Title: Exploring Rehoming Times and Characteristics Across Dog Breeds

NAME: SUBRAMANIAN MATHUR SEETHARAMAN

1) Introduction

Rehoming is never easy to understand, and many factors come into play, which depend on breed, age, return, visited and health status. This report tries to conduct an analysis of a dataset that involves dog rehomings with the view to understanding how these factors vary across breeds and affect the time dogs spend in a shelter before finding a new home. This analysis tries to show if there is a breed-specific trend in the duration of re-homing, investigates the key characteristics of sampled dogs, and assesses which statistical model fits best for the prediction of the success in rehoming.

Characteristics:

This data involves three different breeds: **Border Collie**, **Dobermann**, and **Bichon Frise**. Each of them is known for certain special characteristics. Border Collies are very famous for being intelligent and highly energetic dogs, Dobermanns are loyal, strong dogs, and Bichon Frises are basically playful and very affectionate. Analyzing such traits with respect to rehoming data will yield several valuable inferences related to the adoption process.

Objectives:

The report starts with the cleaning of data, removing incomplete records to ensure that the analysis is sound. Further, breed-specific variations are examined, statistical models for rehoming time distributions are proposed, and conclusions on average rehomings are drawn. Finally, it discusses findings, focusing on real world examples and pointing out areas of limitation and future improvement.

2) Results

a) Data Cleaning

The data cleaning process involved removing rows with missing observations for either **rehoming time** (coded as 99999) or **breed** (coded as NA). The summary of the observations removed is as follows:

Table 2.1: Data Cleaning Summary
Summary of rows removed based on missing observations for rehoming time or breed

Reason for removal	Number of Rows Removed	Percentage of Total (%)
Missing rehoming time	9	7.6%
Missing breed	6	4.84%
Total Removed	15	12.10%

b) Data Exploration

The dataset was split into three subgroups based on the dog breed: **Bichon Frise**, **Border Collie**, and **Dobermann**. Below are the numerical summaries and graphical visualizations of key variables, including rehoming time, health, and visit counts, grouped by breed.

Table 2.2: Summary of Rehoming Time by Breed
This table presents the **mean**, **standard deviation**, **median**, and **interquartile range (IQR)** of **rehoming time** for each breed.

Breed	Count (n)	Mean Rehomed (weeks)	SD Rehomed (weeks)	Median Rehomed (weeks)	IQR Rehomed (weeks)
Bichon Frise	10	23.4	12.1	19.5	15.8
Border Collie	78	20.5	11.9	18	17.5
Dobermann	21	17.9	10.5	16	11

Table 2.3: Summary of Visit Counts by Breed
This table shows the **mean**, **standard deviation**, **median**, and **interquartile range (IQR)** of **visit counts** for each breed.

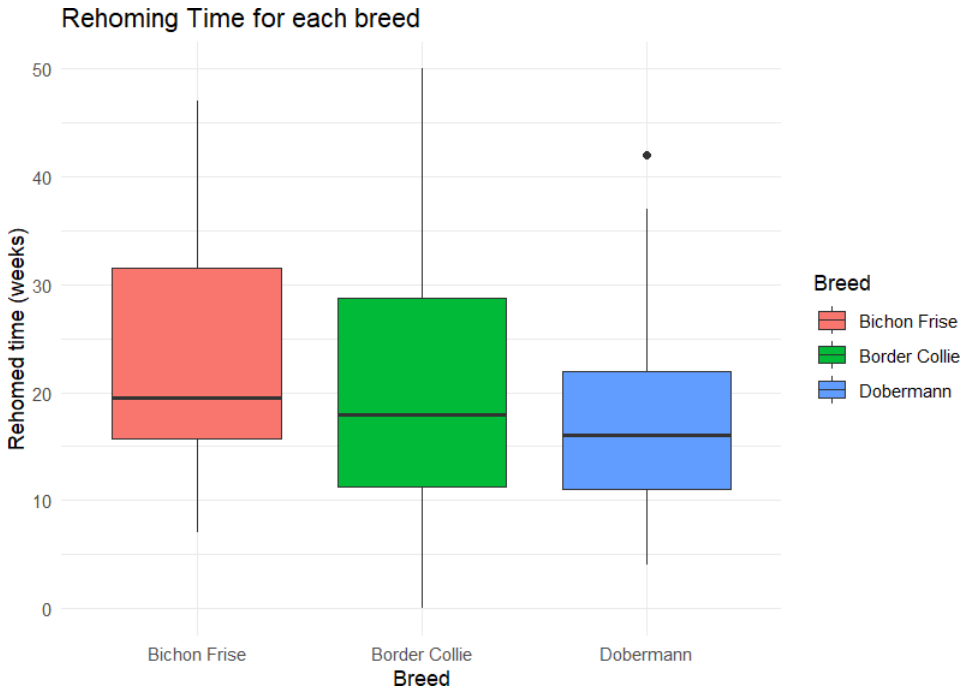
Breed	Count (n)	Mean visits	SD visits	Median visits	IQR visits
Bichon Frise	10	20.5	10.8	17.5	18.5
Border Collie	78	14.1	9.86	10.5	13.8
Dobermann	21	11.4	7.14	10	9

