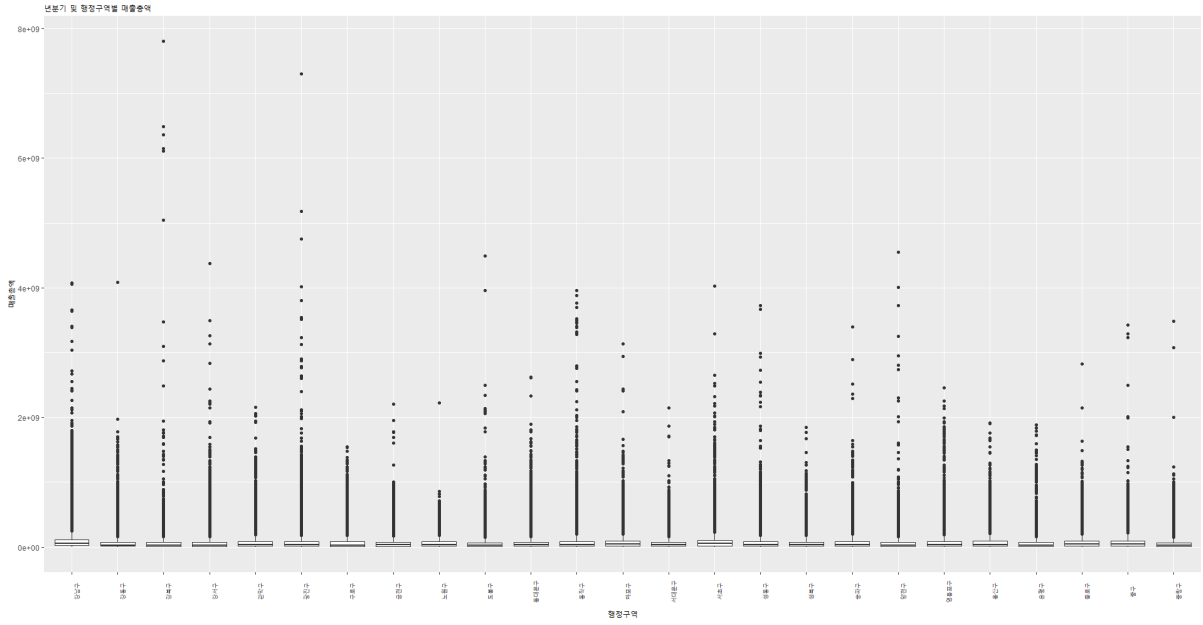
eda

- 년분기 및 행정구역에 따른 점포별 매출총액 박스 플랏

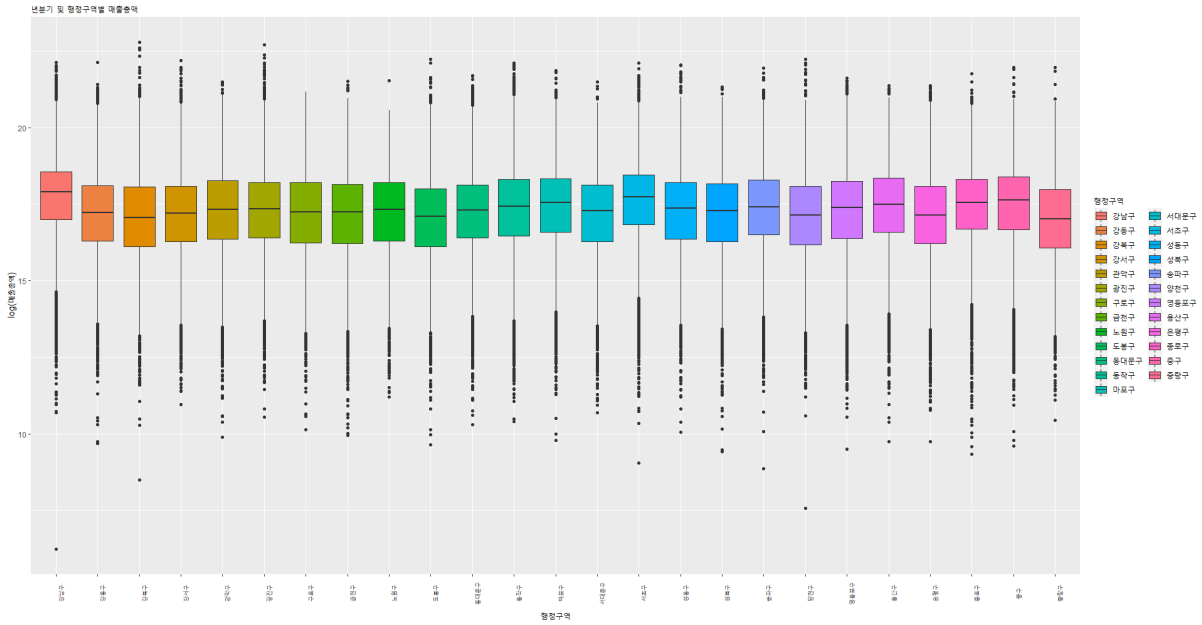
```
smallbz_total_1501_2009 %>%
  mutate(년분기 = as.factor(paste0(년도,"_",분기))) %>%
  filter(년분기 != "2020_3" & 행정동명 !="가락1동") %>%
  mutate(매출총액 = 매출총액/점포수) %>%
  ggplot(aes(x = 행정구역, y = 매출총액, fill = 행정구역))+
  geom_boxplot(position = 'dodge')+
  ggtitle("년분기 및 행정구역별 매출총액")+theme(axis.text.x = element_text(angle = 90))
```



데이터 스케일 및 정규성을 확보하기 위해 매출 및 유동인구수에 자연로그를 취한 값으로 진행

- 년분기 및 행정구역에 따른 점포별 log(매출총액) 박스 플랏

```
smallbz_total_1501_2009 %>%
  mutate(년분기 = as.factor(paste0(년도,"_",분기))) %>%
  filter(년분기 != "2020_3" & 행정동명 !="가락1동") %>%
  mutate(매출총액 = 매출총액/점포수) %>%
  ggplot(aes(x = 행정구역, y = log(매출총액),fill = 행정구역))+
  geom_boxplot(position = 'dodge')+
  ggtitle("년분기 및 행정구역별 매출총액")+theme(axis.text.x = element_text(angle = 90))
```



