Technische Universität Wien
Vienna University of Technology

Visualizing Independence Using Extended Association and Mosaic Plots

Achim Zeileis David Meyer Kurt Hornik

Overview

- The independence problem in 2-way contingency tables
  - Standard approach: χ^2 test
  - Alternative approach: max test
- Visualizing the independence problem
  - Association plots
  - Mosaic plots
- Extensions
  - Visualization & significance testing
  - HCL instead of HSV colors
  - Implementation in `grid`
  - Multi-way tables
- The `vcd` package

The independence problem

Standard approach:

- Analyze the relationship between two categorical variables based on the associated 2-way contingency table.
- Measure the discrepancy between observed frequencies $\{n_{ij}\}$ and expected frequencies under independence $\{\widehat{n}_{ij}\}$ by the Pearson residuals:

$$r_{ij} = \frac{n_{ij} - \widehat{n}_{ij}}{\sqrt{\widehat{n}_{ij}}}.$$

- Use the Pearson X^2 statistic for testing:

$$X^2 = \sum_{ij} r_{ij}^2,$$

which has an asymptotic χ^2 distribution.

The independence problem

Alternative approach(es):

- There are many conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda\left(\left\{r_{ij}\right\}\right)$.

- In particular:

$$M = \max_{ij} \left| r_{ij} \right| .$$

  Then, every residual exceeding the critical value c_α violates the null hypothesis at level α.

- Instead of relying on unconditional limiting distributions, perform a permutation test, either by simulating or computing the conditional permutation distribution of $\lambda\left(\left\{r_{ij}\right\}\right)$.

The independence problem

Relationship between hair color and eye color among 328 female students:

Hair color	Eye color: Brown	Blue	Hazel	Green	Total
Black	36	9	5	2	52
Brown	81	34	29	14	158
Red	16	7	7	7	37
Blond	4	64	5	8	181
Total	137	114	46	31	328

$$X^2 = 112.30 \qquad p = 0$$
$$M = 6.76 \qquad p = 0$$

The independence problem

Home and away goals in the Bundesliga in 1995:

Home goals	Away goals 0	1	2	3	4	5	6
0	26	16	13	5	0	1	0
1	19	58	20	5	4	0	1
2	27	23	20	5	1	1	1
3	14	11	10	4	2	0	0
4	3	5	3	0	0	0	0
5	4	1	0	1	0	0	0
6	1	0	0	1	0	0	0

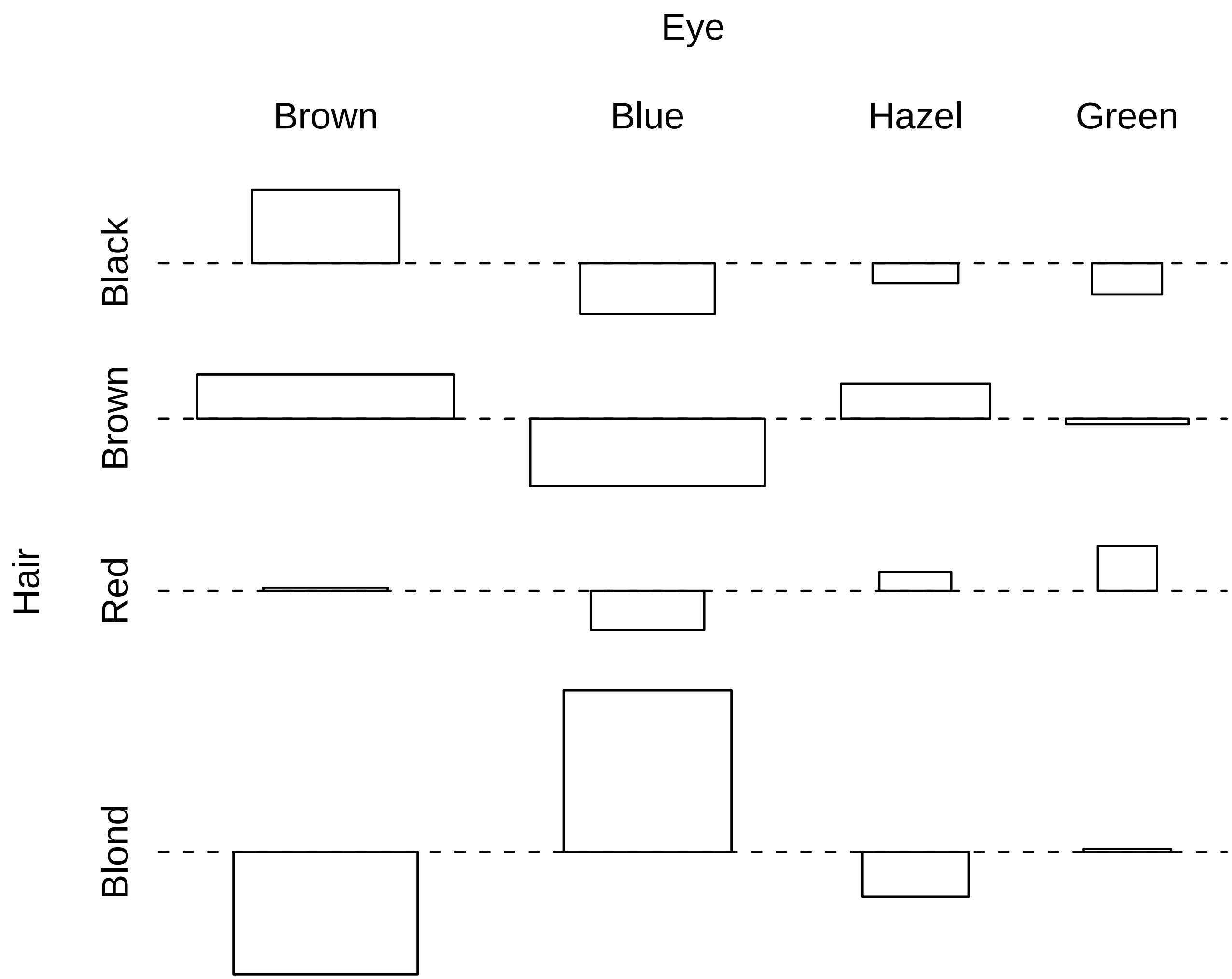
$$X^2 = 46.07 \qquad p = 0.121$$
$$M = 2.87 \qquad p = 0.355$$

Visualization

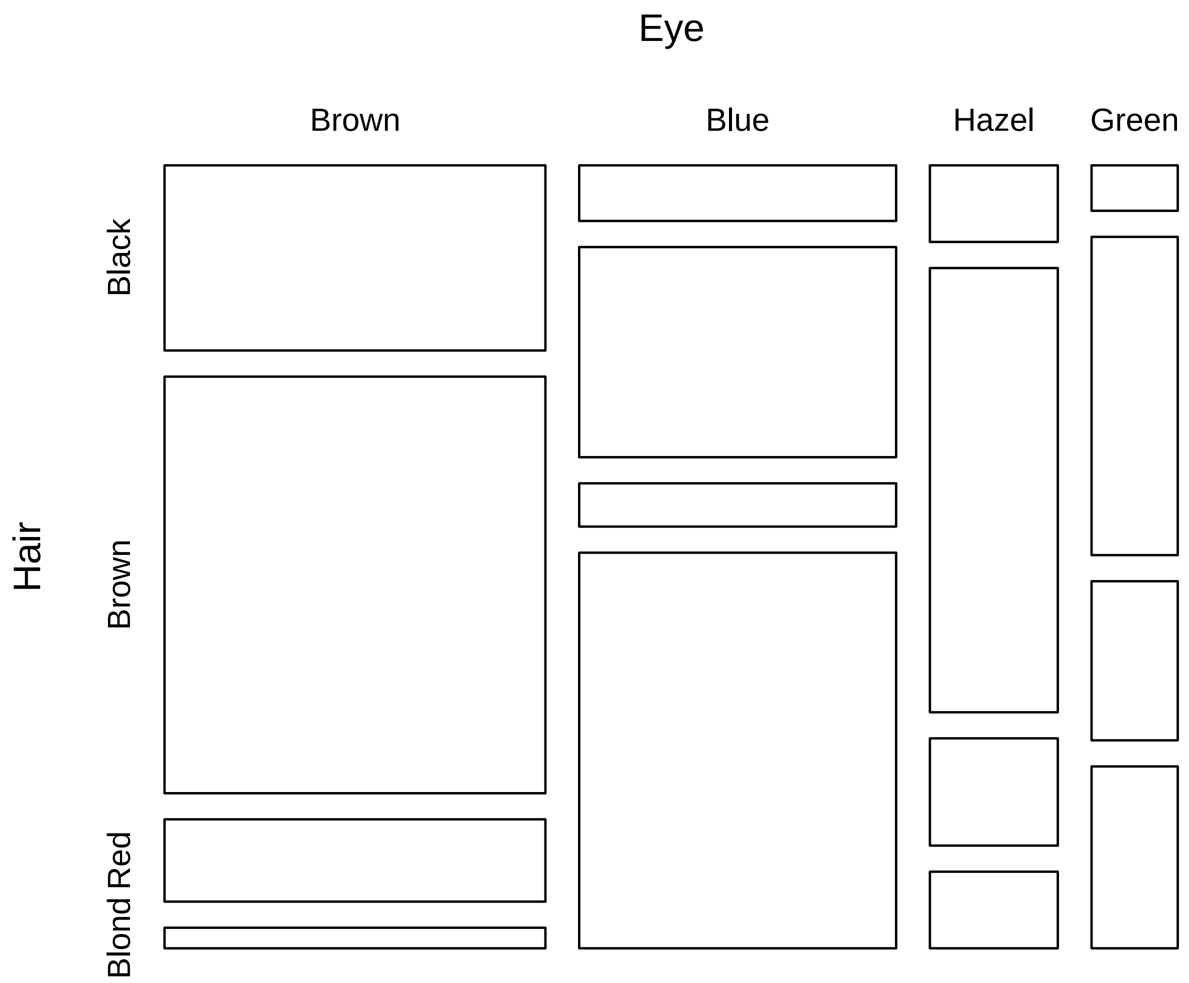
Association plot: display for the Pearson residuals $\{r_{ij}\}$ and the raw residuals $\{n_{ij} - \widehat{n}_{ij}\}$ in an rectangular array.

Mosaic plot: display in which the sizes of the mosaic tiles is proportional to the observed frequencies $\{n_{ij}\}$.

Visualization



Visualization



HSV colors

Colors are commonly used to enhance these plots. In particular, Friendly (1994) suggested shadings for mosaic displays.

HSV colors

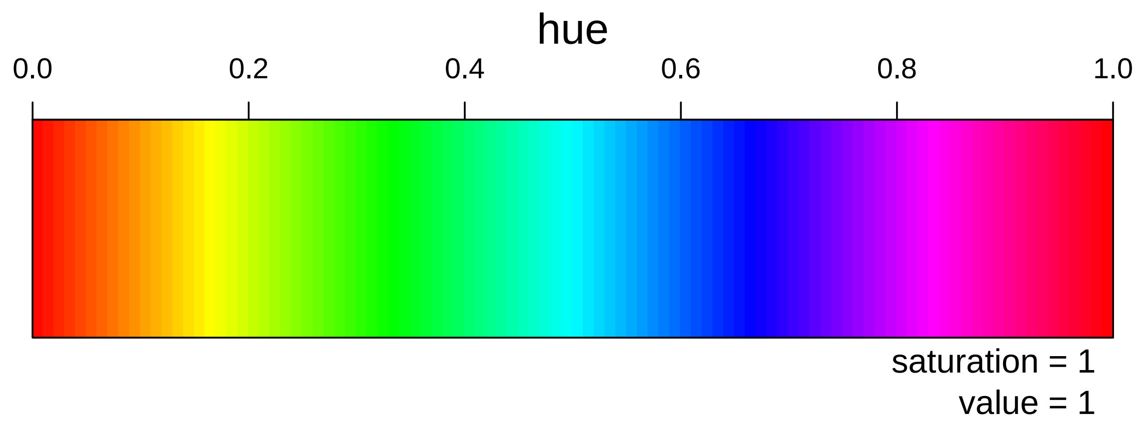
Colors are commonly used to enhance these plots. In particular, Friendly (1994) suggested shadings for mosaic displays.

In R these are implemented based on HSV colors.

The HSV color space is one of the most common implementations of color in many computer packages. Hue, saturation and value range in $[0, 1]$.

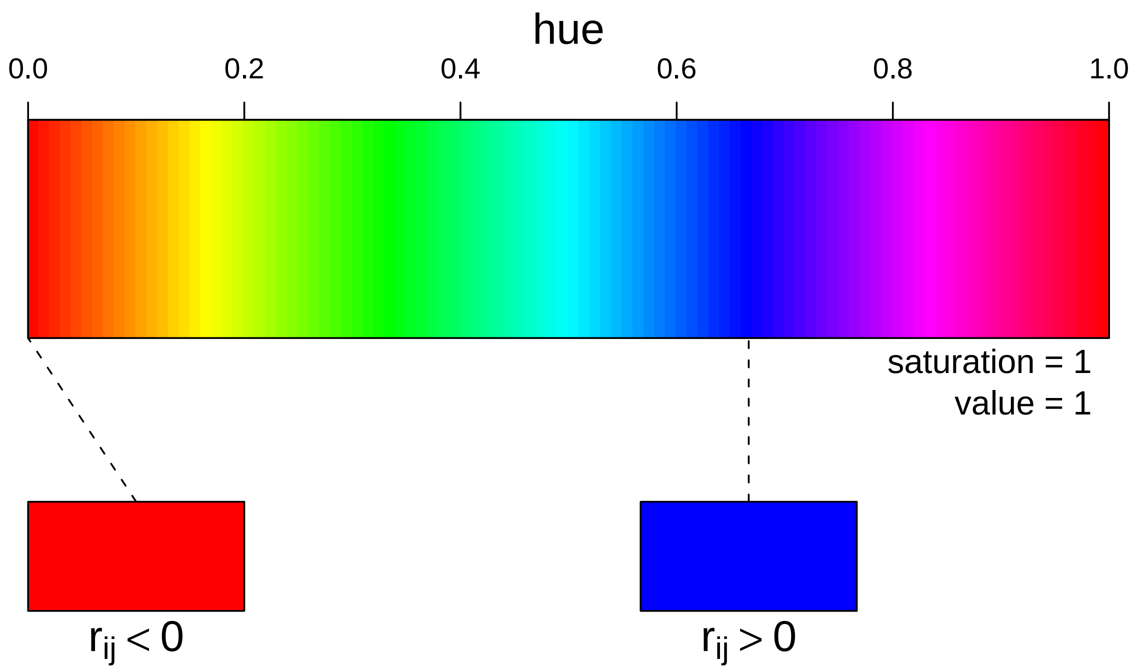
HSV colors

The hue is typically used to code the *sign* of the residuals.



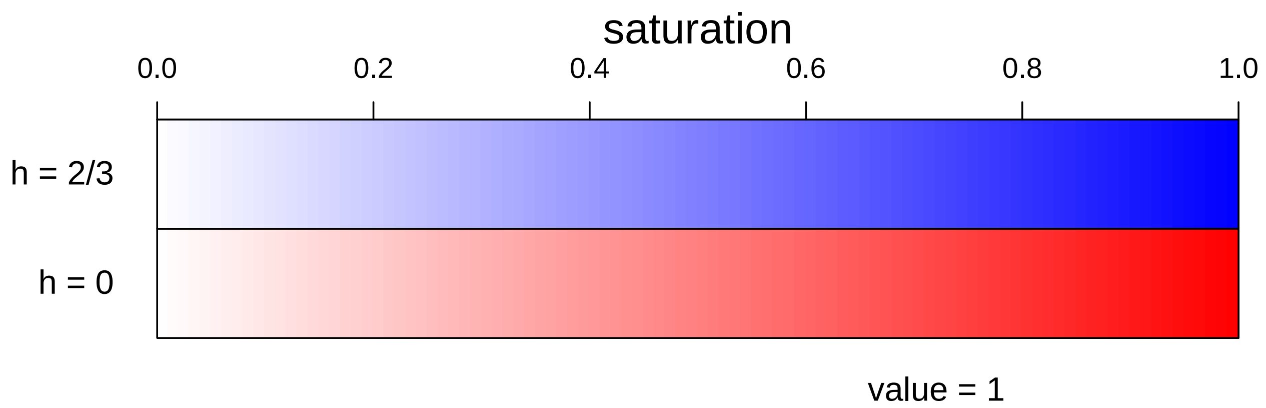
HSV colors

The hue is typically used to code the *sign* of the residuals.



