1. Factor Analysis

Car 데이터는 10개의 자동차에 대한 학생들의 인식을 조사한 데이터이다. 16개의 항목(Exciting, Dependable, Luxurious, ..., Practical)에 대해 각 차의 느낌을 1부터 5까지 평가하였다. 이 16개의 항목들 사이에 포함되어 있는 잠재적인 공통 요인을 뽑아내기 위해 Factor Analysis(FA)를 진행한다. Factor Analysis는 이러한 항목이 많은 설문조사에서 공통 요인이 있는 것끼리 항목들을 묶어 항목 수를 줄일 때 자주 사용되는 기법이다.

먼저 요인의 개수를 정한다. [그림1]는 eigenvalues가 기록되어 있는 Scree Plot으로 요인의 개수를 결정하는 데 쓰인다. 보통 이 Scree Plot이 꺾이는 지점인 elbow 부분에서 요인의 개수를 정한다. 이 경우 elbow 지점이 4이므로 4개의 요인으로 FA를 할 수 있다. 그런데 네 번째 eigenvalue를 보면 너무 작은 값을 갖는다. Eigenvalue를 중요도라고도 볼 수 있기 때문에 작은 eigenvalue는 중요도가 낮다는 것이다. 따라서 elbow 지점은 4에 위치하지만 요인의 개수를 3개로 정한다.

이후 FA를 진행한다. 이때 모두 표준화한 변수들을 사용한다. R에서 함수로 제공하고 있는 방법은 Maximum Likelihood Method 방법이므로, 이 방법으로 FA를 하였다. 이때 해석의 편의성을 위해서 Varimax 방법으로 rotation을 한 후 factor loadings를 구하였다. Communality를 보면 Dependable을 제외하면 모두 0.5 이상의 큰 값을 가지고 있다. 즉 요인 분석이 적합했다는 것을 의미한다. [그림2]는 factor loadings을 나타낸 표이다. Factor 1 축을 보면 Status, Performance, Powerful, Stylish, Fun, Exciting, Sporty가 1에 가까운 값을 가진다. Luxurious도 0.5 정도의 값을 갖는다. 그리고 Family와 Practical은 음수 값을 가진다. 따라서 Factor 1을 '멋있고 기능이 좋은 차'라고 해석할 수 있다. 즉 가족이 타기에 좋고 실용적인 차라고 하기에는 어렵지만, 멋있고 기능이 좋은 차라는 것이다. Factor 2 축을 보면 Safe, Comfortable, Dependable이 1에 가까운 값을 갖는다. 그리고 Family와 Practical이 0.5 정도의 값을 갖는다. 나머지 변수들은 대부분 0에 가까운 값을 가진다. 따라서 Factor 2는 '가족을 위한 안전하고 편리한 차'라고 해석할 수 있다. Factor 3 축을 보면 Outdoorsy과 Rugged가 1에 매우 가까운 값을 가진다. 그리고 Versatile을 제외하면 나머지 변수들이 모두 0에 가까운 값을 가진다. 따라서 Factor 3은 'outdoor activity를 위한 차'라고 해석할 수 있다.

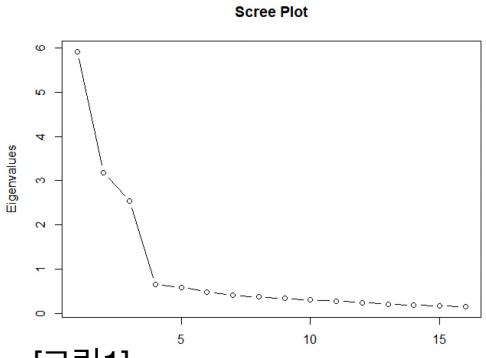
따라서 16개의 항목들을 총 3개의 항목으로 나누어도 좋을 것 같다. 이렇게 항목을 줄이면 다중공산성 문제도 해결할 수 있어 분석의 질이 높아질 수 있다.

