이화여자대학교

2024-1 다변량분석및실습 팀프로젝트

I 새의 뼈와 생태학적 분류

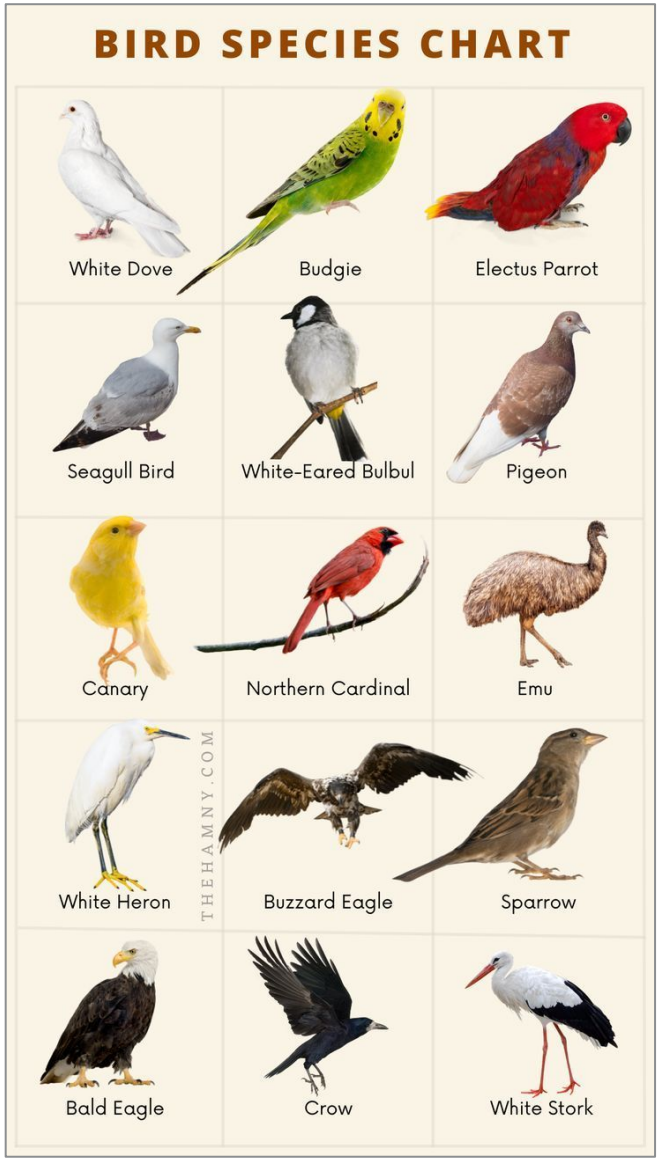
3조 (다알조)

2129006 김경민
2129036 차수빈
2135019 장단
2176247 이도경

EWHA,
THE FUTURE
WE CREATE

목차

1. 주제 및 데이터 소개



📌[Birds' Bones and Living Habits]

- 새의 뼈에 대한 자료로 생태학적 그룹을 분류
- 11개의 독립변수와 1개의 종속변수(type)로 구성
- 420개의 데이터

뼈 길이 (length)

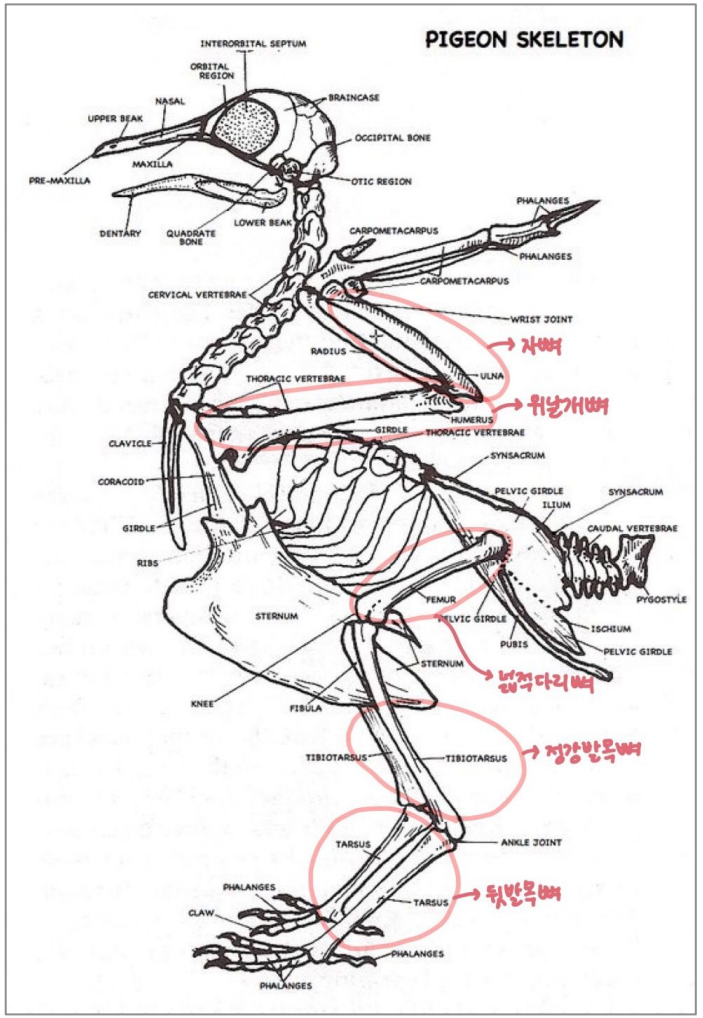
뼈 지름 (width)

EWHA, THE FUTURE WE CREATE

1. 주제 및 데이터 소개

📌id를 제외한 10개의 독립변수(측정값,mm)

- huml: Length of Humerus(위날개뼈)
- humw: Diameter of Humerus
- ulnal: Length of Ulna(자뼈)
- ulnaw: Diameter of Ulna
- feml: Length of Femur(넓적다리뼈)
- femw: Diameter of Femur
- tibl: Length of Tibiotarsus(정강발목뼈)
- tibw: Diameter of Tibiotarsus
- tarl: Length of Tarsometatarsus(뒷발목뼈)
- tarw: Diameter of Tarsometatarsus



EWHA,
THE FUTURE
WE CREATE

1. 주제 및 데이터 소개

📌6개의 종속 변수(Type of birds)

- 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

2. 데이터 전처리와 EDA

📌ID 변수 제거

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

📌결측치 처리

- 총 15개의 결측치 존재, 각 'type'의 기술통계량 활용하여 중앙값으로 보간

```
# type 별로 분리
bird_grouped <- split(bird, bird$type)

# type 별로 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
}
```

EWHA,
THE FUTURE
WE CREATE

2. 데이터 전처리와 EDA

📌이상치 탐지 및 처리

```
outliers <- list()
for (col in names(group_data)) {
  if (col != "type") {
    outliers[[col]] <- get_outlier_indices(group_data[[col]], weight = 3)
  }
}
```

