The Differences in Desired Audio Features Between 3rd and 4th Generation Kpop Groups*

Data mainly incorporates international input

Oluwabusayomi Adekuajo

29 April 2022

Abstract

Kpop has gained rapid popularity globally which is evident through setting new records and receiving public recognition on an international scale. This paper analyzes the possible changes in musicality represented in the shift between 3rd and 4th generation Kpop idols. Finidngs display that 4th generation music is louder, but also has less energy. Understanding current and evolving trends in this economically growing industry creates a plotform to predict popularity in groups to debute in the upcoming 4th generation.

Keywords: music analysis, audio feature, 3rd generation kpop, 4th generation kpop, spotify, generational change

Contents

*Code and data are available at: https://github.com/rubyzero10/Final_Starter_Fold.git.

1 Introduction

K-pop has become a worldwide phenomenon with its groups constantly breaking new records on a global scale. What was once known exclusively in South Korean and Asian countries has found popularity internationally, known as the Korean wave, which is continuously growing. The timeline of this growth can be monitored through generations, for which there are currently four. The First Generation, which dates from the 1990s to the mid 2000s targeted mainly domestic and neighboring youth markets and consisted of family owned companies (Doré and Pugsley (2019)) essentially it was the founding of such culture. During the second generation, things started to expand outside Korea as YouTube became popular in the early 2010s and international fans could be reached. As the third generation consists of debuted groups from 2012 to 2018, this allowed for the starting target audience to be on an international scale. This was demonstrated with the involvement of social media platforms and increase of streaming availability. Represented by those who have debuted from 2018 to present day, the 4th generation is currently the peak of K-pop, with the its fanbase reaching national and international scales.

Other than just the dates of their debut what separates generations of K-pop. Over the course of time, trends in music become distinguishable. In this accelerating cycle, younger listeners, not only adopt particular songs or genres as badges of distinction and identity but also possess the discretionary income that will allow them to express their preferences and became a significant sector in the economy of music's consumption (Ku (2015)). In this paper, we analyze the title tracks of the 3rd generation and 4th generation K-pop groups which are amongst the most followed and popular on Spotify. It is assumed that these groups will have the largest fan bases in the community, contributing the most to the economy along with attracting new fans to the genre as a whole. Top 20 K-pop groups were presented in data collected directly from Spotify, from whcih, 3rd and 4rd generation groups were selected. Other popular groups were added to total observable groups. The goal of the study was to determine the audio features that could predict whether a group was in 3rd or 4th generation. We find that, generally 3rd generation groups have softer music that is more energic and in key G. As 4th generation is suggested to continue till 2025, implications of this study can help groups that will debut within timeframe of interest develop music that will be successful in the market.

The remainder of the paper is split into the following sections. Section 2 explains data source and collection, along with characteristics of selected data Section 3 discusses the methods used to obtain final model investigating the transition of music between generation of K-pop. Section 4 presents the findings from our data analysis and model. Section 5 attempts to understand the results as well as explain the weaknesses and limitation in interpreting findings of the paper. It also showcases next steps to enhance and expand on the findings of the report

2 Data

Data was obtained through the Spotify Web API for developers Thompson et al. (2021). Analysis of report was done through the use of the R statistical programming language R Core Team (2020), along with other packages, ggplot2 Wickham (2016), knitr Xie (2021), kableExtra Zhu (2021), tidyr Wickham and Girlich (2022), dplyr Wickham et al. (2021), modelsummary Arel-Bundock (2022), tidymodels Kuhn and Wickham (2020), patchwork Pedersen (2020) and tidyverse Wickham et al. (2019), that aided to clean and manipulate dataset. This report uses 194 observations of 17 variables from selected popular tracks' audio features.

2.1 Data Source

The Spotify Web API for developers is known as a RESTful API, or a Representational state transfer API. It is an interface that allows for the retrieval and management of data using a standard HTTP protocol to access internet available data. Using REST principles, the Spotify Web API provides a set of endpoints each with its own unique path, that returns metadata in regards to music artists, albums, tracks, playlists and saved music directly from the Spotify Data Catalogue. As Spotify information is protected, authorization is required to access Spotify data and features. A Spotify account, along with a Spotify Developer account are used to inquire a 'Client ID' and a 'Client Secret,' which are needed in order to be permitted to access the Spotify API.

2.2 Strengths

The Spotify Web API allows for huge amount of available data to be collected. Simply through making calls through the Web API any information regarding an artist, track, playlist, genre, and much more can be received. With the ability to gather such a multitude and diversity of music data, there is an endless amount of research that can be done. The process to obtain such data is quite simple as it does not require much effort; access to data is also free. As data is obtained directly through the Spotify catalog it is not prone to measurement error. Spotify is one the most used music platforms, making data extremely relevant as it is the direct output of its users. Variables like followers and popularity represent updated real-life information of the demand of an artist, allowing implications of the study to affect not only current artist, but upcoming artists.

2.3 Weaknesses

While Spotify may be the most popular streaming service it is not available everywhere. The main interest of this report involves Korean popular music. Though K-pop is a growing on an international scale, the main target markets include South Korea itself, China, Japan, along with other neighboring Asian countries. Spotify however had only launched in South Korea as of Feb 2, 2022 and is still not available in China. This indicates that Spotify is not the main source of music streaming in the country that it involves along with one of the biggest markets in Asia. Thus, this data does not capture the true population of listeners, and mainly represents an international viewpoint.

2.4 Data collection

After access was acquired, the appropriate HTTP verb used to retrieve resources was GET. As the main music genre of interest was Korean pop, the string parameter "k-pop" was imputed and a list containing 20 observations of K-pop artists and 12 variables of information on the artists were generated. Only K-pop groups that are considered to be 3rd or 4th generation were of interest, and such all artist in list that did not meet these requirements were removed, this included solo artists as well. This reduced the list to involve 16 K-pop groups, specifically 9 boy groups, 7 girls groups, of which half were split into 3rd and 4th generation.

