EDA

매출 총액을 평균으로 할 지, 중앙값으로 할 지 정하기

```
data <- smallbz_total_1501_2009 %>%
  mutate(년분기 = as.factor(paste0(년도,"_",분기))) %>%
  filter(년분기 != "2020_3") %>%
  group_by(년도,분기,행정구역,대분류,중분류,소분류) %>%
  summarise(매출총액_mean = mean(log(매출총액/점포수)),
                매출총액_median = median(log(매출총액/점포수)))

#매출총액 평균 및 중앙값 박스플랏
ggplot(data = data, aes(x = 행정구역, y = 매출총액_mean, fill = 행정구역))+geom_boxplot(alpha = 0.7)+
  theme(axis.text.x = element_text(angle = 45, size = 10, face="bold"),
        plot.title = element_text(hjust = 0.5,size = 20, face="bold"))+ggtitle("행정구별 매출총액_mean 박스플랏")

ggplot(data = data, aes(x = 행정구역, y = 매출총액_median, fill = 행정구역))+geom_boxplot(alpha = 0.7)+
  theme(axis.text.x = element_text(angle = 45, size = 10, face="bold"),
        plot.title = element_text(hjust = 0.5,size = 20, face="bold"))+ggtitle("행정구별 매출총액_median 박스플랏")
```

Aa 이름	생성일	태그
0. 행정구별 매출총액_mean	@2021년 1월 24일 오후 10:42	
0. 행정구별 매출총액_median	@2021년 1월 24일 오후 10:42	

임대료 기준, 점포수 기준, 행분기 데이터 유무 기준으로 아웃라이어를 최대한 줄인 상태에서 나머지 데이터를 모두 고려하여 결과를 예측할 예정이므로 평균값 mean으로 진행

EDA를 위해 1차 데이터셋 정리

```
dataset <- smallbz_total_1501_2009 %>%
  group_by(년도,분기,행정구역,대분류,중분류,소분류) %>%
  summarise(매출비율_월 = mean(월요일_매출_금액/매출총액),
            매출비율_화 = mean(화요일_매출_금액/매출총액),
            매출비율_수 = mean(수요일_매출_금액/매출총액),
            매출비율_목 = mean(목요일_매출_금액/매출총액),
            매출비율_금 = mean(금요일_매출_금액/매출총액),
            매출비율_토 = mean(토요일_매출_금액/매출총액),
            매출비율_일 = mean(일요일_매출_금액/매출총액),
            매출비율_0006 = mean(시간대_00.06_매출_금액/매출총액),
            매출비율_0611 = mean(시간대_06.11_매출_금액/매출총액),
            매출비율_1114 = mean(시간대_11.14_매출_금액/매출총액),
            매출비율_1417 = mean(시간대_14.17_매출_금액/매출총액),
            매출비율_1721 = mean(시간대_17.21_매출_금액/매출총액),
            매출비율_2124 = mean(시간대_21.24_매출_금액/매출총액),
            매출비율_남성 = mean(남성_매출_금액/매출총액),
            매출비율_여성 = mean(여성_매출_금액/매출총액),
            매출비율_10대 = mean(연령대_10_매출_금액/매출총액),
            매출비율_20대 = mean(연령대_20_매출_금액/매출총액),
            매출비율_30대 = mean(연령대_30_매출_금액/매출총액),
            매출비율_40대 = mean(연령대_40_매출_금액/매출총액),
            매출비율_50대 = mean(연령대_50_매출_금액/매출총액),
            매출비율_60대이상 = mean(연령대_60_이상_매출_금액/매출총액),
            생존율_1년차 = mean(생존률_1년차),
            생존율_3년차 = mean(생존률_3년차),
            생존율_5년차 = mean(생존률_5년차),
            소득분위 = as.factor(round(mean(as.numeric(소득분위)),digits = 0)),
            집객시설수 = mean(집객시설_수),
            매출총액 = mean(log(매출총액/점포수)),
            총유동인구= mean(log(총_유동인구수)),
            총매출건수 = mean(log(총매출건수))) %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  as.data.frame()

#ln() 후 -inf를 0으로 변경
dataset[,"총유동인구"] <- ifelse(is.infinite(dataset[,"총유동인구"])==T, 0, dataset[,"총유동인구"])
```

