CA11

2025-05-15

```r
# 2-way Contingency table:
library(ca)
ratings <- structure(
c(
50, 30, 10, 1, 60, 80, 40, 2,
40, 60, 20, 1, 10, 30, 50, 4),
dim = c(4L, 4L),
dimnames = list(
c("High School", "Bachelor's", "Master's", "Doctorate"),
c("Action", "Drama", "Comedy", "Documentary"))
)
```

```r
# Question 1a:
# Row profile:
row_profile <- prop.table(ratings, margin = 1)
round(row_profile, 4)
```

```
##             Action  Drama Comedy Documentary
## High School 0.3125 0.3750 0.2500      0.0625
## Bachelor's  0.1500 0.4000 0.3000      0.1500
## Master's    0.0833 0.3333 0.1667      0.4167
## Doctorate   0.1250 0.2500 0.1250      0.5000
```

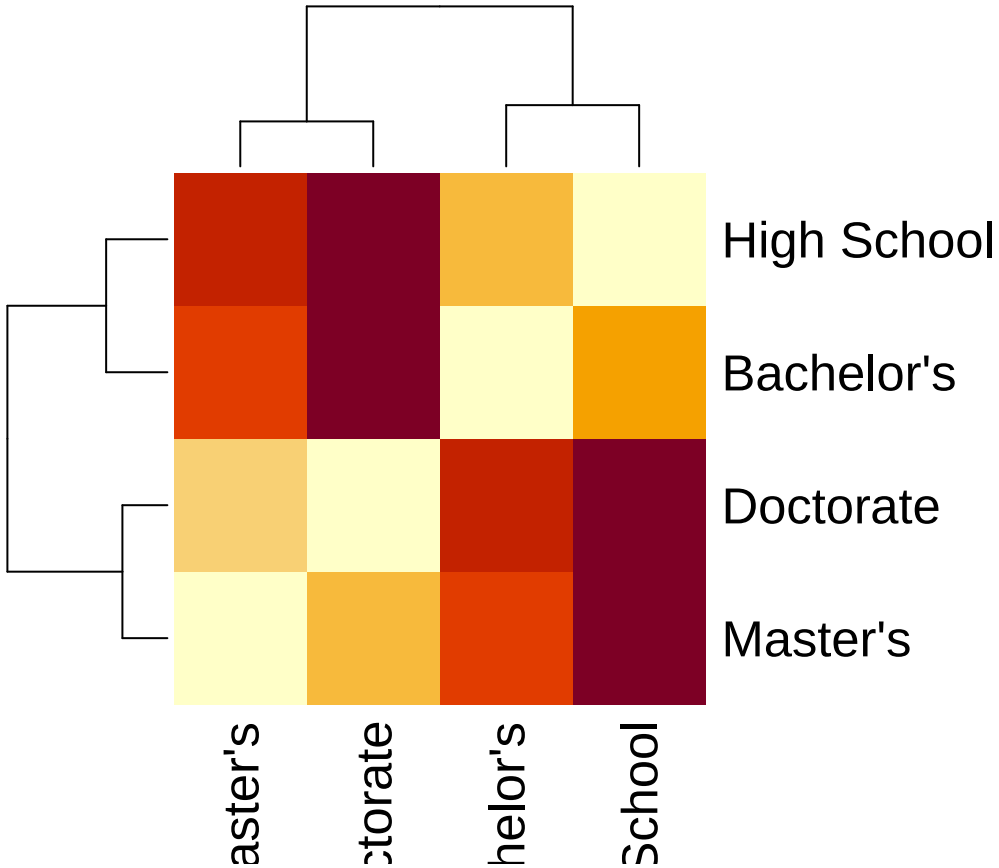
```r
# Column profile:
column_profile <- prop.table(ratings, margin =2)
round(column_profile,4)
```

```
##             Action  Drama Comedy Documentary
## High School 0.5495 0.3297 0.3306      0.1064
## Bachelor's  0.3297 0.4396 0.4959      0.3191
## Master's    0.1099 0.2198 0.1653      0.5319
## Doctorate   0.0110 0.0110 0.0083      0.0426
```

```r
# Question 1b:
#Chi square distance between row profiles
chi_row <- dist(row_profile, method = "euclidean")
chi_row_profile <- as.matrix(chi_row)
print(chi_row_profile)
```

```
##             High School Bachelor's  Master's Doctorate
## High School   0.0000000  0.1928406 0.4320092 0.5077524
## Bachelor's    0.1928406  0.0000000 0.3126944 0.4198214
## Master's      0.4320092  0.3126944 0.0000000 0.1317616
## Doctorate     0.5077524  0.4198214 0.1317616 0.0000000
```

