선형회귀 분석을 이용한 전력수요 예측

ITM 20235575 안지석

1. **서론** 
   1. **연구 배경**

**전력수요 예측**은 전력 시스템 운영과 안정적 에너지 공급에 있어서 매우 중요한 역할을 한다. 전기는 현대사회의 일상 필수품이기 때문이다. 아래는 전력수요 예측이 중요한 이유를 나타냄.

* **공급 계획과 발전 계획을 통해 최적의 자원 할당**: 전력수요 예측은 향후의 수요에 대응하기 위해 적절한 자원 할당이나 발전 계획을 수립하는 데 도움을 주며, 발전소들은 예측된 수요에 맞게 발전 계획을 조정하여 효율적이며 안정적인 에너지 공급 수행.
* **시스템 안정성 유지 및 부하 균형**: 전력수요 예측은 시스템 부하 균형을 유지하는 데 도움이 됩니다. 전력은 실시간으로 공급과 수요가 일치해야 하는 특징을 가지며, 예측치를 기반으로 에너지 생산 및 공급을 유지하여 전력 시스템의 안정성 확보.
* **변동성 신재생 에너지 통합**: 재생에너지의 특성상 발전량이 불안정하므로, 전력수요 예측은 신재생 에너지의 통합에 필수적임. 신재생에너지와 전통적 발전원의 효율적인 운영을 위해서는 정확한 예측이 요구됨.
  + 종합적으로, 전력수요 예측은 전력 시스템의 효율성, 안정성, 경제성, 지속가능성 등 다양한 측면에서 중요한 역할을 하며, 특히 **탄소중립을 위한 에너지 전환 및 신재생 에너지의 도입으로 예측이 더욱 중요해**지고 있음.
  1. **연구 목적**

본 연구는 **공공데이터포털에서 제공하는 전력수요 및 연료원별 전력거래량 데이터를 활용하여 전력수요와 연료원별 전력거래량의 선형회귀분석을 수행하고, 더 나아가 예측 모델 구축, 변수 중요도 평가, 모델의 적합성 검토 등을 수행**함. 선형회귀 모델을 구축하여 통계적으로 해석(설명)하고, 새로운 입력데이터의 미래 반응변수 값을 예측해보고 평가하는 데 목적이 있음

* 1. **연구 범위 및 제한**

2021년 1월 1일부터 동년 12월 31일까지 시간별 전력수요량 데이터와, 연료원별 전력거래량 데이터를 활용한다. 연료원은 24개로 구성되며, 값이 모두 0인 가스압과 RPS는 제외하였음.

1. **문제 정의**

**2.1 연구 주제**

전력수요 및 연료원별 전력거래량 데이터를 활용하여 전력수요와 연료원별 전력거래량의 선형회귀분석을 수행하고 전력수요 예측 모델 구축

**2.2 변수 정의**

- 데이터 출처: 본 분석에 필요한 데이터는 공공데이터포털에서 제공하는 데이터를 활용하였으며, 한국전력거래소\_연료원별 전력거래량, 한국전력거래소\_시간별 전력거래량 두 데이터 셋을 정제하였음

- 종속변수: 2021년 1월 1일부터 동년 12월 31일까지 전력수요

- 독립변수: 동 기간 동안의 24개 연료원별 전력거래량

**2.3 문제 설명**

우리나라의 전력거래 시장은 국가에 의해 통제되는 단일 시장으로, 전력거래소를 통해 이루어짐. 발전사업 허가를 받은 회사만 거래 참여가 가하며, 전력생산량은 발전사가 임의로 결정하는 것이 아니라, 전력거래소가 특정시간의 전력수요를 예측해서 각 발전소에 발전비용이 저렴한 순서대로 발전을 지시함. 이러한 이유로 **전력수요의 예측은 안정적 전력 공급뿐 아니라, 발전사의 운영 측면에서도 중요함. 전력수요 예측 모델 구축을 통해 연료원별 전력거래량과 전력수요와의 관계를 파악해 볼 필요가 있음.**

