HairEyeColor

Boro Kelvin Waweru

2024-03-07

Introduction

The HairEyeColor data set provides information on the frequencies of individuals with different combinations of hair and eye colors. In this report, we conduct a thorough analysis of the data set to explore relationships between these variables and to understand any underlying patterns. The HairEyeColor data set is provided by the '***datasets***' package in R.

Data Overview:

```
library(tidyverse)

data("HairEyeColor")
glimpse(HairEyeColor)
```

```
 'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 ...
 - attr(*, "dimnames")=List of 3
  ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
  ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
  ..$ Sex : chr [1:2] "Male" "Female"
```

The **glimpse()** function from the '***tidyverse***' package gives a concise summary of a data frame. From the results of the **data()** and **glimpse()** functions, it looks like the data is a three-dimensional table representing counts. We should it to a data frame first before continuing with the analysis.

```
HairEyeColor <-as.data.frame(HairEyeColor)

dim(HairEyeColor)
```

```
[1] 32  4
```

```
names(HairEyeColor)
```

```
[1] "Hair" "Eye"  "Sex"  "Freq"
```

```
library(knitr)

# view the HairEyeColor data set
kable(HairEyeColor)
```

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
Brown	Hazel	Male	25
Red	Hazel	Male	7
Blond	Hazel	Male	5
Black	Green	Male	3
Brown	Green	Male	15
Red	Green	Male	7
Blond	Green	Male	8
Black	Brown	Female	36
Brown	Brown	Female	66
Red	Brown	Female	16
Blond	Brown	Female	4
Black	Blue	Female	9
Brown	Blue	Female	34
Red	Blue	Female	7
Blond	Blue	Female	64
Black	Hazel	Female	5
Brown	Hazel	Female	29
Red	Hazel	Female	7

Hair	Eye	Sex	Freq
Blond	Hazel	Female	5
Black	Green	Female	2
Brown	Green	Female	14
Red	Green	Female	7
Blond	Green	Female	8

The HairEyeColor data set consists of categorical variables representing hair color (black, brown, red, blond), eye color (blue, brown, hazel, green), sex(male, female) and frequency counts. It contains 32 observations.

Exploratory Data Analysis:

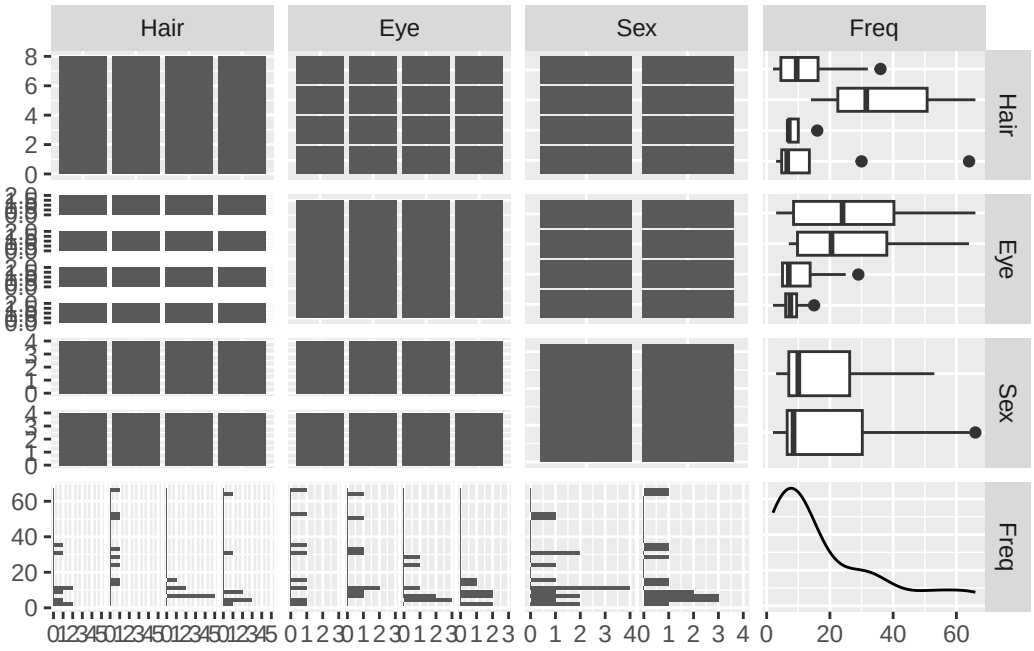
To gain insights into the data set, we employ various visualization techniques. Visualizations aid in identifying trends, patterns, and potential associations in the data.

