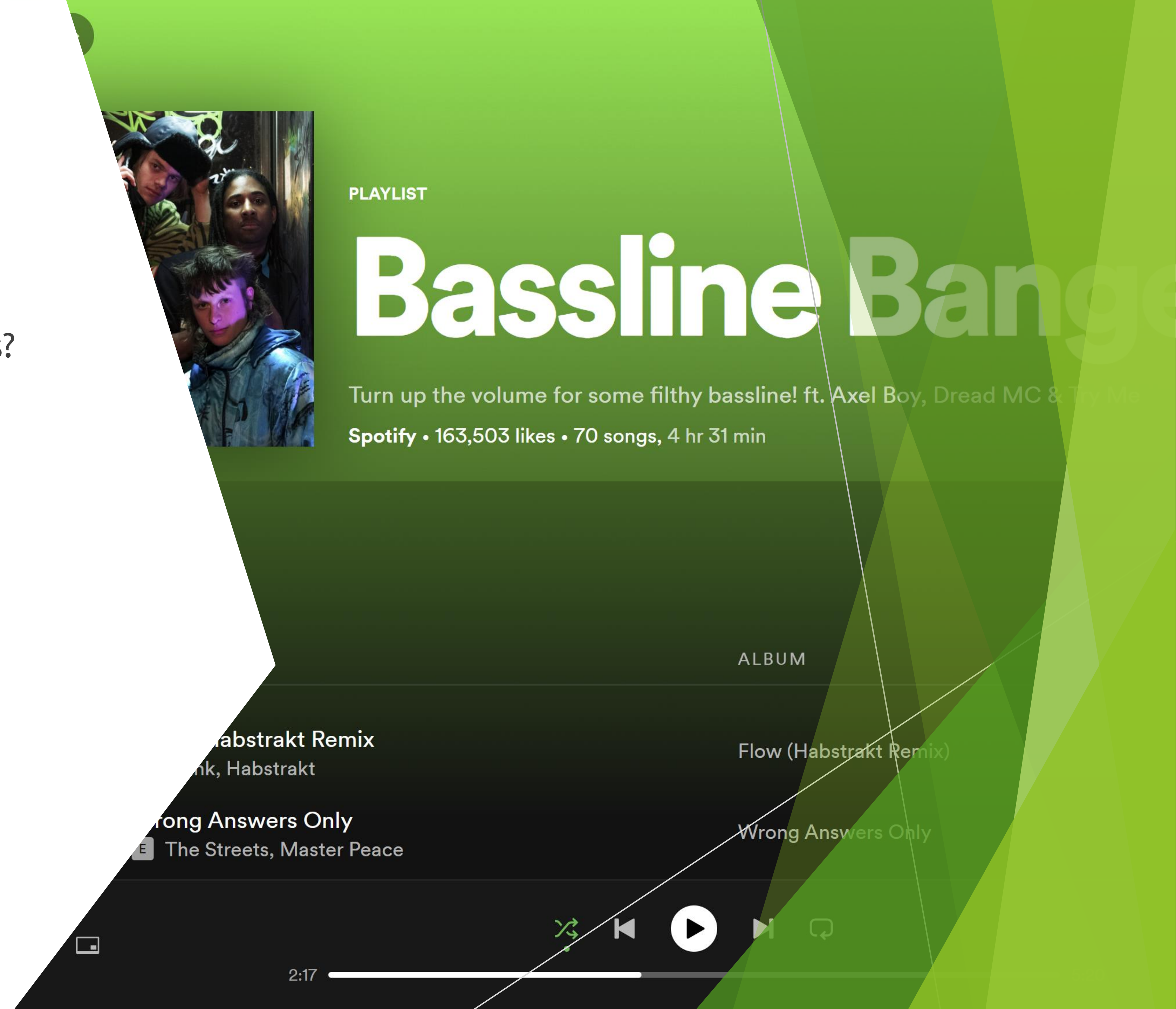
What makes a playlist successful?

Wesley Beckner, PhD

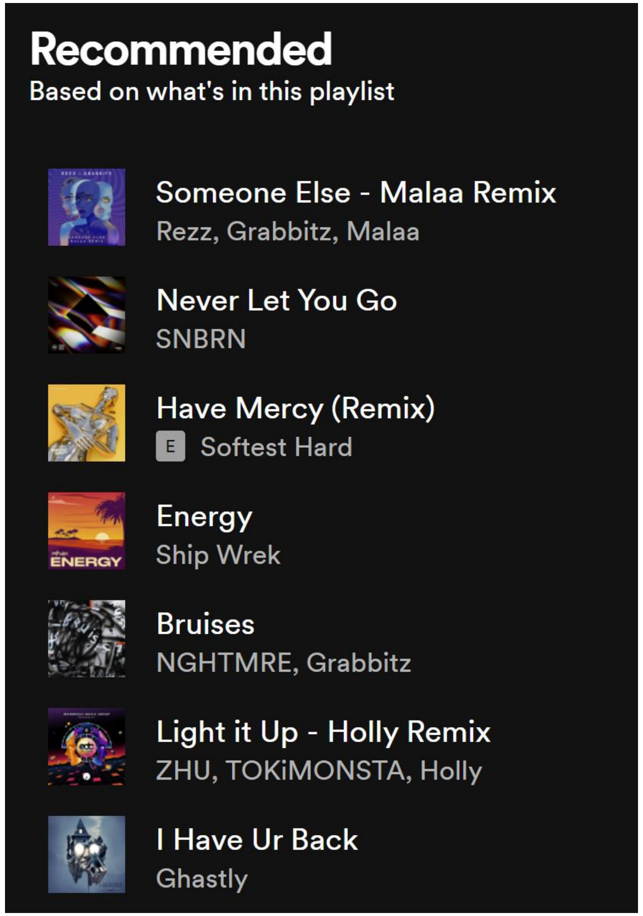
April 26, 2022

Agenda

- Why do we care about success?
- Defining success: popular vs trending
- Features of popular playlists
- Features of trending playlists
- Conclusions
- Future work



Why do we care about success?



Spotify could use the associations with "trendy" playlists in how it recommends songs for playlist creators

Defining success

Popular vs Trending

Popularity

Variable	Reasoning
mau_previous_month	user based likeability over the long term
mau_both_months	user based replay value over the long term
monthly_stream30s	depth measure of likeability over the long term
stream30s	depth measure of likeability over the short term

- A playlist is labeled popular if it is top quartile for *all* variables
- 5% labeled successful

Trendiness

Variable	Derivation
listen_conversions	stream30s / streams
user_conversions	mau / users
user_retention	mau_both_months / mau_previous_month
mau_growing	mau > mau_previous_month

- A playlist is labeled trendy if it is above 50% for conversions and retention and “True” for mau_growing
- 10% labeled successful

