A Review of Categorical Visualisation Techniques

Suggested citation: Trye, D. (2024). *A review of categorical visualisation techniques* [Unpublished manuscript]. University of Waikato, New Zealand. `https://dgt12.github.io/files/catvis.pdf`

1.1 Introduction

Categorical variables are prevalent in real-world datasets, frequently occurring in domains such as the behavioural and social sciences, public health, biomedical science, education, business and marketing (Agresti, 2012). Examples of categorical data include responses to multiple-choice survey questions (e.g., strongly disagree, disagree, neutral, agree, strongly agree), treatment options assigned to participants in a medical trial (e.g., drug A, drug B, placebo), and the biological class to which different animals belong (mammal, bird, reptile, etc.). Categorical variables are even found in highly quantitative fields, such as industrial quality control, where products are rated based on their adherence to specific standards (Agresti, 2019).

When dealing with categorical data, analysts are typically interested in comparing category frequencies and investigating relationships between categories. Like other data types, the amount of categorical data available is continually growing, increasing the need for efficient analysis methods (Johansson Fernstad, 2011). Surprisingly, despite these demands, visualisation techniques for categorical data have received considerably less attention in the literature compared to those for numeric data (Liu et al., 2016; Friendly, 1998). This is especially true when the need arises to visualise more than three categorical variables simultaneously.

Categorical data visualisation presents several challenges. Firstly, nominal categories do not have an intrinsic order or inherent spatial mapping (Cibulková and Kupková, 2022). Secondly, combinations of categories become

increasingly sparse when more variables are added, exemplifying the 'curse of dimensionality' (Hofmann, 2006). Thirdly, variables with a large number of categories may exceed the limits of the visual encoding, or render a visualisation unreadable. Overall, compared to numeric data, categorical visualisation techniques appear to be more sensitive to structural characteristics of the data (Johansson Fernstad, 2011).

This chapter provides a review and taxonomy of categorical visualisation techniques. We begin by defining key terminology (Section 1.2), before detailing the scope of the review and our method for gathering and organising the relevant literature (Section 1.3). The heart of the chapter describes six distinct 'families' of techniques that we have identified, which form the basis of the proposed taxonomy (Section 1.4). We focus on prototypical examples within each family, then introduce nine different types of analysis tasks from a categorical visualisation perspective (Section 1.5). Finally, we compare general strengths and weaknesses of each family and reflect on opportunities for future work (Section 1.6). An interactive repository of the techniques reviewed in this chapter is available at: `https://cat-vis.github.io/`.

1.2 Categorical Data

Categorical data consist of variables that take a fixed set of values, each representing a distinct category or group, such as colour. Due to their unique characteristics, these variables require different analysis methods from numeric data, including specialised visualisation techniques (Friendly and Meyer, 2015). The main advantage of visualising categorical data is the ability to reveal relationships between multiple variables or categories more clearly than tabular or textual representations.

1.2.1 Terminology

A range of terms is used in the literature to refer to categorical data. Our preferred terms within this thesis are emphasised here in bold. Individual **(data) items** may alternatively be called *objects, cases, records, tuples, points, vectors, observations* or *samples*. The properties of each data item are described by a set of **variables**, where a variable is defined as a characteristic that can vary from one item to another. Variables are sometimes also known as *attributes, features* or *dimensions*. The number of distinct values that a variable can take is its **cardinality**, while the values themselves are variously referred to as **categories**, *levels* or *classes*. We refer to a group of two or more orthog-

onal categories as a **combination of categories**. Categorical variables with only two possible values are sometimes referred to as *binary* variables (Agresti, 2019; Friendly and Meyer, 2015).

Following Tan et al. (2006), we consider a categorical variable to be either **nominal**, meaning its categories are unordered, or **ordinal**, meaning they have a natural ordering. Examples of nominal variables include 'gender' and 'continent', whereas 'customer satisfaction' and 'education level' are both ordinal variables. We consider it important for a categorical visualisation tool to accommodate both these data types. Additionally, quantitative (numeric) variables can be **binned**, or *discretized*, to form (typically) ordinal variables, though this process results in a loss of precision. Two common binning strategies are to create categories of equal *width* or *frequency* (Dougherty et al., 1995). For example, 'income' and 'age' are often divided into specific ranges.

Data can be *univariate*, *bivariate*, or *multivariate*, depending on whether they comprise one, two, or more than two variables, respectively. We use the terms *multivariate* and *multidimensional* interchangeably. Multivariate categorical data are relatively common: census data may include variables such as gender, education level, religion and marital status; medical records might include disease types, treatment protocols and patient outcomes; retail databases frequently categorise products by type, payment method and customer demographics. Analysing all categorical variables simultaneously can enhance understanding of complex relationships and support informed decision-making.

Statistical models often distinguish between **response** (or *dependent*) variables and **explanatory** (or *independent* variables). The latter are thought to partially explain the value of the former. Often, a dataset contains a single response variable and several explanatory variables (Theus, 2008). For example, in the Titanic and Mushroom datasets introduced below, the response variables are *Survived* (yes/no) and *Edibility* (poisonous/edible), respectively. Depending on a user's analysis task, it may be beneficial to highlight a response variable within a visualisation by assigning it a prominent position, for instance, or mapping it to colour.

1.2.1.1 Common Datasets and Data Forms

The *Titanic dataset* (Dawson, 1995; see Figure 1.1) is arguably the most well-known dataset in the field of categorical visualisation. This dataset provides socio-historical information about the passengers and crew aboard the RMS *Titanic*, which tragically sank in 1912. Although the dataset has been the subject of considerable attention (see, for example, Symanzik et al., 2019),

and is widely used for illustrative purposes, it is relatively small, containing only 4 variables, 10 categories and 2201 data items. Several different versions of the Titanic data exist, some of which include the names of passengers as an additional string-type (text) variable. We will use the Titanic dataset in most of the examples in this chapter.

The synthetic *Mushroom dataset* (Schlimmer, 1987), describing properties of mushrooms like their colour, odour and stalk shape, is considerably larger than the Titanic dataset. It comprises 22 variables, 119 categories and 8124 data items, making it a popular choice for demonstrating how categorical visualisation techniques can (or cannot) scale to larger and more complex datasets.

At the internal representation level, Friendly and Meyer (2015) refer to three main forms of categorical data: *case form*, *frequency form* and *table form*, which are illustrated in Figure 1.1. Case form provides each data item as a separate entry, with rows corresponding to data items and columns to variables. This allows any data item to be traced back to its individual identifier. In contrast, frequency form collapses identical combinations of categories into a single row, reporting their counts in an additional column. Finally, table form presents data in a contingency table, which involves cross-tabulating some or all of the available variables.

(a)

ID	Class	Age	Sex	Fate
1	first	adult	male	survived
2	first	adult	male	survived
3	first	adult	male	survived
⋮	⋮	⋮	⋮	⋮
2202	crew	adult	female	died

(b)

Class	Age	Sex	Fate	Freq
crew	adult	male	died	670
third	adult	male	died	387
⋮	⋮	⋮	⋮	⋮
second	child	female	died	0

(c)

		Sex / Fate			
		female		*male*	
Age	**Class**	*died*	*survived*	*died*	*survived*
adult	*crew*	3	20	670	192
	first	4	140	118	57
	second	13	80	154	14
	third	89	76	387	75
child	*crew*	0	0	0	0
	first	0	1	0	5
	second	0	13	0	11
	third	17	14	35	13

Figure 1.1: The Titanic dataset shown in (a) case form, (b) frequency form and (c) table form. The *Survived* (yes/no) variable from Dawson's (1995) original dataset has been renamed *Fate* (survived/died) to give the two categories semantically descriptive names.

1.3 Scope and Methodology

In this review, we focus on visualisation techniques that are capable of showing *purely* categorical data, for any number of variables. We limit our analysis to techniques that treat variables as having *flat* and *disjoint* categories. In other words, the categories within each variable lack *sub*-categories, and are mutually exclusive. Datasets that include multi-value categories are likely better modelled as sets (Alsallakh et al., 2016). Furthermore, our review focuses on exploratory data analysis rather than on statistical model building (see Friendly and Meyer, 2015). Categorical data with special properties fall outside the scope of this review, including geospatial and time-oriented data, as well as relational data with categorical attributes.[1]

This chapter synthesises ideas and techniques for visualising categorical data from roughly 120 papers. The literature was extracted by paying special attention to publications from *IEEE Xplore, EuroGraphics, Sage Information Visualization* and the *Journal of Computational and Graphical Statistics* that explicitly mentioned 'categorical' data in the title or keywords. We also expanded our search to include literature cited by these papers, as well as work that cited them. The collected papers were tagged according to their primary contribution, the vast majority (80%) being *technique* papers:

- *technique:* the paper introduces a specific technique or system for visualising categorical data.
- *evaluation:* the paper provides an empirical, algorithmic or theoretical evaluation of visualisation approaches for categorical data.
- *ordering algorithm:* the paper contributes an algorithm for rearranging categorical data.
- *framework:* the paper contributes a framework or paradigm for visualising categorical data.
- *textbook:* a textbook on the topic of visualising categorical data.
- *survey:* the paper presents a survey of categorical data visualisation or a related field.

Technique papers were tagged according to five further attributes that we deemed important, as outlined in Table 1.1.

[1] We do, however, explore this further in our final case study, in Chapter 9.

Table 1.1: Details of the five attributes by which technique papers were tagged.

Category	Description
Family	
Size-Encoding	The technique uses bars (line marks) with the length channel, or wedges (area marks) with the angle or length channels.
Space-Filling	The technique fills the available space and likely imposes a hierarchy of variables.
Table	The technique represents data in a 2D table or matrix, where each cell contains visual encodings.
Glyph	The technique uses glyphs or icons to represent individual items or aggregates in the dataset.
Other	The technique represents frequencies (in line with the CatViz approach) but does not fit into any of the above categories.
Projection	The technique converts categories into numerical values before representing these visually (in line with the QuantViz approach).
Data Type	
Homogeneous	The technique only supports categorical (not quantitative) data.
Heterogeneous	The technique supports a mixture of categorical and quantitative data.
Dimensionality	
Univariate	The technique supports only one categorical variable.
Bivariate	The technique supports up to (or exactly) two categorical variables.
Trivariate	The technique supports up to (or exactly) three categorical variables.
Multivariate	The technique can support more than three categorical variables.
Cardinality	
Very Low	The technique requires at least one binary variable.
Low	The technique supports variables with roughly (only) 2-5 categories.
Moderate	The technique can handle at least one variable with 6-10 categories.
High	The technique is designed to support at least one variable with 10-100 categories.
Very High	The technique is designed to support at least one variable with 100+ categories.
Alignment	
Linear	The technique arranges data along perpendicular or parallel axes.
Radial	The technique is laid out in elliptical form, and likely uses polar coordinates.
Other	The technique does not use a linear or radial layout (e.g., force-directed).

The *families*, which form the basis of our proposed taxonomy, are explained in detail in Section 1.4. It was sometimes necessary to make subjective judgements when assigning these tags, if relevant details were not overtly mentioned

in the paper. We acknowledge that the *interplay* between a technique's supported cardinality and dimensionality is important, though this was not explicitly coded. Our final literature collection can be interactively explored at: https://cat-vis.github.io.[2]

1.3.1 Technique Taxonomy

Given the focus of this thesis on visualisation methods, the technique papers were fundamental to the current review. We have organised this body of literature into a two-level taxonomy, as shown in Figure 1.2. The first-level classification groups techniques into *CatViz* (frequency-based) and *QuantViz* (quantification-based) approaches, following Johansson Fernstad and Johansson (2011). The CatViz approach involves directly mapping the cell counts from a contingency table, using a visual representation suitable for *categorical* data. In contrast, the QuantViz approach projects categories onto a (typically) two-dimensional plane using quantification methods, and then represents the data visually using any technique designed for *numeric* data. The quantification approach aims to preserve relationships, such as distances, similarities and associations between data points. Each approach has its own merits: in an initial user study (ibid), CatViz techniques were found to be superior for *frequency* tasks (e.g., identifying the most frequent category), while QuantViz techniques were found to be better suited for *similarity* tasks (e.g., determining which two categories are most alike).