Pairwise scatterplot matrix:

This visualization allows us to examine potential relationships between different pairs of variables. We look for patterns or clusters that may indicate associations between hair and eye colors.

```
library(GGally)

ggpairs(HairEyeColor)
```

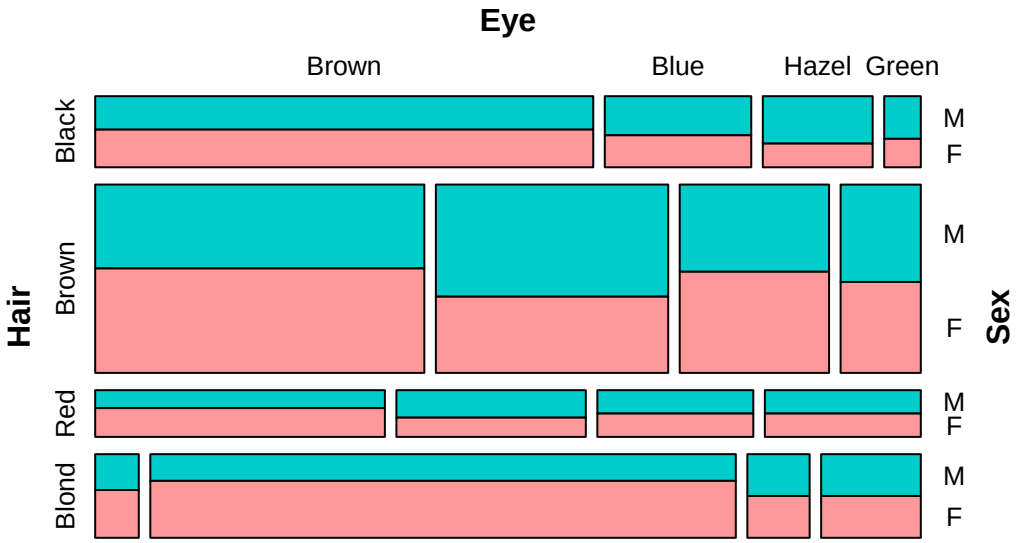


Mosaic plot:

We use mosaic plots to visualize the joint distribution of hair and eye colors, with a focus on differentiating by sex. This helps us identify any patterns or discrepancies in the distribution across categories.

