Data Mining Project : 음악 생존력

12181945 통계학과 전준호

# Problem Forulation

예전에, web craling practicing을 하면서 음악 사이트를 web crawling한 적이 있다.

그때, 재밌는 점이, 굉장히 옛날 노래들이 in200을 유지하고 있는 경우가 많이 보였다는 점이다.

이번 주제는 permanent music : 이 음악이 오래 살아남을 것인가 에 대해 categorical model을 세워보고자 한다.

# Data Preparing

본인이 Genie music을 애용하기도 했고, crawling의 편의성을 이유로 genie를 선택했다.

(genie 사진)

## title artist ranking  
## 1 취중고백 김민석 1  
## 2 겨울잠 아이유 (IU) 2  
## 3 회전목마 (Feat. Zion.T & 원슈타인) (Prod. by Slom) sokodomo 3  
## 4 리무진 (Feat. MINO) (Prod. by GRAY) BE'O (비오) 4  
## rankingdiff date  
## 1 6상승 2022-01-10  
## 2 1상승 2022-01-10  
## 3 2하강 2022-01-10  
## 4 유지 2022-01-10

genie에서는 공식적으로 최근부터 2022년 1월 첫째주까지 in200 순위를 제공하고 있다. 사이트 주소를 임의로 변경해 이전 기록에도 접근 가능하지만, 일단 공식적으로 제공하는 정보만 이용하기로 결정하였다.

2022년 1월 첫째주부터 2024년 4월 첫째주까지 2년이 좀 넘는 dataset을 확보했다.

음원 순위의 기준은 주간으로 정했다. 일간은 dataset이 과도하게 많아질 것 같고, 하루 잠깐 올라왔다 떨어지는 음악이 많아 자료가 unstable하다고 생각했다. 또한 주간 순위라면 충분히 대표성이 있을 것이라고 생각했다.

variables는 title, artist, ranking, rankingdiff, date이다.

타겟 변수 Y는 isPermanent{0, 1}으로, “all-time date동안 in200을 만족하는가”가 조건이다.

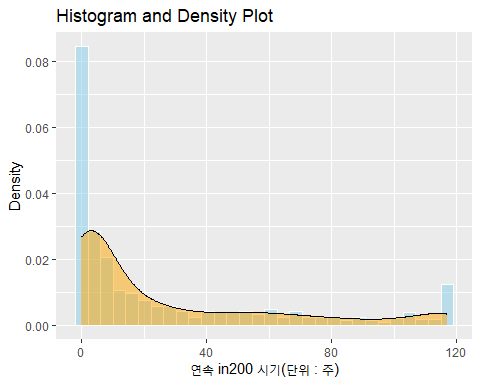
EDA를 해보고 데이터(X)가 추가될수도 있고, 타겟변수의 디테일한 조건이 변경될 수도 있지만 일단은 이러하다.

# EDA

### in 200 유지 기간 히스토그램

먼저, 관심있는 부분은 “얼마나 오래 순위권을 유지하는가”이다. 따라서, 전처리를 거쳐 음악별로 in200 유지 기간(이하 생존기간)이 얼마나 되는지 확인하고자 한다.

## [1] 966



대부분의 음악들이 3주 이상을 버티지 못하고 in 200에서 벗어나는 모습을 관측할 수 있다.

또한, 전체 조회 기간동안 in200에서 생존하는 음악이 group화되어 있는 모습을 확인할 수 있다.

생존 주간을 빈도표로 나타내면 다음과 같다.