In addition, we developed a second-level classification, based on 'families' of visualisation techniques. We have identified six families but, as new techniques emerge, others can be added. Five of the six families relate to the CatViz approach: *size-encoding, space-filling, table, glyph* and *miscellaneous*. The remaining category, *projection*, encompasses any visualisation technique used as part of the quantification approach. The projection family is highly versatile, since converting categories to numbers fundamentally changes what can be done with the visual representation. We note that these families are not mutually exclusive: for instance, *dimensional stacking* (Section 1.4.3.2) can be regarded as a hybrid table/space-filling technique.

[2]The database was created using the *SurVis* template (Beck et al., 2015).

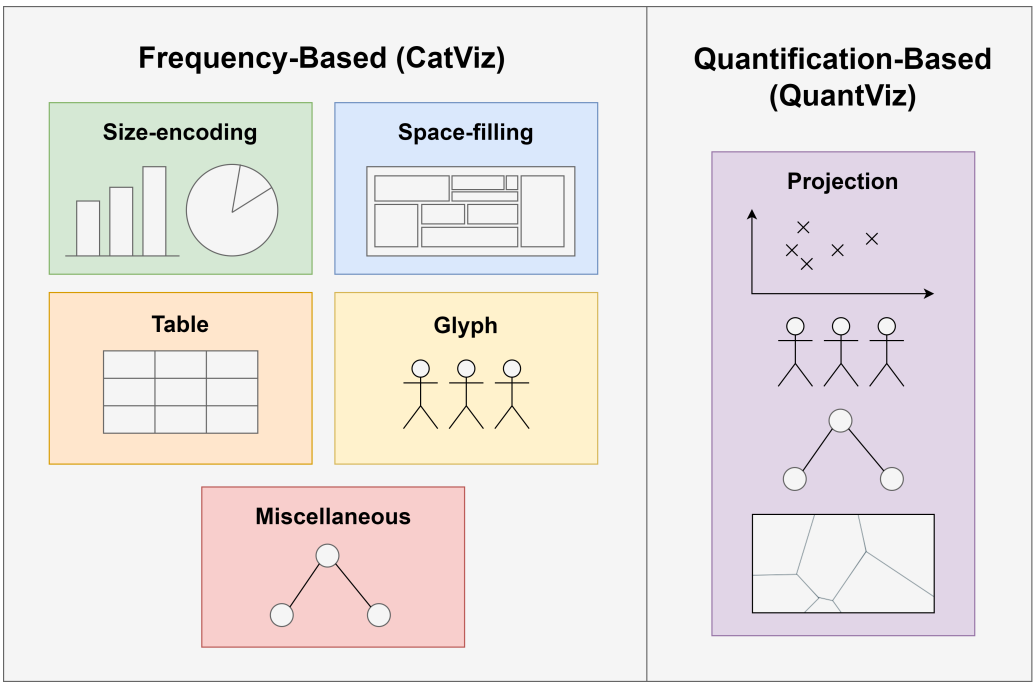

Figure 1.2: Our proposed taxonomy comprises six 'families' of techniques: *size-encoding, space-filling, table, glyph, miscellaneous* (all frequency-based) and *projection* (quantification-based). The rectangle for *projection* is larger to indicate that it encompasses many different possible representations for numeric data.

1.4 Overview of Technique Families

In this section, we describe the six families of techniques, breaking these down into further sub-categories where appropriate. At least one visualisation technique is reviewed in each section, and references are given for related methods.

1.4.1 Size-Encoding Techniques

We define *size-encoding* techniques as those which use *bars* (line marks) with the length channel, or *wedges* (area marks) with the angle or length channels. Consequently, this family can be clearly divided into a bar family and a wedge family. Most techniques in the bar family have *linear* alignment, while those in the wedge family are *radial*. Although equivalent from a mathematical point of view, the wedge family is generally less effective than the bar family, since angles are harder to compare than lengths (**?**). The *Trellis display framework* Becker et al. (1996) can be applied to many size-encoding techniques to encode additional categorical data via faceting.

1.4.1.1 Bar Family

Dating back to the latter half of the 18th century (Playfair, 1786, as cited in Friendly, 2006), the *bar chart* (or *column chart*) is a simple yet powerful technique for encoding categorical data. As well as being easy to create and interpret, bar charts are helpful for highlighting precise differences in category counts. For nominal variables, the categories within a bar chart should generally be sorted by frequency (i.e., bar length); for ordinal variables, it may be preferable to preserve the natural ordering of categories. While the classic bar chart is limited to displaying a single categorical variable, numerous variations exist, many of which enable additional variables to be encoded by leveraging colour, texture and/or faceting. These extensions include:

- *Stacked bar charts* and their variants (see Figure 1.3; Indratmo et al., 2018; Streit and Gehlenborg, 2014):
  - *Grouped bar charts* (also called *clustered bar charts*, *dodged bar charts*, *multiple bar charts*, and *multi-series bar charts*)
  - *100% stacked bar charts* (also called *normalised bar charts*)
  - *Layered bar charts*
  - *Diverging stacked bar charts* (also called a *bidirectional bar chart* if the coloured variable is binary)
  - *Inverting stacked bar charts*
  - *Faceted bar charts*
  - *Relative multiples barcharts (rmb plots)*
- *Linked bar charts* (Hummel, 1996), as implemented in tools like *Mondrian* (Theus, 2002) and *High-D* (Brodbeck and Girardin, 2019)
- *Horizon bars* (Lex et al., 2014)
- *Du Bois wrapped bar charts* (Karduni et al., 2020)
- *Pareto charts* (Wilkinson, 2006)
- *Radial bar charts* (Booshehrian et al., 2011)
- *Circular bar charts* (Skau and Kosara, 2016)

Taking one of the most popular examples from this list, the stacked bar chart (Figure 3, top left) typically encodes the frequency of two categorical variables, rather than just one. The first variable determines the categories for the bars along the x- or y-axis, as in a regular bar chart, while the second variable is broken down into segments within each bar. These segments are typically distinguished by colour and are consistently ordered across all bars. Stacked bar charts show the marginal distribution of the first variable and the conditional distribution of the second variable (i.e., the distribution of the second variable given the first one). This means that reversing the roles of the

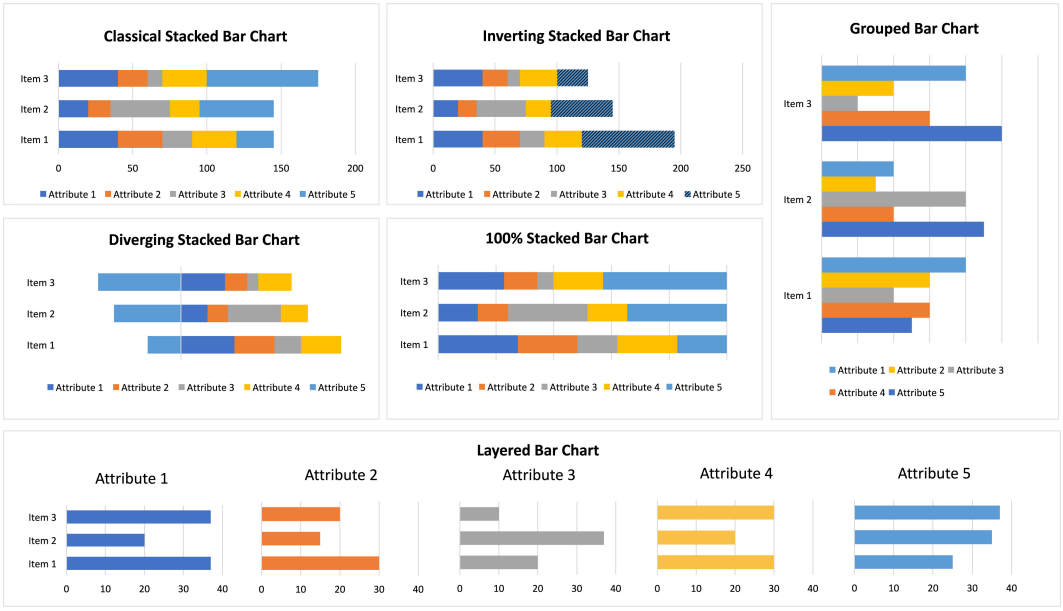

Figure 1.3: Six different variations of stacked bar charts (Indratmo et al., 2018).

variables would result in a different plot and potentially yield different insights.

As with a regular bar chart, the scalability of a stacked bar chart ranges from dozens to hundreds of categories for the axis variable, but is limited to roughly a dozen categories for the second variable **?**. Comparing both the total length and the bottom segment of each bar is straightforward because they share a common baseline, but comparing other segments is more challenging.

Bar charts can display more than two variables by 'chaining' multiple variables along the same or different axes, as shown in Figure 1.4. The bars in the resulting visualisation show the joint frequency of each combination of categories involving all variables. Dozens to hundreds of bars can be shown, and up to roughly eight variables. However, the more variables that are shown, the less room there is to display the labels for each category. This kind of visualisation imposes a hierarchy of variables (like most *space-filling* techniques), which means changing the order of variables can affect the patterns seen, even though the values of the bars remain unchanged. Tooltips and drag-and-drop reordering may help to make sense of patterns in the data.

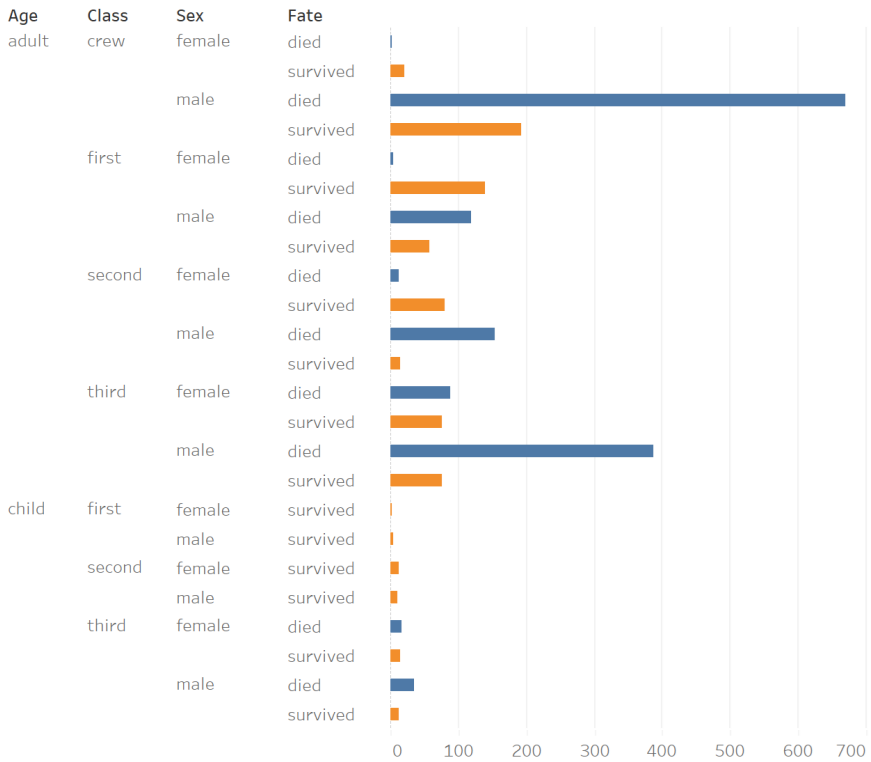

Figure 1.4: 'Multivariate' bar chart showing the joint frequency of all four variables from the Titanic dataset. Colour redundantly encodes Fate (blue = died, orange = survived).

