대학상권 매출액에 영향을 주는 요인 분석 및 2020년도 3분기 매출액 예측

킵고잉
정두리 이정훈

목차

1. 주제선정

분석목표

주제선정 배경

상권지역 선정

대학상권 기준

대학상권 매출액에 영향을 주는 요인을 분석하여 2020년도 3분기 매출액 예측

1. 주제선정

분석목표

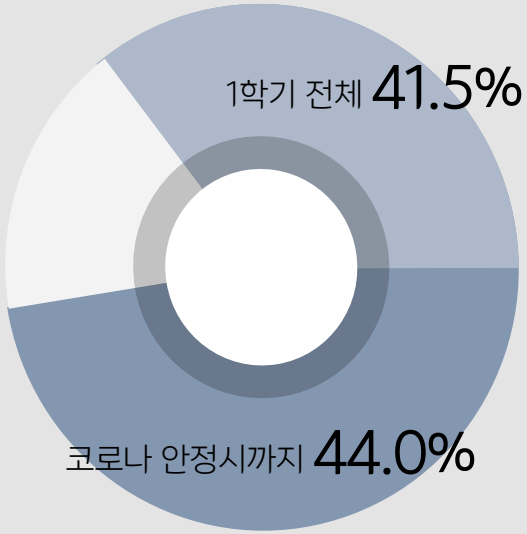
주제선정 배경

상권지역 선정

대학상권 기준

“사회적 거리두기”로 인한 대면수업 중단

온라인 수업 진행 대학 비율 (전국 193개 4년제 대학 중)



<한국사립대학총장협의회의, 5/15기준>

코로나19로 외국인 유학생 6년 만에 감소

<2020 교육기본통계,교육부>

올해 초부터 시작된 코로나19 여파에 국내 대학에 재학 중인 외국인 유학생도 6년만에 감소했다. 2015년 9만1332명이었던 국내 외국인 유학생 수는 2016년 10만 명을 넘어선 뒤 △2017년 12만3900명 △2018년 14만2200명 △2019년 16만200명 등으로 매년 증가세를 보였지만 올해는 15만3695명으로 6년 만에 감소했다.

