Impact of Danceability on Spotify Track Popularity

Datasci W203: Lab 2

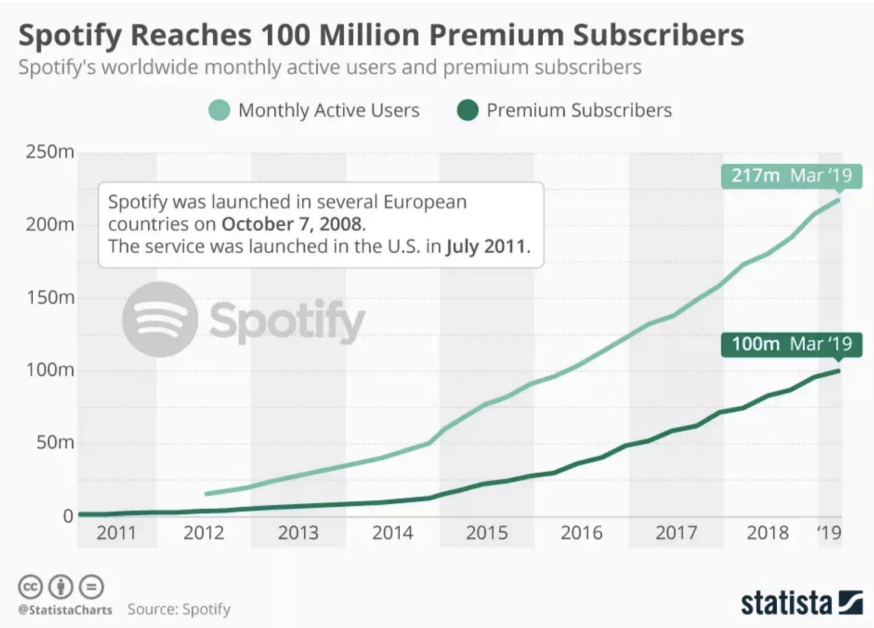
Saniya Lakka, Megan Martin, Andrew Sandico

Contents

1. Introduction

1.1 Motivation

Accurate predicting of music preferences is an important product feature for spotify to maintain user engagement and personalization. By increasing user engagement, users will spend more time on Spotify vs. other platforms leading to increasing revenue streams. Per Spotify's 2021 annual report, the company's first key risk factor is "We face significant competition and we might not be successful at attracting and retaining users; including through predicting, recommending, and playing content that our users enjoy, or monetizing our products and services including podcasts and other non-music content"[1]. To date, Spotify has successfully grown both their monthly active users and premium subscribers year over year:



Spotify monetizes by usage (how many times a song is played and how long a user stays on the app). As part of the data scientist team for Acme, Inc, we are supporting Spotify to optimize their monetization strategy. As number of streams (song plays) is a key metric to Spotify's ongoing revenue and growth plans, we seek to maximize the number of streams by identifying the key audio feature that causes an increased number of streams. Spotify provides robust descriptive metrics, called song features, for each song within its database. While the number of streams on Spotify is not readily available to the public in most situations, a metric called track popularity can be used as a proxy for understanding how much a song/track has been played.

1.2 Research Question

Due to the importance of predicting and recommending track popularity for maintaining users on the spotify platform, our research will be focused on identifying the key audio features for track popularity. With multiple audio features available for songs, this research will be focused on identifying the one key feature that causes track popularity leveraging an explanatory model. Specifically we aim to answer the following research question:

> *How does the danceability score for a spotify song affect its track popularity?*

With multiple types of songs and preferences available to a user, a key audio feature that would encourage repeatable plays would be of interest for optimizing a monetization strategy. Thus, we propose selecting a feature that creates motivation for a user to replay a song. We have selected the danceability feature for this conceptual model due to the connection to social events (going to a party, listening with friends, motivating oneself individually to dance). We will use the understanding between danceability and track popularity to make decisions on Spotify's selection of tracks to share or promote with their users.

