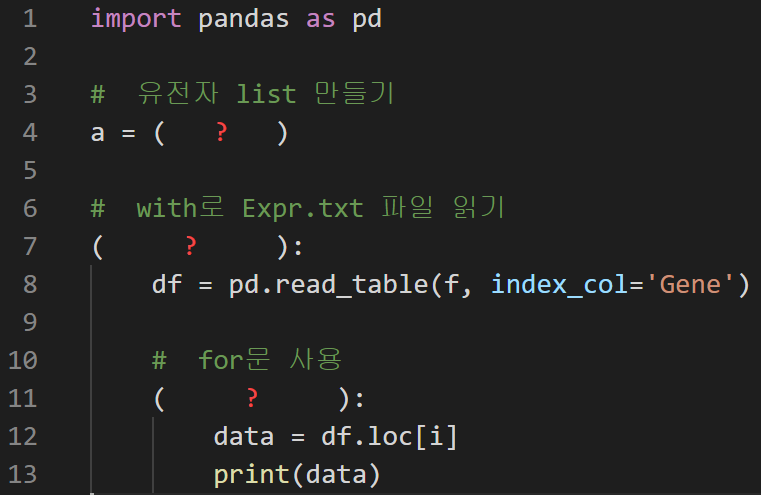
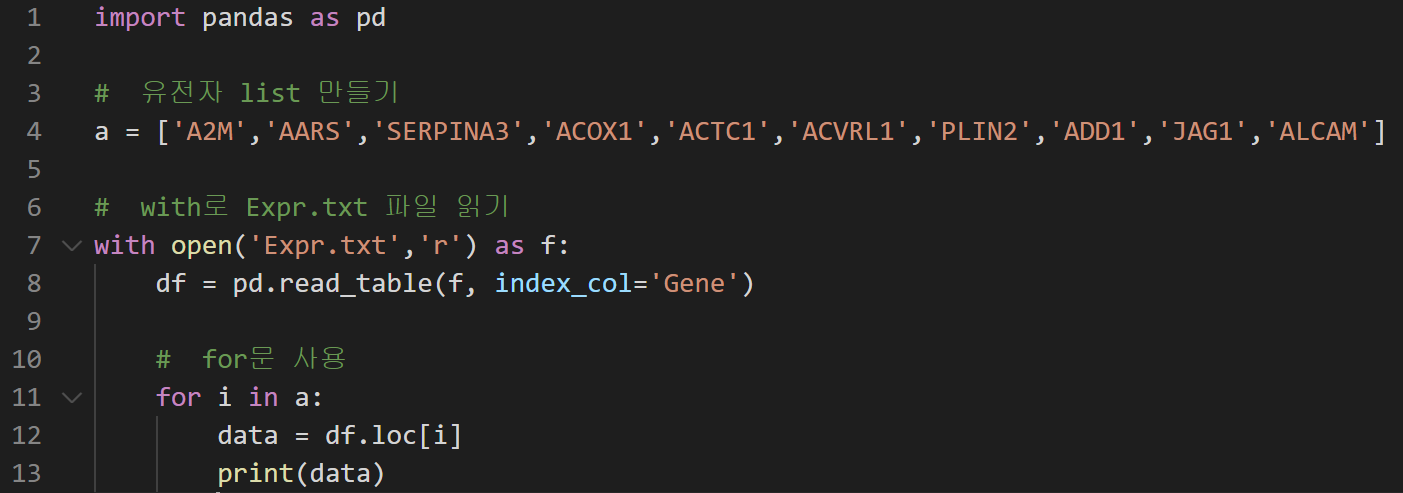
