Analysis of Spotify Dataset

Akai Kaeru
Professor Klaus Mueller
June 2022 ~ September 2022
JeongYoon Lee

Introduction

In this project I tried to find some interesting insights about songs in Spotify which is the biggest music streaming service through the AK Analyst, and make sure to check the software with making some test cases.

Experimental Setup

I used google colab to use python for this project and used AK Analyst for data preprocessing and analyzing.

Extract Dataset

1. Log in to Spotify for Developer (https://developer.spotify.com/dashboard/login).
2. Push the "Create an App" Button and get the Client ID and Client Secret.
3. We can approach with these ID and Password in python.

```
[ ]  !pip install spotipy
     !pip install urllib3 --upgrade
     !pip install requests --upgrade
     !pip install spotipy --upgrade
```

```
import spotipy
from spotipy.oauth2 import SpotifyClientCredentials

sp = spotipy.Spotify(auth_manager=SpotifyClientCredentials(client_id="YOUR_SPOTIFY_ID",
                                                           client_secret="YOUR_SPOTIFY_PW"))
```

```
artist_name =[]
track_name = []
track_popularity =[]
artist_id =[]
track_id =[]
album_id = []
release_date = []
release_date_precision = []
duration_ms = []
artist_genre = []
artist_popularity = []

for i in range(0,1000,50):
    track_results_2021 = sp.search(q='year:2021', type='track', limit=50, offset=i)
    for i, t in enumerate(track_results_2021['tracks']['items']):
        artist_name.append(t['artists'][0]['name'])
        artist_id.append(t['artists'][0]['id'])
        art = sp.artist(t['artists'][0]['id'])
        artist_genre.append(art['genres'])
        artist_popularity.append(art['popularity'])
        track_name.append(t['name'])
        track_id.append(t['id'])
        album_id.append(t['album']['id'])
        duration_ms.append(t['duration_ms'])
        track_popularity.append(t['popularity'])
        release_date.append(t['album']['release_date'])
        release_date_precision.append(t['album']['release_date_precision'])
```

>>Extract released tracks (1000)

```python
import pandas as pd
track_df_2021 = pd.DataFrame({'artist_name' : artist_name, 'track_name' : track_name, 'track_id' : track_id, 'album_id' : album_id, 'artist_id' : artist_id ,'track_popularity' : track_popularity,'artist_popularity':artist_popularity, 'artist_genre': artist
print(track_df_2021.shape)
track_df_2021
```

(1000, 11)

	artist_name	track_name	track_id	album_id	artist_id	track_popularity	artist_popularity	artist_genre	duration_ms	release_date	release_date_precision
0	Doja Cat	Woman	6Uj1ctrBOjOas8xZXGqKk4	1nAQbHeOWTfQzbOoFrvndW	5cj0lLjcoR7YOSnhnX0Po5	92	89	[dance pop, pop]	172626	2021-06-25	day
1	Otmar Eros	Year 2016:, Pt. 4	1sUxW2zAmXA7lXHC8Dxu4s	21AA25PUsBnGgzLmiCFhX6	5XY9JN9PcQ41KQTZqtyhsL	26	22	[]	69839	2021-03-09	day
2	Morgan Wallen	Wasted On You	3cBsEDNhFI9E82vPj3kvi3	6JlCkqkqobGirPsaleJpFr	4oUHIQIBe0LHzYfvXNW4QM	84	83	[contemporary country]	178520	2021-01-08	day
3	Elvis Costello & The Attractions	Pump It Up - 2021 Remaster	3oyc1mIdCBGaU55wX7otqM	4RLIesiAVONV4fOUlOSmr4	4qmHkMxr6pTWh5Zo74odpH	63	54	[art rock, folk rock, mellow gold, new wave po...	196680	1978-03-17	day
4	Olivia Rodrigo	good 4 u	4ZtFanR9U6ndgddUvNcjcG	6s84u2TUpR3wdUv4NgKA2j	1McMsnEElThX1knmY4oliG	91	85	[pop]	178146	2021-05-21	day
...	...	...	...	...	...	...	...	...	...	...	...
995	Above & Beyond	Thing Called Love - Oliver Heldens Remix (Mixed)	7y6idgWHgPB68mJg37pRJP	6K4BHdXN6slKvUDKTED2vP	10gzBoINW3cLJfZUka8Zoe	8	61	[edm, pop dance, progressive house, progressiv...	360000	2021-12-15	day

>>Extract Audio features

```python
track_features = []
for t_id in track_df_2021['track_id']:
    af = sp.audio_features(t_id)
    track_features.append(af)
tf_df_2021 = pd.DataFrame(columns = ['danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'type', 'id', 'uri', 'track_href', 'analysis_url', 'duration_ms', 'time_signature']
for item in track_features:
    for feat in item:
        tf_df_2021 = tf_df_2021.append(feat, ignore_index=True)
tf_df_2021.head()
```

