

| **데이터 분석 모듈프로젝트** |
| --- |

**2023년 1월 18일**

**주제명 : 서울시 구별 아파트 매매가격과 교육기관의 상관관계 분석**

**교육명 : 클라우드 AI 융합 전문가 과정**

**팀번호 : 4조**

**팀원명**  **:** **김태혁, 박상욱**

| **목 차**  1. 프로젝트 주제  2. 기획의도 및 목표  3. 데이터  4. 데이터 전처리, 정제  5. 데이터 분석  6. 결론 도출 |
| --- |

1. **프로젝트 주제**

* 서울시 구별 아파트 매매가격과 교육기관의 상관관계 분석

1. **기획의도 및 목표**

* 서울시 집값과 교육기관의 관계 분석
* 학원과 서울시 집값 데이터와의 관계 분석
* 학교와 서울시 집값 데이터와의 관계 분석
* 서울시 평형별 아파트 매매가격과 교육기관의 관계 분석
* 서울시 아파트 매매가격과 교육기관의 가장 상관관계가 깊은 관계 분석

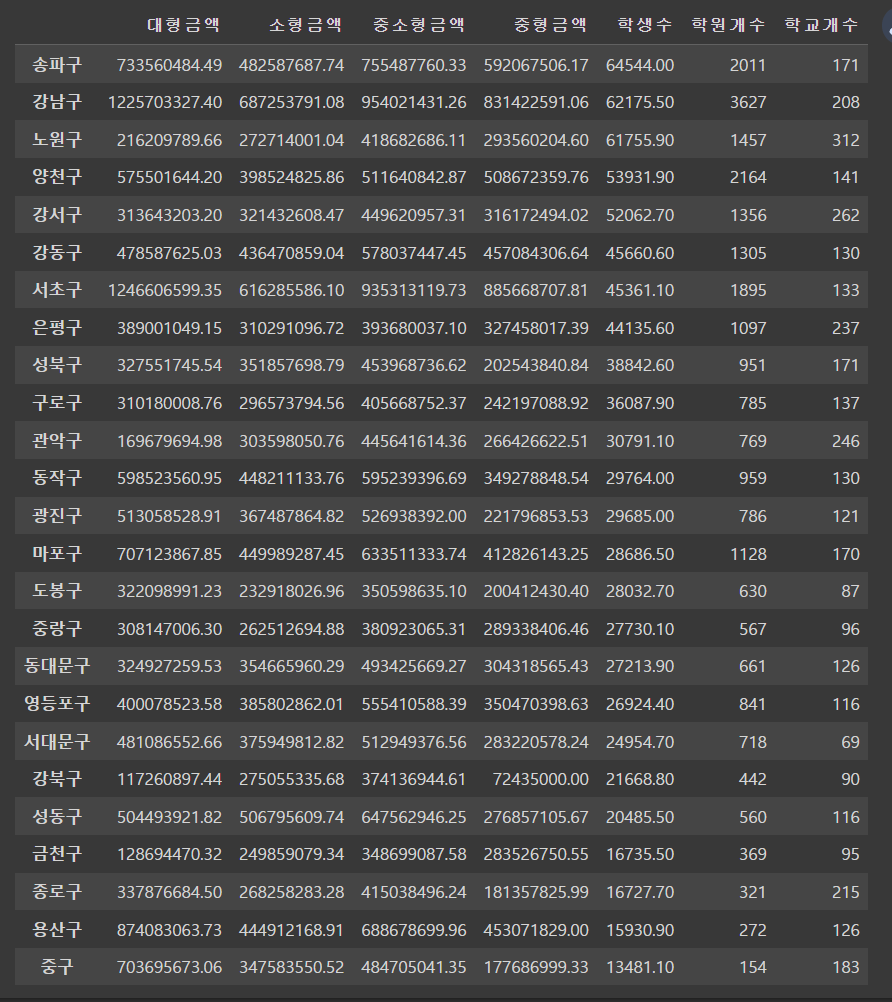
1. **데이터**

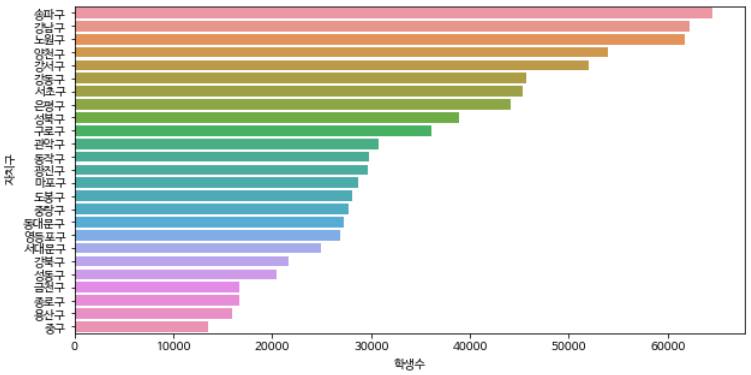
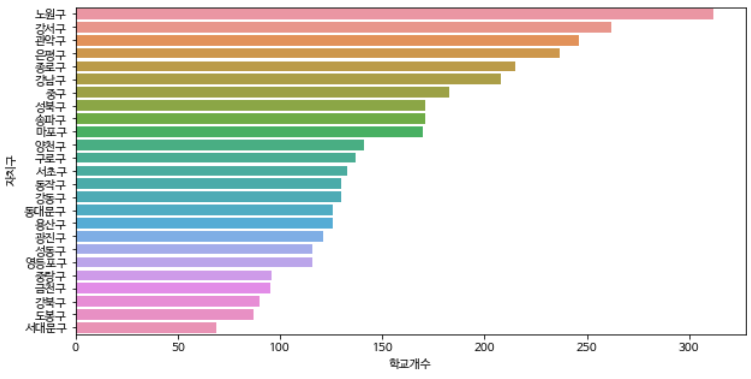
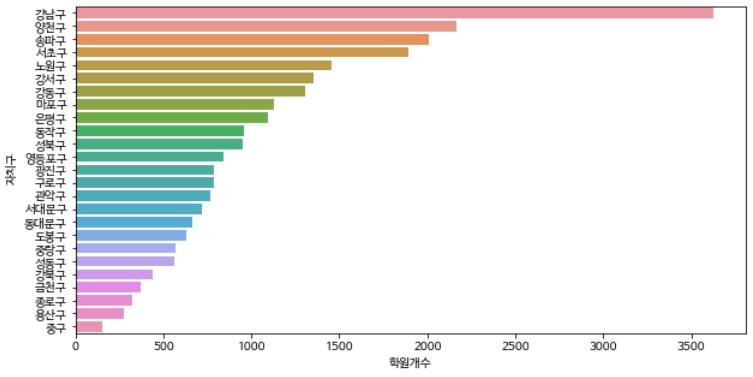
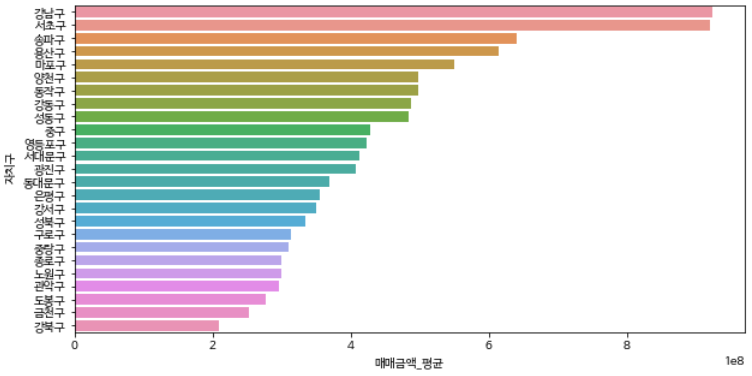
* 서울시 구별 2018 ~ 2020 아파트 매매 현황
* 서울시 구별 학교 데이터
* 서울시 구별 학원 데이터
* 서울시 구별 학급당 학생수

**4. 전처리, 정제**

**4-1) 기본데이터 정제**

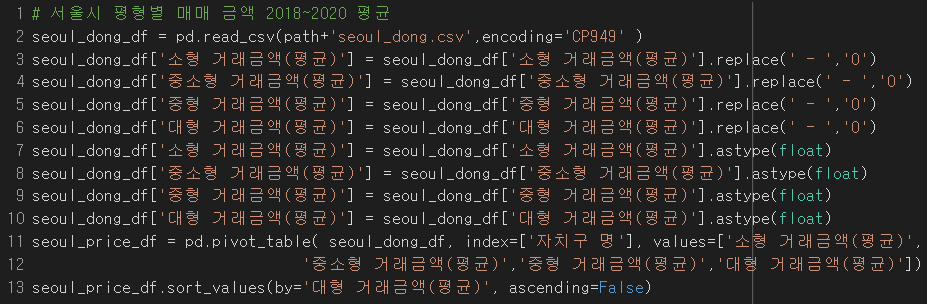
* 서울시 구별 학생수, 학원개수, 학교개수, 2018~2020 아파트 평형별 금액





**4-1-1 서울시 구별 매매금액**

* 서울시 구별 2018 ~ 2020 아파트 매매 소형~대형의 평균거래금액

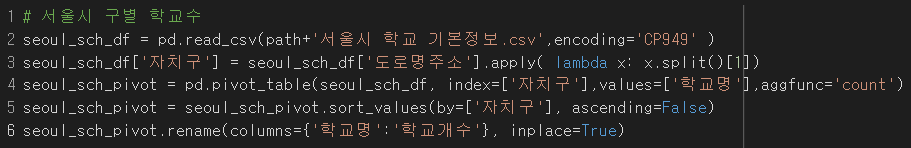


* 평형별 아파트 매매 거래금액 평균을 3개년 평균값으로 바꾸기 위해 ‘-’을 0으로 수정

object 형식인 평형별 거래금액 평균을 float으로 수정

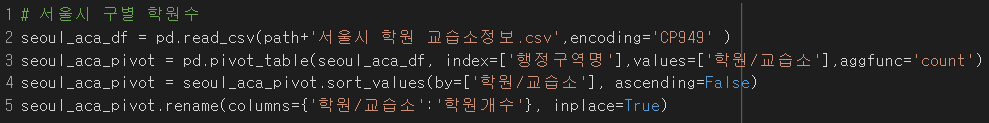
* pivot\_table로 index:자치구 values:평형별 거래금액평균 으로 자동 평균값계산

