MultiCat: A Visualisation Technique for Multidimensional Categorical Data

Suggested citation: Trye, D., Apperley, M., & Bainbridge, D. (2024). *MultiCat: A visualisation technique for multidimensional categorical data.* [Unpublished manuscript]. University of Waikato. `https://dgt12.github.io/files/multicat.pdf`

Abstract

Exploring relationships among several variables is an important analysis task when dealing with multidimensional categorical data. A key challenge in visualising such data lies in ensuring that each variable and its categories can be clearly distinguished, especially when these grow in number and complexity. This paper presents MultiCat, an interactive visualisation technique for analysing categorical datasets comprising roughly 3–20 variables. MultiCat uses a familiar, spreadsheet-like layout to represent both nominal and ordinal variables in frequency form. It incorporates several interactive features, including column-wise sorting, dynamic queries and filtering, and allows rapid calculation of *a priori* and conditional probabilities. MultiCat offers several advantages over existing techniques, including: (1) enhanced clarity in extracting high-dimensional relationships and comparing their frequencies; (2) a non-hierarchical default layout that promotes user-driven analysis; and (3) a structured visual overview of the relative contribution of each category. We validate MultiCat by reporting on the promising results and outcomes of a small-scale usability study. A prototype of MultiCat is available at `https://dgt12.github.io/multicat/`.

1.1 Introduction

Understanding relationships among variables is an important analysis task, whether those variables are categorical or continuous. Categorical variables are frequently encountered in real-world datasets, ranging across such varied domains as behavioural and social sciences, public health, biomedical science, education and marketing (Agresti, 2012). For example, categories can be used to represent patient treatment outcomes (no improvement, some improvement, marked improvement), survey responses (strongly disagree to strongly agree) or customer brand preferences (Brand X, Brand Y, Brand Z). However, despite their prevalence, few visualisation techniques support the analysis of more than three categorical variables at the same time.

A significant challenge in visualising categorical data lies in ensuring that each variable and its categories can be clearly distinguished, regardless of how many there are. In addition, given that nominal variables do not have an intrinsic order, it is difficult to know how best to arrange them. Existing methods for visualising multidimensional categorical data either do not scale well (Hartigan and Kleiner, 1981; Greenacre, 2017; Tenenhaus and Young, 1985; Reza and Watson, 2019) or fail to consider relationships among all variables simultaneously. They typically break down the data into more restricted views, such as pairwise relationships (Rocha and da Silva, 2018; Trye et al., 2023; Friendly, 1999; Im et al., 2013; Greenacre, 2017; Tenenhaus and Young, 1985), or impose a hierarchy of variables (Kosara et al., 2006; Hartigan and Kleiner, 1981; Kolatch and Weinstein, 2001), which affects what insights can be seen. Moreover, these techniques often lack code-free, user-friendly interfaces, limiting their accessibility to a broader audience.

Recognising this gap, we adopt a technique-driven approach (Sedlmair et al., 2012) to design and validate MultiCat, a novel method for visualising multidimensional categorical data. MultiCat allows users to generate new insights and hypotheses about the interplay of categories across as many as 20 variables, by focusing on both individual categories and their higher-dimensional relationships. This is accomplished by using a tabular visualisation of the data in frequency form, coupled with a sidebar that comprises multiple linked bar charts. MultiCat also serves as an interactive probability calculator, helping users to compute and reason about a wide range of *a priori* and conditional probabilities. We validate MultiCat with a small-scale usability study, from which we have used feedback and observations to improve our prototype. MultiCat is generalisable across datasets and domains, and is therefore of interest to anyone who works with categorical data, including social scientists, business analysts and marketing experts.

Throughout the paper, we use the **Titanic dataset** (Dawson, 1995) as our primary example. This dataset, compiled by Robert Dawson in 1995, details socio-historical information about the people aboard the RMS *Titanic* when it tragically sank in 1912. It has been visualised extensively in the context of categorical data analysis (Symanzik et al., 2019). The dataset contains 2,201 observations (people) and comprises four categorical variables: Class (first, second, third and crew), Sex (male, female), Age (child, adult) and Fate (survived, died). This last variable has been renamed from Survived (yes, no) to provide more descriptive category names. Given its modest size and absence of missing values, the Titanic dataset is well-suited to introducing the MultiCat technique. At the same time, its widespread use enables a direct comparison with other methods (see, for instance, Figures 1.1-1.3). When describing MultiCat, we emphasise design features that make it suitable for handling more complex datasets, drawing on other examples where necessary.

The structure of the paper is as follows: We begin by stating contributions, discussing key terminology and surveying related work. Based on the capabilities and limitations of existing techniques, we outline a set of design requirements that informed the development of MultiCat. We then introduce the MultiCat technique by focusing on its spreadsheet view and sidebar, before comparing it with an earlier design. Next, we delve into MultiCat's interactive features, which range from dynamic queries to sorting and filtering. Implementation details of our prototype are provided, followed by a discussion of scalability constraints. We then describe the methodology and results of a user study aimed at identifying usability issues and gathering general feedback. A direct comparison with two other techniques is given, highlighting MultiCat's unique advantages and areas for improvement. Following this, we propose a series of extensions and enhancements for MultiCat. The paper concludes with a summary of the contributions of our research and opportunities for future work.

1.1.1 Contributions

This paper makes the following contributions:

1. The design and implementation of MultiCat, a novel visualisation technique for analysing multidimensional categorical data.
2. A small-scale usability study that sheds light on the value of this technique and highlights opportunities for further improvement.

1.1.2 Terminology

Since a variety of terms are used in the literature in relation to categorical data, we detail our adopted usage here. The term **multidimensional** is used throughout the paper to refer specifically to three or more categorical variables. We primarily refer to each entity in a dataset as a **data item**, rather than as a "record", "case" or "observation". Each discrete set of values is described as a **variable**, rather than as an "attribute" or "dimension", and the values themselves are designated as **categories** rather than "levels" or "classes". We consider a categorical variable to be either **nominal** (unordered) or **ordinal** (categories with a natural ordering), and believe it is important for a categorical visualisation tool to accommodate both types. The term **colour** is used throughout the paper to refer specifically to "hue". Finally, **cardinality** denotes the number of categories belonging to some variable, such that a high-cardinality variable has many (10 or more) categories.

1.2 Related work

We discuss related work in the context of prominent approaches and techniques for visualising multidimensional categorical data, before outlining relevant connections to interactive tables and hypergraphs.

1.2.1 Multidimensional categorical data

Over the past few decades, a number of techniques for visualising multidimensional categorical data have emerged, yet their adoption has not always been widespread (Theus, 2012). Most of these techniques are derived from contingency tables (Alsallakh et al., 2012), either by representing the cell counts directly or projecting categories into a two-dimensional plane. Following previous work (Johansson Fernstad and Johansson, 2011), we refer to these two approaches as "CatViz" and "QuantViz" methods, respectively. In general, CatViz methods are "lossless" (Dimara et al., 2017) and more effective for frequency-based tasks, whereas QuantViz methods are "lossy" and better suited to similarity-based tasks (Johansson Fernstad and Johansson, 2011). Due to limitations of space, we cover only three established techniques here, focusing on those which are most commonly cited in the literature. A far more comprehensive database of relevant techniques is available at `https://cat-vis.github.io`.

1.2.2 Interactive Mosaic Plots

Mosaic Plots (Hartigan and Kleiner, 1981) have been described as the “Swiss Army knife” of categorical data displays (Theus, 2012). These plots fall under the CatViz umbrella and are created by recursively subdividing variables along alternate axes, forming area-proportional tiles. If the tiles are neatly aligned, this means the variables are independent (Friendly, 1999). Residual-based shading of tiles is sometimes also used to visualise loglinear models (Friendly, 1994) and statistical significance of test results (Zeileis et al., 2007). Interactive Mosaic Plots are available in a variety of tools, including Mondrian (Theus, 2002), ViSta (Young and Bann, 1996) and MANET (Unwin et al., 1996), greatly enhancing their exploratory power. For instance, Mondrian allows users to switch between multiple variants of Mosaic Plots (Theus, 2012), add, remove or rotate variables, select different regions and access tooltips for each tile. Moreover, users can probe complex relationships by querying the data via linked bar charts, as shown in Figure 1.1. A major limitation of Mosaic Plots, however, is that they become increasingly difficult to read when displaying more than a handful of variables and/or categories. This leads to an increase in empty combinations and skewed tiles (Hofmann, 2006), exacerbated by low-frequency categories.

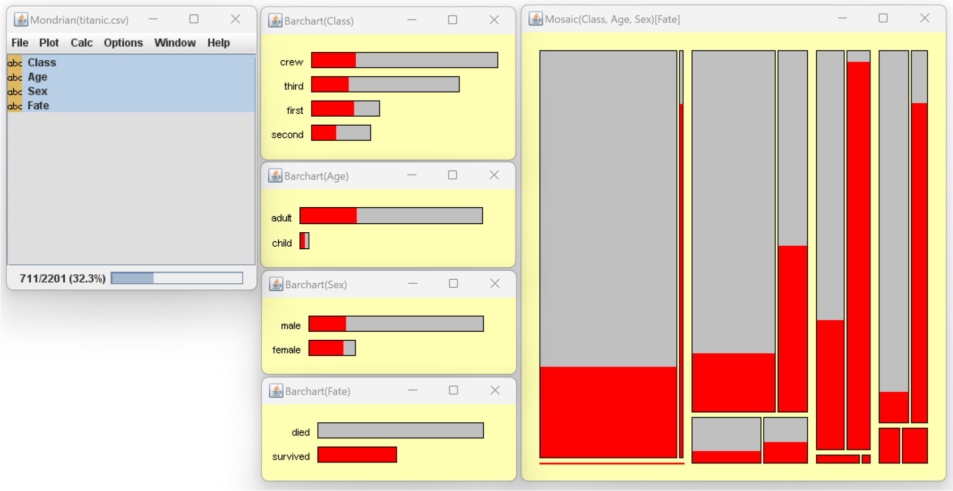


Figure 1.1: The Mondrian interface showing a Mosaic Plot with linked bar charts for the Titanic dataset (Dawson, 1995). Survivors are highlighted in red.

1.2.3 Parallel Sets

Perhaps the most scalable technique for visualising multidimensional categorical data is Parallel Sets (Kosara et al., 2006), reminiscent of Sankey Diagrams (Schmidt, 2006). Another area-proportional, CatViz technique, this method represents variables in stacked "tiers" of equal width. Associations between subsets are then shown using shaded parallelograms connecting adjacent tiers; see Figure 1.2. Parallel Sets visualisations are capable of handling 10–15 categorical variables in an interactive environment (Kosara et al., 2006), and 20–30 categories in total. While Parallel Sets supports rich interaction, including flexible reordering of variables and categories, it invariably suffers from line crossings and perceptual distortions (Hofmann and Vendettuoli, 2013). These are exacerbated by the hierarchical nature of the display. Furthermore, changing the aspect ratio of the visualisation can alter the appearance of the parallelograms, yielding skewed results. These disadvantages are partially addressed by Common Angle Plots (Hofmann and Vendettuoli, 2013) and Hammock Plots (Schonlau, 2003) but, in all cases, the order in which variables are plotted can drastically change the visualisation. Various quality metrics for evaluating Parallel Sets have been proposed, with a view to reducing visual clutter (Dennig et al., 2021; Zhang et al., 2019).

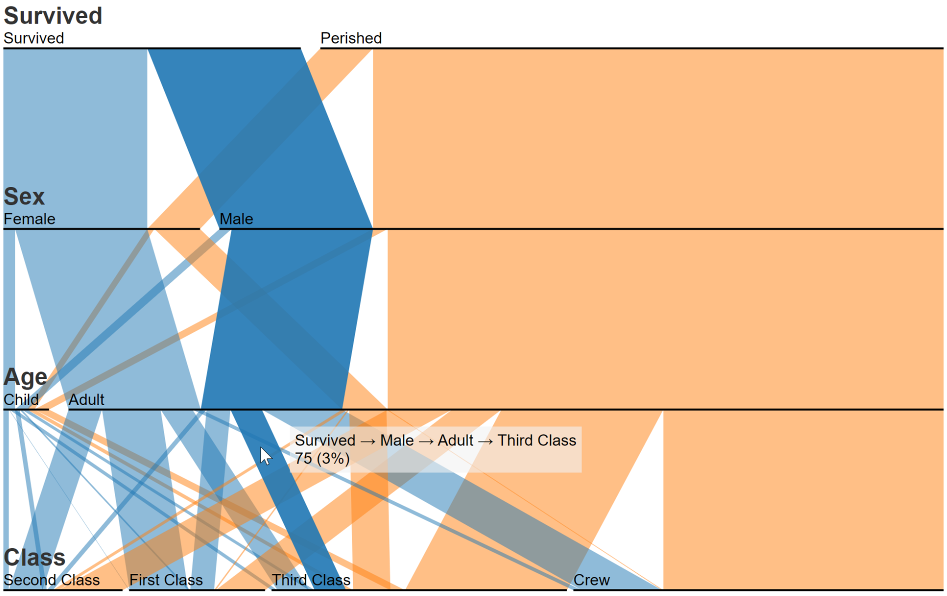


Figure 1.2: A Parallel Sets visualisation of the Titanic dataset (Davies, 2012).

Although an academic prototype was developed for Parallel Sets, it is no longer maintained and does not appear to run on modern machines (par, 2009). However, Parallel Sets has been implemented as a reusable D3.js (Bostock et al., 2011) chart, together with the most important interactive features (Davies, 2012). Related visualisation techniques, such as (Categorical) Treemaps (Kolatch and Weinstein, 2001), are also available in code-free tools like RAWGraphs (Mauri et al., 2017). Additionally, the R package *ggparallel* (Hofmann and Vendettuoli, 2013) creates static, pairwise visualisations of Parallel Sets, Hammock Plots and Common Angle Plots. Nevertheless, these alternatives do not offer the full range of features described in the original Parallel Sets papers (Bendix et al., 2005; Kosara et al., 2006), such as the ability to view histograms, merge categories or select multiple parallelograms to visualise the corresponding proportion of data. This exemplifies a broader issue endemic to the field: the divide between theoretical innovation and practical application of novel visualisation techniques.

1.2.4 Correspondence Analysis

Correspondence Analysis (CA) (Greenacre, 2017) is a widely used QuantViz method that shows associations in a two-way contingency table. The row and column categories in a table are depicted as points on a graph whose positions indicate associations between categories.

Multiple Correspondence Analysis (MCA) (Tenenhaus and Young, 1985) extends this principle to n-way tables, accommodating analyses involving more than two variables (see Figure 1.3). While MCA provides a broader scope than CA, it still focuses on pairwise relationships. MCA typically provides a visual representation of the so-called "Burt Matrix", which encodes the joint, bivariate relations between every pair of variables in a dataset (Friendly and Meyer, 2015).

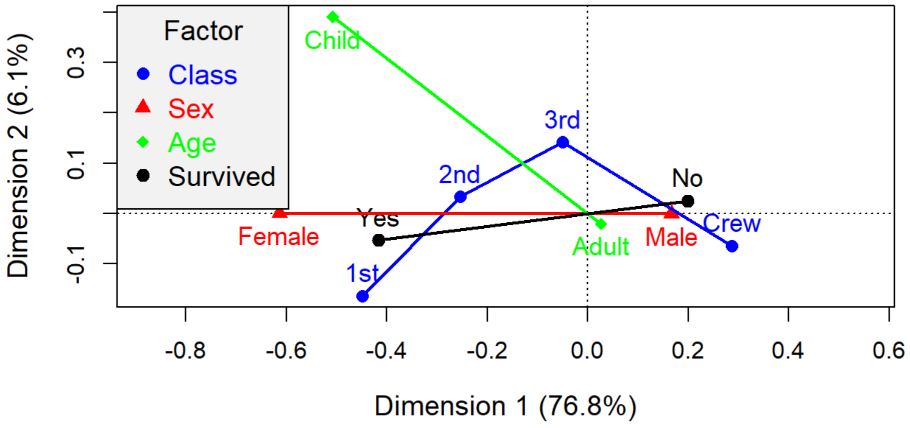


Figure 1.3: MCA Plot of the Titanic dataset (Friendly and Meyer, 2015).

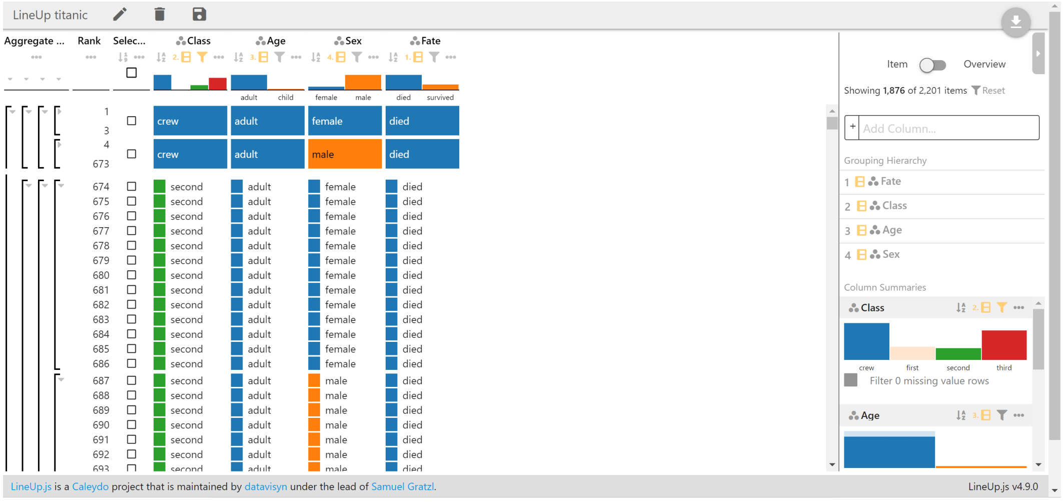
While CA and MCA are useful for capturing structure in high-dimensional categorical datasets, they have a number of drawbacks. Both techniques are difficult for non-experts to interpret, they do not display frequency-related information, or convey the reasons *why* items belong to particular clusters. 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.