2 Data and Methodology

2.1 About the Data

For our analysis we leveraged the Spotify API to generate our dataset. The Spotify API allows retrieval of playlist data by inputting a url of a spotify playlist into an API function. To access the Spotify API, we utilized a python package called spotipy which allowed us to retrieve audio features for each song in a playlist. After retrieving the data from the API we configured it into a dataframe and exported it to a csv for use.

Selection of the playlists to retrieve the audio features was an important consideration. Because existing playlists on the platform are either user-generated or optimized by algorithm, we needed to generate new playlists containing randomly selected songs. We utilized an online web application called randify which randomly selects Spotify songs from an extensive database in order to create new playlists. Due to limitations with the Spotify API, song features could only be retrieved for playlists containing 50 songs at a time. We generated 14 playlists utilizing randify. After generation of the randomized Spotify playlists, song feature data was pulled utilizing the Spotify API. From these playlists, a total of 700 rows (samples) and 22 columns (variables) were used to create our initial dataset.

Below is a table of the variables we will be using in our analysis paired with their definition, as well as reasons why we chose to omit certain variables. Descriptions of spotify's song feature metrics were obtained from here.

Column Name	Definition	Included/Omitted
Track Popularity	Popularity of track calculated by # of plays the track has had and how recent they were played. (0-100)	Included; outcome variable
Danceability	A 0-1 scale. 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. A value of 0.0 is least danceable and 1.0 is most danceable.	Included
Acousticness	A 0-1 scale where closer to 1 means high confidence the track is acoustic.	Included
Artist Popularity	Popularity of an artist from 0-100 where 100 is most popular.	Omitted: Derived from track popularity metric and Spotify algorithm, resulting in reverse causality with track popularity metric.
Duration	Duration of the track.	Omitted: This was a filtered variable in the original dataset.
Energy	A 0-1 scale. How energetic a track is: loud, noisy, upbeat etc.	Included

Column Name	Definition	Included/Omitted
Instrumentalness	A 0-1 scale. How many vocals are in a track compared to instruments. The closer to a score of 1, the more likely the track is purely instrumental.	Included
Key	The key the track is in. Integers map to pitches using standard pitch class notation (0-11; 0 = "C", etc).	Omitted: This is an ordinal variable with 12 outcomes.
Liveness	A 0-1 scale. Detects whether the track was a live performance.	Included
Loudness	-1 to -100. Loudness of a track in decibels.	Omitted: this is a metric used in the energy variable.
Mode	Binary 0 or 1. Whether the track is more major = 1 or minor = 0.	Included
Speechiness	A 0-1 scale. The presence of spoken words a.k.a an audio book or podcast etc.	Omitted: We are choosing to filter out the spoken word tracks, thus this feature is not necessary.
Tempo	Tempo of a track in beats per minute.	Omitted: Closely related to the energy metric.
Time Signature	How many beats are in each bar.	Omitted: Closely related to the energy metric.
Valence	A 0-1 scale. How positive a track is. Tracks with higher scores sound happier, cheerful, euphoric while tracks with low score sound sad, depressed, angry.	Included

2.2 Data Cleansing and Initial Findings

After loading the data into R we began to clean it by checking for missing (n.a.) values, duplicated tracks, and track ids that were not 22 characters in length (per Spotify, the track id should be 22 characters). Next we removed tracks where the "speechiness" scores above .66, because these are spoken word tracks (ie. audio book chapters) and are not useful for our investigation.

In order to remove outliers of song length, we used the interquartile range method which calculated the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of song lengths in the dataset. We defined an observation to be an outlier if it is 1.5 times the interquartile range greater than the third quartile (Q3) or 1.5 times the interquartile range less than the first quartile (Q1). By performing this calculation we only included tracks that were between 48 seconds and 6.11 minutes, which seems to represent typical song lengths. Adding confidence to our method for dropping outliers, this filtered dataset has an average track