1.4.1.2 Wedge Family

Members of the wedge family use area marks, rather than line marks, to show frequency. *Pie charts* (Playfair, 1801) and their close cousins, *donut charts* (Skau and Kosara, 2016), are useful for representing proportions or percentages of a whole when there are 12 categories or fewer. They are effective for comparing one category relative to the whole dataset, but not for comparing the proportion of one category to another, except when the variation is extreme, or there are only two categories. Other members of this family include:

- *Nightingale rose chart* (Nightingale, 1857), also known as *sector graphics, Coxcomb charts* and *polar area diagrams*
- *Wind roses* Sanderson and Peacock (2020)
- *Four-fold displays* (Fienberg, 1975; Friendly, 1995))

Although aesthetically pleasing, perceptually, pie and donut charts are known to be less precise than bar charts. Figure 1.5 provides an example of a *faceted pie chart*, representing three of the four variables in the Titanic dataset. However, such charts should be used with caution. In his book *The Visual Display of Quantitative Information*, Edward Tufte (1983, p. 178) re-

Figure 1.5: A faceted pie chart of the Titanic dataset: Class is shown on the x-axis, Sex on the y-axis and Fate is mapped to colour (blue = died, orange = survived). It is clear that many more men than women died in each class.

marked: "the only thing worse than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies". Despite their limitations, pie and donut charts are pervasive and participants in a user study expressed a subjective preference for them over bar charts (Siirtola, 2014).

1.4.2 Space-Filling Techniques

As the name suggests, space-filling techniques are arranged so that the layout consumes all available space in the view. In the context of multivariate categorical data, these techniques typically use area or containment marks to show different combinations of categories. Space-filling techniques are geared towards high information density, but the fact that they consume all the available space does not necessarily mean they do so efficiently (**?**, p. 175).

A variety of space-filling techniques can be applied to multivariate categorical data by creating a hierarchy of variables (Reza and Watson, 2019; Kosara, 2008). This is despite the fact that the data in question are not inherently hierarchical (i.e., categories do not have sub-categories). The hierarchy is derived by mapping each categorical variable to a different level, with all categories of the first variable at the top level, all categories of the second variable at the second level, and so on. This results in a fully balanced tree whose nodes represent different combinations of categories. Figure 1.6 shows an example for the Titanic data, together with a corresponding *treemap* (see Section 1.4.2.3).

The order of variables in the hierarchy is significant as it affects the user's ability to perceive structures. This ordering becomes even more crucial as the

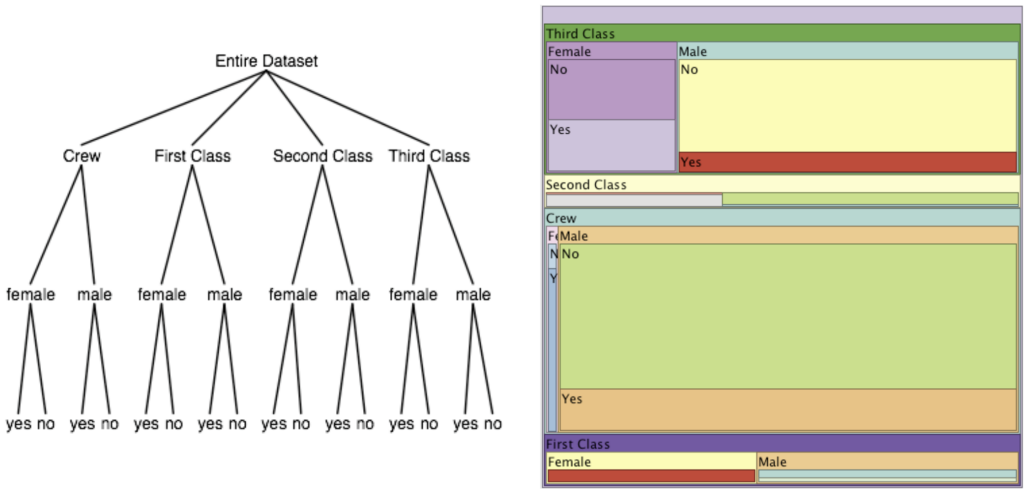

Figure 1.6: Left: Hierarchy derived from three of the four variables in the Titanic dataset, splitting first by Class, then Sex, then Fate. Right: Treemap using the same hierarchical structure, which shows values at the leaves of the tree (the frequency of combinations of all three variables), as well as aggregates at higher levels (Kosara, 2008).

number of variables increases. It is therefore important for the user to be able to reorder, add or remove variables as desired (Kosara, 2008). Relevant factors for determining an appropriate order may include the position of the response variable, the perceived importance of other variables, and the distribution of variable cardinalities. A good ordering for one technique might also differ for another. Colour is commonly used to highlight the response variable.

1.4.2.1 ParSets Family

Several categorical visualisation techniques adapt *parallel coordinates* for numeric data (**?**) by substituting data points with a frequency-based representation. *Parallel Sets* (Kosara, 2010; Kosara et al., 2006), pictured in Figure 1.7, is the most well-known technique among this family. Reminiscent of a *Sankey diagram* (Schmidt, 2006), this technique arranges variables along the y-axis in bands of equal width, which are then partitioned according to category frequencies. Associations between subgroups are shown using shaded parallelograms (or ribbons) that connect categories from adjacent dimensions. The widths of individual categories indicate marginal frequencies, while the widths of parallelograms reflect both joint frequencies (relative to the width of the display) and conditional frequencies (relative to the width of the previous subset). Numeric variables can be binned but not shown directly.

Two variations of Parallel Sets are possible: *hierarchical* and *pairwise* (see

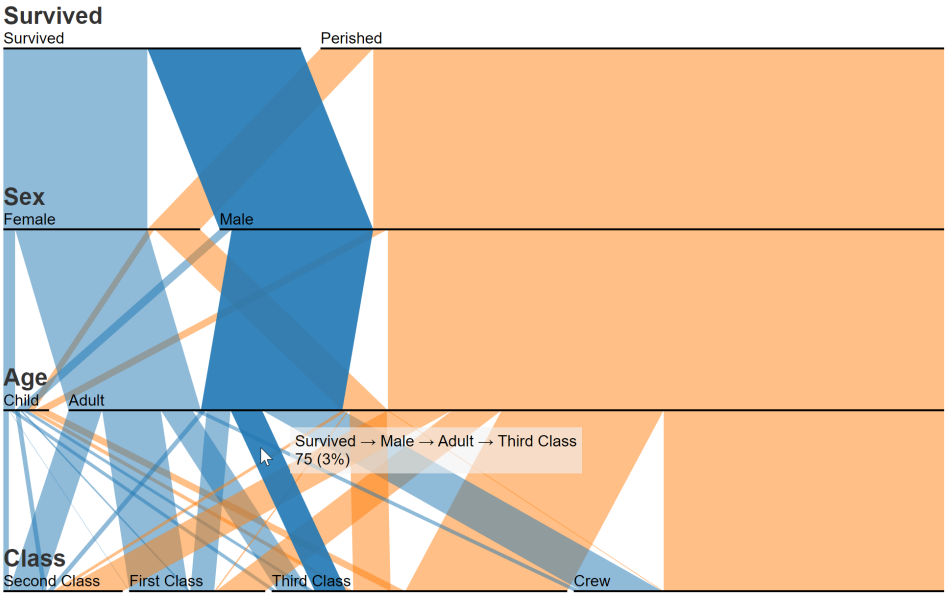

Figure 1.7: A Parallel Sets visualisation of the Titanic dataset, showing all four variables (Davies, 2012).

Hofmann and Vendettuoli, 2013). In the hierarchical variation (described above, and shown in Figure 1.7), the parallelograms are split according to every preceding variable, resulting in increasingly complex, and less frequent, subsets. In contrast, the pairwise variation displays two-dimensional subsets relating to each pair of neighbouring variables. The hierarchical view is more useful for visualising multivariate relationships but is inevitably more cluttered.

The main advantage of Parallel Sets is that it can handle roughly 10–15 variables in an interactive environment and 20-30 categories in total, which exceeds the limits of most frequency-based techniques. In addition, the order in which the hierarchy is derived is clearly readable—from top to bottom—and categories and variables can be flexibly reordered, facilitating detection of complex patterns in the data. Parallel Sets can also display numeric variables by binning them.

Key limitations of Parallel Sets include visual interference from line crossings and poor visibility of small parallelograms representing infrequent combinations. These issues are exacerbated when handling large numbers of categories and variables. For example, the Mushroom dataset requires 22 layers and 8123 combinations, which is untenable (Dennig et al., 2024). To alleviate visual clutter, research has focused on measuring and improving the layouts of Parallel Sets (Alsakran et al., 2014; Dennig et al., 2021; Zhang et al., 2019).

Other techniques in the ParSets family, all of which can display mixed data,

are *Hammock Plots* (Schonlau, 2003, 2024), *CPCP* (Pilhöfer and Unwin, 2013), *GPCP* (VanderPlas et al., 2023), *Parallel Assemblies Plots* (Cantu et al., 2023) and *SET-STAT-MAP* (Wang et al., 2022). Hofmann and Vendettuoli (2013) observed that Parallel Sets and Hammock Plots suffer from the *line width illusion* and *reverse line width illusion*, respectively. They proposed *Common Angle Plots* to overcome these distortions, while Schonlau (2024) suggested a correction to Hammock Plots by replacing the parallelograms with rectangles.

Finally, we note that *chord diagrams* (inspired by Krzywinski, 2009) can be used to visualise relationships between two categorical variables (Humayoun et al., 2018). Chord diagrams are related to techniques in the ParSets family as they emphasise the flow of category subsets, but they are limited to showing only two variables in the same plot.

1.4.2.2 Mosaic Family

Techniques in the Mosaic family are largely area-proportional, with colour often being used to highlight particular variables or statistical information. The technique after which this family is named, the *mosaic plot*, was introduced by Hartigan and Kleiner (1981) and further developed by Friendly (1999). An example of a mosaic plot is given in Figure 1.8. In this technique, area-proportional tiles are created by recursively subdividing the space along the axes based on the categories of each variable. In addition to showing joint frequencies through the size of the tiles, mosaic plots show the marginal proportion of the first variable used for splitting, and the conditional proportions for each subsequent variable based on the previous ones. A useful property of mosaic plots is that the cells are aligned when variables are independent Friendly (1999). Unfortunately, mosaic plots become difficult to read when representing more than three variables, or a large number of categories.

