제 4회 빅스타(빅데이터, 스타트업) 경진대회

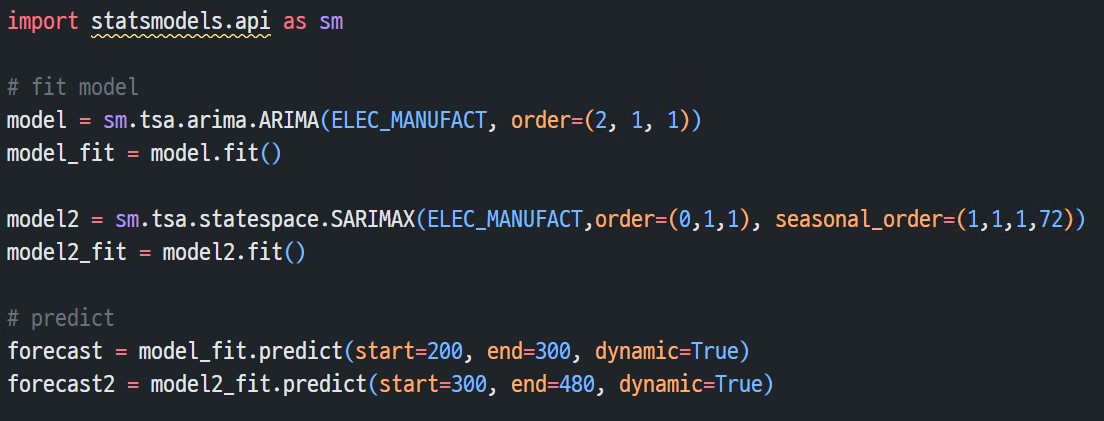
코드 설명자료

2022년 11월 25일

팀명 : 토마토

1. 라이브러리 및 데이터 (Library & Data)

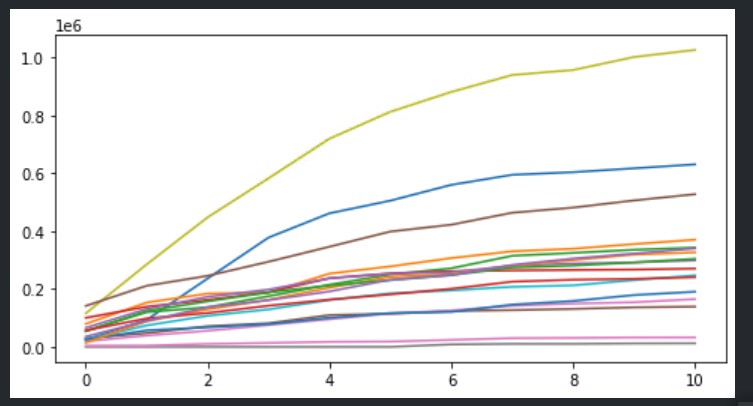
* 라이브러리 및 데이터에 대한 분석 결과에 대한 내용 서술
* 25년간의 데이터로 15년 미래를 예측하라는 과제가 주어졌기 때문에 매우 난감하였다.
* ARIMA, SARIMA 등 단일 시계열 예측을 할 경우 단순 추세, 계절성, 이동평균 등 만을 사용하기 때문에 예측의 근거가 부족하고, 결국 도메인 지식에 의존한 예측값을 얻어내야 하기 때문에 모델을 만드는 의미가 없다.
* ARIMAX, SARIMAX 같은 다변량 시계열 예측을 할 경우에는 추가적인 변수를 반영 할 수 있지만 주어진 데이터가 과거 25년의 데이터밖에 없기 때문에, 사용할 변수들의 미래 값을 다시 예측해야 하는 문제가 발생하며, 예측값을 구하는 과정에서 각각의 변수에 대한 도메인 지식에 의존해야 하는 문제가 발생한다,
* 그러나 대회에서 주어진 과제는 가스공사의 모델이 예측한 값과의 오차를 점수로 하기 때문에 결국 점수를 여러번 제출하며 가스공사가 사용한 미래 변수를 예측해야하는 과제가 발생하였고, 페이스북에서 만든 prophet 모델을 활용하여 다양한 미래 변수 예측값을 각종 연구자료에서 찾은 미래 예측을 바탕으로 산출한 뒤, 이를 통해서 LSTM 모델을 학습시켜 값을 예측하는 방향으로 진행하였다.



(ARIMA 모델을 통한 예측)

2. 데이터 전처리 (Data Cleansing & Pre-Processing)

* 사용한 데이터
  + 기온효과
    - 한국 가스공사에서 기존에 사용 중인 도시가스 수요의 기온효과 자료를 활용한다.
  + TCL
    - 통계청에서 제공하는 전국의 광역시 및 도의 기온과 해당 지역의 도시가스 가수를 곱하고 전국 도시가스 가구 수로 나눈 값들을 모두 합쳐, 실질적인 기온 변화에 따른 지역별 민간 도시가스 사용 예측치를 평균/최고/최고 기온으로 나누어 계산한 Temperature correlate LPG(TCL) 라는 지수를 설정한다.
  + 발전전력량
    - 발전전령양에 사용되는 다양한 분류들 사이에서 기력에 해당하는 데이터의 경우 LNG가 가장 중요한 지표이기 때문에 소분류 단위로 구성, 대분류의 경우 한전 및 타사 발전이 주가 되고 자가발전의 경우 매우 작은 값이므로 사업자 데이터를 사용한다.
  + 용도별판매전력량
    - 전력량에 대한 다양한 용도의 판매량 합산 치를 사용한다.
* 전처리 방식
  + 결측치 제거
    - TCL
      * 전국의 도시가스 가구는 연별로 크게 차이가 나지 않는다는 가정으로 월별로 나누고 결측치의 경우 전체 증가 추세를 반영하여 계산한다.

