MATH5741

Statistical Theory and Methods

Statistical Analysis on Rehoming Time for different Dog Breeds in Animal Shelter

Sathyasri Sudhakar – 201749908

Introduction

Given a collection of records documenting the stray, unwanted, or neglected dogs sent to animal shelters to be rehomed, we analyse their rehoming patterns based on their breeds. We work with a subset of this larger dataset, consisting of roughly 670 records with details on breeds, such as Labrador Retrievers, Greyhounds, and Border Collies. This dataset includes details about the dog's rehoming time, a health indicator, age, breed, and why it was brought to the shelter. The Rehoming and Breed columns are this analysis's main variables of interest.

Data Cleaning

In this section, we begin cleaning the data by removing any unwanted rows containing invalid or missing values. The total number of removed observations and overall removal percentage are as follows:

Column Name	Condition	Number of Observations	Total Number	Percentage of Removal
Rehomed (Rehoming time)	Missing values recorded as "99999"	9	670	1.34%
Breed	Missing values record as "NA"	6		0.90%
Rehomed or Breeds	Missing values recorded as "99999" or "NA"	15		2.24%

Table 1: Tabular representation of number of removed observations and its percentages.

Data Exploration

	Border Collie - 78 records			Greyhound - 515 records			Labrador Retriever - 62 records		
	Visited	Rehomed	Health	Visited	Rehomed	Health	Visited	Rehomed	Health
Mean	14.46	20.47	52.91	13.64	16.82	54.74	14	19.85	53.27
Min	1	0	1	1	1	2	1	5	5
Max	44	50	92	59	61	100	43	43	87

Table 2: Statistical Summaries for different breed.

After splitting the dataset by breeds and performing data exploration on the three different breeds - we identified the following essential features:

- The average time taken for a Greyhound to be rehomed is much less when compared to the other two breeds, and it also contains the maximum time a dog has taken to be rehomed.
- Greyhound makes up the vast majority of the dataset (78.62%).
- The majority of the dogs that are got to the shelter were fully grown.
- Examining why the dogs are in the shelter home, we find that most of them are either neglected or strays.
- Figure 1 provides data on the frequency of dogs rehomed with different time frames.

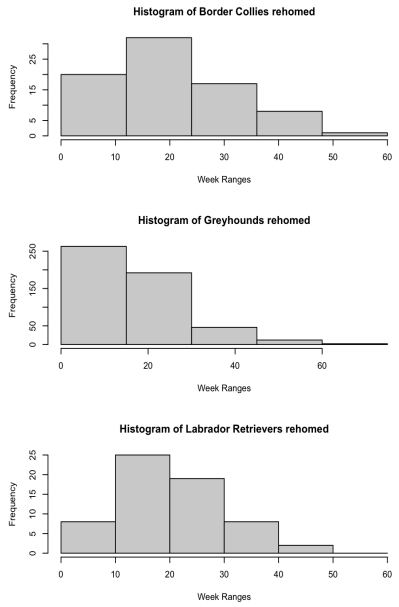

Figure 1: Histogram of Rehomed data.

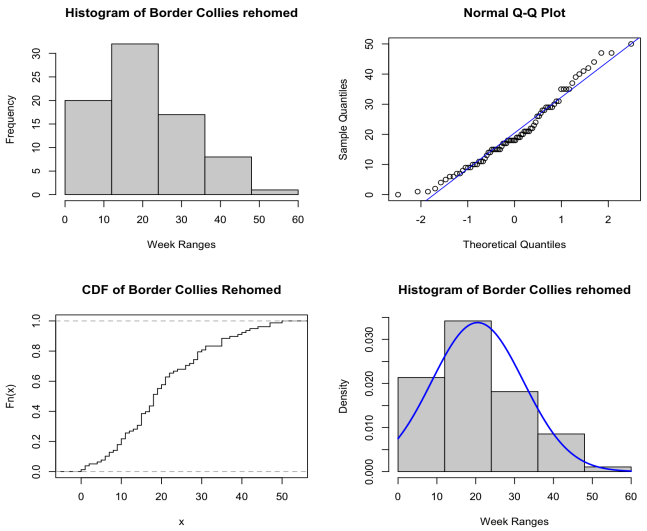

Figure 2: Histogram, Normal QQ plot, CDF plots of Border Collies Rehoming data.

