NAME: NAVEEN SABARINATH BABU
STUDENT ID: 201783589

A COMPREHENSIVE STUDY OF REHOMING DURATION IN DOG BREEDS

1.INTRODUCTION:

This report explains about dog rehoming, examining the process for three sample breeds: Mixed breed, Labrador Retriever, and Border Collie. Our analysis focuses on rehoming durations that exceed 27 weeks for each breed, with the aim of comparing the rehoming periods for each breeds.

The data set [mysample] provides comprehensive details on rehoming dogs, including information on the number of visits needed to rehome each dog, the time taken to rehome them, their health status as a percentage, breed names, age, reasons for being placed in rehoming, and whether they have been returned in the past, denoted by a yes or no response.

2 DATA CLEANING:

In the initial data exploration, we cleaned data for better analysis and removed unwanted data. The rehoming variable contains the negative value and outlier such as "99999", and the breeds variable contains "NA" as a description of the breed for dogs. Cleaning the data makes it more reliable for analysis.

After cleaning, the dimensions of the sample dataset were reduced from 864 observations to 849 observations. This leads to the removal of 15 values, which contains nine instances of rehoming and six from the breed variable. This process results in a data loss of 1.7% and helps to derive factors influencing dog rehoming on the analysis.

3 DATA EXPLORATION:

TOTAL DATA SUMMARY	VISITED	REHOMED	HEALTH (%)
MEAN	14.40	19.17	52.66
MEDIAN	13	18	54
STANDAR DEVIATION	8.65	9.78	15.12

Table 1 (*Summary of sample data*)

From the given sample data, we have explored that, on average, 19.2 weeks are taken to rehome dogs. Furthermore, the overall health status of dogs in the dataset is around 52.7%, with only 57 counts of dogs being returned. The sample data are split based on the dog breeds in sample data sets to explore more.

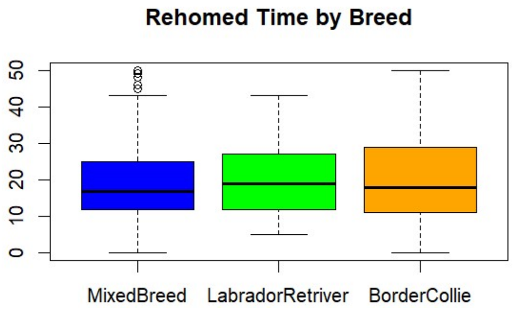
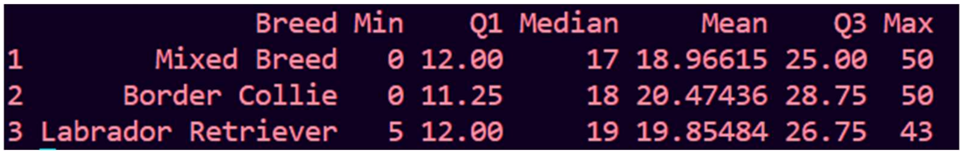


Figure 1 (*Box Plot of rehoming time for each breed*)

After splitting the data based on breeds, a box plot is plotted in Figure 1 to illustrate the rehoming periods. The plot explains that the rehoming duration for each breed falls within the range of 10 to 25 weeks. It also explains each median (represented as the centre line) of the corresponding breeds as 17,20,19. We can also notice the outer layer in mixed breeds away from the rage, stating that the rehoming period of dogs takes more than 40 weeks for few dogs.



```
                Breed Min    Q1 Median     Mean    Q3 Max
1         Mixed Breed   0 12.00     17 18.96615 25.00  50
2       Border Collie   0 11.25     18 20.47436 28.75  50
3 Labrador Retriever    5 12.00     19 19.85484 26.75  43
```

Figure 2 (*Summary Table for rehoming each breed*)

This statistics summary as shown in Figure 2 provides insights into each breed category's rehoming distribution. For this, border collie has a higher average of 20 weeks, suggesting it takes longer period to rehome than other breeds. On other hand mixed breed averages around 18, which takes less duration to be rehomed than other breeds.

4.MODELLING:

To understand the distribution of rehoming periods of different breeds in the data set, we use two essential techniques and test, such as histogram and QQ-plot, to visually understand the data distribution in each breed of the data set.