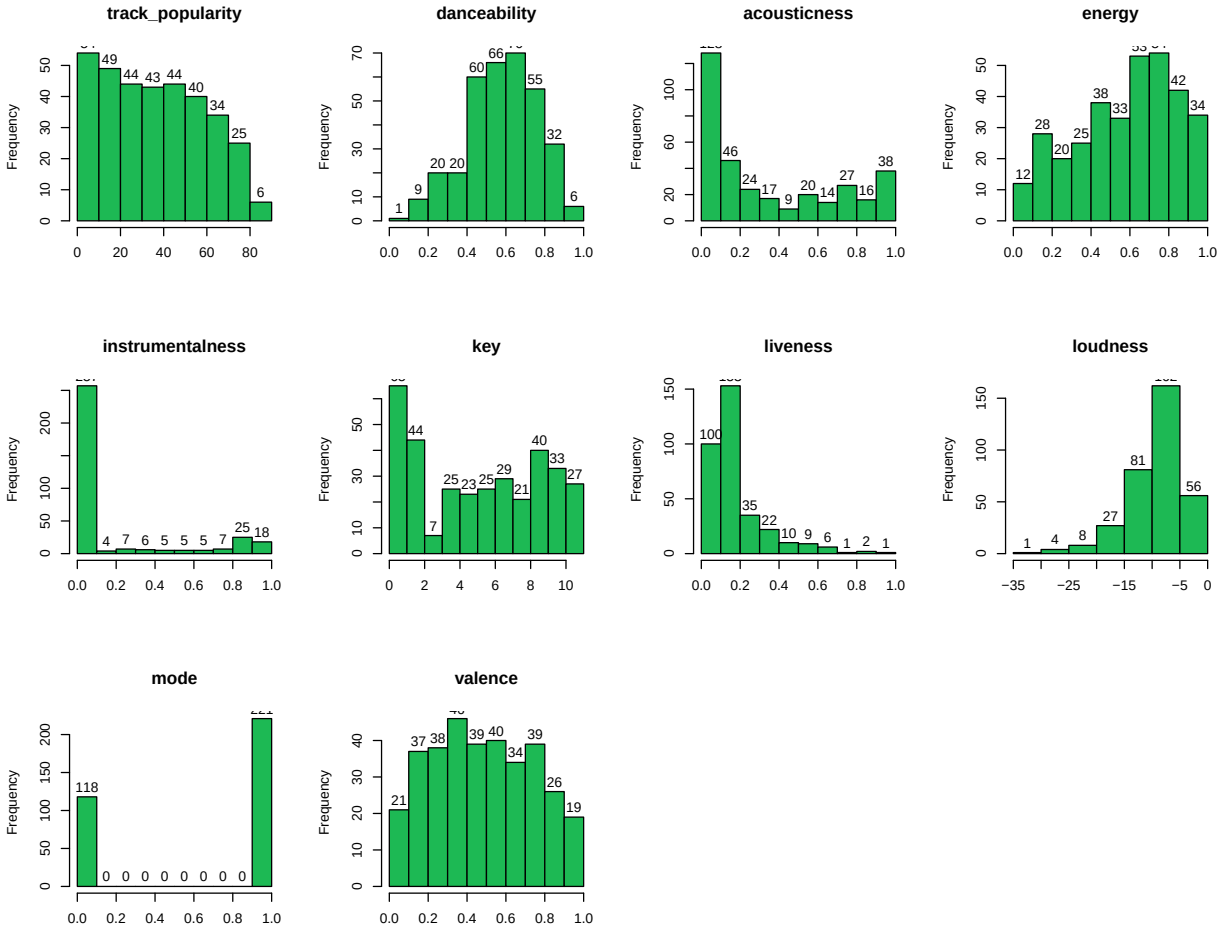
length of ~3.5 minutes, which is the historical average length of popular songs[2].

Finally, we removed tracks that had a track popularity score of 0. Per Spotify, the track popularity is calculated based on number of plays and artist popularity. With a 0 score, the songs are likely to not have been played or have been played very little.

