



Project 2 : 전력사용량 예측

5공대

송현주

오서연

이승학

이용규

이태현

목차

- 01** 시계열 데이터란
- 02** 데이터 소개
- 03** EDA
- 04** 모델링
- 05** 평가
- 06** 발전 방향성

01

시계열 데이터란

시계열 데이터란?

일정한 시간 동안 수집된 일련의 **순차적으로** 정해진 데이터 셋의 집합
데이터에서 법칙성을 발견해 이를 모형화, 추정된 모형을 통하여 **미래의 값을 예측**

01

시계열 데이터란

추세변동

- 시계열의 장기간에 걸친 점진적이고 지속적인 변화 상태(상승 경향, 하강 경향)

계절변동

- 분기별, 월별 자료에서 기후 등과 같은 자연의 조건, 사회적 관습, 혹은 제도 등의 영향 받아서 계절적인 차이를 나타낸 것
- 주기적인 패턴을 갖고 반복적으로 나타나는 주기변동

순환변동

- 수년간의 간격을 두고 상승과 하락이 주기적으로 나타나는 변동
- 계절변동으로 설명되지 않는 장기적인 주기변동

불규칙변동

- 사전적으로 예상할 수 없는 특수한 사건에 의해 야기 되는 변동
- 명확히 설명될 수 없는 요인에 의해 발생되는 우연변동

02

데이터 소개

num(건물)	date_time	기온(°C)	...	태양광 보유	전력사용량(kWh)
1	2020-06-01 00	17.6	...	X	8179.056
1	2020-06-01 01	17.7	...	X	8135.64
1	2020-06-01 02	17.5	...	X	8107.128
2	2020-06-01 00	17.1	...	X	977.184

122400 rows x 10 columns

02

데이터 소개

왜 전력사용량을 예측해야 하는가?

공급의 안정화 및 건물 운영 예산을 측정하는데 도움

03

EDA

독립 변수

종속 변수

num

date_time

기온(°C)

풍속(m/s)

습도(%)

전력사용량
(kWh)

강수량(mm)

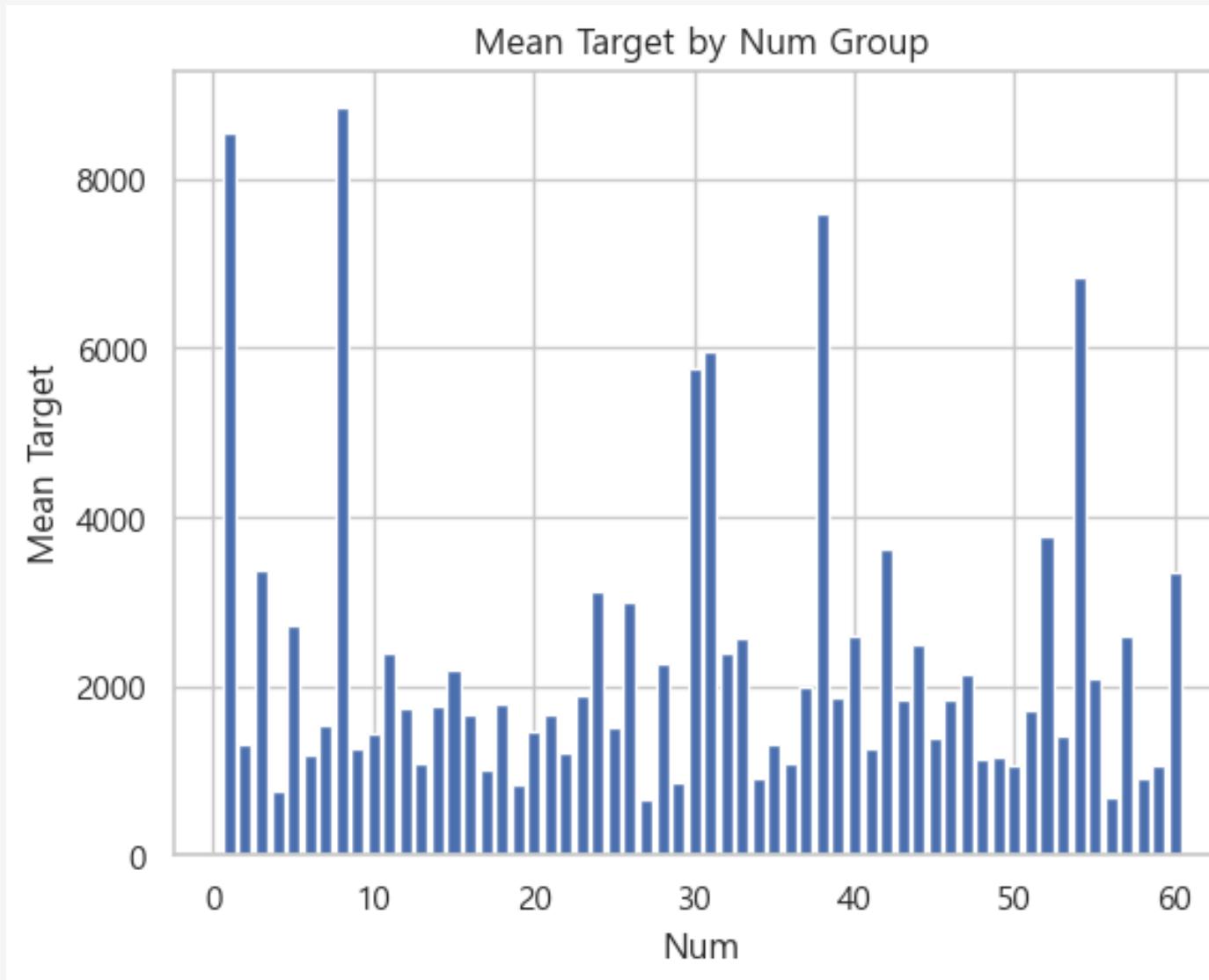
일조(hr)

비전기냉방설비운영

태양광 보유

03

EDA - 독립 변수 설명



건물 번호

✓ df["date_time"].head(10) ...

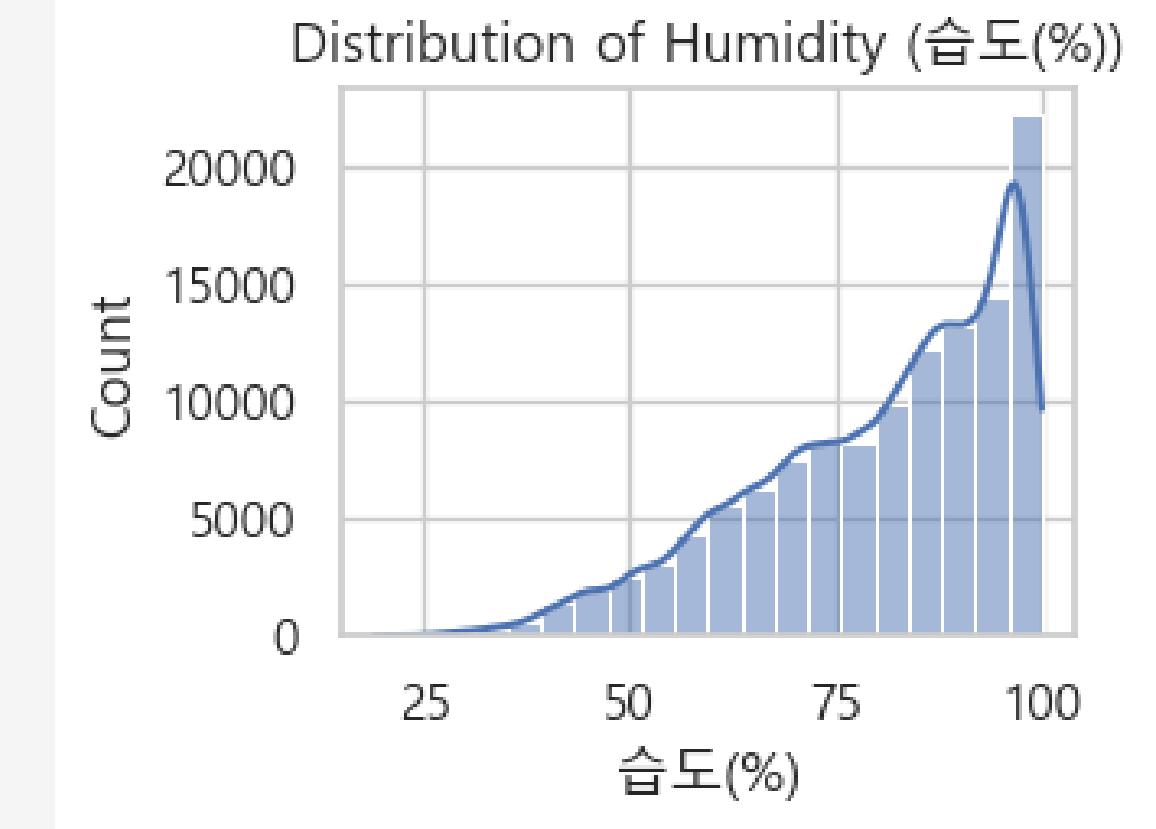
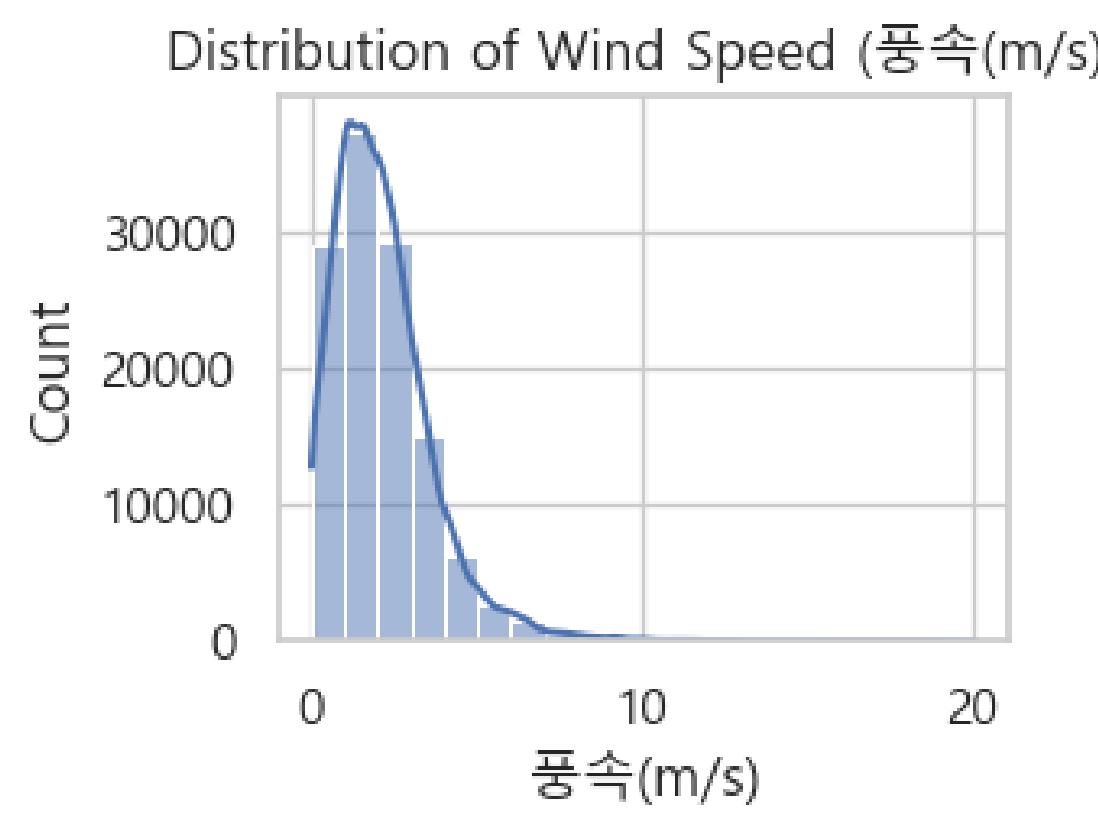
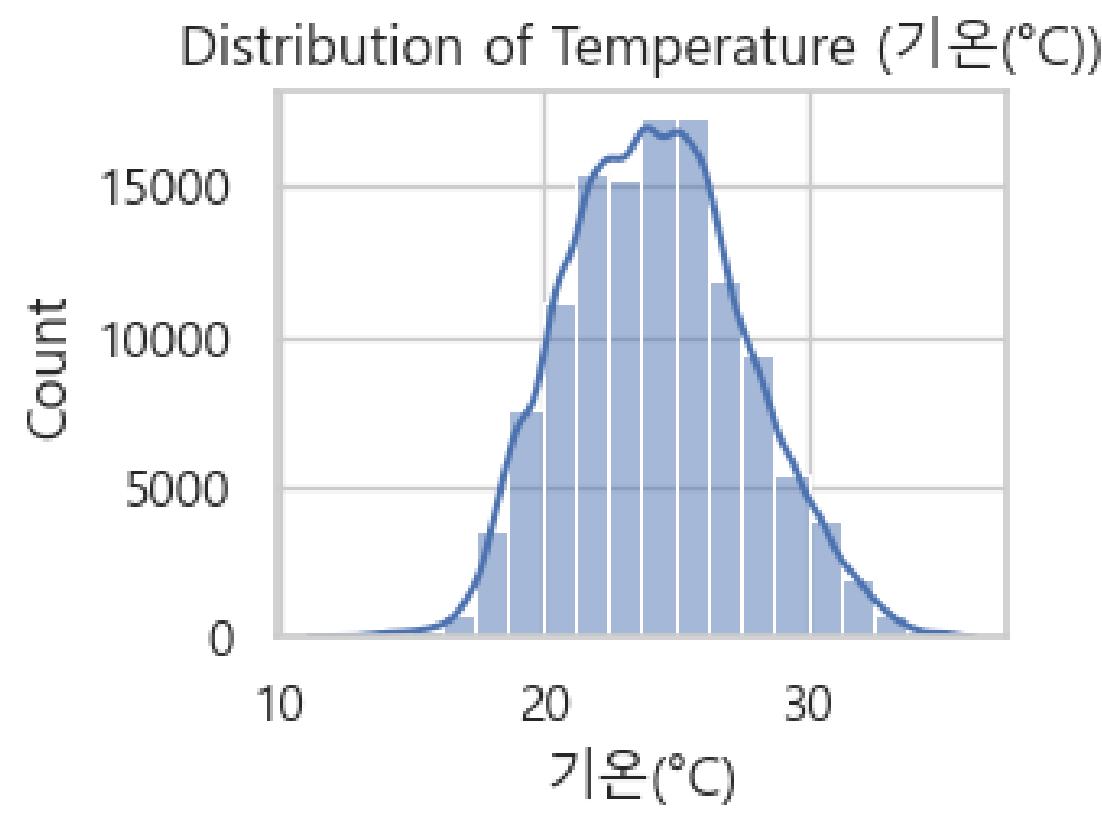
0	2020-06-01 00
1	2020-06-01 01
2	2020-06-01 02
3	2020-06-01 03
4	2020-06-01 04
5	2020-06-01 05
6	2020-06-01 06
7	2020-06-01 07
8	2020-06-01 08
9	2020-06-01 09

