다변량 데이터 분석 프로젝트

- 국민체력실태조사(2017)

2019. 6. 18

7 조

1602045 박정현

1602050 박지원

1602069 예지혜

1602073 유은지

목차

1. 서론

(1) 개요

본 팀 프로젝트는 '다변량분석및실습' 수업에서 배운 다양한 분석 방법을 통해 자료를 적절히 분석하고, 의미 있는 해석 결과를 도출하는 것을 목표로 하였다. 목표를 달성하기 위해 본 팀은 국민체력실태조사 데이터를 선정하였다. 정수형, 숫자형, 범주형 등 다양한 형태의 변수들과 충분한 자료 수가 갖춰져야 다양한 방향으로 분석이 가능할 것이라고 판단하여 해당 데이터를 선정하게 되었다.

진행 과정은 다음과 같다. 팀원 모두가 데이터에 대한 간단한 EDA 를 하여 데이터에 대한 전반적인 이해를 한 뒤, 흥미로워 보이는 분석 방향에 대한 의견을 내고 토의를 거쳐 유의미한 결과를 도출해낼 수 있을 것으로 생각되는 가설들을 목적별로 설정하였다. 가설 설정 후, 전처리 및 상세한 EDA, 분석 방법 별 가설 검증을 분담하여 분석을 실시하였다. 최종적으로 네 가지 가설에 대해 분석해 본 결과를 바탕으로 의미 있는 결론을 도출하고 이를 통해 본 프로젝트의 의의와 한계점에 대해 함께 평가하는 시간을 가졌다.

다음과 같이 역할을 분배하여 프로젝트를 진행하였다.

박정현 : EDA, 가설 1, 2 에 관한 주성분분석, 자료 정리

박지원 : 가설 3, 4 에 관한 정준상관분석, MANOVA, 주성분분석 응용, PPT

예지혜 : 가설 3, 4 에 관한 정준상관분석, 군집화, MANOVA, 결론

유은지 : 서론, 가설 1, 2 에 관한 판별분석

(2) 자료 설명

[데이터] 전국 17 개 시, 도 19~64 세 성인의 국민체력실태조사

- 데이터 출처 : https://mdis.kostat.go.kr/index.do
- 검색 경로 : 교육 문화 > 국민체력실태조사 > 성인(제공) > 2017
- 변수 16 개, 관측값 4291 개
-

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

[변수 참고사항]

- 연령: 19~64 세
- 연령집단: 19~24 세, 25~29 세, 30~34 세, 35~39 세, 40~44 세, 45~49 세, 50~54 세, 55~59 세, 60~64 세
- BMI: 18.5 미만 저체중, 18.5~22.9 정상, 23~24.9 과체중, 25~30 경도 비만, 30~35 중등도 비만, 35 이상 고도 비만
- 20m 왕복오래달리기: 시간내 도달 성공 횟수, 10m 왕복달리기: 2 회 왕복 걸린 시간

2. EDA

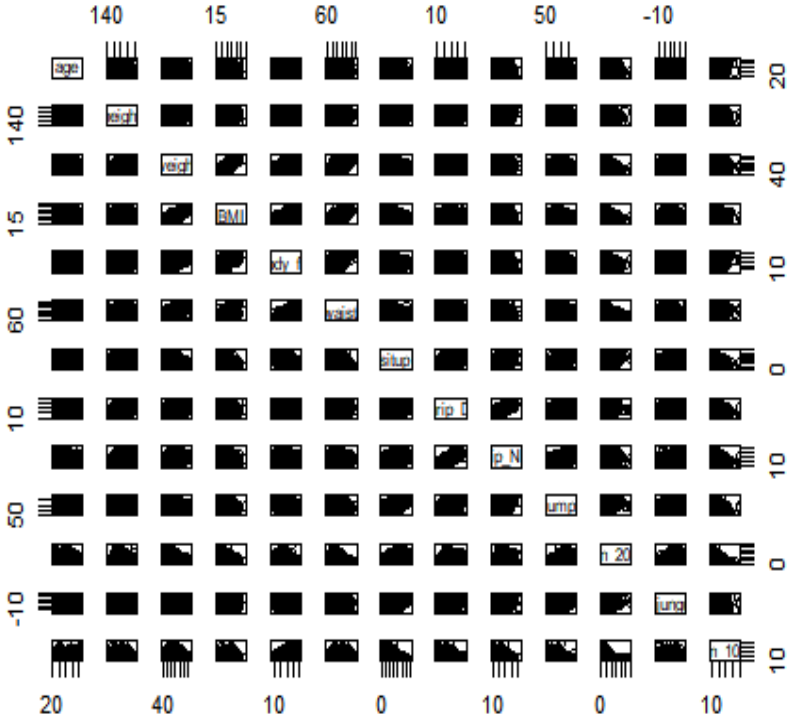
(1) 자료 살펴보기

1) 데이터 불러오기

```
data<-read.table("c:/Temp/data.txt",header=FALSE, sep=',')
colnames(data)<-
c("district","age","sex","height","weight","BMI","body_fat","waist","situp
","grip_D","grip_ND","jump","run_20m","jajungul","run_10m","age_group")
data<-data.frame(data)
data$sex<-as.factor(data$sex)
data$district<-as.factor(data$district)
data$age_group<-as.factor(data$age_group)
library(reshape2)
library(ggplot2)


summary(data)
```

```
##     district        age        sex         height          weight
##  3      : 828   Min.   :19.00   1:2146   Min.   :141.4   Min.   : 38.5
##  1      : 746   1st Qu.:30.00   2:2146   1st Qu.:159.2   1st Qu.: 56.4
##  13     : 330   Median :41.00            Median :165.8   Median : 64.5
##  12     : 248   Mean   :41.08            Mean   :165.9   Mean   : 65.8
##  15     : 248   3rd Qu.:52.00            3rd Qu.:172.2   3rd Qu.: 73.6
##  2      : 164   Max.   :64.00            Max.   :198.0   Max.   :136.7
##  (Other):1728
##       BMI           body_fat         waist            situp
##  Min.   :15.54   Min.   : 6.10   Min.   : 57.00   Min.   : 0.0
##  1st Qu.:21.50   1st Qu.:20.90   1st Qu.: 75.10   1st Qu.:21.0
##  Median :23.47   Median :25.80   Median : 81.00   Median :31.0
##  Mean   :23.78   Mean   :26.23   Mean   : 81.58   Mean   :31.1
##  3rd Qu.:25.60   3rd Qu.:31.82   3rd Qu.: 87.26   3rd Qu.:41.0
##  Max.   :44.23   Max.   :52.00   Max.   :137.90   Max.   :78.0
##
##      grip_D         grip_ND          jump          run_20m
##  Min.   :10.30   Min.   : 8.90   Min.   : 50.0   Min.   :  1.00
##  1st Qu.:25.10   1st Qu.:23.40   1st Qu.:139.0   1st Qu.: 15.00
##  Median :32.80   Median :30.50   Median :167.0   Median : 24.00
##  Mean   :34.08   Mean   :32.05   Mean   :168.7   Mean   : 28.53
##  3rd Qu.:42.70   3rd Qu.:40.20   3rd Qu.:199.0   3rd Qu.: 38.00
##  Max.   :69.90   Max.   :67.80   Max.   :295.0   Max.   :117.00
##
##     jajungul         run_10m        age_group
##  Min.   :-20.000   Min.   : 8.28   1      : 508
##  1st Qu.:  6.675   1st Qu.:11.30   7      : 508
##  Median : 13.200   Median :12.74   8      : 508
##  Mean   : 12.186   Mean   :12.89   6      : 506
##  3rd Qu.: 19.000   3rd Qu.:14.12   2      : 488
##  Max.   : 39.000   Max.   :30.00   3      : 488
##                                    (Other):1286
```

2) 변수간 관련성



- 나이가 많아지면 situp, jump, 20m 달리기에서 성적이 줄어드는 모습이 존재한다.
- weight, BMI, bodyfat, waist 사이의 양이 상관성이 존재한다.
- 위 변수들의 대부분의 운동 변수들과 음의상관성을 보이지만, grip 과 10m 달리기의 경우엔 그 정도가 덜하고 양의 상관성을 가지는 모습을 보인다.
- situp, grip, jump, 20m 와도 양의 상관성이 존재한다.
- 20m 달리기와 10m 달리기는 음의 상관성이 있다. 10m 달리기가 다른 운동변수들과 음의 상관성을 보이는 경향이 있다. (20m 달리기 횟수이기 때문이다.)

3) 남녀데이터 나눠 살펴보기

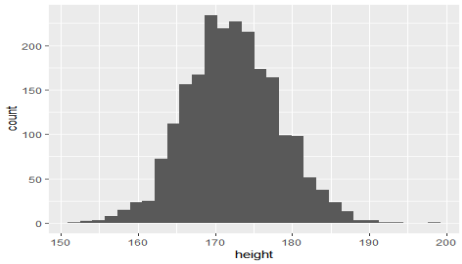
```
ggplot(data_men,aes(height))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
data_men<-data[data$sex==1,]
data_women<-data[data$sex==2,]
```

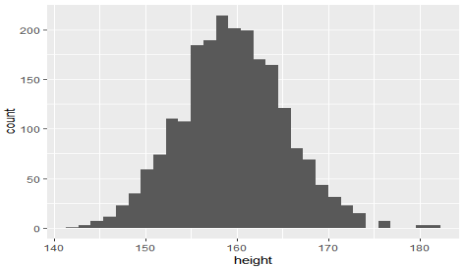
```
#1. 남녀 키
ggplot(data_men,aes(height))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data_women,aes(height))+geom_histogram()
```

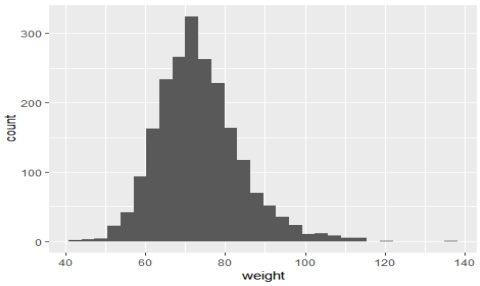
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



남녀 키 모두 정규분포를 따르는 모양새를 보인다.

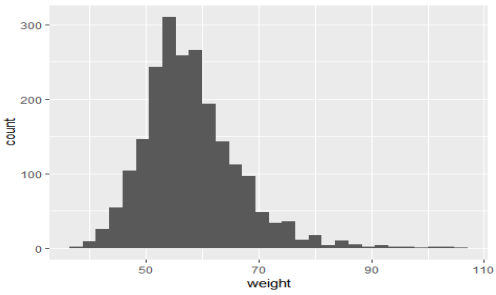
```
#2. 남녀 몸무게
ggplot(data_men,aes(weight))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



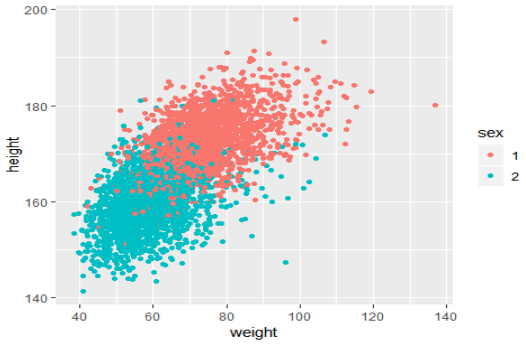

```
ggplot(data_women,aes(weight))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



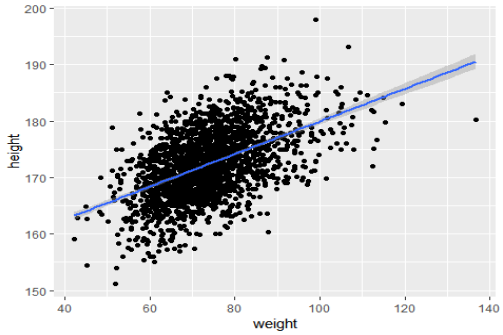
남녀 몸무게 모두 왼쪽으로 치우친 모양새를 보인다(비만 인원이 더 작다).

```r
#3. 키와 몸무게의 성별별 산점도
ggplot(data,aes(weight,height,colour=sex))+geom_point()
```

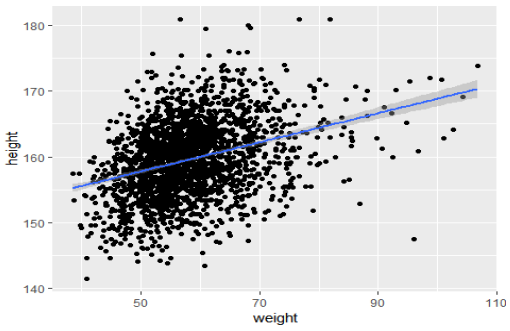


키와 몸무게별로 plot 을 그려본 결과, 남성과 여성의 차이가 매우 큰 것을 확인할 수 있다.

```r
ggplot(data_men,aes(weight,height))+geom_point()+geom_smooth(method =
"lm")
```