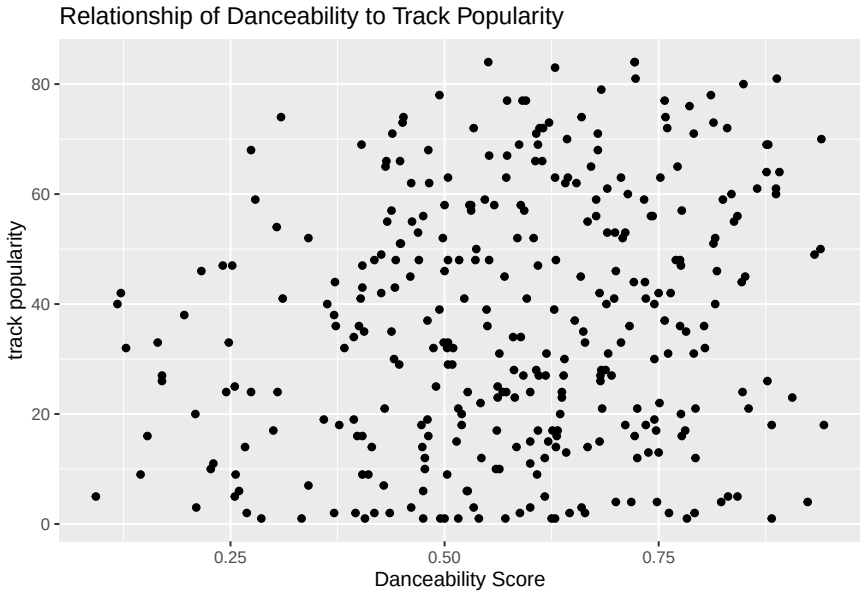
A description of the sample size due to the various data cleansing methods is provided below.

Cause	Number of Samples for Analysis (After Removal for Cause)	Removed Number of Samples for Cause
Original	700	
Speechiness > .66	673	27
Song Length Outliers	613	60
Track Popularity = 0	485	128
Final total	485	

After data cleansing, the distributions of feature scores are as follows. Note that appears to be heavy skewing in the features (liveness, acousticness, loudness, instrumentalness). Log transformations were conducted on these variables, however, it did not result in a more normalized distribution for these features. Thus, log transformations were not utilized for the independent variables.



Finally, as part of the exploratory data analysis, we looked at the relationship between our independent variable (danceability) and our dependent variable (track popularity). There does not appear to be any significant clustering between our two variables of interest.



2.3 Research Design

Our aim is to understand the relationship between the danceability metric and track popularity score. We will focus on danceability as the main independent variable of interest. For further analysis we will also investigate how the following audio features (variables) will also affect track popularity:

- Energy
- Valence
- Accousticness
- Instrumentalness
- Key
- Liveness
- Loudness
- Mode

We hypothesize that danceability will influence a song's track popularity score. Our approach is to start with a danceability conceptual "short" model and create two additional models to determine the best fit model for validating our explanatory analysis. The two other models will consist of Model 2 a control model, Model 3 a full or best fit model, which contains all included metrics in the previously described audio features table. This approach will be sequential and built to each model to limit the risk of p-hacking and overfitting. More details of each model will be covered in the following model section. Additionally, the details of which variables were omitted will be described in the omitted variables section 5.2.

As this research is being conducted without specific domain knowledge of Spotify or the music industry, the research will be conducted in two steps:

Step 1: A sequential model approach for 3 models

Step 2: An exploration and testing data set

Each model will retain the track popularity metric as the outcome variable. The sequential model approach will go through:

1. Model 1 - A danceability conceptual model will be used to verify danceability as an independent variable connected to the dependent variable track popularity.
2. Model 2 - A control model will be used to identify other features related to danceability, based on conceptual understanding of the danceability metric (such as energy and valence variables).
3. An exploration process to build model 3 included an exploratory process to validate the coefficient values of each audio feature included in the data set, excluding omitted variables:

- Testing violations (i.e discovered loudness to be incorrectly included)
- Look at correlation matrix (to validate that there weren't more right hand errors)
- Conceptual models (re-reviewed features and identified that we should remove key because of ordinal value (12 scale)
- Checked the coefficients to confirm statistical significance and variable "power"

4. Model 3 - A full model that will be used from the exploration process to validate the best fit model of the variance (R2).

An exploration and testing data set will be created before research and model creation to validate the learnings from each of the models and help identify any potential issues or mistaken violations in the approach of this research (IDD, conceptual model, p-hacking, and/or overfitting). This train and test will also preserve IDDs and ensure there is no violation of "stopping rules".

