Final Report

DATA MINING

212STG02 고정욱

212STG04 김이현

212STG18 예지혜

목차

I. 문제 정의

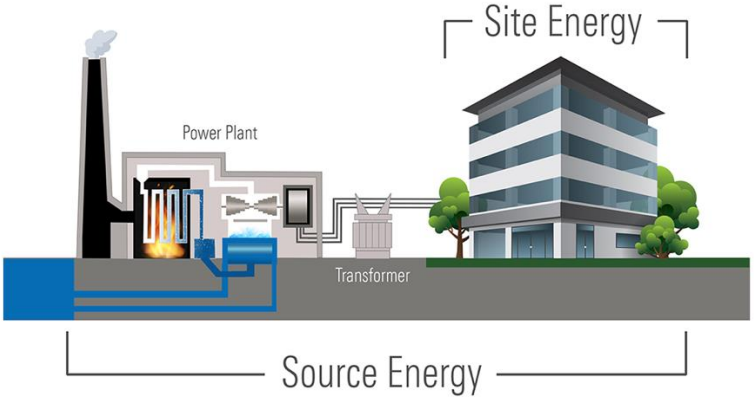
에너지 문제는 유엔의 지속가능발전목표(SDG)로 선정된 지 수년 째다. 에너지 소비량 증가는 최근 들어 심각한 사회문제로 대두되고 있다. 전 세계 에너지 소비량의 약 40%가 건물에서 소비되고 있다고 한다. 이런 상황에서 그린뉴딜 탄소중립 실현의 핵심 사업으로 건물 에너지 관리 시스템 (BEMS)이 주목받고 있다. 특히 세계 각국이 탄소중립 의지를 보이면서, BEMS(Building Energy Management System)은 건물 내 주요 공간과 설비에 센서를 부착하여, 실시간으로 에너지 사용 데이터를 수집 및 분석하여 에너지소비 절감과 건물의 쾌적한 실내환경 유지에 활용하는 최첨단 ICT 시스템이다. 전기 사용의 절대량을 줄이는 것이 아니라, 효율적으로 에너지를 사용하는 것이 보다 중요해지면서, 건물을 "운영"하는 차원에서 지속적인 절감효과를 유지할 수 있는 BEMS 가 각광받고 있는 것이다.

데이터와 인공지능(AI)을 통해 도시에 '숨어 있는' 에너지 문제를 찾아내기도 한다. 특히 건물 에너지를 절감하기 위한 도시 단위 에너지 진단을 통해 에너지 낭비가 심하거나 절감 효과가 높은 건물을 찾아내는 '에너지 맵'을 만들어 제공하고 있다. 에너지 맵은 수많은 건물 중 비용 투자 대비 에너지 절감, 탄소 저감 효과가 높은 건물을 찾아내는 기술이다. 건물의 크기·위치·용도·에너지 사용량 등 정형적인 정보와 날씨 변화, 이용자 수, 전력 이용 패턴 등 비정형적인 정보를 분석한다. 에너지 낭비가 심하거나 절감이 필요한 건물을 찾고 에너지 효율을 높이는 방법을 제시한다. 이를 통해 불필요하거나 과도한 설비투자를 방지하고, 투자 대비 에너지를 최대한 절감하고 탄소를 감축하는 효과를 거둘 수 있다. 이처럼 사회문제로 대두된 전력 소비 감축을 위해, 다양한 기술들이 활용되고 있으며, 기술을 활용해 알아낸 전력 과잉 소비 패턴은 실질적인 감축으로 이어지고 있다.

본 분석에서는 건물이 위치한 지역의 건물 특성과 기후 및 날씨 변수를 설명하는 변수들로 구성된 데이터를 활용하여 각 건물의 1 년간 에너지 소비량을 예측하고자 한다. 에너지 소비량을 예측하는 목적에는, 소비량의 절대적인 값을 더 정확하게 예측하는 것도 있겠지만, 그 이면에는 건물이 갖는 어떤 특징적인 요소들이 에너지 발생량에 더 큰 영향을 주는지를 파악하는 데 있다. 따라서 에너지 소비량 예측 모형을 설계하는 것은, 에너지 소비량에 가장 영향력이 큰 변수들을 파악함으로써, 건물 관리 차원에서 에너지 효율 개선에 도움이 될 수 있으며, 미래 에너지 소비량을 예측함으로써 효과적인 전력 사용량 관리에 도움을 줄 것으로 기대할 수 있다. 나아가 최근 많은 기업들이 ESG 경영에 투자하고 있으며, 국가 차원에서도 에너지 감축을 권장하고 있는 상황이므로, 에너지 소비량 감소 정책 입안에도 기여할 것으로 본다.

II. 데이터 설명

가장 대표적으로 에너지 사용량을 측정하는 2 가지 기준에는, source EUI 와 site EUI 가 있다. EUI 란, Energy Use Intensity 의 약자로, square foot 단위면적에 대한 연간 전력 소비량을 표현한다. 빌딩의 연간 총 전력 소비량을 빌딩의 연면적으로 나눈 값으로 산출된다. Source EUI 는, 해당 건물에서 사용되는 에너지 소비량만이 아니라, 그 에너지를 공급하기 위해 건설 현장에서 전력회사 및 기타 기업이 소비하는 에너지 소비량까지 계산해 더한 값이다. Site EUI 란, 연간 평방 피트당 사용된 에너지 양이다. 1 년 동안 건물에 의해 소비된 총 에너지를 kBtu 또는 GJ 로 측정하여, 건물의 총 바닥 면적(평방 피트 또는 평방 미터로 측정)으로 나누어 산출한다. Site EUI 는 연중 내내 측정되며, 건물 공간에 난방이나 냉방을 가하면서 발생하는 에너지 소비량과, 전기, 수도 사용량에서 발생하는 에너지 소비량도 포함된다. 이외에도, 건물 설비나 기타 전기 제품들로 소비되는 요소들에 대해서 모두 측정되기 때문에, 건물의 특징적인 요소들이 서로 영향을 주고 받을 것으로 보인다. 따라서 변수들간 관계와, 여러 가설을 검정해보는 과정을 통해 전력량에 영향을 미치는 요인을 분석해보고자 한다.



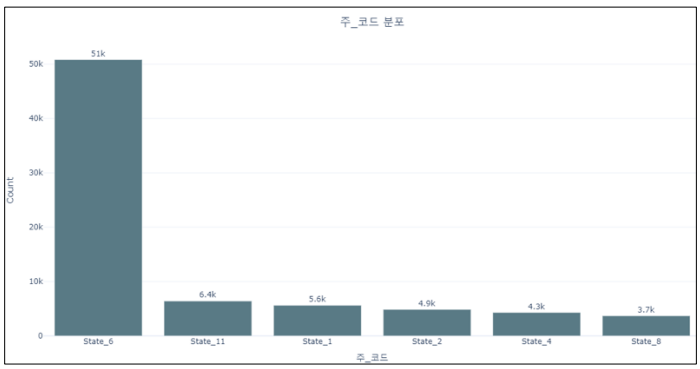
<그림 1> source energy vs site energy

본 분석에서는 데이터 셋에 내재한 site EUI 를 예측 타겟 변수로 선정하였다. 설명변수로 사용될 기타 변수들로는 총 61 개의 변수들이 존재한다. 변수들을 큰 범주에서 나누어 보면, 건물과 관련한 특성 , 월별 기온 정보, 그리고 날씨 정보로 크게 3 가지 범주로 나누어 볼 수 있다. 각 범주에는 순서대로 8 개, 36 개, 16 개의 설명변수들이 존재한다. (<표 1> 참고) 우선 건물 정보로는, 건물의 에너지 발생량을 관찰한 관측년도, 건물의 위치, 건물 주거 상업 여부, 건물 주 용도, 건물 연면적, 건물 건축년도, 건물 에너지 등급, 건물 위치의 고도가 존재한다. 변수들 중 관측년도와, 건물의 위치(주, State)는 마스킹 되어있는 변수로, 해당 변수가 갖는 영향력을 깊이 파악하는데 한계가 있었다. 또한 36 개의 월별 기온 정보는, 건물이 위치한 곳의 최저, 최고, 평균 기온을 월별로 제시한 변수이다. 또한 날씨정보로는, 냉각이 필요한 날을 기준으로 정하고, 기준을 만족하는 날들에 한해 그날의 화씨 도수를 모두 합한 변수와, 난방에 대해서도 이를 동일하게 적용한 변수가 각각 존재한다. 또한, 섭씨 기준 -1 도, -6 도, -12 도, -17 도 이하로 떨어진 총 일수와, 섭씨 26 도, 32 도, 37 도, 43 도를 넘어선 총 일수의 정보를 담고 있는 변수도 존재한다. 최대 풍향, 최대 풍속 그리고 안개 일수 등 기타 기상 정보와 관련한 변수들도 존재했지만, 결측치가 80%를 넘어 의미 있는 분석으로 이어지기는 힘들 것으로 보인다. 이외로 연간 강수량, 강설량, 적설량 변수와, 연간 평균 기온을 분석에 함께 활용할 예정이다. 각 변수들에 대한 설명과 함께 주목해 볼만한 특장점들은 III. 탐색적 데이터 분석에서 소개하도록 하겠다.

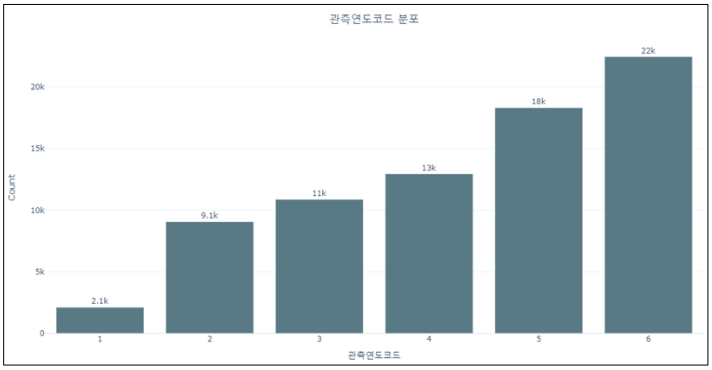
III. 탐색적 데이터 분석

a. 건물 정보 변수

- 건물의 위치 정보 및 에너지 소비량 관측년도



<표 1> 건물의 위치정보 주(State)를 마스킹한 정보



<표 2> 건물의 전력소비량 관측년도(Year)를 마스킹한 정보

먼저, 건물 정보 변수로 관측 연도와 건물위치(주, State)를 살펴보면, 건물 위치 변수는 'State_1', 'State_2', 'State_4', 'State_6', 'State_8', 'State_11'까지 총 6 개의 주에 대한 정보가 주어졌다. 그 중 'State_6'의 도수가 5 만개로 다른 주들은 평균 5000 개의 관측치를 갖고 있음에 비해 상대적으로 굉장히 많은 관측치가 하나의 주에 몰려 있음을 알 수 있다. 또한, 건물의 전력소비량을 관측한 시점에 대한 정보 또한 마스킹된 코드 값으로 주어졌다. 관측년도는 1 부터 6 까지 6 개의 범주로 주어졌는데, 큰 수의 범주에 가까울수록 더 많은 관측치를 보유함을 확인할 수 있다.

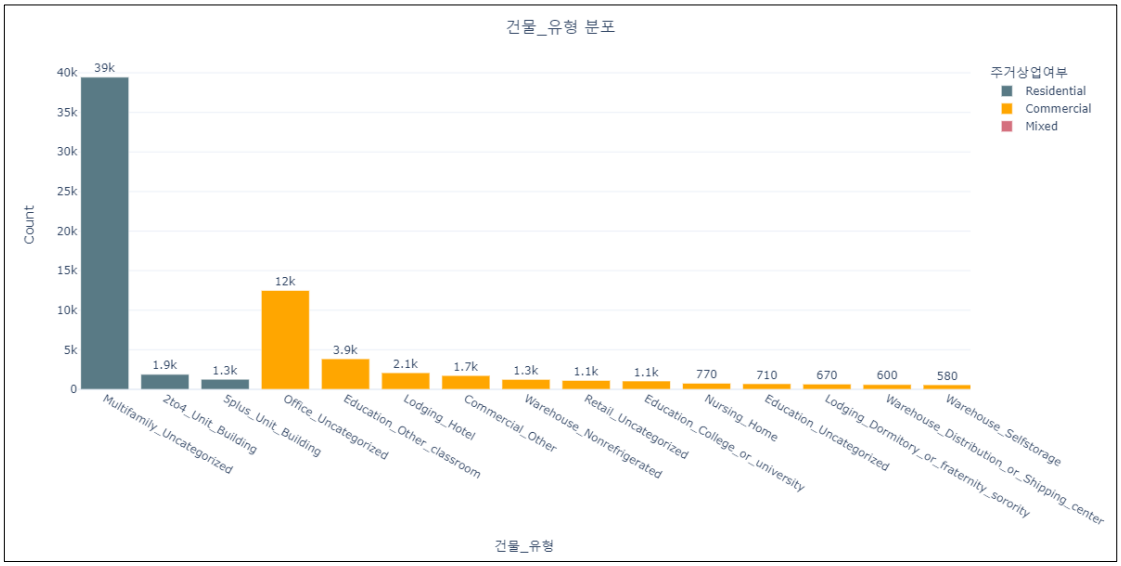
- 건물의 주 사용 용도

'건물 유형'은 건물 주 사용 용도로 유형화한 변수로, 총 61 개의 하위 분류를 갖는다. <표 3> 주거용 빌라, 사무실, 교육기관이 가장 많은 관측치를 보유하고 있다. <차트 1> '주거상업여부' 변수는, '건물 유형' 변수를 주거, 상업, 주상복합 세 가지 분류로 재범주화하여 만든 변수로, 다음과 같은 분포를 갖는다. <차트 2>

