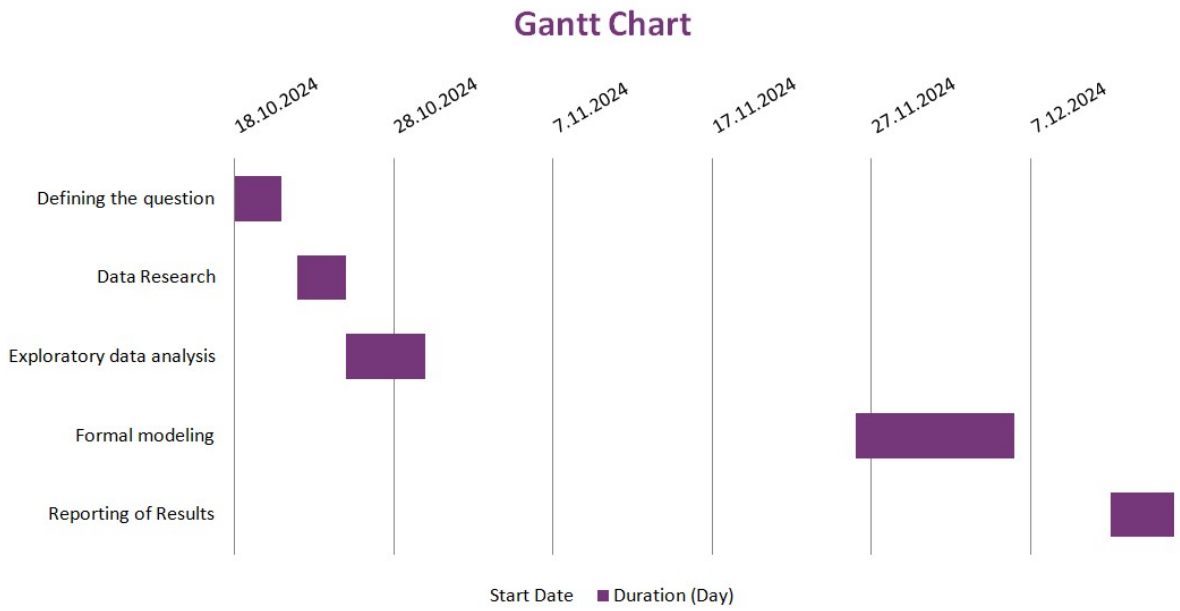
Data Science Project (Report)



1. Project Questions/Problems

Objective: We aim to uncover relationships that can reveal important information about different breeds and possible predictions of cat characteristics.

- **Age and Weight Prediction:** Can we predict the gender of cats based on their age and weight?
- **Colour and Gender Relationship:** Is there a relationship between colour and sex of cats?
- **Breed and Weight Distribution:** How does the weight distribution of different cat breeds differ? will be analysed.

2. Exploratory Data Analysis (EDA)

View the first six rows

```
> head(cats_data)
            Breed Age..Years. Weight..kg.         Color Gender
1     Russian Blue          19           7 Tortoiseshell Female
2 Norwegian Forest          19           9 Tortoiseshell Female
3        Chartreux           3           3         Brown Female
4          Persian          13           6         Sable Female
5          Ragdoll          10           8         Tabby   Male
6           Ocicat           9           8          Blue Female
```