입력변수 : 요일

```
######## 1. 년&분기별 요일에 따른 매출비율 비교 #line
vars <- dataset %>% distinct(년분기) %>% arrange(년분기)
text <- dataset %>%
  filter(년분기 !="2020_3") %>% group_by(년분기) %>%
  summarise(월요일 = mean(매출비율_월),
                화요일 = mean(매출비율_화),
                수요일 = mean(매출비율_수),
                목요일 = mean(매출비율_목),
                금요일 = mean(매출비율_금),
                토요일 = mean(매출비율_토),
                일요일 = mean(매출비율_일)) %>%
  gather(key = 요일,value = 매출비율, 월요일:일요일) %>%  group_by(요일) %>%
  mutate(num = row_number(),요일 = factor(요일, levels = c('월요일','화요일','수요일','목요일','금요일','토요일','일요일'))) %>%
  filter(num ==22) %>% select(-년분기)

dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기) %>%
  summarise(월요일 = mean(매출비율_월),
                화요일 = mean(매출비율_화),
                수요일 = mean(매출비율_수),
                목요일 = mean(매출비율_목),
                금요일 = mean(매출비율_금),
                토요일 = mean(매출비율_토),
                일요일 = mean(매출비율_일)) %>%
  gather(key = 요일,value = 매출비율, 월요일:일요일) %>%
  group_by(요일) %>%
  mutate(num = row_number(),요일 = factor(요일, levels = c('월요일','화요일','수요일','목요일','금요일','토요일','일요일'))) %>%
  ggplot(aes(x = num, y =매출비율, group = 요일, col = 요일))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 요일별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text,mapping = aes(x = num+0.5, y = 매출비율, label = 요일), fontface = "bold", size = 6)

######## 2. 요일별 매출 비율 #box plot
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 요일,value = 비율, 매출비율_월:매출비율_일) %>%
  mutate(요일 = factor(요일, levels = c('매출비율_월','매출비율_화','매출비율_수','매출비율_목',
                                '매출비율_금','매출비율_토','매출비율_일'))) %>%
  ggplot(aes(x = 요일, y = 비율, fill = 요일))+geom_boxplot()+
  ggtitle("요일별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

######## 3. 요일별 매출 비율
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 요일,value = 비율, 매출비율_월:매출비율_일) %>%
  mutate(요일 = factor(요일, levels = c('매출비율_월','매출비율_화','매출비율_수','매출비율_목',
                                '매출비율_금','매출비율_토','매출비율_일'))) %>%
  ggplot(aes(x = 비율, y = 요일, fill = 요일))+geom_density_ridges(alpha = 0.4)+
  theme_classic()+ggtitle("요일별 매출비율")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))

######## 4. 요일별 산점도도
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 요일,value = 비율, 매출비율_월:매출비율_일) %>%
  mutate(요일 = factor(요일, levels = c('매출비율_월','매출비율_화','매출비율_수','매출비율_목',
                                '매출비율_금','매출비율_토','매출비율_일'))) %>%
  ggplot(aes(x = 비율, y = 매출총액, col = 요일))+geom_point(alpha = 0.4)+
  theme_classic()+ggtitle("요일별 매출비율")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
1-1. 년분기별 요일에 따른 매출비율	@2021년 1월 24일 오후 10:49	
1-2. 요일별 매출비율 박스플랏	@2021년 1월 24일 오후 10:49	
1-3. 요일별 매출비율 밀도	@2021년 1월 24일 오후 10:49	
1-4. 요일별 매출총액 산점도	@2021년 1월 24일 오후 11:21	

입력변수 : 시간대

```
######## 1. 년&분기별 시간대에 따른 매출 비율
text <- dataset %>%
  filter(년분기 =="2020_2") %>%
  group_by(년분기) %>%
  summarise(매출비율_0006 = mean(매출비율_0006),
                매출비율_0611 = mean(매출비율_0611),
```

```
                  매출비율_1114 = mean(매출비율_1114),
                  매출비율_1417 = mean(매출비율_1417),
                  매출비율_1721 = mean(매출비율_1721),
                  매출비율_2124 = mean(매출비율_2124)) %>%
  gather(시간대,비율,매출비율_0006:매출비율_2124)

dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기) %>%
  summarise(매출비율_0006 = mean(매출비율_0006),
                  매출비율_0611 = mean(매출비율_0611),
                  매출비율_1114 = mean(매출비율_1114),
                  매출비율_1417 = mean(매출비율_1417),
                  매출비율_1721 = mean(매출비율_1721),
                  매출비율_2124 = mean(매출비율_2124)) %>%
  gather(key = 시간대,value = 매출비율, 매출비율_0006:매출비율_2124) %>%
  group_by(시간대) %>%
  mutate(num = row_number(),시간대 = factor(시간대)) %>%
  ggplot(aes(x = num, y =매출비율, group = 시간대, col = 시간대))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 시간대별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text, aes(x = rep(22.8,nrow(text)), y = 비율, label=시간대), size = 4, fontface = "bold")

######## 2. 시간대별 매출 비율
colnames(dataset)
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 시간대,value = 비율, 매출비율_0006:매출비율_2124) %>%
  ggplot(aes(x = 시간대, y = 비율, fill = 시간대))+geom_boxplot()+
  ggtitle("시간대별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

######## 3. 시간대별 매출 비율
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 시간대,value = 매출, 매출비율_0006:매출비율_2124) %>%
  mutate(비율 = 매출/매출총액) %>% select(시간대,비율) %>%
  mutate(시간대 = factor(시간대, levels = c("매출비율_0006","매출비율_0611","매출비율_1114",
                                    "매출비율_1417","매출비율_1721","매출비율_2124"))) %>%
  ggplot(aes(x = 비율, y = 시간대, fill = 시간대))+geom_density_ridges(alpha = 0.4)+
  theme_classic()+ggtitle("요일별 매출비율")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))

