**Imputation & Outlier Process**

**결측치 처리(Imputation)**

**Theory**

결측치란 입력이 누락된 값을 의미하며, NA로 출력됩니다. 그런데 이 결측치가 데이터에 포함되어 있을 경우 산술 연산을 수행할 수 없고, 분석 결과가 왜곡될 수 있습니다. 따라서 본격적인 EDA를 수행하기 전에 데이터에 결측치가 존재하는지를 확인하고 처리하는 절차가 필수적으로 필요합니다.

**Imputation Strategy**

단순 대치법 (Single Imputation): 단일 값으로 대치 or 결측치가 존재하는 행을 삭제

* Complete analysis: 결측값이 존재하는 행을 삭제
* 평균 대치법(Average Imputation): 관측 데이터의 평균으로 대치
* 단순확률 대치법(Simple Probability Imputation): Hot-deck 방법, nearest neighbor 방법

선형 보간법 (Linear Interpolation): 시계열 데이터에서 주로 쓰이는 방법으로 선형 보간 수행.

**이상치 처리(Outlier Process)**

**Theory**

이상치(Outlier)란 관측된 데이터의 범위에서 많이 벗어나 있는 아주 작거나 아주 큰 값으로서 정상범위 밖에 있는 데이터를 뜻합니다. 이러한 이상치는 잘못 입력된 값일 수도 있으나 실제로 존재하는 값일 수도 있으며, 전체 데이터 분포에 큰 영향을 미칩니다.

이상치가 존재할 경우 분석 결과가 특정한 쪽으로 치우치거나 왜곡될 수 있습니다. 따라서 데이터에서 이상치를 인식한 후, 이를 제외하고 분석을 할지 혹은 포함하거나 조정하여 분석을 할지 판단한 후 분석을 수행해야합니다.

**IQR Module**

‘IQR Module’은 사분위수를 활용한 대표적인 이상치 처리 방법입니다. 사분위수(Quaantiles)란 데이터 표본을 가장 작은 값부터 가장 큰 값으로 정렬한 후, 4등분 했을 때 각 등위에 해당하는 값을 의미하며, 사분위 범위(IQR: Interquantile Range)는 제1사분위수와 제3사분위수 간의 거리로, 데이터의 중앙에 있는 50%구간의 범위에 해당됩니다.

‘IQR Module’은 숫자들 중 특정 값이 [중앙값-1.5\*IQR ~ 중앙값+1.5\*IQR] 범위에서 벗어나 있는 경우 해당 데이터를 이상치로 간주하는 이상치 처리 방법입니다.