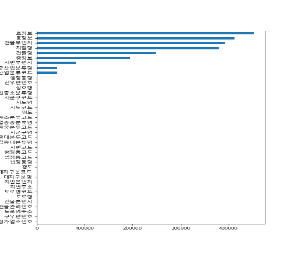
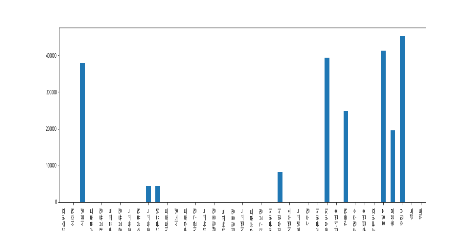
**1번부터 14번**까지 실행하시오.

제출 파일 : 본인이름사례연구.ipynb, 그래프설명.png들

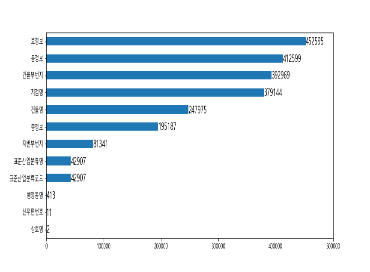
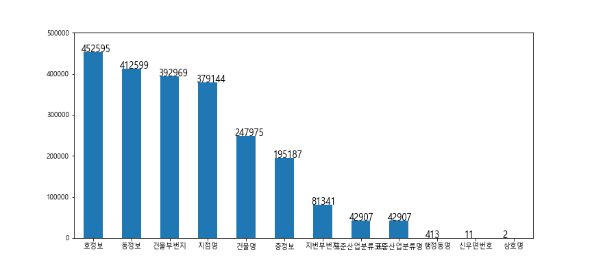
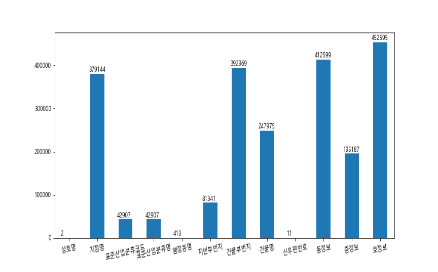
1. 필요한 라이브러리를 로드하고 시각화에서 디스플레이를 선명하게 표시되도록 하고, 한글 폰트가 나오도록 설정하고 확인한다. 한글깨짐 여부를 그래프로 test
2. <https://www.data.go.kr>에서 “상가(상권)정보”를 다운 받아, 서울과 부산데이터만 df 변수에 읽어들인다.
3. df 데이터 셋의 결측치 확인 및 시각화

[ResidentMario/missingno: Missing data visualization module for Python.](https://github.com/ResidentMario/missingno)

* 1. df 변수의 컬럼들을 확인하고 상위 3줄, 하위3줄을 출력한다.
  2. df 변수의 결측치를 제외한 데이터 개수 및 dtype들을 출력(dtypes함수)하고 변수가 사용되는 메모리 사용량 확인한다.
  3. 결측치를 확인하고 결측치가 없는 컬럼을 포함하여 시각화(정렬전 bar plot, barh plot, 정렬 후 bar plot, barh plot)



* 1. 결측치가 없는 컬럼는 제외하고 시각화(정렬전 bar plot, barh plot, 정렬 후 bar plot, barh plot).



1. df 데이터 셋에서 사용하지 않을 컬럼은 제거
   1. 결측치가 너무 많은 컬럼은 제거한다. 결측치가 너무 많은 상위 9개 컬럼은 제거함.
   2. 컬럼명에 “코드”나 “번호”가 있는 컬럼은 분석에 사용되지 않을 예정입니다. 제거함( df.columns.str.contains(“코드|번호”) 이용).

( df.columns.str.find(‘코드”)!=-1 ) | ( df.columns.str.find(‘번호”)!=-1 )

* 1. 컬럼을 제거 후 제거 전과 memory usage를 확인하고 줄어든 columns수도 확인한다.

1. df 데이터 셋의 값을 가져온다(loc함수, iloc함수등을 사용하기도 하고 사용하지 않기도 하며 서브셋을 가져온다)
   1. 상호명 필드만 가져온다
   2. 상호명의 종류별 데이터수를 가져온다(df.상호명.value\_counts()이용)
   3. “상호명”과 "도로명주소” 컬럼 가져오기
   4. 0~2행을 출력하라 (head()함수를 이용하지 않고, loc과 iloc을 이용)
2. 기술 통계값 보기

- [Descriptive statistics - Wikipedia](https://en.wikipedia.org/wiki/Descriptive\_statistics)

- [Computations / descriptive stats](https://pandas.pydata.org/docs/reference/frame.html#computations-descriptive-stats)

