7.6 군집화 실습 - 고객 세그먼테이션

고객 세그먼테이션의 정의와 기법

- 정의 : 다양한 기준으로 고객을 분류하는 기법
  - CRM이나 마케팅의 중요 기반 요소
- 고객 분류 요소 : 지역/결혼 여부/성별/소득 등…
- 더 중요한 분류 요소 : 어떤 상품을 얼마나 많은 비용을 써서 얼마나 자주 사용하는가에 기반한 정보
- 주요 목표 : 타깃 마케팅
  - 고객을 여러 특성에 맞게 세분화하여 그 유형에 따라 맞춤형 마케팅이나 서비스를 제공하는 것
- RFM 기법
  - Recency(R) 가장 최근 상품 구입 일에서 오늘까지의 기간, Frequency(F) 상품 구매 횟수, Monetary Value(M)총 구매 금액을 합한 것

데이터 세트 로딩과 데이터 클렌징

- 엑셀 파일을 판다스의 read_excel() 함수를 이용해 DataFrame으로 로드함
- 데이터 세트 : 제품 주문 데이터 세트
  - Invoice(주문번호) + StockCode(제품코드)를 기반으로 주문량, 주문 일자, 제품 단가, 주문 고객 번호, 주문 고객 국가 등의 칼럼으로 구성됨
- CustomerID의 Null 값이 너무 많고, 다른 칼럼의 경우도 오류 데이터 존재 → 사전 정제 작업 필요
  - Null 데이터는 필요가 없기에 제거함.
  - 오류 데이터는 분석의 효율성을 위해 삭제함
- Country 칼럼 사전 정제
  - 주요 고객은 영국인데 이외 나라도 포함되어 있음 → 다른 국가 데이터 모두 제외
- RFM 기반 데이터 가공
  - ‘UnitPrice’ * ‘Quantity’ = 주문 금액 데이터
  - CustomerNo : float형을 int형으로 변경

    - Top-5 주문 건수와 주문 금액을 가진 고객 데이터 추출
    - 주문번호+상품코드 기준의 데이터 → RFM 데이터로 변경
        - 주문번호 기준의 데이터 → 개별 고객 기준 데이터로 Group by
        - groupby 호출해 반환된 DataFrameGroupby 객체에 agg() 이용
        - Frequency : ‘CustomerID’로 groupby()해서 ‘InvoiceNo’의 count() ago로 구함
        - Monetary value : ‘CustomerID’로 groupby()해서 ‘sale_amout’의 sum() agg로 구함
        - Recency : ‘CustomerID’로 groupby()해서 ‘InvoiceDate’ 칼럼의 max()로 고객별 가장 최근 주문 일자를 먼저 구함 → 가공 작업 별도로 수행
            - 오늘 날짜를 현재 날짜로 해서는 안되기 때문에 현재 날짜를 정해서 가장 최근 주문 일자를 뺀 데이터에서 일자 데이터만 추출해서 생성해야 함
- RFM 기반 고객 세그먼테이션
    - 온라인 판매 데이터 세트 : 소매업체의 대규모 주문 포함 → 왜곡된 데이터 분포도를 가지는지 칼럼별 히스토그램 확인 필요
        - Frequency, Monetary는 왜곡 정도가 매우 심함 → 왜곡 정도가 매우 높은 데이터 세트에 K-평균 군집을 적용하면 중심의 개수를 증가시켜도 변별력이 떨어지는 군집화가 수행됨
            - 먼저 데이터 세트를 StandardScaler로 평균과 표준편차를 재조정 → 군집 수가 3개 이상일 때부터 데이터 세트의 개수가 너무 작은 군집이 만들어짐
                - 이 소수의 데이터 세트가 특정 소매점의 대량 주문 구매 데이터 → 군집 수를 늘려봐야 이 군집만 지속적으로 분리됨 → 의미 없는 군집화 결과로 이어짐
    - 데이터 세트 왜곡을 줄이기 위한 방법 : 데이터 값에 로그를 적용하는 로그 변환
        - 데이터를 로그 변환한 뒤 K-평균 알고리즘 적용 → 더 균일하게 군집화 수행됨