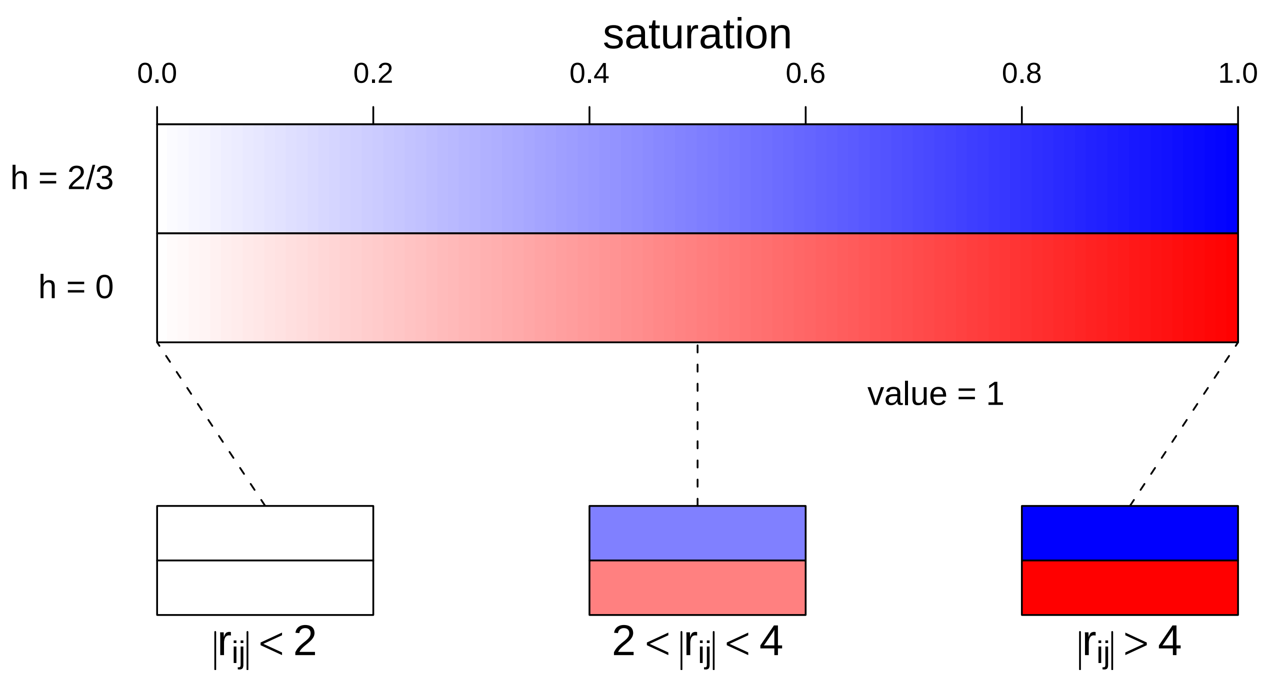
HSV colors

Friendly's extended mosaic displays use the saturation to code the *absolute size* of the residuals.



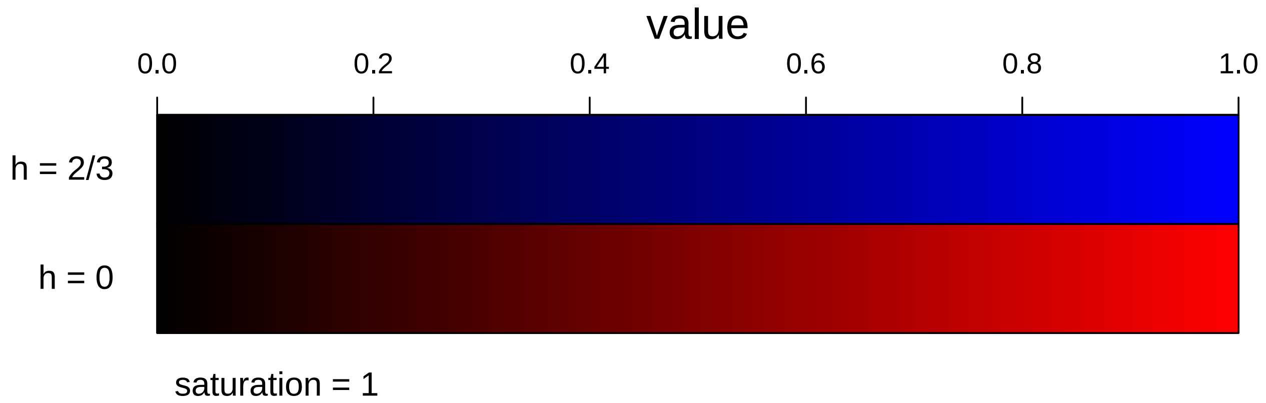
HSV colors

Friendly's extended mosaic displays use the saturation to code the *absolute size* of the residuals.



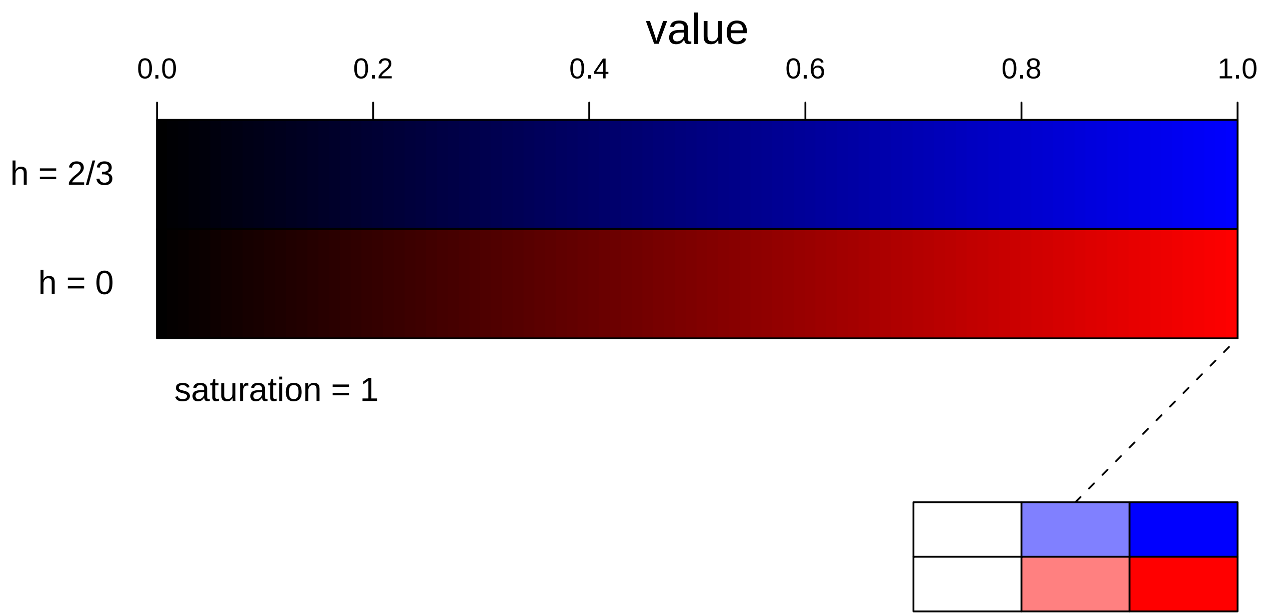
HSV colors

Value is currently not used for coding, always set to 1.

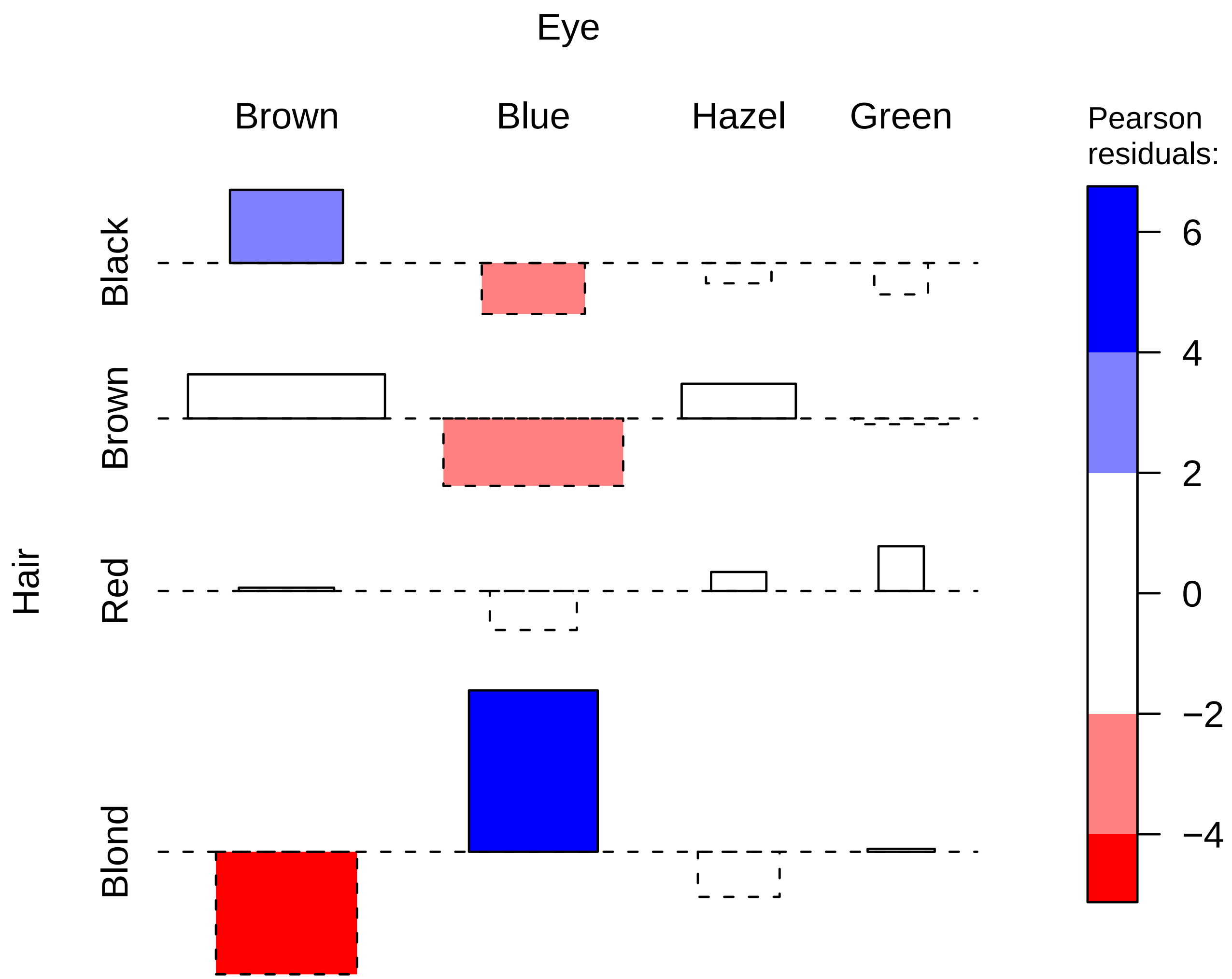


HSV colors

Value is currently not used for coding, always set to 1.

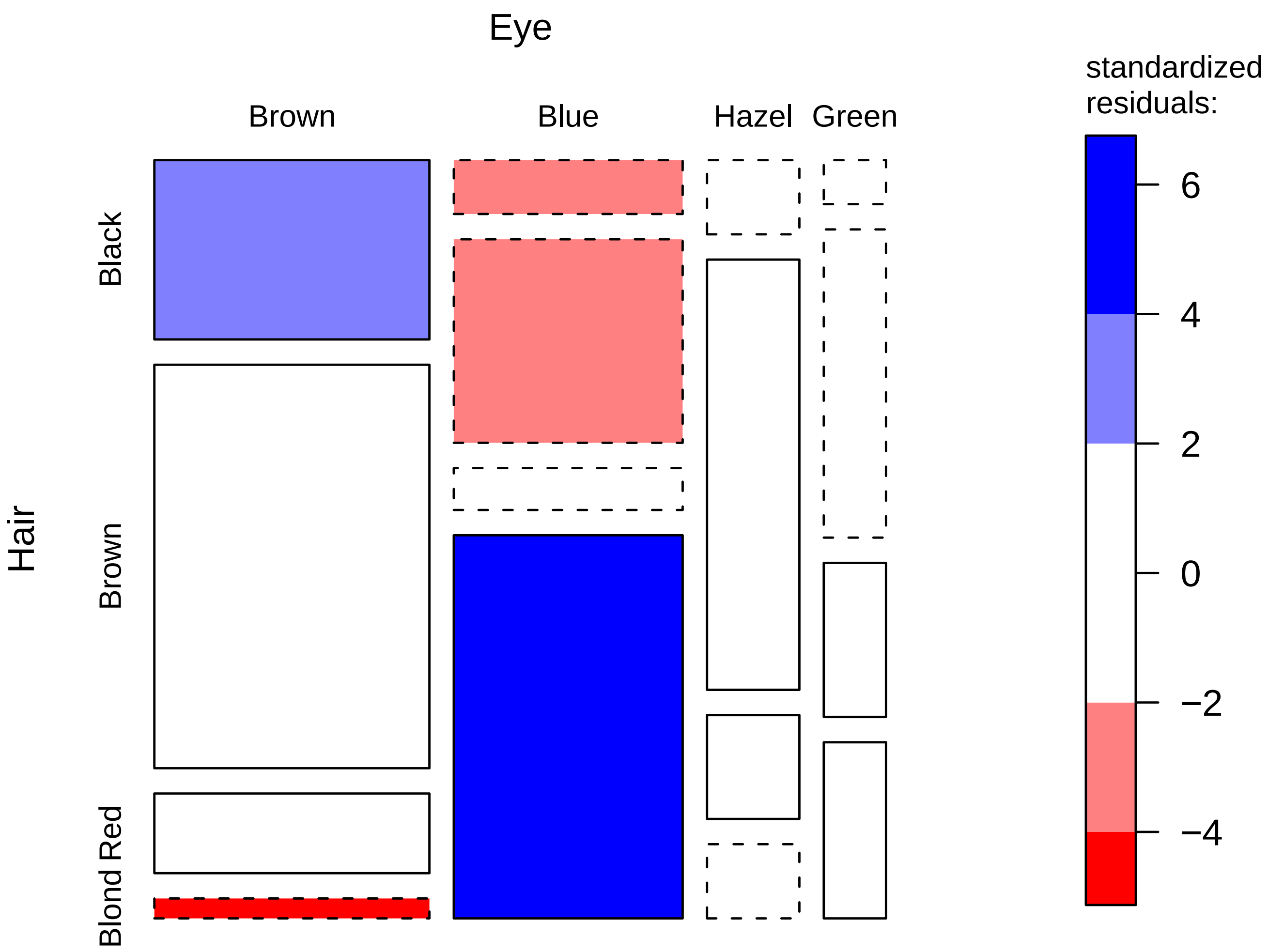


HSV colors

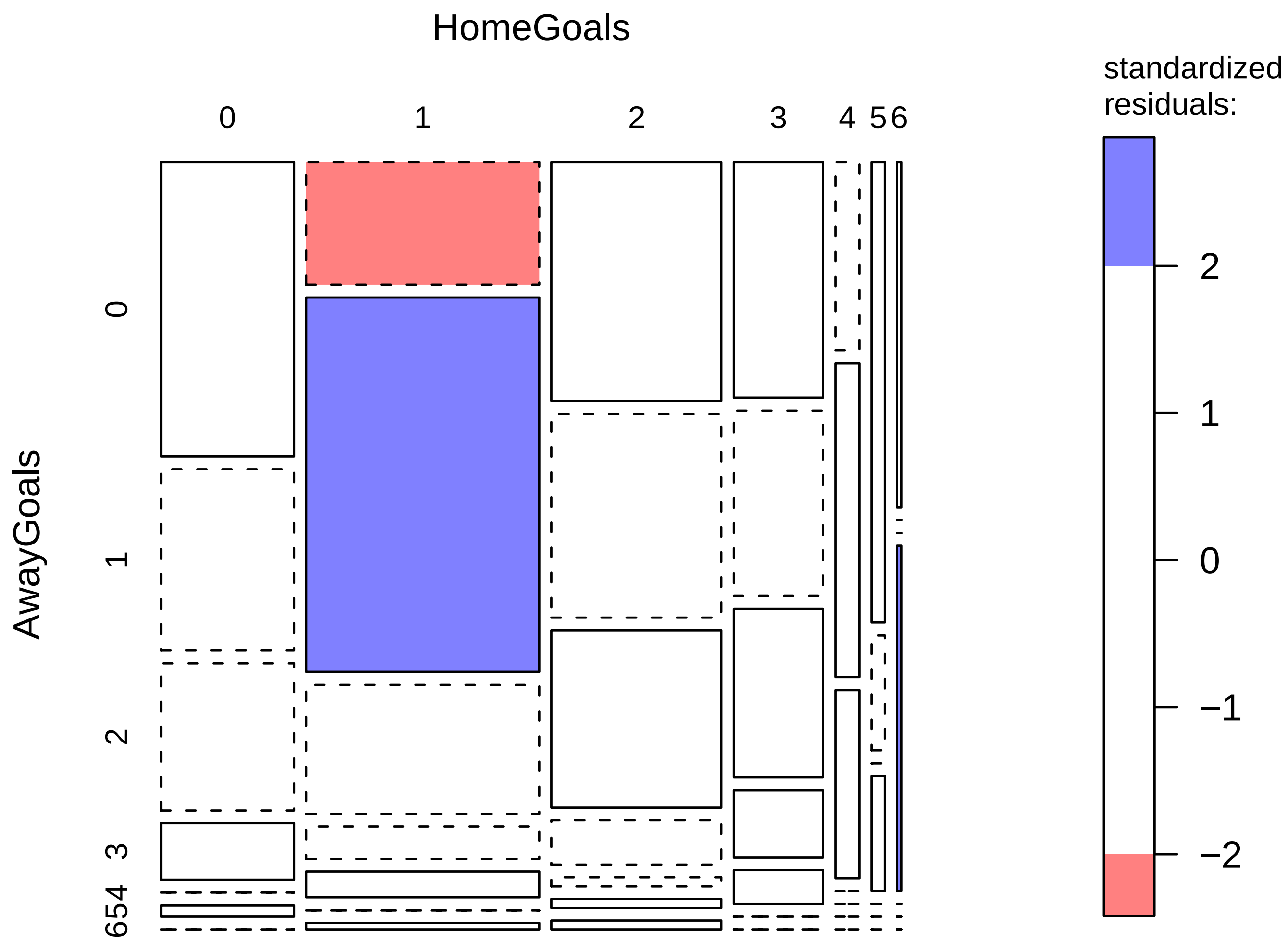


Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
Pearson
residuals:
6
4
2
0
−2
−4

HSV colors


Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
standardized
residuals:
6
4
2
0
−2
−4

Visualization & testing



Visualization & testing

Intuition: colored cells convey the impression that there is significant dependence.

Visualization & testing

Intuition: colored cells convey the impression that there is significant dependence.

Currently this is not true. But it can be achieved by using the 90% and 99% critical values for the max statistic M instead of 2 and 4.

