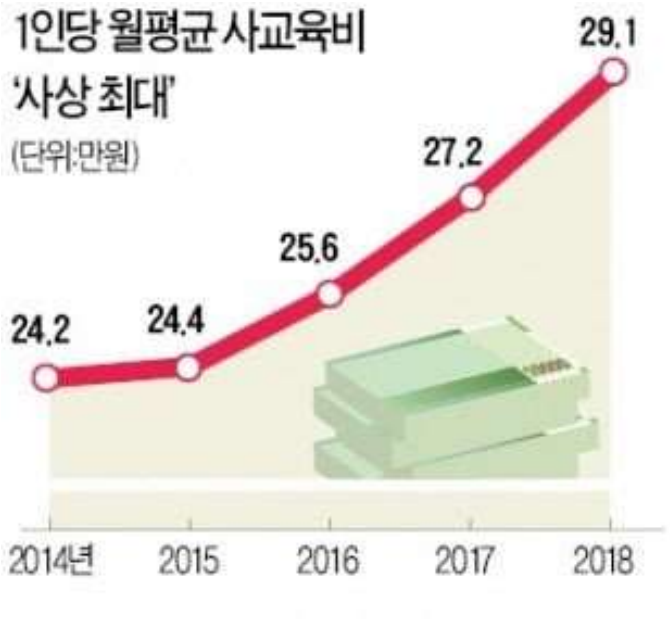
서울특별시 자치구별 학구열 분석 & 서울런 교육 사업

헬로월드조
정가연 김지은 배지환 박수현 허지원

CONTENTS

주제 배경

출처 : 한경닷컴



우리나라가 교육에 대해 매우 민감하고 예민하다는 것을 여러 매체들을 통해서 알 수 있다

사용 데이터 소개

메인 데이터	서브 데이터	외부 데이터	제작 데이터
서울특별시 학원 교습소정보.csv	서울시 가구형태별 가구 및 가구원 (동별) 통계.csv	2021 서울시 학급당 학생 수 통계. csv	helloworld_data_set.xlsx
서울특별시 학교 기본정보.csv	서울시 주민등록인구 (동별) 통계.csv		
서울특별시 유치원 일반현황.csv	서울특별시 유치원 통학차량 현황.csv		
서울시 주민등록연앙인구(연령별동별) 통계.csv	자치구별 1인당 지역내총생산 및 수준지수.csv		

사용 데이터 소개

메인 데이터	서브 데이터	외부 데이터	제작 데이터
서울특별시 학원 교습소정보.csv	서울시 가구형태별 가구 및 가구원 (동별) 통계.csv	2021 서울시 학급당 학생 수 통계. csv	helloworld_data_set.xlsx
서울특별시 학교 기본정보.csv	서울시 주민등록인구 (동별) 통계.csv		
서울특별시 유치원 일반현황.csv	서울특별시 유치원 통학차량 현황.csv		
서울시 주민등록연앙인구(연령별동별) 통계.csv	자치구별 1인당 지역내총생산 및 수준지수.csv		

사용 데이터 소개

메인 데이터	서브 데이터	외부 데이터	제작 데이터
서울특별시 학원 교습소정보.csv	서울시 가구형태별 가구 및 가구원 (동별) 통계.csv	2021 서울시 학급당 학생 수 통계. csv	helloworld_data_set.xlsx
서울특별시 학교 기본정보.csv	서울시 주민등록인구 (동별) 통계.csv		
서울특별시 유치원 일반현황.csv	서울특별시 유치원 통학차량 현황.csv		
서울시 주민등록연앙인구(연령별동별) 통계.csv	자치구별 1인당 지역내총생산 및 수준지수.csv		

사용 데이터 소개

메인 데이터	서브 데이터	외부 데이터	제작 데이터
서울특별시 학원 교습소정보.csv	서울시 가구형태별 가구 및 가구원 (동별) 통계.csv	2021 서울시 학급당 학생 수 통계. csv	helloworld_data_set.xlsx
서울특별시 학교 기본정보.csv	서울시 주민등록인구 (동별) 통계.csv		
서울특별시 유치원 일반현황.csv	서울특별시 유치원 통학차량 현황.csv		
서울시 주민등록연앙인구(연령별동별) 통계.csv	자치구별 1인당 지역내총생산 및 수준지수.csv		

전처리 과정

Main Data 1

01

```
# 도로명상세주소 + 도로명주소 + 교습계열명 + 교습과정명 + 일시수용능력인원합계 결측치 포함 행 제거
academy = academy.dropna(subset=['도로명상세주소', '교습계열명', '일시수용능력인원합계'])

academy['행정구역명'] = academy['행정구역명'].fillna(academy['도로명주소'].str.split(' ').str[1])
academy['인당수강료내용'] = academy['인당수강료내용'].fillna('알수없음')
academy['기숙사학원여부'] = academy['기숙사학원여부'].fillna('알수없음')
academy['휴원시작일자'] = academy['휴원시작일자'].fillna(0)
academy['휴원종료일자'] = academy['휴원종료일자'].fillna(0)

# '서울특별시마포구'같이 띄어쓰기가 안 되어 있는 데이터가 있어 '서울특별시'를 제거하고 앞의 공백을 제거
academy['도로명주소'] = academy['도로명주소'].str.replace('서울특별시', '')
academy['도로명주소'].str.lstrip().tolist()
```

결측치 제거

02

```
academy.duplicated().sum() # 중복값 1개 확인
```

```
1
```

```
academy.shape # 중복값 제거 전 행,열 개수
```

```
(24984, 16)
```

```
academy.drop_duplicates(inplace=True) # 중복값 제거
```

```
academy.shape # 중복값 제거 후 행, 열 개수
```

```
(24983, 16)
```

중복값 제거

03

```
academy['휴원종료일자'].unique() # 이상치 99999999
```

```
# 이?미용 -> 이 · 미용, 보습?논술 -> 보습 · 논술 / 데이터 출처 찾아서 ?를 · 로 수정
academy['교습과정명'] = academy['교습과정명'].str.replace('?', ' · ')
academy['교습계열명'] = academy['교습계열명'].str.replace('?', ' · ')
```

```
# 등록일자, 개설일자 Dtype object에서 datetime으로 변경
academy['등록일자'] = pd.to_datetime(academy['등록일자'], format='%Y%m%d')
academy['개설일자'] = pd.to_datetime(academy['개설일자'], format='%Y%m%d')
```

```
# 전처리와 이상치 처리 후 인덱스 번호 재설정
academy.reset_index(drop=True, inplace=True)
```

이상치 처리

전처리 과정

Main Data 2

01

```
# 해당 열들은 대부분의 값이 비어있음  -> 유용하지 않은 열이므로 삭제
# 고등학교구분명, 특수목적고등학교계열명, 계열명, 학과명
del school['고등학교구분명']
del school['특수목적고등학교계열명']
del school['계열명']
del school['학과명']
del school['도로명상세주소']
```

```
school['주소구']= school['도로명주소'].str.split().str[1] # 자치구만 뽑아내기
school['주소구']
```

```
# 결측치 행 제거
school= school.dropna(subset=['도로명주소', '주소구'])
```

결측치 제거

02

```
school.duplicated().sum()
```

```
school.shape
```

```
(3867, 15)
```

```
school.drop_duplicates(inplace=True)
```

```
school.shape
```

```
(1446, 15)
```

중복값 제거

03

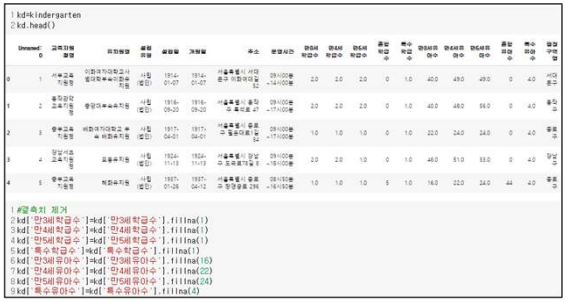
```
#이상치 없음
```

이상치 처리

전처리 과정

Main Data 3

01



```
kd=kindergarten
kd.head()
```

Unnamed: 0	교육지원청명	유치원명	설립유형	설립일	개원일	주소	운영시간	만0세학급수	만1세학급수	만2세학급수	혼합학급수	특수학급수	만0세유아수	만1세유아수	만2세유아수	혼합유아수	특수유아수	입학구비일
0	1	서부교육지원청	사립(법인)	1914-01-07	1914-01-07	서울특별시 서대문구 이화여자대학교	09시00분~14시00분	2.0	2.0	2.0	0	1.0	40.0	49.0	49.0	0	4.0	서대문구
1	2	동작관악교육지원청	사립(법인)	1916-09-20	1916-09-20	서울특별시 동작구 흑석로 47	09시00분~17시00분	2.0	2.0	2.0	0	1.0	48.0	48.0	56.0	0	4.0	동작구
2	3	중부교육지원청	사립(법인)	1917-04-01	1917-04-01	서울특별시 중구 동호이로 [illegible]	09시00분~17시00분	1.0	1.0	1.0	0	1.0	22.0	24.0	24.0	0	4.0	중구
3	4	양교서초교육지원청	사립(법인)	1924-11-13	1924-11-13	서울특별시 양남구 도곡로 [illegible]	09시00분~16시00분	2.0	2.0	1.0	0	1.0	46.0	51.0	33.0	0	4.0	양남구
4	5	중부교육지원청	사립(법인)	1927-01-26	1927-04-12	서울특별시 중구 청계천로 [illegible]	08시50분~16시50분	1.0	1.0	1.0	5	1.0	16.0	22.0	24.0	44	4.0	중구

```
#결측치 제거
kd['만3세학급수'].fillna(1)
kd['만4세학급수'].fillna(1)
kd['만5세학급수'].fillna(1)
kd['특수학급수'].fillna(1)
kd['만3세유아수'].fillna(16)
kd['만4세유아수'].fillna(22)
kd['만5세유아수'].fillna(24)
kd['특수유아수'].fillna(4)
```