Modelling and estimation

Based on the graphical summaries in Figure 1, we try to determine the best distribution and check the suitability of the proposed model one by one.

1. Border Collies

The data in Figure 2 is a normal distribution. We are analysing this with different tests and plots. The output of the tests is as follows:

Kolmogorov Smirnov Test		Shapiro - Wilk test		Chi - Square	
D - value	p - value	W - value	p - value	P - value	p - value
0.11055	0.2962	0.965	0.03056	12.769	0.1733

Table 3: Suitability Check outputs for Border Collie data.

In relation to Figure 2:

- Q-Q plot shows a slight deviation from the straight line, but most points lie on the line, so that it may be a normal distribution.
- Histogram – The histogram contains a single peak (unimodal) and is roughly symmetrical, indicating that it is a normal distribution.

In relation to Table 3:

- Kolmogorov Smirnov (KS) test – here, the p-value is 0.2962, indicating that there is not a statistically significant difference between the Cumulative distribution function (CDF) of the data and the CDF of a normal distribution.
- Shapiro-Wilk test – Here, the p-value is 0.03056, slightly less than 0.05, suggesting that the data might not be normally distributed.
- Chi-squared test – A p-value of 0.1733 indicates no significant difference between the data and a normal distribution.

Overall, the test results are mixed for everything other than the Shapiro-Wilk test in favour of it being a Normal distribution.

2. Labrador Retriever

The data is normally distributed, even if the plot looks skewed and non-symmetric. The output of the tests are as follows :

Kolmogorov Smirnov Test		Shapiro - Wilk test		Chi - Square	
D - value	p - value	W - value	p - value	P - value	p - value
0.10523	0.4985	0.94918	0.01219	10.742	0.2168

Table 4: Suitability checks output for Labrador Retriever data.

Concerning Figure 3:

- Q-Q plot shows a good fit of the data to the straight line, indicating that it is a normal distribution.
- Histogram – The histogram contains a single peak (unimodal) and is roughly symmetrical, indicating that it is a normal distribution.

Concerning Table 4:

- KS test – here, the p-value is 0.4985, indicating that there is not a statistically significant difference between the CDF of the data and the CDF of a normal distribution.
- Shapiro-Wilk test – Here, the p-value is 0.01219, slightly less than 0.05, suggesting that the data might not be normally distributed. This test is more powerful at detecting non-normality in smaller samples, so the fact that the p-value is less does not necessarily mean that the data is not normally distributed.
- Chi-squared test – A p-value of 0.02168 indicates no difference between the data and a normal distribution.

Overall, it suggests that the data is normally distributed.

3. Greyhounds

The output of the different tests are as follows:

Normal Distribution						Exponential Distribution	
Kolmogorov Smirnov Test		Shapiro - Wilk test		Chi - Square		Kolmogorov Smirnov Test	
D - value	p - value	W - value	p - value	P - value	p - value	D - value	p - value
0.11322	3.69E-06	0.91912	5.47E-16	172.04	< 2.2e-16	0.1872	4.44E-16

Table 5:Suitability check outputs for Greyhound data.

This data is none of the distributions because:

1. Normal Distribution – This data is not normally distributed as:
   - The p-values for the KS and Shapiro-Wilk tests are very small (Table 5), indicating that the data is from a normal distribution.
   - The chi-square test also rejects the null hypothesis of normality with a minimal value of p.
   - The Q-Q plot in Figure 4 shows a significant deviation from the straight line than expected for a normal distribution.
2. Uniform Distribution – This data is not uniformly distributed.
   - When the density plot of the data was plotted in Figure 4 it did not contain relatively constant line across a range of values.
3. Exponential Distribution – This data is not exponentially distributed.
   - From the p-value obtained from the KS test (Table 5), it is very small, which indicates that we can reject that the data is exponentially distributed.

