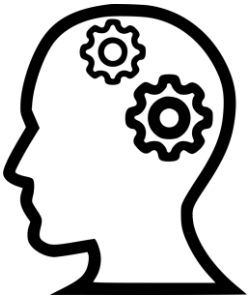
Extending the Heatmap Matrix: Pairwise Analysis of Multivariate Categorical Data

David Trye, Mark Apperley & David Bainbridge | University of Waikato, New Zealand

1

Motivation & Aim

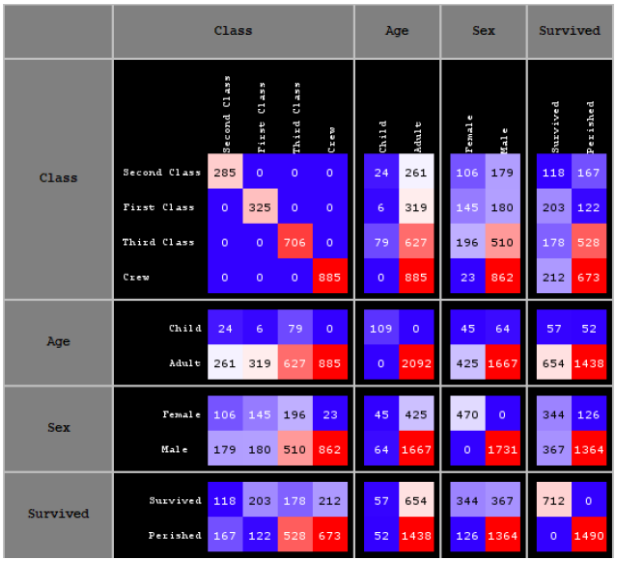


- **Categorical variables** are common in real-world datasets (Agresti, 2013; Friendly & Meyer, 2016)
- Few visualisation techniques facilitate *multiple* categorical variables
- Existing techniques have limited scalability and interaction, and are under-developed compared to those for continuous data
- Our aim is to build upon the strengths of an existing technique: the *Heatmap Matrix*

2

Heatmap Matrix (Rocha & da Silva, 2018; 2022)

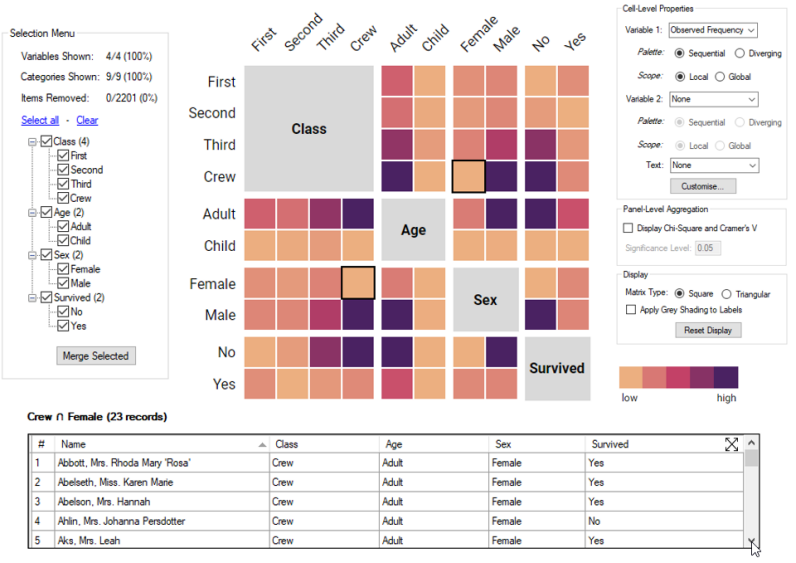
- Displays all possible $n \times m$ contingency tables as heatmaps
- Categories are grouped by variable
- Each (non-diagonal) heatmap 'panel' represents a distinct pair of variables
- Each 'cell' shows the frequency of the corresponding categories



3

Heatmap Matrix Explorer

- Matrix View
- Main Menu
- Selection Menu
- Linked Table View



4

Heatmap Matrix Explorer

- **Matrix View**
- Main Menu
- Selection Menu
- Linked Table View

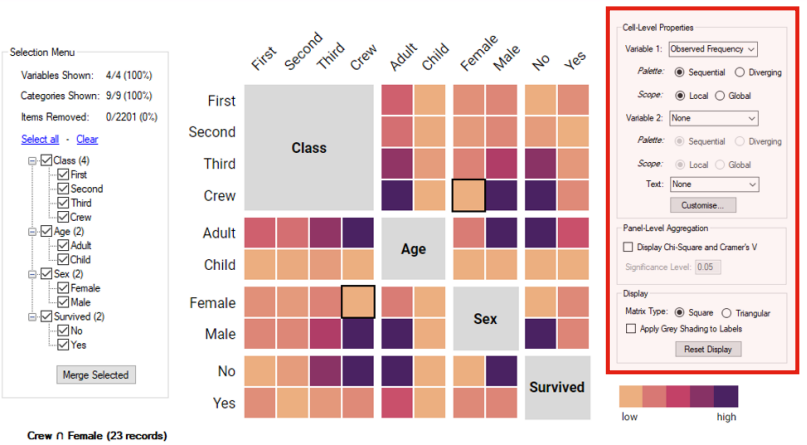
Crew ∩ Female (23 records)

