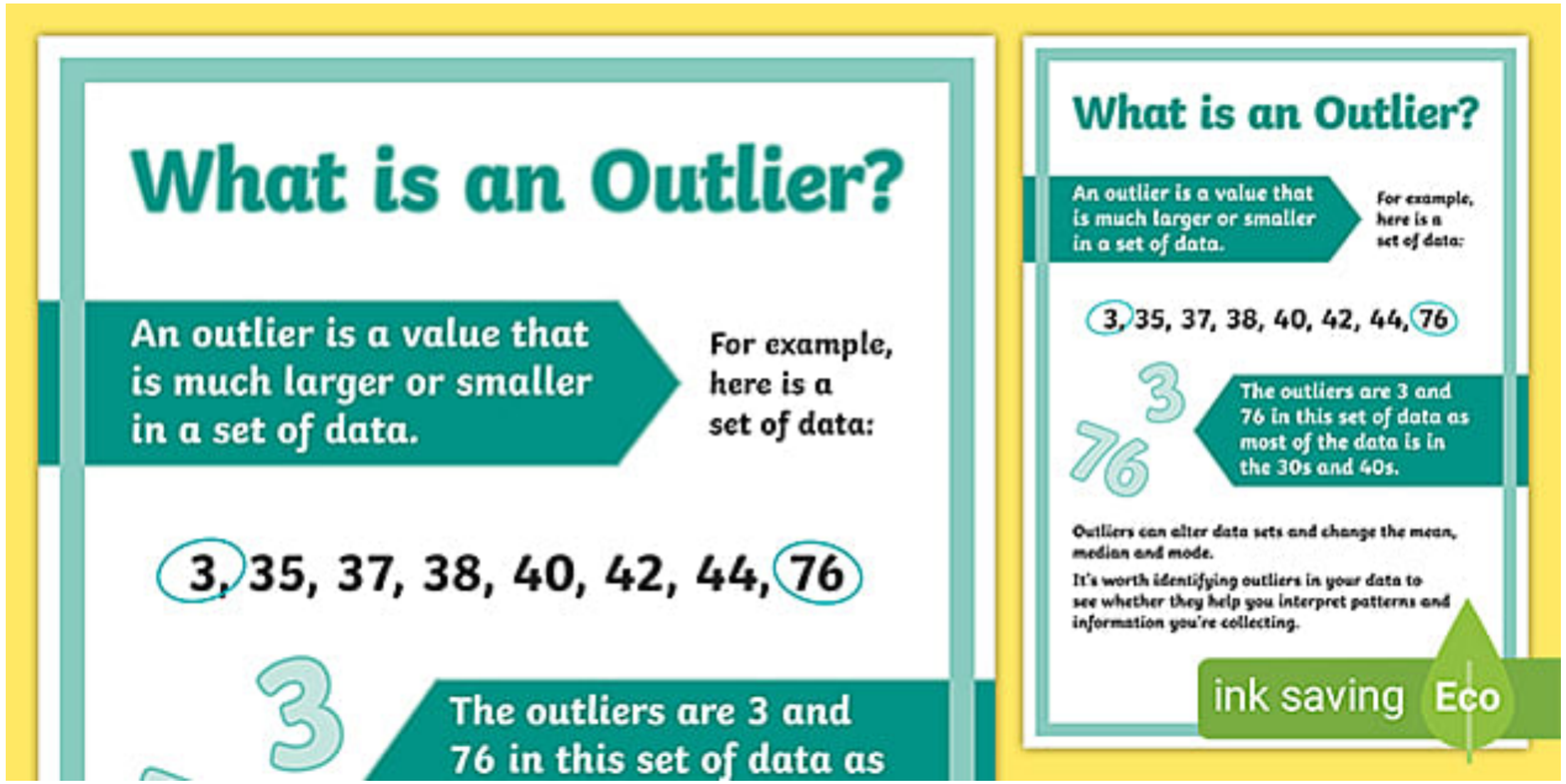
이상치 (Outlier)