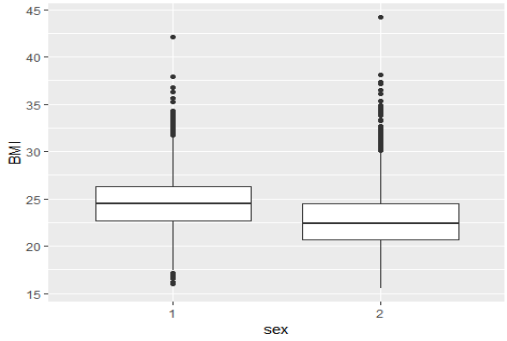



```r
ggplot(data_women,aes(weight,height))+geom_point()+geom_smooth(method =
"lm")
```



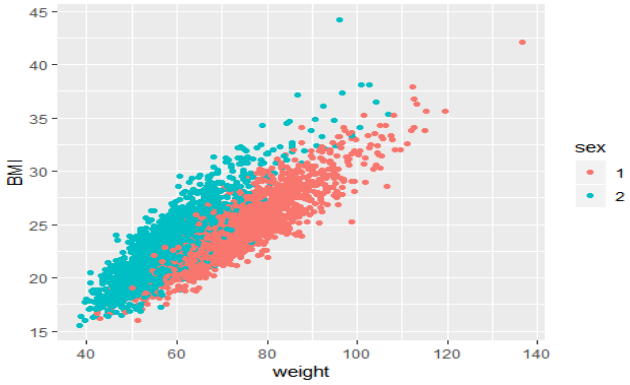
전반적으로 남성이 같은 몸무게에 대해 키의 분포가 모여 있는 편이다.

```r
#4.BMI
ggplot(data,aes(sex,BMI))+geom_boxplot()
```



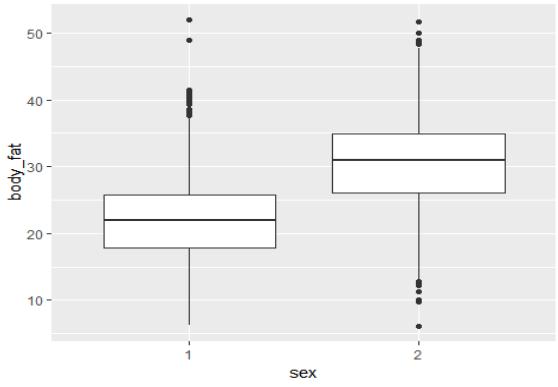
여성이 남성보다 BMI 가 낮은 편이다.

```
ggplot(data,aes(weight,BMI,colour=sex))+geom_point()
```



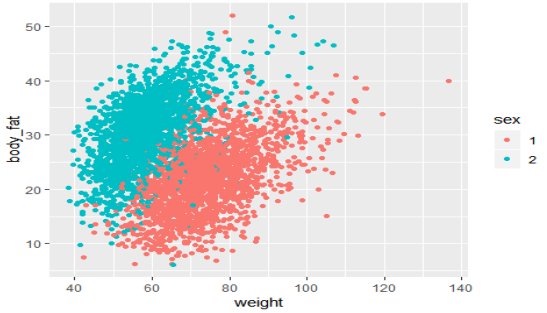
같은 몸무게일땐 상대적으로 키가 작은 여성이 BMI 가 더 높은 편이다.

```
#5. 체지방률
ggplot(data,aes(sex,body_fat))+geom_boxplot()
```

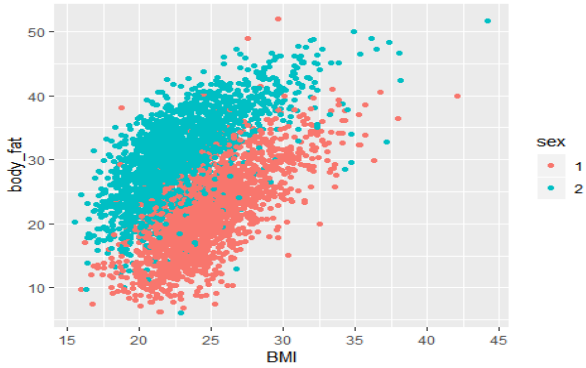


체지방률은 확연히 여성이 남성보다 높은 것을 확인할 수 있다. 신체적 차이인 것으로 보인다.

```
ggplot(data,aes(weight,body_fat,colour=sex))+geom_point()
```

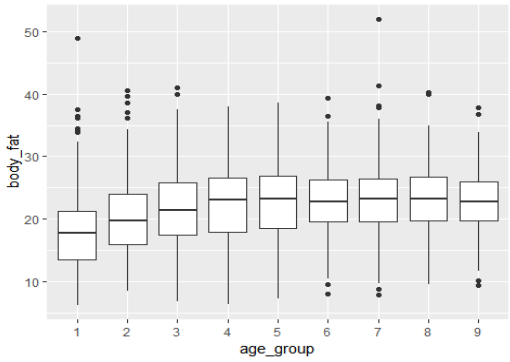



```
ggplot(data,aes(BMI,body_fat,colour=sex))+geom_point()
```

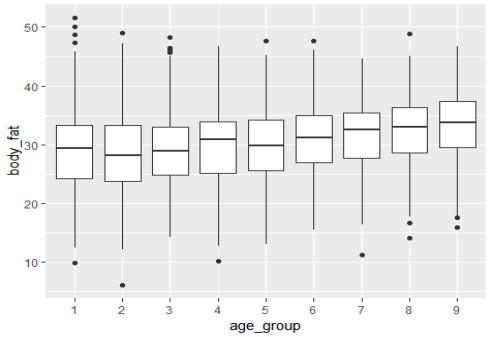


체지방률은 몸무게보다는 BMI 와 연관성이 더 높은 것으로 보인다.

```
ggplot(data_men,aes(age_group,body_fat))+geom_boxplot()
```

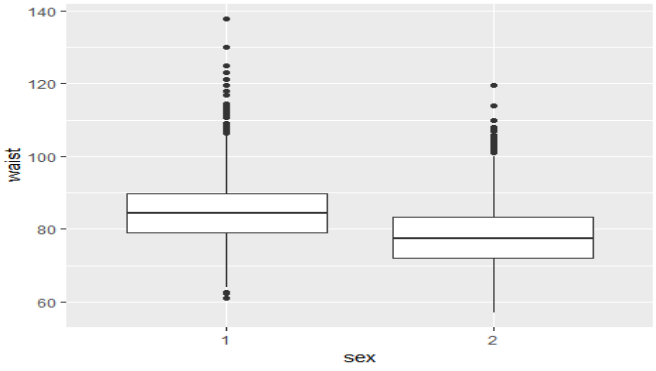



```
ggplot(data_women,aes(age_group,body_fat))+geom_boxplot()
```

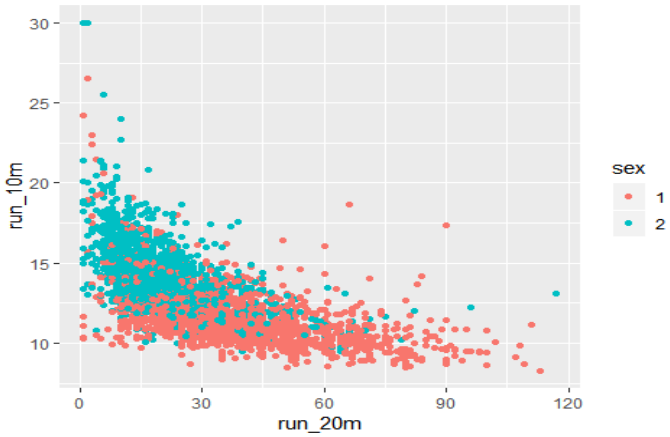


남성과 여성 모두 나이가 들면 평균 체지방률이 높아진다.

```
#6. 허리둘레
ggplot(data,aes(sex,waist))+geom_boxplot()
```

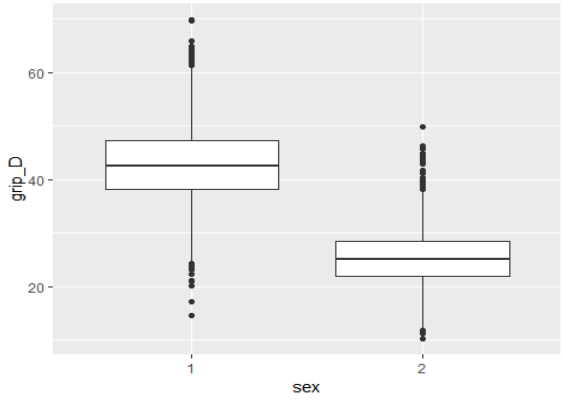



```
#7. 20m,10m
ggplot(data,aes(run_20m,run_10m,colour=sex))+geom_point()
```

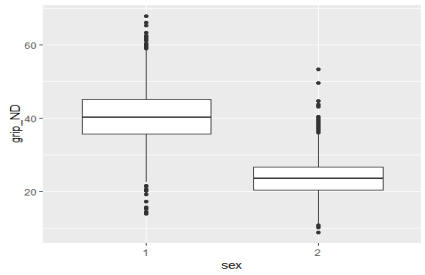


대체적으로 여성보다 남성이 훨씬 기록이 좋은 편이다. 특히 20m 왕복 오래 달리기 횟수가 남성이 훨씬 높다.

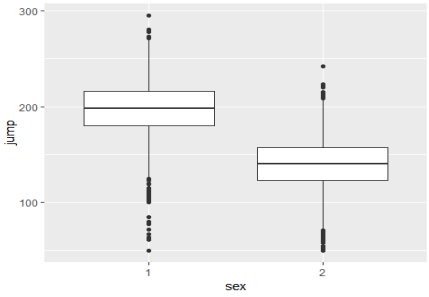
```
#8. 운동
ggplot(data,aes(sex,grip_D))+geom_boxplot()
```



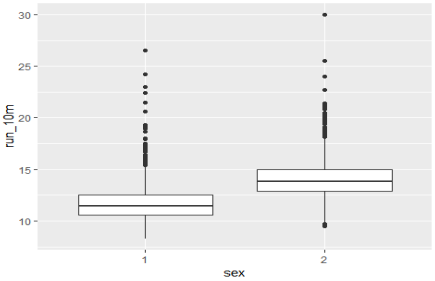

```
ggplot(data,aes(sex,grip_ND))+geom_boxplot()
```



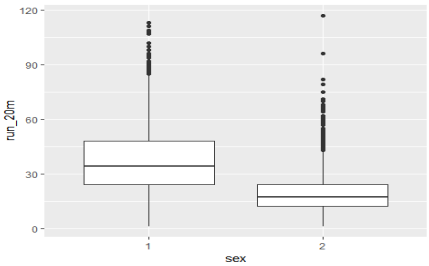

```
ggplot(data,aes(sex,jump))+geom_boxplot()
```



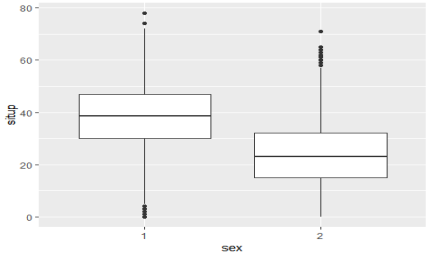

```
ggplot(data,aes(sex,run_10m))+geom_boxplot()
```




```
ggplot(data,aes(sex,run_20m))+geom_boxplot()
```

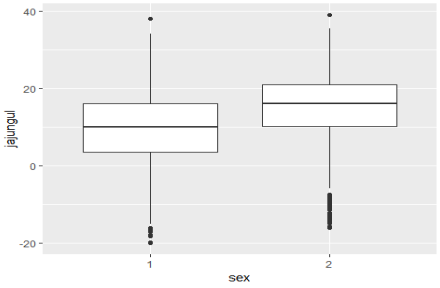



```
ggplot(data,aes(sex,situp))+geom_boxplot()
```



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

```
ggplot(data,aes(sex,jajungul))+geom_boxplot()
```