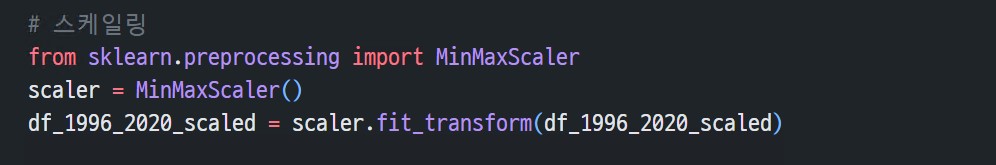


(전국 도시가스 가구 수 시각화)

* + - 발전전력량
      * 문제
        + 수력, 화력, 원자력, 기력, 내연력의 경우 96년에서 97년까지 총합만 존재한다.
        + 집단에너지, 대체에너지, 신재생에너지의 경우 96년도에서 14년도의 경우 대체 에너지만 존재, 15년도에서 20년도에는 신재생 에너지만 존재한다.
        + 기력 발전 중 LNP와 무연탄, 중유, 유연탄으로 세분화한다.
      * 해결책
        + 수력, 화력, 원자력, 기력, 내연력의 경우 소분류별 총합을 통해서 96년과 97년을 합친다.
        + 타사 발전에서는 대체에너지와 집단에너지를 합한 값만 존재, 통계청의 경우 이 세가지 유형을 유사하게 판단하여 제공, 실제로 다른 발전 유형들에 비해 유사하므로 하나의 컬럼으로 통합, 대체에너지와 신재생에너지는 과거에는 발전량이 적으므로 결측치의 경우 0으로 대체한다.
        + LNG와 그 외 2가지 컬럼으로 통합한다.
      * 최종적으로 ‘수력, 화력, 원자력, 내연력, LNG, LNG 제외한 기력, 집단 + 대체 + 신재생에너지’를 사용하며 Min-Max 스케일링을 통해 자료값들의 크기 차이를 해결한다.

