eda

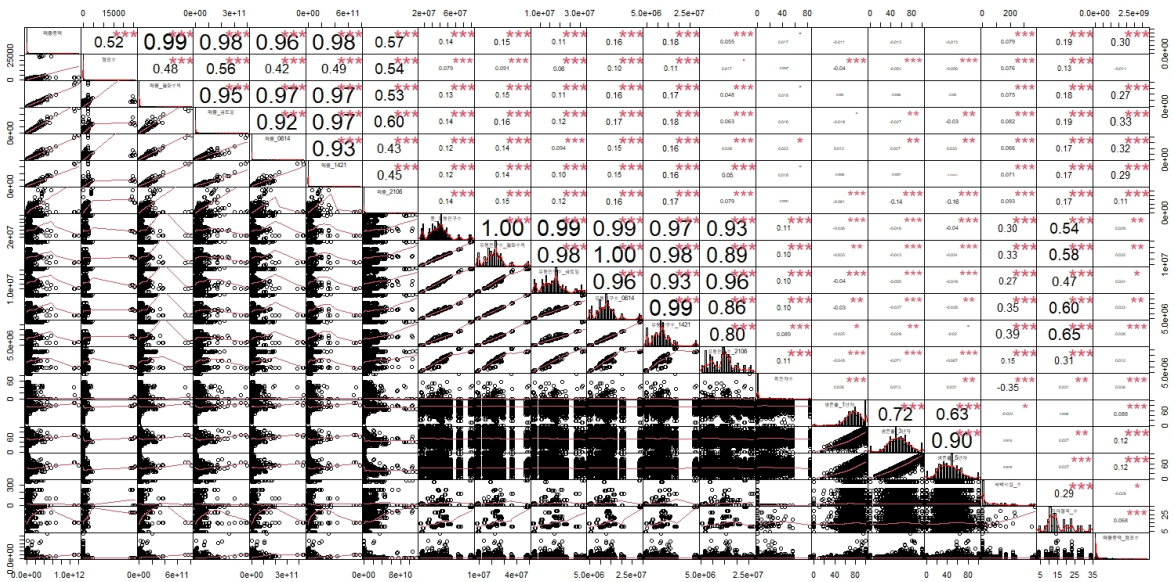
eda를 위한 trainset 데이터 셋 정리

```
setwd("C:/Users/ChangYong/Desktop/나노디그리/1.정규강의 학습자료/1차 프로젝트/소상공인/데이터")
load("dataset.rda")
summary(smallbz_total)
smallbz_total_dummy <- smallbz_total %>% filter(점포수 !=0)
train_dummy <- trainset %>% filter(점포수 != 0)
test_dummy <- testset %>% filter(점포수 !=0)

#거리&상권별 값이 다른 컬럼은 sum / 거리&상권에 동일한 값을 적용한 컬럼은 mean
train_dummy <- train_dummy %>%
  group_by(년도,분기,행정구역,대분류,소분류,년분기) %>%
  summarise(매출총액 = sum(매출총액),
                점포수 = sum(점포수),
                매출_월화수목 = sum(매출_월화수목),
                매출_금토일 = sum(매출_금토일),
                매출_0614 = sum(매출_0614),
                매출_1421 = sum(매출_1421),
                매출_2106 = sum(매출_2106),
                총_유동인구수 = mean(총_유동인구수),
                유동인구수_월화수목 = mean(유동인구수_월화수목),
                유동인구수_금토일 = mean(유동인구수_금토일),
                유동인구수_0614 = mean(유동인구수_0614),
                유동인구수_1421 = mean(유동인구수_1421),
                유동인구수_2106 = mean(유동인구수_2106),
                확진자수 = mean(확진자수),
                생존률_1년차 = mean(생존률_1년차),
                생존률_3년차 = mean(생존률_3년차),
                생존률_5년차 = mean(생존률_5년차),
                숙박시설_수 = mean(숙박시설_수),
                지하철역_수 = mean(지하철역_수)) %>%
  mutate(매출총액_점포수 = 매출총액/점포수) %>%
  as.data.frame()

#매출 및 유동인구수 scale 조정 및 데이터 정규분포화를 위한 자연로그 적용
vars <- c(7,9:19,26)
train_dummy_log <- train_dummy
train_dummy_log[,vars] <- log(train_dummy_log[,vars])
for(i in vars){
  train_dummy_log[,i] <- ifelse(is.infinite(train_dummy_log[,i])==T,0,train_dummy_log[,i])
}
```

