HEALTH DATA VISUALISATION

August 16, 2015

Athanasios Anastasiou
for the Farr Institute
Swansea School of Medicine
Swansea University

Acknowledgements

A number of people have contributed in one or another way in the making of this handbook:

1. The rest of the FARR Institute Leads Catharine Goddard, Colin McCowan, Georgina Moulton and Paul Taylor provided the opportunity to run this workshop on health data visualisation.

2. Georgina Moulton and Colin McCowan provided a number of useful academic papers making use of the Quality Outcomes Framework datasets and also Georgina Moulton proof read and provided feedback on intermediate versions of this handbook.

3. Karen Tingay, provided a set of more suitable alternative references, to the ones I initially had in mind, on visual perception and typography.

4. Richard Fry, provided a set of helpful comments on pre-processing the Ordnance Survey's post-code dataset.

Contents

1 Principles of Visualisation **4**

1.1 Perception, Perceptual Organisation and Attention 5

1.1.1 Using a layout grid 5

1.1.1.1 Useful resources 6

1.1.2 Composing the visualisation 7

1.1.2.1 The law of Proximity 8

1.1.2.2 The law of Similarity 8

1.1.2.3 The law of Closure 8

1.1.2.4 The law of Good Continuation 8

1.1.2.5 The law of Symmetry 8

1.1.2.6 The law of Simplicity 9

1.1.2.7 Useful resources 9

1.1.3 The role of typography 10

1.1.4 Useful resources 13

1.1.5 The role of colour 13

1.1.6 Useful resources 14

2 Visualising Data With Tableau Public **22**

2.1 Obtaining & Installing Tableau Public 22

2.2 Basic Concepts 23

2.2.1 The Visualisation Story 24

2.2.2 The Dashboard 26

2.2.3 The Sheet 26

2.2.3.1 Data Source 26

2.2.3.2 Content Visualisation 28

2.3 Basic Operations 28

2.3.1 Loading A Data File 28

2.3.2 Joining Two Data Sources . . . 32
2.3.3 Producing Calculated Attributes . . . 32
2.3.4 Composing A Very Simple Visualisation . . . 33
2.3.5 Filtering Data . . . 34
2.3.6 Customising a visualisation . . . 36
2.3.7 Putting together a Dashboard . . . 39
2.3.8 Putting together a Story . . . 41
2.3.9 Exporting and accessing the visualisation online. . . . 44
2.4 Conclusion & Outlook . . . 45

3 Working with a realistic dataset: UK NHS Quality Outcomes Framework 46
3.1 Introduction . . . 46
3.2 Specific QoF datasets . . . 47
3.3 Other datasets enabling rich visualisation . . . 48
3.4 Publications . . . 48

Chapter 1

Principles of Visualisation

The objective of this chapter is to introduce a minimal set of universal principles which can be used to structure 'content' in a way that it becomes easily perceived by an audience of human beings. These principles are stemming out of the study of the human visual system and psychology and they therefore apply to any kind of medium whether a printed article, poster, web page or other.

Specifically, the material presented in this chapter are drawn from three large areas of study:

Visual Perception The way by which the human visual system perceives the real world

Attention The way by which the human brain selectively focuses on parts of perceived information

Perceptual Organisation The way by which the human brain organises perceived information and derives meaning

Each of these areas has received an extraordinary amount of research due to its importance to the way human beings perceive, interpret and navigate spaces around them, whether physical or virtual. For this reason, each point is presented here *in addition to the activities carried out during the theoretical session that takes place on the second day of the PhD Symposium* in a compact form, providing an opportunity for further exploration rather than becoming the definitive guide for structuring scientific content.

1.1 Perception, Perceptual Organisation and Attention

1.1.1 Using a layout grid

It is impossible to convey the utter importance of ***using a layout grid*** with words . . . in any written language.

A layout grid is used to structure content and figure 1.1 depicts its main components. Margins are used to delineate the space occupied by the content. As if divided by recursion, content space is further divided into smaller rectangular areas called Modules, surrounded by Gutters. In this way, modules are naturally organised into rows and columns. A group of closely located modules forms a Zone.

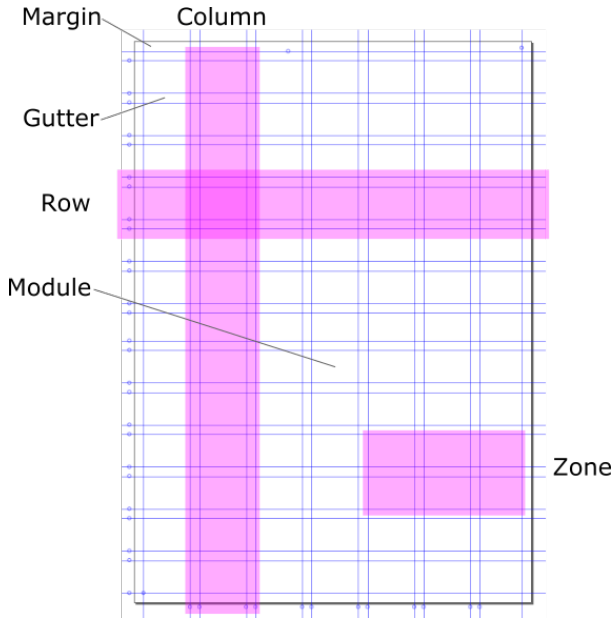


Figure 1.1: A layout grid used to structure content along with a set of key terms associated with it.

Figure 1.2 shows a sample poster with the same grid overlaid on top of the content. Three horizontal zones have been used to present three different aspects of the research. Each zone is further subdivided into text and image content and within each of these subzones, elements such as text and images still conform to the spatial division effected by the grid.

Layout does not necessarily have to be rectangular. Alternative layouts include (but are not limited to) the Column layout, Rule of thirds layout, Z layout, the Perspective layout, the Circular layout and others.

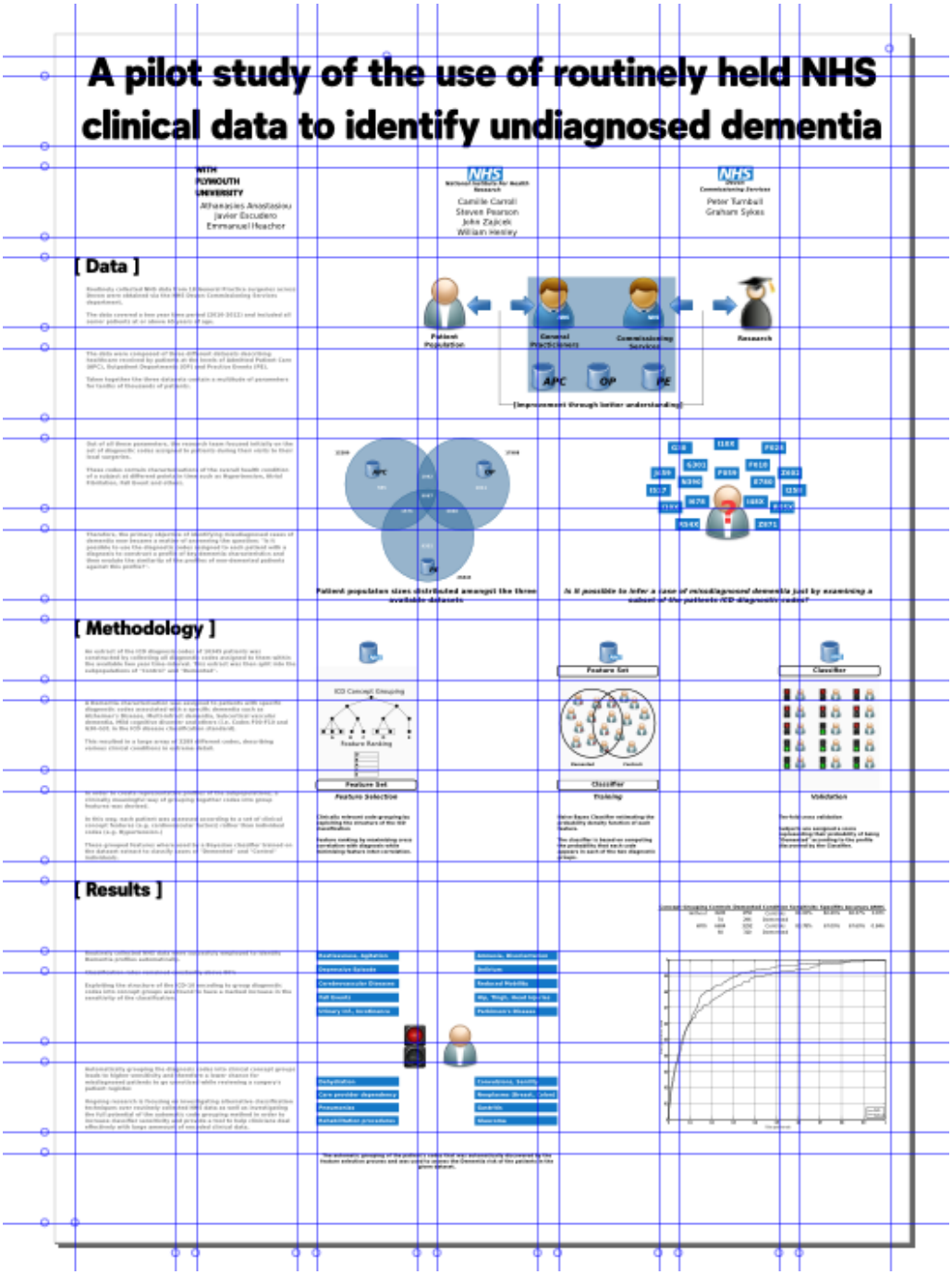


Figure 1.2: The same layout grid, presented in figure 1.1 with the content it provides a structure for.

1.1.1.1 Useful resources

1. Movie posters are a great source of inspiration and universal layout conventions. The project Movie Posters By Numbers, is an interactive gallery of movie posters alongside descriptive statistics about the conventions used in their design.

2. A large number of books entirely devoted on the subject of layout grids exist already. Amongst them, 'Grid Systems in Graphic Design' by Josef Mülller-Brockmann (ISBN: 978-3721201451) and 'Making and Breaking the Grid' by Timothy Samara (ISBN:978-1592531257) are highly recommended for readers who would be interested in finding out more about grid layouts.

3. The website The Grid System is an ever expanding resource about Grid layout systems for both print and electronic media.

4. Two popular layouts for laying out web pages are Twitter's bootstrap and 960 pixels. The projects offer templates with the layout already set and leave placeholders for filling in the content.

5. Finally, automated layout is also emerging as a research field. For an example of this, see O'Donovan and Agarwala's paper on 'Learning Layouts for Single-Page Graphic Designs'.

1.1.2 Composing the visualisation

This section essentially makes a reference to Gestalt laws of perception and its objective is to group together a set of principles about structuring content 'properly'. 'Properly', is used here to denote content structured by taking into account evidence about the way human beings perceive and comprehend information visually. It has no connotations to aesthetics or what could be considered 'beautiful' or 'nice'. These are subjective qualities.

The objective of a composition is to maximise the perceived legibility and readability of the information presented. These two terms have a specific meaning [1]:

Legibility The term refers to *... a reader's ability to easily recognize letter-forms and the word-forms built from them. (We do not read by recognising one letter at a time, but by recognising the shapes of whole words and phrases)*

Readability The term refers to *... the facility and comfort with which text can be comprehended. Text with good readability must also be legible, but mere legibility does not make a text readable.*

The composition of a set of elements on a page affects the way they are perceived. In fact, Gestalt psychology, an early 20th century theory of psychology, has derived a set of laws

[1] Please see reference 2b (p 104) in section 1.1.4

from relevant experiments, that capture the way the human brain groups elements in a visual scene[2]:

1.1.2.1 The law of Proximity

Elements placed closely together tend to be perceived as a group

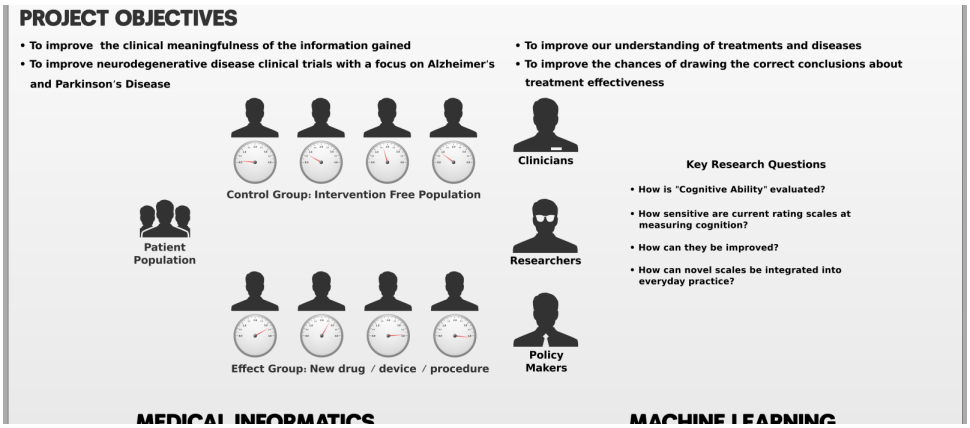


Figure 1.3: The law of proximity is demonstrated here on a poster fragment where, each instrument (representing a measurable quantity from a patient) is grouped with an avatar and each series of avatars further forms a group.

1.1.2.2 The law of Similarity

Similar elements tend to be grouped, this tendency can even dominate grouping due to proximity

1.1.2.3 The law of Closure

Elements tend to be grouped into complete figures

1.1.2.4 The law of Good Continuation

Elements tend to be grouped as to minimize change or discontinuity

1.1.2.5 The law of Symmetry

Regions bound by symmetrical boarders tend to be perceived as coherent figures

[2] Please see reference 2a (p 110) in section 1.1.2.7

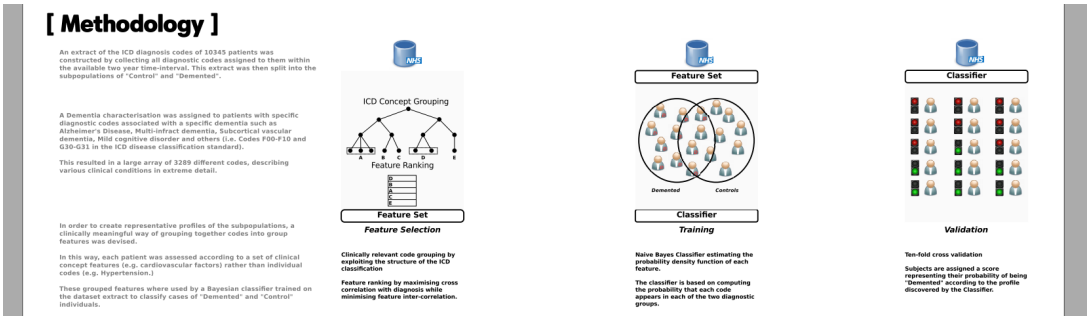


Figure 1.4: The law of similarity is demonstrated here on a poster fragment where, repeating traffic light and avatar symbols are perceived as single columns. An additional example of the law of similarity is available in figure 1.3

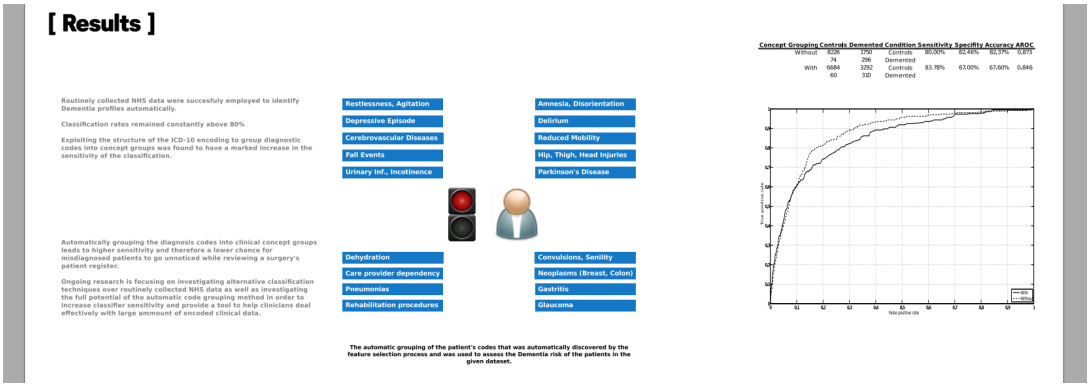


Figure 1.5: The law of closure is demonstrated here on a poster fragment where, sets of grouped labels appear as four distinct rectangles.

