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

```
data = {
    'Hair': ['Black', 'Brown', 'Red', 'Blond', 'Black', 'Brown', 'Red', 'Blond',
      'Black', 'Brown', 'Red', 'Blond', 'Black', 'Brown', 'Red', 'Blond'],
    'Eye': ['Brown', 'Brown', 'Brown', 'Brown', 'Blue', 'Blue', 'Blue', 'Blue',
     'Hazel', 'Hazel', 'Hazel', 'Hazel', 'Green', 'Green', 'Green', 'Green'],
    'Freq': [32, 53, 10, 3, 11, 50, 10, 30, 10, 25, 7, 5, 3, 15, 7, 8]
}

hair_eye = pd.DataFrame(data)
contingency_table = hair_eye.pivot_table(values='Freq', index='Hair', columns='
     Eye', fill_value=0)
contingency_table

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

Summary

To perform a correspondence analysis, we can create the function:

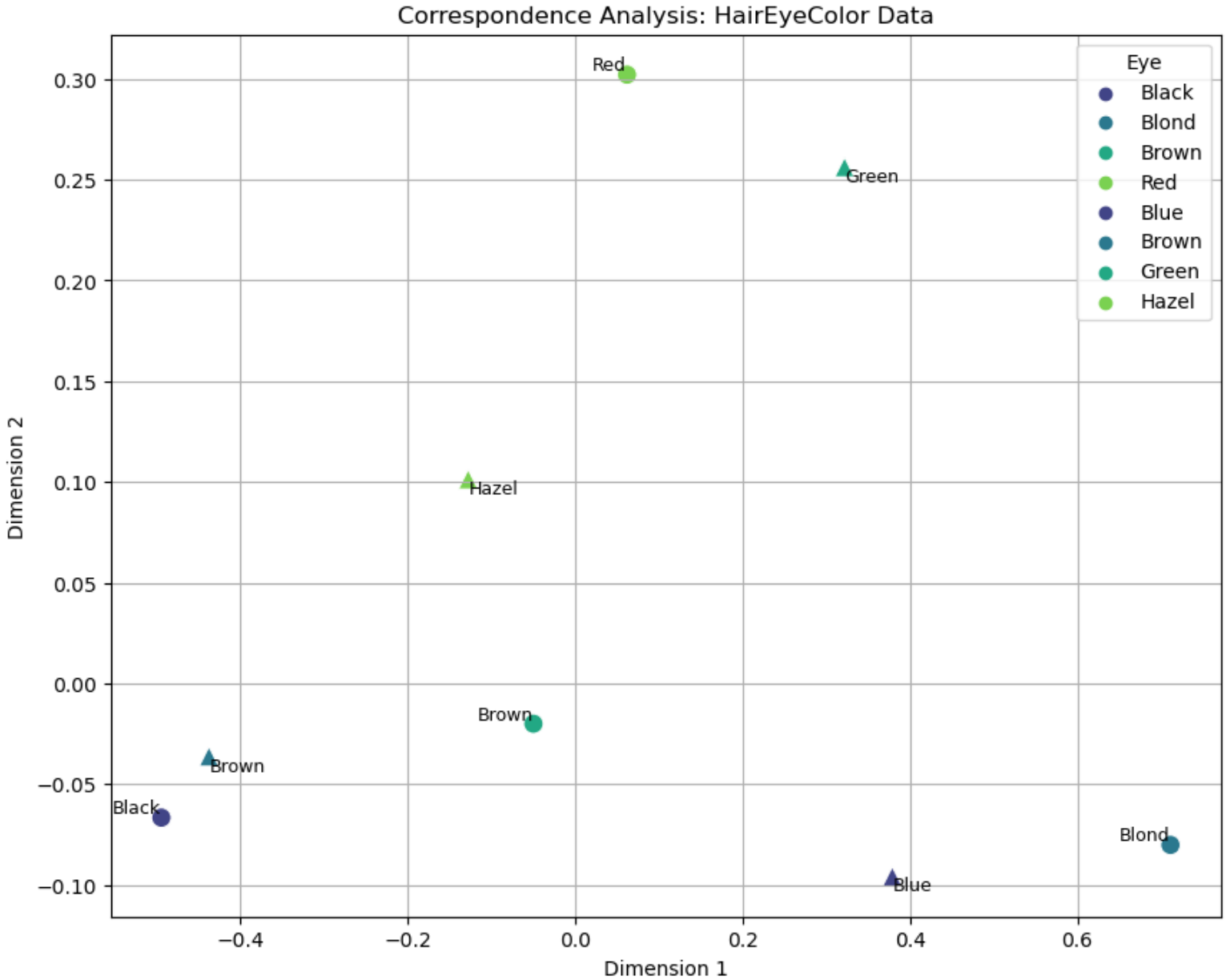
```
# Correspondence Analysis
def correspondence_analysis(table):
    table = table.values
    n = np.sum(table)
    P = table / n
    r = np.sum(P, axis=1)
    c = np.sum(P, axis=0)
    R = np.diag(r)
    C = np.diag(c)
    S = P - np.outer(r, c)
    R_inv_sqrt = np.linalg.inv(np.sqrt(R))
    C_inv_sqrt = np.linalg.inv(np.sqrt(C))
    Z = R_inv_sqrt @ S @ C_inv_sqrt
    U, D, Vt = svd(Z, full_matrices=False)
    F = R_inv_sqrt @ U @ np.diag(D)
    G = C_inv_sqrt @ Vt.T @ np.diag(D)
    return F, G, D

F, G, D = correspondence_analysis(contingency_table)

row_coords = pd.DataFrame(F[:, :2], columns=['Dim1', 'Dim2'])
row_coords['Hair'] = contingency_table.index

col_coords = pd.DataFrame(G[:, :2], columns=['Dim1', 'Dim2'])
col_coords['Eye'] = contingency_table.columns
```

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

Python:
https://www.python.org/