종합평가 데이터 설명 및 문제

```
# 문제 풀이에 필요한 함수가 있는 패키지를 전부 불러옵니다. 아래 코드를 실행해주세요
library(ggplot2)
library(dplyr)
library(caret)
set.seed(1004)
```

[1~2번 문제] bank data

- 데이터 출처 : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing 원본 데이터 다운로드 후 변형하였음
- 데이터 설명

  포르투갈 어느 은행의 마케팅 캠페인 데이터로, 마케팅 캠페인은 휴대전화로 진행되었으며 은행 정기 예금에 가입을 원하는지 여부를 물었다. 변수 설명은 다음과 같습니다.

1. ID : 고객 ID (식별자)
2. age : 나이
3. job : 직업 종류

   (범주 : admin(관리자), blue-collar(생산직 근로자), entrepreneur(기업가), housemaid(가사도우미), management(경영간부), retired(은퇴), self-employed(자영업자), services(서비스직), student(학생), technician(기술자), unemployed(실업자), unknown(모름))
4. marital : 혼인여부(범주 : divorced(이혼), married(기혼), single(미혼))
5. education : 교육수준 (범주 : secondary, primary, tertiary, unknown)
6. default : 신용불량 여부 (범주 : no(아니오), yes(예), unknown(모름))
7. balance : 연평균 잔액 (단위 : 유로)
8. housing : 주택 담보 대출 여부 (범주 : no(아니오), yes(네))
9. loan : 개인 대출 여부 (범주 : no(아니오), yes(예))
10. contact : 접촉한 통신기기 종류 (범주 : cellular(휴대전화), telephone(유선전화), unknown(모름))
11. month : 마지막으로 접촉한 월 (범주 : jan, feb, mar, ..., dec)
12. day : 마지막으로 접촉한 달의 일자 (1, 2, ... 31)
13. duration : 마지막 접촉 시간(초)
14. campaign : 캠페인 동안 해당 고객에게 접촉한 연락 수
15. y : 정기예금 가입 여부 (범주 : no(아니오), yes(예))

```
# 아래 코드를 실행하여 데이터를 불러오세요. 경로는 적절하게 설정하세요
bank1 = read.table("./bank1.csv", sep = ";", header = TRUE)
bank2 = read.table("./bank2.csv", sep = ";", header = TRUE)
bank = merge(bank1, bank2, by = "ID")
```

(5점) 1-(1). 새로운 데이터는 여러가지를 검토하면서 데이터를 파악해야 합니다. 데이터를 검토하는 함수를 2개 이상 작성하세요.

(5점) 1-(2). 데이터에 결측치가 있는지 확인하세요. (참고로 결측치는 없습니다.)

(5점) 1-(3). 20세부터 59세까지의 고객만 고려하려고 합니다. age변수가 20세 이상 60세 미만인 데이터만 추출하세요.

(5점) 1-(4). 직업 종류(job)에 따른 연평균 잔액(balance)의 평균을 구하고 내림차순으로 정렬하세요.

(5점) 1-(5) 교육수준(education)에 따른 정기예금 가입(y) 비율을 sum_by_edu라는 데이터에 저장하세요.
Hint. 1과 0으로 이루어진 변수는 mean을 이용하면 비율을 구할 수 있습니다.

(5점) 1-(6) 위 문제에서 만든 sum_by_edu 데이터를 이용하여 교육수준에 따른 정기예금 가입 비율을 막대그래프로 표현하세요.

(7)~(8) 나이와 평균 연평균 잔액의 관계를 살펴보고 싶습니다.
(5점) 1-(7) 먼저 나이(age)별로 연평균 잔액(balance)의 평균을 구하고 이를 age_bl라는 데이터에 저장하세요.

(5점) 1-(8) age_balance데이터를 이용하여 나이와 평균 연평균 잔액의 산점도를 그리세요.

(5점) 2-(1). 나이를 범주화 하여 age_group 변수로 지정해주세요.
이 때, 범주는 20대, 30대, 40대, 50대로 나누어주시기 바랍니다.

(5점) 2-(2) 연령 그룹에 따라 연평균 잔액의 평균이 다른지 알고 싶습니다. 알맞은 가설을 세워주세요.

귀무가설 :
대립가설 :

(10점) 2-(3) (2)에서 세운 가설에 맞게 가설검정을 진행하세요. 관련 코드와 통계량에 따른 결론을 모두 작성해주셔야 합니다.
단, 정규성과 등분산성을 모두 만족한다고 가정합니다.

결론을 아래에 작성하세요.

[3~5번 문제] HR_new data

- 데이터 출처 : https://github.com/ryankarlos/Human-Resource-Analytics-Kaggle-Dataset 에서 원본 데이터 다운로드 후 데이터 전처리 및 샘플링 진행하였음
- 데이터 설명

  회사를 떠나는 직원을 예측하기 위한 문제입니다. 각 변수에 대한 설명은 아래와 같습니다. (변수명과 대치시켜서 생각해주시면 됩니다.)

  - Employee satisfaction level
  - Last evaluation
  - Number of projects
  - Average monthly hours
  - Time spent at the company
  - Whether they have had a work accident
  - Whether they have had a promotion in the last 5 years
  - Department
  - Salary
  - Whether the employee has left

```
# 아래 코드를 실행해주세요 (경로는 각자 환경에 맞게 지정하세요.)
hr <- read.csv("./HR_new.csv", na.string = c("", " "))
hr$left <- as.factor(hr$left)
summary(hr)
```

(5점) 3. 결측치가 last_evaluation, time_spend_company에 있다고 할 때, 이들 두 변수에 있는 결측치를 제거하는 코드를 작성해주세요. (현재 샘플링으로 인해 없어진 상태이지만 있다고 했을 때에 사용하는 코드를 작성해주시면 됩니다.)

*hr 데이터는 left를 제외하고 모두 연속형 변수로 이루어져 있습니다.

(10점) 4. 연속형 변수들을 이용하여 차원축소를 진행하려고 합니다. 차원 축소 후, 각 주성분이 분산의 몇%를 설명하는지 까지 확인하는 코드도 작성하세요.

*hr 데이터의 종속변수는 left로, 직장을 떠나는지의 여부입니다.
(15점) 5. 분류에 해당하는 많은 알고리즘을 배웠습니다. 본인이 사용하고 싶은 알고리즘을 이용하여 분석을 진행하세요.
결과 해석(test 데이터를 이용한 예측) 까지 진행하시기 바랍니다.
Hint. 데이터 분할 / 데이터 분석 / 결과 예측 순서대로 진행하시기 바랍니다.

[6~7번 문제] Adsp 관련 문제

설명에 맞는 단어를 작성해주세요.

(5점) 6. 아래에서 설명하고 있는 빅데이터 활용 기본 테크닉은 무엇인가요?

- 자연세계의 진화과정에 기초한 계산 모델로서 최적화 문제를 해결하는 기법의 하나이다. 생물의 진화를 모방한 진화 연산의 대표적인 기법으로, 실제 진화의 과정에서 많은 부분을 차용하였다.
- 어떤 미지의 함수 Y = f(x)를 최적화하는 해 x를 찾기 위해 진화를 모방한 탐색 알고리즘이라고 말할 수 있다.

정답 :

(5점) 7. KDD 분석 방법론에서 분석 목적에 맞게 변수를 생성, 선택하는 등 데이터마이닝을 할 수 있도록 데이터를 변경하는 단계와 유사한 CRISP-DM 분석 방법론의 단계는 무엇인가요?
정답 :