Motivated by the limitations of these existing solutions, we adopt a technique-driven approach in this work to design and validate a novel method for exploring and analysing multidimensional categorical data.

1.2.5 Connection to tabular data

Multidimensional categorical data lends itself to tabular representations. Tabular visualisation techniques employ a spreadsheet-like layout, where rows correspond to individual data items and columns correspond to variables. These techniques are geared towards understanding the properties of items by considering all variables simultaneously. Cells use visual channels such as position, length and colour to enhance readability and facilitate exploration of higher-level trends. The two tabular visualisation techniques most closely related to MultiCat are Taggle (Furmanova et al., 2020) and TableLens (Rao and Card, 1994), both of which support heterogeneous data (i.e., both continuous and categorical variables). However, while very powerful, these techniques are not optimised for categorical data, as is evident in Figure 1.4. TableLens, for instance, does not support aggregation, limiting its scalability. Taggle, while offering aggregation and a height-proportional layout that reflects frequencies in its overview mode, does not provide the compactness of MultiCat's aligned bar chart encoding. Furthermore, Taggle can be confusing to navigate when visualising purely categorical data, as many of its features were not intended for such data. MultiCat provides a more focused analysis by removing extraneous features and streamlining its workflow for categorical data.

It has been convincingly argued that interactive tables are an important visualisation technique in their own right (Bartram et al., 2021). Spreadsheet applications like Microsoft Excel and Google Sheets play a critical role in helping users to make sense of data, incorporating powerful features like sorting and filtering. However, at the same time, these applications lack custom interaction, provide limited support for visual encoding of cells, make assumptions about how the data should be handed (e.g., sorting categorical values alphabetically rather than by frequency) and do not offer multiple coordinated



LineUp titanic

Aggregate ...	Rank	Selec...	Class	Age	Sex	Fate
	1 3	□	crew	adult	female	died
	4 673	□	crew	adult	male	died
	674	□	second	adult	female	died
	675	□	second	adult	female	died
	676	□	second	adult	female	died
	677	□	second	adult	female	died
	678	□	second	adult	female	died
	679	□	second	adult	female	died
	680	□	second	adult	female	died
	681	□	second	adult	female	died
	682	□	second	adult	female	died
	683	□	second	adult	female	died
	684	□	second	adult	female	died
	685	□	second	adult	female	died
	686	□	second	adult	female	died
	687	□	second	adult	male	died
	688	□	second	adult	male	died
	689	□	second	adult	male	died
	690	□	second	adult	male	died
	691	□	second	adult	male	died
	692	□	second	adult	male	died
	693	□	second	adult	male	died

Item Overview

Showing **1,876** of 2,201 items Reset

Add Column...

Grouping Hierarchy

1 Fate

2 Class

3 Age

4 Sex

Column Summaries

Filter 0 missing value rows

LineUp.js is a Caleydo project that is maintained by datavisyn under the lead of Samuel Gratzl.

LineUp.js v4.9.0

Figure 1.4: Taggle is not optimised for purely categorical data.

views (Gratzl et al., 2013). MultiCat seeks to overcome these limitations from a categorical data perspective, while preserving the essence and functionality of an interactive table.

1.2.6 Connection to hypergraphs

Finally, we note that multidimensional categorical data can be accurately represented in a *hypergraph* (Fischer et al., 2021). This structure is an extension of a traditional graph: the key difference is that its edges, termed *hyperedges*, can connect any number of vertices. A hyperedge is therefore equivalent to a set. Our initial design for MultiCat, depicted in Figure 1.9, was inspired by PAOHVis (Valdivia et al., 2021), which employs a matrix layout, displaying the vertices of a hypergraph as circular nodes on one axis, and hyperedges as connecting lines on the other. While experimenting with this technique, we realised that the categories in a dataset could be treated as vertices and their orthogonal combinations as hyperedges. Previous work linking hypergraphs with multidimensional categorical data has done the opposite, representing data items as vertices and categories as hyperedges (Nguyen and Mamitsuka, 2020). Although the final design of MultiCat differs significantly from that of PAOHVis, this early inspiration was crucial, and we believe that modelling categorical data as hypergraphs may be fruitful in a variety of other contexts.

1.3 Design requirements

We developed the following set of design requirements for MultiCat by assessing the capabilities and limitations of the above techniques. From the outset, we decided to adopt an aggregation-based approach that provided a direct representation of category counts.

R1: Aggregate categories. Provide a compact visual representation of the data in frequency form, ensuring that the categories within each combination/aggregate are easily readable.

R2: Show category distributions. Include univariate summaries for each variable that can be readily compared. Users should be able to extract absolute values and marginal frequencies for each category.

R3: Support multiple variables. Allow the user to visualise 3–20 categorical variables. Variables should be treated as equally as possible, and changing their order should not drastically alter the display. Additionally, the layout used should be relatively independent of the number of data items.

R4: Support high-cardinality variables. Ensure that the technique can handle variables with potentially large numbers of categories, while accentuating the most important/frequent categories within each variable. The cardinality of variables may differ considerably, but most variables will be expected to have between two and ten categories.

R5: Handle ordinal variables. Ordinal variables should also be supported, and they should be visually distinct from nominal ones. For ordinal variables, the inherent order of categories should be apparent in the visualisation.

R6: Allow interactive refinement and visual feedback. Users should be able to dynamically add and remove variables, and to select/query different subsets of categories. The percentage of selected data should always be visible, and the display should update immediately when the user interacts with it.

R7: Incorporate filtering. The interface should allow users to filter the data and compute conditional probabilities from the resultant subsets.

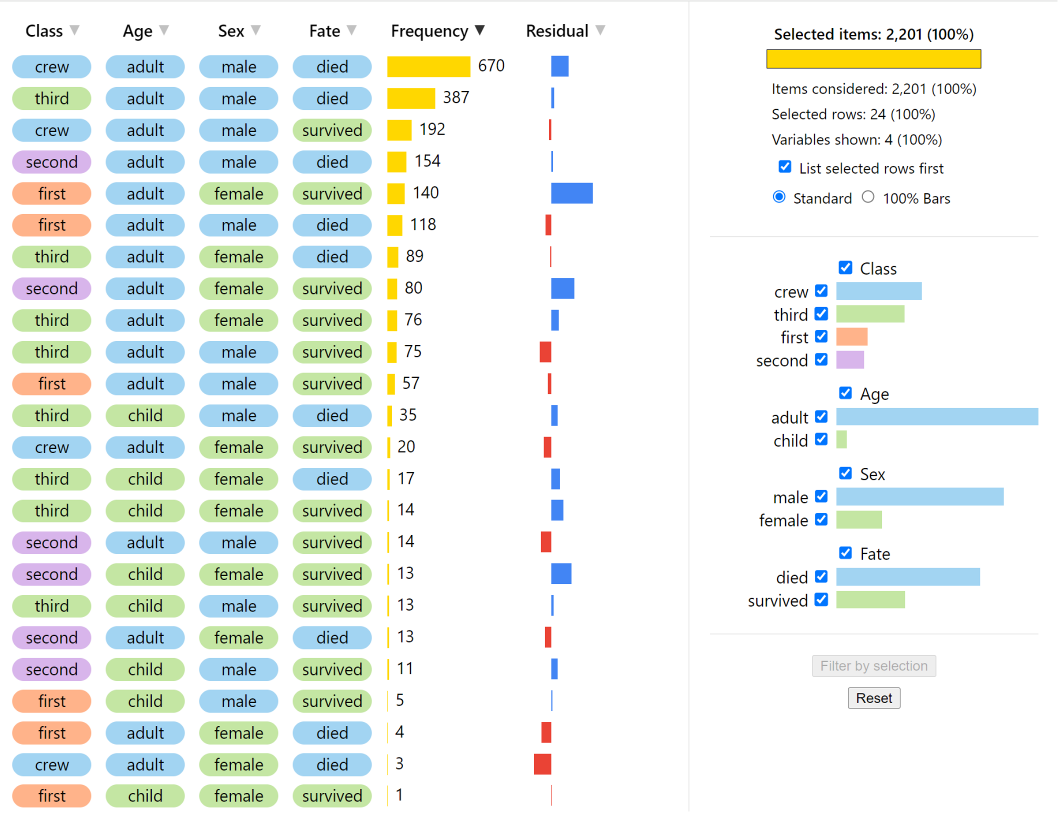
R8: Incorporate sorting. Users should be able to efficiently sort categorical and numeric values. It should be possible to sort by multiple columns in order to break ties at higher levels.

R9: Use a minimalist design. The interface should avoid unnecessary features that detract from the above requirements.

These nine requirements have guided the development and evaluation of our proposed technique for visualising multidimensional categorical data.

1.3.1 Assumptions

We make the following assumptions about the data to be analysed within MultiCat:



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

Selected items: 2,201 (100%)

Items considered: 2,201 (100%)
Selected rows: 24 (100%)
Variables shown: 4 (100%)

☑ List selected rows first

● Standard ○ 100% Bars

☑ Class
crew ☑
third ☑
first ☑
second ☑

☑ Age
adult ☑
child ☑

☑ Sex
male ☑
female ☑

☑ Fate
died ☑
survived ☑

Figure 1.5: MultiCat visualisation of the Titanic dataset (Dawson, 1995). The spreadsheet view on the left shows every observed combination of categories, aggregated and sorted by frequency. Positive (blue) residuals and negative (red) residuals indicate over- and under-represented combinations, respectively. The sidebar on the right summarises univariate category distributions, with categories grouped by variable and ordered by frequency.

1. The input dataset contains only nominal and ordinal variables.
2. Categories belonging to the same variable are mutually exclusive.
3. Categories belonging to different variables are not necessarily independent.
4. Any missing values are coded as "Unknown".

1.4 The MultiCat technique

In this section, we describe the MultiCat technique in detail and justify our design decisions with reference to visualisation theory and existing tools. We have endeavoured to use perceptually efficient visual encodings in our design, but the complexity of the data meant there were a number of trade-offs involved. Figure 1.5 shows the MultiCat interface with the Titanic dataset (Dawson, 1995) loaded in. MultiCat consists of two coordinated views: a spreadsheet view on the left, which is the main display, and a sidebar on the right. The spreadsheet view shows distinct combinations of orthogonal categories (rows) associated with a chosen set of variables (columns). The sidebar, on the other hand, displays information about individual categories and affords an intuitive means of selecting and filtering different subsets of the data. We provide more detail about the layout of each of these components, before addressing their interactive capabilities. The descriptions given here reflect our final prototype, `https://dgt12.github.io/multicat/`, which differs slightly from the version used by participants in our formative user study.

1.4.1 Spreadsheet view

The spreadsheet view in MultiCat provides a compact visual representation of categorical data in *frequency form* (Friendly and Meyer, 2015), displaying each combination of orthogonal categories for the selected variables. Rows represent distinct category combinations and columns represent variables. In the two right-most columns, frequency and Pearson residual values are shown as embedded bar charts (Gratzl et al., 2013) in a similar manner to UpSet's "Cardinality" and "Deviation" metrics (Lex et al., 2014). This arrangement aggregates items with shared characteristics, fulfilling design requirement R1. For instance, the top rows in Figure 1.5 indicate that the largest groups of people on board the Titanic were 670 male adult crew members who tragically died and 387 male adult third-class passengers who suffered the same fate. The categories within each combination are represented by colour-coded "stickers" with text labels. For nominal data, the colour assignment is based on the frequency ranking within each variable: the most frequent category is blue, followed by green, and so on, with categories beyond fifth position being coloured grey. To maintain uniform column widths, variable and category labels are truncated as necessary, with tooltips displaying the full names.

MultiCat's category stickers enhance perceptual processing by combining colour and text within a single component. In the visualisation literature, the closest counterparts to these stickers are Taggle's (Furmanova et al., 2020) five "Item Visualization" options for categorical data. However, these approaches either separate colour from text, decreasing visual immediacy (see Figure 1.6, left column), or align category icons horizontally, which does not accommodate fixed text labels or high-cardinality variables (see Figure 1.6, right column). MultiCat addresses these issues by effectively balancing perceptual recognition with spatial efficiency.

The right-most column in the spreadsheet view represents each combina-

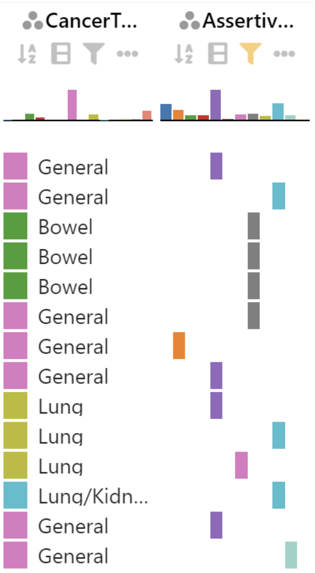
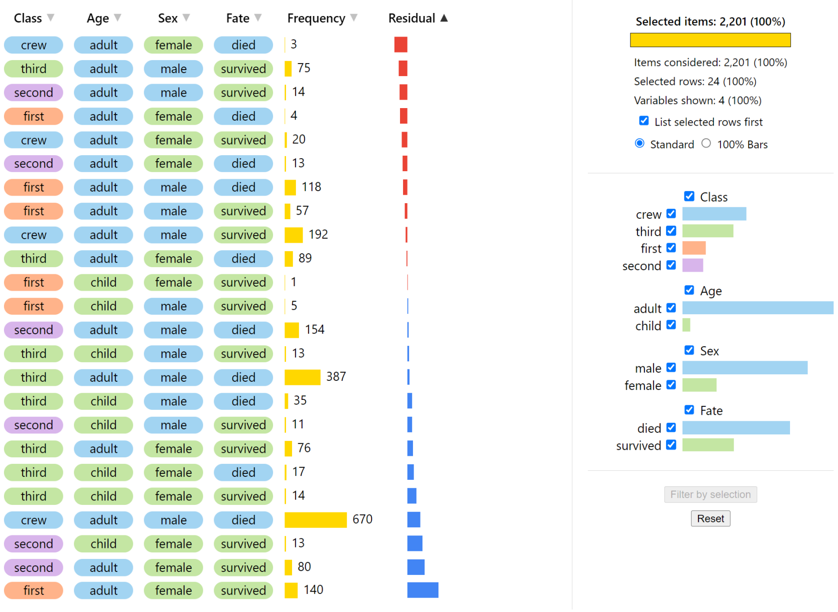


Figure 1.6: Taggle's "Color & Label" encoding (left column) and "Matrix" encoding (right column) are two alternatives to MultiCat's "Sticker" approach.

tion's Pearson residual (Friendly, 1994) in a diverging bar chart. This measure shows the deviation of a combination's observed frequency from its expected frequency, assuming mutual independence between categories. Akin to UpSet's (Lex et al., 2014) "Deviation" metric, combinations occurring more or less frequently than expected are represented by blue (positive) or red (negative) bars, respectively. As illustrated in Figure 1.7, the smallest negative residuals for the Titanic dataset are linked to deceased female adults and surviving male adults, while the largest positive residuals are associated with female survivors.

Combinations in the spreadsheet view are initially sorted by descending frequency, not by the categories themselves. This is demonstrated by the decreasing size of yellow bars in Figure 1.5. Sorting in this way ensures that variables are treated relatively equally, aligning with design requirement R3. More importantly, it simplifies gaining an overview of the distribution of category combinations, and facilitates identification of the most and least frequent aggregates. The least frequent combinations reveal anomalies in the data, such as one girl in first class who survived the Titanic disaster and three female crew members who did not (see bottom rows of Figure 1.5). Crucially, the spreadsheet view leverages users' familiarity with interactive tables, harnessing their depth of meaning and structural benefits, as detailed by Bartram et al. (2021).



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

Selected items: 2,201 (100%)

Items considered: 2,201 (100%)
Selected rows: 24 (100%)
Variables shown: 4 (100%)
☑ List selected rows first
● Standard ○ 100% Bars

☑ Class
crew ☑
third ☑
first ☑
second ☑

☑ Age
adult ☑
child ☑

☑ Sex
male ☑
female ☑

☑ Fate
died ☑
survived ☑

Figure 1.7: Sorting the Titanic data by residuals confirms expected patterns: male survivors were under-represented, whereas female survivors were over-represented.

