전처리

- 데이터 추가 확보 19.1분기~20.3분기 ⇒ 15.1분기~20.3분기
  이상치 처리 : 매출액이 마이너스(-)인 년도,분기,상권코드 데이터 삭제

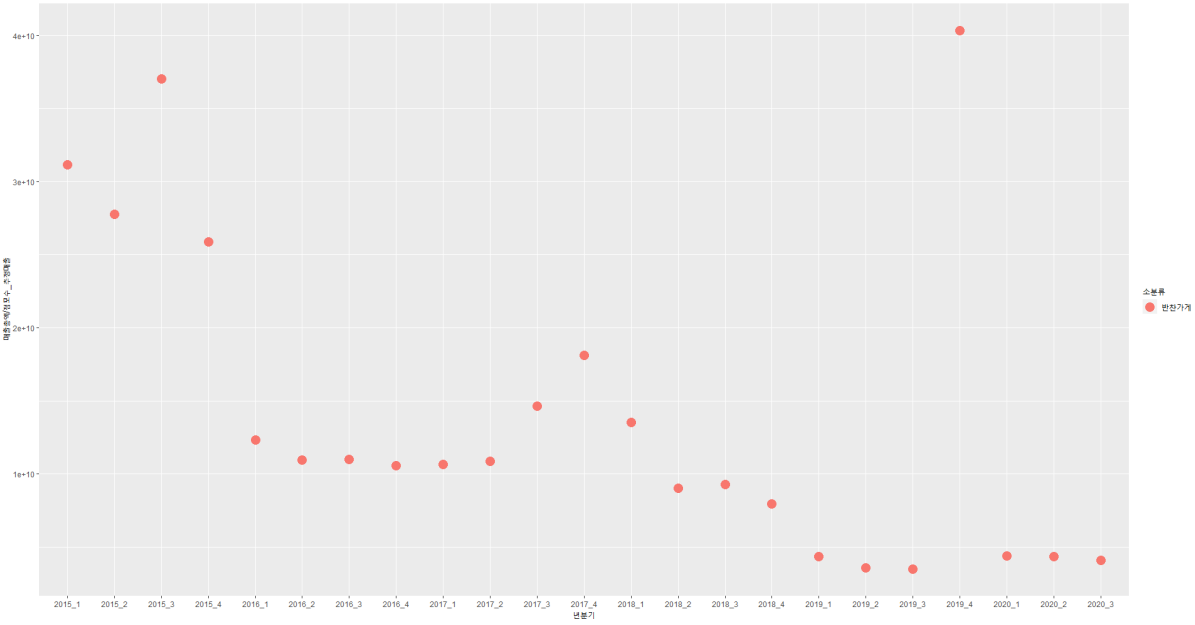
```
outlier_minus <- data.frame()
for(i in 36:48){
  outlier_minus <- rbind(outlier_minus, smallbz_sales[smallbz_sales[,i]<0,c(1,2,5,7)])

}
outlier_minus <- outlier_minus %>% distinct(상권_코드,서비스_업종_코드)
outlier_minus$사용여부 <- 1

smallbz_sales <- merge(x = smallbz_sales,y = outlier_minus, by = c("상권_코드","서비스_업종_코드"), all.x = T)
smallbz_sales <- smallbz_sales %>% filter(is.na(사용여부) == T)
smallbz_sales <- smallbz_sales[,-81]
```

- 15.1~20.3분기 데이터가 모두 있는 상권만 선택

```
selected <- smallbz_total_1501_2009 %>%
              count(ADSTRD_CD,상권_코드,소분류) %>%
              filter(n == 23)
smallbz_total_1501_2009 <- merge(x = smallbz_total_1501_2009,
                                 y = selected,
                                 by = c('ADSTRD_CD','상권_코드','소분류'),all.x=T)
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>%
                           filter(n==23)%>%
                           select(-n)
```