- [표준 편차 - 위키백과, 우리 모두의 백과사전](https://ko.wikipedia.org/wiki/표준편차)

\* describe()를 사용하면 데이터 요약을 볼 수 있음(기술통계량)

- 기본적으로 수치형 데이터를 요약

- include, exclude 옵션으로 다른 데이터 타입의 요약 수치 출력

- count, min, max, mean, median, 1사분위수, 3사분위수

### \*\*개별 기술 통계값 구하기\*\*

\* [Computations / descriptive stats](https://pandas.pydata.org/docs/reference/frame.html#computations-descriptive-stats)

\* count : 결측치를 제외한 값 갯수

\* min, max: 최솟값, 최댓값

\* argmin, argmax : 최솟값 인덱스, 최댓값 인덱스 반환

\* quantile 특정 사분위수에 해당하는 값을 반환 (0~1 사이)

\* 0.25 : 1사분위 수

\* 0.5 : 2사분위수 (quantile 의 기본 값)

\* 0.75 : 3사분위수

\* sum 수치 데이터의 합계

\* mean 평균

\* median 중앙값(중간값:데이터를 한 줄로 세웠을 때 가운데 위치하는 값, 중앙값이 짝수일 때는 가운데 2개 값의 평균을 구함)

\* std, var 표준편차, 분산을 계산

\* cumsum 맨 첫 번째 성분부터 각 성분까지의 누적합을 계산 (0 번째 부터 계속 더해짐)

\* cumprod 맨 첫번째 성분부터 각 성분까지의 누적곱을 계산 (1 번째 부터 계속 곱해짐)

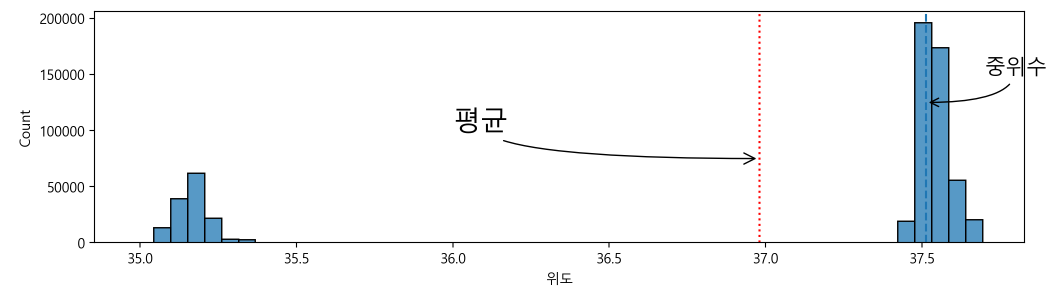
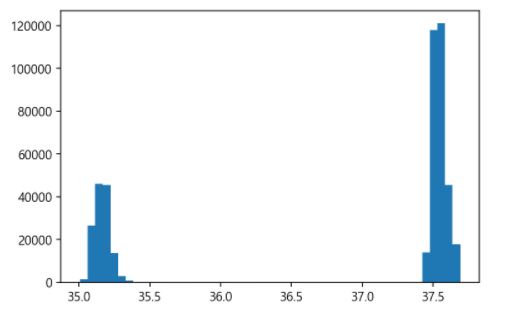
\* cov, corr() : 공분산, 상관계수

\* skew 왜도 (오른쪽으로 치우침= 왜도<0, 왼쪽으로 치우침=왜도>0)

\* kurt 첨도

* 1. df 데이터셋의 요약기술 통계량
  2. df 데이터셋의 “지번본번지", "건물본번지” 컬럼의 데이터 개수와 dtype 메모리 사용량을 확인한다.
  3. “위도", "경도” 컬럼만의 요약 기술통계량

1. 단변량 수치형 변수 시각화
   1. 위도의 빈도표를 시각화 : plot.hist, hist, sns.displot, sns.histplot

* 1. 경도의 빈도표를 시각화 : plot.hist, hist, sns.displot, sns.histplot

1. 상관계수

- 두 변수 간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법

- [상관 분석 - 위키백과, 우리 모두의 백과사전M](https://ko.wikipedia.org/wiki/상관\_분석)

- 결과의 해석 : r 값은 X 와 Y 가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일 하면 –1 을 가진다. 결정계수(coefficient of determination)는 $r^2$로 계산하며 이것은 X 로부터 Y를 예측할 수 있는 정도를 의미한다.