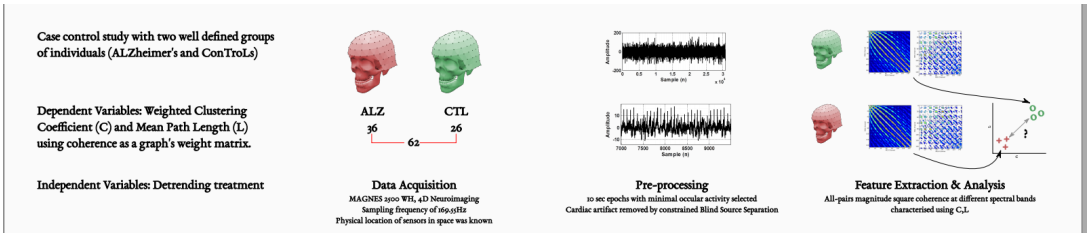


Figure 1.6: The law of symmetry is demonstrated here on a poster fragment where, the two symmetrical line fragments group the population of the experiment.

1.1.2.6 The law of Simplicity

Ambiguous elements tend to be resolved in favor of the simplest

1.1.2.7 Useful resources

1. Gestalt laws reveal a great deal about the way human perception works. An even greater deal is revealed by the way human perception misinterprets a scene. The lottolab

has compiled an extensive list of relevant research evidence that stand not only as impressive examples of misinterpretations of visual scenes but also as what to avoid when structuring a visualisation.

2. In terms of bibliography, the following books at the intersection of art and science are highly recommended:

   (a) Visual Perception by Bruce & Green (ISBN:0-86377-146-7)

   (b) Semiology of Graphics: Diagrams, Networks, Maps by Jacques Bertin (ISBN: 978-1589482616)

   (c) The visual display of quantitative information by Edward Tufte (ISBN: 978-0961392147)

1.1.3 The role of typography

Typography is another example of a vast field that due to its effortless availability through modern technology remains largely unappreciated. Just as an indication, apart from the general concept of a 'font-size' (which itself is not the most straightforward quantity to comprehend), there are another 26 terms [3] that refer to parts of a font and have their own importance in choosing a typeface.

Typefaces are essentially sets of symbols. The alphabetic part of them represents a particular sound, pictorially. Many symbols together form words, phrases, chapters, books and masterpieces of literature. But also, at the level of a single page, many symbols together create one large mosaic image with its own appearance, rhythm and characteristics.

Therefore, the choice of a typeface is not only affected by the practical objective of just printing some text. It also provides an overall look-and-feel to text zones and offers a way of capturing and controlling the reader's attention. As an example of this, please see figures 1.7, 1.8, 1.9 depicting the same poster fragment rendered with different fonts. These are actual variations that were made to assess the suitability of different font choices.

So, what are some basic considerations regarding typesetting? These are presented here in a multiscale way starting from the level of a single symbol and expanding outwards to larger bodies of text.

A Symbol (character)

- In typography, the geometrical extents of the characters vertically are determined by 5 parallel lines:

  Ascent The absolute highest point a character can reach

[3] Please see Anatomy of a typeface

Cap The highest line an upper case letter can reach

Baseline The line on which characters are resting

Descent The absolute lowest point a character can reach

- A typeface's vertical 'size' is determined by the distance between the Cap and Descent lines.
- The geometrical width of the characters is a slightly more complicated issue. From the point of view of a typeface's width, the classification is between fixed-width and variable-width typefaces. Fixed-width typefaces use the same width for all characters whether that is a thin **I** or a wide **M** and variable-width typefaces take the shape of the character into account.
- The overall size of a single character is another complicated issue in typography. In typography, characters are measured in 'points' (e.g. this text is typeset at 12-points). The definition of the 'point' has been varying since the first letterpress was invented but ever since computers took over the world [4], a 'point' is equivalent to $\frac{1}{72}$ of the ***international*** inch.
- Because of this fluid nature of determining the size of the symbols, the task of typesetting in different fonts is becoming even more difficult because typefaces of equal 'size' end up having different physical sizes.
- Symbol size is one of the strongest tools at the disposal of a designer to enforce hierarchy in a document and drive the attention of the reader. (Think about Title, subtitle, headings, key points).

A Set of Symbols (word)

- Symbols in *close proximity* are perceived as a set. In the case of type, this translates to 'words'. The matter of 'spacing' between symbols of the same set is a . . . slightly complicated issue in typography. Symbol spacing has two distinct varieties, 'tracking' and 'kerning'. Tracking is, intuitevely enough, the fixed width between symbols of the same group. However, due to their shape, some symbol combinations might appear further apart than intended. This in turns means that a fixed width between symbols might produce an uneven result where some symbols appear closer than others affecting the way they are perceived. The typical example is that between the letters **V** and **A**. Kerning is that *conditional* manipulation of tracking depending on symbol co-occurence.

[4] Please see Skynet

- Symbol spacing affects the way they are perceived visually. As a rule of thumb, small size symbols deserve larger spacing so that individual symbols can be perceived more clearly and vice versa. For an example of this, please see figure 1.10 and for another example please see figure 1.11.
- Setting symbol spacing is . . . a slightly complicated issue in typography. Specifically, spacing can be defined in proportions of **M**. A full **M** is the width of symbol **M** at a given size. Because history.

A Group of Symbol Sets (paragraph)

- The extra level of text organisation offered by the paragraph brings with it an additional set of parameters, mostly associated with spacing. Predictably enough, a paragraph of text has word-spacing and line-spacing, with a self-explanatory meaning. Similar considerations apply here as well regarding proximity and perception. Large word-spacing (compared to line-spacing) can end up in text being perceived as composed of columns of words, rather than rows and vice versa.
- Setting word-spacing is . . . a slightly complicated issue in typography. Specifically, word-spacing can be defined as the space occupied by a lowercase **i**. It is worth noting here that word-spacing should be set at the least amount required to distinguish words but not further than that.
- Overall, a paragraph also has a width that can be measured in units of distance (e.g. mm), **M** units (e.g. number of characters per line) or average number of words per line. As a rule of thumb, human beings tend to comprehend text better when it is no more than 50-80 symbols per line wide with words having between 5 to 10 letters on average. These guidelines are not absolute but provide an excellent starting point.
- Finally, a paragraph also has 'alignment' and 'rag' (which is a by-product of alignment). Paragraph alignment refers to the alignment of each individual line of text within the paragraph's 'virtual' bounding box. Paragraphs where text always begins at left are left-aligned, paragraphs where every line's mid-point is adjusted to the mid-line of the bounding box are center-aligned and paragraphs where ever line ends exactly at the right bounding line are right-aligned. It is worth noting here the 'justified' alignment which appears to be achieving simultaneously left and right alignment. This is achieved by constantly modulating word-spacing in a line of text and if taken to the extreme it can have a detrimental effect to legibility. The aligned side, creates a perceived clean vertical line where characters either

start from or end at. The opposite side of that appears un-even depending on text content. That un-even effect is called 'rag'. The rag's shape is determined by a number of factors but the objective is to achieve a consistent 'shallow' rag giving the impression of a mostly *continuous rectangle of text*.

1.1.4 Useful resources

1. Identification and 'proper' use of font families is a difficult task, especially for the uninitiated. It can also become quite an addictive habbit. Whatever the case, a small set of links that can help in identifying font-families is as follows:
   (a) Fonts in Use and associated Flickr pool and forum.
   (b) What the font, is a similar website with a slightly more automated workflow for identifying font families.
2. In terms of bibliography, the following books on typography are strongly recommended:
   (a) Typography Workbook by Timothy Samara, (ISBN:978-1-59253-301-5)
   (b) The complete manual of typography by James Felici (ISBN: 0-321-12730-7)
   (c) Digital typography by Donald Knuth (ISBN:1-57586-010-4)

1.1.5 The role of colour

Colour is another huge discussion involving many subtle points such as capturing, describing, manipulating and reproducing colour consistently across different media. And this would be just the part covering the technical points involved in working with colour. Another part, of equal if not greater magnitude, would be the one involving the human-factor and how do humans perceive and react to colour, involving elements of their culture.

The importance of colour is paramount to data visualisation. From the point of view of creating representations of data, the key objective is to make the message clearly perceived and this immediately hints at contrast. In fact, the relative brightness of an element in a scene, does not only depend on its colour but also its 'frequency' of usage. In other words, less frequently used colours will appear darker in a composition, even if they would be considered as 'bright'.

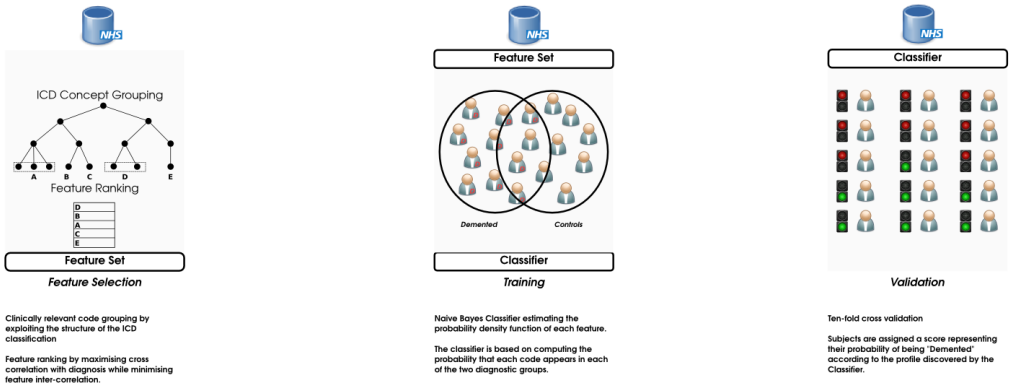
Here is a set of key considerations and observations regarding colour:

- Coloured elements are perceived as 'floating' in space with blue appearing to recede, yellow to advance and red to be floating somewhere in the middle of blue and yellow.
- Colour and size can be used to indicate hierarchy and guide the viewer's attention. An example of this is depicted in figure 1.12.
- Specific colours are associated with universal and culture specific meanings. For a comprehensive guide, please see figure 1.13 reproduced from the Information Is Beautiful website.
- Colour models are used for the purpose of describing particular colours. These are mathematical models that describe a colour as a set of parameters in a mathematical model. For example, in the RGB colour model, colours are described as triplets of Red, Green and Blue colours, while in the HSV colour model, colours are described as triplets of Hue, Saturation and Value (a.k.a Brightness) values. Other colour models are the Yellow, Magenta, Cyan (YMC) and Cyan, Magenta, Yellow, blacK (CMYK).
- Visualisations of colour models can be helpful in depicting relationships between colours. One such visualisation is the colour wheel. It depicts the three primary colours (Red, Yellow, Blue) and their combinations and it can be used to create colour-schemes.
- Specific colour models such as the Pantone colour model have become standards in representing colour faithfully across media. Choosing such a standardised colour (for instance a 'True Red' or a 'Sand Dollar') for a design guarantees that it will be reproduced faithfully across calibrated devices (monitors but more importantly printers).
- A colour scheme is a combination of colours using some criterion for their selection. As a rule of thumb, regular n-polygons inscribed to a colour wheel result in the selection of n 'matching' colours.

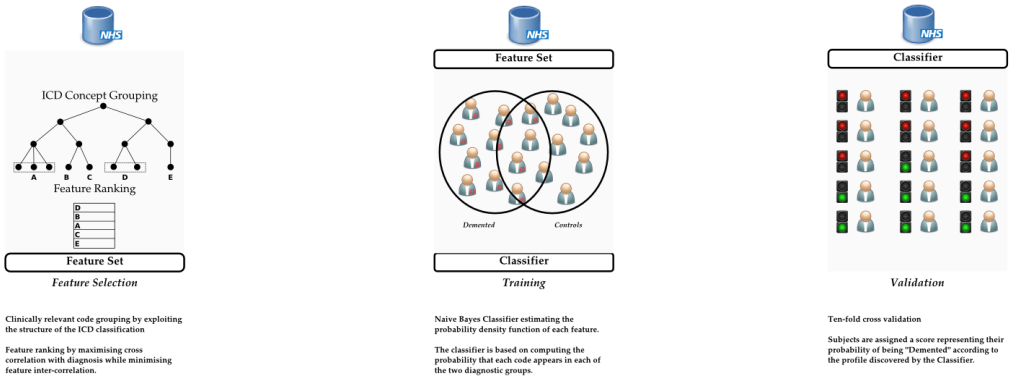
1.1.6 Useful resources

1. Similarly to choosing a font, choosing a colour combination can be a daunting or addictive task. The following list brings together a number of relevant useful tools:
   (a) I want hue is perhaps one of the best tools specifically designed for the selection of harmonious and distinct colours for data visualisation purposes.
   (b) Coolors takes a more 'evolutionary' approach to colour scheme generation.

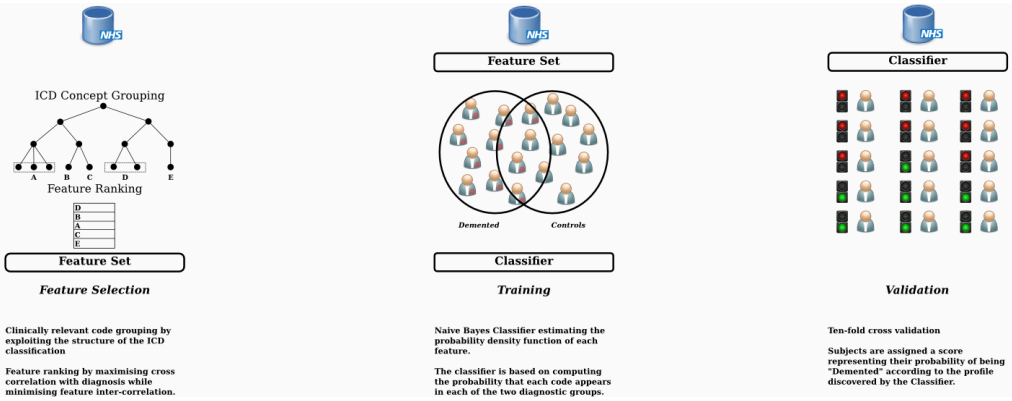
(c) COLOURlovers is a very simple colour scheme ‘gallery’ of user (human) generated colour scheme.



