20200407\_서점의\_고객데이터에\_대한\_가상사례

areum

2020 4 7

### 패키지 로드

### 데이터 불러오기

cs0 <- read.csv("../수업DATA/cust\_seg\_smpl\_280122.csv")  
# 데이터 확인  
head(cs0)

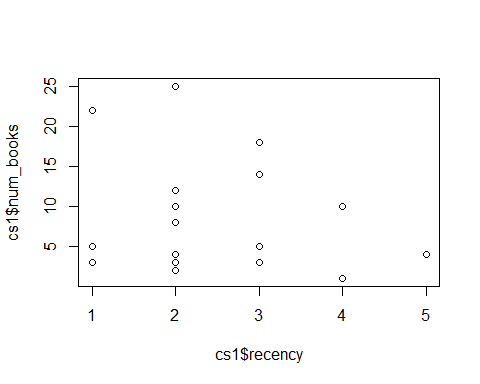
## 고객명 성별 연령 지역 구매일수 최종구매후기간 구매서적수 서적구매액  
## 1 한지민 여 17 부산 5 2 8 96,000  
## 2 장이성 남 19 서울 2 2 3 36,000  
## 3 이수문 남 21 서울 20 1 22 286,000  
## 4 오진호 남 25 인천 4 3 5 60,000  
## 5 염시훈 남 26 서울 3 5 4 160,000  
## 6 진정희 여 26 서울 4 2 12 144,000  
## 기타상품구매액 총구매액 관심쟝르 구매쟝르수 가입기간 SMS수신여부  
## 1 38,400 134,400 소설 2 12 수신  
## 2 14,400 50,400 과학 1 24 수신  
## 3 214,500 500,500 소설 4 15   
## 4 24,000 84,000 시 1 16 수신  
## 5 64,000 224,000 소설 3 9 수신  
## 6 201,600 345,600 소설 4 6 수신

names(cs0)

## [1] "고객명" "성별" "연령" "지역"   
## [5] "구매일수" "최종구매후기간" "구매서적수" "서적구매액"   
## [9] "기타상품구매액" "총구매액" "관심쟝르" "구매쟝르수"   
## [13] "가입기간" "SMS수신여부"

cs1 <- cs0  
  
# 고객명 / 성별 / 연령 / 지역 / 구매일수 / 최종구매후기간 / 구매서적수  
# / 서적구매액 / 기타상품구매액 / 총구매액 / 관심장르 / 구매장르 / 가입기간 / SMS수신여부  
names(cs1) <- c("cust\_name", "sex", "age", "location", "days\_purchase", "recency", "num\_books",  
 "amt\_books", "amt\_non\_book", "amt\_total", "interest\_genre", "num\_genre", "membership\_period", "sms\_option")

# 최종 구매 후 기간 recency와 구매한 서적의 수간의 관계 확인



## [1] 0.995 0.988 1.010 1.209 1.161 1.180 1.215 3.013 3.017 3.016

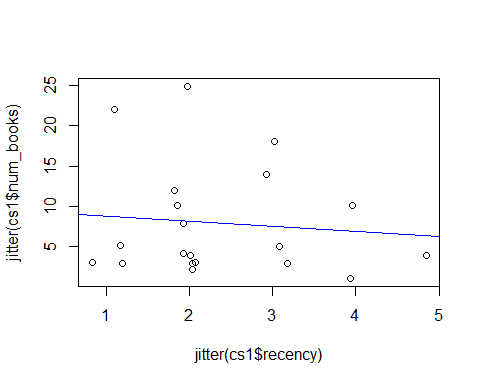
# 가설 1

# 보조선인 회귀선을 본다면 최근성이 낮을수록,

# 즉 구매한지 오래되었을 수록 구매한 서적의 수가 많음

결과 : 구매한지 오래되지 않은 사람일수록 구매한 서적의 수가 많다.

plot(jitter(cs1$recency), jitter(cs1$num\_books))  
abline(lm(cs1$num\_books~cs1$recency), col="blue")

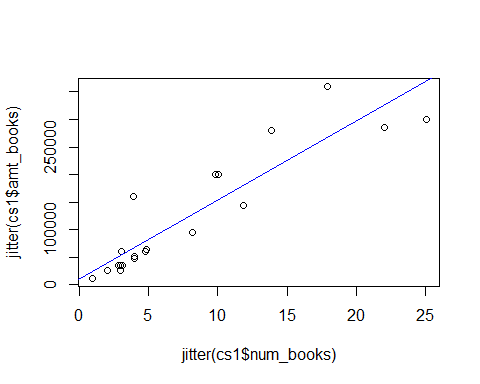


#—————————————————————————— #

**가설 2 구매한 책의 수가 많을수록 구매금액이 큼**

결과 : 구매한 책의 수가 많을수록 구매금액이 크다.