3 Modeling

3.1 Model 1 - Danceability Conceptual (Short) Model

In this first model, we are establishing the independent variable of danceability without covariates to test the main hypothesis. How does the danceability score for a Spotify song affect its track popularity?

$$Track\ Popularity = \beta_0 + \beta_1 danceability$$

Based on the results, we determined that danceability has a 24.815 relationship to track popularity which signifies a 0.036 adjusted R2, indicating that 3.6% of the variance in track popularity (outcome variable) is being described in the conceptual short model.

3.2 Model 2 - Danceability Control Model

In Model 2 we adjust the model to ensure we are not inflating the results in Model 1 by adding additional explanatory variables related to danceability. The creation of this new conceptual model is to start with a theoretical approach to ensure no overfitting or bias in relation to the model. The two added variables are energy and valence.

$$Track\ Popularity = \beta_0 + \beta_1 danceability + \beta_2 energy + \beta_3 valence$$

These two variables are selected based on the definitions and the conceptual approach that these two covariates would have a strong relationship to the danceability metric. For example, a high energy or a more positive sounding song may be more likely to have a change to the track popularity and would have an inflation effect on danceability if excluded.

3.3 Model 3 - Full and best fit Model (Long Model)

In model 3, we add non-omitted variables. To test and validate our previous models, non-omitted values are included to confirm accuracy of the conceptual model and to check for any further potential inflation of results. Note that mode is an ordinal (binary) metric, thus factor was utilized within the model so that this variable is not treated as a continuous metric.

$$Track\ Popularity = \beta_0 + \beta_1 danceability + \beta_2 energy + \beta_3 valence + \beta_4 acoustincness + \beta_5 instrumentalness + \beta_6 liveness + factor\beta_8\ mode$$

4 Results

The results of the three models are as follows:

Table 3:

	Dependent variable:		
	track_popularity		
	(1)	(2)	(3)
danceability	24.815***	27.319***	22.075**
	(6.365)	(7.491)	(7.854)
acousticness			−4.458
			(5.539)
energy		9.290	2.432
		(4.746)	(7.518)
instrumentalness			−8.950*
			(3.697)
liveness			−5.192
			(8.310)
factor(mode)1			−0.648
			(2.600)
valence		−7.727	−6.581
		(5.660)	(5.707)
Constant	22.216***	19.089***	29.969***
	(3.736)	(4.268)	(7.465)
Observations	339	339	339
R^2	0.038	0.050	0.067
Adjusted R^2	0.036	0.042	0.047
Residual Std. Error	22.692 (df = 337)	22.620 (df = 335)	22.551 (df = 331)
F Statistic	13.454*** (df = 1; 337)	5.897*** (df = 3; 335)	3.406** (df = 7; 331)
Note:			*p<0.05; **p<0.01; ***p<0.001

In all three models we see a relevant significance of danceability as an explanatory variable to the outcome variable track popularity. Therefore, we will approach model selection by determining the best fit model (as defined by R2 value) that represents the most explanatory coefficients and statistical significance.

Looking at the conceptual model 1 and control model 2, both have strong statistical significance of p <.001. Model 2 also has a stronger explanatory power (0.042 adjusted R2) than Model 1 (0.036). However, based on the results provided, model 3 suggests an inflation of the danceability variable for model 1 and 2 since the scores of the coefficient decrease from Model 1 (24.815) and Model 2 (27.319) to Model 3 (22.075).

Model 3 has a higher explanatory power with an adjusted R2 value of 0.047 compared to model 1's (0.036) and model 2's (0.042). Although model 3's p-value has a higher p value and thus lower statistical significance compared to the other 2 models, we suggest that model 3's practical significance is stronger. Model 1 and model 2 are inflating the effect size of danceability due to the missing coefficients causing the danceability

coefficient to move away from zero.