(a) TeXGyreAdventor

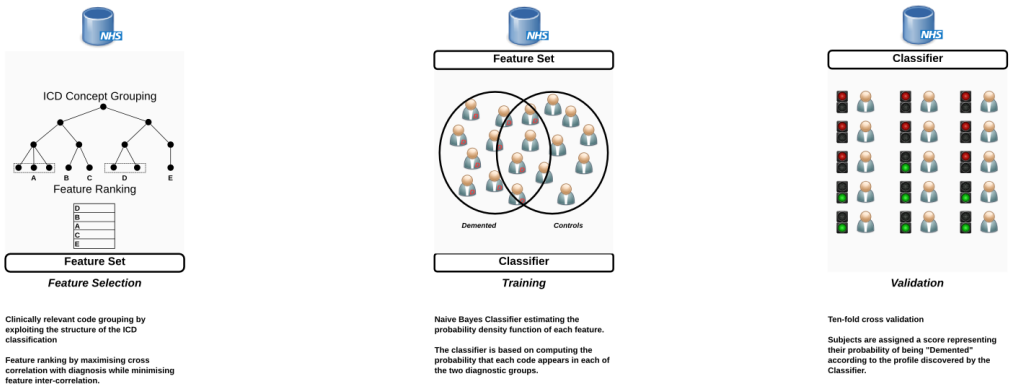


(b) URW Palladio L

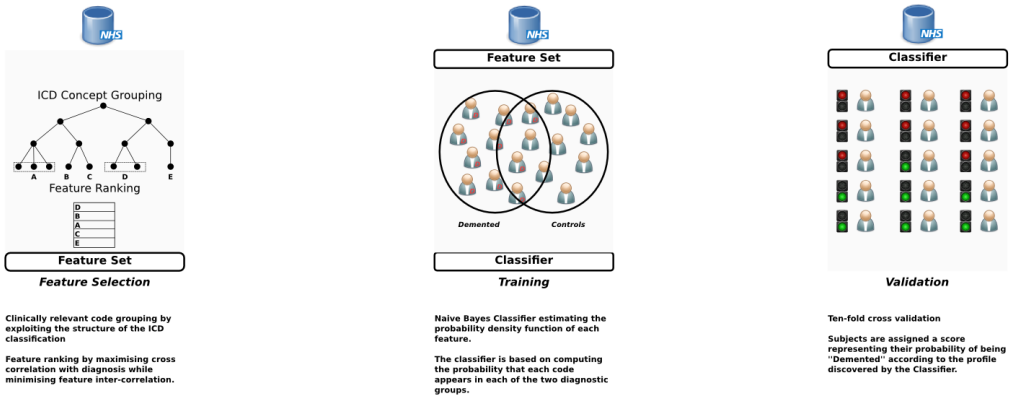


(c) DeJaVu Serif

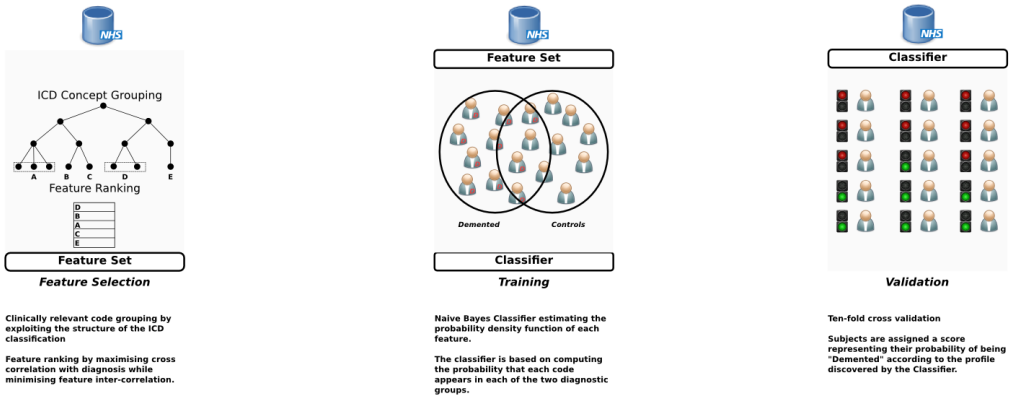
Figure 1.7: The same poster fragment rendered with three different fonts, notice how the appearance of text, including text flow, justification and number of words per paragraph, changes. All 8 variations listed in these figures were actually considered as alternatives for the final result that made it to print. The final result made use of 'Dejavu Sans' and this choice represented the majority vote amongst the co-authors. All fonts freely available from the websites mentioned in the resources section.



(a) Liberation Sans

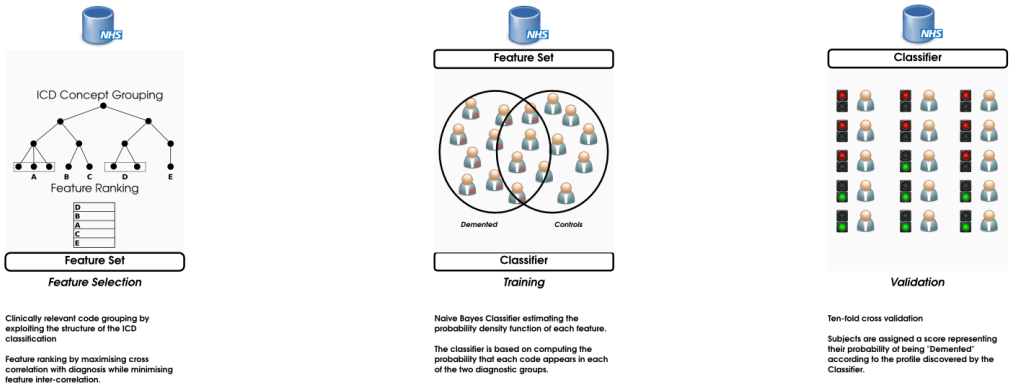


(b) esint10

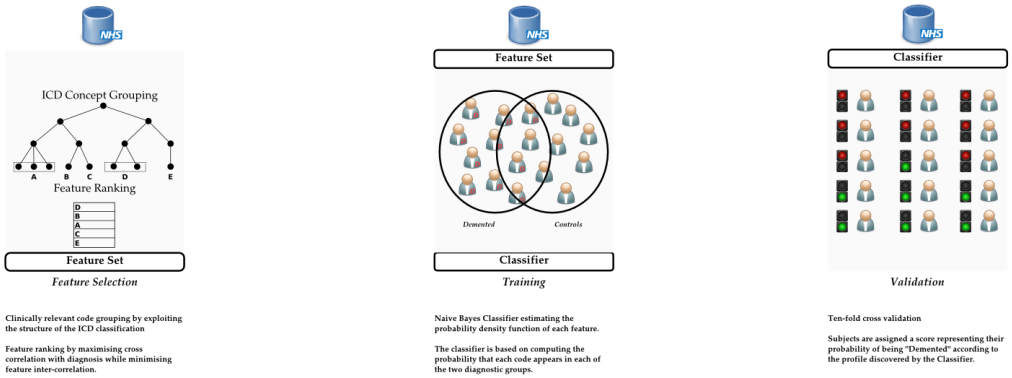


(c) DejaVu Serif

Figure 1.8: The same poster fragment rendered with another three different fonts.



(a) DejaVu Sans



(b) URW Bookman L

Figure 1.9: The same poster fragment rendered with a final two different fonts.

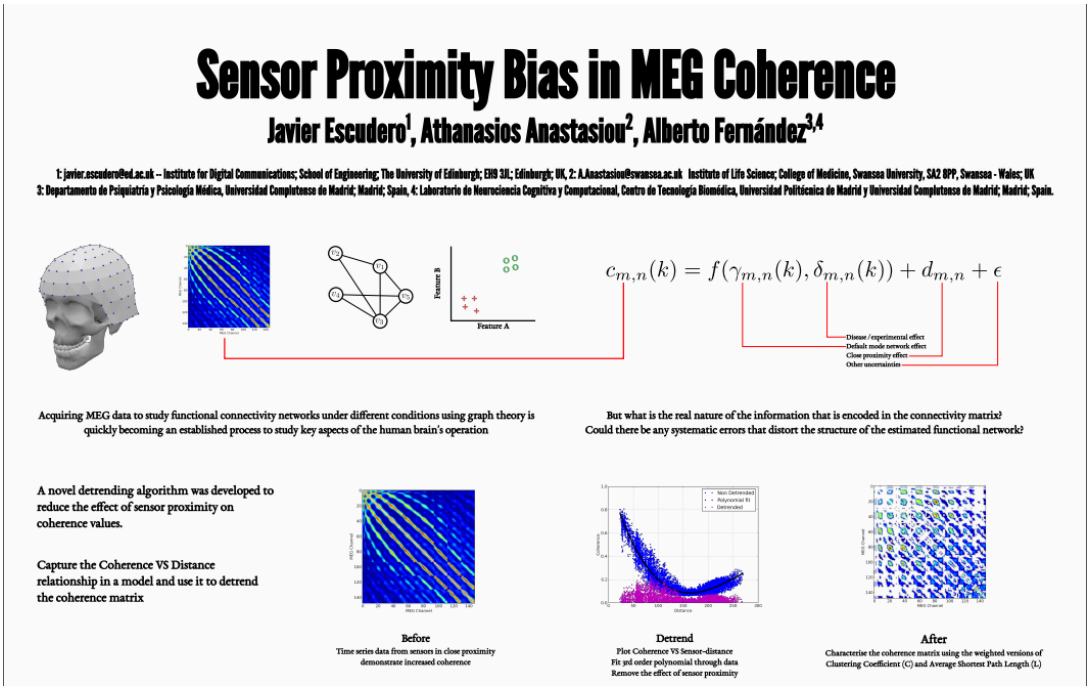


Figure 1.10: An example of reducing symbol spacing depending on symbol size.

Figure 1.11: Poor choice of symbol spacing can have a catastrophic result on perception. Image reproduced from a retweet by Catherine Dixon.

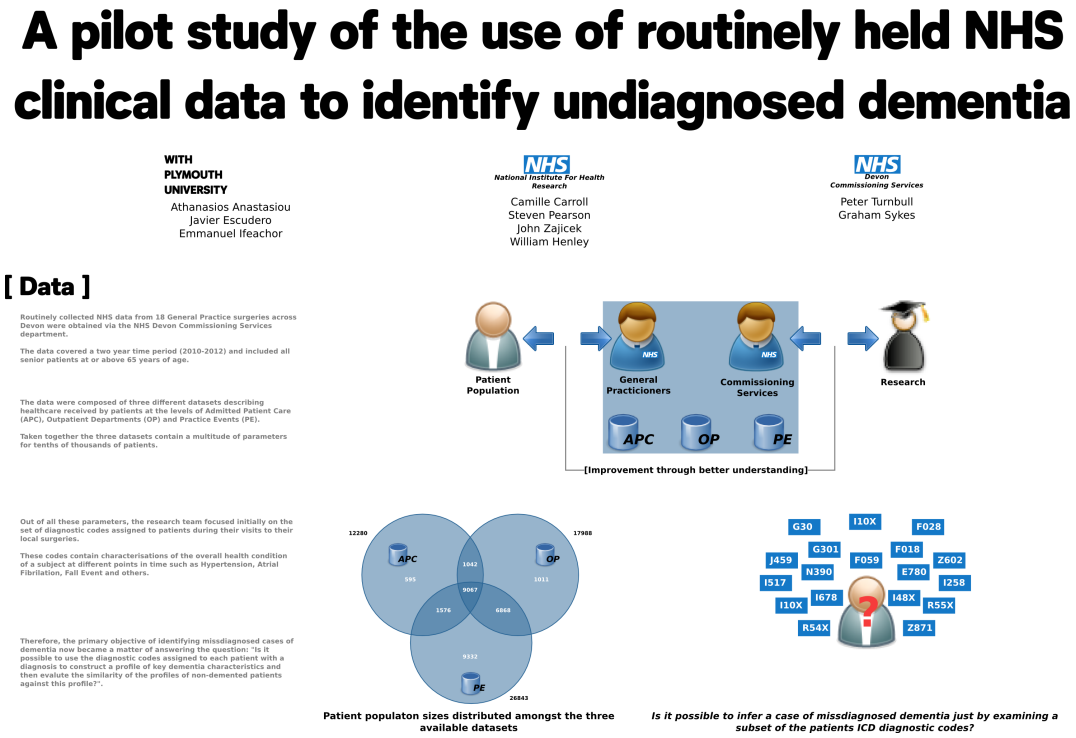


Figure 1.12: Hierarchy through the use of size and colour. The title, header and text are typeset in different sizes and colours which make them legible from different distances.

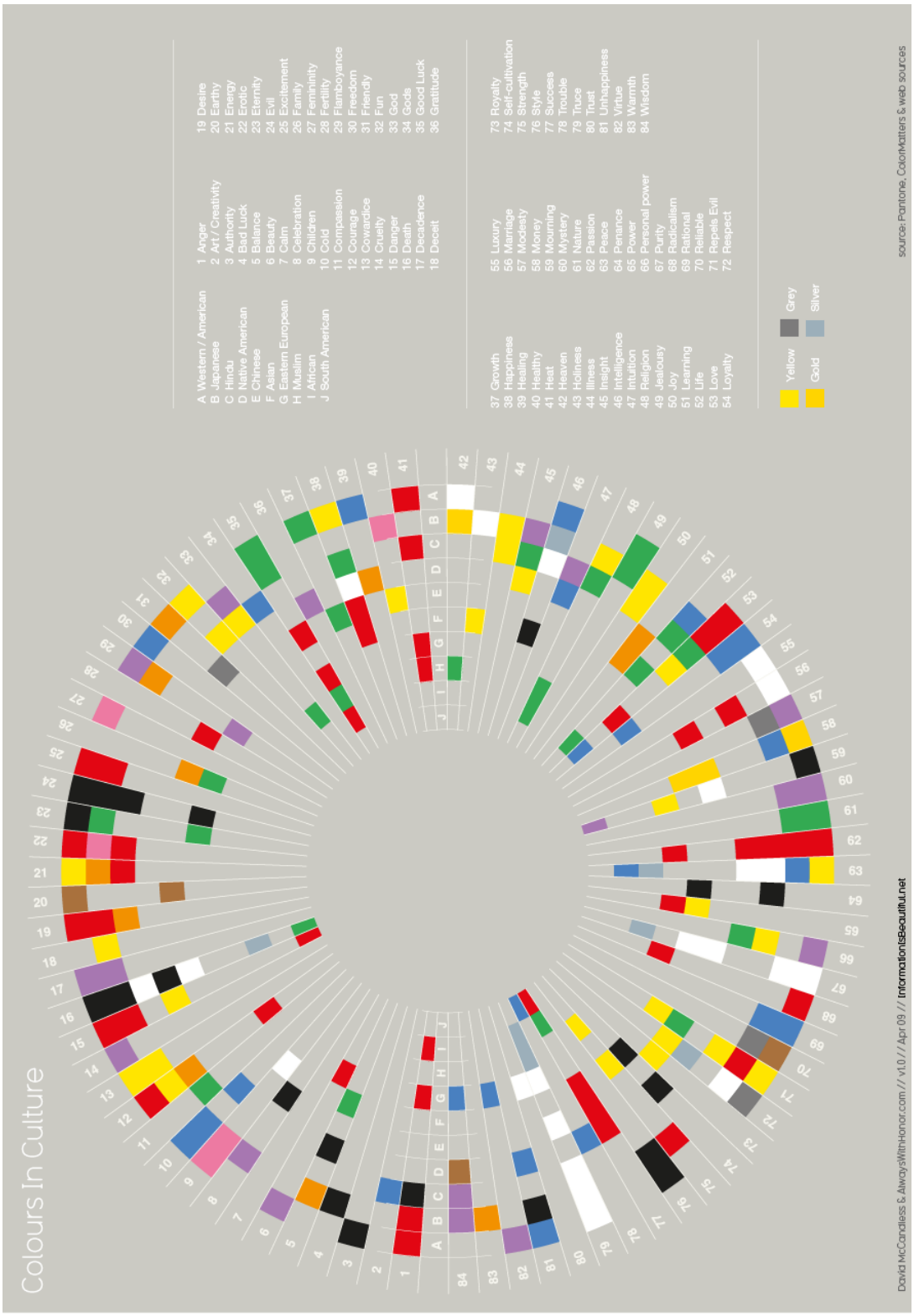


Figure 1.13: Colours and their different meanings across different cultures. The image is reproduced from the Information is Beautiful website.

Chapter 2

Visualising Data With Tableau Public

INTRODUCTION

The objective of this chapter is to introduce 'Tableau Public' to the complete novice. The chapter begins with a brief but thorough exposition to basic concepts governing the operation of the 'free' version of Tableau, codenamed 'Tableau Public' [1] and proceeds with basic operations that showcase its functionality.

From an educational point of view, the objective of this chapter **is** to help readers with their first steps in visualising data with Tableau Public. The objective of this chapter **is not** to turn the reader to a complete Tableau Public expert.

2.1 OBTAINING & INSTALLING TABLEAU PUBLIC

'Tableau' is the name chosen by 'Tableau Software' for its complete data visualisation solution and the various different suffixes it appears with denote different versions and capabilities.

This series of 'Health Data Visualisation' sessions is designed with 'Tableau Public' in mind. This is a freely available version of Tableau with a minimal but still powerful set of capabilities available to a researcher.