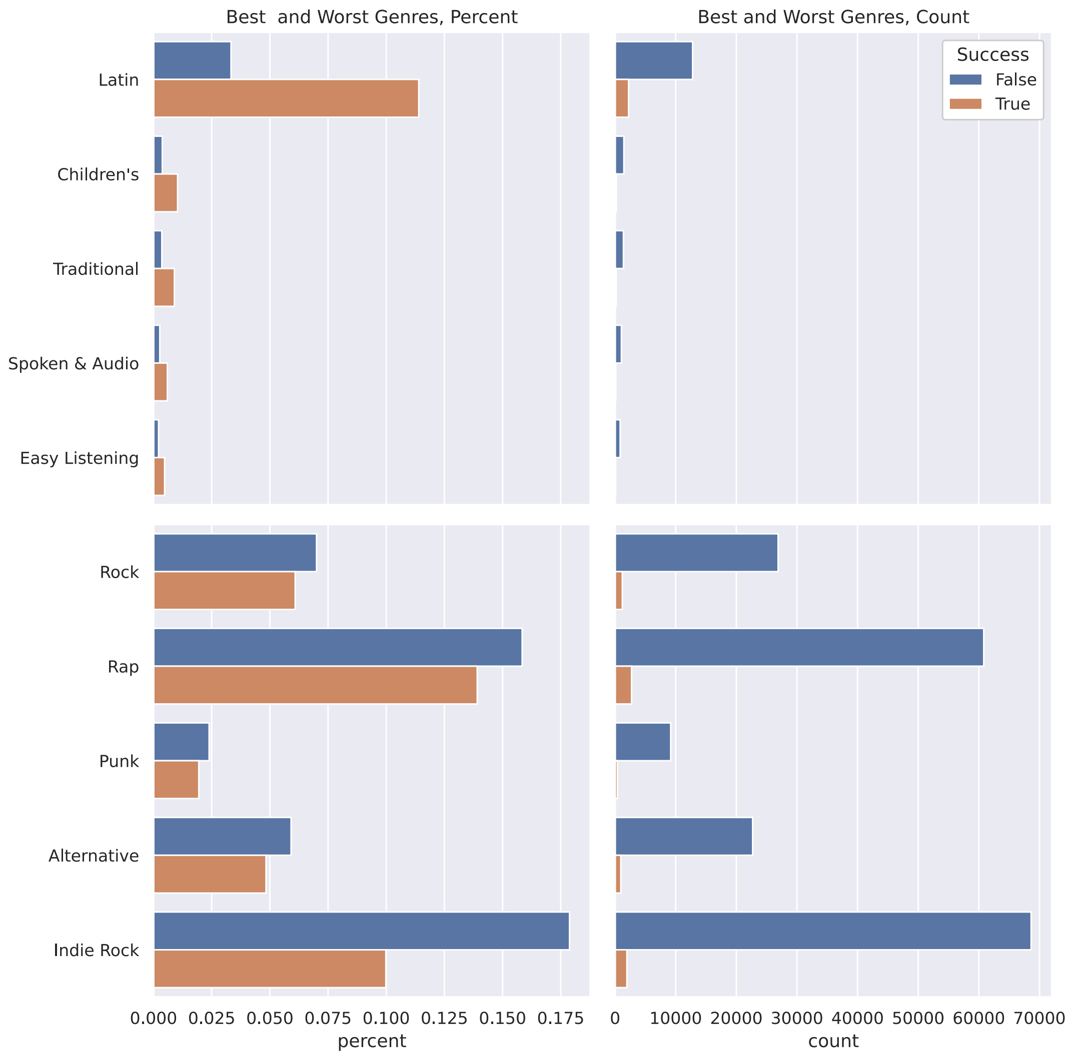
Another way to think of the two definitions: popularity is at the top of the mountain, trendiness is ascending the base of the mountain

Features of popular playlists

Chi-square, t-tests, logistic regression

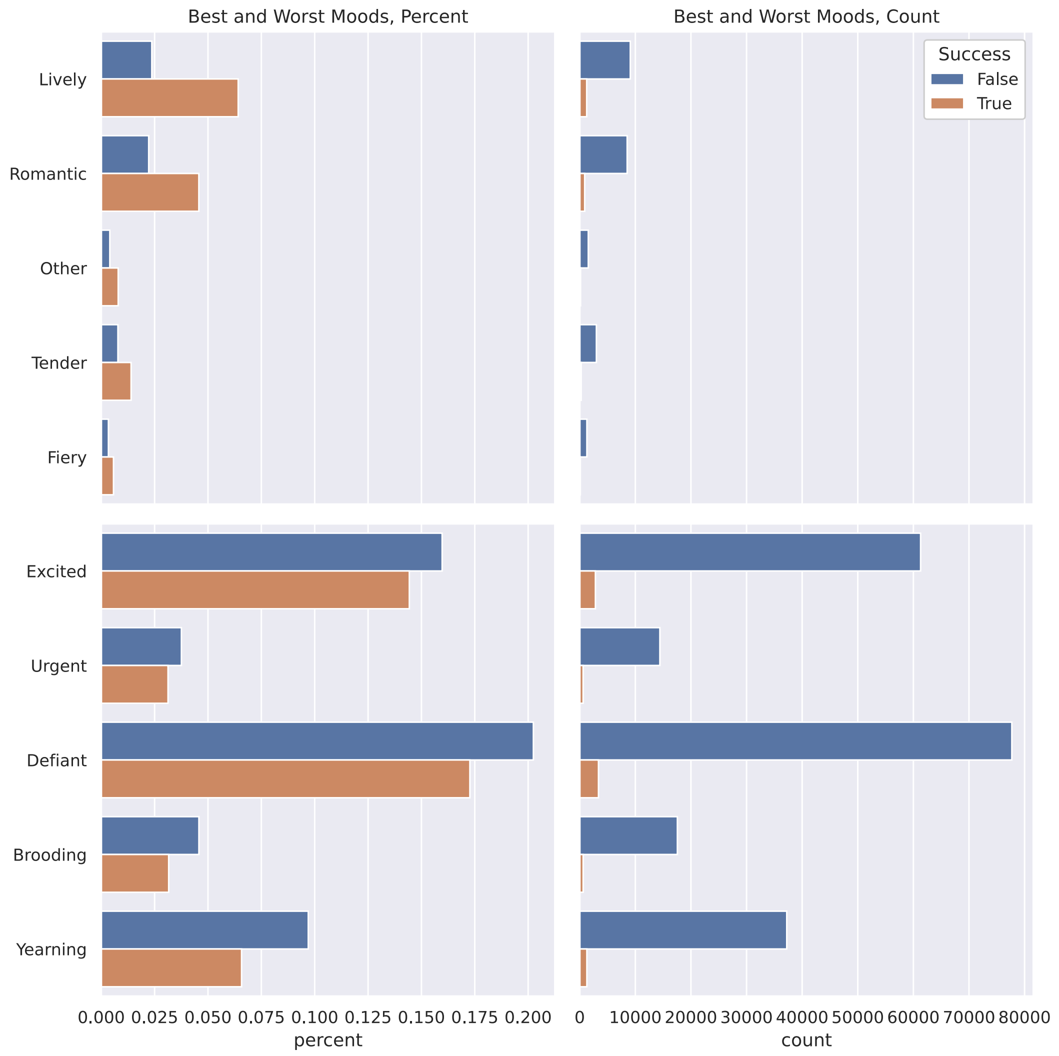
Chi-square: Latin and Children's **genres** will increase the likelihood of being successful by 3.7x and 2.8x, respectively

group	chi	p-value	multiplier
Latin	3372	0	3.73
Children's	196	1.67E-44	2.76
Traditional	139	3.48E-32	2.52
Spoken & Audio	68	2.05E-16	2.21
Easy Listening	53	3.01E-13	2.20
…			
Rock	24	9.59E-07	0.86
Rap	52	5.39E-13	0.86
Punk	16	7.68E-05	0.81
Alternative	39	4.36E-10	0.81
Indie Rock	802	2.3E-176	0.51

