future developments identified by the author. One thing to note is the project involves the application of data science principles and web application development principles. The author is a software engineer by experience, it is the project is an opportunity for the author to apply data science principles learnt, discover the insights and discuss the findings.

Aims and Objectives

The aim of the project is to analyze prescription data and correlate with other healthcare datasets available in public. The report details the datasets available and the ones considered for analysis as part of the project.

The initial objective is to explore the datasets available and then build on top of it to find the prescription behaviours by General Practices and discover rationale. The other core objective is to visualize the prescription

costs data along with other datasets. The aim is to help the personnel in clinical commission groups (CCG) in elaborate analysis and decision making.

The project takes a phased approach and focuses on drug groups provided for Diabetes disease to generate insights for specific behaviours between CCGs or GPs. The outcome of the project is an understanding of the datasets independent of GPs with their prescription data i.e. it details the independent and dependent variables linked to GP prescription data. The domain knowledge of the data and their insights form are core part of project's outcome. The insights discovered during the project are detailed. Based on the datasets and their correlation, further learning about the data is done using cluster analysis. Each of the clusters formed is analyzed individually to discover a general practice's behaviour around prescriptions and related costs.

The application developed provides an intuitive way to look at a general practice with aspects of patient size profile, health profile and prescription profile with respect to the specific drug groups linked to Diabetes. The application-built forms a base framework or a platform to incorporate newly related datasets to give a combined view of the correlations formed.

The report details the challenges faced in the process of applying data science principles especially in the areas of finding independent and dependent variables to find out specific segmentation amongst General Practices. It also highlights the challenges faced and changes in initial scope and actual outcomes from the software application built.

In the end, the report details the aspects to be developed and extended in future. The insights of domain knowledge learnt can be further used with

other related datasets in future to extend the model and broaden the application of them.

Project Overview

Background

Healthcare Data

Historically the organisations within healthcare domain were bound by policies set by the government. Due to the compliance measures, they had to maintain records in several forms leading to huge volumes of data. This data can be

- Used by clinical groups to find cost-effective ways to treat patients.
- Used by public health to discover disease patterns, predict and respond to disease outbreaks proactively.
- Help GPs to do patient profile analysis to predict lifestyle changes and prescriptions required in advance. However, the latest trend is data digitization which requires developing systems to explore and discover the data in a systematic way. [\[1\]](#)

For this project, the focus is on the practice level prescription data. The volumes

Of the data under consideration is approximate ~1 to 1.25 GB per month (with ~20 million records of prescription data). Considering we have multiple datasets its important follow a systematic way. The steps broadly include define questions or core concern areas, define & explore the datasets, discover insights, build a methodology and develop the application.

Second important aspect is to find how does prescription costs vary from one practice to another. Based on the study by a team of researchers in Korea [\[2\]](#). A prescription for a general disease such as influenza varies based on the patient's condition. Same is applicable for the variety of diseases in

general. It is also possible that several diseases might have a common set of prescription patterns. The study was to uncover these prescription patterns from a set of data to find a relationship between diseases and medicines. It also states that prescription patterns can be linked to other factors such as patient age and other independent factors. As part of this project, in addition to the prescription dataset there are other datasets related to health indicators and deprivation indicators which are also considered.

The third aspect of the prescription is the usage of generic and branded medicines. The trend of medical practices is to switch back from branded to generic drugs [\[3\]](#). As part of the study conducted by researchers in the USA, a patient population which was using branded drugs for certain group of medicines were identified when they switched to a generic medicine soon they were introduced. A switchback is an option available to choose a specific medicine which has same ingredients and formulae available in branded as well as generic medicines. As part of the study, the patients who opted for generic medicines were followed to switch to branded ones. As an outcome “adjusted switchback rates were consistently lower for patients who switched from branded to authorized generic drug products compared with branded to generic drug products in the primary cohort (pooled hazard ratio 0.72, 95% confidence interval 0.64 to 0.81)” [\[3\]](#). It will be essential to understanding the usage of generic medicines versus branded medicines and their impact on overall prescription costs of medical practices. In line with this study, CCGs within NHS provide a “Recommendations for appropriate prescribing of generic and branded Medicines” [\[12\]](#) which acts as a guideline for general practitioners.

The fourth aspect is whether the competition between the GPs has any impact prescriptions. Based on the study by Catherine Schaumans, TILEC, CentER, Tilburg University, Netherlands. It is possible that a GP can

prescribe all the medicines need from the pharmacy including over-the-counter as well as prescription medicines. The probable reason for this behaviour is to meet the patient's expectations. Another aspect is the volume of prescriptions, the study mentions that "The higher the percentage of female GPs, the lower number of GP contacts (which is in line with female GPs being more likely to work part-time) and the higher the percentage GPs with a lot of experience (>20 years), the lower the consumption of primary care. Finally, the more GPs perform home visits, the higher is the consumption of care" [4]. As part of the conclusion, there are several aspects to be considered for the behaviours such as the volume of prescriptions, GP competition and GP characteristics.

Data Sources

A list of data sources considered for the project are as below:

1. NHS Digital – This is a digital platform from NHS providing information systems including infrastructure for health and care. It provides several datasets in relation to prescriptions. However only below two datasets are considered for this project:

- Practice Level Prescribing in England
- Clinical Commissioning Group (CCG) prescribing data

This data contains practice level data for each month within England. "For each practice in England, the following prescription information is presented at presentation level for each medicine, dressing and appliance that has been subsequently dispensed in the community in the UK:

- Items - the total number of items prescribed and dispensed
- NIC - the total net ingredient cost
- ACT - the total actual cost
- Quantity - the total quantity (in terms of number of tablets or millilitres, for example)" [NHS Digital]

For each month's data released (generally after 2 months of actual data is ready), comes with three distinct files as below:

- “Practice Prescribing Data file” – it is the primary data file of focus with the practice level prescribing data at presentation level using the full 15-digit BNF code. This data file is larger size ~1 GB with more than 1.5 million records.
- “GP prescribing chemical substance file” – it gives the chemical substance or section as appropriate using the 9-digit (Drugs) or 4-digit (Appliances) BNF code.
- “GP prescribing practice address file” - this is a lookup data with details of the practice name and address.

As of now, GP prescribing chemical substance and GP prescribing practice address are considered as available from NHS BSA. Only the latest datasets are considered.

2. NHS Business Service Authority's Information Portal – This data source provides additional details of the practices and medical formulations i.e. BNF. As a system it has “reports and data to help NHS customers track trends, inform decisions and support policy”. Below datasets are considered:

BNF Code Information:

BNF Code is a unique code given to a prescription. This is a 15-digit code sub-divided into 8 sub-sections:

Tradorec XL Tablets 300mg

Chapter	Section	Paragraph	Sub paragraph	Chemical substance	Product	Strength and formulation	Generic equivalent
04	07	02	0	40	BI	AC	AM
Central Nervous System	Analgesics	Opioid Analgesics	Opioid Analgesics	Tramadol Hydrochloride	Tradorec	Tradorec XL_Tab 300mg	

Tramadol HCl Tablets 300mg (generic)

Chapter	Section	Paragraph	Sub paragraph	Chemical substance	Product	Strength and formulation	Generic equivalent
04	07	02	0	40	AA	AM	*
Central Nervous System	Analgesics	Opioid Analgesics	Opioid Analgesics	Tramadol Hydrochloride	Tramadol HCl	Tramadol HCl_Tab 300mg M/R	

refers to

'AA' here always means 'generic'

Figure 1 BNF Code Structure (source: <https://ebmdatalab.net/prescribing-data-bnf-codes/>)

In total there are 23 chapters with each of them linked to a specific body function. Each of the chapters is divided into section, paragraph and sub-paragraphs.

- Patient List Size Information:

This data consists of general information about a GP such as a region information, its name, its code and more importantly age distribution of the patients as below:

General Information	Patient Age Distribution
Regional Office Name	Male 0-4
Regional Office Code	Female 0-4
Area Team Name	Male 5-14
Area Team Code	Female 5-14
PCO Name	Male 15-24
PCO Code	Female 15-24

Practice Name	Male 25-34
Practice Code	Female 25-34
	Male 35-44
	Female 35-44
	Male 45-54
	Female 45-54
	Male 55-64
	Female 55-64
	Male 65-74
	Female 65-74
	Male 75+
	Female 75+

From the patient age distribution, total patient list size of GP is calculated.

- It also provides Detailed Prescribing Information as offline data (same as – NHS Digital Prescription data)

3. Public Health Data England or FingerTips data:

This is an important data source with information on disease prevalence and other health profile indicators for England. As a system, it provides profiles which “are a rich source of indicators across a range of health and wellbeing themes that have been designed to support JSNA and commissioning to improve health and wellbeing and reduce inequalities. With these profiles you can:

- Browse indicators at different geographical levels
- Benchmark against the regional or England average
- Export data to use locally” [<https://fingertips.phe.org.uk/>]

This dataset is used to link the prescription with the outcomes and effectiveness of health care policies. From this dataset following indicators are considered:

Indicator ID	Indicator	Definition
241	Diabetes: QOF prevalence (17+)	It is the percentage of patients aged 17 years+ and with diabetes mellitus, as recorded on practice disease registers.
219	Hypertension: QOF prevalence (all ages)	The percentage of patients with established hypertension, as recorded on practice disease registers (proportion of total list size).
91872	Deprivation score (IMD 2015)	Index of Multiple Deprivation Measure

There are several other indicators which can be considered in future as necessary.

Related Work

This section details the systems developed around prescription data.

1. OpenPrescribing –

This is a system developed in collaboration between several partners namely The Health Foundation, University of Oxford, NIHR and others. It also encompasses other data sources. This system provides dashboards from the perspective of a CCG and Practices. It also provides users to analyse specific formulations of interests within a CCG or Practice. It also provides a public open API -

<https://openprescribing.net/api/>

2. <http://www.prescribing.info/>

This is another system built using Microsoft BI. This primarily focuses on prescription data and provides a drill-up and drill down views which is an important reference for the project.

However, the systems above do not correlate the prescription data with other datasets related age or diseases. This is crucial from the clinical questions under consideration. The need is to have a system which will allow the CCG personnel to visualize the data in a combination of different profiles.

Data Analytics

The core of this project is data mining and analysis of prescription dataset and others under consideration. As a choice there were a lot of tools/frameworks to select from, some of them are: Hadoop, MapReduce, Weka or R etc. However, the author as a preference has chosen Anaconda data science platform as a framework of choice. The next section will detail the components or libraries available as part of the platform with reasons of using for the project.

Anaconda Data Science Platform

Anaconda is one of the most popular data science platforms based on Python.

It allows to setup several environments in a seamless and integrated manner

The version is 1.6.5.

It was selected due to the reasons below:

- Author's personal interest to learn and work using Python as a programming language
- The courses attended by the author used Python and frameworks extensively.

- More importantly, notebooks can be used to collaborate the ideas within the team. This feature was crucial to get inputs and guidance from stakeholders involved in the project.
- It provides a platform of libraries essential for analysis, visualization and machine learning. Below sections will detail libraries chosen for the project.

