Discrete Data Analysis

A Friendly Guide to Visualising Categorical Data for Machine Learning Practitioners

Julian Hatwell

18 February, 2019

Introduction

Aim and Objectives

Democratise Machine Learning

Turn You On to Discrete Data

- Share tips and tricks
- Build intuition, focus on visual analytics
- By example
- Relevant to ML project life-cycle
- Simple, reproducible code
- Share simple code snippets
- Plain English
- Avoid theory and formulas where possible

Target Audience

Knowledge professionals who

- have some experience of classification on tabular data sets
- understand different data-types (continuous, nominal, ordinal, count)
- are aware of ML project life-cycle, in particular:
  - Exploratory Data Analysis
  - Classification
  - Evaluating models
- know a little stats (you know what a χ^2 test is)

You can't teach an old dog new tricks[1]

About me

- BSc (Hons) Microbiology, University of Leeds
- Over 15 years database design and development through to senior management roles (mostly in higher education sector), UK and Singapore
- MSc (Distinction) Business Intelligence, Birmingham City University
- Research to PhD (in progress) in Machine Learning, Birmingham City University

[1] Oh, yes you can!

$\mathbb{E}$(ML project) = ?

1. Not fun - Getting and cleaning data, exploratory analysis, feature engineering
2. Fun! - Training Models, XVal, Param Tuning
3. Not fun - Reporting results

This might be what we want, but is it realistic?

Frequently Observed ML Projects

1. Not fun - Getting and cleaning data, exploratory analysis, feature engineering
2. Fun! - Training Models, XVal, Param Tuning
3. Not fun - Reporting results

You can teach old tricks to a new dog!

Focus on those head and tail activities:

- Developing a better EDA strategy for categorical data sets
- Exploring classification results in more detail
- Demonstrating rigorous and robust, yet visually intuitive reporting of results
- Correct handling of ordinal classification results

Theme for today:

- Stats based techniques, applied:
    - in new ways
    - to new problems

Essential Tools

- Plots and Charts
- Tables and Arrays
- χ^2 Test
- Log Odds Ratios
- Discrete Distributions

Out of Scope

No time to develop theories and proofs

Predictive and Explanatory Models

- Logistic Regression
- Cumulative Odds Models
- Loglinear Models
- Generalised Linear Models

No discussion of R software itself

Credits

Michael Friendly is a pioneer in this field and has contributed to the development of modules and libraries for SAS and R.

Code examples based on material contained in the book **Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data** by Michael Friendly.

Shorter, valuable tutorial on these topics in the vcd vignette. Just type **vignette("vcd")** at the R console.

This **video** https://www.youtube.com/watch?v=qfNsoc7Tf60 is a much more in depth lecture, by Michael Friendly, on topics covered in the book.

Example One: Exploring Data with Area Based Plots

Why is Exploratory Analysis so Important?

"Get to know" the data. This always pays off:

- Get rid of noise variables
- Identify most useful variables early
- Guide model selection
- Identify anomalies
- Develop intuition prior to modeling
- Develop a research question, if you don't have one

In 3D with Hair Colour, Eye Colour and Gender

592 stats students, self-categorised, University of Delaware, 1974

No	Name	Levels
1	Hair	Black, Brown, Red, Blond
2	Eye	Brown, Hazel, Green, Blue
3	Sex	Male, Female

Hard to Parse Lots of Numbers

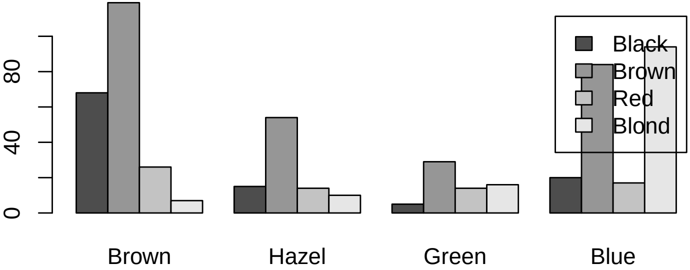
```
##              Hair Black Brown Red Blond
## Sex    Eye
## Male   Brown         32    53  10     3
##        Hazel         10    25   7     5
##        Green          3    15   7     8
##        Blue          11    50  10    30
## Female Brown         36    66  16     4
##        Hazel          5    29   7     5
##        Green          2    14   7     8
##        Blue           9    34   7    64
```

