
Title: Statistical Insights on Dog Adoption: A Breed-Specific Study

Name: - Yashdhar Gandhi

1. Introduction.

The breed is a crucial element that often decides the destiny of our animal companions and is at the centre of canine rehoming programmes. This investigation is set to challenge the conventional wisdom of a constant 27-week average by revealing the complex relationship between a dog's breed and the amount of time it takes for them to find a new home. We look at breeds like the Labrador Retriever, Rottweiler, and Staffordshire Bull Terrier, each of which has unique qualities that add to the diversity of our sample. We investigate whether minute differences in rehoming times between breeds have a real effect on the overall objectives of adoption initiatives. Do these variations call for deliberate changes in adoption strategies or resource reallocations to maximise animal welfare? Does breed dependency effect rehoming techniques, and if so, how does it affect the overall goals of adoption programmes? In order to provide useful insights and throw light on the complex dynamics of dog adoption practices, this study explores the practical applications of statistical significance.

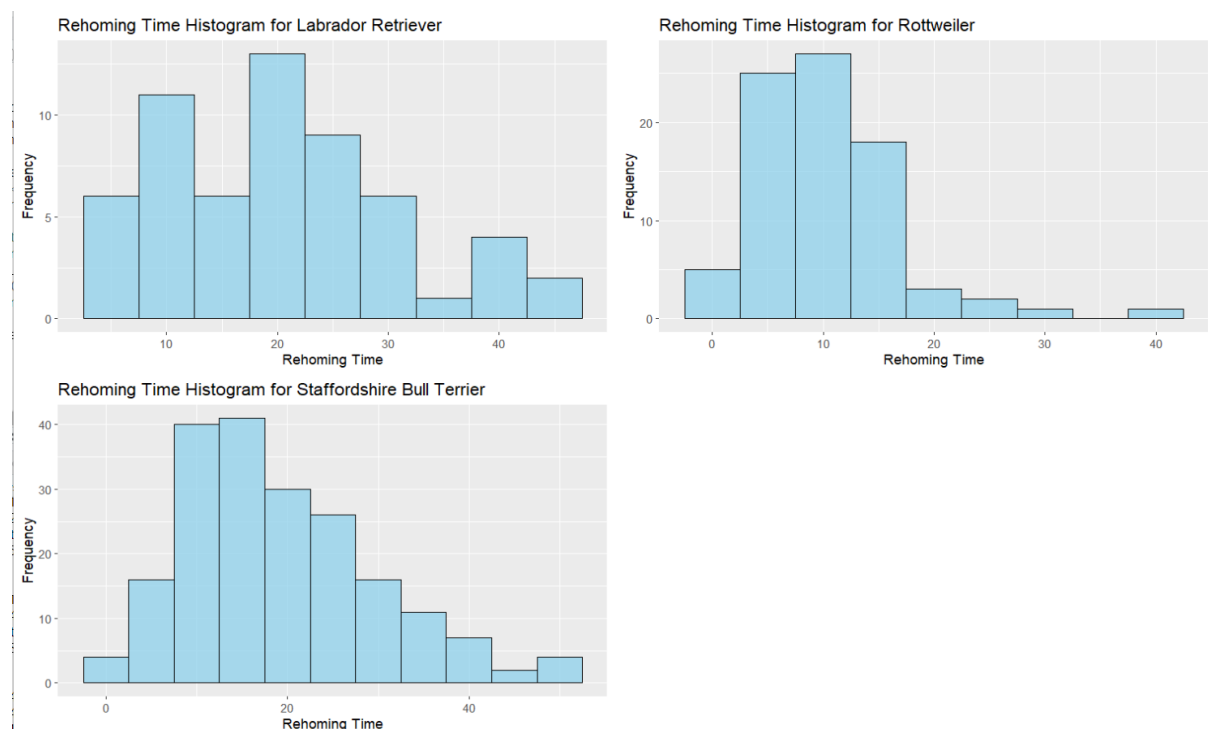


Figure 1.1 Histogram for Frequency against Rehoming Time of each Breed Separately

