Lab Sheet 3

Introduction to Data Analytics

Ian Nabney

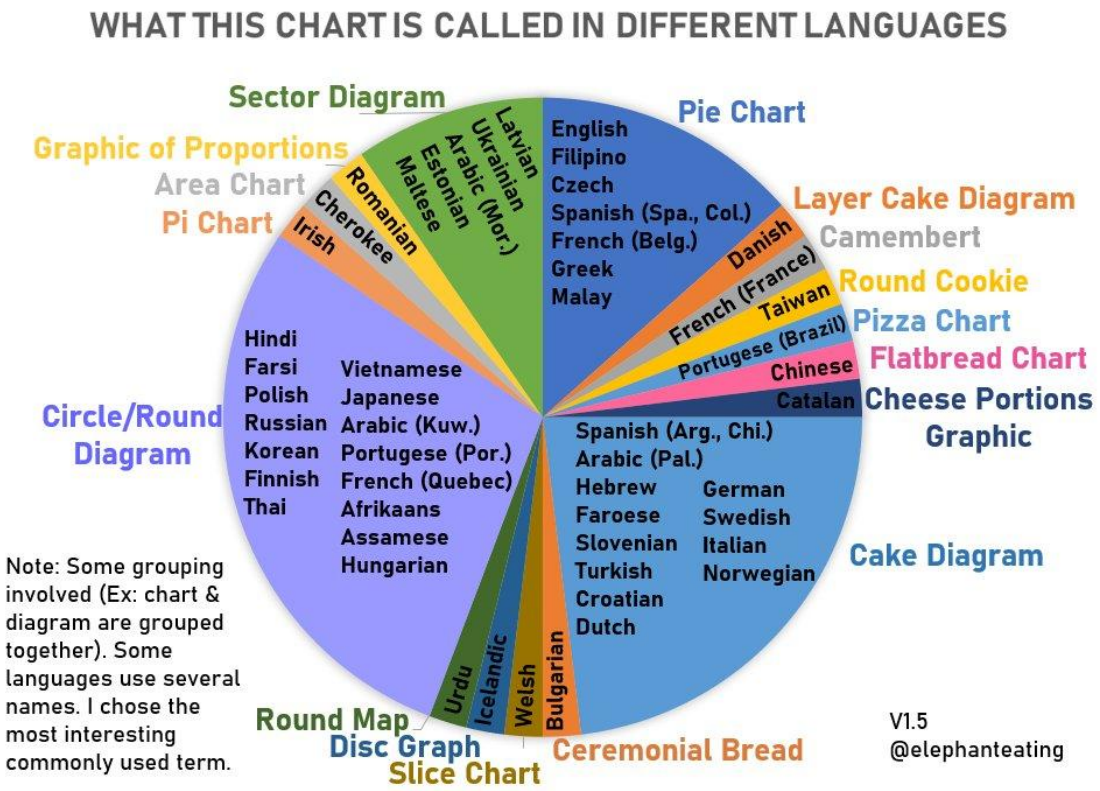
1 Tableau charts

(Partly based on a tutorial presented at the Alan Turing Institute).

We will now continue our survey of Tableau chart types from where we left off in Lab 2. Start by opening the "World Indicators" dataset and create a new Worksheet.

1.1 Pie charts

No, I am not going to tell you how to create them. But I must pass on this graphic (thanks to Eric Hittinger for creating it – see https://twitter.com/ElephantEating/status/1360647164012027907).



I think that the prize for the best name goes to the French, for naming it after a famous cheese.

Now to more serious matters.

1.2 Scatterplots

How do internet and mobile phone usage relate? One way of investigating that is to create a scatter plot, but by default Tableau only shows one data point – the total of all the values in the dataset. Drag *Internet Usage* to *Columns* and *Mobile Phone Usage* to *Rows.* The aggregation measure defaults to 'Sum' (at least, in my installation of Tableau: change this by selecting the pulldown menu at the right-hand end of the pills and then choosing Average from the Measure sub-menu (see Figure 1(a)). You know that the dataset actually contains a value for each combination of country and year, so to create a scatter plot that shows the average usage for each country drag *Country/Region* into the *Detail* box of the *Marks* card (see Figure 1).