#	Name	Class	Age	Sex	Survived
1	Abbott, Mrs. Rhoda Mary 'Rosa'	Crew	Adult	Female	Yes
2	Abelseth, Miss. Karen Marie	Crew	Adult	Female	Yes
3	Abelson, Mrs. Hannah	Crew	Adult	Female	Yes
4	Ahlin, Mrs. Johanna Persdotter	Crew	Adult	Female	No
5	Aks, Mrs. Leah	Crew	Adult	Female	Yes

5

Heatmap Matrix Explorer

- Matrix View
- **Main Menu**
- Selection Menu
- Linked Table View



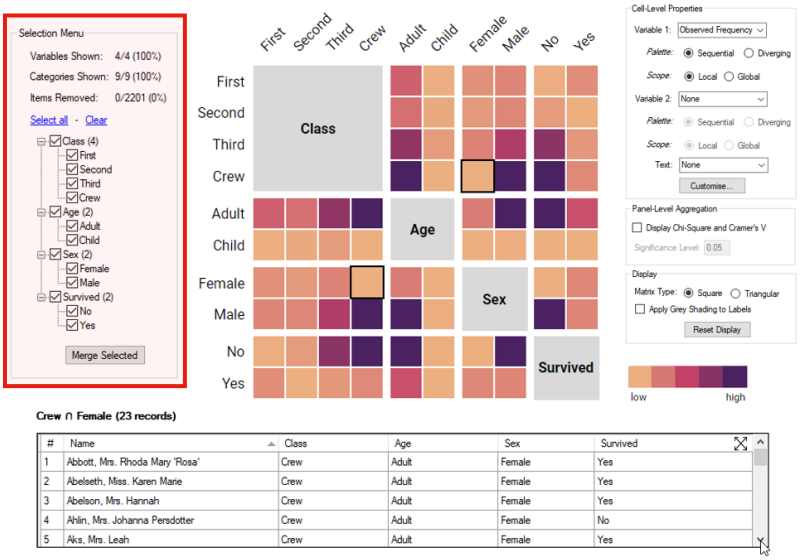
Crew ∩ Female (23 records)

#	Name	Class	Age	Sex	Survived
1	Abbott, Mrs. Rhoda Mary 'Rosa'	Crew	Adult	Female	Yes
2	Abelseth, Miss. Karen Marie	Crew	Adult	Female	Yes
3	Abelson, Mrs. Hannah	Crew	Adult	Female	Yes
4	Ahlin, Mrs. Johanna Persdotter	Crew	Adult	Female	No
5	Aks, Mrs. Leah	Crew	Adult	Female	Yes

6

Heatmap Matrix Explorer

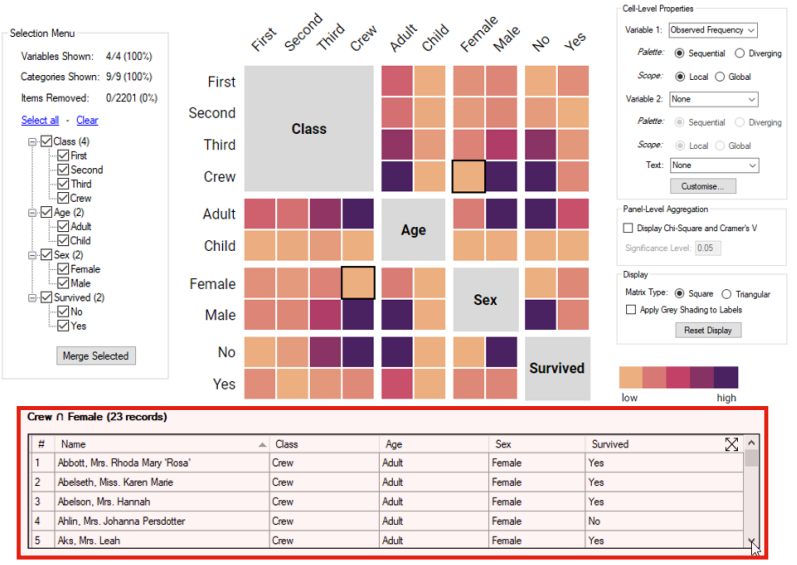
- Matrix View
- Main Menu
- **Selection Menu**
- Linked Table View



