Rhethwika Narayanan Kutty

201751000

Rehoming Period For Different Dog Breeds

Introduction

The dataset used here provides information about the time gap until the dog is visited by a potential owner (in weeks), rehoming time (in weeks), age, health, breed, whether it is returned, and reason for it. The sample consists of 646 rows which are details of dogs and 7 columns. There are 3 varieties of breeds in this sample such as Doberman, Greyhound, and Rottweiler. These dogs are either puppy or fully grown. This project aims to analyze whether the rehoming time for each breed is 27 weeks and to find out whether the rehoming period is different for different breeds. So the main columns of interest are 'Rehomed' and 'Breed'.

Analysis and Results:

1. **Data Cleaning:** Columns 'Reason' and 'Breed' have 40 null values in total and there are 9 '99999' values in the column 'Rehomed'. There are 2 '-1' values in 'Rehomed' which is invalid data as time can't be a negative integer. All these rows are removed which constitutes approximately 7.9% of the total data.

2. **Summaries:** From the numerical summary found, it is visible, in Figure 2.1, that Rottweiler has the highest mean health (65/100) and the lowest mean rehoming time (10 weeks) with the smallest standard deviation. This implies people tend more to adopt dogs with better health.

	Breed	mean_health	sd_health	skewness_health	Age_mode	mean_Visited	sd_Visited	skewness_Visited	mean_Rehomed	sd_Rehomed	skewness_Rehomed
1	Dobermann	51.33333	15.69182	-0.1437309	Fully grown	11.380952	7.144762	0.8724214	17.85714	10.541754	0.766366
2	Greyhound	54.63489	17.91490	-0.5915369	Fully grown	13.541582	9.040829	1.2754081	16.83976	10.936842	1.097929
3	Rottweiler	65.87952	12.21694	-0.5534952	Fully grown	8.180723	5.120844	1.1294785	10.34940	6.453272	1.534758

>

Figure 2.1: Comparison of numerical Summaries of each dog breed.

From the Quantiles also, Table 2.1, it's visible that the Rottweiler has the lowest median, Fig 2.2, for rehome weeks and the highest median for health with the lowest IQR spread.

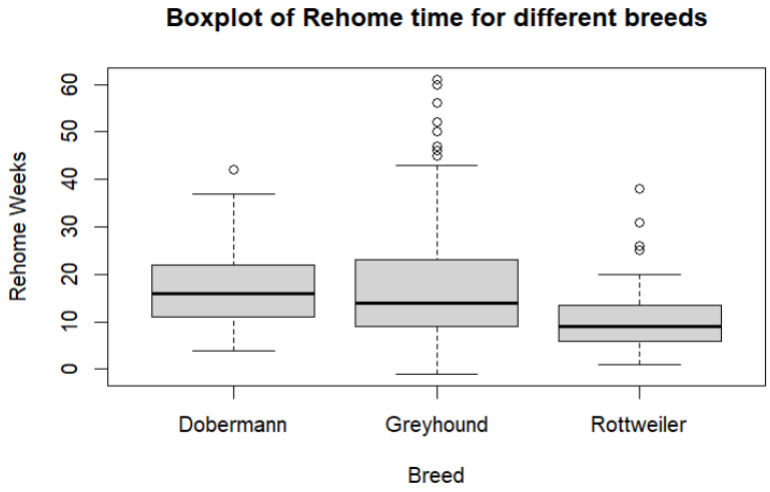


Figure 2.2: Boxplot of 'Rehomed' for different breeds

Rhethwika Narayanan Kutty
201751000

Breed	Q1_Rehomed	Q2_Rehomed	Q3_Rehomed	IQR_Rehomed	Q2_Health
Doberman	11	16	22	11	53
Greyhound	9	14	23	14	57
Rottweiler	6	9	13.5	7.5	66

Table 2.1: Quantile Summary of 'Rehomed' and Median of 'Health' for different breeds

To find the distribution model for each of the breed, density plots are plotted to find the shape of the distribution. Since time is a continuous variable density plot against 'Rehomed' is plotted. Figures 2.3, 2.4, 2.5 shows density plots of Doberman, Greyhound and Rottweiler respectively.