1.4.2 Sidebar

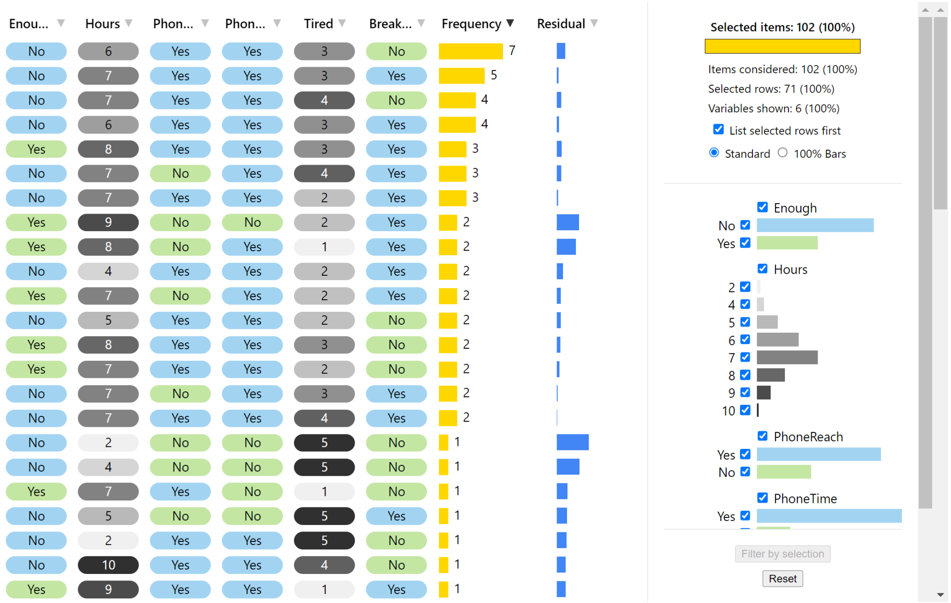
The sidebar, positioned on the right-hand side of the display, offers a succinct summary of category distributions with respect to the current filter, satisfying design requirement R2. It features a series of horizontal bar charts, where categories are grouped by variable and sorted by frequency for nominal variables, or inherent order for ordinal variables. Each variable group includes a heading with the variable name and a checkbox, followed by individual category labels with their own checkboxes and bars. Category labels are truncated if they do not fit within the allocated space. All category bars share a common baseline and scale, enabling direct comparisons within and across variables. When no filter has been applied, as in Figure 1.5, the category bar lengths correspond to the marginal distribution of each variable. For instance, the sidebar in this figure highlights that the crew accounted for a surprisingly large proportion of people on board the Titanic, that only a small proportion of passengers were children or female, and that roughly twice as many people died as survived. The sidebar also serves as a quick reference for identifying the number of categories per variable (based on the number of bars) and determining the sequence in which colours are assigned to nominal categories (blue first, then green, etc.). This in turn helps the user to establish the relative rank of the categories within each variable when inspecting the combinations in the spreadsheet view.

The sidebar interacts with the spreadsheet view in simple yet powerful ways, as explained in the section on Interaction. At the top of the sidebar, four statistics provide useful context about the current state of the display. These statistics are expressed both as absolute values and percentages. They relate to items that are currently selected (”Selected items”), items that are currently visible but not necessarily selected (“Items considered”), distinct combinations that are currently selected (“‘Selected rows”), and active variables (“Variables shown”). The “Selected items” statistic is considered the most important, and as such, it also has a yellow bar chart representing its value. The highlighted combination frequencies in the spreadsheet view necessarily sum to the number of selected items in the sidebar. There are two buttons at the bottom of the sidebar, including a “Reset” button which provides a convenient means of returning to the original display.

Our sidebar is inspired by Taggle’s “Data Selection Panel” (Furmanova et al., 2020), but we have made several changes to optimise the readability and scalability of categorical data. Both sidebars display category distributions and enable direct category selection by clicking on the bars. However, as Taggle’s use of vertical bars in a fixed space can lead to overcrowding with high-cardinality variables, in MultiCat we employ horizontal bars of uniform height. This maintains readability even for variables with many categories, providing a scrollbar in the situation where not all bars are visible at once. Another point of difference is that MultiCat respects the inherent order of ordinal variables, as discussed below.

1.4.3 Ordinal variables

As per design requirement R5, MultiCat can handle ordinal variables as well as nominal ones. Variables in the input dataset whose category names begin with Arabic numerals (e.g., “1 Small”, “2 Medium”, “3 Large”) are treated as ordinal. There are two key differences regarding the appearance and behaviour of ordinal variables, which are illustrated in Figure 1.8. Firstly, these variables are depicted using greyscale values instead of hue, following Mackinlay’s recommendation for a more precise visual encoding (Mackinlay, 1986). This design choice not only aligns with best practice but also ensures that ordinal variables are instantly distinguishable from nominal ones, allowing users to quickly gauge the number of variables of each type. Categories beginning with larger numbers are represented by progressively darker shades, with white text being used on darker backgrounds to ensure that category stickers remain



Enou... ▼	Hours ▼	Phon... ▼	Phon... ▼	Tired ▼	Break... ▼	Frequency ▼	Residual ▼
No	6	Yes	Yes	3	No	7	
No	7	Yes	Yes	3	Yes	5	
No	7	Yes	Yes	4	No	4	
No	6	Yes	Yes	3	Yes	4	
Yes	8	Yes	Yes	3	Yes	3	
No	7	No	Yes	4	Yes	3	
No	7	Yes	Yes	2	Yes	3	
Yes	9	No	No	2	Yes	2	
Yes	8	No	Yes	1	Yes	2	
No	4	Yes	Yes	2	Yes	2	
Yes	7	No	Yes	2	Yes	2	
No	5	Yes	Yes	2	No	2	
Yes	8	Yes	Yes	3	No	2	
Yes	7	Yes	Yes	2	No	2	
No	7	No	Yes	3	Yes	2	
No	7	Yes	Yes	4	Yes	2	
No	2	No	No	5	No	1	
No	4	No	No	5	No	1	
Yes	7	Yes	No	1	No	1	
No	5	No	No	5	Yes	1	
No	2	Yes	Yes	5	No	1	
No	10	Yes	Yes	4	No	1	
Yes	9	Yes	Yes	1	Yes	1	

Selected items: 102 (100%)

Items considered: 102 (100%)

Selected rows: 71 (100%)

Variables shown: 6 (100%)

☑ List selected rows first

● Standard ○ 100% Bars

☑ Enough
No ☑
Yes ☑

☑ Hours
2 ☑
4 ☑
5 ☑
6 ☑
7 ☑
8 ☑
9 ☑
10 ☑

☑ PhoneReach
Yes ☑
No ☑

☑ PhoneTime
Yes ☑

Figure 1.8: MultiCat uses greyscale values for ordinal variables. This dataset about people's sleeping habits (Lomuscio, 2020) comprises two ordinal variables and four nominal ones. Ordinal categories are displayed in their inherent order in the sidebar.

readable. Secondly, the sorting of ordinal variables in MultiCat maintains their inherent sequential order, rather than applying a frequency-based ranking. This natural ordering is also used for categories in the sidebar. A limitation of the greyscale mapping is its reduced effectiveness in differentiating between selected categories with lighter shades and non-selected categories with darker shades, due to the interplay of transparency with saturation (Munzner, 2014). However, the category checkboxes in the sidebar and partial transparency of entire rows (including nominal variables) allow the user to ascertain which categories have and have not been selected.

1.4.4 Colour coding

Colour coding is an effective means of displaying category information (Ware, 2019), with colour differences being more readily perceived than shape differences (Wolfe and Horowitz, 2017). However, the number of colours should be limited to between five and ten to ensure they can be rapidly distinguished (Healey, 1996; Ware, 2019). With this in mind, MultiCat employs a maximum of six colours for nominal variables, using grey for all categories beyond the fifth most frequent one. While this makes less frequent categories harder to differentiate, we consider this to be a good compromise since these categories are generally less important and tend to be dispersed across fewer distinct combinations (rows). This conservative use of colour is also in keeping with design requirement R9. The six colours chosen for MultiCat's qualitative palette were inspired by Google Sheets' drop-down presets, which are well-balanced, have similar intensity and are suitable for reading black text. Repeating colours across variables does make it harder to identify related information in each view, yet this approach is preferable to assigning a unique hue per category, which would quickly result in a palette of indistinguishable shades. It also means that the category rankings are shown consistently, as explained below.

MultiCat's colour usage aims to reduce the viewer's cognitive load by leveraging preattentive processing (Ware, 2019). Although the visualisation is comprehensible in greyscale due to its text labels and interactivity, colour enhances usability by: (1) distinguishing categories within variables; (2) linking categories across views; and (3) indicating their frequency ranking or natural order. Given that hue does not have an inherent perceptual ordering (Munzner, 2014; Muth, 2021), MultiCat prioritises this first point over the third one. Yet, because the sidebar initially shows the order in which colours are assigned, this enables users to discern patterns based on the colour of stickers within each row. This accords with the Pearson residuals in the right-most column. For example, in visualisations containing only nominal variables, the most frequent combinations would be expected to have predominantly blue stickers, as these represent the most frequent categories. Indeed, the most frequent combination in Figure 1.5 features only categories with blue stickers. It is also apparent when relatively infrequent categories occur within a relatively frequent combination (for example, the 13 girls in second class who survived). The distribution of colours within combinations therefore aids in confirming expected trends and identifying unexpected patterns, further strengthening the use of colour in MultiCat.

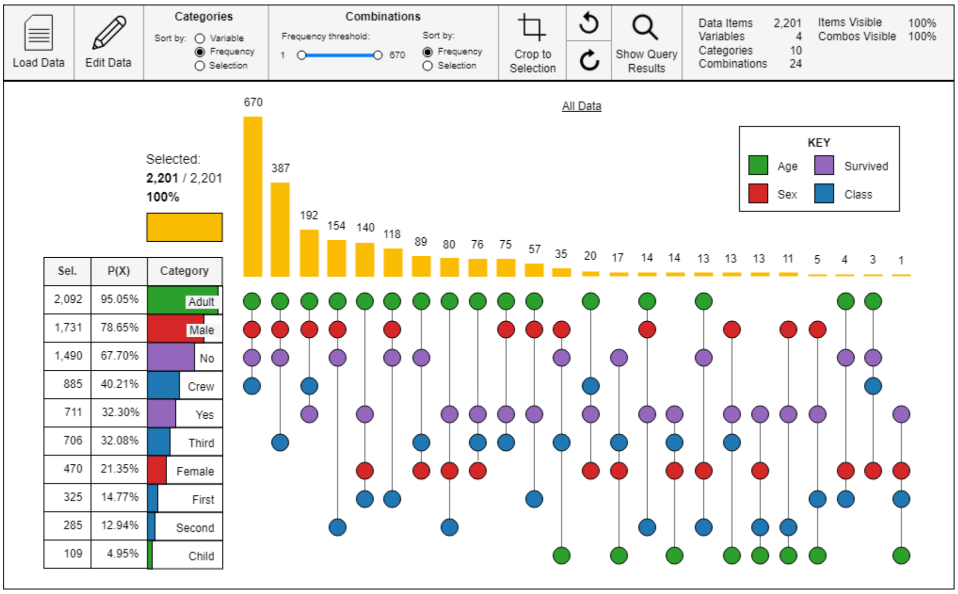
1.5 Comparison with earlier design

An early version of MultiCat was proposed in a poster paper (Trye, 2022). Inspired by set-based techniques such as PAOHVis (Valdivia et al., 2021) and UpSet (Lex et al., 2014), rather than tabular approaches like Taggle (Furmanova et al., 2020), the original design stemmed from the observation that high-dimensional combinations of categories can be represented as hyperedges in a hypergraph. It is useful to compare our new design with this older one, reproduced in Figure 1.9, to show how and why it has evolved over time. This provides clear evidence of our iterative design process, while also helping to illustrate the rationale behind specific choices. We identify and explain six key changes below.

Firstly, we have transposed the matrix and re-positioned the sidebar. Previously, rows represented categories and columns represented combinations. Flipping this orientation helps to improve scalability since a multidimensional categorical dataset will typically contain a larger number of distinct combinations than variables, and it is easier to scroll vertically than horizontally. This also better aligns with a spreadsheet layout, where columns typically represent variables, rather than items or groups of items.

Secondly, text labels have been embedded within each node, resulting in the coloured "stickers" described in the previous section. This change was motivated by the need for a more readable representation. In the original design, the smaller nodes made the layout more compact but ultimately much harder to decipher, particularly in non-interactive settings. This is because each node had to be decoded by manually tracking its position along both axes to find the corresponding label, which required a much higher cognitive load than reading category labels directly.

Thirdly, we removed the connecting lines between the categories in each combination. These were superfluous and detracted from the interactive spreadsheet metaphor. For instance, if horizontal lines were present in our new design, this might deceive users into thinking that sorting happens with respect to entire combinations rather than individual columns.



Categories — Sort by: ○ Variable ● Frequency ○ Selection

Combinations — Frequency threshold: 1 – 670; Sort by: ● Frequency ○ Selection

Data Items	2,201	Items Visible	100%
Variables	4	Combos Visible	100%
Categories	10		
Combinations	24		

Selected:
2,201 / 2,201
100%

Sel.	P(X)	Category
2,092	95.05%	Adult
1,731	78.65%	Male
1,490	67.70%	No
885	40.21%	Crew
711	32.30%	Yes
706	32.08%	Third
470	21.35%	Female
325	14.77%	First
285	12.94%	Second
109	4.95%	Child

Figure 1.9: An early design of MultiCat, which looked less like a spreadsheet and more like a custom set-based representation (Trye, 2022).

Fourthly, space is now allocated on a per-variable rather than per-category basis. This makes the layout more efficient, especially if the dataset contains one or more high-cardinality variables. Of course, the stickers in the new design are also much wider than their node predecessors, so laying them all out side-by-side would necessarily consume a lot more space.

The fifth change concerns the assignment of colour to variables. Originally, each variable was assigned a distinct colour, which was then shared among all categories belonging to that variable. This approach made it easy to differentiate between variables, but difficult to distinguish the categories within each one. To address this, we experimented with using different shades of the same hue for categories belonging to the same variable. However, this conflicted with having different opacities for selected and non-selected items; for instance, it was difficult to distinguish lighter categories from non-selected ones.

Another drawback of using different shades of the same hue was that this led to unintended salience effects. Darker shades appeared more influential within combinations, despite all categories in a combination having the same frequency. They also stood out disproportionately in infrequent combinations. Furthermore, this method of colour allocation only works for categories with a small number of categories as it is difficult to distinguish more than three shades of the same colour (Muth, 2021).

To overcome these issues, we decided to allocate the same set of colours to each variable, following the same order of assignment. This method aligns with the default behaviour of TableLens (Rao and Card, 1994) and Taggle (Furmanova et al., 2020). Since we sort the categories within each variable, this also implicitly conveys their relative rankings, albeit in a less intuitive way than a sequential scale.

Our sixth and final change was to add the column for residuals next to the frequencies. Inspired by UpSet's (Lex et al., 2014) "Deviation" measure, this shows the extent to which combinations are over- or under-represented within the dataset, which may lead to additional insights.

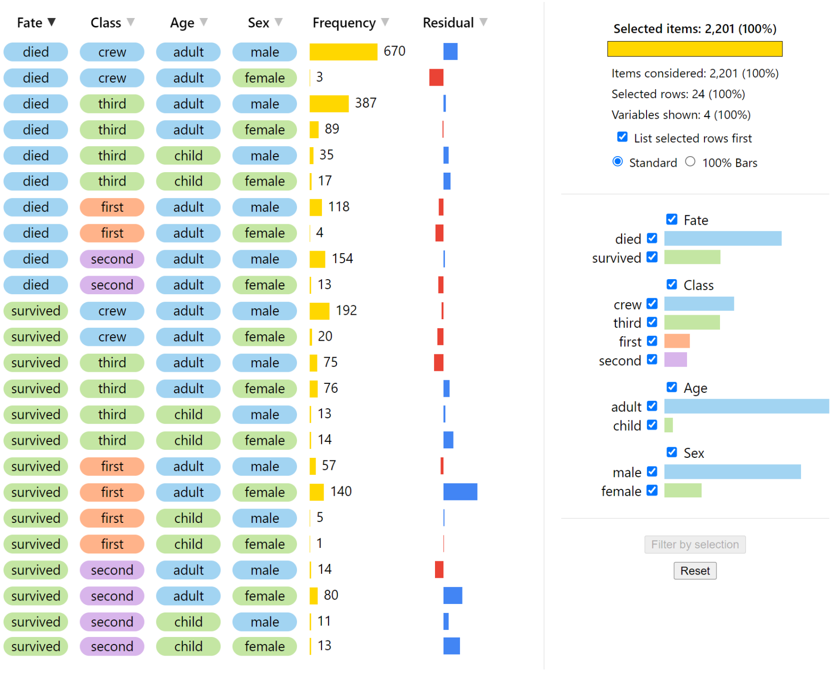
1.6 Interaction

This section describes the rich interactive features supported by MultiCat, which can be used in conjunction to highlight salient patterns, trends and relationships in the data. Users are initially presented with a high-level overview of categories and their combinations, but they may wish to interactively explore the data to gain a deeper understanding, either in a directed or undirected manner. The use of common spreadsheet operations, such as sorting and reordering columns, helps to consolidate the user's sense-making process (Bartram et al., 2021), while features such as selection and filtering enable rapid visualisation and comprehension of user-defined queries.

1.6.1 Sorting

MultiCat's sort functionality, which relates to design requirement R8, can be used to reveal relationships between category subsets and combinations. By default, combinations (rows) in the spreadsheet view are sorted by descending frequency, with residuals breaking ties (see Figure 1.5). Interacting with column headers rearranges the combinations, and through such exploration, enables the user to discover potentially revealing ways of viewing the data. Each column header is marked with a small triangle, indicating its ability to be sorted and reflecting the current state of the display. The triangle of the most recently sorted column is black, whereas all others are light grey. A single click on a column sorts it in descending order, with the sorting method varying by data type: nominal variables by rank frequency; ordinal variables by the number preceding the category name; and frequency and residual columns by numeric value. Clicking again on the same column switches to ascending order (Figure 1.7). For nominal and ordinal variables, this sorting mirrors or inversely matches the top-to-bottom order of categories in the sidebar. Sorting by frequency and residual columns quickly highlights minima, maxima and outliers. Multi-column sorting is enabled by clicking on several column headers in succession, creating a user-defined hierarchy where the last sorted variables are prioritised. This is evident in Figure 1.10, where the four categorical variables have been sorted from right to left, with Fate at the top of the hierarchy. Note how the frequencies and residuals fluctuate considerably from row to row.