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

(2) 주성분분석

```
#9.전체 데이터 주성분 분석

data<-read.table("c:/Temp/data.txt",header=FALSE, sep=',')
colnames(data)<-
c("location","age","sex","height","weight","BMI","bodyfat","waist","situp"
,"grip_D","grip_ND","jump","run_20","flexion","run_10","age_group")
data<-data.frame(data)
data$sex<-as.factor(data$sex)
data$location<-as.factor(data$location)
data$age_group<-as.factor(data$age_group)

data1<- data[,c(-1,-3,-16)]
S<-var(data1)
S
```

```
##                age     height     weight        BMI     bodyfat
## age      161.450547 -27.372224  -8.108742  5.1495965   17.923243
## height   -27.372224  74.789027  73.048903  4.7075409  -34.126239
## weight    -8.108742  73.048903 150.982420 32.9705687   -1.496673
## BMI        5.149597   4.707541  32.970569 10.5496359    9.334991
## bodyfat   17.923243 -34.126239  -1.496673  9.3349907   58.576826
## waist     15.972780  27.681868  90.394384 24.5198119   21.740005
## situp    -70.299041  59.593290  48.466038  0.1182671  -64.303184
## grip_D   -15.525789  67.972384  87.989406 11.9821188  -43.978547
## grip_ND  -11.716972  64.610521  83.374148 11.2840691  -43.013570
## jump    -171.416820 227.404022 212.762262 11.0500762 -203.408556
## run_20   -76.955286  69.972278  47.966861 -2.7391851  -79.085957
## flexion   -1.110531 -21.223983 -30.691515 -4.9434876   -2.229371
## run_10    10.438087  -9.551429  -8.012694 -0.1151876    9.296068
```

```
##              waist       situp    grip_D   grip_ND        jump     run_20
## age      15.9727800 -70.2990408 -15.52579 -11.71697 -171.416820 -
76.955286
## height   27.6818682  59.5932900  67.97238  64.61052  227.404022
69.972278
## weight   90.3943842  48.4660383  87.98941  83.37415  212.762262
47.966861
## BMI      24.5198119   0.1182671  11.98212  11.28407   11.050076  -
2.739185
## bodyfat  21.7400045 -64.3031843 -43.97855 -43.01357 -203.408556 -
79.085957
## waist    85.4181512  -1.9331355  36.44910  34.64897   45.685879  -
7.019092
## situp    -1.9331355 212.2030508  91.79718  87.60601  432.724746
177.629769
## grip_D   36.4490985  91.7971792 115.77184 105.54241  326.630925
107.048343
## grip_ND  34.6489681  87.6060145 105.54241 107.77754  313.435280
102.353909
## jump     45.6858787 432.7247457 326.63093 313.43528 1626.391512
497.565501
## run_20   -7.0190919 177.6297691 107.04834 102.35391  497.565501
309.892580
## flexion -22.8912136  16.8575694 -12.43979 -10.09381   -3.251873
15.767191
## run_10   -0.4388183 -22.0483625 -14.05455 -13.38593  -69.938809 -
25.299908
##             flexion      run_10
## age      -1.1105308  10.4380866
## height  -21.2239829  -9.5514291
## weight  -30.6915146  -8.0126939
## BMI      -4.9434876  -0.1151876
## bodyfat  -2.2293706   9.2960679
## waist   -22.8912136  -0.4388183
## situp    16.8575694 -22.0483625
## grip_D  -12.4397886 -14.0545506
## grip_ND -10.0938122 -13.3859347
## jump     -3.2518731 -69.9388094
## run_20   15.7671905 -25.2999081
## flexion  88.4914961  -0.6465164
## run_10   -0.6465164   4.5586865
```

분산공분산 행렬 S를 살펴본 결과, 각 변수들의 분산의 크기가 크게 다른 것을 알 수 있다. 따라서 분산공분산 행렬이 아닌 상관행렬을 사용해 전체데이터 주성분 분석을 실시하기로 한다.

```
PC.result.R<-princomp(data1,cor=TRUE)
PC.result.R
```

```
## Call:
## princomp(x = data1, cor = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
```

```
## 2.44635127 1.73288617 1.02065839 0.98890695 0.70732716 0.58381340
##     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
## 0.54466226 0.53704404 0.44064941 0.41800366 0.37320474 0.23328987
##    Comp.13
## 0.06294828
##
##  13  variables and  4292 observations.
```

```
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
```

```
PC.result.R$loadings
```

```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## age     0.132  0.190  0.190  0.824  0.321         0.279  0.138  0.103
## height -0.332         0.291        -0.518 -0.112  0.443  0.176  0.255
## weight -0.268  0.412               -0.164         0.104        -0.143
## BMI    -0.109  0.499 -0.305         0.184        -0.193        -0.396
## bodyfat  0.261  0.322 -0.291 -0.240  0.169                       0.736
## waist  -0.127  0.501 -0.121                       0.290
## situp  -0.314 -0.192 -0.219         0.262        -0.167  0.833  0.150
## grip_D -0.370               0.208               -0.428 -0.187  0.250
## grip_ND -0.366              0.243 -0.102        -0.428 -0.192  0.237
## jump   -0.368 -0.116                0.111  0.277  0.128 -0.109 -0.199
## run_20 -0.302 -0.209 -0.148         0.344 -0.778  0.191 -0.282
## flexion        -0.211 -0.766  0.368 -0.462         0.123
## run_10  0.327  0.169  0.123  0.102 -0.325 -0.547 -0.363  0.270 -0.144
##        Comp.10 Comp.11 Comp.12 Comp.13
## age     0.111
## height  0.196                  -0.416
## weight  0.370                   0.734
## BMI     0.350                  -0.537
## bodyfat         0.327
## waist  -0.710  -0.343
## situp
```

```
## grip_D                  0.713
## grip_ND                -0.701
## jump    -0.337   0.761
## run_20
## flexion
## run_10  -0.239   0.390
##
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.077  0.077  0.077  0.077  0.077  0.077  0.077  0.077
## Cumulative Var  0.077  0.154  0.231  0.308  0.385  0.462  0.538  0.615
##               Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
## SS loadings     1.000   1.000   1.000   1.000   1.000
## Proportion Var  0.077   0.077   0.077   0.077   0.077
## Cumulative Var  0.692   0.769   0.846   0.923   1.000
```

전체데이터 주성분 분석 결과를 해석해보면 첫번째로, Comp.1 변수의 loading 에 따르면 가장 많은 설명을 하는 주성분 1 은 나이와 체지방률 대비 나머지 변수의 효과를 나타낸 것을 알 수 있다. run_10 의 경우 달리기가 빠를수록 기록이 작기 때문에 음수로 나타난 것으로 보인다.
두번째로, 주성분 2 는 몸무게, BMI, 체지방률, 허리둘레와 같이 신체의 비만도를 정량적으로 나타내는 변수들의 효과를 나타냈다.
세번째로, 주성분 3 는 유연성의 효과를 나타내고 있다.

(3) 가설설정 및 분석방향 제시

자료에는 범주형 변수가 성별, 연령, 지역으로 총 세 가지가 있다. 본 팀은 성별, 연령, 지역이라는 집단별로 체력조사 데이터에 유의미한 차이가 존재하는지 알아보기 위해 가설을 다음과 같이 설정하였다.

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.

위 네 가지의 가설을 검정하기 위해 주성분분석, 판별분석, MANOVA, 군집분석과 같은 분석 방법을 사용하였다.

3. 분석 과정 및 결과

(1) 성별별로 차이가 있을 것이다.

- 분석 방법 : 주성분 분석, 판별분석, K-평균법

① 주성분 분석

1) 남자데이터 주성분 분석

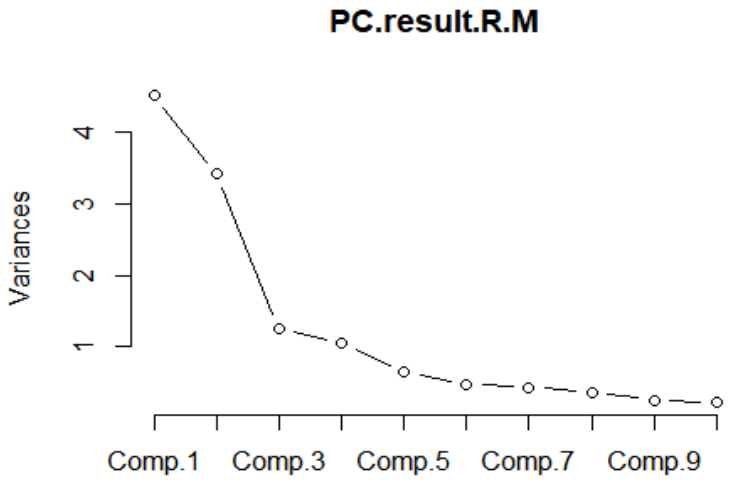
```
data_M <- data[data$sex==1,]
data_M1<- data_M[,c(-1,-3,-16)]
PC.result.R.M<-princomp(data_M1,cor=TRUE)
PC.result.R.M
```

```
## Call:
## princomp(x = data_M1, cor = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 2.1239536 1.8464598 1.1173526 1.0275294 0.8051078 0.6880795 0.6648803
##    Comp.8    Comp.9   Comp.10   Comp.11   Comp.12   Comp.13
## 0.6056382 0.5104315 0.4824984 0.4279815 0.4074050 0.0458978
##
##  13  variables and  2146 observations.
```

```
summary(PC.result.R.M)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.1239536 1.8464598 1.11735258 1.02752936
## Proportion of Variance 0.3470138 0.2622626 0.09603668 0.08121666
## Cumulative Proportion  0.3470138 0.6092764 0.70531303 0.78652969
##                           Comp.5     Comp.6     Comp.7    Comp.8
## Standard deviation     0.80510783 0.68807949 0.66488027 0.6056382
## Proportion of Variance 0.04986143 0.03641949 0.03400506 0.0282152
## Cumulative Proportion  0.83639112 0.87281061 0.90681567 0.9350309
##                           Comp.9    Comp.10    Comp.11   Comp.12
## Standard deviation     0.51043152 0.48249840 0.42798150 0.4074050
## Proportion of Variance 0.02004156 0.01790805 0.01408986 0.0127676
## Cumulative Proportion  0.95507244 0.97298050 0.98707035 0.9998380
##                           Comp.13
## Standard deviation     0.0458978001
## Proportion of Variance 0.0001620468
## Cumulative Proportion  1.0000000000
```

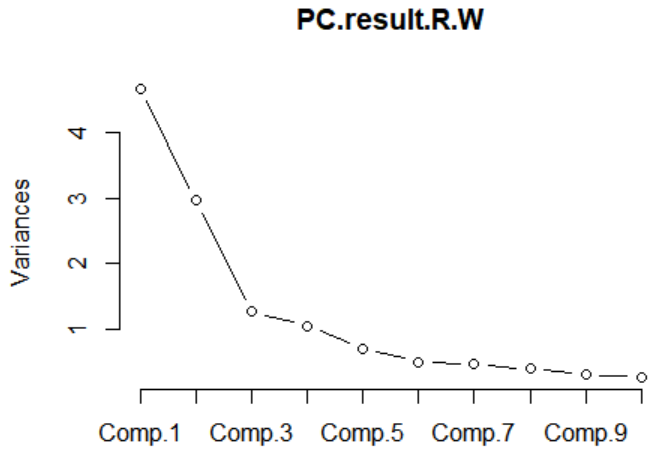
```
screeplot(PC.result.R.M,type="l")
```



2) 여자데이터 주성분 분석

```
data_W <- data[data$sex==2,]
data_W1<-data_W[,c(-1,-3,-16)]
PC.result.R.W<-princomp(data_W1,cor=TRUE)
PC.result.R.W

## Call:
## princomp(x = data_W1, cor = TRUE)
##
## Standard deviations:
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## 2.16101071 1.72296072 1.12713033 1.02905426 0.84271383 0.71189716
##     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
## 0.68602733 0.63262498 0.55615067 0.52199462 0.46783873 0.37370037
##    Comp.13
## 0.06280576
##
##  13  variables and  2146 observations.

summary(PC.result.R.W)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4    Comp.5
## Standard deviation     2.1610107 1.7229607 1.12713033 1.0290543
0.8427138
## Proportion of Variance 0.3592283 0.2283534 0.09772483 0.0814579
0.0546282
## Cumulative Proportion  0.3592283 0.5875816 0.68530644 0.7667643
0.8213925
##                           Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.71189716 0.68602733 0.63262498 0.55615067
## Proportion of Variance 0.03898443 0.03620258 0.03078572 0.02379258
## Cumulative Proportion  0.86037697 0.89657954 0.92736527 0.95115785
##                           Comp.10    Comp.11    Comp.12     Comp.13
```

```
## Standard deviation     0.52199462 0.46783873 0.37370037 0.062805761
## Proportion of Variance 0.02095988 0.01683639 0.01074246 0.000303428
## Cumulative Proportion  0.97211772 0.98895411 0.99969657 1.000000000

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



두 성별 모두 screeplot 의 elbow 가 Comp.3 에서 Comp.4 로 변하는 시점이므로 Comp.3 까지 의 주성분을 비교해보기로 한다.

3) 성별 주성분 비교

```
PC.result.R.M$loadings
PC.result.R.W$loadings
```

```
          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
```

<남성 주성분 loadings>

```
          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
```

<여성 주성분 loadings>

[주성분 1 비교]

두 변수 모두 EDA 에서 살펴본 전체 데이터의 주성분 1 과 같이 신체적인 측정 대비 체력적인 측정 변수의 효과를 나타내나 남성의 경우 여성과 다르게 몸무게(weight)의 효과가 없는 것으로 나타났다.

[주성분 2 비교]

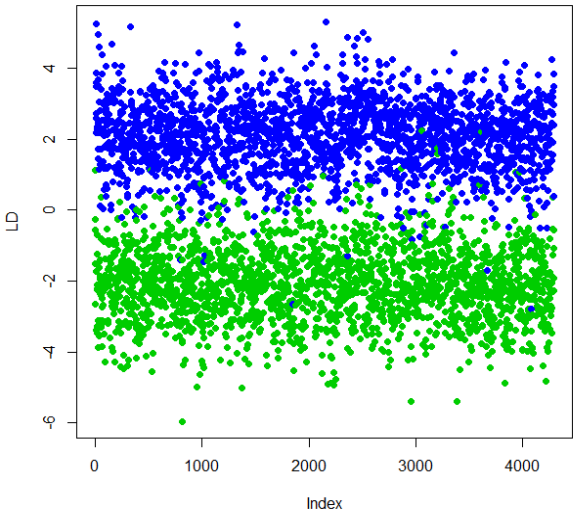
여성의 주성분 2 의 경우 몸무게, BMI, 체지방률, 허리둘레와 같이 전체데이터에서 영향력 있었던 변수들과 함께 악력과 멀리뛰기와 같은 체력 데이터도 포함된 것을 알 수 있다. 특히 윗몸일으키기와 멀리뛰기와 같은 경우에는 그 효과가 크지는 않지만, 남성 주성분 데이터에서는 해당 변수들의 효과가 아예 나타나지 않는다는 점에서는 의미를 가진다.

[주성분 3 비교]

여성 데이터의 경우에는 남성 데이터에는 포함되지않은 BMI 가 포함된다. 이는 여성의 경우 피하지방양이 많고, 남성은 근육양이 많으므로, 다음과 같은 결과가 나온 것으로 추정된다.

② 판별 분석

