

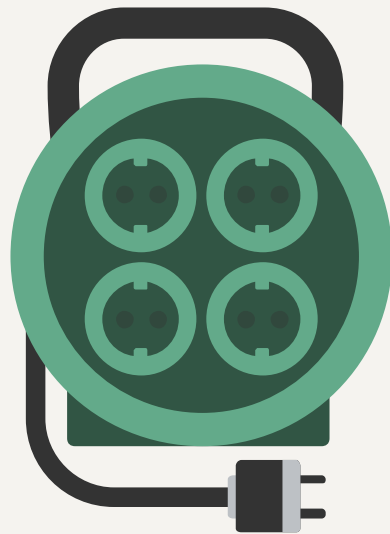
# 기상에 따른 공동주택 전력수요 예측 개선

김동희 진재언 김태현 송다은 이승아



01

# 주제 설명 및 선정 배경





# 주제 설명 및 선정 배경

## 주제 선정 배경

### 여름철 폭염 기간에 노후 아파트의 정전은 계속 증가하는 추세

정전 사태에 특히 취약한 곳은 지어진 지 40년 이상 된 노후 아파트들이다. 가정용 전력의 전압은 110V에서 220V로 1973년부터 순차적으로 승압됐는데 과거에 지은 노후아파트의 경우 집마다 변압기가 설치된 경우가 많다. 가구 내 전력 사용량이 변압기 용량을 초과하면 온 집안 전력이 끊기는 일이 생긴다. 110V 변압기는 부품을 구하기도 어려워 복구에 하루 이상의 긴 시간이 걸리기도 한다.

“변압기 용량 초과로 잠시 후...” 노후아파트, 날마다 ‘정전 주의보’,  
천호성, 한겨레



한편, 2020년 인구주택총조사에 따르면 국내 아파트의 42.7%가 20년 이상, 9.6%는 30년 이상 된 아파트인 것으로 나타났다. 오래된 아파트일수록 세대별로 설계된 전력용량이 낮아 전력사용 과다에 따른 고장에 취약하며, 폭염기간 정전으로 국민 불편이 가중되고 있다.

“한전-기상청, 전력·기상 빅데이터 활용 안정적인 전력 공급 기여”  
전력경제신문

오래된 아파트일수록 세대당 설계된 전력용량이 낮아 전력사용 과다에 따른 고장에 취약하며

**동·하계 전력수요가 급증하는 시기에** 과부하 정전이 빈번하게 발생하여 국민의 불편이 가중되고 있는 상황



# 주제 설명 및 선정 배경

## 주제 설명

따라서, 정확하게 전력 수요를 예측함으로써 전력 공급이 끊기지 않도록 하여  
정전을 사전에 예방하는 것이 주된 과제!



### 최종 목표 ①

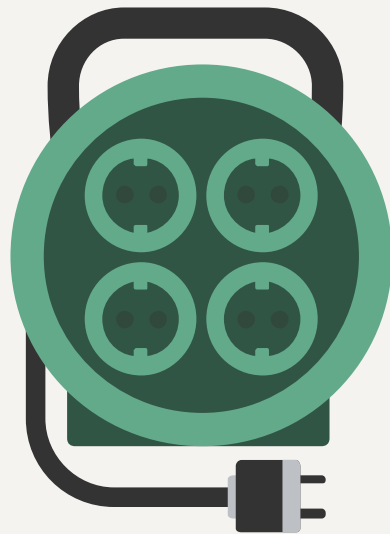
기상 변수 및 공공데이터 등을 활용하여  
**공동주택 전력수요 증감** 영향 요인 분석

### 최종 목표 ②

**계절, 지역**에 따른 모델 세분화를 통한  
공동주택 전력 수요 예측 (전력기상지수)  
최적모델 개발

02

# 데이터 설명 및 EDA





# 데이터 설명 및 EDA

## 분석데이터 소개 | 학습데이터 (전력)



원격검침이 이루어지는 공동주택 중 각 연도별 결측치가 없는 단지를 대상으로  
기상예보 격자별로 산출된 2020년, 2021년, 2022년 **전력통계값**으로  
총 7593355개의 행으로 이루어짐  
(격자내 공동주택이 10개 이상인 격자에 대해서만 공개됨)



# 데이터 설명 및 EDA

## 학습데이터 (전력) | TM, HH24

tm

0~23시 시간을 포함한 공동주택 전력부하 측정 날짜

hh24

공동주택 전력부하 측정시간(1~24)

ex. 5시는 4시 1분부터 5시 00분까지의 전력부하 의미

electric_train.num	electric_train.tm	electric_train.hh24	electric_train.n	...	electric_train.elec
4821	2021-01-01 1:00	1	11	...	99.56
4821	2021-01-01 2:00	2	11	...	91.78



# 데이터 설명 및 EDA

## 학습데이터 (전력) | elec

### 전력기상지수

기상변화에 따른 지역별 공동주택의 예상되는 전력수요변화를  
기상예보처럼 국민들이 쉽게 인지할 수 있도록 수치화하여 최대 72h 예측해주는 서비스

### 산출식

$$A\text{격자 } 00\text{시각의 전력기상지수} = \frac{A\text{격자 } 00\text{시각의 전력수요 (또는 예상 전력수요)}}{A\text{격자 해당년도 전력수요}}$$

국민들이 사전에 예방할 수 있도록 지원하고자  
해당지역(5km x 5km) 공동주택의 연중 평균부하를 100으로 하였을 때,  
특정 시각의 전력수요 또는 예상전력수요를 상대비율로 표현한 값





# 데이터 설명 및 EDA

## 학습데이터 (전력) | elec

electric_train.num	electric_train.tm	electric_train.sum_load	electric_train.n_mean_load	...	electric_train.elec
4821	2021-01-01 1:00	751.32	68.60645	...	99.56
4821	2021-01-01 2:00	692.6	68.60645	...	91.78
4821	2021-01-01 3:00	597.48	68.60645	...	79.17
4821	2021-01-01 4:00	553.48	68.60645	...	73.34
4821	2021-01-01 5:00	526.24	68.60645	...	69.73
4821	2021-01-01 6:00	538	68.60645	...	71.29

격자 4821의 2021년 1월 1일 6:00 전력기상지수



# 데이터 설명 및 EDA

## 학습데이터 (전력) | elec

전일 또는 지난주 대비 전력수요의 증감비율을 통해  
전력수요 변화량을 예상할 수 있으므로 활용해볼 수 있음!



