Spotify 인기 요소의 상관 관계 분석

스트리밍 수가 높은 음악적 특징, 월별/요일별 분석과
스트리밍에 영향을 미치는 요인 파악을 통한 음악 시장 전략

- 03기 제인팀 2조 -

김가람, 박현서, 백승훈

목 차

1. 개요

1.1. 주제 설명 및 프로젝트 목적

1. 주제 선정 배경

음악은 과거에서부터 지속되어 현대까지 이어진 전세계 공통 문화로 자리 잡았습니다. 일상 어디에서든 쉽게 접할 수 있는 음악은 디지털 시대에 접어들면서 더욱 접근이 용이해졌고, 이에 따라 음악의 수요와 소비가 올라갔습니다. 그중에서 해외에서 가장 큰 스트리밍 사이트인 Spotify의 최다 스트리밍 리스트가 담긴 데이터를 선정하였습니다.

해당 데이터를 분석함으로써 얻는 이점은, 성공한 음악 간의 연결 고리를 찾아 소비층의 선호도와 트렌드를 분석할 수 있다는 점입니다. 또한, 스트리밍에 영향을 주는 요소를 찾아낸다면 향후 음악 사업의 성공에 기여할 수 있을 것이라 생각하여 해당 데이터셋을 선정하였습니다.

따라서 큰 주제로는 'Spotify 인기 요소의 상관 관계 분석'을 주제로 설정하였고, 개인별 세부 주제를 통하여 각자의 주제에 맞게 심층 분석하였습니다.

2. 프로젝트 세부 주제

가람님은 스트리밍 수가 높은 곡들 간에 음악적 요소 공통점이 있을 것이라고 가정하여, 스트리밍 상위 100위 곡들의 연관성 높은 요소 간의 상관 관계를 분석하였습니다. 스트리밍 상위 100 곡의 음악 특징 조합 및 분석을 통해 리스트의 전체 곡과 상위권의 곡들에서 특징적으로 나타나는 음악적 요소의 분포도를 분석했습니다.

현서님은 Spotify 내 상위 100개의 곡들을 월과 요일로 나누어 분석하였습니다. 이를 통해 곡의 흥행과 연관되는 스트리밍과 관련된 요소를 분석해 보고자 했습니다. 첫 번째로, 월별/요일별 스트리밍 수 가장 높은 월/요일을 도출하였고. 해당 날짜들의 스트리밍 수와 가장 밀접한 관계가 요소를 시각화하여 스트리밍과의 상관 관계를 분석하였습니다.

승훈님은 장르별 인기 요소 분석을 진행하였습니다. EDA의 하위 목표로써 관심 패턴 감지를 통해 음악적 특성의 패턴이 비슷한 데이터들로 클러스터링하여 장르를 예측하고, 클러스터별 인기 요인을 분석하였습니다. 장르 전체에서 아티스트, 전체적인 데이터에서 개별 값까지 확인하였습니다. 효율적인 정보 전달을 위해 적절한 시각화 사용하며 시각화 이전에 비시각화를 통해 데이터의 정확한 값을 파악하였습니다.

개인별 주제를 통합해 보면, 모두 스트리밍이 높은 음악을 분석하였습니다. 스트리밍이 높다는 것은 음악의 성공과 직결된다는 것을 알 수 있습니다. 따라서 스트리밍 분석을 통해 음악 시장의 활성화 및 미래에 발매될 음악들의 수요 예측이 도움이 될 수 있을 것이라고 생각합니다.

2. 데이터

2.1. 데이터 칼럼 설명

Data : **Most Streamed Spotify Songs 2023 (24 columns, 10 rows)**
https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023

columns:

column	설명	예시
track_name	곡 제목	Seven(feat.Latto) (Explicit Ver.)
artitst_name	곡의 아티스트 이름	Latto, Jung Kook
artist_count	곡에 기여한 아티스트 수	2
released_year	곡이 나온 해	2023
released_month	곡이 발매된 달	7
released_day	곡이 발매된 날	14
in_spotify_playlists	곡이 포함된 Spotify 재생 목록 수	553
in_spotify_chars	Spotify 차트에서 곡의 순위	147
streams	Spotify의 총 스트리밍 수	141381703
in_apple_playlists	곡이 포함된 Apple Music 재생 목록 수	43
in_apple_charts	Apple Music 차트에서의 곡의 순위	263
in_deezer_playlists	곡이 포함된 Deezer 재생 목록 수	45
in_deezer_charts	Deezer 차트에서의 곡의 순위	10
in_shazam_charts	Shazam차트에서의 곡의 순위	826
bpm	분당 비트 수	125
key	곡의 키	B
mode	major or minor	Major
danceability	댄스 음악인지 나타내는 백분율	80
valence	음악적 내용에 대한 긍정성	89
energy	곡의 에너지 수준	83
acousticness	곡의 어쿠스틱 사운드의 양	31
instrumentalness	곡의 포함된 악기의 양	0
liveness	라이브 공연 요소 존재	8
speechiness	노래에 나오는 단어의 양	4

2.2. 데이터 전처리

1. 인코딩

원본 데이터 중 라틴어와 스페인어 등으로 쓰인 글자들에서 인코딩 문제가 발생하였습니다. 이에 대해 단일대체법-일치대응대체법의 방법을 사용하였습니다.

단일대체법-일치대응대체법이란?
결측치를 다른 조사 자료로부터 얻을 수 있는 경우, 외부 자료 값을 대체하는 방법입니다. 이 방법을 통해 key 칼럼에 존재하는 Nan 값의 경우, C코드가 누락되었음을 확인할 수 있었고, 해당 누락값을 대체하였습니다.

2. 불필요한 문자 처리와 타입 변형
컬럼 사용의 효율성을 위해 컬럼에 존재하는 불필요한 문자 (energy_% -> energy) 를 제거하였고, float 데이터 타입을 int로 통일하였습니다.

3. 데이터 분석

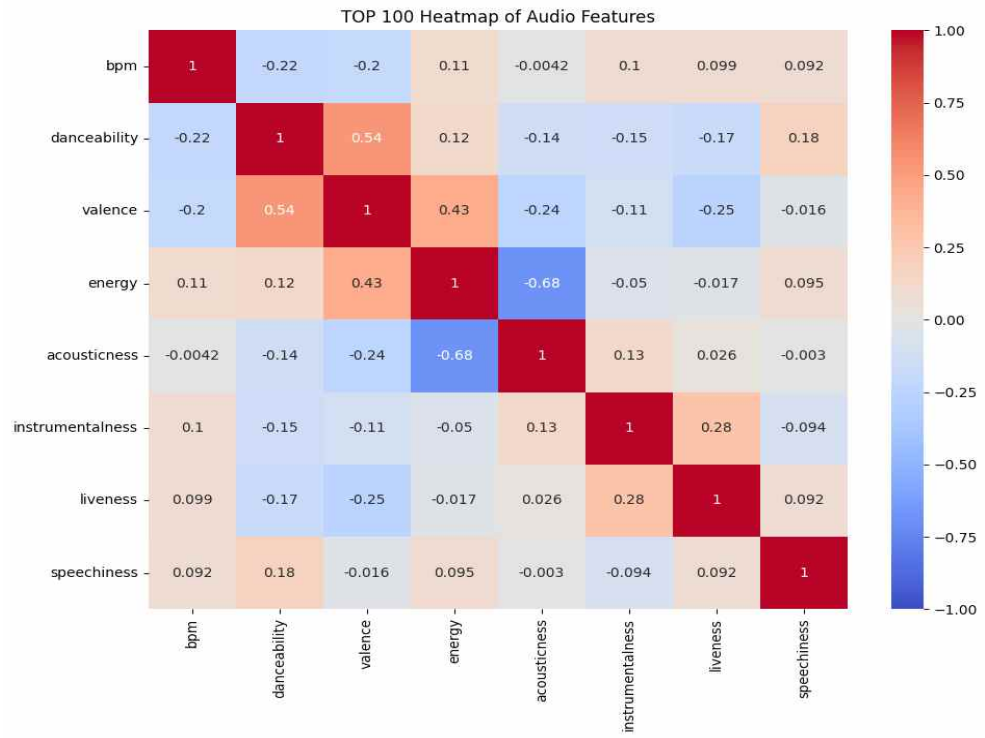
3.1. 인기 음악 요소 분석

1. 인기 음악 요소의 특징 탐구

스트리밍이 높은 곡들 간 음악 요소의 공통점이 있을 것이라고 가정하였습니다. 따라서 스트리밍이 높은 곡의 Top 10, 50, 100, 300, 500, 700으로 그룹을 나누었습니다. 사용된 음악 요소는 bpm ,danceability, valence, energy, acousticness, instrumentalness, liveness, speechiness 칼럼들입니다.

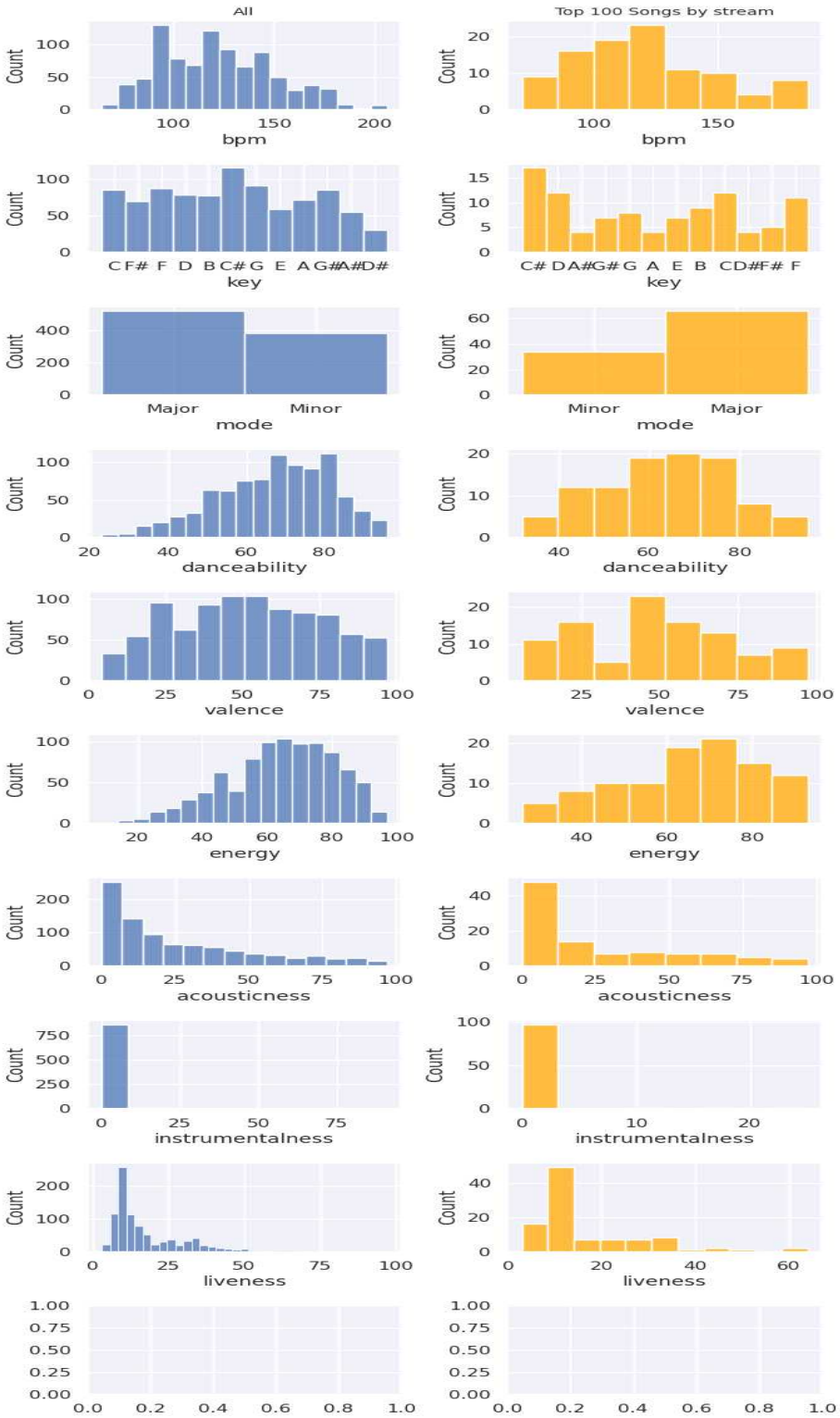
1.1. top 100 그룹의 음악 요소 분석

우선, top 100 그룹을 선택하여 해당 그룹에 포함된 곡들의 연관성이 높을 것이라고 가설을 설정하였습니다. 인기가 좋은 곡들은 특정한 성향을 띌 것이라고 생각하였고, 전체 음악적 요소의 상관 관계를 분석하였습니다.