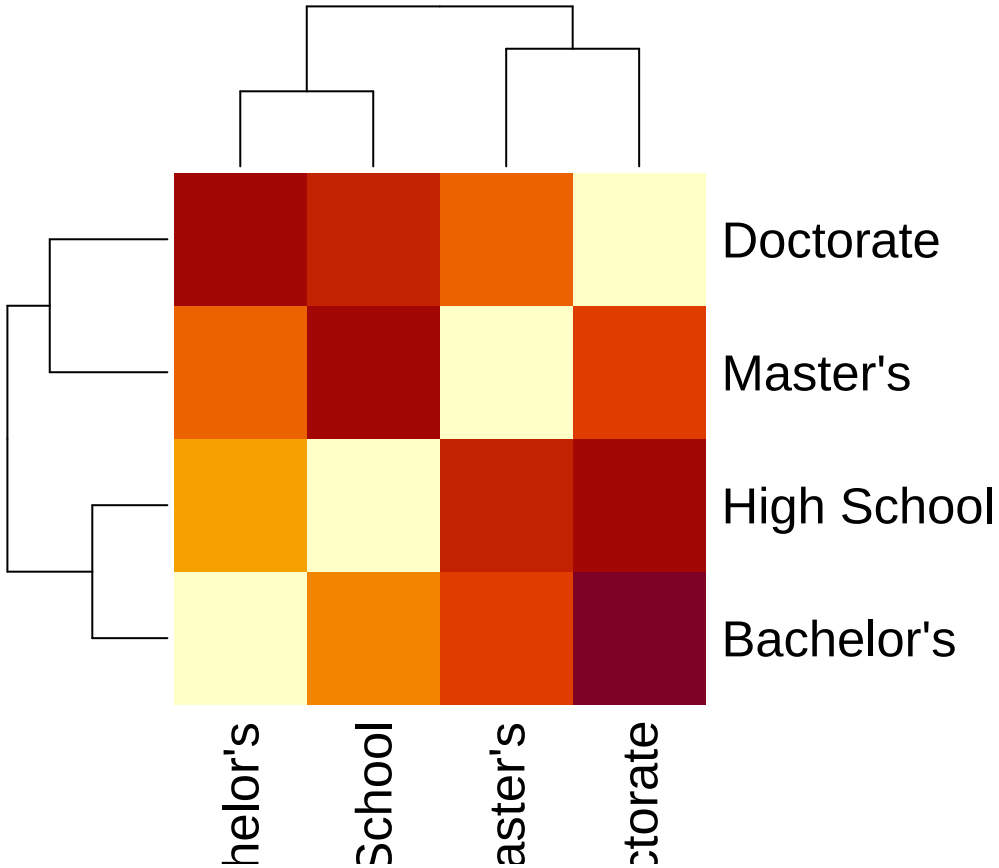
```r
heatmap(chi_row_profile)
```




```
#Chi square distance between column profiles
chi_col <- dist(column_profile, method = "euclidean")
chi_col_profile <- as.matrix(chi_col)
print(chi_col_profile)
```

```
##             High School Bachelor's  Master's Doctorate
## High School   0.0000000  0.3646492 0.6431852 0.7067243
## Bachelor's    0.3646492  0.0000000 0.5011569 0.7742696
## Master's      0.6431852  0.5011569 0.0000000 0.5634775
## Doctorate     0.7067243  0.7742696 0.5634775 0.0000000
```

```
heatmap(chi_col_profile)
```



Interpretation: The Chi-squared distances between row profiles and column profiles are shown in the heatmap, emphasising how similar or different the movie tastes of education-level groups are. Darker shades indicate greater deviations, while lighter shades indicate smaller deviations(more similarity). Diagonal cells are lightest as each movie preference is identical to itself, so distance is 0.

```
# Question 1c:
expected_value <- outer(rowSums(ratings), colSums(ratings))/sum(ratings)
deviation <- ((ratings - expected_value)^2)/expected_value
deviation
```

```
##                Action       Drama     Comedy Documentary
## High School 13.6272744 0.001801477 0.00270966   14.064353
## Bachelor's   1.4269501 0.392361737 2.18520526    1.886292
## Master's     6.8459136 0.505014112 3.19762453   31.270782
## Doctorate    0.1621329 0.324265898 0.48773879    3.923962
```

```
which.max(deviation)
```

```
## [1] 15
```

Interpretation: Greatest deviation is observation 15, which is 31.2708. This is the masters(row) and documentary(column) combination/cell.

