다변량 분석 및 실습 Group project

새의 뼈와 생태학적 분류

3조 – 다알조

2129006 김경민 | 2129036 차수빈 | 2135019 장단 | 216247 이도경

1) 주제 및 데이터 소개

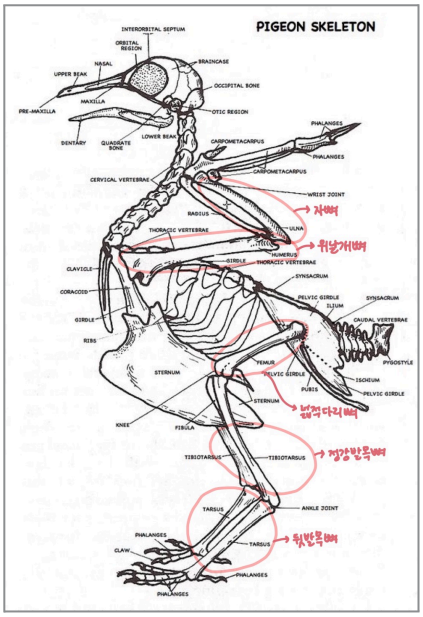
작성자: 장단

I. 주제

본 연구에서는 새의 뼈의 길이 및 지름에 대한 정보를 바탕으로 새의 생태학적 그룹을 분류한 데이터의 변수 분포를 살펴보고 해당 변수들의 특징을 탐색하여 다양한 분석을 시행하고자 한다. 데이터는 11개의 연속형 변수와 1개의 범주형 변수로 구성되며, 총 420개의 데이터 건수를 가지고 있다. 결측치가 거의 존재하지 않고, 종속변수 중 범주형 변수가 없는 6-class classification 문제이기에, 너무 단순하지도 복잡하지도 않은 데이터셋이라고 생각되어 해당 데이터를 선정하였다. 또한, 하나의 뼈에 대해 길이(length)와 지름(diameter) 속성이 명시적으로 구분되기에 정준상관분석 수행에 적합한 데이터라 판단하였다.

II. 데이터 설명

해당 연구에서 사용한 <Birds' Bones and Living Habits> 데이터셋은 생활 환경 및 생활 습관에 따라 8가지 생태학적 그룹으로 분류된 새들의 뼈 길이와 뼈 지름의 측정값(mm)을 담고 있다. 420개의 관측치(새)가 존재하며, 각 관측치는 아래의 10개의 측정값(특징)으로 표현된다.



Id를 제외한 10개의 피처는 다음과 같다.

- huml / humw: 위날개뼈 길이/ 지름
- ulnal/ ulnaw: 자뼈 길이/ 지름
- feml/ femw: 넓적다리뼈 길이/ 지름
- tibl/ tibw: 정강발목뼈 길이/ 지름
- tarl/ tarw: 뒷발목뼈 길이/ 지름

모든 측정값(mm)은 수치형 변수이며, 결측치는 모두 공백으로 표시되어 있다.

각 새는 해당하는 생태학적 그룹에 대한 레이블(Type of birds)을 가지고 있는데, 이는 데이터셋의 target 변수에 해당한다.

- SW: Swimming Birds (수영하는 새) → 오리, 백조, 펭귄 등
- W: Wading Birds (물가의 새) → 왜가리, 두루미, 해오라기 등
- T: Terrestrial Birds (지상조류) → 참새, 비둘기, 닭 등
- R: Raptors (맹금류) → 독수리, 매, 올빼미 등
- P: Scansorial Birds (나무를 오르는 새) → 딱따구리 등
- SO: Singing Birds (노래하는 새) → 지빠귀, 종달새 등

총 8가지의 생태학적 그룹 중 위의 6가지 그룹을 제외한 나머지 2가지 그룹(Cursorial Birds: 주행성 새, Marine Birds: 해양 새)은 데이터셋에 포함되지 않았다.

서로 다른 생태학적 그룹에 속하는 새들은 서로 다른 생태학적 특성을 지닌다. 예를 들어 비행하는 새들은 강한 날개를 가지고 있고, 물가 새들은 긴 다리를 가지고 있다. 이를 통해, 그들의 생활 습관은 어느 정도 그들의 뼈 모양에 반영되어 있다고 유추할 수 있다. 따라서 본 연구를 통해 뼈 크기와 생태학적 그룹 간의 근본적인 관계를 조사하고, 뼈 모양으로 새의 생태학적 그룹을 인식하는 분석을 수행할 수 있을 것이라 예상한다.

2) 데이터 전처리

작성자: 장단

I. 데이터 확인

1. 데이터 불러오기

```
bird <- read.csv("C:/Temp/bird.csv")
```

2. 일부 데이터 확인하기

head() 함수를 통해, 처음 몇 개 행의 데이터를 확인할 수 있다.

```
head(bird)
```

```
##   id  huml humw ulnal ulnaw  feml femw  tibl tibw  tarl tarw type
## 1  0 80.78 6.68 72.01  4.88 41.81 3.70  5.50 4.03 38.70 3.84   SW
## 2  1 88.91 6.63 80.53  5.59 47.04 4.30 80.22 4.51 41.50 4.01   SW
## 3  2 79.97 6.37 69.26  5.28 43.07 3.90 75.35 4.04 38.31 3.34   SW
## 4  3 77.65 5.70 65.76  4.77 40.04 3.52 69.17 3.40 35.78 3.41   SW
## 5  4 62.80 4.84 52.09  3.73 33.95 2.72 56.27 2.96 31.88 3.13   SW
## 6  5 61.92 4.78 50.46  3.47 49.52 4.41 56.95 2.73 29.07 2.83   SW
```

3. 데이터 형태 확인

```
# 차원
dim(bird)
```

```
## [1] 420  12
```

420개의 행, 12개의 변수로 구성되어 있다.

```
# 데이터셋 구조 확인
str(bird)
```

```
## 'data.frame':    420 obs. of  12 variables:
##  $ id   : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ huml : num  80.8 88.9 80 77.7 62.8 ...
##  $ humw : num  6.68 6.63 6.37 5.7 4.84 ...
##  $ ulnal: num  72 80.5 69.3 65.8 52.1 ...
##  $ ulnaw: num  4.88 5.59 5.28 4.77 3.73 3.47 4.5 4.55 6.13 7.05 ...
##  $ feml : num  41.8 47 43.1 40 34 ...
##  $ femw : num  3.7 4.3 3.9 3.52 2.72 4.41 3.41 3.78 5.45 7.44 ...
##  $ tibl : num  5.5 80.2 75.3 69.2 56.3 ...
##  $ tibw : num  4.03 4.51 4.04 3.4 2.96 2.73 3.56 3.81 5.58 7.31 ...
##  $ tarl : num  38.7 41.5 38.3 35.8 31.9 ...
##  $ tarw : num  3.84 4.01 3.34 3.41 3.13 2.83 3.64 3.81 4.37 6.34 ...
##  $ type : chr  "SW" "SW" "SW" "SW" ...
```

id 변수 외 10개의 feature 변수, 1개의 target 변수로 구성되어 있음을 확인할 수 있다. 또한, feature 변수의 경우 5개 뼈(위날개뼈, 자뼈, 넓적다리뼈, 정강발목뼈, 뒷발목뼈)에 대한 길이와 지름에 대한 속성으로 구성되어 있음을 확인할 수 있다.

id 삭제

id 변수는 분석에 활용되지 않기에, 삭제하기로 결정하였다.

```
bird <- bird[,-1]
```

데이터 요약 정보 확인

```
summary(bird[,-11])
```

```
##       huml            humw            ulnal           ulnaw      
##  Min.   :  9.85   Min.   : 1.140   Min.   : 14.09   Min.   : 1.000  
##  1st Qu.: 25.17   1st Qu.: 2.190   1st Qu.: 28.05   1st Qu.: 1.870  
##  Median : 44.18   Median : 3.500   Median : 43.71   Median : 2.945  
##  Mean   : 64.65   Mean   : 4.371   Mean   : 69.12   Mean   : 3.597  
##  3rd Qu.: 90.31   3rd Qu.: 5.810   3rd Qu.: 97.52   3rd Qu.: 4.770  
##  Max.   :420.00   Max.   :17.840   Max.   :422.00   Max.   :12.000  
##  NA's   :1        NA's   :1        NA's   :3        NA's   :2       
##       feml            femw            tibl            tibw      
##  Min.   : 11.83   Min.   : 0.930   Min.   :  5.50   Min.   : 0.870  
##  1st Qu.: 21.30   1st Qu.: 1.715   1st Qu.: 36.42   1st Qu.: 1.565  
##  Median : 31.13   Median : 2.520   Median : 52.12   Median : 2.490  
##  Mean   : 36.87   Mean   : 3.221   Mean   : 64.66   Mean   : 3.182  
##  3rd Qu.: 47.12   3rd Qu.: 4.135   3rd Qu.: 82.87   3rd Qu.: 4.255  
##  Max.   :117.07   Max.   :11.640   Max.   :240.00   Max.   :11.030  
##  NA's   :2        NA's   :1        NA's   :2        NA's   :1       
##       tarl            tarw       
##  Min.   :  7.77   Min.   : 0.660  
##  1st Qu.: 23.04   1st Qu.: 1.425  
##  Median : 31.74   Median : 2.230  
##  Mean   : 39.23   Mean   : 2.930  
##  3rd Qu.: 50.25   3rd Qu.: 3.500  
##  Max.   :175.00   Max.   :14.090  
##  NA's   :1        NA's   :1
```

데이터에 일부 결측치가 존재함을 확인할 수 있다. 또한, 대부분 오른쪽 꼬리가 긴(왼쪽으로 치우친), 왜곡된 분포를 지님을 확인할 수 있다.

II. 결측치 처리

1. 결측치 확인

```
# 전체 데이터에서 결측치가 몇 건인지 확인
sum(is.na(bird))
```

```
## [1] 15
```

```
# 각 변수별로 결측치가 얼마나 존재하는지 확인
colSums(is.na(bird))
```

```
##  huml  humw ulnal ulnaw  feml  femw  tibl  tibw  tarl  tarw  type
##     1     1     3     2     2     1     2     1     1     1     0
```

```
bird[!complete.cases(bird),]
```

```
##      huml humw  ulnal ulnaw  feml femw   tibl tibw  tarl tarw type
## 161 76.43 4.11  86.79  3.84    NA   NA  67.13 2.48 41.65 2.10    W
## 205 63.76 4.74     NA    NA 57.33 4.88  75.67 4.33 60.19 3.82    R
## 208 98.08 7.77 113.04  5.76 82.04 7.17 107.47 6.65    NA   NA    R
## 343    NA   NA     NA    NA 32.54 2.65  55.06 2.81 38.94 2.25   SO
## 379 20.10 1.86     NA  1.52 17.21 1.22     NA   NA 18.46 0.91   SO
## 397 16.51 1.47  20.56  1.43 15.88 1.27     NA 1.19 17.63 1.02   SO
## 405 20.36 1.87  22.19  1.60    NA 1.77  37.47 1.64 25.54 1.34   SO
```

총 15개의 결측치가 존재하고, 7개의 데이터에 대해 결측인 부분이 15개이다.

2. 결측치 처리

type(어떤 생태학적 특성을 가지는 조류인지)에 대한 정보는 모든 데이터에 대해 제공된 상황이다. 따라서, 각 type의 기술통계량을 활용하여 결측치를 처리하기로 결정하였다. 또한 대부분의 변수가 왜곡된 분포를 보이고 있기에, 중앙값으로 결측치를 보간하기로 결정하였다.

```
bird_grouped <- split(bird, bird$type)

# type별로 데이터프레임 그룹화하여 결측치를 중앙값으로 보간
for (type in names(bird_grouped)) {
  group_data <- bird_grouped[[type]]
  for (col in names(group_data)[!names(group_data) %in% "type"]) {
    group_data[[col]][is.na(group_data[[col]])] <- median(group_data[[col]],
na.rm = T)
  }
  bird[bird$type == type, ] <- group_data
}

sum(is.na(bird))
```

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

III. 이상치 처리

1. 이상치 탐지 함수

이상치 탐지를 위해, IQR(Inter Quartile Range)을 활용하여 데이터의 존재 범위를 지정하는 함수를 작성하였다.

```
### 이상치가 있는 데이터의 인덱스를 반환

get_outlier_indices <- function(df_col, weight = 1.5) {
  q <- quantile(df_col, c(0.25, 0.75))
  iqr <- q[2] - q[1]

  lowest_val <- q[1] - iqr * weight
  highest_val <- q[2] + iqr * weight

  outlier_indices <- which(df_col < lowest_val | df_col > highest_val)

  return(outlier_indices)
}
```

2. 이상치 탐지

이상치 탐지 시, weight를 3으로 설정하였다.

- 최댓값: $Q_3 + IQR * 3$
- 최솟값: $Q_1 - IQR * 3$

```
## type별로 분리
# 위에서 결측치 보간을 수행하였기에 보간된 데이터로 다시 분할해 준다.

bird_grouped <- split(bird, bird$type)

# 각 type별로 이상치 비율 저장할 리스트 생성
outlier_ratios <- list()
# 이상치 index를 저장할 리스트 생성
outlier_idx <- list()

# 각 type별로 이상치 탐지 및 제거
for (type in names(bird_grouped)) {
  group_data <- bird_grouped[[type]]

  # 각 변수에 대해 이상치 탐지
  outliers <- list()
  for (col in names(group_data)) {
    if (col != "type") {
      outliers[[col]] <- get_outlier_indices(group_data[[col]], weight = 3)
    }
  }

  # 이상치 인덱스를 하나의 리스트에 모음
  outlier_idx[[type]] <- unique(unlist(outliers))

  # 이상치 비율 계산
  total_rows <- nrow(group_data)
  total_outliers <- length(outlier_idx[[type]])
  outlier_ratio <- total_outliers / total_rows
  outlier_ratios[[type]] <- outlier_ratio

  cat("Type:", type, "- 이상치 비율:", outlier_ratio, "\n")
}
```

```
## Type: P - 이상치 비율: 0.05263158
## Type: R - 이상치 비율: 0
## Type: SO - 이상치 비율: 0.0078125
## Type: SW - 이상치 비율: 0.00862069
## Type: T - 이상치 비율: 0.173913
## Type: W - 이상치 비율: 0.03076923
```

확인 결과, 대부분 group에서 이상치는 거의 존재하지 않는 것을 확인할 수 있다. Type T의 경우 해당하는 데이터의 건수 자체가 적기에, 다른 type에 비해 이상치 비율이 높게 측정되었다고 볼 수 있다.

3. 이상치 제거

이상치가 포함된 경우 이는 유의미한 분석에 악영향을 줄 가능성이 존재하기에 적절한 처리가 필요하다. 대부분의 경우 이상치의 비율이 작기에 이상치를 삭제하는 방법을 선택하였다. 다만 `bird_type == T`인 경우 해당하는 데이터 건수도 매우 적고, 이상치의 비율 또한 높게 나타났다. 이는 해당 범주를 구분하는 유의미한 데이터 분석에 방해가 될 것이라 생각되어, 일단 분석 시 제외하였다.

```
# type == T 제거
bird <- bird[-which(bird$type == "T"), ]

# 이상치 데이터의 인덱스
all_outliers <- unique(unlist(outlier_idx))

# 이상치 제거
bird <- bird[-all_outliers, ]

dim(bird)

## [1] 387  11
```

데이터가 420개에서 387개로 감소하였다.

IV. 데이터 저장

전처리가 완료된 데이터를 별도의 파일로 내보냈다.

```
write.csv(bird, "C:/Temp/bird_preprocessed.csv", row.names = FALSE)
```

3) 탐색적 데이터 분석(EDA)

작성자: 장단

I. 라이브러리 및 데이터 불러오기

```
bird = read.csv("C:/Temp/bird_preprocessed.csv")
```

II. 데이터 구조 확인

```
head(bird)
```

```
##     huml  humw  ulnal ulnaw  feml femw   tibl tibw  tarl tarw type
## 1  79.97  6.37  69.26  5.28 43.07 3.90  75.35 4.04 38.31 3.34   SW
## 2  77.65  5.70  65.76  4.77 40.04 3.52  69.17 3.40 35.78 3.41   SW
## 3  79.73  5.94  67.39  4.50 42.07 3.41  71.26 3.56 37.22 3.64   SW
## 4  86.98  5.68  74.52  4.55 44.46 3.78  76.02 3.81 37.94 3.81   SW
## 5 118.20  7.82 116.64  6.13 59.33 5.45 110.00 5.58 61.62 4.37   SW
## 6 145.00 10.42 144.00  7.05 70.96 7.44 120.00 7.31 78.67 6.34   SW
```

```
dim(bird)
```

```
## [1] 387  11
```

```
str(bird)
```

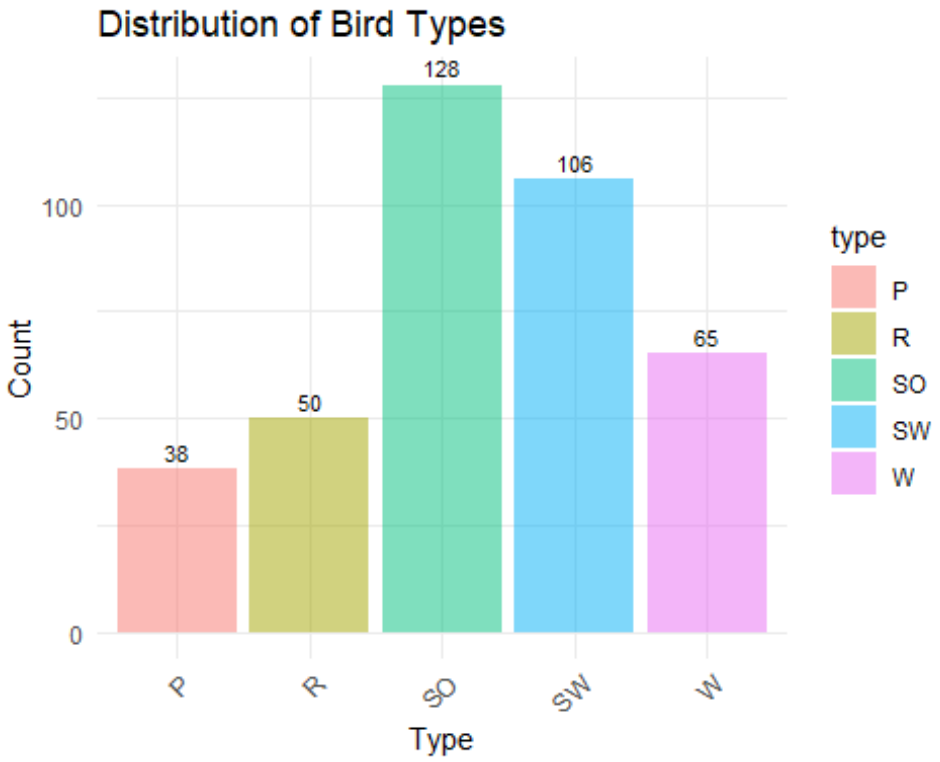
```
## 'data.frame':    387 obs. of  11 variables:
##  $ huml : num  80 77.7 79.7 87 118.2 ...
##  $ humw : num  6.37 5.7 5.94 5.68 7.82 ...
##  $ ulnal: num  69.3 65.8 67.4 74.5 116.6 ...
##  $ ulnaw: num  5.28 4.77 4.5 4.55 6.13 7.05 8.68 8.76 8.43 6.5 ...
##  $ feml : num  43.1 40 42.1 44.5 59.3 ...
##  $ femw : num  3.9 3.52 3.41 3.78 5.45 7.44 7.85 7.02 6.68 6.33 ...
##  $ tibl : num  75.3 69.2 71.3 76 110 ...
##  $ tibw : num  4.04 3.4 3.56 3.81 5.58 7.31 8.25 8.07 9.62 6.68 ...
##  $ tarl : num  38.3 35.8 37.2 37.9 61.6 ...
##  $ tarw : num  3.34 3.41 3.64 3.81 4.37 6.34 6.63 4.59 5.5 4.24 ...
##  $ type : chr  "SW" "SW" "SW" "SW" ...
```

III. EDA 수행

1. 'type'의 분포 확인

분석: 차수빈, 김경민, 이도경

```
# 시각화를 위해 ggplot2 로드
library(ggplot2)


# type 변수의 분포를 시각화하고 막대 위에 count 표시
ggplot(bird, aes(x = type, fill = type)) +
  geom_bar(alpha = 0.5) +
  geom_text(stat = 'count', aes(label = after_stat(count)),
            vjust=-0.5, size = 3) +
  labs(title = "Distribution of Bird Types", x = "Type", y = "Count") +
theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



SO(노래하는 조류)와 SW(수영하는 조류)가 전체 데이터의 약 60% 정도를 차지하며 P(산악지대 조류)가 가장 적게 관측되었다.

2. 독립변수들의 분포 확인(1)

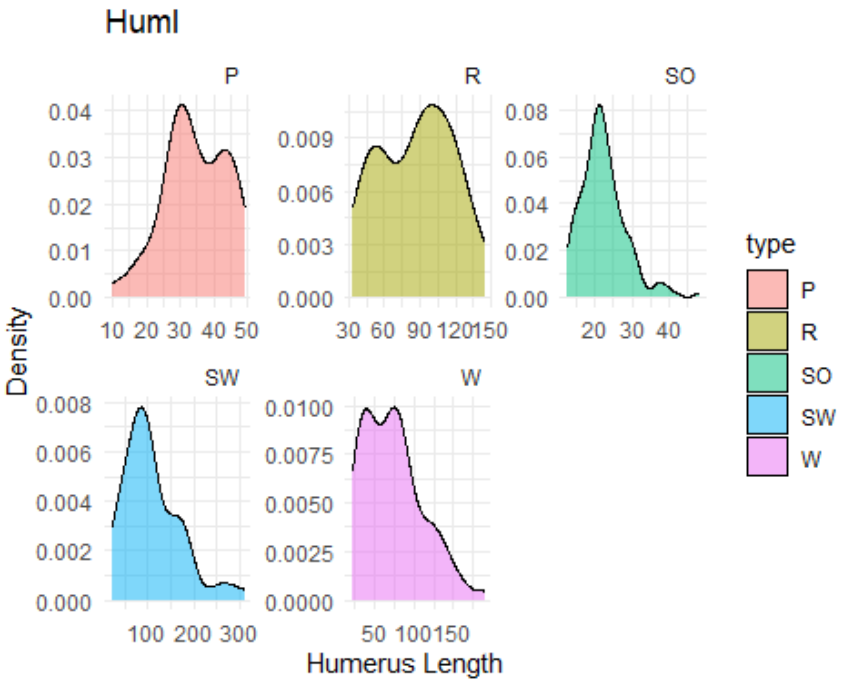
분석: 차수빈, 김경민

먼저 density plot을 통해 type에 따른 개별 변수의 분포를 확인한 후, type별로 뼈의 길이(length)와 지름(width)의 관계를 파악하였다.

2-1. humerus(위날개뼈)

길이

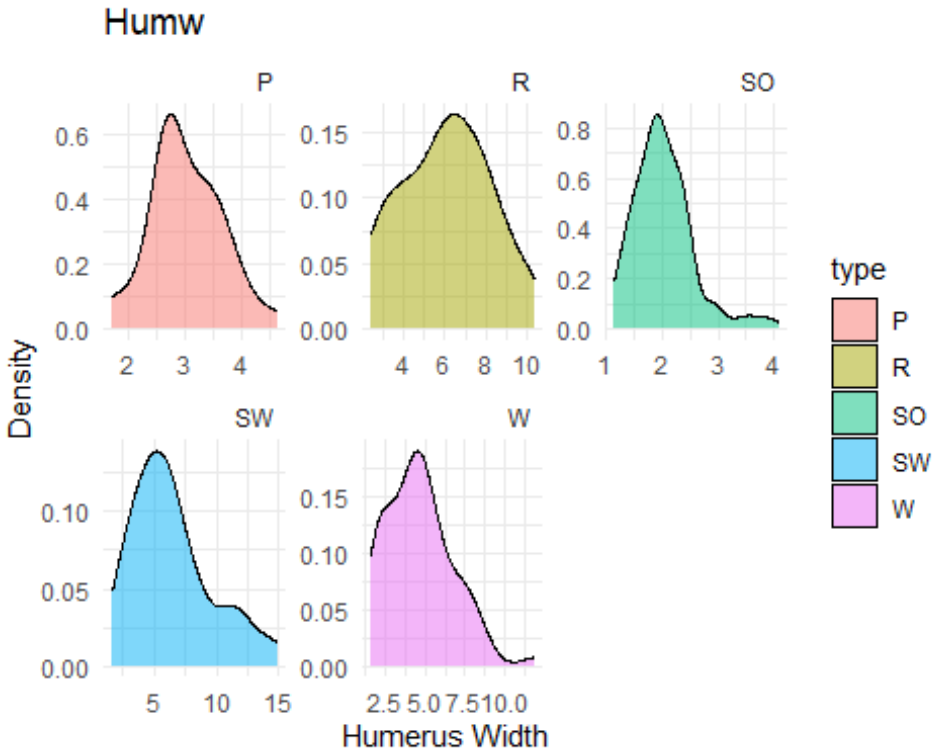
```
# type별 huml의 분포 시각화
ggplot(bird, aes(x = huml, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Huml", x = "Humerus Length", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



type별로 위날개뼈 길이의 분포가 다르며, 위날개뼈 길이의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. SO(노래하는 조류), P(산악지대에 서식하는 조류)의 경우 다른 type에 비해 위날개뼈의 길이가 짧고, SW(수영하는 조류)의 경우 다른 type에 비해 위날개뼈의 길이가 길다. 해당 그룹을 제외한 나머지 type은 비슷한 길이 범위를 보임을 확인할 수 있다.