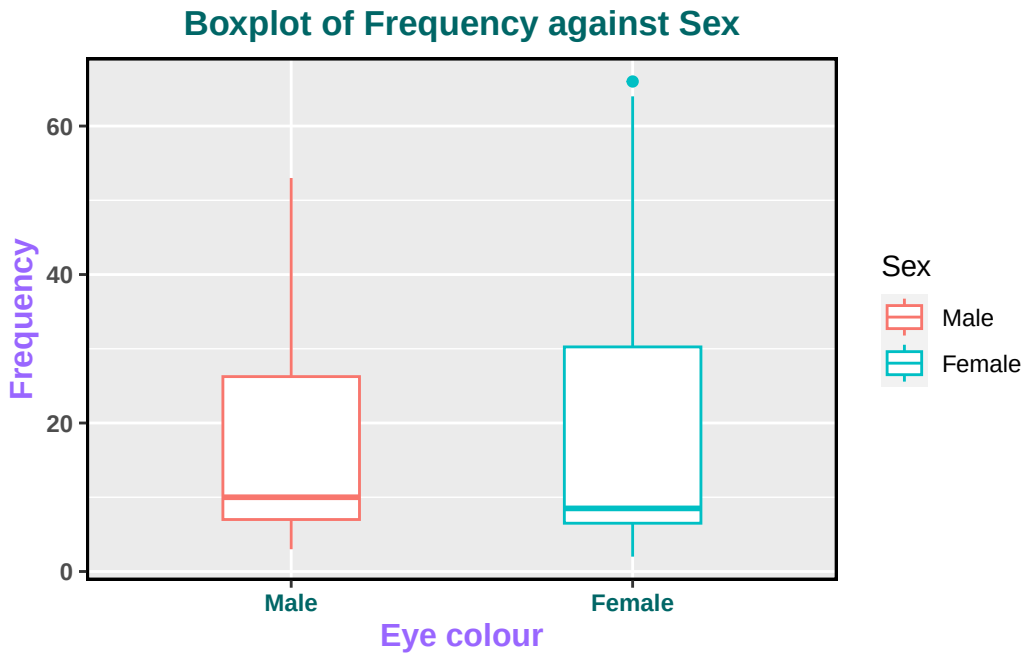
```
library(vcd)

mosaic(~ Hair + Sex + Eye, data = HairEyeColor,
       direction = c("h","h","v"),
       highlighting = "Sex",
       highlighting_fill = c("#00CCCC", "#FF9999"),
       gp_labels = gpar(fontsize = 9.5),
       rot_labels = c(0, 0, 0, 90),
       abbreviate_labs = c(6, 6, 1),
       pos_labels = c("center"))
```

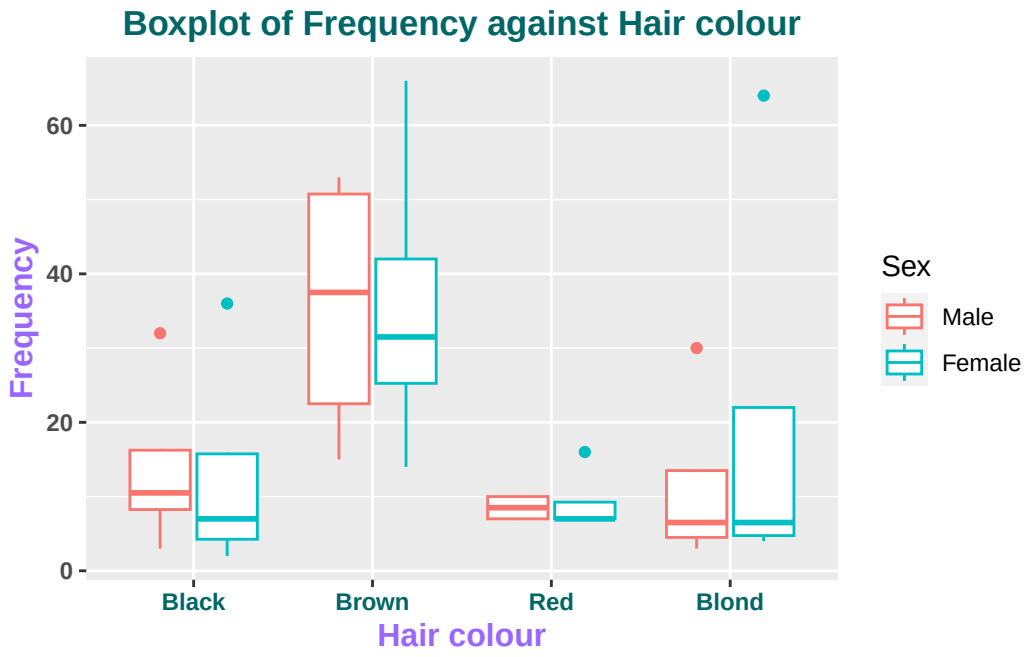


Boxplots:

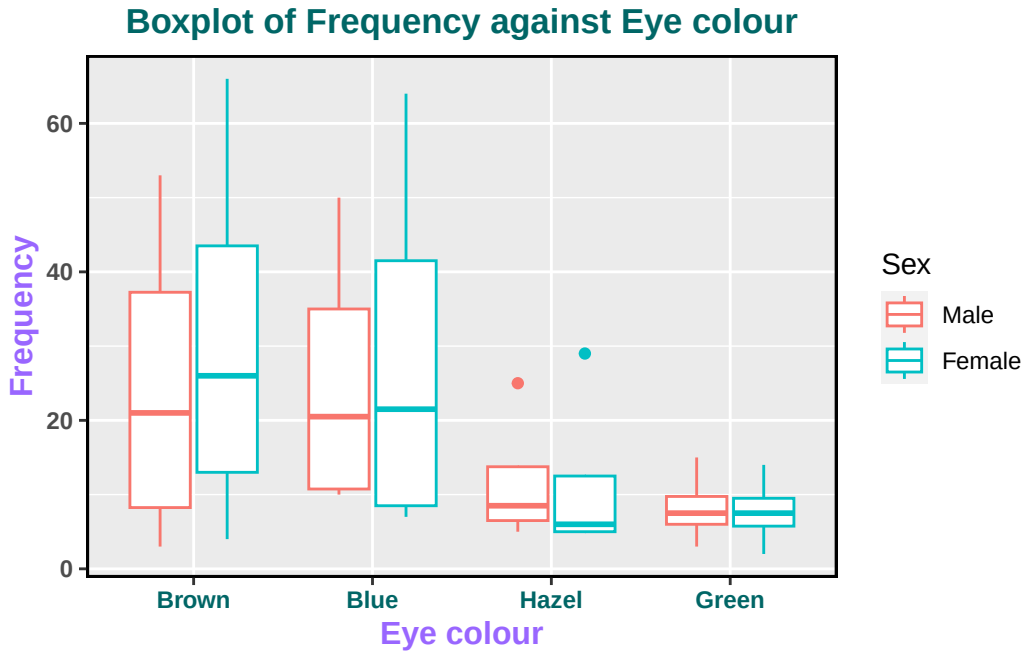
Boxplots are employed to visualize the distribution of frequency counts across different categories of hair and eye colors. We look for variations in frequency distributions across these categories.



The median frequency count for both males and females is slightly lower in females than in males. However, the range of frequency counts for females appears slightly wider than that of males, suggesting greater variability in the number of individuals with specific hair and eye colour combinations among females.



The median frequency counts vary across different hair colours, with some colours having higher median counts than others. Variability in frequency counts varies across hair colour categories, as evidenced by the width of the boxplots. Outliers are also observed in either groups, indicating extreme frequency counts that deviate from the typical distribution.



Eye colour categories such as brown and blue appear to have higher median frequency counts compared to hazel and green. The median frequency for males with brown and blue eye colours is similar suggesting comparable distributions. The median and the range of frequencies for individuals with green eye colour, for both genders, is very similar suggesting close symmetry in their distributions.

Statistical Analysis:

Chi-Squared Test:

We perform a chi-squared test to assess the independence between hair and eye colors. This test helps determine if there is a significant association between the variables. We first create a contingency table using the **xtabs()** function where the rows represent different hair colors, the columns represent different eye colours, and the cells contain the corresponding frequencies from the HairEyeColor data set.

```
HairEyeColor_table <- xtabs(Freq ~ Hair + Eye, data = HairEyeColor)
```

```r
chisq_results <- chisq.test(HairEyeColor_table)
print(chisq_results)
```

```
    Pearson's Chi-squared test

data:  HairEyeColor_table
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

The null hypothesis of Pearson's chi-squared test is that there is no association between the two categorical variables. Since the p-value is less than the significance level (typically 0.05), we reject the null hypothesis. In other words, there is evidence to suggest that the distribution of hair color and eye color is not independent; they are associated with each other in some way. Therefore, we conclude that there is a statistically significant association between the Hair and Eye colours, X^2 (df=9, N=592) = 138.29, p< 2.2e-16.

Generalized Linear Model (GLM):

We fit a Poisson regression model to analyze the relationship between hair, eye colors, and frequency counts. Poisson regression is chosen due to the count nature of the response variable. Interaction terms between `hair` and `eye` colors are included to assess potential modification effects. This is because there may be theoretical reasons to believe that the relationship between `Hair` colour and `Freq` differs depending on the `Eye` colour, and vice versa. For example, the effect of `hair` colour on `eye` colour might influence the frequency of certain combinations of `hair` and `eye` colours.