**4-1-2 서울시 구별 학교수**



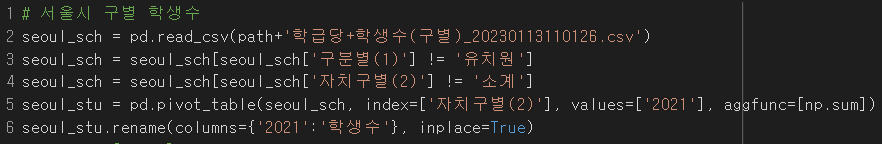
* 서울시 학교정보를 pivot\_table로 index:자치구 values:학교명 으로 count

**4-1-3 서울시 구별 학원수**



* 서울시 학원정보를 pivot\_table로 index:행정구역명 values:학원/교습소 로 count

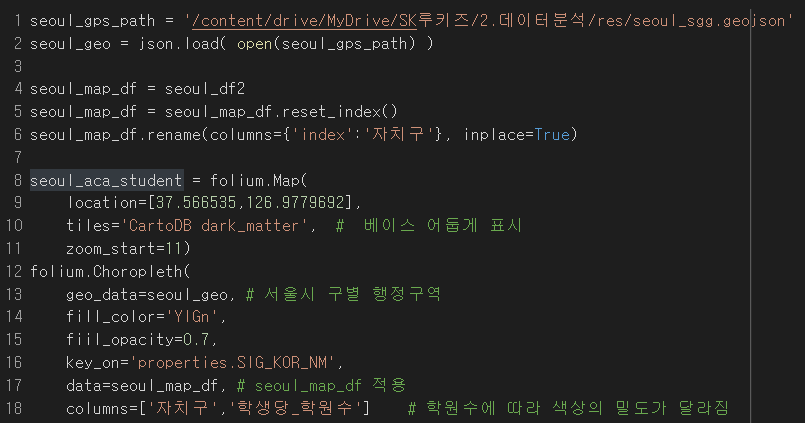
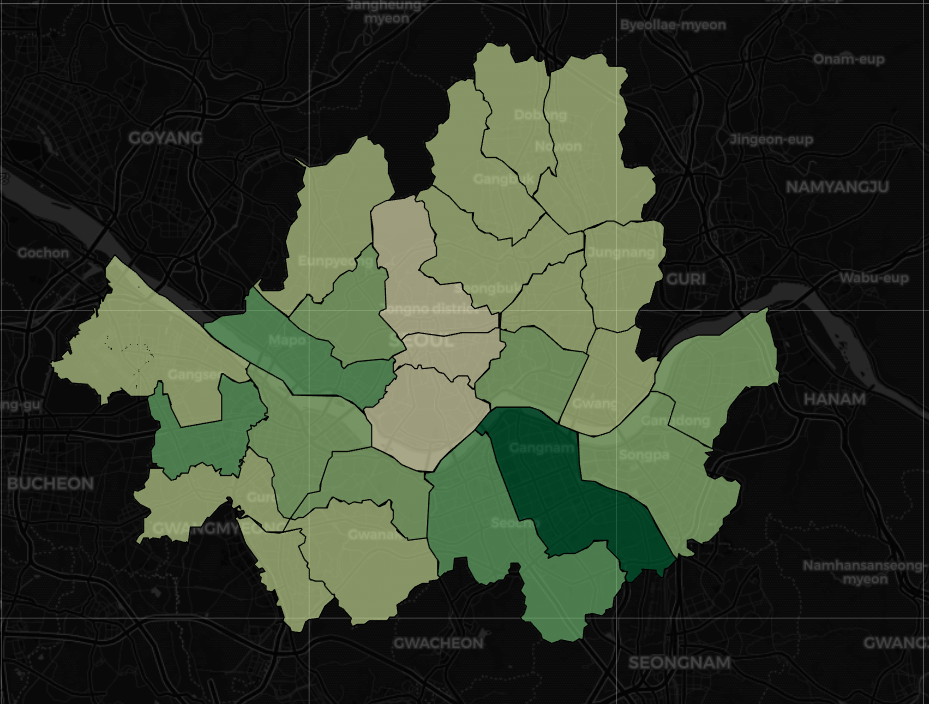
**4-1-4서울시 구별 학생수**



* 서울시 학생수 정보를 pivot\_table로 index:자치구 values:소계 로 np.sum

(단, 학급구분 != 유치원)

**4-2) 서울시 구별 학생당 학원수를 지도로 시각화**

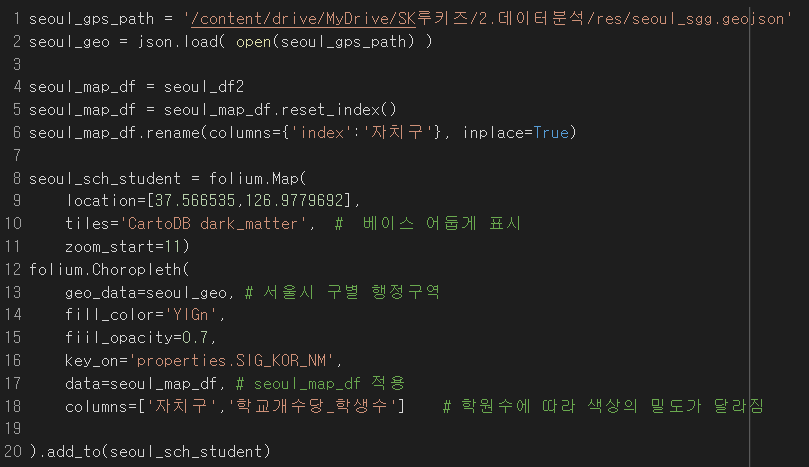
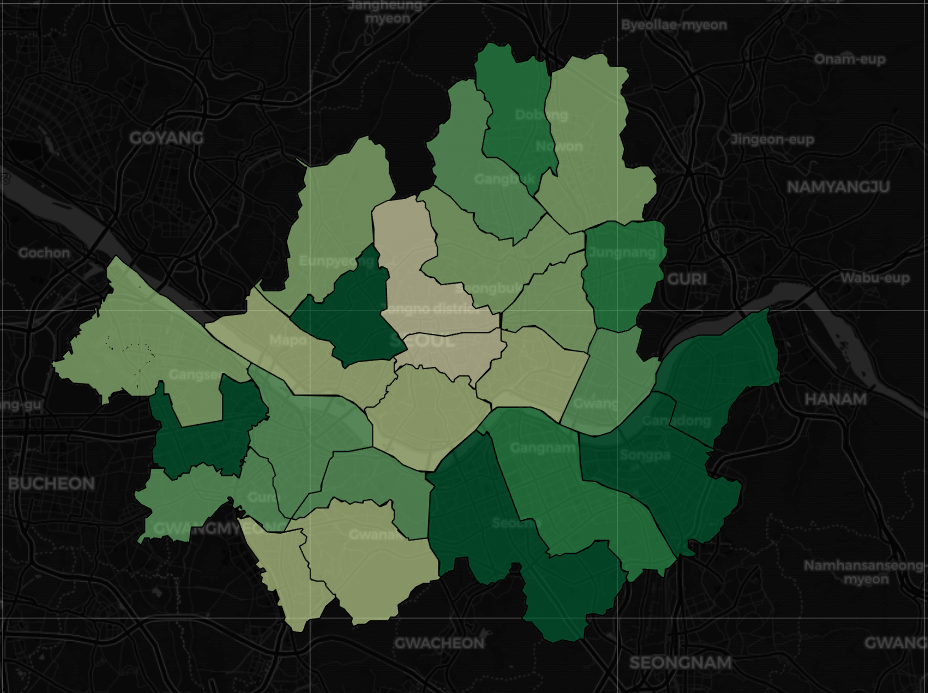


* 서울시 구별 학생당 학원수 시각화 자료로 학원의 실수요자인 **학생이 등교하는 지역에 비해 학원이 많은 지역**을 의미한다.

1~5위 강남구, 서초구, 양천구, 마포구, 동작구 순이다.

* 코드설명 : folium 모듈을 사용해 지도로 시각화 하였으며, seoul\_sgg.geojson과 학생당 학원수를 자치구로 묶어 색상의 밀도로 구분한 지도이다

**4-3) 서울시 구별 학교당 학생수를 지도로 시각화**



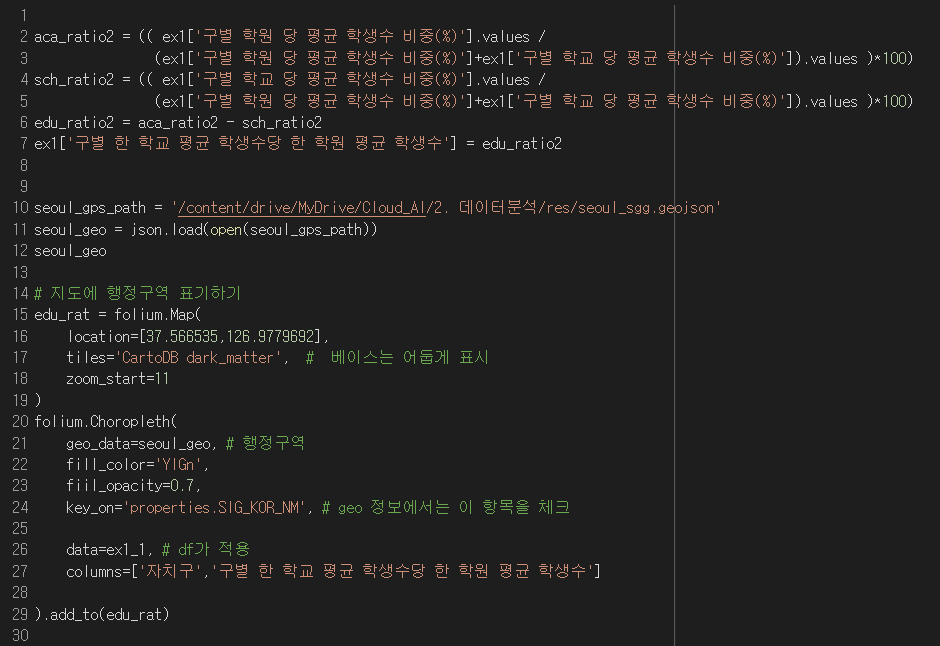
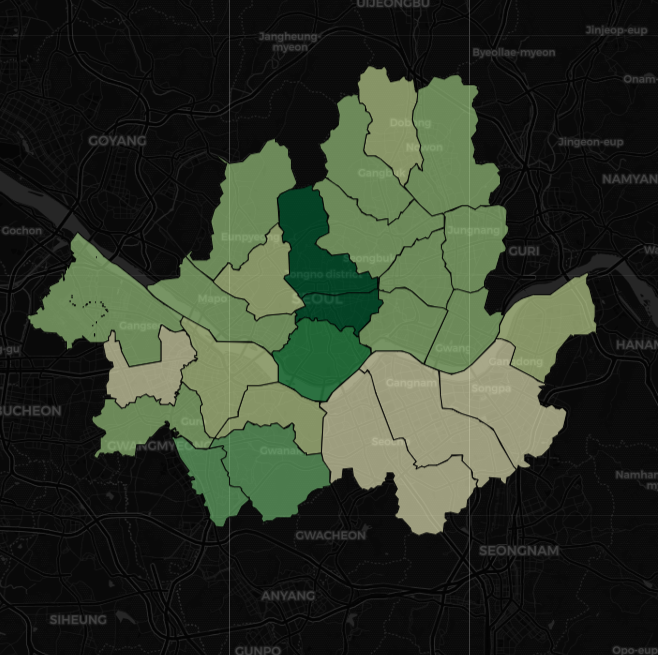
* 서울시 구별 학교당 학생수로 **학교수에 비해 학생이 많은 지역**을 의미하며,