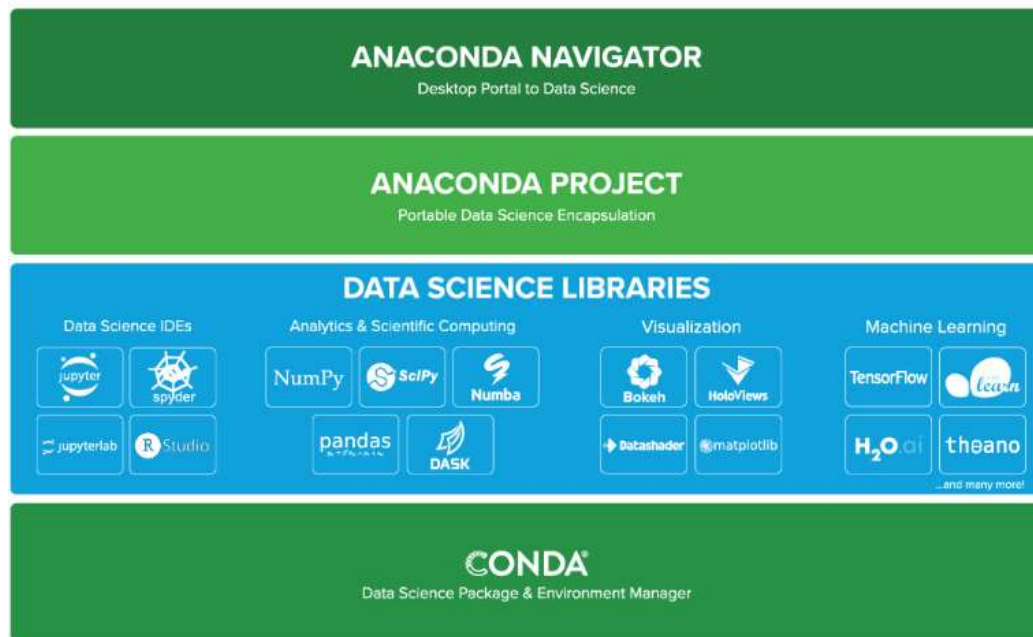


Figure 2: Anaconda Data Science Platform

(source: <https://www.anaconda.com/distribution/>)

Python

Python is used as a general-purpose programming language for data exploration activities. The version is 3.6.3.

SciPy

This is a python based eco-system. It encompasses following core libraries. It also provides a library for scientific and mathematical computing. The version is 0.19.1.

Pandas

This is powerful data analysis toolkit based on the platform. It provides a high-performance library to load the large volume of data in an efficient way. It allows the importing huge volumes of datasets from various sources and perform various operations on the data in the form of panda series and data frames.

It provides functions to slice, merge or cleanse the data. In general, it has support for time-series data, visualizations, data variety support, data integration, data transformation, descriptive statistics.

The version is 0.22.0.

Numpy

This library is useful in processing for numbers. It complements the Pandas toolkit for advanced operations on data being processed. It provides speed and functionality with multidimensional arrays known as “ndarray” and built-in array operations. The version is 1.13.3

Jupyter notebooks

This is a web-based, interactive computing notebook environment. It provides collaboration platform to share ideas as python code, markup, documentation and visualizations. In general, it can be used for documented data Science with markdown, presentation of results as HTML or pdf, in addition to python it supports Julia, and R. The version is 5.0.0.

Orange

This is workflow-based data analysis and visualization framework. It provides a component-based data mining framework with interactive workflows and an extensive toolbox for broader data science activities.

Most of the data analysis activities are done using Jupyter notebooks and libraries available. However, there are instances where a flow-based approach needs to be taken to analyse a problem. In the project, it used it to:

- perform chapter based on cluster and subchapter categorization for prescription costs in a generic way.
- to develop a data model for test and validation for per patient cost prediction. The version is 3.11.0.

Machine Learning

Next core activity of the project is to learn from the data aggregated from the various datasets. It requires the application of the various machine learning algorithm to process and discover insights. The project has two distinct aspects to be covered. The first being to distinguish practices based on selected independent variables and find a way to perform segmentation or distribution. The second is a process of generating a data model to discover a correlation between certain dependent variables to predict a target which is related to prescription cost. Thus, the project involves both supervised and unsupervised learning required from the data.

Supervised learning is a concept where the target variable is known. The process to generate a data model which could train the data to learn about the target variable and predict the output for any new data. In case of supervised learning, to build a robust data model a trained data is used to build the model and make it learn. The robustness of the model is validated with the test data and as necessary the model is fitted to be considered for prediction. With the help of this learning, the model built can make predictions for the new data and generate the target. Some of the techniques are:

- Regression – used to predict a numeric or continuous value given a set of input variables for example: predicting stock price
- Classification – used to categorize the data and decide to which it belongs to. For example, classify whether the day is sunny, or cloudy or stock price of a company will grow or dip.

Unsupervised or representational learning is a concept where the primary focus is to explore the data and learn from it. Unlike supervised learning, the target variables are unknown. As part of this technique, we rely on various machine learning algorithms and use them to discover patterns from the data. Some of the techniques are:

- Clustering – process data and form groups of it based on given independent parameters. For example: customer segmentation, grouping counties or organizations.

Scikit

This library provides set of modules for machine learning and data mining. The framework is built on top of NumPy, SciPy and matplotlib. Although it can be set for both types of learning, in this project context it is used for unsupervised learning only. As part of the project it is used within Jupyter Notebooks package with Anaconda framework. The version is 0.19.1.

Orange

Orange as a framework also provides components for applying machine learning to the data. As part of this project, this tool is used for supervised learning of the data in the form of workflow. Some of the objects used for data model are:

- Rank – ranking features based on the correlation, assists in feature selection

- Data selection – a selection of data based on columns or row filters.
- Data sampler – sample the data into training and testing sets.
- Model objects for learning the data based on Linear Regression, Support Vector Machine and Nearest Neighbour – Machine learning models to learn from the sample data.
- Test & Score – Score the results performed by models and assist in choosing the model for the problem domain.
- Predictions – Represents the model underneath to validate the test data and model itself.

Data Visualization

The next vital part of this project is data visualization of healthcare data.

In a study[\[5\]](#) the focus was prescription behaviour for specific physicians, medicines and set of diseases. The primary entities of the study were the relationship between physician, medicine and patient. Although the work is related electronic health record visualization, the main aim was to find out the problems faced in providing user interfaces/visualizations which are intuitive and are effective for physicians. The objective was to help them understand their behaviours when providing prescriptions at different stages of consultation.

Based on the findings, they had concluded set of requirements such as:

“the system must show multivariate data”

“the system must show data from multiple perspectives” to facilitate different viewpoints

“the system must enable to select entities or a time interval of interest to get relevant statistics about this selection”

“the system must facilitate comparison between these entities”.

As a solution, it also mentions about “Three table view” which gives a combined view of physicians, patients and medicines. The above

requirements and “three tabled view” (3TV) forms a good base for this project.

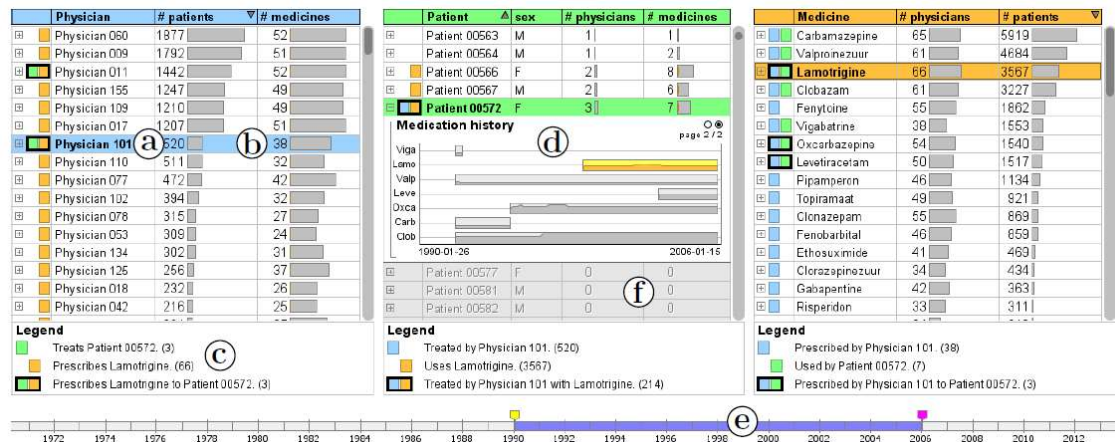


Figure 3 Three tabled view [6]

In a paper [6], the focus was to use “time-oriented cartographic treemaps” to show multivariate data which is hierarchical and uses space available in the most optimized way. It uses the results from the study to explore trends of health diagnoses overlaid in geospatial view. It exploits both static and animated view.

In the below view, the conventional treemap is used at geographic map level to show the data trends at CCG level.

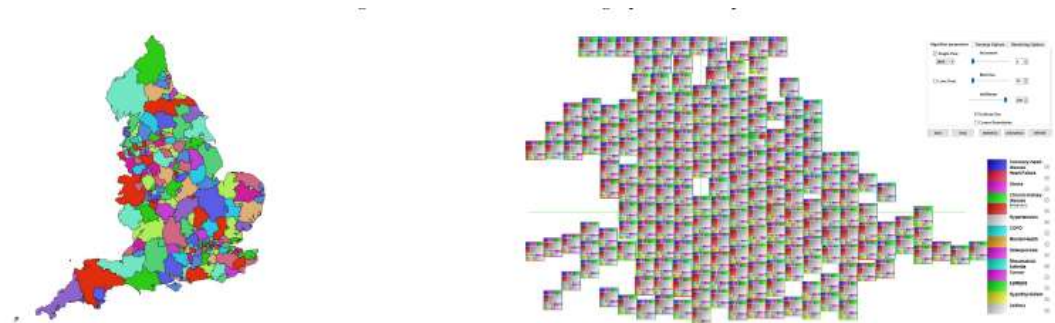


Figure 4 Cartographic Treemap is covering more space as compared to a conventional map. [6]

This is a useful reference to use the conventional visualization technique to use along with space and time complexities thus providing dynamic content.

In terms of recent advances in data visualization, a survey [7] discusses the information visualization research area which aims to help users to explore, understand the data via visual exploration. The paper discusses, a “visualization pipeline”

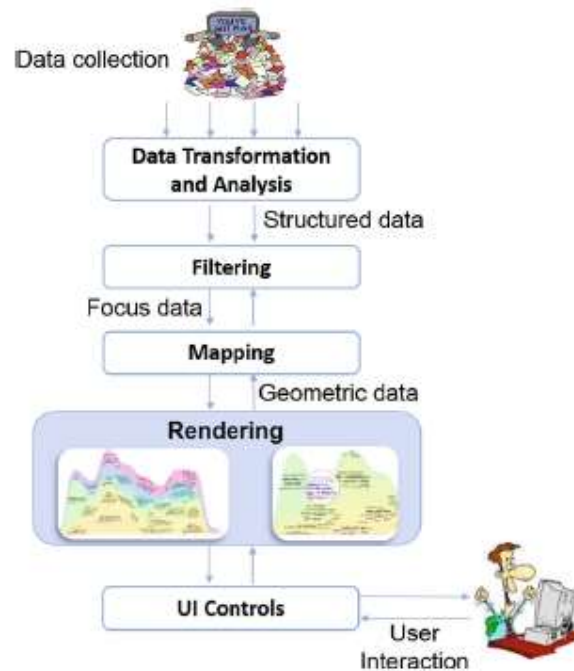


Figure 5 Visualization pipeline [7]

It suggests a structured way to handle the visualization based on the user interaction. The steps include “data transformation”, “filtering”, “mapping” and “rendering” on the UI controls. In terms of architecture for the project, the data transformation and filtering should be handled by the systems which are aware of the structure of the data whereas mapping and rendering should be handled by the layer responsible for visualization or presentation.

The survey also discusses several frameworks such as “InfoVis”, “Prefuse” etc. which are based on JavaScript and provide interactive visualizations. However, the document-driven document (D3) library is a recent and a popular web-based library. It provides direct manipulation of the document

which updated the web element and drives the visualization. Using this library, the application can provide folding interactions, theme river, Stacking-based trajectory visualizations to name a few. From a development perspective, D3 is available in JavaScript and comes with a learning curve but can be incorporated into modern web application frameworks.