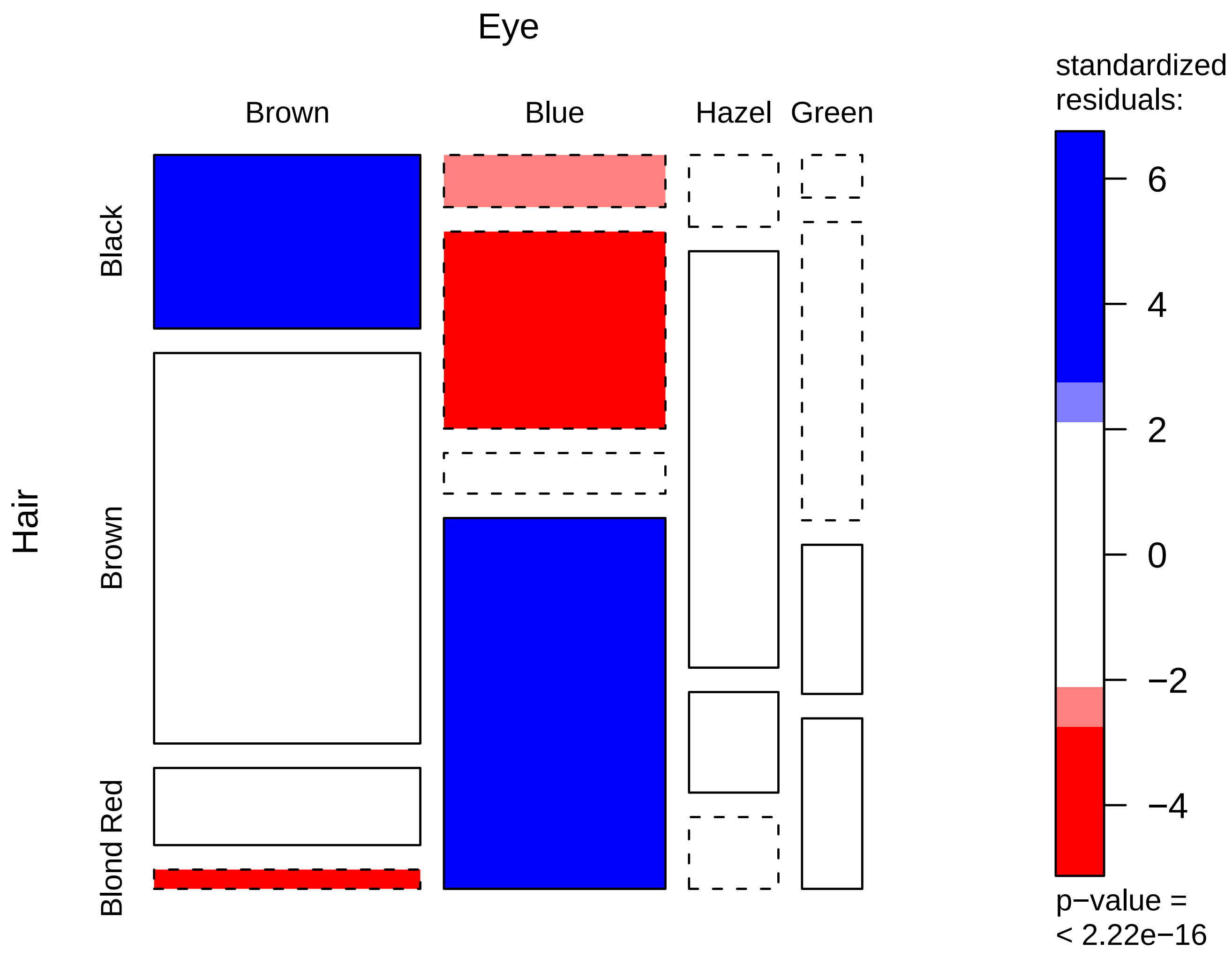
Advantage:

- color $\Leftrightarrow$ significance
- highlights the cells which "cause" the dependence (if any).

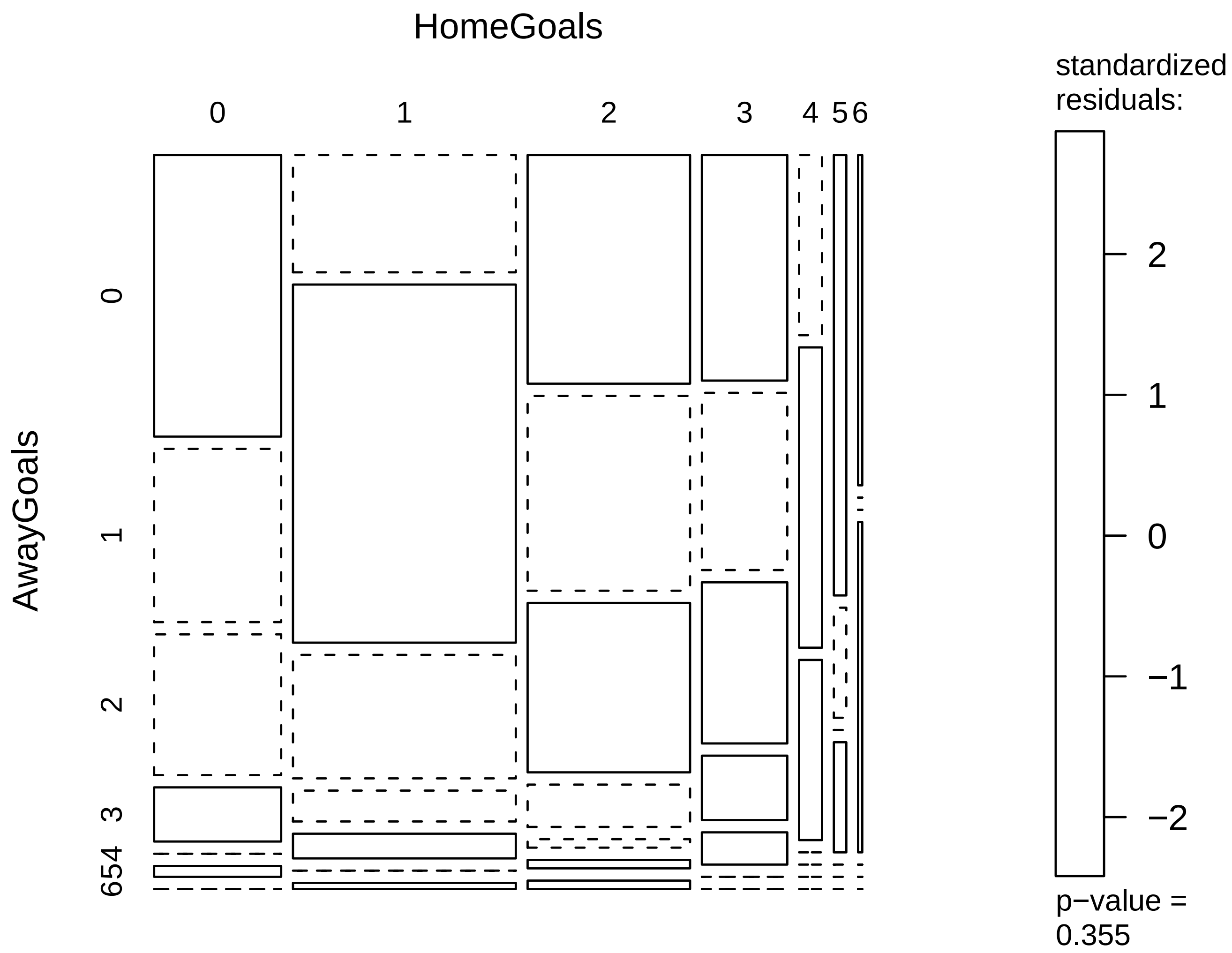
Disadvantage:

- does not work for the χ^2 test (or any other functional $\lambda(\cdot)$).

Visualization & testing

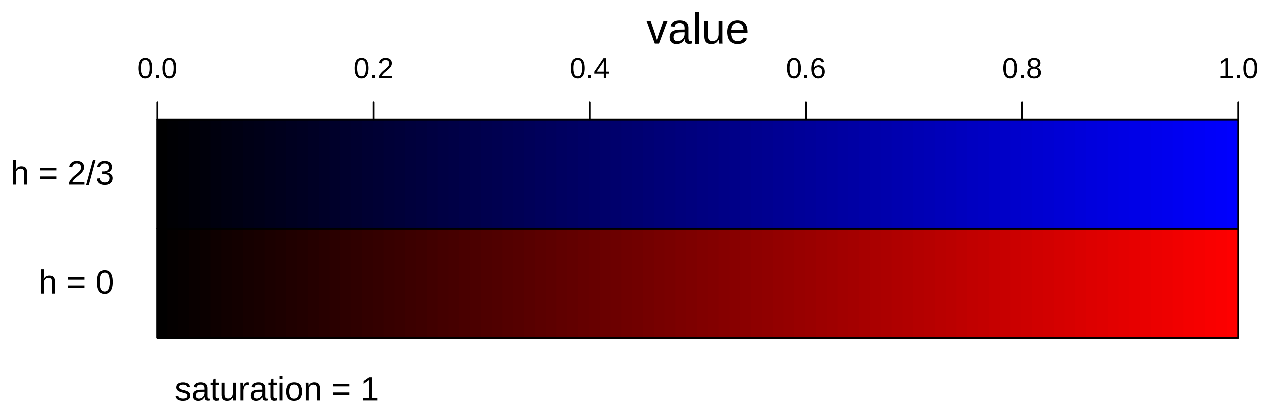


Visualization & testing



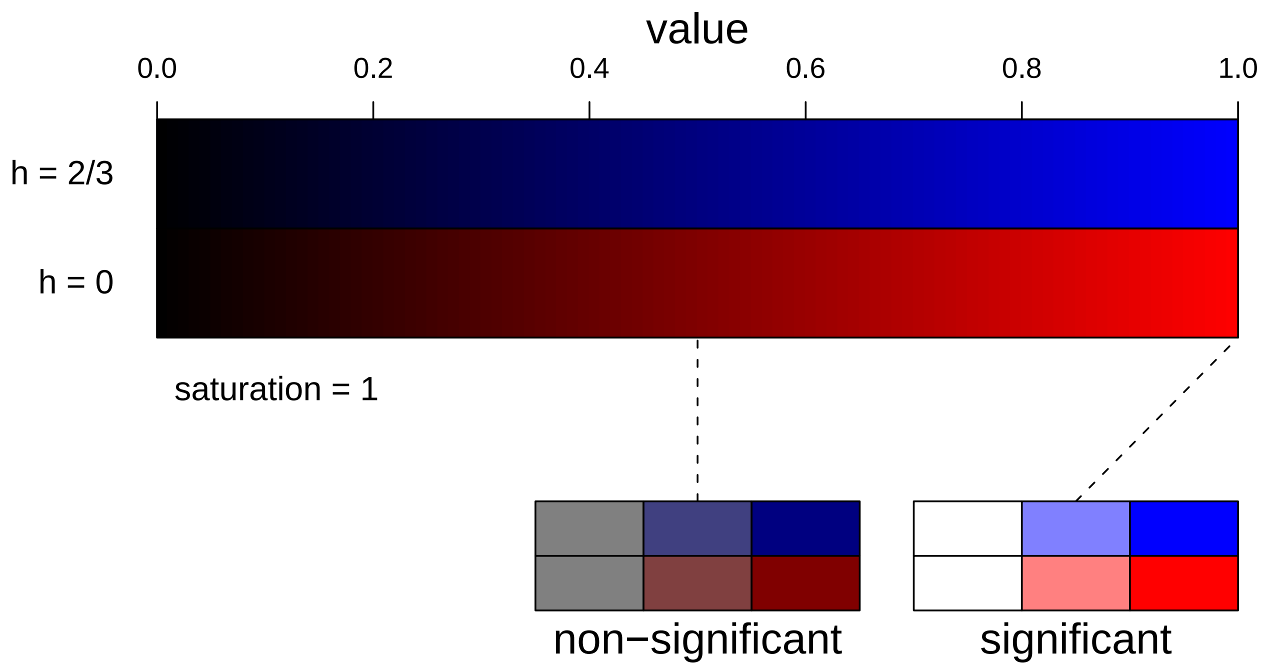
Visualization & testing

Use value to code the *result of a significance test* for independence.

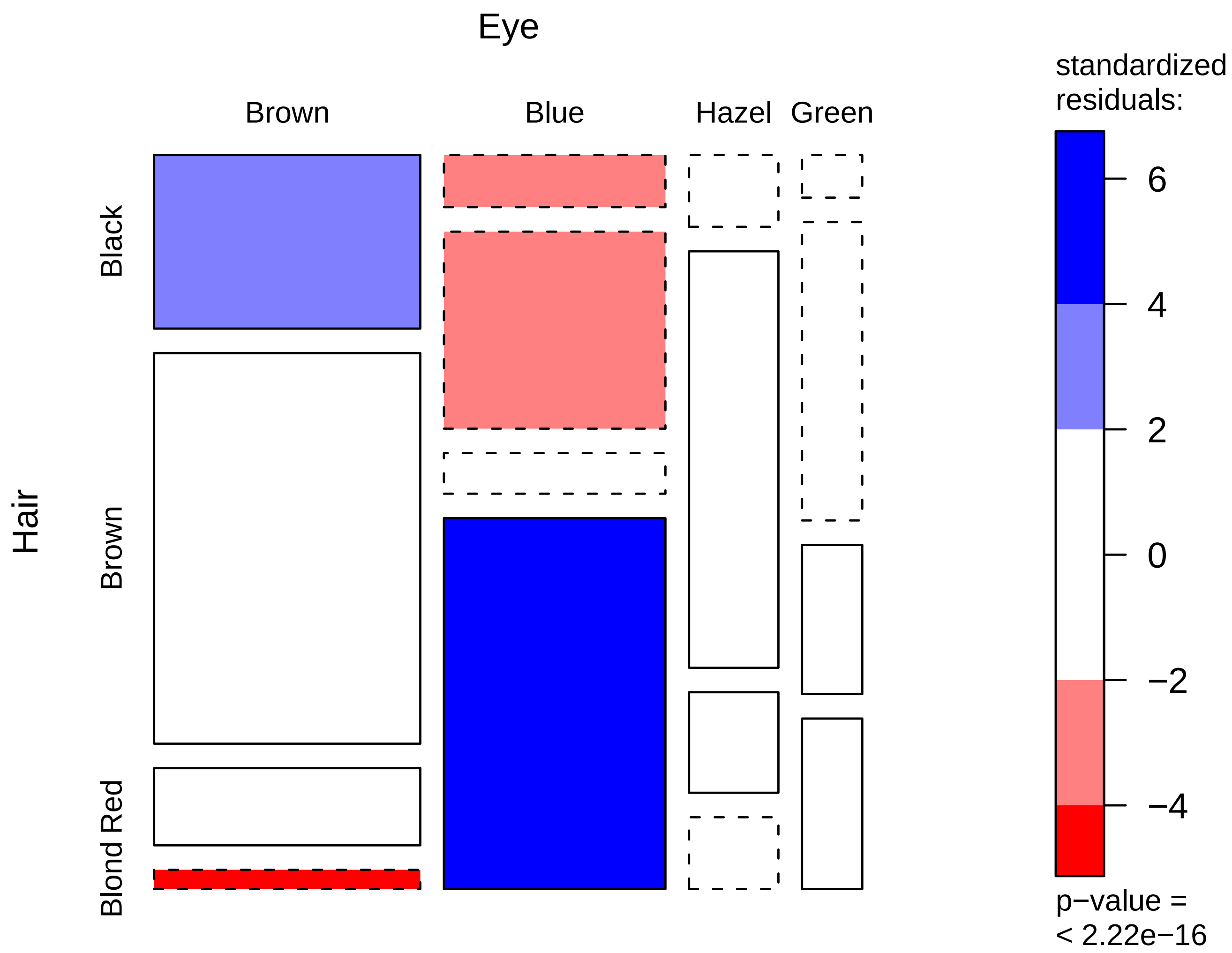


Visualization & testing

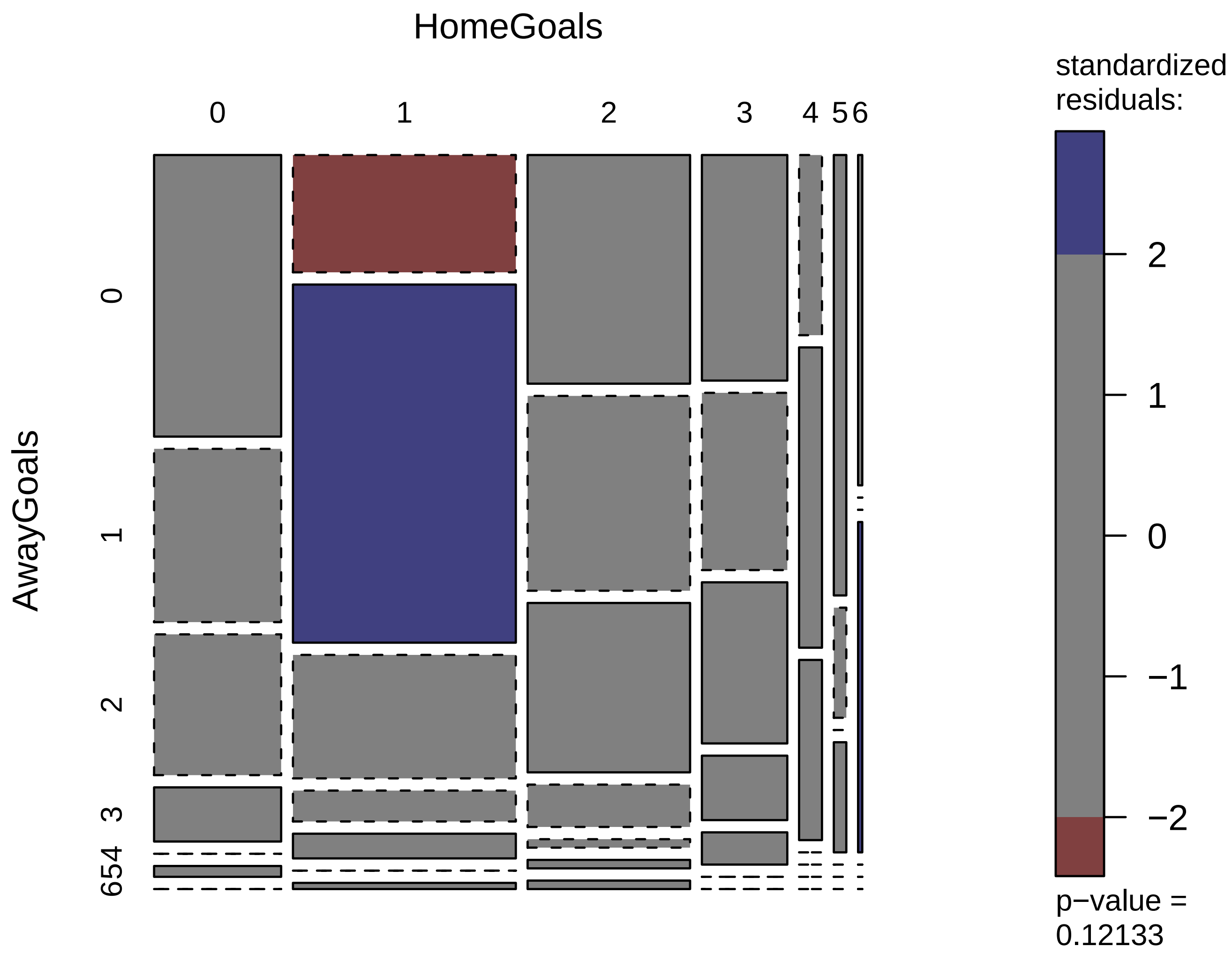
Use value to code the *result of a significance test* for independence.



