Discrete Data Analysis

A Friendly Guide to Visualising Categorical Data for Machine Learning Practitioners

Julian Hatwell

19 February, 2019

Introduction

About Me

- Over 15 years in the for-profit education sector, UK and Singapore - various roles from DBA to Senior Management
- MSc (Distinction) Business Intelligence, Birmingham City University
- Research to PhD (in progress) in Machine Learning, Birmingham City University

About You

Knowledge professionals who

- have some experience of classification on tabular data sets
- understand different data-types (continuous, nominal, ordinal, count, etc. . .)
- 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)

Aim and Objectives

Tips from the Trenches

- Share Practical Experience
- Relevant to ML project life-cycle
- Build intuition
- Focus on visual analytics
- Plain English
- By example
- Avoid theory and formulas
- Simple, reproducible code

This prez is based on a hands-on tutorial that I deliver at BCU. Ask me for the lecture notes if you're interested.

Head and tail activities:

- Developing a better EDA strategy for categorical data sets
- Exploring classification results in more detail
- Correct handling of ordinal classification results

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

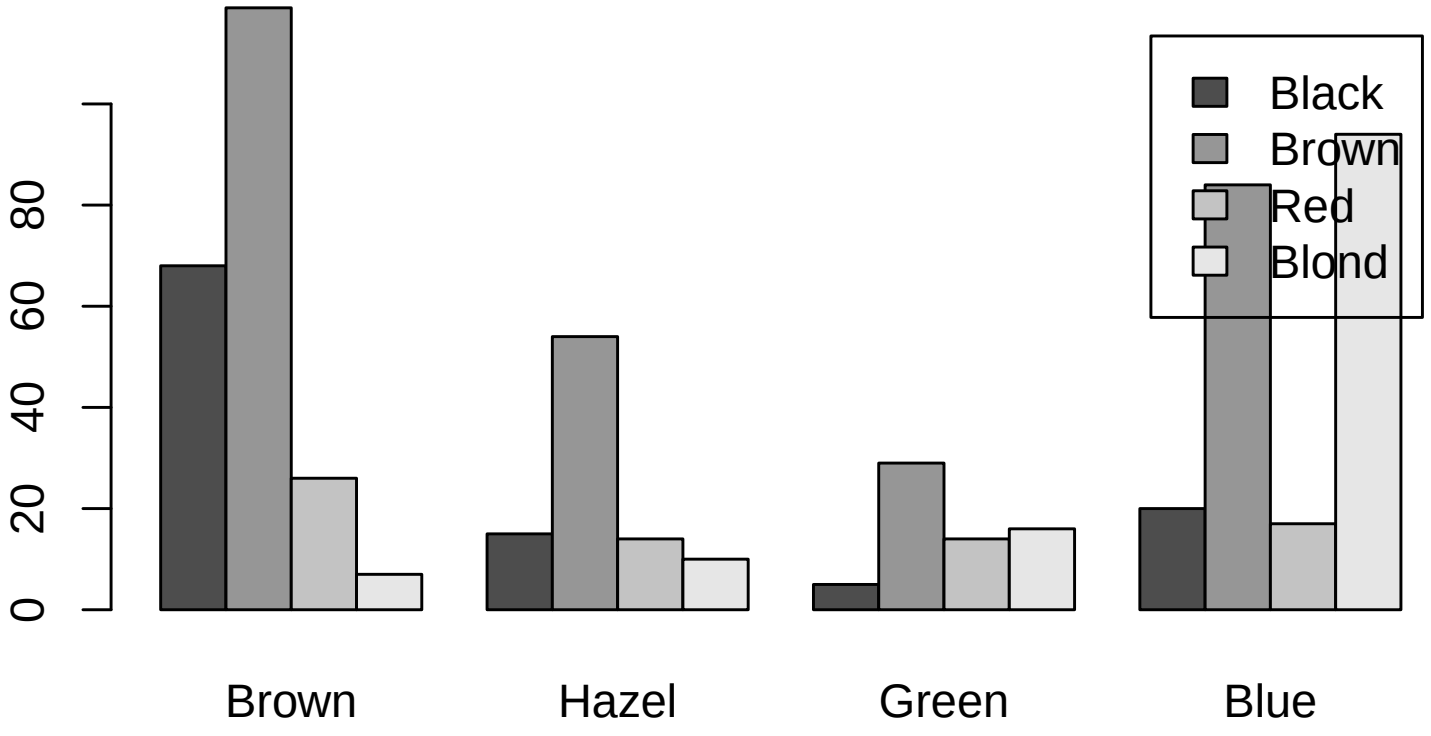
```
##                Hair Black Brown Red Blond
## Sex    Eye
## Male   Brown          32    53  10     3
##        Hazel          10    25   7     5
##        Green           3    15   7     8
##        Blue           11    50  10    30
## Female Brown          36    66  16     4
##        Hazel  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

What You See Depends on the Pivot

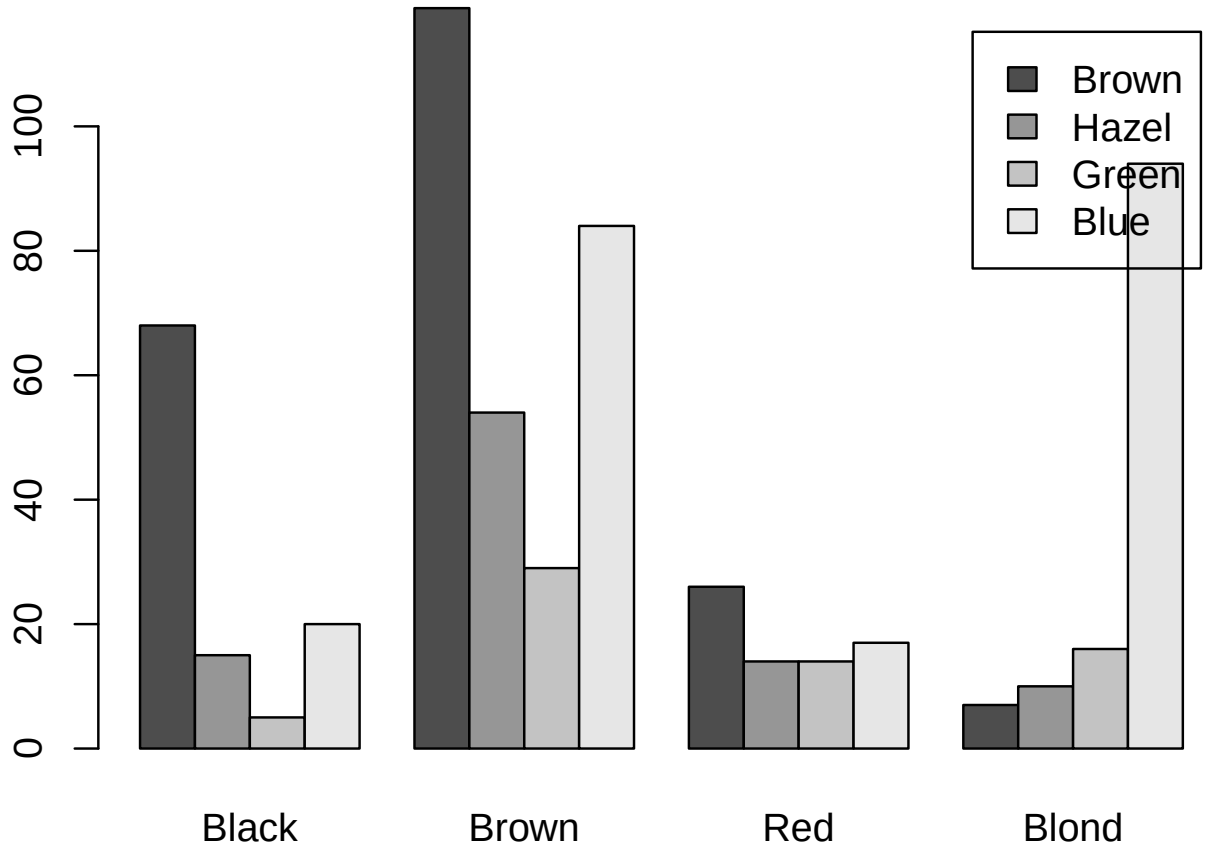
```
##                Sex Male Female
## Hair  Eye
## Black Brown       32     36
##       Hazel       10      5
##       Green        3      2
##       Blue        11      9
## Brown Brown       53     66
##       Hazel       25     29
##       Green       15     14
##       Blue        50     34
##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

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

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

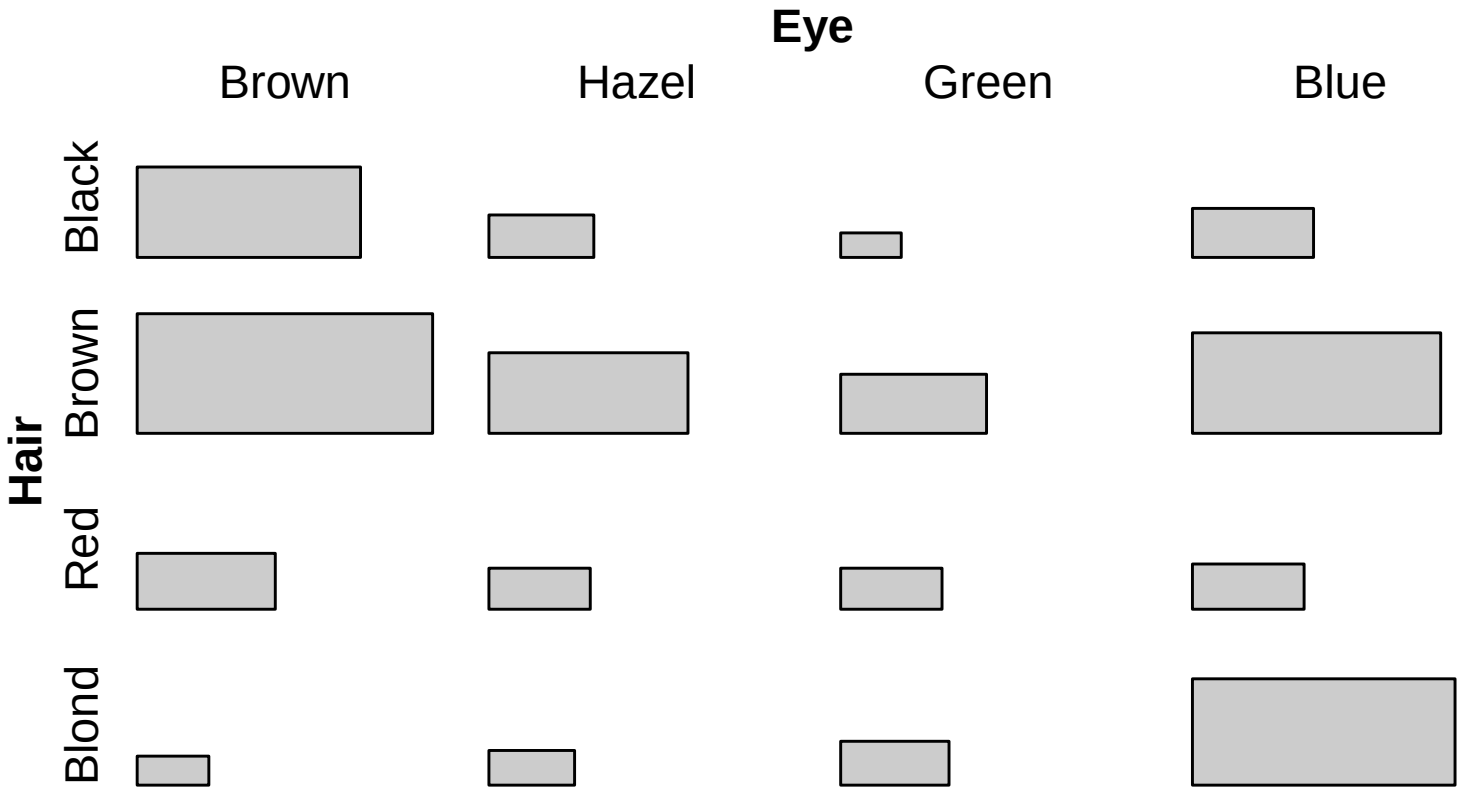


Or does eye colour depend on hair colour? Comparing between groups is tricky.