- ([수학기호 사이트 참조](https://matplotlib.org/2.0.2/users/mathtext.html))

- 일반적으로

\* r이 -1.0과 -0.7 사이이면, 강한 음적 선형관계,

\* r이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,

\* r이 -0.3과 -0.1 사이이면, 약한 음적 선형관계,

\* r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,

\* r이 +0.1과 +0.3 사이이면, 약한 양적 선형관계,

\* r이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,

\* r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계로 해석한다.

<img src="https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation\_examples2.svg/220px-Correlation\_examples2.svg.png" width="600">

이미지 출처 : [위키백과](https://ko.wikipedia.org/wiki/상관\_분석)

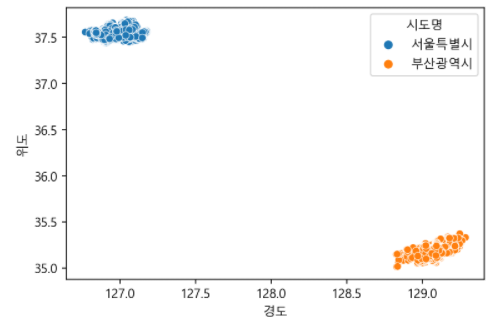
* 1. 전체 숫자 컬럼끼리의 상관계수
  2. 상관계수를 이용하여 heatmap 시각화

참조 <http://seaborn.pydata.org/examples/many_pairwise_correlations.html>

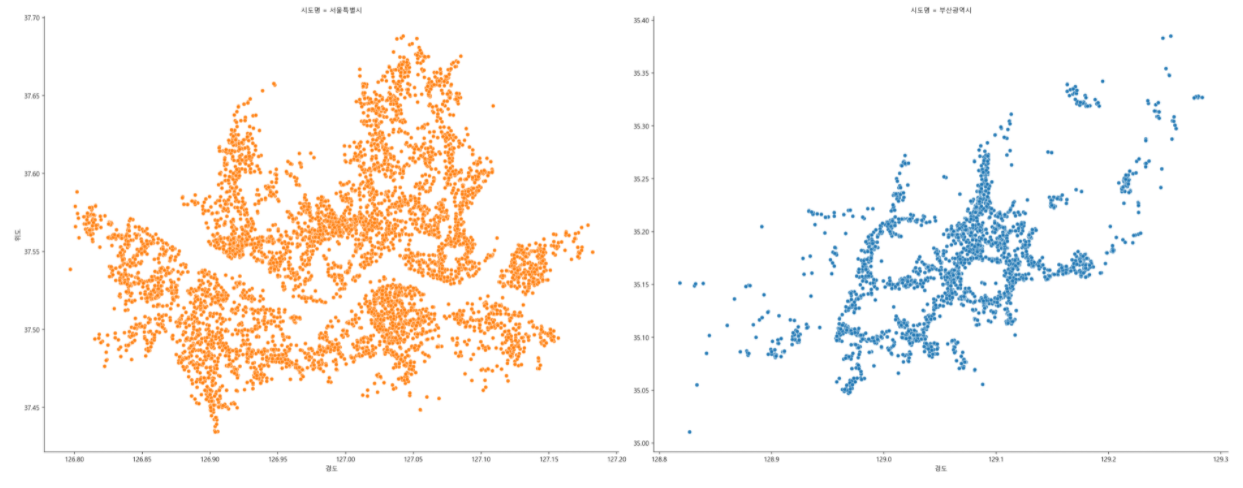
1. 경도와 위도 컬럼을 이용하여 산점도

[Anscombe’s quartet — seaborn 0.10.0 documentation](https://seaborn.pydata.org/examples/anscombes_quartet.html)