**시각적으로 강남 3구가 상반되는 데이터에서 같은 밀도를 보여 데이터간 관계가 없을 것으로 예상된다.**

1~5위 양천구, 송파구, 서대문구, 강동구, 서초구 순이다.

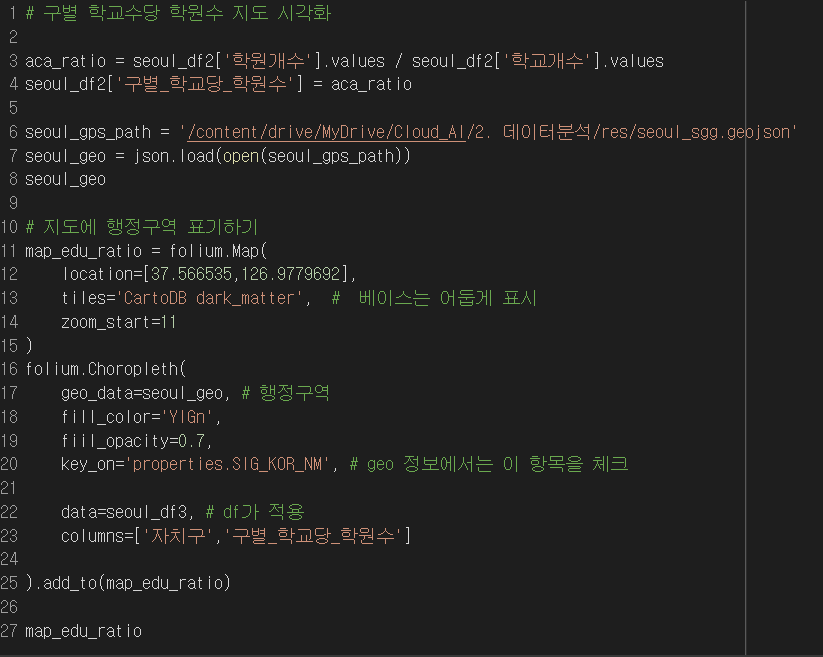
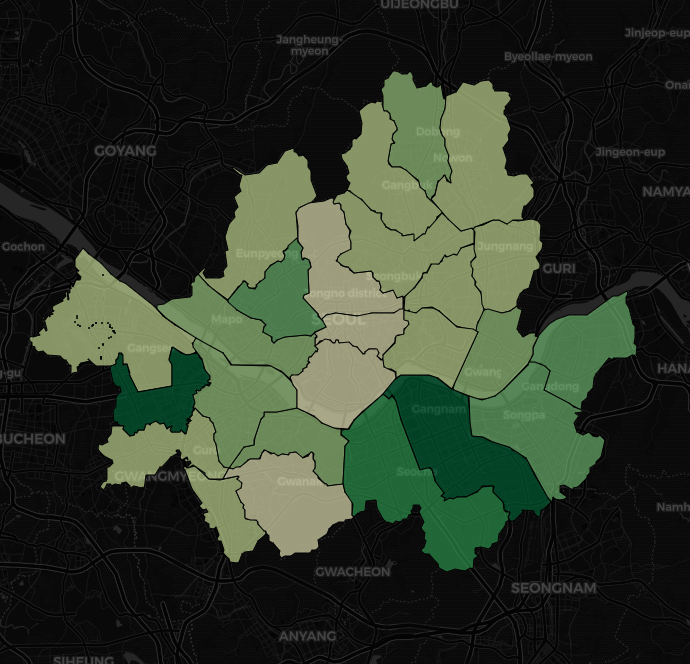
* 코드설명 : folium 모듈을 사용해 지도로 시각화 하였으며, seoul\_sgg.geojson과 학교당 학생수를 자치구로 묶어 색상의 밀도로 구분한 지도이다

**4-4) 서울시 구별 학생수기준 학원 비중과 학교 비중 차이**



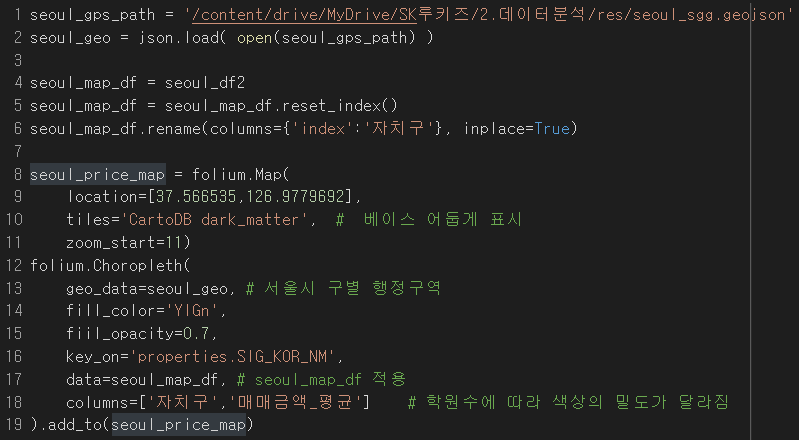
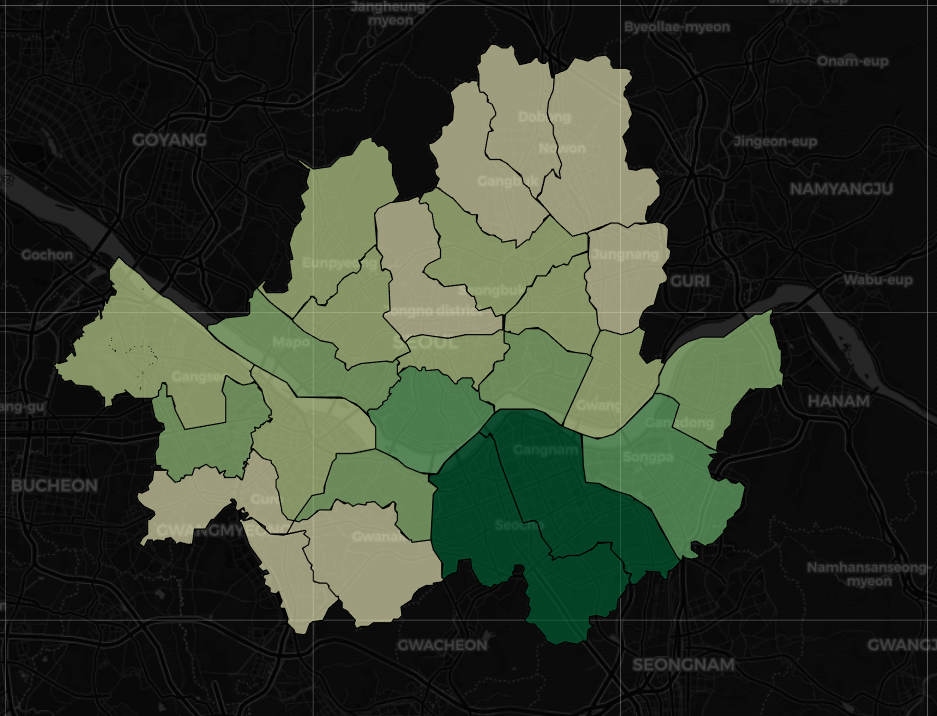
* 학생수기준 (학원 비중 - 학교 비중) 이 음수 값으로 나오는 구 는 학교에 다니는 학생수 비중이, 학원에 다니는 비중보다 크다는 것.
* 즉 학원에 다니는 학생수 비중이 낮다는것은 그 구에 학원이 많다고 해석.
* 지도에서 색이 옅을수록 학원에 다니는 학생수 비중이 낮은 것.
* 학원에 다니는 학생수 비중이 작은순으로 강남구,서초구,마포구.. 순이다.
* 이 순서대로 매매가격이 높을 것이라 예상.
* **코드설명**: folium 모듈을 사용해 지도로 시각화 하였으며, 구별 학생수 기준 학원비중과 학교비중의 차이를 자치구로 묶어 색상의 밀도로 구분

**4-5) 서울시 구별 학교수당 학원수**



* 학교수당 학원수 값이 양수 값으로 나오는 구는 학원수가 학교수 보다 많다는 것이다.
* 즉, 지도에서 색이 진할수록 다른 구 대비 학교수당 학원수가 높다.
* 이 값이 클수록 매매가격이 높을 것이라 예상.
* **코드설명**: folium 모듈을 사용해 지도로 시각화 하였으며, 구별 학교수당 학원수 비중을 자치구로 묶어 색상의 밀도로 구분

