Chickadee Pathogen Data Statistical Analysis

Junsoo Park

2023-09-02

TRU 4001 - Biostatistics - Final Project

Introduction

Data from nest boxes at 47 mountain chickadee sites was collected to examine microbial community compositions associated with chickadees in urban, semi-urban, and rural environments. DNA sequencing was used to determine the relative abundance of different microbial taxa present in swabs from either chickadee nests or feathers in nest boxes in these three habitat types. Nest boxes were set up to encourage nesting in the sample sites. For the feather data, physical characteristics of the birds were recorded. The data was also used to determine two different measures of microbial species richness (alpha diversity) at each site.

1. Numerical Summary

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(emmeans)
library(GGally)
library(factoextra)
library(ggrepel)
library(vegan)
library(Hmisc)
library(car)
library(pheatmap)
library(ggrepel)
library(reshape2)
library(rgl)
library(vcd)
library(ade4)
library(visdat)
library(gridExtra)
theme_set(theme_classic())
```

```
# Read the data
chickadeeData <- read.csv("ChickadeeData.csv")
dissimilaritiesData <- read.csv("ChickadeeDissimilarities.csv")

# Check for Missing Data in the Entire Dataset
sum(is.na(chickadeeData))
```

```
## [1] 85
```

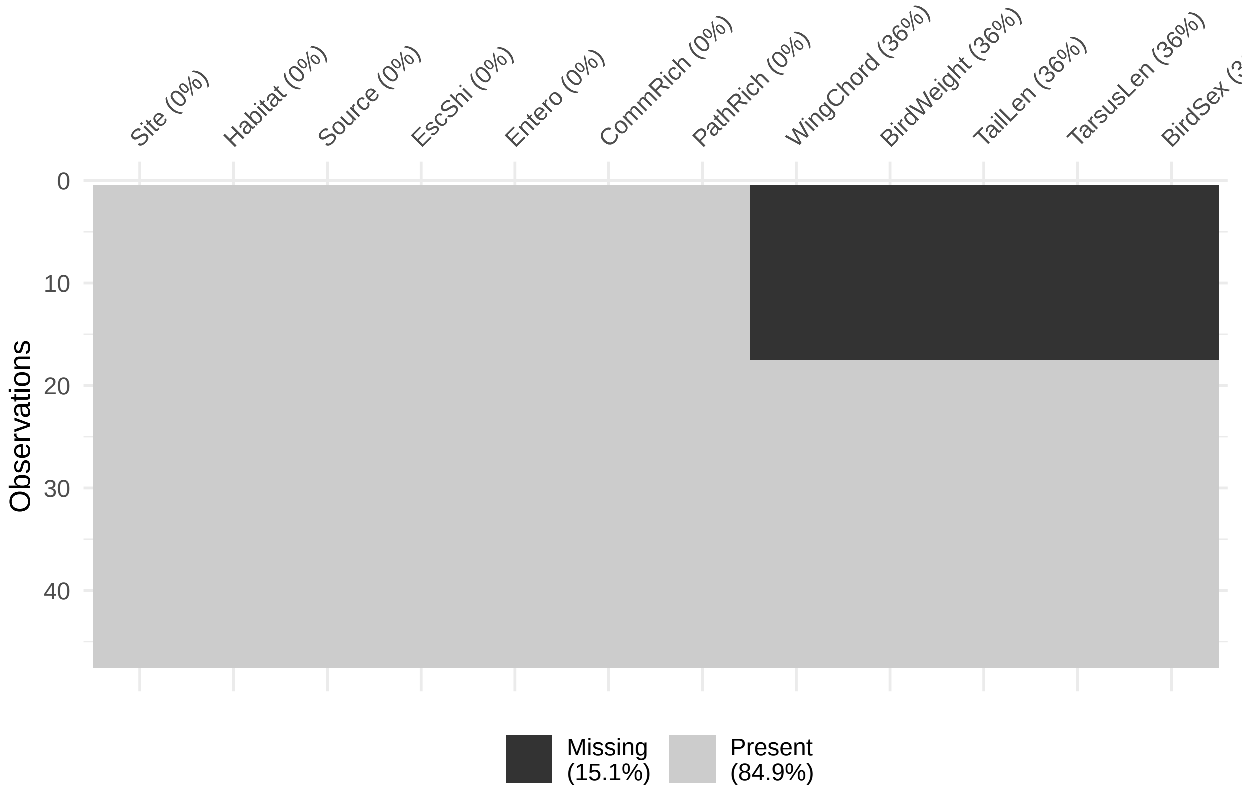
```
# Check for Missing Data Column-wise
colSums(is.na(chickadeeData))
```

```
##       Site    Habitat     Source     EscShi     Entero   CommRich   PathRich
##          0          0          0          0          0          0          0
##  WingChord BirdWeight    TailLen  TarsusLen    BirdSex
##         17         17         17         17         17
```

```
# Check for Missing Data Row-wise
rowSums(is.na(chickadeeData))
```

```
##  [1] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0 0 0 0 0 0 0 0
```

```
# Visualizing Missing Data
vis_miss(chickadeeData)
```



```
# dissimilarities data

# Check for Missing Data in the Entire Dataset
sum(is.na(dissimilaritiesData))
```

```
## [1] 0
```

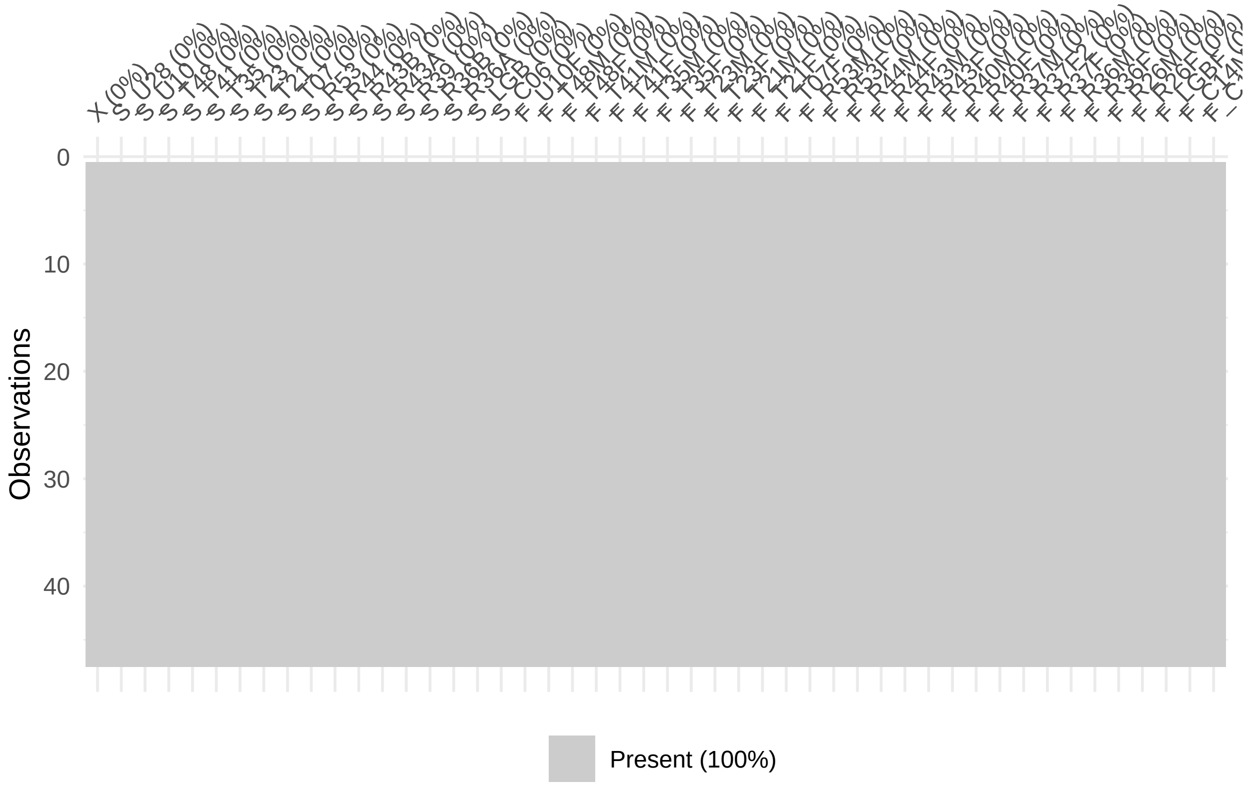
```
# Check for Missing Data Column-wise
colSums(is.na(dissimilaritiesData))
```

```
##       X   S_U28   S_U10   S_T48   S_T41   S_T35   S_T23   S_T21   S_T07   S_R53 
##       0       0       0       0       0       0       0       0       0       0 
##   S_R44  S_R43B  S_R43A   S_R39  S_R36B  S_R36A   S_LGB   S_C06  F_U10F  F_T48M 
##       0       0       0       0       0       0       0       0       0       0 
##  F_T48F  F_T41M  F_T41F  F_T35M  F_T35F  F_T23M  F_T23F  F_T21M  F_T21F  F_T07F 
##       0       0       0       0       0       0       0       0       0       0 
##  F_R53M  F_R53F  F_R44M  F_R44F  F_R43M  F_R43F  F_R40M  F_R40F  F_R37M F_R37F2 
##       0       0       0       0       0       0       0       0       0       0 
##  F_R37F  F_R36M  F_R36F  F_R26M  F_R26F   F_LGBF  F_C14M  F_C14F 
##       0       0       0       0       0       0       0       0
```

```
# Check for Missing Data Row-wise
rowSums(is.na(dissimilaritiesData))
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0 0 0 0 0 0 0 0
```

```
# Visualizing Missing Data
vis_miss(dissimilaritiesData)
```



```r
# Summarize data
summary(chickadeeData)
```

```
##      Site             Habitat             Source              EscShi      
##  Length:47          Length:47          Length:47          Min.   :   0.0  
##  Class :character   Class :character   Class :character   1st Qu.:  25.0  
##  Mode  :character   Mode  :character   Mode  :character   Median : 112.0  
##                                                           Mean   : 577.9  
##                                                           3rd Qu.: 432.5  
##                                                           Max.   :6520.0  
##                                                                           
##      Entero         CommRich            PathRich      WingChord    
##  Min.   :  0.00   Length:47          Min.   :11.0   Min.   :60.00  
##  1st Qu.:  9.00   Class :character   1st Qu.:36.0   1st Qu.:63.00  
##  Median : 20.00   Mode  :character   Median :41.0   Median :65.00  
##  Mean   : 92.74                      Mean   :44.3   Mean   :64.78  
##  3rd Qu.: 48.50                      3rd Qu.:53.5   3rd Qu.:66.75  
##  Max.   :687.00                      Max.   :88.0   Max.   :69.00  
##                                                     NA's   :17     
##    BirdWeight    TailLen        TarsusLen       BirdSex         
##  Min.   :10   Min.   :53.00   Min.   :16.80   Length:47         
##  1st Qu.:11   1st Qu.:55.00   1st Qu.:17.93   Class :character  
##  Median :11   Median :57.00   Median :18.25   Mode  :character  
##  Mean   :12   Mean   :57.23   Mean   :18.25                     
##  3rd Qu.:12   3rd Qu.:60.00   3rd Qu.:18.68                     
##  Max.   :17   Max.   :62.00   Max.   :19.50
##  NA’s   :17   NA’s   :17      NA’s   :17
```

```
head(chickadeeData)
```

```
##    Site Habitat Source EscShi Entero CommRich PathRich WingChord BirdWeight
## 1 S_U28   rural   nest     86    557      low       41        NA         NA
## 2 S_U10   rural   nest   1429    546      low       36        NA         NA
## 3 S_T48   rural   nest      0     84      low       31        NA         NA
## 4 S_T41   rural   nest      1     22      low       41        NA         NA
## 5 S_T35   rural   nest    248      4      low       38        NA         NA
## 6 S_T23   rural   nest      8     13      low       31        NA         NA
##   TailLen TarsusLen BirdSex
## 1      NA        NA    <NA>
## 2      NA        NA    <NA>
## 3      NA        NA    <NA>
## 4      NA        NA    <NA>
## 5      NA        NA    <NA>
## 6      NA        NA    <NA>
```

```
# See column names
names(chickadeeData)
```

```
##  [1] "Site"       "Habitat"    "Source"     "EscShi"     "Entero"    
##  [6] "CommRich"   "PathRich"   "WingChord"  "BirdWeight" "TailLen"   
## [11] "TarsusLen"  "BirdSex"
```

```
# Check unique values of a specific column
unique(chickadeeData$Source)
```

```
## [1] "nest"    "feather"
```

```
# Detailed summary of dataframe
str(chickadeeData)
```

```
## ’data.frame’:    47 obs. of  12 variables:
##  $ Site      : chr  "S_U28" "S_U10" "S_T48" "S_T41" ...
##  $ Habitat   : chr  "rural" "rural" "rural" "rural" ...
##  $ Source    : chr  "nest" "nest" "nest" "nest" ...
##  $ EscShi    : int  86 1429 0 1 248 8 14 4 0 4 ...
##  $ Entero    : int  557 546 84 22 4 13 22 20 15 9 ...
##  $ CommRich  : chr  "low" "low" "low" "low" ...
##  $ PathRich  : int  41 36 31 41 38 31 45 48 49 32 ...
##  $ WingChord : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BirdWeight: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TailLen   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TarsusLen : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BirdSex   : chr  NA NA NA NA ...
```

```
# Summarize each variable using Hmisc
describe(chickadeeData)
```

```
## chickadeeData
##
##  12  Variables      47  Observations
## --------------------------------------------------------------------------------
## Site
##        n  missing distinct
##       47        0       47
##
## lowest : F_C14F F_C14M F_LGBF F_R26F F_R26M, highest: S_T35  S_T41  S_T48  S_U10  S_U28
## --------------------------------------------------------------------------------
## Habitat
##        n  missing distinct
##       47        0        3
##
## Value           rural semi-urban      urban
## Frequency          20         14         13
## Proportion      0.426      0.298      0.277
## --------------------------------------------------------------------------------
## Source
##        n  missing distinct
##       47        0        2
##
## Value      feather    nest
## Frequency       30      17
## Proportion   0.638   0.362
## --------------------------------------------------------------------------------
## EscShi
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       47        0       42        1    577.9    905.8      1.0      3.2
##      .25      .50      .75      .90      .95
##     25.0    112.0    432.5   1348.0   2941.7
##
## Value          0    50   100   150   200   250   300   350   450   850   900
## Frequency     16     6     3     3     3     2     1     1     2     1     1
## Proportion 0.340 0.128 0.064 0.064 0.064 0.043 0.021 0.021 0.043 0.021 0.021
##
## Value       1200  1250  1400  2250  3200  3750  6500
## Frequency      1     2     1     1     1     1     1
## Proportion 0.021 0.043 0.021 0.021 0.021 0.021 0.021
##
## For the frequency table, variable is rounded to the nearest 50
## --------------------------------------------------------------------------------
## Entero
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       47        0       32    0.998    92.74    142.4      0.0      1.6
##      .25      .50      .75      .90      .95
##      9.0     20.0     48.5    322.4    553.7
##
## Value          0     5    10    15    20    25    30    35    45    60    80
## Frequency      8     6     7     2     5     2     1     1     4     2     1
## Proportion 0.170 0.128 0.149 0.043 0.106 0.043 0.021 0.021 0.085 0.043 0.021
##
## Value        200   285   370   545   555   615   685
## Frequency      1     2     1     1     1     1     1
## Proportion 0.021 0.043 0.021 0.021 0.021 0.021 0.021
## 
## For the frequency table, variable is rounded to the nearest 5
## --------------------------------------------------------------------------------
## CommRich 
##        n  missing distinct 
##       47        0        2 
##                       
## Value       high   low
## Frequency     19    28
## Proportion 0.404 0.596
## --------------------------------------------------------------------------------
## PathRich 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##       47        0       30    0.998     44.3    15.48     24.8     30.0 
##      .25      .50      .75      .90      .95 
##     36.0     41.0     53.5     59.0     63.0 
## 
## lowest : 11 17 23 29 30, highest: 58 59 63 66 88
## --------------------------------------------------------------------------------
## WingChord 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##       30       17       10    0.983    64.78    2.599    62.00    62.00 
##      .25      .50      .75      .90      .95 
##    63.00    65.00    66.75    68.00    68.00 
## 
## Value      60.00 61.98 62.43 62.97 63.96 64.95 65.94 66.93 67.92 69.00
## Frequency      1     3     1     6     3     4     4     4     3     1
## Proportion 0.033 0.100 0.033 0.200 0.100 0.133 0.133 0.133 0.100 0.033
## 
## For the frequency table, variable is rounded to the nearest 0.09
## --------------------------------------------------------------------------------
## BirdWeight 
##        n  missing distinct     Info     Mean      Gmd 
##       30       17        7    0.899       12    1.802 
## 
## Value      10.00 10.98 11.96 12.94 14.97 15.95 17.00
## Frequency      3    13     8     1     3     1     1
## Proportion 0.100 0.433 0.267 0.033 0.100 0.033 0.033
## 
## For the frequency table, variable is rounded to the nearest 0.07
## --------------------------------------------------------------------------------
## TailLen 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##       30       17       10    0.975    57.23    2.834    54.00    54.00 
##      .25      .50      .75      .90      .95 
##    55.00    57.00    60.00    60.00    60.55 
## 
## Value      53.00 53.99 54.98 55.97 56.96 57.95 58.94 59.93 60.92 62.00
## Frequency      1     3     5     4     5     2     1     7     1     1
## Proportion 0.033 0.100 0.167 0.133 0.167 0.067 0.033 0.233 0.033 0.033
## 
## For the frequency table, variable is rounded to the nearest 0.09
## --------------------------------------------------------------------------------
## TarsusLen
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       30       17       17    0.995    18.25   0.7779    16.90    17.08
##      .25      .50      .75      .90      .95
##    17.92    18.25    18.67    19.00    19.11
##
## Value      16.800 16.881 17.097 17.583 17.880 17.988 18.096 18.177 18.285
## Frequency       1      2      1      1      3      2      3      2      1
## Proportion  0.033  0.067  0.033  0.033  0.100  0.067  0.100  0.067  0.033
##
## Value      18.393 18.474 18.582 18.690 18.879 18.987 19.176 19.500
## Frequency       1      2      3      1      2      3      1      1
## Proportion  0.033  0.067  0.100  0.033  0.067  0.100  0.033  0.033
##
## For the frequency table, variable is rounded to the nearest 0.027
## --------------------------------------------------------------------------------
## BirdSex
##        n  missing distinct
##       30       17        2
##
## Value          F     M
## Frequency     17    13
## Proportion 0.567 0.433
## --------------------------------------------------------------------------------
```

```
# Use a table to display example categorical variable
chickadeeTableEscShi <- table(chickadeeData$EscShi)
chickadeeTableEscShi
```

```
##
##    0    1    2    4    8   14   17   22   28   30   31   33   54   56   61   63
##    2    2    1    2    2    1    1    1    1    1    1    1    1    1    1    1
##   78   86  105  112  113  173  180  191  203  232  248  261  288  305  379  486
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2
##  882  906 1228 1267 1294 1429 2269 3230 3771 6520
##    1    1    1    1    1    1    1    1    1    1
```

```
chickadeeTableEntero <- table(chickadeeData$Entero)
chickadeeTableEntero
```

```
##
##   0   1   2   3   4   7   8   9  10  11  12  13  15  19  20  22  28  31  39  45
##   4   1   1   1   1   1   1   4   2   1   1   3   1   1   3   2   2   1   1   2
##  48  49  63  84 202 286 288 374 546 557 615 687
##   1   1   2   1   1   1   1   1   1   1   1   1
```

The data consists of 47 observations with 12 variables: Site, Habitat, Source, EscShi, Entero, CommRich, PathRich, WingChord, BirdWeight, TailLen, TarsusLen, and BirdSex. Below are the summary statistics for key variables:

Categorical Variables:

Habitat: 3 types (Rural: 20, Semi-urban: 14, Urban: 13)

Source: 2 types (Feather: 30, Nest: 17)

CommRich: 2 types (High: 19, Low: 28)

BirdSex: 2 types (F: 17, M: 13) [missing: 17]

Numerical Variables:

EscShi: Min: 0, 1st Quartile: 25, Median: 112, Mean: 577.9, 3rd Quartile: 432.5, Max: 6520

Entero: Min: 0, 1st Quartile: 9, Median: 20, Mean: 92.74, 3rd Quartile: 48.5, Max: 687

PathRich: Min: 11, 1st Quartile: 36, Median: 41, Mean: 44.3, 3rd Quartile: 53.5, Max: 88

Variables with Missing Data:

WingChord: 30 non-missing (Min: 60, Median: 65, Max: 69) [missing: 17]

BirdWeight: 30 non-missing (Min: 10, Median: 11, Max: 17) [missing: 17]

TailLen: 30 non-missing (Min: 53, Median: 57, Max: 62) [missing: 17]

TarsusLen: 30 non-missing (Min: 16.8, Median: 18.25, Max: 19.5) [missing: 17]

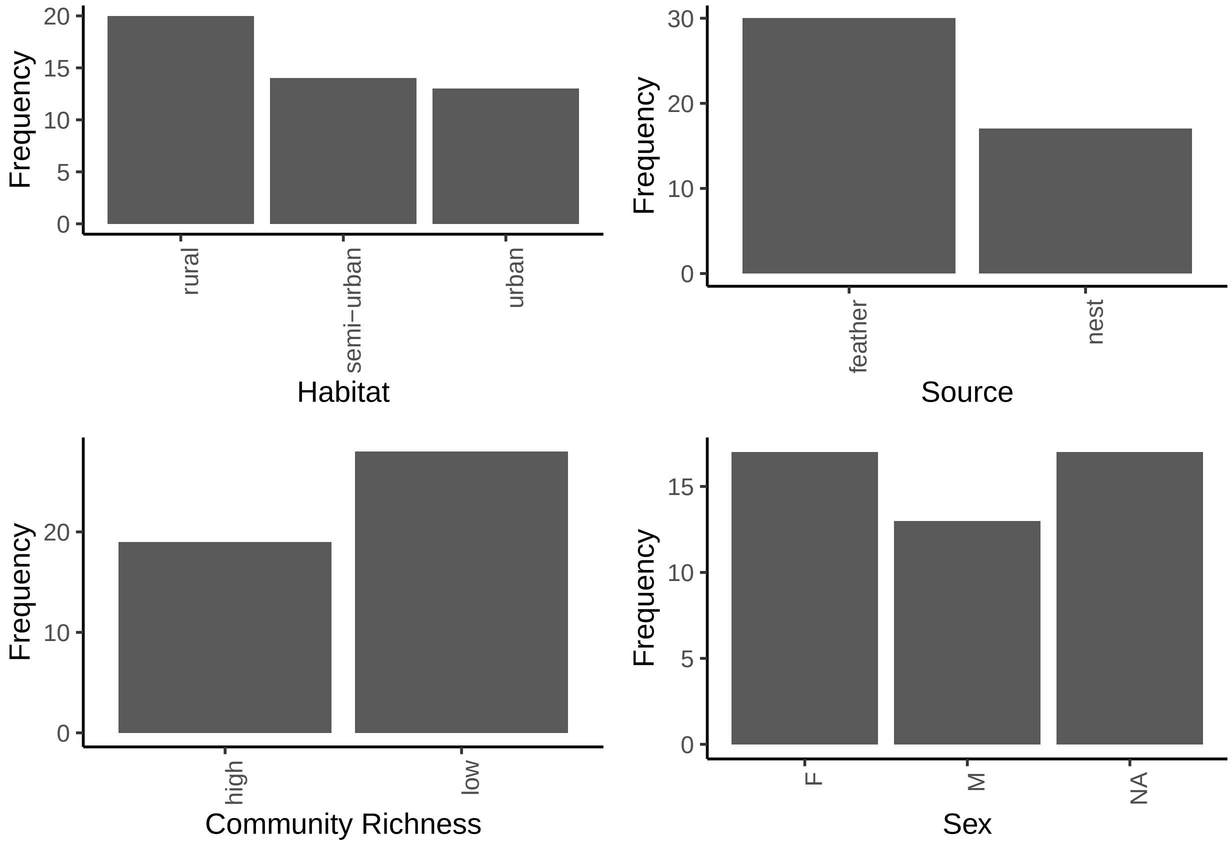
We find that 15.1% of the chickadee dataset is missing, while none of the dissimilarities dataset is.

The data shows a variety of bird habitats and sources, with most coming from feathers. The data also suggests wide variations in the numerical variables, like EscShi and Entero. There are also quite a few missing values for variables related to bird physical features, which should be taken into account in any subsequent analysis.

2. Graphical Summary

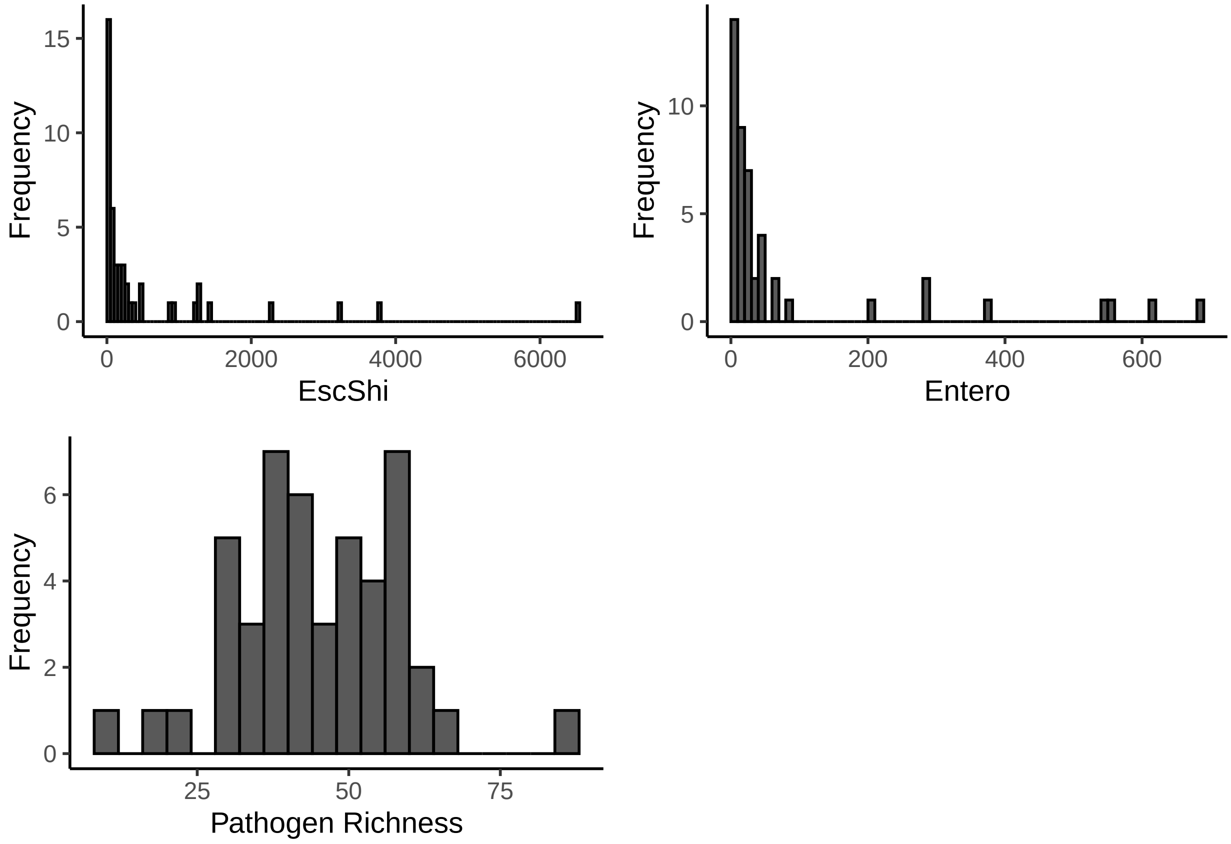
```
# Visualize frequency distributions of single variables

# Bar charts for categorical variables

# Visualize frequency distributions of single variables

# Bar charts for categorical variables

# Habitat
p1 <- ggplot(data = chickadeeData, aes(x = Habitat)) +
geom_bar(stat = "count") +
labs(x = "Habitat", y = "Frequency") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Source
p2 <- ggplot(data = chickadeeData, aes(x = Source)) +
geom_bar(stat = "count") +
labs(x = "Source", y = "Frequency") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))

# CommRich
p3 <- ggplot(data = chickadeeData, aes(x = CommRich)) +
geom_bar(stat = "count") +
```

```r
labs(x = "Community Richness", y = "Frequency") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))