Residual-based shading of the tiles in a mosaic plot can visually indicate the lack of fit of a specific log-linear model (Friendly, 1994) or the statistical significance of test results (Zeileis et al., 2007). Commonly, two shades for both positive (blue) and negative (red) residuals are used. The shading usually either reflects significance at 90% or 99% confidence levels, or employs fixed cut-offs at ± 2 and ± 4, corresponding to *individual* significance at alpha levels of $\alpha = 0.05$ and $\alpha = 0.001$, respectively (Friendly, 1994). The use of residuals works well for large tiles but not for smaller ones as it is difficult to make out the colours. Moreover, the difference of size and colour may lead to misinterpretations of the data; for instance, if two tiles have the same colour but are drastically different sizes, a viewer may mistakenly believe the larger

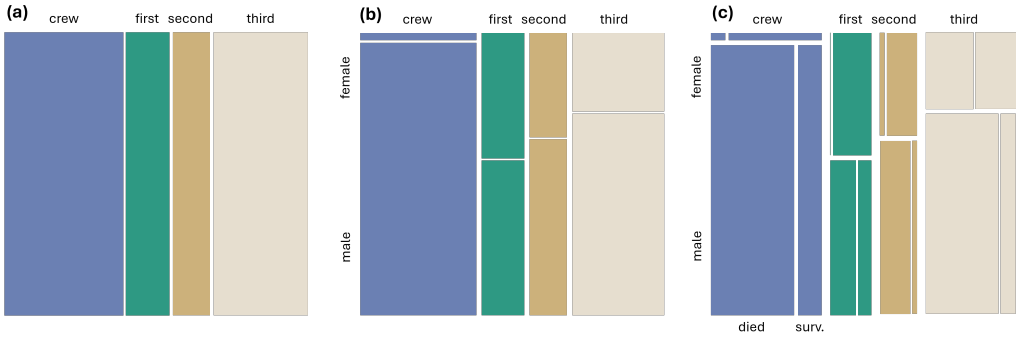

Figure 1.8: Mosaic plot of the Titanic dataset, illustrating the splitting process for three variables: (a) first by Class, (b) then by Sex, (c) then by Survived. Age is not shown.

one has a larger residual.

In addition to the traditional mosaic plot, the mosaic family comprises the following chart types:

- *Spine plots* (Hummel, 1996, Figure 1.8a)
- *Line mosaic plots* (Huh, 2004)
- *Marimekko charts* (Miyamoto et al., 2022)
- *Eikosograms* (Cherry and Oldford, 2003)
- *Double-decker plots* (Hofmann et al., 2000; Hofmann, 2001)
- *Sieve plots* or *parquet diagrams* (Riedwyl and Schüpbach, 1994)
- *Association plots* (Cohen, 1980)
- *Fluctuation diagrams* (Hofmann et al., 2000)
- *Equal bin size plots* (Hofmann et al., 2000)
- *Faceted mosaic plots* (Meyer et al., 2008)
- Additional variations resulting from the *Product Plots* framework (Wickham and Hofmann, 2011)

These charts have different strengths and weaknesses. For example, fluctuation diagrams and equal bin size plots are useful for emphasising patterns related to data sparsity, including empty combinations Hofmann et al. (2000). Sönning and Schützler (2023) suggest that double-decker plots may be preferable to traditional mosaic plots when a dataset comprises three or more variables, as this avoids comparisons of non-aligned tile lengths. In turn, *rmb plots* (Section 1.4.1.1) are generally a better option than double-decker plots when both the frequencies of combinations of explanatory variables vary considerably, and the conditional relative frequencies of response categories, or the difference between them, is small (Pilhöfer and Unwin, 2013).

Some techniques within the mosaic family represent observed frequencies

less directly than traditional mosaic plots, either by emphasising expected frequencies (e.g., sieve plots) or deviations from expected independence (e.g., association plots). Association plots and fluctuation diagrams were classified within the mosaic family, rather than the size-encoding family, since both the width and heights of the bars vary.

1.4.2.3 Implicit Tree Family

Implicit tree visualisations constitute another relevant type of space-filling technique. These visualisations represent hierarchies without explicitly showing parent-child relationships, instead using positional encodings of nodes, such as node overlap or containment (Schulz et al., 2010). The techniques that work best for multivariate categorical data emphasise the *size* of nodes within a visualisation, corresponding to combination frequencies, more than they do the *structure* of the tree.

A prominent technique in the Implicit Tree family is the *sunburst diagram* (Stasko and Zhang, 2000). This technique shows the proportion of different categories and combinations of categories via a series of concentric rings. Each ring corresponds to a different variable, with the angle of each slice being proportional to the frequency of the category (first level) or combination of categories (subsequent levels) that it represents. Figure 1.10 illustrates two examples for the Titanic dataset. Outer levels are conditioned on inner levels, effectively showing conditional relative frequencies. If too many variables are shown, the slices or rectangles invariably become thin and unreadable. However, zoomable versions of sunburst diagrams can help to accommodate a larger number of categories and variables.

Other implicit tree techniques that can be applied to multivariate categorical data include:

- *Categorical Treemaps* (Johnson, 1993), including *CatTrees* (Kolatch and Weinstein, 2001)[3]
- *Voronoi treemaps* (Balzer and Deussen, 2005)
- *Circular treemaps* (Wang et al., 2006), also called *circle packing*, *packed circles* and *pebbles*
- *Icicle plots* (Kruskal and Landwehr, 1983)
- *Radial Icicle Trees* (Jin et al., 2023)
- *Hi-D Maps* (Reza and Watson, 2019)

[3]Although devised independently, these are similar to mosaic plots.

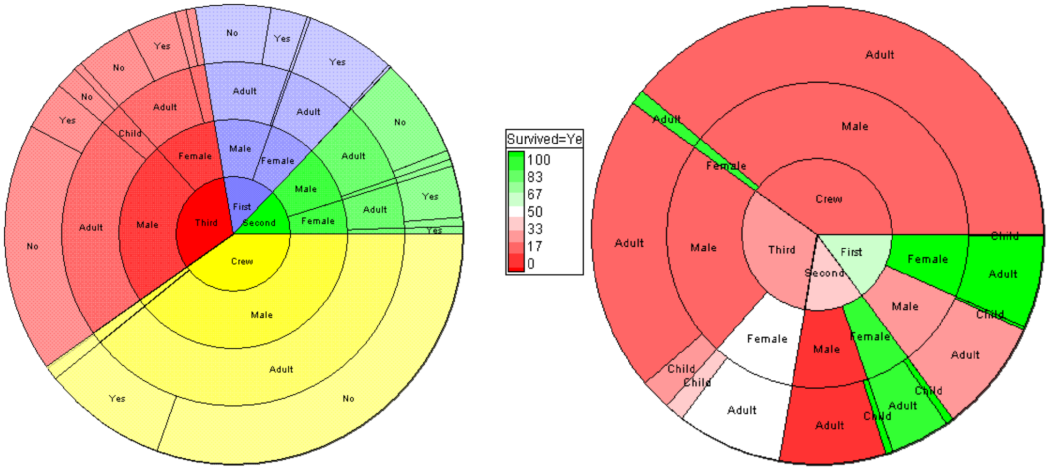

Figure 1.9: Left: Sunburst diagram showing all four variables of the Titanic dataset. Right: One of the variables (Survived) is removed from the sunburst itself and instead emphasised using colour (Clark, 2006).

1.4.2.4 Miscellaneous Space-Filling Techniques

A handful of space-filling techniques for categorical data do not fall neatly into the above families. These include *Karnaugh-Veitch-Maps* (*KVMaps*; May et al., 2007; 2010), *Nested Rings* (*NRV*; Vivacqua and Garcia, 2008), the *Attribute Map View* (Liu et al., 2009) and *concentric pie charts* (Wickham and Hofmann, 2011). On the surface, Nested Rings appear similar to sunburst diagrams but they are not recursively subdivided; instead, they show marginal (univariate) frequencies for each variable. This is also how the Attribute Map View differs from a regular treemap.

1.4.3 Table Techniques

Techniques in the table family utilise visual encodings within each cell of a table, such as colours and bars, instead of displaying only raw text. We divide these techniques into three sub-categories: *tabular, graphical contingency tables* and *pairwise matrices*. Tabular techniques and pairwise matrices are generally well-suited to heterogeneous data, while graphical contingency tables are designed for purely categorical data.

1.4.3.1 Tabular Family

Tabular visualisations leverage the intuitiveness of a spreadsheet, with each row (or column) representing a data item or aggregate, and each column (or row) representing a variable. Prominent examples of tabular techniques

that accommodate multiple categorical variables include *Table Lens* (Rao and Card, 1994) and *Taggle* (Furmanova et al., 2020). Table Lens displays each categorical variable as a 'blip'—a horizontal coloured line aligned with the category's name—while Taggle provides additional multi-form encodings, including a 'matrix' arrangement and 'colour' square with an adjacent text label. Both techniques support common spreadsheet operations, such as sorting and filtering, as well as overview and detail displays. Another notable technique in this family is the *Tableplot* (Tennekes and de Jonge, 2013; Tennekes et al., 2013), which requires a numeric variable for sorting but supports several high-cardinality categorical variables. The data and legend are not the focus here; the figure is simply included to provide the general look and feel of this technique.

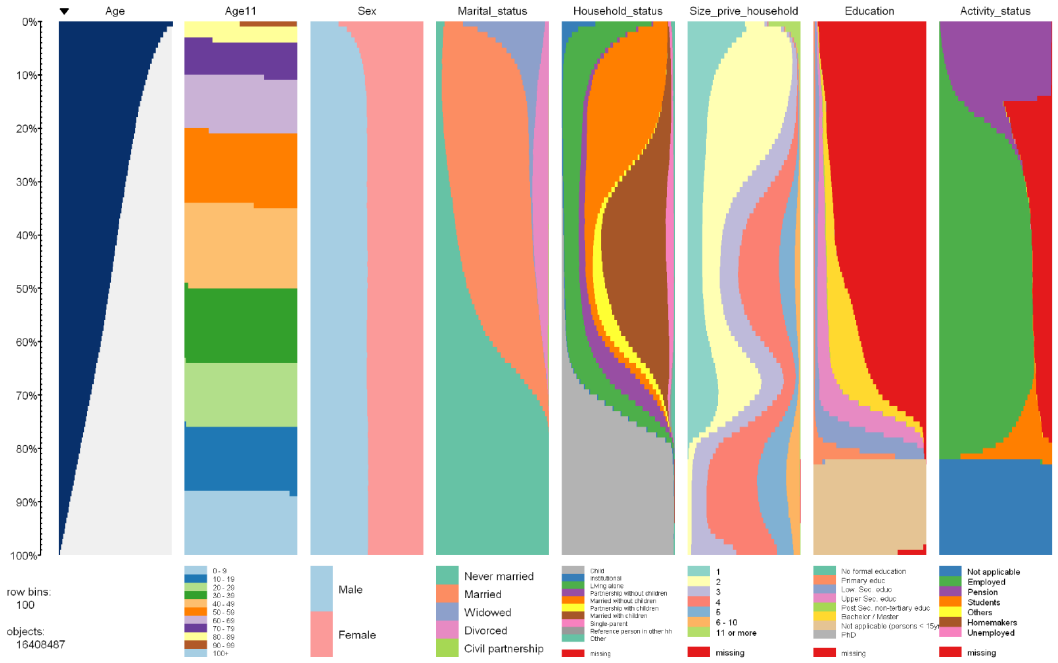

Figure 1.10: Tableplot of census data showing seven categorical variables (Tennekes and de Jonge, 2013).

1.4.3.2 Graphical Contingency Tables

Graphical contingency tables provide a visual representation of an n-way table. The arrangement of variables and categories within the table affects the patterns that can be seen.

Notable examples of techniques in this family are *dimensional stacking* (LeBlanc et al., 1990), *colour-coded text tables* and *balloon plots* (Jain and Warnes, 2006). Dimensional stacking produces a heatmap, like the one in Figure 1.11, by embedding grids within grids. The heatmap contains one cell for

each possible combination of categories, and is helpful for identifying clusters, outliers and patterns in the data. Dimensional stacking can be implemented in spreadsheet software using *Pivot tables* in conjunction with conditional formatting.