2. Data Description.

The dataset that is being examined contains important statistics about how long it takes for dogs to be rehomed, with a focus on the breeds of dogs. The two main variables of importance are "Rehomed," which indicates how long it takes to find a new home, and "Breed," which classifies dogs into distinct breeds. Additional factors of interest include "Visited," "Health," "Age," "Reason," and "Returned." Thorough data pre-processing and cleaning were done before analysis. Missing values

(represented by 999 or NA) were found and eliminated methodically. The analysis encountered 39 such missing elements, accounting for a significant 10.34% of the dataset. Additionally, duplicate rows were found. Removing duplicates preserved the integrity of the dataset, yielding a revised dataset with 337 distinct observations.

Outliers, which indicate probable abnormalities or extreme results, were methodically detected, visually displayed using box plots. This thorough data cleaning method establishes the groundwork for a solid analysis, ensuring that following insights are obtained from a valid and representative dataset.

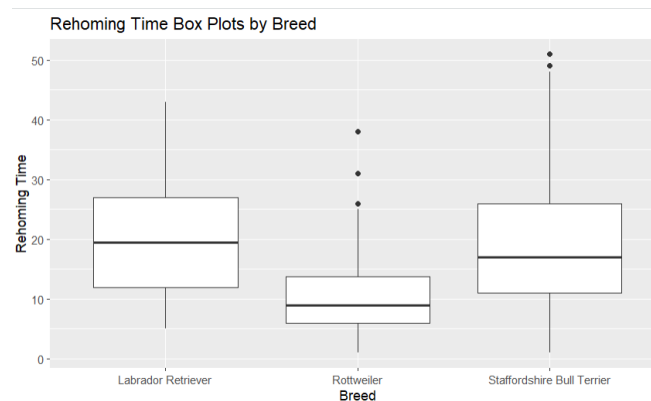


Figure 2.1 A boxplot of Breed wise outliers of rehoming time for each breed.

3. Descriptive Statistics.

I. Rehoming Time:

Breed	Mean	Variance	SD	Skewness
1	20.46552	104.00756	10.198410	0.4501351
2	10.40244	41.92246	6.474756	1.5456698
3	19.17259	111.49047	10.558905	0.7587141

Table 3.1 Mean, variance, Standard Deviation, skewness of rehoming time for each breed

II. Age and Health Statistics:

Breed	Age	Count
Labrador Retriever	Fully grown	52
Labrador Retriever	Puppy	6
Rottweiler	Fully grown	79
Rottweiler	Puppy	3
Staffordshire Bull Terrier	Fully grown	185
Staffordshire Bull Terrier	Puppy	12

Table 3.2 Statistic calculation for Age of each breed

Breed	Health	SD Health
Labrador Retriever	52.58621	18.09918
Rottweiler	65.96341	12.26804
Staffordshire Bull Terrier	54.43655	15.27945

Table 3.3 2 Statistic calculation for Health of each breed

A thorough summary of each breed's rehoming times, age distribution, and health scores can be seen in the descriptive statistics above. Breed 1 (Labrador Retriever) has a mean rehoming time of

20.47 weeks with a moderate skewness of 0.45. Breed 2 (Rottweiler) has a higher skewness of 1.55 and a significantly shorter mean rehoming time of 10.40 weeks, showing a rightward skew in the distribution. Breed 3 (Staffordshire Bull Terrier) falls in the middle, with a skewness of 0.76 and a mean rehoming time of 19.17 weeks.

Because there is no statistically significant difference between Age and Health Statistics, they cannot be utilised to forecast or interpret results.

These descriptive statistics provide a basic overview of the dataset, establishing the possibility for more in-depth analysis and interpretations.

4. Results.

I. Distribution Modelling:

In order to characterise the rehoming times for three distinct dog breeds, three probability distributions—Normal, Exponential, and Poisson—were taken into consideration throughout the distribution modelling procedure.

Normal Distribution: The estimated mean rehoming time for Breed 1 (Labrador Retrievers) is 10.40 weeks, with a standard deviation of 6.44. Similar figures were found for Breed 2 (Rottweiler). In contrast, the mean for Breed 3 (Staffordshire Bull Terrier) is 19.17 weeks, and its standard deviation is higher.