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	type	id	uri	track_
0	0.824	0.764	5	-4.175	0	0.0854	0.088800	0.002940	0.1170	0.881	107.998	audio_features	6Uj1ctrBOjOas8xZXGqKk4	spotify:track:6Uj1ctrBOjOas8xZXGqKk4	https://api.spotify.com/v1/tracks/6Uj1ctrBOj
1	0.608	0.841	9	-8.354	1	0.0293	0.000672	0.000017	0.0704	0.185	129.994	audio_features	4S4ZY1yKo3WUtzsg3O6hcf	spotify:track:4S4ZY1yKo3WUtzsg3O6hcf	https://api.spotify.com/v1/tracks/4S4ZY1yKo3\
2	0.505	0.657	11	-5.240	0	0.0318	0.373000	0.001070	0.1260	0.252	196.000	audio_features	3cBsEDNhFI9E82vPj3kvi3	spotify:track:3cBsEDNhFI9E82vPj3kvi3	https://api.spotify.com/v1/tracks/3cBsEDNhF
3	0.645	0.809	11	-6.120	1	0.0385	0.009210	0.001080	0.1060	0.966	138.978	audio_features	3oyc1mIdCBGaU55wX7otqM	spotify:track:3oyc1mIdCBGaU55wX7otqM	https://api.spotify.com/v1/tracks/3oyc1mIdCE
4	0.563	0.664	9	-5.044	1	0.1540	0.335000	0.000000	0.0849	0.688	166.928	audio_features	4ZtFanR9U6ndgddUvNcjcG	spotify:track:4ZtFanR9U6ndgddUvNcjcG	https://api.spotify.com/v1/tracks/4ZtFanR9U6

Merge genres

→ Since there are too many genres, I tried to group them into a small number of genres

```python
total_idx = 0
idx = 0

for i_list in grouping_genre["artist_genre"]:
  for j in i_list:
    if "korean pop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "k-pop"
    elif "korean electropop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "k-pop"
    elif "k-pop girl group" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "k-pop"
    elif "k-pop boy group" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "k-pop"
    elif "britpop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "uk-pop"
    elif "classic uk pop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "uk-pop"
    elif "british alternative rock" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "uk-pop"
    elif "pop rock" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "rock"
    elif "alternative pop rock" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "rock"
    elif "country pop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "country"
    elif "latin pop" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "latin"
    elif "pop r&b" in j:
      grouping_genre['artist_genre'][total_idx][idx] = "r&b"
```

[Code for grouping genres]

freq_new ×

```
Genre,Freq
pop,1764
edm,1041
rap,1025
etc,939
hip hop,654
rock,593
,480
folk,434
r&b,232
indie,197
latin,127
blues,101
dance,93
techno,84
k-pop,75
reggae,73
metal,72
funk,60
punk,58
classic,50
uk-pop,43
jazz,40
lo-fi,36
country,28
band,27
lo-fi beats,26
singer-songwriter,24
soundtrack,20
```

[The number of tracks per each genres]