```
library(MASS)
datas<-data[,c(-1,-3,-16)]
datas.LDA<-lda(datas,data$sex)
datas.LDA

## Call:
## lda(datas, data$sex)
##
## Prior probabilities of groups:
##   1   2
## 0.5 0.5
##
## Group means:
##       age   height   weight      BMI  bodyfat    waist    situp   grip_D
## 1 41.05359 172.2930 73.50558 24.72588 21.93817 85.06010 38.30895
42.86302
## 2 41.10997 159.5751 58.09733 22.82697 30.52160 78.09714 23.88583
25.28956
##    grip_ND     jump   run_20  flexion   run_10
## 1 40.43467 197.3815 37.30475  9.29199 11.70492
## 2 23.67100 140.0406 19.75582 15.08002 14.06873
##
## 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

error.list<-which(datas.LDA$group!=
                  predict(datas.LDA)$class)
LD<-predict(datas.LDA)$x
plot(LD,type='n')
points(LD,col=as.numeric(data$sex)+2,pch=16)
points(LD[error.list,],col=2,pch=14,cex=1.5)
```



성별 판별분석의 경우, 두 성별이 LD1 하나로 거의 정확하게 판별되는 것을 확인할 수 있었다. 특히 BMI 가 많은 영향을 끼치는 것을 알 수 있다.

③ K-평균법

본 자료의 개체들은 크게 성별별, 연령별, 지역별로 나눌 수 있다. 그 중, 성별별 관측치의 거리가 가장 길 것으로 예상되므로 kmeans 방법을 통해 2 개의 군집으로 나누어보고, 이를 실제 데이터의 성별과 비교해보았다.

```
data2<-data[,c(4:15)] #범주형 변수를 뺀 자료
data2.kmeans <- kmeans(data2,center=2) #kmeans 방법으로 2 개의 그룹으로 분류

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

## [1] 0.1102307

data2.kmeans$centers

##     height   weight      BMI  bodyfat    waist    situp   grip_D  grip_ND
## 1 172.0961 72.63813 24.47188 21.59631 84.07920 40.85415 42.90816
40.52663
## 2 160.5454 59.82537 23.16916 30.28364 79.39435 22.55745 26.35400
24.64297
##       jump   run_20  flexion   run_10
## 1 204.0301 39.94655 10.94561 11.35937
## 2 137.8014 18.54391 13.26733 14.22403

a<-data[,c(3:15)] %>% group_by(sex) %>%
summarize(medheight=median(height),medweight=median(weight),medBMI=median(
BMI),medbodyfat=median(bodyfat),medwaist=median(waist),medsitup=median(sit
up),medgrip_D=median(grip_D),medgrip_ND=median(grip_ND),medjump=median(jum
p),medrun_20=median(run_20),medflextion=median(flexion),medrun_10=median(r
un_10))

as.data.frame(a)

##   sex medheight medweight   medBMI medbodyfat medwaist medsitup
medgrip_D
## 1   1     172.0      72.5 24.53266       21.9    84.30       38      42.6
## 2   2     159.5      57.0 22.34915       30.9    77.35       23      25.1
##   medgrip_ND medjump medrun_20 medflextion medrun_10
## 1       40.2     198        34        10.0    11.470
## 2       23.5     140        17        16.1    13.865
```

오분류율이 0.11 로 약 90%의 자료가 실제와 같게 군집화 되었다. 즉, 성별별 거리 차이가 확연히 존재함을 알 수 있다. 또한, 군집화 된 그룹별 변수들의 중심과, 실제 성별 별 변수들의 중앙값을 비교해본 결과, 거의 값이 비슷하다.

(2) 연령별로 차이가 있을 것이다

- 분석 방법 : 주성분 분석, 판별분석, 계층적 군집분석

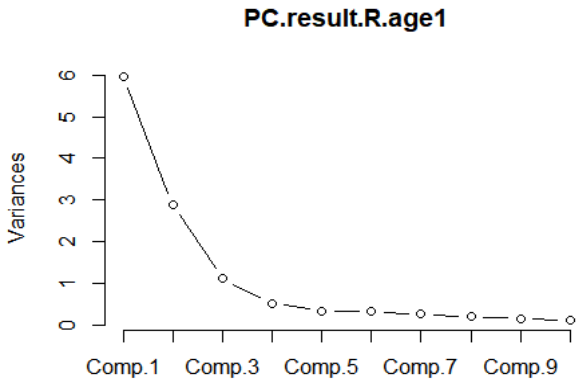
① 주성분 분석

1) 연령별 데이터 나누기

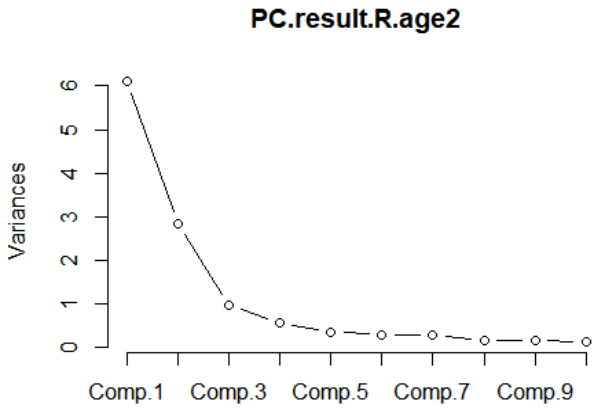
```
data$age_group2<-
ifelse(data$age_group==1|data$age_group==2,1,data$age_group)
data$age_group2<-
ifelse(data$age_group==3|data$age_group==4,2,data$age_group2)
data$age_group2<-
ifelse(data$age_group==5|data$age_group==6,3,data$age_group2)
data$age_group2<-
ifelse(data$age_group==7|data$age_group==8|data$age_group==9,4,data$age_gr
oup2)

data_age1<-data[data$age_group2==1,]
data_age2<-data[data$age_group2==2,]
data_age3<-data[data$age_group2==3,]
data_age4<-data[data$age_group2==4,]

#group1 : 19 세 ~ 29 세
#group2 : 30 세 ~ 39 세
#group3 : 40 세 ~ 49 세
#group4 : 50 세 ~ 64 세
```

1) 주성분 분석 후 screeplot 그리기

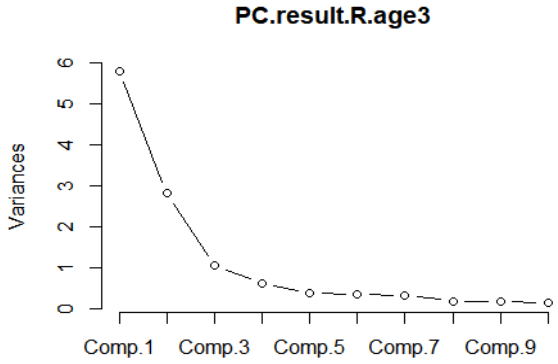
```
data_age1_0<- data_age1[,c(-1,-2,-3,-16,-17)]
PC.result.R.age1<-princomp(data_age1_0,cor=TRUE)
screeplot(PC.result.R.age1,type="l")
```



```
data_age1_0<- data_age2[,c(-1,-2,-3,-16,-17)]
PC.result.R.age2<-princomp(data_age1_0,cor=TRUE)

screeplot(PC. result.R.age2,type="l")
```

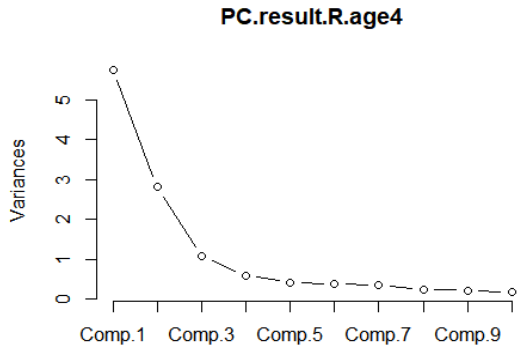



```
data_age1_0<- data_age3[,c(-1,-2,-3,-16,-17)]
PC.result.R.age3<-princomp(data_age1_0,cor=TRUE)

screeplot(PC.result.R.age3,type="l")
```




```
data_age1_0<- data_age4[,c(-1,-2,-3,-16,-17)]
PC.result.R.age4<-princomp(data_age1_0,cor=TRUE)
```

```
screeplot(PC.result.R.age4,type="l")
```



3) 주성분 변수 해석

```
PC.result.R.age1$loadings
PC.result.R.age2$loadings
PC.result.R.age3$loadings
PC.result.R.age4$loadings
```

```
         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
```

<group1:19 세~29 세>

```
         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
```

<group2:30 세~39 세>

```
         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
```

<group3:40 세~49 세>

```
         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
```

<group4:50 세~64 세>

Screeplot 을 그려본 결과 네 그룹 모두 elbow 가 같기 때문에 전부 Comp.1, Comp.2, Comp.3 의 세 변수를 사용하기로 한다. 세 변수는 전체 자료의 80%를 설명한다

네 그룹에서 각 주성분변수들의 효과에 별다른 차이점이 보이지 않지만, 50-60 대 데이터에서는 제 1 주성분의 BMI 가 빠지는 차이를 확인할 수 있다.

② 판별분석

1) 판별분석

```
library(MASS)
datas<-data[,c(-1,-2,-3,-16,-17)]
datas<-scale(datas)
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
```

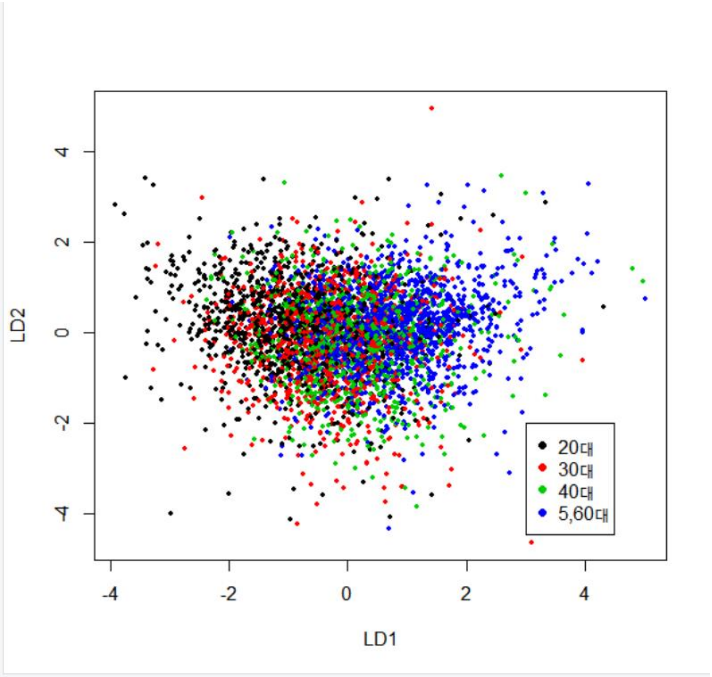
2) 판별분석 LDA 계수

```
##
## 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
## run_10   0.41312446  0.33160024 -0.1031137
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.9302 0.0561 0.0138
```

3) 판별분석 그림 그리기