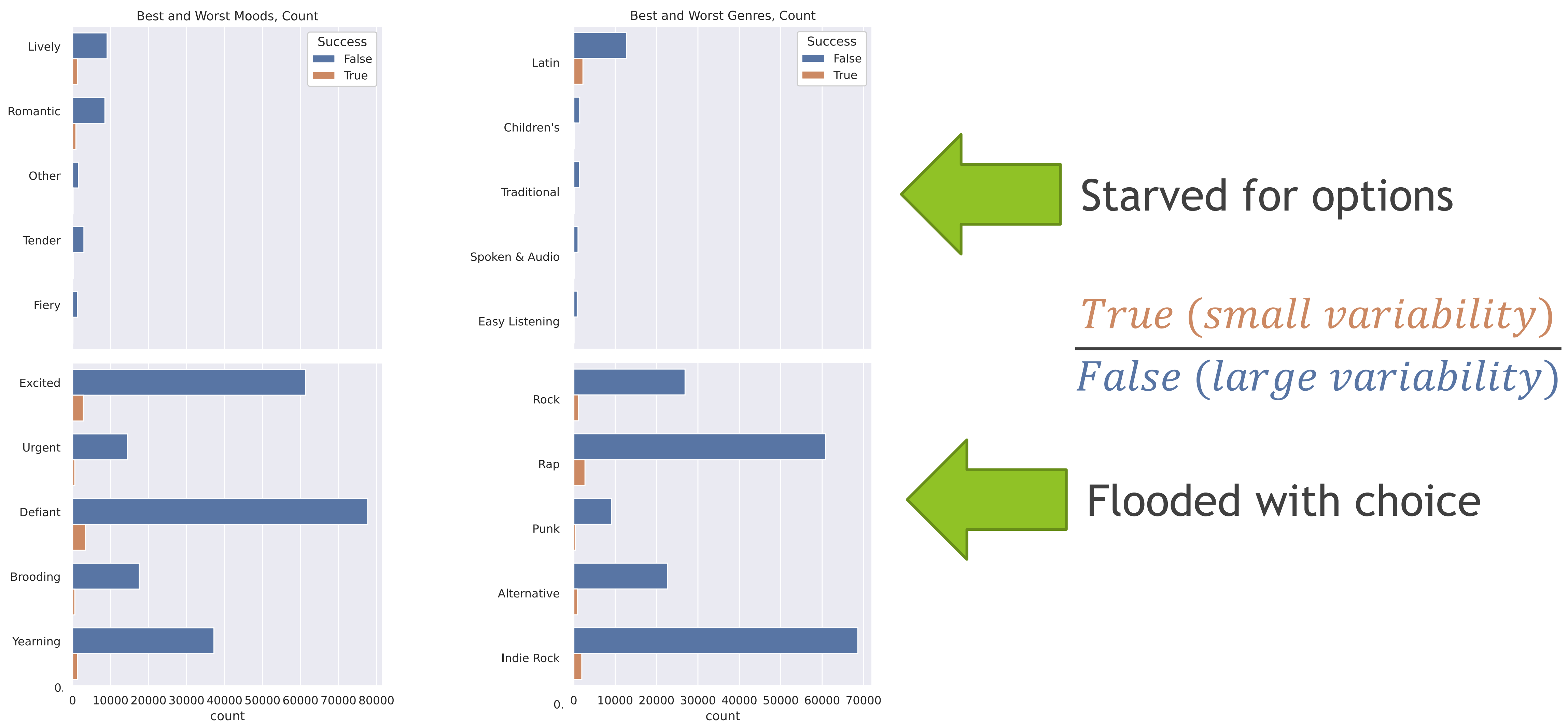


Chi-square: Lively and Romantic **moods** will increase the likelihood of being successful by 2.8x and 2.1x, respectively

group	chi	p-value	multiplier
Lively	1216	2E-266	2.83
Romantic	449	1.3E-99	2.11
Other	67	2.16E-16	1.99
Tender	87	9.73E-21	1.80
Fiery	30	4.98E-08	1.71
	…		
Excited	33	1.06E-08	0.89
Urgent	20	7.12E-06	0.83
Defiant	102	4.99E-24	0.82
Brooding	86	1.59E-20	0.68
Yearning	210	1.56E-47	0.66

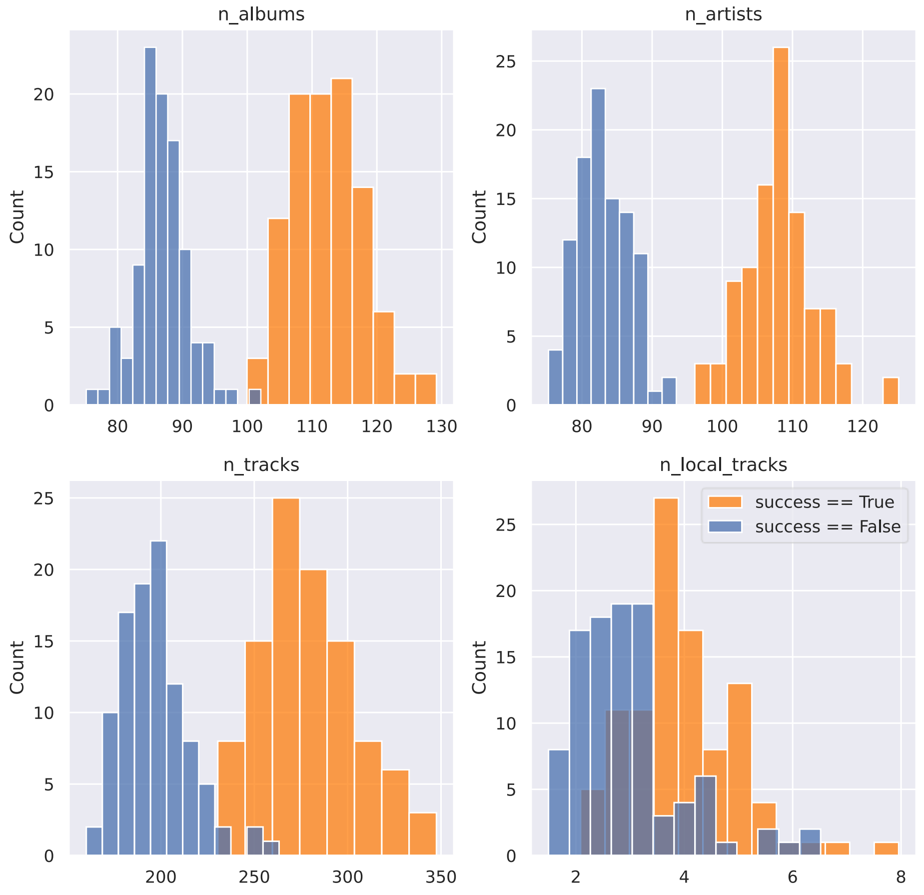


Both genre and mood have minority classes that impact popularity positively and majority classes that impact popularity negatively



T-Test: Continuous features such as albums, artists, tracks, and local tracks have a clear relationship with success

feature	test stat	p-value	upper q avg	lower q avg
n_artists	40	1.19E-90	107.9	83.0
n_albums	35	7.34E-83	112.2	86.9
n_tracks	26	2.03E-63	279.0	197.2
n_local_tracks	7	1.52E-11	4.0	2.9



Logistic regression: comparatively, genre is more important than mood, though both have selected features that push the playlist toward the successful class

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-3.1965	0.009	-363.622	0.000	-3.214	-3.179
genre_1_Latin	1.1350	0.028	39.882	0.000	1.079	1.191
n_artists	0.0845	0.014	5.993	0.000	0.057	0.112
genre_3_-	0.4321	0.022	19.626	0.000	0.389	0.475
genre_1_Soundtrack	0.5298	0.048	11.054	0.000	0.436	0.624
genre_1_Children's	1.0654	0.076	13.965	0.000	0.916	1.215
genre_1_Traditional	0.9924	0.082	12.096	0.000	0.832	1.153
mood_3_Romantic	0.3785	0.038	9.978	0.000	0.304	0.453
genre_1_Spoken & Audio	0.8005	0.099	8.074	0.000	0.606	0.995
genre_1_Easy Listening	0.7925	0.111	7.149	0.000	0.575	1.010
genre_3_Traditional	0.4275	0.063	6.801	0.000	0.304	0.551
genre_1_Jazz	0.4414	0.067	6.543	0.000	0.309	0.574
mood_2_Sophisticated	0.2599	0.040	6.521	0.000	0.182	0.338
mood_2_Romantic	0.2940	0.038	7.745	0.000	0.220	0.368
n_albums	0.0920	0.014	6.418	0.000	0.064	0.120
mood_1_Romantic	0.2532	0.040	6.262	0.000	0.174	0.332

