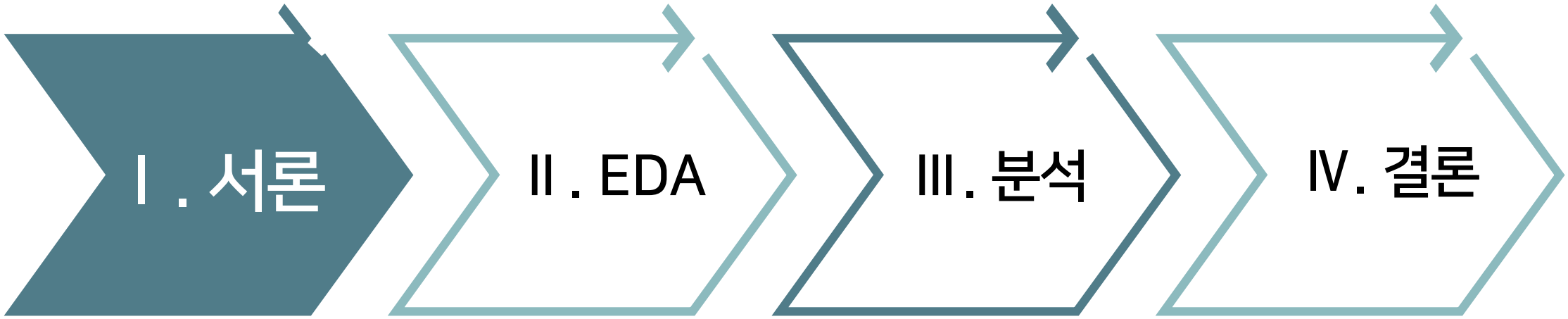
다변량 데이터 분석 프로젝트

- 국민체력실태조사 자료

6조
1602045 박정현
1602050 박지원
1602069 예지혜
1602073 유은지

목차

Ⅰ. 서론



1.자료 설명

Ⅰ. 서론 – (1) 자료 설명

1. 데이터

- 전국 17개 시,도 19~64세 성인의 국민체력실태조사

- 변수 16개, 관측값 4291개

변수 이름	변수 타입	단위
지역	CHAR , 범주형	
연령	INT	
연령집단	CHAR, 범주형	
성별	INT, 범주형	
신장	INT	0.1cm 단위 측정
체중	INT	0.1kg 단위 측정
BMI	INT	체중(kg)/신장(㎡)
체지방률	INT	0.1% 단위
허리둘레	INT	0.1cm 단위
윗몸일으키기	INT	회/1분
악력(D)	INT	악력(0.1kg 단위) ,D: 쓰는 손 ,ND: 안 쓰는 손
악력(ND)	INT	
제자리멀리뛰기	INT	0.1cm 단위
20m 왕복오래달리기	INT	회
앉아윗몸앞으로굽히기	INT	0.1cm 단위
10m 왕복달리기	INT	0.01초 단위

Ⅱ. EDA

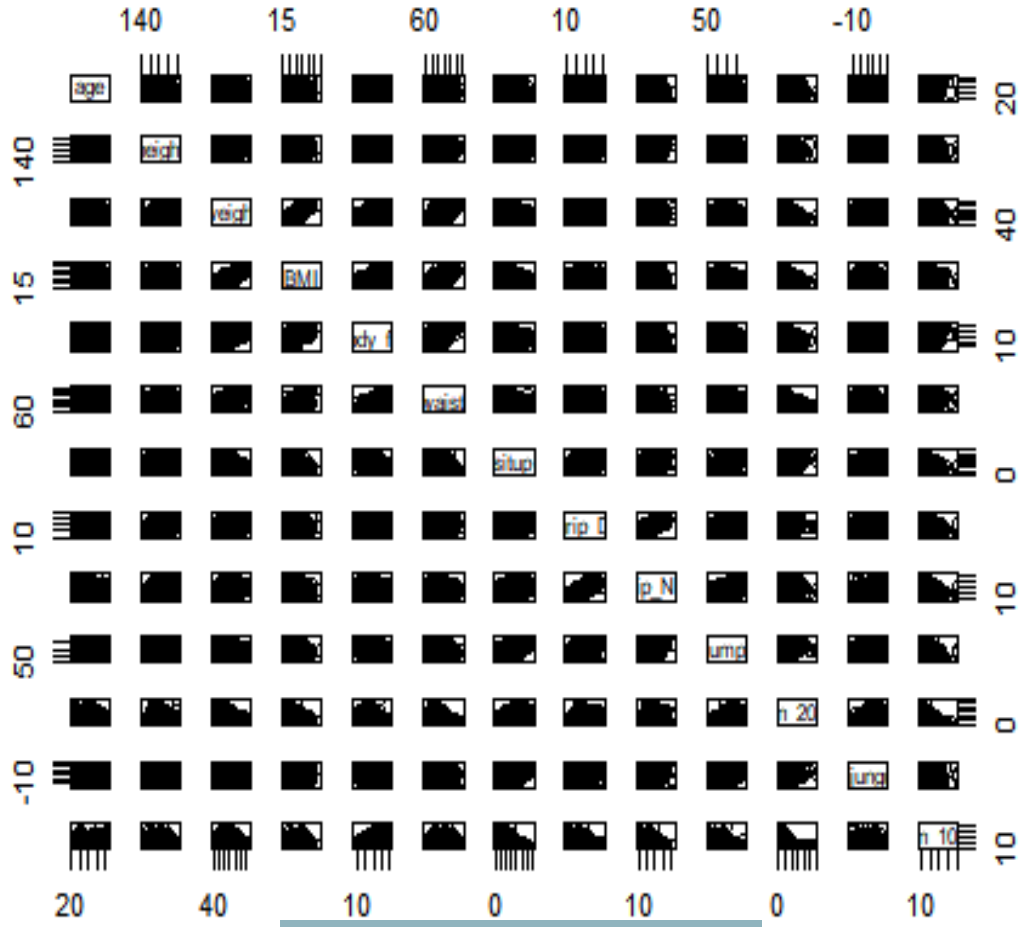
1. 자료 살펴보기

2. 주성분 분석

3. 가설 설정 및 분석 방향 제시

II . EDA – (1) 자료 살펴보기

1. 변수 간 상관관계



상관관계 표

나이 ↑

윗몸일으키기, 멀리뛰기, 20m왕복오래달리기성적 ↓

체중, BMI, 체지방률, 허리둘레 사이

➔ 양의 상관성이 존재

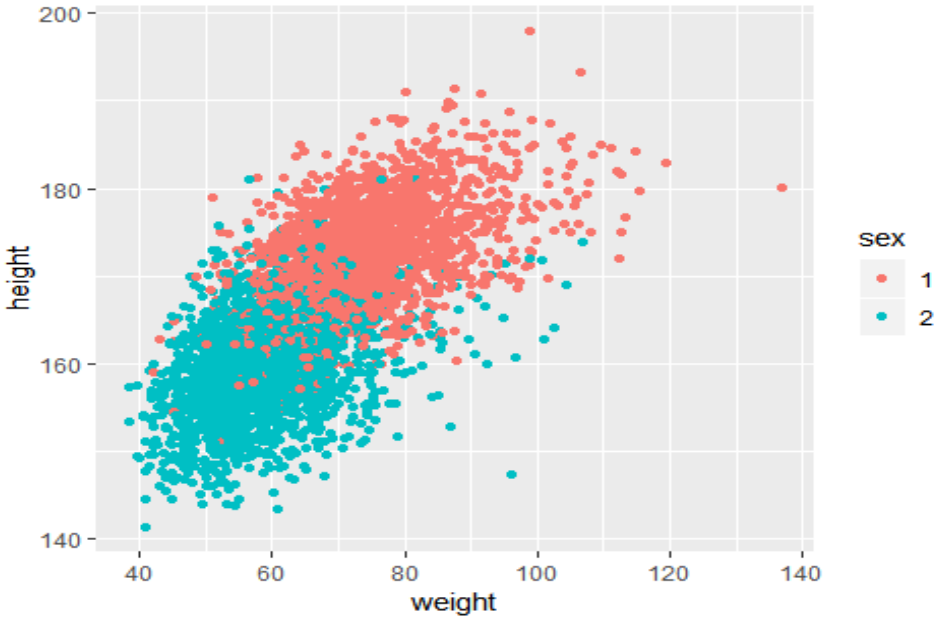
변수들의 대부분은 운동 변수들과 음의 상관성을 보이지만, 악력과 10m 양의 상관성을 가지는 모습

