전처리

1. 아웃라이어는 15.1~22.2분기 사이에 1회라도 임대료보다 매출이 적은 분기가 있으면 처리하기로 했습니다.
   다만 매출 데이터는 BC카드+신한카드+국민카드 데이터를 통한 추정 자료입니다.
   따라서 임대료를 하기와 같이 계산하여 임대료 컬럼을 생성하여 비교했습니다.
   임대료 = **임대료 X 결제 방법 중 카드결제 비율 X 전체 카드사 비율 중상기 3개 카드사의 비율**
2. 점포수가 적을 경우 bias가 생길 수 있으므로, 점포수가 5개 이상인 지역&업태 데이터만 선택
3. 데이터가 15.1분기~20.3분기 ⇒ 총 23분기 데이터가 모두 있는 데이터만 선택

```
#임대면적 데이터 불러오기
setwd("C:/Users/ChangYong/Desktop/나노디그리/1.정규강의 학습자료/1차 프로젝트/소상공인/데이터/원본데이터")
area_1501_1612 <- read.xlsx("상권별_집합_상가_기본정보_14011612.xlsx")
area_1701_1812 <- read.xlsx("상권별_집합_상가_기본정보_17011812.xlsx")
area_1901_1912 <- read.xlsx("상권별_집합_상가_기본정보_19011912.xlsx")
area_2001_2009 <- read.xlsx("상권별_집합_상가_기본정보_20012009.xlsx")

#데이터 전처리
area_modi <- function(data,type){
  if(type==1){
    data <- data[,2:ncol(data)]
  }
  data <- as.data.frame(t(data))
  colnames(data) <- data[1,]
  data$년도 <- str_sub(string = rownames(data),1,4)
  data$분기 <- str_sub(string = rownames(data),6,6)
  vars <- c(6,7,2,3,4,5)
  data <- data[,vars] %>% slice(-1)
  return(data)
}
area_1501_1612 <- area_modi(area_1501_1612,0)
area_1701_1812 <- area_modi(area_1701_1812,1)
area_1901_1912 <- area_modi(area_1901_1912,1)
area_2001_2009 <- area_modi(area_2001_2009,1)
colnames(area_1501_1612)[3:6] <- c("임대면적_도심지역","임대면적_강남지역","임대면적_신촌마포지역","임대면적_기타")
colnames(area_1701_1812)[3:6] <- c("임대면적_도심지역","임대면적_강남지역","임대면적_신촌마포지역","임대면적_기타")
colnames(area_1901_1912)[3:6] <- c("임대면적_도심지역","임대면적_강남지역","임대면적_신촌마포지역","임대면적_기타")
colnames(area_2001_2009)[3:6] <- c("임대면적_도심지역","임대면적_강남지역","임대면적_신촌마포지역","임대면적_기타")
area <- rbind(area_1501_1612, area_1701_1812, area_1901_1912, area_2001_2009)

#도심,강남,신촌마포,기타지역으로 데이터 구분
area_1 <- area[,c(1,2,3)] #서울시 도심지역
area_2 <- area[,c(1,2,4)] #서울시 강남지역
area_3 <- area[,c(1,2,5)] #서울시 신촌마포지역
area_4 <- area[,c(1,2,6)] #서울시 기타지역
area_list_1 <- c("중구","종로구")
area_list_2 <- c("강남구","서초구")
area_list_3 <- c("서대문구","마포구")
area_list_4 <- c("강동구","강북구","강서구","관악구","광진구","구로구","금천구","노원구","도봉구","동대문구",
                 "동작구","성동구","성북구","송파구","양천구","영등포구","용산구","은평구","중랑구")

#상권분석 데이터 임대료 데이터 추가
smallbz_total_1501_2009$임대면적 <- NA
for(i in unique(smallbz_total_1501_2009$년도)){
  for(j in unique(smallbz_total_1501_2009$분기)){
    for(k in unique(smallbz_total_1501_2009$행정구역)){
      if(k %in% area_list_1){
        smallbz_total_1501_2009[smallbz_total_1501_2009$년도==i & smallbz_total_1501_2009$분기==j,"임대면적"] <- area_1[area_1$년도==
      } else if(k %in% area_list_2){
        smallbz_total_1501_2009[smallbz_total_1501_2009$년도==i & smallbz_total_1501_2009$분기==j,"임대면적"] <- area_2[area_2$년도==
      } else if(k %in% area_list_3){
        smallbz_total_1501_2009[smallbz_total_1501_2009$년도==i & smallbz_total_1501_2009$분기==j,"임대면적"] <- area_3[area_3$년도==
      } else {
        smallbz_total_1501_2009[smallbz_total_1501_2009$년도==i & smallbz_total_1501_2009$분기==j,"임대면적"] <- area_4[area_4$년도==
      }
    }
  }
}
smallbz_total_1501_2009$임대면적 <- as.numeric(smallbz_total_1501_2009$임대면적)

#카드 결제 비율 계산
sales_card <- as.data.frame(t(read.xlsx("금융감독원_금융통계정보시스템_카드사.xlsx", colNames = F)))
sales_bank <- as.data.frame(t(read.xlsx("금융감독원_금융통계정보시스템_은행사.xlsx", colNames = F)))
colnames(sales_card) <- c("년월","우리카드","국민카드","롯데카드","비씨카드","삼성카드","신한카드","하나카드","현대카드")
colnames(sales_bank) <- c("년월","은행")
sales_card <- sales_card %>% slice(-1)