Final Input Dataset Overview

artist_name	track_name	track_id	album_id	artist_id	track_popularity	artist_popularity	artist_genre	duration_ms	release_date	release_date_precis
Steve Lacy	Dark Red	37y7iDayfwm3WXn5BiAoRk	5fvUFzgVEni3L7769OabqQ	57vWImR43h4CaDao012Ofp	87	76	['etc', 'pop']	173104	2017-02-20	day
Otmar Eros	Year 2016:, Pt. 4	1sUxW2zAmXA7IXHC8Dxu4s	21AA25PUsBnGgzLmiCFhX6	5XY9JN9PcQ41KQTZqtyhsL	26	22	['']	69839	2021-03-09	day
Tyler, The Creator	See You Again (feat. Kali Uchis)	7KA4W4McWYRpgf0fWsJZWB	2nkto6YNI4rUYTLqEwWJ3o	4V8LLVI7PbaPR0K2TGSxFF	85	84	['hip hop', 'rap']	180386	2017-07-21	day
Mark Ronson	Uptown Funk	4rmFRTmHa2bWUmMLIRVEXQ	6ndaa5yzks3YifHX1u5Esl	3hv9jJF3adDNsBSIQDqcjp	49	74	['pop']	269666	2017-12-22	day
Lil Uzi Vert	20 Min	0uxSUdBrJy9Un0EYoBowng	0zicd2mBV8HTzSubByj4vP	4O15NlyKLIASxsJ0PrXPfz	84	84	['rap']	220586	2017-11-17	day
Dark.D	Year 2017	4PNPLAWzF3qCF2Z8Xur1pc	7jX3f3rXHx5PWSIXL48U64	13fEC4mCM6Ddu07ydQRcRq	0	0	['']	348754	2017-07-05	day
Ruth B.	Dandelions	2eAvDnpXP5W0cVtil0PUxV	6FgtuX3PtiB5civjHYhc52	2WzaAvm2bBCf4pEhyuDgCY	91	75	['r&b', 'pop']	233720	2017-05-05	day
Otmar Eros	Year 2016:, Pt. 1	48Mk9nvXiHZ9eSDqkX8sSZ	21AA25PUsBnGgzLmiCFhX6	5XY9JN9PcQ41KQTZqtyhsL	22	22	['']	59443	2021-03-09	day
A Boogie Wit da Hoodie	Drowning (feat. Kodak Black)	1f5cbQtDrykjarZVrShaDI	3HHp5I6Q6SEyU5bkvoCtnV	31W5EY0aAly4Qieq6OFu6I	81	80	['rap']	209269	2017-09-29	day
Ten Years After	50,000 Miles Beneath My Brain - 2017 Remaster	0CflMLPy8lYSdgjuoLTMtq	1WQORrTyf78zuJCBziHfQg	7nkLRaWHImCvWGHdNGnhVE	41	52	['rock', 'blues', 'folk']	457449	1970-04-01	day
XXXTENTACION	Revenge	5TXDeTFVRVY7Cvt0Dw4vWW	5VdyJkLe3yvOs0l4xXbWp0	15UsOTVnJzReFVN1VCnxy4	87	87	['hip hop', 'rap']	120026	2017-08-25	day
Schoolgirl Byebye	Year,2015	0UsmyJDsst2xhX1ZiFF3JW	5gWxh24iphqQ8WDh8MBMfe	6kfcndVsu8F9Y5gL5xc717	24	34	['pop', 'etc', 'indie']	74301	2020-09-16	day
Drake	Passionfruit	5mCPDVBb16L4XQwDdbRUpz	1IXY618HWkwYKJWBRYR4MK	3TVXtAsR1Inumwj472S9r4	84	95	['hip hop', 'pop', 'rap']	298940	2017-03-18	day
Billy Joel	Miami 2017 (Seen the Lights Go Out On Broadway)	5Bgs8sHxL7zbNMyEAiSkMq	4nFLLh5qSlp2z2FuLpVERX	6zFYqv1mOsgBRQbae3JJ9e	33	75	['rock', 'folk', 'singer-songwriter']	314620	2022-04-08	day

[new_with_genre_final.csv]

danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	type	id	uri	
0.603	0.783	6	-4.023	1	0.062	0.449	7.96E-06	0.119	0.775	172.041	audio_features	37y7iDayfwm3WXn5BiAoRk	spotify:track:37y7iDayfwm3WXn5BiAoRk	https://api.spotify.
0.241	0.064	5	-24.272	1	0.0602	0.994	0.95	0.0715	0.0372	141.739	audio_features	1sUxW2zAmXA7IXHC8Dxu4s	spotify:track:1sUxW2zAmXA7IXHC8Dxu4s	https://api.spotify.
0.558	0.559	6	-9.222	1	0.0959	0.371	7.49E-06	0.109	0.62	78.558	audio_features	7KA4W4McWYRpgf0fWsJZWB	spotify:track:7KA4W4McWYRpgf0fWsJZWB	https://api.spotify.
0.856	0.609	0	-7.223	1	0.0824	0.00801	8.15E-05	0.0344	0.928	114.988	audio_features	4rmFRTmHa2bWUmMLIRVEXQ	spotify:track:4rmFRTmHa2bWUmMLIRVEXQ	https://api.spotify.
0.773	0.75	8	-4.009	0	0.117	0.109	0	0.174	0.783	123.426	audio_features	0uxSUdBrJy9Un0EYoBowng	spotify:track:0uxSUdBrJy9Un0EYoBowng	https://api.spotify.
0.885	0.494	3	-8.004	0	0.0565	0.00107	0.664	0.0514	0.432	128.005	audio_features	4PNPLAWzF3qCF2Z8Xur1pc	spotify:track:4PNPLAWzF3qCF2Z8Xur1pc	https://api.spotify.
0.609	0.692	1	-2.958	1	0.0259	0.0157	0	0.0864	0.454	116.959	audio_features	2eAvDnpXP5W0cVtil0PUxV	spotify:track:2eAvDnpXP5W0cVtil0PUxV	https://api.spotify.
0.314	0.0855	9	-15.775	1	0.0342	0.969	0.795	0.16	0.161	69.893	audio_features	0UsmyJDsst2xhX1ZiFF3JW	spotify:track:0UsmyJDsst2xhX1ZiFF3JW	https://api.spotify.
0.839	0.81	5	-5.274	0	0.0568	0.501	0	0.117	0.814	129.014	audio_features	1f5cbQtDrykjarZVrShaDI	spotify:track:1f5cbQtDrykjarZVrShaDI	https://api.spotify.
0.344	0.83	9	-7.67	1	0.0569	0.0133	0.036	0.101	0.435	116.883	audio_features	0CflMLPy8lYSdgjuoLTMtq	spotify:track:0CflMLPy8lYSdgjuoLTMtq	https://api.spotify.