With model 3 being the best fit the following would be true:

1. Overall .047 for R2 or 4.7% of variance explained.
2. Model 3 corrects danceability inflation. The effect size of danceability is 22.075, indicating that a one unit increase in danceability score increases the track popularity by 22.075 units. Interestingly, all other audio features except for energy are a negative relationship, implying that model 3 shows the more accurate coefficient relationship by likely reducing overinflation.
3. An increase of 2.21 on a scale of 0-100 for track popularity for each .10 unit increase in danceability (scale 0-1) is non-trivial, suggesting that danceability does have an effect on track popularity. In other words if a song had a max score of 1 for danceability, it would result in ~22 points in track popularity.

Notably, only one other feature (instrumentalness) reached statistical significance with a contribution of -0.9 in track popularity score for each .10 increase in instrumentalness score. This intuitively makes sense as most popular songs today contain some singing from an artist and a purely instrumental song would seem to be less likely to be popular.

Looking at the danceability variable, it accurately predicts track popularity with about a 22.551 error on average. More precisely, we can say that 68% of the predicted track popularity will be within 22.551 of the real values.

Overall we find the model to be relevant with danceability being an explanatory variable to track popularity. Danceability effect size changes with each model we evaluated but it remains the highest effect size in all models. Adding more coefficients provides more accuracy to the danceability effect size while maintaining the statistical significance. The explanatory variable is 4.7% or R2 0.047 as described by danceability. It is important to note that the intercept coefficient is high (29.969) in Model 3, meaning if all explanatory variables in the model equal zero, there would be a mean track popularity score of ~30. This suggests that there are important variables related to track popularity that are not included in the best-fit model. These other variables are described below in the limitations and omitted variables section. However, for the purpose of this analysis, due to the interest of Spotify looking to understand a predictor of what songs to recommend and/or play to maintain engagement of users, these results suggest the importance of leveraging danceability in predicting track popularity.

Reproducibility As stated above, the cleansed dataset was randomly split into an exploratory subset of 339 samples (represented in the analysis above) and a testing subset of 146. Testing Model 3 on the test subset for reproducibility did not result in significance for any of the variables except for instrumentalness (coefficient of -10.226). The danceability coefficient was 7.220 in this test set, but did not reach significance. Thus, we were unable to reproduce Model 3 outcomes in this smaller testing subset.