In order to get an even amount of boy and girl groups an additional two 3rd generation girl groups and one girl and boy group part of the 4th generation was added. These additional artist were chosen through research, and searched on Spotify.

In order receive a response from the API the following parameter, Spotify ID, which is a base-62 identifier found at the end of a Spotify URI, was used to obtain relevant information regarding researched artist of interest. Relevant information was extracted from artist information and added to previous K-pop list. This list only contained information about the artist, like their followers, id, etc., however data regarding musical requirements of a song was the interest of study. Thus the top tracks for each artist was obtained along with the tracks audio features using GET verb.

2.5 Variable selection

Many of the variables in the first list, containing information about the artist, provided information regarding location of artist on Spotify, like their id and url, which was not relevant to study. Thus only the artist name, popularity and followers were selected. As the groups in the list separate between gender and generation, which was not available through Spotify Web API, additional columns were manually constructed and labeled accordingly. Figure 1 shows the distribution of popularity of K-pop groups based on generation. It can be seen that 3rd generation groups occupy the ends of the graph, being the most and least popular on the Spotify platform, while 4th generation group are mainly in the middle. Figure 2 shows the distribution of Spotify followers which has a much larger overall difference between most followed and least followed. The most followed groups are 3rd generation while some of the least followed groups are of 4th generation. In fact, it can be seen that almost all 3rd generation groups have more followers the almost all 4th generation groups. This could be possibly due to the later debut date of 4th generation idols and a represent a growing fandom. The second data list contained information on the audio features of top tracks along with information of track location. Interested in understanding what defining audio features separate 4th generation from 3rd generation, the danceability, energy, key, mode, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, name, and popularity of top tracks were selected, along with album release date to compare any changes of features of music. Distributions of each variable is shown in Table 1

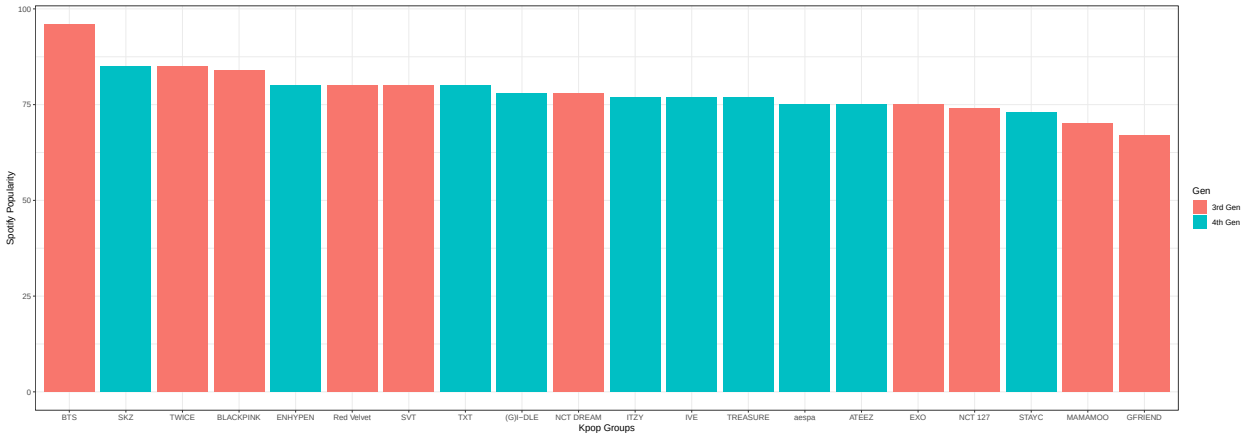

Figure 1: Distrubution of Spotify Popularity

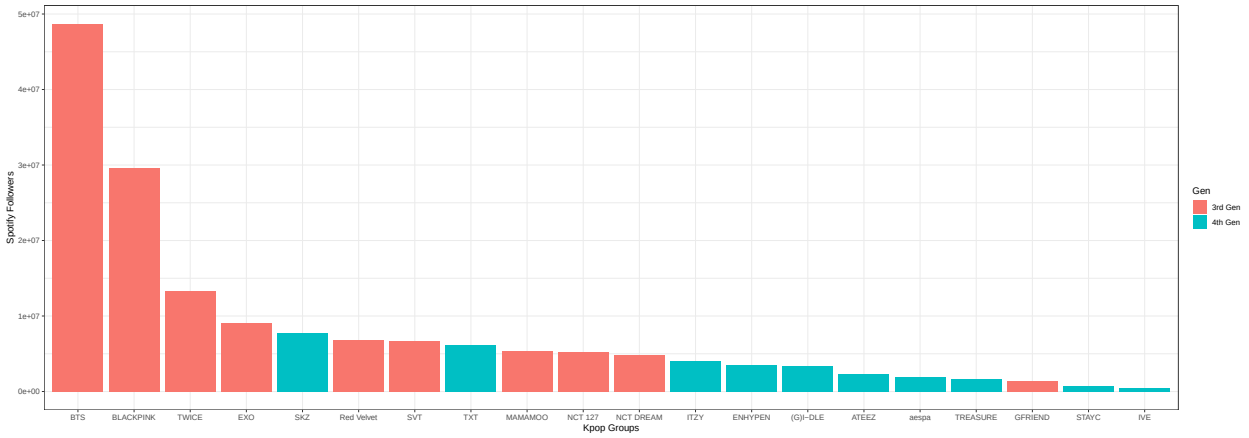

Figure 2: Distrubution of Spotify Followers

Table 1: Summary of audio features ditrubution

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
danceability	146	0	0.7	0.1	0.4	0.7	0.9	
energy	151	0	0.8	0.1	0.5	0.8	1.0	
loudness	191	0	−3.9	1.5	−8.7	−3.8	−0.2	
speechiness	171	0	0.1	0.1	0.0	0.1	0.4	
acousticness	184	0	0.1	0.1	0.0	0.1	0.7	
instrumentalness	53	0	0.0	0.0	0.0	0.0	0.0	
liveness	170	0	0.2	0.1	0.0	0.1	0.7	
valence	169	0	0.6	0.2	0.1	0.6	1.0	
tempo	192	0	124.5	25.4	73.0	123.0	192.1	
popularity	35	0	68.6	7.3	50.0	67.0	87.0	

