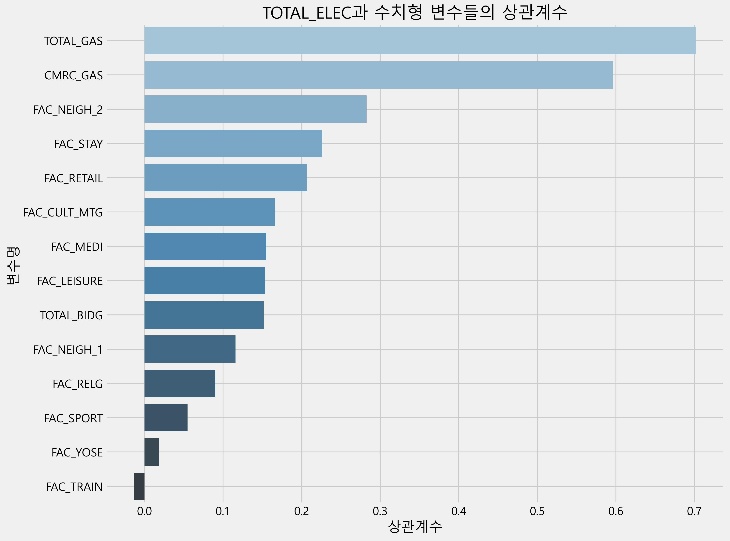
**데이터 시각화**

1. **상관관계 확인 : 연속형 변수**

* 상관계수 **0.6 이상이면 강한 양의 상관** : TOTAL\_GAS, CMRC\_GAS 같은 애들
* 0.3~0.6: 적당한 관련성
* 0~0.3: 거의 없는 수준

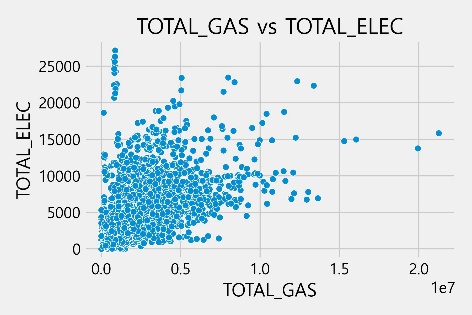
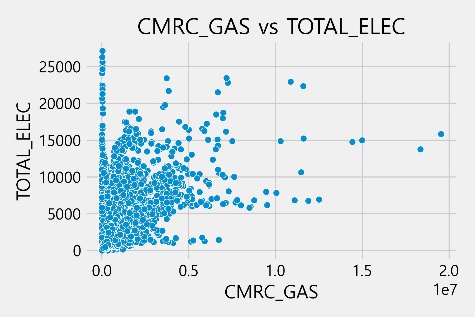
텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

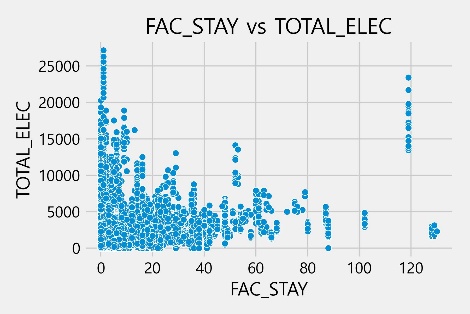
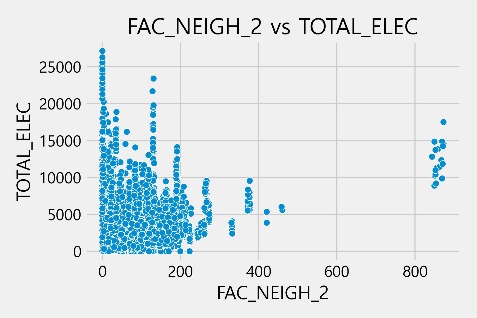
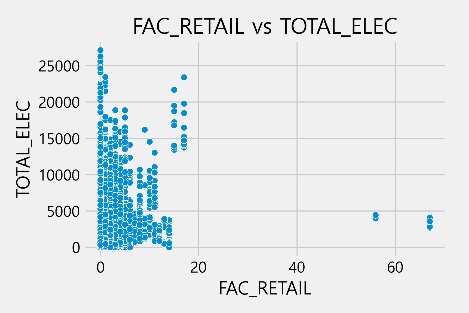
AI가 생성한 콘텐츠는 부정확할 수 있습니다.

**✅ 수치형 변수 vs 타겟 상관관계 해석**  
TOTAL\_GAS, CMRC\_GAS : TOTAL\_ELEC와 높은 양의 상관 → 매우 중요한 피처  
그 외 FAC\_NEIGH\_2, FAC\_STAY, FAC\_RETAIL : 0.2~0.3 사이 → 보조 피처로 유의미