Check the structure of the data

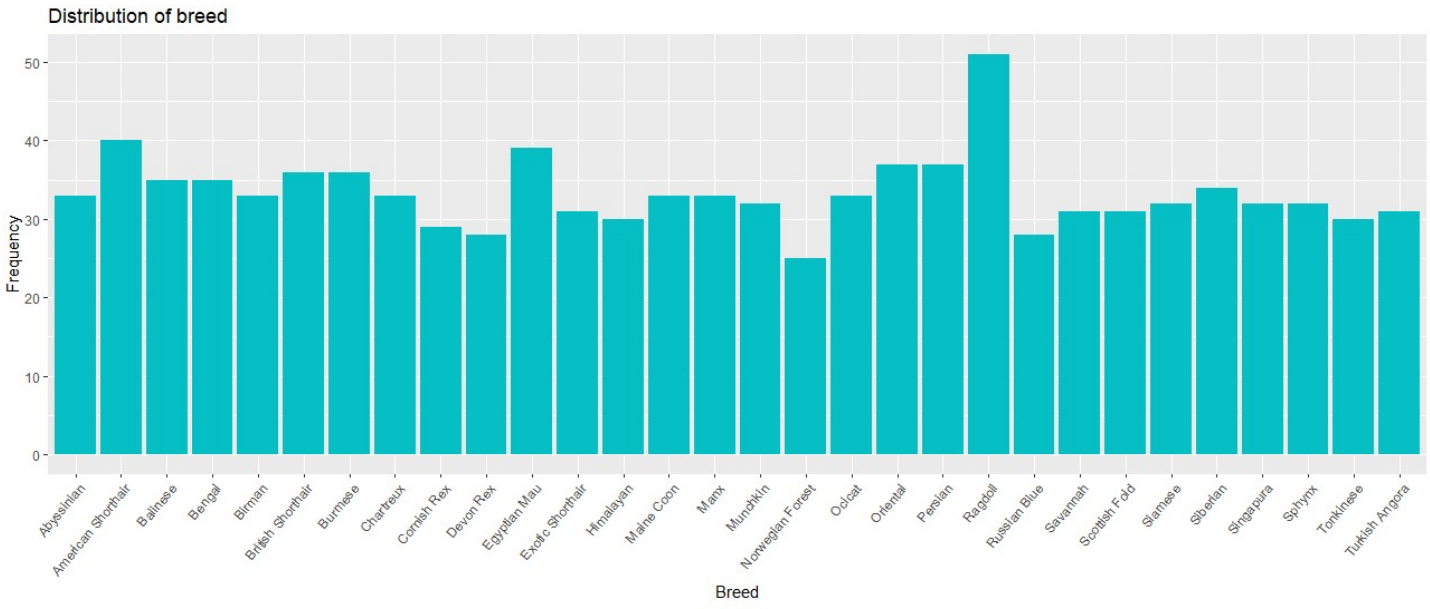
```
> str(cats_data)
'data.frame':   1000 obs. of  5 variables:
 $ Breed      : chr  "Russian Blue" "Norwegian Forest" "Chartreux" "Persian" ...
 $ Age..Years.: int  19 19 3 13 10 9 6 12 2 12 ...
 $ Weight..kg.: int  7 9 3 6 8 8 5 3 7 3 ...
 $ Color      : chr  "Tortoiseshell" "Tortoiseshell" "Brown" "Sable" ...
 $ Gender     : chr  "Female" "Female" "Female" "Female" ...
```

Summary statistics for numeric columns (Age and Weight)

```
> summary(cats_data)
    Breed           Age..Years.      Weight..kg.      Color
 Length:1000        Min.   : 1.00   Min.   :2.00   Length:1000
 Class :character   1st Qu.: 5.00   1st Qu.:4.00   Class :character
 Mode  :character   Median :10.00   Median :6.00   Mode  :character
                    Mean   :10.21   Mean   :5.55
                    3rd Qu.:15.00   3rd Qu.:7.00
                    Max.   :19.00   Max.   :9.00

    Gender
 Length:1000
 Class :character
 Mode  :character
```