########  4. 시간대별 산점도
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 시간대,value = 매출, 매출비율_0006:매출비율_2124) %>%
  mutate(비율 = 매출/매출총액) %>%
  mutate(시간대 = factor(시간대, levels = c("매출비율_0006","매출비율_0611","매출비율_1114",
                                    "매출비율_1417","매출비율_1721","매출비율_2124"))) %>%
  ggplot(aes(x = 비율, y = 매출총액, col = 시간대))+geom_point(alpha = 0.4)+
  theme_classic()+ggtitle("시간대별 매출총액 산점도")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
3-1. 년분기별 시간대에 따른 매출비율	@2021년 1월 24일 오후 10:52	
3-2. 시간대별 따른 매출비율	@2021년 1월 24일 오후 10:52	
3-3. 시간대별 따른 매출비율	@2021년 1월 24일 오후 10:52	
3-4. 시간대 산점도	@2021년 1월 24일 오후 11:21	

입력변수 : 연령대

```
######## 1. 년&분기별 연령대에 따른 매출 비율
text <- dataset %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  filter(년분기 =="2020_2") %>%
  group_by(년분기) %>%
  summarise(매출비율_10대 = mean(매출비율_10대),
                  매출비율_20대 = mean(매출비율_20대),
                  매출비율_30대 = mean(매출비율_30대),
                  매출비율_40대 = mean(매출비율_40대),
                  매출비율_50대 = mean(매출비율_50대),
                  매출비율_60대이상 = mean(매출비율_60대이상)) %>%
  gather(key = 연령대,value = 매출비율, 매출비율_10대:매출비율_60대이상)
```

```
dataset %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기) %>%
  summarise(매출비율_10대 = mean(매출비율_10대),
                매출비율_20대 = mean(매출비율_20대),
                매출비율_30대 = mean(매출비율_30대),
                매출비율_40대 = mean(매출비율_40대),
                매출비율_50대 = mean(매출비율_50대),
                매출비율_60대이상 = mean(매출비율_60대이상)) %>%
  gather(key = 연령대,value = 매출비율, 매출비율_10대:매출비율_60대이상) %>%
  group_by(연령대) %>%
  mutate(num = row_number(),연령대 = factor(연령대)) %>%
  ggplot(aes(x = num, y =매출비율, group = 연령대, col = 연령대))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 연령대별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text, mapping = aes(x = rep(22.8,nrow(text)),y = 매출비율, label = 연령대), size = 4, fontface = "bold")

######## 2. 연령대별 매출 비율
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 연령대,value = 비율, 매출비율_10대:매출비율_60대이상) %>%
  ggplot(aes(x = 연령대, y = 비율, fill = 연령대))+geom_boxplot()+
  ggtitle("연령대별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

######## 3. 연령대별 매출 비율
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 연령대,value = 매출, 매출비율_10대:매출비율_60대이상) %>%
  mutate(비율 = 매출/매출총액) %>% select(연령대,비율) %>%
  mutate(시간대 = factor(연령대, levels = c("매출비율_10대","매출비율_20대","매출비율_30대",
                                      "매출비율_40대","매출비율_50대","매출비율_60대_이상"))) %>%
  ggplot(aes(x = 비율, y = 연령대, fill = 연령대))+geom_density_ridges(alpha = 0.4)+
  theme_classic()+ggtitle("연령대별 매출비율")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))

########  4. 연령대별 산점도
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 연령대,value = 매출, 매출비율_10대:매출비율_60대이상) %>%
  mutate(비율 = 매출/매출총액) %>%
  mutate(시간대 = factor(연령대, levels = c("매출비율_10대","매출비율_20대","매출비율_30대",
                                      "매출비율_40대","매출비율_50대","매출비율_60대_이상"))) %>%
  ggplot(aes(x = 비율, y = 매출총액, col = 연령대))+geom_point(alpha = 0.4)+
  theme_classic()+ggtitle("연령대별 산점도")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
4-1. 년분기별 연령대에 따른 매출비율	@2021년 1월 24일 오후 10:55	
4-2. 연령대별 매출비율	@2021년 1월 24일 오후 10:55	
4-3. 연령대별 매출비율	@2021년 1월 24일 오후 10:55	
4-4. 연령대별 산점도	@2021년 1월 24일 오후 11:22	

입력변수 : 대분류

```
######## 1. 년&분기 별 대분류에 따른 매출비율 비교
dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기, 대분류) %>%
  summarise(매출총액 = mean(매출총액)) %>%
  group_by(대분류) %>% mutate(num = row_number()) %>%
  ggplot(aes(x = num, y = 매출총액, group = 대분류, col = 대분류))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 대분류별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))

######## 2. 대분류별 매출 총액
text <- dataset %>% group_by(대분류) %>% summarise(매출총액 = mean(매출총액))
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 대분류, y =매출총액, fill = 대분류))+
  geom_boxplot()+theme_bw()+  ggtitle("대분류별 매출비율")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),axis.text.x = element_text(vjust = 0.6,size = 13, face
  geom_text(data = text, mapping = aes(x = 대분류,y = 매출총액+0.3, label = round(매출총액,2)),size = 8)
```

