진행과정

1. 데이터 전처리 순서

- missing value 제거 후 outlier 처리함

- null data나 문자열 데이터가 있는 경우 숫자를 기준으로 처리하는 outlier 적용이 어려움

- 만약 null data를 평균값 등의 보간법을 적용하면 자칫 데이터의 신뢰도가 떨어질 수 있어 missing value를 먼저 처리함

2. missing value 처리 방법

- missing value (null)가 있는 row는 모두 제거함

- 보간법을 적용하여 숫자를 채워 넣을 수 있으나 자칫 오류를 발생할 수 있기 때문에 missing value가 있는 row는 모두 제거함

3. missing value 처리 과정

- 데이터 information (df.info() method 사용)

- 각 컬럼의 non-null row의 개수와 date type을 보여줌

<class 'pandas.core.frame.DataFrame'>

Int64Index: 200000 entries, 0 to 199999

Data columns (total 19 columns):

viewCount 165369 non-null float64

editCount 165369 non-null float64

shareCount 165369 non-null float64

searchCount 165369 non-null float64

coworkCount 165369 non-null float64

add 63166 non-null float64

del 63166 non-null float64

move 63166 non-null float64

rename 63166 non-null float64

adddir 63166 non-null float64

movedir 63166 non-null float64

visdays 184305 non-null float64

openCount 149090 non-null float64

saveCount 149090 non-null float64

exportCount 149090 non-null float64

viewTraffic 149090 non-null float64

editTraffic 149090 non-null float64

exportTraffic 149090 non-null float64

traffic 149090 non-null float64

dtypes: float64(19)

memory usage: 30.5 MB

- dropna(how=’any’) method를 적용하여 missing value가 하나라도 있는 row는 모두 제거함

- missing value 제거 후 데이터: 54,722개의 row만 남음

<class 'pandas.core.frame.DataFrame'>

Int64Index: 54722 entries, 1 to 199998

Data columns (total 19 columns):

viewCount 54722 non-null float64

editCount 54722 non-null float64

shareCount 54722 non-null float64

searchCount 54722 non-null float64

coworkCount 54722 non-null float64

add 54722 non-null float64

del 54722 non-null float64

move 54722 non-null float64

rename 54722 non-null float64

adddir 54722 non-null float64

movedir 54722 non-null float64

visdays 54722 non-null float64

openCount 54722 non-null float64

saveCount 54722 non-null float64

exportCount 54722 non-null float64

viewTraffic 54722 non-null float64

editTraffic 54722 non-null float64

exportTraffic 54722 non-null float64

traffic 54722 non-null float64

dtypes: float64(19)