결측치 제거

02

```
In [14]: kd.duplicated().sum()
Out[14]: 0
```

중복값 제거

03

```
#이상치 처리
kd['혼합학급수']=kd['혼합학급수'].clip(0,10)
kd['혼합유아수']=kd['혼합유아수'].clip(0, 44)
```

이상치 처리

전처리 과정

Main Data 4

01

```
#결측치 중앙값으로 대체
pop_age['0~4세'] = pop_age['0~4세'].fillna(pop_age['0~4세'].median())
pop_age['0~4세'] = pop_age['0~4세'].astype('int')
pop_age['5~9세'] = pop_age['5~9세'].fillna(pop_age['5~9세'].median())
pop_age['5~9세'] = pop_age['5~9세'].astype('int')
pop_age['10~14세'] = pop_age['10~14세'].fillna(pop_age['10~14세'].median())
pop_age['10~14세'] = pop_age['10~14세'].astype('int')
pop_age['15~19세'] = pop_age['15~19세'].fillna(pop_age['15~19세'].median())
pop_age['15~19세'] = pop_age['15~19세'].astype('int')
pop_age['20~24세'] = pop_age['20~24세'].fillna(pop_age['20~24세'].median())
pop_age['20~24세'] = pop_age['20~24세'].astype('int')
pop_age['30~34세'] = pop_age['30~34세'].fillna(pop_age['30~34세'].median())
pop_age['30~34세'] = pop_age['30~34세'].astype('int')
pop_age['45~49세'] = pop_age['45~49세'].fillna(pop_age['45~49세'].median())
pop_age['45~49세'] = pop_age['45~49세'].astype('int')
pop_age['50~54세'] = pop_age['50~54세'].fillna(pop_age['50~54세'].median())
pop_age['50~54세'] = pop_age['50~54세'].astype('int')
pop_age['100세 이상+'] = pop_age['100세 이상+'].fillna(0)
pop_age['100세 이상+'] = pop_age['100세 이상+'].astype('int')
```

결측치 제거

02

```
#중복값 삭제
#자치구열에 합계 행 삭제
all = pop_age[pop_age['자치구'].str.contains('합계')].index
pop_age.drop(all, inplace=True)
```

중복값 제거

03

```
#이상치 없음
```

이상치 처리

전처리 과정

Sub Data 1

01

```
#첫번째 행 삭제
family = family.drop([0,1])

#총가구를 수치형으로 변경
family['총가구'] = family['총가구'].astype('int')

#1인가구를 수치형으로 변경
family['1인가구'] = family['1인가구'].astype('int')

#2인가구를 수치형으로 변경
family['2인가구'] = family['2인가구'].st
try:
    family['2인가구'] = family['2인가구
except:
    family['2인가구'] = family['2인가구'].astype('float')
```

```
#결측치 확인 -> 없음
family.isnull().sum()
```

```
    family['4인가구'] = family['4인가구'].astype('float')
except:
    family['7인이상가구'] = family['7인이상가구'].astype('float')
```

결측치 제거

02

```
# 중복값 처리
family = family.drop_duplicates()
```

중복값 제거

03

```
#이상치 없음
family['자치구'].value_counts()
```

```
송파구      28
강남구      23
관악구      22
성북구      21
강서구      21
노원구      20
강동구      19
서초구      19
영등포구     19
양천구      19
종로구      18
성동구      18
은평구      17
마포구      17
구로구      17
중랑구      17
용산구      17
중구       16
동작구      16
광진구      16
도봉구      15
동대문구     15
서대문구     15
강북구      14
금천구      11
Name: 자치구, dtype: int64
```

이상치 처리

전처리 과정

Sub Data 2

01

```
#첫번째,두번째,세번째 행 삭제
pop_dong = pop_dong.drop([0,1,2])
pop_dong.reset_index(drop=True)
```

```
pop_dong['세대'] = pop_dong['세대'].str.replace(',','')
pop_dong['세대'].unique()
try:
    pop_dong['세대'] = pop_dong['세대'].astype('int')
except:
    pop_dong['세대'] = pop_dong['세대'].astype('float')

pop_dong['총인구
pop_dong['총인구
try:
    pop_dong['총
except:
    pop_dong['총

pop_dong['남자인구'] = pop_dong['남자인구'].str.replace(',','')
pop_dong['남자인구'].unique()
try:
    pop_dong['남자인구'] = pop_dong['남자인구'].astype('int')
except:
    pop_dong['남자인구'] = pop_dong['남자인구'].astype('float')
```

```
#결측치 확인 -> 없음
pop_dong.isnull().sum()
```

결측치 제거

02

```
# 중복값 처리
pop_dong = pop_dong.drop_duplicates()
```

중복값 제거

03

```
#이상치 없음
pop_dong['자치구'].value_counts()
```

```
#이상치 없음
pop_dong['동'].value_counts()
```

이상치 처리

전처리 과정

Sub Data 3

01

```
#결측치 확인 -> 없음
kinder_bus.isnull().sum()
```

결측치 제거

02

```
# 중복값 처리
kinder_bus = kinder_bus.drop_duplicates()
```

중복값 제거

03

```
#이상치 없음
kinder_bus['교육지원청명'].value_counts()
kinder_bus['유치원명'].value_counts()
kinder_bus['설립유형'].value_counts()
kinder_bus['주소'].value_counts()
kinder_bus.describe().T
```

이상치 처리

전처리 과정

Sub Data 4

01

```
#결측치 확인 -> 없음
grdp.isnull().sum()
```

결측치 제거

02

```
# 중복값 처리
grdp = grdp.drop_duplicates()
```

중복값 제거

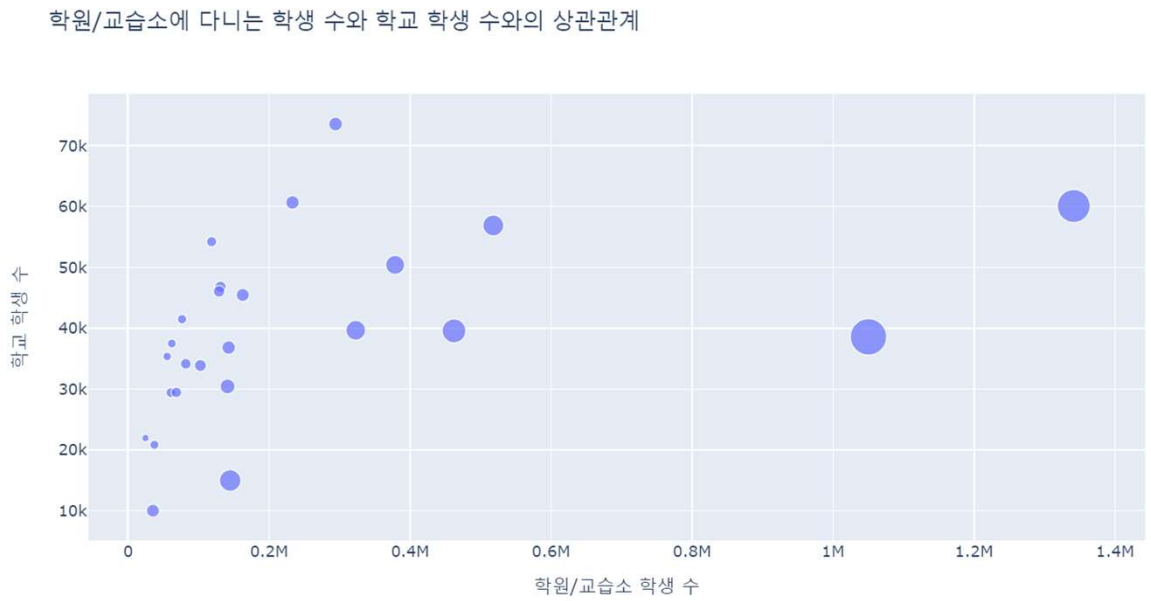
03

```
#이상치 확인 -> 없음
grdp.value_counts()
```