- > 대학상권 분석을 통해 매출액을 예측하고 상권을 활성화 시킬 수 있는 방안을 찾고자 했음

1. 주제선정

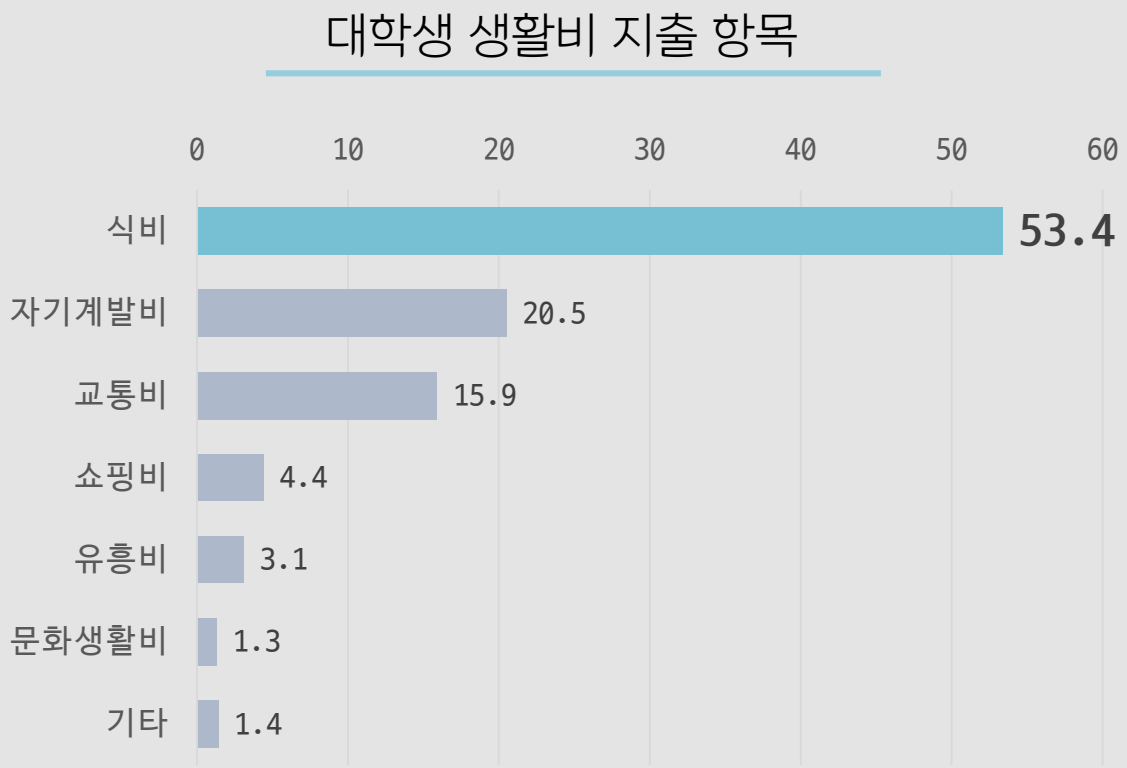
분석목표

주제선정 배경

상권지역 선정

대학상권 기준

대학생 생활비 지출항목 1위는 식비



'대학생 월평균 생활비 설문조사'
조사기관 : 잡코리아X알바몬 통계센터, 2020

1. 주제선정

분석목표

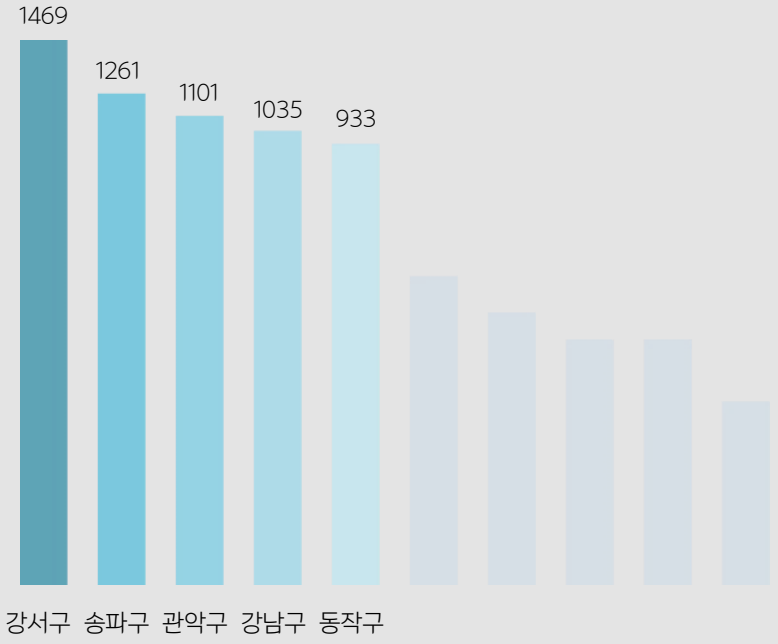
주제선정 배경

상권지역 선정

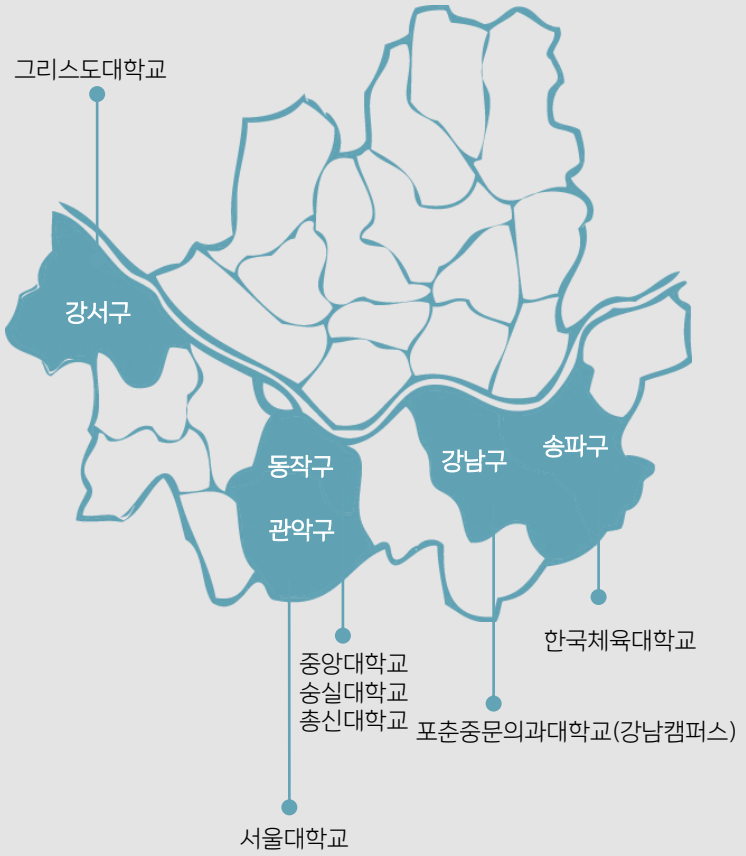
대학상권 기준

서울시 코로나 누적 확진자 상위 5개 자치구

서울시에 있는 25개의 자치구 중 코로나 누적 확진자 상위 5개 자치구를 기준으로 폭넓은 대학상권을 EDA해보기 위해 동작구를 선정



21.2.1기준
(서울특별시_I_SEOUL_U)



1. 주제선정

분석목표

주제선정 배경

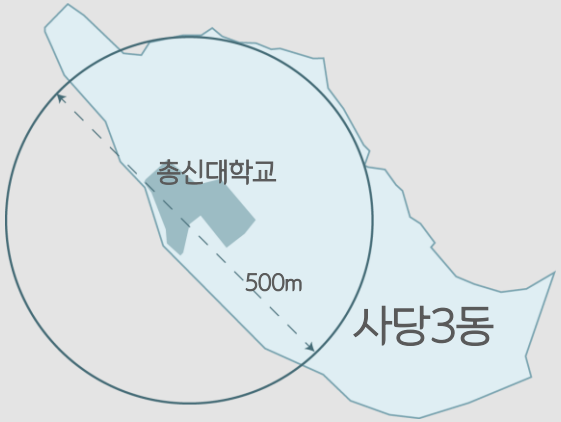
상권지역 선정

대학상권 기준

세가지 기준으로 범위설정

1. 대학교가 위치한 행정동
2. 정문 기준 반경 500m내외

   보행상권에 초점을 두고 기존 많은 연구에서 200m~1,000m까지 다양하게 설정이 되었지만, 소규모 점포들이 밀집해 있는 대학상권의 특징과 대학생들의 생활패턴을 반영하여 반경 500m로 지정
3. 점포수 10개 이하 상권 제외



2. 데이터

데이터 수집

데이터 전처리

데이터 수집 출처 및 사용 변수

서울 열린 데이터 광장에서 제공하는 우리마을가게 상권분석 서비스/ 공공데이터포털 / our World in Data
(http://data.seoul.go.kr/dataList/OA-15572/S/1/datasetView.do)

데이터	사용변수	시점	파일형식
상권영역	기준년코드,기준분기코드,상권코드	2015-2020 3분기	csv
추정매출	서비스업종코드,분기매출금액,분기매출건수	2015-2020 3분기	csv
추정유동인구	총유동인구수,성별유동인구수,연령별유동인구수,시간별유동인구수,요일별유동인구수	2015-2020 3분기	csv
상주인구	총상주인구수,성별상주인구수,연령별상주인구수	2015-2020 3분기	csv
직장인구	총직장인구수,성별직장인구수,연령별직장인구수	2015-2020 3분기	csv
점포	점포수,개폐업점포수,프랜차이즈점포수	2015-2020 3분기	csv
집객시설	집객시설수	2015-2020 3분기	csv
서울시 코로나 확진자수	서울시 코로나 확진자	2020_1월 ~ 3분기	csv
코로나 확진자	코로나 확진자_수	2020_1월 ~ 3분기	csv

추정매출 : 카드승인금액 / 보정비율
추정유동인구 : 서울시와 KT가 공공빅데이터와 통신데이터를 이용하여 추계한 인구

2. 데이터

데이터 수집

데이터 전처리

중앙대 (흑석로9길,흑석로13길,서달로8가길,서달로15길,서달로14길,흑석시장)
숭실대 (상도로61길,상도로62길,상도로47길,상도로37길,상도전통시장)
총신대 (사당로2차길,사당로8길,사당로16가길,사당로29길,남성역골목시장)

필요한 상권 남기기

›

trainset (2015-2020 2분기)
testset (2020 3분기)

데이터셋 나누기

›

상권명(중앙대,숭실대,총신대)
방학개월수 -> 대학특징 반영
(1분기 = 2, 2분기 = 0, 3분기는 2, 4분기는 0으로 구분)

코로나(전국/서울 확진자_수, 분기별)