Table 2: Discribtion of Audio Features

name	discribtion
danceability	A value of 0.0 is least danceable and 1.0 is most danceable.
energy	Measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity
key	The key the track is in using standard Pitch Class notation.
mode	Mode indicates the modality (major or minor) of a track. Major is represented by 1 and minor is 0.
loudness	The overall loudness of a track in decibels (dB) averaged across the entire 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
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.
tempo	The overall estimated tempo of a track in beats per minute (BPM).

3 Model

3.1 Logistic regression

The final dataset allowed for the investigation of what musical factors contribute to the separation of K-pop generation, in specific 3rd and 4th generation groups. As the variable of interest is binary, the regression model used was logistic. What we will get out of this regression is the probability of whether a group is 3rd or 4th generation based of the predictor variables. We are interested in estimating the equation

$$Pr(yi = 1) = \beta_0 + \beta_1 x_i + ... + \beta_j x_{ij}$$

where,
- yi is the response variable,
- xij are the possible predictor variables,
- β_0 the 'intercept,'
- β_1 is the 'slope'

3.2 Assumptions

Before fitting any models, it was important to perform an exploratory data analysis in order to investigate the distributions of all variable. This was done through plotting each predictor variable seen in Table 2 against the response, generation of idol. This allowed for potential issues in the data set to be observed, as well as insight into possible predictor variable to include. As we are interested in fitting a Logistic Model, residual plots were not necessary, instead we were interested in checking the follow assumptions, dependant variable was binary, the observations to be impendent of each other, and there to be little or no multicollinearity among the independent variables. As the dependent variable is generation which can either be 3rd or 4th generation it is a binary variable. Multicollinearity was observed during model selection, focusing on achieving smallest AIC to maintain assumptions.

As we are focused on prediction, if we fit are model with all observations, it is possible to over fit data. In order to alleviate the limits on claims made, we divided out dataset into two independent datasets, the training set, which goes through the selection process, and the test set, which we will use to evaluate and validate the performance of produced model. In order to select the final model, we started with a model with one predictor and added variables one at a time and compare AIC to view multicollinearity. Variables to add were determined through results of relationship to response.

4 Results

In order to determine what music features differ to define each generation, it was first necessary to take a look on the popularity of songs produced by each generation. Figure 3 shows the distribution of song popularity divided by generation. It can be seen in each generation, most of the song are in range 60 - 70 in popularity. 3rd generation occupies the most and least popular songs while 4th is in the middle. This pattern was observed earlier with popularity of groups, indicating the likelihood that popular groups make popular songs. This is further seen through Figure 4 as the groups with the more popular songs roughly correlates to the groups seen in Figure 1 that are more popular. Due to fact the 4th generation has had less time to grow then 3rd, Figure 4 does not provide information of specific time period of popularity, only showcasing overall popularity. Figure 5 helps against this problem by comparing song popularity to the release date of the album it was a part of. First we can see how 3rd generation groups have much more time to build the popularity of their songs. In 2019, the first record of a 4th generation song in this dataset, the average song popularity of 3rd gen can be seen above 4th generation. This is the result of a stead increase of popularity from their first record of song seen in 2014. Continually looking at the 3rd generation information, they have an increasing pattern observable until 2020 where they peak, and start to decrease in average popularity. 4th generation, however, has a steadily increasing pattern until 2021 and then a sharp increase in which they pass 3rd generation in popularity in 2022. This indicated a demand for 4th generation music from 2021 on wards.

