Multivariate Data Analysis Assignment1

2014150137 통계학과

박정진

**2.3**

(a)

> w <- c(0.25, 0.25, 0.5)

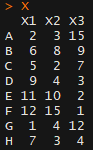
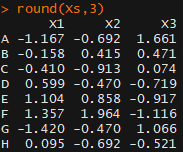
> w <- w/sqrt(sum(w^2))

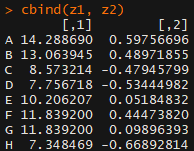
> w

[1] 0.4082483 0.4082483 0.8164966

w’=(0.4082, 0.4082, 0.8164)

(b)



(c)

> data.frame(z1 = z1[c(3,4),], z2 = z2[c(3,4),])

z1 z2

C 8.573214 -0.4794580

D 7.756718 -0.5344498

Z1과 Z2 모두에서 C의 값이 더 크다. 이는 원 데이터 X와 표준화된 Xs에서 세 번째 변수만 C가 큰데, 곱해진 w에서 세 번째 변수가 가장 큰 가중치를 가지고 있기 때문이다.

**2.4**

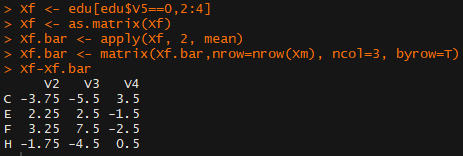
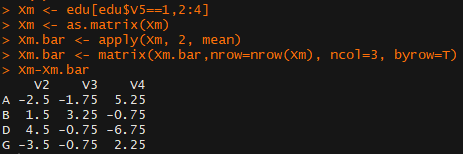
(a)



for male = (4.5, 4.75, 9.75)

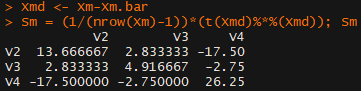


for female = (8.75, 7.50, 3.50)

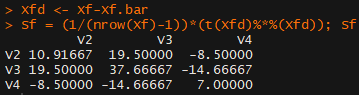


(b)

Covariance matrix for male



Covariance matrix for female



(c)

Language Aptitude와 Analogical Reasoning의 경우 여성이 평균적으로 높으며

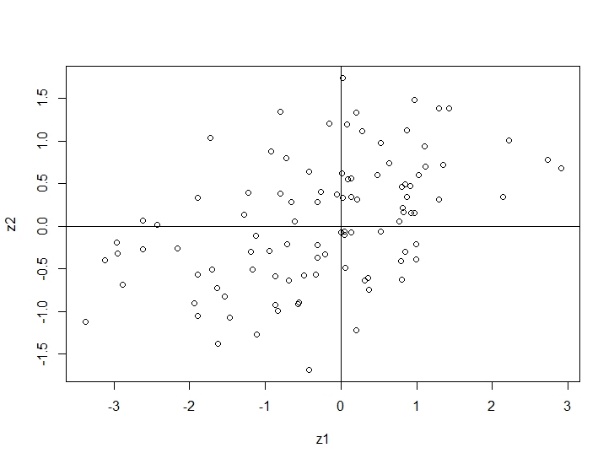
Geometric Reasoning은 남성의 평균이 더 높은 것으로 나타난다.

분산의 경우, Language Aptitude와 Geometric Reasoning에서는 남성이 더 높으며,

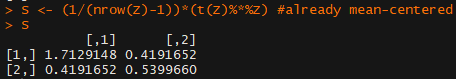
Analogical Reasoning의 경우 여성이 더 높다.

**2.5**

(a)



(b)



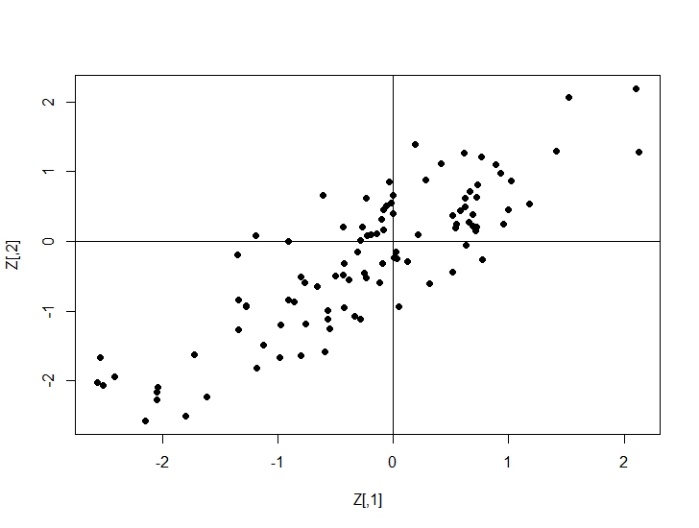
(c)



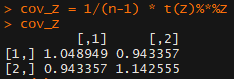
W의 determinant 값이 1에 가깝기 때문에 거의 동일함을 알 수 있다.

**2.6**

(a)



(b)



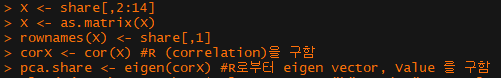
(c)



W의 determinant 값이 0.5 정도이기 때문에 cov(Z)의 determinant 값이 훨씬 줄어들었음을 알 수 있다.