Name: date_time, dtype: object

날짜와 시간

03

EDA - 독립 변수 설명

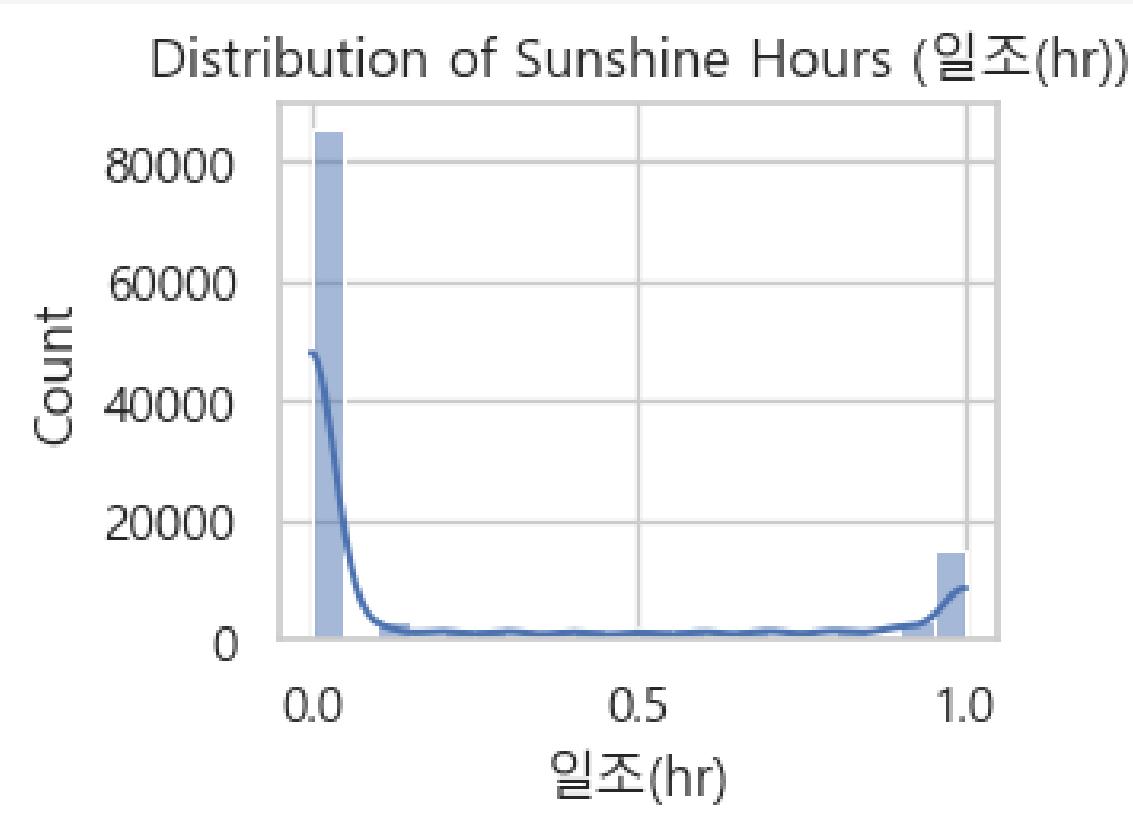


약 15°C ~ 30°C 분포
평균 기온 약 25°C

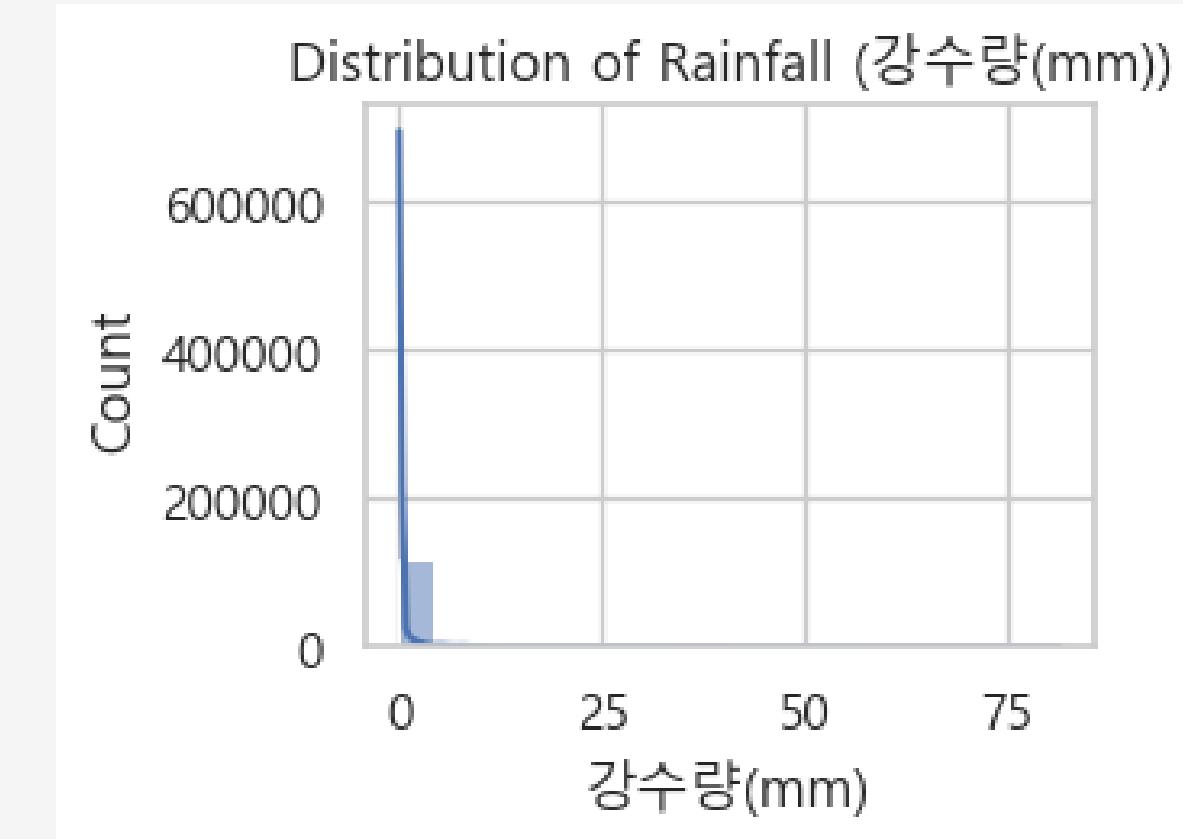
대부분의 풍속 0 ~ 5 m/s 분포

습도 80~100%에 데이터가 몰려 있음
이는 습한 환경이 자주 나타남을 의미

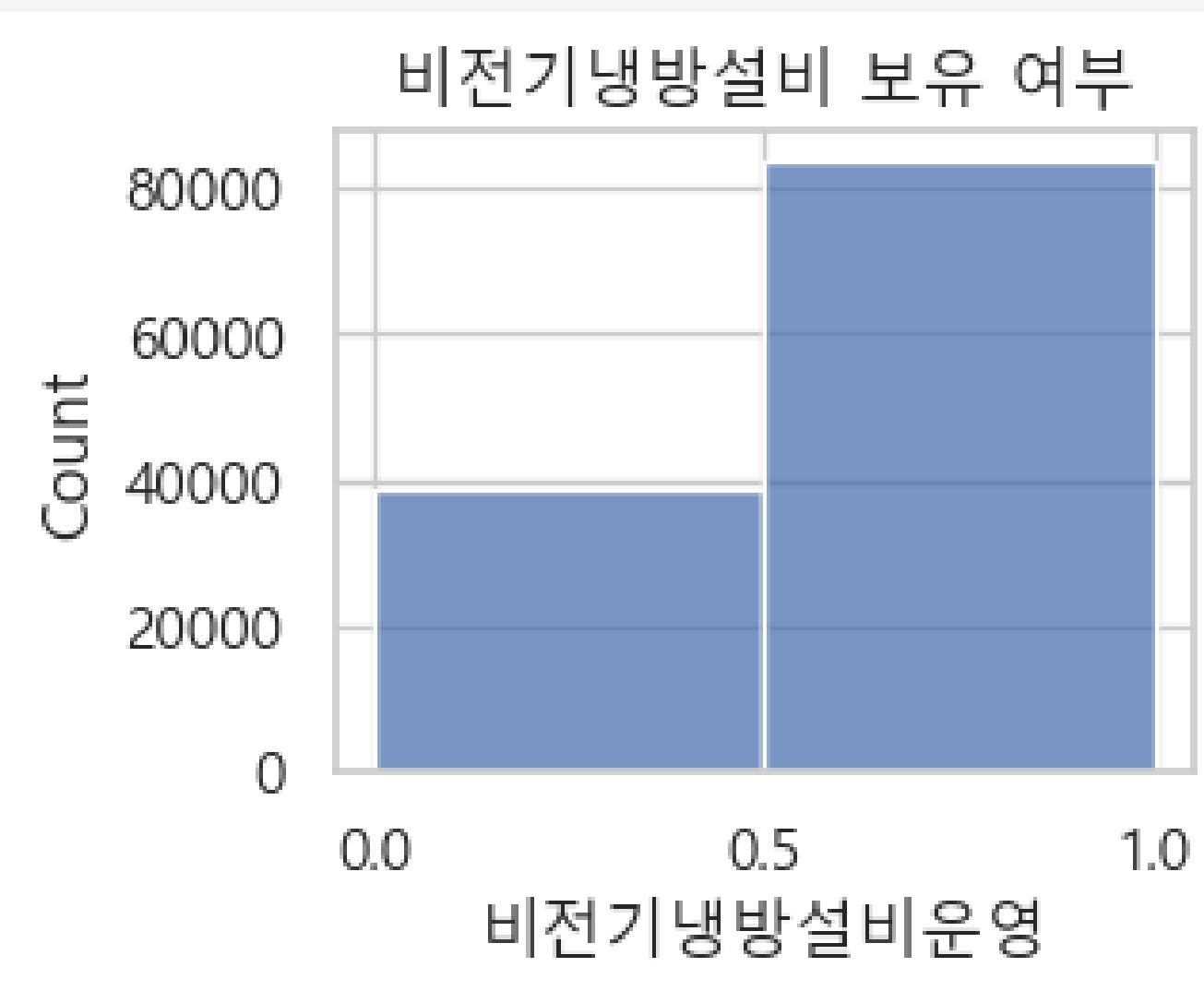
EDA - 독립 변수 설명



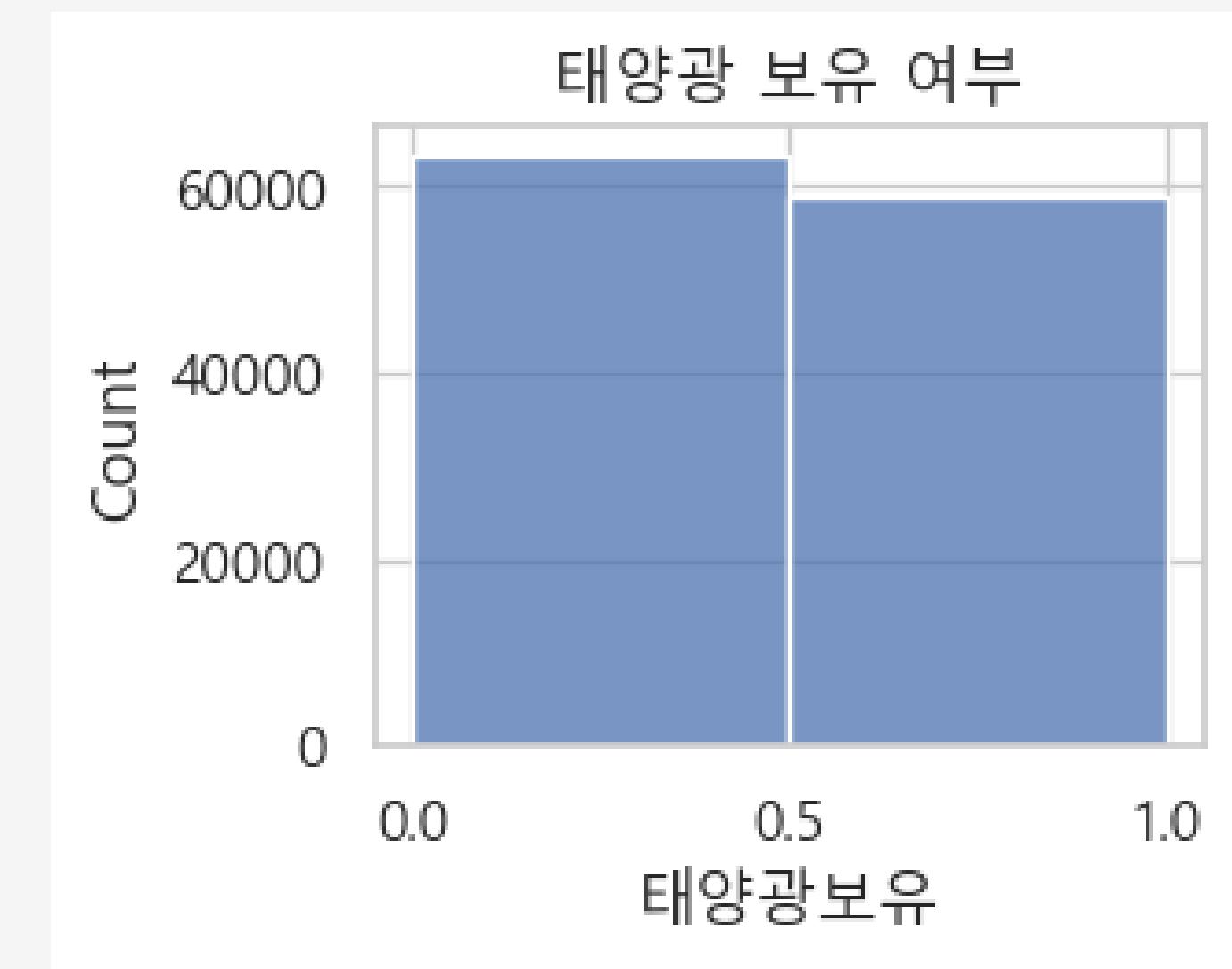
대부분 0시간 근처에 몰려 있음



대부분의 값이 0에 매우 가까움



비전기냉방설비운영을 가진 건물이 약 2배 많음



태양광 보유 여부/미여부 수 비슷함

03

EDA - 독립 변수 설명

비전기 냉방설비에 대하여

공공기관 에너지이용 합리화 추진에 관한 규정

제10조(에너지 수급 안정 및 효율 향상을 위한 전력수요관리시설 설치)

① 각 공공기관에서 연면적 1,000㎡ 이상의 건축물을 신축하거나 연면적 1,000㎡ 이상을 증축하는 경우 또는 냉방설비를 전면 개체할 경우에는 **냉방설비용량의 60%이상**을 심야전기를 이용한 축냉식, 도시가스를 이용한 냉방방식, 집단에너지사업 허가를 받은 자로부터 공급되는 집단에너지를 이용한 지역냉방방식, 소형 열병합발전을 이용한 냉방방식, 신.재생에너지를 이용한 냉방방식 등 **전기를 사용하지 아니한 냉방방식으로 냉방설비를 설치하여야 하며, 냉방설비를 증설 또는 부분 개체할 경우에는 전기를 사용하지 아니한 냉방방식의 냉방설비용량이 전체의 60% 이상이 되도록 유지하여야 한다.** 다만, 다음 각 호에 해당하는 경우는 제외한다.

1. 도시철도법에 의해 설치하는 지하철역사
2. 냉방공간의 연면적 합계가 500㎡ 미만인 경우
 도시가스 미공급 지역에 건축하는 시설 중 연면적 3,000㎡ 미만인 경우
 건축법 시행령」별표 1의 제2호에 따른 공동주택
 건축법 시행령」별표 1의 제23호 라목에 따른 국방.군사시설 중 병영생활관, 간부숙소
 「공공주택특별법 시행령」제4조에 따른 공공준주택
 그 밖에 산업통상자원부장관이 인정하는 경우

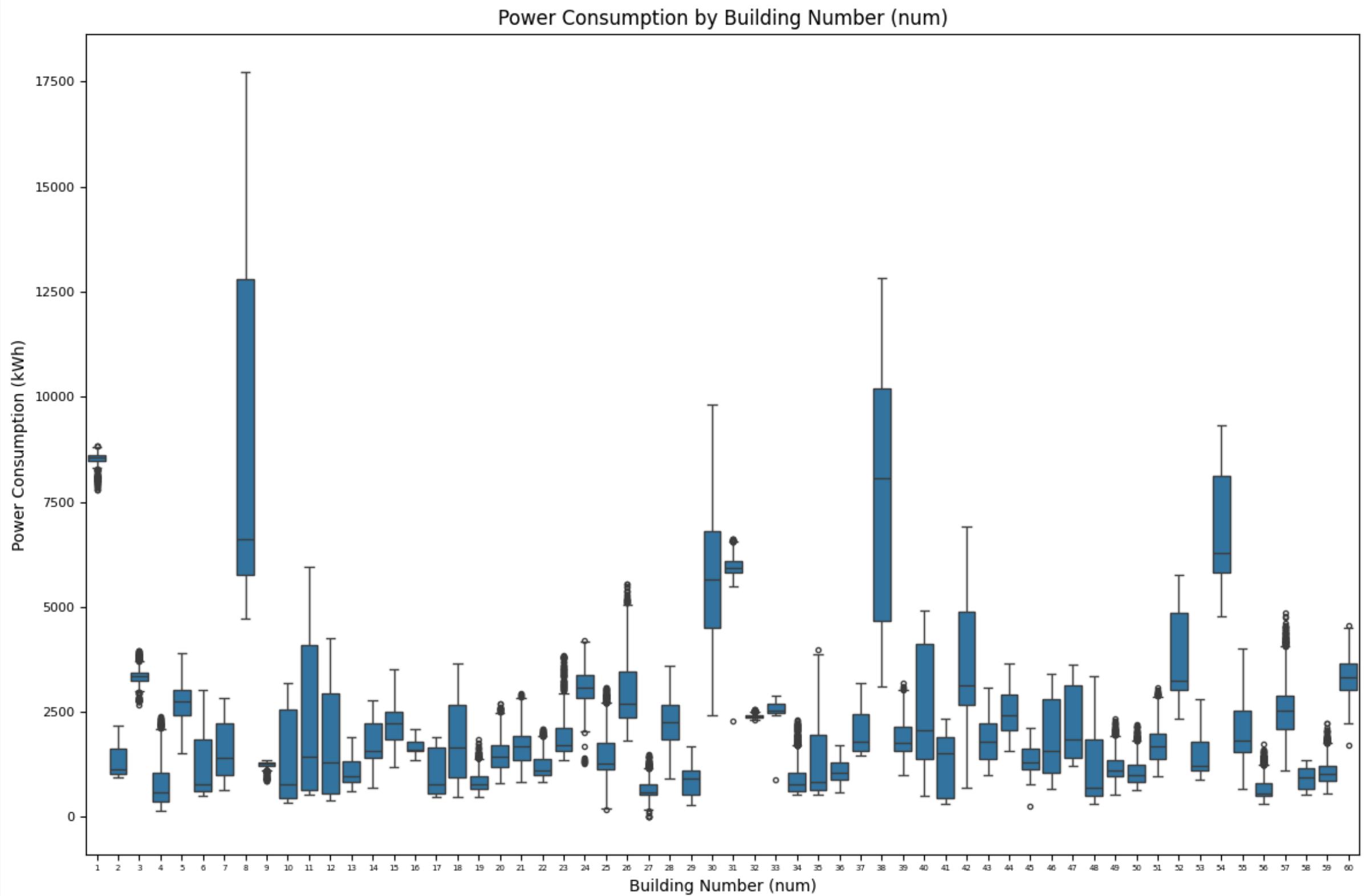
② 제1항에도 불구하고 수직 수평 증축의 경우, 기존 건축물의 전기를 사용하지 아니한 냉방방식의 냉방설비용량이 수직.수평 증축되는 연면적을 포함하더라도 전체의 60% 이상이 될 경우에는 제1항을 적용하지 아니할 수 있다.

