데이터 통계량을 확인한 결과 -999 값과 음수 값이 존재한 것을 확인할 수 있었다. 먼저, -999 값은 일반적으로 누적된 값으로 다른 값들과 구분을 하기 위해 전혀 연관되지 않은 값으로 볼 수 있었다. 따라서 -999 값은 일괄 결측치로 처리한 후 결측치 보간을 진행 할 것이다. 확인 결과 자외선지수에 53,207개, 그 외의 밴드들에는 18,060개의 결측치가 있음을 확인하였다. 다양한 결측치 보간법 중 하나를 선택하기 위해 회귀모델 중 엘라스틱넷을 통해 값의 결과가 가장 좋은 interpolate 결측치 보간법을 이용하기로 하였다. interpolate란 값에 선형으로 비례하는 방식으로 결측값 보간하는 함수이다. 결측치를 보간하는 과정에서 전체 데이터와 자외선 결측치가 포함된 행을 지운 데이터의 결과를 비교하기 위해 두 데이터의 결측치를 보간한 후 회귀모델인 엘라 스틱 넷을 통해 더 유의미한 결과인 전체 데이터를 사용하기로 하였다. 다음으로, 가시채널과 근적외채널의 밴드에 음수값이 존재하였다. 멘토링 결과 두 채널의 특성상 야간에는 측정이 불가능하고 음수값이 존재할 수 없다는 것을 확인할 수 있었다. 이를 주간 야간으로 구분하기 위해 변수 중 태양천정각을 이용했으며 태양천정각의 90도를 기준으로 90도 아래로는 주간, 90도 위로는 야간임을 확인했다. 이를 통해 가시채널과 근적외채널 중 야간의 값과 주간의 음수값을 0으로 처리했다. 이와 관련하여 오입력된 자료들을 삭제하지 않고 0으로 처리한 이유는 데이터 삭제 없이 최대한의 데이터를 분석하기 위함이다.

데이터의 범위는 검증데이터의 기간이 6월이므로 2020년과 2021년의 5,6, 7월 데이터만 분석에 이용할 예정이다. 변수를 줄이기 위해 5개의 채널별 밴드들의 통계량 수치와 상관관계를 확인한 후 평균을 이용하여 하나로 묶어주었다. 또한 단파적외채널, 수증기채널, 적외채널의 단위가 K(Kelvin)으로 같으므로 세 개의 채널을 하나로 다시 묶어주었다. 변수 중요도 확인한 결과 채널1과 채널2는 높은 값의 결과가 나왔기 때문에 추가 전처리하지 않았다. 현재 변수 중 태양천정각을 기준으로 자외선지수가 달라질 것으로 예상하여 0~45 이하, 45~90 이하, 90~135 이하, 135~180 이하로 나눈 후 새로운 컬럼 'time'을 생성하였다. 변수 중 관측소지점과 지면타입, time은 범주형 데이터이므로 숫자형으로 처리하기 위해 원핫인코딩을 진행하였다.

+ EDA를 위한 컬럼 생성 내용 (추가 혹은 삭제?)
날짜의 월 기준으로 3-5월 : 봄, 6-8월 : 여름, 9-11월 : 가을, 1,2,12월 : 겨울로 구분해준 후 계절 컬럼을 생성해 주었다.