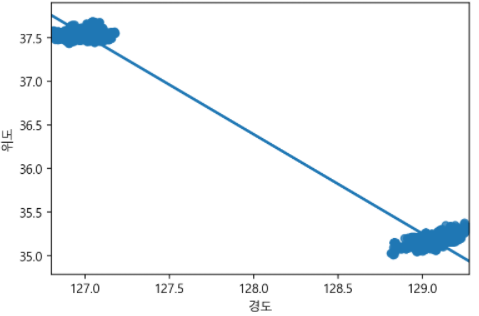
* 1. X축에는 경도, y축에는 위도 컬럼을 산점도로 시각화



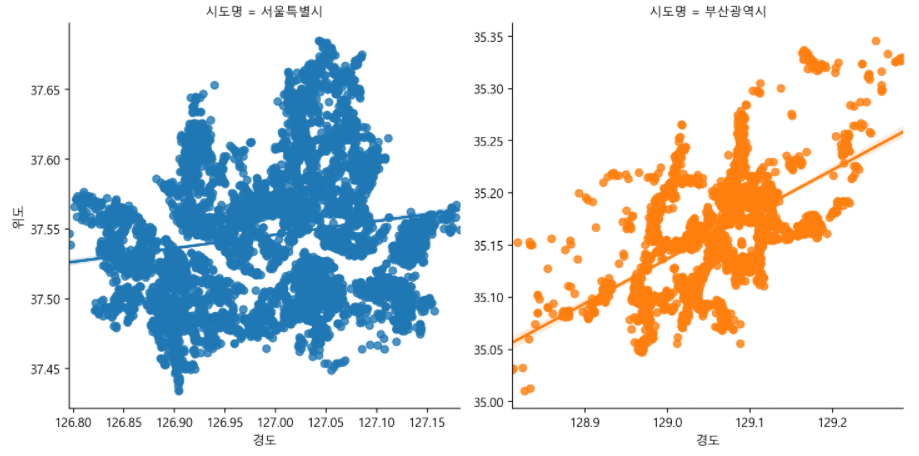
* 1. 경도와 위도의 산점도를 “시도명” 컬럼별로 서브플롯으로 시각화



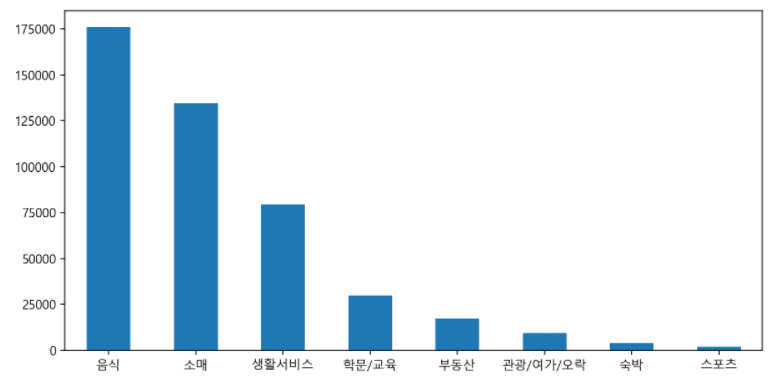
* 1. 위의 1번에서 시각화된 산점도를 바탕으로 회귀선을 그린다.



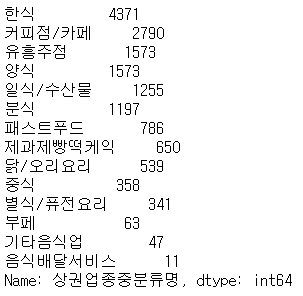
* 1. 위의 2번에서 시각화된 산점도를 바탕으로 회귀선을 서브 플롯으로 그린다.



1. 상권업종대분류명별 상호명의 개수를 도출하고 시각화하기



1. 상권업종대분류명이 음식인 서브셋을 이용한 분석
   1. “상권업종대분류명”이 음식인 서브셋을 변수 df\_food에 할당하고 확인
   2. “상권업종대분류명”이 음식이면서, "시군구명”이 강남구 데이터만 가져와 “상권업종중분류명”별로 빈도수를 구함(loc함수를 이용 vs loc함수 이용안함)



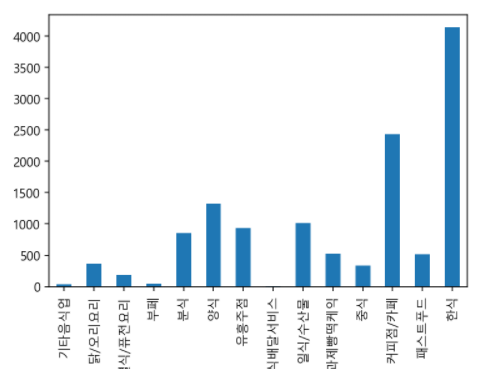
1. df 데이터셋에서 “상권업종대분류명”이 음식인 데이터 중 서울특별시 데이터 서브셋

<https://seaborn.pydata.org/tutorial/categorical.html> : 범주형그래프

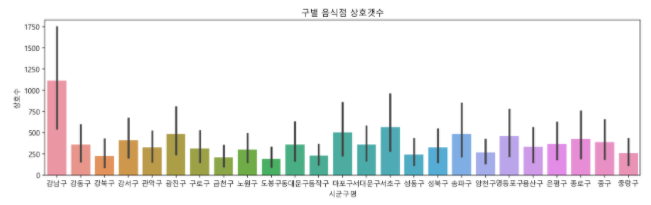
* 1. “상권업종대분류명”이 음식인 서브셋 중 서울특별시 데이터만 변수 df\_seoul\_food에 할당하고 확인
  2. df\_seoul\_food 데이터 셋을 시군구명, 상권업종중분류명으로 그룹화하여 상점수를 count한 내용을 food\_gu 변수에 할당.
  3. food\_gu 변수를 다음과 같은 스타일의 표로 출력(food\_gu 이용).