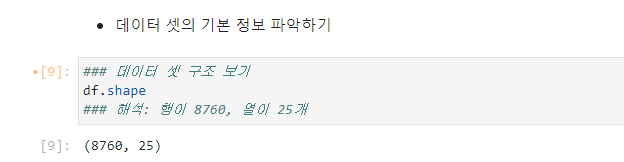
1. **EDA(Exploratory Data Analysis)**

**3.1 데이터 셋 기본 정보 파악**

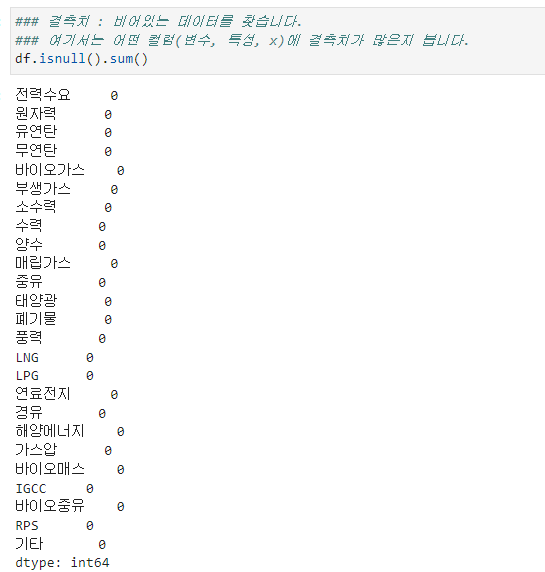
위에서 설명하였듯이 본 분석에서 활용한 데이터는 공공데이터포털에서 제공하는 데이터로, 출처는 한국전력거래소로 2021년 시간대별 전력수요와 연료원별 전력거래량임. 데이터셋을 불러와 데이터의 구성을 살펴보기 위해 df.tail(10)을 활용하여 마지막 10개 데이터를 불러왔으며, 데이터는 8760개 인 것을 확인할 수 있음(2021년 365일 x 24시간). 전력수요 값과, 원자력, 유연탄, 무연탄 등 24개 연료원별 전력거래량 데이터가 포함되어 있는 것을 아래 그림으로부터 확인 가능함.



데이터셋의 구조는 df.shape를 통해 확인할 수 있으며, 8,760개 행과 25개 열로 이루어져 있으며, 종속변수(전력수요) 1개를 제외한 24개 독립변수로 구성되어 있음을 다시 확인하였음.



df.isnull().sum()을 활용하여 결측값을 포함하고 있는 데이터를 확인할 수 있으며, 본 데이터에는 결측 데이터가 없음을 확인하였다. 다만, 가스압과, RPS는 모든 데이터가 0으로 되어 있어 독립변수에서 제외하였음

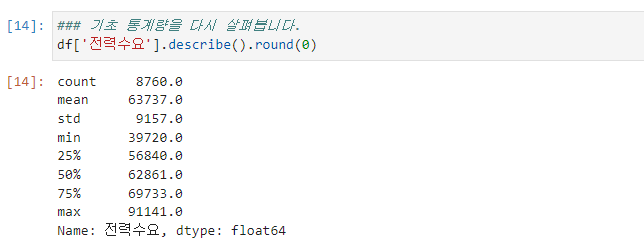


모든 변수는 숫자형이고 df.describe().round(0)을 통해 소수점 첫째자리에서 반올림한 기본 통계를 확인할 수 있음. 평균 전력수요는 63,737MW이며, 주요 연료원 두 개만 살펴보면 유연탄의 평균 전력거래량은 21,370MW, 원자력의 평균 전력거래량은 17,175MW임을 확인할 수 있음.

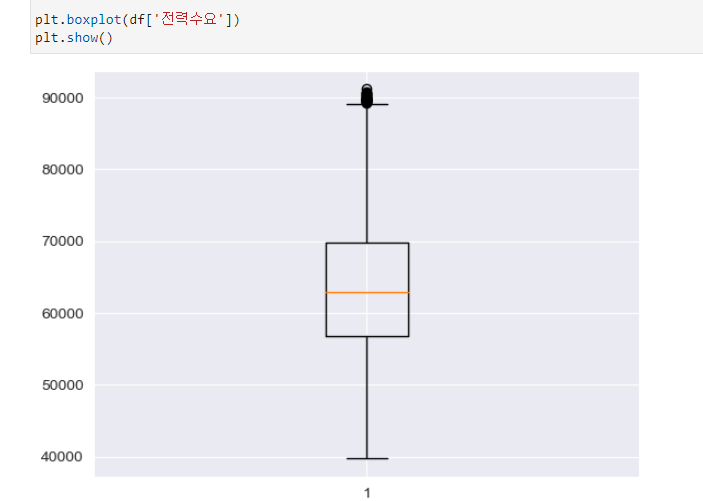
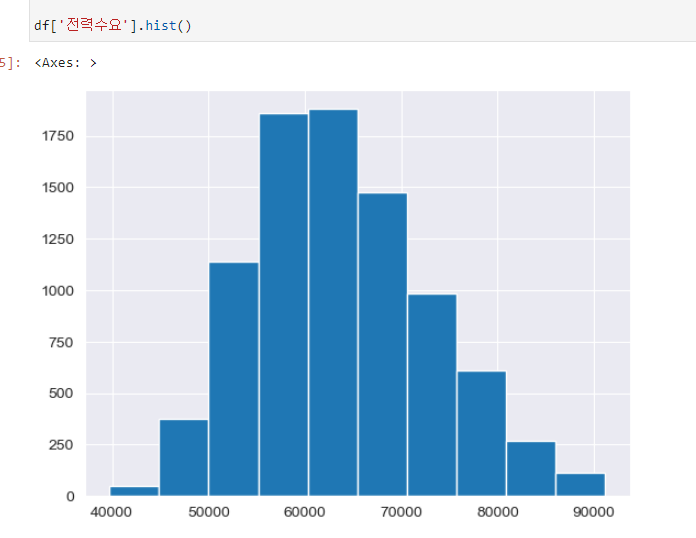


