563Project

Team Name: Hackberry Penguins; Team: Weiyi You, Peiyuan Gao

2025-04-06

Data

```
data <- read.csv("penguins_cleaned.csv")
head(data)
```

```
##   studyName Sample.Number                             Species Region    Island
## 1   PAL0708             2 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 2   PAL0708             3 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 3   PAL0708             5 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 4   PAL0708             6 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 5   PAL0708             7 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 6   PAL0708             8 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
##                 Stage Individual.ID Clutch.Completion  Date.Egg Culmen_Length
## 1 Adult, 1 Egg Stage          N1A2               Yes 2011/11/7          39.5
## 2 Adult, 1 Egg Stage          N2A1               Yes  11/16/07          40.3
## 3 Adult, 1 Egg Stage          N3A1               Yes  11/16/07          36.7
## 4 Adult, 1 Egg Stage          N3A2               Yes  11/16/07          39.3
## 5 Adult, 1 Egg Stage          N4A1                No  11/15/07          38.9
## 6 Adult, 1 Egg Stage          N4A2                No  11/15/07          39.2
##   Culmen_Depth Flipper_Length Body_Mass    Sex Delta.15.N Delta.13.C
## 1         17.4            186      3800 FEMALE    8.94956  -24.69454
## 2         18.0            195      3250 FEMALE    8.36821  -25.33302
## 3         19.3            193      3450 FEMALE    8.76651  -25.32426
## 4         20.6            190      3650   MALE    8.66496  -25.29805
## 5         17.8            181      3625 FEMALE    9.18718  -25.21799
## 6         19.6            195      4675   MALE    9.46060  -24.89958
```

```
data$Island <- as.factor(data$Island)
data$Sex <- as.factor(data$Sex)
data$Species <- as.factor(data$Species)
```

```
summary(data)
```

```
##   studyName         Sample.Number
##  Length:324         Min.   :  1.00
##  Class :character   1st Qu.: 31.00
##  Mode  :character   Median : 59.50
##                     Mean   : 64.60
##                     3rd Qu.: 96.25
##                     Max.   :152.00
##                                        Species       Region
##  Adelie Penguin (Pygoscelis adeliae)      :139   Length:324
##  Chinstrap penguin (Pygoscelis antarctica): 67   Class :character
##  Gentoo penguin (Pygoscelis papua)        :118   Mode  :character
##
##
##
##        Island       Stage           Individual.ID      Clutch.Completion
##  Biscoe   :162   Length:324         Length:324         Length:324
##  Dream    :119   Class :character   Class :character   Class :character
##  Torgersen: 43   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Date.Egg         Culmen_Length    Culmen_Depth   Flipper_Length
##  Length:324         Min.   :32.10   Min.   :13.10   Min.   :172.0
##  Class :character   1st Qu.:39.50   1st Qu.:15.57   1st Qu.:190.0
##  Mode  :character   Median :44.95   Median :17.30   Median :197.0
##                     Mean   :44.05   Mean   :17.13   Mean   :201.2
##                     3rd Qu.:48.70   3rd Qu.:18.60   3rd Qu.:213.0
##                     Max.   :59.60   Max.   :21.50   Max.   :231.0
##    Body_Mass        Sex         Delta.15.N       Delta.13.C
##  Min.   :2700   FEMALE:163   Min.   : 7.632   Min.   :-27.02
##  1st Qu.:3550   MALE  :161   1st Qu.: 8.304   1st Qu.:-26.33
##  Median :4050                Median : 8.659   Median :-25.84
##  Mean   :4214                Mean   : 8.740   Mean   :-25.69
##  3rd Qu.:4800                3rd Qu.: 9.181   3rd Qu.:-25.06
##  Max.   :6300                Max.   :10.025   Max.   :-23.89
```

```
#install.packages("summarytools")  # Run if not installed
library(summarytools)
```

```
dfSummary(data)
```

```
## Data Frame Summary
## data
## Dimensions: 324 x 16
## Duplicates: 0
##
## -----------------------------------------------------------------------------------------------------
## No   Variable            Stats / Values                   Freqs (% of Valid)    Graph
## ---- ------------------- ------------------------------ --------------------- ----------------------
## 1    studyName           1. PAL0708                        95 (29.3%)           IIIII
##      [character]         2. PAL0809                       113 (34.9%)           IIIIII
##                          3. PAL0910                       116 (35.8%)           IIIIIII
##
## 2    Sample.Number       Mean (sd) : 64.6 (40.2)          152 distinct values   . : :
##      [integer]           min < med < max:                                       : : : : .
##                          1 < 59.5 < 152                                         : : : : : :
##                          IQR (CV) : 65.2 (0.6)                                  : : : : : : :
##                                                                                 : : : : : : : :
##
```

```
## 3    Species             1. Adelie Penguin (Pygosceli   139 (42.9%)           IIIIIIII
##      [factor]            2. Chinstrap penguin (Pygosc    67 (20.7%)           IIII
##                          3. Gentoo penguin (Pygosceli   118 (36.4%)           IIIIIII
##
## 4    Region              1. Anvers                      324 (100.0%)          IIIIIIIIIIIIIIIIIIII
##      [character]
##
## 5    Island              1. Biscoe                      162 (50.0%)           IIIIIIIIII
##      [factor]            2. Dream                       119 (36.7%)           IIIIIII
##                          3. Torgersen                    43 (13.3%)           II
##
## 6    Stage               1. Adult, 1 Egg Stage          324 (100.0%)          IIIIIIIIIIIIIIIIIIII
##      [character]
##
## 7    Individual.ID       1. N13A1                         3 ( 0.9%)
##      [character]         2. N13A2                         3 ( 0.9%)
##                          3. N18A1                         3 ( 0.9%)
##                          4. N18A2                         3 ( 0.9%)
##                          5. N21A1                         3 ( 0.9%)
##                          6. N21A2                         3 ( 0.9%)
##                          7. N22A1                         3 ( 0.9%)
##                          8. N22A2                         3 ( 0.9%)
##                          9. N23A1                         3 ( 0.9%)
##                          10. N23A2                        3 ( 0.9%)
##                          [ 178 others ]                 294 (90.7%)           IIIIIIIIIIIIIIIIII
##
## 8    Clutch.Completion   1. No                           34 (10.5%)           II
##      [character]         2. Yes                         290 (89.5%)           IIIIIIIIIIIIIIIII
##
## 9    Date.Egg            1. 11/27/07                     18 ( 5.6%)           I
##      [character]         2. 2011/9/8                     15 ( 4.6%)
##                          3. 11/18/09                     14 ( 4.3%)
##                          4. 11/16/07                     13 ( 4.0%)
##                          5. 11/13/08                     12 ( 3.7%)
##                          6. 11/21/09                     12 ( 3.7%)
##                          7. 2011/4/8                     12 ( 3.7%)
##                          8. 2011/6/8                     12 ( 3.7%)
##                          9. 11/14/08                     10 ( 3.1%)
##                          10. 11/15/09                    10 ( 3.1%)
##                          [ 40 others ]                  196 (60.5%)           IIIIIIIIIIII
##
## 10   Culmen_Length       Mean (sd) : 44.1 (5.5)         159 distinct values             . :
##      [numeric]           min < med < max:                                       . : : : : :
##                          32.1 < 45 < 59.6                                       : : : : : :
##                          IQR (CV) : 9.2 (0.1)                                   : : : : : :
##                                                                               . : : : : : : : .
##
## 11   Culmen_Depth        Mean (sd) : 17.1 (2)           79 distinct values              :
##      [numeric]           min < med < max:                                       .   . : :
##                          13.1 < 17.3 < 21.5                                     : . : : : .
##                          IQR (CV) : 3 (0.1)                                   . : : : : : : :
##                                                                               : : : : : : : . .
##
## 12   Flipper_Length      Mean (sd) : 201.2 (14)         53 distinct values          :
```

```
##      [integer]           min < med < max:                                         . :
##                          172 < 197 < 231                                          : : :   : .
##                          IQR (CV) : 23 (0.1)                                      : : :   : : :
##                                                                                  : : : : : : : : :
##
## 13   Body_Mass           Mean (sd) : 4214 (809.3)       93 distinct values        :
##      [integer]           min < med < max:                                       . :
##                          2700 < 4050 < 6300                                     : : : :
##                          IQR (CV) : 1250 (0.2)                                  : : : : : .
##                                                                                . : : : : : :
##
## 14   Sex                 1. FEMALE                      163 (50.3%)           IIIIIIIIII
##      [factor]            2. MALE                        161 (49.7%)           IIIIIIIII
##
## 15   Delta.15.N          Mean (sd) : 8.7 (0.6)          324 distinct values       . :
##      [numeric]           min < med < max:                                         : : : . .
##                          7.6 < 8.7 < 10                                           : : : : : : :
##                          IQR (CV) : 0.9 (0.1)                                   : : : : : : : : .
##                                                                                : : : : : : : : : :
##
## 16   Delta.13.C          Mean (sd) : -25.7 (0.8)        324 distinct values       :
##      [numeric]           min < med < max:                                         :
##                          -27 < -25.8 < -23.9                                    : : . : .
##                          IQR (CV) : 1.3 (0)                                     : : : : : .
##                                                                                 : : : : : :
## ----------------------------------------------------------------------------------------------------
```

Multivariate Analysis

Principle Components

plots showing relationships between variables and multivariate normality

```r
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(PerformanceAnalytics)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

```
library(heplots)
```

```
## Loading required package: broom
```

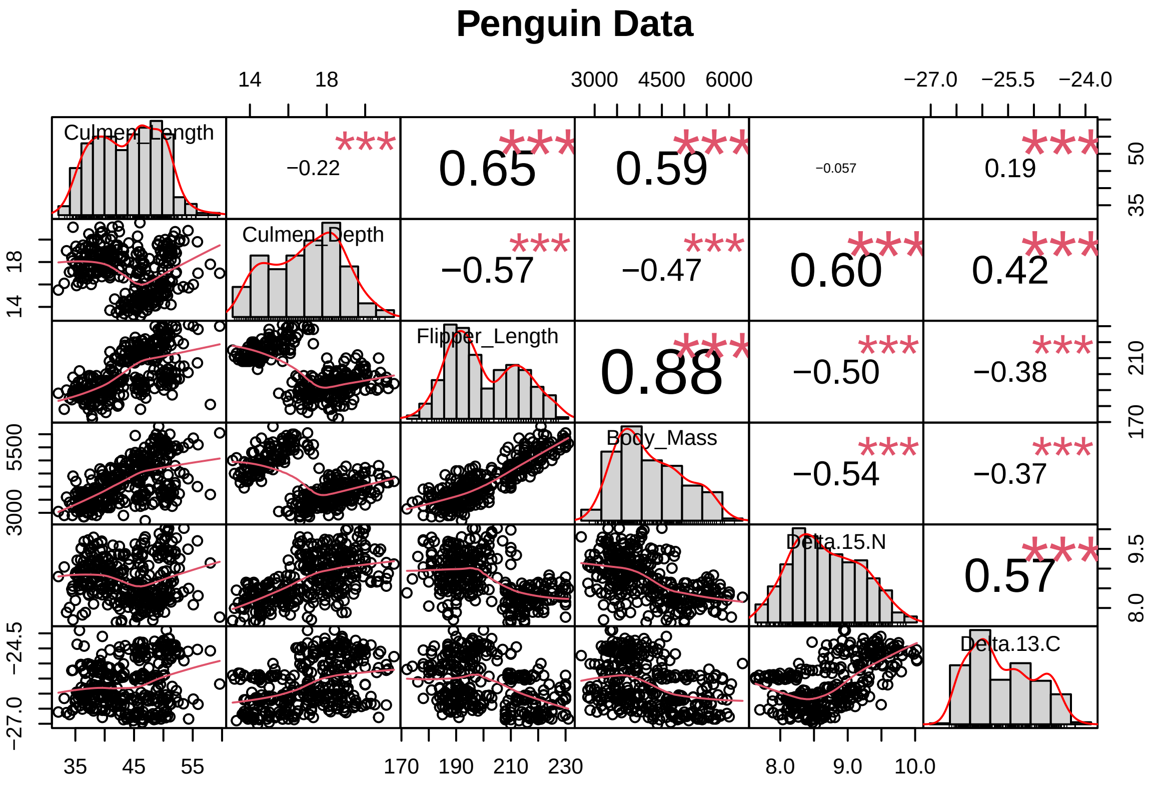
```
library(FactoMineR)
library(dplyr)
```

```
##
## ######################### Warning from 'xts' package ##########################
## #                                                                             #
## # The dplyr lag() function breaks how base R's lag() function is supposed to  #
## # work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or       #
## # source() into this session won't work correctly.                            #
## #                                                                             #
## # Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
## # conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop           #
## # dplyr from breaking base R's lag() function.                                #
## #                                                                             #
## # Code in packages is not affected. It's protected by R's namespace mechanism #
## # Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
## #                                                                             #
## ###############################################################################

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:xts':
##
##     first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_PCA <- data[,c("Individual.ID","Species","Island","Culmen_Length","Culmen_Depth","Flipper_Length","B
levels(df_PCA$Species) <- c("Adelie", "Chinstrap", "Gentoo")
head(df_PCA)
```

```
##   Individual.ID Species    Island Culmen_Length Culmen_Depth Flipper_Length
## 1          N1A2  Adelie Torgersen          39.5         17.4            186
## 2          N2A1  Adelie Torgersen          40.3         18.0            195
## 3          N3A1  Adelie Torgersen          36.7         19.3            193
## 4          N3A2  Adelie Torgersen          39.3         20.6            190
## 5          N4A1  Adelie Torgersen          38.9         17.8            181
## 6          N4A2  Adelie Torgersen          39.2         19.6            195
##   Body_Mass Delta.15.N Delta.13.C
## 1      3800    8.94956  -24.69454
## 2      3250    8.36821  -25.33302
## 3      3450    8.76651  -25.32426
## 4      3650    8.66496  -25.29805
## 5      3625    9.18718  -25.21799
## 6      4675    9.46060  -24.89958
```

```
#look for non-linearity, get correlation, make histograms.
chart.Correlation(df_PCA[,-c(1,2,3)], main = "Penguin Data")
```



Pairwise correlation matrix plot includes:

Correlations

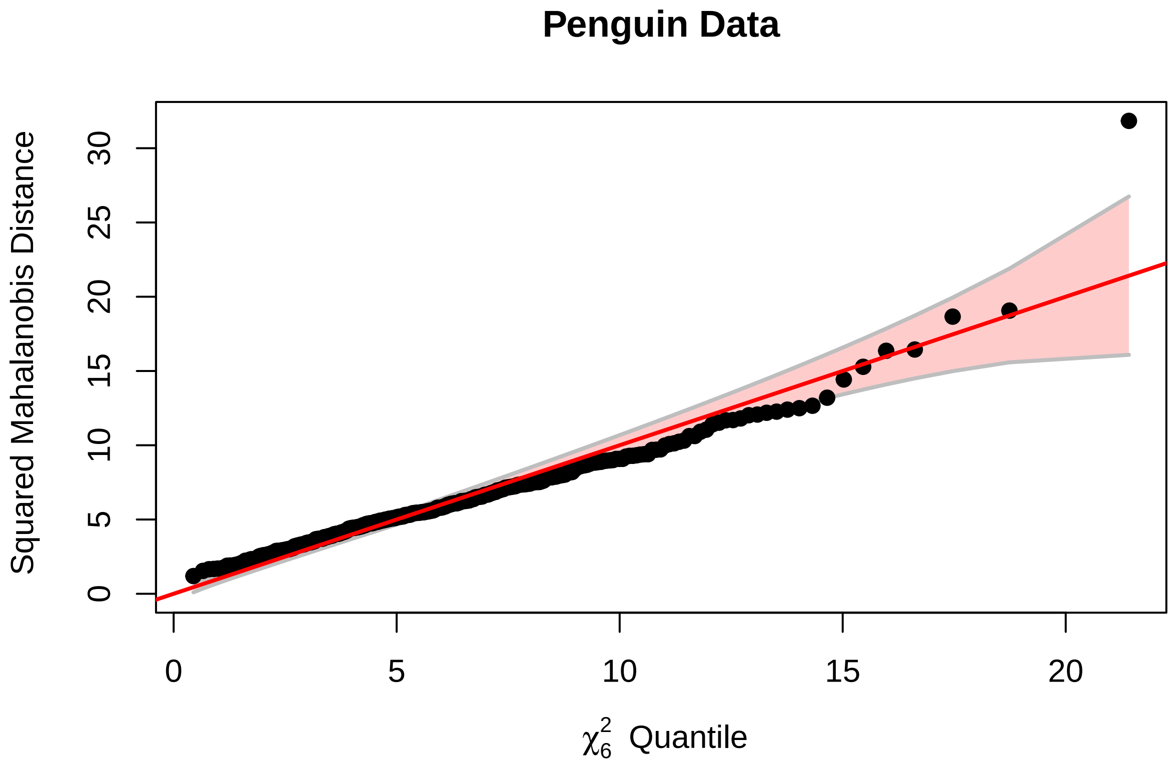
Strong Positive Correlations (Strong Linear Relationships (Positive)): Flipper_Length vs Body_Mass: r = 0.88, penguins with longer flippers tend to weigh more — very strong and statistically significant. A tight, upward-trending scatter of points indicates a strong positive linear relationship. Culmen_Length vs Flipper_Length: r = 0.65, longer bills correlate with longer flippers. Also shows a relatively straight line trend — good linearity. Delta.15.N vs Delta.13.C: r = 0.57, suggests that nitrogen and carbon isotope values tend to vary together, possibly linked to diet. Moderate linear relationship — not as strong as the others, but the points still loosely form a linear cloud.

Negative Correlations (Negative Linear Relationships): Culmen_Depth vs Culmen_Length: r = -0.57, longer culmen tends to have shallower depth — interesting morphological trade-off. The downward trend in the scatterplot confirms the negative correlation (r = -0.57 ***). The pattern is linear. Culmen_Depth vs Flipper_Length: r = -0.47; Culmen_Depth vs Body_Mass: r = -0.50. These indicate that deeper Culmen(bill) are associated with shorter flippers and lower body mass. Also show negative linearity — as depth increases, body mass and flipper length tend to decrease.

Weak or No Correlation(Non-linear or Weak Relationships): Culmen_Length vs Delta.15.N: r = -0.057 (no asterisk) No statistically significant relationship here. No visible linear pattern — points are scattered randomly

multivariate normality before transformation

```r
# chi-square quantile plot
cqplot(df_PCA[,-c(1,2,3)], main = "Penguin Data")
```

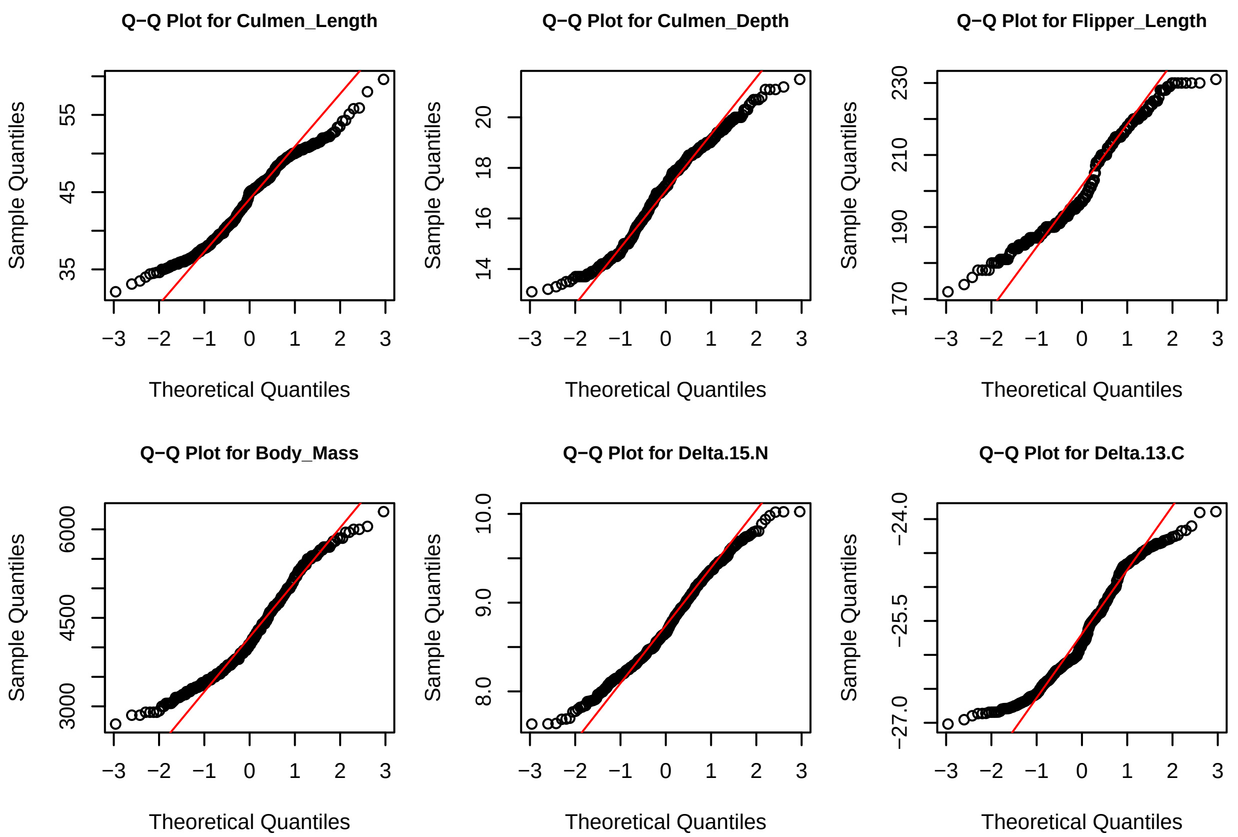


Chi-Square Q-Q Plot of Mahalanobis Distance is used to assess multivariate normality

Majority of Points Are Close to the Line. A few points toward the right deviate above the line (may be potential multivariate outliers). Chi-Square Quantile plots in Figure show that continuous variables in the data set seem to follow a multivariate normal distribution.

We use QQ plot to check for univariate normality before log transformation.