# BirdSex
p4 <- ggplot(data = chickadeeData, aes(x = BirdSex)) +
geom_bar(stat = "count") +
labs(x = "Sex", y = "Frequency") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Arrange the plots on one or more pages
grid.arrange(p1, p2, p3, p4, ncol = 2)
```



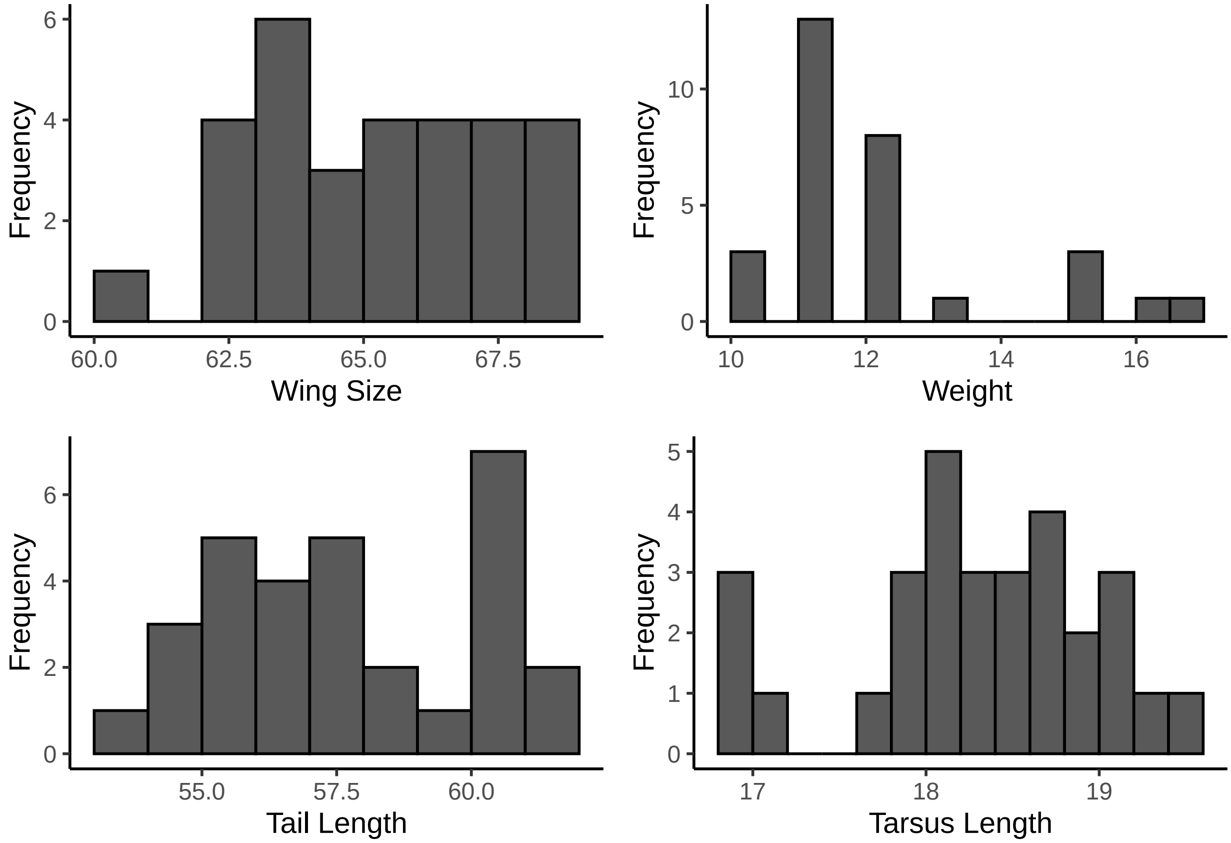
```r
# Histograms for numerical variables

# EscShi
p5 <- ggplot(data = chickadeeData, aes(x = EscShi)) +
geom_histogram(col = "black", binwidth = 50,
boundary = 0, closed = "left") +
labs(x = "EscShi", y = "Frequency")

# Entero
p6 <- ggplot(data = chickadeeData, aes(x = Entero)) +
geom_histogram(col = "black", binwidth = 10,
boundary = 0, closed = "left") +
labs(x = "Entero", y = "Frequency")
```

```
# PathRich
p7 <- ggplot(data = chickadeeData, aes(x = PathRich)) +
geom_histogram(col = "black", binwidth = 4,
boundary = 0, closed = "left") +
labs(x = "Pathogen Richness", y = "Frequency")

grid.arrange(p5, p6, p7, ncol = 2)
```

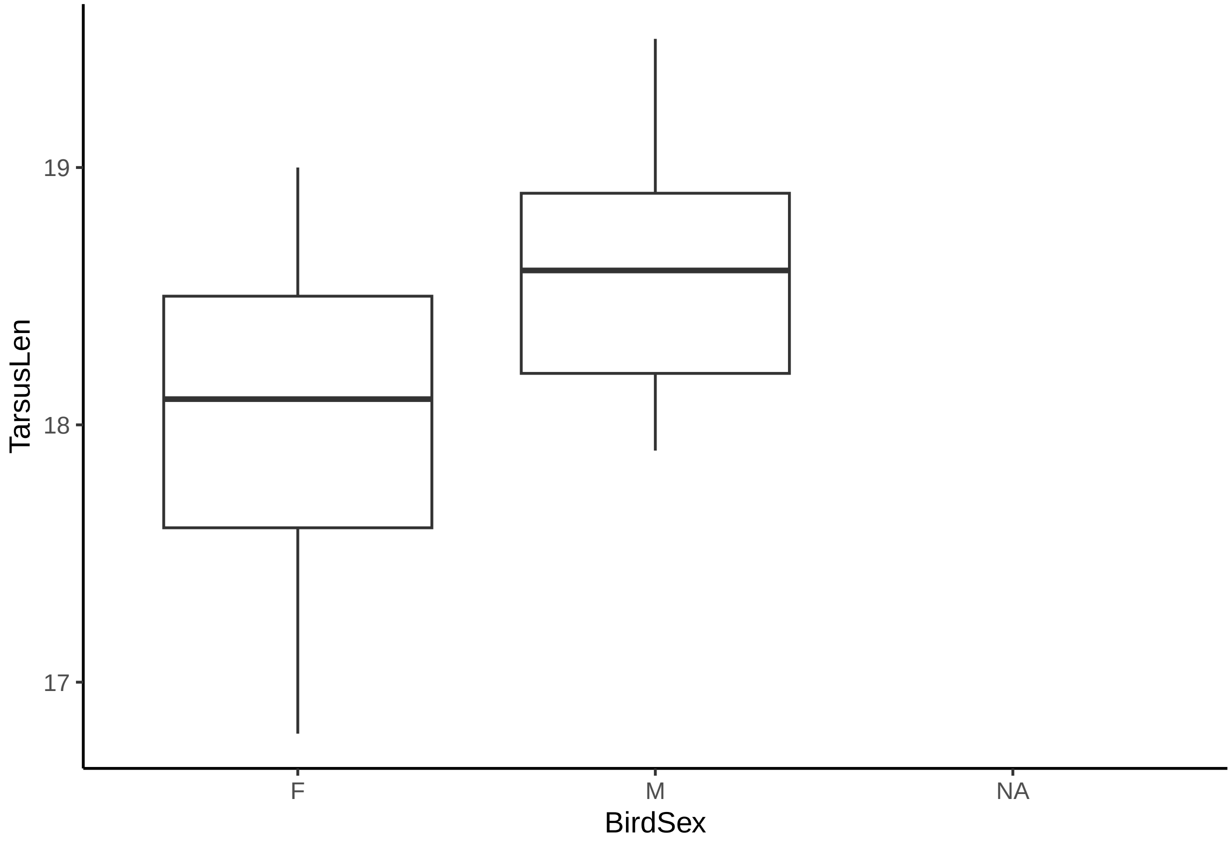


```
# WingChord
p8 <- ggplot(data = chickadeeData, aes(x = WingChord)) +
geom_histogram(col = "black", binwidth = 1,
boundary = 0, closed = "left") +
labs(x = "Wing Size", y = "Frequency")

# BirdWeight
p9 <- ggplot(data = chickadeeData, aes(x = BirdWeight)) +
geom_histogram(col = "black", binwidth = 0.5,
boundary = 0, closed = "left") +
labs(x = "Weight", y = "Frequency")

# TailLen
p10 <- ggplot(data = chickadeeData, aes(x = TailLen)) +
geom_histogram(col = "black", binwidth = 1,
boundary = 0, closed = "left") +
labs(x = "Tail Length", y = "Frequency")
```

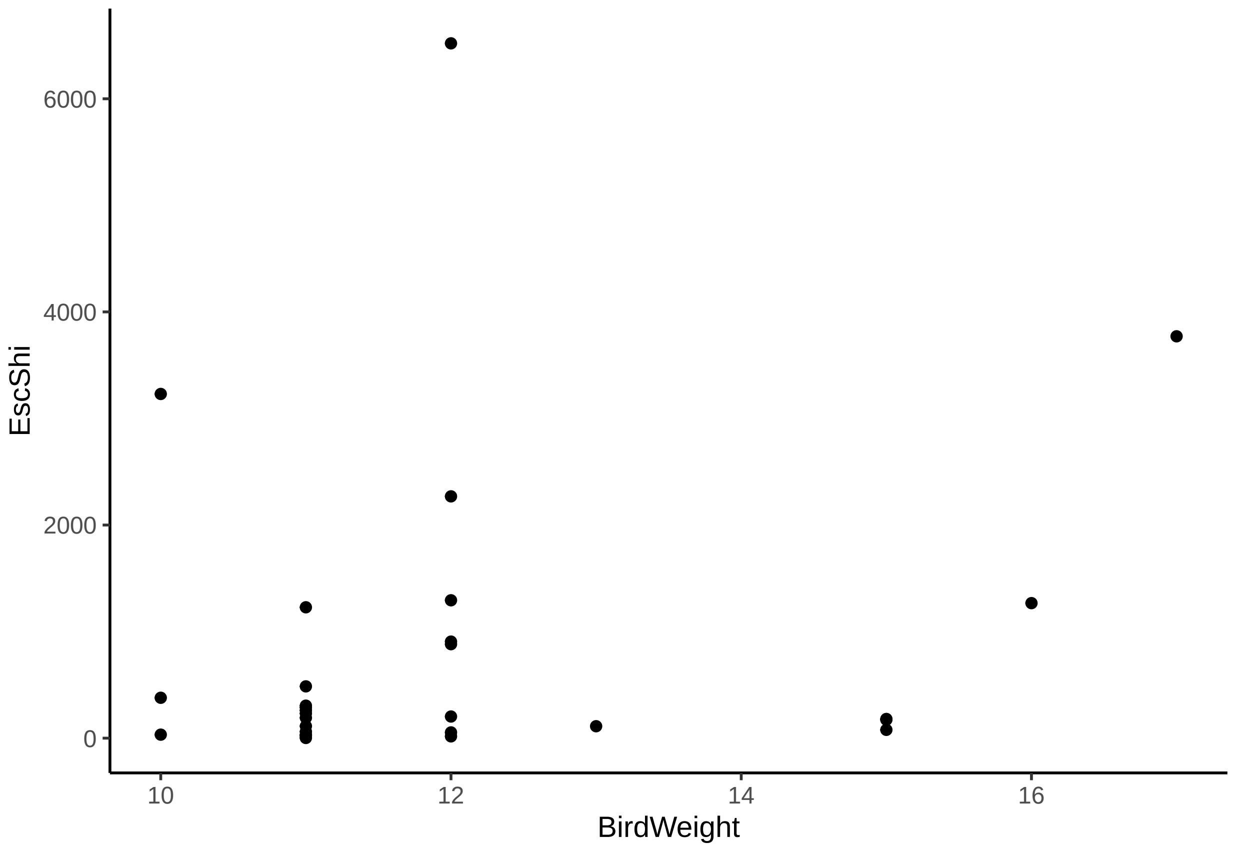
```r
# TarsusLen
p11 <- ggplot(data = chickadeeData, aes(x = TarsusLen)) +
geom_histogram(col = "black", binwidth = 0.2,
boundary = 0, closed = "left") +
labs(x = "Tarsus Length", y = "Frequency")

grid.arrange(p8, p9, p10, p11, ncol = 2)
```



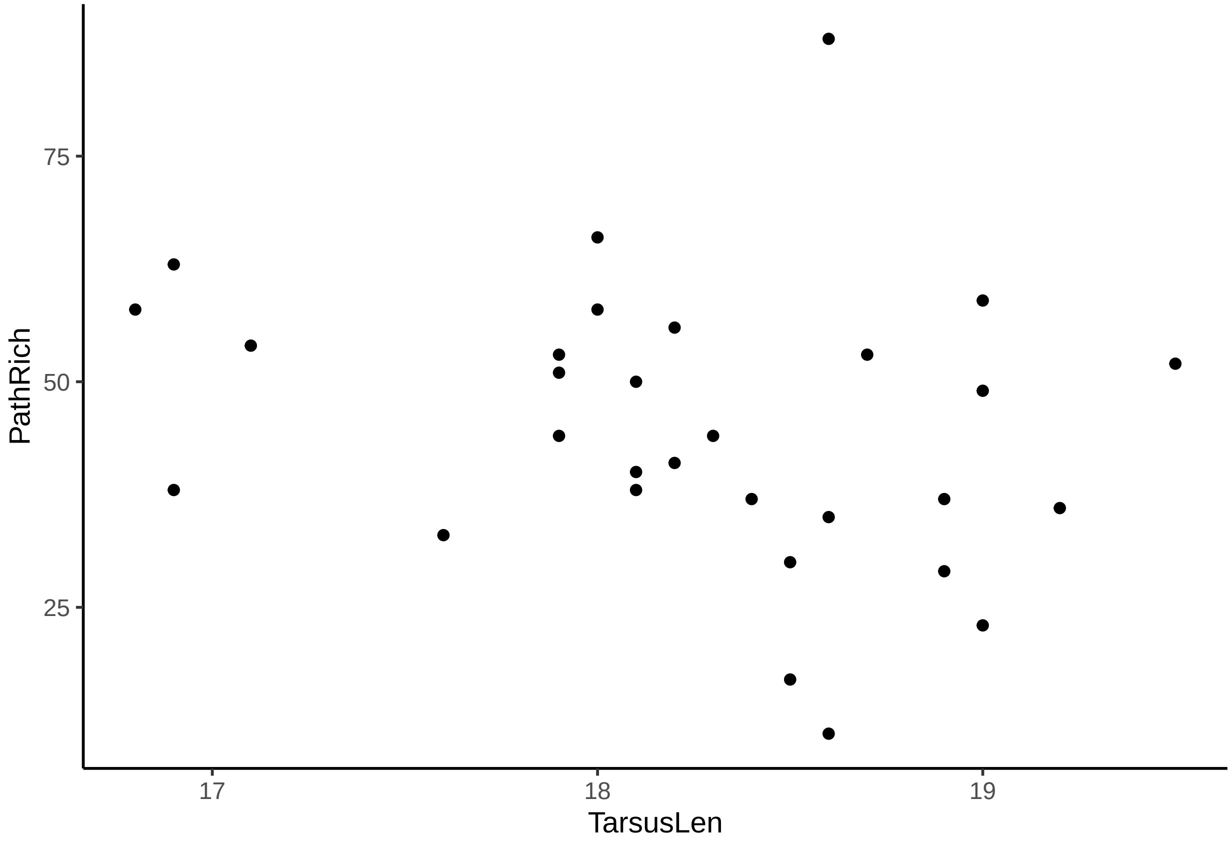
```r
# Relationship between 2 variables

# Box Plot for numerical variable vs categorical variable
ggplot(chickadeeData, aes(x=BirdSex, y=TarsusLen)) + geom_boxplot()
```



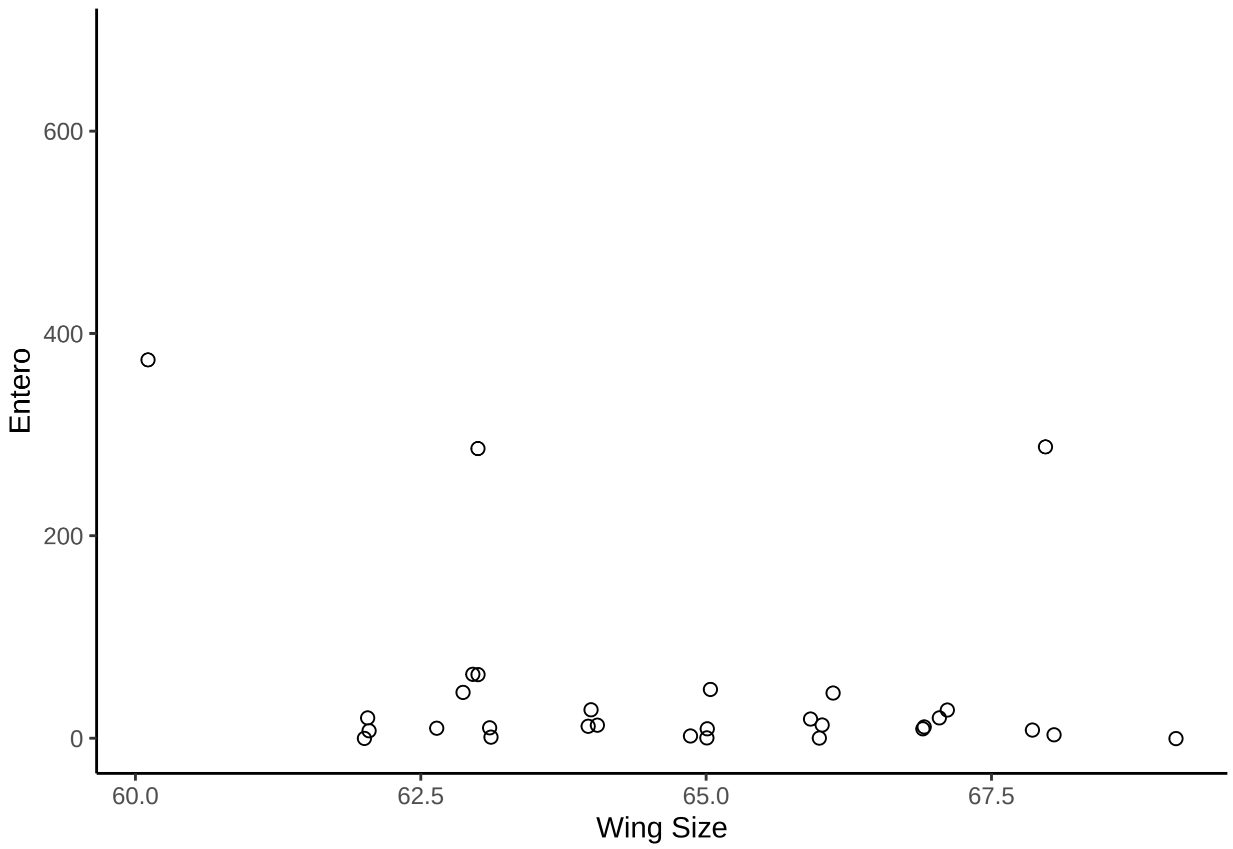
```
# Scatter Plots
ggplot(chickadeeData,
  aes(x = BirdWeight,
  y = EscShi)) +
  geom_point()
```



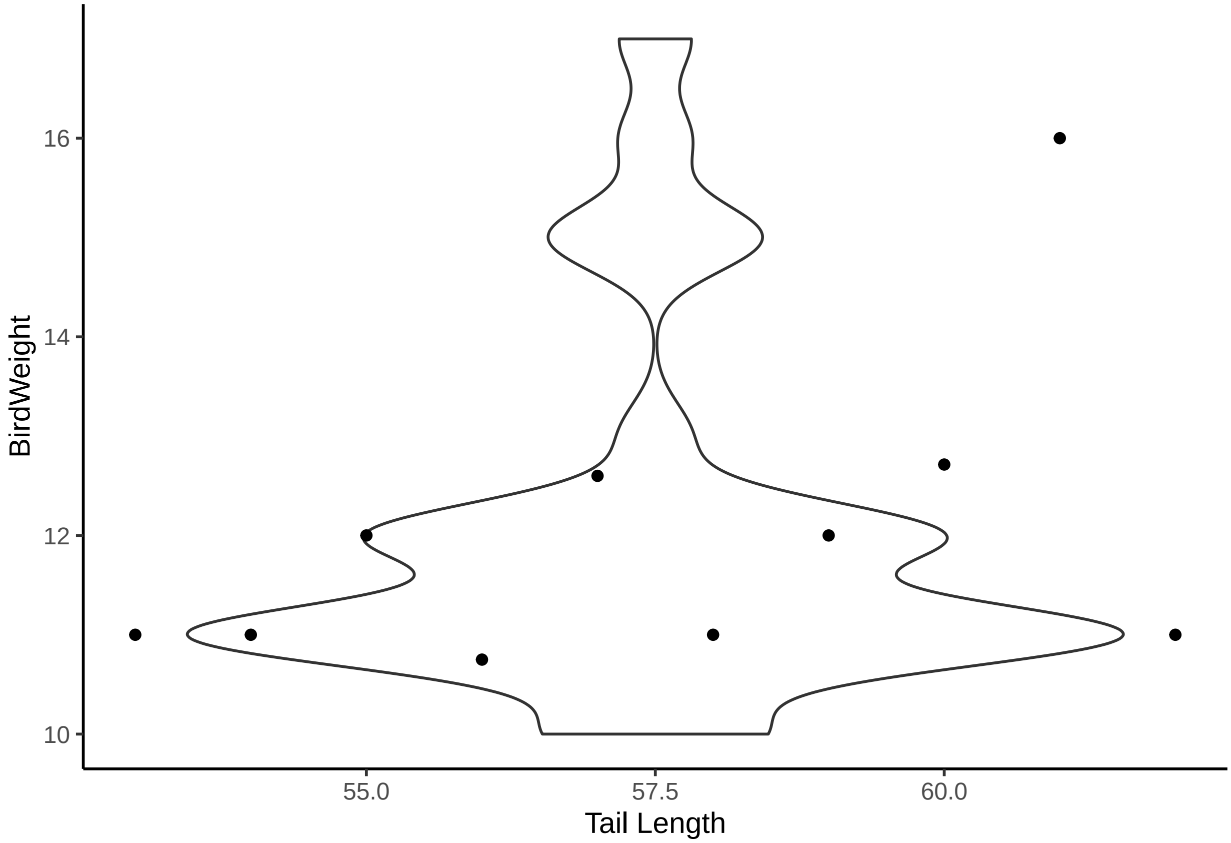
```
ggplot(chickadeeData,
  aes(x = TarsusLen,
  y = PathRich)) +
  geom_point()
```



```
# Strip Chart
ggplot (data = chickadeeData, aes(x = WingChord, y = Entero)) +
geom_jitter(shape = 1, size = 2, width = 0.15) +
labs(x = "Wing Size", y = "Entero")
```



```
# Violin Plot
ggplot(data = chickadeeData, aes(y = BirdWeight, x = TailLen)) +
geom_violin() +
labs(x = "Tail Length", y = "BirdWeight") +
stat_summary(fun = mean, geom = "point")
```



Upon visual inspection, we find that *Escherichia/Shigella* and *Enterococcus* distributions are right-skewed, and that pathogen richness, wing size, tail length, and tarsus length are approximately normally distributed.

3. Two-Sample *t*-Test For *Escherichia/Shigella* Between Nest and Feathers

```r
# Natural Log transformation + 1 (to handle zeroes)
EscShi_transformed <- log(chickadeeData$EscShi + 1)

# Two sample t-test of transformed data, assuming equal group variance
t.test(EscShi_transformed ~ Source, data = chickadeeData, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  EscShi_transformed by Source
## t = 4.0756, df = 45, p-value = 0.0001842
## alternative hypothesis: true difference in means between group feather and group nest is not equal t
## 95 percent confidence interval:
##  1.213690 3.585231
## sample estimates:
## mean in group feather    mean in group nest
##              5.425583              3.026123
```

In the two-sample t-test conducted to determine whether the mean abundance of bacteria identified as genus *Escherichia/Shigella* differs significantly between nests and feathers, our null hypothesis is that the true difference in means between the feather group and the nest group is equal to 0. The alternative hypothesis is that this difference is not equal to 0.

Based on the test results for the natural logarithm-transformed variable, the test statistic t is 4.0756, and the p-value is 0.0001842. Given that the p-value is less than the alpha level of 0.05, we reject the null hypothesis.

We conclude that there is a statistically significant difference in the mean abundance of bacteria identified as genus *Escherichia/Shigella* between nests and feathers, based on this data.

4. Mann-Whitney U-Test For *Escherichia/Shigella* Between Nest and Feathers

```
# Mann-Whitney U-Test
wilcox.test(EscShi ~ Source, data = chickadeeData)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  EscShi by Source
## W = 405.5, p-value = 0.0008952
## alternative hypothesis: true location shift is not equal to 0
```

In the Mann-Whitney U-test conducted to determine whether the population distribution of the abundance of bacteria identified as genus *Escherichia/Shigella* differs significantly between nests and feathers, our null hypothesis is that the true location shift between the feather group and the nest group is equal to 0. The alternative hypothesis is that this location shift is not equal to 0.

Based on the test results for the untransformed data, the test statistic W is 405.5, and the p-value is 0.0008952. Since the p-value is less than the alpha level of 0.05, we reject the null hypothesis.

We conclude that there is statistically significant evidence to suggest that the distributions of abundance for bacteria identified as genus *Escherichia/Shigella* differ between nests and feathers, based on this data.

The Mann-Whitney U-test is useful in this context as it does not require the normal distribution assumption that the two-sample t-test does. Therefore, it allows us to extend our conclusions to the original, untransformed population, providing further evidence for a difference in bacterial abundance between the two groups.

5. Two-Sample *t*-Test For *Enterococcus* Between Nest and Feathers

```
# Natural Log transformation + 1 (to handle zeroes)
Entero_transformed <- log(chickadeeData$Entero + 1)

# Two sample t-test of transformed data, assuming equal group variance
t.test(Entero_transformed ~ Source, data = chickadeeData, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Entero_transformed by Source
## t = -2.6195, df = 45, p-value = 0.01196
## alternative hypothesis: true difference in means between group feather and group nest is not equal to
## 95 percent confidence interval:
##  -2.2836566 -0.2983701
## sample estimates:
## mean in group feather    mean in group nest
##              2.670757              3.961771
```

In the two-sample t-test conducted to assess whether the mean abundance of bacteria identified as genus *Enterococcus* differs significantly between nests and feathers, our null hypothesis is that the true difference in means between the feather group and the nest group is equal to 0. The alternative hypothesis is that this difference is not equal to 0.

Based on the results of the test on the natural logarithm transformed data, the test statistic t is -2.6195, and the p-value is 0.01196. Since the p-value is less than the alpha level of 0.05, we reject the null hypothesis.

We conclude that there is statistically significant evidence to suggest that the mean abundance of bacteria identified as genus *Enterococcus* differs between nests and feathers, based on this data. Additionally, the test indicates that the mean abundance of this bacterial genus is higher in nests compared to feathers.

6. Mann-Whitney U-Test For *Enterococcus* Between Nest and Feathers