The order of combinations (rows) in the spreadsheet view can be configured to either prioritise the user's current selection or remain independent of it. The sidebar features a "List selected rows first" checkbox, which is enabled by default. When this option is selected, highlighted combinations are grouped at the top of the display, with the current sort criteria being applied separately to selected and non-selected items, as demonstrated in Figure 1.11. Conversely, when this option is unchecked, all rows follow a global sort order, regardless of selection status, as depicted in Figure 1.12. These settings emphasise different aspects of the data: the former facilitates direct comparisons of combinations of interest, while the latter reveals their distribution within the broader context of the dataset.



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

Selected items: 2,201 (100%)

Items considered: 2,201 (100%)
Selected rows: 24 (100%)
Variables shown: 4 (100%)
☑ List selected rows first
● Standard ○ 100% Bars

☑ Fate
died ☑
survived ☑

☑ Class
crew ☑
third ☑
first ☑
second ☑

☑ Age
adult ☑
child ☑

☑ Sex
male ☑
female ☑

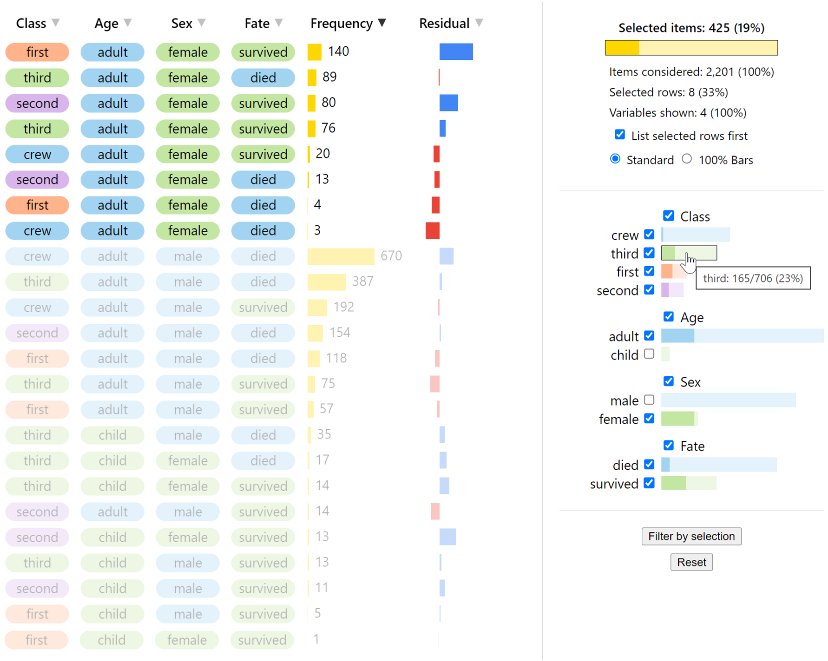
Figure 1.10: The Titanic dataset with combinations sorted by all four categorical variables. The columns have also been reordered, with Fate now appearing on the left.

1.6.2 Reordering

Columns representing categorical variables can be reordered by dragging and dropping their headers to a new position. The top-to-bottom ordering of categorical variables in the sidebar matches their left-to-right ordering in the spreadsheet view. This means that variables further to the left in the spreadsheet view appear higher in the sidebar. Users may wish to organise variables in a specific manner, or ensure that the response variable occupies the left-most column, so that it is prominent in both views. The columns in Figure 1.10 have been rearranged so that the response variable (Fate) comes first rather than last.

1.6.3 Customising category bar charts

The two radio buttons in the sidebar control the appearance of the individual category bar charts. The default “Standard” option scales each bar’s length according to the most frequent category in the dataset (e.g., “adult” in the Titanic dataset). In contrast, the “100% Bars” option normalises bar lengths, so that the selected proportions of each category can be directly compared, as shown in Figure 1.12. This feature—reminiscent of Mondrian’s (Theus, 2002) built-in support for converting bar charts to Spine plots—effectively conveys



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

Selected items: 425 (19%)

Items considered: 2,201 (100%)
Selected rows: 8 (33%)
Variables shown: 4 (100%)
☑ List selected rows first
● Standard ○ 100% Bars

☑ Class
crew ☑
third ☑
first ☑
second ☑
third: 165/706 (23%)

☑ Age
adult ☑
child ☐

☑ Sex
male ☐
female ☑

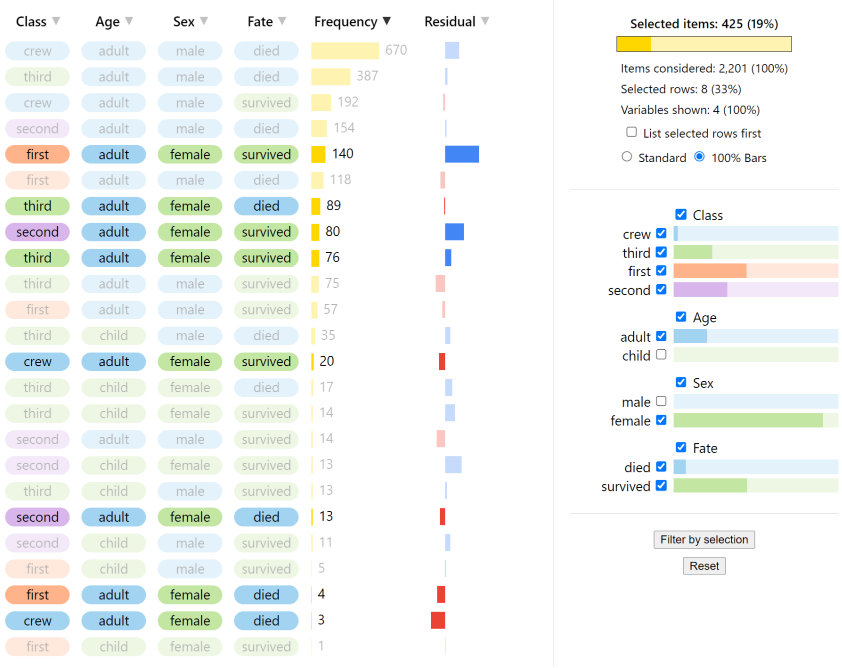
☑ Fate
died ☑
survived ☑

Figure 1.11: The Titanic dataset with all 425 female adults selected. The “Selected items” bar chart indicates that the joint probability of someone on the Titanic being female and an adult is 19%.

part-whole relationships and is particularly useful for visualising relatively infrequent categories, which might otherwise be difficult to discern.

1.6.4 Brushing and linking

MultiCat uses brushing and linking (Hearst, 1999) to capture the association between selected items in the spreadsheet view and sidebar. Selected items in MultiCat are fully opaque, whereas non-selected items are rendered partially transparent to reduce their salience. This is exemplified in Figure 1.11, which highlights female adults in the Titanic dataset. Whenever a selection is made, three updates occur simultaneously: matching combinations (rows) in the spreadsheet view are highlighted; the statistics in the sidebar are updated, with the yellow chart showing the selected items as a proportion of the current filter; and the individual category bars in the sidebar reflect the corresponding proportion within each category. Together, these features play a critical role in revealing complex interactions between multiple categories and their combinations, helping users to see higher-dimensional features and structures in the data.



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

Selected items: 425 (19%)

Items considered: 2,201 (100%)
Selected rows: 8 (33%)
Variables shown: 4 (100%)
□ List selected rows first
○ Standard ● 100% Bars

☑ Class
crew ☑
third ☑
first ☑
second ☑

☑ Age
adult ☑
child □

☑ Sex
male □
female ☑

☑ Fate
died ☑
survived ☑

Figure 1.12: The Titanic dataset highlighting female adults, as per Figure 1.11, but with the “List selected rows first” checkbox disabled and 100% bar charts displayed in the sidebar.

1.6.5 Tooltips

Hovering over most visual elements in MultiCat produces an informative tooltip, yielding “details-on-demand” (Shneiderman, 1996). In the spreadsheet view, hovering near a yellow frequency bar displays that combination’s relative contribution to the current filter, whereas hovering near the red and blue residuals shows their exact values. Explanations of these two metrics are accessible via tooltips associated with their column headers. Additionally, tooltips reveal the full names of items on column headers or stickers, which is helpful for truncated text. In the sidebar, tooltips detail the selected proportion of each category in the format “third: 165/706 (23%)” (see Figure 1.11), where the numerator shows the exact number of selected instances and the denominator reflects the total number of occurrences of the category within the current filter. Finally, tooltips for the inactive “Filter by selection” button explain the criteria for its activation.

1.6.6 Dynamic queries

In MultiCat, dynamic queries facilitate the exploration of specific groups of categories, providing a logical and intuitive means for users to drill down into the data. These queries integrate interactive refinement and visual feedback (Shneiderman, 1994), as per requirement R6. Users can form simple Boolean queries involving AND/OR logic by manipulating the category checkboxes in the sidebar. These checkboxes can be toggled directly, or the user can click on or besides the category bars to select *only* that category from within its parent variable. This is a useful shortcut for isolating one or a few categories within a high-cardinality variable. Alternatively, the user can click on category stickers in the spreadsheet view; this has the same effect as toggling the checkboxes, unless all categories for the parent variable are already selected, in which case it acts like the bar shortcut.

MultiCat employs straightforward Boolean logic in its queries: it uses OR (union) logic for categories within the same variable and AND (intersection) logic across different variables. This design prioritises simplicity over expressiveness, maximising ease of use and reducing the risk of logical errors (Spoerri, 1995). It leverages the principle that AND-ing categories within the same variable, under mutual exclusivity, always leads to an empty intersection (i.e., no matching records). Generally, selecting more categories in MultiCat broadens a query's scope, while choosing fewer categories narrows it. Each represented variable must have at least one selected category for matches to occur. Currently, it is not possible to formulate complex queries in MultiCat that feature multiple levels of nesting or incorporate more sophisticated Boolean operators such as XOR.

Dynamic queries allow users to adjust their selection based on their information needs. The interactive and exploratory nature of these queries encourages users to ask questions of the data that they might not otherwise consider, such as "Are there more items with characteristics X than Y?" or "What happens if I select or deselect this checkbox?" In this process, users may uncover strongly associated categories, or one-way dependencies where a less frequent category is nearly always accompanied by a more frequent one, but not vice versa.

The "Selected items" bar chart functions as a real-time probability calculator for the current query, displaying empirical values based on actual data observations. This chart shows *a priori* probabilities when no filter is applied ("Items considered" is 100%) and conditional probabilities otherwise. For instance, Figure 1.11, in which all data items are represented, shows the proportion of women (adult + female) on board the Titanic.

As seen in Figure 1.1, Mondrian allows users to formulate Boolean queries through linked bar charts (Theus, 2002), similar to MultiCat. While Mondrian offers a wider range of Boolean operators than MultiCat, and allows more flexibility in applying them—for instance, OR-ing categories from different variables is permitted—it lacks any visual clues as to a query's internal representation. This can easily lead to user errors, especially when constructing complex queries involving multiple variables and operations. In contrast, MultiCat's simple design ensures that the syntax of a query can always be inferred from the active checkboxes in the display.

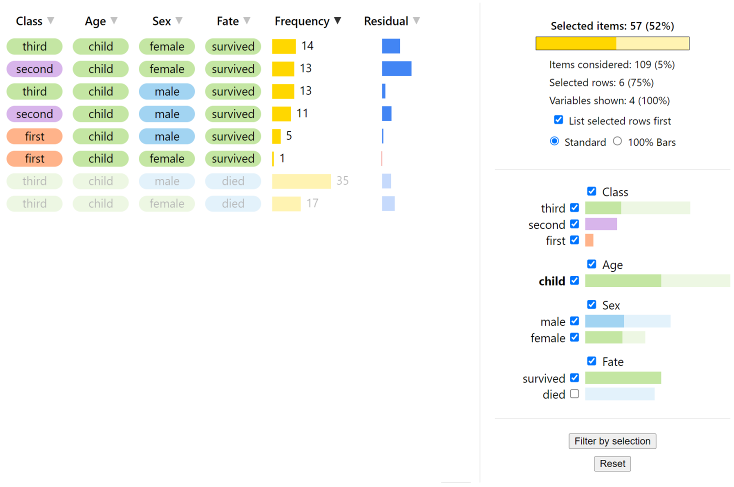
1.6.7 Filtering

Filtering is a useful strategy for reducing the size and complexity of a categorical dataset. In MultiCat, it is possible to filter out data items that are not part of the current category selection, as well as entire variables. This aligns with design requirements R6 and R7. Filtering differs from selection in that excluded items are removed from the display, rather than merely being faded out.

Category-based filtering is accomplished by selecting a subset of categories and clicking on the "Filter by selection" button; this removes non-selected combinations (rows) from the spreadsheet view and the corresponding items from the sidebar. To prevent confusion, categories always remain the same colour, even if their ranking changes within a new filter. However, the order of categories within the sidebar updates accordingly. Filtering criteria can be progressively refined to reveal more in-depth relations in the data. Categories that are no longer represented in the filtered data are removed from the sidebar to save space, whereas categories that form the basis of the filter are emphasised in bold; these correspond to the "given" part of a conditional query.

As an example, Figure 1.13 shows the Titanic dataset filtered by "child". Since children accounted for only a small proportion of passengers, their distribution is quite different from the overall dataset shown in Figure 1.5. For instance, there are many more children in second class than first class, and there is a relatively even split between children who died and survived. The query in Figure 1.13 specifically shows the conditional probability $P(\mathit{survived}|\mathbf{child})$; that is, the percentage of children who survived: 52%, or 57 children as an absolute value. The opacity of the bars shows that all children in first and second class survived (although there were relatively few children in these classes), whereas only a third of children in third class survived. Notably, while a similar *number* of boys and girls survived, a greater *proportion* of boys died, which is also true (and more pronounced) for males and females in general.

In addition to category filtering, MultiCat supports the removal of variables via the variable checkboxes in the sidebar, which are positioned centrally. The



Class ▼	Age ▼	Sex ▼	Fate ▼	Frequency ▼	Residual ▼
third	child	female	survived	14	
second	child	female	survived	13	
third	child	male	survived	13	
second	child	male	survived	11	
first	child	male	survived	5	
first	child	female	survived	1	
third	child	male	died	35	
third	child	female	died	17	

Selected items: 57 (52%)

Items considered: 109 (5%)

Selected rows: 6 (75%)

Variables shown: 4 (100%)

☑ List selected rows first

● Standard ○ 100% Bars

☑ Class
third ☑
second ☑
first ☑

☑ Age
child ☑

☑ Sex
male ☑
female ☑

☑ Fate
survived ☑
died ☐

Filter by selection

Reset

Figure 1.13: The Titanic dataset filtered by children and showing the conditional probability $P(survived|\textbf{child})$.

user can examine as many or as few variables at a time as desired. When variables are removed, combinations in the spreadsheet view are instantly updated, including their frequencies and residuals. The category bars for disabled variables are removed from the sidebar, but the variable checkboxes remain visible so that they can be re-selected. Generally, reducing the number of variables increases combination frequencies in the spreadsheet view as there are fewer permutations; however, the number and distribution of categories within these variables is still a significant factor.

1.6.8 Scrolling

Scrolling becomes necessary when there is too much data to fit everything in the available space. If there are too many combinations (rows) in the spreadsheet view, a vertical scrollbar appears in the main window. The header row of the table is frozen, so that the column names remain visible when the user scrolls. Likewise, a horizontal scrollbar is added when there are too many variables (columns). As previously mentioned, an internal vertical scrollbar is added to the sidebar when there is insufficient space to show all category bars at once. The summary statistics and buttons in the sidebar are also fixed in place. Datasets that require excessive scrolling in either direction increase the user's cognitive load and are therefore not recommended, though the user may be able to quickly derive meaningful subsets using the above interactive features. Zooming out may also help in some cases.

1.7 Implementation

A live demo of MultiCat, preconfigured with the Titanic dataset (Dawson, 1995) for demonstration purposes, is available at `https://dgt12.github.io/multicat/`. Researchers wishing to visualise their own categorical datasets can do so by downloading the source code, replacing the input file, and running the application on a local server. Detailed instructions and all necessary code are provided in the project GitHub repository: `https://github.com/dgt12/multicat`.

MultiCat is implemented in Svelte (`https://svelte.dev/`) and has been tested in Google Chrome with datasets containing up to 20 categorical variables and 500,000 data items. The data are currently read in as an array of JavaScript objects, where each object represents a single item. All data are converted from case form to frequency form before the visualisation is rendered. The spreadsheet view is simply an HTML table.

Regarding limitations, our prototype does not currently integrate with other tools, support more than one input format, provide edit history (undo/redo) or generate publication-ready figures. There are also important scalability constraints, which we discuss below.

1.8 Scalability

MultiCat is built on the premise that aggregation is crucial for creating scalable visualisations. Consequently, its efficacy depends on the presence of many recurrent combinations of categories; this is what exposes meaningful structure within the spreadsheet view. As the number and diversity of categories increase, combination frequencies typically decrease, except in cases where variables are highly associated. Introducing a single high-cardinality variable with weak associations drastically reduces the number of recurrent combinations in MultiCat, underscoring the importance of strategic variable selection. Nevertheless, even in datasets comprising predominantly unique records, MultiCat's interactive features, coupled with the sidebar, remain valuable for exploring the data in a task-driven manner.

The screen space needed for a MultiCat visualisation is a function of the number of categorical variables (horizontal space) and the distinct combinations they form (vertical space). As with many other techniques for visualising categorical data (Hofmann, 2006), MultiCat's display is relatively independent of the number of data items. The maximum number of combinations (rows) for a given set of variables can be calculated by multiplying the variables' cardinalities. This represents the worst-case scenario for vertical space allocation, but it is unlikely that every possible combination will appear in a real-world dataset with several variables. While scrolling allows every combination to be viewed if necessary, other features such as sorting, filtering and querying allow the user to better utilise the available screen space.