```
######## 3.년&분기 별 대분류에 따른 매출비율 비교
data <- dataset %>% mutate(년분기 = paste0(년도,"_",분기)) %>% filter(년분기 !="2020_3") %>% select(-년분기)
text = data.frame(text = c("서비스업","외식업","소매업"), x = c(15.8, 17.9, 19.1), y = c(0.34, 0.7, 0.34))
ggplot()+
  geom_density(data = data, mapping = aes(x = 매출총액, fill = 대분류),alpha = 0.4)+
  geom_text(data=text, mapping = aes(x = x, y = y, label = text, color = text),size = 7, fontface = "bold")+theme_classic()
```

Aa 이름	생성일	태그
2-1. 년분기별 대분류에 따른 매출총액	@2021년 1월 24일 오후 10:57	
2-2. 대분류별 매출총액 박스플랏	@2021년 1월 24일 오후 11:35	
2-3. 대분류별 매출총액 밀도	@2021년 1월 24일 오후 10:57	

입력변수 : 중분류

```
######## 1. 년&분기 별 중분류에 따른 매출비율 비교
vars <- dataset %>% distinct(년분기) %>% arrange(년분기)
text <- dataset %>% filter(년분기 =="2020_2") %>%
  group_by(중분류) %>% summarise(매출총액 = mean(매출총액))
dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기, 중분류) %>%
  summarise(매출총액 = mean(매출총액)) %>%
  group_by(중분류) %>% mutate(num = row_number()) %>%
  ggplot(aes(x = num, y = 매출총액, group = 중분류, col = 중분류))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 중분류별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text, aes(x = rep(22.5,20),y = 매출총액, label = 중분류), fontface = "bold", size = 5)

######## 2. 중분류별 매출 총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 중분류, y =매출총액, fill = 중분류))+
  geom_boxplot()+theme_bw()+  ggtitle("중분류별 매출비율")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
        axis.text.x = element_text(vjust = 0.6,angle = 20))

######## 3. 중분류에 따른 매출비율 비교
dataset %>% mutate(년분기 = paste0(년도,"_",분기)) %>% filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 매출총액,y = 중분류, fill = 중분류))+
  geom_density_ridges(alpha = 0.4)+theme_classic()+ggtitle("중분류별 매출비율")+
  theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
6-1. 년분기별 중분류에 따른 매출총액	@2021년 1월 24일 오후 10:59	
6-2. 중분류별 매출총액	@2021년 1월 24일 오후 10:59	
6-3. 중분류별 매출총액	@2021년 1월 24일 오후 10:59	

입력변수 : 소분류

```
######## 1. 년&분기 별 소분류에 따른 매출비율 비교
vars <- dataset %>% distinct(년분기) %>% arrange(년분기)
text <- dataset %>% filter(년분기 =="2020_2" & 대분류 == "서비스업") %>%
  group_by(소분류) %>% summarise(매출총액 = mean(매출총액))
dataset %>%
  filter(년분기 !="2020_3" & 대분류 == "서비스업") %>%
  group_by(년분기, 소분류) %>%
  summarise(매출총액 = mean(매출총액)) %>%
  group_by(소분류) %>% mutate(num = row_number()) %>%
  ggplot(aes(x = num, y = 매출총액, group = 소분류, col = 소분류))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 소분류별 매출총액")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text, aes(x = rep(22.5,nrow(text)),y = 매출총액, label = 소분류), fontface = "bold", size = 5)+
  scale_y_continuous(breaks = seq(10,30,0.5), limits = c(14.75,20.25))
```

```
######## 2. 소분류별 매출 총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 소분류, y =매출총액, fill = 소분류))+
  geom_boxplot()+theme_bw()+  ggtitle("소분류별 매출총액")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
        axis.text.x = element_text(vjust = 0.6,angle = 60))

######## 3. 소분류에 따른 매출비율 비교
dataset %>%
  filter(년분기 !="2020_3" & 대분류 == "서비스업") %>%
  ggplot(aes(x = 매출총액,y = 소분류, fill = 소분류))+
  geom_density_ridges(alpha = 0.4)+theme_classic()+ggtitle("소분류별 매출총액")+
  theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
15-1. 년분기 및 소분류에 따른 매출총액(서비스업)	@2021년 1월 24일 오후 11:01	
15-1. 년분기 및 소분류에 따른 매출총액(소매업)	@2021년 1월 24일 오후 11:01	
15-1. 년분기 및 소분류에 따른 매출총액(외식업)	@2021년 1월 24일 오후 11:01	
15-2. 소분류에 따른 매출총액	@2021년 1월 24일 오후 11:01	
15-3. 소분류에 따른 매출총액(서비스업)	@2021년 1월 24일 오후 11:01	
15-3. 소분류에 따른 매출총액(소매업)	@2021년 1월 24일 오후 11:01	
15-3. 소분류에 따른 매출총액(외식업)	@2021년 1월 24일 오후 11:01	

입력변수 : 생존율 1년차 / 3년차 / 5년차