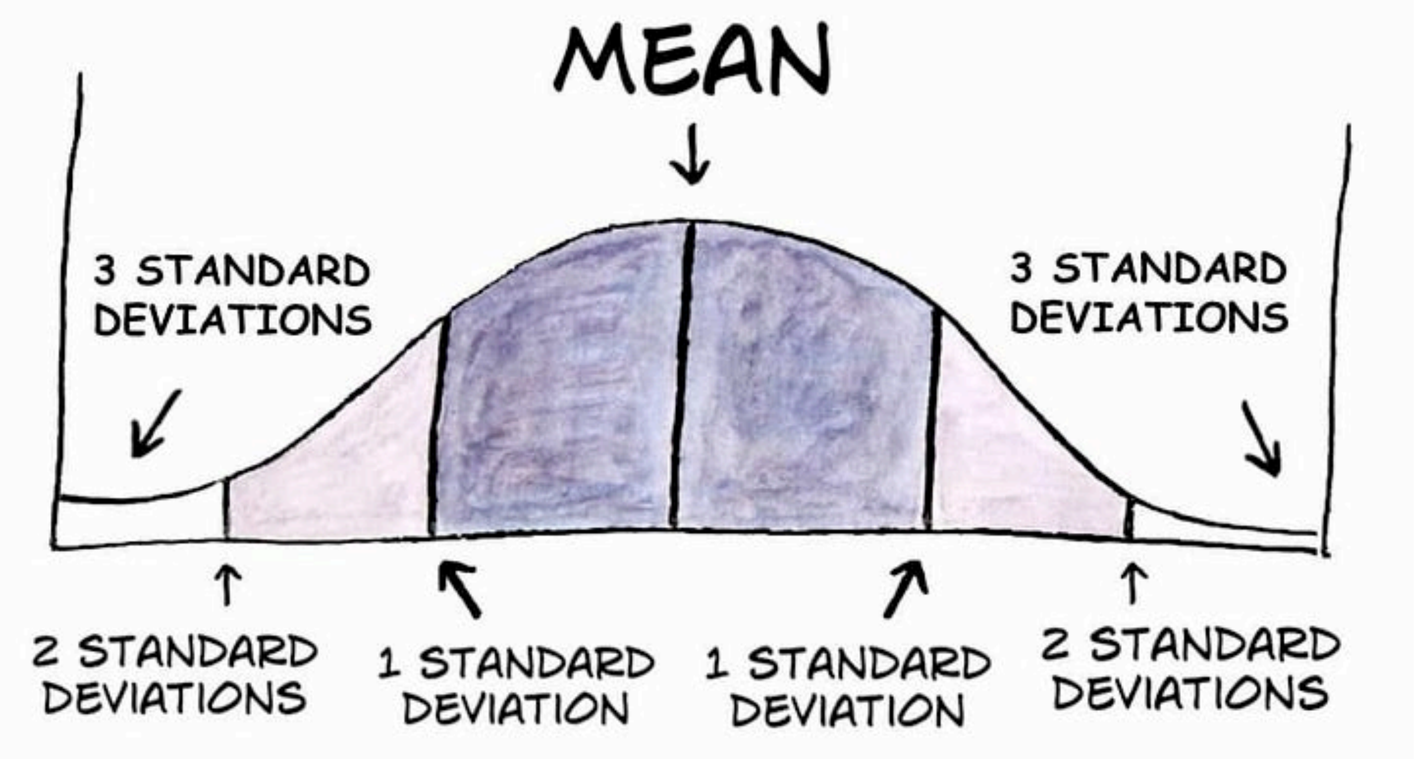
데이터 분석 전처리 (판다스)

이상치가 발생하는 이유?

1. 데이터 입력 오류(Data Entry Errors)
2. 측정오류(Measurement Errors)
3. 샘플링 오류(Sampling Errors)
4. 자연 발생 이상치(Natural Outliers)
5. 데이터 처리 오류(Data Processing Errors)