박스플랏 결과, 아웃라이어가 다수 보이지만, 해당 값은 업태별 매출액 차이므로 평균값으로 대체하여 사용

- 데이터셋 정리

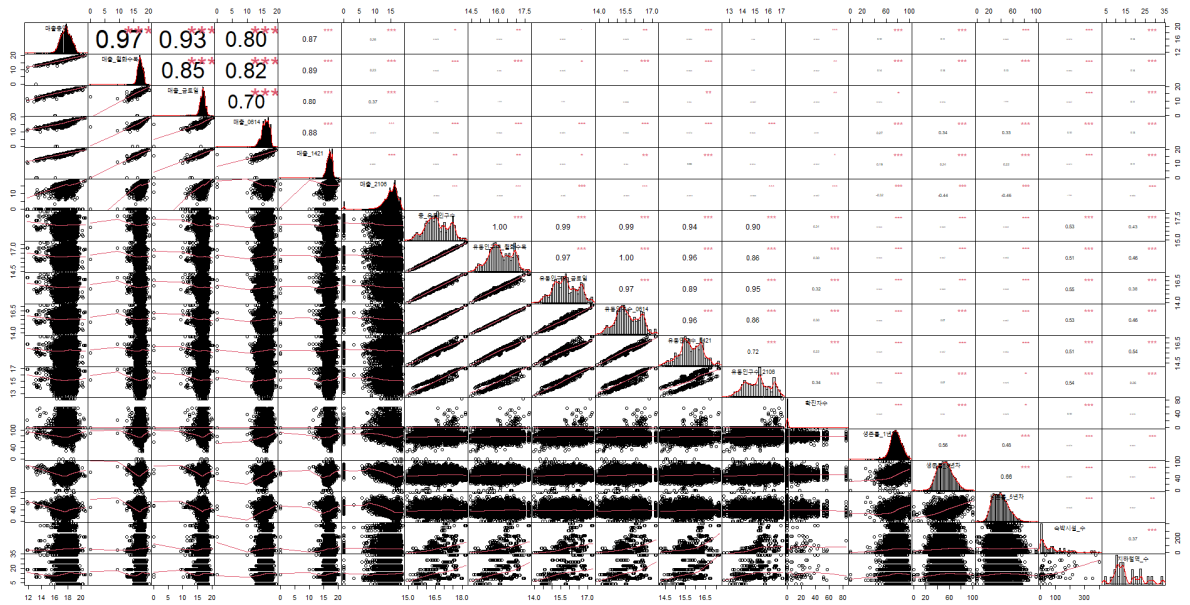
```
#모든 값을 각 중분류의 평균값으로 적용하고,
#스케일 조정 및 정규분포화를 위해 매출 및 유동인구 데이터는 ln 적용
dataset <- smallbz_total_1501_2009 %>%
  group_by(년도,분기,행정구역,대분류,중분류) %>%
  summarise(매출총액 = log(mean(매출총액)/mean(점포수)),
                매출_월화수목 = log(mean(매출_월화수목)/mean(점포수)),
                매출_금토일 = log(mean(매출_금토일)/mean(점포수)),
                매출_0614 = log(mean(매출_0614)/mean(점포수)),
                매출_1421 = log(mean(매출_1421)/mean(점포수)),
                매출_2106 = log(mean(매출_2106)/mean(점포수)),
                총_유동인구수 = log(mean(총_유동인구수)),
                유동인구수_월화수목 = log(mean(유동인구수_월화수목)),
                유동인구수_금토일 = log(mean(유동인구수_금토일)),
                유동인구수_0614 = log(mean(유동인구수_0614)),
                유동인구수_1421 = log(mean(유동인구수_1421)),
                유동인구수_2106 = log(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)) %>%
  as.data.frame()

#일부 컬럼에서 값이 ln(0)로 -inf인 인덱스는 0 입력
```

```
vars <- 6:17
dataset[,vars] <- lapply(X = dataset[,vars],FUN = function(x){ifelse(x < 0,0,x)})
#train # test set 나누기
dataset <- dataset %>% mutate(년분기 = as.factor(paste0(년도,"_",분기)))
trainset <- dataset %>% filter(년분기 !='2020_3')
testset <- dataset %>% filter(년분기 =='2020_3')
```

- 연속형 데이터 상관분석

```
chart.Correlation(R = trainset[,6:23])
```



목표변수인 매출총액과 매출_월화수목 , 매출_금토일 , 매출_0614 , 매출_1421 데이터는 상관관계가 있으므로 입력변수로 채택

- 연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정
- 목표변수 : 매출총액 / 입력변수 : 년도

```
table(trainset$년도)
bartlett.test(formula = 매출총액~년도, data = trainset)
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)
```

1. 정규성 검정 : 관측치 개수가 년도별 900개 이상이므로 중심극한정리에 의하여 정규성이라 가정
2. 분산 검정 : p-value 0.05미만이므로 귀무가설 채택 ⇒ 이분산
3. 아노바 검정 : p-value 0.05 미만이며, 사후검정 결과,
   2015,2016,2017 / 2018 / 2019,2020 세그룹으로 확인됨

⇒ 후보 입력변수로 채택하고 입력변수 유무일 때 결과 비교

- 연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정
- 목표변수 : 매출총액 / 입력변수 : 년도

```
table(trainset$분기)
bartlett.test(formula = 매출총액~분기, data = trainset)
oneway.test(formula = 매출총액~분기, data = trainset, var.equal=T)
duncan.test(y = aov(formula = 매출총액~분기, data = trainset),
            alpha = 0.05,
            trt = "년도",
            group=T,
            console = T)
```