column 추가

›

코로나 2015~2019 - > 0
점포수 0 -> 1
집객시설_수, 직장인구 NA값 0 으로 대체

NA값 처리

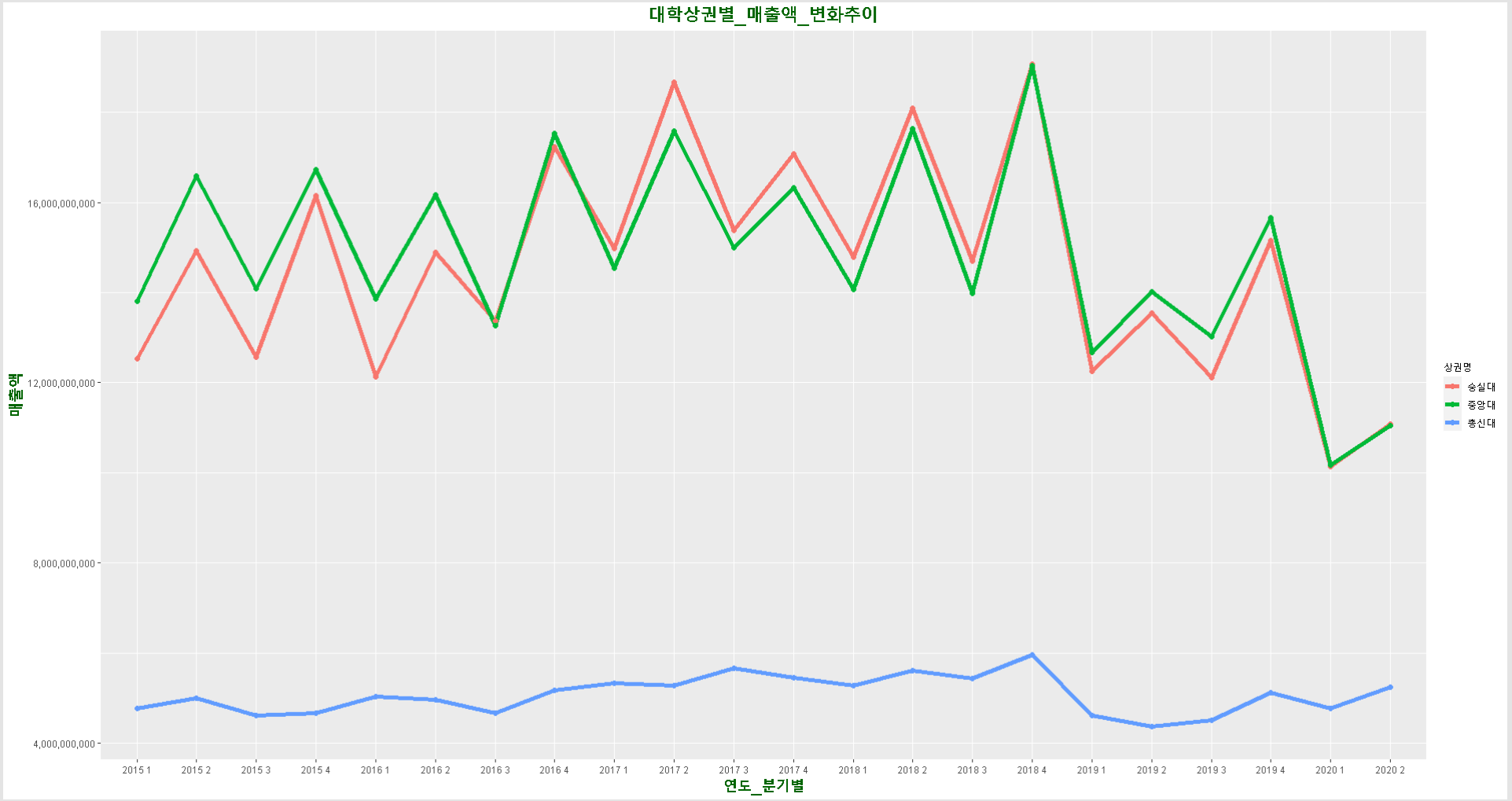
3. EDA

선정상권 특징

중앙대

숭실대

총신대



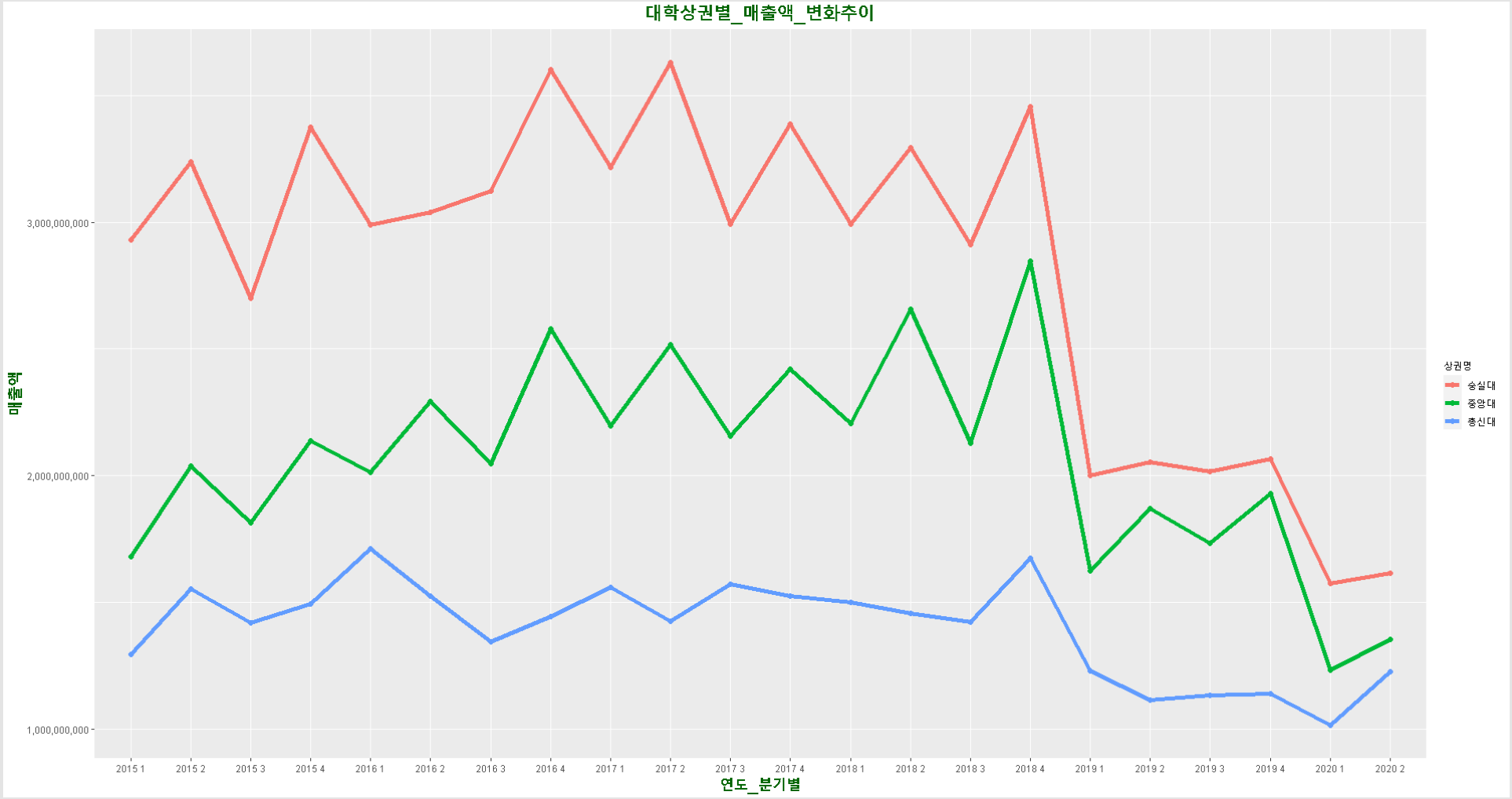
3. EDA

선정상권 특징

중앙대

숭실대

총신대



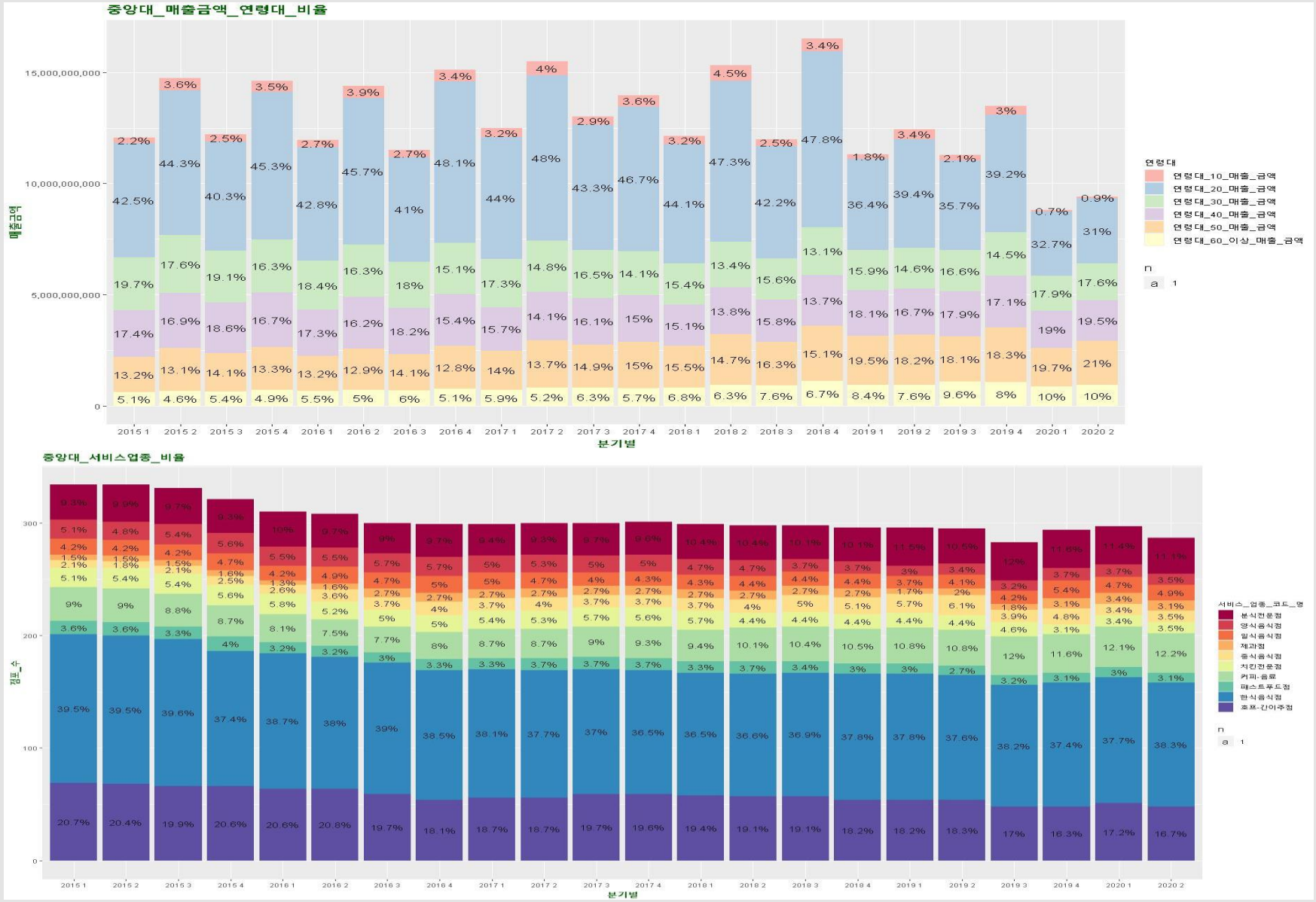
3. EDA

선정상권 특징

중앙대