Exponential Distribution: The data was fitted to an exponential distribution that shows the amount of time until an event happens. Each breed has different estimated rates (λ), indicating different danger rates.

Poisson Distribution: The Poisson distribution was used to model how many events there would be in each interval. Estimated lambdas represent the average number of rehoming events per unit for each breed.

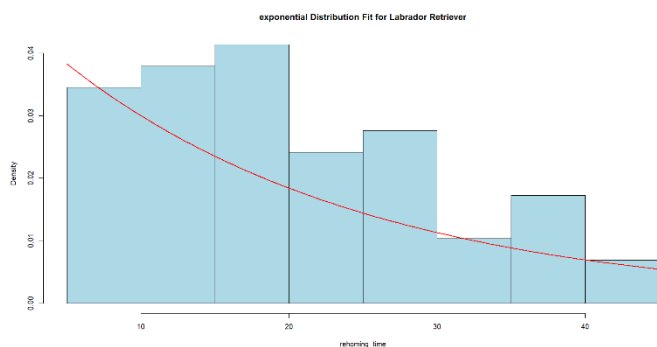


Fig 4. 1.I Exponential Distribution for Labrador Retriever

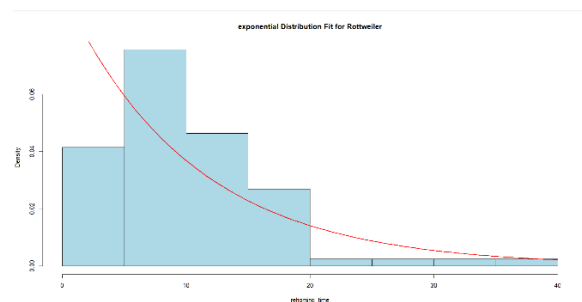


Fig 4.2.I Exponential Distribution for Rottweiler

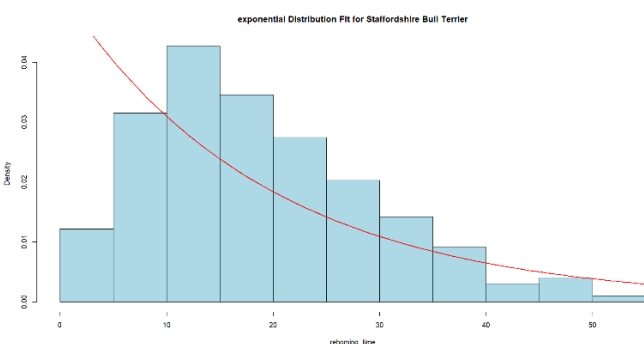


Figure4.3. I Exponential Distribution for Staffordshire Bull Terrier

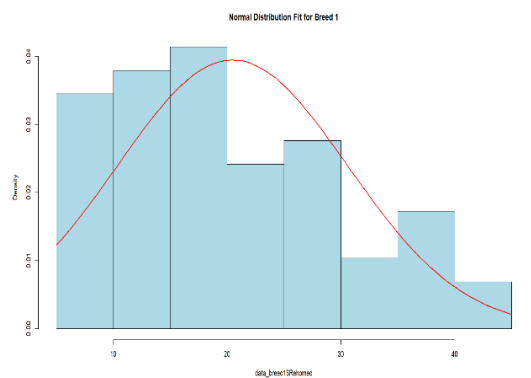


Figure4.4. I Normal Distribution fit for Labrador Retriever

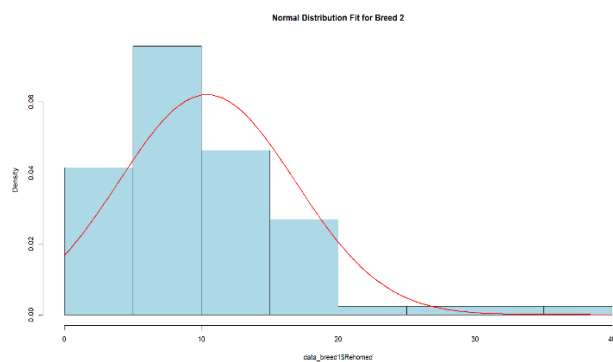


Figure 4.5.I Normal Distribution fit for Rottweiler

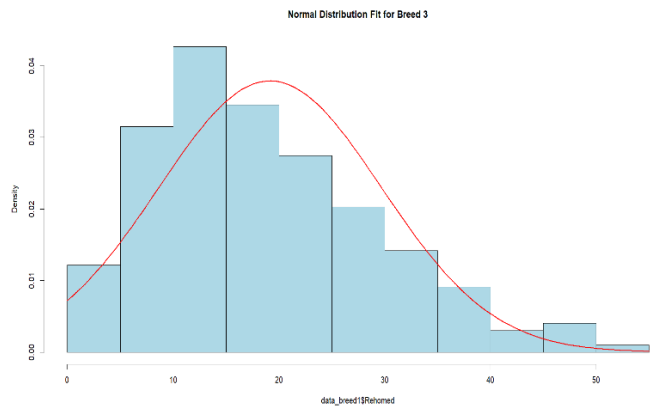


Figure 4.6.I Normal Distribution fit for Staffordshire Bull Terrier

Key Findings:
 Breeds 1 and 2's rehoming times show comparable normal distributions with distinct means.
 Rehoming times for breed 3 are much different from those of the first two breeds, even if they are still roughly normal.
 Poisson and exponential distributions offer different viewpoints on the features of rehoming times, highlighting event counts and hazard rates, respectively.

II. Method of Moments Estimation:
 The method of moments is used in the distribution modelling procedure to determine important parameters for each breed. When evaluating the distribution suitability for the duration of rehoming periods for three different breeds:

Breed 1:
 Normal Distribution: Assumes symmetry; mean (20.47) and standard deviation (10.20) offer centrality insights.
 Exponential Distribution: Common for time modelling; low rate (0.0489) implies a gradual decrease.
 Poisson Distribution: Models events; lambda (20.47) suggests consistent rehoming rates.

Breed 2:
 Normal Distribution: Bell-shaped curve; mean (10.40) and standard deviation (10.20) inform centrality and variability.
