군집화 실습 - 고객 세그멘테이션
고객 세그멘테이션은 다양한 기준으로 고객을 분류하는 기법을 지칭함. 고객 세그먼테이션은 CRM이나 마케팅의 중요 기반 요소임.

고객을 분류하는 요소는 여러 가지가 있음. 개인의 신상 데이터가 이를 위해 사용될 수도 있지만, 고객 분류가 사용되는 대부분의 비즈니스가 상품 판매에 중점을 두고 있기 때문에 더 중요한 분류 요소는 어ㅍ너 상품을 얼마나 많은 비용을 써서 자주 사용하는가에 기반한 정보를 분류하는 것이 보통임.

고객 세그멘테이션의 주요 목표 : 타깃 마케팅(고객을 여러 특성에 맞게 세분화해서 그 유형에 따라 맞춤형 마케팅이나 서비스를 제공하는 것)

[RMF 기법]
Recency(가장 최근 상품 구입일에서 오늘까지의 시간), Frequency(상품 구매 횟수), Monetary Value(총 구매 금액)

크게 왜곡된 데이터 세트의 도출은 굳이 군집화를 이용하지 않고도 간단한 데이터 분석만으로도 충분히 가능함. -> 지나치게 왜곡된 데이터 세트는 K-평균과 같은 거리 기반 군집화 알고리즘에서 지나치게 일반적인 군집화 결과를 도출하게 됨.

비지도학습 알고리즘의 하나인 군집화의 기능적 의미는 숨어 있는 새로운 집단을 발견하는 것임. 새로운 군집 내의 데이터 값을 분석하고 이해함으로써 이 집단에 새로운 의미르 ㄹ부여할 수 있음. 이를 통해 전체 데이터를 다른 각도로 바라볼 수 있게 만들어줌.

데이터 세트의 왜곡 정도를 낮추기 위해 자주 사용되는 방법 -> 데이터 값에 로그를 적용하는 로그 변환.

실루엣 스코어는 절대치가 중요한 게 아님. 어떻게 개별 군집이 더 균일하게 나뉠 수 있는지가 더 중요함!!

Online Retail K-Means & Hierarchical Clustering

K-Means Clustering

[algorithm works as follows :]
1. 무작위로 평균이라고 불리는 k개의 점을 초기에 설정한다.
2. 각 항목을 가장 가까운 평균으로 분류하고, 해당 평균으로 분류된 항목의 평균인 평균의 좌표를 업데이트한다.
3. 주어진 반복 횟수동안 이 과정을 반복하면 cluster가 만들어진다.

[최적의 군집 수 찾기]

- elbow curve
자율적인 알고리즘의 기본 단계는 클러스터가 클러스터링 될 수 있는 최적의 클러스터 수를 결정하는 것임. elbow 방법은 k의 죄적 값을 결정하는 가장 인기 있는 방법 중 하나임.

- Silhouette Analysis
Silhoutte score = (p-q)/max(p,q)
p : 데이터 점이 포함되지 않는 가장 가끗운 군집의 점까지의 평균 거리
q : 자체 군집의 모든 점에 대한 군집 내 평균 거리

범위 : -1에서 1사이.
점수가 1에 가까우면 데이터 점이 클러스터의 다른 점과 매우 유사하다는 것을 나타냄.
점수가 -1에 가까우면 데이터 점이 클러스터의 데이터 점과 유사하지 않음을 나타냄.

[Hierarchical Clustering]
계층적 군집화는 위에서 아래로 미리 결정된 순서를 가진 군집을 만드는 것을 포함함.
예를 들어, Disk의 모든 파일 및 폴더는 계층 구조로 구성됨. 계층적 클러스터링에는 두 가지 유형이 있음.
1. Divisive
2. Agglomerative.

[Single Linkage]
단일 연결 계층적 군집화에서 두 군집 사이의 거리는 각 군집의 두 점 사이의 최단 거리로 정의됨.

[Complete Linkage]
완전한 연결 계층적 군집화에서 두 군집 사이의 거리는 각 군집의 두 점 사이의 가장 긴 거리로 정의됨.

[Average Linkage]
평균 연결 계층적 군집화에서 두 군집 사이의 거리는 한 군집의 각 점에서 다른 군집의 모든 점 사이의 평균 거리로 정의됨.

Customer Segmentation : Clustering

[프로젝트 소개]
식료품 회사의 데이터베이스에서 고객 기록의 비지도 데이터 군집 수행.
고객 세분화는 각 클러스터의 고객 간 유사성을 반영하는 그룹으로 고객을 분리하는 관행임.
각 고객이 비즈니스에 미치는 중요성을 최적화하기 위해 고객을 부문별로 나눔.
고객의 고유한 요구와 행동에 따라 제품을 수정함. 또한 비즈니스에서 다양한 유형의 고객의 우려에 부응할 수 있도록 지원함.

[imensionality reduction 하는 이유]
feature의 수가 많을수록 해당 feature로 작업하기가 더 어려워짐. 이러한 feature 중 다수는 상관 관계가 있으므로 중복됨.

PCA : 이러한 데이터 세트의 차원을 줄이고 해석 가능성을 높이는 동시에 정보 손실을 최소화하는 기술임.
1. PCA를 통한 dimensionality 감소
2. 축소된 데이터 프레임 plotting

[Agglomerative clustering]
계층적 군집화 방법. 원하는 수의 군집이 달성될때까지 병합함.

<Steps involved in the Clustering>
1. 형성할 군집 수를 결정하는 Elbow 방법
2. 응집 군집화를 통한 군집화
3. 산점도를 통해 형성된 군집 분석