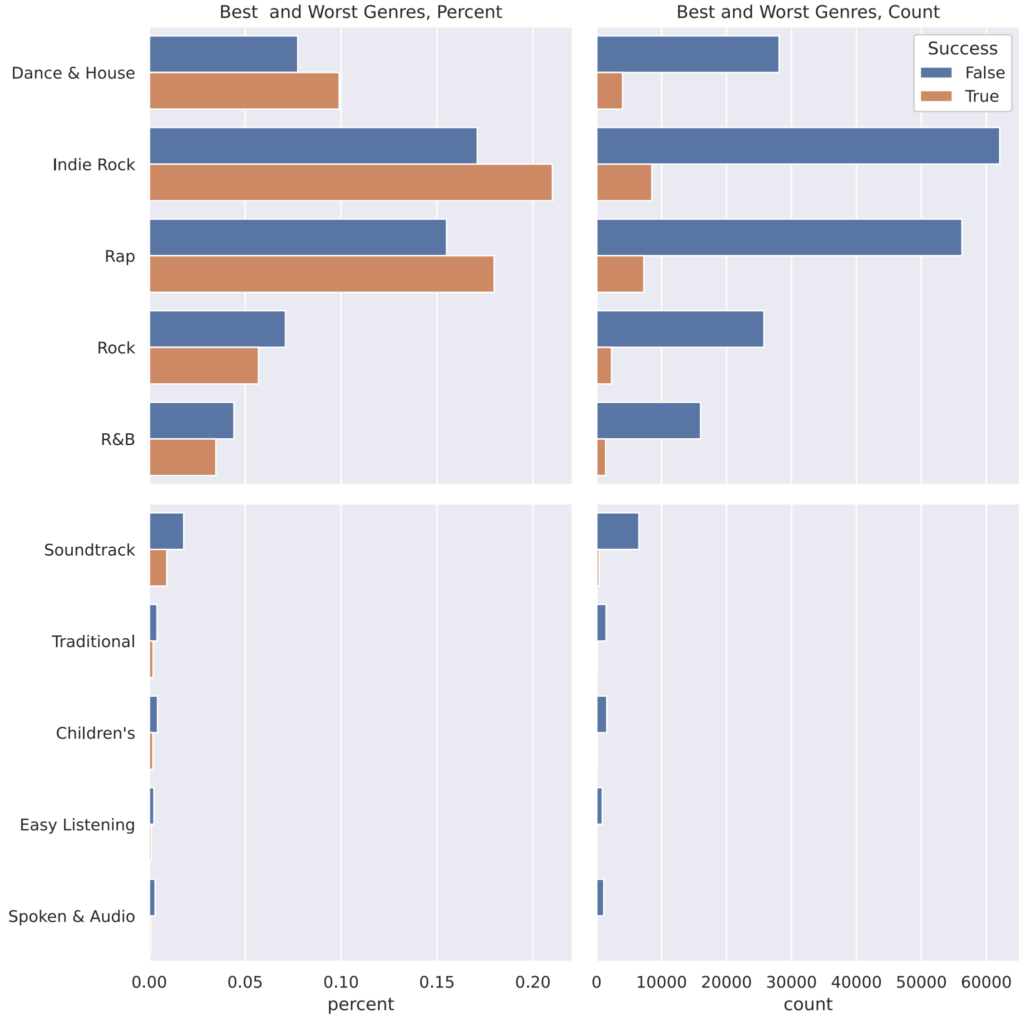
some features missing in chi-square tests signify collinearity (ex: mood_1, lively)

Features of trending playlists

Chi-square, t-tests, logistic regression

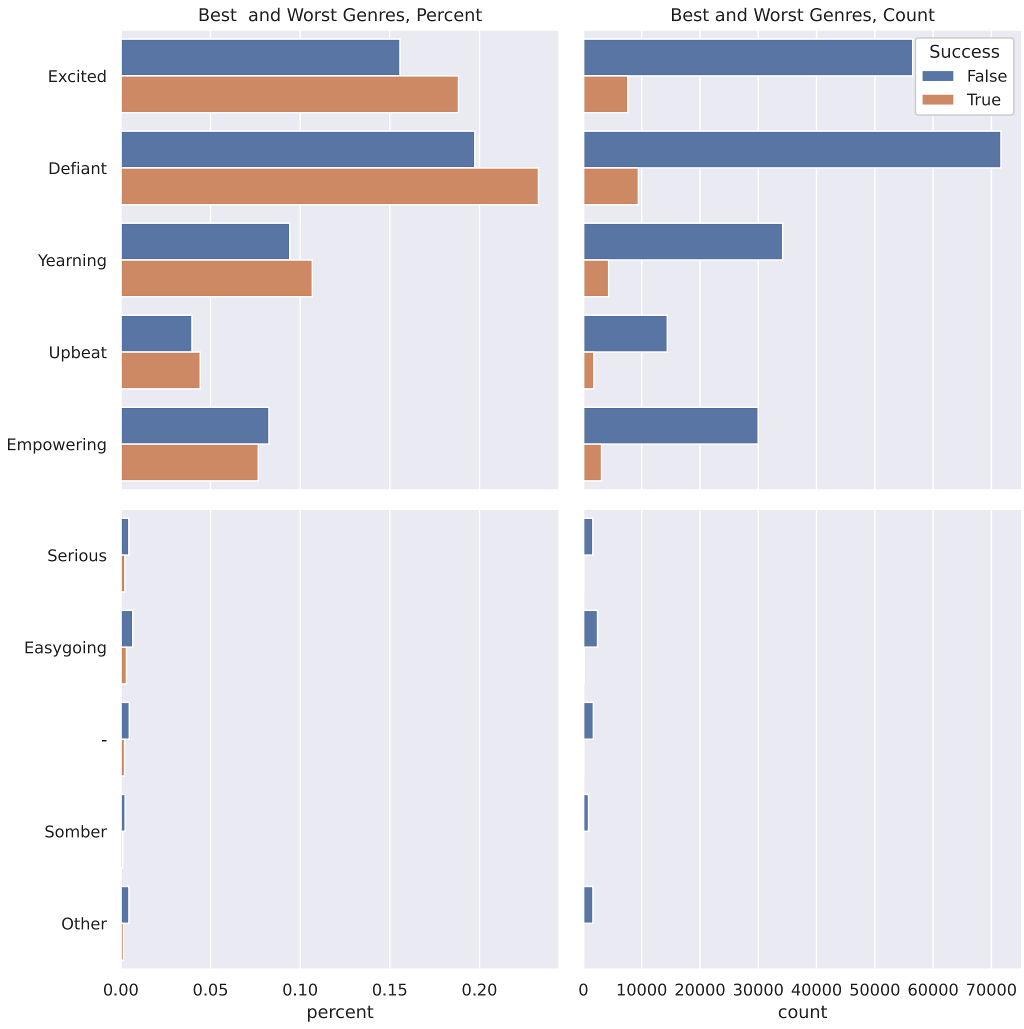
A reversal of roles: Indie Rock, Rock, Rap now +, Children’s, Traditional, Easy Listening now -

group	chi	p-value	multiplier
Dance & House	231	3.22E-52	1.31
Indie Rock	386	5.21E-86	1.29
Rap	169	1.45E-38	1.20
Rock	111	6.44E-26	0.79
R&B	78	1.31E-18	0.78
	…		
Soundtrack	169	1.2E-38	0.50
Traditional	39	3.77E-10	0.49
Children's	52	5.04E-13	0.44
Easy Listening	31	3.15E-08	0.42
Spoken & Audio	59	1.76E-14	0.30