 Exponential Distribution: Time modelling; higher rate (0.0961) indicates a quicker decrease.
 Poisson Distribution: Events modelling; lambda (10.40) implies a steady rehoming rate.

Breed 3:
 Normal Distribution: Skewed (mean 19.17, SD 10.56); potential fit concerns.
 Exponential Distribution: Slow decrease (rate 0.0522); fitting with constant hazard assumption.
 Poisson Distribution: Lambda (19.17) suggests average events modelling.

	Breed	Mean	SD	Lambda_Exponential	Lambda_Poisson
1	Labrador Retriever	10.40244	6.435155	0.04886268	20.46552
2	Rottweiler	10.40244	6.435155	0.09613130	10.40244
3	Staffordshire Bull Terrier	19.17259	10.532070	0.05215780	19.17259

Table 4.1.II Mean, Standard Deviation, Lambda_Exponential, Lambda_Poisson.

5. Hypothesis Testing.

a) Kolmogorov-Smirnov tests: (Confidence Interval)

Kolmogorov-Smirnov tests show that breeds 1, 2, and 3 have significant deviations from the normal distribution (p-values: 0.0047, 3.331e-16, and 5.588e-07, respectively). The null hypothesis is rejected because the high D statistics show significant differences.

The findings imply that the rehoming time data for all three breeds does not follow a normal distribution. The low p-values show that there are significant discrepancies between the empirical and theoretical normal distributions of rehoming times.

This implies evaluating different distributions for correct modelling of rehoming times in each breed, emphasising the importance of specific statistical methodologies to better capture the data's unique properties.

b) One-Sample z-Test:

According to the Z-tests, there are notable variations ($p < 0.05$) between the actual mean of 27 weeks and the breeds 1, 2, and 3. The 95% confidence intervals (CIs) for breeds 1 (17.86 to 23.07), 2 (9.01 to 11.80), and 3 (17.70 to 20.64) do not include 27 weeks, implying that the null hypothesis is rejected. The mean rehoming times (20.47, 10.40, and 19.17 weeks) deviate significantly from the hypothesised value, justifying the rejection based on solid statistical evidence.