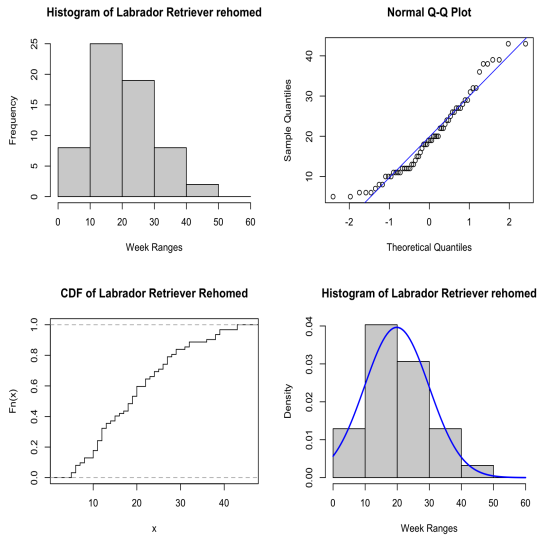

Figure 3: Histogram, Normal Q-Q plot, CDF of Labrador Retrievers data.

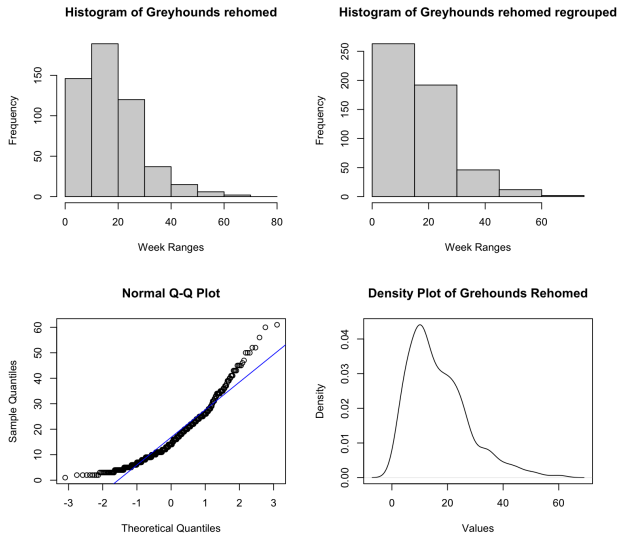

Figure 4: Histograms, Normal Q-Q plots, and Density plots for Greyhounds data.

Inference

Calculating the confidence level for each breed by:

1. Border Collie and Labrador Retrievers– The z test was applied to calculate the confidence level.
    a. The standard deviation of the dataset is known.
    b. Z test assumes that the data is a normal distribution, and both the datasets here are normally distributed.
    c. N is the length of the dataset, which is > 30.
    d. Finally, the z-test is more powerful and has a better performance.
2. Greyhound – applied the t-test to calculate the confidence level.
    a. It does not need a standard deviation, even if we know the standard deviation of the dataset.
    b. It does not assume the data is normally distributed; here, we do not know our data's distribution.

Interpretation made from the Confidence Interval Forest Plot – Figure 5

- From the forest plot, none of the confidence intervals plotted contain the mean value 27.
- The confidence interval for Labrador Retriever and Border Collie is wide, suggesting that the sample size is small and uncertainty in our estimation. Whereas for Greyhound, it is narrow, stating that the sample size is large.
- As the confidence interval for all three breeds lies to the left of the hypothesized mean 27, the observed sample mean is lower than the hypothesized mean.

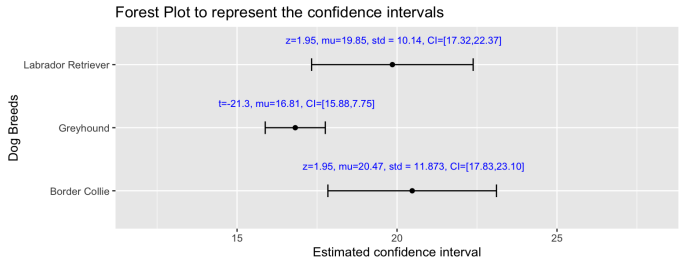

Figure 5: Forest Plot to understand the confidence intervals of the different breeds.

Comparison

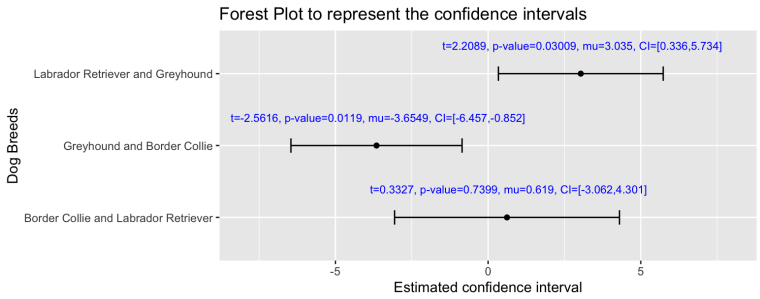

Figure 6: Forest plot to represent the Confidence Intervals for different pairs of data.