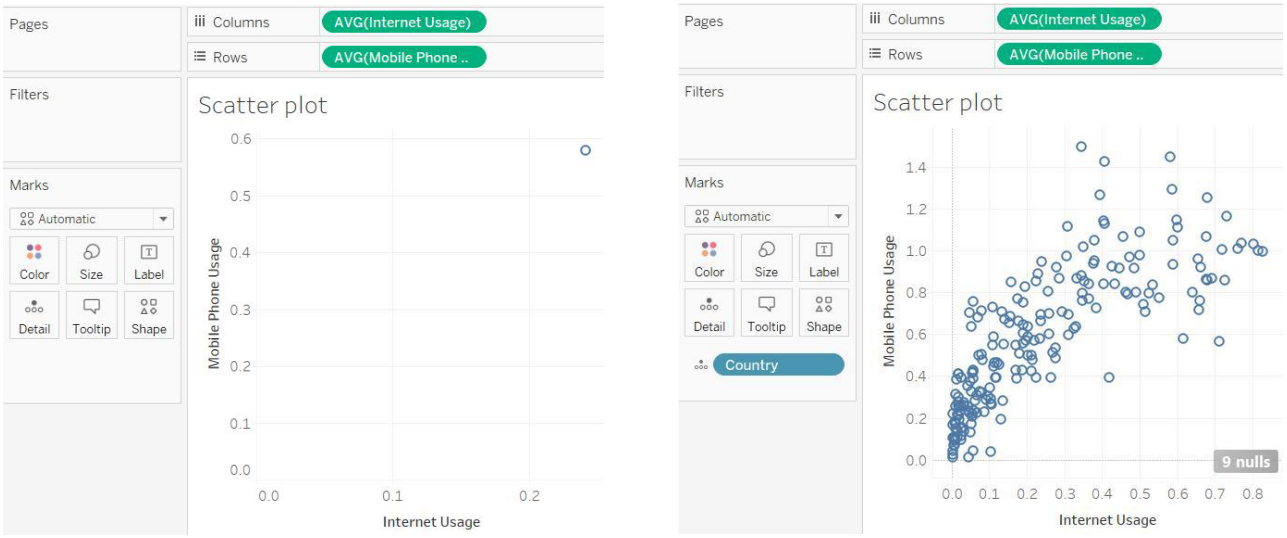


Figure 1: A scatter plot showing: (a) the default level of detail, and (b) one data point per country.

Note that there are 9 nulls: these are missing entries in Internet usage.

In general, variables on the *Columns* shelf are mapped to the x-axis in a graph while variables on the *Rows* shelf are mapped to the y-axis.

1.3 Box-and-whisker plot

The scatter plot indicates that the internet and mobile phone usage data are both skewed. A box plot lets us see the distribution of the data for the countries in each region (see Figure 2). Remove Mobile Phone Usage from the rows, and then choose the box-and-whisker plot from the *ShowMe* palette on the right-hand side and drag *Region* to *Columns*.

Figure 2: A box plot showing the average Internet Usage for the countries in each region.

1.4 Histogram

Another way of looking at data distributions is to create a histogram. To do that, remove *Region* from the *Columns Shelf* and then drag *Internet Usage* onto the *Columns Shelf* and *Country/Region* from *Detail* (note how this displays the points scattered along a line) and select *Histogram* from *Show Me* (see Figure 3).