7

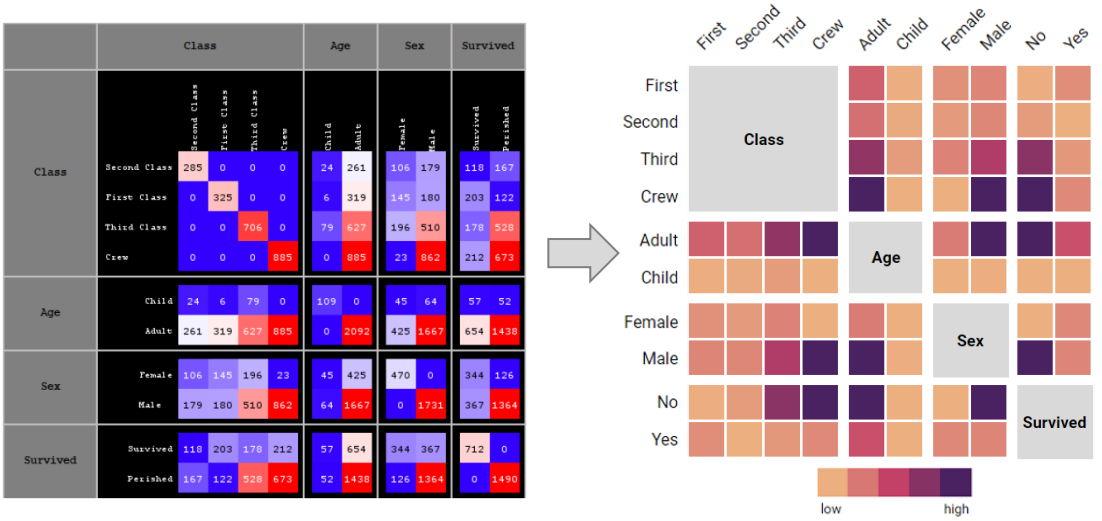
Heatmap Matrix Explorer

- Matrix View
- Main Menu
- Selection Menu
- **Linked Table View**



8

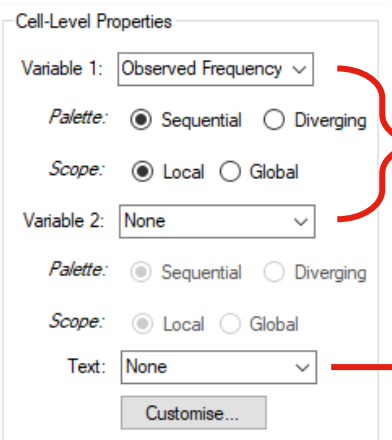
Matrix View: Layout Refinements



9

Main Menu: Supported Metrics

1. Observed Frequency
2. Expected Frequency
3. Row Percentages
4. Column Percentages
5. Pearson Residuals
6. Cell Chi-Square Values

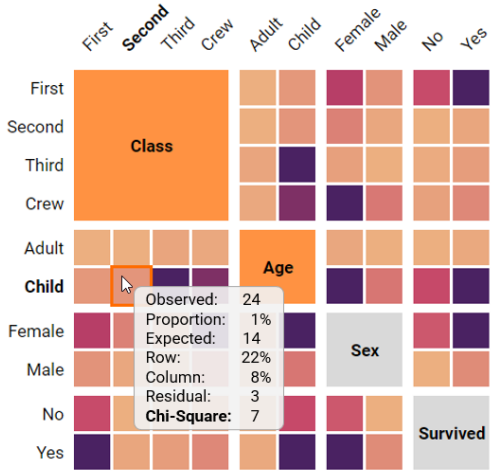


Colour (1-2 metrics)

Cell labels (optional)

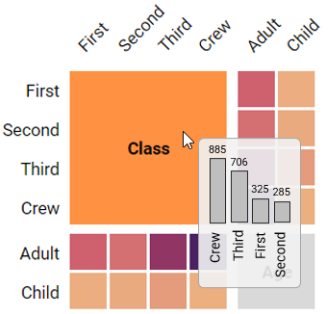
10

Tooltips & Associative Highlighting



'Cell' Tooltip

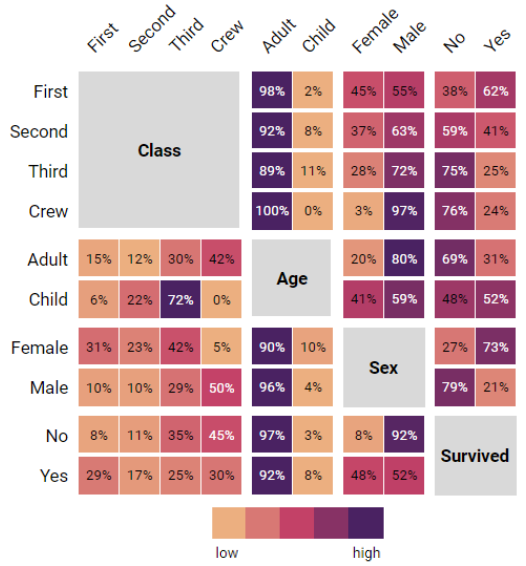
- Provides details-on-demand (Shneiderman, 1996)
- Helps to orient the viewer



'Box' Tooltip

11

Row Percentages

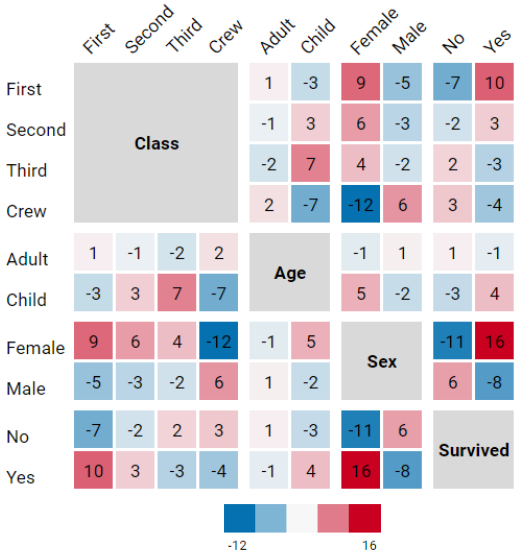


- Rows within each panel sum to 100%
- Each cell shows P(X|Y), where X and Y are the categories on each axis
- Interpretation: What percentage of Y is X?
- "Column Percentages" is simply the transpose