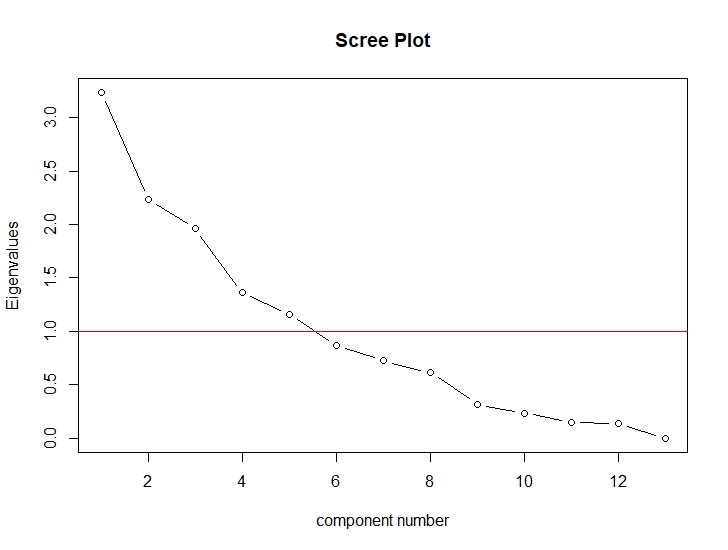
**4.3**

(a)



Eigen Value = 차지하는 분산





분산이 1보다 큰 5개의 Componet를 사용하는 것이 적절해보인다.

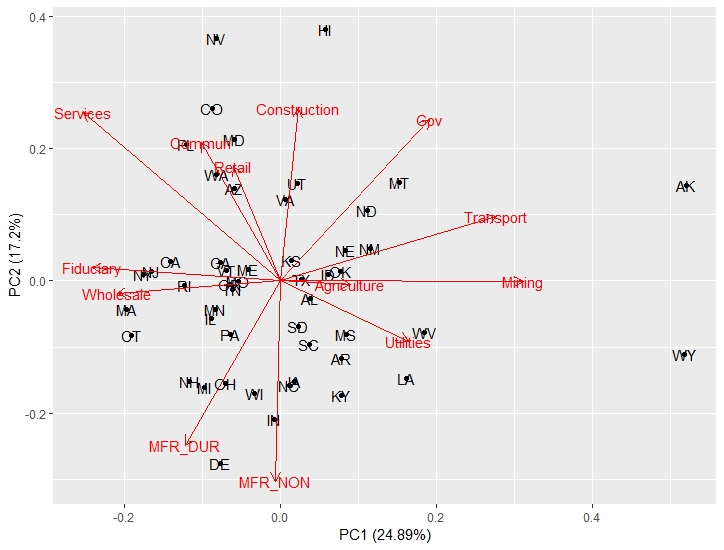


5개를 사용했을 때의 설명력은 77%정도임.

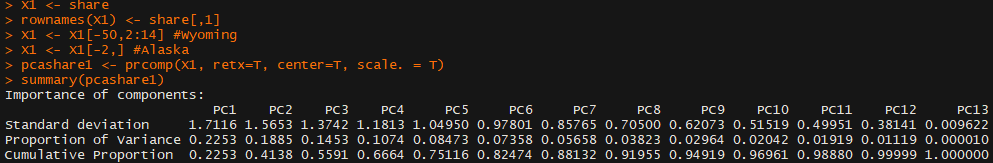


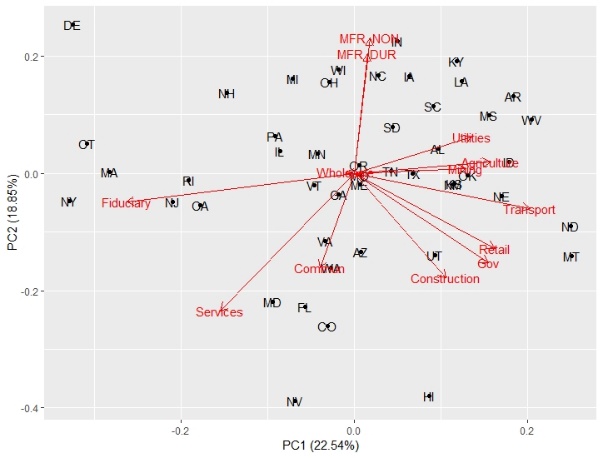
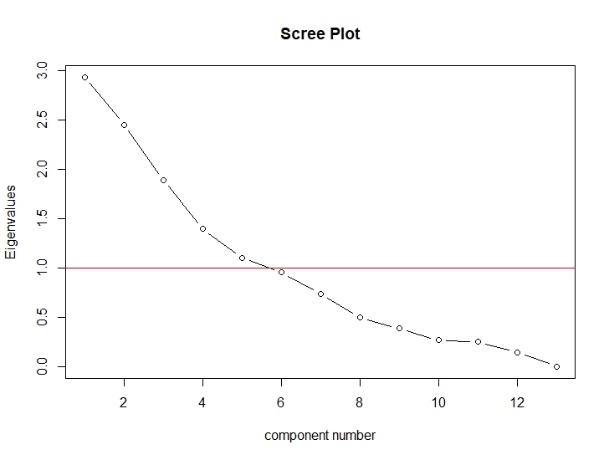
Loadings =



  
Score plot을 그렸을 때, AK와 WY가 크게 벗어난 것을 볼 수 있음 즉 일종의 Outlier로 볼 수 있음.

(b)



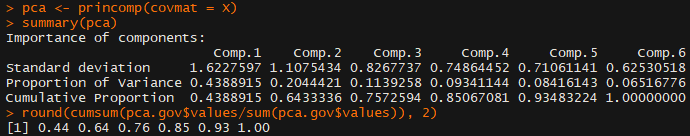


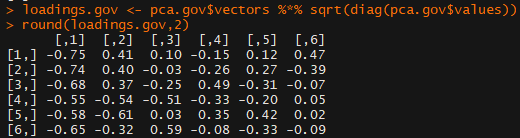
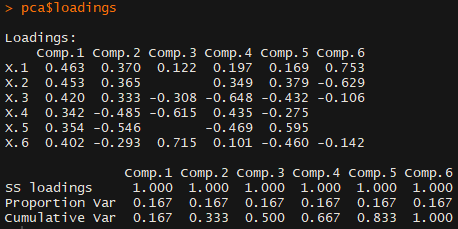
Outlier로 분류한 AK, WY를 제거하고 분석함.

역시나 분산이 1 이상인 5개 정도의 Factor를 사용하는 것이 적당하며, 이때 75%정도의 설명력을 지닌다. Score plot의 데이터들이 원점을 중심으로 고루 퍼져있음을 볼 수 있다.