**3.2 종속변수의 탐색**

종속변수이자 타겟변수인 전력수요의 기초통계량은 df[‘전력수요’].describe().round(0)을 통해 소수점 첫째자리에서 반올림한 값으로 살펴볼 수 있음.

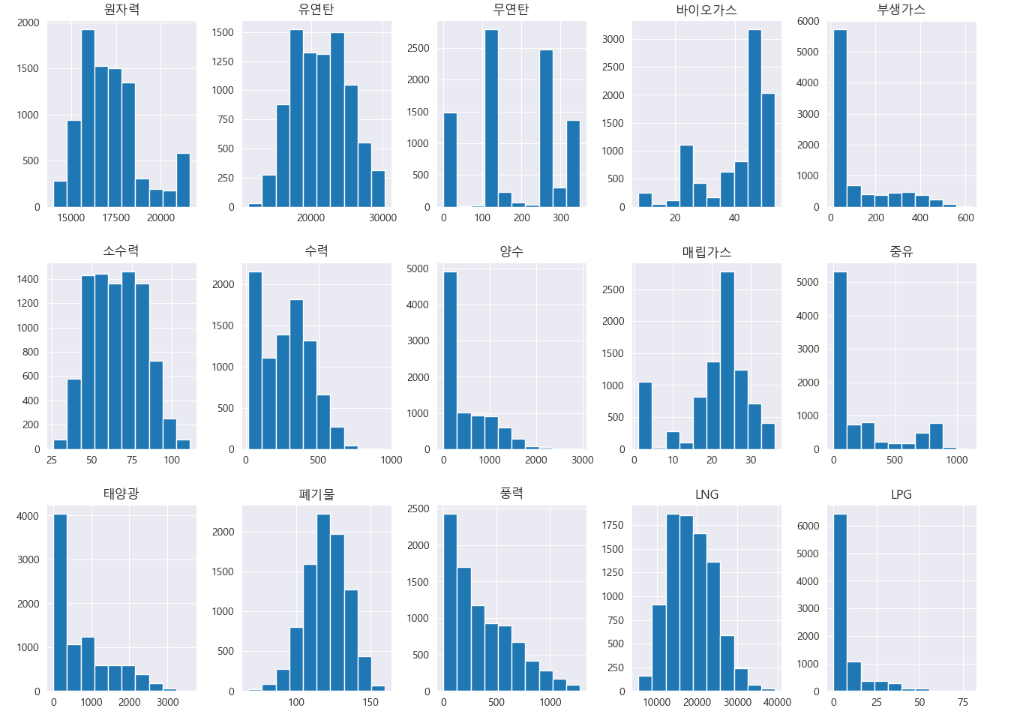


통계 수치로만 데이터를 파악하는 것이 아니라 데이터 시각화를 통해 실제 데이터의 분포를 파악하는 것도 필요함. 히스토그램, 박스플롯 등으로 시각화하여 데이터의 분포를 한눈에 볼 수 있음.