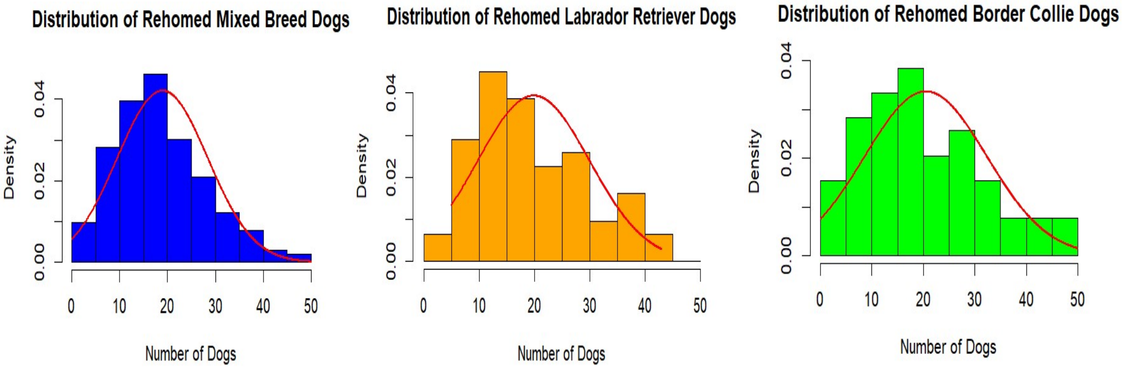


Figure 3 (*Histogram of rehoming period for each breed showing its distribution*)

From the histogram, we can see that the tail of the distribution is stretched towards the right for each breed in the plot, which is a characteristic of positive skewness. As an initial overview of the data, we can see the difference in the variance. In mixed breeds, it is lesser compared to other breeds. Additional statistical tests and techniques are used to assess the distribution.

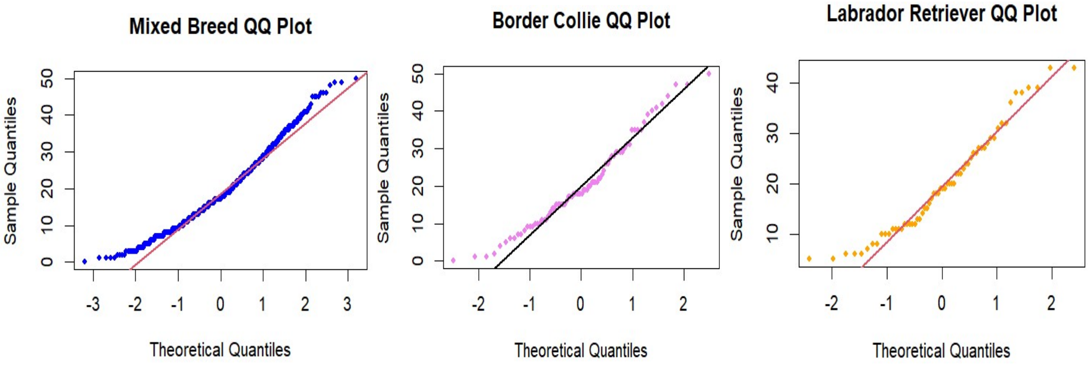


Figure 4 (*QQ-plot of rehoming period for each breed showing the distribution***)**

As shown in figure 4 QQ-plots suggest that the Mixed Breed dataset approximates normality with some deviations in the tails and suggest as skewed distribution. On the other hand, the Border Collie data shows a strong alignment with normal distribution. Additionally, the rehoming times for Labrador Retrievers exhibit a good alignment with normality, which support normal distribution assumption.

5.TESTING:

We used different statistical tests to analyse the data samples of different dog breeds. Specifically, we utilized the Kolmogorov-Smirnov test for smaller datasets and the Shapiro test for larger ones. The purpose was to determine the normal distribution of each sample. The Labrador Retriever and Border Collie datasets were smaller, while the Mixed breed dataset was larger.

BREED	TEST	P-VALUE
Mixed Breed	**Kolmogorov-Smirnov (KS)**	2.304e-05
Labrador Retriever	**Shapiro-Wilk**	0.01219
Border Collie	**Shapiro-Wilk**	0.03056

Table 2 (*Checking normal distribution using Kolmogorov-Smirnov (KS)and Shapiro-Wilk***)**

As shown in Table 2, the P-values for each breed are below the 5% (0.05) threshold. This suggests that the sample data does not confirm the normal distribution assumption. Notably, labrador retriever and border collie has closest value to be a normal distribution, leading to the interpretation that the data does not necessarily follow a normal distribution.