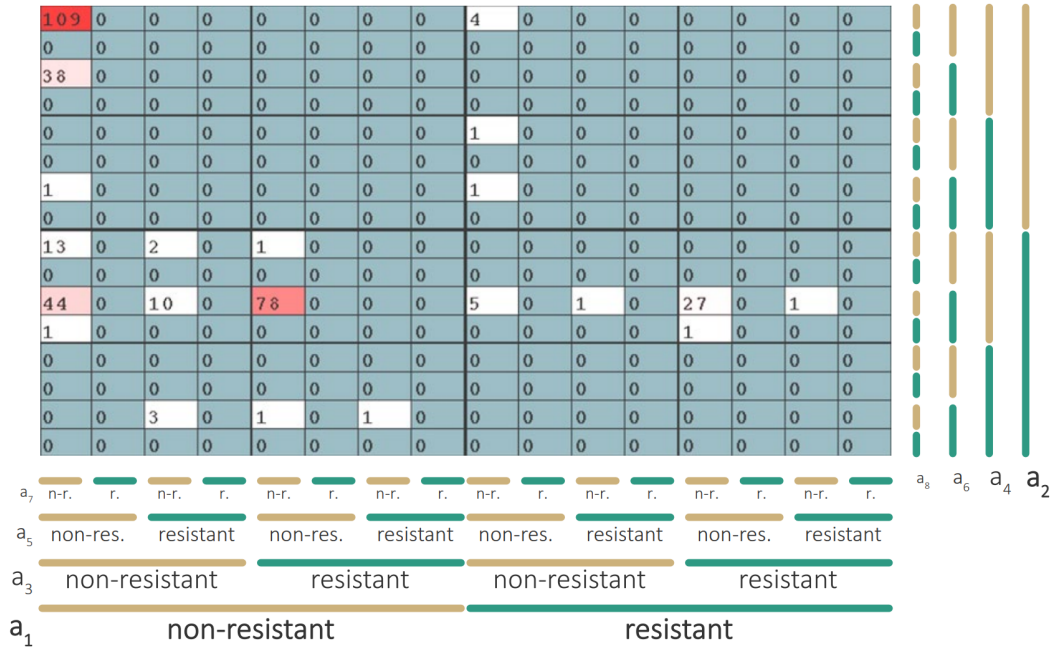

Figure 1.11: Dimensional stacking showing Bacteria resistance against eight antibiotics, labelled $a_1 - -a_8$ (Tominski and Schumann, 2020).

In terms of scalability, dimensional stacking should be limited to nine variables, each with no more than roughly five categories (Hoffman and Grinstein, 2001). Balloon plots are similar to dimensional stacking, but they display coloured circles in each cell, which are sized according to frequency. The colour of the circles can either redundantly encode this value or highlight the categories of a particular variable of interest.

1.4.3.3 Pairwise Matrices

The final type of table technique that we identified is pairwise matrices. These techniques feature a plot for each pair of variables in the data, thereby displaying all possible bivariate relationships. Univariate summaries can optionally be shown along the main diagonal. Examples that support purely categorical data are the *Heatmap Matrix* (Rocha and da Silva, 2018, 2022) and *Mosaic Matrix* (Friendly, 1999), while the *GPLOM* (Im et al., 2013) and *Generalized Pairs Plot* (Emerson et al., 2013) are suitable for mixed data. The GPLOM uses a heatmap matrix for pairs of categorical variables, whereas the Generalised Pairs Plot offers a choice between a mosaic plot, fluctuation diagram, or

faceted bar chart. The GPLOM provides the richest interaction out of these techniques.

A shared property of most pairwise matrices—apart from displays involving mosaic plots—is that they are fully symmetrical. This means that only half of the matrix needs to be displayed. Nevertheless, it can be beneficial to keep the full display, so that panels relating to each variable can be identified in a straight line, with the user focusing on either rows or columns. The Heatmap Matrix differs from the other techniques in that it allocates a fixed amount of space per category, rather than per variable. This enhances readability when a small number of variables have more than five categories. One limitation of pairwise matrices is that they do not display multivariate relationships directly, though this can be accomplished via brushing and linking across panels. In all cases, reordering rows and columns can be helpful for identifying relevant patterns.

1.4.4 Glyph Techniques

Glyphs and icons can also be used to represent categorical data, including pictorial, associative and geometric symbols (Robinson et al., 1984, p. 288). When designing glyphs for categorical data, it is important to consider the number of variables and internal categories to be represented, as well as suitable combinations of variables and encodings. Individual glyphs may be created for individual items, or for each combination of categories. In the latter case, the frequency of each combination can be mapped to the size of the glyph (e.g., Dennig et al., 2024). Incorporating a reference glyph can aid viewers in decoding the mappings (Maguire et al., 2012). Additionally, sorting glyphs by one or more variables can be beneficial (Chung et al., 2015; Ancker et al., 2011).

Examples of glyph techniques include *Star plots* (Coekin, 1969), *Autoglyphs* (Beddow, 1990) and *Chernoff faces* (Chernoff, 1973), but see Ward (2002) for a detailed list. Chernoff faces involve mapping variables to facial features, such as mouth size and face colour, and they support roughly a dozen variables. They are well suited to low-cardinality categorical data where not many values have to be discriminated. Disadvantages of Chernoff faces include that the mappings can be unnatural, and may convey unintended emotional states. De Soete and Do Corte (1985) found that only some facial features were clearly perceptible to users. Consequently, they recommended using those features for encoding the most important variables.

An advantage of glyphs over other techniques is that they enable designs that leverage metaphors and semantic relations. Domain-specific encodings

promote 'natural mappings' (Siirtola, 2005), which increases understanding of glyphs and their memorisation (Maguire et al., 2012; Borgo et al., 2013). An example of metaphorical glyphs, applied in the context of hearing loss context, is shown in Figure 1.12 (Ramos et al., 2023).

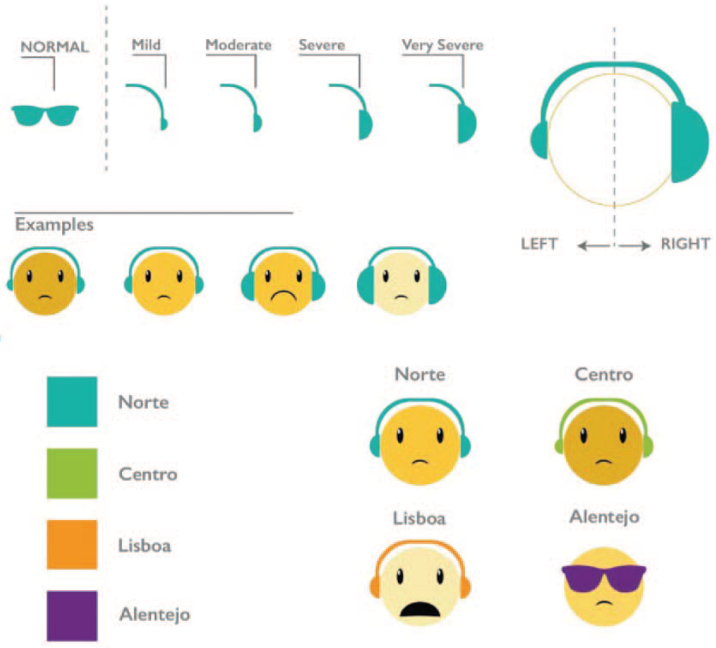

Figure 1.12: Metaphorical 'emoji' glyph, where each glyph represents a person. Several categorical variables are encoded: hearing loss in left and right ears (sunglasses or headphones), region (colour), ear test appointment status (facial expression) and age (face colour).

When representing categorical data, glyphs are typically only feasible if there is a relatively small number of categories per variable. Other, more general disadvantages of glyphs relate to their size, the limited capacity of visual channels and the cognitive demand they place on the viewer (Borgo et al., 2013).

Alternatively, instead of using complex glyphs, simple icons can be organised within a grid display, typically just varying the use of colour and/or shape. This approach is exemplified by *frequency grids* (Micallef et al., 2012), *Gatherplots* (Park et al., 2023), and *icon plots* (Wolf, 2021). Figure 1.13 illustrates an icon plot of the Titanic dataset, in which each full-sized icon represents 100 people. Such plots are relatively simple to interpret.

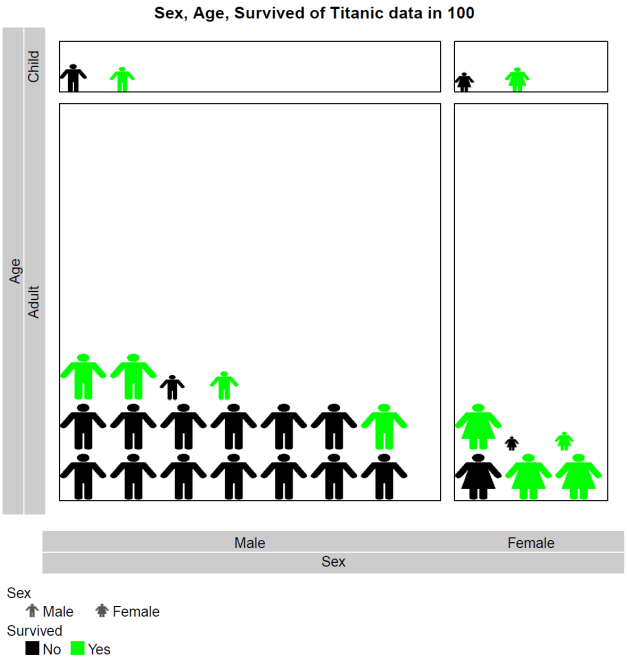

Figure 1.13: Icon plot of the Titanic dataset where each full-sized item represents 100 people.

1.4.5 Miscellaneous Techniques

Several other frequency-based (CatViz) visualisation techniques for categorical data do not fit into the above groups. These include but are not limited to:

- *Cleveland dot plots* (Cleveland, 1984) and *lollipop charts*
- *Spreadplots* as implemented in ViSta (Valero-Mora et al., 2003)
- *Granular Representation* (Shiraishi et al., 2009)
- *Kinetica* (Rzeszotarski and Kittur, 2014)
- *Cobweb diagrams* (Upton, 2000)
- *CatNetVis* (Thane et al., 2023)
- *Conditional Inference Trees* (Hothorn et al., 2006)
- Multivariate bar charts with an explicit *tree diagram* Kosara (2007)
- *ContingencyWheel* and *ContingencyWheel++* (Alsallakh et al., 2011, 2012)
- *Worlds within worlds* (Feiner and Beshers, 1990)
- *Trilinear plots* (Allen, 2002)
- *Tetrahedrons* (Fienberg and Gilbert, 1970)
- Various *set* and *hypergraph* representations, where categories are represented as sets (e.g., *RectEuler*; Paetzold et al., 2023) or hyperedges (e.g., *PAOHVis*; Valdivia et al., **?**)

CatNetVis (Thane et al., 2023), shown in Figure 1.14, represents categories

as nodes in a force-directed network. Connected nodes are attracted to each other, and non-connected nodes are repelled. An advantage of this layout is that no order needs to be specified for either the categories or variables. The size of each node represents its frequency, while its colour is determined by the mode response category. Node labels show the name of the corresponding category and variable, with font size denoting entropy. The width of each edge is proportional to the overlap between the corresponding categories, as calculated by the Jaccard Index. Edges can be filtered by entropy to reduce clutter and home in on specific communities, aided by zooming and tooltips. With these interactive capabilities, CatNetVis can be used to explore dozens of variables and hundreds of categories.