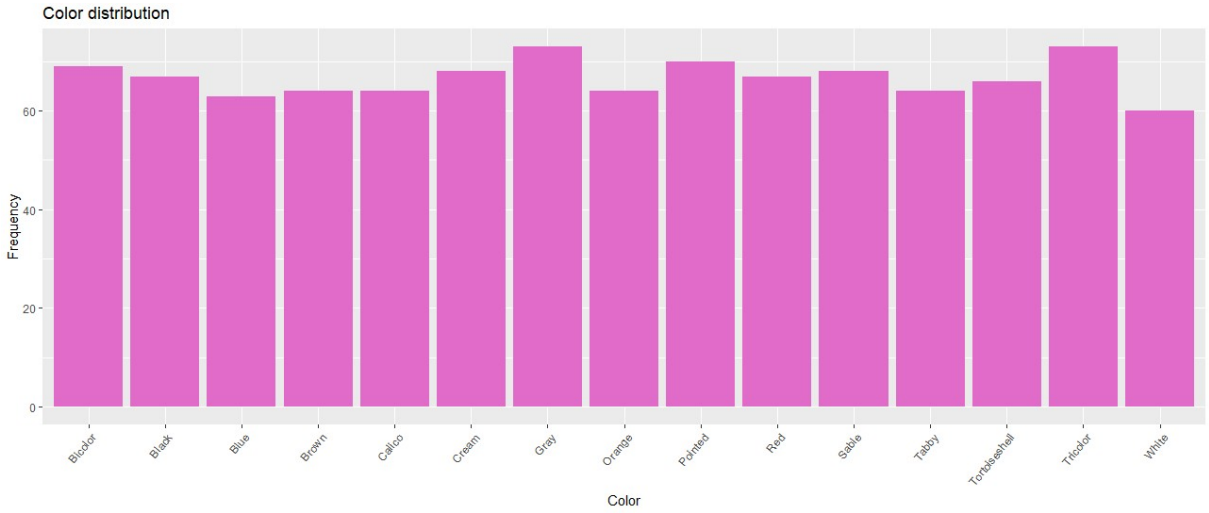
Distribution of breed



There are no major differences between the frequencies of cats by breed. However, some breeds (e.g. Ragdolls) are distinctly different from others. - There are about 30 different

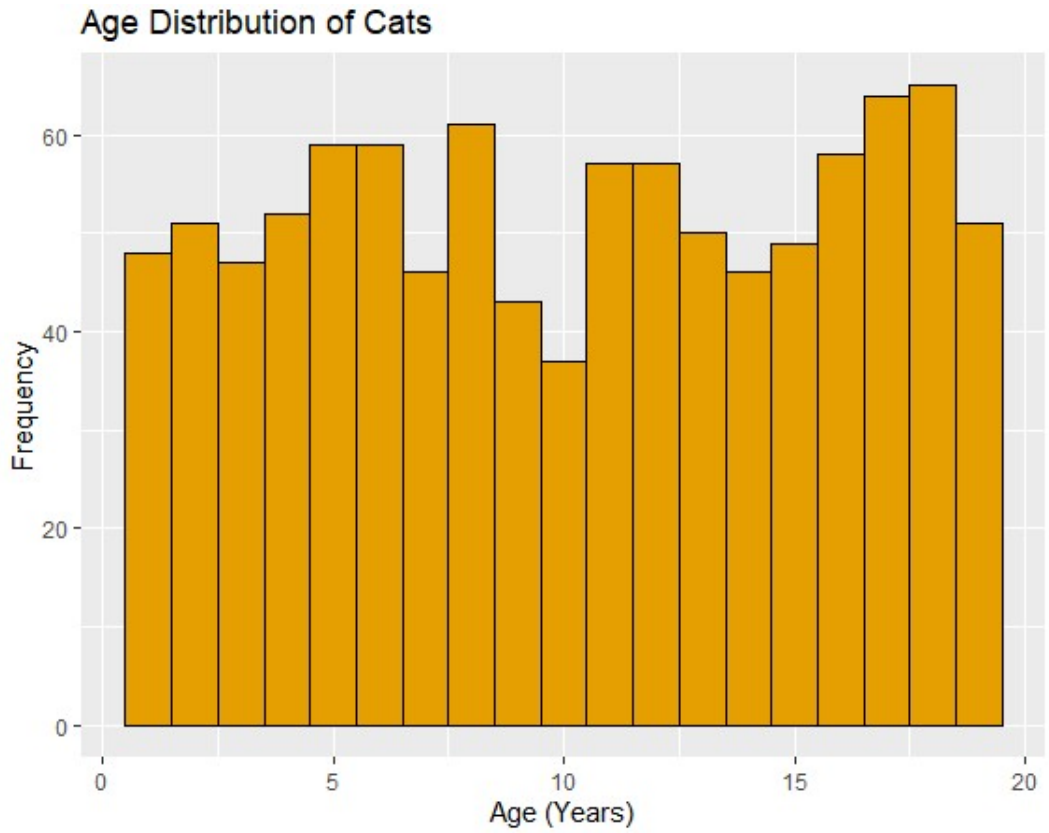
cat breeds in the dataset and the majority of the distribution is concentrated in the 30-40 frequency range.

