# 클린업 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1주차 패키지 문제의 조건 및 힌트는 Python을 기준으로 하지만, R을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **20일 목요일 23:59까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실 시도 퇴출이니 유의해주세요.

## Chapter 1. Data Preprocessing & EDA

클린업 1주차에서는 데이터분석 프로젝트의 A to Z를 간소하게나마 경험해보고자 합니다. 데이터 전처리부터 모델링, 해석까지 전반적인 데이터 분석이 어떻게 진행되는지 공부해보도록 하겠습니다. 데이터 분석 공모전에서는 주제를 풀어나가는 논리가 데이터를 기반으로 얼마나 잘 도출이 되는지가 핵심인데요, 다양한 기법을 활용하고 해당 결과들이 다음 단계로 어떻게 이어지는지에 대한 흐름을 잘 느껴보시길 바라겠습니다.

여러분은 중증외상센터 병원장입니다. 새로운 의원을 개업하기 위해 오피스 입지 선정을 진행해보고자 합니다. 먼저, 데이터를 탐색하는 EDA 과정이 잘 선행되어야 이후에 분석을 위한 중요한 인사이트를 얻을 수 있습니다. 다양한 관점에서의 EDA를 진행하여 어떠한 지역구에 병원이 들어서면 좋을지 파악해보겠습니다.

문제 1. data_gu를 불러온 뒤, df_gu로 저장해주세요. 데이터의 구조 및 변수들을 파악해보세요.

문제 2. EDA 시 자주 사용되는 함수에 대해 공부해보고, 각 기능 및 파이썬 기본 문법을 정리해서 작성해주세요.

```
ex. groupby() : 그룹화 이후 그룹마다 연산 적용
# 스팟별 사용자의 평균 나이 구하기
df.groupby("Spot_ID")["Age"].mean()
```

(Hint) drop(), astype(), fillna(), rename() 등이 있습니다.

문제 3. 추후 데이터 변환 등의 문제에서 자유롭기 위해 변수명을 영어로 통일해주겠습니다. 다음과 같이 변수명을 변경해주세요.

| 변경 전 | 변경 후 |
|---|---|
| 연월 | STD_YM |
| 도명 | DO_NM |
| 지역구 | GU_NM |
| 지역구 코드 | GU_CD |
| 업종분류 | CLASS |
| 총매출 | SALE |
| 인구수 | FLOW_POP |

문제 4. 불필요한 변수를 제거하겠습니다. DO_NM, GU_CD 컬럼을 삭제해주세요.

문제 5. 2025년의 자료만 추출하여 df_2025를 만들어주세요. 이후, 지역구별로 매출 및 유동인구 평균을 계산해주세요.

(Hint) STD_YM 변수를 datetime 형식으로 변환해주세요.

문제 6. 문제 5.에서 계산된 2025년의 평균 매출액과 유동인구의 관계를 scatterplot으로 표현하고, 상관계수를 계산하여 유동인구와 매출액의 관계를 해석해주세요.

문제 7. pyplot에서 사용되는 객체인 figure와 axes에 대해 조사해주세요. 한 번에 여러 그래프를 그리기 위해서 plt.subplots()를 이용하게 되는데, 인자로 받는 값은 무엇인지, 반환해주는 객체는 무엇인지 작성해주세요.

문제 8. 지역구별 업종별 매출 분포를 파악하겠습니다. 파이차트로 표현한 뒤, 일반적인 동향에 대해 해석해주세요. 이후, 의료기관 매출액이 전체 매출에서 차지하는 비율이 상대적으로 높은 지역을 추출해주세요. (df_gu를 사용해주세요)

문제 9. 피벗 테이블에 대해 1-2줄로 정리한 뒤, 의료기관 매출을 지역구별로 파악할 수 있는 피벗 테이블을 생성해주세요. 이후, 원하는 지역구들을 선정해 자유롭게 시각화해주세요.

(Hint) index = “GU_NM”, columns = “STD_YM”

기존 EDA를 통해서는 지역구별 차이를 한눈에 파악하기에는 한계가 있는 것 같습니다. 파이썬의 Folium 패키지를 통하여 지도에 시각화를 진행해보겠습니다.

