**[문제 1~4]**

삼성 온라인쇼핑몰은 미국 아마존에 대항할 세계 최고의 온라인 쇼핑몰로 발돋움하기 위해 실력 있다는 Data Scientist를 총 집합시켜 데이터를 분석하기로 하였다.

삼성 온라인쇼핑몰로 초빙된 당신은 다음 데이터를 활용하여 분석을 하고자 한다.

다음은 데이터 모델링을 위해 주어지는 파일 목록이다.

**ds\_purchase\_log.csv** : 삼성 온라인쇼핑몰에서 8월 31일 하루 동안 발생한 구매 로그 데이터

cust\_no : 회원번호(상품을 구매한 회원번호)

item\_no : 상품번호(회원이 구매한 상품번호)

**ds\_customer\_mst.csv** : 회원 마스터로, 회원정보가 수정되면 기존 정보는 그대로 있고, 수정된 정보가 추가된다.

cust\_no : 회원번호

age\_gender : 회원 연령대, 성별(연령대: 20대, 30대, 40대, 50대 이상)(성별: M-남, F-여)

update\_dt : 수정일자

**ds\_product\_mst.csv** : 상품 마스터

item\_no : 상품번호

item\_ctg1 : 상품의 대카테고리

item\_ctg2 : 상품의 중카테고리

item\_amt : 매출(구매금액)

분석에 앞서, 다음의 전처리를 수행하여 base\_table을 만든다.

ds\_purchase\_log.csv 로그를 기준으로 ds\_customer\_mst.csv에서 회원 연령대, 성별(age\_gender) 정보를 가져오고, ds\_product\_mst.csv에서 카테고리(item\_ctg1, item\_ctg2)와 매출(item\_amt) 정보를 가져온다.

이 때, ds\_customer\_mst.csv에서 회원 연령대, 성별(age\_gender)를 가져올 때, 고객별 가장 최신의 정보(고객별 update\_dt가 가장 큰 정보)를 가져온다. 단, 동일 회원번호에 대해 update\_dt가 동일한데, age\_gender 컬럼의 값이 다른 경우, max(age\_gender)값으로 가져오시오.

1. (base\_table을 활용하여 분석을 수행하시오.) 성별에 따라 어떤 대카테고리에 더 많은 매출이 발생하는지 비교하고자 한다.

매출의 비교는 대카테고리별 인당 매출금액 기준으로 하며, 인당 매출금액은 매출 총 금액을 매출을 발생시킨 회원수를 나눈 값으로 한다. 남녀 간 비교를 위해 회원 연령대, 성별(age\_gender)을 활용하여 성별 컬럼을 만들고, 성별 대카테고리별 인당 구매금액을 산출하여, 남성 1인당 매출액 2위인 카테고리(A)와, 여성 1인당 매출액 2위인 카테고리(B)를 차례로 쓰시오. (10점)

(답안예시) 가전,생활용품

2. (base\_table 또는 ds\_purchase\_log.csv 을 활용하여 분석을 수행하시오.) 상위 20퍼센트의 회원이 일으킨 매출은 전체 매출의 몇 퍼센트를 차지하며, 상위 20퍼센트의 회원이 일으킨 매출은 하위 20퍼센트의 회원이 일으킨 매출의 몇 배인지 분석하고자 한다.

상위 20퍼센트와 하위 20퍼센트를 산정하는 방식은 다음과 같다.

(a) 전체 회원에 대해 회원별 구매금액에 대해 내림차순으로 순위를 매긴다. 순위는 동차발생시 동차로 등수를 주고 그 다음 순위는 동차 사람수만큼 더한 이후 그 다음 순위를 준다 (아래 예시 참조)

<순위 예시>

금액 순위

5300 1

5300 1

5300 1

5200 4

5200 4

(b) 상위 20퍼센트는 Rank <= N\*0.2 인 회원을 의미하며, 하위 20퍼센트는 Rank >= N\*0.8 인 회원을 의미한다. (여기서 N은 base\_table 또는 ds\_purchase\_log.csv에 구매이력이 존재하는 회원수, Rank는 위 (a)에서 만든 순위를 의미한다)

상위 20퍼센트의 회원이 일으킨 매출은 전체 매출의 몇 퍼센트(A)이며, 상위 20퍼센트의 회원이 일으킨 매출은 하위 20퍼센트의 회원이 일으킨 매출의 몇 배(B) 인지 차례대로 쓰시오. (10점)

- Hint : Brightics는 hive 구문 rank() over(order by) 활용

- A, B는 소수점 아래는 버리고, 정수형으로 기술하시오.

(답안예시) 15,2

3. (base\_table을 활용하여 분석을 수행하시오.) 회원별 구매금액이 연령대별로 차이가 있는지 비교하고자 회원 연령대, 성별(age\_gender)을 활용하여 연령대(age\_cls) 컬럼을 만들고 다음 분석을 수행한다.

각 회원별 구매한 총 구매금액(y)을 산출하고, 회원의 연령대 그룹을 매핑한다. 연령대 그룹(x)은 age\_cls가 '30대'이면 그룹이 ‘G1’, age\_cls가 '40대'이면 그룹이 ‘G2’, age\_cls가 그 외 연령대이면 그룹이 ‘G3’이다.

연령대 그룹(x)에 따라 회원별 구매한 총 구매금액(y)이 차이가 있는지 ANOVA를 수행하고, ANOVA의 F-통계량 추정치(A)를 구하시오. 또한, 사후검정으로 Duncan Test(alpha=0.05)를 수행하여 평균구매금액 순으로 오름차순하고, 그 결과를 등호, 부등호로 표기하시오. (15점)

- A는 소수점 아래는 버리고, 정수형으로 기술하시오.

- B는 회원별 구매한 총 구매금액(y)의 평균이 그룹 G1, G2, G3 순으로 크고, Duncan Test 결과가 1, 2가 같다면 G1=G2>G3으로 표기한다.

(답안예시) 3,G1=G2>G3

4. (base\_table을 활용하여 분석을 수행하시오.) 함께 구매되는 상품이 어떤지 분석하여 상품 상세 페이지에 ‘함께 구매하는 상품’으로 추천하기 위해, 구매 로그의 중카테고리 기준으로 연관성 분석(Association Rule)을 수행하고자 한다. 하나의 중카테고리만 구매한 이력이 있는 회원은 제외하고, 둘 이상의 중카테고리를 구매한 회원의 구매 이력을 기반으로 연관성 분석을 수행하고, 중카테고리가 ‘농산물’인 상품의 상세 페이지에 함께 노출되면 좋을 Lift 기준의 가장 적합한 중카테고리를 구하시오. (단, Min Support는 0.001 이상, Min Confidence는 0.001 이상 이어야 한다.) (15점)

(답안예시) 스포츠의류

**[답안]**

5 : 의류,화장품

6 : 49,9

7 : 11,G2>G3=G1

8 : 가공식품