- 대부분의 경우 이상치 비율이 작음 ⇒ 이상치 제거

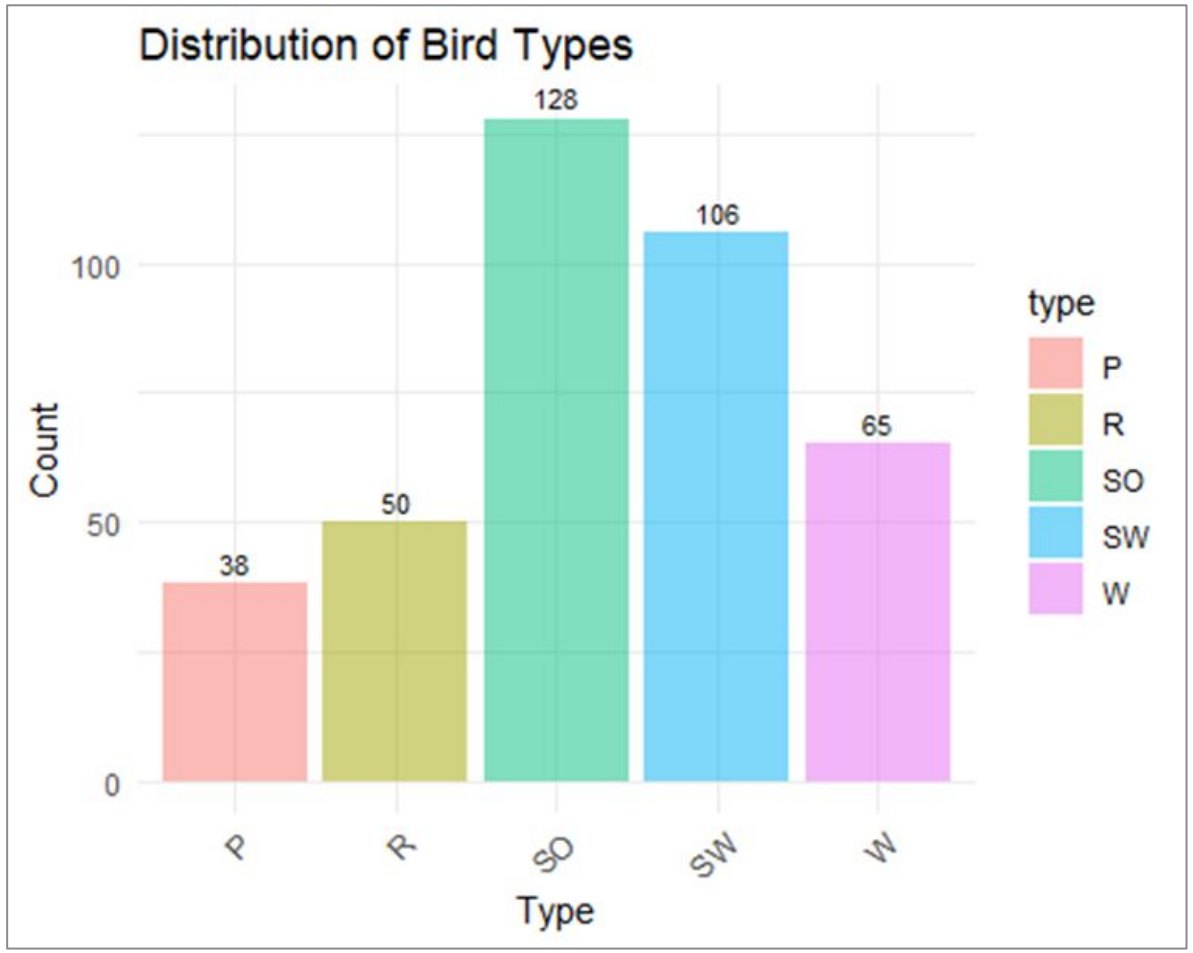
```
## Type: P - 이상치 비율: 0.05263158
## Type: R - 이상치 비율: 0
## Type: SO - 이상치 비율: 0.0078125
## Type: SW - 이상치 비율: 0.00862069
## Type: T - 이상치 비율: 0.173913
## Type: W - 이상치 비율: 0.03076923
```

Type 'T'의 데이터 건 수 ↓, 이상치 비율 ↑

⇒ 분석에 방해가 될 수 있어 제외 결정

2. 데이터 전처리와 EDA

📌Type(종속변수) 별 분포 확인: 막대그래프



- SO(노래하는 조류)와 SW(수영하는 조류)가 전체 데이터의 약 60% 정도를 차지
- P(산악지대 조류)가 가장 적게 관측

EWHA,
THE FUTURE
WE CREATE

2. 데이터 전처리와 EDA

📌독립변수들의 분포 확인: Boxplot

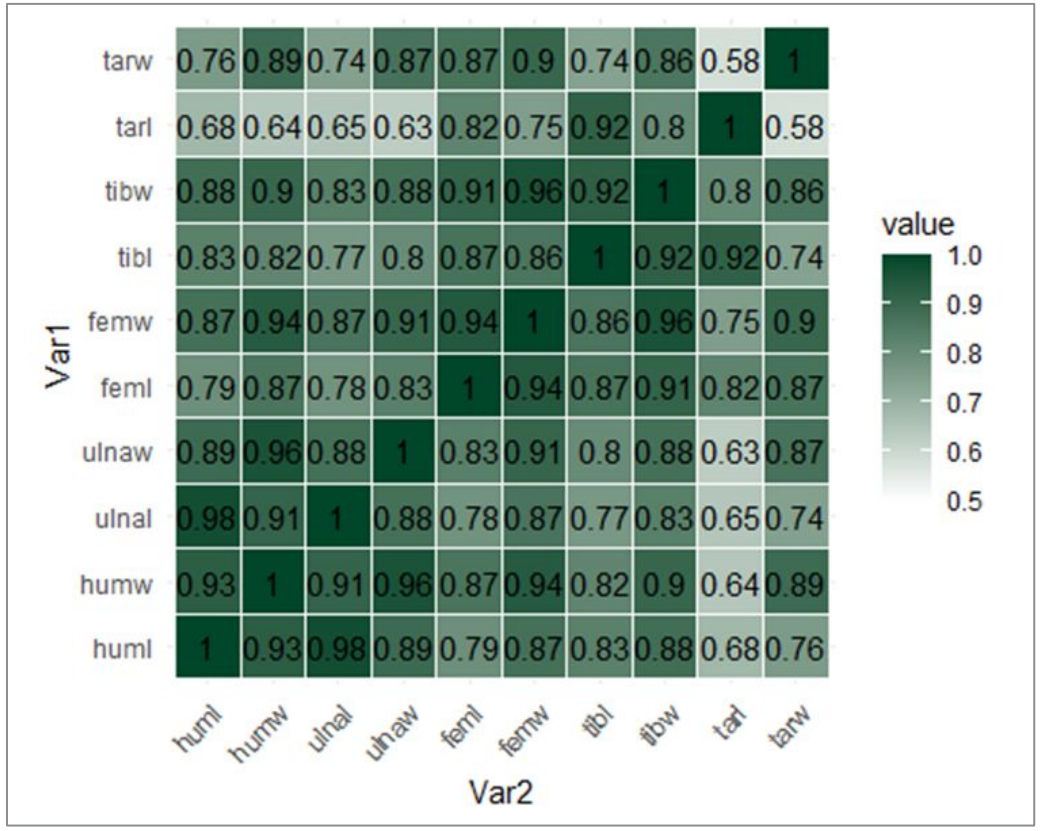
- type 별로 길이/지름의 분포가 다르게 나타나며 그 범위에서도 차이가 드러난다.
- P, SO는 좁게 분포하고 값이 작은 경향을 보임
- R, SW는 분포가 넓고 값이 큰 경향을 보임
- 대부분의 변수에서 비슷한 양상이 드러나는 것을 알 수 있음

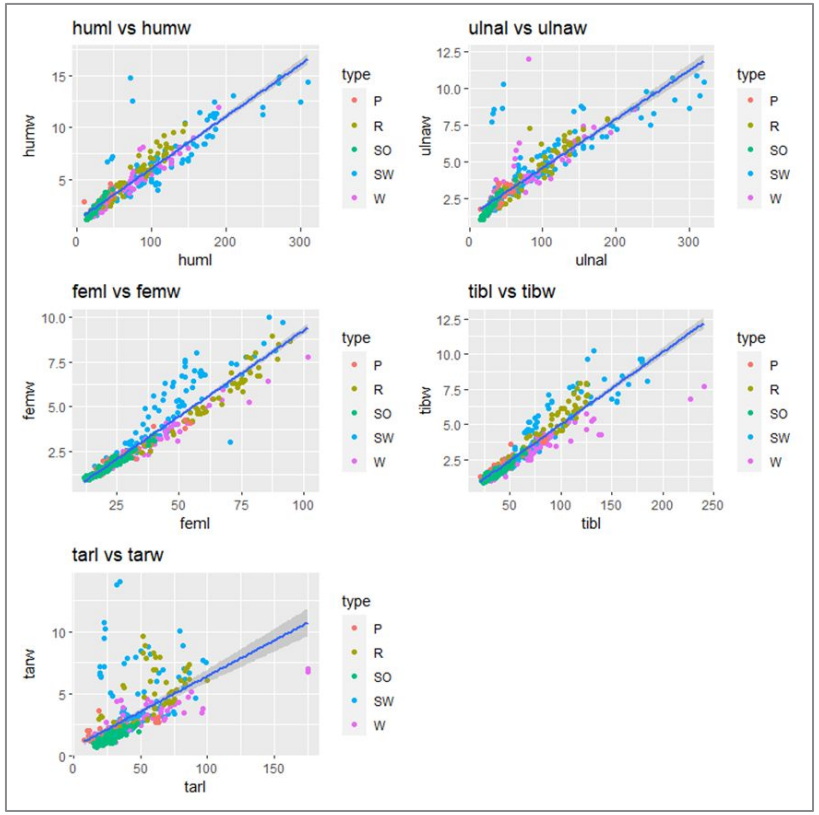
EWHA,
THE FUTURE
WE CREATE

2. 데이터 전처리와 EDA

📌독립변수들 간 correlation 확인

- 강한 양의 상관관계를 가지며, 특히 같은 뼈끼리 길이-지름 간의 상관성이 높게 나타난다.





