Spotify Dataset Analysis

Data Science 1130: Final Project Report

Done by: Eshal Farrukh

Overall Research Topic: Spotify Tracks Engagement Analysis: The Influence of Popularity, Collaborations and Musical Attributes.

Subtopic 1 - Influence of Popular Tracks on Respective Albums

Background/Introduction of the problem:

The analysis presented investigates the influence popular tracks have on other songs within the same album. We hypothesize that there should be a positive correlation, since popular tracks tend to boost overall listener engagement of the album.

Data Analysis:

To answer the question, we manually inspected the first few hundred entries in the provided Spotify Dataset, potentially searching for missing values/errors. We concluded that the given dataset does not provide enough information so we loaded another dataset obtained from Kaggle. It contained information about track streams, chart ranks and playlists number. To avoid mingled results, we organised overlapping tracks found in both datasets and performed our analysis on them only.

First, for a track to be considered "popular", we created a popularity threshold which determines whether a track is popular or not (See Appendix A-1). We decided that if a song track has a popularity of **80<=**, it is considered popular on Spotify.

A problem we identified in the provided dataset were duplicate entries. It was found that these tracks only differed by popularity levels and track id; their musical attributes (danceability, energy, tempo etc) were the same. For example, "Winter Wonderland" by Jason Mraz and "You and Me on the Rock" by Brandi Carlile; Lucius are tracks which appeared over 5 times under different album names. We realised that these duplicates likely result from Spotify's "personalised" albums. To fix this problem, we decided to ignore the track id and average each duplicate track's popularity.

Next, we calculated stream numbers, chart ranks and playlists but another issue we faced were "singles" from artists (See Appendices A-2 to A-4). These tracks have no respective albums, such as "Hunger" by Ross Cooperman. Since these singles have the same popularity as the album and the track, we couldn't have included them in our results. Therefore, we decided to drop these singles.

There were some track names which did not exist; "é€æ˜Žã ã‡ãŸä¸–•Œ" by Motohiro Hata does not exist on Spotify. Therefore, we identified and processed non-existent tracks containing symbols/characters like the above as outliers and excluded them.

Main Results/Conclusions:

After cleaning and processing the data, we created histograms which tested the relationship of popular and less popular songs on the same album for streams, popularity, chart ranking, and playlists (See Appendixes A-5 to A-8). The results proved our hypothesis. From the graphs, it seems that highly popular tracks cause a dramatic increase in overall popularity, stream counts, chart ranks and playlists. We conclude popular tracks play a massive role in determining listener engagement of the whole album.

Drawbacks of the Analysis performed and Any Concerns:

1. *Random symbols and characters:* Many unusual track names which do not seem to appear on Spotify were listed in the given dataset. Since we can neither confirm nor deny their existence, we chose to ignore these entries since they are only a fraction in a vast dataset, and do not severely affect our results.
2. *Playlist Algorithms:* Album playlists based on popular tracks don't always reflect a track's playlist popularity. For example, Spotify tailors playlists to tastes, but tracks may also appear in genre or mood-based playlists, not just due to popularity.

Subtopic 2 - Solos vs Collaborations

Background/Introduction of the problem:

This part of the analysis examines the changes in musical attributes of solos and collaborations for an artist who does both. We tried to figure out which musical elements are maintained and which are changed in a collaboration, and hypothesized that the resulting collaborated track will show different musical attributes as each artist brings their individuality to it.

Data Analysis:

We saw in the dataset that collaborations haven't been differentiated from solos but rather had two or more artists separated with ";". An example of this is the track "Party of one" which has artists Brandi Carlile; Sam Smith. So we differentiated collaborations from solos as our first step. We also found there were tracks which did not have more than one artist, instead the collaborations were marked as "Feat. xyz", such as "In My Veins (Feat. Erin Mccarley)" by Andrew Belle. So, we included them in our analysis as well.

Afterwards, we decided to only test the relationship between artists who have collaborated and their solo performances in order to avoid mingled results (See B-1). For these artists, we first discovered their primary genre and calculated average musical attributes of all artists in the same genres for when they worked solo. For collaborations however, we calculated average musical attributes in the specific genre the track lies within and then compared the results.

An additional factor that severely affects our analysis are "remixes". These tracks do not qualify as collaborations since a third party combines/alters one or more tracks. To resolve this problem, we dropped these entries as they are unnecessary and alter our results.

We also identified entries which did not qualify as genres. For example, some genres were represented as numbers such as 0.043, 0.831, 120.044 (See B-2). We assumed that these numbers rather belonged in a musical attribute (tempo, time signature etc) and discounted them from our analysis.

Main Results/Conclusions:

Based on the track genre distribution of solos and collaborations (See B-3), we used a line graph to show the change in genre (See B-4). The two genres that change the most in this graph are classical and grindcore, respectively. In addition, all lines in the line graph have slopes >0 or <0, which means that most solo artists change their genre choice in collaborations. Next, we display the results for the changes in musical attributes between solos and collabs (See B-5). For instance, the danceability levels in collaborations are slightly higher than solos, whereas energy levels remain more or less the same. Additionally, in the comparison of six musical attributes, there were generally only subtle differences between solos and collaborations (See B-6). It seems that "loudness" was the only element which had a drastic difference. This concludes that collaborations do not typically result in higher/lower musical attributes, but rather depend on the genre.

Drawbacks of the Analysis performed and Any Concerns:

1. *Genre Misclassification:* The presence of unclassified genres (i.e. Brazil, British, Guitar etc) deeply affects our analysis by lengthening and mingling our results.
2. *Remixes:* Remixes are unofficial collaborations that typically result from a third party changing a tracks' musical attributes, usually adding and altering loudness and beats; they can also be formed by combining two or more tracks to create mashups. These remixes severely affect our results.

Subtopic 3 - Hit Tracks vs. Album Cohesion: Do Musical Attributes Align or Diverge?

Background/Introduction of the problem:

The question was, "When an album has one or more hit tracks, do they share similar musical attributes with other tracks, or do they significantly stand out?". This is a sub-topic of problem 1, and the purpose is to find out how many musical elements popular tracks in an album share with other tracks. We wanted to test the relationship between the musical attributes of highly popular tracks in an album with the unpopular tracks to determine whether particular musical attributes increase an artist's chances of success.

Data Analysis:

First, there are 15 musical attributes columns in the dataset which provide musical properties that each track represents. Among them, we selected 6 musical properties with the most diverse ranges that can easily distinguish the distribution differences between each property. These included danceability, energy, valence, tempo, liveness and loudness. The distribution of musical attributes of popular tracks and tracks other than those in the entire data is not comparing tracks within the same album, but comparing popular tracks with the rest of the data, which may distort our perspective (See Appendix C-1), so we had to filter the data. In order to use a clear sample group of this data, we added identifying information only to albums that included tracks with a popularity having a threshold of 80 or higher. After setting the sample, we analyzed the data by obtaining the distribution of musical attributes of popular tracks and the rest of the tracks within the same album.

Main Results/Conclusions:

We decided to use density overlap graphs as it would expectedly show a clearer view of the overlap between the attributes of highly popular tracks and the other tracks in an album. From the results (See Appendix C-2), we see that highly popular tracks have greater density in all attributes except "Liveness" than the latter; the most significant difference being energy density and loudness density. It seems popular tracks tend to have energy which usually is kept between 0.5 to 0.9 as popular tracks have an energy density of over 2.0 which is less spread. Meanwhile, lesser popular tracks have a higher spread and their energy exceeds to 1.0 but their density does not go beyond 1.6. This proves the fact that energy plays an important role in determining a tracks' popularity.