```r
inter_model <- glm(Freq ~ Hair * Eye, data = HairEyeColor, family= poisson)

summary(inter_model)
```

```
Call:
glm(formula = Freq ~ Hair * Eye, family = poisson, data = HairEyeColor)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.5264     0.1213  29.079  < 2e-16 ***
HairBrown           0.5596     0.1520   3.681 0.000232 ***
HairRed            -0.9614     0.2306  -4.170 3.05e-05 ***
HairBlond          -2.2736     0.3969  -5.728 1.02e-08 ***
```

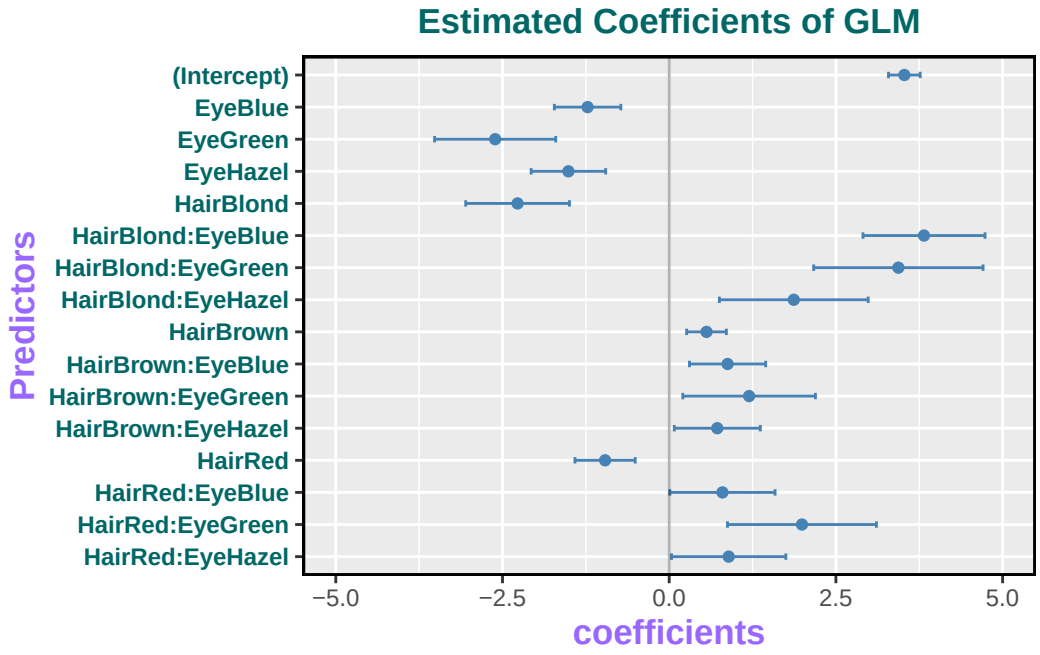
```
EyeBlue             -1.2238     0.2544  -4.811 1.50e-06 ***
EyeHazel            -1.5115     0.2853  -5.299 1.17e-07 ***
EyeGreen            -2.6101     0.4634  -5.633 1.77e-08 ***
HairBrown:EyeBlue    0.8755     0.2916   3.003 0.002677 **
HairRed:EyeBlue      0.7989     0.4025   1.985 0.047153 *
HairBlond:EyeBlue    3.8212     0.4671   8.180 2.83e-16 ***
HairBrown:EyeHazel   0.7213     0.3291   2.192 0.028386 *
HairRed:EyeHazel     0.8924     0.4373   2.041 0.041293 *
HairBlond:EyeHazel   1.8681     0.5694   3.281 0.001035 **
HairBrown:EyeGreen   1.1982     0.5075   2.361 0.018230 *
HairRed:EyeGreen     1.9910     0.5697   3.495 0.000475 ***
HairBlond:EyeGreen   3.4367     0.6481   5.303 1.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 475.12  on 31  degrees of freedom
Residual deviance:  21.81  on 16  degrees of freedom
AIC: 192.69