Ex)

전력기상지수는 본 프로젝트의 종속변수(Y)이며

기상변수를 통해서 이를 예측해야 함

전일 최고 전력기상지수가 100,

당일 최고 전력기상지수가 125이면,

당일 최대수요는 전일대비  $125/100 = 1.25$ 배(25%) 증가를 예상 가능

즉, 전력기상지수 예측값과 실제 전력수요와의

상관계수를 높이는 것이 목표!



# 데이터 설명 및 EDA

학습데이터 (전력) | ⑦ elec

## 모델 평가 지표

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

전력기상지수 예측값과 실제 전력수요의 **상관계수** 전격자 평균

상관계수를 높이는 것이 목표!



# 데이터 설명 및 EDA

## 학습데이터 (기상) | AWS 측정 데이터

nph\_ta

AWS 관측 자료에서 지형효과를 반영하여 격자형태로 생산한 **기온** 자료이며, 단위는 °C

nph\_hm

AWS 관측 자료에서 지형효과를 반영하여 격자형태로 생산한 **상대습도** 자료이며, 단위는 %

nph\_ws\_10m

AWS 관측 자료에서 지형효과를 반영하여 격자형태로 생산한 **풍속** 자료이며, 단위는 m/s



# 데이터 설명 및 EDA

## 학습데이터 (기상) | AWS 측정 데이터

nph\_rn\_60m

AWS 관측 자료에서 지형효과를 반영하여 격자형태로 생산한 **강수량** 자료이며, 단위는 mm

nph\_ta\_chi

AWS 관측 자료에서 지형효과를 반영하여 격자형태로 생산한 **체감온도** 자료이며, 단위는 °C

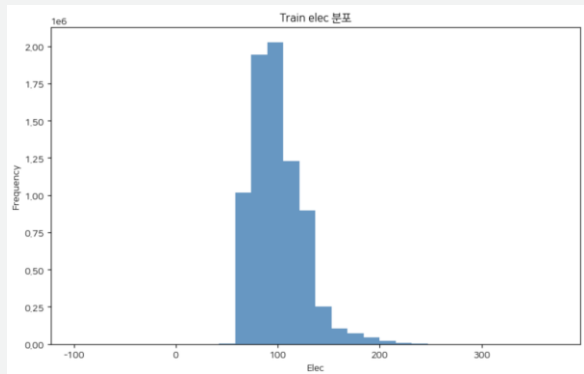
⋮

이후 단계에서 공동주택의 전력수요와 가장 밀접한 관계를 보이는 요인을 찾아볼 예정

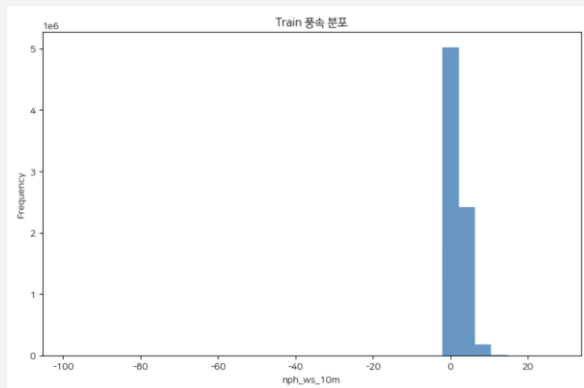


# 데이터 설명 및 EDA

## EDA



```
electric train.elec
-99.00  5
48.90  1
49.18  1
49.28  1
49.69  1
..
270.02  1
270.12  1
272.00  1
284.40  1
373.06  1
```



```
electric train.nph_ws_10m
-99.0  170
0.0  105950
0.1  242837
0.2  299642
0.3  312935
...
23.2  3
23.3  1
23.9  6
24.0  1
27.4  1
```

변수들의 분포 확인 결과,  
Y 변수인 전력기상지수와  
X 변수인 풍속의 분포를 통해  
**결측치와 이상치가 존재함을 확인**



이때, Y와 X의 결측치는  
**다르게** 처리해야 함!

한국전력공사에서는 결측치를 -99로 표현



# 데이터 설명 및 EDA

## 결측치 처리 | 반응변수

설명변수				반응변수	
electric_train.num	electric_train.tm	electric_train.stn	...	electric_train.nph_ta_chi	electric_train.elec
11412	2020-02-10 12:00:00	899	...	7.9	-99
14258	2021-03-10 09:00:00	138	...	10.5	-99
15735	2022-08-28 17:00:00	136	...	26.2	-99
18680	2020-10-27 23:00:00	511	...	12.8	-99
19724	2020-05-11 20:00:00	99	...	15.9	-99

반응변수의 결측치는 imputation method보다  
deletion method가 더 정확하기에 전부 제거하기로 결정



# 데이터 설명 및 EDA

## 결측치 처리 | 반응변수

설명변수				반응변수	
electric_train.num	electric_train.tm	electric_train.stn	...	electric_train.nph_ta_chi	electric_train.elec
11412	2020-02-10 12:39:00	899	...	7.9	-99
14258	2021-03-10 09:00:00	138	...	10.5	-99
15735	2022-06-29 17:55:00	...	...	...	-99
18680	2020-02-10 25:00:00	...	...	12.8	-99
19724	2020-02-10 20:00:00	...	...	15.9	-99

It is common to discard observations having missing Y.

“반응변수에 imputation method을 활용하게 된다면 overfitting과 같은 모델링 전반에 문제가 생길 수 있기에 제거를 하는 것이 바람직합니다.”



2024년 5월 7일 통계적 모델링과 머신러닝 수업

반응변수의 결측치는 imputation method보다 deletion method가 더 정확하기에 전부 제거하기로 결정





# 데이터 설명 및 EDA

## 결측치 처리 | 설명변수

설명변수				반응변수	
electric_train.num	electric_train.tm	electric_train.stn	...	electric_train.nph_ws_10m	electric_train.elec
8994	2022-05-31 00:00:00	261	...	-99	95.30
10069	2022-05-31 00:00:00	162	...	-99	88.79
10069	2022-08-05 17:00:00	162	...	-99	147.08
319271	2022-05-31 00:00:00	294	...	-99	96.90
...	...	...	...	...	...

설명변수 중 풍속에 총 170개의 결측치가 있음을 확인하였음



# 데이터 설명 및 EDA

## 결측치 처리 | 설명변수

설명변수				반응변수	
electric_train.num	electric_train.tm	electric_train.stn	...	electric_train.nph_ws_10m	electric_train.elec
8994	2022-05-31 00:00:00	261	...	2.15	95.30
10069	2022-05-31 00:00:00	162	...	3.15	88.79
10069	2022-08-05 17:00:00	162	...	2.45	147.08
319271	2022-05-31 00:00:00	294	...	2.95	96.90
...	...	...	...	...	...