EWHA,
THE FUTURE
WE CREATE

3. 정준상관분석(CCA)

📌분석 목표

- 각 뼈의 길이 vs 지름
- 날개 뼈(길이, 지름) vs 다리 뼈(길이, 지름)

3. 정준상관분석(CCA)

📌새의 뼈의 길이와 지름 간의 관계 분석 - 정준상관계수

- Zx : 표준화된 뼈의 길이, 5차원
- Zy : 표준화된 뼈의 지름, 5차원

```
cc1 = cc(Zx,Zy)

# 정준상관계수
cc1$cor

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

- 첫 번째 정준상관계수(0.9822)가 거의 1에 가까움

  ⇒ 첫번째 정준변수 쌍(U1, V1)이 아주 강한 양의 상관관계를 가짐

- 두 번째 정준상관계수(0.7926) 또한 0.8에 가까운 상관계수로

  ⇒ 두 번째 정준변수 쌍(U2, V2)의 강한 양의 상관관계를 보여줌

3. 정준상관분석(CCA)

📌새의 뼈의 길이와 지름 간의 관계 분석 - 정준계수

- 원래의 변수들이 정준변수에 상대적으로 기여하는 정도를 알 수 있다.

```
# 정준변수 U 의 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 V 의 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
```

EWHA,
THE FUTURE
WE CREATE

3. 정준상관분석(CCA)

📌새의 뼈의 길이와 지름 간의 관계 분석 - 정준적재

	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

X 와 U 의 상관관계이다. X 의 변수 모두 첫 번째 정준변수 U1 과의 상관관계가 매우 높다.

	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

Y 와 U 의 상관관계이다. Y 의 변수 모두 첫 번째 정준변수 U1 과의 상관관계가 매우 높다.

	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

X 와 V 의 상관관계이다. X 의 변수 모두 첫 번째 정준변수 V1 과의 상관관계가 매우 높다.

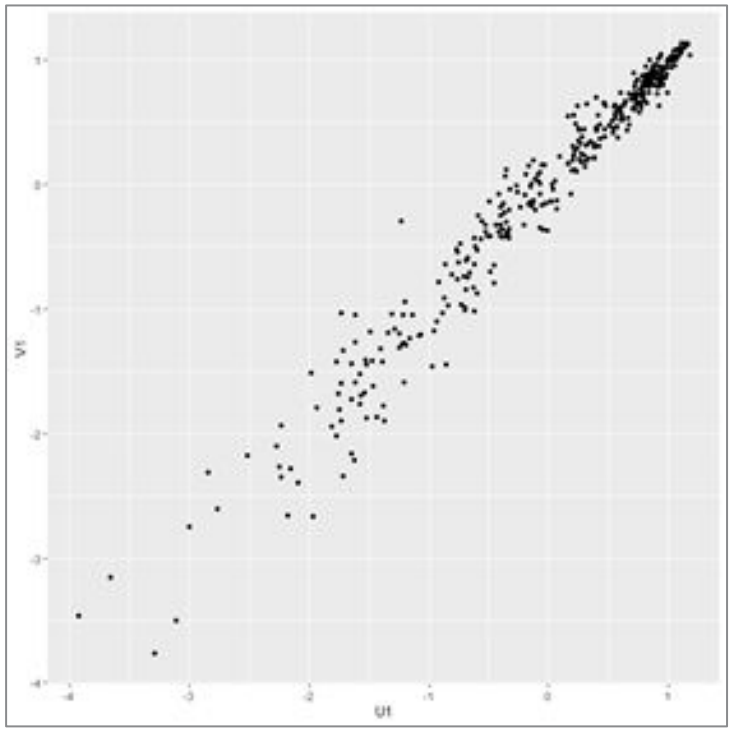
	V1	V2	V3	V4	V5
humw	-0.9779697	-0.06999025	0.18906779	0.053634736	0.007302480
ulnaw	-0.9458520	-0.05201181	0.16888231	0.069710119	-0.263207312
femw	-0.9738103	-0.18075418	-0.12220623	-0.063584000	-0.006643858
tibw	-0.9728088	0.07169426	-0.21994835	-0.001356518	0.011130672
tarw	-0.8882854	-0.28509616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

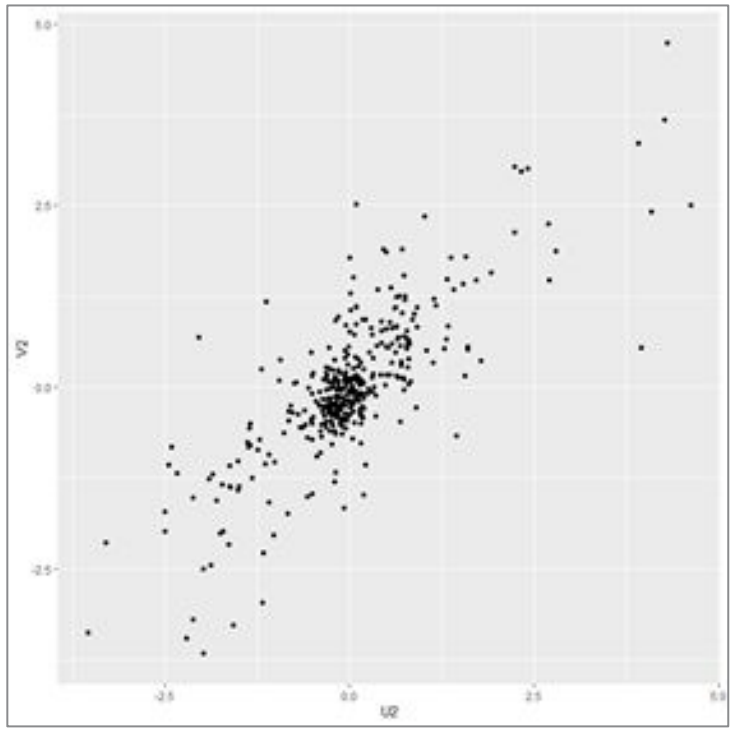
Y 와 V 의 상관관계이다. Y 의 변수 모두 첫 번째 정준변수 V1 과의 상관관계가 매우 높다.

- 원래의 변수들과 정준변수의 상관관계
- X와 Y 모두 첫 번째 정준변수와의 상관성이 매우 크다.
- 첫 번째 정준변수 쌍(U1,V1)만 사용하여
  두 집단 간의 관계의 대부분을 설명하는 것이 가능하다.

3. 정준상관분석(CCA)

📌새의 뼈의 길이와 지름 간의 관계 분석 - 결과 시각화





- 첫 번째 정준변수 쌍의 상관관계가 매우 높음
- 두 번째 정준변수 쌍 부터는 상관성이 낮아짐

3. 정준상관분석(CCA)

📌새의 날개뼈와 다리뼈의 관계 분석 - 정준상관계수

- Zx : 표준화된 날개뼈의 길이 및 지름, 4차원
- Zy : 표준화된 다리뼈의 길이 및 지름, 6차원

```
# 정준상관계수
bird_cc$cor

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

- 첫 번째 정준상관계수(0.96)가 거의 1에 가까움

  ⇒ 첫번째 정준변수 쌍(U1, V1)이 아주 강한 양의 상관관계를 가짐

- 두 번째 정준상관계수(0.77) 또한 0.7 이상의 상관계수를 가짐

  ⇒ 두 번째 정준변수 쌍(U2, V2)의 강한 양의 상관관계를 보여줌

3. 정준상관분석(CCA)

📌새의 날개뼈와 다리뼈의 관계 분석 - 정준계수

- 원래의 변수들이 정준계수에 상대적으로 기여하는 정도를 알 수 있다.

```
# 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
```

EWHA,
THE FUTURE
WE CREATE

3. 정준상관분석(CCA)

📌새의 날개뼈와 다리뼈의 관계 분석 - 정준적재

```
##              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
```

X 와 U 의 상관관계이다. X 의 변수 모두 첫 번째 정준변수 U1 과의 상관관계가 매우 높다.

```
##              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
```

X 와 V 의 상관관계이다. X 의 변수 모두 첫 번째 정준변수 V1 과의 상관관계가 매우 높다.

```
##              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
```

Y 와 U 의 상관관계이다. Y 의 변수 모두 첫 번째 정준변수 U1 과의 상관관계가 매우 높다.

