# Phenotypic Association Analysis in HairEyeColor

**A contingency-table study with publication-style visualization**

Daniel Harrod

2026-02-17

## Table of contents

# 1 Abstract

This report analyzes the relationship between hair color and eye color using `datasets::HairEyeColor`, a built-in categorical dataset in R.
The analysis applies contingency-table inference through chi-square tests of independence, standardized residual diagnostics, and Cramer's V effect sizes, with results contextualized by sex.
Visual outputs are designed to match academic presentation standards, with consistent typography, colorblind-aware palettes, and high-resolution exports.

# 2 Introduction

Categorical trait combinations are commonly analyzed in population studies to identify non-random structure in observed frequencies.
Although `HairEyeColor` is not a direct genotype dataset, it is an appropriate pedagogical proxy for studying trait-association methodology central to genetics-oriented data science.

This project addresses three questions:

1. Are hair color and eye color independent in the pooled sample?
2. Does this association persist when stratified by sex?
3. Which trait pairings contribute most strongly to departures from independence?

# 3 Data and Preprocessing

The source data are a three-way contingency table (`Hair`, `Eye`, `Sex`) with observed counts (`Freq`).

Table 1: First 12 rows of the long-format HairEyeColor table.

| Hair | Eye | Sex | Freq |
|---|---|---|---|
| Black | Brown | Male | 32 |
| Brown | Brown | Male | 53 |
| Red | Brown | Male | 10 |
| Blond | Brown | Male | 3 |
| Black | Blue | Male | 11 |
| Brown | Blue | Male | 50 |
| Red | Blue | Male | 10 |
| Blond | Blue | Male | 30 |
| Black | Hazel | Male | 10 |

| Hair | Eye | Sex | Freq |
|---|---|---|---|
| Brown | Hazel | Male | 25 |
| Red | Hazel | Male | 7 |
| Blond | Hazel | Male | 5 |

# 4 Methods

## 4.1 Independence Testing

For each contingency table (overall, male, and female), a chi-square test of independence is used:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where ( O_{ij} ) and ( E_{ij} ) are observed and expected counts.

## 4.2 Effect Size

Cramer's V quantifies association strength while accounting for table size:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

## 4.3 Residual Diagnostics

Standardized residuals identify cells with unusually high or low counts under independence assumptions.
As a practical heuristic, absolute residuals greater than 2 indicate meaningful local deviations.

