Multivariate Data Analysis Assignment #8

Dataset: College Dataset

해당 데이터셋은 미국 각 대학들의 신입생 선발에 대한 정보를 포함하는 데이터셋이다. 총 18개의 변수로 구성되어 있으며 아래 스크립트를 실행하여 데이터셋을 불러올 수 있다.

```
install.packages("ISLR")
library(ISLR)
data(College)
View(College)
```

이 중에서 첫 번째 column인 Private을 제외한 총 17개의 수치형 변수에 대해서 군집화를 수행하고 그 결과를 해석하고자 한다.

[K-Means Clustering]

K-Means clustering은 각 변수의 값의 범위에 영향을 받을 수 있으므로 scale( )함수를 사용하여 모든 변수의 평균을 0, 표준편차를 1로 만드는 정규화를 수행하시오.

[Q1-1] clValid( ) 함수를 사용하여 K-Means Clustering의 군집 수를 2개부터 10개까지 증가시켜 가면서 internal 및 stability 관련 타당성 지표 값들을 산출하시오. 총 소요 시간은 얼마인가? Dunn index와 Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

[Q1-2] K=3으로 군집화를 10회 반복 수행하고 회차마다 각 군집의 Center와 Size를 확인해보시오. 10회 반복 수행 시 모두 동일한 군집이 생성되었는가?

[Q1-3] K=10으로 군집화를 10회 반복 수행하고 회차마다 각 군집의 Center와 Size를 확인해보시오. 10회 반복 시 모두 동일한 군집이 생성되었는가?

[Q1-4] K=3으로 군집화를 수행한 뒤, 각 변수들에 대해 정규화 이후의 값들을 이용하여 Radar chart를 도시하시오. 이 중에서 상대적으로 더 유사한 두 개의 군집 쌍과 가장 다른 두 개의 군집 쌍을 판별해 보시오.

[Q1-5] 세 개의 두 군집 조합(Cluster 1 vs. Cluster 2, Cluster 1 vs. Cluster 3, Cluster 2 vs. Cluster 3)에 대해서 각 변수별 차이의 유의성에 대해서 t-test를 수행하고 그 결과를 해석하시오.

[Q1-6] K-Means Clustering의 결과물을 시각화할 수 있는 방법을 웹에서 찾아 스스로 적용해보시오.

[Hierarchical Clustering]

College Dataset에 대해 scale( )함수를 사용하여 모든 변수의 평균을 0, 표준편차를 1로 만드는 정규화를 수행하시오.

[Q2-1] clValid( ) 함수를 사용하여 Hierarchical Clustering의 군집 수를 2개부터 10개까지 증가시켜 가면서 internal 및 stability 관련 타당성 지표 값들을 산출하시오. 총 소요 시간은 얼마인가? K-Means Clustering과 비교할 경우 소요 시간의 차이가 있는가? Dunn index와 Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

[Q2-2] Distance matrix는 실습 자료에서 제공하는 spearman correlation을 이용한 방식으로 산출한 뒤, hclust( ) 함수의 method 옵션을 다양하게 조절해 가면서 dendrogram을 그려보시오. 군집화를 수행한다고 할 때 군집의 크기가 가장 극단적으로 차이가 날 것으로 예상되는 옵션은 무엇인가?

[Q2-3] 생성된 Dendrogram으로부터 10개의 군집을 찾은 후 각 변수들에 대한 정규화 이후의 값들을 이용하여 Radar chart를 도시하시오. 본인이 판단하기에 가장 차이가 극명하게 나타나는 두 군집 조합과 가장 유사한 것으로 보이는 두 군집 조합을 선택한 뒤 각 조합에 대해서 변수별 차이의 통계적 유의성을 t-test를 통해 검증하시오.

[Q2-4] Hierarchical Clustering의 결과물을 heatmap을 사용하여 도시하는 방법을 웹에서 찾아 적용해보시오.

[DBSCAN]

실습에서 사용한 "Personal Loan.csv" 파일을 사용하여 다음 절차를 수행한 뒤 이후 물음에 답하시오. (1) 1번, 5번, 10번 column 제외, (2) 평균 = 0, 표준편차 = 1로 정규화.

[Q3-1] dbscan( )함수의 eps 옵션과 minPts 옵션에 대해 각각 최소 5개 이상의 서로 다른 값(총 25개 이상의 조합)을 사용하여 생성된 군집의 수와 어느 군집에도 할당되지 않은 Noise points 수를 아래 표와 같이 정리하시오.

eps	minPts	군집 수	Noise 수

[Q3-2] 최소 3개 이상의 군집이 판별된 eps/minPts 조합에 대하여 군집별 변수 Radar chart를 도시하고 각 군집의 특성을 파악해 보시오.

[Q3-3] 원래 데이터를 PCA를 이용하여 2차원으로 축소시킨 후 [Q3-2]에서 선택한 군집의 수를 이용하여 군집들과 Noise points를 2차원 평면에 도시해 보시오.

[Extra Question] 지금까지 제시된 문제 이외에도 군집화 방법론들에 대한 흥미로운 시각화/결과 해석 기법을 찾아보고 적용해보시오.