```r
# Mann-Whitney U-Test
wilcox.test(Entero ~ Source, data = chickadeeData)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Entero by Source
## W = 148, p-value = 0.01826
## alternative hypothesis: true location shift is not equal to 0
```

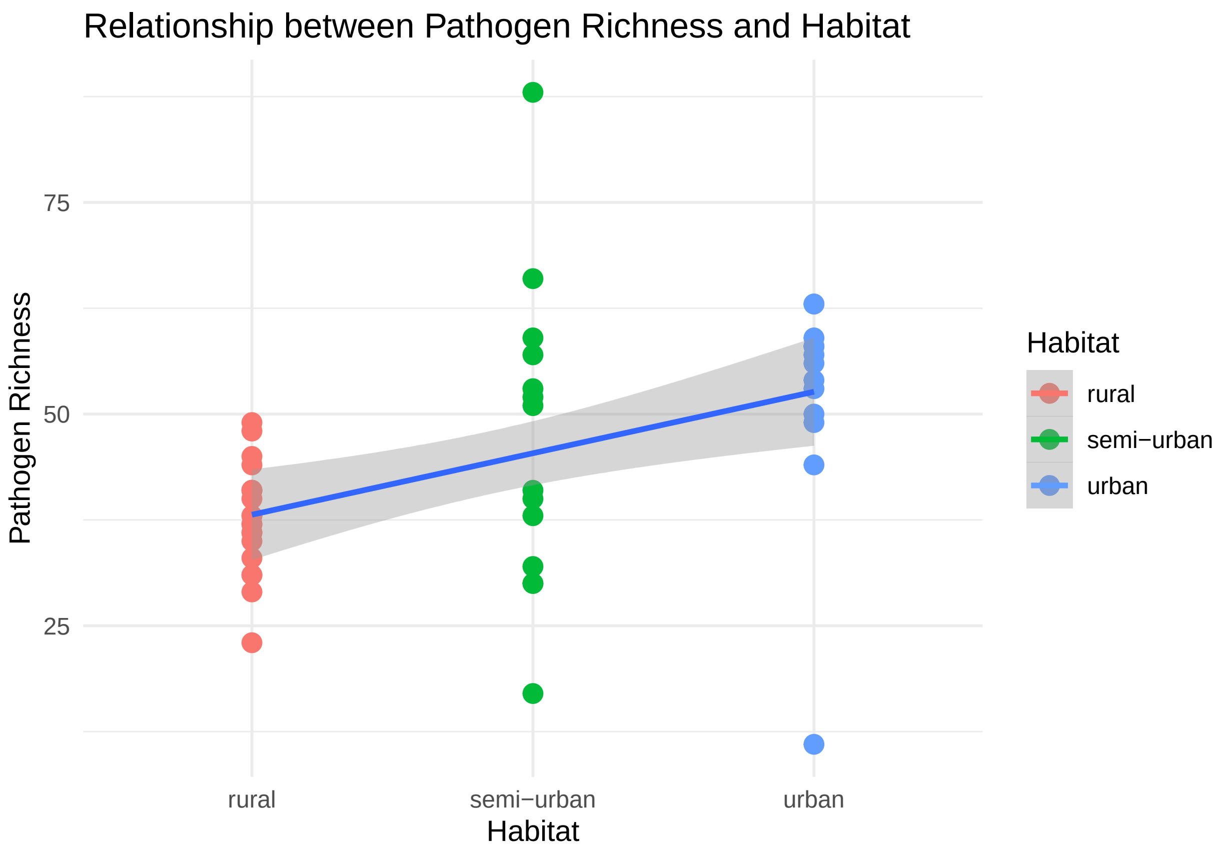
In the Mann-Whitney U-test to examine whether the population distribution of the abundance of bacteria identified as genus *Enterococcus* varies significantly between nests and feathers, our null hypothesis is that the true location shift between the feather and nest groups is equal to 0. The alternative hypothesis is that this location shift is not equal to 0.

From the test results, the test statistic W is 148, and the p-value is 0.01826. Given that the p-value is less than the alpha level of 0.05, we reject the null hypothesis.

We conclude that there is statistically significant evidence to suggest that the population distribution of the abundance of bacteria identified as genus *Enterococcus* differs between nests and feathers. Furthermore, based on this evidence and previous data, we infer that the abundance of *Escherichia/Shigella* is higher in feathers, while the abundance of *Enterococcus* is higher in nests.

7. Analysis of Variance for Pathogen Richness and Habitat

```
# Scatter Plot with linear model
ggplot(chickadeeData, aes(x=Habitat, y=PathRich)) +
  geom_point(aes(color=Habitat), size=3) +            # Scatter plot
  geom_smooth(method='lm', aes(group=1, color=Habitat)) + # Add a linear fit, colored by Habitat
  labs(title="Relationship between Pathogen Richness and Habitat",
       x="Habitat",
       y="Pathogen Richness") +
  theme_minimal()
```



```
# Simple Linear Regression
PathRich_Habitat_Regression <- lm(PathRich ~ Habitat, data = chickadeeData)
summary(PathRich_Habitat_Regression)
```

```
##
## Call:
## lm(formula = PathRich ~ Habitat, data = chickadeeData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.923  -6.182   1.077   6.181  41.286
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           37.650      2.856  13.181  < 2e-16 ***
## Habitatsemi-urban      9.064      4.451   2.036  0.04777 *
## Habitaturban          14.273      4.551   3.136  0.00305 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 12.77 on 44 degrees of freedom
## Multiple R-squared:  0.1934, Adjusted R-squared:  0.1567
## F-statistic: 5.275 on 2 and 44 DF,  p-value: 0.008844
```

```
# ANOVA
anova(PathRich_Habitat_Regression)
```

```
## Analysis of Variance Table
##
## Response: PathRich
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Habitat    2 1721.5  860.75  5.2745 0.008844 **
## Residuals 44 7180.3  163.19
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

In the analysis of variance (ANOVA), we assess whether the mean pathogen richness is significantly related to habitat. The null hypothesis is that there is no relationship between habitat and pathogen richness, while the alternative hypothesis is that such a relationship does exist.

According to the linear model results, the F-statistic is 5.275 with a corresponding p-value of 0.008844. Given that this p-value is below the alpha level of 0.05, we reject the null hypothesis.

The model shows statistically significant coefficients for both semi-urban and urban habitats, with p-values of 0.04777 and 0.00305 respectively. This suggests that the type of habitat is a meaningful predictor of pathogen richness.

The Multiple R^2 value of 0.1934 indicates that approximately 19.34% of the variability in pathogen richness is explained by habitat. While statistically significant, the model accounts for less than one-fifth of the variance, suggesting that other factors may also be influential.

We conclude that habitat is a statistically significant predictor of pathogen richness, although it explains only a limited portion of the variability. The trend in the estimates of coefficients implies that pathogen richness tends to increase in more densely populated human environments, such as semi-urban and urban areas, based on this data.

8. Kruskal-Wallis Test for Pathogen Richness and Habitat

```
# Kruskal-Wallis Test
kruskal.test(PathRich ~ Habitat, data = chickadeeData)
```

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  PathRich by Habitat
## Kruskal-Wallis chi-squared = 13.587, df = 2, p-value = 0.001121
```

In the Kruskal-Wallis test, we assess whether mean pathogen richness varies significantly across different habitats. The null hypothesis is that there are no differences in mean pathogen richness across habitats, while the alternative hypothesis is that at least one habitat exhibits a difference in mean pathogen richness.

The Kruskal-Wallis X^2 test statistic is 13.587, with 2 degrees of freedom, and a p-value of 0.001121. Given that the p-value is well below the alpha level of 0.05, we reject the null hypothesis.

We conclude that there is a statistically significant difference in mean pathogen richness across habitats, based on this data. This evidence aligns with the findings from the linear model, supporting the significance of habitat as a factor influencing pathogen richness.

9. Tukey-Kramer Test for Pathogen Richness and Specific Habitat

```r
# Tukey-Kramer test
pathogenPairs <- emmeans(PathRich_Habitat_Regression, specs = "Habitat")
pathogenUnplanned <- contrast(pathogenPairs, method = "pairwise",
adjust = "tukey")

pathogenUnplanned
```

```
##  contrast              estimate   SE df t.ratio p.value
##  rural - (semi-urban)     -9.06 4.45 44  -2.036  0.1155
##  rural - urban           -14.27 4.55 44  -3.136  0.0084
##  (semi-urban) - urban     -5.21 4.92 44  -1.059  0.5445
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

In the Tukey-Kramer post-hoc test, we assess which specific habitats differ significantly in terms of mean pathogen richness. The test provides pairwise comparisons between the three habitat types: rural, semi-urban, and urban.

Rural vs. Semi-Urban: The estimated mean difference is -9.06, with a p-value of 0.1155. Since this p-value exceeds the alpha level of 0.05, the difference in pathogen richness between rural and semi-urban habitats is not statistically significant.

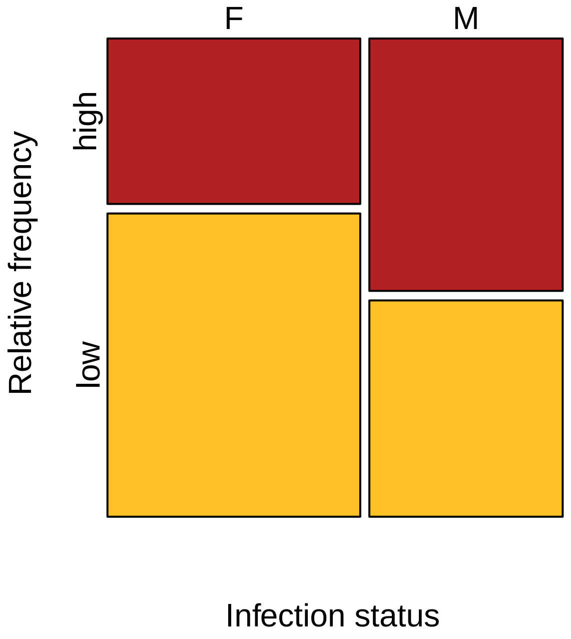
Rural vs. Urban: The estimated mean difference is -14.27, with a p-value of 0.0084. This p-value is less than the alpha level of 0.05, indicating a statistically significant difference in pathogen richness between rural and urban habitats.

Semi-Urban vs. Urban: The estimated mean difference is -5.21, with a p-value of 0.5445. As this p-value is greater than 0.05, the difference in pathogen richness between semi-urban and urban habitats is not statistically significant.

Based on the Tukey-Kramer test, we can conclude that pathogen richness is significantly higher in urban areas compared to rural areas. However, no statistically significant differences in pathogen richness were found between rural and semi-urban habitats, or between semi-urban and urban habitats. These findings are consistent with those from the linear regression model and the Kruskal-Wallis test, further substantiating the influence of habitat on pathogen richness.

10. Contingency Table Analysis and Chi-Squared Test for Community Richness and Sex

```
# Filter df for Source = feather
df_feather <- filter(chickadeeData, Source == 'feather')
#df_feather

# Create 2x2 Contingency Table for community richness and sex
CommRich_BirdSex_table <- table(df_feather$CommRich, df_feather$BirdSex)
#CommRich_BirdSex_table
#str(CommRich_BirdSex_table)

# Create a mosaic plot to visualize the data
par(pty = "s") # makes a square plot
mosaicplot(t(CommRich_BirdSex_table), col = c("firebrick", "goldenrod1"),
cex.axis = 1, main = "",
sub = "Infection status", ylab = "Relative frequency")
```



```
# Conduct a Chi-squared Test of Independence of Two Categorical Variables
Xsq <- chisq.test(CommRich_BirdSex_table, correct = FALSE)
Xsq
```

```
##
##  Pearson's Chi-squared test
##
## data:  CommRich_BirdSex_table
## X-squared = 1.0325, df = 1, p-value = 0.3096
```

In this contingency table analysis using Pearson's Chi-squared test, we assess whether community richness on feathers is independent of the sex of mountain chickadees. The null hypothesis is that community richness and the sex of the bird are independent variables, while the alternative hypothesis is that they are not independent.

According to the test results, the Chi-squared statistic X^2 is 1.0325 with 1 degree of freedom, and the p-value is 0.3096. Given that the p-value exceeds the alpha level of 0.05, we fail to reject the null hypothesis.

We conclude that there is insufficient evidence to suggest that community richness on feathers is dependent on the sex of mountain chickadees, based on this data.

11. Linear Regression of Morphological Features and Pathogen Richness

```r
# Fit a linear model to each relationship
fit_WingChord <- lm(df_feather$PathRich ~ df_feather$WingChord)
fit_BirdWeight <- lm(df_feather$PathRich ~ df_feather$BirdWeight)
fit_TailLen <- lm(df_feather$PathRich ~ df_feather$TailLen)
fit_TarsusLen <- lm(df_feather$PathRich ~ df_feather$TarsusLen)

# Summarize the fits
summary(fit_WingChord)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$WingChord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.208  -9.151   0.889   7.422  41.663
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -38.829     83.710  -0.464    0.646
## df_feather$WingChord      1.290      1.291   0.999    0.326
##
## Residual standard error: 15.65 on 28 degrees of freedom
## Multiple R-squared:  0.03443,    Adjusted R-squared:  -5.351e-05
## F-statistic: 0.9984 on 1 and 28 DF,  p-value: 0.3262
```

```r
summary(fit_BirdWeight)
```

```
##
## Call:
```

```
## lm(formula = df_feather$PathRich ~ df_feather$BirdWeight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.724  -7.775  -0.272   8.725  43.233
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            32.8943    19.7948   1.662    0.108
## df_feather$BirdWeight   0.9894     1.6319   0.606    0.549
##
## Residual standard error: 15.82 on 28 degrees of freedom
## Multiple R-squared:  0.01296,    Adjusted R-squared:  -0.02229
## F-statistic: 0.3676 on 1 and 28 DF,  p-value: 0.5492
```

```
summary(fit_TailLen)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$TailLen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.856  -6.229   1.144   8.374  37.291
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -78.153     64.848  -1.205   0.2382
## df_feather$TailLen    2.148      1.132   1.897   0.0682 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.99 on 28 degrees of freedom
## Multiple R-squared:  0.1139, Adjusted R-squared:  0.08226
## F-statistic: 3.599 on 1 and 28 DF,  p-value: 0.06816
```

```
summary(fit_TarsusLen)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$TarsusLen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.823  -7.761  -1.573  10.140  45.177
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           145.140     76.015   1.909   0.0665 .
## df_feather$TarsusLen   -5.501      4.163  -1.321   0.1971
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 15.45 on 28 degrees of freedom
## Multiple R-squared:  0.0587, Adjusted R-squared:  0.02508
## F-statistic: 1.746 on 1 and 28 DF,  p-value: 0.1971
```

```
# Alternatively, fit a multiple linear regression model
fit_all <- lm(PathRich ~ WingChord + BirdWeight + TailLen + TarsusLen, data = df_feather)
summary(fit_all)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + BirdWeight + TailLen + TarsusLen,
##     data = df_feather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.510  -7.162   0.677   7.818  39.183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.6554   105.1841   0.253    0.802
## WingChord    -0.3523     2.3143  -0.152    0.880
## BirdWeight    0.3688     1.8250   0.202    0.841
## TailLen       2.2744     1.9022   1.196    0.243
## TarsusLen    -5.1332     4.5403  -1.131    0.269
##
## Residual standard error: 15.35 on 25 degrees of freedom
## Multiple R-squared:  0.1708, Adjusted R-squared:  0.03816
## F-statistic: 1.288 on 4 and 25 DF,  p-value: 0.3015
```

In this analysis, simple linear regression was used to examine the linear relationship between four bird features (Wing Chord, Bird Weight, Tail Length, and Tarsus Length) and pathogen richness on feathers. Additionally, multiple linear regression was conducted to examine the combined effect of these predictors.

Wing Chord: With an R^2 of 0.0344 and a p-value of 0.3262, Wing Chord is not a statistically significant predictor of pathogen richness.

Bird Weight: Similarly, Bird Weight shows a low R^2 of 0.0130 and a high p-value of 0.5492, indicating it is also not a statistically significant predictor.

Tail Length: Among the features, Tail Length had the highest R^2 value of 0.1139 and the lowest p-value of 0.0682, which approaches significance. While not statistically significant at the alpha = 0.05 level, this variable shows the most promise as a predictor among those tested.

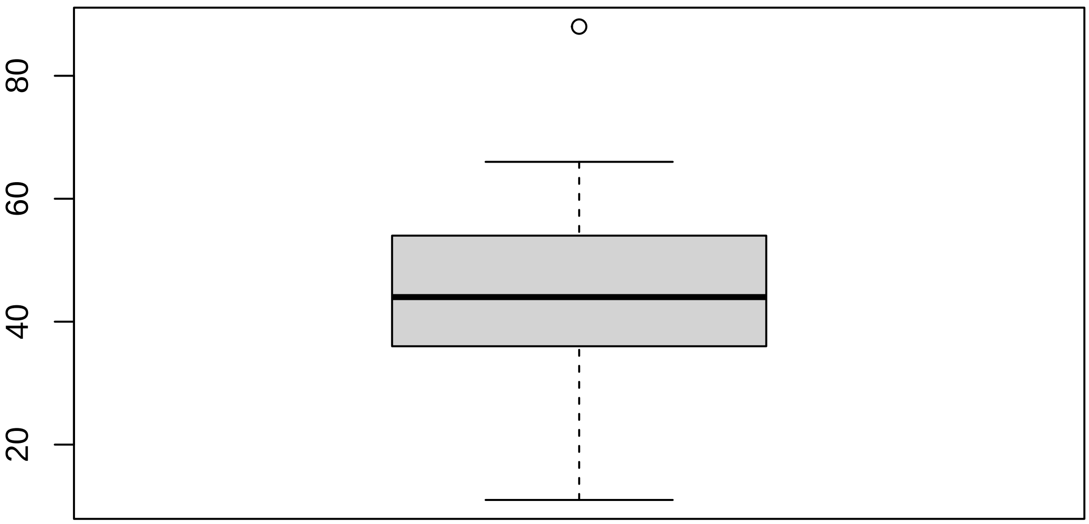
Tarsus Length: With an R^2 value of 0.0587 and a p-value of 0.1971, Tarsus Length was also not found to be a statistically significant predictor but does have a negative relationship with pathogen richness.

None of the single predictors are statistically significant based on their p-values exceeding the alpha level of 0.05.

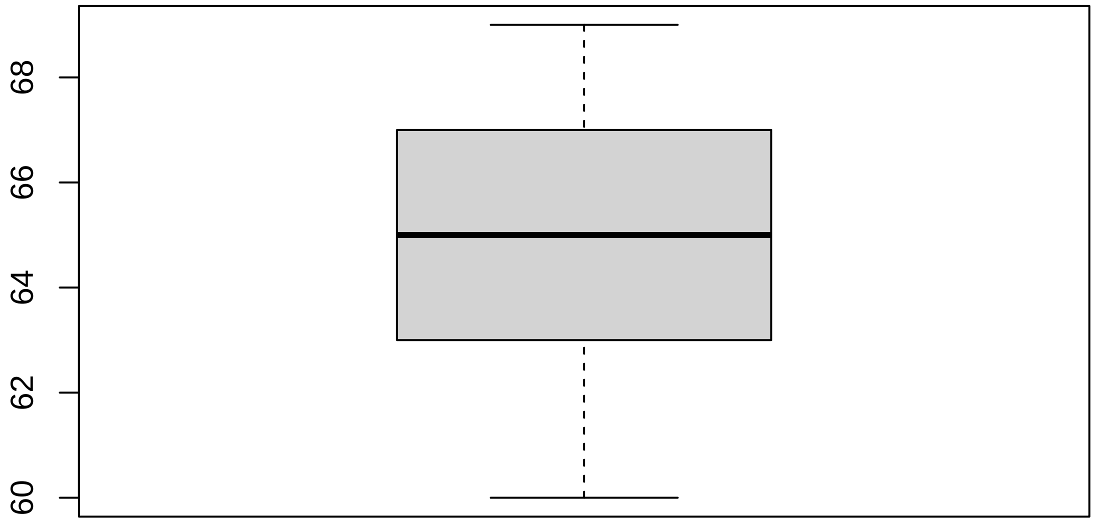
The multiple regression model yielded a p-value of 0.3015 and an R^2 of 0.1708, indicating that collectively, the predictors are not statistically significant at explaining the variability in pathogen richness. The multiple regression model also failed to reach statistical significance, supporting the null hypothesis that none of these predictors significantly influence pathogen richness on feathers.

12. Linear Regression of Morphological Features and Pathogen Richness With Outliers Removed

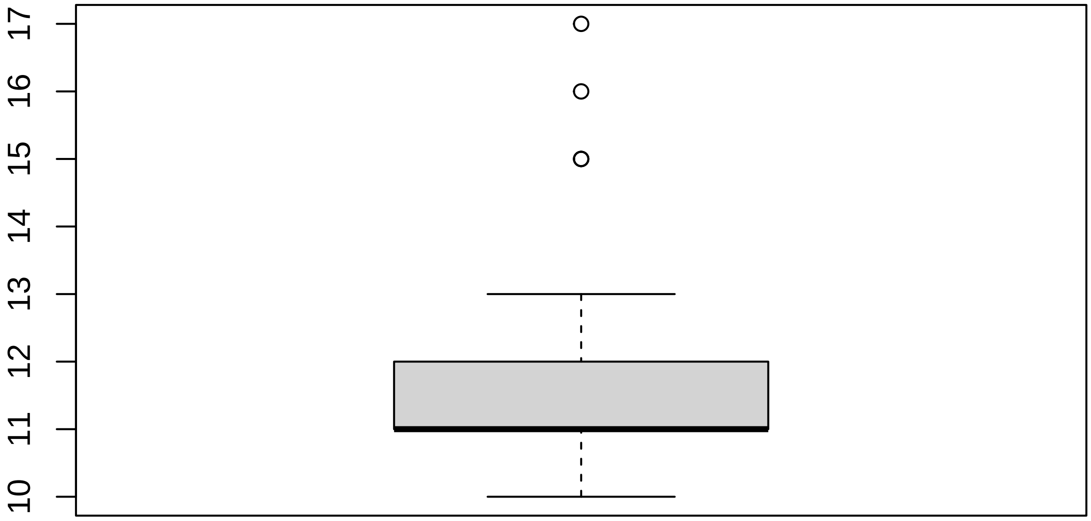
```
# Individually
boxplot(df_feather$PathRich) # Outlier > 60
```



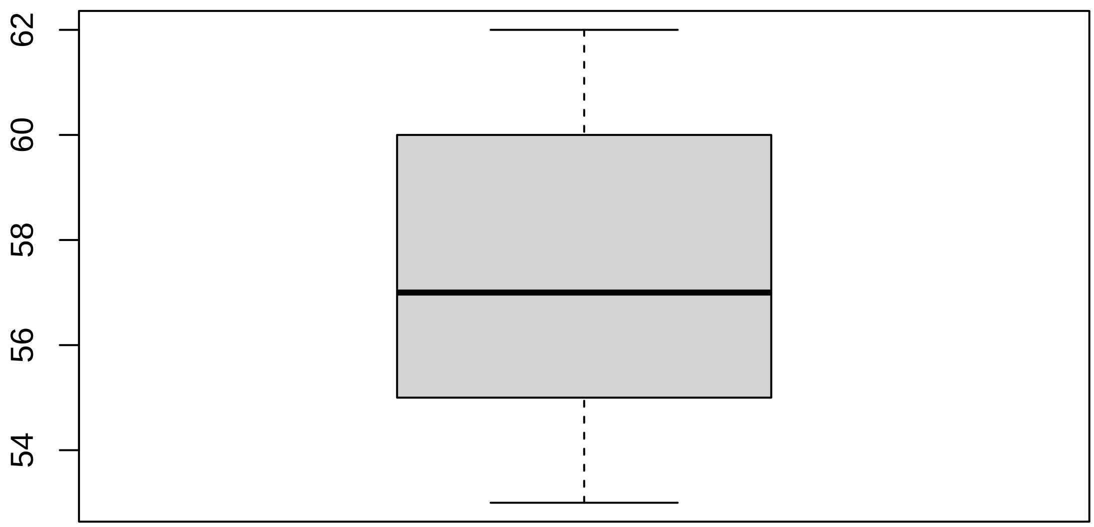
```
boxplot(df_feather$WingChord)
```



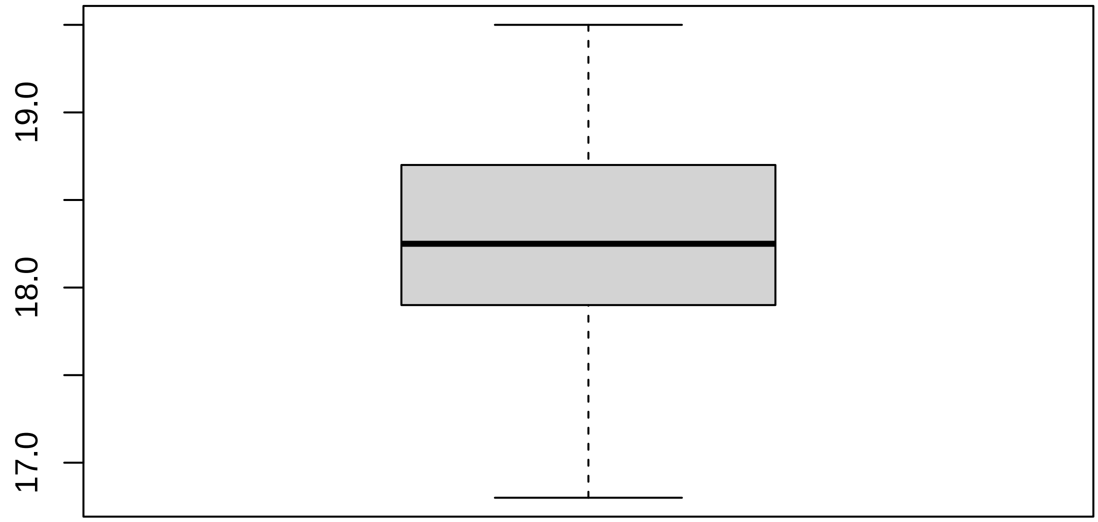
```
boxplot(df_feather$BirdWeight) # Outliers > 13
```



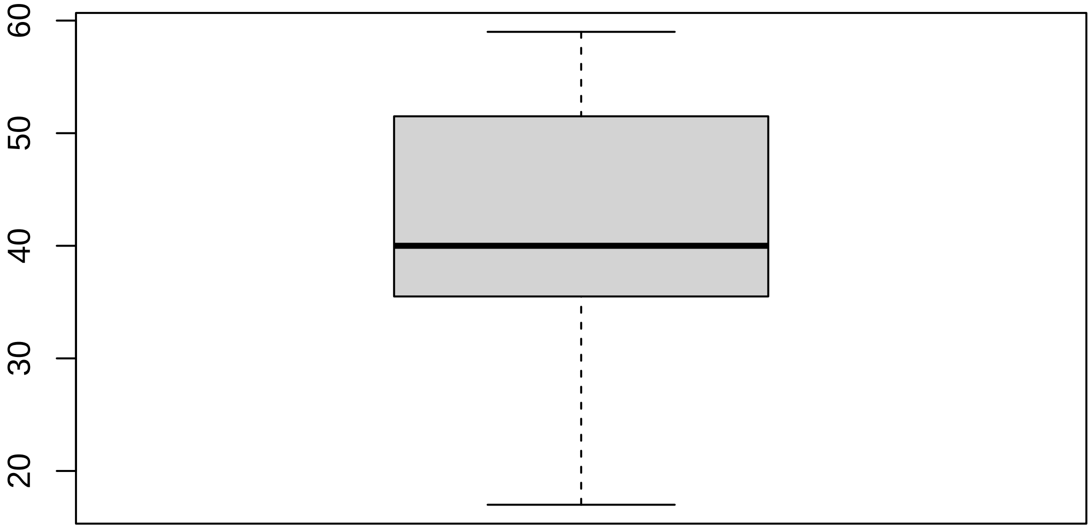
```
boxplot(df_feather$TailLen)
```



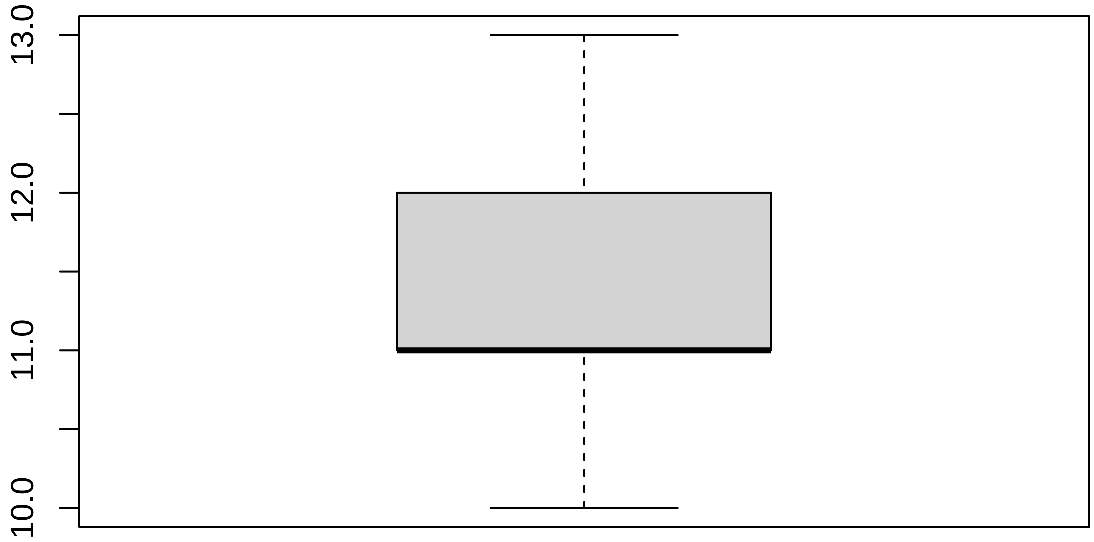
```
boxplot(df_feather$TarsusLen)
```



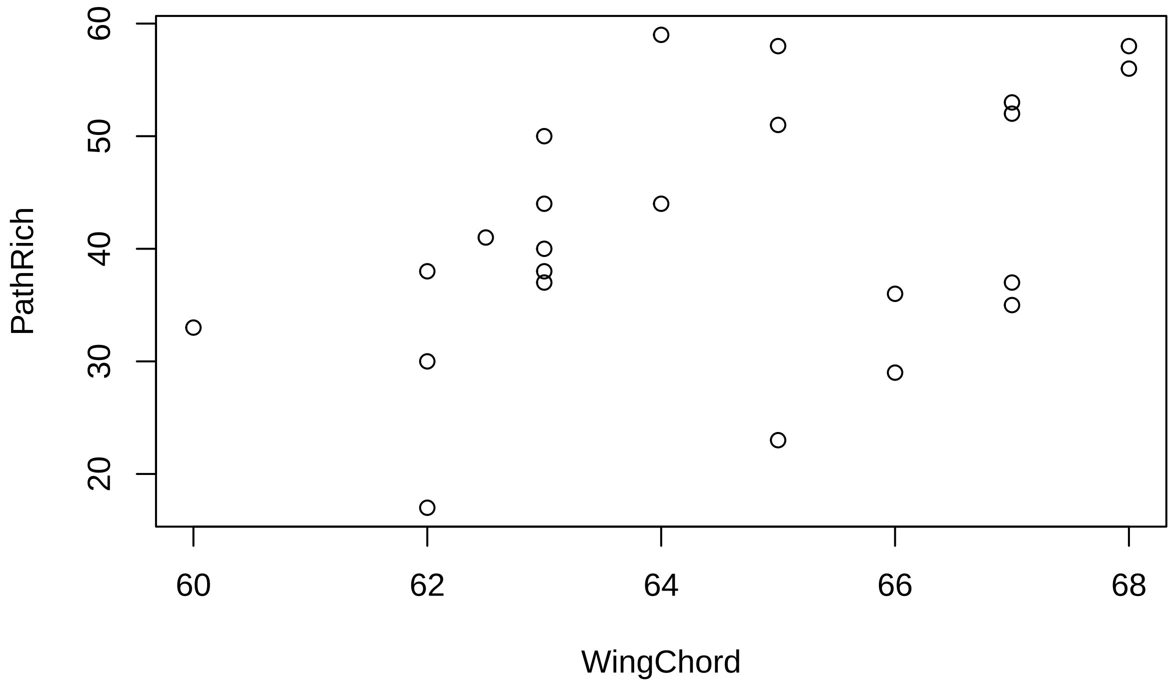
```
# Filter outliers
df_feather <- filter(df_feather, PathRich < 66) # Outlier > 60
df_feather <- filter(df_feather, BirdWeight < 14) # Outlier > 13

# Check if filters were successful
boxplot(df_feather$PathRich) # Outlier > 60
```