# 5 Results

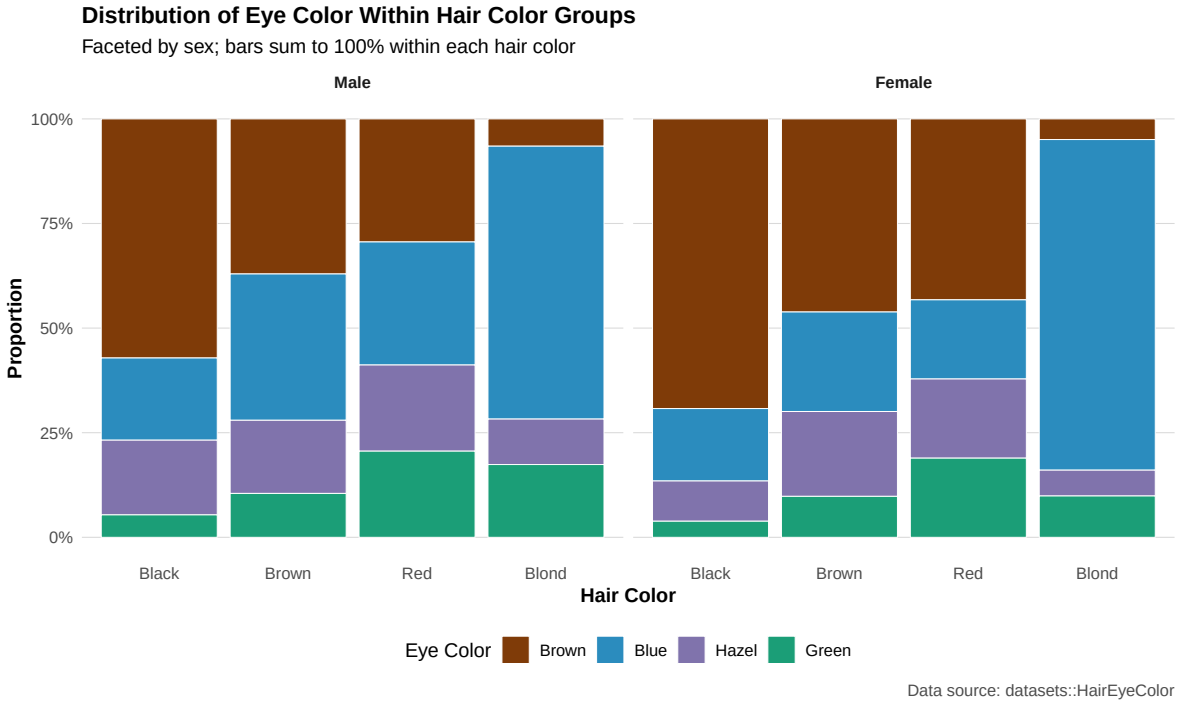
## 5.1 Statistical Test Summary

Table 2: Chi-square results and effect sizes by model.

| model | chi_square | df | p_value | cramers_v |
|---|---|---|---|---|
| Overall | 138.29 | 9 | 0.0e+00 | 0.279 |
| Male | 41.28 | 9 | 4.4e-06 | 0.222 |
| Female | 106.66 | 9 | 0.0e+00 | 0.337 |

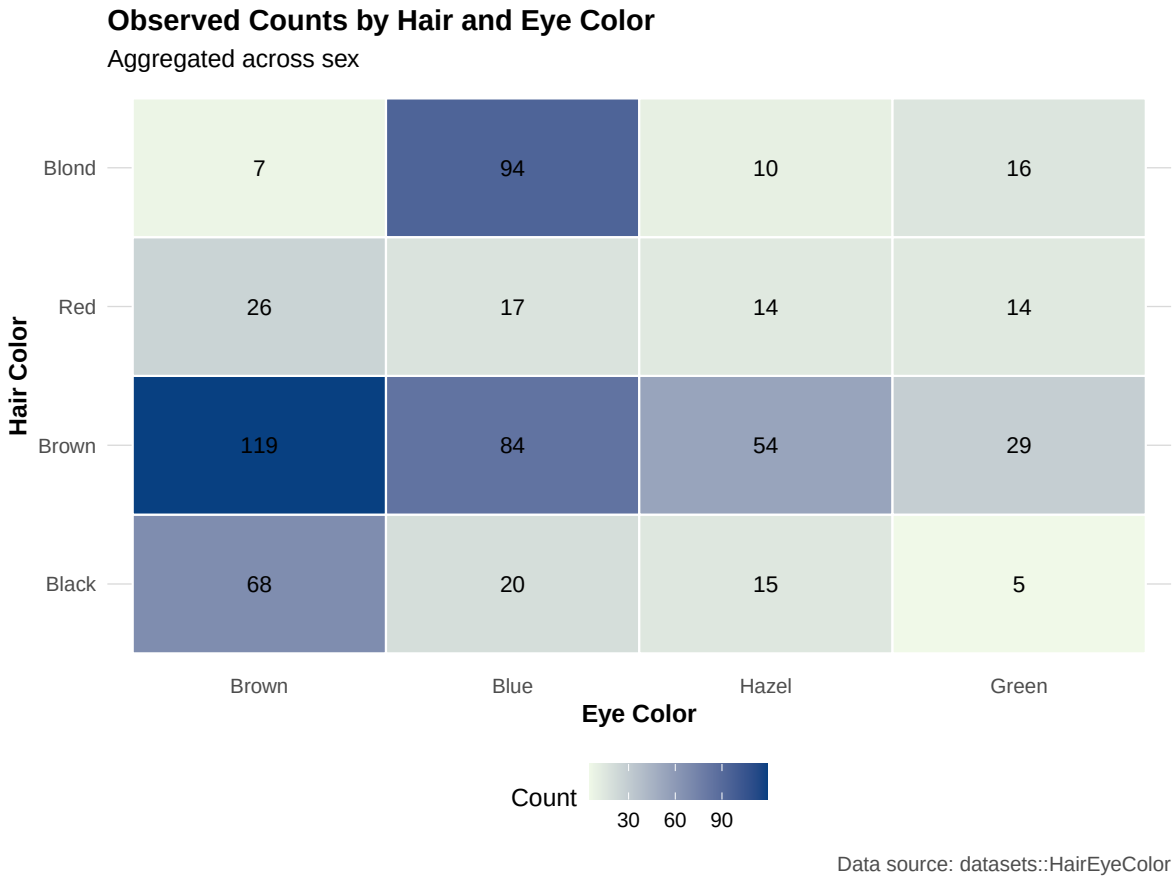
The table shows whether trait association is statistically detectable and how large that association is in practical terms.