윗몸일으키기, 악력, 제자리멀리뛰기, 20m왕복오래달리기와도 양의 상관성이 존재한다

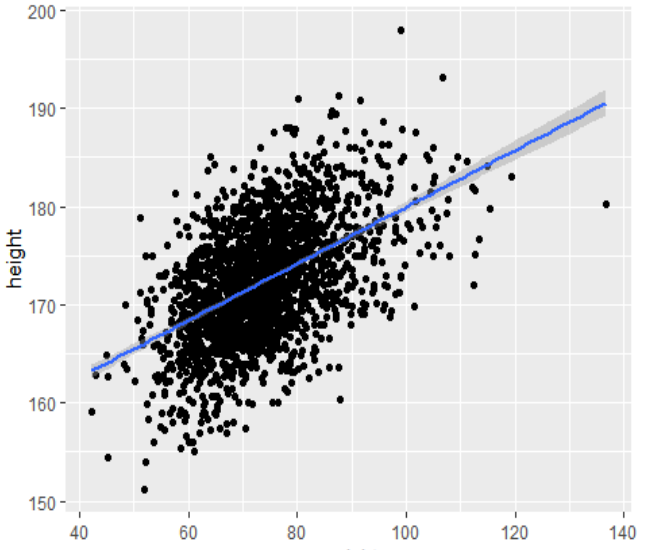
II . EDA – (1) 자료 살펴보기

2. 자료 시각화

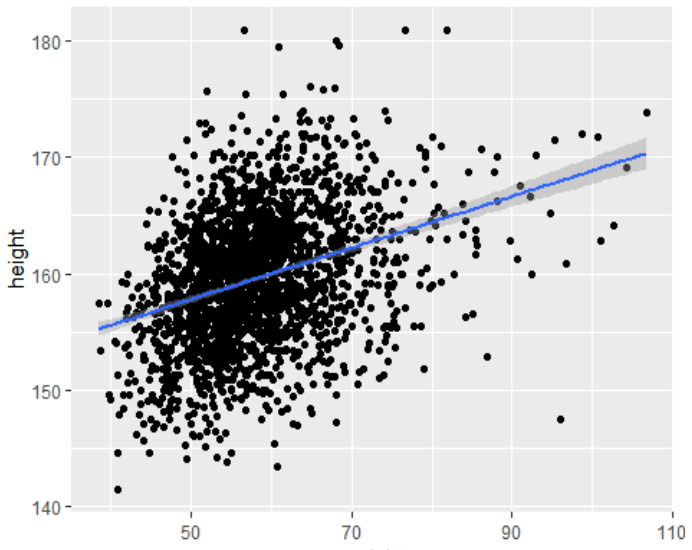
[남녀 데이터 나누어 살펴보기] – 키와 몸무게의 성별 별 산점도



키와 몸무게별로 산점도를 그려본 결과,
남성과 여성의 차이가 매우 큰 것을 확인 가능



남성 키/몸무게 산점도



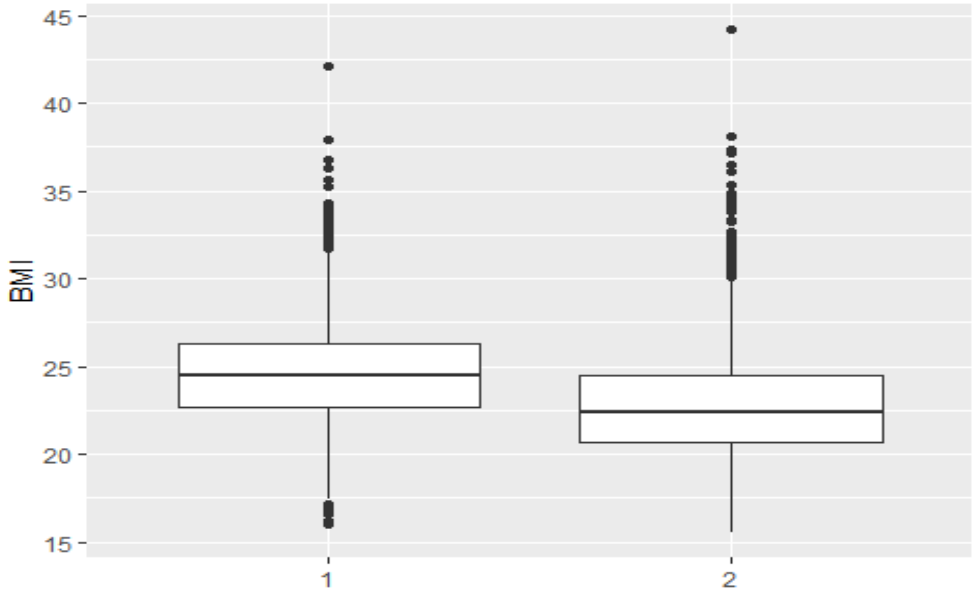
여성 키/몸무게 산점도

전반적으로 남성이 같은 몸무게에 대해 키의 분포가 모여 있는 편

II . EDA – (1) 자료 살펴보기

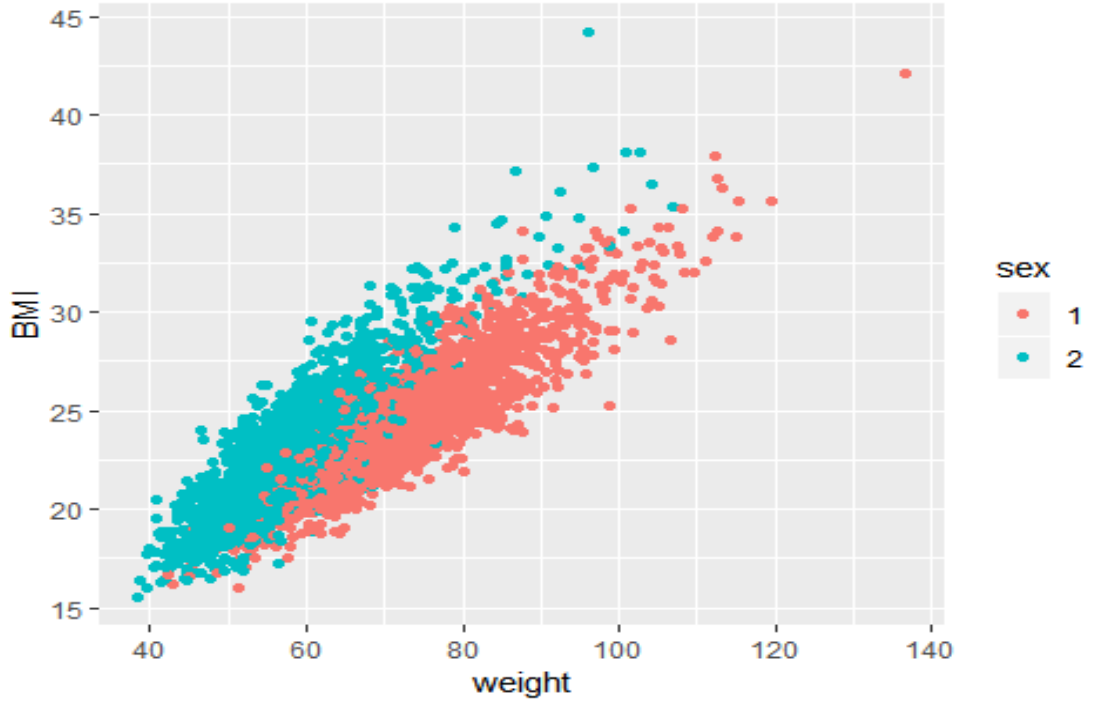
2. 자료 시각화

[남녀 데이터 나누어 살펴보기] – BMI



BMI boxplot

여성이 남성보다 BMI가 낮은 편

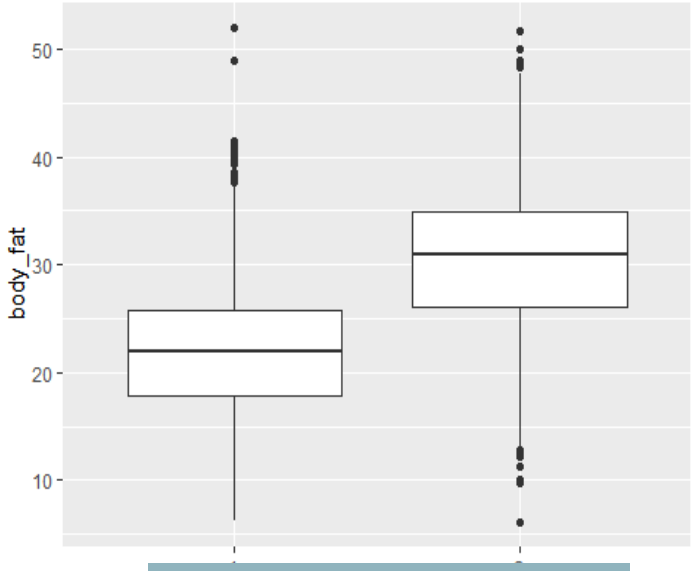


같은 몸무게일때는 상대적으로 키가 작은 여성의 BMI가 더 높은 편

II . EDA – (1) 자료 살펴보기

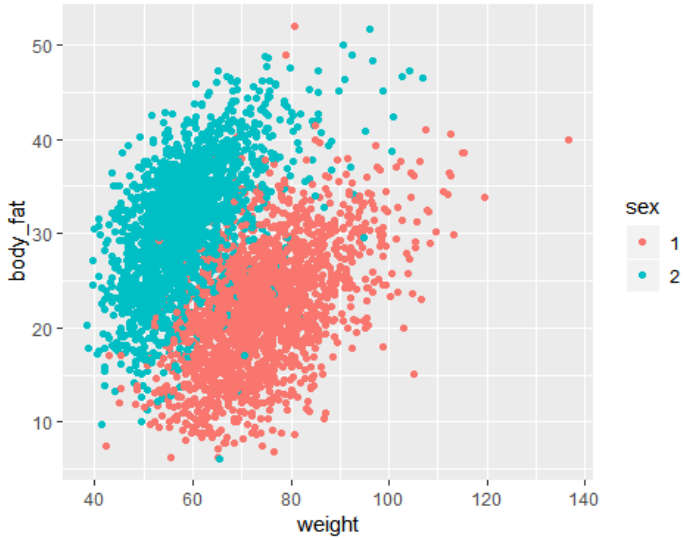
2. 자료 시각화

[남녀 데이터 나누어 살펴보기] – 체지방률

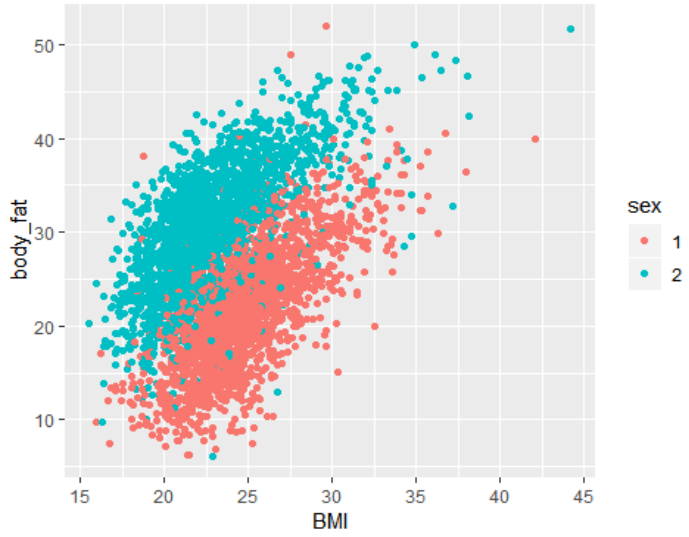


체지방률 boxplot

체지방률은 확연히 여성이 남성보다 높음



체지방률/몸무게 산점도



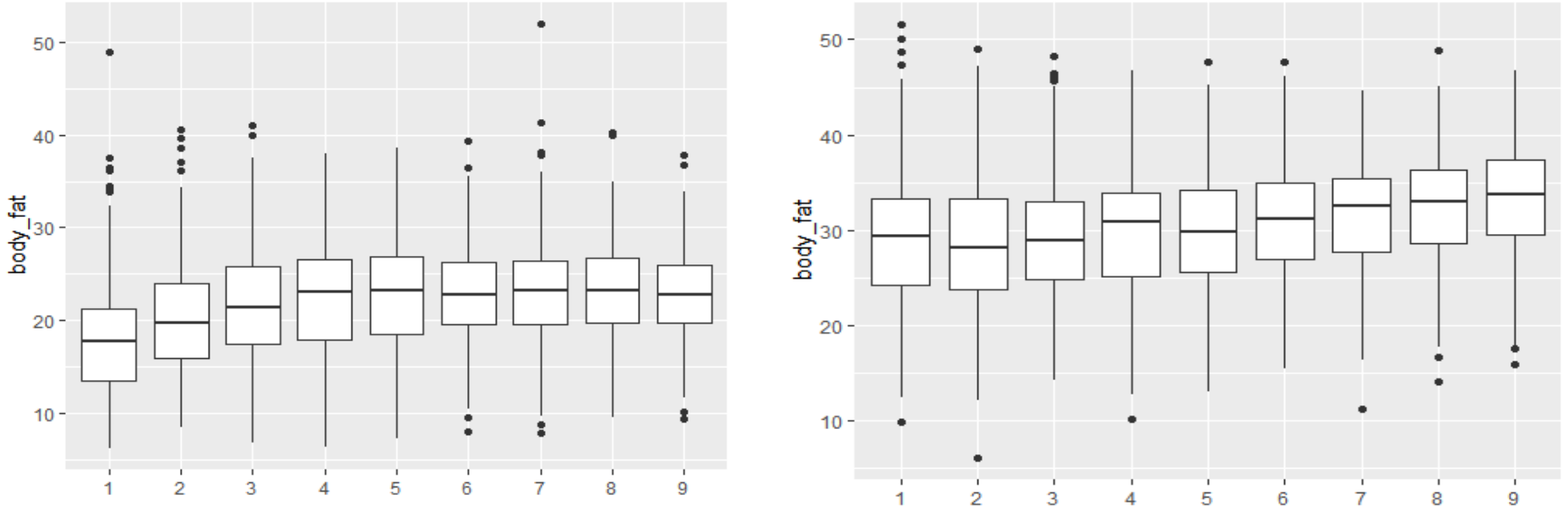
체지방률BMI 산점도

체지방률은 몸무게보다는 BMI와 연관성이 더 높은 것으로 보임

II . EDA – (1) 자료 살펴보기

2. 자료 시각화

[연령 데이터 나누어 살펴보기] – 체지방률



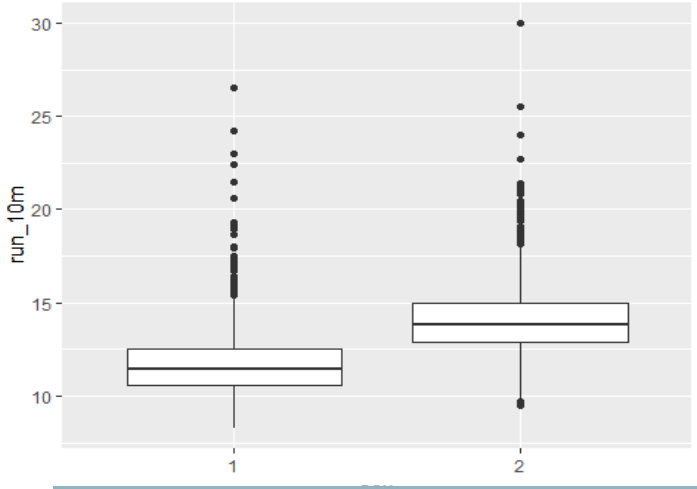
연령그룹별 체지방률 boxplot

남성과 여성 모두 나이가 들면 평균 체지방률이 높아진다.