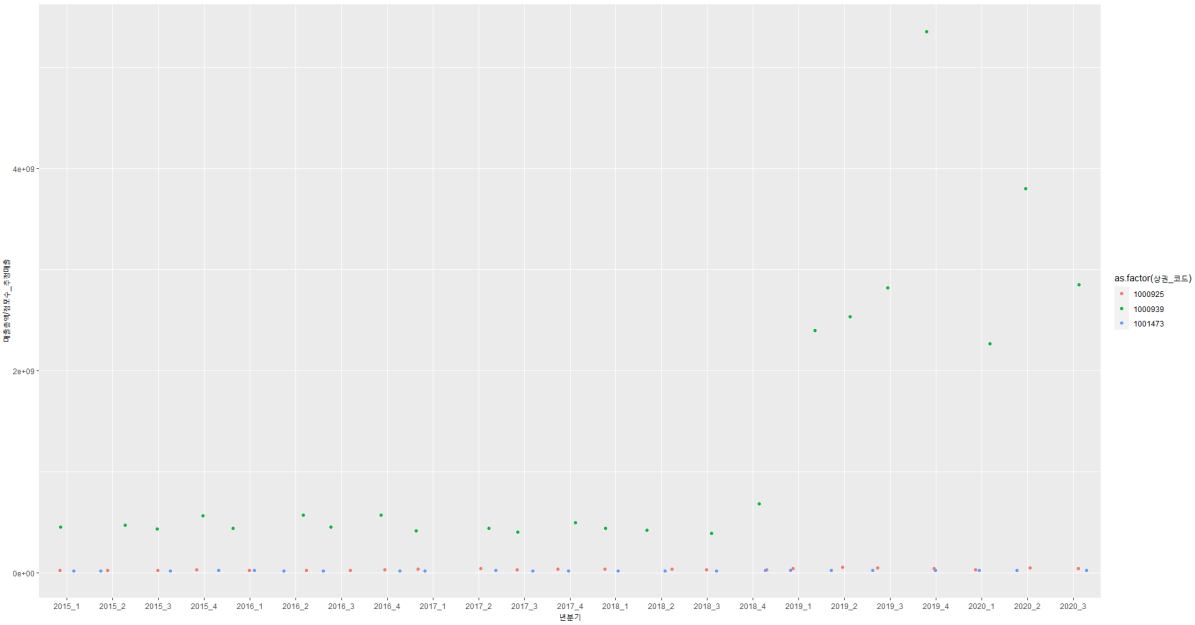
이상치 데이터가 있는 송파구 가락1동 반찬가게 데이터 모두 삭제

```
smallbz_total_1501_2009 %>%
  filter(행정구역 == '송파구' & 소분류 %in% c("반찬가게") & 행정동명 == "가락1동") %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  ggplot(aes(x = 년분기, y=매출총액/점포수_추정매출,color = 소분류),)+
  geom_point(size = 5)
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>%
                              filter(행정동명 != "가락1동" | !소분류 %in% c("반찬가게"))
```



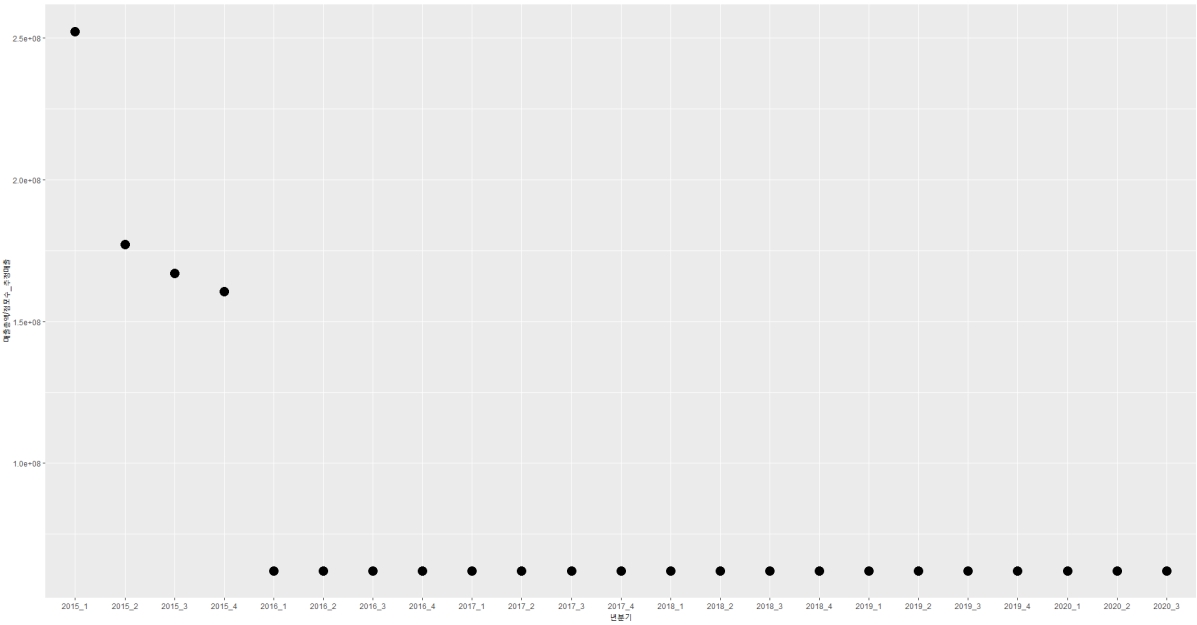
- 강남구 청담동 문구 데이터 중 이상치라 판단되는 상권코드 1000939의 데이터 제거

```
smallbz_total_1501_2009 %>%
  filter(행정구역 =="강남구" & 소분류 =="문구" & 행정동명 == '청담동') %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  ggplot(aes(x = 년분기, y = 매출총액/점포수_추정매출,color = as.factor(상권_코드)))+
  geom_point(position = position_jitter())
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>%
                           filter(행정동명 != "청담동" | 소분류 !="문구" | 상권_코드!=1000939)
```