12

Pearson Residuals

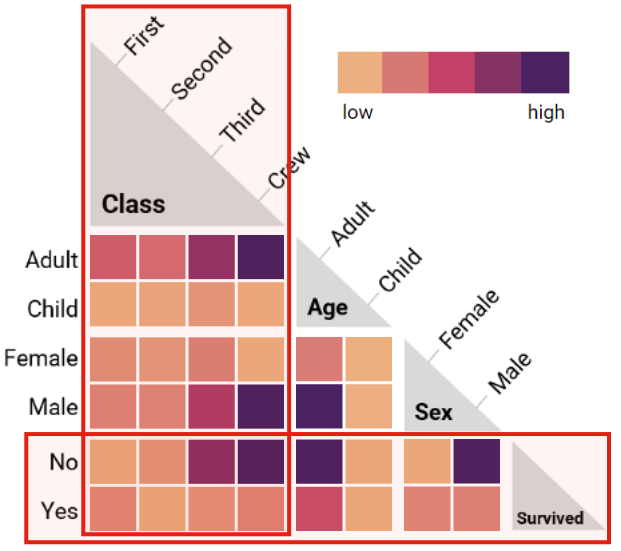
- Shows deviations from independence
- Formula: $r_{ij} = (O_{ij} - E_{ij}) / \sqrt{E_{ij}}$
- Diverging colour palette shows both magnitude and direction, like in a Mosaic Matrix (Friendly, 1999)
- Large residuals (in either direction) may warrant further investigation



13

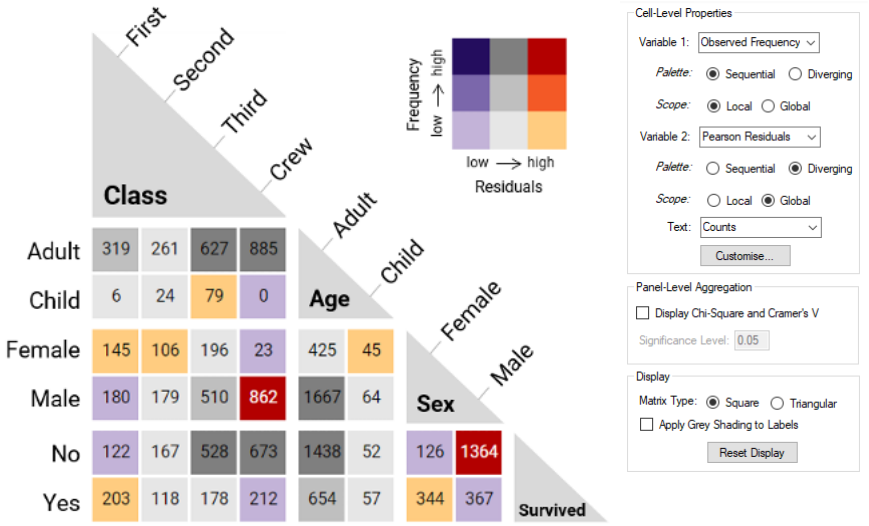
Triangular Matrix

- Less cluttered
- Removes redundant information
- Outer variables are special cases
- Square matrix is a better choice for comparing several categories or variables



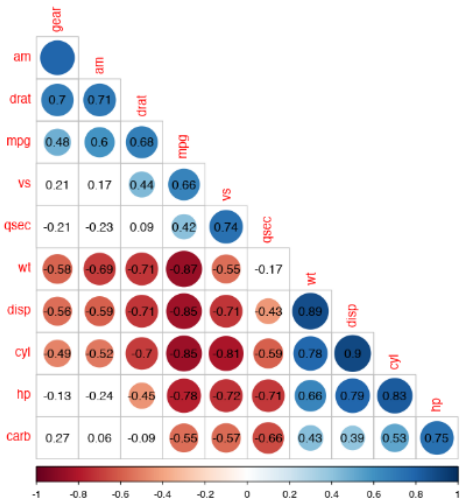
14

Bivariate Colour Map



15

Visualising Statistical Tests



- Statistical test results can be represented graphically
- For instance, *corrgrams* (Friendly, 2002) can be used to show correlation coefficients and *p*-values for continuous data

https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

16

Chi-Square Test of Independence

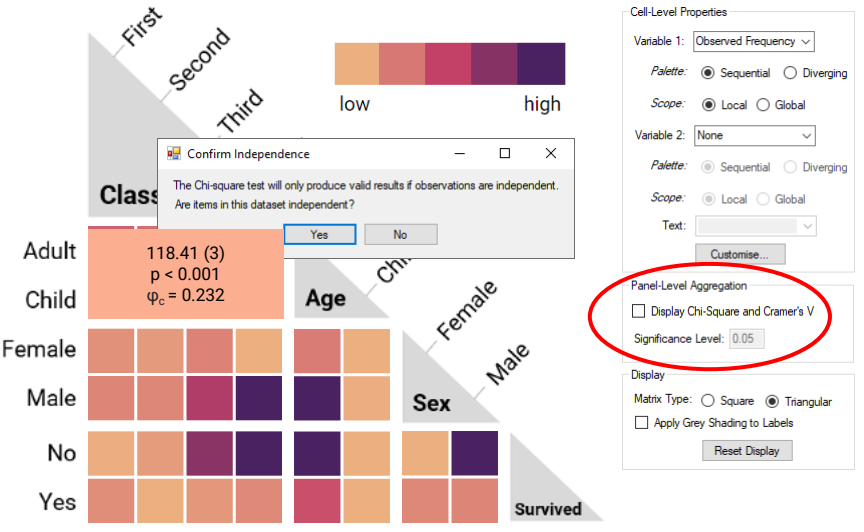
1. Is there a *significant* association between the two variables?
   → Chi-square test
2. If so, how *strong* is it?
   → Cramér's V

Four test conditions:

1. Categorical variables (preferably nominal)
2. Independent observations
3. Mutually exclusive categories
4. *Expected* frequency >0 in all cells
   >4 in at least 80% of cells