지름

```r
# type별 humw의 분포 시각화
ggplot(bird, aes(x = humw, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Humw", x = "Humerus Width", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



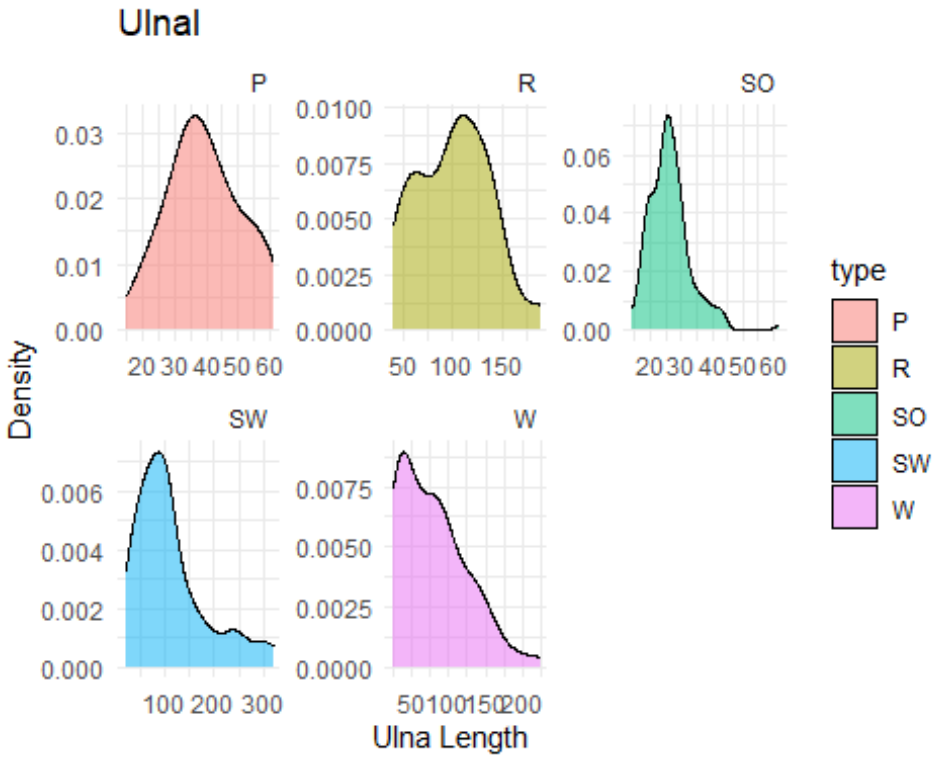
type별로 위날개뼈 지름의 분포가 다르며, 위날개뼈 지름의 범위 또한 차이가 존재함을 확인할 수 있다. SO(노래하는 조류), P(산악지대에 서식하는 조류)의 경우 다른 type에 비해 위날개뼈의 지름이 짧고, SW(수영하는 조류)의 경우 다른 type에 비해 위날개뼈의 지름이 길다. 해당 그룹을 제외한 나머지 type은 비슷한 지름 범위를 보임을 확인할 수 있다.

2-2. ulna(자뼈)

자뼈란 날개뼈 바로 위의 뼈를 말한다.

길이

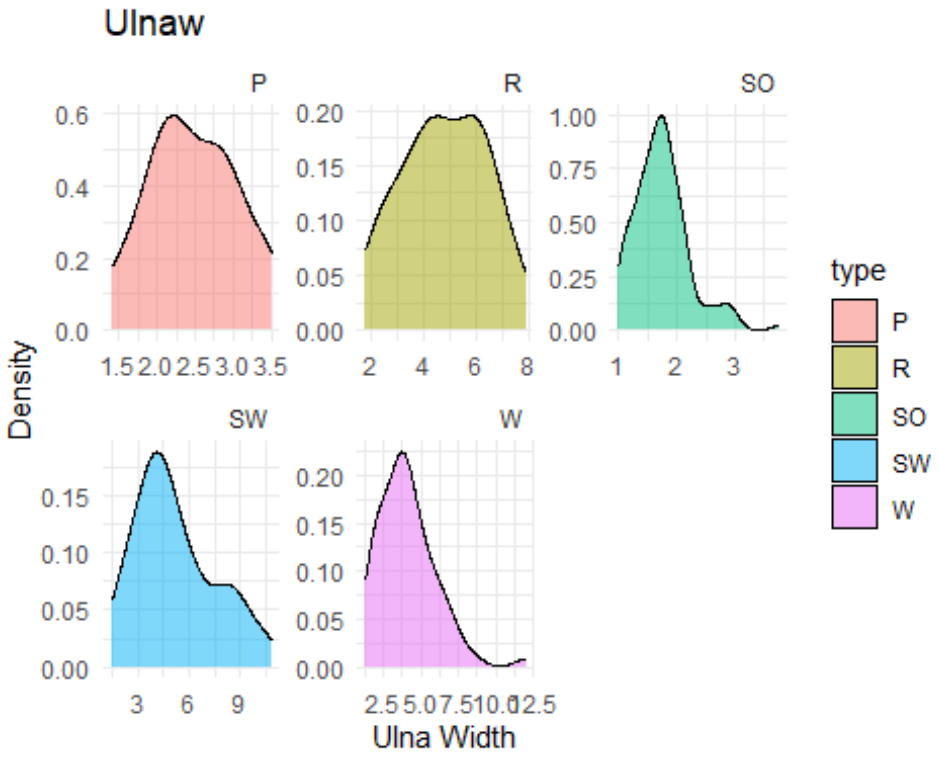
```
# type별 ulnal의 분포 시각화
ggplot(bird, aes(x = ulnal, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Ulnal", x = "Ulna Length", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



type별로 자뼈 길이의 분포가 다르며, 자뼈 길이의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. SO(노래하는 조류), P(산악지대에 서식하는 조류)의 경우 다른 type에 비해 자뼈의 길이가 짧고, SW(수영하는 조류)의 경우 다른 type에 비해 자뼈의 길이가 길다. 해당 그룹들을 제외한 나머지 type은 비슷한 길이 범위를 보임을 확인할 수 있다.

지름

```
# type별 humw의 분포 시각화
ggplot(bird, aes(x = ulnaw, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Ulnaw", x = "Ulna Width", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```

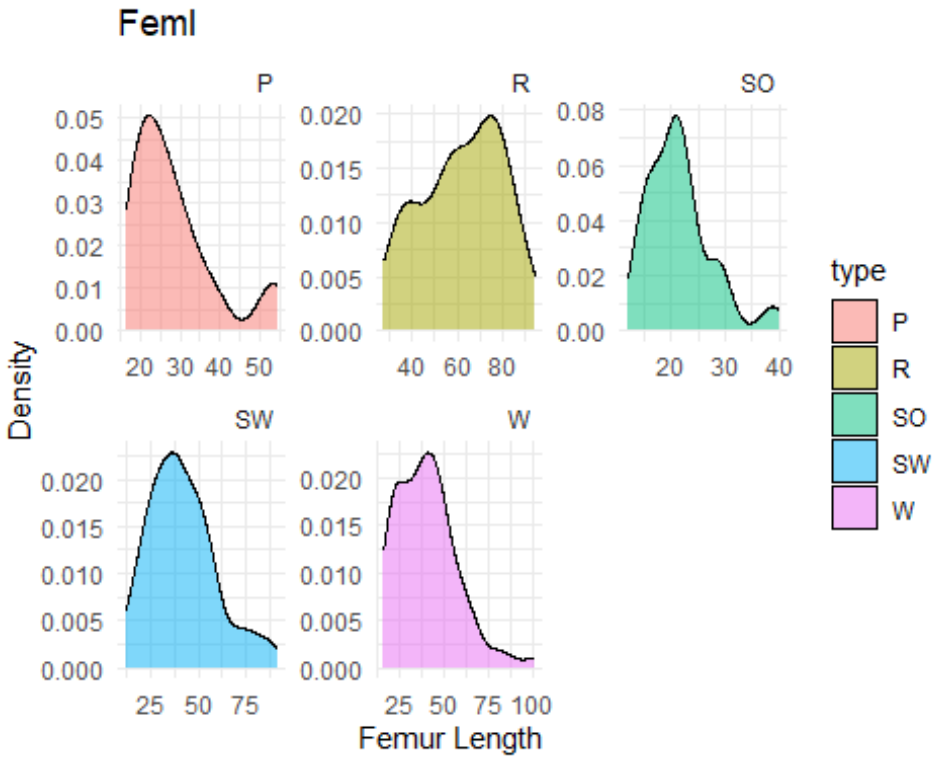


type별로 자뼈 지름의 분포가 다르며, 자뼈 지름의 범위 또한 차이가 존재함을 확인할 수 있다. SO(노래하는 조류), P(산악지대에 서식하는 조류)의 경우 다른 type에 비해 자뼈의 지름이 짧고, SW(수영하는 조류)의 경우 다른 type에 비해 자뼈의 지름이 길다. 해당 그룹들을 제외한 나머지 type은 비슷한 지름 범위를 보임을 확인할 수 있다.

2-3. femur(넓적다리뼈)

길이

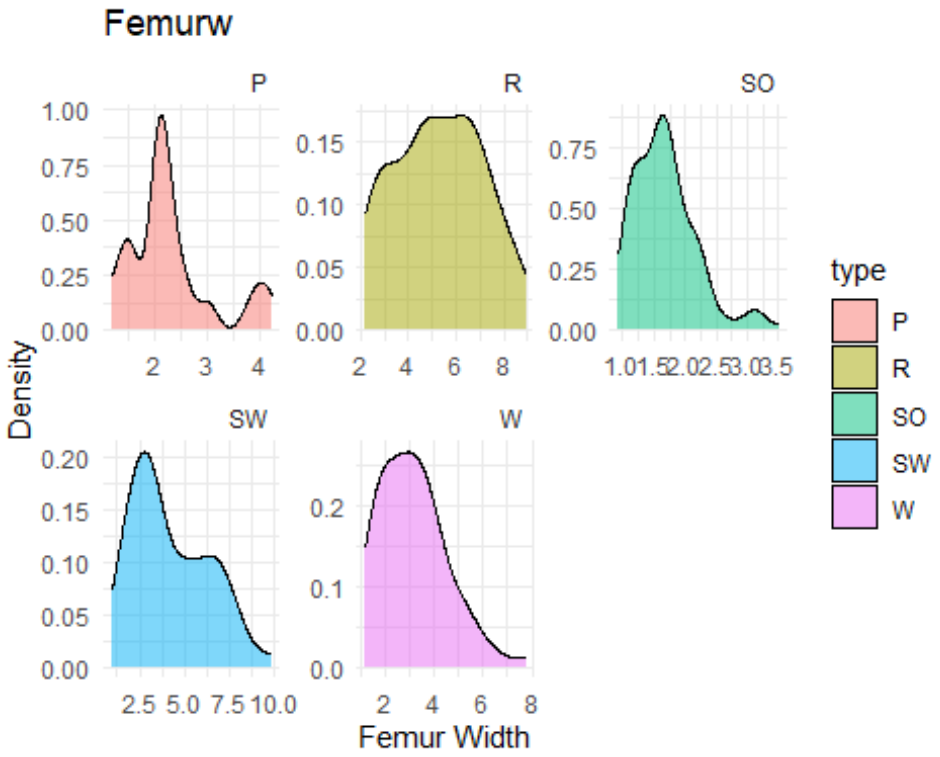
```
# type별 feml의 분포 시각화
ggplot(bird, aes(x = feml, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Feml", x = "Femur Length", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



type별로 넓적다리뼈 길이의 분포가 다르며, 넓적다리뼈 길이의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류), W(물가에 서식하는 조류)의 경우 다른 type에 비해 넓적다리뼈의 길이가 길다. 해당 그룹들을 제외한 나머지 type은 비슷한 길이 범위를 보임을 확인할 수 있다.

지름

```
# type별 femw의 분포 시각화
ggplot(bird, aes(x = femw, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Femurw", x = "Femur Width", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```

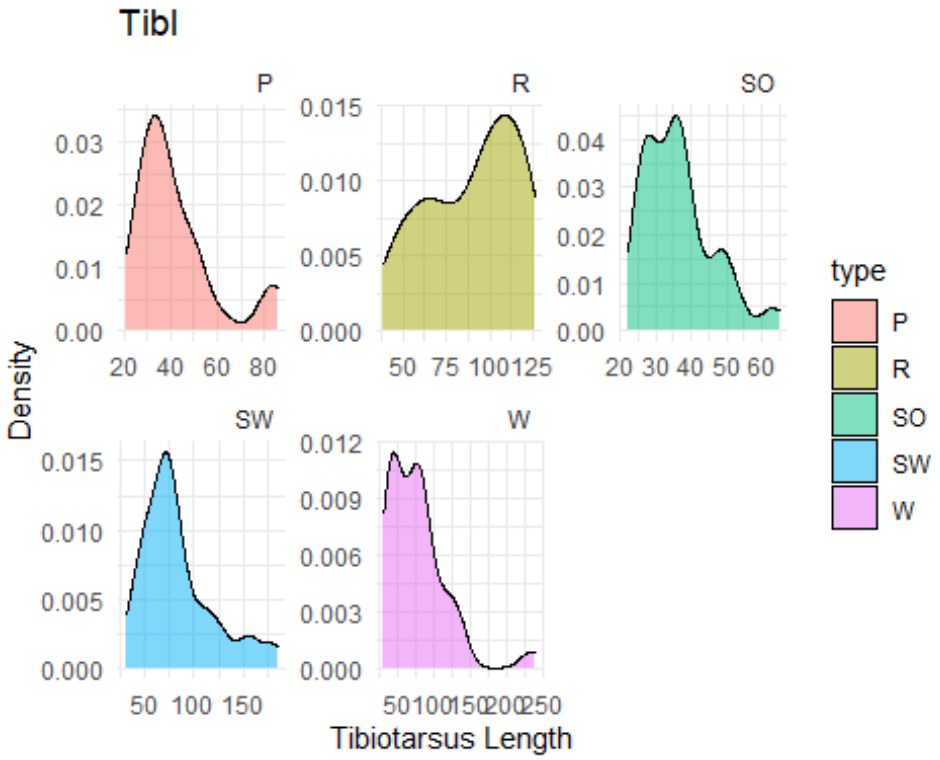


type별로 넓적다리뼈 지름의 분포가 다르며, 넓적다리뼈 지름의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류), W(물가에 서식하는 조류)의 경우 다른 type에 비해 넓적다리뼈의 지름이 길다. 해당 그룹들을 제외한 나머지 type은 비슷한 범위를 보임을 확인할 수 있다.

2-4. tibiotarsus(정강발목뼈)

길이

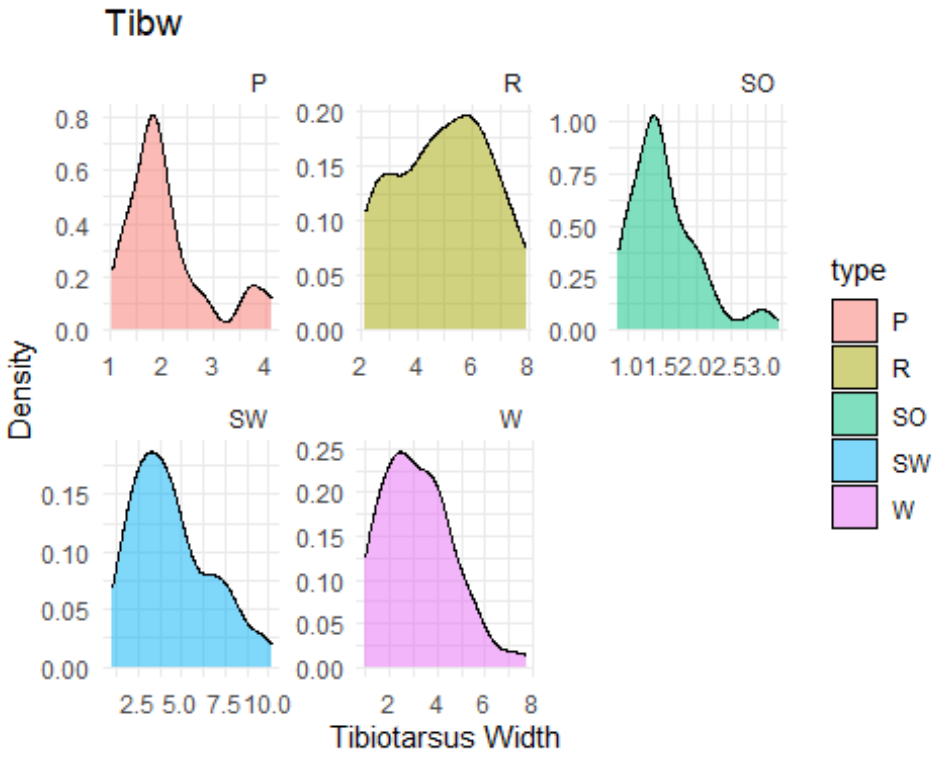
```
# type별 tibl의 분포 시각화
ggplot(bird, aes(x = tibl, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tibl", x = "Tibiotarsus Length", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



type별로 정강발목뼈 길이의 분포가 다르며, 정강발목뼈 길이의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류), W(물가에 서식하는 조류)의 경우 다른 type에 비해 넓적다리뼈의 길이가 길다. 특히, W(물가에 서식하는 조류)의 경우 비교적 길이가 긴 R과 SW에 비해서도 정강발목뼈 길이의 최댓값이 2배 정도 더 크다. 이는 물가에 서식하는 조류에게서 보이는 주요한 생태학적 특징임을 짐작할 수 있다. 나머지 type은 비슷한 길이 범위를 보임을 확인할 수 있다.

지름

```
# type별 tibw의 분포 시각화
ggplot(bird, aes(x = tibw, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tibw", x = "Tibiotarsus Width", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```

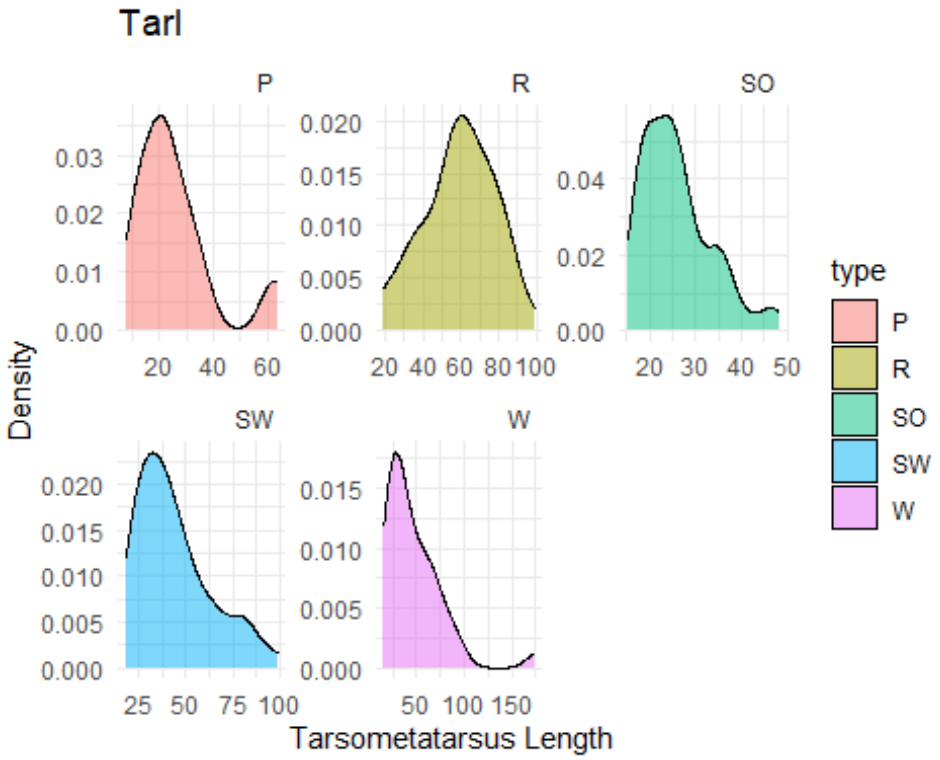


type별로 정강발목뼈 지름의 분포가 다르며, 정강발목뼈 지름의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류), W(물가에 서식하는 조류)의 경우 다른 type에 비해 정강발목뼈의 지름이 길다. 해당 그룹들을 제외한 나머지 type은 비슷한 범위를 보임을 확인할 수 있다.

2-5. tarsometatarsus(뒷발목뼈)

길이

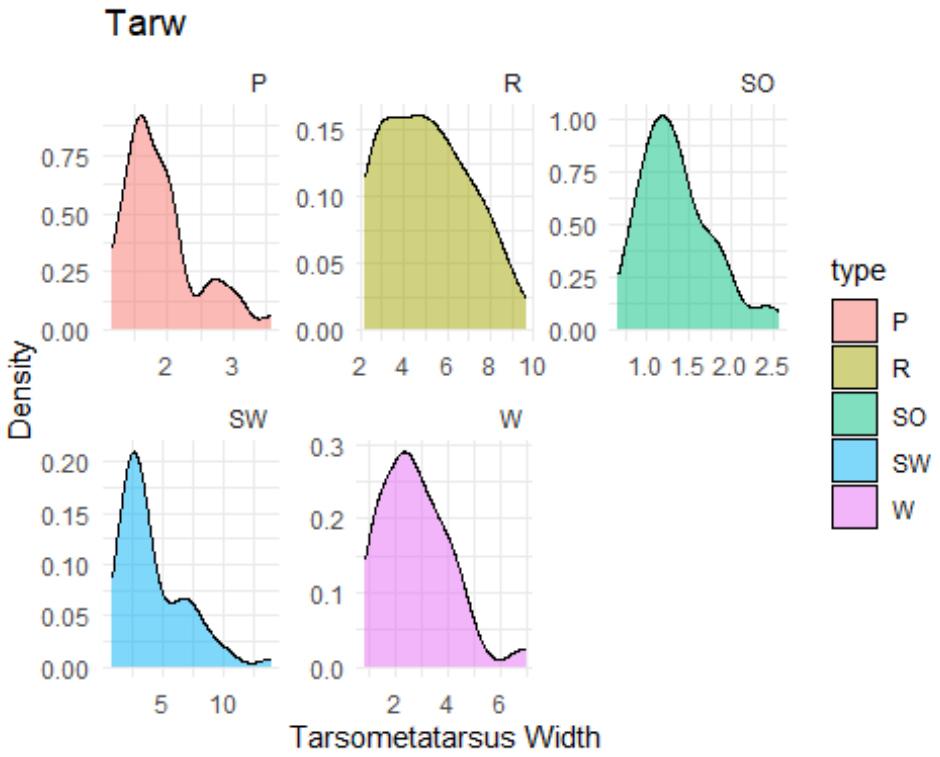
```
# type별 tarl의 분포 시각화
ggplot(bird, aes(x = tarl, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tarl", x = "Tarsometatarsus Length", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```



type별로 뒷발목뼈 길이의 분포가 다르며 뒷발목뼈 길이의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류), W(물가에 서식하는 조류)의 경우 다른 type에 비해 넓적다리뼈의 길이가 길다. 특히 W(물가에 서식하는 조류)의 경우 비교적 길이가 긴 R과 SW에 비해서도 뒷발목뼈 길이의 최댓값이 1.5배 정도 더 크다. 이는 물가에 서식하는 조류에게서 보이는 주요한 생태학적 특징임을 짐작할 수 있다. 해당 그룹들을 제외한 나머지 type은 비슷한 길이 범위를 보임을 확인할 수 있다. 또한, SW(수영하는 조류)의 경우 뒷발목뼈의 분포가 약간 왜곡되었다고 볼 수 있다.

지름

```
# type별 tarw의 분포 시각화
ggplot(bird, aes(x = tarw, fill = type)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tarw", x = "Tarsometatarsus Width", y = "Density") +
  theme_minimal() +
  facet_wrap(~ type, nrow = 2, ncol = 3, scales = "free") +
  theme(strip.text.x = element_text(angle = 0, hjust = 1))
```

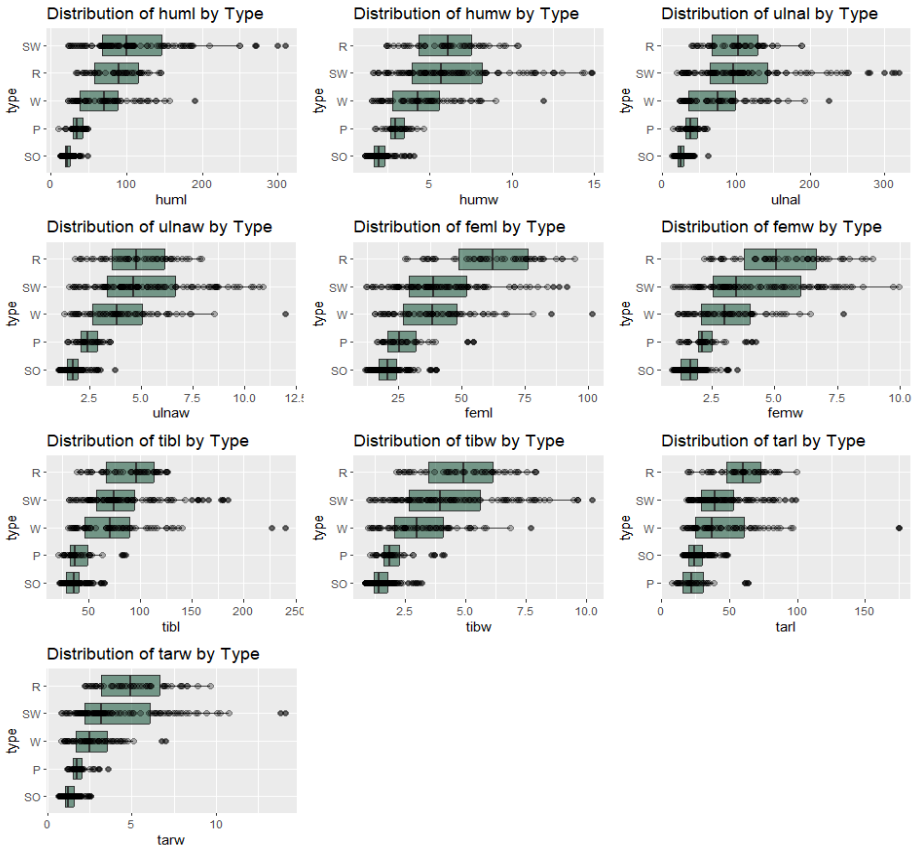


type별로 뒷발목뼈 지름의 분포가 다르며, 뒷발목뼈 지름의 범위 또한 차이가 꽤나 큰 것을 확인할 수 있다. R(사냥하는 조류), SW(수영하는 조류)의 경우 다른 type에 비해 뒷발목뼈의 지름이 길다. 그에 반해 P(산악지대에 서식하는 조류), SO(노래하는 조류)의 경우 다른 type에 비해 뒷발목뼈의 지름이 짧은 것을 확인할 수 있다.