Table 2.4: Summary of Health Scores by Breed
This table summarizes the **mean**, **standard deviation**, **median**, and **interquartile range (IQR)** of the **health scores** for each breed.

Breed	Count (n)	Mean Health	SD Health	Median Health	IQR Health
Bichon Frise	10	49.8	21.1	57.5	17
Border Collie	78	52.9	21.2	54.5	27.2
Dobermann	21	51.3	15.7	53	20

Graphical Summaries of Rehoming Time by Breed

Figure 2.1: Box Plot of Rehoming Time by Breed
The plot is a **box plot** that visualizes the distribution of **rehoming times (in weeks)** for three dog breeds: **Bichon Frise, Border Collie, and Dobermann**. Each breed is represented with a distinct color.



- **Bichon Frise** has consistent rehoming times, with no outliers and a median of 19.5 weeks.
- **Border Collie** displays the widest variability, with rehoming times ranging from 10 to 45 weeks and a median of 18 weeks.
- **Dobermann** has shorter rehoming times overall, with a median of 16 weeks, but includes one outlier above 35 weeks.

c) Modelling and estimation

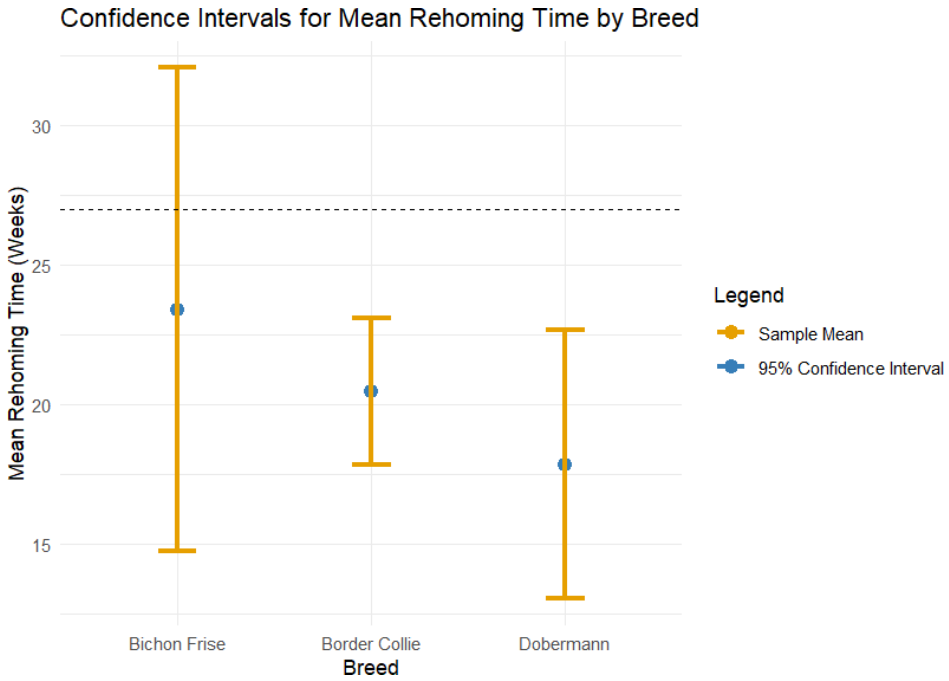
Bichon Frise
Graphical summaries suggest the Normal distribution fits well, judged by alignment in the Q-Q plot, and empirical CDF, while the Exponential does poorly, failing to capture how spread out the data is. The statistical tests confirm this view: the Shapiro-Wilk test does not reject Normality, with p = 0.66, while Kolmogorov-Smirnov rejects the Exponential model, as p = 0.28. **Conclusion:** Normal distribution is likely to be a good model.