17

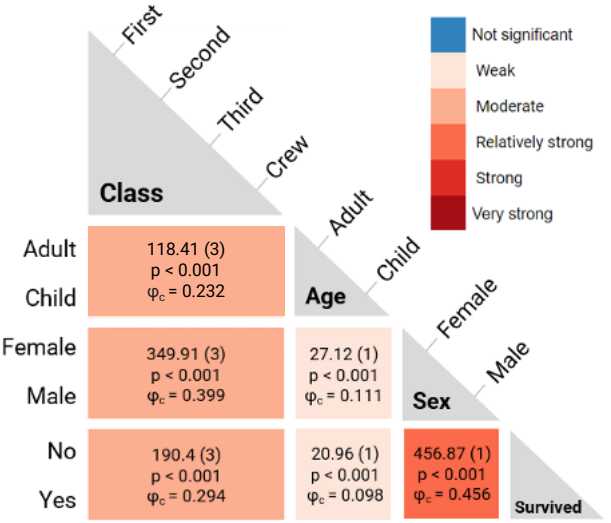
Panel-Level Aggregation



18

Advantages of Panel-Level Aggregation

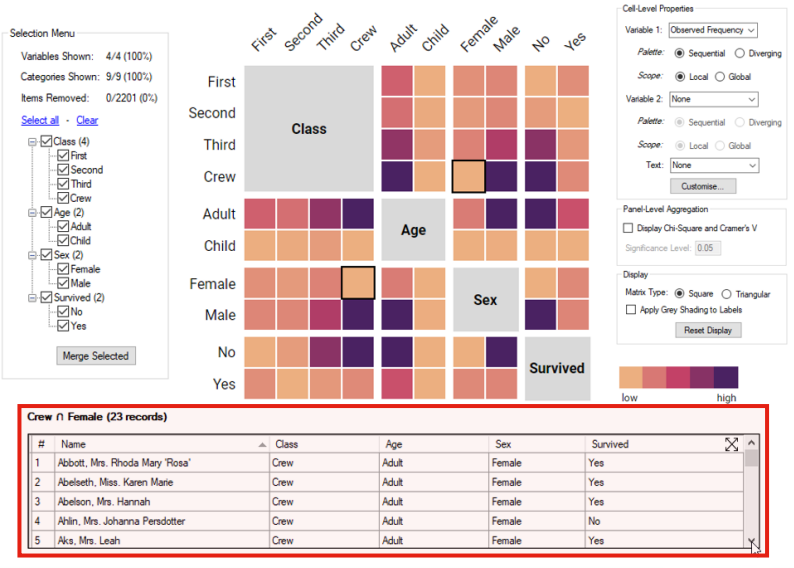
1. Visually reinforces correct interpretations
2. Removes burden of manual computation
3. Groups all test results in one place
4. Can be used to generate hypotheses about cell-level associations, which can then be explored in other views



19

Linked Table View

- Interface between heatmap and underlying data
- Useful for displaying ID-type values, especially when cells have low counts



20

Covid Directives Dataset (Burnette & Calude, 2022)

- 754 directives (e.g. "Stay home!") from tweets containing *#covid19nz*

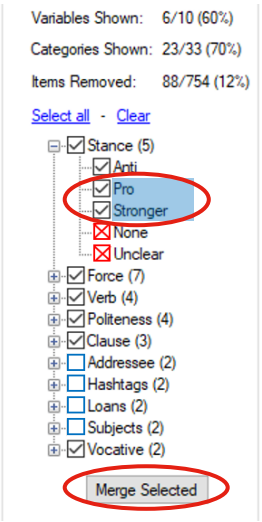
Variable	Categories
Stance	against, pro, for stronger measures, neutral, unclear
Force	advice, criticism, indirect, offer, plea, prototypical, well wishers
Politeness	no redress, on record negative, on record positive, off record
Verb	let, main verb, modal, no
Clause	declarative, imperative, interrogative
Addressees	explicit, implicit
Hashtags	none, yes
Loanwords	none, yes
Subjects	individuated, non-individuated
Vocative	none, yes

21

Covid Directives Dataset (Burnette & Calude, 2022)

22

Selection Menu



- Enables users to work with manageable subsets
- Filter categories
  - show (black)
  - hide (blue)
  - exclude (red)

- Filter variables
  - show (black)
  - hide (blue)
- Merge categories
- Sort: Manually reorder categories and variables
- Undo/redo

Wrapping Up

- Our goal was to enhance the readability, functionality and scalability of the Heatmap Matrix
- Cosmetic changes include:
  - White background, removal of gridlines, fresh colour palettes, different use of main diagonal, triangular matrix design
- Novel features include:
  - Univariate or bivariate colour mapping for six different metrics
  - Visualising the Chi-square test and Cramér's V
  - *Linked Table View* for displaying matching records
  - Interactive filtering & data transformation via *Selection Menu*

Thank you for listening!

- Extending the Heatmap Matrix: Pairwise Analysis of Multivariate Categorical Data
  - David Trye, dgt12@students.waikato.ac.nz

- Thanks to the University of Waikato for funding this research

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

References

- Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.
- Burnette, J., & Calude, A. S. (2022). Wake up New Zealand! Directives, politeness and stance in Twitter #Covid19NZ posts. *Journal of Pragmatics*, *196*, 6-23.
- Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, *8*(3), 373-395.
- Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, *56*, 316–324
- Friendly, M., & Meyer, D. (2016). *Discrete data analysis with R: visualization and modeling techniques for categorical and count data* (Vol. 120). CRC Press.
- Rocha, M. M. N., & da Silva, C. G. (2022). Heatmap matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis. https://doi.org/10.13140/RG.2.2.36619.57126
- Rocha, M. M. N., & da Silva, C. G. (2018). Heatmap matrix: A multidimensional data visualization technique. In *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*.
- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages* (pp. 336-343). IEEE.