**4-6) 서울시 구별 아파트 매매 거래금액 평균**



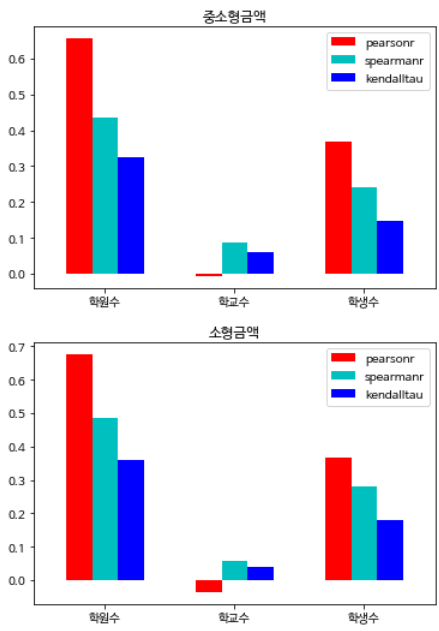
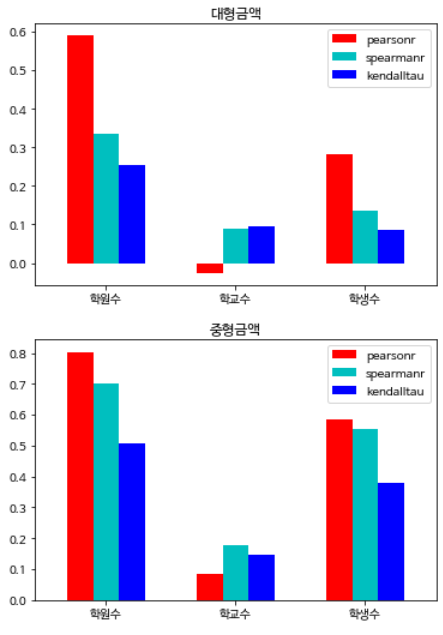
* 서울시 구별 평균 아파트 매매 거래금액을 추출 하였으며,

1~5위 강남구, 서초구, 송파구, 용산구, 마포구 순이다

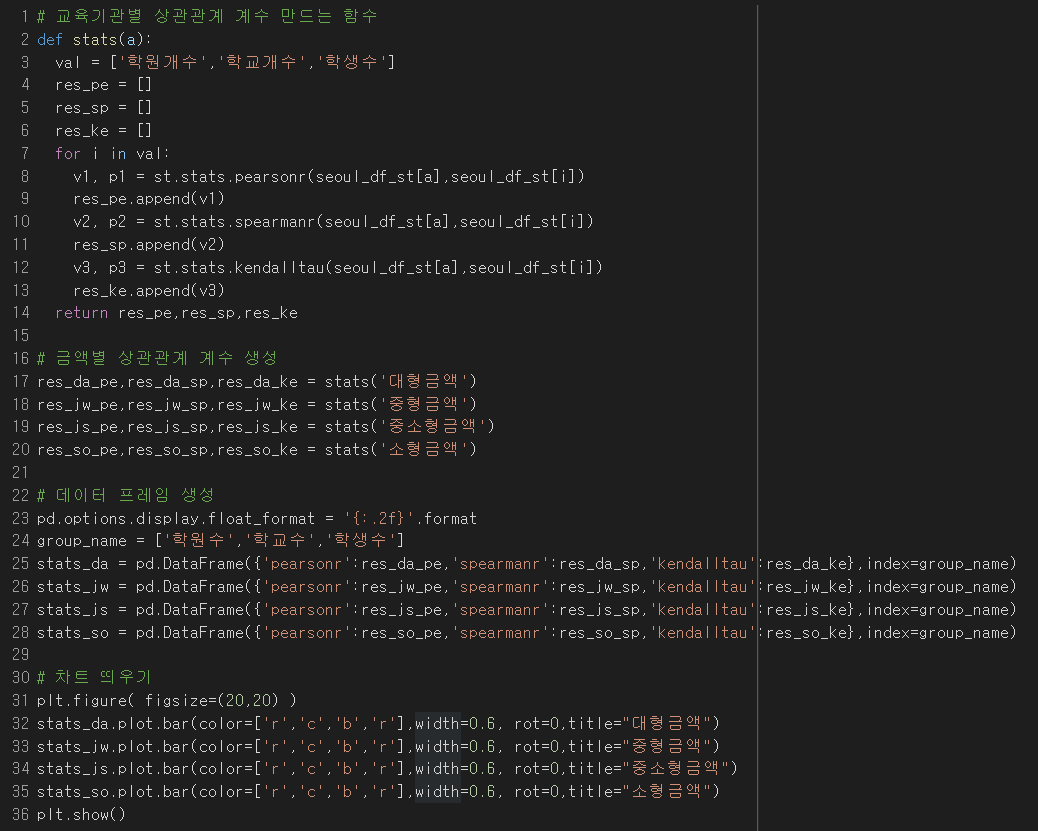
* 코드설명 : folium 모듈을 사용해 지도로 시각화 하였으며, seoul\_sgg.geojson과 학교당 학생수를 자치구로 묶어 색상의 밀도로 구분한 지도이다.

**5. 데이터 분석**

**5-1) 서울시 아파트 평형별 매매가격과 서울시 구별 학원수, 학교수, 학생수와의 상관계수 그래프**



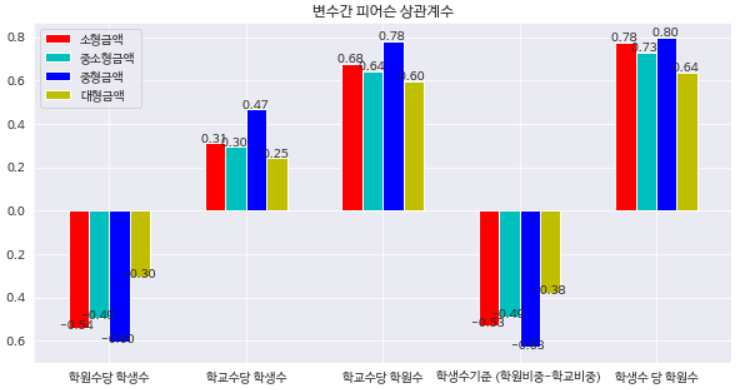
* 위의 사진은 서울시 아파트 평형별 매매가격과 서울시 구별 학원수, 학교수, 학생수와의 상관계수 그래프이다
* 4개의 평형별금액 그래프에서 학원수의 상관계수가 높았으므로 기본데이터 컬럼인 학원수, 학교수, 학생수 에서 서울시 아파트 매매가격과 **가장 관계가 있는 데이터**는 **학원수** 임을 알수있다.
* 학교수는 금액별로 전부 관련이 많이 없으며, 학생수는 0.4~0.5 관련이 있으며 중형에서 0.6으로 가장 높은 수치를 보였다.
* 4개 평형별금액 그래프에서 금액별로 학원수가 가장 높은 상관관계를 가졌으며 학교수는 가장 낮은 상관관계를 보이고 학생수는 대부분 학원수의 상관계수의 반절 정도이나 중형금액에서 0.6으로 **학생수는 중형금액에서 가장 높은 상관관계**를 가지고 있음을 볼수있다.

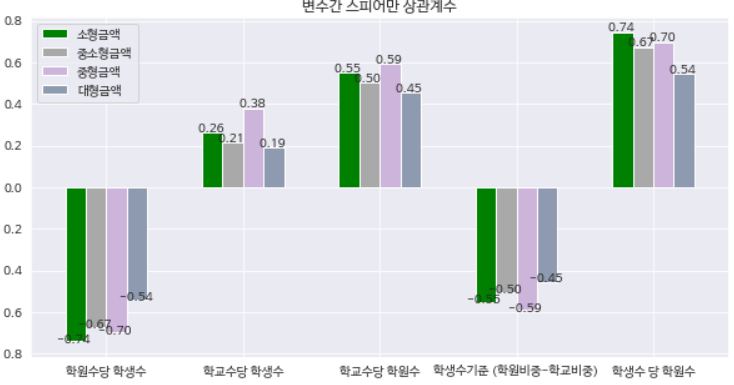


**코드설명**

* stats의 pearsonr, spearmanr, kendalltau로 교육기관별 상관관계 계수 생성 함수 stats() 생성
* stats()를 사용해 평형별 금액과 교육기관별 상관계수 생성
* 평형별 금액과 교육기관별 상관계수를 DataFrame으로 생성
* 생성한 DataFrame을 plot.bar차트를 사용하여 시각화 한다.