이상치 처리

분석 과정

주제 1 - 자치구별 학교 다니는 학생수 비율, 자치구별 학원 다니는 학생수 비율



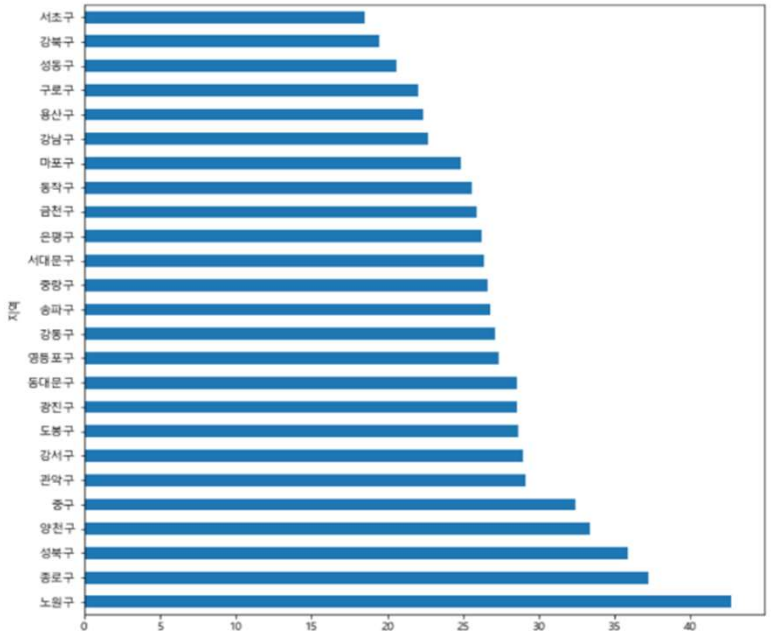
1) 자치구별 grdp 총 수와 자치구별 학교수의 상관관계

1. 자치구별 유치원, 초등학교, 중학교, 고등학교에 다니는 학생 수 산출
2. 자치구별 연령별 총 인구 수 산출
3. 연령별로 학교에 다니는 학생 수/ 전체 학생 수(교육열)를 구함
4. 교육열을 정렬한 뒤 막대그래프로 시각화
5. 연령별 총 학생 수와 학교에 다니는 학생 수 간의 상관관계를 나타냄