This version can be obtained from Tableau's website in exchange for a valid email address. The same page contains a brief comparison of the three different versions of Tableau where the limitations of each are outlined but the most interesting of those are also listed below:

- Tableau Public can accept input from a number of different data sources such as Coma Separated Files (CSV), Microsoft (MS) Access databases, MS Excel spreadsheets and

[1] Please see: `https://public.tableau.com/s/`

remote data servers conforming to the Open Data Protocol.

- Irrespectively of the data source, the 'Public' version imposes **a limitation of 1,000,000 (one million) cases to the amount of data it can handle**.
- Tableau Public **cannot save its output to a local file** in the disk of the computer running the software. Visualisations produced by the 'Public' version are saved online, at Tableau's servers and can be turned into interactive presentations that can be accessible by the public via Tableau's Gallery or a direct link.
- The online space available to 'Tableau Public's' users has a capacity of 1GB.
- Due to the fact that the 'Public' version **cannot save its output to a local file** but only an online location, it is important to confirm that visualisations do not pose any **disclosure risk** to the research. It is also important to go through "Tableau's' data and privacy policy as well as the relevant terms of service.

Installer packages are available for the MS Windows and Apple Mac OS operating systems. This guide assumes the use of the MS Windows version and the typical 'Tableau Public' download (version 9) in this case is approximately 123 MB. An additional 500MB of disk space will be required for a complete installation.

There are no special points of note regarding the installation process which should proceed smoothly.

2.2 BASIC CONCEPTS

Just like any kind of software, Tableau Public deals with a particular type of 'Object'. The 'Object' of this piece of software is the **Visualisation Story** (from here onwards referred to simply as, the **Story**).

From Tableau Public's point of view, a **Story** can be made up from an individual **Sheet** or a **Dashboard**.

Good knowledge of these top level concepts, as well as a small number of secondary ones, is essential for putting together an effective visualisation through Tableau Public. These concepts and their relationships are summarised in figure 2.1 and the most important of those are explained in more detail in the following subsections.

The information that each of these objects holds is modified via a **View** of the same name (e.g. The **Story View**, the **Dashboard View** and so on). The **View** is essentially the Graphical User Interface (GUI) that the software uses to communicate with the user.

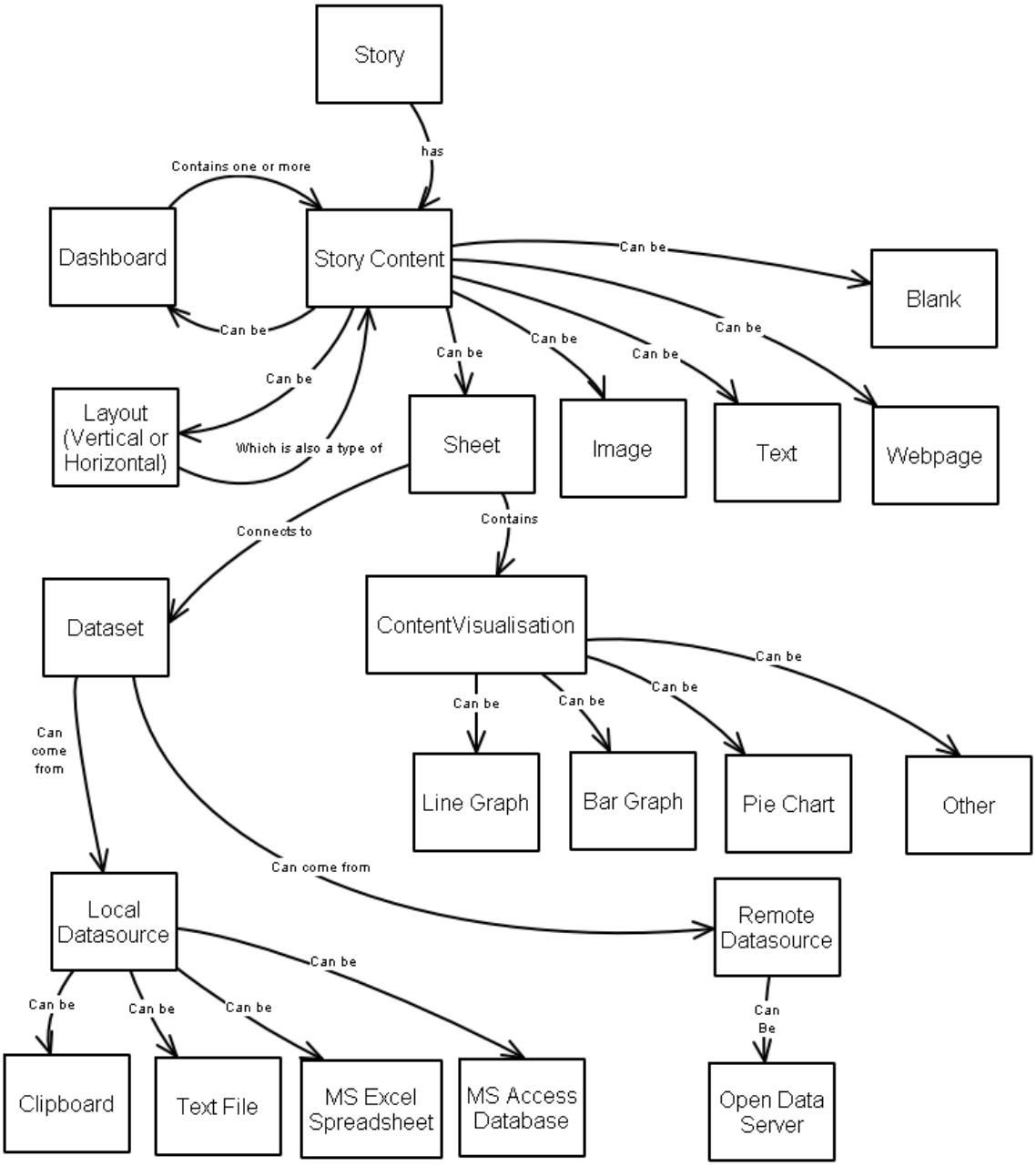


Figure 2.1: A conceptual diagram of the entities that make up a visualisation in Tableau Public

2.2.1 The Visualisation Story

A **Story** can be made up of a single **Sheet** or **Dashboard** and is primarily defined via a small set of attributes:

Title A short but descriptive title of the visualisation.

Key Points A set of zero or more key points that provide the reader / viewer of the visualisation the context regarding the story.

Description(s) A set of zero or more descriptive pieces of text with full control of their appearance (i.e. font, colour, placement and other)

Size The size of the story is an important parameter when it comes to rendering the visualisation in different media. Tableau already offers a wide range of popular dimensions to choose from such as 'A4 Landscape', 'A3 Landscape', 'iPad landscape' and others.

The placement of these elements to the **Story View** is depicted in figure 2.2.

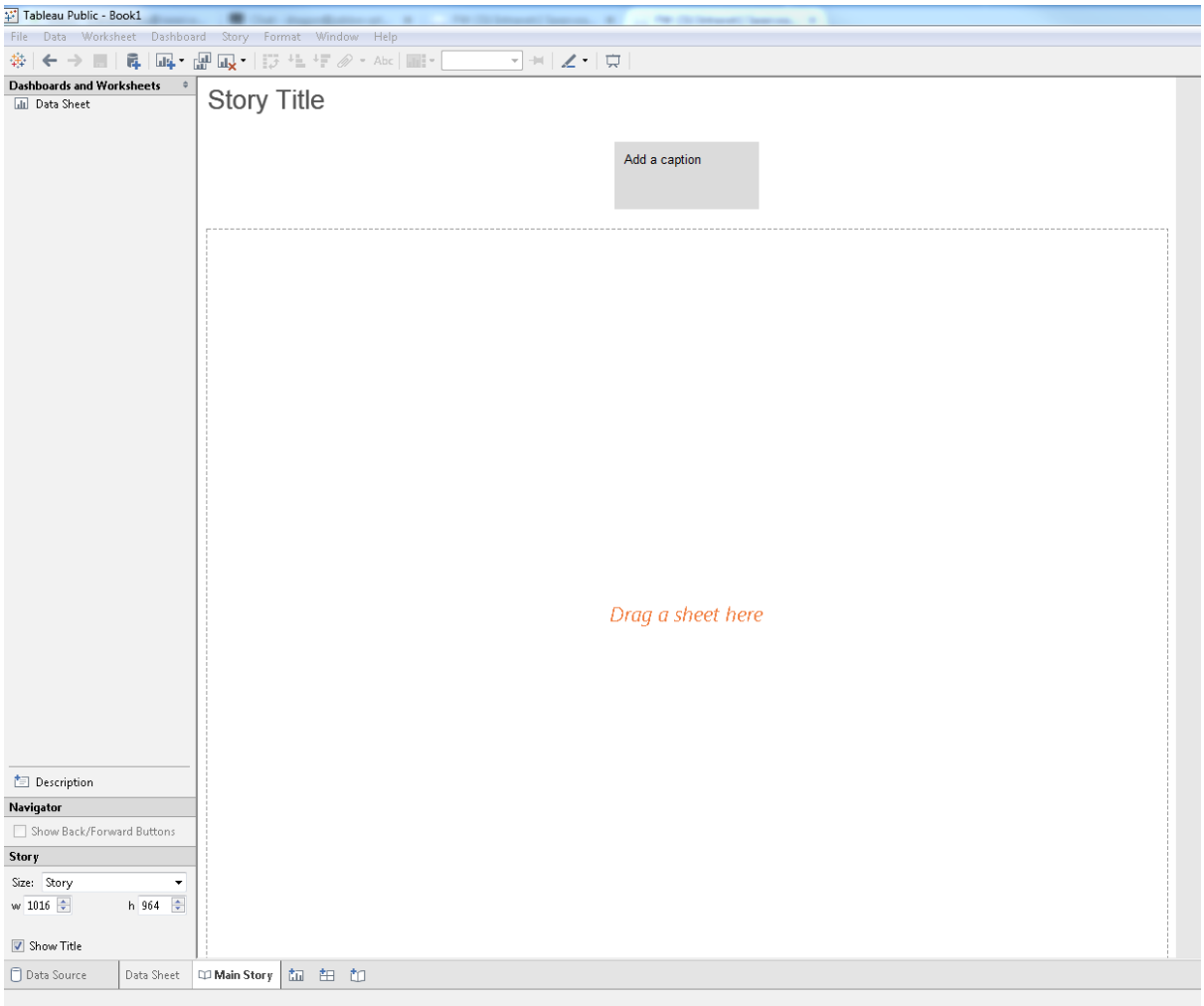


Figure 2.2: The Story View through which a visualisation story is constructed

2.2.2 The Dashboard

A **Dashboard** can be made up primarily from a set of one or more **Sheets** which are laid out either in vertical or horizontal order. Other elements that can appear in a layout along with **Sheets** are *Text* (with fully customisable appearance) , *Images, Web pages* (i.e. the rendered view of a website) and finally *Blank* elements.

The placement of these elements to the **Dashboard View** is depicted in figure 2.3.

2.2.3 The Sheet

A **Sheet** is one of the most important concepts in Tableau Public and the accompanying view (the **Sheet View**) is the one that users probably spend most of their time on when creating a visualisation story.

A **Sheet** maintains connections to the **Data Sources** and also handles their **Content Visualisation**. Data sources and content visualisation have a number of important concepts associated with them and is worth expanding on those in the following paragraphs.

2.2.3.1 Data Source

The Public version of Tableau can connect to various **local** (in the form of a file) and **remote** (a server) data sources.

Local access is limited to the following file data sources:

- Clipboard (e.g. Copying tabular from another software and pasting them into Tableau).
- Microsoft (MS) Excel spreadsheet
- Text File (e.g. Coma Separated Value (CSV), Tab Delimited (TSV) and others)
- MS Access database.

Remote access is currently limited to the Open Data Protocol [2]

In addition to these limitations, please note that the Public version of Tableau has a data volume limitation of 1 million rows per datasource.

The datasets retrieved from data sources are very similar in structure to a spreadsheet with a number of columns representing different **Cases** (or Samples) and a number of rows representing different **Attributes** (or Features, or Variables) for each case.

[2] For more information please see `https://en.wikipedia.org/wiki/Open_Data_Protocol`

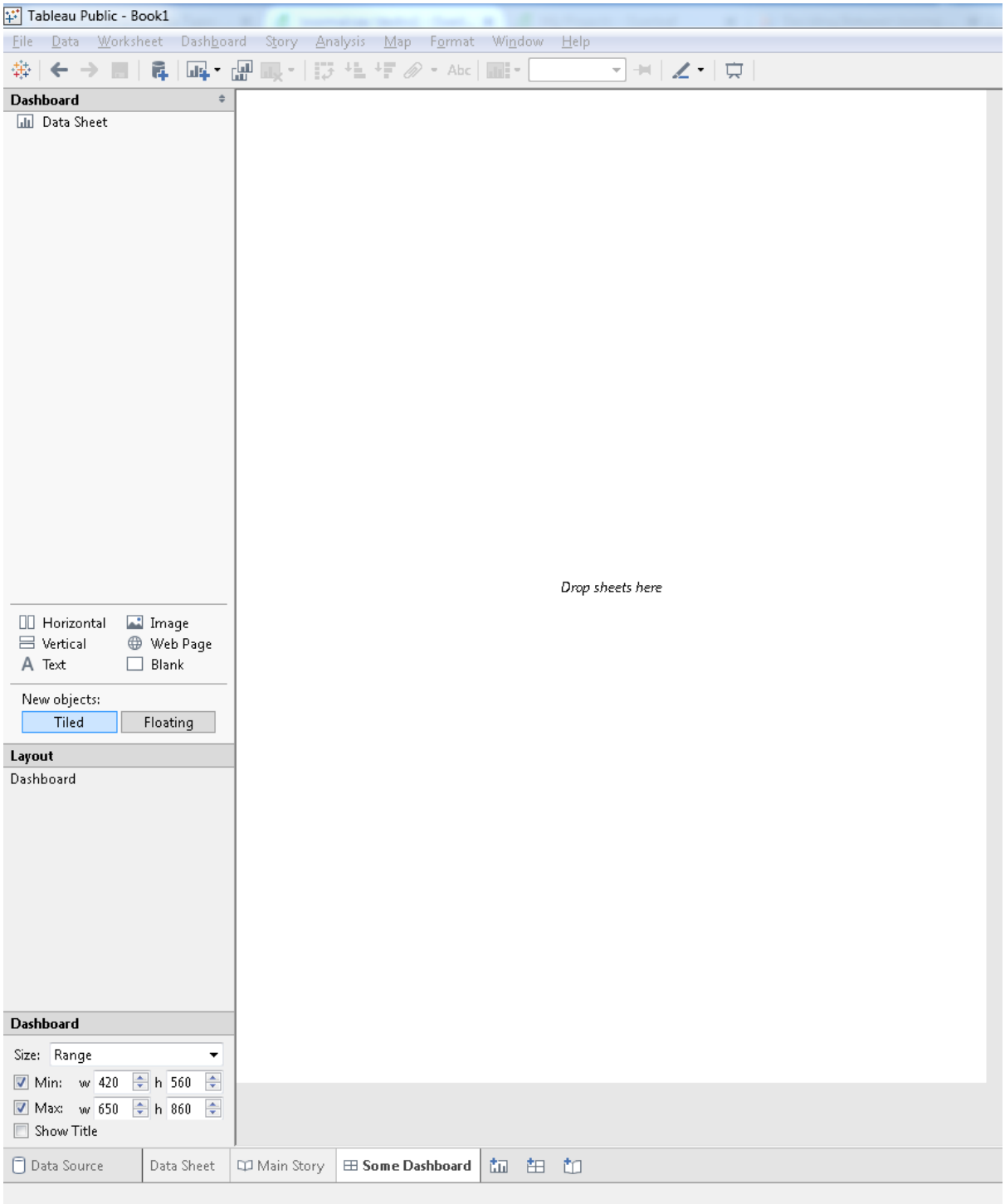


Figure 2.3: The Dashboard View through which a set of Data Sheets get laid out

In addition to commonly used data types such as numeric, date and string, Tableau uses another classification for data which helps when structuring a visualisation.

Attributes are primarily divided into:

Dimension The word 'Dimension' is used within Tableau to denote categorical data.