위의 그림을 보면, top 100 그룹의 전체 음악 요소의 유의미한 공통점이 보이지 않는 것을 알 수 있습니다. 따라서 히스토그램으로 시각화하여 전체 곡의 음악 요소와 top 100 그룹의 음악 요소를 비교해 보았습니다.

1.2. 전체 곡과 top 100 그룹의 음악 요소 비교 분석

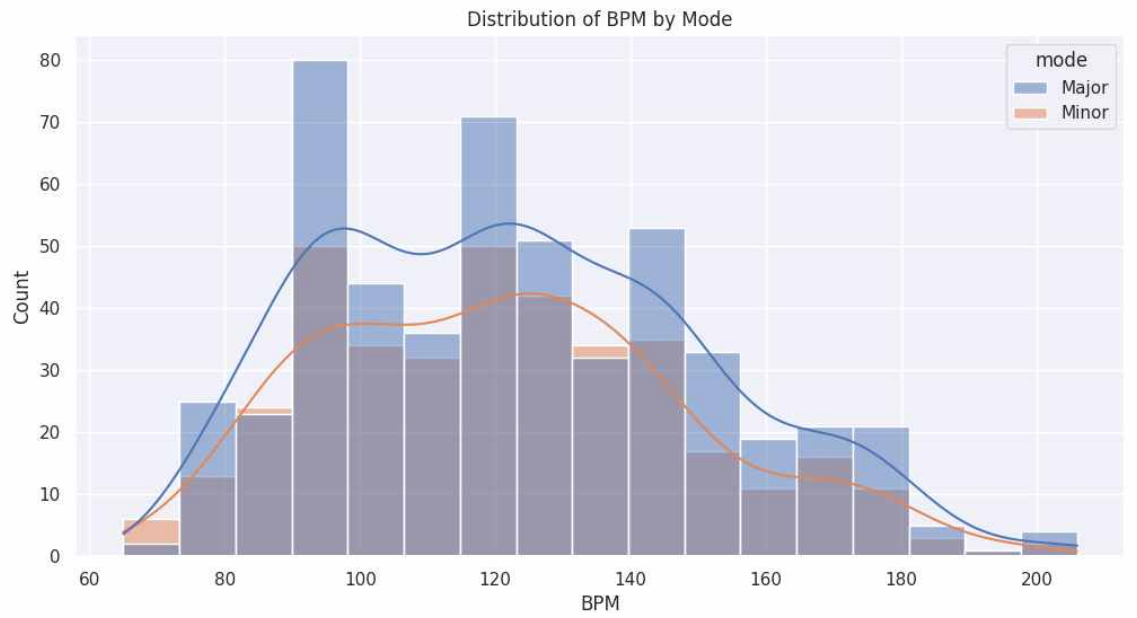


전체 곡과 top 100그룹을 비교한 결과, 몇 가지 특징들이 발견되었습니다.

1. bpm: 전체 곡에서는 100 이하의 곡이 가장 많은 반면, Top 100 그룹은 주로 100~120의 bpm 분포를 보였습니다.
2. key: 전체 곡에서는 C# 곡이 가장 많았고, D#을 제외한 전체 키의 고른 분포를 보이지만 Top 100 그룹에서는 C#, D, C, F 키에 주로 분포를 보였습니다.
3. danceability: 전체 곡에서는 약 50%~80%의 비율을 차지하지만, Top 100 그룹은 60%~80%에 분포되어 있습니다.
4. valence: 전체 곡에서는 어두운 곡을 제외하고 꽤 고르게 분포되어 있으나, Top 100 그룹의 경우는 어두운 곡이 없고 50% 정도의 밝기에 집중되어 있습니다.
5. energy: 전체 곡에서는 에너지가 적은 곡보다 상대적으로 55% 이상 많은 곡들 위주로 분포되어 있으나, Top 100 그룹에서는 60%~80%에 분포를 보였습니다.

2. monde 비교

곡의 무드를 결정하는 major 장조와 minor 장조의 분포도를 비교해 보았습니다. 이는 bpm 분포도와 유사한 흐름을 가질 것이라고 예측하였습니다. 따라서 bpm이 느리면 단조, 빠를수록 장조가 많을 것이라는 가설을 세웠습니다.

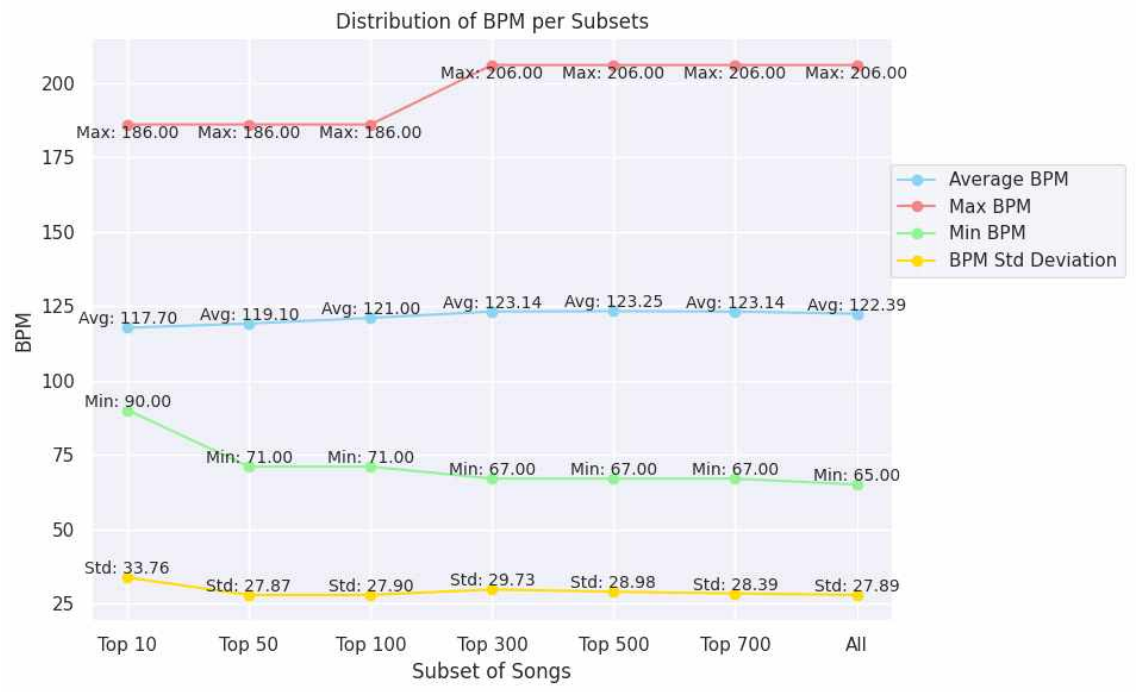


분석 결과, bpm이 느린 곡들이 minor인 경우가 많이 있지만, 전체적으로 major인 곡이 주류를 이룬다는 것을 알 수 있었습니다. 따라서 해당 가설은 기각되었습니다.

3. bpm 분포도 분석

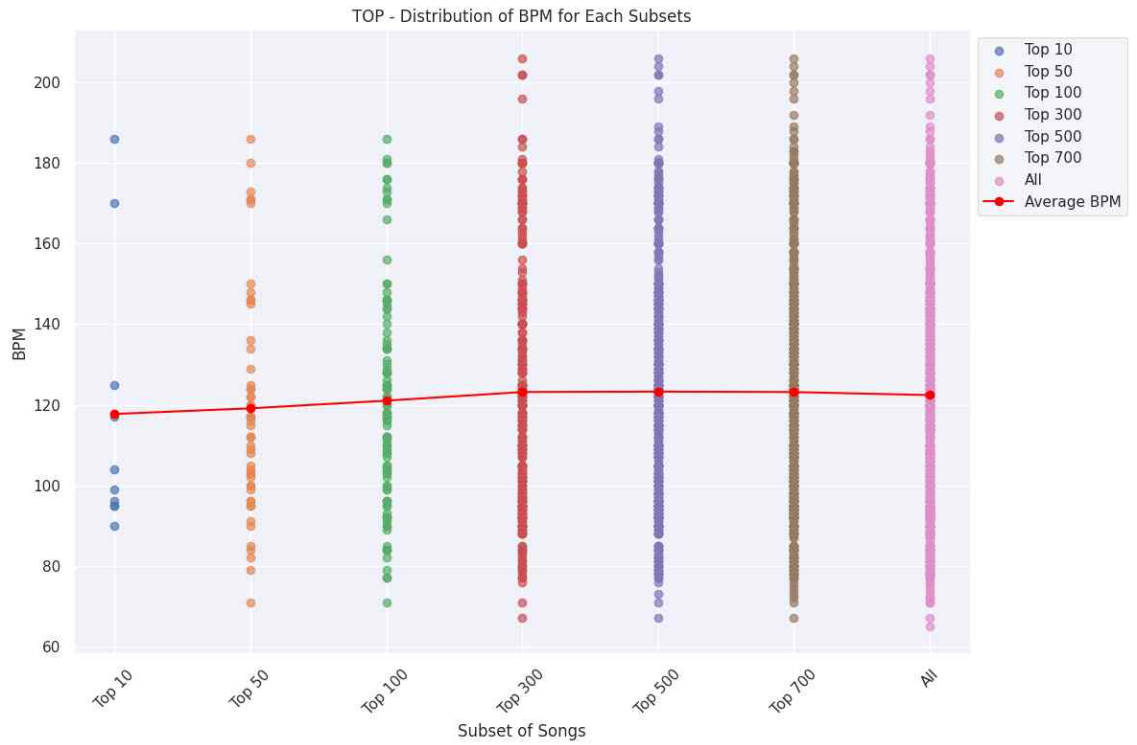
3.1. bpm 분포도 분석

1.2.의 차트에서와 같이 전체 곡과 top 100의 분포도가 달라 각 스트리밍이 많은 구간별 분포도를 파악하고자 bpm의 최저, 최고, 평균, 표준편차를 계산하여 bpm 속도에 따른 선호도가 어느 속도에서 상승폭을 가지는지 알아보았습니다.



top 300 그룹부터 빠른 곡과 느린 곡의 다양성이 넓어진 것을 알 수 있습니다. 그러나 가장 느린 곡은 top 700 그룹 안쪽 순위로는 들지 못하였습니다. 따라서 상대적으로 느린 곡에서 선호도가 낮다는 결과를 도출할 수 있었습니다.