2. Clustering

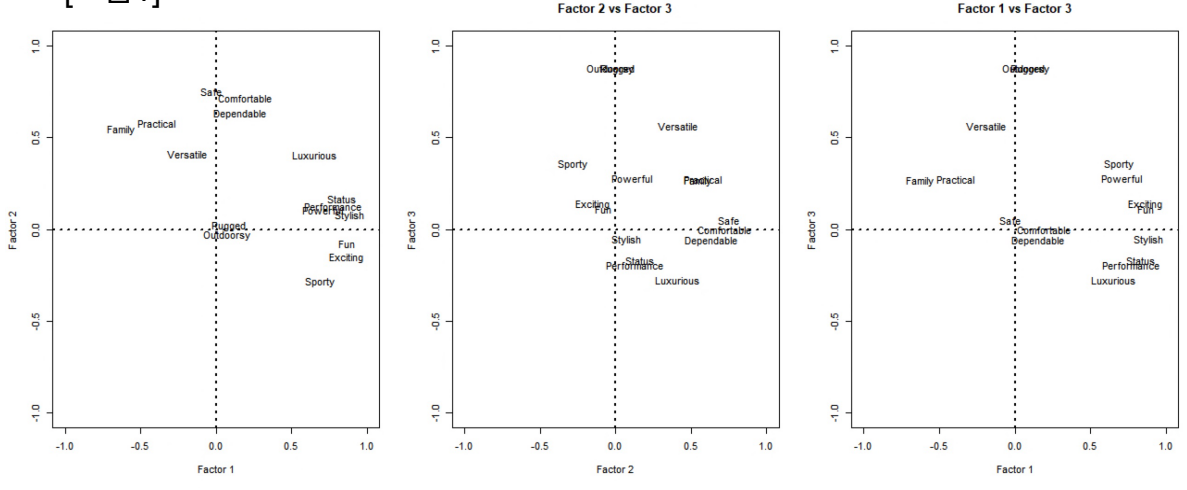
같은 데이터로 clustering을 진행하였다. Clustering이란 서로 다른 성질들을 가진 요소들로 이루어진 집단에서 공통적인 특성을 지닌 요소들끼리 cluster를 이루게 하여 분류하는 것이다.

먼저 계층적 군집분석 중 하나인 Ward's method를 통해 clustering을 했다. 이 방법은 두 개의 cluster를 합할 때 집단 내의 분산과 집단 간의 분산의 비율을 최대화하도록 cluster를 구성하는 방법이다. 이 방법은 분류 결과가 좋으며 가장 명확한 cluster가 만들어질 확률이 높다고 알려져 있기 때문에 이 방법을 택했다. 먼저 cluster의 개수를 구한다. [그림3]은 cluster의 개수 별로 total within-cluster sum of squares를 나타낸 그래프이다. 이 그래프에서 elbow 지점을 cluster의 개수로 정한다. [그림3]에서 elbow 지점은 4이다. 따라서 4개의 cluster를 만든다. [그림4]는 dendogram이다. 모든 데이터를 가지고 dendogram을 만들면 그래프가 겹치고 보기가 어렵기 때문에 sampling을 해서 일부의 데이터만을 가지고 dendogram을 만들었다. 4개의 cluster가 만들어지도록 하는 cutoff 지점을 알 수 있다. 이 경우 144가 cutoff 지점이다. 그리고 [그림5]에서 cluster를 표시하기 위해 PCA를 하였고 그룹 별로 색을 다르게 하여 scatter plot을 그렸다. 그리고 그 점이 어떤 car_id에 해당하는 점인지 나타내기 위해 car_id를 표시하였다. Group 1의 car_id는 대부분 1, 3, 5, 7이고 group 2의 car_id는 대부분 6이다. Group 3의 car_id는 대부분 9이고 group 4의 car_id는 2, 4이다. 공통적인 특성을 지닌 요소들끼리 cluster가 잘 된 것을 확인할 수 있다. 뿐만 아니라 car_id 2와 4인 자동차가 유사하며 car_id가 1, 3, 5, 7인 자동차들끼리 유사하다는 것을 알 수 있다.