Plotly

Plotly is a web visualization library built on top of D3.js. It is an open-source JavaScript library used to create sophisticated and visual interactions.

As a package, it provides several 2D and 3D visualization components with interactive features such as zoom, selection, save which are additions to the D3.js library. For the project, although D3.js was the initial choice for providing visualization. In addition, Plotly.js provides extensive user documentation and very active developer forum. This is key to reduce the learning curve and rapid prototype development. Due to these reasons, Plotly.js is selected to develop the application prototype for the project.

2D Visualization

Under this category, the data is visualized as scatter plots, pie, radar and spider charts. In addition to Plotly, the 2D visualization was used in following tools/frameworks:

- in Jupyter notebooks, “matplotlib” and “seaborn” libraries are used for visualization.
- The orange framework is also used as necessary.
- MS Office Excel using its extensive 2-D charts

3D Visualization

In most situations 2D visualization suffices when the number of dimensions is limited. When dealing with multiple datasets within a problem domain, it's essential to present the data as a rich source of information. In which case, showing data in the 2D plane will not be enough, hence the 3-D visualization is required. The Plotly.js provides a way to visualize the data in the form of 3D scatter plots and 3D surface plots. In this project, 3-D scatter plot is used to visualize the cluster analysis of GPs based independent variables. The 3-D visualization makes the clustering profound with the help of zooming, filtering and transformation in a 3-D space.

Web Technologies

In addition to data mining and analysis, a web application is essential to evaluate the results of the data, experiment it. An application will then enable the end users or external stakeholders such as clinical commissioning personnel and domain experts to explore the outcomes and share feedback. The below sections will detail the technologies selected for the application prototype.

Wireframes using Draw.io

A wireframe is a layout detailing the main elements, content and description of a user interface.

This project is executed in an iterative model involving continuous feedback primarily focusing on data mining and analysis aspects. The requirements for the application prototype were not decided at the time of project's inception. Hence, wireframes were created for the application to propose a look & feel of it for review and feedback within the team.

As part of UI design, “draw.io” was used. This is an online design and collaboration tool. Although it provides ways to generate software designs such as UML and ER Diagrams. For the project, it was for mock-creation. As a tool it provides options to create the user interfaces using available libraries such as HTML elements and Bootstrap components. The mocks created were shared in the form of HTML and pdf within the team.

HTML

HTML stands for hypertext mark-up language. HTML5 is a markup language used for structuring and presenting content on the World Wide Web.

It's matured standard with evolving extensions for handheld devices with different screen resolutions based on different platforms.

Javascript

JavaScript is a high-level, dynamic, un-typed, and interpreted programming language. It has been standardized in the ECMAScript language specification. Alongside HTML and CSS, JavaScript is one of the core technologies used to create World Wide Web content production; the majority of websites employ it, and all modern Web browsers support it without the need for plug-ins. JavaScript is prototype-based with first-class functions, making it a multi-paradigm language, supporting object-oriented, imperative, and functional programming styles. It has an API for working with text, arrays, dates and regular expressions.

jQuery

jQuery is a cross-platform JavaScript library designed to simplify the client-side scripting of HTML. jQuery is the most popular JavaScript library in use today, with installation on 65% of the top 10 million

highest-trafficked sites on the Web. jQuery is free, open-source software licensed under the MIT License.

jQuery's can be used for easy navigation, modification and updation HTML in an intuitive way. It helps in developing modular HTML based applications with dynamic content in combination with JavaScript.

Bootstrap

Bootstrap is an open source toolkit for developing web applications based on HTML, CSS, and JS. It helps to develop prototype quickly using rapid application development approached. It provides several powerful inbuilt components and plugin built on jQuery technology. It is a framework of chouce to build responsive, mobile-first projects on the web which makes it the world's most popular front-end component library.

JSON

“JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the [JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999](#). JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.” [\[18\]](#)

Ajax

AJAX stands for Asynchronous JavaScript and XML. It uses XMLHttpRequest i.e XHR to communicate with service side components and scripts. It supports information in a variety of formats some of them include JSON, XML, HTML, CSV and even raw files. Its main characteristic is the feature of asynchronous communication which enables developers to build dynamic content for full page or partial page contents. [17]

Ajax is used in this project for making requests to the server without reloading the page, receiving and working with data from the server.

HTML, JavaScript, AJAX and jQuery are industry standards, advanced and mature technologies. Thus making this set toolset of choice for the project.

Azure

From a deployment perspective, the websites and web-services will be deployed in the Azure environment. For the project, the Azure account provided via Microsoft Imagine licence of the University is used. However, when a decision to deploy publicly appropriate environment needs to be selected.

Visual Studio

Visual Studio from Microsoft is one of the most popular IDE (IDE) availability in the market. It is available in various editions, which are licensed and have a cost. For this project, Visual Studio 2015 community edition is used. As an IDE it can be used to develop desktop, web and mobile applications. It comes with Microsoft Azure support to create and publish applications seamlessly from the IDE.

Learning

This section details the areas of learning achieved or attempted by the author for the project. The author is a software developer (C# and Java) by experience, however the area of Data Science and Machine Learning is an area of learning.

As part of the project, following courses were considered for learning.

- Python and Anaconda Data Science Platform – Python and all the libraries & tools listed as part of Anaconda platform are learnt and used as necessary. It involved programming in python 3 within Jupyter notebooks developed for the project.
- Big Data & Predictive Analysis – CO3093 module from the University of Leicester
- edX course – Python for Data Science – This is one of the modules a part of Data Science MicroMasters program delivered by The University of California, San Diego
- Bootstrap – This project was the first instance where Bootstrap was used by the author. It was used for the application prototypes.
- Plotly.js – The prototype developed uses the Plotly.js javascript visualization library. The visualization code written in javascript is used by client side in the browser and had a steep learning curve for the project.

From a domain perspective, the healthcare domain and related datasets for the project were new. The topics include various healthcare data sources at a high level, NHS Digital Practice and CCG level prescription costs data, BNF, Public Health England health profiles and indicators. The domain knowledge is vital to contribute to this project and develop the idea in the future.

Requirements

This section details the requirements considered for the project their status and any remarks if any.

Data Mining

The table below is the set of requirements carried for data analysis aspect of the project.

Req. ID	Requirement Description	Status	Remarks
Essential – Priority 1			
D-1	Define data sets and the parameters for consideration	Achieved	
D-2	To prepare localized data sets for pilot	Achieved	
D-3	Explore the data sets with reference to clinical questions under consideration	Achieved	
D-4	Explore and generate correlations of data sets	Achieved	
D-5	A set of clinical questions to be considered in the order of priority are as below: Practices – Patient Size and Prescription cost variation	Achieved	
D-5-a	Practices level insights with respect prevalence of drugs and diseases	Partial	Only for Diabetes and Drug Groups provided by Prof. Umesh
D-5-b	CCG level view – practices & prevalence of drugs	Partial	Data exploration only. CCG level prescription costs were not considered for analysis and application.
Essential – Priority II			
D-6	Develop a methodology and generate a model to normalize, visualize and correlate the data	Achieved	Normalization of datasets and Cluster Analysis.
D-7	Visualizations for correlations	Achieved	3D Cluster Analysis and further analysis into each of the cluster to discover insights.
D-8	Export the correlated data in visually rich formats	Achieved	2D Scatter plot covers five dimensions - x-axis parameter

	<ul style="list-style-type: none"> Combining Patient Size, Prescription Volume and Patient Distribution to bring insights 		<ul style="list-style-type: none"> - y-axis parameter - size of the dot - the colour of the dot - the shape of the dot
Recommended			
D-9	Cost per head calculation and its relation to patient size and prescription volume	Achieved	Per patient cost was analysed as part of the data model. As part of the data model, it's relation to other parameters are discussed.
D-10	Practices – Prescription and Distribution among drug groups	Partial	Only for Diabetes and Drug Groups provided by Prof. Umesh
Optional			
D-11	Develop a (used in epidemiology) technique for statistical calculation of prevalence – i.e. to develop a prevalence estimate which is age adjusted	Not Achieved	This will require further analysis and domain knowledge of datasets with respect to profile indicators.
D-12	Build a technique to predict the demand for volume for a certain drug	Modified	Instead a model to predict per-patient cost was carried out to find out dependent variables.
D-13	3D Visualization of data from multiple datasets	Achieved	Via 3-D Cluster Analysis linked to 2-D plot and multiple tables for General Information, Costs information overall at the chapter level.

Software

The table below lists the requirements considered for the web application prototype developed as part of the project. These requirements were considered later in the project, due to the fact that knowledge of the domain & data during the course of the project.

Req. ID	Requirement Description	Status
A-1	It must provide a way to search a GP using a free text	Achieved
A-2	The search pattern could be practice code, CCG code, postcode or area name	Achieved
A-3	The result of a search should show the entries with general information: practice or CCG name, code and postcode.	Achieved

A-4	Each search result item should provide two options: View Trends or Add to Focus	Achieved
A-5	View Trends provides an option to view information for a GP	Achieved
A-6	Add to Focus provides an option to add an item to focus group for further analysis	Achieved
A-7	It must provide a specific page to visualize the clusters	Achieved
A-8	Once a cluster is selected, it should provide all the parameters to select either as a value for x or y-axes.	Achieved
A-9	The application provides an option to choose the Drug group for analysis of costs	Achieved
A-10	When visualization of GP data is shown, hovering over a GP instance should show general GP information, Total costs information, Chapter and Sub information	Achieved
A-11	When a GP is viewed it should show all the profiles general, patient list and health profiles.	
A-12	When GP is viewed, it could provide an option to view the costs trend for a specific drug group or total for each month across the year	Achieved
A-14	When a focus group is viewed, it should show all the profiles applicable to group	Achieved
A-15	When a focus group is viewed, it should show all the costs of across all the BNF chapters for the latest data for the items in focus group	Achieved
A-16	When a focus group is viewed, it should show allow the user to select a drug group and view the costs for the last year	Achieved

Approach

This section details the approaches considered for data mining and analysis of the datasets.

It gives detailed step-by-step activities carried out for the core part of the project.

The questions or topics considered are as below:

- Discover correlation between volume of patients with the volume of prescriptions?
- Are prescription behaviours linked to the type of patient population such as age, demographics or independent of them?
- Understand cost variation between practices

- What proportion of practice population has diabetes, hypertension and heart problems and how does this affect prescription practices?
- Present cumulative view of the datasets
- Usage of generic and specific or branded drugs
- GP Competition

Data Exploration

Data Exploration started with the process of finding the offline data for the datasets under consideration. All of the datasets were available in CSV format as direct links or downloaded via a Web interface.

Jupyter notebooks created for this stage are available under [DataExploration] folder within the project repository. They also been exported as .py and HTML under [Python] & [HTML] folders respectively.

Datasets linking

The first step was to link various datasets under consideration. As discussed, NHS Digital, NHS BSA and PHE are the data sources under consideration. Below table shows the structure and features of the data within each data source.