[final_audio_feature.csv]

Final Input File Feature Details

1) **new_with_genre_final.csv**

   artist_name : The name of artist
   track_name : The name of track (music)
   track_id : Track's id number (*unique value)
   album_id : Album's id number
   track_popularity : Popularity of track
   artist_popularity :Popularity of artist
   artist_genre : A list of the genres the artist is associated with.
   duration_ms : The track length in milliseconds.
   release_date : The date the album was first released.
   release_date_precision : The precision with which "realese_date" value is known.

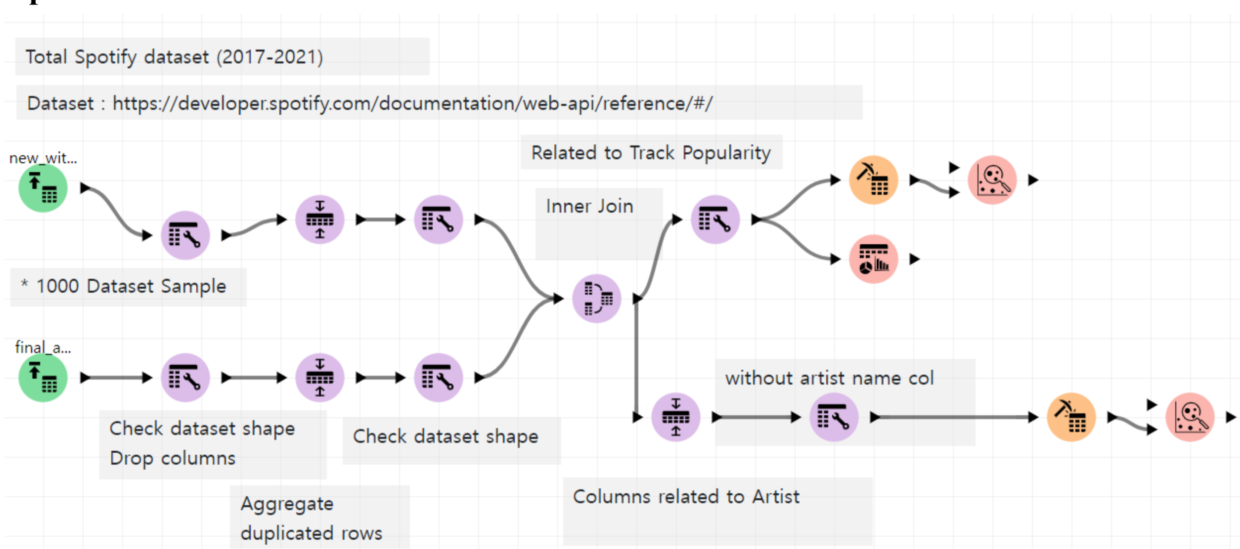
2) **final_audio_feature.csv**

   danceability : Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.

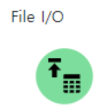
energy : Represents a perceptual measure of intensity and activity.
key : The key the track is in.
loundness : The overall loudness of a track in decibels (dB).
mode : Mode indicates the modality (major or minor) of a track.
speechiness : Detects the presence of spoken words in a track.
acousticness : A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
instrumentalness : Predicts whether a track contains no vocals.
liveness : Detects the presence of an audience in the recording
valence : A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. 1.0 is the most positive, and 0.0 is the most negative.
tempo : The overall estimated tempo of a track in beats per minute (BPM).
type : The object type
id : Track ID
uri : The Spotify URI for the track.
track_herf : A link to the Web API endpoint providing full details of the track.
analysis_url : A URL to access the full audio analysis of this track.
duration_ms : The duration of the track in milliseconds.
time_signatiure : An estimated time signature.

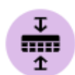
*More details :
https://developer.spotify.com/documentation/web-api/reference/#/operations/get-track

Pipeline



Use of Software

1. Import Data
   Can import the data using the green icon on the left bar.
   

2. Clean and Transform Data
   Can drop columns or change the column's name or transform lots of columns with this icon.
   
3. Aggregate Dataset Rows
   Using this icon, we can aggregate the datasets to make one column a unique value.
   
4. Merge Datasets
   Can join two tables with each data column.
   
5. AK Miner
   Can do data mining using this icon. Select the miner method between FP Miner and Bayesian miner, and it will result in some pattern of the data.
   
6. AK Pattern Browser
   It will show the result of the analysis based on the pattern that was already found using AK Miner.
   Or, it will recommend some feature combination when clicking "Launch Feature Explorer".
   