Interpretation made from the Confidence Interval Forest Plot for pairs of breeds – Figure 6

- Only the Border Collie and Labrador Interval confidence interval includes a zero, indicating that the actual difference between the two means is zero, which means that they are not significantly different from each other.
- For the other two pairs – Labrador Retriever & Greyhound and Greyhound & Border Collie we see that the confidence interval does not contain 0, indicating they are significantly different.
- The wider the confidence interval is, the less precise the estimate of the difference in means is. The smaller the sample size, the wider the confidence interval; therefore, we see a wider confidence interval for Border Collie and Labrador Retriever than the other two.

Discussions

Through statistical analysis, we were able to determine the extent of difference between the various dog breeds' rehoming patterns, estimate the distributions of the rehoming data, and calculate confidence intervals. Greyhounds are among the most popular breeds for rehoming, as evidenced by the analysis's finding that most dogs belong to this breed. Additionally, given the length of the dataset, it would be simpler to determine with certainty from the analysis above whether a greyhound is being rehomed.

Limitations :

- Data Quality – the accuracy and the completeness of the data. If there are any missing or inaccurate data, then it would impact the analysis.
- Variations – Over time, adoption preferences, dog breeds, and other factors change. To reflect the current circumstances, this data must therefore be updated on a regular basis.
- Relation – Although we are only working with rehoming data in this case, the dogs' chances of finding new homes vary greatly depending on a number of factors, including their age, health, and the reason they are in the shelter. There are no underlying relationships between the single data we are working with and the dataset's other columns.

Future Work :

- Employ advanced statistical methods, like machine learning, to find more underlying connections and patterns among the variables influencing rehoming.
- Predictive models could be developed to estimate the rehoming time for dogs based on their characteristics.

