데이터과학 Homework1 보고서

21600685 조예성

Question1

은행 데이터를 가져오기 위해 read.csv()를 실행해주고 모든 string데이터는 factor형 변수이기에 stringsAsFactors = T로 인자를 설정합니다. 해당 데이터프레임은 bank라는 변수에 저장합니다

nrow()를 통해 고객들의 수를 파악합니다

전체고객들의 수 : 4521명

이 중 나이가 30보다 어린 고객과 50보다 많은 고객들의 데이터를 포함하기 위해 데이터 프레임을 해당 기준으로 subset()을 이용해 분리합니다

bank\_under\_thiry <- 30대 이하의 고객 데이터

bank\_over\_fifty <- 50대 이상의 고객 데이터

nrow()를 통해 두 데이터 프레임의 고객의 수를 더합니다

30대 이하, 50대 이상의 고객 수 : 1409명

Question 1 정답 : 4521명, 1409명

Question 2

‘$’ 을 이용해 bank$balance\_kw 변수 설정, 해당 변수에 기존 유로화인 bank$balalnce 에 1200을 곱해 한화 수 구함

(정답은 스크립트 파일을 참고)

Question 3

정기 예금 상품에 가입한 고객의 수를 확인하기 위해 subset() 이용

* bank\_yes <- subset(bank, bank$y == “yes”)

정기 예금 상품에 가입한 고객 수 : bank\_yes 변수 설정 nrow()를 통해 고객 수 파악

* nrow(bank\_yes)

예금 상품에 가입한 고객수는 521명

기존 bank의 고객 수 대비 예금 상품에 가입한 고객 수 비율 구함

* nrow(bank\_yes)/nrow(bank)

정답 : 521명, 0.11524

Question 4

-1이 있는 개수를 파악하기 위해 which() 명령어 사용, 이전에 연락한 적이 없는 고객들의 인덱스 반환

=> which(bank$pdays == -1)

해당 인덱스의 수를 이용해 length()를 통한 고객 수 파악

* length(bank$pdays[which(bank$pdays == -1)])

해당 값을 NA로 전환

* bank$pdays[which(bank$pdays == -1)] <- NA

NA 개수는 3705개

정답 : 3705개

Question 5

직업 군 별 고객수를 구하기 위해 summary() 사용

* summary(bank$job)

정답 :

* admin. : 478
* blue-collar : 946
* entrepreneur : 168
* housemaid : 112
* management : 969
* retired : 230
* self-employed : 183
* services : 417
* student : 84
* technician : 768
* unemployed : 128
* unknown : 38

Question 6

우선, cut()을 사용하기 위한 가장 고령 나이를 파악합니다. 이를 위해서는 summary() 사용

* summary(bank$job)

명령 실행 결과 87세가 가장 고령

87세를 포함한, 연령대별 범주형 데이터를 형성하기 위해 cut() 사용, 20~29같은 범위이므로 right = F로설정

연령대에 따른 고객군 분포는 summary를 통해 계산

* summary(bank$age\_group)

20대 이하 : 4

20대 : 478

30대 : 1808

40대 : 1203

50대 : 854

60대 이상 : 174

고로 가장 많은 고객이 속한 연령대는 30대이다

정답 : 30~39세 연령대

연령대에 따른 고객 수를 벡터화하면

* age\_groups <- c(4, 478, 1808, 1203, 854, 174)

이를 통해 추후에 비율을 구할 수 있다.

가입 했는 지의 여부를 count하고 싶지만, yes no라는 factor형이므로 계산이 되는 수치형으로 전환. 가입한 경우를 1, 가입 안 한 경우를 0으로 계산

* bank$y\_int <- ifelse(bank$y == “yes”, 1, 0)

해당 내용을 이용하여 연령대 대비 가입자 수 계산

* aggregate(y\_int~age\_group, data = bank, FUN = sum)

20대 이하 : 2명

20대 : 72명

30대 : 185명

40대 : 123명

50대 : 84명

60대 이상 : 55명

위에서 구했던 고객 수의 분포를 이용하여 비율을 계산해준다.

* age\_group\_yes <- c(2, 72, 185, 123, 84, 55)

비율의 결과값을 반환해준다.

* results <- age\_group\_yes / age\_groups

value값이 가리키는 값을 명확히 하기 위해 naming을 해준다

name(results) <- c(“under 20s”, “20s”, “30s”, “40s”, “50s”, “over 60s”)

연령대별 정기 예금 가입 비율

20대 이하 : 0.5

20대 : 0.15062762

30대 : 0.10232301

40대 : 0.10224439

50대 : 0.09836066

60대 이상 : 0.31609195

마케팅 성공율이 가장 높은 연령대는 20대 이하

Question 8

고객의 마케팅 접촉 유형 별 평균 접촉 기간을 구하는 group by를 위해 aggregate() 사용

* aggregate(duration~contact, data = bank, FUN = mean)

정답 :

* celluar : 267.1126
* telephone : 243.3555
* unknown : 261.7530

Question 9

고객의 나이를 오름차순으로 정렬하기 위해 sort()함수를 사용해서 데이터 프레임에 적용시킨다.

* bank$age <- sort(bank$age)

Question 10

작업한 bank 데이터 프레임을 RData 파일로 저장한다.

* save(bank, file=”Homework.RData”)
* load(“Homework.RData”)