```
# Set up a 3x2 layout (6 plots in total)
par(mfrow = c(2, 3), mar = c(4, 4, 4, 1))  # Adjust margins as needed

# Loop through the first 10 columns of df2
for (i in 4:(min(ncol(df_PCA), 15))) {  # Adjust upper limit for fewer columns if needed
  qqnorm(df_PCA[[i]], main = paste("Q-Q Plot for", colnames(df_PCA)[i]), cex.main = 0.9)
  qqline(df_PCA[[i]], col = "red")
}
```



```
# Reset plotting layout to default after plotting
par(mfrow = c(1, 1))
```

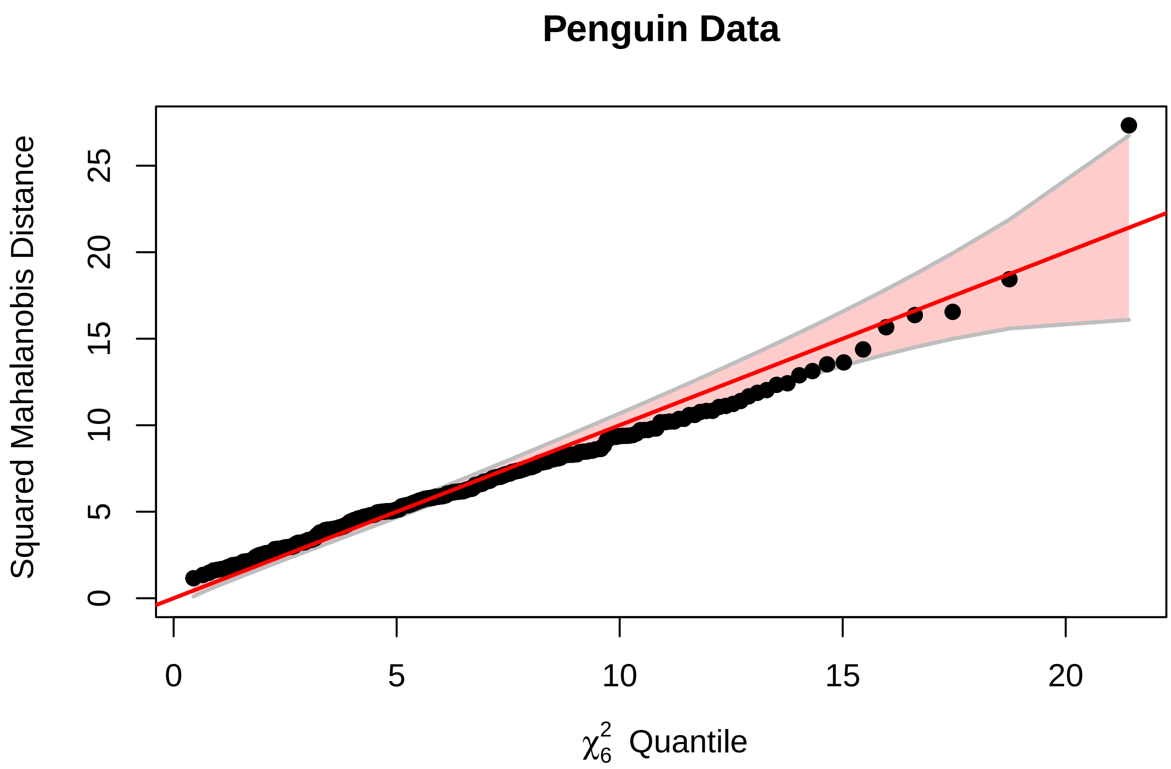
Q-Q plots provide a clear check of univariate normality for each of the six continuous variables. (Q-Q (quantile-quantile) plots compare the sample quantiles of a variable to the theoretical quantiles of a normal distribution.)

None of six variables show strictly normally distributed. This suggests Log-transform would be applied.

multivariate normality after log-transform

```
levels(df_PCA$Species) <- c("Adelie", "Chinstrap", "Gentoo")
df_PCA$Culmen_Length_log <- log(df_PCA$Culmen_Length)
df_PCA$Culmen_Depth_log <- log(df_PCA$Culmen_Depth)
df_PCA$Flipper_Length_log <- log(df_PCA$Flipper_Length)
df_PCA$Body_Mass_log <- log(df_PCA$Body_Mass)
df_PCA$Delta.15.N_log <- log(df_PCA$Delta.15.N)

# ensures all values are positive while preserving the relative structure

# Find the minimum value
min_val_Delta.13 <- min(df_PCA$Delta.13.C, na.rm = TRUE)

# Shift so the minimum becomes slightly above zero (e.g., 1)
shift_constant <- abs(min_val_Delta.13) + 1  # e.g., if min is -27.02 → shift by 28.02

# Step 3: Apply log-transform
df_PCA$Delta.13.C_log <- log(df_PCA$Delta.13.C + shift_constant)
# df_PCA$Delta.13.Cs_log <- log(df_PCA$Delta.13.C) # negative value
```

```
cqplot(df_PCA[,c("Culmen_Length_log", "Culmen_Depth_log", "Flipper_Length_log", "Body_Mass_log","Delta.
       main = "Penguin Data")
```



After Log Transformation, majority of points lie closer to the red line, especially in the mid-range quantiles. Fewer points deviate strongly in the upper-right tail compared to the previous plot. The outlier at the far upper-right (likely a multivariate outlier) is still present, but its distance is reduced compared to before. Log transformation improved multivariate normality.

principal components

Perform Principal components analysis using the Correlation matrix (standardized variables). Think about how many principal components to retain. To make this decision look at: • Total variance explained by a given number of principle components. 80% as threshold • The 'eigenvalue > 1' criteria • The 'scree plot elbow' method (turn in the scree plot) • Parallel Analysis: think about whether this is appropriate based on what you discover in question 1.

```
head(df_PCA)
```

```
##   Individual.ID Species    Island Culmen_Length Culmen_Depth Flipper_Length
## 1          N1A2  Adelie Torgersen          39.5         17.4            186
## 2          N2A1  Adelie Torgersen          40.3         18.0            195
## 3          N3A1  Adelie Torgersen          36.7         19.3            193
## 4          N3A2  Adelie Torgersen          39.3         20.6            190
## 5          N4A1  Adelie Torgersen          38.9         17.8            181
## 6          N4A2  Adelie Torgersen          39.2         19.6            195
##   Body_Mass Delta.15.N Delta.13.C Culmen_Length_log Culmen_Depth_log
## 1      3800    8.94956  -24.69454          3.676301         2.856470
## 2      3250    8.36821  -25.33302          3.696351         2.890372
## 3      3450    8.76651  -25.32426          3.602777         2.960105
## 4      3650    8.66496  -25.29805          3.671225         3.025291
## 5      3625    9.18718  -25.21799          3.660994         2.879198
## 6      4675    9.46060  -24.89958          3.668677         2.975530
##   Flipper_Length_log Body_Mass_log Delta.15.N_log Delta.13.C_log
## 1           5.225747      8.242756       2.191604      1.2011689
## 2           5.273000      8.086410       2.124440      0.9878744
## 3           5.262690      8.146130       2.170939      0.9911310
## 4           5.247024      8.202482       2.159287      1.0008120
## 5           5.198497      8.195610       2.217809      1.0298158
## 6           5.273000      8.449984       2.247136      1.1374996
```

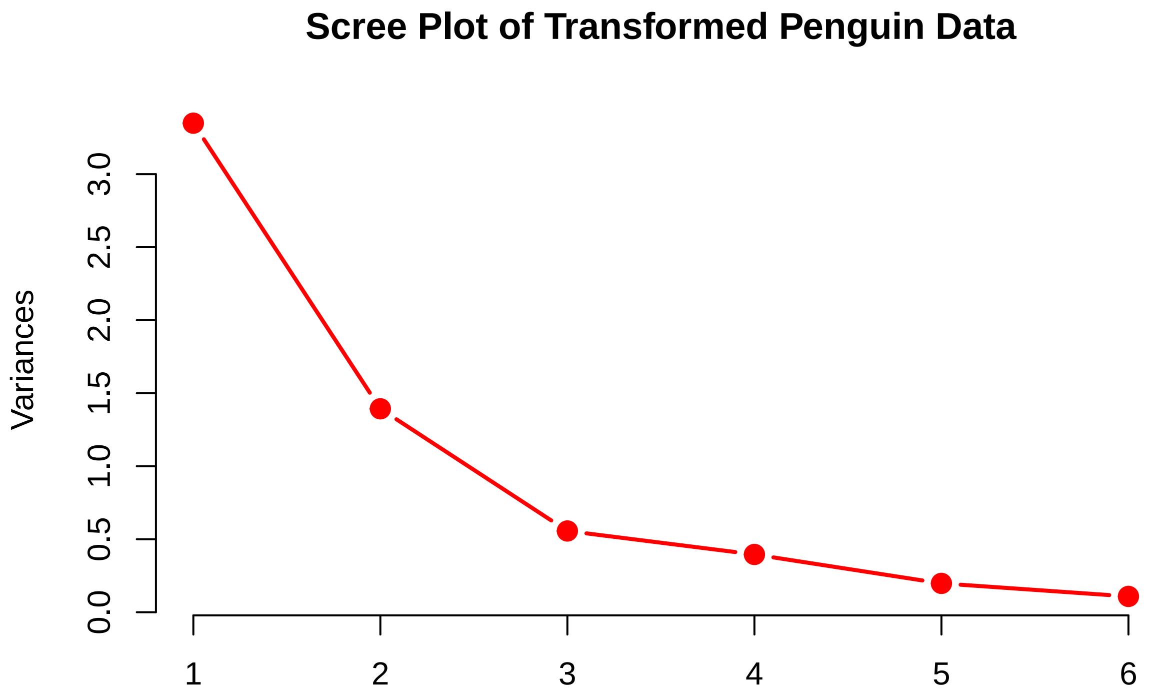
```
#scale. = TRUE means run on the correlation matrix, i.e. standardize the variables.
pc1 <- prcomp(df_PCA[, 10:15], scale. = TRUE)
summary(pc1)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     1.8301 1.1804 0.74590 0.62891 0.44417 0.32895
## Proportion of Variance 0.5582 0.2322 0.09273 0.06592 0.03288 0.01803
## Cumulative Proportion  0.5582 0.7904 0.88316 0.94908 0.98197 1.00000
```

```
summary.PCA.JDRS <- function(x){
  sum_JDRS <- summary(x)$importance
  sum_JDRS[1, ] <- sum_JDRS[1, ]^2
  attr(sum_JDRS, "dimnames")[[1]][1] <- "Eigenvals (Variance)"
  sum_JDRS
}
#print results -
#Here are eigenvalues
round(summary.PCA.JDRS(pc1),2)
```

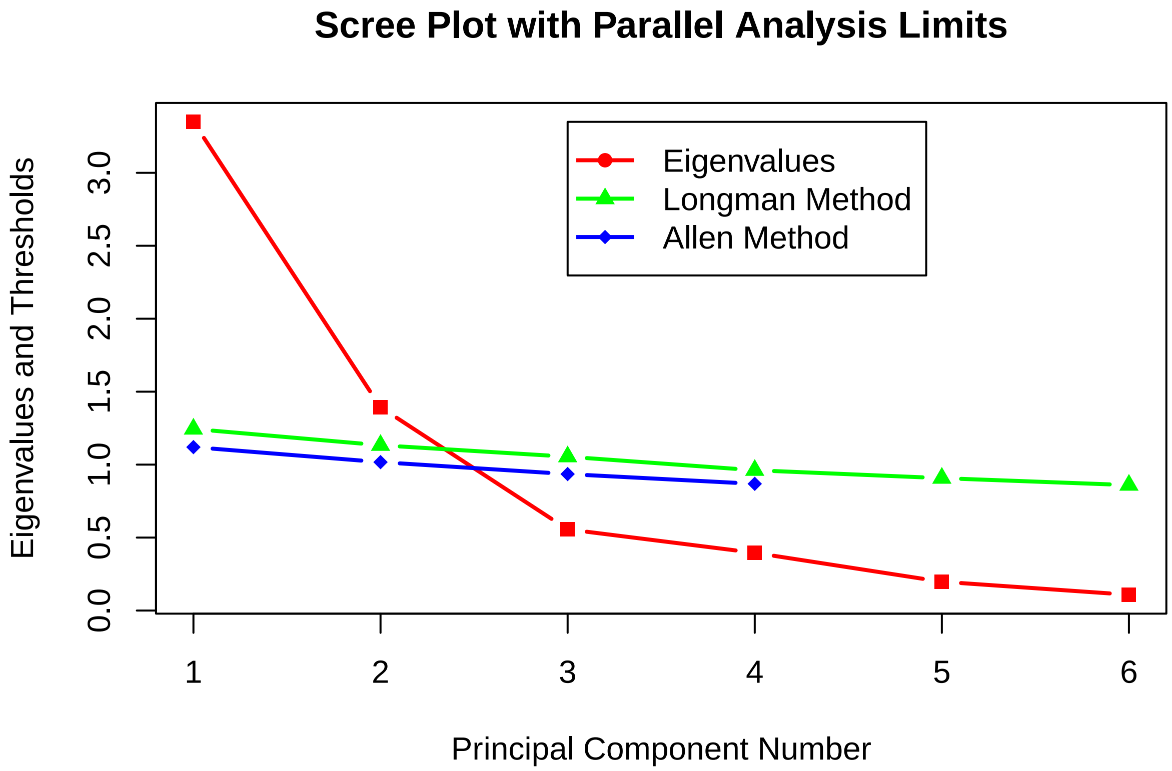
```
##                           PC1  PC2  PC3  PC4  PC5  PC6
## Eigenvals (Variance)    3.35 1.39 0.56 0.40 0.20 0.11
## Proportion of Variance 0.56 0.23 0.09 0.07 0.03 0.02
## Cumulative Proportion  0.56 0.79 0.88 0.95 0.98 1.00
```

```
screeplot(pc1, type = "lines", col = "red", lwd = 2, pch = 19, cex = 1.2,
          main = "Scree Plot of Transformed Penguin Data")
```



```
#get function from online
source("https://raw.githubusercontent.com/jreuning/sds363_code/refs/heads/main/parallel.r.txt")
parallelplot(pc1)
```

```
##   pcompnum   longman     allen
## 1        1 1.2442844 1.1200013
## 2        2 1.1328433 1.0170629
## 3        3 1.0543484 0.9353226
## 4        4 0.9615026 0.8684552
## 5        5 0.9074322        NA
## 6        6 0.8597180        NA
```



We choose 4 methods to decide how many principal components to retain • Total variance explained by a given number of principle components. 80% as threshold. The cumulative proportion of variance reaches 88.31% at PC3, cumulative proportion of variance at PC2 reaches 79% would also be good. • The 'eigenvalue > 1' criteria. PCs 1–2 are clearly above 1 • The 'scree plot elbow' method (turn in the scree plot). Scree plot has elbow at two and three which would argue for retaining 1 or 2 component. • Parallel Analysis: based on the chi-square quantile plot of our transformed data, our data shows multivariate normality, we can use can perform parallel analysis. The plot shows PC2 is slightly above the green and blue thresholds. PC3 to PC6 are all below both threshold lines. It supports 2 principal components to be chosen. After consideration in order of least desirable to most desirable. We decided to keep two components

```
#Get loadings
round(pc1$rotation[, 1:2], 2)
```

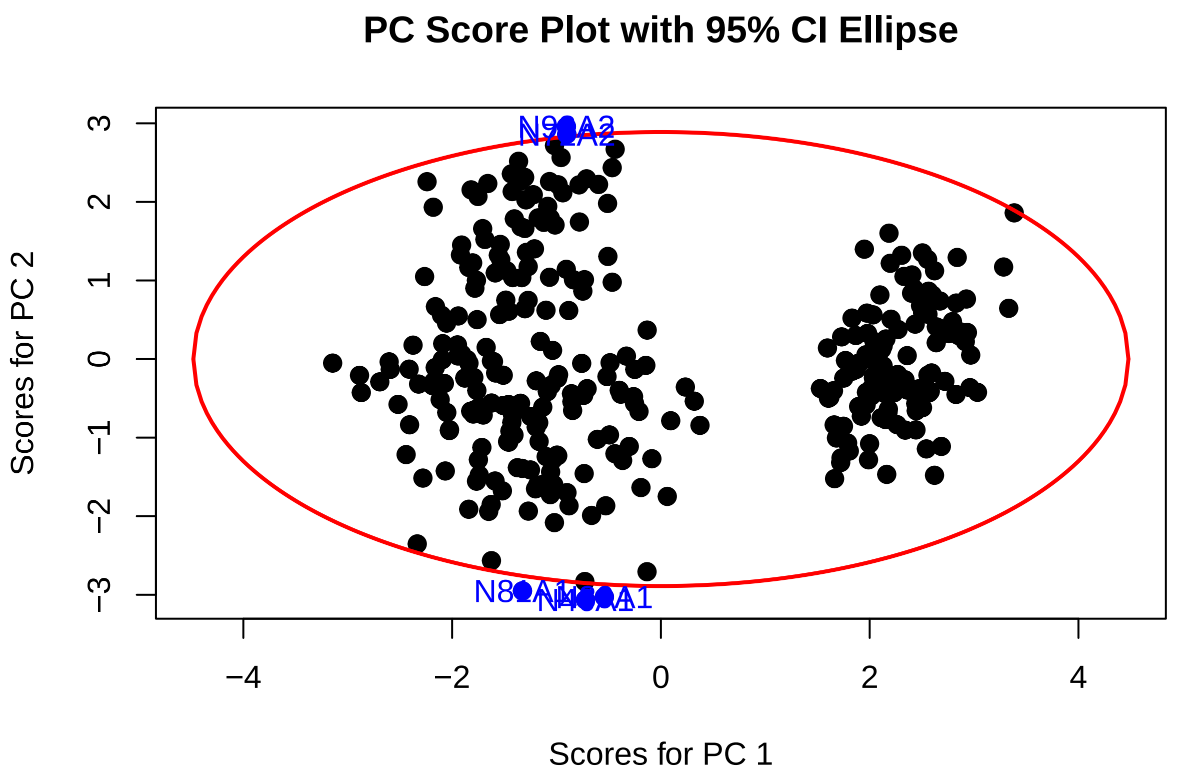
```
##                      PC1  PC2
## Culmen_Length_log   0.30 0.66
## Culmen_Depth_log   -0.41 0.21
## Flipper_Length_log  0.50 0.22
## Body_Mass_log       0.48 0.20
## Delta.15.N_log     -0.40 0.38
## Delta.13.C_log     -0.31 0.54
```

Interpretation of Each Component: PC1 would be interpreted as "Size vs Trophic Structure" Dimension. High positive loadings: Flipper_Length_log (0.50), Body_Mass_log (0.48), Culmen_Length_log (0.30). High negative loadings: Culmen_Depth_log (-0.41), Delta.15.N_log (-0.40) (used to measure food chain length and the trophic level of a given organism), Delta.13.C_log (widely used for the reconstruction of past diets, particularly to see if marine foods or certain types of plants were consumed.) (-0.31). Interpretation:

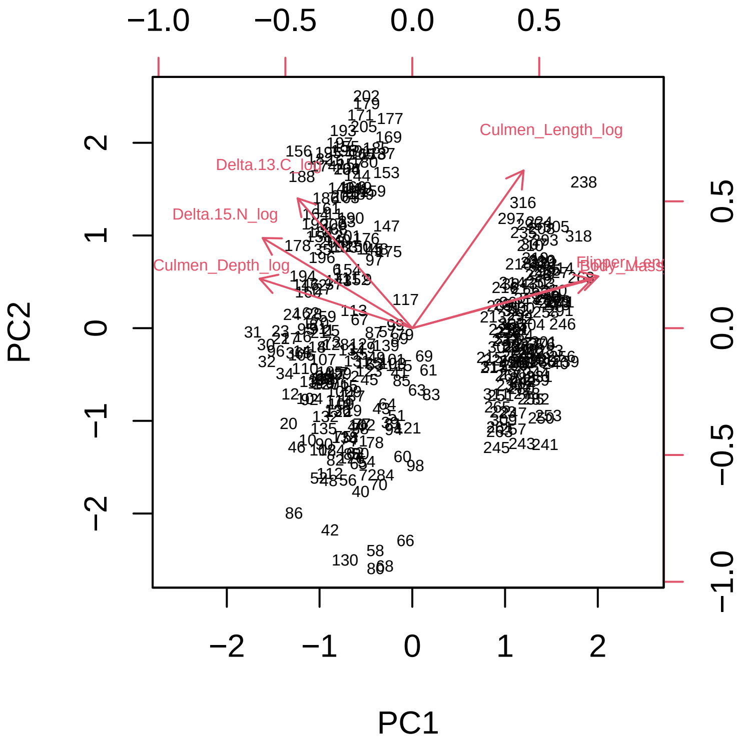
PC1 separates penguins based on physical size (larger flippers and body mass) vs. bill depth and stable isotope ratios. Penguins with high PC1 scores are physically larger, but lower in trophic level indicators (isotopic values), and shallower bills. Penguins with low PC1 scores may have deeper bills, higher nitrogen and carbon isotope values, and smaller body size.

PC2 would be interpreted as "Morpho-Isotopic Axis". High loadings: Culmen_Length_log (0.66), Delta.13.C_log (0.54), Delta.15.N_log (0.38), and low loadings for others. Interpretation: PC2 captures variation in bill length and diet (isotope signatures). High PC2 scores indicate penguins with longer bills and higher delta13C/delta15N values — possibly suggesting different foraging behavior or habitat. This component seems to blend morphological and ecological (dietary) information.

```r
source("https://raw.githubusercontent.com/jreuning/sds363_code/refs/heads/main/ciscoreplot.r.txt")
# use components 1 and 2, and ID to represent the points
ciscoreplot(pc1, c(1, 2), df_PCA[, 1])
```



```r
#ciscoreplot(pc1, c(1, 2), df_PCA[, 2])
#ciscoreplot(pc1, c(1, 2), df_PCA[, 3])
# make a biplot for first two components
biplot(pc1, choices = c(1, 2), pc.biplot = T, cex = 0.5)
```



```
# exact outlier points
x = pc1
comps = c(1, 2)
namevec = df_PCA[, 1]
y1<-sqrt(5.99*(x$sdev[comps[1]]^2))
ymod<-y1-y1%%.05
y1vec<-c(-y1,seq(-ymod,ymod,by=0.05),y1)
y2vecpos<-sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)*x$sdev[comps[2]]^2)
y2vecneg<--sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)*x$sdev[comps[2]]^2)
y2vecpos[1]<-0
y2vecneg[1]<-0
y2vecpos[length(y2vecpos)]<-0
y2vecneg[length(y2vecneg)]<-0
outliers<-((x$x[,comps[1]]^2)/(x$sdev[comps[1]]^2)+(x$x[,comps[2]]^2)/(x$sdev[comps[2]]^2))>5.99
namevec[outliers]
```

```
## [1] "N39A1" "N46A1" "N81A1" "N72A2" "N98A2"
```

By plots above, the points appear strong or obvious cluster structure which can be clearly separated by PC1 axis (PC1 appears to capture the biggest separation. It might be related to overall body size or morphology, as previously inferred from the loadings.). It suggests components one and two capture distinct size and trophic profiles within the dataset.

Using a 95% confidence ellipse in the PC1 vs. PC2 space can be a quick visual check for potential outliers (and our data basically follows roughly multivariate normal distribution). By PC Score Plot with 95%

CI Ellipse, a few labeled points with their Individual.ID (outlier points: “N39A1”, “N46A1”, “N81A1”, “N72A2”, “N98A2”) are shown as possible outliers or representative samples.

By PCA Biplot, it shows 1. Flipper_Length_log and Body_Mass_log are the main vectors in the positive PC1 direction. This suggests that individuals with higher flipper length and body mass will tend to have higher PC1 scores.

2. Culmen_Length_log also points positively along PC1 (though slightly toward PC2 as well). Individuals with longer culmen lengths also tend to have higher PC1 scores, and higher PC2 value.
3. Culmen_Depth_log, Delta.15.N_log, and Delta.13.C_log arrows point toward the upper left quadrant — i.e., more negatively on PC1 and positively on PC2. Individuals scoring high on these variables tend to have lower PC1 and higher PC2 scores.

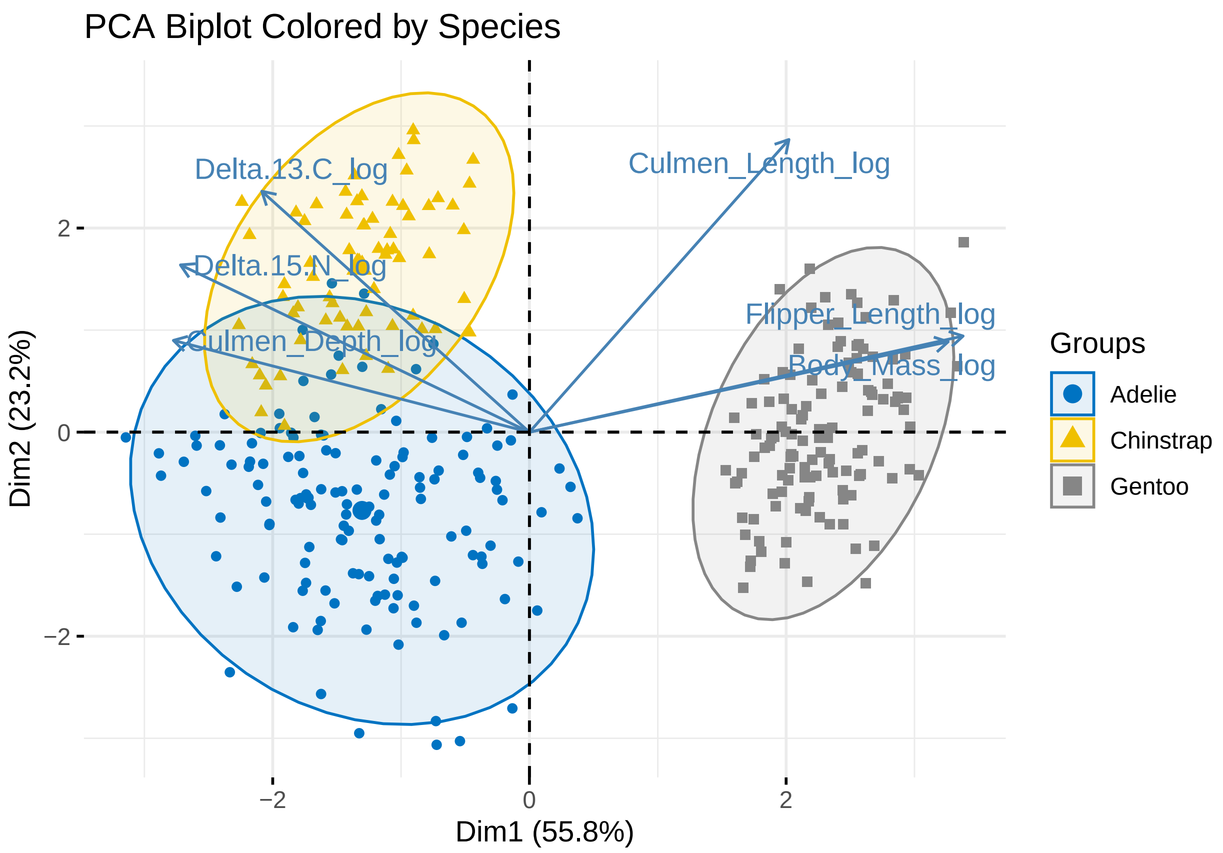
Variables pointing in the same direction (e.g., Flipper_Length_log and Body_Mass_log) suggest a strong positive correlation in this PC1–PC2 space. Variables pointing in opposite directions (e.g., Flipper_Length_log vs. Culmen_Depth_log) suggest a negative correlation. This aligns with earlier correlation matrix: deeper bills tend to occur with shorter flippers and lighter body mass.