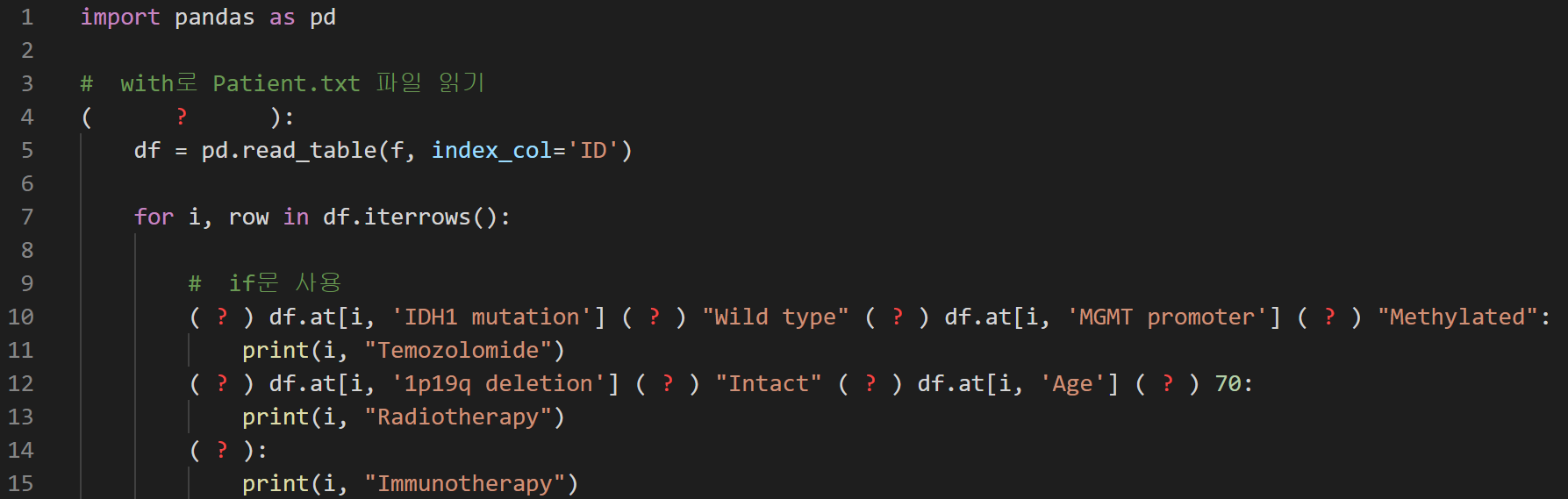
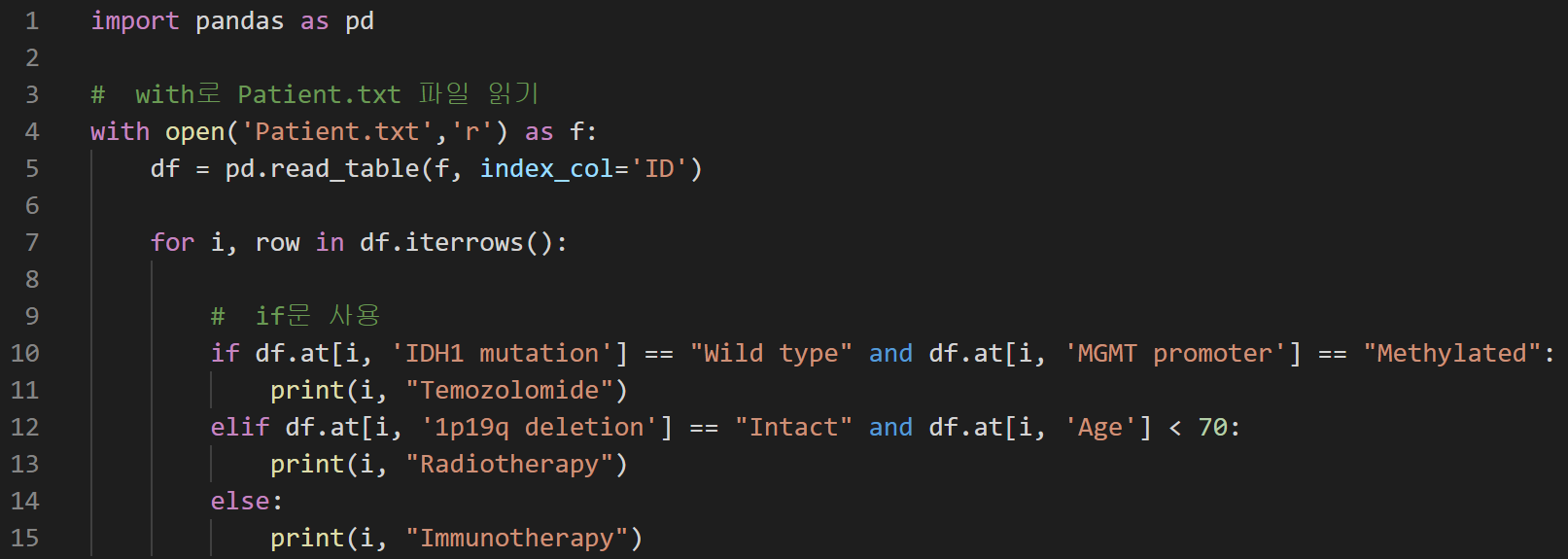
**Quiz 2. Python & R**

