문제1 : 온라인 소비자 분석(군집분석)

데이터 설명

해당 온라인 소매 영업점은 영국에 기반을 두고 등록된 비점포 매장입니다.

2010 년 1 월 12 일부터 2011 년 12월 9 일 사이에 발생하는 모든 거래를 포함하는 초국적 데이터 세트입니다. 이 회사는 주로 독특한 행사 선물을 판매합니다. 회사의 많은 고객이 도매상입니다.

사업 목표

우리는 온라인 거래 데이터 세트를 사용하여 클러스터링을 구축하고 회사가 목표로 삼아야하는 최상의 고객 세트를 선택할 것입니다.

주어진 데이터(OnlineRetail_after)는 아래 조건에 따라 소비자(CustomerID)를 기준으로 Recency, Frequency, Amount 변수에 대한 merge를 수행하여 데이터 프레임을 만든 것입니다.(시간이 되시면 최초 데이터(OnlineRetail_RAW)를 통해 데이터셋을 만들어 보시오)

- R (Recency): Number of days since last purchase
- F (Frequency): Number of tracsactions
- A (Amount): Total amount of transactions (revenue contributed)

hint : Quantity * UnitPrice

결과(이러한 형태의 데이터 프레임 출력)

CustomerID	*Amount*	*Frequency*	*Recency*
고객1	*X1*	*Y1*	*Z1*
고객2	*X2*	*Y2*	*Z2*
고객3	*X3*	*Y3*	*Z3*
고객4	*X4*	*Y4*	*Z4*
…	*…*	*…*	*…*

("시각화 필수")

1. EDA를 통해 소비자 특성 등 인사이트를 도출하시오.
2. 계층적 군집분석을 통해 고객 집단을 그룹핑하고 집단의 특성을 설명하시오.
   - random_state=123(=set.seed(123)
   - 비지도학습 : 'Amount', 'Frequency', 'Recency'를 변수로 사용
3. 비계층적 분석을 통해 고객 집단을 그룹핑하고 집단의 특성을 설명하시오.
   - random_state=123(=set.seed(123))
   - 비지도학습 : 'Amount', 'Frequency', 'Recency'를 변수로 사용
4. 최종적으로 고객에 대한 마케팅 전략을 수립하시오.
5. 추가적인 고객 분석을 하고자 할 때 어떤 방식이 좋을지를 논하시오.

문제2 : '금시세와 코로나19 바이러스는 관련이 있을까?'(시계열 분석)
(외생변수 : 미국 워싱턴 기준 코로나 19 확진자수, 종속변수 : 금 종가)

("시각화 필수")

'covid19_wc.csv'는 2020년 1월 21일부터 2022년 5월 13일까지의 코로나 미국 워싱턴주의 데이터셋이다. 'gold.csv'는 2000년부터 현재 2022년 5월말 까지의 금 시세 데이터셋이다.

1. 위 두 데이터셋을 불러와서 covid19_wc 데이터셋의 날짜 기준으로 merge(inner join)하고 ['date', ' cases', ' Close'] 3개의 변수를 갖는 데이터셋으로 만드시오.

2. 코로나 확진자수에 따라 금시세가 변하는지 회귀분석을 실시하시오.

3. 훈련 데이터와 테스트 데이터 셋을 80:20으로 분할하여 진행하시오.

3-1. 금시세를 예측하기 위해서 금시세(종가 기준) 한 개 변수를 가지고 시계열 모델을 적합하시오.
3-2. 금시세를 예측하기 위해서 금시세(종가 기준)와 외생변수(코로나 확진자수)를 통해 ARIMA-X 모형을 접하시오.
3-3. 위 두 모델에 대한 예측결과를 시각화 하시오
3-4. 위 도메델에 대한 잔차분석을 시행하시오.