1. 정규성 검정 : 관측치 개수가 최소 2350개 이상이므로 중심극한 정리에 의해 정규성 가정
2. 분산 검정 : p-0.05 초과로 귀무가설 채택 불가 ⇒ 등분산
3. 이노바 검정 : p-value 0.05미만이며, 사후검정 결과,
   1분기 : a, 2분기 : ab, 3분기 : b, 4분기 : c

⇒ 후보 입력변수로 채택하고 입력변수 유무일 때 결과 비교

연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정

- 목표변수 : 매출총액 / 입력변수 : 행정구역

```
table(trainset$행정구역)
bartlett.test(formula = 매출총액~행정구역, data = trainset)
oneway.test(formula = 매출총액~행정구역, data = trainset, var.equal=F)
duncan.test(y = aov(formula = 매출총액~행정구역, data = trainset),
            alpha = 0.05,
            trt = "년도",
```

```
            group=T,
            console = T)
```

1. 정규성 검정 : 관측치 개수가 최소 400개 이상이므로 중심극한 정리에 의해 정규성 가정
2. 분산 검정 : p-0.05 미달로 귀무가설 채택 불가 ⇒ 이분산
3. 이노바 검정 : p-value 0.05미만이며, 사후검정 결과, 일부 행정구역 별로 차이가 있다고 판단됨

⇒ 후보 입력변수로 채택하고 입력변수 유무일 때 결과 비교

연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정

- 목표변수 : 매출총액 / 입력변수 : 대분류

```
table(trainset$행정구역)
bartlett.test(formula = 매출총액~대분류, data = trainset)
oneway.test(formula = 매출총액~대분류, data = trainset, var.equal=F)
duncan.test(y = aov(formula = 매출총액~대분류, data = trainset),
            alpha = 0.05,
            trt = "년도",
            group=T,
            console = T)
```

1. 정규성 검정 : 관측치 개수가 최소 2200개 이상이므로 중심극한 정리에 의해 정규성 가정
2. 분산 검정 : p-0.05 미달로 귀무가설 채택 불가 ⇒ 이분산
3. 이노바 검정 : p-value 0.05미만이며, 사후검정 결과, 대분류 각각이 a,b,c로 나뉨

⇒입력변수로 채택

연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정

- 목표변수 : 매출총액 / 입력변수 : 중분류

```
table(trainset$행정구역)
bartlett.test(formula = 매출총액~중분류, data = trainset)
oneway.test(formula = 매출총액~중분류, data = trainset, var.equal=F)
duncan.test(y = aov(formula = 매출총액~중분류, data = trainset),
            alpha = 0.05,
            trt = "년도",
```

```
            group=T,
            console = T)
```

1. 정규성 검정 : 관측치 개수가 최소 2200개 이상이므로 중심극한 정리에 의해 정규성 가정
2. 분산 검정 : p-0.05 미달로 귀무가설 채택 불가 ⇒ 이분산
3. 이노바 검정 : p-value 0.05미만이며, 사후검정 결과, 중분류가 다양한 그룹으로 나뉨

⇒ 후보 입력변수로 채택

연속형인 목표형 데이터와 명목형인 입력변수 아노바 검정

- 목표변수 : 매출총액 / 입력변수 : 년분기

```
table(trainset$행정구역)
bartlett.test(formula = 매출총액~년분기, data = trainset)
oneway.test(formula = 매출총액~년분기, data = trainset, var.equal=F)
duncan.test(y = aov(formula = 매출총액~년분기, data = trainset),
            alpha = 0.05,
            trt = "년도",
            group=T,
            console = T)
```

1. 정규성 검정 : 관측치 개수가 최소 450개 이상이므로 중심극한 정리에 의해 정규성 가정
2. 분산 검정 : p-0.05 미달로 귀무가설 채택 불가 ⇒ 이분산
3. 이노바 검정 : p-value 0.05미만이며, 사후검정 결과, a~f그룹으로 나뉨

⇒ 후보 입력변수로 채택