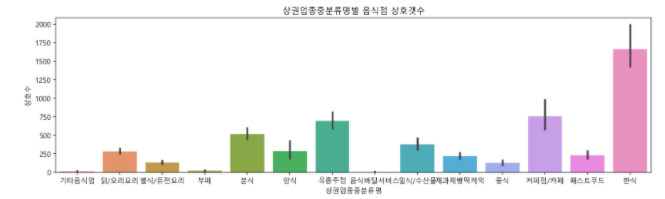
* 1. 위 3번 스타일의 표를 pivot\_table함수를 이용하여 출력
  2. 3번의 결과 중 강남구 데이터만 뽑아 barplot으로 시각화(판다스 plot이용)



* 1. 3번 food\_gu를 seaborn을 이용하여 구별 음식점 상호 개수를 시각화

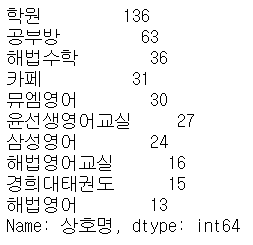


* 1. 상권업종중분류명별 음식점 상호갯수

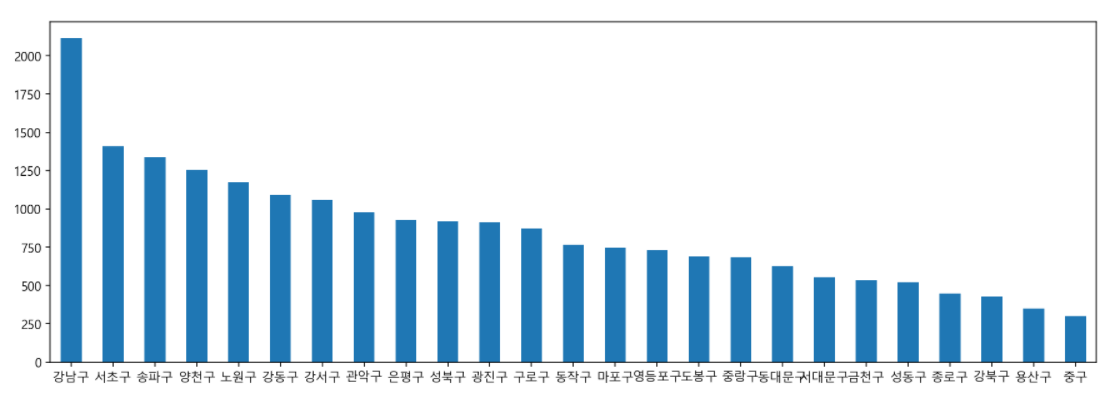


* 1. Seaborn의 catplot을 이용하여 상권업종중분류별 음식점을 구별로 상권업종중분류명별 bar차트를 서브플롯으로 시각화
  2. Seaborn의 catplot을 이용하여 구별 음식점을 상권업종중분류명별로 서브 플롯으로 시각화

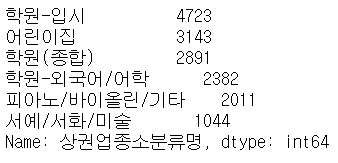
1. 구별로 학원수 비교 : 서울 대치동이나 목동에 사교육이 발달되었다는 가설을 뒷받침할 수 있는 분석
   1. 서울시 교육(상권업종대분류명 이용) 데이터를 df\_academy 변수에 할당하고 확인
   2. df\_academy 데이터 셋을 상호명별로 빈도수 출력(value\_counts()함수 이용하거나 groupby이용)
   3. df\_academy 데이터 셋을 상호명별로 빈도수 상위 10개 출력



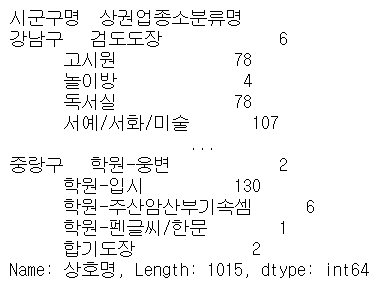
* 1. df\_academy 데이터 셋을 시군구명 별로 빈도수 출력(학원이 가장 많은 구부터 출력)