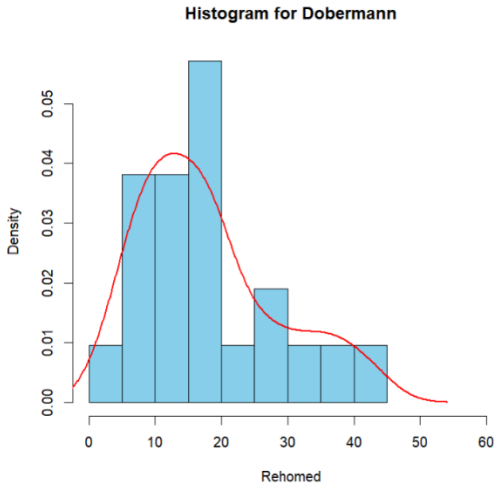


Figure2.3: Density plot of 'Rehomed' for Doberman

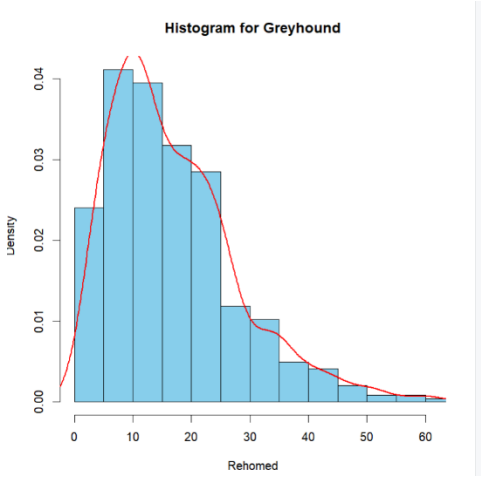


Figure 2.4: Density plot of 'Rehomed' for Greyhound

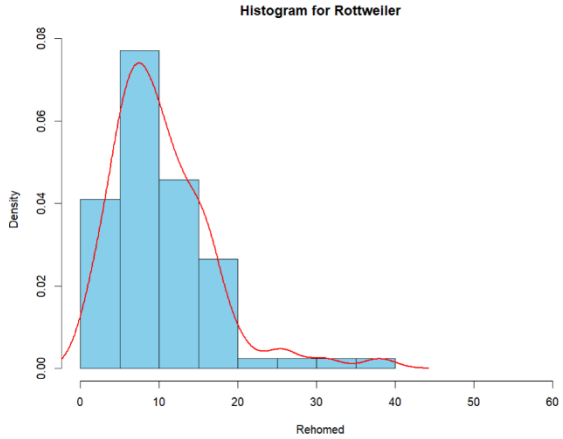


Figure 2.5: Density plot of 'Rehomed' for Rottweiler

3. **Modelling:** From the density plots, it's visible that the data has a strong positive skewness. The shape of the curve for Doberman is almost symmetrical which extends the possibility of normal distribution. For the other two breeds, the curve is not symmetrical and it is highly skewed which can be validated by Figure 2.1. This can be further clarified by using Kolmogorov-Smirnov test and Q-Q plots. From the Q-Q plot, Figure 3.1, it is visible that for

Doberman, most of the points lie in a straight. This can be confirmed by the p-value from the K-S test. The p-value achieved for Doberman is 0.9 which is greater than the significance level (0.05). So there is a possibility that Doberman fits under Normal distribution. The parameters obtained for Doberman are Mean = 17.8571 and Standard deviation = 10.5417.