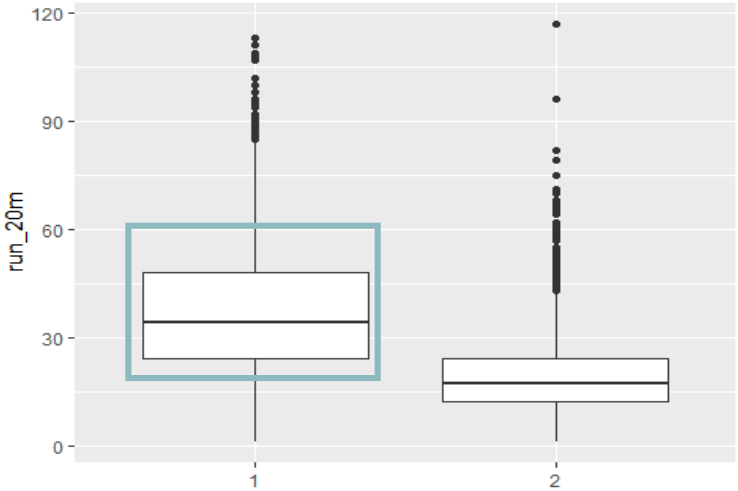
Ⅱ. EDA – (1) 자료 살펴보기

2. 자료 시각화

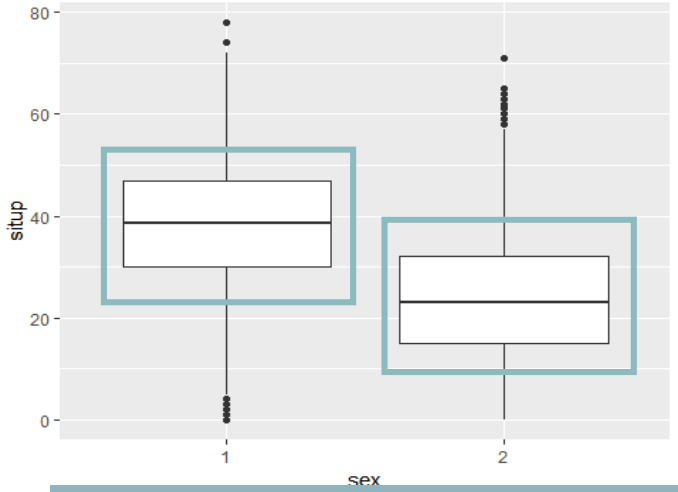
[남녀 데이터 나누어 살펴보기] – 체력데이터



성별 10m 왕복달리기 횟수 boxplot



성별 20m 왕복달리기 기록 boxplot



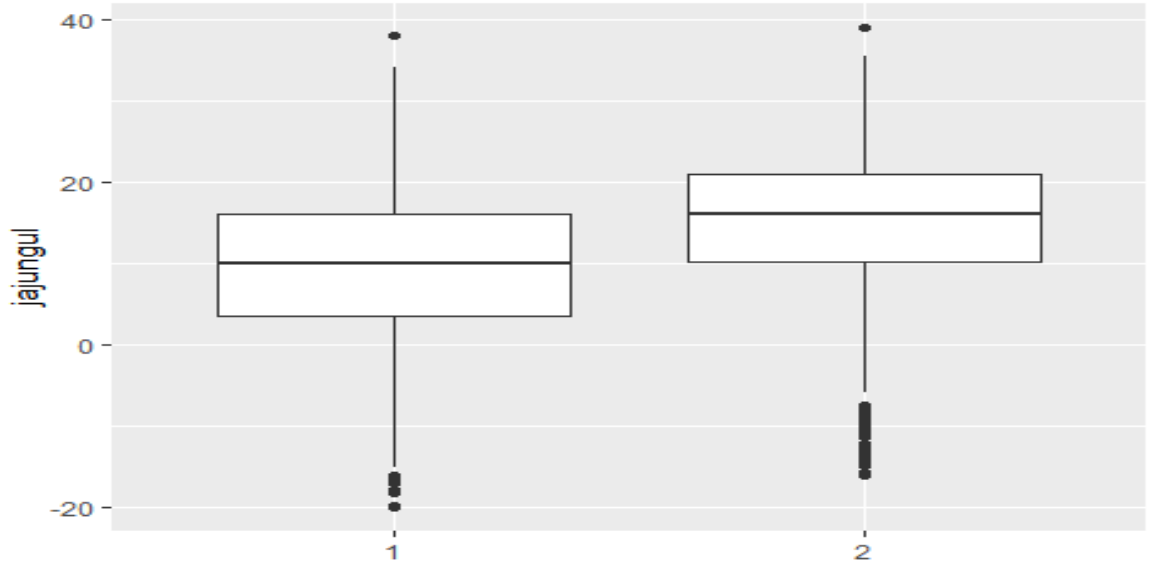
성별 윗몸일으키기 횟수boxplot

체력적 기록은 성별 차이가 확연히 나는 편이다. 달리기 같은 경우 개인간의 편차도 심하다.
10m 달리기 기록과 달리 20m왕복 오래 달리기 횟수의 경우 남성 분포의 분산이 굉장히 큰 편이다.
윗몸일으키기(situp)는 남녀 모두 개인차가 크다.

II . EDA – (1) 자료 살펴보기

2. 자료 시각화

[남녀 데이터 나누어 살펴보기] – 체력데이터



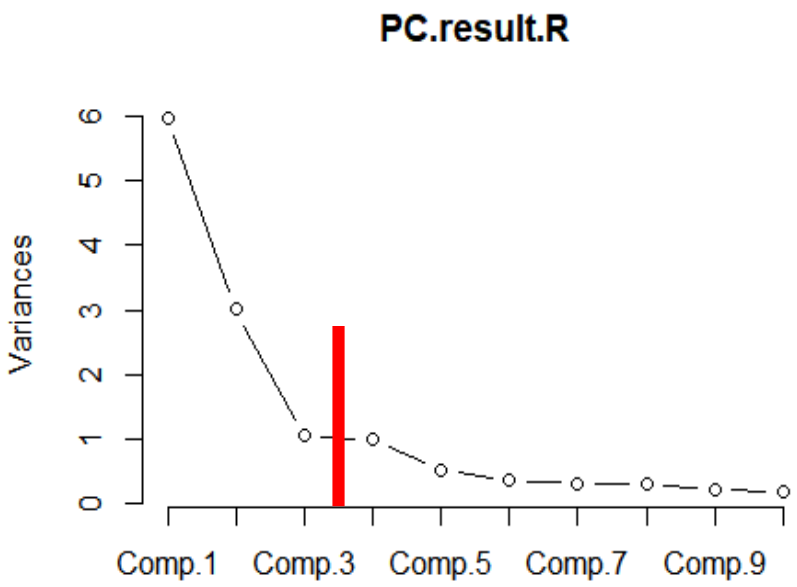
성별 윗몸 앞으로 굽히기 기록boxplot

유연성을 나타내는 앉아 윗몸 앞으로 굽히기는 여성이 남성보다 더 높은 기록을 보인다.

II . EDA – (2) 주성분 분석

1. 전체 데이터 주성분 분석

[주성분 개수 선택]




```
summary(PC.result.R)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.4463513 1.7328862 1.02065839 0.98890695
## Proportion of Variance 0.4603565 0.2309919 0.08013412 0.07522592
## Cumulative Proportion  0.4603565 0.6913484 0.77148250 0.84670842
##                           Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     0.70732716 0.58381340 0.54466226 0.53704404
## Proportion of Variance 0.03848552 0.02621831 0.02281977 0.02218587
## Cumulative Proportion  0.88519394 0.91141226 0.93423202 0.95641789
##                           Comp.9    Comp.10    Comp.11     Comp.12
## Standard deviation     0.4406494 0.41800366 0.37320474 0.233289872
## Proportion of Variance 0.0149363 0.01344054 0.01071398 0.004186474
## Cumulative Proportion  0.9713542 0.98479474 0.99550872 0.999695193
##                           Comp.13
## Standard deviation     0.0629482815
## Proportion of Variance 0.0003048066
## Cumulative Proportion  1.0000000000
```

➔ 설명력이 높은 3개의 주성분 선택

II . EDA – (2) 주성분 분석

1. 전체 데이터 주성분 분석

[주성분 loading]

```
PC.result.R$loadings

##
## Loadings:
##          Comp.1 Comp.2 Comp.3
## age       0.132  0.190  0.190
## height   -0.332         0.291
## weight   -0.268  0.412
## BMI      -0.109  0.499 -0.305
## bodyfat   0.261  0.322 -0.291
## waist    -0.127  0.501 -0.121
## situp    -0.314 -0.192 -0.219
## grip_D   -0.370
## grip_ND  -0.366
## jump     -0.368 -0.116
## run_20   -0.302 -0.209 -0.148
## flexion         -0.211 -0.766
## run_10    0.327  0.169  0.123
```

[주성분 변수 해석]

주성분	해석
1	나이, 체지방률 대비 나머지 변수
2	몸무게, BMI, 체지방률, 허리둘레 = 비만도
3	세번째 주성분 Comp.3 = 유연성

Ⅱ . EDA – (3) 가설 설정 및 분석 방향 제시

가설 1. 성별별로 차이가 있을 것이다.

가설 2. 연령별로 차이가 있을 것이다.

가설 3. 지역별로 차이가 있을 것이다.

가설 4. 체격 특징과 체력적 특징의 관련성이 존재할 것이다.