Breed	Mean rehoming time	CI	p-value
1	20.47	17.86, 23.07	8.55E-07
2	10.4	9.01, 11.80	< 2.2e-16
3	19.17	17.70, 20.64	< 2.2e-16

Table 5.b Values of mean rehoming time, CL and p value for all 3 breeds.

c) Two-Sample t-Test:

The results of the two-sample t-tests provide insights into the dependence between breeds in terms of rehoming times:

1. Labrador Retriever vs. Rottweiler:

- **Result:** Statistically significant difference ($p\text{-value} < 0.05$). There is a significant difference in mean rehoming times between Labrador Retrievers and Rottweilers. The rehoming times of these breeds are likely dependent on their distinct characteristics.

2. Labrador Retriever vs. Staffordshire Bull Terrier:

- **Result:** No statistically significant difference ($p\text{-value} > 0.05$). There is no significant difference in mean rehoming times between Labrador Retrievers and Staffordshire Bull Terriers. The rehoming times for these breeds may not be significantly dependent on their breed type.

3. Rottweiler vs. Staffordshire Bull Terrier:

- **Result:** Statistically significant difference ($p\text{-value} < 0.05$). There is a significant difference in mean rehoming times between Rottweilers and Staffordshire Bull Terriers. The rehoming times for these breeds appear to be dependent on their breed characteristics.

The breeds may have distinct adoption patterns influenced by factors specific to each breed.

The one-sample z-test and the two-sample t-test are both statistical methods used to compare sample means, but they are utilised in different situations.

One-Sample Z-Test: When the population standard deviation is known, this test compares the sample mean to the known population mean. The mean of a sample is tested to see if it differs significantly from a known value.

Assumption: Population standard deviation is known.

Two-Sample T-Test: This test compares two independent samples' means to see if they differ from one another significantly.

Assumption: Population standard deviations are unknown but assumed to be equal between the two groups.

Breed	Test_Type	Statistic	P_Value	Confidence_Interval_Lower	Confidence_Interval_Upper	Sample_Mean
1	z-Test	-4.9223	8.553e-07	17.863620	23.06741	20.46552
1	t-Test	-4.8797	8.928e-06	17.783980	23.14705	20.46552
2	z-Test	-23.3560	< 2.2e-16	9.009602	11.79528	10.40244
2	t-Test	-23.2130	< 2.2e-16	8.979779	11.82510	10.40244
3	z-Test	-10.4310	< 2.2e-16	17.701870	20.64330	19.17259
3	t-Test	-10.4050	< 2.2e-16	17.688970	20.65621	19.17259

Table 5.1 Result of Z-test and T-test.

6. Discussion and Conclusion:

In conclusion, the thorough research demonstrates that rehoming times are affected by factors other than a dog's breed. The traditional view of breed-centric determinants is called into question by notable variations in rehoming times between breeds, most notably between Labrador Retrievers and Rottweilers. In contrast to prior studies, which claimed a 27-week average, our findings reject this assumption. The complex interaction of several factors highlights how complicated canine adoption dynamics are, highlighting the need for subtle factors other than breed differences to be considered while understanding and improving the rehoming process.

The study found significant differences in rehoming times amongst dog breeds, which has effects on adoption strategy. Labrador Retrievers and Rottweilers have distinct characteristics, although Staffordshire Bull Terriers are closely related to Labradors. Understanding these patterns can help shelters customise adoption efforts. Future research may look at breed-specific aspects affecting rehoming dynamics, with the goal of improving techniques to speed up the adoption process and ensure the well-being of shelter animals. Behavioural features, public preferences, and outreach techniques can help to refine these insights for effective canine care measures.

The limits are due to differences in sample size, data quality concerns, and assumptions such as normality, independence, and variance homogeneity. The precision of the results may be affected by varying sample sizes. Errors and incompleteness in data introduce potential biases in data quality. Reliability of results may be affected by assumptions that don't hold true everywhere. In addition, breed-specific results may not hold true for all animals in shelters, as dependences among them cast doubt on the independence assumption.