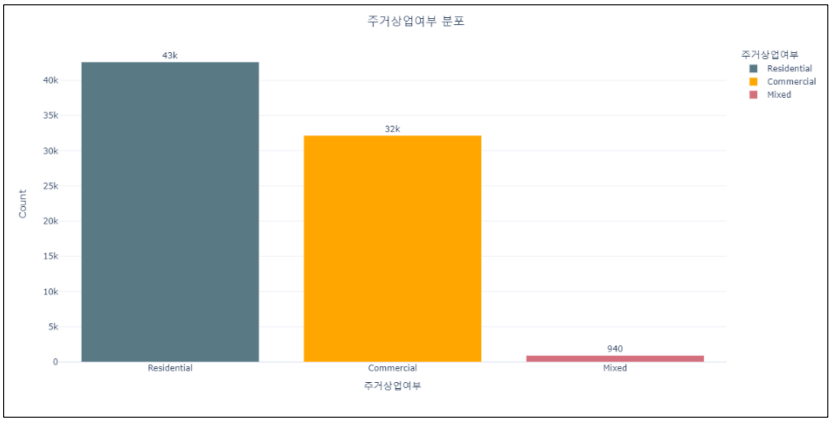
건물_유형	설명	건물_유형	설명
Warehouse Uncategorized	기타 창고	Office Medical non diagnostic	의료 사무공간 - 진단 미포함
Warehouse Selfstorage	개인 창고	Office Bank or other financial	은행 및 금융기관
Warehouse Refrigerated	냉장용 창고	Nursing Home	양로원
Warehouse Nonrefrigerated	냉장 아닌 창고	Multifamily Uncategorized	빌라 - 미분류
Warehouse Distribution or Shipping center	물류 창고	Mixed Use Predominantly Residential	주거상업복합 - 주거위주
Service Vehicle service repair shop	카센터	Mixed Use Predominantly Commercial	주거상업복합 - 상업위주
Service Uncategorized	기타 서비스센터	Mixed Use Commercial and Residential	주거상업 혼합
Service Drycleaning or Laundry	세탁소	Lodging Uncategorized	숙박 기타
Retail Vehicle dealership showroom	차량(및 기타 운송수단 판매)	Lodging Other	숙박 기타
Retail Uncategorized	기타 소매상	Lodging Hotel	호텔
Retail Strip shopping mall	실외 쇼핑센터	Lodging Dormitory or fraternity sorority	호스텔
Retail Enclosed mall	실내 복합쇼핑센터	Laboratory	실험실
Religious worship	종교 기관	Industrial	산업체
Public Safety Uncategorized	기타 공공 기관	Health Care Uncategorized	헬스케어기관
Public Safety Penitentiary	교도소	Health Care Outpatient Uncategorized	외래 기타
Public Safety Fire or police station	소방서 또는 경찰서	Health Care Outpatient Clinic	외래 클리닉
Public Safety Courthouse	법원	Health Care Inpatient	입원 병동
Public Assembly Uncategorized	공공 시설 - 미분류	Grocery store or food market	식료품점
Public Assembly Stadium	공공 시설 - 스타디움	Food Service Uncategorized	음식점 - 미분류
Public Assembly Social meeting	공공 시설 - 회의실	Food Service Restaurant or cafeteria	음식점 - 식당
Public Assembly Recreation	공공 시설 - 대중 오락	Food Service Other	음식점 - 기타
Public Assembly Other	공공 시설 - 기타	Food Sales	식료품점
Public Assembly Movie Theater	영화관	Education Uncategorized	교육기관 기타

Public Assembly Library	도서관	Education Preschool or daycare	유치원 어린이집
Public Assembly Entertainment culture	문화센터	Education Other classroom	교실 교육기관
Public Assembly Drama theater	극장	Education College or university	대학
Parking Garage	주차공간	Data Center	데이터 센터
Office Uncategorized	기타 사무공간	Commercial Unknown	상업용 기타
Office Mixed use	복합 사무실	Commercial Other	상업용 기타
2to4 Unit Building	빌라 - 2~4 unit	5plus Unit Building	빌라 - 5plus unit

<표 3> 건물 주 사용 용도 구분 (건물 유형)



<차트 1> 건물 유형별 분포

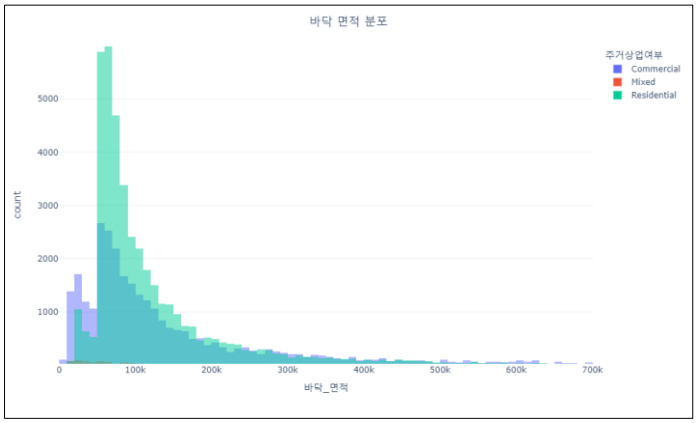


<차트 2> 주거상업여부 분포

- 건물 연면적 및 건축 년도

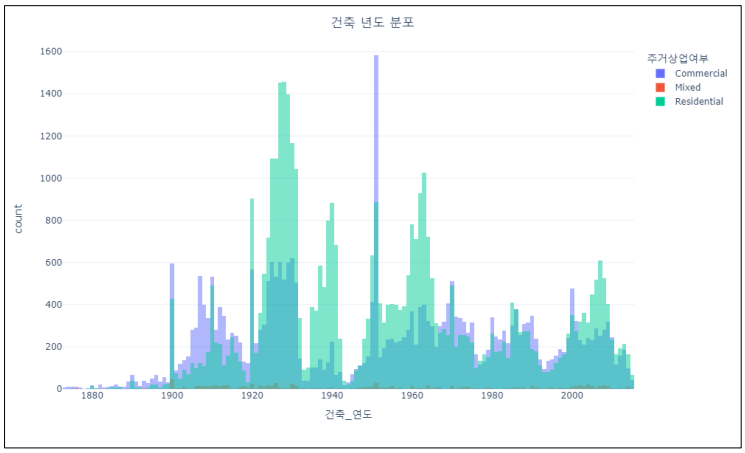
건물 바닥 면적은 건물의 한 층 바닥 면적을 건물의 총 층수로 곱해서, 대지에 들어선 건축물 내부의 모든 바닥 면적을 합한 크기라고 볼 수 있다. 연면적의 경우, 5 만 제곱미터 이하의 건물인 경우와, 5 만 제곱미터 이상인 건물들의 전력 소비량이 서로 다른 경향을 보이는 것으로 나타났다. 따라서 해당 값을 기준으로 그룹화하여 분석에 추가적으로 활용하였다.

건물 건축 년도는 해당 건물이 건축된 연도를 뜻한다. 특히 건축 년도 변수는 1900 년도 이전부터 2015 년까지의 값을 갖는다. 이 중 전력량 발생이 더 많은 특정 시기들이 존재하는 것으로 나타나, 비슷한 분포를 가지는 시기별로 그룹화하였다. 그룹별로 그려본 전력량 분포는 다음과 같다. <차트 5>



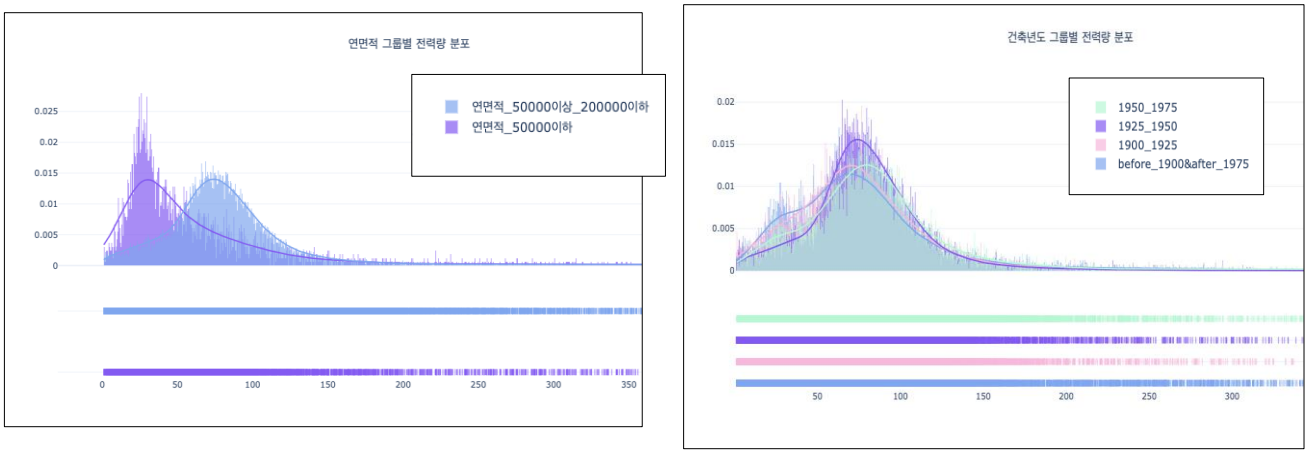
통계량	값
Mean	165959.16011
Standard deviation	246889.21715
Minimum	943
25% Quantile	62371.25
50% Quantile (Median)	91360
75% Quantile	165938.25
Maximum	6385382

<차트 3> 건물 연면적 분포 및 통계량



통계량	값
Mean	1952.306764
Standard deviation	37.053619
Minimum	0
25% Quantile	1927
50% Quantile (Median)	1951
75% Quantile	1977
Maximum	2015

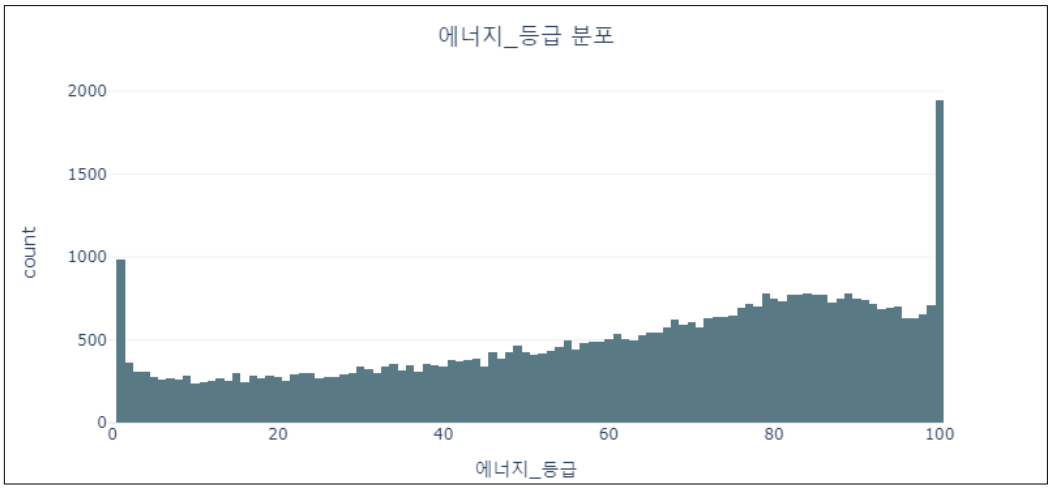
<차트 4> 건축년도 분포 및 통계량



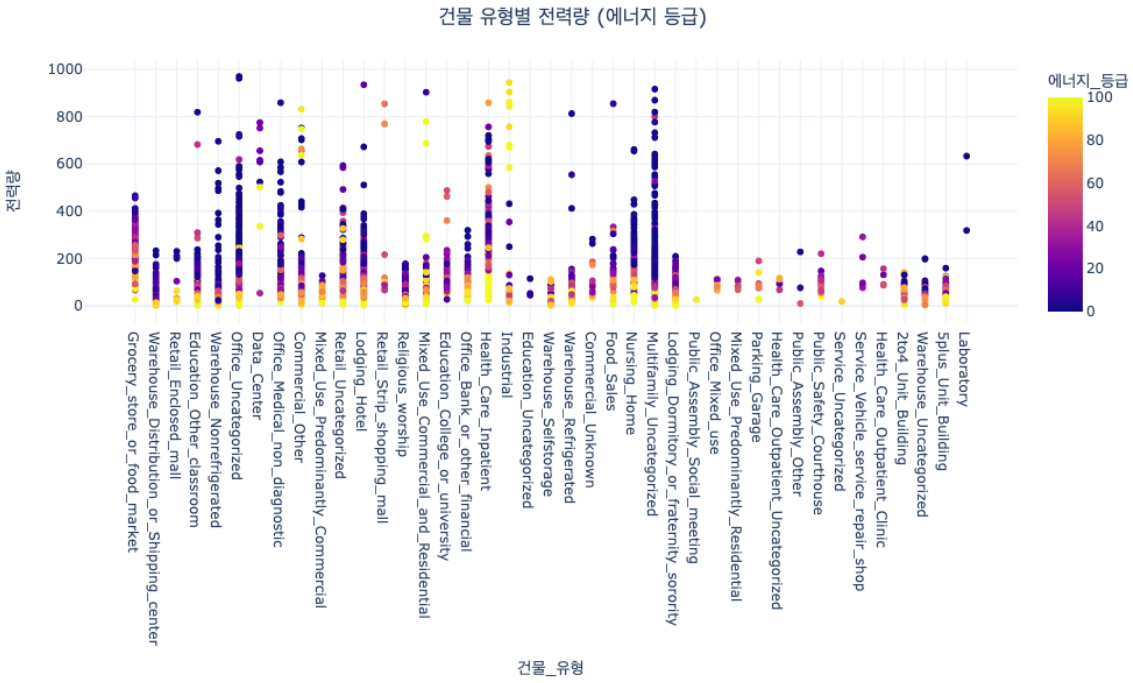
<차트 5> 그룹화된 변수들의 그룹별 전력량 분포 (좌 – 연면적 / 우 – 건축년도)

