**문제 유형3**

다음은 영국에 위치한 온라인 쇼핑몰 U사의 2010년~2011년의 판매 데이터이며, 판매 영수증 정보는 다음과 같다.

Online2.txt (구분자: tab(‘\t’), 541,909 Rows, 8 Columns, UTF-8 Encoding)

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **데이터 타입** |
| InvoiceNo | 영수증 번호 | Double |
| StockCode | 물품 번호 | String |
| Description | 제품 설명 | String |
| Quantity | 구매 수량 | Double |
| InvoiceDate | 구매 시간, 2010년의 경우 12 월만 존재 | String |
| UnitPrice | 물품 단가(단위: 파운드) | Double |
| CustomerID | 구매자 ID | Double |
| Country | 배송지 국가 | String |

**\* 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

**단계 1-1**. 영수증 번호(InvoiceNo)가 수치형이 아닌 데이터 Row는 제거하시오.

**단계 1-2**. 물품 단가(UnitPrice) 또는 구매 수량(Quantity)이 0 이하(<=0)인 Row는 제거하시오.

**단계 1-3**. 구매액(변수명: TotOrder) 변수를 추가하시오. TotOrder = Quantity\*UnitPrice

**\* 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 530,103 Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1을 이용) ‘영수증 번호(InvoiceNo)별 총 구매액(변수명: InvTotOrder)’의 평균이 국가별로 차이가 있는지 다음의 조건들을 참고하여 검정을 수행하고 이 때의 검정 통계량(F-value)을 기술하시오. 이때 처리 순서는 다음 제시된 조건 순서대로 진행하시오.

[조건]

- 하나의 InvoiceNo는 하나의 Country에 대응

- 영수증 번호별 총 구매액(InvTotOrder)은 각 InvoiceNo에 대한 TotOrder의 합으로 정의 (19959개)

- 국가별 구매 건수(InvoiceNo 수)가 50이상(>=50) 이고 400이하(<=400)인 국가만을 대상으로 검정 수행, 자유도(df)=14

- 주어진 데이터는 등분산 조건을 만족한다고 가정

- 검정 통계량은 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술 (답안 예시) 1.234

|  |  |
| --- | --- |
| 정답 | 3.7262665618641413 |

2. (basetable1을 이용) ‘주문날짜(InvoiceDate)’와 ‘영수증 번호(InvoiceNo)별’ 총 구매액’(InvTotOrder, 1번 참고) 테이블을 구성한 후 요일(Weekday)변수를 생성하시오. (InvoiceNo, InvTotOrder Weekday 컬럼을 가지는 데이터 프레임, 19959 rows). 이 때, 요일별 총 구매액이 상위 20%에 속하는 데이터만을 대상으로 총 구매액의 요일별 평균을 구하시오. 평균이 가장 낮은 요일과 평균 값을 기술하시오

- 평균 값은 정수 부분만 기술하시오.

- 영수증 번호별 총 구매액(InvTotOrder)은 각 InvoiceNo에 대한 TotOrder의 합으로 정의

- quantile(0.8) 함수를 사용하되 함수의 parameter는 기본값 사용

(답안예시) 일요일, 1

|  |  |
| --- | --- |
| 정답 | 일요일, 942 |

3. (basetable1을 이용) 구매자(CustomerID)가(와) 주문 날짜(InvoiceDate) 기준으로 동시에 구매한 물품(StockCode)들에 대한 연관성 분석을 다음 조건과 같이 수행하고자 합니다

- 분석 대상: 물품 단가(UnitPrice)의 평균이 4 이상(>=4)이고 구매(판매) 수량(Quantity)의 합이 600 이상(>=600)인 물품

- 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 물품 목록은 동일목록 존재 시 한 품목만 포함

- 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 한 종류의 물품만을 구매한 구매자는 제외

- 연관성 분석 시 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 물품목록을 추출한 후 연관성 분석(basket, transaction) 단위로 함

Note: 물품(StockCode)의 UnitPrice는 구매 시점마다 상이할 수 있습니다.

- 구매자 (CustomerID)를 확인할 수 없는 경우(결측치)는 분석에서 제외합니다.

- (Association Rule 모델 가이드) Min Support: 0.01, Min Confidence: 0.01, 그 외: Default

연관성 분석 결과 중, 선행(antecedent)과 후행(consequent)이 단일 아이템으로 구성된 rule만 추출하시오. 이 결과를 활용하여, 물품 '23012'을 주문한 고객에게 추천하기 적합한 품목을Lift 기준(내림차순)으로 선정하고, 그 때의 Lift를 기술하시오.

- Lift는 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술하시오.

(답안예시) 1.234

|  |  |
| --- | --- |
| 정답 | 23013, 42.52577319587629 |