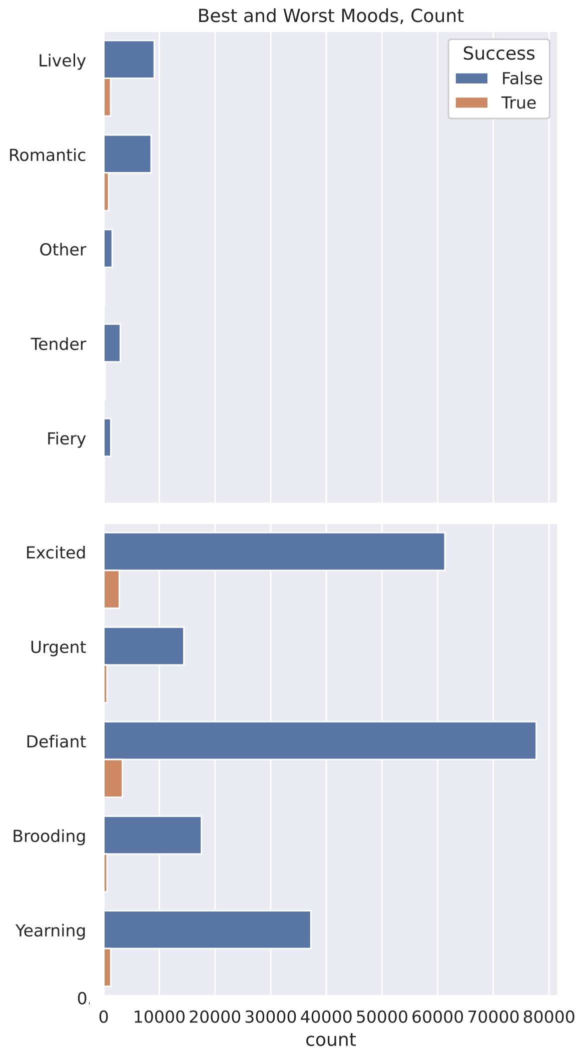


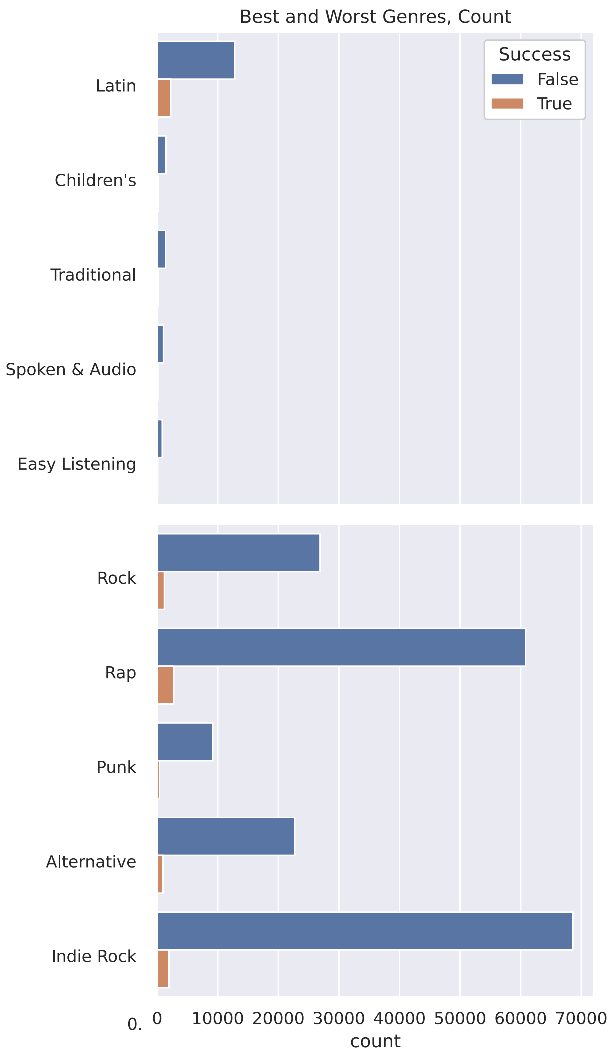
A reversal of roles: Excited, Defiant, Yearning now +

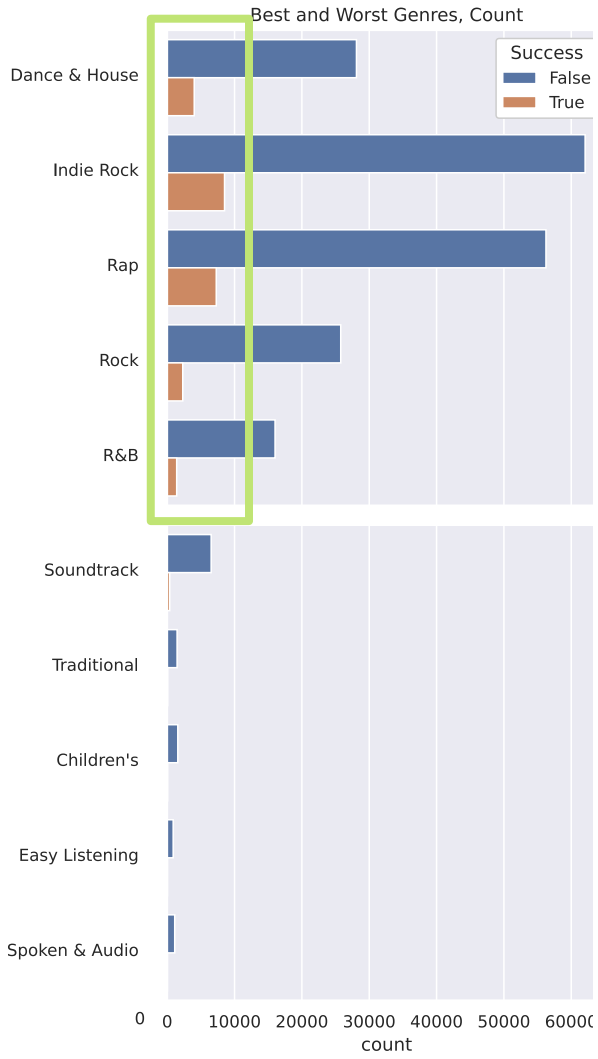
group	chi	p-value	multiplier
Excited	290	5.44E-65	1.26
Defiant	285	6.06E-64	1.23
Yearning	67	3.16E-16	1.15
Upbeat	20	8.12E-06	1.12
Empowering	17	3.27E-05	0.92
	…		
Serious	42	8.92E-11	0.50
Easygoing	73	1.45E-17	0.47
-	58	3.21E-14	0.43
Somber	31	2.78E-08	0.42
Other	82	1.5E-19	0.32