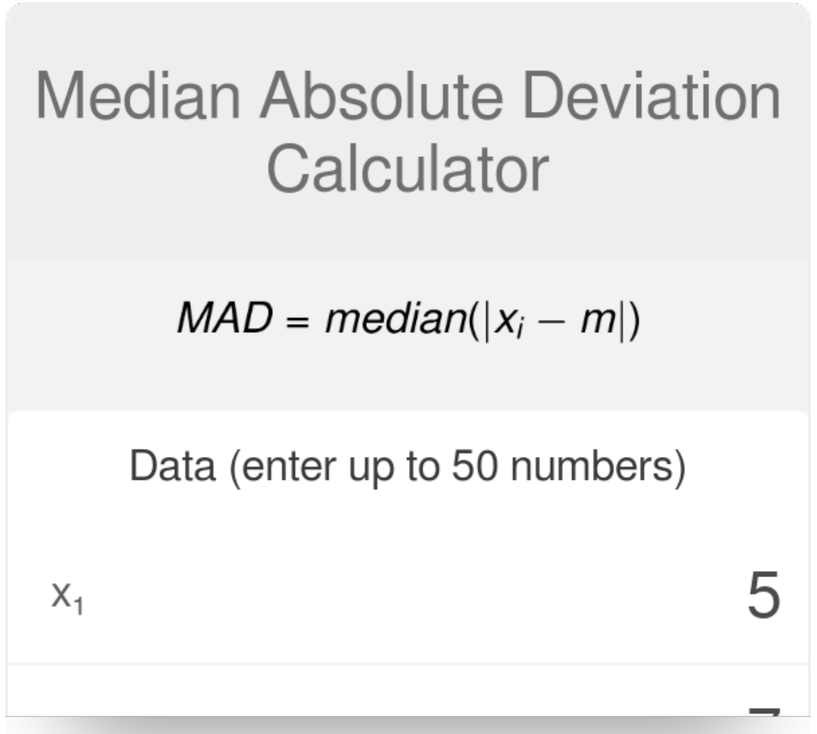


Outlier detection

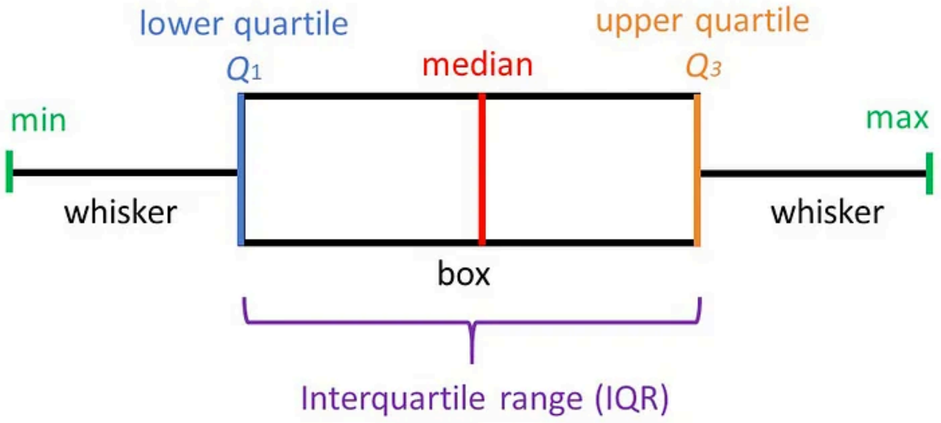


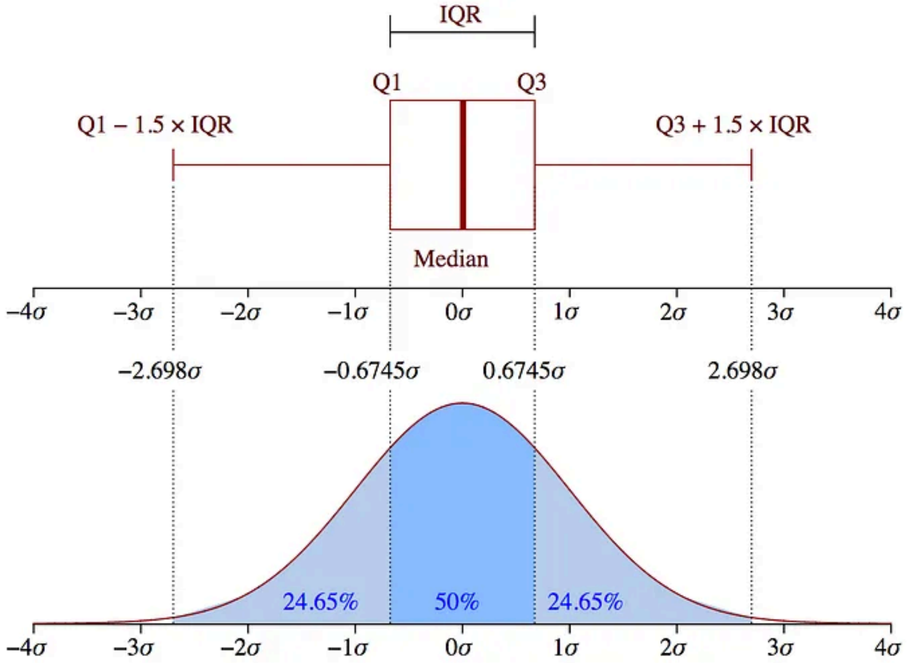
MAD

$$MAD = median(|x_i - m|)$$

