Data EDA

* **Outlier(Kills에 대해서, 나머지 수치형 변수에도 적용 필요)**

1. LOF(local factor) – 밀도 기반으로 다차원에서 outlier detecting에 사용 – outlier의 결과 제대로 설명하지 못함
2. ISF(Isolation forest) – 랜덤포레스트 기반으로 다차원에서 outlier detecting에 사용 – outlier의 결과 제대로 설명하지 못함
3. Boxplot 결과 2.5킬 이상이 outlier – 충분히 가능한 킬 수로 판단

Z-score : 분포가 정규분포를 따른다는 가정하에 실행.

1. Z-score > 3 => 6킬 : 1순위
2. Z-score > 2 => 4킬 : 2순위
3. IQR > 3rd + 1.5\*IQR => 3킬 표본의 0.01 : 3순위
4. 0킬 제외한 IQR : 7킬 이상 0.001percent

IQR 방식은 75% percentile \* 1.5 이상이거나 25 percentile\* 1.5 이하인 경우 극단치로 처리하는 방식이다. 이해하기 쉽고 적용하기 쉬운 편

IQR이나 MAD와 같이 Median(중앙치)을 고려한 방법은, 극단적으로 0에 편중된 분포를 보일 경우 0에서 조금만 수치가 벗어나도 이상치로 고려되기 때문에 0을 제외하고 IQR 구해서 처리해야 함.

위 b,c,d를 고려하여 분석에서 MAE를 최소화 하는 값으로 아웃라이어 제거

* **변수제거(컬럼제거)**

1. Covariance
2. 도메인 인사이트를 통한 사전 제거
3. Pairplot
4. PCA(차원 감소, 다중공선성 문제 해결) => PCA 성분을 이용한 데이터 분석 필요.

PCA 및 요인분석 모두 사전 Scaling은 필수입니다.

* **데이터 단위 변환(단위가 다른 변수 스케일 작도록)**

1. Normalization - 평균이 0, 분산이 1인 분포로 변환
2. Minmaxscaler - 특정 범위 (예, 0~1)로 모든 데이터를 변환
3. Box-Cox: 여러 k 값중 가장 작은 SSE 선택
4. Robust\_scale: median, interquartile range 사용(outlier 영향 최소화)

* **결측치 처리 – 결측치가 1건 존재, 넘어감**
* **데이터 변환**

1. PUBG의 독립, 종속변수의 경우 정규분포가 아님(특히 왼쪽에 치우친 변수들이 많음) => 선형 모델의 경우 log변환을 통해 정규분포 형태로 변환 후 학습시킨다.

통계 분석 방법을 기반으로 한 모델의 경우, 대부분의 통계 분석 모델은 정규성 가정을 기반으로 하기 때문에 데이터 변환을 통해 정규분포를 따르도록 만드는 것이 좋음.

대부분의 모델은 변수가 특정 분포를 따른다는 가정을 기반으로 한다. 예를 들어 선형 모델의 경우, 설명 및 종속변수 모두가 정규분포와 유사할 경우 성능이 높아지는 것으로 알려져 있다. 자주 쓰이는 방법은 Log, Exp, Sqrt 등 함수를 이용해 데이터 분포를 변환하는 것이다. -

* left\_distribution: X^3
* mild\_left: X^2
* mild\_right: sqrt(X)
* right: ln(X)
* servere right: 1/X

1. headshotkills => headshotkills rate 컬럼으로

처리!!!

**변수제거**

groupId,matchId,killplace,(killpoint, rankpoint, winpoint => Weighted),damagedealt,matchduration, numgroups, roadkills, teamkill,vehicle destroys, maxplace

**이상치 제거(솔로에 대해서만, duo, squad는 min,max값 보고)**

matchType은 이상치 없음

'assists\_mean', 'boosts\_mean', 'headshotKills\_mean', 'heals\_mean', 'kills\_mean', 'killStreaks\_mean','longestKill\_mean', 'rideDistance\_mean', ’swimDistance\_mean', 'walkDistance\_mean','weaponsAcquired\_mean' => 상위 0.005퍼센트 이상치 제거

모든 데이터의 경우 walkingdistance==0이고 winplaceperc > 0.5 은 제거(핵쟁이일 가능성)

**데이터 단위 변환**

Walkingdistance, swimdistance, ridingdistance, longestkill – log 변환

**데이터 변환**

Headshotkills => headshotkillsrate = headshotkill/kill

Team game이기 때문에(팀은 같은 등수를 가짐) 팀의 스탯으로 분석.

Kill, assist, boost, DBNOs, headshot, heal, killstreaks, revive, Weaponsacquired, Walkingdistance, swimdistance, ridingdistance, longestkill => groupby([‘matchid’,’groupid’]) 같다면 mean값 사용

매치아이디 사이즈

matchId별 평균 변수

solo의 경우 DBNOs, revives 컬럼 제거

duo인데 다른 아이디 4개, 같은 그룹아이디 1개, 같은 매치 아이디 1개, 승률마저 똑같음..?

팅겼다가 다시 들어온다?, Id 새로 부여? => 평균값으로 해야함..(로그데이터)

assists\_mean/assists\_match\_mean 으로 비율형태로 만들기.