시계열 데이터이므로 앞 1시간, 뒤 1시간 풍속의 평균으로 대체하였음

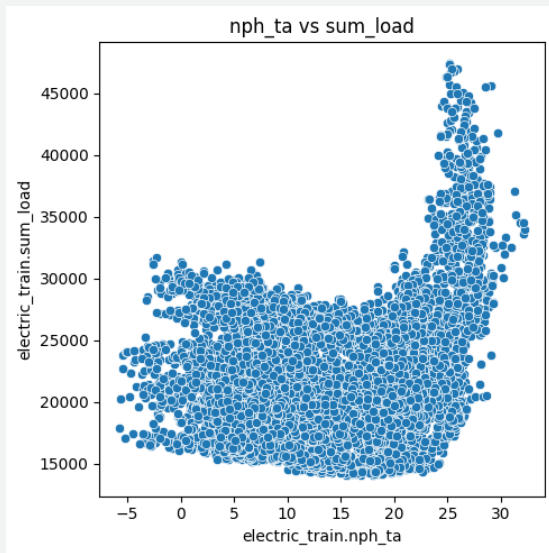
Ex) 2022-05-31 00:00:00에 결측치가 있다면

2022-05-30 23:00:00과 2022-05-31 01:00:00의 평균 풍속

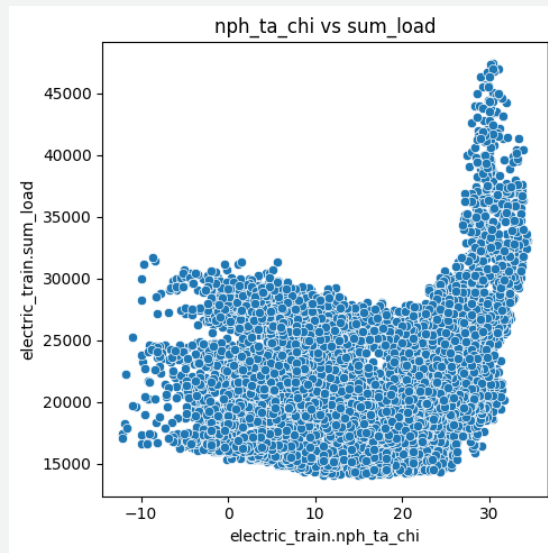


# 데이터 설명 및 EDA

## 기상변수와 전력수요의 관계



기온



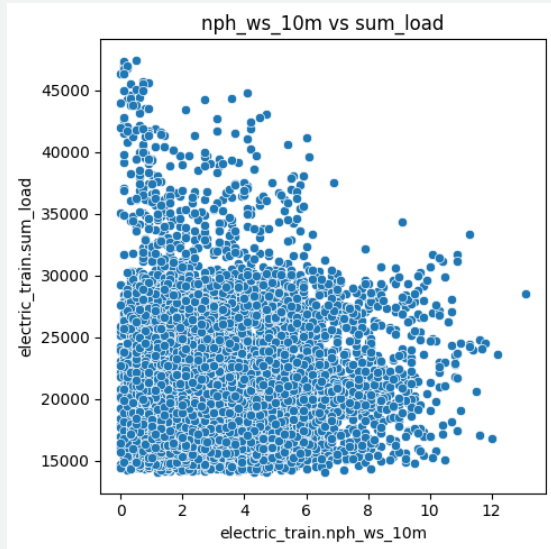
체감온도

공동주택의 전력수요는 기상변수 중 **기온과 체감온도**와 가장 밀접한 관계를 보이며,  
여름철과 겨울철의 냉난방 전력수요의 변화가 주요 원인인 것으로 파악됨