```
dataset %>%
  ggplot(aes(x = 생존율_1년차, y = 매출총액, col = 매출총액))+geom_point()+
  xlab("생존율(%)")+ggtitle("1년차 신생기업 생존율에 따른 매출")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

dataset %>%
  ggplot(aes(x = 생존율_3년차, y = 매출총액, col = 매출총액))+geom_point()+
  xlab("생존율(%)")+ggtitle("3년차 신생기업 생존율에 따른 매출")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

dataset %>%
  ggplot(aes(x = 생존율_5년차, y = 매출총액, col = 매출총액))+geom_point()+
  xlab("생존율(%)")+ggtitle("5년차 신생기업 생존율에 따른 매출")+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))
```

Aa 이름	생성일	태그
5-1. 1년차 생존율에 따른 매출총액 산점도	@2021년 1월 24일 오후 11:03	
5-1. 3년차 생존율에 따른 매출총액 산점도	@2021년 1월 24일 오후 11:03	
5-1. 5년차 생존율에 따른 매출총액 산점도	@2021년 1월 24일 오후 11:03	

입력변수 : 성별

```
######## 1. 년&분기별 성별에 따른 매출비율 비교
dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기) %>%
  summarise(매출비율_남성 = mean(매출비율_남성),
                    매출비율_여성 = mean(매출비율_여성)) %>%
  gather(key = 성별,value = 매출비율, 매출비율_남성:매출비율_여성) %>%
  group_by(성별) %>%
  mutate(num = row_number(),성별 = factor(성별)) %>%
  ggplot(aes(x = num, y =매출비율, group = 성별, col = 성별))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 성별별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                  axis.text.x = element_text(angle = 50, vjust = 0.6))
```

```
######## 2. 성별에 따른 매출 비율
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 성별,value = 비율, 매출비율_남성:매출비율_여성) %>%
  ggplot(aes(x = 성별, y = 비율, fill = 성별))+geom_boxplot()+
  ggtitle("성별에 따른 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))

######## 3. 성별에 따른 매출 비율
colnames(dataset)
dataset %>%
  filter(년분기 !="2020_3") %>%
  gather(key = 성별,value = 매출, 매출비율_남성:매출비율_여성) %>%
  mutate(비율 = 매출/매출총액) %>% select(성별,비율) %>%
  ggplot(aes(x = 비율, y = 성별, fill = 성별))+geom_density_ridges(alpha = 0.4)+
  theme_classic()+ggtitle("성별에 따른 매출비율")+theme(plot.title = element_text(face = "bold", size = 15,hjust = 0.5))
```

Aa 이름	생성일	태그
7-1. 년분기별 성별에 따른 매출비율	@2021년 1월 24일 오후 11:04	
7-2. 성별에 따른 매출비율	@2021년 1월 24일 오후 11:04	
7-3. 성별에 따른 매출비율	@2021년 1월 24일 오후 11:04	
7-4. 성별 산점도	@2021년 1월 24일 오후 11:23	

입력변수 : 매출건수

```
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 총매출건수, y = 매출총액, col = 매출총액))+geom_point()
```

Aa 이름	생성일	태그
11. 매출건수별 매출총액 산점도	@2021년 1월 24일 오후 11:05	

입력변수 : 년도

```
########  1. 년도에 따른 매출비율 비교
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 년도, y = 매출총액,fill = 년도))+geom_boxplot()

######## 2 년도에 따른 매출
dataset %>% filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 매출총액, y = 년도,fill = 년도))+geom_density_ridges(alpha = 0.4)
```

Aa 이름	생성일	태그
8-1. 년도에 따른 매출비율	@2021년 1월 24일 오후 11:24	
8-2. 년도에 따른 매출비율	@2021년 1월 24일 오후 11:24	

입력변수 : 분기

```
#######  1. 분기에 따른 매출비율 비교
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 분기, y = 매출총액,fill = 분기))+geom_boxplot()

######## 2 분기에 따른 매출
dataset %>% filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 매출총액,y = 분기,fill = 분기))+geom_density_ridges(alpha = 0.4)
```

Aa 이름	생성일	태그
9-1. 분기에 따른 매출비율	@2021년 1월 24일 오후 11:26	
9-2. 분기에 따른 매출비율	@2021년 1월 24일 오후 11:26	

입력변수 : 행정구역

```
######## 1 행정구역 및 년분기별 매출총액
text <- dataset %>% filter(년분기 !="2020_3") %>%
  group_by(년분기, 행정구역) %>% summarise(매출총액 = mean(매출총액)) %>% filter(년분기=="2020_2")
dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기, 행정구역) %>%
  summarise(매출총액 = mean(매출총액)) %>%
  group_by(행정구역) %>% mutate(num = row_number()) %>%
  ggplot(aes(x = num, y = 매출총액, group = 행정구역, col = 행정구역))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 행정구역별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                    axis.text.x = element_text(angle = 50, vjust = 0.6))+
  geom_text(data = text, aes(x = rep(22.5,nrow(text)),y = 매출총액, label = 행정구역), fontface = "bold", size = 5)

######## 2행정구역별 매출총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 매출총액, y = 행정구역, fill = 행정구역))+geom_density_ridges(alpha = 0.4)
```

Aa 이름	생성일	태그
10-1.행정구역 및 년분기별 매출총액 추이	@2021년 1월 24일 오후 11:28	
10-2.행정구역별 매출총액	@2021년 1월 24일 오후 11:28	

입력변수 : 유동인구수