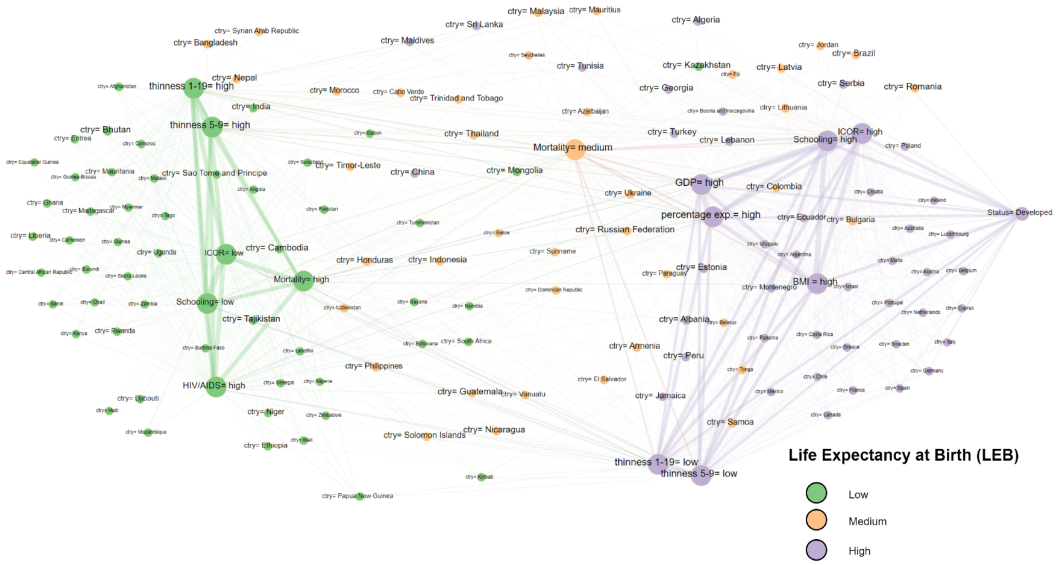

Figure 1.14: CatNetVis showing life expectancy data with a filter applied. There are two main communities, relating to under-developed countries (left) and developed countries (right).

1.4.6 Projection Techniques

Following the QuantViz approach mentioned in Section 1.3, techniques in the projection family map high-dimensional data into a low-dimensional space. The goal is to preserve relationships in the data, such as distances, similarities and associations between categories. At the heart of this approach are two key steps: a dimensionality reduction technique *transforms* categories into numbers, and a visualisation technique *represents* the result of this transformation. For a detailed review of projection techniques, see Johansson Fernstad (2011).

The most well-known dimensionality reduction techniques for categorical

data are *Correspondence Analysis* (*CA*; Greenacre, 2017) for two categorical variables, and *Multiple Correspondence Analysis* (*MCA*; Tenenhaus and Young, 1985) for greater numbers of variables. These dimensionality reduction techniques have many different names and variations.

Popular choices for visualising the results of Correspondence Analysis are *CA Maps* and *Biplots* Gabriel (1971), which are both types of scatterplots. Other visualisation techniques used for CA and MCA include:

- *Contribution Biplots* (Greenacre, 2013)
- *Moon Plots* (Bock, 2011)
- *Voronoi Diagrams* (Broeksema et al., 2013; Dennig et al., 2024)
- *Andrews Curves* (Rovan, 1994)
- *Dendrograms* (Beh and Lombardo, 2014)
- *Z-Plots* (Choulakian et al., 1994)
- *Chernoff Faces* (Beh and Lombardo, 2014)

While Correspondence Analysis and Multiple Correspondence Analysis are useful for capturing structure in high-dimensional categorical datasets, they have a number of drawbacks. Both techniques are difficult for non-experts to interpret, they do not display frequency-related information, or convey the reasons *why* items belong to particular clusters. Furthermore, CA and MCA quickly become cluttered when the number of categories increases, since the individual category labels are usually shown next to the points themselves. When there are large numbers of variables in MCA, it is also difficult to determine which categories belong to which variables. These plots result in overlapping labels when there are many categories, and are generally sensitive to outliers.

Some interactive tools combine both stages of the QuantViz process in a user-controlled way. MiDAVisT (Johansson Fernstad, 2009; Johansson and Johansson, 2010), shown in Figure 1.15, is one such approach. The figure has been chosen to show all views at once, thereby highlighting the interactive capabilities; the details of the categories and text are not important. This tool provides suggestions for numeric representations to the user, which they can adjust interactively. The user can then explore the results using a range of visualisation methods in multiple coordinated views, as well as algorithmic analyses, such as clustering and correlation analysis.

Figure 1.15: Multiple view environment within MiDAVisT, consisting of a scatterplot matrix (top left), Table Lens visualisation (top right) and parallel coordinates display (bottom).

1.5 Analysis Tasks for Categorical Data Visualisation

Before moving on to a general comparison of the techniques in our taxonomy and a discussion of avenues for future work, we provide an overview of nine kinds of analysis tasks associated with categorical data visualisation. An awareness of such tasks is helpful for designing, improving, evaluating and comparing categorical visualisation techniques. We note that not all of these categories are applicable to all techniques.

Overview tasks constitute a useful first step in any analysis of categorical data: these include determining the total (or selected) number of data items, variables and categories in a dataset, as well as inspecting the distribution of variable cardinalities. Users should be familiar with the basic structure and characteristics of a dataset before proceeding with their analysis.

Drawing inspiration from Unwin and Pilhoefer (2020), *missing value tasks* are concerned with obtaining a summary of missing values in the data, so that these can be dealt with appropriately. Missing value tasks for categorical data may relate to variables or data items (records). For example, a user may wish to summarise the number of missing values across all variables, or determine the number of data items that are incomplete, before filtering or removing

them. Missing value tasks may be seen as a subset of overview tasks, since they involve gaining a preliminary understanding of the structure of the data.

Identification tasks focus on contextualising individual data units. Examples include determining which variable a particular category belongs to, determining which categories belong to a particular variable, and identifying any response variables within the dataset. An example involving multiple variables is identifying which categories are present in a given combination of categories, such as a particular tile in a mosaic plot.

Frequency tasks, which were a key part of Johansson Fernstad and Fernstad's (2011) study, play a crucial role in categorical data analysis. These tasks involve determining, comparing and ranking the frequency of particular categories or combinations of categories. The tasks may be univariate (involving *marginal* frequencies of one or more variables) or multivariate (involving *cross-tabulation* of two or more variables). Examples of univariate tasks are inspecting the marginal distribution of each variable, determining the n-th most (or least) frequent category across the entire dataset and ranking all categories within a particular variable by frequency. Examples of multivariate tasks are comparing the joint frequencies of two or more combinations of categories, determining conditional frequencies of a target variable for each combination of explanatory variables, and identifying the number of empty combinations involving n variables.

Similarity tasks, also explicitly mentioned by Johansson Fernstad and Fernstad (2011), involve identifying structural patterns and clusters within the data. These tasks operate at both the category and variable levels. Clustering seeks to identify groups of items that are similar to each other and different to items belonging to other clusters. Examples of similarity tasks include identifying the n most similar categories within a given variable, finding clusters of combinations of categories, and identifying the n most similar variables to a given variable. In certain contexts, it may also be helpful to identify which variable is *least* similar to all others.

To support these tasks, various similarity measures can be used, such as the overlap similarity measure, Jaccard index, Cosine distance and mutual information (Boriah et al., 2008 discuss additional measures). The most effective approach for computing some of these measures involves converting variables into multidimensional binary attributes through one-hot encoding, then comparing the resultant vectors across variables.

Co-occurrence tasks combine elements of the previous two task types, investigating conditions under which two or more categories appear together

within the dataset. Examples include finding categories across a given set of variables that occur together at least $p\%$ of the time, and finding n categories from any other variables that a given category occurs with most.

Association tasks explore the relationships between variables or categories, aiming to determine if and how they are associated. Investigating category frequencies by themselves can be misleading if uncorrelated variables exist. Examples of association tasks include discerning global associations between variables, detecting individual associations between categories of different variables, and identifying one-way dependencies where one category nearly always occurs with another, but not *vice versa*. Several association measures are available for analysing categorical and ordinal data, including Cramer's V and the Goodman-Kruskal tau index (Goodman et al., 1979).

Deviation tasks involve determining the extent to which the observed data deviate from expected values. They can be helpful for identifying patterns and outliers in the data, and determining the lack-of-fit of a log-linear model. Common examples of deviations include Pearson residuals, Standardised residuals and Deviance residuals. Typical tasks are finding the combination with the smallest/largest deviation, finding the deviation of a given combination of categories and examining the distribution of deviations for all combinations involving n variables.

Finally, *data item tasks* are related to the records in a dataset, and are only applicable when a dataset contains individual identifiers (e.g., passenger names are included in some versions of the Titanic dataset). Example tasks include looking up a data item based on its identifier, comparing categories among two or more data items, and summarising category counts for a particular group of data items. Currently, only a few categorical visualisation techniques support analysis of individual data items.

1.6 Discussion and Future Work

We now consider general strengths and weaknesses of the visualisation families reviewed in this chapter, as well as possible avenues for future work in the area of categorical data visualisation. The different families of techniques in our taxonomy have different strengths and weaknesses, affecting their suitability for different contexts and analysis tasks. An overview of key points is provided in Table 1.2.

Table 1.2: Comparison of visualisation families.

Family	**Strengths**	**Weaknesses**
Size-encoding (e.g., bar charts, pie charts)	• Intuitive (no training required) • Bars support precise comparisons • Can be faceted to show extra variables • Useful for part-to-whole comparisons for a single variable	• Limited to few categories per variable or few variables • Wedges suffer from perceptual distortions • Linking becomes complicated with many variables
Space-filling (e.g., ParSets, mosaic plots)	• Area (spatial regions) well-suited to categorical data • Optimize the space used • Useful for emphasising a response variable • Relatively independent of number of data items	• Quickly become cluttered • Different orders vastly change the display / sensitive to ordering • Suffer from visual interference (e.g., line-crossings)
Glyphs (e.g., Chernoff faces, metaphoric glyphs)	• Visually emphasise items as individual objects • Can use semantically meaningful representations • Suitable for both dense and sparse structures	• Poor scalability (if using one glyph per item) • Usually requires carefully chosen variable-to-glyph mapping • Learning and memorisation can be cognitively demanding • Not all glyphs suitable for nominal variables
Table (e.g., Table Lens, Taggle)	• Utilise a familiar, spreadsheet-like layout • Fairly scalable in terms of both categories and variables • Direct representation of individual records • Well-suited for heterogeneous data • Pairwise matrices provide a compact overview	• Pairwise matrices limited to bivariate relationships • May confuse frequency in heatmap with similarity • Desired properties not always possible (many frequent combinations)

Continued on next page

Table 1.2 continued from previous page

Family	Strengths	Weaknesses
Projection (e.g., M/CA, biplots)	• Excel at similarity tasks • Can handle many variables • Useful for cluster analysis • Any visualisation technique for numeric data can be used	• Lack of frequency information • Not easily interpretable • Sensitive to outliers • Distortion (e.g., horseshoe) effects • Cluttered when there are lots of categories • First two dimensions may not capture sufficient variance

Regarding future work, the analysis tasks outlined in the previous section can be used to identify gaps in existing work, such as the lack of explicit consideration of missing values within most categorical visualisation techniques. There is also potential for visualising the results of a wider range of similarity and association measures for sets of two or more categorical variables.

A major limitation of the reviewed techniques is their general lack of scalability. Most techniques scale exponentially when a categorical variable is added, quickly leading to visual clutter and increased computation time. High-cardinality variables are also problematic, not least because channels like colour can only show about 6-8 categories effectively. On the other hand, the scalability of categorical visualisation techniques is relatively independent of the number of data items, except when these are displayed individually, as is the case for various table and glyph techniques. Some of the reviewed techniques accommodate only a small number of variables, while others support multiple variables but only consider pairwise relationships. Rarely does a technique enable visualisation of relationships between many variables and categories simultaneously. However, techniques like CatNetVis are promising recent developments. Even so, there remains a need for more powerful categorical visualisation techniques that make use of visual channels in perceptually efficient ways.