3. 독립변수들의 분포 확인(2)

분석: 차수빈, 김경민, 장단, 이도경

```
library(rlang)
library(patchwork)

cols <- colnames(bird[-11])
plot_list <- list()  # 빈 리스트 생성

for (col in cols) {
  plot_title <- paste("Distribution of", col, "by Type")
  p <- ggplot(bird, aes(x=!!rlang::sym(col), y=reorder(type, !!rlang::sym(col),
FUN=median))) +
    geom_boxplot(fill="#00462A", alpha=0.5) +
    geom_point(alpha=0.3, size=2) +
    labs(title=plot_title, x=col, y="type")
  plot_list[[length(plot_list) + 1]] <- p  # 플롯을 리스트에 추가
}

# Combine all plots into one screen
combined_plot <- wrap_plots(plot_list, ncol=3, nrow=4)
print(combined_plot)
```

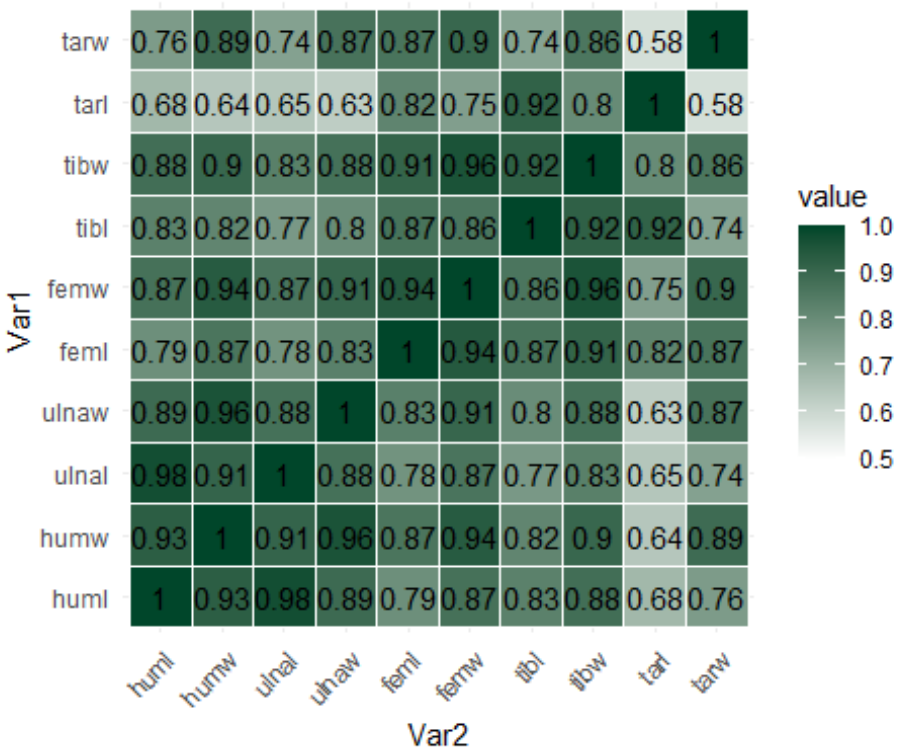


대부분의 변수에서 비슷한 양상이 드러난다. R, SW는 분포가 넓고 값이 큰 경향을 보이고 P, SO는 좁게 분포하고 값이 작은 경향을 보인다.

4. 독립변수들 간 correlation

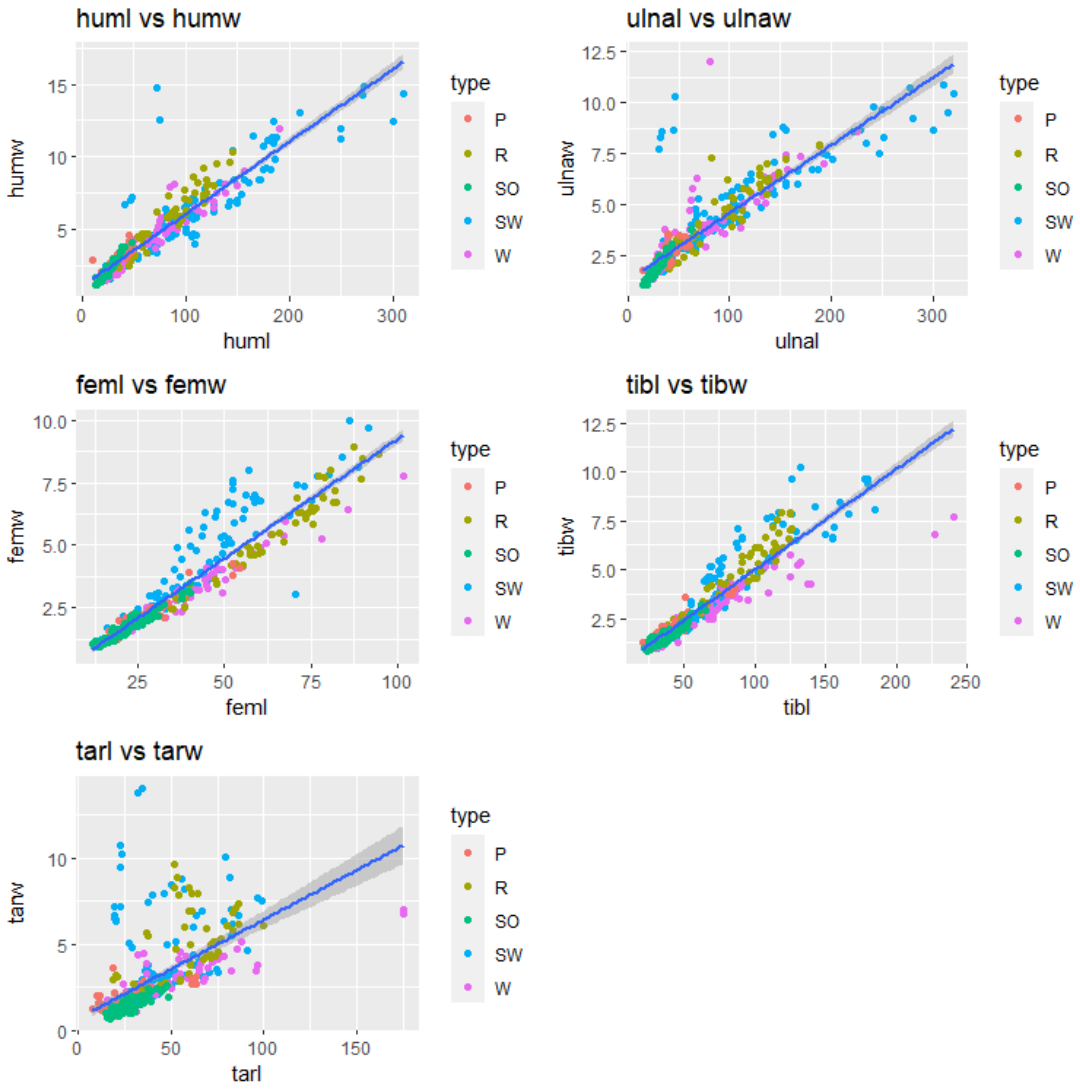
분석: 김경민, 이도경, 차수빈

```
corr <- cor(bird[-11])

library(reshape2)

corr_bird <- melt(corr)

ggplot(corr_bird, aes(Var2, Var1, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(value, 2)), color = "black") +
  scale_fill_gradient(low = "white", high = "#00462A", limits = c(0.5, 1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_fixed()
```



```r
plot1 <- ggplot(bird, aes(x=huml, y=humw)) +
  geom_point(aes(color = type)) +
  geom_smooth(method="lm") +
  xlab("huml") +
  ylab("humw") +
  labs(colour="type") +
  ggtitle("huml vs humw")

plot2 <- ggplot(bird, aes(x=ulnal, y=ulnaw)) +
  geom_point(aes(color = type)) +
  geom_smooth(method="lm") +
  xlab("ulnal") +
  ylab("ulnaw") +
  labs(colour="type") +
  ggtitle("ulnal vs ulnaw")

plot3 <- ggplot(bird, aes(x=feml, y=femw)) +
  geom_point(aes(color = type)) +
  geom_smooth(method="lm") +
  xlab("feml") +
  ylab("femw") +
  labs(colour="type") +
  ggtitle("feml vs femw")

plot4 <- ggplot(bird, aes(x=tibl, y=tibw)) +
  geom_point(aes(color = type)) +
  geom_smooth(method="lm") +
  xlab("tibl") +
  ylab("tibw") +
  labs(colour="type") +
  ggtitle("tibl vs tibw")

plot5 <- ggplot(bird, aes(x=tarl, y=tarw)) +
  geom_point(aes(color = type)) +
  geom_smooth(method="lm") +
  xlab("tarl") +
  ylab("tarw") +
  labs(colour="type") +
  ggtitle("tarl vs tarw")

combined_plot <- wrap_plots(plot1, plot2, plot3, plot4, plot5, ncol = 2)
combined_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



대부분의 변수들에서 강한 양의 상관관계를 보이고 있다. 특히, 같은 종류의 뼈인 경우 길이와 지름이 강한 상관관계를 가짐을 확인할 수 있다. 다만 tar(뒷발목뼈)의 경우 다른 변수들의 비해 상관관계가 상대적으로 약했다.

4) 다변량 분석

Chapter 1. 정준상관분석(CCA)

작성자: 이도경

I. 서론

정준상관분석(CCA, Canonical Correlation Analysis)은 두 개의 다변량 집단간의 관계를 분석하는 데 사용되는 통계 기법이다. 본 항목에서는 type을 제외하고 날개 뼈와 다리 뼈의 길이 및 지름 변수들을 이용하여 두 번의 정준상관분석을 통해 변수들간의 관계를 파악하고자 하였다.

II. 분석 목표 및 예측

본 항목에서는 다음의 목표를 가지고 정준상관분석을 진행하였다.

1. 새의 뼈의 길이와 지름 간의 관계 분석
2. 새의 날개 뼈의 길이 및 지름과 다리 뼈의 길이 및 지름 간의 관계 분석

앞서 진행한 EDA 결과에서 모든 변수들간의 상관관계가 매우 높았기에, 설정한 두 집단 사이에 강한 상관성이 존재할 것이라고 예측된다.

III. 정준상관분석 수행

1. 뼈의 길이와 지름 간 관계

1-1. 정준상관분석을 위한 데이터 준비

뼈의 길이와 관련된 변수들을 X, 뼈의 지름과 관련된 변수들을 Y로 분리하였다.

- X : 뼈의 길이, 5차원
- Y : 뼈의 지름 , 5차원

```r
X = bird[c(1,3,5,7,9)]
head(X)
```

```
##     huml  ulnal  feml   tibl  tarl
## 1  79.97  69.26 43.07  75.35 38.31
## 2  77.65  65.76 40.04  69.17 35.78
## 3  79.73  67.39 42.07  71.26 37.22
## 4  86.98  74.52 44.46  76.02 37.94
## 5 118.20 116.64 59.33 110.00 61.62
## 6 145.00 144.00 70.96 120.00 78.67
```

```r
Y = bird[c(2,4,6,8,10)]
head(Y)
```

```
##    humw ulnaw femw tibw tarw
## 1  6.37  5.28 3.90 4.04 3.34
## 2  5.70  4.77 3.52 3.40 3.41
## 3  5.94  4.50 3.41 3.56 3.64
## 4  5.68  4.55 3.78 3.81 3.81
## 5  7.82  6.13 5.45 5.58 4.37
## 6 10.42  7.05 7.44 7.31 6.34
```

뼈의 길이 변수와 지름 변수의 범위 차이가 분석 결과에 영향을 미칠 가능성이 있기에, 평균이 0, 분산이 1이 되도록 표준화를 진행하여 새로운 변수 Zx와 Zy를 만들었다.

- Zx : 표준화 된 뼈의 길이, 5차원
- Zy : 표준화 된 뼈의 지름, 5차원

```r
Zx = scale(X)
```

```r
mean(Zx); diag(var(Zx))
```

```
## [1] -4.803604e-18
```

```
##  huml ulnal  feml  tibl  tarl
##     1     1     1     1     1
```

```r
Zy = scale(Y)
```

```r
mean(Zy); diag(var(Zy))
```

```
## [1] 2.406811e-17
```

```
##  humw ulnaw  femw  tibw  tarw
##     1     1     1     1     1
```

1-2. 정준상관분석 수행

표준화된 변수 Zx와 Zy로 정준상관분석을 진행하여 얻은 정준상관계수는 다음과 같다.

```
cc1 = cc(Zx,Zy)

# 정준상관계수
cc1$cor

## [1] 0.9821844 0.7925731 0.6537193 0.4980619 0.1161470
```

정준상관계수는 두 변수 집단 간의 상관관계를 나타내는 척도이다. 첫 번째 정준상관계수(0.9822)가 거의 1에 가깝기에 첫번째 정준변수 쌍(U1, V1)이 아주 강한 양의 상관관계를 갖는다는 것을 알 수 있다. 즉, X와 Y가 거의 완벽한 선형 관계를 보이므로 두 집단 간의 패턴이 유사하다는 것을 알 수 있다.

두 번째 정준상관계수(0.7926) 또한 0.8에 가까운 상관계수로 두 번째 정준변수 쌍(U2, V2)의 강한 양의 상관관계를 보여준다. 즉, 두 번째 정준변수 쌍 역시 두 집단 간의 관계를 주요하게 설명하며, X와 Y 사이에 다차원적인 상관관계가 존재한다는 것을 알 수 있다.

다음은 정준상관분석을 통해 얻은 전체 정준변수의 선형 결합식이다.

$$U_1 = -0.4391X_1 - 0.0018X_2 - 0.4645X_3 - 0.4519X_4 + 0.3424X_5$$

$$U_2 = 2.7679X_1 - 2.2334X_2 - 1.5598X_3 + 0.5050X_4 + 0.5394X_5$$

$$U_3 = 0.1672X_1 + 1.2042X_2 - 0.6398X_3 - 0.0732X_4 - 0.8023X_5$$

$$U_4 = 1.8176X_1 - 2.5190X_2 + 1.0655X_3 + 1.2256X_4 - 1.9996X_5$$

$$U_5 = 5.6246X_1 - 4.3641X_2 + 1.0706X_3 - 4.1506X_4 + 1.9095X_5$$

$$V_1 = -0.5170Y_1 - 0.0378Y_2 - 0.0805Y_3 - 0.4552Y_4 + 0.0704Y_5$$

$$V_2 = 0.7607Y_1 + 0.3078Y_2 - 3.4378Y_3 + 3.0999Y_4 - 0.7913Y_5$$

$$V_3 = 2.7102Y_1 + 0.0807Y_2 - 0.7632Y_3 - 1.5126Y_4 - 0.5766Y_5$$

$$V_4 = 0.1813Y_1 - 0.0673Y_2 - 3.2146Y_3 + 1.0324Y_4 + 2.2655Y_5$$

$$V_5 = 3.4944Y_1 - 3.6908Y_2 - 0.3721Y_3 + 0.3262Y_4 + 0.1334Y_5$$

```
# 정준변수 U의 coefficient
cc1$xcoef
```

```
##              [,1]       [,2]        [,3]      [,4]      [,5]
## huml  -0.43905335  2.7678977  0.16718756  1.817610  5.624604
## ulnal -0.00180951 -2.2333517  1.20415772 -2.519040 -4.364141
## feml  -0.46450381 -1.5597944 -0.63976702  1.065549  1.070647
## tibl  -0.45191783  0.5050242 -0.07319958  1.225577 -4.150633
## tarl   0.34244555  0.5393571 -0.80225136 -1.999647  1.909545
```

```
# 정준변수 V의 coefficient
cc1$ycoef
```

```
##              [,1]       [,2]        [,3]        [,4]       [,5]
## humw  -0.51697679  0.7606603  2.71020014  0.18125535  3.4944333
## ulnaw -0.03768535  0.3078232  0.08065738 -0.06727986 -3.6907568
## femw  -0.08058408 -3.4377918 -0.76317230 -3.21455715 -0.3721291
## tibw  -0.45521430  3.0999234 -1.51256649  1.03243119  0.3261988
## tarw   0.07040808 -0.7913376 -0.57657207  2.26547162  0.1334186
```

정준상관계수가 가장 큰 첫 번째 정준변수 쌍인 U1에 X1(huml), X3(feml), X4(tibl), X5(tarl)의 상대적 기여도가 높고, V1에는 Y1(humw), Y4(tibw)의 상대적 기여도가 높은 것을 정준 계수를 통해 확인할 수 있다. 또한 정준상관계수가 두 번째로 큰 U2에 X1(huml), X2(ulnal), X3(feml)의 기여도가, V2에는 Y3(femw), Y4(tibw)의 기여도가 크다는 사실을 알 수 있다.

정준적재는 원래의 변수들과 정준변수 간의 상관관계를 나타낸다. 다음은 정준상관분석을 통해 얻은 정준적재 값이다.

```
# corr(X,U)
colnames(cc1$scores$corr.X.xscores) <- c("U1","U2","U3","U4","U5")
cc1$scores$corr.X.xscores
```

```
##               U1          U2         U3          U4           U5
## huml  -0.9498050  0.14176107  0.2266230 -0.14474527  0.073926406
## ulnal -0.9185339 -0.01786942  0.2919175 -0.26562584  0.014254264
## feml  -0.9255078 -0.23235638 -0.2887236 -0.04350932  0.064741522
## tibl  -0.9077391  0.22788918 -0.3057024 -0.10501240 -0.139981167
## tarl  -0.7557480  0.16722395 -0.5029640 -0.38458310 -0.002055427
```

```r
# corr(Y,U)
colnames(cc1$scores$corr.Y.xscores) <- c("U1","U2","U3","U4","U5")
cc1$scores$corr.Y.xscores
```

```
##                U1          U2          U3             U4             U5
## humw  -0.9605466 -0.05547239  0.12359727  0.0267134210  0.0008481614
## ulnaw -0.9290011 -0.04122316  0.11040163  0.0347199576 -0.0305707489
## femw  -0.9564613 -0.14326090 -0.07988857 -0.0316687711 -0.0007716644
## tibw  -0.9554776  0.05682294 -0.14378449 -0.0006756302  0.0012927945
## tarw  -0.8724601 -0.22595955 -0.05444061  0.1741149243 -0.0026598480
```

```r
# corr(X,V)
colnames(cc1$scores$corr.X.yscores) <- c("V1","V2","V3","V4","V5")
cc1$scores$corr.X.yscores
```

```
##                V1          V2         V3          V4             V5
## huml  -0.9328837  0.11235601  0.1481478 -0.07209211  0.0085863329
## ulnal -0.9021697 -0.01416282  0.1908321 -0.13229812  0.0016555905
## feml  -0.9090193 -0.18415942 -0.1887442 -0.02167034  0.0075195358
## tibl  -0.8915672  0.18061883 -0.1998436 -0.05230268 -0.0162583975
## tarl  -0.7422839  0.13253721 -0.3287973 -0.19154621 -0.0002387318
```

```r
# corr(Y,V)
colnames(cc1$scores$corr.Y.yscores) <- c("V1","V2","V3","V4","V5")
cc1$scores$corr.Y.yscores
```

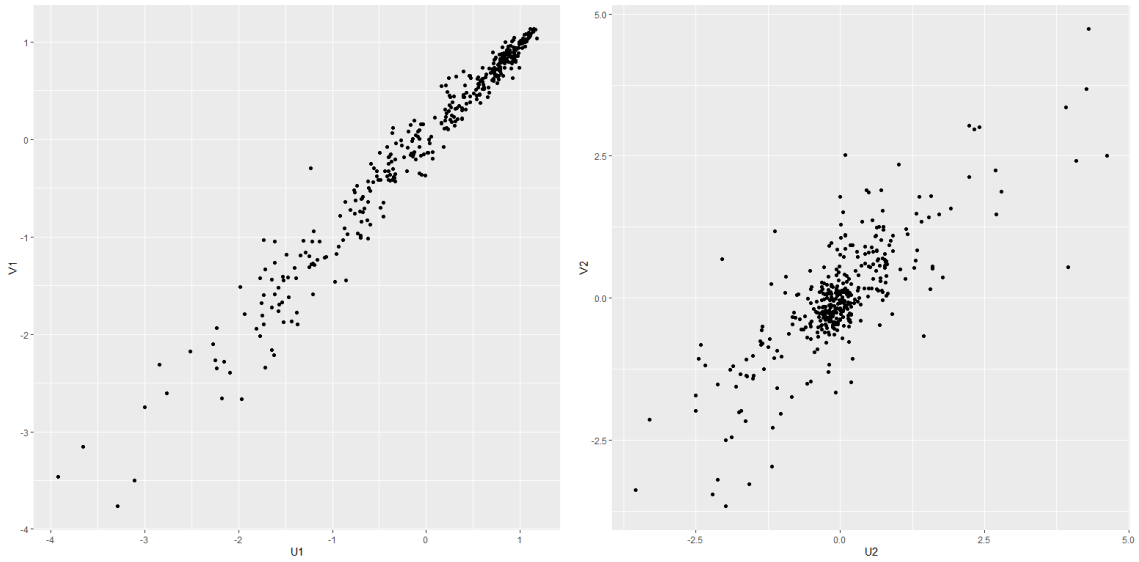
```
##                V1          V2          V3           V4           V5
## humw  -0.9779697 -0.06999025  0.18906779  0.053634736  0.007302480
## ulnaw -0.9458520 -0.05201181  0.16888231  0.069710119 -0.263207312
## femw  -0.9738103 -0.18075418 -0.12220623 -0.063584000 -0.006643858
## tibw  -0.9728088  0.07169426 -0.21994835 -0.001356518  0.011130672
## tarw  -0.8882854 -0.28509616 -0.08327826  0.349584875 -0.022900697
```

U1과 V1에 대한 정준적재의 절대값이 나머지 정준변수에 비해 절대적으로 크기에, 첫번째 정준변수 쌍에 미치는 X와 Y의 영향 또한 가장 크다는 것을 알 수 있다. 따라서 X와 Y의 관계를 분석할 때 첫 번째 정준변수 쌍만 사용하여 두 집단 간의 관계의 대부분을 설명하는 것이 가능하다.

1-3. 시각화

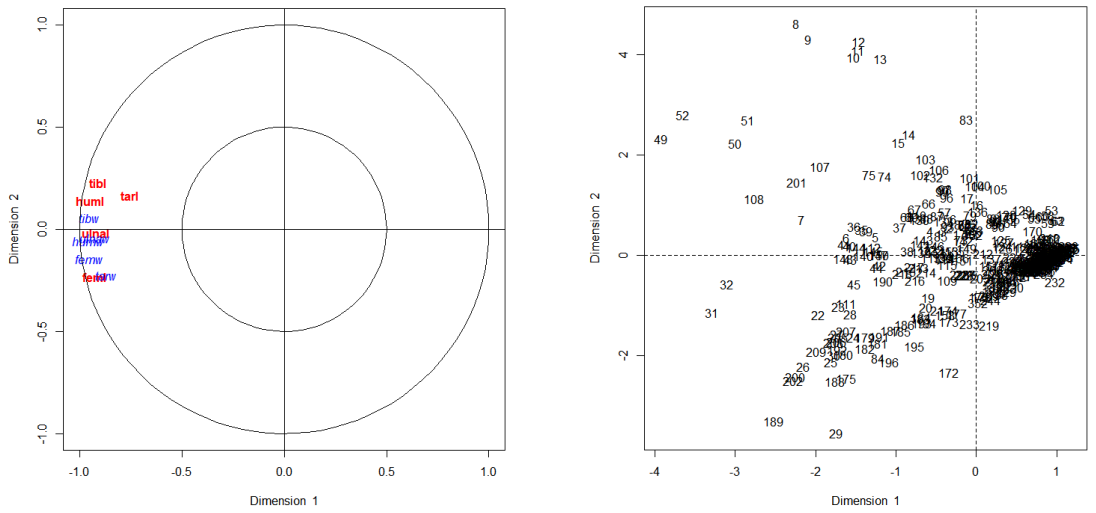
앞서 진행한 정준상관분석의 결과를 한 눈에 알아보기 쉽도록 시각화 하였다.

```
U1 = cc1$scores$xscores[,1]
V1 = cc1$scores$yscores[,1]
ggplot(data.frame(U1,V1), aes(U1,V1))+geom_point()

U2 = cc1$scores$xscores[,2]
V2 = cc1$scores$yscores[,2]
ggplot(data.frame(U2,V2), aes(U2,V2))+geom_point()
```



이 두 그래프는 정준상관계수가 가장 높은 정준변수 쌍인 (U1, V1)과 (U2, V2)의 산점도이다. 정준상관계수를 통해 알 수 있듯이 U1과 V1은 거의 완벽한 선형 관계를 보이며, U2와 V2는 양의 선형관계를 보이지만 첫 번째 정준변수 쌍보다는 약하다.

```
plt.cc(cc1, type="v", var.label=TRUE)
plt.cc(cc1, type="i", var.label=TRUE)
```



왼쪽의 그래프는 처음 두 쌍의 정준변수에 대한 원래 변수들의 상관관계를 나타낸 것이다. 첫 번째 정준변수에 대해서는 대부분의 변수의 영향력이 비슷하나, X5(tarl)의 영향력이 약간 적다는 사실을 확인할 수 있다. 두 번째 정준변수에 대해서는 모든 변수의 상관관계가 0.5 이하로, 그 영향력이 크지 않다는 것을 알 수 있다.

오른쪽의 그래프는 처음 두 쌍의 정준변수에 대해 표준화된 각 관측치의 위치를 나타낸 것이다. 관측치들의 분포 양상을 통해 U1, U2 간의 상관관계가 거의 보이지 않음을 확인할 수 있다.

1-4. 상관성 검정 및 시각화

X 집단과 Y 집단 사이에 충분한 상관성이 존재하지 않으면 정준상관분석의 결과가 유의하다고 판단하기 어렵다. 따라서 앞서 진행한 정준상관분석이 유의한지 확인하기 위해 유의수준 0.05 하에서 상관성 검정을 진행하였다.

$$H_0 : \sum XY = \mathbf{0} \text{ (상관성이 없다.)}$$

$$H_1 : \sum XY \neq \mathbf{0} \text{ (상관성이 있다.)}$$

$$\Lambda = \frac{|S|}{|S_{XX}||S_{YY}|} = \frac{|R|}{|R_{XX}||R_{YY}|} = \Pi_{i=1}^{p}(1 - \hat{\rho}_i^{*2})$$

$$-\left(n - \frac{1}{2(p+q+3)}\right) log\Lambda \sim \chi_{pq}^2$$

위의 식을 통해 얻은 검정통계량(Λ) 및 χ_{pq}^2와 p-value는 다음과 같다.

```
rho = cc1$cor
n = nrow(df)
p = ncol(X)
q = ncol(Y)
alpha = 0.05
(Lambda = prod(1-rho^2))
```

```
## [1] 0.005577724
```

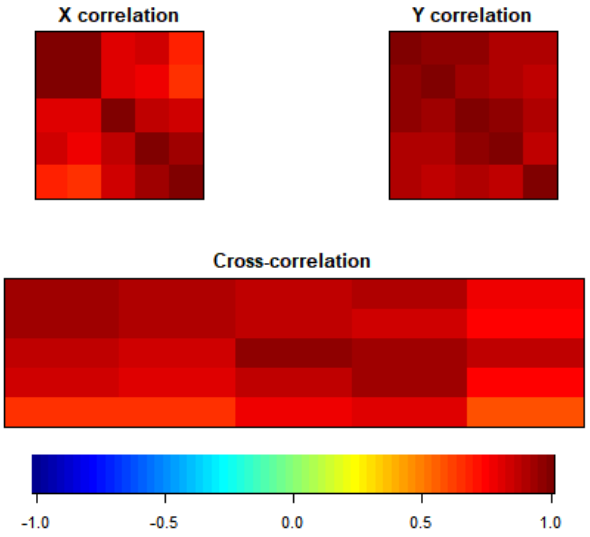
```
term = -(n-(p+q+3)/2)

# 통계량
chi_stat = term*log(Lambda)

# chi-square
chi = qchisq(1 - alpha, p*q)

# p-value
p_value = 1 - pchisq(chi_stat, p*q)
c(chi_stat, chi, p_value)
```

```
## [1] 1974.40478   37.65248    0.00000
```

term $\cdot log\Lambda(1974.40) > \chi_{pq}^2(37.6525)$ 이므로 귀무가설을 기각하고, X와 Y 사이에 상관성이 존재한다는 결론을 얻을 수 있다.

다음은 X의 변수들과 Y의 변수들 간의 상관관계를 한 눈에 볼 수 있도록 시각화한 그래프이다.