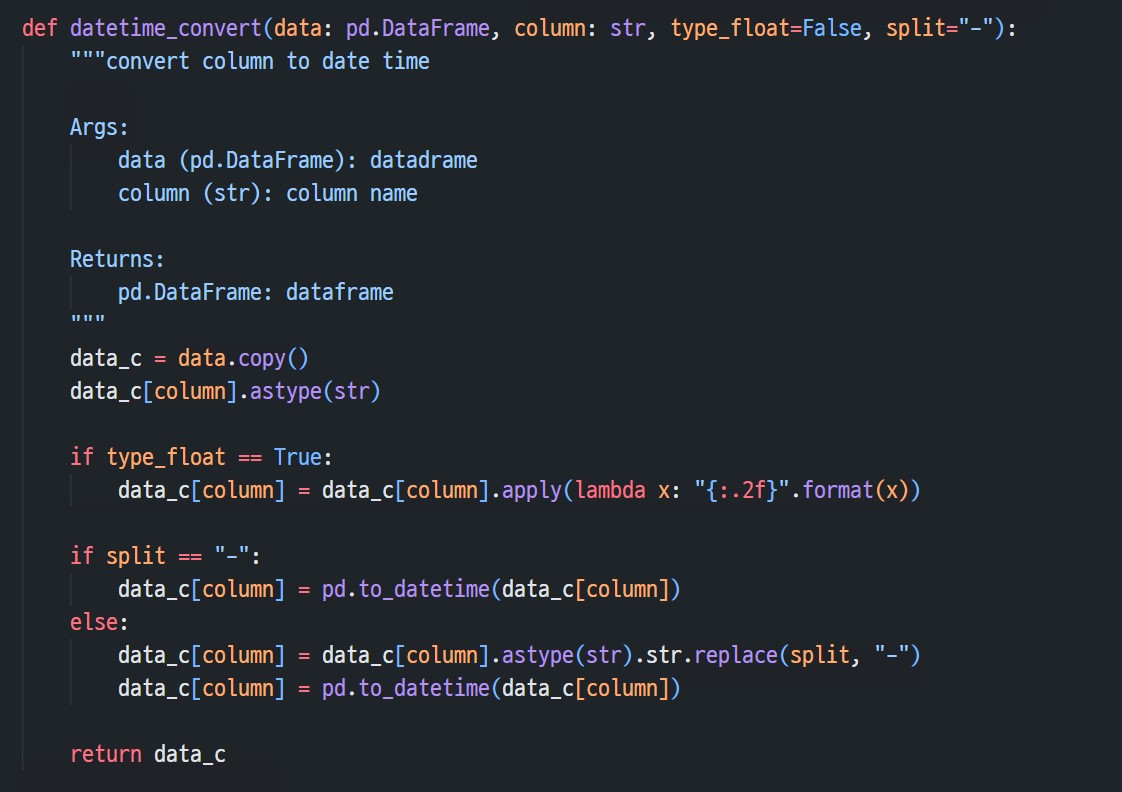


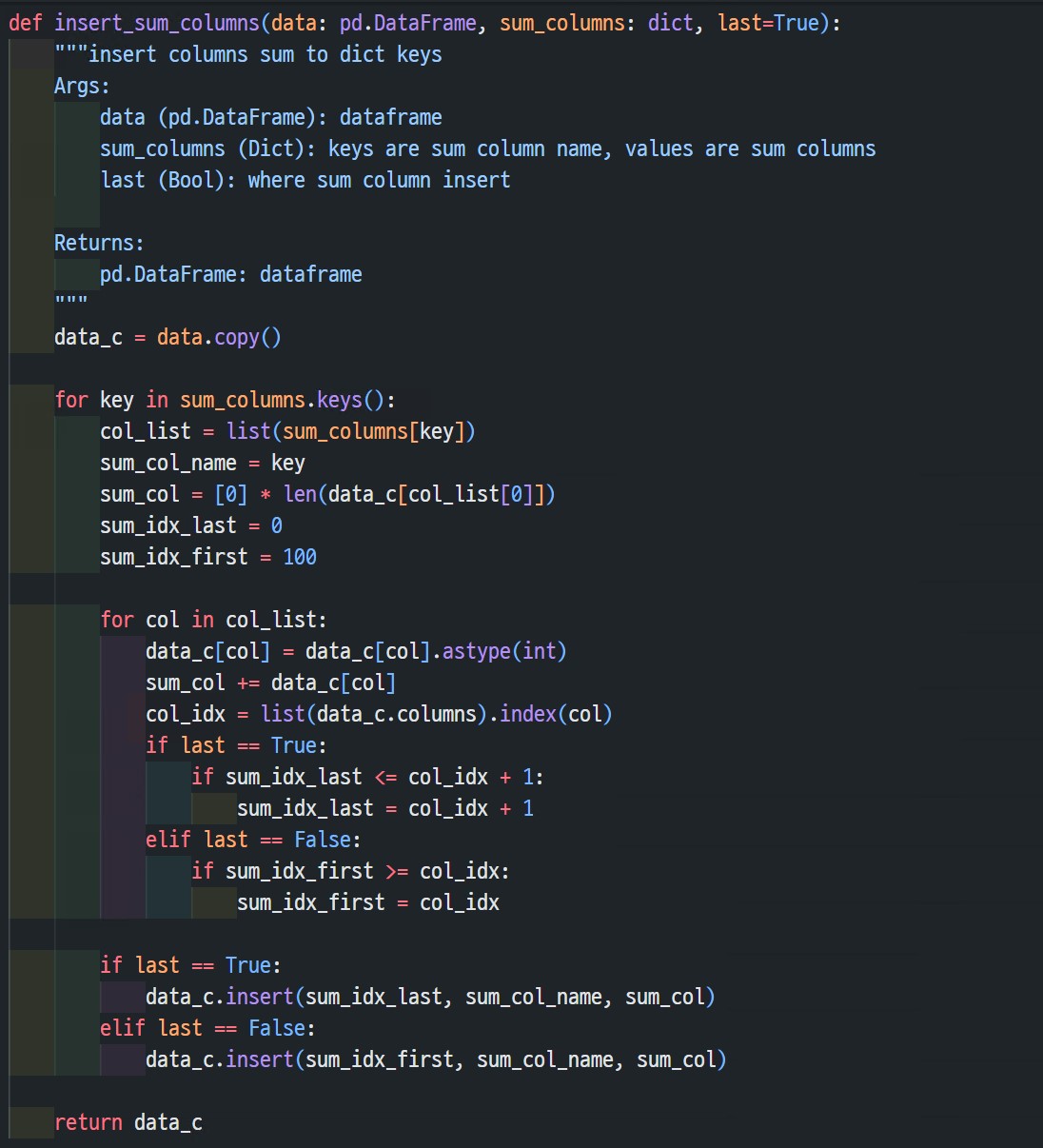
(보간법을 통한 결측치 대체)



(스케일링을 통한 데이터 통합)

* + - 용도별판매전력량
      * 1997년부터 2021년까지의 데이터가 있음.
      * 1995, 1996년의 데이터가 누락되어 있으며, 2015, 2016년의 데이터는 연도별 데이터만 존재함
      * 누락된 1995, 1996년의 데이터는 선형 보간법을 사용하여 채웠으며 2015, 2016년의 데이터는 2014, 2017년의 월별 분포를 참고하여 데이터를 월별로 분리함