7. Visualize Data
   It will show the plot of the dataset. We can select X and Y and the type of the plot.
   

Analysis of Track Popularity

- Dataset Preprocessing
    1. Drop the non-numerical columns
    2. Aggregate the rows to make the key column as unique value.
    3. Inner Join two tables with music track's id to analyze with music features.
    4. Drop the meaningless columns

- Using Bayesian Miner and finding Pattern

  Target : Track Popularity

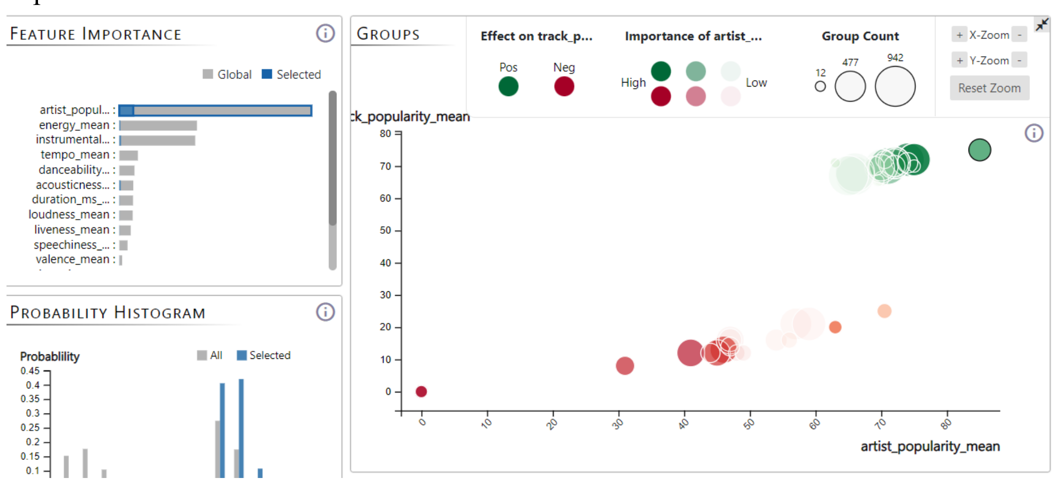
  ACTION OUTPUT

  Mining Results

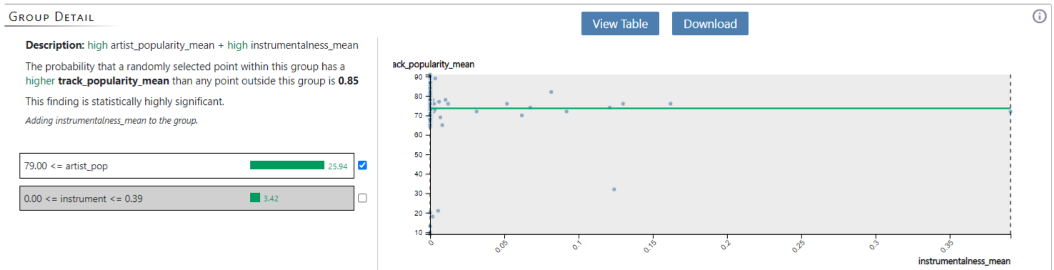
Input Data Properties	
Item Count	**1000**
Feature Count	**13**

Mined Patterns - Details	
Pattern Count	**44**
Largest Pattern	**801**
Smallest Pattern	**10**
Maximum Feature Count	**3**
Minimum Feature Count	**1**

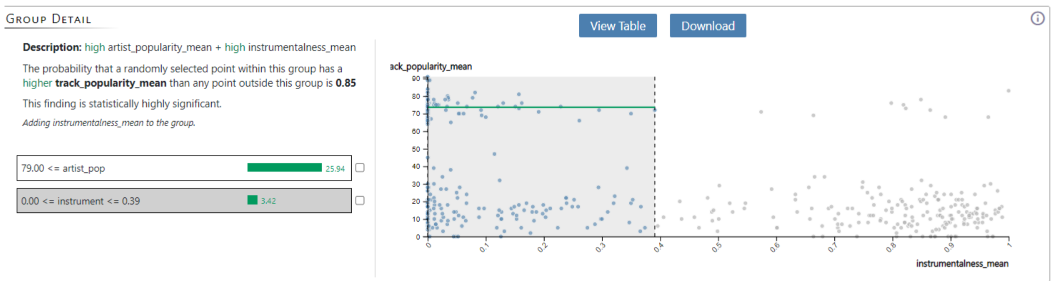
- Insight
    1. When we see the feature importance and groups, it is shown that the track popularity and the artist popularity have big positive correlation, and the artist popularity is the most important feature.



When click the group at the top of right side, the instrumentalness is also important in the other attribute section.