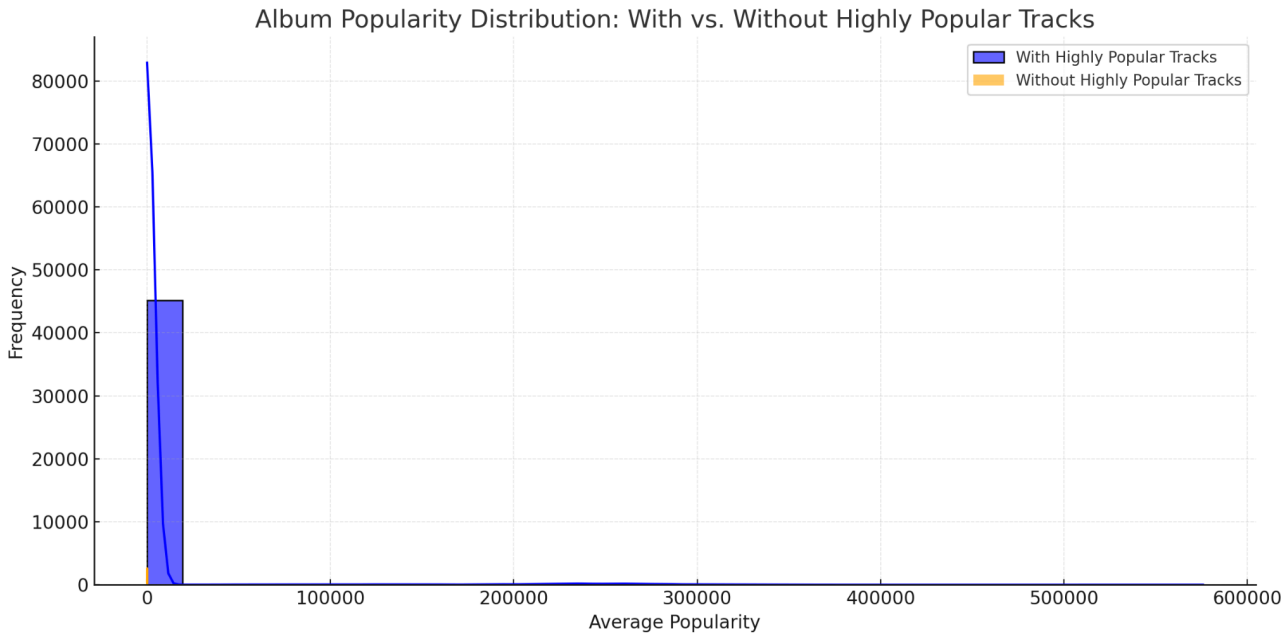
Another attribute to look into would be loudness. We see that a huge portion of highly popular tracks have a loudness which remains from 0 to -10. There is a drastic difference in loudness density between highly popular and lesser popular tracks which shows that loudness too plays a significant role in determining a track's success. We also notice "Valence" for popular tracks is higher and less spread than lesser known tracks, where for popular tracks it on average stays between 0.4 and 0.8. Whereas for less popular tracks, it remains between 0.1 to 0.6. Be that as it may, the other main attributes do not significantly affect the popularity of a track and more or less remain the same.

Resources:

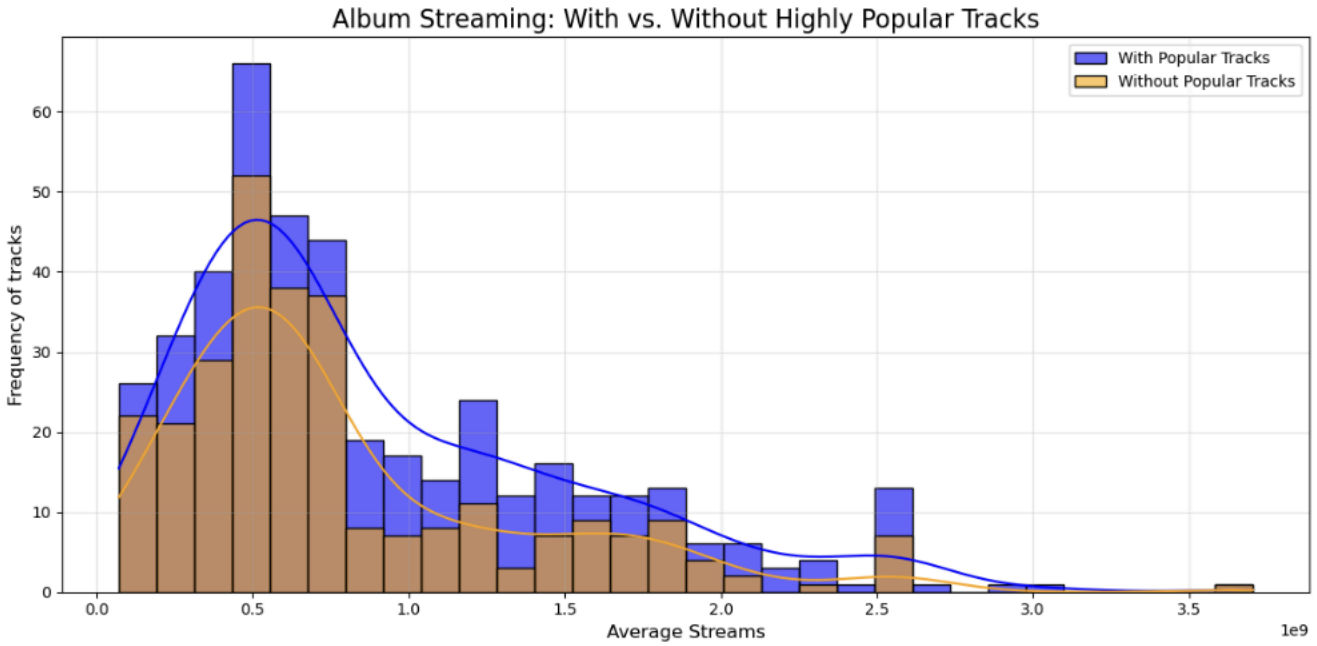
https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs

Appendix -

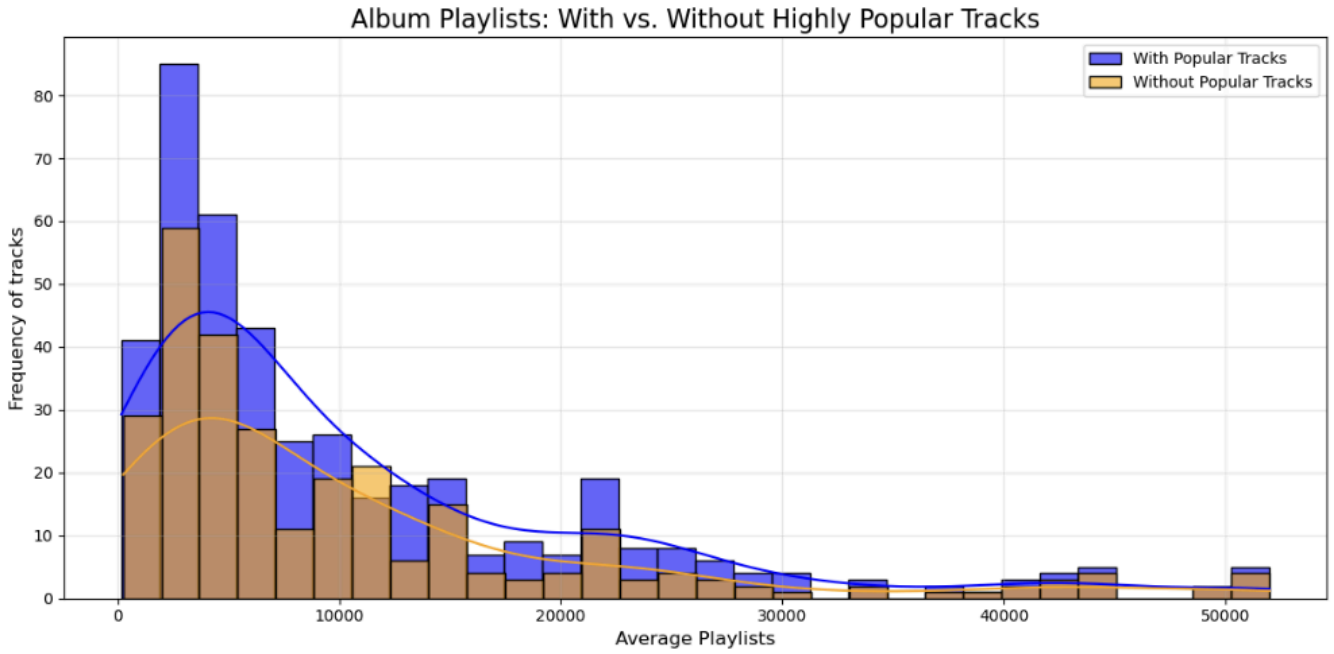
A-1



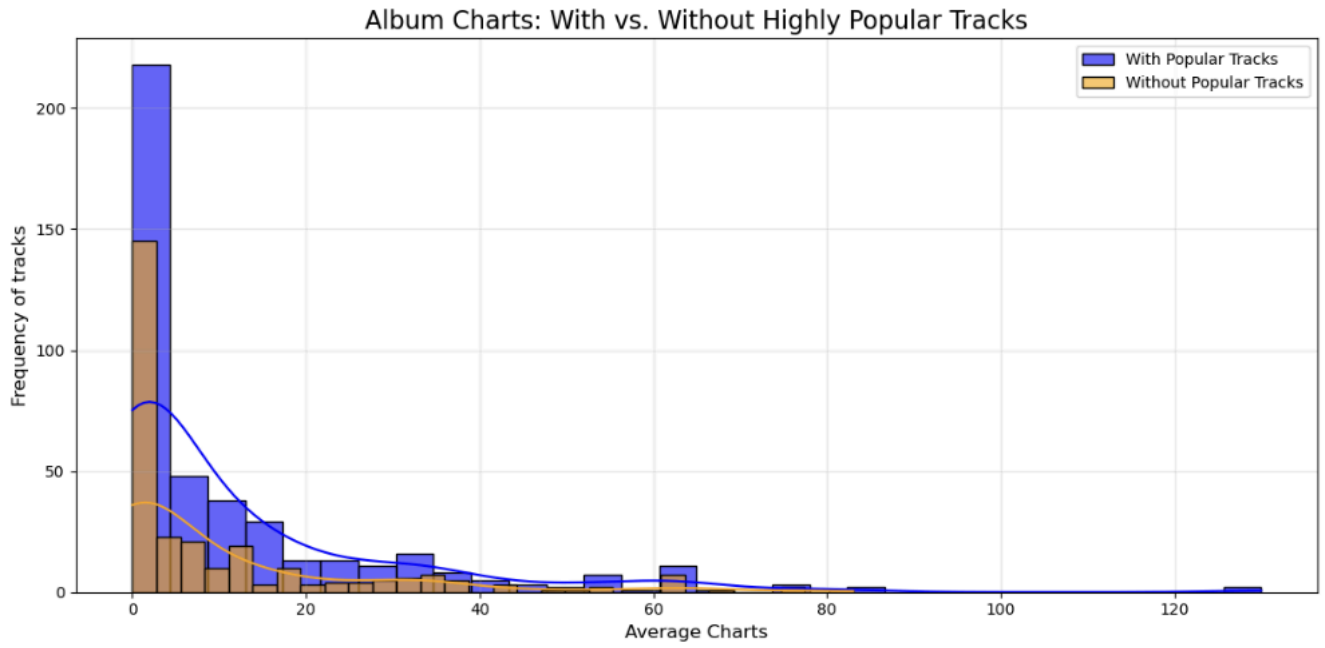
A-2



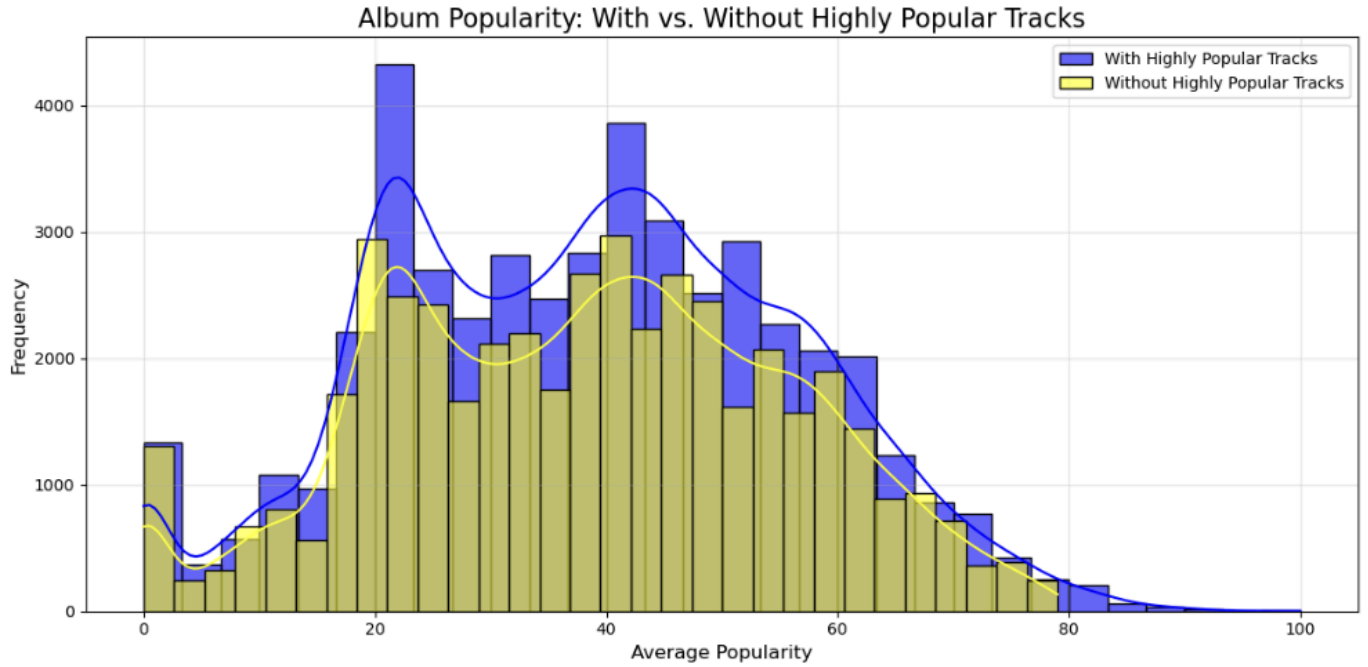
A-3


Album Playlists: With vs. Without Highly Popular Tracks
With Popular Tracks
Without Popular Tracks
Frequency of tracks
Average Playlists
0
10
20
30
40
50
60
70
80
0
10000
20000
30000
40000
50000

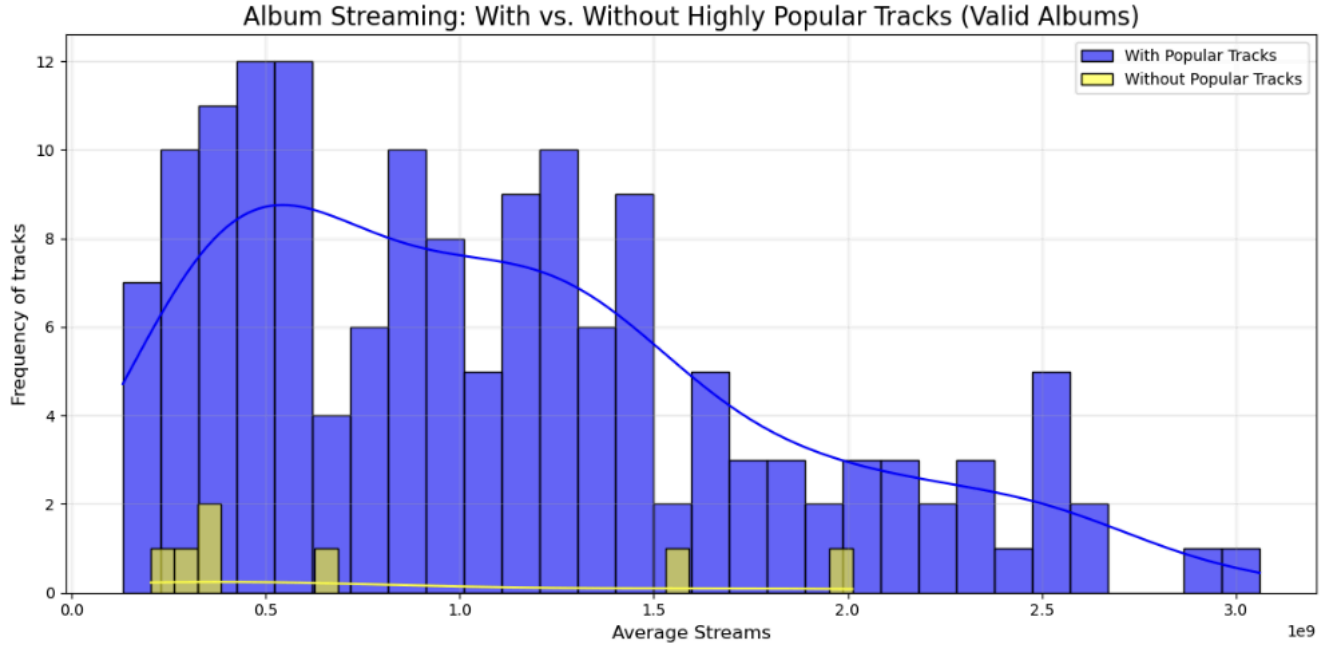
A-4


Album Charts: With vs. Without Highly Popular Tracks
With Popular Tracks
Without Popular Tracks
Frequency of tracks
Average Charts
0
50
100
150
200
0
20
40
60
80
100
120