6.INFERENCE & COMPARISONS:

After estimating the distribution of the rehoming period for dogs, our main objective is to determine whether the time exceeds 27 weeks. We plan to achieve this by checking the confidence interval using the z-test. We have chosen the z-test over the t-test for all three sample breeds for two primary reasons. Firstly, the size of the dataset for each breed is larger than 30, making the z-test more appropriate. Secondly, we have access to both the population mean and standard deviation, which makes the z-test more efficient. These factors combined make the z-test the preferred statistical method for our analysis.

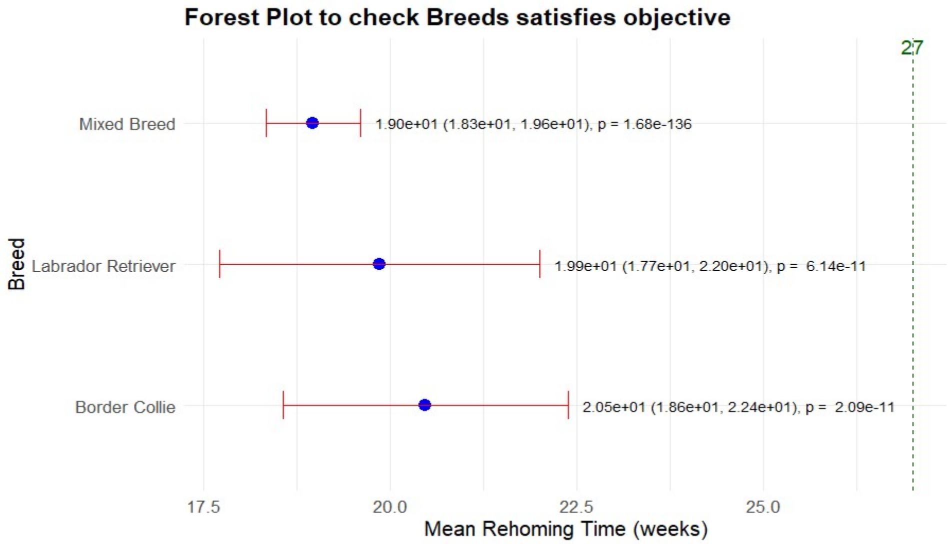


Figure 9(*Forest plot to show comparison using z-test to check whether the CI is above 27***)**

In figure 9 each breed category significantly falls below 27 weeks as indicated by green line. notably, all breeds fall below the range of 24 weeks in the plot where the labrador retriever is the closest value near 27 weeks. This implies that all rehoming periods for breeds lies between the range of 17 to 24 weeks.

BREED PAIR	CONFIDENCE INTERVAL	P-VALUE
Mixed Breed vs Border Collie	(-4.273, 1.256)	0.282
Mixed Breed vs Labrador Retriever	(-3.555, 1.777)	0.509
Border Collie vs Labrador Retriever	(-3.071, 4.302)	0.740

Table 3(*Breeds Comparison using t-test***)**

The analysis compared rehoming times for Mixed Breed, Border Collie, and Labrador Retriever. The confidence intervals for differences in mean rehoming weeks include zero for all breed pairs (Mixed Breed vs. Border Collie, Mixed Breed vs. Labrador Retriever, Border Collie vs. Labrador Retriever), indicating no statistically significant differences. P-values further support this, all being greater than 0.05. Overall, the results suggest that rehoming times are similar across these breeds, with some variability.

7.DISCUSSION:

The analysis conducted showed the real-life implications of each breed's rehoming period. It can accessing them based on the rehoming period of breeds and prioritizing the dogs breeds based on that, which helps rehoming management to shorten the rehoming process and select specific breeds with a shorter period of rehoming. However, there were some limitations to the analysis. The sample data for each breed were uneven, which may lead to a biased interpretation of the conclusions. Also, the age of dogs was measured in only three different groups, which may lead to confusion in measuring their exact age. Additionally, more external factors that correlate with the rehoming of dogs need to be considered.

The main limitation affecting the current analysis is the data size, where two categories of data are smaller than one breed, which may lead to a different interpretation than the actual one. Checking for a module is tricky as there are different ways to explore the distribution, and assumptions can cause potential flaws in the analysis.

On the other hand, the analysis provides a strong foundation for understanding rehoming among dog breeds, which helps manage the resources for dogs rehoming. As a future factor, we can analyse the health condition's effectiveness in rehoming for different breeds.

8.CONCLUSION:

Therefore, based on this analysis, we do not find substantial evidence to support the notion that rehoming weeks for these breeds consistently below 27 weeks. The average weeks for rehoming the 3 dogs breeds where with in the rage of 17 to 24 weeks and we conclude that different breed dogs have similar rehoming time .