PC1과 Agriculture, Mining, Construction, Transport, Utilities, retail, Gov는 양의 상관 관계를, Fiduciary, Service는 음의 상관 관계를 보인다. 또한 PC2는 Mfr\_dur, Mfr\_non과 양의 상관 관계를 and Construc, Commun, Retail, service, Gov과 음의 상관관계를 보인다.

**4.4.**

****

****

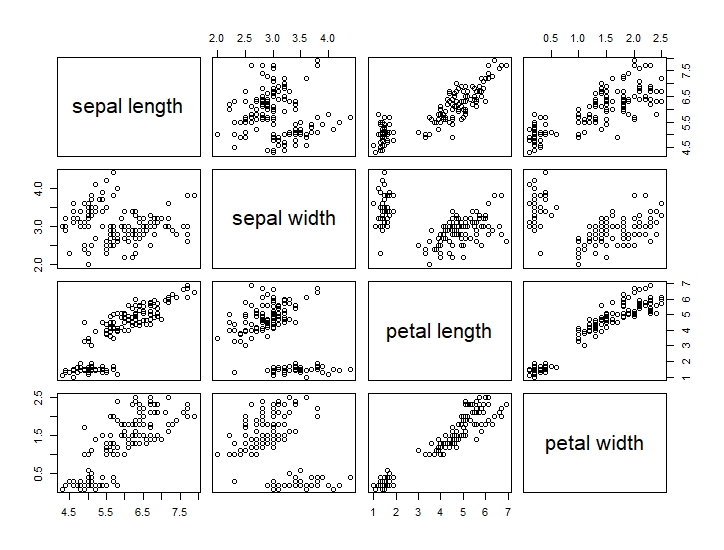
첫 두 Component(PC1 44%, PC2 20%)에 의해서는 64%정도의 variation이 설명된다.

첫 component는 X1~X6의 정책에 모두 같은 방향의 설명력을 보인다.

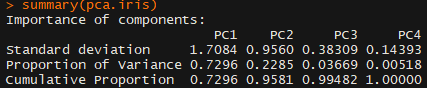
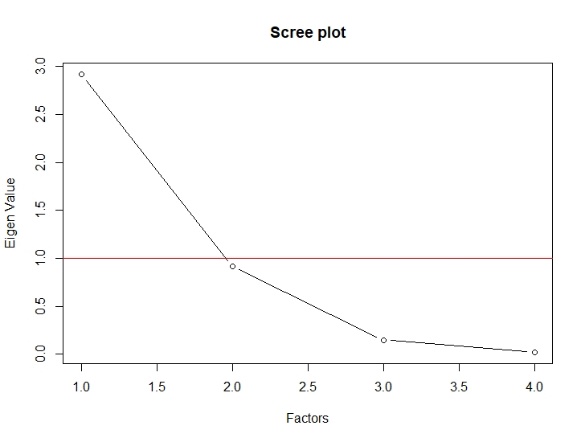
두 번째 component는 (X1, X2, X3), (X4, X5, X6) 간에 반대 방향의 설명력을 보여준다.

**4.5**

(a)

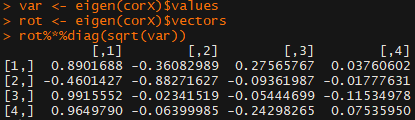


몇몇 변수들 간에 상관관계가 보인다.



첫 번째 component만이 variation이 1이 넘고, 72%이상의 설명력을 지닌다. 즉 이것만으로도 데이터를 설명할 수 있다.

또는 두 번째 component도 variation이 1에 근접하므로 포함해도 괜찮다고 생각한다.



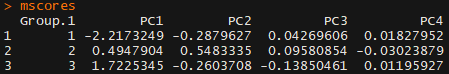
Loading matrix를 보았을 때,

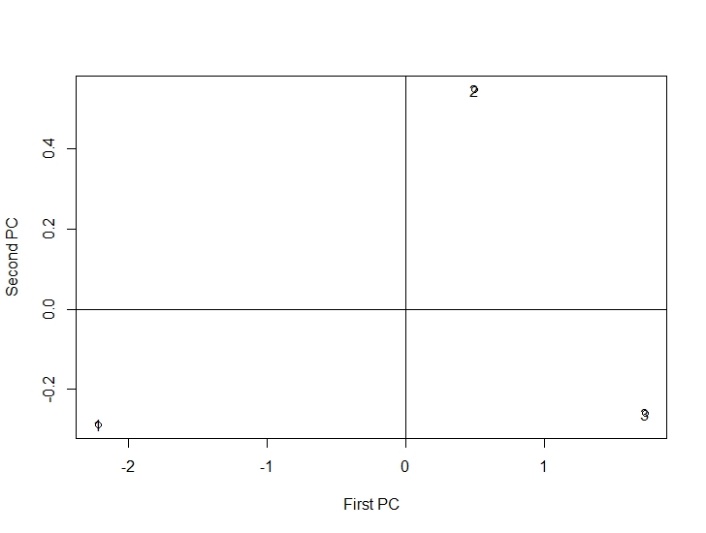
첫 번째 component는 X2, X4, X5에 대하여 강한 설명력을 나타내며

두 번째 component는 X3와 강한 설명력을 나타낸다.