비전기 냉방설비가 있으면 냉방에 필요한 전력 소비가 감소

- 2011년부터 일정 규모 이상의 공공기관과 민간 건축물에 대해 **비전기식 냉방 설비 설치가 의무화**
- 연면적 300평(1,000㎡) 이상의 건축물을 신축하거나 증축할 경우, 냉방 설비 용량의 60% 이상을 비전기식(예: 가스냉방, 축냉식 등)으로 설치해야 한다는 규정 적용
- 이 규정은 주로 여름철 전력 피크 수요를 완화하기 위한 조치로, 전력 의존도를 줄이고 에너지 효율을 높이기 위해 도입

전력 사용량 (kWh)

1 kWh는 1시간 동안
1kW의 전력을 소비했을 때
사용되는 전력

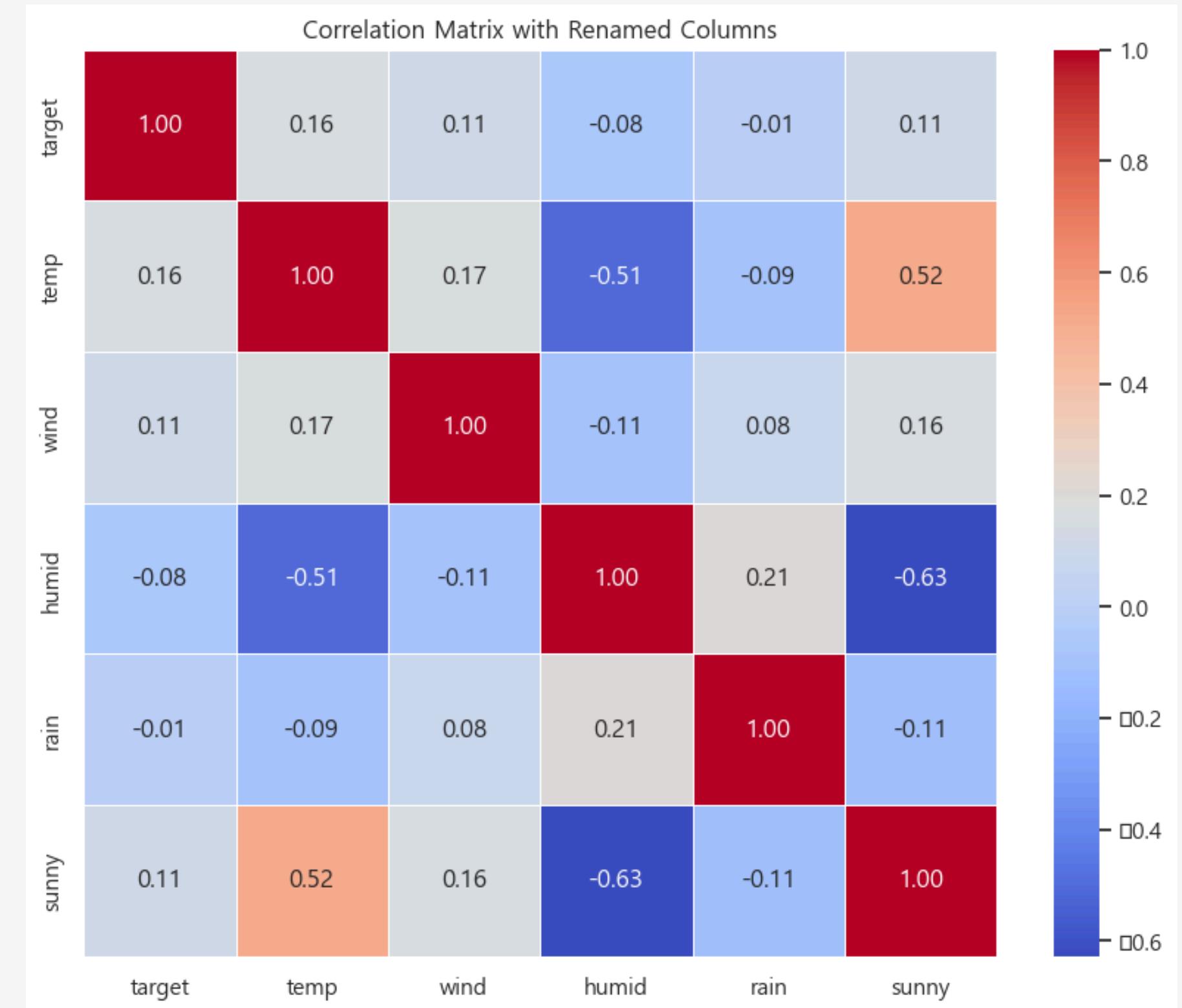


03

EDA

상관행렬

상관 행렬을 보면, 기후 변수들과 전력량 사이의 상관관계가 대체로 낫거나 약한 편이라는 것을 알 수 있음

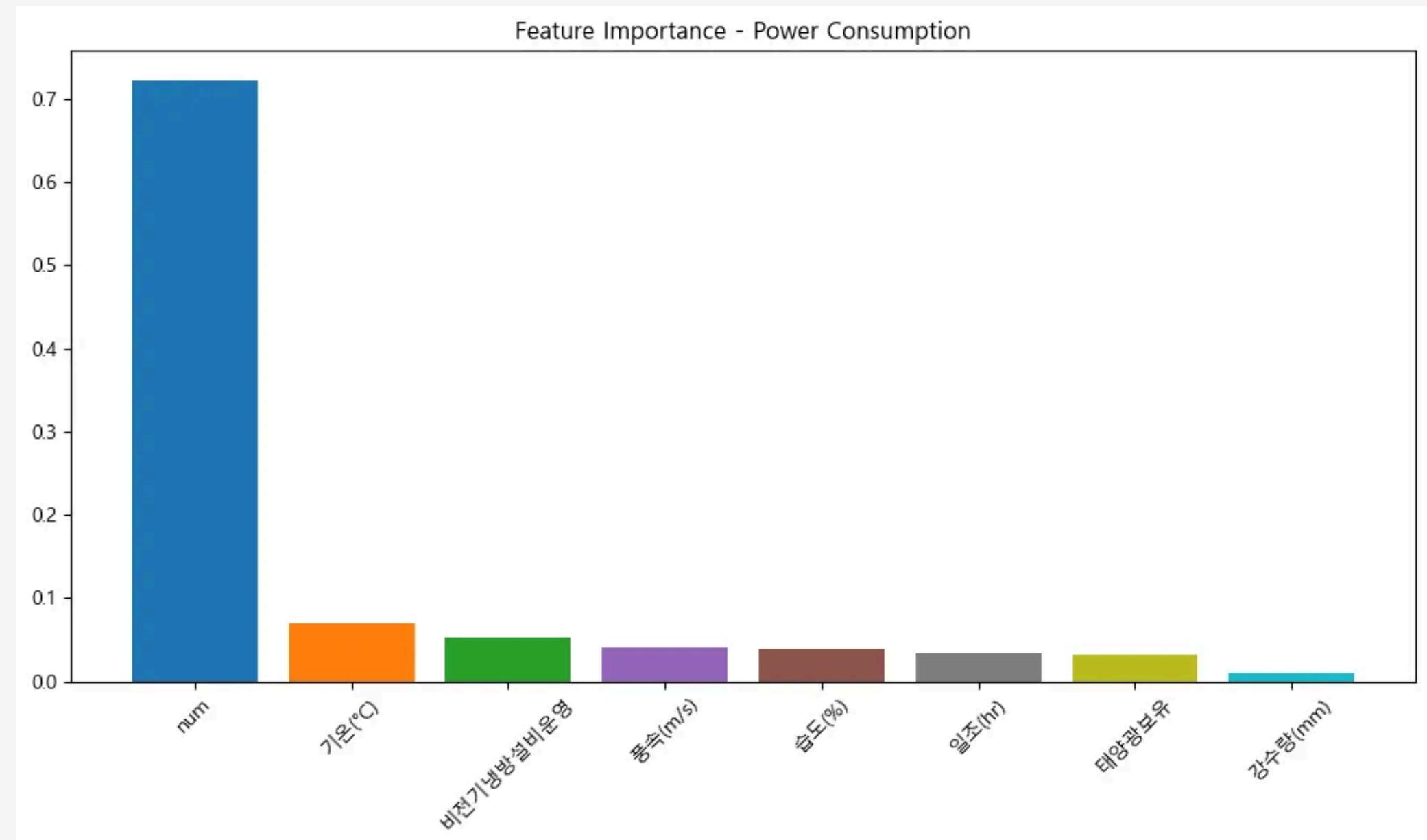


03

EDA

전력 사용량과 다른 독립변수의
Feature Importance 확인
이때, **건물 번호**가 가장 중요한 변수!

1. **Feature 'num' Importance: 0.721**
2. **Feature '기온(°C)' Importance: 0.071**
3. **Feature '비전기냉방설비운영' Importance: 0.053**
4. Feature '풍속(m/s)' Importance: 0.041
5. Feature '습도(%)' Importance: 0.038
6. Feature '일조(hr)' Importance: 0.034
7. Feature '태양광보유' Importance: 0.031
8. Feature '강수량(mm)' Importance: 0.010



03

EDA

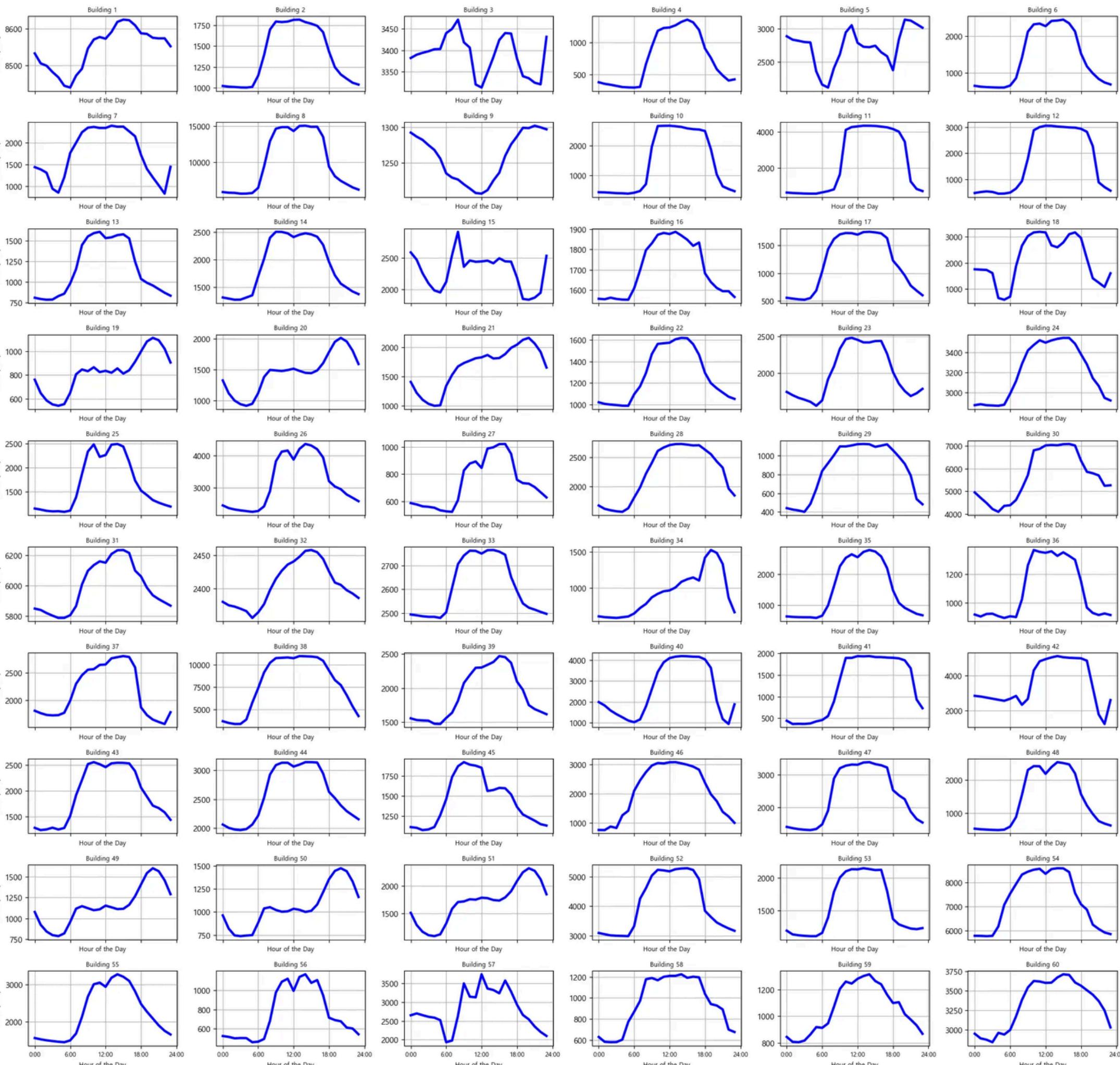
(6 ~ 8월) 평일 평균 전력 사용량

건물마다 어떤 특징이 있는지 확인하기 위해
평일 평균 그래프를 시각화

비슷한 패턴을 보이는 그래프가 보여서 수동으로 그룹화를
진행하여 비교 및 분석 진행

※ 사람이 직접 분류를 진행하였기에, 그룹 1과 그룹 2가 정확히 구분되지 않은
그래프가 있을 수 있다는 점 양해 부탁 드립니다.

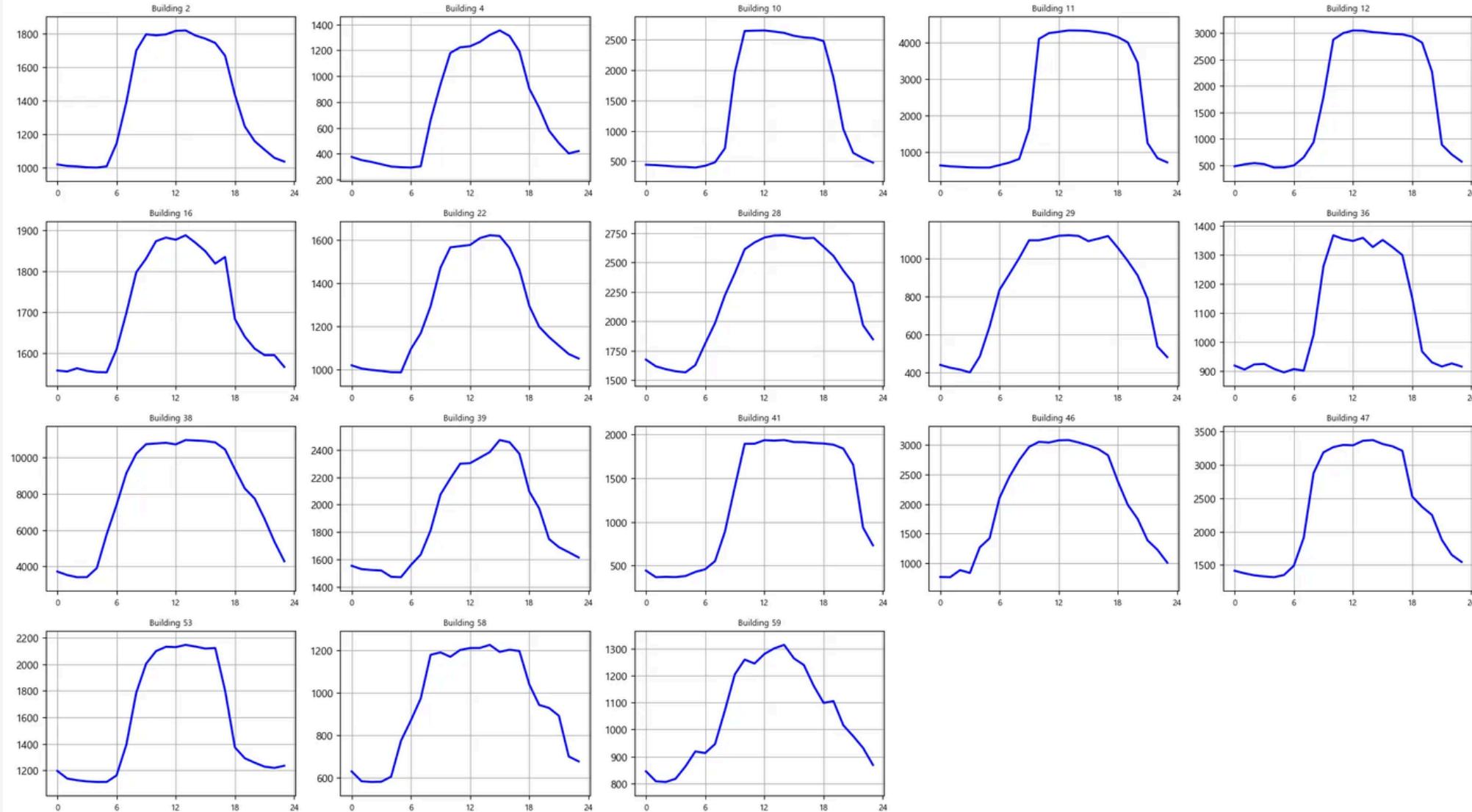
Weekday Hourly Average Power Consumption by Building



