The average time taken to rehome dogs of different breeds

Introduction

This report examines the average time it takes to rehome dogs at animal shelters using a dataset that includes information on dogs with three different breeds: Mixed Breed, Shih Tzu, and Staffordshire Bull Terrier. While the author of the research claims that the average time to rehome a dog is around 27 weeks for all breeds, other researchers suggest that it depends heavily on the breed of dog. The report will investigate whether the variable 'Rehomed', which represents the number of weeks from the dog's arrival at the shelter until being rehomed, supports the claim that the average time to rehome dogs varies based on breed.

Results

First, the data needs to undergo data cleaning, as it contains some missing values. Specifically, there are six missing values in 'Breed', marked as 'NA', and nine missing values in 'Rehomed', marked as '99999'. Since the percentage of missing values is only 1.57%, all rows containing missing values are removed from the dataset.

To analyse the average time taken to rehome for each breed, the dataset is divided into three separate samples based on breed. Each sample has a different number of data, with 710 for Mixed Breed, 24 for Shih Tzu, and 207 for Staffordshire Bull Terrier. Significant characteristics are shown by analysing each sample. As the sample size for Shih Tzu is significantly small compared to the others, most variables have different distributions. For instance, the distributions of categorical variables such as 'Age' and 'Reason' of Shih Tzu differ from the others, as shown in Table 1.

Table 1 Frequency counts of the variable 'Reason' for each sample

	Dangerous	Health condition	Neglect	Stray	Unwanted
Mixed Breed	5	6	433	234	30
Shih Tzu	-	-	21	1	2
Staffordshire Bull Terrier	3	2	92	97	13

Additionally, it can be observed from Table 2 that the ranges of the numerical variables 'Visited' and 'Health' are much smaller than those of the other two samples.

Table 2 Numerical summaries of the variable 'Visited' for each sample

	Min	1st Quartile	Median	Mean	3rd Quartile	Max

Mixed Breed	0	45	54	52.56	61	100
Shih Tzu	9	46.5	56	50.58	59.5	67
Staffordshire Bull Terrier	3	47	56	54.69	64	89

After further analysis using the variable 'Rehomed', some notable characteristics are discovered. Table 3 demonstrates a similar pattern as before, with a smaller range of 'Rehomed' and variance. Furthermore, Figure 1 illustrates significant differences in the distributions between Mixed Breed and Staffordshire Bull Terrier, and Shih Tzu. By visualising the distributions, it is found that the samples of Mixed Breed and Staffordshire Bull Terrier are normally distributed with bell-shaped distributions, whereas the distribution of Shih Tzu does not seem to be normally distributed.

Table 3 Numerical summaries of the variable 'Rehomed' for each sample

	Mean	Variance	Standard deviation	Skewness
Mixed Breed	18.9380	90.5124	9.5138	0.6166
Shih Tzu	19.5000	77.9130	8.8268	0.7666
Staffordshire Bull Terrier	19.3382	111.5938	10.5638	0.7270

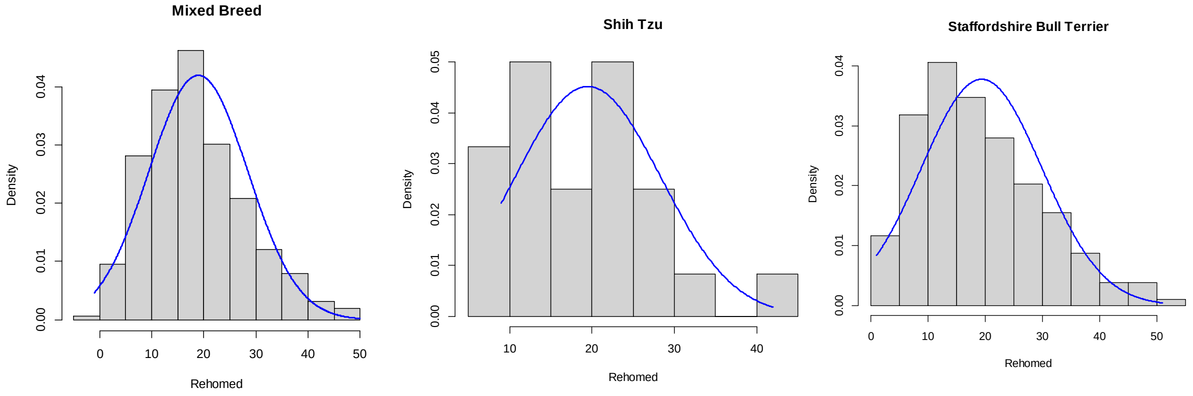

Figure 1 Distributions of the variable 'Rehomed' for each sample

It is necessary to determine further whether the samples can be considered normally distributed. One way to do this is by examining the Q-Q plot, which shows how closely the data points follow a straight line. In Figure 2, the Q-Q plots for Mixed Breed and Staffordshire Bull Terrier appear to have nearly straight lines, indicating that the samples may be normally distributed. Another method to verify the normal distribution is by comparing the Fn(x) and G(x) values and checking if they are close to each other. For instance, by calculating Fn(30) and G(30) for both breeds, it can be concluded that both samples are normally distributed, as shown in Table 4. Visualising the cumulative distribution function in Figure 3 also suggests that the samples are normally distributed.