(b)



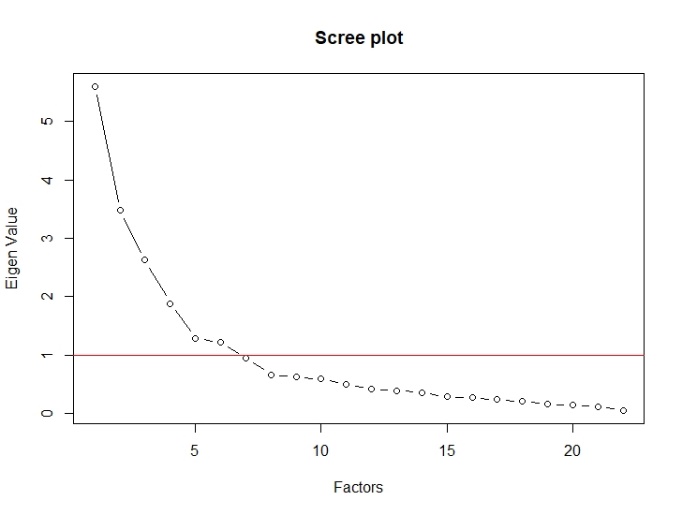
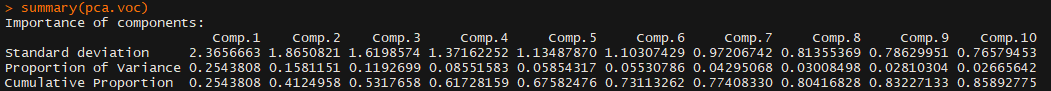




1(Iris Setosa)와 3(Iris Virginica)는 2번째 component와는 별 관련이 없지만, 1번째 component에 따라 차이가 나뉜다.

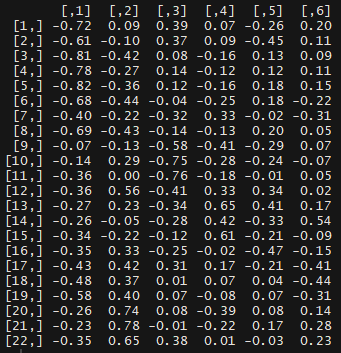
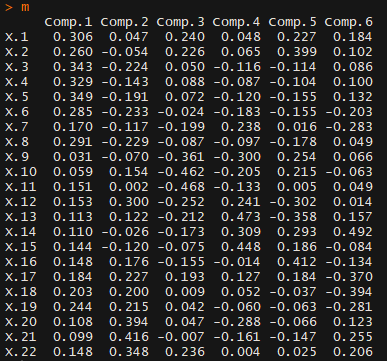
2(Iris Versicolor)는 1번째 component의 값은 거의 0으로 상관이 없지만, 2번째 component에 크게 영향을 받고 있다.

**4.7**



Variation이 1이 넘는 6개의 Components들을 선택하면 적절할 것으로 보인다.



Component 별로,

PC1 : X1~X6, X8 / 말하기, 상업적 활동

PC2 : X12, X20~X22 / 예술적 활동

PC3 : X9~X12 / 수학, 공학적 활동

PC4 : X13~X15 / 모험적인 활동

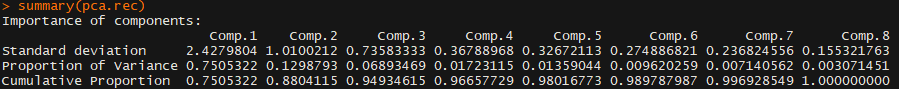
PC5 : X16/ 의료 관련 활동

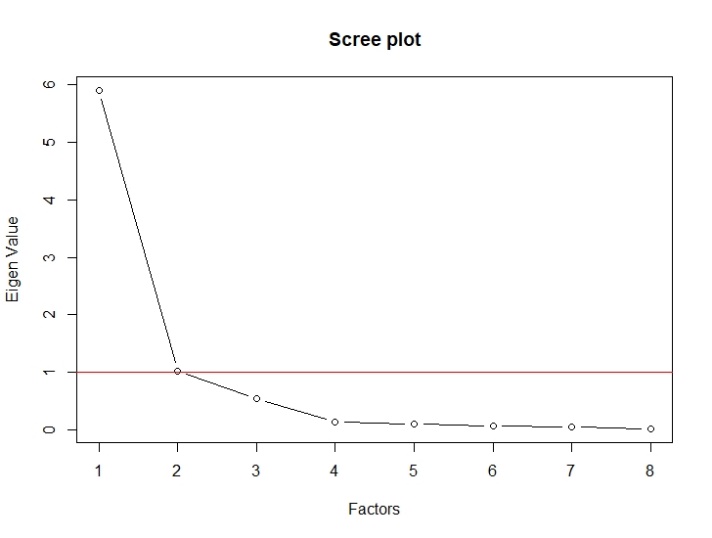
PC6 : X7, X14 X17~X19 / 군사, 사회, 봉사적 활동

에 대하여 높은 설명력을 보여주고 있다.

**4.8**

(a)



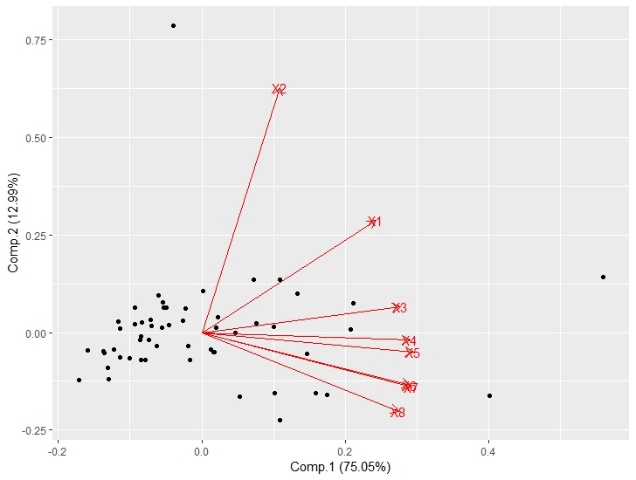
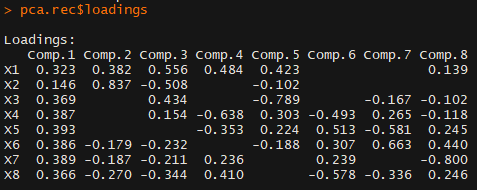


2개의 component로 88%에 육박하는 variation이 설명된다.

(b)

2개의 component면 된다.