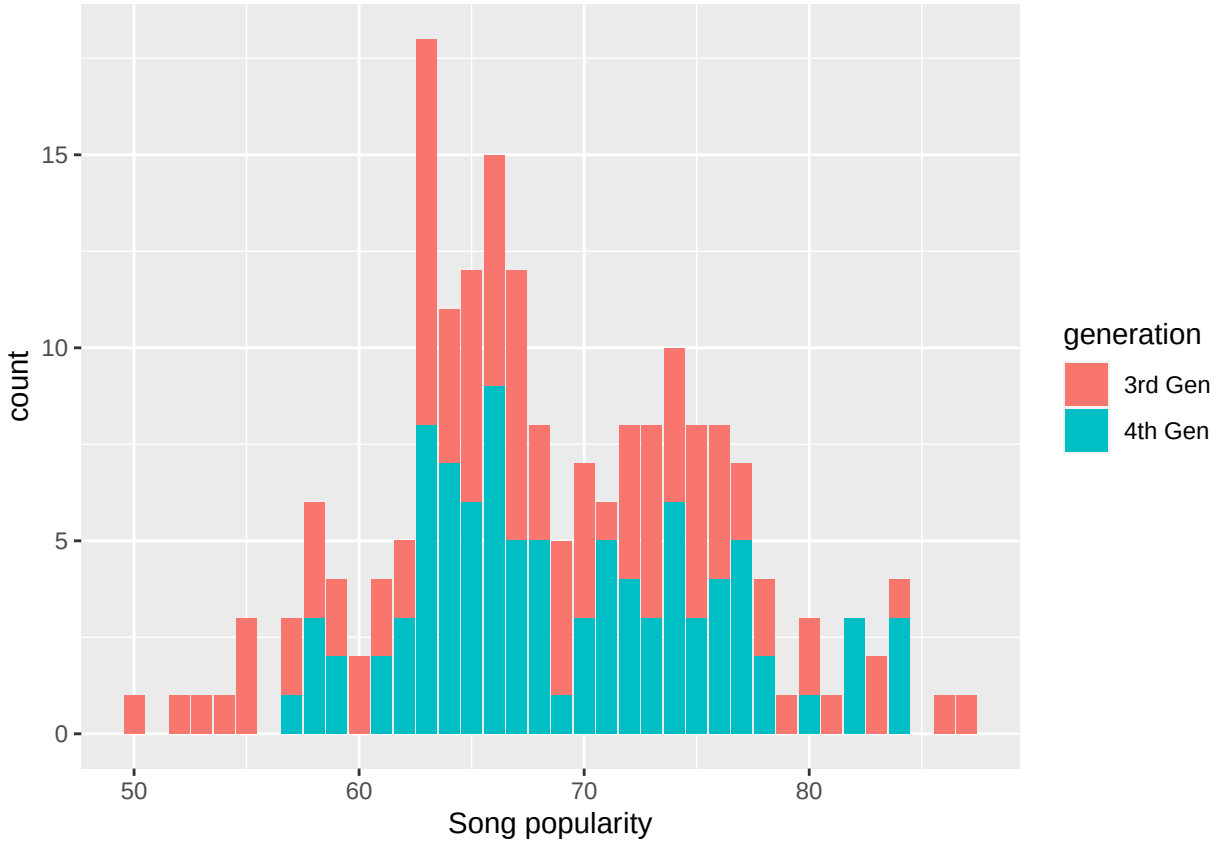

Figure 3: Distribution of song popularity in generation

Now, with the interest of making our model, audio features were plotted against popularity to see if there are any discernible patterns different between generations. Figure 6 containing the continuous audio features, show multiple graphs with same direction and slope. However, with further inspection, loudness speechiness,