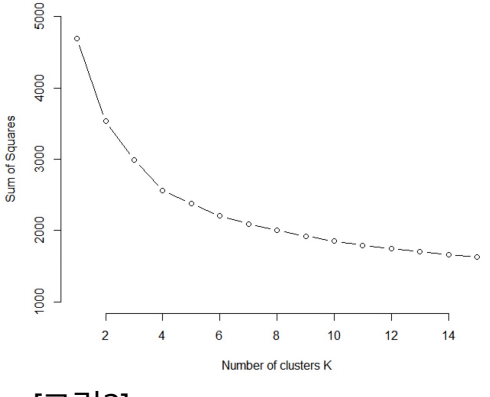
그런데 분석 대상의 자료가 많을 경우 계층적 군집분석은 계산상의 어려움이 존재한다. 따라서 비계층적 군집분석인 K-Means method를 쓴다. [그림6]의 세 그래프는 각각 3개의 그룹으로 나누었을 때, 4개의 그룹으로 나누었을 때, 5개의 그룹으로 나누었을 때의 그래프이다. 4개의 그룹으로 나누었을 때 그래프를 아까 Ward's method로 clustering한 그래프와 비교해보면 가운데 부분만 다르고 양 끝에 있는 cluster은 결과가 비슷하게 나온 것을 확인할 수 있다. [그림6]의 세 그래프 중 첫 번째 그래프가 가장 겹치는 부분이 덜하고 clustering이 잘 된 것 같다. 따라서 이번에는 cluster의 개수를 3개로 하고 K-Means method로 clustering을 한다. [그림7]이 그 결과를 나타내며 각 점이 어떤 car_id에 해당하는 것인지를 표시해주었다. Group 1은 9, 1, 8, 5, 7이 많다. Group 2는 대부분 6이다. 종종 3과 10도 보인다. 그리고 group 3은 2, 4가 많다. 공통적인 특성을 지닌 요소들끼리 cluster가 잘 된 것을 확인할 수 있다. 이 경우에는 car_id 2와 4이 유사하며 1, 9, 8, 5, 7이 유사하다는 것을 알 수 있다. 위의 Wald's method의 결과와 비교해보면 비슷하다.



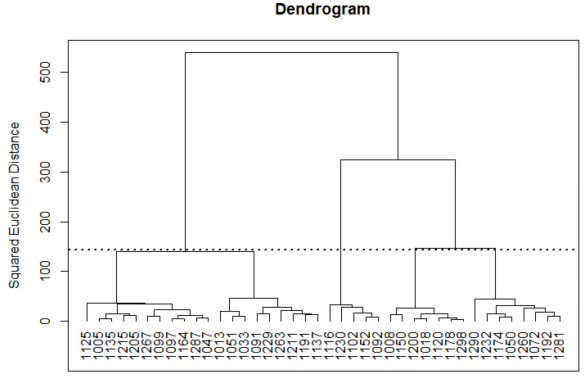
[그림1]



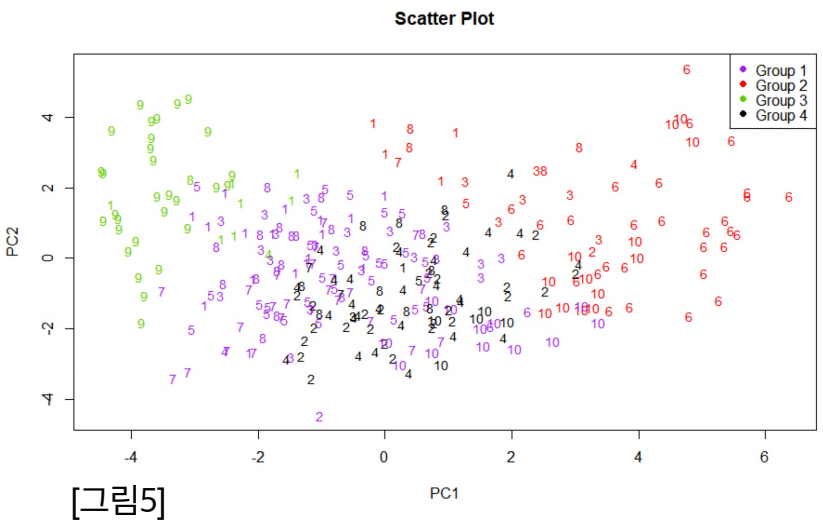
[그림2]