NHS DIGITAL			
CCG	PRACTICE LEVEL		
CCG Prescription Data	Practice level	Prescribing Practice Presentation Level	Presentation
Quarter	PERIOD	SHA	CHEM SUB
Area Team Code	Practice Code	PCT	NAME
Area Team Name	Practice Name	PRACTICE	
CCG Code	Address 1	BNF CODE	
CCG Name	Address 2	BNF NAME	
Chapter	Address 3	ITEMS	
Chapter Name	Address 4	NIC	
Section	Postcode	ACT COST	
Section Name		QUANTITY	
Items		PERIOD	
Actual Cost			
NIC			

NHS BSA		PHE	
		Area Code	

PATEINT LIST SIZE	PRACTICE GP COUNT	GP PROFILES
Patient List Size	PCO	IMD
Regional Office Name	PracticeCode	Disease Prevalence
Regional Office Code	PracticeName	Health Profiles
Area Team Name	Address1	
Area Team Code	Address2	
PCO Name	Address3	
PCO Code	Address4	
Practice Name	PostCode	
Practice Code	Dispensing	
Male 0-4	Prescribing	
Female 0-4	Total	
Male 5-14	GP Count	
Female 5-14	GP practice	
Male 15-24		
Female 15-24		
Male 25-34		
Female 25-34		
Male 35-44		
Female 35-44		
Male 45-54		
Female 45-54		
Male 55-64		
Female 55-64		
Male 65-74		
Female 65-74		
Male 75+		
Female 75+		

Although data is available through various sources the data they can be linked via common codes which are unique for a GP or CCG throughout the UK. As highlighted above with arrows green and blue arrows practice codes and area codes can be used to link the data.

GP Profile data

This section details the exploration of GP related data sources.

Patient List size :

This gives the information about the distribution of patient ages within a practice.

The distribution of patients to vital to understand the whether the patient population young, mid-age or older. This could help us to help us to correlate with the costs and variations between it.

From the available data below additional features are generated:

- Total Size: It the total patient list size of a GP. It is the sum of the patient count in all the age groups.

Details are in: PatientListSize notebook.

Statistical highlights of patient total list size for Dec 2017 data are :

Total number of GPs	7327
Mean. number of patients	8017
Std. number of patients	5120
Min. number of patients	1
Max. number of patients	72151

- Average Age Profile: As mentioned earlier, to differentiate one of GP from another mere age distribution of patients would not suffice.

To enable the comparison, the average age profile should be calculated Based on the information available it was calculated as simple formula as below:

$$\text{Average Age Profile} = \frac{\sum \text{Number of patients in an age range} * \text{Max of age range}}{\text{Total Patient List Size}}$$

GP count:

This information has details about GP address, postcode and a number of general practitioners i.e. Doctors appointed for a GP. This is an important measure to decide whether the number of doctors appointed have any effect on prescription behaviour and impact the measure of health deprivation.

Prescription costs data

Considering the data volume, the practice level prescription data was localized for a CCG. The CCG selected was Q59 of Leicestershire and Lancashire. So “Area Code” was used as a filter.

The localized data was used as a basis for understanding. Some of the highlights are as below:

- It contains individual prescriptions for each of the practices. However for the same presentation, the data is not aggregated as assumed.
- To find chapter and sub-chapter level costs, the data needs to be updated with additional columns chapter and sub-chapter. Once this information is added, the data can be grouped at chapter and sub-chapter level to find the aggregates.

The required aggregation of data was carried out, details are in PrescriptionDatasetsCorrelation Jupyter notebook.

As part of exploration for a specific disease, Diabetes was considered and related drug groups:

- 0601021* Oral hypoglycaemics
- 0601022* Metformin
- 0601023* Gliptins
- 0601023AKAAA * Alogliptin – costlier options
- 0601011* Insulin
- 0601012*Insulin
- 0704050* Viagra and others

Drugs for Diabetes fall under Chapter 6 - endocrine system. Each of the drug groups above is prescribed based on the patient condition.

- Gliptins have generic and branded options. The usage of the branded drugs could impact the total cost of prescriptions by a practice.

- Viagra and others : this group is available as off-the counter (OTC) drugs. However, prescriptions of this drug group could increase the costs.
- Based on the domain knowledge, the relation between Metformin and Gliptins should be relative to the diabetic population. Higher prescription of gliptins could imply specific prescription behaviour by a practice.
- All the drug groups above except the last, relating to the percentage of the diabetic population for a GP.

Above pointers could be used to understand the prescription behaviours for a specific disease and relation between generic & branded drugs between GPs with similar characteristics.

The required aggregation of costs at the chapter level are available in Prescription data-ChapterBased-Analysis notebook.

Public Health Indicators & Index of Deprivation Measures

The second aspect of a general practice profile is the health profiles available from Public Health England dataset. This data is available offline in the form of CSV data via download as well as public API via fingertips interface.

The dataset is available annually since 2008 but it provides crucial information which covers multiple facets as below:

- Index of multiple deprivation measures
- QoF information for various diseases such as “patient count” of a specific disease can be found.
- The data also captures what is the distribution of patients w.r.t ratings of specific diseases such a Hb1Ac level, Blood Pressure levels. This is crucial to understand the trends and health profile of a specific GP or an area annually.
- GP patient data – this dataset is available as a separate public data. It provides satisfaction indicators which are generated from the surveys or feedback questionnaires filled by the patients.
- Other crucial information of interest is “% of 65+ age” patients and the number of patients with long-term illness.

The disease prevalence for diabetes can be linked to GP patient population and the prescription costs of the drug groups for a GP or a group of GPs for comparison.

The % of 65+ aged feature could be used to understand its impact on overall costs.

Details are in PHE_PracticeProfileIndicators_Analysis notebook

Combining Datasets

This section details the merging of the data across sources and analyses them.

Details are in PrescriptionCost_PatientListSize-Indicators-Correlation notebook

As part of Data exploration, GP patient list size and GP count datasets were combined. Average age profile was introduced. The insights from this combined dataset are as below:

- at UK level – map view (to be updated)
- at Leicester level – map view (to be updated)

Next step was to find a correlation between patient list size and costs, the data for Q59 CCG was visualized. The scatter plot of total prescription costs for Dec 2017 (NHS Digital) vs Total patient list size (NHS BSA) is as below:

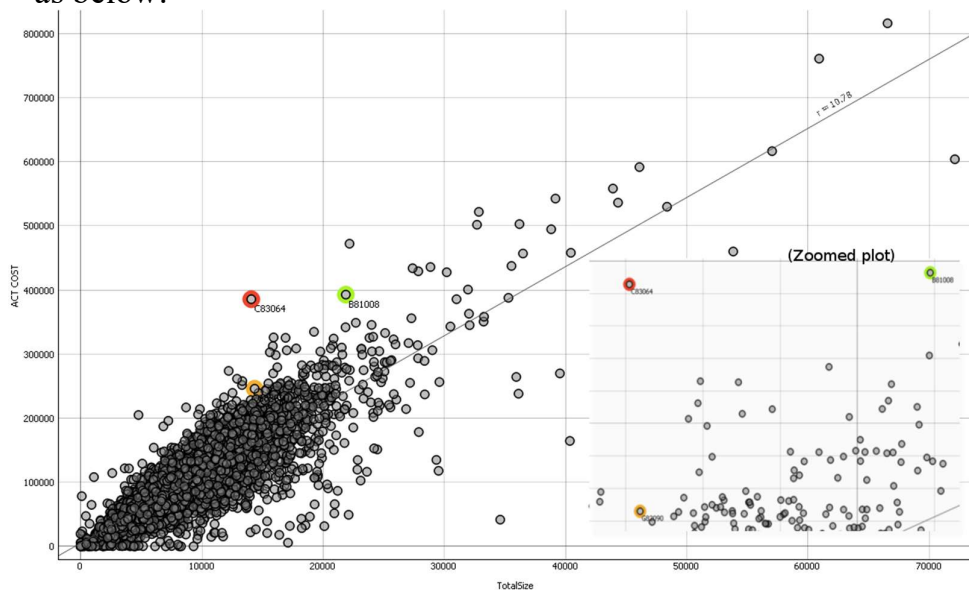


Figure 6: Scatter plot for Patient list size vs Actual Costs

As a sample GPs were considered based on below factors:

- comparison of GPs with same list size but different costs
- comparison of GPs with different list sizes and similar costs

As highlighted in the plot above, selected GPs are: C83064, G82090, B81008.

C83064 has a patient list size same as G82090 but prescription costs (in GBP) are close to B81008



Figure 7: Comparison of patient list sizes and costs for selected practices

To understand the reasons for the variance, the distribution of patient list size, additional factors such as disease prevalence (PHE: Diabetes, CVD profiles) were considered:

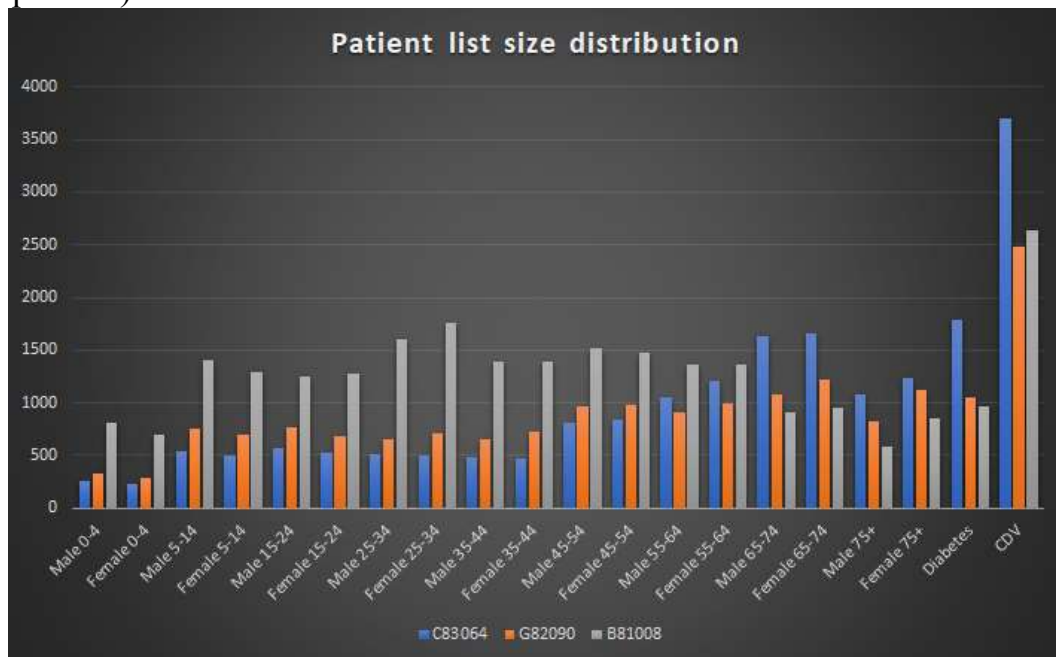


Figure 8: Distribution of patient across ages, diabetes and Cardio-vascular diseases

From the above plot, C83064 differs from other two GPs:

1. It has a higher patient list size in range 55 to 75+
2. Diabetes and Cardio Vascular Disease significantly higher than others.

To explore additional parameters, Index of multiple deprivations (PHE: IMD 2015) and GP Count (NHS BSA) were considered.

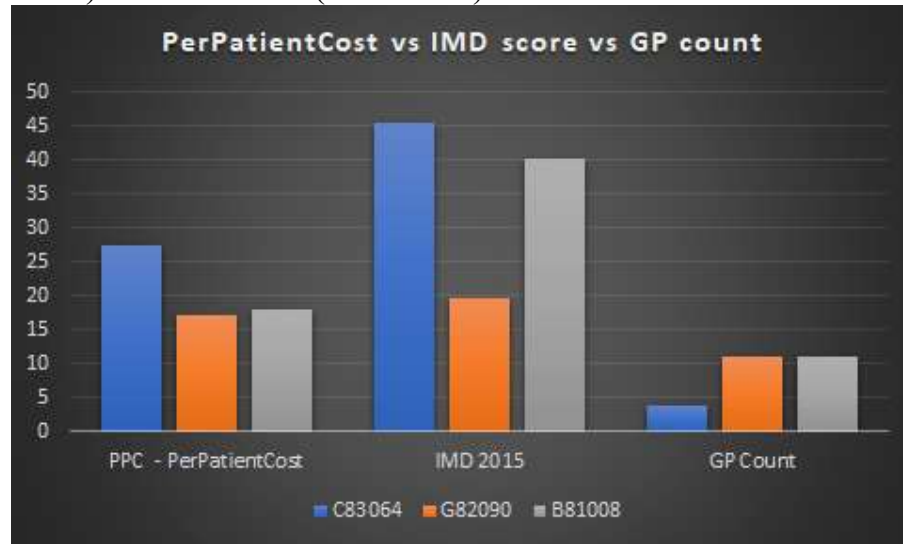


Figure 9: Per Patient Cost vs Index of Multiple Deprivation vs GP Count

From the above plot, with 19% IMD – G82090 is less deprived than C83064 which has a higher deprivation index. However, B81008 also highly deprived area but in terms of GP count, C83064 has only 4 GPs as compared to 11 in others.