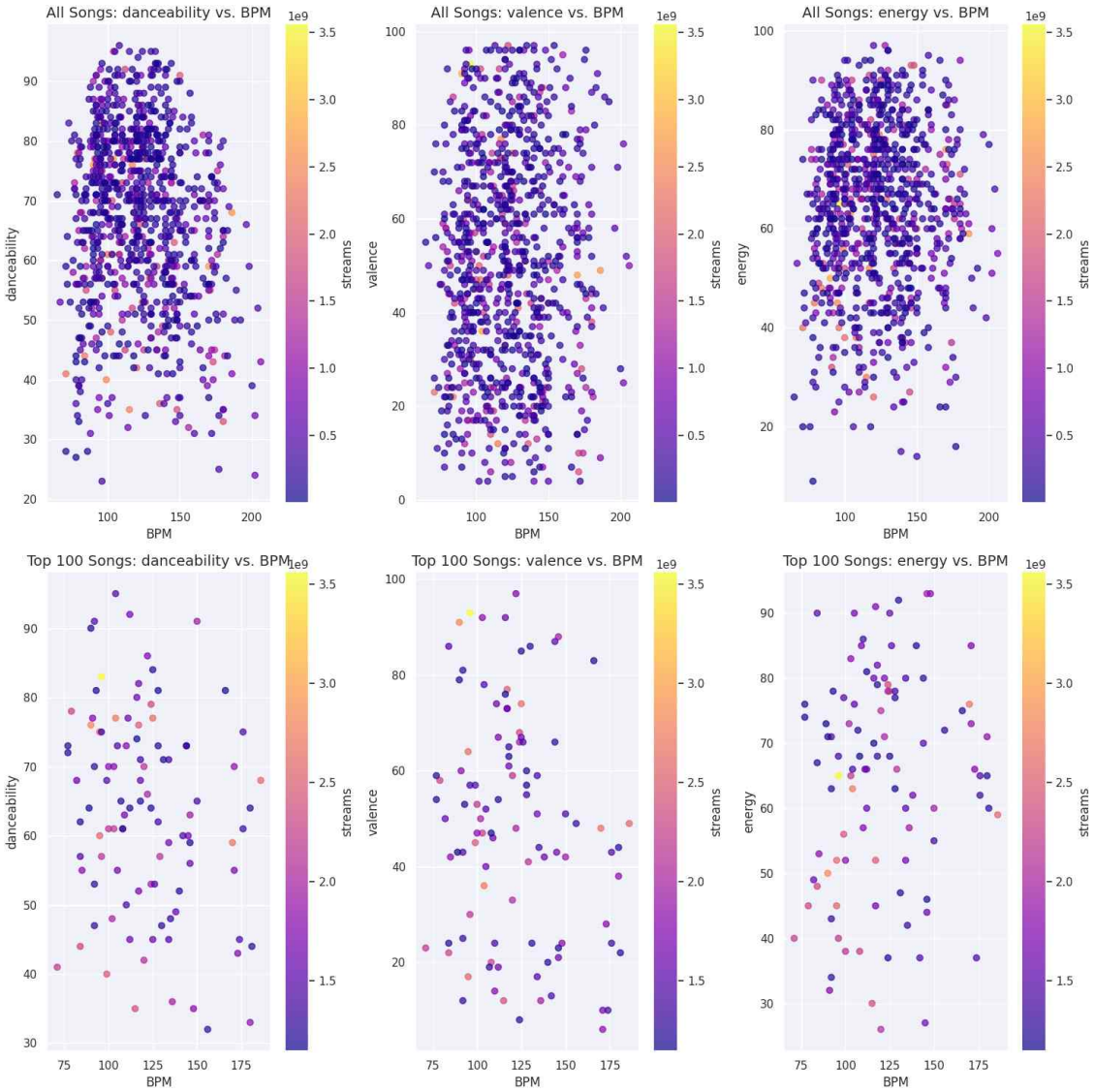
3.2. top 그룹별 선호하는 bpm 분석



위의 결과를 바탕으로 각 순위별 bpm의 밀집도에 차이를 분석해 본 결과, top 10 그룹의 경우 특징적인 부분이 없었습니다. top 50, 100, 300 그룹의 경우 평균값 근처인 120 수준의 100~140에서 분포를 보였습니다. 또한 아주 빠르거나, 아주 느린 곡의 개수 자체가 적은 것을 알 수 있습니다.

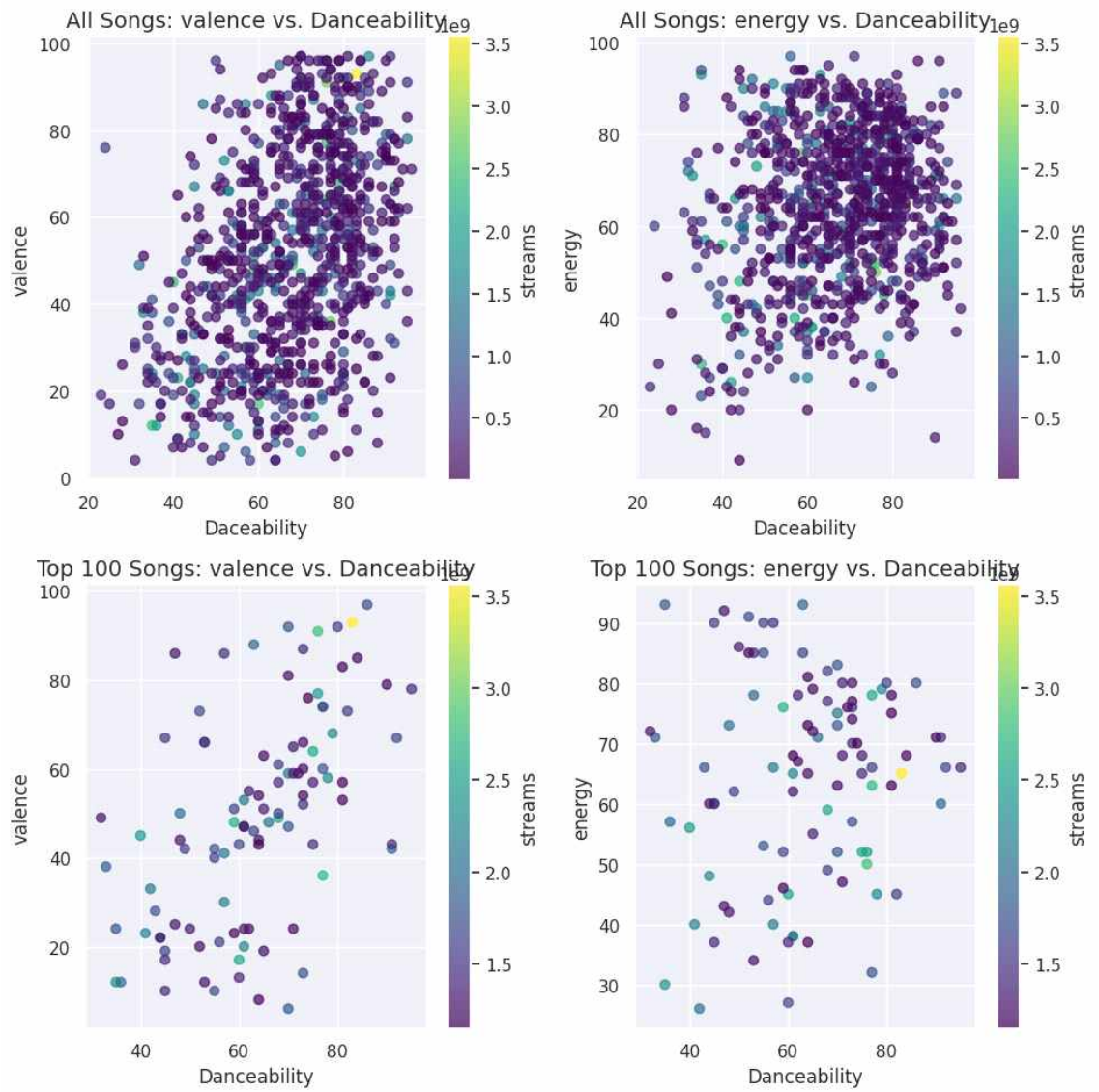
해당 결과를 기반으로 bpm의 빠르기에 따라서 음악적 요소들이 관련 있을 것이라고 가정하였습니다. 즉, danceability, valence, energy 수준이 높을수록 bpm이 높고 인기가 좋은 곡에 몰려있을 것이라고 예상해 보았습니다.

3.3. 전체 곡과 top 100 그룹의 bpm에 따른 danceability, valence, energy 분포



분석 결과, bpm과 danceability, valence, energy에 대한 연관성이 매우 낮다는 것을 알 수 있었습니다. 또한 이러한 조합들이 스트리밍에 큰 영향을 끼치지 않았습니다.

그렇다면 danceability, valence, energy 이 세 가지 음악 요소는 어떤 특징과 연관성이 높은지 비교해 보겠습니다.



전체 곡과 top 100 그룹의 뚜렷한 연관성을 보이지 않지만, bpm에 비해 상대적으로 danceability가 높을수록 valence, energy 높은 경향이 나타났습니다.

그렇다면 top 100 그룹 내에서 가장 이상적인 음악 요소들을 갖추고 있는 곡은 몇 곡인지 알아겠습니다. top 100 그룹 내 bpm은 100~120, key 값은 C#, D, C, F, danceability는 60%~80%, valence는 50%. energy는 60%~80%인 곡의 개수를 출력해 보았습니다

3.4. 이상적인 음악 요소를 가진 곡의 개수

```
Number of songs =: 0
```

결과적으로, 이러한 이상적인 요소를 모두 갖춘 곡은 0개였습니다. 따라서 top 100 그룹에는 많은 공통 요소들이 존재하지만, 모든 음악 요소가 전부 동일하지 않다는 결과를 도출할 수 있었습니다.

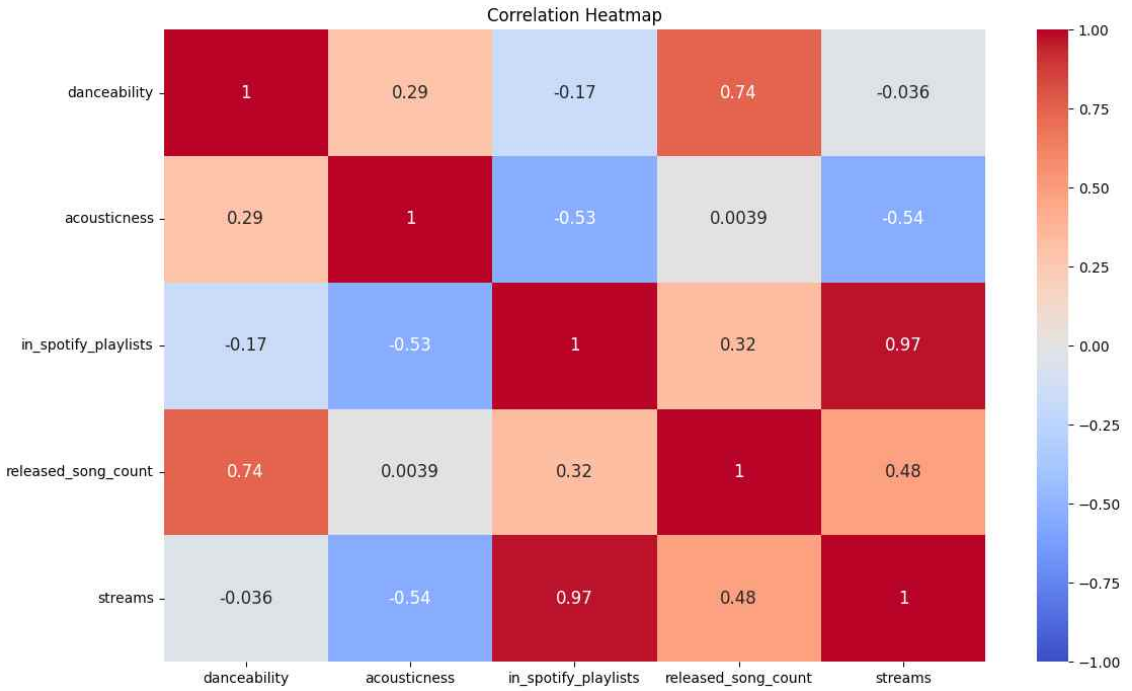
3.2. Spotify 스트리밍 top 100 음악의 월별/요일별 분석

월별 및 요일별로 스트리밍 수를 분석함으로써 소비자들이 어느 시기에 가장 많이 음악을 듣는지, 특정 기간 동안 어떤 변화가 있는지 파악할 수 있습니다. 또한, 각 월별/요일별로 발매된 곡의 수와 스트리밍 수를 비교하여 음악 산업 종사자들이 가장 효과적인 발매 시기를 결정하는 데 도움을 줄 수 있습니다. 예를 들어, 특정 월에 스트리밍 수가 급증하거나 급감하는 패턴이 발견된다면, 그 원인을 분석하고 발매 전략을 최적화할 수 있습니다.

1. 월별 분석

1.1. Spotify top 100 곡의 음악 요소 상관 관계 분석

이전에 미리 분석하여 유의미한 상관 관계를 보인 요소들만 시각화하였습니다. 음악 요소는 danceability, acousticness, in_spotify_playlists, released_song_count, streams 입니다.