```
mtc = matcor(X,Y)
img.matcor(mtc, type=2)
```



푸른색에 가까울수록 강한 음의 상관관계를, 붉은색에 가까울수록 강한 양의 상관관계를 나타낸다. X 변수 간의 상관관계, Y 변수 간의 상관관계, X와 Y 변수 간의 상관관계를 나타내는 색이 모두 붉은색이므로, 상관성 검정 결과에서 확인했듯이 두 집단 간의 상관성이 매우 높다는 것을 알 수 있다.

1-5. 요약 및 의미

해당 분석을 통해 새의 뼈 길이(X)와 지름(Y) 간의 상관관계를 파악할 수 있었다. 정준상관분석 결과, X, Y와의 상관관계가 가장 높은 첫 번째 정준변수의 정준상관계수는 0.98로, 매우 강한 양의 상관관계를 보인다. 또한, 첫 번째 정준변수 쌍을 통해 두 집단 간의 관계를 대부분 설명할 수 있다. 특히, 윗날개뼈 길이(X1, huml), 넓적다리뼈 길이(X3, feml), 정강발목뼈 길이(X4, tibl) , 윗날개뼈 지름(Y1, humw) , 정강발목뼈 지름(Y4, tibw)의 기여도가 높은 것을 확인할 수 있었다.

2. 날개 뼈와 다리 뼈 간 관계 분석

2-1. 정준상관분석을 위한 데이터 준비

날개 뼈와 관련된 변수들을 X, 다리 뼈와 관련된 변수들을 Y로 분리하였다.

- X : 날개 뼈의 길이 및 지름, 4차원
- Y : 다리 뼈의 길이 및 지름 , 6차원

```
# X : 새의 날개 뼈, Y : 새의 다리 뼈

X = bird[,1:4]
Y = bird[,5:10]
```

뼈의 길이 변수와 지름 변수의 범위 차이가 분석 결과에 영향을 미칠 가능성이 있기에, X와 Y를 평균이 0, 분산이 1이 되도록 표준화 하여 새로운 변수 Zx와 Zy를 생성하였다.

- Zx : 표준화된 날개 뼈의 길이 및 지름, 4차원
- Zy : 표준화된 다리 뼈의 길이 및 지름 , 6차원

```
# X와 Y 표준화
Zx = scale(X)
Zy = scale(Y)

mean(Zx); diag(var(Zx))

## [1] 4.13649e-17

##  huml  humw ulnal ulnaw
##     1     1     1     1

mean(Zy); diag(var(Zy))

## [1] -1.037878e-17

## feml femw tibl tibw tarl tarw
##    1    1    1    1    1    1
```

2-2. 정준상관분석 수행

표준화된 변수 Zx와 Zy로 정준상관분석을 진행하여 얻은 정준상관계수는 다음과 같다.

```
bird_cc = cc(Zx,Zy)

# 정준상관계수
bird_cc$cor

## [1] 0.9600571 0.7705648 0.4911247 0.1274248
```

정준상관계수는 두 변수 집단 간의 상관관계를 나타내는 척도이다. 첫 번째 정준상관계수(0.9600)가 거의 1에 가깝기에, 첫번째 정준변수 쌍(U1, V1)이 매우 강한 양의 상관관계를 갖는다는 것을 알 수 있다. 즉, X와 Y가 거의 완벽한 선형 관계를 보이므로 두 집단 간의 패턴이 유사하다는 것을 알 수 있다.

두 번째 정준상관계수(0. 7706) 또한 0.8에 가까운 상관계수로 두 번째 정준변수 쌍(U2, V2)의 강한 양의 상관관계를 보여준다. 즉, 두 번째 정준변수 쌍 역시 두 집단 간의 관계를 주요하게 설명 가능하며 X와 Y 사이에 다차원적인 상관관계가 존재한다는 것을 알 수 있다.

다음은 정준상관분석을 통해 얻은 전체 정준변수의 선형 결합식이다.

$$U_1 = -0.1360X_1 - 0.9532X_2 + 0.2426X_3 - 0.1445X_4$$

$$U_2 = -5.0255X_1 + 1.3852X_2 + 3.6982X_3 \_ 0.0692X_4$$

$$U_3 = -0.2573X_1 - 1.5825X_2 + 2.6023X_3 - 0.5081X_4$$

$$U_4 = -0.2823X_1 - 3.8084X_2 + 0.5842X_3 + 3.6445X_4$$

$$V_1 = 0.0442891Y_1 - 0.7871Y_2 - 0.9651Y_3 + 0.1982Y_4 + 0.4835Y_5 - 0.2006Y_6$$

$$V_2 = 0.3002Y_1 + 2.5464Y_2 - 1.0573Y_3 - 2.6407Y_4 + 0.4777Y_5 + 0.3559Y_6$$

$$V_3 = -1.0477Y_1 + 3.0705Y_2 - 1.0334Y_3 - 0.3020Y_4 + 1.0861Y_5 - 1.7100Y\_6$$

$$V_4 = -2.8278Y_1 + 2.4092Y_2 + 1.7740Y_3 - 3.4341Y_4 + 0.9676Y_5 + 1.3608Y\_6$$

```
# U의 정준계수
bird_cc$xcoef
```

```
##             [,1]        [,2]       [,3]       [,4]
## huml  -0.1360200 -5.02545237 -0.2573206 -0.2822968
## humw  -0.9532141  1.38518058 -1.5824762 -3.8084307
## ulnal  0.2426448  3.69824029  2.6023249  0.5841556
## ulnaw -0.1444534 -0.06918585 -0.5080989  3.6445323
```

```
# V의 정준계수
bird_cc$ycoef
```

```
##            [,1]       [,2]       [,3]       [,4]
## feml  0.0442891  0.3001517 -1.0476941 -2.8277813
## femw -0.7871159  2.5464488  3.0704958  2.4092708
## tibl -0.6950653 -1.0573250 -1.0333872  1.7739608
## tibw  0.1981612 -2.6407375 -0.3019789 -3.4341317
## tarl  0.4834804  0.4777407  1.0861365  0.9675962
## tarw -0.2005673  0.3558737 -1.7099892  1.3607928
```

정준상관계수가 가장 큰 첫 번째 정준변수 쌍인 U1에 대해 X2(humw)의 상대적 기여도가 높고, V1에 대해서는Y2(femw), Y3(tibl)의 상대적 기여도가 높은 것을 확인할 수 있다.
또한 정준상관계수가 두 번째로 큰 U2에 X1(huml), X3(ulnal)의 기여도가, V2에는 Y2(femw), Y4(tibw)의 기여도가 크다는 사실을 알 수 있다.

정준적재는 원래의 변수들과 정준변수 간의 상관관계를 나타낸다. 다음은 정준상관분석을 통해 얻은 정준적재 값이다.

```
# X와 U의 correlation
colnames(bird_cc$scores$corr.X.xscores) = c("U1","U2",'U3','U4')
bird_cc$scores$corr.X.xscores
```

```
##               U1           U2         U3           U4
## huml  -0.9137117 -0.193115412 0.35740528  0.009945751
## humw  -0.9971940  0.025427784 0.06361672 -0.030172663
## ulnal -0.8862818 -0.001323936 0.46221755  0.029286115
## ulnaw -0.9707571  0.011825244 0.02007589  0.238930382
```

```
# X와 V의 correlation
colnames(bird_cc$scores$corr.X.yscores) = c("V1","V2",'V3','V4')
bird_cc$scores$corr.X.yscores
```

```
##               V1           V2          V3           V4
## huml  -0.8772153 -0.148807933 0.175530570  0.001267336
## humw  -0.9573632  0.019593754 0.031243744 -0.003844746
## ulnal -0.8508812 -0.001020179 0.227006464  0.003731778
## ulnaw -0.9319822  0.009112116 0.009859768  0.030445659
```

```r
# Y와 U의 correlation
colnames(bird_cc$scores$corr.Y.xscores) = c("U1","U2",'U3','U4')
bird_cc$scores$corr.Y.xscores
```

```
##               U1          U2          U3             U4
## feml -0.8629279  0.05534640  0.03973065 -0.0214906446
## femw -0.9343640  0.04541638  0.08305228 -0.0087279709
## tibl -0.8229005 -0.25867952  0.07851611  0.0143374546
## tibw -0.9065858 -0.15825209  0.06827345 -0.0139763186
## tarl -0.6405972 -0.19273099  0.16890723  0.0142429390
## tarw -0.8985482  0.10949627 -0.11464224  0.0003453878
```

```r
# Y와 V의 correlation
colnames(bird_cc$scores$corr.Y.yscores) = c("V1","V2",'V3','V4')
bird_cc$scores$corr.Y.yscores
```

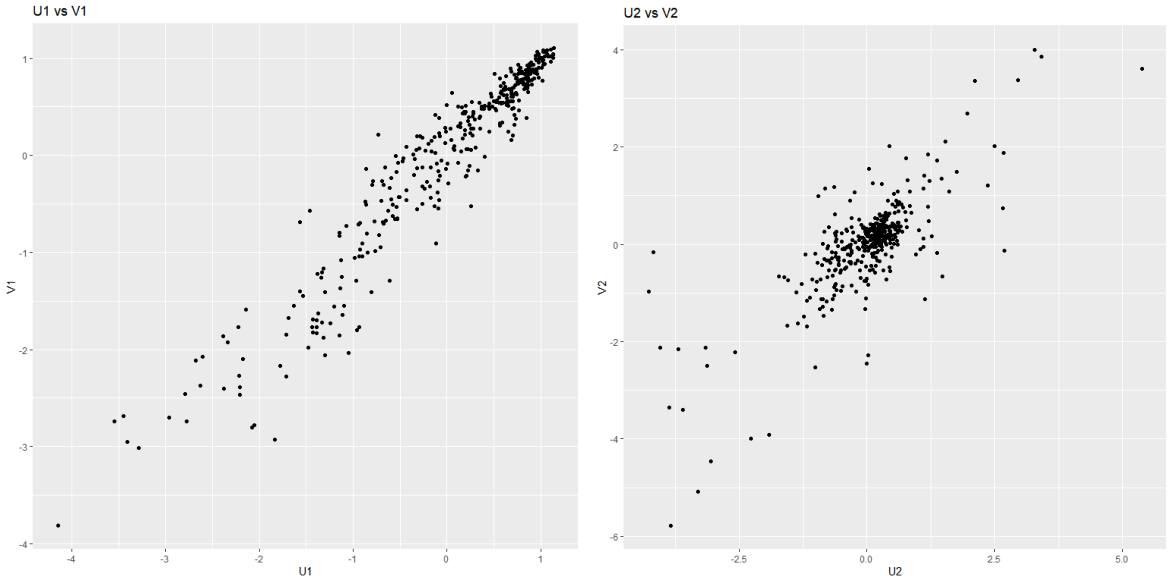
```
##               V1          V2          V3           V4
## feml -0.8988298  0.07182576  0.08089728 -0.168653530
## femw -0.9732380  0.05893908  0.16910629 -0.068495065
## tibl -0.8571371 -0.33570121  0.15987001  0.112516975
## tibw -0.9443041 -0.20537157  0.13901448 -0.109682865
## tarl -0.6672491 -0.25011654  0.34391921  0.111775239
## tarw -0.9359320  0.14209873 -0.23342796  0.002710523
```

U1과 V1에 대한 정준적재의 절대값이 나머지 정준변수에 비해 절대적으로 크기에, 첫 번째 정준변수 쌍에 미치는 X와 Y의 영향 또한 가장 크다는 것을 알 수 있다. 따라서 X와 Y의 관계를 분석할 때 첫 번째 정준변수 쌍만 사용하여 두 집단 간의 관계의 대부분을 설명할 수 있다.

2-3. 시각화

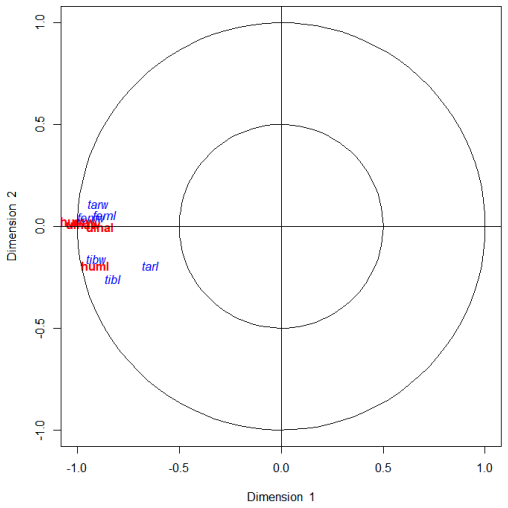
앞서 진행한 정준상관분석의 결과를 한 눈에 알아보기 쉽도록 시각화 하였다.

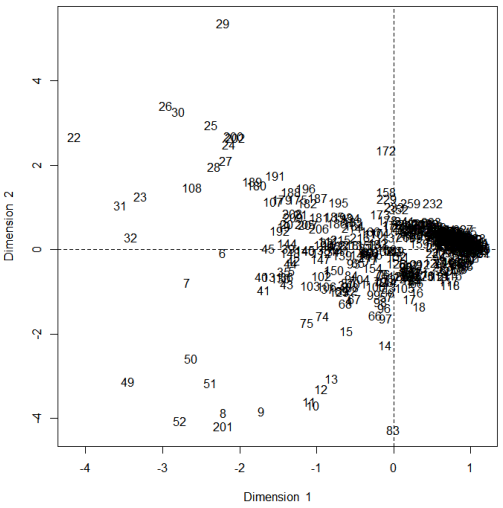
```
# U1과 V1의 correlation을 보기 위한 산점도
U1 = bird_cc$scores$xscores[,1]
V1 = bird_cc$scores$yscores[,1]
ggplot(data.frame(U1, V1), aes(U1, V1)) +
  geom_point() +
  labs(title = "U1 vs V1")

# U2과 V2의 correlation을 보기 위한 산점도
U2 = bird_cc$scores$xscores[,2]
V2 = bird_cc$scores$yscores[,2]
ggplot(data.frame(U2, V2), aes(U2, V2)) +
  geom_point() +
  labs(title = "U2 vs V2")
```



위 두 그래프는 정준상관계수가 가장 높은 정준변수 쌍인 (U1, V1)과 (U2, V2)의 산점도 plot이다. 정준상관계수를 통해 알 수 있듯이 U1과 V1은 거의 완벽한 선형 관계를 보이며, U2와 V2는 첫 번째 정준변수 쌍보다는 약한 양의 선형관계를 보인다.

```
plt.cc(bird_cc, type="v", var.label=TRUE)
plt.cc(bird_cc, type="i", var.label=TRUE)
```





왼쪽의 그래프는 처음 두 쌍의 정준변수에 대하여 원래의 변수들의 상관관계를 나타낸 것이다. 첫 번째 정준변수에 대해서는 대부분의 변수의 영향력이 비슷하나, Y5(tarl)의 영향력이 약간 적다는 사실을 확인할 수 있다. 두 번째 정준변수에 대해서는 모든 변수의 상관관계가 0.5 이하로, 그 영향력이 크지 않다는 것을 알 수 있다.

오른쪽의 그래프는 처음 두 쌍의 정준변수에 대하여 각 관측치의 위치를 나타낸 것이다. 관측치들이 분포되어있는 모습을 통해 U1과 U2의 상관관계가 존재하지 않는다는 사실을 확인할 수 있다.

2-4. 상관성 검정 및 시각화

X 집단과 Y 집단 사이에 충분한 상관성이 존재하지 않으면 정준상관분석의 결과가 유의하다고 판단하기 어렵다. 따라서 앞서 진행한 정준상관분석이 유의한지 확인하기 위해 유의수준 0.05 하에서 상관성 검정을 진행하였다.

$$H_0 : \sum XY = \mathbf{0} \text{ (상관성이 없다.)}$$

$$H_1 : \sum XY \neq \mathbf{0} \text{ (상관성이 있다.)}$$

$$\Lambda = \frac{|S|}{|S_{XX}||S_{YY}|} = \frac{|R|}{|R_{XX}||R_{YY}|} = \Pi_{i=1}^{p}(1 - \hat{\rho}_i^{*2})$$

$$-\left(n - \frac{1}{2(p+q+3)}\right) log\Lambda \sim \chi_{pq}^2$$

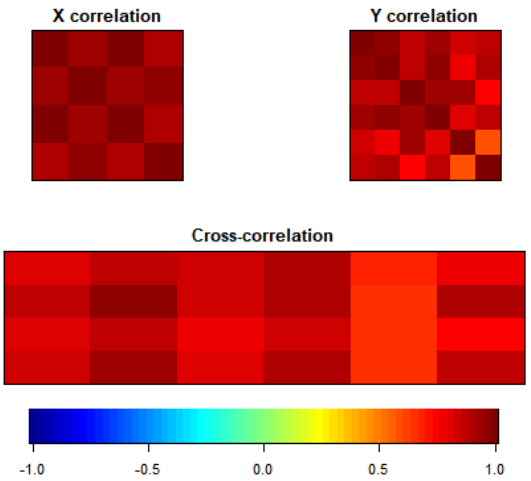
위의 식을 통해 얻은 검정통계량(Λ) 및 χ_{pq}^2와 p-value는 다음과 같다.

```
n = dim(X)[1]
p = dim(X)[2]
q = dim(Y)[2]
rho = bird_cc$cor
alpha = 0.05
Lambda = prod(1-rho**2)


# 통계량
chi_t = -(n-(p+q+3)/2)*log(Lambda)

# chi-square
chi = qchisq(1-alpha, p*q)

# p-value
p_value = 1-pchisq(chi_t, p*q)
c(chi_t, chi, p_value)

## [1] 1423.28236   36.41503    0.00000
```

term • $log\Lambda(1423.28) > \chi_{pq}^2(36.42)$ 이므로 귀무가설을 기각하고, X와 Y 사이에 상관성이 존재한다는 결론을 얻을 수 있다.

다음은 X의 변수들과 Y의 변수들 간의 상관관계를 한 눈에 볼 수 있도록 시각화 한 결과이다.

```
mtc = matcor(X,Y)
img.matcor(mtc,type=2)
```



푸른색에 가까울수록 강한 음의 상관관계를, 붉은색에 가까울수록 강한 양의 상관관계를 나타낸다. X 변수 간의 상관관계, Y 변수 간의 상관관계, X와 Y 변수 간의 상관관계를 나타내는 색이 모두 붉은색이므로, 상관성 검정 결과에서 확인했듯이 두 집단 간의 상관성이 매우 높다는 것을 알 수 있다.

2-5. 요약 및 의미

해당 분석을 통해 새의 날개 뼈(X)와 다리 뼈(Y) 간의 상관관계를 조사하였다. 정준상관분석 결과, X, Y 와의 상관관계가 가장 높은 첫 번째 정준변수의 정준상관계수는 0.96으로, 이를 통해 두 집단 간의 관계의 대부분을 설명할 수 있다. 특히, 윗날개뼈 지름(X2, humw)과 넓적다리뼈 지름(Y2, femw), 정강발목뼈 길이(Y3, tibl)가 이러한 관계에서 중요한 역할을 하는 것으로 나타났다.

IV. 결론

본 항목에서는 정준상관분석을 활용하여 새의 뼈의 길이와 지름 간의 관계, 그리고 날개 뼈와 다리 뼈 간의 관계를 조사하였다. 분석 결과, 새의 생물학적 구조에서 뼈의 길이와 지름이 상호 영향을 미치며, 날개 뼈와 다리 뼈의 구조 또한 상호 연관성이 있음을 확인할 수 있었다.

다만, 본 분석에서 사용한 데이터는 변수들 간의 상관성이 높아 추가적으로 유의미한 결과를 도출하는 데 어려움이 있었다. 따라서, 분석 진행 전 예측한 것과 같이 이번 분석은 EDA에서 확인된 결과를 재확인하는 수준에 그쳤다.

Chapter 2. 군집분석(Clustering)

작성자: 차수빈

I. 서론

군집분석은 군집의 개수나 구조에 대한 가정 없이 다변량 데이터로부터 거리 기준에 의해 자발적인 군집화를 유도하는 다변량 분석 방법이다. 본 항목에서는 데이터에 본래 존재하는 범주인 type을 제외한 10개의 변수들(뼈의 길이, 뼈의 지름)을 활용하여 군집분석을 수행하였다. 이를 통해 데이터들을 적절한 군집으로 나누어 각 군집의 특성, 군집 간의 차이 등에 대한 탐색적 분석을 수행하고, 해당 결과를 본래의 범주인 type과 비교해 보고자 한다.

II. 분석 설계 및 예측

본 항목에서는 다음의 목표를 가지고 군집분석을 진행하였다.

1. 같은 type에 속하는 데이터(= 새)들은 같은 군집에 속할 것이다.
2. 먼저 군집화 되는 type의 경우 다른 type과는 구분되는, 두드러지는 특징이 존재할 것이다.

앞서 진행한 EDA 결과를 통해 알 수 있듯이, 각 type별로 뼈의 길이와 지름의 분포가 상이한 것을 확인할 수 있었다. 따라서, 각 type별로 군집이 형성될 것이라 예측된다.

III. 군집분석 수행

```
library(tidyverse)

library(ggplot2)
```

0. 데이터 불러오기

```
bird <- read.csv('./data/bird_preprocessed.csv')

head(bird)
```

```
##     huml  humw  ulnal ulnaw  feml femw   tibl tibw  tarl tarw type
## 1  79.97  6.37  69.26  5.28 43.07 3.90  75.35 4.04 38.31 3.34   SW
## 2  77.65  5.70  65.76  4.77 40.04 3.52  69.17 3.40 35.78 3.41   SW
## 3  79.73  5.94  67.39  4.50 42.07 3.41  71.26 3.56 37.22 3.64   SW
## 4  86.98  5.68  74.52  4.55 44.46 3.78  76.02 3.81 37.94 3.81   SW
## 5 118.20  7.82 116.64  6.13 59.33 5.45 110.00 5.58 61.62 4.37   SW
## 6 145.00 10.42 144.00  7.05 70.96 7.44 120.00 7.31 78.67 6.34   SW
```

```
dim(bird)
```

```
## [1] 387  11
```

```
### feature 변수들만 선택
X <- bird[,-11]
```

1. 거리 계산

EDA 시 변수들 간의 스케일 차이가 큰 것을 확인할 수 있었다. 왜곡 정도가 매우 높은 데이터 세트의 경우 변별력이 없는 군집화가 수행될 수 있기에, 변수들 간의 스케일 차이를 반영한 표준화 거리를 활용하기로 결정하였다.