What You See Depends on the Pivot

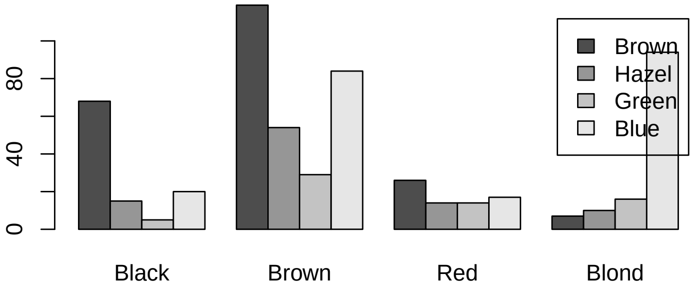
```
##              Sex Male Female
## Hair  Eye
## Black Brown        32     36
##       Hazel        10      5
##       Green         3      2
##       Blue         11      9
## Brown Brown        53     66
##       Hazel        25     29
##       Green        15     14
##       Blue         50     34
## Red   Brown        10     16
##       Hazel         7      7
##       Green         7      7
##       Blue         10      7
## Blond Brown         3      4
##       Hazel         5      5
```

Naive Approach: Barplot Count in 2-D (ignoring Sex)

You're forced to favour one variable over the other. Does hair colour depend on eye colour?

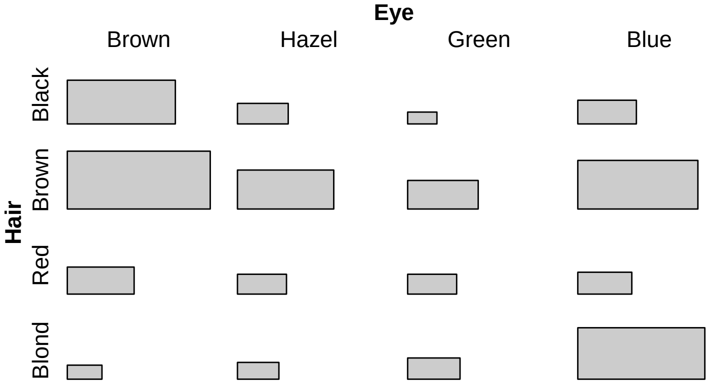


Or does eye colour depend on hair colour? Comparing between groups is tricky.



Tile Plot - Preserve Table Structure

```
# vcd package: one line of code!
# (table haireye prepared earlier)
tile(haireye)
```



Evidence of a Relationship

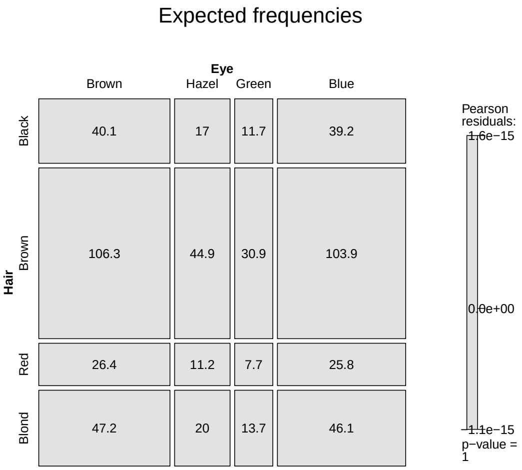
Relationship between hair colour and eye colour is evident.

The tile plot is exploratory. We need to be rigorous.

To demonstrate pattern is real, not sampling error, we perform a χ^2 test of independence:

Check observed counts against expected counts.