- 에너지 스타 등급

에너지 스타 등급 변수는 비율 변수로, 1 과 100 사이의 상대적인 값으로 존재하며, 건물의 에너지 소비량이 전국의 유사한 다른 건물들에 비해 얼마나 높은지를 이해할 수 있는 척도이다. 에너지 등급은 건물의 위치, 고도와 같은 건물의 물리적 특성과, 건물 내부에 거주하는 기업이나, 세입자, 거주민들, 건물에서 사용되는 디지털 기기의 갯수 정도나, 건물의 주된 사용 목적 등 건물이 가진 다양한 측면의 정보들을 종합하여 산정된다. 보통 이 척도를 통해 건물의 설비 상태를 파악하기도 하며, 일반적으로 50 점은 평균 성능, 75 점 이상은 최고 성능임을 나타낸다. 따라서, 건물의 설비 상태나, 에너지 발생량을 예측하는 데 높은 설명력을 갖는 변수일 것으로 판단된다. 건물 유형별 에너지 등급에 따른 전력량 분포를 살펴보면, 에너지 등급이 높을 수록 전력량 소비가 적은 것을 확인할 수 있다. <차트 7>



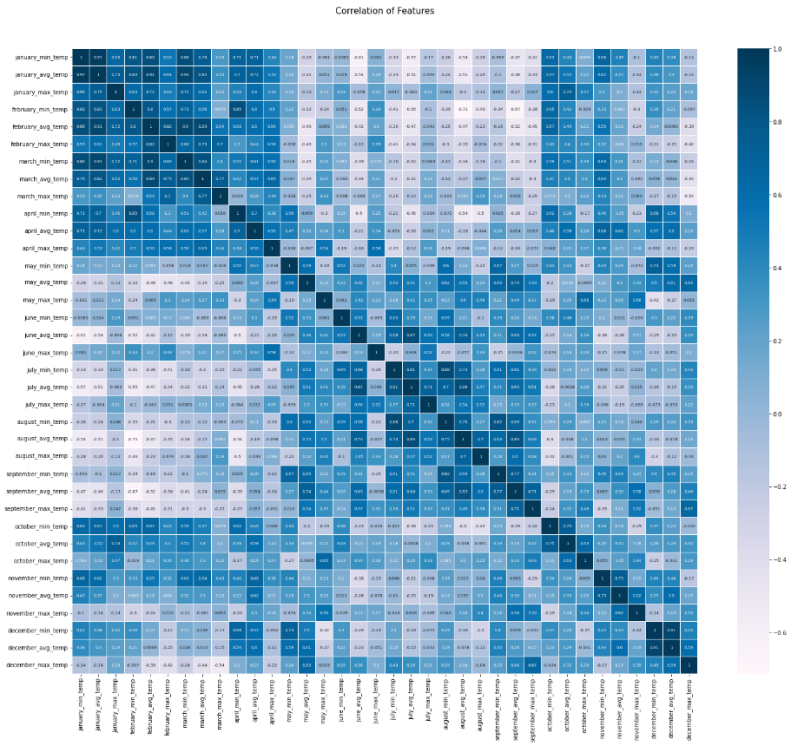
<차트 6> 에너지 등급 분포



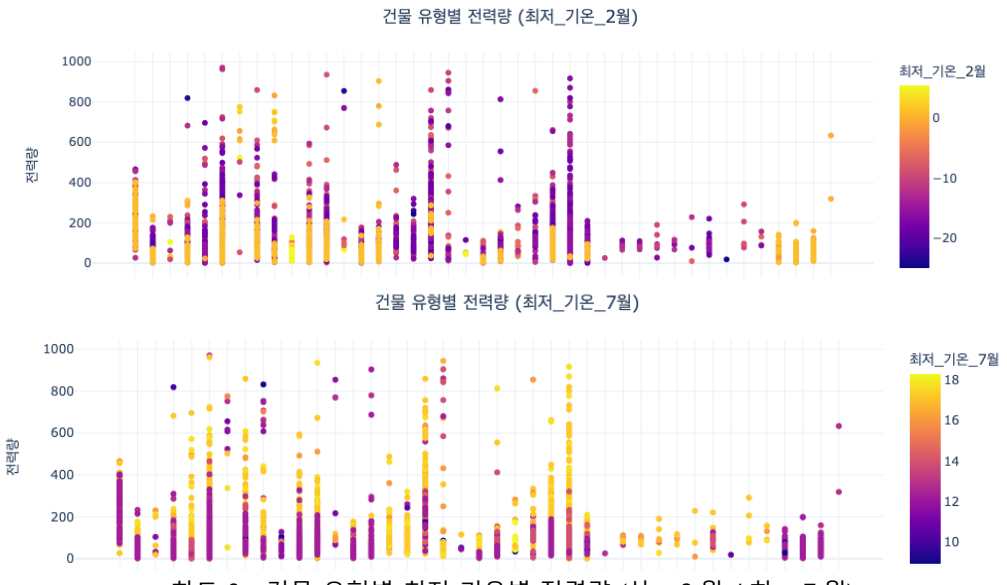
<차트 7> 건물 유형별 에너지 등급별 전력량

b. 월별 기온 변수

1 월부터 12 월까지 최저, 최고, 평균 기온에 대한 변수가 주어짐에 따라, 이들 간의 상관관계를 확인해보고자 히트맵으로 시각화 해보았다. <차트 8> 1 월-3 월의 상관관계가 특히 큰 것으로 나타났고, 6 월-9 월 동안의 상관관계도 마찬가지로 큰 것으로 나타났다. 계절별로 달라지는 기온의 흐름을 확인할 수 있었으며, 나아가 계절의 기온별 전력 소비량은 어떨지 확인해보았다. 겨울에 해당하는 2 월의 최저 기온에 따른 전력량과, 여름에 해당하는 7 월의 최저 기온에 따른 전력량은 다음과 같다. <차트 9> 2 월의 경우, 최저 기온이 높은 곳일 수록 전력 소비량이 낮음을 확인할 수 있고, 반대로 여름에 해당하는 7 월은, 최저 기온이 높은 곳일수록 전력 소비량이 높음을 확인할 수 있디. 따라서 계절별 온도에 따른, 냉난방 시설이 전력량 소비에 큰 영향력을 가지는 요인이 될 것으로 기대할 수 있다.



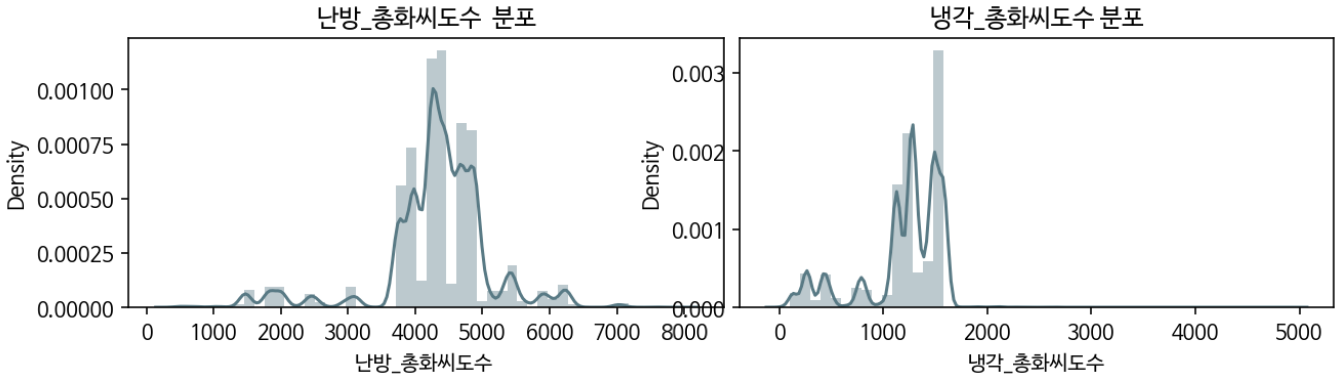
<차트 8> 월별 최저, 최고, 평균 기온 히트맵



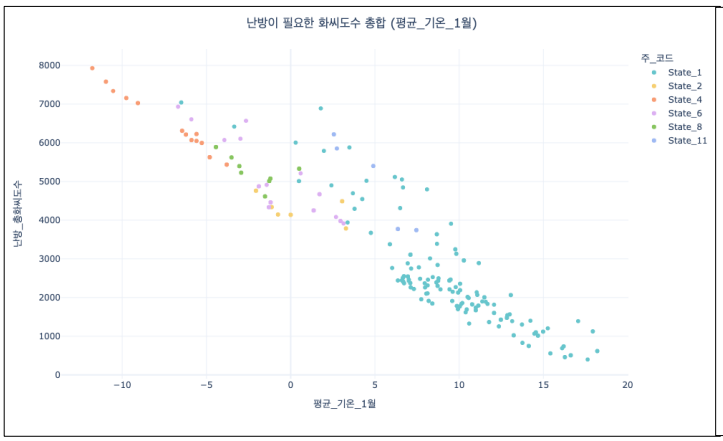
<차트 9> 건물 유형별 최저 기온별 전력량 (상 – 2 월 / 하 – 7 월)

- 난방 총 화씨도수 / 냉방 총 화씨도수

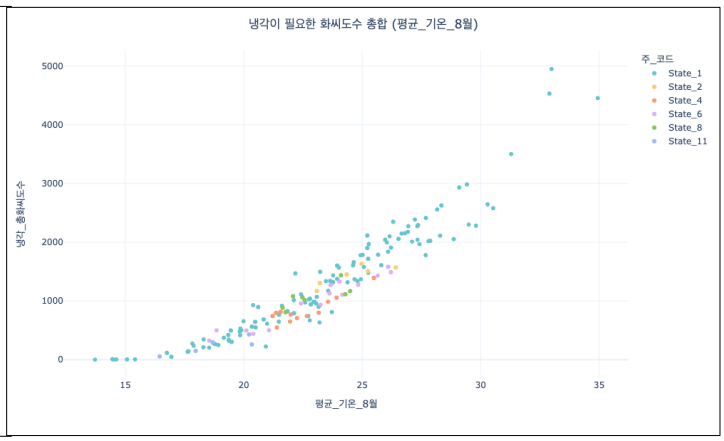
다음은 앞서 언급했던 냉난방 총 화씨도수에 관한 분석이다. '냉각 총 화씨도수'는 섭씨기준 65 도, 약 섭씨 18 도를 냉방이 필요한 기준으로 정하고, 최저 기온이 기준 온도를 넘어간 날들에 한해 그날의 화씨 도수를 모두 합한 변수이다. 난방에 대해서도 이를 동일하게 적용한 변수가 '난방 총 화씨도수'로 존재한다. 섭씨 기준 18 도 밑으로 떨어진 날들에 대해 화씨 도수를 모두 합한 값을 갖는다. 매달 합산하여 건물마다의 연간 총계를 산출한다. 1 월의 평균기온과 난방 화씨도수가 음의 관계를 보이는 것을 확인할 수 있고, 8 월의 평균 기온과 냉방 화씨 도수가 양의 관계를 보이는 것을 확인할 수 있다.



<차트 10> 냉난방 총화씨도수 (좌-난방 / 우-냉방)



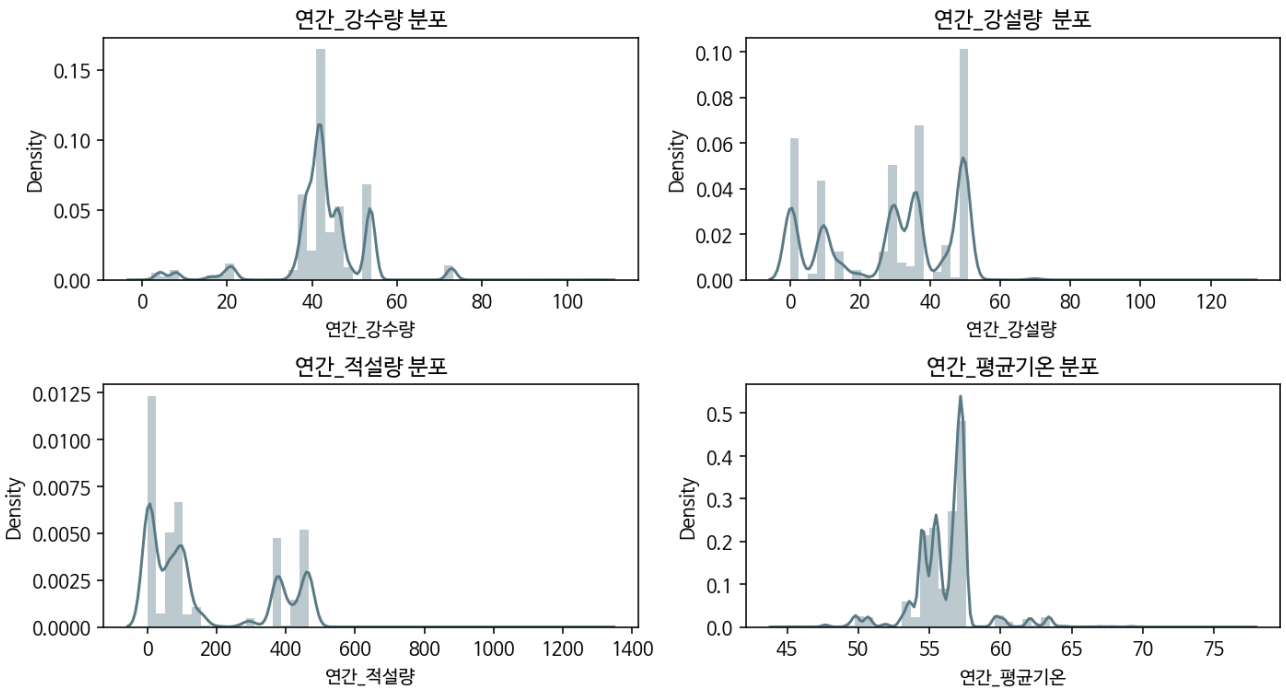
<차트 11> 1 월 평균 기온별 난방 총화씨도수



<차트 12> 8 월 평균 기온별 냉방 총화씨도수

c. 기상 변수

기상 관련 변수로는 연간 강수량, 연간 강설량, 연간 적설량, 연간 평균 기온이 주어졌으나, 결측치들이 80% 가까이 존재해 분석에 다방면으로 활용하기 어렵다는 한계가 있었다. 건물의 지역적 위치나, 관측치의 관측년도에 관한 정보가 주어졌다면 외부에서 유용한 정보를 추가적으로 기상 변수로 활용할 수 있었을 것 같다. 또한, 기상정보는 전력량 소비에 다양한 측면에서 영향을 줄 수 있을 것으로 기대된다. 예를 들어, 강수량이 높은 일자에 건물 내부에서도 전등을 더 많이 사용하기 때문에 전력 소비량이 증가할 수 있을 것으로 생각된다. 본 분석은 일자별 예측이 아니고, 연간 소비량 예측이었으며, 각 관측치의 지역과 관측 시기 정보를 마스킹하여 제공했기 때문에, 기상 변수를 풍부하게 활용하지 못했다. 추후 분석이 이루어질 경우, 기상 변수 활용이 전력 소비에 있어 주목할 만한 가설과 결과들을 많이 제공할 것으로 기대된다.