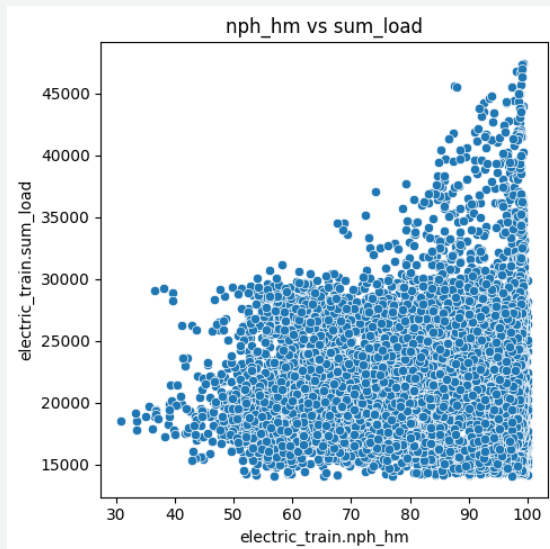


# 데이터 설명 및 EDA

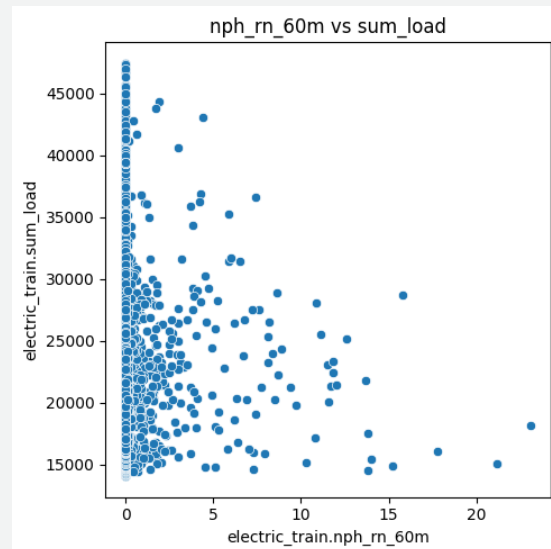
## 기상변수와 전력수요의 관계



풍속



상대습도



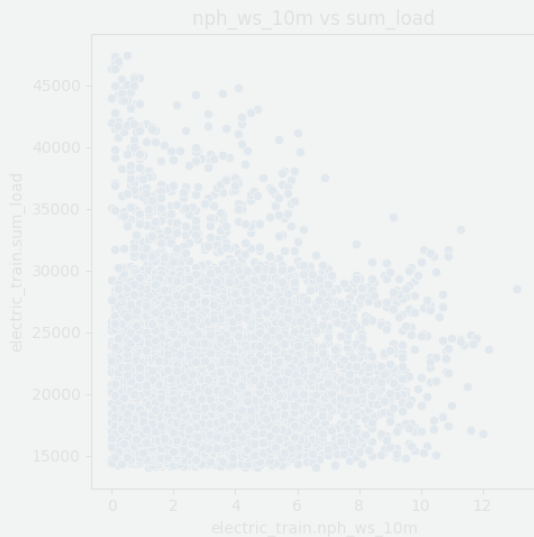
강수량

다른 기상변수들은 기온과 체감온도에 비해 덜 밀접한 관계가 보임을 알 수 있음

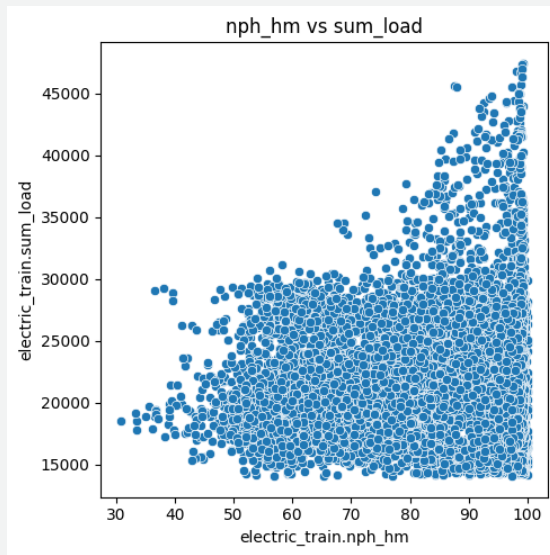


# 데이터 설명 및 EDA

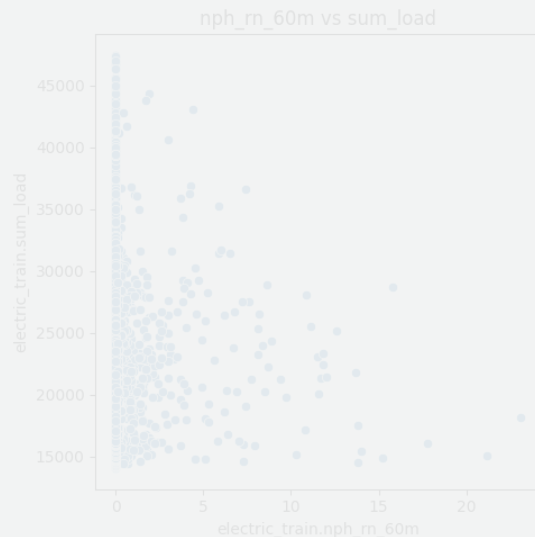
## 기상변수와 전력수요의 관계



풍속



상대습도



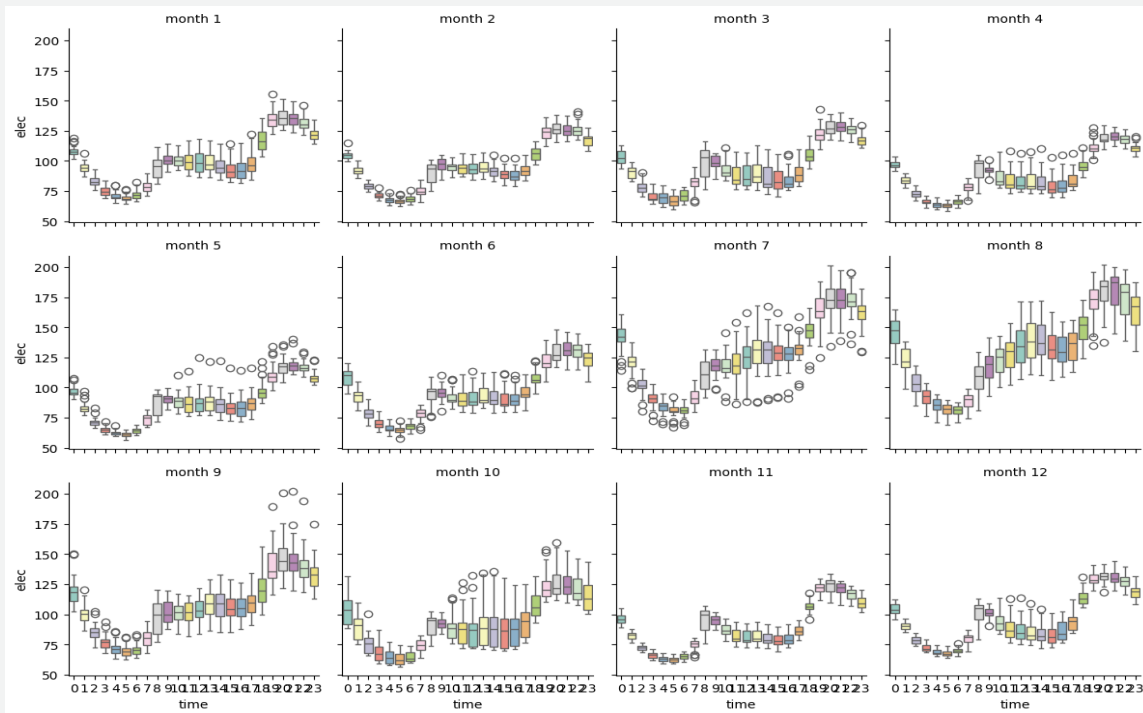
강수량

그러나, 상대습도도 기온과 체감온도와 마찬가지로  
특정 시점을 기준으로 급격히 증가하는 비선형적 관계를 보인다는 특이점도 발견할 수 있음



# 데이터 설명 및 EDA

## 월별 전력수요 분포



특정 격자번호 기준 월별 전력수요 시간별 분포 Boxplot

공동주택의

일 최대 전력수요는 대체로

약 21시경에 발생하며

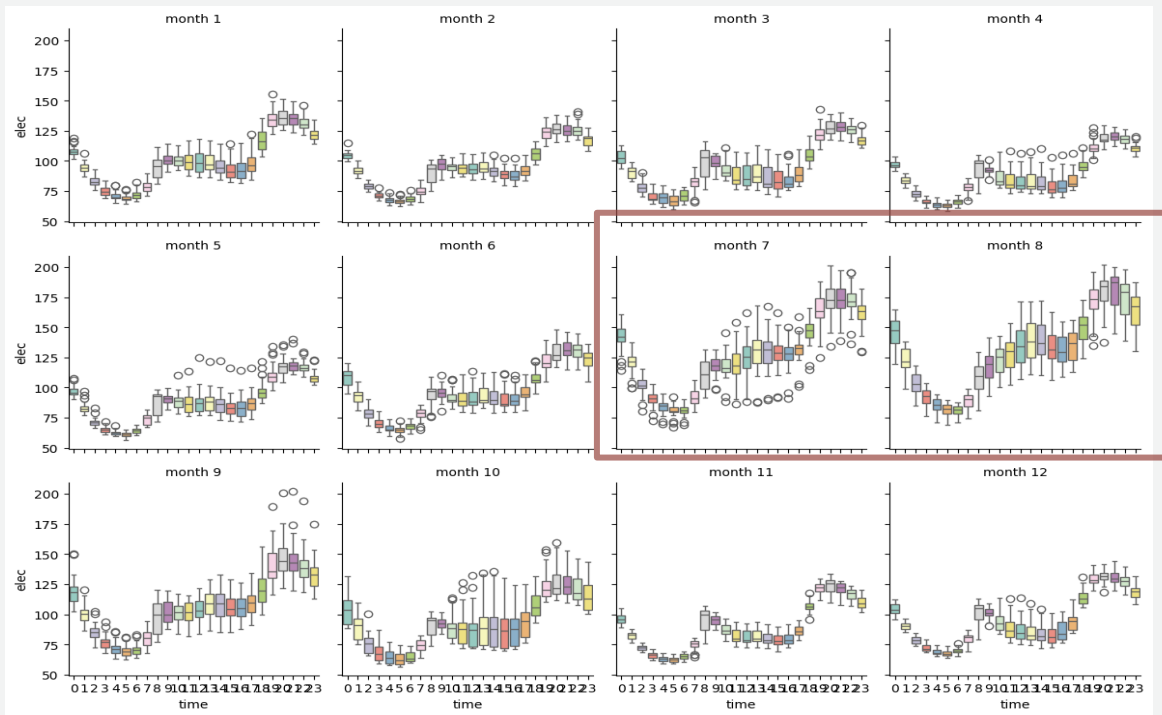
일정한 패턴이 보임을

알 수 있음



# 데이터 설명 및 EDA

## 월별 전력수요 분포

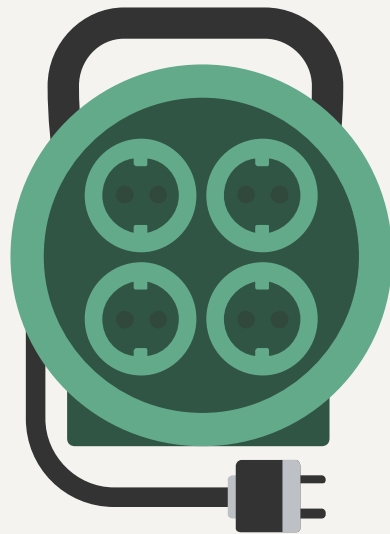


그러나, 하계(7,8월)에는  
시간대별 전력수요의  
변동폭이 급격히  
커지는 현상 존재

특정 격자번호 기준 월별 전력수요 시간별 분포 Boxplot

03

## 모델 세분화







# 모델 세분화

## 예측 모델 구상

계절과 지역에 따른 모델 세분화를 통한  
전력기상지수 예측 최적모델 개발

### 계절별 구분

봄 / 여름 / 가을 / 겨울 시기

### 지역별 구분

기상 및 전력지표 관측 위치

계절/지역별로 세분화한 모델마다  
데이터셋을 추출하여 개별적으로 예측