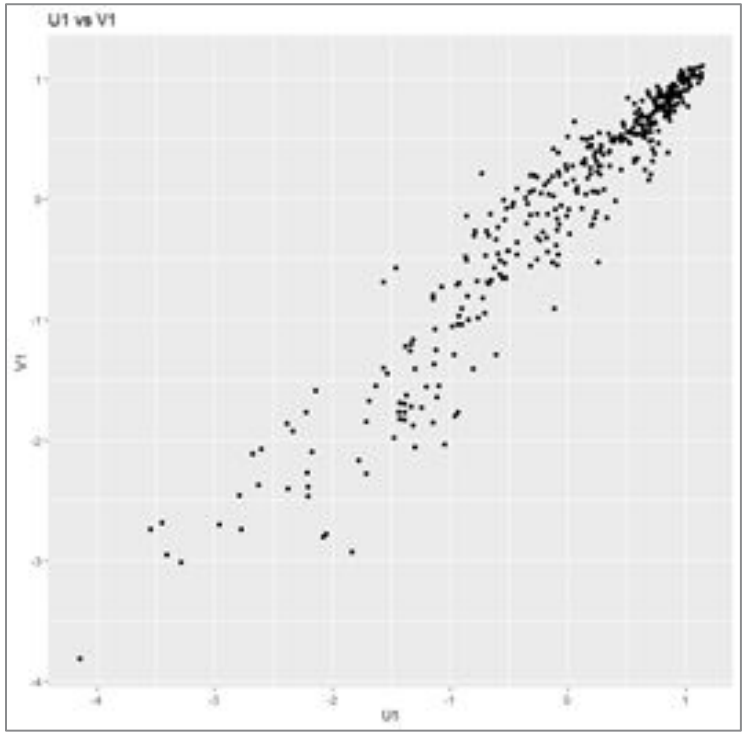
```
##              V1          V2          V3           V4
## feml  -0.8988298  0.07182576  0.08089728 -0.168653530
## femw  -0.9732380  0.05893908  0.16910629 -0.068495065
## tibl  -0.8571371 -0.33570121  0.15987001  0.112516975
## tibw  -0.9443041 -0.20537157  0.13901448 -0.109682865
## tarl  -0.6672491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

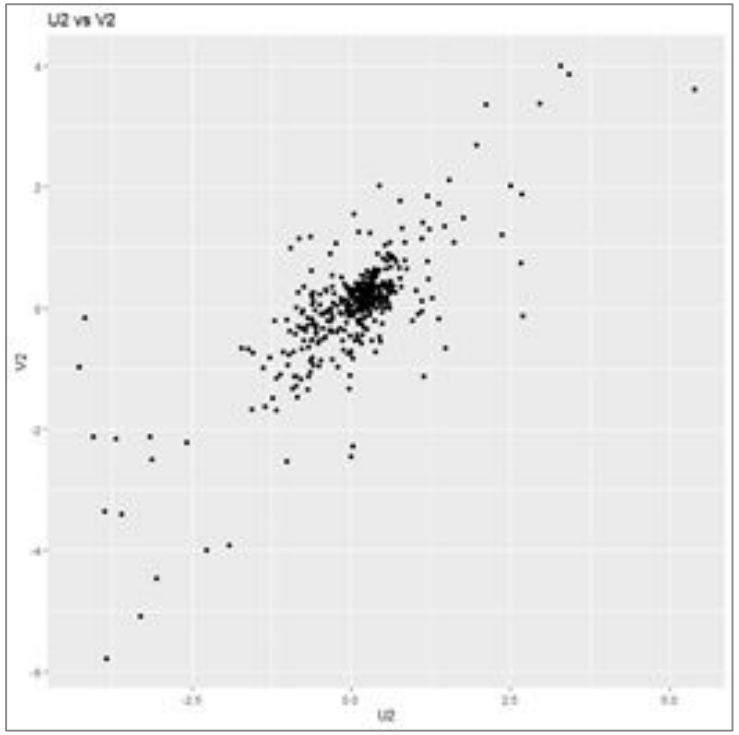
Y 와 V 의 상관관계이다. Y 의 변수 모두 첫 번째 정준변수 V1 과의 상관관계가 매우 높다.

- 원래의 변수들과 정준변수의 상관관계
- X와 Y 모두 첫 번째 정준변수와의 상관성이 매우 크다.
- 첫 번째 정준변수 쌍(U1,V1)만 사용하여
  두 집단 간의 관계의 대부분을 설명하는 것이 가능하다.

3. 정준상관분석(CCA)

📌새의 날개뼈와 다리뼈의 관계 분석 - 결과 시각화





- 첫 번째 정준변수 쌍의 상관관계가 매우 높음
- 두 번째 정준변수 쌍부터는 상관성이 낮아짐

EWHA,
THE FUTURE
WE CREATE

3. 정준상관분석(CCA)

📌정준상관분석 결론

- 새의 생물학적 구조에서 뼈의 길이와 지름이 상호 영향을 강하게 미치며, 날개 뼈와 다리 뼈의 구조 또한 상호 연관성이 강함을 확인할 수 있었다.

4. 군집분석(Clustering)

📌소개

- 데이터에 본래 존재하는 범주인 type을 제외한
  10개의 변수들(뼈의 길이, 뼈의 지름)을 활용하여 군집분석을 수행
- 데이터들을 적절한 군집으로 나누어
  각 군집의 특성, 군집 간의 차이 등에 대한 탐색적 분석을 수행
- 결과를 본래의 범주인 type과 비교

4. 군집분석(Clustering)

📌분석 설계

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

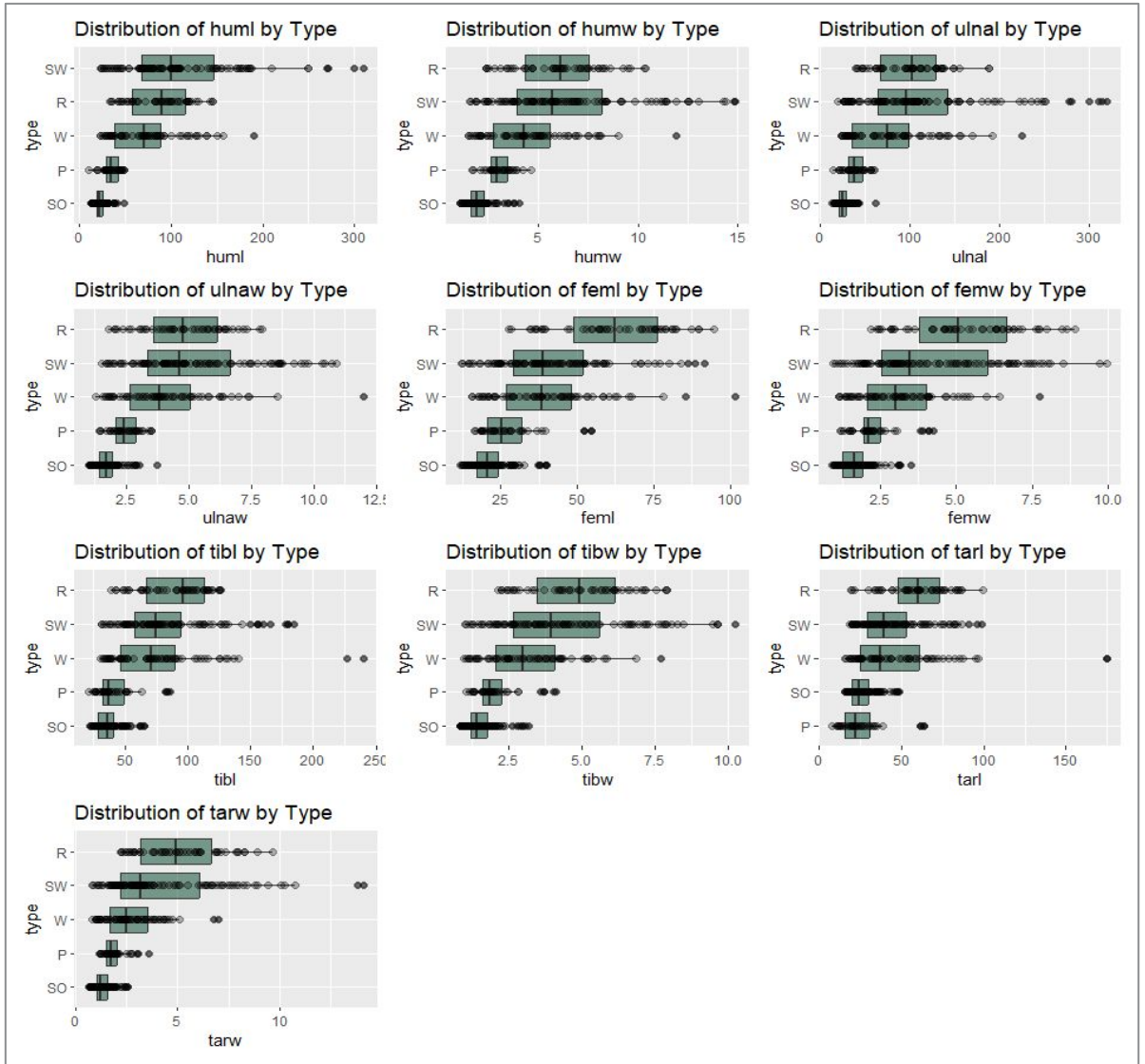
⇒ 각 type별로 군집이 형성될 것이다.

4. 군집분석(Clustering)

📌분석 1) 거리 계산

- 변수들 간의 스케일 차이가 존재

  ⇒ 표준화 거리 활용



4. 군집분석(Clustering)

📌분석 2) 군집화 방법 선택

- 계층적 군집화
  - ✓ 최단연결법
  - ✓ 최장연결법
  - ✓ 평균연결법

- 비계층적 군집화
  - ✓ K-Means

4. 군집분석(Clustering)

📌분석 2) 군집화 방법 선택

계층적 군집화

⇒ 최단연결법, 평균연결법: 대부분의 데이터가 하나의 군집으로 군집화 됨

⇒ 최장연결법: 지나치게 세분화되는 군집 존재

▼ 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
```