숭실대

총신대



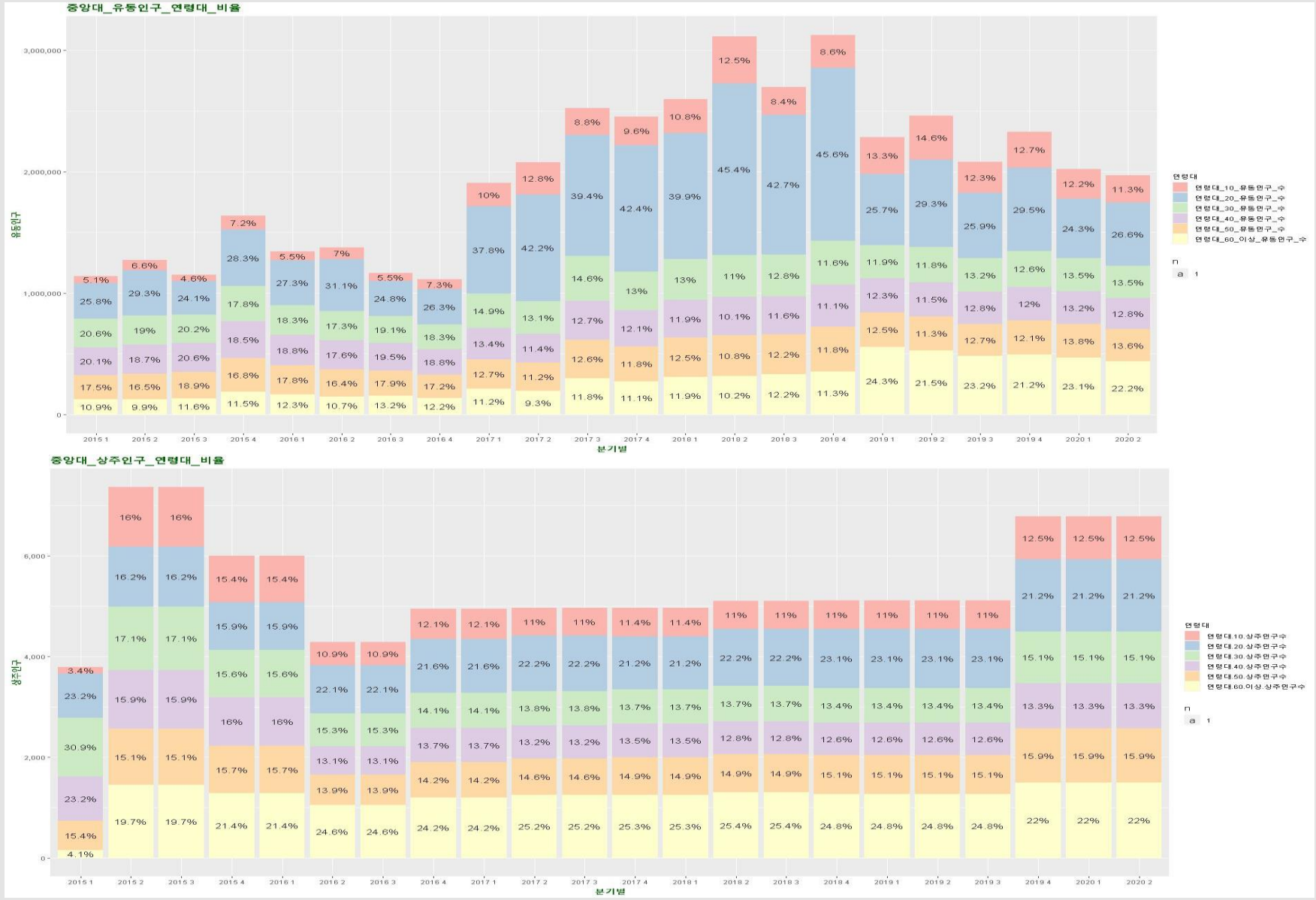
3. EDA

선정상권 특징

중앙대

숭실대

총신대



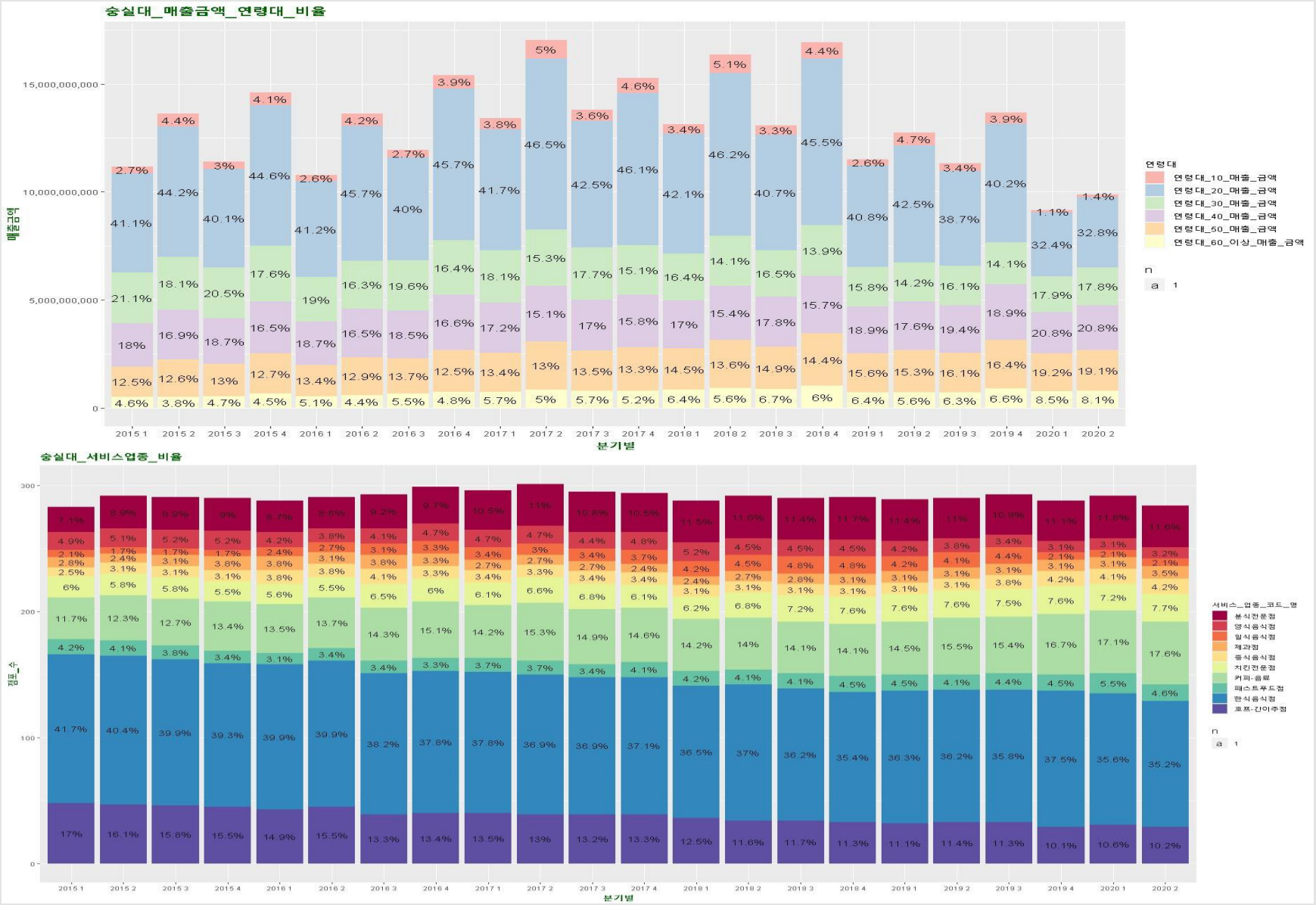
3. EDA

선정상권 특징

중앙대

숭실대

총신대



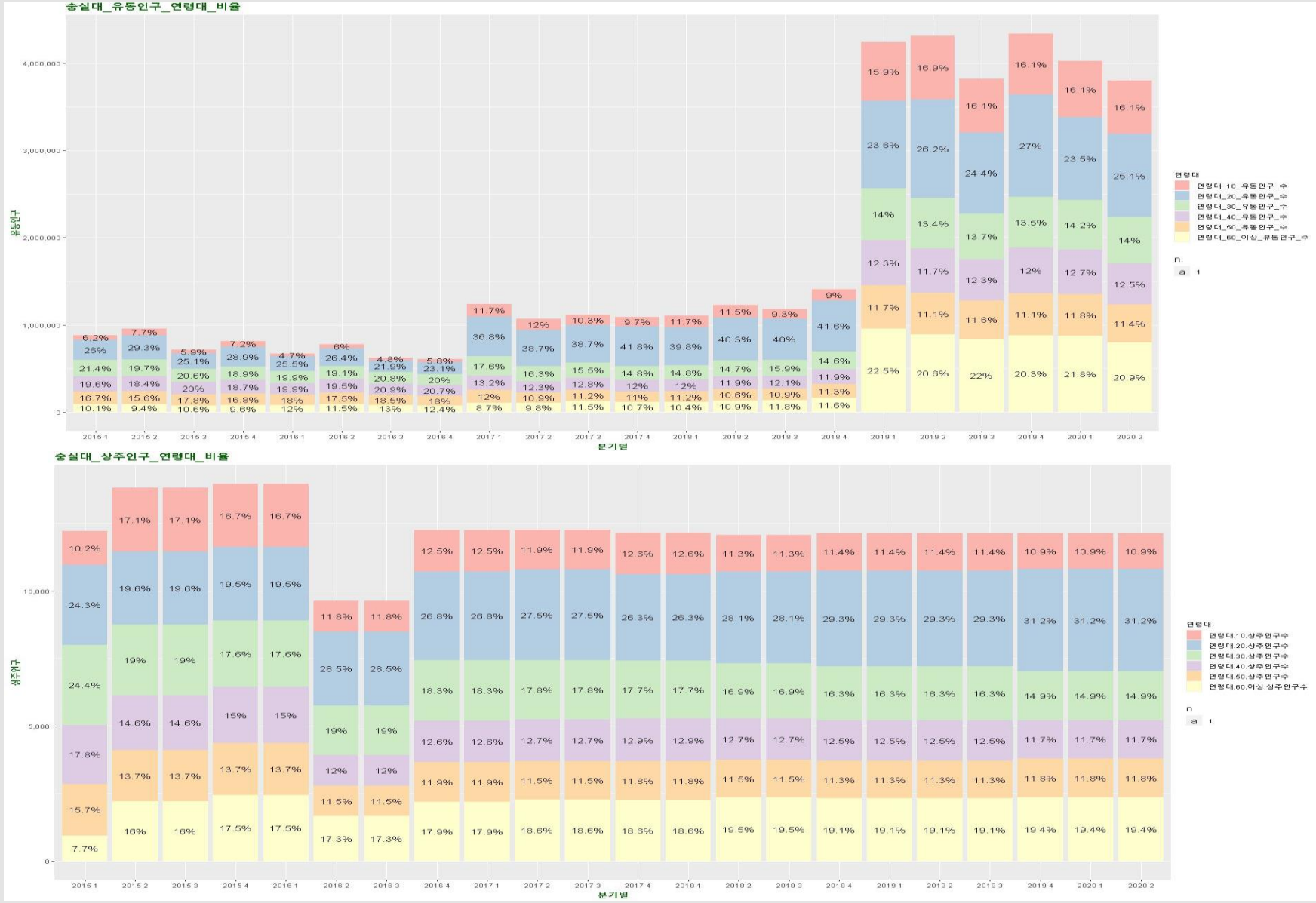
3. EDA

선정상권 특징

중앙대

숭실대

총신대



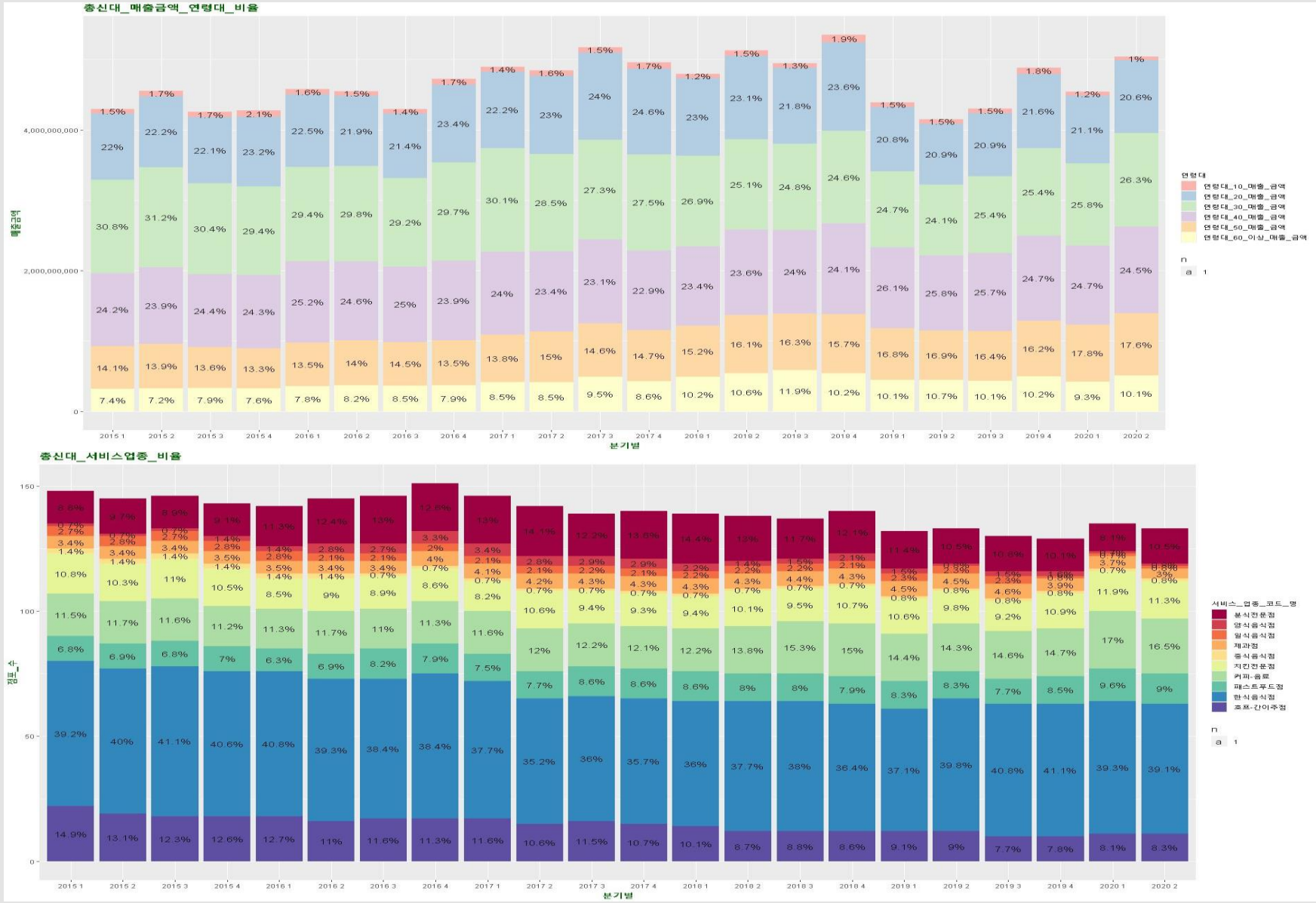
3. EDA