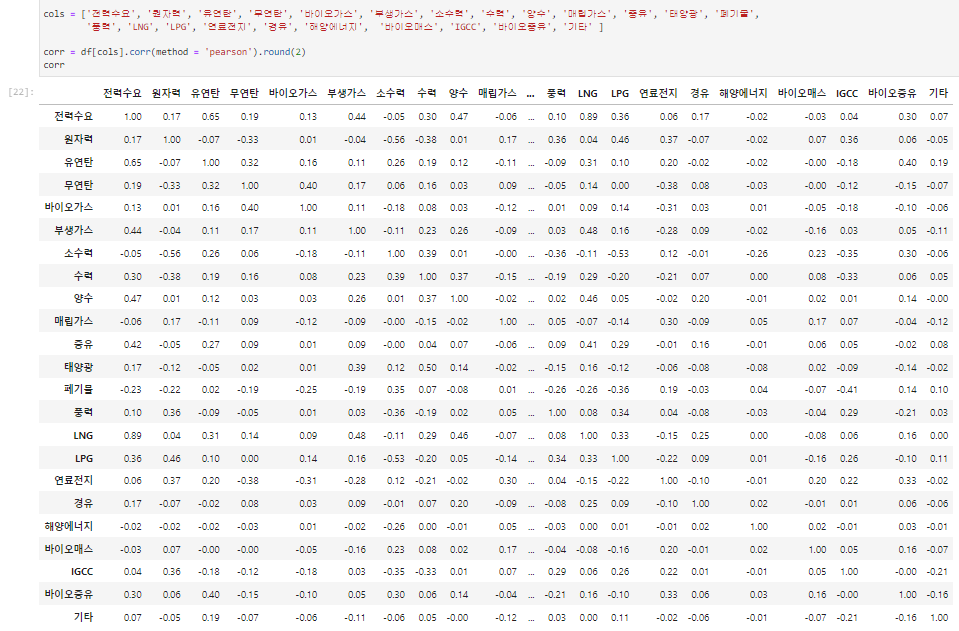
**3.3 독립변수의 탐색**

독립변수들의 분포도 확인할 수 있음. 독립변수가 많기에 개별 그래프로 나타내어 확인할 수 있으나, 분포를 확인할 때 연료원별로 전력거래량의 축 범위가 다르다는 것을 명심해야 함. 예를 들어 원자력의 경우는 만 단위를 나타내나, 무연탄의 경우 백 단위를 나타냄. 동일한 축으로 그래프를 제작하여 제시하면 정보 전달에 있어 왜곡을 피할 수 있음.