**5-2) 서울시 아파트 평형별 매매가격과 서울시 구별 학원수당 학생수, 학교수당 학생 수, 학교수당 학원수, 학생수기준 학원과학교 비중차이, 학생수당 학원수 와의 상관관계 그래프**

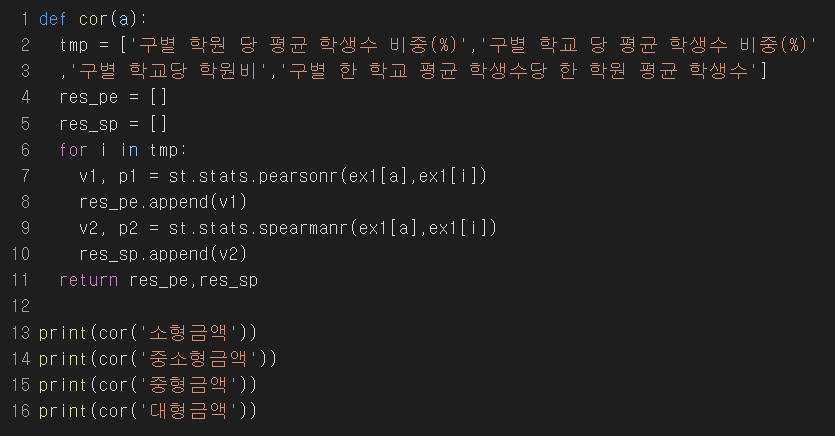


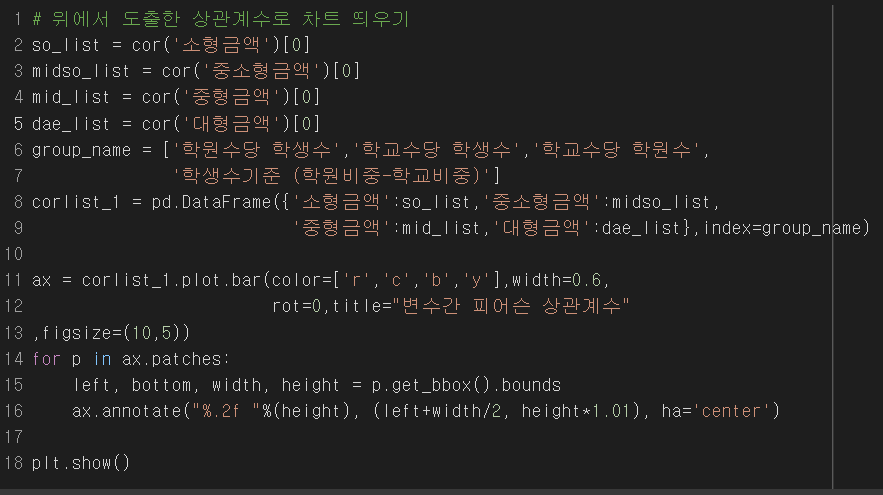


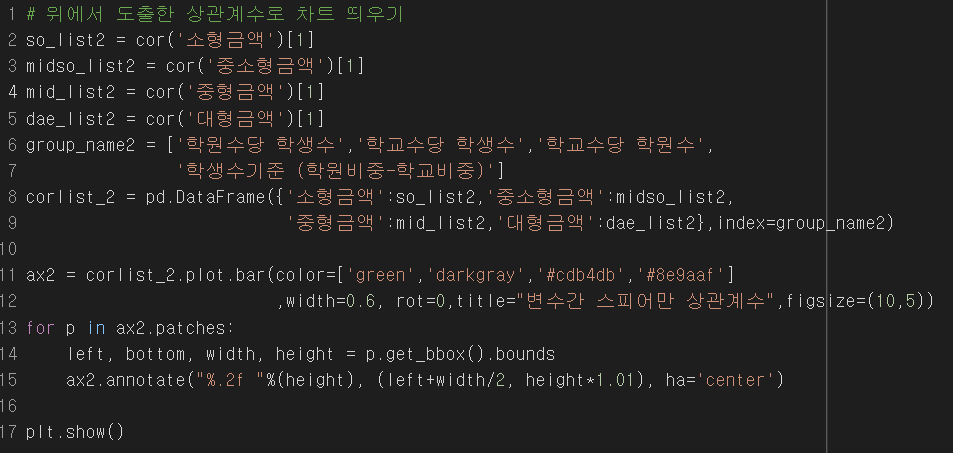
* 위의 그래프는 학원수당 학생수, 학교수당 학생 수, 학교수당 학원수, 학생수기준 (학원비중-학교비중), 학생수당 학원수 변수와 아파트별 평수(소형,중소형,중형,대형) 매매가격과의 피어슨 상관계수와 스피어만 상관계수 값을 나타낸 그래프이다.
* 상관계수가 양의값으로 1과 가까질수록 두 변수간 상관관계가 높은 것이고(양의 기울기), 반대로 음의값으로 1과 가까워질수록 두 변수간 반대의 상관관계를 갖는다(음의 기울기)

**그래프 분석**

* **학원수당 학생수**: 이 변수와 평수별 매매가격과의 관계는 모두 음의 상관계수가 나왔다. 이는 학원에 학생수가 많을수록 매매가격이 떨어졌음을 의미하고, 다시 말해 학원당 학생수가 적을수록 매매가격이 올라감을 알 수 있다. 또한 이변수와 중형매매가격과의 계수값이 가장 큰데, 이것은 다른 평수 매매가격에 비해 중형평수가 학원당 학생수의 영향을 가장 많이 받는다는 것을 의미.
* **학교수당 학생수**: 이 변수와 평수별 매매가격과의 관계는 모두 양의 상관계수이다. 이는 학교당 학생수가 많을수록 매매가격이 올라감을 의미한다. 하지만 다른 변수들의 비해 상관계수 값이 작기 때문에 학교수와는 매매가격과 연관성이 떨어진다는 것을 의미.
* **학교수당 학원수**: 이 변수와 평수별 매매가격과의 관계는 모두 양의 값이 나왔다. 이는 구별 학교수 대비 학원수가 많을수록 매매가격이 올라감을 의미한다. 중형금액과의 상관계수가 가장 높다.
* **학생수기준 학원과 학교 비중 차이**: 이 변수와 평수별 매매가격과의 관계는 모두 음의 값이 나왔다. 이는 학원에 다니는 학생수 비중이 학교에 다니는 학생수 비중보다 작을수록 매매가격이 올라감을 의미.
* **학생수당 학원수:** 이 변수와 평수별 매매가격과의 관계는 모두 양의 값으로, 학생수당 학원수가 많을수록 평수별 매매가격이 올라감을 의미한다. 특이점으로는 이 변수가 다른 변수에 비해 평수별 매매가격과의 상관관계 값이 가장 높은것으로 보여진다. 그중에서도 중형가격과의 상관관계가 가장 높다.







**코드설명**

* cor의 pearsonr, spearmanr 으로 각 변수간 상관관계 계수 생성
* 변수간의 상관계수를 DataFrame으로 생성
* 생성한 DataFrame을 plot.bar차트를 사용하여 시각화 한다.

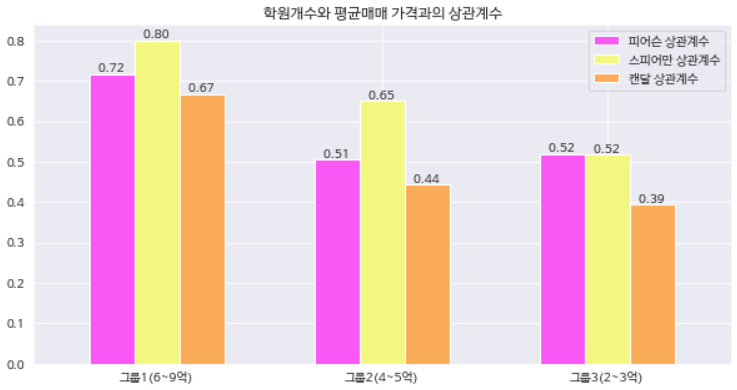
**5-3) 매매 가격별 그룹당 학군개수와 집값과의 상관관계**

* 학원, 학교개수와 매매가격과의 상관관계를 더욱 신뢰도 있게 비교하기 위해 기존의 외부요인들로 인해 구의 가치 가격이 올라가 있는 문제를 해결해야 한다.

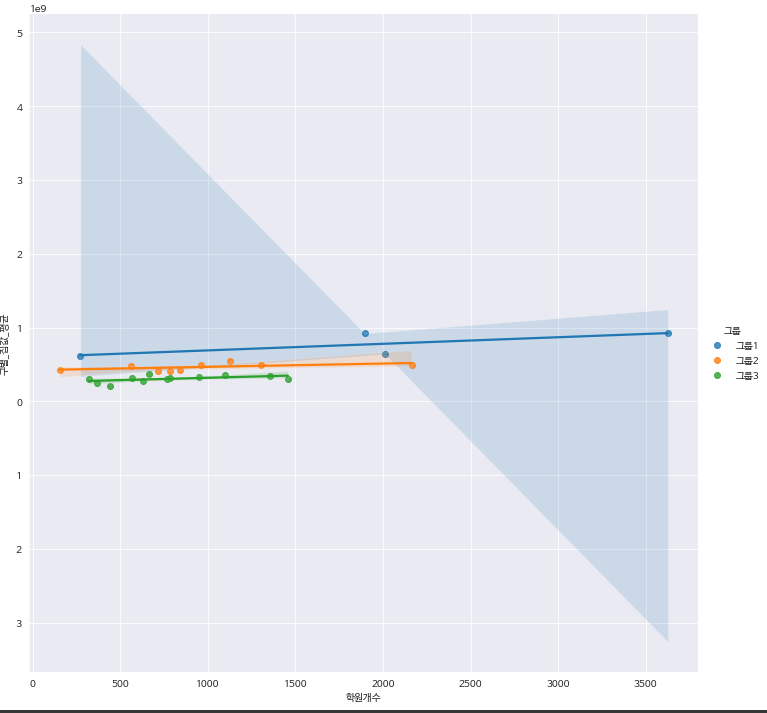
때문에 서울시 구별로 평균 매매가격이 비슷한 구들로 3그룹으로 그룹화를 시켜(편향고려) 학군개수와 평균매매 가격과의 상관관계를 알아본다.

* 평균매매가격별로 그룹1(6~9억), 그룹2(4~5억), 그룹3(2~3억)

**5-3-1 그룹별 학원개수와 평균매매 가격과의 상관계수**

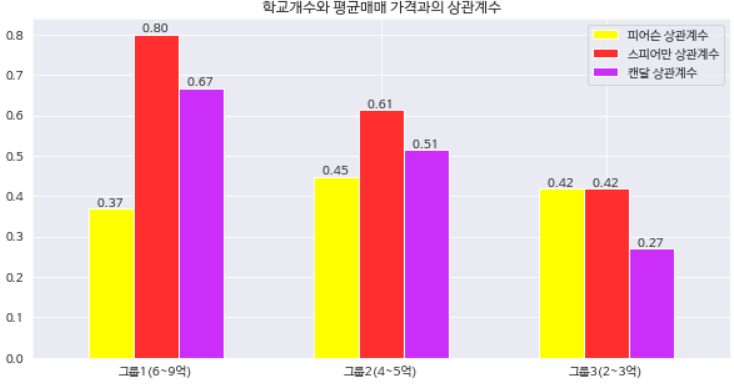


* 위의 그래프는 **그룹별 학원개수와 평균매매가격과의 상관계수**이다.
* 전체적으로 보았을때, 모든 그룹이 학원개수가 많아 질 수록 평균 매매가격이 올라간 것을 볼 수 있다(상관계수값 양수)
* 평균매매가격이 높은 그룹이 낮은 그룹보다 학원개수와 매매가격관의 상관정도가 더 높음을 알 수 있다.