Measure The word 'Measure' is used within Tableau to denote ratio types of data (e.g. a quantity

The placement of these elements to the **Sheet View** is depicted in figure 2.4

2.2.3.2 Content Visualisation

Tableau supports a wide range of visualisations out of the box such as tables, heat maps, pie charts, bar and line plots, geographical maps and others (please see figure 2.5).

For data to be represented and visualised by these elements, they first need to be specified as sets of **Visualisation Rows** and **Visualisation Columns**. The definition of these Rows and Columns is different from their data counterparts. Essentially, a **Visualisation Row / Column** can contain any **Attribute** from the dataset or a simple transformation of it.

In this way, it is possible to construct complex multidimensional visualisations and also apply aggregate functions on the attributes of the dataset. Out of the box, Tableau supports the Sum, Average, Median, Count, Count(Distinct), Min, Max and St.Dev / Variance.

Furthermore, for each of the chosen visualisation elements, additional information (or context sensitive information) might appear that enables further customisation.

Finally, Tableau offers Filtering and Joining functionality which can be used to isolate, merge and visualise parts of a dataset. This functionality is described in sections Joining Two Data Sources and Filtering Data.

The placement of these elements to the GUI is depicted in figure 2.6

2.3 BASIC OPERATIONS

The objective of this section is to introduce a (small) set of operations that are elementary to the operation of Tableau Public. These are presented here in the form of very simple walk-throughs with additional guidance wherever this is required.

2.3.1 Loading A Data File

Loading a Data File or connecting to or adding a **Data Source** to a **Sheet** is a relatively straightforward process:

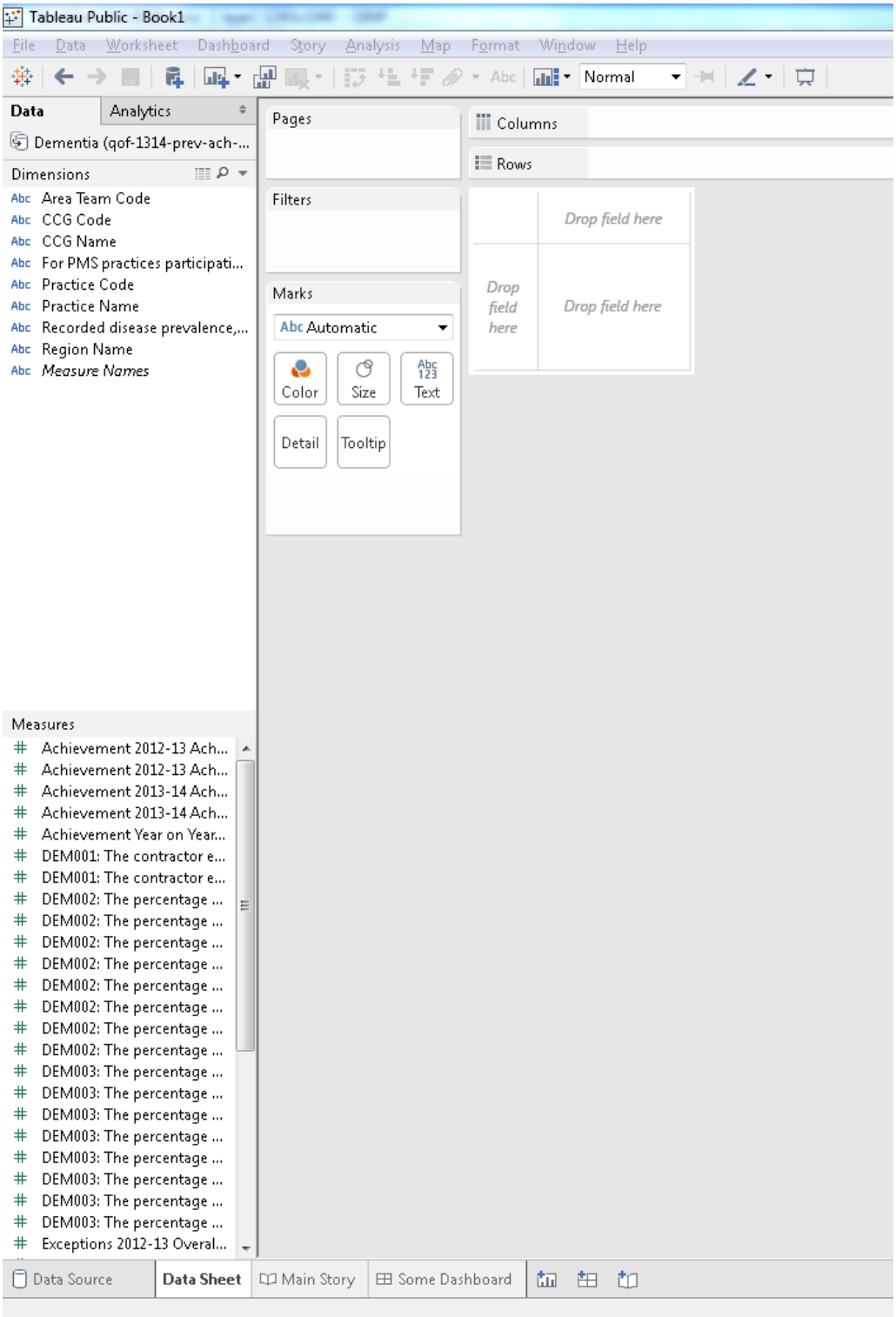


Figure 2.4: The Sheet View through which the actual visualisation and filtering of data is performed

Figure 2.5: Different ways of representing data available in Tableau Public.

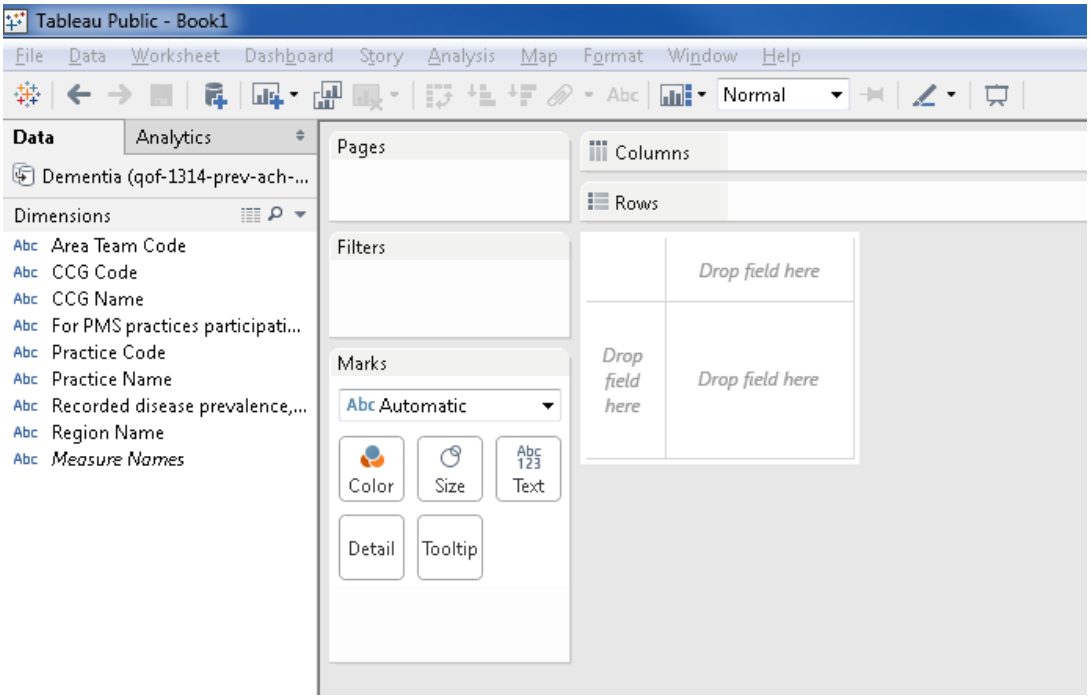


Figure 2.6: A focused aspect of the Sheet View showing the 'Dimensions' and 'Measures'. Also visible in this aspect, are the 'Filter' and 'Marks' controls that enable data filtering and content visualisation customisation.

1. Switch to the Data Sheet View.
2. Click on the database symbol bearing a plus sign.
3. From the dialog box that opens, navigate to the appropriate location on the disk.
4. Select the correct File Type.
5. Select the required Data File.
6. Click Open.

This will start the data import process which takes place in a separate window (please see figure 2.7).

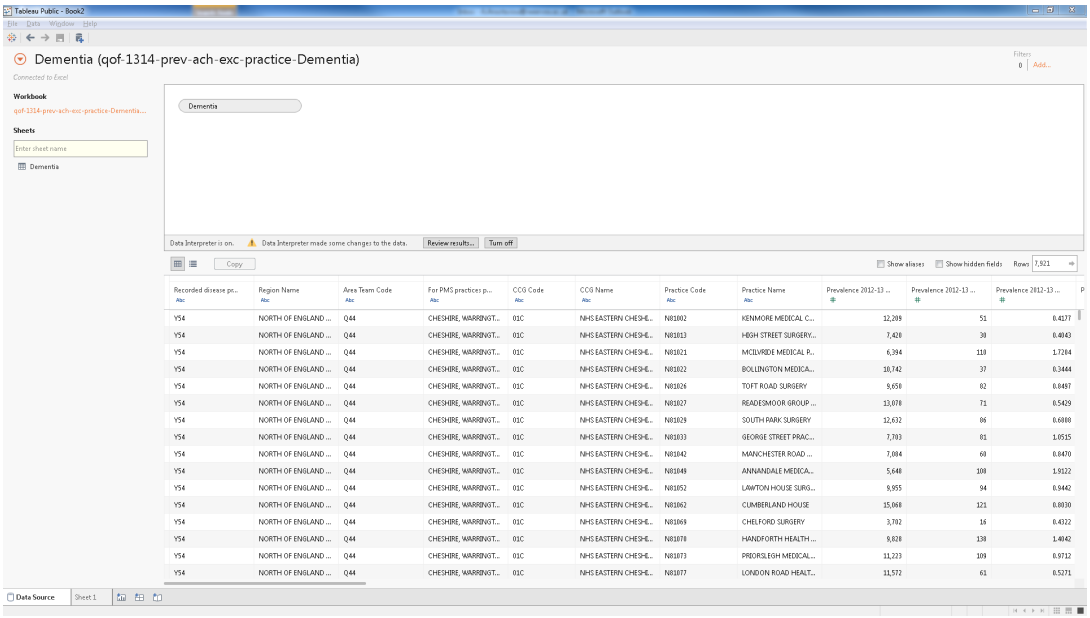


Figure 2.7: The Data Import View which governs the process of importing a dataset from a selected datasource. This view allows the user to select a subset of the dataset's attributes or cases to be imported. Of particular note in this example is that the QoF dataset is loaded with the 'Data Interperter' option turned on.

The example depicted in figure 2.7 was generated by importing the 'Dementia' QoF dataset.

Please Note: To import QoF data it will be necessary to enable Tableau's *Data Interperter* which is required to make sense of the data format.

2.3.2 Joining Two Data Sources

To establish a relationship, or *join*, two data sources, these first have to be already loaded. For this purpose, please follow the section Loading A Data File twice to load two different datasets.

Once two (or more) datasets are loaded within Tableau:

1. From the application menu, click on 'Data'
2. From the 'Data' sub-menu click on 'Edit Relationship'
3. From the dialog box that opens, establish the source and target fields that will be used to connect the datasets.

Please Note: Tableau Public will attempt to establish relationships automatically on the basis of the dataset's attribute names. However, to establish a relationship between two attributes that do not share the same name across two different data sets, the option 'Custom' must first be selected. The placement of these options is depicted in figure 2.8.

2.3.3 Producing Calculated Attributes

Calculated attributes are attributes that do not originally exist within the data but are created as the result of expressions between existing attributes.

To create a calculated attribute, a dataset must already have been connected to a Sheet (To do this, please see Loading A Data File). Once this is done:

1. Right-click on a **Dimension** or **Measure** of the dataset.
2. From the context menu that appears, click on 'Create'.
3. From the sub-menu that appears, click on 'Calculated Field'.

From the dialog box that opens, specify the name of the new attribute and a mathematical expression that defines its value as a function of existing attributes and finally click 'Create'. The process is depicted in figure 2.9 where a new field that is the average of two values is created.

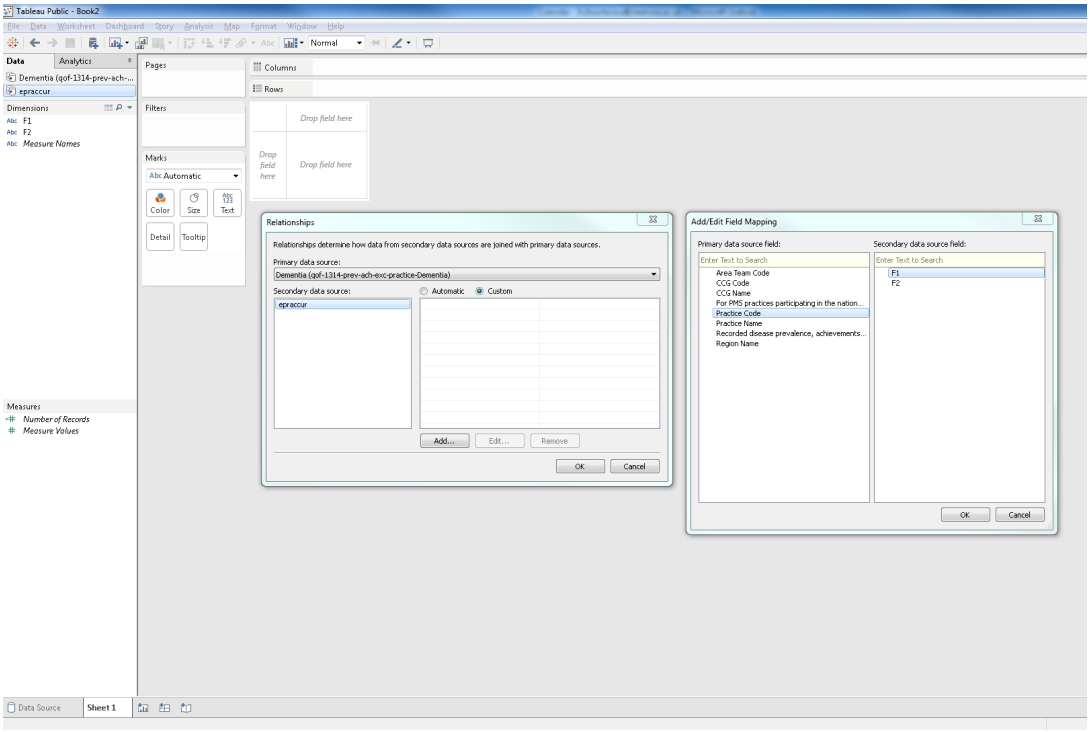


Figure 2.8: Two dialog boxes that open sequentially (from left to write) and are used to establish relationships between datasets. In this example, a relationship is established between the 'General Practice Code' and its counterpart in the 'General Practices List' which contains much more information about the General Practice than it is available in the original QoF dataset.

2.3.4 Composing A Very Simple Visualisation

One of the simplest visualisations on the Dementia QoF dataset is that of disease prevalence across different regions of the United Kingdom (UK).

Before following the steps below, the 'Dementia QoF' dataset must have been loaded by following the process described in Loading A Data File

Once the dataset is loaded:

1. Drag and drop the dimension entitled 'Region Name' to the 'Columns' of the visualisation. This will be used as the X-Axis of the visualisation.

2. Drag and Drop the measure entitled 'Prevalence 2013-14 Prevalence (per cent)'. This will be used as the Y-Axis of the visualisation. **Please Note:** The original dataset contains data at the level of 'General Practices' and each 'Region Name' contains more than one 'General Practice'. Tableau public will therefore apply, automatically, an aggregation function so that one single value is produced. The default aggregation function is 'SUM'. The result of these actions is depicted in figure 2.10.