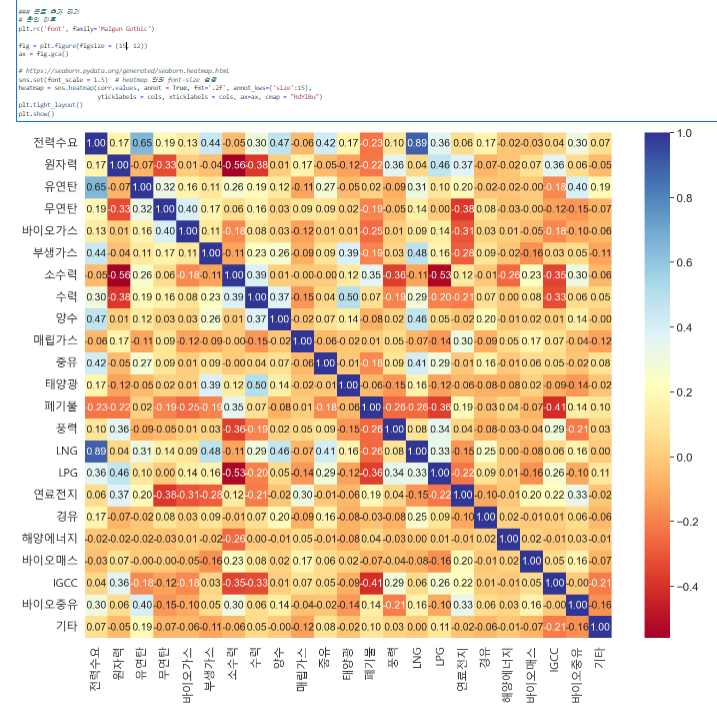


**3.4 독립변수(x)와 종속변수(y)의 관계 탐색**

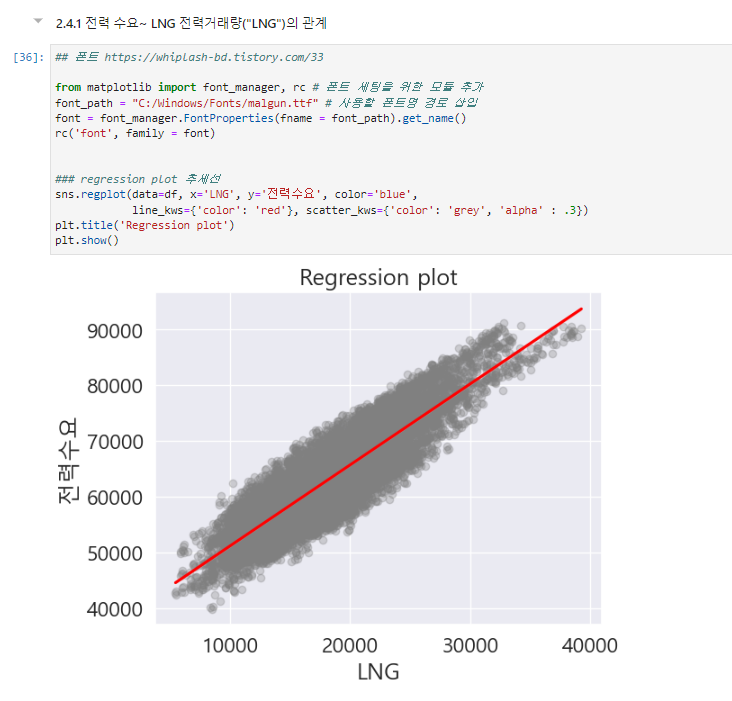
독립변수와 종속변수의 관계를 탐색하기 위해서는 상관관계 분석에서 사용하는 대표적인 지표인 pearson 계수를 사용하였으며, 변수간의 상관관계를 파악할 수 있음.

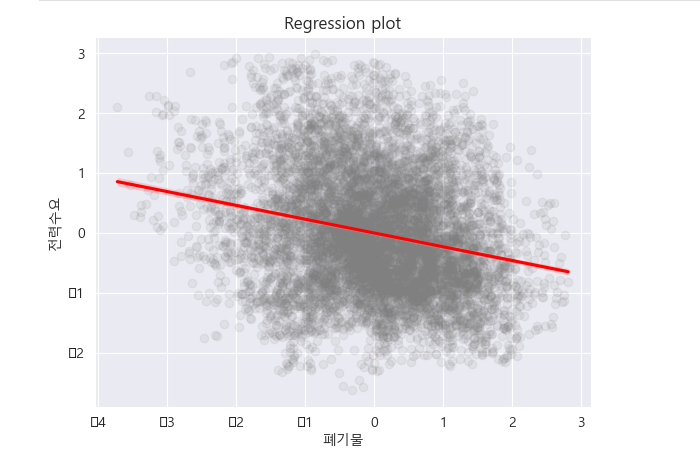


숫자로는 변수들의 상관관계를 한눈에 파악하기 어려우며, 데이터 시각화가 필요함. 대표적인 시각화 라이브러리인 seaborn을 통해 히트맵으로 시각화함. 색이 파란색일수록 양의 상관관계가 크며, 색이 붉을수록 음의 상관관계를 가짐. 우리의 관심사이자 종속 변수인 ‘전력수요’와 다른 독립 변수간의 상관관계를 살펴보면, 전력수요는 LNG와 0.89의 가장 높은 양의 상관관계를 보이며, 유연탄과는 0.65의 상관관계를 보임. 주요 연료원인 원자력은 0.17로 낮은 상관관계를 보였음.



전력수요와 상관관계가 가장 높은 LNG를 예시로 들어 scatter plot으로 시각화하고, 회귀선을 나타내보면 아래와 같음. 시각화를 통해 전력수요가 증가할수록 LNG 전력거래량도 증가하는 것을 직관적으로 확인할 수 있음.