From above insights, it can be concluded that comparison of prescription costs will need consideration of health profiles data. This is an important insight volume of patients, the volume of prescriptions and type of patient population such as age, disease prevalence

Cluster Analysis

Next step was to process the correlated data and learn about it. Since we don't have specific target variables defined for the data unsupervised machine learning technique is used.

For the problem under consideration cluster analysis was used to determine segmentation based on following independent parameters:

- Average age profile
- % of diabetic patients
- index of multiple Deprivation scores.

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import pandas as pd
import numpy as np
```



```

from itertools import cycle, islice
import matplotlib.pyplot as plt
from pandas.plotting import parallel_coordinates
%matplotlib inline

mergedData = pd.read_csv('./UKMergedDataSelected_u2.csv', sep=',', encoding
= "ISO-8859-1")

features = ['AvgAgeProfile', '% aged 65+ years', '% of Diabetes patients',
'Deprivation score (IMD 2015)', ]

sampleData = sampleData[features]

X = StandardScaler().fit_transform(sampleData)
kmeans = KMeans(n_clusters=4, init='k-means++', n_init=20, max_iter=500,)

model = kmeans.fit(X) print("model\n", model)

centers = model.cluster_centers_

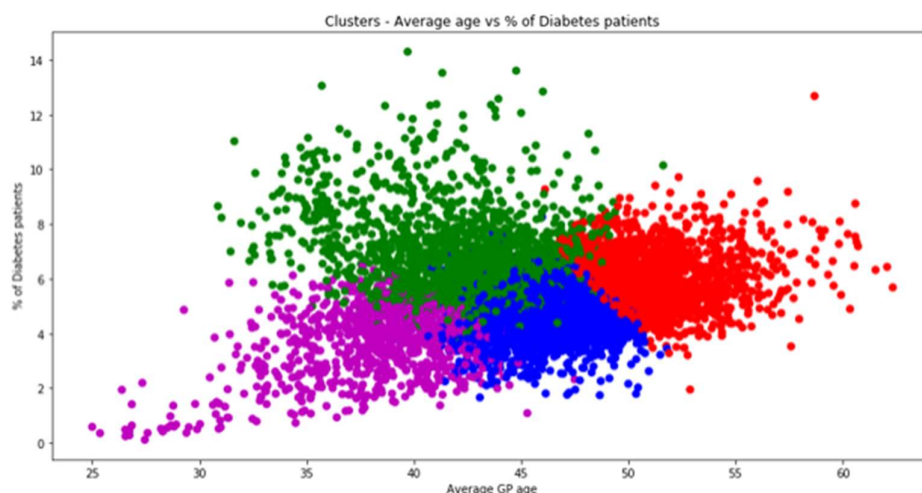
# Plot clusters - Average age + vs % of Diabetes patients
plt.figure(figsize=(14,7))
colormap = np.array(['b', 'g', 'r', 'm', 'c', 'y', 'k', 'fuchsia',
'peachpuff', 'k', 'm'])
plt.scatter(sampleData['AvgAgeProfile'],
sampleData['% of Diabetes patients'],
c=colormap[model.labels_], s=40)

plt.title('Clusters - Average age vs % of Diabetes patients')
plt.xlabel("Average GP age") plt.ylabel("% of Diabetes patients")