# 모델 세분화

## 예측 모델 구상

모델 구축에 앞서

계절 및 지역 구분에 대한 기준 확립 필요



### 계절별 구분

봄 / 여름 / 가을 / 겨울 시기

후보 ① 한 계절 = 3개월로 간주

(3, 6, 9, 12월에 계절 변화)

후보 ② 월평균기온 기반 분산분석

후보 ③ 일평균기온 기반 자연계절 구분

(by 기상청)

### 지역별 구분

기상 및 전력지표 관측 위치

후보 ① 행정구역으로 구분

후보 ② 기온분포에 따른 지역 클러스터링



# 모델 세분화

## 모델 세분화 | 계절별 구분

### 계절 길이

통상적인 기준에 따라

**3개월로 간주**

ANOVA 활용

월평균기온 기반 분산분석

기상청 분석에 따른

일평균기온 기준



봄 (3~5월)

여름 (6~8월)

가을 (9~11월)

겨울 (12~2월)

### 계절 시작일 산출 기준

계절 시작일 산출 기준	
봄 시작일	3.1.
여름 시작일	6.1.
가을 시작일	9.1.
겨울 시작일	12.1.



# 모델 세분화

## 모델 세분화 | 계절별 구분

### 계절 길이

통상적인 기준에 따라  
3개월로 간주

ANOVA 활용  
월평균기온 기반 분산분석

기상청 분석에 따른  
**일평균기온 기준**



계절 시작일 산출 기준	
봄 시작일	일평균기온이 5°C 이상 올라간 후 다시 내려가지 않는 첫날
여름 시작일	일평균기온이 20°C 이상 올라간 후 다시 내려가지 않는 첫날
가을 시작일	일평균기온이 20°C 미만으로 내려간 후 다시 올라가지 않는 첫날
겨울 시작일	일평균기온이 5°C 미만으로 내려간 후 다시 올라가지 않는 첫날



# 모델 세분화

## 모델 세분화 | 계절별 구분

### 계절 길이

통상적인 기준에 따라  
3개월로 간주

ANOVA 활용  
월평균기온 기반 분산분석

기상청 분석에 따른  
일평균기온 기준



계절 시작일 산출 기준	
봄 시작일	3.1.
여름 시작일	5.31.
가을 시작일	9.26.
겨울 시작일	12.4.