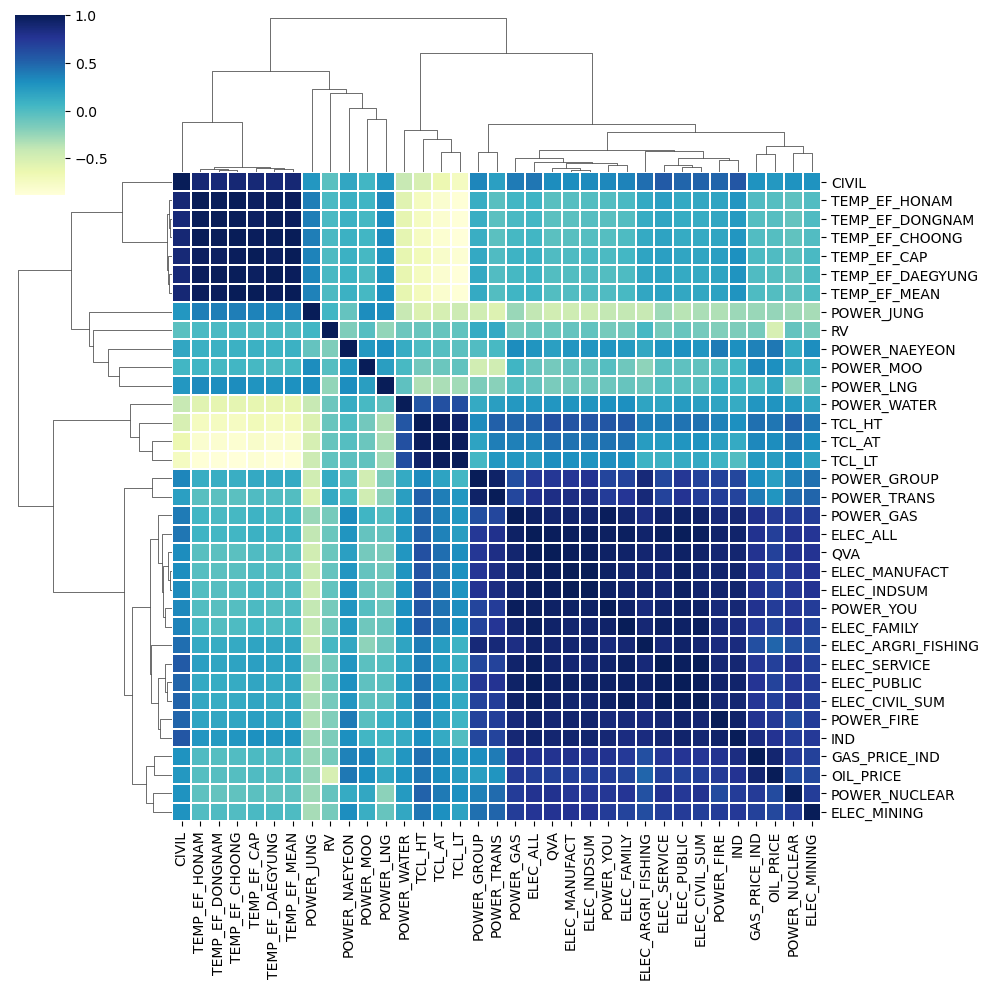


(산업용, 개인용 합계 구하는 함수)