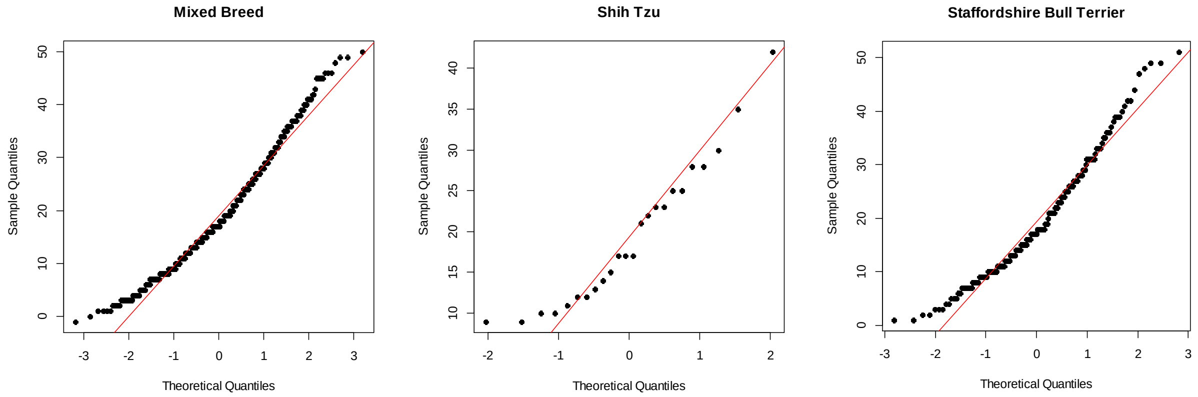

Figure 2 Q-Q plots of the variable 'Rehomed' for each sample

Table 4 Values for Fn(30) and G(30) for each sample

	Fn(30)	G(30)
Mixed Breed	0.8746	0.8775
Shih Tzu	0.9167	0.8436
Staffordshire Bull Terrier	0.8357	0.8436

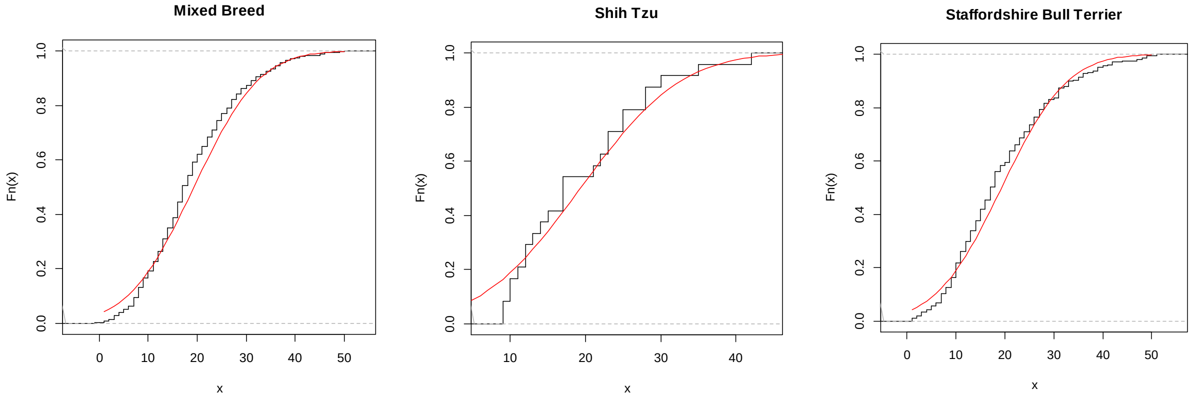

Figure 3 CDF visualizations for each sample

The analysis of the samples indicates that the Mixed Breed and Staffordshire Bull Terrier samples have normal distributions. However, the sample of Shih Tzu is not normally distributed. This can be observed by the lack of a bell-shaped distribution, the non-linear Q-Q plot, and the significant difference between the values of Fn(x) and G(x) in the CDF. Based on these findings, it is clear that the sample of Shih Tzu does not follow a normal distribution.

