Multivariate Data Analysis Proposal

8조, 김지영, 김자영, 정석원, 임지원, 박정진

1. 개요

세상에는 많은 종류의 사람들이 존재하며, 각각은 다른 종류의 다양한 관심사를 가지고 있습니다. 사람들을 나누는 여러 가지 기준들이 존재하지만, 관심사의 묶음에 따라서도 사람들을 분류할 수 있다고 생각합니다. 저희는 20대 젊은 사람들을 관심사에 따라서 몇 가지 분류로 나누고 그렇게 분류된 사람들의 각 집합이 어떠한 특성을 지닌 사람들로 구성되어있는지 알아보고자 합니다.

1. 이용 데이터 셋

Kaggle의 Young People Survey 데이터 셋을 활용합니다. 1010명의 젊은 사람들에게 음악 선호, 영화 선호, 취미&관심사, 공포, 건강&습관, 개인적인 특성&가치관, 소비습관 등에 대한 질문으로 이루어져 있습니다. 답변은 모두 1-5의 정도로 나타나 있습니다.

Data (scale of 1-5)

- Music preferences (19 items)

- Movie preferences (12 items)

- Hobbies & interests (32 items)

- Phobias (10 items)

- Personality traits, views on life, & opinions (57 items)

- Spending habits (7 items)

- Demographics (4 items)

(<https://www.kaggle.com/miroslavsabo/young-people-survey#responses.csv>)

1. 진행 계획

* EDA를 통하여 응답자의 성별, 교육수준, 출신 도시 등의 분포를 확인, 결측값 제거 등 전처리
* Music Preferences, Movie preferences, Hobbies & interests, Phobias와 관련된 변수들을 PCA, EFA, CFA를 통하여 유의미한 변수들로 묶어냄
* 줄어든 변수들의 특성을 파악
* 변수들을 통하여 Hierarchical, KNN, K-means clustering으로 응답자들을 분류
* 각 군집들이 어떠한 특성을 갖고 있는지 파악
* 간 군집에 소속되어있는 사람들이 어떠한 특성을 갖고 있는지 줄어든 변수들을 이용하여 파악
* Personality traits, views on life, & opinions, Demographics의 변수들을 통하여 추가적인 설명
* 시각화