문제 10. 필요한 데이터를 준비하겠습니다.

문제 10-1. Folium 패키지를 설치한 뒤, 다음의 코드를 실행하여 서울시 지역구를 나타내는 polygon에 대한 json 데이터를 받아주세요.

```
res=requests.get("https://raw.githubusercontent.com/southkorea/seoul-maps/master/kostat/2013/json/seoul_municipalities_geo_simple.json")
contents=res.content
seoul_geo=json.loads(contents)
```

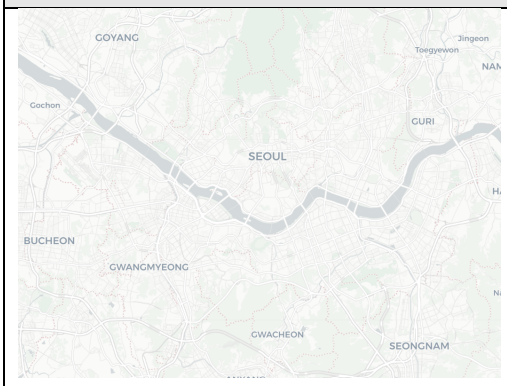
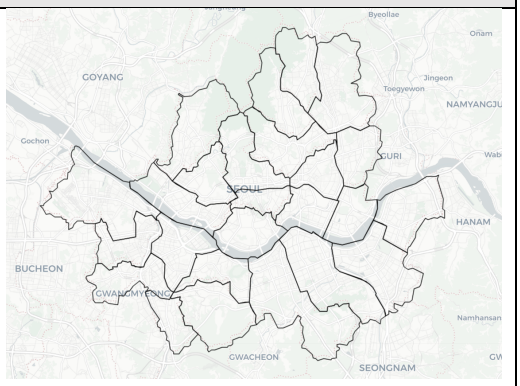
문제 10-2. 2025년 2월의 데이터만 필터링해주세요.

문제 11. 서울시 지도를 출력하겠습니다.

문제 11-1. Folium의 GeoJson 객체를 이용하여 받아온 Json 데이터를 출력해주세요.

(HINT) 위치는 [37.5665, 126.9780]을 중심으로 하고, zoom_start=11, max_bounds=True, tiles = 'CartoDB positron'을 설정하세요.

문제 11-2. Folium의 GeoJson 객체를 이용하여 받아온 Json 데이터를 출력해주세요.

| 문제 11-1 | 문제 11-2 |
| --- | --- |
|  |  |

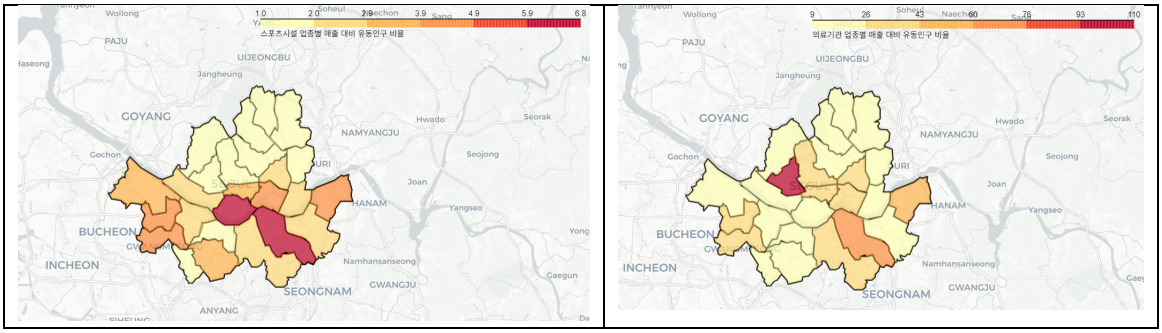


문제 12. Choropleth Map에 대해 조사한 후 간단히 작성해주세요.

문제 13. 각 업종 카테고리별 Choropleth Map을 생성하겠습니다. 지역구별 유동인구 대비 매출(SALE/FLOW_POP )를 기준으로 시각화해주세요.