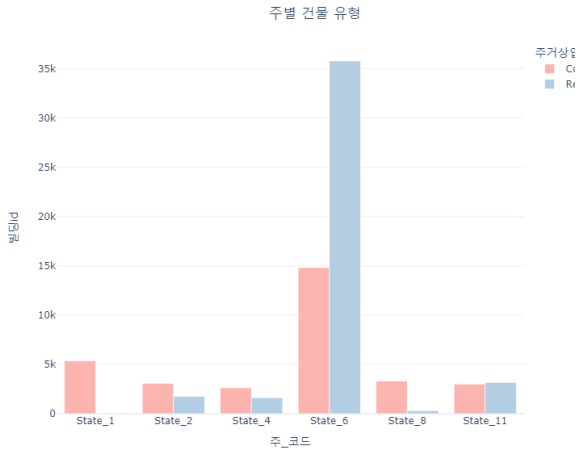


<차트 13> 연간 기상 변수별 분포

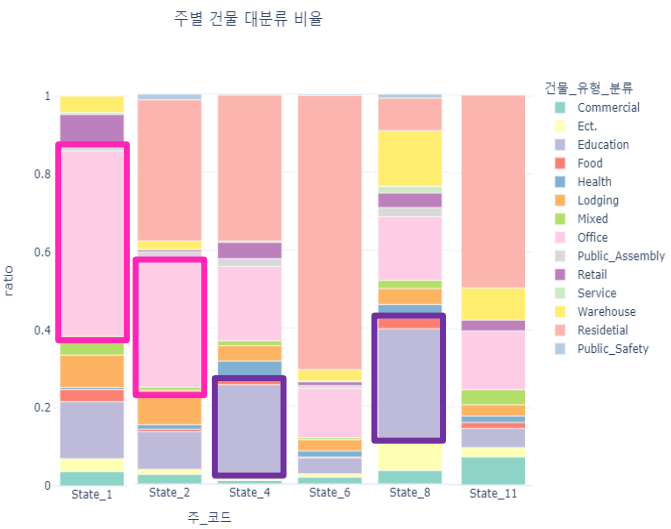
IV. 가설 검정

a. 마스킹 된 지역코드 (주, State) 변수에 관한 해석

가설 1. 지역별 건물 유형은 어떨까



<차트 14> 주별 주거상업여부별 건물 분포



<차트 15> 주별 건물 유형별 건물 분포

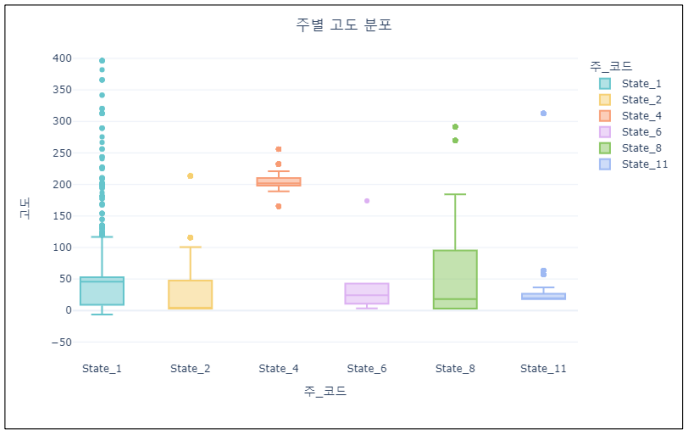
'State 6'에 해당하는 관측치는 약 5 만개 정도로 압도적으로 많고, 다른 지역들은 5 천개의 선에서 비슷하게 관측된다. <차트 14> 지역별 건물 유형의 빈도를 살펴보면, 'State 1'과 'State 8'은 거의 상업지구로 보이고, 'State 6'은 유일하게

거주용 빌딩이 압도적으로 많은 지역으로 보인다. 구체적으로 'State 1'과 'State 2' 지역에는 사무용 건물이 많고, 'State 4'와 'State 8'은 교육용 건물들이 많은 것을 확인할 수 있다.

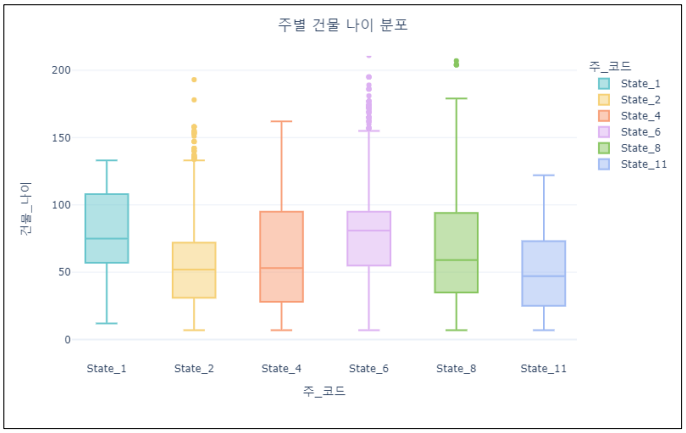
가설 2. 지역별 고도와 건물 건축년도를 통해 지역을 특정할 수 있을까?

주별로 고도와 건물 건축년도를 살펴보면, 'State 4'는 다른 지역들에 비해 고도가 매우 높은 지역임을 확인할 수 있다. 'State 1'도 분포의 꼬리가 매우 길어 높은 고도에 위치한 건물들이 소수 존재함을 알 수 있었다. 이 건물들의 고도는 'State 4'보다도 높은 것으로 보인다. 주로 높은 산에 형성된 마을이 있을 수 있다고 보여지며, 이 관측값들의 건물 유형을 확인해 본 결과 모두 'Education Uncategorized(교육-미분류)'에 속한 건물들인 것으로 나타났다. 여기서 주목할 점은, 'State 4'의 고도가 높음에 따라, 겨울기간의 기온들이 타 지역에 비해 뚜렷하게 낮은 것으로 드러났고, 이에 대해 해당 시기의 전력량 소비도 다른 지역들보다 높음을 확인할 수 있었다.

또한, 건물의 건축년도를 활용해 짐작 가능한 건물의 나이를 계산하여 분포를 비교해보았다. <차트 17> 건물 나이는 30 년에서 100 년 사이에 분포하고 있는데, 그 중에서도 'State 2'는 비교적 젊은 지역에 속했고, 'State 1'은 비교적 오래된 지역으로 볼 수 있다. 추가적으로, 신식 건물일수록 고층 빌딩이 많을 것으로 판단하여, 연면적과 건물 나이와의 분포도 확인해보았으나, 별다른 패턴을 확인할 수 없었다.

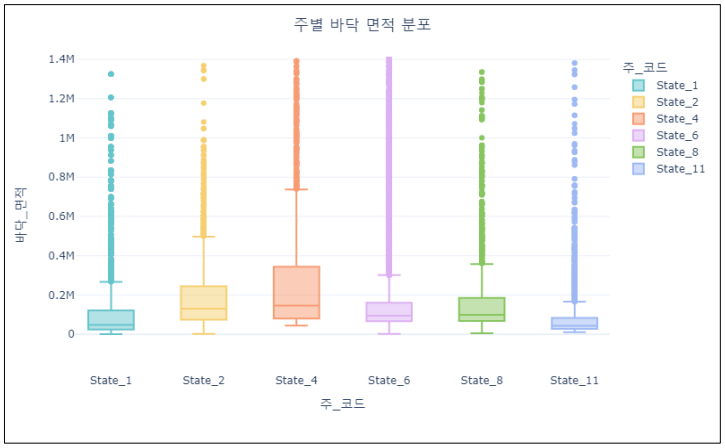


<차트 16> 주별 고도 분포

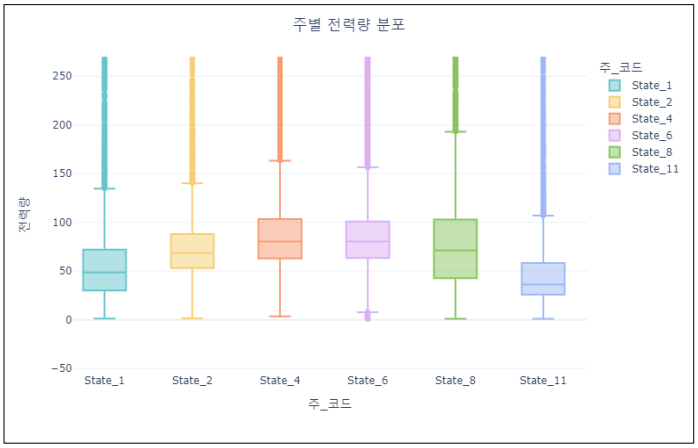


<차트 17> 주별 건물 나이 분포

- 그 외 지역적 특성 파악



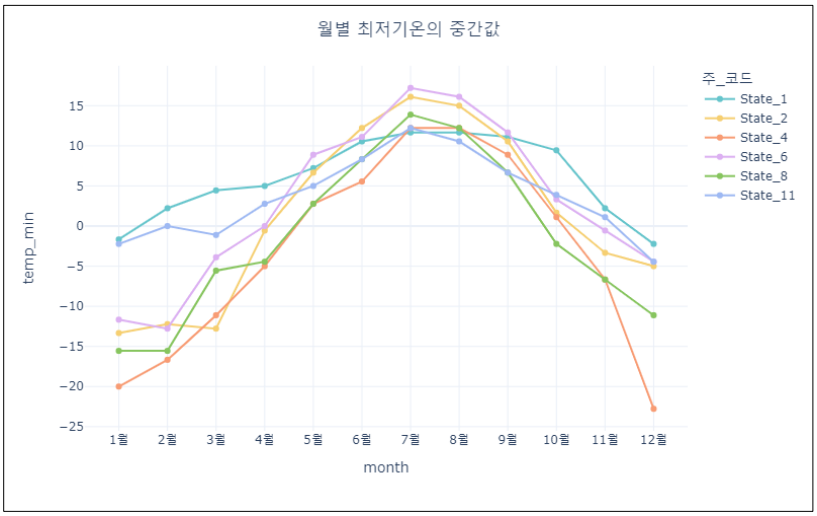
<차트 18> 주별 연면적 분포

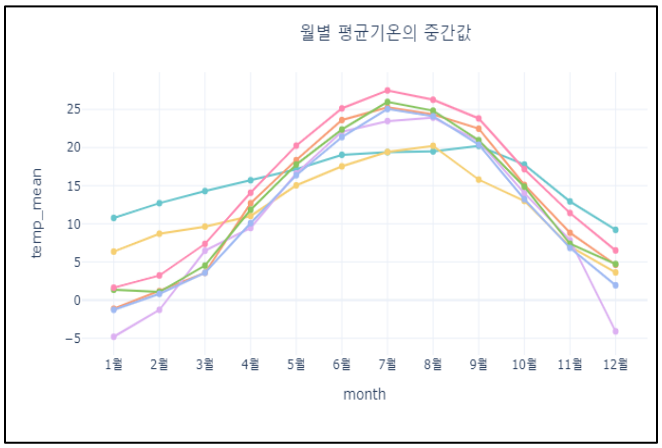


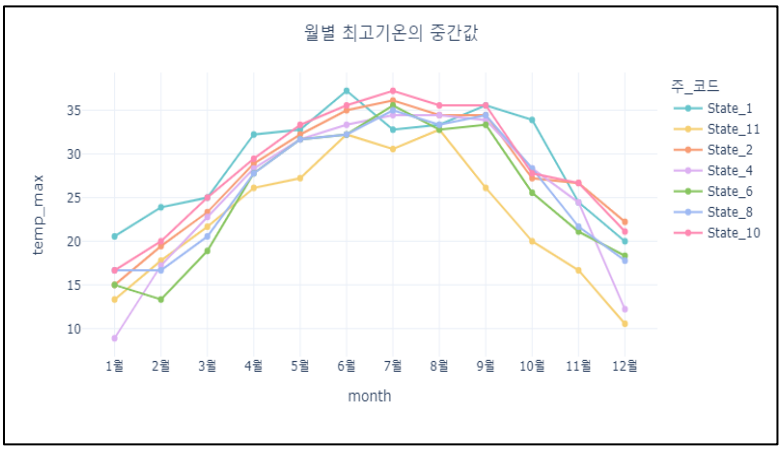
<차트 19> 주별 전력량 분포

이외로도, 주별로 건물 연면적 분포를 확인해보았다. <차트 18> 대부분 꼬리가 아주 긴 분포를 가지고 있었고, 'State 1'과 'State 11'은 특히나 작은 건물이 많은 지역임을 확인할 수 있었다. 또한 주별 전력량 분포도 확인해보았는데, 대체적으로 비슷한 분포를 보이나, 'State 4'와 'State 6'에서 많은 전력량을 사용하고 있음을 알 수 있었다. 'State 8'은 가장 넓은 4 분위 범위(IQR)를 보였고, 'State 1'과 'State 11'이 가장 적은 전력량을 소비하는 것을 알 수 있었다. <차트 19>

가설 3. 각 주의 날씨는 어떨까?