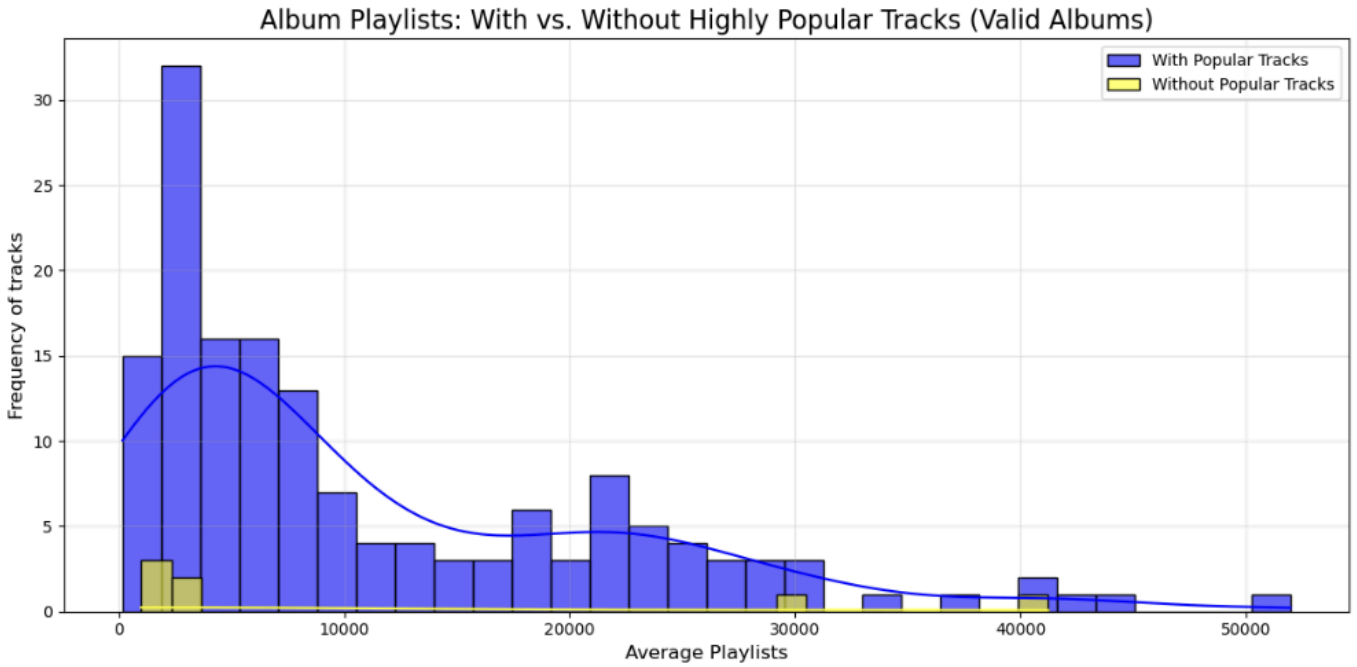
A-5


Album Popularity: With vs. Without Highly Popular Tracks
With Highly Popular Tracks
Without Highly Popular Tracks
Frequency
Average Popularity
0
1000
2000
3000
4000
0
20
40
60
80
100

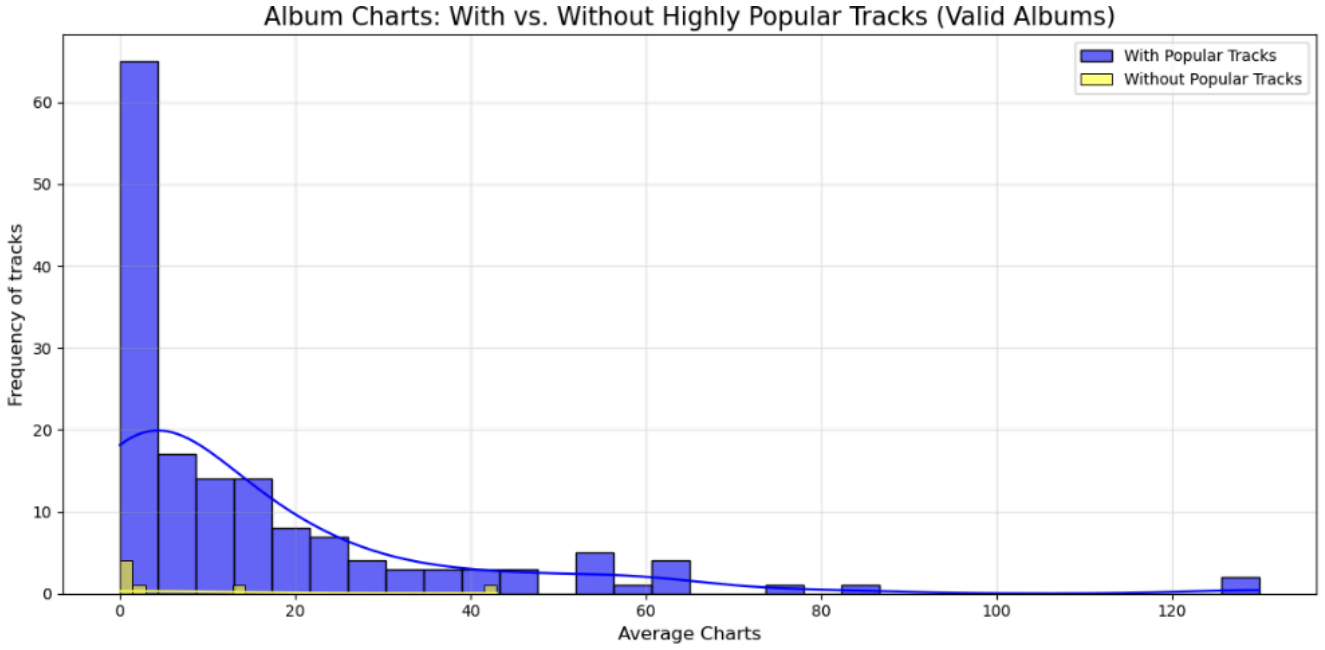
A-6


Album Streaming: With vs. Without Highly Popular Tracks (Valid Albums)
With Popular Tracks
Without Popular Tracks
Frequency of tracks
Average Streams
1e9
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
3.0