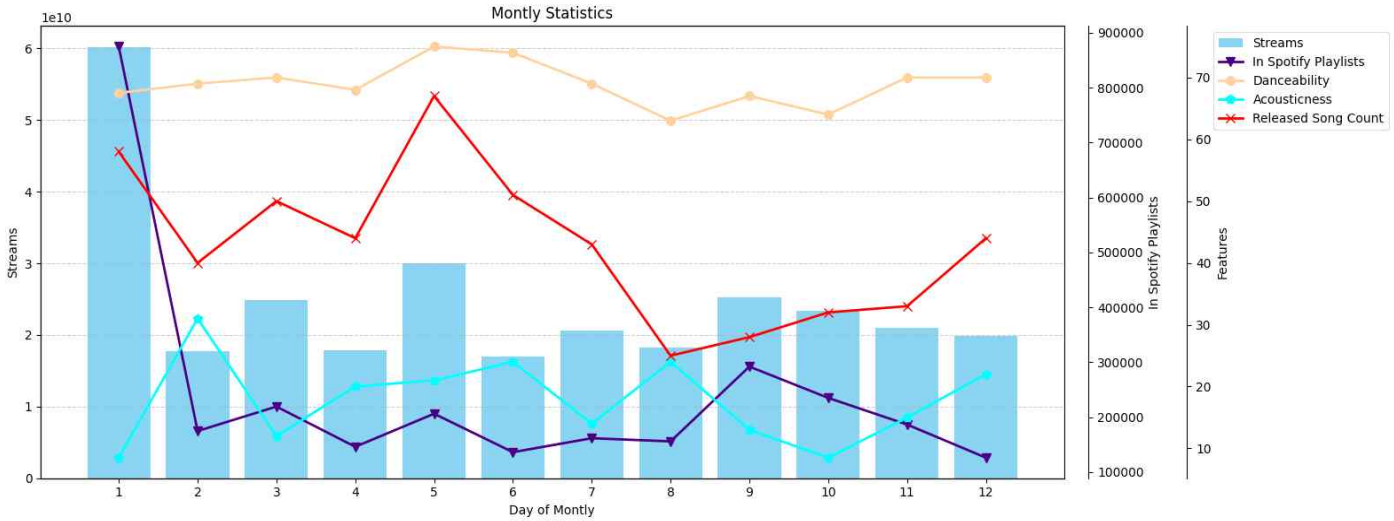


히트맵의 경우 상관 관계 계수가 0.7 이상이면 높은 상관 관계를 가진다고 할 수 있습니다. 따라서 히트맵에서 상관 관계 계수가 0.7 이상인 값을 설명하겠습니다.

1. danceability와 released_song_count (0.74)
: 발매된 곡들의 댄스곡 비중이 높습니다.
2. in_spotify_playlists와 streams (0.97)
: 재생 목록에 많이 포함된 곡일수록 스트리밍 수도 높습니다.

해당 분석 목적에서 가장 중점을 두는 요소는 streams입니다. 이는 음악의 흥행에 직결되는 요소이므로, 앞으로의 분석에서도 in_spotify_playlists 항목 또한 주의 깊게 볼 것입니다.

1.2. Spotify top 100 곡의 월별 통계 시각화

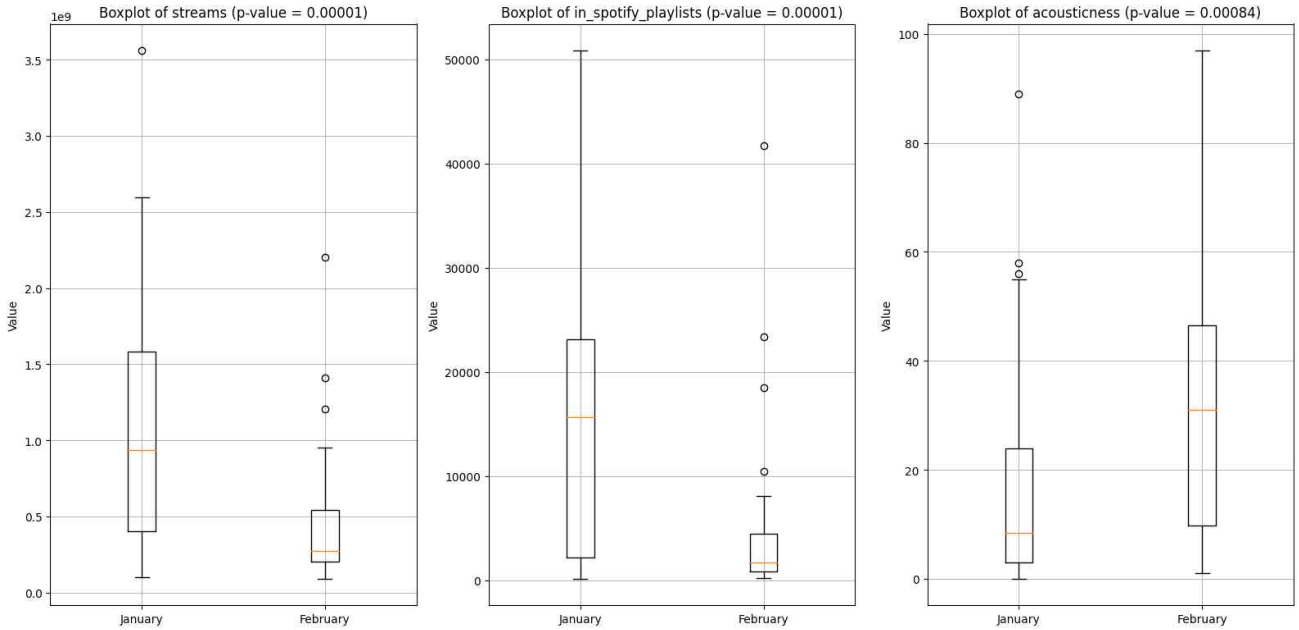


streams와 in_spotify_playlists, released_song_count의 증감이 유사한 패턴을 보이고 있습니다. 특히, 1월의 streams와 in_spotify_playlists의 값 눈에 띕니다. 이는 보통 새로운 음악이 매달 초에 많이 발매되기 때문일 가능성이 큽니다.

released_song_count는 9월에 다시 증가하는 경향을 보입니다. 이는 상반기 이후로 침체되어 있던 음악 시장의 새로운 음악 발매 시기임을 의미합니다. 또, 히트맵에서 이외의 음악 요소는 streams와 상관관계가 낮음을 파악했으므로, 단순히 확인만 하겠습니다.

따라서 첫째로, streams이 가장 많았던 1월과, 바로 다음 달이지만 1년 중 streams이 가장 낮은 2월의 Anova 분석을 통해 더 자세히 살펴보고자 합니다. 둘째로, streams 수가 1, 2등인 1월과 5월의 Anova 분석을 시행해 보겠습니다.

1.3. 1월과 2월의 anova 분석



p-value에 대한 설명
p-value는 귀무 가설을 지지할 확률을 나타냅니다. p-value의 값이 작을수록 귀무 가설(월별 차이가 없다)을 기각하고 대립 가설(월별 차이가 있다)을 지지합니다.

세 박스플롯 모두 p-value가 0.05 이하로 매우 낮은 값이 도출되었습니다. 첫 번째 그래프를 예시로 설명하겠습니다. 해당 그래프의 귀무 가설은 1월과 2월의 streams 평균에 차이가 없다이고, 대립 가설은 이에 반대입니다. 이때, p-value의 값이 0.00001이므로, 이는 귀무 가설을 기각합니다. 즉, 1월과 2월의 streams의 평균에 대한 유의미한 차이가 있다는 결과를 도출할 수 있습니다.

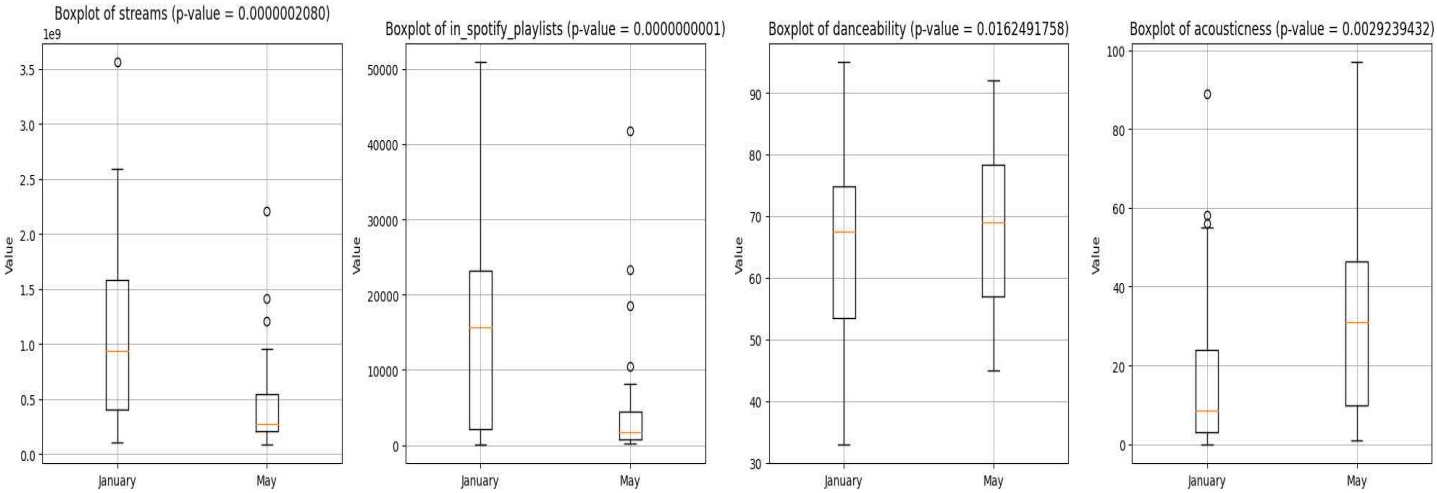
3-1) streams
1월의 streams 중앙값이 약 10억으로 2월의 중앙값인 약 5억보다 두 배 높습니다. 또한, 1월의 스트림 수는 3.5억 이상의 이상치가 몇 개 존재하는 반면, 2월의 이상치는 이보다 현저히 적습니다. 이는 1월에 스트림 수가 월등히 높다는 것을 보여줍니다.

3-2) in_spotify_playlists
1월의 in_spotify_playlists 중앙값은 약 20,000으로, 2월의 중앙값인 약 10,000보다 약 두 배 정도 높습니다. 또한, 1월의 상위 사분위수는 50,000에 달하며 몇 개의 이상치가 존재합니다. 이는 1월에 더 많은 곡들이 Spotify 플레이리스트에 포함되었음을 나타냅니다.

3-3) Acousticness
1월의 Acousticness 중앙값은 약 20으로, 2월의 중앙값인 약 40보다 약 두 배 낮습니다. 이는 2월에 발매된 곡들이 1월에 비해 Acousticness 특성이 더 강함을 나타냅니다. 또한, 1월의 어쿠스틱 지수는 60 이상의 몇몇 이상치가 존재하며, 2월의 분포는 상대적으로 균등합니다.

분석을 통해 1월과 2월의 스트리밍 데이터 및 음악적 특성을 비교한 결과, 1월에는 streams, in_spotify_playlists 수가 더 많았으며, 이는 1월에 발매된 곡들이 더 많은 인기를 얻었음을 의미합니다. 반면, 2월에는 Acousticness 곡들이 더 많이 발매되었고, 이에 따른 청취 패턴의 변화를 반영합니다.

1.4. 1월과 5월의 anova 분석



4-1) streams
1월의 중앙값이 약 10억으로, 5월의 중앙값인 약 2억보다 약 5배 높습니다. 또한, 1월의 스트림 수는 3.5억 이상의 이상치가 몇 개 존재하는 반면, 5월의 이상치는 이보다 훨씬 적습니다. 이는 1월에 streams 수가 월등히 높다는 것을 보여줍니다.

4-2) in_spotify_playlists
1월의 중앙값은 약 20,000으로, 5월의 중앙값인 약 5,000보다 약 4 배 높습니다. 또한, 1월의 상위 사분위수는 50,000에 달하며 몇 개의 이상치가 존재합니다. 이는 1월에 더 많은 곡들이 Spotify 플레이리스트에 포함되었음을 나타냅니다.