<차트 20> 주별 월별 기온의 중간값 (위-최저 / 왼쪽 아래 – 평균 / 오른쪽 아래 – 최고)

월별 최저기온으로 각 주의 날씨를 살펴보았다. <차트 20> 'State 1'과 'State 11'은 겨울 최저기온이 0 도에 가깝고, 여름에도 기온이 그리 높지 않은 지역임을 확인할 수 있다. 'State 4'는 앞서 살펴봤듯이, 고도가 높은 지역적 특성에 맞물려, 겨울에 가장 춥고, 여름에 가장 시원한 지역임이 확인된다. 'State 6'와 'State 8' 그리고 'State 2'의 경우, 1 월에서 3 월까지의 경향이 매우 비슷하다. 특히 1 월에 비해 2 월에 날씨가 따듯해지지 않는 것이 특이점으로 보인다.

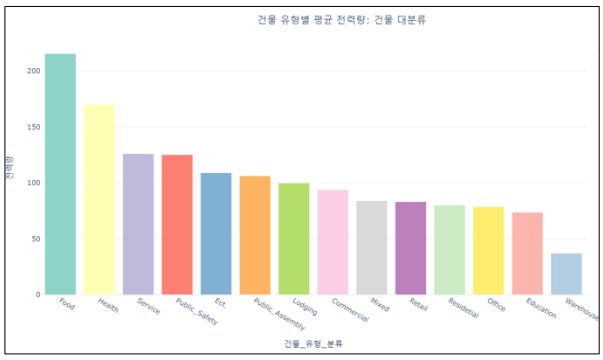
주별로 월별 최저기온을 확인해본 결과와 전력 소비량 간의 상관관계를 파악해보았을 때, 추워지는 시기와 해당 지역의 전력량이 높아지는 패턴을 가짐을 확인할 수 있었다. 해당 분석 결과는 자명해보이지만, 더워지는 시기에 전력량이 늘어나는 것보다 더욱 뚜렷한 패턴을 보였기 때문에, 분석에서 얻은 의미 있는 결과로 판단된다. 결과적으로 정리해본 주 코드별 결과는 다음과 같다. <표 4>

	전력량	관측치 개수	주거/상업	고도	건물 나이	바닥 면적	기온
State 1	낮음		상업지구		오래된 건물	작은 건물 많음	따뜻, 연교차 작음
State 2					최신 건물		
State 4	약간 높음			매우 높은 지역		가장 다양한 분포	매우 추운 지역
State 6	약간 높음	다른 지역의 약 10 배	주거지 70%				
State 8			상업지구				
State 11	낮음					작은 건물 많음	따뜻, 연교차 작음

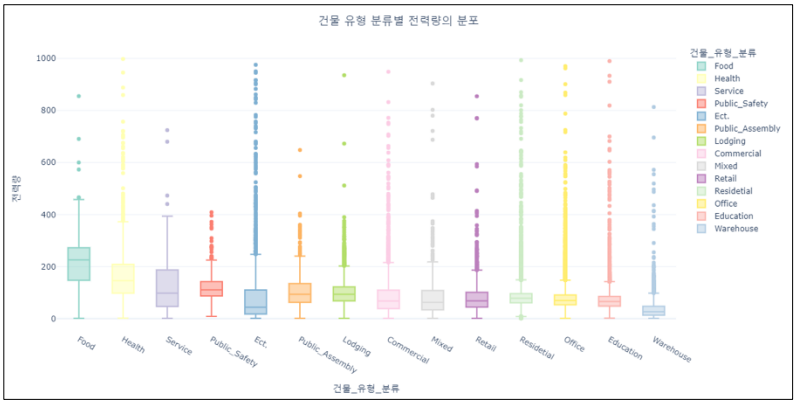
<표 4> 지역별 특징 분류 표

b. 건물 유형별 전력 소비량 해석

가설 2. 어떤 용도의 건물이 전력을 많이 소비하는가?



<차트 21> 건물 유형 대분류별 평균 전력량



<차트 22> 건물 유형 대분류별 전력량 분포

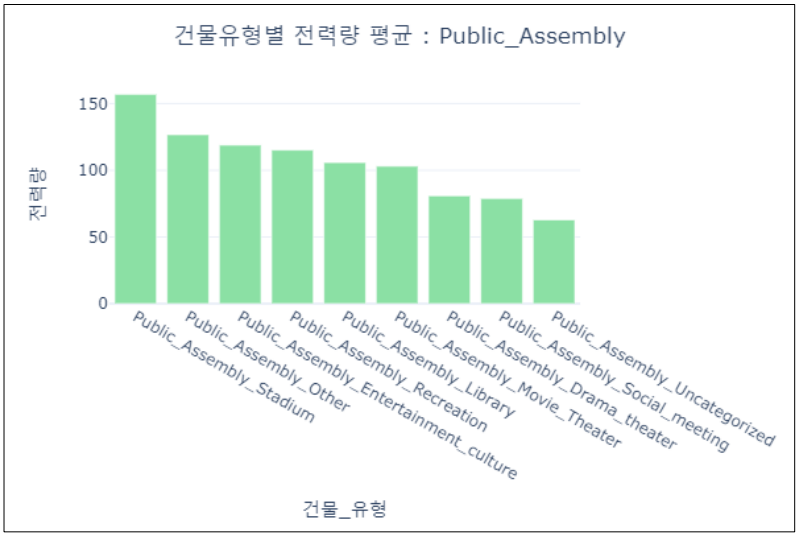
건물 유형 변수는 총 60 개의 분류를 가진다. 이를 14 개의 대분류로 구분하고, 대분류별 전력량을 파악해보았다. 건물 유형 대분류별 평균 전력량을 살펴본 결과, 가장 많은 전력량을 소비하는 분류는 순서대로 음식, 건강, 서비스, 공공 보안, 기타 분류였다. <차트 21> 1 위인 food 는 식당. 식료품점 등이 포함된 대분류로, 음식을 보관해야 하므로 전력 소모가 많다고 유추해 볼 수 있다. 2 위는 헬스케어인데, 입원 병동, 병원과 같은 각종 의료 기관이 포함되어 있기 때문에 전력량 소비가 높은 것으로 보인다. 다음은 서비스 분류인데, 세탁소, 카센터 등의 서비스를 하는 기관들이 속해 있다. 다음은 공공 보안과 관련된 대분류로, 소방서, 경찰서, 그리고 교도소까지 다양한 공공기관들을 포함하고 있다. 5 위는 기타로 설정한 분류인데, 데이터 센터, 실험실, 산업체, 주차공간이 기타 분류에 속한다. 특히 데이터 센터나 실험실과 같이 실제로 건물 유형별 전력 소비량을 살펴봤을 때, 가장 전력 소비가 많았던 건물 유형들이 기타 분류에 해당된다. 상위 5 개의 건물 유형 대분류에 속한 하위 건물 유형들은 다음과 같다. <표 5>

Grocery store or food market	식료품점	Health Care Uncategorized	기타 헬스케어기관
Food Service Uncategorized	음식점	Health Care Outpatient Uncategorized	외래 미분류

Food Service Restaurant or cafeteria	음식점(식당)	Health Care Outpatient Clinic	외래 클리닉
Food Service Other	음식점 기타	Health Care Inpatient	입원 병동
Food Sales	식료품점	Nursing Home	양로원
Public Safety Uncategorized	기타 공공 기관	Data Center	데이터 센터
Public Safety Penitentiary	교도소	Laboratory	실험실
Public Safety Fire or police station	소방서 또는 경찰서	Industrial	산업체
Public Safety Courthouse	법원	Parking Garage	주차공간
Service Vehicle service repair shop	카센터	Service Drycleaning or Laundry	세탁소
Service Uncategorized	기타 서비스센터		

<표 5> 상위 5 개 건물 유형 대분류 하위 분류

평균 전력량 6 위부터 10 위까지를 차지한 대분류는 순서대로 공공시설, 숙박, 상업용 건물들, 주거상업복합 시설, 소매 및 쇼핑센터들이 있다. 공공시설에는 영화관, 도서관, 극장과 같은 곳들이 해당하며, 이 중 스타디움이 가장 규모가 큰 시설이다. 해당 대분류에 속하는 건물 유형 중 스타디움이 가장 큰 전력 소비량을 차지했는데, 공간의 크기가 큰 만큼, 그리고 많은 사람들이 사용하는 공간인 만큼, 전력 소비량이 많은 것을 확인할 수 있다. <차트 23> 또한 10 위를 차지한 소매 및 쇼핑센터들에는 차량 및 기타 운송수단을 판매하는 쇼룸과, 기타 소매상, 길거리 소매상, 실내 복합 쇼핑센터에 위치한 소매상들이 속해 있는데, 이 시설들은 공통적으로 낮시간에 주로 운영되는 건물들로 전력량이 다른 대분류보다 낮은 것으로 해석해볼 수 있다.



<차트 23> "건물 유형 대분류 – 공공시설"의 전력량 분포

		Lodging Uncategorized	숙박 미분류
		Lodging Other	숙박 기타
Public Assembly Uncategorized	공공 장소 미분류	Lodging Hotel	호텔
Public Assembly Stadium	공공 스타디움	Lodging Dormitory or fraternity sorority	호스텔
Public Assembly Social meeting	공공 모임 장소	Commercial Unknown	상업용 미분류
Public Assembly Recreation	공공 휴게 장소	Commercial Other	상업용 기타
Public Assembly Other	기타 공공 장소	Mixed Use Predominantly Residential	복합 – 주거위주
Public Assembly Movie Theater	영화관	Mixed Use Predominantly Commercial	복합 – 상업위주
Public Assembly Library	도서관	Mixed Use Commercial and Residential	복합 – 기타
Public Assembly Entertainment culture	문화센터	Retail Vehicle dealership showroom	차량(및 기타 운송수단 판매)
Public Assembly Drama theater	극장	Retail Uncategorized	기타 소매상
		Retail Strip shopping mall	실외 쇼핑센터
		Retail Enclosed mall	실내 복합쇼핑센터

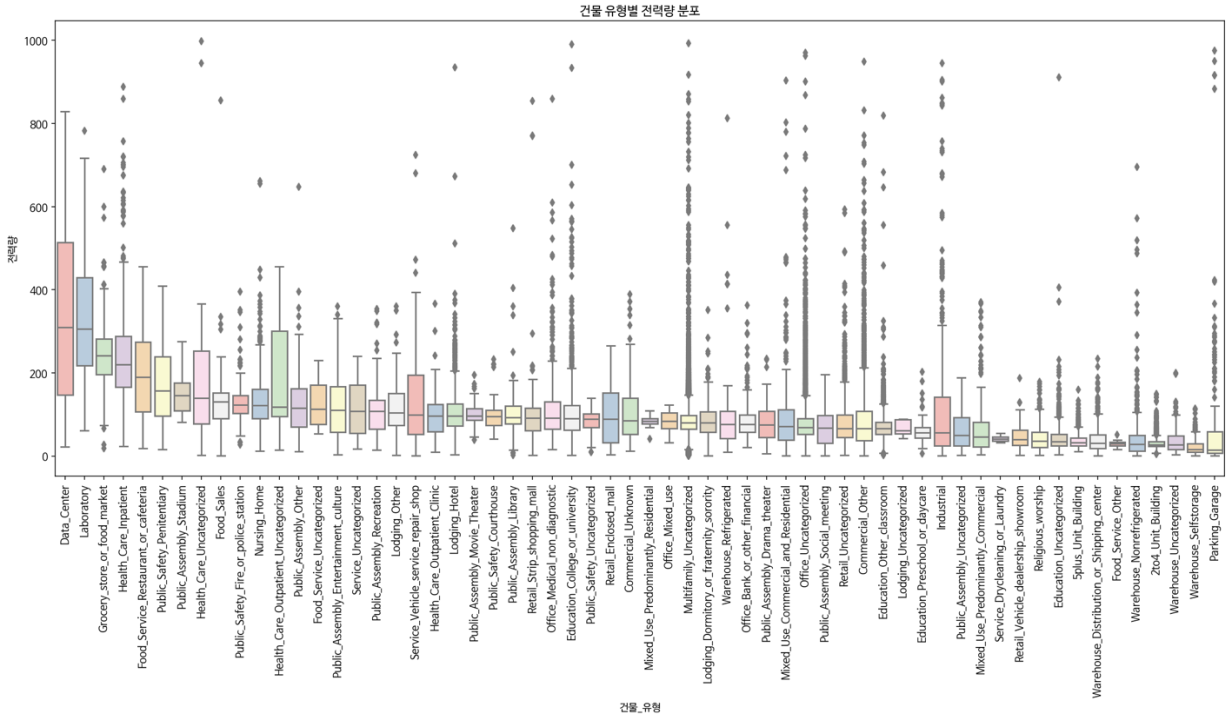
<표 6> 6 위-10 위 건물 유형 대분류 하위 분류

마지막으로 11 위-14 위까지는, 순서대로 주거용 건물, 사무용 건물, 교육용 건물, 창고용 건물이 차지했다. 특히 창고용 건물들은 냉장용인 창고와 그렇지 않은 창고들로 분류가 되는데, 냉장용 창고가 다른 창고의 약 2 배정도 큰 전력 소비량을 보여주고 있다. 60 개 모든 건물 유형의 전력량 분포를 확인해보았다. <차트 25> 또한 건물 유형별 전력량의 중간값을 기준으로 워드 클라우드를 그려보면, 데이터센터, 실험실이 가장 독보적인 크기를 가짐을 확인할 수 있었다.