▼ 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
```

▼ 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
```

4. 군집분석(Clustering)

📌분석 2) 군집화 방법 선택

비계층적 군집화

⇒ 대부분의 데이터가 특정한 군집으로 군집화 됨

```
### 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
```

4. 군집분석(Clustering)

📌분석 3) 군집 분리 과정 확인 - type에 따른 분리

✓ type SO 제거

⇒ type P가 다른 type들과 비교적 잘 분리됨

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

4. 군집분석(Clustering)

📌분석 3) 군집 분리 과정 확인 - type에 따른 분리

✓ type SO, P 제거

⇒ type 간 군집화 시의 차이점을 찾기 어려움

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

4. 군집분석(Clustering)

📌분석 3) 군집 분리 과정 확인 - 변수 선택에 따른 분리

✓ 날개뼈 관련

⇒ type P, type SO의 데이터가 비교적 잘 분리됨

⇒ type SW의 경우 모든 군집에 걸쳐 산발적으로 분포

##		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

4. 군집분석(Clustering)

📌분석 3) 군집 분리 과정 확인 - 변수 선택에 따른 분리

✓ 다리뼈 관련

⇒ type P, type SO의 데이터가 비교적 잘 분리됨

⇒ type SW의 경우 모든 군집에 걸쳐 산발적으로 분포

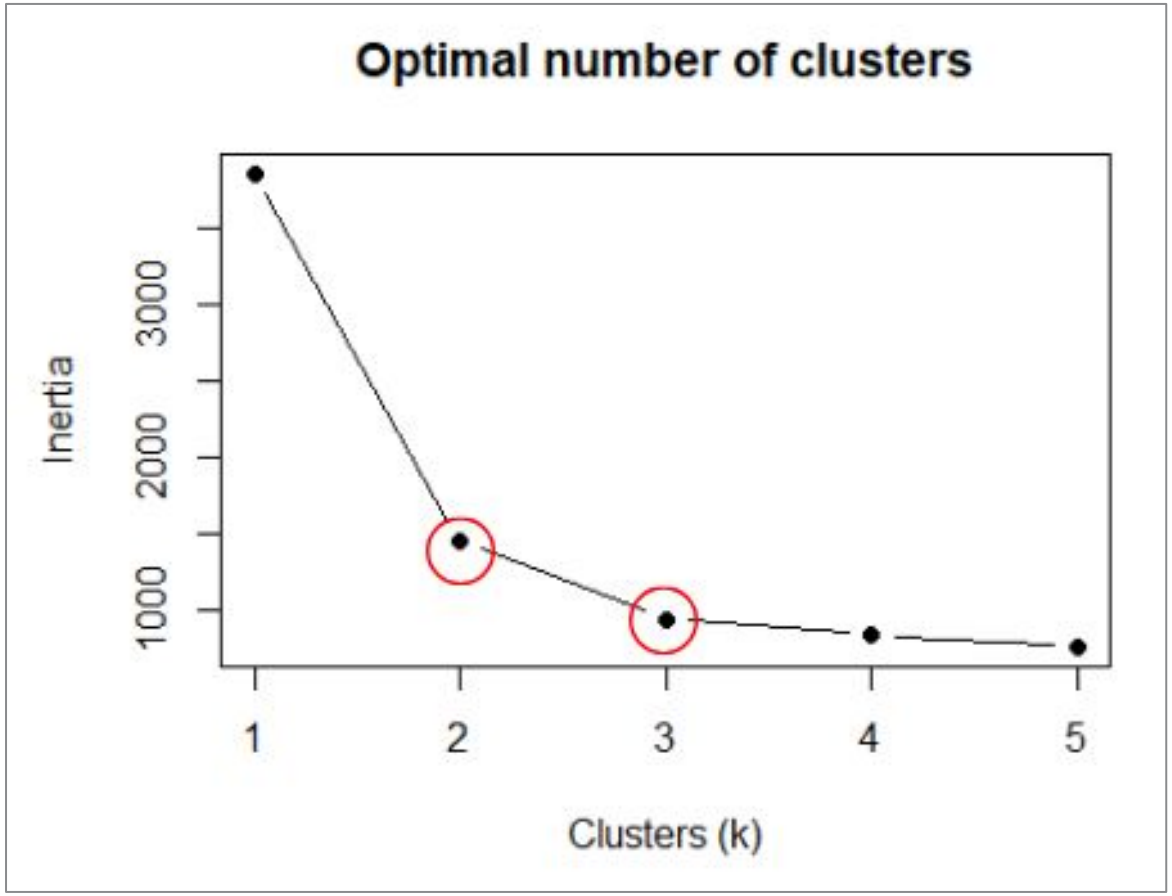
	P	R	SO	SW	W
1	0	11	0	12	2
2	21	0	99	10	15
3	12	10	26	27	18
4	5	9	3	28	20
5	0	20	0	29	10

4. 군집분석(Clustering)

📌분석 4) 적절한 군집 개수 설정

Elbow Plot

⇒ 2~3개의 군집이 적절



EWHA,
THE FUTURE
WE CREATE

4. 군집분석(Clustering)

📌분석 4) 적절한 군집 개수 설정

군집화

○ 2개 ▶

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

○ 3개 ▶

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

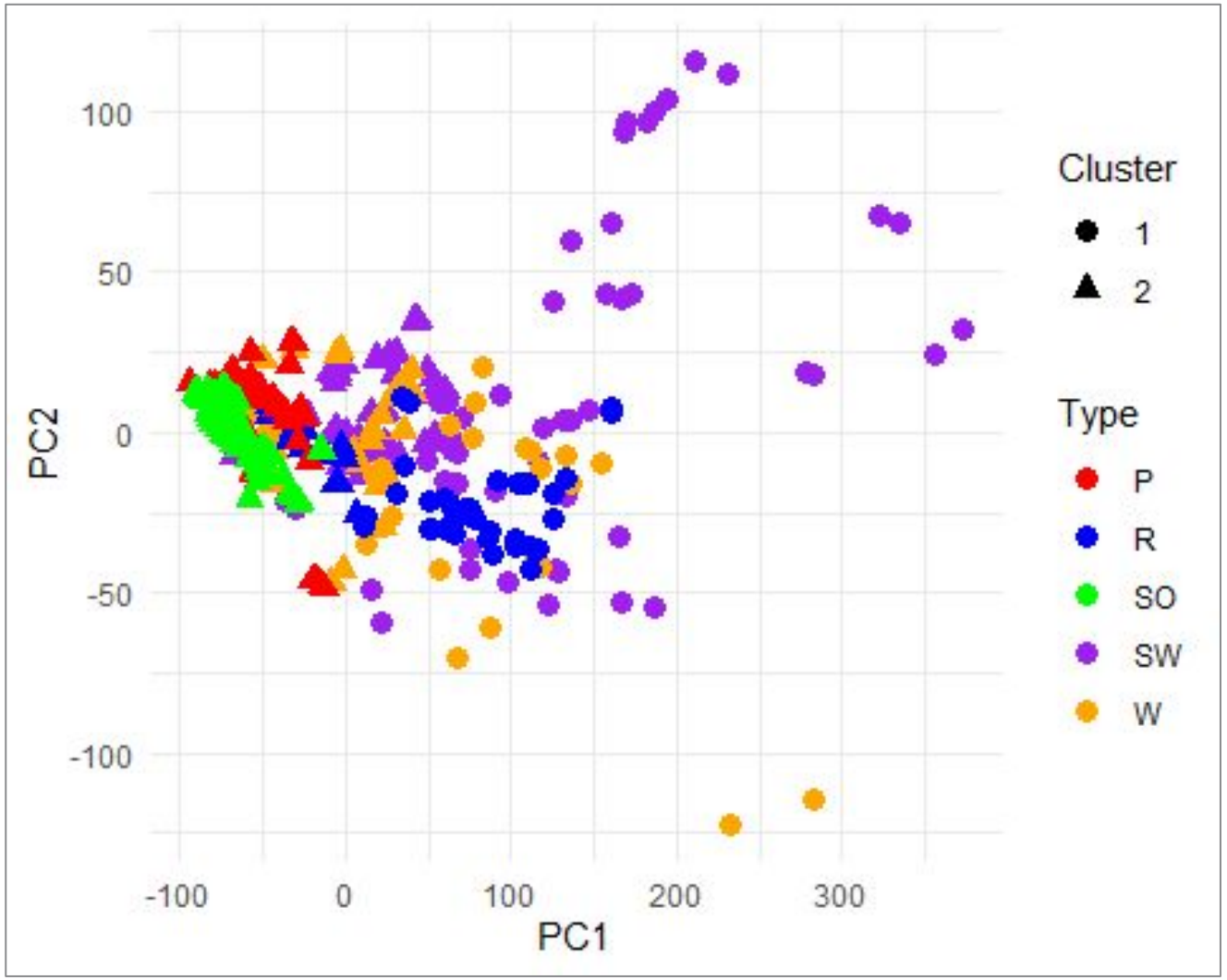
4. 군집분석(Clustering)