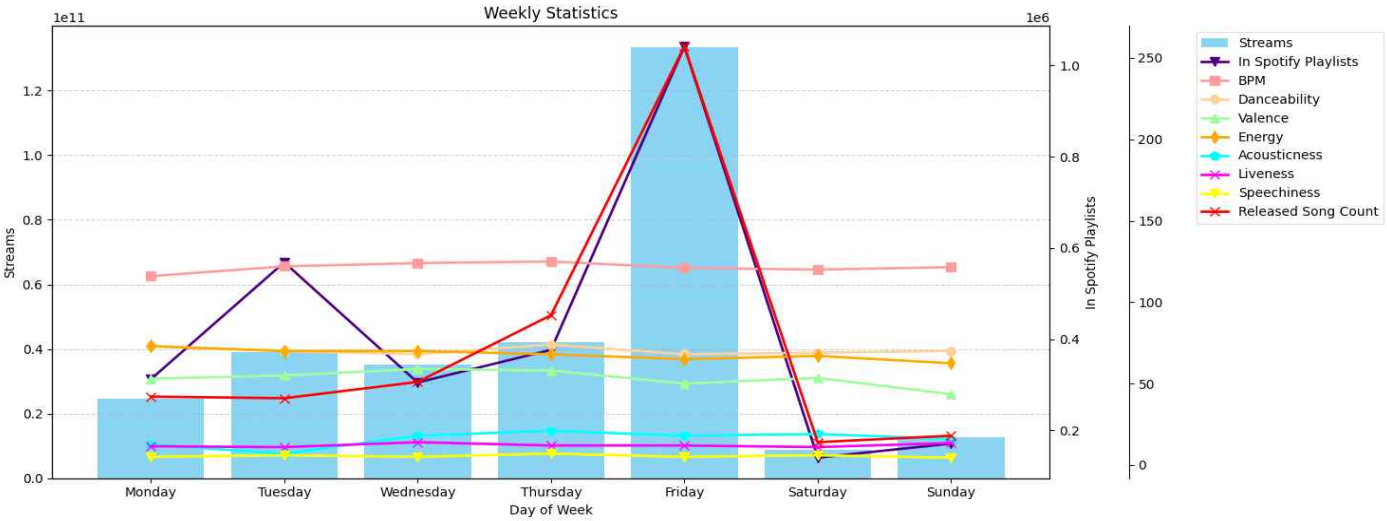
4-3) Danceability
1월과 5월의 중앙값은 비슷하지만, 5월의 값이 조금 더 높습니다. 이는 5월에 발매된 곡들이 1월에 비해 Danceability 특성이 약간 높은 것을 알 수 있습니다.

4-4) Acousticness
1월의 중앙값은 약 20으로, 5월의 중앙값인 약 40보다 2 배 정도 낮습니다. 이는 5월에 발매된 곡들이 1월에 비해 Acousticness 특성이 더 강함을 나타냅니다. 또한, 1월의 어쿠스틱 지수는 60 이상의 몇몇 이상치가 존재하며, 5월의 분포는 상대적으로 균등합니다.

분석을 통해 1월과 5월의 streams, in_spotify_playlists, Danceability, 그리고 Acousticness 지수에 유의미한 차이가 있음을 확인할 수 있었습니다. 예를 들어, 5월에는 더 높은 Danceability와 Acousticness 특성을 가진 곡을 발매하는 전략을 세울 수 있습니다.

2. 요일별 분석

2.1. Spotify top 100 곡의 요일별 통계 시각화

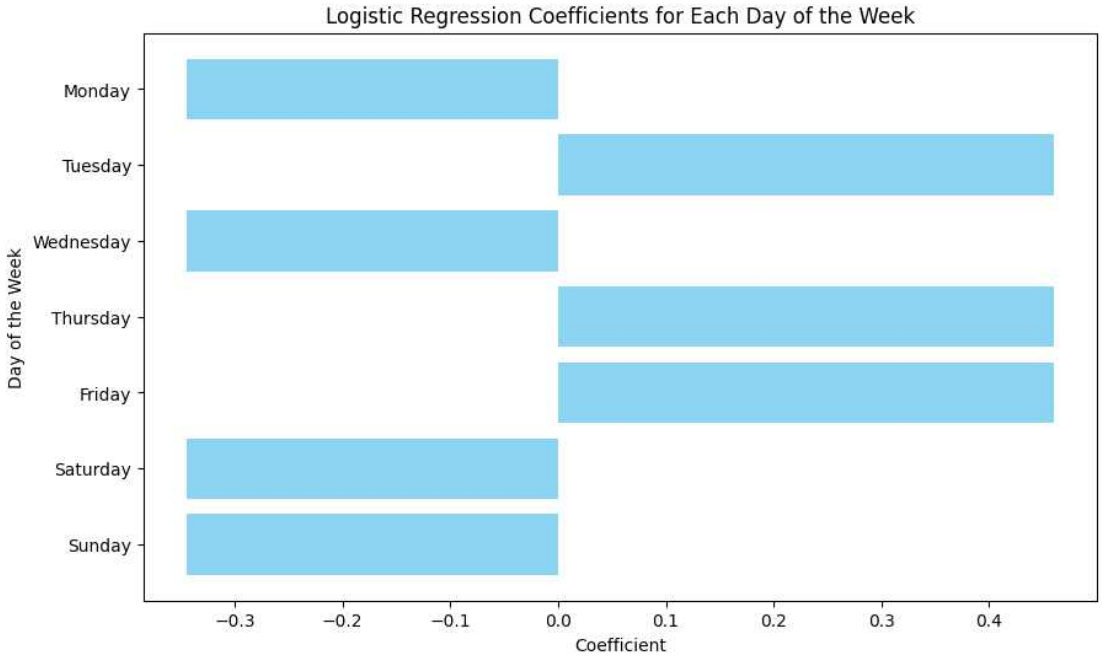


월별 분석과 마찬가지로 요일별 분석에서도 streams와 in_spotify_playlists, released_song_count이 유사한 흐름을 보이고 있습니다. 특히 금요일에 세 요소 모두 급격히 증가한 현상이 눈에 띕니다. 이는 금요일이 새로운 음악 발매일로 자주 선택되면서 나타났을 가능성이 큽니다.

월별 분석과 한 가지 다른 점은, BPM, Danceability, Valence, Energy, Acousticness, Liveness, Speechiness와 같은 음악적 요소들이 요일에 관계 없이 거의 일정한 값을 유지한다는 것입니다. 이는 특정 요일에 음악의 요소가 크게 변하지 않음을 시사합니다.

이 분석을 통해 금요일이 음악 소비와 발매에 있어서 중요한 날임을 확인할 수 있으며, 음악의 요소는 요일에 상관없이 일관되게 유지됨을 알 수 있습니다.

2.2. 요일별 회귀 분석



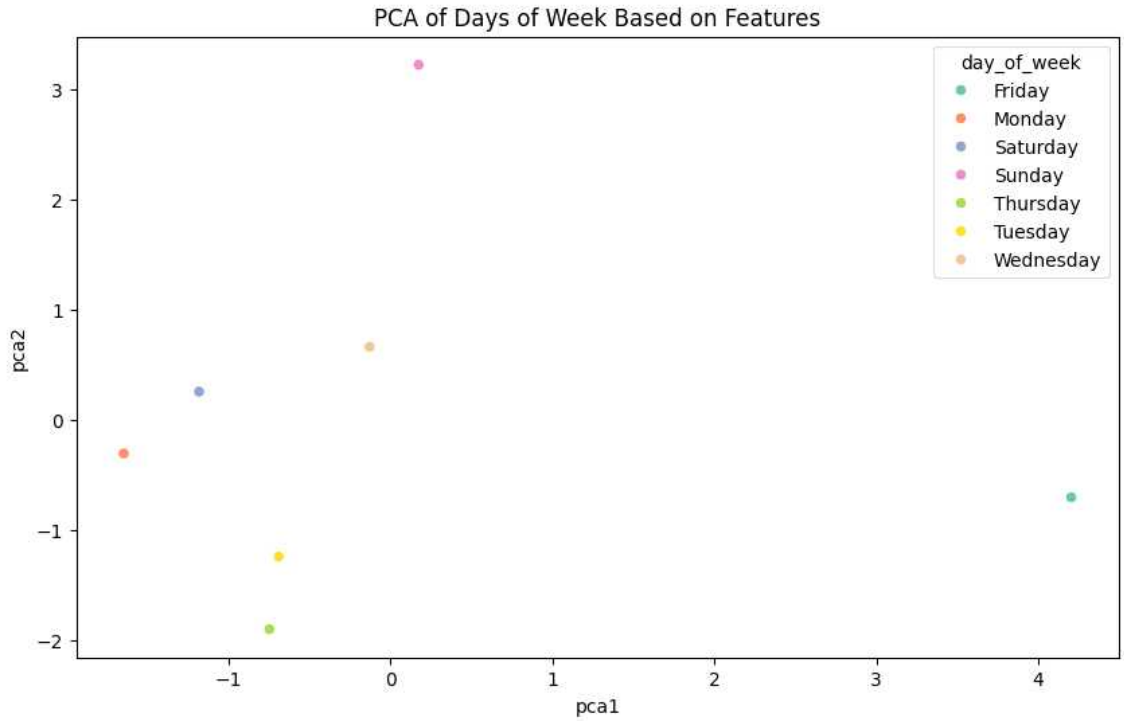
Logistic Regression에 대한 설명
Coefficients: [[0.45958091, -0.34472259, -0.34472259, -0.34472259, 0.45958091, 0.45958091, -0.34472259]]
Intercept: [-0.29771258]
Score: 1.0

1. Coefficients: 음수의 경우 해당 독립 변수가 종속 변수의 값을 높일 확률을 감소
   양수의 경우 해당 독립 변수가 종속 변수의 값을 높일 확률을 증가
2. Intercept: 모든 독립 변수가 0일 때, 종속 변수의 값
3. Score: 모델의 정확도 (0.8 이상이면 높다고 평가된다.)

화요일, 목요일, 금요일의 계수는 양수로 나타나 있습니다. 이는 해당 요일들이 종속 변수에 긍정적인 영향을 미친다는 것을 의미합니다. 월요일, 수요일, 토요일, 일요일은 종속 변수에 부정적인 영향을 미친다는 것을 알 수 있습니다.

이 그래프를 통해 요일별로 특정 결과에 미치는 영향을 비교할 수 있으며, 이를 바탕으로 특정 요일에 대한 전략을 수립할 수 있습니다. 예를 들어, 화요일, 목요일, 금요일이 긍정적인 영향을 미치므로, 이 요일에 중요한 이벤트나 활동을 집중시키는 전략을 고려해 볼 수 있습니다.

2.3. 요일별 pca 분석



PCA에 대한 설명
PCA를 통해 고차원 데이터를 2차원으로 축소함으로써, 데이터의 주요 패턴을 시각적으로 파악할 수 있습니다.

X축 (pca1): 주성분 1. 데이터의 분산을 최대한 설명하는 방향입니다.
Y축 (pca2): 주성분 2. 주성분 1과 직교하며, 데이터의 추가적인 분산을 설명합니다.

분석 결과, 상대적으로 streams, in_spotify_playlists, realsed_song_count의 값이 높은 금요일이 다른 요일들과 확연히 구분되는 위치에 있습니다. 이는 금요일의 데이터 특성이 다른 요일과 크게 다름을 시사합니다. 나머지 요일들은 비교적 가까운 위치에 있으며, 이는 이들의 데이터 특성이 비슷함을 나타낼 수 있습니다.

이를 통해 금요일이 다른 요일과 구분되는 특징적인 패턴을 보유하고 있으며, 나머지 요일들은 비교적 유사한 특성을 공유하고 있음을 알 수 있습니다.

3. 결론

이번 분석을 통해 다음과 같은 결론을 도출할 수 있습니다. 1월, 특히 금요일에 새로운 곡을 발매하는 것이 높은 스트리밍 수와 재생목록 포함 수를 유도하는 데 효과적일 수 있습니다. 이는 새해 초와 앨범 발매율이 가장 높은 금요일에 음악 소비가 활발히 이루어지는 패턴을 반영합니다. 높은 스트리밍 수를 유도하려면 재생목록 포함 수를 늘리는 전략이 중요합니다. Danceability와 같은 특정 음악 요소도 발매 전략에 고려할 요소입니다. 화요일, 목요일, 금요일에 중요한 이벤트나 프로모션을 집중시키는 전략이 효과적일 수 있습니다.