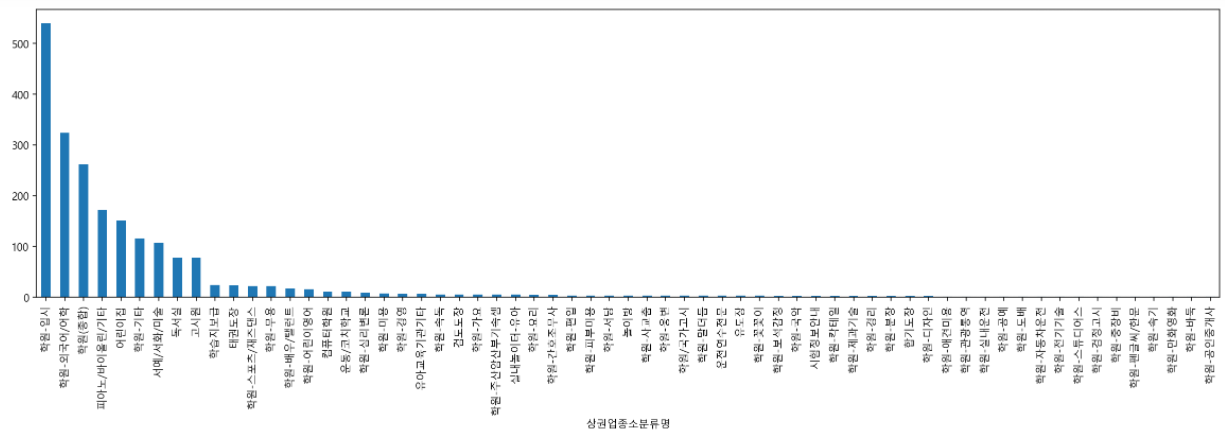
* 1. df\_academy 데이터 셋에서 어떤 종류의 학원들이 많은지 상위 10개만 academy\_count변수에 할당하고 출력(상권업종소분류명 컬럼 이용)
  2. df\_academy 데이터셋에서 상권업종소분류명별로 빈도수를 구했을 때 빈도가 1000이상인 데이터만 따로 academy\_count\_1000변수에 할당

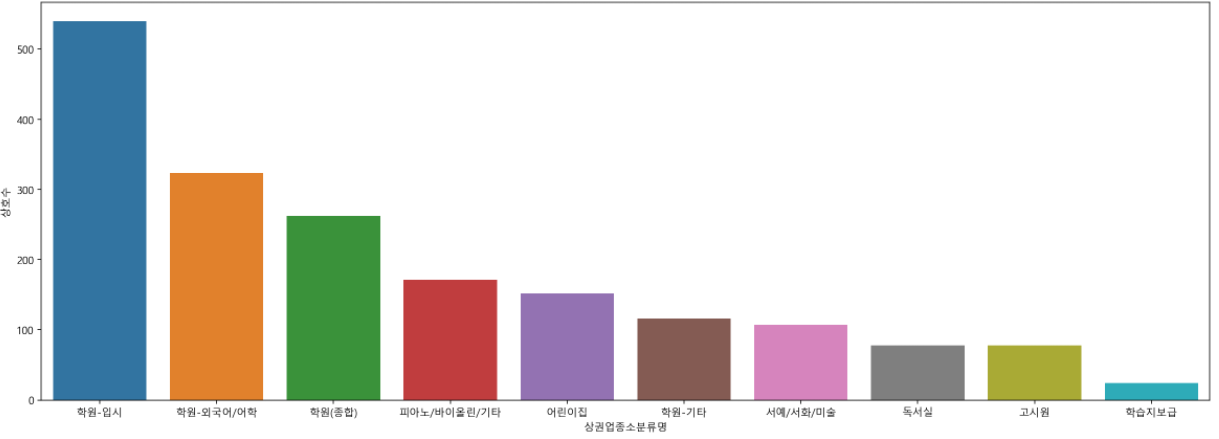


* 1. df\_academy 데이터셋을 “시군구명”, "상권업종소분류명” 별 상호명 빈도수를 academy\_group 변수에 할당 출력

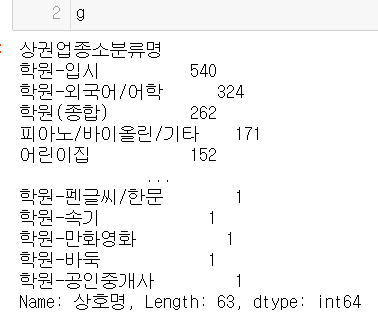


* 1. academy\_group 데이터셋에서 강남구 데이터만 출력 및 시각화(barplot)