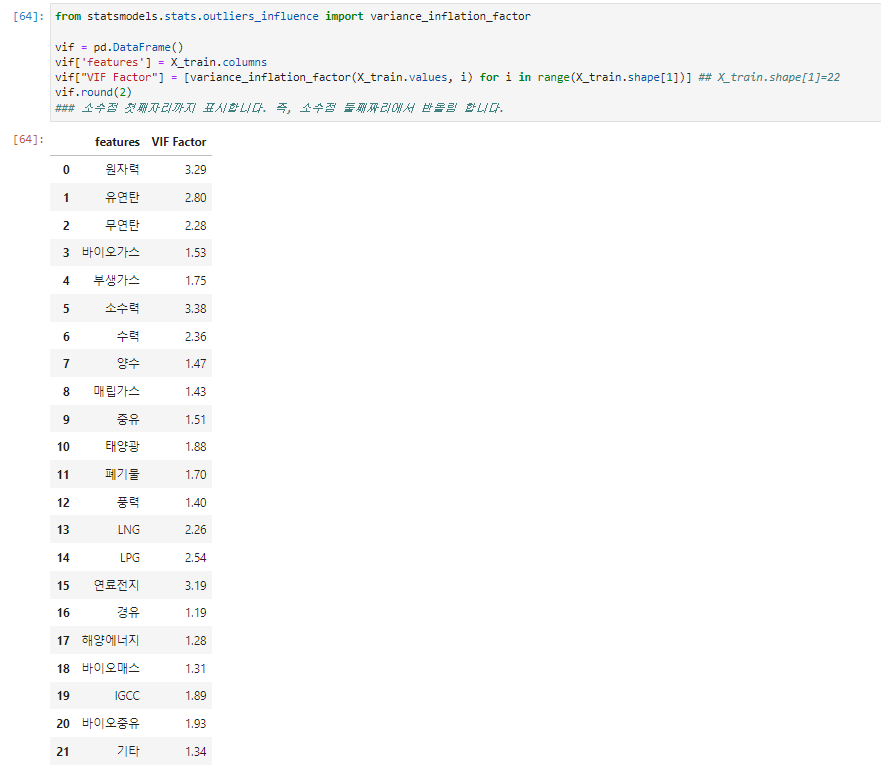
1. **전력수요 예측 모델링(회귀분석)**

**4.1 모델 구축을 위한 데이터 전처리**

변수들을 활용하여 전력수요를 예측하는 회귀모델을 만들기 위해서는, 변수들의 표준화가 필요함(이 경우의 경우에는 독립변수와 종속변수가 모두 MW 단위로 동일하기에 필요 없을 수도??). Scikit-learn 라이브러리를 통해 평균 0, 표쥰편차 1로 표준화를 한 결과는 다음과 같음.



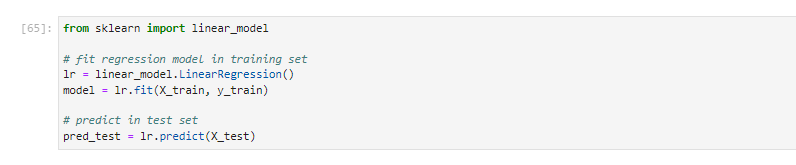
모델이 학습할 데이터를 training set, 모델의 성능을 테스트하기 위해 사용할 데이터를 test set을 구분함. 역시, Scikit-learn 라이브러리를 통해 train set과 test set으로 구분하였으며, test size는 0.2로 설정하였음. 하나의 예측 변수가 다른 변수와의 상관관계가 높은 경우(다중공산성이 존재하면), 회귀 분석 결과에 부정적 영향을 미칠 수 있기에, 모델링 전에 다중공산성의 존재 여부를 확인함.

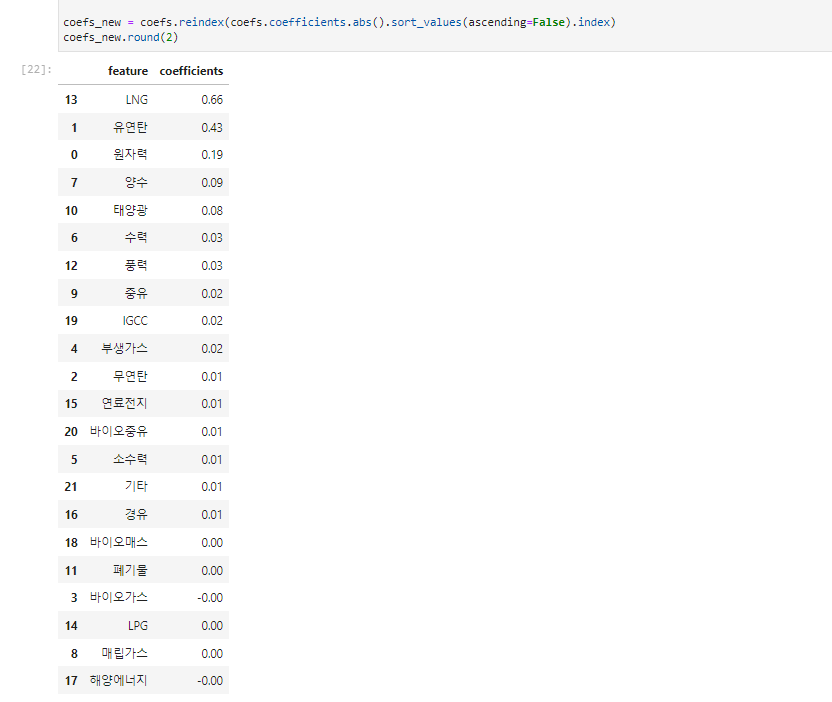


일반적으로 VIF Factor 값이 10 이상일 경우 다중공산성이 있다고 판단하는데, 본 분석 대상 데이터는 VIF Factor 값이 가장 높은 것이 원자력 3.29로, 본 독립변수들 사이에서는 다중공산성이 존재하지 않는 것으로 파악하였음. 따라서, 모델링에서 다중공산성으로 인해 분석에서 제외할 독립변수는 없는 것으로 판단하였음.