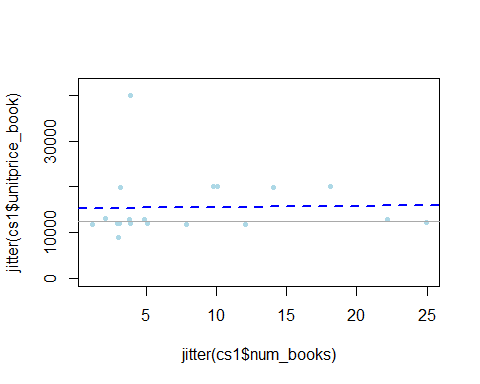
cs1$amt\_books <- as.numeric(gsub(",", "", as.character(cs1$amt\_books)))  
cs1$amt\_non\_book <- as.numeric(gsub(",", "", as.character(cs1$amt\_non\_book)))  
  
plot(jitter(cs1$num\_books), jitter(cs1$amt\_books))  
abline(lm(cs1$amt\_books~cs1$num\_books), col="blue")



# 주로 비싼 책을 샀는지를 파악하기 위해 평균금액을 계산

결과 : 소비자는 저렴한 책을 더 많이 구매한 것으로 보아, 앞으로 저렴한 책 위주로 판매하는 것이 더 효과적이다.

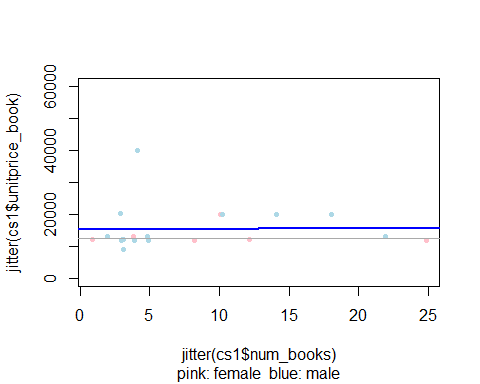
cs1$unitprice\_book <- cs1$amt\_books / cs1$num\_books  
  
plot(jitter(cs1$num\_books),  
 jitter(cs1$unitprice\_book),  
 pch=19,  
 col="lightblue",  
 cex=0.7,  
 ylim=c(0, max(cs1$unitprice\_book)\*1.05))  
  
abline(lm(cs1$unitprice\_book~cs1$num\_books), col="blue", lwd=2, lty=2)  
  
abline(h=median(cs1$unitprice\_book), col="darkgrey")



# 성별을 구분해서 특성 차이 비교

결과 : 남자가 여자보다 책을 더 많이 구매했다.

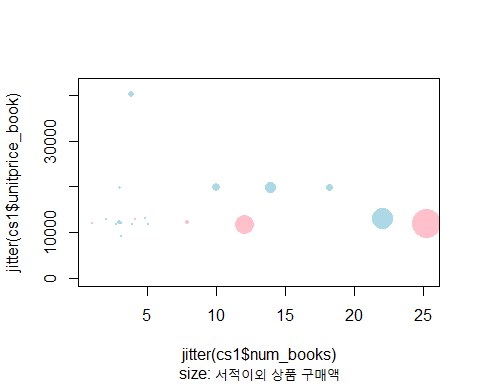
plot(jitter(cs1$num\_books),  
 jitter(cs1$unitprice\_book),  
 pch=19,  
 cex=0.7,  
 col=ifelse(cs1$sex=="여", "pink", "lightblue"),  
 ylim=c(0, max(cs1$unitprice\_book)\*1.5),  
 sub="pink: female blue: male")  
  
abline(lm(cs1$unitprice\_book~cs1$num\_books), col="blue", lwd=2, lty2)  
abline(h=median(cs1$unitprice\_book), col="darkgrey")



# 서적과 서적이외 구매액 비교

결과 : 여자가 남자보다 서적이외 상품 구매금액이 더 크다.