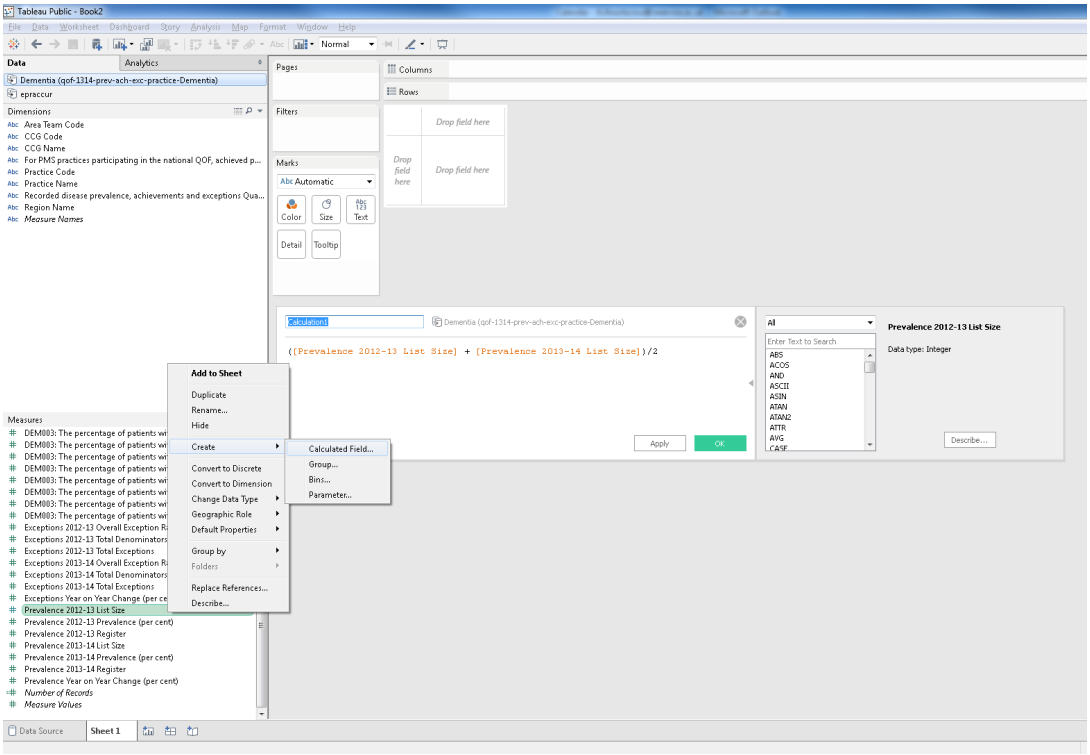


Figure 2.9: Creating a new calculated field as the mean between two existing fields. In this example, the calculated field will be the average value of two existing fields (i.e. the average list size of a general practice between two consecutive years).

3. Prevalence is expressed as a percentage and no assumptions should be held regarding the underlying distribution of the data. For this reason, it would be better to switch this function to the 'MEDIAN'. To do this, click on the small downwards pointing arrow of the 'Row' and select 'Measure / Median'. The result of these actions is depicted in figure 2.11.

In this example, Tableau Public has made a number of decisions for the user. The first of which was the type of visualisation which was automatically set to a 'Bar chart'. Further customisation is possible and it is described in section Customising a visualisation.

2.3.5 Filtering Data

Tableau public offers a very simple way by which data that are represented within a visualisation can be filtered. For a filter to become effective, it has to be added on an existing visualisation, for this reason, please see Composing A Very Simple Visualisation.

Once this task is over:

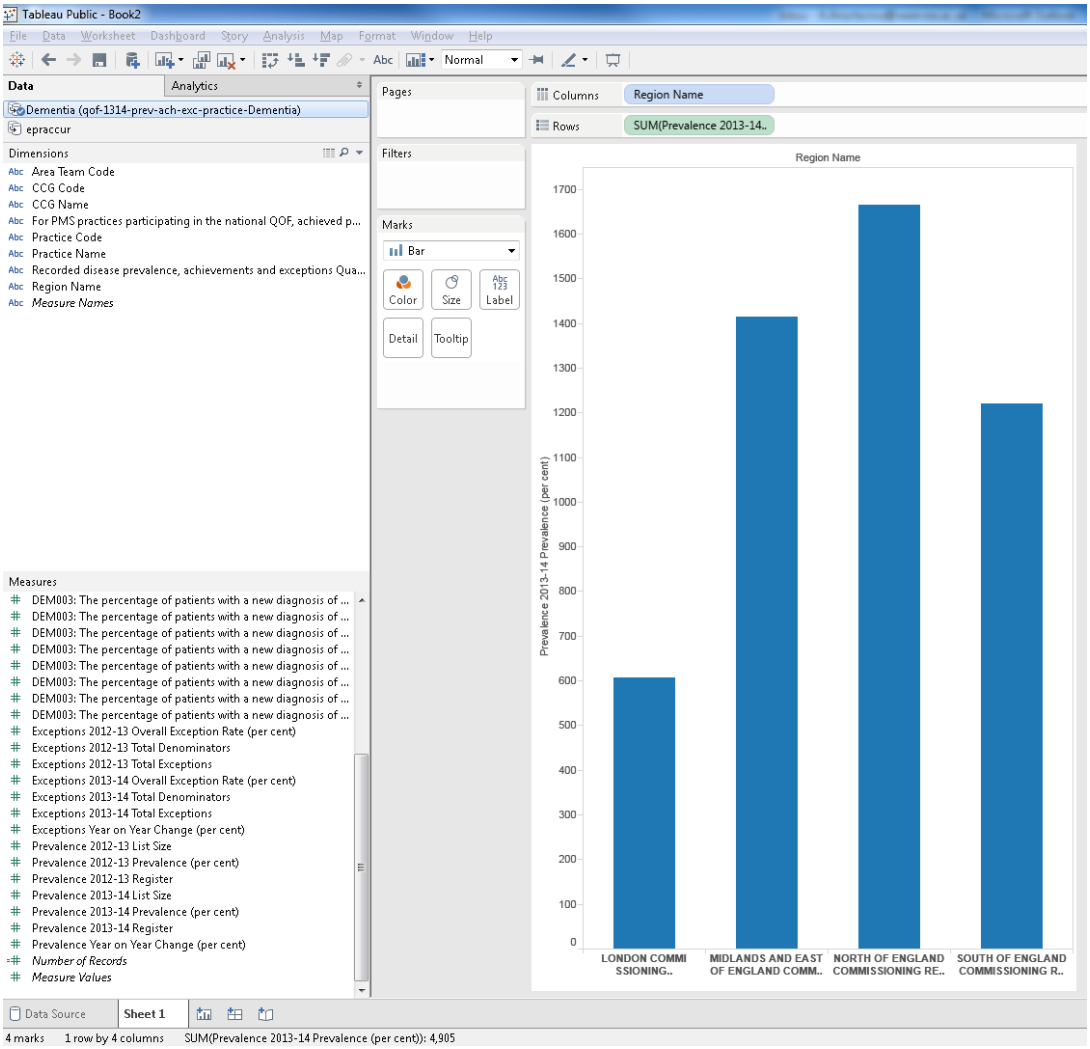


Figure 2.10: Creating a very simple bar chart that shows the SUM of Dementia Prevalence across different UK regions.

1. Drag and drop a **Dimension** or **Measure** from the original dataset to the 'Filter' box. For this example, a filter will be used to exclude the 'London Comissioning Region'...for no particular reason, other than to demonstrate the filtering functionality. To do this, drag and drop the 'Region Name' dimension to the 'Filters' box. This will automatically open the 'Filter' dialog box which is depicted in figure XX and allows the full specification of an expression that is used to filter the data. To get back into the 'Filter' dialog box at any time:

   (a) Click on the little downwards arrow that appears next to the **Attribute's** name.

   (b) Click 'Filter'.

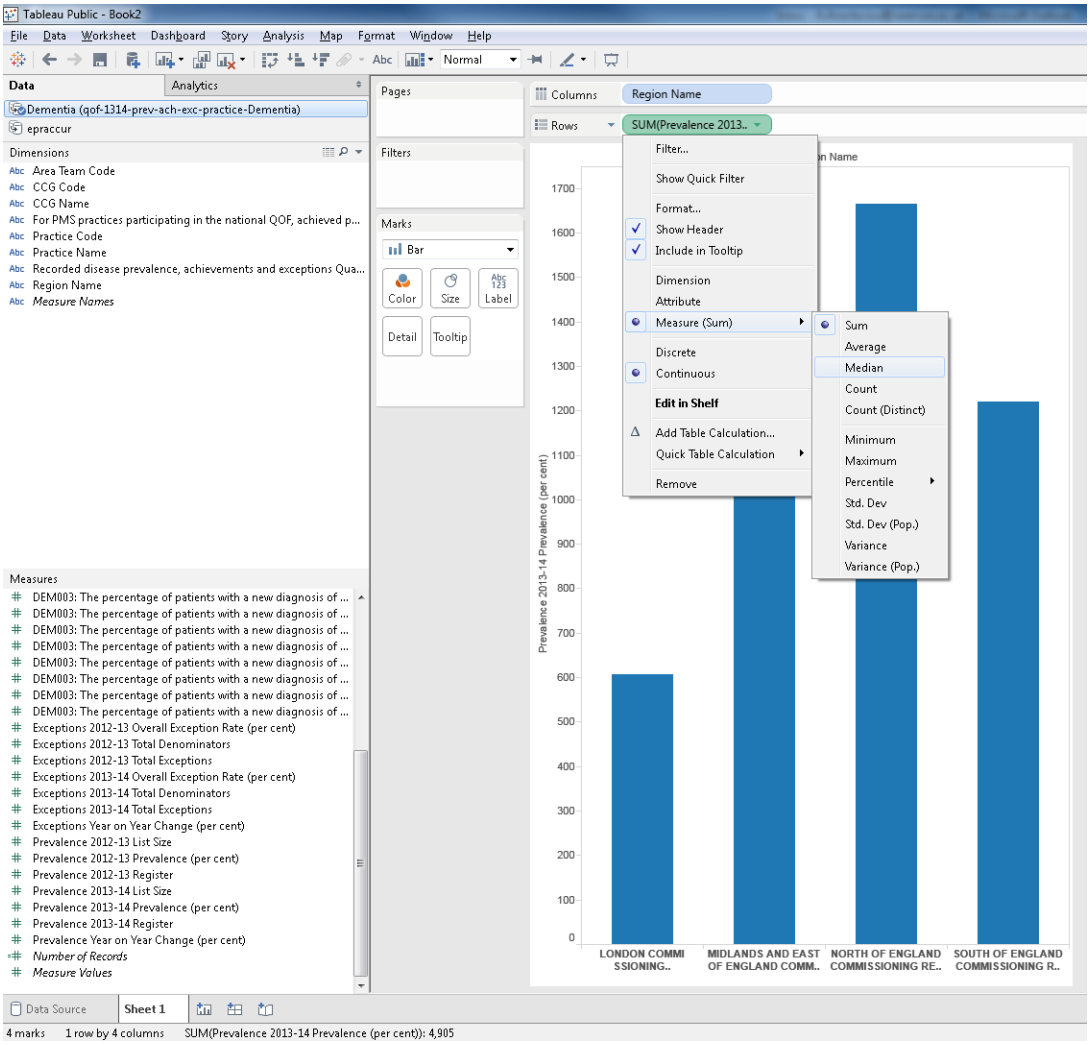


Figure 2.11: Switching the aggregation function to 'Median' for a better depiction of the prevalence across UK geographical regions.

The final and rather simple in this case result of these actions is depicted in figure 2.12.

Please Note: Filters on different variables can be cascaded. In this case, 'AND' logic is implied between the attributes.

2.3.6 Customising a visualisation

Please Note: The dataset is assumed UNFILTERED in this section and therefore the 'London Comissioning Region' that was excluded at a previous step will be appearing in these results.

There are a number of different ways by which the appearance of a visualisation can be customised in Tableau Public. This section will deal with a minimal set of four customisa-

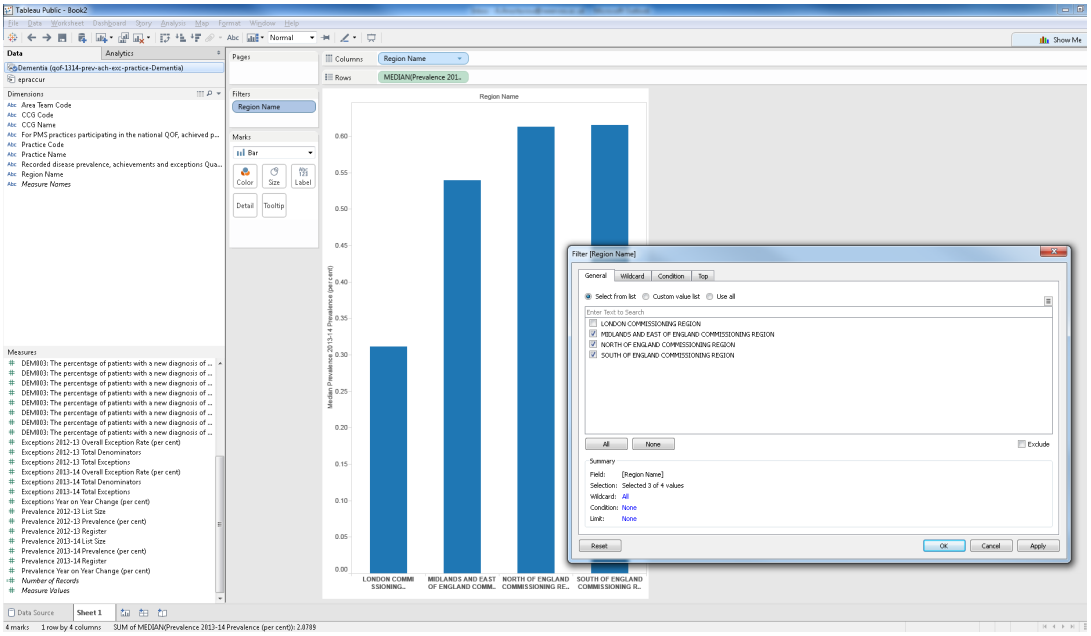


Figure 2.12: The Filter View allows for very flexible filtering of the data.

tions which are broad enough to enable further experimentation with different datasets and visualisation types.

These customisations are:

1. Sort order modifications.
2. Axis modifications.
3. Customising how each data item is represented.
4. Further customising the representation of data items by assigning their Color, element size, label and other attributes to particular data attributes.

Sorting of a Column or Row is modified by clicking the small downwards arrow right next to a Column or Row and selecting ‘Sort’ (please see figure 2.14).

To sort the bar graph produced by section Composing A Very Simple Visualisation, so that regions appear in order of descending median percentage of prevalence:

1. Select ‘Descending’
2. Select ‘Sort By / Field’ and ‘Prevalence 2013-14 Prevalence (per cent)’ and finally the aggregate function (here ‘Median’).