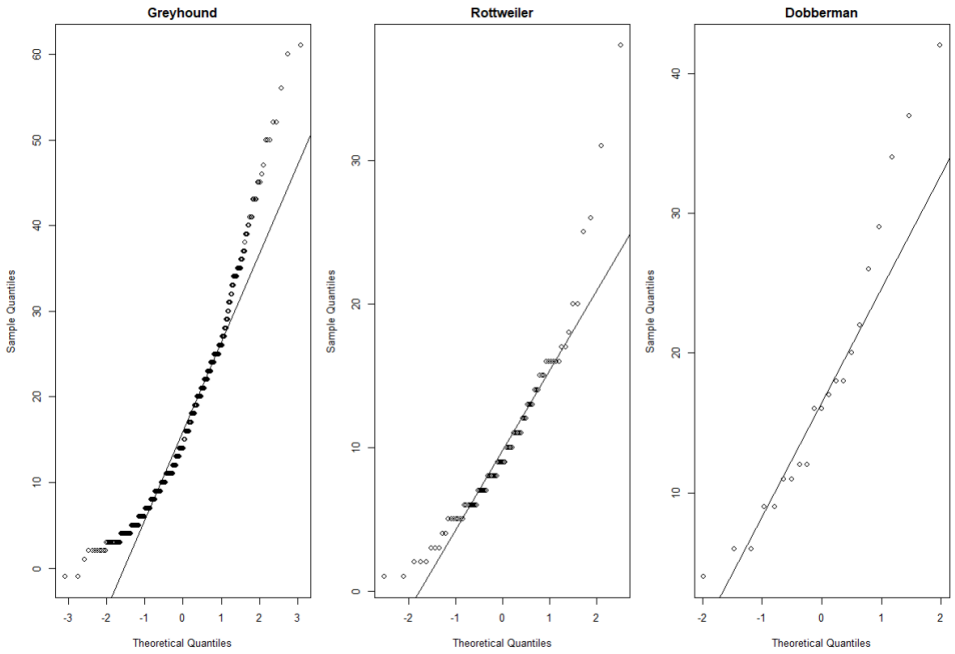
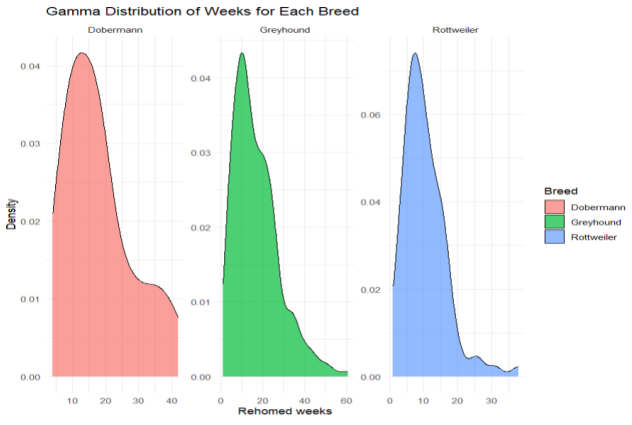


Figure 3.1: Q-Q plots of 'Rehomed' for different breeds

The Q-Q plots for Greyhound and Rottweiler seem exponential but the p-value for the K-S Test for both breeds is less than the significance level for Normal, exponential, and uniform Distribution. So, none of the distribution given in the module fits Greyhound and Rottweiler. A further analysis of different types of continuous distribution led to realize that Gamma distribution, a continuous time-event distribution that is used to model time until a particular event occurs, might be suitable for both of these breeds. It has two parameters- shape parameter and scale parameter. When the shape parameter =1, it becomes an exponential distribution. Figure 3.2 shows the Gamma distribution. This is just a trial to see how would the data will fit in this model.