```
# 데이터 표준화
X_scaled <- scale(X)

# 표준화 거리 계산
d = dist(X_scaled)
```

2. 군집화 방법 선택

군집화 방법은 크게 계층적 군집화 방식과 비계층적 군집화 방식으로 구분된다.

- 계층적 군집방법(hierarchical clustering)
  - n개의 군집에서 시작하여 점차 군집의 개수를 줄여나가는 방법
  - 군집간 거리에 대한 정의가 필요
  - 최단연결법, 최장연결법, 평균연결법이 존재
- 비계층적 군집화
  - 그룹의 개수 g를 정하고 n개의 개체를 g개의 군집으로 나누는 모든 가능한 방법을 점검하여 최적의 군집을 형성하는 방법

각각의 군집화 방법을 적용해 본 후, 군집별 특성이 더 두드러지는 군집화 방법을 선택하고자 한다.

2-1. 계층적 군집화

최단연결법

```
### 군집화 수행
hc1 <- hclust(d^2, method = "single")

### 결과 확인
table(cutree(hc1, k = 5), bird$type)
```

```
##
##        P   R  SO  SW   W
##   1   38  50 128 101  62
##   2    0   0   0   3   0
##   3    0   0   0   2   0
##   4    0   0   0   0   2
##   5    0   0   0   0   1
```

- 대부분의 데이터가 하나의 군집으로 군집화 되는 것을 확인할 수 있다.

최장연결법

```
### 군집화 수행
hc2 <- hclust(d^2, method = "complete")

### 결과 확인
table(cutree(hc2, k = 5), bird$type)

##
##        P   R  SO  SW   W
##   1   38  18 128  64  55
##   2    0  32   0  33   8
##   3    0   0   0   5   0
##   4    0   0   0   4   0
##   5    0   0   0   0   2
```

- 군집 1, 군집 2 이후에 생성되는 군집들의 크기가 매우 작은 것을 확인할 수 있다.
    - 군집 개수 조정이 필요함을 시사한다.

평균연결법

```
### 군집화 수행
hc3 <- hclust(d^2, method = "average")

### 결과 확인
table(cutree(hc3, k = 5), bird$type)

##
##        P   R  SO  SW   W
##   1   38  30 128  75  61
##   2    0  20   0  22   2
##   3    0   0   0   5   0
##   4    0   0   0   4   0
##   5    0   0   0   0   2
```

- 대부분의 데이터가 하나의 군집으로 군집화 되는 것을 확인할 수 있다.

2-2. 비계층적 군집화

K-Means

```
### 군집화 수행
bird_k <- kmeans(X_scaled, centers = 5)

### 결과 확인

table(bird_k$cluster, bird$type)

##
##        P   R  SO  SW   W
##    1  17  10  25  22  15
##    2   4  16   0  38  26
##    3  17   0 103  10  13
##    4   0  21   0  15  11
##    5   0   3   0  21   0
```

- 대부분의 데이터가 특정한 군집으로 군집화 됨을 확인할 수 있다.
- 다만, 일부 군집의 경우 지나치게 세분화되는 양상이 보인다.
    - 적절한 군집 개수의 선택이 필요하다고 판단된다.

⇒ 여러 군집화 방법 중 K-Means 군집화 방법을 활용하기로 결정하였다.

3. 군집 분리 과정 확인

3-1. type에 따른 분리 과정

- type SO의 경우 다른 type들과 비교적 잘 구분되는 것을 확인할 수 있다.
    - SO : Singing Birds (노래하는 조류)
    - 다른 type에 비해 전체적으로 뼈의 길이와 지름이 작은 type이다.
- 나머지 type들의 경우 모든 군집에 데이터가 일정 비율로 군집화 되는 것을 확인할 수 있다.

⇒ type을 하나씩 제거해 보며 군집 분리 과정을 확인해 보고자 한다.

a) type SO 제거

```
### SO 제거
bird_woSO = bird[bird$type != "SO",]
X_woSO = bird_woSO[,-11]
X_woSO_scaled = scale(X_woSO)

### 군집화
bird_woSO_k = kmeans(X_woSO_scaled, centers = 4)
table(bird_woSO_k$cluster, bird_woSO$type)

##
##      P  R SW  W
##   1 33  9 26 26
##   2  5 11 39 27
##   3  0 26 18 10
##   4  0  4 23  2
```

- SO를 제외한 4개의 type 중에서는 P가 다른 type들과 비교적 잘 분리되는 것을 확인할 수 있다.

b) type P 제거

```
### P 제거
bird_woP = bird_woSO[bird_woSO$type != "P",]
X_woP = bird_woP[,-11]
X_woP_scaled = scale(X_woP)

### 군집화
bird_woP_k = kmeans(X_woP_scaled, centers = 3)
table(bird_woP_k$cluster, bird_woP$type)

##
##      R SW  W
##   1 19 44 31
##   2 21 33  8
##   3 10 29 26
```

- SO, P를 제외한 3개의 type들 간에는 군집의 분리에 영향을 미치는 차이점을 찾기 어렵다.
    - 각 군집에 데이터가 골고루 퍼져 있음을 확인할 수 있다.

3-2. 변수 선택에 따른 분리 과정

데이터에 포함된 뼈는 크게 날개뼈와 다리뼈로 구분할 수 있다. 따라서, 모든 변수를 활용하는 대신 일부 변수만 선택적으로 활용하여 군집분석을 수행해 보았다.

a) 날개뼈 관련

```
### 날개뼈와 관련된 변수만 선택
# huml, humw, ulnal, ulnaw
X_wing <- bird[,c(1:4)]
X_wing_scaled <- scale(X_wing)

### 군집화
wing_k = kmeans(X_wing_scaled, centers = 5)
table(wing_k$cluster, bird$type)

##
##       P   R  SO  SW   W
##   1   0   0   0  13   1
##   2   0  18   0  41  19
##   3   0  16   0  21  10
##   4  17  12   6  19  19
##   5  21   4 122  12  16
```

- type P와 SO에 속하는 데이터의 경우 비교적 다른 type과 잘 분리되는 것을 확인할 수 있다.
- type SW에 속하는 데이터의 경우 모든 군집에 걸쳐 산발적으로 분포함을 확인할 수 있다.

b) 다리뼈 관련

```
### 다리뼈와 관련된 변수만 선택
# feml, femw, tibl, tibw, tarl, tarw
X_leg <- bird[,c(5:10)]
X_leg_scaled <- scale(X_leg)

### 군집화
leg_k = kmeans(X_leg_scaled, centers = 5)
table(leg_k$cluster, bird$type)

##
##      P  R SO SW  W
##   1  0 11  0 12  2
##   2 21  0 99 10 15
##   3 12 10 26 27 18
##   4  5  9  3 28 20
##   5  0 20  0 29 10
```

- type P와 SO에 속하는 데이터의 경우 비교적 다른 type과 잘 분리되는 것을 확인할 수 있다.
- type SW에 속하는 데이터의 경우 모든 군집에 걸쳐 넓게 분포함을 확인할 수 있다.

c) 날개뼈 vs 다리뼈

- 날개뼈와 관련된 변수만을 활용하여 군집화를 수행할 때가 다리뼈와 관련된 변수만을 활용하여 군집화를 수행할 때보다 군집이 더 잘 분리되는 것을 확인할 수 있다.
    - 새의 생태학적 특징을 구분할 때 날개뼈의 차이가 다리뼈의 차이보다 좀 더 두드러진다고 말할 수 있다.

4. 적절한 군집 개수 설정

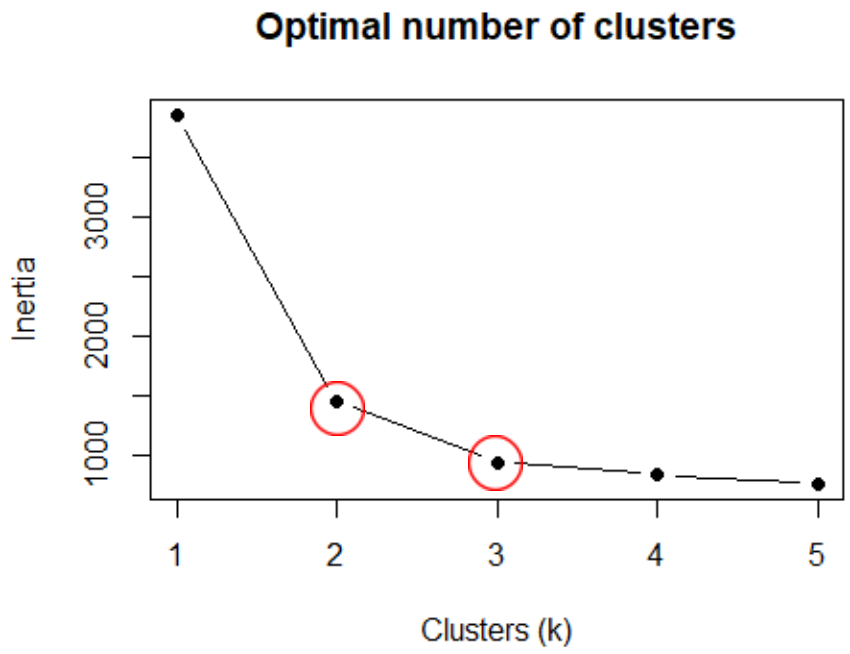
군집의 개수를 type과 동일하게 5개로 설정하니 군집이 지나치게 세분화되는 경향을 보인다.
따라서, 군집의 개수를 조정하여 어떠한 type이 비슷한지, 또 다른지 비교해 보고자 한다.

아래와 같이 이너셔(Inertia) 방법을 활용하여 적절한 군집 개수를 선택하고자 한다.

- K-Means 군집화의 성능 지표를 이너셔(Inertia)라고 한다.
    - 이너셔는 각 샘플과 가장 가까운 센트로이드 사이의 평균 제곱 거리를 측정한 수치이다.
- 클러스터 수와 이너셔는 반비례 관계에 있다.
    - 클러스터 수가 늘어감에 따라 이너셔가 급격하게 감소하고 어느 지점에서는 완만하게 감소한다.
    - 이너셔 감소 기울기가 급격하게 변하는 지점이 있는데, 이를 엘보우라고 하고 해당 지점 근처를 최적 군집 수로 결정한다.

4-1. Elbow Plot

```r
### Elbow plot을 그리는 함수 정의
plot_elbow <- function(data, max_k) {
  wcss <- numeric(max_k)

  # 군집 수에 따른 WCSS 계산
  for (i in 1:max_k) {
    kmeans_model <- kmeans(data, centers = i)
    wcss[i] <- kmeans_model$tot.withinss
  }

  # Elbow plot 그리기
  plot(1:max_k, wcss,
       type = "b", pch = 19,
       xlab = "Clusters (k)", ylab = "Inertia",
       main = "Optimal number of clusters")
}

## 최대 군집 수 설정
max_clusters <- 5
## Elbow plot 그리기
plot_elbow(X_scaled, max_clusters)
```



⇒ **2 ~ 3개**의 군집이 가장 적절해 보인다.

a) 2개 군집으로 군집화

```
### 군집화 수행
kmeans_2 <- kmeans(X_scaled, centers = 2)

### 결과 확인
table(kmeans_2$cluster, bird$type)
```

```
##
##        P   R  SO  SW   W
##   1    0  36   0  50  21
##   2   38  14 128  56  44
```

b) 3개 군집으로 군집화

```
### 군집화 수행
kmeans_3 <- kmeans(X_scaled, centers = 3)

### 결과 확인
table(kmeans_3$cluster, bird$type)
```

```
##
##        P   R  SO  SW   W
##   1    6  20   4  47  30
##   2   32   8 124  26  26
##   3    0  22   0  33   9
```

3개 군집으로 군집화 하는 경우, 지나치게 세분화되는 경향이 일부 존재하기에, 2개의 군집으로 군집화 하는 것이 더 적절하다고 판단하였다.

5. 최종 군집화

- 최종적으로 **2개**의 군집으로 군집화 후 결과를 분석해 보았다.

5-1. 차원 축소

- 10차원의 데이터를 2차원 평면에 시각화 하기 위해 2차원으로 차원 축소를 진행하였다.

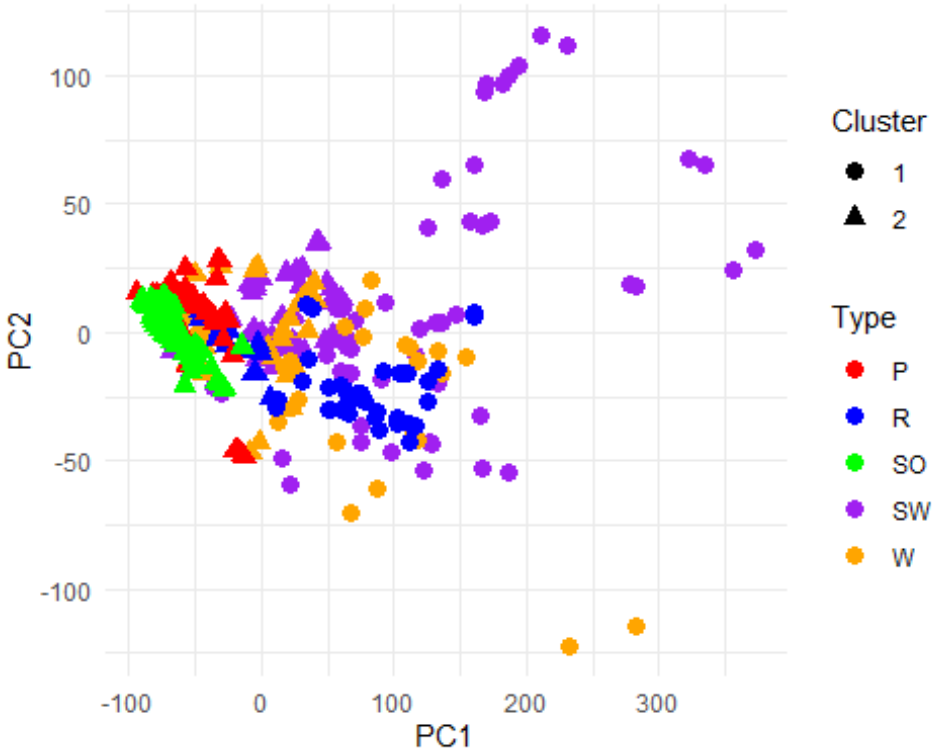
```
# PCA를 통한 차원 축소
pca_result <- princomp(bird[,-11], cor = FALSE)

# 원래 데이터프레임에 PCA 결과(주성분 점수) 추가
bird$pca_comp1 <- pca_result$scores[, 1] # Comp 1
bird$pca_comp2 <- pca_result$scores[, 2] # Comp 2
```

5-2. 결과 시각화

```
### 군집화 결과 저장

bird$cluster <- kmeans_2$cluster
head(bird)
```

```
##     huml  humw  ulnal ulnaw  feml femw   tibl tibw  tarl tarw type
pca_comp1
## 1  79.97  6.37  69.26  5.28 43.07 3.90  75.35 4.04 38.31 3.34   SW
15.370184
## 2  77.65  5.70  65.76  4.77 40.04 3.52  69.17 3.40 35.78 3.41   SW
8.288697
## 3  79.73  5.94  67.39  4.50 42.07 3.41  71.26 3.56 37.22 3.64   SW
12.047802
## 4  86.98  5.68  74.52  4.55 44.46 3.78  76.02 3.81 37.94 3.81   SW
23.412076
## 5 118.20  7.82 116.64  6.13 59.33 5.45 110.00 5.58 61.62 4.37   SW
89.857975
## 6 145.00 10.42 144.00  7.05 70.96 7.44 120.00 7.31 78.67 6.34   SW
133.289819
##    pca_comp2 cluster
## 1  -6.400217       2
## 2  -2.142498       2
## 3  -3.694797       2
## 4  -3.389881       2
## 5 -18.089891       1
## 6 -19.746065       1
```

```r
# 시각화
ggplot(bird, aes(x = pca_comp1, y = pca_comp2,
                 color = type, shape = as.factor(cluster))) +
  geom_point(size = 3) +
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange")) +
  scale_shape_manual(values = c(19, 17)) +
  labs(x = "PC1", y = "PC2", color = "Type", shape = "Cluster") +
  theme_minimal()
```




```r
table(kmeans_2$cluster, bird$type)
```

```
##
##        P   R  SO  SW   W
##   1    0  36   0  50  21
##   2   38  14 128  56  44
```

5-3. 결과 해석

- type P와 type SO는 다른 type들과 확연하게 구분된다.
    - P: Scansorial Birds (산악지대에 서식하는 조류)
    - SO : Singing Birds (노래하는 조류)
    - 이들은 다른 type에 비해 전체적으로 뼈의 길이와 지름이 작은 종류이다.
- type SW의 경우 군집 1과 군집 2에 속하는 데이터의 비율이 거의 50:50이다.
    - SW: Swimming Birds (수영하는 조류)
    - 현재 데이터에서 3분의 1에 해당하는 조류의 type이 SW이다.
    - 또한, 모든 변수들에 대해 SW의 데이터 분포가 넓었던 점을 감안할 때 적절한 군집화 결과라고 판단된다.
- type R의 경우 대부분의 데이터가 군집 1에 속하였고, 일부 데이터만이 군집 2에 속함을 확인할 수 있다.
    - R: Raptors (사냥하는 조류)
    - type R에 해당하는 새들의 경우 대부분 뼈의 길이와 지름이 크다는 특징을 지니는데, 해당 새들 중 일부 작은 새들이 군집 2로 분류된 것이라고 판단할 수 있다.
- type W의 경우 데이터의 3분의 2가 군집 2에 속하였고, 일부 데이터만이 군집 1에 속함을 확인할 수 있다.
    - W: Wading Birds (물가에 서식하는 조류)
    - EDA 시 boxplot 시각화 결과로 미루어 볼 때, type W에 속하는 데이터들 중 일부 뼈의 길이와 지름이 큰 새들이 존재함을 확인할 수 있었다.
    - 따라서, 비교적 크기가 큰 새들이 군집 1로 군집화 되었음을 짐작할 수 있다.

Chapter 3. 주성분분석(PCA)

작성자: 김경민

I. 서론

주성분 분석은 P차원의 데이터를 P보다 작은 m개의 선형결합식으로 설명하고자 하는 방법이다. PCA를 통해 정보의 손실은 최소화하면서 더 작은 차원에서 데이터를 해석할 수 있다. 우리의 데이터는 새의 골격에 대한 정보를 담고 있는 10개의 변수와, 새를 5가지의 생태학적 그룹으로 분류하는 'type' 변수로 이루어져 있다. 먼저 데이터를 type별로 분리한 후에, 10차원의 자료를 잘 설명하는 주성분을 찾아 그 의미를 해석해보고자 했다. 추가로 전체 데이터에 대해 PCA를 수행하고 2차원으로 축소된 데이터를 가지고 분류 분석도 진행해보았다. 이때 뼈의 길이(l)와 지름(w)의 범위가 서로 다르기 때문에 분산-공분산 행렬 대신 상관행렬 R을 이용하였다.

II. 주성분분석 수행

0. 데이터 불러오기

```
library(ggplot2)

df = read.csv("./bird_preprocessed.csv")
head(df)
```

```
##     huml  humw  ulnal ulnaw  feml femw   tibl tibw  tarl tarw type
## 1  79.97  6.37  69.26  5.28 43.07 3.90  75.35 4.04 38.31 3.34   SW
## 2  77.65  5.70  65.76  4.77 40.04 3.52  69.17 3.40 35.78 3.41   SW
## 3  79.73  5.94  67.39  4.50 42.07 3.41  71.26 3.56 37.22 3.64   SW
## 4  86.98  5.68  74.52  4.55 44.46 3.78  76.02 3.81 37.94 3.81   SW
## 5 118.20  7.82 116.64  6.13 59.33 5.45 110.00 5.58 61.62 4.37   SW
## 6 145.00 10.42 144.00  7.05 70.96 7.44 120.00 7.31 78.67 6.34   SW
```

1. type별 분석

1-1. 'type' P(Scansorial Birds, 산악지대 서식 조류)

```
df.P = df[df$type=="P",-11]
dim(df.P)
```

```
## [1] 38 10
```

```
R = cor(df.P)
eigen(R)$values
```

```
##  [1] 7.556143491 1.830681050 0.275156540 0.142299020 0.100146273 0.043686551
##  [7] 0.024679633 0.012654265 0.010582504 0.003970673
```

```
PC.result.P = princomp(df.P, cor=TRUE)
summary(PC.result.P)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.7488440 1.3530266 0.52455366 0.3772254 0.31645896
## Proportion of Variance 0.7556143 0.1830681 0.02751565 0.0142299 0.01001463
## Cumulative Proportion  0.7556143 0.9386825 0.96619811 0.9804280 0.99044264
##                           Comp.6      Comp.7      Comp.8     Comp.9
## Standard deviation     0.209013279 0.157097527 0.112491177 0.10287130
## Proportion of Variance 0.004368655 0.002467963 0.001265426 0.00105825
## Cumulative Proportion  0.994811292 0.997279256 0.998544682 0.99960293
##                           Comp.10
## Standard deviation     0.0630132759
## Proportion of Variance 0.0003970673
## Cumulative Proportion  1.0000000000
```

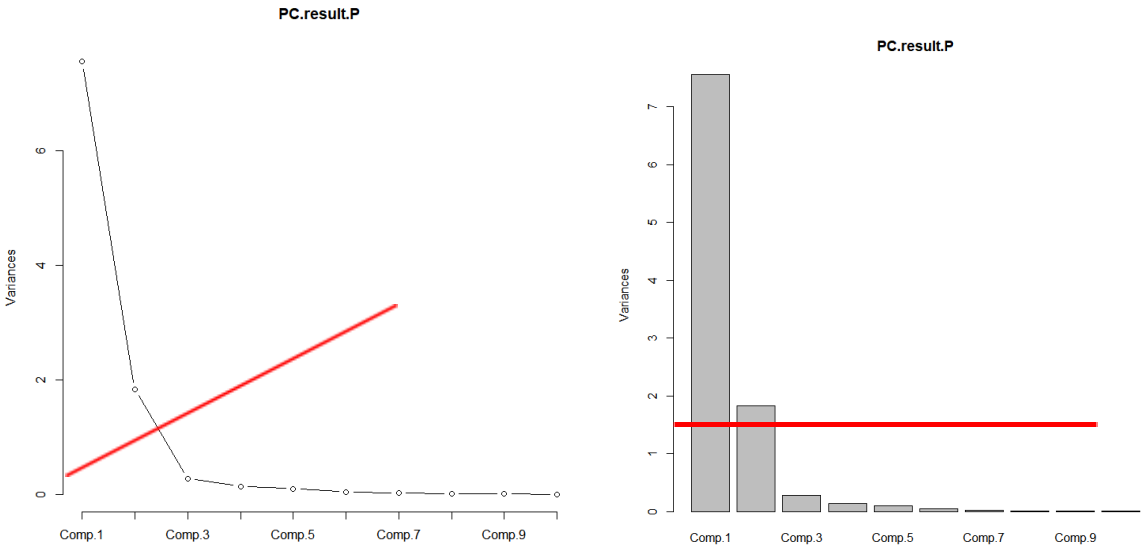
상관행렬 R을 이용한 PCA를 진행하기 위해 princomp의 cor 옵션을 TRUE로 설정하였다. 먼저 주성분 분석의 요약 정보를 확인해보자. 첫 번째 주성분이 전체 변동 중 75.6%를 설명하고 있고, 두 번째 주성분이 전체 변동 중 18.3%를 설명하고 있다. 첫 번째 주성분만으로도 전체 데이터를 충분히 설명할 수 있다는 의미이다. 상관행렬의 첫 번째 고유값이 두 번째 고유값보다 약 4배 이상 크다는 것 역시 같은 의미로 해석할 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이는 변수들 간 공선성 문제가 존재함을 의미한다. 모든 변수들 간 상관관계가 높은 데이터이기 때문에 PCA에서도 공선성 문제를 확인할 수 있었다.

```
PC.result.P$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.315  0.307  0.422  0.345         0.150  0.556  0.227  0.284  0.210
## humw   0.267  0.445 -0.283 -0.572  0.539         0.115
## ulnal  0.223  0.563  0.264  0.253        -0.439 -0.387 -0.206 -0.213 -0.245
## ulnaw  0.336  0.233        -0.222 -0.442  0.670 -0.361
## feml   0.342 -0.232                0.293  0.288  0.202 -0.329        -0.704
## femw   0.348 -0.161        -0.206 -0.432 -0.196  0.416        -0.618  0.152
## tibl   0.331 -0.276  0.160  0.171  0.369  0.118 -0.313 -0.389 -0.122  0.587
## tibw   0.345 -0.194        -0.205 -0.284 -0.384        -0.307  0.685
## tarl   0.301 -0.375  0.387 -0.198  0.122 -0.194 -0.281  0.651        -0.156
## tarw   0.330        -0.695  0.528                       0.337
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                Comp.10
## SS loadings        1.0
## Proportion Var     0.1
## Cumulative Var     1.0
```

다음 loadings를 이용해서 주성분의 의미를 분석해보자. 첫 번째 주성분은 뼈 크기의 가중평균을 나타낸다. 전반적인 골격 크기의 주요 변동을 나타내는 축이다. 두 번째 주성분은 huml(위날개뼈 길이), humw(위날개뼈 지름), ulnal(자뼈 길이)과 tarl(뒷발목뼈 길이)의 대비를 나타낸다. 이때 양의 부호를 가진 변수들은 날개와 관련된 뼈이고 음의 부호를 가진 변수들은 다리와 관련된 뼈이므로, 두번째 주성분을 날개뼈와 다리뼈의 대비를 나타내는 축이라고 분석했다. 2개의 주성분만으로 전체 변동의 93.6%를 설명할 수 있기 때문에 두번째 주성분까지만 분석을 진행했다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : 날개 관련 뼈와 다리 관련 뼈의 대비를 나타내는 축

```
screeplot(PC.result.P, type="lines")
screeplot(PC.result.P)
```

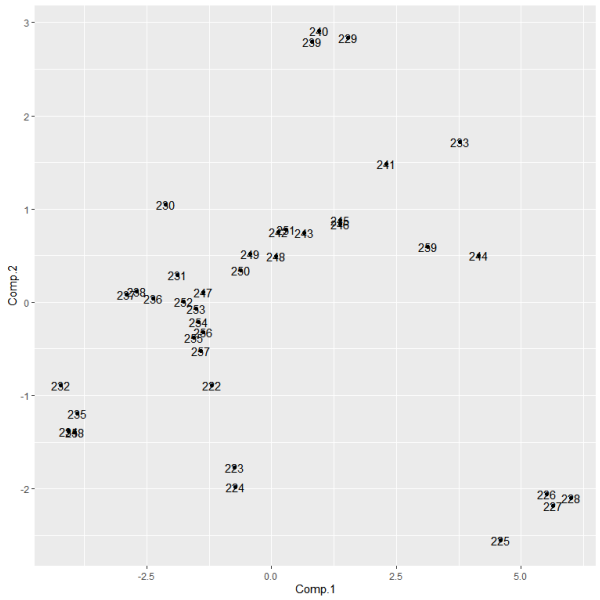


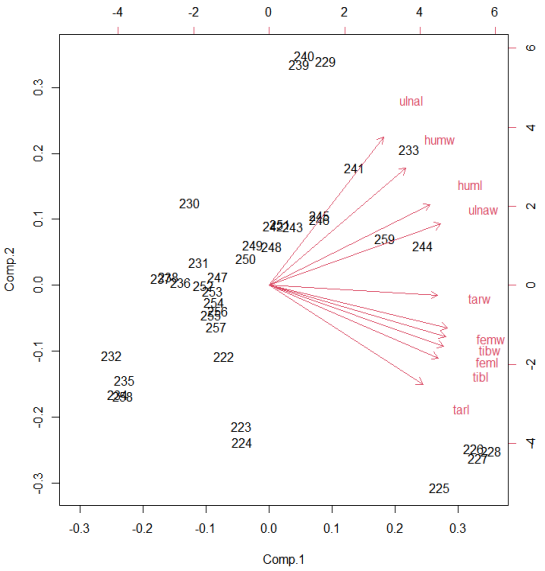
screeplot을 이용해 적절한 주성분 개수를 찾아보자. 왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 급격히 감소하고, Comp.2와 Comp.3사이에서도 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1과 Comp.2의 분산이 1 이상임을 확인할 수 있다. 첫번째 주성분이 전체 변동의 70% 이상을 설명하지만 고유값이 1이상인 Comp.2까지 두 개의 주성분을 선택한다.