(Hint) 유동인구 대비 매출 변수명은 FLOW_PER_POP입니다.

(Hint) fill_color="YlOrRd", fill_opacity=0.7, line_opacity=0.2, key_on="feature.properties.name"로 설정하세요.



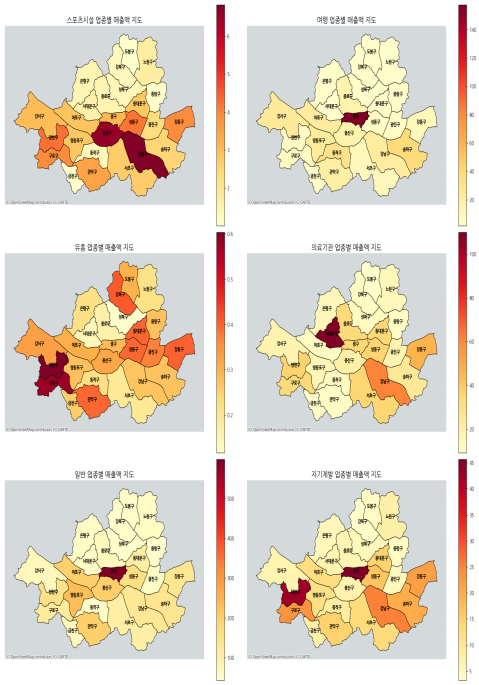

문제 14. 각 업종별로, 매출이 높은 상위 8개 지역구들을 파악해주세요. 대해서만 마커 객체를 추가해주세요.

(BONUS) 업종별 상위 8개 지역구에 대해 마커 객체를 추가해주세요.

문제 15. '의료시설'에 대해서만 유동인구 대비 매출이 유난히 낮은 지역을 선별하고, 데이터에서 삭제해주세요.

(Hint) 다른 시설에서는 값이 높은 편이나, 의료시설 지도에서만 값이 낮게 나오는 지역구를 선별해주세요.

(BONUS) 또 다른 시각화툴 geopandas 패키지를 활용해 문제 13.을 수행해주세요.



다른 카테고리에 비해 중구, 송파구, 관악구, 마포구, 영등포구에서 의료시설에 대한 매출이 부진한다고 해석할 수 있겠습니다. 해당 지역구를 제외하고, 최종 지역구를 선정해보겠습니다.

## Chapter 2. Modeling

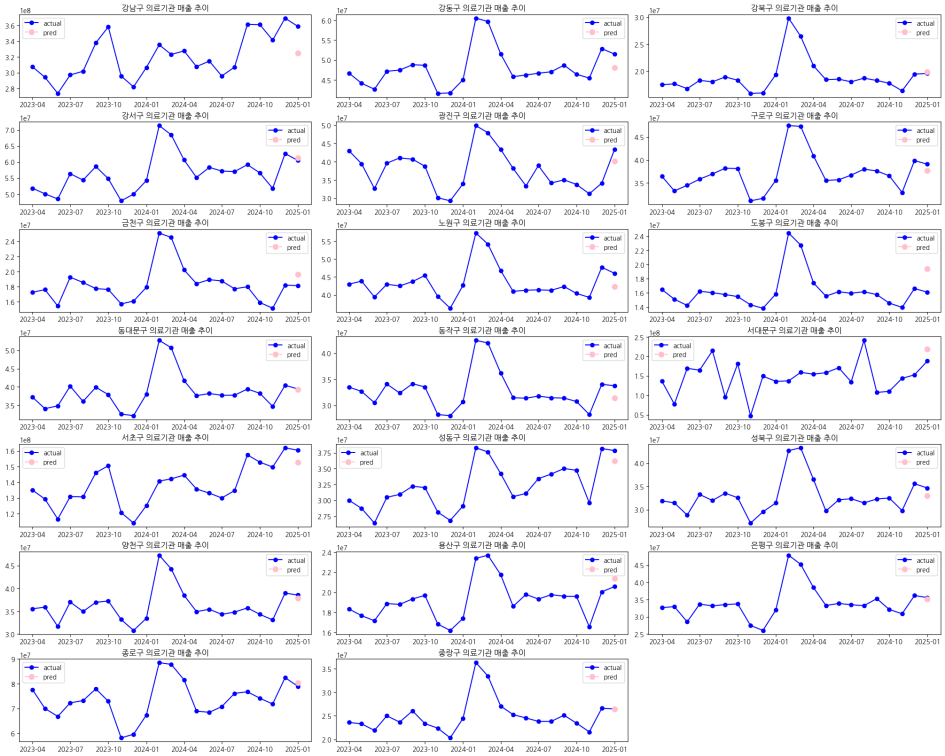
데이터 분석 프로젝트에서 모델링은 해당 모델을 선정한 이유와, 모델링을 통해 도출할 수 있는 결과가 얼마나 논리적인지가 중요합니다. 매출 예측 모델링을 통하여 최종 지역구 선정을 의료기관 잠재 매출이 증대될 것으로 예상되는 지역구를 최종 선정하겠습니다.