```
######## 1 년분기 및 총유동인구별 매출총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  group_by(년분기) %>%
  summarise(매출총액 = mean(매출총액), 총유동인구 = mean(총유동인구)) %>%
  mutate(num = row_number()) %>%
  gather(구분,값,매출총액:총유동인구) %>%
  ggplot(aes(x = num, y = 값, group = 구분, color = 구분))+geom_line(lwd = 2)+
  scale_x_continuous(breaks = seq(1,23), labels = unique(dataset$년분기))+theme_classic()+
  ggtitle("년분기별 총유동인구수에 따른 매출총액")+theme(plot.title = element_text(hjust = 0.5, size = 15, face = "bold"))

######## 2 총유동인구별 매출총액 산점도
dataset %>%
  filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 총유동인구, y = 매출총액, col = 매출총액))+geom_point()
```

Aa 이름	생성일	태그
13-1. 년분기별 총유동인구에 따른 매출총액	@2021년 1월 24일 오후 11:29	
13-2. 총유동인구_산점도	@2021년 1월 24일 오후 11:29	

입력변수 : 집객시설수

```
filter(년분기 !="2020_3") %>%
  ggplot(aes(x = 집객시설수, y = 매출총액, col = 매출총액))+geom_point()
```

Aa 이름	생성일	태그

Aa 이름	생성일	태그
14. 집객시설수와 매출총액 산점도	@2021년 1월 24일 오후 11:29	

입력변수 : 소득분위

```
######## 1 년분기 및 소득분위별 매출총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  mutate(소득분위 = factor(소득분위, levels = c("9","8","7","6","5","4","3","2","1"))) %>%
  group_by(년분기, 소득분위) %>%
  summarise(매출총액 = mean(매출총액)) %>%
  group_by(소득분위) %>% mutate(num = row_number()) %>%
  ggplot(aes(x = num, y = 매출총액, group = 소득분위, col = 소득분위))+geom_line(size = 1.2)+
  scale_x_continuous(breaks = seq(1:23), labels = vars[,])+theme_bw()+
  ggtitle("년분기 및 소득분위별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
                                    axis.text.x = element_text(angle = 50, vjust = 0.6))
######## 2 년분기 및 소득분위별 매출총액
dataset %>%
  filter(년분기 !="2020_3") %>%
  mutate(소득분위 = factor(소득분위, levels = c("9","8","7","6","5","4","3","2","1"))) %>%
  ggplot(aes(x = 소득분위, y = log(매출총액), group = 소득분위, fill = 소득분위))+geom_boxplot()+theme_bw()+
  ggtitle("년분기 및 소득분위별 매출비율")+theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 15))
```

Aa 이름	생성일	태그
12-1. 소득분위	@2021년 1월 24일 오후 11:37	
12-2. 소득분위	@2021년 1월 24일 오후 11:37	

통계적 분석 - 연속형

```
#데이터 변경, 범주형 및 연속형
vars <- c("년도","분기","행정구역","대분류","중분류","소분류","년분기","소득분위")
vars <- which(colnames(dataset) %in% vars)
dataset[,vars] <- map_df(.x = dataset[,vars], .f = as.factor)
dataset[,-vars] <- map_df(.x = dataset[,-vars], .f = as.numeric)

#trainset & testset 나누기
trainset <- dataset %>% filter(년분기 !='2020_3')
testset <- dataset %>% filter(년분기 =='2020_3')

#연속형 컬럼을 나누어 상관관계플랏 생성
vars_numeric <- colnames(trainset)[-vars]
num <- length(vars_numeric)
set1 <- c(vars_numeric[c(1:(num/2))],"매출총액")
set2 <- vars_numeric[(num/2+1):num]
corrplot(cor(trainset[,set1],use = "na.or.complete"),method = "number")
corrplot(cor(trainset[,set2],use = "na.or.complete"),method = "number")

# 시간대 06-14 / 17-24 컬럼을 생성하여 상관관계 재 분석
#시간대별 매출비율 피쳐 생성
dataset <- smallbz_total_1501_2009 %>%
  group_by(년도,분기,행정구역,대분류,중분류) %>%
  summarise(매출비율_월 = mean(월요일_매출_금액/매출총액),
                  매출비율_화 = mean(화요일_매출_금액/매출총액),
                  매출비율_수 = mean(수요일_매출_금액/매출총액),
                  매출비율_목 = mean(목요일_매출_금액/매출총액),
                  매출비율_금 = mean(금요일_매출_금액/매출총액),
                  매출비율_토 = mean(토요일_매출_금액/매출총액),
                  매출비율_일 = mean(일요일_매출_금액/매출총액),
                  매출비율_10대 = mean(연령대_10_매출_금액/매출총액),
                  매출비율_20대 = mean(연령대_20_매출_금액/매출총액),
                  매출비율_30대 = mean(연령대_30_매출_금액/매출총액),
                  매출비율_40대 = mean(연령대_40_매출_금액/매출총액),
                  매출비율_50대 = mean(연령대_50_매출_금액/매출총액),
                  매출비율_60대이상 = mean(연령대_60_이상_매출_금액/매출총액),
                  매출비율_0614 = mean((시간대_06.11_매출_금액+시간대_11.14_매출_금액)/매출총액),
                  매출비율_1724 = mean((시간대_17.21_매출_금액+시간대_21.24_매출_금액)/매출총액),
                  매출비율_남성 = mean(남성_매출_금액/매출총액)*(매출비율_10대+매출비율_20대),
                  매출비율_여성 = mean(여성_매출_금액/매출총액)*(매출비율_40대),
                  생존율_1년차 = mean(생존률_1년차),
                  생존율_3년차 = mean(생존률_3년차),
```