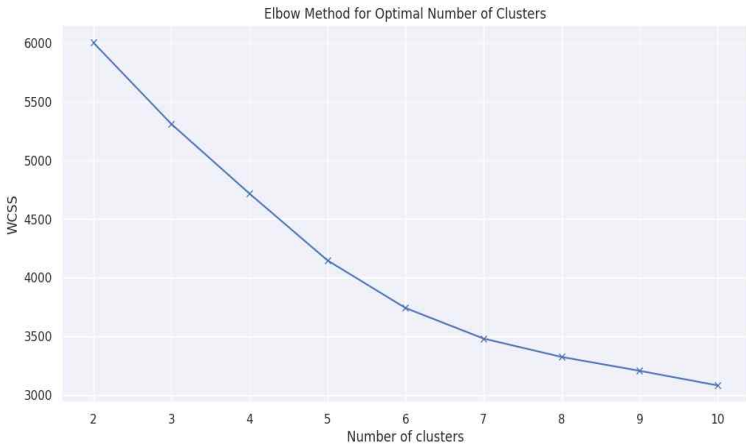
3.3 음악 장르별 인기 요소 분석

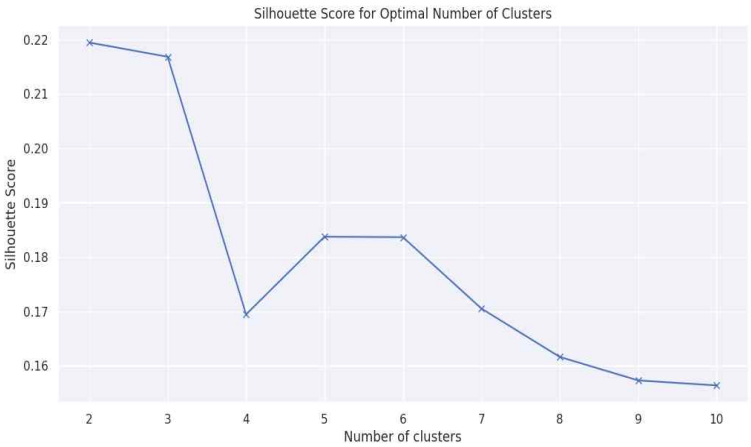
장르 내 스트리밍 순위에 영향을 미치는 음악적 요소가 있을 것이라고 생각하였고, 해당 요소를 중심으로 분석해 보고자 했습니다. 사용된 음악의 요소는 bpm, danceability, valence, energy, acousticness, instrumentalness, liveness, speechiness입니다.

1. 적정 클러스터 개수 파악

가설에 대한 분석을 진행하기 위해 우선 데이터셋에는 장르에 대한 데이터가 없기 때문에, 우선 군집화를 통해 장르 예측을 하겠습니다. 군집화 방법으로 K-means Clustering을 사용합니다. 진행하기 앞서, K-means Clustering의 적정 클러스터 개수를 파악하기 위해 Elbow Method와 Silhouette Score를 사용하였습니다.

Elbow Method : K-Means 군집화에서 최적의 군집(클러스터) 개수를 찾기 위한 방법
Silhouette Score : 군집화 결과의 품질을 평가하는 지표





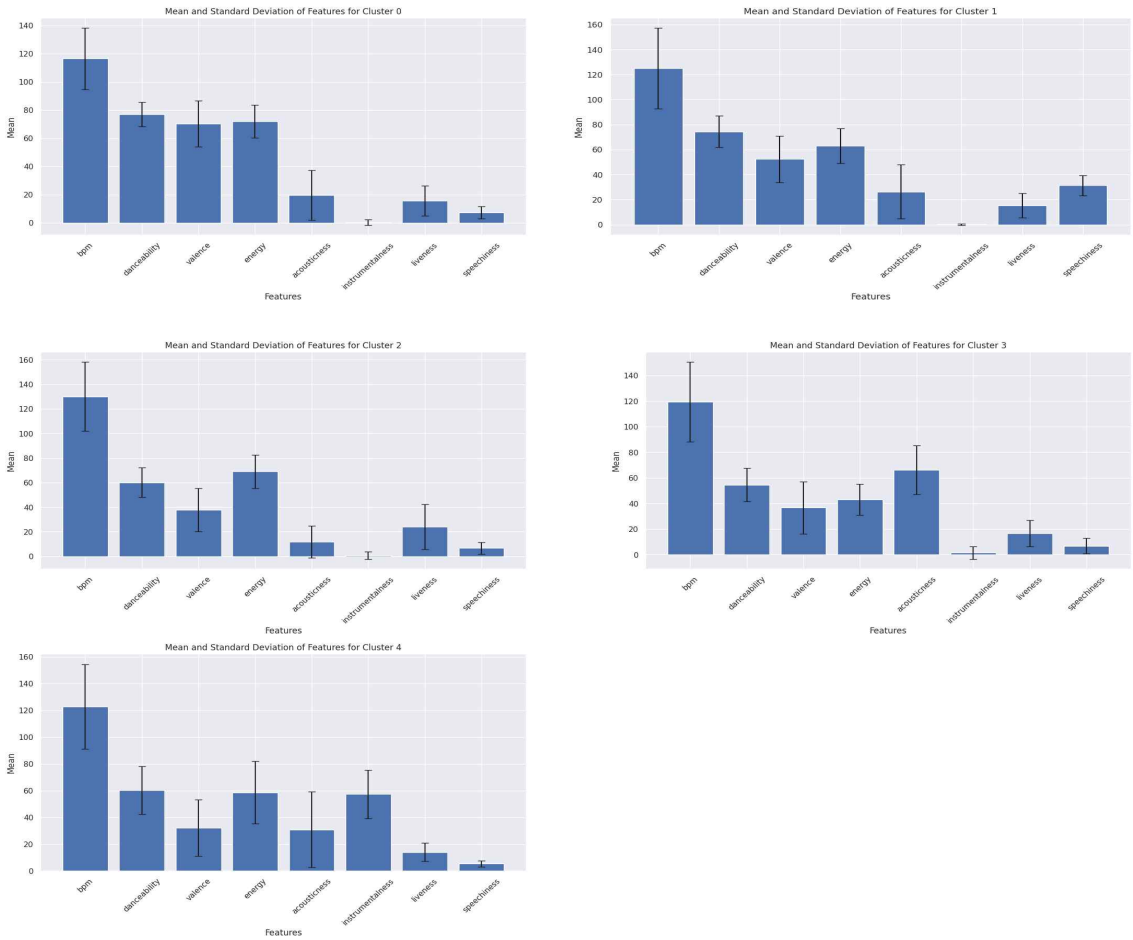
	Number of Clusters	WCSS	Silhouette Score
0	2	6003.738875	0.219478
1	3	5309.788062	0.216843
2	4	4717.630899	0.169441
3	5	4148.464111	0.183741
4	6	3742.657985	0.183655
5	7	3482.259303	0.170529
6	8	3325.727964	0.161614
7	9	3207.681263	0.157301
8	10	3083.069686	0.156374

Elbow Method는 그래프에서 WCSS가 급격히 감소하다가 완만해지는 지점을 찾습니다. 단어 그대로 팔꿈치처럼 꺾이는 모습을 보이는 지점이 적절한 개수를 의미합니다. 첫 세 구간보다 비교적 완만해지는 구간인 5 부근이 적정 클러스터 개수입니다. SilhouetteScore는 클러스터의 품질을 나타내며, 값이 클수록 클러스터가 잘 구분되었음을 의미합니다. 그래프에서 2, 3 지점이 높은 실루엣 점수를 가지고 있어서 적절한 클러스터 개수일 수 있지만, 너무 적거나 너무 많은 클러스터는 적합하지 않은 개수일 수 있기 때문에 다음으로 높은 값인 5개의 클러스터로 진행하였습니다.

2. K-means Clustering

	bpm		danceability		valence		energy		acousticness		instrumentalness		liveness		speechiness	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
cluster																
0	116.55	21.8	76.81	8.69	70.34	16.35	71.9	11.69	19.6	17.76	0.29	1.96	15.61	10.51	7.19	4.29
1	125.09	32.27	74.32	12.57	52.32	18.54	63.06	13.89	26.28	21.52	0.08	0.57	15.35	9.73	31.28	8.0
2	130.04	28.22	60.07	12.16	37.71	17.68	68.96	13.42	11.84	12.91	0.69	3.03	24.02	18.42	6.62	4.7
3	119.4	31.16	54.49	12.92	36.55	20.4	42.93	12.07	65.94	19.1	1.35	4.95	16.44	10.34	6.7	6.11
4	122.76	31.61	60.35	17.84	32.24	20.98	58.65	23.36	31.0	28.22	57.41	18.15	14.12	6.95	5.41	2.43

5개의 클로스터로 클러스터링하여 각 음악 요소들의 평균값과 표준편차를 계산한 데이터 프레임으로 표현하였습니다. 참고할 점은, 데이터 포인트가 20개 미만인 cluster 4의 경우 데이터 포인트 수의 부족으로 신뢰성이 떨어진다는 점입니다.

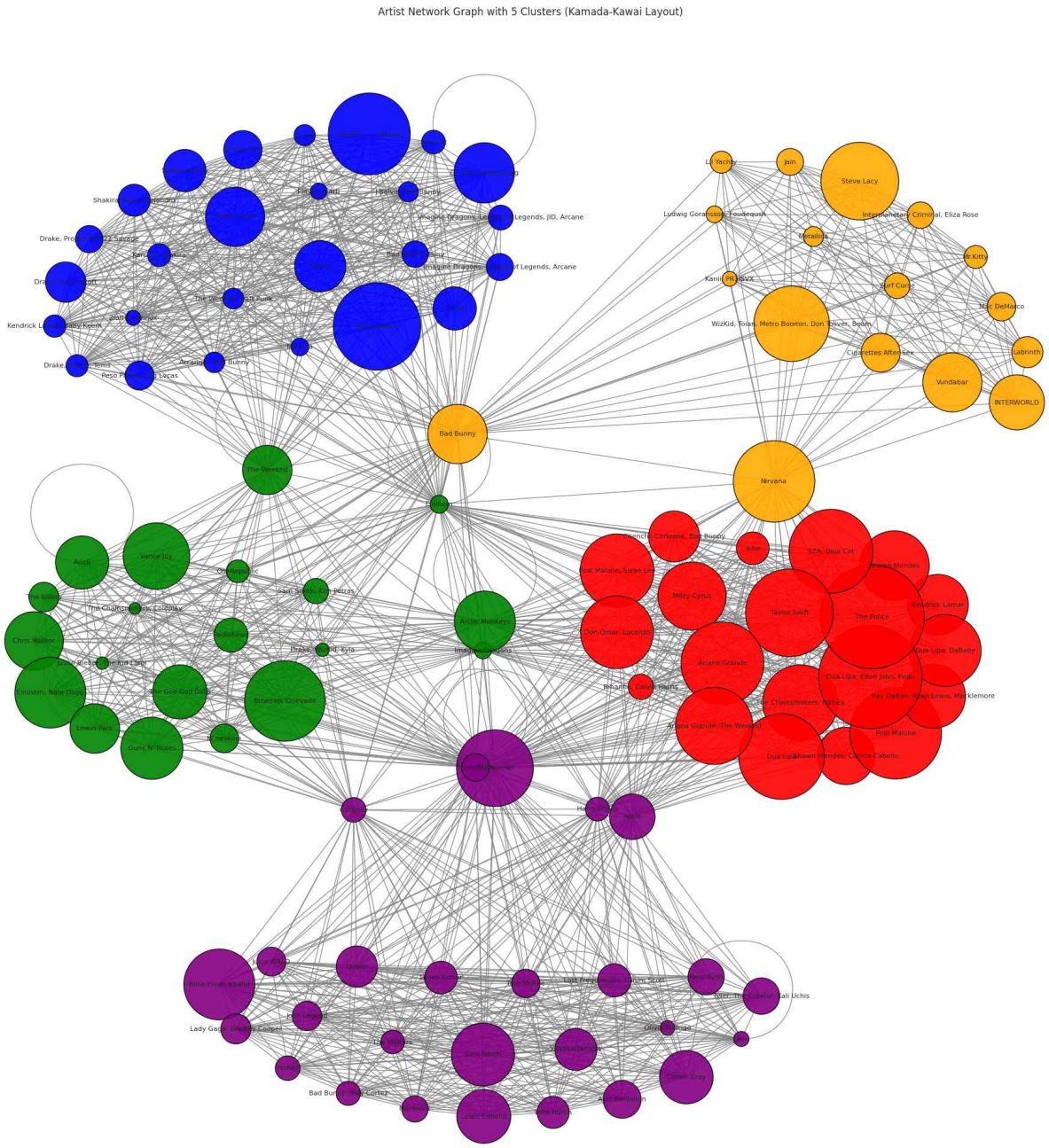


수치로 확인하기 전, 어떤 음악 요소가 두드러지는 값을 띄는지 확인하기 위해 시각화를 진행하였습니다. 막대그래프와 에러바를 사용하여 각 음악 요소들의 표준편차도 확인할 수 있도록 하였습니다. 해당 그래프를 통해 다음과 같이 장르를 예측해 볼 수 있었습니다.