Tile Plot - Preserve Table Structure

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



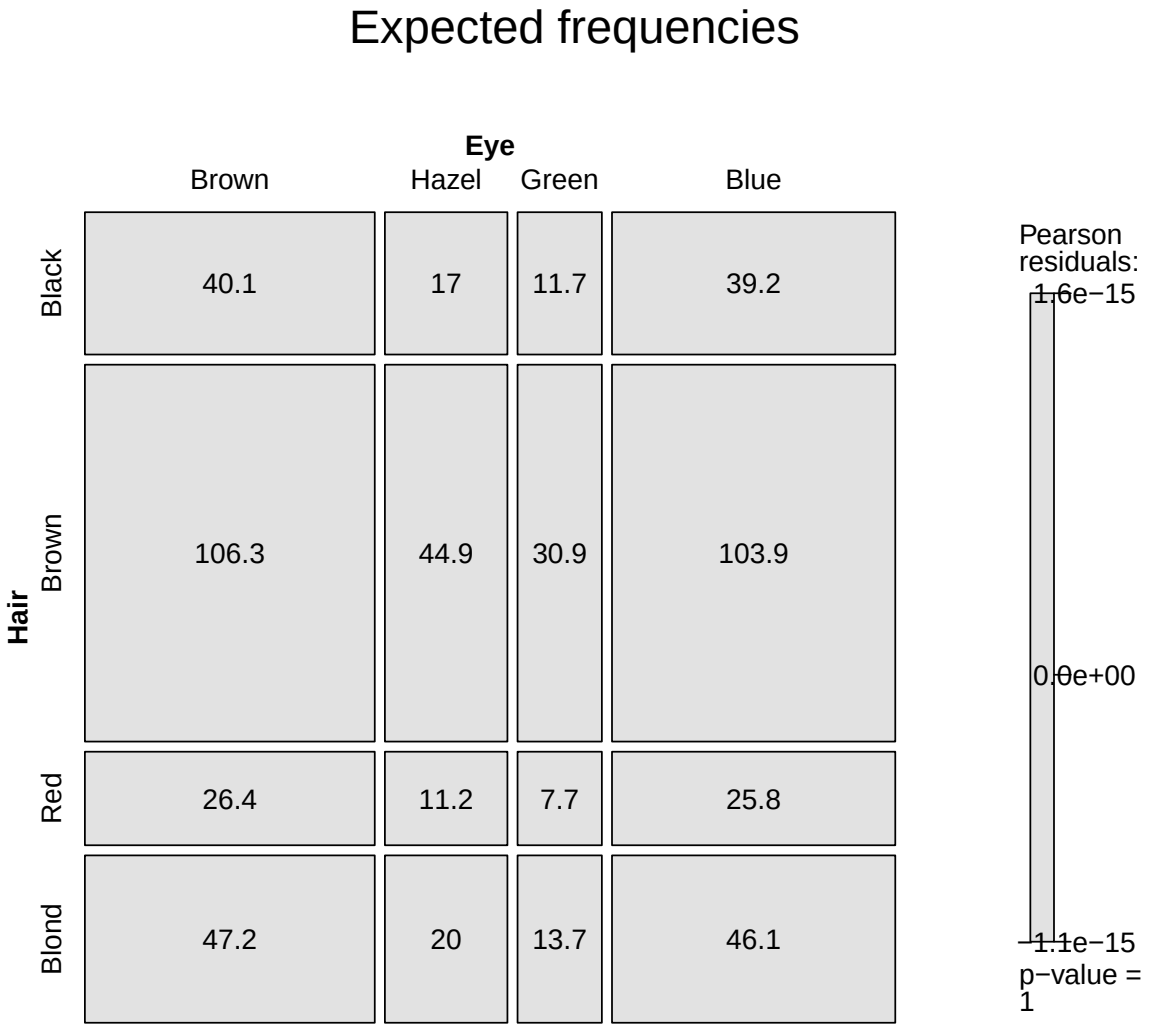
H_0: No Association Between Hair and Eye

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

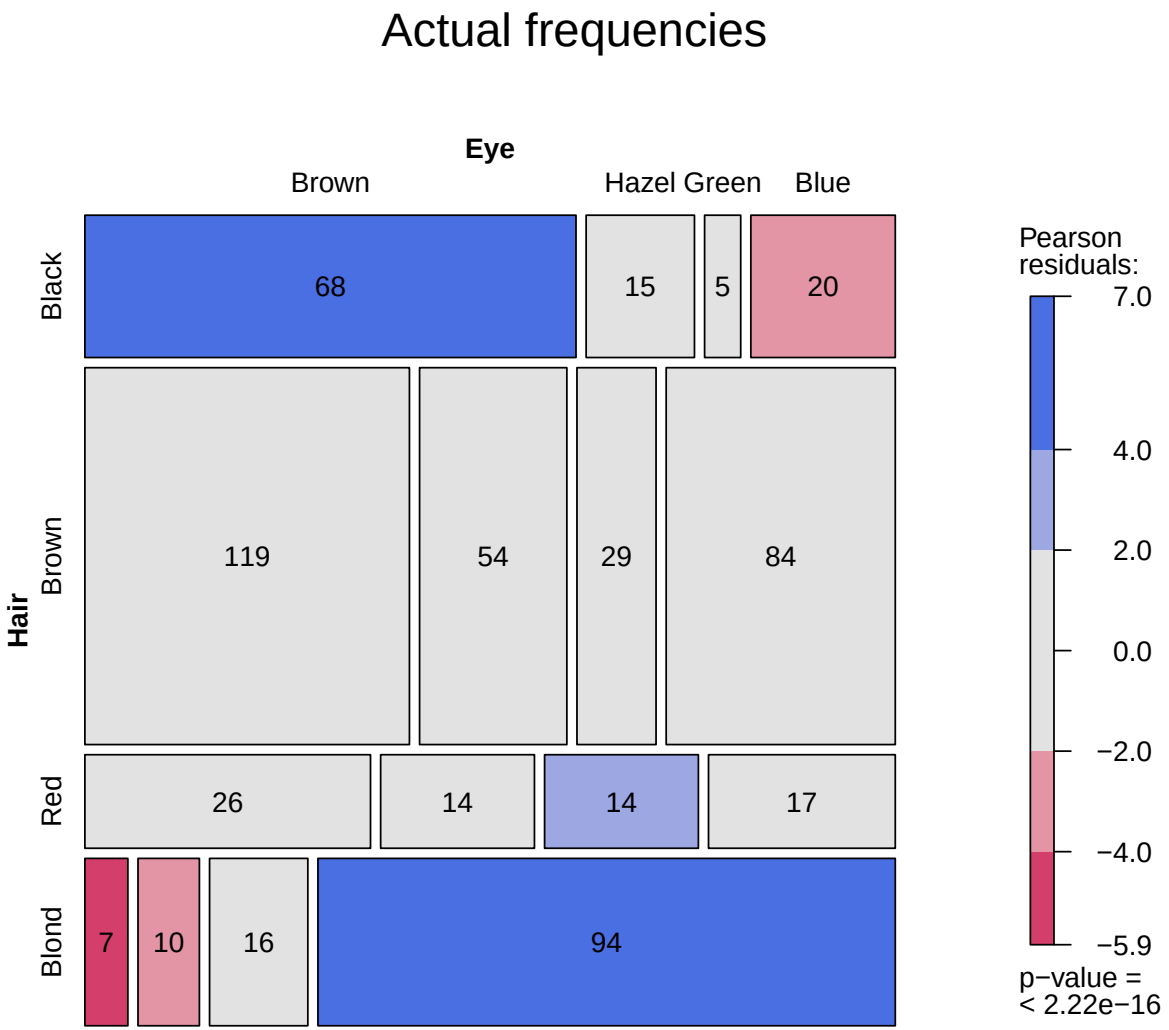
Rejected, obviously. There clearly is a relationship.