Theoretically, design aspects such as scrollbars, fixed column widths and systematic allocation of colours mean that MultiCat can accommodate any number of categories and variables. In practice, however, having too many of these can lead to highly fragmented groups in the spreadsheet view, limiting its effectiveness. Moreover, MultiCat visualisations become challenging to interpret when the number of variables exceeds screen capacity. While scrolling allows access to additional variables, perceiving differences among category combinations becomes difficult as users must remember parts of the display that are not immediately visible.

In terms of cardinality, for best results, most variables should have only a handful of categories, with none exceeding ten. MultiCat is capable of handling datasets that go beyond these limits, as per design requirement R4. However, operating within the stated limits not only aids in keeping recurrent combinations together, but also ensures the number of distinct grey stickers per variable is minimised. This is helpful since the grey stickers are harder to differentiate than those with different colours. Additionally, this approach declutters the sidebar, reducing the need for extensive scrolling and enabling easier comparison of variable distributions.

1.9 Formative user study

This section outlines the procedure and findings of our small-scale user study of MultiCat.[1] The study aimed to detect general usability issues, collect participant feedback, improve the prototype and assess users' ability to interpret the visualisation without prior training. Based on the outcomes of the study, we made several refinements to the prototype, which are detailed below.

1.9.1 Procedure

Our study employed a within-subjects design. Six participants with a background in computer science were recruited, including four students and two staff members. Following other research, this was deemed to be enough participants to identify significant usability issues in a cost-effective manner (Peña-

[1] Ethics approval for this study is given in Appendix A.

Araya et al., 2022). All participants completed the study in the University of Waikato's Usability Lab using the same Windows 11 laptop and mouse. The study was divided into an *Exploratory Phase*, where participants familiarised themselves with the layout and functionality of the MultiCat prototype, and a *Task Phase* comprising the same seven tasks for each of two datasets. Subsequently, participants completed a short online questionnaire. They were encouraged to think aloud throughout the study, and a screen and audio recording was captured for detailed analysis. Each session took approximately 30 minutes, with participants giving informed consent at the beginning.

During the initial *Exploratory Phase*, participants were introduced to the MultiCat prototype with the Titanic dataset loaded in (Dawson, 1995). Having only been told that MultiCat was designed for visualising multiple categorical variables, and that they were viewing data related to the Titanic disaster, participants were asked to explain what they thought the visualisation was showing. They were then encouraged to interact freely with the prototype. The interviewer facilitated this exploration by answering questions and guiding participants towards features they had not yet encountered, using prompts such as “What happens when you hover over one of these frequency bars?” This approach ensured that participants actively engaged with and verbalised their understanding of the functionality, rather than simply being told what it did. This in turn provided valuable insights into which features were and were not intuitive.

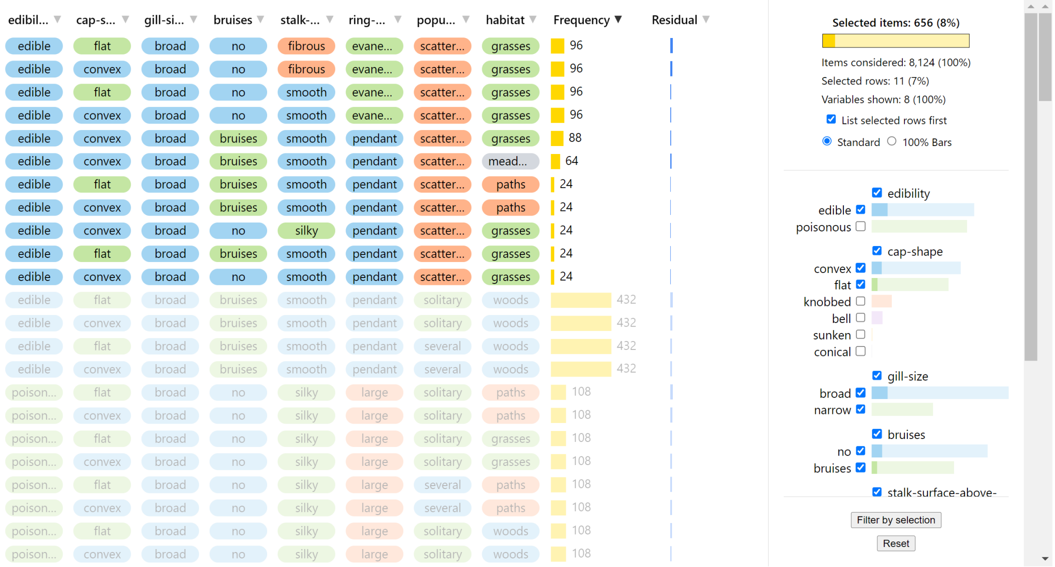
In the subsequent *Task Phase*, participants were asked questions about two publicly available datasets of varying complexity. First, they revisited the Titanic dataset from the *Exploratory Phase*, which comprises four variables, 10 categories and 2,201 observations. The second dataset was a simplified version of the Mushroom dataset (Schlimmer, 1987), representing a hypothetical collection of 8,124 mushrooms belonging to 23 species. We selected only eight of the original 22 variables to ensure all columns were visible in MultiCat without the need for scrolling (see Figure 1.14). These variables, featuring between two and seven categories for a total of 34 categories, encompass a diverse range of properties, including the mushrooms' edibility, physical characteristics, population and habitat. As neither dataset in the study incorporates ordinal variables, we did not evaluate MultiCat's capabilities for handling such data.

Table 1.1 details the tasks that participants were asked to carry out, including the specific questions posed for each dataset. These tasks varied in complexity and were chosen to reflect common activities in categorical data analysis. The first task, while straightforward, served to acquaint users with the size and structure of the dataset. Other tasks integrated different visual elements and features, enabling observation of user strategies. Participants did not receive any feedback on their answers and were asked to reset the display between tasks to ensure a fresh start each time. To mitigate potential memorisation effects, the sequence of tasks for the Mushroom dataset was pseudo-randomised.

After completing the *Task Phase*, participants filled out an online questionnaire. The first set of questions in this survey was about users' familiarity with related tools and their knowledge of statistical concepts, while the next set required them to provide a subjective rating of their experience and impressions using MultiCat. Finally, there were three open-ended questions asking users to identify things they liked and disliked about MultiCat, as well as any suggestions they had for enhancing the interface.

Table 1.1: Task descriptions and associated questions for the Titanic and Mushroom datasets.

Task	**Titanic Dataset (Dawson, 1995)**	**Mushroom Dataset (Schlimmer, 1987)**
T0: Summarise dataset	How many items (in this case, *people*) does the dataset contain? How many categorical variables does the dataset contain?	How many items (in this case, *mushrooms*) does the dataset contain? How many categorical variables does the dataset contain?
T1: Identify key N-way relationship(s)	What is the most frequent combination of categories involving all variables and how often does it occur? What proportion of the total dataset does this combination account for?	How often do the most frequent combinations of categories involving all variables occur? How many combinations with this frequency are there? Do they share any of the same characteristics? If so, what are they?
T2: Find absolute value and (marginal) frequency for a category or subset of categories	How many children were on board the Titanic? What percentage of the data do the children account for?	How many mushrooms have a pendant ring type? What percentage of the data do they account for?
T3: Compare frequencies of categories or subsets involving different variables	Which category is *more* frequent: “female” or “first” class?	Which category is the *least* frequent out of “convex” cap-shape, “broad” gill-size and “no” bruises?
T4: Find non-conditional probability	What proportion of people on board the Titanic were female passengers (i.e., non-crew) who survived?	What proportion of mushrooms are edible, have a convex or flat cap, and reside in scattered populations?
T5: Find conditional probability	What is the probability (as a percentage) that someone was in first class, given that they were female?	What is the probability (as a percentage) that a mushroom does not have a smooth stalk surface, given that it is edible and has no bruises?
T6: Explore a (binary) response variable with respect to all other variables	Let’s say you are particularly interested in the people who survived the Titanic disaster. Do you notice any trends among this group of people? How about with respect to over- or under-represented groups?	Assume you are particularly interested in edible mushrooms, and you want to avoid the poisonous ones. How many edible mushrooms are there? For which categories/properties can you be certain that a mushroom will be edible rather than poisonous?



edibil... ▼	cap-s... ▼	gill-si... ▼	bruises ▼	stalk-... ▼	ring-... ▼	popu... ▼	habitat ▼	Frequency ▼	Residual ▼
edible	flat	broad	no	fibrous	evane...	scatter...	grasses	96	
edible	convex	broad	no	fibrous	evane...	scatter...	grasses	96	
edible	flat	broad	no	smooth	evane...	scatter...	grasses	96	
edible	convex	broad	no	smooth	evane...	scatter...	grasses	96	
edible	convex	broad	bruises	smooth	pendant	scatter...	grasses	88	
edible	convex	broad	bruises	smooth	pendant	scatter...	mead...	64	
edible	flat	broad	bruises	smooth	pendant	scatter...	paths	24	
edible	convex	broad	bruises	smooth	pendant	scatter...	paths	24	
edible	convex	broad	no	silky	pendant	scatter...	grasses	24	
edible	flat	broad	bruises	smooth	pendant	scatter...	grasses	24	
edible	convex	broad	no	smooth	pendant	scatter...	grasses	24	
edible	flat	broad	bruises	smooth	pendant	solitary	woods	432	
edible	convex	broad	bruises	smooth	pendant	solitary	woods	432	
edible	flat	broad	bruises	smooth	pendant	several	woods	432	
edible	convex	broad	bruises	smooth	pendant	several	woods	432	
poison...	flat	broad	no	silky	large	solitary	paths	108	
poison...	convex	broad	no	silky	large	solitary	paths	108	
poison...	flat	broad	no	silky	large	solitary	grasses	108	
poison...	convex	broad	no	silky	large	solitary	grasses	108	
poison...	flat	broad	no	silky	large	several	paths	108	
poison...	convex	broad	no	silky	large	several	paths	108	
poison...	flat	broad	no	silky	large	solitary	woods	108	
poison...	convex	broad	no	silky	large	solitary	woods	108	

Selected items: 656 (8%)

Items considered: 8,124 (100%)

Selected rows: 11 (7%)

Variables shown: 8 (100%)

☑ List selected rows first

● Standard ○ 100% Bars

☑ edibility

edible ☑

poisonous □

☑ cap-shape

convex ☑

flat ☑

knobbed □

bell □

sunken □

conical □

☑ gill-size

broad ☑

narrow ☑

☑ bruises

no ☑

bruises ☑

☑ stalk-surface-above-

Figure 1.14: The Mushroom dataset highlighting items that are edible, have a convex or flat cap and reside in scattered populations, as per task T4.

1.9.2 Results

We now detail the results of our user study, providing general observations from each phase, a discussion of factors influencing task completion, and a summary of participants' responses to the post-study questionnaire.

The self-reported prior knowledge of our six participants is presented in Table 1.2. All participants indicated at least moderate familiarity with categorical data, with half of them reporting high familiarity. They all described themselves as being very familiar with spreadsheet applications, such as Microsoft Excel and Google Sheets, and possessed at least a basic understanding of statistical concepts. Finally, while most participants considered themselves moderately familiar with visualisations, none identified as an expert in this area.

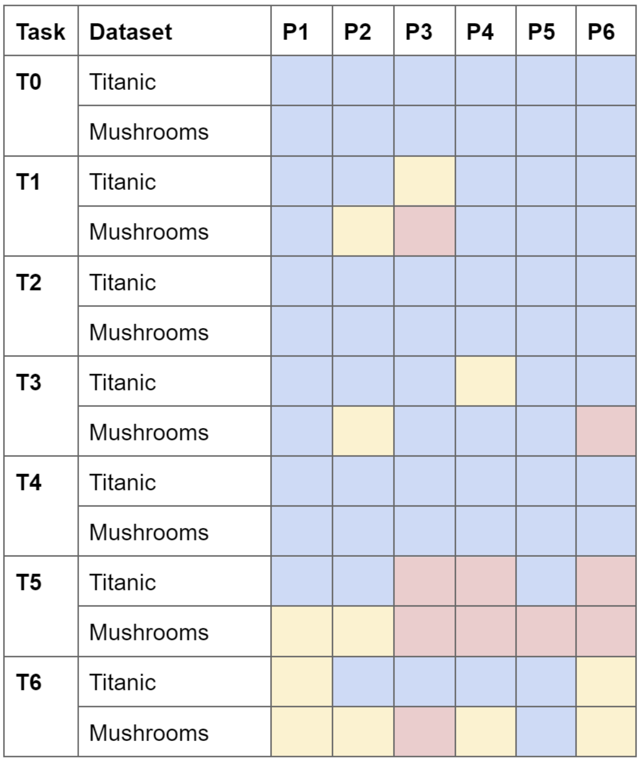
Table 1.2: Summary of participants' self-reported familiarity with relevant tools and concepts.

Topic (1=unfamiliar, 5=extremely familiar)	**Median**	**Mode**
Bar charts	3.5	4
Spreadsheet applications (Microsoft Excel, Google Sheets)	4	4
Visualisations (in general)	3	3
Categorical data	3.5	3, 4
Joint, conditional and marginal probabilities	3	3
Observed frequencies, expected frequencies and deviations	3	3

In the *Exploratory Phase* of the study, five of the six participants quickly and accurately described the key features of MultiCat by themselves, while the sixth participant needed some guidance. For example, one participant commented within a matter of seconds “Ahh, so this [row] is like the combination, so it’s saying 670 people were crew, adult, male and died”. Participants also made relevant observations about the sidebar: “I see you’ve colour-coded the values and this [sidebar] is like a legend to go with it” and, after making a selection, “I would imagine this filled in bit [of each bar] is the data that’s being actually used, and the whole thing is the total amount of data”. Participants explored several interactive features on their own, often correctly deducing that it was possible to sort the data by clicking on the column headers and that deselecting the category checkboxes in the sidebar would remove them from the selection.

Regarding points of confusion, a few participants tried sorting the spreadsheet view by multiple columns but did not find this process intuitive. For example, one participant described the sorting order as “back-to-front”. Two participants tried clicking on the category stickers, expecting this to filter the data, but this feature had not yet been implemented. There were a few instances of “change blindness”, whereby users made a selection and immediately noticed that the combinations in the spreadsheet view had changed, but not the content in the sidebar. However, once they realised the two views were linked, this greatly enhanced their understanding of the interface. The “Filter by selection” button was another source of confusion, as it had no effect when participants clicked it without having made a selection. As one participant noted, “I find the filtering a little confusing, but I think if I used it and played with it, it would make more sense”. Most participants incorrectly assumed that the “Deviation” metric—which is what the “Residual” column was previously called—was based on the standard deviation, until its actual function was clarified. Finally, the tooltips were quite delayed, which resulted in some participants missing relevant information on their first attempt to hover over different components.

Figure 1.15 summarises results from the *Task Phase*, broken down by dataset and participant (P1-P6). For the most part, participants were able to complete tasks quickly and successfully, but they encountered similar issues and consistently struggled with tasks T5 and T6. Among the 14 task iterations, the number of correct answers per participant ranged from 9 to 13, with everyone succeeding at tasks T0, T2 and T4 across both datasets. For correctly solved tasks, participants mostly used the expected strategies given in Appendix B. During the first iteration of T6, for example, participants extracted trends relating to Titanic survivors by first selecting the ‘Survived’ category and then detecting patterns in the spreadsheet view, sometimes sorting by categories and/or residuals to facilitate this process.



Task	Dataset	P1	P2	P3	P4	P5	P6
T0	Titanic	🟦	🟦	🟦	🟦	🟦	🟦
	Mushrooms	🟦	🟦	🟦	🟦	🟦	🟦
T1	Titanic	🟦	🟦	🟨	🟦	🟦	🟦
	Mushrooms	🟦	🟨	🟥	🟦	🟦	🟦
T2	Titanic	🟦	🟦	🟦	🟦	🟦	🟦
	Mushrooms	🟦	🟦	🟦	🟦	🟦	🟦
T3	Titanic	🟦	🟦	🟦	🟨	🟦	🟦
	Mushrooms	🟦	🟨	🟦	🟦	🟦	🟥
T4	Titanic	🟦	🟦	🟦	🟦	🟦	🟦
	Mushrooms	🟦	🟦	🟦	🟦	🟦	🟦
T5	Titanic	🟦	🟦	🟥	🟥	🟦	🟥
	Mushrooms	🟨	🟨	🟥	🟥	🟥	🟥
T6	Titanic	🟨	🟦	🟦	🟦	🟦	🟨
	Mushrooms	🟨	🟨	🟥	🟨	🟦	🟨

Figure 1.15: Matrix of user study results showing tasks as rows, differentiated by dataset, and participants as columns. Blue cells signify correct responses, yellow cells denote partially correct responses (right approach, wrong answer) and red cells signify incorrect responses.

When selecting categories, five participants effectively used the bar shortcut in the sidebar, while the remaining participant preferred to toggle the checkboxes individually. The use of sorting varied among participants, with some relying on it quite heavily and others not using it at all. One participant incorrectly answered two questions in the Titanic dataset after inadvertently scrolling past the top two combinations. At that time, the “Reset” button did not reposition the scrollbar, which meant this had a flow-on effect (this issue has since been addressed; see *Refinements* below).

Participants were sometimes uncertain which part of the interface they should use for specific tasks. They tended to focus on the spreadsheet view, even for T2 and T3, which involved univariate category frequencies and were therefore better suited to the sidebar. This was also the case for T6 in the Mushroom dataset, where participants needed to identify categories unique to edible mushrooms. Most participants attempted to answer this question by manually sifting through the combinations to find categories that were present in the "edible" selection, but absent from the non-selected (i.e., poisonous) data. While entirely possible, a much faster strategy—which one participant employed—was to look for fully opaque categories in the sidebar after selecting "edible" mushrooms.