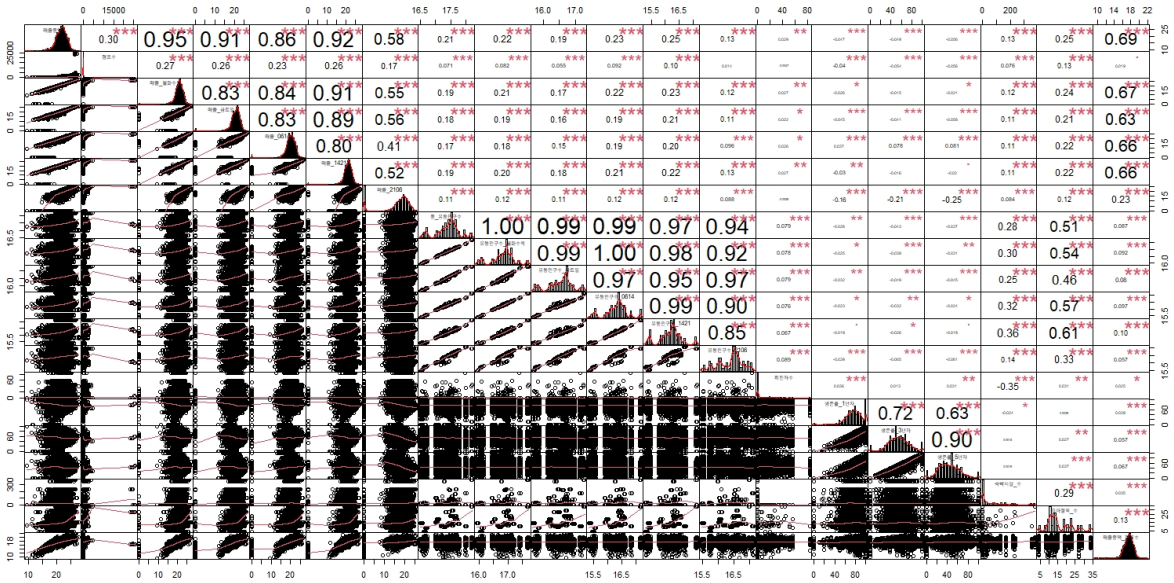
log 적용 전 값이 0인 경우 자연로그를 취하면 무한대가 되는데 이걸 0으로 처리해도 무방한지..

train_dummy 자연로그 적용&미적용 데이터 상관관계 분석

자연로그 미적용



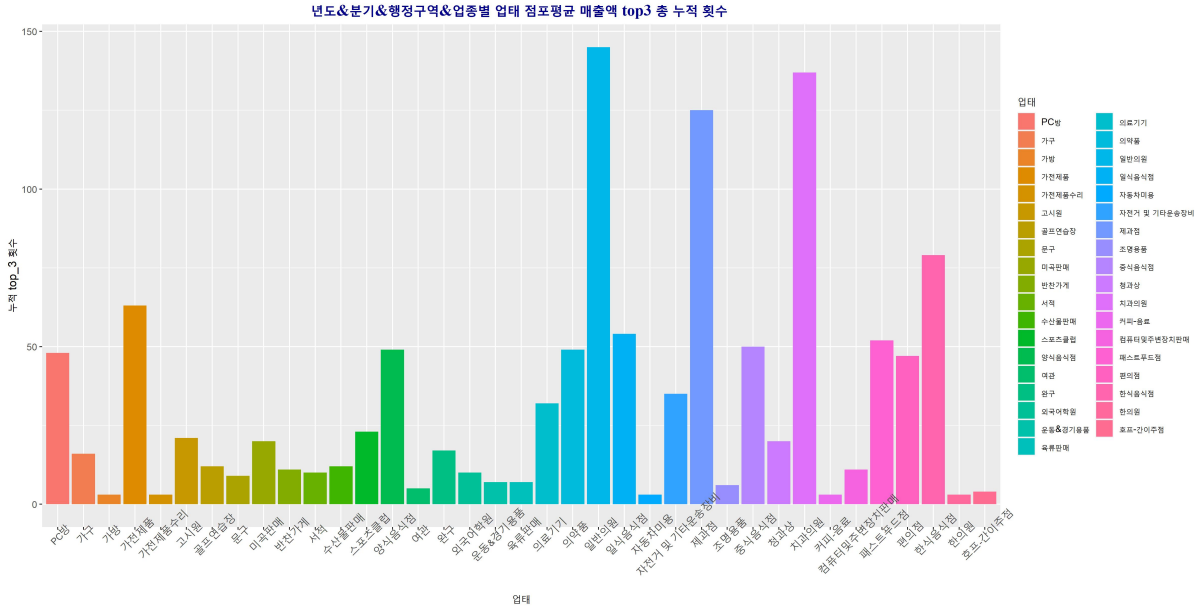
*자연로그 적용



매출 데이터의 경우 분기 및 업태별 매출액이 다르게 표기되어 있으나, 유동인구 & 코로나 확진자수 & 생존율 & 숙박시설 & 지하철 개수 데이터는 분기별로만 데이터가 있어서 상대적으로 데이터가 부족한데, 상관관계가 작아도 사용해도 괜찮은지..