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

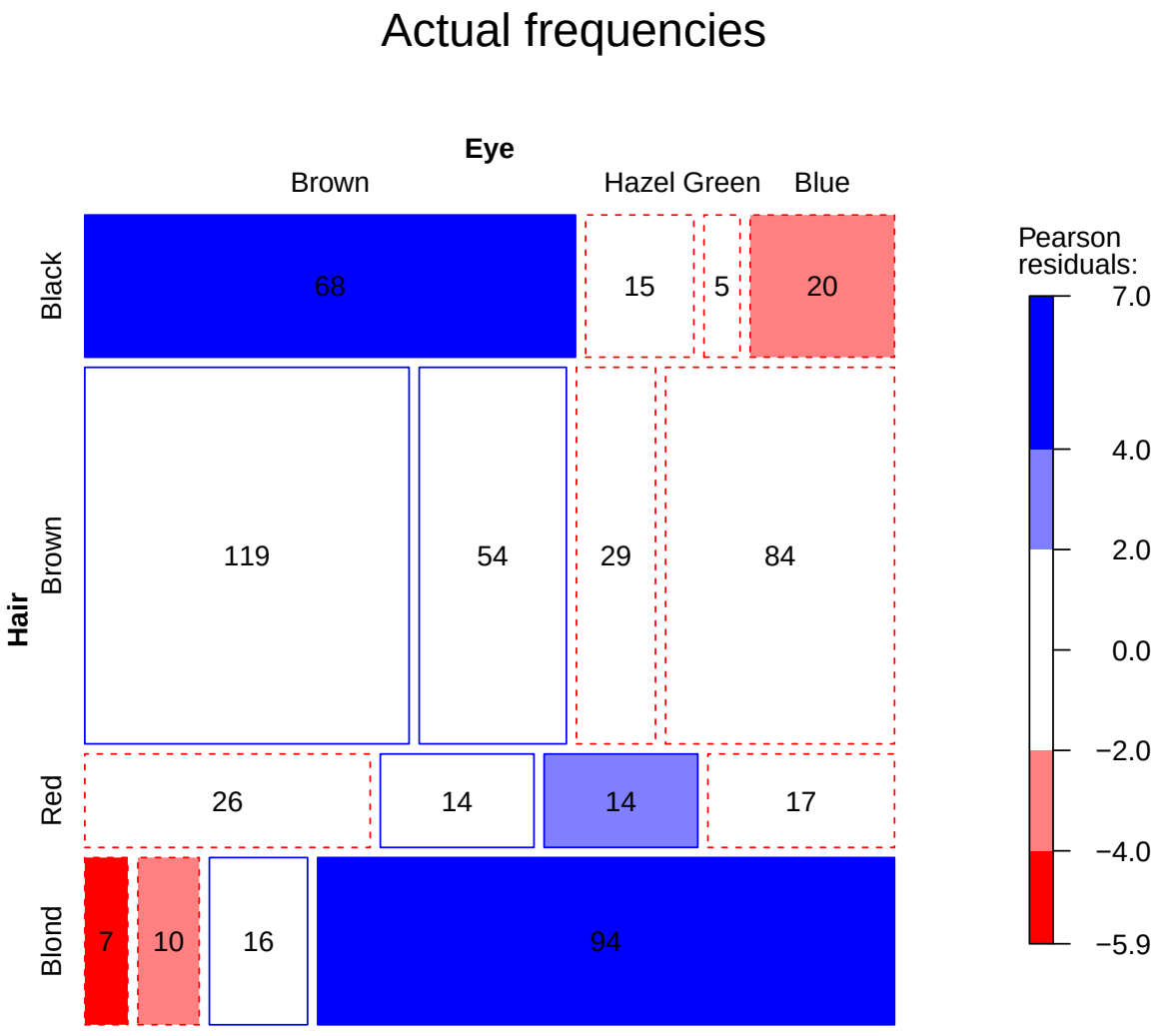
Mosaic Plot - Expected Counts



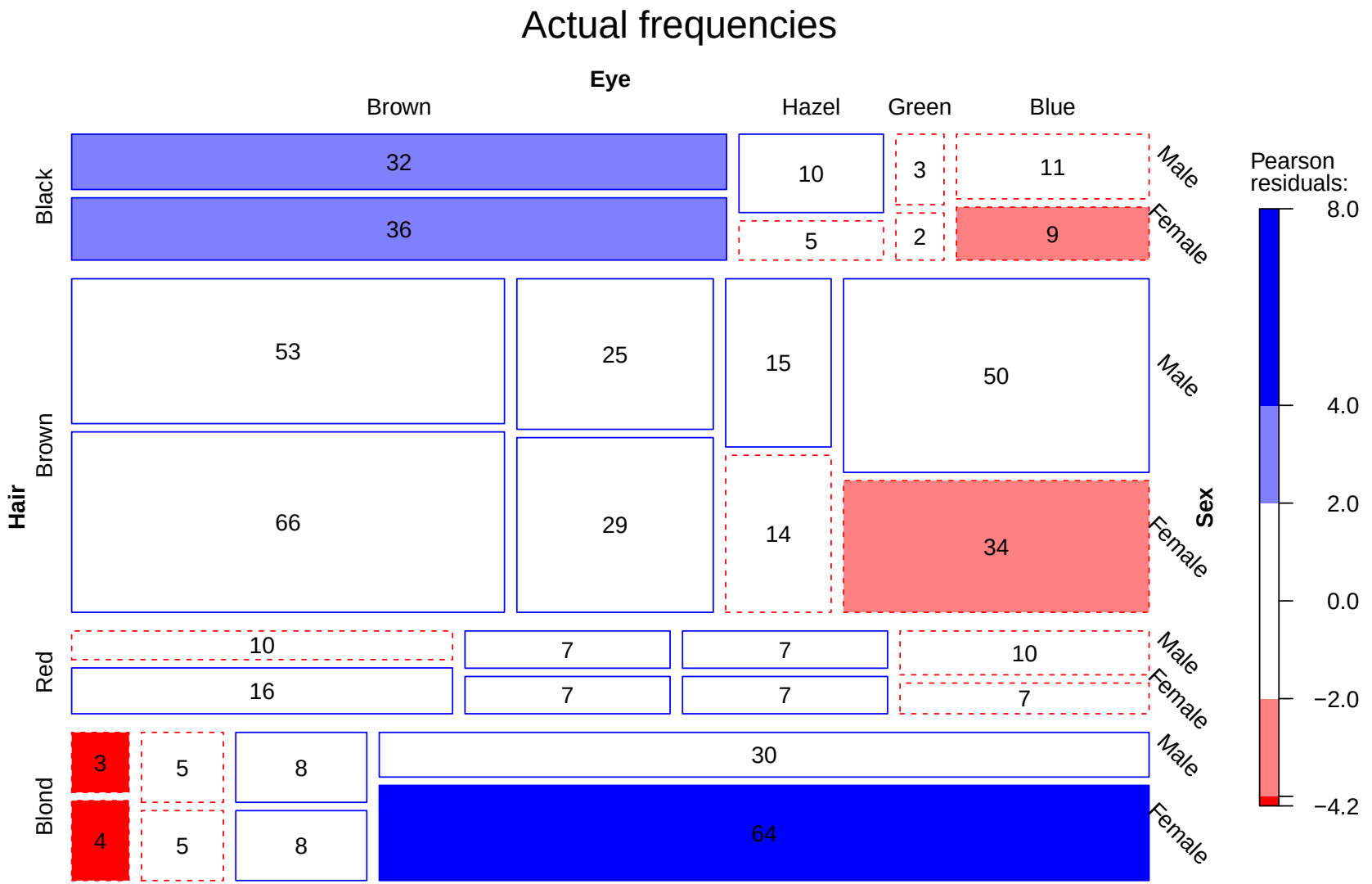
Mosaic Plot - Observed Counts



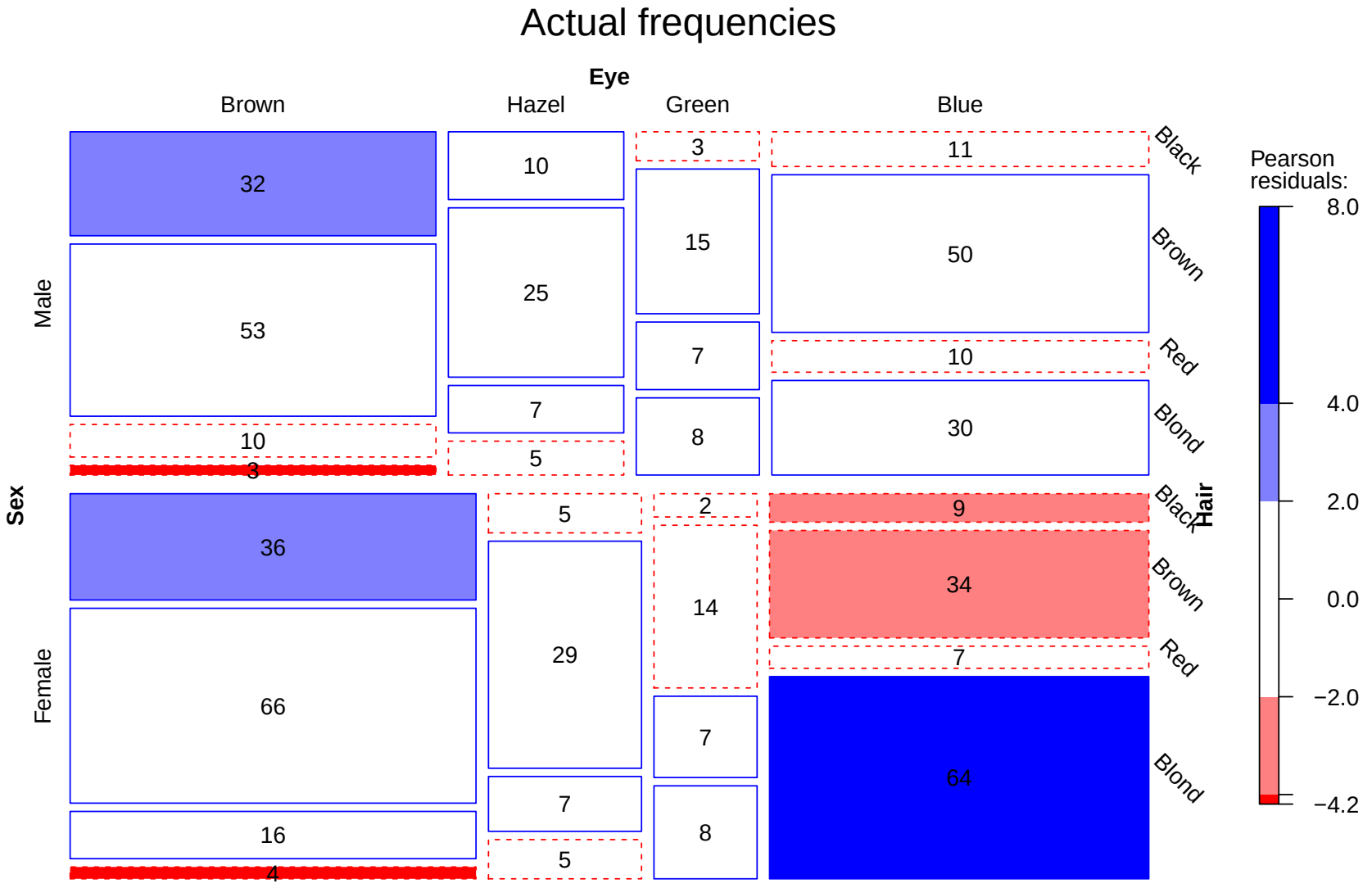
Mosaic Plot - Friendly Colour Scheme



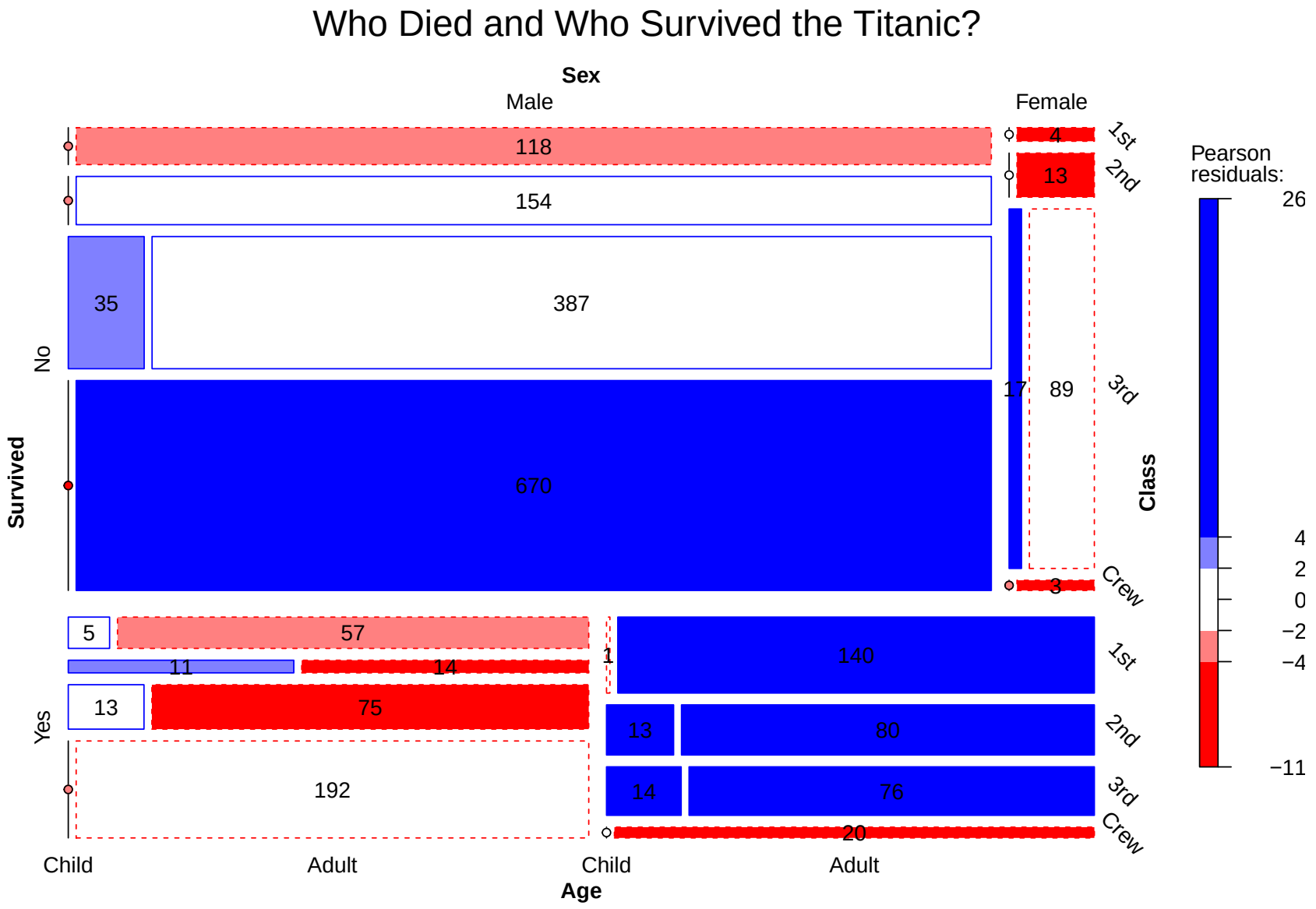
Previous + Sex Feature: Now in 3D



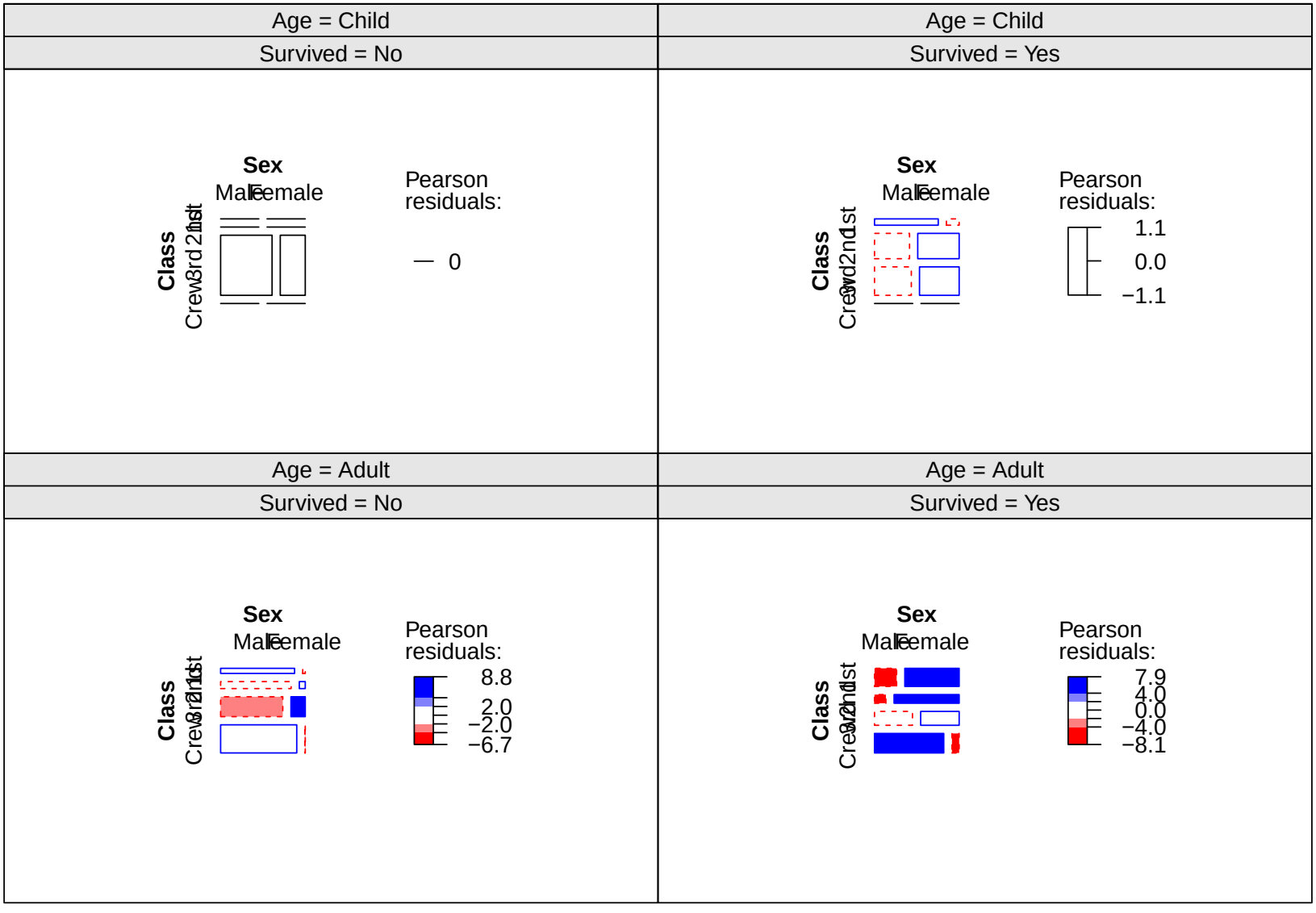
What You See Depends on the Pivot



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:

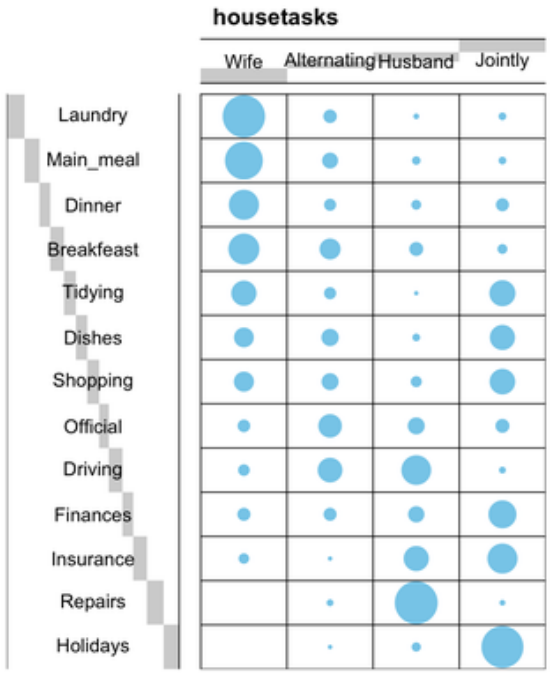
- Mosaic plots; area $\propto$ cell count
- Fill colour by size of deviance residual
- Option for outline colour by deviance residual sign
- Scales well to 4-D
- At least a further 2-D can elevate up to facet
- Allows visual exploration of n-way interactions

Example Two: Clustering and Dimension Reduction with Correspondence Analysis

What is Correspondence Analysis?

Think of CA as somewhere between correlation analysis and PCA but for categorical 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).

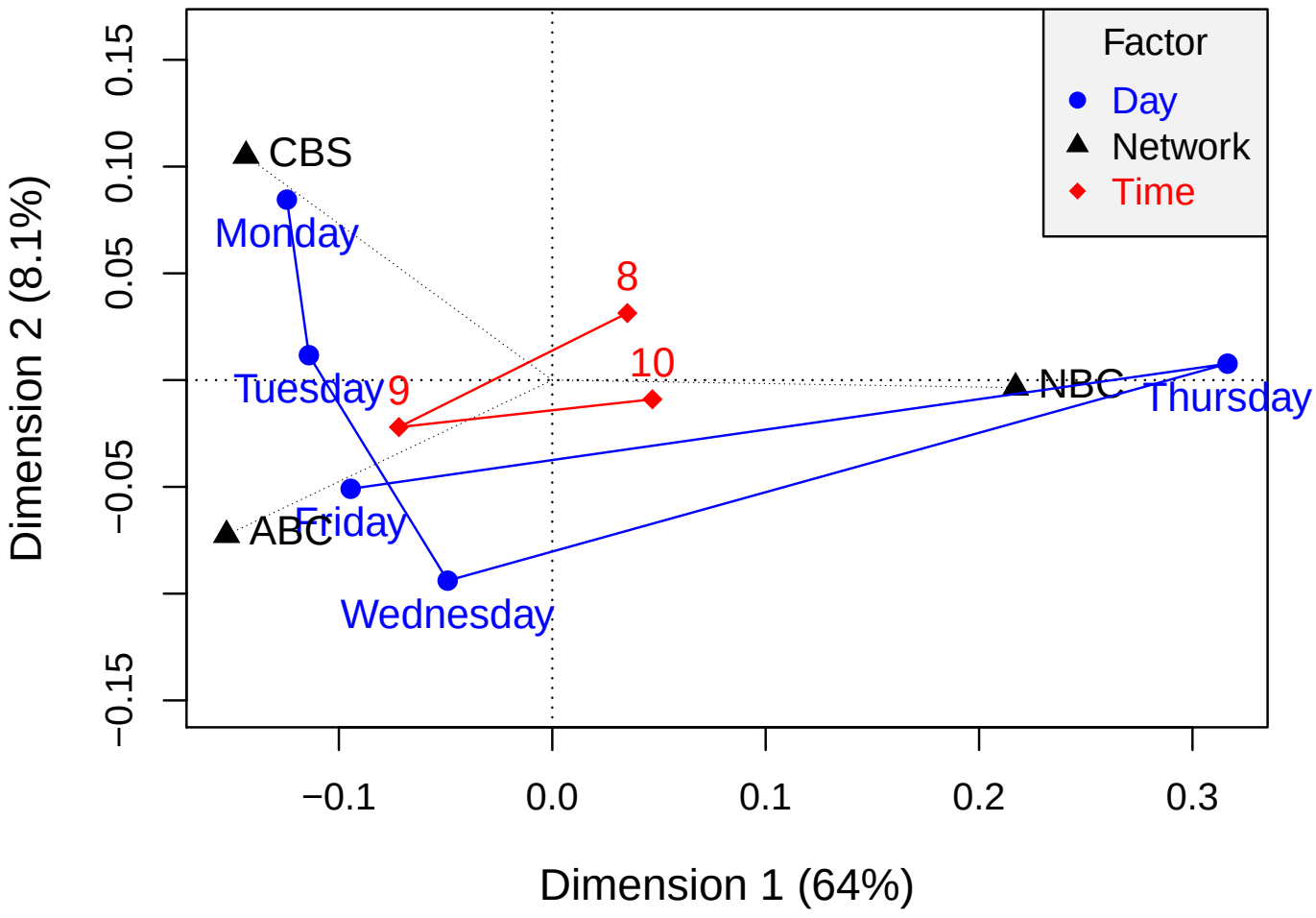
Really, just a little and all base R graphics.