After confirming that the samples of Mixed Breed and Staffordshire Bull Terrier are normally distributed, the parameters of each model can be estimated. A normal distribution has two parameters, which are the mean and variance. In Figure 4, two estimators of each mean and variance are examined to determine which one to use. By comparing estimators between the sample mean and the 50% quantile, it is clear that the sample mean should be selected since the second estimator appears to be biased. Moreover, when estimating the variance between the sample variance and the maximum likelihood estimator, the second estimator is biased, so the sample variance is the appropriate choice. The estimators for both samples can be found in Table 3.

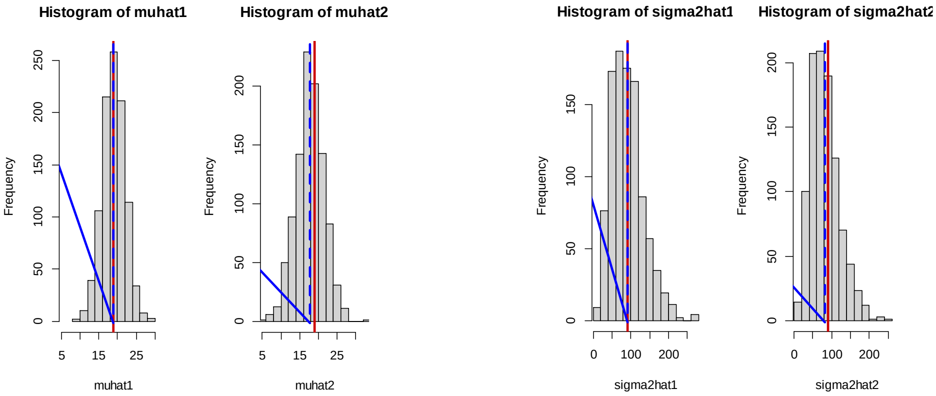

Figure 4 Histograms to estimate the mean and variance

To examine whether the author's claim is true, based on the analysis of the samples, the confidence intervals to test whether the average rehoming time is 27 weeks can be calculated. As the samples of Mixed Breed and Staffordshire Bull Terrier are found to be normally distributed, the z-test can be used to find the confidence intervals. On the other hand, for the sample of Shih Tzu, the t-test can be used instead, as it is not normally distributed, and the t-test is based on quantiles. Figure 5 visualises the 95% confidence interval for each sample. This shows that any of the three samples is evident for a statement that the average time taken to rehome is 27 weeks for all breeds, as none of the confidence intervals consists of 27.

In order to test the author's initial claim, the confidence intervals need to be calculated to determine if the average rehoming time is 27 weeks. The Mixed Breed and Staffordshire Bull Terrier samples are normally distributed, so the z-test can be used to find the confidence intervals. However, the sample for Shih Tzu is not normally distributed, so the t-test can be used instead as it is based on quantiles. Figure 5 displays the 95% confidence interval for each sample. The results show that none of the confidence intervals include the value of 27 with the far left intervals, which means that the claim that the average time taken to rehome is 27 weeks for all breeds cannot be confirmed.

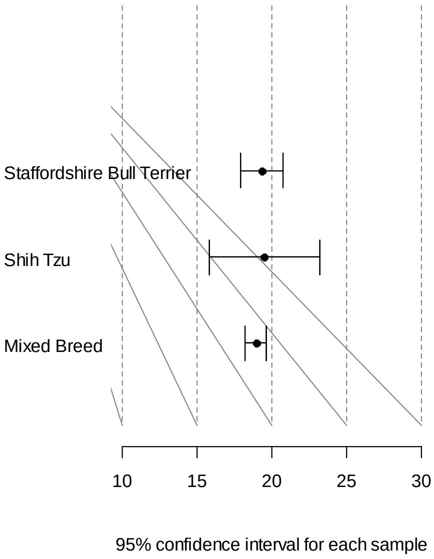

Figure 5 95% confidence interval for each sample

For further analysis, the confidence interval for each pair of samples can be calculated to determine the difference in mean rehoming time. Assuming equal variances, the t-test can be applied to calculate the confidence intervals. Figure 6 illustrates the confidence intervals in the following order: Mixed Breed and Staffordshire Bull Terrier, Shih Tzu and Staffordshire Bull Terrier, and Mixed Breed and Shih Tzu, respectively from the top. This indicates that the paired sample of Mixed Staffordshire Bull Terrier is relatively stable with a significant narrow interval.