5 Model Limitations

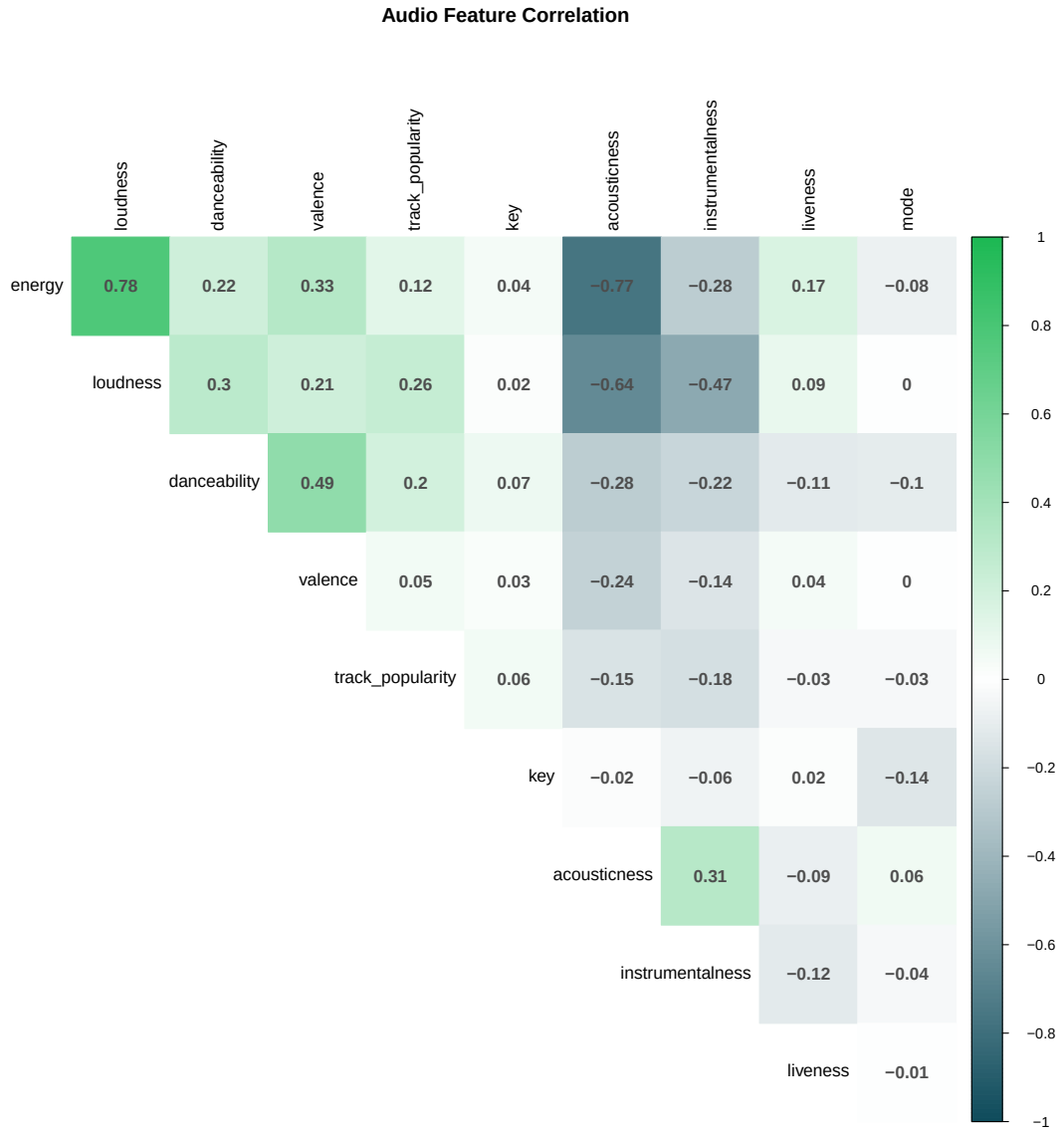
5.1 Large Sample Assumptions

1. Independent and Identically Distributed (I.I.D.) Data:

The data we are working with satisfies I.I.D. because the randomized form of gathering the tracks allowed our data to be independent of each other. By ensuring that there were no duplicate track ids and using the randify web application we were able to satisfy the I.I.D assumption.

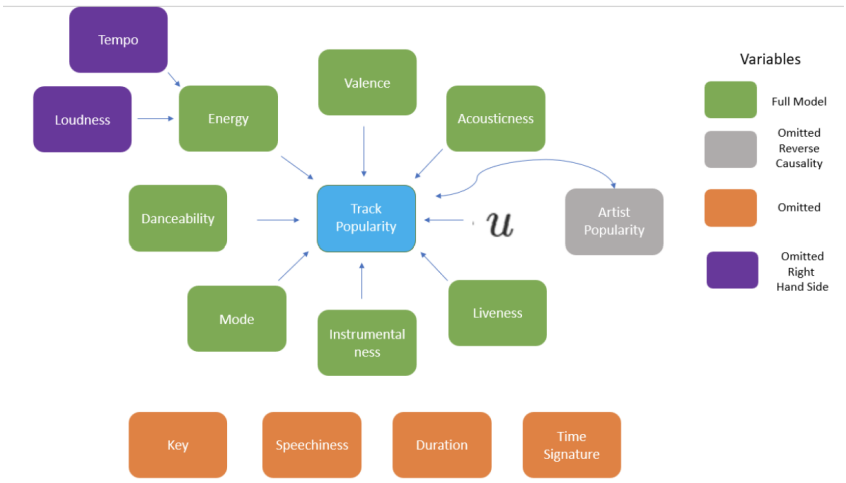
2. Unique BLP Exists:

To assess whether a unique BLP exists we must first check that there is no perfect collinearity. Our correlation matrix below indicates that there is no perfect collinearity:



5.2 Structural Limitations - Ommitted Variables

Through our research and reviewing the data we identified a few variables that needed to be omitted that could conceptually impact results of the model. The ommitted variables are outlined in the causal diagram below:



Intentional Ommitted Variables:

1. Duration & Speechiness:

Duration and Speechiness were utilized as a filter in the data cleansing process. Since speechiness measures the spoken words within the track and not necessarily singing, we felt this variable was not relevant to songs. For song duration, we filtered out outliers and the final distribution of song duration seems consistent with standard songs. Conceptually, it is challenging to justify that a longer song would be more popular over a shorter song, thus we did not think it was a relevant variable to include in the model.