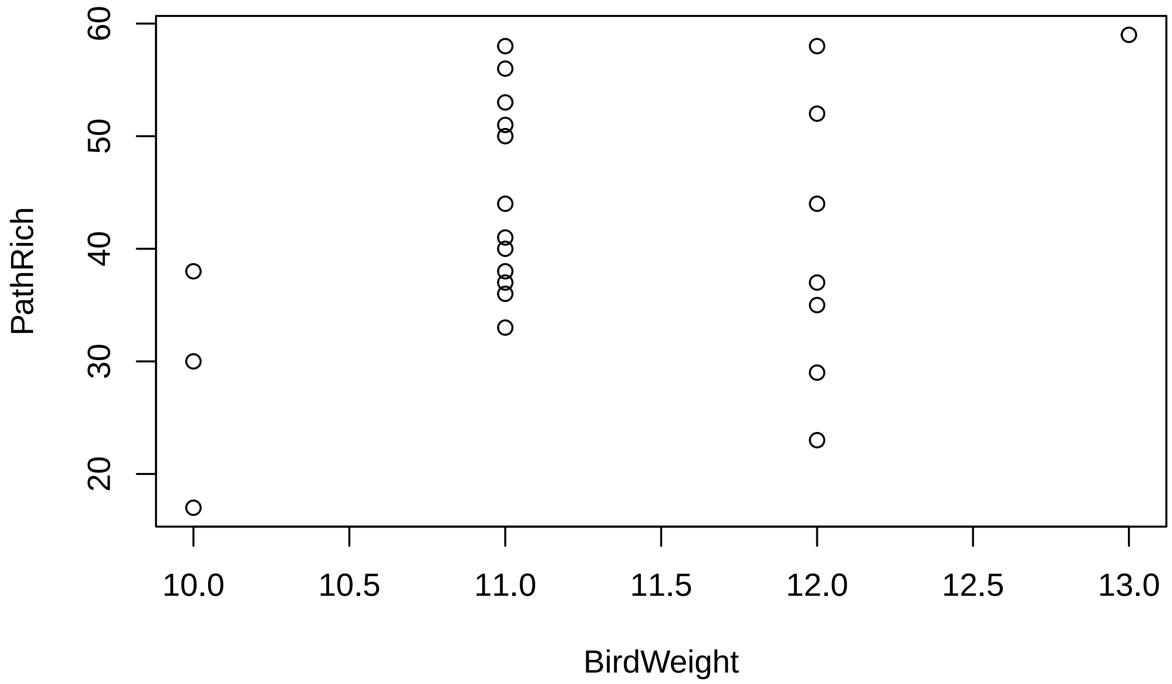
```
boxplot(df_feather$BirdWeight) # Outliers > 13
```



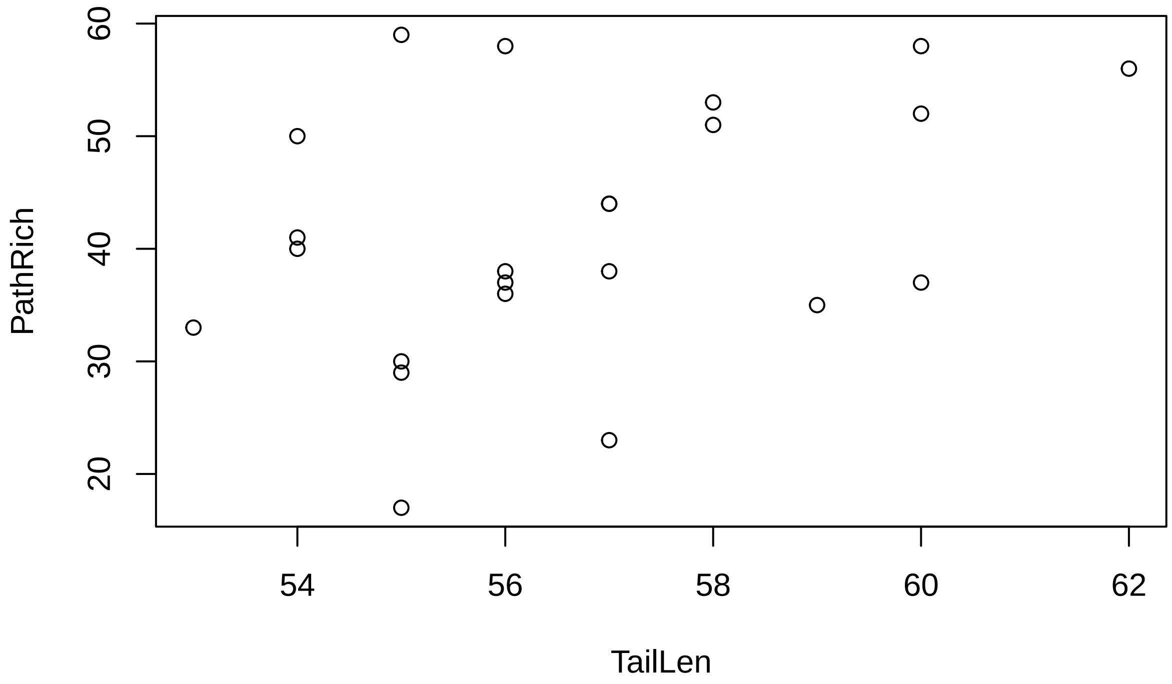
```
# Check Relationships for more potential outliers
plot(PathRich ~ WingChord, data = df_feather)
```



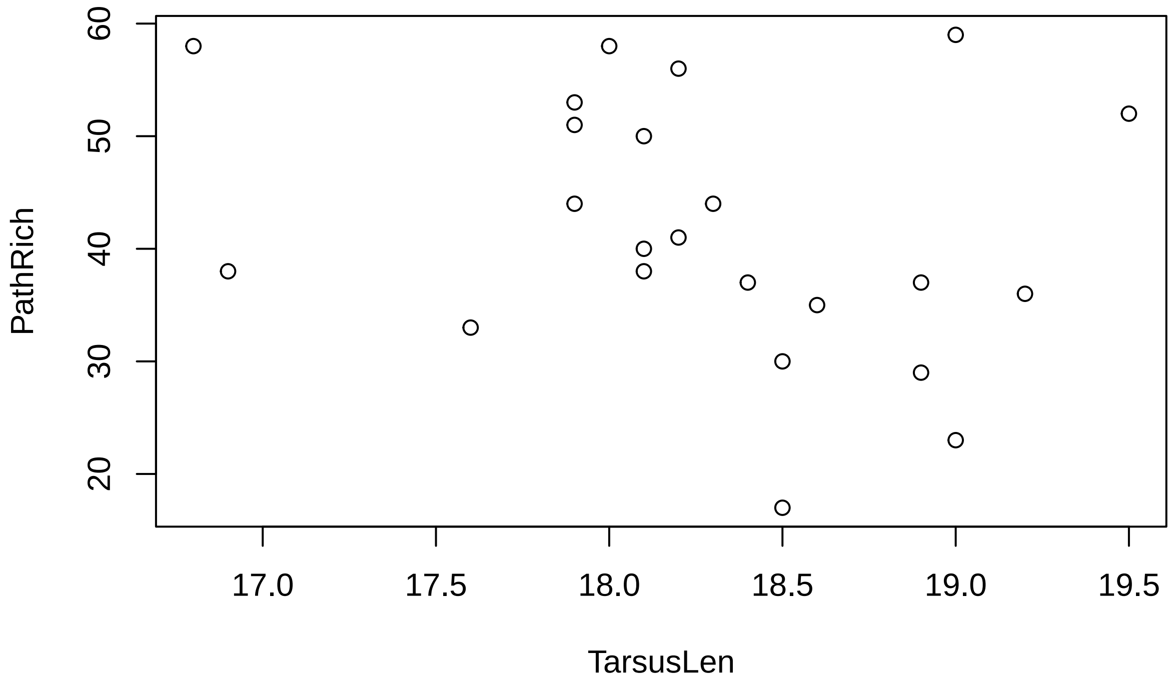
```
plot(PathRich ~ BirdWeight, data = df_feather)
```



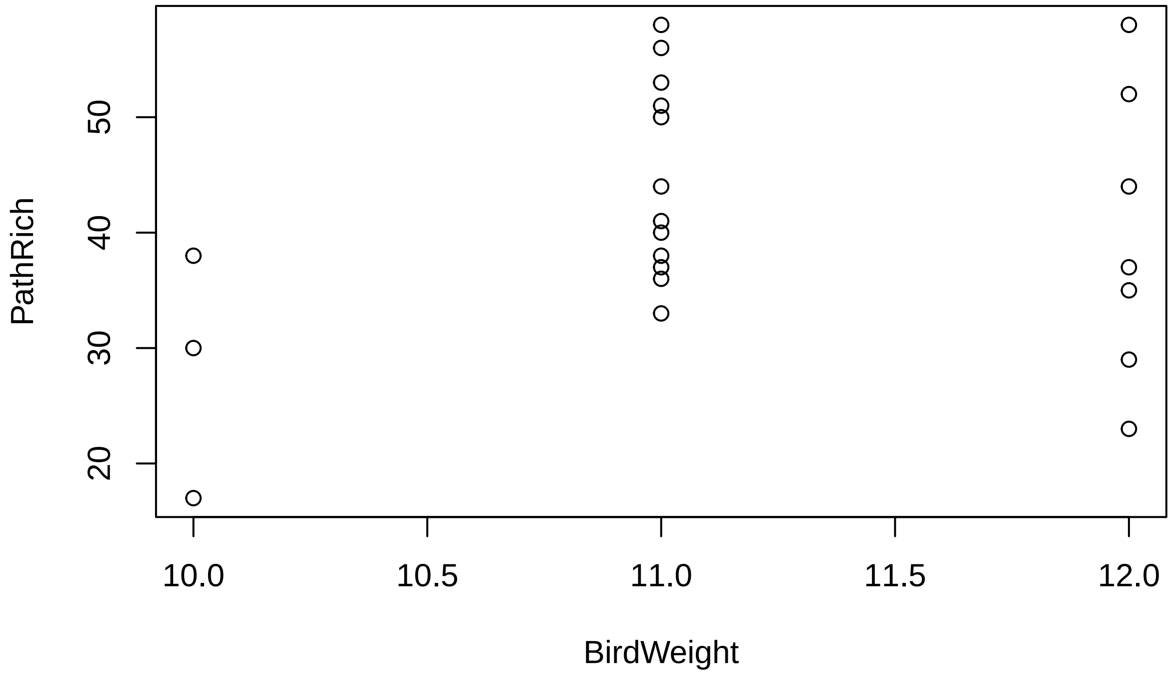
```
plot(PathRich ~ TailLen, data = df_feather)
```



```
plot(PathRich ~ TarsusLen, data = df_feather)
```



```
# BirdWeight only has 1 value at 13, treat it as an outlier
df_feather <- filter(df_feather, BirdWeight < 12.5)
plot(PathRich ~ BirdWeight, data = df_feather)
```



```
# After removing the outliers, fit the linear model again

# Fit a linear model to each relationship
fit_WingChord <- lm(df_feather$PathRich ~ df_feather$WingChord)
fit_BirdWeight <- lm(df_feather$PathRich ~ df_feather$BirdWeight)
fit_TailLen <- lm(df_feather$PathRich ~ df_feather$TailLen)
fit_TarsusLen <- lm(df_feather$PathRich ~ df_feather$TarsusLen)

# Summarize the fits
summary(fit_WingChord)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$WingChord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.091  -7.591   2.561   6.420  15.909
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -104.8539    62.4870  -1.678   0.1089
## df_feather$WingChord    2.2607     0.9686   2.334   0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 10.12 on 20 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.1748
## F-statistic: 5.448 on 1 and 20 DF,  p-value: 0.03014
```

```
summary(fit_BirdWeight)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$BirdWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5000  -7.0000   0.0833   9.3750  17.6667
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.500     41.246   0.133    0.895
## df_feather$BirdWeight    3.167      3.682   0.860    0.400
##
## Residual standard error: 11.21 on 20 degrees of freedom
## Multiple R-squared:  0.03566,    Adjusted R-squared:  -0.01256
## F-statistic: 0.7395 on 1 and 20 DF,  p-value: 0.4
```

```
summary(fit_TailLen)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$TailLen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.030  -6.124   1.470   5.909  18.782
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -83.3272    53.4809  -1.558   0.1349
## df_feather$TailLen   2.1883     0.9412   2.325   0.0307 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.13 on 20 degrees of freedom
## Multiple R-squared:  0.2128, Adjusted R-squared:  0.1734
## F-statistic: 5.405 on 1 and 20 DF,  p-value: 0.03072
```

```
summary(fit_TarsusLen)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$TarsusLen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3499  -6.8525  -0.0381   8.0127  18.8869 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)  
## (Intercept)           154.731     64.695   2.392   0.0267 *
## df_feather$TarsusLen   -6.237      3.543  -1.760   0.0936 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 10.63 on 20 degrees of freedom
## Multiple R-squared:  0.1342, Adjusted R-squared:  0.09088 
## F-statistic: 3.099 on 1 and 20 DF,  p-value: 0.09362
```

```
# Alternatively, fit a multiple linear regression model
fit_all <- lm(PathRich ~ WingChord + BirdWeight + TailLen + TarsusLen, data = df_feather)
summary(fit_all)
```

```
## 
## Call:
## lm(formula = PathRich ~ WingChord + BirdWeight + TailLen + TarsusLen, 
##     data = df_feather)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -12.671  -5.117   1.243   3.198  14.665 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  15.4010    63.3530   0.243  0.81084   
## WingChord     2.5854     1.6562   1.561  0.13694   
## BirdWeight    0.8171     3.7004   0.221  0.82786   
## TailLen       0.7090     1.3911   0.510  0.61684   
## TarsusLen   -10.4426     3.1050  -3.363  0.00369 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 8.331 on 17 degrees of freedom
## Multiple R-squared:  0.5476, Adjusted R-squared:  0.4412 
## F-statistic: 5.144 on 4 and 17 DF,  p-value: 0.006673
```

After identifying and removing the most extreme outliers from our dataset, we refit our simple linear regression models and found that two predictor variables remained statistically significant.

Wing Chord: The model estimates a coefficient of 2.2607 for Wing Chord, with a p-value of 0.0301, which is significant at the 0.05 level. The Multiple R^2 value for this model is 0.2141.

Tail Length: The model estimates a coefficient of 2.1883 for Tail Length, with a p-value of 0.0307, also significant at the 0.05 level. The Multiple R^2 value for this model is 0.2128.

When we used all four predictors in a multiple linear regression model, the overall model was significant with a p-value of 0.006673, and a Multiple R^2 value of 0.5476. This suggests that our model explains approximately 54.76% of the variability in the response variable, which is a substantial improvement from the individual models.

Therefore, after excluding the most extreme outliers, we reject the null hypothesis that the predictors have no effect on the response variable for both the individual and multiple regression models. We conclude that there exists at least one predictor that has a statistically significant effect on the response variable.

13. Multiple Linear Regression of Morphological Features and Pathogen Richness to Find Best Two Predictors

```r
# WingChord + BirdWeight
fit_WingChord_BirdWeight <- lm(PathRich ~ WingChord + BirdWeight, data = df_feather)
summary(fit_WingChord_BirdWeight)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + BirdWeight, data = df_feather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.399  -8.294   3.374   6.619  15.109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -106.802     63.452  -1.683   0.1087
## WingChord      2.790      1.267   2.203   0.0402 *
## BirdWeight    -2.879      4.348  -0.662   0.5159
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.27 on 19 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.1509
## F-statistic: 2.867 on 2 and 19 DF,  p-value: 0.08165
```

```r
# WingChord + TailLen
fit_WingChord_TailLen <- lm(PathRich ~ WingChord + TailLen, data = df_feather)
summary(fit_WingChord_TailLen)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + TailLen, data = df_feather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.834  -5.858   2.797   5.557  17.347
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -107.154     63.386  -1.690    0.107
## WingChord      1.256      1.739   0.722    0.479
## TailLen        1.181      1.689   0.699    0.493
```

```
##
## Residual standard error: 10.25 on 19 degrees of freedom
## Multiple R-squared:  0.2338, Adjusted R-squared:  0.1532
## F-statistic: 2.899 on 2 and 19 DF,  p-value: 0.07965
```

```
# WingChord + TarsusLen
fit_WingChord_TarsusLen <- lm(PathRich ~ WingChord + TarsusLen, data = df_feather)
summary(fit_WingChord_TarsusLen)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + TarsusLen, data = df_feather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0403  -4.1876   0.3558   2.7972  15.6981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.9477    59.0703   0.270 0.790087
## WingChord     3.3360     0.8147   4.095 0.000617 ***
## TarsusLen   -10.4184     2.8393  -3.669 0.001630 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.946 on 19 degrees of freedom
## Multiple R-squared:   0.54,  Adjusted R-squared:  0.4916
## F-statistic: 11.15 on 2 and 19 DF,  p-value: 0.000625
```

```
# BirdWeight + TailLen
fit_BirdWeight_TailLen <- lm(PathRich ~ BirdWeight + TailLen, data = df_feather)
summary(fit_BirdWeight_TailLen)
```

```
##
## Call:
## lm(formula = PathRich ~ BirdWeight + TailLen, data = df_feather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.453  -6.465   1.491   5.986  18.745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -81.6078    56.8264  -1.436   0.1672
## BirdWeight   -0.4436     3.8317  -0.116   0.9091
## TailLen       2.2454     1.0840   2.071   0.0522 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.39 on 19 degrees of freedom
## Multiple R-squared:  0.2133, Adjusted R-squared:  0.1305
## F-statistic: 2.576 on 2 and 19 DF,  p-value: 0.1024
```

```r
# BirdWeight + TarsusLen
fit_BirdWeight_TarsusLen <- lm(PathRich ~ BirdWeight + TarsusLen, data = df_feather)
summary(fit_BirdWeight_TarsusLen)
```

```
##
## Call:
## lm(formula = PathRich ~ BirdWeight + TarsusLen, data = df_feather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7225  -6.8188  -0.6157   7.3957  16.9310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  131.582     61.670   2.134   0.0461 *
## BirdWeight     7.081      3.610   1.961   0.0647 .
## TarsusLen     -9.307      3.666  -2.539   0.0200 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.941 on 19 degrees of freedom
## Multiple R-squared:  0.2799, Adjusted R-squared:  0.2042
## F-statistic: 3.693 on 2 and 19 DF,  p-value: 0.04415
```

```r
# TailLen + TarsusLen
fit_TailLen_TarsusLen <- lm(PathRich ~ TailLen + TarsusLen, data = df_feather)
summary(fit_TailLen_TarsusLen)
```

```
##
## Call:
## lm(formula = PathRich ~ TailLen + TarsusLen, data = df_feather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8531  -4.5860  -0.1362   5.6269  15.4642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.2383    62.7149   0.658  0.51871
## TailLen       2.7648     0.8283   3.338  0.00346 **
## TarsusLen    -8.6189     2.9728  -2.899  0.00919 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.655 on 19 degrees of freedom
## Multiple R-squared:  0.4542, Adjusted R-squared:  0.3968
## F-statistic: 7.906 on 2 and 19 DF,  p-value: 0.003174
```

To identify the best multiple linear regression model for predicting pathogen richness on feathers from two predictors among Wing Chord, Bird Weight, Tail Length, and Tarsus Length, we evaluated several models based on their p-values, Multiple R^2, Adjusted R^2, and Residual Standard Error (RSE).

Based on these criteria, three models emerged as significant:

Wing Chord + Tarsus Length Model:

P-value: 0.000625 (Significant)

Multiple R^2: 0.54

Adjusted R^2: 0.4916

RSE: 7.946

Bird Weight + Tarsus Length Model:

P-value: 0.04415 (Significant)

Multiple R^2: 0.2799

Adjusted R^2: 0.2042

RSE: 9.941

Tail Length + Tarsus Length Model

P-value: 0.003174 (Significant)

Multiple R^2: 0.4542

Adjusted R^2: 0.3968

RSE: 8.655

The model that includes Wing Chord and Tarsus Length as predictors stands out as the best model based on all the criteria we considered. It not only has the lowest p-value but also the highest Multiple R^2 and Adjusted R^2 values, along with the lowest RSE. This suggests that the model is both statistically significant and explains a substantial portion of the variance in pathogen richness.

This finding is consistent with our single linear regression analysis, which also identified Wing Chord as a significant predictor. Although Tail Length was identified as a significant predictor in the single linear regression, the multiple regression analysis showed that a model including Wing Chord and Tarsus Lenth is superior based on our evaluation criteria.

14. Two-Sample *t*-Test for Pathogen Richness on Feathers between Sexes

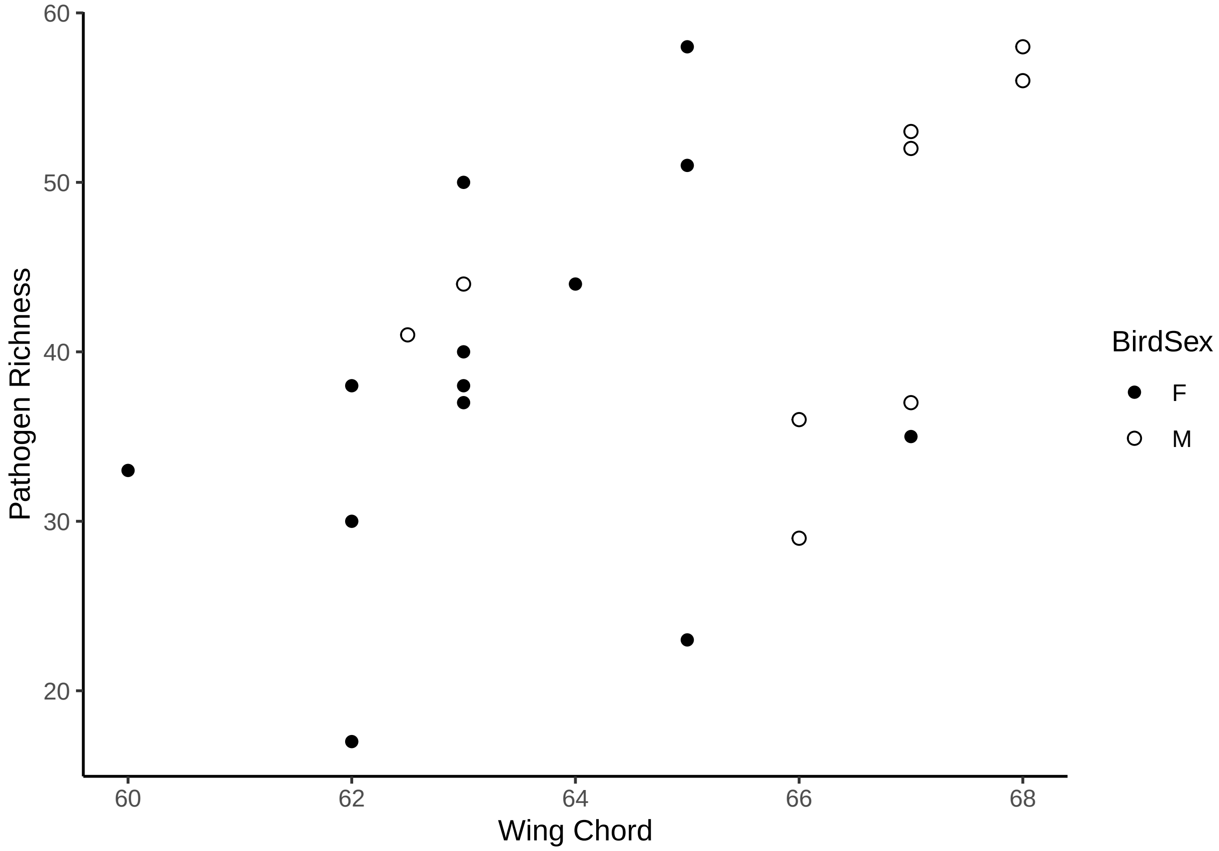
```r
# t-test assuming equal variance
t.test(PathRich ~ BirdSex, data = df_feather, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  PathRich by BirdSex
## t = -1.5165, df = 20, p-value = 0.1451
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -16.892824   2.670602
## sample estimates:
## mean in group F mean in group M
##        38.00000        45.11111
```

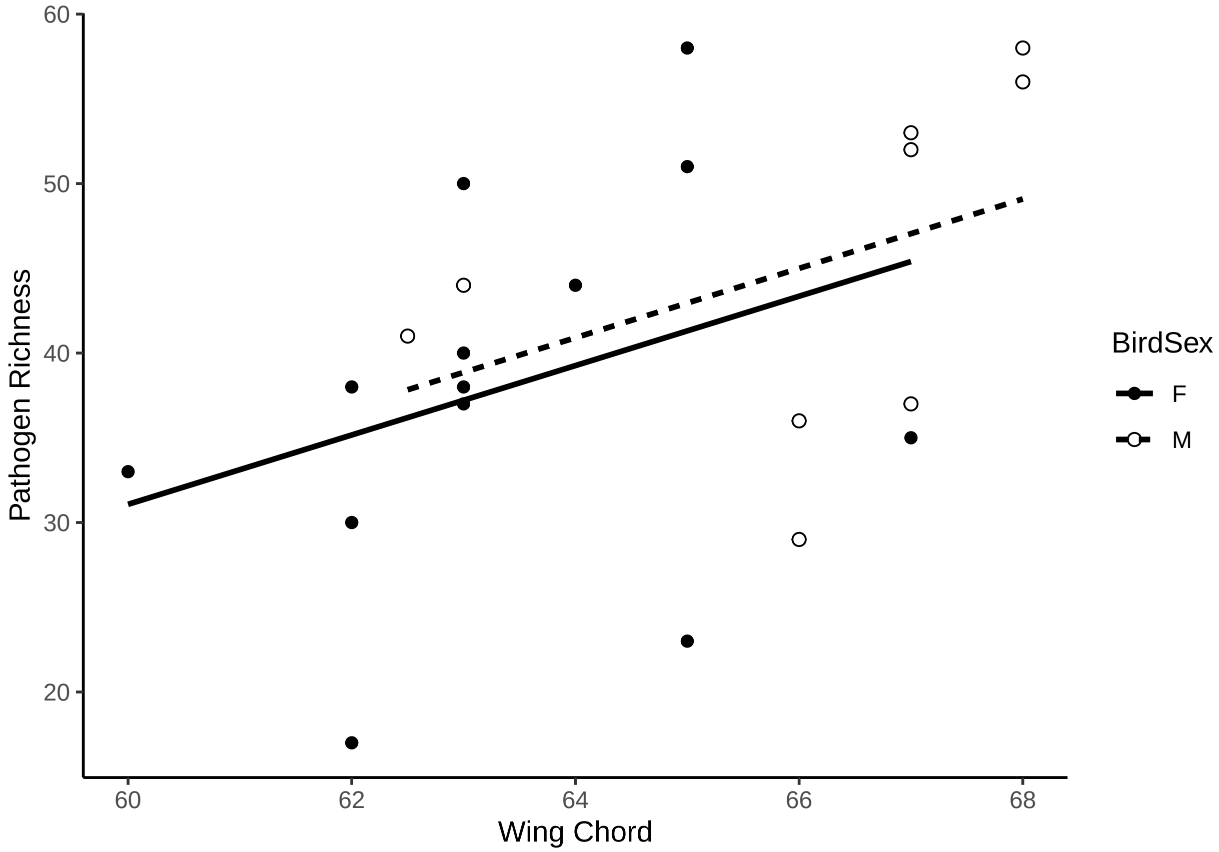
In this study, we assessed whether there is a significant difference in the mean pathogen richness on feathers between male and female birds. To do so, we used a two-sample t-test. The null hypothesis is that there is no significant difference between the means, while the alternative hypothesis asserts otherwise. Our analysis yielded the t-value of -1.5165 with 20 degrees of freedom, and a p-value of 0.1451. We find that the p-value exceeds the alpha level of 0.05. Thus, there is insufficient evidence to reject the null hypothesis. We conclude that there is no statistically significant difference in the mean pathogen richness on feathers between male and female birds based on this data.