9.APPENDEX:

```
load("rehoming.RData")
createsample(201783589)
save(mysample,file="mysample.RData")
load("mysample.RData")

#creating a data frame
data = data.frame(mysample)

#Data cleaning
data =data[data$Rehomed !=99999 & data$Rehomed>=0,]
nadata=is.na(data$Breed)
data =data[nadata=="FALSE",]
dim(data)

#splitting the data based on breeds
mixedbreed  = subset(data,Breed == "Mixed Breed")
bordercollie = subset(data, Breed == "Border Collie")
labradorretriver = subset(data,Breed == "Labrador Retriever")

#mean median and standard deviation
mean(data$Visited)
mean(data$Rehomed)
mean(data$Health)

median(data$Visited)
median(data$Rehomed)
median(data$Health)

sd(data$Visited)
sd(data$Rehomed)
sd(data$Health)

#box plot for each breeds
ggplot2::ggplot(data, ggplot2::aes(x = Breed, y = Rehomed, fill = Breed)) +
  ggplot2::geom_boxplot() +
  ggplot2::labs(title = "Rehomed Time by Breed", x = "Breed", y = "Rehomed Time") +
  ggplot2::scale_fill_manual(values = c("blue", "green", "orange")) +
  ggplot2::theme_minimal()
boxplot(list(MixedBreed=mixedbreed$Rehomed, LabradorRetriver=labradorretriver$Rehomed,
BorderCollie=bordercollie$Rehomed), main="Rehomed Time by Breed", col=c("blue", "green",
"orange"),names=c("MixedBreed", "LabradorRetriver", "BorderCollie"))

# Assuming mixedbreed, bordercollie, and labradorretriever are your data frames
summary_mixedbreed <- summary(mixedbreed$Rehomed)
summary_bordercollie <- summary(bordercollie$Rehomed)
summary_labradorretriever <- summary(labradorretriver$Rehomed)

# Create a data frame
summary_table <- data.frame(
  Breed = c("Mixed Breed", "Border Collie", "Labrador Retriever"),
  Min = c(summary_mixedbreed[1], summary_bordercollie[1], summary_labradorretriever[1]),
  Q1 = c(summary_mixedbreed[2], summary_bordercollie[2], summary_labradorretriever[2]),
  Median = c(summary_mixedbreed[3], summary_bordercollie[3],
summary_labradorretriever[3]),
  Mean = c(summary_mixedbreed[4], summary_bordercollie[4], summary_labradorretriever[4]),
  Q3 = c(summary_mixedbreed[5], summary_bordercollie[5], summary_labradorretriever[5]),
  Max = c(summary_mixedbreed[6], summary_bordercollie[6], summary_labradorretriever[6])
)

# Print the summary table
print(summary_table)

#model fitting
#Mixeed Breed
```

```
mum <- mean(mixedbreed$Rehomed)
sigmam <- sd(mixedbreed$Rehomed)
hist(mixedbreed$Rehomed, breaks=seq(from=0, to=50, by=5), col = "blue", freq=FALSE,
     main="Distribution of Rehomed Mixed Breed Dogs", xlab="Number of Dogs",
ylab="Density")
x <- seq(from=min(mixedbreed$Rehomed),to=max(mixedbreed$Rehomed), length.out = 1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y, col="red", lwd=2)

#bordercollie
mub <- mean(bordercollie$Rehomed)
sigmab <- sd(bordercollie$Rehomed)
hist(bordercollie$Rehomed, breaks=seq(from=0, to=50, by=5), col = "green", freq=FALSE,
     main="Distribution of Rehomed Border Collie Dogs", xlab="Number of Dogs",
ylab="Density")
x <- seq(from=min(bordercollie$Rehomed),to=max(bordercollie$Rehomed), length.out = 1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y, col="red", lwd=2)

#labradorretriver
mul <- mean(labradorretriver$Rehomed)
sigmal <- sd(labradorretriver$Rehomed)
hist(labradorretriver$Rehomed, breaks=seq(from=0, to=50, by=5), col = "orange",
freq=FALSE,
     main="Distribution of Rehomed Labrador Retriever Dogs", xlab="Number of Dogs",
ylab="Density")
x <- seq(from=min(labradorretriver$Rehomed),to=max(labradorretriver$Rehomed), length.out =
1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y, col="red", lwd=2)

#QQ-plot
qqnorm(mixedbreed$Rehomed, main = "Mixed Breed QQ Plot", xlab = "Theoretical Quantiles",
ylab = "Sample Quantiles", pch = 20, col = "blue")
qqline(mixedbreed$Rehomed, col=2,lwd=2)

qqnorm(bordercollie$Rehomed, main = "Border Collie QQ Plot", xlab = "Theoretical
Quantiles", ylab = "Sample Quantiles", pch = 20, col = "violet")
qqline(bordercollie$Rehomed, col="black",lwd=2)

qqnorm(labradorretriver$Rehomed[labradorretriver$Breed == "Labrador Retriever"], main =