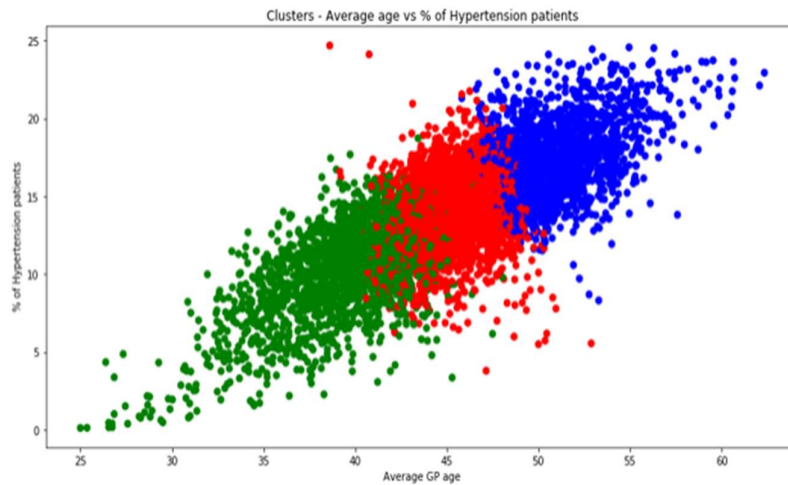
```

For clustering, Kmeans ++ algorithm is applied.

Below is the clusters for Diabetes vs Average Age



Below are the clusters for cardiovascular disease: Hypertension vs Average Age



However, it was not very evident whether the clusters formed were profound. Hence 3-D visualization was carried. Plotly as a visualization library had technical issues while using the Jupyter notebook. So as part of the project and HTML page was developed.

The code is available under [Prototype] folder in ClusterAnalysis.html and clusters.js

Cluster Analysis of General Practices in UK

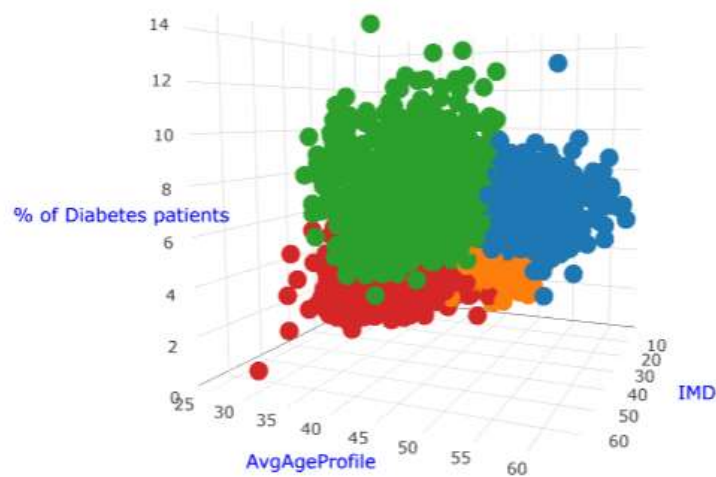


Figure 10: 3D Scatter Plot of Cluster Analysis

Based on the above clustering, it was evident that the clustering is possible

and segmentation can be done. Of the 3 features, the clear indicators are age profile and deprivation measures.

In terms of naming, they can be termed as below:

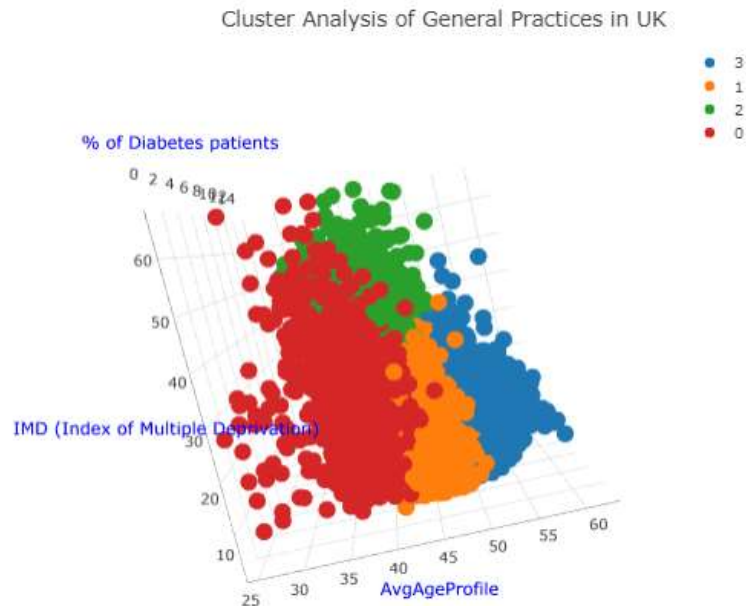


Figure 11: 3D Scatter plot Average Age Profile vs IMD

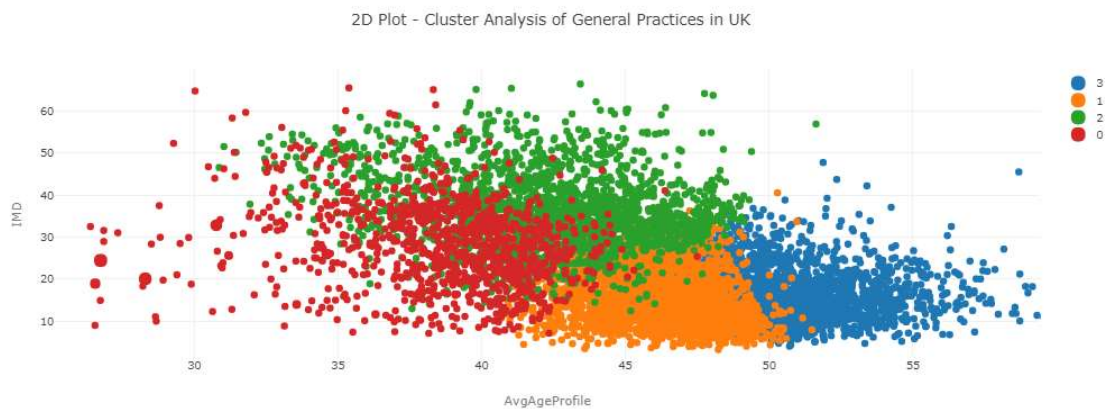


Figure 12: 2D Scatter plot: Average Age Profile vs IMD (zoomed)

- Cluster 0 is relative population which is young with some medium-age living in affluent areas
- Cluster 1 is medium-age population living in affluent areas
- Cluster 2 is relatively less young to medium-age, deprived
- Cluster 3 is older population living in affluent areas

In line with the clusters, when per-patient cost and % of 65+ age are considered. Following insights can be discovered:

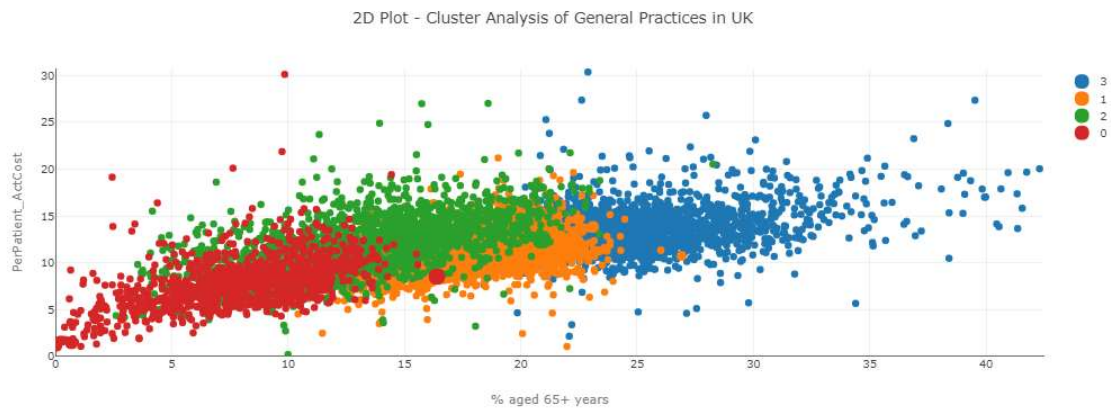


Figure 13: 2D Scatter plot of % aged 65+ years vs Per Patient Cost

From the above plot, Cluster 2 has lesser % aged 65 than compared to Cluster 1. But the per patient cost of 2 is higher on average as compared to cluster 1. This can be used to conclude that older the population and deprived, per patient cost will be higher as compared to medium to the older population living in affluent areas. This insight can be used to prove that prescription behaviours are linked to the type of patient population such as age, demographics.

Details are in ClusterAnalysis notebook

As described earlier, as part of the project-specific drug groups were considered for Diabetes as a disease. Each of the clusters and costs for each drug groups was analysed. For detailed analysis the distribution of costs for drug groups was done at subchapter level w.r.t formulation names. This helped to understand the variations in prescriptions of generic or specific drugs which vary in costs thus affecting overall costs incurred by a practice.

To generate costs for each drug group based on the cluster. The orange framework is used to build a generic workflow as below:

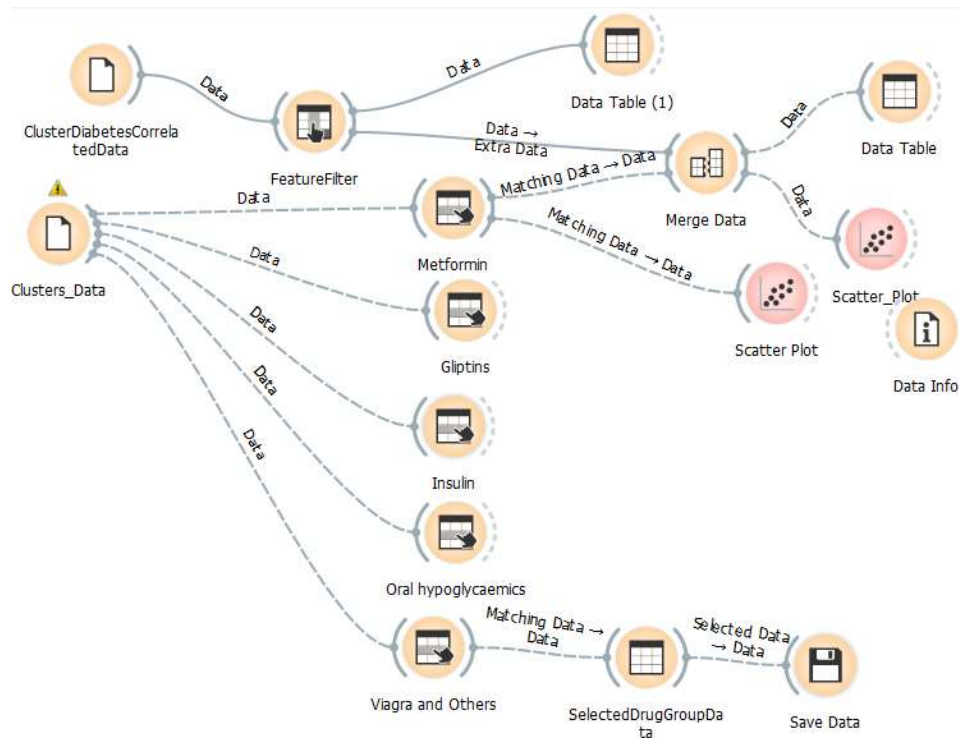


Figure 14: Generic Cluster Analysis data for selected Drug Groups

It used file, column and row selectors to merge the data. Save the selected data into an external file.

Jupyter notebooks

As part of the data exploration phase, for tasks notebooks developed are as below:

- PatientListSize
- PrescriptionDataExploration
- PHE_PracticeProfileIndicators_Analysis
- AreaCost_PatientListSize-Analysis
- PrescriptionCost_PatientListSize-Indicators-Correlation
- ClusterAnalysis
- PrescriptionData-ChapterBased-Cluster-Correlation
- PrescriptionData-ChapterBased-Cluster-SubChapterCategorization
- DataMergingForPrototype
- PrescriptionData-ChapterBased-Analysis
- PrescriptionDatasetsCorrelation

They are available under [DataExploration] folder within the the project repository.

Data Model

Based on the understanding of the datasets, “per patient cost” variable is considered for the data model. This data model is generated in Orange framework.

The workflow to generate the data model is as below:

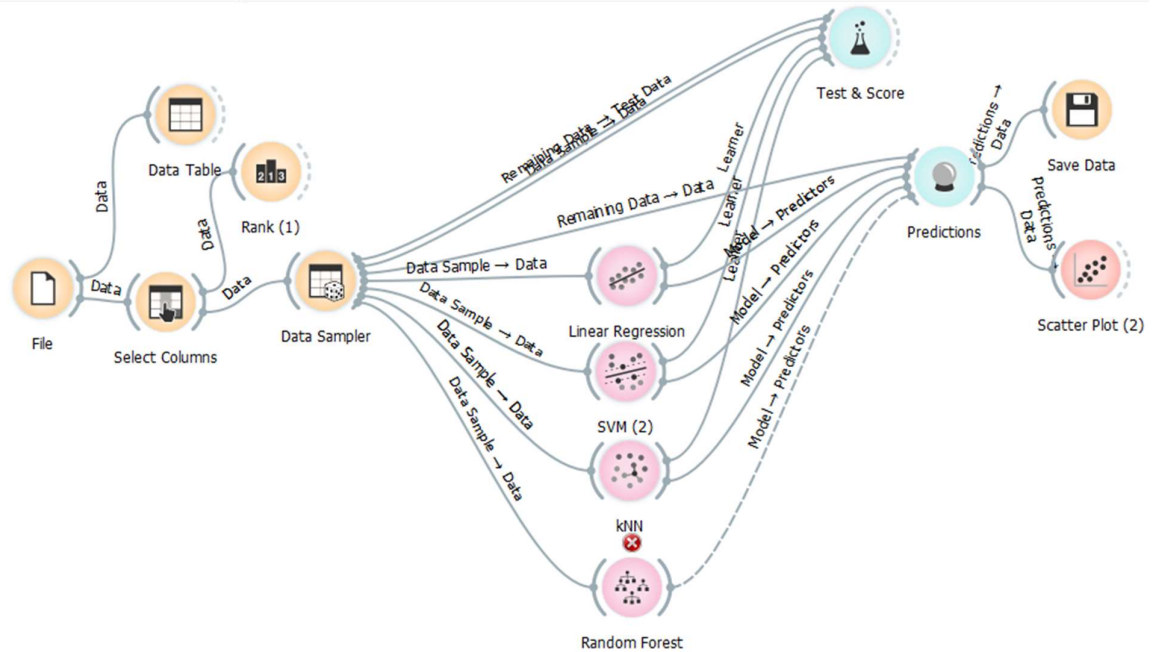


Figure 15: Data model to predict per patient cost

To begin the features under consideration were ranked and below is the standing:

	#	Univar. reg.
N % of Female 75+		3544.025
N % of 75+		3439.531
N % aged 65+ years		3419.078
N % of Male 75+		2996.560
N AvgAgeProfile		2877.742
N % of 65-74		2842.336
N % of Female 65-74		2796.499
N % of Male 65-74		2747.831
N % of Female 55-64		2448.783

Figure 16: Ranking of features with a level of correlation

Later top 4 features were used as the basis of the generating the data model. By using the Data Sampler object, train and tests data were created. 90/10% fixed proportion technique was used to generate the sample data. The trained data is passed to the models which are supervised learning to predict per patient cost. The techniques considered are:

- Linear Regression
- Simple vector machines
- kNearest Neighbours model

The generated scores were:

On Trained Data					On Test Data				
	Method	MSE	RMSE	MAE		Method	MSE	RMSE	MAE
	Linear Regression	8.935	2.989	2.001		Linear Regression	7.095	2.664	1.944
	SVM	19.531	4.419	3.337		SVM	15.657	3.957	3.188
	kNN	6.978	2.642	1.738		kNN	7.853	2.802	2.092

When the models were used to predict the cost, below image shows the results:

	Linear Regression	SVM	kNN		PerPatient_ActCost	PRACTICE
1	13.547184154	9.054280336	14.697753234		14.370155970	M92011
2	11.163228884	9.280253932	10.957356170		11.229362140	F81656
3	12.903743401	9.220704287	13.287368976		12.145693830	N81088
4	12.837693319	9.136766267	12.617268404		14.509943180	G81048
5	12.702142820	9.024744464	12.266311684		13.803995470	B82041