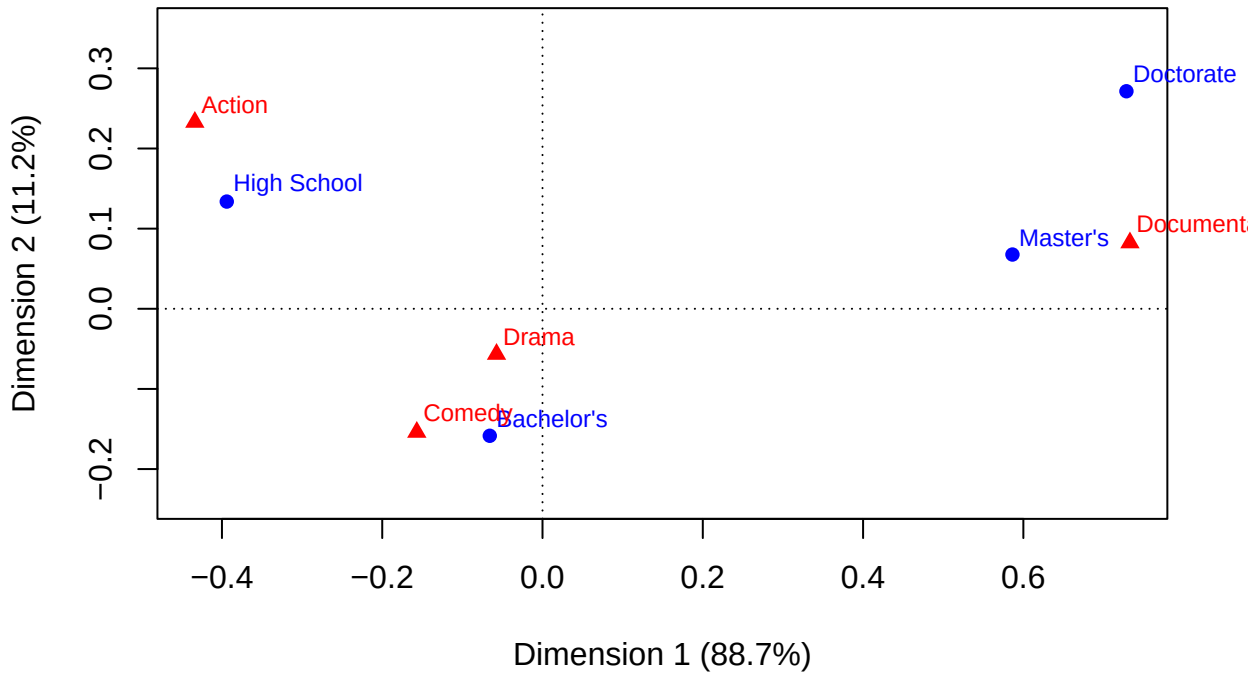
```
# Question 1d:
# First principles:
N <- ratings
P <- N/sum(N)
row_mass <- apply(P, MARGIN = 1, sum)
column_mass <- apply(P, MARGIN = 2, sum)
```

```r
Dr <- diag(row_mass)
Dc <- diag(column_mass)
S <- sqrt(solve(Dr))%*%(as.matrix(P)-row_mass%*%t(column_mass))%*%sqrt(solve(Dc))
svd_S <- svd(S)
U <- svd_S$u
D <- svd_S$d
V <- svd_S$v

# Principal row coordinates
principal_row <- sqrt(solve(Dr)) %*% U %*% diag(D)

# Principal column coordinates
principal_column <- sqrt(solve(Dc)) %*% V %*% diag(D)

# plot
plot <- plot(ca(ratings))
```




```r
plot
```

```
## $rows
##                    Dim1       Dim2
## High School -0.39396301  0.1337581
## Bachelor's  -0.06580259 -0.1584814
## Master's     0.58638171  0.0676958
## Doctorate    0.72859931  0.2714369
##
```

```
## $cols
##                    Dim1        Dim2
## Action      -0.43381302  0.23286915
## Drama       -0.05738487 -0.05665754
## Comedy      -0.15686541 -0.15394541
## Documentary  0.73299729  0.08242526
```