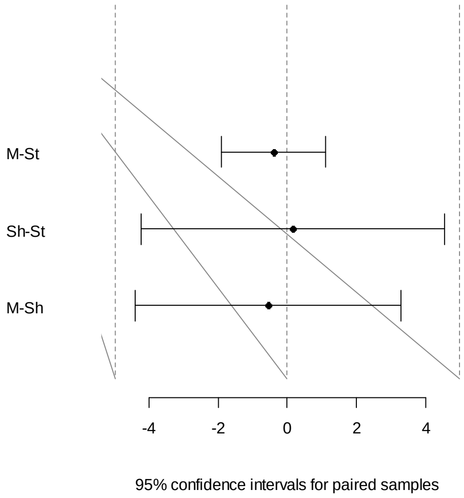

Figure 6 95% confidence intervals for paired samples

Conclusion

By analysing three different samples using various methods, the initial claim that the average rehoming time is around 27 weeks, regardless of the breed is considered. As a result, the estimated parameters show that all three samples' average rehoming time is about 19 weeks. The paired confidence intervals also indicate that the average time for rehoming is constant for all three breeds. Furthermore, none of the confidence intervals include 27. Thus, it is evident that the claim should be rejected.

During the analysis, it was discovered that the sample size for Shih Tzu is significantly small. This means that the numerical summaries may be inaccurate due to insufficient data. In a future work, if more data is collected in the sample, the analysis results would be more precise. With more accurate analysis and further investigation, the data can be utilised to decrease the average rehoming time. This would enable more dogs to be rehomed.

Appendix A

Removing missing values

```
mysample <- mysample[!(is.na(mysample$Breed)), ]
mysample <- mysample[!mysample$Rehomed == 99999, ]
```

Numerical summaries for each sample

```
summary(df_MixedBreed$Rehomed)
mean(df_MixedBreed$Rehomed)
var(df_MixedBreed$Rehomed)
sd(df_MixedBreed$Rehomed)
quantile(df_MixedBreed$Rehomed, type=1)
skewness(df_MixedBreed$Rehomed)
```

Distributions for each sample, Figure 1

```
hist(df_MixedBreed$Rehomed, freq=FALSE, xlab=”Rehomed”, main=”Mixed Breed”)
x <- seq(from = min(df_MixedBreed$Rehomed), to = max(df_MixedBreed$Rehomed),
          by = 0.1)
lines(x, dnorm(x, mean = 18.94, sd = 9.51), lwd = 2, col = "blue")
```

Q-Q plot of the variable ‘Rehomed’, Figure 2

```
qqnorm(df_MixedBreed$Rehomed, pch=16, main="Mixed Breed")
```

CDF visualization for each variable, Figure 3

```
Fn <- ecdf(df_MixedBreed$Rehomed)
plot(Fn, verticals = TRUE, pch = NA)
Fn(30)
sum(df_MixedBreed$Rehomed<=30)/length(df_MixedBreed$Rehomed)
G <- function(x){
        return(pnorm(x, mean=mu, sd=sigma))
}
G(30)
x <- 1:50
lines(x, G(x), col = "red")
```

Estimating parameter: mean, Figure 4

```
mu <- mean(df_MixedBreed$Rehomed)
sigma <- sd(df_MixedBreed$Rehomed)
muhat1 <- rep(NA, 1000)
muhat2 <- rep(NA, 1000)

for (i in 1:1000){
   x <- rnorm(n=10, mean=mu, sd=sigma)
   muhat1[i] <- mean(x)
   muhat2[i] <- quantile(x, type=1)[3]

par(mfrow=c(1, 2))
hist(muhat1, xlim=range(c(muhat1, muhat2)))
abline(v=mu, col="red3", lwd=3)
abline(v=mean(muhat1), col="blue", lty=2, lwd=3)

hist(muhat2, xlim=range(c(muhat1, muhat2)))
abline(v=mu, col="red3", lwd=3)
abline(v=mean(muhat2), col="blue", lty=2, lwd=3)
```

Estimating parameters: variance, Figure 4

```
sigma2hat1 <- rep(NA, 1000)
sigma2hat2 <- rep(NA, 1000)

for (i in 1:1000){
  x <- rnorm(n=10, mean=mu, sd=sigma)
  sigma2hat1[i] <- sd(x)^2
  sigma2hat2[i] <- (9/10)*sd(x)^2
}

par(mfrow=c(1, 2))
hist(sigma2hat1, xlim=range(c(sigma2hat1, sigma2hat2)))
abline(v=sigma^2, col="red3", lwd=3)
abline(v=mean(sigma2hat1), col="blue", lty=2, lwd=3)
hist(sigma2hat2, xlim=range(c(sigma2hat1, sigma2hat2)))
abline(v=sigma^2, col="red3", lwd=3)
abline(v=mean(sigma2hat2), col="blue", lty=2, lwd=3)
```