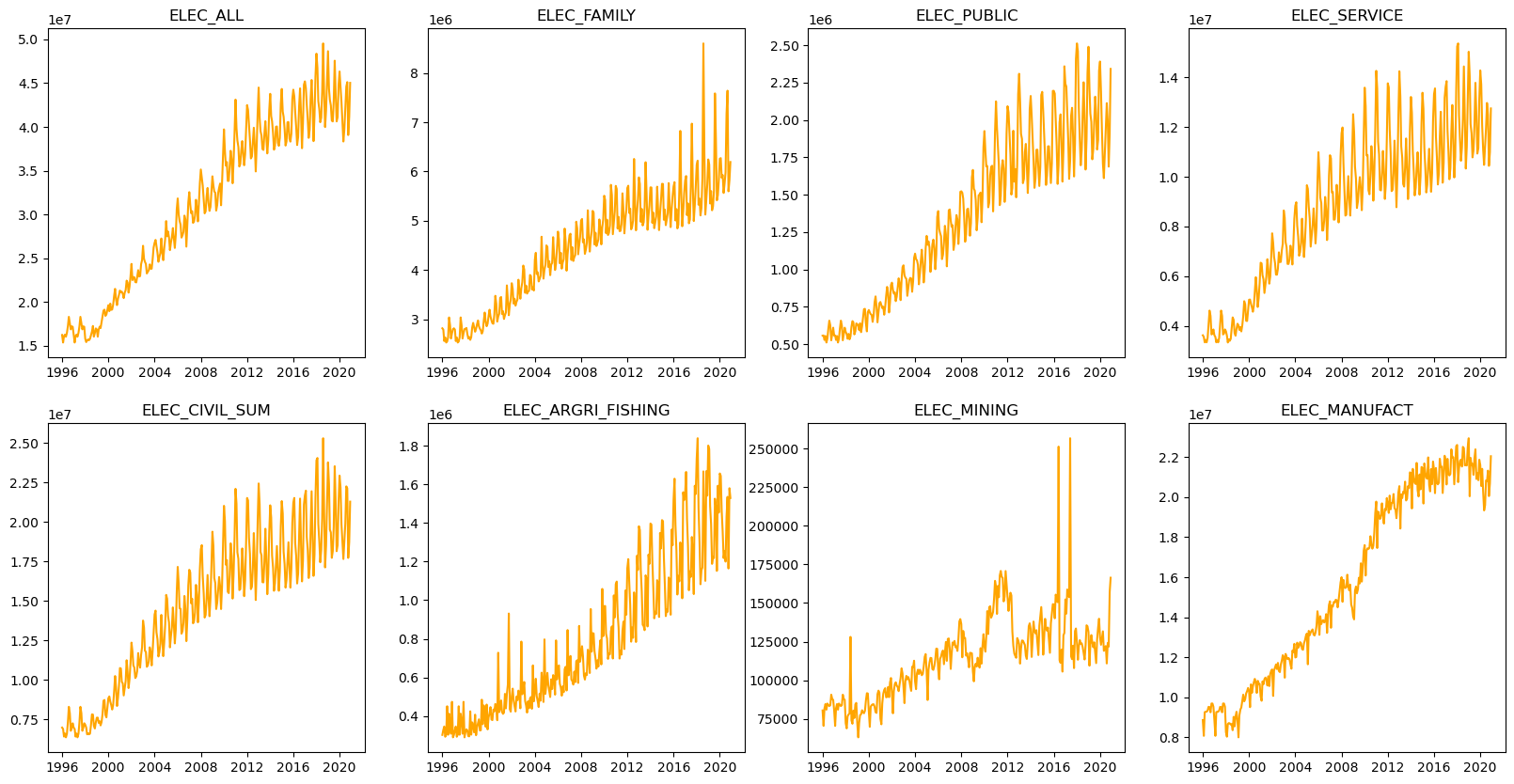


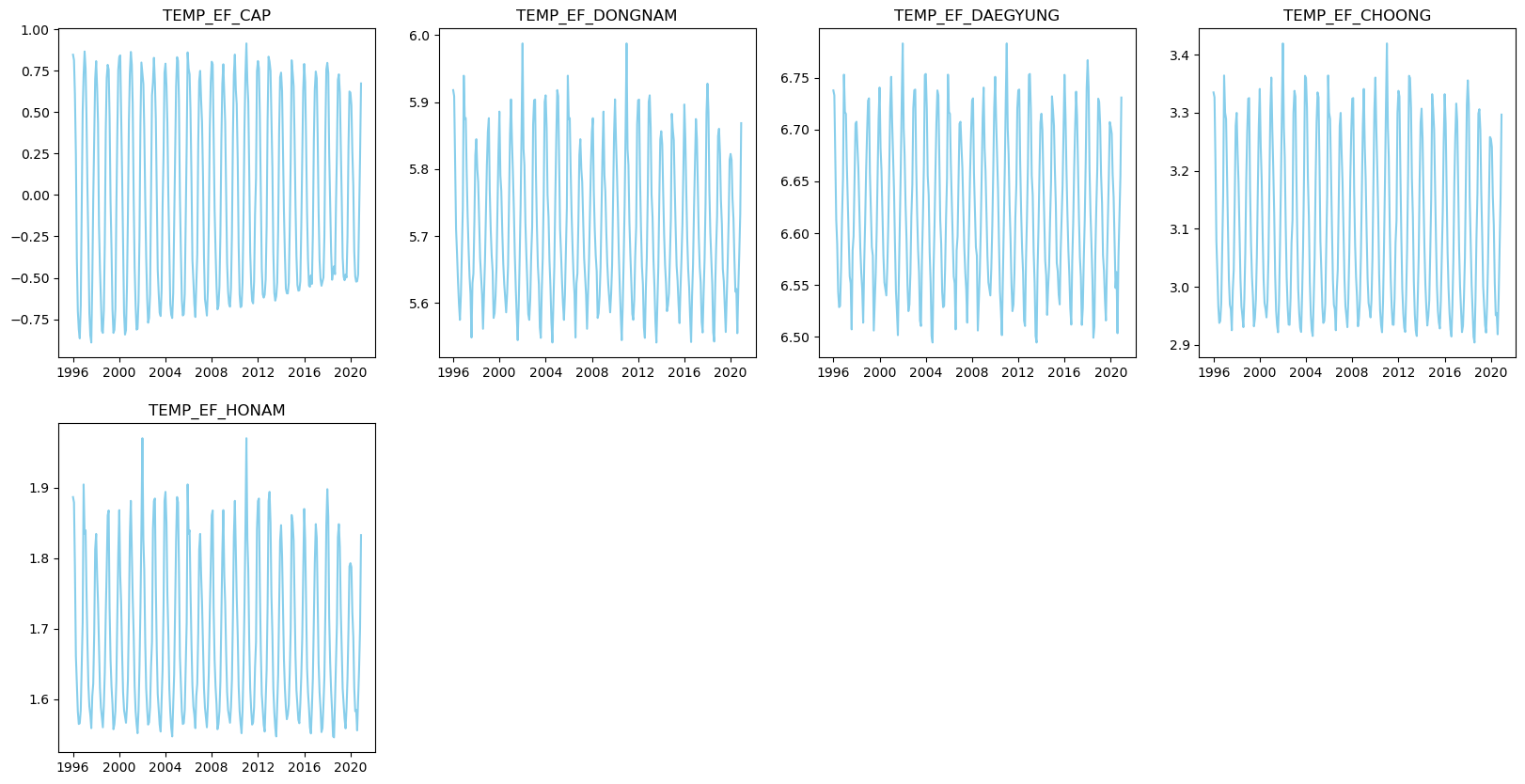
3. 탐색적 자료 분석 (Exploratory Data Analysis)