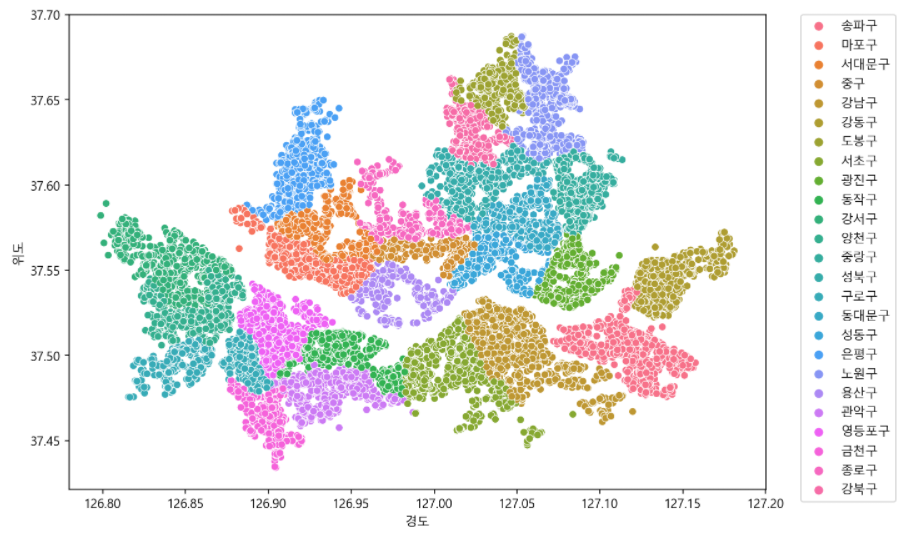


* 1. df\_academy데이터 중 “법정동명”컬럼이 “대치동”과 “목동”인 데이터만 가져와 상권업종소분류명별 빈도수 출력
  2. “상권업종소분류명”별 "시군구명” 별 상호명 빈도수를 g변수에 할당하고 출력

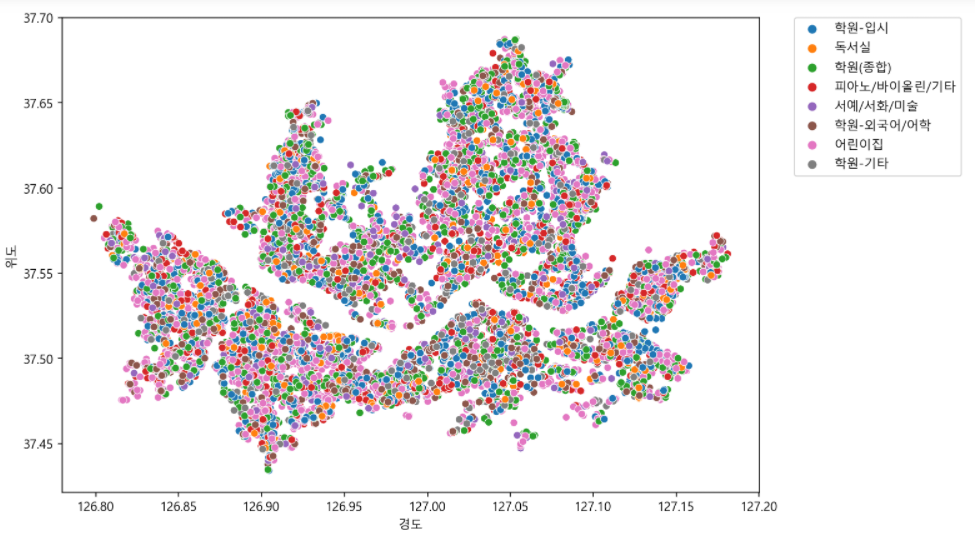


* 1. g변수의 내용중 "상권업종소분류명” 컬럼이 “입시·교과학원”데이터만 시각화(pandas의 plot.bar, pandas의 barh, seaborn의 barplot)

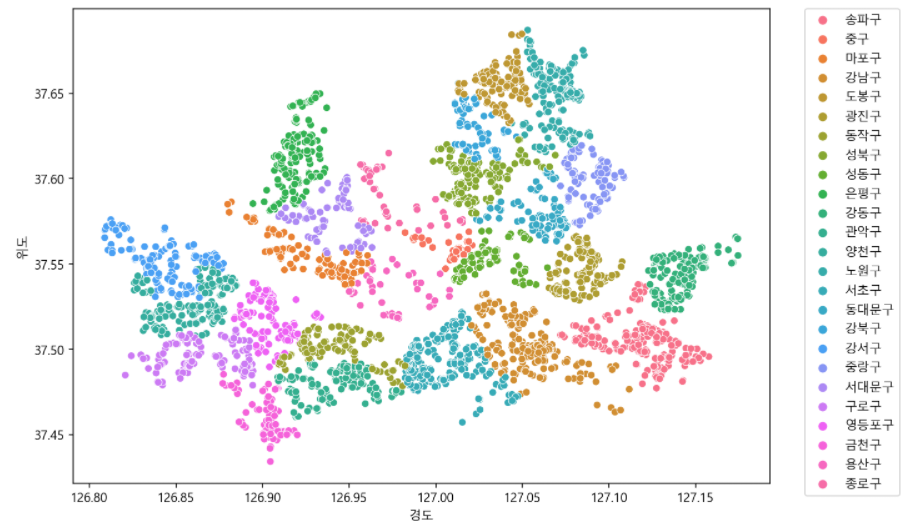
1. 서울시 데이터만 경도와 위도를 산점도로 시각화
   1. df\_academy 데이터셋의 경도와 위도를 “시군명”별로 색상을 다르게 scatterplot으로 시각화