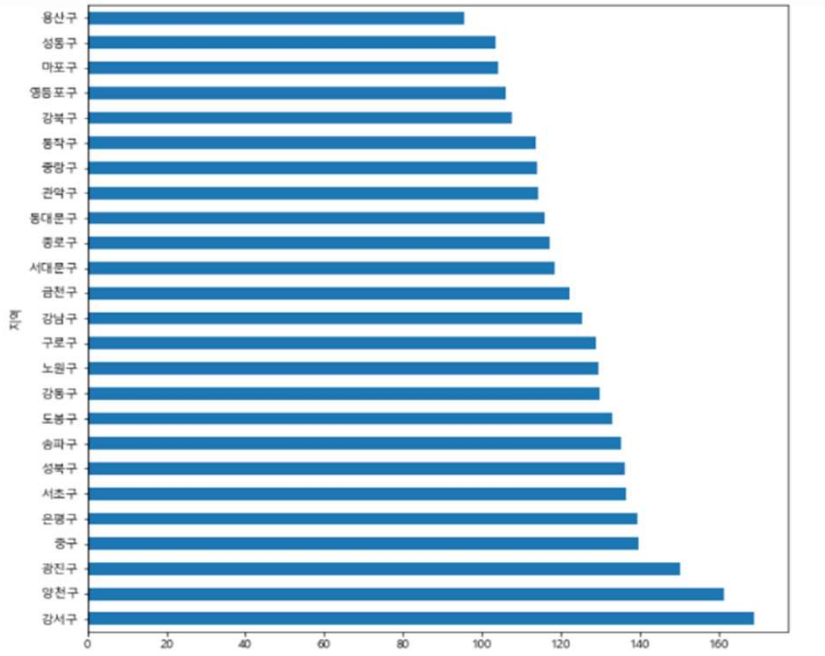
분석 과정

주제 1 - 자치구별 학교 다니는 학생수 비율, 자치구별 학원 다니는 학생수 비율

2) 유·초·중·고별 교육열 시각화



유치원교육열

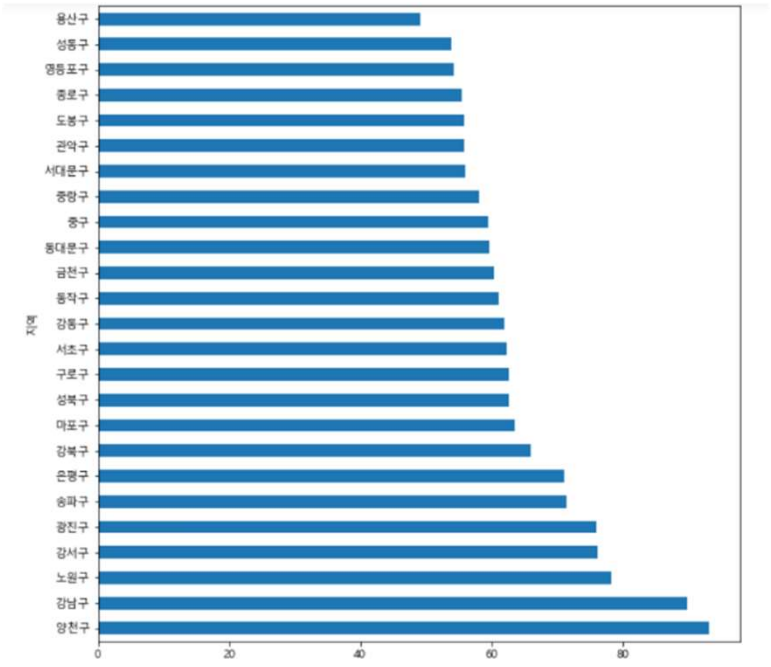


초등교육열

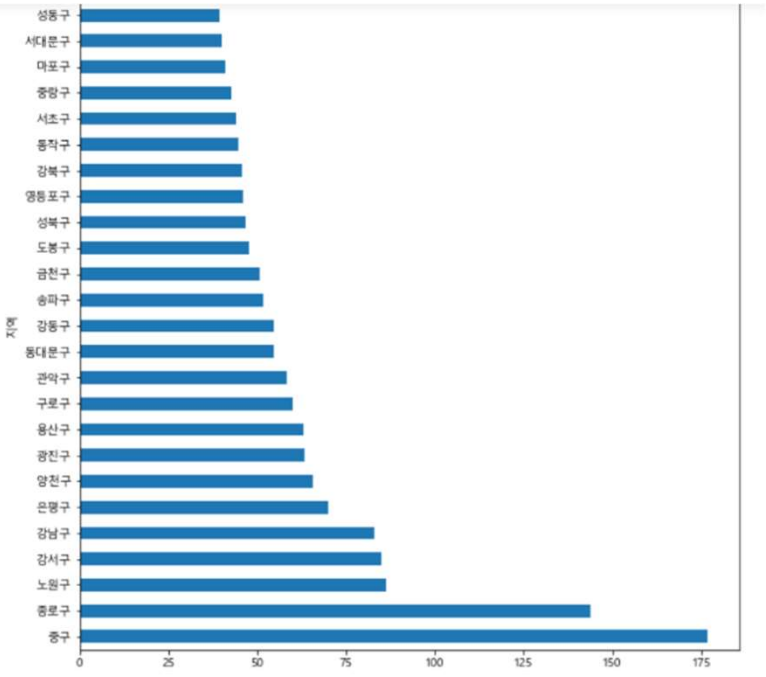
분석 과정

주제 1 - 자치구별 학교 다니는 학생수 비율, 자치구별 학원 다니는 학생수 비율

2) 유·초·중·고별 교육열 시각화



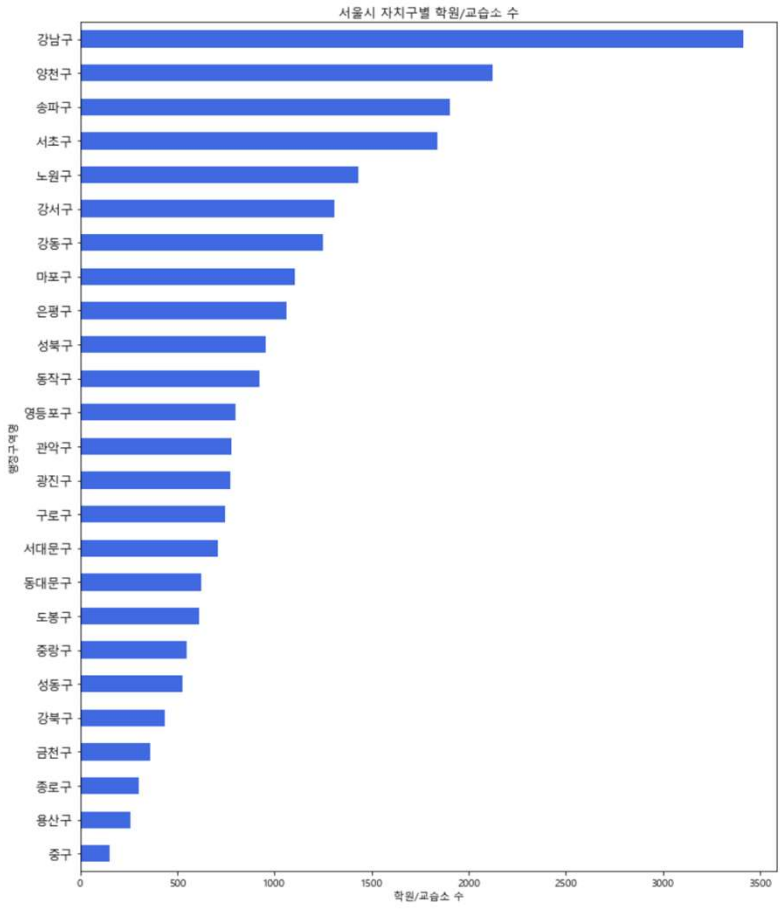
중등교육열



고등교육열

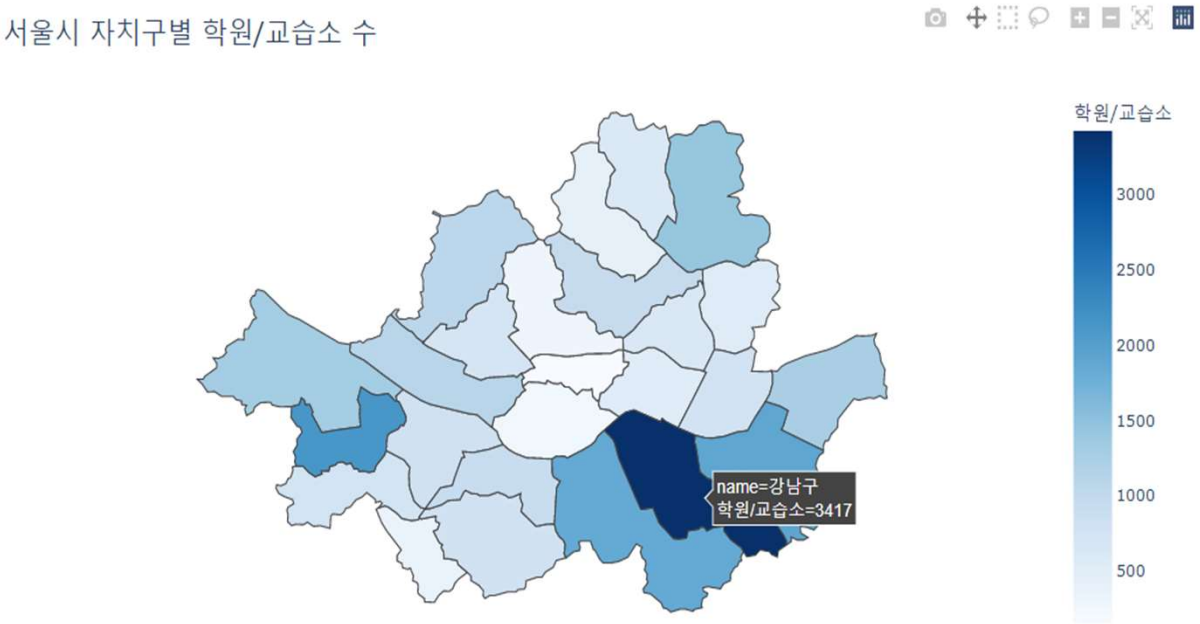
분석 과정

주제 2 – 강남 8학군은 실제로 학구열이 높을까?



1) 자치구별 학원/교습소 수 비교

1. 자치구별 학원/교습소 수 합계 산출



분석 과정

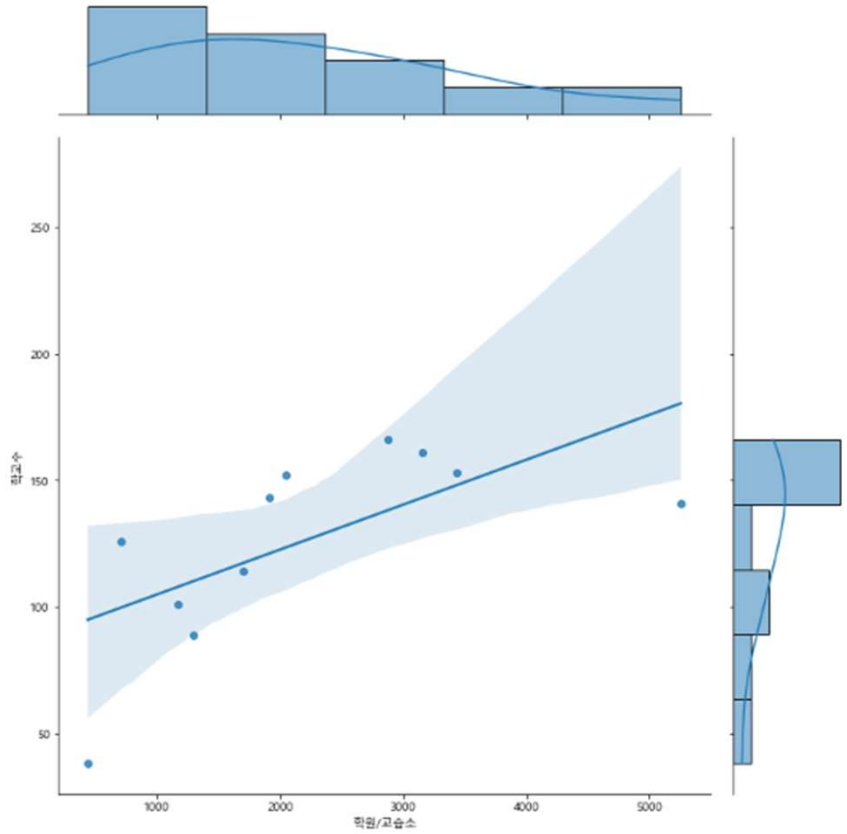
주제 2 – 강남 8학군은 실제로 학구열이 높을까?

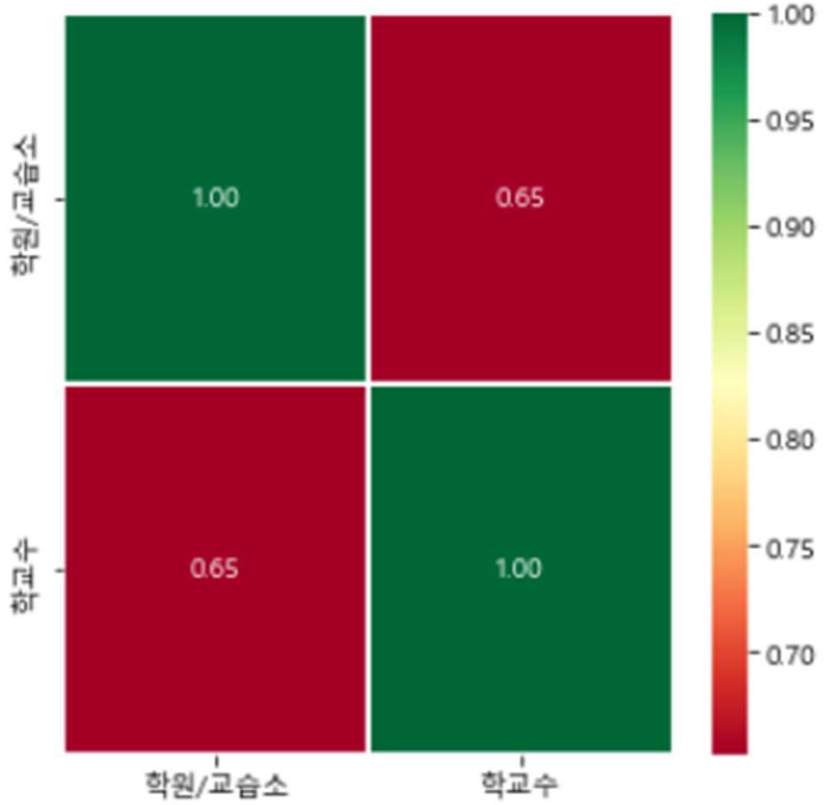
2) 학군별 학원/교습소 수 & 학교 수 비교

1. 자치구별 해당하는 학군 열 추가 (강남구, 서초구 〉 8학군)
2. 학군별 학원 수 합계 산출

분석 과정

주제 2 – 강남 8학군은 실제로 학구열이 높을까?