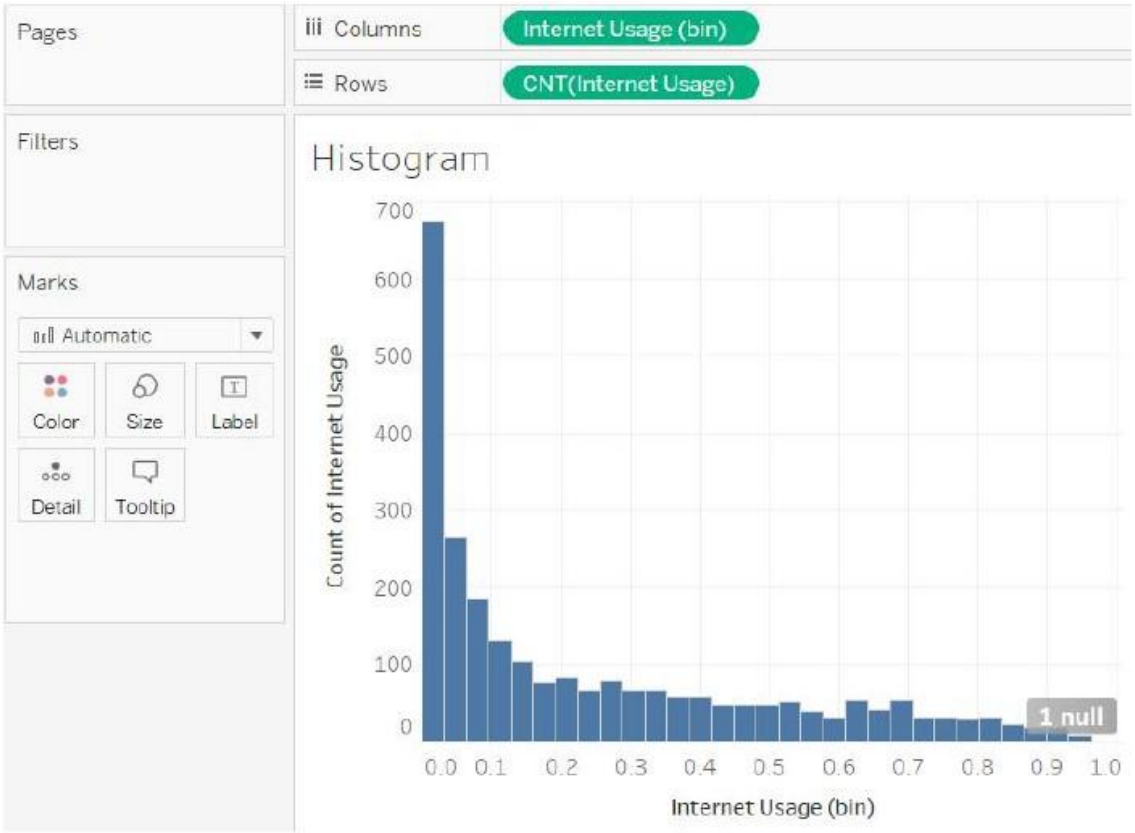


Figure 3: A histogram of internet usage.

This histogram shows the data for every combination of country and year. There are many ways that you can subdivide the data, and one is to show a separate histogram for each *Region* (drag that dimension onto the *Rows Shelf*) and animate the *Year* (drag that dimension onto the *Pages Shelf* in the top-left of the Tableau display). See Figure 4. You can then scroll through the years in turn using the control.

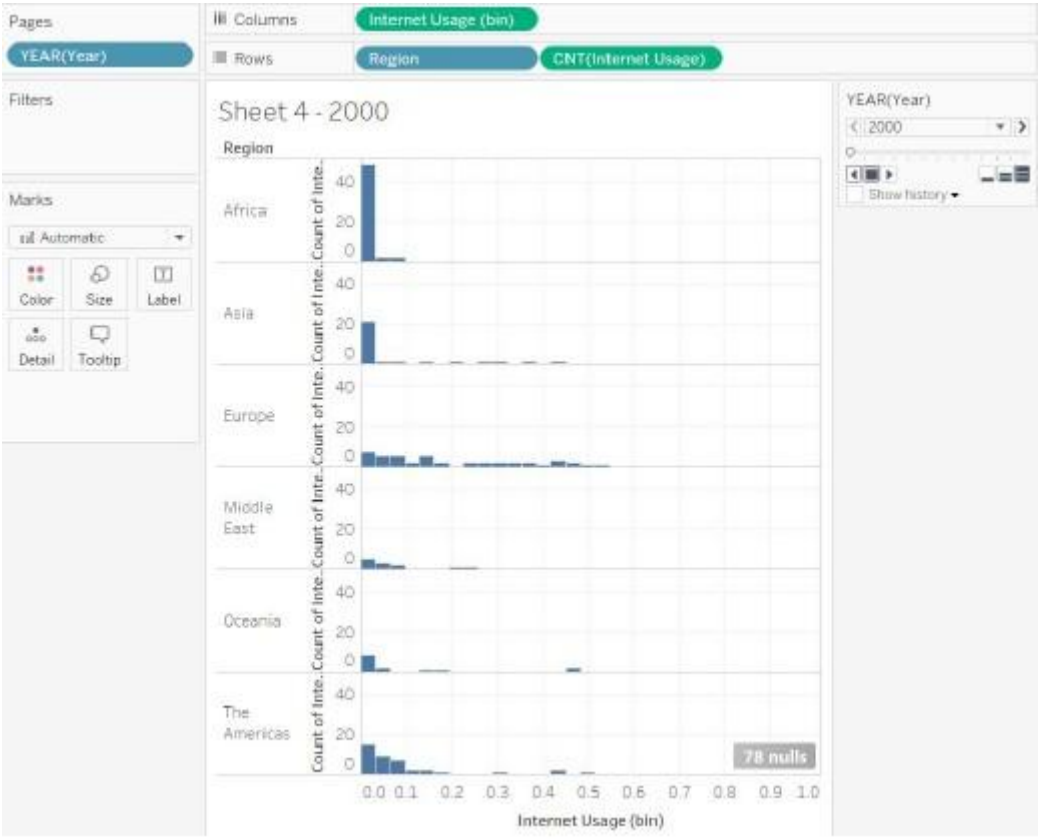


Figure 4: Using the Pages Shelf to display separate histograms for each year.

1.5 Heatmaps

Heat maps are helpful to gain an overview of how a measure (e.g., Internet Usage) varies with two dimensions (e.g. Country/Region and Year). By choosing an appropriate colour map you can emphasise certain differences (e.g. low vs. high usage; see Figure 5). The figure uses the orange-blue diverging colour map. Look at the graphic to see how it was built up and copy these steps.