Additional Slides

Interactive Enhancements (Rocha & da Silva, 2022)

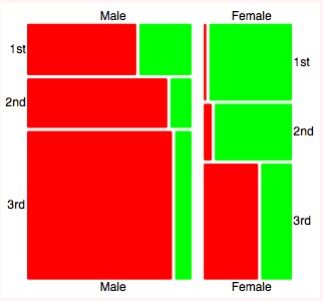
1. **Reorder Matrix:** Sort heatmap to reveal patterns concerning:
   - all pairs of variables (all panels)
   - a single pair of variables (one chosen panel)
   - a subset of variables
   - all variables
2. **Filter Data:** Extract subset using Spearman's correlation coefficient and/or association rules
3. **Discretize Variables:** Make continuous variables in the dataset categorical by creating bins of equal width/frequency
4. **Adjust Scope:** Switch between a local (panel) or global (matrix) colour mapping

- BUT no prototype or GUI documentation available…

Related Work: Contingency Tables

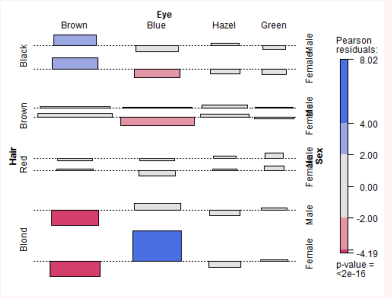
Visualisations of contingency tables can be classified into three types (Alsallakh et al., 2012):

Frequency Representations



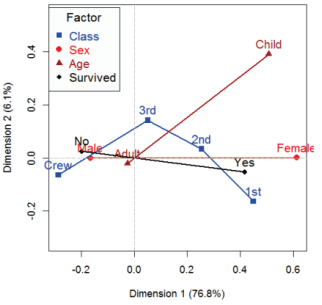
http://www.harold.thimbleby.net/mosaic/

Deviation Representations



http://rstudio-pubs-static.s3.amazonaws.com/

Intermediate Representations

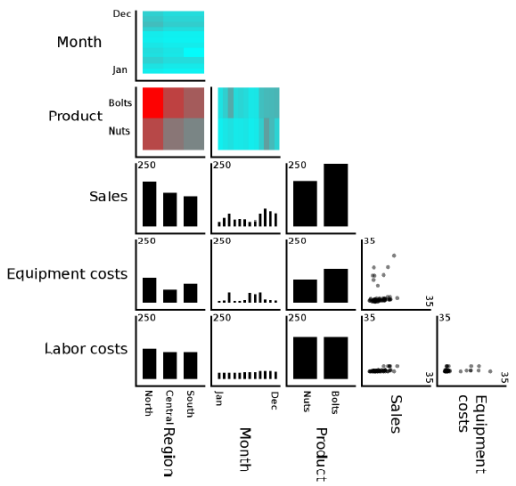


https://friendly.github.io/psy6136/

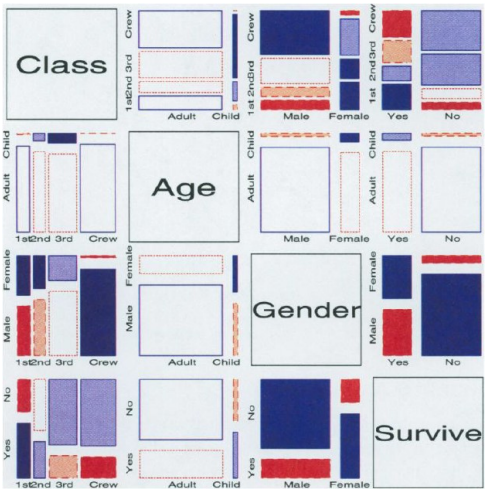
29

Related Work: Pairwise Techniques

GPLOM (Im et al., 2013)



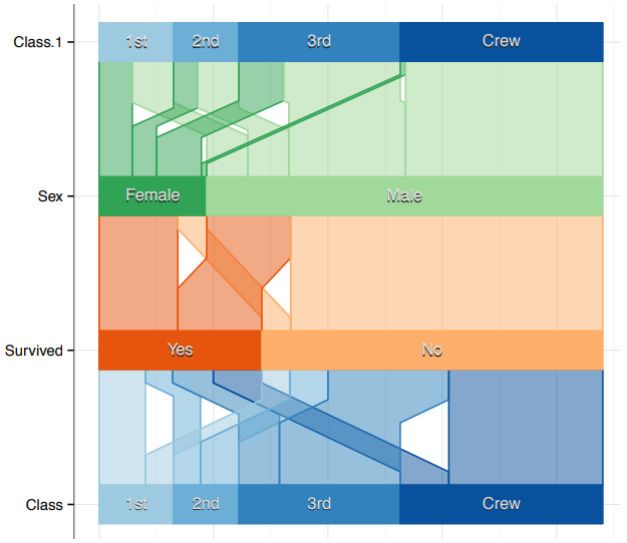
Mosaic Matrix (Friendly, 1999)



30

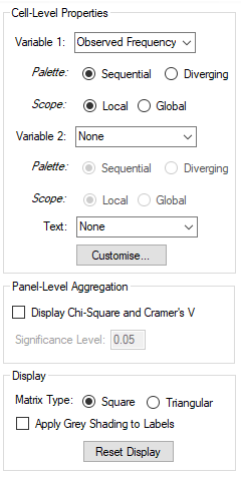
Related Work: Pairwise Techniques

- Parallel Sets and Common Angle Plots can also be used to explore pairwise associations
  (Kosara et al., 2006; Hofmann & Vendettuoli, 2013)