Another observation is that participants sometimes hid variables that were not directly related to a task's requirements, especially within the Titanic dataset. This may have been motivated by a desire to simplify the visualisation as much as possible. For example, in task T5, which asked about the proportion of females in first class, one participant excluded the variables Age and Fate. While not incorrect, this was not necessary for completing the task, as retaining all categories for non-mentioned variables would not affect the relevant details in the sidebar. In practical scenarios, removing variables can be counterproductive as it reduces the dimensionality of combinations in the spreadsheet view, obscuring potential insight into more complex relationships. However, this did not matter within the context of our study, especially since the display was reset after each task.

The task with the lowest success rate was T5, which required participants to calculate a conditional probability. The most common approach was to select the mentioned categories without applying a filter, then read the resultant probability from the "Selected items" bar. This yielded the correct numerator but an incorrect denominator, leading to an incorrect answer. The "Filter by selection" feature was largely overlooked, being used by only two participants. This perhaps reflects a gap in participants' understanding of conditional probabilities, although there is clearly also room for supporting these better. Interestingly, one participant with a strong background in statistics extracted the numerator and denominator for each conditional probability from the non-filtered display, choosing to give their answer as a fraction. The same participant sometimes manually added the frequencies of relevant combinations rather than selecting them in the sidebar.

Overall, participants performed better with the Titanic dataset and found it much easier to navigate than the Mushroom one. This was to be expected, given that the Mushroom dataset had significantly more categories and variables, and required vertical scrolling in both views. Participants' familiarity with the category-variable relationships in each dataset may have been another important factor, though our study did not control for this.

In the questionnaire, participants rated MultiCat very highly, as shown by their responses in Table 1.3. Most participants commented in the open-ended questions that they liked the appearance of MultiCat and found the interface easy to use and understand. For instance, one participant remarked "MultiCat is very intuitive. I really enjoyed the visual aspects, being able to visually see the categories, the relationships between them, etc.", while another stated "The visualisations made it easy to see data at a glance. The tooltips were really helpful."

Regarding things they disliked, three participants noted that they found it comparatively difficult to navigate the Mushroom dataset, with one participant saying "I guess I found it a little hard to answer questions with the mushroom dataset in terms of finding the categories". However, as another participant observed, this is to be expected when analysing more complicated data: "The more complex interactions were a little tricky on the first try. BUT this is allowing you to visualize and analyze more complex relationships, so it makes sense that it wouldn't be as straightforward as the more simple visualizations. I could imagine this being an incredibly useful tool!"

There were three suggestions for improving MultiCat: (1) provide more informative tooltips; (2) allow the user to formulate queries using the category stickers in the main visualisation; and (3) allow dynamic resizing of column widths to view the full variable and category names, without having to inspect the tooltip. These first two suggestions have been incorporated into the updated prototype, as noted in *Refinements* below.

Table 1.3: Summary of participants' responses to different statements about MultiCat. Values marked with an asterisk have been adjusted to enable direct comparison with other questions, where higher values are better.

Statement (1=strongly disagree, 5=strongly agree)	**Median**	**Mode**
I found MultiCat easy to use.	4	4
I was able to complete the tasks.	4	4
I felt confident using MultiCat.	4	4
I thought that the main visualisation and the sidebar worked well together.	5	5
I thought some features were unnecessarily complicated *(lower is better).*	2 (4*)	2 (4*)
I thought the interactive features (sorting, querying, filtering) were useful.	5	5
I thought the interactive features worked well together.	4.5	4, 5
I think that I would need assistance to use MultiCat again *(lower is better).*	2 (4*)	2 (4*)
I think most people would learn to use MultiCat fairly quickly.	4	4
I would like to use MultiCat again in the future.	4.5	4, 5
Overall rating (1=unusable, 5=exceptional)	5	5

Overall, reflecting on our study, participants found the concept of MultiCat compelling and were enthusiastic about using it again in the future. They succeeded in performing a wide range of tasks, but clearly found some features (like filtering) less intuitive than others. While some issues with the prototype were identified, these do not overshadow MultiCat's potential as a valuable tool for analysing categorical data.

1.9.3 Refinements

Based on observations and feedback elicited from the user study, we made the following changes to the MultiCat prototype, which were already accommodated in our prior explanation of the technique:

1. Renamed the "Deviation" column to "Residual" to avoid confusion with the standard deviation.
2. Added more informative tooltips to the "Frequency" and "Residual" column headers.
3. Extended the query functionality to allow clicking on category stickers within the spreadsheet view.
4. Adjusted the scaling of the category bar charts in the sidebar to enable direct comparison across different variables. Previously, the bars were scaled according to the most frequent category within each variable, rather than the global maximum.
5. Added the two radio buttons to the sidebar, instead of just offering the "Standard" view for category bar charts.
6. Greyed out the "Filter by selection" button when it has no effect.
7. Modified the "Reset" button to reconfigure the vertical scrollbars for the spreadsheet view and sidebar.

1.10 Comparison with existing techniques

In this section, we compare the strengths and weaknesses of MultiCat with two existing techniques: Parallel Sets (Kosara et al., 2006) and (Interactive) Mosaic Plots (Hartigan and Kleiner, 1981; Theus, 2002). We have chosen these techniques for three reasons: (1) they are established methods for visualising multidimensional categorical data; (2) they directly encode cells in contingency tables, rather than employing dimensionality reduction techniques, meaning they are CatViz, not QuantViz, techniques; (Johansson Fernstad and Johansson, 2011) and (3) they preserve higher-order relationships, unlike, for instance, the Heatmap Matrix (Rocha and da Silva, 2018; Trye et al., 2023), Mosaic Matrix (Friendly, 1999), or GPLOM (Im et al., 2013), which only explicitly show pairwise relationships. While MultiCat meets these last two criteria, it differs from Mosaic Plots and Parallel Sets in that it does not use an inherently hierarchical or area-proportional layout. This has important implications for its relative strengths and weaknesses, as discussed below.

All three techniques—MultiCat, Parallel Sets and Mosaic Plots—facilitate quick identification of the most frequent combinations of categories involving N variables. In MultiCat, these combinations are prominently displayed in the top rows of the default spreadsheet view; in Mosaic Plots and Parallel Sets they are shown by the largest tiles and largest parallelograms in the bottom "tier", respectively. Of these techniques, Mosaic Plots have the most intuitive semantic structure, as combinations involving subsets of categories are logically laid out side-by-side. Mosaic Plots are also unique in that the tiles align when variables are independent; (Friendly, 1999) this cannot be so easily discerned from MultiCat or Parallel Sets. However, at the same time, Mosaic Plots scale poorly when there are more than four variables because this means more than two variables have to be plotted on the same axis, increasing the potential for confusion.

MultiCat excels at helping users to identify outliers, namely combinations that were only observed once or a handful of times. By default, these combinations are situated at the bottom of the spreadsheet view, but they can be easily brought to the top by reversing the sort applied to the "Frequency" column. In contrast, identifying such rare combinations in Mosaic Plots and Parallel Sets is more challenging due to their area-proportional layouts, which result in very small tiles and parallelograms. MultiCat overcomes this limitation with its tabular layout, where all rows are of uniform height, guaranteeing their readability.

One limitation that MultiCat shares with Parallel Sets is the inability to display non-observed combinations (i.e., those with a frequency of 0). In certain situations, including sanity checks, it is useful to identify or estimate the number of non-occurring combinations. Some implementations of Mosaic Plots address this by representing non-occurring combinations with a small circle, making them distinguishable (Hofmann, 2000).

As alluded to above, both Parallel Sets and Mosaic Plots employ a hierarchical layout, but they do so in distinct ways. These hierarchies emphasise conditional relationships between variables. In Parallel Sets, the order in which the subsets are derived can easily be ascertained by following the variables from the top tier down to the bottom. Only the bottom tier of a Parallel Sets visualisation shows relationships involving all variables simultaneously. The upper tiers can be useful for revealing interactions among fewer variables, but they occupy additional space and privilege variables that are higher up, potentially biasing the viewer's interpretation. Furthermore, changing the order of variables and/or categories alters the appearance of the display, sometimes drastically, which can in turn influence the insights derived. Similarly, with Mosaic Plots, the order in which variables are split affects what can be seen in the visualisation (Hofmann, 2006). The order in this case is less obvious than in Parallel Sets, but can be deduced from the category labels usually found along the external edges of the display. However, tools like Mondrian only display category labels for the (two) outermost variables, necessitating the use of interactive tooltips to identify labels for nested variables. This can make it difficult to perceive the full structure of the nested data at a glance (Hofmann, 2000).

In contrast, MultiCat allows users to discern patterns without being constrained by a predefined hierarchy. Although the order of categorical variables (columns) might subtly influence the interpretation of combinations, it does not change the content of each combination and therefore does not profoundly impact the display. The initial order of combinations (rows) is determined by the combination frequency rather than by related groups of categories. This approach ensures that variables are treated as equally as possible, unless the user decides to sort the combinations by one or more categorical variables, thereby specifying a hierarchy of their own. If users do want to focus on a particular subset of variables, they can query the data or filter out certain variables. MultiCat thus promotes user-driven exploration of important variables and relationships in a relatively undirected manner.

Extracting complete combinations of categories is arguably more straightforward in MultiCat than either Parallel Sets or Mosaic Plots. MultiCat's use of coloured stickers explicitly names each category within a combination, and this feature remains effective even for large numbers of variables. In contrast, Parallel Sets and Mosaic Plots typically require interactive tooltips for decoding combinations as other strategies are cognitively demanding. Moreover, since tooltips can usually only be accessed one at a time, MultiCat is more efficient for comparing multiple combinations involving a large number of variables at the same time.

In general, it seems easier to accurately compare combination frequencies in MultiCat than the other techniques. The lengths of bars in MultiCat's frequency bar chart are easier to compare than the areas of differently sized and shaped tiles in Mosaic Plots, or the varied angles of parallelograms in Parallel Sets. Parallel Sets also invariably suffer from line crossings, which create visual interference, particularly in datasets with a high diversity of categories and variables. Mosaic Plots, while free from line crossings, are hard to read when there is an abundance of small tiles. MultiCat circumvents these issues since additional combinations can be scrolled vertically if they do not fit in the available screen space, and additional variables can be scrolled horizontally.

Both MultiCat and Mosaic Plots incorporate Pearson residuals, which are valuable for determining whether particular combinations of categories are over- or under-represented in the data. In Mosaic Plots, these residuals are typically shown by applying discrete (Friendly, 1999) or continuous (Zeileis et al., 2007) colour shading to the tiles. This works well for large tiles but not for small ones as it is difficult to make out the colours. Moreover, the different use 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 one has a larger residual. MultiCat achieves a more precise encoding for the residuals by using a diverging bar chart that is separate from the bar chart for frequencies. Any residuals that are difficult to see have smaller absolute values and are therefore less important.

To summarise, MultiCat is useful for identifying combinations of any frequency and Pearson residuals of any size. It differs from the other techniques because it uses a tabular, non-hierarchical layout that does not encode frequencies in an area-proportional way. Parallel Sets clearly shows the order in which a hierarchy is formed, but this may impact what the user perceives in the visualisation. It does not support Pearson residuals and is less efficient for identifying infrequent combinations. Mosaic Pots, on the other hand, do support Pearson Residuals, and exhibit unique features such as the alignment of tiles for independent variables. However, they present readability challenges for small tiles and their colour-based representation of residuals is less perceptually accurate than MultiCat's length-based encoding. Moreover, Mosaic Plots do not scale well to more than four categorical variables. Thus, while each technique has its merits, MultiCat's approach can be seen to offer a more versatile and user-friendly solution for exploring categorical data.

1.11 Possible extensions

We propose the following extensions to MultiCat, which we believe would further enhance its usability:

Direct data manipulation: MultiCat could provide a means of accessing and editing the raw data. Users should be able to modify or delete specific items, merge existing categories, derive new variables from existing ones, and so on.

Heterogeneous datasets: In practice, datasets often comprise both categorical and continuous variables, rather than being limited to only one type (Zhang et al., 2014). Recognising this, MultiCat could be extended to also handle continuous variables. One possible approach is to categorise all continuous variables into bins (Wickham and Hofmann, 2011; Rocha and da Silva, 2018), treating these bins as ordinal categories. This would allow them to be integrated into each combination of categories. Alternatively, drawing inspiration from tools like UpSet (Lex et al., 2014), continuous variables could be kept in their original form and presented alongside categorical combinations as aggregated visualisations, such as box-and-whisker plots. This would afford insights into the extent to which data items within and across combinations of categories vary with respect to their continuous characteristics.

Drill down to individual records: Akin to Taggle (Furmanova et al., 2020), a drill-down feature would enable users to explore individual data items—in mini scrollable tables, for instance—within the spreadsheet view. An expand/collapse icon could be positioned next to each category combination. Clicking on this icon would reveal unique identifiers about the corresponding records, such as passengers' names in the Titanic dataset (one row per record). Additionally, a text search feature could visually highlight matching items. A global “Expand/Collapse All” option could also be included, with the parent combinations always remaining visible.

Response variables: When analysing categorical datasets, users may wish to examine a response variable in relation to all other variables (Agresti, 2012). To facilitate this, the sidebar could incorporate a drop-down menu for selecting a response variable. This menu would list the names of all columns, with “None” selected by default. Upon choosing a response variable, the corresponding column would be removed from the spreadsheet view. Instead, each of its categories would be given individual frequency/proportion and residual columns appearing alongside each combination. To visually distinguish them from other variables, these new bars for the response variable could employ hatching patterns instead of colour. This feature would enhance users' understanding of how categories within the response variable are distributed among combinations of all other variables. For instance, in the Titanic dataset, selecting Fate as the response variable would help to answer questions about mortality rates, such as “Which combinations had significantly more fatalities than survivors (or vice versa)?”

More powerful queries: The MultiCat interface could be adapted to support more expressive queries, by taking inspiration from tools such as ComBiNet (Pister et al., 2023), 2dSearch (Russell-Rose and Gooch, 2018) and AI-STARS (Anick et al., 1989). Currently, users are not able to OR categories across different variables, or create compound queries of arbitrary complexity. Following ComBiNet, it might be helpful to offer a synchronised text-based representation of queries, allowing users to edit *either* the selected visual elements or corresponding text. Additionally, MultiCat could incorporate the ability to save and reload queries, and even allow logical operations to be applied to the queries themselves.

Automatic feature selection: Upon loading a dataset in MultiCat, a “Configuration Selector” tool could be introduced to assist users in strategically choosing variables to include in their analysis. This tool would offer a visual summary of key characteristics of different subsets of variables, providing insights and automatic recommendations about different possible analysis paths. For example, it could display ranked information about the number of distinct combinations for each set of variables and their median frequency. This would be particularly useful for identifying influential relationships, especially in cases where certain variables add undue complexity by fragmenting frequent category combinations. As discussed in the section on Scalability, the impact of a single variable can be substantial, particularly if it encompasses a large number of categories or exhibits significant variation with respect to other variables. Therefore, the “Configuration Selector” would be valuable not only for setting up the initial display, but also for guiding users in making informed decisions about which variables and categories to include in their subsequent explorations.

Missing values: Settings could be added to MultiCat to provide an overview of missing values across all variables and to filter these out in a controlled way.

1.12 Conclusion

This paper has introduced MultiCat, a novel visualisation technique for exploring multidimensional categorical data. MultiCat combines the strengths of a tabular layout with multiple coordinated views, supporting the user in rapid data observation, hypothesis testing and exploratory information seeking. MultiCat distinguishes itself from other techniques through its: (1) high readability of category labels, which notably includes high-dimensional relationships and low-frequency combinations; (2) non-hierarchical default layout; (3) visual summary of individual category contributions; and (4) separate treatment of nominal and ordinal variables. The spreadsheet view provides a comprehensive overview of multidimensional relationships, complemented by sorting operations that enable task-driven analysis of typical observations and outliers alike. The sidebar helps to bridge the gap between individual categories and multidimensional combinations by summarising category distributions and indicating proportions of selected subsets. Furthermore, dynamic queries in MultiCat enable fast computation of absolute values and empirical probabilities, providing a natural and intuitive means of drilling down into the data. We validated MultiCat by conducting a small-scale user study, in which participants rated their experience highly and successfully performed a diverse range of tasks. The results of this study suggest that MultiCat would be a valuable tool for data analysts, while hinting at its advantages over traditional techniques.

Future work could focus on implementing and evaluating the proposed extensions, from direct data manipulation to special treatment of response variables. An in-depth comparative study between MultiCat and established techniques for visualising multidimensional categorical data would also be valuable, in order to better understand the relative strengths and weaknesses of each approach for various analysis tasks.

Acknowledgements

We would like to thank our study participants for generously giving up their time to contribute to this research. We are grateful to Andreea Calude for her comments on the draft. DT thanks the University of Waikato for funding this research through a Doctoral Scholarship.

5.14 References

(2009). parsets. `https://code.google.com/archive/p/parsets/downloads`. Accessed January 25, 2024.

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

Alsallakh, B., Aigner, W., Miksch, S., and Gröller, M. E. (2012). Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data. *IEEE Trans Vis Comput Graph*, 18(12):2849–2858.

Anick, P. G., Brennan, J. D., Flynn, R. A., Hanssen, D. R., Alvey, B., and Robbins, J. M. (1989). A direct manipulation interface for boolean information retrieval via natural language query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, page 135–150, New York, NY, USA. Association for Computing Machinery.

Bartram, L., Correll, M., and Tory, M. (2021). Untidy data: The unreasonable effectiveness of tables. *IEEE Trans Vis Comput Graph*, 28(1):686–696.