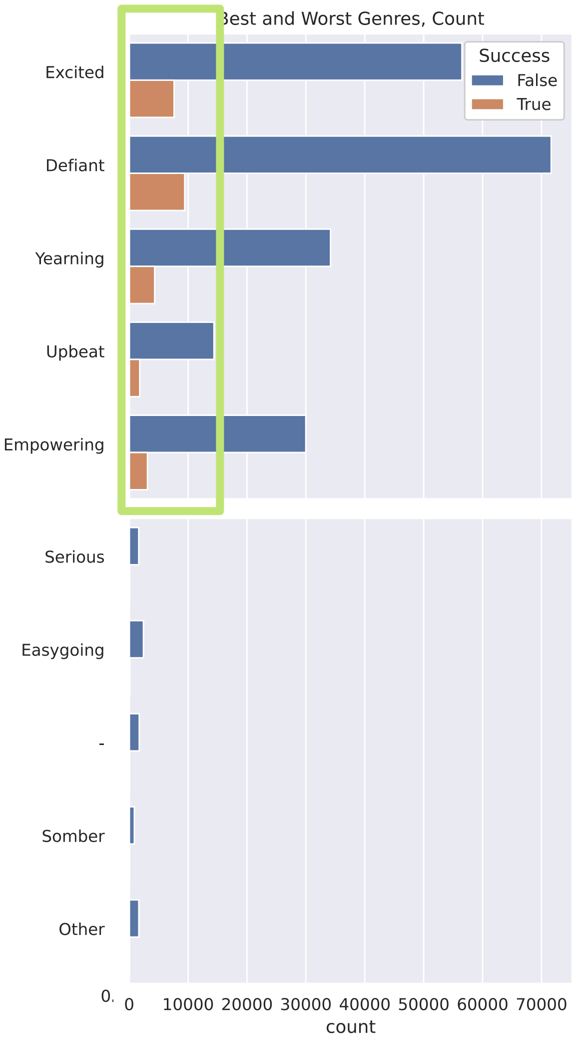


Success is now attributed to growth in numerator







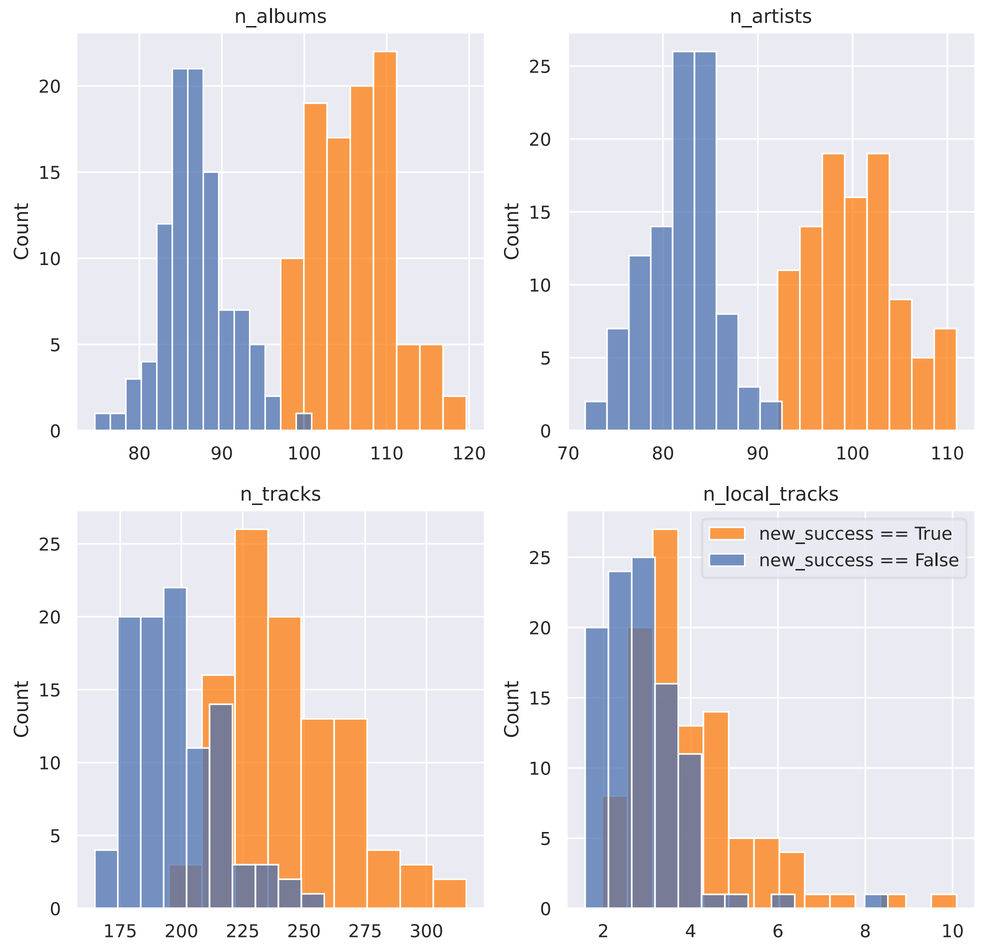


albums, artists, tracks, and local tracks have a clear relationship with success, although strength of associations is slightly weaker

feature	test stat	p-value	upper q avg	lower q avg
n_artists	30	3.33E-75	100	82
n_albums	30	8.03E-75	106	87
n_tracks	15	2.12E-34	243	198
n_local_tracks	6	1.09E-08	4	3

recall popular playlist results…

n_artists	40	1.19E-90	107.9	83.0
n_albums	35	7.34E-83	112.2	86.9
n_tracks	26	2.03E-63	279.0	197.2
n_local_tracks	7	1.52E-11	4.0	2.9



Logistic regression: compared to popular playlist model: higher representation of mood, more negative associations, influence of genre_3 = - is flipped, artists and albums are proxy for one another

coef	
-3.1965	intercept
1.1350	genre_1_Latin
0.0845	n_artists
0.4321	genre_3_-
0.5298	genre_1_Soundtrack
1.0654	genre_1_Children's
0.9924	genre_1_Traditional
0.3785	mood_3_Romantic
0.8005	genre_1_Spoken & Audio
0.7925	genre_1_Easy Listening
0.4275	genre_3_Traditional
0.4414	genre_1_Jazz
0.2599	mood_2_Sophisticated
0.2940	mood_2_Romantic
0.0920	n_albums
0.2532	mood_1_Romantic

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-2.4336	0.012	-201.725	0.000	-2.457	-2.410
genre_3_-	-0.6766	0.025	-27.158	0.000	-0.725	-0.628
n_albums	0.1399	0.015	9.597	0.000	0.111	0.169
genre_1_Indie Rock	0.2702	0.016	17.240	0.000	0.240	0.301
mood_1_Defiant	0.2505	0.018	14.035	0.000	0.215	0.285
genre_1_Dance & House	0.3042	0.021	14.388	0.000	0.263	0.346
mood_1_Excited	0.1917	0.017	11.607	0.000	0.159	0.224
mood_1_Upbeat	0.2698	0.028	9.713	0.000	0.215	0.324
genre_2_Indie Rock	0.1527	0.019	7.854	0.000	0.115	0.191
genre_1_Rap	0.1876	0.019	9.843	0.000	0.150	0.225
genre_1_Religious	0.2676	0.030	8.877	0.000	0.209	0.327
mood_2_Romantic	-0.2858	0.044	-6.533	0.000	-0.372	-0.200
mood_1_Yearning	0.1965	0.020	9.809	0.000	0.157	0.236
mood_1_Romantic	-0.2540	0.045	-5.620	0.000	-0.343	-0.165
mood_3_Romantic	-0.2249	0.042	-5.304	0.000	-0.308	-0.142

Conclusions

Relationship between success and moods and genres is dependent on how we define success

The top of the mountain (popularity) is capped for each genre/mood and so associations with "success" are dominated by denominator effect – genres/moods with many constituents are simply less likely to have a seat at the helm

The sides of the mountain (trendiness) are not capped for each genre/mood – and here we even see a reversal of associations compared to the popularity success metric