Appendix

List of Figures

List of Tables

Code

```
rm(list=ls())
load("rehoming.Rdata")
createsample("201749908")
save(mysample,file="mysample.RData")
mysample
#---------------------- Data Cleaning----------------------#
unique(mysample$Visited)#None
unique(mysample$Rehomed)### Has 99999 and -1 present
unique(mysample$Health)#None
unique(mysample$Breed)### Has NA present
unique(mysample$Age)#None
unique(mysample$Reason)### Has NA values present
unique(mysample$Returned)### Has NA values present
### Getting the percentage of observations removed
## Breed
len_breed=length(mysample$Breed)
len_breed###no. of samples in breed
len_na_breed=sum(is.na(mysample$Breed))
len_na_breed
per_rem_breed=(len_na_breed/len_breed)*100
per_rem_breed
## Rehoming
len_rehoming=length(mysample$Rehomed)
len_rehoming
len_na_rehoming=nrow(mysample[mysample$Rehomed=='99999',])
len_na_rehoming
per_rem_rehoming=(len_na_rehoming/len_rehoming)*100
per_rem_rehoming
## Removing the columns
new_samp<-mysample[!is.na(mysample$Breed),]
new_samp<-new_samp[new_samp$Rehomed!='99999',]
### Also removing the -1 present in the Rehoming column
new_samp<-new_samp[new_samp$Rehomed!='-1',]
### Replacing the Negative values in the visited column with NA
new_samp$Visited[new_samp$Visited <= 0] <- "NA"
#---------------------- Data Exploration----------------------#
new_samp$Visited<-as.numeric(new_samp$Visited)
new_data_grouped<-split(new_samp, new_samp$Breed)
summary(new_data_grouped$`Border Collie`)
summary(new_data_grouped$Greyhound)
summary(new_data_grouped$`Labrador Retriever`)
A<-new_data_grouped$`Border Collie`$Age
```

```
table(A)
B<-new_data_grouped$Greyhound$Age
table(B)
C<-new_data_grouped$`Labrador Retriever`$Age
table(C)
A<-new_data_grouped$`Border Collie`$Reason
table(A)
B<-new_data_grouped$Greyhound$Reason
table(B)
C<-new_data_grouped$`Labrador Retriever`$Reason
table(C)
A<-new_data_grouped$`Border Collie`$Returned
table(A)
B<-new_data_grouped$Greyhound$Returned
table(B)
C<-new_data_grouped$`Labrador Retriever`$Returned
table(C)
# Rehoming - graphically
BC_R<-new_data_grouped$`Border Collie`$Rehomed
LR_R<-new_data_grouped$`Labrador Retriever`$Rehomed
G_R<-new_data_grouped$Greyhound$Rehomed
par(mfrow=c(3,1))
hist(BC_R, breaks=seq(from=0, to=60, by=12),right=FALSE, freq=TRUE,
    main="Histogram of Border Collies rehomed", xlab="Week Ranges")
hist(G_R,breaks=seq(from=0, to=80, by=15),right=FALSE, freq=TRUE,
    main="Histogram of Greyhounds rehomed", xlab="Week Ranges")
hist(LR_R, breaks=seq(from=0, to=60, by=10),right=FALSE,freq=TRUE,
    main="Histogram of Labrador Retrievers rehomed", xlab="Week Ranges")
par(mfrow=c(1,1))
#----------------------Modeling and Estimation----------------------#
#----Border Collie--------#
par(mfrow=c(2,2))
#Histogram
hist(BC_R, breaks=seq(from=0, to=60, by=12),right=FALSE, freq=TRUE,
    main="Histogram of Border Collies rehomed", xlab="Week Ranges")
#QQ plot
qqnorm(BC_R)
abline(a=mean(BC_R), b=sd(BC_R), col="blue")
##CDF
Fn<-ecdf(BC_R)
plot(Fn,verticals = TRUE,pch=NA, main="CDF of Border Collies Rehomed")
library(MASS)
fit_normal <- fitdistr(BC_R, "normal")
hist(BC_R, breaks=seq(from=0, to=60, by=12),right=FALSE, freq=FALSE,
    main="Histogram of Border Collies rehomed", xlab="Week Ranges")
x<-c(BC_R)
curve(dnorm(x,mean=fit_normal$estimate[1], sd=fit_normal$estimate[2]),
     col = "blue", lwd = 2, add = TRUE)
##Kolmogorov Smirnov Test
ks.test(x=BC_R, y="pnorm", mean=mean(BC_R), sd=sd(BC_R))
##Shapiro Wilk test
shapiro.test(BC_R)
##Chi squared test
library(nortest)
pearson.test(BC_R)
#----Labrador Retriever--------#
par(mfrow=c(2,2))
#Histogram
hist(LR_R, breaks=seq(from=0, to=60, by=10),right=FALSE, freq=TRUE,
    main="Histogram of Labrador Retriever rehomed", xlab="Week Ranges")
#QQ plot
qqnorm(LR_R)