```r
#install.packages("factoextra")  # if not installed
library(factoextra)
```

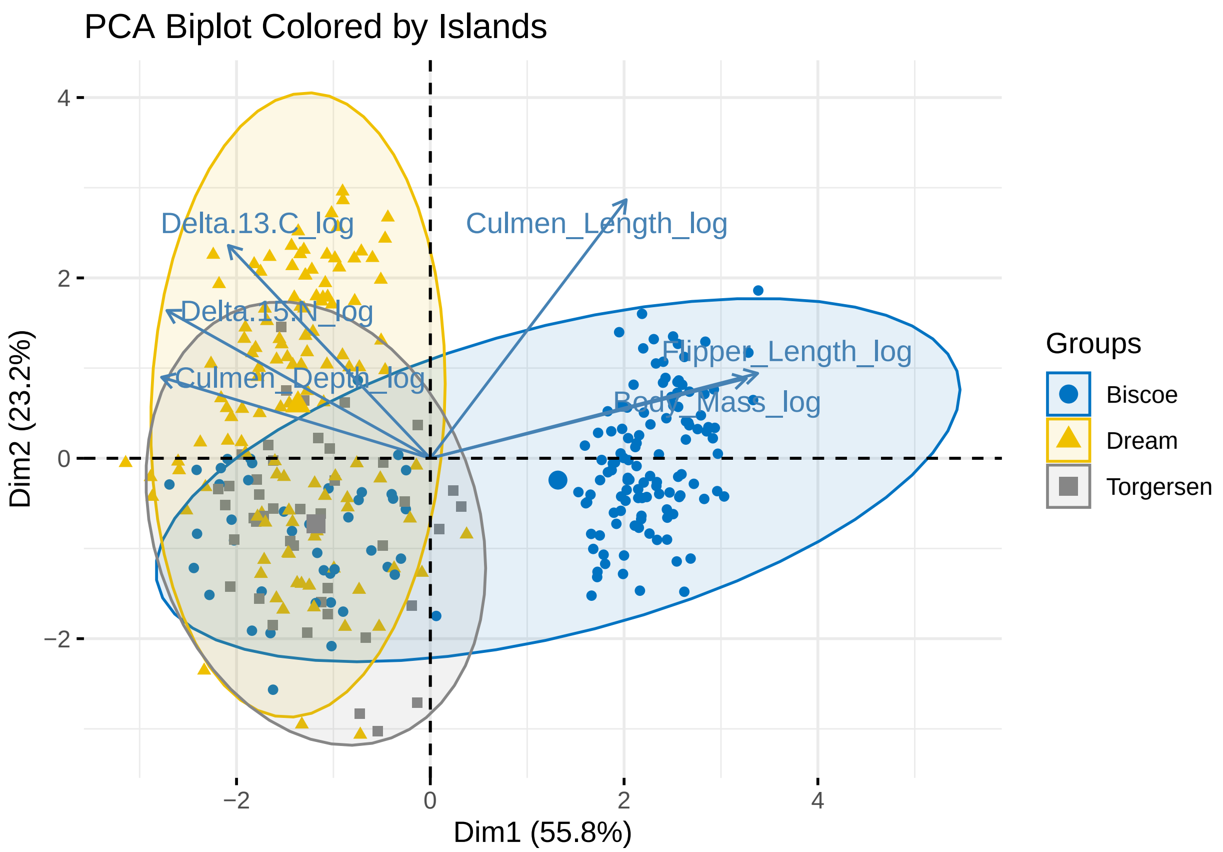
```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
# PCA with factoextra biplot
fviz_pca_biplot(pc1,
                label = "var",      # show variable arrows
                habillage = df_PCA$Species,  # color by species
                addEllipses = TRUE,
                ellipse.level = 0.95,
                palette = "jco",
                repel = TRUE,
                title = "PCA Biplot Colored by Species")
```



```r
# PCA with factoextra biplot
fviz_pca_biplot(pc1,
                label = "var",      # show variable arrows
                habillage = df_PCA$Island,  # color by species
                addEllipses = TRUE,
                ellipse.level = 0.95,
                palette = "jco",
                repel = TRUE,
                title = "PCA Biplot Colored by Islands")
```



PCA Biplot Colored by Species, The three species of penguins, Adelie (blue), Chinstrap (yellow), and Gentoo (gray) - form clearly separated clusters in the PCA space. The first principal component (PC1) explains 55.8% of the variance and clearly separates: Gentoo penguins with High PC1 scores (far right), Adelie penguins and Chinstrap with Low PC1 scores (far left) and the second principal component (PC2) would be considered used to make separation between Adelie penguins and Chinstrap. Flipper_Length_log and Body_Mass_log point strongly in the PC1 direction, showing Gentoo penguins are larger birds. Delta.15.N_log, Delta.13.C_log, and Culmen_Depth_log point to the upper left, suggesting Chinstrap penguins, indicating higher trophic levels, longer food chains (also from more fishes eat C4 plants) and width bill depth. And for Adelie, Culmen_Length_log point strongly to upper right(high in both PC1 and PC2) suggests they have shorter bill.

Species differences are strongly associated with morphology and diet, and PCA effectively separates species based on these traits. Providing a promising result for coming Discriminant Analysis.

By PCA Biplot Colored by Island, it doesn't show clear pattern that PC1 and PC2 would clearly separate the islands penguins lives. This may indicates islands have sharing features for penguins to survive. But we do see that Biscoe island does provide penguins lower chance for higher trophic levels, longer food chains. Also, by comparing two colored biplots, we notices Adelie penguins existing in all three islands, while Chinstrap only stay in Dream Island and Gentoo only stay in Biscoe islands.

summary

PCA effectively revealed structure in the penguin dataset, with two principal components capturing 79% of the total variance. PC1 represented a "Size vs. Trophic Structure" axis, separating larger-bodied penguins from those with deeper bills and higher isotope values. PC2 reflected variation in bill length and diet. Species were clearly separated in the PCA space, especially along PC1, while island-based grouping showed more overlap. Overall, PCA successfully reduced dimensionality and highlighted key ecological and morphological differences among penguin species.

Discriminant Analysis

Goal: Discrimination - identify variables that 'best' discriminate between three known species and use data on variables from known species to develop a rule for classifying future observations.

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(biotools)
```

```
## ---
## biotools version 4.2

##
## Attaching package: 'biotools'

## The following object is masked from 'package:heplots':
##
##     boxM
```

```r
library(klaR)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

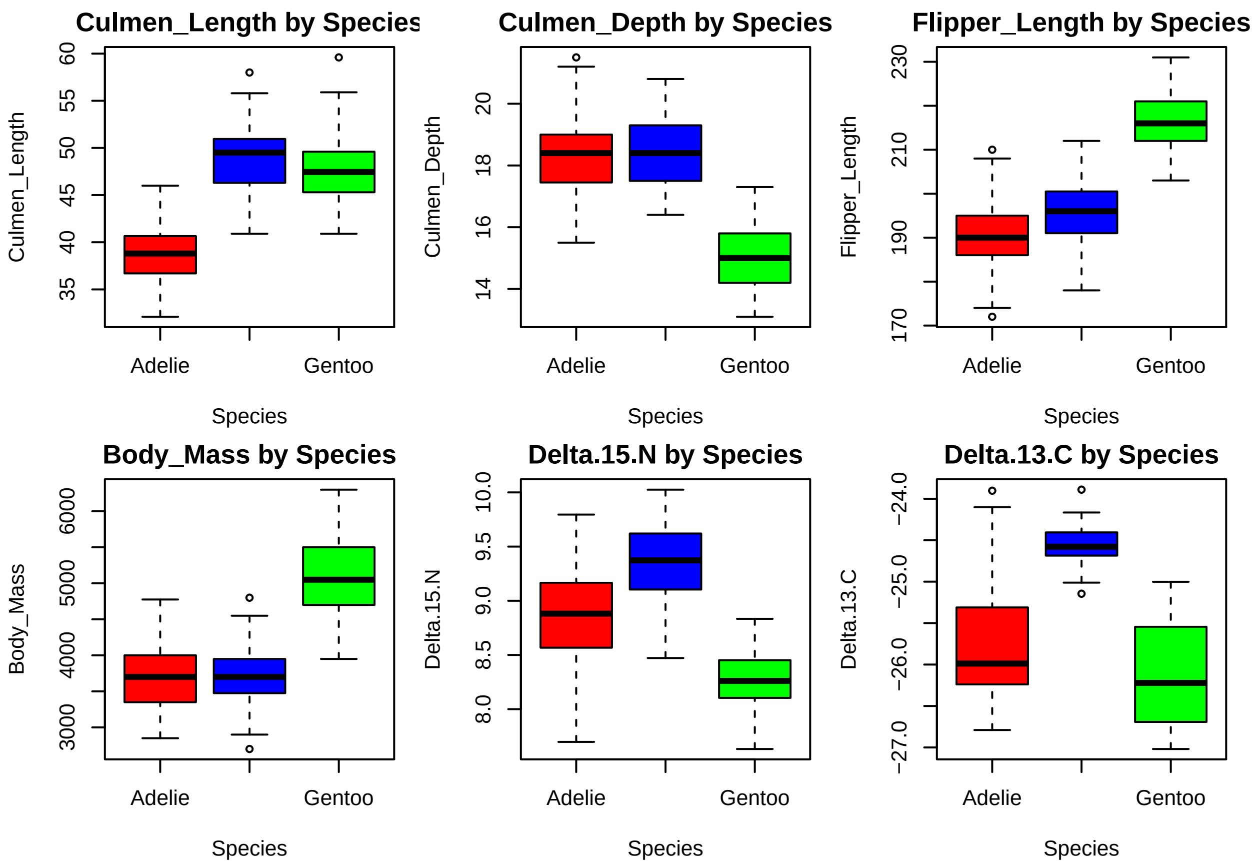
```r
library(ggplot2)
library(ggExtra)
library(heplots)
library(corrplot)
library(PerformanceAnalytics)
```

```r
head(data)
```

```
##   studyName Sample.Number                             Species Region    Island
## 1   PAL0708             2 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 2   PAL0708             3 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 3   PAL0708             5 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 4   PAL0708             6 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 5   PAL0708             7 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 6   PAL0708             8 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
##                Stage Individual.ID Clutch.Completion  Date.Egg Culmen_Length
## 1 Adult, 1 Egg Stage          N1A2               Yes 2011/11/7          39.5
## 2 Adult, 1 Egg Stage          N2A1               Yes  11/16/07          40.3
## 3 Adult, 1 Egg Stage          N3A1               Yes  11/16/07          36.7
## 4 Adult, 1 Egg Stage          N3A2               Yes  11/16/07          39.3
## 5 Adult, 1 Egg Stage          N4A1                No  11/15/07          38.9
## 6 Adult, 1 Egg Stage          N4A2                No  11/15/07          39.2
##   Culmen_Depth Flipper_Length Body_Mass    Sex Delta.15.N Delta.13.C
## 1         17.4            186      3800 FEMALE    8.94956  -24.69454
## 2         18.0            195      3250 FEMALE    8.36821  -25.33302
## 3         19.3            193      3450 FEMALE    8.76651  -25.32426
## 4         20.6            190      3650   MALE    8.66496  -25.29805
## 5         17.8            181      3625 FEMALE    9.18718  -25.21799
## 6         19.6            195      4675   MALE    9.46060  -24.89958
```

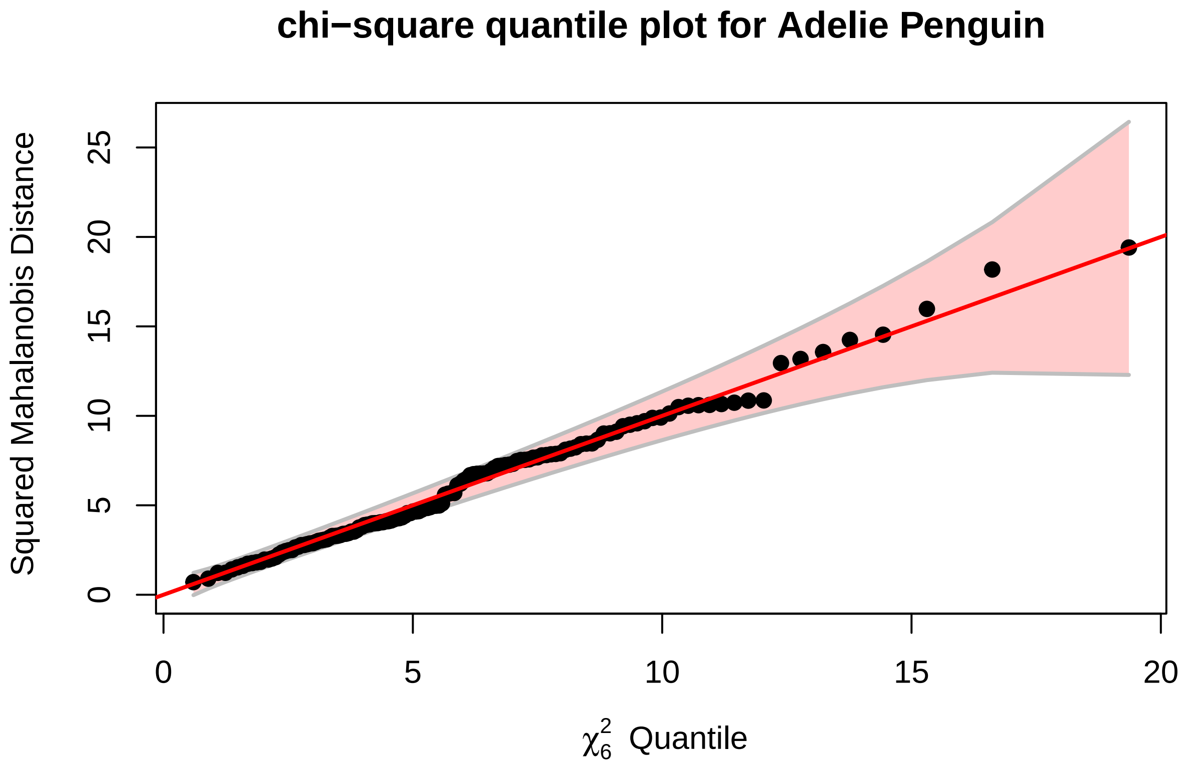
Discriminant Analysis Assumptions: The observations within each group represent a sample from a multivariate normal distribution. The covariance matrices of each group are assumed to be identical.

```r
# Boxplot
# Set up the plotting area with multiple panels (2 rows, 3 columns)
par(mfrow = c(2, 3), mar = c(4, 4, 2, 1))
levels(data$Species) <- c("Adelie", "Chinstrap", "Gentoo")

# Boxplot for Culmen_Length grouped by Species
boxplot(Culmen_Length ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Culmen_Length by Species", ylab = "Culmen_Length")

# Boxplot for Culmen_Depth grouped by Species
boxplot(Culmen_Depth ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Culmen_Depth by Species", ylab = "Culmen_Depth")

# Boxplot for Flipper_Length grouped by Species
boxplot(Flipper_Length ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Flipper_Length by Species", ylab = "Flipper_Length")

# Boxplot for Body_Mass grouped by Species
boxplot(Body_Mass ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Body_Mass by Species", ylab = "Body_Mass")

# Boxplot for Delta.15.N grouped by Species
boxplot(Delta.15.N ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Delta.15.N by Species", ylab = "Delta.15.N")

# Boxplot for Delta.13.C grouped by Species
boxplot(Delta.13.C ~ Species, data = data, col = c('red', 'blue', 'green'),
        main = "Delta.13.C by Species", ylab = "Delta.13.C")
```

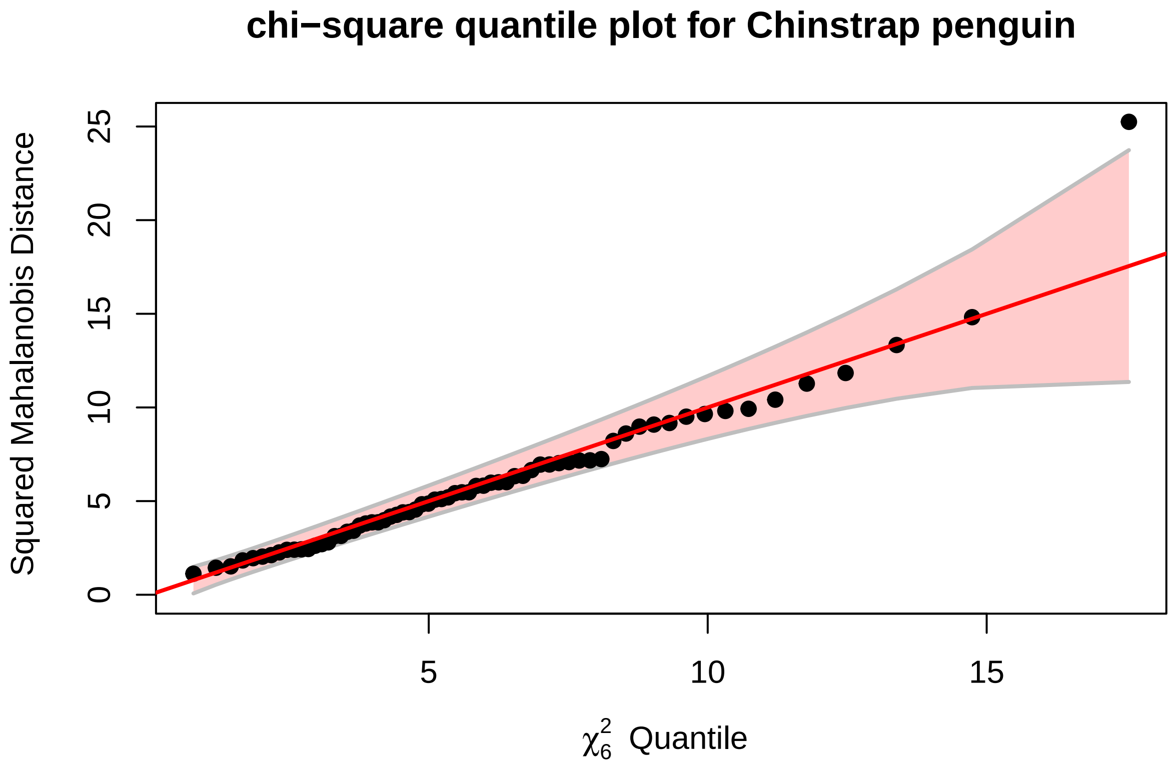


#### chi-square quantile plot

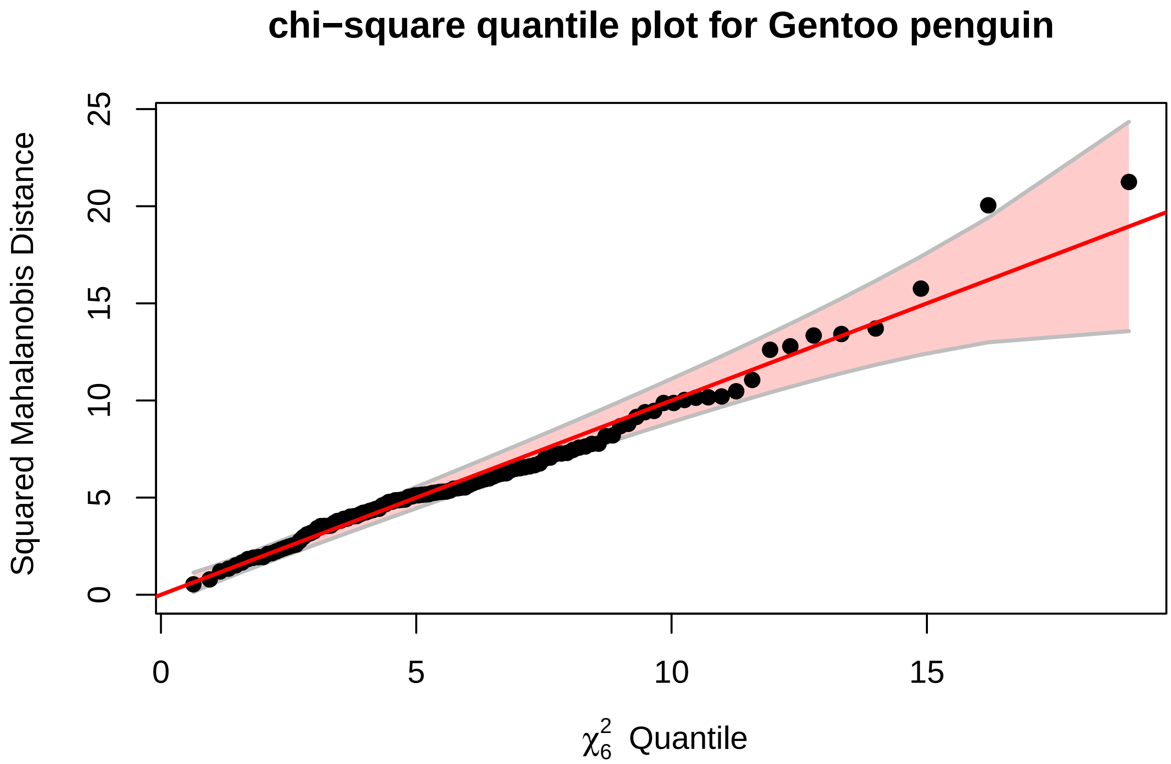
```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Species  == 'Adelie', c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Mass"
```



```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Species  == 'Chinstrap', c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Ma
```



```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Species  == 'Gentoo', c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Mass"
```



According to chi-square quantile plots of each species, most points are within the 95% range area of the plots. We can conclude that The observations within each group represent a sample from a multivariate normal distribution.