```
                  생존율_5년차 = mean(생존률_5년차),
                  소득분위 = as.factor(round(mean(as.numeric(소득분위)),digits = 0)),
                  집객시설수 = mean(집객시설_수),
                  총유동인구= mean(log(총_유동인구수)),
                  총매출건수 = mean(log(총매출건수)),
                  매출총액 = mean(log(매출총액/점포수))) %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  as.data.frame()

#ln() 후 -inf를 0으로 변경
dataset[,"총유동인구"] <- ifelse(is.infinite(dataset[,"총유동인구"])==T, 0, dataset[,"총유동인구"])

#데이터 변경, 범주형 및 연속형
vars <- c("년도","분기","행정구역","대분류","중분류","소분류","년분기","소득분위")
vars <- which(colnames(dataset) %in% vars)
dataset[,vars] <- map_df(.x = dataset[,vars], .f = as.factor)
dataset[,-vars] <- map_df(.x = dataset[,-vars], .f = as.numeric)

#trainset & testset 나누기
trainset <- dataset %>% filter(년분기 !='2020_3')
testset <- dataset %>% filter(년분기 =='2020_3')

corrplot(cor(trainset[,-vars],use = "na.or.complete"),method = "number")
```

Aa 이름	생성일	태그
16. 상관분석_1	@2021년 1월 24일 오후 11:45	
16. 상관분석_2	@2021년 1월 24일 오후 11:45	
16. 상관분석_3_시간대 합계 피쳐 생성	@2021년 1월 25일 오전 12:07	

상관계수가 0.25인 변수 선택 ⇒ 시간대_0614, 시간대_1724, 총매출건수

통계적 분석 - 범주형

입력변수 : 년도, 분기, 행정구역, 대분류, 중분류, 소분류, 년분기, 소득분위

```
#연속형인 목표형 데이터와 명목형인 입력변수 t검정 및 아노바 검정
#년도(2015~2020 => anova)
tapply(X = trainset$매출총액,INDEX = trainset$년도, FUN = shapiro.test)
table(trainset$년도)
#n>30 으로 정규성이라 가정하고 진행
bartlett.test(formula = 매출총액~년도, data = trainset)
#p-value 0.05미만으로 귀무가설을 채택하지 못하므로 이분산
oneway.test(formula = 매출총액~년도, data = trainset, var.equal=F)
duncan.test(y = aov(formula = 매출총액~년도, data = trainset),
            alpha = 0.05,
            trt = "년도",
            group=T,
            console = T)
#p-avlue 0.05이하로 년도에 따른 매출총액 값의 평균 중 적어도 하나이상은 다름
#3개 그룹으로 나뉨


#분기(1,2,3,4 => anova)
library(nortest)
by(data = trainset$매출총액, INDICES = trainset$분기, FUN = ad.test)
table(trainset$분기)
#n>30으로 정규성이라 가정하고 진행
bartlett.test(formula = 매출총액~분기, data = trainset)
tapply(X = trainset$매출총액,INDEX = trainset$분기,FUN = mean)
#등분산
summary(aov(formula = 매출총액~분기, data = trainset))
#p-value 0.05 이하로 분기 그룹에 따른 매출총액의 차이가 있음
#사후검정
library(agricolae)
duncan.test(y = aov(formula = 매출총액~분기, data = trainset),
            trt = "분기",
            alpha = 0.05,
            group = T,
            console = T)
#1분기 /2,3분기 / 4분기 그룹으로 나뉨
#모델의 입력변수로 추가하여 있을 경우와 없을 경우 비교


#행정구역(25개 => anova)
by(data = trainset$매출총액, INDICES = trainset$행정구역, FUN = shapiro.test)
table(trainset$행정구역)
#종로구, 송파구, 노원구, 구로구는 shapiro.test를 통과하지 못했으나
```

```r
#n>30 초과로 정규분포라 가정
bartlett.test(formula = 매출총액~행정구역, data = trainset)
#p-value 0.05 이하로 귀무 채택, 즉 이분산
oneway.test(formula = 매출총액~행정구역,data = trainset, var.equal = F)
#적어도 한 그룹 이상의 매출총액  평균이 다름 => 사후검정 진행
#이분산의 사후검정
library(agricolae)
duncan.test(y = aov(formula = 매출총액~행정구역, data = trainset),
            trt = "행정구역",
            alpha = 0.05,
            group = F,
            console = T)
#행정구에 따라 평균이 같거나 다른 경우가 있음
#입력변수 적용 전후 모델 성능 비교해보기