Border Collie
The data for Border Collies do not distribute normally or exponentially. Normally distributed Q-Q plot shows evidence of skew; the fitted exponential model does not approximate the empiric CDF. Poorly fitted models are supported by statistically significant results: Shapiro-Wilk p = 0.03; Kolmogorov-Smirnov p = 0.0017. Skewness and multi-modality suggest that neither distribution is appropriate. **Conclusion:** None of the distributions tested fits well.

Dobermann
Q-Q and P-P plots for Dobermanns show that Normal distribution describes these data quite well. As expected, the CDF-and density plots also show reasonable agreement for this model, while poor for the Exponential one. This result is further supported by the statistical tests: a Shapiro-Wilk p value is 0.09, Kolmogorov-Smirnov p equals 0.12. **Conclusion:** The Normal distribution can be an appropriate model.

c) Inference

Figure 2.2: Confidence Intervals for Mean Rehoming Time by Breed
The graph displays the 95% confidence intervals for the mean rehoming times (in weeks) of three dog breeds: Bichon Frise, Border Collie, and Dobermann. The points represent the sample means, while the vertical lines indicate the confidence intervals for each breed. A dashed horizontal line at 27 weeks marks the hypothesized mean for comparison. The color-coded legend distinguishes between the sample mean, confidence intervals, and the reference line, providing a clear visual representation of the analysis.



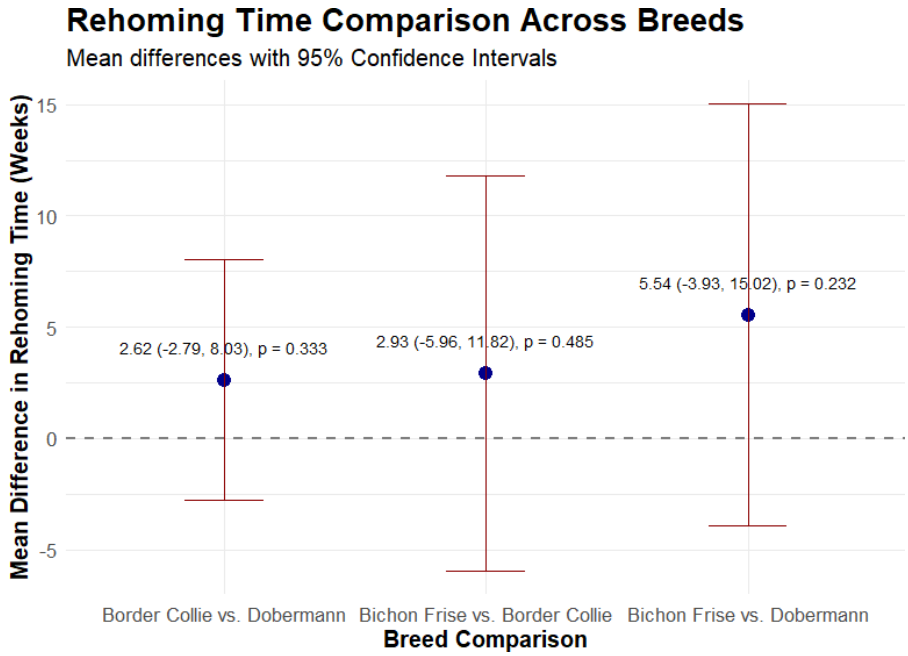
This analysis examined whether the mean rehoming time for three dog breeds (Bichon Frise, Border Collie, and Dobermann) is 27 weeks. A 95% confidence interval was calculated for each breed using the most suitable statistical method based on sample characteristics.