Ⅲ. 분석

Ⅰ. 서론 → Ⅱ. EDA → **Ⅲ. 분석** → Ⅳ. 결론

1. 성별별 차이

2. 연령별 차이

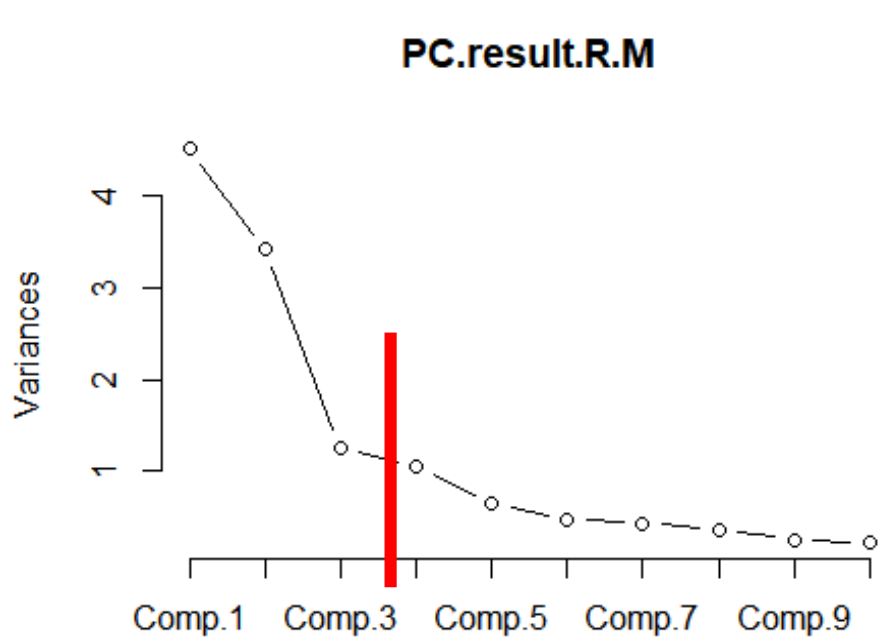
3. 지역별 차이

4. 체력과 체격 데이터의 연관성

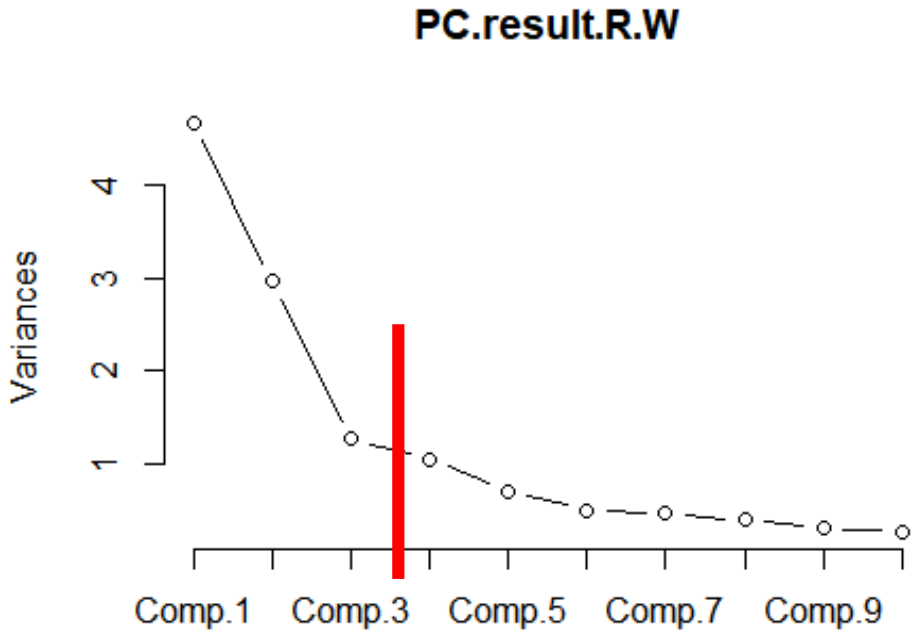
Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[주성분 분석] - 주성분 개수 선택



남성 데이터 screeplot



여성 데이터 screeplot

➔ 설명력이 높은 3개의 주성분 선택

Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[주성분 분석] – 남녀 데이터 주성분 비교

[남자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.270	0.107	0.517
height	-0.151	-0.250	-0.393
weight		-0.514	-0.109
BMI	0.162	-0.455	
bodyfat	0.320	-0.266	
waist	0.202	-0.433	
situp	-0.369		
grip_D	-0.246	-0.307	0.325
grip_ND	-0.242	-0.300	0.376
jump	-0.398		
run_20	-0.367		
flexion	-0.220		0.532
run_10	0.364		0.129

[여자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.243		0.515
height	-0.151	-0.214	-0.540
weight	0.185	-0.497	-0.145
BMI	0.273	-0.407	0.126
bodyfat	0.349	-0.196	
waist	0.280	-0.381	
situp	-0.351	-0.114	
grip_D	-0.211	-0.395	0.183
grip_ND	-0.214	-0.377	0.216
jump	-0.360	-0.147	
run_20	-0.338		
flexion	-0.169		0.565
run_10	0.351	0.115	

[주성분 1 비교]

두 변수 모두 전체 데이터의 주성분 1과 같이 신체적인 측정 대비 체력적인 측정 변수의 효과를 나타냄

남성의 경우 여성과 다르게 체중의 효과가 없음

Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[주성분 분석] – 남녀 데이터 주성분 비교

[남자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.270	0.107	0.517
height	-0.151	-0.250	-0.393
weight		-0.514	-0.109
BMI	0.162	-0.455	
bodyfat	0.320	-0.266	
waist	0.202	-0.433	
situp	-0.369		
grip_D	-0.246	-0.307	0.325
grip_ND	-0.242	-0.300	0.376
jump	-0.398		
run_20	-0.367		
flexion	-0.220		0.532
run_10	0.364		0.129

[여자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.243		0.515
height	-0.151	-0.214	-0.540
weight	0.185	-0.497	-0.145
BMI	0.273	-0.407	0.126
bodyfat	0.349	-0.196	
waist	0.280	-0.381	
situp	-0.351	-0.114	
grip_D	-0.211	-0.395	0.183
grip_ND	-0.214	-0.377	0.216
jump	-0.360	-0.147	
run_20	-0.338		
flexion	-0.169		0.565
run_10	0.351	0.115	

[주성분 2 비교]

여자 주성분 2: 전체 데이터 주성분 2에서 영향력 있었던 변수
: 몸무게, BMI, 체지방률, 허리둘레
+
체력데이터 :악력과 멀리뛰기

윗몸일으키기와 멀리뛰기는 그 효과가 크지는 않다.

남성 주성분 데이터에서는 해당 변수들의 효과가 아예 나타나지 않는다는 점에서는 의미를 가짐

Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[주성분 분석] – 남녀 데이터 주성분 비교

[남자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.270	0.107	0.517
height	-0.151	-0.250	-0.393
weight		-0.514	-0.109
BMI	0.162	-0.455	
bodyfat	0.320	-0.266	
waist	0.202	-0.433	
situp	-0.369		
grip_D	-0.246	-0.307	0.325
grip_ND	-0.242	-0.300	0.376
jump	-0.398		
run_20	-0.367		
flexion	-0.220		0.532
run_10	0.364		0.129

[여자 주성분 loading]

	Comp.1	Comp.2	Comp.3
age	0.243		0.515
height	-0.151	-0.214	-0.540
weight	0.185	-0.497	-0.145
BMI	0.273	-0.407	0.126
bodyfat	0.349	-0.196	
waist	0.280	-0.381	
situp	-0.351	-0.114	
grip_D	-0.211	-0.395	0.183
grip_ND	-0.214	-0.377	0.216
jump	-0.360	-0.147	
run_20	-0.338		
flexion	-0.169		0.565
run_10	0.351	0.115	

[주성분 3비교]

여성 데이터의 경우에는 남성 데이터에는 포함되지않은 **BMI**가 포함

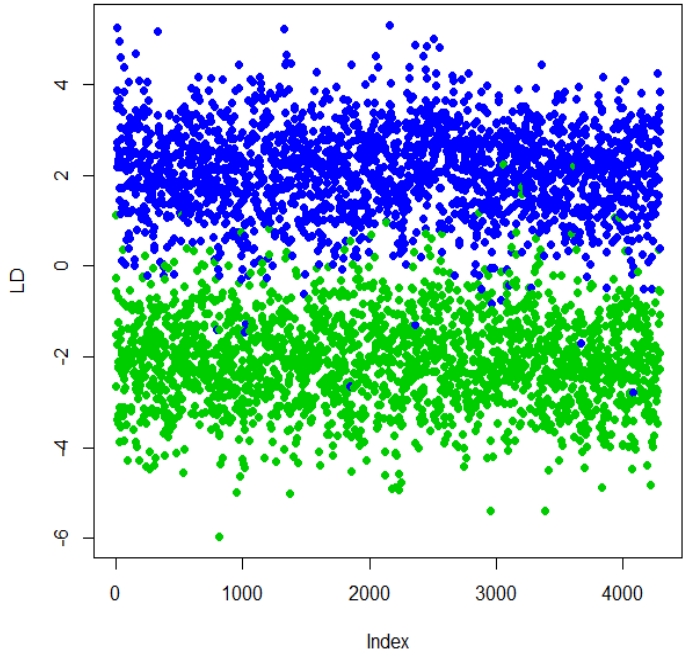
➔ 여성의 경우 피하지방률이 높고, 남성은 근육률이 높으므로, 다음과 같은 결과가 나온 것으로 추정

Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[판별 분석]

```
## Coefficients of linear discriminants:
##                 LD1
## age      -0.035071897
## height   -0.099673202
## weight    0.064693630
## BMI      -0.259652619
## bodyfat   0.088518620
## waist    -0.030780775
## situp     0.001145135
## grip_D   -0.052183348
## grip_ND  -0.021155194
## jump     -0.012674682
## run_20   -0.004900292
## flexion   0.052475517
## run_10    0.029147624
```



두 성별이 LD1하나로 거의 정확하게 판별되는 것을 확인할 수 있었다. 특히 BMI가 많은 영향을 끼치는 것을 알 수 있다

Ⅲ. 분석 – (1) 성별 별 차이

가설 1. 성별별로 차이가 있을 것이다

[K-평균 군집분석]

```
data2.kmeans <- kmeans(data2,center=2) #kmeans 방법으로 2 개의 그룹으로 분류

sum(data2.kmeans$cluster!=data[,3])/nrow(data2)

## [1] 0.1102307
```

[군집분석 결과]

- 오분류율이 0.11로 약 90%의 자료가 실제와 같게 군집화 되었다.
- 군집화 된 그룹별 변수들의 중심과, 실제 성별 별 변수들의 중앙값을 비교해본 결과, 값이 거의 비슷하다.

K-평균법 결과

	Height	Weight	BMI	Bodyfat	Waist	Situp	Grip_d	Grip_nd	Jump	Run_20	Flexion	Run_10
남자	172.1	72.6	24.5	21.6	84.1	40.9	42.9	40.5	204.0	39.9	10.9	11.4
여자	160.5	59.8	23.2	30.3	79.4	22.6	26.4	24.6	137.8	18.5	13.3	14.2

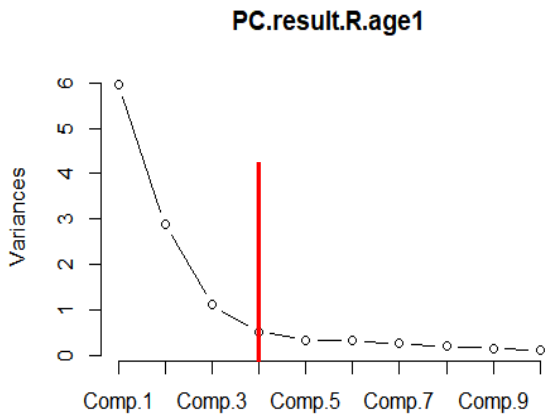
실제 성별별 평균

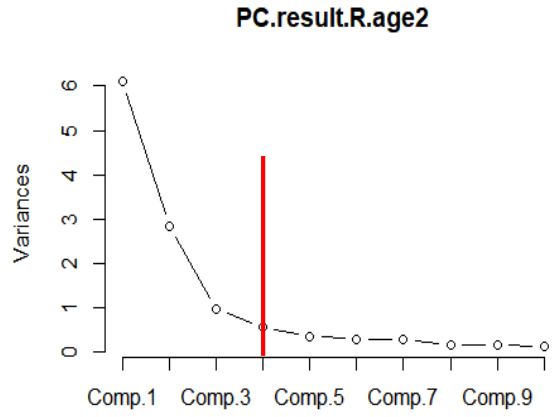
	Height	Weight	BMI	Bodyfat	Waist	Situp	Grip_d	Grip_nd	Jump	Run_20	Flexion	Run_10
남자	172.0	72.5	24.5	21.9	84.3	38	42.6	40.2	198	34	10.0	11.5
여자	159.5	57.0	22.3	30.9	77.4	23	25.1	23.5	140	17	16.1	13.9