```
group1_cov <- cov(data[data$Species == 'Adelie', c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "
group2_cov <- cov(data[data$Species == 'Chinstrap', c("Culmen_Length", "Culmen_Depth", "Flipper_Length"
group3_cov <- cov(data[data$Species == 'Gentoo', c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "
```

```
cov_rat <- group2_cov/group1_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

```
##                Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N
## Culmen_Length            1.5          1.9            1.7       1.1        2.1
## Culmen_Depth             1.9          1.2            1.8       1.3        1.5
## Flipper_Length           1.7          1.8            1.2       1.2       -4.2
## Body_Mass                1.1          1.3            1.2       1.4        1.3
## Delta.15.N               2.1          1.5           -4.2       1.3        1.3
## Delta.13.C               1.7          3.9           -5.3      -4.5      -10.7
##                Delta.13.C
## Culmen_Length         1.7
## Culmen_Depth          3.9
## Flipper_Length       -5.3
## Body_Mass            -4.5
## Delta.15.N          -10.7
## Delta.13.C            7.0
```

```
cov_rat <- group2_cov/group3_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

```
##                Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N
## Culmen_Length            1.1          1.2            1.3       1.6        1.9
## Culmen_Depth             1.2          1.3            1.0       1.4        1.1
## Flipper_Length           1.3          1.0            1.1       1.3        1.4
## Body_Mass                1.6          1.4            1.3       1.7       -4.2
## Delta.15.N               1.9          1.1            1.4      -4.2        1.9
## Delta.13.C               2.1         -1.6           -1.8      17.0        7.1
##                Delta.13.C
## Culmen_Length         2.1
## Culmen_Depth         -1.6
## Flipper_Length       -1.8
## Body_Mass            17.0
## Delta.15.N            7.1
## Delta.13.C            6.0
```

```
cov_rat <- group1_cov/group3_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

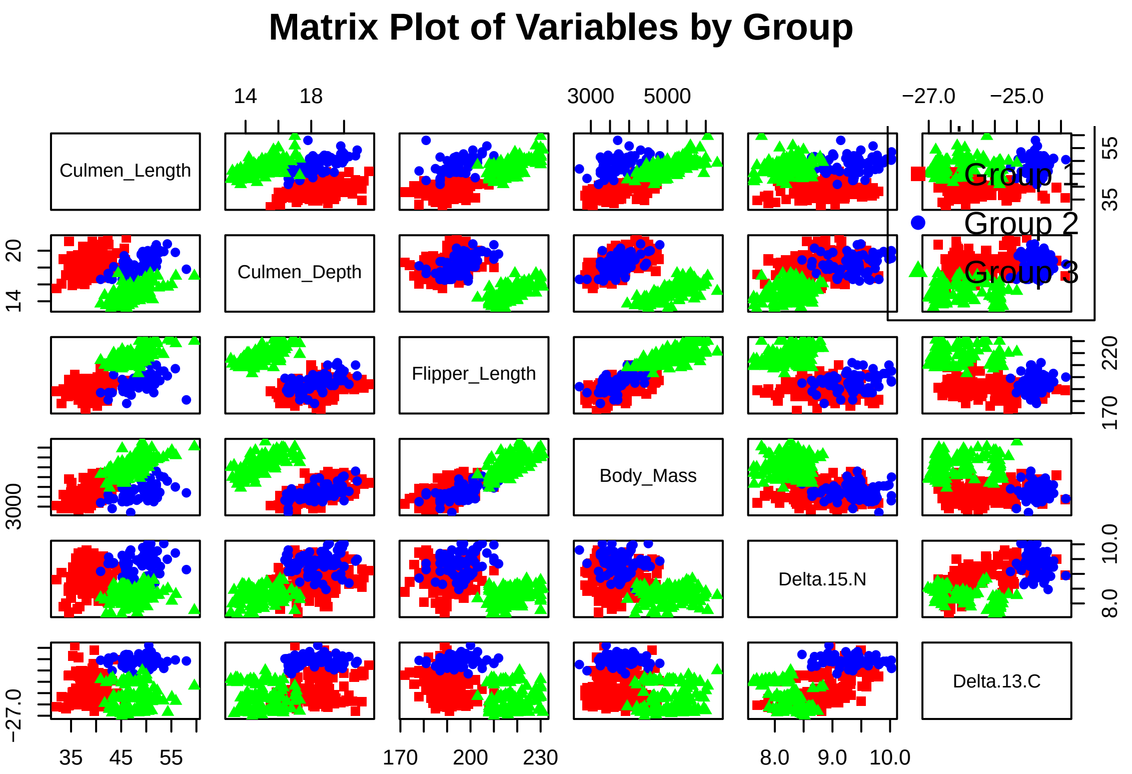
```
##                Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N
## Culmen_Length            1.3          1.5            2.2       1.5        1.1
## Culmen_Depth             1.5          1.5            1.8       1.1        1.6
## Flipper_Length           2.2          1.8            1.0       1.6       -3.0
## Body_Mass                1.5          1.1            1.6       1.2       -5.6
## Delta.15.N               1.1          1.6           -3.0      -5.6        2.6
## Delta.13.C               1.2         -2.4            3.0      -3.8       -1.5
##                Delta.13.C
## Culmen_Length         1.2
## Culmen_Depth         -2.4
## Flipper_Length        3.0
## Body_Mass            -3.8
## Delta.15.N           -1.5
## Delta.13.C            1.2
```

```
# Box's M statistic
boxM(data[,c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Mass", "Delta.15.N", "Delta.13.C"
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data[, c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Mass",     "Delta.15.N", "De
## Chi-Sq (approx.) = 272.44, df = 42, p-value < 2.2e-16
```

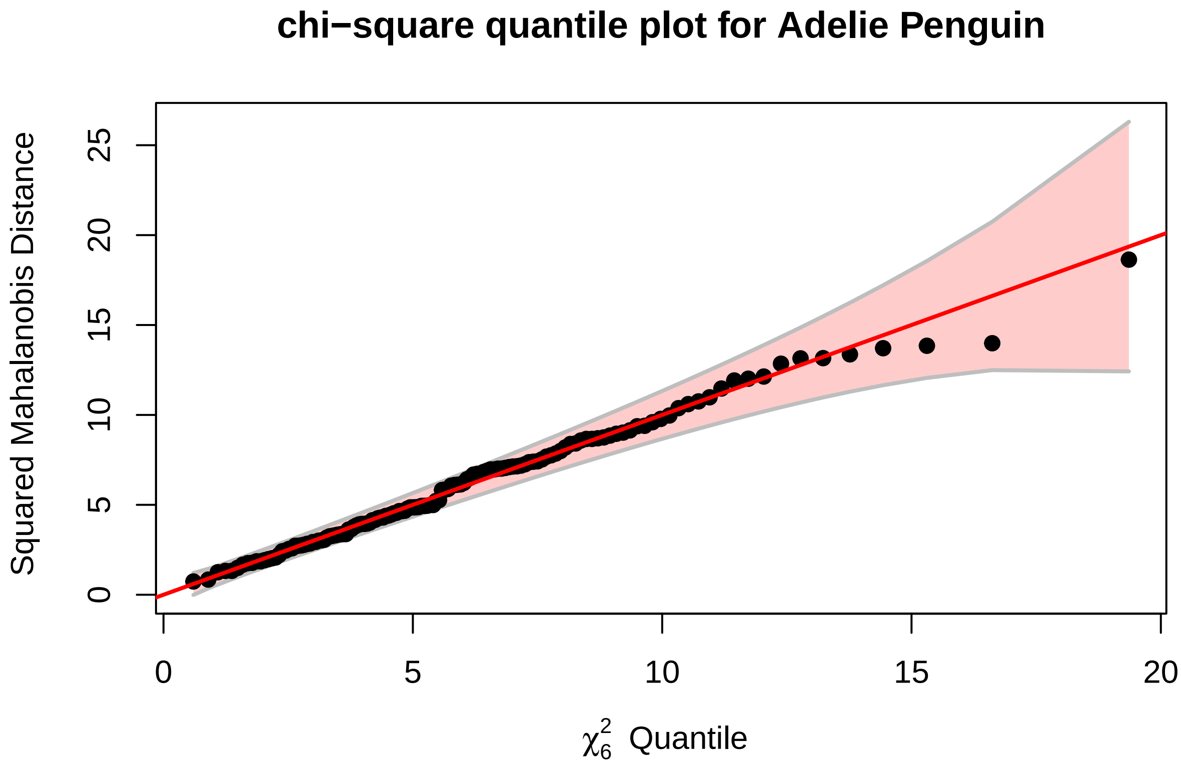
Ratios of largest to smallest elements of the covariance matrices are not smaller than 4 (Body_Mass and Delta.13.C when comparing group 2 and group 3) and based on the Box's M-test, the p-value < 2.2e-16 which is smaller than 0.01 or 0.05, so we reject the null and conclude that there is significant difference between the each pair of groups' covariance matrices. This indicates that we should consider quadratic discrimination functions.

```r
# Assign colors and symbols to each group
group_colors <- c('red', 'blue', 'green')[data$Species]
group_symbols <- as.numeric(data$Species) + 14  # Use symbols 15, 16, 17 for each group

# Matrix plot
pairs(data[,c("Culmen_Length", "Culmen_Depth", "Flipper_Length", "Body_Mass", "Delta.15.N", "Delta.13.C
      col = group_colors, pch = group_symbols,
      main = "Matrix Plot of Variables by Group")

# Add a legend
legend("topright", legend = c("Group 1", "Group 2", "Group 3"),
       col = c('red', 'blue', 'green'), pch = 15:17, title = "Species")
```

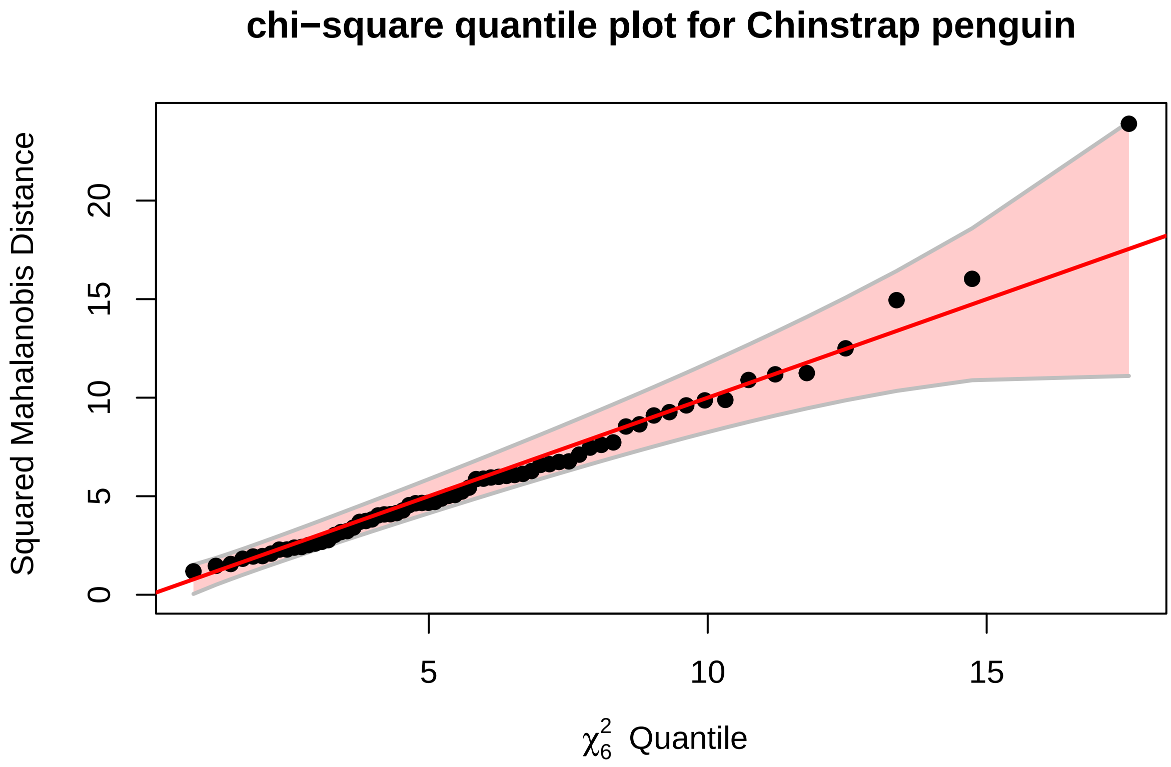


pairs plot
test whether some log-transformation help with make covariance matrices simila

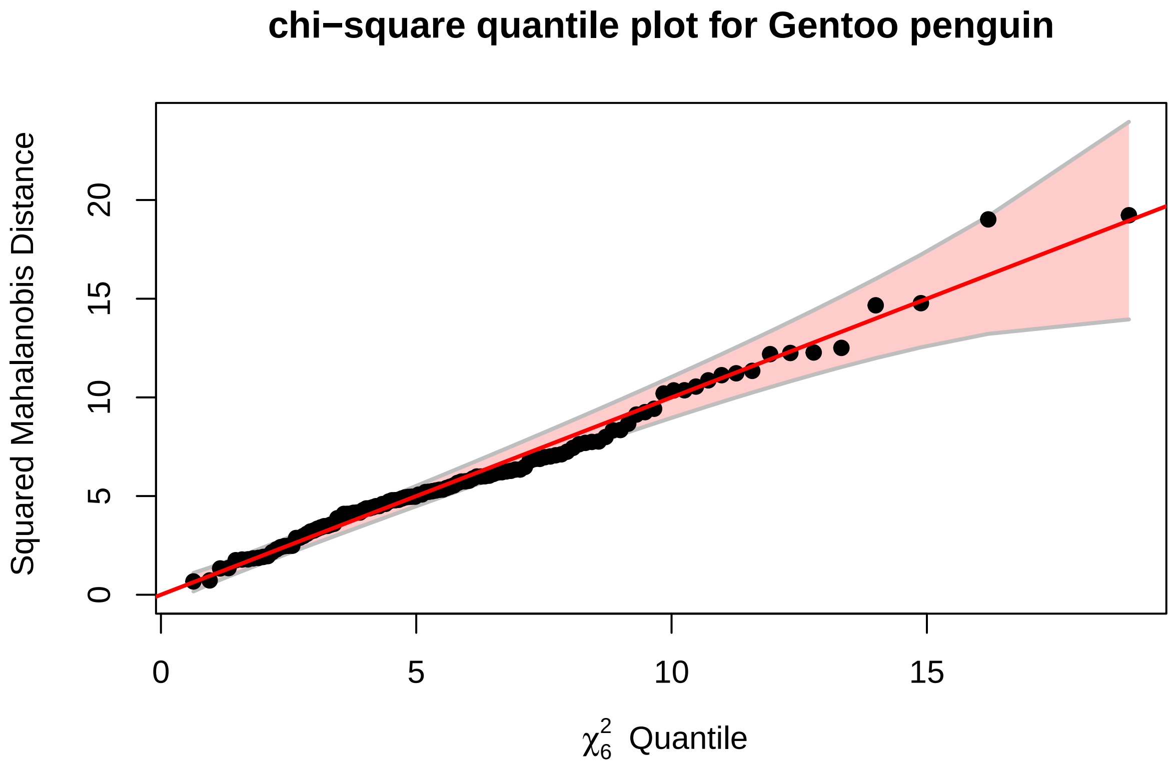
```r
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(df_PCA[df_PCA$Species  == 'Adelie', 10:15], main = "chi-square quantile plot for Adelie Penguin"
```



```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(df_PCA[df_PCA$Species  == 'Chinstrap', 10:15], main = "chi-square quantile plot for Chinstrap pe
```



```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(df_PCA[df_PCA$Species  == 'Gentoo', 10:15], main = "chi-square quantile plot for Gentoo penguin")
```



```r
# Box's M statistic
boxM(df_PCA[,10:15], df_PCA$Species)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  df_PCA[, 10:15]
## Chi-Sq (approx.) = 336.68, df = 42, p-value < 2.2e-16
```

By above plots and Box's M-test, it seems like the transformation won't make covariances matrices to be seem as similar and even make normality worse.

Inclusion, multivariate normality within each group hold and covariance matrices are different between groups, quadratic discrimination functions would be a better choice. Yet since there is not need to force all variables used for discrimination, we can try step-wise LDA first for result.

```r
# Convert tibble to data frame
df_PCA <- as.data.frame(df_PCA)

# Stepwise LDA
# Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N Delta.13.C
step_lda <- stepclass(Species ~ Culmen_Length + Culmen_Depth + Flipper_Length + Body_Mass + Delta.15.N
                      data = df_PCA,
                      method = "lda",
                      direction = "both",
                      fold = nrow(df_PCA)) # LOOCV
```

First, perform step-wise LDA

```
##  ‘stepwise classification’, using 324-fold cross-validated correctness rate of method lda’.
```

```
## 324 observations of 6 variables in 3 classes; direction: both
```

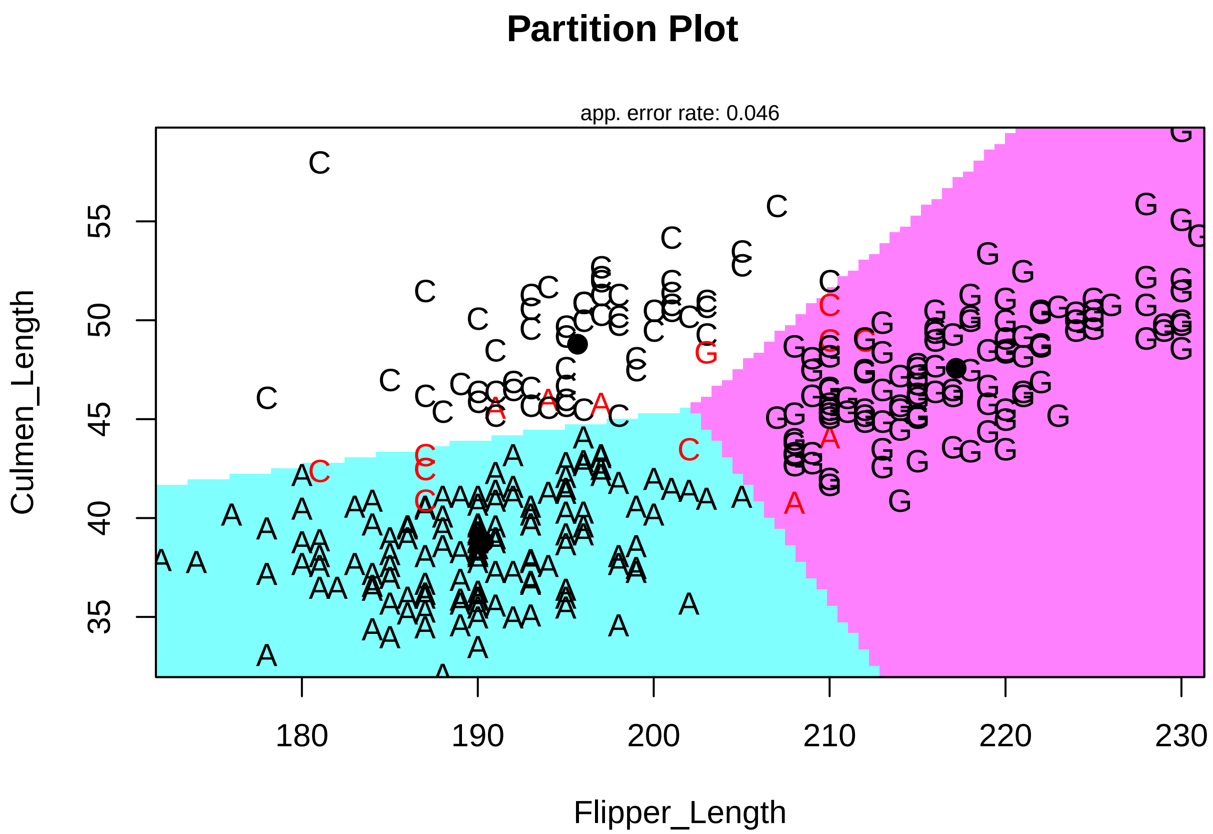
```
## stop criterion: improvement less than 5%.
```

```
## correctness rate: 0.79321;  in: "Flipper_Length";  variables (1): Flipper_Length
## correctness rate: 0.9537;  in: "Culmen_Length";  variables (2): Flipper_Length, Culmen_Length
##
##  hr.elapsed min.elapsed sec.elapsed
##        0.00        0.00        5.39
```

```
print(step_lda)
```

```
## method      : lda
## final model : Species ~ Culmen_Length + Flipper_Length
## <environment: 0x000001d33735fac0>
##
## correctness rate = 0.9537
```

```
partimat(Species ~ Culmen_Length + Flipper_Length,data=df_PCA,, method="lda",
                      direction = "both",
                      fold = nrow(df_PCA))
```



```
# Perform stepwise QDA
step_qda <- stepclass(Species ~ Culmen_Length + Culmen_Depth + Flipper_Length + Body_Mass + Delta.15.N
                      data = df_PCA,
                      method = "qda",
                      direction = "both",
                      fold = nrow(df_PCA)) # LOOCV
```

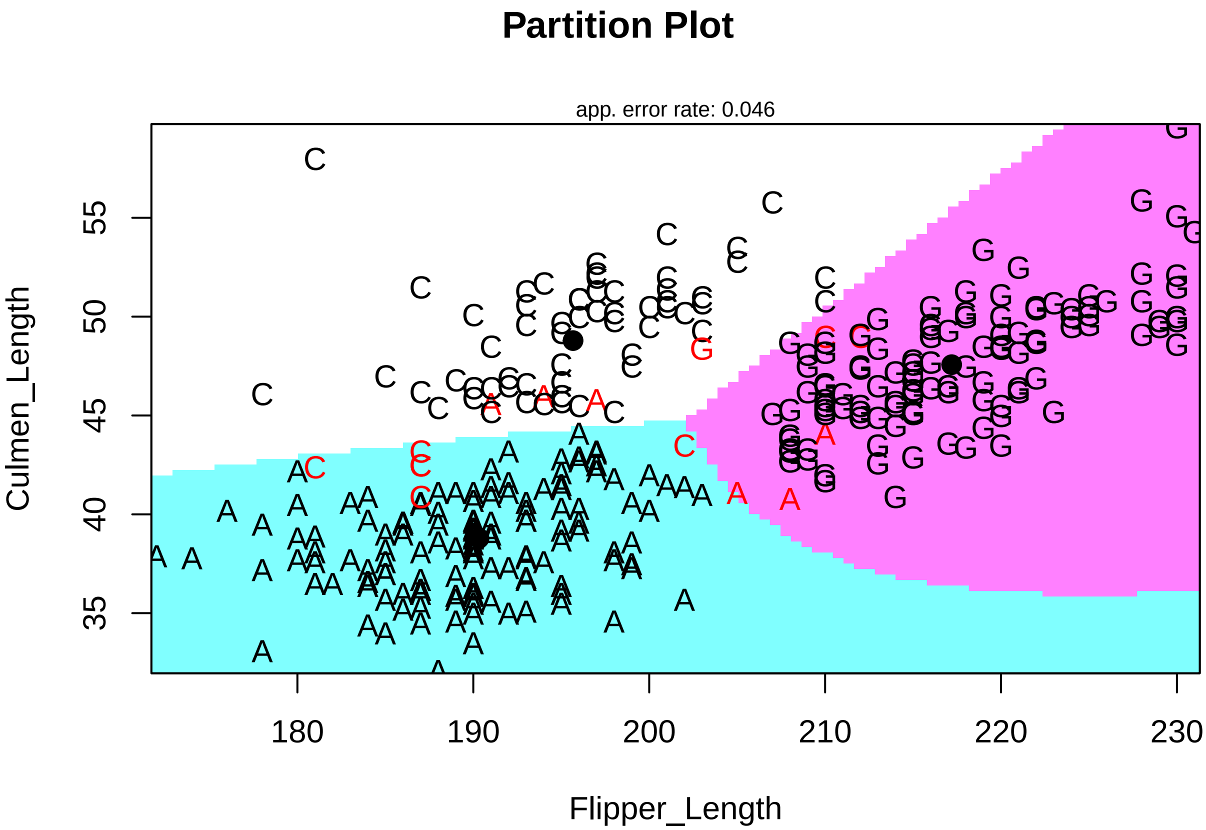
Then, perform step-wise QDA

```
##  'stepwise classification', using 324-fold cross-validated correctness rate of method qda'.

## 324 observations of 6 variables in 3 classes; direction: both

## stop criterion: improvement less than 5%.

## correctness rate: 0.80556;  in: "Flipper_Length";  variables (1): Flipper_Length
## correctness rate: 0.9537;  in: "Culmen_Length";  variables (2): Flipper_Length, Culmen_Length
##
##  hr.elapsed min.elapsed sec.elapsed
##        0.00        0.00        5.14
```

```
# Print stepwise QDA results
print(step_qda)
```

```
## method      : qda
## final model : Species ~ Culmen_Length + Flipper_Length
## <environment: 0x000001d32cd124d8>
##
## correctness rate = 0.9537
```

```
# partimat(group ~ hirecall + lirecall,data=df_alz, method="lda")
partimat(Species ~ Culmen_Length + Flipper_Length , data = df_PCA, method = "qda",
                      direction = "both",
                      fold = nrow(df_PCA))
```



Also, recheck covariance matrices similarity for this two variables.

```
group1_cov <- cov(data[data$Species == 'Adelie', c("Culmen_Length", "Flipper_Length")])
group2_cov <- cov(data[data$Species == 'Chinstrap',  c("Culmen_Length", "Flipper_Length")])
group3_cov <- cov(data[data$Species == 'Gentoo',  c("Culmen_Length", "Flipper_Length")])
cov_rat <- group2_cov/group1_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

```
##                Culmen_Length Flipper_Length
## Culmen_Length            1.5            1.7
## Flipper_Length           1.7            1.2
```

```
cov_rat <- group1_cov/group3_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

```
##                Culmen_Length Flipper_Length
## Culmen_Length            1.3            2.2
## Flipper_Length           2.2            1.0
```

```
cov_rat <- group2_cov/group3_cov
cov_rat[abs(cov_rat) < 1] <- 1/(cov_rat[abs(cov_rat) < 1])
round(cov_rat, 1)
```

```
##                Culmen_Length Flipper_Length
## Culmen_Length            1.1            1.3
## Flipper_Length           1.3            1.1
```

Ratios of largest to smallest elements of the covariance matrices are smaller than 4, for this two variables.