6	13.485102340	8.788416280	14.863973480		9.314308145	B81666
7	9.936346337	9.184598691	12.209015217		12.435457880	P83621
8	9.547108546	9.234573941	10.546753736		14.040391150	N85629
9	11.638660833	9.018072599	11.091845852		11.640871740	H81095
10	10.713572010	9.204851964	10.551440804		10.059172370	G82768
11	10.218822166	8.927187704	9.801330310		10.985990930	A89025
12	11.531634167	9.172000969	12.582550108		12.778697830	P92605
13	10.625600234	9.191477508	10.693367080		7.704318336	H85101

Figure 17: Predicted Per-Patient Costs

Of the selected model, both Linear Regression and kNN look suitable. Both the models will need parameter fitting. However due to variation from the actual costs, confidence on the generated models was low. Hence the model generation will require further analysis and work to create accurate models.

Technical Specifications

This section details the proposed architecture of the application to be developed for the project. The current scope of the project might be limited, but in a long-term application needs to handle multiple datasets and handle correlation between them.

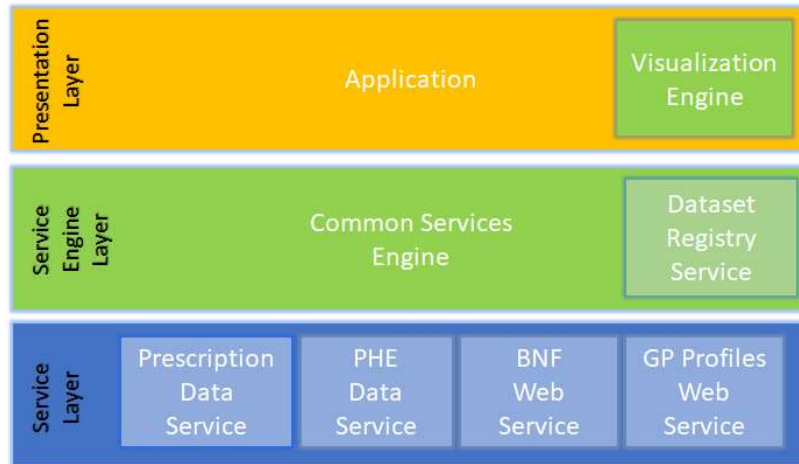


Figure 18: Software Architecture

The overview of the architecture is as below:

- The service layer represents the datasets and the web services providing the information.
- Each of the data services registers itself via Dataset Registry Service. During registration, the service will submit its schema information and as well as correlation with other datasets. If the related datasets available, the correlation information will be used. Each of the data services will be a rest API service, so that if required it can be deployed as public API.
- The registry service is an integral part of common services engine. Through this service any new dataset service can be registered.
- The common service engine abstracts the communication with actual datasets and their services from the application perspective.
- Since the application is built on MVC framework. Only controllers will be responsible for communicating with common services engine.
- Visualization engine has the intelligence to build data visualization and is used by Views of the application. It abstracts all the complex logic to create document driven documents required as part of the plotly.js library.

Considering the simplicity of the current requirements of the project, the pages are developed using HTML and JavaScript. Most of the business logic communicates with to OpenPrescribing.net dataset using AJAX.

Application prototype

Wireframes

Based on the areas of interest for the prescription data and related datasets the mock screens were generated. They are detailed in following sections.

The screen below shows the search screen and results returned with the option to view trends and Add to focus.

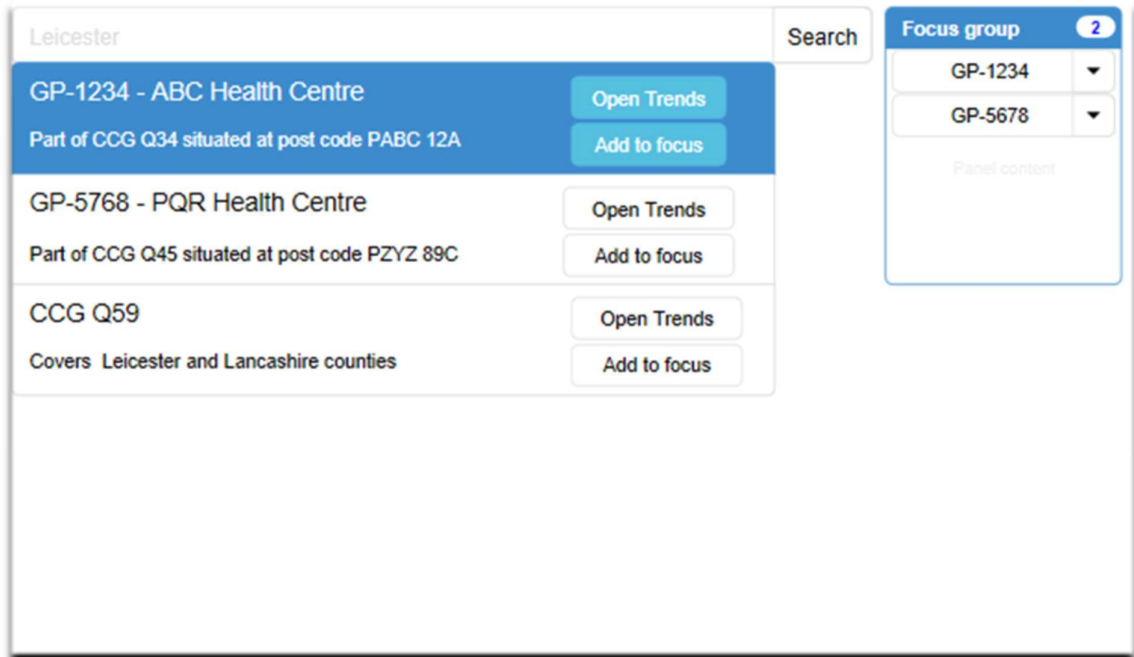


Figure 19 : Mock for search, results and focus group

The screen below shows the information shown on the items selected in the focus group. The details are general information, patient profile, practice profile and other indicators of interest.

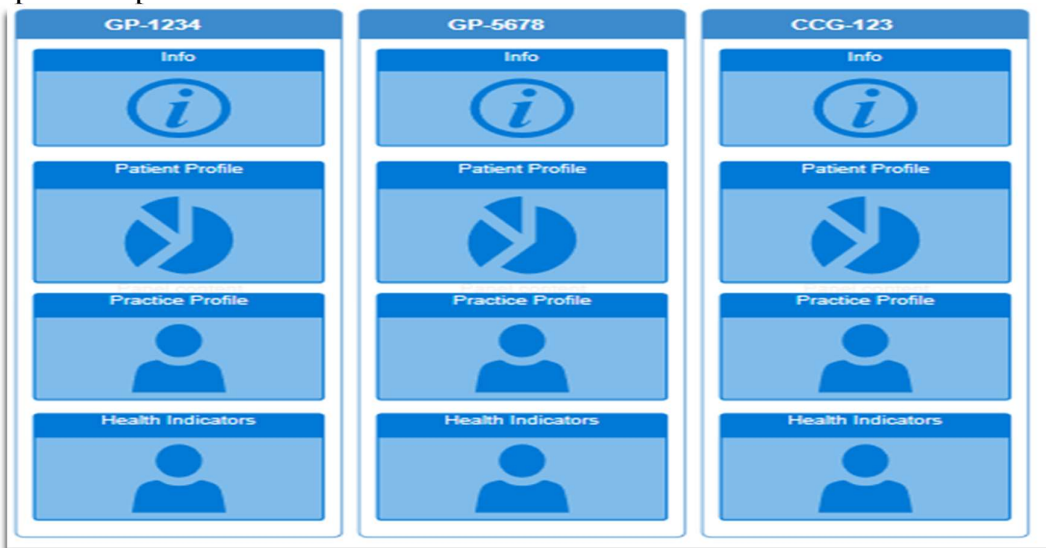


Figure 20: Mock to view the items selected in focus group

The screen below shows the specific page of GP or an item in a focus group. It provides monthly information for each chapter costs and provides a filter for a drug group.

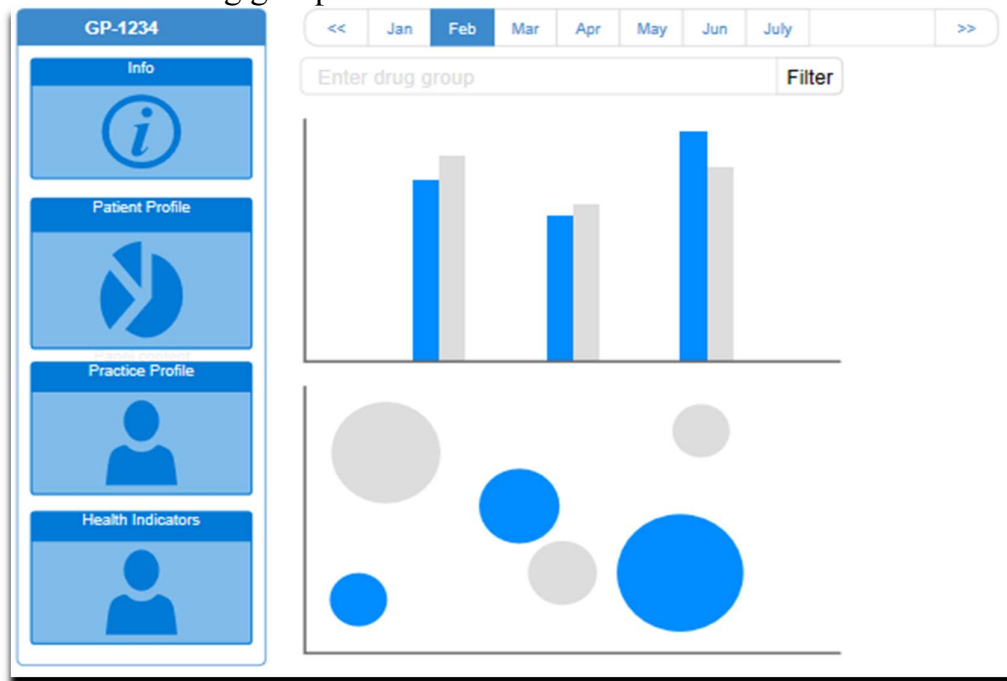


Figure 21: Specific page mock for an individual item such as GP or CCG

The screen below is used to view the clusters generated as part of the project. It provides an overview of the hovered item. The data can be filtered based on a cluster or a specific filter.

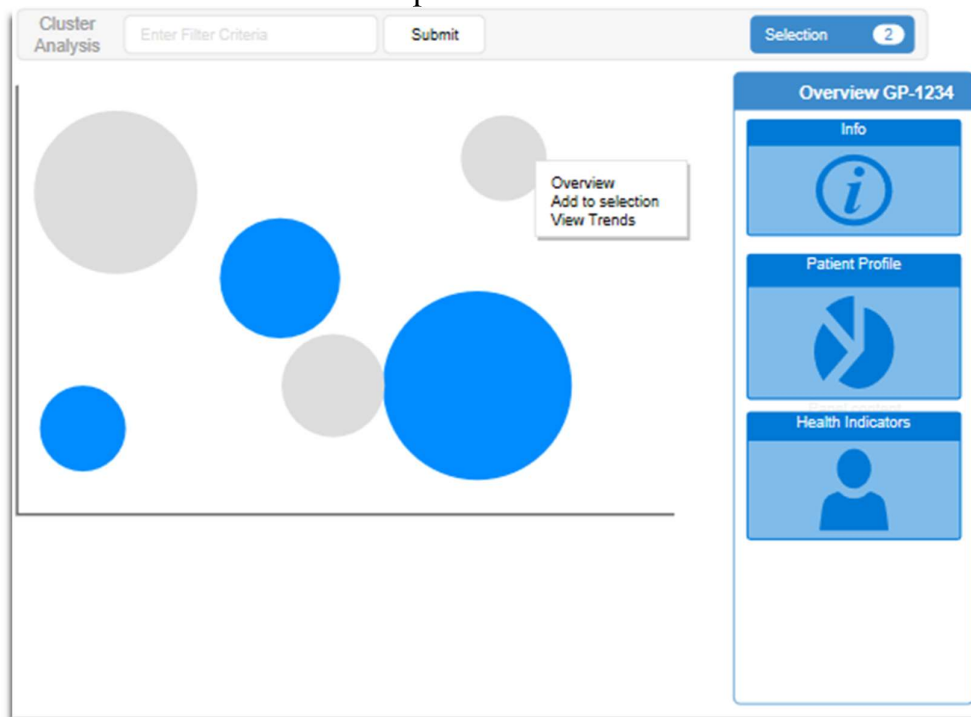


Figure 22: Cluster Analysis mock with an overview for hovered item

Final Prototype

This section details the application prototype developed with the screenshots.

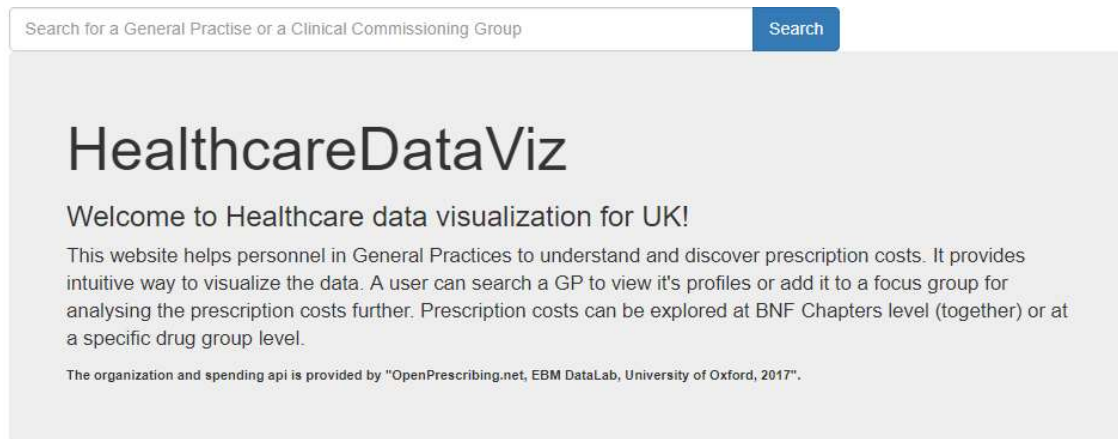


Figure 23: Home Page

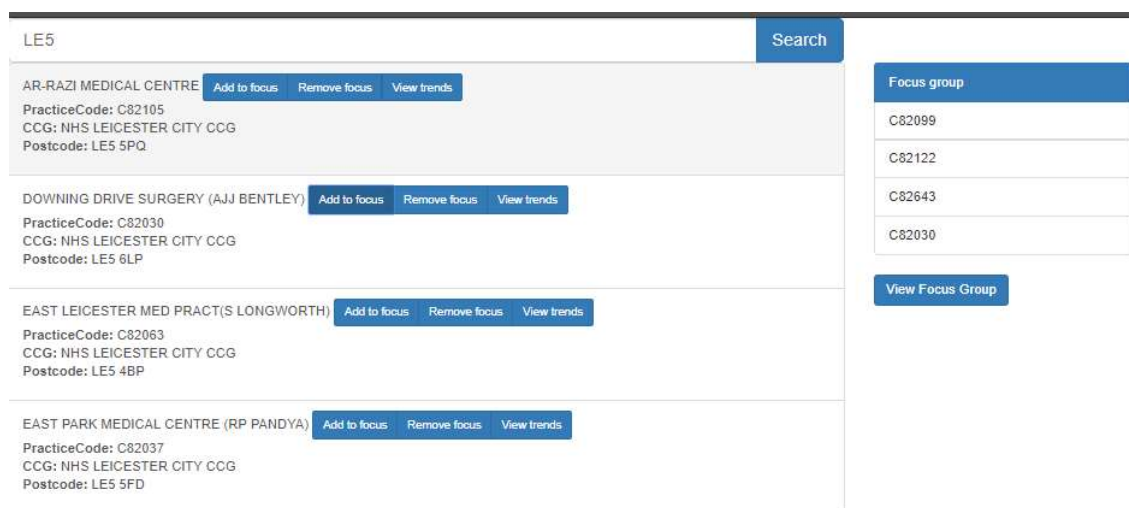


Figure 24: Search Results and items added to Focus group



Figure 25: View Trends for a single GP: costs for all chapters and costs for selected drug groups

In addition, a single page prototype was created for cluster analysis. The screenshots are as below:

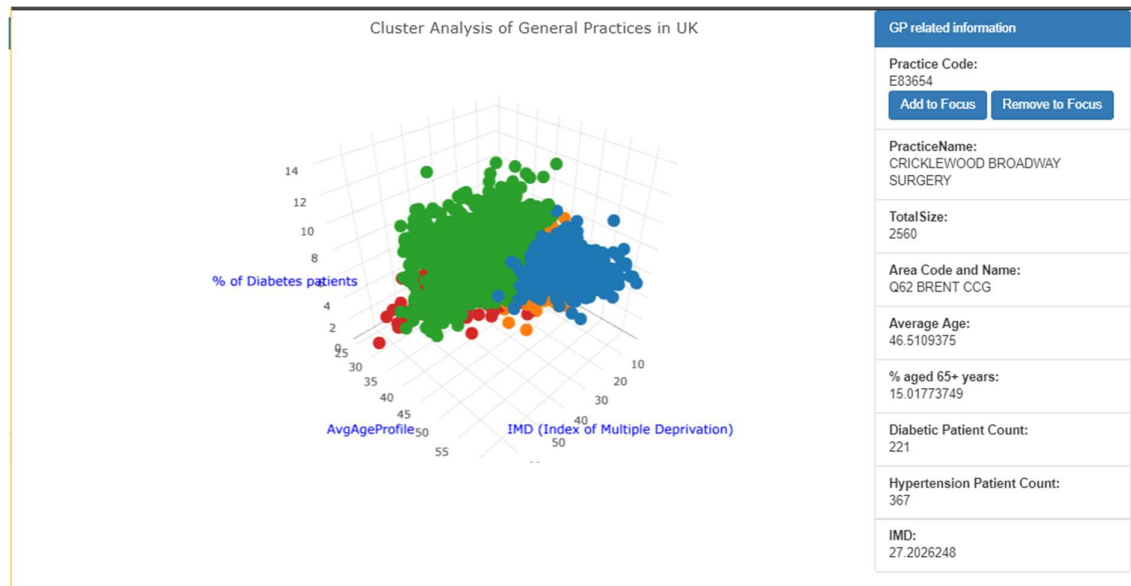


Figure 26: 3D Cluster View with general GP information



Figure 27: 2D Plot view with a selection of x and y parameters. Costs overview.

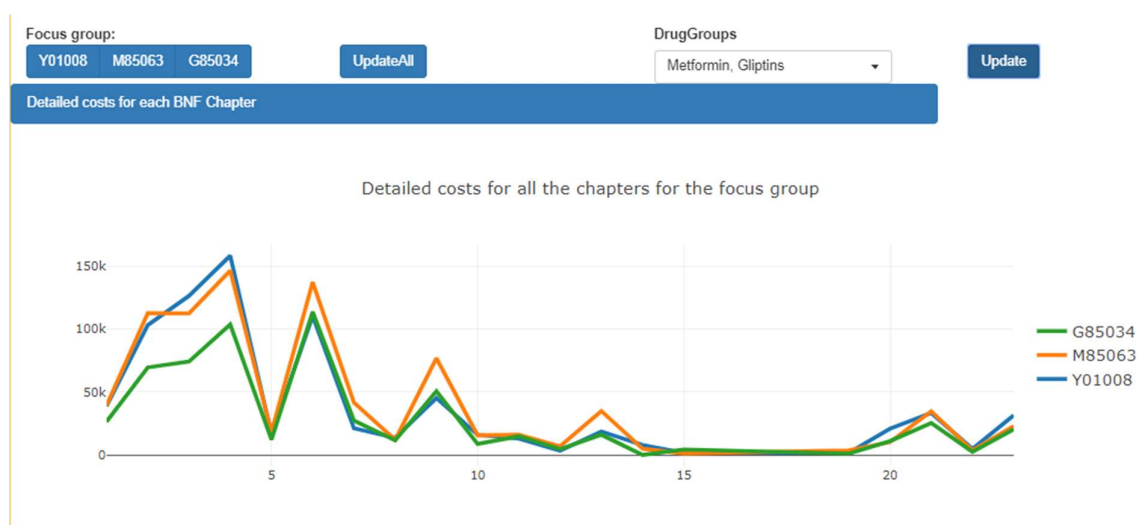


Figure 28: Drug groups costs view for the focus group

The application prototype and cluster analysis prototype have been developed using standard technologies i.e. HTML, JavaScript, JSON, Bootstrap. Plotly.js is used as the charting library which is open source. The functionality provided is intuitive and meet the requirements set as part of the project.

Conclusion

Project

The primary scope to correlate the prescription costs data with disease prevalence health indicators is achieved. The project concisely discusses the primary concerns and gives a detailed account of the datasets explored and insights from them. As a conclusion, the prescription costs and the behaviours exhibited by the medical practices are linked to numerous factors such as patient age profile, area demographics such as deprivation measures and other health indicators. So, when comparing the practices