Crucially, the field would benefit from more comprehensive user studies that focus specifically on multivariate categorical data. Our review of the literature suggests that there have been few developments in this area since this gap was identified by Johansson Fernstad and Johansson (2011), apart from the study carried out by Hofmann and Vendettuoli (2013). The proposed task and technique classifications in this survey paper provide a framework for designing such studies: representative techniques from different groups in our taxonomy (Section 1.3) can be compared with respect to key analysis

tasks (Section 1.5), using datasets of varying complexity. Online user studies for multivariate categorical data could be facilitated by the ReVISit software framework (2023).[4].

Only a small proportion of the reviewed techniques have publicly available implementations that do not require programming skills and which allow users to analyse their own datasets. Few interactive tools are available for techniques that were proposed more than ten years ago (e.g., we could not find implementations for Nested Rings, Granular Representation or KVMaps). Even some more recent techniques suffer from this problem (e.g., the Heatmap Matrix and CatNetVis). User-friendly tools for other techniques, like Parallel Sets and ContingencyWheel++, were previously available but are no longer maintained, making them less accessible, or inaccessible, to non-computer scientists. The development of modern, code-free tools is needed to democratise access to these techniques.

Existing visualisation techniques offer significant potential for enhancement through increased interaction. For example, allowing flexible changes to data mapping can increase the readability of glyph-based techniques like Chernoff faces. Similarly, the ability to reorder variables is crucial for techniques where a hierarchical structure is imposed, such as Parallel Sets, mosaic plots, sunburst diagrams, since these changes can drastically alter the display. For such techniques, providing an interactive, separate view of the imposed tree structure could facilitate understanding and exploration of different configurations of the visualisation. This could be implemented as a classic tree diagram with drag-and-drop functionality. More generally, since it does not always make sense to incorporate all variables at once Theus (2008), it is beneficial to allow user-controlled inclusion and exclusion of individual variables. There should be flexibility to (re-)display variables that are not currently visible, unless the user has explicitly removed them from the dataset.

Related to this, of all the techniques reviewed, only MiDaVist and ViSta (spreadplots) appear to integrate multiple coordinated views. In general, combining different representations can highlight different aspects of the data, provided the display is not overly cluttered. In fact, we only identified two techniques that combine CatViz and QuantViz representations (Valero-Mora et al., 2003; Dennig et al., 2024)). Given that these two approaches are useful for different analysis tasks (Johansson Fernstad and Johansson, 2011), connecting them in visualisation systems offers potential to harness their relative strengths. For example, it would be interesting to be able to view Parallel Sets

[4]See `https://revisit.dev/`

and CA plots side-by-side, and to enable linked interactions between them. Even if plots cannot be shown side-by-side, due to lack of screen space, it is helpful to be able to switch between different representations while preserving selections

Furthermore, apart from tabular techniques such as Table Lens and glyph-based methods like Chernoff faces, few categorical visualisation techniques support the display or analysis of individual data items (in line with the data item tasks detailed in Section 1.5). Providing access to individual identifiers and any additional string-type (text) variables in the raw data enables users to address micro-questions about specific records. While it is possible to display limited text about each data item in area-proportional visualisations of categorical data (e.g., as demonstrated by Brath, 2018, p. 155), a more scalable solution involves displaying the text within a coordinated table view (following Liu et al. (2009)). For instance, clicking on different visual elements (e.g., a bar in a multivariate bar chart, a tile in a mosaic plot or a parallelogram in Parallel Sets) could highlight or isolate the corresponding records in the table view. This could be powerfully assisted by search functionality that targets the identifier. Many existing categorical visualisation techniques could be extended in this way.

We stated at the beginning of the chapter that, in our view, categorical visualisation techniques should support both nominal and ordinal data. Some QuantViz techniques, including several variants of Correspondence Analysis, take the order of categories into consideration (Beh and Lombardo, 2014). Surprisingly, however, we did not encounter any CatViz tools where ordinal variables were treated or displayed differently from nominal variables. For instance, it may be more appropriate to use greyscale for ordinal variables instead of colour, in accordance with perceptual guidelines (Mackinlay, 1986).

1.7 Postscript

This chapter has reviewed existing techniques for visualising categorical data. After explaining our scope and methodology, we introduced a two-level technique taxonomy, providing a foundation for understanding and comparing different approaches. Situated within the established CatViz/QuantViz framework, this taxonomy organises six distinct families: *size-encoding*, *space-filling, table, glyph, miscellaneous* (all frequency-based), and *projection* (quantification-based). We discussed prominent examples from each family, ranging in complexity from simple bar charts to much more sophisticated tools like CatNetVis

and MiDAVisT.

In Section 1.5, nine different kinds of analysis tasks for dealing with categorical data were proposed, from *overview tasks* to *frequency* and *association tasks*. This was followed by a summary of the general strengths and weaknesses of each family of techniques. Finally, we pinpointed areas for future research, emphasising the need for more scalable solutions, empirical user studies, code-free tools and enhanced interaction. We also advocated for better support for individual data items, as well as for handling ordinal variables alongside nominal ones. The remaining chapters in Part II (Chapters 4 & 5) present new and adapted techniques that seek to address some of these gaps, with a particular focus on improving scalability and interaction.

3.8 References

Agresti, A. (2012). *Categorical data analysis*. John Wiley & Sons, 3rd edition.

Agresti, A. (2019). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 3 edition.

Allen, T. (2002). Using and interpreting the Trilinear plot. *Chance*, 15(3):29–35.

Alsakran, J., Huang, X., Zhao, Y., Yang, J., and Fast, K. (2014). Using entropy-related measures in categorical data visualization. In *2014 IEEE Pacific Visualization Symposium*, pages 81–88. IEEE.

Alsallakh, B., Aigner, W., Miksch, S., and Gröller, M. E. (2012). Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2849–2858.

Alsallakh, B., Gröller, M. E., Miksch, S., and Suntinger, M. (2011). Contingency wheel: Visual analysis of large contingency tables. In *EuroVA 2011*, pages 53–56. Eurographics.

Alsallakh, B., Micallef, L., Aigner, W., Hauser, H., Miksch, S., and Rodgers, P. (2016). The state-of-the-art of set visualization. In *Computer Graphics Forum*, volume 35, pages 234–260. Wiley Online Library.

Ancker, J. S., Weber, E. U., and Kukafka, R. (2011). Effect of arrangement of stick figures on estimates of proportion in risk graphics. *Medical Decision Making*, 31(1):143–150.

Balzer, M. and Deussen, O. (2005). Voronoi treemaps. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 49–56. IEEE.

Beck, F., Koch, S., and Weiskopf, D. (2015). Visual analysis and dissemina-

tion of scientific literature collections with survis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):180–189.

Becker, R. A., Cleveland, W. S., and Shyu, M.-J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155.

Beddow, J. (1990). Shape coding of multidimensional data on a microcomputer display. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pages 238–246. IEEE.

Beh, E. J. and Lombardo, R., editors (2014). *Correspondence Analysis*. Wiley Series in Probability and Statistics. John Wiley Sons, Ltd.

Bock, T. (2011). Improving the display of correspondence analysis using moon plots. *International Journal of Market Research*, 53(3):307–326.

Booshehrian, M., Möller, T., Peterman, R. M., and Munzner, T. (2011). Vismon: facilitating risk assessment and decision making in fisheries management. Technical report, Tech. Rep. TR 2011-05. School of Computing Science, Simon Fraser University ....

Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramee, R. S., Hauser, H., Ward, M. O., and Chen, M. (2013). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In Sbert, M. and Szirmay-Kalos, L., editors, *Eurographics 2013 - State of the Art Reports*, pages 39–63. Eurographics.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 243–254. SIAM.

Brath, R. (2018). *Text in visualization: extending the visualization design space*. PhD thesis, London South Bank University.

Brodbeck, D. and Girardin, L. (2019). High-d. `https://www.high-d.com/`. [Online; accessed 13-October-2023].

Broeksema, B., Telea, A. C., and Baudel, T. (2013). Visual analysis of multi-dimensional categorical data sets. In *Computer Graphics Forum*, volume 32, pages 158–169. Wiley Online Library.

Cantu, A., Micó-Amigo, M. E., Del Din, S., and Johansson Fernstad, S. (2023). Parallel assemblies plot, a visualization tool to explore categorical and quantitative data: application to digital mobility outcomes. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, pages 21–30. IEEE.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

Cherry, W. and Oldford, R. (2003). Picturing probability: The poverty of venn diagrams, the richness of eikosograms.

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). Cramér-von mises statistics for discrete distributions. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 125–137.

Chung, D. H., Legg, P. A., Parry, M. L., Bown, R., Griffiths, I. W., Laramee, R. S., and Chen, M. (2015). Glyph sorting: Interactive visualization for multi-dimensional data. *Information Visualization*, 14(1):76–90.

Cibulková, J. and Kupková, B. (2022). Review of visualization methods for categorical data in cluster analysis. *Statistika: Statistics & Economy Journal*, 102(4):396–408.

Clark, J. (2006). Multi-level pie charts. `https://www.neoformix.com/2006/MultiLevelPieChart.html`.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, 38(4):270–280.

Coekin, J. (1969). A versatile presentation of parameters for rapid recognition of total state. In *Proceedings International Symposium on Man-Machine Systems, IEEE Conference Record*, volume 69.

Cohen, A. (1980). On the graphical display of the significant components in two-way contingency tables. *Communications in Statistics-Theory and Methods*, 9(10):1025–1041.

Davies, J. (2012). Parallel Sets. `https://www.jasondavies.com/parallel-sets/`. Accessed Jaunary 12, 2024.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3).

De Soete, G. and Do Corte, W. (1985). On the perceptual salience of features of Chernoff faces for representing multivariate data. *Applied psychological measurement*, 9(3):275–280.

Dennig, F. L., Fischer, M. T., Blumenschein, M., Fuchs, J., Keim, D. A., and Dimara, E. (2021). Parsetgnostics: Quality metrics for parallel sets. In *Computer Graphics Forum*, volume 40, pages 375–386. Wiley Online Library.

Dennig, F. L., Joos, L., Paetzold, P., Blumberg, D., Deussen, O., Keim, D. A., and Fischer, M. T. (2024). Toward the categorical data map. *Preprint*.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pages 194–202. Elsevier.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hof-

mann, H., and Wickham, H. (2013). The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91.

Feiner, S. K. and Beshers, C. (1990). Worlds within worlds: Metaphors for exploring n-dimensional virtual worlds. In *Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, SIGGRAPH, pages 76–83.

Fienberg, S. E. (1975). Perspective canada as a social report. *Social Indicators Research*, 2:153–174.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65(330):694–701.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200.

Friendly, M. (1995). A fourfold display for 2 by 2 by k tables. Technical report, Technical Report 217, Psychology Department, York University.

Friendly, M. (1998). Conceptual models for visualizing contingency table data. In *Visualization of categorical data*, pages 17–I. Elsevier.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395.

Friendly, M. (2006). A brief history of data visualization. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Computational Statistics: Data Visualization*, volume III. Springer-Verlag, Heidelberg.

Friendly, M. and Meyer, D. (2015). *Discrete data analysis with R: visualization and modeling techniques for categorical and count data*, volume 120. CRC Press.

Furmanova, K., Gratzl, S., Stitz, H., Zichner, T., Jaresova, M., Lex, A., and Streit, M. (2020). Taggle: Combining overview and details in tabular data visualizations. *Information Visualization*, 19(2):114–136.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Goodman, L. A., Kruskal, W. H., Goodman, L. A., and Kruskal, W. H. (1979). *Measures of association for cross classifications*. Springer.

Greenacre, M. (2013). Contribution biplots. *Journal of Computational and Graphical Statistics*, 22(1):107–122.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *Computer science and statistics: Proceedings of the 13th symposium on the*

interface, pages 268–273. Springer.

Hoffman, P. E. and Grinstein, G. G. (2001). A survey of visualizations for high-dimensional data mining. In *Information visualization in data mining and knowledge discovery*, pages 47–82.