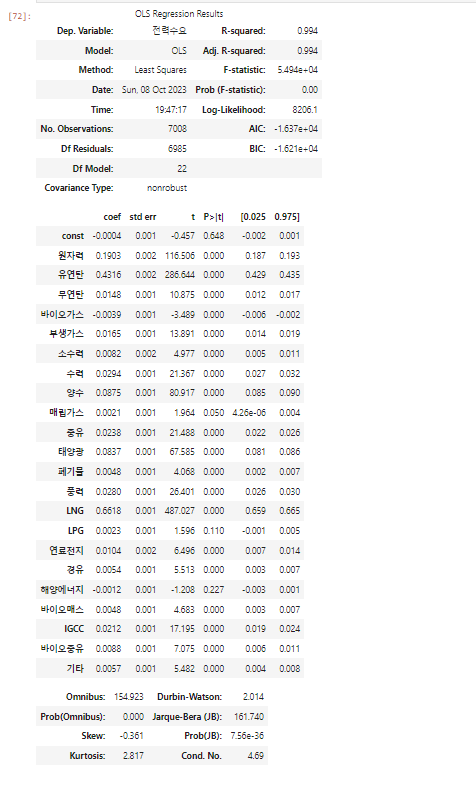
**4.2 회귀모델링 및 모델 해석하기**

회귀모형을 구축하고 해석하기 위해서는 우선 training set을 대상으로 선형회귀 예측모형을 학습시키고, 도출된 모델을 test set에 적용해 전력수요를 예측하여야 함. 이 예측 결과는 실제 전력수요 값과 비교하여 모델의 예측 성능을 평가하는데 활용함. Scikit-learn의 linear model을 통해 모형을 세우고, 독립변수별 회귀계수를 절대값의 내림차순으로 확인한 결과는 다음과 같음.





변수가 유의미한지 확인하기 위한 유의성 검정을 실시함. Statsmodels의 OLS(Ordinary Least Seuares)를 사용하여 선형회귀분석을 수행함. OLS는 가장 기본적인 결정론적 선형회귀방법으로 잔차제곱합을 최소화하는 가중치를 구하는 방법임. R-square는 결정계수로 모형의 성능을 나타내며, 회귀계수 coef는 x가 한단위 증가할 때 y가 얼마나 증가하는지 확인할 수 있음.