Figure 3.2: Gamma distribution of different breeds



4. Hypothesis Testing:

Even though the distribution is not normal for Rottweiler and Greyhound, having sample sizes 83 and 493 respectively, we can perform the z-test hypothesis since the sample population for both breeds is large (n>30) and the variance is known. According to the central limit theorem, the mean of a distribution will be approximately normal if the sample is large regardless of the distribution of original data. So, the z-test can be used for these 2 breeds. The sample size of Doberman is very small (n= 21), but the standard deviation of the population is known and has an approximately normal distribution. In this case, z-test can be performed but there is a risk since the distribution is not perfectly normal. When the sample size is small, the distribution of the sample mean might not be approximated correctly by the normal distribution, whereas the t-test has heavier tails, which is suitable in case of uncertainty caused by small samples. For a significance level of 5 %, and a two-tailed test, the null hypothesis will be rejected if the z-test statistics fall below 1.96 or above 1.96(critical values). z-test statistics of Rottweiler and Greyhound fall below -1.96 so we reject the null hypothesis which was mean of rehoming weeks for all breeds is 27 weeks. Table 4.1 shows z-test statistics and p-values of Rottweiler and Greyhound. For Doberman, the t-test is performed and the p-values are less than 0.05. So null hypothesis is rejected. So, for all breeds, the mean rehoming period is not 27 weeks. Table 4.2 shows the t-test scores of Doberman. Figure 4.1 shows an interval plot showing the confidence interval. If the population mean is to the left of the confidence interval, It suggests the true population mean is smaller than the lower boundary of the interval. The width of the confidence interval provides information about the uncertainty of the estimate. Wide confidence Interval means higher uncertainty. A narrow confidence interval suggests lower uncertainty which means higher precision. So, from the interval plot, it is visible that the breed of Doberman has the highest precision.

Table 4.1 : z-test for Rottweiler and Greyhound

Breed	z	p-value	Confidence interval(95%)	Mean
Rottweiler	-2.0499	0.04037	-5.5705 , 26.2693	10.3494
Greyhound	-3.0486	0.002299	10.3076, 23.3719	16.8376

Table 4.2: t-test for Doberman

Breed	z	p-value	df	Confidence interval(95%)	Mean
Doberman	-3.9745	0.0007468	20	13.0585, 22.655	17.8571

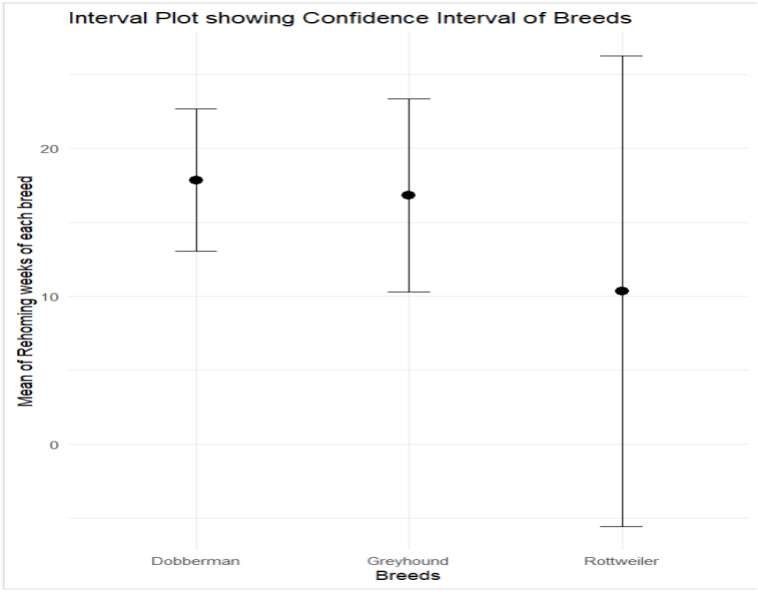


Figure 4.1: Confidence interval on each breed

5. **Comparison:** Performed two-sample t-tests to compare between breeds. Table 6.1 shows p-values for comparison between 2 breeds. Here, the null hypothesis states the mean of two breeds is equal. Since the p-value is below the significance level(5%) for the first and last comparisons in the table, the null hypothesis is rejected. But for samples greyhound and Doberman, the p-value is greater than the significance level. So failed to reject the null hypothesis. The mean for these are very close.