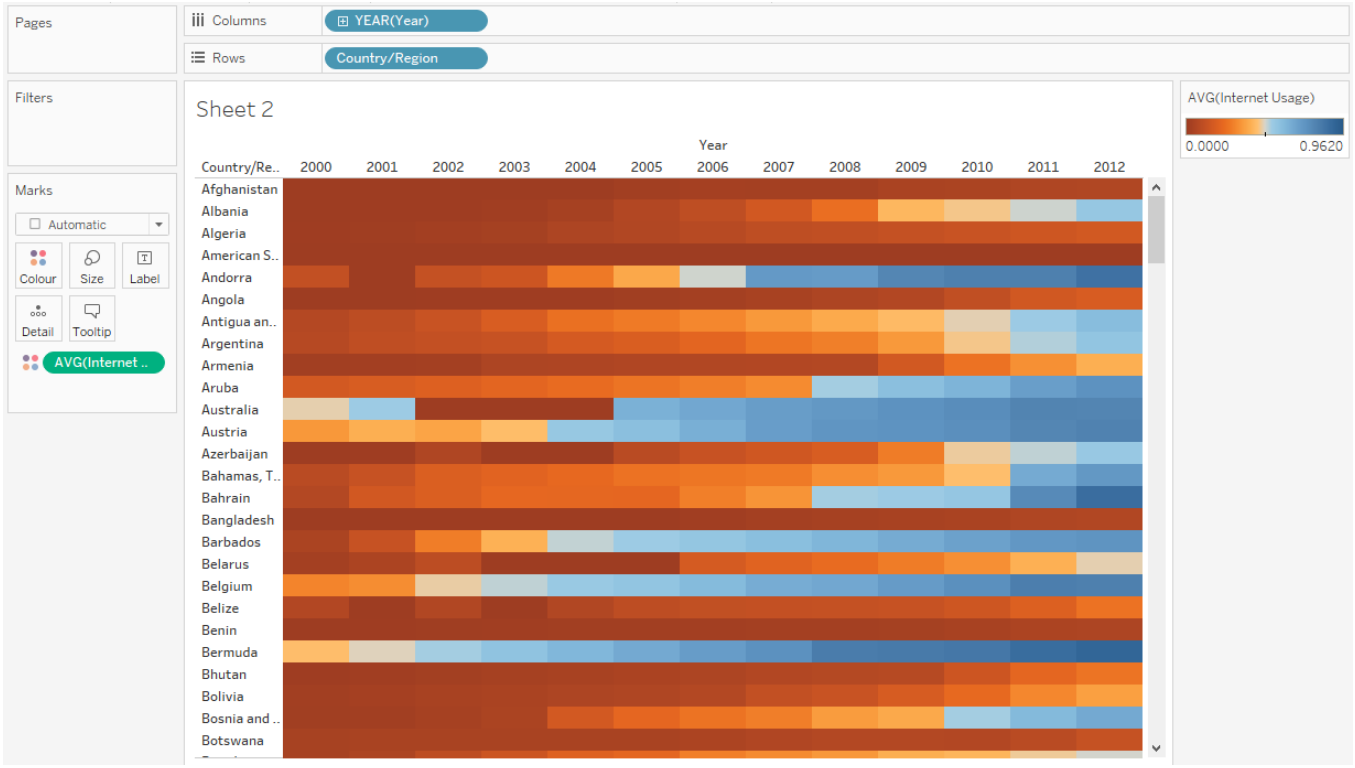


Figure 5: A heatmap showing Internet Usage for European countries from 2000 – 2012.

What general trends can you see in the data as the date changes? What anomalies can you notice in the data? (Hint: look at the Isle of Man, Kosovo, and San Marino.) What possible explanations are there for these outliers?