## # A tibble: 113 × 2  
## `in200 기간` count  
## <fct> <dbl>  
## 1 0 177  
## 2 1 100  
## 3 2 53  
## 4 117 44  
## 5 4 33  
## 6 7 30  
## 7 5 28  
## 8 3 27  
## 9 6 23  
## 10 10 18  
## # ℹ 103 more rows

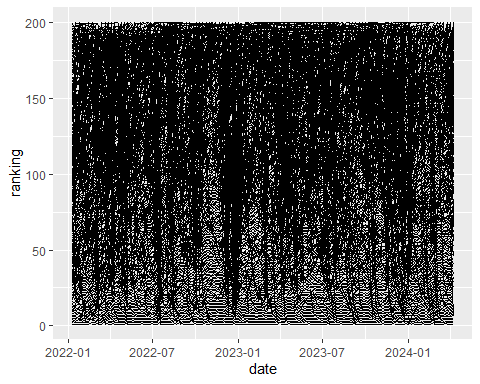
대부분이 3주 이상 in200에서 버티지 못하는 것이 수치적으로도 보인다. 전체 기간에서 in200을 유지하는 곡이 특이하게 많이 보인다.

### 순위 ~ 기간 plot

전체 기간에서 순위변동의 흐름이 존재하는지 파악하기 위해 순위 ~ 기간 plot을 그려보았다.

변동이 심하다면 세로선으로 왔다갔다 하는 경향이 강할 것이고, 변동이 적다면 가로로 평행하게 보이는 경향이 있을 것이다.

## [1] 966



음악이 매우 많기 때문에(966개) 그래프가 매우 난잡하게 나타나는 모습을 확인할 수 있다. 하지만 그럼에도 불구하고 자세히 보면 약간의 경향성을 확인할 수 있다.

먼저, 세로선과 가로선이 보인다. 높은 순위권의 음악이 낮은 순위를 거쳐오기 때문으로 보인다. 고 순위권(1~25)에서는 세로선의 비중이 비교적 적고 가로선의 비중이 큰 것을 확인할 수 있다.

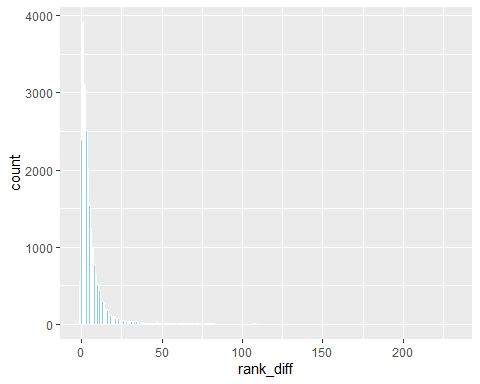
이것으로 고 순위권 생태계가 상대적으로 안정된 기간을 유지하는 경향이 있다라고 추측할 수 있다.

또한, 연말 연초에 해당하는 기간에 유난히 음원 변동이 큰 것으로 보인다. 이는 show me the money같은 유명 음악 프로의 영향일 수도 있다. 혹은 겨울 계절음악이 타 계절에 비해 유행을 잘 타기 때문이라고 생각할 수 도 있다.

“세로선과 가로선”은 음원순위 변동성에 관한 이야기이다. 이를 좀 더 자세히 살펴보기로 했다.

### 음원 순위 변동성 plot

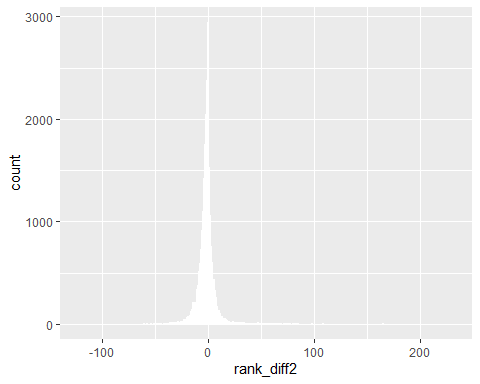
음원랭킹 변동수치를 가지고 histogram을 그려보았다.