Color distribution



While the color distribution appears to be fairly balanced, there are some small differences in color preference. However, there is no strong indication of color preference. Gray is slightly more frequently observed color.

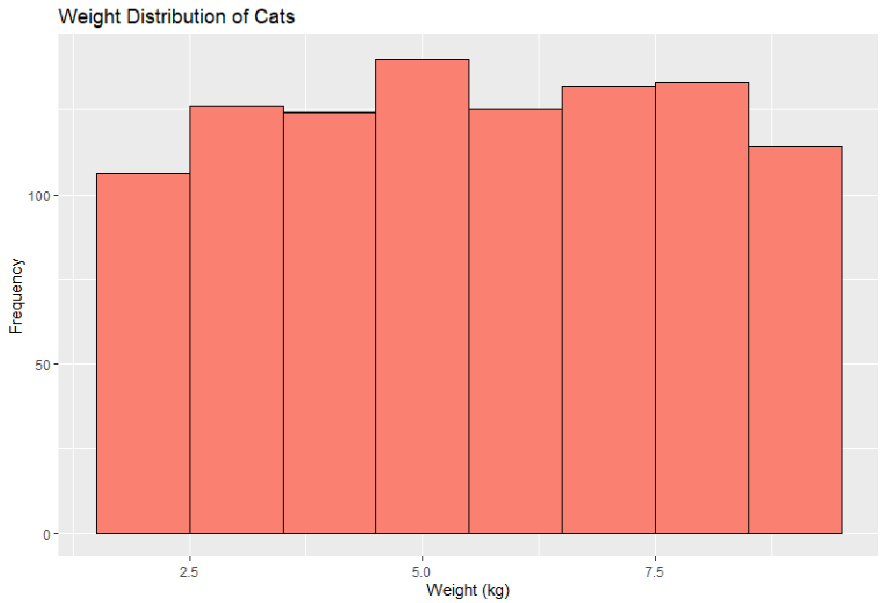
Histogram of Age



Age Distribution:

- The age of cats in the dataset ranges from 1 to 19 years, with an average of about 10 years.
- The age distribution shows variability, but the data is relatively balanced.

Histogram of Weight

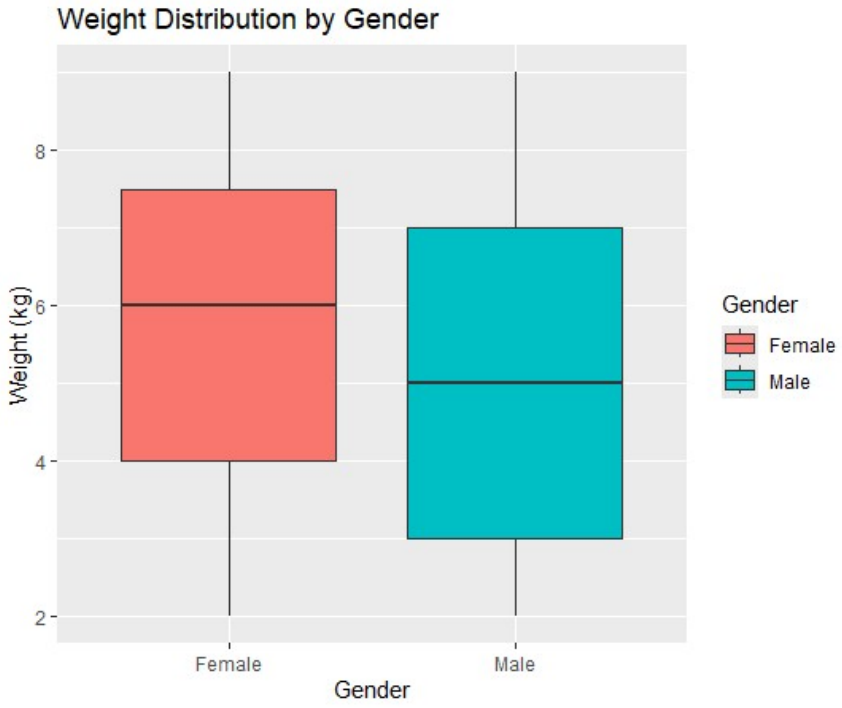


Weight Distribution:

- The weight of cats ranges from 2 to 9 kg, with an average weight of about 5.55 kg.
- The weight distribution is moderately spread, showing a fairly even distribution across different weight categories.