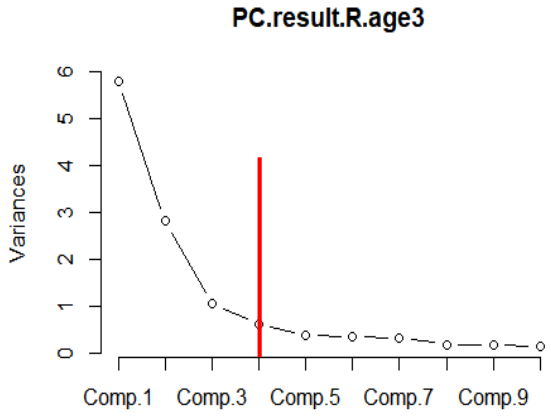
Ⅲ. 분석 – (2) 연령별 차이

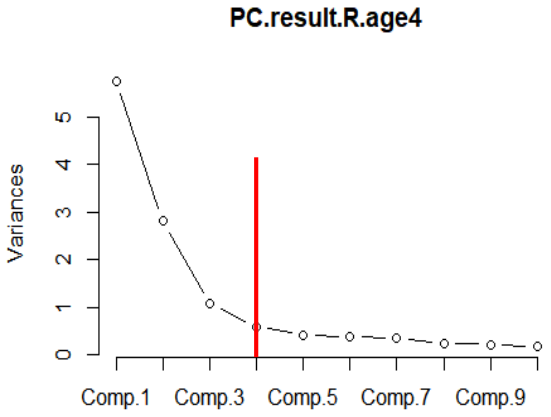
가설 2. 연령별로 차이가 있을 것이다

[연령집단별 주성분 분석 – 주성분 개수 선택]









- Screeplot을 그려본 결과 네 그룹 모두 elbow 위치 동일 →전부 Comp.1, Comp.2, Comp.3의 세 변수 사용
- 세 변수는 전체 자료의 80%를 설명

Ⅲ. 분석 – (2) 연령별 차이

가설 2. 연령별로 차이가 있을 것이다

[연령집단별 주성분 분석]

	Comp.1	Comp.2	Comp.3
height	0.317		0.368
weight	0.274	0.418	
BMI	0.149	0.495	-0.282
bodyfat	-0.247	0.374	-0.256
waist	0.160	0.497	-0.105
situp	0.314	-0.183	-0.213
grip_D	0.376		
grip_ND	0.375		
jump	0.368	-0.149	
run_20	0.302	-0.227	-0.102
flexion		-0.215	-0.793
run_10	-0.328	0.173	0.143

group1: 19세-29세

	Comp.1	Comp.2	Comp.3
height	0.325		0.306
weight	0.297	0.380	
BMI	0.181	0.473	-0.301
bodyfat	-0.226	0.405	-0.244
waist	0.195	0.471	-0.118
situp	0.292	-0.238	-0.197
grip_D	0.377		
grip_ND	0.378		
jump	0.358	-0.141	
run_20	0.286	-0.242	-0.118
flexion		-0.266	-0.826
run_10	-0.320	0.182	

group2: 30세-39세

	Comp.1	Comp.2	Comp.3
height	0.331		0.348
weight	0.300	0.388	
BMI	0.161	0.484	-0.317
bodyfat	-0.242	0.381	-0.275
waist	0.176	0.489	-0.123
situp	0.283	-0.245	-0.246
grip_D	0.382		
grip_ND	0.382		
jump	0.363	-0.138	
run_20	0.284	-0.243	-0.125
flexion		-0.236	-0.760
run_10	-0.314	0.180	0.147

group3: 40세-49세

	Comp.1	Comp.2	Comp.3
height	0.343		0.279
weight	0.270	0.432	
BMI		0.522	-0.297
bodyfat	-0.282	0.329	-0.235
waist	0.117	0.514	-0.105
situp	0.310	-0.163	-0.252
grip_D	0.384		
grip_ND	0.381		
jump	0.364	-0.122	
run_20	0.285	-0.209	-0.229
flexion		-0.193	-0.787
run_10	-0.313	0.183	0.157

group4: 50세-64세

[주성분 변수 해석]

네 그룹에서 각 주성분 변수들의 효과에 별다른 차이점이 보이지 않지만,
Group4: 50세-64세 데이터에서 제 1주성분에서 BMI가 빠지는 차이가 보인다.

Ⅲ. 분석 – (2) 연령별 차이

가설 2. 연령별로 차이가 있을 것이다

[판별 분석]

```
datas.LDA<-lda(datas,data$age_group2)

datas.LDA

## Call:
## lda(datas, grouping = data$age_group2)
##
## Prior probabilities of groups:
##         1         2         3         4
## 0.2320596 0.2273998 0.2315937 0.3089469
##
## Group means:
##        height      weight         BMI     bodyfat       waist       situp
## 1  0.26737030 -0.02894614 -0.243495463 -0.28778095 -0.27247326
0.50031277
## 2  0.20817077  0.12341257 -0.001088583 -0.03927823  0.04803233
0.16655065
## 3 -0.02767569  0.06993996  0.112307808  0.05771549  0.08510350 -
0.03852309
## 4 -0.33330759 -0.12152386  0.099509786  0.20180708  0.10551352 -
0.46951207
##        grip_D     grip_ND         jump      run_20     flexion      run_10
## 1  0.05678203  0.01670172  0.394060038  0.47502798  0.05762076 -
0.46660790
## 2  0.13962544  0.13842495  0.194900729  0.14931592 -0.01056287 -
0.21713818
## 3  0.06024850  0.06513821 -0.002691114 -0.09750218 -0.08933787
0.02362835
## 4 -0.19058547 -0.16326172 -0.437429821 -0.39362220  0.03146375
0.49259559
```

[LDA 계수]

```
##
## Coefficients of linear discriminants:
##                LD1        LD2       LD3
## height  -0.20907236  1.83875432  0.5654367
## weight  -1.10313772 -3.22954444  0.3580253
## BMI      0.73344051  2.79492034 -1.1961927
## bodyfat -0.65812582 -0.99420956  0.4189274
## waist    0.54818947 -0.04130853  0.7297345
## situp   -0.50770991  0.08022343 -0.5095399
## grip_D   0.21552847 -0.18874488 -0.2651819
## grip_ND  0.75751607 -0.89240145  0.3715867
## jump    -0.45022271 -0.38017888 -0.5168920
## run_20  -0.31419345  0.54007300  0.5269873
## flexion  0.07989771  0.15703062  0.6274437
```

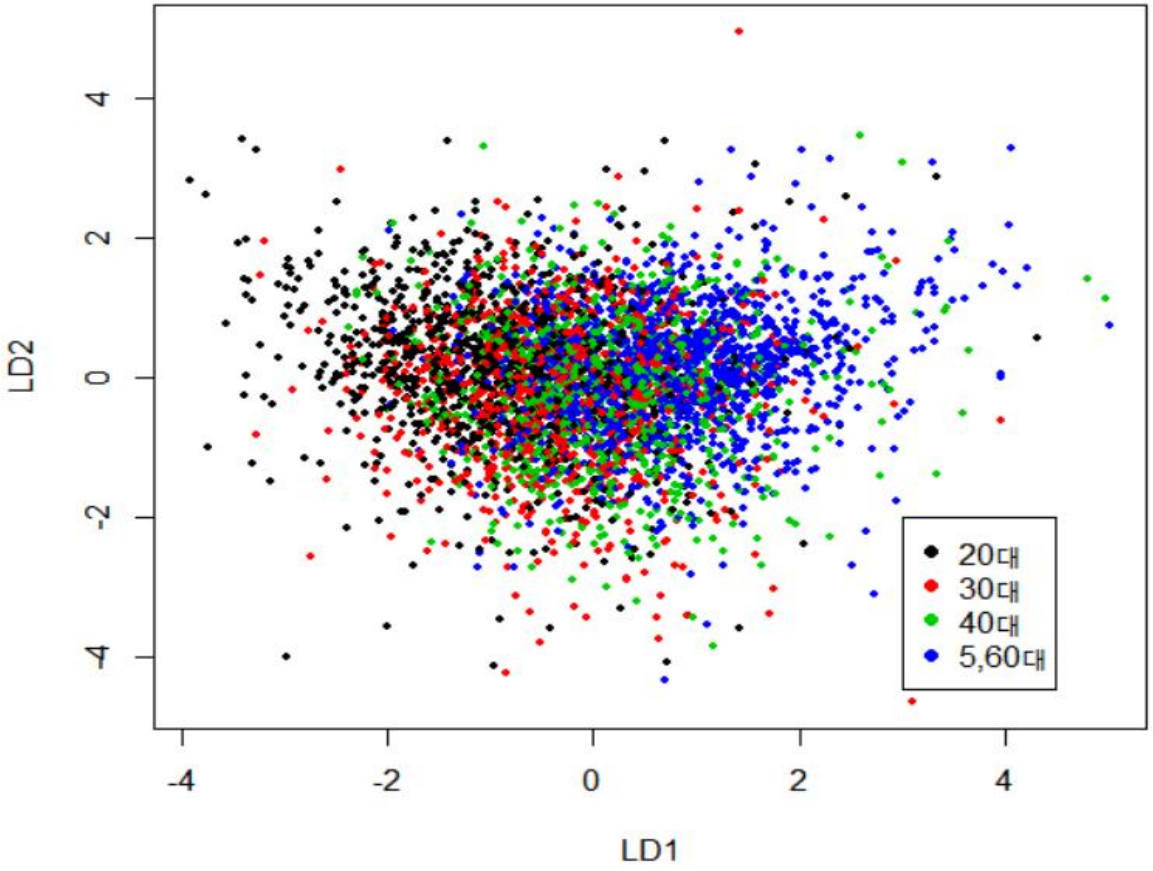
```
## Proportion of trace:
##    LD1    LD2    LD3
## 0.9302 0.0561 0.0138
```

- LD1이 자료의 93%, LD2는 5%를 설명
- 대부분의 데이터가 LD1에 의해 설명되고 있음

Ⅲ. 분석 – (2) 연령별 차이

가설 2. 연령별로 차이가 있을 것이다

[판별 분석]



[분석결과 해석]

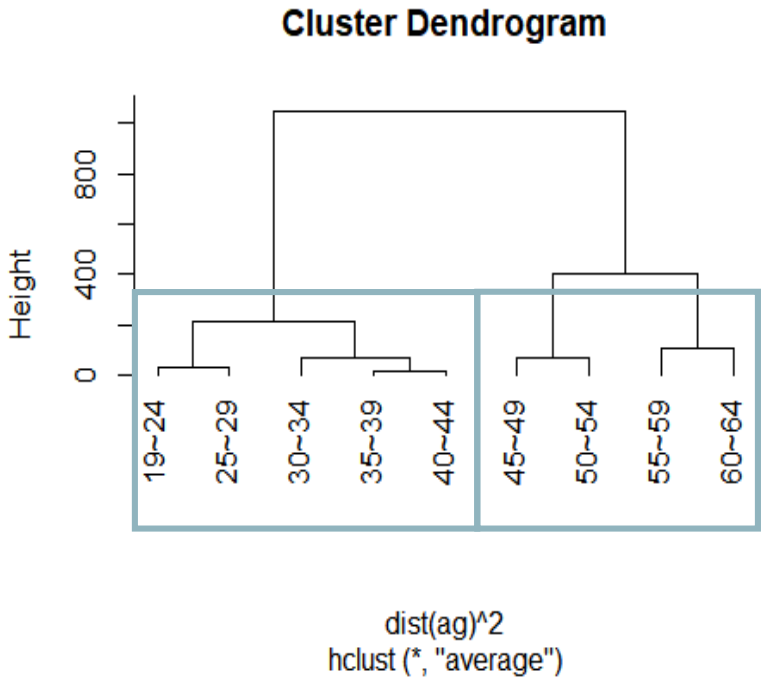
- LD1과 LD2를 축으로 그래프를 그려본 결과, 20대, 5-60대의 경우 서로의 반대편에 위치하여 다른 그룹들에 비해 분포의 확연한 차이를 확인 가능