행정구역&년도&분기&업종별 업태 점포 평균 매출액 top_3 누적횟수

```
train_dummy %>%
  group_by(년도,분기,행정구역,대분류,소분류) %>%
  summarise(평균매출액 = mean(매출총액_점포수)) %>%
  arrange(년도,분기,행정구역,대분류,desc(평균매출액)) %>%
  top_n(3) %>%
  ungroup() %>% count(행정구역,대분류,소분류) %>%
  group_by(행정구역,대분류) %>% arrange(행정구역,대분류,desc(n)) %>%
  top_n(3) %>%
  ggplot(aes(x = 소분류, y = n, fill = 소분류)) +
  geom_bar(stat = "identity")+
  ggtitle("년도&분기&행정구역&업종별 업태 점포평균 매출액 top3 총 누적 횟수 ")+
  theme(axis.text.x = element_text(angle = 45),
        plot.title = element_text(family = "serif", face = "bold",
                                  hjust = 0.5, size = 15,color = "darkblue"))+
  labs(fill = "업태",x ="업태", y="누적 top_3 횟수")+
  ggsave(path = eda_path,filename = "sales_top3.jpg")
```



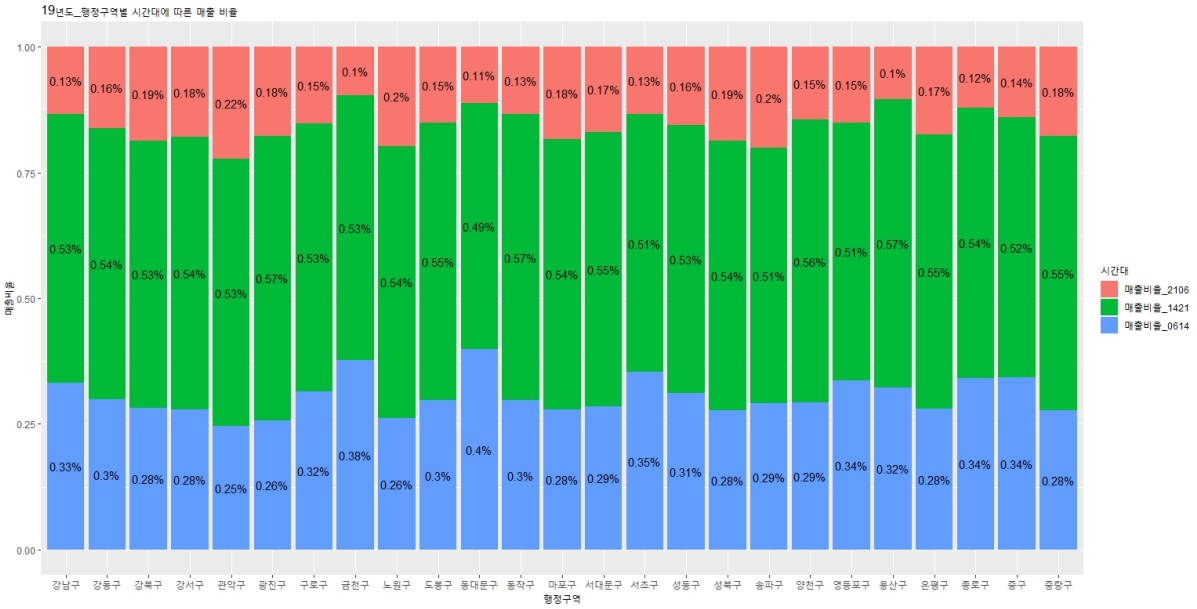
업태 기준으로, 일반의원(지역병원), 치과, 제과점이 6분기 동안 최대 top_3 누적

행정구역별 시간대에 따른 매출 비율

```
train_dummy %>%
  group_by(행정구역) %>%
  filter(년도 == 2020) %>%  #2019 / 2020
  summarise(매출총액 = sum(매출총액),
            매출_0614 = sum(매출_0614),
            매출_1421 = sum(매출_1421),
            매출_2106 = sum(매출_2106)) %>%
  mutate(매출비율_0614 = 매출_0614/매출총액,
              매출비율_1421 = 매출_1421/매출총액,
              매출비율_2106 = 매출_2106/매출총액) %>%
  select(행정구역,매출비율_0614,매출비율_1421,매출비율_2106) %>%
```

```
gather(시간대,매출비율,매출비율_0614:매출비율_2106) %>%
arrange(행정구역,시간대,매출비율) %>%
group_by(행정구역) %>%
mutate(시간대 = factor(x = 시간대, levels = c("매출비율_2106","매출비율_1421","매출비율_0614")),
       매출비율_cumsum = cumsum(매출비율)-(0.5*매출비율)) %>%
ggplot(aes(x = 행정구역, y = 매출비율, fill = 시간대))+
geom_bar(stat = "identity")+
geom_text(aes(label = paste0(round(매출비율,digits=2L),"%"), y=매출비율_cumsum))
```

2019년



2020년