15. Analysis of Covariance on Linear Models between Sexes

```
# ANCOVA

# Wing Chord

# Scatterplot by Sex
ggplot(df_feather, aes(WingChord, PathRich, shape = BirdSex)) +
geom_point(size = 2) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Wing Chord", y = "Pathogen Richness")
```



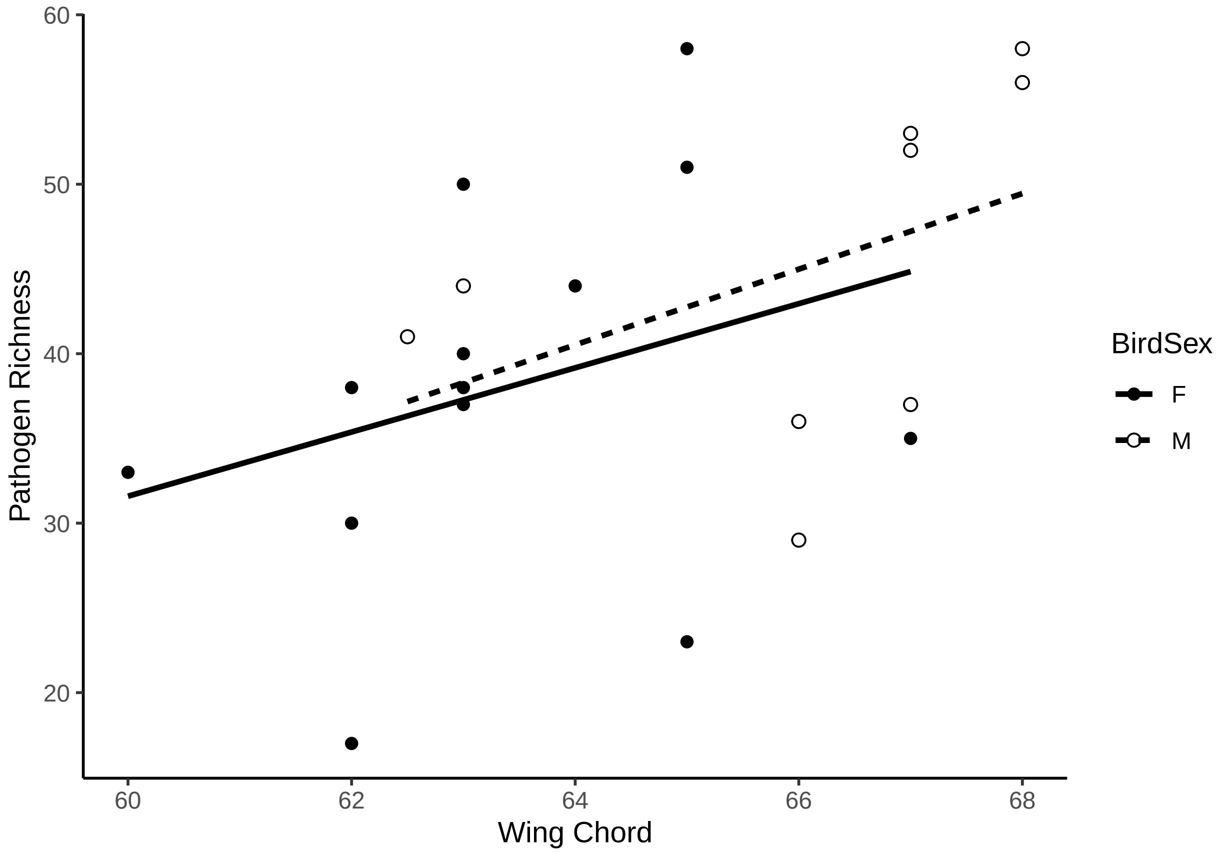
```
# Fit the main effects model (with no interaction term)
featherNoInteractModel <- lm(PathRich ~ WingChord + BirdSex,
data = df_feather)
df_feather$fit0 <- predict(featherNoInteractModel)
ggplot(df_feather, aes(WingChord, PathRich, colour = BirdSex,
shape = BirdSex, linetype=BirdSex)) +
geom_line(aes(y = fit0), size = 1, color = "black") +
geom_point(size = 2) +
scale_colour_manual(values = c("black", "black")) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Wing Chord", y = "Pathogen Richness")
```



```
summary(featherNoInteractModel)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + BirdSex, data = df_feather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.307  -8.040   2.811   5.753  16.693
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -91.744     77.838  -1.179    0.253
## WingChord      2.047      1.227   1.668    0.112
## BirdSexM       1.644      5.562   0.296    0.771
##
## Residual standard error: 10.36 on 19 degrees of freedom
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.1353
## F-statistic: 2.643 on 2 and 19 DF,  p-value: 0.09709
```

```
# Fit the interaction model
featherInteractModel <- lm(PathRich ~ WingChord * BirdSex,
data = df_feather)
ggplot(df_feather, aes(WingChord, PathRich, colour = BirdSex,
shape = BirdSex, linetype=BirdSex)) +
geom_smooth(method = "lm", size = 1, se = FALSE, col = "black") +
geom_point(size = 2) +
scale_colour_manual(values = c("black", "black")) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Wing Chord", y = "Pathogen Richness")
```



```
summary(featherInteractModel)
```

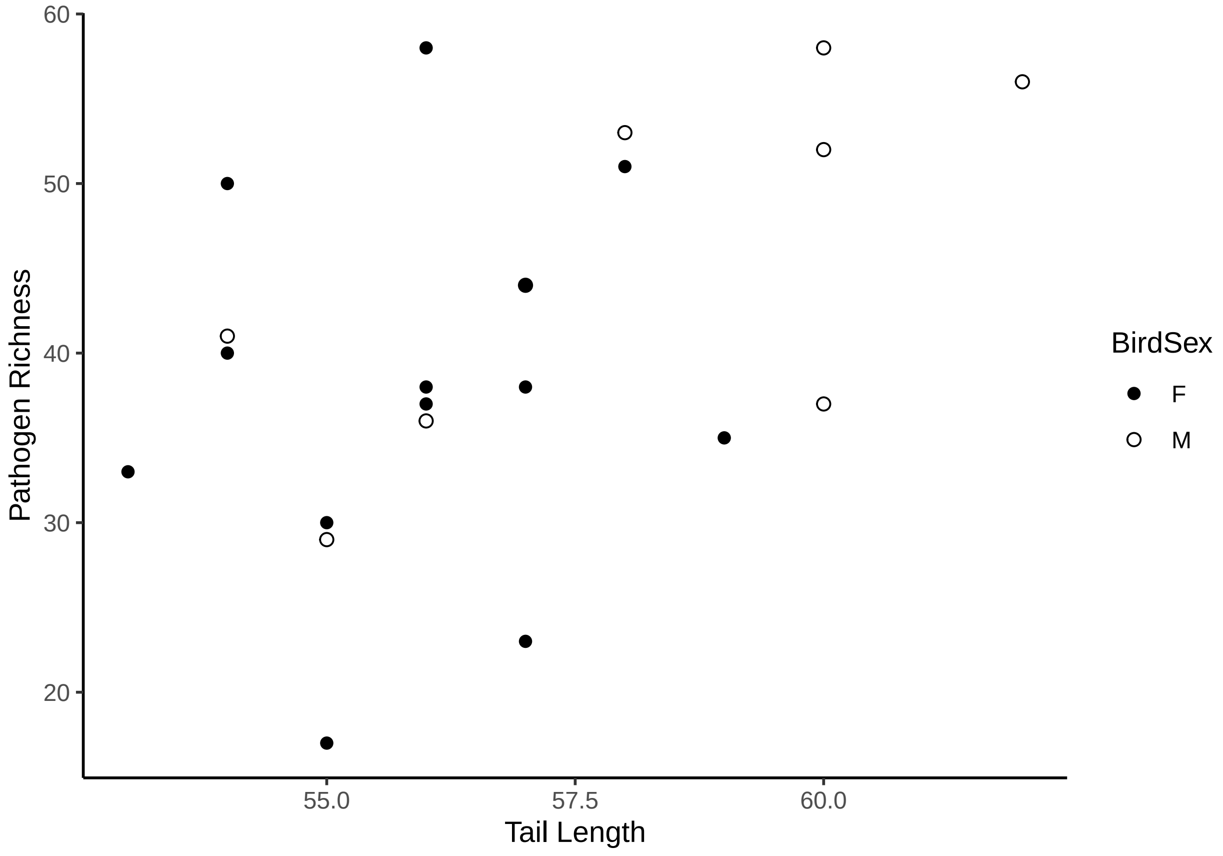
```
##
## Call:
## lm(formula = PathRich ~ WingChord * BirdSex, data = df_feather)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.378  -8.085   2.675   5.763  16.941
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -82.0315   107.9356  -0.760    0.457
## WingChord            1.8937     1.7022   1.112    0.281
## BirdSexM           -20.3435   164.3029  -0.124    0.903
## WingChord:BirdSexM   0.3391     2.5321   0.134    0.895
##
## Residual standard error: 10.64 on 18 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.0882
## F-statistic: 1.677 on 3 and 18 DF,  p-value: 0.2075
```

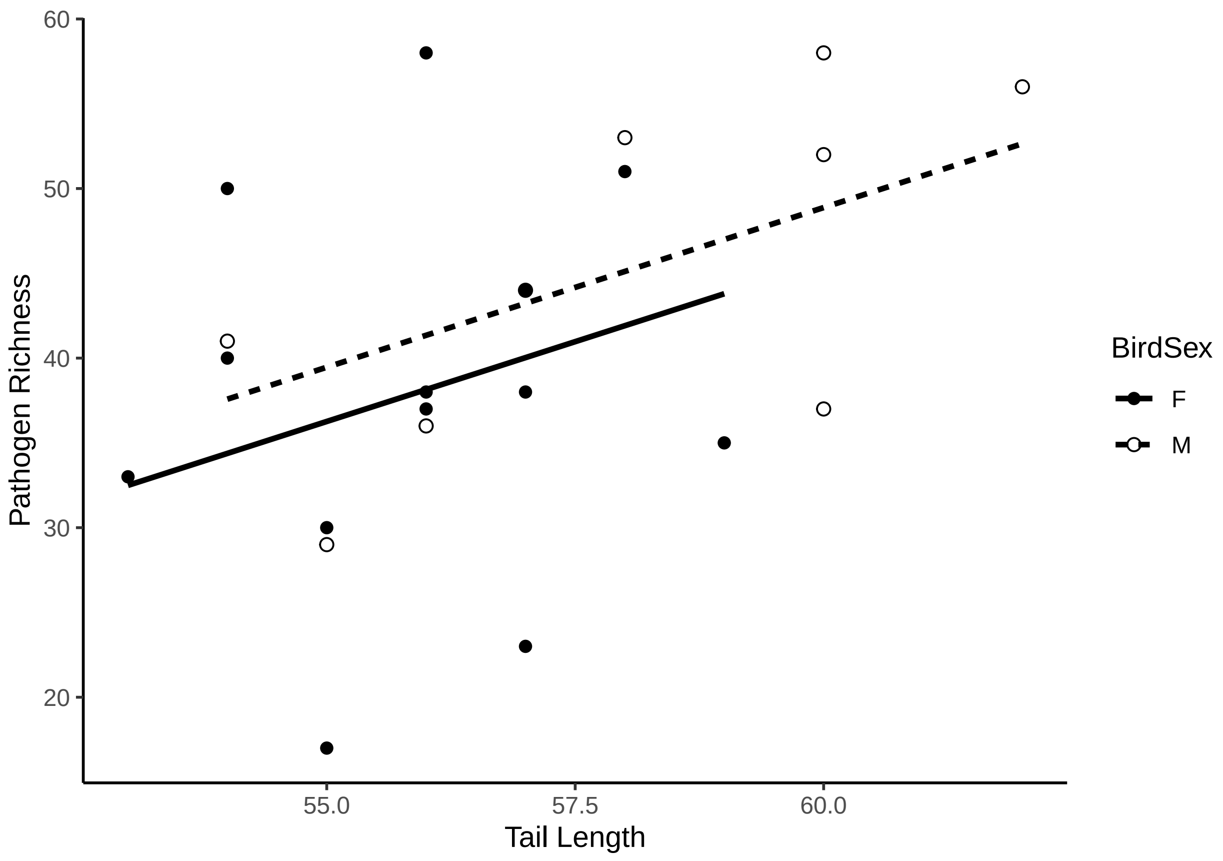
```
# ANOVA table
anova(featherNoInteractModel, featherInteractModel)
```

```
## Analysis of Variance Table
##
## Model 1: PathRich ~ WingChord + BirdSex
## Model 2: PathRich ~ WingChord * BirdSex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 2040.2
## 2     18 2038.1  1    2.0302 0.0179  0.895
```

```
# No significant difference

# Tail Length
# Scatterplot by Sex
ggplot(df_feather, aes(TailLen, PathRich, shape = BirdSex)) +
geom_point(size = 2) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Tail Length", y = "Pathogen Richness")
```



```
# Fit the main effects model (with no interaction term)
featherNoInteractModel <- lm(PathRich ~ TailLen + BirdSex,
data = df_feather)
df_feather$fit0 <- predict(featherNoInteractModel)
ggplot(df_feather, aes(TailLen, PathRich, colour = BirdSex,
shape = BirdSex, linetype=BirdSex)) +
geom_line(aes(y = fit0), size = 1, color = "black") +
geom_point(size = 2) +
scale_colour_manual(values = c("black", "black")) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Tail Length", y = "Pathogen Richness")
```

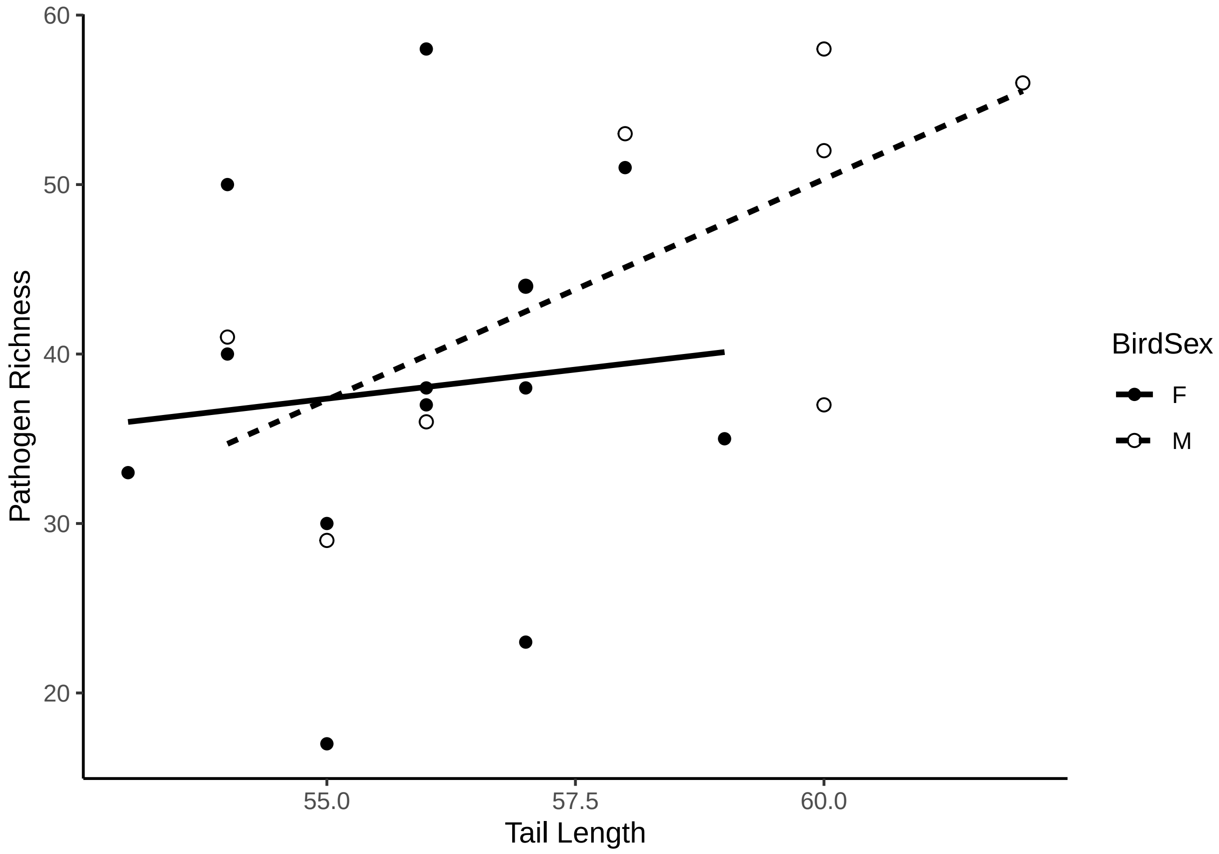


```
summary(featherNoInteractModel)
```

```
##
## Call:
## lm(formula = PathRich ~ TailLen + BirdSex, data = df_feather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2616  -6.0323   0.6386   5.2092  19.8551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -67.319     59.727  -1.127   0.2737
## TailLen        1.883      1.067   1.765   0.0936 .
## BirdSexM       3.200      4.979   0.643   0.5282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.28 on 19 degrees of freedom
## Multiple R-squared:  0.2295, Adjusted R-squared:  0.1484
## F-statistic:  2.83 on 2 and 19 DF,  p-value: 0.08401
```

```
# Fit the interaction model
featherInteractModel <- lm(PathRich ~ TailLen * BirdSex,
data = df_feather)
```

```
ggplot(df_feather, aes(TailLen, PathRich, colour = BirdSex,
shape = BirdSex, linetype=BirdSex)) +
geom_smooth(method = "lm", size = 1, se = FALSE, col = "black") +
geom_point(size = 2) +
scale_colour_manual(values = c("black", "black")) +
scale_shape_manual(values = c(16, 1)) +
labs(x = "Tail Length", y = "Pathogen Richness")
```



```
summary(featherInteractModel)
```

```
##
## Call:
## lm(formula = PathRich ~ TailLen * BirdSex, data = df_feather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3656  -4.8120   0.2111   6.0420  19.9471
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.4317    98.0126  -0.004    0.997
## TailLen             0.6872     1.7519   0.392    0.699
## BirdSexM         -105.4572   125.8369  -0.838    0.413
## TailLen:BirdSexM    1.9162     2.2174   0.864    0.399
##
```

```
## Residual standard error: 10.35 on 18 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.1369
## F-statistic:  2.11 on 3 and 18 DF,  p-value: 0.1346
```

```
# ANOVA table
anova(featherNoInteractModel, featherInteractModel)
```

```
## Analysis of Variance Table
##
## Model 1: PathRich ~ TailLen + BirdSex
## Model 2: PathRich ~ TailLen * BirdSex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 2009.3
## 2     18 1929.3  1     80.04 0.7468 0.3989
```

```
# No significant difference
```

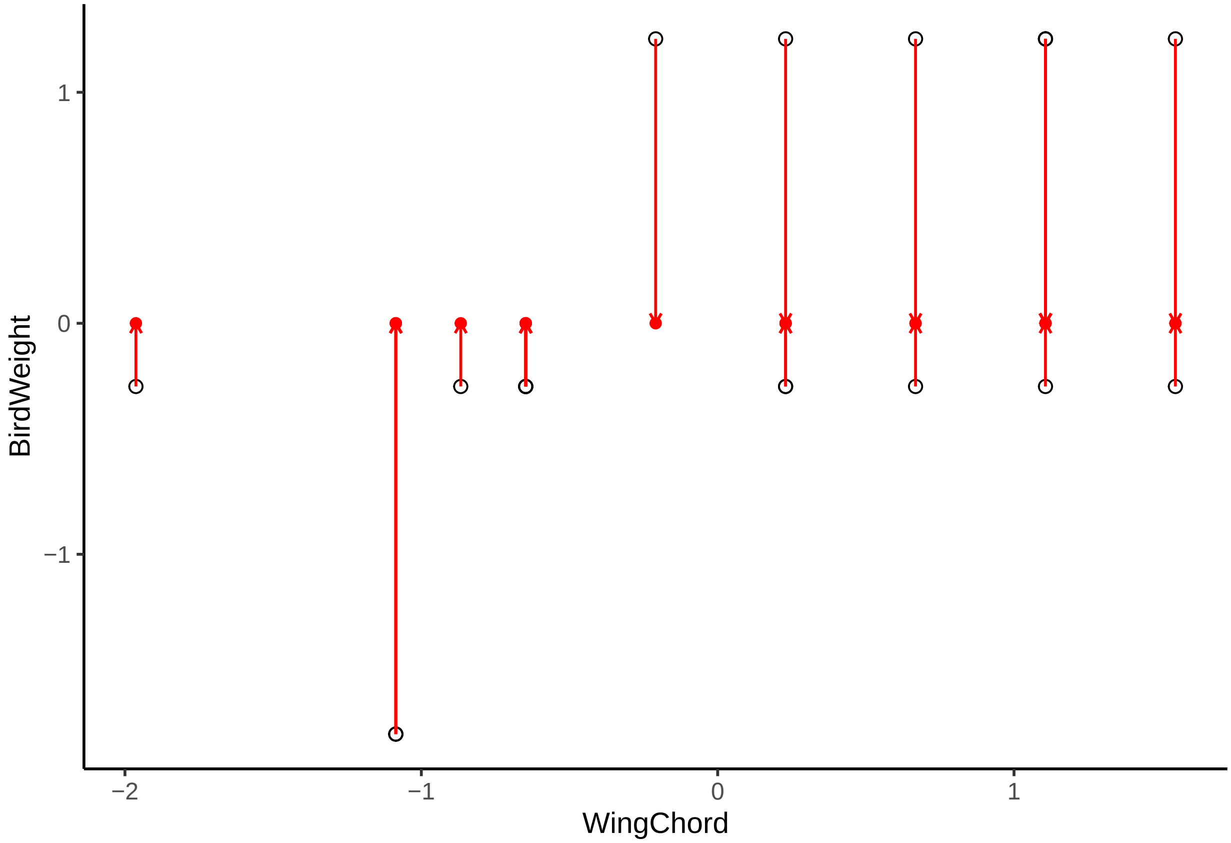
In this study, we used analysis of covariance to investigate whether the linear association in the model differs for male and female birds.

For the model focusing on Wing Chord and Bird Sex, the p-value for the effect of Bird Sex was 0.771, suggesting no significant difference between sexes. Additionally, an interaction term between Wing Chord and Bird Sex was added to the model, resulting in an F-statistic of 0.0179 and an associated p-value of 0.895. Both of these p-values are above the 0.05 significance level, indicating no significant interaction. Similarly, in the model that incorporated Tail Length and BirdSex, the p-value for the Bird Sex effect was 0.5282. When the interaction term between Tail Length and Bird Sex was considered, it yielded an F-statistic of 0.7468 and a p-value of 0.399, both of which also exceed the 0.05 significance level. In both cases, the Analysis of Variance tables suggested that including the interaction term did not significantly improve the fit of the models, as evidenced by their respective p-values of 0.895 and 0.3989.

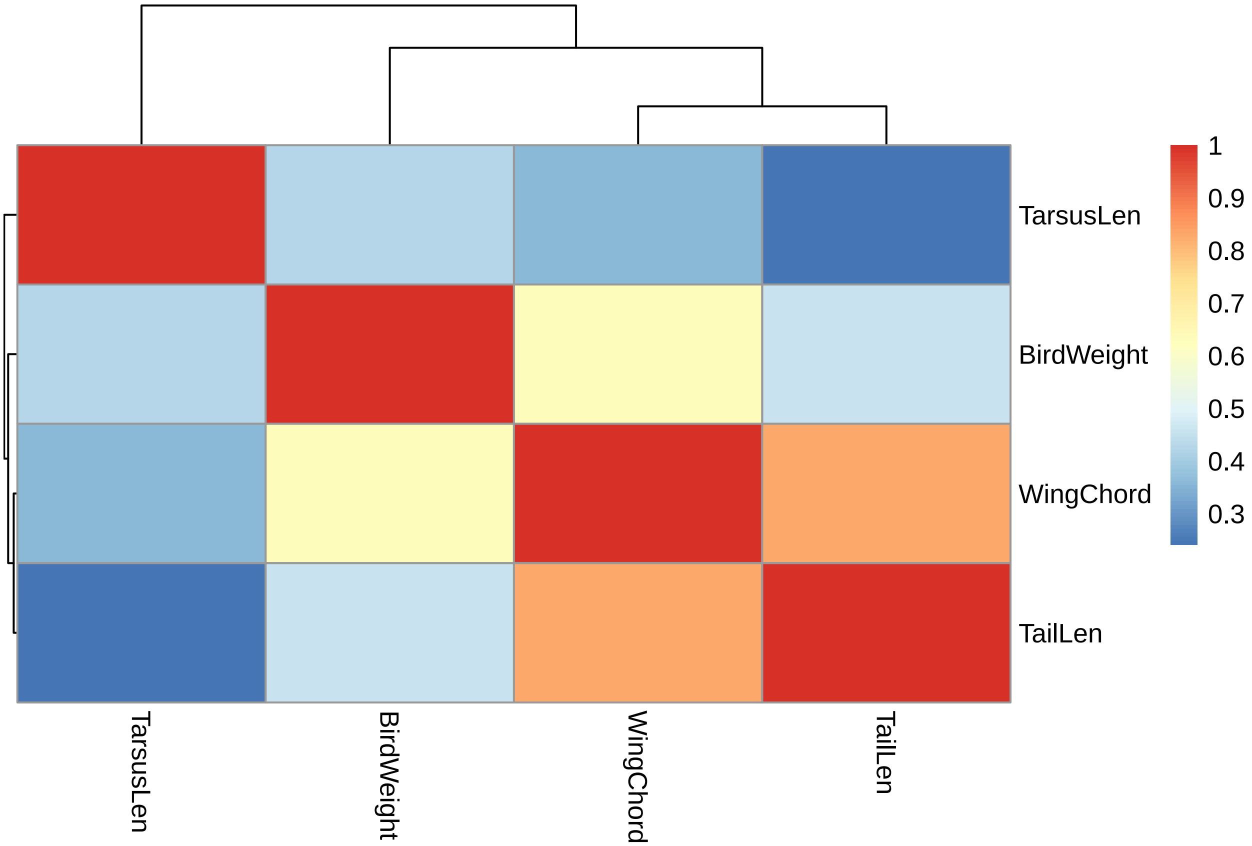
Given these findings, there is insufficient evidence to reject the null hypothesis, which is that there is no difference in the linear associations between pathogen richness and either Wing Chord or Tail Length, for male and female birds. Therefore, we conclude that the linear associations do not significantly differ between sexes in this dataset.