Boxplot for Age by Gender

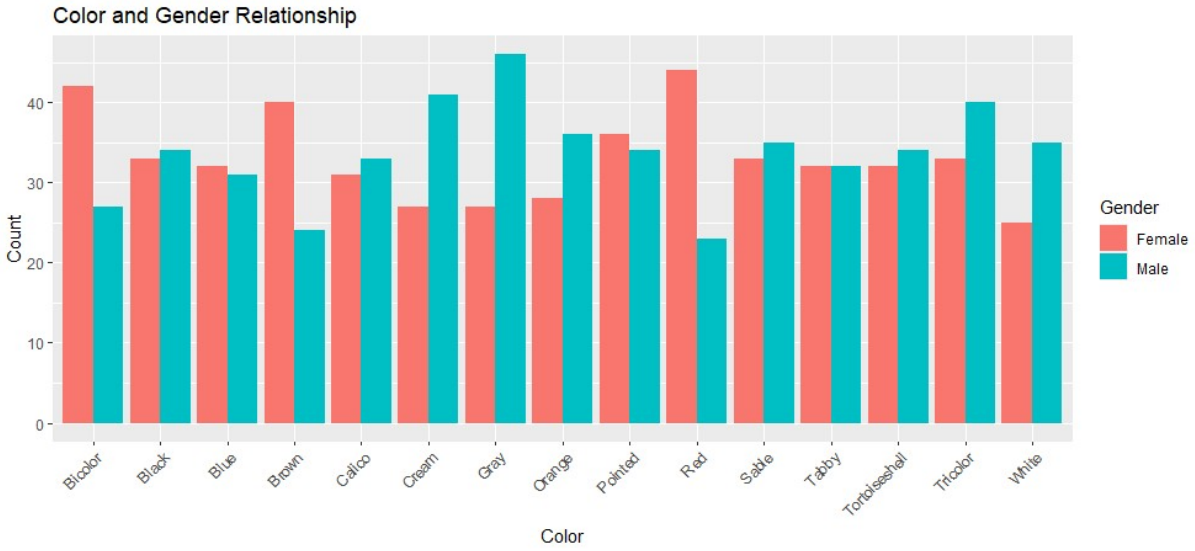
Boxplot for Weight by Gender



Age and Weight by Gender:

- The boxplots reveal that the age and weight distributions are relatively similar across genders, but with some variation in age spread for each gender.
- Both genders show a similar median and spread in weight, though male cats may show slightly higher weight variations.