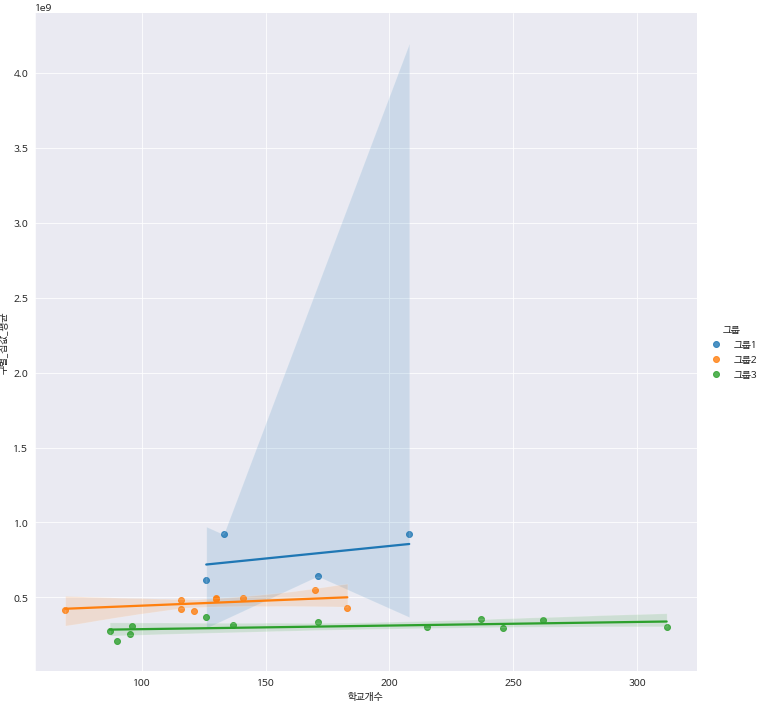


* 위의 그래프는 학원개수와 평균매매가격과의 관계를 Implot차트를 띄워 그룹별로 나타낸 것이다.
* 모두 양의 기울기를 갖음으로, 학원개수가 많을수록 평균 매매가격이 올라감을 알 수 있다.
* 그룹별로 학원개수와 평균금액과의 상관계수값이 높게 측정됨으로, 기울기가 1에 가깝게 기울여 보여야 하지만 그래프의 오차범위 표시로 인해 상대적으로 기울기 정도가 낮아보여서 그룹별로 확연한 차이를 보기 힘들다.
* 오차범위가 넓게 나온 이유는 그룹별로 학원개수의 차이가 크고 샘플들이 적게 존재함으로 오차범위가 넓다(신뢰도가 떨어진다)

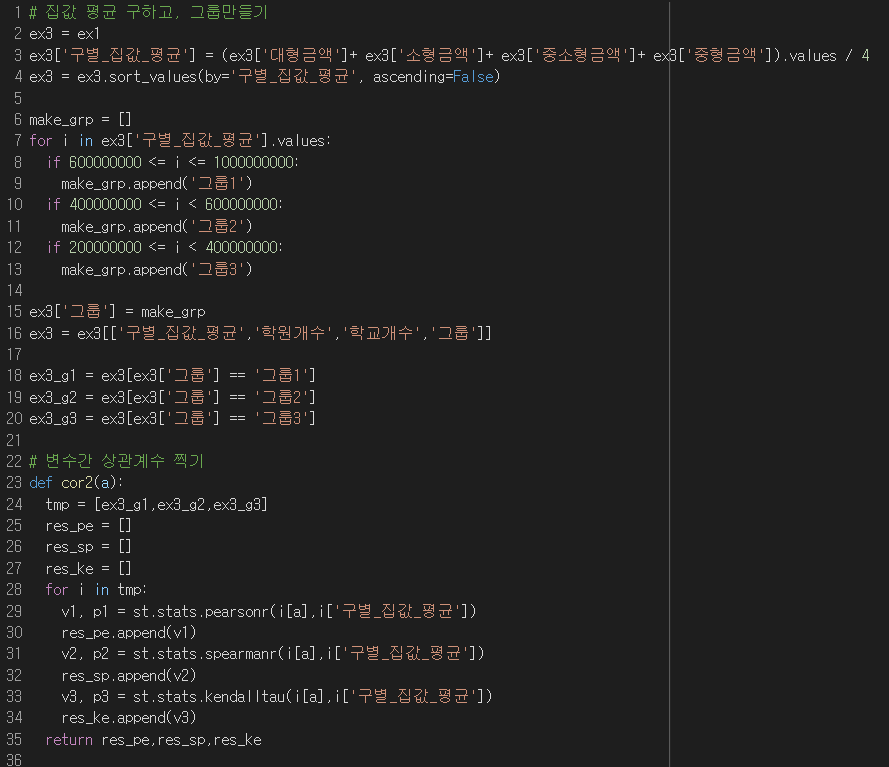
**5-3-2 그룹별 학교개수와 평균매매 가격과의 상관계수**

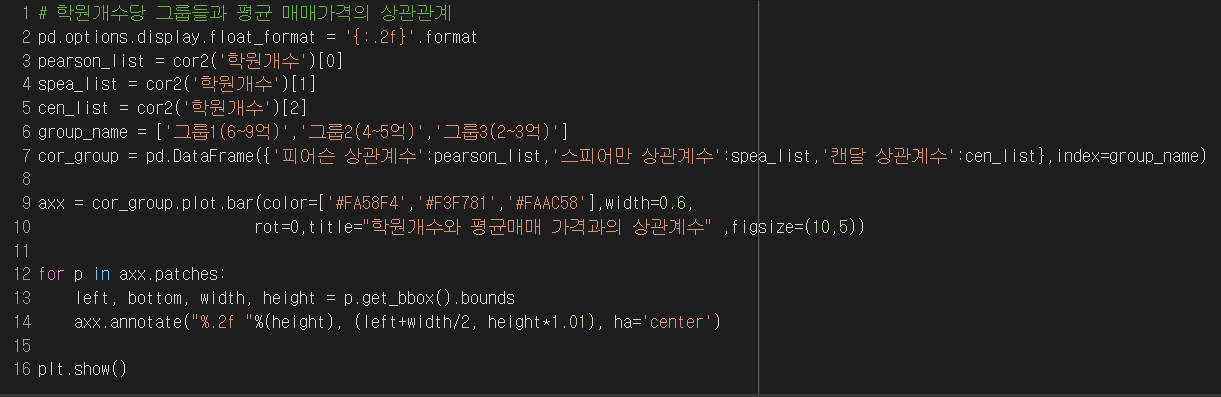


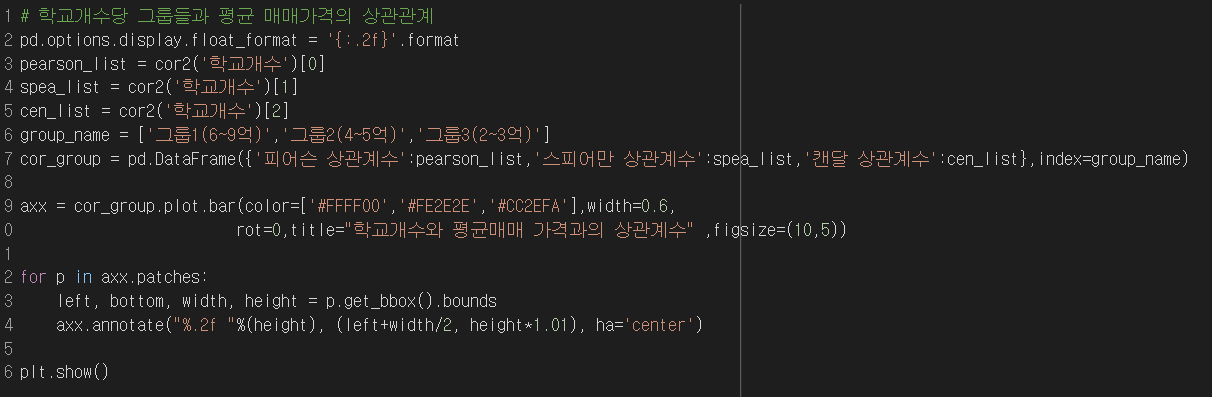
* 위 그래프는 그룹별 학교개수와 평균매매가격과의 상관관계를 나타낸 수치이다.
* 전체적으로 보았을때, 모든 그룹이 학교개수가 많아 질 수록 평균 매매가격이 올라간 것을 볼 수 있다.
* 평균매매가격이 높은 그룹이 낮은 그룹보다 학교개수와 매매가격관의 상관정도가 더 높음을 알 수 있다.
* 특이점으로는, 그룹화를 하지 않은 서울시 전체 구별 학교 개수에 따른 평균 매매가격과의 상관관계값이 작게 나온 반면, 그룹화를 시켜 그룹에서의 학교개수와 집값 평균 매매금액과의 상관관계 값은 높게 측정 되었다.



* 위의 그래프는 학교개수와 평균매매가격과의 관계를 Implot차트를 띄워 그룹별로 나타낸 것이다.
* 모두 양의 기울기를 갖음으로, 학교개수가 많을수록 평균 매매가격이 올라감을 알 수 있다.
* 그룹별로 학교개수와 평균금액과의 상관계수값이 높게 측정됨으로, 기울기가 1에 가깝게 기울여 보여야 하지만 그래프의 오차범위 표시로 인해 상대적으로 기울기 정도가 낮아보여서 그룹별로 확연한 차이를 보기 힘들다.
* 오차범위가 넓게 나온 이유는 그룹별로 학원개수의 차이가 크고 샘플들이 적게 존재함으로 오차범위가 넓다(신뢰도가 떨어진다)