# 동그라미 크기 비율 조정  
plot(jitter(cs1$num\_books),  
 jitter(cs1$unitprice\_book),  
 pch=19,  
 cex=4\*cs1$amt\_non\_book/max(cs1$amt\_non\_book),  
 col=ifelse(cs1$sex=="여", "pink", "lightblue"),  
 ylim=c(0, max(cs1$unitprice\_book)\*1.05),  
 sub="size: 서적이외 상품 구매액")

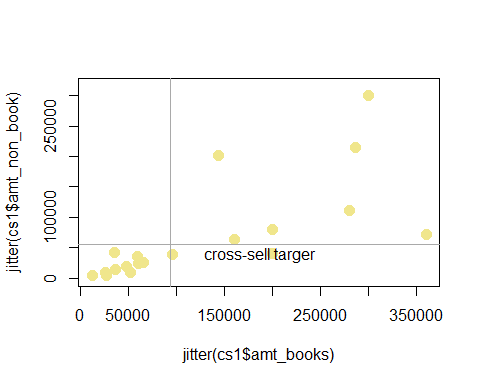


결과 : 구매서적금액이 10만원 이하, 기타상품금액이 5만원 정도 하는 고객들 위주로 집중 타겟 (구매서적 보다 기타상품 구매율이 낮은 고객층)

# 서적 구매는 많으나 기타 상품 구매가 약한 집단을 선정해

# 집중적 cross-selling 노력 기울이는 것이 필요해보인다.

plot(jitter(cs1$amt\_books),  
 jitter(cs1$amt\_non\_book),  
 pch=19,  
 col="khaki",  
 cex=1.5,  
 ylim=c(0, max(cs1$amt\_non\_book)\*1.05))  
  
abline(h=median(cs1$amt\_non\_book)\*1.5, col="darkgrey")  
abline(v=median(cs1$amt\_books)\*1.5, col="darkgrey")  
  
text(median(cs1$amt\_books)\*1.5 \*2,  
 median(cs1$amt\_non\_book)\*1.5 \*0.7, "cross-sell targer")



# 대상 집단 조건 - 시각적으로 설정했던 기준선 영역에 해당하는 고객 리스트 추출

target <- cs1[cs1$amt\_books > median(cs1$amt\_books)\*1.5 &  
 cs1$amt\_non\_book < median(cs1$amt\_non\_book)\*1.5 ,]  
  
nrow(target)

[1] 2

paste("size of target = ",  
 as.character(100\* nrow(target)/ nrow(cs1)),  
 " % of customer base")

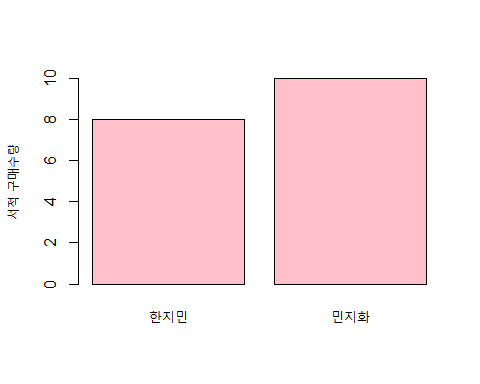
[1] “size of target = 10 % of customer base”

# => 타겟층 10%

# 선정된 집단의 프로파일 시각적으로 확인

# 서적 구매수량과 성별 분포 확인 (여성은 pink)

barplot(target$num\_books,  
 names.arg=target$cust\_name,  
 col=ifelse(target$sex=="여", "pink", "lightblue"),  
 ylab="서적 구매수량")



# 한지민, 민지화

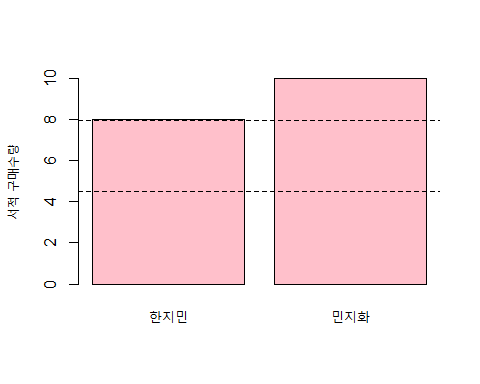
# 전체고객의 평균/중위수 서적구매수량과 비교

결과 : 중위수와 평균에 비해서 서적구매수량이 많은 고객 성별은 여성 고객.