문제 1. XGBoost 모델에 대해 조사한 후, XGBoost 모델의 각 하이퍼파라미터에 대해 설명해주세요.

문제 2. 다음 코드를 실행하여, 지역구별 2025 년 2 월의 매출을 예측해주세요.

https://smart-crayon-83a.notion.site/Chapter-2-1b800d5053148034986aedbe40f682a5?pvs=4

문제 3. 분석 결과를 다음과 같이 시각화하고, 의료시설의 잠재 매출 증대 지역구를 추출해주세요. (시각화 형식은 자유입니다.)




저는 서대문구를 의료기관 매출 증가 잠재 지역으로 판단하고, 서대문구 내 오피스 건물 중 최적의 입주 건물을 선정하겠습니다.

# Chapter 3. AHP(Analytic Hierarchy Process)

서대문구 내 오피스 건물 중에서, 최적의 입주 건물을 선정하고자 합니다. 일반적인 입지선정 프로세스의 경우, 다양한 지표와 거리 등을 고려하여 행정동을 결정하지만, 이번 패키지에서는 특정 지역구 내 개별 건물들을 후보록 삼아 보다 정밀한 평가를 진행하려 합니다. 이를 위해, 입지 선정 시 고려해야 할 여러 기준의 중요도를 반영하여 최적의 대안을 도출할 수 있는 AHP(Analytic Hierarchy Process) 기법을 활용하고자 합니다.

AHP는 다양한 평가 기준을 계층적으로 구조화하고, 각 기준 간 중요도를 상대적으로 비교하여 최적의 대안을 도출하는 방법입니다. 이를 통해 입주 건물 선택 시 다양한 요소를 체계적으로 분석하고, 가중치를 반영하여 최적의 건물을 객관적으로 선정해보겠습니다.

추가로, 2023-2 선형대수학팀 주제분석 4주차에 AHP를 실제 분석한 사례가 있으니 참고해주시면 공부에 도움이 되실 것 같습니다.

문제 1. AHP 알고리즘에 대해 조사한 후, 어떤 단계로 수행해야 하는지 간략히 적어주세요.

문제 2. data_sdm을 불러와 df_office에 저장해주세요.

문제 2-1. 병원은 건물유형(bld_type)이 집합인 경우 들어설 수 없습니다. 해당 행들은 삭제해주세요.

**가격 / 컨디션 / 서비스** 세 가지 기준에 대해서 적합도를 평가하고자 합니다.
각 기준별 고려할 컬럼은 다음과 같습니다.
**가격**: dealt_yr, price_tr
**컨디션**: gfa_dlt, land_area, floor_gr, floor_bm, year_con
**서비스**: MON_OPER_TIME~SUN_OPER_TIME, STN_PROXIMITY_AT, STN_PROXIMITY_AT, PARKNG_POSBL_AT, WIFI_HOLD_AT,AC_HOLD_AT,HEATER_HOLD_AT

문제 3. 인플레이션을 반영하여 매매가를 보정해주도록 하겠습니다. 다음 원리로 조정 가격을 계산하겠습니다.

$$\text{조정 가격} = \text{매매가} \times \prod_{t=\text{거래연도}}^{2023}(1 + \textit{연간 변동률})$$

ex)

| 연도 | 변동률 |
|---|---|
| 2010년 | 0 |
| 2011년 | 10.0 |
| 2012년 | 9.09 |
| 2013년 | 16.67 |

2011년에 거래된 11억 건물의 2023년 조정 가격:
11억 x (1.1 x 1.0909 x 1.1667) = 15.4억

: 보정계수

문제3-1. dealt_yr(거래 연도)별 평균 매매가를 계산하여 price_avg 데이터를 만들어주세요.