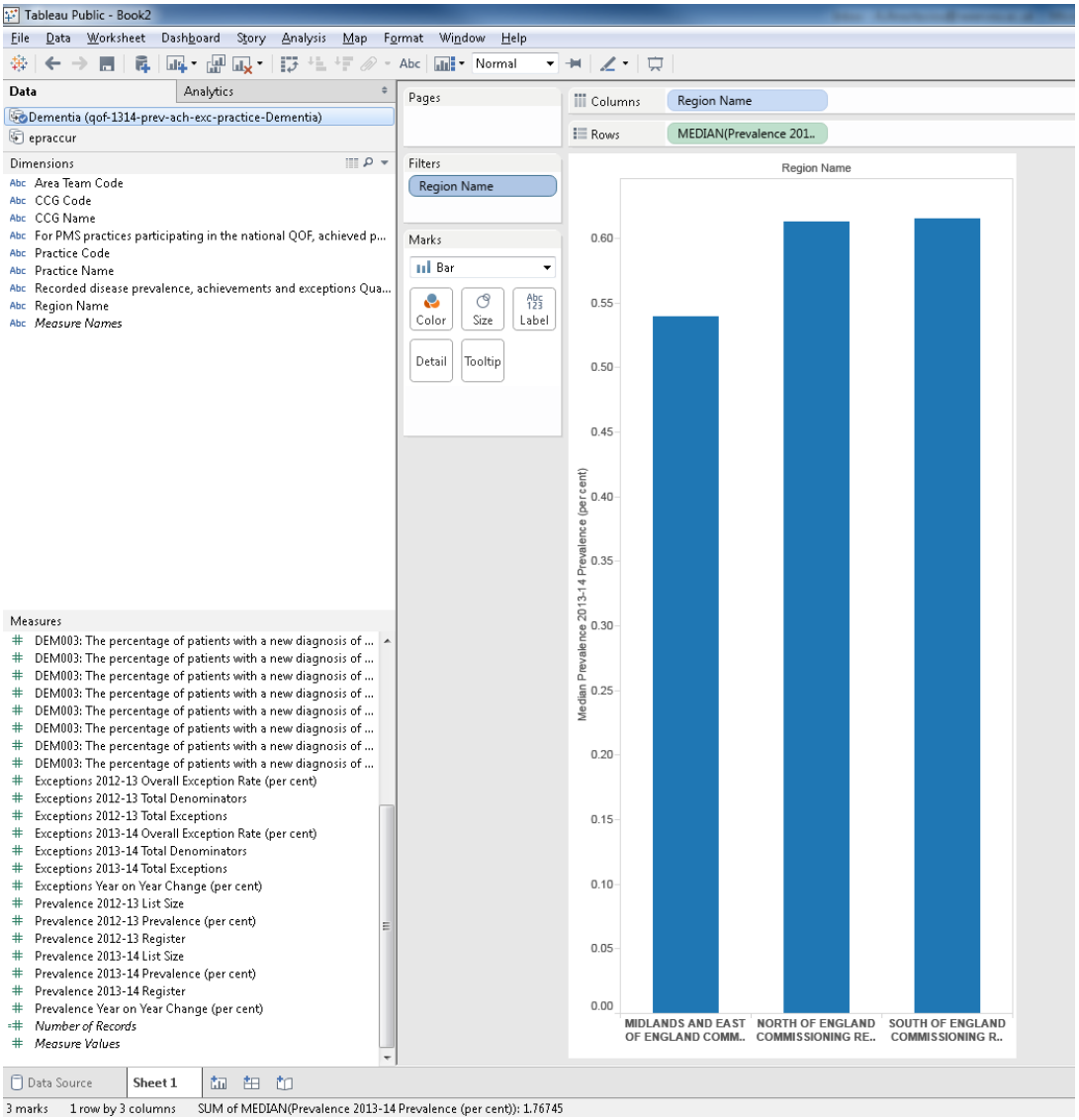


Figure 2.13: The final result of the Filtered dataset. Please note the absence of the 'London Comissioning Region'.

3. Click 'Apply' or 'OK'.

The result of this operation is depicted in figure 2.15.

To modify the way that an Axis represents information, simply right click on the axis and select 'Edit Axis'. The dialog box that appears (please see figure 2.16) is self explanatory regarding the different axis customisation options.

Further customisations on the way that data items are represented are available via the 'Marks' element (please see figure 2.17). Specifically, 'Marks' enables the customisation of how each data item mark appears (for example, 'Bar' or 'Line').

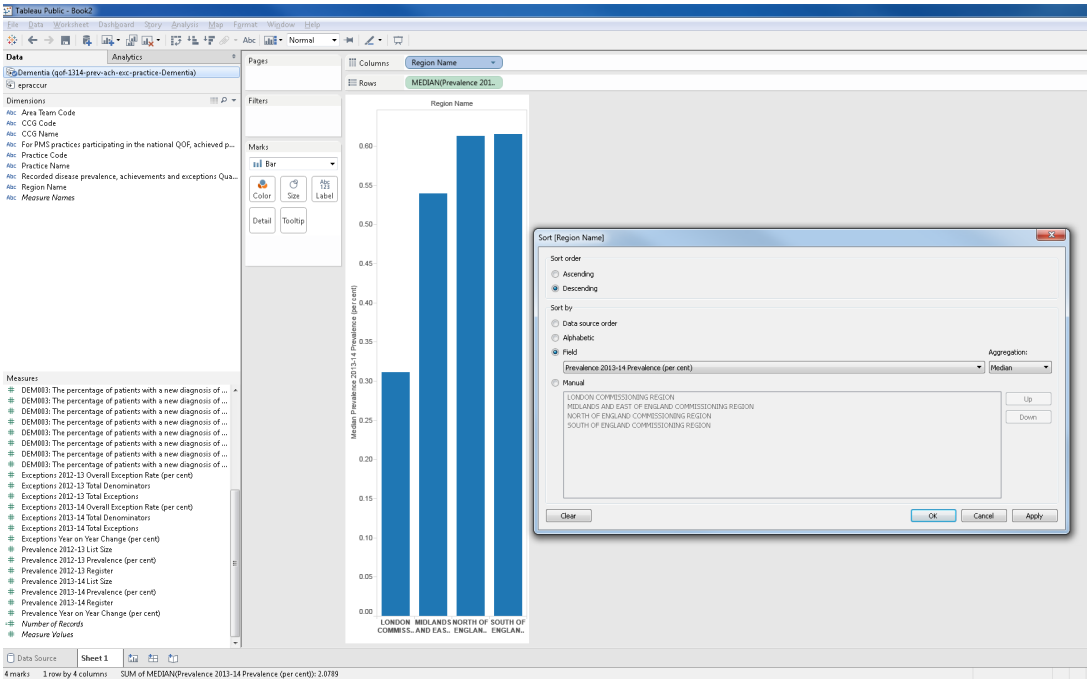


Figure 2.14: An aspect of the sorting dialog box depicting all the different options available to sort the elements of a Dimension.

Other parameters of the visualisation are simply mapped to Dimensions and Measures by dragging and dropping them on each element. For example, to have each UK region rendered under a different colour, drag and drop the 'Region Name' Dimension exactly on top of the 'Colour'. The same drag and drop idea can be applied to assign data attribute values to other elements of the visualisation such as 'Detail' and extra information to appear in the 'Tooltip'. The end result of this operation is depicted in figure 2.18.

2.3.7 Putting together a Dashboard

Having worked on a simple visualisation of Dementia Prevalence, it is now time to work on its layout and finally embed it into a Story.

Aesthetics can vary across authors. For the purposes of this example, a Dashboard with the overall structure depicted in figure 2.19 was created. To achieve a similar result:

1. Click on the plus sign next to the dashboard icon at the bottom of Tableau Public's window.

2. Right click on the dashboard that was created and give it a more descriptive name 'Dementia Across Regions'.

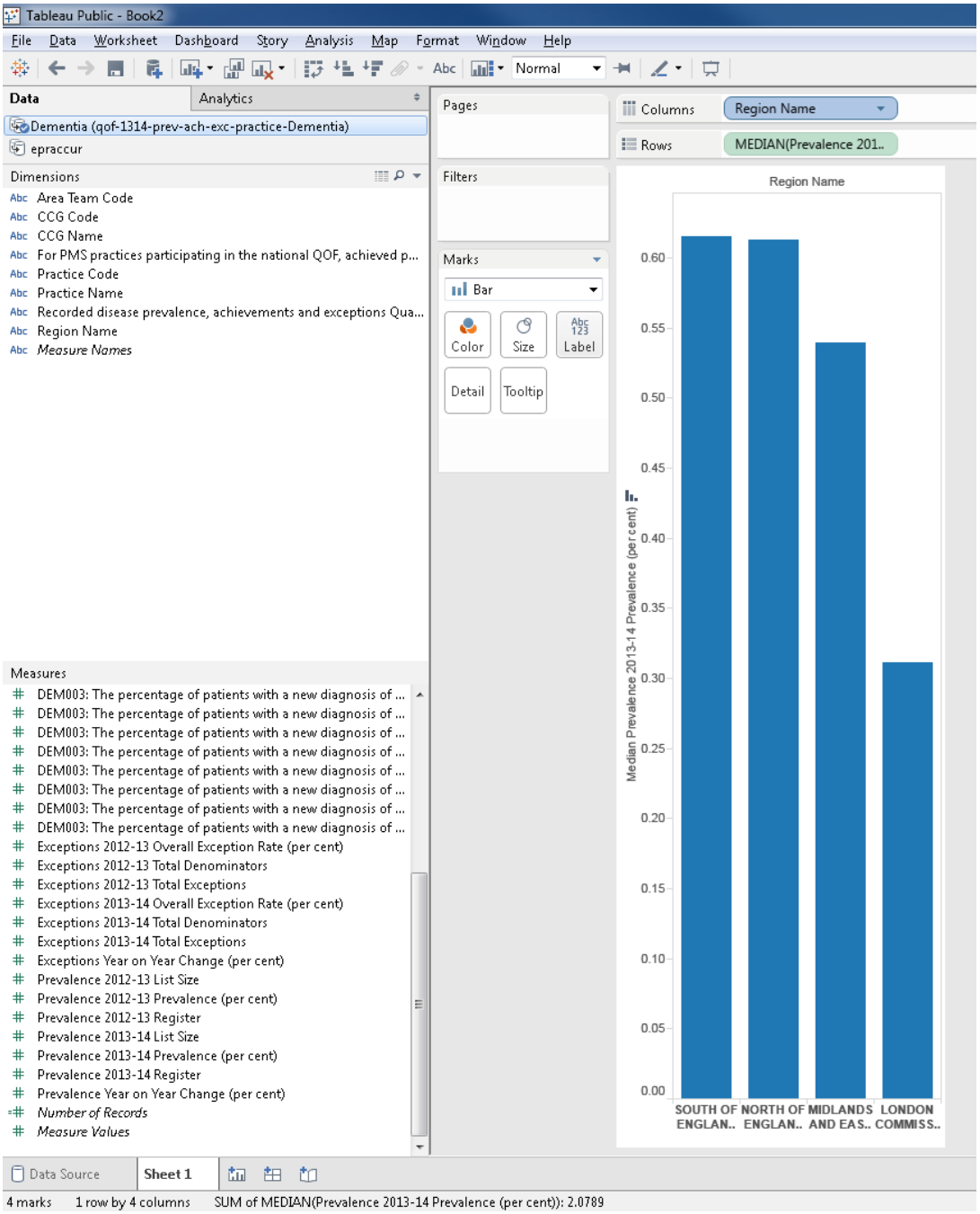


Figure 2.15: The final result of the sorting operation with the regions sorted in decreasing order of Dementia prevalence.

3. Drag and drop the ‘Vertical’ layout component in the form.

4. Drag and drop the ‘Dementia Prevalence’ sheet at the top compartment.

5. Drag and drop another ‘Vertical’ layout at the bottom part of the dashboard.

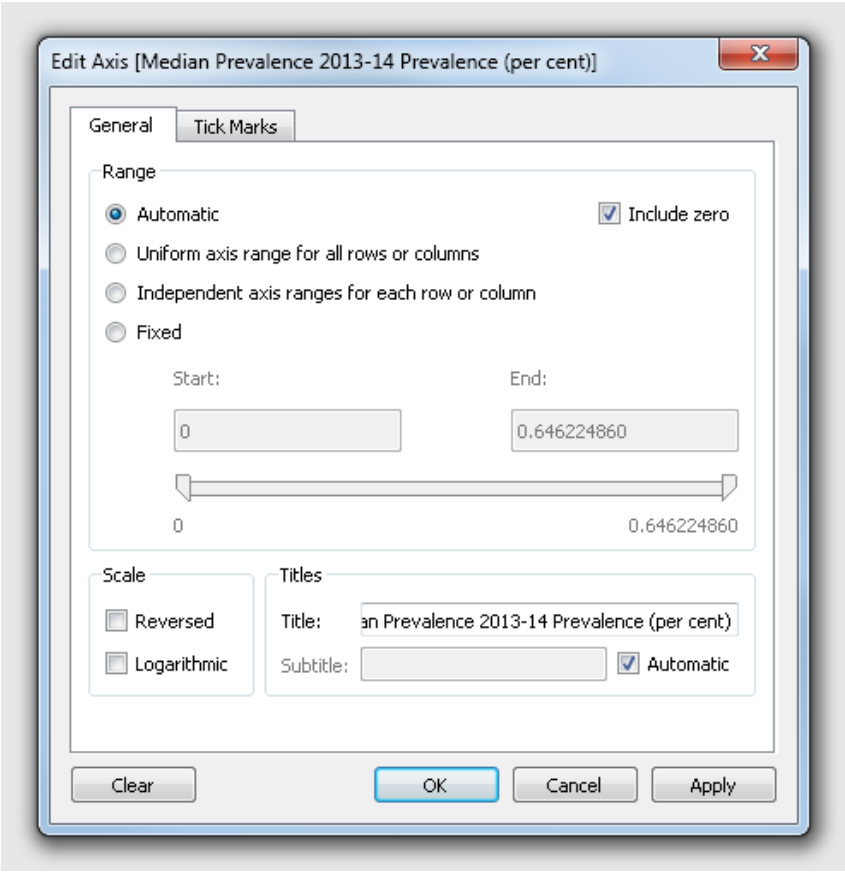


Figure 2.16: Modifying the appearance of an axis bearing a Measure.

6. Drag and drop a 'Text' element on the top part of the newly added 'Vertical' layout container. This serves as a caption in this case. The text that was entered for the purposes of this example was 'Median prevalence of Dementia across different comissioning regions in the UK'.

7. Go to Google images and download a copyright-free image with some resemblance to the data being visualised [3].

2.3.8 Putting together a Story

Having created a simple visualisation and a layout for it, it is now time to create the final product which corresponds to the 'Story' level.

[3] For the purposes of this example, the image was tracked down via 'Google Images' and linked from `http://psychone.net/blogs/wp-content/uploads/2012/02/elderly5.jpg`

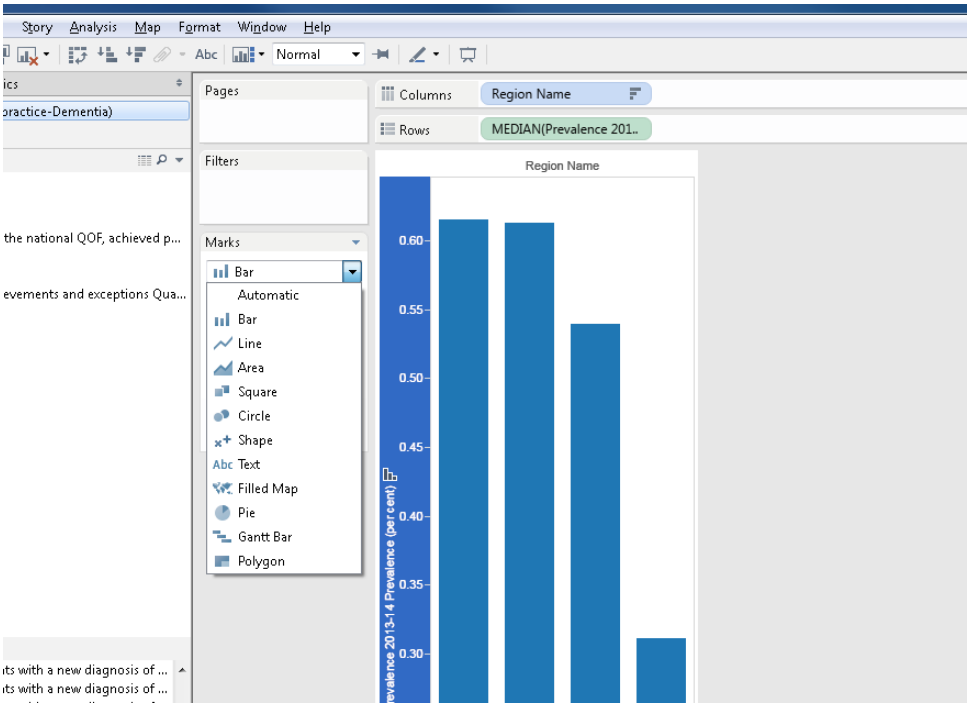


Figure 2.17: A close-up of the Sheet View depicting the Marks panel that can be used to customise the appearance of a visualisation.