📌분석 5) 최종 군집화

- 차원축소

  ⇒ 10차원 → 2차원

- 결과 시각화



4. 군집분석(Clustering)

📌분석 5) 최종 군집화

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

결과

- type P와 SO는 다른 type들과 확연히 구분됨
- type SW의 경우 군집 1과 군집 2에 속하는 데이터의 비율이 50:50임
- type R의 경우 대부분의 데이터가 군집 1에 속함
- type W의 경우 데이터의 2/3가 군집 1에 속하고, 나머지가 군집 2에 속함

⇒ 군집 1) 크기가 비교적 큰 새

⇒ 군집 2) 크기가 비교적 작은 새

5. 주성분분석(PCA)

📌분석 방법

1) type별로 데이터를 분리한 후, 10차원의 자료를 잘 설명하는 주성분 탐색
2) 분산공분산행렬(S) 대신, 상관행렬(R)을 이용
3) 2차원으로 축소된 데이터로 분류 분석 진행

📌분석 목표

“10차원의 자료를 잘 설명하는 주성분을 찾고, 그 의미를 해석해보자”

5. 주성분분석(PCA)

📌산악지대 서식 조류(type P) 데이터에 대한 PCA 결과

```
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
```

- 마지막 고유값이 0에 가까움 → 변수들 간 공선성 문제 존재
- 첫번째 주성분만으로 전체 변동 중 75.6% 설명 가능

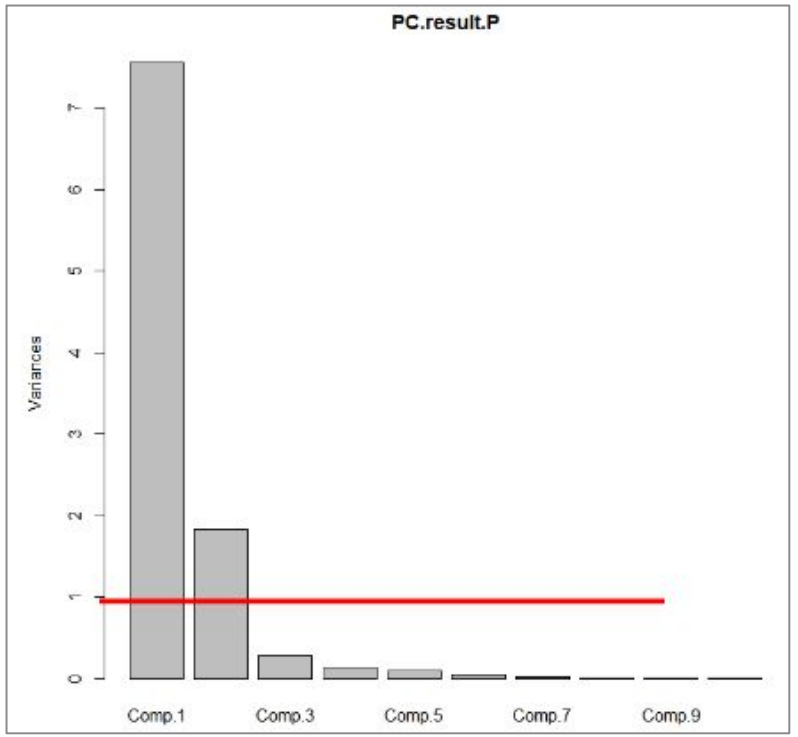
5. 주성분분석(PCA)

📌산악지대 서식 조류(type P) 데이터에 대한 PCA 결과

`PC.result.P$loadings`

주성분 번호	1	2	1	2
huml	0.315	0.307	+	+
humw	0.267	0.445	(+)	+
ulnal	0.223	0.563	(+)	+
ulnaw	0.336	0.233	+	(+)
feml	0.342	-0.232	+	(-)
femw	0.348	-0.161	+	
tibl	0.331	-0.276	+	(-)
tibw	0.345	-0.194	+	
tarl	0.301	-0.375	+	-
tarw	0.330		+	
누적변동비율(%)	75.6	93.9		

- PC1 : 전반적인 골격 크기의 가중평균을 나타내는 축
- PC2 : 날개 관련 뼈(huml, humw, ulnal)와 다리 관련 뼈(tarl)의 대비



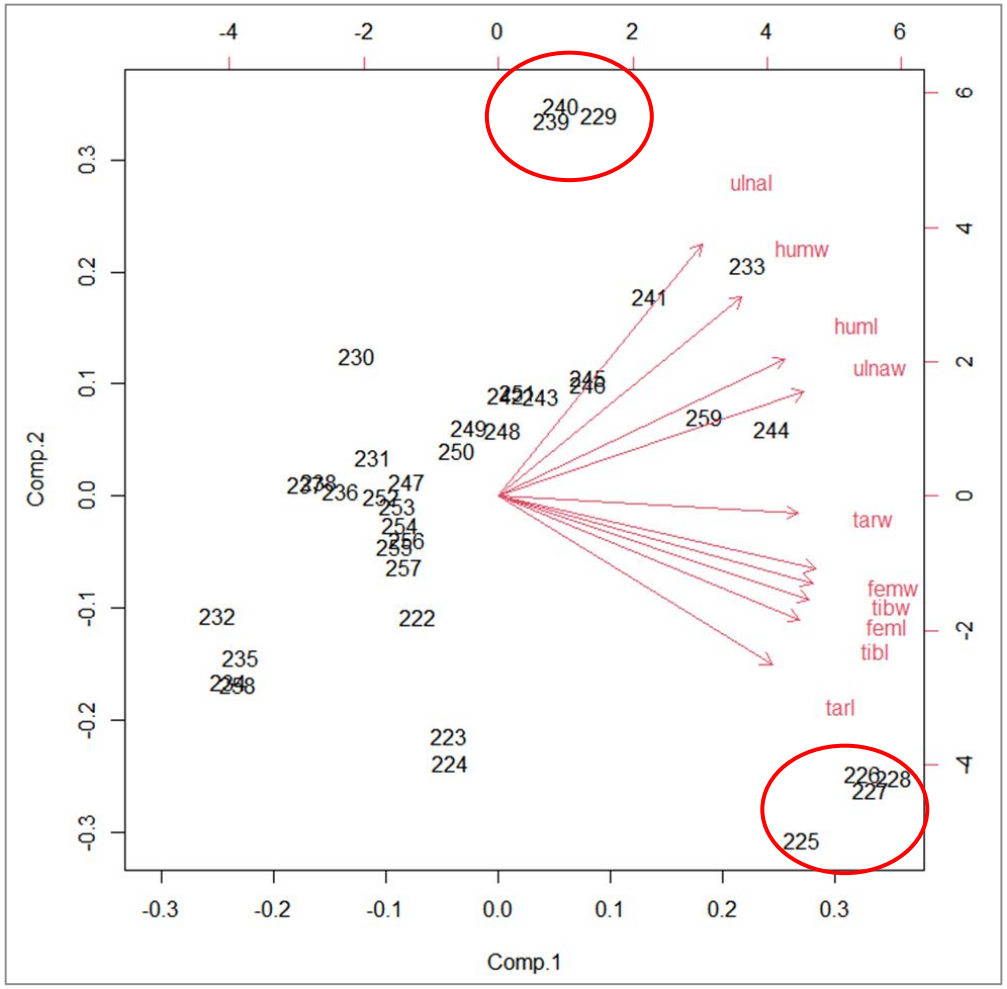
⇒ 고유값이 1 이상인

Comp.2까지 주성분 2개 선택

EWHA,
THE FUTURE
WE CREATE

5. 주성분분석(PCA)