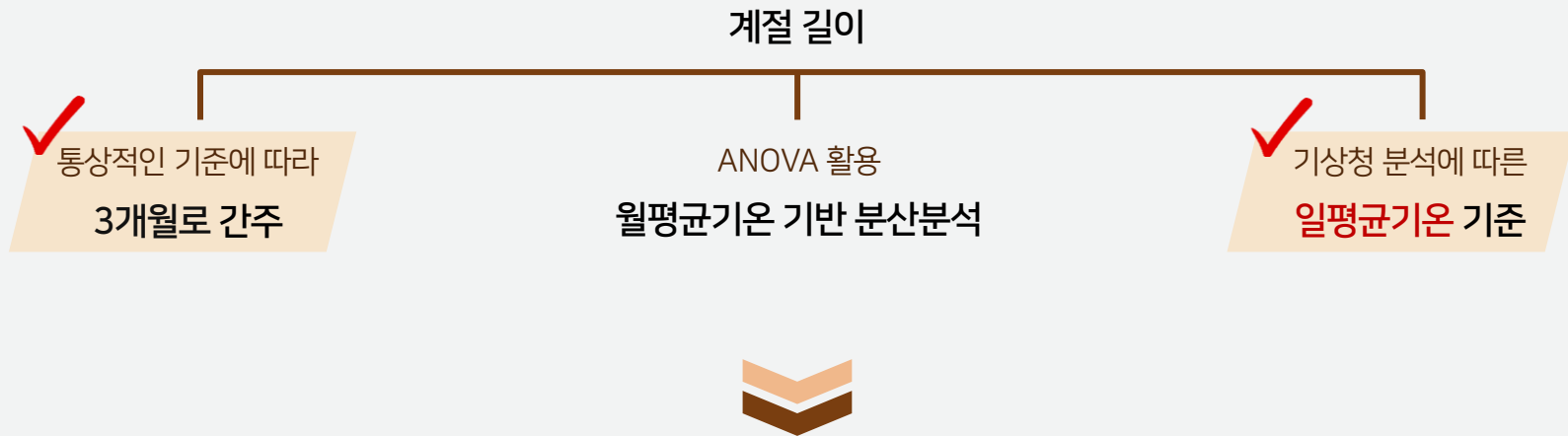


상당 부분  
직관과 부합



# 모델 세분화

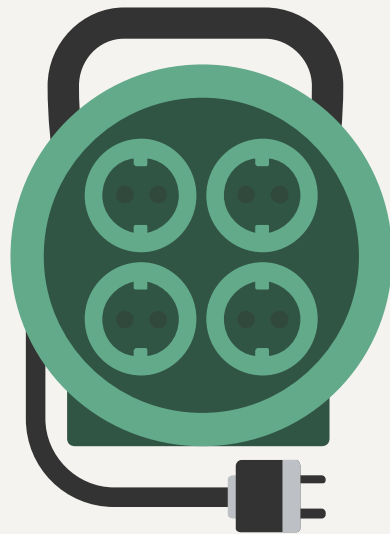
## 모델 세분화 | 계절별 구분



검증에 실패한 경우를 제외한 계절 길이 구분 방법을 각각 적용한 모델 생성

04

## 모델 성능 비교





# 모델 성능 비교

## 모델링 전제 조건

파생변수 유무에 따른 모델 성능을 비교하기 위해  
파생변수를 제외한 **기본변수만을 이용한 모델링** 일차적으로 진행

⋮

### 기본변수

'electric\_train.nph\_ta', 'electric\_train.nph\_hm', 'electric\_train.nph\_ta\_chi',  
'electric\_train.nph\_ws\_10m', 'electric\_train.nph\_rn\_60m',  
'electric\_train.hh24', 'electric\_train.num', 'electric\_train.weekday'





# 모델 성능 비교

## 3차 다항회귀모델

체감온도를 활용한 3차 다항회귀모델의 상관계수가 가장 높았다는  
선행 연구 결과에 따라, 3차 다항회귀모델 개발

⋮

### 2.4 전력기상지수 예측 모델 개발

전력기상지수 예측값 산출을 위한 예측 모델은 총 3종의 분석 모델의 비교분석을 통해 개발되었으며 활용된 분석모델은 아래와 같다.

- 모델 1 : 체감온도를 활용한 3차 다항회귀모델
- 모델 2 : 기온, 습도, 풍속성분을 활용한 다중회귀모델
- 모델 3 : 기온, 습도, 풍속성분을 활용한 랜덤 포레스트 모델

기상 데이터와 공동주택의 부하율은 <그림 3>과 같이 비선형적인 관계에 있으므로 연중 전체의 데이터를 사용하지 않고 아래와 같이 각 모델별로 예측일 주변 기간의 과거년도 데이터를 추출하여 모델 학습에 활용하였다.

<표 3> 예측 지역의 규모에 따른 상관계수

구 분	도/특별시/ 광역시	시/군/구	읍/면/동	5km 격자별
모델 1	0.947	0.919	0.932	0.940
모델 2	0.916	0.892	0.906	0.912
모델 3	0.917	0.894	0.905	0.912

활용된 3개 분석모델 모두 최소 서비스 제공 단위인 읍/면/동에서 상관계수가 0.9 이상으로 비교적 높은 상관성을 보였으며, 그 중에서 모델 1에서 산출된 전력기상지수가 모든 서비스 단위에서 전력부하와 높은 상관계수를 보였다. 그에 따라 해당 모델을 기반으로 대국민 서비스를 개발하여 제공하고 있다.



# 모델 성능 비교

## 3차 다항회귀모델

### 다항회귀모델

독립변수가 1차항으로 구성된 것이 아닌, 2차항, 3차항 등으로 구성되어 있는 회귀모델

$$\hat{y} = b_0 + b_1x_i + b_2x_i^2 + \dots + b_px_p^p$$



체감온도, 체감온도의 제곱, 체감온도의 세 제곱을 파생변수로 추가하여 다항회귀모델 생성



# 모델 성능 비교

## 3차 다항회귀모델 | 모델성능비교

통상 기준 계절, 기온 기준 지역 클러스터링

Model	MAE	MSE	상관계수
spring_cluster_2	9.244	119.437	0.798
...	...	...	...
summer_cluster_2	13.727	90.068	0.839
...	...	...	...
autumn_cluster_2	10.213	151.108	0.793
...	...	...	...
winter_cluster_2	9.78	136.935	0.797