abline(a=mean(LR_R), b=sd(LR_R), col="blue")
##CDF
Fn<-ecdf(LR_R)
plot(Fn,verticals = TRUE,pch=NA, main="CDF of Labrador Retriever Rehomed")
library(MASS)
```

```
fit_normal <- fitdistr(LR_R, "normal")
hist(LR_R, breaks=seq(from=0, to=60, by=10),right=FALSE, freq=FALSE,
    main="Histogram of Labrador Retriever rehomed", xlab="Week Ranges")
x<-c(LR_R)
curve(dnorm(x,mean=fit_normal$estimate[1], sd=fit_normal$estimate[2]),
     col = "blue", lwd = 2, add = TRUE)
##Kolmogorov Smirnov Test
ks.test(x=LR_R,  y="pnorm", mean=mean(LR_R), sd=sd(LR_R))
##Shapiro Wilk test
shapiro.test(LR_R)
##Chi squared test
library(nortest)
pearson.test(LR_R)
#----Greyhound--------#
par(mfrow=c(2,2))
#Histogram
hist(G_R, breaks=seq(from=0, to=75, by=15),right=FALSE, freq=TRUE,
    main="Histogram of Greyhounds rehomed", xlab="Week Ranges")
#QQ plot
x=G_R
x2<-rexp(G_R)
qqnorm(x2, main = "Normal Q-Q plot: exponential data")
abline(a=mean(rexp(G_R)), b=mean(rexp(G_R)), col="blue")
##CDF
Fn<-ecdf((G_R))
plot(Fn,verticals = TRUE,pch=NA, main="CDF of Greyhounds Rehomed")
library(MASS)
fit_exp<-fitdistr(G_R, "exponential")
hist(G_R, breaks=seq(from=0, to=75, by=15),right=FALSE, freq=FALSE,
    main="Histogram of Greyhounds rehomed", xlab="Week Ranges")
x<-c(G_R)
curve(dexp(x,rate=fit_exp$estimate[1]),
     col = "red", lwd = 2, add = TRUE)
##Kolmogorov Smirnov Test
ks.test(x=G_R, y="pexp", rate=1/mean(G_R))
#Not an exponential distribution
#----Greyhound--------#
par(mfrow=c(2,2))
#Histogram
hist(G_R, breaks=seq(from=0, to=80, by=10),right=FALSE, freq=TRUE,
    main="Histogram of Grehound rehomed", xlab="Week Ranges")
#QQ plot
qqnorm((G_R))
abline(a=mean((G_R)), b=sd((G_R)), col="blue")
##CDF
Fn<-ecdf(G_R)
plot(Fn,verticals = TRUE,pch=NA, main="CDF of Greyhounds Rehomed")
library(MASS)
fit_normal <- fitdistr(G_R, "normal")
hist(G_R, breaks=seq(from=0, to=80, by=10),right=FALSE, freq=FALSE,
    main="Histogram of Greyhounds rehomed", xlab="Week Ranges")
x<-c(G_R)
curve(dnorm(x,mean=fit_normal$estimate[1], sd=fit_normal$estimate[2]),
     col = "blue", lwd = 2, add = TRUE)
##Kolmogorov Smirnov Test
ks.test(x=G_R,  y="pnorm", mean=mean(G_R), sd=sd(G_R))
##Shapiro Wilk test
shapiro.test(G_R)
##Chi squared test
library(nortest)
pearson.test(G_R)
#-----Plot for Greyhound--------#
par(mfrow=c(2,2))
hist(G_R, breaks=seq(from=0, to=80, by=10),right=FALSE, freq=TRUE,
    main="Histogram of Greyhounds rehomed", xlab="Week Ranges")
hist(G_R, breaks=seq(from=0, to=80, by=15),right=FALSE, freq=TRUE,
    main="Histogram of Greyhounds rehomed regrouped", xlab="Week Ranges")
```

```
qqnorm((G_R))
abline(a=mean((G_R)), b=sd((G_R)), col="blue")
plot(density(G_R), main = "Density Plot of Grehounds Rehomed", xlab = "Values")
#-----------------------Inference-----------------------------#
# Using z intervals
# 1. Border Collie - Z test seems best
BC_R<-c(BC_R)
n <- length(BC_R)
sample_mean <- mean(BC_R)
sample_sd <- sd(BC_R)
confidence_level <- 0.95
z_value <- qnorm((1 + confidence_level) / 2)
z_value
margin_error <- z_value * (sample_sd / sqrt(n))
lower_bound <- sample_mean - margin_error
upper_bound <- sample_mean + margin_error
cat(
  paste(
    "Sample Mean:", sample_mean, "\n",
    "Sample Standard Deviation:", sample_sd, "\n",
    "Confidence Interval (", confidence_level * 100, "%): [", lower_bound, ", ", upper_bound, "]"
  )
)
#2. Labrador Retriever - Z test seems best
LR_R<-c(LR_R)
n <- length(LR_R)
sample_mean <- mean(LR_R)
sample_sd <- sd(LR_R)
confidence_level <- 0.95
z_value <- qnorm((1 + confidence_level) / 2)
z_value
margin_error <- z_value * (sample_sd / sqrt(n))
lower_bound <- sample_mean - margin_error
upper_bound <- sample_mean + margin_error
cat(
  paste(
    "Sample Mean:", sample_mean, "\n",
    "Sample Standard Deviation:", sample_sd, "\n",