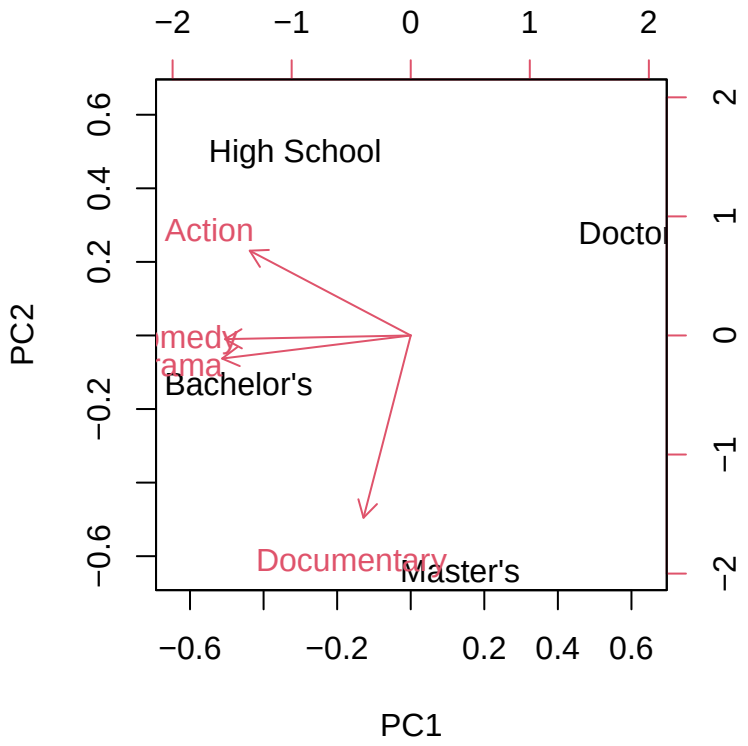
Interpretation: The CA biplot illustrates a strong association between movie preference and educational level. Doctorate holders have a strong preference for documentaries, while recent high school graduates tend to like action movies. Respondents with bachelor's degrees exhibit neutral, mediocre profiles and have poor associations with comedy and drama. The primary difference between lower and higher education choices is captured by Dimension 1, which accounts for 88.7% of the inertia, whereas Dimension 2 only adds a little amount of variation (11.2%).

```
# Question 1e:
summary(ca(ratings))
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.145916  88.7  88.7  **********************
##  2      0.018494  11.2  99.9  ***
##  3      0.000148   0.1 100.0
##         -------- -----
##  Total: 0.164558 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | HghS |  328 1000  345 | -394 897 349 |  134 103 317 |
## 2 | Bchl |  410 1000   73 |  -66 147  12 | -158 853 557 |
## 3 | Mstr |  246 1000  521 |  586 987 579 |   68  13  61 |
## 4 | Dctr |   16  987   61 |  729 867  60 |  271 120  65 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Actn |  186 1000  275 | -434 776 241 |  233 224 547 |
## 2 | Dram |  373  967   15 |  -57 490   8 |  -57 478  65 |
## 3 | Cmdy |  248  995   73 | -157 507  42 | -154 488 318 |
## 4 | Dcmn |  193 1000  637 |  733 987 709 |   82  12  71 |
```