(c)

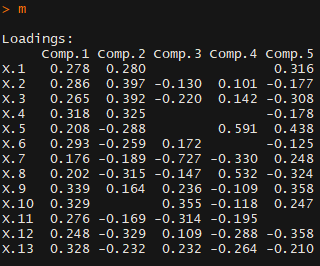
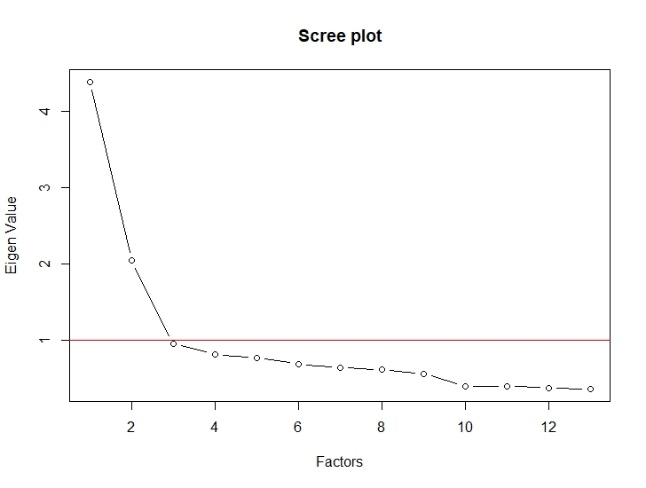


Variation이 적어도 1이 넘어가는 2번째 Components 까지를 선택하겠다.

Component1은 모든 변수와 같은 방향의 상관관계를 가지며, 전체적인 성적 자체를 설명한다.

Component2는 100m, 200m와 상대적으로 큰 상관관계를 가지며, 단거리 성적을 이야기 할 수 있다.

**4.9**



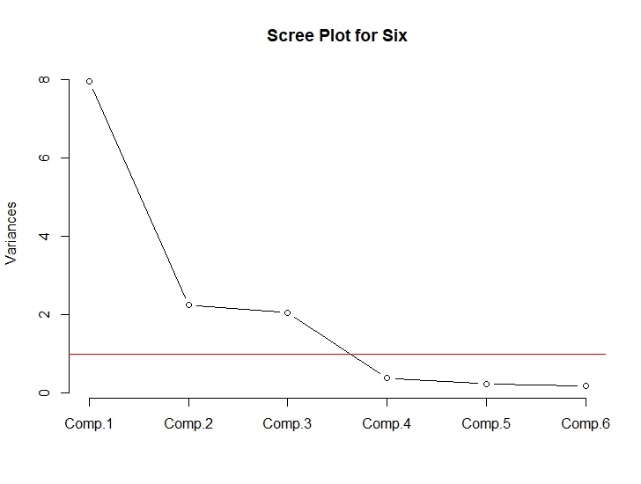
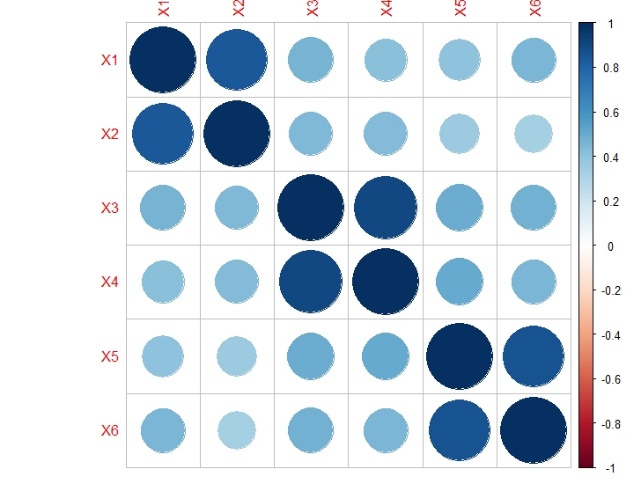
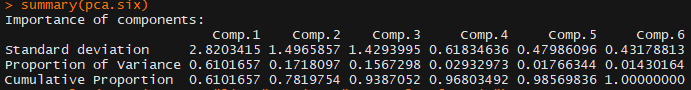
분석을 통하여, Variation이 1을 넘어가는 2번째 Component까지를 사용한다.

Component1는 모든 약물과 같은 방향의 상관관계를 가지고 있다.

Component2는 (술, 담배) (마약류) 간에 반대의 상관관계를 가진다.

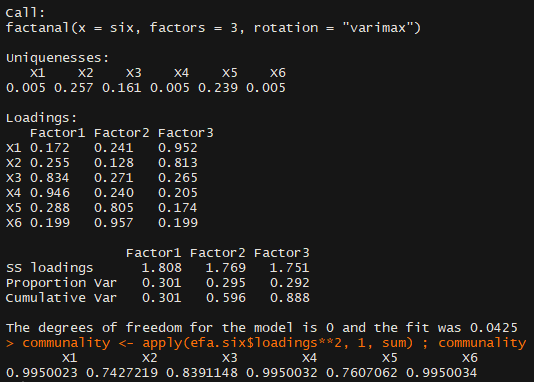
**5.4**

(a)



PCA에서 Variation이 1보다 큰 3개를 추출한다. Correlation을 보아도 크게 3뭉치로 묶인 것을 볼 수 있다.

(b)



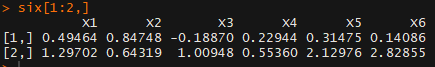
Factor1은 30.1%, Factor2는 29.5%, Factor3는 29.2%를 설명하며, 총 89%정도가 새로운 3가지 Factor에 의해 설명된다.

(c)

Factor1은 X3, X4 / Factor2는 X5, X6 / Factor3는 X1, X2를 대표하는 것이라고 할 수 있다.