    "Confidence Interval (", confidence_level * 100, "%): [", lower_bound, ", ", upper_bound, "]"
  )
)
#3. Grey Hound - T test seems best - as we dont know the distribution
G_R<-c(G_R)
checking_mean<-27
result_G_R<-t.test(G_R, mu=checking_mean)
print(result_G_R)
## Confidence Interval Plot
library(ggplot2)
options(
  repr.plot.width = 2,    # Width of the plot in inches
  repr.plot.height = 2   # Height of the plot in inches
)
data <- data.frame(
  Study = c("Border Collie", "Labrador Retriever", "Greyhound"),
  Estimate = c(20.47435, 19.85483, 16.81942), #mean
  lci = c(17.839, 17.32968, 15.88044),      # Lower confidence interval
  uci = c(23.109, 22.3799, 17.75840),# Upper confidence interval
  values=c("z=1.95, mu=20.47, std = 11.873, CI=[17.83,23.10]",
           "z=1.95, mu=19.85, std = 10.14, CI=[17.32,22.37]",
           "t=-21.3, mu=16.81, CI=[15.88,7.75]")
  )
ggplot(data, aes(y = Study, x = Estimate, xmin = lci, xmax = uci)) +
  geom_point() +
  geom_errorbarh(height = 0.2) +
  labs(x = "Estimated confidence interval", y = "Dog Breeds") +
  ggtitle("Forest Plot to represent the confidence intervals")+
  coord_cartesian(xlim = c(12, 28))+# Setting x-axis limits)
```

```
  geom_text(aes(label = paste("", values)), vjust = -3, size = 3, color = "blue")
#----------------------Comparison--------------------------#
#1. Border Collie and Labrador Retriever
result_BC_LR<-t.test(BC_R,LR_R)
print(result_BC_LR)
diff_mean<-mean(BC_R)-mean(LR_R)
diff_mean
diff_strder <-sqrt(var(BC_R)/length(BC_R) + var(LR_R)/length(LR_R))
margin_error <- qt(0.975, df = length(BC_R) + length(LR_R) - 2) * diff_strder
lower_bound <- diff_mean - margin_error
upper_bound <- diff_mean + margin_error
cat("Confidence Interval for the Difference in Means:", "[", lower_bound, ",", upper_bound, "]\n")
#2. Labrador Retriever and Greyhound
result_LR_R_G_R<-t.test(LR_R,G_R)
print(result_LR_R_G_R)
diff_mean<-mean(LR_R)-mean(G_R)
diff_mean
diff_strder <-sqrt(var(LR_R)/length(LR_R) + var(G_R)/length(G_R))
margin_error <- qt(0.975, df = length(LR_R) + length(G_R) - 2) * diff_strder
lower_bound <- diff_mean - margin_error
upper_bound <- diff_mean + margin_error
cat("Confidence Interval for the Difference in Means:", "[", lower_bound, ",", upper_bound, "]\n")
#3. Greyhound and Border Collie
result_G_R_BC_R<-t.test(G_R,BC_R)
print(result_G_R_BC_R)
diff_mean<-mean(G_R)-mean(BC_R)
diff_mean
diff_strder <-sqrt(var(G_R)/length(G_R) + var(BC_R)/length(BC_R))
margin_error <- qt(0.975, df = length(G_R) + length(BC_R) - 2) * diff_strder
lower_bound <- diff_mean - margin_error
upper_bound <- diff_mean + margin_error
cat("Confidence Interval for the Difference in Means:", "[", lower_bound, ",", upper_bound, "]\n")
#-----------------------------------------------------------------#
#Confidence plot for pairs
data <- data.frame(
  Study = c("Border Collie and Labrador Retriever", "Labrador Retriever and Greyhound"
            , "Greyhound and Border Collie"),
  Estimate = c(0.619, 3.0354, -3.654941), # diff means
  lci = c(-3.0623, 0.33642, -6.4572),      # Lower confidence interval
  uci = c(4.3014, 5.7344, -0.85264),      # Upper confidence interval
  values=c("t=0.3327, p-value=0.7399, mu=0.619, CI=[-3.062,4.301]",
           "t=2.2089, p-value=0.03009, mu=3.035, CI=[0.336,5.734]",
           "t=-2.5616, p-value=0.0119, mu=-3.6549, CI=[-6.457,-0.852]")
)
ggplot(data, aes(y = Study, x = Estimate, xmin = lci, xmax = uci)) +
  geom_point() +
  geom_errorbarh(height = 0.2) +
  labs(x = "Estimated confidence interval", y = "Dog Breeds") +
  ggtitle("Forest Plot to represent the confidence intervals")+
  coord_cartesian(xlim = c(-8,8)) + # Setting x-axis limits)
  geom_text(aes(label = paste("", values)), vjust = -3, size = 3, color = "blue")
#########################################################
```