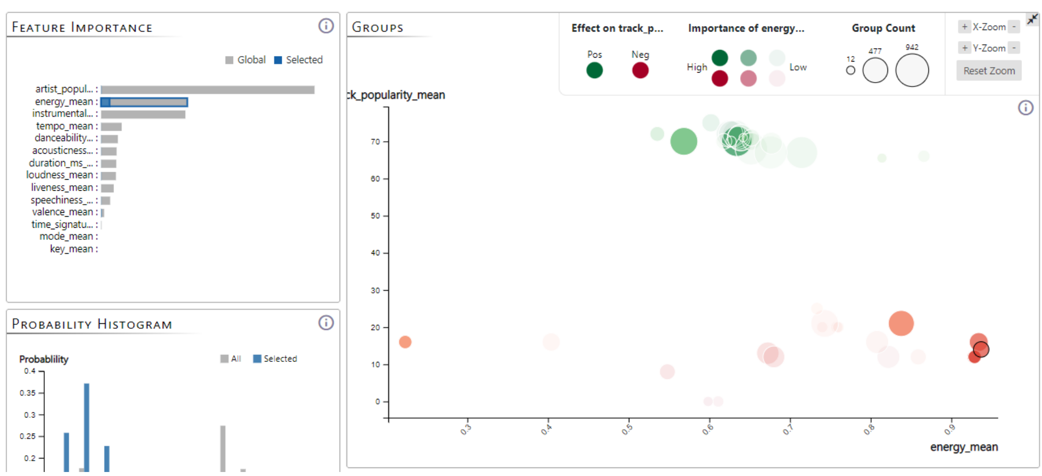


It is showed that this group contains low instrumentalness, and high track popularity. But we cannot say that most of tracks which has small instrumentalness are mostly popular tracks.

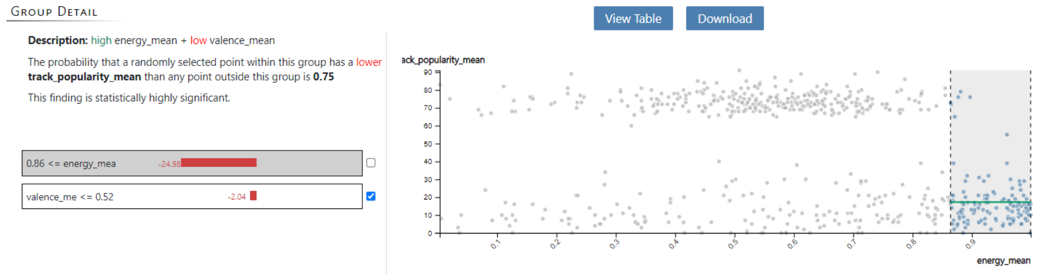


When we look at the entire track points, there are many tracks that have small instrumentalness but low track popularity. But still we can say that most of the popular tracks have small instrumentalness since there are not many tracks for the big instrumentalness with high track popularity.

2. This group has small track popularity and large energy.



What I found in this group detail is that these tracks have high energy and the valence is less the medium value, which means that those tracks sounds more negative (sad, depressed, angry).



So, I was wondering that the valence can affect to the track popularity, and the result is not actually.

When I click the valence in the feature importance, the groups seem that there are no big correlation between track popularity and valence. However, most of the tracks with high popularity don't have low valence value, so I can guess that tracks with low valence are difficult to be popular tracks.

Analysis of Artist Genre

- Dataset Preprocessing
    1. Drop the non-numerical columns
    2. Aggregate the rows to make the key column as unique value.
    3. Inner Join two tables with music track's id to analyze with music features.
    4. Drop the meaningless columns and artist name columns
    5. Make it as one columns for each genre using Cell Split.