should be selected with similar characteristics for unbiased study. An attempt made to generate the data model to predict was not conclusive and will need further analysis. In terms of outcome, the application prototypes developed useful from the end user's perspective. The initial feedback is encouraging with scope to develop further on the idea and domain knowledge attained.

Challenges

- The data domain considered for the project is new for the author. Hence it was relatively difficult for the author to judge the results and needed guidance from a domain expert.
- Application of data science principles on multivariate data in some instances was difficult, since the number of features was quite high.
- Understanding of the BNF or medical formulations is crucial to understand generic and branded drugs. This was taken up later in the project.
- Usage of Visualization library has a steeper learning curve than anticipated.
- The discussion on the application prototype and feedback was necessary. This was initiated very late in the project and should have planned earlier.

Future Developments

Considering the challenges faced and the understanding of the prescription data and other datasets is a new initiative. It is vital to capitalize on the domain knowledge built and work on next steps.

Following steps are proposed:

- To complete the data model for prescription cost prediction
- To conduct user interactions with different stakeholders and understand the need for advanced visualizations.
- Work on the feedback provided by the end-users during the demo. This should be developed in an iterative way with continuous involvement from the teams and stakeholders.
- To build the architecture of the system, so that it can be incorporated into other NHS platforms or OpenPrescribing.net from EDM Data labs
- In terms of visualization, it will be interesting to generate animated trends in the timeline for a group of GP or CCGs in terms of:
 - Average Age Profile
 - % of Diabetic Patients or any specific disease
 - Prescription Cost
 - Patient List Total Size
 - Usage of specific medicine vs another
 - Viagra Usage
- In terms of new datasets, below datasets should be explored:

- Prescription cost analysis
- Dispensing data (Pharmacies data) link it to prescriptions and practices.
- ONS data – to link it to outcomes such as mortality rate and other parameters.
- Various health indicators provided by Public Health England
- GP survey data from the questionnaires answered by Patients

Glossary

GP	General Medical Practice
CCG	Clinical Commissioning Group
NHS	National Health Service
NHS BSA	National Health Service Business Service Authority
PHE	Public Health England
BNF	British National Formulary
IMD	Index of Multiple Deprivation
NIC	Net Ingredient Cost
ACT	Actual Cost
IDE	Integrated development environment

References

1. Raghupathi, W. and Raghupathi, V. (2014) 'Big data analytics in healthcare: promise and potential', *Health information science and systems*, 2(1), pp. 3.
2. Park, S., Choi, D., Kim, M., Cha, W., Kim, C. and Moon, I. (2017) 'Identifying prescription patterns with a topic model of diseases and medications', *Journal of Biomedical Informatics*, 75, pp. 35-47.
3. Desai, R.J., Sarpatwari, A., Dejene, S., Khan, N.F., Lii, J., Rogers, J.R., Dutcher, S.K., Raofi, S., Bohn, J., Connolly, J., Fischer, M.A., Kesselheim, A.S. and Gagne, J.J. (2018) 'Differences in rates of switchbacks after switching from branded to authorized generic and branded to generic drug products: cohort study', *BMJ*, 361.
4. Schaumans, C. (2015) 'Prescribing behavior of General Practitioners: Competition matters', *Health Policy*, 119(4), pp. 456-463.
5. Corput, van der, PNA Paul, Arends, J.J. and Wijk, van, JJ Jarke = Jack (2014a) 'Visualization of medicine prescription behavior', *Computer Graphics Forum*, 33(3), pp. 161-170.
6. Laramée, B. (2017) *Time-Oriented Cartographic Treemaps for the Visualization of Public Healthcare Data*. (Accessed: Feb 21, 2018).
7. Liang, J. & Huang, M.L. (2010) 'Highlighting in Information Visualization: A Survey', , pp. 79.
8. Inseok Ko, MS and Hyejung Chang (eds.) (2017) *Interactive Visualization of Healthcare Data Using Tableau*.
9. Chandarana, P. & Vijayalakshmi, M. (2014) 'Big Data analytics frameworks', IEEE, pp. 430.
10. Diances, J.A. *Data Science with Python & R: Dimensionality Reduction and Clustering* | Codementor. Available at: <https://www.codementor.io/jadianes/data-science-python-pandas-r-dimensionality-reduction-du1081aka> (Accessed: Apr 3, 2018).
11. Frtunić, G., Pufović, D., Stevanoska, E., Jevtović, S., Velinov, G. and Stoimenov, L. (2017) 'Interactive map visualization system based on integrated semi-structured and structured healthcare data', 10649 LNBI, pp. 94-108.
12. Recommendations for appropriate prescribing of generic and branded medicines (<http://pad.res360.net/Content/Documents/Generic%20Prescribing%20policy%20-%20SHCCG%20July%202015.pdf>)
13. OpenPrescribing.net, EBM DataLab, University of Oxford, 2017 (<https://openprescribing.net/api/>)
14. NHS Digital Prescribing (<https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing>)
15. NHS Business Service Authority (<https://www.nhsbsa.nhs.uk/>)
16. Public Health England Data (<https://fingertips.phe.org.uk/api#/Data>)
17. "Getting Started," Mozilla Developer Network, 2017, accessed 23 April 2018, https://developer.mozilla.org/en-US/docs/AJAX/Getting_Started.
18. The JSON Data Interchange Standard. (<https://www.json.org/>)