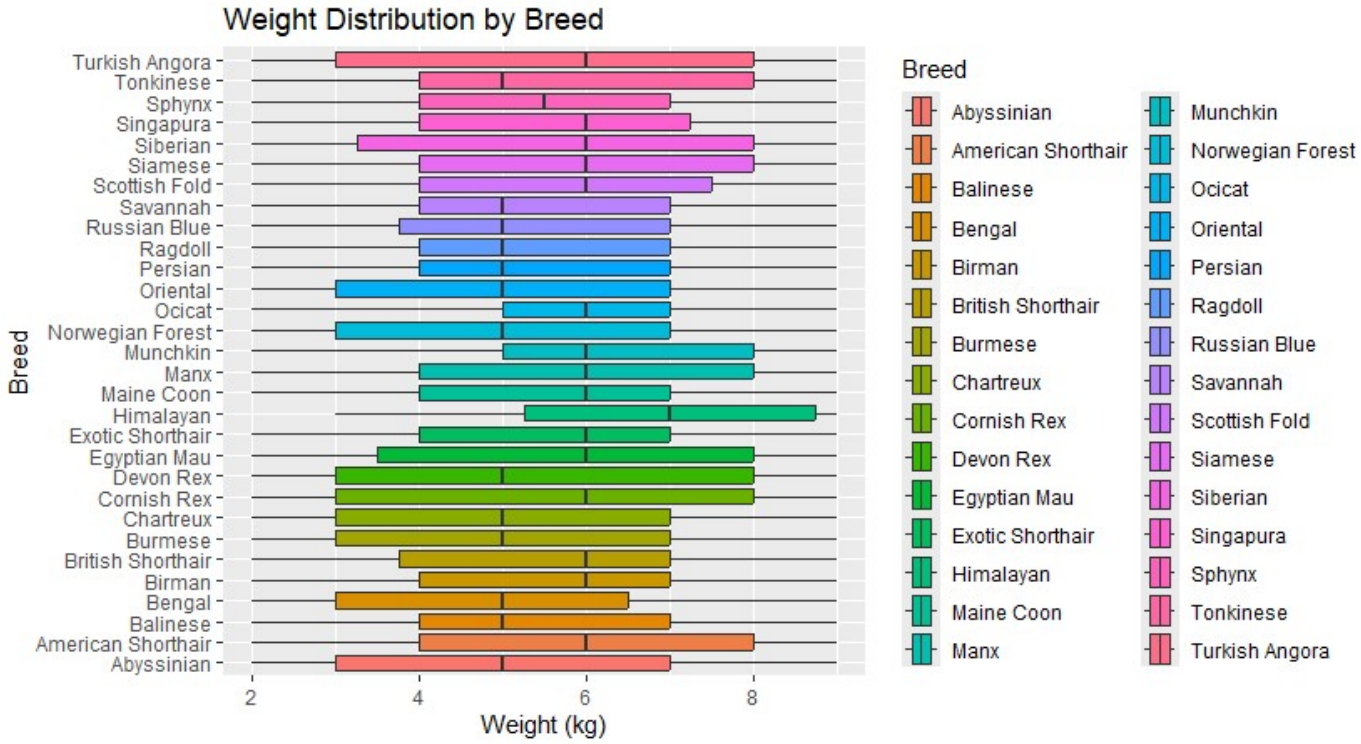
Color and Gender Relationship



- The color distribution across genders appears fairly balanced, with some slight variations in color preference by gender. However, there is no strong indication of a gender-specific color preference.

Weight Distribution by Breed

```
> ggplot(cats_data, aes(y = Breed, x = `weight..kg.`, fill = Breed)) +
+   geom_boxplot() +
+   labs(title = "Weight Distribution by Breed", y = "Breed", x = "Weight (kg)") +
+   theme(axis.text.y = element_text(angle = 0, hjust = 1))
```



- The weight distribution varies across different breeds, with some breeds having wider weight ranges than others.
- Certain breeds have a consistently higher or lower median weight, suggesting breed-specific weight characteristics.

3. Formal Modelling

Can we predict the gender of cats based on their age and weight?

```
> # Logistic regression model
> logistic_model <- glm(gender ~ `age(years)` + `weight(kg)`,
+                       data = cats_dataset,
+                       family = "binomial")
> summary(logistic_model)

Call:
glm(formula = gender ~ `age(years)` + `weight(kg)`, family = "binomial",
    data = cats_dataset)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.27521    0.20783   1.324    0.185
`age(years)` -0.01127    0.01145  -0.984    0.325
`weight(kg)` -0.02525    0.02845  -0.887    0.375

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.2  on 999  degrees of freedom
Residual deviance: 1384.5  on 997  degrees of freedom
AIC: 1390.5

Number of Fisher Scoring iterations: 3
```

Significance: The p-values of both age (age(years)) and weight (weight(kg)) variables are greater than 0.05, therefore age and weight of cats do not show a significant effect in predicting their sex.

Is there a relationship between the color and gender of cats?

```
> # Contingency table for Chi-Square test
> color_gender_table <- table(cats_dataset$color, cats_dataset$gender)
> # Chi-Square test
> chi_test <- chisq.test(color_gender_table)
> chi_test

        Pearson's Chi-squared test

data:  color_gender_table
X-squared = 25.181, df = 14, p-value = 0.03282
```

