분석용 데이터랑 trian_y 값이 서로 다른 이유.
1. 소수점 반올림 차이
소수점 반올림 차이라기엔, 값의 차이가 너무 크다.

~~2. 결측치, 이상치 존재~~
SCADA 데이터에 NaN이 섞여 있거나 측정 오류(0 또는 비정상 값)가 있을 경우.
데이터를 보면 NaN, 측정 오류가 섞여 있지 않음. 코드로 확인함.

파일만 보고 결측치 판단 어려운 이유 Excel에서 눈으로 봤을 때 셀이 비어 있어도
pandas로 읽을 때 NaN으로 처리되지 않을 수 있다. (특히 공백 문자열 등)

반대로 어떤 값이 있어 보이지만 실제로는 숫자 포맷 파싱 실패로 인해
NaN으로 읽힐 수 있습니다.

~~3. WTG01 ~ WTG09번까지가 정확히 맞는지_WTG10번이 존재하는 건 아닐지?~~
말이 안 됨. 아니라면 보조 터빈, 시험 설비, 계통손실 보정 등등 고려하지 않았는지?

4. train_y 데이터 값이 정제된 값일 수 있다. (가장 유력)
즉, train_y는 SCADA 데이터가 아니라 모델 학습을 위한 전처리된 값일 수 있음.
(센서 에러 제거, 이상치 보정 등등) - 하지만, 이 부분은 직접 확인할 수 없다.

5. 누적값과 순간값의 차이
누적값 - 최종 발전량에서 시작 발전량을 빼주는 것 (풍력 발전량을 측정할 때 누적값 활용)
즉, 계산 방식이 잘못되면 실제보다 훨씬 크게 계산되거나 작게 계산돼.

순간값 - 그때 그때 발전량을 다 더해주는 것

분석용 데이터와 train_y 데이터 오차가 있다면
분석용 데이터를 train_y 데이터에 맞춰서 보정하는 것이 일반적