## 5.2 Figure 1: Proportional Composition by Sex



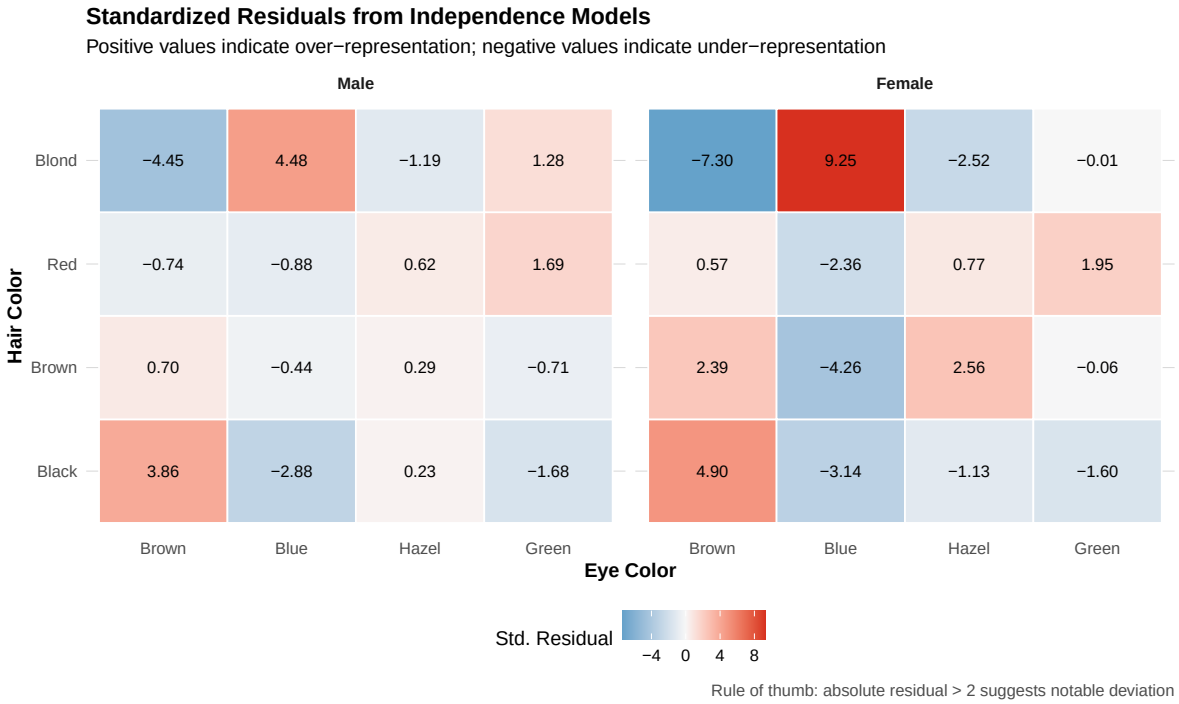

This chart compares eye-color composition within each hair-color category, separated by sex. Differences in stacked proportions indicate heterogeneous trait association patterns.