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

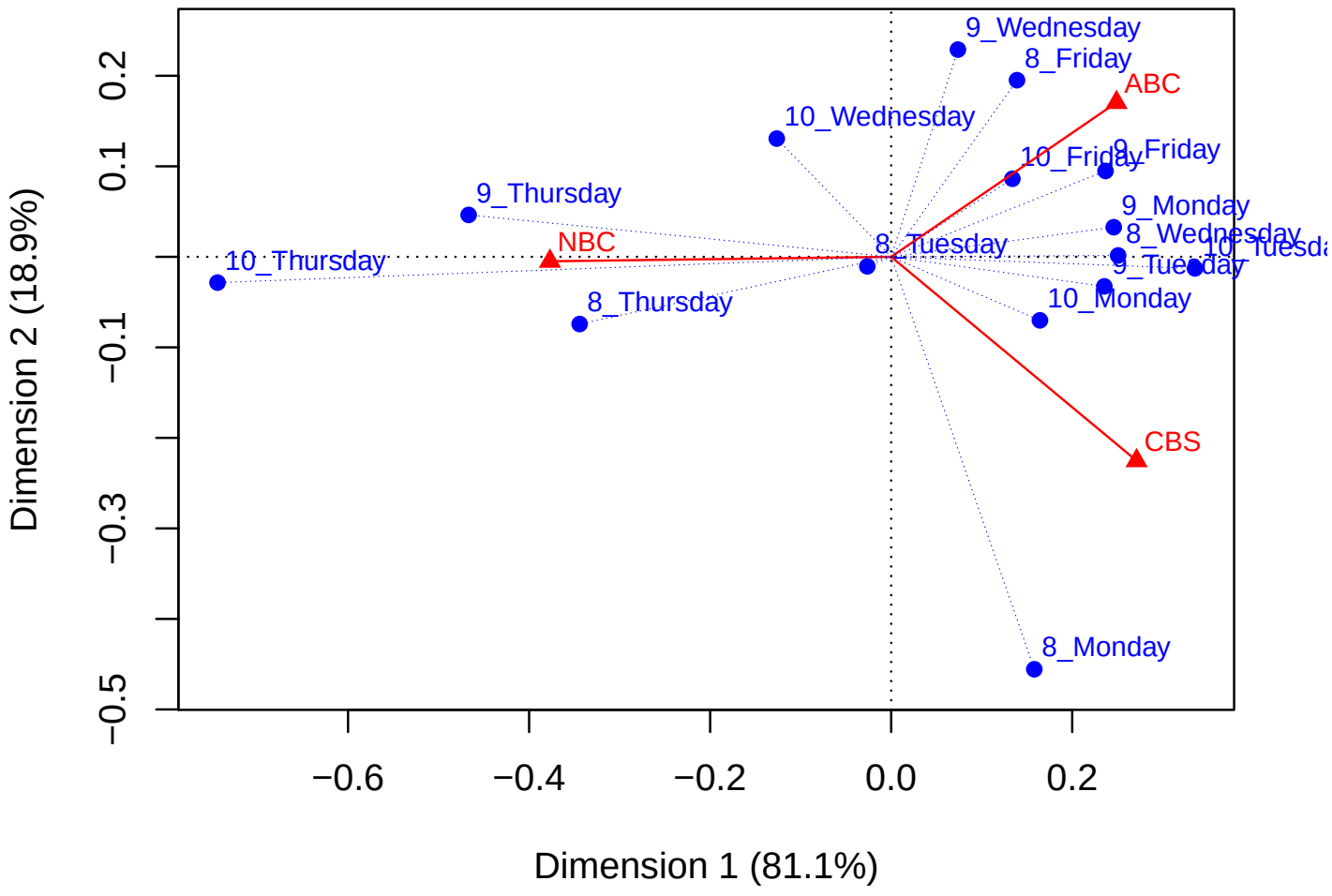


CA Converts Categorical to Continuous

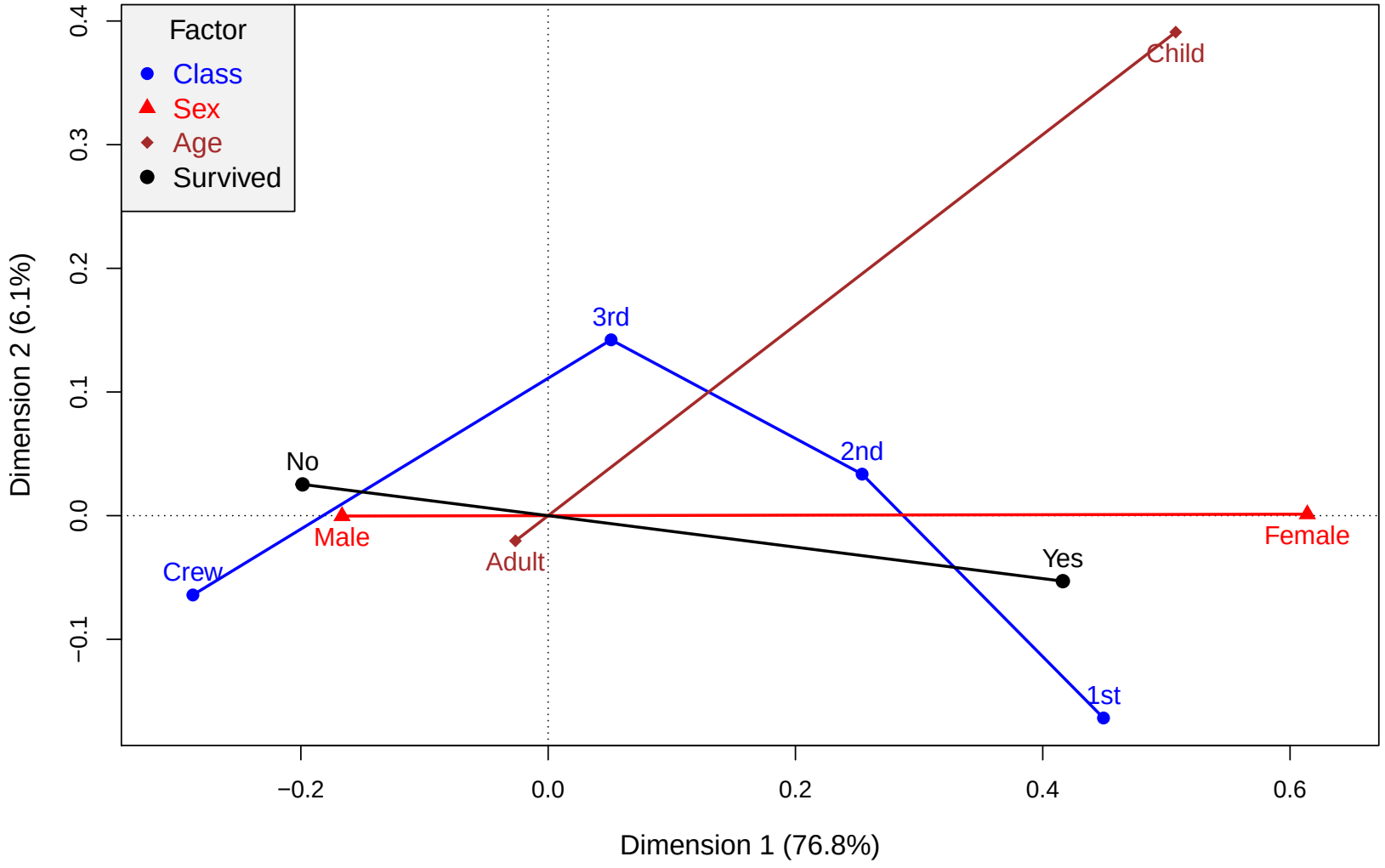
Order is NOT arbitrary!

```
##                          Dim1         Dim2
## Day:Thursday    0.31642576  0.007689690
## Day:Wednesday -0.04902112 -0.093931518
## Day:Friday    -0.09446945 -0.050910135
## Day:Tuesday   -0.11404788  0.011668791
## Day:Monday    -0.12435027  0.084518729
## Network:NBC    0.21708887 -0.003522689
## Network:CBS   -0.14349390  0.105025169
## Network:ABC   -0.15261339 -0.072574817
## Time:10        0.04694020 -0.008975690
## Time:8         0.03518235  0.031389305
## Time:9        -0.07187607 -0.022020743
```

Simple Correspondence Analysis Plot



Multiple Correspondence Analysis Plot in 4-D



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 dimensions
- Converting categorical dimensions to continuous, while preserving information
- Sort data by CA dimension rather than natural ordering: [*Monday, Tuesday, Wednesday, Thursday, Friday*] $\nRightarrow$ [*1,2,3,4,5*]

Example Three: Slicing a Confusion Matrix by Important Features

Overview - The Nursery Data Set

Four possible classes

Three evenly balanced

Fourth is a tiny minority class

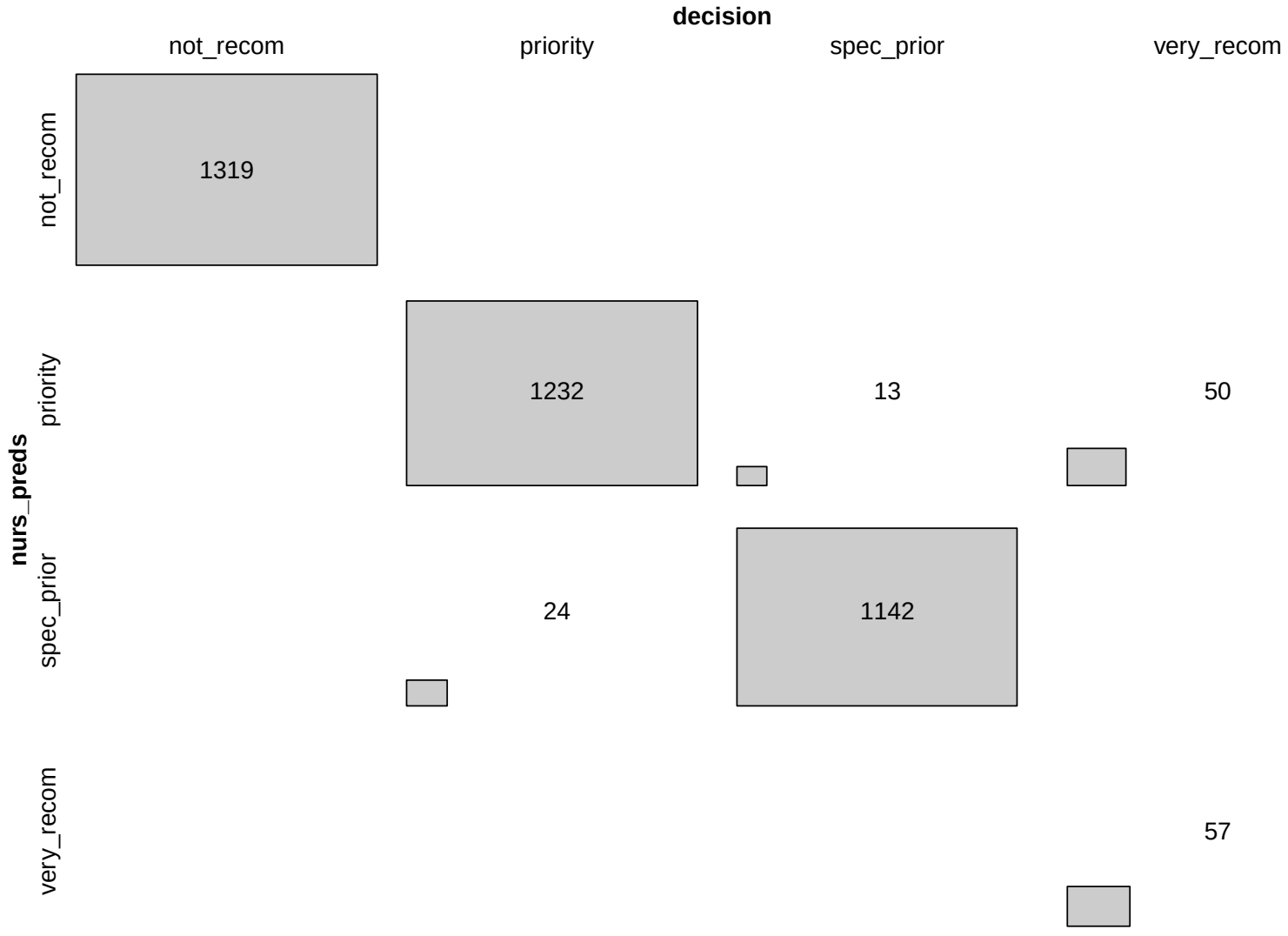
All the predictors are discrete

A random forest is trained on 70% of the data and evaluated on the remaining 30%

```
##          parents                has_nurs                form
##  great_pret :4320   critical   :2592   complete  :3238
##  pretentious:4320   improper   :2592   completed :3240
##  usual      :4318   less_proper:2592   foster    :3240
##                     proper     :2590   incomplete:3240
##                     very_crit  :2592
##
##          housing                finance                social
##  convenient:4318   convenient:6478   nonprob      :4319
##  critical  :4320   inconv    :6480   problematic  :4320
##  less_conv :4320                     slightly_prob:4319
##
##
##
##           health                decision
##  not_recom  :4320   not_recom :4320
##  priority   :4320   priority  :4266
##  recommended:4318   spec_prior:4044
##                     very_recom: 328
##
```

```
##                decision
## nurs_preds    not_recom priority spec_prior very_recom
##   not_recom        1319        0          0          0
##   priority            0     1232         13         50
##   spec_prior          0       24       1142          0
##   very_recom          0        0          0         57
```

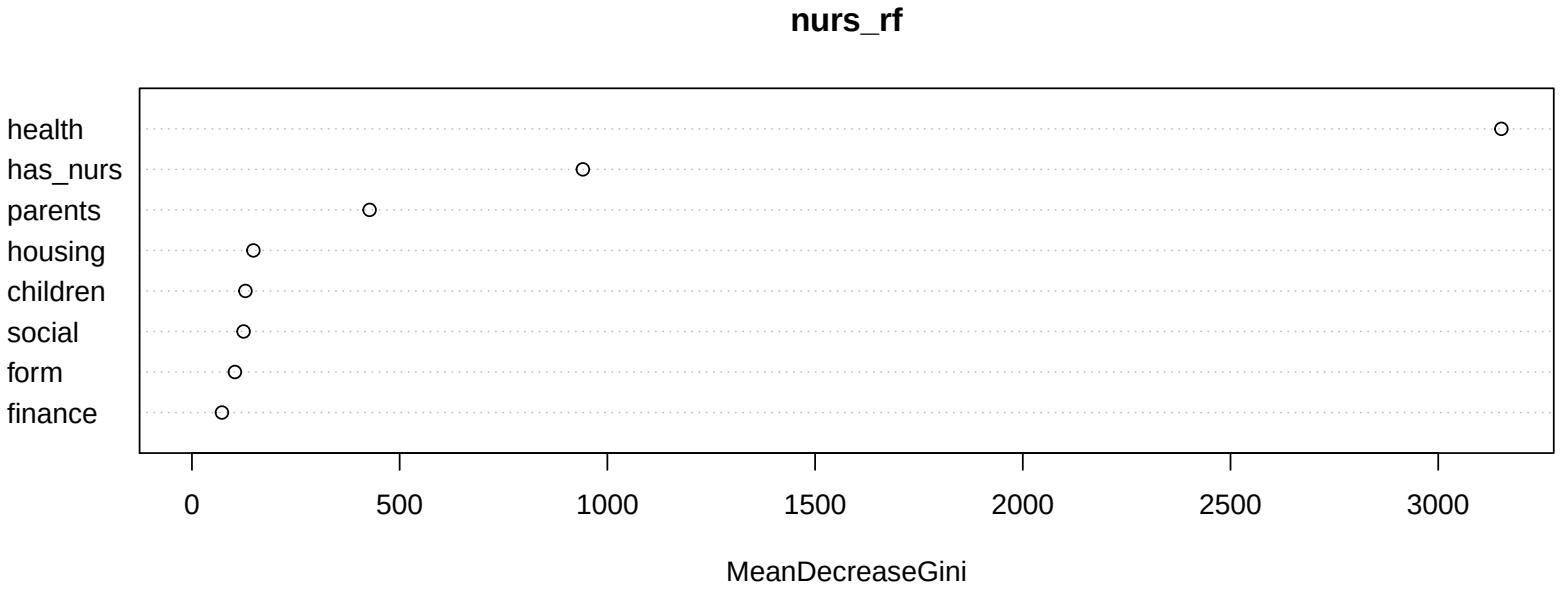
Tile Plot of Confusion Matrix



Variable Importance Plots

Important in what way?

What does anyone do with this information?



Note the elbow. Could we do without half the features?

What Makes A Variable Important?