Future work and Business Context

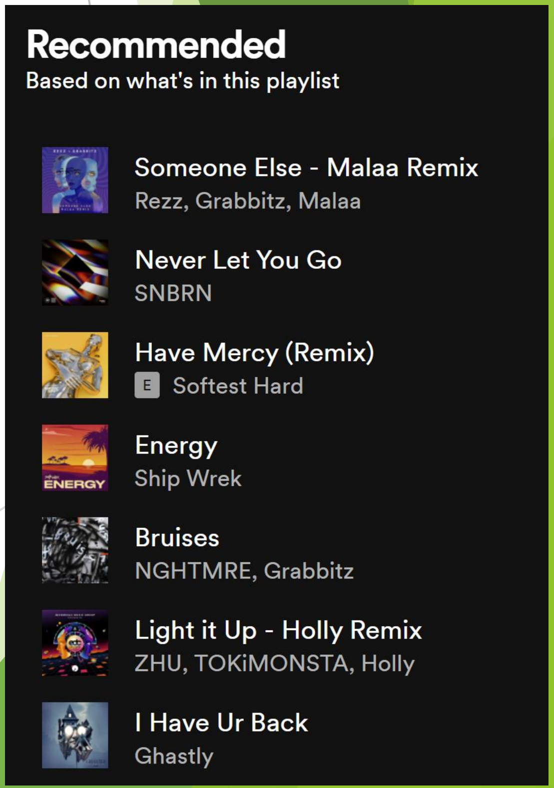
Spotify could use the associations with "trendy" playlists in how it recommends songs for playlist creators

Ex: remember that genre_3 = "-" and how it had a negative coefficient?

Perhaps this indicates that a trendy playlist has diversity in genre (i.e. – is a wildcard that gets assigned when there is not a strong 3rd genre)

Spotify would make recommendations that diversify the playlist in the context of the current 2 dominant genres

This will increase the number of hill climbers at the base of the mountain, and ideally lead to an overall greater number of playlist listeners!

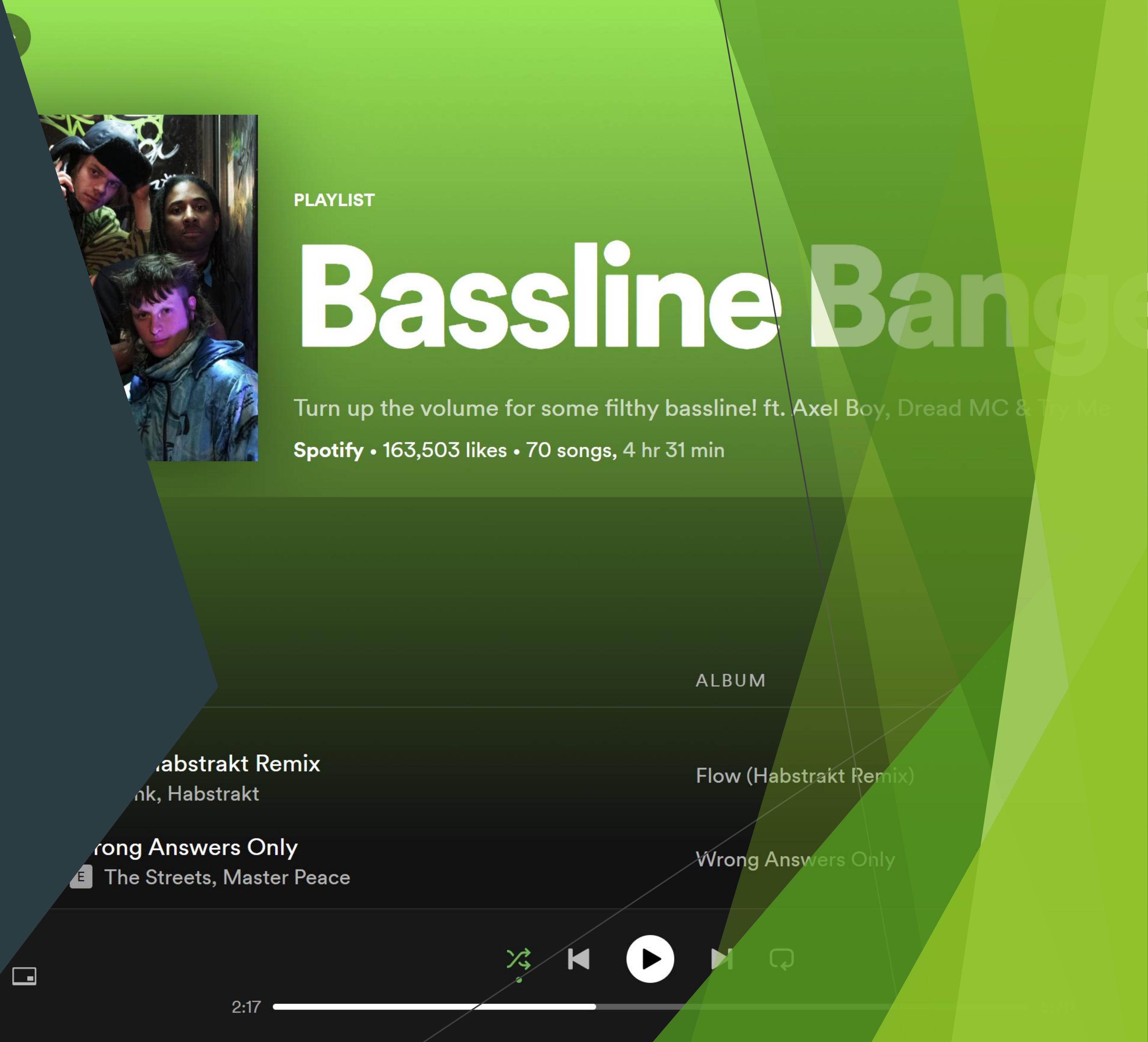


Business Context Summary

We care about recommending relevant playlists because this means the freshest beats are on the home screen – leading consumers to choose Spotify over competitors

Appendices

- example dataset slides
- moods/genres by count rank
- why did I choose the independent variables that I did
- major assumptions and limitations
- data I would like to have
- unanswered questions
- what other features would I consider
- alternative methods
- preprocessing and model training
- additional chi-square tests