```
error.list<-which(datas.LDA$group!=
                    predict(datas.LDA)$class)
LD<-predict(datas.LDA)$x
plot(LD,type='n')
points(LD,col=as.numeric(data$age_group2),pch=16,cex=0.5)
legend(3,-2,c("20 대","30 대","40 대","5,60 대
"),pch=c(16,16,16,16),col=c(1,2,3,4))
```



연령대별 판별분석을 시행한 결과, proportion of trace 를 보면 LD1 이 자료의 95%, LD2 는 5%를 설명함을 알 수 있다. 즉, 대부분의 데이터가 LD1 에 의해 설명되고 있다. 설명비율이 거의 0%에 가까운 LD3 를 제외하고 LD1 과 LD2 를 축으로 그래프를 그려본 결과, 상대적으로 다른 group 에 비해 20 대와 5,60 대 group 의 경우 서로의 반대편에 위치함을 확인할 수 있다. 반면에 30 대, 40 대의 경우 전체 평면에 고루 분포하여 판별분석을 통한 해석이 불분명하다. 본 팀이 추출한 데이터에는 19 세 미만과 65 세이상의 데이터가 없으나 해당 데이터가 있다면 초년, 중년, 장년, 노년층의 판별이 더 뚜렷할 것으로 추정한다.

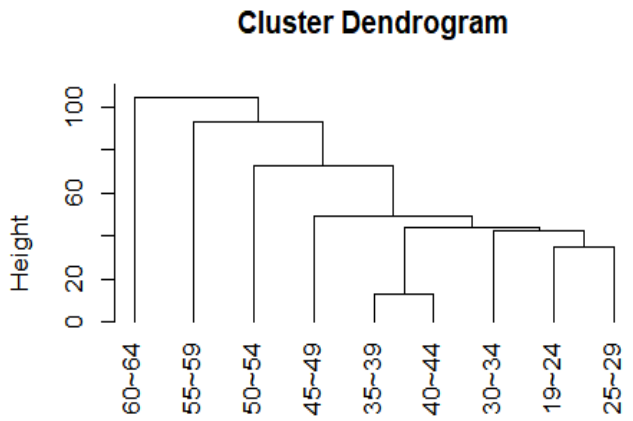
③ 계층적 군집분석

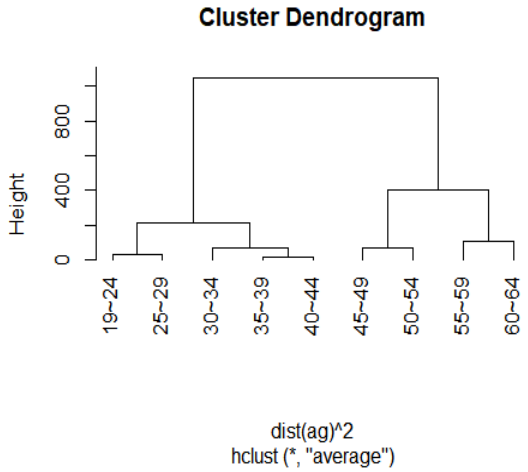
본 데이터에 존재하는 연령그룹 9 개를 최단연결법, 평균연결법, 최장연결법을 이용하여 군집화 해보았다.

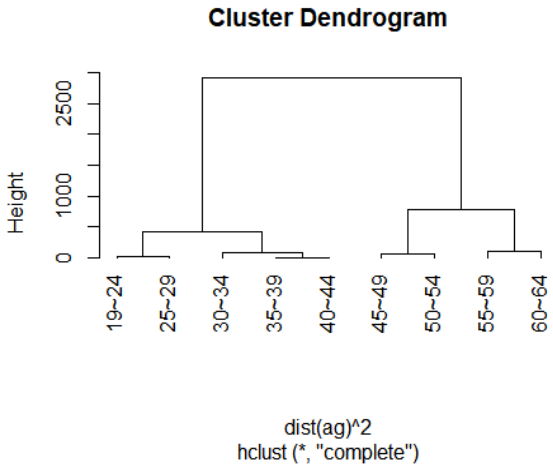
```
data_agegroup<-data
data_agegroup$age_group<-
dplyr::recode(data$age_group,`1`="19~24",`2`="25~29",`3`='30~34',`4`='35~39',`5`='40~44',`6`='45~49',`7`='50~54',`8`='55~59',`9`='60~64')

ag<-data_agegroup %>% group_by(age_group) %>%

summarize(avgheight=mean(height),avgweight=mean(weight),avgBMI=mean(BMI),avgbodyfat=mean(bodyfat),avgwaist=mean(waist),avgsitup=mean(situp),avggrip_D=mean(grip_D),avggrip_ND=mean(grip_ND),avgjump=mean(jump),avgrun_20=mean(run_20),avgflextion=mean(flexion),avgrun_10=mean(run_10))
ag<-as.data.frame(ag)
row.names(ag)<-ag[,1]
ag<-ag[,-1]
ag
```

```
##      avgheight avgweight   avgBMI avgbodyfat avgwaist avgsitup avggrip_D
## 19~24  167.9118  64.61105 22.79795   23.67949 78.03536 39.21696 34.42899
## 25~29  168.6018  66.32820 23.18350   24.38961 80.13320 37.50615 34.96721
## 30~34  167.9895  68.08111 23.93917   25.70741 82.11882 35.14754 36.04147
## 35~39  167.4791  66.55467 23.60661   26.15113 81.92627 31.89959 35.11578
## 40~44  166.2935  67.00365 24.07592   26.40266 81.98839 31.61475 35.12336
## 45~49  165.1172  66.33024 24.20416   26.93100 82.72853 29.49605 34.33992
## 50~54  164.1266  65.01933 24.04259   27.40482 82.17888 26.72047 32.98602
## 55~59  162.8224  64.36024 24.17304   27.90227 82.70164 23.93110 31.91732
## 60~64  161.6655  63.05774 24.07284   28.17061 82.92591 20.75806 30.62935
##       avggrip_ND  avgjump avgrun_20 avgflextion avgrun_10
## 19~24   31.68249 186.2968  38.42604    12.90375  11.83097
## 25~29   32.79836 182.7869  35.30943    12.52998  11.95639
## 30~34   33.93125 178.6756  32.58197    12.96193  12.30998
## 35~39   33.04857 174.4666  29.73566    11.21135  12.53644
## 40~44   33.15799 171.5148  28.20902    11.38715  12.73832
## 45~49   32.31542 165.7939  25.46838    11.30553  13.12915
## 50~54   31.28839 158.2559  24.31890    12.06921  13.55575
## 55~59   30.14665 149.9906  20.99803    12.66130  13.89693
## 60~64   29.17935 141.0639  18.13548    12.86455  14.63413
```

```
hc<-hclust(dist(ag)^2,method="single")
plot(hc,hang=-1)

hc<-hclust(dist(ag)^2,method="average")
plot(hc,hang=-1)

hc<-hclust(dist(ag)^2,method="complete")
plot(hc,hang=-1)
```







평균연결법과 최장연결법의 결과가 동일하므로 이것을 해석해보겠다. 확연하게는 19 세~44 세와 45 세~64 세 두개의 그룹으로 나뉜다. 국민체력이 40 대 중반을 기준으로 많은 변화가 생긴다고 해석할 수 있다. 4 개의 그룹으로 나누면 10~20 대, 30~40 대 중반, 40 대 중반~50 대 중반, 50 대 중반~60 대 중반 그룹으로 나눌 수 있다. 최단연결법과 함께 보아도, 연령의 흐름에 따라 비슷한 경향을 보이는 것을 확인할 수 있다.

(3) 지역별로 차이가 있을 것이다.

- 분석 방법: MANOVA, 주성분을 이용한 ANOVA, 계층적 군집화

① MANOVA

지역별로 비만도에 차이가 있는지 알아보기 위해 MANOVA 를 실시하기로 하였다. MANOVA 를 실시할 때 어떤 변수들을 이용하여 지역별 차이를 검정할지 알아보기 위하여 상관행렬을 구하였다.

```
> 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
```

비만도에 영향을 주리라 기대되는 변수들인 weight, BMI, bodyfat, waist 중에서 bodyfat 을 제외한 나머지 변수들 간 상관관계가 크므로 weight, BMI, waist 의 세 변수를 이용하여 MANOVA 를 실시한다.

MANOVA 를 실시하기에 앞서 등분산 가정을 만족하는지 검정하였다. 관측수 4000 개 이상으로, 표본의 크기가 충분히 크기 때문에 정규성은 만족한다고 보았다.

[분산에 대한 동일성 검정]

```
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':
##
##     some

# weight*waist*BMI
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 가 0.18 로 유의수준 0.05 에서 기각되지 않는다. 따라서 등분산 가정을 만족한다고 볼 수 있다.

다음으로 MANOVA 를 실시하였다.

```
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
```

```
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
```

Pillai's trace 를 이용한 검정과 Wilk's lambda 를 이용한 검정에서 p-value 의 값이 모두 매우 작게 나왔으므로 유의하다. 즉 지역에 따라 비만도의 차이가 있다고 할 수 있다.

② 주성분을 이용한 ANOVA

앞서 전체 데이터에 대한 주성분 분석 결과, 두번째 주성분은 몸무게, BMI, 체지방률, 허리둘레와 같이 신체의 비만도를 정량적으로 나타내는 변수들의 효과를 나타내는 것을 알 수 있었다. 이를 이용하여 비만도를 나타내는 새로운 변수 obesity 를 생성시키고 지역간 비만도 차이를 보기 위해 ANOVA 를 실시하였다.

```
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
```

Obesity (두번째 주성분) = 0.412*체중 + 0.499*BMI + 0.322*체지방률 + 0.501*허리둘레

새로운 변수 obesity 를 포함한 새로운 데이터 data2 를 생성하였다.

```
library(dplyr)
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

head(data2)

##   location age sex height weight      BMI bodyfat waist situp grip_D
## 1        1  19   1  164.3  58.04 21.50068    23.6  75.2    46   29.0
## 2        1  19   1  167.9  52.90 18.76525    11.7  69.5    38   27.1
## 3        1  19   1  171.2  51.20 17.46877    12.1  67.9    36   26.6
## 4        1  19   1  172.6  57.70 19.36842    17.1  73.0    28   37.3
## 5        1  19   1  172.7  59.80 20.05010     8.9  71.2    39   38.2
## 6        1  19   1  175.5  73.60 23.89591    18.3  83.0    50   40.2
##   grip_ND jump run_20 flexion run_10 age_group  obesity
## 1    28.7  212     32    20.3   9.98         1 79.91572
```

```
## 2    26.5  224     26     3.7  11.68         1 69.74556
## 3    24.6  210     44     7.4  10.61         1 67.72542
## 4    37.5  179     34   -11.2  12.10         1 75.51644
## 5    35.7  207     54    -9.5   9.90         1 73.17960
## 6    37.4  234     50    15.2   9.77         1 89.72286
```

```
summary(data2$obesity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.68   80.43   87.28   88.29   94.81  159.30
```

ANOVA 를 실시하기에 앞서 등분산 가정을 만족하는지 알아보기 위해 분산에 대한 동일성 평가를 하였다. 관측수 4000 이상으로 표본의 크기가 충분히 크기 때문에 정규성은 만족하는 것으로 보았다.

[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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value 가 0.03 으로, 유의수준 0.01 에서 기각되지 않으므로 등분산성을 만족한다고 볼 수 있다.

지역에 따라 비만도(obesity)에 차이가 있는지 검정하기 위해 ANOVA 를 실시하였다.

```
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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

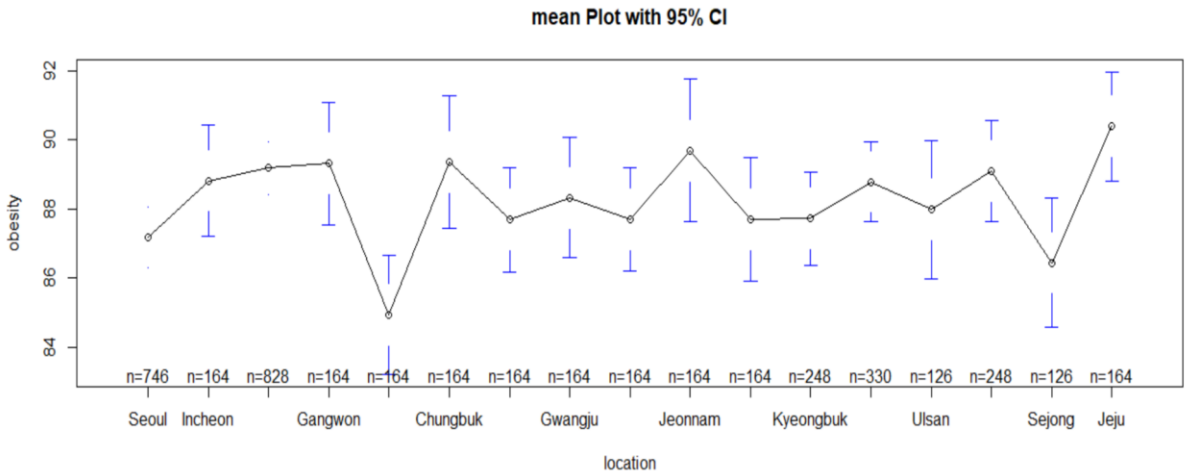
p-value 가 매우 작은 값을 가지므로 유의하다. 즉, 지역에 따라 비만도의 차이가 있다고 할 수 있다.

그래프를 통해 확인하면 다음과 같다.

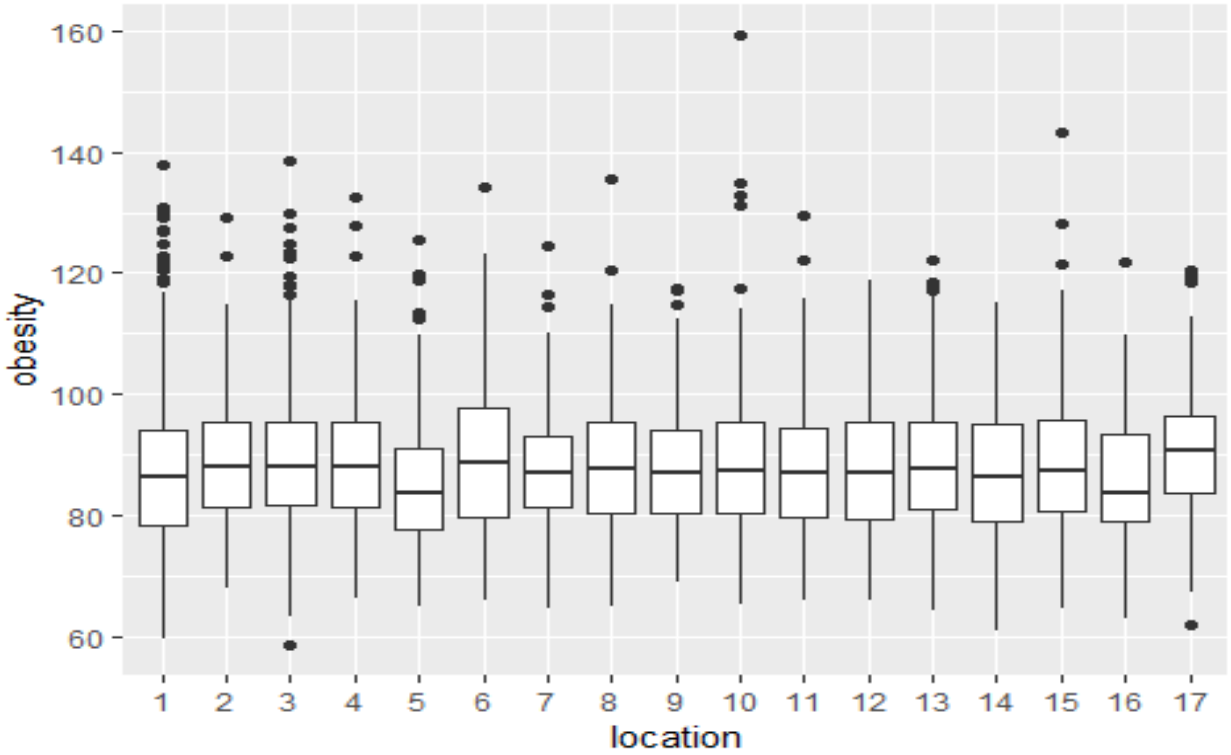
```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess

par(mfrow=c(1,1))
plotmeans(data2$obesity~data2$location,xlab="location",ylab="obesity",main
="mean Plot with 95% CI")
```




```
ggplot(data2,aes(x=location,y=obesity))+geom_boxplot()
```



③ 계층적 군집화

어떤 지역끼리 비슷한 체력 수준으로 묶이는지 알아보기 위하여 계층적 군집분석을 실시하였다.

1) 17 개 지역을 전체 변수에 대하여 군집화

지역별 전체적인 국민체력 차이를 보기위해, 지역별로 변수마다 평균을 내어 정리하였다.

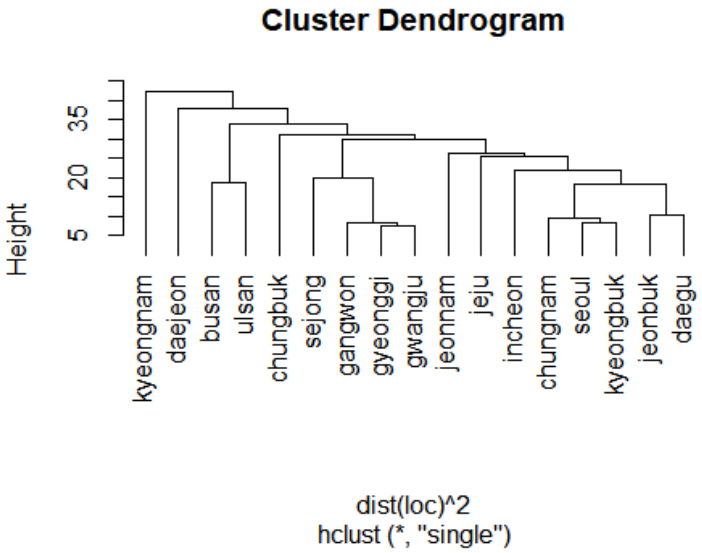
변수 평균을 가지고 최단연결법, 평균연결법, 최장연결법을 통해 군집화 해보았다.

```
library(car)

data_location<-data
data_location$location<-
dplyr::recode(data$location,`1`="seoul",`2`='incheon',`3`='gyeonggi',`4`='
gangwon',`5`='daejeon',`6`='chungbuk',`7`='chungnam',`8`='gwangju',`9`='je
onbuk',`10`='jeonnam',`11`='daegu',`12`='kyeongbuk',`13`='busan',`14`='uls
an',`15`='kyeongnam',`16`='sejong',`17`='jeju')
loc<-data_location %>% group_by(location) %>%
summarize(avgheight=mean(height),avgweight=mean(weight),avgBMI=mean(BMI),a
vgbodyfat=mean(bodyfat),avgwaist=mean(waist),avgsitup=mean(situp),avggrip_
D=mean(grip_D),avggrip_ND=mean(grip_ND),avgjump=mean(jump),avgrun_20=mean(
run_20),avgflextion=mean(flexion),avgrun_10=mean(run_10))
loc<-as.data.frame(loc)
row.names(loc)<-loc[,1]
loc<-loc[,-1]
loc
```

```
##           avgheight avgweight   avgBMI avgbodyfat avgwaist avgsitup
## seoul      165.6910  64.94566 23.50979   25.83094 80.58013 32.32215
## incheon    164.8573  65.42488 23.95865   28.80183 81.09756 29.65244
## gyeonggi   165.9365  66.44734 24.00275   27.11280 82.02917 30.68478
## gangwon    166.0018  65.79524 23.81634   26.48242 83.41402 29.96951
## daejeon    167.6244  64.67744 22.90929   25.01829 77.44939 31.56098
## chungbuk   166.3543  66.03244 23.73394   26.50122 83.36271 29.76220
## chungnam   166.1957  64.95866 23.39870   25.67643 81.80419 32.89634
## gwangju    166.0585  66.27073 23.90119   25.65366 81.50122 31.00000
## jeonbuk    165.6927  65.09512 23.59042   24.92439 81.99831 31.55488
## jeonnam    165.9122  67.72159 24.44628   26.51530 81.91829 31.90854
## daegu      165.7628  65.56476 23.70959   24.74774 81.61975 30.32317
## kyeongbuk  166.2645  65.78726 23.69175   26.07512 80.62128 32.37903
## busan      165.9662  66.35921 23.98494   26.27576 81.86109 31.56061
## ulsan      165.9802  65.83730 23.78827   24.27472 82.16349 29.00794
## kyeongnam  166.4347  65.90169 23.68866   26.07653 83.27558 28.74194
## sejong     165.9103  64.46000 23.31189   26.85794 79.04841 30.49206
## jeju       164.6439  66.89841 24.57135   26.74146 83.72851 31.53659
##           avggrip_D avggrip_ND  avgjump avgrun_20 avgflextion avgrun_10
## seoul      33.10604   31.23208 170.8698  30.11678   12.308094  12.79397
## incheon    32.67195   30.49817 170.7256  28.20732   11.829268  12.86884
## gyeonggi   33.57021   31.32959 166.7371  25.92150   11.682246  12.82157
## gangwon    34.27805   32.49177 168.4713  25.81707   11.934390  12.98447
## daejeon    33.88476   31.93720 167.1829  28.50610    9.981098  12.56927
## chungbuk   35.39512   33.13354 168.6280  31.20732   10.049268  12.98701
## chungnam   34.77561   33.96402 173.6646  31.96341   13.251707  13.03902
## gwangju    35.02012   32.67622 166.6220  25.07317   12.060976  12.91685
```

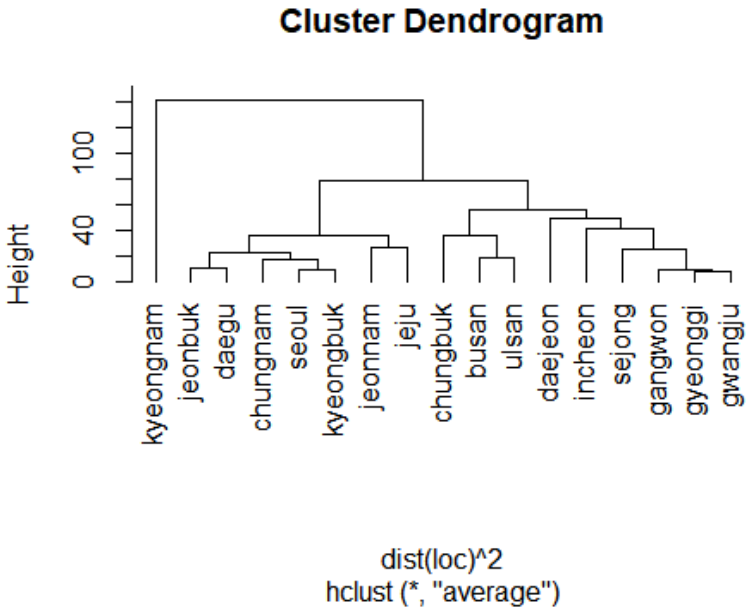
```
## jeonbuk     33.41341   33.56402 174.9518  28.36585   14.065244  12.56384
## jeonnam     36.68841   33.96463 170.2921  29.31707   13.535366  13.04543
## daegu       34.48841   32.59268 173.9451  28.59146   11.752378  12.61256
## kyeongbuk   33.94395   31.98194 172.7540  30.83065   13.720040  12.86480
## busan       34.26939   32.01727 163.6461  30.73030   12.339515  13.02298
## ulsan       32.79921   31.18571 165.6825  31.38889   12.714286  13.48103
## kyeongnam   34.79315   32.65242 159.6855  26.95161   11.353266  13.39790
## sejong      33.47222   30.52937 166.0556  23.65079   11.707143  12.91175
## jeju        37.31341   34.21524 174.6159  30.23171   13.932256  12.56073
```

```
hc<-hclust(dist(loc)^2,method="single")
plot(hc,hang=-1)
```

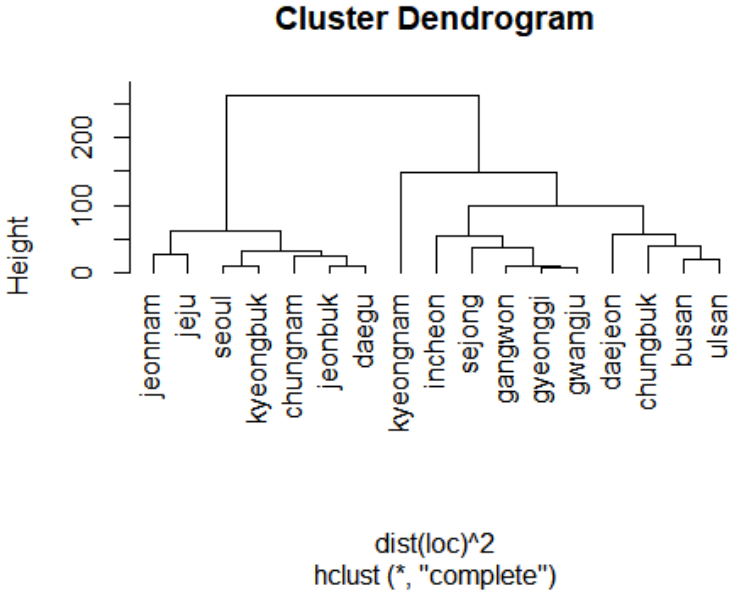


명확한 군집을 보기 어려워 다음 2 개의 결과를 해석하겠다.