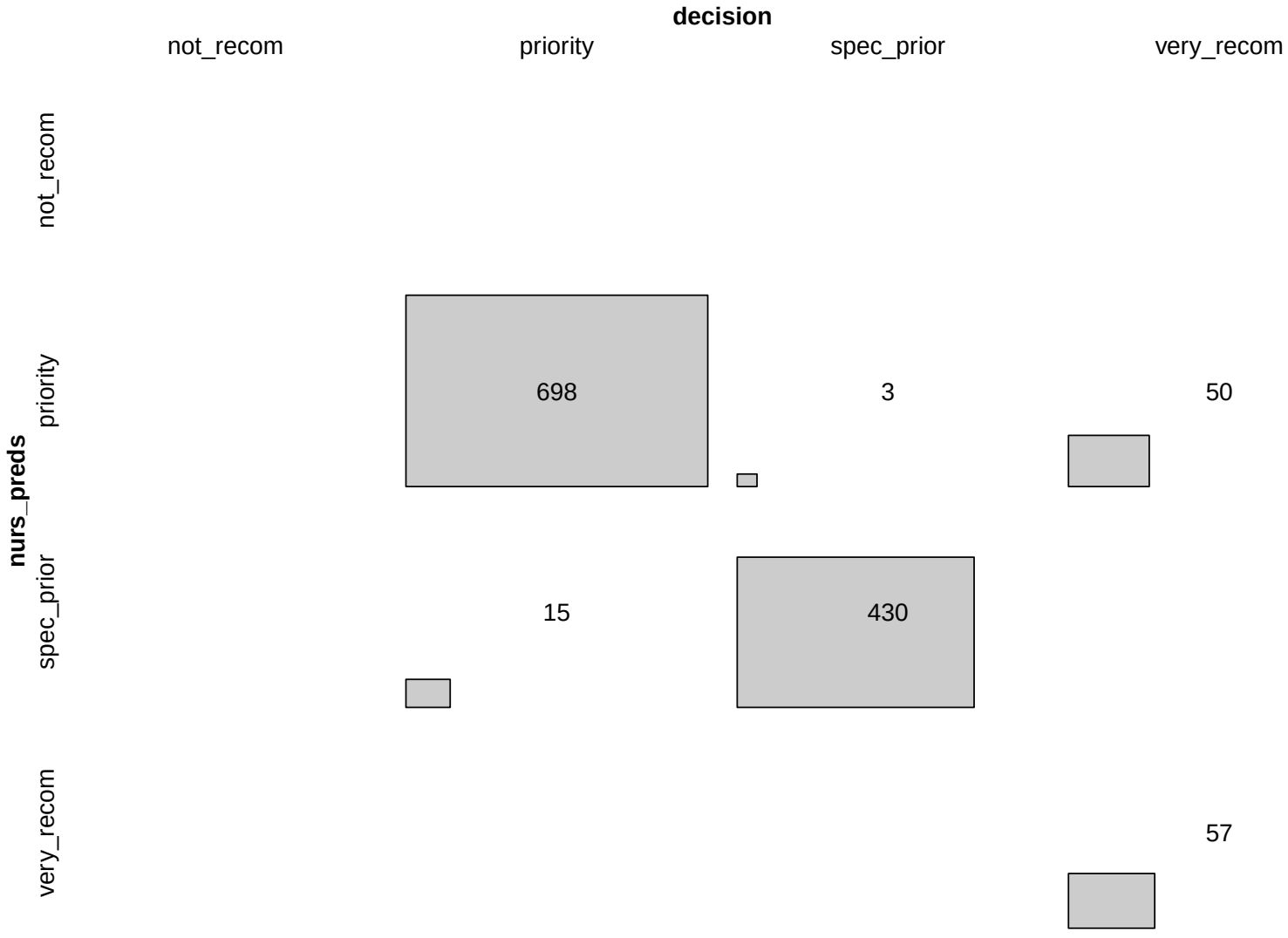
A categorical feature c has high importance for classification.

Assumption: c encodes useful information - association.

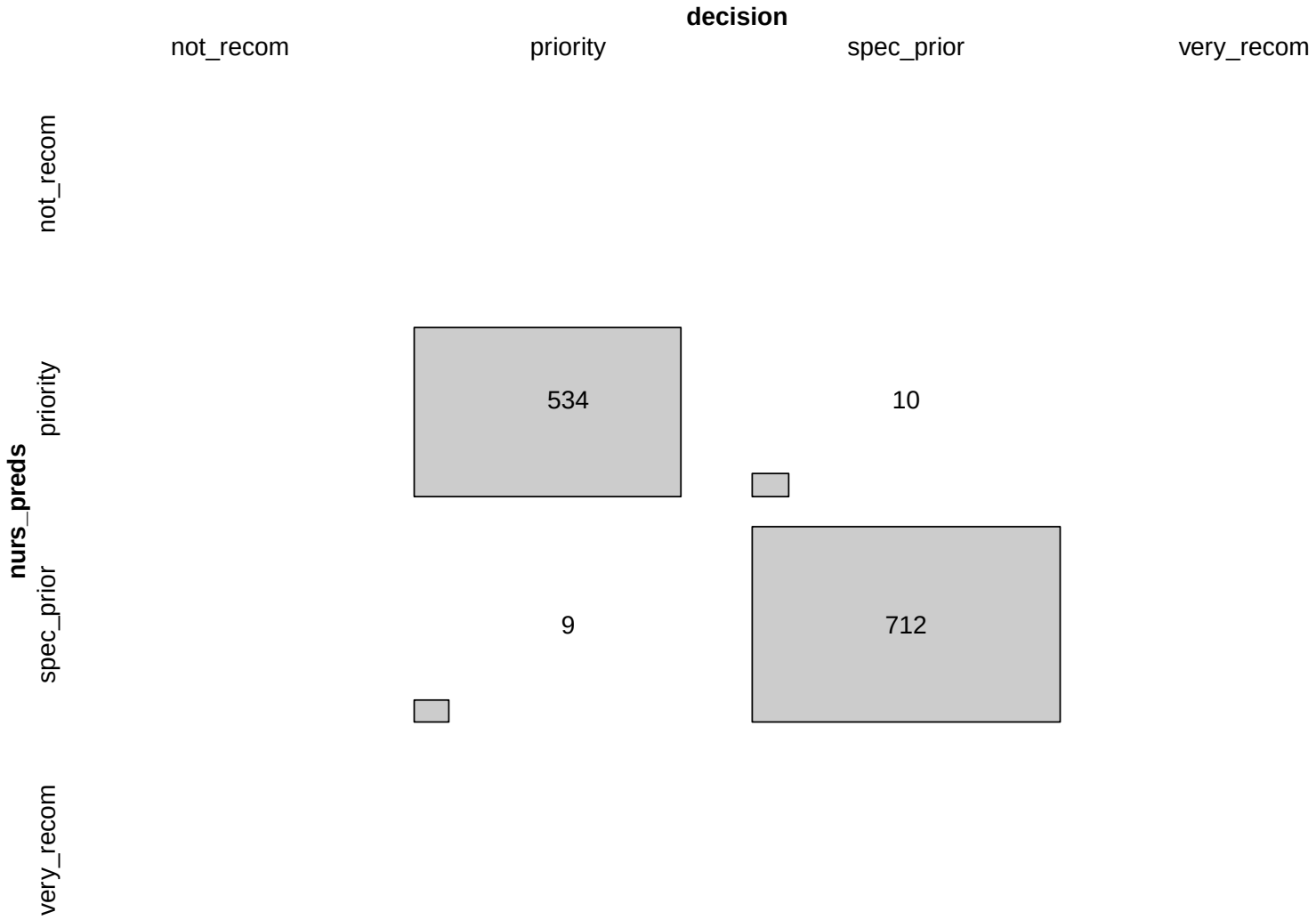
Significant changes in ratios of the class labels must be associated with different categories of c

The following plots are generated separately because there is no cotab_tile() :-(

Confusion Matrix By health == “recommended”



Confusion Matrix By health == “priority”



Confusion Matrix By health == “not_recom”

What Makes A Variable Important? - Revisited

health feature - very important, and now we know why

health == "not_recom" is a useful decision stump: take outside model?

Thorough EDA should find this prior to modeling

Applied to Binary Classification

This situation can't come about during binary classification. One feature would be a perfect representation of target.

Can this technique still help?

Binary Classification 2×2 Table

Special case 2×2 table

vcd has a fourfold plot function: compares odds ratios by standardising pie areas

Force the fourfold to represent counts on the radius:

$$count \propto \sqrt{area}$$

fourfold also automatically stratifies by a third variable! Just a snippet of code required.

This is a very nifty shortcut! vcd always wants to preserve the given table structure. Any other plotting method could not be done in one line of code.

German Data Set - Credit Rating

A mix of discrete and continuous. Target is rating: "bad" (30%) or "good" (70%)

A predicted bad rating on a true bad rating - True Positive

A predicted good rating on a true bad rating - False Negative

Very high cost per False Negative: Offer of credit to a likely defaulter!

Random Forest can only optimise 0-1 loss

Training data is balanced by over sampling.

```
##                rating
## german_preds bad good
##          bad   34   28
##          good  42  185
```

```r
# accuracy
sum(diag(confmat))/sum(confmat)
```

```
## [1] 0.7577855
```

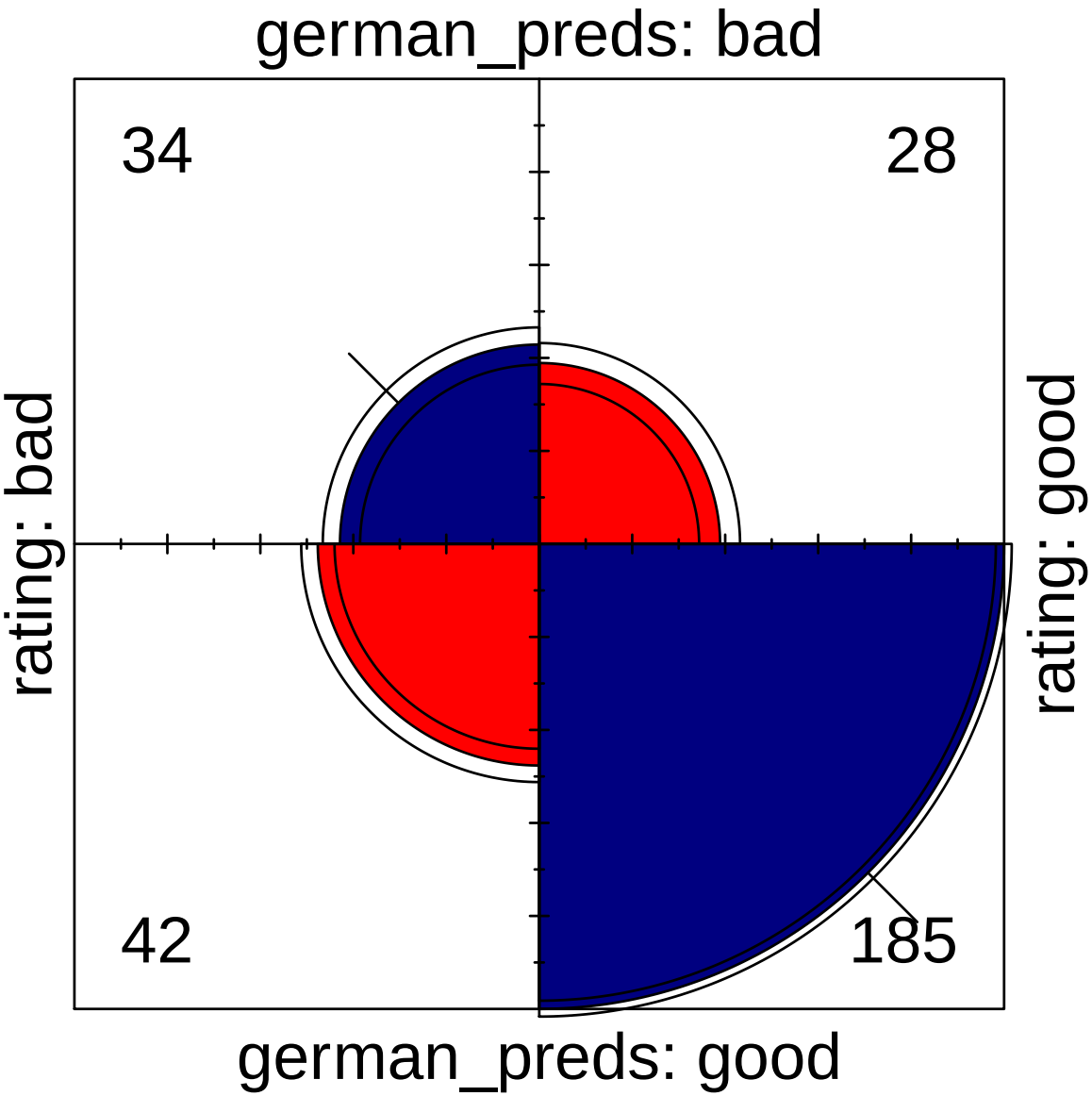
```r
# False Negative Rate FN / (TP + FN)
confmat[2, 1]/sum(confmat[, 1])
```