- Bichon Frise: A t-test was used due to the smaller sample size and unknown population variance. The confidence interval was [14.75, 32.05], which includes 27 weeks. This means there is no significant evidence to suggest the mean rehoming time differs from 27 weeks.
- Border Collie: A z-test was applied because of the larger sample size and stable variance. The confidence interval was [17.84, 23.11], which does not include 27 weeks. This indicates the mean rehoming time is significantly shorter than 27 weeks.
- Dobermann: A t-test was used due to the smaller sample size and unknown population variance. The confidence interval was [13.06, 22.66], which also does not include 27 weeks. This suggests the mean rehoming time is significantly shorter than 27 weeks.

Conclusion: The mean rehoming time for Border Collies and Dobermanns is significantly shorter than 27 weeks, while for Bichon Frise, it is not significantly different.
The graph shows the sample means and confidence intervals, with the dashed line representing 27 weeks for comparison.

d) Comparison

Figure 2.3: Rehoming Time Comparison Across Breeds
The plot is a caterpillar plot showing the mean differences in rehoming times (in weeks) between three dog breeds: Bichon Frise, Border Collie, and Dobermann. Each comparison is represented with a blue point for the mean difference, accompanied by red error bars illustrating the 95% confidence intervals. The dashed horizontal line at zero represents no difference in rehoming times. Labels indicate the mean difference, confidence interval, and p-value for each comparison.



Assumptions:

- **Bichon Frise vs. Border Collie**:
  Assumed that rehoming times are independent, and sampling distributions are approximately normal. Welch's t-test was used to account for unequal variances. No significant difference was found ($p = 0.485$).
- **Bichon Frise vs. Dobermann:**
  Assumed independence and normality, similar to the first comparison. Welch's t-test was applied, and no significant difference was observed ($p = 0.232$).
- **Border Collie vs. Dobermann:**
  Assumptions of normality and independence were maintained. The t-test indicated no significant difference in rehoming times ($p = 0.333$).

Conclusions:

- For all pairwise comparisons, confidence intervals included zero, and p-values exceeded 0.05, indicating no significant influence of breed on rehoming time. The accompanying graph visually confirms these findings, as the confidence intervals overlap with zero for all comparisons, further supporting the lack of statistically significant differences.

3) Discussion

Practical significance: Practically, these findings can help shelters prioritize resources. For example, breeds like Bichon Frise, which take longer to rehome, might benefit from targeted adoption campaigns. Understanding why Border Collies and Dobermanns are adopted faster could help improve strategies for other breeds.
Strength: The analysis has several strengths. It appropriately tailored statistical tests to the sample characteristics and provided confidence intervals to highlight the range of plausible values for rehoming times.
Limitations: However, there are limitations. The analysis assumes the samples are representative of the broader population, which may not always be true. It also relies on assumptions about normality and variance that might not fully hold, and external factors affecting rehoming times (e.g., shelter policies or adopter preferences) were not considered.
Future research: should examine other factors influencing rehoming times, such as age or health, and include larger sample sizes to increase reliability. Exploring alternative statistical methods may also help address concerns about data assumptions. Despite these limitations, the analysis provides valuable insights and a basis for further exploration.

4) Appendix

Data cleaning:

```
# Here Data cleaning is done to Remove rows where 'Rehomed' is 99999 or 'Breed' is NA
data_cleaned <- mysample[mysample$Rehomed != 99999 & !is.na(mysample$Breed), ]


# Calculate removals for each condition
removed_rehomed <- nrow(mysample[mysample$Rehomed == 99999, ])
removed_breed <- nrow(mysample[is.na(mysample$Breed), ])
total_removed <- nrow(mysample) - nrow(data_cleaned)

# Percentages of removals
removed_rehomed_pct <- (removed_rehomed / nrow(mysample)) * 100
removed_breed_pct <- (removed_breed / nrow(mysample)) * 100


# Display results for removed rows
cat("Removed due to 'Rehomed' == 99999:", removed_rehomed, "(", removed_rehomed_pct, "%)\n")
cat("Removed due to missing 'Breed':", removed_breed, "(", removed_breed_pct, "%)\n")
cat("Total removed rows:", total_removed, "(", (total_removed / nrow(mysample)) * 100, "%)\n")

# Save the cleaned dataset
save(data_cleaned, file = "data_cleaned.RData")

# Load the cleaned dataset
load("data_cleaned.RData")
```