\* Excel quiz 때와 같은 데이터셋을 사용합니다 (tab-delimited text file).

\* 순서대로 문제를 풀어야합니다.

1. (Python) Expr.txt 파일로부터 A2M, AARS, SERPINA3, ACOX1, ACTC1, ACVRL1, PLIN2, ADD1, JAG1, ALCAM 10개 유전자에 대한 92명 환자의 expression 값을 출력하는 python code를 작성하고자 합니다. 아래 code의 괄호 안에 들어갈 알맞은 내용을 작성하세요.  
     
   답안  
   
2. (Python) Patient.txt의 정보를 이용하여 각 환자별로 적절한 치료법을 제시하고자 합니다. (1) *IDH1* mutation 여부가 wild-type이면서 *MGMT* promoter가 methylated이면 “Temozolomide”로 표기; (2) 나머지 환자들 중 1p19q deletion이 없으면서 나이 (Age)가 70세 미만인 환자는 “Radiotherapy”로 표기; (3) 그 외의 환자들은 “Immunotherapy”로 표기하는 python code의 빈칸에 들어갈 알맞은 내용을 적으세요.  
     
   답안  
   
3. (R) read.table 함수를 이용해서 Expr.txt 파일과 Patient.txt 파일을 불러오는 코드를 작성하세요.
   * 불러온 정보를 각각 expr, patient 라는 이름의 변수에 저장하세요.
   * Hint 1: header = TRUE 인자를 사용하면 파일의 첫 행을 colname으로 사용
   * Hint 2: row.names = 1 인자를 사용하면 파일의 첫 열을 rowname으로 사용

**expr <- read.table("Expr.txt", sep = "\t", header = TRUE, row.names = 1)  
patient <- read.table("Patient.txt", sep = "\t", header = TRUE, row.names = 1)**

1. (R) correlation이라는 이름의 벡터를 하나 생성해서 유전자의 발현량과 환자의 생존기간과의 상관관계 (Pearson correlation)를 저장하고, 벡터의 각 요소에 해당 유전자 이름을 할당하는 코드를 작성하세요.
   * Hint 1: 새로운 벡터는 c() 함수로 생성할 수 있고, 벡터의 각 요소에 대한 이름은 names() 함수로 불러올 수 있으며, 데이터프레임의 행 이름은 rownames(expr)과 같은 형태로 불러올 수 있음
   * Hint 2: 상관관계는 cor() 함수로 계산할 수 있으며, 모든 유전자에 대해 생존기간과의 상관관계를 계산하기 위해 for문을 사용 (expr의 행 수만큼 반복문 수행)
   * Hint 3: 데이터프레임의 특정 행을 불러오면 벡터 형태로 반환되지 않음. 상관관계 계산 시 숫자 벡터 형태로 인자를 입력해야하므로 인자에 as.numeric() 함수 사용

**correlation <- c()  
for (i in 1:nrow(expr)) {  
 correlation[i] <- cor(as.numeric(expr[i, ]), patient[, “Survival”])  
}  
names(correlation) <- rownames(expr)**

1. (R) correlation 벡터에 저장된 정보를 통해 상관관계가 가장 낮은 (환자의 예후에 가장 나쁜 영향을 끼칠 것으로 예상되는) 유전자를 찾아낸 후, target이라는 이름의 데이터프레임을 생성하여 해당 유전자의 환자별 발현량과 생존기간 정보를 저장하는 코드를 작성하세요.
   * 첫번째 열의 이름을 해당 유전자 이름으로, 두번째 열의 이름을 Survival로 생성하세요.
   * Hint 1: which.min() 함수를 사용하여 벡터 내 최소값을 가지는 위치를 반환
   * Hint 2: 2번 문제와 마찬가지 이유로 데이터프레임 생성시 as.numeric() 함수 사용

**which.min(correlation)  
target <- data.frame(PARP1 = as.numeric(expr["PARP1", ]), Survival = patient[, "Survival"])**

1. (R) ggplot() 함수를 사용해서 3번 문제에서 저장한 target 변수에 저장된 유전자-생존기간 정보를 scatter plot 그래프로 표현하고 추세선을 그리는 코드를 작성하세요.
   * Hint 1: ggplot2 패키지를 설치하고 호출해야 ggplot() 함수 실행 가능
   * Hint 2: scatter plot을 그리기 위해 geom\_point(), 추세선을 그리기 위해 geom\_smooth() 함수 사용

**ggplot(data = target, aes(x = PARP1, y = Survival)) + geom\_point() + geom\_smooth()**