Similarly to the case of the Dashboard, the aesthetics of the actual layout of a Story vary across authors. For the purposes of this example, a Story with the overall structure depicted in figure 2.20 was created. To achieve a similar result:

1. Click on the plus sign next to the Story icon at the bottom of Tableau Public's window.
2. Right click on the Story that was created and give it a more descriptive name ('Living With Dementia')
3. Create one blank point for each message conveyed by the Story. For this example, one point was created with the following text

   Dementia is a common condition. Your risk of developing dementia increases as you get older, and the condition usually occurs in people over the age of 65. Coincidentally, this is too close to retirement age and lifelong plans for retiring to a different location than one's birthplace. Consequently, this increases the pressure on NHS systems for specific services in certain regions.
4. For each of the blank points, drag and drop the corresponding Dashboard or Sheet to

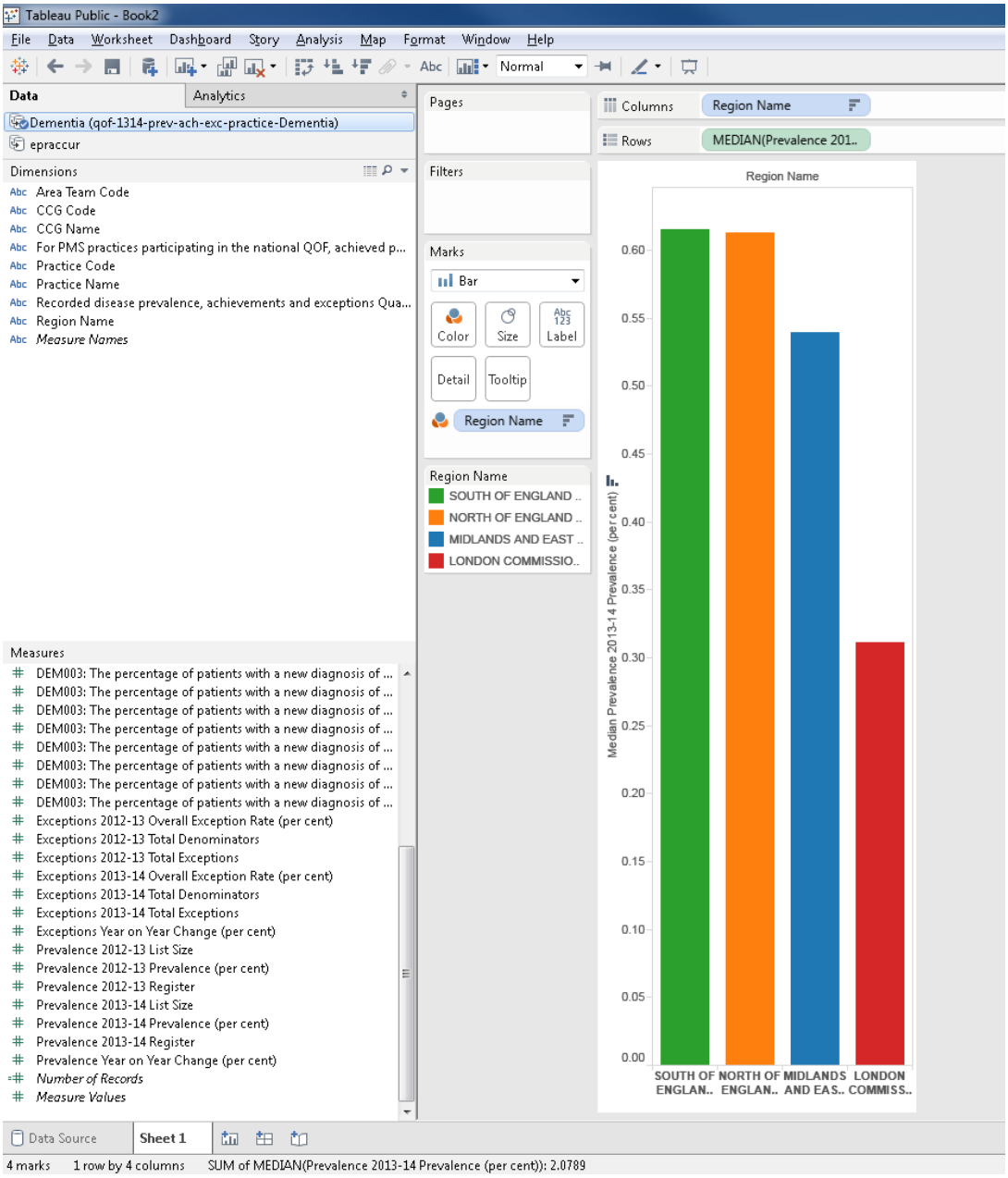


Figure 2.18: The final result of the Filtered dataset. Please note the absence of the 'London Comissioning Region'.

the main body. In this example, the story is composed of just one dashboard 'Dementia Across Regions'.

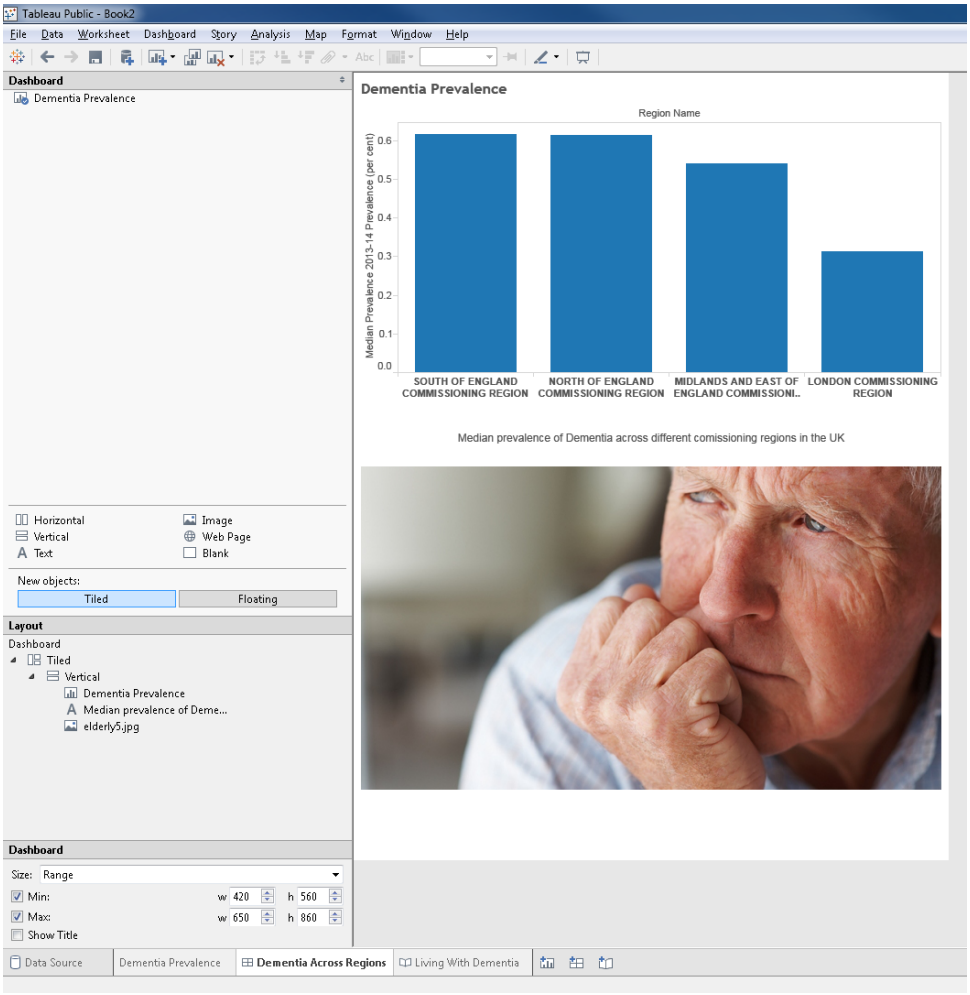


Figure 2.19: Assigning a different colour to each region by binding the Colour parameter to the 'Region Name' Dimension.

2.3.9 Exporting and accessing the visualisation online.

Finally, the last step to creating a story is to publish it online and let it be viewed by others. Tableau public offers the infrastructure to do this right from the desktop application as follows:

1. From the application menu, click on 'File / Save to Tableau Public as'

2. Provide a descriptive name

Tableau will export all the data and metadata of the Story to Tableau's servers and provide the Universal Resource Location (URL) to a web page that contains a 'Live' view of the story. The final result for this example is accessible at `https://public.tableau.com/views/`

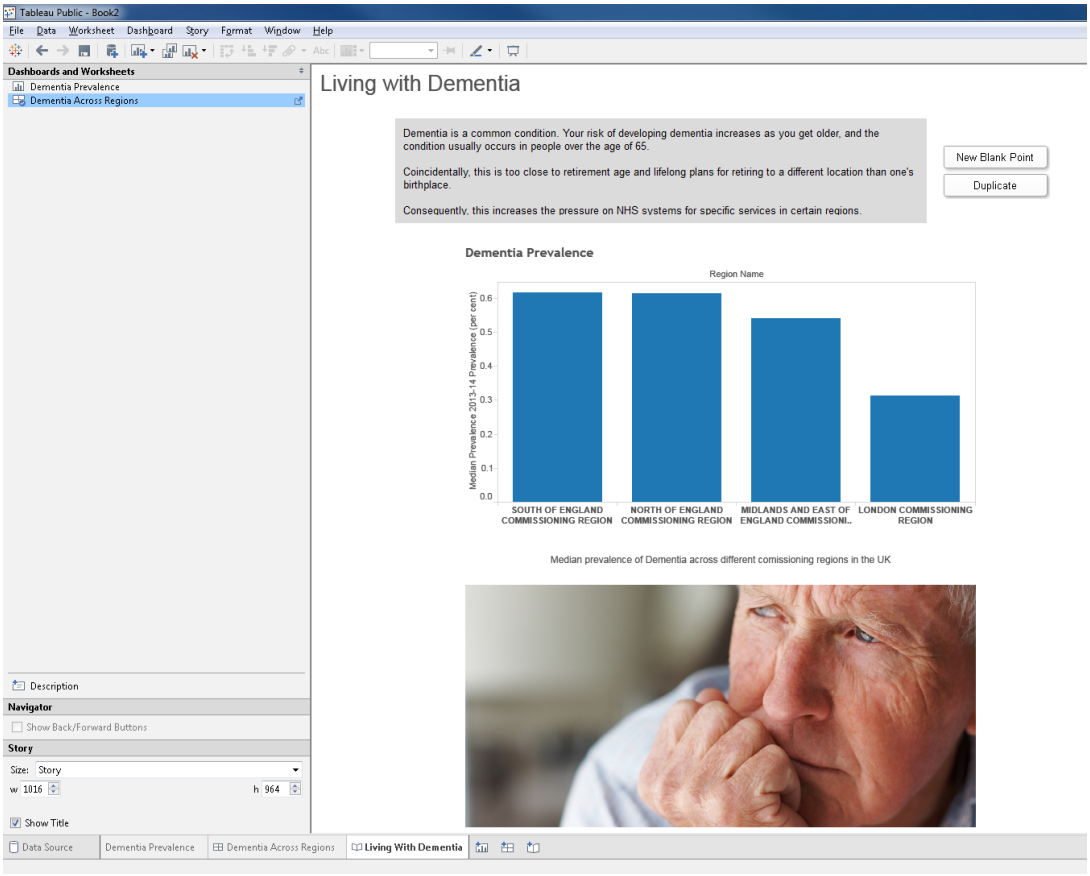


Figure 2.20: A potential dashboard with a number of different graphical elements to support a story about Dementia prevalence across the UK.

`LivingWithDementia/LivingWithDementia?:embed=y&:showTabs=y&:display_count=yes`

2.4 Conclusion & Outlook

This chapter introduced the functionality of Tableau Public by briefly explaining the basic concepts and operations that govern its operation. The interested reader is now invited to go through chapter Working with a realistic dataset: UK NHS Quality Outcomes Framework which provides pointers to real complex datasets to find and visualising emerging stories.

Chapter 3

Working with a realistic dataset: UK NHS Quality Outcomes Framework

3.1 Introduction

The objective of this chapter is to provide a brief overview of the United Kingdom's (UK) National Health Service (NHS) Quality outcomes Framework (QoF) which is *an annual reward and incentive programme detailing General Practice (GP) achievement results*.

A large amount of detailed information and data can be obtained directly from the relevant QoF pages of the Health and Social Care Information Centre (HSCIC). However, this guide expands only on a few key points that were deemed 'necessary knowledge' for the purposes of the Health Data Visualisation sessions.

The purpose of the NHS's QoF programme is essentially to provide a feedback mechanism to monitor population health as well as GP Practice performance in terms of service delivery. These two broad aspects are quantified by a total of 121 performance indicators. The data which are used to derive these performance indicators are provided by the General Practice Extraction Service (GPES). The service extracts data directly from GP systems on a regular basis and through a number of data processing steps calculates indicators such as the prevalence of common chronic diseases (e.g. Dementia and Diabetes), as well as aspects of the NHS's service delivery such as child health and maternity service usage and minimum GP service timings.

A key point in the actual data collection process in the field is that GP practices receive a financial reward both for capturing the data as well as performing well, in providing quality healthcare. As an indication of the magnitude of the reward, within the period 2013-2014, general practices received £156.92, on average, for each point they achieved in the key

performance indicators.

While the data used to derive the performance indicators are coming directly from the patient data stored in GP systems across the country, the QoF data that are made available to the public through the HSCIC are aggregate data up to the level of detail of a single GP Practice. In addition, certain exceptions to reporting might apply to very special medical cases that are excluded from the QoF data and reports. These cases amount to 1% of the total reported population.

The rest of this chapter focuses on the structure of the GP Practice level data for the time period 2013-2014 that are used throughout this guide.

3.2 SPECIFIC QOF DATASETS

This guide focuses on GP level aggregate data for the time period 2013-2014 covering indicators such as prevalence, achievements and exceptions for the following groups:

- Cardiovascular
- Respiratory
- Lifestyle
- High dependency and other long term conditions
- Mental health and neurology
- Musculoskeletal
- Fertility, obstetrics and gynaecology.

Each one of these groups is represented via a compressed archive which contains one or more MS Excel spreadsheets describing the actual data with the following broad categories of attributes:

- Region identification (code,name)
- Area identification (code, name)
- Clinical Commissioning Group identification (code, name)
- General Practice identification (code, name)

- Prevalence data
- Achievements data
- Exceptions data
- One or more disease specific performance indicators (for example, *HYP002: The percentage of patients with hypertension in whom the last blood pressure reading (measured in the preceding 9 months) is 150/90 mmHg or less*)

3.3 OTHER DATASETS ENABLING RICH VISUALISATION

In addition to the above datasets that are generally made available through the HSCIC, there are a number of other datasets that can be useful in visualising health data and are provided by a number of other UK government agencies. A brief list of those is available below:

Ordnance Survey Ordnance Survey are *Britain's mapping agency* and produce various spatial products with open access for the public. Among these are administrative and electoral boundary data for different regions, post-codes cross referenced with their physical locations and others.

UK Office of National Statistics (ONS) The ONS maintains a large and diverse collection of data. Among them:

- Various different boundaries used for the purposes of the national census
- A very large collection of health related data.

Eurostat Eurostat is the European equivalent of ONS and aggregates and analyses data from across Europe on a number of different parameters, including some very interesting statistics on health.

3.4 PUBLICATIONS

The quality outcomes framework has already been used in a number of publications. Some of these are listed below:

- McLean, G., M. Sutton, and B. Guthrie. "Deprivation and quality of primary care services: evidence for persistence of the inverse care law from the UK Quality and Outcomes Framework." Journal of Epidemiology and Community Health 60.11 (2006): 917-922.
- Sutton, Matt, et al. "Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers." Health economics 19.1 (2010): 1-13
- Kontopantelis E, Springate DA, Ashworth M, Webb R, Buchan I and Doran T. (2015). Investigating the relationship between quality of primary care and premature mortality in England: a spatial whole-population study