분석 과정

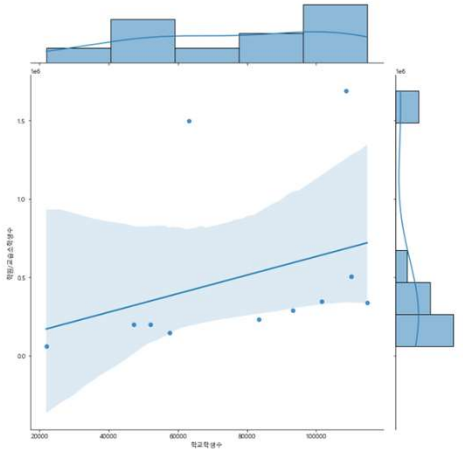
주제 2 – 강남 8학군은 실제로 학구열이 높을까?

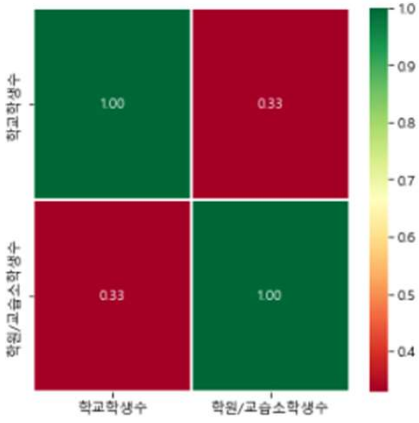
학군별 학원/교습소 학생 수 / 학교 학생 수 비율

3) 학군별 학원/교습소 학생 수 & 학교 학생 수 비교

1. 온라인, 원격 키워드 포함한 학원명 행 제거
(실제 등원하는 학생 수 데이터가 필요)
2. 학군별 학원/교습소 학생 수 합계 산출

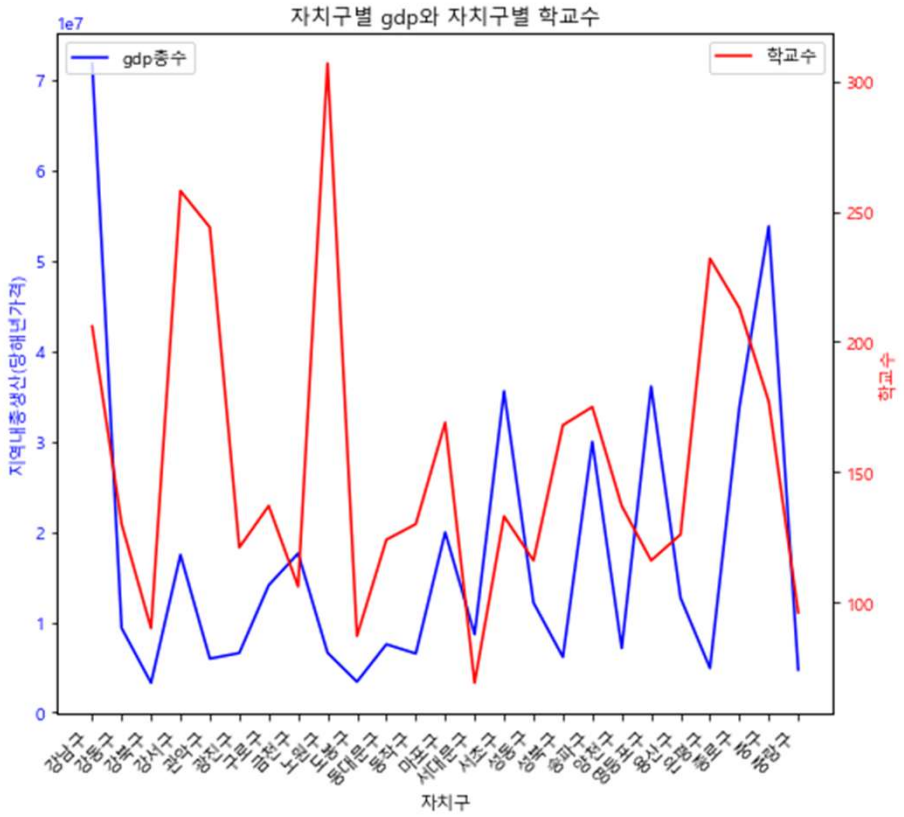
서울시 학군별 학원/교습소 & 학교 학생 수





분석 과정

주제 3 - grdp가 높은 자치구 학구열 높을까?

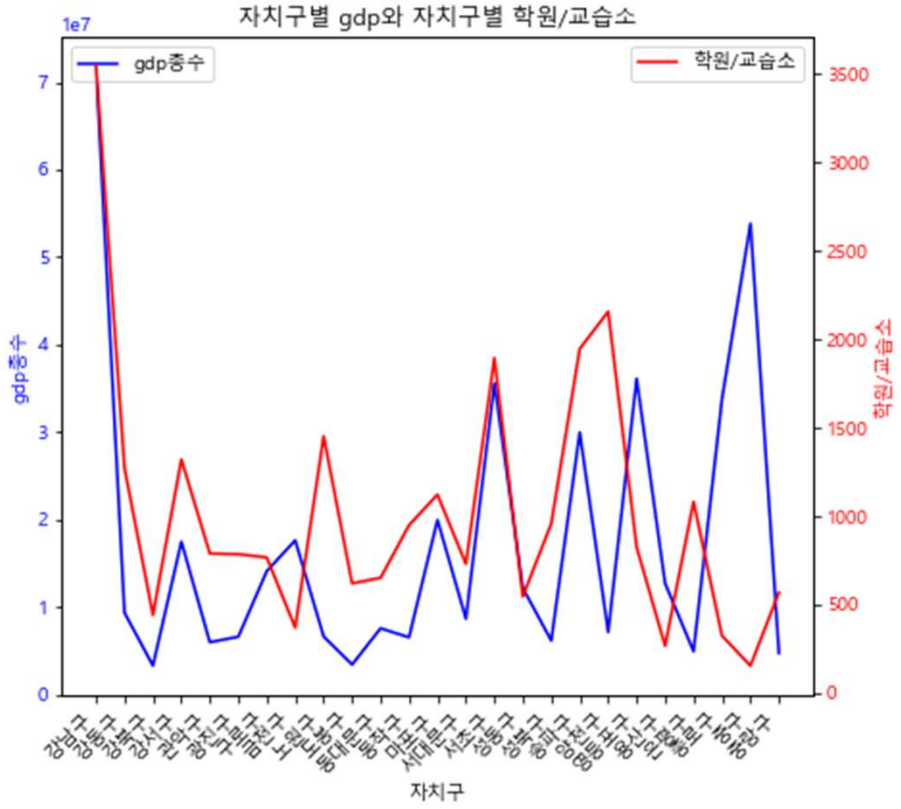


1) 자치구별 grdp 총 수와 자치구별 학교수의 상관관계

1. grdp 데이터에서 자치구별로 총가구수의 개수(sum) 구해서 새로운 데이터프레임 생성
2. school 데이터에서 자치구별 학교 개수 구해서 새로운 데이터 프레임 생성
3. 둘의 데이터를 자치구 기준으로 concat 해준 뒤, 새로운 데이터 프레임 생성

분석 과정

주제 3 – grdp가 높은 자치구 학구열 높을까?



