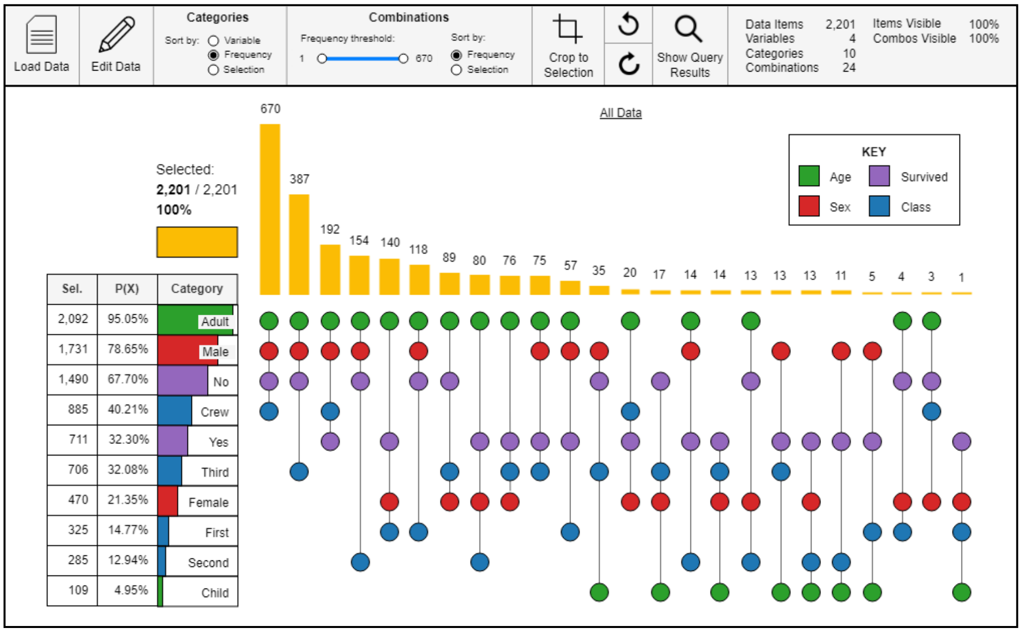
Visualising Multivariate Categorical Data

David Trye*

University of Waikato



Categories Sort by: ○ Variable ● Frequency ○ Selection

Combinations 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: MultiCat visualisation of the Titanic dataset, drawn using diagrams.net.

Abstract

Despite categorical dimensions being common in real-world datasets, few visualisation techniques support the analysis of multiple categorical variables at the same time. Those methods that do exist do not scale well, or do not consider relationships between all variables simultaneously, instead breaking them down into more restricted views or reflecting a hierarchy of variables. Drawing inspiration from set-based tools, this paper introduces a novel technique for visualising multivariate categorical data, by aggregating different combinations of categories. Advantages of this approach include the ability to easily compare frequencies among both variable categories and their combinations, the absence of line crossings, and a non-hierarchical layout that does not privilege one variable above all others.

Keywords: Categorical data, multivariate data, multidimensional data, interaction, set-based data.

1 Introduction and Motivation

Categorical data are prevalent in a wide range of disciplines, including the behavioural and social sciences, public health, biomedical science, education and marketing [1]. Furthermore, given datasets often contain a mixture of data types, categorical data can provide important context for understanding continuous variables [2]. In spite of this, few visualisation techniques exist for properly dealing with categorical data, and fewer still facilitate analysis of a large number of categorical variables.

Mosaic Plots are a prominent technique for visualising categorical data [3]. A major limitation, however, is that they quickly become difficult to read when displaying more than three variables. *Parallel Sets* [2] offer a more scalable alternative, capable of handling 10-15 categorical variables in an interactive environment; however, the technique invariably suffers from line crossings and perceptual distortions [4]. These disadvantages are partially addressed by *Hammock Plots* [5] and *Common Angle Plots* [4], but as with the other two techniques, the order in which dimensions are plotted can drastically alter the display. Motivated by these limitations, we propose *MultiCat*, a novel technique for visualising and interactively exploring large categorical datasets.

2 The MultiCat Technique

In categorical datasets, it is likely that certain combinations of orthogonal variables will occur more frequently than others. The MultiCat technique leverages this fact to reveal high-dimensional category associations. Figure 1 shows how the data are aggregated in two distinct ways. First, frequency information is given for all categories in the table summary on the left-hand side. Second, the frequency of all combinations involving those categories is shown in the main display. The main visualisation can be thought of as a matrix in which 'rows' represent categories and 'columns' represent combinations. Colour is used to denote category membership, thereby helping to distinguish variables and reinforce connections between them.

We exemplify MultiCat using the well-known Titanic dataset [6] in Figures 1 and 2. When interpreting these figures, the table summary provides an important overview of category distributions. For each category, the number of currently selected items is shown, alongside the marginal probability of the category and a bar chart that visually encodes this information, using the fixed width of the column to represent the entire dataset. The table in Figure 1 helps highlight that the crew accounted for a surprisingly large proportion of people on board, that only a small

* dgt12@students.waikato.ac.nz

proportion of passengers were women and children, and that the people on board were twice as likely to die than survive.