Data exploration:

```
# Density plot
# a. Bichon Frise
ggplot(Breed_group1, aes(x = Rehomed, color = Breed, fill = Breed)) +
  geom_density(alpha = 0.4) +
  labs(
    title = "Density Plot of Rehoming Time for Bichon Frise ",
    x = "Rehoming Time (week)",
    y = "Density"
  ) +
  theme_minimal()

#b. Border Collie
ggplot(Breed_group2, aes(x = Rehomed, color = Breed, fill = Breed)) +
  geom_density(alpha = 0.4) +
  labs(
    title = "Density Plot of Rehoming Time for Border Collie Breed",
    x = "Rehoming Time (week)",
    y = "Density"
  ) +
  theme_minimal()

#c. Dobermann
```

Parameter Estimation:

```
parameters_by_breed <- lapply(breed_samples, estimate_parameters, column = "Rehomed")

# Display parameter estimates
for (breed in names(parameters_by_breed)) {
  cat("\nBreed:", breed, "\n")
  cat("Normal Distribution Parameters (Mean, SD):",
      paste(names(parameters_by_breed[[breed]]$Normal), parameters_by_breed[[breed]]$Normal, sep = " = ", collapse = ", "), "\n")
  cat("Exponential Distribution Parameter (Rate):",
      paste(names(parameters_by_breed[[breed]]$Exponential), parameters_by_breed[[breed]]$Exponential, sep = " = ", collapse = ",
"), "\n")
}

# Function to perform goodness-of-fit tests for each breed
perform_tests <- function(data, column) {
rehomed_times <- na.omit(data[[column]])
  rehomed_times <- jitter(rehomed_times)  # Add jitter if there are ties

  # Perform Shapiro-Wilk Test (Normality)
  shapiro <- shapiro.test(rehomed_times)

  # Perform Kolmogorov-Smirnov Test (Exponential)
  ks_exp <- ks.test(rehomed_times, "pexp", rate = 1 / mean(rehomed_times))

  # Perform Chi-squared Goodness-of-Fit Test (Normal Distribution)
  breaks <- seq(min(rehomed_times), max(rehomed_times), length.out = 10)
  observed <- hist(rehomed_times, breaks = breaks, plot = FALSE)$counts
  expected <- diff(pnorm(breaks, mean = mean(rehomed_times), sd = sd(rehomed_times))) * length(rehomed_times)

  # Ensure expected frequencies are valid
  expected[expected < 1] <- 1

  chi_sq <- chisq.test(x = observed, p = expected / sum(expected))

  # Return results
  return(list(
    shapiro = shapiro,
    ks_exp  = ks_exp,
    chi_sq = chi_sq
  ))
}

# Apply the function for each breed
results_by_breed <- lapply(breed_samples, perform_tests, column = "Rehomed")
ggplot(Breed_group3, aes(x = Rehomed, color = Breed, fill = Breed)) +
geom_density(alpha = 0.4) +
  labs(
    title = "Density Plot of Rehoming Time for Dobermann Breed",
    x = "Rehoming Time (week)",
    y = "Density"
  ) +
  theme_minimal()
```