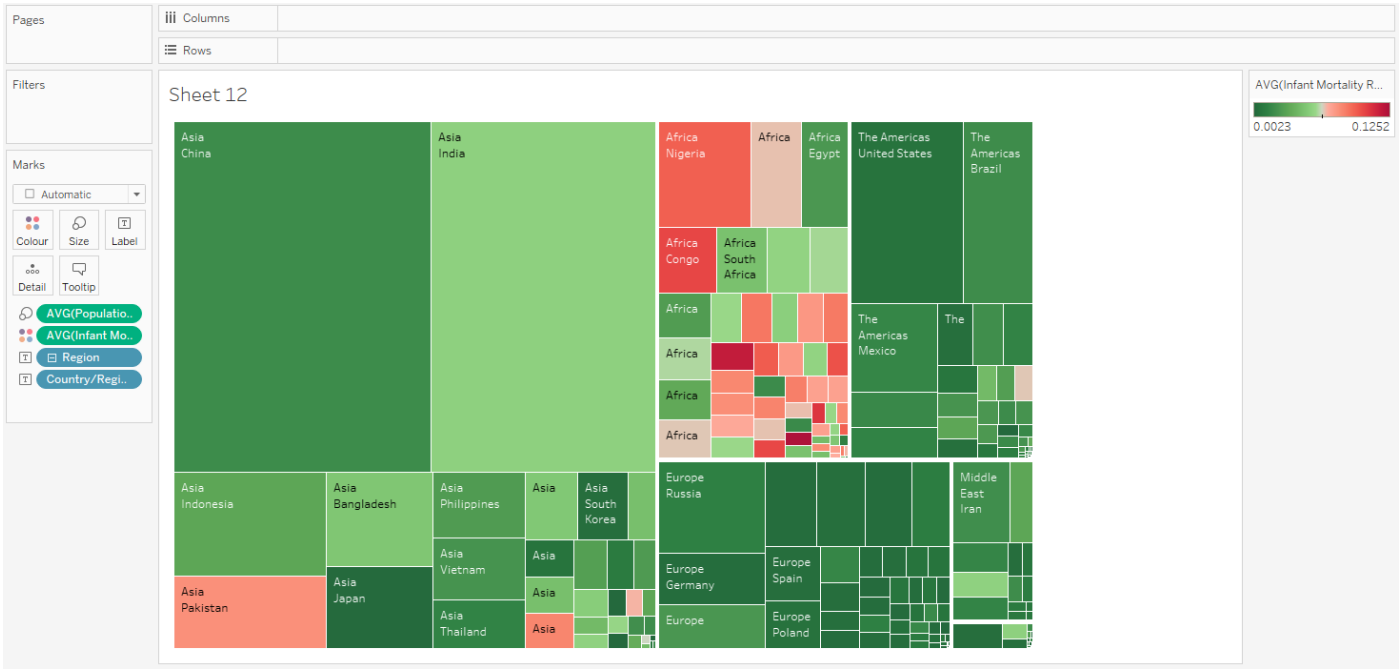
1.6 Tree Maps

Treemaps are useful for hierarchical data, for example the population of countries in regions. Drag *Population Total* onto both *Size* and *Color* Marks. Add *Region* and *Country/Region* to the *Text* marks and then select Tree Map from ShowMe (see Figure 6(a)). By default Tableau uses both area and colour to encode one measure, but if you drag another measure onto *Color* in the *Marks Card* then you can see contrasts between the measures (see Figure 6 (b)). In Figure 6 (b), total population is on size and infant mortality is on colour.

Note how the automatically provided tooltips enable you to see the details on demand, which is particularly useful since few of the boxes are large enough to read the text directly.

There are several other graph types in Tableau, but I shall leave you to explore them in your own time if you wish because there are other important features of Tableau for you to understand that will help you in dealing with practical problems (such as the group mini-project).

(a)



(b)

Figure 6: A treemap showing: (a) the Population of each region and country, and (b) the Population and Infant Mortality Rate (with a diverging, reversed colour map high-mortality countries are emphasized in shades of red for 'danger').

2 Tableau data handling

One great strength of Tableau is the way that it supports seamless interconnection with a wide range of different data sources. On this page https://help.tableau.com/current/pro/desktop/en-us/exampleconnections_overview.htm there is an overview of all the different connectors to data stores that are available: there are 93 listed on the page (it was 90 in 2022 and 84 in 2021), so you should usually be able to read in your data more or less no matter what its format.

2.1 Tableau data model

Tableau's ability to handle complex datasets is built on a two-layer data model (of the type that will be familiar if you have studied databases). This becomes relevant when a data source contains more than one data table. You can think of a data model as a diagram that tells Tableau how it should query data in the connected database tables.

To understand these layers better, first open the Bookshop Excel file in Excel. You can see that there are 13 tabs containing data: each one of these is a table with each row representing a different data entry and the columns represent variables. The tables mainly have distinct fields, but some are in common – we will use these common fields to bring together information from multiple tables to analyse.

Open the Bookshop as an Excel source (after closing the workbook you have built up in Section 1 of this worksheet) and bring up the data source canvas. Now drag the Book table to the right-hand canvas, you should see something like Figure 7.