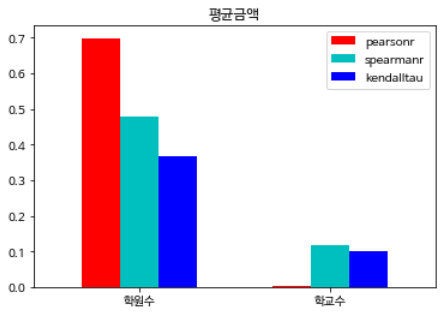


**코드설명**

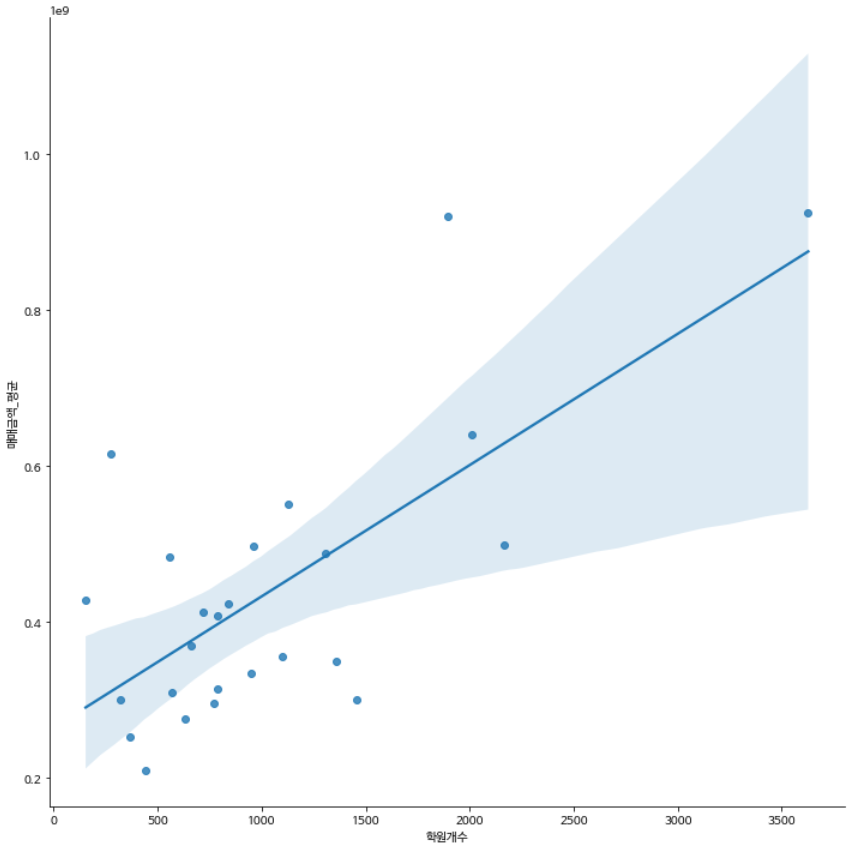
* 평수별 금액으로 부터 평균 매매가격을 구하고, make\_grp 리스트를 생성해 설정 가격별로 그룹화 한다
* 학원/학교 개수와 평균 매매가격관의 상관관계를 그룹별로 찍어주는 cor2() 생성.
* 변수간의 상관계수를 DataFrame으로 생성
* 생성한 DataFrame을 plot.bar와, plot.Implot 을 사용하여 시각화 한다.

**6. 결론 도출**

**결론1. 서울시 아파트 매매가격 평균과 교육기관의 상관관계**

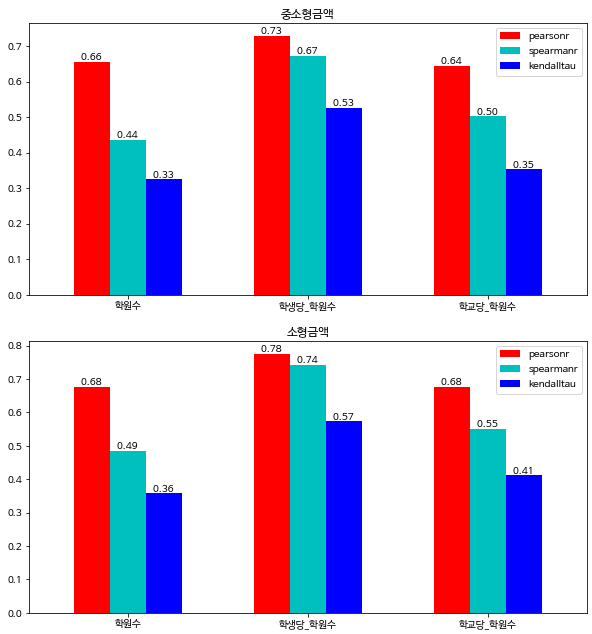
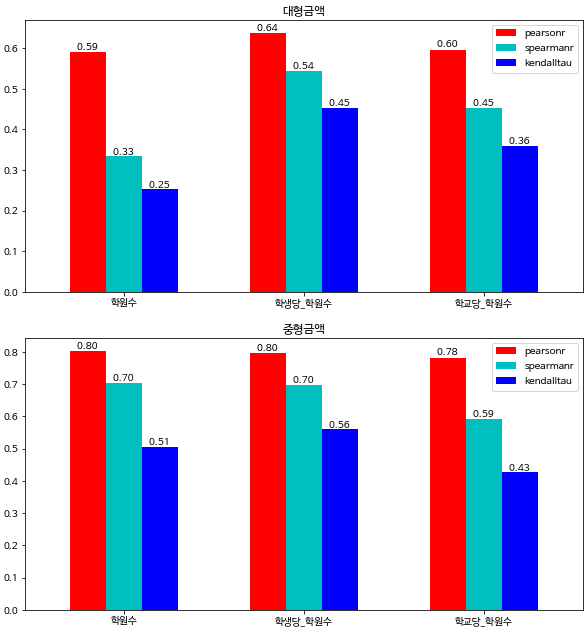


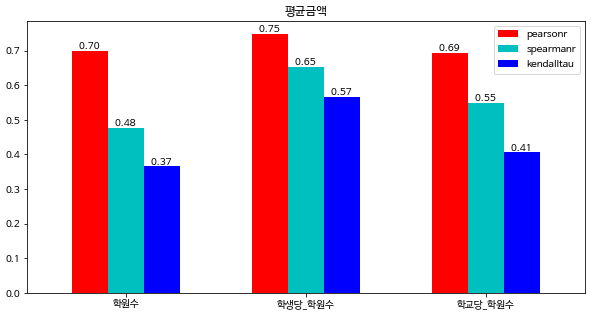
* 사진은 **서울시 평균 아파트 매매가격과 교육기관별 상관계수**이며 서울시 아파트 **매매가격은 학교수와 상관관계가 없음**을 알 수 있으며, 학원과 서울시 아파트 매매가격은 stats.pearsonr 기준 **0.699% 상관관계가 있음**을 알수있다



* 서울시 평균 아파트 매매가격과 학원수가 상관관계가 있음을 알수있었으며, 사진은 학원과 서울시 아파트 구별 매매가격과의 데이터 분포도로 **학원 개수가 많아짐에 따라 서울시 아파트 매매가격도 높아짐**을 알수있다

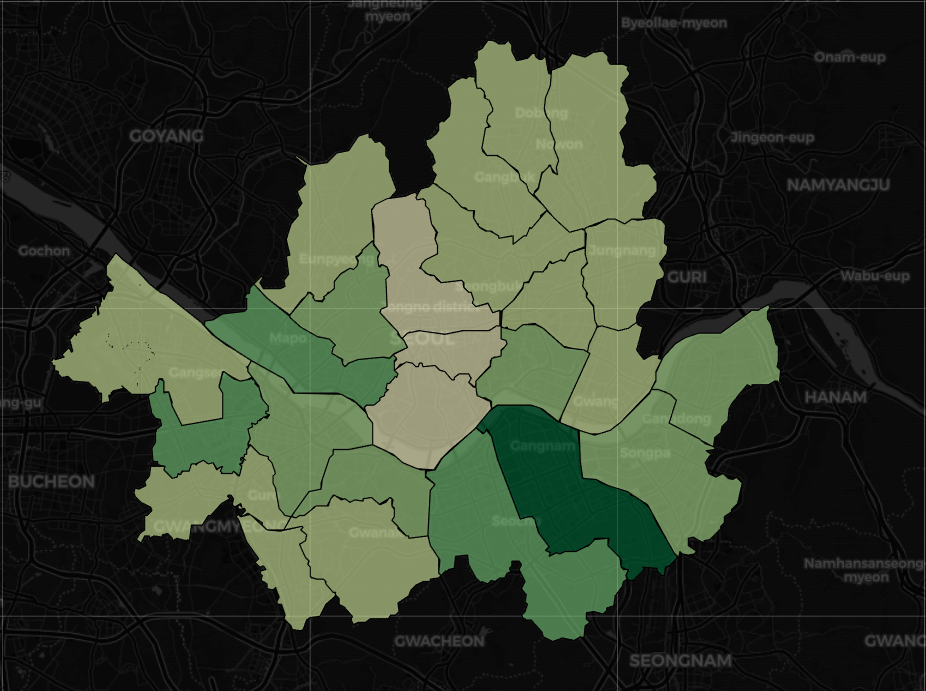
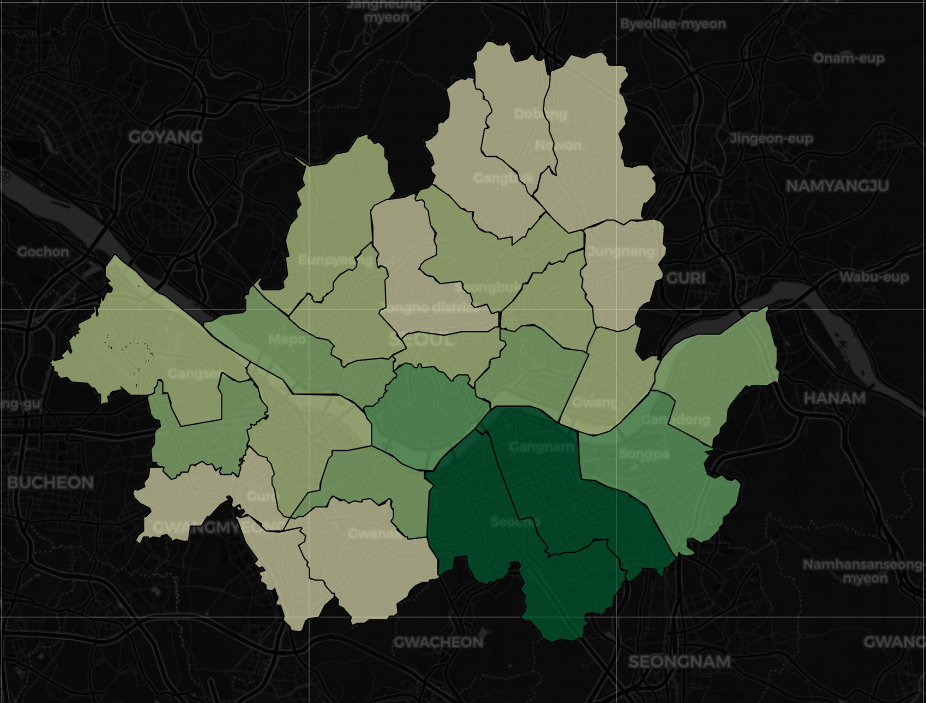
**결론2. 서울시 아파트 평형별, 평균 금액과 학원의 상관관계**