## # A tibble: 172 × 2  
## rank\_diff freq  
## <dbl> <int>  
## 1 1 3915  
## 2 2 3109  
## 3 3 2512  
## 4 0 2395  
## 5 4 2022  
## 6 5 1541  
## 7 6 1246  
## 8 7 992  
## 9 8 772  
## 10 9 651  
## # ℹ 162 more rows

굉장히 right skewed된 plot이 보인다.

plot으로 보면 25 이하 변화에서 거의 노는 모습을 볼 수 있다. 최초 등록시 급상승하는데, 이 “급상승”은 정해진 수치가 없고 range가 넓으므로 넓게 퍼지는 것으로 생각된다. 이후 소량변동의 경우 range가 좁기 때문에 extremely skewed로 나타나는 것 같다.

변화량은 절댓값이므로, 이번엔 증감을 적용해서 plot을 그려보았다. 

크게 다르진 않아보인다.

증가의 경우 200+도 있는데, 이는 genie에서 200위권 밖의 음악이 200위권 이내로 진입했을 때 랭킹 변동 200+의 수치도 적용해주기 때문으로 보인다. 랭킹 감소의 경우, -200 이하는 web crawling에 적용되지 않기 때문에 보이지 않는 모습이다.

지금까지 내용을 통해, 예측변수 Y를 in200을 조회 기간(117주)동안 유지하는가 하지 못하는가에 초점을 둘 수도 있다.

하지만 Y를 117주 동안 in200으로 두면 문제가 생긴다. 바로 조회 기간동안 in200을 처음 기록한 음악들이다. 이런 음악들은 실제로 앞으로 계속 in200을 기록하게 되어도 permanent music에 속하지 못한다.

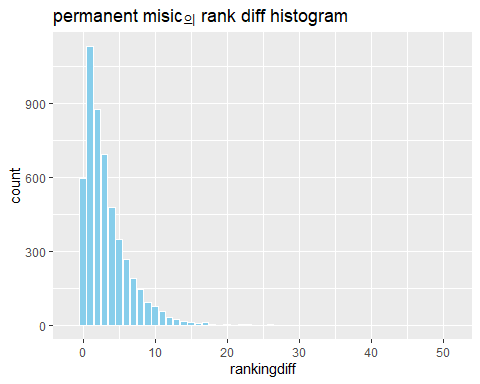
따라서, 일차적으로 “2024년 4월 첫째주까지 생존한 음악”을 permanent music으로 두겠다.

### Permanent music

가장 먼저 드는 생각은, “permanent music은 안정적으로 그들의 순위를 유지할 것이다”이다. 실제로 그럴지 확인해보자.

전처리를 통해 permanent music을 구하고, 그들의 rank difference를 histogram으로 그려보았다.

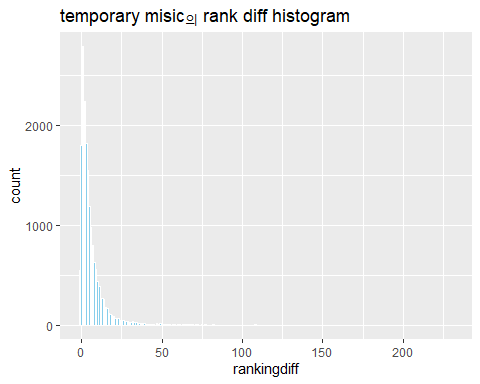
## [1] "에잇 (Prod. & Feat. SUGA of BTS)" "OHAYO MY NIGHT"   
## [3] "내 손을 잡아"



여전히 right skewed이다. 즉, 안정적으로 그들의 순위를 유지한다. 실제로 10 이상을 벗어나는 값이 거의 없는 모습을 확인할 수 있다.

### Temporary music

permanent가 아닌 음악을 temporary music이라고 표현하겠다. permanent music의 rank 변동성을 확인했으니, 그렇지 않은 음악의 rank 변동성도 확인해보았다.



이들도 right skewed로 안정성이 있다고 이야기할 수 있지만, permanent보다 훨씬 큰 값에 data가 존재함을 알 수 있다.

수치통계량도 확인해보았다.

## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## -1 0 1 1 2 2 3 4 5 8 51

## [1] 0.9611384

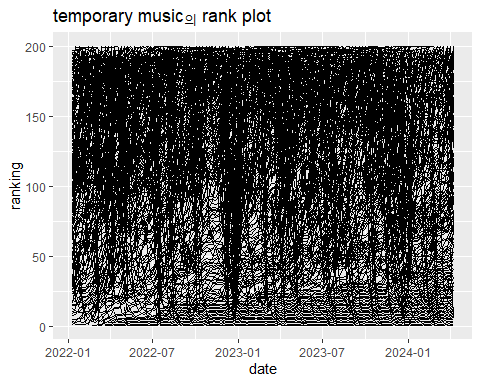
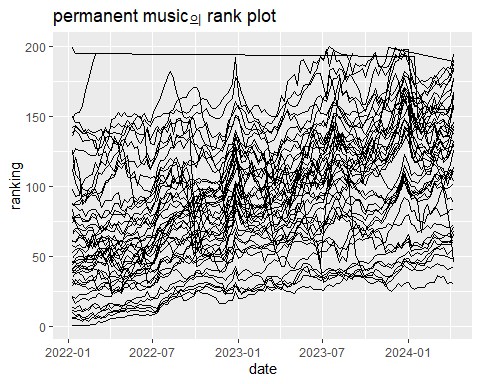
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## -1 0 1 2 3 4 5 7 9 15 231

## [1] 0.8281546

변화량 10 이내의 비율이 permanent는 약약 96%이고, temporary는 약 83%이다. 두 차이가 크다면 크고, 작다면 작다고 이야기 할 수 있어보인다.

### Permanent vs. Temporary

permanent의 특성과 temporary의 특성을 확인해보기 위해, 이 둘의 ranking 자체변동 역시 비교해보자.



permanent music은 확실히 횡보하는 경향이 나타난다. temporary music은 처음에 봤던 그 복잡한 경향이 나타난다.

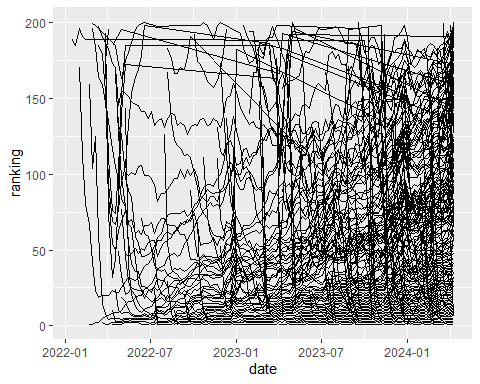
### Candidate music

EDA하면서 든 생각은, permanent music의 정의이다.

지금까지 분석은 permanent music을 “117주 동안 in200에 차트인한 음악”으로 정의했다.

하지만 이러면 중간에 차트인 하고 앞으로 117주 이상 차트인할 수 있는, 다시 말해 “유망한”음악을 temporary로 분류하게 된다.

따라서, 이러한 음악들을 candidate, 즉 “유망한” 음악으로 정의하고 다시 한번 plot들을 그려보았다.



permanent와 temporary의 속성이 모두 들어가 있는 모습이 보인다.

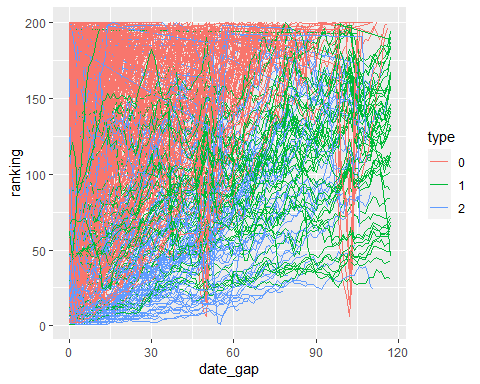
세로선의 경향은 temporary의 속성이며, 최신 시간에 가까우질수록 횡보하는 경향이 보이는데 이것은 permanent의 속성이다.