* 전체 데이터들 관계에서 타겟 값과의 상관관계와 다른 요소들 간에 상관관계를 보고 가장 적절한 요소들을 선택
* 전체적인 데이터들의 시계열 흐름을 살펴 이상치 등에 대한 접근 방식 설정

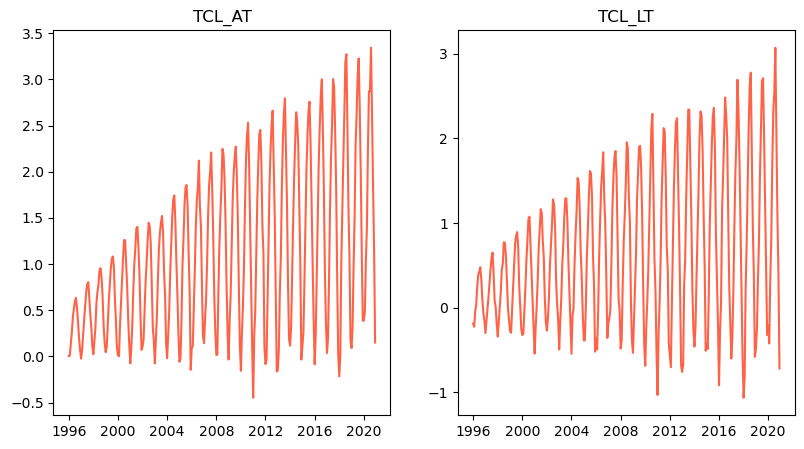


(변수 간 상관관계 EDA)



(발전 전력량 EDA)  
 

(기온효과에 대한 EDA)



(TCL에 대한 EDA)

4. 변수 선택 및 모델 구축 (Feature Engineering & Initial Modeling)

**4-1.변수 선택(Feature Engineering)**

1. 상관관계와 그래프 시각화를 통해 target과 높은 상관관계의 feature를 선정

2. 기온 관련 feature의 경우, 계절성을 대표하므로 산업용 target과 상관관계가 높지 않지만 민간용, 산업용 예측 모두 사용함.

* 산업용 가스수요 예측 feature

TEMP\_EF\_MEAN(기온효과\_평균),POWER\_FIRE(화력발전량),POWER\_GAS(가스발전량),ELEC\_ALL(전체 전력수요량), ELEC\_PUBLIC(공공용 전력수요량), ELEC\_MANUFACT(제조 전력수요량), ELEC\_INDSUM(산업용 전체 전력수요량)

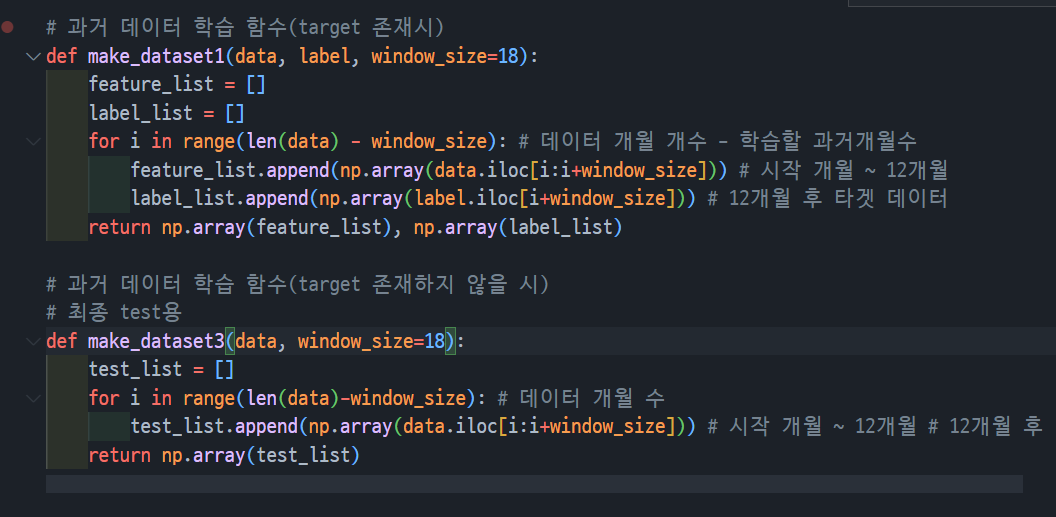
* 민수용 가스수요 예측 feature

TCL\_LT(기온최저),TCL\_HT(기온최대),TEMP\_EF\_CAP(수도권기온효과),TEMP\_EF\_MEAN(기온효과평균),POWER\_FIRE(화력발전),POWER\_GAS(가스발전),ELEC\_ALL(전체전력수요량),ELEC\_PUBLIC(공공전력수요량),ELEC\_CIVIL\_SUM(민간전력수요량),OIL\_PRICE(원유가격)