Both LDA and QDA choose Culmen_Length and Flipper_Length as significant discriminating variables and provides same correctness rate = 0.9537. And after recheck the covariance matrices similarity, both LDA and QDA would be good choice.

Let’s try how DA based on PC works

```
pc1 <- prcomp(df_PCA[, 10:15], scale. = TRUE)

# Extract the first k PCs - in your case, all 6
df_all_pc <- data.frame(pc1$x[, 1:6], Species = df_PCA$Species)
```

```
# Box's M statistic
boxM(df_all_pc[,1:3], df_all_pc$Species)
```

```
##
##  Box’s M-test for Homogeneity of Covariance Matrices
##
## data:  df_all_pc[, 1:3]
## Chi-Sq (approx.) = 165.46, df = 12, p-value < 2.2e-16
```

```
# Stepwise LDA
# Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N Delta.13.C
step_lda <- stepclass(Species ~ PC1 + PC2 + PC3+ PC4 + PC5 + PC6,
                      data = df_all_pc,
                      method = "lda",
                      direction = "both",
                      fold = nrow(df_all_pc)) # LOOCV
```

```
##  ‘stepwise classification’, using 324-fold cross-validated correctness rate of method lda’.
```

```
## 324 observations of 6 variables in 3 classes; direction: both
## stop criterion: improvement less than 5%.
```

```
## correctness rate: 0.79321;  in: "PC1";  variables (1): PC1
## correctness rate: 0.96296;  in: "PC2";  variables (2): PC1, PC2
##
##  hr.elapsed min.elapsed sec.elapsed
##        0.00        0.00        5.03
```

```r
print(step_lda)
```

```
## method      : lda
## final model : Species ~ PC1 + PC2
## <environment: 0x000001d343e8cba0>
##
## correctness rate = 0.963
```

```r
# Stepwise LDA
# Culmen_Length Culmen_Depth Flipper_Length Body_Mass Delta.15.N Delta.13.C
step_qda <- stepclass(Species ~ PC1 + PC2 + PC3+ PC4 + PC5 + PC6,
                      data = df_all_pc,
                      method = "qda",
                      direction = "both",
                      fold = nrow(df_all_pc)) # LOOCV
```

```
##  'stepwise classification', using 324-fold cross-validated correctness rate of method qda'.
```

```
## 324 observations of 6 variables in 3 classes; direction: both
```

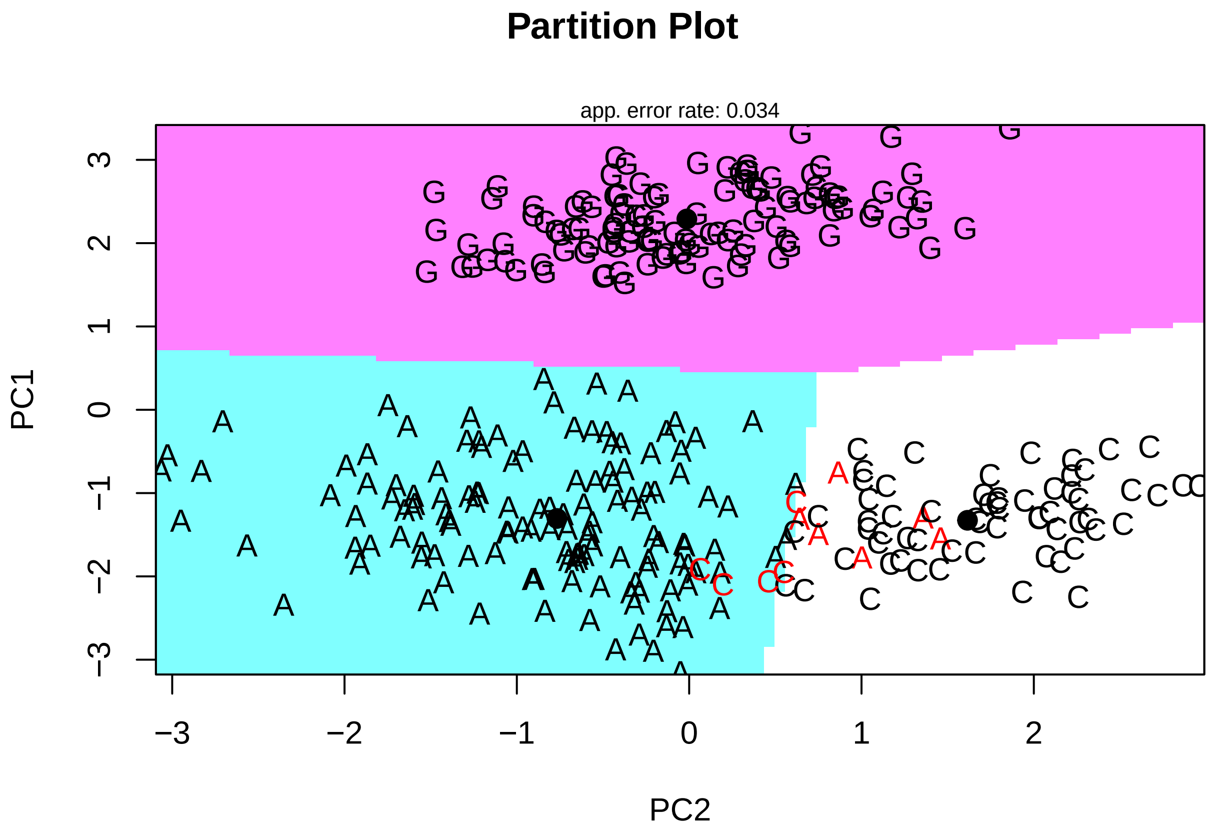
```
## stop criterion: improvement less than 5%.
```

```
## correctness rate: 0.79321;  in: "PC1";  variables (1): PC1
## correctness rate: 0.9537;  in: "PC2";  variables (2): PC1, PC2
##
##  hr.elapsed min.elapsed sec.elapsed
##        0.00        0.00        5.25
```

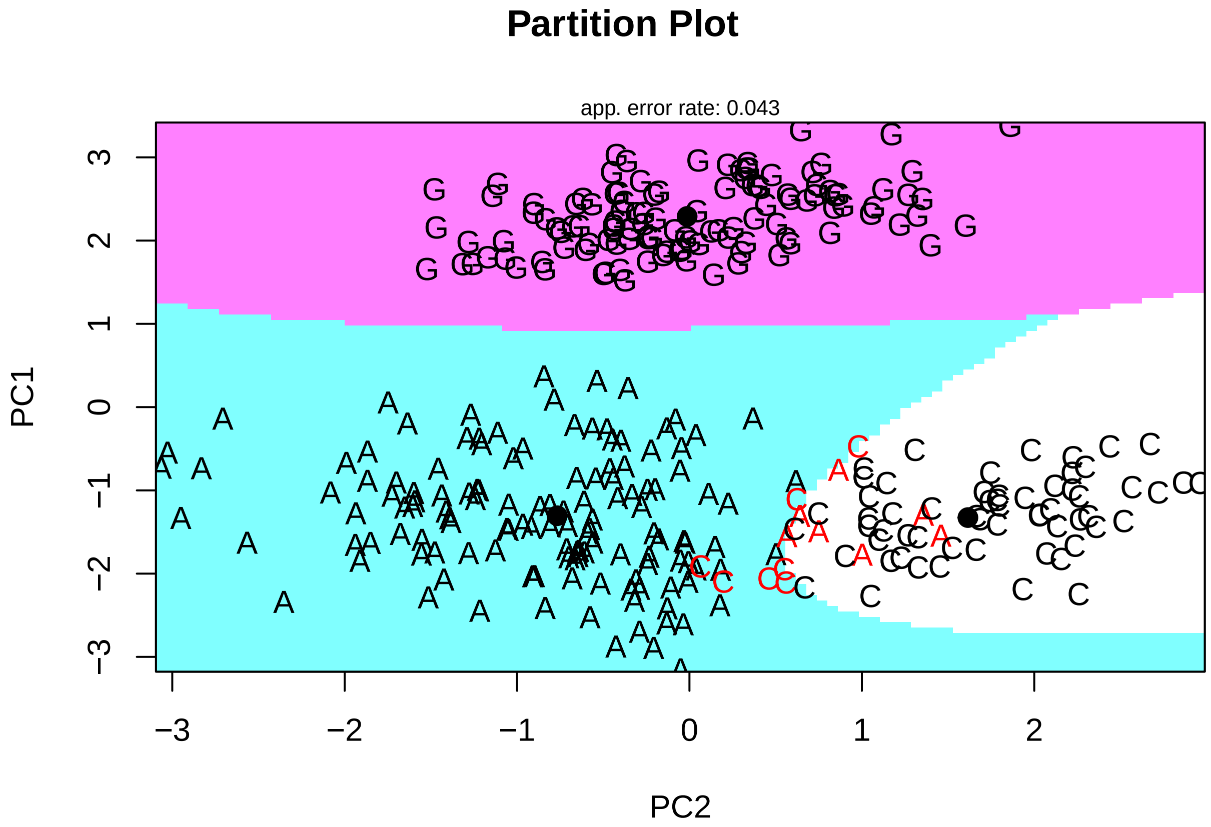
```r
print(step_qda)
```

```
## method      : qda
## final model : Species ~ PC1 + PC2
## <environment: 0x000001d337659388>
##
## correctness rate = 0.9537
```

```r
# Create data frame with PC1, PC2, and Species
df_pc_scores <- data.frame(PC1 = pc1$x[,1],
                           PC2 = pc1$x[,2],
                           Species = df_PCA$Species)

# Plot partition plots for LDA
# partimat(Species ~ PC1 + PC2, data = df_pc_scores, method = "lda")
partimat(Species ~ PC1 + PC2,data=df_pc_scores, method="lda",
                      direction = "both",
                      fold = nrow(df_pc_scores))
```



```r
partimat(Species ~ PC1 + PC2,data=df_pc_scores, method="qda",
                      direction = "both",
                      fold = nrow(df_pc_scores))
```



Basically, based on the stepwise discriminant analysis, the QDA suggest PC1 and PC2 are significant discriminating variable. LDA provides correctness rate of 0.9537, and qda provides correctness rate of 0.9537. Although Box's M test indicated differences in covariance matrices among groups, LDA showed better classification performance than QDA in leave-one-out cross-validation, likely due to its lower model complexity and better generalization.

In conclusion, using principal components does improve the performance, yet the improvement is relatively low. We still consider to use Culmen_Length and Flipper_Length for discriminant analysis

whether there is statistical evidence that the multivariate group means are different using the multivariate Wilk's Lambda test

```
data.manova<-manova(as.matrix(df_PCA[,c("Culmen_Length","Flipper_Length")])~ df_PCA$Species) # Species
summary.manova(data.manova,test= "Wilks")
```

```
##                   Df    Wilks approx F num Df den Df    Pr(>F)
## df_PCA$Species     2 0.087547   380.75      4    640 < 2.2e-16 ***
## Residuals        321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-values < 2.2e-16 which is smaller than 0.05 or 0.01, then we reject the null hypothesis and concludes that the multivariate group means are different while using Culmen_Length and Flipper_Length as discriminators.

```
source("https://raw.githubusercontent.com/jreuning/sds363_code/refs/heads/main/discrim.r.txt")
```

Use LDA with Culmen_Length + Flipper_Length

```
discriminant.significance(df_PCA[, c("Culmen_Length","Flipper_Length")], df_PCA$Species)
```

```
##   Test of Function(s) Wilks Lambda Approximate F p-value
## 1         1 through 2       0.0875      380.7549       0
## 2                   2       0.3923      497.1769       0
```

There are 2 discriminant functions. Both discriminant functions are significant(both p-value<0.05).

For Test of Function(s) 1 through 2:

The Wilks' Lambda value of 0.0875 indicates that the combination of both functions explains a significant portion of the variance in group separation (smaller values indicate better discrimination). The F-statistic (380.7549) and a very small p-value (0.0000) suggest that the functions jointly are statistically significant (i.e., they help distinguish between species).

For Test of Function 2:

The Wilks' Lambda is 0.3923, indicating that function 2 explains smaller variance than the combination of both functions, yet explains significant portion of the variance in group separation. The F-statistic (497.1769) is larger, and the p-value is statistically significant ($p < 0.05$), meaning that function 2 does significantly contribute to group separation.

Provide some evidence as to which of your original variables are the 'best' discriminators amongst your groups (look at standardized discriminant coefficients)

```
df_scaled <- lda(scale(df_PCA[, c("Culmen_Length","Flipper_Length")]), grouping = df_PCA$Species)  # df

df_scaled
```

```
## Call:
## lda(scale(df_PCA[, c("Culmen_Length", "Flipper_Length")]), grouping = df_PCA$Species)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4290123 0.2067901 0.3641975
##
## Group means:
##           Culmen_Length Flipper_Length
## Adelie       -0.9611630     -0.7803250
## Chinstrap     0.8638898     -0.3972014
## Gentoo        0.6417038      1.1447259
##
## Coefficients of linear discriminants:
##                      LD1       LD2
## Culmen_Length  0.3916693 -2.070900
## Flipper_Length 1.8350621  1.568639
##
## Proportion of trace:
##    LD1    LD2
## 0.6921 0.3079
```

LD1 (First Discriminant Function) explains 69.21% of the variance, making it the primary function responsible for group separation. LD2 (Second Discriminant Function) explains only 30.79% of the variance, contributing relatively smaller to discrimination between groups. This indicates that most of the discriminating power is concentrated in the first function, yet both functions together contributing almost all of the discriminating power.

```r
print("Standardized Coefficients")
```

```
## [1] "Standardized Coefficients"
```

```r
# Coefficients of linear discriminants:
round(df_scaled$scaling, 2)
```

```
##                  LD1   LD2
## Culmen_Length  0.39 -2.07
## Flipper_Length 1.84  1.57
```

Culmen_Length has moderate contribution to LD1, but a strong negative contribution to LD2 — important for separating groups along LD2. Flipper_Length has the largest positive influence on both LD1 and LD2 — the strongest overall discriminator between species. Flipper_Length is clearly the best overall discriminator — it contributes strongly to both LD1 and LD2. Culmen_Length is important primarily for LD2, but less so for LD1. This aligns with our earlier PCA findings: flipper length (and body size) drives much of the separation, especially for Gentoo penguins, while culmen traits play a bigger role in differentiating Chinstrap and Adelie penguins along LD2.

standard and CV difference: no difference as concluded

For LDA,

```r
# lda(scale(df_PCA[, c("Culmen_Length","Flipper_Length")]), grouping = df_PCA$Species)
alz_lda <- lda(df_PCA[, c("Culmen_Length","Flipper_Length")],grouping = df_PCA$Species)
# raw results
(raw <- table(df_PCA$Species, predict(alz_lda)$class))
```

```
##
##             Adelie Chinstrap Gentoo
##   Adelie       134         3      2
##   Chinstrap      6        58      3
##   Gentoo         0         1    117
```

```r
# total percent correct
round(sum(diag(prop.table(raw))), 2)
```

```
## [1] 0.95
```

```r
alz_ldaCV<-  lda(df_PCA[, c("Culmen_Length","Flipper_Length")],grouping = df_PCA$Species, CV = TRUE)

#cross validated results
(CV <- table(df_PCA$Species, alz_ldaCV$class))
```

```
##
##               Adelie Chinstrap Gentoo
##   Adelie         134         3      2
##   Chinstrap        6        58      3
##   Gentoo           0         1    117
```

```
# total percent correct
round(sum(diag(prop.table(CV))), 2)
```

```
## [1] 0.95
```

For QDA,

```
alz_qda <- qda(df_PCA[, c("Culmen_Length","Flipper_Length")],grouping = df_PCA$Species)
# raw results
(raw <- table(df_PCA$Species, predict(alz_qda)$class))
```

```
##
##               Adelie Chinstrap Gentoo
##   Adelie         133         3      3
##   Chinstrap        6        59      2
##   Gentoo           0         1    117
```

```
# total percent correct
round(sum(diag(prop.table(raw))), 2)
```

```
## [1] 0.95
```

```
alz_qdaCV <- qda(df_PCA[, c("Culmen_Length","Flipper_Length")],grouping = df_PCA$Species, CV = TRUE)

#cross validated results
(CV <- table(df_PCA$Species, alz_qdaCV$class))
```

```
##
##               Adelie Chinstrap Gentoo
##   Adelie         133         3      3
##   Chinstrap        6        59      2
##   Gentoo           0         1    117
```

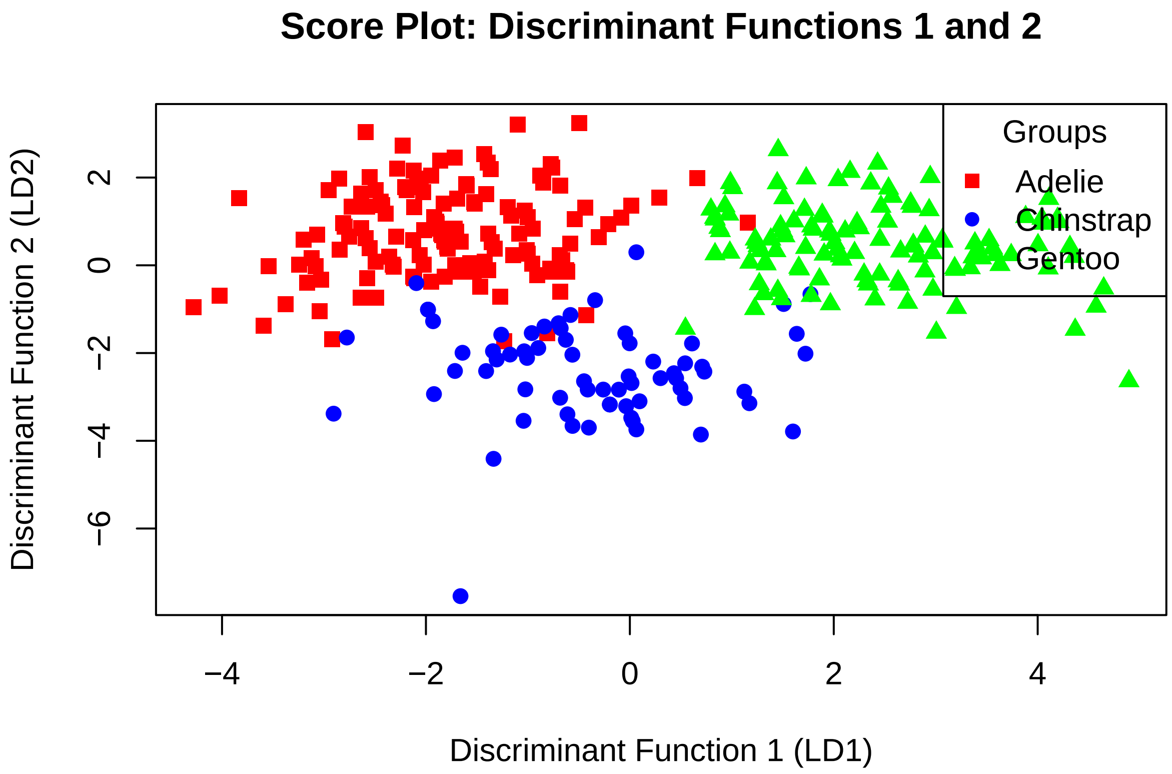
```
# total percent correct
round(sum(diag(prop.table(CV))), 2)
```

```
## [1] 0.95
```

All four models lda, ldaCV,qda, qdaCV provides same correctness.

Make score plots for the first two or three DA function scores

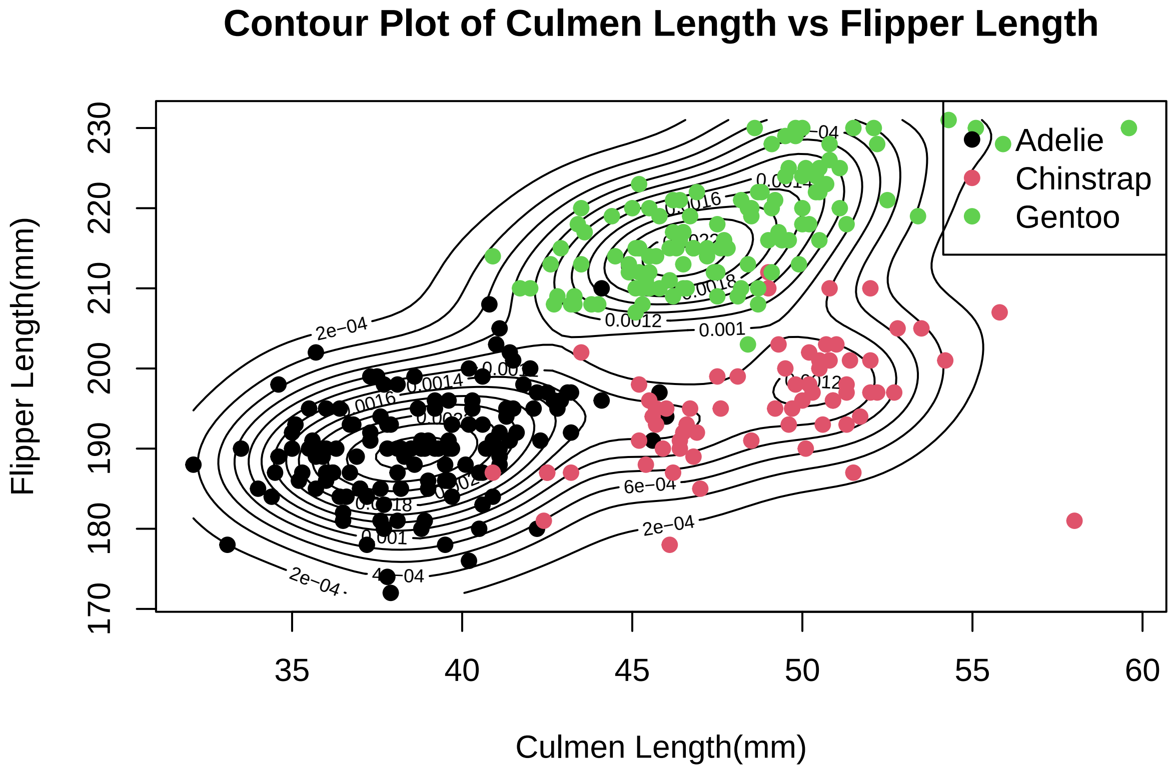
```
# Get LDA scores
lda_scores <- predict(alz_lda)$x  # Scores for the first two discriminant functions
levels(df_PCA$Species) <- c("Adelie", "Chinstrap", "Gentoo")
group_labels <- df_PCA$Species       # Group labels