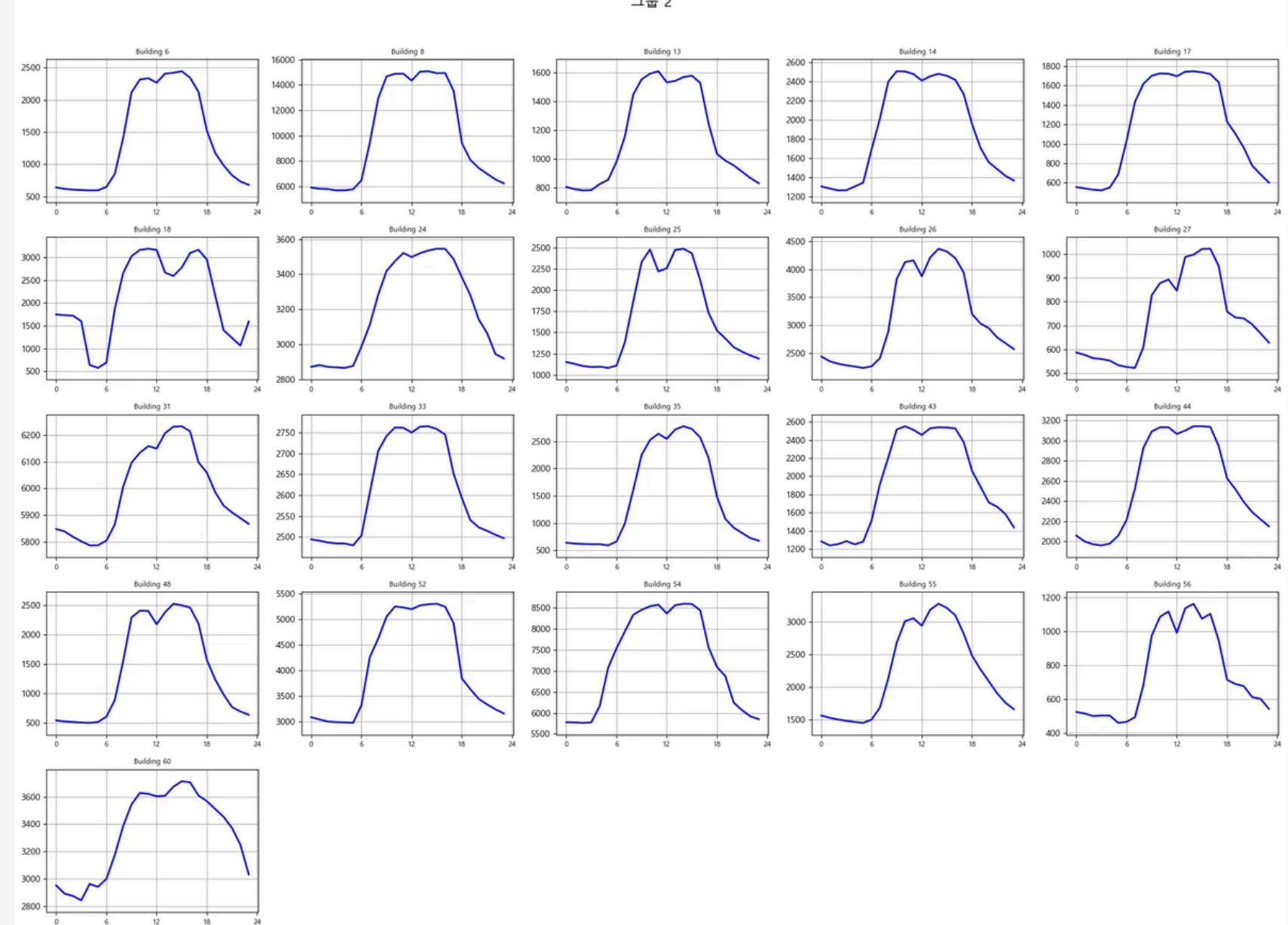
03

EDA

그룹 1



그룹 2



그룹 1

- 6~18시 까지 전력 사용이 집중되는 그래프 형태

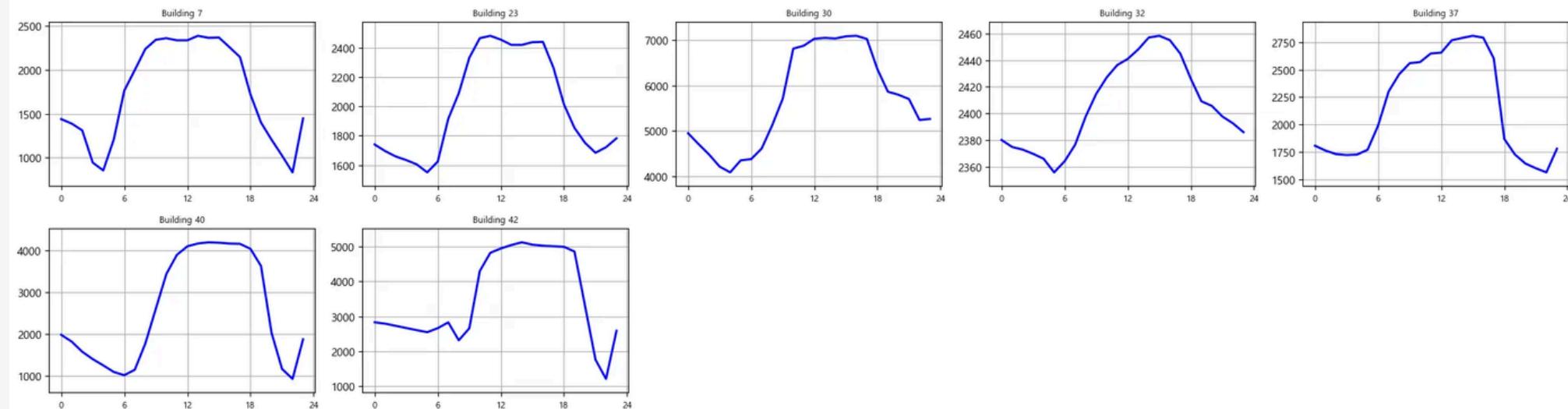
그룹 2

- 6~18시 까지 전력 사용이 집중되는 그래프 형태
- 12시에 잠깐 전력 사용량이 하락하는 특징

03

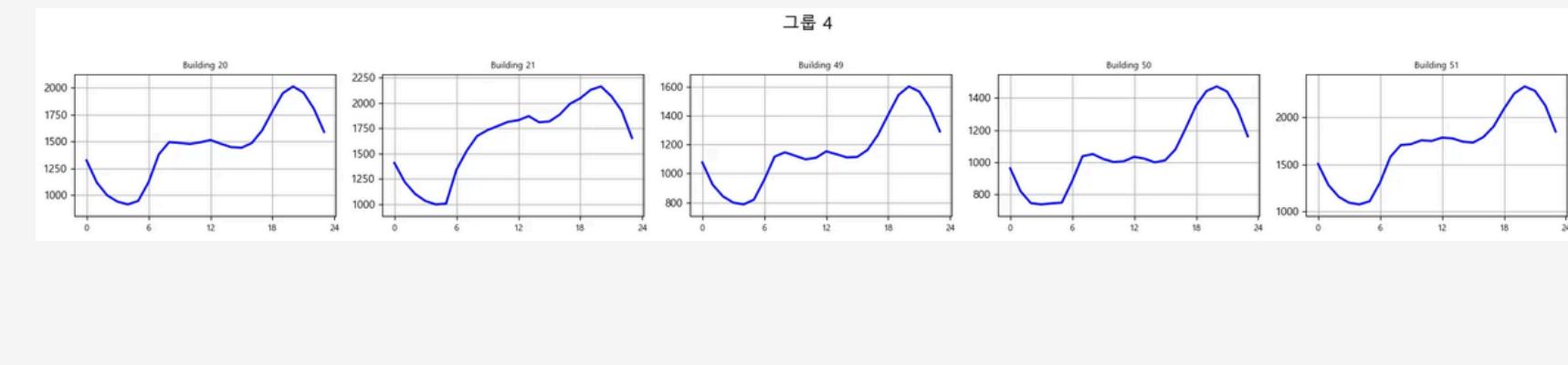
EDA

그룹 3



그룹 3

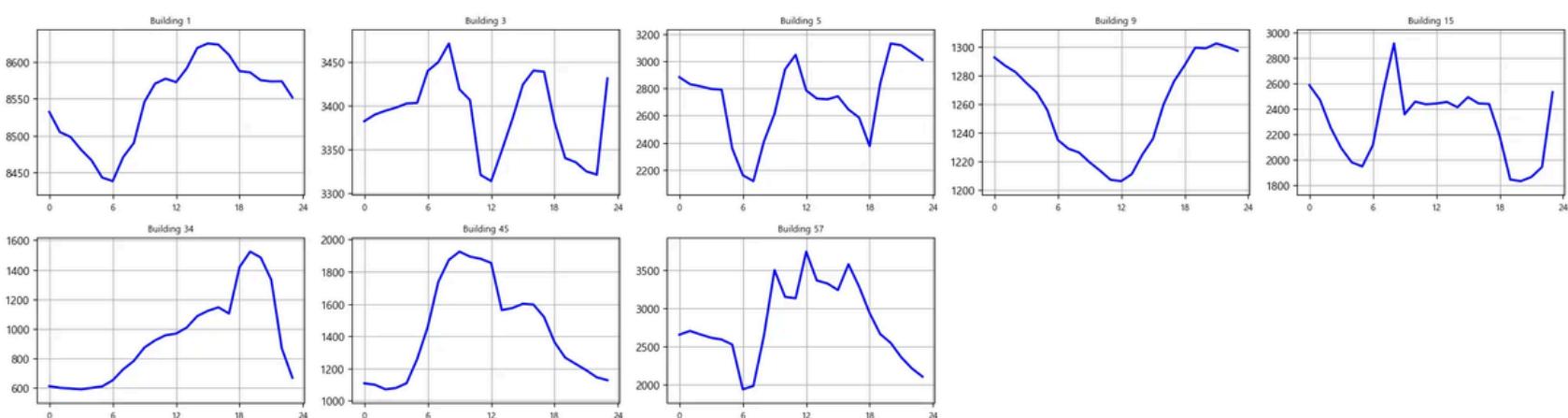
- 6~18시 까지 전력 사용이 집중되는 그래프 형태
- 새벽 시간에 전력 사용량이 다시 증가하는 특징



그룹 4

- 0~6시 감소, 6~18시 유지, 18시에 증가했다가 감소하는 그래프 형태

그룹 5



그룹 5

- 그룹 1~4에 속하지 않는 패턴의 그래프

03

EDA

비슷한 패턴을 보이는 산업 찾기

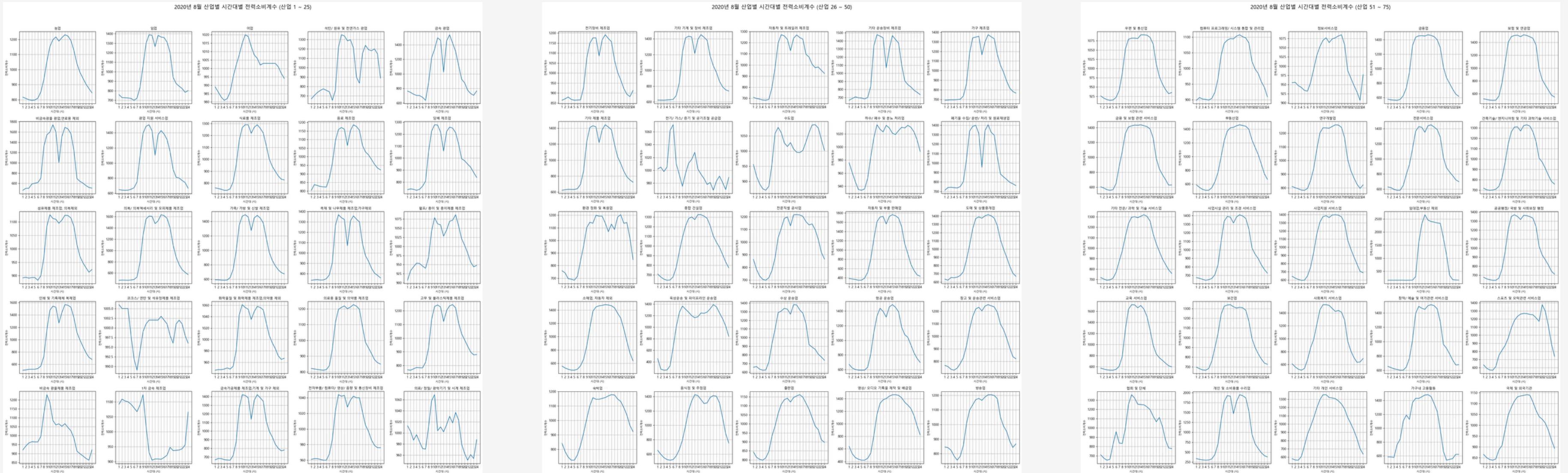
[통계청]

산업별(표준산업코드 중분류) 월별 1~24시 전력소비계수

다양한 산업의 전력소비를 확인하기 위해 2020년 8월 기준
75개 산업군의 시간대별 전력소비계수를 시각화

03

EDA



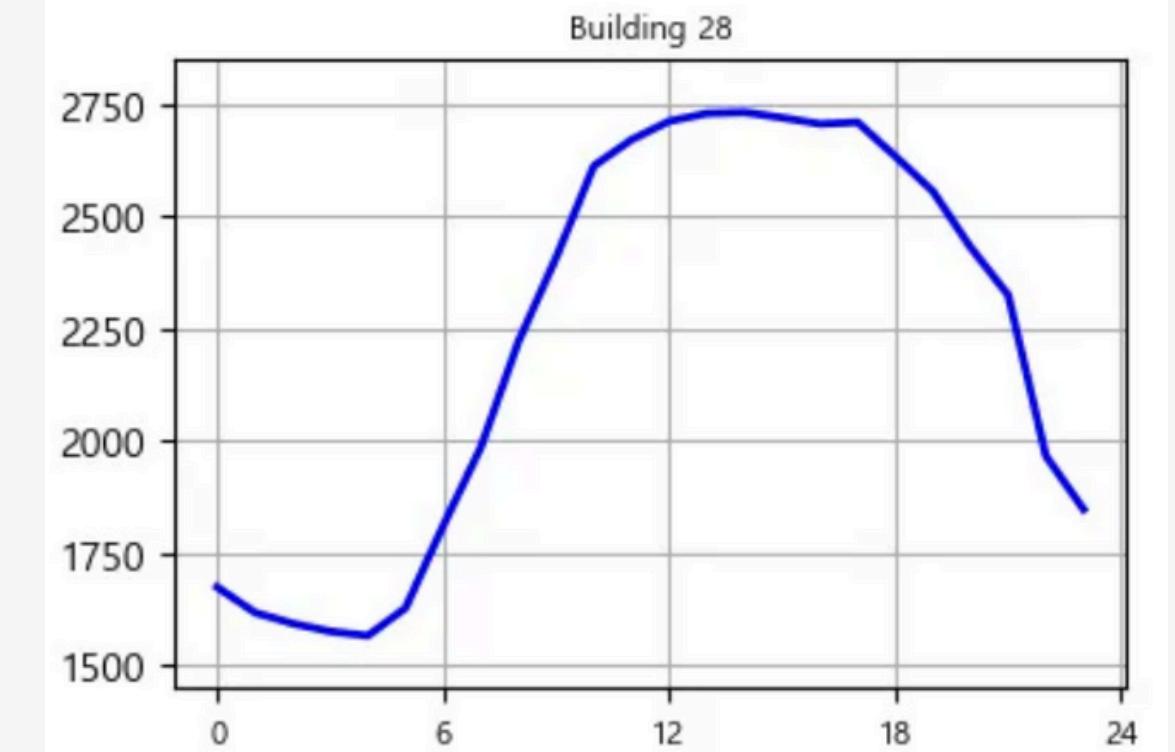
산업별 시간대별 전력소비계수

03

EDA

그룹 1과 비슷한 그래프

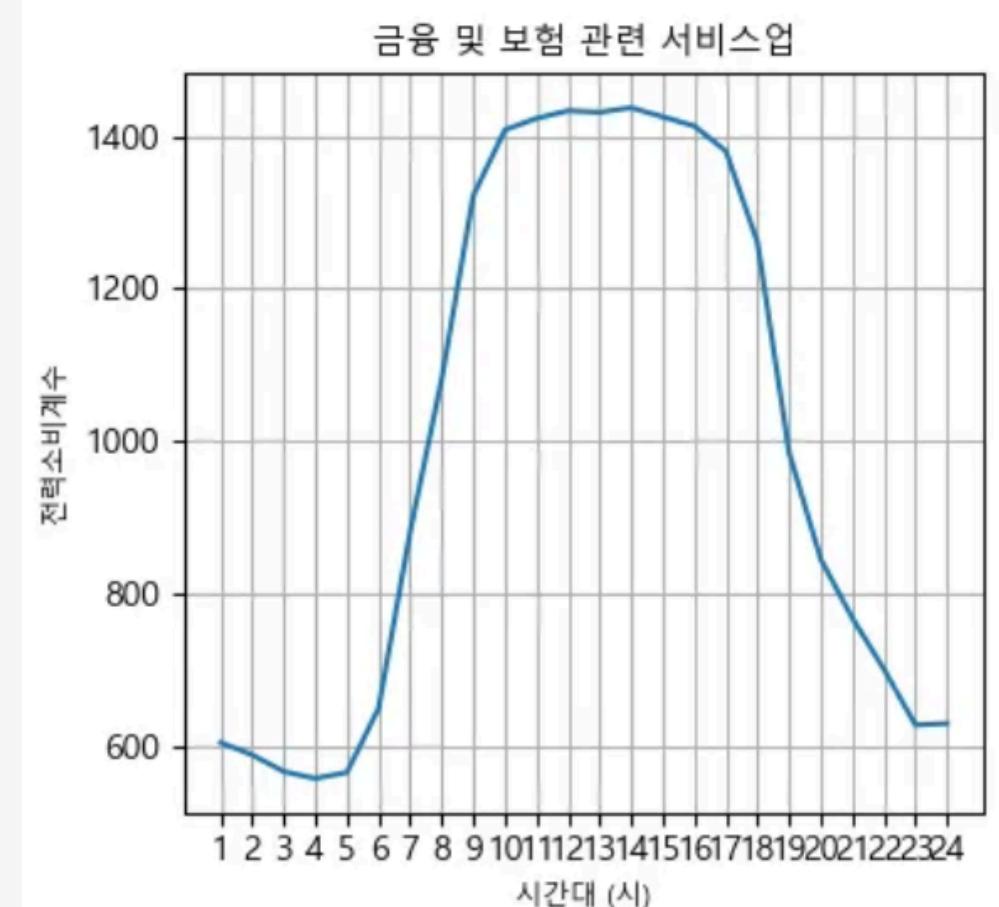
- 도매업, 소매업, 숙박업
- 영상 기록물 제작 및 배급업
- 통신업, 금융 및 보험 관련 서비스업, 부동산업, 연구개발업



대표 예시) 건물 11번

특징

- 6~18시 까지 전력 사용이 집중되는 그래프 형태



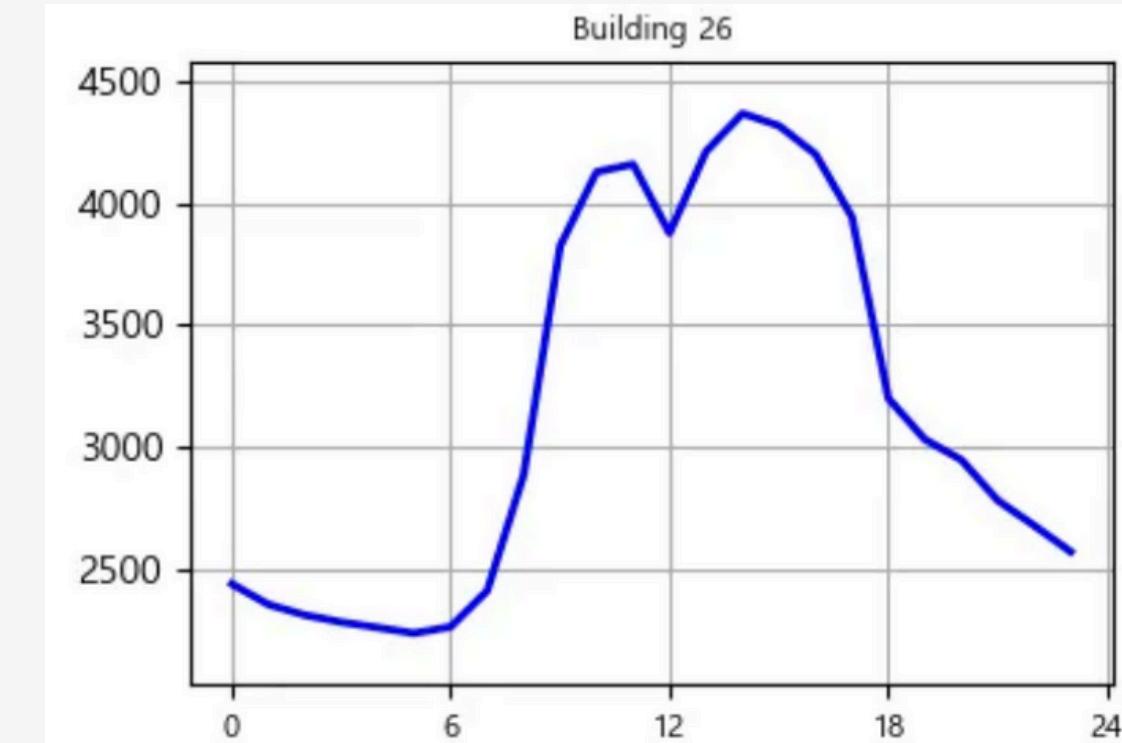
대표 예시) 금융 및 보험 관련 서비스업

03

EDA

그룹 2과 비슷한 그래프

- 농업, 임업 등
- 광업 : 금속, 비금속광물
- 제조업 : 식품, 음료, 자동차, 전자 부품 등



대표 예시) 건물 26번

특징)