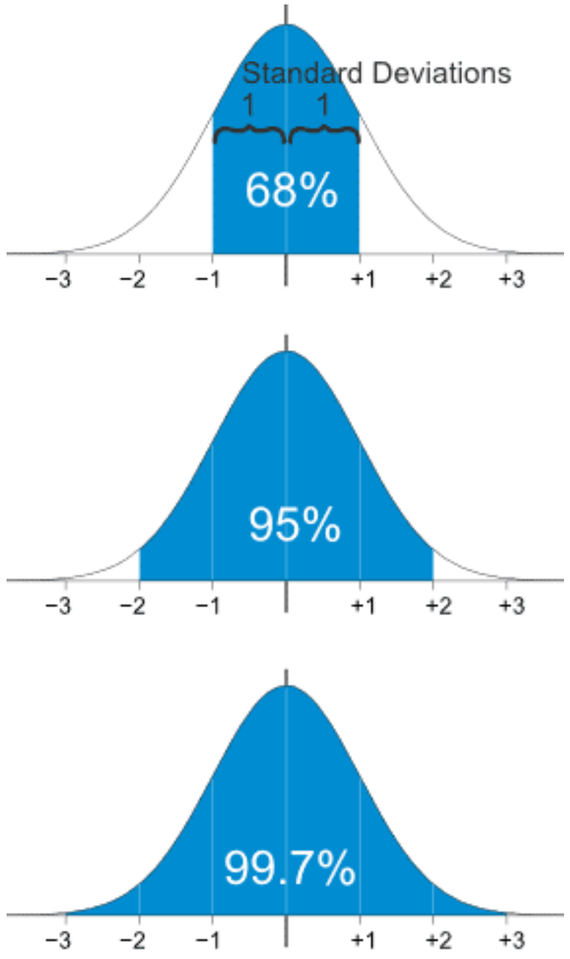


IQR

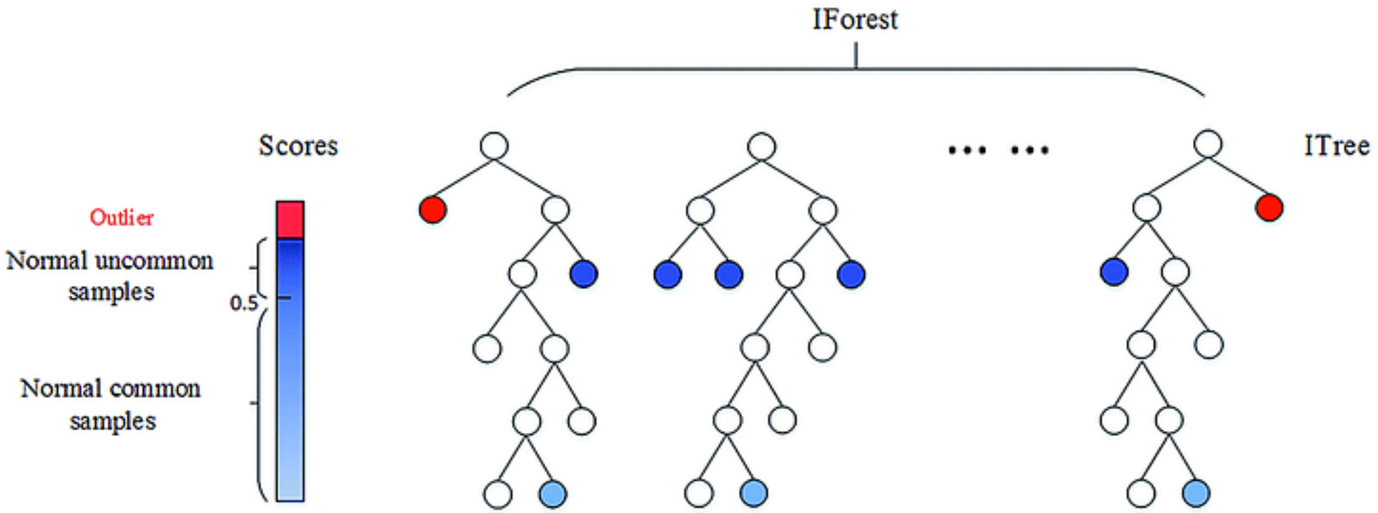

lower quartile
Q_1
median
upper quartile
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%

Standard Deviation