16. Principal Component Analysis on Morphological Features

```
# Quick PCA

# Gather the numeric variables
numeric_variables <- df_feather[, c("WingChord", "BirdWeight", "TailLen", "TarsusLen")]

scaledBirds <- data.frame(scale(numeric_variables))

ggplot(scaledBirds, aes(x = WingChord, y = BirdWeight)) +
geom_point(size = 2, shape = 21) +
geom_point(aes(y = 0), colour = "red") +
geom_segment(aes(xend = WingChord, yend = 0),
colour = "red",
arrow = arrow(length = unit(0.15, "cm")))
```



```
scaledBirds_clean_rows <- na.omit(scaledBirds)
# clustered heatmap of correlations
pheatmap(cor(scaledBirds_clean_rows), treeheight_row = 0.2)
```



```
# Perform PCA
pca_result <- prcomp(scaledBirds_clean_rows, center = TRUE, scale. = TRUE)
# center and scale. are usually TRUE for standardized data, but since the data is already scaled, they

summary_pca <- summary(pca_result)
print(summary_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4
## Standard deviation     1.5859 0.9233 0.6999 0.37753
## Proportion of Variance 0.6288 0.2131 0.1225 0.03563
## Cumulative Proportion  0.6288 0.8419 0.9644 1.00000
```

```
# Explain the first 2 principal components (rotation matrix)
pca_result$rotation[, 1:2]
```

```
##                 PC1        PC2
## WingChord  0.5821770  0.2671332
## BirdWeight 0.5027543 -0.2030167
## TailLen    0.5257270  0.4645204
## TarsusLen  0.3632067 -0.8195394
```

```
pca_result$sdev # standard deviations
```

```
## [1] 1.5859144 0.9232940 0.6999104 0.3775303
```

```
pca_result$sdev^2 # eigenvalues/variances
```

```
## [1] 2.5151245 0.8524718 0.4898745 0.1425292
```

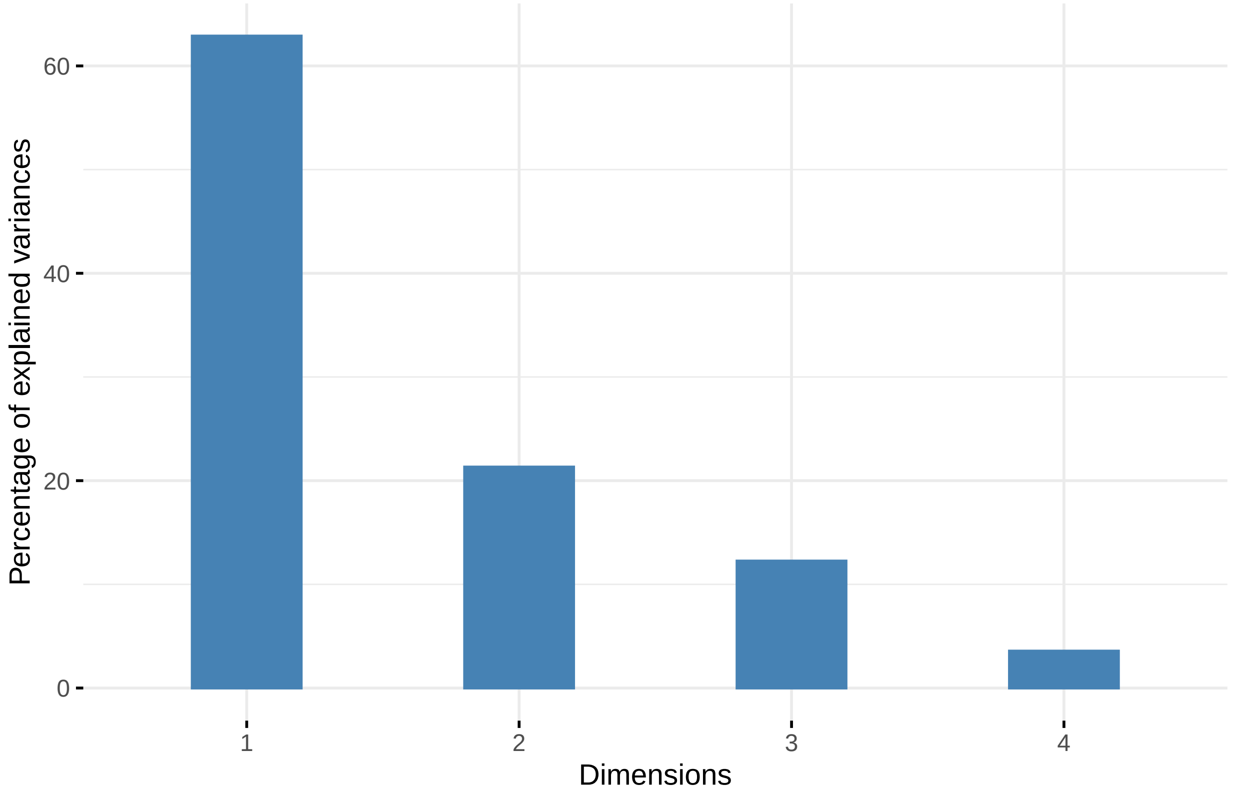
```
pca_result$sdev^2 / sum(pca_result$sdev^2) # proportion of variance
```

```
## [1] 0.62878114 0.21311794 0.12246863 0.03563229
```

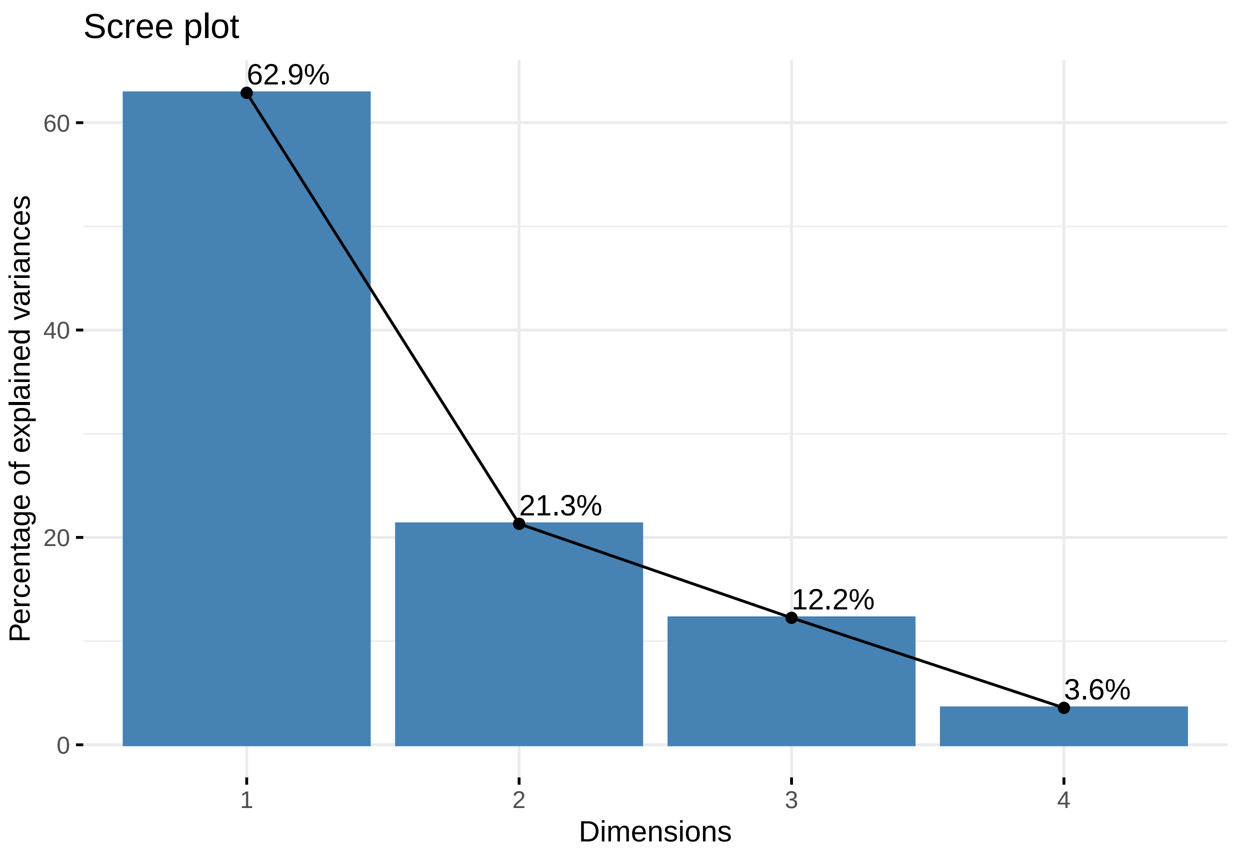
```
get_eig(pca_result) # eigenvalues/variances
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.5151245        62.878114                    62.87811
## Dim.2  0.8524718        21.311794                    84.18991
## Dim.3  0.4898745        12.246863                    96.43677
## Dim.4  0.1425292         3.563229                   100.00000
```

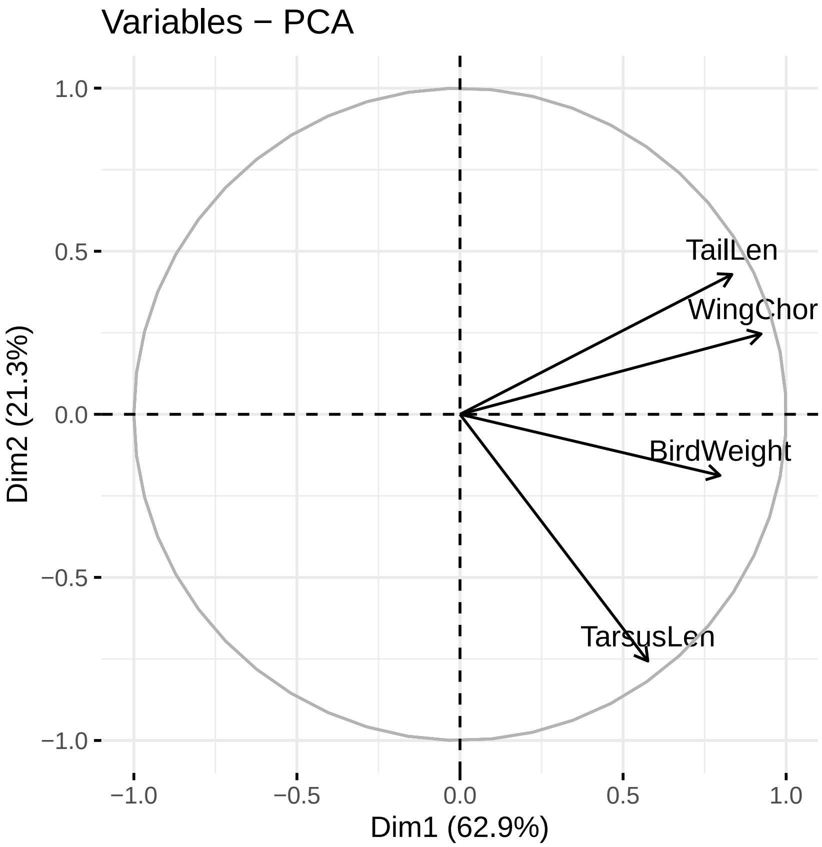
```
fviz_eig(pca_result, geom = "bar", bar_width = 0.4) +
ggtitle("")
```



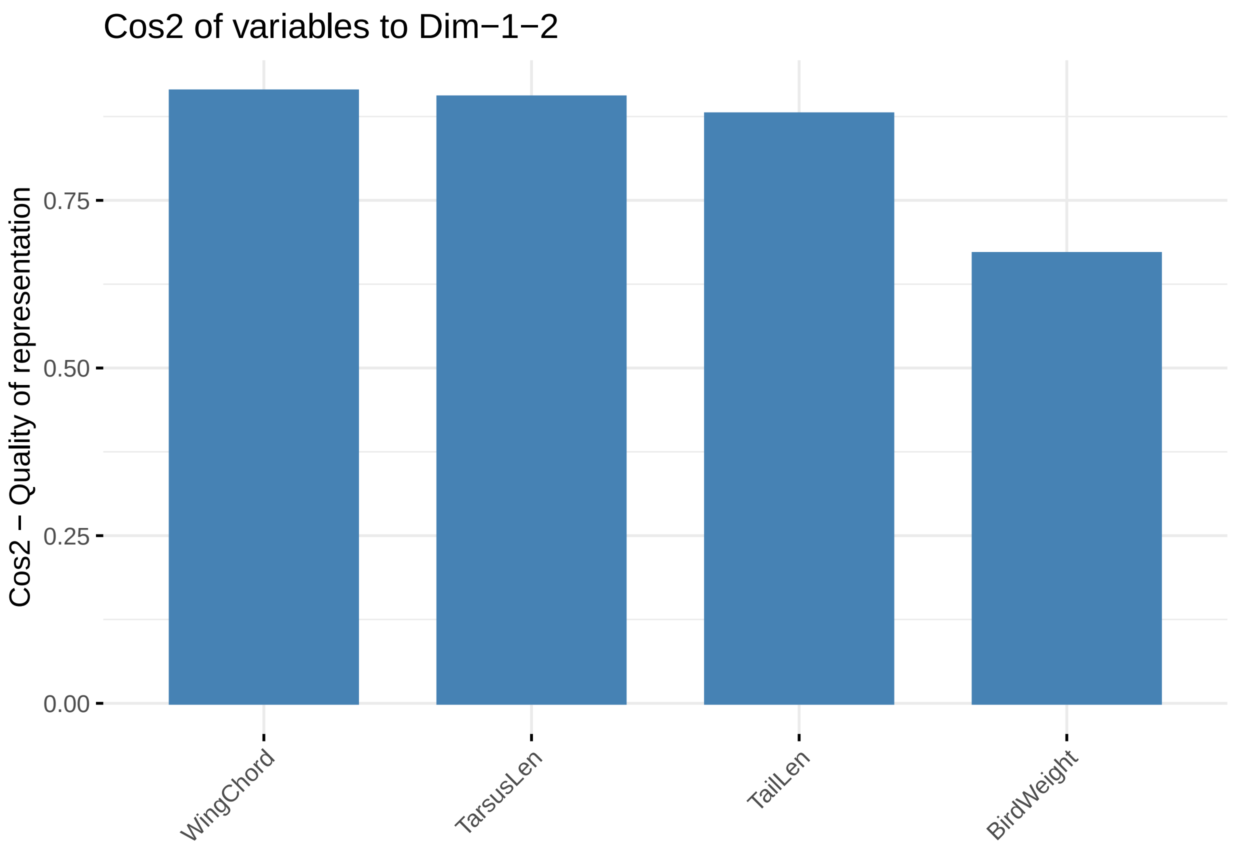
```
# Scree Plot
fviz_eig(pca_result, addlabels = TRUE)
```



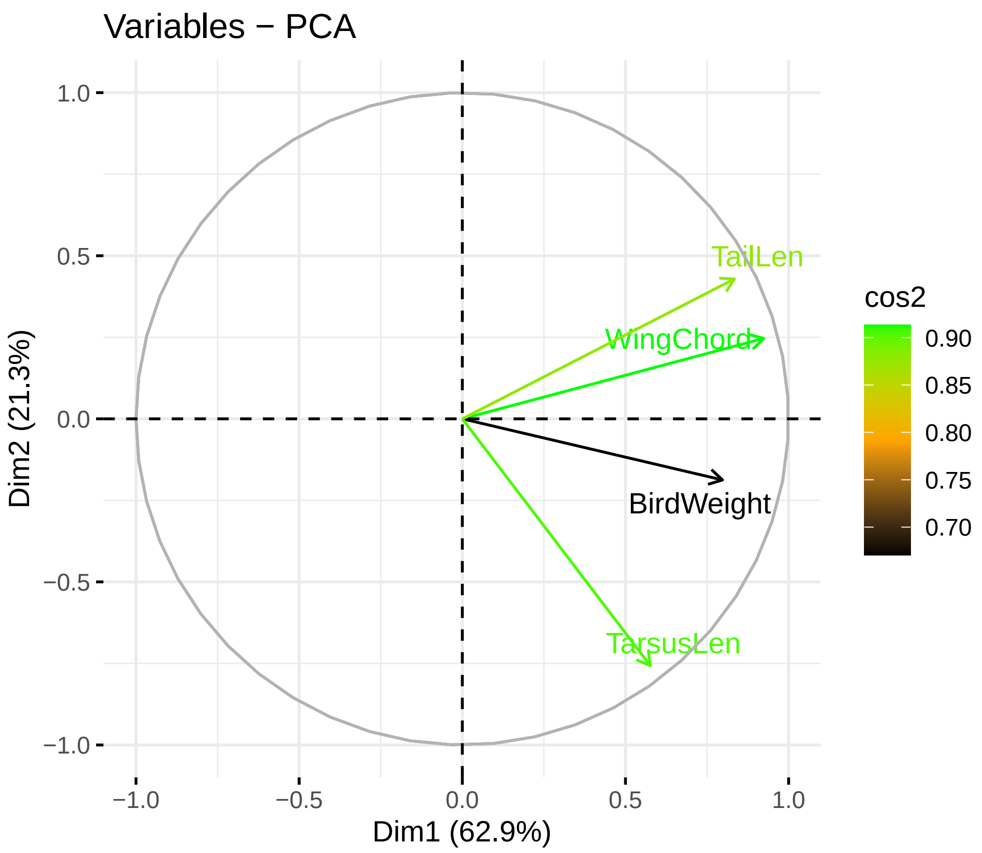
```
# Biplot of attributes
fviz_pca_var(pca_result, col.var = "black")
```



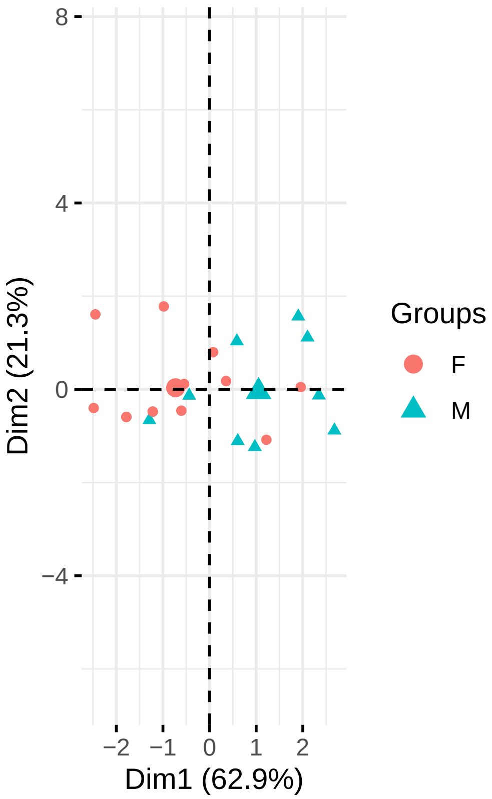
```
# Contribution of each variable
fviz_cos2(pca_result, choice = "var", axes = 1:2)
```



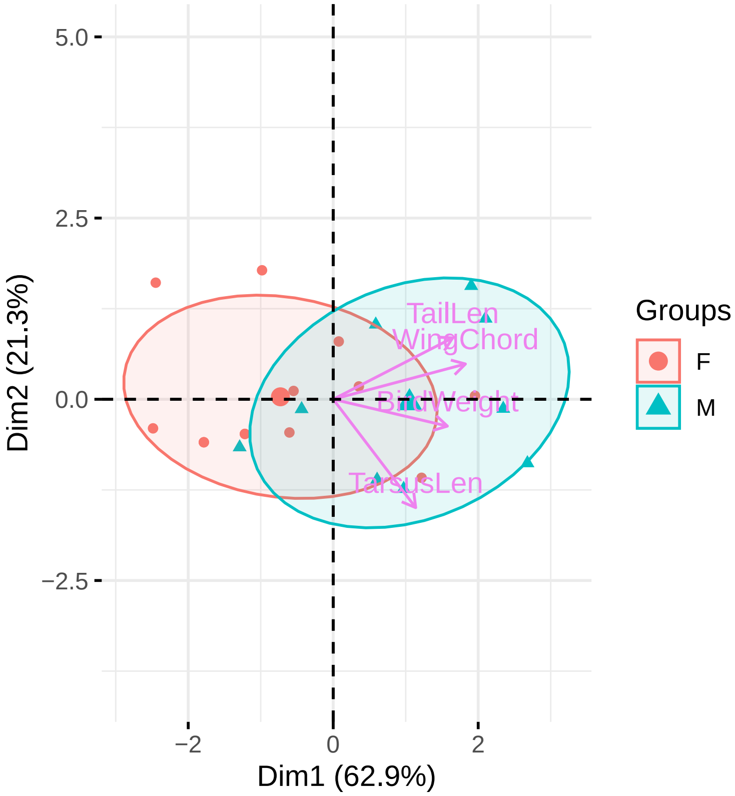
```
# Biplot combined with cos2
fviz_pca_var(pca_result, col.var = "cos2",
             gradient.cols = c("black", "orange", "green"),
             repel = TRUE)
```



```
# PCA Scatter Plot
chickadee_clean_rows <- na.omit(df_feather) # we need to ensure the rows are equal from the dataset not

fviz_pca_ind(pca_result, habillage = chickadee_clean_rows$BirdSex,
geom = "point") +
ggtitle("") +
ylim(c(-6.5,7.5)) +
coord_fixed()
```



```
# PCA BiPlot
fviz_pca_biplot(pca_result, geom = "point",
habillage = chickadee_clean_rows$BirdSex,
col.var = "violet", addEllipses = TRUE,
ellipse.level = 0.69) +
ggtitle("") +
ylim(c(-4,5)) +
coord_fixed()
```



In this study, Principal Component Analysis (PCA) was applied to four variables: Wing Chord, Bird Weight, Tail Len, and Tarsus Length. The first principal component (PC1) accounted for approximately 62.88% of the total variance, with a standard deviation of 1.586. The second principal component (PC2) accounted for 21.31% of the variance, having a standard deviation of 0.923. Together, these two components captured 84.19% of the total variance in the data.

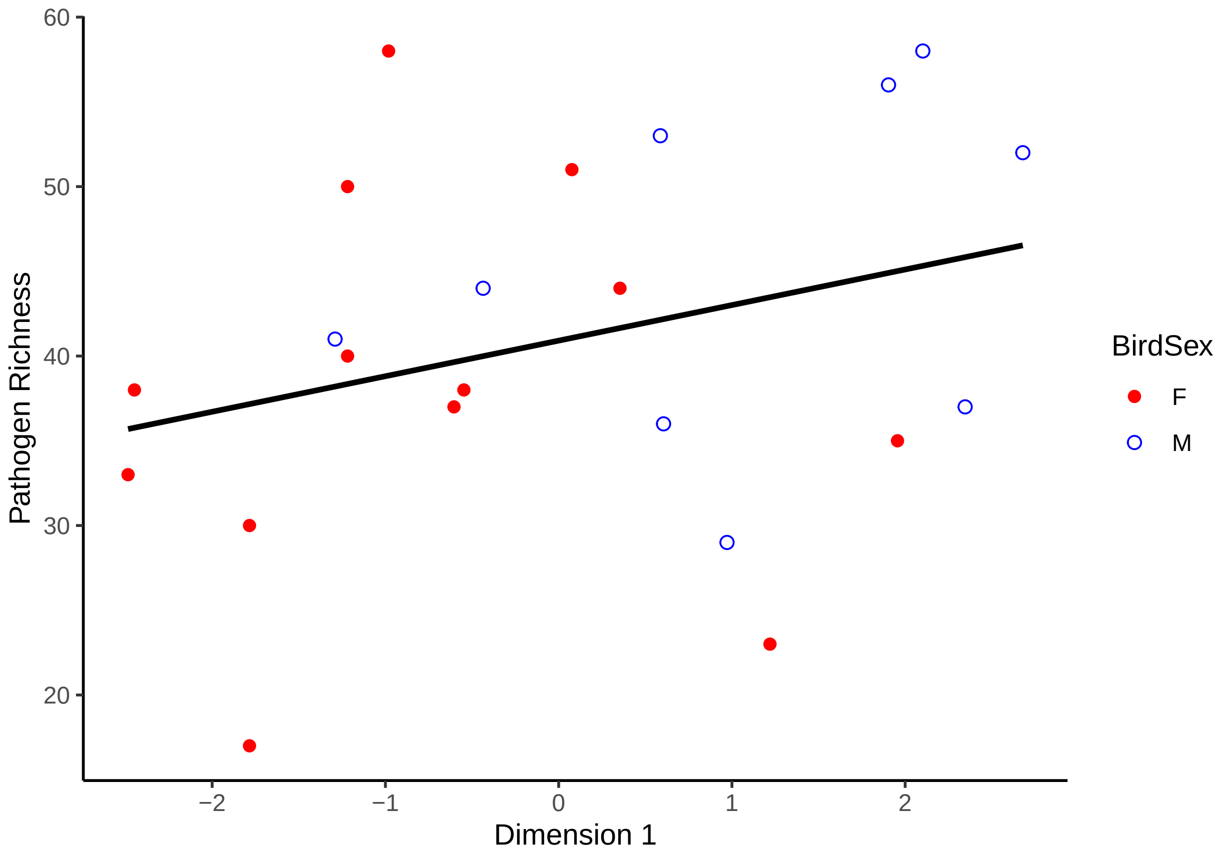
The biplot revealed that the female group was generally lower on both PC1 and PC2 as compared to the male group. The variables Wing Chord, Tail Length, and Bird Weight are positively correlated to each other.

In summary, the PCA indicates that Wing Chord, Bird Weight, and Tail Length are the major contributors to the first principal component and are positively correlated with each other. Tarsus Length, however, exhibits different behavior, especially in its significant contribution to the second principal component.