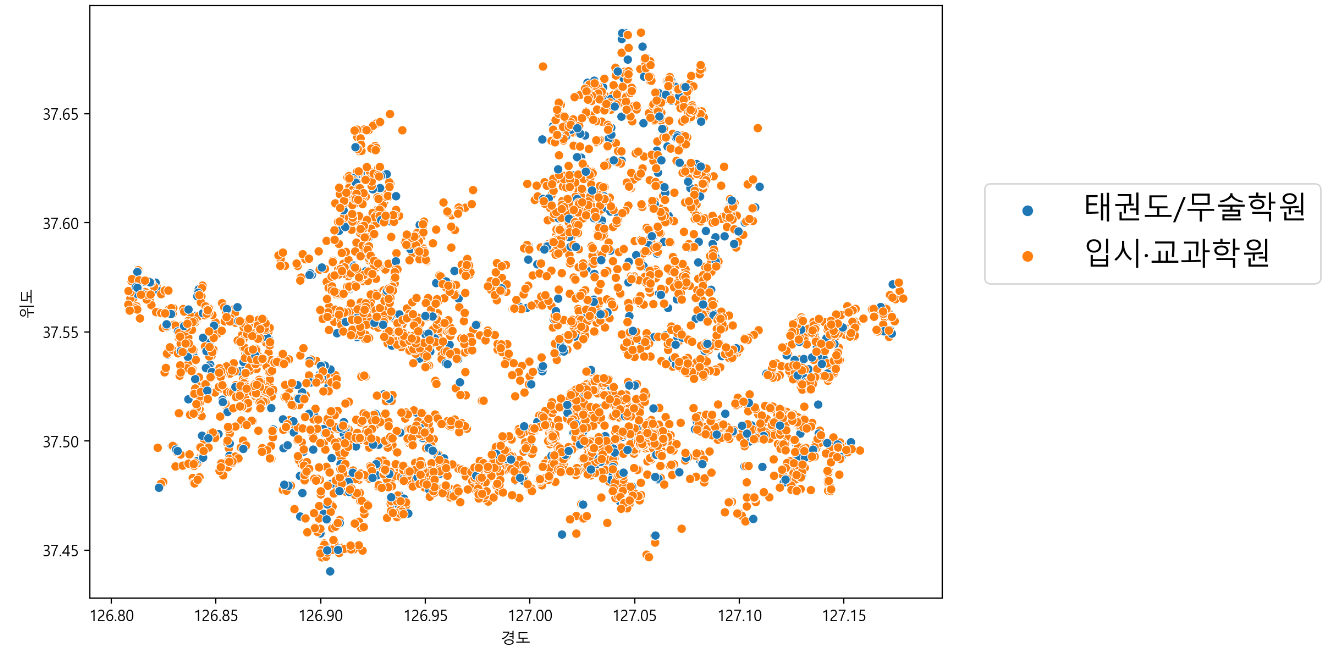
* 1. df\_academy 데이터셋의 경도와 위도를 “상권업종소분류명”별로 색상을 다르게 scatterplot으로 시각화



* 1. df\_academy 데이터셋 중 “입시·교과학원” 데이터만, 경도와 위도를 “시군구명”별로 색상을 다르게 scatterplot으로 시각화



* 1. df\_academy 데이터셋 중 “**태권도/무술학원**” 데이터만, 경도와 위도를 “시군명”별로 색상을 다르게 scatterplot으로 시각화
  2. df\_academy 데이터셋 중 “입시·교과학원” 데이터와 “**태권도/무술학원**” 데이터만, 경도와 위도를 “상권업종소분류명”별로 색상을 다르게 scatterplot으로 시각화



1. 수업시간에 할 지도 시각화 참조 사이트
   * docs : <https://python-visualization.github.io/folium/latest/getting_started.html?utm_source=chatgpt.com>
   * Quickstart : <https://python-visualization.github.io/folium/version-v0.9.1/quickstart.html?utm_source=chatgpt.com>