평균 상관계수 : 0.798

평균 MAE : 10.999

평균 MSE : 187.619



# 모델 성능 비교

## 3차 다항회귀모델 | 모델성능비교



모델 성능의 평균을 비교해보았을 때, 체감온도로 지역을 클러스터링한 것과  
기온으로 지역을 클러스터링한 것이 유의미한 차이를 보임

상관계수 : 체감온도 < 기온

MAE, MSE : 체감온도 > 기온



모델링 소요시간 등을 고려하여, 앞으로의 모델링에서는

**기온**으로 지역을 클러스터링한 것만을 사용하기로 결정!



# 모델 성능 비교

## XGBoost, LGBM, RF 모델 성능 비교

### XGBoost

평균 상관계수 : 0.962

평균 MAE : 4.776

평균 MSE : 49.197

### LGBM

평균 상관계수 : 0.963

평균 MAE : 4.754

평균 MSE : 47.340

### RF

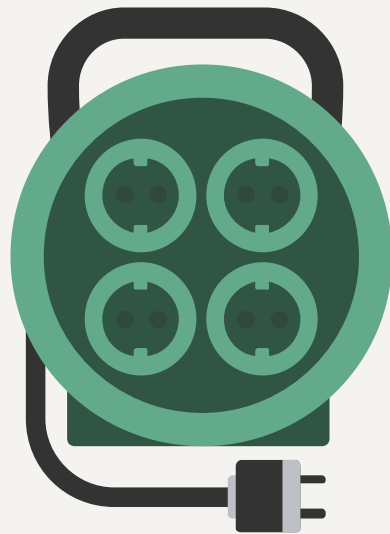
평균 상관계수 : 0.958

평균 MAE : 4.999

평균 MSE : 54.624

05

파생변수 생성





# 파생변수 생성

## 시간 관련 파생 변수

`electric_train.year`

datetime 형식으로 변환된 'electric\_train.tm'에서 연도를 정수형으로 추출하여 변환한 변수

`electric_train.month`

datetime 형식으로 변환된 'electric\_train.tm'에서 월을 정수형으로 추출하여 변환한 변수

`electric_train.day`

datetime 형식으로 변환된 'electric\_train.tm'에서 일을 정수형으로 추출하여 변환한 변수



# 파생변수 생성

## 시간 관련 파생 변수

sin\_hour

'electric\_train.hh24' 열에 기반한 **sin** 값을 계산한 변수

cos\_hour

'electric\_train.hh24' 열에 기반한 **cos** 값을 계산한 변수



'electric\_train.hh24' 는 시간을 1~24시로 표현한 변수이며  
푸리에 특징(Fourier Features)를 사용하여 표현





# 파생변수 생성

## 시간 관련 파생 변수

electric\_train.week

'electric\_train.tm' 열에 주 단위로 1~52의 숫자를 부여한 변수

⋮



격자 4821의 2021년 1월 1일의 주 변수 1

electric_train.num	electric_train.tm	electric_train.sum_load	electric_train.n_mean_load	...	electric_train.week
4821	2021-01-01 1:00	751.32	68.60645	...	1
4821	2021-01-01 2:00	692.6	68.60645	...	1
4821	2021-01-01 3:00	597.48	68.60645	...	1
4821	2021-01-01 4:00	553.48	68.60645	...	1



# 파생변수 생성

## 지역 관련 파생 변수

지역마다 온도의 분포가 다르기 때문에 추가한 변수

loc\_label

AWS 좌표를 바탕으로 지역을 라벨링한 변수

⋮

0 서울, 1 부산, 2 대구, 3 인천, 4 광주, 5 대전, 6 울산, 7 세종, 8 경기, 9 강원, 10 충북, 11 충남, 12 전북, 13 전남, 14 경북, 15 경남, 16 제주

electric_train.num	electric_train.tm	electric_train.stn	...	electric_train.elec	loc_label
4821	2021-12-01 21:00:00	884	...	128.51	16
...	...	...	...	...	...
20947	2021-12-27 21:00:00	138	...	-145.35	9



# 파생변수 생성

## 기상 관련 파생 변수

electric\_train.di

날씨에 따라서 사람이 느끼는 불쾌감의 정도를 간단한 수식으로 표현하는 **불쾌지수**를 표현한 변수

⋮

산출식

$$0.81 * \text{섭씨온도} + 0.01 * \text{상대습도}(\%) \quad (0.99 * \text{섭씨온도} - 14.3) + 46.3$$



# 파생변수 생성

## 기상 관련 파생 변수

electric\_train.cw

아침 최저기온이 영하 12도 이하인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **한파**를 나타낸 변수

electric\_train.hw

오후 6시부터 오전 9시 동안 25도 이상인 행과 오전 9시부터

오후 6시 동안 33도 이상인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭염**을 나타낸 변수

electric\_train.rain

강수량이 30 이상인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭우**를 나타낸 변수

electric\_train.storm

풍속이 13.9m/s인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭풍**을 나타낸 변수



# 파생변수 생성

## 기상 관련 파생 변수

electric\_train.cw

아침 최저기온이 연한 12도 이하인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **한파**를 나타낸 변수

**binary(0,1) 인코딩 변수를 새로 만든 이유?**

electric\_train.hw

오후 6시부터 오전 9시 동안 25도 이상인 행과 오전 9시부터

오후 6시 동안 33도 이상인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭염**을 나타낸 변수

기상 지표 중 이상치로 보이는 극단적인 값들이 다수 관찰됨

electric\_train.rain

이를 보정하기 위해 0과 1 값을 부여하는 과정을 거치며

강수량이 30 이상인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭우**를 나타낸 변수

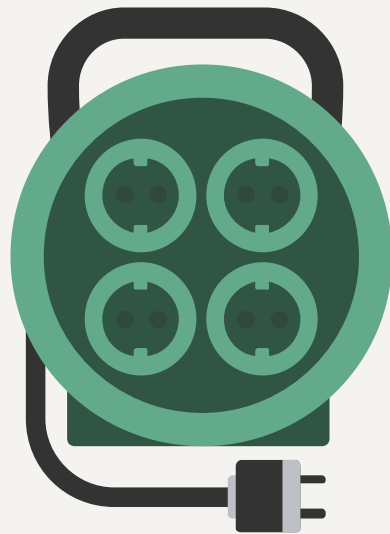
**예측 모델의 성능 향상**

electric\_train.storm

풍속이 13.9m/s인 행에 1을 부여하고 그렇지 않은 행에 0을 부여하여 **폭풍**을 나타낸 변수

06

모델 개선





# 모델 개선

## Auto ML

AutoML (Automated Machine Learning)은 시간 소모적이고  
반복적인 기계 학습 모델 개발 작업을 자동화하는 프로세스

Model	MAE	MSE	RSME	...	R2
Catboost	4.2997	39.6584	6.2975	...	0.9360
Knn	4.2816	39.9091	6.3174	...	0.9356
Lasso	13.1949	319.4010	17.8718	...	0.4848
Huber	12.8749	324.6681	18.0185	...	0.4763
Rf	2.7133	20.9377	4.0927	...	0.7662



# 모델 개선

## 파생변수 선택

여러가지 파생변수 조합을 고려해본 결과, 다음의 변수 조합에서 성능이 가장 괜찮은 것으로 판단!