* 위 사진은 결론1. 의 결과 학원의 상관계수가 높았기에, 학원 관련된 컬럼과 평형별, 평균 금액과의 상관계수 그래프이다
* 평형별 금액과 **평균금액 결과** **학생당\_학원수가 가장 높은 상관계수**를 보였으며,

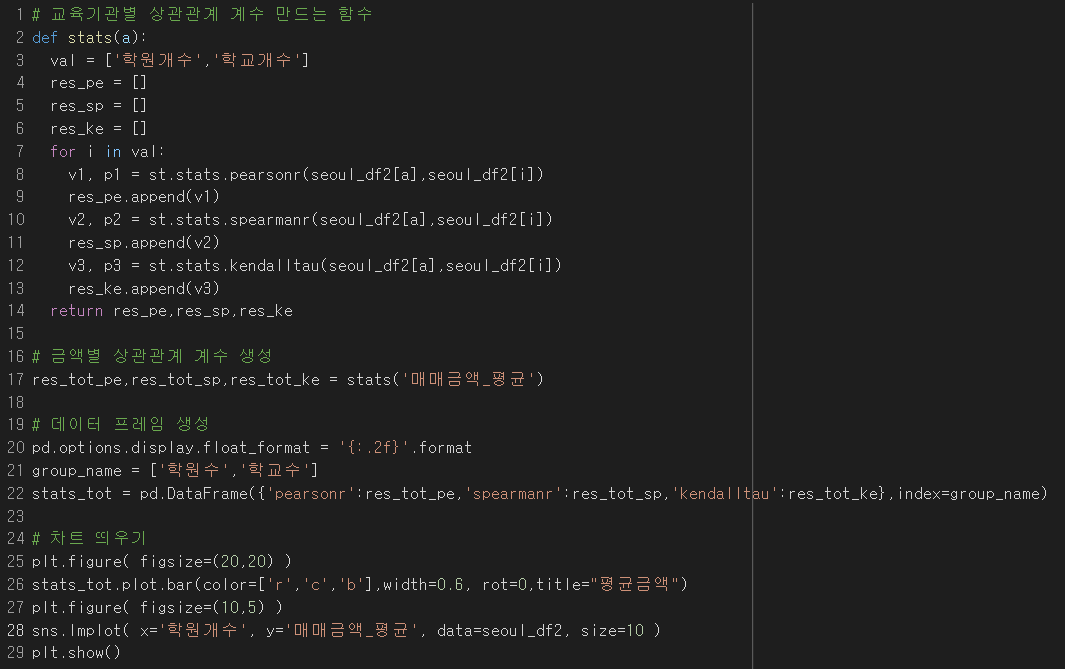
**평형별 가장 높은 상관계수는 중형금액별 학생당\_학원수** 이다.



< 서울시 구별 아파트 매매 평균 지도 > < 서울시 구별 학생당\_학원수 지도 >

**결론 소스코드**

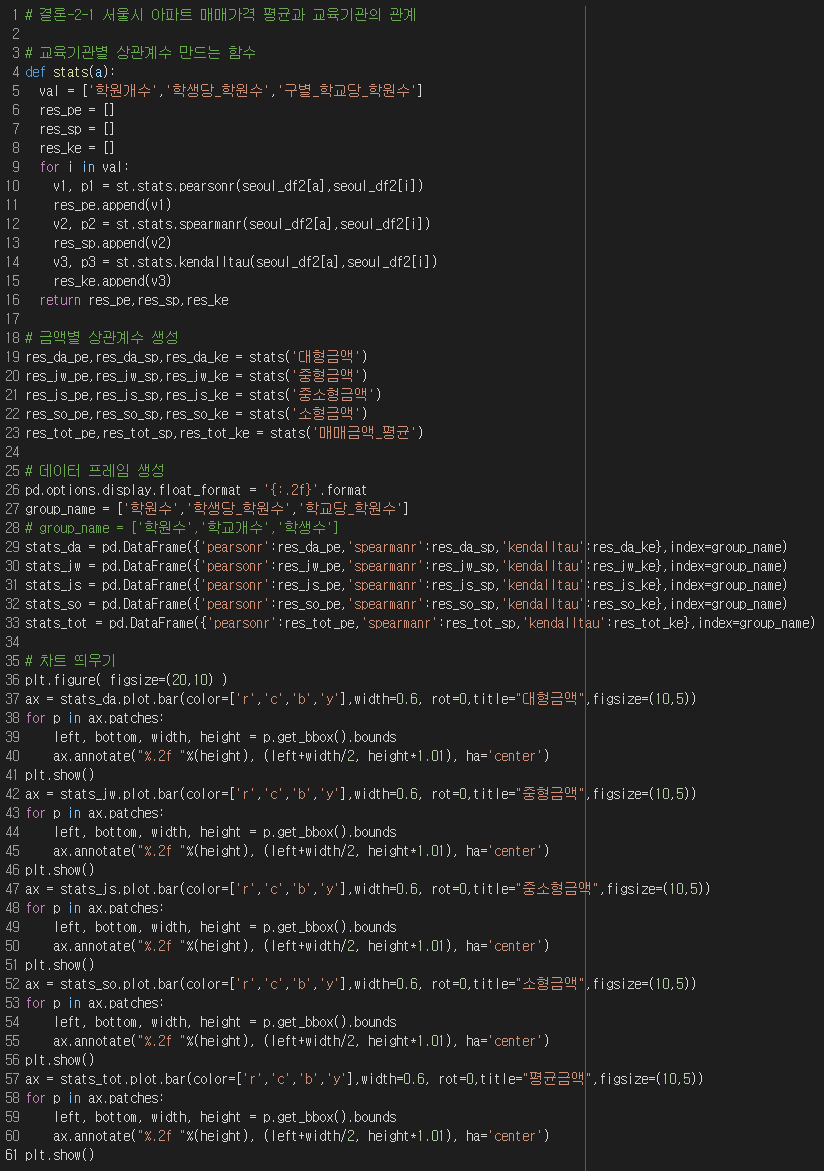
**결론1. 서울시 아파트 매매가격 평균과 교육기관의 상관관계**



**코드설명**

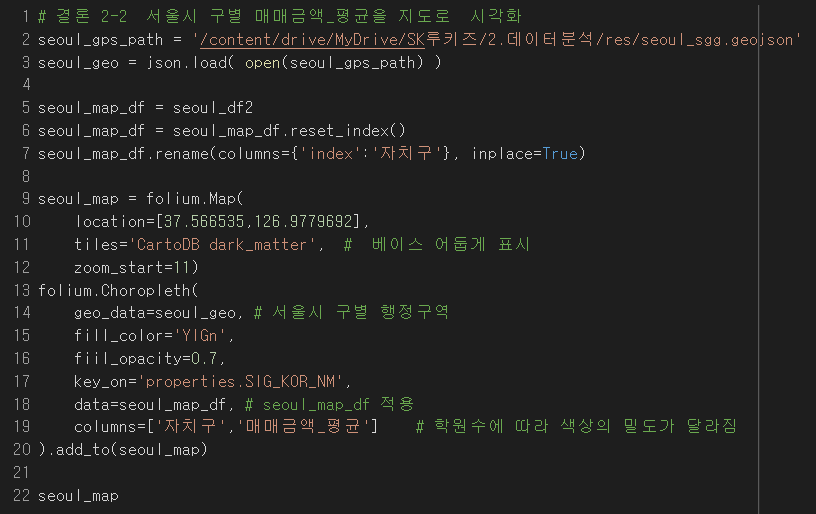
* stats의 pearsonr, spearmanr, kendalltau 상관계수를 만드는 **함수 stats()를 생성**
* res = stats() **함수**로 매매금액\_평균의 **상관계수 생성**
* res를 **데이터 프레임**으로 생성
* 데이터 프레임을 **plot.bar 차트**로 띄우기

**결론2-1. 서울시 아파트 매매가격 평균과 교육기관의 관계**

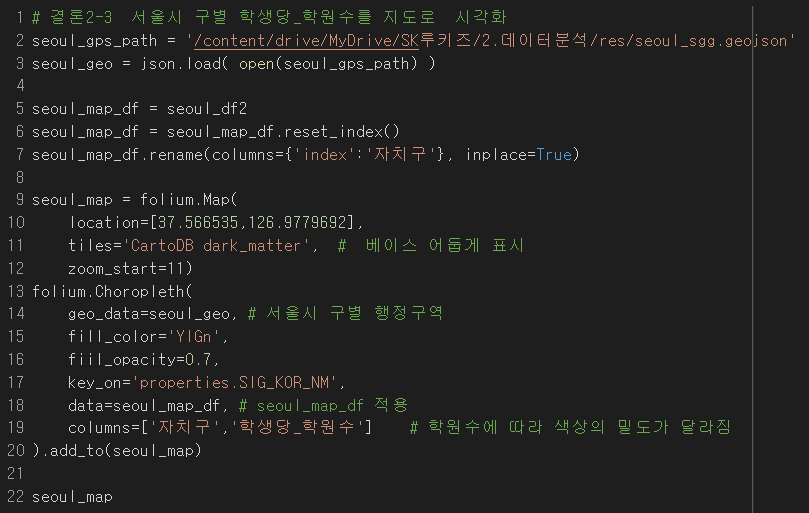


* stats의 pearsonr, spearmanr, kendalltau 상관계수 값을 리턴하는 stats()함수 생성
* stats() 사용하여 평형별 금액의 상관계수 생성
* 상관계수를 데이터프레임으로 생성
* plot.bar 사용하여 데이터프레임을 차트로 띄우기
* 값을 차트 bar위에 띄우기

**결론 2-2. 서울시 구별 매매금액\_평균을 지도로 시각화**



**결론 2-3 서울시 구별 학생당\_학원수를 지도로 시각화**



* 결론 2-2, 2-3은 folium 모듈을 사용해 지도로 시각화 하였으며, seoul\_sgg.geojson과 (학생당\_학원수, 매매금액\_평균)을 자치구로 묶어 색상의 밀도로 구분한 지도이다.