cluster	장르
cluster0	팝, 댄스
cluster1	랩, 힙합
cluster2	록, 일레트로닉
cluster3	인디, 어쿠스틱
cluster4	재즈, 클래식

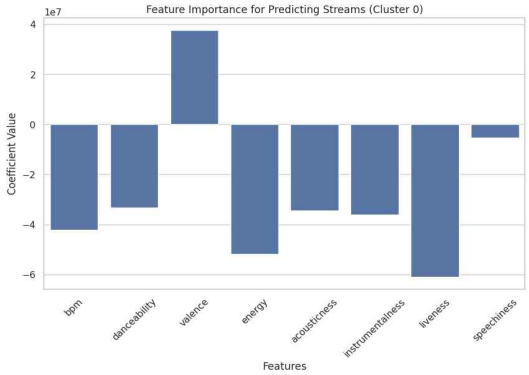
3. 음악성 유사도 네트워크 그래프

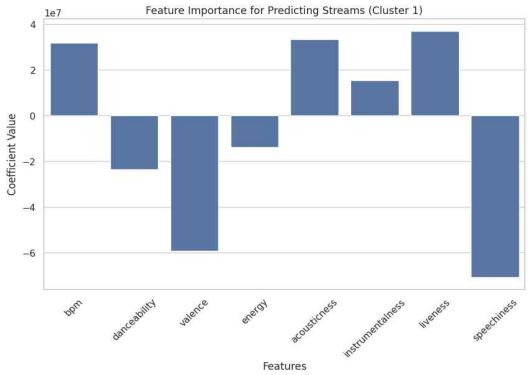
예측된 장르를 바탕으로 음악성이 비슷한 아티스트 간의 네트워크 그래프를 구성하였습니다. 해당 그래프를 통해 각 장르의 아티스트와 다양한 장르를 아우르는 아티스트들을 확인할 수 있었습니다.

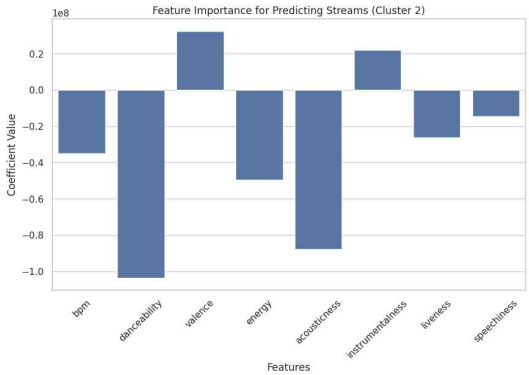


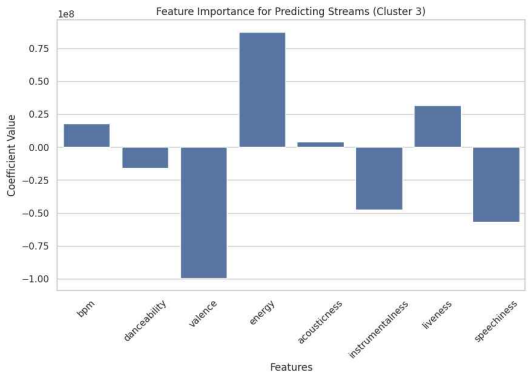
4. 다중회귀 분석

스트리밍에 영향을 주는 음악의 요소가 있을 것이라는 가설을 바탕으로 다중 회귀 분석 모델을 적용해 보았습니다. 다중 회귀 분석은 여러 독립 변수가 종속 변수에 미치는 영향을 평가하기 위한 기본적인 분석 방법으로 사용됩니다. 해당 분석에서는 음악의 요소를 독립변수로, 스트리밍을 종속 변수로 설정하였습니다.









Cluster 0 Summary:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	1.595
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.125
Time:	04:49:09	Log-Likelihood:	-7043.2
No. Observations:	328	AIC:	1.410e+04
Df Residuals:	319	BIC:	1.414e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.771e+08	2.87e+07	16.636	0.000	4.21e+08	5.34e+08
x1	-4.224e+07	3.01e+07	-1.403	0.162	-1.01e+08	1.7e+07
x2	-3.328e+07	3.03e+07	-1.099	0.273	-9.29e+07	2.63e+07
x3	3.765e+07	3.16e+07	1.190	0.235	-2.45e+07	9.99e+07
x4	-5.19e+07	3.13e+07	-1.657	0.099	-1.14e+08	9.73e+06
x5	-3.456e+07	3.07e+07	-1.124	0.262	-9.5e+07	2.59e+07
x6	-3.61e+07	2.9e+07	-1.245	0.214	-9.31e+07	2.09e+07
x7	-6.095e+07	2.97e+07	-2.055	0.041	-1.19e+08	-2.62e+06
x8	-5.308e+06	2.95e+07	-0.180	0.858	-6.34e+07	5.28e+07

Omnibus:	170.795	Durbin-Watson:	1.698
Prob(Omnibus):	0.000	Jarque-Bera (JB):	845.355
Skew:	2.239	Prob(JB):	2.71e-184
Kurtosis:	9.466	Cond. No.	1.65

Cluster 1 Summary:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.053
Model:	OLS	Adj. R-squared:	-0.012
Method:	Least Squares	F-statistic:	0.8209
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.586
Time:	04:49:10	Log-Likelihood:	-2700.6
No. Observations:	127	AIC:	5419.
Df Residuals:	118	BIC:	5445.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.76e+08	3.83e+07	9.825	0.000	3e+08	4.52e+08
x1	3.189e+07	4.04e+07	0.790	0.431	-4.81e+07	1.12e+08
x2	-2.362e+07	4.29e+07	-0.550	0.583	-1.09e+08	6.14e+07
x3	-5.913e+07	4.23e+07	-1.398	0.165	-1.43e+08	2.46e+07
x4	-1.375e+07	4.34e+07	-0.317	0.752	-9.96e+07	7.22e+07
x5	3.35e+07	4.2e+07	0.798	0.426	-4.96e+07	1.17e+08
x6	1.556e+07	3.95e+07	0.394	0.694	-6.27e+07	9.38e+07
x7	3.707e+07	3.86e+07	0.960	0.339	-3.94e+07	1.14e+08
x8	-7.069e+07	4.1e+07	-1.724	0.087	-1.52e+08	1.05e+07

Omnibus:	83.421	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	390.691
Skew:	2.431	Prob(JB):	1.45e-85
Kurtosis:	10.084	Cond. No.	1.85

Cluster 2 Summary:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.072
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	2.511
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.0121
Time:	04:49:11	Log-Likelihood:	-5754.6
No. Observations:	268	AIC:	1.153e+04
Df Residuals:	259	BIC:	1.156e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.791e+08	3.18e+07	15.062	0.000	4.16e+08	5.42e+08
x1	-3.514e+07	3.53e+07	-0.997	0.320	-1.05e+08	3.43e+07
x2	-1.037e+08	3.39e+07	-3.056	0.002	-1.7e+08	-3.69e+07
x3	3.238e+07	3.69e+07	0.879	0.380	-4.02e+07	1.05e+08
x4	-4.954e+07	3.5e+07	-1.418	0.158	-1.18e+08	1.93e+07
x5	-8.777e+07	3.28e+07	-2.675	0.008	-1.52e+08	-2.32e+07
x6	2.204e+07	3.24e+07	0.680	0.497	-4.18e+07	8.59e+07
x7	-2.615e+07	3.47e+07	-0.753	0.452	-9.45e+07	4.22e+07
x8	-1.448e+07	3.37e+07	-0.430	0.668	-8.09e+07	5.19e+07

Omnibus:	105.270	Durbin-Watson:	1.657
Prob(Omnibus):	0.000	Jarque-Bera (JB):	289.741
Skew:	1.823	Prob(JB):	1.21e-63
Kurtosis:	6.557	Cond. No.	1.94

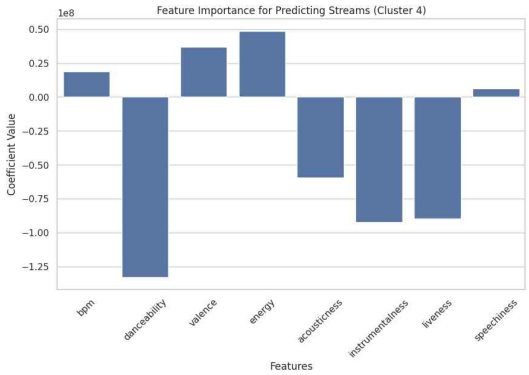
Cluster 3 Summary:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.064
Model:	OLS	Adj. R-squared:	0.015
Method:	Least Squares	F-statistic:	1.315
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.240
Time:	04:49:12	Log-Likelihood:	-3489.8
No. Observations:	162	AIC:	6998.
Df Residuals:	153	BIC:	7025.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.232e+08	4.44e+07	11.792	0.000	4.36e+08	6.11e+08
x1	1.803e+07	4.63e+07	0.390	0.697	-7.33e+07	1.09e+08
x2	-1.614e+07	4.82e+07	-0.335	0.738	-1.11e+08	7.92e+07
x3	-9.951e+07	4.79e+07	-2.077	0.039	-1.94e+08	-4.87e+06
x4	8.732e+07	5.06e+07	1.727	0.086	-1.26e+07	1.87e+08
x5	4.149e+06	5.07e+07	0.082	0.935	-9.61e+07	1.04e+08
x6	-4.735e+07	4.59e+07	-1.032	0.304	-1.38e+08	4.33e+07
x7	3.183e+07	4.67e+07	0.682	0.497	-6.04e+07	1.24e+08
x8	-5.706e+07	4.61e+07	-1.237	0.218	-1.48e+08	3.4e+07

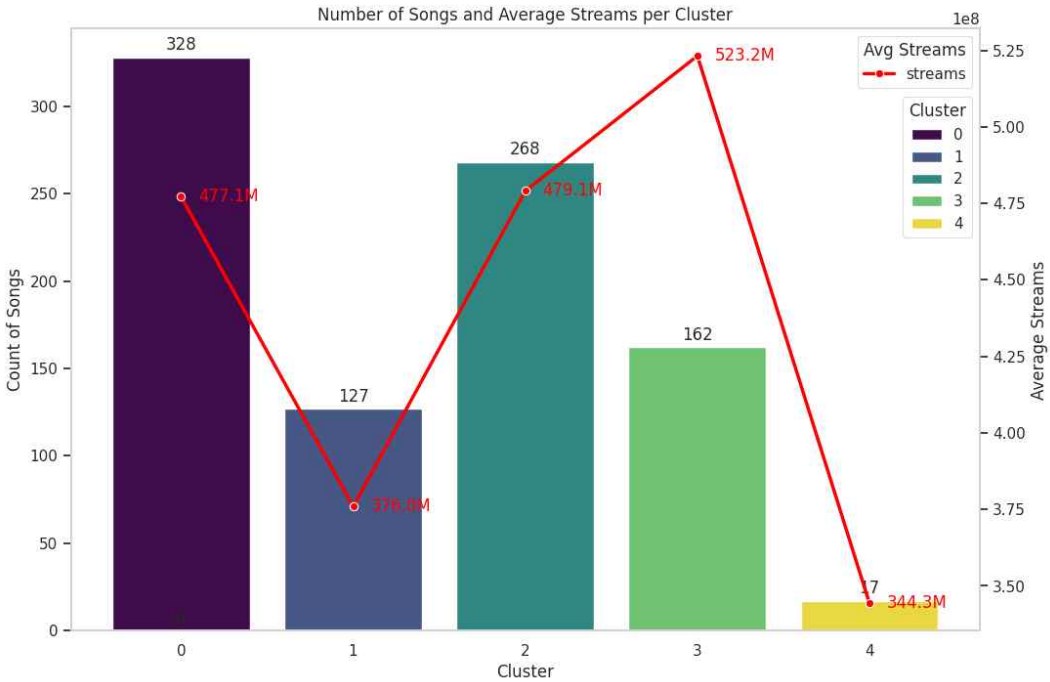
Omnibus:	55.363	Durbin-Watson:	1.697
Prob(Omnibus):	0.000	Jarque-Bera (JB):	107.879
Skew:	1.624	Prob(JB):	3.75e-24
Kurtosis:	5.330	Cond. No.	1.84




```
Cluster 4 Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.678
Model:                            OLS   Adj. R-squared:                  0.355
Method:                 Least Squares   F-statistic:                     2.102
Date:                Fri, 14 Jun 2024   Prob (F-statistic):              0.157
Time:                        04:49:13   Log-Likelihood:                -343.31
No. Observations:                  17   AIC:                             704.6
Df Residuals:                       8   BIC:                             712.1
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.443e+08   5.04e+07      6.827      0.000    2.28e+08    4.61e+08
x1          1.894e+07   1.04e+08      0.183      0.859    -2.2e+08    2.58e+08
x2         -1.329e+08   9.35e+07     -1.421      0.193   -3.48e+08    8.27e+07
x3          3.712e+07   7.64e+07      0.486      0.640   -1.39e+08    2.13e+08
x4          4.868e+07   1.22e+08      0.399      0.700   -2.32e+08     3.3e+08
x5         -5.956e+07   7.78e+07     -0.765      0.466   -2.39e+08     1.2e+08
x6         -9.232e+07    7.8e+07     -1.183      0.271   -2.72e+08    8.76e+07
x7         -8.985e+07   1.09e+08     -0.825      0.433   -3.41e+08    1.61e+08
x8          6.471e+06   9.69e+07      0.067      0.948   -2.17e+08     2.3e+08
==============================================================================
Omnibus:                        0.744   Durbin-Watson:                   1.595
Prob(Omnibus):                  0.689   Jarque-Bera (JB):                0.749
Skew:                          -0.367   Prob(JB):                        0.688
Kurtosis:                       2.279   Cond. No.                         6.42
==============================================================================
```

모든 장르에 대해서 클러스터별 분석한 결과, 음악의 요소들은 일부 영향이 있을 수 있지만, 스트리밍 수에 크게 유의미한 영향을 주지 않는다는 결과를 확인할 수 있었습니다.