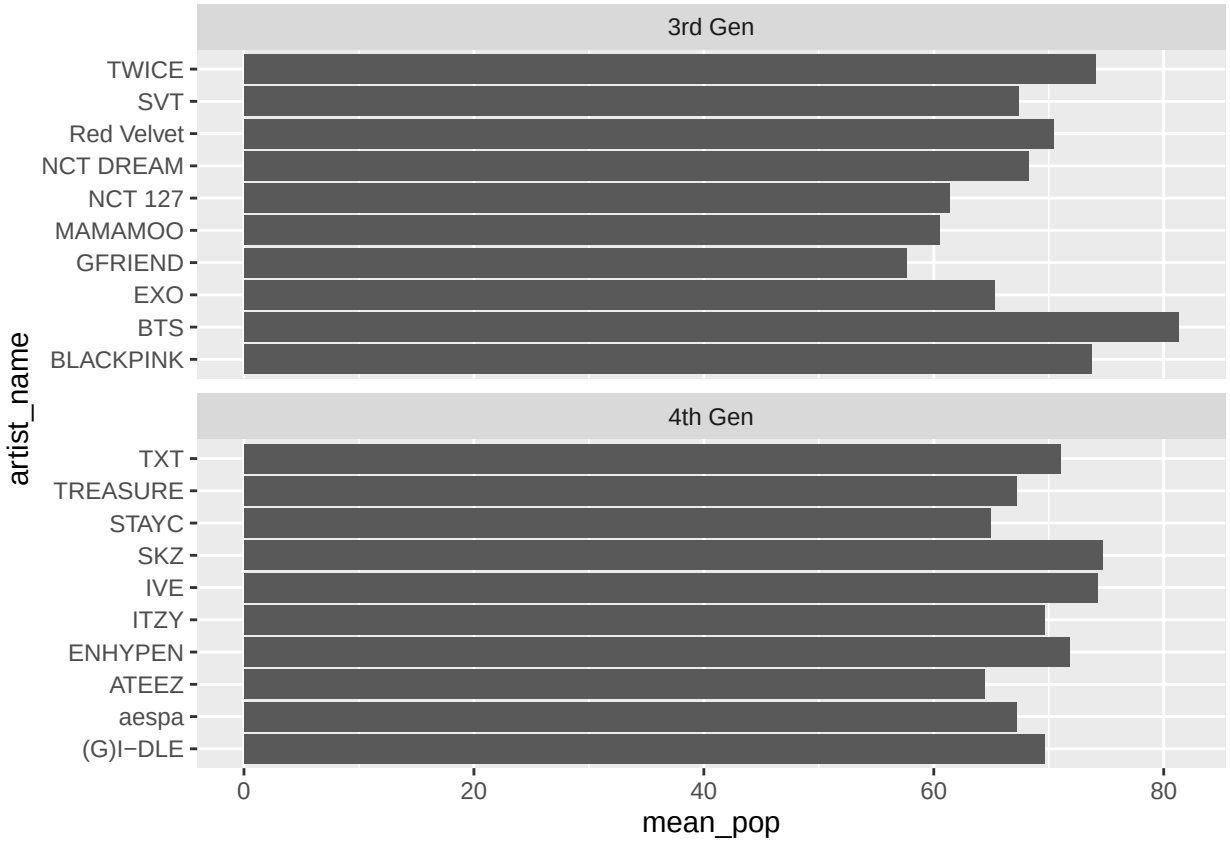

Figure 4: Popularity of song based on group

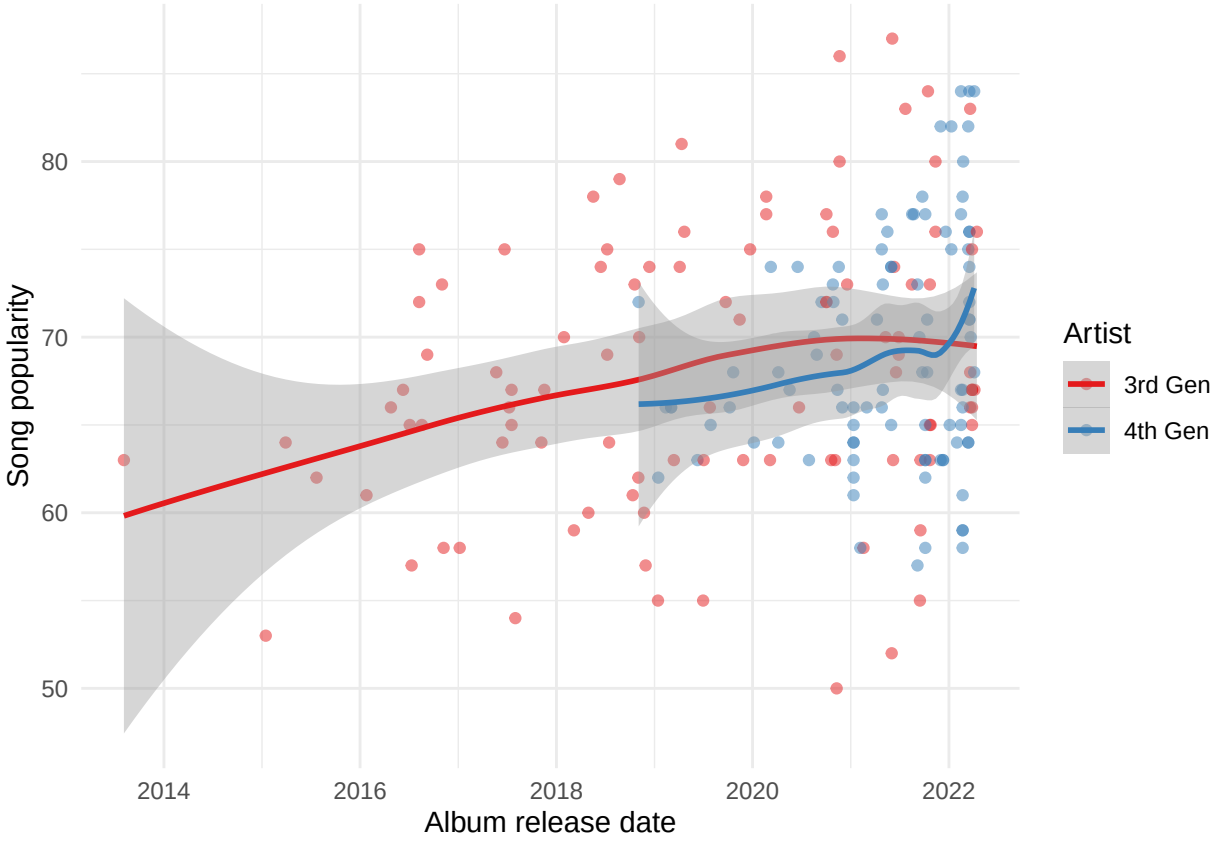

Figure 5: Popularity of song based on time

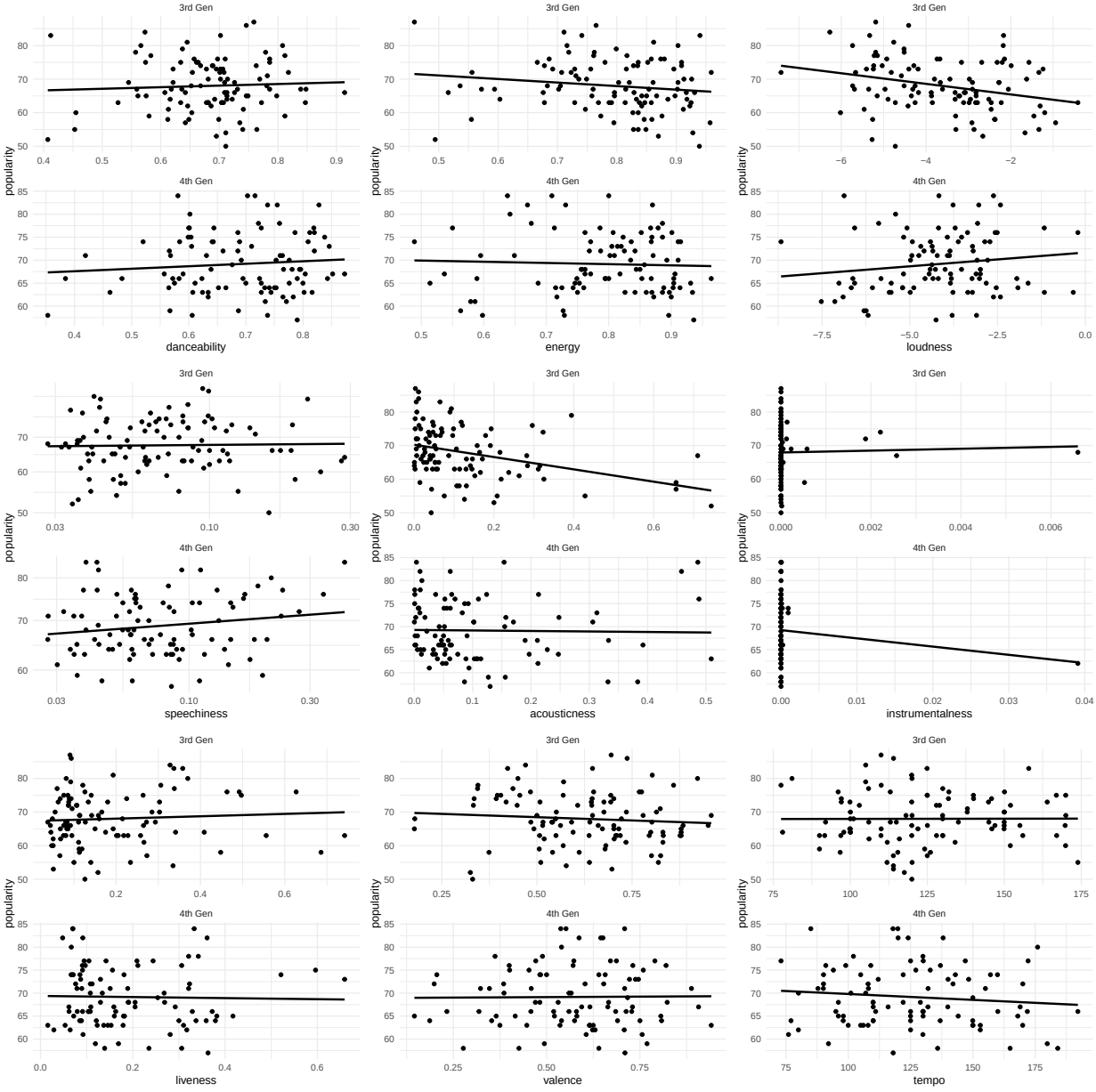

Figure 6: Audio features against popularity

and liveness have opposite patterns. Loud 3rd generation music is not popular, while loud 4th generation is. The more appearance of liveness and speechiness does not seem to affect 3rd generation groups, however the present of live audio in 4th generation groups leads to less popularity of song, and the presents of speech in a song leads to more popular 4th generation music. While these results relate to popularity, further analysis