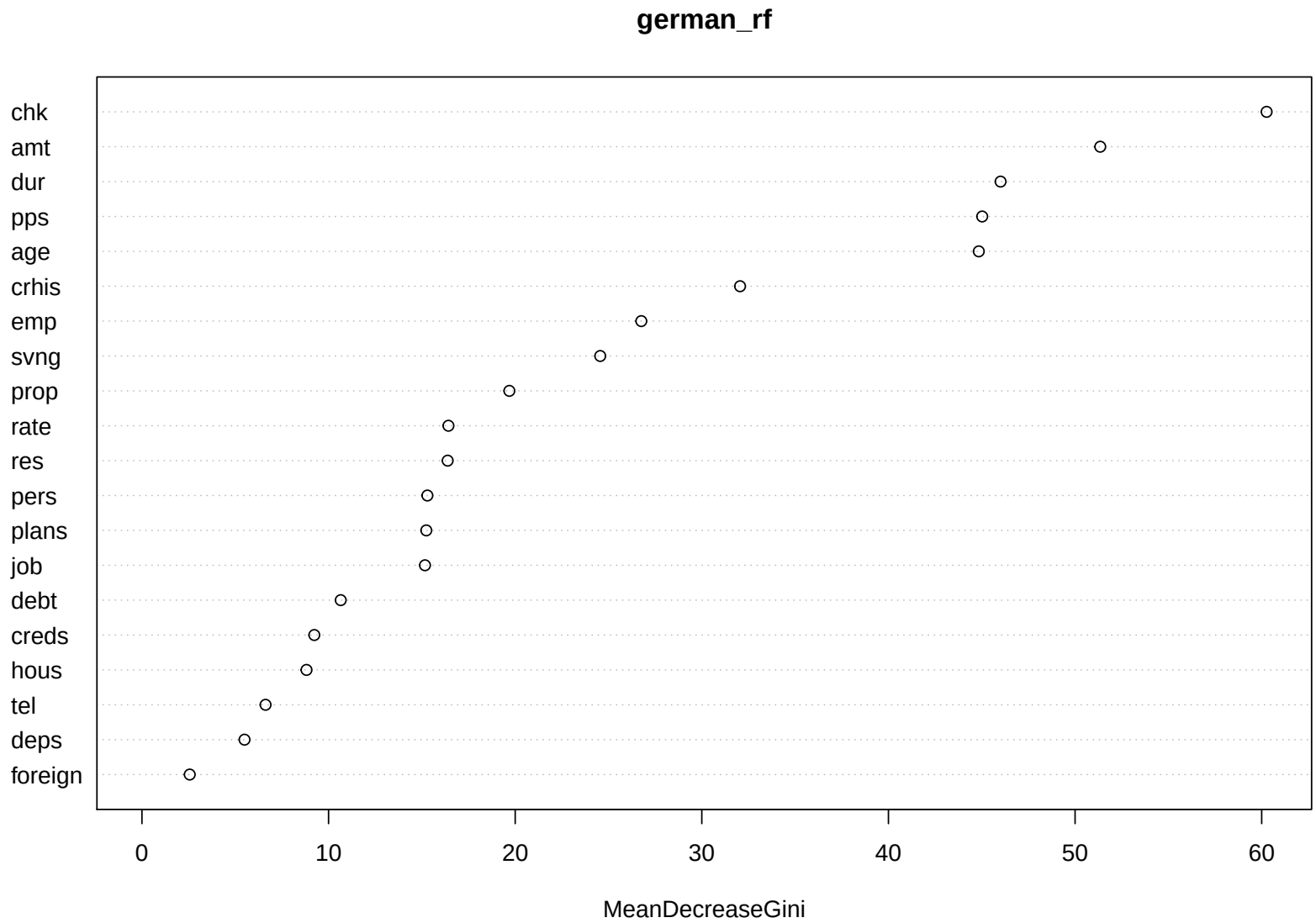
```
## [1] 0.5526316
```

```r
# False Ommision Rate FN / (TN + FN)
confmat[2, 1]/sum(confmat[2, ])
```

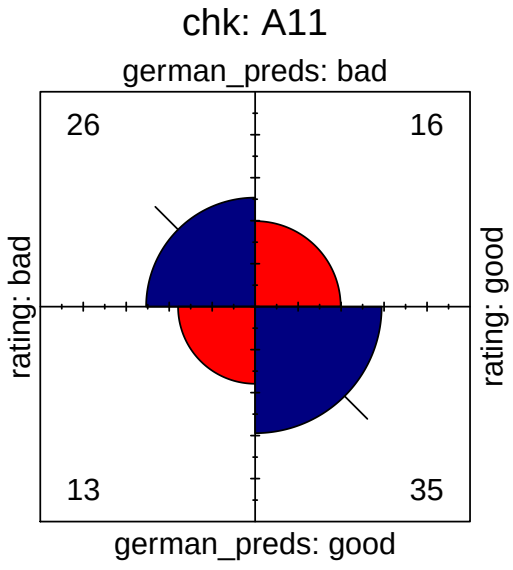
```
## [1] 0.185022
```

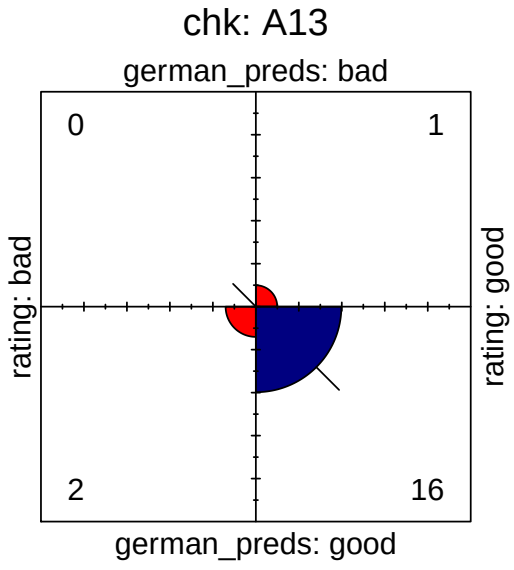
```r
# std = "all.max" suppresses balancing the areas
# behaves like a rose plot
fourfold(confmat, std = "all.max")
```

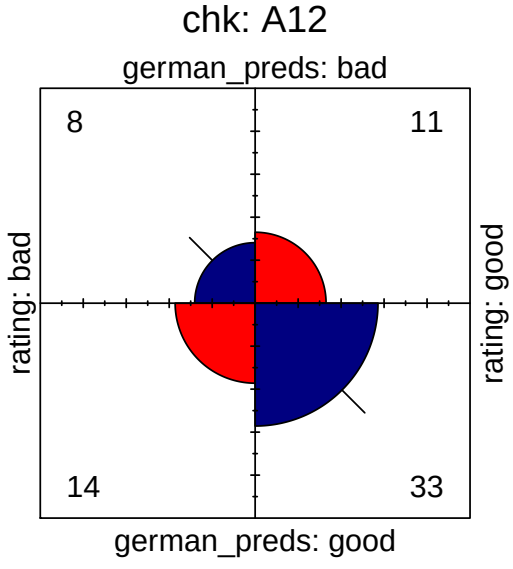


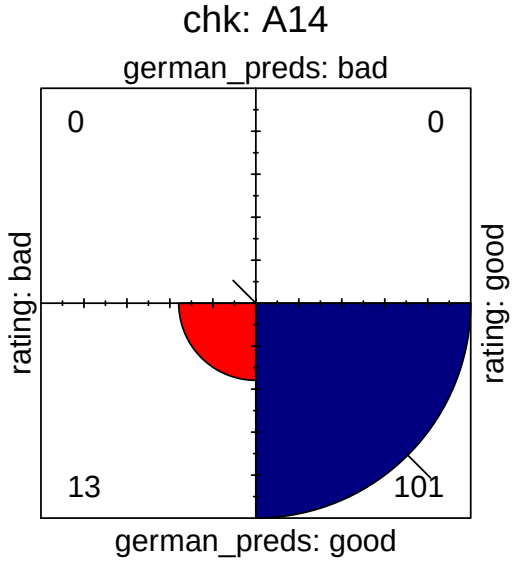

german_rf
chk
amt
dur
pps
age
crhis
emp
svng
prop
rate
res
pers
plans
job
debt
creds
hous
tel
deps
foreign
0
10
20
30
40
50
60
MeanDecreaseGini