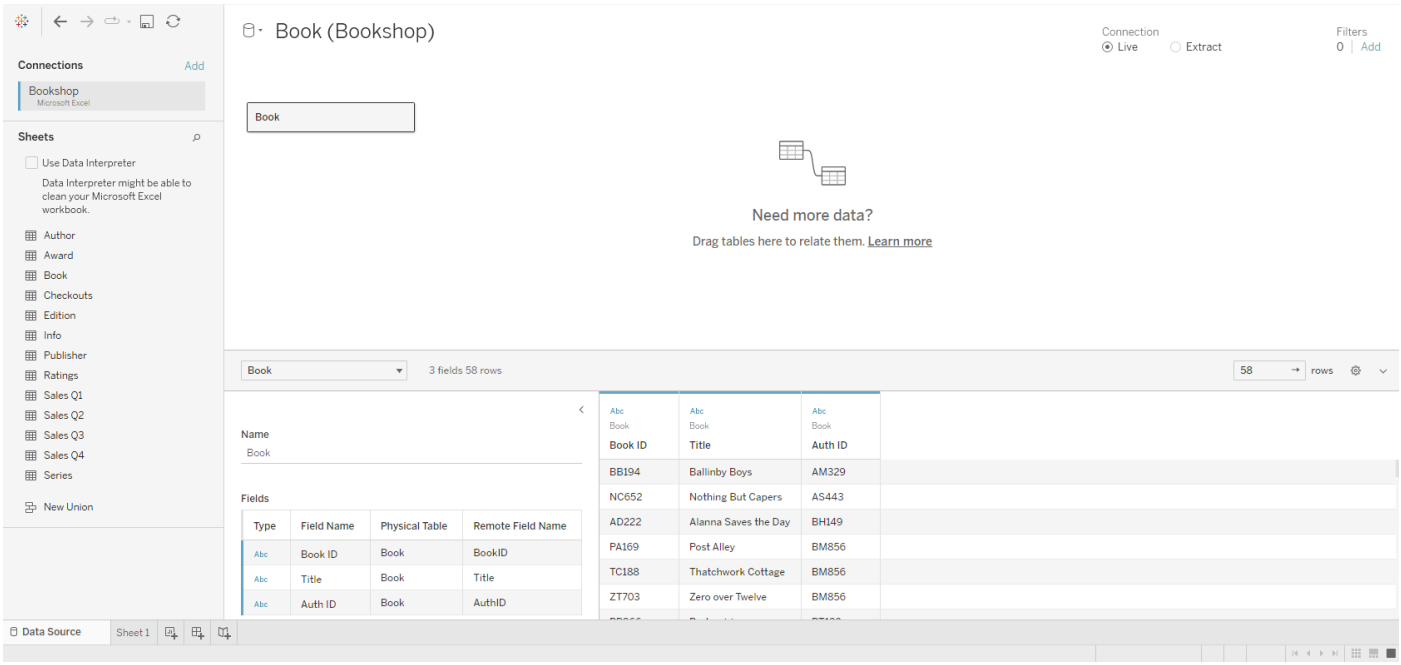


Figure 7: What you see when you open up the Data canvas. Left pane shows the connected data source and other details about your data (e.g. the tables it contains). Canvas (top centre) shows the logical layer. Data grid (lower right) displays up to the first 1000 rows of the data. Metadata grid (to the left of the data grid) shows the fields in your data source.

You can drag one or more tables to the canvas which means that it has two layers[1]:

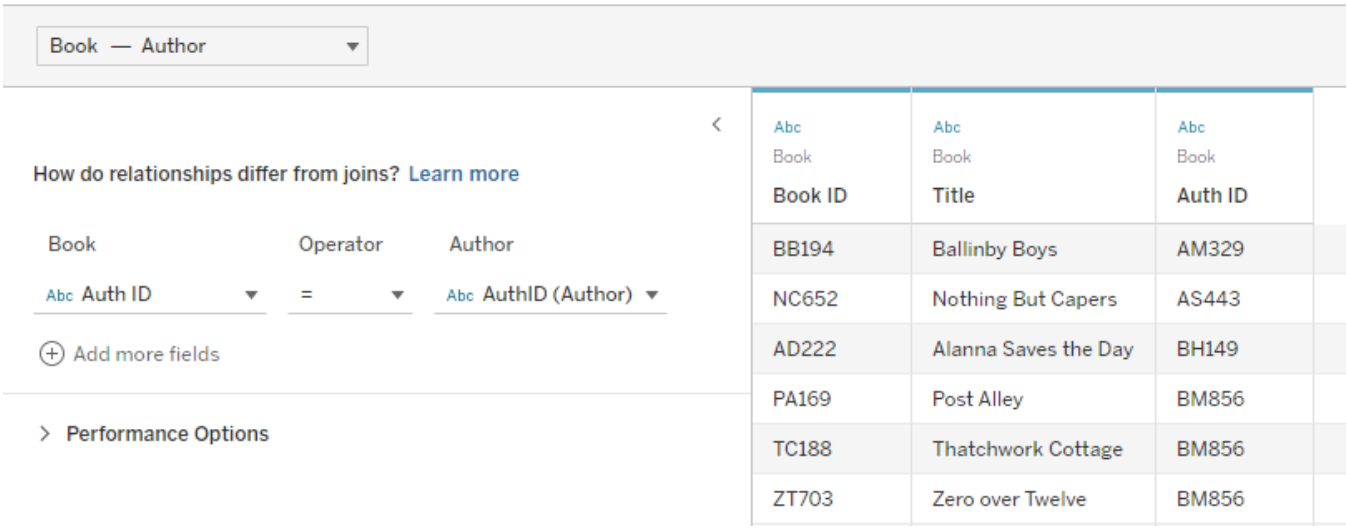
The ***logical layer*** of the data source. You combine data in the logical layer using relationships. Think of this layer as the Relationships canvas in the Data Source page.

The ***physical layer***. You combine data between tables at the physical layer using joins and unions. Each logical table contains at least one physical table in this layer. Think of the physical layer as the Join/Union canvas in the Data Source page.

[1] This two-layer model was introduced with Tableau 2020.2, so older tutorials will not include this idea. Also, older datasets may work slightly differently if stored in Tableau format.