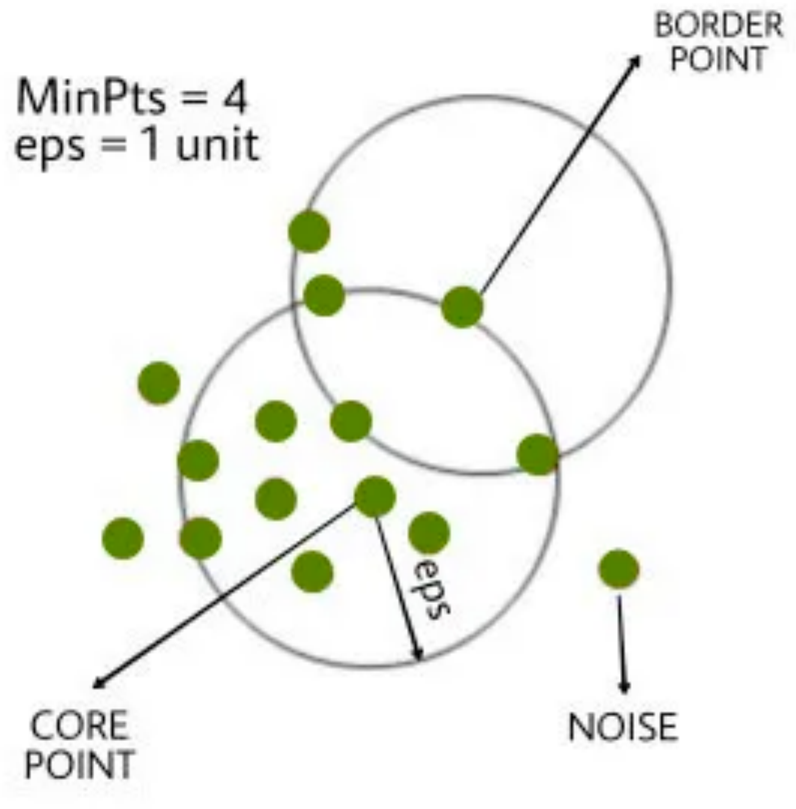


Isolation Forest



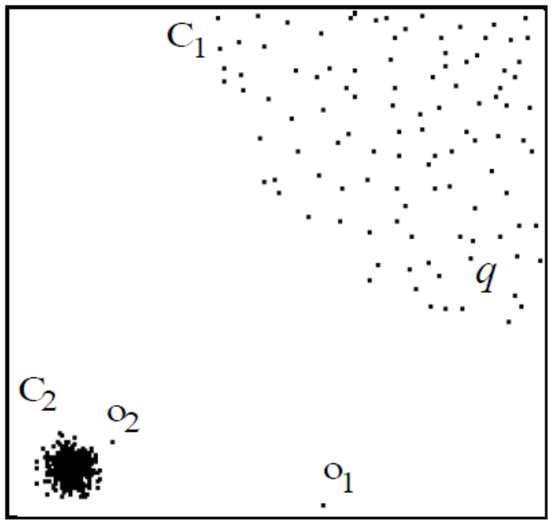
Isolation Forest는 데이터 포인트를 '고립'시키는 데 필요한 분할의 수를 기반으로 이상치를 식별
정상적인 관측치보다 이상치가 고립되는 데 적은 분할이 필요