Fourfold Rose Plot - Confidence Matrix By chk

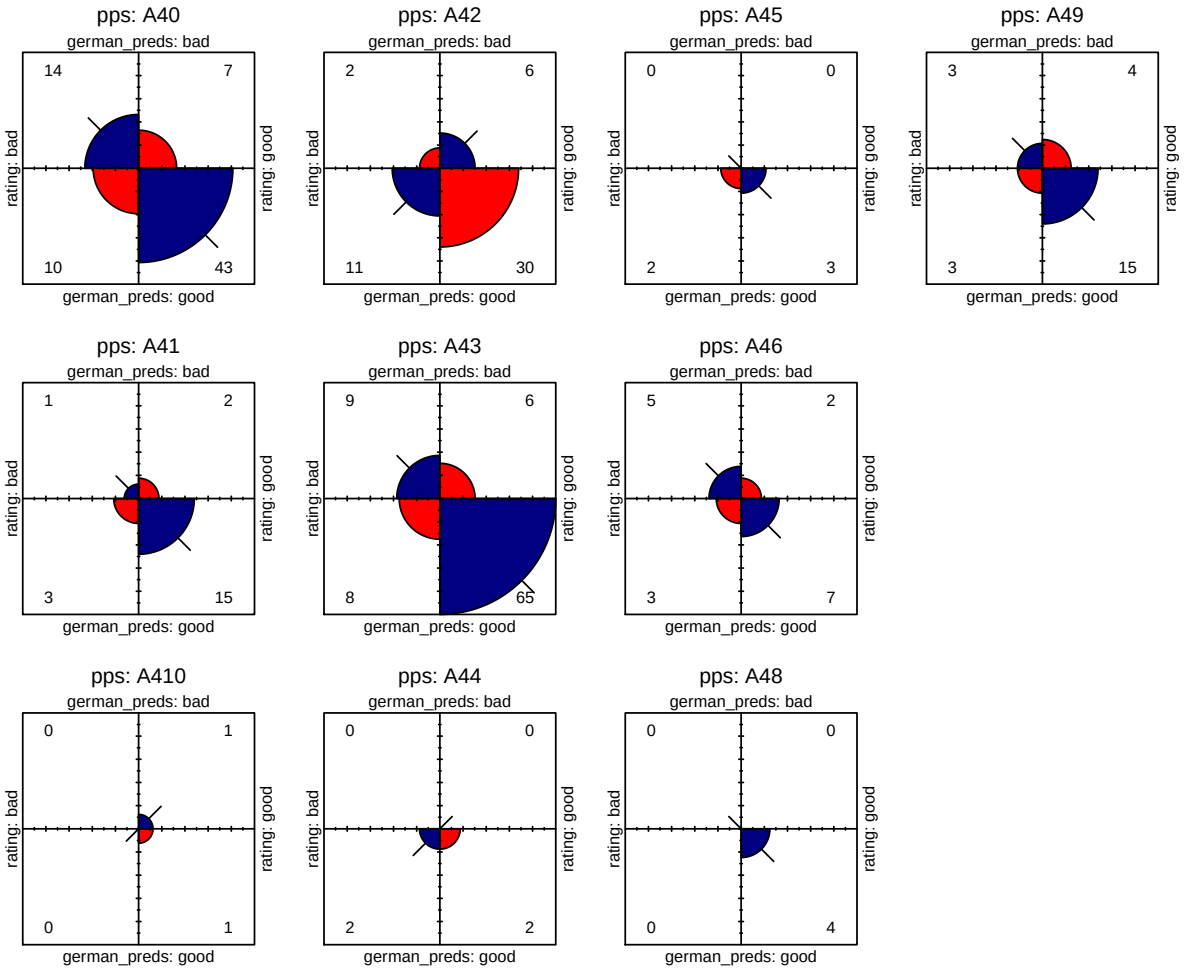




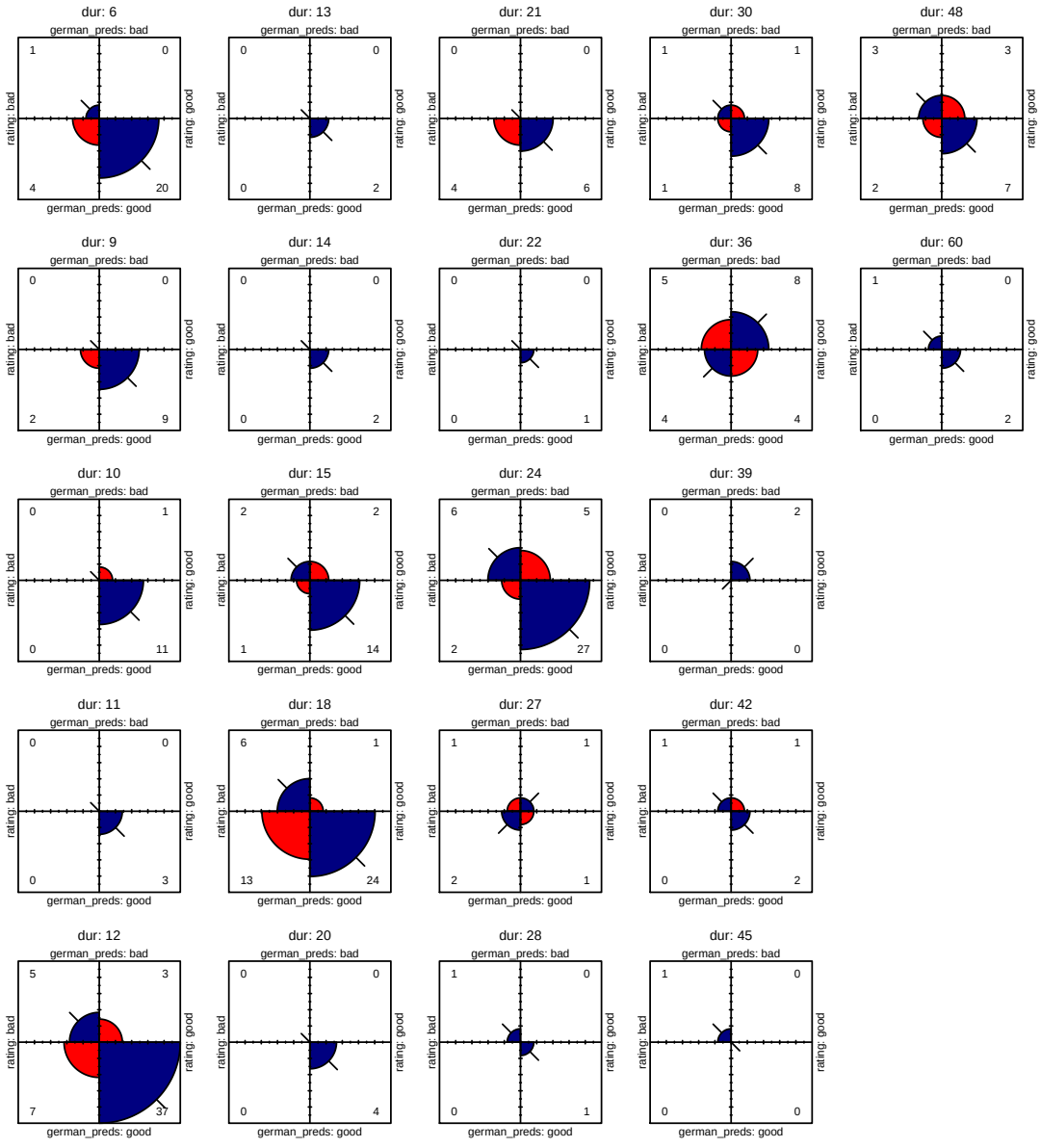




This method REALLY scales!



Integer Valued Feature with 22 Unique Values



2 & 3-Way Interactions Possible - Log Odds Ratio

The odds ratio for the binary confusion matrix conveniently rearranges to:

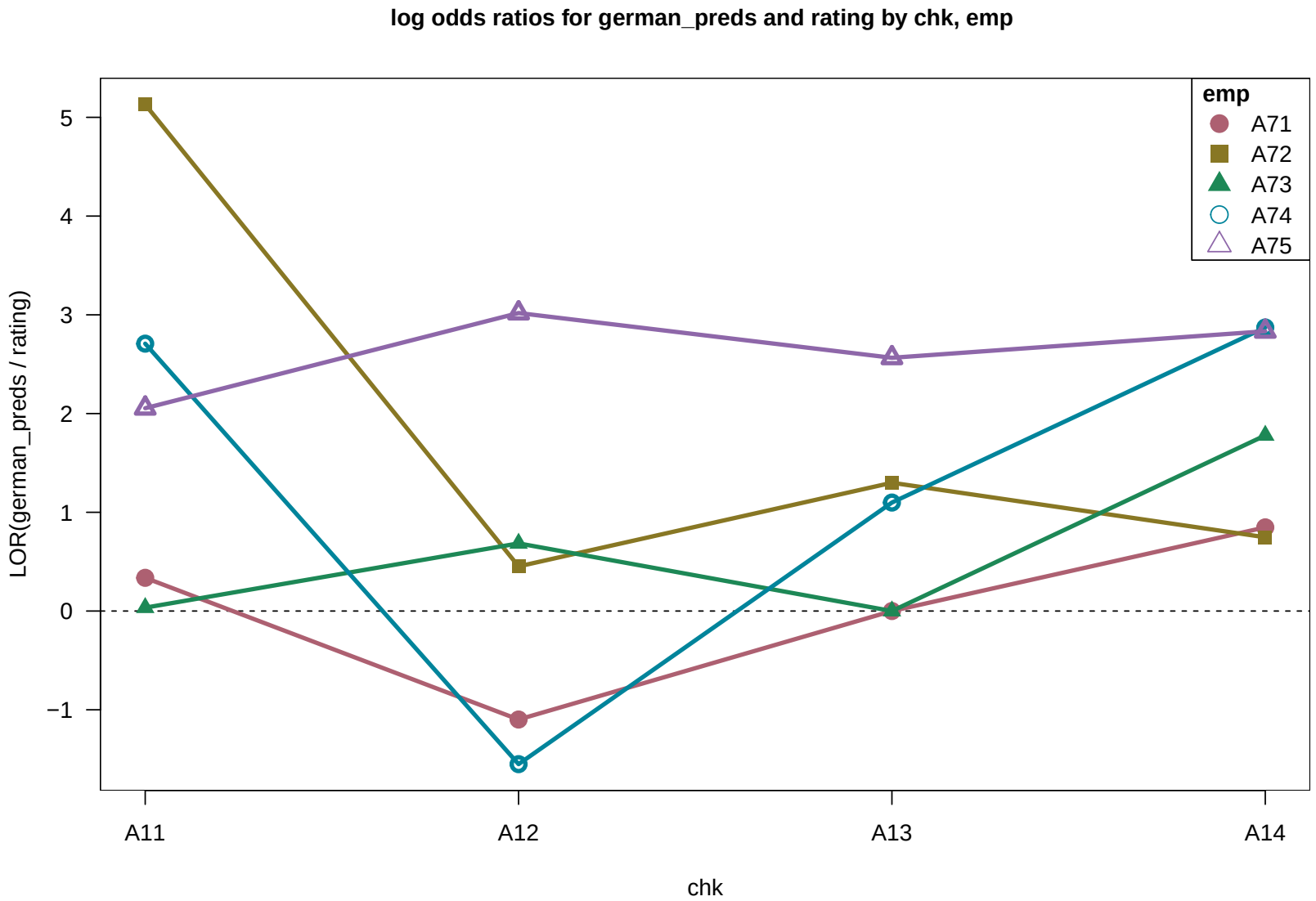
$$\phi = \log\left(\frac{\mathrm{TP} \times \mathrm{TN}}{\mathrm{FP} \times \mathrm{FN}}\right)$$

Note, all the T in the numerator and all the F in the denominator.

Very easy to interpret:

- $\phi \in \mathbb{R}^+$ is good but check confidence intervals
- $\phi > 3 \implies$ ratio of T:F $\approx 20:1$

Bonus: confidence intervals!


log odds ratios for german_preds and rating by chk, emp
LOR(german_preds / rating)
5
4
3
2
1
0
−1
A11
A12
A13
A14
chk
emp
A71
A72
A73
A74
A75

Slicing a Confusion Matrix Summary

Important points:

- Use whenever you have suitable categorical variables
- Feature importance measure; a helpful guide but not essential.
- Drill into sources of bias
- Rose plots are intuitive for 2×2 tables and easy to produce
- vcd fourfold scales to tens of levels on one category
- Explore 2 and 3-way interactions with log odds ratios
- Log odds ratios compromise clarity on accuracy measures such as FNR

Example Four: Handling Ordinal Classification Correctly

MSPatients Overview

Two Neurologists diagnose the same 218 patients on a severity scale

This is an inter-rater agreement table. We'll pretend it's a confusion matrix. Classes not completely balanced.

Baseline Score

```
## tru
##  Certain Probable Possible Doubtful
##       95       66       22       35

## tru
##   Certain  Probable  Possible  Doubtful
## 0.4357798 0.3027523 0.1009174 0.1605505

##   Certain
## 0.4357798

## Accuracy - sum(diag)/total

## [1] 0.4449541
```

The MSP Confusion Matrix

```
##             tru
## prd          Certain Probable Possible Doubtful
##   Certain         43        8        0        1
##   Probable        36       22        7        0
##   Possible        12       27        8       10
##   Doubtful         4        9        7       24
```

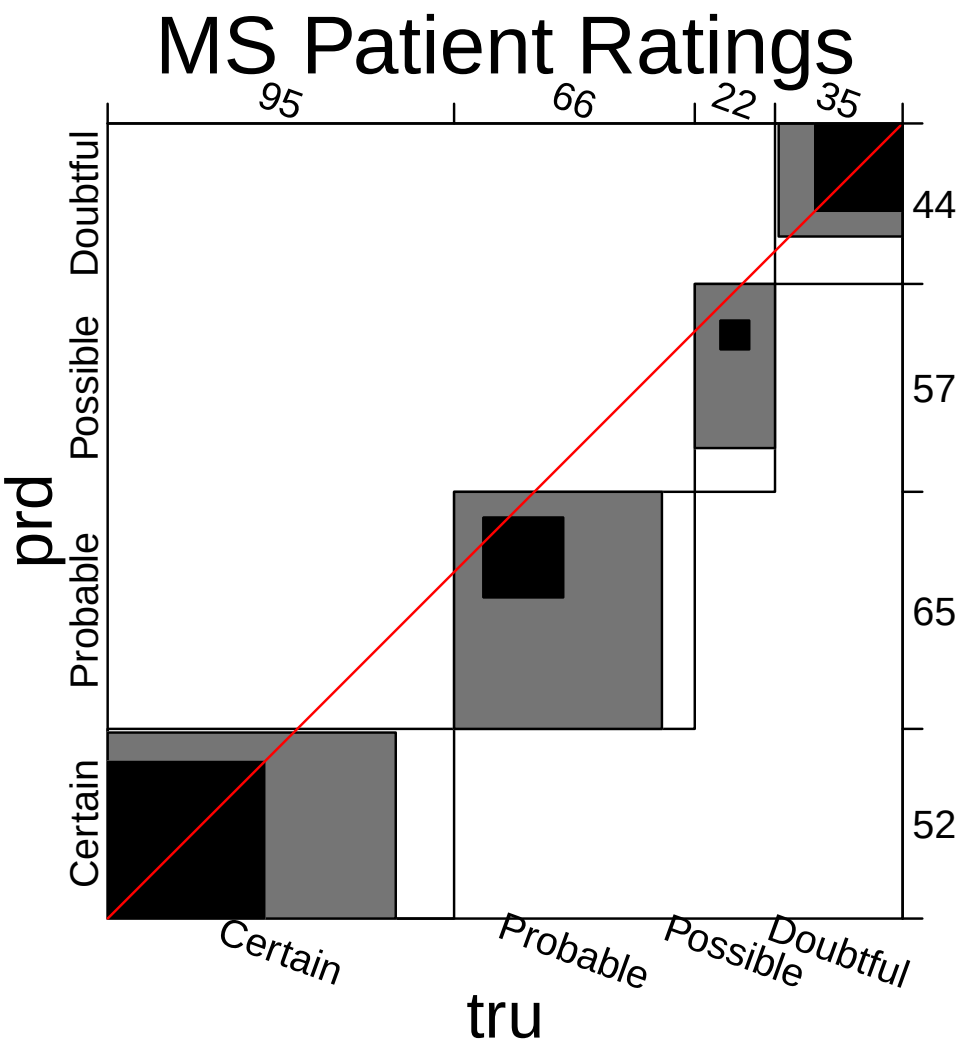
Obviously it's not a constant model and doesn't look like a random guess. So what's wrong?

Check Cohen's κ to correct for class imbalance

```
##
## Kappa                 lwr       upr
##   Unweighted 0.1728083 0.3411072
##   Weighted   0.4987456 0.6785713
```

Weighted indicates that off-by-n errors are important

Agreement Plot




```
##                Bangdiwala Bangdiwala_Weighted
```

Handling Ordinal Classification - Summary

- Ordinal classification is a special case
- Near disagreement needs to be weighted
- Weighted κ and Bangdiwala stats useful
- Agreement plot is ideal
- Plot reveals differences in inter-rater marginal totals, where statistics are insensitive

It's easy to rush into using the latest tools and technologies. Most Machine Learning is based on years of statistical research. That work is still just as relevant as ever. It's worth reflecting on tried and tested techniques that might easily address today's challenges.

End