```
head(PC.result.P$scores)
```

```
##         Comp.1     Comp.2      Comp.3     Comp.4      Comp.5       Comp.6
## 222 -1.1998340 -0.8915445  0.04506642 -0.1465436  0.10766516  0.350707326
## 223 -0.7351940 -1.7760515  0.19658082  0.4465844  0.61107129  0.045851772
## 224 -0.7255312 -1.9814966  0.15812951  0.5598719  0.38750504 -0.009273193
## 225  4.5872835 -2.5560772  0.39103773  0.2133254  0.13074810 -0.077058927
## 226  5.5169239 -2.0593428  0.49090967 -0.3042561 -0.07176955  0.530257477
## 227  5.6390168 -2.1799861 -0.24273940 -0.1817185  0.08382362 -0.262775853
##          Comp.7       Comp.8       Comp.9       Comp.10
## 222  0.11631976  0.032042272  0.041486731 -0.0585944687
## 223  0.15959929  0.119575395 -0.039207475  0.0005003072
## 224  0.14418267  0.085183153  0.008748827  0.0799752824
## 225  0.02083917 -0.095117305  0.029363307 -0.0761322499
## 226 -0.06767612  0.002985729 -0.118701058 -0.0477051178
## 227  0.17515317 -0.057700874  0.044856442  0.1313161147
```

```
ggplot(data.frame(PC.result.P$scores), aes(Comp.1,Comp.2))+
  geom_point()+
  geom_text(aes(label=rownames(df.P)))

biplot(PC.result.P)
```





주성분을 이용해 자료를 해석해보자. PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 날개뼈에 비해 다리뼈(특히, 뒷발목뼈)가 큰 개체, 값이 크면 날개뼈에 비해 다리뼈가 작은 개체라고 해석할 수 있다. 225, 226, 227, 228번 개체는 상대적으로 PC1 값이 크고 PC2 값이 작으므로 골격이 크고 다리뼈가 더 발달한 개체이다. 반대로 230번 개체는 상대적으로 PC1값이 작고 PC2 값이 크므로 골격이 작고 날개뼈가 더 발달한 개체이다.

biplot을 통해서, PC1 축에 가장 영향을 많이 미치는 변수는 ‘tarw’, 영향을 덜 미치는 변수는 ‘ulnal’과 ‘tarl’ 임을 확인할 수 있다. 반대로 PC2 축에 영향을 많이 미치는 변수는 ‘ulnal’, ‘tarl’이고 덜 미치는 변수는 ‘tarw’ 이다.

1-2. ‘type’ R(Raptors, 사냥하는 조류)

```
df.R = df[df$type=="R",-11]
dim(df.R)

## [1] 50 10

R = cor(df.R)
eigen(R)$values

##  [1] 8.900045739 0.642965684 0.211850916 0.086287521 0.064001416 0.045574689
##  [7] 0.024664227 0.013836826 0.006093018 0.004679966

PC.result.R = princomp(df.R, cor=TRUE)
summary(PC.result.R)

## Importance of components:
##                            Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
## Standard deviation     2.9832944 0.80185141 0.46027265 0.293747376 0.252985011
## Proportion of Variance 0.8900046 0.06429657 0.02118509 0.008628752 0.006400142
## Cumulative Proportion  0.8900046 0.95430114 0.97548623 0.984114986 0.990515127
##                            Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     0.213482292 0.157048487 0.117630038 0.0780577840
## Proportion of Variance 0.004557469 0.002466423 0.001383683 0.0006093018
## Cumulative Proportion  0.995072596 0.997539019 0.998922702 0.9995320034
##                           Comp.10
## Standard deviation     0.0684102741
## Proportion of Variance 0.0004679966
## Cumulative Proportion  1.0000000000
```

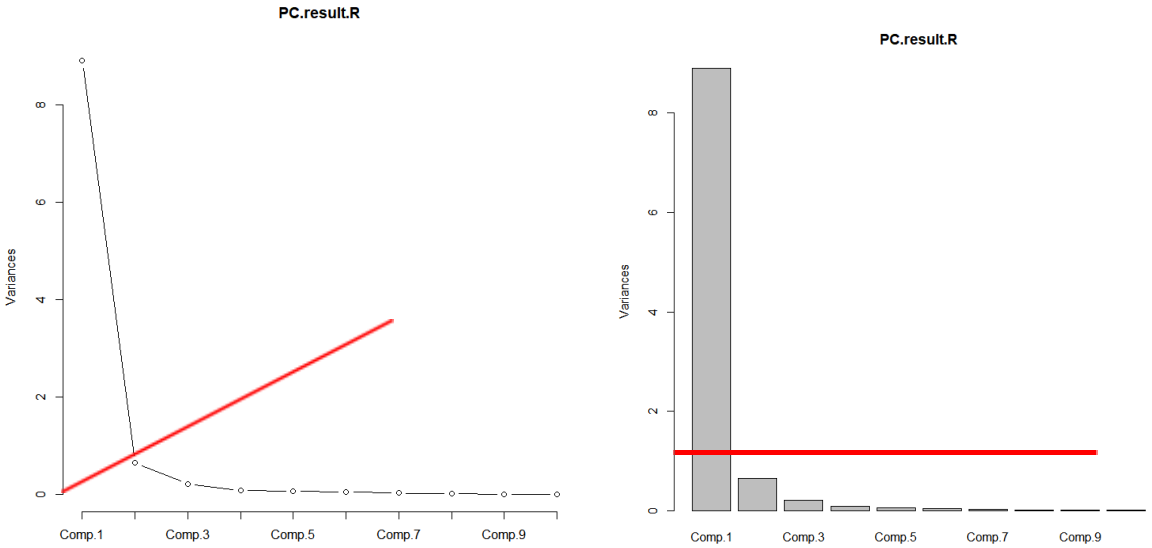
주성분 분석의 요약 정보를 확인해보자. 첫 번째 주성분이 전체 변동 중 89.0%를, 두 번째 주성분이 전체 변동 중 6.4%를 설명하고 있다. type R 데이터 역시 첫 번째 주성분 만으로도 전체 데이터를 충분히 설명할 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이는 변수들 간 공선성 문제가 존재함을 의미한다. 모든 변수들 간 상관관계가 높은 데이터이기 때문에 역시 공선성 문제가 확인되었다.

```
PC.result.R$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.324  0.138  0.389  0.400  0.206  0.124  0.117  0.435  0.523  0.176
## humw   0.327  0.179 -0.134         0.533                0.386 -0.629 -0.108
## ulnal  0.314  0.114  0.551 -0.744 -0.106  0.128
## ulnaw  0.326  0.150 -0.252 -0.135  0.405 -0.374 -0.406 -0.400  0.401
## feml   0.325 -0.135 -0.326        -0.427  0.324 -0.583  0.334        -0.164
## femw   0.326        -0.431 -0.197         0.160  0.374                0.699
## tibl   0.326 -0.106  0.287  0.420         0.377 -0.149 -0.593 -0.312
## tibw   0.331        -0.245                0.178  0.555 -0.164  0.181 -0.656
## tarl   0.240 -0.859  0.146               -0.409
## tarw   0.312  0.378         0.229 -0.551 -0.595               -0.185
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                Comp.10
## SS loadings        1.0
## Proportion Var     0.1
## Cumulative Var     1.0
```

주성분의 의미를 간략하게 정리하면 다음과 같다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : tarl(뒷발목뼈 길이)과 tarw(뒷발목뼈 지름)의 대비를 나타내는 축

```r
screeplot(PC.result.R, type="lines")

screeplot(PC.result.R)
```

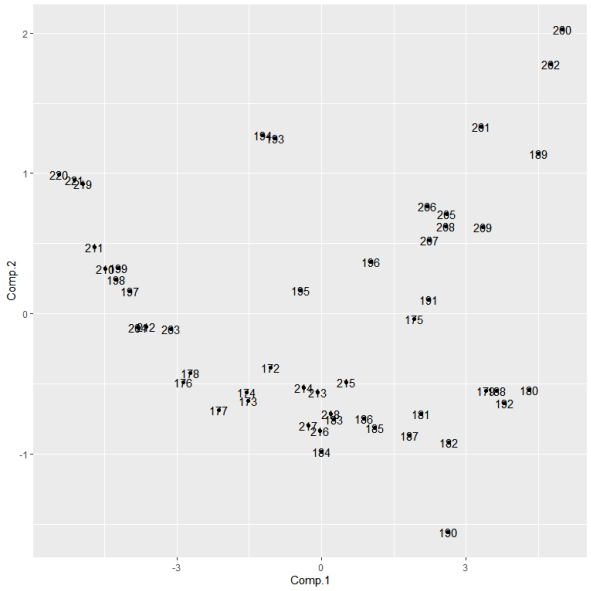


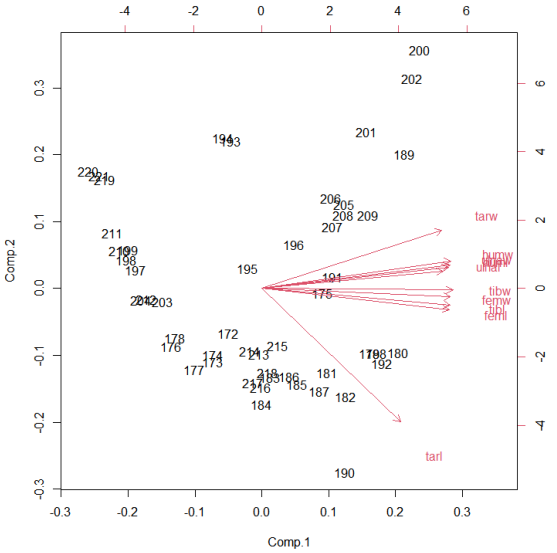
왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1의 분산만이 1 이상임을 확인할 수 있다. 첫번째 주성분만으로 전체 변동 중 89.0%를 설명할 수 있으므로 1개의 주성분을 선택한다.

```r
head(PC.result.R$scores)
```

```
##         Comp.1      Comp.2      Comp.3      Comp.4       Comp.5       Comp.6
## 172 -1.060872 -0.38323053 -0.10284277 -0.67284000 -0.007120978 -0.087743360
## 173 -1.527357 -0.62155834 -0.49175781  0.05807457 -0.117497775 -0.004708917
## 174 -1.547632 -0.56129931 -0.38368748  0.10687939 -0.113763722 -0.039211354
## 175  1.920812 -0.03998966 -0.79032735 -0.11257418  0.134941130  0.810089582
## 176 -2.863553 -0.48947061 -0.07599983 -0.02520934  0.157533768 -0.094369070
## 177 -2.124411 -0.68992994 -0.23552556  0.06539484 -0.119994589  0.070624151
##          Comp.7       Comp.8      Comp.9     Comp.10
## 172 -0.17559369 -0.243014732  0.19878468  0.01179980
## 173 -0.06794481 -0.222273080 -0.12330473  0.02352340
## 174 -0.03870859 -0.074596410 -0.04171132 -0.03631187
## 175 -0.02845754 -0.002206784 -0.03176685 -0.09189173
## 176 -0.16856643  0.149105943 -0.11671755 -0.02507428
## 177 -0.14422259 -0.027411663 -0.09167006  0.03230665
```

```r
ggplot(data.frame(PC.result.R$scores), aes(Comp.1,Comp.2))+
  geom_point()+
  geom_text(aes(label=rownames(df.R)))

biplot(PC.result.R)
```





PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 뒷발목뼈의 길이가 지름보다 긴 개체, 값이 크면 뒷발목뼈의 길이가 지름보다 짧은 개체라고 해석할 수 있다. 200번 개체는 상대적으로 PC1과 PC2 값이 크므로 골격이 크고 뒷발목뼈의 길이가 지름보다 짧은, 즉 짧고 두꺼운 뒷발목뼈를 가진 개체이다. 190번 개체는 상대적으로 PC2 값이 작으므로 길고 얇은 뒷발목뼈를 가진 개체이다.

biplot을 통해서, PC1 축에 가장 영향을 덜 미치면서 PC2 축에 가장 영향을 많이 미치는 변수가 ‘tarl’ 임을 확인할 수 있다.

1-3. ‘type’ SO(Singing Birds, 노래하는 조류)

```
df.SO = df[df$type=="SO",-11]
dim(df.SO)

## [1] 128  10

R = cor(df.SO)
eigen(R)$values

##  [1] 9.08504666 0.46423735 0.18221787 0.07078562 0.05678457 0.04993438
##  [7] 0.03797842 0.02296137 0.01871339 0.01134036

PC.result.SO = princomp(df.SO, cor=TRUE)
summary(PC.result.SO)

## Importance of components:
##                            Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
## Standard deviation     3.0141411 0.68134965 0.42686986 0.266055672 0.238295141
## Proportion of Variance 0.9085047 0.04642373 0.01822179 0.007078562 0.005678457
## Cumulative Proportion  0.9085047 0.95492840 0.97315019 0.980228750 0.985907208
##                             Comp.6      Comp.7      Comp.8      Comp.9
## Standard deviation     0.223460027 0.194880522 0.151530093 0.136796888
## Proportion of Variance 0.004993438 0.003797842 0.002296137 0.001871339
## Cumulative Proportion  0.990900646 0.994698488 0.996994625 0.998865964
##                            Comp.10
## Standard deviation     0.106491139
## Proportion of Variance 0.001134036
## Cumulative Proportion  1.000000000
```

첫 번째 주성분이 전체 변동 중 90.9%를, 두 번째 주성분이 전체 변동 중 4.6%를 설명하고 있다. type SO 데이터 역시 첫 번째 주성분 만으로도 전체 데이터를 충분히 설명할 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이를 통해 변수들 간 공선성 문제가 존재함을 확인할 수 있다.

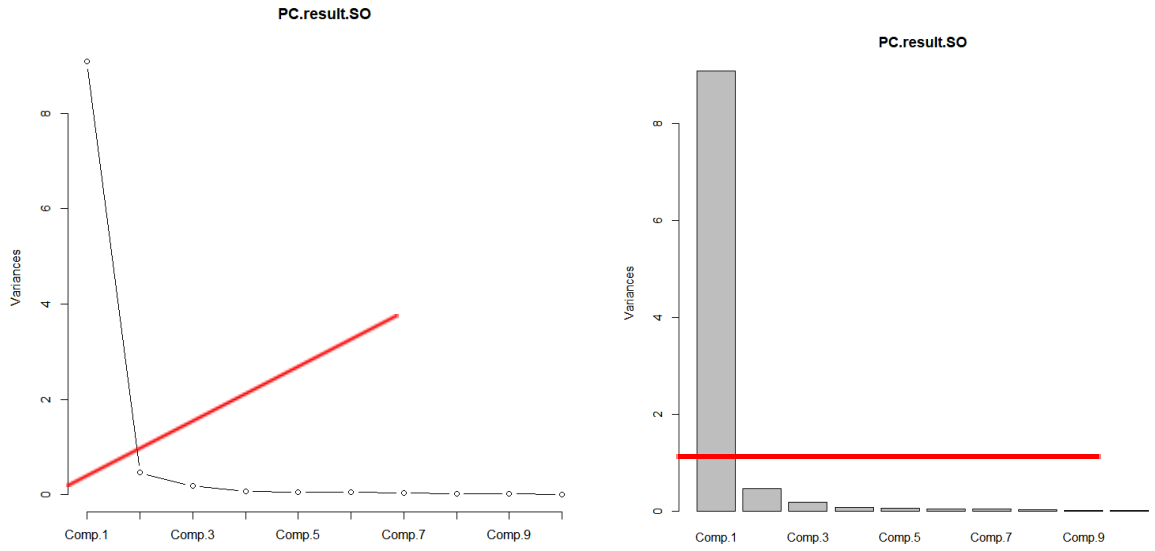
```
PC.result.SO$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.324  0.248         0.137         0.383         0.412  0.176  0.677
## humw   0.316  0.368  0.133 -0.159 -0.500         0.141 -0.643  0.153  0.111
## ulnal  0.312  0.436  0.202  0.270  0.121  0.273 -0.457        -0.298 -0.458
## ulnaw  0.312  0.414                0.272 -0.730  0.290  0.144
## feml   0.325 -0.145        -0.318  0.302  0.349  0.317         0.496 -0.458
## femw   0.324 -0.136 -0.110 -0.343 -0.535         0.175  0.468 -0.435 -0.146
## tibl   0.316 -0.331  0.319         0.421         0.266 -0.318 -0.542  0.209
## tibw   0.322 -0.165 -0.200 -0.487  0.161 -0.219 -0.678         0.102  0.197
## tarl   0.303 -0.456  0.493  0.414 -0.267 -0.252 -0.148  0.115  0.337
## tarw   0.307 -0.241 -0.732  0.498                      -0.220
##
```

```
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings        1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var     0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var     0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                 Comp.10
## SS loadings         1.0
## Proportion Var      0.1
## Cumulative Var      1.0
```

주성분의 의미를 간략하게 정리하면 다음과 같다. PC2는 날개뼈와 다리뼈의 대비를 나타내는 축이라고도 해석할 수 있다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : humw(윗날개뼈 길이), ulnal(자뼈 길이), ulnaw(자뼈 지름)과 tibl(정강발목뼈 길이), tarl(뒷발목뼈 길이)의 대비를 나타내는 축

```
screeplot(PC.result.SO, type="lines")

screeplot(PC.result.SO)
```

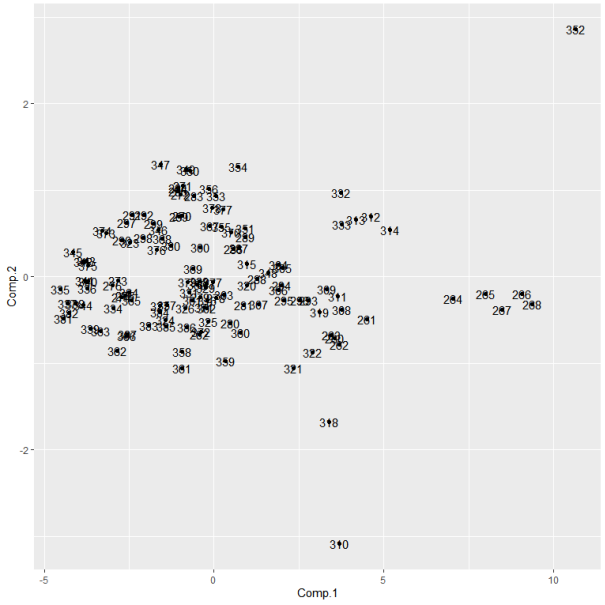


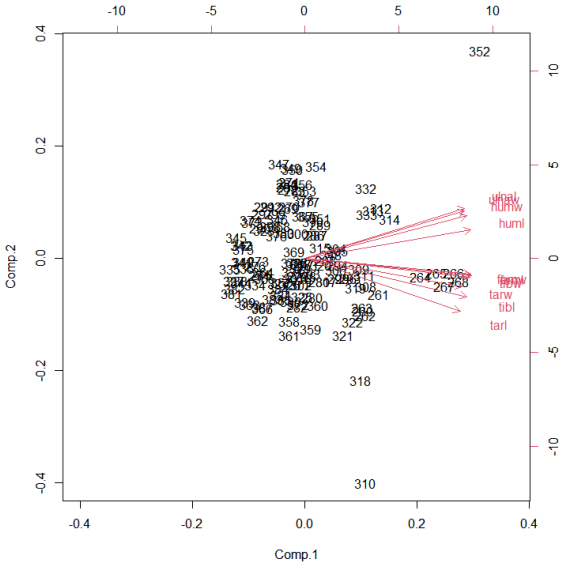
왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1의 분산만이 1 이상임을 확인할 수 있다. 첫번째 주성분만으로 전체 변동 중 90.9%를 설명할 수 있으므로 1개의 주성분을 선택한다.

```r
head(PC.result.SO$scores)
```

```
##        Comp.1     Comp.2     Comp.3      Comp.4      Comp.5     Comp.6
## 260 3.552646 -0.7138560  0.4596303  0.18868476  0.31602581  0.4576263
## 261 4.516321 -0.4929873  0.7590956 -0.38065206 -0.23055478  0.1901300
## 262 3.694503 -0.7799647  0.2670824 -0.06649617  0.12817638 -0.1404958
## 263 3.467779 -0.6711055  0.1852269  0.12067771  0.08764376  0.3268078
## 264 7.037138 -0.2538693 -0.3002376  0.04177805 -0.33307079  0.4401197
## 265 8.005176 -0.1966513  0.2138508 -0.46097746 -0.20334273 -0.2098904
##           Comp.7      Comp.8      Comp.9      Comp.10
## 260  0.089768546  0.11192726  0.05947184 -0.103662540
## 261 -0.086792994 -0.08697247 -0.01210837 -0.004692989
## 262 -0.046270302  0.14064681  0.02984570 -0.098191860
## 263  0.041800148 -0.02027326 -0.10165750 -0.064285989
## 264  0.005484937 -0.02993922  0.44116430 -0.170933125
## 265  0.104534807  0.01763407 -0.02132889  0.041103613
```

```r
ggplot(data.frame(PC.result.SO$scores), aes(Comp.1,Comp.2))+
  geom_point()+
  geom_text(aes(label=rownames(df.SO)))

biplot(PC.result.SO)
```





PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 날개뼈에 비해 다리뼈가 큰 개체, 값이 크면 날개뼈에 비해 다리뼈가 작은 개체라고 해석한다. 대부분의 데이터가 (0,0) 근처에 모여있는 것과 달리 310번, 352번 데이터는 분포에서 떨어져있는 것을 확인할 수 있다. 이 두 데이터는 이상치라고 판단한다.

1-4. ‘type’ SW(Swimming Birds, 수영하는 조류)

```
df.SW = df[df$type=="SW",-11]
dim(df.SW)

## [1] 106  10

R = cor(df.SW)
eigen(R)$values

##  [1] 7.763148401 1.224846804 0.633385840 0.142148370 0.098878924 0.061808085
##  [7] 0.028400709 0.024751594 0.017424528 0.005206745

PC.result.SW = princomp(df.SW, cor=TRUE)
summary(PC.result.SW)

## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4      Comp.5
## Standard deviation     2.7862427 1.1067280 0.79585541 0.37702569 0.314450194
## Proportion of Variance 0.7763148 0.1224847 0.06333858 0.01421484 0.009887892
## Cumulative Proportion  0.7763148 0.8987995 0.96213810 0.97635294 0.986240834

##                             Comp.6      Comp.7      Comp.8      Comp.9
## Standard deviation     0.248612319 0.168525100 0.157326392 0.132002000
## Proportion of Variance 0.006180809 0.002840071 0.002475159 0.001742453
## Cumulative Proportion  0.992421642 0.995261713 0.997736873 0.999479325
##                            Comp.10
## Standard deviation     0.0721577807
## Proportion of Variance 0.0005206745
## Cumulative Proportion  1.0000000000
```

첫 번째 주성분이 전체 변동 중 77.6%를, 두 번째 주성분이 전체 변동 중 12.2%를 설명하고 있다. type SW 데이터 역시 첫 번째 주성분 만으로도 전체 데이터를 충분히 설명할 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이를 통해 변수들 간 공선성 문제가 존재함을 확인할 수 있다.

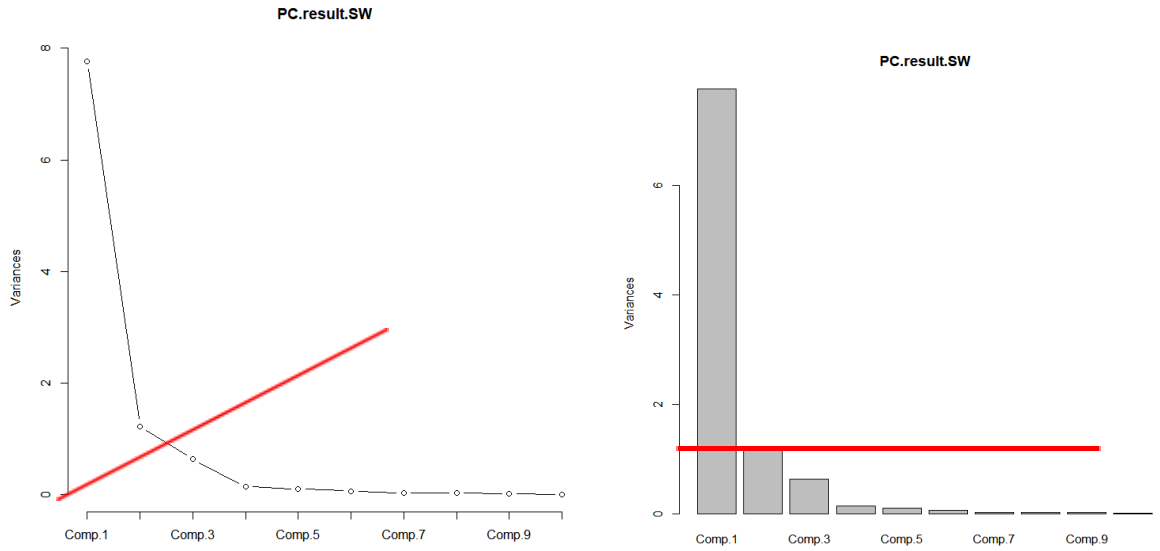
```
PC.result.SW$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.328         0.489                 0.356  0.193                 0.688
## humw   0.339 -0.244        -0.160 -0.342 -0.182  0.161 -0.720  0.320
## ulnal  0.304 -0.119  0.635  0.110  0.183                       -0.138 -0.645
## ulnaw  0.328 -0.297        -0.504 -0.249  0.135 -0.496  0.427  0.198
## feml   0.340        -0.157  0.582 -0.505 -0.321  0.142  0.367
## femw   0.349               -0.138  0.354 -0.515 -0.225 -0.131 -0.567  0.260
## tibl   0.292  0.466 -0.212 -0.446 -0.197  0.193  0.470        -0.352 -0.164
## tibw   0.336  0.222 -0.195         0.533 -0.256  0.204  0.207  0.603
## tarl   0.249  0.635         0.258         0.167 -0.596 -0.267
## tarw   0.282 -0.392 -0.493  0.280  0.279  0.565        -0.141 -0.142
##
```

```
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings        1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var     0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var     0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                 Comp.10
## SS loadings         1.0
## Proportion Var      0.1
## Cumulative Var      1.0
```

주성분의 의미를 간략하게 정리하면 다음과 같다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : tibl(정강발목뼈 길이), tarl(뒷발목뼈 길이)와 tarw(뒷발목뼈 지름)의 대비를 나타내는 축
- PC3 : huml(윗날개뼈 길이), ulnal(자뼈 길이)와 tarw(뒷발목뼈 지름)의 대비를 나타내는 축

```
screeplot(PC.result.SW, type="lines")

screeplot(PC.result.SW)
```

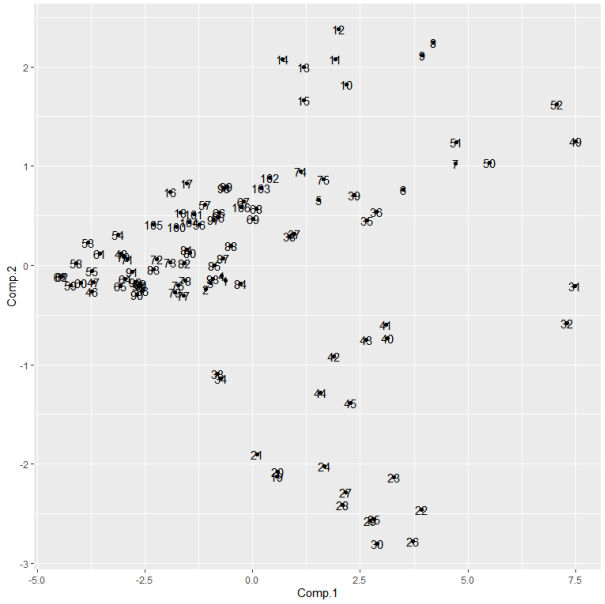


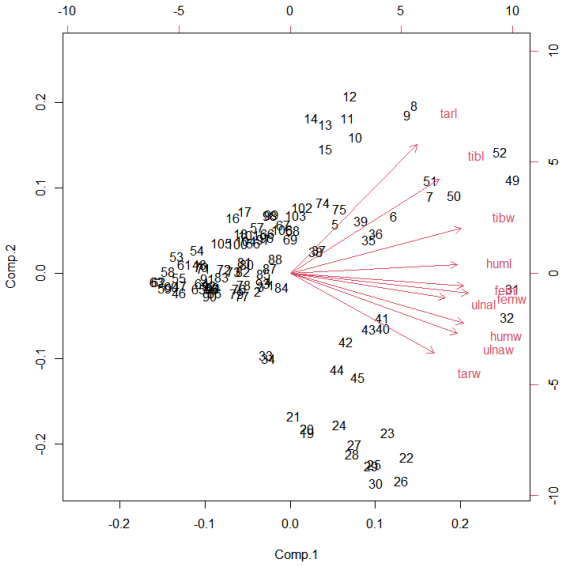
왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1과 Comp.2의 분산이 1 이상임을 확인할 수 있다. 첫번째 주성분만으로 전체 변동 중 77.6%를 설명할 수 있으나 고유값이 1보다 큰 Comp.2까지 두 개의 주성분을 선택한다.