선정상권 특징

중앙대

숭실대

총신대



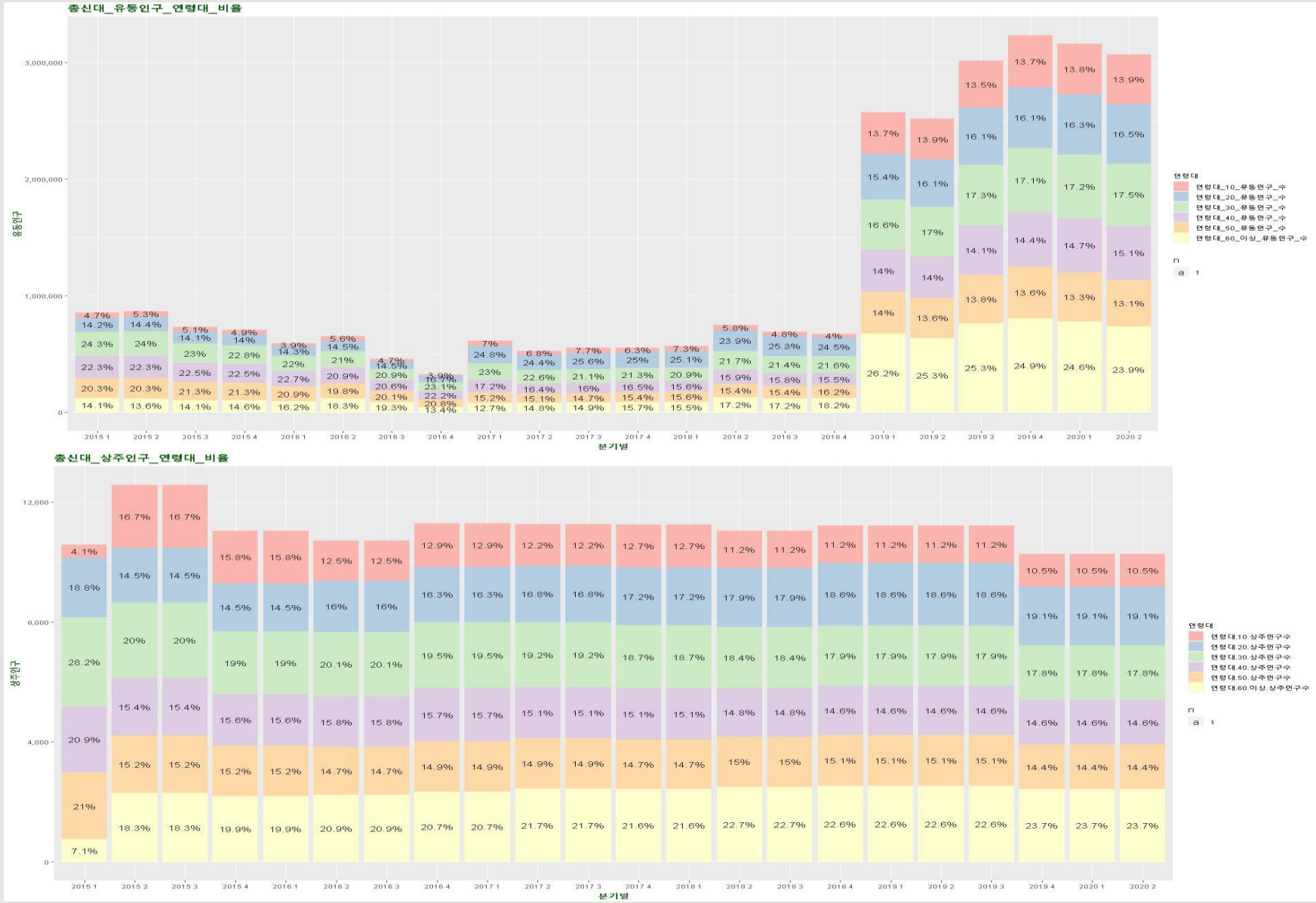
3. EDA

선정상권 특징

중앙대

숭실대

총신대



4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

map_lgl() 이용한 상관분석 실시

```r
train_data <- train_data %>% mutate(분기_평균매출금액 = 분기_매출_금액 / 점포_수)
colnames(train_data)

train_data <- train_data %>% select(-분기_매출_건수, -점포_수,-분기_매출_금액)

cor_data<-train_data[,-c(1:6,54)]
cor <- map_lgl(.x = cor_data, .f = function(x) {
  test<- cor.test(x= x,y= train_data$분기_평균매출금액)
  result<- test$p.value>0.05
  return(result)
})
cor
```

cor.test
귀무가설 : 변수끼리의 상관관계가 없다.
대립가설 : 변수끼리의 상관관계가 있다.