#대분류
by(data = trainset$매출총액,INDICES = trainset$대분류, FUN = shapiro.test)
table(trainset$대분류)
#정규분포 검정을 만족하지 못하지만, n>30이므로 정규성 가정
bartlett.test(formula = 매출총액~대분류, data = trainset)
#p-value 0.05 이하로 이분산
oneway.test(formula = 매출총액~대분류, data = trainset, var.equal = F)
#p-value 0.05 미만으로 대분류 그룹간 평균 차이 있음
#입력변수로 채택
duncan.test(y = aov(formula = 매출총액~대분류, data = trainset),
            trt = "대분류",
            console = T,
            group = T,
            alpha = 0.05)
#사후 검정을 통해 업태별로 다른 그룹으로 판별됨


#중분류
tapply(X = trainset$매출총액, INDEX = trainset$중분류, FUN = shapiro.test)
table(trainset$중분류)
#20개 주운류 중주점업,오락관련서비스,스포츠,숙박,부동산,기타생활용품,교육을 제외하고 정규성 없음
#단, n>30으로 정규성 가정하고 진행
bartlett.test(formula = 매출총액~중분류, data = trainset)
#p-value 0.05 이하로 이분산
oneway.test(formula = 매출총액~중분류, data = trainset,var.equal = F)
#중분류별 매출총액의 평균은 차이가 있다는, 귀무가설 채택
duncan.test(y = aov(formula = 매출총액~중분류, data = trainset),
            trt = "중분류",
            alpha = 0.05,
            group = T,
            console = T)
#사후검정 시 중분류 20개가 16개 그룹으로 나뉨


#소득분위
tapply(X = trainset$매출총액, INDEX = trainset$소득분위, FUN = shapiro.test)
table(trainset$소득분위)
#20개 주운류 중주점업,오락관련서비스,스포츠,숙박,부동산,기타생활용품,교육을 제외하고 정규성 없음
#단, n>30으로 정규성 가정하고 진행
bartlett.test(formula = 매출총액~소득분위, data = trainset)
#p-value 0.05 이하로 이분산
oneway.test(formula = 매출총액~소득분위, data = trainset,var.equal = F)
#중분류별 매출총액의 평균은 차이가 있다는, 귀무가설 채택
duncan.test(y = aov(formula = 매출총액~소득분위, data = trainset),
            trt = "소득분위",
            alpha = 0.05,
            group = T,
            console = T)
#소득분위가 현재 3~7로 총 5개 그룹이 있고, 사후검정 시 3그룹으로 나뉨


#년분기
tapply(X = trainset$매출총액, INDEX = trainset$년분기, FUN = shapiro.test)
table(trainset[trainset$년분기!="2020_3",]$년분기)
#정규성이 없지만 N>30 이므로 정규성 가성
bartlett.test(formula = 매출총액~년분기, data = trainset)
#p-value 0.05미만으로 이분산
oneway.test(formula = 매출총액~년분기, data = trainset, var.equal = F)
#p-value 0.05 미만이지만으로 적어도 하나 다름 => 사후검정
duncan.test(y = aov(formula = 매출총액~년분기, data = trainset),
            trt = "년분기",
            console = T)
#22개 분기가 총 5그룹으로 나뉨

library(nortest)
data <- trainset %>% gather(성별,매출비율,매출비율_남성,매출비율_여성) %>% select(성별,매출비율)
tapply(X = data$매출비율,INDEX = data$성별, FUN = ad.test)
var.test(매출비율~성별, data = data)
t.test(매출비율~성별, data = data, var.equal=F)
by()
#n>30 이상이므로 정규성 가정
```

```
#var.test => 이분산
#t.test => 대립 채택 못함, 성별에 따른 차이 없음
```

상기와 같이 EDA, 상관분석, 아노바 검정 결과 다음과 같이 입력변수를 채택함

연속형 : 총매출건수, 매출비율_0614, 매출비율_1724

범주형 : 년도, 분기, 행정구역, 대분류, 중분류, 소분류, 년분기, 소득분위

```
최종 데이터 셋
dataset <- smallbz_total_1501_2009 %>%
  group_by(년도,분기,행정구역,대분류,중분류) %>%
  summarise(매출비율_0614 = mean((시간대_06.11_매출_금액+시간대_11.14_매출_금액)/매출총액),
            매출비율_1724 = mean((시간대_17.21_매출_금액+시간대_21.24_매출_금액)/매출총액),
            소득분위 = as.factor(round(mean(as.numeric(소득분위)),digits = 0)),
            총매출건수 = mean(log(총매출건수)),
            매출총액 = mean(log(매출총액/점포수))) %>%
  mutate(년분기 = paste0(년도,"_",분기)) %>%
  as.data.frame()

#데이터 변경, 범주형 및 연속형
vars <- c("년도","분기","행정구역","대분류","중분류","소분류","년분기","소득분위")
vars <- which(colnames(dataset) %in% vars)
dataset[,vars] <- map_df(.x = dataset[,vars], .f = as.factor)
dataset[,-vars] <- map_df(.x = dataset[,-vars], .f = as.numeric)

#trainset & testset 나누기
trainset <- dataset %>% filter(년분기 !='2020_3')
testset <- dataset %>% filter(년분기 =='2020_3')
```