집중 타겟 고객 : 한지민, 민지화

기타 상품 중 여성 선호 상품을 찾아 제안하는 방식으로 cross-sell 캠페인 진행 필요해보임

barplot(target$num\_books,  
 names.arg=target$cust\_name,  
 col=ifelse(target$sex=="여", "pink", "lightblue"),  
 ylab="서적 구매수량")  
  
abline(h=mean(cs1$num\_books), lty=2)  
abline(h=median(cs1$num\_books), lty=2)



#군집분석을 활용한 고객세분화

cs2 <- cs1[, names(cs1) %in% c("days\_purchase", "recency",  
 "num\_books", "amt\_books",  
 "unitprice\_book", "amt\_non\_book",  
 "num\_genre", "membership\_period")]  
  
  
kmm1 <- kmeans(cs2, 3)  
  
table(kmm1$cluster)

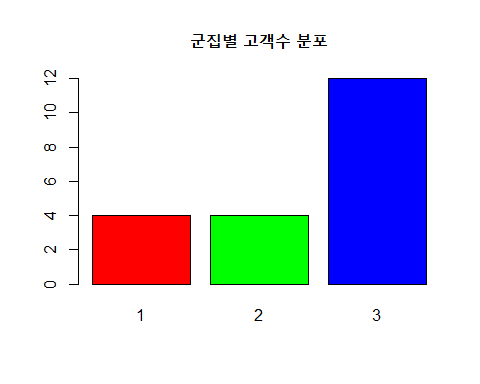
1 2 3 4 4 12

# 고객집단을 표시할 색상을 임의로 지정

# 번호순의 색상 이름 벡터 생성

# 각 고객의 소속 집단이 어디인가에 따라 색상 표시

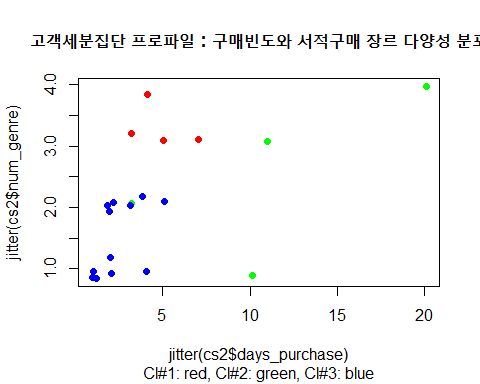
cols <- c("red", "green", "blue")  
barplot(table(kmm1$cluster),  
 names.arg=names(table(kmm1$cluster)),  
 col=cols,  
 main="군집별 고객수 분포")



# 구매빈도와 서적구매 장르 다양성 분포

결과 : 서적구매 장르가 많으면 구매빈도수가 낮다. 장르를 간소화 시켜 빈도수를 높이는 방법이 효과적일 것이다.

plot(jitter(cs2$days\_purchase),  
 jitter(cs2$num\_genre),  
 col=cols[kmm1$cluster],  
 pch=19,  
 main="고객세분집단 프로파일 : 구매빈도와 서적구매 장르 다양성 분포",  
 sub="Cl#1: red, Cl#2: green, Cl#3: blue")



#구매 빈도와 서적구매량대비 장르 다양성 분포

결과 : 구매빈도가 낮을수록 서적구매량대비 구매장르수 비율이 높다.

따라서 구매빈도가 낮은 고객층에겐 다양한 장르의 서적을 추천하는 것이 매출을 올리는데 효과적일 것이다.

plot(jitter(cs2$days\_purchase),  
 jitter(cs2$num\_genre/cs2$num\_books),  
 col=cols[kmm1$cluster],  
 pch=19,  
 main="고객세분집단 프로파일 : 구매빈도와 서적구매 장르 다양성 분포",  
 sub="Cl#1: red, Cl#2: green, Cl#3: blue",  
 ylab="서적구매량대비 구매장르수 비율")