(d)



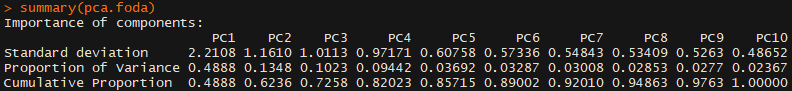
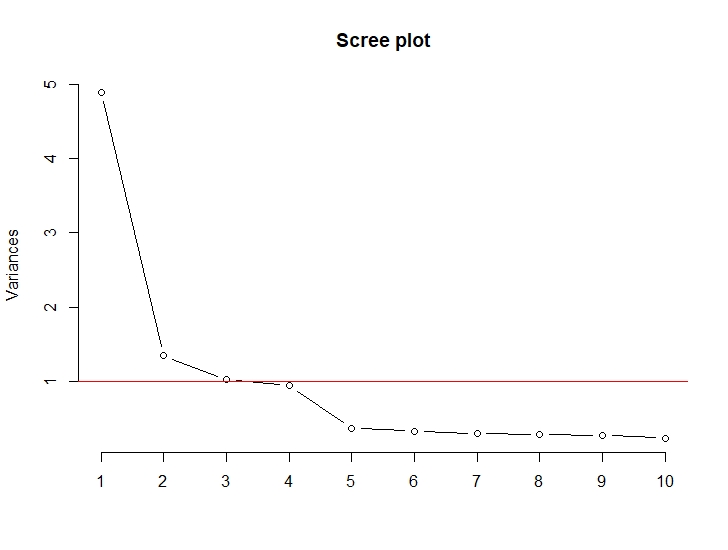


Obs1은 Factor3에 대해 높은 score를 갖고 있으며, X1, X2 변수에 대해 높은 값을 가지고 있다.

Obs2는 Factor2에 대해 높은 score를 갖고 있으며, X5, X6 변수에 대해 높은 값을 가지고 있다.

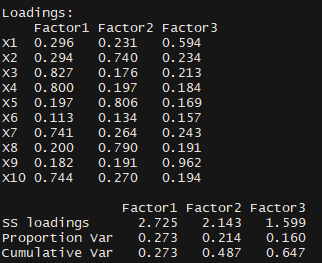
**5.6**

(a)

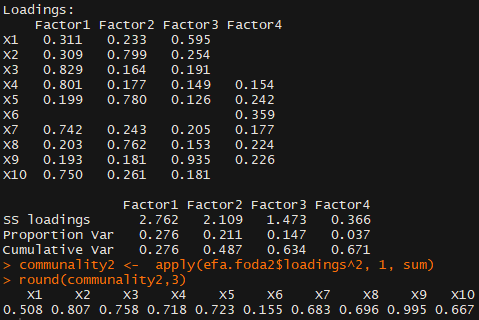


PCA를 통하여 variation이 1보다 큰 3개 정도의 Component를 이용하는 것이 좋다고 판단한다

변수의 설명력을 살릴 수 있는 EFA를 이용한다.

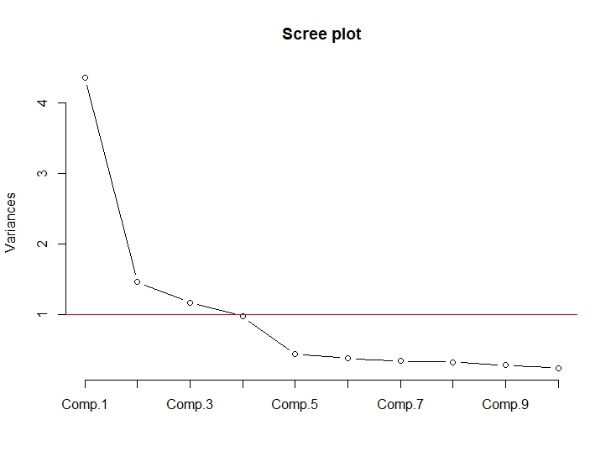
 

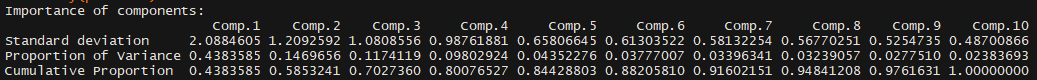
Communality를 통해 Factor 3개를 이용할 시에는 나머지 변수들은 잘 설명 되지만, X6이 제대로 설명이 되지 않는 모습을 보인다. 그래서 Factor를 4개로 늘려서 분석한다.



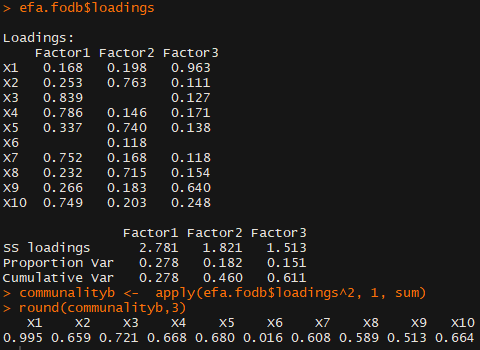
Factor를 4개로 늘릴 시에, X6의 communality가 여전히 낮긴 하지만 개선이 되었다. EFA를 사용하여서 3개 또는 4개의 Factor를 사용하도록 하는 것이 좋아보인다.

(b)





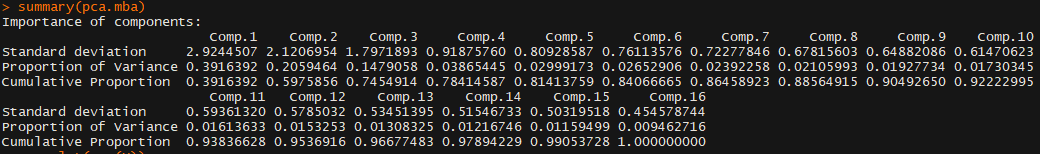
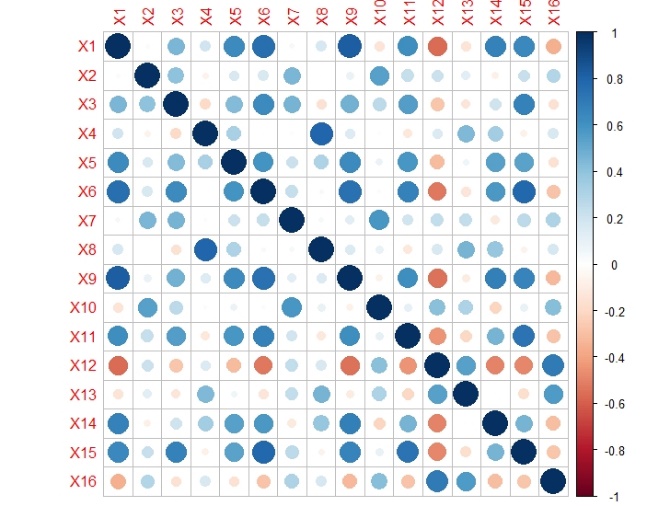
A 데이터와 마찬가지로 3개의 Component가 variation이 1 이상이므로 EFA에서 3개의 Factor를 사용한다.