📌산악지대 서식 조류(type P) 데이터에 대한 PCA 결과



[225, 226, 227, 228번 개체]

- 상대적으로 PC1 값이 크고, PC2 값이 작은 개체
- 전체적인 골격이 크고, 날개뼈에 비해 다리뼈가 더 발달한 개체

[229, 239, 240번 개체]

- 상대적으로 PC2 값이 큰 개체
- 다리뼈에 비해 날개뼈가 더 발달한 개체

EWHA,
THE FUTURE
WE CREATE

5. 주성분분석(PCA)

📌전체 데이터에 대한 PCA 결과

```
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
```

- 마지막 고유값이 0에 가까움 → 변수들 간 공선성 문제 존재
- 첫번째 주성분만으로 전체 변동 중 85.6% 설명 가능

5. 주성분분석(PCA)

📌전체 데이터에 대한 PCA 결과

`PC.result$loadings`

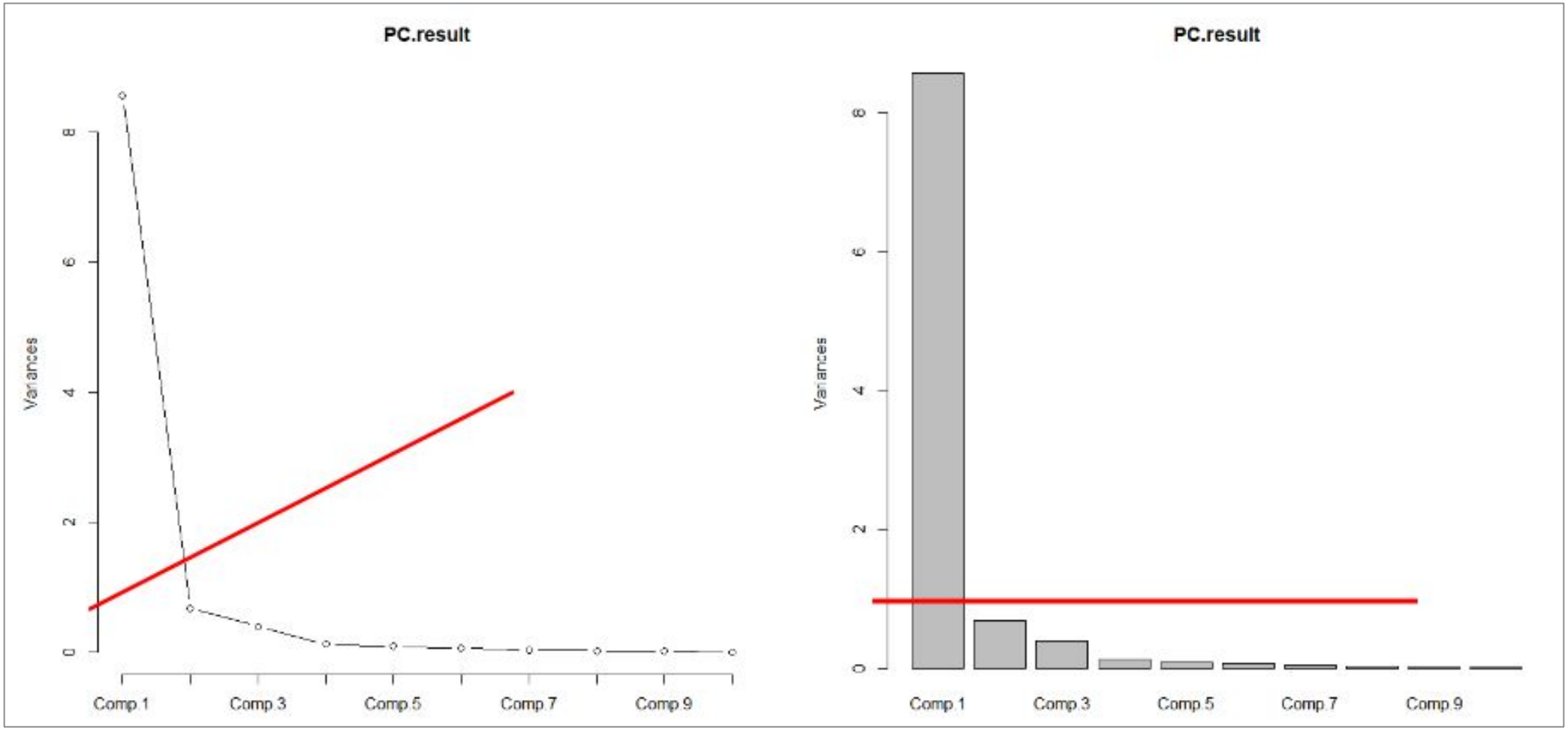
주성분 번호	1	2	3	1	2	3
huml	0.319	0.189	0.480	+	(+)	+
humw	0.328	0.277		+	(+)	
ulnal	0.311	0.256	0.498	+	(+)	+
ulnaw	0.320	0.287		+	(+)	
feml	0.321	-0.178	-0.321	+		-
femw	0.333		-0.191	+		
tibl	0.315	-0.398	0.102	+	-	
tibw	0.331	-0.103		+		
tarl	0.275	-0.695	0.138	+	-	
tarw	0.304	0.235	-0.587	+	(+)	-
누적변동비율(%)	85.6	92.4	96.4			

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

5. 주성분분석(PCA)

📌전체 데이터에 대한 PCA 결과

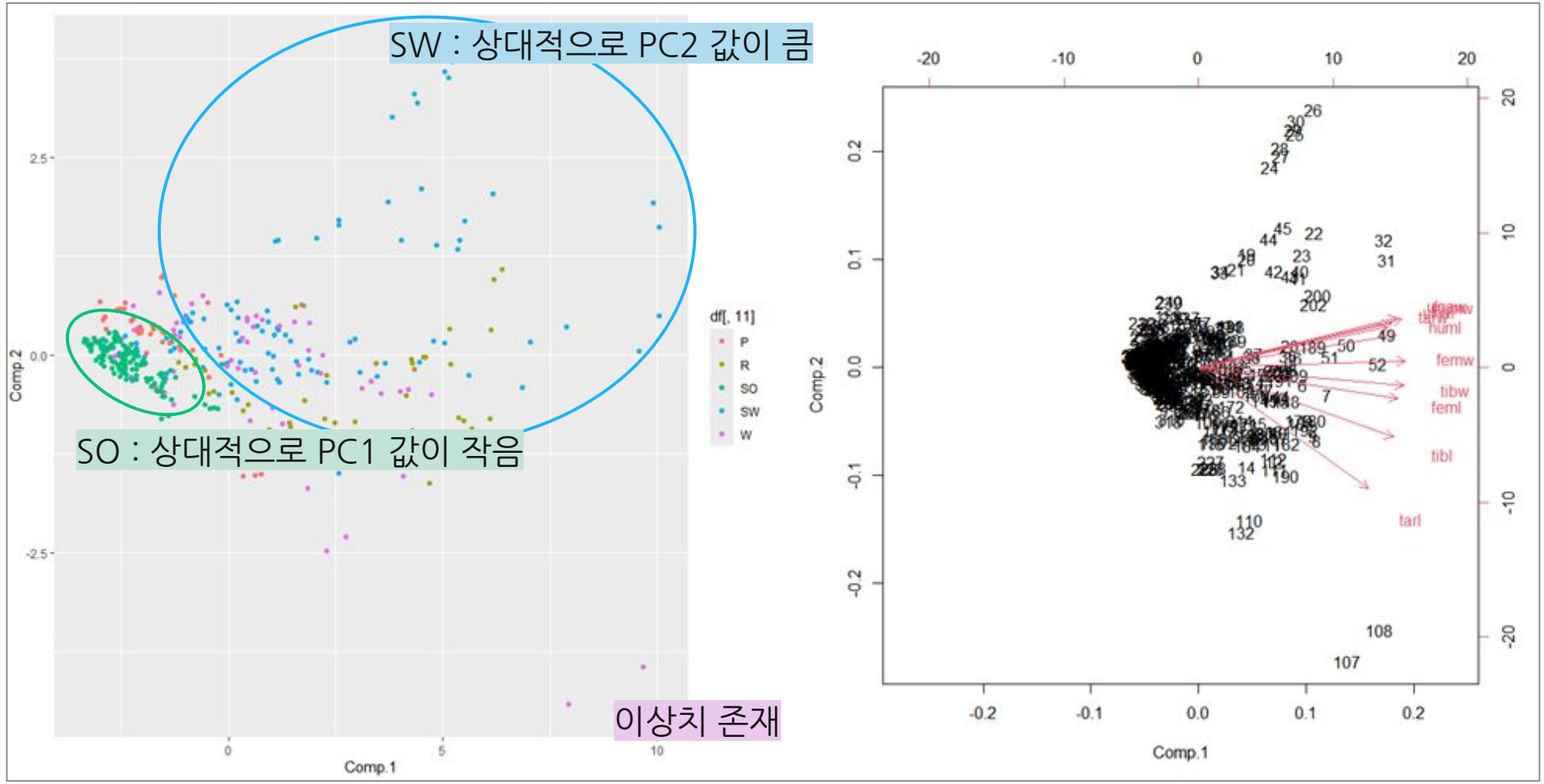
주성분 1개로 충분히 설명 가능



EWHA,
THE FUTURE
WE CREATE

5. 주성분분석(PCA)