귀무, 대립가설의 기준이 되는 p-value = 0.05 를 사용하여 p값이 0.05이하가 되는 변수를 목표변수와 상관관계가 있다고 판단하여 컬럼축소 및 데이터세팅

4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

p.value>0.05 컬럼 삭제

총 유동인구수	남성 유동인구수	여성 유동인구수	10대 유동인구수	20대 유동인구수	30대 유동인구수	40대_유동 인구수	50대 유동인구수	60대 유동인구수	00~06시 유동인구
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
06_11시 유동인구	11_14시 유동인구	14_17시 유동인구	17_21시 유동인구	21_24시 유동인구	월요일 유동인구수	화요일 유동인구수	수요일 유동인구수	목요일 유동인구수	금요일 유동인구수
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
토요일 유동인구수	일요일 유동인구수	총 상주인구수	남성 상주인구수	여성 상주인구수	10대 상주인구수	20대 상주인구수	30대 상주인구수	40대 상주인구수	50대 상주인구수
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
60대 상주인구수	총 직장인구수	남성 직장인구수	여성 직장인구수	10대 직장인구수	20대 직장인구수	30대 직장인구수	40대 직장인구수	50대 직장인구수	60대 직장인구수
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
총 집객시설수	개업 점포_수	폐업 점포_수	총 확진자_수	서울 확진자수	방학 개월수				
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE				

4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

범주형 입력변수_ANOVA 분석 실시

귀무 : 그룹간의 평균의 차이가 없다.
대립 : 적어도 하나의 그룹의 평균이 다르다.

1. 서비스업종 분산분석 실시

가. 정규성 검증

정규성 검증은 관측값이 30개 이상이므로 중심극한정리에 의하여 생략한다.

나. 등분산성 검증

```
bartlett.test(분기_평균매출금액~as.factor(서비스_업종_코드_명), data = train_data)
```

```
        Bartlett test of homogeneity of variances

data:  분기_평균매출금액 by as.factor(서비스_업종_코드_명)
Bartlett's K-squared = 1880.8, df = 9, p-value < 2.2e-16
```

-> p값이 0.05이하 이므로 등분산성을 만족하지 않는다. -> val.equal = F로 진행

4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

범주형 입력변수_ANOVA 분석 실시

다. 분산분석 실시

```
oneway.test(분기_평균매출금액~as.factor(서비스_업종_코드_명), data = train_data, var.equal = F)
```

```
        One-way analysis of means (not assuming equal variances)

data:  분기_평균매출금액 and as.factor(서비스_업종_코드_명)
F = 55.405, num df = 9.00, denom df = 929.23, p-value < 2.2e-16
```

- p값이 0.05이하 이므로 대립가설 채택
- 대립가설 : 적어도 하나의 그룹간 평균이 다르다.

라. 사후검정 실시

```
Independent Variable:  as.factor(서비스_업종_코드_명)
 Factors        Means
 제과점         124831828.00914  a
 중식음식점     69907255.7750754  b
 패스트푸드점   69524436.6892552  b
 일식음식점     65508091.7027149  bc
 한식음식점     56858284.3861847   cd
 치킨전문점     50917011.3291309    d
 양식음식점     50831359.9817123    d
 호프-간이주점  36923831.3557489     e
 분식전문점     26681212.1771125     ef
 커피-음료      25285641.8291908      f
```

매출액에 대한 비슷한 패턴을 가지고 있는 그룹확인

4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

범주형 입력변수_ANOVA 분석 실시

2. 대학상권 분산분석 실시

가. 정규성 검증

정규성 검증은 관측값이 30개 이상이므로 중심극한정리에 의하여 생략한다.

나. 등분산성 검증

```
bartlett.test(분기_평균매출금액~as.factor(상권명), data = train_data)
```

```
        Bartlett test of homogeneity of variances

data:  분기_매출_금액 by as.factor(상권명)
Bartlett's K-squared = 743.1, df = 2, p-value < 2.2e-16
```

- p값이 0.05이하 이므로 등분산성을 만족하지 않는다. -> val.equal = F로 진행

4. 모형

상관관계분석

회귀나무

랜덤포레스트

총신대

범주형 입력변수_ANOVA 분석 실시

다. 분산분석 실시

```
oneway.test(분기_매출_금액~as.factor(상권명), data = merge_data, var.equal = F)
```

```
        One-way analysis of means (not assuming equal variances)

data:  분기_매출_금액 and as.factor(상권명)
F = 28.411, num df = 2.0, denom df = 1656.2, p-value = 7.386e-13
```

- p값이 0.05이하 이므로 대립가설 채택
- 대립가설 : 적어도 하나의 그룹간 평균이 다르다.

라. 사후검정 실시

```
Independent Variable:  as.factor(상권명)
 Factors  Means
 숭실대  66331382.501233  a
 중앙대  49057227.8080132  b
 총신대  41806081.988777    c
```

대학상권마다 다른 그룹으로 나타나는 것을 확인

4. 모형

상관관계분석

회귀나무

랜덤포레스트

회귀나무모형적합

```
#정지규칙
ctrl<- rpart.control(minsplit = 10,
                     cp = 0.001,
                     maxdepth = 30)
#모형적합
set.seed(1234)
fit1<-rpart(formula = (분기_매출_금액/점포_수)~.,
            data = train_mod,
            control = ctrl)
#가지치기
which.min(fit1$cptable[,4])
cp<- fit1$cptable[62,1]
fit2<- prune.rpart(tree=fit1,cp=cp)
#실제값 할당
real<- (test_mod$분기_매출_금액/test_mod$점포_수)
pred2<- predict(fit2, newdata = test_mod, type = 'vector')
```

```
Variable importance

연령대.30.상주인구수 연령대.10.상주인구수 연령대.20.상주인구수 서비스_업종_코드
           12           12           10            9
상권_코드 총_집객시설_수 총_유동인구_수 연령대_30_유동인구_수
           7            6            6            5
연령대_60_이상_직장인구_수 시간대_17_21_유동인구_수 여성_유동인구_수 남성_유동인구_수
           4            3            3            3
시간대_11_14_유동인구_수 시간대_14_17_유동인구_수 연령대_40_유동인구_수 연령대_20_유동인구_수
           3            3            3            2
기준_년_코드 연령대_50_유동인구_수 화요일_유동인구_수 월요일_유동인구_수
           2            2            2            1
수요일_유동인구_수 상권명 금요일_유동인구_수
           1            1            1
```