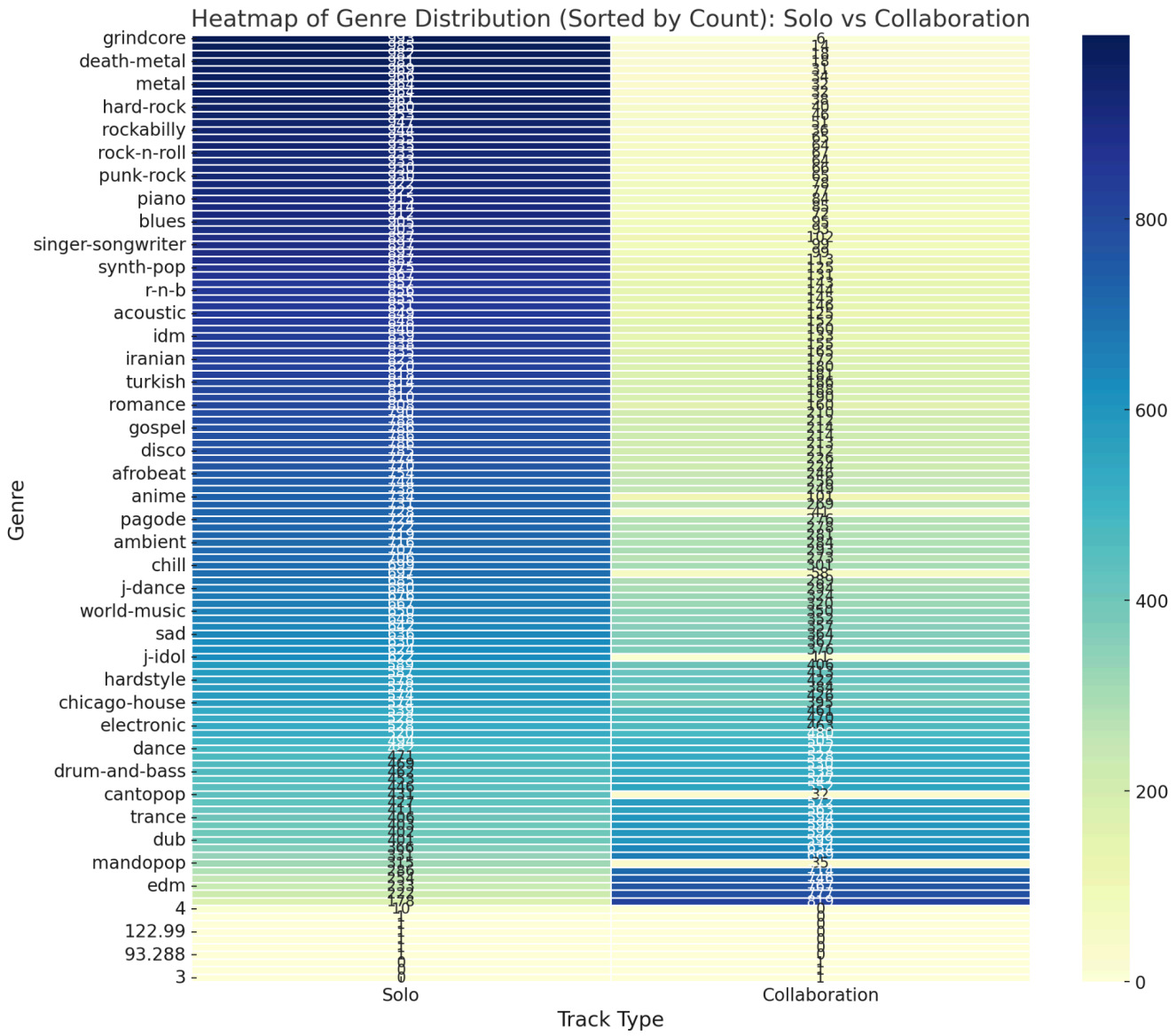
A-7


Album Playlists: With vs. Without Highly Popular Tracks (Valid Albums)
With Popular Tracks
Without Popular Tracks
Frequency of tracks
Average Playlists
0
5
10
15
20
25
30
0
10000
20000
30000
40000
50000

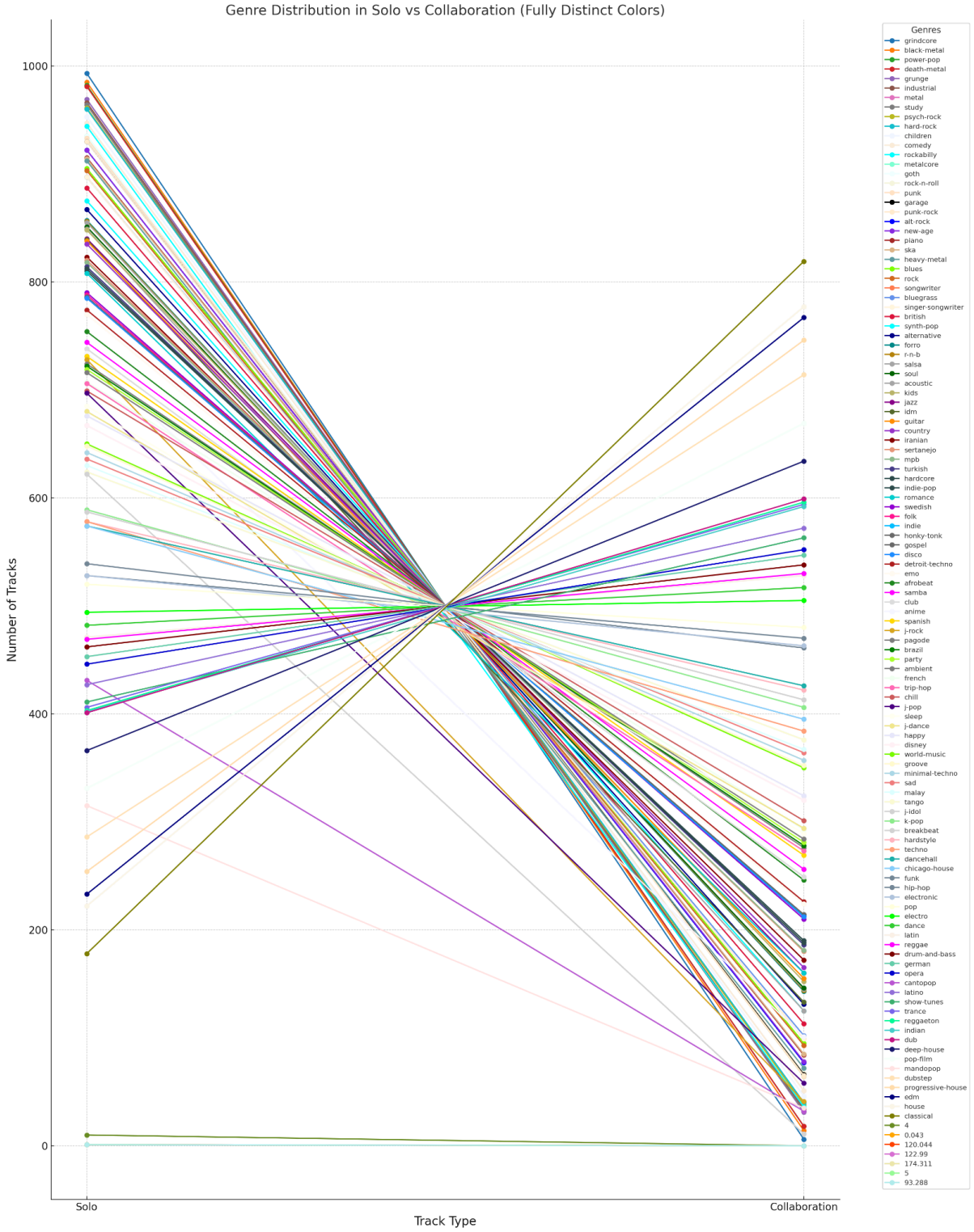
A-8


Album Charts: With vs. Without Highly Popular Tracks (Valid Albums)
With Popular Tracks
Without Popular Tracks
Frequency of tracks
Average Charts
0
10
20
30
40
50
60
0
20
40
60
80
100
120