Bendix, F., Kosara, R., and Hauser, H. (2005). Parallel Sets: visual analysis of categorical data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 133–140. IEEE.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3: data-driven documents. *IEEE Trans Vis Comput Graph*, 17(12):2301–2309.

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).

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

Dimara, E., Bezerianos, A., and Dragicevic, P. (2017). Conceptual and methodological issues in evaluating multidimensional visualizations for decision support. *IEEE Trans Vis Comput Graph*, 24(1):749–759.

Fischer, M. T., Frings, A., Keim, D. A., and Seebacher, D. (2021). Towards a survey on static and dynamic hypergraph visualizations. In *2021 IEEE visualization conference (VIS)*, pages 81–85. IEEE.

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

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Comput Graph Stat*, 8(3):373–395.

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.

Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). LineUp: Visual analysis of multi-attribute rankings. *IEEE Trans Vis Comput Graph*, 19(12):2277–2286.

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.

Healey, C. G. (1996). Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96*, pages 263–270. IEEE.

Hearst, M. A. (1999). User interfaces and visualization. *Modern information retrieval*, pages 257–323.

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika*, 51:11–26.

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

Hofmann, H. and Vendettuoli, M. (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE Trans Vis Comput Graph*, 19(12):2297–2305.

Im, J.-F., McGuffin, M. J., and Leung, R. (2013). GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Trans Vis Comput Graph*, 19(12):2606–2614.

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.

Kolatch, E. and Weinstein, B. (2001). CatTrees: Dynamic visualization of categorical data using treemaps. `https://cat-vis.github.io/src/data/papers_pdf/kolatch2001cattrees.pdf`.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE Trans Vis Comput Graph*, 12(4):558–568.

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.

Lomuscio, M. (2020). Sleep study. Accessed January 11, 2024.

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

Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAWGraphs: a visualisation platform to create open outputs. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter*, pages 1–5.

Munzner, T. (2014). *Visualization analysis and design.* CRC press.

Muth, L. C. (2021). When to use quantitative and when to use qualitative color scales. Accessed January 11, 2024.

Nguyen, C. H. and Mamitsuka, H. (2020). Learning on hypergraphs with sparsity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2710–2722.

Peña-Araya, V., Xue, T., Pietriga, E., Amsaleg, L., and Bezerianos, A. (2022). HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62.

Pister, A., Prieur, C., and Fekete, J.-D. (2023). ComBiNet: Visual query and comparison of bipartite multivariate dynamic social networks. In *Comput Graph Forum*, volume 42, pages 290–304. Wiley Online Library.

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.

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).*

Russell-Rose, T. and Gooch, P. (2018). 2dSearch: A visual approach to search strategy formulation. In *Proceedings of the 1st Biennial Conference on Design of Experimental Search and Information Retrieval Systems.*

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.*

Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans Vis Comput Graph*, 18(12):2431–2440.

Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE software*, 11(6):70–77.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pages 336–343. IEEE.

Spoerri, A. (1995). *InfoCrystal, a visual tool for information retrieval.* PhD thesis, Massachusetts Institute of Technology.

Symanzik, J., Friendly, M., and Onder, O. (2019). The unsinkable Titanic data. In *2019 Joint Statistical Meetings (ASA) Conference Proceedings*, Denver, USA. [Online]. Available: `https://www.datavis.ca/papers/JSM-2019-proceedings-final.pdf`.

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.

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

Theus, M. (2012). Mosaic plots. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):191–198.

Trye, D. (2022). Visualising multivariate categorical data. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, Tsukuba, Japan.

Trye, D., Apperley, M., and Bainbridge, D. (2023). Extending the Heatmap Matrix: Pairwise analysis of multivariate categorical data. In *2023 27th International Conference Information Visualisation (IV)*, pages 29–36.

Unwin, A., Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive graphics for data sets with missing values—MANET. *Comput Graph Stat*, 5(2):113–122.

Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., and Fekete, J.-D. (2021). Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE Trans Vis Comput Graph*, 27(1):1–13.

Ware, C. (2019). *Information visualization: perception for design.* Morgan Kaufmann.

Wickham, H. and Hofmann, H. (2011). Product plots. *IEEE Trans Vis Com-*

put Graph, 17(12):2223–2230.

Wolfe, J. M. and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):0058.

Young, F. W. and Bann, C. M. (1996). ViSta: The visual statistics system. Technical report, Technical Report 94–1 (c), UNC LL Thurstone Psychometric Laboratory Research ....

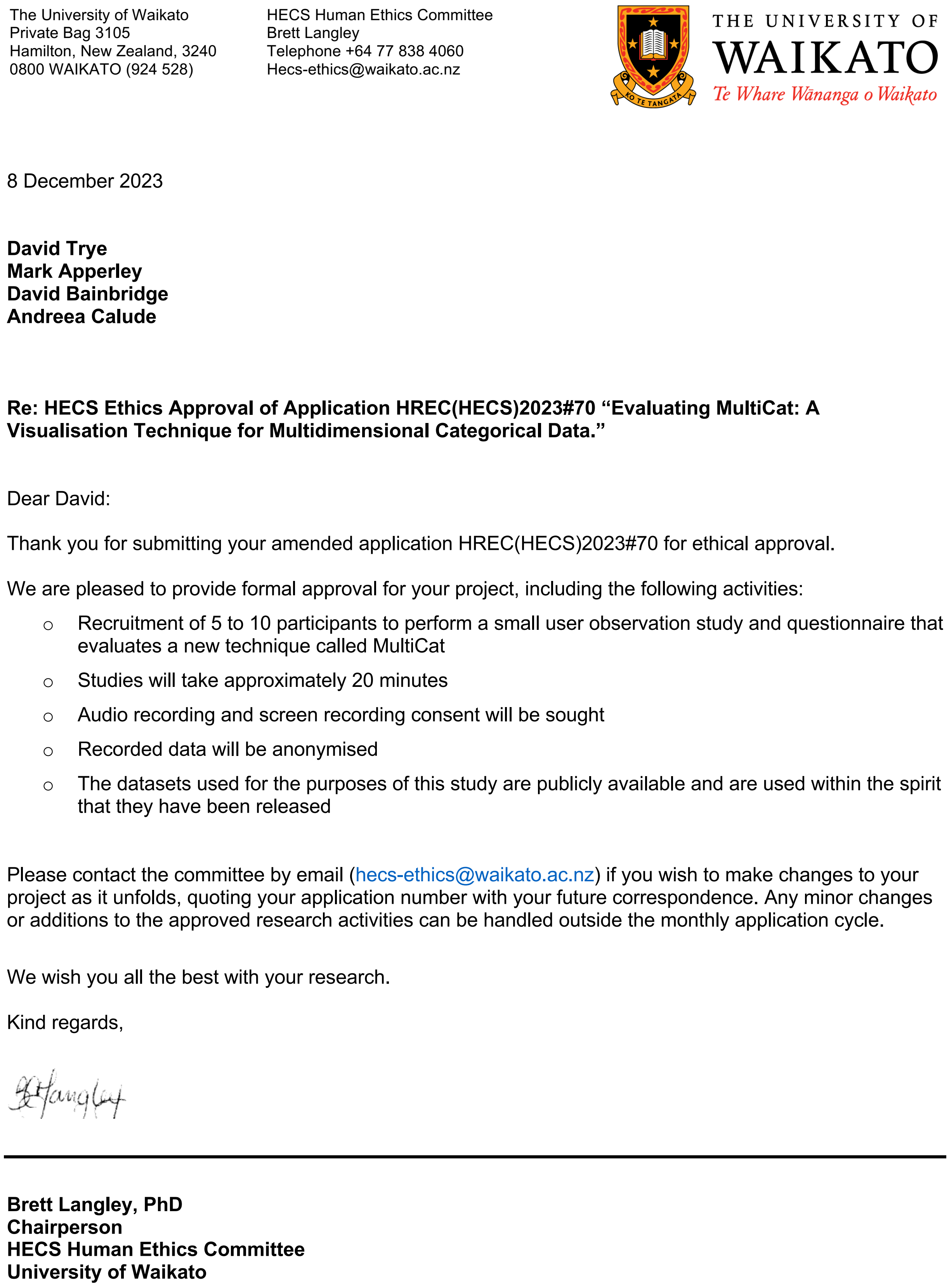
Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Comput Graph Stat*, 16(3):507–525.

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.

Zhang, Z., McDonnell, K. T., Zadok, E., and Mueller, K. (2014). Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Trans Vis Comput Graph*, 21(2):289–303.

Appendix A

Ethics Approval



The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

8 December 2023

David Trye
Mark Apperley
David Bainbridge
Andreea Calude

Re: HECS Ethics Approval of Application HREC(HECS)2023#70 "Evaluating MultiCat: A Visualisation Technique for Multidimensional Categorical Data."

Dear David:

Thank you for submitting your amended application HREC(HECS)2023#70 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

- Recruitment of 5 to 10 participants to perform a small user observation study and questionnaire that evaluates a new technique called MultiCat
- Studies will take approximately 20 minutes
- Audio recording and screen recording consent will be sought
- Recorded data will be anonymised
- The datasets used for the purposes of this study are publicly available and are used within the spirit that they have been released

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

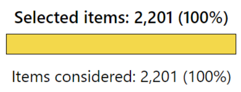
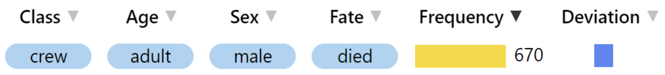
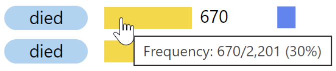
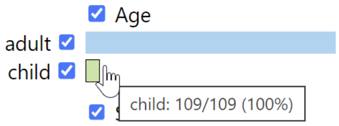
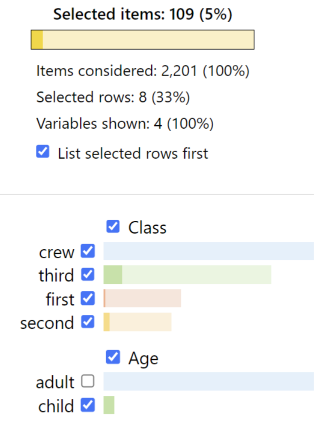
Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato

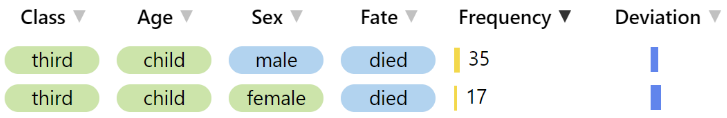
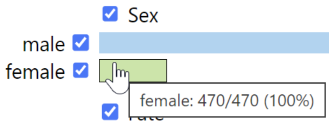
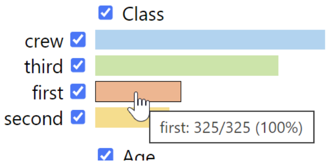
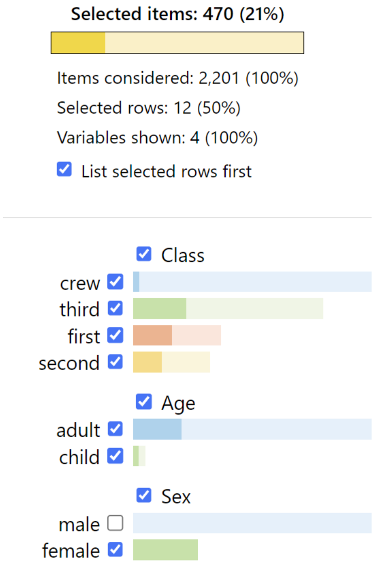
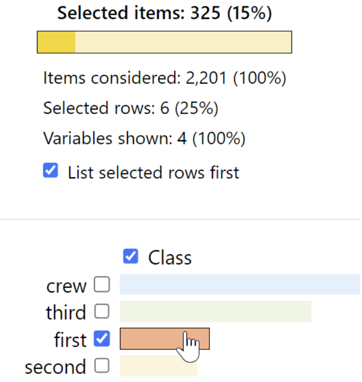
Appendix B

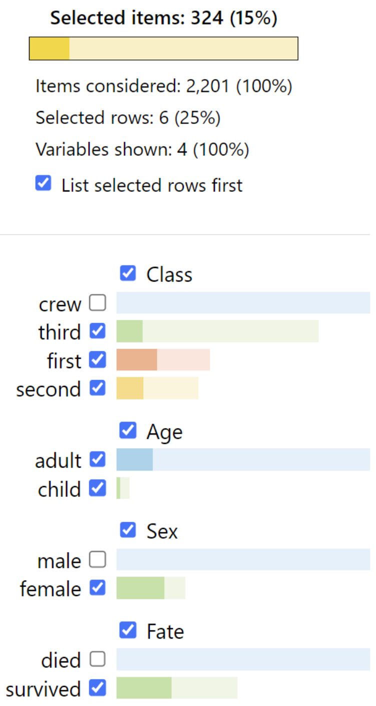
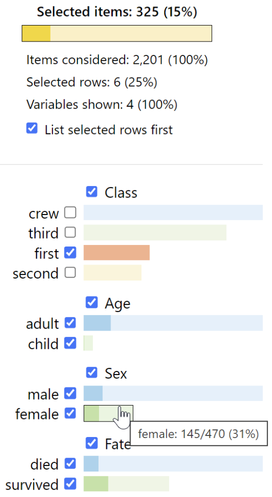
MultiCat User Study Tasks

Supplemental Material: MultiCat User Study Tasks

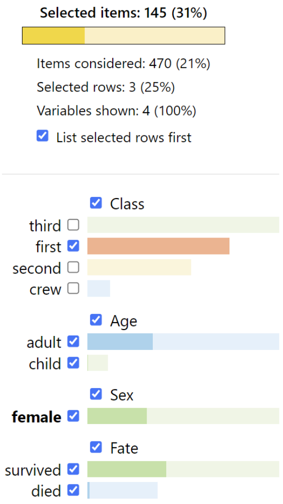
Titanic Dataset (Dawson, 1995):

	Task Description	**Question (Q), Solution/Working (S) and Answer (A)**
T0	Summarise dataset	Q0: How many items (in this case, *people*) does the dataset contain? S: Read directly off the top right of the screen (top of the sidebar).  Selected items: 2,201 (100%) Items considered: 2,201 (100%) A: 2,201
T1	Identify key *N*-way relationship(s)	Q1a: What is the most frequent combination of categories (involving all four categorical variables) and how often does it occur? S: Look at the top-most row of the main visualisation.  Class ▼ Age ▼ Sex ▼ Fate ▼ Frequency ▼ Deviation ▼ crew adult male died 670 A: {crew, adult, male, died}, 670 Q1b: What proportion of the total dataset does this combination account for? S: Hover over the combination's frequency bar to reveal the tooltip, then read the percentage.  died 670 died Frequency: 670/2,201 (30%) A: 30%
T2	Find absolute value and marginal frequency for a particular category	Q2a: How many children were on board the Titanic? S1: Hover over the 'child' category in the sidebar.  ☑ Age adult ☑ child ☑ ☑ child: 109/109 (100%) S2: Select 'child', then look at the 'Selected items' bar chart.  Selected items: 109 (5%) Items considered: 2,201 (100%) Selected rows: 8 (33%) Variables shown: 4 (100%) ☑ List selected rows first ☑ Class crew ☑ third ☑ first ☑ second ☑ ☑ Age adult □ child ☑ S3 (inefficient): Remove all variables except 'Age', then read off the yellow frequency bar for children. A: 109

		Q2b: What percentage of the data do the children account for? S: After S2, read percentage (can't get answer directly from S1). A: 5% Q2c: What do you notice about the two most frequent combinations involving children? S: Look at the top two combinations (assuming data is still sorted by frequency) and identify columns where both stickers are the same. Participants may also comment on the deviation column (now called 'Residual').  Class ▼ Age ▼ Sex ▼ Fate ▼ Frequency ▼ Deviation ▼ third child male died 35 third child female died 17 A: They both relate to children in third class who died (boys first, then girls). Both combinations are slightly over-represented in the dataset.
T3	Compare frequencies of categories belonging to different variables	Q3: Which category is more frequent: 'female' or 'first' class? S1: Hover over tooltips for each category in the sidebar. *Don't* simply compare bar lengths as the bars for each variable are scaled independently (this is no longer the case; users *can* simply compare bar lengths).  ☑ Sex male ☑ female ☑ female: 470/470 (100%) ☑ Fate  ☑ Class crew ☑ third ☑ first ☑ second ☑ first: 325/325 (100%) ☑ Age S2: Select each category in turn and look at the 'Selected items' bar chart.  Selected items: 470 (21%) Items considered: 2,201 (100%) Selected rows: 12 (50%) Variables shown: 4 (100%) ☑ List selected rows first ☑ Class crew ☑ third ☑ first ☑ second ☑ ☑ Age adult ☑ child ☑ ☑ Sex male □ female ☑  Selected items: 325 (15%) Items considered: 2,201 (100%) Selected rows: 6 (25%) Variables shown: 4 (100%) ☑ List selected rows first ☑ Class crew □ third □ first ☑ second □ A: 'female' (470 > 325)
T4	Find non-conditional probability involving multiple categories	Q4: What proportion of people on board the Titanic were female passengers (i.e. non-crew) who survived? S: Select checkboxes for Class={first, second, third} (can just deselect 'crew'), Sex='female' and Fate='survived'. Read proportion from top right of screen.