Mosaic Plot - Expected Counts



H_0: No Association Between Hair and Eye

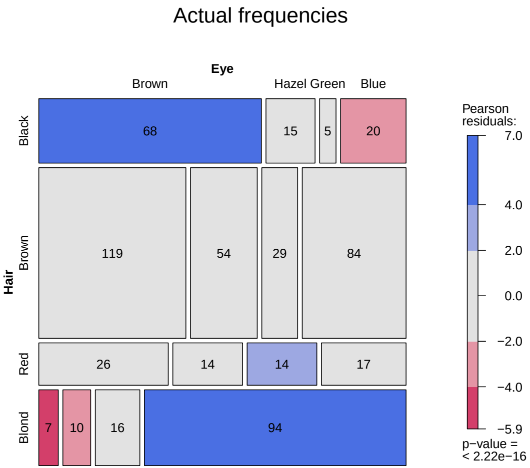
```
##
##  Pearson's Chi-squared test
##
## data:  haireye
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Rejected, obviously. There clearly is a relationship.

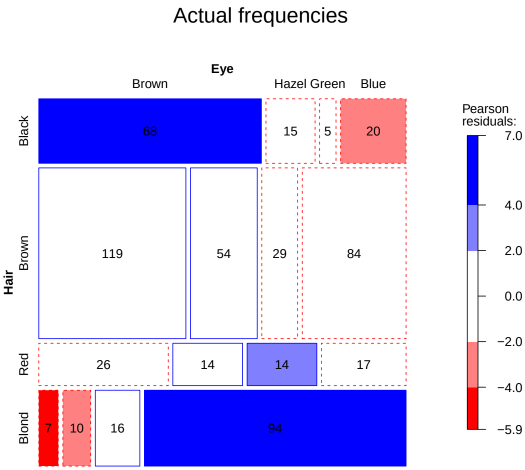
χ^2 test gives no details. How to describe it?

χ^2 residuals contain a lot of information!

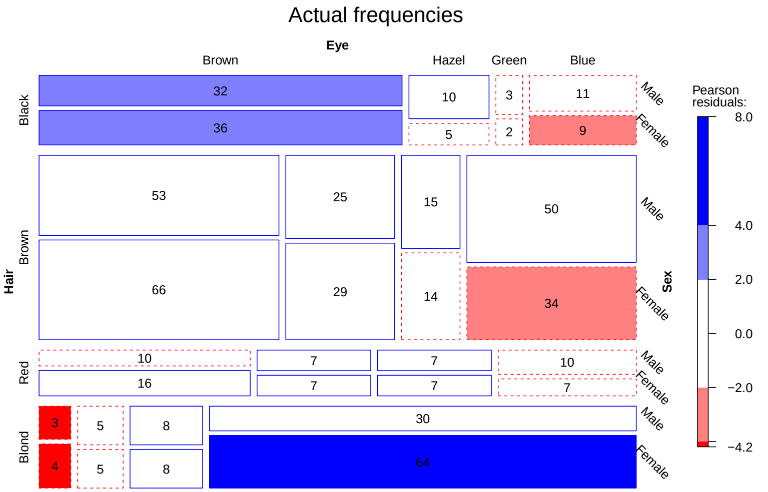
Mosaic Plot - Observed Counts



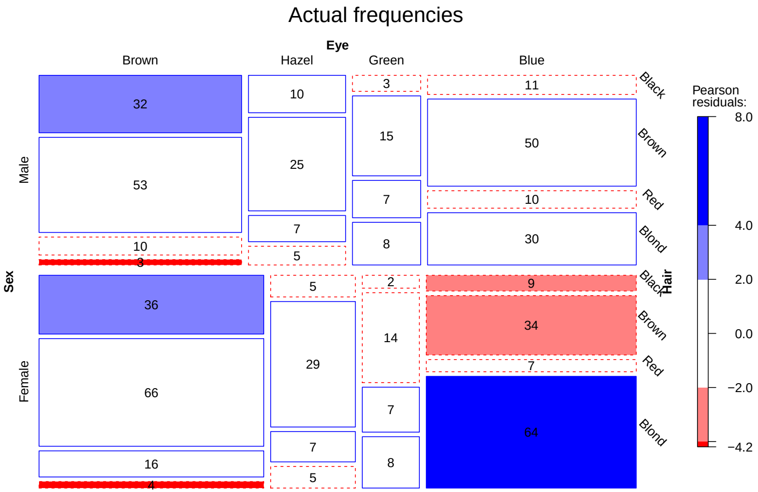
Mosaic Plot - Friendly Colour Scheme



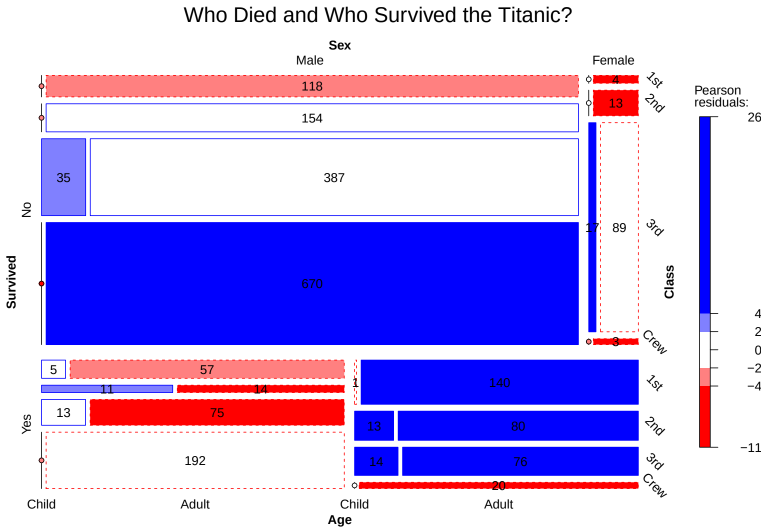
Previous + Sex Feature: Now in 3D



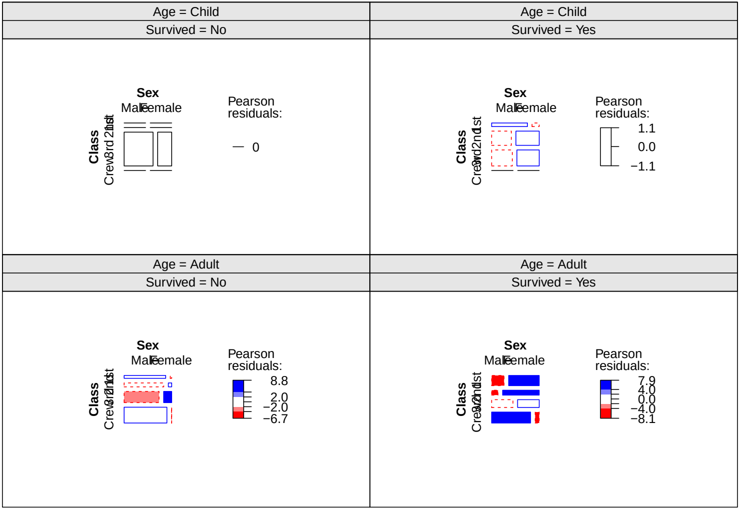
What You See Depends on the Pivot



Mosaic Plot in 4D - Who Survived the Titanic?



Mosaic in n-D - Faceting



Mosaic - Summary

The final plot didn't render nicely on these slides. Screen real-estate at run time was used more efficiently.

Important points:

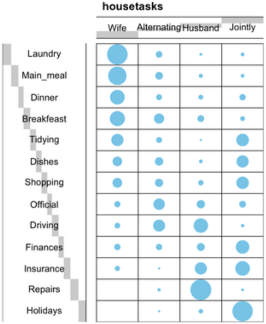
- Mosaic plots; area $\propto$ cell count
- Fill colour by size of deviance residual
- Option for outline colour by deviance residual sign
- Scales well to 4-D
- At least a further 2-D can elevate up to facet
- Allows visual exploration of n-way interactions

Example Two: Clustering and Dimension Reduction with Correspondence Analysis

What is Correspondence Analysis?

Think of CA as somewhere between correlation analysis and PCA for continuous data.

Features and categories that change together, move together. Cells with the largest values have the strongest influence.



Audience Viewing Data

Audience viewing data from Neilsen Media Research for the week starting November 6, 1995

It is a 3-D array cross-tabulating the viewing figures for three networks, between 8-11pm, Monday to Friday. The features and their levels are as follows:

No	Name	Levels
1	Day	Monday, Tuesday, Wednesday, Thursday, Friday
2	Time	8, 9, 10
3	Network	ABC, CBS, NBC

CA - A Cinch