<차트 24> "건물 유형 대분류 – 창고"의 전력량 분포

		Education Uncategorized	교육기관 미분류
Multifamily Uncategorized	주거 공간 미분류	Education Preschool or daycare	유치원 어린이집
5plus Unit Building	빌라 - 5plus unit	Education Other classroom	교실 교육기관
2to4 Unit Building	빌라 - 2~4 unit	Education College or university	대학
Office Uncategorized	기타 사무공간	Warehouse Uncategorized	기타 창고
Office Mixed use	복합 사무실	Warehouse Selfstorage	개인 창고
Office Medical non diagnostic	의료 사무실	Warehouse Refrigerated	냉장용 창고
Office Bank or other financial	은행 및 금융기관	Warehouse Nonrefrigerated	냉장 아닌 창고
		Warehouse Distribution or Shipping center	물류 창고

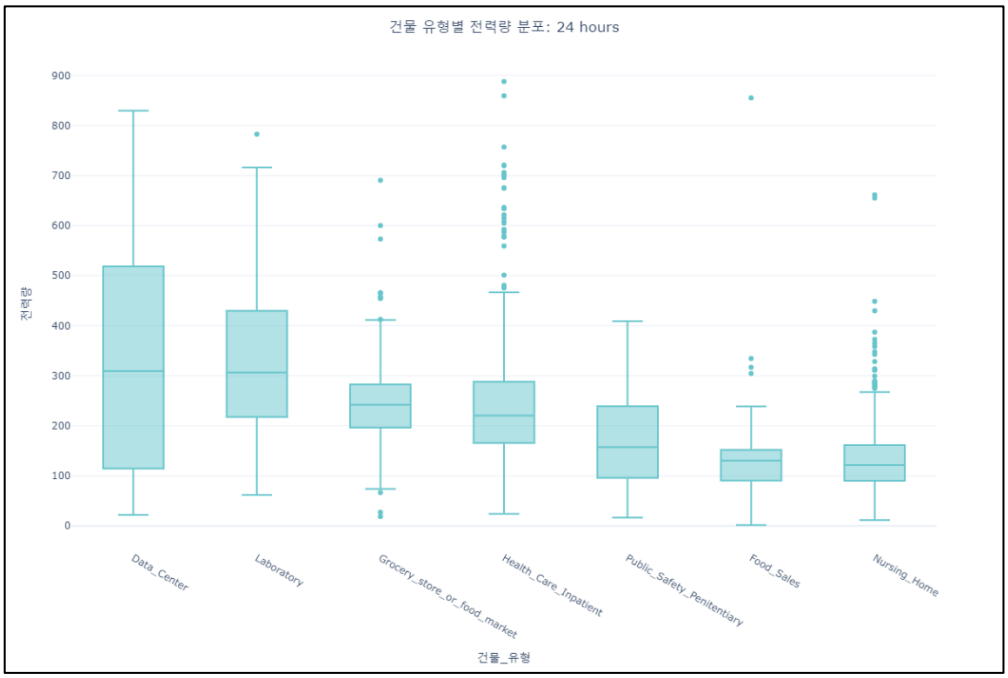
<표 7> 11 위-14 위 건물 유형 대분류 하위 분류



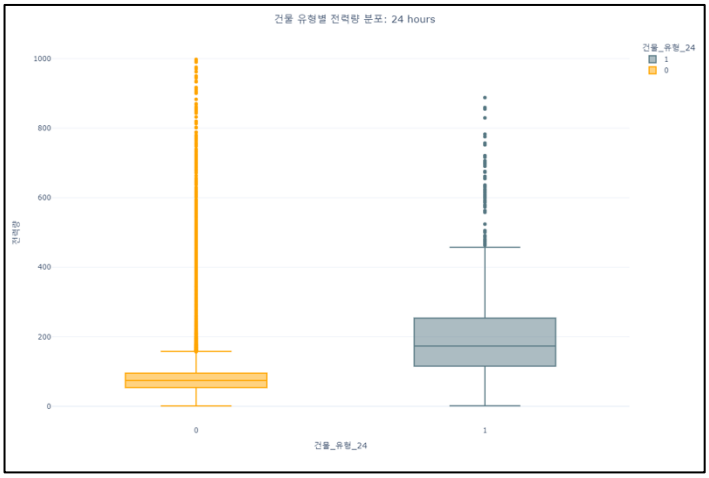
<차트 25> 건물 유형별 전력량 분포

건물 유형 대분류의 분석을 통해, 건물 유형들 중에서 24 시간동안 많은 전력을 사용해야하는 건물들이 존재하며, 해당 건물들의 전력 소비량이 높은 것을 유추해볼 수 있었다. 따라서 많은 전력이 상시 소비되는 건물들을 분류하여 '건물 유형_24'라는 새로운 범주형 파생변수를 만들어 분석에 활용하였다. 총 7 개의 건물 유형이 해당 분류에 속하였고, 이들은 주거공간으로 분류되지는 않지만 요양원이나 교도소처럼 많은 사람들이 24 시간 움직이지 않고 건물 내에서 생활하는 등 주거공간 이상의 역할을 하는 공간들임을 확인할 수 있다. 그리고 식료품점, 실험실, 데이터 센터처럼 24 시간동안 전력이 소비되면서 관리가 필요한 공간들을 특수하게 분류해낼 수 있다. <차트 26> 또한 24 시간 운영되는 건물들이 그렇지 않은

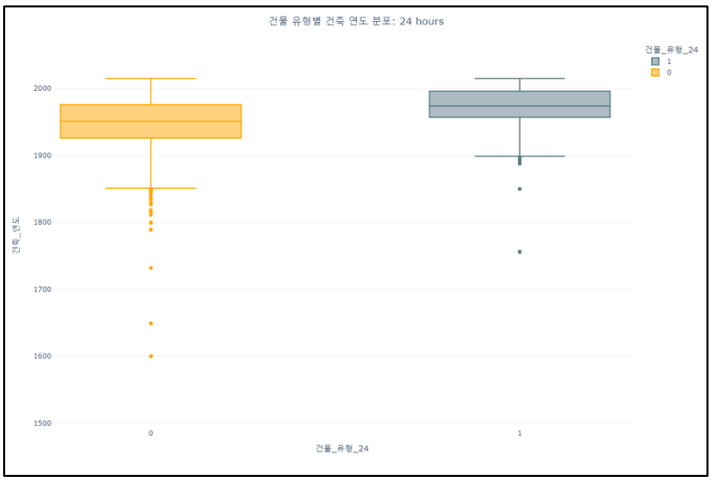
건물들에 비해 에너지 소비량 (중간값 기준)이 높은 것을 알 수 있었다. <차트 27> 또한 건축 연도가 비교적 최근인 건물들로 파악됐다. <차트 28>



<차트 26> 건물 유형_24 에 속한 건물들의 전력량 분포



<차트 27> 건물 유형별 전력량 분포



<차트 28> 건물 유형별 건축년도 분포

c. 기온 변수에 대한 효율적 분석

이 데이터 중 거의 절반에 해당하는 36 개의 변수가 기온에 관한 변수이다. 월별 최저, 평균, 최고 기온 변수는 각 건물이 위치한 지역의 기후를 설명하고, 특히 겨울에는 난방, 여름에는 에어컨을 가동하기 때문에 전력량과 밀접한 연관이 있다. 이를 효율적으로 설명하고자 PCA 와 time series clustering 기법을 사용하였다.

가설 1. 기온 변수 중 가장 많은 설명력을 가진 변수는 무엇일까?

기온 변수는 월별로 관측되었고, 3 개의 통계량으로 설명되어 변수 간의 상관계수가 매우 높다. 이를 효과적으로 축소하고, 중요한 변수를 파악하고자 PCA 를 진행하였다. 모든 변수가 기온을 의미하기 때문에 따로 표준화는 진행하지 않았다.

PC 1	PC 2	PC 3
47.54%	16.24%	11.79%

PC1 계수 (절대값 큰 순)	
변수명	계수
2 월 최저기온	0.457
3 월 최저기온	0.369
1 월 최저기온	0.362
2 월 평균기온	0.348
1 월 평균기온	0.277
2 월 최고기온	0.248
3 월 평균기온	0.228
11 월 최저기온	0.179
4 월 최저기온	0.169
1 월 최고기온	0.157
10 월 최저기온	0.143

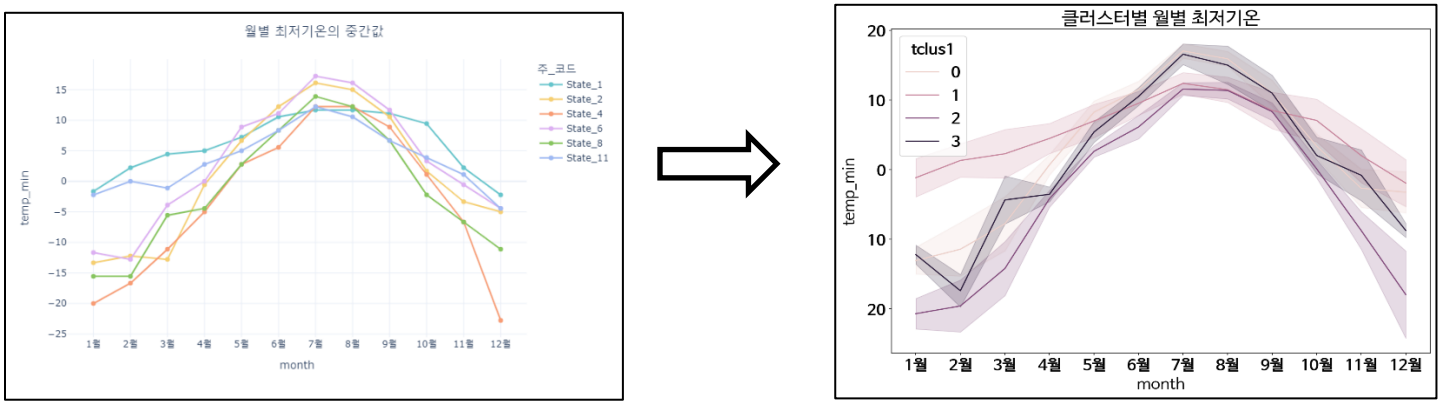
<표 8> 기온 변수 PCA 결과

PCA 결과, 첫 번째 주성분이 전체의 47.54%, 두 번째 주성분은 16.24%, 세 번째 주성분은 11.79%를 설명해 3 개의 변수만으로 총 75.56%를 설명할 수 있다. 이는 이후 선형 모형 적합에서 변수 간의 상관도를 낮추기 위한 목적으로 모델링의 설명 변수로도 활용하였다.

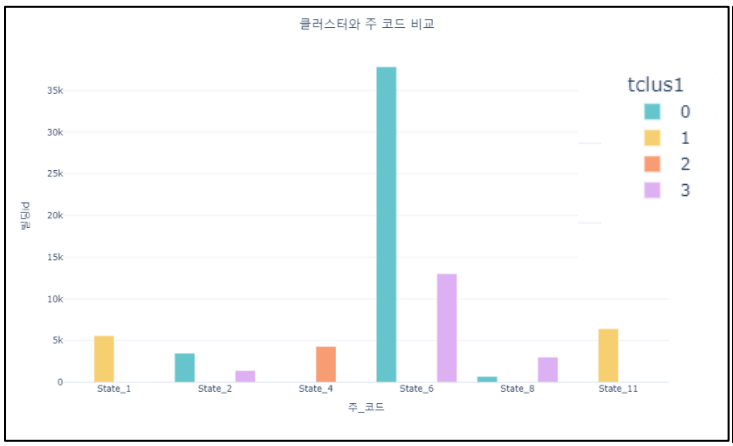
첫 번째 주성분의 계수를 살펴보면, 2 월 최저기온, 3 월 최저기온, 1 월 최저기온 순으로 중요하다. 상위 8 개의 변수가 겨울을 설명하는 것으로 보아 겨울의 기온이 많은 정보를 담고 있다고 볼 수 있다. PCA 에 전력량 정보가 들어가지는 않지만, 상식적으로 겨울의 전력량은 중요한 요소이기 때문에 더욱 주목할 필요가 있다.

가설 2. 같은 주는 같은 기온 흐름을 가질까?

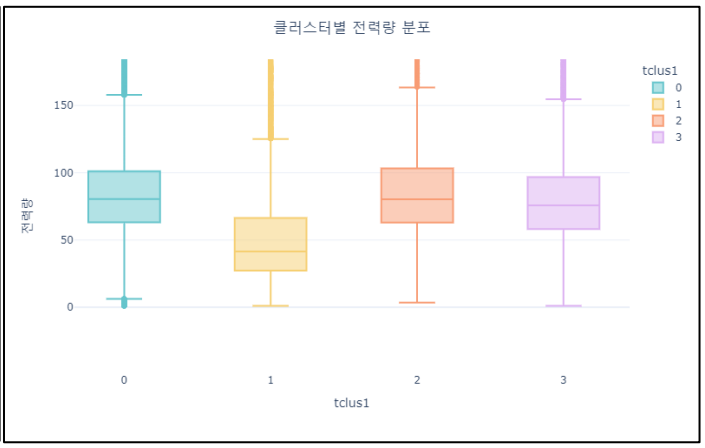
기온 변수는 월별 정보이므로 그 흐름 또한 중요한 정보일 수 있다. 여름에 비슷한 기온 양상을 띠더라도 항상 더운 지역과 여름에만 더운 지역은 전력량 소비 패턴이 다를 것이다. 따라서 기온 정보가 가지고 있는 시간적 흐름을 반영하고자 Time Series Clustering 을 진행하였다. tslearn 패키지의 유클리디안 거리 기반의 클러스터링 방법을 사용하였다. 월별 최저기온 12 개 변수, 평균 기온 12 개 변수, 최고 기온 12 개 변수로 총 3 가지를 시도하였으며 PCA 결과에서도 보았듯 최저기온이 가지고 있는 정보의 가장 설명력이 좋았다.