**4-2. 모델 선정 및 구축(Initial Modeling)**

1. 모델선정 : LSTM

lstm은 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망의 구조를 가지고 있고, 입력값을 현재와 과거에 유의성에 비교를 통해 가중치를 변화시키기 때문에 적절하다고 판단하여 사용하였다. 또한 미래예측은 먼 과거보다는 가까운 현재시점에 영향을 크게 받으므로 직전 18개월 학습 단위 차원을 추가로 생성하여 정확도를 높였다.Bidirectional lstm의 경우, 오히려 예측의 변동성을 과하게 감소시켜 사용하지 않았다.

****

****

(직전 18개월 과거학습 단위 차원 생성 & layer 구성)

1-1. 산업용 LSTM

- 학습 범위: 1996~2020년

- layer 및 epoch구성: input node=64, Dense=(64, 32, 16, 1), epoch=120

1-2. 민수용 LSTM

- 학습 범위: 1996~2020년

- layer 및 epoch 구성: node=32, epoch=120

2. prophet

2-1. target을 제외한 학습 feature들을 페이스북이 만든 prophet 시계열 예측모델을 사용하여 2034년까지 예측하였다.

2-2. Prophet 은 한층 더 진보적인 방법으로 트렌드와 주기적 특성뿐 아니라 예외적이고 이벤트와 같은 휴가철 상황까지도 모델링할 수 있도록 ARIMA 알고리즘을 확장한 것입니다. prophet은 facebook에서 만든 시계열 예측 라이브러리로 장점으로 사용법이 쉽다.

5. 모델 학습 및 검증 (Model Tuning & Evaluation)