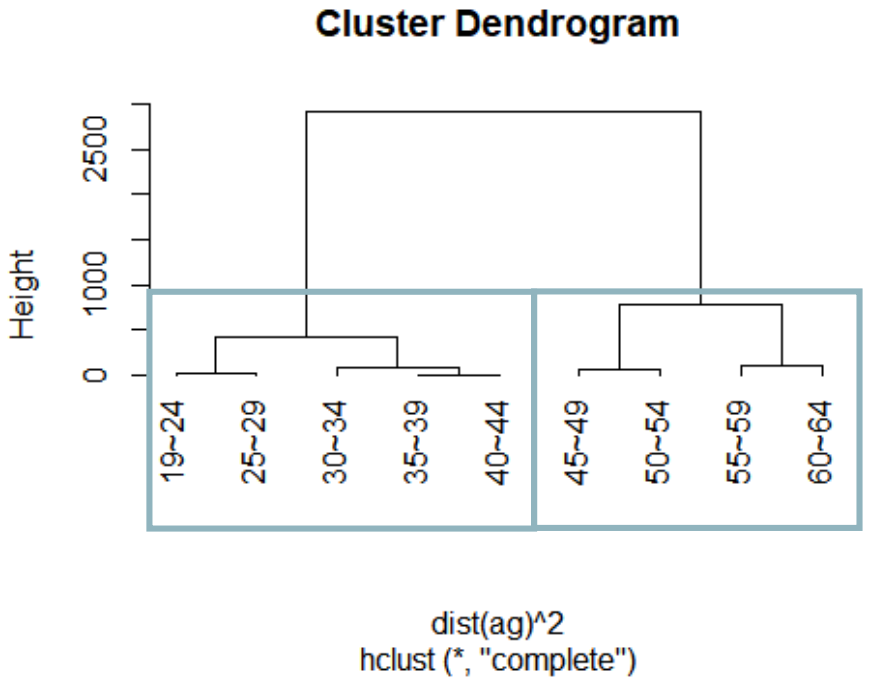
- 30대, 40대의 경우 전체 평면에 고루 분포하여 판별분석을 통한 해석이 불분명

Ⅲ. 분석 – (2) 연령별 차이

가설 2. 연령별로 차이가 있을 것이다

[계층적 군집분석]





- 크게 19세~44세와 45세~64세 두개의 그룹으로 나뉨
- 국민 체력이 40대 중반을 기준으로 많은 변화가 생긴다고 해석 가능
- 4개의 그룹으로 나누면 10~20대, 30~40대 중반, 40대 중반~50대 중반, 50대 중반~60대 중반 그룹으로 구분 가능

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[MANOVA - 지역에 따른 비만도 차이]

```
> cor(data1)
                age      height      weight         BMI     bodyfat       waist
age      1.000000000 -0.2490987 -0.05193624  0.124777049  0.18430364  0.13601464
height  -0.249098652  1.0000000  0.68743489  0.167593211 -0.51559247  0.34633898
weight  -0.051936243  0.6874349  1.00000000  0.826122618 -0.01591478  0.79598187
BMI      0.124777049  0.1675932  0.82612262  1.000000000  0.37551928  0.81681427
bodyfat  0.184303644 -0.5155925 -0.01591478  0.375519280  1.00000000  0.30734180
waist    0.136014643  0.3463390  0.79598187  0.816814269  0.30734180  1.00000000
situp   -0.379799010  0.4730452  0.27076874  0.002499594 -0.57675816 -0.01435858
grip_D  -0.113561853  0.7304863  0.66552708  0.342857318 -0.53404333  0.36653061
grip_ND -0.088824237  0.7196481  0.65358802  0.334643864 -0.54135037  0.36111962
jump    -0.334519259  0.6520286  0.42935731  0.084359499 -0.65901265  0.12257287
run_20  -0.344043302  0.4596227  0.22175454 -0.047906789 -0.58699022 -0.04314200
flexion -0.009290949 -0.2608898 -0.26552435 -0.161794571 -0.03096482 -0.26329521
run_10   0.384752343 -0.5172846 -0.30541861 -0.016609910  0.56887455 -0.02223770
```

- 서로 상관관계가 큰 변수들인 체중, BMI, 허리둘레를 이용하여 지역에 따라 비만도 차이가 있는지 MANOVA 검정

[분산에 대한 동일성 검정]

```
leveneTest(weight*waist*BMI~location,data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group   16  1.3076 0.1822
##       4275
```

p-value > 0.01
분산에 대한 동일성 만족
→ MANOVA 실시

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[Pillai's trace를 이용한 MANOVA]

```
summary(manova(cbind(weight,waist,BMI)~location,data=data))

##             Df   Pillai approx F num Df den Df    Pr(>F)
## location    16 0.055587   5.0442     48  12825 < 2.2e-16 ***
## Residuals 4275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Wilk's lambda를 이용한 MANOVA]

```
summary(manova(cbind(weight,waist,BMI)~location,data=data),test=c("Wilks"))

##             Df   Wilks approx F num Df den Df    Pr(>F)
## location    16 0.94502   5.0824     48  12710 < 2.2e-16 ***
## Residuals 4275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MANOVA 결과 P-value 값 유의

→ 지역에 따라 비만도의 차이가 있다!

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[주성분을 이용한 ANOVA]

```
PC.result.R$loadings

##
## Loadings:
##          Comp.1 Comp.2 Comp.3
## age       0.132  0.190  0.190
## height   -0.332         0.291
## weight   -0.268  0.412
## BMI      -0.109  0.499 -0.305
## bodyfat   0.261  0.322 -0.291
## waist    -0.127  0.501 -0.121
## situp    -0.314 -0.192 -0.219
## grip_D   -0.370
## grip_ND  -0.366
## jump     -0.368 -0.116
## run_20   -0.302 -0.209 -0.148
## flexion         -0.211 -0.766
## run_10    0.327  0.169  0.123
```

두 번째 주성분 = 0.412*체중 + 0.499*BMI + 0.322*체지방률 + 0.501*허리둘레
= '비만도'를 의미하는 새로운 변수 생성

```
data2<-mutate(data,obesity=0.412*weight+0.499*BMI+0.322*bodyfat+0.501*waist)
head(data2[,c(1,17)])

##   location  obesity
## 1        1 79.91572
## 2        1 69.74556
## 3        1 67.72542
## 4        1 75.51644
## 5        1 73.17960
## 6        1 89.72286
```

비만도를 의미하는 새로운 변수를 이용하여
지역에 따라 비만도의 차이가 있는지 분산분석

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[분산에 대한 동일성 검정]

```
leveneTest(obesity~location,data=data2)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group   16  1.7134 0.03754 *
##       4275
```

p-value > 0.01
분산에 대한 동일성 만족
→ ANOVA 실시

[ANOVA]

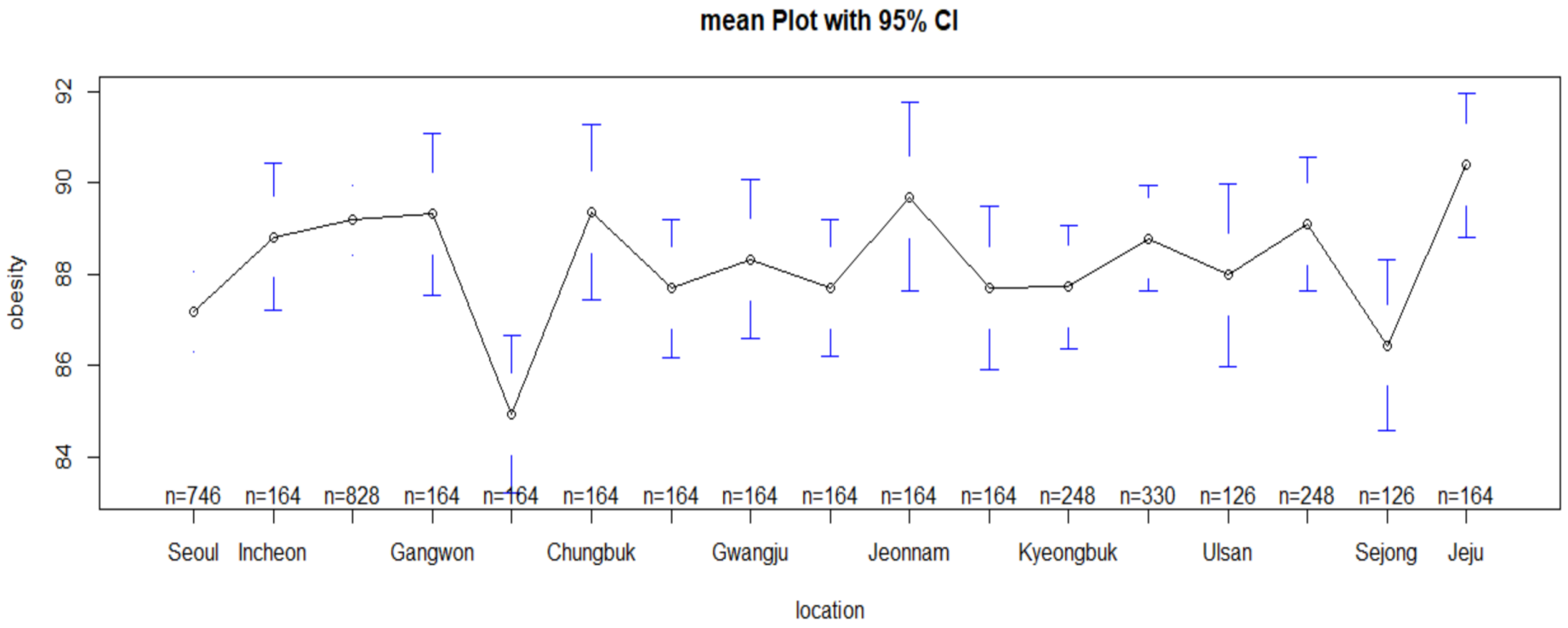
```
obesity.anova<-aov(obesity~location,data=data2)
summary(obesity.anova)

##               Df Sum Sq Mean Sq F value    Pr(>F)
## location      16   5810   363.1   2.845 0.000123 ***
## Residuals   4275 545666   127.6
## ---
```

ANOVA 결과 P-value 값 유의
지역별 비만도의 차이 존재!