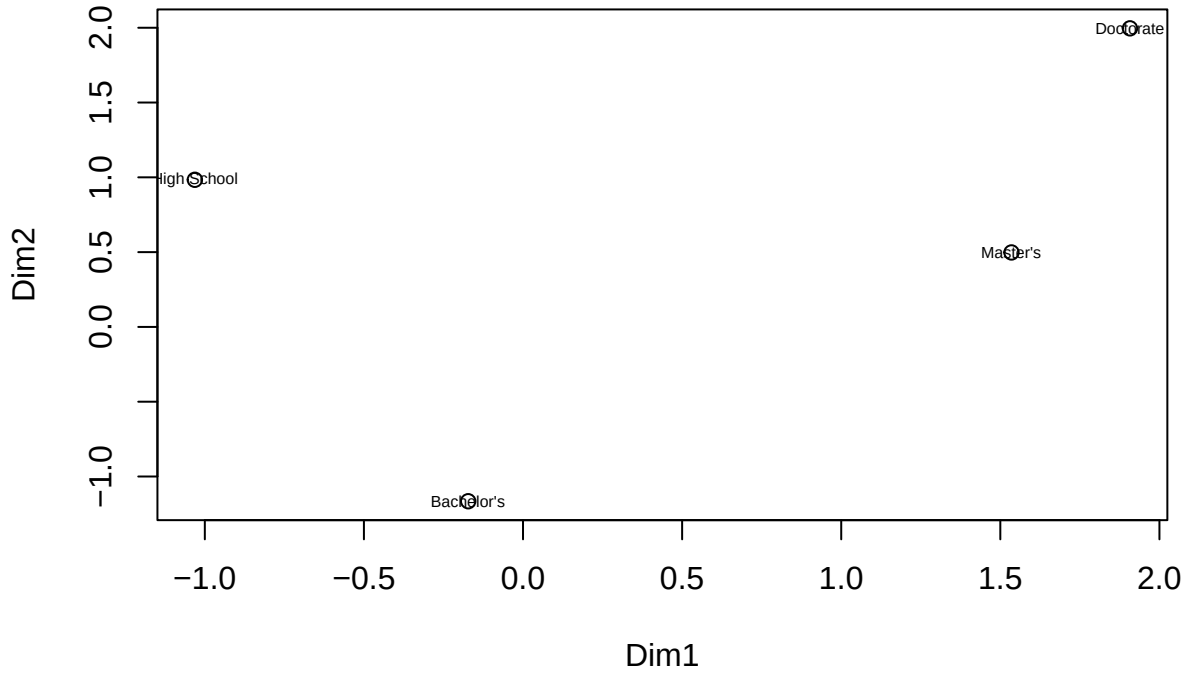
Interpretation: An outstanding low-dimensional representation of the data is shown by the fact that the first two dimensions account for 99.9% of the total inertia. The majority of the relationship structure (88.7%) is explained by the first dimension alone, with the second dimension adding an extra 11.2% to capture small distinctions. As a result, the two-dimensional CA biplot presents a highly accurate overview of the relationships between movie preferences and educational levels.

```
# Question 1f:
pca <- prcomp(ratings, scale. = TRUE)
biplot(pca)
```



Interpretation: The construction and interpretation of the PCA biplot are different from those of the CA biplot. PCA focusses on explaining variance in scaled data, whereas CA depicts association based on chi-squared distances from independence (i.e., row and column profiles). Consequently, the PCA plot highlights high-frequency patterns, which may overstate the impact of marginal totals. Documentary, for instance, has a significant variance and falls sharply downward in the PCA biplot, however, according to association structure, it is located close to Master's and Doctorate degrees in CA. CA is better suited for categorical data like this since PCA does not maintain the dual nature of rows and columns. CA biplot shows assosciation while PCA biplot shows linear structure.

```
# Question 1g:
# Focusing on education levels:
fit <- ca(ratings)
plot(fit$rowcoord[,1:2])
text(fit$rowcoord[,1:2], labels = rownames(ratings), cex=0.5)
```



Question 1h: Standardised column coordinates make it difficult to visually evaluate the relationships between columns and rows when column profiles are very similar or located around the centroid. Furthermore, the biplot becomes cluttered or uninformative if the inertia is low or the column masses are too small.