불쾌지수(electric\_train.di), 폭염(electric\_train.hw), sin\_hour, cos\_hour,  
한파(electric\_train.cw), 년도(electric\_train.year), 월(electric\_train.month)  
변수를 추가함





# 모델 개선

## Case 1 | 계절 및 지역 세분화, 파생변수 추가

파생변수를 추가한 후에 계절 및 지역으로 세분화한 모델의 성능을 비교

CatBoost, 통상적인 기준의 계절

Model	MAE	MSE	상관계수
spring_cluster_2	2.902	14.850	0.9786
...	...	...	...
summer_cluster_2	6.886	90.068	0.9587
...	...	...	...
autumn_cluster_0	3.774	27.5005	0.9676
...	...	...	...
winter_cluster_2	3.710	24.1285	0.9680



# 모델 개선

## Case 2 | 계절 세분화, 파생변수 추가

클러스터링을 통한 지역 세분화 대신 계절 세분화만 진행  
지역에 대한 차이는 loc\_label 변수를 추가하여 보정함

**LGBM**, 통상적인 기준의 계절

Model	MAE	MSE	상관계수
Spring_model	4.03	28.192	0.9571
Summer_model	9.18	156.532	0.9289
Autumn_model	5.05	45.1916	0.9426
Winter_model	4.39	32.24	0.9573



# 모델 개선

## Case 3 | 전체 데이터셋 사용, 파생변수 추가

모델을 세분화 하지 않고, 전체 데이터 셋을 이용하여 파생변수 등을 활용하여 모델을 학습 시킴

Model	MAE	MSE	상관계수
CatBoost	4.3086	38.35	0.9713
XGBoost	4.601	41.22	0.9683
LGBM	4.75	41.98	0.9659
Random Forest	4.89	42.31	0.9645
KNN	5.59	49.35	0.9522



# 모델 개선

## Case 3 | 전체 데이터셋 사용, 파생변수 추가

모델을 세분화 하지 않고, 전체 데이터 셋을 이용하여 파생변수 등을 활용하여 모델을 학습 시킴

CatBoost Regressor가 매우 좋은 성능을 보임을 확인할 수 있음!!

Model	다른 모형들도 CatBoost와 비슷하게 좋은 성능을 보임		상관계수
CatBoost	4.3086	39.35	0.9713
XGBoost	4.601	41.22	0.9683
LGBM	4.75	양상불 방법을 고려해보기로 함!	
Random Forest	4.89		
KNN	5.59		



# 모델 개선

## 앙상블

앙상블 (ensemble)은 여러 개의 모델을 결합하여 하나의 모델보다 더 나은 성능을 얻는 방법



CatBoost, LGBM, XGBoost 3개의 모델을 **Voting Regressor** 기법을 이용하기로 함!



# 모델 개선

## 앙상블

CatBoost, LGBM, XGBoost 3개의 모델을 **Voting Regressor** 기법을 이용하기로 함!

Model	MAE	MSE	상관계수
Ensemble	4.303	37.33	<b>0.9717</b>

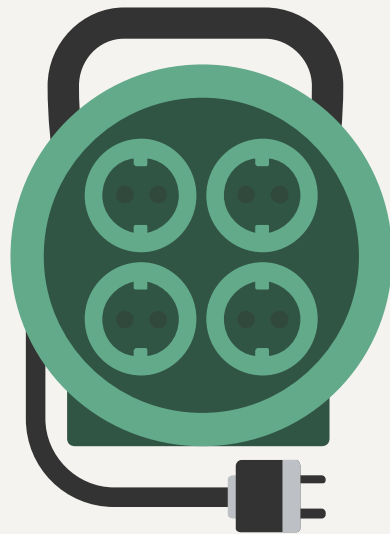


가장 좋은 성능을 보임!



07

## 결과 해석





# 결과 해석

## 최종 모델 선정

### 모델링 시나리오

계절세분화  
지역세분화  
파생변수 추가

계절세분화  
파생변수 추가

✓  
전체 데이터셋  
파생변수 추가  
(계절 변수 포함)



각 경우마다의 성능평가를 통해  
“전체 데이터셋 + 계절변수를 포함한 파생변수 추가” 경우를  
최종 모델링 시나리오로 선정





# 결과 해석

## 최종 모델 선정

### 모델링 시나리오

계절세분화  
지역세분화  
파생변수 추가

계절세분화  
파생변수 추가

✓  
전체 데이터셋  
파생변수 추가  
(계절 변수 포함)



각 경우마다의 성능평가를 통해  
좋은 성능을 보인 상위 3개 모델을  
"전체 데이터셋 + 계절변수를 포함한 파생변수 추가" 경우를  
다시 앙상블한 **VotingRegressor** 사용  
최종 모델링 시나리오로 선정



# 결과 해석

## 한계 및 의의

### 모델 세분화 전후 예측결과 비교

지역/계절 세분화 모델링 결과


전체 데이터셋 모델링 결과

ex) XGBoost

	MAE	MSE	상관계수
봄	3.671	23.117	0.967
여름	8.22	129.1	0.945
가을	4.385	36.407	0.956
겨울	4.145	29.772	0.96

ex) XGBoost

	MAE	MSE	상관계수
봄	3.874	27.081	0.958
여름	8.891	154.704	0.924
가을	4.654	38.115	0.95
겨울	4.327	31.686	0.958

기온에 따른 계절/지역 세분화가  
전반적인 성능 **향상**에 영향을 끼침을 확인할 수 있음 



# 결과 해석

## 한계 및 의의

### 모델 세분화 예측결과

ex) LGBM 예측 성능

	MAE	MSE	상관계수
봄	3.665	22.753	0.968
여름	7.923	118.591	0.949
가을	4.341	35.112	0.959
겨울	4.088	28.653	0.962

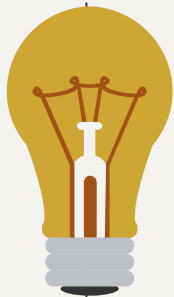
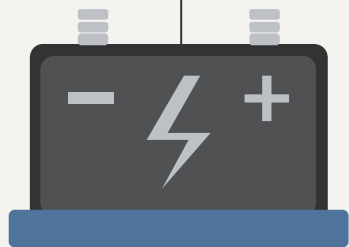
### 여름 클러스터 예측 성능이 상대적으로 낮음

세분화 모델 전체 성능평가 시 전체 지표의 평균치로 산정  
여름 시점에 대한 예측력이 전체 성능을 깎아내리는 문제 발견



**여름** 시점의 예측력을 높이는

새로운 전처리 및 모델링 필요



지금까지 회귀분석팀이었습니다!!

감사합니다!