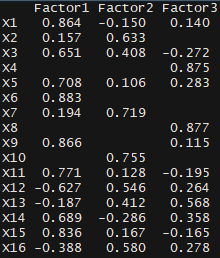
A와 마찬가지로 X6이 설명이 잘 안되고 있다. 그리고 변수 X1과 X9의 Communality가 바뀌어서 나오는데, 연구자가 두 변수를 바꾸어서 입력한 것은 아닐지 조언을 생각해보는 것도 좋겠다.

**5.7**

(a)



PCA를 통하여 Variation이 1 이상인 Component가 3개임을 알 수 있으며, 3~4개 정도의 factor를 사용하도록 한다.



전체적으로 나쁘지 않은 Communality를 보여준다.

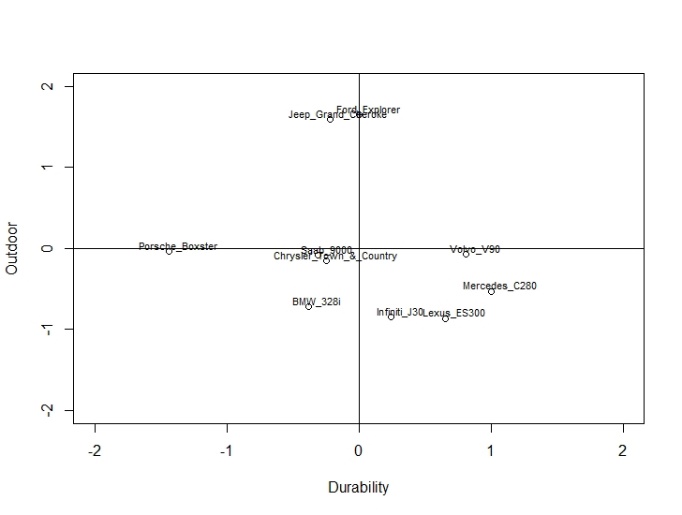
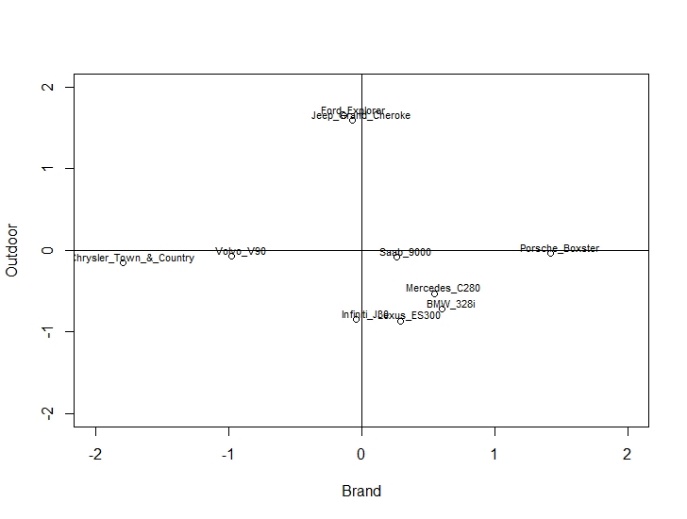
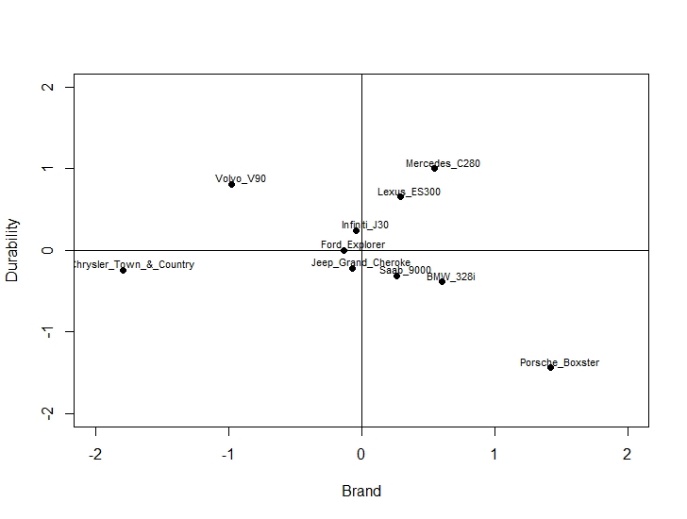
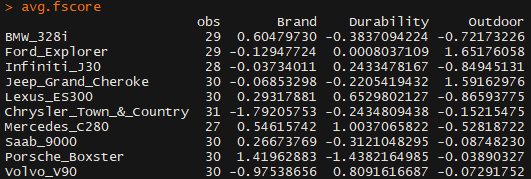
Factor1은 X1(Exciting), X3(Luxurious), X5(Powerful), X6(Stylish), X9(Fun), X11(Performance),

-X12(Family), X14(Sports), X15(Status)와 큰 관련이 있는 브랜드 이미지와 관련한 변수이다.

Factor2는 X2(Dependable), X7(Comfortable), X10(Safe), X12(Family), X16(Practical)과 큰 관련이 있으며 안정성과 관련이 있는 변수이다.