```
hc<-hclust(dist(loc)^2,method="average")
plot(hc,hang=-1)
```




```r
hc<-hclust(dist(loc)^2,method="complete")
plot(hc,hang=-1)
```



평균연결법	최장연결법
경남	경남
전북, 대구, 충남, 서울, 경북, 전남, 제주	전남, 제주, 서울, 경북, 충남, 전북, 대구
충북, 부산, 울산, 대전, 인천, 세종, 강원, 경기, 광주	인천, 세종, 강원, 경기, 광주, 대전, 충북, 부산, 울산

지역별로 전체 변수들의 평균을 내어 지역 간의 군집화를 진행하였다. 최단연결법, 평균연결법, 최장연결법의 세 가지 방법에 의해 군집화를 진행한 결과는 조금씩 다르나, 경남지역은 일관적으로 가장 마지막에 분류되는 것을 발견할 수 있다. 경남 지역의 신체적 특징이 다른 지역과는 구별된다고 해석할 수 있다. 또한, 군집화가 극단적으로 일어나는 최단연결법을 제외하고, 평균연결법과 최장연결법의 결과를 비교해보면, 3 개의 그룹으로 나누었을 때 완벽히 일치하는 것을 볼 수 있다. 그 이유를 파악하는 데에는 이 자료만으로 한계가 있다.

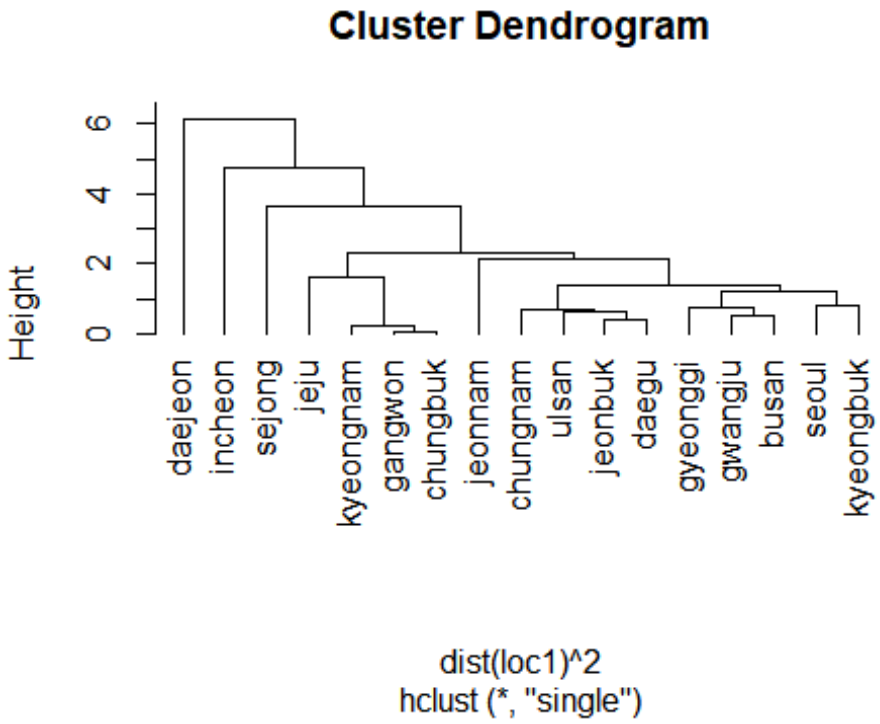
2) 17 개 지역을 비만도에 대하여 군집화

앞서 주성분으로 새롭게 만들었던 '비만도'를 구성하는 네 개의 변수 weight, BMI, bodyfat, waist 만을 가지고 다시 군집화 해보았다. 각 변수의 계수가 비슷하므로 각 변수에 대한 평균을 내어 군집화에 이용하였다.

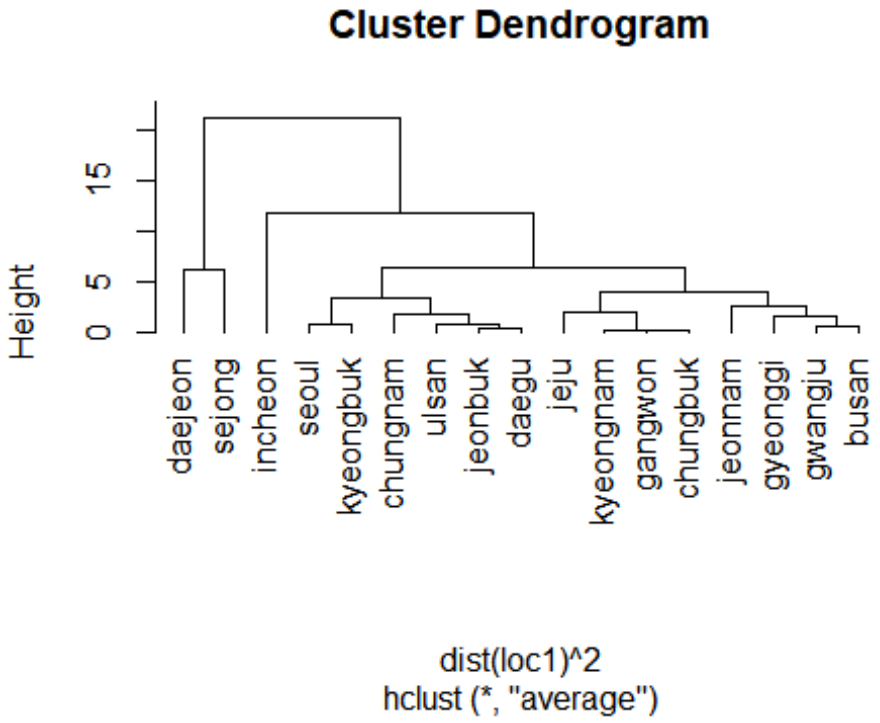
```
loc1<-data_location %>% group_by(location) %>%
  summarize(avgweight=mean(weight),avgBMI=mean(BMI),avgbodyfat=mean(bodyfa
t),avgwaist=mean(waist))
loc1<-as.data.frame(loc1)
row.names(loc1)<-loc1[,1]
loc1<-loc1[,-1]
loc1
```

```
##           avgweight   avgBMI avgbodyfat avgwaist
## seoul      64.94566 23.50979   25.83094 80.58013
## incheon    65.42488 23.95865   28.80183 81.09756
## gyeonggi   66.44734 24.00275   27.11280 82.02917
## gangwon    65.79524 23.81634   26.48242 83.41402
## daejeon    64.67744 22.90929   25.01829 77.44939
## chungbuk   66.03244 23.73394   26.50122 83.36271
## chungnam   64.95866 23.39870   25.67643 81.80419
## gwangju    66.27073 23.90119   25.65366 81.50122
## jeonbuk    65.09512 23.59042   24.92439 81.99831
## jeonnam    67.72159 24.44628   26.51530 81.91829
## daegu      65.56476 23.70959   24.74774 81.61975
## kyeongbuk  65.78726 23.69175   26.07512 80.62128
## busan      66.35921 23.98494   26.27576 81.86109
## ulsan      65.83730 23.78827   24.27472 82.16349
## kyeongnam  65.90169 23.68866   26.07653 83.27558
## sejong     64.46000 23.31189   26.85794 79.04841
## jeju       66.89841 24.57135   26.74146 83.72851
```

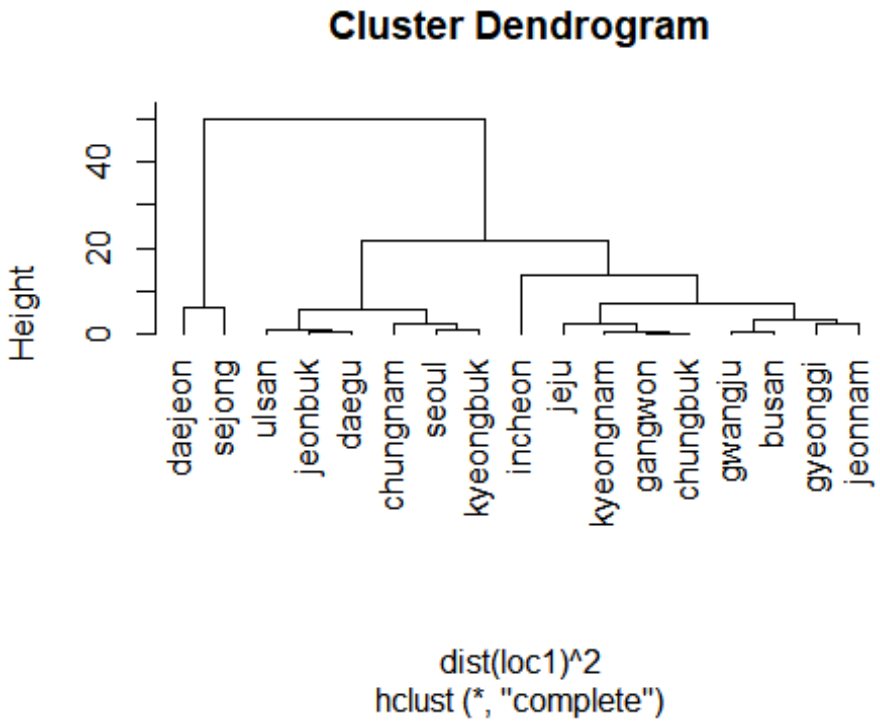
```
hc<-hclust(dist(loc1)^2,method="single")
plot(hc,hang=-1)
```




```
hc<-hclust(dist(loc1)^2,method="average")
plot(hc,hang=-1)
```




```
hc<-hclust(dist(loc1)^2,method="complete")
plot(hc,hang=-1)
```



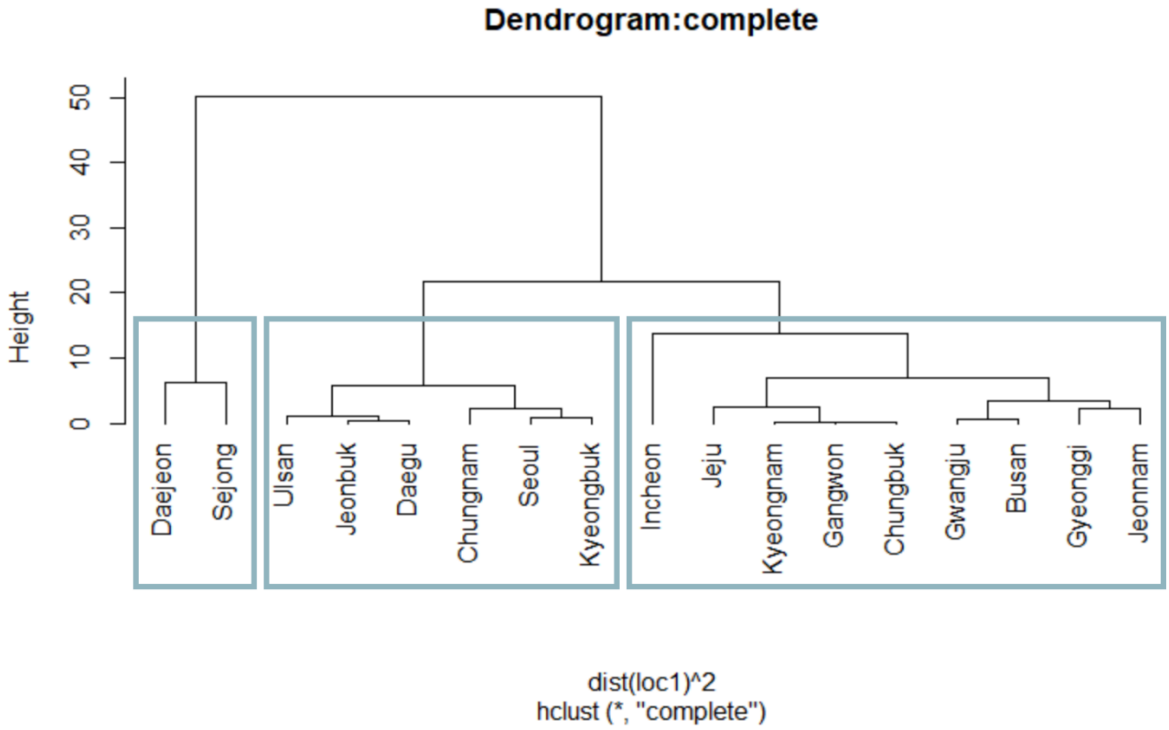
세 개의 결과 중 가장 고르게 군집화 된 최장연결법의 결과와 앞서 새롭게 만든 주성분의 지역별 결과를 비교해보겠다.

최장연결법 군집화
인천, 제주, 경남, 강원, 충북, 광주, 부산, 경기, 전남
울산, 전북, 대구, 충남, 서울, 경북
대전, 세종

```
> 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
```

오른쪽 그림은 앞서 생성한 새로운 변수 obesity 에 대하여 각 지역의 평균 비만도 (mean_obesity)가 큰 순서대로 정렬한 것이다. 이것을 최장연결법을 이용한 군집화 결과와 비교해보면 비만도가 높은 그룹, 비만도가 중간인 그룹, 비만도가 낮은 그룹 총 세 개의 그룹으로 잘 군집화 되었음을 알 수 있다. 비만도가 높은 그룹에는 인천, 제주, 경남, 강원, 충북, 광주,

부산, 경기, 전남, 비만도가 중간인 그룹에는 울산, 전북, 대구, 충남, 서울, 경북, 비만도가 낮은 그룹에는 대전과 세종이 포함된다.



(4) 체력과 체격 데이터의 연관성

- 분석 방법: 정준상관분석

① 정준상관분석

본 데이터의 변수는 크게 3 가지로 분류할 수 있다. 각 개체의 연령, 지역, 성별 등 특성을 나타내는 범주형 변수와 키, 몸무게 등 체격 특징을 나타내는 변수, 윗몸일으키기, 악력, 왕복달리기 등 체력적 특징을 나타내는 변수이다. 이 중, 체격 특징과 체력적 특징의 관계를 알아보기 위해 정준상관분석을 사용하였다. 체격 변수 그룹에는 키(height), 몸무게(weight), BMI, 체지방률(bodyfat), 허리둘레(waist) 변수가 해당된다. 체력적 특징 그룹은 그 성격에 따라 힘과 운동 능력으로 나누었다. 자주 쓰는 손의 악력(grip_D), 반대 손의 악력(grip_ND) 변수는 악력 변수 그룹으로, 윗몸일으키기(situp), 20m 왕복 오래 달리기(run_20), 10m 왕복 오래 달리기(run_10) 변수는 체력 변수 그룹으로 결정하였다. 각 변수들의 단위가 모두 다르기 때문에 표준화하여 사용하였으며, 정준상관계수는 척도불변이기 때문에 변하지 않는다.

1) 체격과 체력의 관련성

```
library(CCA)

library(ggplot2)
data.std<-scale(data1)