Confidence interval for each sample: z-test

```
# z-test
n <- nrow(df_MixedBreed)
sigma <- sd(df_MixedBreed$Rehomed)
xbar <- mean(df_MixedBreed$Rehomed)
c <- qnorm(0.975) # 5% level test
CI <- xbar + c(-1, 1)*c*sqrt(sigma^2/n)
CI

z <- (xbar-27)/sqrt(sigma^2/n)
2*pnorm(z) # p-value
```

Confidence interval for each variable: t-test

```
# t-test
n <- nrow(df_ShihTzu)
s <- sd(df_ShihTzu$Rehomed)
xbar <- mean(df_ShihTzu$Rehomed)
t <- qt(p=0.975, df=n-1)
CI <- xbar + c(-1, 1)*t*sqrt(s^2/n)
CI
t <- (xbar-27)/sqrt(s^2/n)
2*(1-pt(q=abs(t), df=n-1)) # p-value
```

Visualisation of the confidence intervals, Figure 5

```
analysis = c("Mixed Breed", "Shih Tzu", "Staffordshire Bull Terrier")
estimate = c(18.9380, 19.5000, 19.3382)
upper = c(19.63783, 23.22725, 20.77724)
lower = c(18.23823, 15.77275, 17.89909)
pval = c(6.874979e-113, 0.0003754547, 1.713764e-25)


par(mfrow=c(1,1))
par(mar=c(6,6,1,6))
plot(x=0, xlim=c(10,30), ylim=c(0,5), type="n", xaxt="n", yaxt="n", xlab=NULL, ylab=NULL,
     ann=FALSE, bty="n")

axis(side=1, cex.axis=1)
mtext("95% confidence interval for each sample", side = 1, line = 4)

for (i in c(10, 15, 20, 25, 30)){
  lines(c(i, i), c(0, 5), lty=2, col="gray53")
}
verticalpos=1:3
mtext(text=analysis, at=verticalpos, side=2, line=5, outer=FALSE, las=1, adj=0)
points(estimate, verticalpos, pch=16)
```

```
for(i in 1:3){
  lines(c(lower[i], upper[i]), c(verticalpos[i], verticalpos[i]))
  lines(c(lower[i], lower[i]), c(verticalpos[i]+0.2, verticalpos[i]-0.2))
  lines(c(upper[i], upper[i]), c(verticalpos[i]+0.2, verticalpos[i]-0.2))
}
```

Paired samples

```
x1 <- df_MixedBreed$Rehomed
x2 <- df_ShihTzu$Rehomed
n1 <- nrow(df_MixedBreed)
n2 <- nrow(df_ShihTzu)

meandiff <- mean(x1)-mean(x2)
t <- qt(0.975, df=n1+n2-2)
sp <- sqrt(((n1-1)*sd(x1)^2+(n2-2)*sd(x2)^2)/(n1+n2-2))
meandiff + c(-1, 1)*t*sp*sqrt(1/n1+1/n2)
```

Visualisation of the confidence intervals of paired samples, Figure 6

```
analysis = c("M-Sh", "Sh-St", "M-St")
estimate = c(-0.5619718, 0.1618357, -0.4001361)
upper = c(3.303705, 4.571650, 1.111917)
lower = c(-4.427648, -4.247978, -1.912189)
pval = c(0.7755, 0.9425, 0.6039)

par(mfrow=c(1,1))
par(mar=c(6,6,1,6))
plot(x=0, xlim=c(-5,5), ylim=c(0,5), type="n", xaxt="n", yaxt="n", xlab=NULL, ylab=NULL,
ann=FALSE,
    bty="n")

axis(side=1, cex.axis=1)
mtext("95% confidence intervals for paired samples", side = 1, line = 4)
```

```
for (i in c(-5, 0, 5)){
  lines(c(i, i), c(0, 5), lty=2, col="gray53")
}

verticalpos=1:3
mtext(text=analysis, at=verticalpos, side=2, line=5, outer=FALSE, las=1, adj=0)
points(estimate, verticalpos, pch=16)

for(i in 1:3){
  lines(c(lower[i], upper[i]), c(verticalpos[i], verticalpos[i]))
  lines(c(lower[i], lower[i]), c(verticalpos[i]+0.2, verticalpos[i]-0.2))
  lines(c(upper[i], upper[i]), c(verticalpos[i]+0.2, verticalpos[i]-0.2))
}
```