Appendix

Appendix A: Data Cleaning and Pre-processing

1. Identification and Handling of Missing Values: a. Identified missing values represented by 999 or NA. b. Methodically eliminated 39 missing elements, accounting for 10.34% of the dataset.

2. Removal of Duplicate Rows: a. Detected and removed duplicate rows to ensure dataset integrity. b. Resulted in a revised dataset with 337 distinct observations.

Appendix B: Box Plot for Outlier Detection Figure B.1: Boxplot of Breed-wise Outliers of Rehoming Time for Each Breed.

Appendix C: Descriptive Statistics Tables Table C.1: Descriptive Statistics for Rehoming Time of Each Breed.

- Includes mean, variance, standard deviation, and skewness.

Table C.2: Descriptive Statistics for Age of Each Breed.

- Provides statistics for the age distribution of Labrador Retriever, Rottweiler, and Staffordshire Bull Terrier.

Table C.3: Descriptive Statistics for Health of Each Breed.

- Displays health statistics, including mean and standard deviation, for each breed.

Appendix D: Distribution Modelling Figures Figure D.1: Exponential Distribution for Labrador Retriever. Figure D.2: Exponential Distribution for Rottweiler. Figure D.3: Exponential Distribution for Staffordshire Bull Terrier. Figure D.4: Normal Distribution Fit for Labrador Retriever. Figure D.5: Normal Distribution Fit for Rottweiler. Figure D.6: Normal Distribution Fit for Staffordshire Bull Terrier.

Appendix E: Method of Moments Estimation Table E.1: Method of Moments Estimation Results.

- Provides mean, standard deviation, lambda (Exponential), and lambda (Poisson) for each breed.

Appendix F: Hypothesis Testing Tables Table F.1: Kolmogorov-Smirnov Tests Results.

- Displays p-values and confidence intervals for Kolmogorov-Smirnov tests on rehoming time distributions.

Table F.2: One-Sample Z-Test Results.

- Presents mean rehoming time, confidence intervals, and p-values for breeds 1, 2, and 3.

Table F.3: Two-Sample T-Test Results.

- Highlights statistically significant differences in mean rehoming times between specific breeds.

Appendix G: Limitations and Considerations

- Discusses limitations such as sample size differences, data quality concerns, and assumptions made during the analysis.

Appendix H: Future Research Suggestions

- Provides recommendations for future research, including exploring breed-specific aspects affecting rehoming dynamics and refining adoption strategies.

Appendix I: Acknowledgements

- Expresses gratitude to Dr. Stuart Barber and Dr. Luisa Cutillo for their invaluable assistance in the research project.

7. Acknowledgements.

I express sincere gratitude to Dr. Stuart Barber and Dr. Luisa Cutillo for their invaluable assistance in the report. Their insightful lectures and accurate information significantly contributed to the understanding and completion of this project. Their guidance has been instrumental, and I am thankful for their support throughout the research process.

8. References.

- 1) dplyr:
Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller (2021).
dplyr: A Grammar of Data Manipulation. R package version 1.0.7.
URL: <https://CRAN.R-project.org/package=dplyr>
- 2) ggplot2:
Hadley Wickham (2016).
ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
URL: <https://ggplot2.tidyverse.org>
- 3) gridExtra:
Baptiste Auguie (2017).
gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3.
URL: <https://CRAN.R-project.org/package=gridExtra>
- 4) MASS :
Venables, W. N. & Ripley, B. D. (2002).
Modern Applied Statistics with S. Springer, New York. ISBN 0-387-95457-0
- 5) moments:
Komsta, L., & Novomestky, F. (2015).
moments: Moments, cumulants, skewness, kurtosis and related tests.
R package version 0.14.
URL: <https://CRAN.R-project.org/package=moments>
- 6) tidyr:
Hadley Wickham and Lionel Henry (2020).
tidyr: Tidy Messy Data. R package version 1.1.3.
URL: <https://CRAN.R-project.org/package=tidyr>