- Using Bayesian Miner and find Pattern
  Target : Artist Popularity

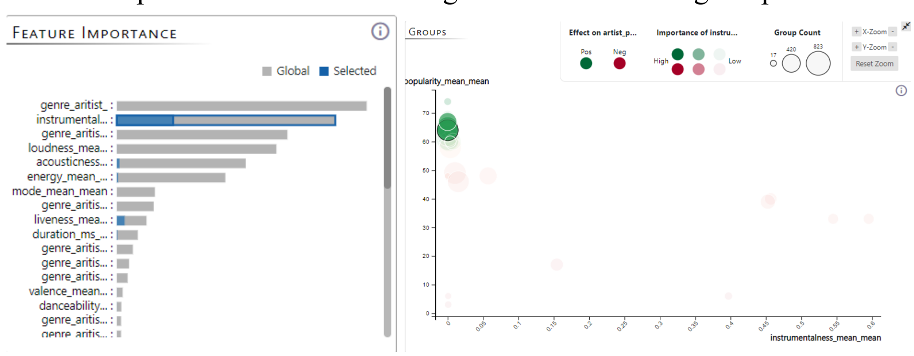
ACTION OUTPUT

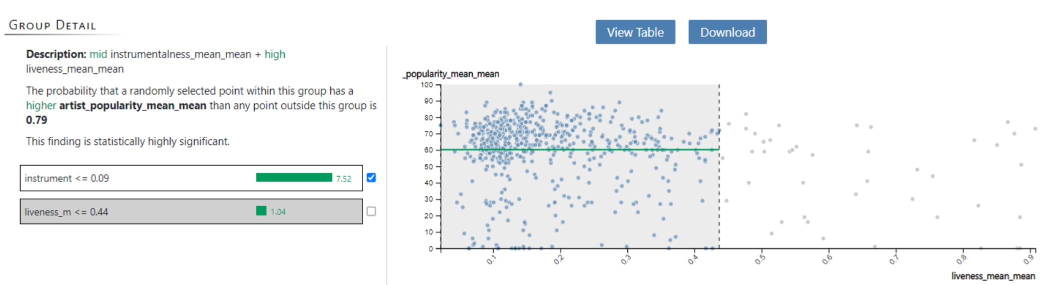
Mining Results

Input Data Properties	
Item Count	**1000**
Feature Count	**40**

Mined Patterns - Details	
Pattern Count	**33**
Largest Pattern	**894**
Smallest Pattern	**11**
Maximum Feature Count	**2**
Minimum Feature Count	**1**

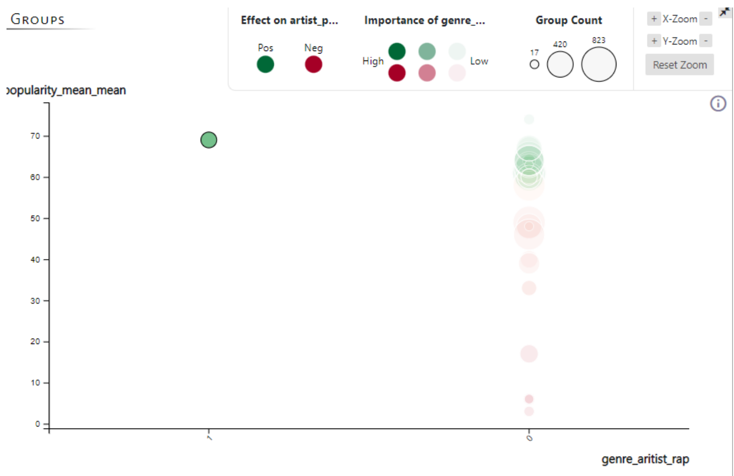
- Insight
    1. Firstly, the instrumentalness feature has the highest importance with artist popularity. When I choose the biggest positive group in the groups section, this group is shown having high relativeness with liveness, and medium relativeness with instrumentalness. So, it can be said that most of the tracks from popular artist has low instrumentalness with medium importance and medium range of liveness with high importance.



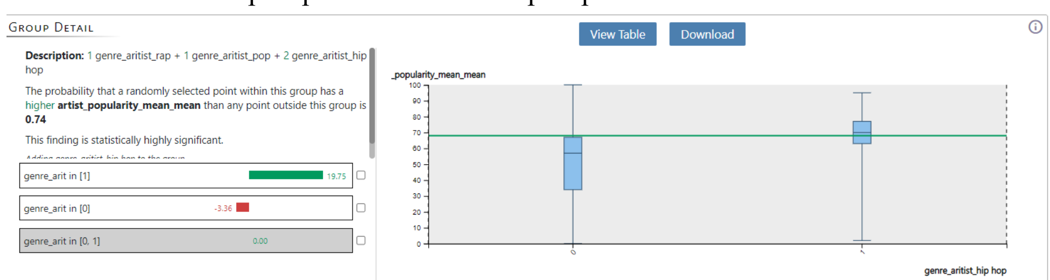


2.

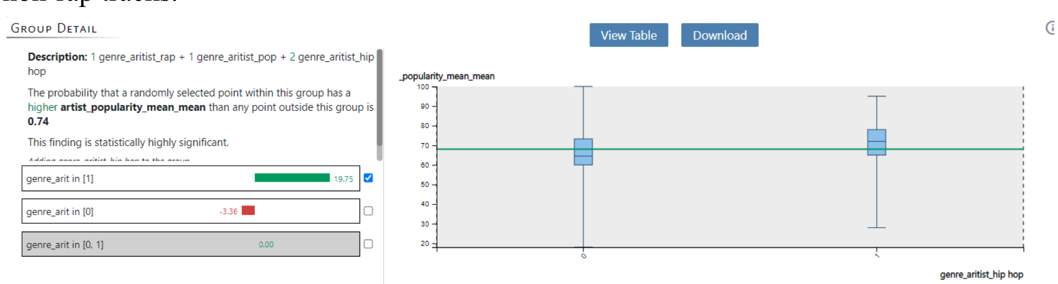
This group contains rap music.