Visualization & testing



Visualization & testing



HCL colors

Disadvantages of HSV colors:

- device dependent,
- not copierproof,
- flashy colors good for drawing attention to a plot, but hard to look at.

HCL colors

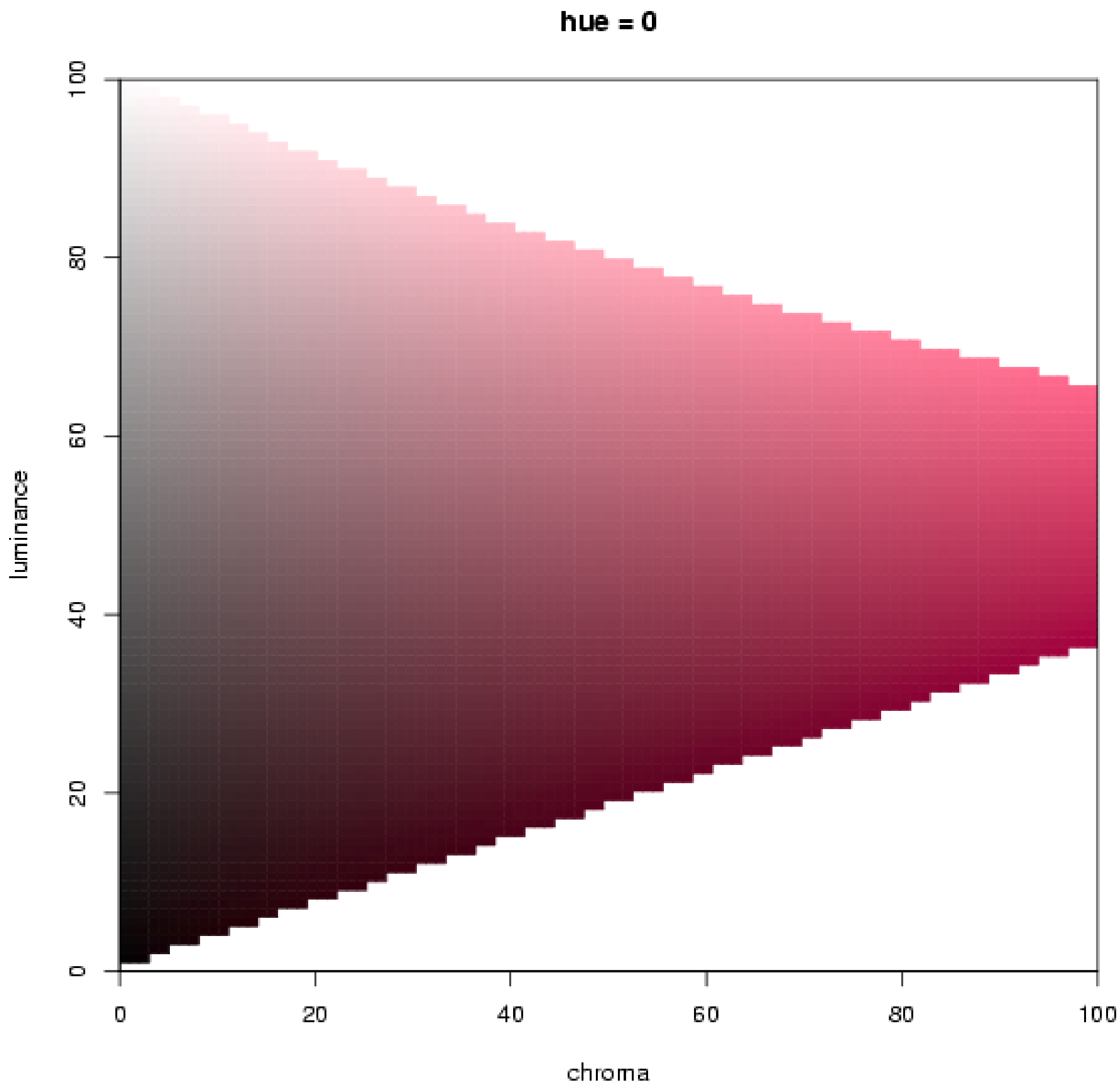
Disadvantages of HSV colors:

- device dependent,
- not copierproof,
- flashy colors good for drawing attention to a plot, but hard to look at.

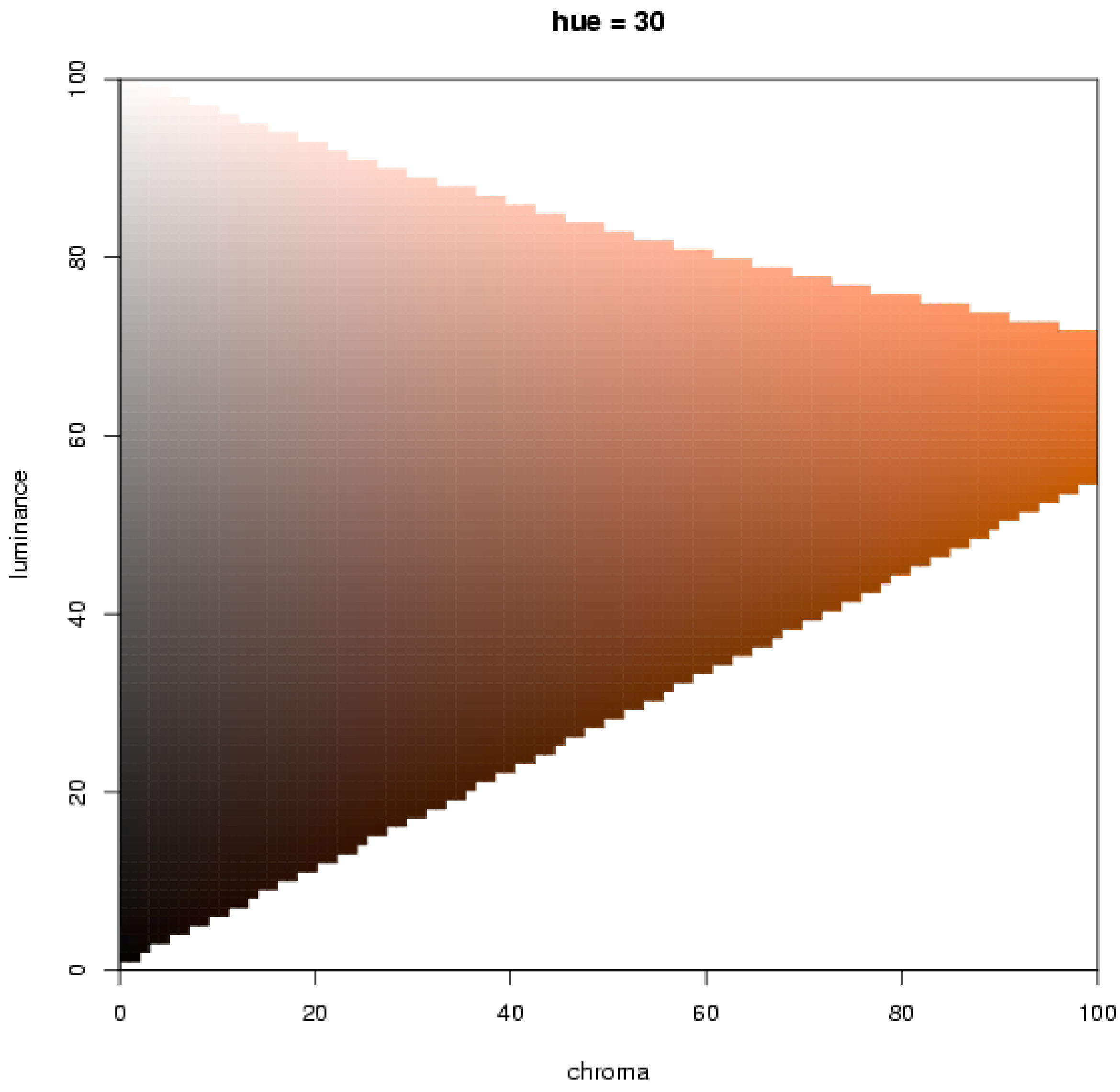
Alternative: use HCL colors instead (see Ihaka, 2003).

HCL colors are defined by hue (in $[0, 360]$), chroma and luminance (in $[0, 100]$). HCL space essentially looks like a double cone.