- 동일한 데이터가 입력되어 있는, 중랑구 가전제품수리 데이터 삭제

```
smallbz_total_1501_2009 %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  filter(행정구역 =="중랑구" & 소분류 =="가전제품수리") %>%
  arrange(년도,분기) %>%
  ggplot(aes(x = 년분기, y=매출총액/점포수_추정매출))+geom_point(size = 5)
```



점포수 개수 조정

```
setwd("C:/Users/ChangYong/Desktop/나노디그리/1.정규강의 학습자료/1차 프로젝트/소상공인/데이터/원본데이터")
data1 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2015년.csv")
data2 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2016년.csv")
data3 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2017년.csv")
data4 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2018년.csv")
data5 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2019년.csv")
data6 <- read.csv("서울시_우리마을가게_상권분석서비스(상권-점포)_2020년.csv")
jeompo <- rbind(data1,data2,data3,data4,data5,data6)

#점포 데이터에 행정동 및 행정구 컬럼 추가
jeompo <- merge(x = jeompo,
                y = sangkwon_loc,by.x = '상권_코드',
                by.y = 'TRDAR_CD', all.x=T)
jeompo <- merge(x = jeompo,
                y = sangkwon_gu,
                by.x = 'ADSTRD_CD',
                by.y = '행자부행정동코드', all.x=T)

#점포 데이터 컬럼 최적화
vars <- c(2,3,4,17,9,10,11,13,15)
jeompo <- jeompo[,vars]
colnames(jeompo)[2:9] <- c("년도","분기","행정구역","소분류","점포수","점포수_유사업종","점포수_개업","점포수_폐업")

#업태(소분류)의 문자열 변환
jeompo$소분류 <- str_replace_all(jeompo$소분류,"/","&")

#매출데이터와 결합
smallbz_total_1501_2009 <- merge(x = smallbz_total_1501_2009,
                                 y= jeompo,
                                 by=c("상권_코드","년도","분기","행정구역","소분류"),
                                 all.x=T)

#점포수 처리, NA=>0
```

```
na_0 <- function(x){
  x <- ifelse(is.na(x)==T,0,x)
  return(x)
}
vars <- c(13,28,29,30,31)

#결합 후 점포 컬럼에 NA가 있을 경우 0으로 변환
for(i in vars){
  smallbz_total_1501_2009[,i] <- na_0(smallbz_total_1501_2009[,i])
}

#점포수는 점포수 추정 매출이 유사업종 수보다 큰 경우 점포수_추정매출을, 반대면 점포수_유사업종
#점포가 분기 중간에 폐업한 경우 점포수에서 제외됨
#매출액은 있는데 두 점포수 모두 없는 경우 1개 적용
smallbz_total_1501_2009 <- smallbz_total_1501_2009 %>%
  mutate(점포수 = ifelse(test = 점포수_추정매출 > 점포수_유사업종,yes = 점포수_추정매출,
                       no = ifelse(점포수_유사업종 >= 1,yes = 점포수_유사업종, no = 1)))
```

- 대분류(업종) & 소분류(업태) 사이 중분류 추가

⇒ 대분류 3개 / 중분류 20개 / 소분류 63 개

```
MD_category <- list(오락관련서비스 = c("PC방","노래방","볼링장","전자게임장"),
                    개인및소비용품수리 = c("가전제품수리", "미용실","자동차수리","통신기기수리"),
                    숙박 = c("고시원","여관"),
                    스포츠 = c('골프연습장','당구장','스포츠클럽'),
                    개인 = c('네일숍','세탁소','자동차미용','피부관리실'),
                    교육 = c('스포츠 강습','예술학원','외국어학원','일반교습학원'),
                    보건 = c('일반의원','치과의원','한의원'),
                    부동산 = c("부동산중개업"),
                    기타상품전문 = c('시계및귀금속','안경','애완동물','의료기기','의약품','화장품','화초','예술품'),
                    기타생활용품 = c('가구','인테리어','조명용품','철물점','악기'),
                    무점포 = c('전자상거래업'),
                    오락및여가용품 = c('문구','서적','완구','운동&경기용품','자전거 및 기타운송장비'),
                    음식료품및담배 = c('미곡판매','반찬가게','수산물판매','슈퍼마켓','육류판매','청과상'),
                    의류 = c('가방','섬유제품','신발','일반의류','한복점','유아의류'),
                    전자제품 = c('가전제품','컴퓨터및주변장치판매','핸드폰'),
                    종합소매 = c('편의점'),
                    기타음식점 = c('분식전문점','제과점','치킨전문점','패스트푸드점'),
                    비알콜음료점 = c("커피-음료"),
                    일반음식점 = c('양식음식점','일식음식점','중식음식점','한식음식점'),
                    주점업 = c('호프-간이주점'))

smallbz_total_1501_2009$중분류 <- 0
for(i in 1:length(MD_category)){
  smallbz_total_1501_2009[smallbz_total_1501_2009$소분류 %in% MD_category[[i]],]$중분류 <- names(MD_category[i])
  cat(round(i/length(MD_category),digits=4L)*100,"% 완료\n")
}
```