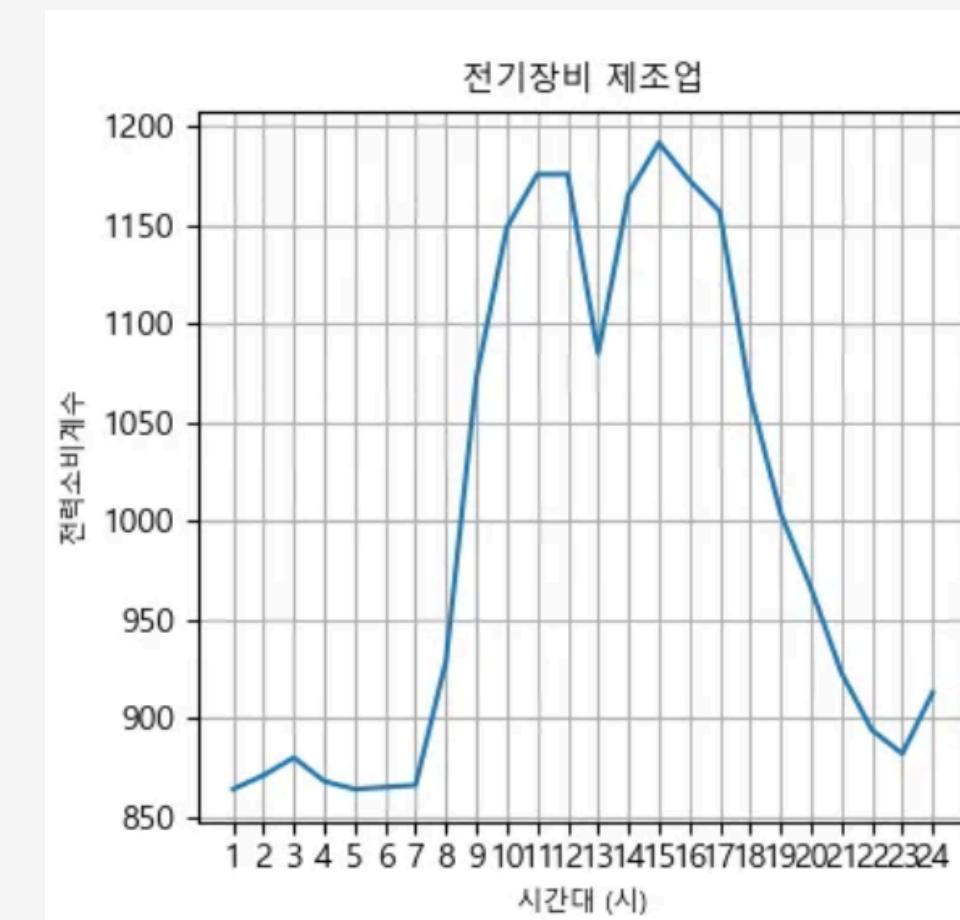
- 6~18시 까지 전력 사용이 집중되는 그래프 형태
- 12시에 잠깐 전력 사용량이 하락하는 특징

12시에 전력 사용량이 하락 하는 이유

위 산업과 같은 산업의 건물이라 유추 가능

따라서, 점심시간 동안 각 산업군의 작업 및 기계 가동이 일시 중단되면서 전력 소비가 감소

이는 전력 소비 패턴에서 12시경 하락이 나타나는 주요 원인



대표 예시) 전기장비 제조업

03

EDA

그룹 3과 비슷한 그래프

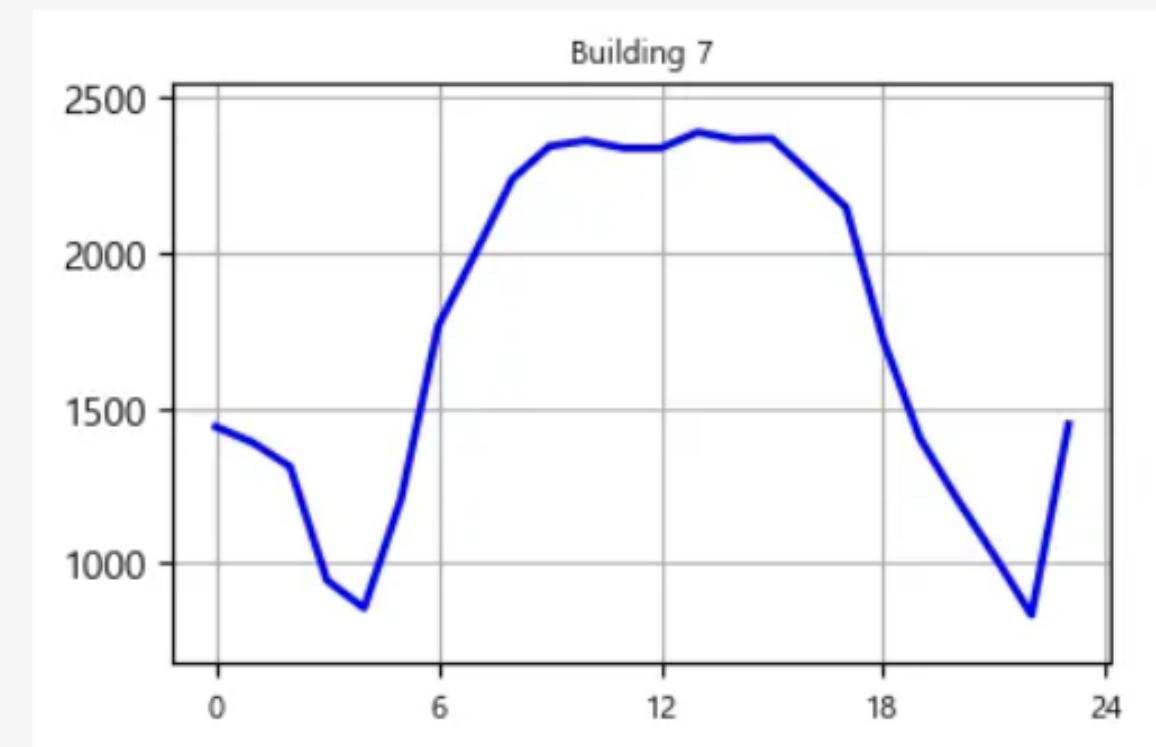
- 방송업, 정보 서비스업, 사회복지 서비스업 등

특징)

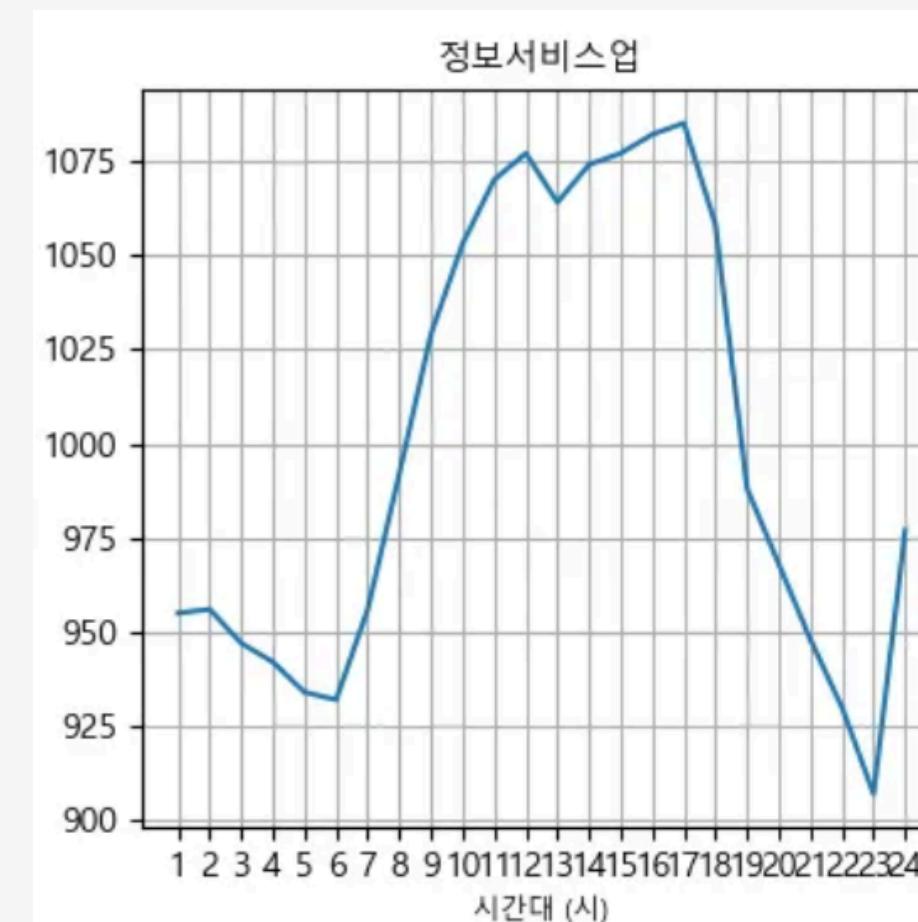
- 6~18시 까지 전력 사용이 집중되는 그래프 형태
- 새벽에 전력 사용량이 급증 및 유지하는 특징

정보 서비스업의 종류

- IT 서비스 및 소프트웨어 개발
- 인터넷 서비스
- 클라우드 및 데이터 센터 서비스
- 통신 서비스
- 디지털 콘텐츠 제공



대표 예시) 건물 7번



대표 예시) 정보서비스업

03

EDA

그룹 4과 비슷한 그래프

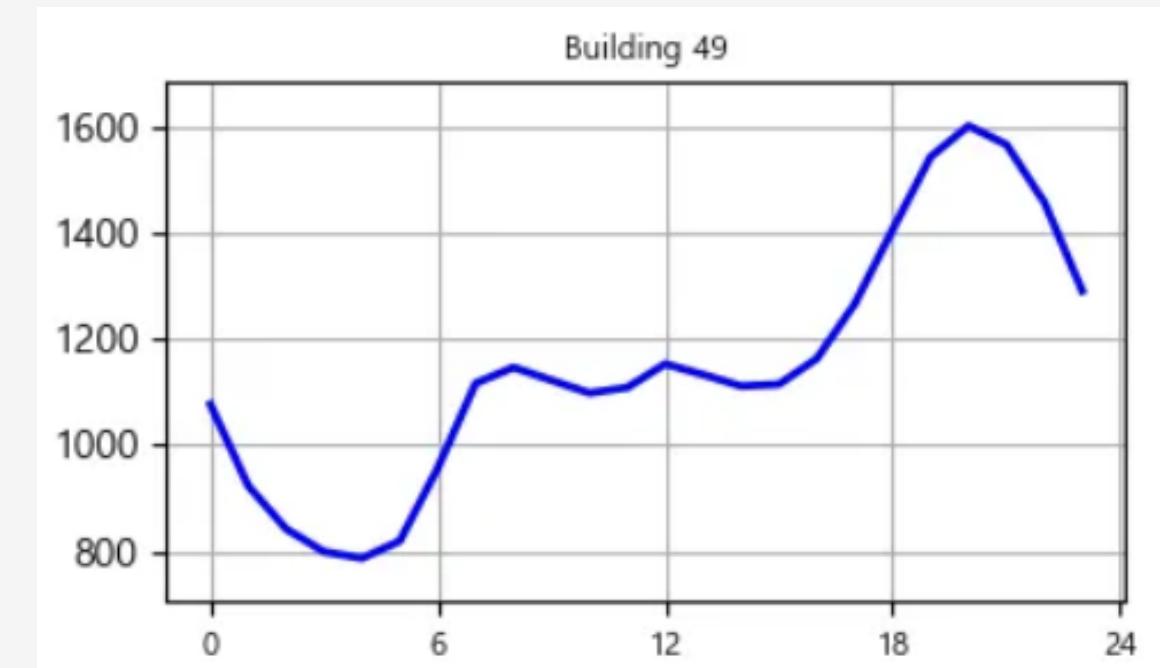
- 주택용 건물

특징)

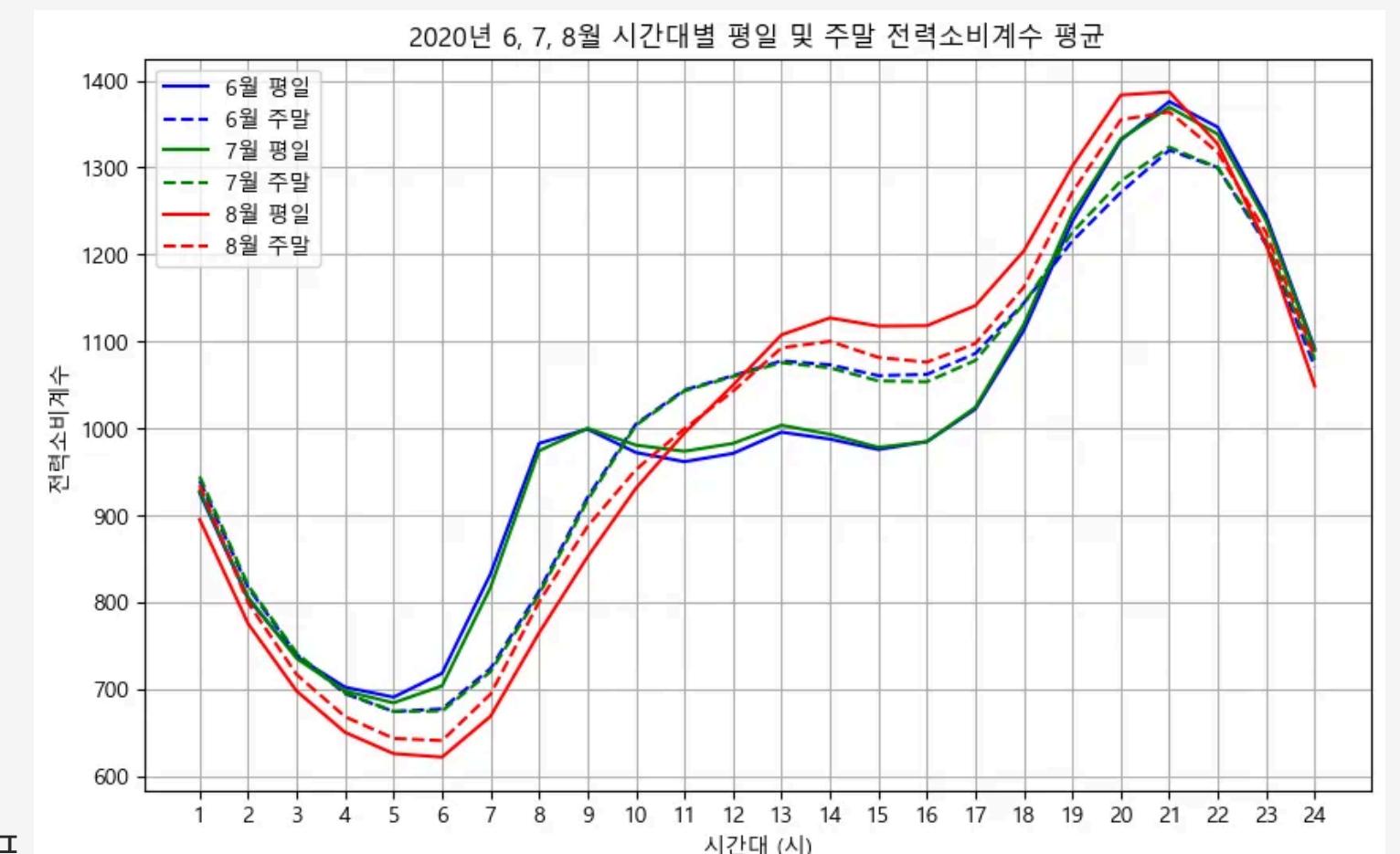
- 0~6시 감소 : 취침 시간
- 6~18시 유지 : 외출 및 근무 시간
- 18시에 증가(퇴근시간 후 복귀)했다가 감소하는 그래프 형태

