종합 평가 답안지

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


# 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")

# 각 변수에 대한 설명은 문제 pdf파일을 참고하시기 바랍니다.

#
# (5점) 1-(1). 새로운 데이터는 여러가지를 검토하면서 데이터를 파악해야 합니다. 데이터를 검토하는 함수를 2개 이상 작성하세요.
head(bank)
str(bank)
dim(bank)
summary(bank)
...

# (5점) 1-(2). 데이터에 결측치가 있는지 확인하세요. (참고로 결측치는 없습니다.)
sum(is.na(bank))
또는
table(is.na(bank)
를 통하여 확인하실 수 있습니다.
직접 한 변수씩 확인해도 되지만 우선 이런식으로 전체 데이터 내에 NA가 있는지 파악해본 뒤, 만약 결측치가 있다면
추가로 개별 변수를 살펴보는 것이 시간을 단축시킬 수 있습니다.

# (5점) 1-(3). 20세부터 60세까지의 고객만 고려하려고 합니다. age변수가 20세 이상 60세 미만인 데이터만 추출하세요.
bank <- bank %>% filter(age >=20 & age < 60)


# (5점) 1-(4). 직업 종류(job)에 따른 연평균 잔액(balance)의 평균을 구하고 내림차순으로 정렬하세요.
bank %>% group_by(job) %>%
  summarise(mean_bal = mean(balance)) %>%
  arrange(desc(mean_bal))


# (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.
sum_by_edu <- bank %>%
  group_by(education) %>%
  summarise(prop = mean(y == "yes"))


# (5점) 1-(6) 위 문제에서 만든 sum_by_edu 데이터를 이용하여 교육수준에 따른 정기예금 가입 비율을 막대그래프로 표현하세요.
ggplot(sum_by_edu, aes(x = education, y = prop, fill = education)) +
  geom_col()
평균표를 만들어서 그래프를 만들었으므로 geom_bar 대신 geom_col을 사용합니다.

* 만약 geom_bar를 사용하고 싶다면
ggplot(sum_by_edu, aes(x = education, y = prop, fill = education)) +
  geom_bar(stat = "identity")
이렇게 (stat='identity') 를 추가해주면 동일하게 표현 가능합니다.
geom_bar 내의 stat인자의 기본 값은 count로 빈도수를 계산하는데, stat='identity'로 지정하면 y축의 높이를 데이터의 값으로 하는
bar그래프의 형태로 지정한다는 뜻입니다.
Tip. 그래프를 좀 더 보기 좋게 표현하기 위해서 제목, 범례의 이름 등을 직접 지정할 수 있습니다.
아래 함수도 한 번 실행해보세요!

ggplot(sum_by_edu, aes(x = education, y = prop, fill = education)) +
  geom_bar(stat = "identity") +
  ggtitle("교육 정도에 따른 정기예금 가입 비율") +
  xlab("교육수준") + ylab("비율") +
  scale_fill_discrete(name = "교육수준")


# (5점) 1-(7) 나이와 평균 연평균 잔액의 관계를 살펴보고 싶습니다.
# 먼저 나이(age)별로 연평균 잔액(balance)의 평균을 구하고 이를 age_bl라는 데이터에 저장하세요.
age_bl <- bank %>% group_by(age) %>% summarise(mean_bal = mean(balance))
```

```
# (5점) 1-(8) age_balance데이터를 이용하여 나이와 평균 연평균 잔액의 산점도를 그리세요.
ggplot(data = age_bl, aes(x = age, y = mean_bal)) + geom_point()
또는
plot(age_bl$age, age_bl$mean_bal)

# (5점) 2-(1). 나이를 범주화 하여 age_group 변수로 지정해주세요.
# 이 때, 범주는 20대, 30대, 40대, 50대로 나누어주시기 바랍니다.
bank$age_group = ifelse(bank$age < 30, '20대',
                        ifelse(bank$age < 40, '30대',
                               ifelse(bank$age < 50,'40대','50대')))

많은 분들이
bank$age_group <- ifelse(bank$age<30, '20대',
                        ifelse(bank$age<40 & bank$age >= 30, '30대',
                               ifelse(bank$age<50 & bank$age >= 40, '40대','50대')))
이렇게 중복 논리문을 사용해주셨는데, 정답 코드 처럼 작성해주셔도 동일하게 작동합니다.
가장 안 쪽의 ifelse문 ifelse(bank$age < 50,'40대','50대') 이 먼저 실행되어 50세 미만은 전부 40대로 처리되고,
그 다음 ifelse문에서 이 중 40세 미만은 30대로 처리되기 때문으로 이해해주시면 됩니다.

# (5점) 2-(2) 연령 그룹에 따라 연평균 잔액의 평균이 다른지 알고 싶습니다. 알맞은 가설을 세워주세요.
귀무가설 : 연평균 잔액의 평균은 연령 집단에 따라 차이가 없다.
대립가설 : 연평균 잔액의 평균은 연령 그룹에 따라 차이가 있다.

# (10점) 2-(3) (2)에서 세운 가설에 맞게 가설검정을 진행하세요. 관련 코드와 통계량에 따른 결론을 모두 작성해주셔야 합니다.
# 단, 정규성과 등분산성을 모두 만족한다고 가정합니다.
aov_test = aov(balance ~ age_group, data = bank)
summary(aov_test)

# 결론을 아래에 작성하세요.
p-value가 2e-16으로 유의수준 0.05보다 작으므로 귀무가설을 기각한다.


# 3~5번 문제는 HR_new 데이터를 이용하여 풀어주세요.
# 데이터에 대한 설명은 문제 pdf를 참고해주세요.
# 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. summary(hr)을 통해 데이터에 일부 결측치가 존재함을 알게 되었습니다.
# 결측치를 제거하는 코드를 작성하세요. 두 변수 모두 결측치를 제거하여야 합니다.
hr <- hr[!is.na(hr$time_spend_company), ]
hr <- hr[!is.na(hr$last_evaluation), ]
or
hr <- hr %>% filter(!is.na(time_spend_company) & !is.na(last_evaluation))
or
hr <- na.omit(hr)

# hr 데이터는 left와 salary를 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.
pca_data <- hr %>% select(-left, -salary)
hr.pca <- prcomp(pca_data, center = T, scale. = T)
summary(hr.pca)

# 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.
datatotal <- sort(sample(nrow(hr), nrow(hr)*0.7))
train <- hr[datatotal,]
test <- hr[-datatotal,]

ctrl <- trainControl(method="repeatedcv",repeats = 5)
rf_fit <- train(left ~ .,
                data = train,
                method = "rf",
                trControl = ctrl,
                metric="Accuracy")

rf_pred <- predict(rf_fit, newdata=test)
confusionMatrix(rf_pred, test$target)


# 6~7 번은 Adsp 강의와 관련된 내용입니다.
# 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 :  유전자 알고리즘 (Genetic Algorithms)

# (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 : 데이터 준비 (Data preparation)
```