# Set up colors and symbols for each group
colors <- c("red", "blue", "green")       # Colors for the groups
symbols <- c(15, 16, 17)                  # Symbols for each group

# Plot the scores for the first two discriminant functions
plot(lda_scores[, 1], lda_scores[, 2],
     col = colors[group_labels],
     pch = symbols[group_labels],
     xlab = "Discriminant Function 1 (LD1)",
     ylab = "Discriminant Function 2 (LD2)",
     main = "Score Plot: Discriminant Functions 1 and 2",cex=1.1)

# Add a legend to differentiate groups
legend("topright", legend = levels(group_labels),col = colors, pch = symbols, title = "Groups")
```



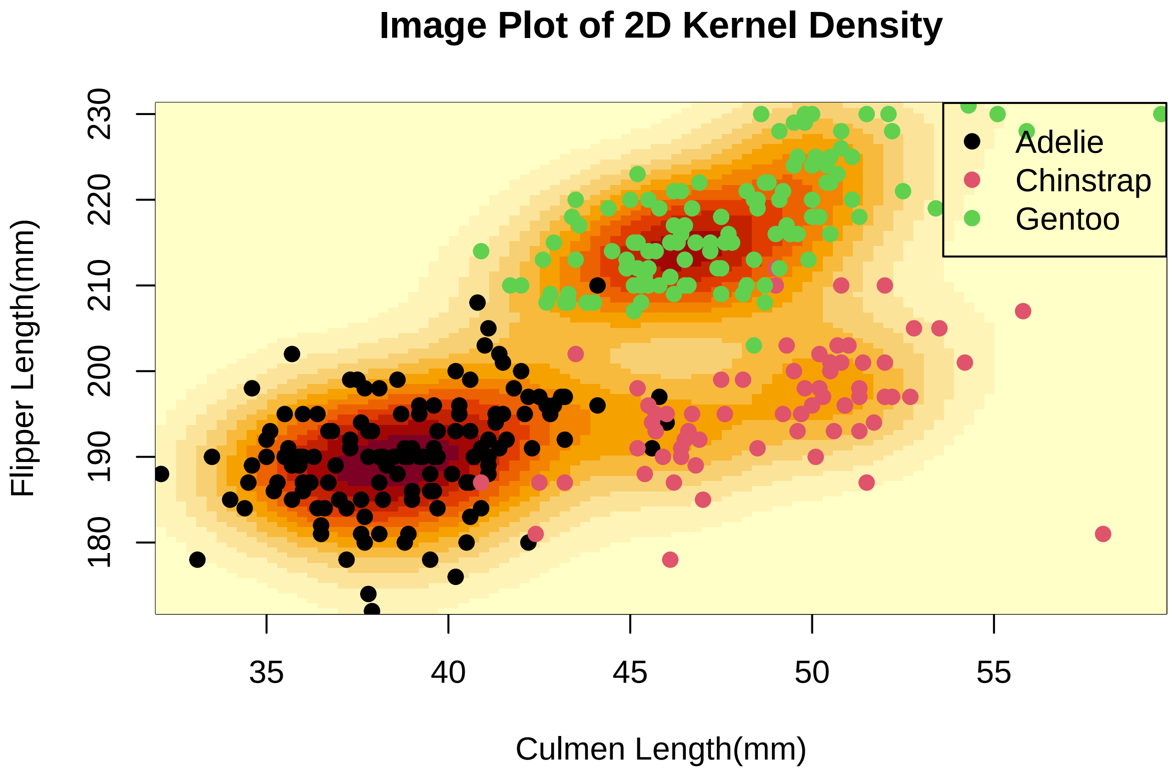
Based on the score plots for the first two DA function scores, it shows that LD1 effectively separates Group 1 (red) from Groups 3 (Gentoo), then with the help of LD2, Adelie and Chinstrap are clearly separated. The distinct clustering of groups suggests that LDA is highly effective in classifying these species based on the selected features.

try kernel smoothing or k-nearest neighbors.

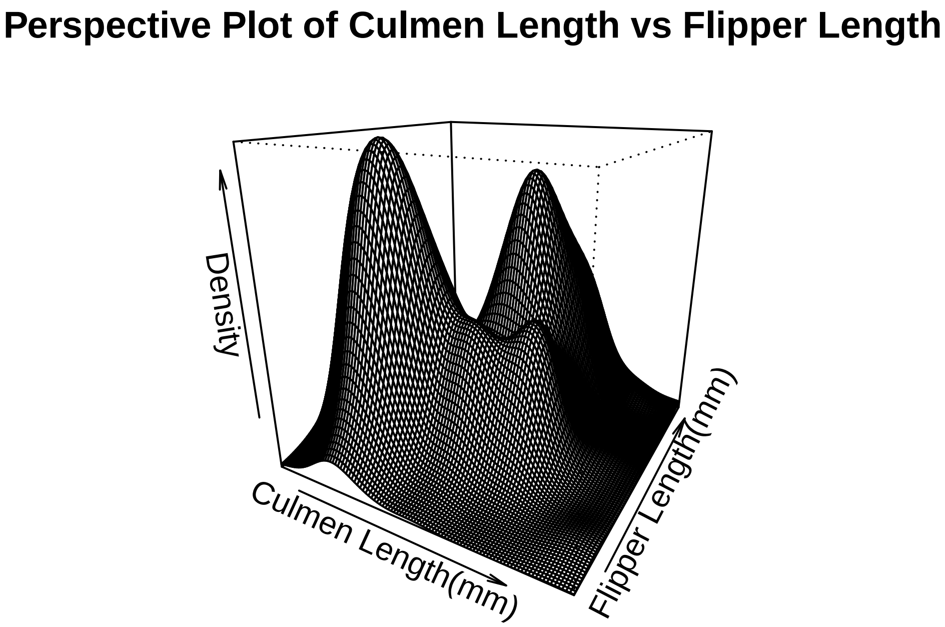
```
# Create kernel density estimate for
# c("Culmen_Length","Flipper_Length")
kde_result <- kde2d(df_PCA$Culmen_Length, df_PCA$Flipper_Length, n = 100)
# Contour plot to visualize density
contour(kde_result,
        xlab = "Culmen Length(mm)",
        ylab = "Flipper Length(mm)",
        main = "Contour Plot of Culmen Length vs Flipper Length")

# Add group points to the plot
points(df_PCA$Culmen_Length, df_PCA$Flipper_Length, col = as.numeric(df_PCA$Species), pch = 19)
legend("topright", legend = levels(df_PCA$Species), col = 1:3, pch = 19)
```



```
# Image plot with color gradient for density
image(kde_result,
      xlab = "Culmen Length(mm)",
      ylab = "Flipper Length(mm)",
      main = "Image Plot of 2D Kernel Density")

# Add points for each observation
points(df_PCA$Culmen_Length, df_PCA$Flipper_Length, col = as.numeric(df_PCA$Species), pch = 19)
legend("topright", legend = levels(df_PCA$Species), col = 1:3, pch = 19)
```



```
# Perspective (3D surface) plot
persp(kde_result,
      xlab = "Culmen Length(mm)",
      ylab = "Flipper Length(mm)",
      zlab = "Density",
      theta = 30, phi = 20,
      main = "Perspective Plot of Culmen Length vs Flipper Length")
```



The KDE plots visually reinforce what I found in LDA

Choose the kernel smoothing with bandwidths calculated based on Silverman's Rule of Thumb. Above Contour, Perspective, and Image plots show:

Culmen length and flipper length provide excellent discrimination of all three species. Culmen length and flipper length provide excellent discrimination of Gentoo penguins from the other two species, but we also notice these variables offer limited separation between Adelie and Chinstrap.

All in all, Culmen length and flipper length would be best two variables when considering making discrimination of all three species. However, if we have all the data, use Principle Components in Discriminant Analysis would provider better result.

MANOVA

Whether geographical environment and sex has effect on penguins' diet and Culmen size

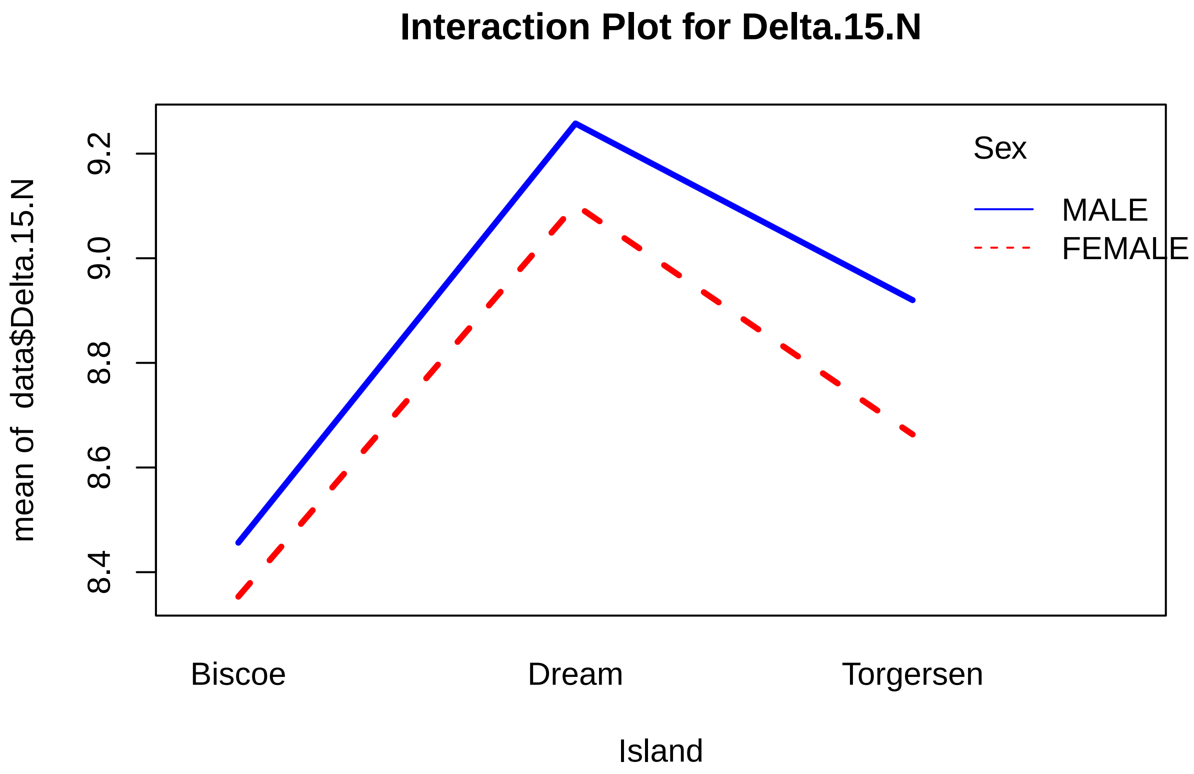
Sex of the penguin may affect their eating habits.

we will use Delta.15.N and Culmen length, Culmen depth as our response variables, and use Island, Sex as our two categorical predictors, with flipper length (body size) as our extra continuous predictor.

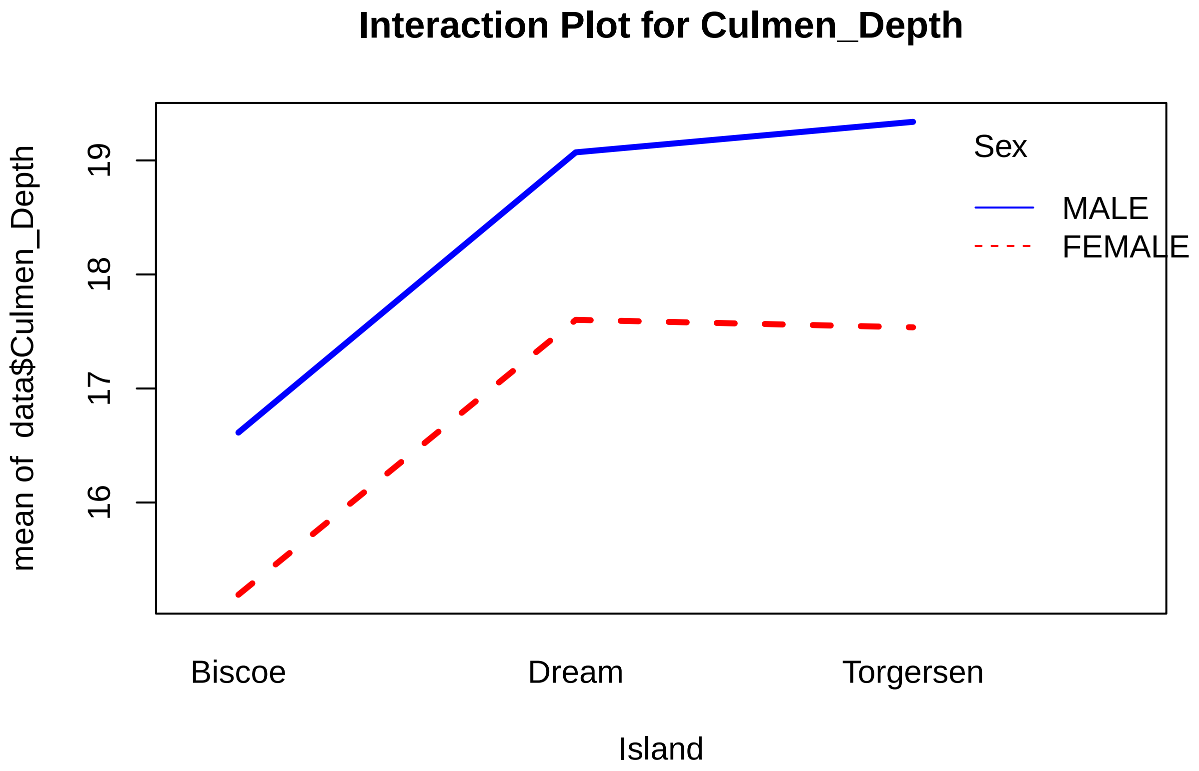
```r
library(MASS)
library(biotools)
library(heplots)
library(klaR)
library(car)
```

make interaction plots

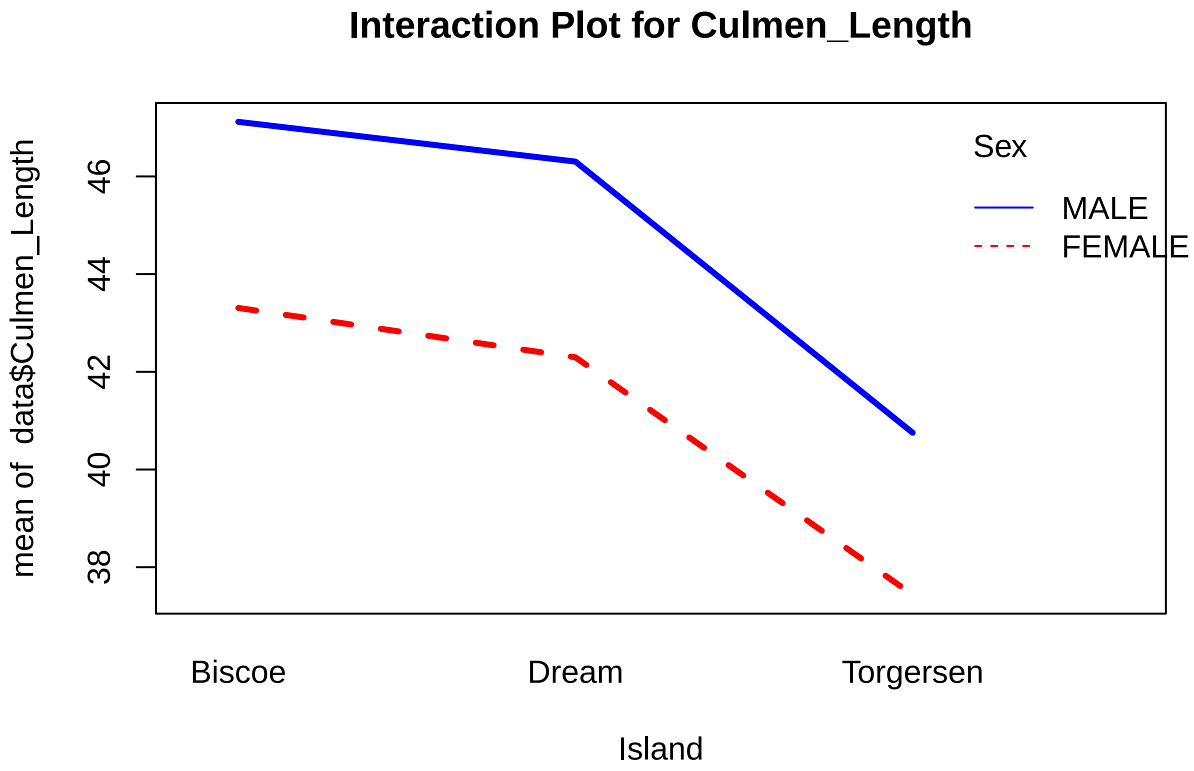
```
interaction.plot(data$Island, data$Sex, data$Delta.15.N,
  lwd=3, col=c("red", "blue", "black"), trace.label="Sex",
  xlab="Island", main="Interaction Plot for Delta.15.N")
```



```
interaction.plot(data$Island, data$Sex, data$Culmen_Depth,
  lwd=3, col=c("red", "blue", "black"), trace.label="Sex",
  xlab="Island", main="Interaction Plot for Culmen_Depth")
```



```r
interaction.plot(data$Island, data$Sex, data$Culmen_Length,
  lwd=3, col=c("red", "blue", "black"), trace.label="Sex",
  xlab="Island", main="Interaction Plot for Culmen_Length")
```



All three plots show parallel trend, which means there may not be interaction effect between Island (geographical environment),Sex for Delta.15.N, Culmen size. For Delta.15.N, the lines show a sharp increase from Biscoe to Dream and then a slighter decrease for Torgersen, this indicates penguins have higher to lower trophic levels or longer to shorter food chains in Dream, Torgersen, and Biscoe. For Culmen Depth, Torgersen and Dream island's penguins share simlar values and much higher than Biscoe' island's penguins. For Culmen Length, Biscoe and Dream island's penguins share simlar values and much higher than Torgersen' island's penguins. Also by all three plots, male penguins shows higher trophic levels, longer food chains(Delta.15.N), bigger culmen(bill) than female penguins.

Notice: higher Delta.15.N indicating higher trophic levels, longer food chains (Delta.13.C also indicates they eat more fishes that eat C4 plants)

Two-Way MANOVA for these two categorical factors

```
options(contrasts=c("contr.sum", "contr.poly"))

penguinsMAOV <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island*Sex,
                data=data)

#Multivariate and univariate results
summary(Anova(penguinsMAOV, type=3), univariate=T)
```

```
##
## Type III MANOVA Tests:
##
```

```
## Sum of squares and products for error:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      55.51976     61.95268     -34.89668
## Culmen_Depth    61.95268    589.58956    -752.79834
## Culmen_Length  -34.89668   -752.79834    7213.91354
##
## ------------------------------------------
##
## Term: (Intercept)
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      18379.99     36706.68      89618.63
## Culmen_Depth    36706.68     73306.93     178977.42
## Culmen_Length   89618.63    178977.42     436969.83
##
## Multivariate Tests: (Intercept)
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1     0.998 51802.94      3    316 < 2.22e-16 ***
## Wilks             1     0.002 51802.94      3    316 < 2.22e-16 ***
## Hotelling-Lawley  1   491.800 51802.94      3    316 < 2.22e-16 ***
## Roy               1   491.800 51802.94      3    316 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## ------------------------------------------
##
## Term: Island
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      41.33768     132.6715     -61.02527
## Culmen_Depth   132.67146     490.5780    -473.67152
## Culmen_Length  -61.02527    -473.6715    1281.60735
##
## Multivariate Tests: Island
##                  Df test stat  approx F num Df den Df     Pr(>F)
## Pillai            2 0.6911535  55.79866      6    634 < 2.22e-16 ***
## Wilks             2 0.3926778  62.75891      6    632 < 2.22e-16 ***
## Hotelling-Lawley  2 1.3331311  69.98938      6    630 < 2.22e-16 ***
## Roy               2 1.1470064 121.20035      3    317 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## ------------------------------------------
##
## Term: Sex
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      1.753617     15.97147      37.88654
## Culmen_Depth   15.971471    145.46389     345.06044
## Culmen_Length  37.886536    345.06044     818.53102
##
```

```
## Multivariate Tests: Sex
##                  Df test stat approx F num Df den Df     Pr(>F)    
## Pillai            1 0.3580715 58.75557      3    316 < 2.22e-16 ***
## Wilks             1 0.6419285 58.75557      3    316 < 2.22e-16 ***
## Hotelling-Lawley  1 0.5578060 58.75557      3    316 < 2.22e-16 ***
## Roy               1 0.5578060 58.75557      3    316 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## ------------------------------------------
##  
## Term: Island:Sex 
## 
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     0.2081161    0.4828945    -0.5302624
## Culmen_Depth   0.4828945    1.2434648    -1.7757252
## Culmen_Length -0.5302624   -1.7757252     3.7690361
## 
## Multivariate Tests: Island:Sex
##                  Df test stat  approx F num Df den Df  Pr(>F)
## Pillai            2 0.0047338 0.2506957      6    634 0.95902
## Wilks             2 0.9952679 0.2501137      6    632 0.95925
## Hotelling-Lawley  2 0.0047530 0.2495303      6    630 0.95948
## Roy               2 0.0043698 0.4617402      3    317 0.70920
## 
##  Type III Sums of Squares
##              df Delta.15.N Culmen_Depth Culmen_Length
## (Intercept)   1 1.8380e+04   73306.9304    436969.832
## Island        2 4.1338e+01     490.5780      1281.607
## Sex           1 1.7536e+00     145.4639       818.531
## Island:Sex    2 2.0812e-01       1.2435         3.769
## residuals   318 5.5520e+01     589.5896      7213.914
## 
##  F-tests
##             Delta.15.N Culmen_Depth Culmen_Length
## (Intercept)  105274.87     19769.35      19262.28
## Island          118.38       264.60         28.25
## Sex              10.04        39.23         36.08
## Island:Sex        0.60         0.67          0.08
## 
##  p-values
##             Delta.15.N Culmen_Depth Culmen_Length
## (Intercept) < 2.22e-16 < 2.22e-16   < 2.22e-16   
## Island      < 2.22e-16 < 2.22e-16   5.1047e-12   
## Sex         0.0016769  5.9253e-16   5.1694e-09   
## Island:Sex  0.5516187  0.4134304    0.9203045    
```

Multivariate MANOVA Results: Island: Very strong multivariate effect. Island significantly affects the combination of Delta.15.N, Culmen Depth, and Culmen Length. Sex: Strong multivariate effect. Penguin sex also significantly affects the combination of traits. Island:Sex: No significant interaction. The effect of sex on the traits does not differ across islands.

Both Island and Sex are significant predictors, independently affecting penguin traits. The interaction is not significant, so the effects of Island and Sex are additive, not synergistic.

Univariate ANOVA Results: For island, all three traits differ significantly by Island. Island likely reflects environmental variation or geographic isolation that influences penguin morphology and diet (stable isotope). For sex, male and female penguins differ significantly in all three traits. This likely reflects sexual dimorphism — males tend to have larger bill and may occupy slightly different foraging niches (reflected in isotope values). For Interaction of Island and Sex, no significant interaction effects — the difference between sexes is consistent across islands, and vice versa.

All three traits are influenced by both factors — suggesting that both geographic location and biological sex shape penguin traits.

Perform multivariate and univariate contrasts to compare levels of a particular factor(island here).