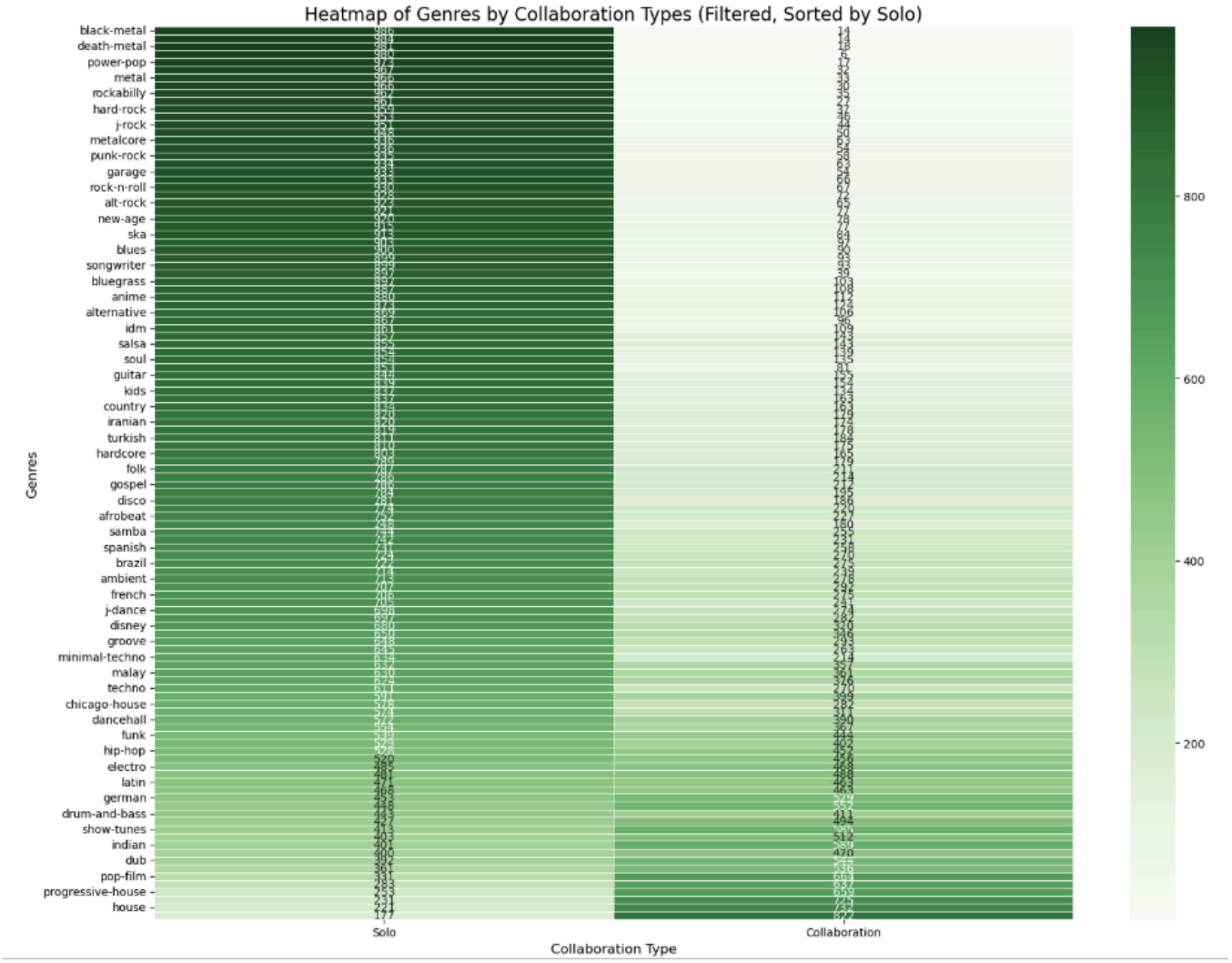
B-1


Heatmap of Genre Distribution (Sorted by Count): Solo vs Collaboration
Genre
grindcore
death-metal
metal
hard-rock
rockabilly
rock-n-roll
punk-rock
piano
blues
singer-songwriter
synth-pop
r-n-b
acoustic
idm
iranian
turkish
romance
gospel
disco
afrobeat
anime
pagode
ambient
chill
j-dance
world-music
sad
j-idol
hardstyle
chicago-house
electronic
dance
drum-and-bass
cantopop
trance
dub
mandopop
edm
4
122.99
93.288
3
Solo
Collaboration
Track Type
800
600
400
200
0

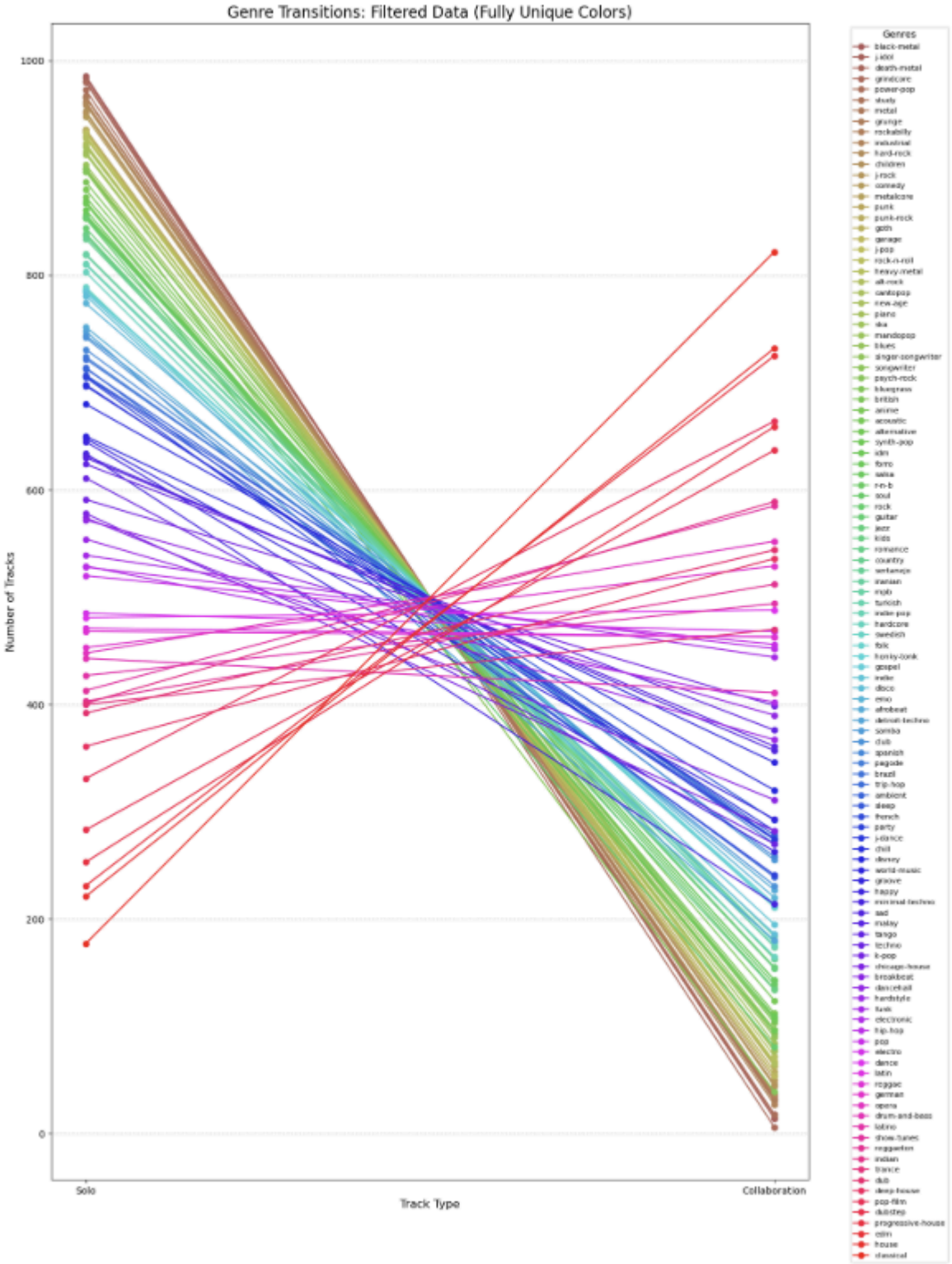
B-2


Genre Distribution in Solo vs Collaboration (Fully Distinct Colors)
Number of Tracks
Track Type
Solo
Collaboration
0
200
400
600
800
1000
Genres
grindcore
black-metal
power-pop
death-metal
grunge
industrial
metal
study
psych-rock
hard-rock
children
comedy
rockabilly
metalcore
goth
rock-n-roll
punk
garage
punk-rock
alt-rock
new-age
piano
ska
heavy-metal
blues
rock
songwriter
bluegrass
singer-songwriter
british
synth-pop
alternative
forro
r-n-b
salsa
soul
acoustic
kids
jazz
idm
guitar
country
iranian
sertanejo
mpb
turkish
hardcore
indie-pop
romance
swedish
folk
indie
honky-tonk
gospel
disco
detroit-techno
emo
afrobeat
samba
club
anime
spanish
j-rock
pagode
brazil
party
ambient
french
trip-hop
chill
j-pop
sleep
j-dance
happy
disney
world-music
groove
minimal-techno
sad
malay
tango
j-idol
k-pop
breakbeat
hardstyle
techno
dancehall
chicago-house
funk
hip-hop
electronic
pop
electro
dance
latin
reggae
drum-and-bass
german
opera
cantopop
latino
show-tunes
trance
reggaeton
indian
dub
deep-house
pop-film
mandopop
dubstep
progressive-house
edm
house
classical
4
0.043
120.044
122.99
174.311
5
93.288