<차트 29> 범주별 월별 최저 기온 (좌 – 주 코드/우 – 클러스터)



<차트 30> 클러스터 별 주 분포



<차트 31> 클러스터 별 전력량 분포

<차트 29>에 따르면, 클러스터링 결과 주별 최저기온의 양상과 비슷하여 함께 비교해보았다. 클러스터 1 은 겨울엔 따뜻하고 여름에 비교적 시원한 지역으로 state 1 과 11 이 완벽하게 함께 묶였다. 클러스터 2 는 겨울에 가장 추운 지역으로 고도가 가장 높았던 state 4 가 혼자 묶였다. 클러스터 3 은 1 월보다 2 월에 급격히 추워지고, 여름에는 가장 더운 지역으로, 대부분 state 6 에 해당한다. 클러스터 0 은 기온이 항상 중간에 해당하는 평균적인 지역으로 state 6 의 많은 관측치가 이곳에 해당한다. 대체로 state 2, 6, 8 이 기온 특성에 따라 클러스터 0 과 3 으로 분리되었다.

<차트 31>의 클러스터별 전력량을 확인해보면 겨울에 따뜻하고 여름에 시원한 클러스터 1 의 전력량의 확실히 낮은 분포를 띤다. 기온을 기반으로 한 범주가 전력량도 설명할 수 있어, 이를 모델링에도 활용하였다. 주 코드와 비슷하면서도, 매우 몸집이 컸던 state6 을 효과적으로 분리시키고, 비슷한 특징을 가지는 주는 한 범주로 묶을 수 있었다.

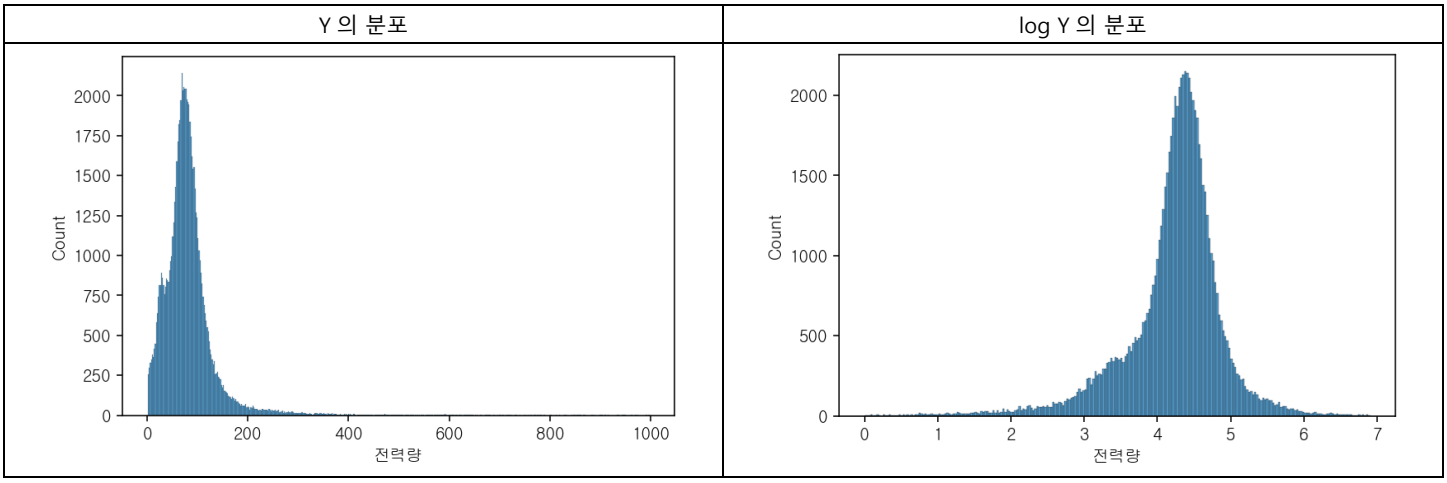
V. 모델링

모델링은 먼저 결측치를 처리하고, 크게 선형 적합 모형과 트리 기반 모형을 시도하였다. 또한, 여러 모형의 장점을 함께 반영하기 위해 앙상블 모형을 시도하였다. 앞서 살펴본 EDA 와 각 모형의 변수 중요도를 기반으로 다양한 변수 집합을 시도하였고, 가장 좋은 성능 지표를 보인 모형을 최종 모형으로 선정하였다. 모형의 성능 지표로는 RMSE(평균 제곱근 오차)를 사용하였고, 전체 데이터 중 랜덤하게 80%를 선택하여 학습 데이터로, 나머지 20%를 검증 데이터로 사용하였다.

a. 결측치 처리

이 데이터에는 6 개의 변수에서 결측치가 발견된다. 그 중, 최대풍속풍향, 돌풍풍속풍향, 최대풍속, 안개일수 등 4 개의 변수는 결측치 비율이 50%가 넘어, 사용하지 않고 제거하였다. 건축년도는 약 2.4%의 결측치가 있어 주별 중간값으로 채워 넣었고, 에너지등급은 약 35%의 결측치를 전체 데이터의 중간값으로 채워 넣었다. 35%도 많은 비율이지만, 에너지등급이라는 변수 의미가 전력량에 매우 큰 관계가 있어 결측치를 채워 사용하였고, 주별 중간값을 사용하기엔 아예 데이터가 존재하지 않는 주가 있어 전체 데이터의 중간값을 사용하였다.

b. 선형 모형



<차트 32> 종속 변수의 분포

선형 모형으로는 다중 회귀 모형과 라쏘 모형을 사용하였다. 두 가지 방법론에 대하여 Y 값 모델링, logY 모델링, PC 변수집합을 사용한 경우 총 3 가지 경우의 수를 시도하여 총 6 개의 모형을 비교하였다. 선형 모형 특성상 정규분포 가정을 따르는데, 전력량 분포는 오른쪽 꼬리가 매우 긴 분포이다. 따라서, 조금 더 정규분포의 형태를 띠는 log Y 값 모델링도 시도하였다. 기온 변수 간의 상관 관계를 줄여 모델을 해석하기 위해 36 개의 기존 기온 변수 대신 3 개의 PC 변수를 사용하는 방법 또한 시도한다. 중요한 변수를 해석하기 위해 모든 변수는 표준화하여 사용하였다.

Y 모델링		log Y 모델링		PCA 기온 변수 사용	
다중 회귀	라쏘	다중 회귀	라쏘	다중 회귀	라쏘
50.727	50.702	52.399	52.507	52.732	52.728

<표 9> 선형 모델링 결과

<표 9>에 따르면 모델링 결과, 로그 변환을 한 것보다 기존의 Y 값을 그대로 모델링한 경우가 가장 성능이 좋았고, 라쏘로 몇 가지 변수가 제거되었을 때 test RMSE 가 약간 더 개선되었다. PCA 기온 변수를 사용한 경우, 기존 정보의 75%만 사용하기 때문에 성능이 더 나빠진 것으로 보이나, 변수 개수가 30 개 이상 줄었음에도 RMSE 는 2 정도만 증가한 것을 확인할 수 있다.

라쏘 모형 결과		
변수명	**계수 (절댓값 큰 순)**	**계수가 0 인 변수**
에너지 등급	-20.146	9 월 최고 기온, 7 월 최고 기온, 난방 총화씨도수, 4 월 최저 기온, 6 월 최저 기온, 연중 최저 기온, 최저 기온 clus, 10 월 최저 기온, 11 월 최고 기온 등 14 개 변수
7 월 평균 기온	-15.920	
건물 유형 24	15.42	
11 월 평균 기온	-11.685	
1 월 평균 기온	11.496	
연간 적설량	9.302	
2 월 최저 기온	-8.920	

<표 10> 라쏘 모형 결과

모든 변수를 표준화시켰기 때문에 라쏘 모형의 계수값으로 변수 중요도와 중요하지 않은 변수를 알 수 있다. (Figure 2) 에너지 등급, 건물 유형이 매우 중요했고, 기온 변수 중에서는 7 월, 11 월, 1 월 등 여름과 겨울의 기온이 중요하게 작용했다. 또한, 기온 변수간 상관관계가 매우 높아 많은 기온 관련 변수의 계수가 0 으로 수렴하여 제거된 것을 확인할 수 있다.

c. 트리 기반 모형

트리 기반 모형으로는 랜덤포레스트와 그래디언트부스팅, XGBoost, LightGBM, CatBoost 총 다섯 가지 모형을 시도하였다. 각 모형에 대해 그리드 서치를 통해 하이퍼 파라미터를 튜닝하였고, 대부분 RMSE 40 초반대를 보이며 선형 모형보다 좋은 성능을 보였다. 간소한 차이로 캣부스트가 가장 좋은 결과를 보였으며, 더 좋은 성능을 내기 위해 다양한 앙상블과 변수 조합을 추가로 시도하였다.

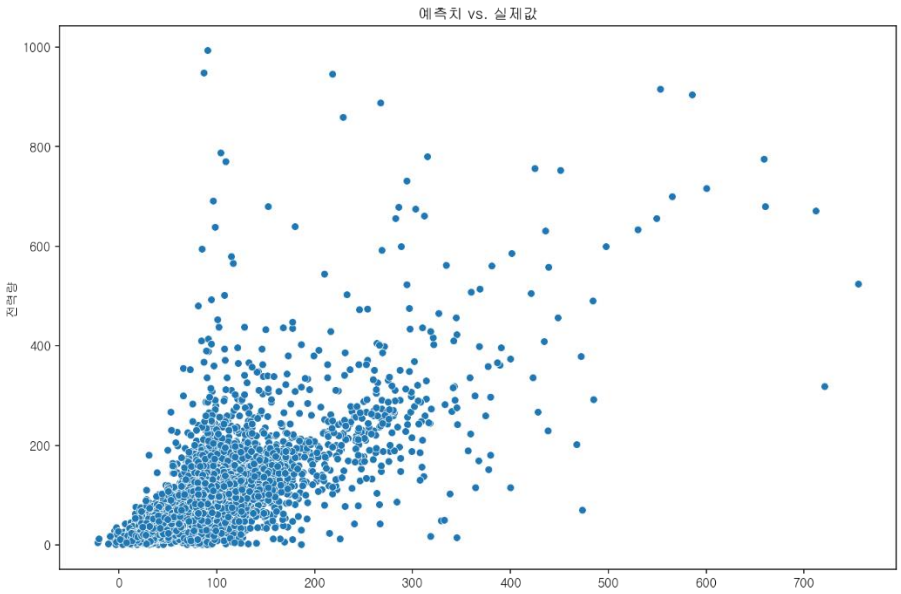
	랜덤포레스트	GBM	XGBoost	LightGBM	CatBoost
그리드서치 결과	max depth 20 min samples leaf 5 n estimators 100	learning rate 0.1 max depth 10 min samples leaf 10 n estimators 200	learning rate 0.2 max depth 20 min child weight 10	learning rate 0.1 max depth 10 min child samples 10 n estimators 200	learning rate 0.1 l2 leaf reg 0.01
test RMSE	41.767	40.505	46.771	40.923	40.372