문제3-2. price_avg를 이용하여 연도별 가격 변동률을 계산한 후, price_change_rate 컬럼을 추가

해주세요.

(Hint) .pct_change() 함수를 사용해주세요. 첫 번째 연도는 0으로 채웁니다.

문제3-3. 2023년을 기준으로 한 보정계수를 계산하겠습니다. cumprod() 함수를 활용해 계산한 후, adjust_factor 컬럼을 저장해주세요.

문제 3-1 ~ 3-3

| | dealt_yr | price_avg | price_change_rate | adjust_factor |
|---|---|---|---|---|
| 0 | 2014 | 17492.307692 | 0.000000 | 1.000000 |
| 1 | 2015 | 19455.555556 | 0.112235 | 1.112235 |
| 2 | 2016 | 20275.000000 | 0.042119 | 1.159081 |
| 3 | 2017 | 19181.818182 | -0.053918 | 1.096586 |
| 4 | 2018 | 23184.615385 | 0.208677 | 1.325418 |
| 5 | 2019 | 22945.454545 | -0.010315 | 1.311745 |
| 6 | 2020 | 26440.000000 | 0.152298 | 1.511522 |
| 7 | 2021 | 38230.769231 | 0.445944 | 2.185576 |
| 8 | 2022 | 35793.750000 | -0.063745 | 2.046257 |
| 9 | 2023 | 49875.000000 | 0.393400 | 2.851253 |

문제3-4. adjust_factor 컬럼을 활용하여 df_office의 adjusted_price 컬럼을 생성해주세요. 이후, dealt_yr 컬럼과 price_tr 컬럼을 삭제해주세요.

문제4. 컨디션 기준에 해당하는 컬럼들의 기초 통계량을 확인하고, 각 변수별 boxplot을 그려 이상치를 확인해주세요.

문제5. 각 변수마다 어떤 스케일링을 진행하면 좋을지 스케일링의 종류에 대해 알아본 후, 서술해주세요.

문제6. 본인이 판단하기에 적합한 스케일링 방식을 선택하여 스케일링을 진행해주세요. 저는 Robust scaling을 적용하겠습니다.

문제7. 컨디션 변수 내에서 어떤 변수가 더 중요한지에 대한 가중치를 반영하고 싶었으나, 컨디션 세부 항목에 대해 설문하는 것을 누락하였습니다. 대신 각 변수가 매매가에 미치는 영향을 회귀 분석을 통해 추정하고, 이를 가중치로 사용하고자 합니다. 동일한 방식으로 스케일링된 변수들을 적합한 회귀식의 계수를 가중치로 반영해도 괜찮은 이유를 서술해주세요. (단, 기본적인 회귀 조건은 만족하고 있고, 충분한 설명력을 지니고 있으며 다중공선성으로부터 자유롭다고 가정합니다. )

(Hint) 회귀 계수를 그대로 변수의 상대적 중요도로 해석하는 것이 항상 타당한지 여부를 고려하여 답변해주세요.

문제 8. 아래 회귀계수 가중치를 반영하여 condition_score를 계산해주세요. (Σ *가중치 × 스케일링된변수값*)

```
weights={
"gfa_dlt_scaled": 0.4015,"land_area_scaled": 0.3488,"floor_gr_scaled": 0.2237,
"floor_bm_scaled": 0.0122,"year_con_scaled": 0.0013}
```

문제 9. 서비스 기준에 대한 전처리를 진행하겠습니다. (  )_OPER_TIME 컬럼의 이용시간이 00:00~24:00인 경우 'full time', 13시간 이상 24시간 미만인 경우 most time, 13시간 미만인 경우 half time, 휴무인 경우 closed로 처리하겠습니다. 시간대별로 범주를 나누는 함수 categorize_time()을 만들고, 적용해주세요.