```
# multiple CA - one line of code!
TV3.mca <- mjca(TV3)

# Flatten to 2-D by stacking time and day
TV3s <- as.matrix(structable(Network~Time+Day
                              , TV3))

# simple CA - one line of code!
TV3s.ca <- ca(TV3s)
```

Other Considerations

Constructing a plot needs a little bit more work (not shown).

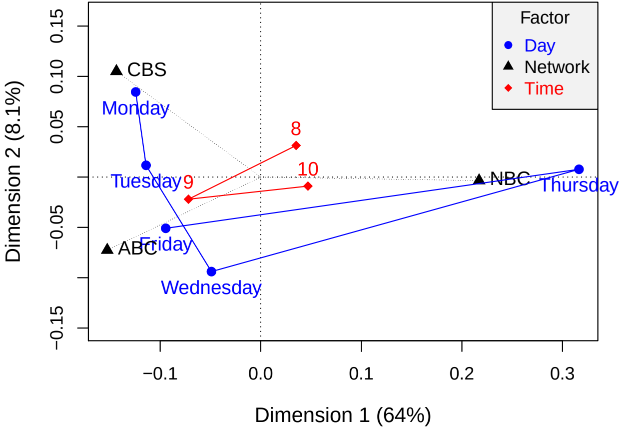
Really, just a little and all base R graphics.

When you've done it once, it's easy to customise for your needs.

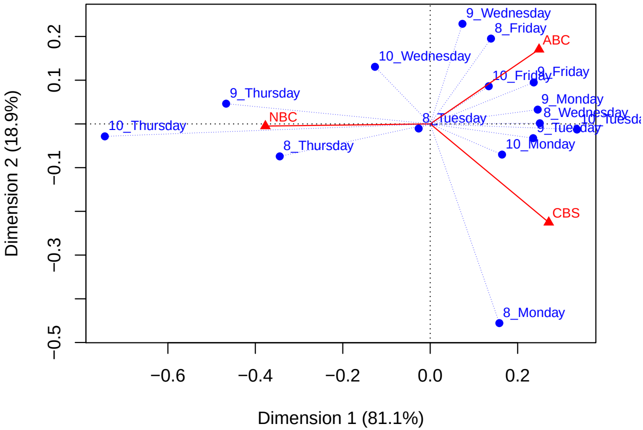
Multiple and Joint CA examines relationships among all features at once and can be used for dimension reduction.

Simple CA only supports 2D data to start with. However, smart use of pivots can actually reveal more information because there are more free points. It will take a bit of trial and error.

Multiple Correspondence Analysis Plot



Simple Correspondence Analysis Plot



Correspondence Analysis - Summary

- CA is a very powerful technique based on matrix decomposition
- Offers additional perspective for exploring data
- Complex, non-parametric relationships are easily visualised can be explored
- Useful for reducing dimenions
- Converting categorical dimensions to continuous, while preserving information
- Sort data by CA dimension rather than natural ordering: [*Monday, Tuesday, Wednesday, Thursday, Friday*] $\not\Rightarrow$ [*1,2,3,4,5*]