```r
head(PC.result.SW$scores)
```

```
##       Comp.1     Comp.2     Comp.3      Comp.4     Comp.5     Comp.6
## 1 -0.6296036 -0.1537032 -0.3657234 -0.11026303 -0.3240111 -0.3674872
## 2 -1.0909615 -0.2418443 -0.2891393  0.00202338 -0.2854379 -0.1956525
## 3 -0.9808991 -0.1723267 -0.3357815  0.14361605 -0.3093046 -0.1692811
## 4 -0.7231459 -0.1056674 -0.3301376  0.18856740 -0.2250741 -0.1980579
## 5  1.5306121  0.6611487 -0.3959009  0.17420064 -0.4543481 -0.3868259
## 6  3.5044834  0.7623314 -0.6843425  0.47710762 -0.2630246 -0.5843814
##        Comp.7      Comp.8      Comp.9     Comp.10
## 1 -0.05562892  0.07037513  0.14480493  0.10491061
## 2 -0.03543660  0.04044672  0.02931648  0.12258150
## 3  0.06940875 -0.02226586  0.06482809  0.09152236
## 4  0.11696740  0.10074545 -0.05728372  0.12411057
## 5 -0.16249308  0.07067100 -0.08468367 -0.01864994
## 6 -0.45799061 -0.32491904 -0.03608902  0.04288209
```

```r
ggplot(data.frame(PC.result.SW$scores), aes(Comp.1,Comp.2))+
  geom_point()+
  geom_text(aes(label=rownames(df.SW)))

biplot(PC.result.SW)
```





PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 tibl, tarl에 비해 tarw가 긴 개체, 값이 크면 tibl, tarl에 비해 tarw가 작은 개체이다. 49, 52번 개체는 상대적으로 PC1과 PC2 값이 크므로 골격 크기가 크고 tibl, tarl에 비해 tarw가 작은 개체이다.

1-5. 'type' W(Wading Birds, 물가에 서식하는 조류)

```
df.W = df[df$type=="W",-11]
dim(df.W)

## [1] 65 10

R = cor(df.W)
eigen(R)$values

##  [1] 8.789906808 0.593287398 0.268911884 0.198091549 0.069423750 0.033302580
##  [7] 0.023869182 0.013278792 0.005894151 0.004033906

PC.result.W = princomp(df.W, cor=TRUE)
summary(PC.result.W)

## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4      Comp.5
## Standard deviation     2.9647777 0.77025152 0.51856714 0.44507477 0.263483871
## Proportion of Variance 0.8789907 0.05932874 0.02689119 0.01980915 0.006942375
## Cumulative Proportion  0.8789907 0.93831942 0.96521061 0.98501976 0.991962139
##                            Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     0.182489946 0.154496543 0.115233641 0.0767733770
## Proportion of Variance 0.003330258 0.002386918 0.001327879 0.0005894151
## Cumulative Proportion  0.995292397 0.997679315 0.999007194 0.9995966094
##                           Comp.10
## Standard deviation     0.0635130345
## Proportion of Variance 0.0004033906
## Cumulative Proportion  1.0000000000
```

첫 번째 주성분이 전체 변동 중 87.9%를, 두 번째 주성분이 전체 변동 중 5.9%를 설명하고 있다. type W 데이터 역시 첫 번째 주성분 만으로도 전체 데이터를 충분히 설명할 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이를 통해 변수들 간 공선성 문제가 존재함을 확인할 수 있다.

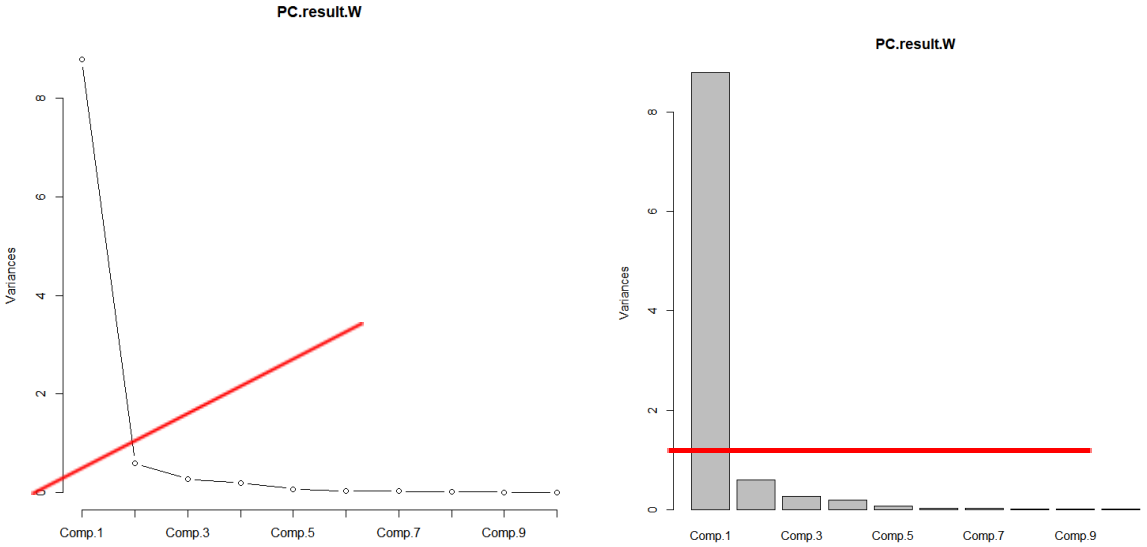
```
PC.result.W$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.322  0.265  0.329  0.245         0.108                 0.103  0.792
## humw   0.318  0.294         0.258 -0.797         0.124        -0.178 -0.222
## ulnal  0.313  0.234  0.584         0.313  0.325 -0.111                -0.526
## ulnaw  0.275  0.617 -0.353 -0.620  0.170
## feml   0.328 -0.198 -0.144  0.165  0.323 -0.235  0.613  0.505
## femw   0.333                0.199  0.171 -0.333  0.163 -0.812 -0.113
## tibl   0.322 -0.326  0.145 -0.258 -0.232                        0.791
## tibw   0.331        -0.192  0.181  0.120 -0.447 -0.727  0.255
## tarl   0.304 -0.422  0.279 -0.540 -0.171        -0.101        -0.541  0.152
## tarw   0.313 -0.281 -0.506  0.143         0.716 -0.133
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                Comp.10
```

```
## SS loadings         1.0
## Proportion Var      0.1
## Cumulative Var      1.0
```

주성분의 의미를 간략하게 정리하면 다음과 같다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : ulnaw(자뼈 지름)와 tibl(정강발목뼈 길이), tarl(뒷발목뼈 길이)의 대비를 나타내는 축

```
screeplot(PC.result.W, type="lines")

screeplot(PC.result.W)
```

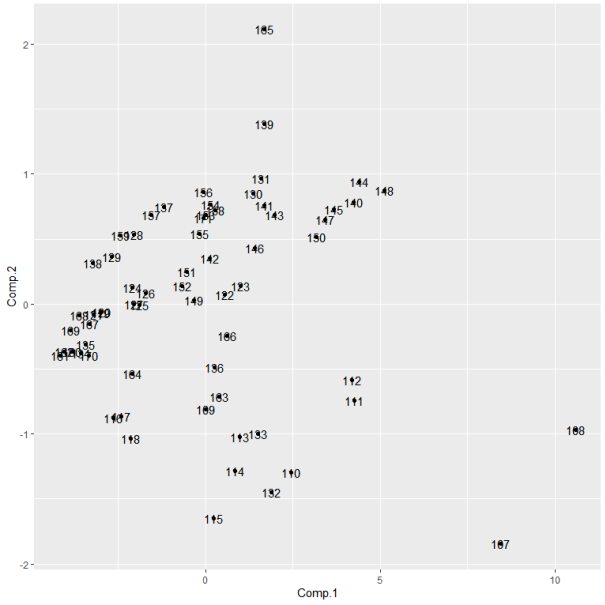


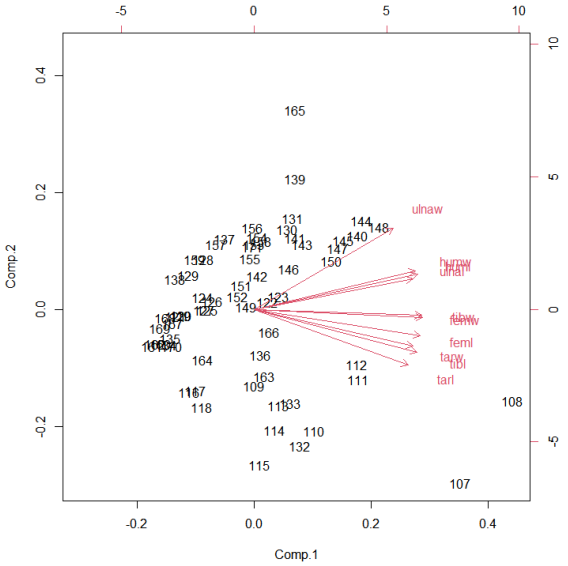
왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1의 분산만이 1 이상임을 확인할 수 있다. 첫번째 주성분만으로 전체 변동 중 87.9%를 설명할 수 있으므로 1개의 주성분을 선택한다.

```r
head(PC.result.W$scores)
```

```
##             Comp.1     Comp.2      Comp.3     Comp.4      Comp.5      Comp.6
## 107  8.4403694380 -1.8463645  0.55105676 -0.9815362 -0.34479138  0.41782521
## 108 10.5794976096 -0.9693131  0.49045867 -0.4405133 -0.59133596  0.02970348
## 109 -0.0006049534 -0.8090813  0.08082072  0.2834335  0.36942134 -0.03857595
## 110  2.4460959874 -1.2940959  0.53263673 -0.3331646 -0.09524134 -0.03066167
## 111  4.2567560834 -0.7441607 -0.10253619  0.5318587  0.38505132 -0.18298299
## 112  4.2007950625 -0.5841883 -0.21236238  0.3488381  0.15855213  0.17688577
##           Comp.7       Comp.8      Comp.9      Comp.10
## 107 -0.16442482  0.077119950 -0.01210827  0.030979030
## 108  0.28883337  0.070614144 -0.07151209 -0.103800844
## 109  0.11600674  0.277292811 -0.03257680  0.008847268
## 110 -0.05239000  0.003690971  0.12915218 -0.022769950
## 111  0.35919845  0.291948123 -0.03294137 -0.103271865
## 112  0.03192227 -0.129732500 -0.08825735  0.042990117
```

```r
ggplot(data.frame(PC.result.W$scores), aes(Comp.1,Comp.2))+
  geom_point()+
  geom_text(aes(label=rownames(df.W)))

biplot(PC.result.W)
```





PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 ulnaw에 비해 tibl, tarl이 긴 개체, 값이 크면 ulnaw에 비해 tibl, tarl이 짧은 개체이다. 107, 108번 개체는 상대적으로 PC1 값이 크고 PC2 값이 작으므로 골격 크기가 크고 정강발목뼈와 뒷발목뼈의 길이가 비교적 길 것이라고 해석할 수 있다. 105번 개체는 상대적으로 PC2 값이 작으므로 자뼈가 비교적 두꺼울 것이라고 해석할 수 있다.

2. 전체 데이터에 대해 R을 이용한 PCA 진행

```
PC.result = princomp(df[,-11], cor=TRUE)

R = cor(df[,-11])
eigen(R)$values

##  [1] 8.564899229 0.677476859 0.397631703 0.123044329 0.089119958 0.063466541
##  [7] 0.035320451 0.024919862 0.016836824 0.007284244

summary(PC.result)

## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4      Comp.5
## Standard deviation     2.9265849 0.82308982 0.63058045 0.35077675 0.298529660
## Proportion of Variance 0.8564899 0.06774769 0.03976317 0.01230443 0.008911996
## Cumulative Proportion  0.8564899 0.92423761 0.96400078 0.97630521 0.985217208
##                            Comp.6      Comp.7      Comp.8      Comp.9
## Standard deviation     0.251925665 0.187937359 0.157860262 0.129756788
## Proportion of Variance 0.006346654 0.003532045 0.002491986 0.001683682
## Cumulative Proportion  0.991563862 0.995095907 0.997587893 0.999271576
##                           Comp.10
## Standard deviation     0.0853477798
## Proportion of Variance 0.0007284244
## Cumulative Proportion  1.0000000000
```

마지막으로 전체 데이터에 대한 주성분 분석을 진행했다. 주성분 요약 정보를 살펴보자. 첫 번째 주성분이 전체 변동 중 85.6%를, 두 번째 주성분이 전체 변동 중 6.8%를 설명하고 있다. type별 분석에서 모든 경우 첫번째 주성분이 자료의 대부분을 설명하고 있었는데, 전체 데이터에서도 역시 첫번째 주성분이 자료의 대부분을 설명하고 있음을 볼 수 있다. 추가로 마지막 고유값이 0에 가까운 값을 가지는데, 이를 통해 변수들 간 공선성 문제가 존재함을 확인할 수 있다. 상관관계가 높은 데이터이기 때문에 공선성 문제가 있다는 것을 알 수 있다.

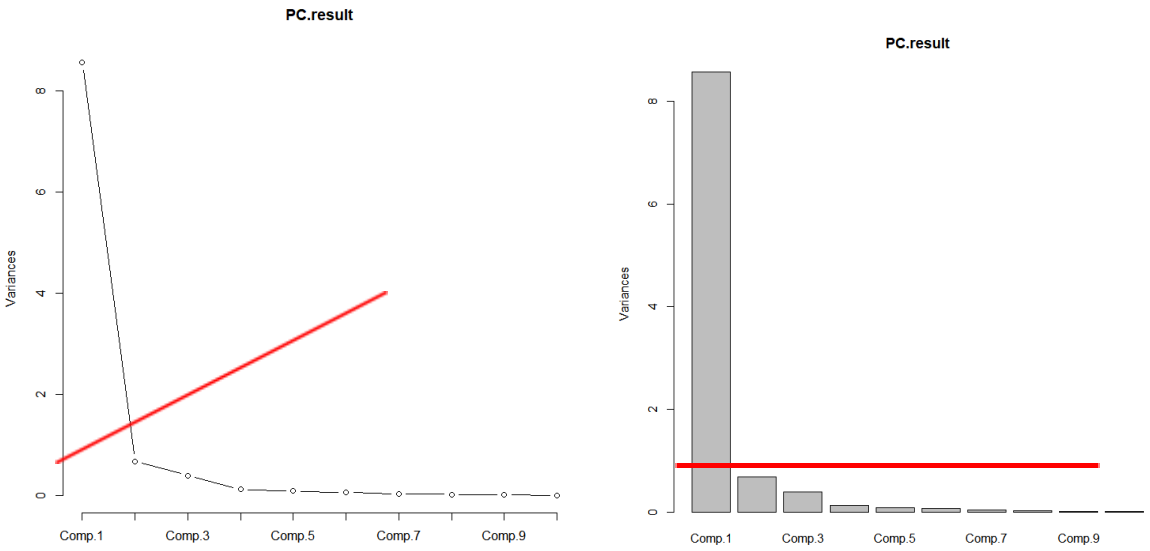
```
PC.result$loadings

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## huml   0.319  0.189  0.480         0.204  0.184  0.195  0.232  0.278  0.622
## humw   0.328  0.277        -0.145 -0.157         0.353 -0.713  0.321 -0.163
## ulnal  0.311  0.256  0.498  0.416         0.194 -0.173        -0.312 -0.500
## ulnaw  0.320  0.287        -0.464 -0.553 -0.276 -0.306  0.350
## feml   0.321 -0.178 -0.321  0.521 -0.284 -0.218  0.491  0.340
## femw   0.333        -0.191  0.243  0.211 -0.406 -0.332 -0.331 -0.425  0.426
## tibl   0.315 -0.398  0.102 -0.480  0.111  0.121  0.405        -0.555
## tibw   0.331 -0.103        -0.151  0.630 -0.309 -0.152  0.223  0.370 -0.384
## tarl   0.275 -0.695  0.138        -0.287  0.214 -0.396 -0.194  0.302
## tarw   0.304  0.235 -0.587                0.690 -0.136
```

```
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
##                Comp.10
## SS loadings        1.0
## Proportion Var     0.1
## Cumulative Var     1.0
```

주성분의 의미를 간략하게 정리하면 다음과 같다. PC2의 경우는 다리뼈의 크기를 나타내는 축이라고도 해석할 수 있다. PC3의 경우는 날개뼈와 다리뼈의 대비를 나타내는 축이라고도 해석할 수 있다.

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : tibl(정강발목뼈 길이)와 tarl(뒷발목뼈 길이)의 가중평균을 나타내는 축
- PC3 : huml(윗날개뼈 길이), ulnal(자뼈 길이)와 feml(넓적다리뼈 길이), tarw(뒷발목뼈 지름)의 대비를 나타내는 축

```r
screeplot(PC.result, type="lines")

screeplot(PC.result)
```



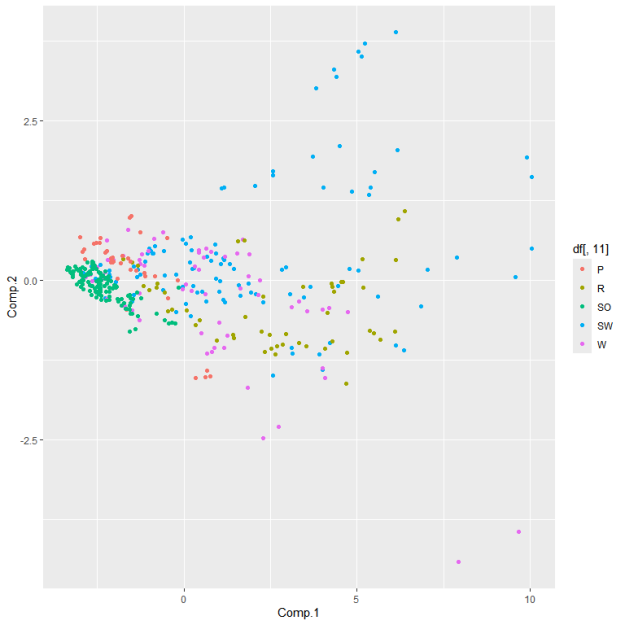
왼쪽의 그림을 보면 Comp.1과 Comp.2 사이에서 분산이 가파르게 감소하는 것을 볼 수 있다. 오른쪽 그림에서는 Comp.1의 분산만이 1 이상임을 확인할 수 있다. 첫번째 주성분만으로 전체 변동 중 85.6%를 설명할 수 있으므로 1개의 주성분을 선택한다.

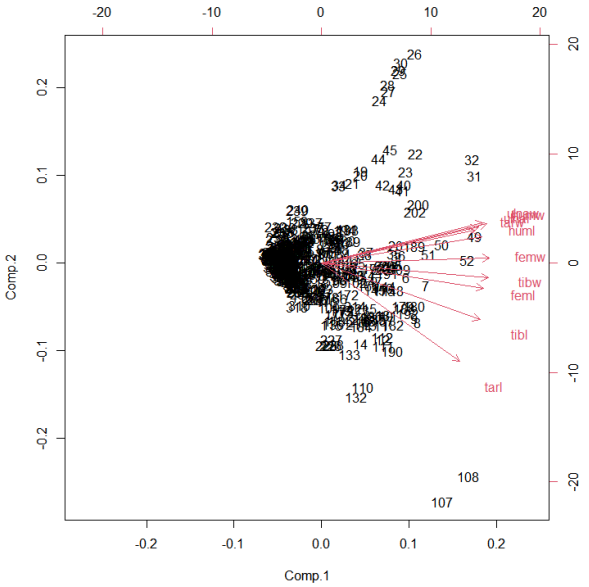
```r
head(PC.result$scores)
```

```
##          Comp.1     Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
## [1,] 1.1578903  0.3228349 -0.17995572 -0.42121656 -0.15643076 -0.43589475
## [2,] 0.6753005  0.3728378 -0.16577167 -0.31208255 -0.17704183 -0.18174088
## [3,] 0.7944200  0.3034628 -0.20602436 -0.25117784 -0.10815481 -0.07449413
## [4,] 1.0716886  0.2579475 -0.19226036 -0.16559710  0.01423168 -0.08881255
## [5,] 3.4617928 -0.2687418  0.05963016 -0.19946831 -0.06657104 -0.37509218
## [6,] 5.5996429 -0.2591069 -0.30958480  0.00986165  0.12875526 -0.38510606
##           Comp.7      Comp.8     Comp.9    Comp.10
## [1,]  0.17665798 -0.06881534 0.15451886 0.03296987
## [2,]  0.17243892 -0.03058137 0.10206144 0.12140062
## [3,]  0.28541479 -0.05987723 0.18002702 0.06108683
## [4,]  0.26064514  0.05719272 0.05938821 0.12556582
## [5,]  0.18739344 -0.07908348 0.04307805 0.02366709
## [6,] -0.07898634 -0.50338317 0.32533570 0.08558373
```

```r
ggplot(data.frame(PC.result$scores), aes(Comp.1,Comp.2,color=df[,11]))+geom_point()
```

```r
biplot(PC.result)
```





PC1 값이 작으면 전체적으로 골격 크기가 작은 개체, 값이 크면 전체적으로 골격 크기가 큰 개체라고 해석할 수 있다. 또, PC2 값이 작으면 정강발목뼈와 뒷발목뼈가 큰 개체, 값이 크면 정강발목뼈와 뒷발목뼈가 작은 개체이다. type과 함께 주성분의 분포를 확인해보자. SO 개체는 상대적으로 작은 PC1 값을 가지는 것으로 보아, SO에 속하는 개체들은 골격이 작을 것이라고 분석할 수 있다. SW 개체는 굉장히 넓은 분포를 보이는데, 대부분이 큰 PC2 값을 가진다. 즉, SW에 속하는 개체들은 정강발목뼈와 뒷발목뼈가 작을 것이라고 분석할 수 있다. 추가로 W 중 2개의 데이터가 분포에서 벗어나 큰 PC1 값, 작은 PC2 값을 가진다. 이 두 개체는 골격 크기가 상대적으로 크고, 정강발목뼈와 뒷발목뼈 역시 큰 개체이며, 이상치로 판단할 수 있다.

III. 결론

모든 PCA에서 첫번째 주성분으로 대부분 설명이 가능했고, 일부는 고유값이 1 이상인 두번째 주성분이 존재했다. 그리고 모든 PCA에서 첫번째 주성분이 설명하고자 하는 것은 전체적인 골격의 크기, 즉 overall mean이다. 첫번째 주성분 값이 크면 개체의 크기가 비교적 크고, 값이 작으면 개체의 크기가 비교적 작다고 해석할 수 있다.

각 type별로 데이터를 분류하여 PCA를 진행했을 때, 두번째 주성분이 설명하고자 하는 것은 서로 달랐다. P와 SO의 두번째 주성분은 날개뼈와 다리뼈의 대비를 의미했고, 이외에도 여러 뼈 사이의 대비나 가중평균을 의미했다. 2개의 주성분을 통해서 10차원의 원자료의 분포를 2차원으로 확인할 수 있었고, 이 과정에서 이상치가 발견되기도 하였다.

마지막으로 전체 데이터에 대한 PCA에서는 SO와 SW type에 속하는 개체들의 분포 특성을 확인할 수 있었다. SO 개체는 다른 type과는 달리 조밀하게 모여 있었고, 작은 골격을 가진다는 특성이 있다. 반대로 SW 개체는 가장 분포가 넓은 것을 확인할 수 있었다. type별 분포가 겹쳐있는 부분이 많아서 5개의 그룹으로 분류하는 것이 쉽지 않을 것이라고 예상된다.

Chapter 4. 선형판별분석(LDA)

작성자: 김경민

I. 서론

판별 분석은 주어진 정보로부터 집단을 분류할 수 있는 판별 규칙을 생성하고, 이를 이용해 새로운 관측치를 분류해내고자 하는 방법이다. 우리의 데이터는 새의 생태학적 그룹에 따라 총 5가지의 type으로 나뉜다. 골격에 대한 정보를 이용해서 새의 type을 분류 해내는 것이 우리의 목표이다. 주성분을 이용한 LDA, 원본 데이터 전체를 사용한 LDA, 훈련용 데이터와 테스트용 데이터를 분리하여 진행한 LDA, 총 3번의 LDA를 진행했다. 3가지의 결과를 성능 지표를 활용해 비교/분석해보고자 한다.

II. 선형판별분석 수행

1. 주성분 2개를 이용한 LDA

전체 데이터에 대한 PCA를 통해 하나의 주성분만으로 데이터의 대부분을 설명할 수 있음을 확인했다. 차원 축소한 데이터가 원 데이터에 대한 정보를 얼마나 보존하였는지 확인하기 위해서 LDA를 진행하고, 원본 데이터를 사용했을 때의 결과와 비교해보고자 한다. 따라서 해당 분석에서는 2개의 주성분을 이용해 LDA를 진행했다.