DBSCAN(Density Based Spatial Clustering of Application with noise)

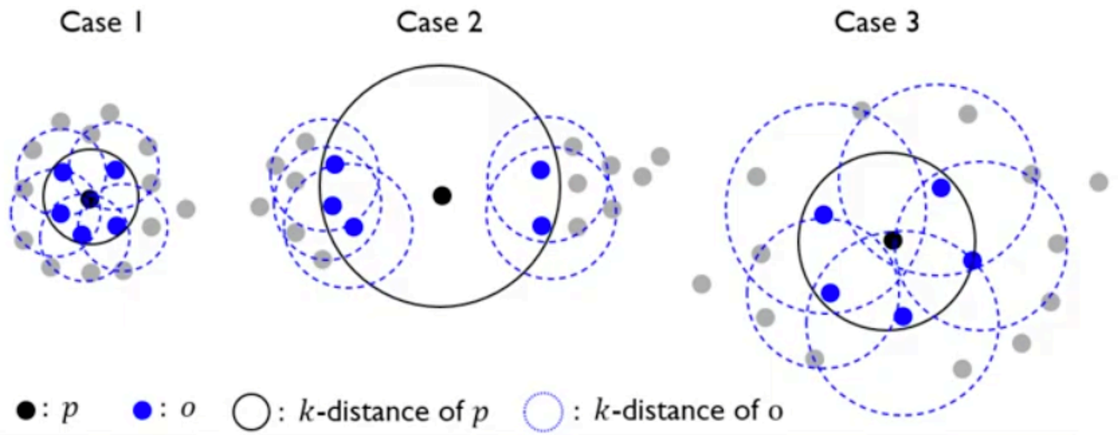


밀도 기반 클러스터링 알고리즘으로, 밀도가 높은 지역의 점들을 클러스터로 그룹화하고, 밀도가 낮은 지역의 점들을 이상치로 간주

LOF(Local Outlier Factor)



$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} = \frac{\frac{1}{lrd_k(p)} \sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|}$$



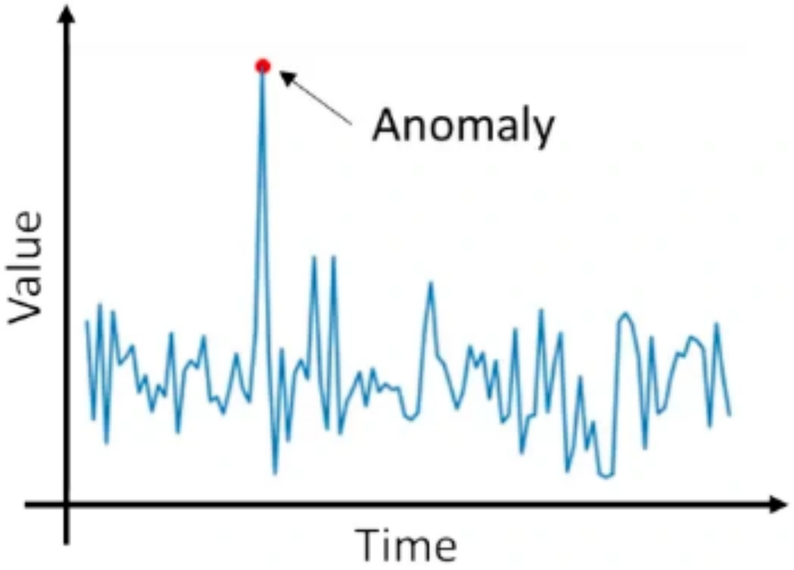
Case	$lrd_k(p)$	$lrd_k(o)$	$LOF_k(p)$
Case 1	Large	Large	Small
Case 2	Small	Large	Large
Case 3	Small	Small	Small

13

밀도기반으로 이상값 찾는 알고리즘, 다차원 공간 안에서 이상값들은 주위 밀도에 비해 밀도가 상당히 낮다는 점을 통해 시작

슈하르트 (시계열 데이터 이상치 탐지)