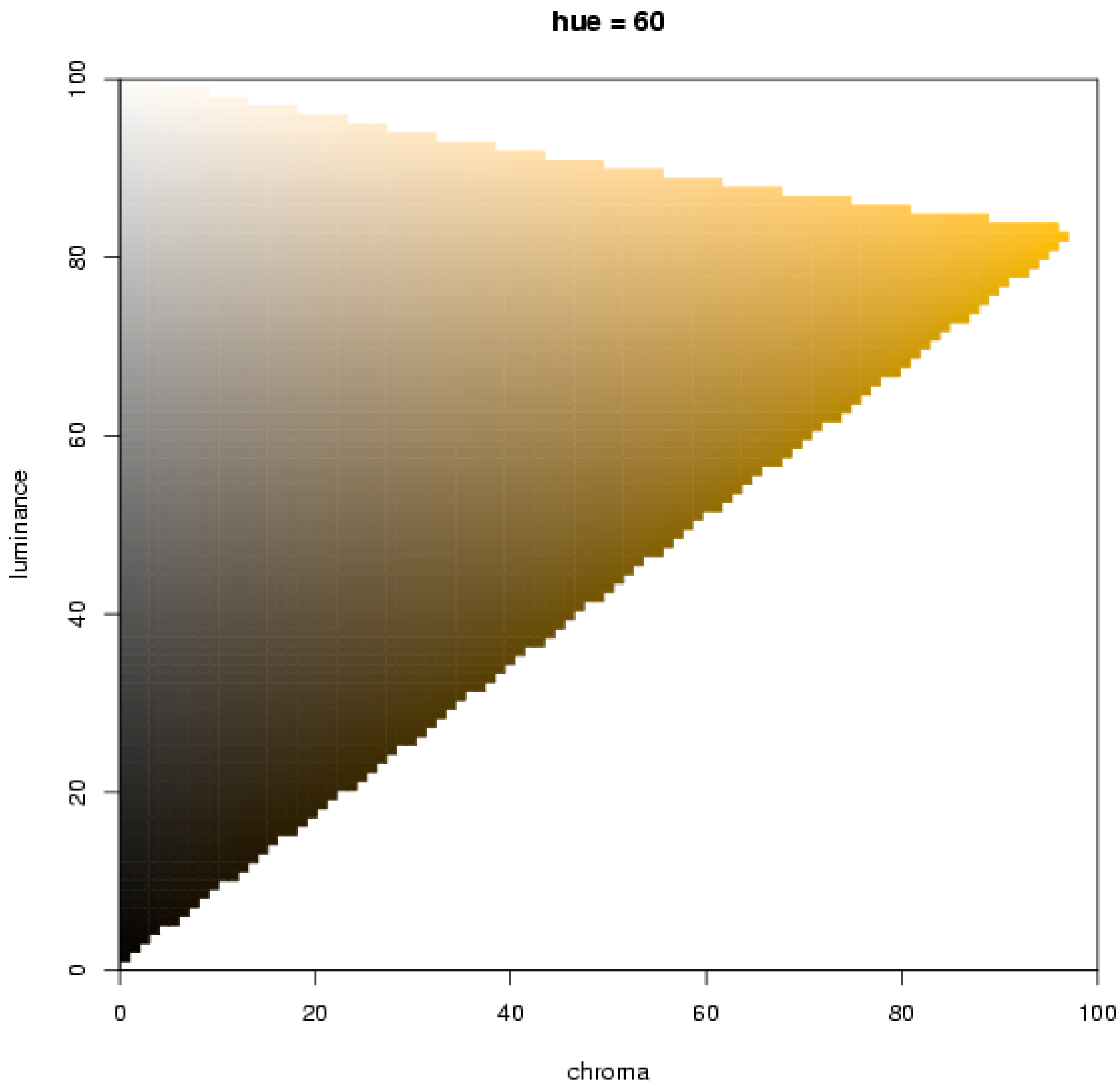
HCL colors



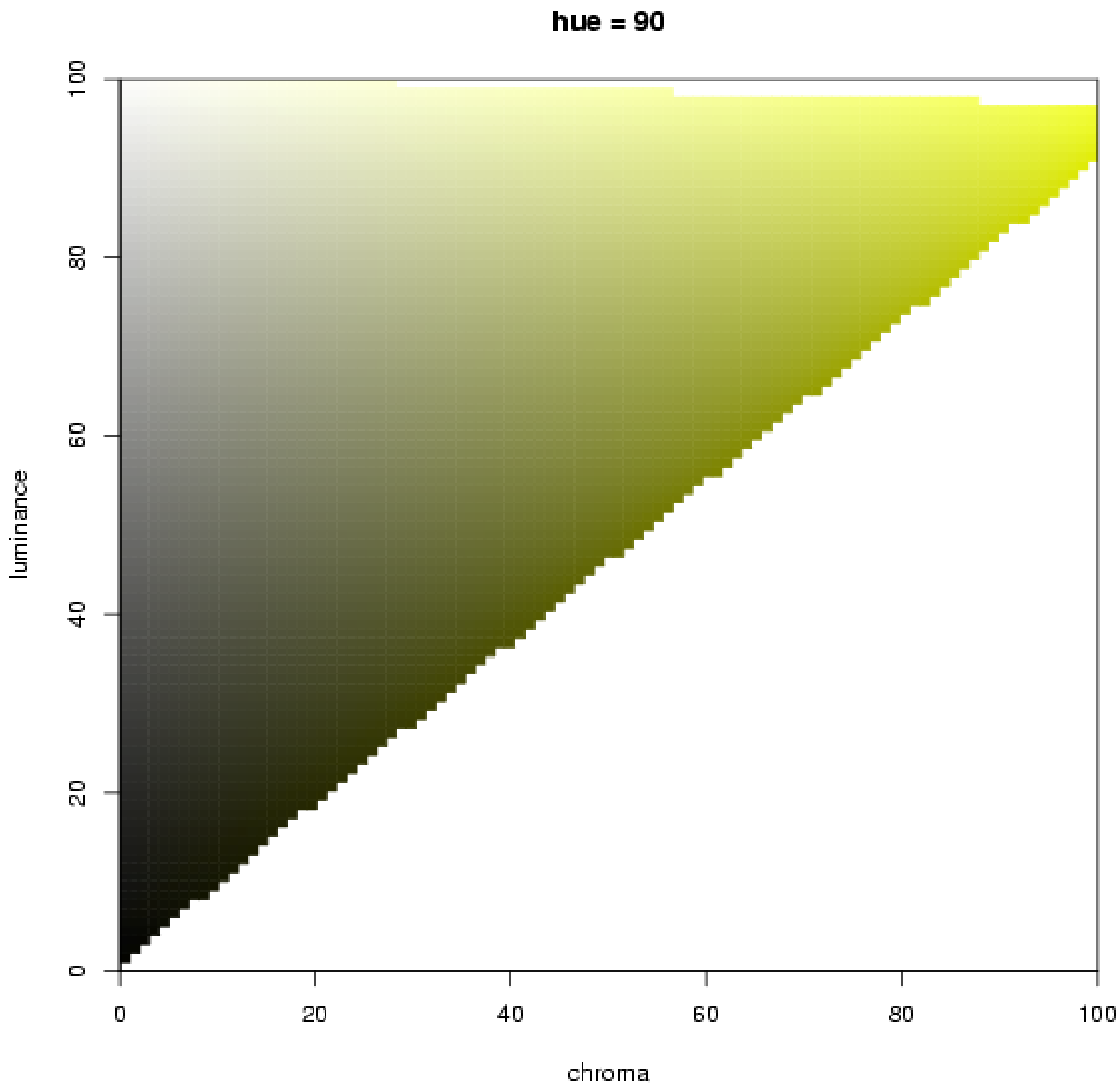
HCL colors



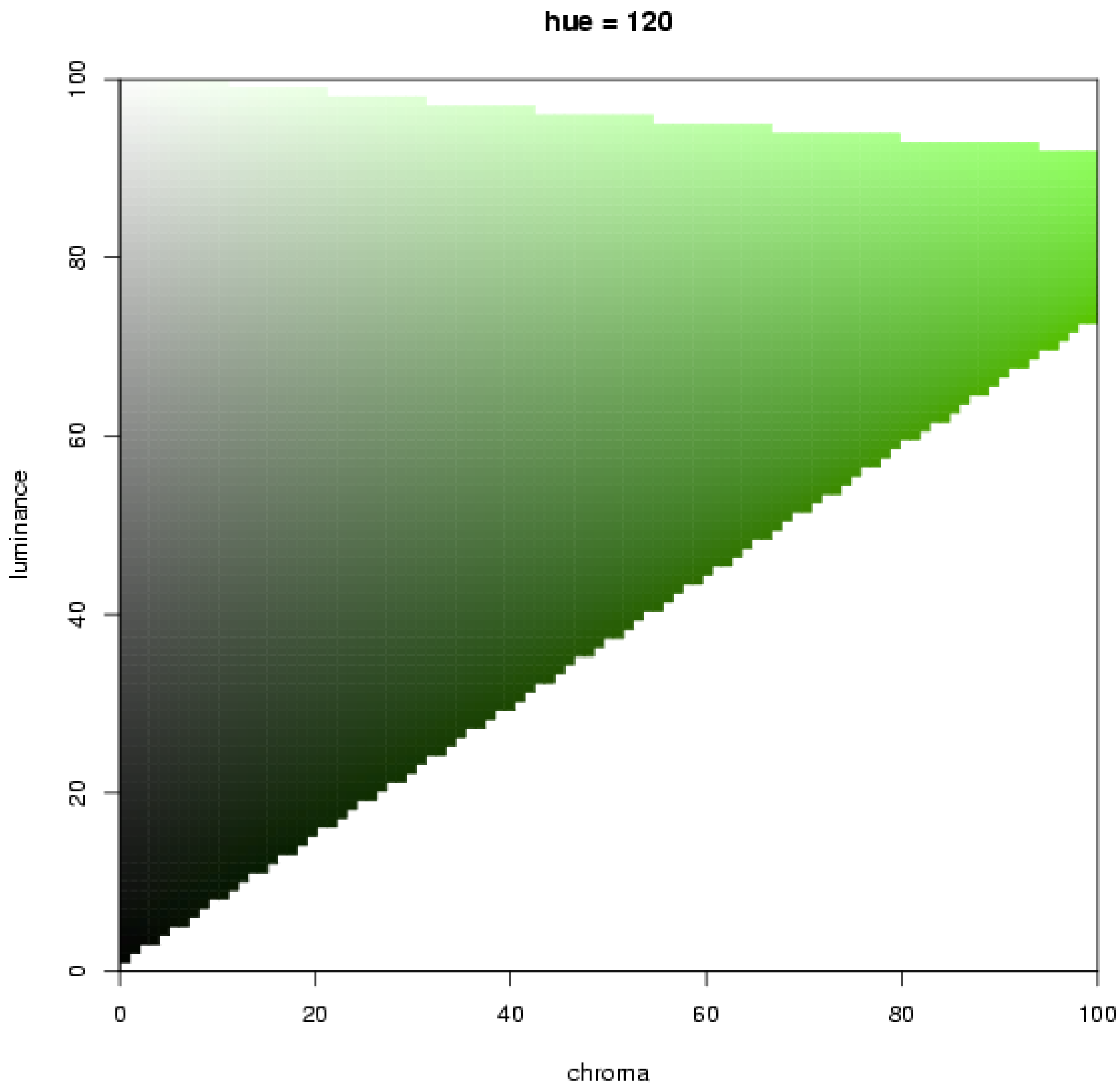
HCL colors



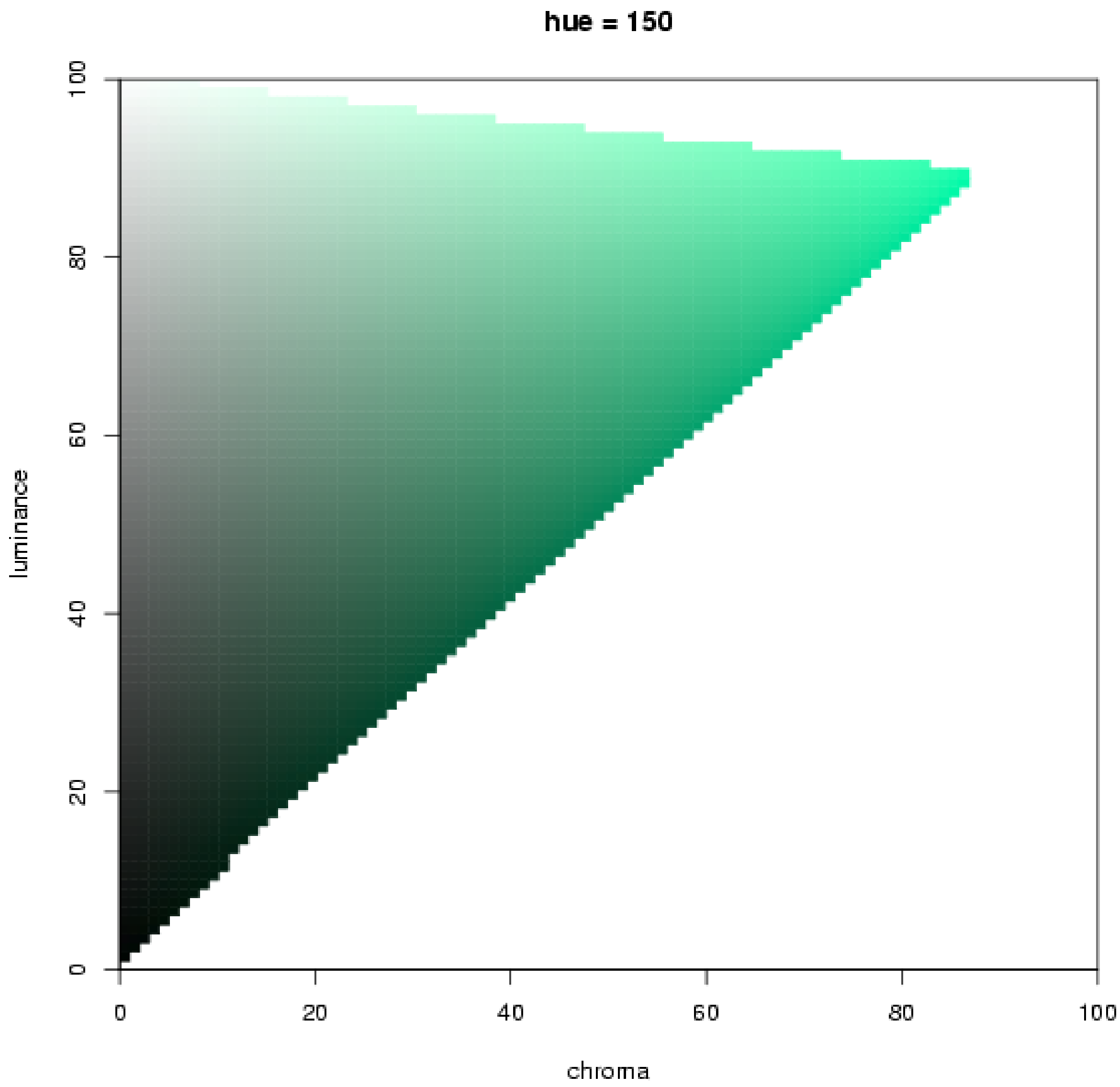
HCL colors



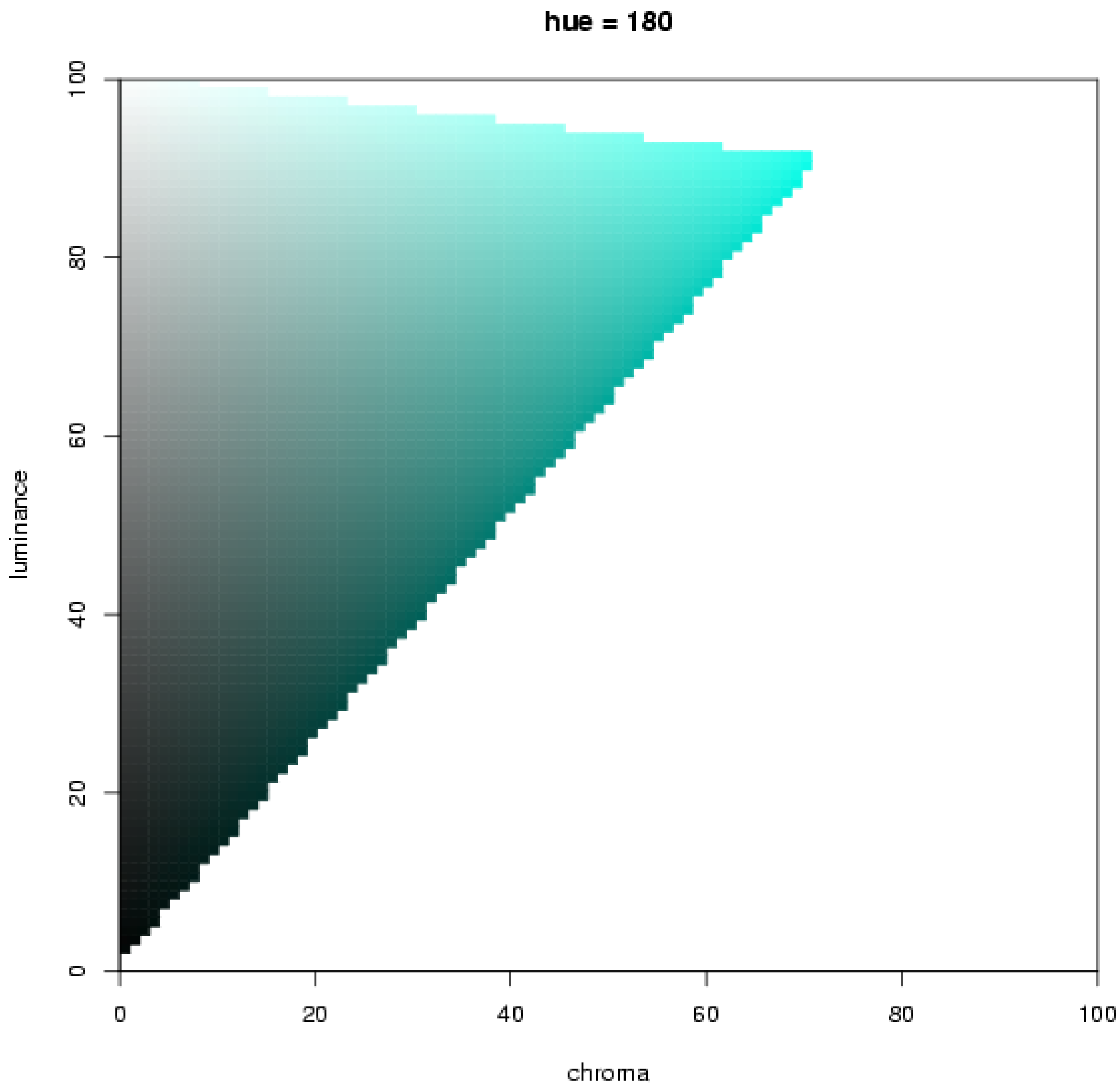
HCL colors



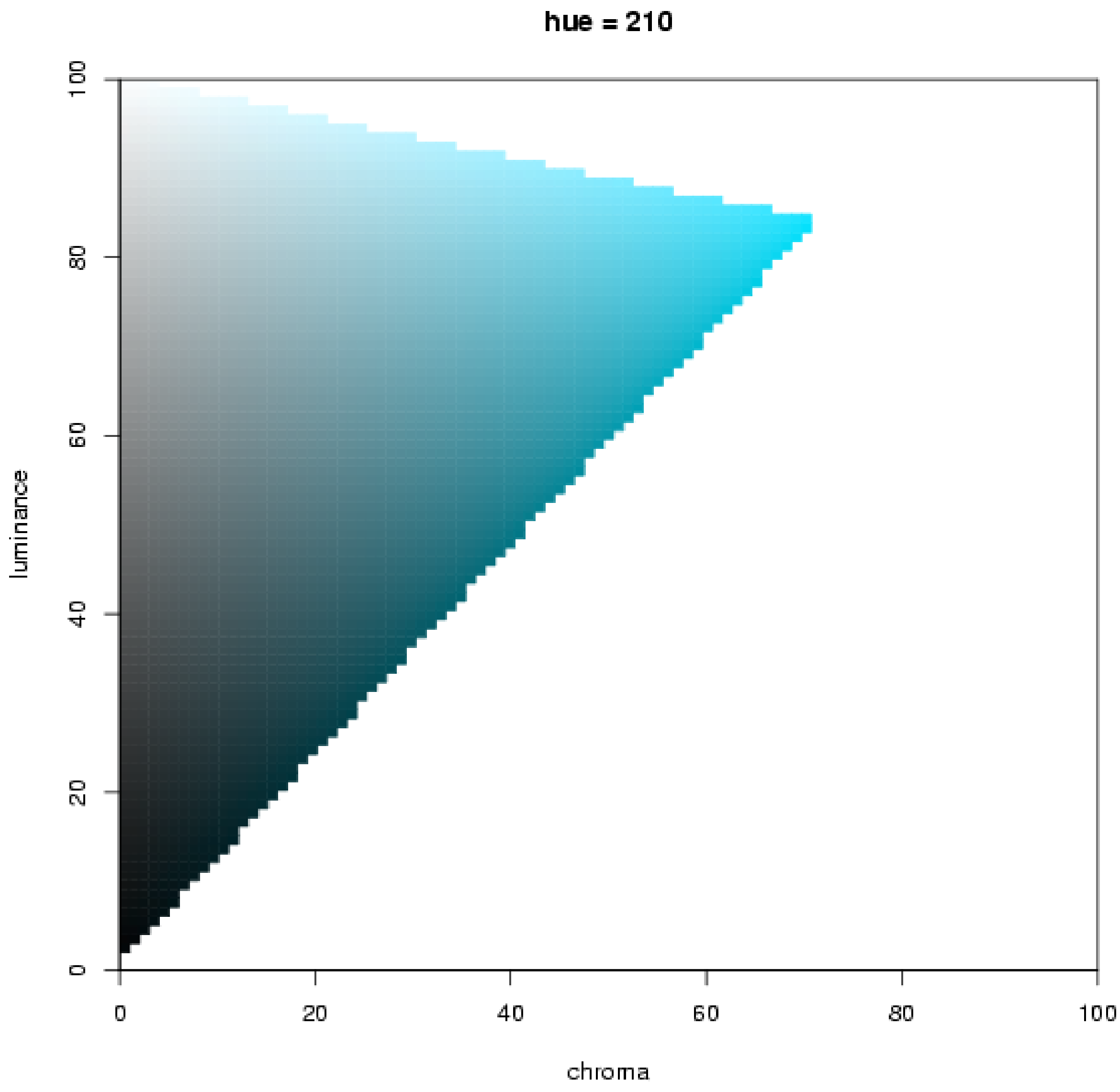
HCL colors



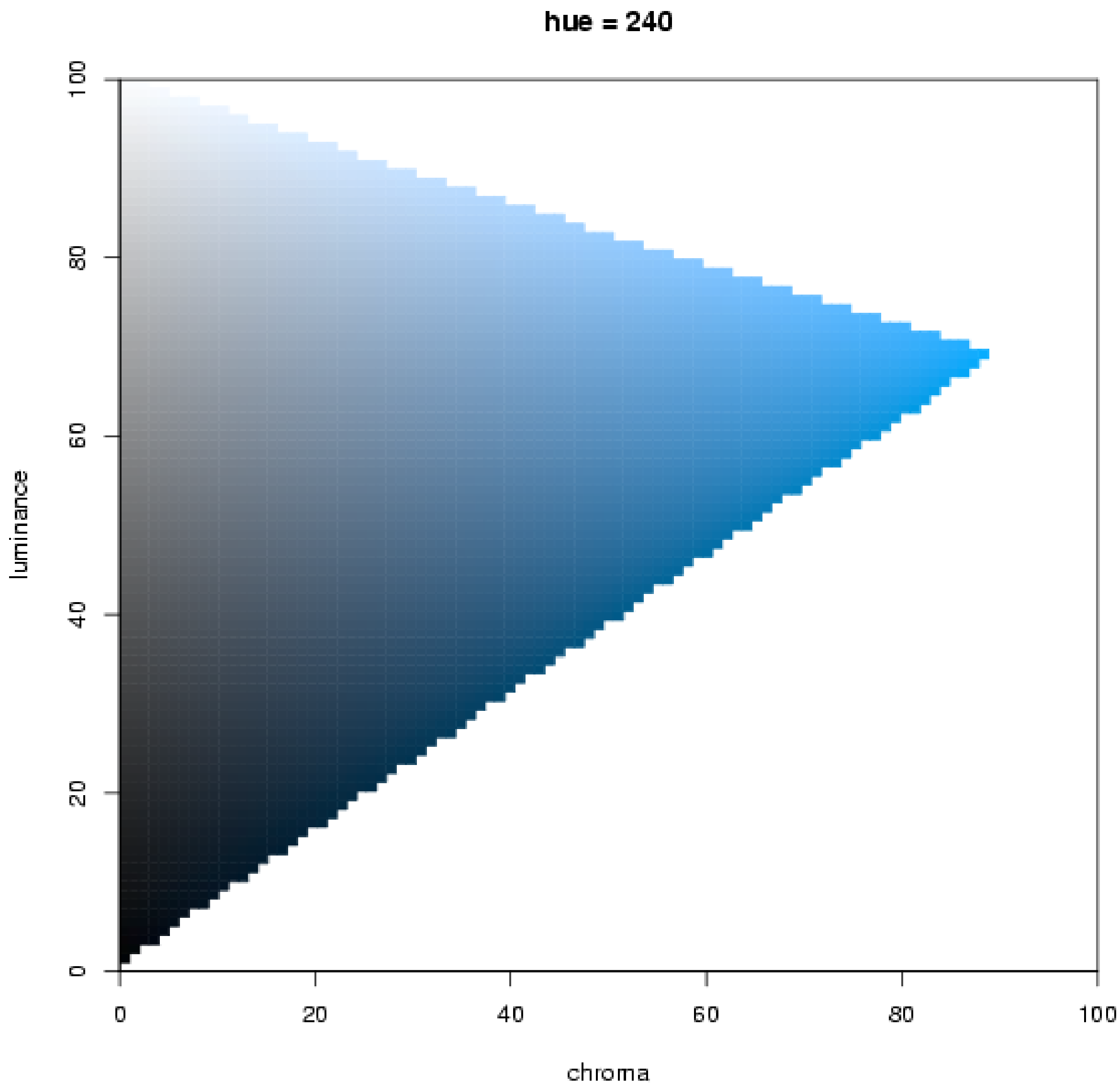
HCL colors



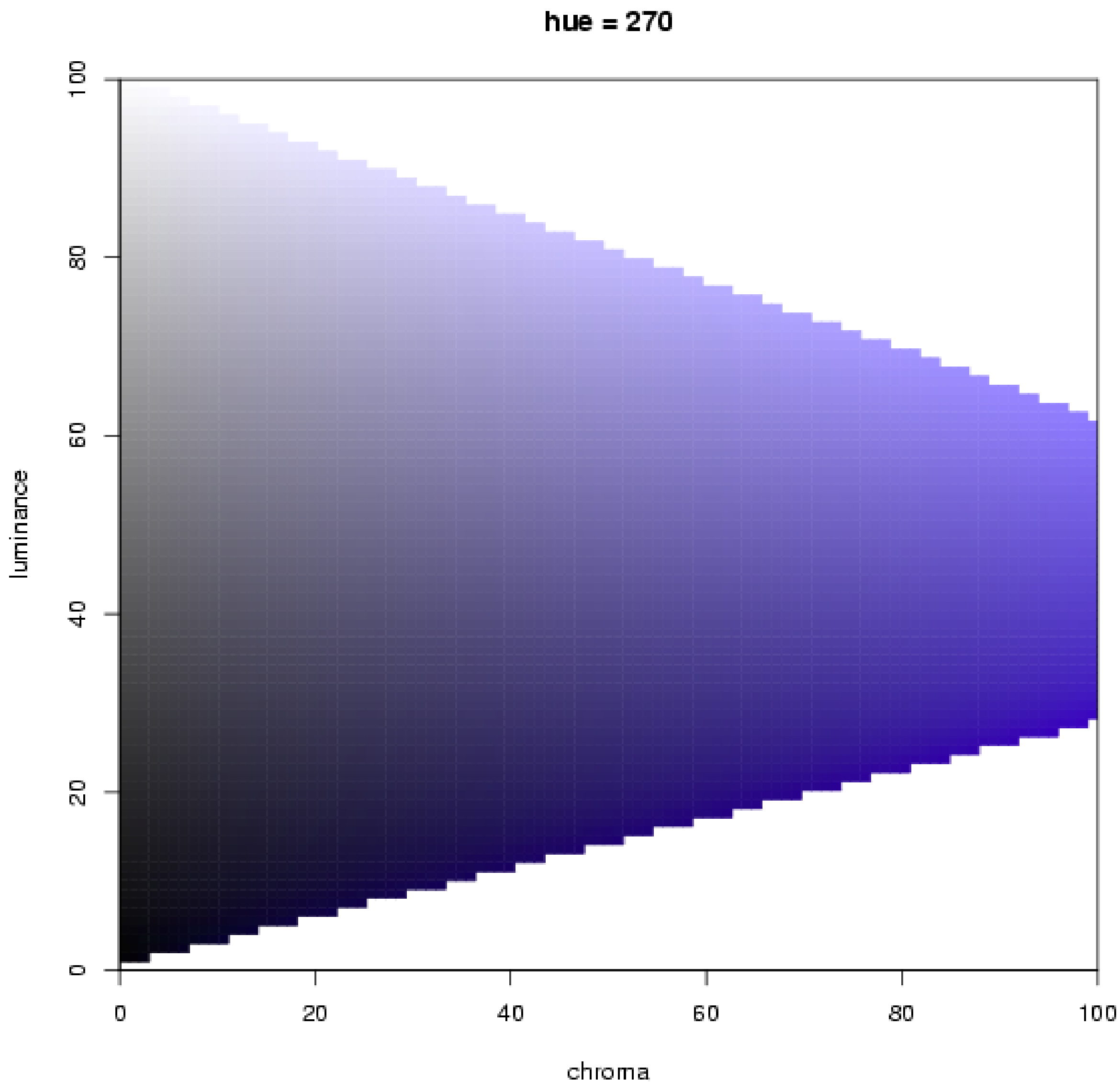
HCL colors



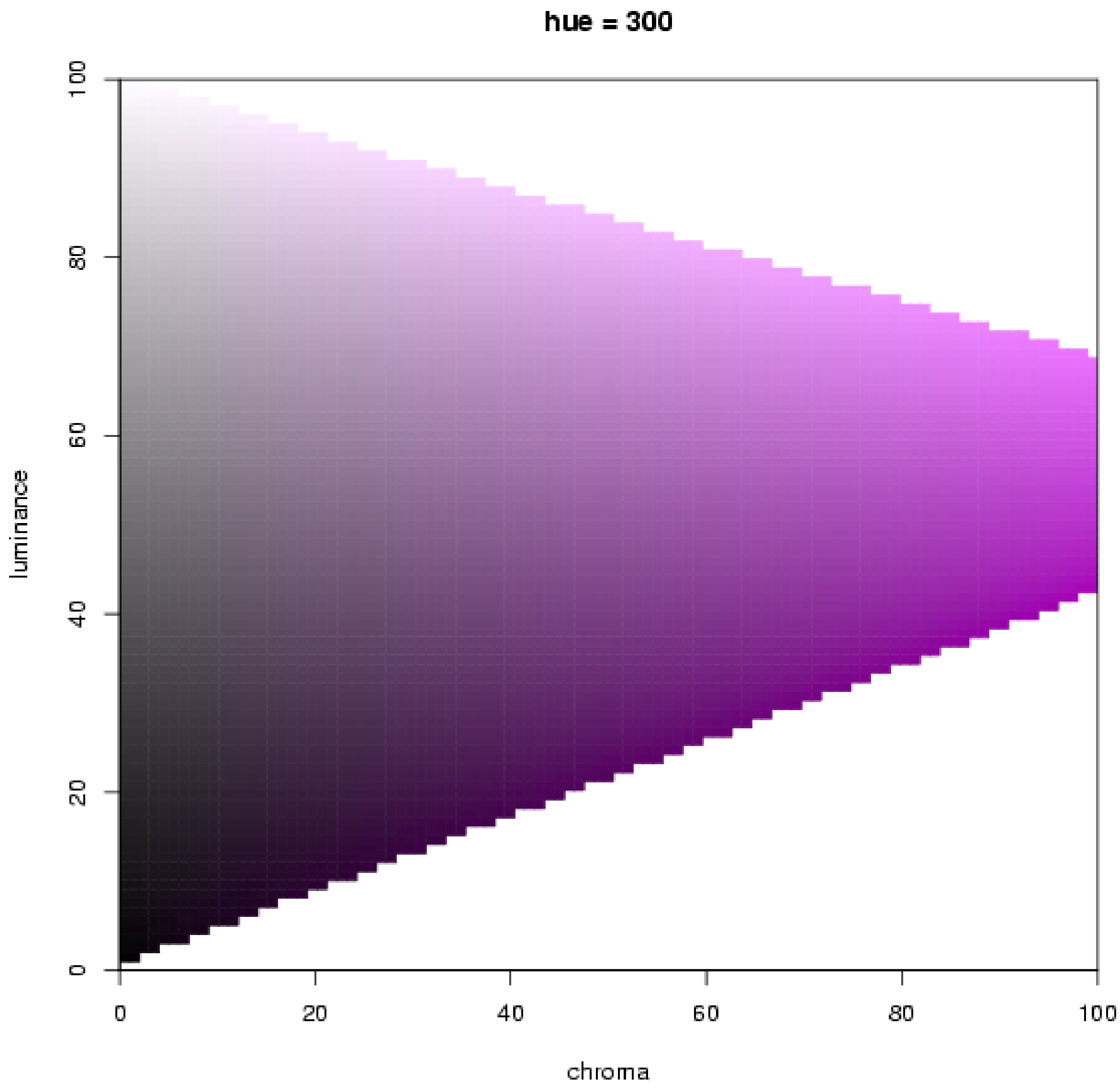
HCL colors