Additional Analysis

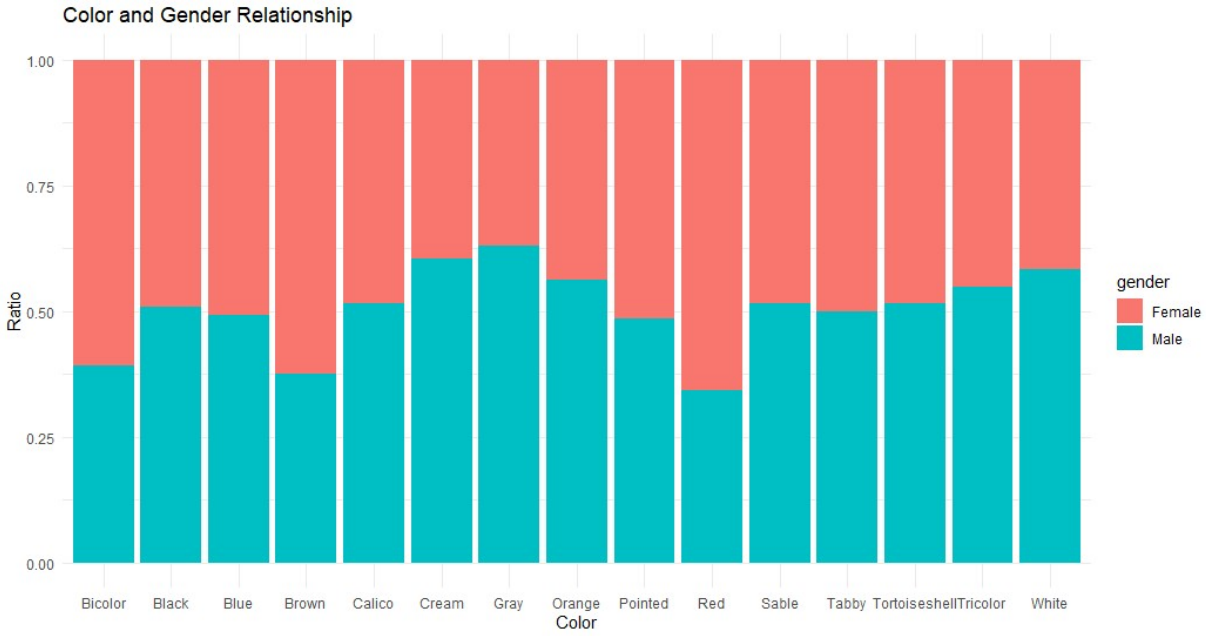
```
> chi_test$residuals

                   Female        Male
Bicolor         1.34235011 -1.32899306
Black          -0.02865127  0.02836618
Blue            0.14594349 -0.14449128
Brown           1.47819163 -1.46348289
Calico         -0.12081374  0.11961158
Cream          -1.14793467  1.13651215
Gray           -1.51965332  1.50453201
Orange         -0.65381553  0.64730974
Pointed         0.22934124 -0.22705918
Red             1.88143368 -1.86271248
Sable          -0.11375929  0.11262733
Tabby           0.05685352 -0.05628780
Tortoiseshell -0.11721960  0.11605321
Tricolor      -0.52152306  0.51633365
White         -0.86242162  0.85384009
```

Since the p-value of this test is 0.03282, at the generally accepted 5% significance level ($\alpha = 0.05$), the following conclusion is reached: H_0 is rejected. It is concluded that there is a significant relationship between the color of the cats and their gender.

Look at the Signs of Residuals

- Positive residuals: More observations compared to expected values.
- Negative residuals: Fewer observations compared to expected values. Example: - Residual 1.881 for "Red - Female": There are many more female cats of the color "Red" than expected.
- Residual for "Cream Female" -1.147: There are fewer she-cats of the color "Cream" than expected.



- "Red" color was significantly more common in female cats (1.881).
- The color "Gray" was found to be significantly more common in male cats (1.504).
- "Cream" color was found to be less common in female cats than expected (-1.147).

How does the weight distribution of different cat breeds differ?

```
> # ANOVA
> anova_model <- aov(`weight(kg)` ~ breed, data = cats_dataset)
> summary(anova_model)
             Df Sum Sq Mean Sq F value Pr(>F)
breed        29    127   4.378    0.88   0.65
Residuals   970   4827   4.976
```

Interpretation

- p-Value: Since p = 0.65, we cannot reject hypothesis H_0. This means that the breed variable (independent variable) has no significant effect on the weight of the cats (dependent variable).
- F-Test Statistic: The F-value (0.88) is low, indicating that the breed variable does not have a very strong effect in explaining the variance in the model.