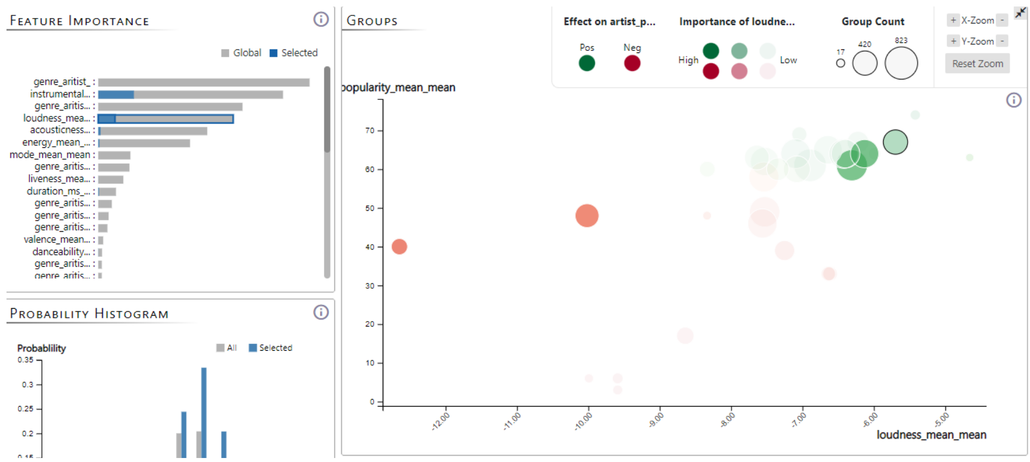
And then, I choose the feature "genre_artist_hiphop" which is hip hop music. And I can find there are both hip-hop music and non-hip-hop music.



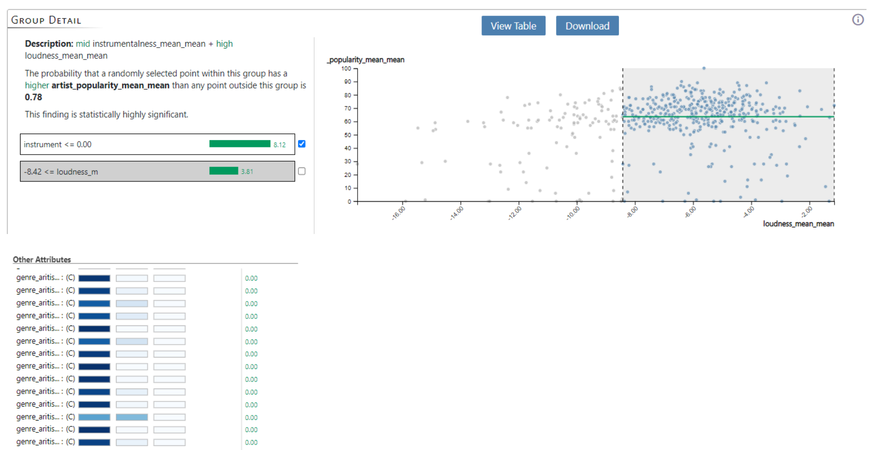
When I check the first check box, which is about genre_artist_rap, the artist popularity for the non-hip-hop group increased, and the hip-hop points don't have any change. In this group, the track which has hip-hop and rap genre at the same time are mostly popular. And also, the track with non-hip-hop and rap is much popular than hip-hop non-rap tracks.



3. When select the loudness feature, it is showed that the group with large loudness usually has high popularity. Unfortunately, for the left two red circle group don't have any special features, so I'm trying to select the highest popularity circle with high importance.



In the group detail, this group almost don't have any instrumentalness and have large loudness since instrumentalness means that predicting whether a track contains no vocals. So we can guess those points contain rap music or acapella musics.



In the other attributes section, it is shown that this group contains rap, pop, hip-hop, rock and etc group. One thing I surprised is that there are rock music in this group even though these tracks barely have instrumentalness.

I guess there's some rock music only with voice without instruments.

Code

https://github.com/JeongYoon-L/Spotify-Analysis/blob/main/Spotify_dataset.ipynb

Conclusions

I was trying to check what kind of genres there are in the tracks, but it was difficult since there are too many specific genres per track. Also, the result could have been biased since genres are focused on popular genres such as pop genre.

While doing my project, I realized that the distribution of data should be even, but my dataset was distorted in terms of genre. I made the genres as categorical variables, so it was hard to find results and make visualization results. For future analysis, I want to analyze specific genre like pop or edm or soundtrack and find something interesting within them.

References

https://developer.spotify.com/documentation/web-api/reference/#/

Attached Files

Pipeline : 0902.aka

Google Colab ipynb file : Spotify_dataset.ipynb

Input Dataset : new_with_genre_final.csv, final_audio_feature.csv