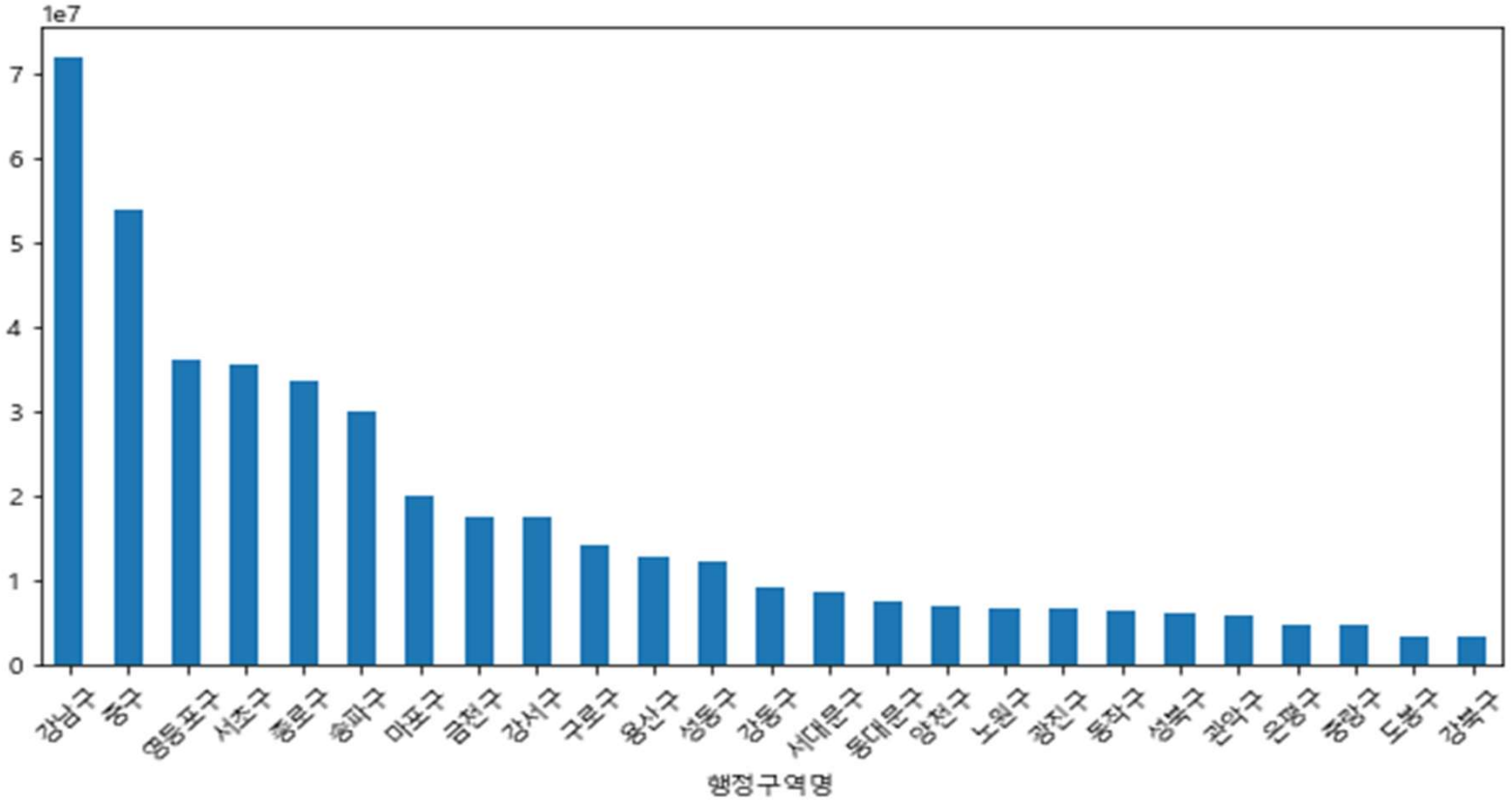
2) 자치구별 grdp 총수와 자치구별 학원 및 교습소의 상관관계

1. grdp 데이터에서 자치구별로 총가구수의 개수(sum) 구해서 새로운 데이터프레임 생성
2. academy 데이터에서 자치구별 학원/교습소 개수 구해서 새로운 데이터프레임 생성
3. 둘의 데이터를 자치구 기준으로 concat 해준 뒤, 새로운 데이터프레임 생성

분석 과정

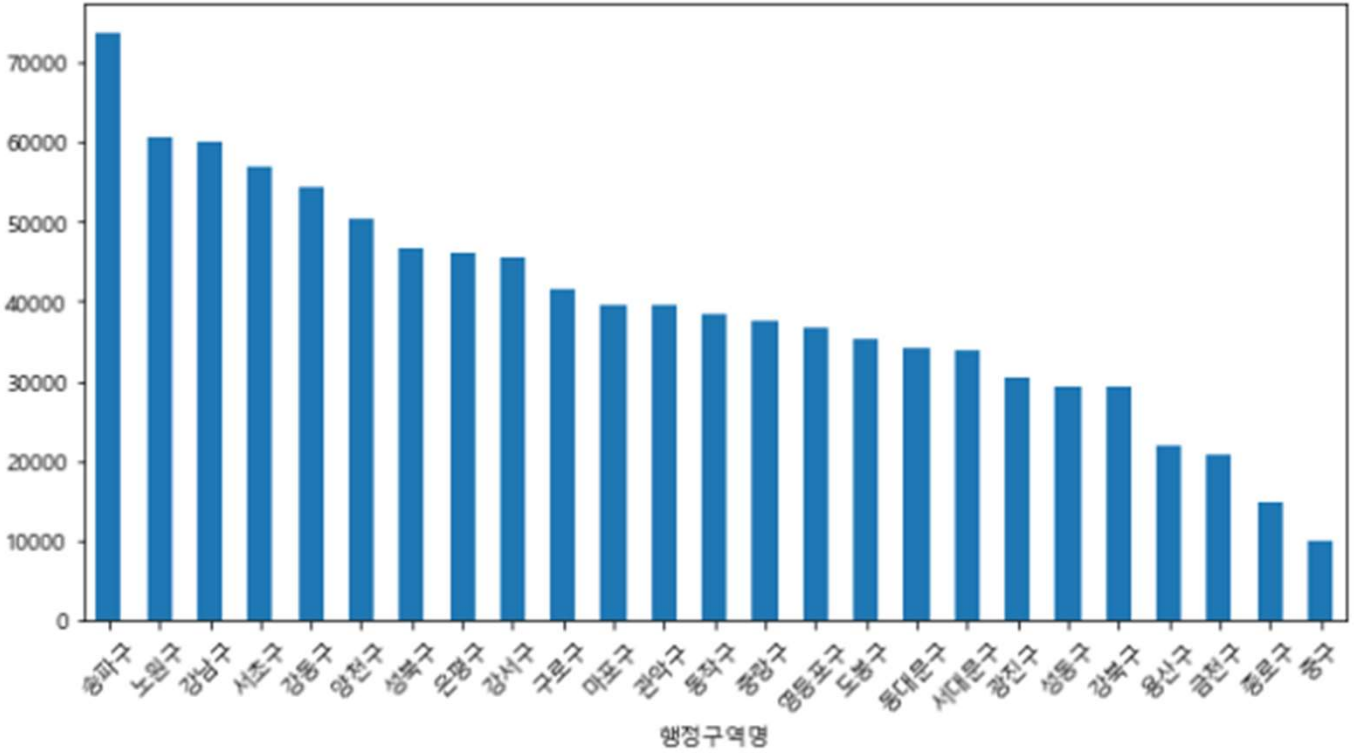
주제 4- 서울 외곽 지역에 학원 수가 많은 이유는?

1) 자치구별 grdp 지수 분포



분석 과정

주제 4- 서울 외곽 지역에 학원 수가 많은 이유는?