Turning to the main display, each distinct combination of categorical values is depicted as a vertical line, with circles showing connections between variables. Being a fully orthogonal dataset, every combination contains exactly four circles, one of each colour. The yellow bars above combinations encode their frequency, and these are sorted in descending order from left to right. Of course, these bars could be re-coloured to encode additional information or to emphasise a particular variable of interest. In the figure, we see that the largest group of people on board the Titanic (which considers all four variables) was 670 male adult crew members who tragically died, followed by 387 male adult third class passengers who suffered the same fate. The least frequent combinations (on the right-hand side) reveal outliers, such as one girl in first class who survived (she was, in fact, the *only* girl in first class), and three female crew members who did not. While all relationships can be read directly from the graph, interaction can play a critical role in reducing the user's cognitive load, by showing only *relevant* connections (Figure 2).

In terms of scalability, spatial requirements for MultiCat are correlated with 'categorical diversity' (number of categories and observed combinations) rather than number of data items.

3 Connection to Set-Based Techniques

MultiCat is inspired by two existing visualisation techniques: *PAOHVis* [7] and *UpSet* [8]. Respectively, these tools are designed for analysing sets (hypergraphs) and set intersections, rather than visualising categorical data. However, some mapping of concepts between these approaches is possible for the purposes of comparison. Key differences with MultiCat include how the horizontal bar chart is used, the use of colour to distinguish variables and the potential for advanced statistical queries.

In PAOHVis terms, the main display is a *hypergraph* or family of sets, in which categories are vertices/elements and combinations are hyperedges/sets. More precisely, the display is a *k-uniform multi-hypergraph*, because repeated hyperedges are permitted and all hyperedges contain exactly *k* vertices, where *k* is the number of variables under investigation. Unlike PAOHVis, we automatically consolidate identical hyperedges and encode their frequency using a bar chart (rather than line thickness), so that all combinations have the same width.

In UpSet, each line depicts a set intersection corresponding to an individual segment of a Venn diagram. Applying this conceptual model, each category is a set, and combinations are set intersections. MultiCat constitutes a restricted use case of UpSet, whereby only certain intersections are possible: specifically, *k*-set exclusive intersections involving one category per variable. The 'cardinality' of a set intersection relates to the frequency of a particular category combination.

4 Interaction

We envisage MultiCat supporting a rich set of interaction methods, two of which are described below. Other methods not covered here include filtering and advanced statistical queries.

4.1 Selection and Highlighting

Figure 2 illustrates how category selections could help highlight relationships in MultiCat. Combinations matching a selection are accentuated with thick lines, and all other combinations are faded out. Category bars and numerical summaries are also updated to reflect selected category proportions. The horizontal yellow bar at the top left of the display shows the proportion of selected items across the entire dataset. When multiple categories are selected, the system takes the intersection of categories across *different* variables, and the union of categories across the *same* variable. Selecting one category for a variable negates all its other levels unless they are also manually selected. This functionality supports complex queries, enabling, for instance, a comparison of the mortality rate of male passengers with male crew members.

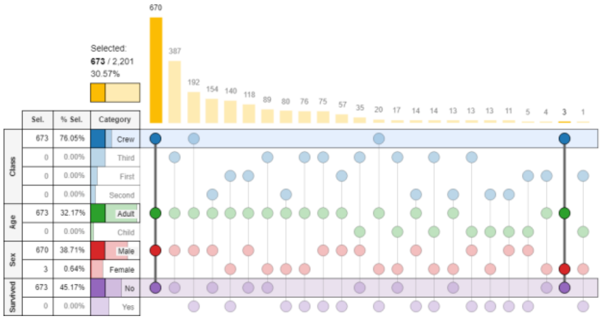


Figure 2: Highlighting in MultiCat.

4.2 Sorting

Sorting can be used to emphasise relationships and better utilise the available screen space. Users can sort categories along the y-axis in three ways: 1) by variable, then frequency; 2) by overall frequency; 3) by category counts for the current selection. Similarly, combinations can be sorted along the x-axis by overall frequency, or so that selected combinations appear on the left-hand side, and non-selected ones to their right.

5 Conclusion

This paper has introduced MultiCat, a visualisation technique for exploring inter-variable associations alongside categorical distributions. By drawing connections with PAOHVis and UpSet, we have shown that categorical data can be fruitfully conceptualised as a special kind of set-based problem. MultiCat overcomes limitations of traditional displays like Mosaic Plots, by enabling a clear comparison of frequencies and avoiding a hierarchy of variables. Future work will focus on developing a fully functional prototype, testing it with datasets of varying complexity, and making changes based on stakeholder feedback.

Acknowledgements

Thanks to Mark Apperley, David Bainbridge and Andreea Calude for their close reading of the paper, and to the University of Waikato for funding this research.

References

[1] A. Agresti. *Categorical data analysis*, 3rd ed. John Wiley & Sons, 2013.

[2] R. Kosara, F. Bendix, and H. Hauser. "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 4, pp. 558–568, 2006.

[3] M. Theus. "Mosaic plots," *WIREs Comput Stat*, vol. 4, no. 2, pp. 191-198. 2012. doi: 10.1002/wics.1192

[4] H. Hofmann, and M. Vendettuoli, "Common angle plots as perception-true visualizations of categorical associations," *IEEE Trans. Vis. Comput. Graphics*, vol 19, no. 12, pp. 2297-2305. 2013.

[5] M. Schonlau, "Visualizing categorical data arising in the health sciences using hammock plots," in *Proc. of the Sect. on Stat. Graphics*, 2003.

[6] R. J. Dawson. The 'unusual episode' data revisited. *Journal of Statistics Education*, 3(3), 1995.

[7] P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, and J.-D. Fekete, "Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, pp. 1–13, 2021, doi:10.1109/TVCG.2019.2933196.

[8] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "UpSet: Visualization of intersecting sets," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.