Below you should see the data (just 58 rows as that is all there is in this table). There are three fields: BookID, Title, and AuthID. The two ID variables are just codes (and in fact provide links to other tables), while the title is the useful information in this table. Now move the Author table to the canvas. As you do so, you should see an orange joining line (called a 'noodle' by Tableau). Once the table is in position, you can click on the relationship and the lower-left pane will now show information about the relationship (Figure 8) and where you can also edit the relationship. What has happened is that Tableau has noticed that one of the fields in the new table is also called AuthID and has made the guess that these two fields should be used to relate the two tables.

(a)



(b)

Figure 8: (a) Relationship in Canvas. (b) Relationship information.

You can leave this as it is (or override it if it is not correct) – performance options are only to be dealt with by experts, and certainly aren't needed with this small dataset.

Click on the Author box in the Canvas. The data grid now shows the Author table: it has six fields with First Name, Last Name, Birthday, Country of Residence, and Hours Writing per Day in addition to AuthID. If you click on the Book block in the Canvas, the data grid returns to the Book table.

Unlike joins, relationships do not merge tables together: we simply set up the relationship and use the fields in the view. Relationships have additional features, such as handling null values smoothly. We will leave joins for now – there is plenty of help material on the Tableau website if you need to know more.

Continue to add tables to the canvas to build up the relationships as shown in Figure 9 (leave out Sales for a moment and omit Series, which doesn't have a suitable field to link to). The sales information comes as four separate tables, and there is value in joining them together in a single table. This can be done by creating a *Union* for them. Double-click on the *New Union* icon on the left pane. Now drag each of the four Sales tables (Sales Q1 etc.) onto the dialog box. When you have finished, click Apply. Once the union has

been placed (in relationship with the Edition table), you can right-click on it and rename it from *Union* to *Sales*.

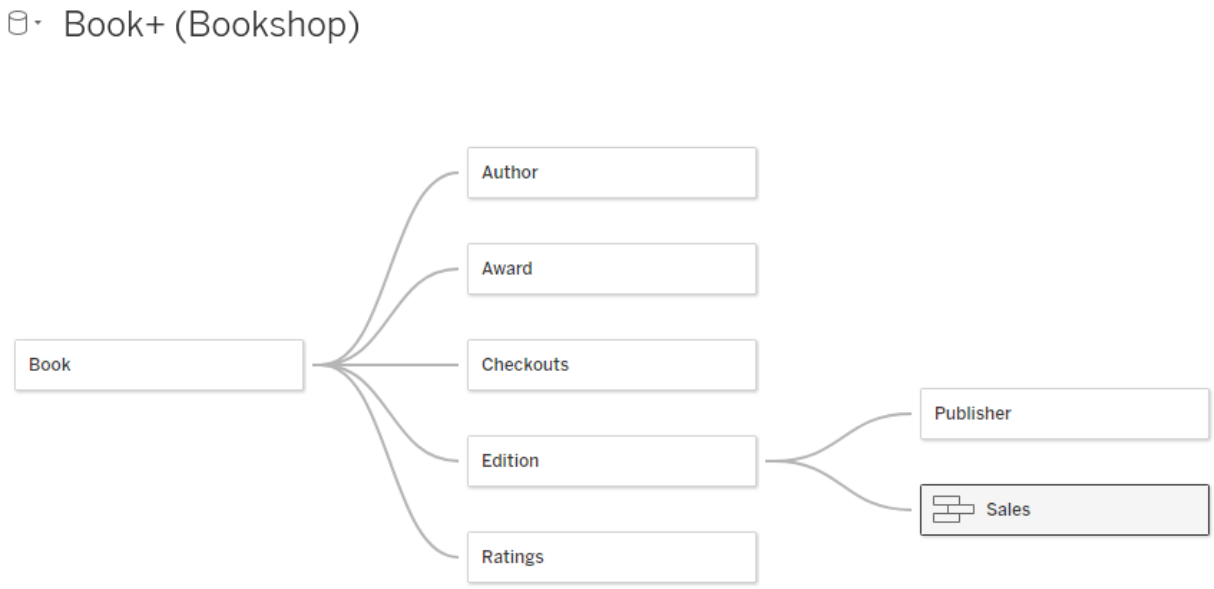


Figure 9: Bookshop table relationships.

2.2 Tableau Relationships

Open a new worksheet. Move the Book numeric field from the Book table to the Columns shelf and the Last Name field from the Author table to the Rows shelf. You should see a horizontal bar chart where every author is listed (in alphabetical order) with the count of the numbers of books that they have written. Tableau is showing all values in the domain, even if there are no matches in the linked table. For example, if you look at the Author table, you will see that the author Wendell Barton has AuthID WB149, which is not a value in the Book table. Rather than stating this is an error, Tableau has treated it as simply a value without any match and assigned a count of zero.