```
df.pca = data.frame(PC1=PC.result$scores[,1], PC2=PC.result$scores[,2], type=df$type)


ld.pca = lda(type~PC1+PC2, data=df.pca)
ld.pca
```

```
## Call:
## lda(type ~ PC1 + PC2, data = df.pca)
##
## Prior probabilities of groups:
##          P          R         SO         SW          W
## 0.09819121 0.12919897 0.33074935 0.27390181 0.16795866
##
## Group means:
##            PC1         PC2
## P  -1.5689235  0.13606149
## R   2.5829799 -0.42537927
## SO -2.3729060 -0.08171884
## SW  1.9028447  0.43110014
```

```
## W   0.5000082 -0.29443041


##
## Coefficients of linear discriminants:
##           LD1       LD2
## PC1 0.4636831 -0.045303
## PC2 0.2066028  1.285404
##
## Proportion of trace:
##    LD1    LD2
## 0.8549 0.1451
```

PC1, PC2, 개체의 type 정보를 담고 있는 데이터프레임(df.pca)을 생성하고 이를 이용해서 LDA를 진행한 결과이다. LD1의 trace는 85.5%, LD2의 trace는 14.5%인 것으로 보아 분류에서 LD1의 영향이 크다는 것을 알 수 있다. 선형판별식을 구하면 다음과 같다.

- $LD_1 = 0.464 * PC_1 + 0.207 * PC_2$
- $LD_2 = -0.045 * PC_1 + 1.285 * PC_2$

```
pca.pred.c = predict(ld.pca)$class
table(df.pca$type, pca.pred.c)
```

```
##     pca.pred.c
##        P   R  SO  SW   W
##   P    0   0  33   1   4
##   R    0  19  10  14   7
##   SO   0   0 128   0   0
##   SW   0  11  30  58   7
##   W    0  11  27  18   9
```

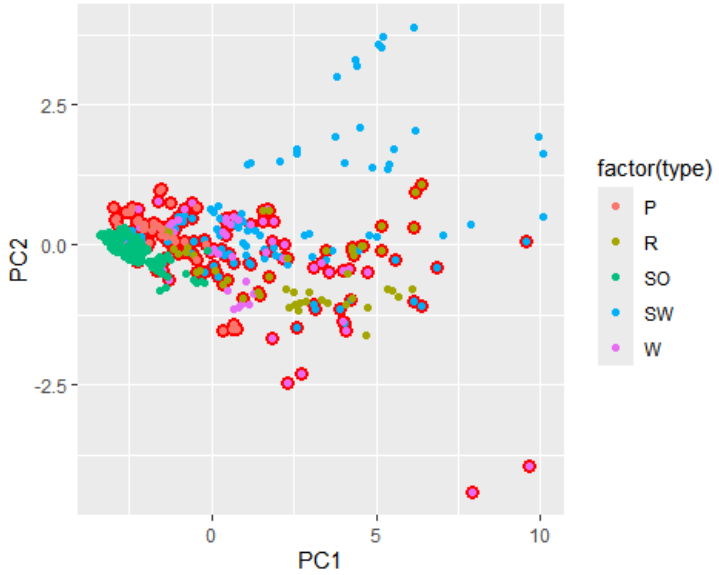
```
(error.rate = mean(df.pca$type!=pca.pred.c))
```

```
## [1] 0.4470284
```

```
(correct.rate = mean(df.pca$type==pca.pred.c))
```

```
## [1] 0.5529716
```

```
df.pca$pred = pca.pred.c
df.pca$miss = (df.pca$type!=pca.pred.c)
plot.data = data.frame(df.pca, predict(ld.pca)$x)

ggplot(df.pca,aes(PC1,PC2))+
  geom_point(data=df.pca[df.pca$miss,], col="red", size=3)+
  geom_point(aes(color=factor(type)))
```



predict 함수를 이용해 class를 분류한 결과를 확인해보자. 혼동행렬을 보면 type P의 경우는 옳게 분류된 경우가 하나도 없고, R, SW, W 역시 잘못 오류율이 높음을 확인할 수 있다. 그러나, SO에 해당하는 128개의 데이터는 모두 옳게 분류되었다. 오류율은 44.7%, 정확도는 55.3%로 낮은 성능을 보이고 있다. 그래프에서도 이상치라고 판단했던 데이터를 비롯한 절반 가까이 되는 데이터가 오분류된 것을 확인할 수 있다. SO 데이터는 모두 옳게 분류되었음을 확인할 수 있다.

2. 원본 데이터를 이용한 LDA

다음은 원본 데이터 전체를 학습과 예측에 모두 사용하여 LDA를 진행한 결과이다.

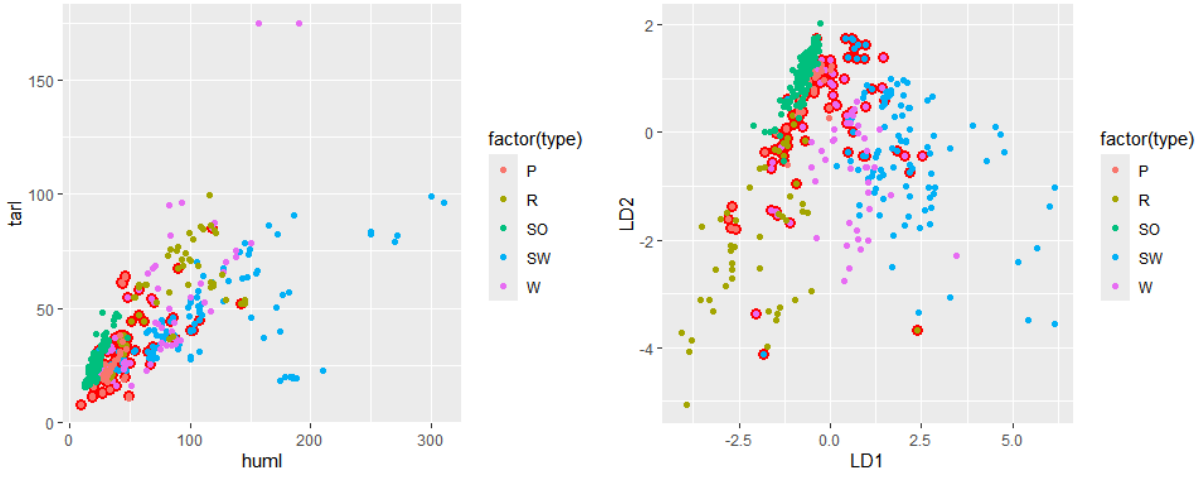
```
ld.df = lda(type~., data=df[1:11])
ld.df
```

```
## Call:
## lda(type ~ ., data = df[1:11])
##
## Prior probabilities of groups:
##          P          R         SO         SW          W
## 0.09819121 0.12919897 0.33074935 0.27390181 0.16795866
##
## Group means:
##         huml     humw     ulnal    ulnaw     feml     femw     tibl     tibw
## P   34.42395 3.039211  39.17737 2.476316 28.21737 2.307895 41.88132 2.095789
## R   86.93440 6.065600 100.38860 4.813000 62.02000 5.266800 89.87740 4.949200
## SO  22.35461 2.028984  26.38625 1.743594 21.37789 1.680313 36.31281 1.547813
## SW 108.70877 6.398774 111.43311 5.207642 41.68642 4.218774 83.43538 4.409717
## W   73.13308 4.607077  78.10138 4.102615 40.09323 3.117077 76.15000 3.179231
##        tarl     tarw
## P  25.78737 1.902632
## R  59.18680 5.071400
## SO 25.84266 1.349375
## SW 43.98858 4.178396
## W  47.54338 2.760308
##
## Coefficients of linear discriminants:
##                LD1          LD2         LD3         LD4
## huml   0.062231118 -0.013078320  0.02425932 -0.01064470
## humw   0.255368987  0.523778536  0.07206236  0.77669723
## ulnal -0.033778567  0.002929654 -0.01885365 -0.01173681
## ulnaw  0.411600599 -0.298185707 -0.81601191 -0.02326180
## feml  -0.116299684 -0.120976710 -0.09837675  0.09655330
## femw  -0.648400670  0.183525341  1.48239665 -0.23887299
## tibl   0.012393570 -0.004884387 -0.04394057  0.01100047
## tibw   0.073231537 -0.064640842  0.47371581  0.42036793
## tarl   0.006617079  0.029734953  0.01195460 -0.08669048
## tarw   0.246044164  0.030830145  0.35983819 -0.97642524
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.5604 0.3145 0.1108 0.0143
```

```
df.pred.c = predict(ld.df)$class


table(df$type, df.pred.c)

##     df.pred.c
##        P   R  SO  SW   W
##   P    2   4  32   0   0
##   R    0  38  11   1   0
##   SO   1   0 127   0   0
##   SW   0   0  12  85   9
##   W    0   5  20   5  35
```

```
(error.rate = mean(df$type!=df.pred.c))

## [1] 0.2583979

(correct.rate = mean(df$type==df.pred.c))

## [1] 0.7416021
```

```
df$pred = df.pred.c
df$miss = (df$type!=df.pred.c)
plot.data = data.frame(df, predict(ld.df)$x)

ggplot(df,aes(huml,tarl))+
  geom_point(data=df[df$miss,], col="red", size=3)+
    geom_point(aes(color=factor(type)))

ggplot(plot.data,aes(LD1,LD2))+
  geom_point(data=plot.data[plot.data$miss,], col='red', size=3)+
  geom_point(aes(color=factor(type)))
```



혼동행렬을 통해서 type P는 대부분 SO로 잘못 분류된 것을 확인할 수 있다. SO는 하나의 데이터를 제외하고는 SO로 옳게 분류되었다. P의 대부분이 SO로 분류되었다는 것은 P의 특성이 SO와 구별하기 어렵다고도 해석할 수 있겠다. 오류율은 25.8%, 정확도는 74.2%로 PCA로 진행한 LDA의 결과보다 좋은 성능을 보인다.

3. 훈련용 데이터와 테스트용 데이터를 분리하여 진행한 LDA

다음은 전체 데이터 중 70%를 훈련용 데이터로, 30%를 테스트용 데이터로 랜덤하게 분리하여 LDA를 진행한 결과이다.

```
p = 0.7
n = nrow(df)

set.seed(2024)
train.ind = sample(n, as.integer(n*p))
df_train = df[train.ind,1:11]
df_test = df[-train.ind,1:11]

ld.result = lda(type~., data=df_train)
ld.result
```

```
## Call:
## lda(type ~ ., data = df_train)
##
## Prior probabilities of groups:
##         P         R        SO        SW         W
## 0.1037037 0.1370370 0.3222222 0.2740741 0.1629630
##
## Group means:
##          huml     humw     ulnal    ulnaw     feml     femw     tibl     tibw
## P    34.18821 3.022143  37.99321 2.488214 28.94393 2.367143 43.17821 2.142143
## R    88.49919 6.095676 101.64378 4.842703 62.60676 5.286216 90.79054 4.972162
## SO   21.96425 1.999195  25.85736 1.717586 21.19747 1.666437 36.05103 1.531839
## SW 114.43892 6.591486 118.03892 5.371216 42.64041 4.332568 86.45095 4.577432
## W    75.24182 4.782500  80.47773 4.273182 41.85795 3.252045 80.91773 3.317273
##         tarl     tarw
## P  27.17250 1.895357
## R  58.83459 5.154054
## SO 25.77149 1.325747
## SW 45.94878 4.188378
## W  50.18318 2.932955
##
## Coefficients of linear discriminants:
##                LD1          LD2         LD3           LD4
## huml   0.04799828 -0.004143370  0.03169806  0.0007291957
## humw   0.27484055  0.481869596 -0.07541123  0.8511186020
## ulnal -0.02284593 -0.002175299 -0.02000491 -0.0272412365
## ulnaw  0.38448716 -0.295365525 -0.66324864 -0.0016816921
## feml  -0.10787864 -0.109020490 -0.09453591  0.0696540997
## femw  -0.73590892  0.192979935  1.42644397  0.2661209366
## tibl   0.01075111 -0.012551737 -0.05793771  0.0099370398
## tibw   0.16880477 -0.071346739  0.48644673  0.0350218002
## tarl   0.00838051  0.035904985  0.02677796 -0.0648754137
## tarw   0.20034282 -0.007513546  0.37691083 -0.9839746701
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.5416 0.3209 0.1236 0.0139
```

먼저 그룹별 데이터의 비율을 살펴보자. train data의 'Prior probabilities of groups'은 전체 데이터에서의 비율과 거의 비슷하다는 것을 알 수 있다. 따라서 type을 stratify하여 데이터를 분리한 후 LDA를 진행하는 과정은 생략했다.

```
### train error
train.pred.c = predict(ld.result)$class
table(df_train$type, train.pred.c)
```

```
##     train.pred.c
##       P  R SO SW  W
##   P   1  4 22  0  1
##   R   0 27  9  1  0
##   SO  0  0 87  0  0
##   SW  0  0 11 59  4
##   W   0  3 13  5 23
```

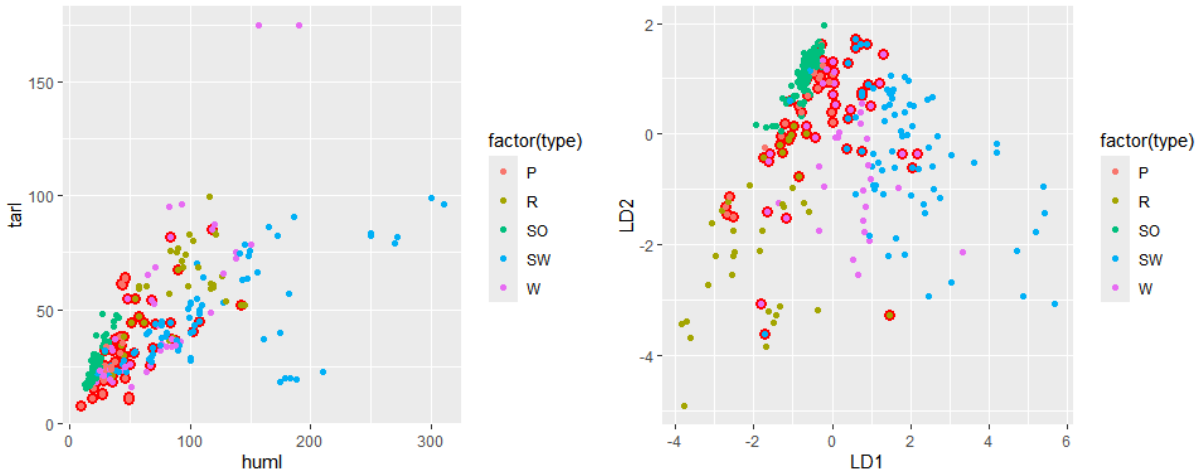
```
(error.rate = mean(df_train$type!=train.pred.c))
```

```
## [1] 0.2703704
```

```
(correct.rate = mean(df_train$type==train.pred.c))
```

```
## [1] 0.7296296
```

```
df_train$pred = train.pred.c
df_train$miss = (df_train$type!=train.pred.c)
plot.data = data.frame(df_train, predict(ld.result)$x)

ggplot(df_train,aes(huml,tarl))+
  geom_point(data=df_train[df_train$miss,], col="red", size=3)+
  geom_point(aes(color=factor(type)))

ggplot(plot.data, aes(LD1, LD2))+
  geom_point(data=plot.data[plot.data$miss,],col='red', size=3)+
  geom_point(aes(color=factor(type)))
```



train 데이터에서 LDA 모델을 학습한 후, 다시 train 데이터로 성능을 평가한 결과이다. 이번에도 역시 SO는 모두 옳게 분류 되었고, P는 단 하나의 데이터만 옳게 분류되었다. 오류율은 27.0%, 정확도는 73.0%로, 2)의 경우보다 조금 낮은 성능을 보인다.

```
### test error
test.pred.c = predict(ld.result, newdata=df_test[,-11])$class
table(df_test$type, test.pred.c)
```

```
##     test.pred.c
##       P  R SO SW  W
##   P   1  0  9  0  0
##   R   1  9  3  0  0
##   SO  0  0 40  0  1
##   SW  0  0  6 21  5
##   W   0  1  8  0 12
```

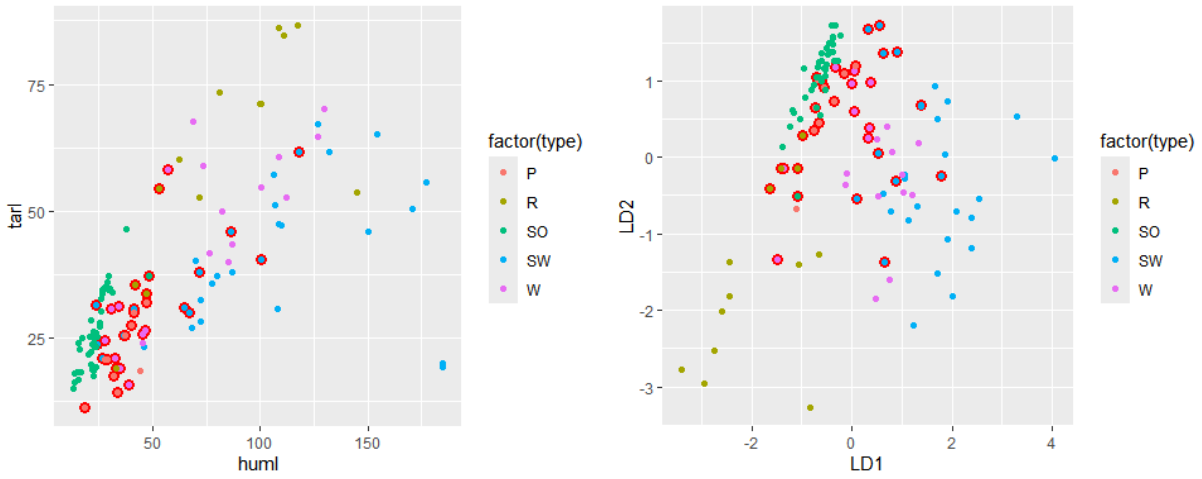
```
(error.rate = mean(df_test$type!=test.pred.c))
```

```
## [1] 0.2905983
```

```
(correct.rate = mean(df_test$type==test.pred.c))
```

```
## [1] 0.7094017
```

```
df_test$pred = test.pred.c
df_test$miss = (df_test$type!=test.pred.c)

ggplot(df_test,aes(huml,tarl))+
  geom_point(data=df_test[df_test$miss,], col="red", size=3)+
  geom_point(aes(color=factor(type)))

plot.data = data.frame(df_test, predict(ld.result, newdata=df_test[,-11])$x)
ggplot(plot.data, aes(LD1, LD2))+
  geom_point(data=plot.data[plot.data$miss,],col='red', size=3)+
  geom_point(aes(color=factor(type)))
```



이번에는 test 데이터를 새로운 관측치로 주었을 때 옳게 분류하는 지를 확인하였다. test 데이터 성능을 보면 오류율은 29.1%, 정확도는 70.9%로, train에서의 결과보다 오분류율이 높은 것을 알 수 있다.

III. 결론

LDA의 결과를 정리하면 아래의 표와 같다.

LDA	주성분 **2**개 이용	원본 데이터 이용	훈련용 데이터로 예측	테스트용 데이터로 예측
오분류율	44.70%	27.04%	25.84%	29.06%
정확도	55.30%	72.96%	74.16%	70.94%

주성분을 이용했을 때보다 원본 데이터에서 더 정확하게 분류했음을 알 수 있다. 주성분 분석은 정보 손실을 최소화하면서 차원을 축소해서 효율적으로 자료를 분석하기 위한 것이다. 2개의 주성분이 전체 변동의 90% 이상을 설명한다고 하더라도 원본 데이터에 비해 정보 손실이 발생했기 때문에 분류 성능이 낮은 것이라고 해석해 볼 수 있다. 즉, 해당 데이터에서는 주성분을 이용한 차원 축소가 모델의 일반화 능력을 향상시키지 못한다. 원본 데이터를 이용한 모델은 훈련용 데이터로 예측했을 때 가장 높은 정확도를 보였으나, 테스트용 데이터에서는 약간 낮은 정확도를 보인다. 이는 모델이 훈련용 데이터에 더 적합하다는 것을 의미하고 약간의 과적합이 발생했을 가능성이 있다고 볼 수 있다.

IV. 이차판별분석 추가 수행

그룹별 분산의 차이가 존재하는 10차원의 데이터를 선형판별식으로만 분류하는 것은 분명히 한계가 존재한다. 다음은 원본 데이터 전체와 훈련용/테스트용으로 분리한 데이터에서 QDA를 진행하고 성능을 확인한 결과이다.

```
### org.data

qda1 = qda(type~., data=df[1:11])

qda1.pred.c = predict(qda1, df)$class
table(df$type, qda1.pred.c)

##     qda1.pred.c
##        P   R  SO  SW   W
##   P   37   0   1   0   0
##   R    2  47   0   0   1
##   SO   2   2 124   0   0
##   SW   1   0   2  81  22
##   W    7   1   0   2  55

(error.rate = mean(df$type!=qda1.pred.c))

## [1] 0.1111111

(correct.rate = mean(df$type==qda1.pred.c))

## [1] 0.8888889
```

```
### train/test split

p = 0.7
n = nrow(df)

set.seed(2024)
train.ind = sample(n, as.integer(n*p))
df_train = df[train.ind,1:11]
df_test = df[-train.ind,1:11]
```

```
# train data
qda.train = qda(type~., data=df_train)

qda.train.pred.c = predict(qda.train, df_train)$class
table(df_train$type, qda.train.pred.c)

##     qda.train.pred.c
##       P  R SO SW  W
##   P  27  0  1  0  0
##   R   1 36  0  0  0
##   SO  1  0 86  0  0
##   SW  1  0  1 59 13

##   W   5  0  0  0 39
```

```
(error.rate = mean(df_train$type!=qda.train.pred.c))
```

```
## [1] 0.08518519
```

```
(correct.rate = mean(df_train$type==qda.train.pred.c))
```

```
## [1] 0.9148148
```

```
# test data
qda.test.pred.c = predict(qda.train, newdata=df_test[,-11])$class
table(df_test$type, qda.test.pred.c)
```

```
##     qda.test.pred.c
##       P  R SO SW  W
##   P   8  1  0  0  1
##   R   0 13  0  0  0
##   SO  1  1 39  0  0
##   SW  0  0  1 23  8
##   W   3  0  0  0 18
```

```
(error.rate = mean(df_test$type!=qda.test.pred.c))
```

```
## [1] 0.1367521
```

```
(correct.rate = mean(df_test$type==qda.test.pred.c))
```

```
## [1] 0.8632479
```

QDA	원본 데이터 이용	훈련용 데이터로 예측	테스트용 데이터로 예측
오분류율	11.11%	8.52%	13.68%
정확도	88.89%	91.48%	86.32%

원본 데이터 전체를 이용했을 때 오류율 11.11%, 정확도 88.89%로 LDA보다 높은 성능을 보인다. 혼동행렬을 살펴보면 LDA로는 분류가 잘 되지 않았던 type P가 하나의 데이터를 제외하고는 모두 옳게 분류되었음을 볼 수 있다. 훈련용 데이터로 예측한 결과의 정확도는 91.48%로 가장 높으나, 테스트용 데이터로 예측한 결과와 비교했을 때 과적합이 발생했음을 알 수 있다.

EDA에서 확인했듯이, 우리의 데이터는 집단마다 분산의 차이가 존재하는 데이터이다. 집단의 분산이 다를 때는 선형판별규칙(LDA)보다는 이차판별규칙(QDA)을 이용한 분류 방법의 성능이 더 좋다는 것을 확인했다. 추가로 LDA 모델보다 QDA 모델이 더 과적합된 것으로 보아 이차판별분석에서 과적합 문제에 더 주의해야 함을 확인하며 분석을 마무리했다.

5) 시사점

이번 보고서에서는 새의 뼈 길이와 지름 데이터를 활용하여 생태학적 그룹을 분류하는 다양한 분석 기법을 적용하였다. 각 분석 단계에서 얻은 시사점을 종합하여 다음과 같이 정리할 수 있다.

I. 데이터 전처리

먼저 데이터셋에 존재하는 결측치는 각 그룹별 중앙값으로 보완하여 데이터의 신뢰성을 높였고, 이를 통해 분석의 일관성을 유지할 수 있었다.

이상치는 주로 IQR 방식을 통해 탐지하였으며, 비율이 낮은 경우에는 제거하였다. 이는 데이터의 대표성을 유지하면서도 왜곡된 결과를 방지하는 데 기여했다.

II. 탐색적 데이터 분석 (EDA)

데이터의 분포와 주요 특징을 파악하기 위해 다양한 시각화 기법을 활용하였고, 이를 통해 각 변수의 분포, 상관 관계 및 데이터의 전반적인 특성을 이해할 수 있었다. 또한, 변수 간 상관성을 통해 뼈의 길이와 지름이 생태학적 그룹을 분류하는 데 중요한 역할을 한다는 점을 확인하였다.

III. 정준상관분석 (CCA)

CCA를 통해 새의 뼈의 길이와 지름 간의 관계, 그리고 날개 뼈와 다리 뼈 간의 관계를 조사하였다. 이를 통해 변수들 간의 숨겨진 관계를 발견하고, 중요 변수를 식별하는 데 도움이 되었다. 다만, 본 분석에서 사용한 데이터는 변수들 간의 상관성이 높아 추가적으로 유의미한 결과를 도출하는 데 어려움이 있었다. 따라서, 분석 진행 전 예측한 것과 같이 이번 분석은 EDA에서 확인된 결과를 재확인하는 수준에 그쳤다.

IV. 군집분석(Clustering)

다양한 군집화 기법을 적용하여 같은 클러스터링 기법을 활용하여 데이터의 자연스러운 그룹화를 시도하였다. 이를 통해 생태학적 그룹(type) 외에도 데이터 내에서 추가적인 패턴을 발견할 수 있었다. 이후 각 군집의 특성을 분석함으로써, 생태학적 그룹과 일치하는지 여부를 확인할 수 있었다.

V. 주성분 분석 (PCA)

PCA를 통해 데이터의 차원을 축소함으로써, 주요 성분을 추출하였다. 이는 데이터 시각화 및 모델의 효율성을 높이는 데 기여했다. 더불어, 전체 분산의 대부분을 설명하는 소수의 주성분을 사용하여도 원 데이터의 주요 정보를 대부분 보존할 수 있음을 확인했다.

VI. 선형 판별 분석 (LDA) 및 이차 판별 분석 (QDA)

LDA와 QDA를 비교 분석하였다. LDA는 상대적으로 단순하면서도 안정적인 성능을 보였다. QDA가 더 높은 분류 성능을 보였지만, 과적합 문제에 더 민감하다는 점을 발견할 수 있었다. 이는 데이터의 특성과 모델의 복잡도를 고려하여, 적합한 모델을 선택하는 것이 중요하다는 점을 시사한다.

종합적 시사점

데이터 전처리, EDA, CCA, 클러스터링, PCA, LDA, QDA 등 다양한 분석 기법을 통합적으로 활용함으로써, 데이터의 다양한 측면을 이해하고 분석의 신뢰성을 높일 수 있었다. 또한, 모델의 적합성을 높이기 위해 각 기법의 필요성과 장단점을 이해하고, 데이터의 특성에 맞는 적절한 분석 방법을 선택하는 것이 중요하다는 점을 알 수 있었다.

실무에서의 데이터 분석은 여러 단계와 기법을 종합적으로 적용해야 하며, 이는 데이터의 특성, 분석 목표 및 현실적인 제약을 고려한 의사결정을 필요로 한다. 이러한 종합적인 접근 방식을 통해, 더욱 정확하고 신뢰성 있는 분석 결과를 도출할 수 있으며, 이는 다양한 분야에서의 실질적인 응용 가능성을 높이는 데 기여할 것이라 생각된다.

6) 역할

주제 및 데이터 소개, 데이터 전처리, 탐색적 데이터 분석(EDA)

- 분석 : 김경민, 이도경, 장단, 차수빈
- 보고서 작성, PPT 제작 및 발표 : 장단

정준상관분석

1) 뼈 길이와 지름 간의 관계

- 분석 : 김경민
- 보고서 작성, PPT 제작 및 발표 : 이도경

2) 날개뼈와 다리뼈 간의 관계

- 분석, 보고서 작성, PPT 제작 및 발표 : 이도경

군집분석

- 분석 : 장단, 차수빈
- 보고서 작성, PPT 제작 및 발표 : 차수빈

주성분분석

- 분석 : 김경민, 장단
- 보고서 작성, PPT 제작 및 발표 : 김경민

LDA/QDA

- 분석, 보고서 작성, PPT 제작 및 발표 : 김경민

보고서 완성 : 이도경, 차수빈

발표자료 완성 : 김경민, 장단