X<-data.std[,c(2:6)]
Y<-data.std[,c(7,11,13)]
cc1<-cc(X,Y)
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
```

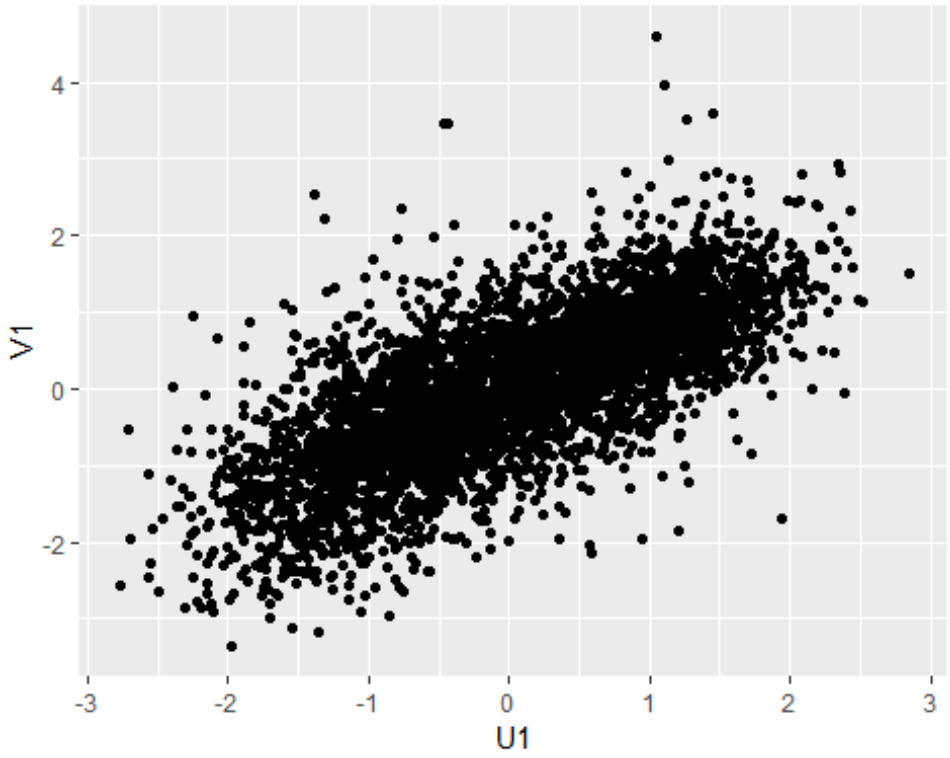
먼저, 체격 특징과 체력의 관계를 알아보기 위해, 체격 변수 집단을 X, 체력 변수 집단을 Y 로 설정하여 정준상관분석을 진행하였다. 분석 결과, 첫번째 정준상관계수는 0.715 로 꽤 높게 나오는 것을 확인할 수 있다. 첫번째 정준변수를 U1 이라 하면, U1 은 키, 몸무게, BMI 대비 체지방률과 허리둘레라고 해석할 수 있고, V1 은 10m 왕복 달리기의 결과는 짧을수록 체력이 좋다는 것을 감안하면, 전체적인 체력의 좋지 않은 정도로 해석할 수 있다. 이때 U1, V1 이 양의 선형관계를 가진다는 것은 체지방이 많을수록 체력이 떨어지는 관계가 존재한다고 해석할 수 있다.

첫번째 정준상관변수 쌍에 대해 그래프를 그려보았다.

```
cc2<-comput(X,Y,cc1)
cc2[3:6]
```

```
## $corr.X.xscores
##               [,1]      [,2]       [,3]
## height  -0.76072910 0.3980123 -0.50380292
## weight  -0.42072854 0.7828639 -0.15922927
## BMI      0.01244335 0.7002581  0.12067601
## bodyfat  0.90471047 0.3608961 -0.05580667
## waist    0.01529380 0.6551342 -0.31650763
##
## $corr.Y.xscores
##             [,1]        [,2]        [,3]
## situp  -0.6391911  0.009405126  0.033435128
## run_20 -0.6271878 -0.055633065 -0.009245693
## run_10  0.6457391 -0.039158088  0.021120568
##
## $corr.X.yscores
##                [,1]       [,2]        [,3]
## height  -0.544206575 0.04759877 -0.038100464
## weight  -0.300978674 0.09362363 -0.012041830
## BMI      0.008901661 0.08374471  0.009126212
## bodyfat  0.647207251 0.04316000 -0.004220420
## waist    0.010940803 0.07834829 -0.023936121
##
## $corr.Y.yscores
##             [,1]        [,2]       [,3]
## situp  -0.8935050  0.07864396  0.4421131
## run_20 -0.8767259 -0.46519363 -0.1222559
## run_10  0.9026582 -0.32743286  0.2792775
```

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

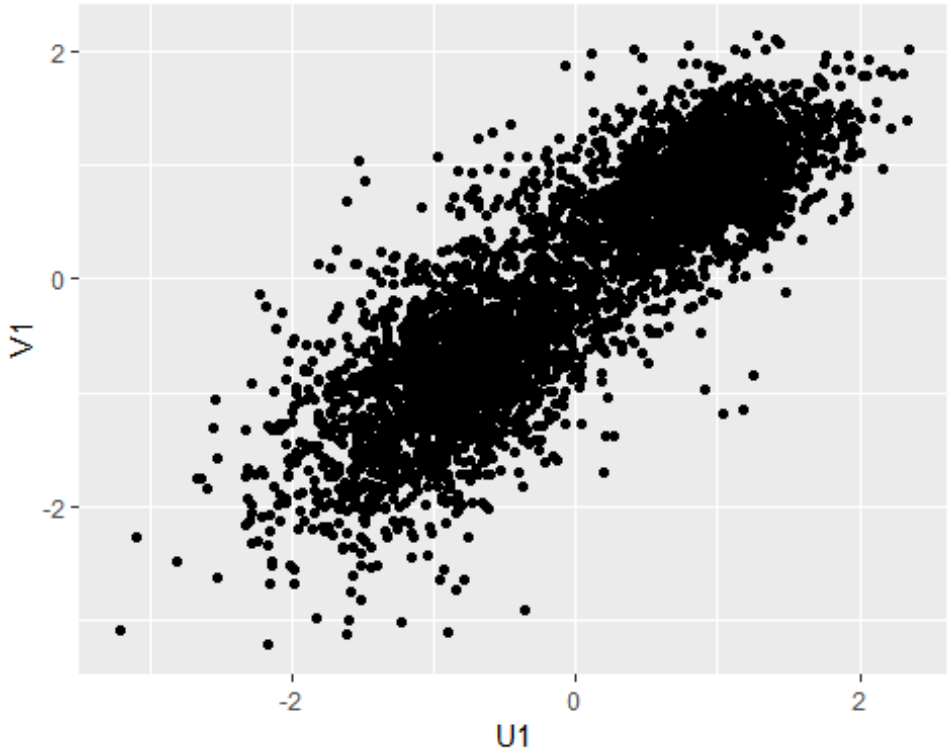


양의 상관관계가 존재하는 것을 확인할 수 있다.

2) 체격과 악력의 연관성

```
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
```

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



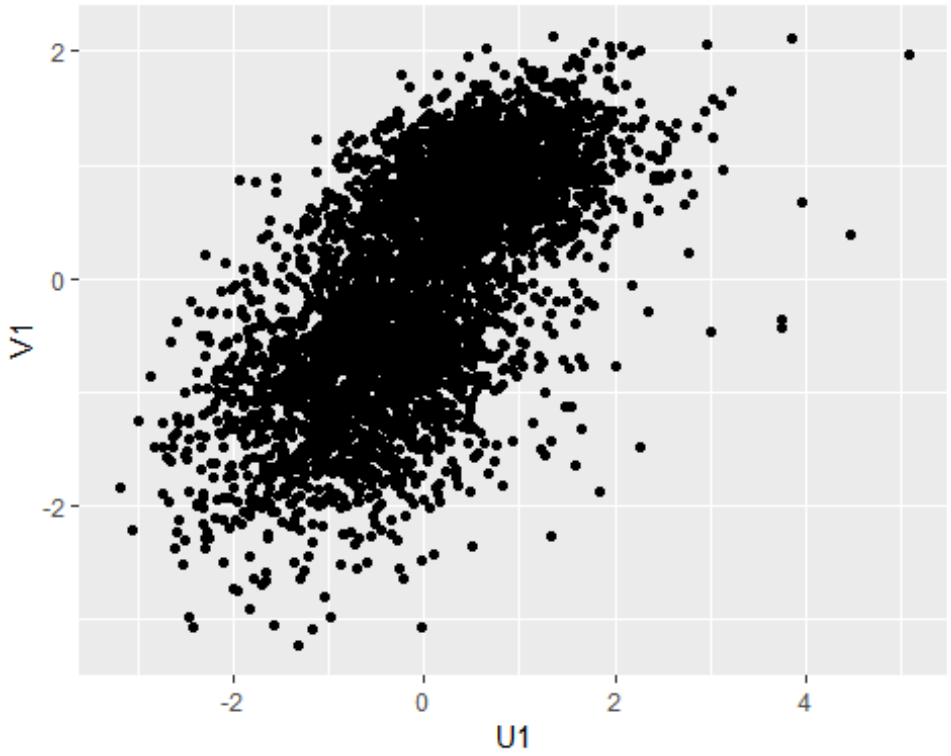
다음은, 체격 특징과 악력의 관계를 알아보기 위해, 체격 변수 집단을 X, 악력 변수 집단을 Y 로 설정하여 정준상관분석을 진행하였다. 분석 결과, 첫번째 정준상관계수는 0.856 으로 상관성이 높게 나온다. 각 정준변수의 의미를 살펴보자. 첫번째 정준변수를 U1 이라 하면, U1 은 키, 몸무게, BMI 대비 체지방률이라고 해석할 수 있고, V1 은 전체적인 악력의 약한 정도로 해석할 수 있다. U1, V1 이 0.86%의 상관계수를 가지므로 체격 대비 체지방이 많을수록 악력이 약해지는 관계가 존재한다고 해석할 수 있다. 앞서 신체적 변수 집단과 체력의 관계와 비교했을 때, 허리둘레가 악력에 큰 영향을 미치지는 않는 것을 발견할 수 있다.

첫번째 정준상관변수 쌍을 그래프로 확인한 결과, 마찬가지로 양의 선형관계를 가진다.

3) 체력과 악력의 연관성

```
X<-data.std[,c(7,11,13)]
Y<-data.std[,c(8:9)]

cc1<-cc(X,Y)
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

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



마지막으로, 체력과 악력의 관계를 알아보기 위해, 체력 변수 집단을 X, 악력 변수 집단을 Y 로 설정하여 정준상관분석을 진행하였다. 분석 결과, 첫번째 정준상관계수는 0.668 로 지금까지의 결과에서는 가장 낮은 상관성을 보인다. 각 정준변수의 의미를 살펴보자. 첫번째 정준변수를 U1 이라 하면, U1 은 전체적인 체력의 좋지 않은 정도, V1 은 전체적인 악력의 약한 정도로 해석할 수 있다. 즉, 체력과 악력이 0.668 정도로 비례하는 것이다. 뚜렷한 선형관계는 아니지만, 어느정도 상관성을 가지는 것으로 해석할 수 있다.

첫번째 정준상관변수 쌍을 그래프로 확인한 결과, 약간 뭉쳐져 있는 양의 선형관계를 확인할 수 있다.

4. 결론

여러가지 다변량 분석 방법으로 국민체력실태조사 데이터를 분석한 결과, 성별별, 연령대별, 지역별 차이를 볼 수 있었다.

먼저, 성별별 차이를 가장 뚜렷하게 볼 수 있었는데, 판별분석과 전체 데이터를 2 개로 나눈 k-means 방법에 의해 성별에 따라 데이터가 2 개로 나뉘는 것을 확인할 수 있었다. 또한, 주성분분석에서도 성별에 따른 신체적 차이에 의해 주성분에 영향을 주는 변수 차이가 있었다.

연령대별 차이는 판별분석과 군집화를 통해 볼 수 있었는데, 판별분석에서는 연령대의 차이가 클수록 확연히 데이터가 분류되는 것을 확인할 수 있었다. 또한, 군집화 결과를 보면 연령의 흐름에 따른 거리 차이가 존재하며 40 대 중반을 기점으로 완벽히 나눠진다 성별별 데이터와 같이 주성분이 명확하게 다르지는 않았으나, 연령 차이가 가장 큰 그룹에 관해서는 어느정도 차이를 가지고 있었다.

지역별 차이 또한 관찰할 수 있었는데, 특히 몸무게, BMI, 허리둘레 – 즉 비만도에 있어서 차이가 났다. 관련 주성분을 응용한 결과 대전과 세종 지역이 가장 비만도가 낮은 것으로 나타났다. 체격, 체력, 악력 변수에 대한 연관성 또한 존재하였다. 셋 중 체격과 악력의 관련성이 가장 높았으며, 세 변수 집단 모두 서로 연관되어 있음을 확인할 수 있었다.

대부분 예상한 결과를 확인할 수 있었으며, 지역별 차이의 경우 표본 수가 고르지 않은 점, 관련 정보가 부족한 점으로 인해 결과에 대한 이해를 완벽히 하기 어려웠다. 지역별 분석을 위해서는 지역적 특성에 대한 이해와 함께 충분하고 고른 표본이 필요하다.