sales_card <- sales_card %>% mutate(년도 = str_sub(년월,1,4), 분기 = case_when(str_sub(년월,-2,-1)==".3" ~ "1",
                                                                                str_sub(년월,-2,-1)==".6" ~ "2",
                                                                                str_sub(년월,-2,-1)==".9" ~ "3",
```

```
                                                                    str_sub(년월,-2,-1)=="12" ~ "4")) %>% select(-년월)
sales_bank <- sales_bank %>% mutate(년도 = str_sub(년월,1,4), 분기 = case_when(str_sub(년월,6,7)=="03" ~ "1",
                                                                    str_sub(년월,6,7)=="06" ~ "2",
                                                                    str_sub(년월,6,7)=="09" ~ "3",
                                                                    str_sub(년월,6,7)=="12" ~ "4")) %>% select(-년월)
sales_merge <- merge(x = sales_bank, y = sales_card, by = c("년도","분기"))

#은행,국민카드,비씨카드, 신한카드 비율 계산
sales_merge[,3:ncol(sales_merge)] <- map_df(.x = sales_merge[,3:ncol(sales_merge)],.f = as.numeric)
sales_merge <- sales_merge %>% group_by(년도,분기) %>% summarise(카드비율_비씨신한국민 = (은행+비씨카드+신한카드+국민카드)/
                        (은행+비씨카드+신한카드+국민카드+롯데카드+삼성카드+현대카드+하나카드+우리카드)) %>% as.data.frame()

sales_merge[,1:2] <- map_df(.x = sales_merge[,1:2],.f = as.factor)

#전체 결제 비율 중 카드결제 비율 데이터 생성
method <- data.frame(년도 = seq(2015,2020), 카드사용비율 = NA)
rate <- c((40.7+14.8)/100,(41.3+12.5)/100.6,(53.8+15.3)/100)
geom <- function(x,y){result = 2*(x*y)/(x+y); return(result)}
for(i in 1:6){
  if(i == 6){
    method[i,2] = (method[i-1,2]*method[i-2,2])/(2*method[i-2,2]-method[i-1,2])
  } else if(i %% 2 == 1){
    method[i,2] = rate[ceiling(i/2)]
  } else {
    method[i,2] = geom(rate[i/2],rate[i/2+1])
  }
}
#카드실적 및 카드결제 비율 데이터 merge
sales_merge <- merge(x = sales_merge,y = method, by="년도")

#임대료로 변환#분기별 전체 결제 방법 중 카드결제 비율 및 카드결제 비율 중 비씨,신한,국민카드 비율)
smallbz_total_1501_2009 <- merge(x = smallbz_total_1501_2009, y = sales_merge, by = c("년도","분기"))
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>%
                           mutate(임대료 = 3*임대료*임대면적*카드비율_비씨신한국민*카드사용비율)%>%
                           select(-c(임대면적,카드비율_비씨신한국민,카드사용비율))

#15.1분기 ~20.2분기의 22개의 분기 중 1번이라도 임대료보다 매출이 적은 경우 제외
remove_data <- smallbz_total_1501_2009 %>% mutate(년분기 = paste0(년도,"_",분기)) %>%
  filter(매출총액 < 임대료 & 년분기!="2020_3") %>%
  count(ADSTRD_CD,상권_코드,소분류)
smallbz_total_1501_2009 <- merge(x = smallbz_total_1501_2009,
                                 y = remove_data,
                                 by = c('ADSTRD_CD','상권_코드','소분류'),all.x=T)
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>% filter(is.na(n)==T) %>% select(-c(n,임대료))


#상권 및 업태별 점포수 개수가 5개 이상인 지역만 선택
jeom_num <- smallbz_total_1501_2009 %>% group_by(상권_코드,소분류)%>%
  summarise(n = mean(점포수)) %>% filter(n>=5) %>% select(-n)

#점포수 개수가 5개 이상인 지역만 선택
smallbz_total_1501_2009 <- merge(x=smallbz_total_1501_2009,y=jeom_num, by = c("상권_코드","소분류"),all.y=T)


#15.1~20.3분기 데이터가 모두 있는 상권 및 업태만 선택
selected <- smallbz_total_1501_2009 %>%
  count(상권_코드,소분류) %>%
  filter(n == 23)
smallbz_total_1501_2009 <- merge(x = smallbz_total_1501_2009,
                                 y = selected,
                                 by = c('상권_코드','소분류'),all.x=T)

smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>% filter(n == 23) %>% select(-n)
```