on relationship of possible predictors was necessary to see if there was an observable difference over time. Figure 7 shows that over time loudness, energy, valence, and liveness of audio have decreased while tempo and speechiness have increased. Other variables do not see to have changed, indicating a possible importance in kpop in general that does not change regardless of time.

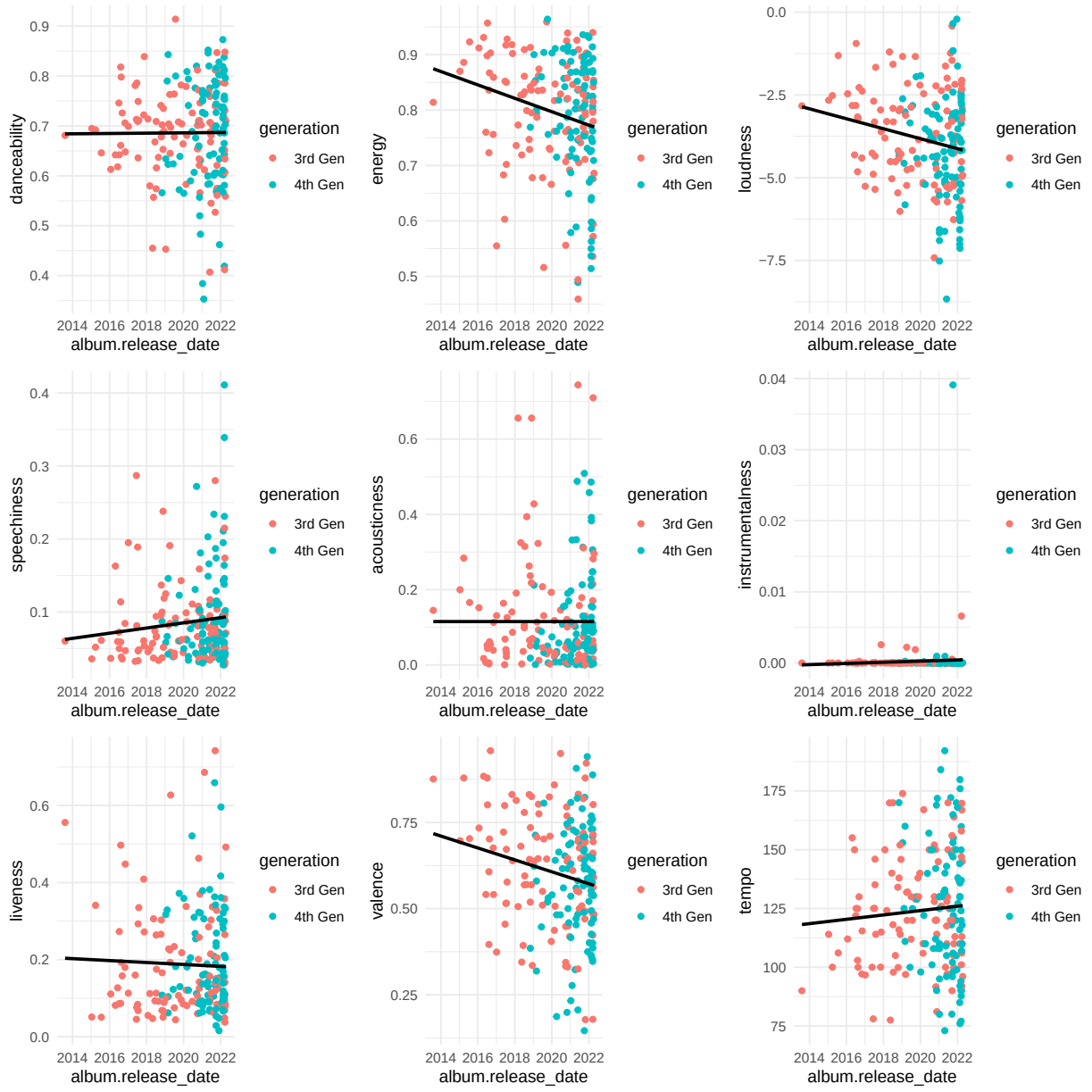

Figure 7: Audio features against time

With the previous information the first predictor variable of interest was loudness, as there is clear dependence on time as well as generation. After a model was fit with single variable, multiple other models containing

