Creating the Most Danceable Song

BA222 E1

May 6, 2023

Nathan Basman U24197638

Ava Garber U43850811

Riley Shafer U32048913

Divya Singh U58274004

James Sullivan U34194946

2 Introduction

What makes a song danceable? Each year, the Billboard top dance song changes to a tune that bears little resemblance to its predecessor from the year before.[1] Being a producer in the music industry, it can be difficult to pinpoint what exactly makes a song "danceable." There are a few things to make known in our quest to answer the question, "what traits should we be focusing on for creating a "danceable" song?" The main topic of our analysis is to sort through the data and identify factors that make a song more danceable, so we can eventually use this information to predict which songs have the potential to become great songs to dance to. In order to do this, we will need to find a well-rounded dataset that matches the requirements of the assignment and will provide us with a collection of songs and variables that enable us to analyze danceability. In the dataset Top Hits Spotify from 2000-2019, we see that there are a wide variety of songs and song characteristics that can help us determine what makes a song danceable.[2]

To determine how to proceed with the project, we decided that it would be best to figure out which variables most affect danceability. We then will use forward regression methods (notably linear techniques), as well as correlation matrices to understand which specific variables are most effective in determining danceability. With this analysis, we are able to determine what factors make a song danceable.

3 Data Description

There are a few crucial pieces of information that help clarify the context of our dataset. Within our dataset, there are 2,000 observations and 18 variables, 12 of which are numerical and 6 are categorical. The breakdown of categorical and numerical variables is as such:
The categorical variables, as listed below, allow us to classify the observations into groups:

artist	song	explicit	year	genre	mode

The numerical variables, as listed below, gives us additional information necessary to determining how danceability is determined:

duration_ms	popularity	danceability	energy	key	loudness
speechiness	acousticness	instrumentalness	liveness	valence	tempo

[1] "Billboard Hot 100," PMC, accessed May 4, 2023, https://www.billboard.com/charts/hot-100/.
[2] Mark Koverha, "Top Hits Spotify from 2000-2019," Kaggle, accessed May 4, 2023, https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019.

Within the dataset, there are 59 duplicate observations. Due to the specific nature of this dataset, there were no observations that would be classified as outliers; however, for specific observations, there were some variables that did have outliers. This is shown in the table below, which highlights numerical variables with the number of outliers for that variable.

duration_ms = 16 outliers	loudness = 21 outliers	instrumentalness = 35 outliers
danceability = 8 outliers	speechiness = 35 outliers	liveness = 51 outliers
energy = 8 outliers	acousticness = 48 outliers	tempo = 7 outliers

Understanding the amount of outliers that each variable has will help us understand which variables may have skewed data as a result, which will enhance our understanding of how to approach the analysis. For variables like liveness, acousticness, instrumentalness, speechiness, and loudness, we can assume that since there is a significantly larger number of outliers, all having 1% or greater of its observations as outliers, that the data for these variables may be distributed more evenly, as there are more observations on either side of the distribution.

3.1 Distribution Analysis

We have taken a typical observation in the dataset and identified multiple typical characteristics. Out of the numerous variables within the dataset, we have chosen to focus on danceability, valence, energy, acousticness, tempo, and year. We chose to work with these variables as they have the greatest correlation and impact on danceability as calculated when using a forward regression, further explained in section 4.1. The table below contains the mean, median and mode of the five most significant numerical variables, in relation to danceability, for a typical observation. We have calculated the mean, median, and mode of these primary variables so as to further understand the variables' distribution and analyze if they are skewed. For clarification regarding variable definitions, please refer to the appendix.

Variable	MEAN	MEDIAN	MODE
Danceability	0.667	0.676	0.688
Valence	0.552	0.558	0.418
Energy	0.720	0.736	0.783
Acousticness	0.129	0.056	0.107, 0.150
Tempo	120.123	120.022	140.022
Year	2009.494	2010	2012

Comparing the variables' mean, median and mode is beneficial to our analysis because we are able to better understand the central tendency of the data. When the mean is less than the

median, there is a negative skew and negative extreme values. In this dataset, danceability is slightly negatively skewed. Energy and valence are more negatively skewed, while the rest of the variables we chose to focus on are positively skewed. The mode is another way for us to measure the central tendency. Since the mode is clearly not the same as the mean and median, it is safe to assume that the data is not symmetric.

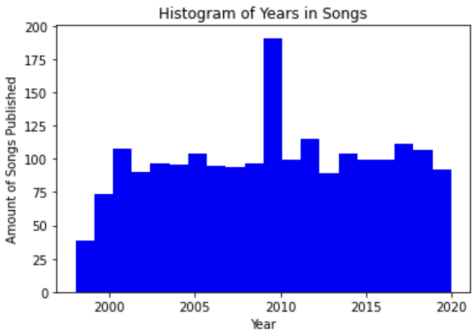
Understanding the central tendency of the dataset allows us to analyze its skew when comparing the central tendency to the data as a whole. For example, although the data is more negatively skewed in terms of valence and energy, we are able to take into account that there are more observations with lower energy and valence.