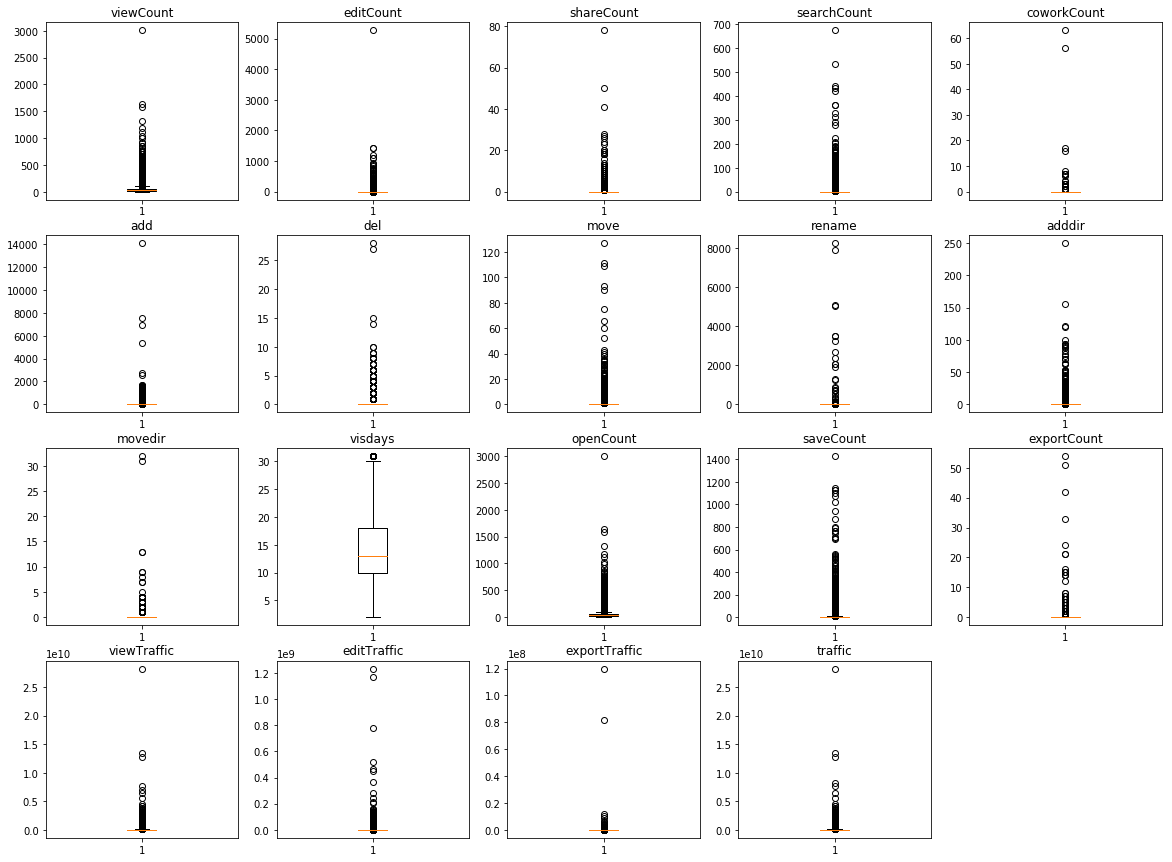
memory usage: 8.3 MB

4. Outlier 처리

- Outlier 처리 전에 string type column인 'iduser', 'mdutype', 'group' 제거

1) Outlier 처리 전 boxplot

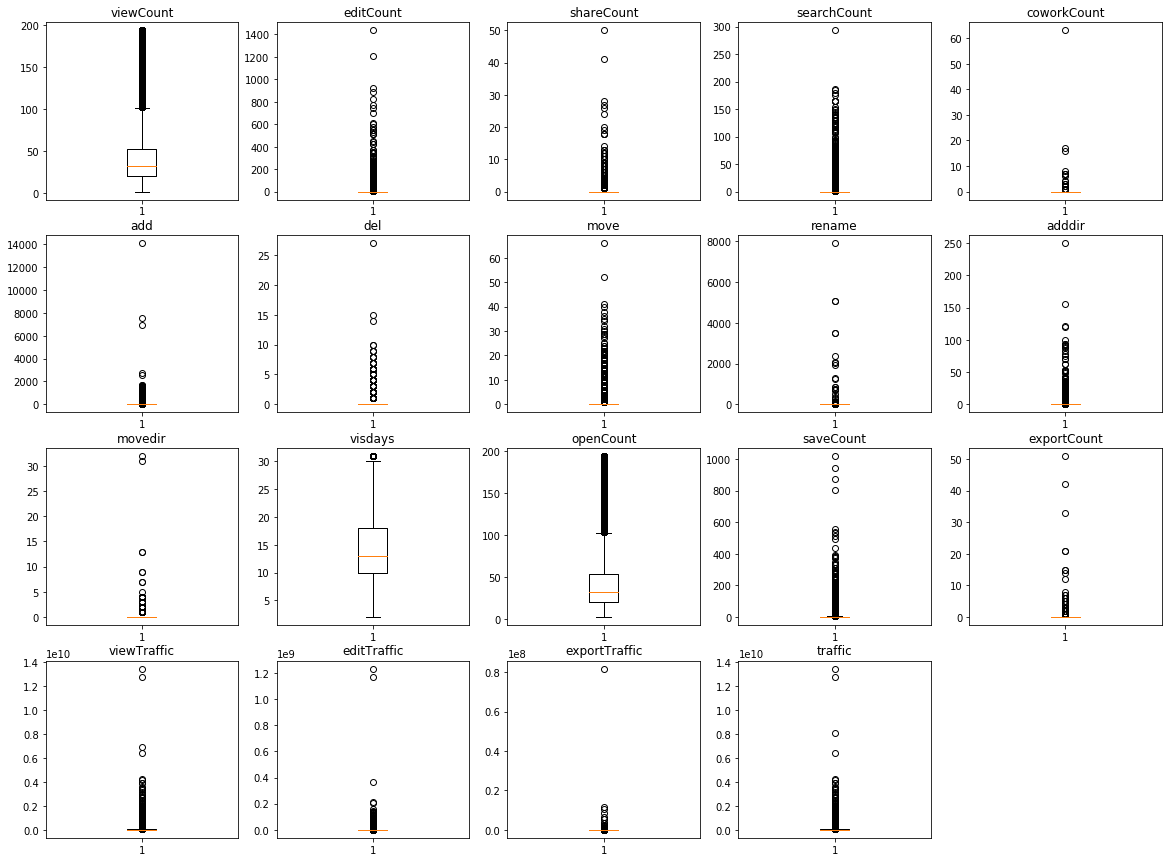
- boxplot에서 ‘visdays’ 컬럼을 제외한 전 컬럼에서 매우 많은 outlier 확인



2) 표준점수방식

- 표준점수(μ=0, σ=1)로 변환 후, ±3σ 넘는 값 제거

- outlier 제거 전보다 data 분포가 줄긴 했지만 여전히 높은 분산을 보이고 있음



3) IQR 방식

- 1Q-1.5IQR 이하와 3Q+1.5IQR 값을 제거하는 방법

- 위 기준의 모든 outlier가 제거된 boxplot을 확인할 수 있음