Number of Fisher Scoring iterations: 4
```

Instead of just printing the results of the Poisson regression model, we can also plot the results for visualization.



The p-values associated with each coefficient indicate the significance of the relationship between the predictor variables and the response variable. Lower p-values (typically < 0.05) suggest that the relationship is statistically significant.

The p-value for `HairBrown` is **<2e-16** indicating a highly significant relationship between `HairBrown` and Freq. The p-value for `HairBlond` is **0.21569**, which is not significant at the 0.05 level, suggesting that the relationship between `HairBlond` and Freq may not be statistically significant.

Each coefficient represents the estimated change in the log of the expected count of the response variable(Freq) associated with a one-unit change in the predictor variable, holding other variables constant.

The coefficient for `HairBrown` is **0.97386**. This suggests that, on average, the log of the expected count of Freq is expected to increase by approximately **0.97386** when comparing `HairBrown` to the reference category (`HairBlack`) holding other variables constant.

The coefficient for `HairRed` is **-0.41945**. This suggests that, on average, the log of the expected count of Freq is expected to decrease by approximately **0.41945** when comparing `HairRed` to the reference category(`HairBlack`), holding other variables constant.

Deviance in the context of chi-squared tests is a goodness-of-fit statistic. The null deviance represents the deviance of the model with only the intercept (null model). The residual deviance represents the deviance of the fitted model after including the predictor variables. A lower residual deviance compared to the null deviance suggests that the fitted model provides a better fit to the data than the null model.

The dispersion parameter measures how much the observed data deviates from the model's predictions. In Poisson regression, the assumption is that the variance of the response variable is equal to the mean. However, in practice, the observed variance may differ from this assumption, leading to the concept of dispersion. If the variance of the response variable is greater than its mean, then the dispersion parameter would be estimated to be greater than 1. Conversely, if the variance is less than the mean, the dispersion parameter would be estimated to be less than 1. When the dispersion parameter is 1, it indicates that the model's assumptions about the relationship between the mean and variance are met.

The Akaike Information Criterion (AIC) is a measure of the relative goodness of fit of the model. Lower AIC values indicate a better fit, considering the trade-off between model complexity and goodness of fit. The value 192.69 looks low but we should perform other tests to assess the model fit just to be sure.

In addition to the AIC, we can also compute the *McFaden's R2* value using the **PR2()** function to assess how well our model fits the data. This number ranges from 0 to 1.

```
library(pscl)