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

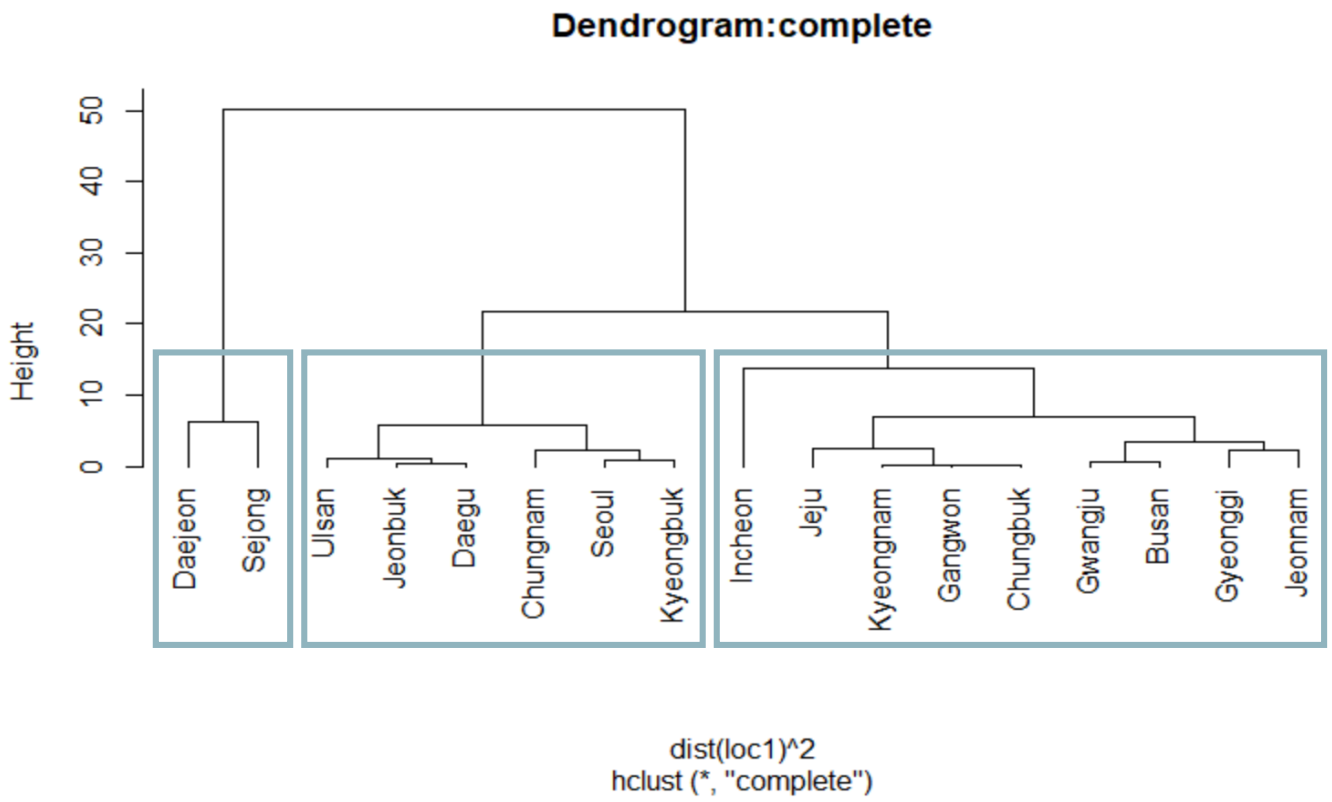
[그래프로 확인]



Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[Clustering - 17개 지역을 비만도에 따라 군집화]



- 체중, 허리둘레, 체지방률, BMI 변수의 평균값 이용
- 계층적 군집분석 실시

- 3개의 군집으로 나눈 결과

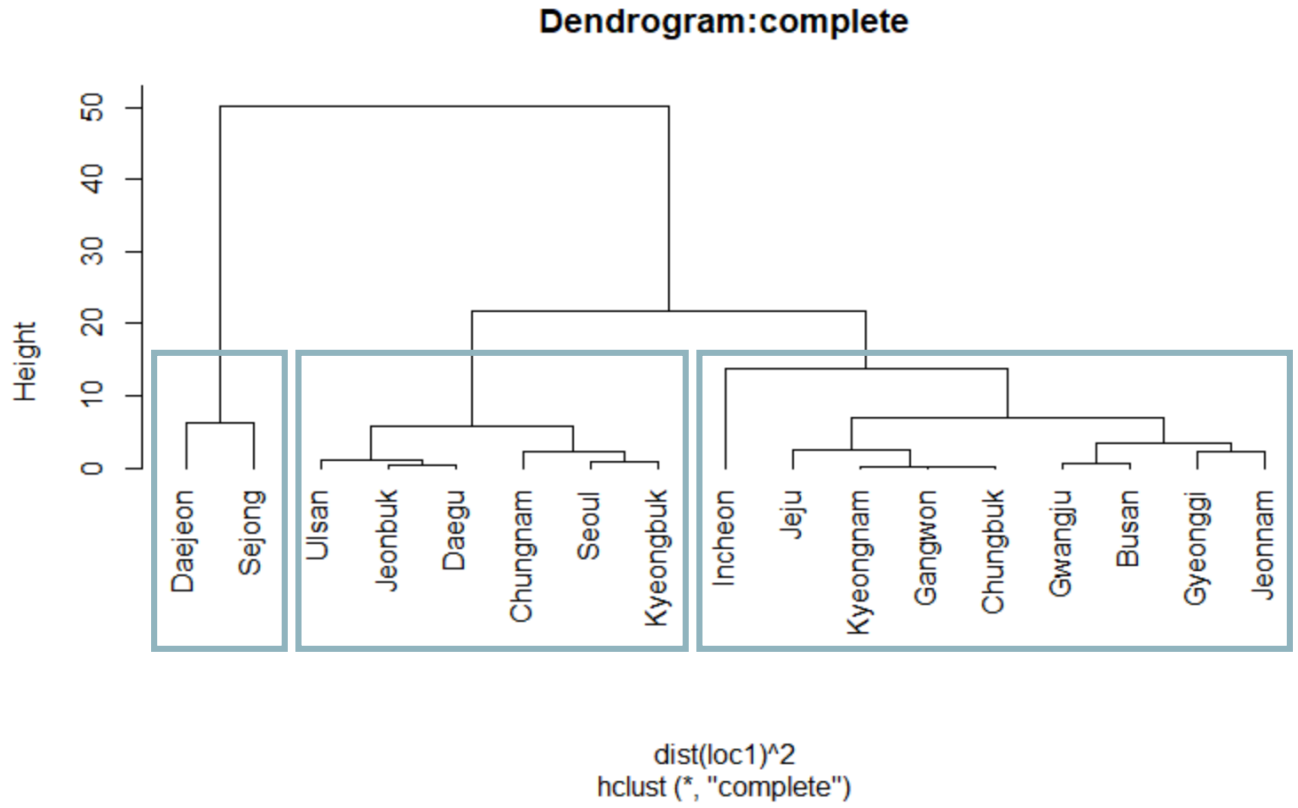
group1: 대전, 세종
group2: 울산, 전북, 대구, 충남, 서울, 경북
group3: 인천, 제주, 경남, 강원, 충북, 광주, 부산, 경기, 전남

Ⅲ. 분석 – (3) 지역 별 차이

가설 3. 지역별로 차이가 있을 것이다

[Clustering - 17개 지역을 비만도에 따라 군집화]

- 주성분 비만도 값이 큰 지역 순서대로 정렬한 데이터와 비교




```
> arrange(data4,desc(mean_obesity))
# A tibble: 17 x 2
   location  mean_obesity
   <fct>            <dbl>
 1 Jeju              90.4
 2 Jeonnam           89.7
 3 Chungbuk          89.3
 4 Gangwon           89.3
 5 Gyeonggi          89.2
 6 Kyeongnam         89.1
 7 Incheon           88.8
 8 Busan             88.8
 9 Gwangju           88.3
10 Ulsan             88.0
11 Kyeongbuk         87.7
12 Daegu             87.7
13 Jeonbuk           87.7
14 Chungnam          87.7
15 Seoul             87.2
16 Sejong            86.4
17 Daejeon           84.9
```

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]

```
cc1$cor

## [1] 0.71537499 0.11959120 0.07562573

cc1$xcoef

##                [,1]       [,2]       [,3]
## height   -0.2543960 -2.8340234 -3.7510263
## weight   -0.1780563  6.2087722  5.3696763
## BMI      -0.2965713 -3.9785689 -2.7729074
## bodyfat   0.8086538  0.5629280 -0.6045865
## waist     0.2388564 -0.2286263 -0.8407856

cc1$ycoef

##               [,1]       [,2]       [,3]
## situp   -0.3577927  0.2831511  1.4884026
## run_20  -0.3510075 -1.3655232 -0.4664797
## run_10   0.4127510 -1.0460136  1.0202314
```

X: 체격 변수 집단
(키, 몸무게, BMI, 체지방률, 허리둘레)
Y: 체력 변수 집단
(윗몸일으키기, 20m 왕복 오래 달리기, 10m 왕복 달리기)

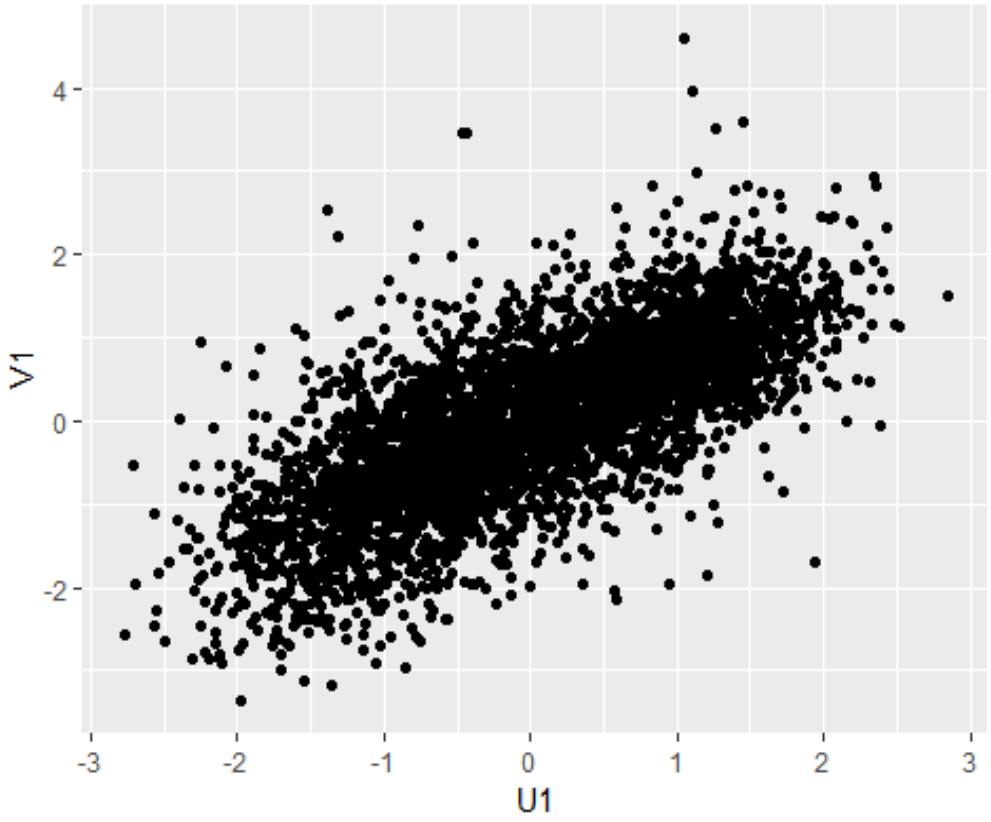
[분석결과 해석]

- 첫번째 정준상관계수 : 0.715
- U1 = 키, 몸무게, BMI 대비 체지방률과 허리둘레
- V1 = 전체적인 체력의 좋지 않은 정도
- U1, V1이 양의 상관관계를 가지므로 체지방이 많을수록 체력이 떨어지는 관계가 존재

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]



X: 체격 변수 집단
(키, 몸무게, BMI, 체지방률, 허리둘레)
Y: 체력 변수 집단
(윗몸일으키기, 20m 왕복 오래 달리기, 10m 왕복 달리기)