"Labrador Retriever QQ Plot", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
pch = 20, col = "orange")
qqline(labradorretriver$Rehomed[labradorretriver$Breed == "Labrador Retriever"],
col=2,lwd=2)


#ks test for normality
ks.test(x=mixedbreed$Rehomed,y="pnorm",mean=mum,sd=sigmam)
ks.test(x=labradorretriver$Rehomed,y="pnorm",mean=mul,sd=sigmal)
ks.test(x=bordercollie$Rehomed,y="pnorm",mean=mub,sd=sigmab)


#shapiro_wilk test
shapiro.test(mixedbreed$Rehomed)
shapiro.test(labradorretriver$Rehomed)
shapiro.test(bordercollie$Rehomed)


#z-testing
library(BSDA)
library(ggplot2)

# Perform z-tests
forest_data <- data.frame(
  Breed = c("Mixed Breed", "Border Collie", "Labrador Retriever"),
  Estimate = c(result_mixed$estimate, result_border$estimate, result_labrador$estimate),
  Lower = c(result_mixed$conf.int[1], result_border$conf.int[1],
result_labrador$conf.int[1]),
  Upper = c(result_mixed$conf.int[2], result_border$conf.int[2],
result_labrador$conf.int[2]),
  p_value = c(result_mixed$p.value, result_border$p.value, result_labrador$p.value)
)

forest_data$Estimate <- as.numeric(forest_data$Estimate)
forest_data$Lower <- as.numeric(forest_data$Lower)
forest_data$Upper <- as.numeric(forest_data$Upper)
forest_data$Breed <- as.factor(forest_data$Breed)

ggplot(forest_data, aes(x = Estimate, y = Breed)) +
  geom_point(size = 3, color = "blue") +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.2, color = "red") +
  geom_vline(xintercept = 27, linetype = "dashed", color = "darkgreen") +
  annotate("text", x = 27, y = Inf, vjust = 1, hjust = 0.5, label = "27", color =
"darkgreen", size = 4) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  ) +
  annotate("text", x = forest_data$Upper + 0.2, y = as.numeric(forest_data$Breed),
           label = paste0(format(forest_data$Estimate, scientific = TRUE, digits = 3), "
(",
                          format(forest_data$Lower, scientific = TRUE, digits = 3), ", ",
                          format(forest_data$Upper, scientific = TRUE, digits = 3),
                          "), p = ", format(forest_data$p_value, scientific = TRUE, digits
= 3)),
           hjust = 0, size = 3) +
  labs(title = "Forest Plot to check Breeds satisfies objective", x = "Mean Rehoming Time
(weeks)", y = "Breed")


# Perform t-tests for comapring breeds
ttest1 <- t.test(mixedbreed$Rehomed, labradorretriver$Rehomed)
print(ttest1)
ttest2 <- t.test(bordercollie$Rehomed, labradorretriver$Rehomed)
print(ttest2)
ttest3 <- t.test(mixedbreed$Rehomed, bordercollie$Rehomed)
print(ttest3)
```

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity for Individual Pieces of Work

I declare that I am aware that as a member of the University community at the University of Leeds I have committed to working with Academic Integrity and that this means that my work must be a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine.

I declare that the attached submission is my own work.

Where the work of others has contributed to my work, I have given full acknowledgement using the appropriate referencing conventions for my programme of study.

I confirm that the attached submission has not been submitted for marks or credits in a different module or for a different qualification or completed prior to entry to the University.

I have read and understood the University's rules on Academic Misconduct. I know that if I commit an academic misconduct offence there can be serious disciplinary consequences.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties to verify that this is my own work, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and I wish to have taken into account.

Student Signature:

Student Number: 201783589

Student Name: Naveen Sabarinath Babu

Date: 14 Dec 2023

Please note:

When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:

"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.

I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity.'"