### 동일한 시작기간에 대한 plot

노래들은 in200한 기간이 모두 다르다. 심지어, in200을 시작한 기간이 모두 다르다. 이때문에 ranking ~ date plot에서 복잡한 모습이 보였다.

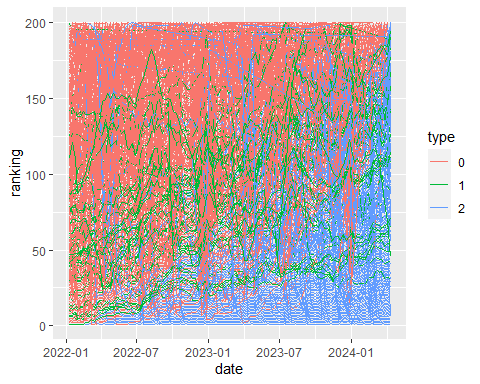
우리가 주목하는 것은 “음악의 생존기간”이기 때문에, 시작점을 모두 일치시켜서 다시 한번 plot을 그려보았다.

또한 이제 세 그룹에 대한 분류를 마쳤으니 한번의 plot에 확인해보고자 한다.



1번(permanent, green)은 대체적으로 횡보한다. 크게 변화하는 구간이 적다. 2번(candidate, blue)은 ranking start point가 굉장히 높은 경향이 있다. 0번(temporary, orange)은 date gap이 굉장히 짧은곳에 몰려있고, 중간에 역주행(발매 이후 시간이 지났는데 갑자기 순위가 급증하는 것) 하는 곡도 보인다. 거의 끝까지 왔다가 permanent에 아쉽게 속하지 못한 곡도 보인다.

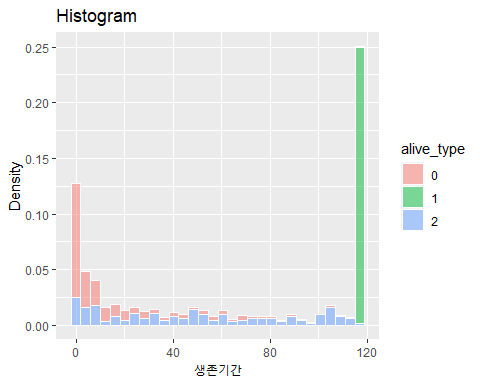
세가지 그룹으로 분류하니, 이전의 복잡한 ranking ~ date plot도 분류할 수 있을 것 같다.



멋있는 그림이 나왔다. 꽤나 명백하게 group이 나뉜 모습을 볼 수 있다. 좌측 상단은 temporary music이, 우측 하단은 candidate music이, 그리고 가운데 횡보하는 permanent music이 보인다.

이 plot을 통해 “음악의 생존기간에 따른 분류”를 할 수 있을 것 같다는 생각이 든다.

최초에 음악의 생존기간에 따른 histogram도 group별로 나누어서 확인해보고자 한다.



histogram으로 본 생존기간 분포 역시 grouping이 가능할 것 처럼 보인다.

# 목표 수립

추가 EDA에 따라 약간의 변동이 생길 수 있겠지만, 일차적인 목표는 “temporary music”의 구분이다.

이 음악이 “temporary”한지, 혹은 “candidate”한지. 잠깐 떳다 지는 음악인지 꽤나 오랜기간 유지할 음악인지. 더 나아가서 2년, 3년, 4년 내내 차트인 할 permanent 음악인지 확인하는 것이 목표가 될 것이다.

그 clustering을 위한 변수는 ranking, rankingdiff가 주된 변수가 될 것이다.

필요하다면 artist의 영향력을 고려하기 위해 artist도 변수로 사용할 수도 있고, 음악 장르를 고려할 수도 있고, 주간 랭킹이 아닌 일별 ranking을 사용할 수도 있겠지만, 일단은 ranking과 rankingdiff를 중점적으로 볼 것이다.