[분석결과 해석]

- 첫번째 정준상관계수 : 0.715
- U1 = 키, 몸무게, BMI 대비 체지방률과 허리둘레
- V1 = 전체적인 체력의 좋지 않은 정도
- U1, V1이 양의 상관관계를 가지므로 체지방이 많을수록 체력이 떨어지는 관계가 존재

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]

```
X<-data.std[,c(2:6)]
Y<-data.std[,c(8:9)]

cc1<-cc(X,Y)
cc1$cor

## [1] 0.8569934 0.0643432

cc1$xcoef

##                [,1]       [,2]
## height   -0.21512580 -1.8238702
## weight   -0.39352450  0.9867893
## BMI      -0.24725492 -1.1731739
## bodyfat   0.62126111 -1.2851064
## waist    -0.03195211  1.0416809

cc1$ycoef

##                [,1]      [,2]
## grip_D   -0.5566548 -3.002159
## grip_ND  -0.4572872  3.018893
```

X: 체격 변수 집단
(키, 몸무게, BMI, 체지방률, 허리둘레)
Y: 악력 변수 집단
(악력_자주 쓰는 손, 악력,_반대 손)

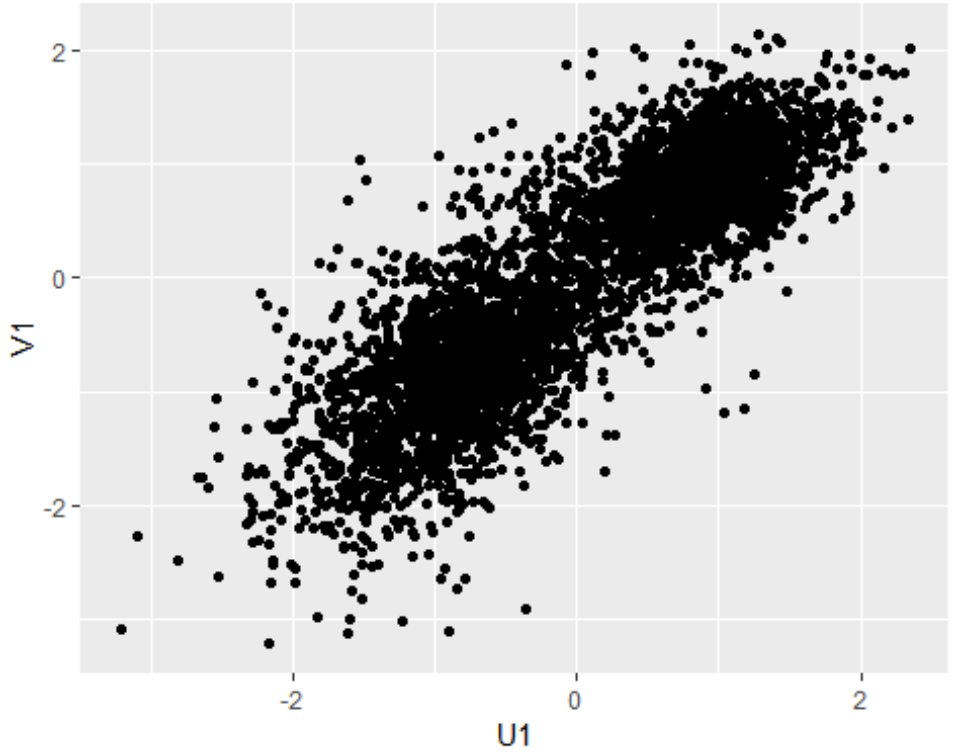
[분석결과 해석]

- 첫번째 정준상관계수 : 0.856
- U1 = 키, 몸무게, BMI (체격) 대비 체지방률
- V1 = 전체적인 악력의 약한 정도
- U1, V10이 양의 상관관계를 가지므로 체격 대비 체지방이 많을수록 악력이 약해지는 관계가 존재
- 허리둘레는 악력에 큰 영향을 미치지 않음

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]



X: 체격 변수 집단
(키, 몸무게, BMI, 체지방률, 허리둘레)
Y: 악력 변수 집단
(악력_자주 쓰는 손, 악력,_반대 손)

[분석결과 해석]

- 첫번째 정준상관계수 : 0.856
- U1 = 키, 몸무게, BMI (체격) 대비 체지방률
- V1 = 전체적인 악력의 약한 정도
- U1, V1이 양의 상관관계를 가지므로 체격 대비 체지방이 많을수록 악력이 약해지는 관계가 존재
- 허리둘레는 악력에 큰 영향을 미치지 않음

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]

```
cc1$cor

## [1] 0.66761212 0.00861644

cc1$xcoef

##                [,1]       [,2]
## situp    -0.3358925 -0.132270
## run_20   -0.2930265 -1.215063
## run_10    0.4890294 -1.250475

cc1$ycoef

##                [,1]       [,2]
## grip_D   -0.6091324  2.991953
## grip_ND  -0.4043612 -3.026436
```

X: 체력 변수 집단
(윗몸일으키기, 20m 왕복 오래 달리기, 10m 왕복 달리기)
Y: 악력 변수 집단
(악력_자주 쓰는 손, 악력,_반대 손)

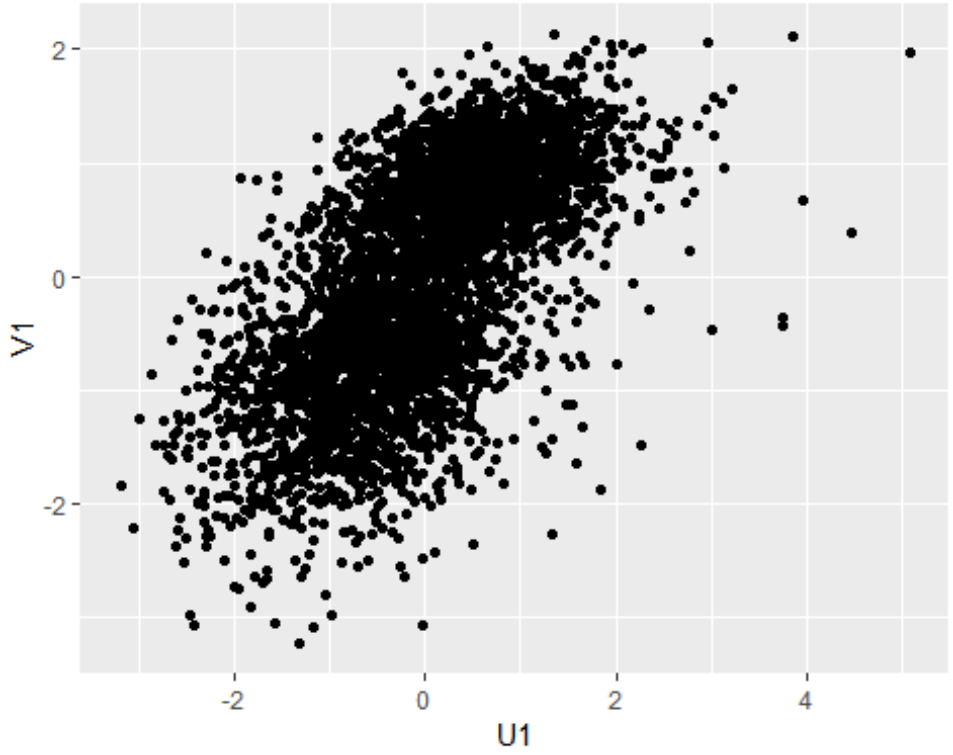
[분석결과 해석]

- 첫번째 정준상관계수: 0.668
- U1 = 전체적인 체력의 좋지 않은 정도
- V1 = 전체적인 악력의 약한 정도
- 체력과 악력이 0.668 정도로 비례하는 관계
- 확실히 앞선 그림들보다 약한 선형 관계를 보임

Ⅲ. 분석 – (4) 체력과 체격 데이터의 연관성

가설 4. 체력과 체격 데이터의 연관성이 높을 것이다

[정준상관분석]

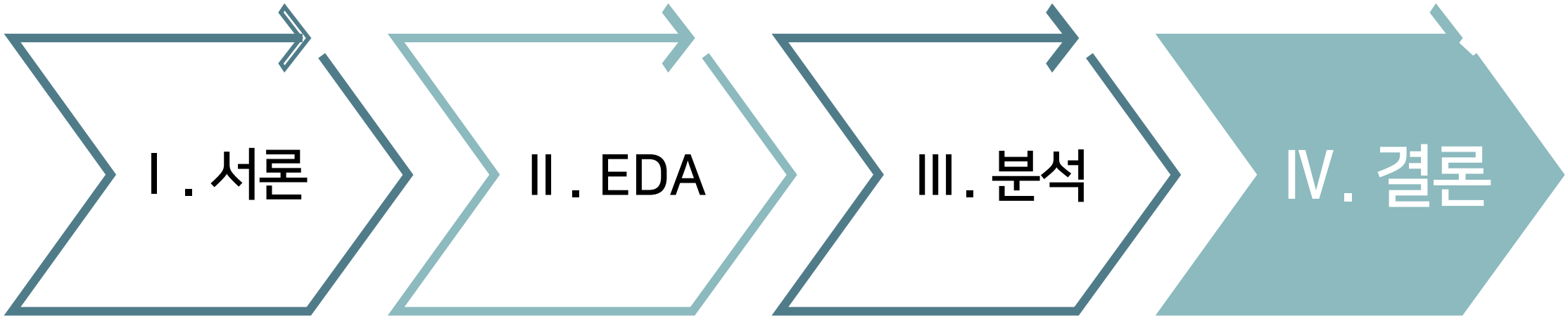


X: 체력 변수 집단
(윗몸일으키기, 20m 왕복 오래 달리기, 10m 왕복 달리기)
Y: 악력 변수 집단
(악력_자주 쓰는 손, 악력,_반대 손)

[분석결과 해석]

- 첫번째 정준상관계수: 0.668
- U1 = 전체적인 체력의 좋지 않은 정도
- V1 = 전체적인 악력의 약한 정도
- 체력과 악력이 0.668 정도로 비례하는 관계
- 확실히 앞선 그림들보다 약한 선형 관계를 보임

Ⅳ. 결론



1.분석 결과

2.한계점

Ⅳ. 결론 – (1) 분석 결과 및 한계점

1. 성별별 차이

- 판별분석, 군집화 결과, 확실히 성별에 따라 데이터가 2개로 나뉘는 것을 확인
- 주성분분석에서도 성별에 따른 신체적 차이에 의해 주성분에 영향을 주는 변수 차이 존재

2. 연령대별 차이

- 판별분석 결과, 연령대의 차이가 클수록 확연히 데이터가 분류되는 것을 확인
- 군집화 결과, 40대 중반을 기점으로 완벽히 나눠짐

3. 지역별 차이

- MANOVA를 통해 지역별로 몸무게, BMI, 허리둘레 차이가 존재하는 것을 확인
- 관련 주성분을 응용한 결과 대전과 세종 지역이 가장 비만도가 낮은 것으로 나타남

4. 체격, 체력, 악력 변수에 대한 연관성 존재

- 셋 중 체격과 악력의 관련성이 가장 높았으며, 세 변수 집단 모두 서로 연관되어 있음을 확인

Ⅳ. 결론 – (1) 분석 결과 및 한계점

- 대부분 예상한 결과를 확인
- 지역별 차이의 경우 표본 수가 고르지 않은 점, 관련 정보가 부족한 점으로 인해 결과에 대한 이해를 완벽히 하기 어려웠음
- 지역별 분석을 위해서는 지역적 특성에 대한 이해와 함께 충분하고 고른 표본 필요

감사합니다.