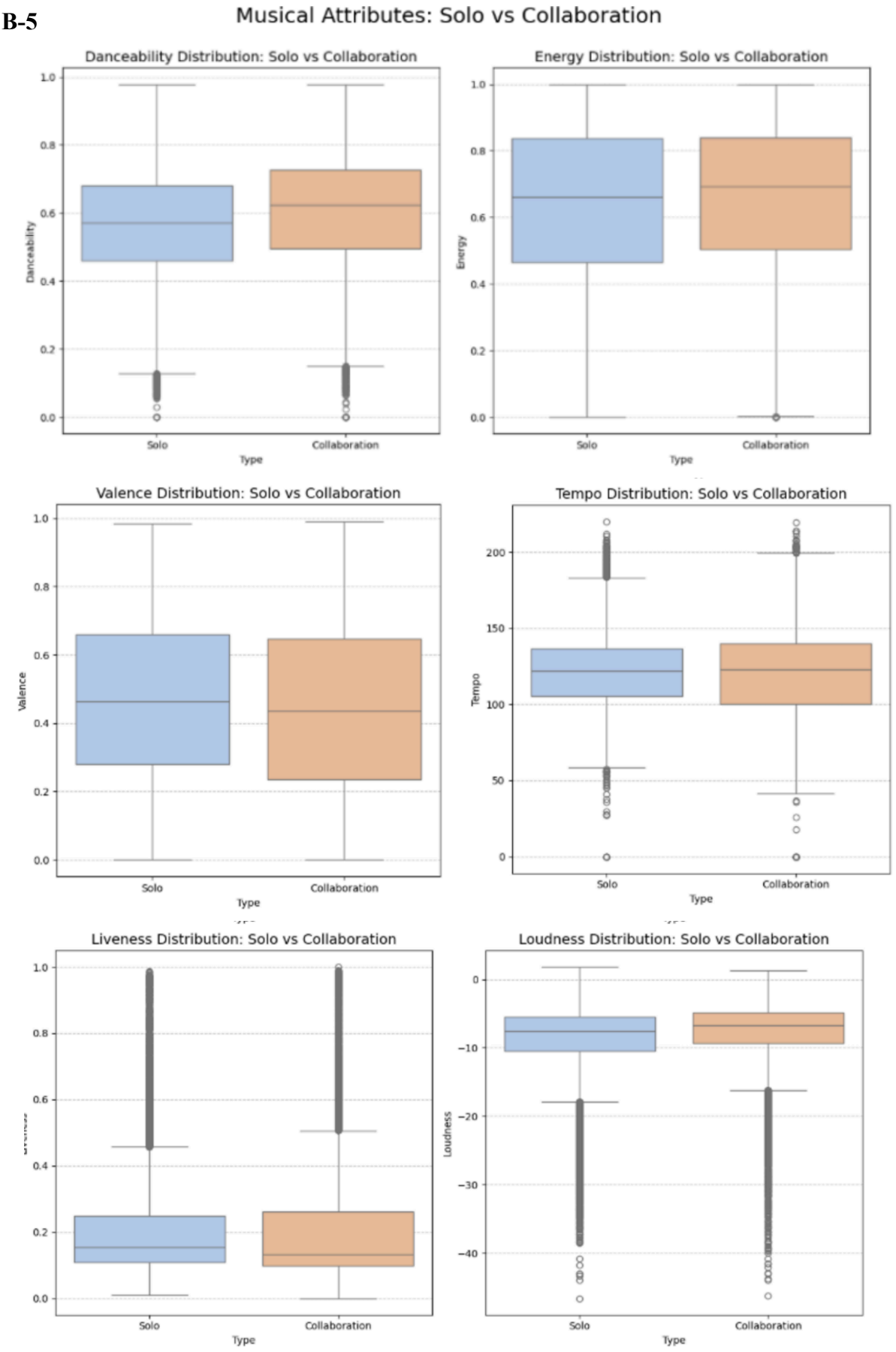
B-3


Heatmap of Genres by Collaboration Types (Filtered, Sorted by Solo)
Genres
black-metal
death-metal
power-pop
metal
rockabilly
hard-rock
j-rock
metalcore
punk-rock
garage
rock-n-roll
alt-rock
new-age
ska
blues
songwriter
bluegrass
anime
alternative
idm
salsa
soul
guitar
kids
country
iranian
turkish
hardcore
folk
gospel
disco
afrobeat
samba
spanish
brazil
ambient
french
j-dance
disney
groove
minimal-techno
malay
techno
chicago-house
dancehall
funk
hip-hop
electro
latin
german
drum-and-bass
show-tunes
indian
dub
pop-film
progressive-house
house
Solo
Collaboration
Collaboration Type
800
600
400
200

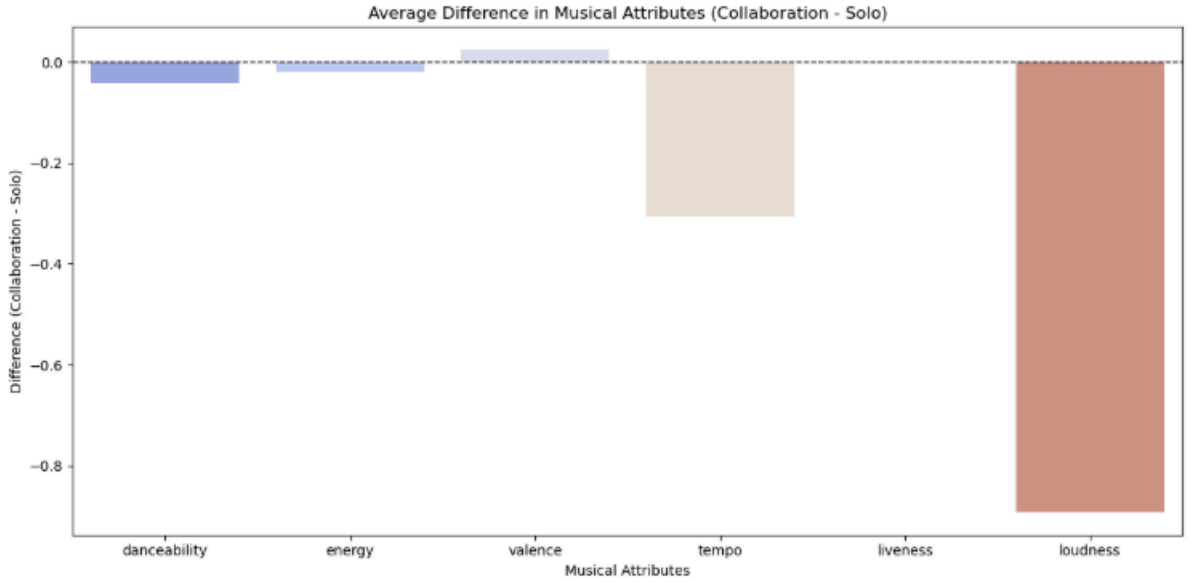
B-4


Genre Transitions: Filtered Data (Fully Unique Colors)
Number of Tracks
Track Type
Solo
Collaboration
0
200
400
600
800
1000
Genres
black-metal
j-idol
death-metal
grindcore
power-pop
study
metal
grunge
rockabilly
industrial
hard-rock
children
j-rock
comedy
metalcore
punk
punk-rock
goth
garage
j-pop
rock-n-roll
heavy-metal
alt-rock
cantopop
new-age
piano
ska
mandopop
blues
singer-songwriter
songwriter
psych-rock
bluegrass
british
anime
acoustic
alternative
synth-pop
idm
forro
salsa
r-n-b
soul
rock
guitar
jazz
kids
romance
country
sertanejo
iranian
mpb
turkish
indie-pop
hardcore
swedish
folk
honky-tonk
gospel
indie
disco
emo
afrobeat
detroit-techno
samba
club
spanish
pagode
brazil
trip-hop
ambient
sleep
french
party
j-dance
chill
disney
world-music
groove
happy
minimal-techno
sad
malay
tango
techno
k-pop
chicago-house
breakbeat
dancehall
hardstyle
funk
electronic
hip-hop
pop
electro
dance
latin
reggae
german
opera
drum-and-bass
latino
show-tunes
reggaeton
indian
trance
dub
deep-house
pop-film
dubstep
progressive-house
edm
house
classical