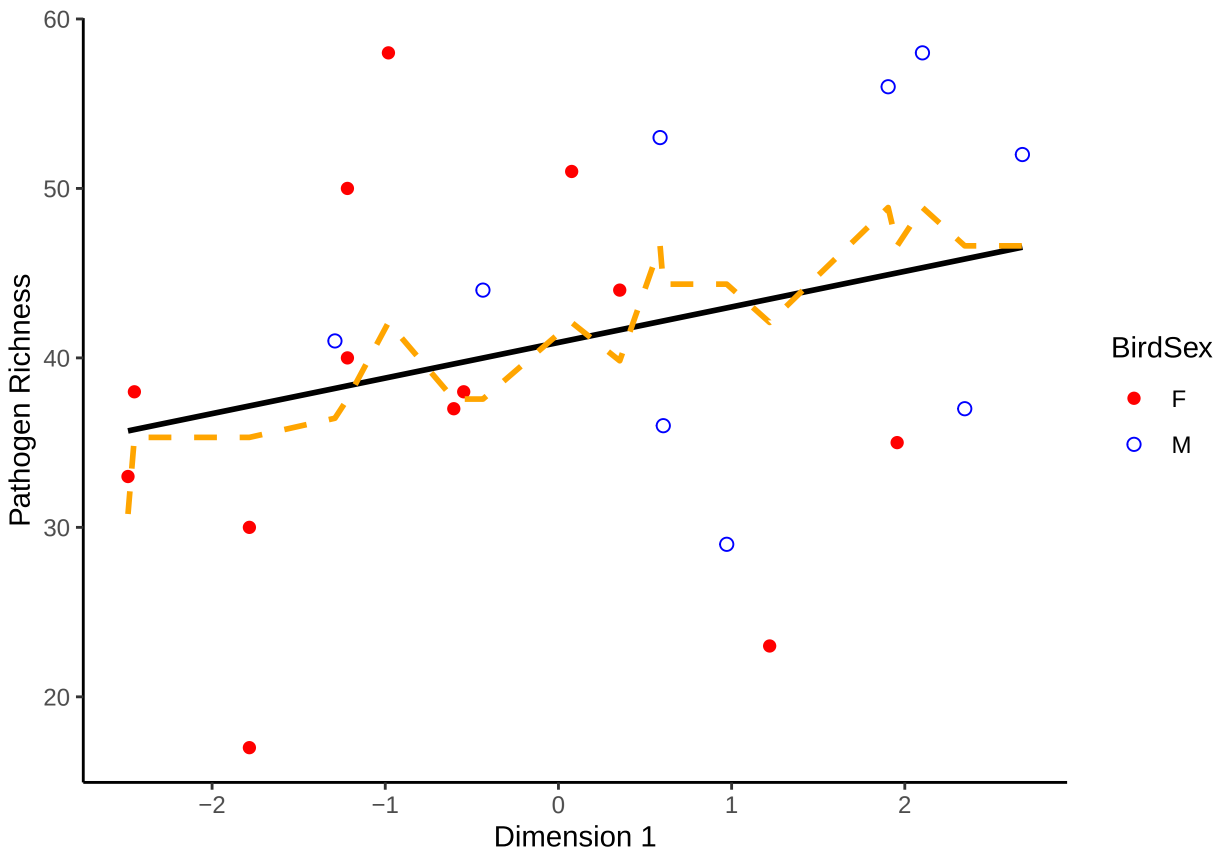
17. Simple Linear Regression Model for Pathogen Richness with the First Principal Component as a Predictor

```
# Extract the principal component scores
pc1_scores <- pca_result$x[, 1]
pc2_scores <- pca_result$x[, 2]

# Fit a linear model to each relationship
fit_PC1 <- lm(chickadee_clean_rows$PathRich ~ pc1_scores)
```

```r
# Make predictions using the linear model
chickadee_clean_rows$pc1_predicted <- predict(fit_PC1, newdata = chickadee_clean_rows)

# Scatterplot by Sex
ggplot(chickadee_clean_rows, aes(x = pc1_scores, y = PathRich)) +
  geom_point(aes(shape = BirdSex, color = BirdSex), size = 2) +
  scale_shape_manual(values = c(16, 1)) +
  geom_line(aes(y = pc1_predicted), size = 1, color = "black") +
  labs(x = "Dimension 1", y = "Pathogen Richness") +
  scale_colour_manual(values = c("red", "blue"))
```



```r
# Make predictions using the previous simple linear model
fit_WingChord_cleaned <- lm(chickadee_clean_rows$PathRich ~ chickadee_clean_rows$WingChord)
chickadee_clean_rows$predicted_WingChord <- predict(fit_WingChord_cleaned, newdata = chickadee_clean_row

# Scatterplot by Sex with both models (along Dimension 1)
ggplot(chickadee_clean_rows, aes(x = pc1_scores, y = PathRich)) +
  geom_point(aes(shape = BirdSex, color = BirdSex), size = 2) +
  scale_shape_manual(values = c(16, 1)) +
  geom_line(aes(y = pc1_predicted), size = 1, color = "black") +
    geom_line(aes(x = pc1_scores, y = predicted_WingChord), size = 1, color = "orange", linetype = "dash
  labs(x = "Dimension 1", y = "Pathogen Richness") +
  scale_colour_manual(values = c("red", "blue"))
```



```
# Summarize the fits
summary(fit_PC1)
```

```
##
## Call:
## lm(formula = chickadee_clean_rows$PathRich ~ pc1_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.469  -6.917   1.940   8.815  19.152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.909      2.323   17.61 1.21e-13 ***
## pc1_scores     2.099      1.499    1.40    0.177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 20 degrees of freedom
## Multiple R-squared:  0.08924,    Adjusted R-squared:  0.0437
## F-statistic:  1.96 on 1 and 20 DF,  p-value: 0.1769
```

```
summary(fit_WingChord_cleaned)
```

```
##
```

```
## Call:
## lm(formula = chickadee_clean_rows$PathRich ~ chickadee_clean_rows$WingChord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.091  -7.591   2.561   6.420  15.909
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -104.8539    62.4870  -1.678   0.1089
## chickadee_clean_rows$WingChord    2.2607     0.9686   2.334   0.0301 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.12 on 20 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.1748
## F-statistic: 5.448 on 1 and 20 DF,  p-value: 0.03014
```

```
summary(fit_WingChord)
```

```
##
## Call:
## lm(formula = df_feather$PathRich ~ df_feather$WingChord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.091  -7.591   2.561   6.420  15.909
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -104.8539    62.4870  -1.678   0.1089
## df_feather$WingChord    2.2607     0.9686   2.334   0.0301 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.12 on 20 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.1748
## F-statistic: 5.448 on 1 and 20 DF,  p-value: 0.03014
```

In this study, we fit a a simple linear regression model with Pathogen Richness as a response variable and predictor variable equal to the first principal component, and compared it to the simple linear regression model with Wing Chord as a predictor.

In the first model, where Pathogen Richness is predicted by PC1, the coefficient for PC1 is 2.099, with a standard error of 1.499, resulting in a t-value of 1.40 and a p-value of 0.177. This indicates that, at the 5% significance level, PC1 is not a significant predictor of Pathogen Richness, as the p-value is greater than 0.05. The model explains only 8.92% of the variance in Pathogen Richness ($R^2 = 0.08924$), which is not a large amount, and the F-statistic of 1.96 with a p-value of 0.1769 further confirms that the model is not a good fit.

In the second and third models, where Pathogen Richness is predicted by Wing Chord, the coefficient for Wing Chord is 2.2607, with a standard error of 0.9686, resulting in a t-value of 2.334 and a p-value of 0.0301. This indicates that Wing Chord is a significant predictor of Pathogen Richness at the 5% significance level. These models explain 21.41% of the variance in Pathogen Richness ($R^2 = 0.2141$), which is higher than the

PC1 model but still not very large. The F-statistic of 5.448 with a p-value of 0.03014 further confirms that the model is a better fit than the PC1 model but still only explains a small proportion of the variance in Pathogen Richness.

In conclusion, Wing Chord is a significant predictor of Pathogen Richness, whereas the first principal component (PC1) is not. However, both models explain a relatively small amount of variance in Pathogen Richness, indicating that other factors not included in these models are likely important in predicting pathogen richness.

18. Multiple Linear Regression Model for Pathogen Richness with the First Two Principal Components as Predictors

```
# Extract the principal component scores
pc1_scores <- pca_result$x[, 1]
pc2_scores <- pca_result$x[, 2]

# Fit a linear model to each relationship for principal components
fit_all_PC <- lm(chickadee_clean_rows$PathRich ~ pc1_scores + pc2_scores)

# Compare summaries
summary(fit_all_PC)
```

```
##
## Call:
## lm(formula = chickadee_clean_rows$PathRich ~ pc1_scores + pc2_scores)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5166  -3.8596   0.9544   4.7858  15.3958
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.909      1.746  23.425 1.77e-15 ***
## pc1_scores     2.099      1.127   1.862 0.078091 .
## pc2_scores     7.840      1.936   4.049 0.000684 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.191 on 19 degrees of freedom
## Multiple R-squared:  0.5111, Adjusted R-squared:  0.4597
## F-statistic: 9.933 on 2 and 19 DF,  p-value: 0.001115
```

```
summary(fit_all)
```

```
##
## Call:
## lm(formula = PathRich ~ WingChord + BirdWeight + TailLen + TarsusLen,
##     data = df_feather)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.671  -5.117   1.243   3.198  14.665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.4010    63.3530   0.243  0.81084
## WingChord     2.5854     1.6562   1.561  0.13694
## BirdWeight    0.8171     3.7004   0.221  0.82786
## TailLen       0.7090     1.3911   0.510  0.61684
## TarsusLen   -10.4426     3.1050  -3.363  0.00369 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.331 on 17 degrees of freedom
## Multiple R-squared:  0.5476, Adjusted R-squared:  0.4412
## F-statistic: 5.144 on 4 and 17 DF,  p-value: 0.006673
```

In this study, we fit a multiple linear regression model with Pathogen Richness as a response variable and predictor variables equal to the first two principal components, and compared the model with the multiple linear regression model with the morphological features as predictor variables.

In the study of the model with PC1 and PC2, PC1 has a coefficient of 2.099 and a p-value of 0.0781, indicating marginal significance. On the other hand, PC2 has a coefficient of 7.840 and a p-value of 0.000684, which is highly significant. The overall model explains 51.11% of the variance in Pathogen Richness ($R^2 = 0.5111$) and is statistically significant, with an F-statistic of 9.933 and a p-value of 0.001115. This suggests that both the first and the second principal components contribute to explaining pathogen richness in the data, with PC2 having a stronger effect than PC1.

In the study of the Multiple Linear Regression Model with Original Variables, Tarsus Length is highly significant with a p-value of 0.00369, while the other variables are not significant. The model explains 54.76% of the variance in Pathogen Richness ($R^2 = 0.5476$) and is statistically significant, with an F-statistic of 5.144 and a p-value of 0.006673.

Both models are statistically significant, but the model with the original variables (Wing Chord, Bird Weight, Tail Length, Tarsus Length) has a slightly higher R^2 (54.76%) compared to the model with PC1 and PC2 (51.11%).

In summary, using the first two principal components provides a lower dimensional, but nearly equally effective way of explaining variance in Pathogen Richness compared to using the original four variables. However, then the model with the original variables provides detail of biological features.

18.2. Simple Linear Regression Model for Pathogen Richness with the Second Principal Component as a Predictor

As the first principal component was not found to be statistically significant in prediction with linear regression, let us test the second principal component:

```
# Fit a linear model to each relationship for principal components
fit_PC2 <- lm(chickadee_clean_rows$PathRich ~ pc2_scores)

# Compare summaries
summary(fit_PC2)
```

```
##
## Call:
## lm(formula = chickadee_clean_rows$PathRich ~ pc2_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.262  -4.527   2.263   3.959  18.002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.909      1.851   22.10 1.59e-15 ***
## pc2_scores     7.840      2.052    3.82  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.682 on 20 degrees of freedom
## Multiple R-squared:  0.4219, Adjusted R-squared:  0.393
## F-statistic:  14.6 on 1 and 20 DF,  p-value: 0.00107
```

The F-statistic is 14.6 with a p-value of 0.00107, below the alpha level of 0.05. We reject the null hypothesis and conclude from this data that this model is statistically significant in explaining the variance in the response variable, pathogen richness. The predictor variable, the second principal component, has a t-value of 3.82, below the alpha level of 0.05. We conclude that the second principal component is statistically significant in predicting pathogen richness. The multiple R-squared value is 0.4219, and this suggests that approximately 42.19% of the variability in pathogen richness is explained by the second principal component.

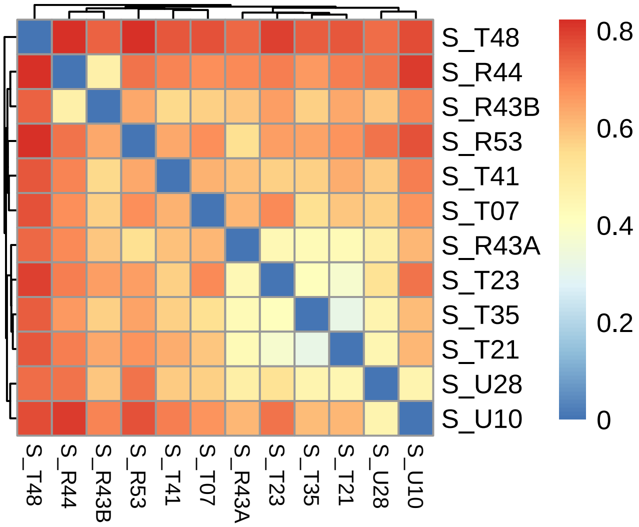
While the first principal component captured more of the variance in the predictor variables, this evidence demonstrates that this does not necessarily relate to how well the component predicts the response variable. While counter-intuitive at first glace - the second principal component in this case was superior to predicting pathogen richness.

19. Metric Multidimensional Scaling (MDS) to Assess the Composition of Microbial Communities Relating to Habitat or Source

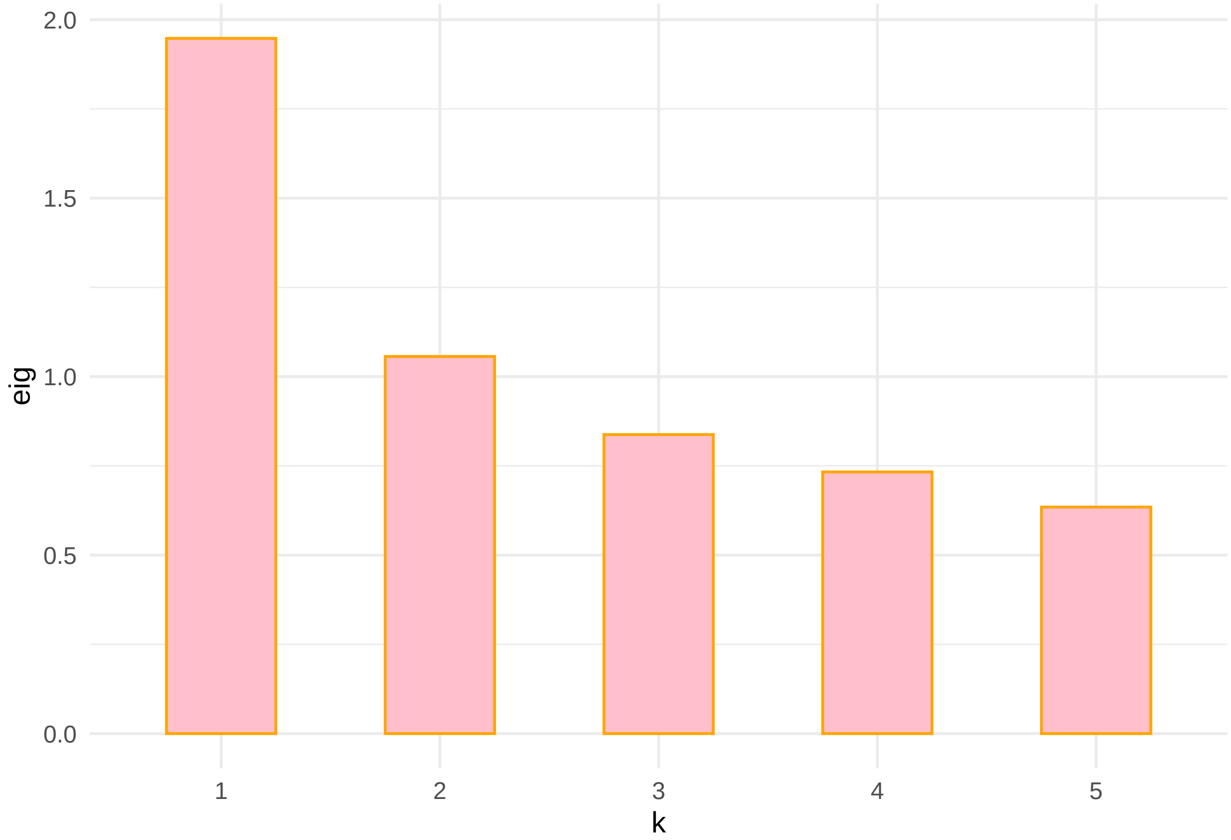
```r
# Microbial community composition data for 2607 taxa was also used to calculate Bray Curtis dissimilari
distChickadee <- read.csv("ChickadeeDissimilarities.csv", row.names=1)
distChickadee[1:6, 1:6]
```

```
##           S_U28     S_U10     S_T48     S_T41     S_T35     S_T23
## S_U28 0.0000000 0.4549776 0.7278924 0.5815247 0.4540807 0.5310314
## S_U10 0.4549776 0.0000000 0.7803587 0.7006278 0.6045740 0.7197309
## S_T48 0.7278924 0.7803587 0.0000000 0.7643946 0.7518386 0.7924664
## S_T41 0.5815247 0.7006278 0.7643946 0.0000000 0.5683408 0.5704036
## S_T35 0.4540807 0.6045740 0.7518386 0.5683408 0.0000000 0.4130942
## S_T23 0.5310314 0.7197309 0.7924664 0.5704036 0.4130942 0.0000000
```

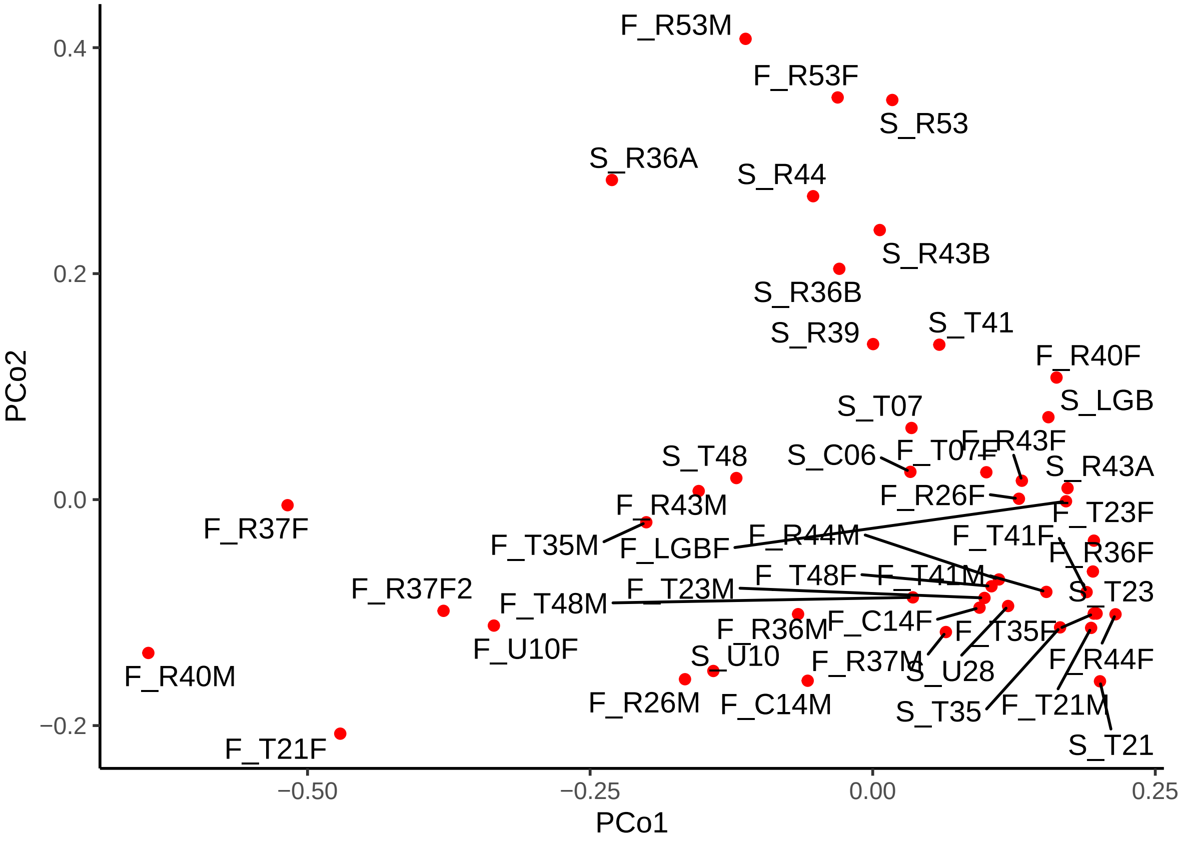
```r
# clustered heatmap
pheatmap(distChickadee[1:12, 1:12], cluster_rows = TRUE,
treeheight_row = 0.0001, treeheight_col = 0.8,
fontsize_col = 8, cellwidth = 13, cellheight = 13)
```



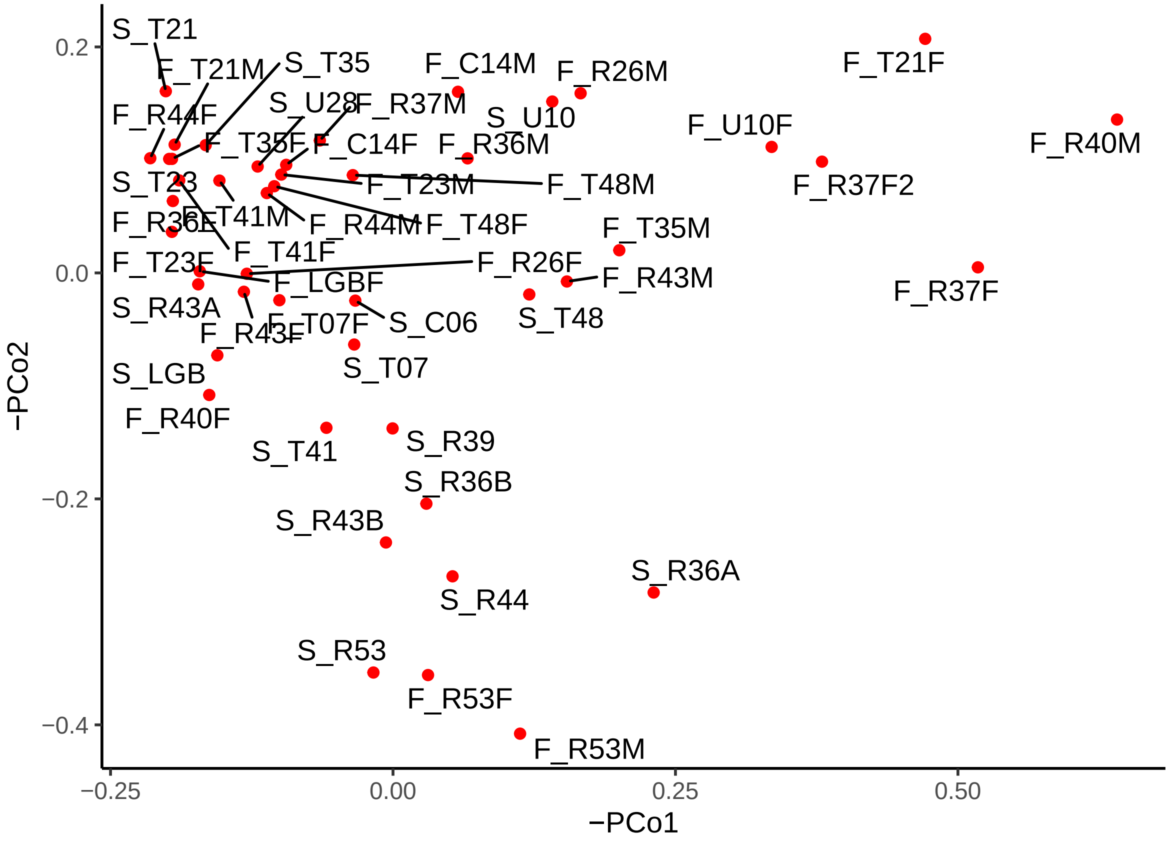
```
# Classical (metric) multidimensional scaling (MDS), also known as principal coordinates analysis
MDSChickadee <- cmdscale(distChickadee, eig = TRUE)

# Create a plotbar function to plot the eigenvalues in a scree plot.
plotbar <- function(res, m = 9) {
ggplot(data.frame(list(eig = res$eig[seq_len(m)],
k = seq(along = res$eig[seq_len(m)]))),
aes(x = k, y = eig)) +
scale_x_discrete("k", limits = factor(seq_len(m))) +
theme_minimal() +
geom_bar(stat="identity", width=0.5, color="orange",
fill="pink")
}

plotbar(MDSChickadee, m = 5)
```



```
# Project the cities onto the first two coordinates created from the distances.
MDSChick <- data.frame(list(PCo1 = MDSChickadee$points[, 1],
PCo2 = MDSChickadee$points[, 2],
labs = rownames(MDSChickadee$points)))

ggplot(MDSChick, aes(x = PCo1, y = PCo2, label = labs)) +
geom_point(color = "red") +
#xlim(-1950, 2000) +
#ylim(-1150, 1200) +
coord_fixed() +
geom_text_repel(size = 4, max.overlaps = 100)
```



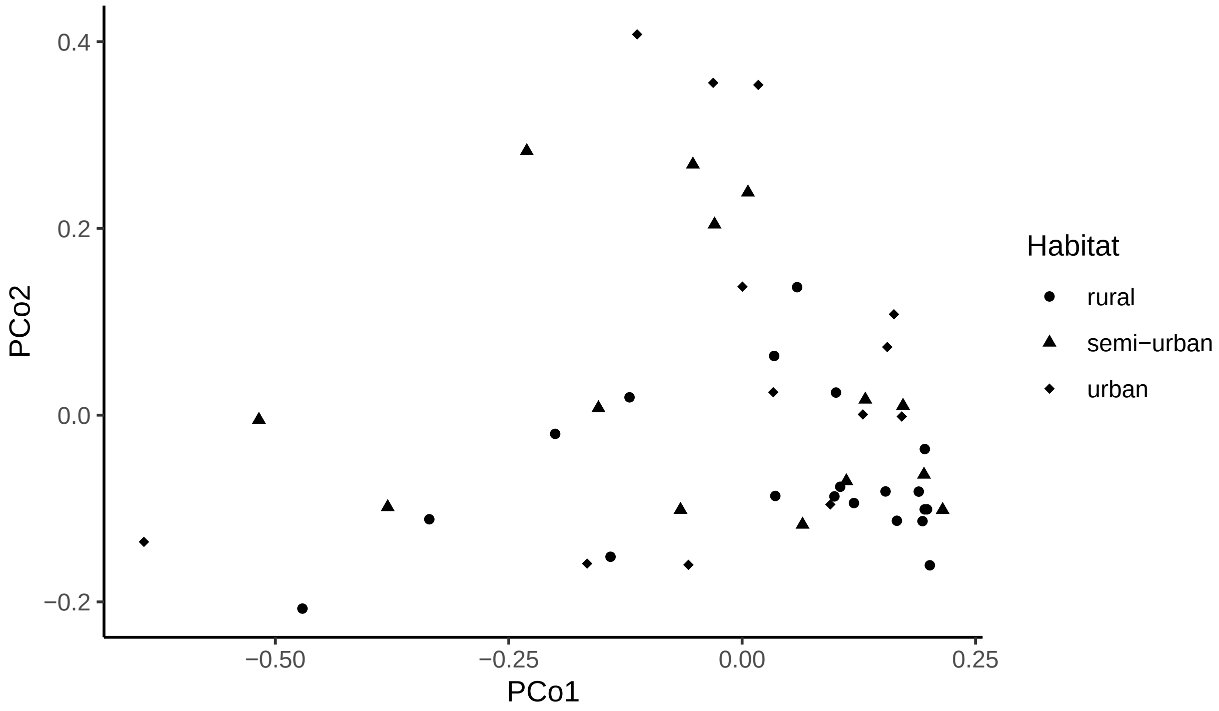
```
#To re-orient the "map" so north is at the top and west is on the left, reverse the signs of the princi
ggplot(MDSChick, aes(x = -PCo1, y = -PCo2, label = labs)) +
geom_point(color = "red") +
coord_fixed() +
geom_text_repel(size = 4, max.overlaps = 100)
```