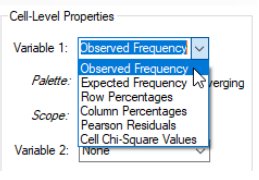
Main Menu

Default Settings



Consists of three sub-menus:

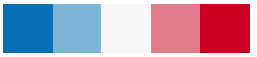
1. Cell-Level Properties:
   Specify colour palette, scope, text labels
2. Panel-Level Aggregation:
   Display Chi-square and Cramér's V
3. Display Settings:
   Change shape, add grey shading



Default Colour Palettes

Univariate:

Sequential  flare (seaborn)

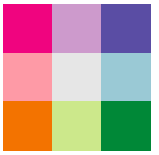
Diverging  blue-white-red

Bivariate: Cynthia Brewer's nine-class bivariate maps

Less precise as fewer shades used for each variable

DivSeq

SeqSeq

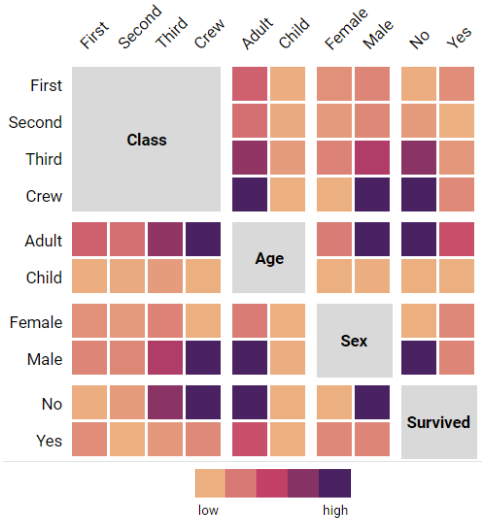


DivDiv

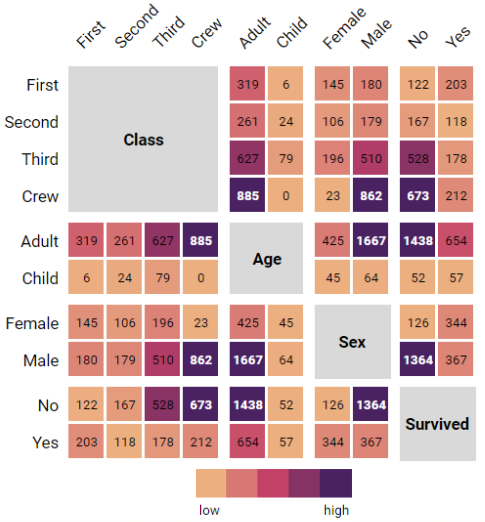
33

Text Labels

None (Default)

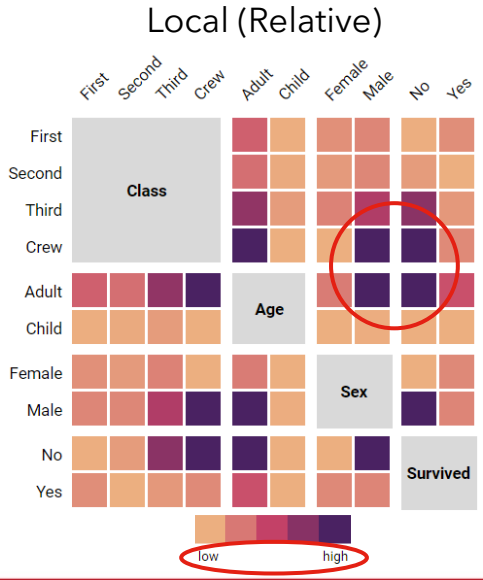


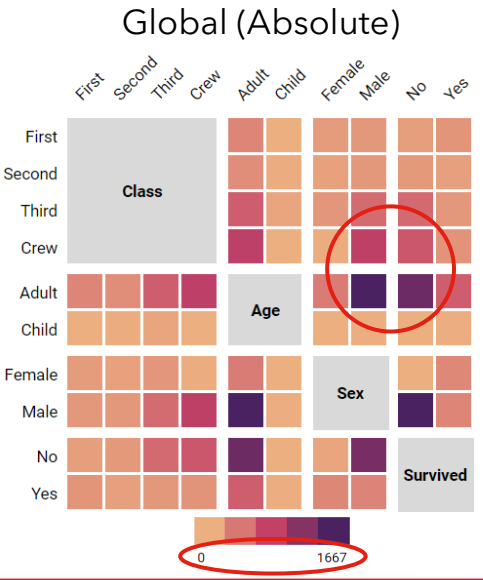
Text (Observed Freq.)



