# Correspondence analysis: introduction

Correspondence Analysis (CA) is a multivariate statistical technique used to analyze and visualize the relationships between categorical variables in a contingency table. It reduces the dimensionality of the data, representing the associations between rows and columns in a low-dimensional space, typically two dimensions, for an easier interpretation. The Chi-squared distance between rows or columns to highlight associations:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(n_{ij} - Np_{ij})^2}{Np_{ij}}$$

where $n_{ij}$ is the observed frequency and $Np_{ij}$ is the expected frequency under the independence assumption.

# 'HairEyeColor' dataset

**HairEyeColor**: dataset of 592 observations x 3 variables.

**Hair**: qualitative variable: Black, Brown, Red, Blond
**Eye**: qualitative variable: Brown, Blue, Hazel, Green
**Sex**: qualitative variable: Male, Female

```
> head(HaiEyeColor)
> HaiEyeColor
> head(HairEyeColor)
, , Sex = Male

        Eye
Hair     Brown Blue Hazel Green
  Black     32   11    10     3
  Brown     53   50    25    15
  Red       10   10     7     7
  Blond      3   30     5     8

, , Sex = Female

        Eye
Hair     Brown Blue Hazel Green
  Black     36    9     5     2
  Brown     66   34    29    14
  Red       16    7     7     7
  Blond      4   64     5     8
```

# Summary

In order to perform a correspondence analysis, we typically call the function 'ca()' from the package 'ca' on a contingency table. The summary of our analysis is displayed below.

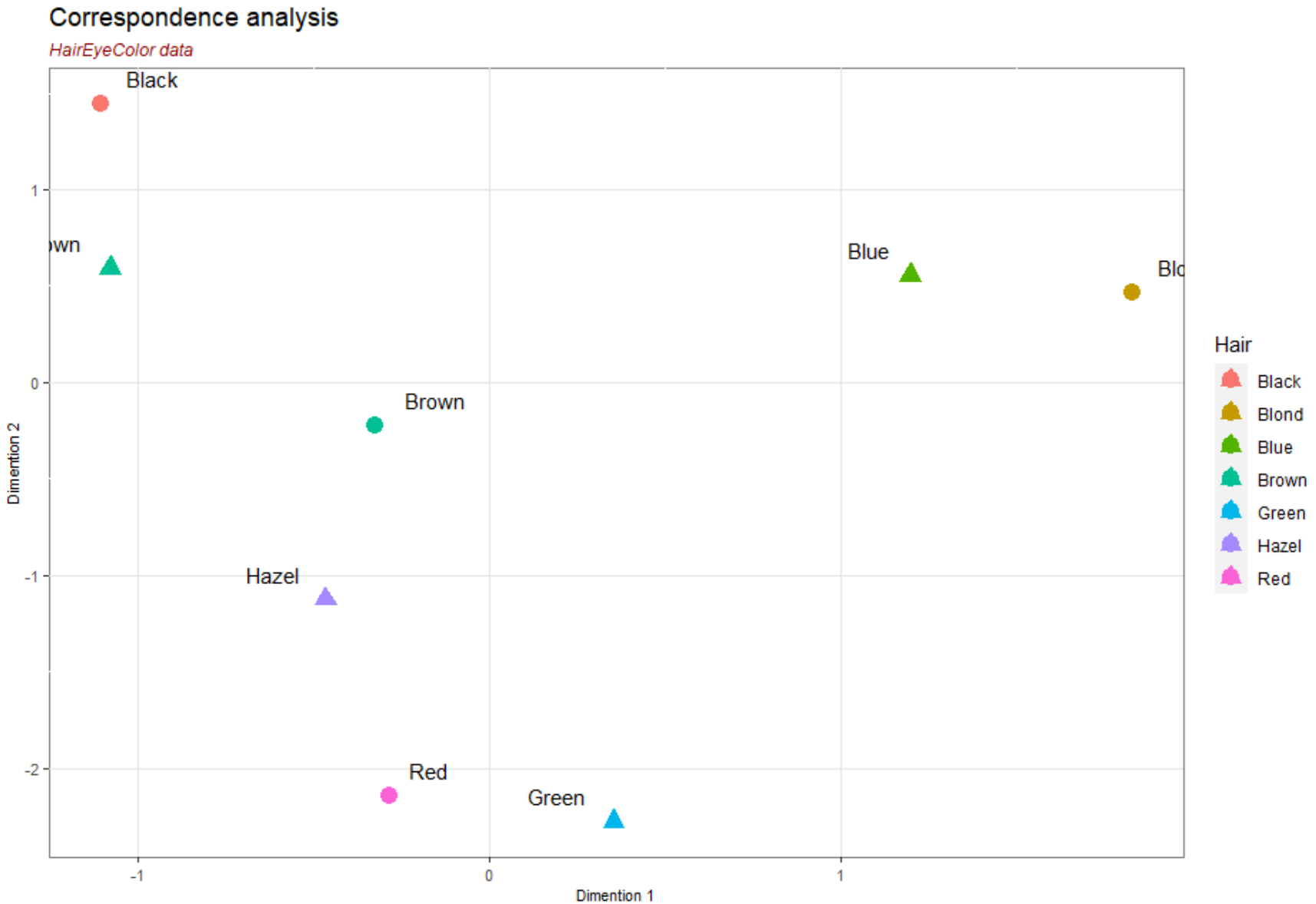
```
> ca_result = ca(contingency_table)
> ca_result

 Principal inertias (eigenvalues):
            1        2        3
Value       0.208773 0.022227 0.002598
Percentage  89.37%   9.52%    1.11%

 Rows:
              Black      Brown        Red     Blond
Mass      0.182432   0.483108   0.119932 0.214527
ChiDist   0.551192   0.159461   0.354770 0.838397
Inertia   0.055425   0.012284   0.015095 0.150793
Dim. 1   -1.104277  -0.324463  -0.283473 1.828229
Dim. 2    1.440917  -0.219111  -2.144015 0.466706

 Columns:
              Brown      Blue      Hazel      Green
Mass      0.371622 0.363176   0.157095   0.108108
ChiDist   0.500487 0.553684   0.288654   0.385727
Inertia   0.093086 0.111337   0.013089   0.016085
Dim. 1   -1.077128 1.198061  -0.465286   0.354011
```

# Plot of factors in main dimentions



# Main observations

- Dimension Reduction: The relationships between hair color and eye color in a lower-dimensional space, here the first two on the plot.
- Association Visualization: Points close to each other in the plot indicate a stronger association between the corresponding hair and eye colors. For example, if "Black Hair" and "Brown Eyes" are close together (frequently observed together).
- Dimensional Interpretation: The axes (Dimension 1 and Dimension 2) represent the principal dimensions that capture the most variance in the data.
- Categorical Differentiation: The plot visually differentiates between hair and eye colors using different shapes and colors, making it easy to interpret the correspondence between categories.

# References

An Introduction to Applied Multivariate Analysis with R, 2011, B. Everitt, T. Hothorn, Springer, e-ISBN 978-1-4419-9650-3

The R Project for Statistical Computing:
https://www.r-project.org/