pR2(inter_model)["McFadden"]
```

```
fitting null model for pseudo-r2
```

```
McFadden
 0.73829
```

We get a value of **0.73829** . This value is very close to 1, which indicates that the model is a very good fit for the data.

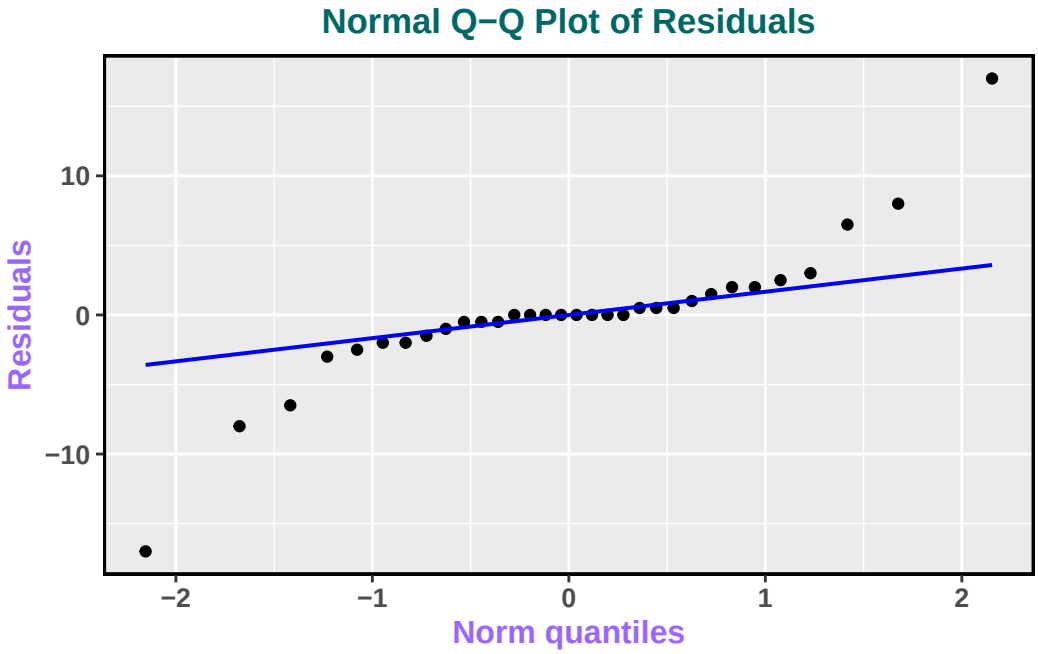
We can also plot a Q-Q plot of the residuals of generalized linear model(GLM) just to observe the distribution of our data. Residuals are the differences between the observed values and the values predicted by the model. This is not necessary, but I'm just doing it to justify the choice for the statistical tests conducted.

```
# Extract residuals
residuals <- residuals(inter_model, type = "response")

# Plot a QQ plot of the residuals using ggplot
residual_data <- data.frame(Residuals = residuals)

ggplot(residual_data, aes(sample = Residuals)) +
  geom_qq() +
  geom_qq_line(colour = "blue", linewidth = .7) +
```

```
labs(title = "Normal Q-Q Plot of Residuals",
     x = "Norm quantiles", y = "Residuals") +
theme(plot.title = element_text(colour = "#006666", face = "bold",
                                hjust = .5),
      panel.border =  element_rect(fill = NA, linewidth = 1),
      axis.title.x = element_text(size = 12, face = "bold",
                                  colour = "#9966FF"),
      axis.title.y = element_text(size = 12, face = "bold",
                                  colour = "#9966FF"),
      axis.text = element_text(face = "bold", size = 9.5))
```



The **residuals()** function in R is used to extract the residuals from a fitted model. This function takes at least two arguments: the fitted model object(inter_model in this case) and an optional type argument. The **`type`** argument specifies the type of residuals to extract. Using **`type = "response"`** ensures that the residuals are on the same scale as the original response variable, which is appropriate for Poisson regression.

As expected, the data doesn't follow a normal distribution. This is because the response variable represents counts of individuals of different combinations of hair and eye colours. This means that the data is binned. We use the generalized linear model when the response variable is not normally distributed.

Conclusion

The analysis of the HairEyeColor data set indicates significant association between hair and eye colours, as evidenced by the results of the chi-squared test. Additionally, the GLM suggests that certain combinations of hair and eye colours are associated with varying frequencies. These findings contribute to our understanding of the relationship between hair and eye colours and provide insights into the distribution of these traits in the population.

Each statistical analysis method is chosen based on the variables and the research questions. The Chi-Squared Test is appropriate for analyzing the association between categorical variables, while the GLM allows for the modeling of count data with categorical predictors.