📌전체 데이터에 대한 PCA 결과



EWHA,
THE FUTURE
WE CREATE

5. 주성분분석(PCA)

📌PCA 분석 결론

1) 첫번째 주성분
   - 모든 경우에서 첫번째 주성분으로 전체 변동의 대부분을 설명 가능 (75% 이상)
   - 첫번째 주성분의 의미 : overall mean

2) 두번째 주성분
   - 각 type별로 설명하고자 하는 것이 조금씩 다름
   - 두번째 주성분의 의미 : 날개뼈와 다리뼈의 대비, 여러 뼈 사이의 대비나 가중평균

3) 전체 PCA를 통해,
   - 이상치 발견
   - SO 개체 : 조밀한 분포, 작은 골격
   - SW 개체 : 넓은 분포, 다리뼈 길이가 짧은 개체

6. 판별분석(LDA)

📌분석 방법

1) 주성분을 이용한 LDA
2) 원본 데이터 전체를 사용한 LDA
3) 훈련용 데이터와 테스트용 데이터를 분리하여 진행한 LDA

⇒ 3개의 결과를 비교 분석

📌분석 목표

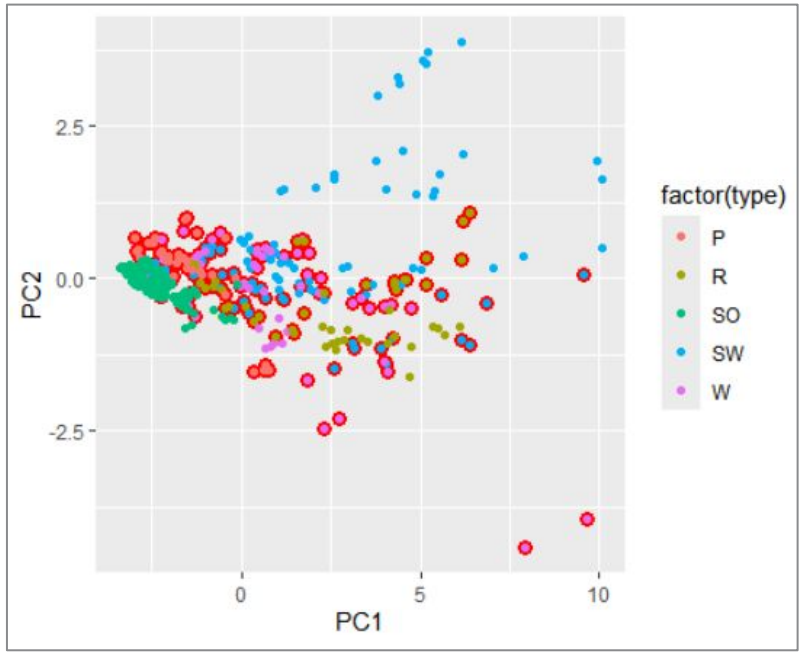
“골격 정보를 이용해서 새의 type을 옳게 분류해보자”

6. 판별분석(LDA)

📌2개의 주성분을 이용한 LDA

```
## 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
```

```
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
```



- 오분류율 : 44.7% / 정확도 : 55.3%
- P : 옳게 분류된 데이터가 하나도 없음
- SO : 모두 옳게 분류됨

EWHA,
THE FUTURE
WE CREATE

6. 판별분석(LDA)

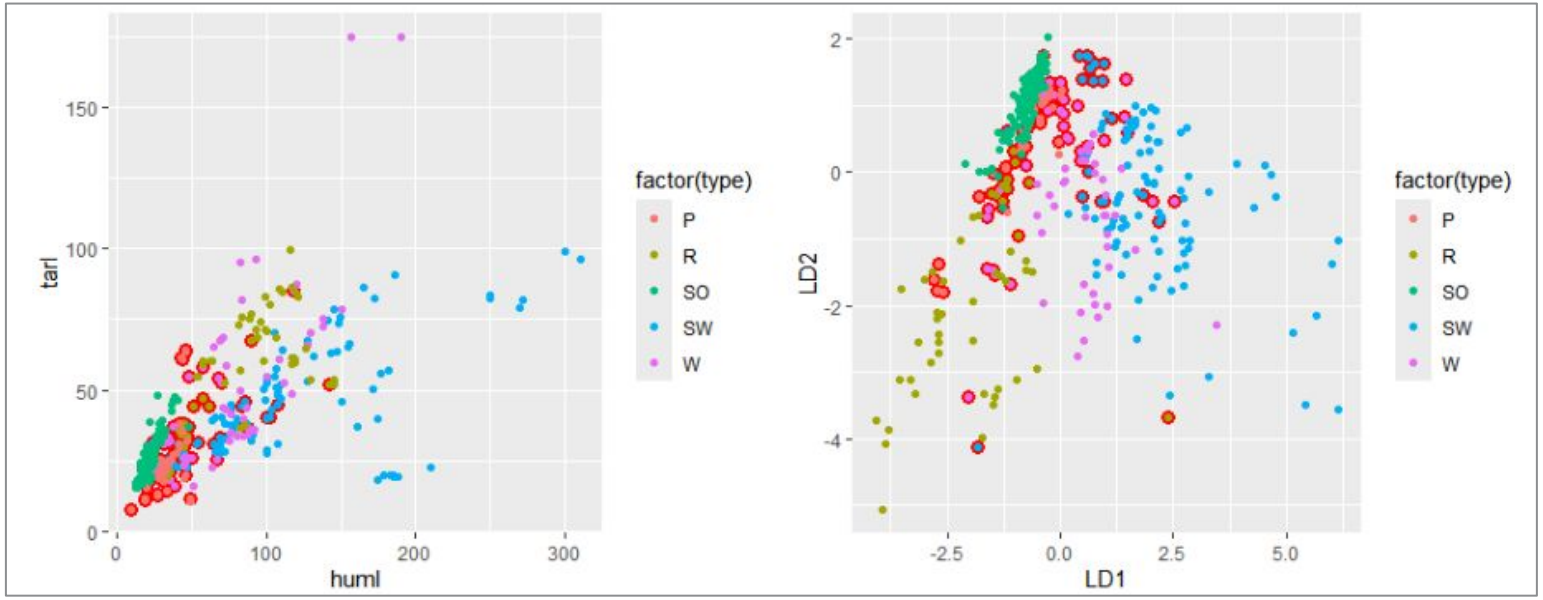
📌 원본 데이터를 이용한 LDA

```
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
```

```
## 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
```

```
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
```

- 오분류율 : 25.8% / 정확도 : 74.2%
- P : 대부분 SO로 분류됨
- SO : 하나의 데이터를 제외하고는 모두 옳게 분류됨



6. 판별분석(LDA)

📌 훈련용/테스트용 분리 LDA

```
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]
```

```
## 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
```

```
## 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
```

```
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
```

```
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
```

[train error]

- 오분류율 : 27.0% / 정확도 : 73.0%

[test error]

- 오분류율 : 29.1% / 정확도 : 70.9%

[train/test 분리 과정]

- train(70%)/test(30%) random split
- stratify 진행 x

EWHA,
THE FUTURE
WE CREATE

6. 판별분석(LDA)

📌 LDA 분석 결론

LDA	주성분 2개 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%

1) 주성분 2개 vs 원본 데이터 :

2개의 주성분이 전체 데이터의 90% 이상을 설명하더라도 정보 손실이 존재

→ 원본 데이터보다 분류 성능이 낮음

2) 훈련용 데이터 vs 테스트용 데이터 :

훈련용 데이터보다 테스트용 데이터에서 성능이 약간 낮음

→ 훈련용 데이터에 더 적합한 모델 (약간의 과적합 가능성)

6. 판별분석(LDA)

📌 (+추가) QDA 결과

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

```
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
```

- 오분류율 : 11.1% / 정확도 : 88.9%
- P : LDA 결과와 달리 옳게 분류됨
- 집단마다 분산이 다를 때는 선형판별규칙(LDA)보다 이차판별규칙(QDA)을 이용한 분류 방법의 성능이 더 좋음