## 5.3 Figure 2: Observed Frequency Heatmap



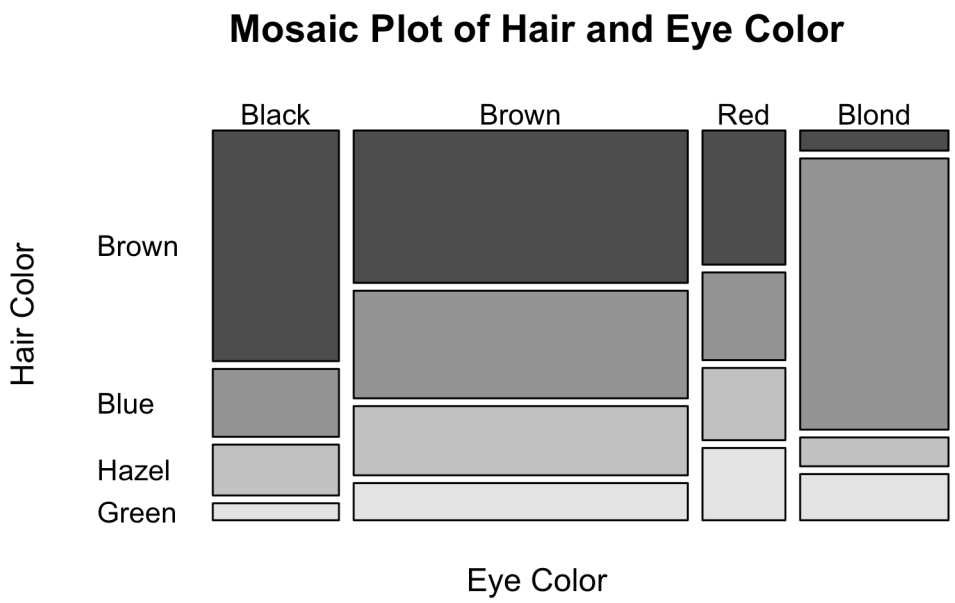

Higher-intensity cells represent more common hair-eye pairings in the pooled sample.

## 5.4 Figure 3: Standardized Residual Heatmap by Sex




Positive residuals indicate over-represented pairings; negative residuals indicate under-represented pairings.

### 5.5 Figure 4: Mosaic Diagnostic Plot




The mosaic plot provides a compact view of cell prevalence and category structure in the pooled contingency table.

## 6 Discussion

The analysis demonstrates how a compact categorical dataset can support rigorous inference and visual diagnostics.
Statistical significance and effect-size interpretation should be considered jointly: a low p-value indicates detectable structure, while Cramer's V contextualizes its magnitude.

### 6.1 Limitations

- This is an aggregated historical dataset, not an individual-level modern cohort.
- Trait categories are coarse and do not represent the full biological complexity of pigmentation genetics.
- The chi-square framework assumes adequate expected counts; sparse cells can influence stability.

## 7 Conclusion

`HairEyeColor` provides a useful teaching framework for genetics-adjacent contingency analysis in R.
A reproducible pipeline combining inferential testing, residual diagnostics, and publication-grade figures can produce a report suitable for graduate-level submission standards.

## 8 Reproducibility

```
R version 4.3.2 (2023-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS 15.7.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  L

locale:
[1] C

time zone: America/Phoenix
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] tidyr_1.3.1   dplyr_1.1.4   ggplot2_4.0.2

loaded via a namespace (and not attached):
 [1] vctrs_0.6.5        cli_3.6.5          knitr_1.50         rlang_1.1.7
 [5] xfun_0.51          png_0.1-8          purrr_1.2.1        generics_0.1.3
 [9] textshaping_0.3.7  S7_0.2.1           jsonlite_1.9.1     labeling_0.4.3
[13] glue_1.8.0         htmltools_0.5.8.1  ragg_1.2.7         scales_1.4.0
[17] rmarkdown_2.29     grid_4.3.2         evaluate_1.0.5     tibble_3.2.1
[21] fastmap_1.2.0      yaml_2.3.10        lifecycle_1.0.4    compiler_4.3.2
[25] RColorBrewer_1.1-3 pkgconfig_2.0.3    systemfonts_1.0.5  farver_2.1.2
[29] digest_0.6.37      R6_2.6.1           tidyselect_1.2.0   dichromat_2.0-0.1
[33] pillar_1.10.1      magrittr_2.0.3     withr_3.0.2        tools_4.3.2
[37] gtable_0.3.6
```