Shewart Control Charts



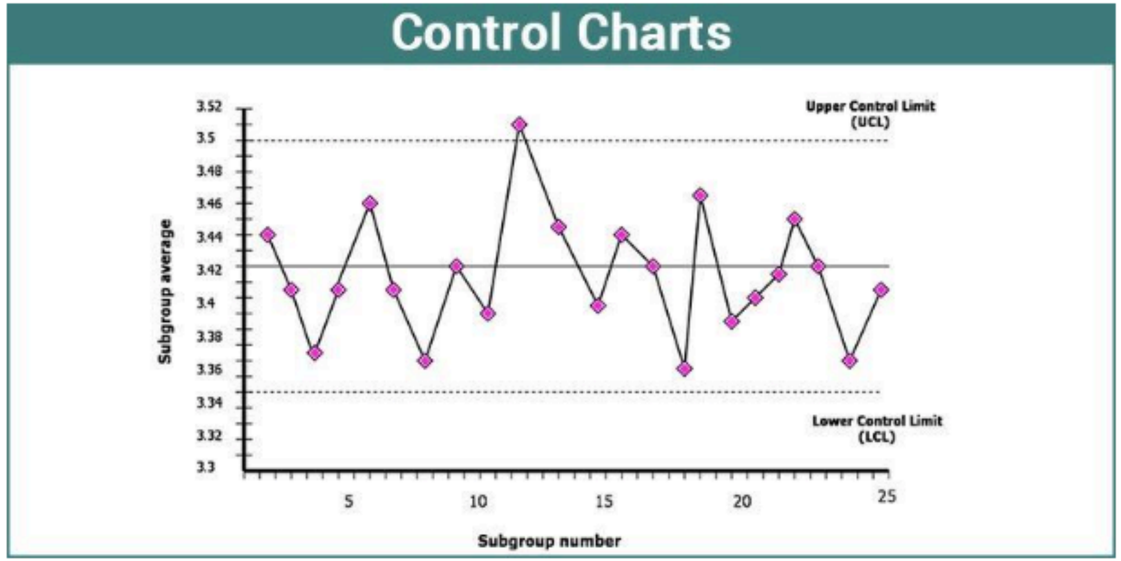


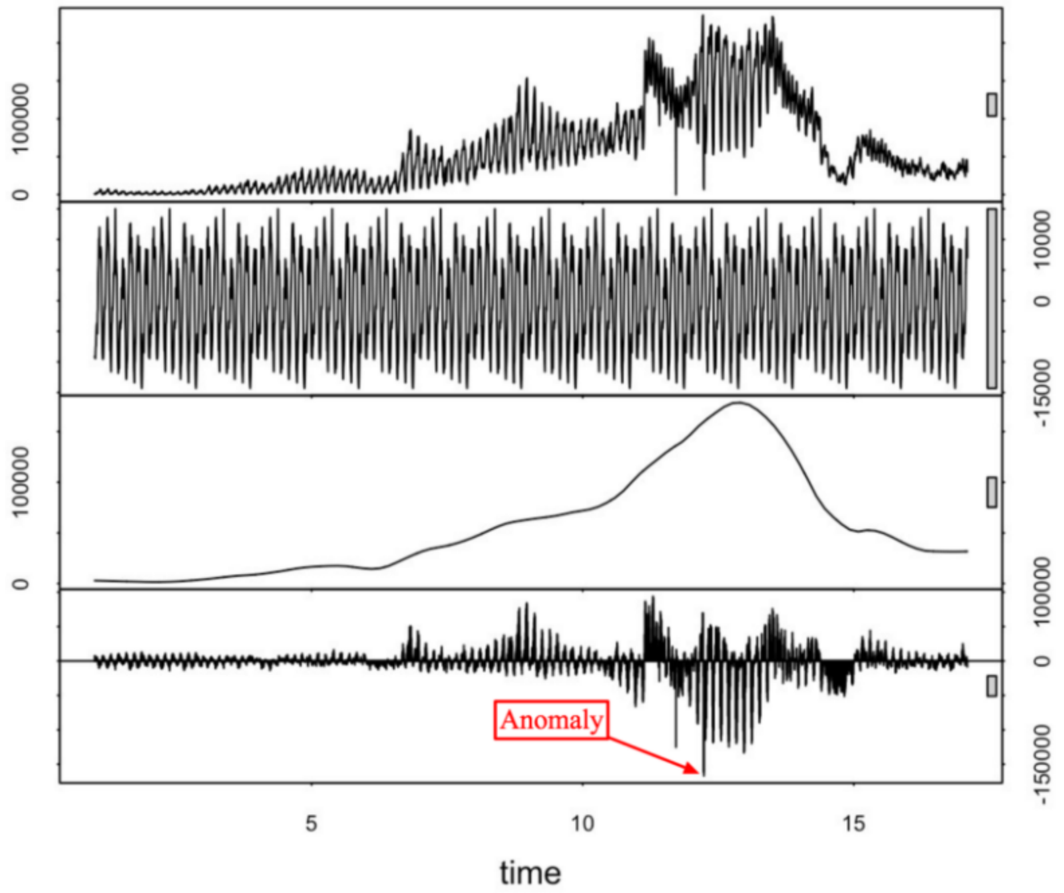
그림1 Control Chart

통계 기반 이상치 탐지 방법론

STL Decomposition

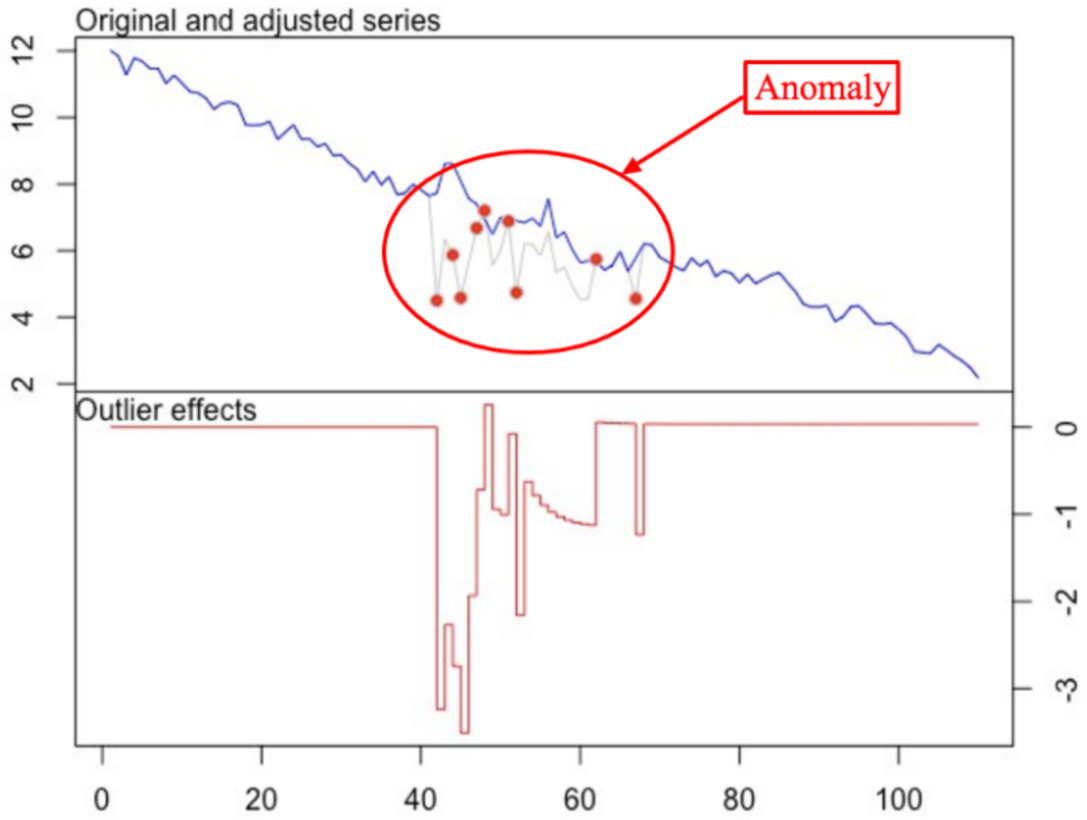
순서대로 original time series, seasonal, trend, residue

맨 위의 origianl time series를 아래의 3 개의 그래프로 분해



From top to bottom: original time series, seasonal, trend and residue parts retrieved using STL decomposition.

ARIMA



Two time series built using original ARIMA model and adjusted for outliers ARIMA model.

데이터 분석가의 도메인 지식을 통한 Outlier 선정