문제 10. 범주형 변수의 인코딩을 진행하겠습니다.

문제 10-1. 평일의 경우, 월~금 중 최빈값을 선택하고, ordinal encoding을 진행하여 weekday_oper_time 컬럼을 생성해주세요. 주말의 경우, 동일한 방식으로 진행하되 '휴무'는 더 큰 패널티를 주기 위해 -1로 인코딩해주세요.

(Hint) mode() 함수를 사용하세요.

(Hint) Scikit-Learn 패키지의 OrdinalEncoder()는 자체적으로 -1로 인코딩해주지 않기 때문에, value 별로 mapping하는 것이 좋습니다.

문제 10-2. STN_PROXIMITY_AT, PARKNG_POSBL_AT, WIFI_HOLD_AT, AC_HOLD_AT, HEATER_HOLD_AT 컬럼에 대해서는 label encoding을 진행해주세요.

문제 11. 인코딩한 값을 모두 합하여 service_score를 계산해주세요.

문제 12. 세 가지 기준(adjusted_price, condition_score, service_score)을 반영한 최종 지수를 계산하겠습니다. AHP의 쌍대비교행렬을 고윳값 및 고유벡터 관점에서 분석해보고자 합니다.

문제 12-1. 세 가지 기준에 대한 쌍대비교를 위해 실행한 설문조사 데이터를 불러와, df_survey에 저장해주세요.

문제 12-2. df_survey는 각 기준을 쌍대비교했을 때 얼마나 더 중요한지에 대한 설문 내역입니다. 기준별 평균을 내어 쌍대비교행렬을 생성하고, 기준 간 쌍대비교를 해주세요.

| | Condition | Price | Service |
|---|---|---|---|
| Condition | 1.000000 | 3.100000 | 3.10 |
| Price | 0.322581 | 1.000000 | 1.36 |
| Service | 0.322581 | 0.735294 | 1.00 |

(Hint) 쌍대비교: A 조건이 B 조건보다 n만큼 중요하다.

문제 12-3. AHP에서 가중치를 도출하는 핵심 원리는 쌍대비교행렬의 최대고윳값과 고유벡터를 구하는 것입니다. 고윳값과 고유벡터의 의미를 생각해본 후, 왜 AHP에서 최대고윳값과 그에 대응하는 고유벡터를 사용하는지 설명해주세요.

문제 12-4. AHP 가중치 벡터를 계산하겠습니다. 최대고윳값에 대응하는 고유벡터를 정규화하여 최종 가중치를 도출해주세요.

(Bonus) 고유벡터를 정규화한 벡터가 각 기준의 최종 가중치 벡터가 되는 이유를 설명해주세요.

문제 12-5. 수행 시 만족해야 하는 일관성 검증에 대해 조사하고, 일관성 검증을 위한 다음 코드를 완성해주세요. 이를 바탕으로 쌍대비교행렬 A의 일관성을 평가해주세요.

```
def ahp_consistency_check(pairwise_matrix):

    n = pairwise_matrix.shape[0]
    eigenvalues, eigenvectors = np.linalg.eig(pairwise_matrix)
    lambda_max = ## 코드를 채워주세요 ##

   # 일관성 지수 (CI) 계산
   CI= ## 코드를 채워주세요 ##

   # 평균 무작위지수
   RI_dict={1: 0.00, 2: 0.00, 3: 0.58, 4: 0.90, 5: 1.12, 6: 1.24, 7: 1.32,
            8: 1.41, 9: 1.45}
   RI= ## 코드를 채워주세요 ##

   # 일관성 비율 (CR) 계산
   CR= ## 코드를 채워주세요 ##

   return CR
```

문제 13. AHP 도출된 가중치 벡터를 활용하여 오피스별 최종 점수를 계산해 최종 입주 후보 건물을 출력해주세요.