실전 종합 문제 1

01 시나리오

서울시 소재의 DS 운수에서는 정기적으로 회사 경영진과 버스기사간 연봉과 근무 처우 관련하여 협상을 한다. 올해부터 데이터 분석을 본격적으로 도입하기로 하였다. 그리하여 각자의 의견을 조율함에 있어 데이터 분석을 기반으로 보다 객관적인 의사결정을 하고자 한다.

02 데이터 설명

서울시 2019년 승차 정보 - Seoul_Bus_2019.csv

변수명	설명
Year_Month	연도와 월
Line_ID	노선 식별자
Line_No	노선 번호
Line_Name	노선 이름
Station_ID	정류소 식별자
Station_Name	정류소 이름
H01 ~ H24	각 시간대별 승차 인원

서울시 버스 정보 - Seoul_Bus_info.csv

변수명	설명
Bus_no	버스 노선 번호
type	구분

03 문제

1번

지선, 간선 버스의 경우 노선당 년 수익이 10억 이하인 경우 배차간격 조정이나 노선 변경 등 수익구조 개선이 필요하다고 한다. 승객 1명당 기대 수익이 천 원이라고 했을 때 몇 개의 버스 노선이 대상인가? (정답 예시: 1)

2번

간선 버스노선 버스기사들은 간선 버스노선이 지선 버스노선 대비 정류장 개수가 많아 버스 기사 확충 또는 배차 간격 조정을 사측에 요구하고 있다. 경영지원팀은 요구를 수용하기 위해 지선과 간선 노선의 정류장 개수를 파악하고 이 차이가 통계적으로 유의한지 확인해보려 한다. 간선 버스노선의 평균 정류장 개수와 지선 버스노선의 평균 정류장 개수의 평균을 적절한 검정을 통해 그 차이를 비교하고 그 검정통계량의 절대값을 소수점 둘 째 자리까지 반올림하여 기술하시오. (정답 예시: 1.23)

3번

출퇴근 시간 배차간격 조정을 위해 우선적으로 각 정류소별 시간대별 출근 시간 승차 패턴을 파악하고자 한다. 지선 버스 노선을 대상으로 각 정류장별 승차인원을 계층적 군집분석을 활용하여 6개의 군집으로 분할하였을 때 출근시간대에 승차 인원이 가장 많은 정류소 군집의 번호는 몇 번인가?
※ 출근 시간대는 오전 7시 부터 9시 까지로 정의
※ 군집분석시 자료는 Min-Max 정규화 실시
※ 거리 계산은 유클리디안 거리로 하고 유사도는 Ward.D 방법을 사용

실전 종합 문제 2

01 시나리오

DS 금융의 올해 목표는 고객의 금융 서비스 이용 패턴 기반의 신규 서비스 런칭과 보다 객관적이고 투명한 고객 신용거래를 지원하기위해 본격적으로 데이터 분석 기법을 도입하기로 하였다. 이를 위하여 1만명의 고객 데이터를 샘플로 하여 파일럿 프로젝트를 실시하기로 하였다.

02 데이터 설명

은행 고객 데이터 - financial_info_10k_persons.csv

변수명	설명
ID	고유 번호
is_attrited	이탈 여부(이탈: 1)
Age	나이
Gender	성별
Dependent_cnt	부양가족 수
Edu_level	교육 수준
Marital_status	결혼 상태
Income	수입
Card	카드 등급
Period_m	가입 기간(월)
Total_rel_cnt	서비스 이용 횟수
Inactive_last_12m	최근 12개월동안 금융 거래가 없었던 기간(월)
Contact_cnt_last_12m	최근 12개월동안 영업점 방문 횟수
Credit_limit	신용 한도
Total_trans_amt	누적 송금액
Total_trans_cnt	누적 송금 횟수

03 문제

1번

고객의 총 송금액이 교육 수준, 혼인 여부에 따라서 어떤 특징을 보이는지 분산분석을 통해 알아보고자 한다.
1회 평균 송금액을 종속변수로 했을 때 독립변수간 교호작용 여부를 알아보고 해당 p-value를 반올림하여 소수점 둘 째 자리까지 기술하시오. (정답 예시: 0.12)

2번

고객의 신용 한도는 다양한 정보를 기반으로 결정된다. 고객의 금융활동이 누적됨에 따라 신용 한도는 바뀌기도 하는데 이를 비교하고자 한다. 고객의 신용한도를 종속변수를 공통으로 하고 부양가족, 수입, 나이, 학력, 성별, 결혼여부를 1번 회귀 모델. 1번 모델에 가입기간과 누적 송금 횟수를 독립변수에 더한 회귀 모델의 결정계수 차이의 절대값을 반올림하여 소수점 셋 째 자리까지 기술하시오. (정답 예시: 0.123)

3번

신규 고객이 개인정보를 입력할 경우 예상 신용 한도를 보여주려고 한다.
부양가족과 수입이 없는 29세 고졸(High School) 미혼의 남성의 경우 예상 신용 한도를 정수 부분만 기술하시오. (정답 예시: 1)

실전 종합 문제 3

01 시나리오

DS 마트의 경영부서는 기존의 매출 데이터를 기반으로 마케팅 팀에 전달한 집중 홍보 상품 목록 선정과 매장 매대 진열 및 소비자 동선 수정을 이번 달 목표로 잡았다. 이를 위해 매출 데이터, 상품 정보, 고객정보를 전산팀으로부터 인계 받아 분석을 준비하였다. 정제된 로그 데이터를 활용하여 다음의 분석을 실시하시오.

02 데이터 설명

마트 매출 데이터 - association_rules_mart.csv

변수명	설명
Date	판매일
ID	고객 식별자
Item	판매 품목

고객 데이터 - association_rules_customers.csv

변수명	설명
ID	고객 식별자
Gender	성별(남자: M, 여자: F)
Age	나이

품목 데이터 - association_rules_products.csv

변수명	설명
product	제품명
price	가격

03 문제

1번

2014년에는 매출이 발생했으나 2015년에는 매출이 발생하지 않은 품목은 총 몇 개 인가? (정답 예시: 1)

2번

전해 12월의 매출은 차년도 매출과 꽤나 연관이 깊다고 한다. 2014년도 12월 매출 상위 3개 품목을 확인하고 해당 품목의 2015년 매출 비중을 반올림하여 소수점 셋째 자리 까지 기술하시오. (정답 예시: 0.123)

3번

남성과 여성의 상품 구매 성향이 다르다는 가정을 확인하기 위해서 2015년 데이터를 기반으로 연관규칙 분석을 실시하고 지지도가 0.05 이상인 규칙 중 향상도가 가장 높은 조건의 결과절(consequent) 품목을 남녀 차례대로 기술하시오.
(정답 예시: water, sugar)

※ 단, 구매 품목이 1건인 회원의 정보는 제외한다.
※ 최초 지지도와 신뢰도 설정은 0.005로 한다.