🧠 해석:  
➡ 전기 사용량이 많은 상권일수록 가스 사용량도 많고, 상업 시설도 많음  
➡ 연료 수요와 편의시설이 전력 수요와 직결

1. **상관관계 확인 : 상위 5개의 상관계수 변수 산점도 확인**





✅ 산점도 분석

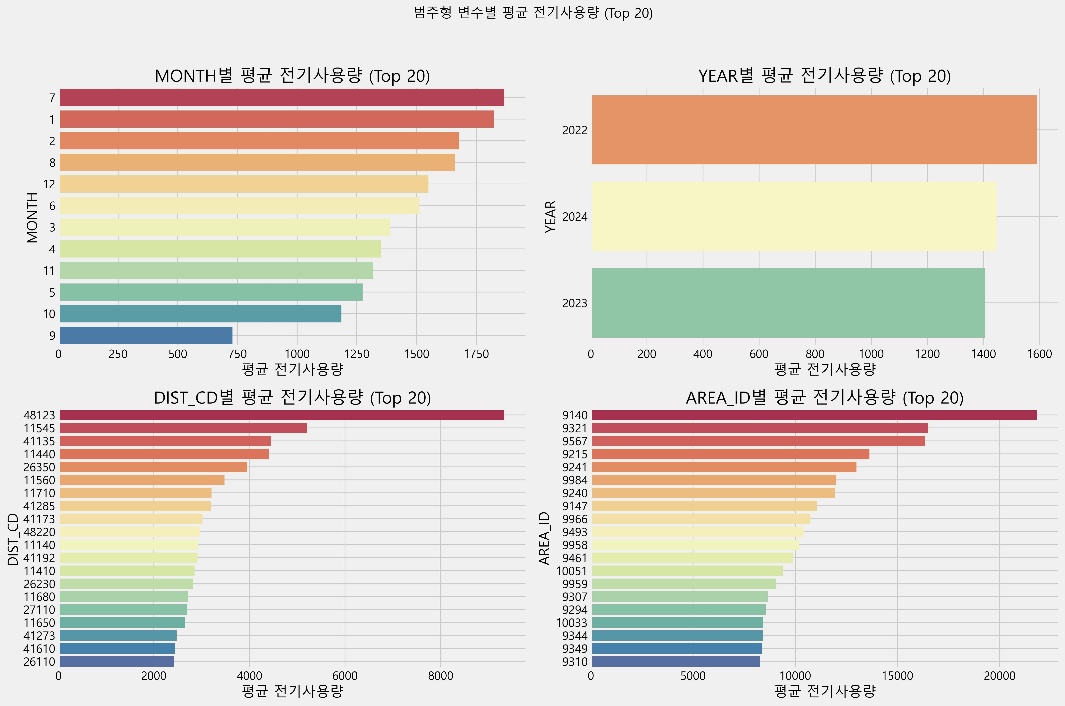
TOTAL\_GAS, CMRC\_GAS : 점들이 뭉치면서 우상향, 비선형성도 조금 있음

→ XGBoost 같은 트리 모델 적합

FAC\_NEIGH\_2나 FAC\_STAY : 일정 범위 넘으면 급증하는 경향

→ 로그 스케일로 변환해보면 좋음

1. **범주형 변수**



✅ 범주형 변수의 분포 해석 (Boxplot)  
MONTH, YEAR는 전기 사용량의 차이가 거의 없음  
➡ seasonality (계절성)가 약하다고 볼 수 있음

DIST\_CD, AREA\_ID는 큰 편차와 이상치가 많음  
➡ 상권/지역 고유 특성이 큼 → one-hot encoding or target encoding 추천

✅ 범주형 변수의 평균 전기 사용량 (Top 20)  
일부 DIST\_CD, AREA\_ID는 평균 전기 사용량이 엄청 높음  
➡ 해당 지역이나 상권 특성이 모델 예측력에 큰 영향 줄 수 있음  
➡ 범주형 변수 중 일부는 예측에 있어 핵심 힌트일 수 있음

1. **인사이트 도출**
2. 파생변수

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. 변수 변환

FAC\_NEIGH\_2, FAC\_STAY, TOTAL\_GAS 등 → log 변환 (np.log1p) 시 분포 안정화 가능

정규화보단 트리 모델(XGBoost)에서 변환 효과 큼

1. 범주형 변수 처리

AREA\_ID, DIST\_CD 등 → 수가 많아서 원-핫보다 타겟 인코딩이 효과적일 수 있음

(단, KFold 타겟 인코딩으로 리크 방지 필수)

MONTH는 sin/cos 변환해서 seasonality 약간 반영해도 무방

1. 모델링 전략

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. **데이터 결측치 처리**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

✅ 결측치 상황 요약  
총 13개 변수에 각각 3개씩 결측치  
전체 데이터에서 0.011% 미만 수준 → 매우 극소수!  
결측된 변수들은 모두 시설 개수 관련 수치형 변수들  
  
📌 처리 제안: 0으로 채우기 (fillna(0))  
시설 수가 결측이라는 건 해당 시설이 없었을 가능성이 높음  
특히 FAC 계열은 실제로 0인 지역이 많은 변수들임  
결측치를 평균/중앙값으로 채우면 의미 없는 값이 들어갈 수 있음

1. **데이터 이상치 처리**

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

✅ 해석 요약

텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

📌 처리 방법

텍스트, 폰트, 영수증, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

🎯 결론  
log1p로 연속형 분포 조정 → RMSE 민감도 ↓  
clip으로 희소한 변수에서 이상하게 튄 값 제한  
최종적으로는 변환 전/후 모델 성능 비교해서 결정