Pairwise Comparison:

```
# Split the cleaned data by Breed
breed_samples <- split(data_cleaned, data_cleaned$Breed)

# Function to calculate 95% confidence interval for difference in means
compare_breeds <- function(breed1_data, breed2_data) {
  # Extract rehomed times for both breeds
  rehomed_breed1 <- na.omit(breed1_data$Rehomed)
  rehomed_breed2 <- na.omit(breed2_data$Rehomed)
```

```
  # Perform a two-sample t-test (Welch's t-test by default, assuming unequal variances)
  test_result <- t.test(rehomed_breed1, rehomed_breed2, conf.level = 0.95)

  # Return the test result
  return(data.frame(
    Mean_Breed1 = mean(rehomed_breed1),
    Mean_Breed2 = mean(rehomed_breed2),
    Difference = test_result$estimate[1] - test_result$estimate[2],
    CI_Lower = test_result$conf.int[1],
    CI_Upper = test_result$conf.int[2],
    P_Value = test_result$p.value
  ))
}

# Compare Bichon Frise vs. Border Collie
result_bichon_border <- compare_breeds(breed_samples$`Bichon Frise`, breed_samples$`Border Collie`)

# Compare Bichon Frise vs. Dobermann
result_bichon_dobermann <- compare_breeds(breed_samples$`Bichon Frise`, breed_samples$`Dobermann`)

# Compare Border Collie vs. Dobermann
result_border_dobermann <- compare_breeds(breed_samples$`Border Collie`, breed_samples$`Dobermann`)

comparison_results <- rbind(result_bichon_border, result_bichon_dobermann, result_border_dobermann)
# Combine the results into a data frame for visualization
comparison_results <- data.frame(
  Comparison = c("Bichon Frise vs. Border Collie",
                 "Bichon Frise vs. Dobermann",
                 "Border Collie vs. Dobermann"),
  Mean_Difference = c(result_bichon_border$Difference,
                      result_bichon_dobermann$Difference,
                      result_border_dobermann$Difference),
  CI_Lower = c(result_bichon_border$CI_Lower,
               result_bichon_dobermann$CI_Lower,
               result_border_dobermann$CI_Lower),
  CI_Upper = c(result_bichon_border$CI_Upper,
               result_bichon_dobermann$CI_Upper,
               result_border_dobermann$CI_Upper),
  P_Value = c(result_bichon_border$P_Value,
              result_bichon_dobermann$P_Value,
              result_border_dobermann$P_Value)
)

# Add formatted label for CI and p-value
comparison_results$Label <- paste0(
round(comparison_results$Mean_Difference, 2), " (",
round(comparison_results$CI_Lower, 2), ", ",
round(comparison_results$CI_Upper, 2),
  "), p = ", round(comparison_results$P_Value, 3)
)

library(ggplot2)

# Updated Caterpillar Plot Code with Adjusted Spacing
ggplot(comparison_results, aes(x = reorder(Comparison, Mean_Difference), y = Mean_Difference)) +
  geom_point(size = 3, color = "darkblue") +  # Larger points for emphasis
  geom_errorbar(aes(ymin = CI_Lower, ymax = CI_Upper), width = 0.3, color = "darkred") +  # Thicker error bars
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray50", size = 0.8) +  # Reference line
  geom_text(
    aes(
      label = paste0(
        sprintf("%.2f", Mean_Difference),
        " (", sprintf("%.2f", CI_Lower), ", ", sprintf("%.2f", CI_Upper), "), p = ", sprintf("%.3f", P_Value)
      )
```

```
  ),
  nudge_y = 1.5,  # Adjust text position above the points
  size = 3,  # Adjust text size
  color = "black"
) +  # Proper placement for text
labs(
  title = "Rehoming Time Comparison Across Breeds",
  subtitle = "Mean differences with 95% Confidence Intervals",
  x = "Breed Comparison",
  y = "Mean Difference in Rehoming Time (Weeks)"
) +
theme_minimal() +  # Clean background theme
theme(
  axis.text.x = element_text(vjust = 0.5, hjust = 0.5, size = 10),  # Horizontal labels
  axis.text.y = element_text(size = 10),  # Clear y-axis text
  plot.title = element_text(size = 16, face = "bold"),  # Larger title for impact
  plot.subtitle = element_text(size = 12),  # Subtitle for context
  axis.title.x = element_text(size = 12, face = "bold"),  # Emphasis on axis titles
  axis.title.y = element_text(size = 12, face = "bold"),
  plot.margin = margin(10, 10, 10, 20)  # Add padding for labels
)
```