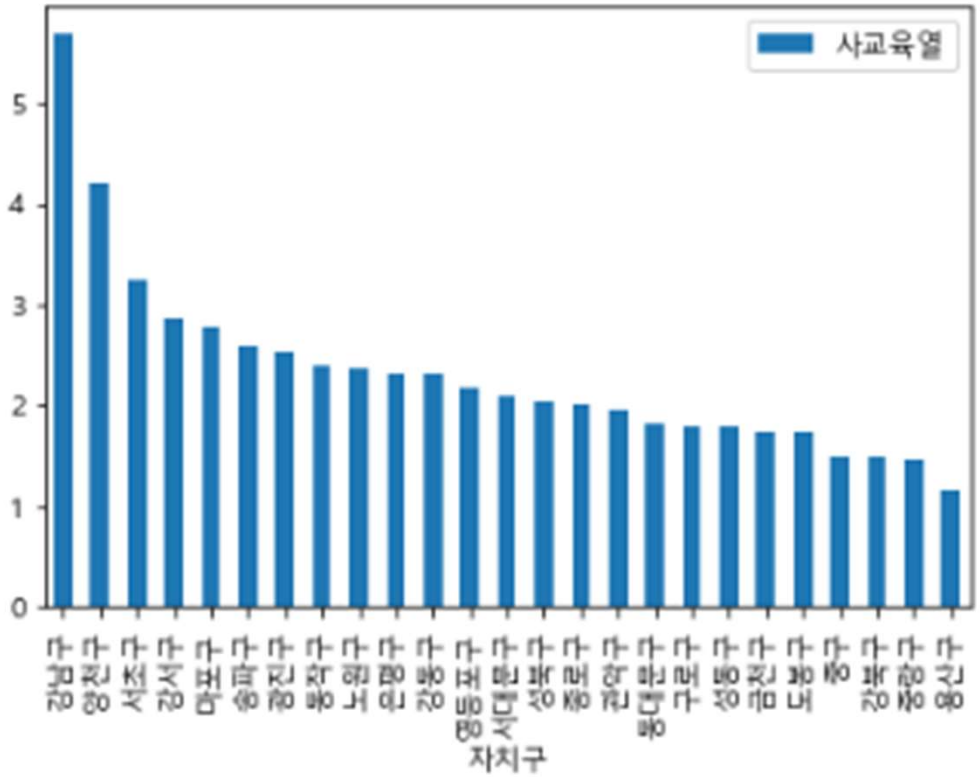


2) 자치구별 학령인구(5세~19세) 분포도

1. 주민등록연앙인구 데이터에서 학령인구의 열만 추출
2. 자치구별 학령인구 총 계 산출

분석 과정

주제 5 - 서울특별시의 교육사각지대와 서울런 사업



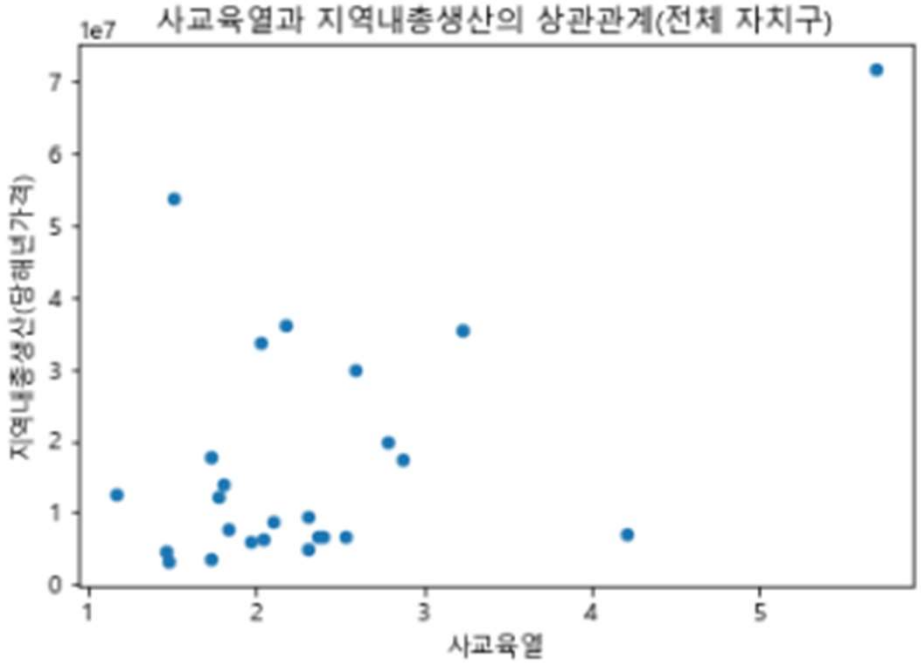
1) 사교육열 시각화

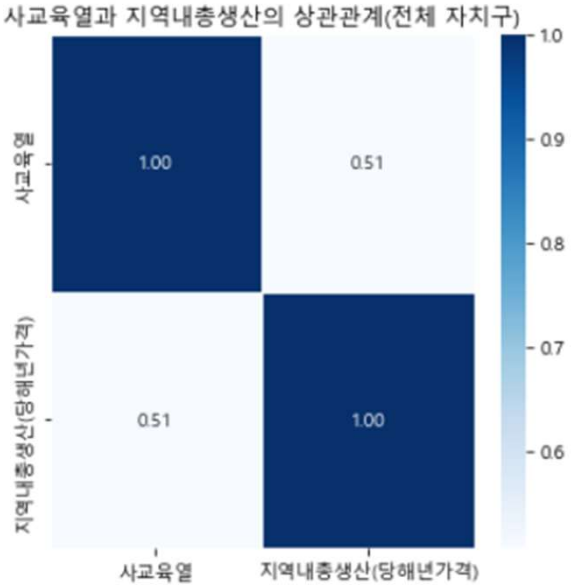
1. 서울특별시 학원 교습소 정보data에서 자치구별로 학원의 개수를 구한다
2. 주민등록연양인구(연령별동별)통계data에서 5~9세, 10~14세, 15~19세 column을 합쳐서 학령인구 column 생성
3. 새롭게 만든 학령인구 column을 자치구별로 나눠서 구한다.
4. 자치구별로 학원의 개수/학령인구*100을 계산하고 이를 사교육열로 정의한다.

분석 과정

주제 5 - 서울특별시의 교육사각지대와 서울런 사업

2) 사교육열과 지역내총생산의 상관관계(전체 자치구)

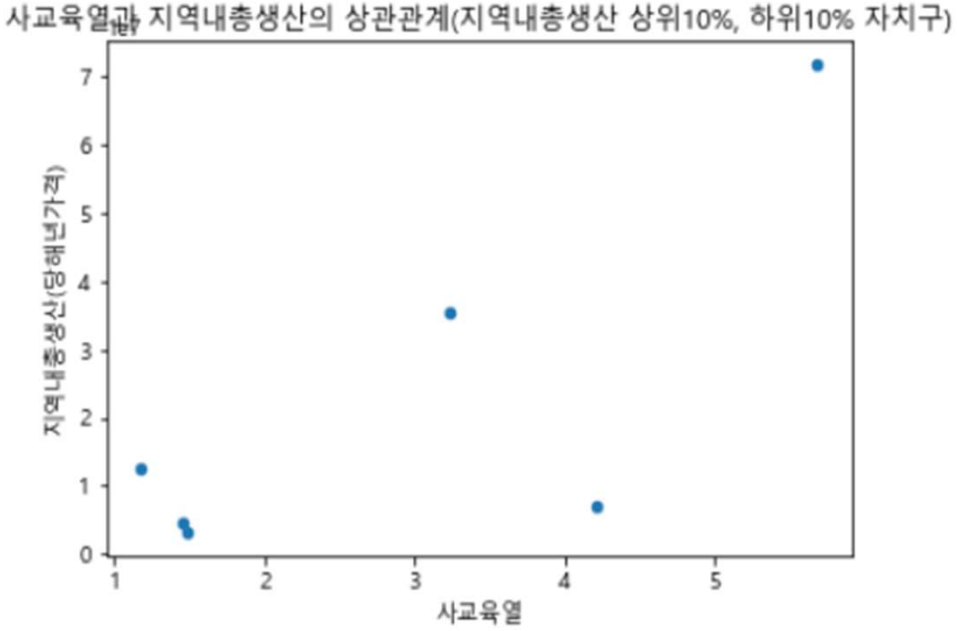


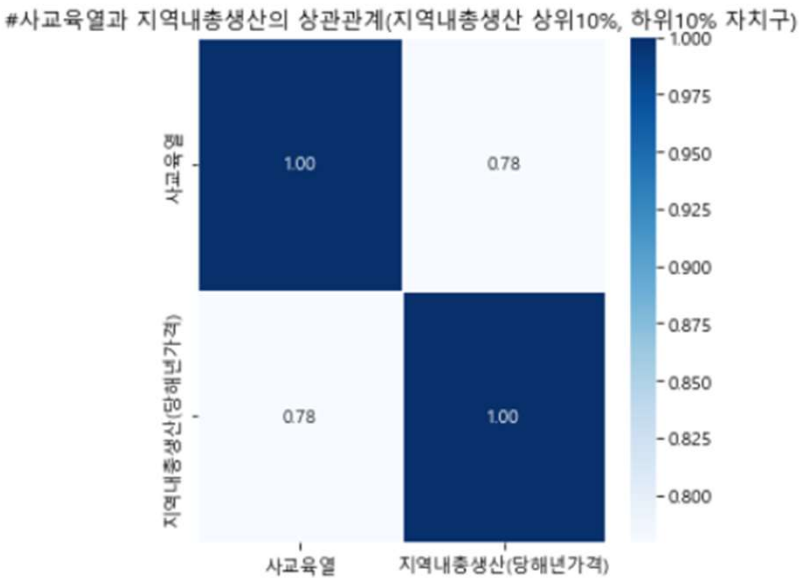


분석 과정

주제 5 - 서울특별시의 교육사각지대와 서울런 사업

3) 사교육열과 지역내총생산의 상관관계 (grdp 상위, 하위 10% 자치구)





결론

주제 1 - 자치구별 학교 다니는 학생수 비율, 자치구별 학원 다니는 학생수 비율

1) 유치원교육열

-높음: 노원구, 종로구, 성북구 / -낮음: 서초구, 강북구, 성동구

2) 초등교육열

-높음: 강서구, 양천구, 광진구 / -낮음: 용산구, 성동구, 마포구

3) 중등교육열

-높음: 양천구, 강남구, 노원구 / -낮음: 용산구, 성동구, 영등포구

4) 고등교육열

-높음: 중구, 종로구, 노원구 / -낮음: 성동구, 서대문구, 마포구

→ 교육열이 낮은 자치구는 **전 연령에서 동일하게 낮은 교육열**을 보이지만
교육열이 높은 자치구는 **연령에 따라서 바뀌는 것**을 알 수 있음.

결론

주제 2 – 강남 8학군은 실제로 학구열이 높을까?