[통계청]
주택용 요일특성별 1~24시 전력소비계수

2020년 6, 7, 8월 평일과 주말의 전력소비계수 평균 그래프



대표 예시) 건물 49번



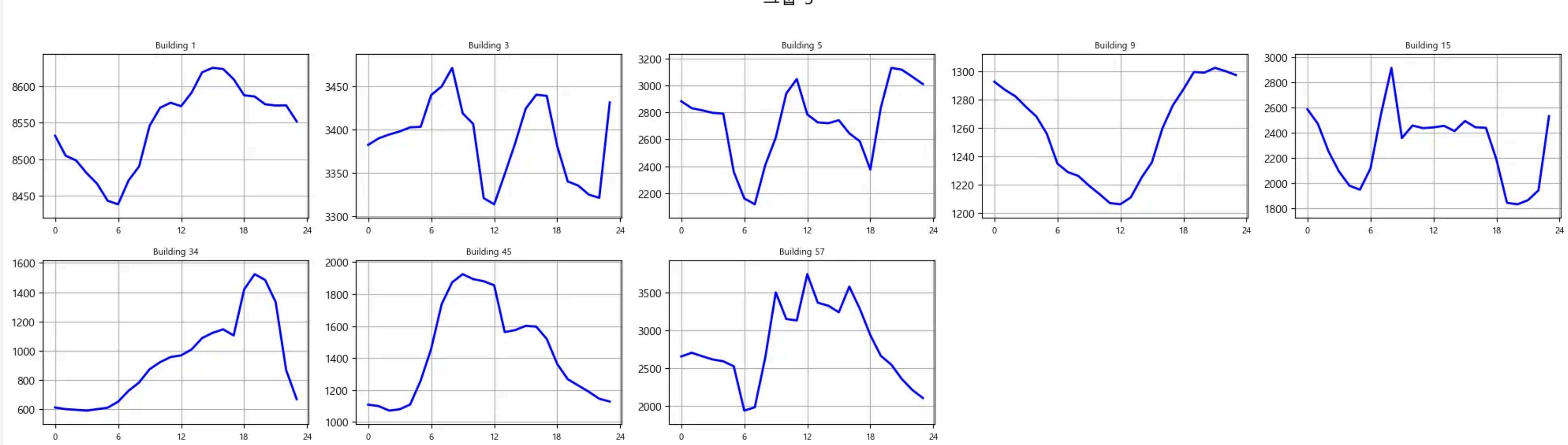
03

EDA

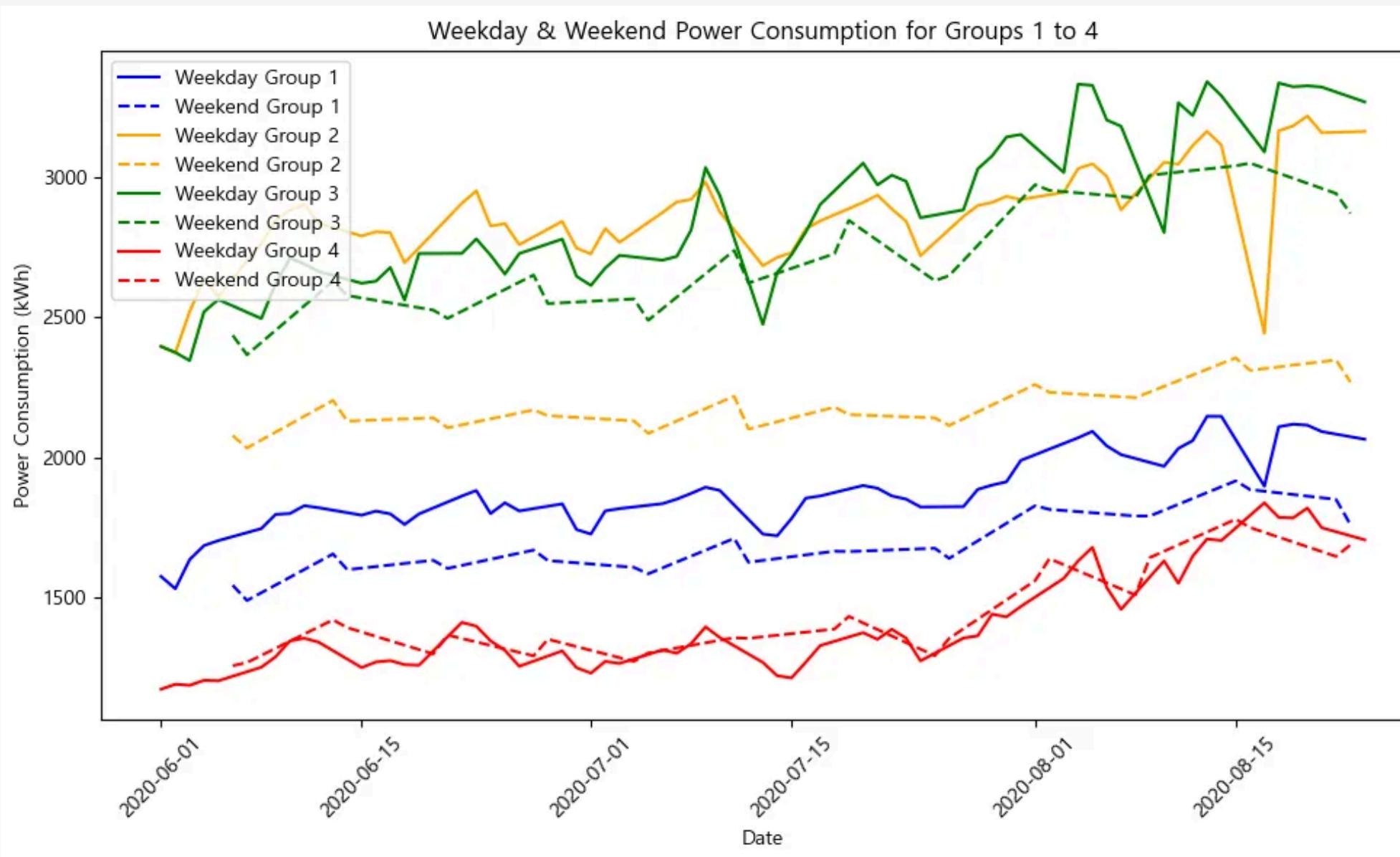
그룹 5

그룹 1~4의 그래프 형태와 유사성이 보이지 않는 나머지 건물

그룹 5



그룹별 특징 시각화



** 그룹 5는 유사성이 없는 건물의 그룹이므로 분석에서 제외

그룹 1: 도매 및 금융업 등

- 1500~2000kWh 사이의 사용량
- 주말은 평일보다 낮은 사용량을 보이지만 1500kWh 부근

그룹 2: 제조업 및 광업 등

- 2500~3000kWh 사이의 사용량
- 주말, 평일 간의 편차가 제일 높은 그룹
- 주말에는 생산 활동을 중단하기 때문인 것으로 추정

그룹 3: 정보 서비스업 등

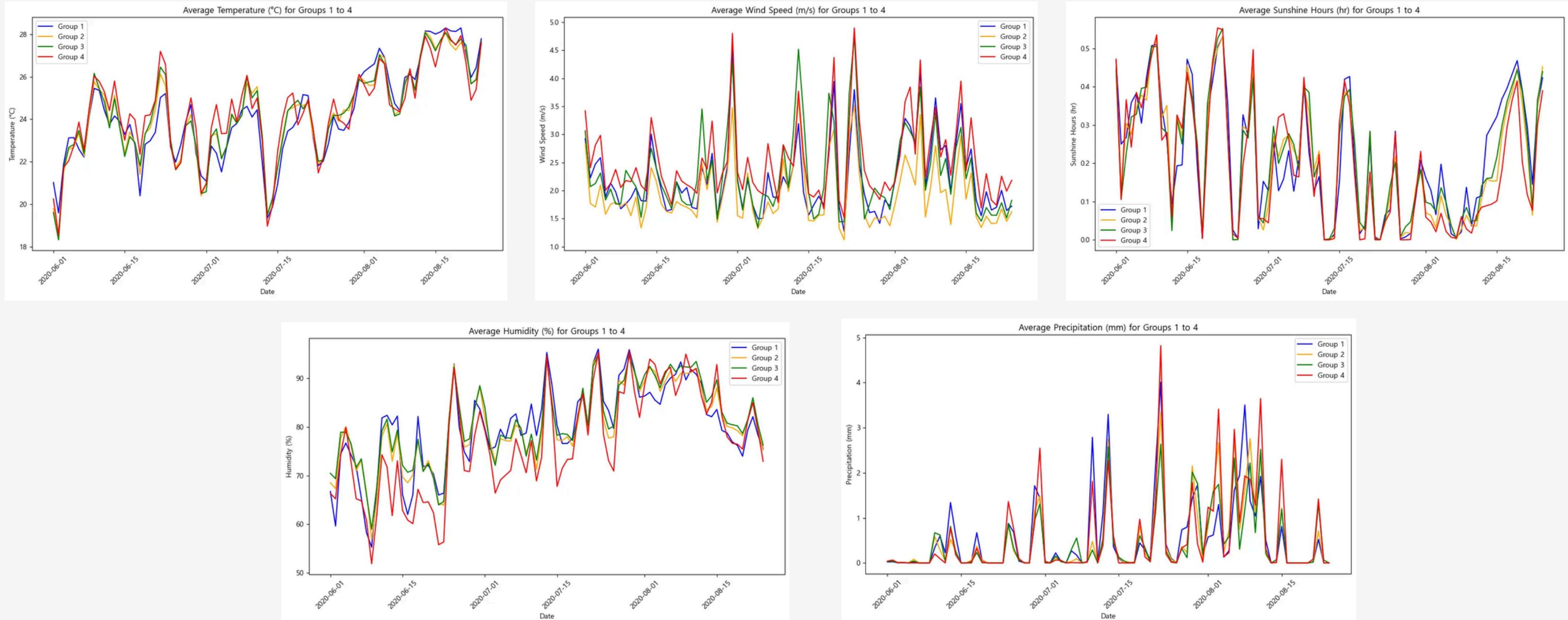
- 2500~3500kWh 사이의 사용량
- 데이터, 인터넷, 통신은 주말에도 전력을 사용하는 산업군
- 평일과 주말 모두 높은 사용량

그룹 4: 주택용 건물

- 1500kWh 부근의 사용량
- 그룹 중 사용량이 제일 낮은 그룹
- 평일과 주말 사이 편차가 제일 낮은 그룹

03

EDA



기온($^{\circ}\text{C}$), 풍속(m/s), 일조(hr), 습도(%), 강수량(mm)

5개 칼럼 모두 시간대별 비슷한 추이

그룹별 특징 시각화

비전기냉방설비운영 여부 및 태양광 보유 비율

그룹 1

- 비전기냉방설비운영: 15/18 (0.83)
- 태양광 보유: 7/18 (0.39)

그룹 2

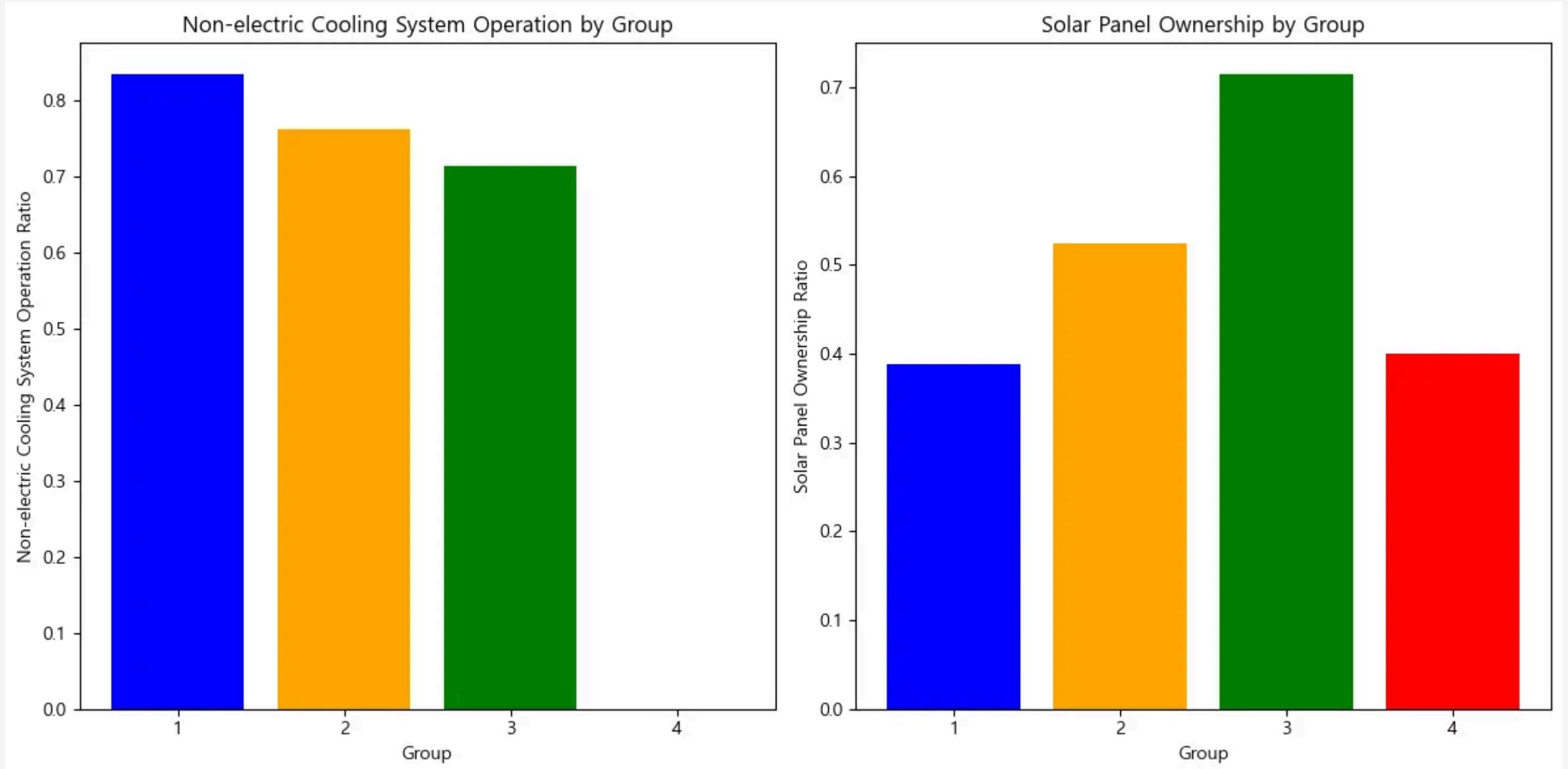
- 비전기냉방설비운영: 16/21 (0.76)
- 태양광 보유: 11/21 (0.52)

그룹 3

- 비전기냉방설비운영: 5/7 (0.71)
- 태양광 보유: 5/7 (0.71)

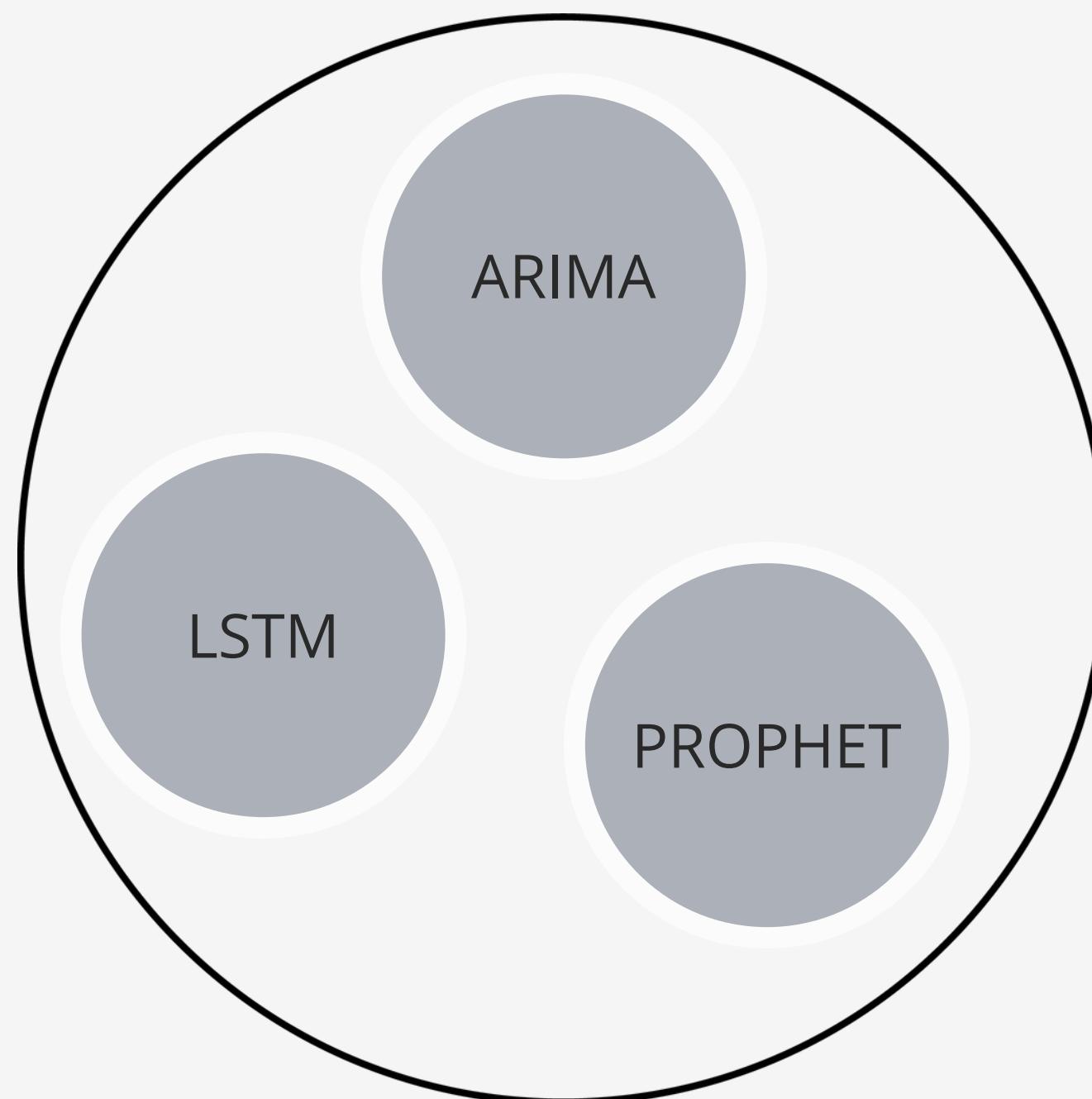
그룹 4

- 비전기냉방설비운영: 0/5 (0.00)
- 태양광 보유: 2/5 (0.40)