34

Scope Comparison

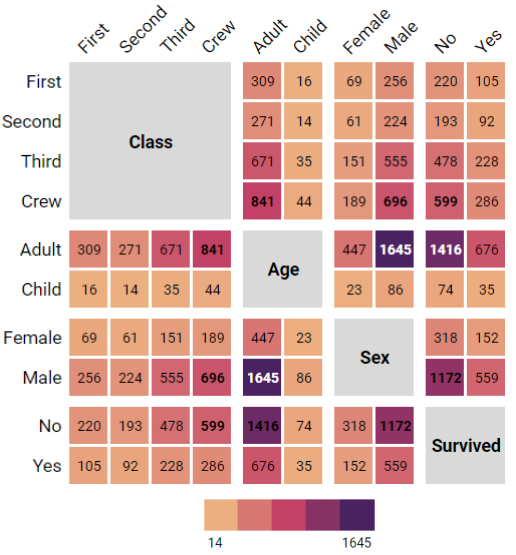




35

Expected Frequencies

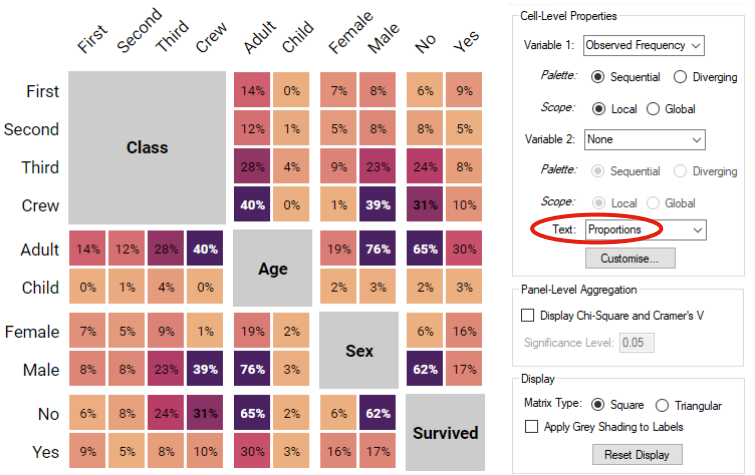
- Shows quantities expected if *no* association between variables
- Calculated using category frequencies
  - E = Row Total x Col Total / *N*



36

Proportions

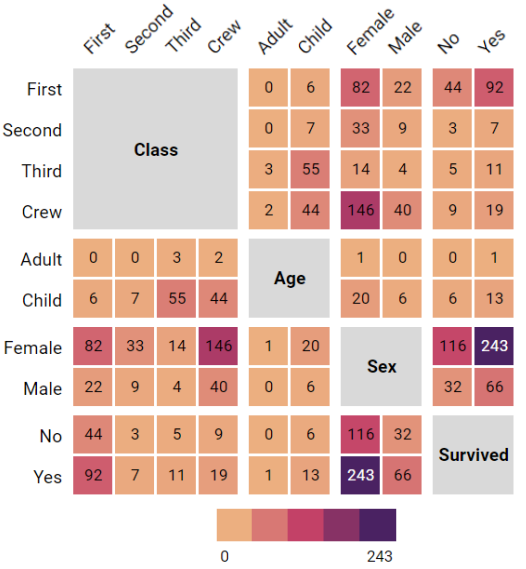
- For text labels, Observed Frequency is split into "Counts" and "Proportions"
- Cells show joint probability, P(X∩Y), where X and Y are categories on each axis



37

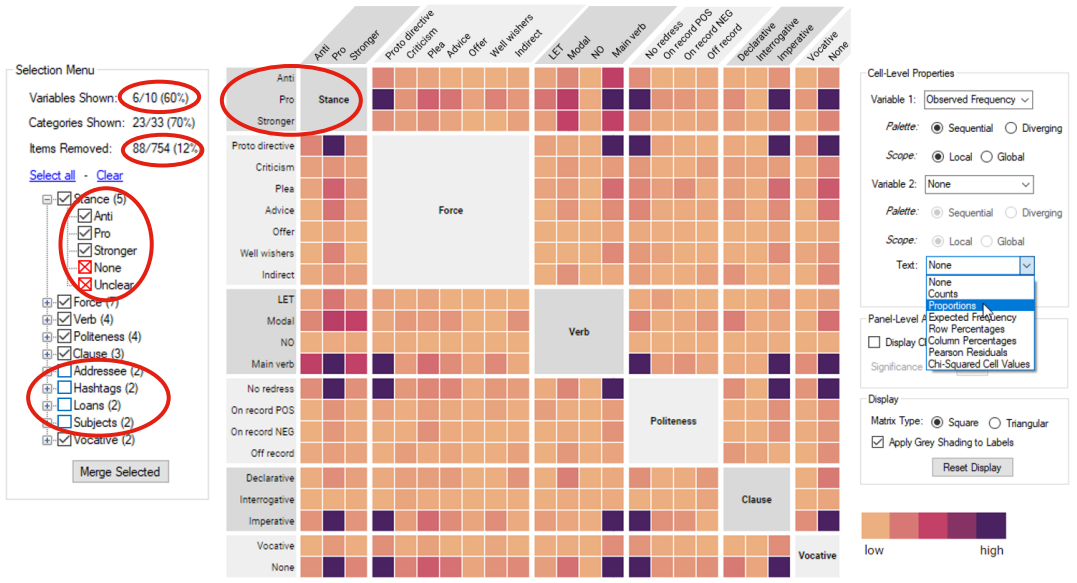
Cell Chi-Square Values

- Shows individual contribution of each cell to overall Chi-square statistic
- Formula: $\chi_{ij} = (O_{ij} - E_{ij})^2 / E_{ij}$
- Large values indicate disparity between observed and expected frequencies



38

Covid Directives Dataset (Burnette & Calude, 2022)



39

Future Work

- Develop a web-based tool where users can upload their own categorical datasets
- Conduct user testing, update design accordingly
- Incorporate features proposed by Rocha & da Silva (2022), especially automated sorting and binning of continuous variables
- Add support for missing values, hierarchical data
- Bigger picture: New methods for visualising (and understanding!) statistical tests for categorical data

40