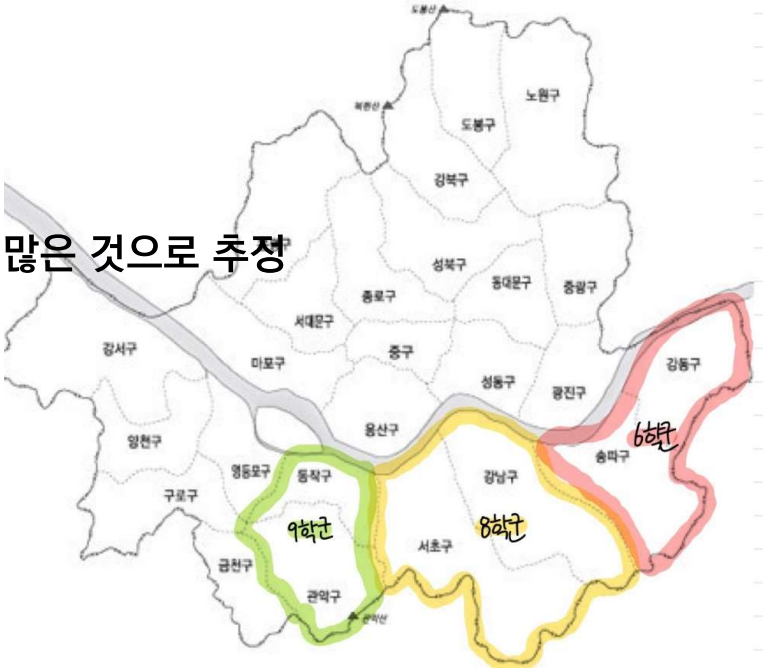
1) 자치구별 학원/교습소 수 비교

-학원/교습소 수 상위 5개 자치구 : 강남구, 양천구, 송파구, 서초구, 노원구, 강서구

→ 학원/교습소는 **6,8,9학군에 집중**되어 있음

-학원/교습소 수는 서울 외곽에 다수 위치

→ 인천과 붙어있는 강서구, 양천구에 **인천 소재 학생 대상 학원이 많은 것으로 추정**



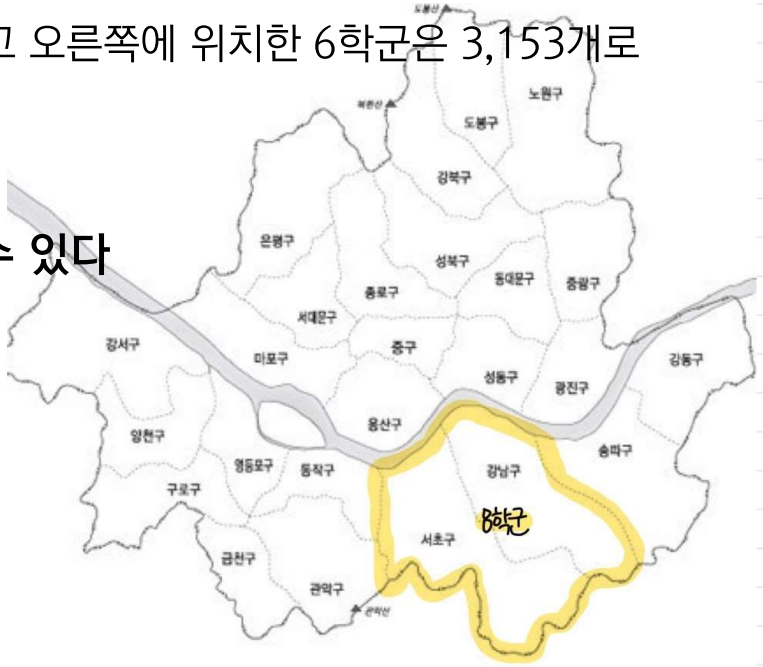
결론

주제 2 – 강남 8학군은 실제로 학구열이 높을까?

2) 학군별 학원/교습소 수 & 학교 수 비교

- 7~9학군을 제외한 나머지 학군의 학원/교습소 수와 학교 수 비율 비슷함
- 지리적으로 8학군 왼쪽에 위치한 9학군의 학원/교습소 수는 1,701개이고 오른쪽에 위치한 6학군은 3,153개로 8학군 5,258개와 큰 차이를 보이고 있다.

→ 다른 학군 자치구에서 8학군 소재 학원에 다니고 있음을 추측할 수 있다



결론

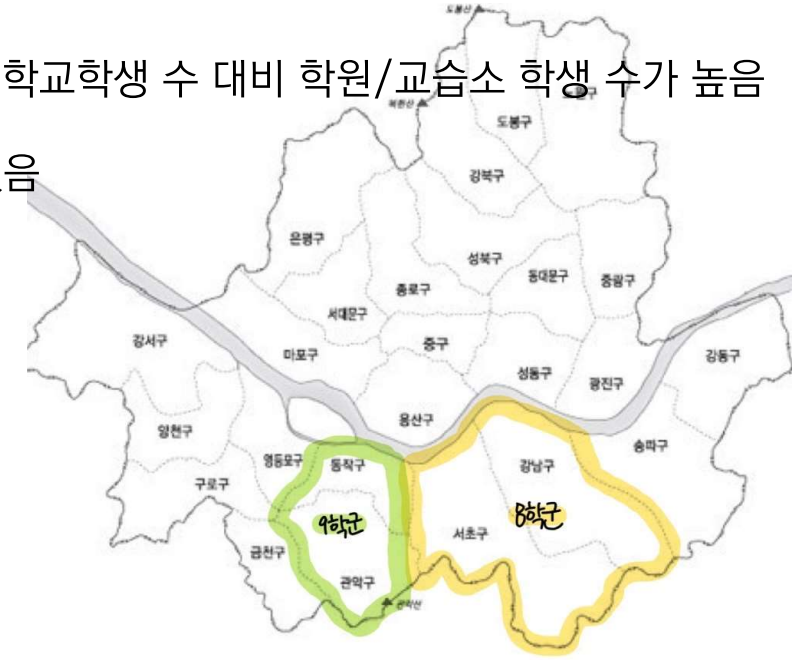
주제 2 – 강남 8학군은 실제로 학구열이 높을까?

3) 학군별 학원/교습소 학생 수 & 학교 학생 수 비교

-모든 학군에서 학교 학생 수보다 **학원/교습소 학생 수가 더 많음**

-특히 **8, 9학군의 학원 학생 수가** 다른 학군에 비해 압도적으로 **많음**. 또한 학교학생 수 대비 학원/교습소 학생 수가 높음

-타 학군 소재 학교 학생들이 8, 9 학군 소재 학원에 등원함을 유추할 수 있음



결론

주제 3 – grdp가 높은 자치구 학구열 높을까?

- 자치구별 grdp가 높은 곳일수록 학교, 학원 및 교습소 수도 많음

  → 강남구/양천구/서초구

- grdp가 높을수록 학구열이 높은 경향이 있음
- 학구열이 높은 7,8학군의 강남구, 서초구, 양천구는 **입시, 검정 및 보습학원의 비율 ↑**

  중랑구는 예체능 학원의 비율 ↑

주제 4 – 서울 외곽 지역에 학원 수가 많은 이유는?

- 학령인구 수 상위 10개의 자치구 중 9군데가 외곽에 위치해 있으므로 외곽지역에 학령인구가 많음
- 외곽 지역의 grdp가 다른 지역의 grdp보다 상대적으로 높음

→ 즉, 소득수준이 높으면 학원에 많이 다니므로 **소득이 높은 외곽지역에 학원 수가 많음**

결론

주제 5 - 서울특별시의 교육사각지대와 서울런 사업

-서울런이란?

: 학습자원 접근이 어려운 청소년들을 대상으로 하는 교육 서비스 사업.

초중고 학생이 가장 선호하는 온라인 교육 콘텐츠를 종합반 수준으로 무제한 수강할 수 있도록 제공하고,

가입자 전원에게 멘토링 서비스를 제공하며 효율적인 학습 관리를 지원하는 서비스

서울런
SEOUL LEARN

결론

주제 5 - 서울특별시의 교육사각지대와 서울런 사업

-자치구 전체를 살펴봤을 때 사교육열과 지역내총생산(grdp)의 상관관계가 약 50% 정도로 낮은 상관관계를 보임

-서울 지역내총생산 상위 10% 자치구와 하위 10% 자치구만 따로 뽑아서 상관관계를 구했더니 78%라는 값 도출

→ 지역내총생산이 평균인 지역들은 지역내총생산과 사교육열 간의 상관관계가 크지 않았지만

지역내총생산이 매우 높은 지역일수록 사교육열이 더욱 높았으며,

지역내총생산이 매우 낮은 지역일수록 사교육열이 줄어드는 것을 확인함

결론

주제 5 - 서울특별시의 교육사각지대와 서울런 사업

지역내총생산이 가장 적은 강북구, 중랑구, 용산구는 교육사각지대로 볼 수 있으며 정부의 교육 지원이 필요하다고 판단

서울런의 가입자수는 계속 늘어날 전망이며 이를 통해 교육격차 해소를 기대해 볼 수 있다

[서울런 가입자 추이]