Table 3: Confusion Matrix

.	Truth	..
Prediction	3rd Gen	4th Gen
3rd Gen	15	6
4th Gen	4	14

Table 4: Final Model Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-62.41	11.94	-5.23	0.00	-88.28	-41.15
loudness	-0.57	0.20	-2.90	0.00	-0.98	-0.20
energy	10.83	3.11	3.48	0.00	5.07	17.35
key1	1.05	0.83	1.27	0.21	-0.56	2.72
key2	0.54	0.98	0.55	0.58	-1.41	2.50
key3	2.20	1.26	1.75	0.08	-0.23	4.78
key4	2.09	1.12	1.87	0.06	-0.01	4.45
key5	1.45	1.23	1.18	0.24	-0.95	3.92
key6	1.16	0.90	1.28	0.20	-0.59	2.99
key7	2.92	1.07	2.72	0.01	0.94	5.23
key8	1.96	1.00	1.96	0.05	0.06	4.03
key9	-0.03	0.86	-0.03	0.97	-1.74	1.68
key10	-0.08	1.11	-0.07	0.94	-2.42	2.06
key11	0.38	0.91	0.42	0.67	-1.41	2.21
valence	-1.96	1.30	-1.51	0.13	-4.61	0.53
album.release_date	0.00	0.00	5.06	0.00	0.00	0.00

additional variables were added and compared to first model. In figure Table 4 the final model suggested was generation ~ loudness +energy + key + valence + album.release_date and Table 3 shows the predicted values in comparison to actual results. The proportion of the estimatation that accurately predicted that the group was in 3rd generation was 79% while the proportion that accurately predicted that the group was 4th gen was 70%. Total efficiency of model, or percentage of the times that the test give the correct answer, was 74%. This model is a good start in predicting generation between 3rd and 4th, however, can be developed further to increase efficiency

5 Discussion

5.1 Age and Generation Interest

The final logistic model results show we can predict whether a group is 3rd or 4th generation based of the loudness of the music, the energy, key it is in, valence, and album release date. The music features shown to significantly predict 3rd generation music is softer with more energy, and often is in G major. This implies that 4th generation music, in comparison to 3rd generation, is much louder with less energy and not so often in G major. Exposure to loud music, especially by young people, has significantly increased in recent years as a result of advancements in technology in terms streaming of music through smart devices (Manchaiah, Zhao, and Ratinaud (2019a)). As 4th generation groups debut later, the average age is much less than that of 3rd generation, assuming they debut at the same average age. Since 3rd generation groups consist of individuals that have debuted from 2012-2016 it likely their fandom has grown with them, which is why the fandom of 4th generation groups can also be assumed as younger. Loud music has the ability to hijack other senses and especially the ability to think. Thus younger individuals tend to use loud music as a means of escape and to repress intense feelings and emotions and escape. Older individual also have means for wanting to repress undesired emotions, the difference however is the outlets presented to them. Loud music is a stimulate that can affects your body temperature and heart rate. Adults will tend to prefer stimulates like caffeine, alcohol, and sometimes other substances in order to achieve same results, these, however, are not available to those that are not of age. It is possible that preference of loud music by fandom is due to age represented and companies of 4th generation groups are aware of such and try to promote to receive popularity. This may also explain why 3rd generation music that is loud is not popular and 4th generation music is.

5.2 Energy and Generation

It was observed that in comparison to 3rd generation, 4th generation music has less energy. Studies show that depressed teens and adolescents spend a lot of their time listening to music. Rather than being the cause of mental illness, depressed adolescents rely on sad music for comfort (Manchaiah, Zhao, and Ratinaud (2019b)). Without a lot of perceived energy to anything, music that relays internal feelings can help provide a relatable outlet and results in momentary relive from depression symptoms. Depression can exist among individuals of any age, and it is not unique to adolescent, however, teenager have less of an outlet to relieve symptoms. Mentioned earlier, many adults rely on substance to provide momentary happiness; these substances are not available for younger individuals. Access to professional care is paid service that young individuals cannot afford by them self as well as symptoms of depression are often dismissed as just being a teenage by guardians and supervisors. The younger audience of 4th generation groups may explain why they tend to offer music that is less energy to suit and comfort fans.

5.3 Dancibility and Generation

The danceability of an audio track was seen to not to have significance in determining generation. It also was seen not to change throughout time, or influence popularity. The reason for such could be that danceability is a requirement for K-pop itself. The audio features selected were the top tracks of the artist, these are

likely to be the promoted tracks or title track of the album. In the K-pop community, a title track will have a music video along with performance. In South Korea there are three large national television networks KBS, MBC, and SBS. SBS has Inkigayo, KBS has Music Bank, MBC has Show, which are popular music programs in which idols will perform their title tracks, along with other songs on the album they are trying to promote. Most performances consist of dancing, with only a few sit down singing acts. With such popularity in performance, danceability of a promoted song would be considered important regardless of generation and explains the results seen why danceability was constant over a multitude of comparisons.