Hofmann, H. (2001). Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, 10(4):628–640.

Hofmann, H. (2006). *Multivariate Categorical Data — Mosaic Plots*, pages 105–124. Springer New York, New York, NY.

Hofmann, H., Siebes, A. P., and Wilhelm, A. F. (2000). Visualizing association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 227–235.

Hofmann, H. and Vendettuoli, M. (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE transactions on visualization and computer graphics*, 19(12):2297–2305.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Huh, M. Y. (2004). Line mosaic plot: Algorithm and implementation. In *COMPSTAT 2004—Proceedings in Computational Statistics: 16th Symposium Held in Prague, Czech Republic, 2004*, Other Conference, pages 277–285. Springer.

Humayoun, S. R., Bhambri, K., and AlTarawneh, R. (2018). Bid-chord: an extended chord diagram for showing relations between bi-categorical dimensional data. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, pages 1–3.

Hummel, J. (1996). Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11(1):23–33.

Im, J.-F., McGuffin, M. J., and Leung, R. (2013). GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2606–2614.

Indratmo, Howorko, L., Boedianto, J. M., and Daniel, B. (2018). The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons. *Visual Informatics*, 2(3):155–165.

Jain, N. and Warnes, G. R. (2006). Balloon plot. *The Newsletter of the R Project Volume 6/2, May 2006*, 6:35.

Jin, Y., de Jong, T. J., Tennekes, M., and Chen, M. (2023). Radial Icicle Tree (RIT): Node separation and area constancy. *arXiv preprint*

arXiv:2307.10481.

Johansson, S. and Johansson, J. (2010). Visual analysis of mixed data sets using interactive quantification. *ACM SIGKDD Explorations Newsletter*, 11(2):29–38.

Johansson Fernstad, S. (2009). Visual exploration of categorical and mixed data sets. In *Proceedings of the acm sigkdd workshop on visual analytics and knowledge discovery: Integrating automated analysis with interactive exploration*, SIGKDD, pages 21–29.

Johansson Fernstad, S. (2011). *Algorithmically guided information visualization: Explorative approaches for high dimensional, mixed and categorical data*. PhD thesis, Linköping University Electronic Press.

Johansson Fernstad, S. and Johansson, J. (2011). A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation*, pages 80–89. IEEE.

Johnson, B. S. (1993). *Treemaps: Visualizing hierarchical and categorical data*. PhD thesis, University of Maryland, College Park.

Karduni, A., Wesslen, R., Cho, I., and Dou, W. (2020). Du bois wrapped bar chart: Visualizing categorical data with disproportionate values. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

Kolatch, E. and Weinstein, B. (2001). CatTrees: Dynamic visualization of categorical data using treemaps.

Kosara, R. (2007). Autism Diagnosis Accuracy - Visualization Redesign. `https://eagereyes.org/blog/2007/autism-diagnosis-accuracy`. [Online; accessed 13-October-2023].

Kosara, R. (2008). Treemaps. `https://eagereyes.org/blog/2008/treemaps`.

Kosara, R. (2010). Turning a table into a tree: Growing parallel sets into a purposeful project. In Steele, J. and Iliinsky, N., editors, *Beautiful Visualization*, pages 193–204. O'Reilly Media.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568.

Kruskal, J. B. and Landwehr, J. M. (1983). Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for

comparative genomics. *Genome research*, 19(9):1639–1645.

LeBlanc, J., Ward, M. O., and Wittels, N. (1990). Exploring n-dimensional databases. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pages 230–237.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph*, 20(12):1983–1992.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., and Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3):1249–1268.

Liu, Z., Stasko, J., and Sullivan, T. (2009). Selltrend: Inter-attribute visual analysis of temporal transaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1025–1032.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions On Graphics (Tog)*, 5(2):110–141.

Maguire, E., Rocca-Serra, P., Sansone, S.-A., Davies, J., and Chen, M. (2012). Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2603–2612.

May, T. (2007). Working with patterns in large multivariate datasets-karnaugh-veitch-maps revisited. In *2007 11th International Conference Information Visualization (IV'07)*, pages 277–285. IEEE.

May, T., Davey, J., and Kohlhammer, J. (2010). Combining details of the chi-square goodness-of-fit test with multivariate data visualization. In *EuroVAST@ EuroVis*, EuroVis, pages 45–50. Eurographics.

Meyer, D., Zeileis, A., and Hornik, K. (2008). Visualizing contingency tables. *Handbook of Data Visualization*, pages 589–616.

Micallef, L., Dragicevic, P., and Fekete, J.-D. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE transactions on visualization and computer graphics*, 18(12):2536–2545.

Miyamoto, A., Allacker, K., and De Troyer, F. (2022). Visual tool for sustainable buildings: A design approach with various data visualisation techniques. *Journal of Building Engineering*, 56:104741.

Nightingale, F. (1857). *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army.* Private Publication, London.

Paetzold, P., Kehlbeck, R., Strobelt, H., Xue, Y., Storandt, S., and Deussen, O. (2023). Recteuler: Visualizing intersecting sets using rectangles. In

Computer Graphics Forum, volume 42 of *Computer Graphics Forum*, pages 87–98. Wiley Online Library.

Park, D., Kim, S.-H., and Elmqvist, N. (2023). Gatherplot: A non-overlapping scatterplot. *arXiv preprint arXiv:2301.10843*.

Pilhöfer, A. and Unwin, A. (2013). New approaches in visualization of categorical data: R package extracat. *Journal of Statistical Software*, 53:1–25.

Playfair, W. (1786). *The commercial and political atlas*. Wallis.

Playfair, W. (1801). *The statistical breviary*. Wallis.

Ramos, B. N., Maçãs, C., Lourenço, N., and Polisciuc, E. (2023). Towards contextual glyph design: Visualizing hearing screenings. In *2023 27th International Conference Information Visualisation (IV)*, pages 96–102. IEEE.

Rao, R. and Card, S. K. (1994). The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322.

Reza, R. M. and Watson, B. A. (2019). Hi-d maps: An interactive visualization technique for multi-dimensional categorical data. In *2019 IEEE visualization conference (VIS)*, pages 216–220. IEEE.

Riedwyl, H. and Schüpbach, M. (1994). Parquet diagram to plot contingency tables. In *Advances in Statistical Software, F. Faulbaum (Ed.).*, pages 293–299. Gustav Fischer.

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muehrcke, P. C. (1984). *Elements of Cartography*. Wiley, New York.

Rocha, M. M. N. and da Silva, C. G. (2018). Heatmap matrix: a multidimensional data visualization technique. In *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*.

Rocha, M. M. N. and da Silva, C. G. (2022). Heatmap matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis. In *IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing 2022 (part of MCCSIS 2022)*, pages 11–18. MCCSIS.

Rovan, J. (1994). Visualizing solutions in more than two dimensions. *Correspondence analysis in the social sciences*, pages 210–229.

Rzeszotarski, J. M. and Kittur, A. (2014). Kinetica: Naturalistic multi-touch data visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, SIGCHI, pages 897–906.

Sanderson, D. and Peacock, D. (2020). Making rose diagrams fit-for-purpose. *Earth-Science Reviews*, 201:103055.

Schlimmer, J. (1987). Mushroom. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5959T.

Schmidt, M. (2006). Der einsatz von sankey-diagrammen im stoffstrommanagement. Technical report, Beiträge der Hochschule Pforzheim.

Schonlau, M. (2003). Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*. ASA.

Schonlau, M. (2024). Hammock plots: Visualizing categorical and numerical variables. *Journal of Computational and Graphical Statistics*, 0(0):1–16.

Schulz, H.-J., Hadlak, S., and Schumann, H. (2010). The design space of implicit hierarchy visualization: A survey. *IEEE transactions on visualization and computer graphics*, 17(4):393–411.

Shiraishi, K., Misue, K., and Tanaka, J. (2009). A tool for analyzing categorical data visually with granular representation. In *Human Interface and the Management of Information. Information and Interaction: Symposium on Human Interface 2009, Held as part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part II*, Other Conference, pages 342–351. Springer.

Siirtola, H. (2005). The effect of data-relatedness in interactive glyphs. In *Ninth International Conference on Information Visualisation (IV'05)*, pages 869–876. IEEE.

Siirtola, H. (2014). Bars, pies, doughnuts & tables–visualization of proportions. In *Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI 2014) 28*, pages 240–245. BCS.

Skau, D. and Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Computer Graphics Forum*, volume 35, pages 121–130. Wiley Online Library.

Stasko, J. and Zhang, E. (2000). Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 57–65. IEEE.

Streit, M. and Gehlenborg, N. (2014). Bar charts and box plots: Creating a simple yet effective plot requires an understanding of data and task. *Nature Methods*, 11(2):117.

Symanzik, J., Friendly, M., and Onder, O. (2019). The unsinkable titanic data.

Sönning, L. and Schützler, O. (2023). Data visualization in corpus linguistics: Critical reflections and future directions. In Sönning, L. and Schützler, O., editors, *Data Visualization in Corpus Linguistics: Critical Reflections and*

Future Directions, number 22 in Studies in Variation, Contacts and Change in English. VARIENG, Helsinki.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison Wesley.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119.

Tennekes, M. and de Jonge, E. (2013). On the exploration of high cardinality categorical data.

Tennekes, M., de Jonge, E., Daas, P. J., et al. (2013). Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, 11(1):43–58.

Thane, M., Blum, K. M., and Lehmann, D. J. (2023). CatNetVis: Semantic Visual Exploration of Categorical High-Dimensional Data with Force-Directed Graph Layouts. In Hoellt, T., Aigner, W., and Wang, B., editors, *EuroVis 2023 - Short Papers*. The Eurographics Association.

Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7:1–9.

Theus, M. (2008). High-dimensional data visualization. In Chen, C.-h., Härdle, W. K., and Unwin, A., editors, *Handbook of Data Visualization*, pages 151–178. Springer, Berlin.

Tominski, C. and Schumann, H. (2020). *Interactive visual data analysis*. AK Peters/CRC Press.

Tufte, E. R. and Graves-Morris, P. R. (1983). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Unwin, A. and Pilhoefer, A. (2020). Visna—visualising multivariate missing values. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*.

Upton, G. J. (2000). Cobweb diagrams for multiway contingency tables. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(1):79–85.

Valero-Mora, P. M., Young, F. W., and Friendly, M. (2003). Visualizing categorical data in vista. *Computational Statistics Data Analysis*, 43(4):495–508. Data Visualization.

VanderPlas, S., Ge, Y., Unwin, A., and Hofmann, H. (2023). Penguins go parallel: a grammar of graphics framework for generalized parallel coordinate plots. *Journal of Computational and Graphical Statistics*, pages 1–16.

Vivacqua, A. S. and Garcia, A. C. B. (2008). Nrv: Using nested rings to interact with categorical data. In *Proceedings of the IADIS International*

Conference on Interfaces and Human Computer Interaction, IADIS, pages 85–92.

Wang, S., Mondal, D., Sadri, S., Roy, C. K., Famiglietti, J. S., and Schneider, K. A. (2022). Set-stat-map: Extending parallel sets for visualizing mixed data. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pages 151–160. IEEE.

Wang, W., Wang, H., Dai, G., and Wang, H. (2006). Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 517–520.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210.

Wickham, H. and Hofmann, H. (2011). Product plots. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2223–2230.

Wilkinson, L. (2006). Revising the pareto chart. *The American Statistician*, 60(4):332–334.

Wolf, H. P. (2021). iconplot: Icon plots for visualization of contingency tables. https://rdrr.io/cran/aplpack/man/iconplot.html.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, 16(3):507–525.

Zhang, C., Chen, Y., Yang, J., and Yin, Z. (2019). An association rule based approach to reducing visual clutter in parallel sets. *Visual Informatics*, 3(1):48–57.