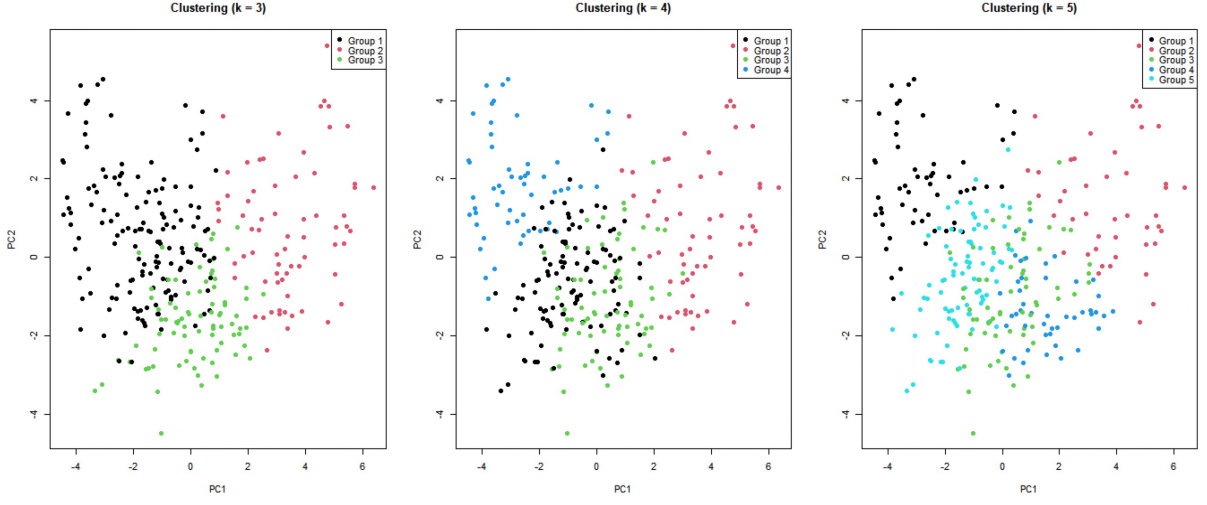
[그림3]



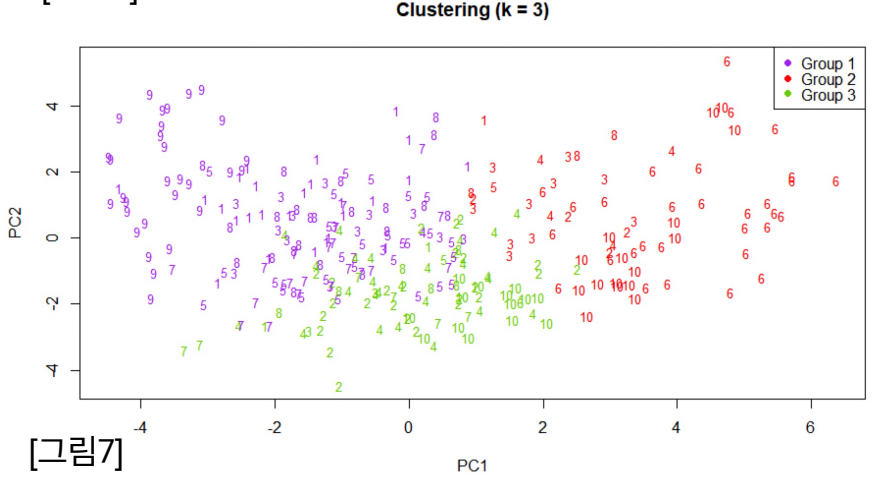
[그림4]


Scatter Plot
Group 1
Group 2
Group 3
Group 4
PC2
PC1

[그림5]


Clustering (k = 3)
Group 1
Group 2
Group 3
Clustering (k = 4)
Group 1
Group 2
Group 3
Group 4
Clustering (k = 5)
Group 1
Group 2
Group 3
Group 4
Group 5
PC1
PC2

[그림6]


Clustering (k = 3)
Group 1
Group 2
Group 3
PC2
PC1

[그림7]