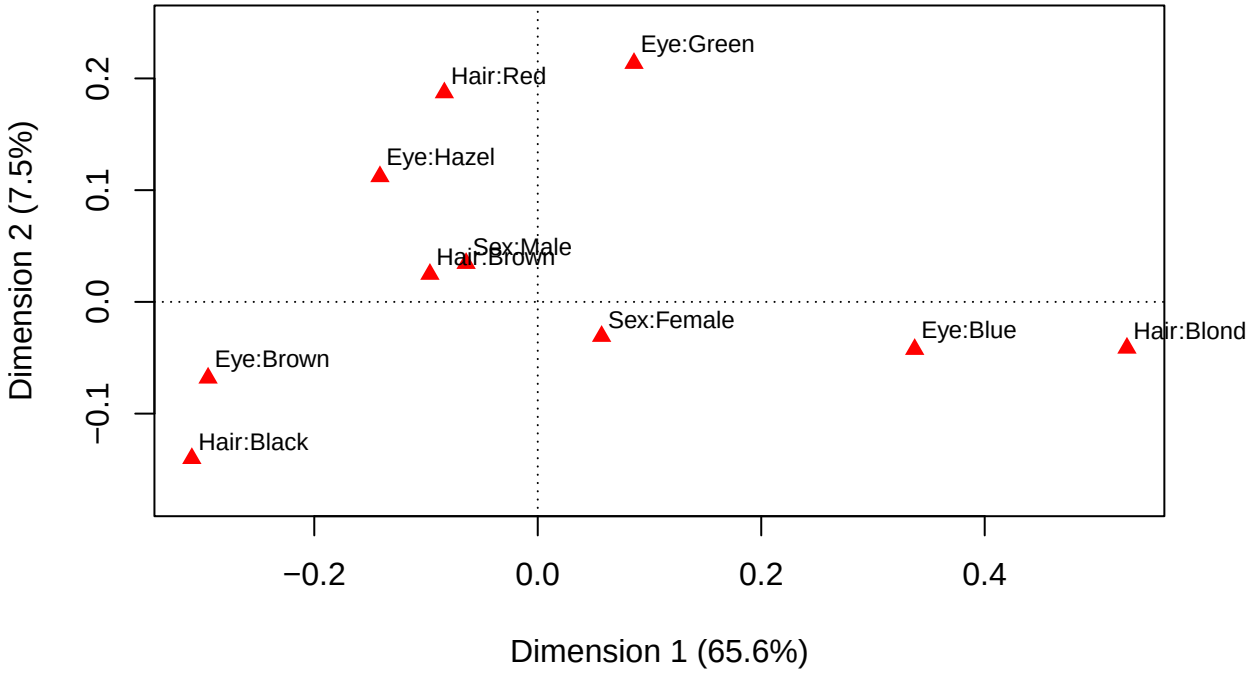
```
# Question 2a:
library(ca)
hair_eye_data <- HairEyeColor
# Transforming for MCA
dat <- as.data.frame(hair_eye_data)
# expanding by frequency
mca_data <- dat[rep(1:nrow(dat), dat$Freq), 1:3]
```

```
mca <- mjca(mca_data, lambda = "adjusted")
summary(mca)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.054579  65.6  65.6  **********************
##  2      0.006263   7.5  73.1  ***
##  3      0.000871   1.0  74.1
##         -------- -----
##  Total: 0.083229
##
##
## Columns:
```

```
##            name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1   | Hair:Black |   61  738  117 | -310 613 107 | -140 125 191 |
## 2   | Hair:Brown |  161  691   71 |  -97 648  27 |   25  43  16 |
## 3   |   Hair:Red |   40  667  119 |  -84 111   5 |  187 556 224 |
## 4   | Hair:Blond |   72  745  126 |  527 741 364 |  -41   5  19 |
## 5   |  Eye:Brown |  124  724   96 | -295 688 198 |  -68  37  92 |
## 6   |   Eye:Blue |  121  754  100 |  337 742 252 |  -42  12  35 |
## 7   |  Eye:Hazel |   52  739  114 | -141 453  19 |  112 286 105 |
## 8   |  Eye:Green |   36  677  121 |   86  95   5 |  214 582 262 |
## 9   |   Sex:Male |  157  588   72 |  -64 456  12 |   35 132  30 |
## 10  | Sex:Female |  176  588   64 |   57 456  11 |  -31 132  27 |
```

```
plot(mca)
```



Interpretation: The MCA biplot shows correlations between sex, eye colour, and hair colour. The main contrast between dark features (Hair:Black, Eye:Brown) and light features (Hair:Blond, Eye:Blue), as well as the special grouping of Hair:Red and Eye:Green, are captured by the first dimension (65.6%). These three groupings are examples of combinations that contrast visually. A secondary contrast is given by Dimension 2, which adds 7.5% and distinguishes Hair: Red and Eye: Green from the others. Females are marginally different on Dimension 2, and sex has an insignificant impact. The biplot provides a visual summary that is moderately informative, capturing 73.1% of the inertia overall.