Now move the Number of Checkouts field from the Checkouts table to the Columns shelf. You should see a pair of axes: the left-hand one as before (count of books) and the right-hand one with the number of checkouts. However, values that haven't matched are listed as null. You can see this by the fact that the first author on the list now has the last name 'Null' (not really a last name, of course). And in the bottom right-hand corner of the screen there is a small grey pill with the phrase '8 nulls' inside it which refers to the fact that eight authors match to null in the Checkout table. We can remove the latter by adding a calculation to the Number of Checkouts pill. Left-click at the end of the pill on the little downward triangle, and select 'Edit in Shelf' from the menu. This brings up the Tableau calculation for this variable in the view: edit it to wrap it in the ZN function (which sets the output to zero if the field is null). It should look like this

```
ZN(SUM([Number of Checkouts]))
```

. Hit return, and the '8 nulls' pill should disappear.

Now remove the two measure variables from the Columns shelf and add the Award Name field (from the Award table) to the Rows shelf. You should see a table with the author names in the first column and their awards in the second column. Note that only the authors with an award are listed. Combining dimensions across tables displays the combinations that exist in your data.

Now add the Count measure from the Book table to the Columns shelf. This shows every award for each author's book, so many of the entries have Null for the award name. Null is a rather unhelpful value name: we can change it by right-clicking on the value and selecting 'Edit Alias' from the drop-down menu. Write in 'No Award' (or similar text). The result should look like Figure 10. We conclude that **all records from measure tables are always retained (i.e. they are in the dataset even if they are not displayed)**. Note that an emergent property of contextual joins is that the set of records in your viz can change as you add or

remove fields. While this may be surprising, it ultimately serves to promote deeper understanding in your data. Nulls are often prematurely discarded, as many users perceive them as “dirty data.” While that may be true for nulls arising from *missing* values, *unmatched* nulls (as was the case here for authors without awards) classify interesting subsets of a relationship.

The graph in the previous step showed only authors who have books (so our good friend Wendell Barton was not included). Adding the Count of Author measure from the Author table onto the Marks shelf (e.g. as a tooltip) includes authors without books. Since Tableau always retains all measure values, you can recover unmatched dimensions by adding a measure from their table into the viz.

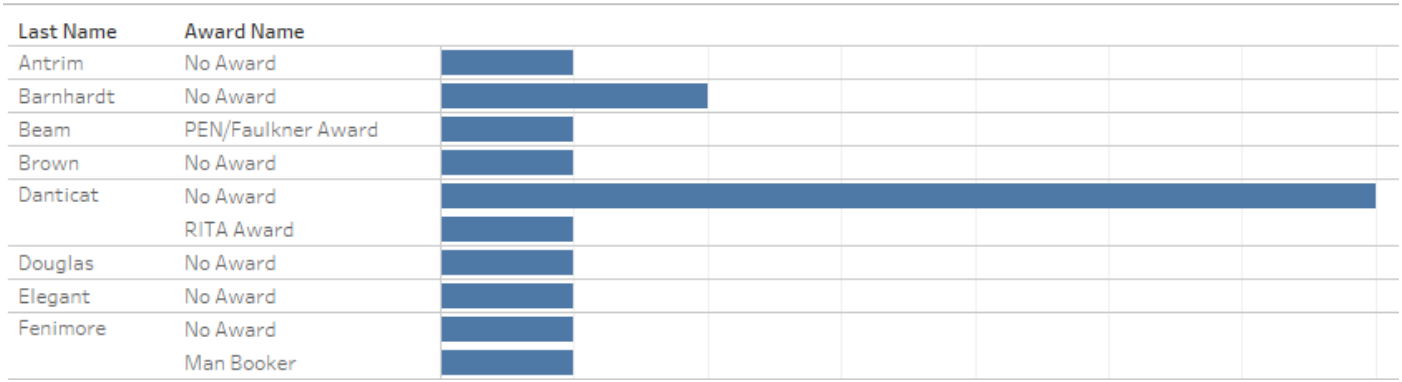


Figure 10: Segment of data for counting the number of books with and without awards for each author.

Remove the attributes from the shelf and the marks shelf. Put Title on the Rows shelf, Publishing House on the Colours mark and then Rating on the Columns shelf. The last is converted into a Sum measure while we want the Average: change it by clicking on the pill in the usual way and change the Measure to Average. You should see one Null title (a book with ratings but no entry) and one zero entry for Cimornul (a book with no ratings). If you hover your mouse close to the title of the x-axis (Avg. Rating) a small symbol should appear. Clicking on it sorts the bars in descending order by that measure (average rating). Clicking again sorts the bars by ascending order. Now click on the AVG(Rating) pill in the Columns shelf and select ‘Show Filter’. You will see that the title Cimornul disappears since the default filter setting is for a non-zero value. Filtering the Count of Ratings, as above, removes books without ratings but preserves reviews that may lack a rating. Excluding null would remove both, because nulls do not discern between missing values and unmatched values. Relationships postpone choosing a join type until analysis; applying this filter is equivalent to setting a right join and purposefully dropping books without ratings. Not specifying a join type from the start enables more flexible analysis.