```
options(contrasts = c("contr.treatment", "contr.poly"))
contrasts(data$Island)
```

```
##           Dream Torgersen
## Biscoe        0         0
## Dream         1         0
## Torgersen     0         1
```

```
#penguinsMAOV <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island*Sex,  data=data)
IslandManova <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island, data = data)
```

```
linearHypothesis(IslandManova, "IslandDream = 0")
```

```
## 
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      40.95158     127.5574     -52.27429
## Culmen_Depth   127.55735     397.3199    -162.82571
## Culmen_Length  -52.27429    -162.8257      66.72761
## 
## Sum of squares and products for error:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     57.376605      79.6482      8.668802
## Culmen_Depth   79.648196     770.5315   -294.217632
## Culmen_Length   8.668802    -294.2176   8397.035871
## 
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)    
## Pillai            1 0.4738718 95.77202      3    319 < 2.22e-16 ***
## Wilks             1 0.5261282 95.77202      3    319 < 2.22e-16 ***
## Hotelling-Lawley  1 0.9006773 95.77202      3    319 < 2.22e-16 ***
## Roy               1 0.9006773 95.77202      3    319 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
linearHypothesis(IslandManova, "IslandTorgersen = 0")
```

```
## 
## Sum of squares and products for the hypothesis:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      4.990734     32.62375     -80.47939
## Culmen_Depth   32.623749    213.25701    -526.08283
## Culmen_Length -80.479392   -526.08283    1297.79156
## 
## Sum of squares and products for error:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     57.376605      79.6482      8.668802
## Culmen_Depth   79.648196     770.5315   -294.217632
## Culmen_Length   8.668802    -294.2176   8397.035871
## 
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)    
## Pillai            1 0.2877836 42.96586      3    319 < 2.22e-16 ***
## Wilks             1 0.7122164 42.96586      3    319 < 2.22e-16 ***
## Hotelling-Lawley  1 0.4040677 42.96586      3    319 < 2.22e-16 ***
## Roy               1 0.4040677 42.96586      3    319 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(IslandManova,"IslandTorgersen - IslandDream=0")
```

```
## 
## Sum of squares and products for the hypothesis:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      4.788736   -1.2130495      63.87678
## Culmen_Depth   -1.213050    0.3072813     -16.18082
## Culmen_Length  63.876778  -16.1808249     852.05012
## 
## Sum of squares and products for error:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     57.376605      79.6482      8.668802
## Culmen_Depth   79.648196     770.5315   -294.217632
## Culmen_Length   8.668802    -294.2176   8397.035871
## 
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)    
## Pillai            1 0.1614964 20.47988      3    319 3.6714e-12 ***
## Wilks             1 0.8385036 20.47988      3    319 3.6714e-12 ***
## Hotelling-Lawley  1 0.1926007 20.47988      3    319 3.6714e-12 ***
## Roy               1 0.1926007 20.47988      3    319 3.6714e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(IslandManova,"2*IslandTorgersen-IslandDream=0") # IslandTorgersen vs others
```

```
## 
## Sum of squares and products for the hypothesis:
##                   Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     0.0003493435   -0.1483641     0.6480748
## Culmen_Depth  -0.1483641366   63.0093728  -275.2335585
## Culmen_Length  0.6480748409 -275.2335585  1202.2578276
##
## Sum of squares and products for error:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N     57.376605      79.6482      8.668802
## Culmen_Depth   79.648196     770.5315   -294.217632
## Culmen_Length   8.668802    -294.2176   8397.035871
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.1751576 22.58018      3    319 2.7796e-13 ***
## Wilks             1 0.8248424 22.58018      3    319 2.7796e-13 ***
## Hotelling-Lawley  1 0.2123528 22.58018      3    319 2.7796e-13 ***
## Roy               1 0.2123528 22.58018      3    319 2.7796e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All three contrasts are statistically significant, indicating meaningful morphological or dietary differences (isotopic signatures) among penguins across the different islands. This reinforces the earlier full-model MANOVA result showing a strong Island main effect. These differences may reflect: Geographic isolation Habitat-driven food web variation (explaining Delta.15.N) Morphological adaptation (explaining culmen traits)

```r
#penguinsMAOV <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island*Sex,  data=data)
IslandManova2 <- lm(cbind(Delta.15.N) ~ Island, data = data)
```

```r
linearHypothesis(IslandManova2, "IslandDream = 0")
```

```
##
## Linear hypothesis test:
## IslandDream = 0
##
## Model 1: restricted model
## Model 2: cbind(Delta.15.N) ~ Island
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    322 98.328
## 2    321 57.377  1    40.952 229.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(IslandManova2, "IslandTorgersen = 0")
```

```
##
## Linear hypothesis test:
## IslandTorgersen = 0
##
## Model 1: restricted model
## Model 2: cbind(Delta.15.N) ~ Island
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    322 62.367
## 2    321 57.377  1    4.9907 27.921 2.338e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
linearHypothesis(IslandManova2,"IslandTorgersen - IslandDream=0")
```

```
##
## Linear hypothesis test:
## - IslandDream  + IslandTorgersen = 0
##
## Model 1: restricted model
## Model 2: cbind(Delta.15.N) ~ Island
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    322 62.165
## 2    321 57.377  1    4.7887 26.791 4.005e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
linearHypothesis(IslandManova2,"2*IslandTorgersen-IslandDream=0") # IslandTorgersen vs others
```

```
##
## Linear hypothesis test:
## - IslandDream  + 2 IslandTorgersen = 0
##
## Model 1: restricted model
## Model 2: cbind(Delta.15.N) ~ Island
##
##   Res.Df    RSS Df  Sum of Sq     F Pr(>F)
## 1    322 57.377
## 2    321 57.377  1 0.00034934 0.002 0.9648
```

```
linearHypothesis(IslandManova2,"IslandTorgersen+IslandDream=0")
```

```
##
## Linear hypothesis test:
## IslandDream  + IslandTorgersen = 0
##
## Model 1: restricted model
## Model 2: cbind(Delta.15.N) ~ Island
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    322 81.085
## 2    321 57.377  1    23.708 132.64 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Dream vs. Biscoe and Torgersen vs. Biscoe are both highly significant (F = 229.11 and 27.92, respectively; $p < 0.001$), indicating clear ecological differences in trophic position.

Torgersen vs. Dream (Torgersen - Dream = 0) is also significant (F = 26.79, $p < 0.001$), confirming distinct foraging niches even between these two smaller islands.

The contrast 2*Torgersen - Dream = 0 (testing if Torgersen differs from the average of the other islands) is not significant (p = 0.9648), suggesting Torgersen sits midway in trophic values between Dream and Biscoe.

The contrast Torgersen + Dream = 0 (i.e., their sum equals Biscoe *2) is highly significant (F = 132.64, p < 0.001), reaffirming that Torgersen and Dream jointly differ from Biscoe.

These tests show that penguins on each island differ in trophic level, with Dream having the highest, Biscoe the lowest, and Torgersen intermediate. This likely reflects differences in local food webs or foraging habitats, supporting the idea of island-specific ecological niches.

```
#penguinsMAOV <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island*Sex,  data=data)
IslandManova3 <- lm(cbind(Culmen_Depth, Culmen_Length) ~ Island, data = data)
```

```
linearHypothesis(IslandManova3, "IslandDream = 0")
```

```
##
## Sum of squares and products for the hypothesis:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      397.3199    -162.82571
## Culmen_Length    -162.8257      66.72761
##
## Sum of squares and products for error:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      770.5315     -294.2176
## Culmen_Length    -294.2176     8397.0359
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.3402308 82.50904      2    320 < 2.22e-16 ***
## Wilks             1 0.6597692 82.50904      2    320 < 2.22e-16 ***
## Hotelling-Lawley  1 0.5156815 82.50904      2    320 < 2.22e-16 ***
## Roy               1 0.5156815 82.50904      2    320 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
linearHypothesis(IslandManova3, "IslandTorgersen = 0")
```

```
##
## Sum of squares and products for the hypothesis:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      213.2570     -526.0828
## Culmen_Length    -526.0828     1297.7916
##
## Sum of squares and products for error:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      770.5315     -294.2176
## Culmen_Length    -294.2176     8397.0359
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.2798889 62.18795      2    320 < 2.22e-16 ***
## Wilks             1 0.7201111 62.18795      2    320 < 2.22e-16 ***
## Hotelling-Lawley  1 0.3886747 62.18795      2    320 < 2.22e-16 ***
## Roy                1 0.3886747 62.18795      2    320 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(IslandManova3,"IslandTorgersen - IslandDream=0")
```

```
##
## Sum of squares and products for the hypothesis:
##               Culmen_Depth Culmen_Length
## Culmen_Depth     0.3072813     -16.18082
## Culmen_Length  -16.1808249     852.05012
##
## Sum of squares and products for error:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      770.5315     -294.2176
## Culmen_Length    -294.2176     8397.0359
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.0923605 16.28144      2    320 1.8456e-07 ***
## Wilks             1 0.9076395 16.28144      2    320 1.8456e-07 ***
## Hotelling-Lawley  1 0.1017590 16.28144      2    320 1.8456e-07 ***
## Roy               1 0.1017590 16.28144      2    320 1.8456e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(IslandManova3,"2*IslandTorgersen-IslandDream=0") # IslandTorgersen vs others
```

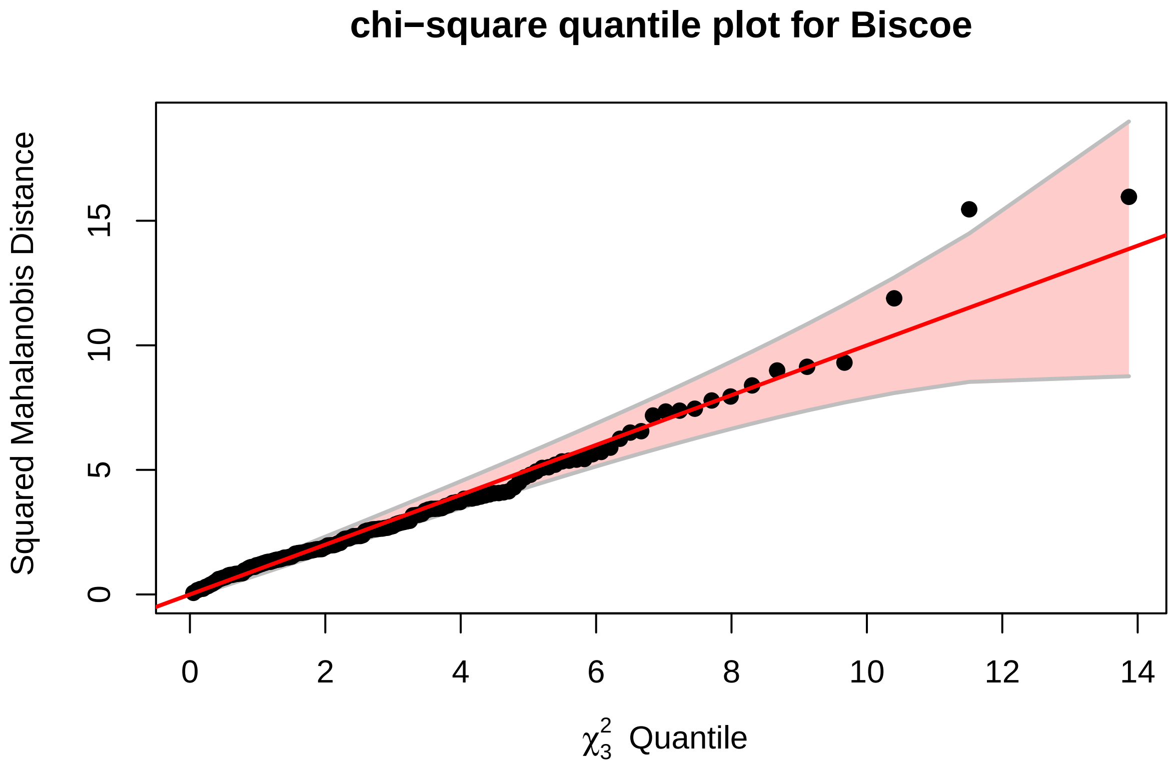
```
##
## Sum of squares and products for the hypothesis:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      63.00937     -275.2336
## Culmen_Length   -275.23356     1202.2578
##
## Sum of squares and products for error:
##               Culmen_Depth Culmen_Length
## Culmen_Depth      770.5315     -294.2176
## Culmen_Length    -294.2176     8397.0359
##
## Multivariate Tests:
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.1684891  32.4208      2    320 1.5096e-13 ***
## Wilks             1 0.8315109  32.4208      2    320 1.5096e-13 ***
## Hotelling-Lawley  1 0.2026300  32.4208      2    320 1.5096e-13 ***
## Roy               1 0.2026300  32.4208      2    320 1.5096e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(IslandManova3,"IslandTorgersen+IslandDream=0")
```

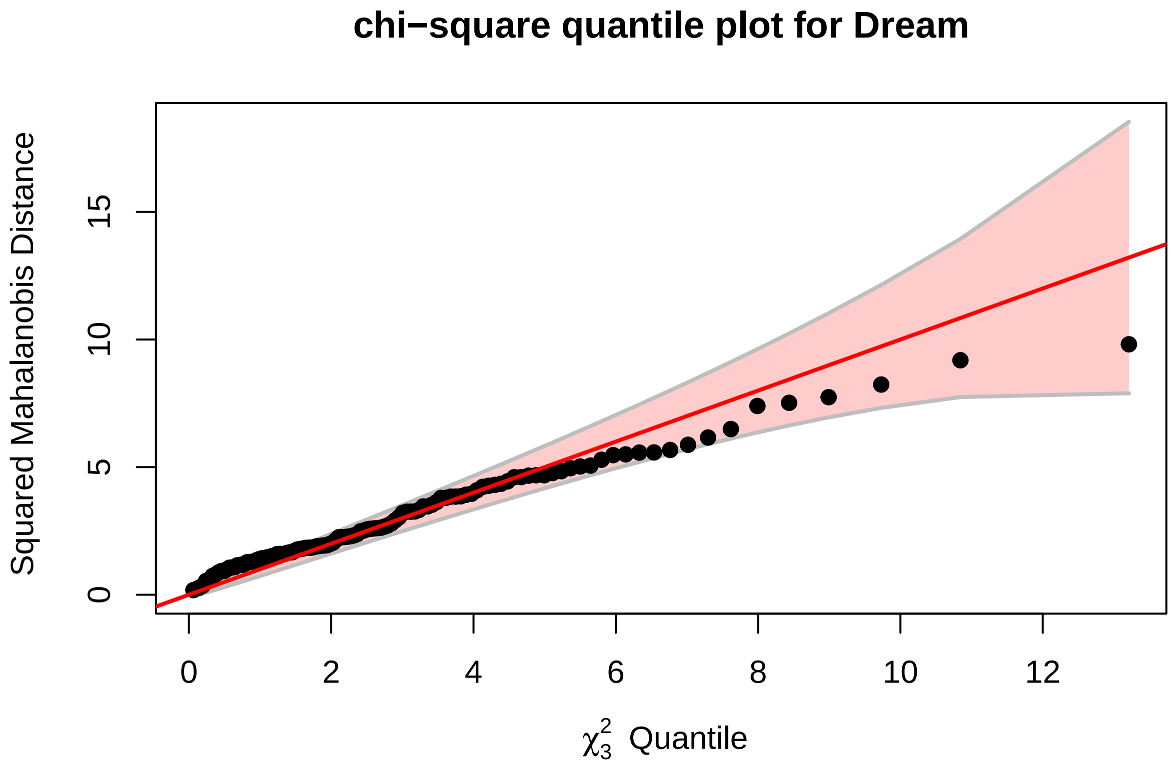
```
##
## Sum of squares and products for the hypothesis:
```

```
##                Culmen_Depth Culmen_Length
## Culmen_Depth       428.1208     -624.6313
## Culmen_Length     -624.6313      911.3415
##
## Sum of squares and products for error:
##                Culmen_Depth Culmen_Length
## Culmen_Depth       770.5315     -294.2176
## Culmen_Length     -294.2176     8397.0359
##
## Multivariate Tests:
##                   Df test stat approx F num Df den Df     Pr(>F)
## Pillai             1 0.3810261 98.49233      2    320 < 2.22e-16 ***
## Wilks              1 0.6189739 98.49233      2    320 < 2.22e-16 ***
## Hotelling-Lawley   1 0.6155770 98.49233      2    320 < 2.22e-16 ***
## Roy                1 0.6155770 98.49233      2    320 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

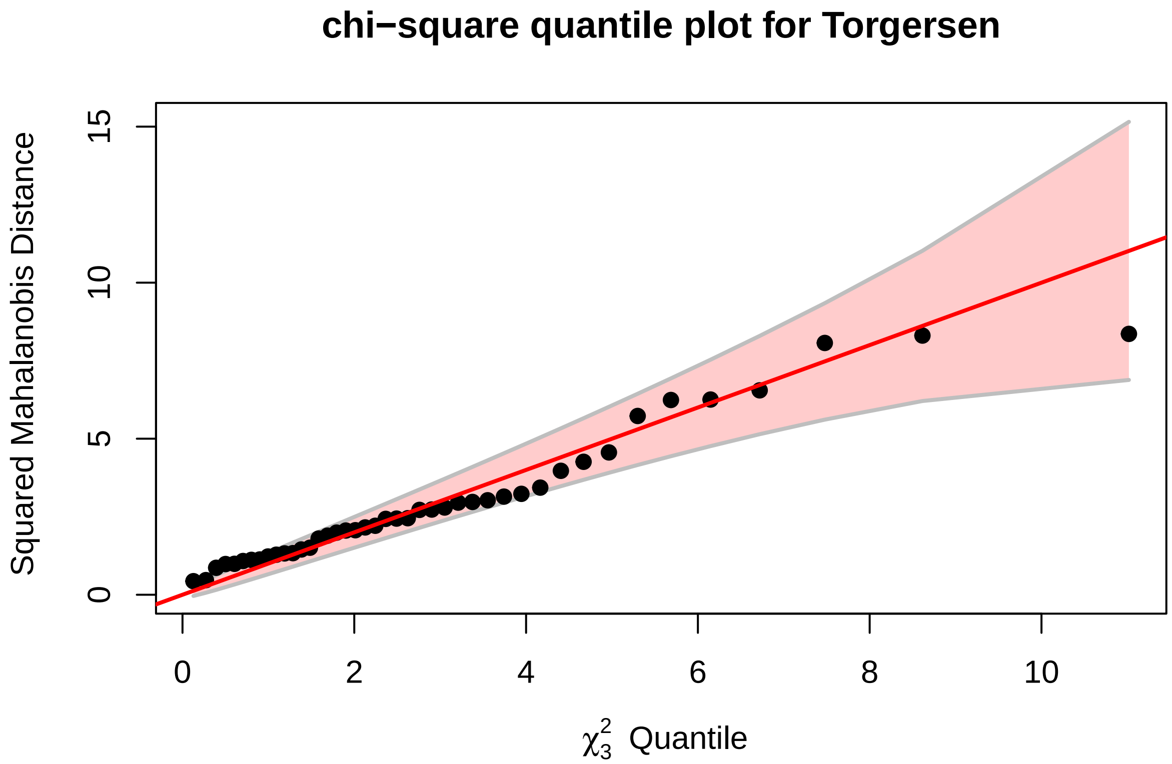
```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Island  == 'Biscoe', c('Culmen_Length','Culmen_Depth','Delta.15.N')], main = "chi-squa
```



```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Island  == 'Dream', c('Culmen_Length','Culmen_Depth','Delta.15.N')], main = "chi-square
```



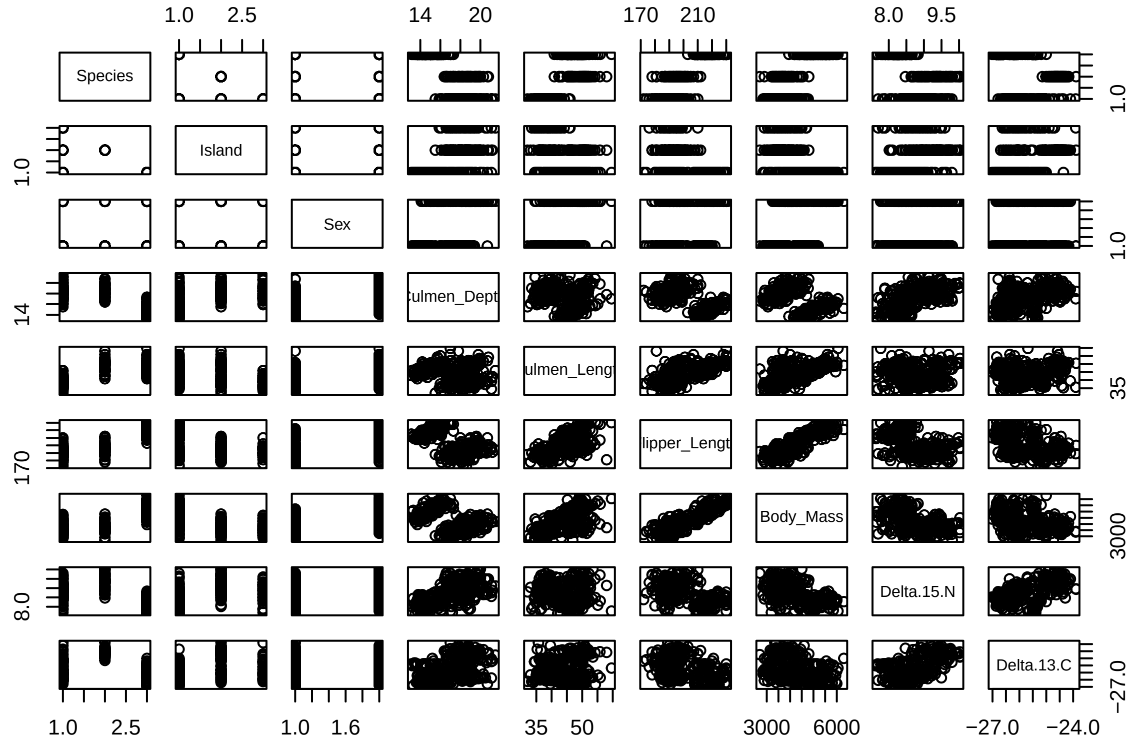
```
# par(mfrow = c(1,3), pty = "s", cex = 0.8)
cqplot(data[data$Island  == 'Torgersen', c('Culmen_Length','Culmen_Depth','Delta.15.N')], main = "chi-sq
```



add a continuous variable to our model and fit as a multiple-response linear model. Before you fit the model, make some plots to see if there are linear relationships between your covariates and your responses. Discuss your multivariate and univariate results.

for this continuous choices: (Flipper_Length, Body_Mass, Delta.13.C)

```
#Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island
pairs(data[,c("Species","Island","Sex", "Culmen_Depth", "Culmen_Length",  "Flipper_Length","Body_Mass",
```



body mass linear increase as flipper_length increase.

```
#SUPER IMPORTANT - include these options below to get correct Type III sum of squares (partial SS)
options(contrasts=c("contr.sum", "contr.poly"))
# #penguinsMAOV <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island*Sex,  data=data)
# two-way MANOVA with interaction
DataMod <- lm(cbind(Delta.15.N, Culmen_Depth, Culmen_Length) ~ Island + Sex + Delta.13.C,
                data=data)
#Multivariate and univariate results
summary(Anova(DataMod, type=3), univariate=T)
```

```
##
## Type III MANOVA Tests:
##
## Sum of squares and products for error:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N       50.57079     53.11866     -97.50087
## Culmen_Depth     53.11866    574.00086    -866.71829
## Culmen_Length   -97.50087   -866.71829    6470.52197
##
## ------------------------------------------
##
## Term: (Intercept)
##
## Sum of squares and products for the hypothesis:
##                Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      37.79504     72.82202      284.2258
## Culmen_Depth    72.82202    140.31065      547.6352
## Culmen_Length  284.22578    547.63524     2137.4312
##
## Multivariate Tests: (Intercept)
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1  0.622522 174.2613      3    317 < 2.22e-16 ***
## Wilks             1  0.377478 174.2613      3    317 < 2.22e-16 ***
## Hotelling-Lawley  1  1.649161 174.2613      3    317 < 2.22e-16 ***
## Roy               1  1.649161 174.2613      3    317 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## ------------------------------------------
##
## Term: Island
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      14.45850     57.20554     -105.7726
## Culmen_Depth    57.20554    283.02773     -643.8336
## Culmen_Length -105.77259   -643.83363     1669.4791
##
## Multivariate Tests: Island
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            2 0.4690862 32.47939      6    636 < 2.22e-16 ***
## Wilks             2 0.5663592 34.74138      6    634 < 2.22e-16 ***
## Hotelling-Lawley  2 0.7030790 37.02883      6    632 < 2.22e-16 ***
## Roy               2 0.5985117 63.44224      3    318 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## ------------------------------------------
##
## Term: Sex
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      1.644092      17.1842      44.00579
## Culmen_Depth   17.184203     179.6109     459.95261
## Culmen_Length  44.005786     459.9526    1177.85942
##
## Multivariate Tests: Sex
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.4710773 94.11049      3    317 < 2.22e-16 ***
## Wilks             1 0.5289227 94.11049      3    317 < 2.22e-16 ***
## Hotelling-Lawley  1 0.8906356 94.11049      3    317 < 2.22e-16 ***
## Roy               1 0.8906356 94.11049      3    317 < 2.22e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## ------------------------------------------
##
## Term: Delta.13.C
```

```
##
## Sum of squares and products for the hypothesis:
##               Delta.15.N Culmen_Depth Culmen_Length
## Delta.15.N      5.157087     9.316916      62.07393
## Culmen_Depth    9.316916    16.832162     112.14423
## Culmen_Length  62.073926   112.144229     747.16061
##
## Multivariate Tests: Delta.13.C
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.2510241 35.41487      3    317 < 2.22e-16 ***
## Wilks             1 0.7489759 35.41487      3    317 < 2.22e-16 ***
## Hotelling-Lawley  1 0.3351565 35.41487      3    317 < 2.22e-16 ***
## Roy               1 0.3351565 35.41487      3    317 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Type III Sums of Squares
##              df Delta.15.N Culmen_Depth Culmen_Length
## (Intercept)   1    37.7950      140.311       2137.43
## Island        2    14.4585      283.028       1669.48
## Sex           1     1.6441      179.611       1177.86
## Delta.13.C    1     5.1571       16.832        747.16
## residuals   319    50.5708      574.001       6470.52
##
##  F-tests
##             Delta.15.N Culmen_Depth Culmen_Length
## (Intercept)     238.41        38.99        105.38
## Island           91.20       157.29         41.15
## Sex              10.37        99.82         58.07
## Delta.13.C       16.27         9.35         36.84
##
##  p-values
##             Delta.15.N Culmen_Depth Culmen_Length
## (Intercept) < 2.22e-16 7.1035e-16   < 2.22e-16
## Island      < 2.22e-16 < 2.22e-16   < 2.22e-16
## Sex         0.0014121  < 2.22e-16   2.9126e-13
## Delta.13.C  1.8769e-07 0.0024129    3.6438e-09
```

After including Delta.13.C as a covariate in the two-way MANOVA model, the results indicate that all predictors — Island, Sex, and now Delta.13.C — have statistically significant multivariate effects on the response variables: Delta.15.N, Culmen_Depth, and Culmen_Length.