4. 모형

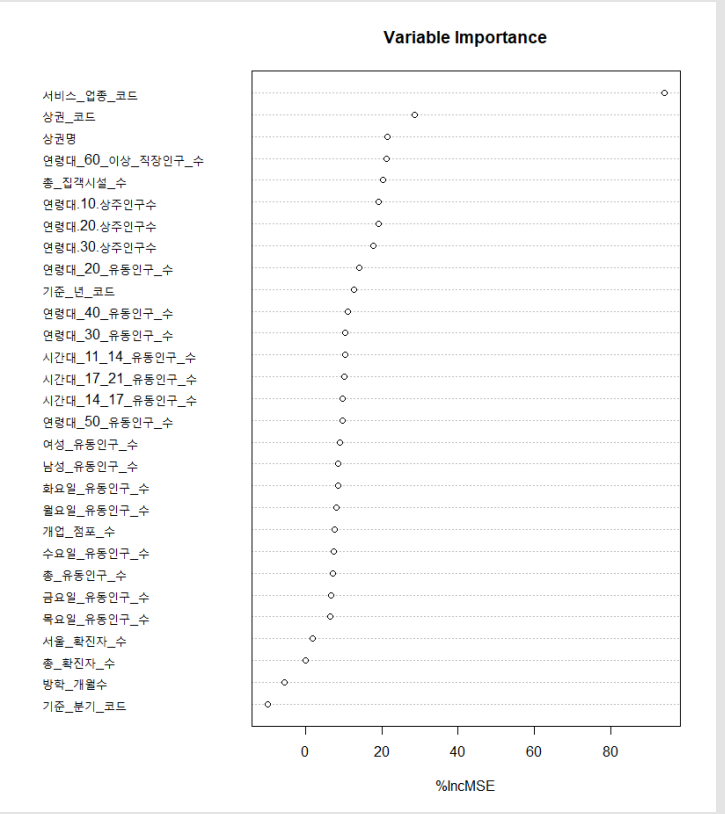
상관관계분석

회귀나무

랜덤포레스트

랜덤포레스트모형적합

```
#모형적합
set.seed(1234)
fit1<-randomForest(formula = (분기_매출_금액/점포_수)~.,
                   data = train_mod,
                   ntree = 1000,
                   mtry = 10,
                   importance = TRUE,
                   do.trace = 50,
                   keep.forest = T)
#그리드생성 후 튜닝
grid<-expand.grid(ntree=c(300,500,700,1000),
                  mtry=c(3:10),
                  error=NA)
n<- nrow(grid)
for(i in 1:n){
  ntree<- grid$ntree[i]
  mtry<- grid$mtry[i]
  disp<- str_glue('현재 {i}번째 행 실행중! [ntree: {ntree},
                   mtry : {mtry}]')
  cat(disp,'\n\n')
  set.seed(1234)
  fit<- randomForest(formula =(분기_매출_금액/점포_수)~.,
                     data = train_mod,
                     ntree = ntree,
                     mtry = mtry)
  grid$error[i] <- tail(x = fit$mse, n = 1)
}
#하이퍼파라미터로 모형적합
loc<- which.min(grid$error)
bestPara<- grid[loc,]
set.seed(1234)
best<-randomForest(formula = (분기_매출_금액/점포_수)~.,
                   data = train_mod,
                   ntree = bestPara$ntree,
                   mtry = bestPara$mtry,
                   importance = TRUE)
pred2<-predict(best,newdata = test_mod,type = 'response')
```



랜덤포레스트모형이 회귀나무모형에 비해
RMSE값 **20.42%** 감소

4. 모형

상관관계분석

회귀나무

랜덤포레스트

주요 입력변수의 영향력 측정

1. 20대 유동인구수, 20대 상주인구수 제거한 모형

:RMSE값 3.54 % 증가 ↑

2. 코로나관련변수(총확진자수, 서울시확진자수) 제거한 모형

:RMSE값 3.84 % 감소 ↓

»»

목표변수인 매출액을 추정하는데 있어

20대 관련 컬럼은 입력변수로서 유의미한 결과 도출

코로나 관련 컬럼은 입력변수로서 유의미하지 않음

5. 결론

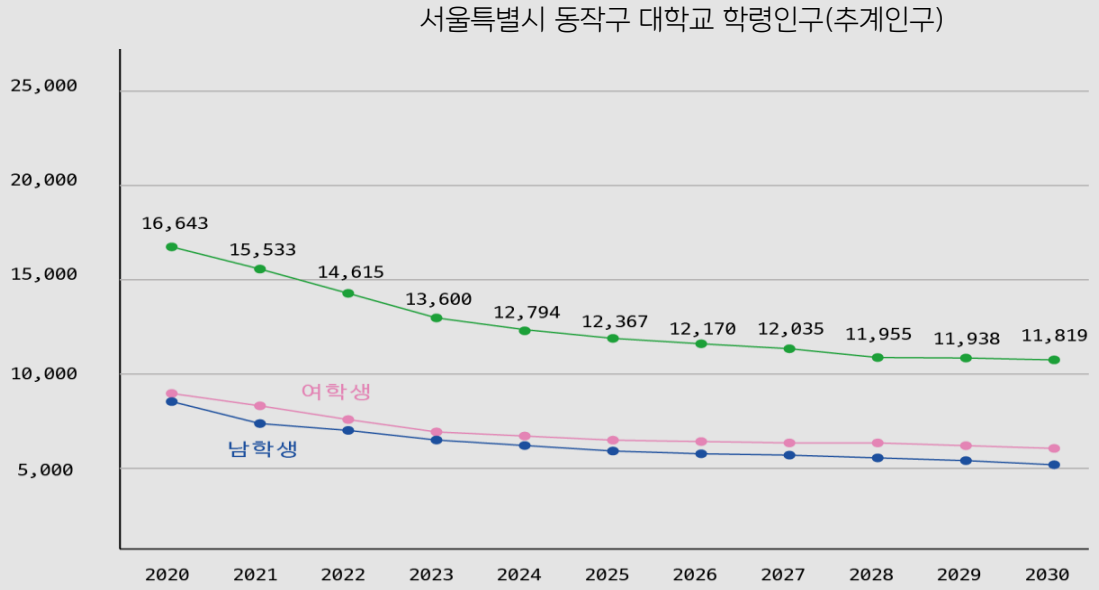
결론 및 제안

한계

1. EDA 결과를 통해 동작구에 위치한 세 개 대학상권의 특징을 파악

2. 모델결과를 통해 매출액에 영향을 주는 중요 변수를 파악하여
2020년 3분기 매출액 예측 및 대학상권 창업 시 참고할 수 있는 모델을 구축

3. 상관분석에서 '목표변수와의 상관관계가 없다' 라는 결론으로 탈락된 변수를
잠재고객, 타겟요소로 재해석하는 새로운 비즈니스 형태를 제안함



서울특별시 빅데이터담당관
「서울특별시 자치구별 장래인구추계
대학교 학령인구」 2020.06

향후 대학교 학령인구의 지속적인 감소가 예측됨에 따라 대학상권의 매출 역시 하락세 보일 것으로 예상

5. 결론

결론 및 제안

한계

1. 데이터 수집의 한계

-상주인구와 직장인구 데이터의 경우 수집빈도가 1년에 2회로 매우 적었음
-점포수 데이터의 경우 매출액과 상응하지 않는 결함 존재
-모든 데이터의 단위가 분기로 설정되어 변화 추세(월별)와 대학상권 특징(개강, 방학)을 반영하는데 한계점이 있었음

2. 데이터 활용의 한계

학습데이터가 2020년 2분기까지로 설정되어 코로나의 추세가 극도로 심해진 2020년 3,4분기를 반영하지 못하였음

3. 모델 활용의 한계

각 대학상권마다 지역적 특성이 분명하기 때문에 보편적으로 사용되기에 어려움이 있다고 판단됨

4. 기타 한계점

정성적 데이터가 추가되면 보다 정확한 분석이 가능할 것으로 추정 (맛, 친절도, 고객만족도 등)

감사합니다.