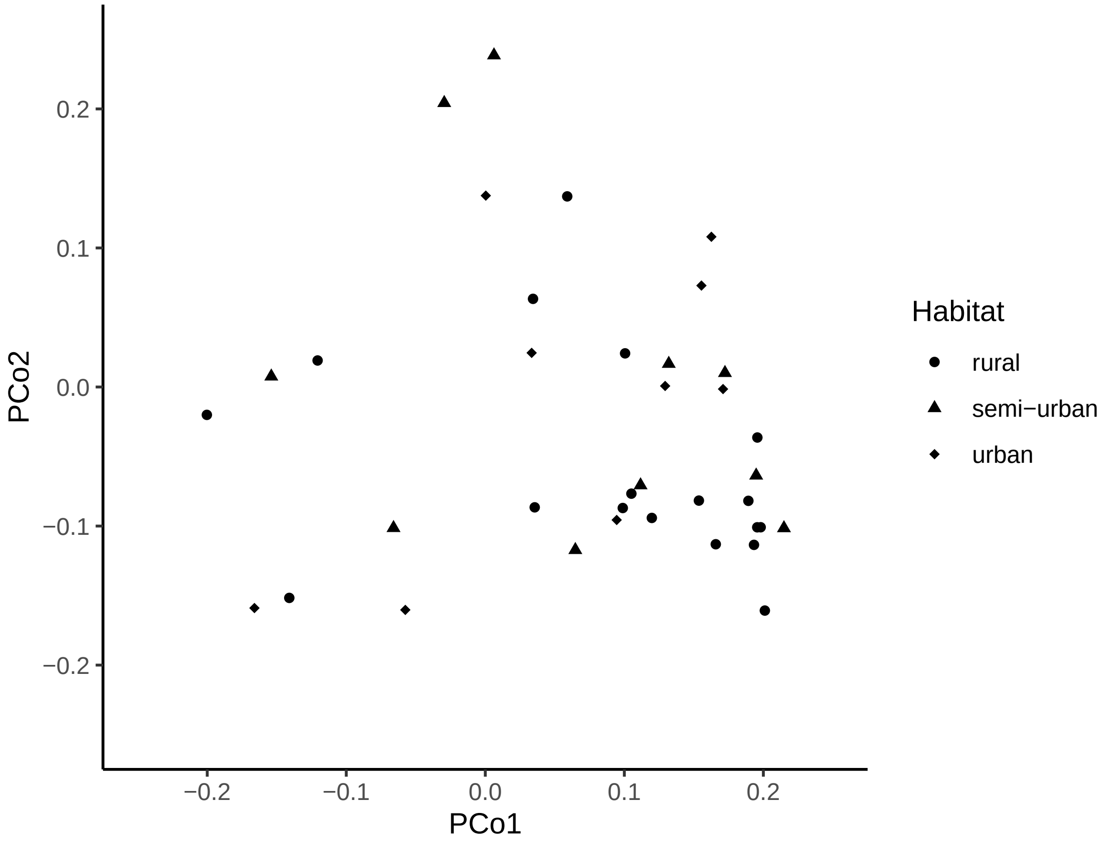
```
# Merge the datasets
MDSChick_merged <- merge(MDSChick, chickadeeData, by.x = "labs", by.y = "Site")
print(MDSChick_merged)
```

```
##       labs          PCo1          PCo2    Habitat  Source EscShi Entero
## 1    F_C14F  0.0945176738 -0.0956366887      urban feather    261    286
## 2    F_C14M -0.0575279670 -0.1603054570      urban feather   2269    288
## 3    F_LGBF  0.1710053413 -0.0014732449      urban feather     54     12
## 4    F_R26F  0.1293637482  0.0007389586      urban feather    173      9
## 5    F_R26M -0.1660559008 -0.1589631739      urban feather   3771      8
## 6    F_R36F  0.1948363956 -0.0636382630 semi-urban feather     31     48
## 7    F_R36M -0.0660126789 -0.1013814758 semi-urban feather   1228     28
## 8    F_R37F -0.5176652570 -0.0049461669 semi-urban feather    379      7
## 9   F_R37F2 -0.3796891196 -0.0984423101 semi-urban feather   3230      0
## 10   F_R37M  0.0647565721 -0.1172100349 semi-urban feather   1294     20
## 11   F_R40F  0.1626247574  0.1080359473      urban feather     78     13
## 12   F_R40M -0.6408373846 -0.1357161329      urban feather   1267      0
## 13   F_R43F  0.1319720024  0.0167027419 semi-urban feather    112     28
## 14   F_R43M -0.1539284826  0.0075834838 semi-urban feather    882     19
## 15   F_R44F  0.2148287227 -0.1014979798 semi-urban feather     33     20
## 16   F_R44M  0.1116849107 -0.0706728572 semi-urban feather    191     13
## 17   F_R53F -0.0309991685  0.3559239560      urban feather     61      2
## 18   F_R53M -0.1124679050  0.4078016917      urban feather    288      3
## 19   F_T07F  0.1005505545  0.0241963920      rural feather      2     10
## 20   F_T21F -0.4710453613 -0.2071537760      rural feather   6520      0
## 21   F_T21M  0.1932527109 -0.1135244212      rural feather     28     10
## 22  F_T23F  0.1956936234 -0.0363216270      rural feather     17     11
## 23  F_T23M  0.0988841563 -0.0870280597      rural feather    906      9
## 24  F_T35F  0.1956076519 -0.1008927097      rural feather    113      1
## 25  F_T35M -0.2002829223 -0.0200427959      rural feather    203      0
## 26  F_T41F  0.1892287367 -0.0818687660      rural feather     22     45
## 27  F_T41M  0.1536271344 -0.0817013388      rural feather    305     45
## 28  F_T48F  0.1050834309 -0.0767038593      rural feather    180     63
## 29  F_T48M  0.0355450025 -0.0865386214      rural feather    232     63
## 30  F_U10F -0.3351211552 -0.1115072455      rural feather    486    374
## 31   S_C06  0.0333484454  0.0245499265      urban    nest     63    687
## 32   S_LGB  0.1554716109  0.0729537142      urban    nest      1      9
## 33  S_R36A -0.2307061634  0.2828755147 semi-urban    nest    486    615
## 34  S_R36B -0.0295427057  0.2042765470 semi-urban    nest     30     39
## 35   S_R39  0.0003591343  0.1376404078      urban    nest    105    202
## 36  S_R43A  0.1723854677  0.0101228407 semi-urban    nest     56     31
## 37  S_R43B  0.0062437764  0.2385802916 semi-urban    nest      8     49
## 38   S_R44 -0.0527041867  0.2685259001 semi-urban    nest      4      9
## 39   S_R53  0.0173372690  0.3536829110      urban    nest      0     15
## 40   S_T07  0.0343270680  0.0633972164      rural    nest      4     20
## 41   S_T21  0.2010769761 -0.1607999373      rural    nest     14     22
## 42   S_T23  0.1980790405 -0.1008693849      rural    nest      8     13
## 43   S_T35  0.1657871944 -0.1130826741      rural    nest    248      4
## 44   S_T41  0.0589319039  0.1370867488      rural    nest      1     22
## 45   S_T48 -0.1206273343  0.0190762073      rural    nest      0     84
## 46   S_U10 -0.1409891998 -0.1516735419      rural    nest   1429    546
## 47   S_U28  0.1197918807 -0.0941588537      rural    nest     86    557
##    CommRich PathRich WingChord BirdWeight TailLen TarsusLen BirdSex
## 1       low       50      63.0          11      54      18.1       F
## 2      high       58      68.0          12      60      18.0       M
## 3      high       44      64.0          12      57      17.9       F
## 4      high       63      65.0          15      60      16.9       F
## 5      high       53      68.0          17      57      18.7       M
## 6       low       51      65.0          11      58      17.9       F
## 7      high       53      67.0          11      58      17.9       M
## 8       low       17      62.0          10      55      18.5       F
## 9       low       30      62.0          10      55      18.5       F
## 10     high       52      67.0          12      60      19.5       M
## 11     high       54      66.0          15      60      17.1       F
## 12      low       11      69.0          16      61      18.6       M
## 13     high       59      64.0          13      55      19.0       F
## 14     high       88      66.0          12      60      18.6       M
## 15      low       38      62.0          10      56      16.9       F
## 16     high       66      64.0          11      60      18.0       M
## 17     high       58      65.0          11      56      16.8       F
## 18     high       56      68.0          11      62      18.2       M
## 19      low       40      63.0          11      54      18.1       F
## 20      low       23      65.0          12      57      19.0       F
## 21      low       41      62.5          11      54      18.2       M
## 22      low       35      67.0          12      59      18.6       F
## 23      low       37      67.0          12      60      18.9       M
## 24      low       38      63.0          11      57      18.1       F
## 25      low       29      66.0          12      55      18.9       M
## 26      low       37      63.0          11      56      18.4       F
## 27      low       36      66.0          11      56      19.2       M
## 28    high       49       63.0         15      55      19.0       F
## 29     low       44       63.0         11      57      18.3       M
## 30     low       33       60.0         11      53      17.6       F
## 31    high       63         NA         NA      NA        NA    <NA>
## 32    high       59         NA         NA      NA        NA    <NA>
## 33     low       30         NA         NA      NA        NA    <NA>
## 34     low       41         NA         NA      NA        NA    <NA>
## 35    high       57         NA         NA      NA        NA    <NA>
## 36    high       57         NA         NA      NA        NA    <NA>
## 37    high       40         NA         NA      NA        NA    <NA>
## 38     low       32         NA         NA      NA        NA    <NA>
## 39     low       49         NA         NA      NA        NA    <NA>
## 40    high       48         NA         NA      NA        NA    <NA>
## 41     low       45         NA         NA      NA        NA    <NA>
## 42     low       31         NA         NA      NA        NA    <NA>
## 43     low       38         NA         NA      NA        NA    <NA>
## 44     low       41         NA         NA      NA        NA    <NA>
## 45     low       31         NA         NA      NA        NA    <NA>
## 46     low       36         NA         NA      NA        NA    <NA>
## 47     low       41         NA         NA      NA        NA    <NA>
```

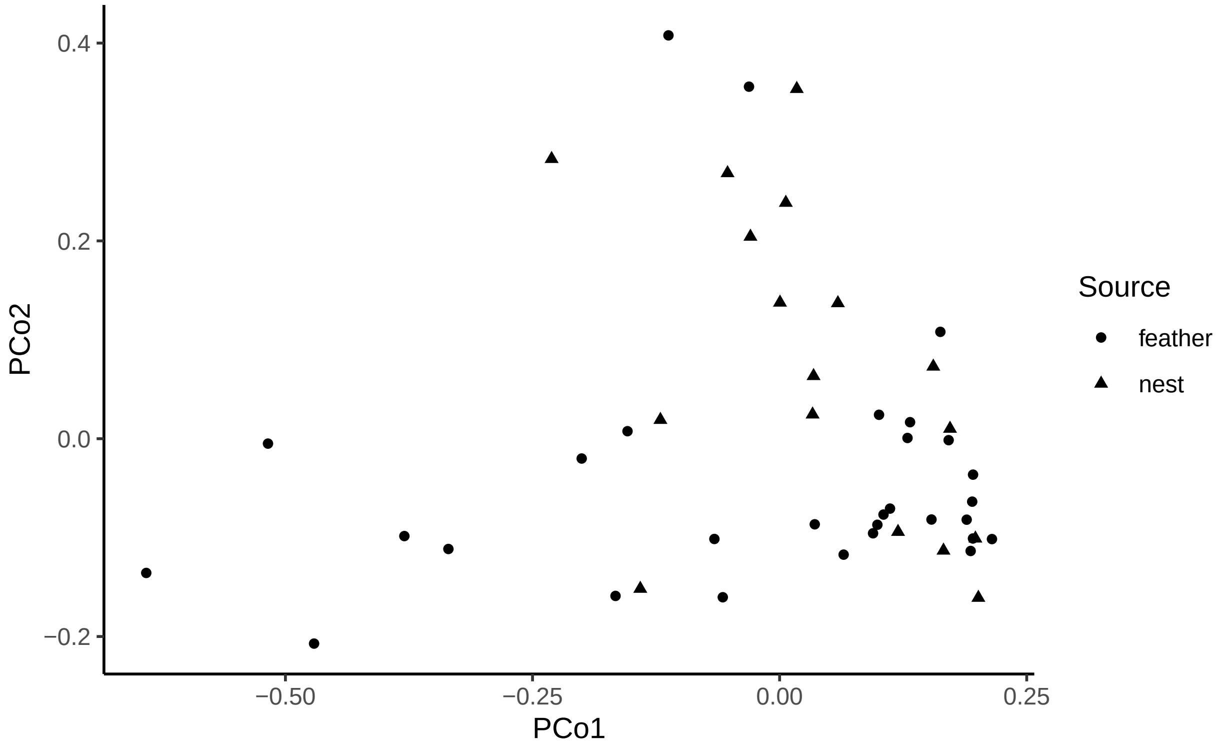
```
# By Habitat
ggplot(MDSChick_merged, aes(x = PCo1, y = PCo2)) +
  geom_point(aes(shape = Habitat)) +
  coord_fixed() +
  scale_shape_manual(values = c(16, 17, 18))
```



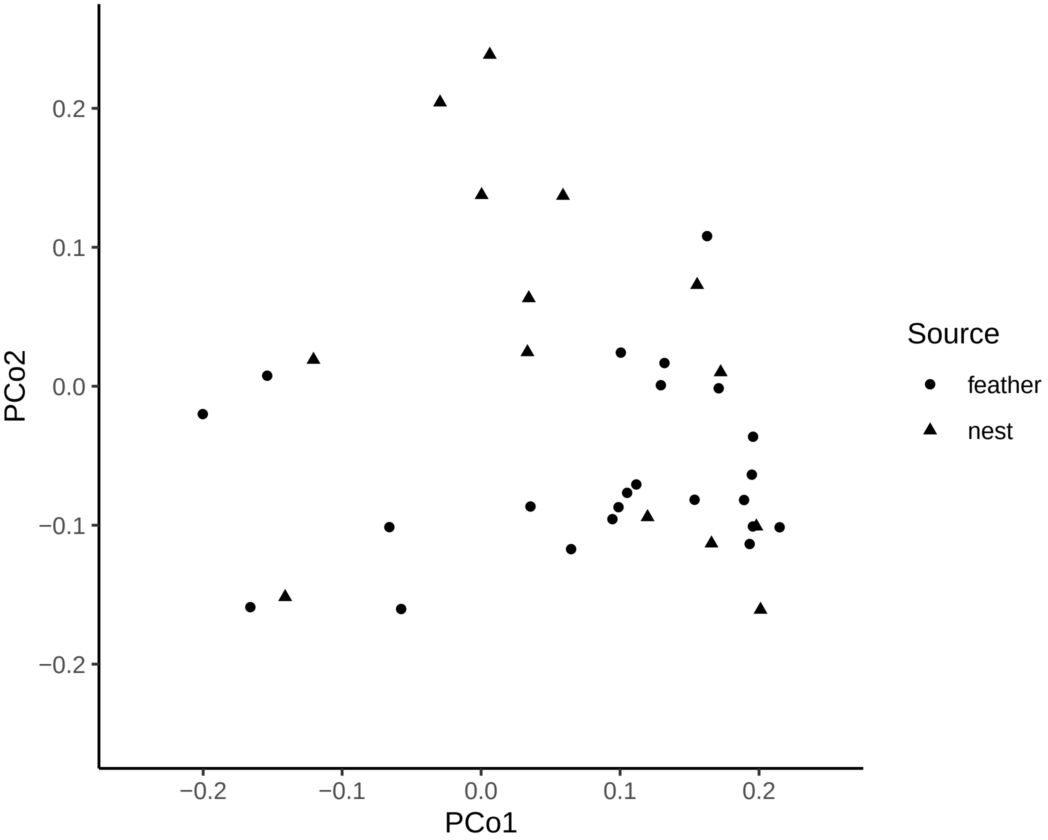
```r
# Zoom in
ggplot(MDSChick_merged, aes(x = PCo1, y = PCo2)) +
  geom_point(aes(shape = Habitat)) +
  xlim(-0.25, 0.25) +
  ylim(-0.25, 0.25) +
  coord_fixed() +
  scale_shape_manual(values = c(16, 17, 18))
```



```r
# By Source
ggplot(MDSChick_merged, aes(x = PCo1, y = PCo2)) +
  geom_point(aes(shape = Source)) +
  coord_fixed() +
  scale_shape_manual(values = c(16, 17))
```

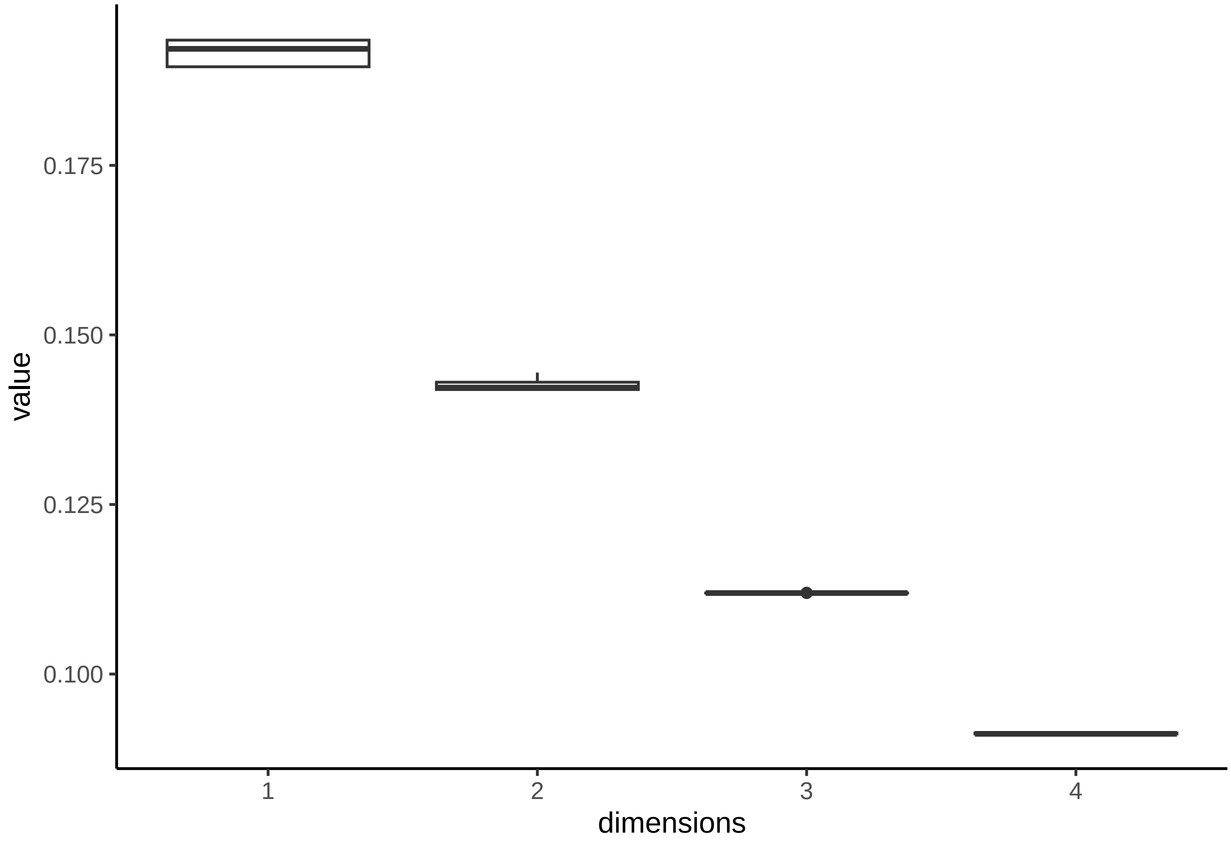


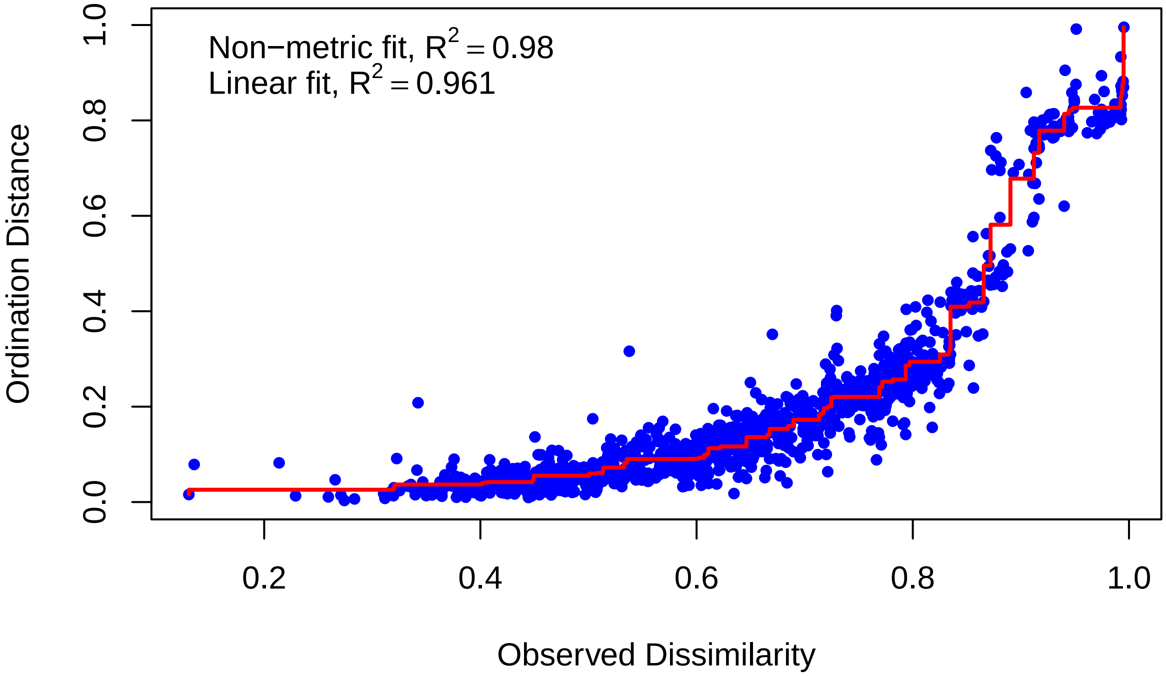
```
# Zoom in
ggplot(MDSChick_merged, aes(x = PCo1, y = PCo2)) +
  geom_point(aes(shape = Source)) +
  xlim(-0.25, 0.25) +
  ylim(-0.25, 0.25) +
  coord_fixed() +
  scale_shape_manual(values = c(16, 17))
```

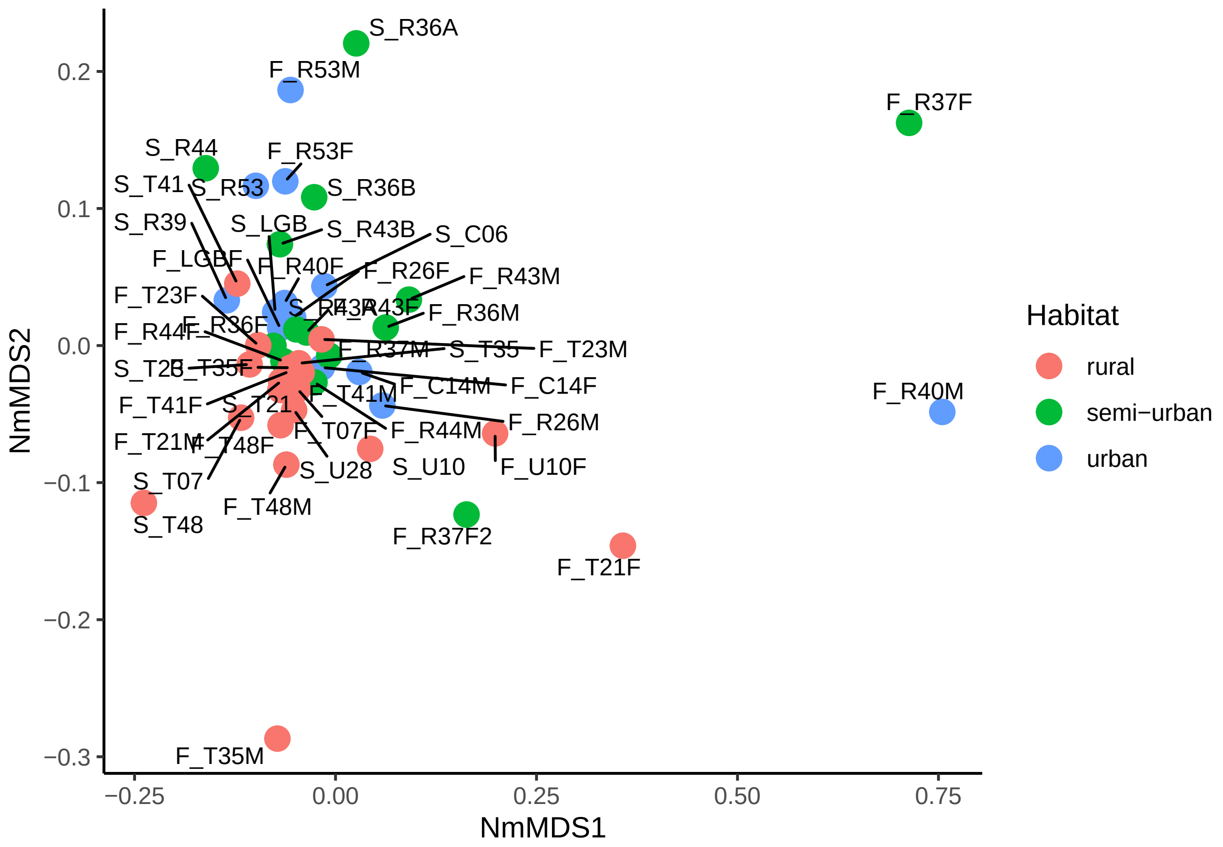


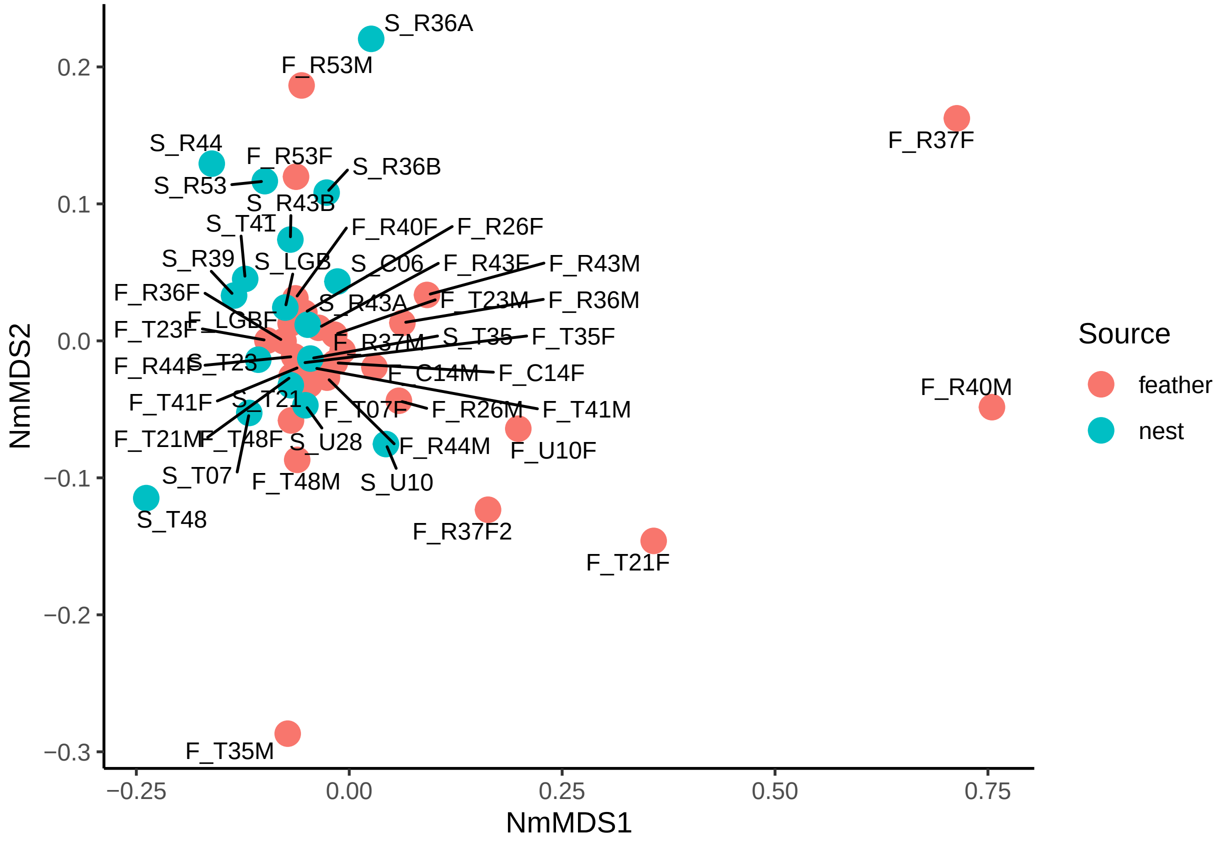
In this study, we applied metric multidimensional scaling (MDS) to the dissimilarities dataset to assess if the composition of the microbial communities appear to be related more to Habitat or Source. Upon visual inspection of the MDS, microbial community composition appears to be more closely associated with Source, as the different categories of Source are more closely clustered.

20. Nonmetric Multidimensional Scaling (NMDS) to Assess the Composition of Microbial Communities Relating to Habitat or Source






S_R36A
F_R53M
F_R37F
S_R44
F_R53F
S_T41
S_R53
S_R36B
S_R39
S_LGB
S_R43B
S_C06
F_LGBF
F_R40F
F_R26F
F_R43M
F_T23F
S_R43A
F_R43F
F_R36M
F_R44F
F_R36F
F_R37M
S_T35
F_T23M
S_T28
F_T35F
F_C14M
F_C14F
F_T41M
F_R40M
F_T41F
S_T21
F_T07F
F_R44M
F_R26M
F_T21M
F_T48F
S_U28
S_U10
F_U10F
S_T07
F_T48M
S_T48
F_R37F2
F_T21F
F_T35M
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.25
0.00
0.25
0.50
0.75
NmMDS1
NmMDS2
Habitat
rural
semi−urban
urban



In this study, we applied non-metric multidimensional scaling (NMDS) to the dissimilarities dataset to assess if the composition of the microbial communities appear to be related more to Habitat or Source.

The Shepard plot suggests that both the linear and non-metric fits are excellent fits, and indicates that the lower-dimensional representation of the data is capturing almost all of the dissimilarity structure in the original high-dimensional data. The higher R^2 value in the non-metric Fit suggests that the NMDS solution is a slightly better model, strictly considering R^2.

From these scatter plots, the composition of the microbial communities appear to be related more to Source, rather than Habitat, as the different categories of Source are more closely clustered.

References

1. www.tru.ca, Thompson Rivers University. "BIOL 4001: Biostatistics." *Thompson Rivers University*, http://www.tru.ca/distance/courses/biol4001.html. Accessed 20 Aug. 2023.

2. *Introduction to R*. https://www.zoology.ubc.ca/~bio501/R/workshops/workshops-intro.html. Accessed 20 Aug. 2023.

3. *Resources for The Analysis of Biological Data*. https://whitlockschluter3e.zoology.ubc.ca/index.html. Accessed 20 Aug. 2023.