Factor3는 X4(Outdoorsy), X8(Rugged), X13(Versatile)과 큰 관련이 있으며, 야외활동과 관련한 변수이다.

(b)



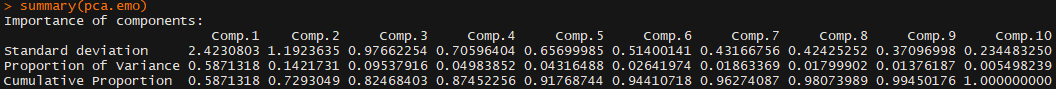
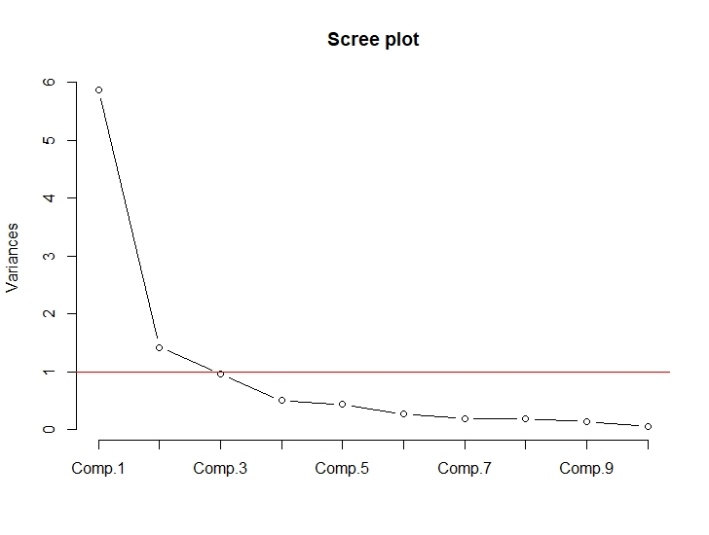
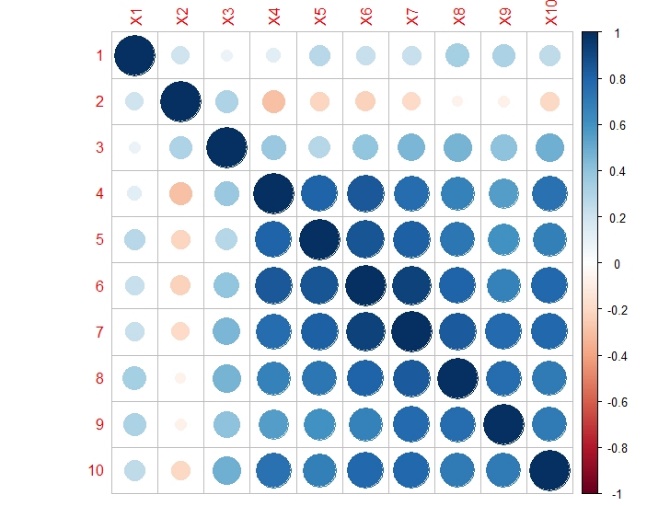
Brand에 있어서는 Porsche Boxster가 가장 독보적이지만 Durability에서는 큰 감점을 받고 있다. 반대로, Chrysler의 모델은 다른 부분은 무난하지만, Brand에 있어서 큰 감점을 받고 있어서 이미지 회복이 시급해 보인다.

많은 회사들이 Durability에 있어서 소비자들에게 큰 감흥을 주지는 못하고 있으나, Volvo와 Mercedes가 상대적으로 강점을 가지고 있다.

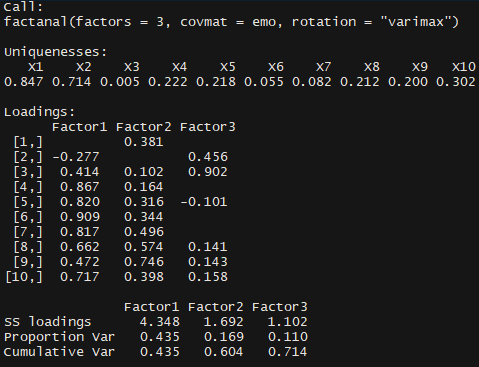
Ford Explorer와 Jeep Grand Cheroke 모델은 Outdoor에 있어서 큰 강점을 가지고 있다.

**5.8**

(a)

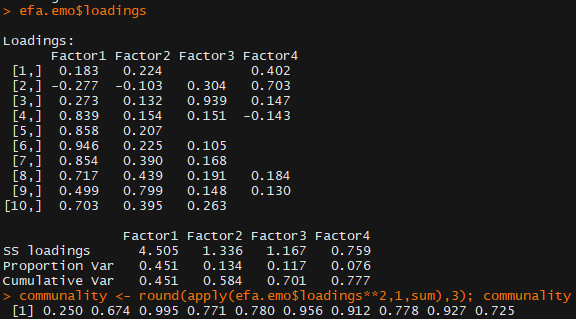
Correlation을 보면 X1, X2는 독립적인 편이고 X3부터는 Correlation이 점차 높아진다.

PCA를 통하여 보면 3번째 Component까지는 variation이 1에 근접하는 것을 볼 수 있다.





3개의 Factor는 X1과 X2에 대하여 좋지 않은 Communality를 보여준다.



Factor를 4개로 늘린 결과, X1은 여전히 Communality가 좋지 않지만, X2의 Communality는 크게 향상되었다.

Factor1은 X4(Sadness), X5(Anger), X6(Disgust), X7(Contempt), X8(Fear), X9(Shame), X10(Guilt)와 관련이 있는 부정적인 감정을 나타낸다.

Factor2는 X8(Fear), X9(Shame)과 관련이 있는 부끄러운 감정을 나타낸다.

Factor3는 X3(Surprise)와 관련되어 놀라움의 감정을 나타낸다.

Factor4는 X1(Interest), X2(Joy)와 관련되어 즐거움의 감정을 나타내어준다.