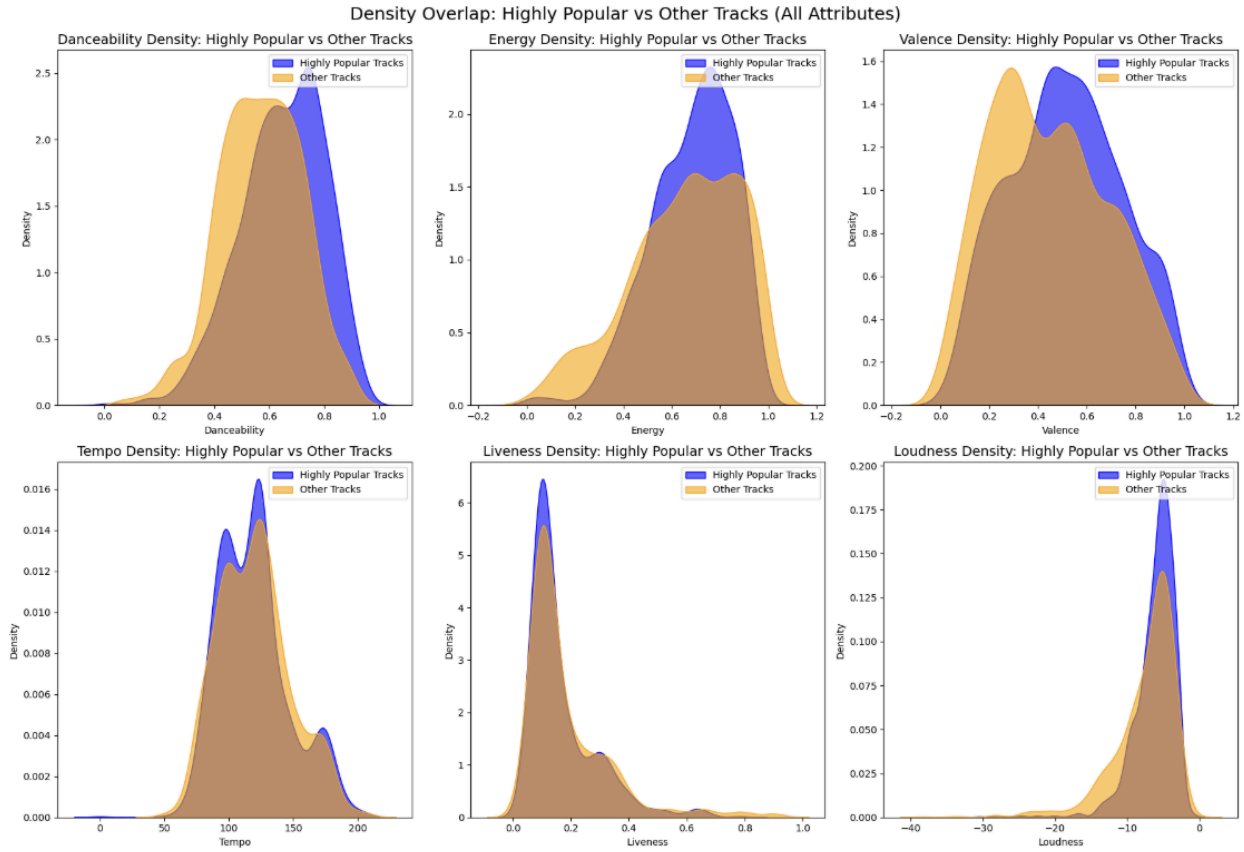
B-5


Musical Attributes: Solo vs Collaboration
Danceability Distribution: Solo vs Collaboration
Danceability
Solo
Collaboration
Type
Energy Distribution: Solo vs Collaboration
Energy
Solo
Collaboration
Type
Valence Distribution: Solo vs Collaboration
Valence
Solo
Collaboration
Type
Tempo Distribution: Solo vs Collaboration
Tempo
Solo
Collaboration
Type
Liveness Distribution: Solo vs Collaboration
Solo
Collaboration
Type
Loudness Distribution: Solo vs Collaboration
Loudness
Solo
Collaboration
Type

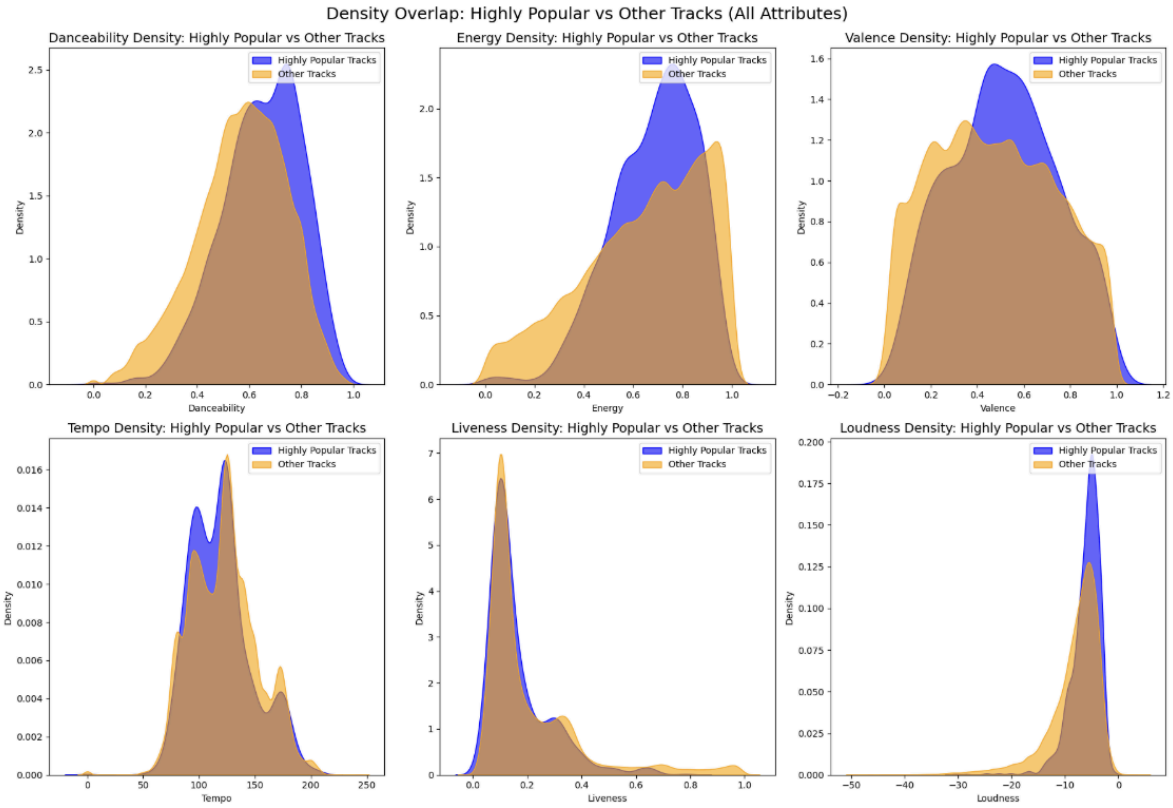
B-6


Average Difference in Musical Attributes (Collaboration - Solo)
Difference (Collaboration - Solo)
0.0
−0.2
−0.4
−0.6
−0.8
danceability
energy
valence
tempo
liveness
loudness
Musical Attributes

C-1


Density Overlap: Highly Popular vs Other Tracks (All Attributes)
Danceability Density: Highly Popular vs Other Tracks
Energy Density: Highly Popular vs Other Tracks
Valence Density: Highly Popular vs Other Tracks
Tempo Density: Highly Popular vs Other Tracks
Liveness Density: Highly Popular vs Other Tracks
Loudness Density: Highly Popular vs Other Tracks
Highly Popular Tracks
Other Tracks
Density
Danceability
Energy
Valence
Tempo
Liveness
Loudness

C-2


Density Overlap: Highly Popular vs Other Tracks (All Attributes)
Danceability Density: Highly Popular vs Other Tracks
Energy Density: Highly Popular vs Other Tracks
Valence Density: Highly Popular vs Other Tracks
Tempo Density: Highly Popular vs Other Tracks
Liveness Density: Highly Popular vs Other Tracks
Loudness Density: Highly Popular vs Other Tracks
Highly Popular Tracks
Other Tracks
Density
Danceability
Energy
Valence
Tempo
Liveness
Loudness