Table 6.1: Two sample t-tests for breeds. GH- greyhound, ROT -Rottweiler, DOB - Doberman

Breed	**p-value**		**Conf.interval(95%)**	**Mean**
GH , ROT	2.765 e-12		4.7875 - 8.1932	16.8397, 10.34940
GH, DOB	0.6696		-5.8978 – 3.863119	16.8397, 17.8514
ROT, DOB	0.00468		-12.4764 – 2.5390	10.3494, 17.8571

6. **Limitation and Practicality:** The high positive skewness, with almost a bell-shaped curve for the breed Doberman, created confusion as skewness is a contradiction to normal distribution. However, the k-s test helped to resolve the ambiguity. As a practical significance, the factors affecting the time for an event to occur are known.

Conclusion

It can be assumed from the findings that the mean rehoming period for the population is not the same for each breed.

Appendix:

Data cleaning:

Removing null values :

```
rows_with_null <- which(apply(mysample, 1, function(x) any(is.na(x))))
mysample <- mysample[-rows_with_null, ]

rows_with_value_99999 <- which(mysample == 99999, arr.ind = TRUE)[,1]
mysample <- mysample[-rows_with_value_99999, ]

value_to_drop <- -1
mysample <- subset(mysample, Rehomed != value_to_drop)
```

segregating:

```
DF_ROT <- mysample[mysample$Breed == "Rottweiler", ]
DF_GH <- mysample[mysample$Breed == "Greyhound", ]
DF_DOB <- mysample[mysample$Breed == "Dobermann", ]
```

Moment Based Summary:

```
grouped_data <- mysample %>%
  group_by(Breed)
breed_summaries <- grouped_data %>%

  summarize(
    mean_health = mean(Health),
    sd_health = sd(Health),
    skewness_health = skewness(Health),
```

```
    Age_mode = mysample$Age[which.max(tabulate(match(mysample$Age,
  mysample$Age)))],


    mean_Visited = mean(Visited),
    sd_Visited = sd(Visited),
    skewness_Visited = skewness(Visited),
    mean_Rehomed= mean(Rehomed),
    sd_Rehomed = sd(Rehomed),
    skewness_Rehomed = skewness(Rehomed)


  )
  print(breed_summaries)
```

Quantile Summary:

```
  quantiles <- mysample %>%
 group_by(Breed) %>%
 summarize(
   q_0.25_rehomed = quantile(Rehomed, probs = 0.25,type=1),
   q_0.5_rehomed = quantile(Rehomed, probs = 0.5,type=1),
   q_0.75_rehomed = quantile(Rehomed, probs = 0.75,type=1),
   IQR_rehomed =IQR(Rehomed),
   q_0.25_Health = quantile(Health, probs = 0.25,type=1),
   q_0.5_Health = quantile(Health, probs = 0.5,type=1),
   q_0.75_Health = quantile(Health, probs = 0.75,type=1),
   IQR_Health = IQR(Health)
 )
quantile_df <- as.data.frame(quantiles)
print(quantile_df)
```

Graphical Summaries:

Boxplot**:**

```
boxplot(Rehomed ~ Breed, data = mysample,
        main = "Boxplot of Rehome time for different breeds",
        xlab = "Breed",
        ylab = "Rehome Weeks")
```

Density plot:

```
selected_column <- "Rehomed"
unique_breeds <- unique(mysample$Breed)
for (i in 1:length(unique_breeds)) {
  breed <- unique_breeds[i]
  subset_data <- mysample[mysample$Breed == breed, ]
  par(mfrow = c(1, 1))

  density_values <- density(subset_data[[selected_column]])
  hist(subset_data[[selected_column]], main = paste("Histogram for", breed),
       xlab = selected_column, col = "skyblue", border = "black",
       xlim = c(0, max(mysample[[selected_column]])), freq = FALSE)
  lines(density_values, col = "red", lwd = 2)
```

