ADP 17회 Review

1. 머신러닝 (data: Housing data – log1p로 정규화) – M/L Regression

- EDA, Preprocessing

-> log1p 정규화(왜도<0으로 인한 데이터 정규화)

EDA : 데이터 산포 확인, 결측치 확인, 이상치 확인
변수간 상관관계 확인, 통계치 확인 (시각화 활용)
Preprocessing : 산포 특이점 시 변환, 결측치 대체
필요시 One-hot encoding 등 변수 변환

- 모델링, 예측

-> Train, Test 데이터 Split하여 M/L 모델링 진행 (**모델 선정**), 예측하여 모델 성능 파악(기준 : RMAE)

-> Xgboost, RandomForest, SVM 등등 다양한 모델 중 데이터에 **알맞은 모델 선정** 필요 (알려진 기준)

- Hyperparameter 조절, RMAE 기준 오차 줄이기

-> GridSearchCV, BaysianOptimization 등 기법 활용하여 Hyperparameter 선정. RMAE 기준 오차 줄이기

2. 시계열분석 및 시각화(data : Covid19 – 일별 확진자수, 일별 완치자수)

- 코로나 위험지수 제작, 설명 -> 위험지수 높은 국가 선정해 시각화

-> 일별 확진자수(시계열 그래프) Decompose 하여 Trend 보기 (평일–많고, 주말-적음, Seasonal – 주기 7 예상)

-> 혹은 누적 확진자수 / 누적완치자수 비율을 구해서 Trend 보기 등등 다양한 정답 예상

- 한국의 코로나 확진자 예측 : 선형 시계열모델, 비선형 시계열 모델

-> 선형 시계열 모델 : 기존 ARIMA / SARIMA 분석

-> 비선형 시계열 모델 : Kalmann Filter / 지수평활법

3. 통계분석(data: 설문조사 – A~S까지의 그룹 설문조사 중간에 반대 문항)

- 그룹별 통계치 계산

-> group_by를 통해 각 설문조사 문항마다 통계치를 계산. Insight 적기

- 탐색적 요인분석을 표로 작성

-> PCA / FA 분석 진행 <-> Bartlett검정(요인분석의 적합성 여부) / KMO 검정(변수들간의 상관관계 지표)