5.4 Weaknesses

Data mainly includes international inputs and thus cannot represent what the South Korean and Asian markets bring to the popularity of track. Tracks selected for observation were top tracks of artist, which may differ depending on popularity in region, and thus represent different features. As well popularity was not considered in the final model but may possibly influence what determines a third and fourth generation group. As new features are being preferred it is possible that dislike for older features exists, which is evident as 4th generation music has become increasingly popular in recent year, taking over 3rd generation just this year. Interest in determining 5th generation groups from this dataset may prove difficult as Asian markets now have access to Spotify which will result in the increase of use in upcoming years. As kpop has become more global, international input becomes more important, but this in no way alleviates the importance of their original market. A logistics regression can predict a binary variable, meaning the results of this study cannot predict what is a 5th generation group. Result, though, can help interpret what features are being preferred in recent years to apply to the debut of future generation K-pop idols.

5.5 Next steps

In response to these limitations, future steps can be taken to measure features of popular K-pop songs in Asian countries and compare with result from present research. In order to get a full interpretation of the change of musical requirements through generation, a study of inactive first- and second-generation groups can be done to further help predict what will succeed in future. What does well in the past does not always do well in the future, but forgotten trends to tend to repeat itself. Additional studies of repeated trends, if any, can be done in order to observe patterns in generational music. Comparison to western artists with the similar debut years may help to determine preferences between two music styles, and predict future deviations.

Appendix

A Datasheet

Extract of the questions from Gebru et al. (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to establish the which audio features present to define generation in Kpop
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The data was created by Oluwabusayomi Adekuajo through the use of Spotify API.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The Spotify Web API is a free service
4. *Any other comments?*
    - NA

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The instances that the data set represent are top audio track features there are 155 instances
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 550 instances of each type oftop audio track features
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The dataset does not contain all possible instances as some are not as related to data collected such as URI of audio track
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - Each instance consisted of raw data fro Spotify catalog
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - There is not target associated with each instance
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- There is no data missing from individual instances

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo which was another instance.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - The recommended data splits includes to split the training set into 80% of the ordingal data to avoid over fitting data.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There is not errors, sources of noise, or redundancies in the dataset
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - Data is linked to external resources,
    - a) It is likely to change overtime as it contantly collects user data
    - b) There are no official archival versions of the complete dataset
    - c) Restriction include you much have a Spotify developers and Spotify account as authentication is needed.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no data that might be considered confidential in dataset
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - There is no data that might be offensive, insulting, threatening, or might otherwise cause anxiety
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - This dataset has sub population divided by gender and generation. Gender include BG or boy group, and GG or girl group as well as 3rd and 4th gen identifiers.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It is not possible to identify individuals.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - Data contains no data that might be considered sensitive in any way

16. *Any other comments?*
    - NA

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data was directly observable
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - Data acquired through use of Spotify Web API
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - NA
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - Data is collected through Spotify catalog and no one is paid
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data used has information from 2014 - 2022
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No ethical review processes
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - Data was obtained through Spotify Web API with codes on how to do so on Spotify developers website
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - Notice of collection is displayed in terms and service of Spotify users.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - Individuals consent was collected through accepting Spotify terms and service.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - It was not necessary to accept Spotify terms and service, if not planning to use Spotify
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - An analysis of the potential impact of the dataset and its use on data subject was not conducted
12. *Any other comments?*
    - NA

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Data was processed as PDF file and converted using R into a usable data fram
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Raw data is saved in inputs/data/raw_data_popular.csv and inputs/data/raw_data_audio.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Yes R software used to clean the data is available at https://www.r-project.org/
4. *Any other comments?*
    - NA

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has not been used for any tasks already
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - https://github.com/rubyzero10/Final-paper-kpop.git
3. *What (other) tasks could the dataset be used for?*
    - The dataset could be used to see what features determine popularity of track.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - There is nothing about the composition of the dataset that might impact future uses

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - There are no tasks for which the dataset should not be used
6. *Any other comments?*
    - NA

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No the dataset will not be distributed to third parties outside of the entity
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be distributed throguh Github
3. *When will the dataset be distributed?*
    - The dataset will be distributed in April 2022
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - Dataset will not be distributed under a copyright or other intellectual property (IP) license
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - Third parties have not imposed IP-based or other restrictions on the data associated with the instances
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No export controls or other regulatory restrictions apply to the dataset or to individual instances
7. *Any other comments?*
    - NA

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Oluwabusayomi Adekuajo
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Via Github
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- Dataset wil not be updated

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - There are no applicable limits on the retention of the data associated with the instances
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Older versions of the dataset will not continue to be maintained?
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - There is no mechanism for other to contribute to the dataset.
8. *Any other comments?*
   - NA

B Additional details

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://CRAN.R-project.org/package=modelsummary.

Doré, Philippa, and Peter C Pugsley. 2019. "Genre Conventions in k-Pop: BTS's 'Dope'music Video." *Continuum* 33 (5): 580–89.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Ku, Robert Ji-Song. 2015. "K-Pop: Popular Music, Cultural Amnesia, and Economic Innovation in South Korea, John Lie (2014)." *East Asian Journal of Popular Culture* 1 (2): 303–8.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Manchaiah, Vinaya, Fei Zhao, and Pierre Ratinaud. 2019a. "Young Adults' Knowledge and Attitudes Regarding 'Music' and 'Loud Music' Across Countries: Applications of Social Representations Theory." *Frontiers in Psychology* 10: 1390.

———. 2019b. "Young Adults' Knowledge and Attitudes Regarding 'Music' and 'Loud Music' Across Countries: Applications of Social Representations Theory." *Frontiers in Psychology* 10: 1390.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots.* https://CRAN.R-project.org/package=patchwork.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Thompson, Charlie, Daniel Antal, Josiah Parry, Donal Phipps, and Tom Wolff. 2021. *Spotifyr: R Wrapper for the 'Spotify' Web API.* https://CRAN.R-project.org/package=spotifyr.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Maximilian Girlich. 2022. *Tidyr: Tidy Messy Data.* https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.