2. Tempo, Loudness, & Time Signature:

We chose to omit these variables because they related to closely to the energy metric. If these were to be included in the model, it would result in a right hand side error, potentially complicating the model outcome interpretation.

3. Artist Popularity:

We recognize that Artists that have a larger following will likely have a higher track popularity score for their songs. Thus, the artist popularity would be an important variable to consider in a model. Spotify API does provide the Artist Popularity score for each song by the artist, however, we found that the Track Popularity metric uses the Artist Popularity score in the calculation. Including the Artist Popularity variable would cause a reverse causality violation so this was intentionally omitted.

Unintentional Ommitted Variable:

1. Track Genre:

Track Genre was a variable missing in the Spotify API that would anticipate making an impact on the model. However, Track Genre is not an available song feature within the Spotify API. Top 100 tracks in the US on Spotify tend to consist of Pop, Rap and R&B songs as opposed to Folk or Jazz genres, so this could be introducing an appreciable omitted variable bias into our model. If the track genre is Pop, Rap or R&B this could have resulted in a higher track popularity, thus this omitted variable bias is positive / away from zero.

2. Reccomender Algorithms:

Spotify uses many ways to recommend songs to users, for instance it suggests newly released tracks from artists the user listens to, it creates albums curated to the user's specific genres of music they enjoy, and it even provides a "hits" only section that enables users to only listen to popular tracks. How this algorithm impacts a song is not a feature that is included in the Spotify API. This recommender algorithm would

introduce bias due to users not being exposed to songs in a randomized manner. Thus, songs may have a lower or higher track popularity score based on the impact of the algorithm.

6 Conclusions

The goal of this analysis was to understand the relationship between a song's danceability score and the track popularity in Spotify. We obtained song features from the Spotify API for a random sample of songs in order to address this research question. After data filtering and ensuring independence in our dataset, the number of samples in the final dataset was 485. This final dataset was randomly split into an exploratory subset of 339 samples and a testing subset of 146.

After building three models in the exploratory subset, we conclude that the track popularity is impacted by the danceability score in Spotify with an increase in a song's danceability by .10 units resulting in an increase of the track popularity by 2.21 units, with all other features being equal. The best fit model (Model 3) included all reasonable ordinal/continuous song features, but only one other feature (instrumentalness) reached statistical significance. Notably, the intercept of the model was high, indicating that there may be other features impacting track popularity that are not represented in this model.

After identifying Model 3 as the best fit model, we re-ran it utilizing the smaller test subset containing 146 songs randomly sequestered from the original cleansed dataset. We were unable to reproduce the outcomes of Model 3 on this subsequent test set. We hypothesize that the lack of reaching significance in the test set is a result of a small dataset. Rather than violate "stopping rules" to add more data to the test set, we recommend that this research design be redone using a larger sample size to determine confidence in reproducibility.

In addition to being unable to be reproduced, based on conceptual knowledge of the music industry, Model 3 is likely missing important features such as the artist's popularity and the genre of the music that would have a significant impact in the track popularity. However, it would still be of interest for Spotify to consider the role of the danceability metric when optimizing for track streams. Based on this analysis, songs having higher danceability scores would be more likely to be popular with all else being equal.

This analysis suggests that the danceability score for a song plays an appreciable role in the track popularity score. While this could be utilized by stakeholders at Spotify to further optimize song recommendations to encourage increasing streams, there are limitations to this analysis. Future analyses would benefit from incorporating additional variables related to the artist's popularity and the song genre and controlling for the impact of the recommender algorithms, which would introduce further robustness to a model.

7 References

1. https://s22.q4cdn.com/540910603/files/doc_financials/2021/q4/0307a021-254e-43c5-aeac-8242b0ea3ade.pdf
2. https://www.digitalmusicnews.com/2019/01/18/streaming-music-shorter-songs-study/