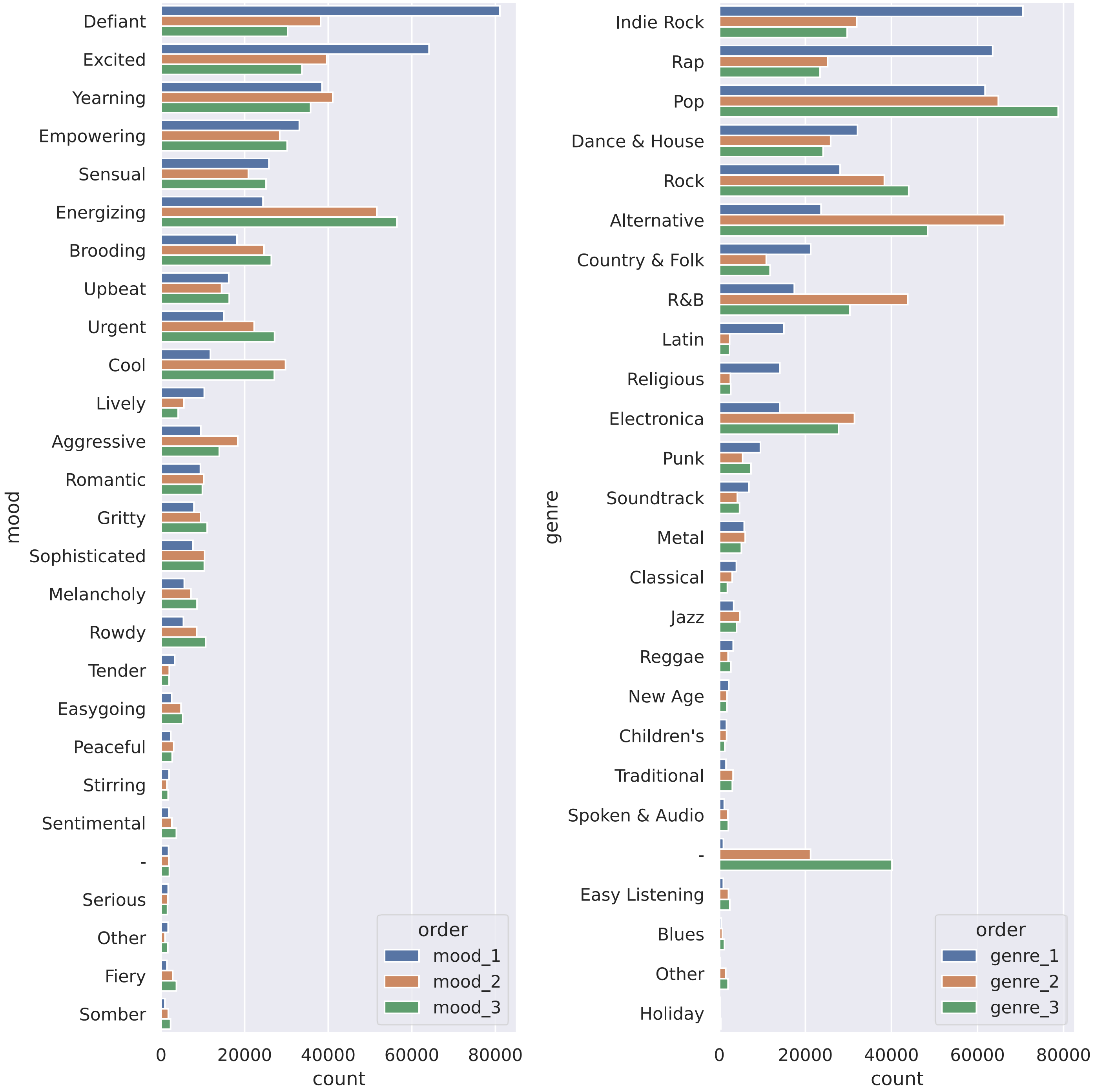


streams	stream30s	dau	wau	mau	mau_previous_month	mau_both_months	users	skippers	owner_country	n_tracks	n_local_tracks	n_artists
59	51	2	10	17	12	5	29	0	US	315	1	107
49158	31127	4711	30856	116427	131460	17878	382105	4520	US	26	0	1
76	22	3	6	18	21	5	51	2	US	213	3	153
52	34	5	27	69	76	12	161	4	US	48	0	46
35	25	2	8	27	27	5	70	1	US	102	0	88
1279	900	66	377	981	645	82	1872	29	US	100	0	64
45	34	2	16	30	18	4	51	2	US	134	0	13
60	53	4	13	34	31	14	56	1	US	555	0	456
107	90	8	42	163	155	18	413	6	US	121	0	4
14	10	2	12	57	41	3	117	3	US	31	0	12
6448	4979	606	8891	34906	28268	2456	91942	546	US	50	0	47
126	98	6	35	93	95	18	225	3	US	210	1	197
42	27	6	28	101	114	4	293	5	US	26	0	25
[illegible]	76	7	38	95	90	20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	15	4	13	42	41	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	13	1	6	18	15	5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

… a “Popular” playlist

streams	stream30s	dau	wau	mau	mau_previous_month	mau_both_months	users	skippers	owner_country	n_tracks	n_local_tracks	n_artists
28	15	1	1	2	1	1	2	0	US	321	0	170
9	5	1	2	2	1	1	2	0	US	373	8	1
32	25	2	3	4	3	3	5	1	US	904	0	81
18	17	1	2	4	3	3	5	1	US	141	1	122
5	5	1	1	2	1	1	3	0	US	84	0	73
1	1	1	1	2	1	1	3	1	US	12	0	12
4	4	1	1	2	1	1	3	0	US	288	0	100
5	3	1	1	2	1	1	2	0	US	76	1	55
1	1	1	2	3	1	1	4	1	US	1124	0	312
8	7	1	2	2	1	1	2	0	US	114	7	72
2	2	1	1	2	1	1	2	0	US	62	0	55
1	1	1	1	2	1	1	2	1	US	25	5	15
1	1	1	2	2	1	1	2	1	US	136	65	64
7	12	1	1	2	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	5	1	1	2	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	10	1	3	5	2	2	6	1	US	106	0	39

… a “Trendy” playlist



moods/genres by count rank

why did I choose the independent variables that I did

- In short, simplicity. Other things I would have liked to consider:
- lemmatized titles
  - produce sentiment score
  - use TFIDF in success/non-success groups
  - use title length as an engineered feature
  - word embeddings to calculate distance from title to genre or title to mood
- use owner listens and/or skippers
  - subtract from the total listens or the number of active users, respectively

major assumptions and limitations

- is “trending” truly trending? We don’t actually know that our trending playlists lead to popularity, and this is a major assumption in the analysis
- no outliers were removed from the data… this is mitigated by taking a categorical modeling approach (i.e. vs linear regression)
- we assume that all playlists were made before the previous month (i.e. all playlists have run for the same duration of data acquisition)

data I would like to have…

- Track level behavior
  - When did users fall off the playlist
  - What are the relationships between new artist releases and playlists that contain those songs
- More historical data
  - My major assumption was that my label for “trendy” was correct and leads to “popular”
  - But with historical data we can validate
  - What is the variability – how big of a swing do playlists typically have in active users, or is it pretty consistent over time
  - Can we model the life cycle of a playlist, do they ever die out?
- Vector representations of genres/moods
  - Distance calculations as an engineered feature
  - Evidence for utility there – is the genre_3 = -
  - This is some measure of diversity
  - i.e. There were only 2 dominant genres vs 3

unanswered questions

- Is there a selection bias in Latin and Children (since it’s an English platform)
  - Ex: if this was a Chinese platform and there was a genre for “English” only the best English songs would make it to the platform
  - But EDM has such a diverse ecosystem none of the songs are overtly popular
  - Same idea with “Traditional” selection bias – we won’t have songs from the 50s-80s that weren’t popular cause why would we

alternative methods

- it might have been better to set a floor for # of users (to target playlists that are slightly more visible)
- rerun logistic model without intercept to check sign flips

preprocessing and model training

- t-tests
  - 1000 bootstrapped samples of x 100
    - 100 averages for each continuous variable in each target group (success/non-success)
- forward selection + logistic model
  - intercept is added
  - 164 features; next feature that best improves pseudo-r square w/o moving p value for any feature past 0.01 is added
  - repeat until 15 features are found

additional chi-square tests

5 metrics

x (top 10% + top 1%)

x (primary, secondary, or tertiary genre/mood)

=

30 chi-square tests

metrics

mau_previous_month

mau_both_months

mau

stream30s

monthly_listen30s

Number of times given mood had a p-value < 0.01 and multiplier effect >2x

Romantic	19
Lively	17
Serious	8
Peaceful	6
Other	5
Fiery	3
Tender	3
Easygoing	3
Sophisticated	2
Somber	1

Number of times given genre had a p-value < 0.01 and multiplier effect >2x

Traditional	16
Children's	16
Jazz	14
Latin	12
Easy Listening	8
Soundtrack	8
New Age	7
Holiday	6
Spoken & Audio	4
Other	2