Multivariate Results (MANOVA) Island continues to show a strong multivariate effect, suggesting substantial morphological and isotopic differences among islands.

Sex also remains a highly significant predictor, reinforcing evidence of sexual dimorphism.

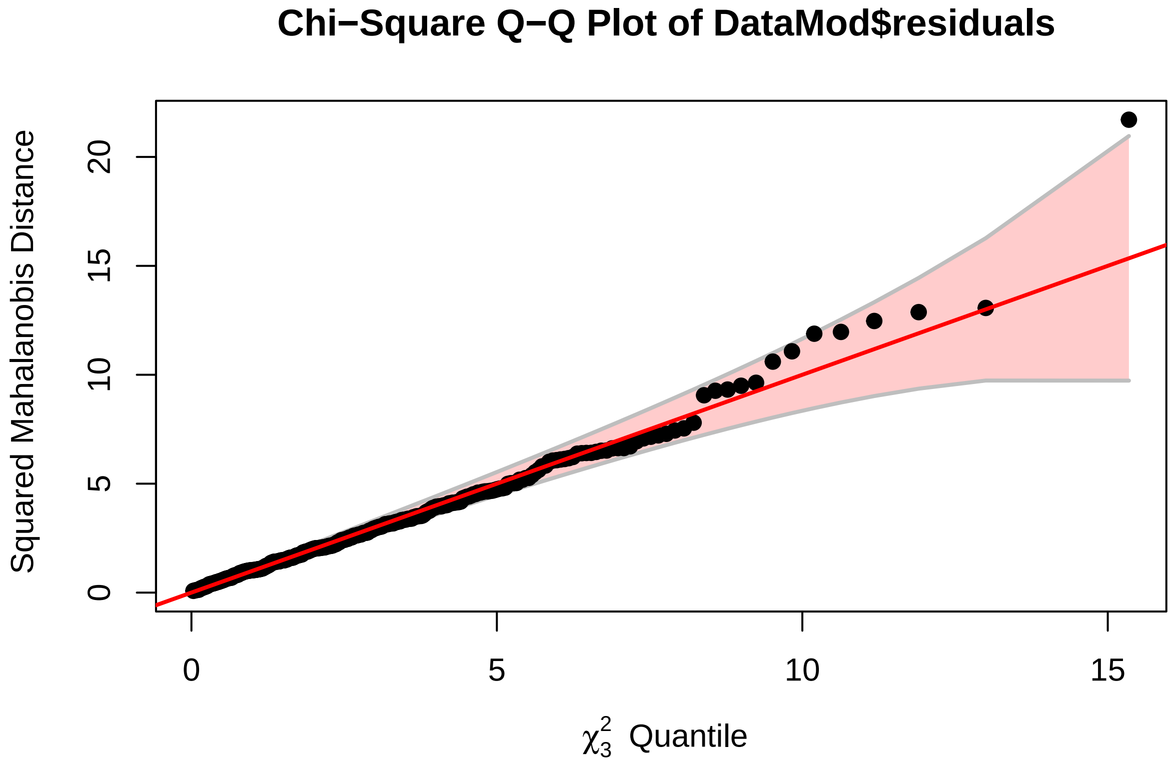
Delta.13.C is newly added and shows a strong multivariate association with the traits (Pillai = 0.25, F = 35.41, p < 2.22e-16), suggesting that carbon isotope ratios are significantly related to penguin diet and morphology.

The inclusion of Delta.13.C in the model enhances explanatory power by capturing additional ecological variation (e.g., foraging habitat). It remains statistically significant even after accounting for Island and Sex, indicating that carbon isotope ratios provides independent information about variation in penguin traits. Meanwhile, Island and Sex remain robust predictors, with strong effects on both isotopic and morphological traits.

These results highlight the multifactorial nature of trait variation in penguins, shaped by geography (Island), biology (Sex), and ecology (carbon isotope ratios diet signatures).

Check model assumptions by making a chi-square quantile plot of the residuals.

```
# let's check our residuals
cqplot(DataMod$residuals, label="Residuals from penguins")
```



Based on the chi-square quantile plot of the residuals above, the original residuals from the multiple-response linear model seems follows the multivariate normality which ensures that the assumption for MANOVA is hold.

run MRPP tests on some form of your data

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
## The following objects are masked _by_ '.GlobalEnv':
##
##     parallel, parallelplot

## Registered S3 methods overwritten by 'vegan':
##   method      from
##   plot.rda    klaR
##   predict.rda klaR
##   print.rda   klaR

##
## Attaching package: 'vegan'

## The following object is masked from 'package:klaR':
##
##     rda
```

```
set.seed(123)
continuous_vars <- data[, c("Culmen_Length",
                   "Culmen_Depth",
                   "Flipper_Length",
                   "Body_Mass",
                   "Delta.15.N",
                   "Delta.13.C")]
```

```
mrpp1 <- mrpp(continuous_vars, data$Island)
print(mrpp1)
```

```
##
## Call:
## mrpp(dat = continuous_vars, grouping = data$Island)
##
## Dissimilarity index: euclidean
## Weights for groups:  n
##
## Class means and counts:
##
##       Biscoe Dream Torgersen
## delta 901.5  464.6 537.3
## n     162    119   43
##
## Chance corrected within-group agreement A: 0.2482
## Based on observed delta 692.7 and expected delta 921.4
##
## Significance of delta: 0.001
## Permutation: free
## Number of permutations: 999
```

mrpp1: All variables (morphology + isotopes) $A = 0.2482$, $p = 0.001$

Interpretation: There is strong evidence that penguins differ multivariately by island in terms of morphology and isotopic profile.

Largest effect size of the three tests — shows that combining all traits leads to the strongest island-level differentiation.

```
mrpp2 <- mrpp(continuous_vars[,1:2], data$Island)
print(mrpp2)
```

```
##
## Call:
## mrpp(dat = continuous_vars[, 1:2], grouping = data$Island)
##
## Dissimilarity index: euclidean
## Weights for groups:  n
##
## Class means and counts:
##
##       Biscoe Dream Torgersen
## delta 6.086  7.171 4.161
## n     162    119   43
##
## Chance corrected within-group agreement A: 0.1162
## Based on observed delta 6.229 and expected delta 7.048
##
## Significance of delta: 0.001
## Permutation: free
## Number of permutations: 999
```

mrpp2: Only bill traits (Culmen_Length, Culmen_Depth) A = 0.1162, p = 0.001 Interpretation: Bill morphology differs significantly by island, but the effect size is smaller than for all traits combined. Conclusion: Culmen traits contribute to group separation but do not fully explain it.

```
mrpp3 <- mrpp(continuous_vars[,5:6], data$Island)
print(mrpp3)
```

```
##
## Call:
## mrpp(dat = continuous_vars[, 5:6], grouping = data$Island)
##
## Dissimilarity index: euclidean
## Weights for groups:  n
##
## Class means and counts:
##
##       Biscoe Dream Torgersen
## delta 0.8336 1.025 0.9534
## n     162    119   43
##
## Chance corrected within-group agreement A: 0.2244
## Based on observed delta 0.9197 and expected delta 1.186
##
## Significance of delta: 0.001
## Permutation: free
## Number of permutations: 999
```

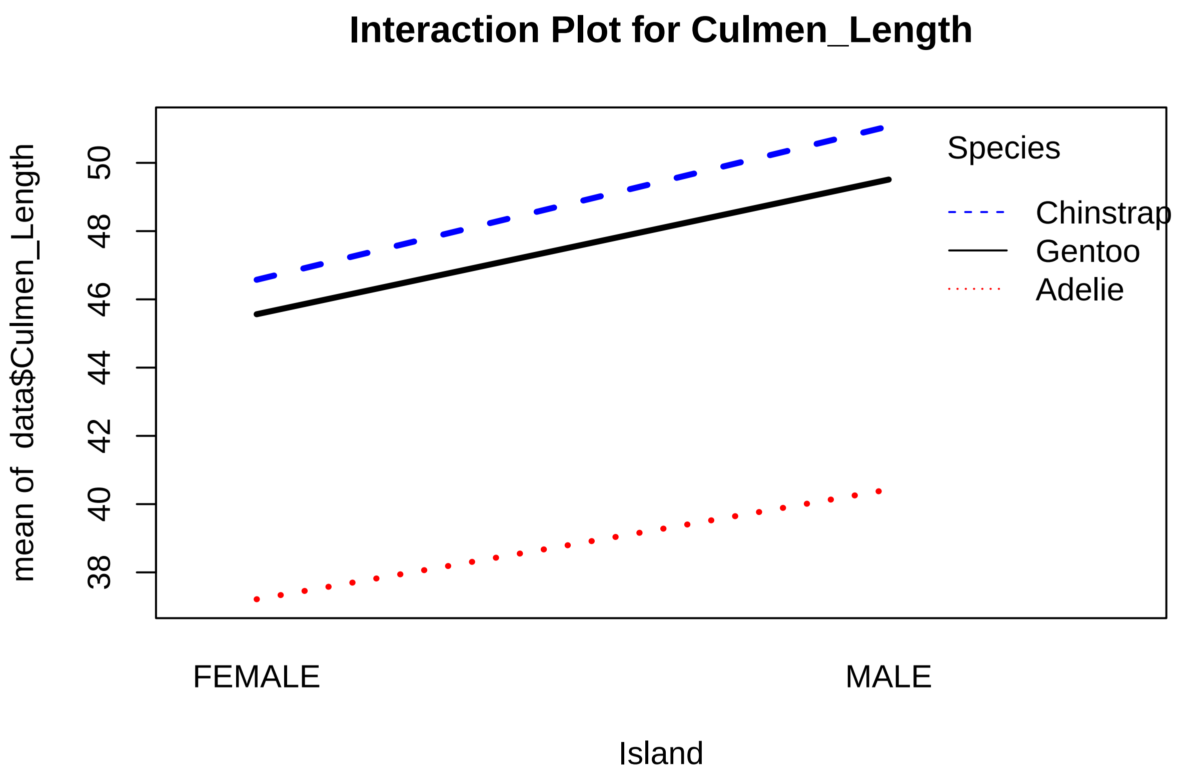
mrpp3: Isotopic variables (Delta.15.N, Delta.13.C) A = 0.2244, p = 0.001

Interpretation: Isotopic values differ significantly by island, with a relatively strong effect. This suggests dietary and trophic variation across habitats is a key driver of group separation.

```
mrpp4 <- mrpp(continuous_vars[,5:6], data$Species)
print(mrpp4)
```

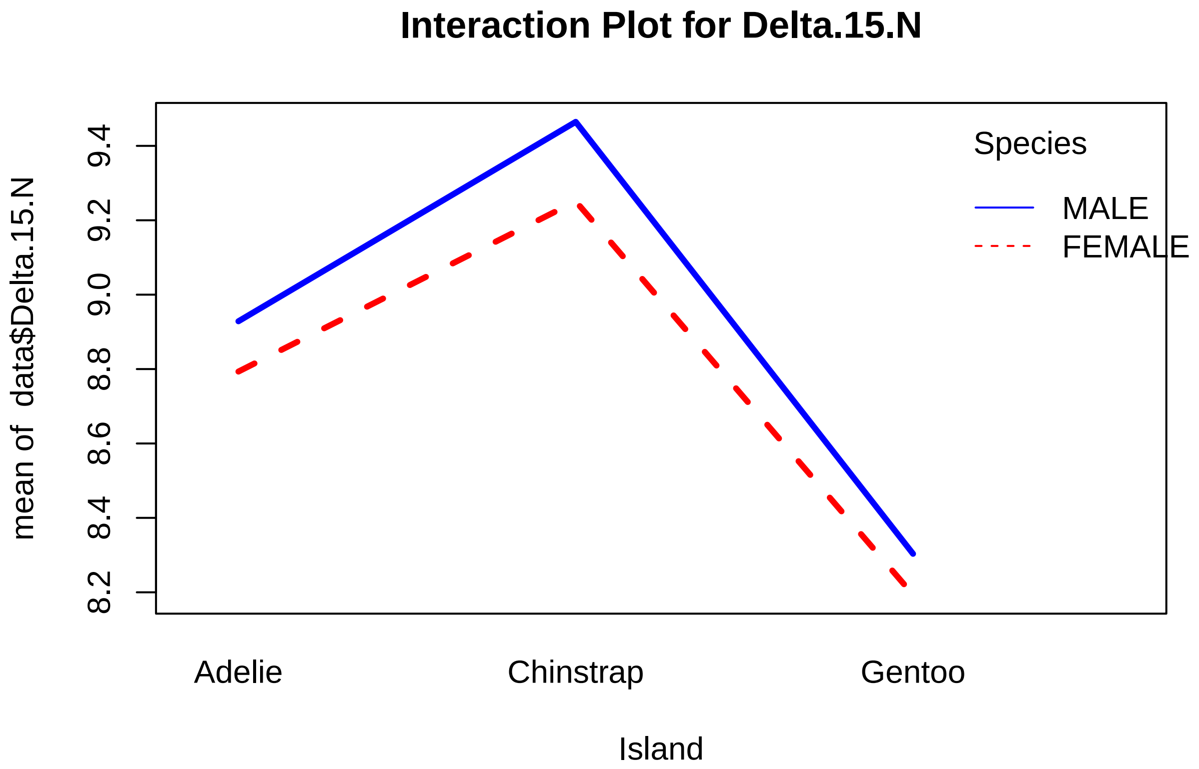
```
##
## Call:
## mrpp(dat = continuous_vars[, 5:6], grouping = data$Species)
##
## Dissimilarity index: euclidean
## Weights for groups:  n
##
## Class means and counts:
##
##       Adelie Chinstrap Gentoo
## delta 0.8865 0.5327    0.7301
## n     139    67        118
##
## Chance corrected within-group agreement A: 0.3621
## Based on observed delta 0.7564 and expected delta 1.186
##
## Significance of delta: 0.001
## Permutation: free
## Number of permutations: 999
```

```
interaction.plot(data$Sex, data$Species, data$Culmen_Length,
  lwd=3, col=c("red", "blue", "black"), trace.label="Species",
  xlab="Island", main="Interaction Plot for Culmen_Length")
```

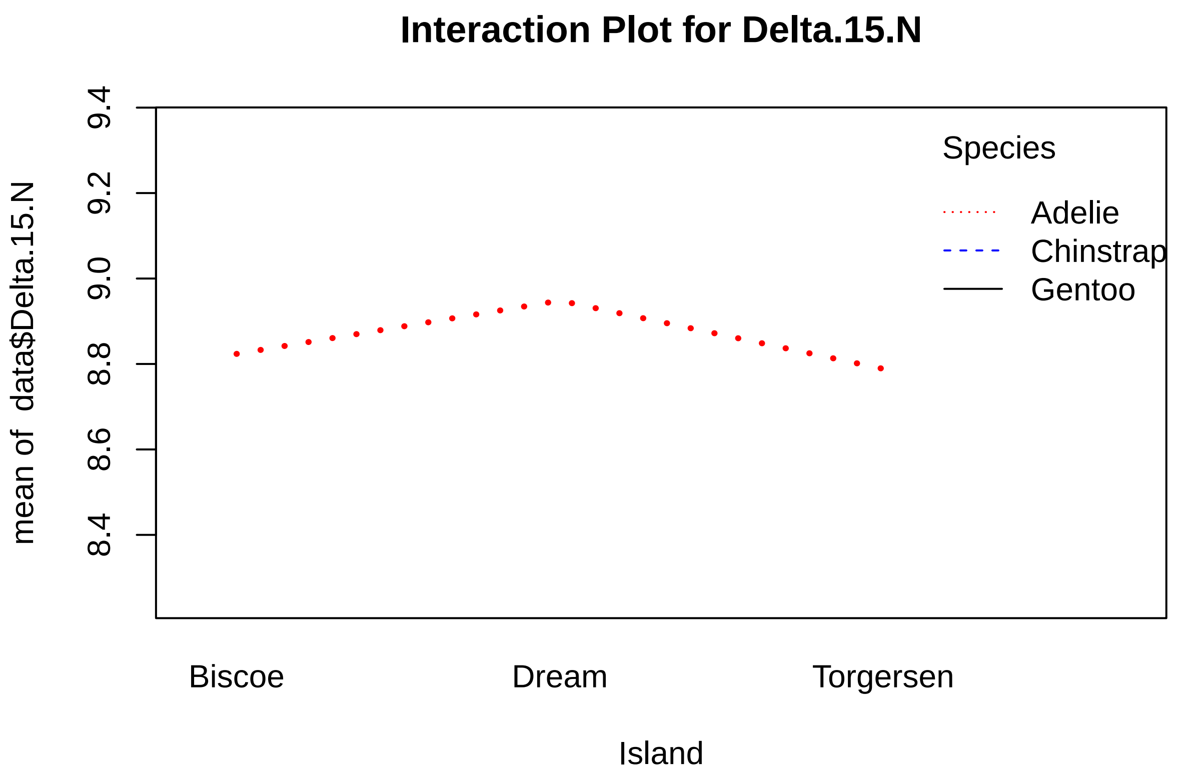


# Delta.15.N, Culmen_Depth, Culmen_Length

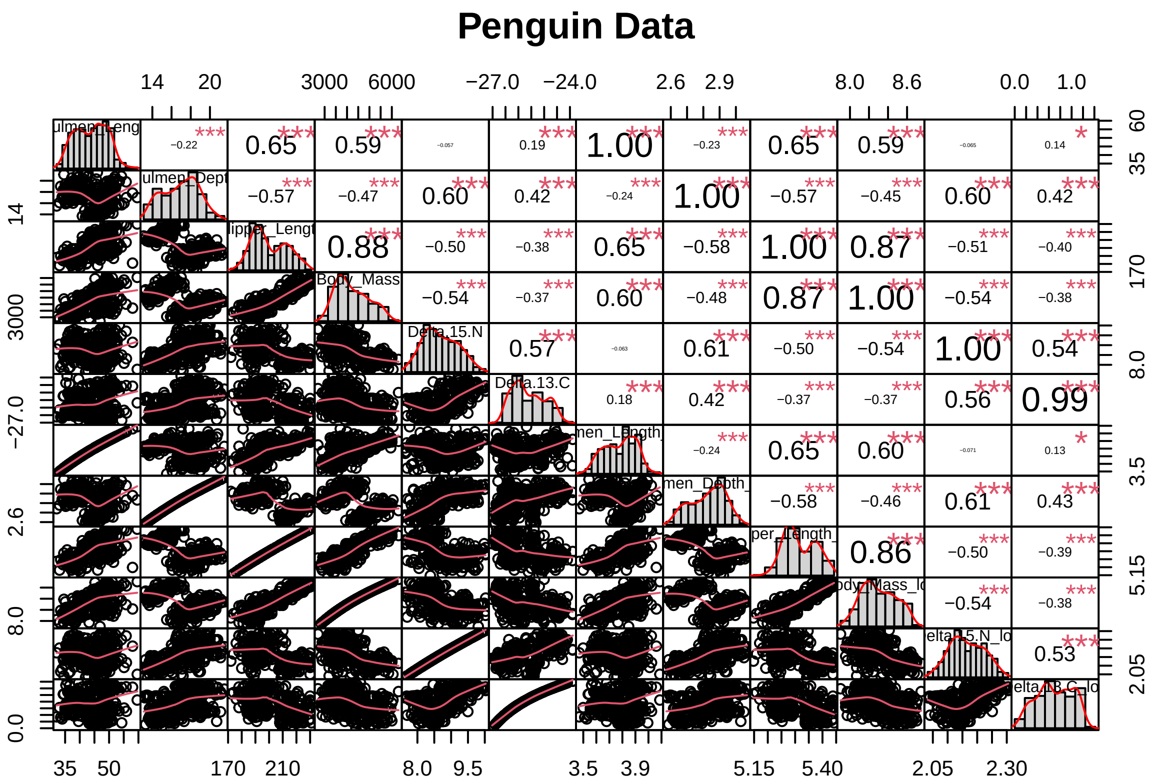
```
interaction.plot(data$Species, data$Sex, data$Delta.15.N,
  lwd=3, col=c("red", "blue", "black"), trace.label="Species",
  xlab="Island", main="Interaction Plot for Delta.15.N")
```



```
interaction.plot(data$Island, data$Species, data$Delta.15.N,
  lwd=3, col=c("red", "blue", "black"), trace.label="Species",
  xlab="Island", main="Interaction Plot for Delta.15.N")
```



```
chart.Correlation(df_PCA[,-c(1,2,3)], main = "Penguin Data")
```



Cluster Analysis (quite similar result which has been shown by PCA and DA, so NO)