Q-Q plot

```
par(mfrow = c(1, 3), mar = c(5, 4, 2, 1))
qqnorm(DF_GH$Rehomed,main="Greyhound")
qqline(DF_GH$Rehomed)
qqnorm(DF_ROT$Rehomed,main="Rottweiler")
qqline(DF_ROT$Rehomed)
qqnorm(DF_DOB$Rehomed,main="Dobberman")
qqline(DF_DOB$Rehomed)
par(mfrow = c(1, 1))
```

MODELLING

#Gamma

```
mysample %>%
 ggplot(aes(x = Rehomed, fill = Breed)) +
 geom_density(alpha = 0.7) +
 labs(title = "Gamma Distribution of Weeks for Each Breed",
     x = "Rehomed weeks",
     y = "Density") +
 theme_minimal() +
 facet_wrap(~Breed, scales = "free")
```

z-test

```
install.packages("BSDA")
library(BSDA)
#result_DOB <- z.test(DF_DOB$Rehomed, mu = 27, sigma.x = 74, alternative = "two.sided",conf.level = 0.95)
result_GH <- z.test(DF_GH$Rehomed, mu = 27, sigma.x = 74, alternative = "two.sided",conf.level = 0.95)
result_ROT <- z.test(DF_ROT$Rehomed, mu = 27, sigma.x = 74, alternative = "two.sided",conf.level = 0.95)
print(result_ROT)
print(result)
```

t-test

```
t.test(DF_DOB$Rehomed, mu = 27, alternative = "two.sided", conf.level = 0.95)
```

Two sample t-test

```
result_ROT_DOB = t.test(DF_ROT$Rehomed,DF_DOB$Rehomed)
print (result_ROT_DOB)
```

K-S Test

```
mu <- mean(DF_GH$Rehomed)
sigma <- sd(DF_GH$Rehomed)
```

Rhethwika Narayanan Kutty
201751000

```
install.packages("nortest")

library(nortest)

ks.test(x = DF_ROT$Rehomed, y = "pexp", rate = 1)

ks.test(x = DF_DOB$Rehomed, y = "pnorm", mean  = mu , sd = sigma ) #normal

ks.test(DF_GH$Rehomed, "punif", min = min(DF_ROT$Rehomed), max = max(DF_ROT$Rehomed))
```

#parameter of doberman for modelling

```
mean <- mean(DF_DOB$Rehomed)

sd <- sd(DF_DOB$Rehomed)
```

Confidence Interval

```
z_test_var1 <- z.test(DF_GH$Rehomed, mu =27, sigma.x = 74,conf.level = 0.95)

z_test_var2 <- z.test(DF_ROT$Rehomed, mu =27, sigma.x = 74,conf.level = 0.95)

z_test_var3 <- t.test(DF_DOB$Rehomed, mu = 27,conf.level = 0.95)

mean_var1 <- z_test_var1$estimate

ci_var1 <- z_test_var1$conf.int


mean_var2 <- z_test_var2$estimate

ci_var2 <- z_test_var2$conf.int


mean_var3 <- z_test_var3$estimate

ci_var3 <- z_test_var3$conf.int


plot_data <- data.frame(
  Variable = c("Greyhound", "Rottweiler", "Dobberman"),
  Mean = c(mean_var1, mean_var2, mean_var3),
  LowerCI = c(ci_var1[1], ci_var2[1], ci_var3[1]),
  UpperCI = c(ci_var1[2], ci_var2[2], ci_var3[2])
)


library(ggplot2)
```

```
ggplot(plot_data, aes(x = Variable, y = Mean, ymin = LowerCI, ymax = UpperCI)) +
  geom_point(position = position_dodge(width = 0.2), size = 3) +
  geom_errorbar(position = position_dodge(width = 0.2), width = 0.2) +
  labs(x = "Breeds", y = "Mean of Rehoming weeks of each breed", title = "Interval Plot showing Confidence Interval of Breeds") +
  theme_minimal()
```