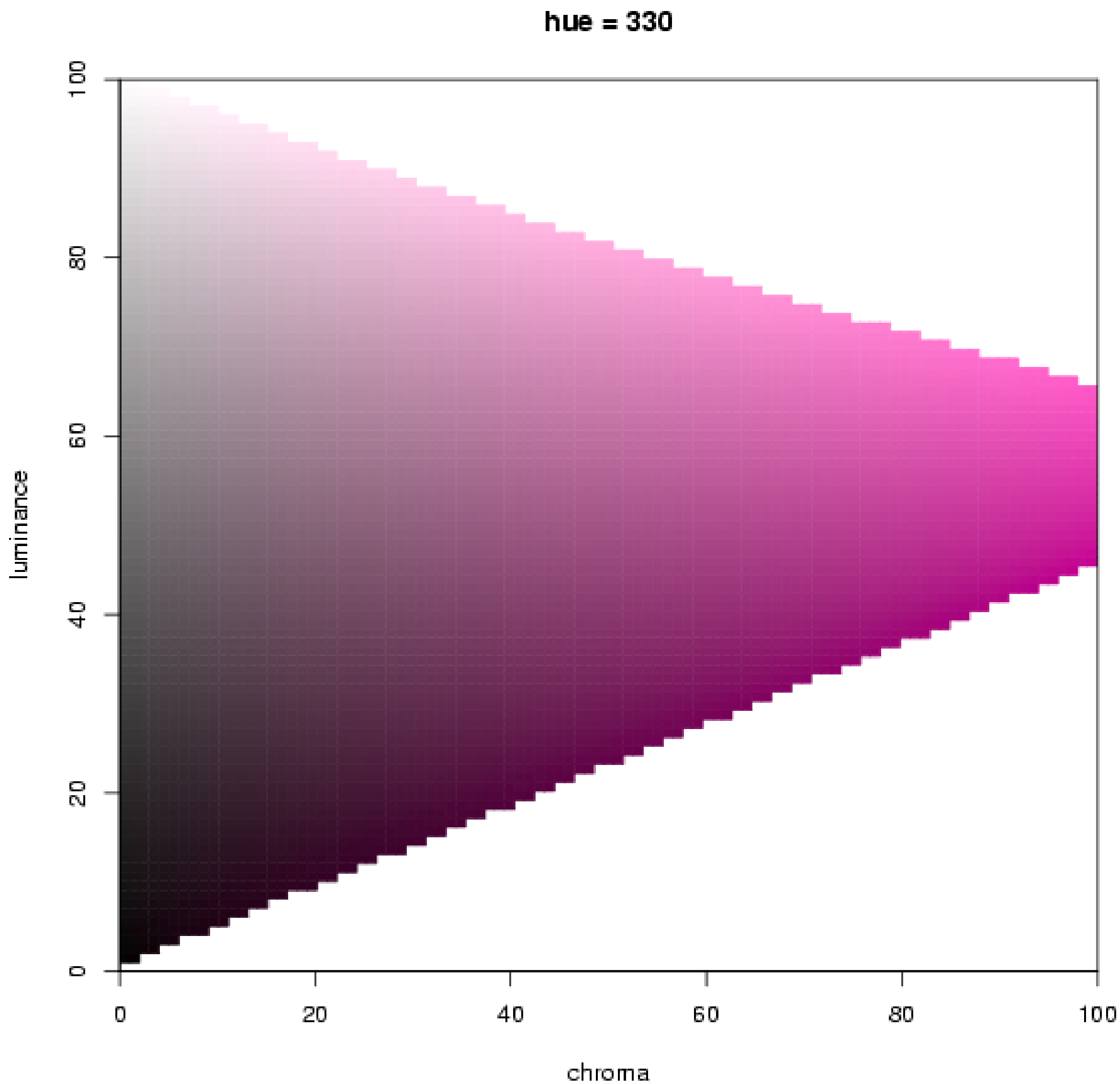
HCL colors



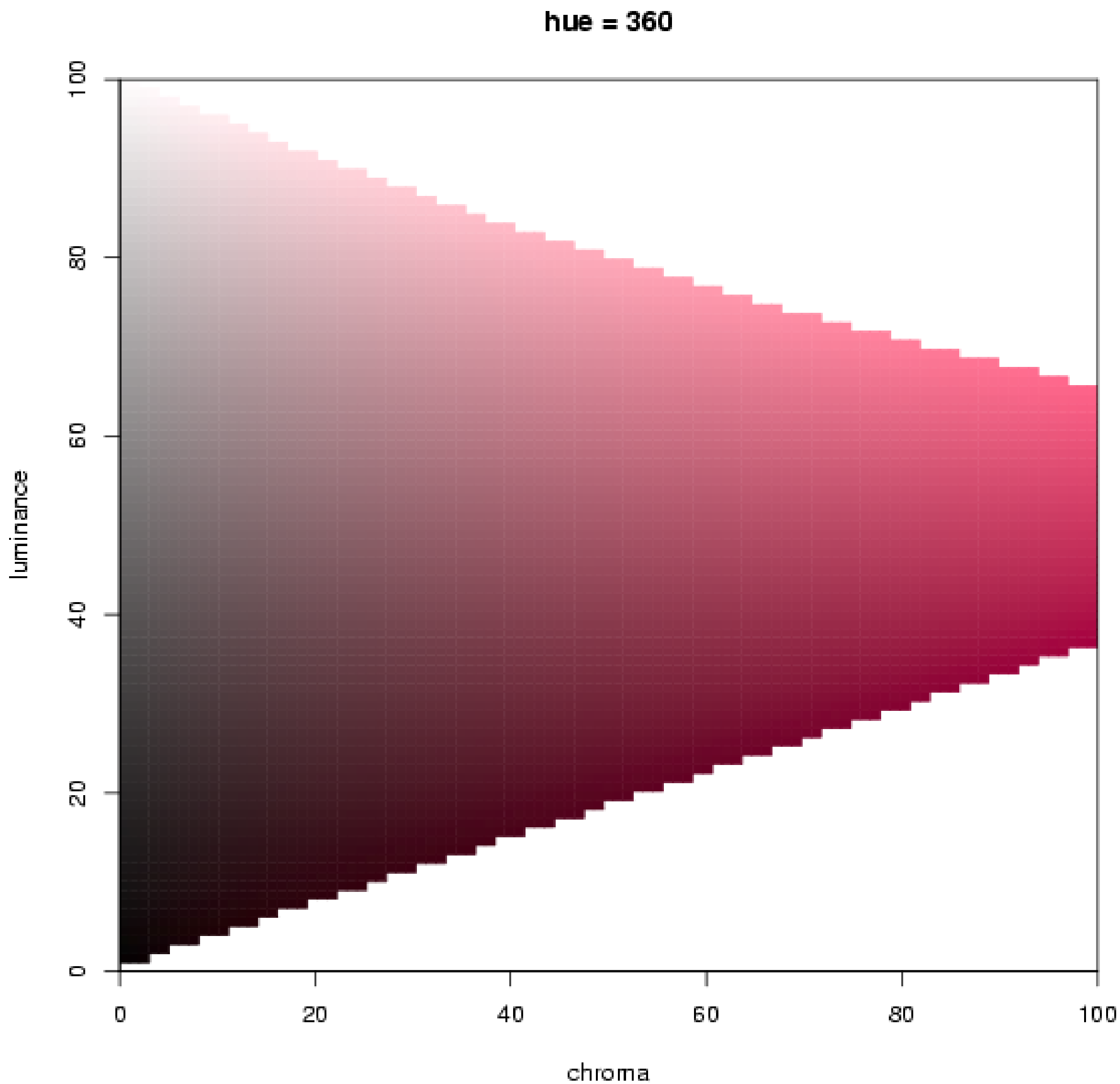
HCL colors



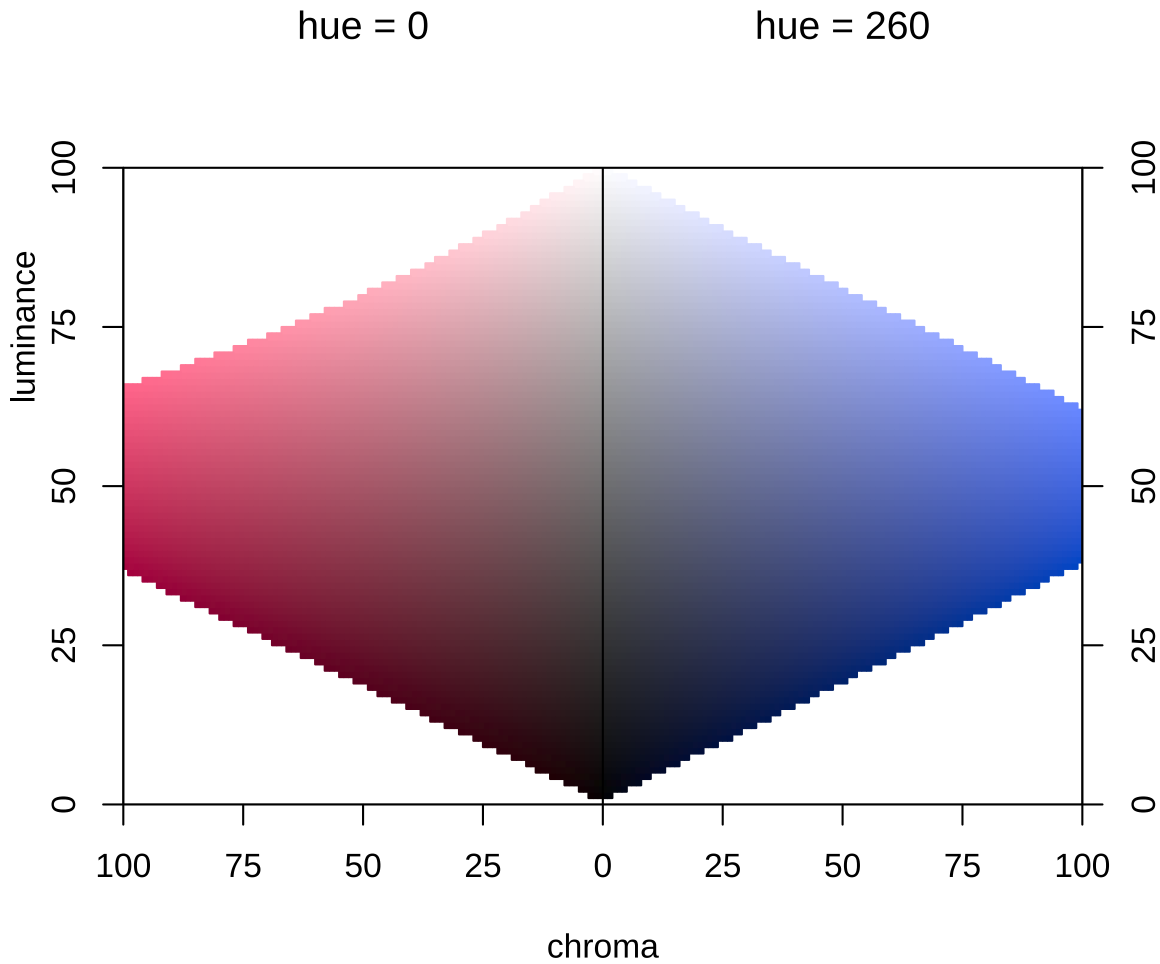
HCL colors



HCL colors

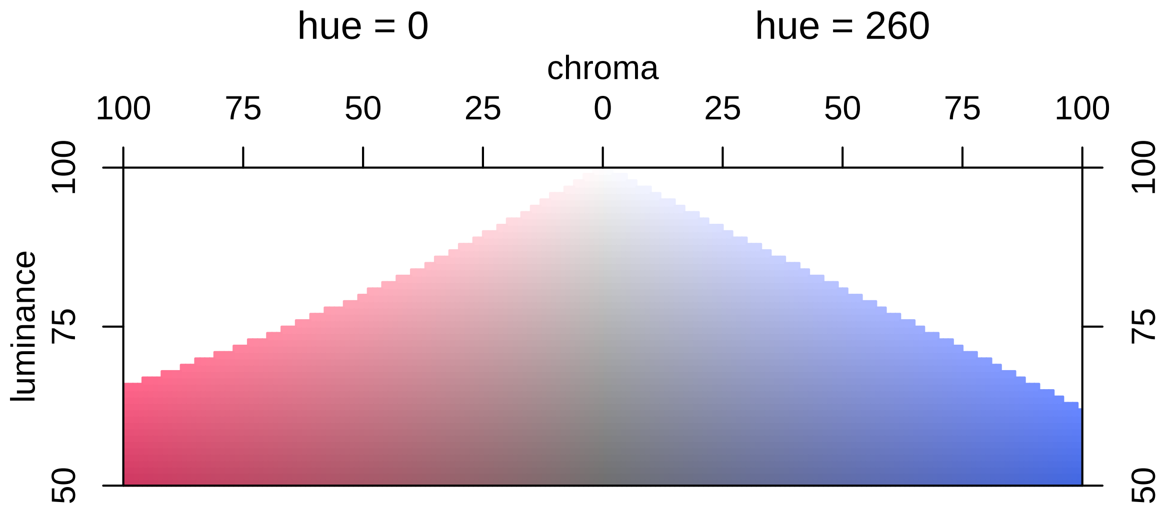


HCL colors



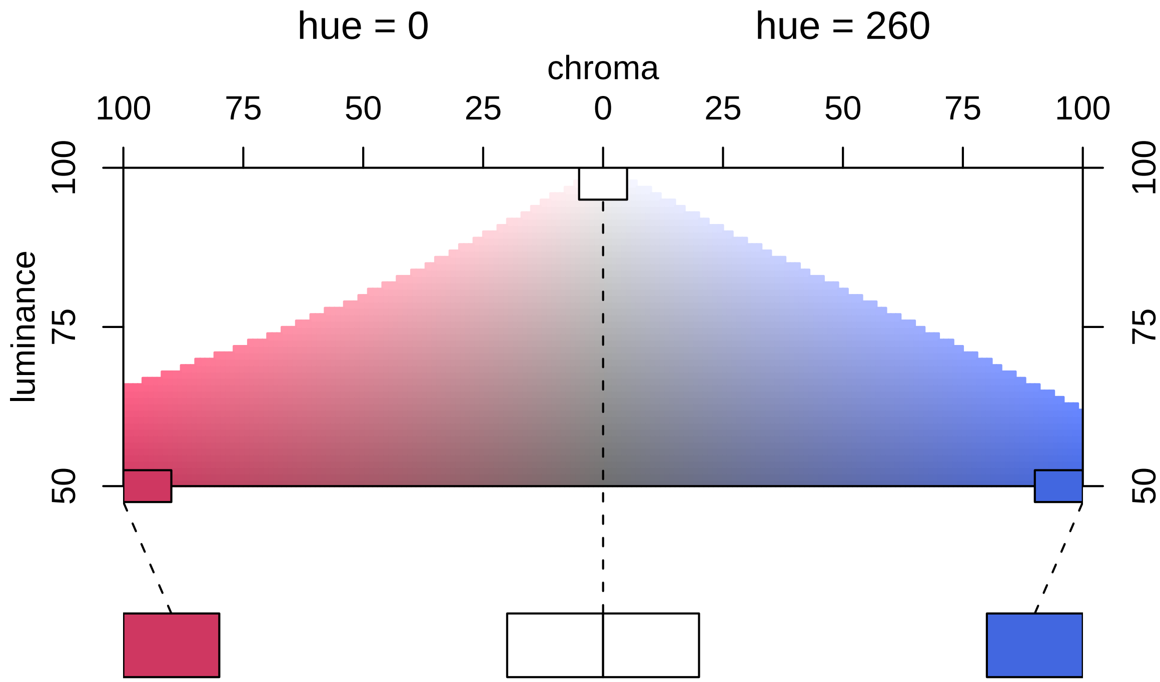
hue = 0
hue = 260
luminance
100
75
50
25
0
100
75
50
25
0
100
75
50
25
0
25
50
75
100
chroma

HCL colors



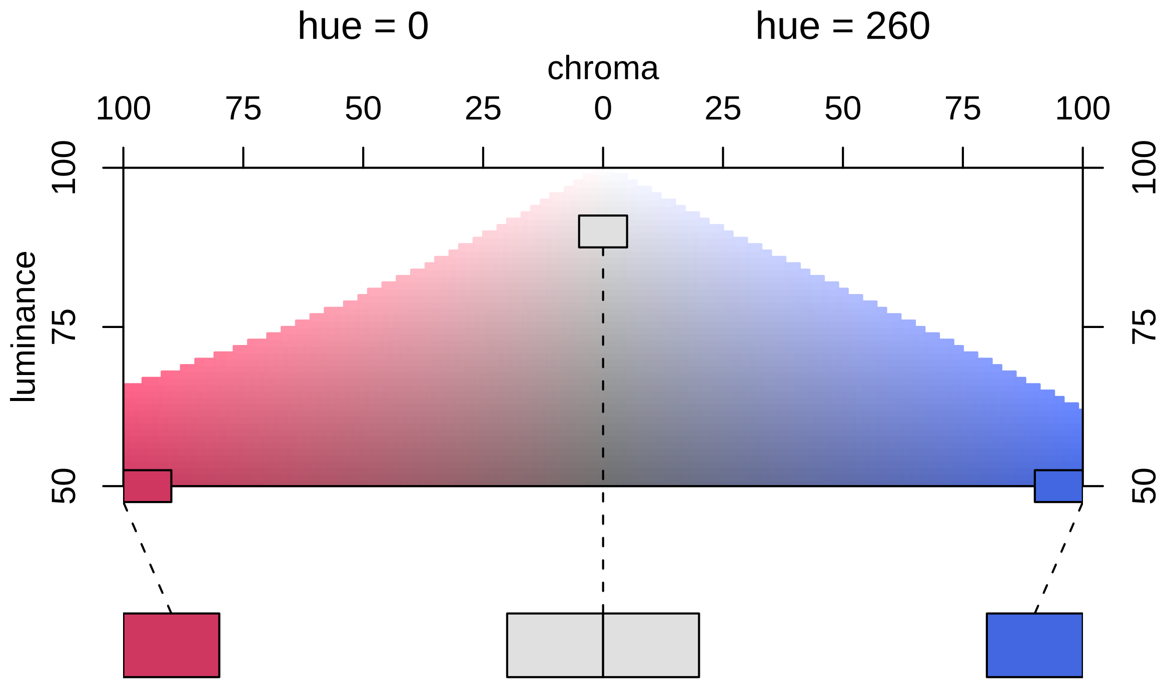
hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
100