		 **Selected items: 324 (15%)** Items considered: 2,201 (100%) Selected rows: 6 (25%) Variables shown: 4 (100%) ☑ List selected rows first ☑ Class crew ☐ third ☑ first ☑ second ☑ ☑ Age adult ☑ child ☑ ☑ Sex male ☐ female ☑ ☑ Fate died ☐ survived ☑ A: 15%
T5	Find conditional probability	Q5: What is the probability (as a percentage) that someone was in first class given that they were female? S1: Another way of phrasing this is *what percentage of females were in first class?* Select first class, then hover over females.  **Selected items: 325 (15%)** Items considered: 2,201 (100%) Selected rows: 6 (25%) Variables shown: 4 (100%) ☑ List selected rows first ☑ Class crew ☐ third ☐ first ☑ second ☐ ☑ Age adult ☑ child ☑ ☑ Sex male ☑ female ☑ female: 145/470 (31%) ☑ Fate died ☑ survived ☑

S2: Select females, click ‘Filter by Selection, then select first class.



Selected items: 145 (31%)

Items considered: 470 (21%)

Selected rows: 3 (25%)

Variables shown: 4 (100%)

☑ List selected rows first

☑ Class

third ☐
first ☑
second ☐
crew ☐

☑ Age

adult ☑
child ☑

☑ Sex

female ☑

☑ Fate

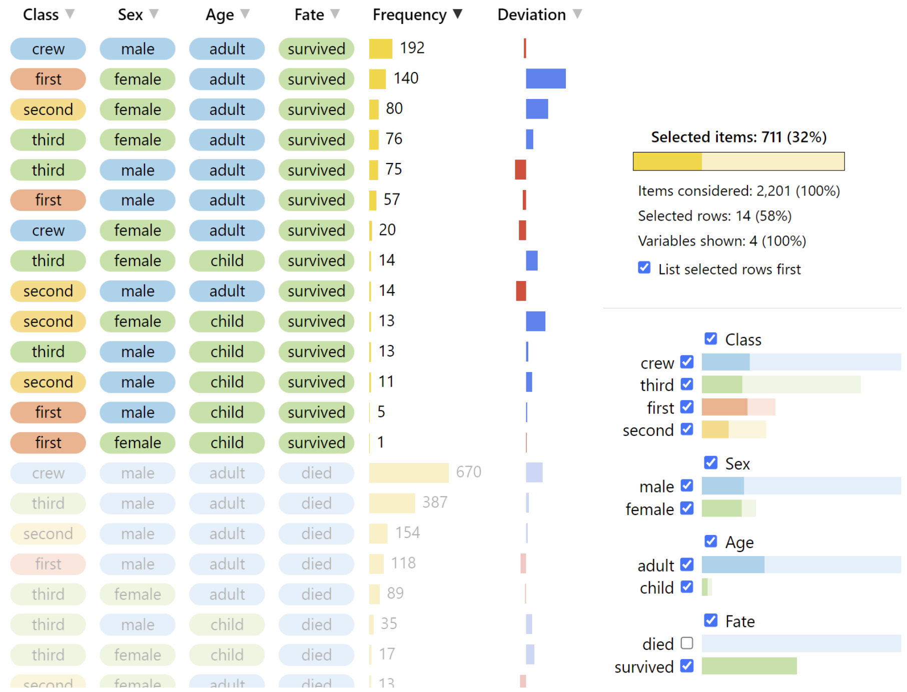
survived ☑
died ☑

A: 145/470 = 31% (not 45% or 7%)

T6

Explore a (binary) response variable w.r.t. all other variables

Q6a: Let’s say you’re particularly interested in the people who *survived* the Titanic disaster. Do you notice any trends among this group of people?
S: Select Fate=‘survived’, then examine selected combinations, as well as the proportion of selected data in the sidebar. Might choose to Filter by Selection (can then see, for instance, that 70% of survivors were passengers (first, second or third class), 30% were crew).



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

Selected items: 711 (32%)

Items considered: 2,201 (100%)

Selected rows: 14 (58%)

Variables shown: 4 (100%)

☑ List selected rows first

☑ Class

crew ☑
third ☑
first ☑
second ☑

☑ Sex

male ☑
female ☑

☑ Age

adult ☑
child ☑

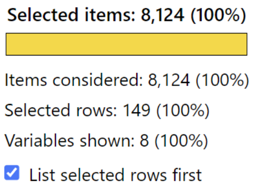
☑ Fate

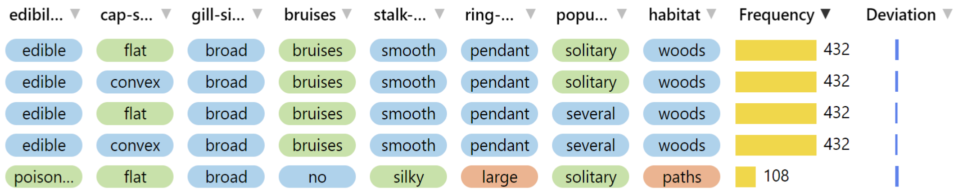
died ☐
survived ☑

The most frequent combinations involve adults. Female class combinations are usually more frequent than corresponding male class combinations (two exceptions being adult crew, which is the most frequent combination, and first-class children). Class is mixed: no obvious trends, but can see female adults are ordered by first, second, third, while children are the opposite (for both sexes), presumably because there were not many children in higher classes. Looking at the sidebar: while a similar *number* of males and females survived, a far greater *proportion* of males died.

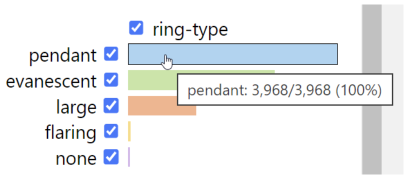
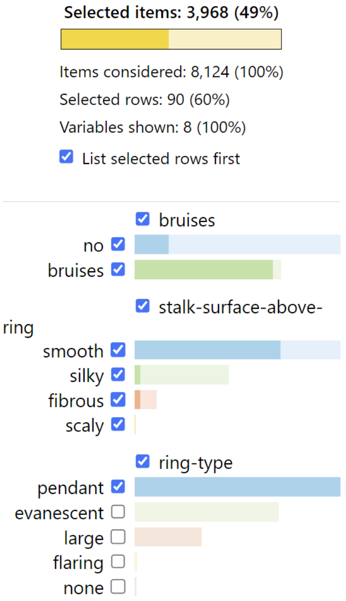
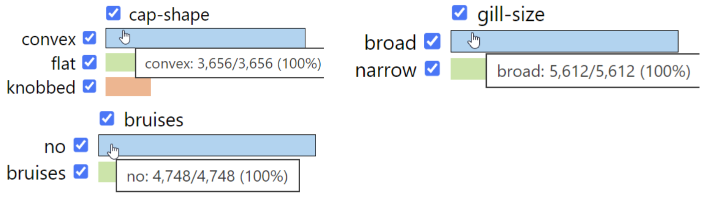
		Q6b: Do you notice any trends among survivors with respect to the deviations (over/under-represented groups)? S: Sort by deviation or manually scan largest values. A: The most over-represented combinations involve female survivors who were passengers (non-crew); conversely, the most under-represented combinations are male adult survivors.

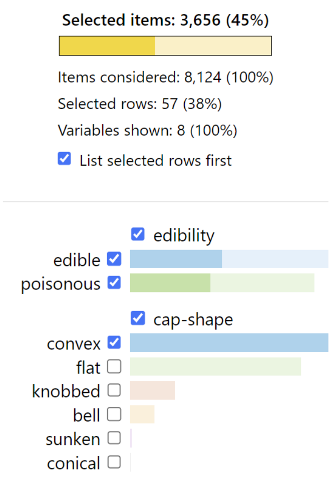
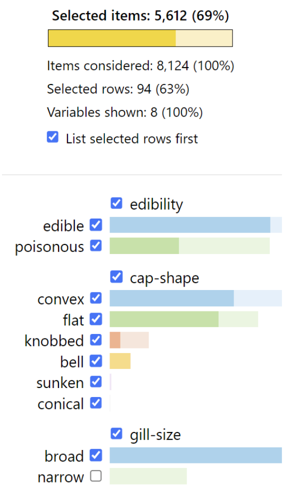
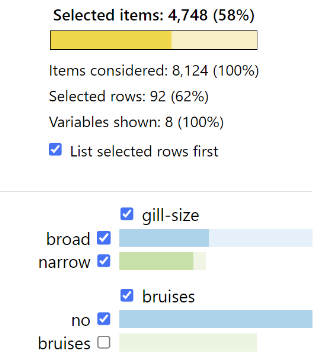
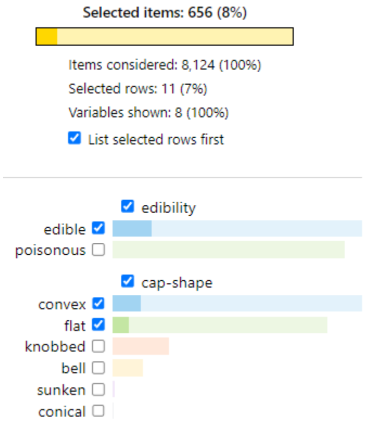
Mushroom Dataset (Condensed): *The order of these questions was randomised.*

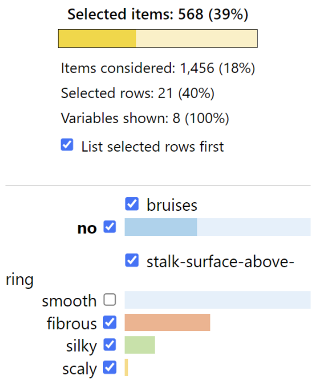
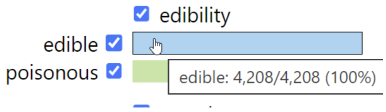
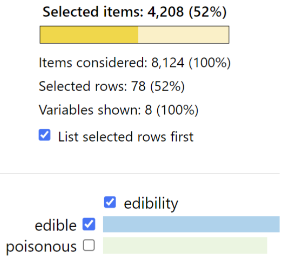
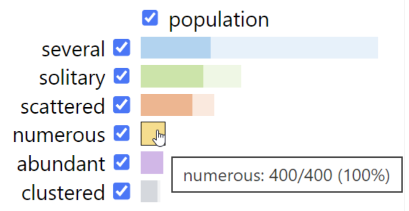
<table>
<tr><th></th><th>Task Description</th><th>Question (Q), Solution (S) and Answer (A)</th></tr>
<tr><td><b>T0</b></td><td>Summarise dataset</td><td>Q0a: How many items (in this case, <i>mushrooms</i>) does the dataset contain?
S: Look at the top right of the screen (top of the sidebar)

<b>Selected items: 8,124 (100%)</b>
Items considered: 8,124 (100%)
Selected rows: 149 (100%)
Variables shown: 8 (100%)
☑ List selected rows first
A: 8124

Q0b: How many categorical variables does the dataset contain?
S: Look at the “Variables shown” metric.
A: 8</td></tr>
<tr><td><b>T1</b></td><td>Identify key <i>N</i>-way relationship(s)</td><td>Q1a: How often do the most frequent combinations of categories occur?
S: Look at the first few rows of the main visualisation (assuming it is still sorted by descending frequency).
A: 432

<table>
<tr><th>edibil...</th><th>cap-s...</th><th>gill-si...</th><th>bruises</th><th>stalk-...</th><th>ring-...</th><th>popu...</th><th>habitat</th><th>Frequency ▼</th><th>Deviation</th></tr>
<tr><td>edible</td><td>flat</td><td>broad</td><td>bruises</td><td>smooth</td><td>pendant</td><td>solitary</td><td>woods</td><td>432</td><td></td></tr>
<tr><td>edible</td><td>convex</td><td>broad</td><td>bruises</td><td>smooth</td><td>pendant</td><td>solitary</td><td>woods</td><td>432</td><td></td></tr>
<tr><td>edible</td><td>flat</td><td>broad</td><td>bruises</td><td>smooth</td><td>pendant</td><td>several</td><td>woods</td><td>432</td><td></td></tr>
<tr><td>edible</td><td>convex</td><td>broad</td><td>bruises</td><td>smooth</td><td>pendant</td><td>several</td><td>woods</td><td>432</td><td></td></tr>
<tr><td>poison...</td><td>flat</td><td>broad</td><td>no</td><td>silky</td><td>large</td><td>solitary</td><td>paths</td><td>108</td><td></td></tr>
</table>
Q1b: How many combinations with this frequency are there?
S: Count the number of combinations whose frequency is 432.
A: 4

Q1c: Do they share any of the same characteristics? If so, what are they?
S: Look for same-coloured stickers in each column for all four combinations.
A: Yes, all four are ‘edible’, ‘broad’, have ‘bruises’, ‘smooth’ stalks, ‘pendant’ rings, and are in ‘woods’ (only variables where categories differ are cap-shape, which is either ‘flex’ or ‘convex’, and population, which is ‘solitary’ or ‘several’)</td></tr>
<tr><td><b>T2</b></td><td>Find absolute value and marginal frequency for a particular category</td><td>Q2a: How many mushrooms have a pendant ring-type?
S1: Hover over the ‘pendant’ category in the sidebar.</td></tr>
</table>

		 ☑ ring-type pendant ☑ evanescent ☑ large ☑ flaring ☑ none ☑ pendant: 3,968/3,968 (100%) S2: Select ‘pendant’, then look at the ‘Selected items’ bar chart.  Selected items: 3,968 (49%) Items considered: 8,124 (100%) Selected rows: 90 (60%) Variables shown: 8 (100%) ☑ List selected rows first ☑ bruises no ☑ bruises ☑ ☑ stalk-surface-above-ring smooth ☑ silky ☑ fibrous ☑ scaly ☑ ☑ ring-type pendant ☑ evanescent ☐ large ☐ flaring ☐ none ☐ A: 3968 Q2b: What percentage of the data do they account for? S: After S2, read percentage (can’t get answer directly from S1). A: 49%
T3	Compare frequencies of categories belonging to different variables	Q3: Which category is the *least* frequent out of convex (cap-shape), broad (gill-size) and *no* bruises (bruises)? S1: Hover over tooltips for each category in the sidebar. *Don’t* simply compare bar lengths as the bars for each variable are scaled independently.  ☑ cap-shape convex ☑ flat ☑ knobbed ☑ convex: 3,656/3,656 (100%) ☑ gill-size broad ☑ narrow ☑ broad: 5,612/5,612 (100%) ☑ bruises no ☑ bruises ☑ no: 4,748/4,748 (100%)

		S2: Select each category in turn and look at the ‘Selected items’ bar chart.  Selected items: 3,656 (45%) Items considered: 8,124 (100%) Selected rows: 57 (38%) Variables shown: 8 (100%) ☑ List selected rows first ☑ edibility edible ☑ poisonous ☑ ☑ cap-shape convex ☑ flat □ knobbed □ bell □ sunken □ conical □  Selected items: 5,612 (69%) Items considered: 8,124 (100%) Selected rows: 94 (63%) Variables shown: 8 (100%) ☑ List selected rows first ☑ edibility edible ☑ poisonous ☑ ☑ cap-shape convex ☑ flat ☑ knobbed ☑ bell ☑ sunken ☑ conical ☑ ☑ gill-size broad ☑ narrow □  Selected items: 4,748 (58%) Items considered: 8,124 (100%) Selected rows: 92 (62%) Variables shown: 8 (100%) ☑ List selected rows first ☑ gill-size broad ☑ narrow ☑ ☑ bruises no ☑ bruises □ A: convex (3656 vs 5612 vs 4748)
T4	Find non-conditional probability involving multiple categories	Q: What proportion of mushrooms are edible, have a convex or flat cap, and reside in scattered populations? S: Select checkboxes for ‘edible’, cap-shape={convex, flat} and population=‘scattered’. Read proportion from top right of screen.  Selected items: 656 (8%) Items considered: 8,124 (100%) Selected rows: 11 (7%) Variables shown: 8 (100%) ☑ List selected rows first ☑ edibility edible ☑ poisonous □ ☑ cap-shape convex ☑ flat ☑ knobbed □ bell □ sunken □ conical □ A: 8% (656)

T5	Find conditional probability	Q5: What is the probability (as a percentage) that a mushroom does not have a smooth stalk surface, given that it is edible and has no bruises? S: Another way of phrasing this is *what percentage of edible mushrooms with no bruises did not have a smooth stalk surface?* We can't use the S1 approach from the Titanic dataset above as we want 'not smooth' and there is no way of hovering over a single merged category representing all other categories. As such, we have to select 'edible', 'no', then 'Filter by selection', then select all but 'smooth' for stalk surface.  Selected items: 568 (39%) Items considered: 1,456 (18%) Selected rows: 21 (40%) Variables shown: 8 (100%) ☑ List selected rows first ☑ bruises no ☑ ☑ stalk-surface-above-ring smooth □ fibrous ☑ silky ☑ scaly ☑ A: 39% (568)
T6	Explore a (binary) response variable w.r.t. all other variables	Q6a: Let's say you're particularly interested in *edible* mushrooms (and you want to avoid the poisonous ones). How many edible mushrooms are there? S1: Hover over 'edible' category. S2: Select 'edible' category, then look at "Selected items".  ☑ edibility edible ☑ poisonous ☑ edible: 4,208/4,208 (100%)  Selected items: 4,208 (52%) Items considered: 8,124 (100%) Selected rows: 78 (52%) Variables shown: 8 (100%) ☑ List selected rows first ☑ edibility edible ☑ poisonous □ A: 4208 Q6b: For which categories/properties can you be certain that a mushroom will be edible rather than poisonous? S: It is not (currently) possible to isolate these categories with a single query. Don't 'Filter by selection' as this means you lose sight of categories that overlap with poisonous mushrooms. Instead, select poisonous mushrooms and look for category bars that are fully opaque, meaning 100% of the category is selected; users can hover over each category in turn to ascertain whether this is the case, which is helpful for smaller bars (in updated prototype, the '100% bars' radio button is useful here)  ☑ population several ☑ solitary ☑ scattered ☑ numerous ☑ abundant ☑ clustered ☑ numerous: 400/400 (100%) A: sunken, flaring, numerous, abundant, waste