1. LSTM

1-1. 산업용 LSTM

- 학습 범위: 1996~2020년

- 과거 학습 단위 차원 생성(WINDOW\_SIZE=18): 직전 18개월 단위 입력으로 다음 1개월 예측

- 과적합이 일어나지 않고, loss가 개선되지 않는 지점 탐색

- layer 및 epoch구성: input node=64, Dense=(64, 32, 16, 1), epoch=120



1-2. 민수용 LSTM

- 학습 범위: 1996~2020년

- 과거 학습 단위 차원 생성(WINDOW\_SIZE=18): 직전 18개월 단위 입력으로 다음 1개월 예측

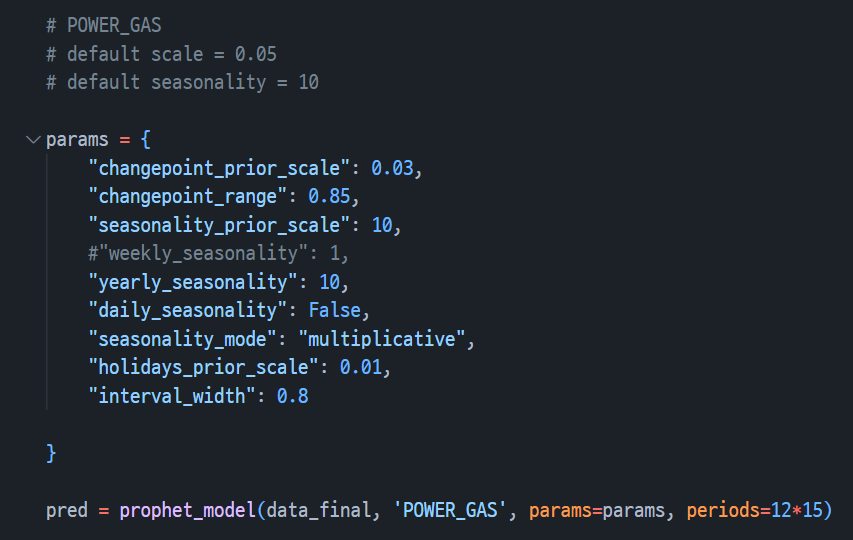
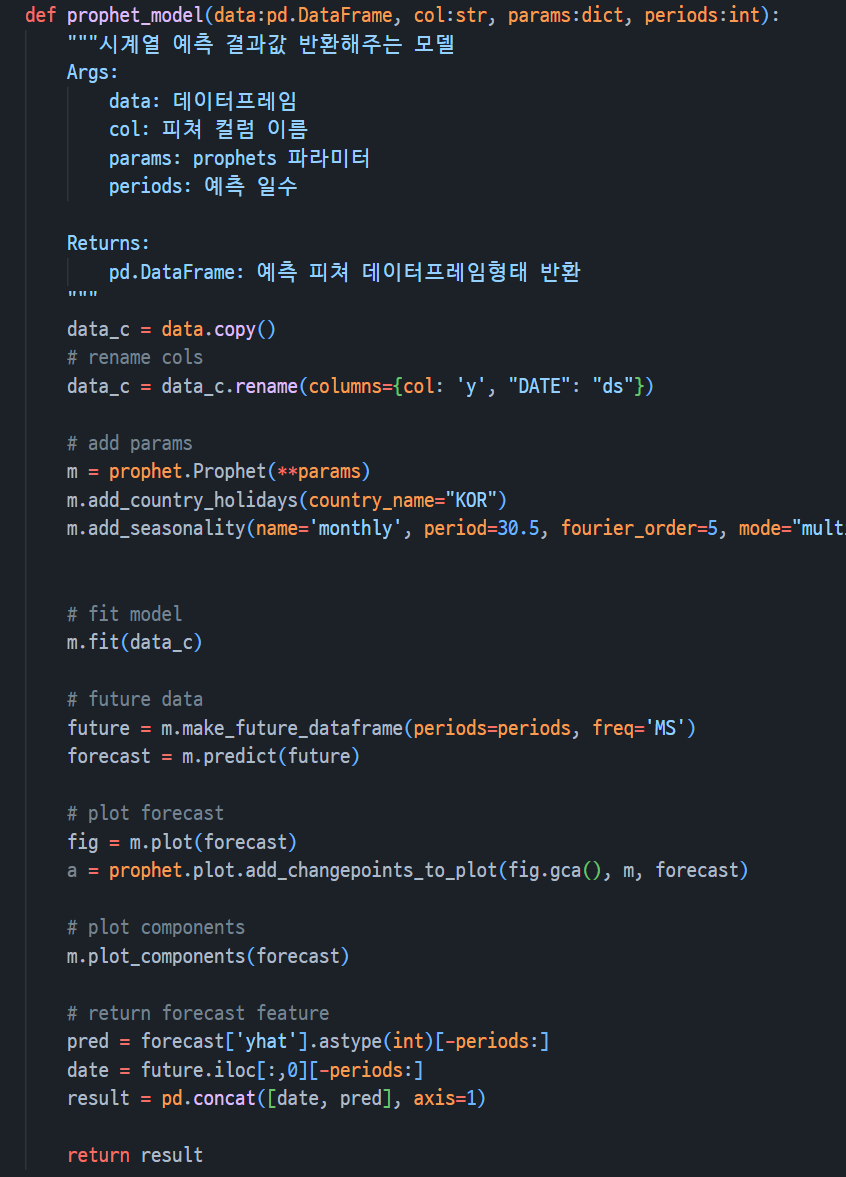
- 과적합이 일어나지 않고, loss가 개선되지 않는 지점 탐색

- layer 및 epoch 구성: node=32, epoch=120



2. prophet

2-1. . 파라미터의 경우, 학습데이터가 월별 데이터이므로 daily\_seasonality, weekly\_seasonality 특성 제외하여 예측하였다.



( prophet함수 생성 & 파라미터 설정)

3. 예측 방법(산업용,민수용 모델 2개 사용)

3-1. 학습 데이터: 1996~2020년 월별 feature들과 한국가스공사가 제공한 target

3-2. input 데이터: prophet을 이용한 2021~2034년 월별feature 예측값

3-3. 최종 예측 데이터: 2021~2034

6. 결과 및 결언 (Conclusion & Discussion)

* 추가 방향성
  + 데이터 늘리기
    - 과적합을 피하기 위해 피처들을 선별하고 선별하여 극히 적은 수의 피처만 사용했음 민간에서 구할 수 있는 데이터뿐만 아니라 가스공사 내의 추가 데이터, 국가 내부 데이터 등을 추가로 사용하여 모델을 개선할 수 있음
      * 환경 정책(저탄소 정책)등도 예측 가능한 피처로 변환할 수 있다면 모델 예측에 유용할 것
      * 천연가스 재고는 가격 형성에 중요한 역할 함. 5년 평균 이하의 재고 수준은 종종 천연 가스 가격 상승과 관련이 있는 반면, 5년 평균 재고 수준을 초과하는 재고 수준은 천연 가스 가격 하락과 관련이 있는 경우가 많음
      * 천연가스 수입의 대부분이 카타르, 호주, 미국이 50% 이상을 차지함 세 국가의 경제 상황에 영향을 미치는 피처를 통해서 천연가스의 수입 가격을 예측하면 예측에 유용한 피처로 사용할 수 있음
  + 가스 수요 예측을 위한 전용 모델을 개발
    - 지금까지의 모델은 기존에 존재하는 모델에서 파라미터 값을 조정했을 뿐 추가적인 레이어나 조정이 없다. 향후에는 기존에 사용한 모델을 기반으로 더 적합한 모델을 개발할 수 있음
  + 산업용 예측 모델에 대한 집중적인 개선(?)
    - 비교적 수요가 온도 효과에 집중된 민수용과 달리 산업용의 경우 수요량의 변동이 크므로 해당 관련 모델에 집중하면 더 좋은 결과를 가져갈 수 있음