75
50

HCL colors

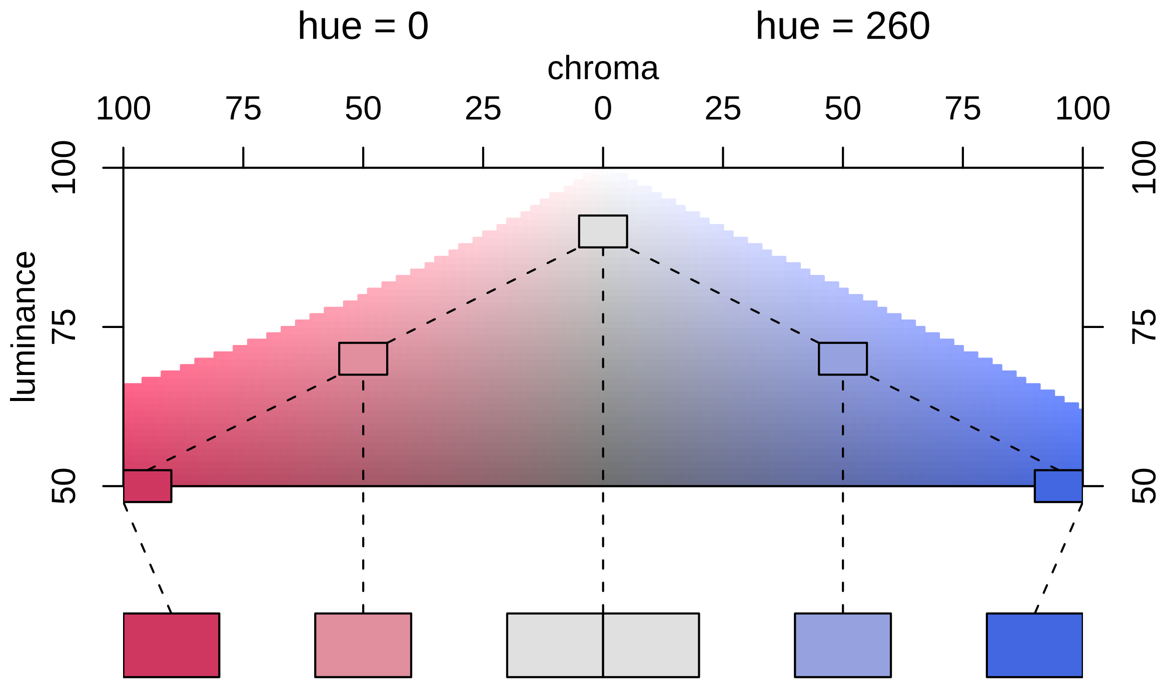


hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
100
75
50

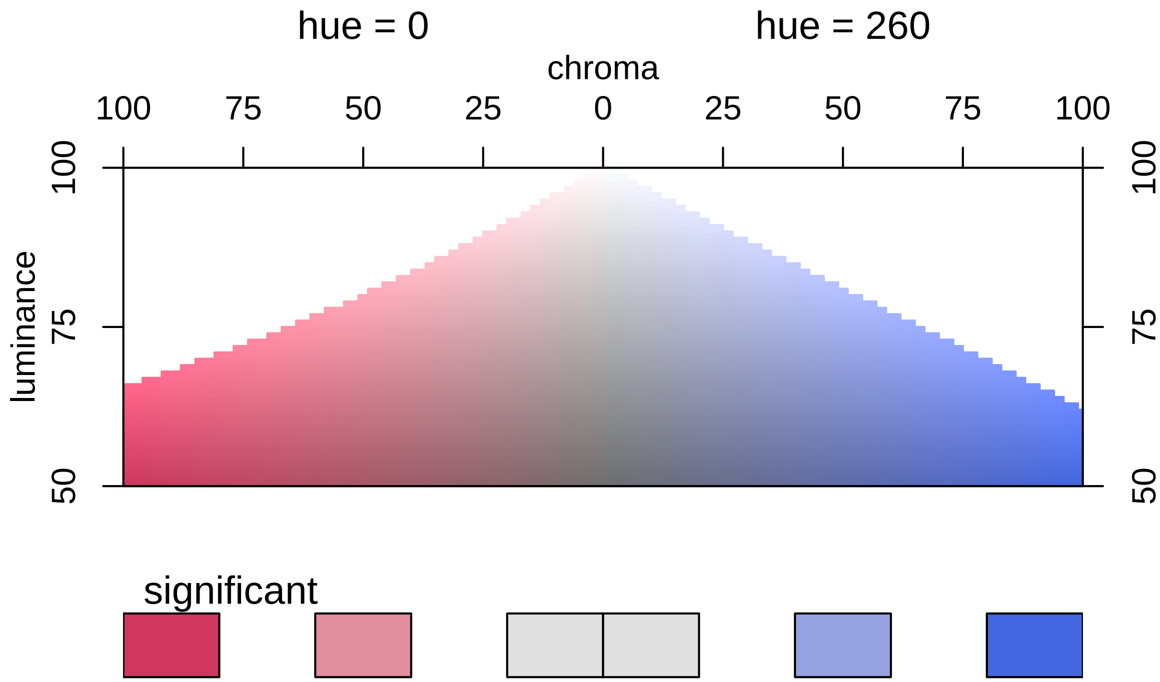
HCL colors


hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
100
75
50

HCL colors

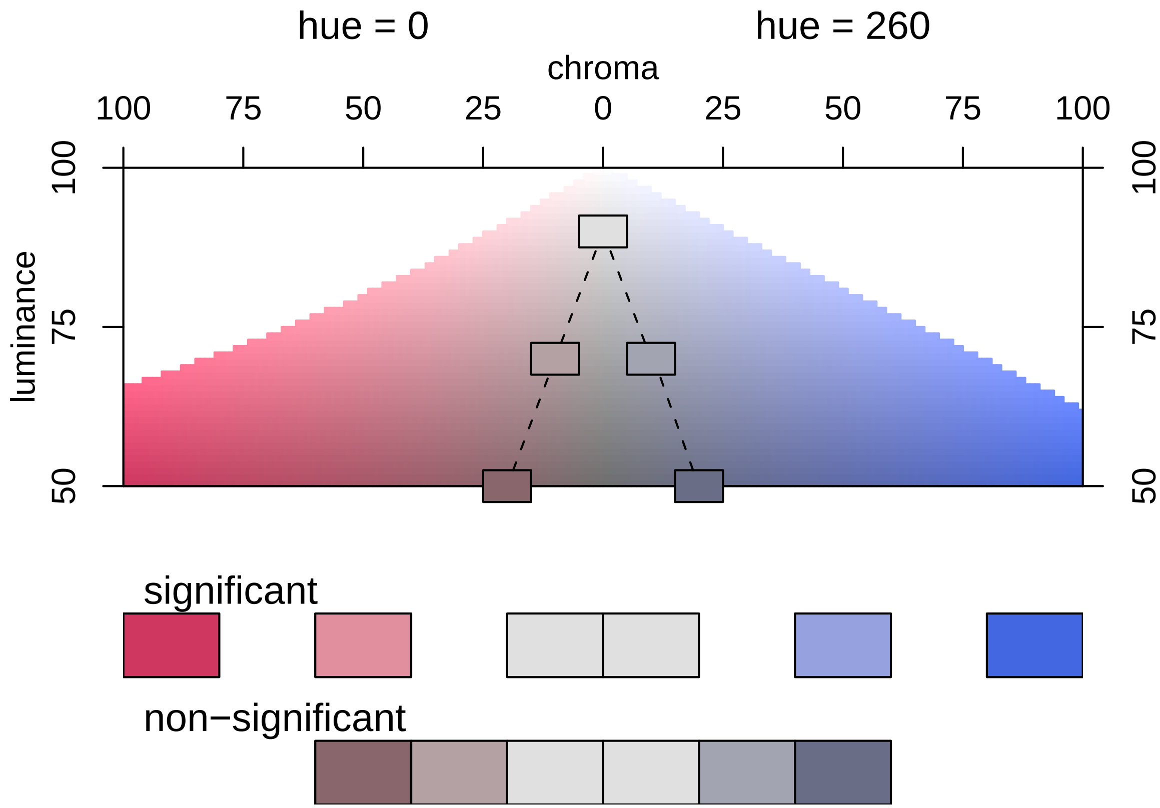

hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
100
75
50

HCL colors



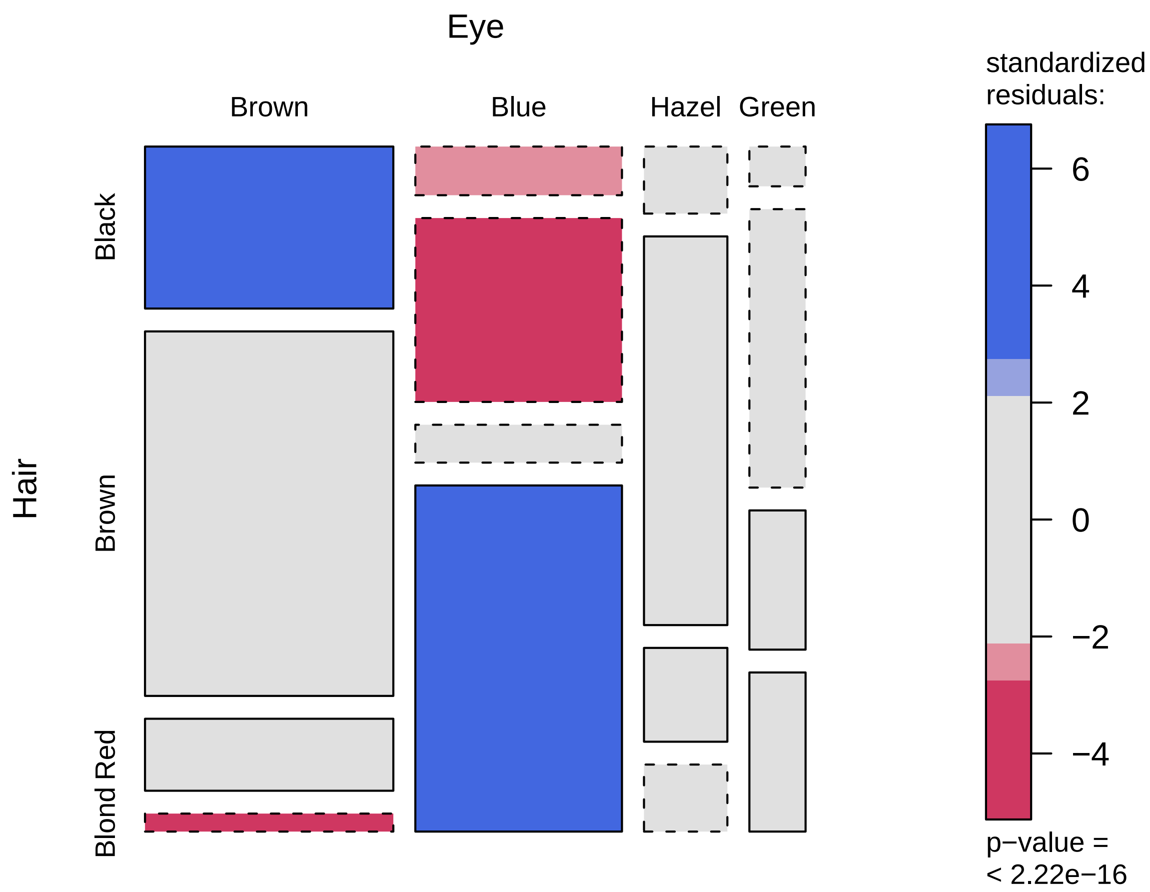
hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
significant