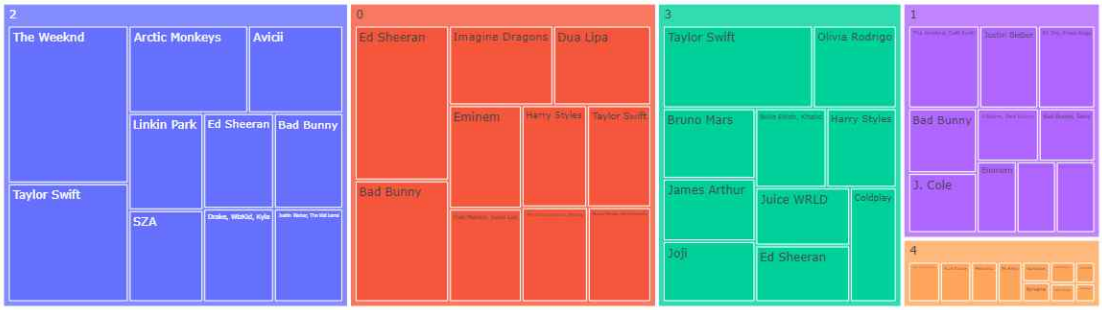
5. 클러스터별 곡 수 & 스트리밍 평균 분석



비교적 곡 수와 스트리밍 수가 비례하는 모습을 보이고 있습니다. 그러나 3번 클러스터의 경우, 어쿠스틱-인디 장르에서 포함된 곡 수에 비해 가장 높은 스트리밍 평균값을 나타내는 것을 확인하였습니다.

6. 아티스트별 스트리밍 수 분석

위 분석을 바탕으로 장르별로 어떤 아티스트가 인기 있는지 알아보려고 합니다. 따라서, 인기 있는 특정 아티스트 곡의 스트리밍이 높을 것이라는 가설을 세우고 분석을 진행해 보았습니다.



Tree map을 사용해 장르별로 아티스트 스트리밍 수를 나타내었습니다. 보이는 것과 같이 대중에게 인기 있는 유명 가수들이 다수 포진되어 있는 것을 알 수 있었습니다.

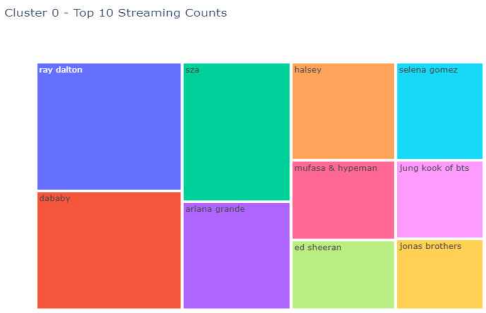
7. 아티스트 곡 수 & 스트리밍 수 분석

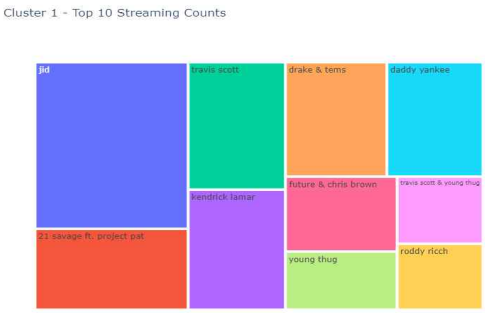
아티스트가 발매한 곡 수와 스트리밍 수에 대해 영향이 있을 것이라는 가설을 세우고 이에 대해 분석을 진행하였습니다.

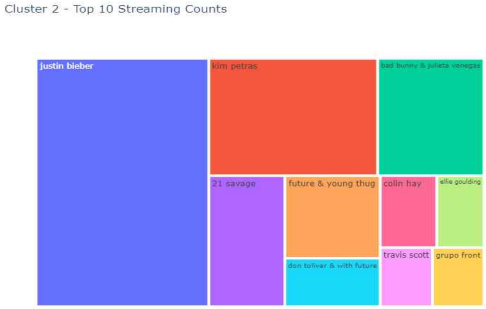
사각형의 크기가 스트리밍을 나타내고, 오른편의 범례의 색은 발매 곡 수를 나타냅니다. Top 10을 제외하고 발매곡 수는 큰 차이를 보이지 않았습니다. 그러나, 특정 아티스트의 스트리밍 수가 압도적으로 높은 것을 알 수 있었습니다. 따라서 곡 수가 스트리밍 수에 영향을 미친다고 판단할 수 있습니다.

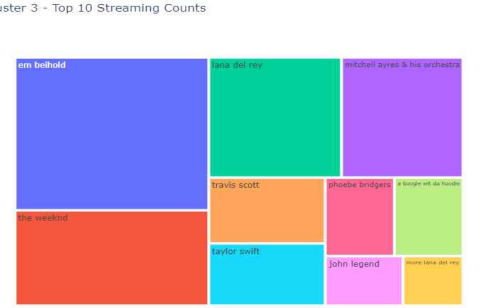
8. 피쳐링 & 스트리밍 수 분석

마지막으로 피쳐링이 스트리밍 수에 영향을 미치는지에 대한 분석을 진행하였습니다

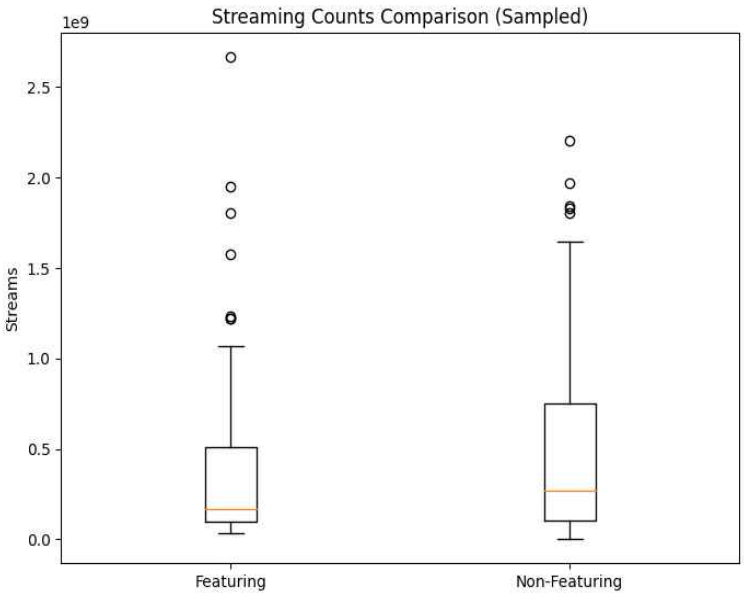








Tree map을 통한 시각화로 장르별로 피쳐링 아티스트가 참여한 곡의 스트리밍 Top10을 확인한 결과, 피처링 된 곡과 그렇지 않은 곡의 곡 수의 차이가 존재했습니다. 따라서, 논 피처링 곡들의 샘플을 통해 회귀 분석과 t-검정을 진행하였습니다. 참고로 클러스터 4번은 신뢰성의 문제로 제외되었습니다.



OLS Regression Results

Dep. Variable:	streams	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	4.060
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.0442
Time:	04:56:24	Log-Likelihood:	-19380.
No. Observations:	902	AIC:	3.876e+04
Df Residuals:	900	BIC:	3.877e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.809e+08	1.82e+07	26.391	0.000	4.45e+08	5.17e+08
is_featuring_int	-1.156e+08	5.74e+07	-2.015	0.044	-2.28e+08	-3.01e+06

Omnibus:	393.835	Durbin-Watson:	1.537
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1557.846
Skew:	2.105	Prob(JB):	0.00
Kurtosis:	7.871	Cond. No.	3.36

각각의 분석을 통해 피처링 여부가 스트리밍 수에 통계적으로 유의미한 영향을 미친다고 나타났습니다. 그러나, 모델의 설명력(R-제곱 값)이 매우 낮아 모델의 설명력과 신뢰성이 부족하다

고 판단됩니다.

9. 결론

K-means clustering 기법을 통한 장르 예측은 효과적으로 수행될 수 있습니다. 그리고 스트리밍은 음악의 요소 외에도 다양한 외부 요인에 의해 결정될 수 있음을 인지해야 하며, 장르별 곡 수와 스트리밍 수의 관계, 유명 아티스트의 영향, 피처링 여부의 등을 종합적으로 고려한다면 음악 산업에서의 성공 요인을 파악하는 데 기여할 수 있습니다.

4. 결론 및 미래 연구 방향

4.1 결론

주요 결론은 다음과 같습니다. 첫째, 인기 음악 요소 분석에서, 스트리밍 상위 100위 곡들은 일반적으로 bpm, danceability, valence, energy의 요소에서 특정 영역에 집중되는 경향을 보였습니다. 뿐만 아니라, 주로 major 키를 사용하는 경우가 많았습니다. 둘째, 월별/요일별 스트리밍 패턴 분석에서는 1월 중 금요일이 스트리밍 수와 재생목록 포함 수에서 특히 높은 값을 보여, 음악 발매일로 적합함을 시사했습니다. 또한 화요일, 목요일, 금요일에 발매 전 프로모션을 진행한다면, 스트리밍에 긍정적인 영향을 준다는 것을 알 수 있었습니다. 셋째, 장르별 인기 요소 분석의 경우 어쿠스틱-인디 장르가 상대적으로 높은 스트리밍 수를 보였습니다. 또, 피처링 여부가 결정적인 요인은 아니나, 피처링된 스트리밍 수가 더 높음을 알 수 있었습니다.

정리하자면, 어쿠스틱-인디 장르의 곡을 특정 영역에 집중된 음악 요소를 사용하여 1월의 발매 전 화요일과 목요일에 미리 프로모션을 진행한 뒤 금요일에 발매한다면 가장 큰 성공 가능성을 얻을 수 있을 것입니다.

4.2. 미래 연구 방향

1. 스트리밍 예측 모델 개발
본 연구에서 파악한 음악적 요소와 스트리밍 간의 관계를 기반으로 스트리밍 예측 모델을 개발하여 향후 발매될 음악의 성공 가능성을 예측할 수 있습니다.

2. 장르별 맞춤형 마케팅과 발매 전략 최적화
각 장르별로 분석된 인기 요소를 바탕으로 맞춤형 마케팅 전략을 수립할 수 있습니다. 또, 월별 및 요일별 분석 결과를 활용하여 최적의 음악 발매 시점을 선정함으로써 스트리밍 수를 극대화할 수 있습니다.

4. 피처링 전략 재고
피처링 여부가 스트리밍 수에 미치는 영향을 고려하여, 피처링을 통한 협업 전략을 재고할 필요가 있습니다. 피처링이 효과적인 장르와 그렇지 않은 장르를 구분하여 전략을 최적화할 수 있습니다.