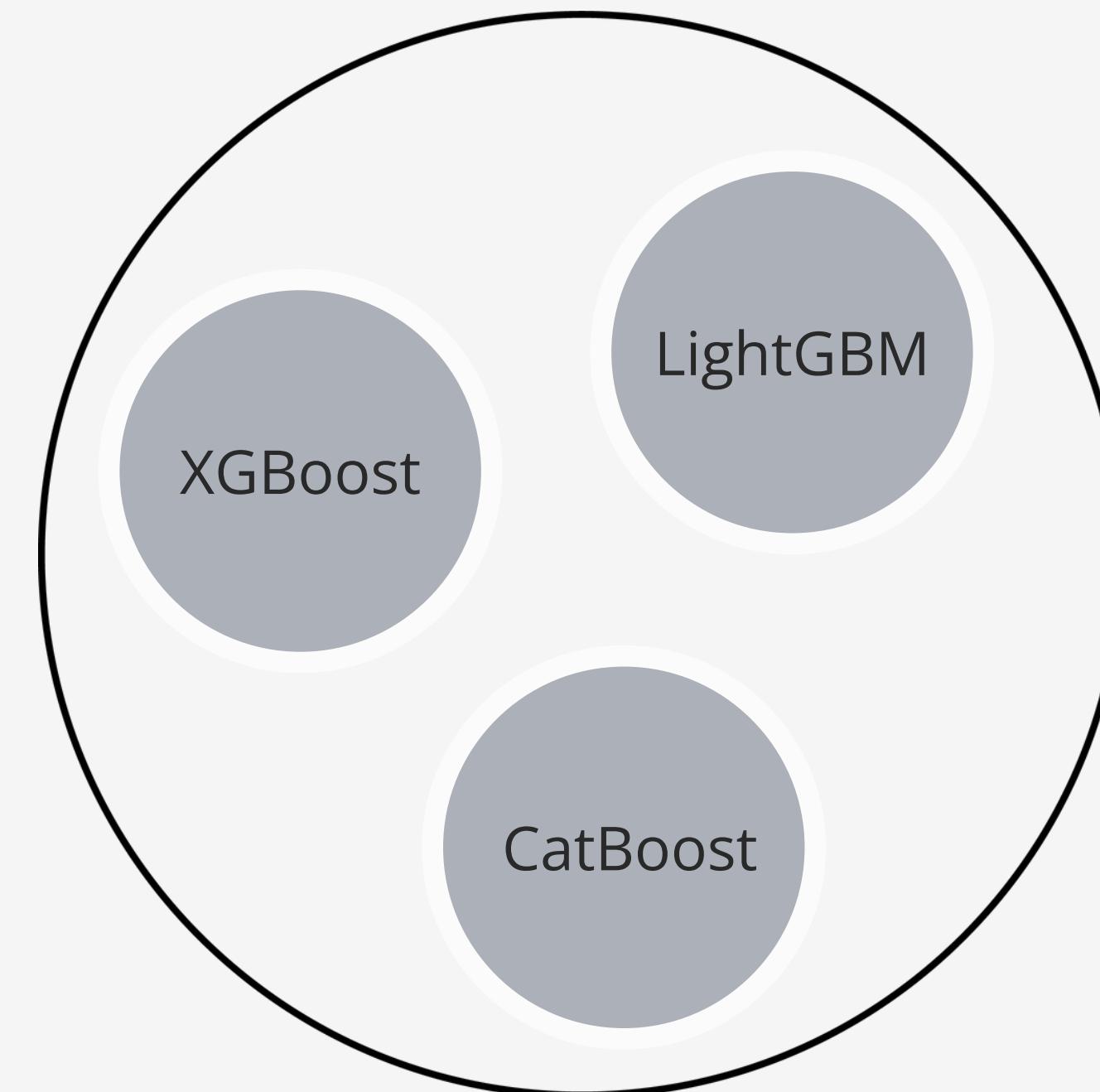


태양광 설비가 있을 경우 자가발전으로 인해 외부에서 공급받는 전력 소비가 감소할 수 있음

시계열 모델



머신러닝 모델



시계열 모델

ARIMA

AR

AutoRegressive, 자기회귀

|

Integrated, 차분

MA

Moving Average, 이동평균

자기회귀, 차분, 이동 평균의

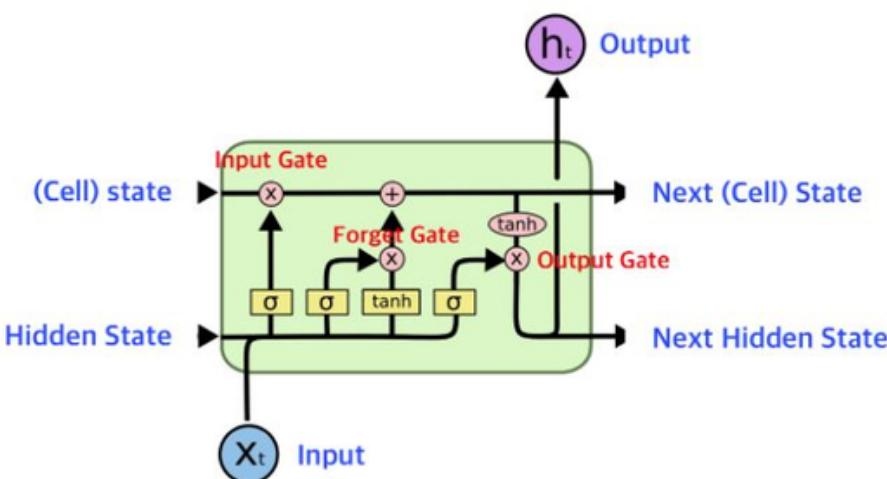
조합으로 구성된 모델

추세가 명확한 시계열 데이터에 적합하며,

데이터가 비정상일 경우

차분을 통해 정상화 가능

LSTM



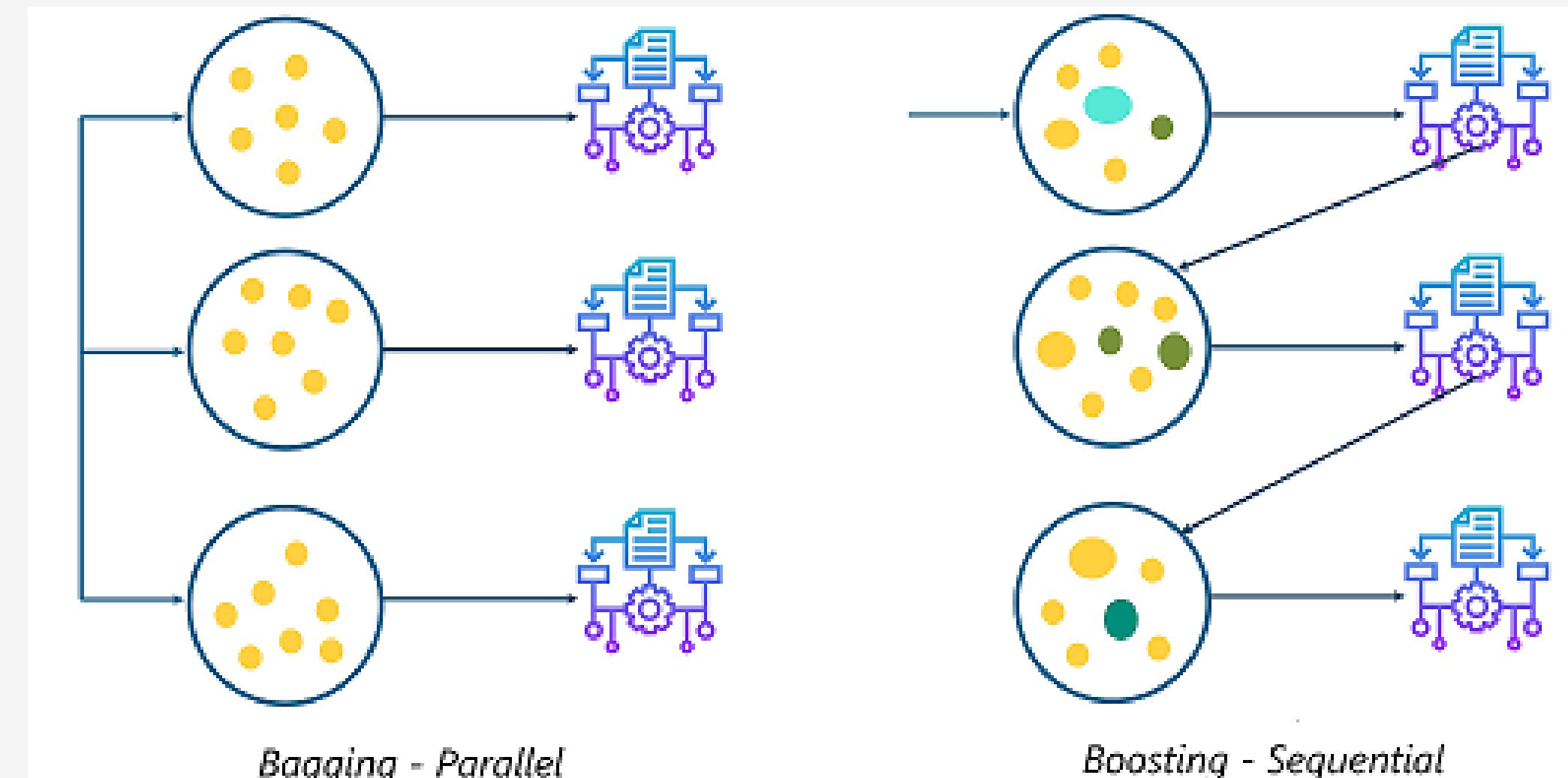
장기 의존성을 잘 학습하는 RNN,
비선형 관계를 모델링하며
다양한 입력 길이에 대응할 수 있는
시계열 데이터 예측에 강력한 성능을 보임

PROPHET

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t.$$

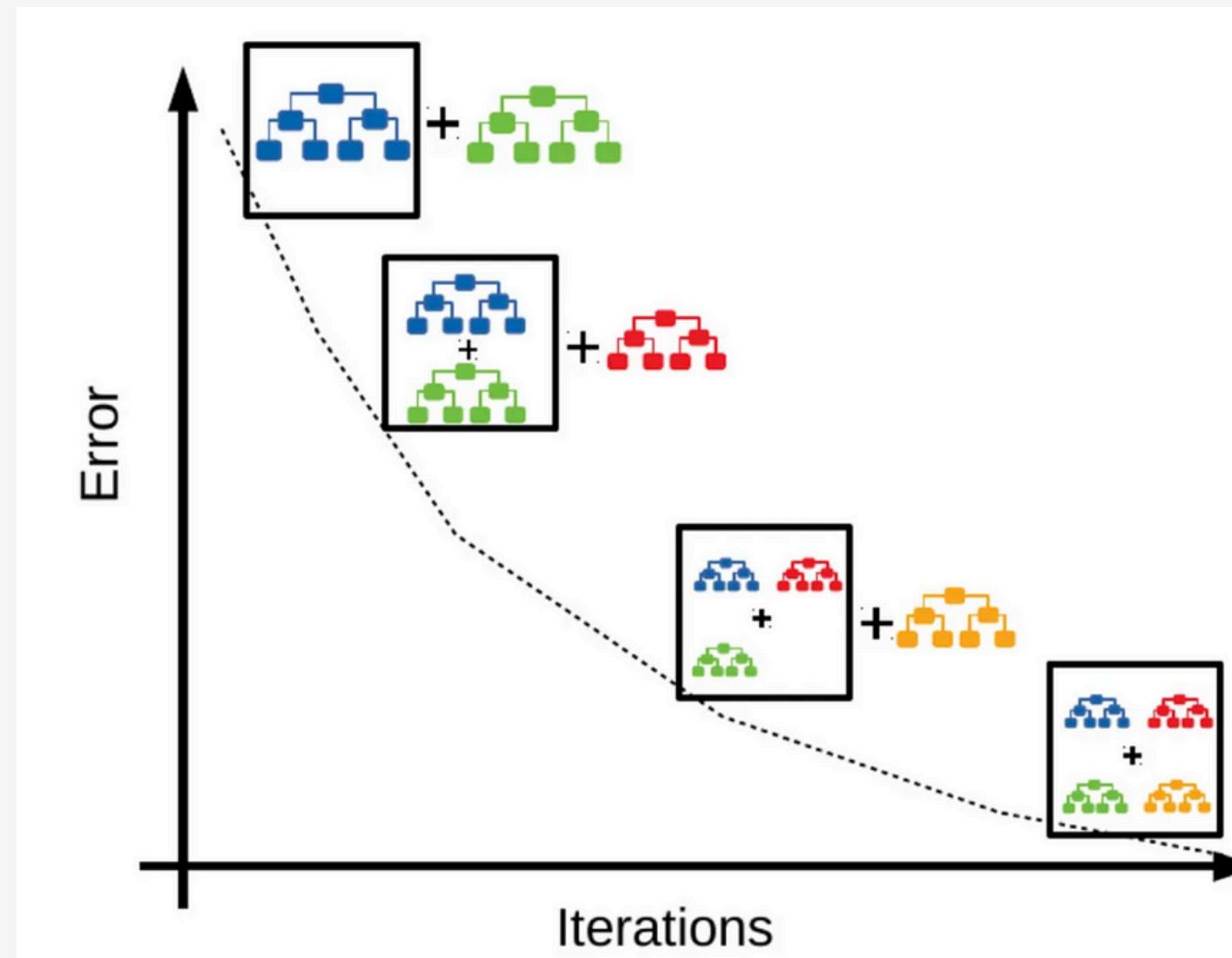
g(t) : 추세, s(t) : 계절, h(t) : 휴가

계절성과 추세 변화를 자동으로 감지하고,
계절적 패턴 예측에 뛰어나며, 수식이 단순
하여 이해와 적용이 쉬움



Bagging: 독립된 모델들의 결과를 합산하여 다수결 투표(Voting)을 통해 결과를 산출

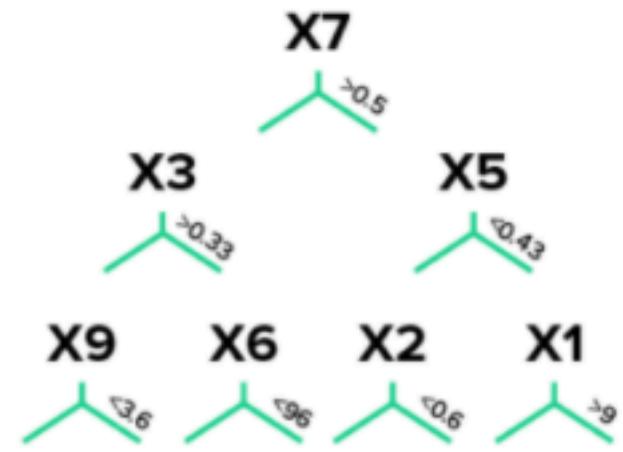
Boosting: 특정 모델의 결과를 다른 모델의 input으로 사용하는 방식, 모델 간 가중치를 부여하여 결과 산출



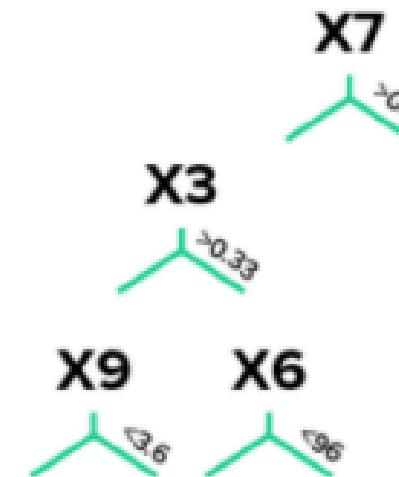
Gradient Boosting: 부스팅 기법을 이용한 앙상블 모델 중 하나

잔차(실제값과 예측값의 차이)를 이용하여 이전 모형의 약점을 보완하는 새로운 모형을 순차적으로 적합한 뒤
이들을 선형 결합하여 얻어진 모형을 생성하는 지도 학습 알고리즘

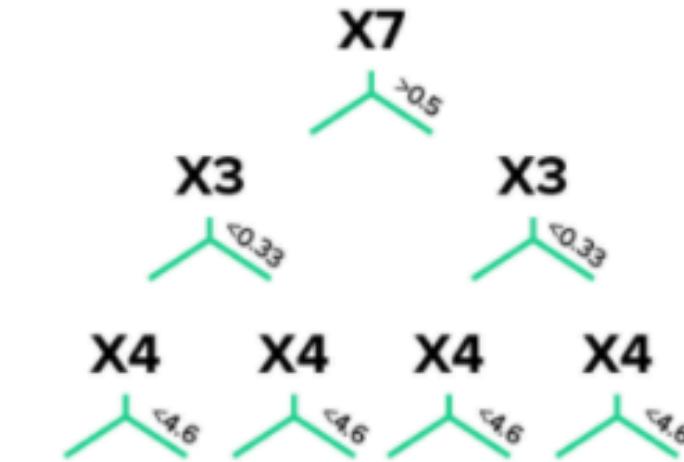
머신러닝 모델

XGBoost

레벨별 확장 방식으로
강력한 성능과 안정성 제공
그러나, 대용량 데이터에서 느리고
메모리 사용이 많음

LightGBM

리프별 확장 방식으로 빠르고 메모리 효율적
과적합 위험이 있으며, 파라미터 튜닝이 필요

CatBoost

범주형 데이터를 직접 처리하며
데이터가 커지면 학습 속도가 느릴 수 있음

성능평가지표

MAE

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

예측값과 실제값 간의 절대 오차를 평균 하는 방법
0 값 문제 없이 계산할 수 있고,
해석이 쉬움

MAPE

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

실제값과 예측값의 절대 오차를 백분율로 나타낸 후 그 평균을 구하는 방식
예측값이 실제값과 얼마나 차이가 나는지를 비율로 평가

SMAPE

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \times 100$$

예측 오차를 비율로 측정하는 성능 지표
시계열 데이터에서 많이 사용되며,
예측값과 실제값의 절대 오차를 상대적인 비율로 평가

MAE

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

절대 오차만 측정하므로, 값이 작아질수록 작은 오차처럼 보임

하지만, 실제로는 상대적인 오차가 매우 클 수 있음

따라서, 값이 작은 경우에도 절대 오차만을 보여주기 때문에

상대적인 예측 성능을 충분히 반영하지 못할 수 있음

그에 반해, SMAPE는 비율로 측정하기 때문에 예측 성능을 해석하기 쉬움

05

평가

Why SMAPE ?

MAPE

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

분모에 0이 들어갈 수 없으므로

비율 계산에서 문제가 발생할 수 있음

	num	target
54684	27	0.0
54685	27	0.0
54686	27	0.0
54687	27	0.0
54688	27	0.0

target 데이터에 0이 존재

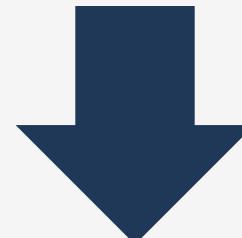
따라서, MAPE 사용에 어려움을 겪으므로 기각

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \times 100$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2}$$

SMAPE는 0 값이 있어도 안정적으로 오차를 계산할 수 있고,
비율적인 차이를 잘 반영하여 계절성 있는 데이터를 평가하는 데 적합

큰 오차에 덜 민감한 SMAPE의 단점을 보완하기 위해
큰 오차를 사용하는 RMSE 추가



RMSE와 SMAPE를 동시에 성능지표로 사용

시계열 모델

Table 1. Time series models

SMAPE(%)	Prophet	14.40
	LSTM	196.21
	Bi-LSTM	196.23
	ARIMA	73.30

SMAPE 점수에 따라 **Prophet** 모델의 성능이 제일 좋은 것으로 확인

머신러닝 모델

Table 2. ML models

			Base	*Additional Variables (1)	Scaling (2)	Parameter Tuning (3)	Clustering (4)
SMAPE (%)	Overall Model	XGBoost	19.54	2.74	2.74	3.82	2.44
		LGBM	17.39	3.04	3.04	3.51	2.52
		CatBoost	15.76	2.51	2.51	3.77	2.34
Building-specific Models		XGBoost	14.25	2.48	2.47		
		LGBM	16.31	2.35	2.35		
		CatBoost	13.52	2.49	2.49		

* holiday, hour_sin, hour_cos, rolling_mean, rolling_std Added variable.