<표 11> 트리 기반 모형 결과

d. 최종 모형 및 결과 해석

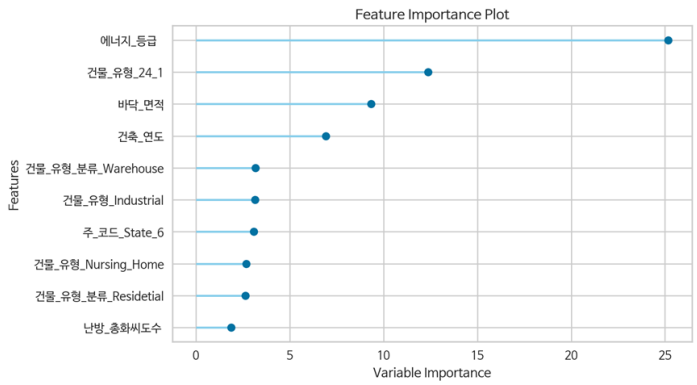
최종 모형 : 캣부스트, 랜덤포레스트, XGBoost 앙상블

test RMSE 37.971

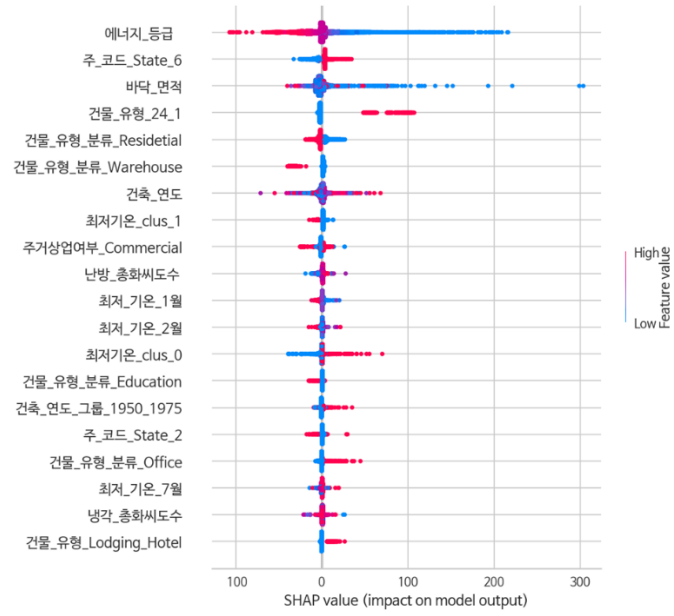


x 축 : 예측치, y 축 : 실제 전력량

변수 중요도



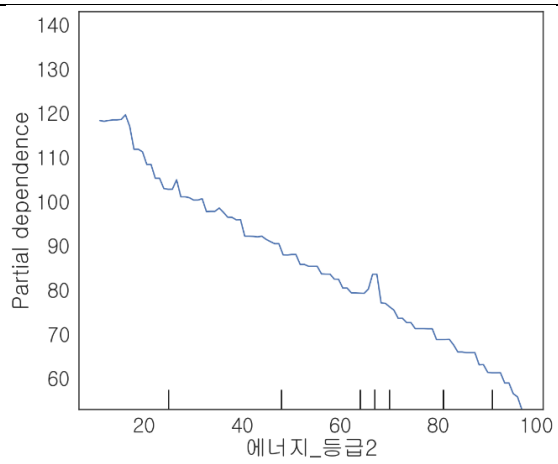
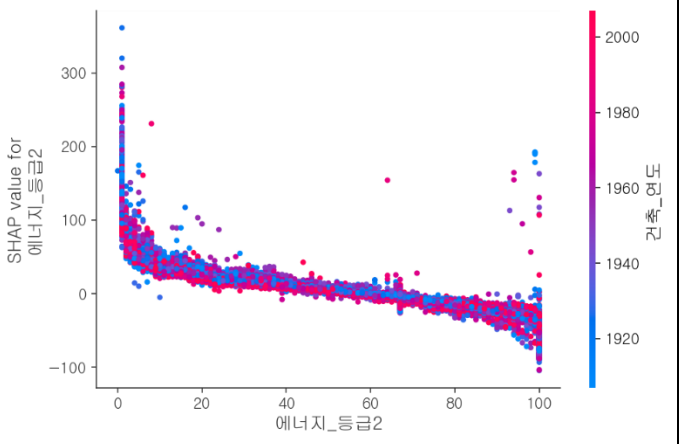
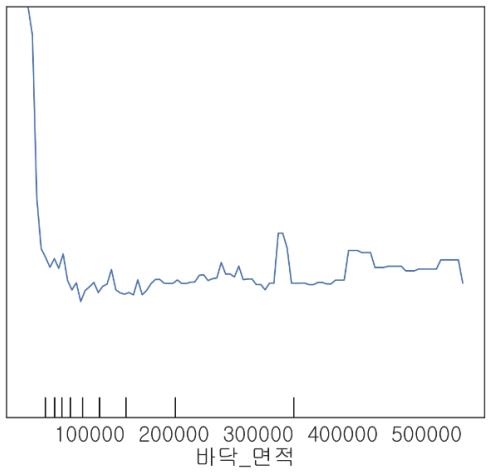
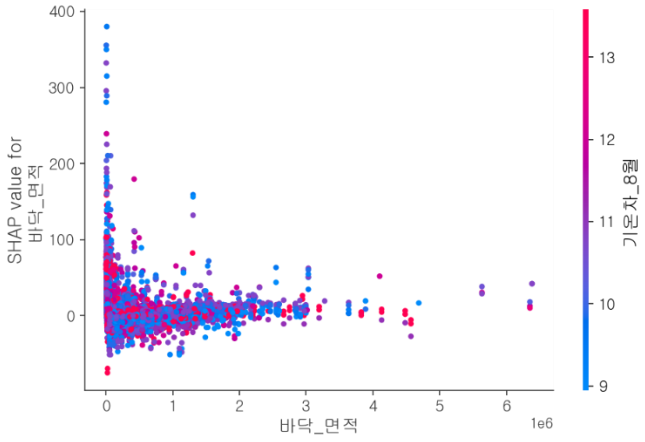
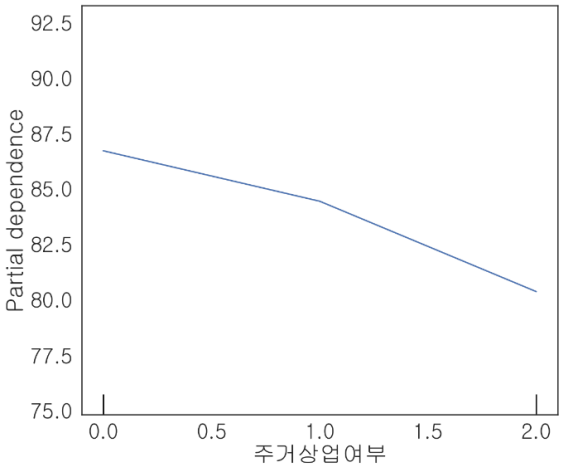
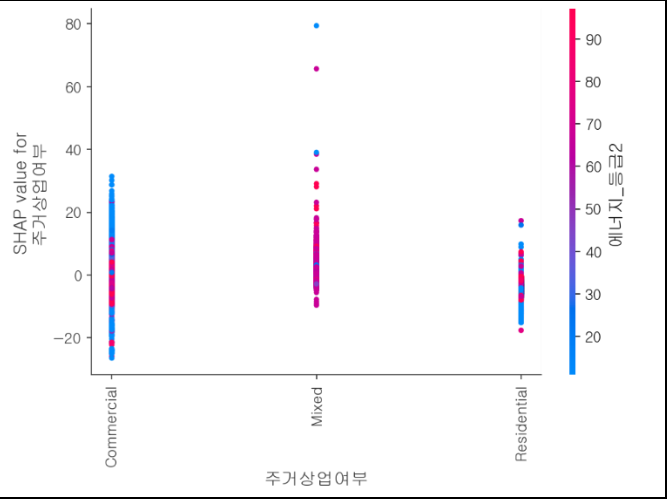
shap value

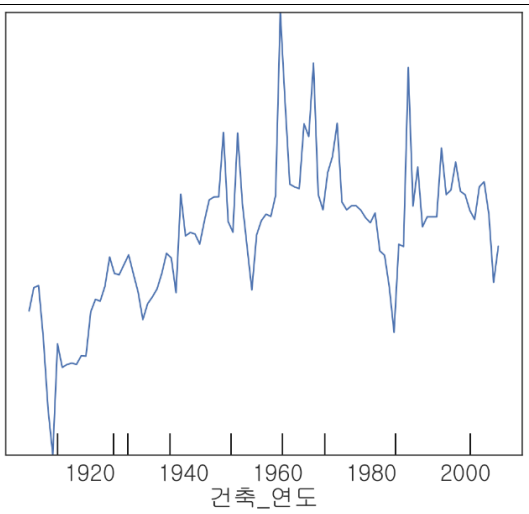
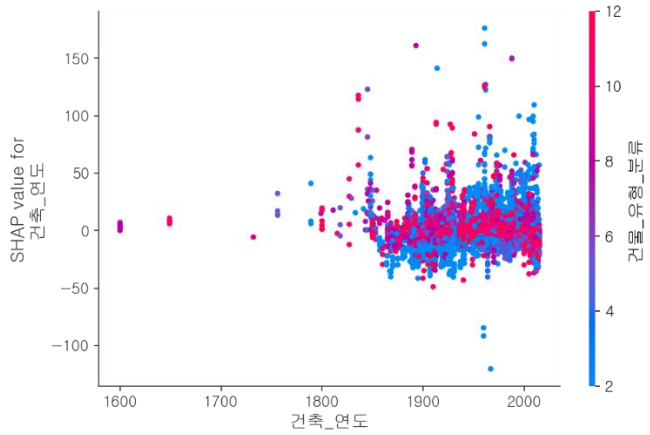


<차트 33> 최종 모형 결과 및 변수 중요도

최종 모형은 캣부스트와 랜덤포레스트, XGBoost 세 모형을 앙상블한 모형으로, test RMSE 는 37.971 이다. 이 모형의 변수 중요도를 살펴보면, 왼쪽은 사이킷런의 변수중요도로, 트리의 각 분기에서 얻은 성능 이득을 기준으로 계산한 값이다. 오른쪽은 shap value 로, 전체 데이터를 고정시켜놓고 해당 변수만 증가시켰을 때의 예측치 차이를 계산한 값, 즉 각 변수의 예측치에 대한 기여도이다.

공통적으로 에너지 등급이 가장 중요하고, 24 시간동안 사용하는 건물 유형과 바닥면적이 중요한 것을 확인할 수 있다. 또한 건물유형은 상업용일때 전력 소모가 많고, 건축연도는 1950 년부터 1975 년 사이인 경우 전력량을 더 높게 예측한다. 그 외에도 창고 유형, 요양원, 최저기온 클러스터 등이 중요하다.

	Partial Dependence Plot	Shap value
에너지 등급		
바닥 면적		
주거 상업 여부		

건축 연도		

<차트 34> 주요 변수 해석

주요 변수들의 Partial dependence plot 과 shap value 를 통해 각 변수와 전력량의 관계를 살펴보자. 먼저 에너지 등급은 높을수록 전력을 적게 사용하고, shap value 의 다른 변수와의 관계를 통해, 같은 에너지 등급 내에서는 최근에 지어진 건물일수록 전력량이 더 감소하는 것을 확인할 수 있다. 바닥 면적은 뚜렷한 선형 관계는 보이지 않지만, 매우 좁은 면적의 건물에서, 전력량이 높게 예측되어 에너지 효율이 좋지 않다는 것을 알 수 있다. 주거 상업 분류에서는 상업용 건물의 전력량이 가장 많고, 주거용 건물이 적게 소모한다.

건축 연도는 특이하게도 1960 년대까지는 최신 건물일수록 전력 소모가 많고, 1950 년에서 1980 년 사이의 전력량이 가장 높게 나타나, 이것이 앞서 변수중요도에 나왔던 이유라고 볼 수 있다. 아마, 더 오래된 건물들은 리모델링 등을 통해서, 좀더 에너지 효율성을 높인 경우가 많다고 해석해볼 수 있다. 또는, 전력량이 높았던 기간의 건축물들의 특징을 좀 더 조사해볼 필요가 있다.

VI. 결론

변수	해석
에너지 등급	에너지 등급이 높을수록 전력량 감소
건물 유형 24	24 시간 돌아가는 건물의 전력량이 높음
바닥 면적	바닥 면적이 매우 좁으면 전력량 증가
주 코드	state 6 일 때 전력량 증가 (많은 표본의 효과) state 2 일 때 전력량 감소 (비교적 최신 건물)
건축 연도	1950 ~ 1975 년에 지어진 건물에서 전력량 높음
주거상업여부	상업용 건물보다 주거용 건물의 전력량이 낮음
건물 유형	창고, 교육 – 낮음 / 오피스, 숙박, 요양원 - 높음
1, 2 월 최저 기온	겨울에 따뜻할수록 전력량 감소

<표 12> 중요 변수 요약

<표 12>은 해석한 내용을 PDP 와 Shap value 를 종합하여 중요한 변수부터 정리한 결과이다. 에너지 등급이 전력 소모량에 대해 잘 평가하고 있는 것을 확인할 수 있었고, 바닥 면적은 매우 좁은 경우에는 전력 효율이 좋지 않았다. 또한, 상업용보단 주거용 건물의 전력량이 낮으며 24 시간 돌아가는 건물이나 오피스, 요양원과 같은 유형의 전력량이 높았다. 여름보다는 겨울의 최저기온이 전력량에 많은 영향을 미치는 것 또한 확인하였다.

건축 연도는 전력량을 선형적으로 설명할 수 없었지만 1950 ~ 1975 년 사이에 지어진 건물의 전력 효율이 좋지 않은 것을 확인하였다. 이 데이터에는 정보가 부족하지만 더 오래된 건물은 리모델링으로 효율을 높였거나, 이 시기에 지어진 건축 방법에 문제가 있을 수 있다. 주 코드는 state 6 일 때 전력량이 증가하고 state 2 일 때 감소하는데, state 6 의 shap value 는 많은 표본의 효과로 보이며, state 2 는 비교적 최신 건물이기 때문으로 생각해볼 수 있다. 그 외에도 주별 정책이나 건물의 특성을 함께 고려하면 좋을 것이다.

이를 종합하면, 상업용 건물, 특히 오피스와 같은 시설에서 에너지 소모를 줄이기 위해 노력해야 하며, 이 데이터가 수집된 전체 지역은 추운 날씨에 전력량 증가 효과가 높아 난방에 사용하는 에너지를 효율화 하면 환경 부담을 많이 줄일 수 있을 것이다. 또한, 에너지 등급을 높이기 위해 노력한다면 전력량도 함께 줄일 수 있을 것이다.

새로운 건물을 건축할 때에, 건물의 위치, 기후, 용도, 크기 등이 비슷한 기존의 데이터를 이용하여 연간 전력량을 예상해보고, 가이드를 제안할 수 있으며 전력량을 줄이기 위해선 어떤 요소를 반영하는 것이 좋은 지 도움을 줄 수 있다. 또한, 기존 건축물에 대해서는 비슷한 그룹 내에서 본인 건물의 위치를 알려주고, 전력량을 줄이기 위한 조언을 줄 수 있다.

- 한계점 및 추후 연구 방향

데이터 분석을 하면서 제한된 건물 정보와 구체적이지 못한 기상 정보에 아쉬움이 있고, 구체적인 위치나 정보를 알 수 없어 외부 데이터를 사용하기도 어려웠다. 건물의 사용 인원, 난방 시스템, 층 수, 전력기기 개수 등 전력량에 영향을 미치는 충분한 데이터가 있다면 더 좋은 성능과 해석이 가능할 것이다. 또한 특정 주의 건물들은 같은 기온 데이터를 가지고 있었기 때문에, 개별 건물의 기상 정보를 알 수 있다면 정확한 해석이 가능할 것이다.

이번 프로젝트에서는 대회를 통해 주어진 데이터만 사용하여 특수한 목적에 따른 분석을 진행할 수는 없었지만, 각 나라에서 개별 국가의 상황이나 특성에 맞는, 필요한 데이터를 수집하거나, 여러 나라의 다양한 스펙트럼의 정보를 분석한다면 목적에 부합하는 더 좋은 분석을 할 수 있을 것이다.