HCL colors



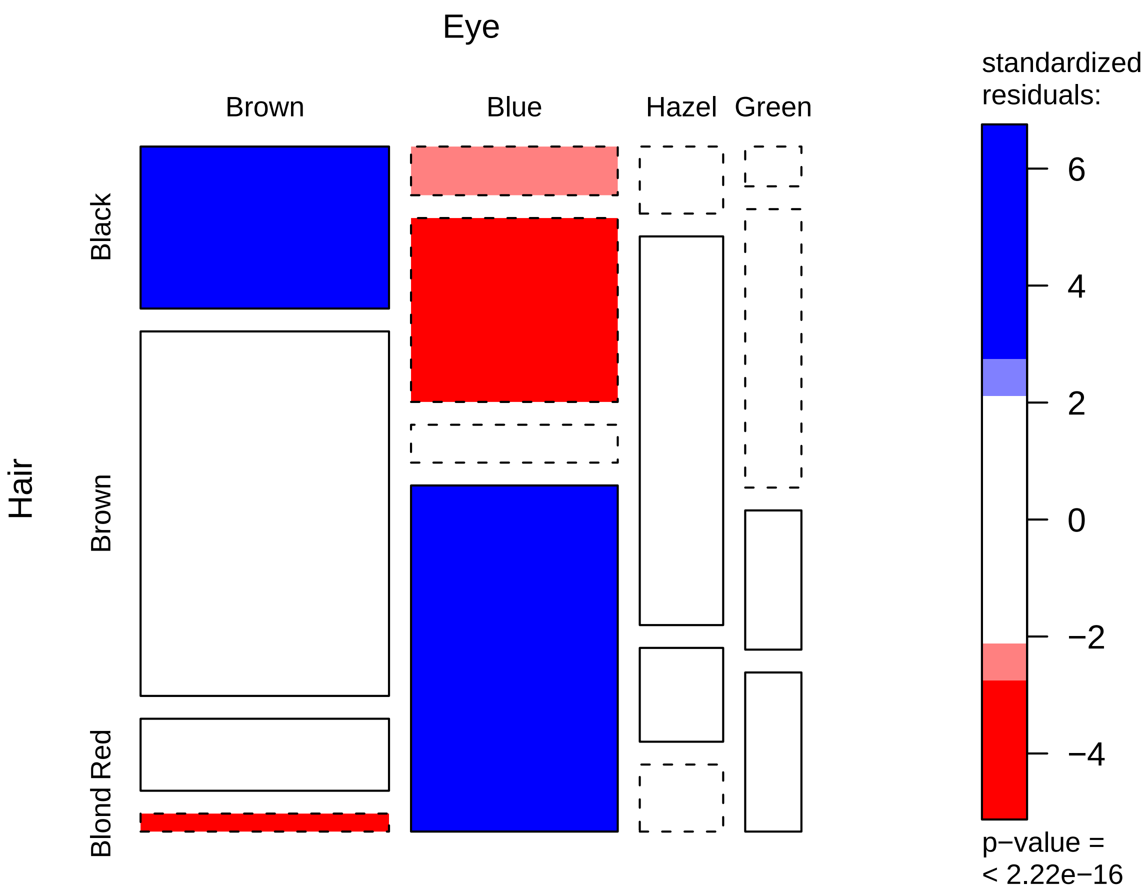
hue = 0
hue = 260
chroma
100
75
50
25
0
25
50
75
100
luminance
100
75
50
significant
non−significant

HCL colors



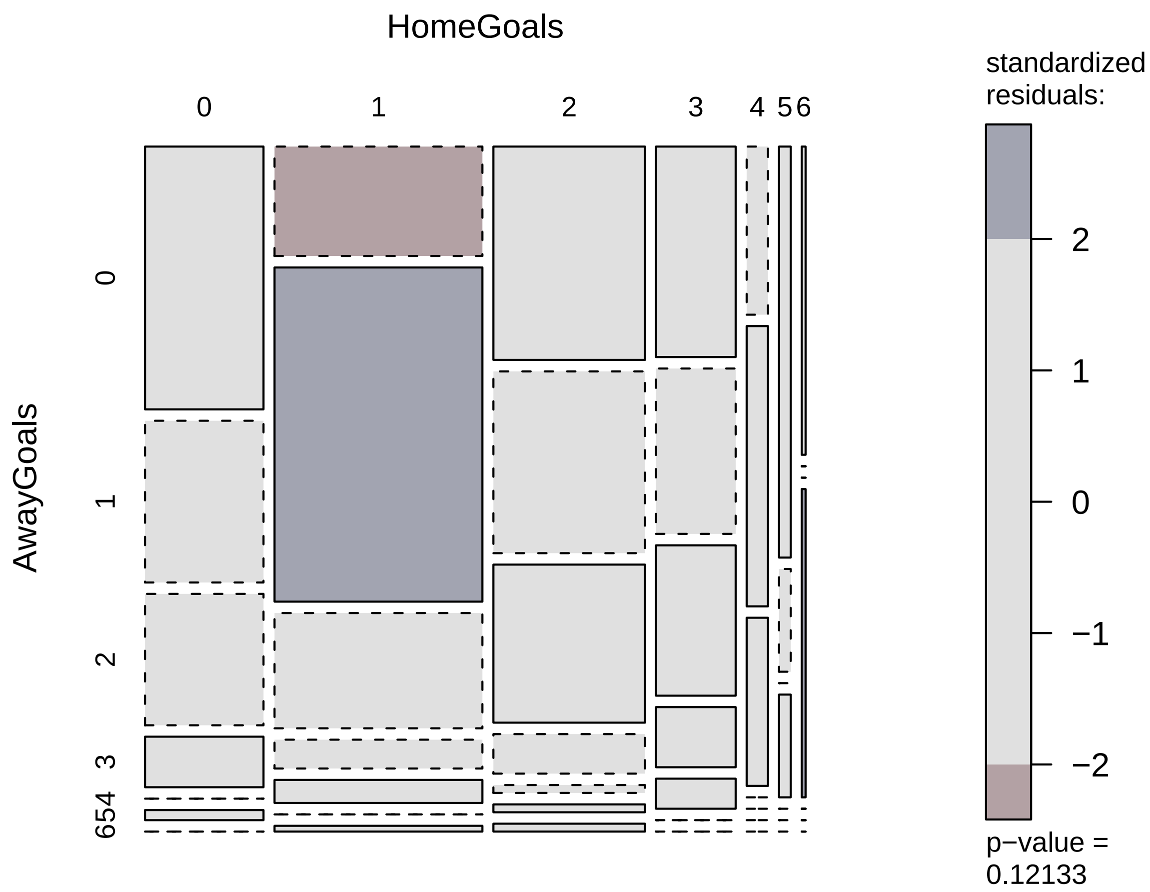
Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
standardized
residuals:
6
4
2
0
−2
−4
p−value =
< 2.22e−16

HCL colors



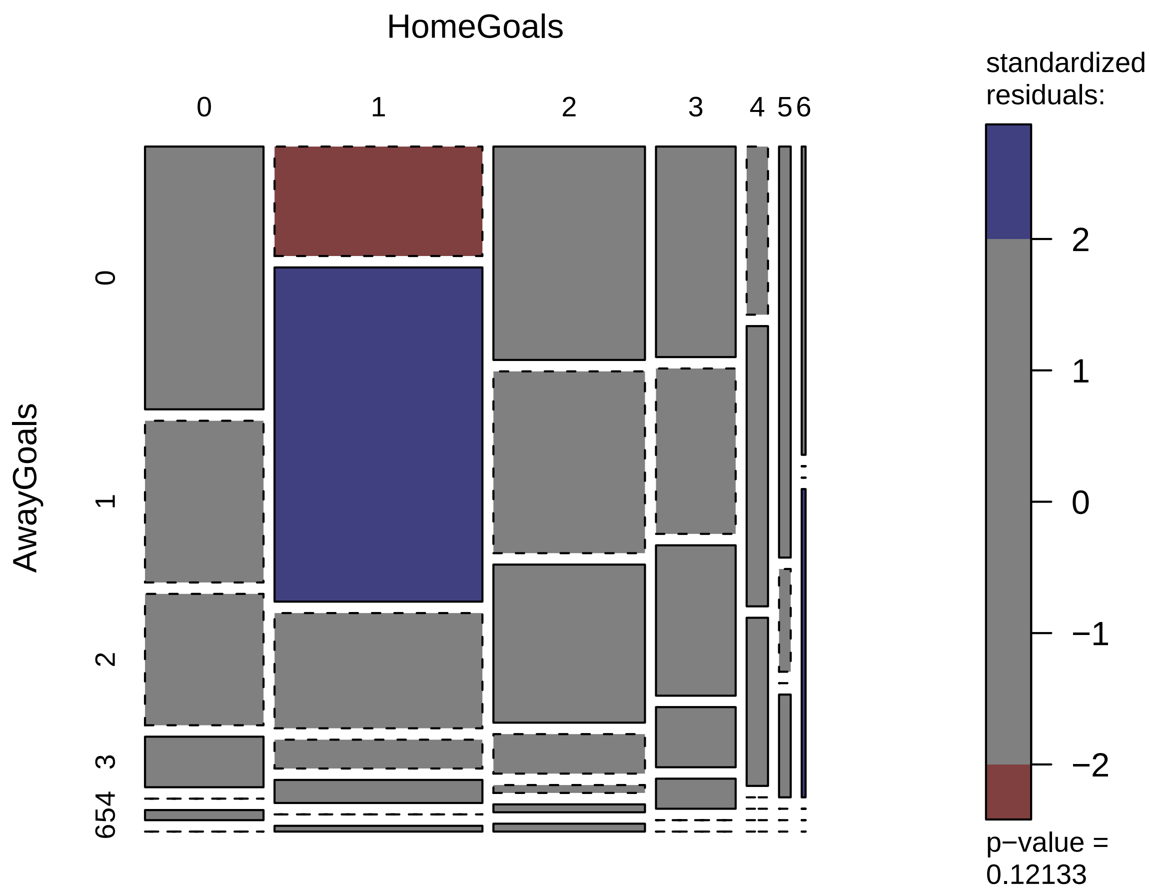
Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
standardized
residuals:
6
4
2
0
−2
−4
p−value =
< 2.22e−16

HCL colors



HomeGoals
0
1
2
3
4
5
6
AwayGoals
0
1
2
3
4
5
6
standardized
residuals:
2
1
0
−1
−2
p−value =
0.12133

HCL colors


HomeGoals
0
1
2
3
4
5
6
AwayGoals
0
1
2
3
4
5
6
standardized residuals:
2
1
0
−1
−2
p−value = 0.12133

Implementation in `grid`

The graphics engine `grid` overcomes the old R concept of plots with a plot region surrounded by a margin. `grid` is

- based on generic drawing regions (viewports),
- allows for plotting to relative coordinates,
- is also the basis for an implementation of Trellis graphics called `lattice`.

(see Murrell, 2002)

Thus, the new implementation of mosaic and association plots makes them easily reusable, e.g., in Trellis-like layouts.

Implementation in `grid`

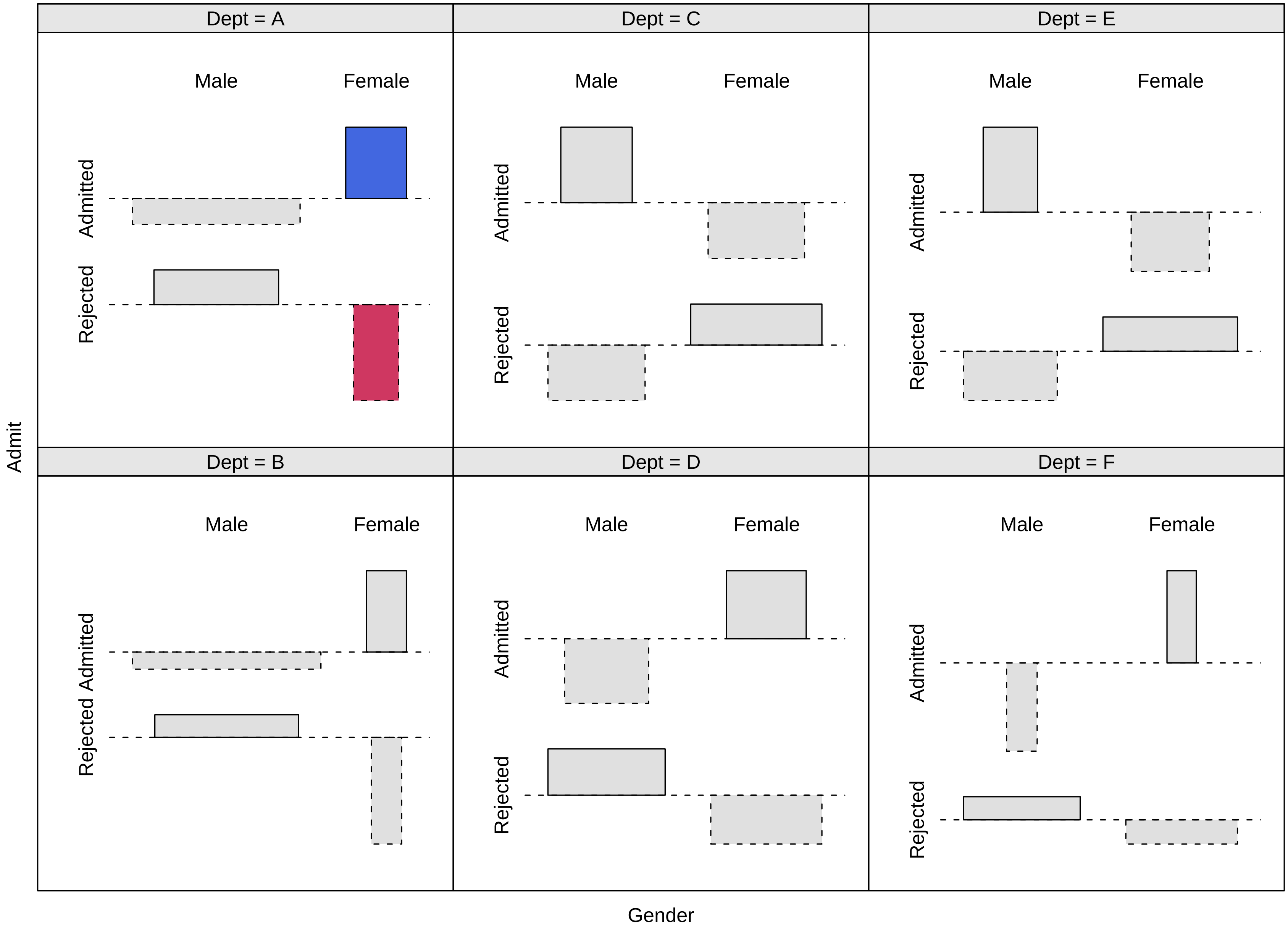
Furthermore, graphics parameters for the rectangles, e.g.,

- fill color,
- line type,
- line color,

can be specified for each cell individually by the user. Each graphics parameter can be an object of the same dimenionality as the original table.

$\longrightarrow$ new shadings can easily be implemented.

Multi-way tables



The `vcd` package

New methods will be available in the package `vcd` for visualizing categorical data.

Currently only in development version. The released version is available from the Comprehensive R Archive Network

http://CRAN.R-project.org/

and it already offers some functionality for

- fitting & graphing of discrete distributions,
- plots for independence and agreement,
- visualization of log-linear models.