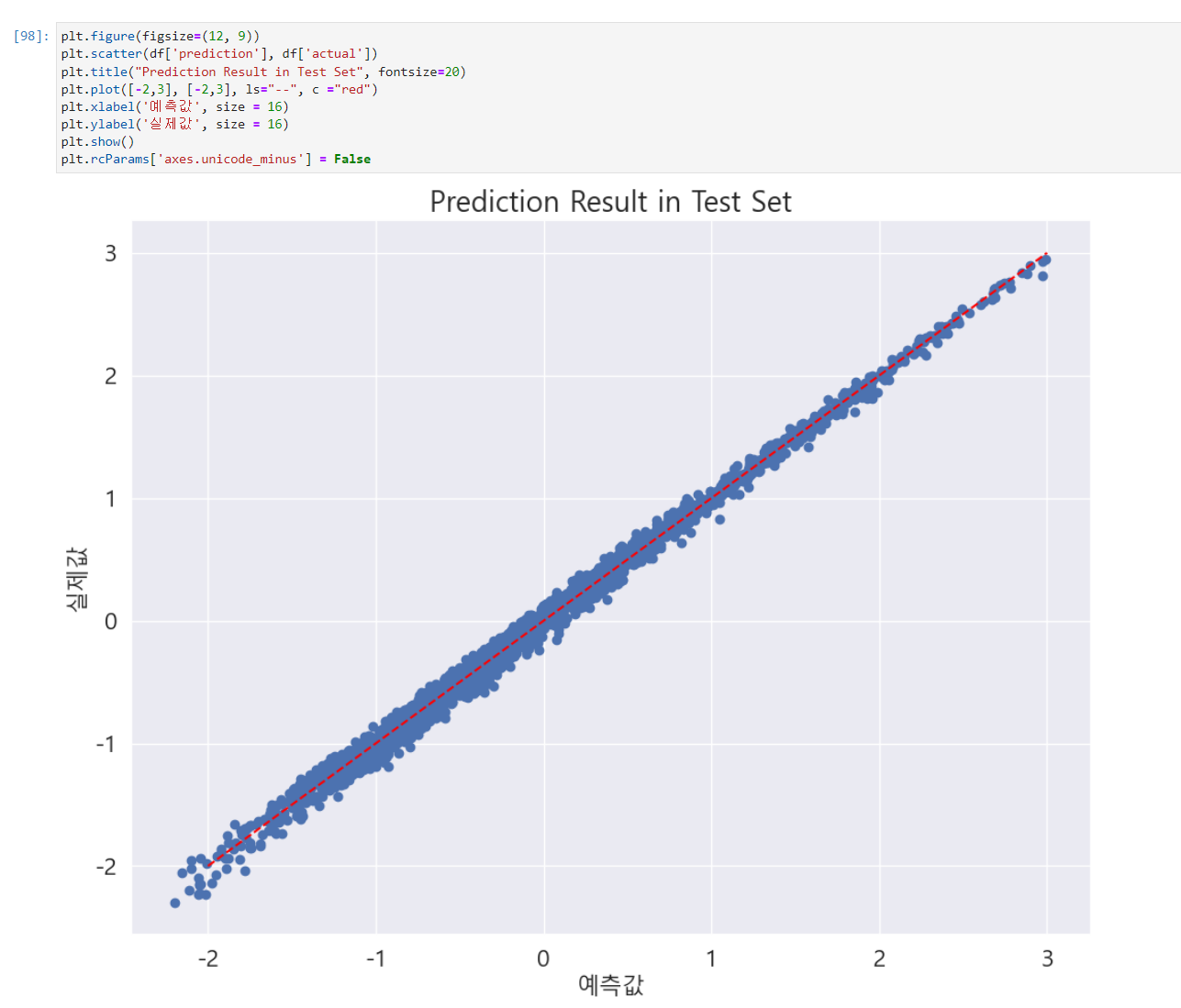
매립가스, LPG, 해양에너지는 p값이 0.05보다 크므로 통계적으로 유의하지 않으며, 그 외 연료원의 경우는 통계적으로 유의하다고 판단할 수 있음. 모델의 성능을 개선하기 위해서는 유의하지 않은 변수들을 제거한 후 다시 모델을 구축해야 하며, 변수 제거 후에도 모델의 성능이 크게 개선되지 않을 수 있으나, 통계량 값들을 유의미하게 해석할 수 있음.

회귀계수 coef를 살펴보면 LNG가 0.6618로 가장 높으며, 이는 전력수요 1이 증가할 때 LNG는 0.6618이 증가한다는 것을 의미함. 유연탄 0.4316, 원자력 0.1903을 보이며, 이 주요 세 연료원을 제외한 연료원은 상관계수가 미미한 것으로 파악됨.

**4.3 모델 예측 결과 및 성능 평가**

학습된 예측 결과를 한눈에 파악하기 위해, 학습한 모델을 test set에 적용하여 전력수요를 예측함. 예측 결과를 확인하기 위해 실제값과 예측값을 한 그래프에 출력하여 시각화함. 예측값과 실제값이 유사한 것을 확인할 수 있음.





Training set과 test set의 결정계수인 R square 값은 0.994로 나타났으며, 이는 본 모델이 전체 데이터의 99.4%의 설명력을 갖고 있다는 것을 뜻 함. RMSE는 실제값과 예측값의 차이를 나타내는 척도로 각각 0.075, 0.074로 나타났음.



1. **결론**

본 분석을 통해 **전력수요와 22개 연료원별 전력거래량의 선형회귀분석을 수행하고, 더 나아가 예측 모델 구축, 변수 중요도 평가, 모델의 적합성 검토 등을 수행**함. 선형회귀 모델을 구축하여 통계적으로 해석(설명)하고, 새로운 입력데이터의 미래 반응변수 값을 예측해보고 평가하였음.

전력수요와 연료원들간의 상관계수를 파악해보면, 전력수요는 LNG와 0.89의 가장 높은 양의 상관관계를 보이며, 유연탄과는 0.65, 원자력과는 0.17의 상관관계를 보였음. 회귀모형을 살펴보면, 매립가스, LPG, 해양에너지의 경우 통계적으로 유의하지 않은 수준으로 보이며, 회귀계수의 경우 LNG, 유연탄, 원자력 순으로 높은 수치를 나타냄. 모델 적합성 관련해서는 R square 값이 0.994로 나타나, 전체 데이터 셋의 99.4%를 설명할 수 있다고 해석할 수 있음.

**전력수요에 영향을 미치는 요인은 기온, 휴일유무 등의 변수도 존재하기에, 향후 분석은 이러한 데이터 확보를 통해 모델을 구축하면 바람직할 것으로 생각함.**