SMAPE 점수에 따라 **LGBM** 모델의 성능이 제일 좋은 것으로 확인

시계열 모델 / 머신러닝 모델

Table 3. Final models

SMAPE(%)	models		
	ML	LGBM (Additional Variables)	2.35
Time series		Prophet	14.40
RMSE	ML	LGBM (Additional Variables)	83.58
Time series		Prophet	338.42

따라서, 최종 모델을 LGBM, Prophet으로 선정

평가 Prophet

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t.$$

- **g(t) (트렌드)**

데이터가 전반적으로 증가하거나 감소하는 장기적인 변화를 설명

- **s(t) (계절성)**

주기적인 변화를 설명

- **h(t) (휴일)**

특정 이벤트나 휴일이 데이터에 미치는 특별한 영향을 반영

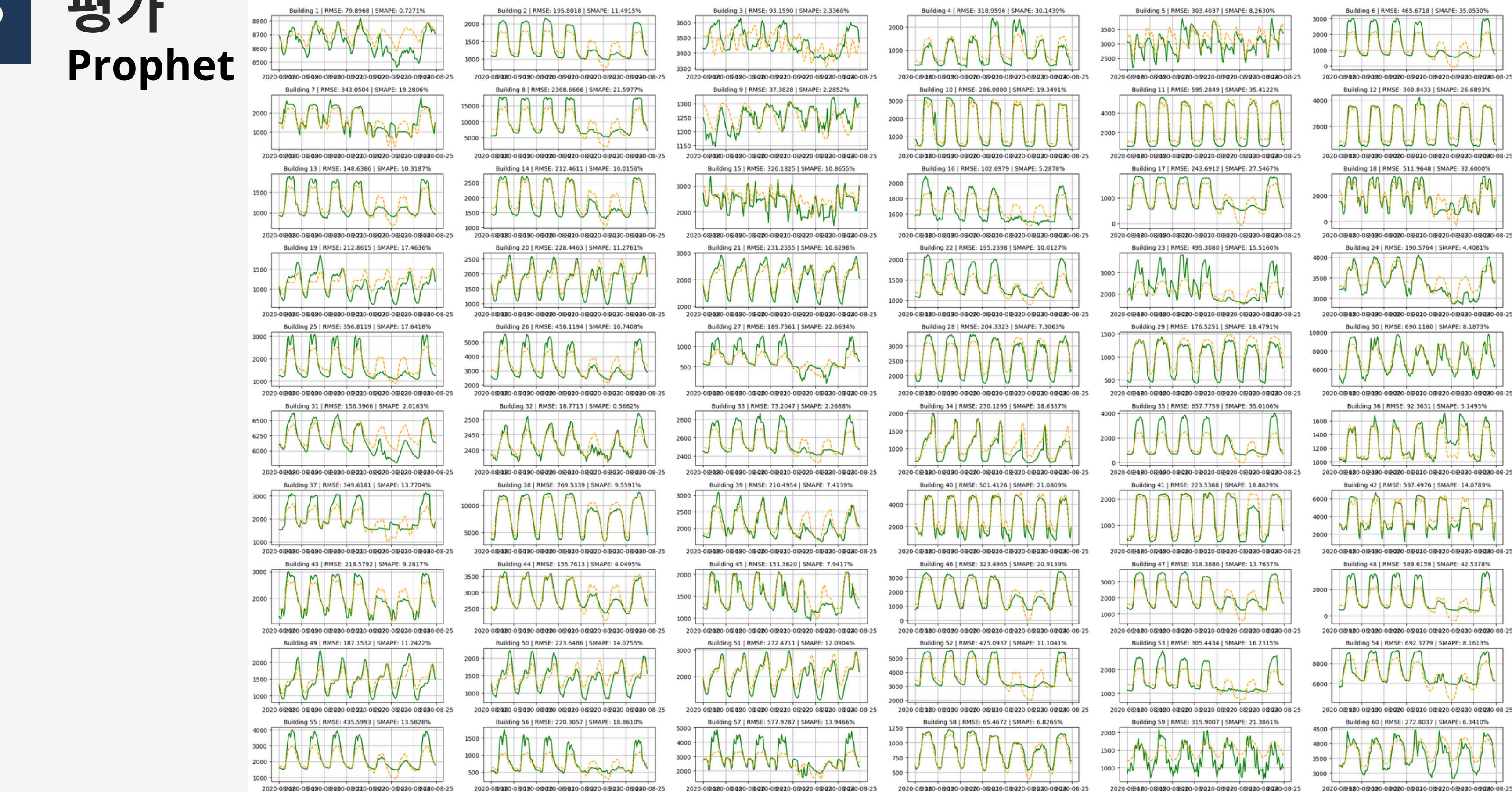
- **$\epsilon(t)$ (오차)**

예측할 수 없는 랜덤한 변동 또는 노이즈를 반영(정규분포 따름)

05

평가 Prophet

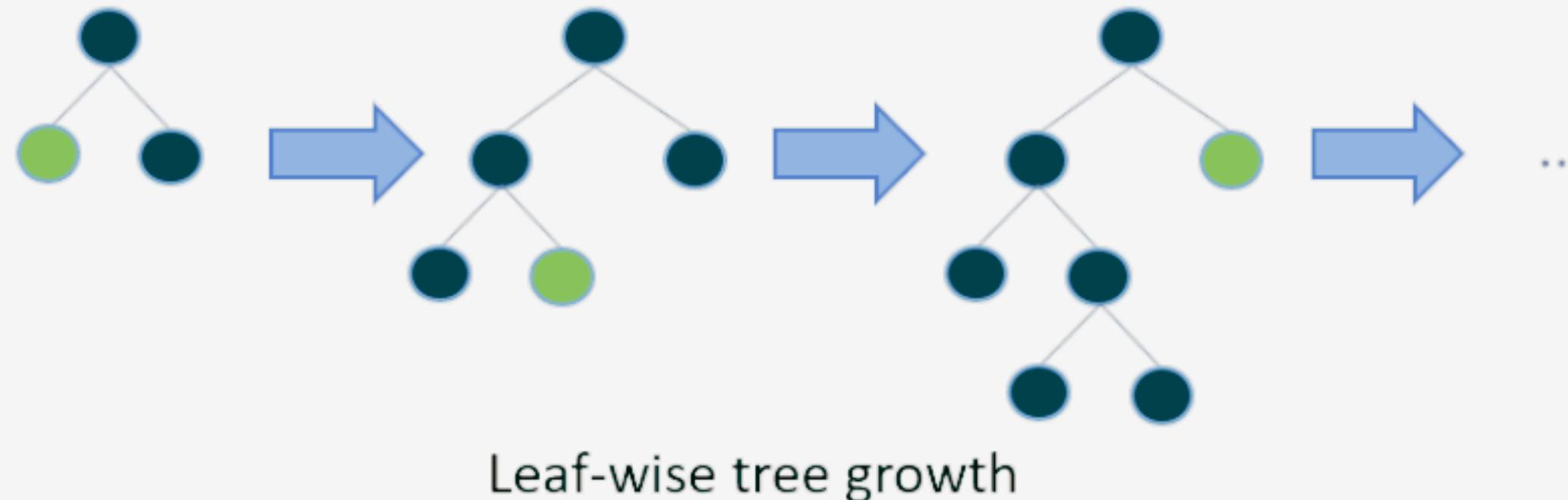
Prophet Model Predictions for All Buildings



05

평가 Prophet

모델	성능	이유	주요 특징
Prophet	✓	<ul style="list-style-type: none"> - 명확한 계절성과 휴일 변동을 반영 - 주기적 패턴만으로도 좋은 예측 가능 	<ul style="list-style-type: none"> - 계절성 및 휴일 효과 반영, 추세가 불분명해도 성능 유지 - 구간별 선형 함수, Fourier series 활용
ARIMA	✗	<ul style="list-style-type: none"> - 추세가 불분명하면 차분을 사용해도 성능 저하 - 계절성을 수동으로 반영해야 하므로 복잡함 	<ul style="list-style-type: none"> - 차분을 통해 정상성 가정, 계절성 처리 한계 - SARIMA 확장 필요
LSTM	✗	<ul style="list-style-type: none"> - 3개월로 데이터 길이가 짧아 장기 패턴 학습에 한계 - 장기적인 패턴이 없을 경우 성능 저하 	<ul style="list-style-type: none"> - 장기 의존성 학습 강점, 데이터 길이 짧으면 과적합 위험 - 학습에 많은 데이터 필요, 복잡한 구조



리프 중심의 트리 확장 알고리즘으로,
손실 최소화 하는 작업이 효율적임
또한, 보다 깊고 복잡한 트리 구조를 가짐
뿐만 아니라 수렴 속도가 빠름
메모리 및 시간 절약 가능

05

평가 LGBM

건물 별로 각각 Feature Importance를 확인한 결과,

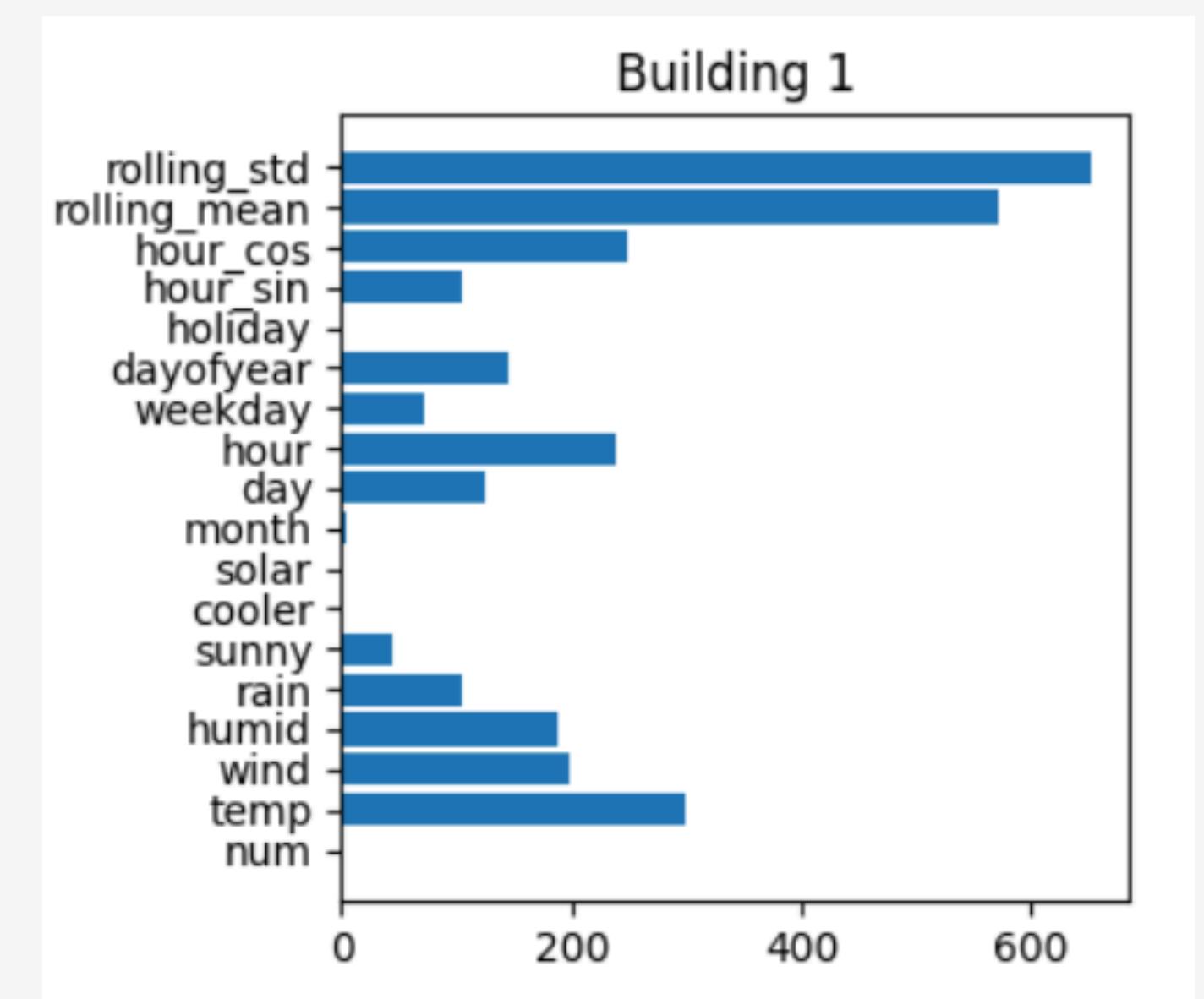
새로 추가한 변수인

“rolling_std”(이동 편차)

“rolling_mean”(이동 평균)

이 두 개가 큰 결정력을 지니고 있었음

따라서, 모델 성능이 향상됨



- **이동 평균**

일정 기간 동안의 값들의 평균을 구한 것



- **이동 표준 편차**

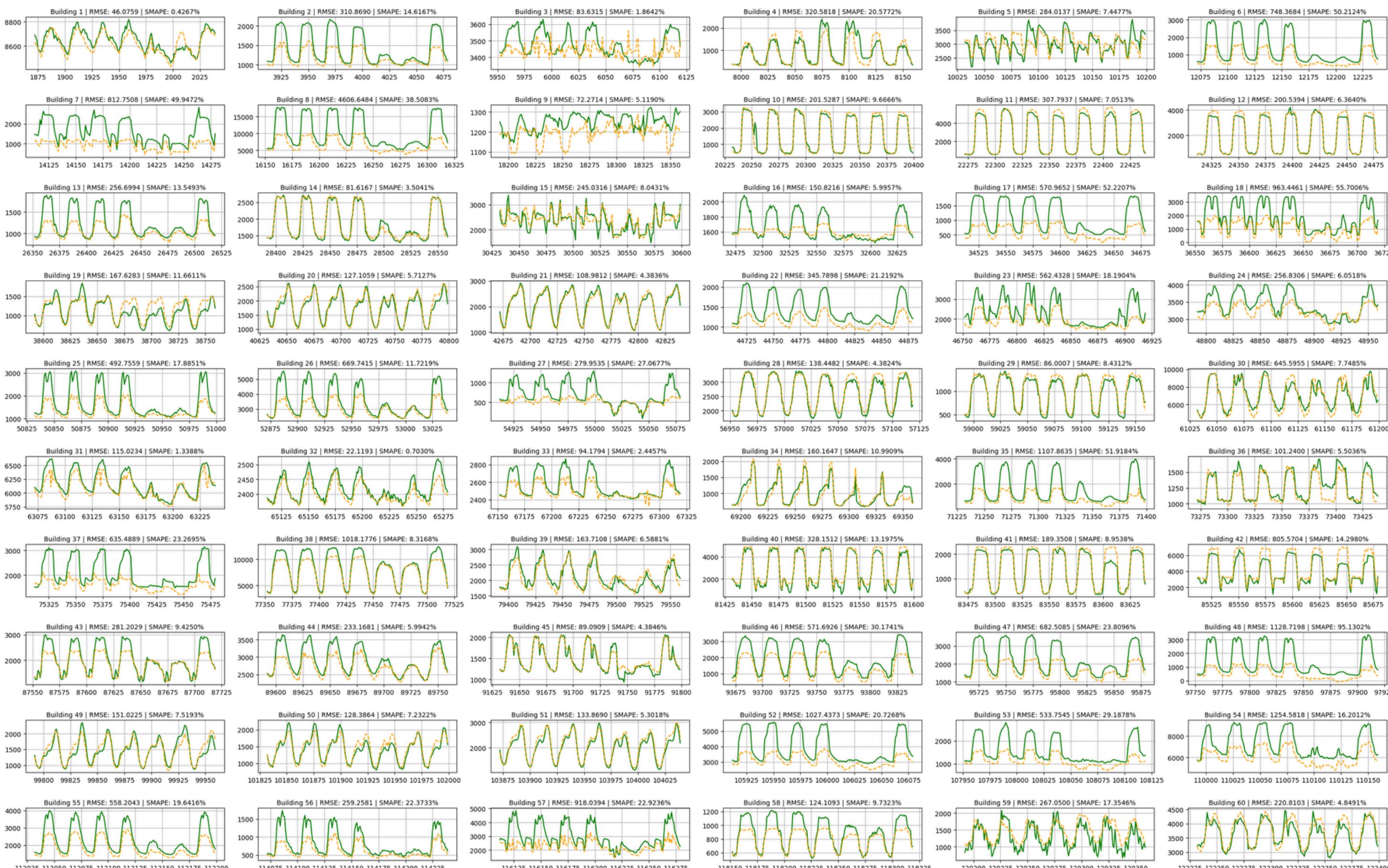
이동 평균과 비슷하게 일정 기간 동안의 값들의 표준 편차를 구함

시계열 데이터 분석에서는 **이동 평균, 표준편차**를 사용하여 시간에 따른 값의 변동을 부드럽게 만들어 **트렌드 파악** 가능

05

평가 LGBM

LightGBM Model Predictions for All Buildings



05

평가

LGBM

모델	성능	장점	단점	소요시간
LGBM	✓	<ul style="list-style-type: none"> - Leaf-wise 방식으로 속도 빠름 - 결정력 있는 변수에 더 민감하게 반응 	<ul style="list-style-type: none"> - 10000개 이하의 데이터에서는 과적합 우려 	3.0s
XGBoost	✗	<ul style="list-style-type: none"> - 과적합 방지 - 결손값 처리 	<ul style="list-style-type: none"> - LGBM보다 속도가 느림 	8.3s
CatBoost	✗	<ul style="list-style-type: none"> - 범주형 데이터를 다루는데 용이함 - 과적합 방지 	<ul style="list-style-type: none"> - 수치형 데이터에서는 비교적 성능이 낮음 	1m11.5s

전략 1

- 현재 데이터에는 외부 요인 변수(온도, 습도, 일조 등)만 존재
- 따라서 건물에 대한 특징(건물의 크기(부지), 건물의 용도(유형)), 주변 인프라 데이터가 있었다면 추가적으로 최적의 건설 위치를 추천해 줄 수 있을 거라 예상

전략 2

- 직접 비슷한 모양의 그래프를 찾는 것이 아닌, **DTW** 처리 후 **K-means** 군집화를 하여 자동으로 유사한 시계열끼리 군집하는 방식 채택

전략 3

- 모델 향상에만 집중
- 모델 향상 뿐만 아니라 다양한 아이디어 바탕으로 여러 관점에서 데이터를 분석하여 보다 의미 있는 인사이트 도출 가능

** DTW는 시계열 데이터의 길이와 변화 속도가 달라도 시간 축을 비틀어 유사성을 측정하는 방법

** DTW 기반 K-평균(K-Means)은 시계열 간 패턴 유사성을 반영해 군집화

감사합니다