	Danceability	Valence	Energy	Acousticness	Tempo	Year
Count	2000	2000	2000	2000	2000	2000
Std	0.140	0.221	0.15	0.173	26.97	5.86
Min	0.129	0.0381	0.0549	.000019	60.019	1998
25%	0.581	0.387	0.622	0.014	98.99	2004
50%	0.676	0.558	0.736	0.0557	120.02	2010
75%	0.764	0.730	0.839	0.176	134.27	2015
Max	0.975	0.973	0.999	0.976	210.85	2020

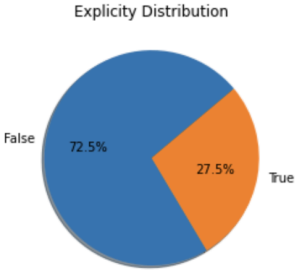
The summary statistics provides us with more valuable information such as standard deviation and confidence intervals. The larger standard deviation of the variable year tells us that there is a wide range of when these songs were produced. Consequently, we are able to observe danceability over a large span of years. However, the smaller standard deviation of the rest of our observed variables show us that the values are more clustered around the mean, relaying that the data points are more consistent. Within the dataset, there are a number of demographic variables:

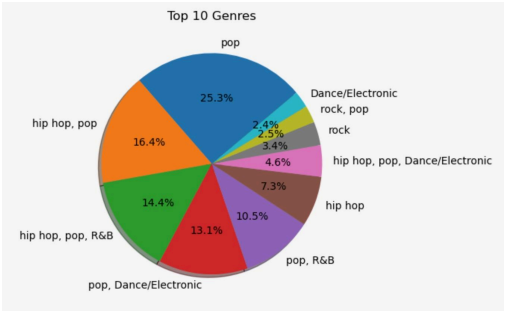
artist	song	year	explicit	genre

Understandably, we can assume that most songs have unique artists, so there would be little to no purpose in creating a data visualization to illustrate 2000 unique variables. Next, we can also assume the same argument regarding song, in which we know that there will be little to no chance of any sort of patterns. For the variable year, we have created a histogram that illustrates the release years for the songs in the dataset, based upon frequency distribution. We can clearly see an increase in the beginning of the histogram shown below, which may suggest the rise of



songs released going into the early 2000s as technology adapted and made personal music-listening more popular. Also, the outlier in 2009 suggests that it was a good year for producers and artists, as this year has nearly double the amount of songs released compared to other years in the dataset. Finally, when observing the variable explicit, we can derive that either songs will be labeled explicit or not, meaning that we should observe the distribution of explicit songs via a pie chart, as seen to the left. We see that with 27.5% of songs within the dataset being explicit, it may suggest that most songs without explicit lyrics tend to chart higher. Finally, due to the numerous types of genres, we decided to narrow down the extensive options and limit it to the top ten most popular genres that are observed in the dataset. As a result, we have created the following pie chart that illustrates the distribution of the top ten genres. We can see in the distribution, that pop and hip hop music tend to be either the primary or overlapping genre in 94.2% of the top 10, which encourages us to consider that when determining the danceability of songs, we should heavily consider that it is likely they will either be pop, hip hop, or an overlap of the two.

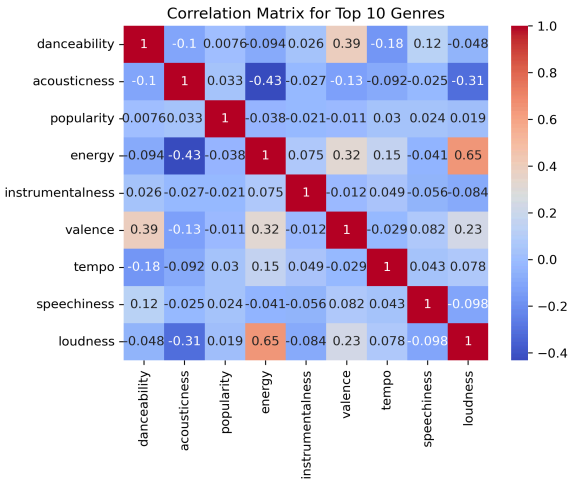




3.2 Variable Relations

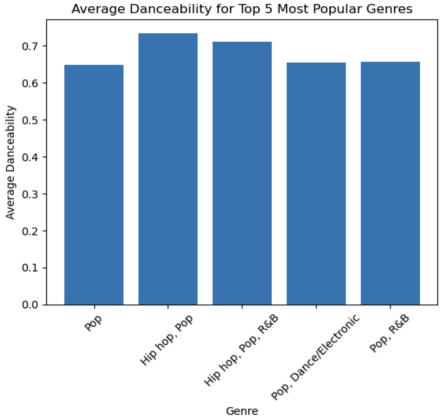
In order to further narrow down the data, we opted to concentrate on the top ten most popular genres (those that occur most frequently) in the dataset. Based on the correlation matrix and heat map for these 10 genres, shown below, it's clear there are several interesting relationships between the numerical variables. Energy and loudness have a strong positive correlation across all genres, indicating that songs with higher energy tend to be louder as well. There is also a moderate positive correlation between valence and danceability, suggesting that high-valence songs are often more danceable. For example, it is more likely that you would dance to Britney Spears' "Oops!.. I Did it Again" (valence: 0.894, danceability: 0.751) than Alexadra Burke's "Hallelujah" (valence: 0.094, danceability: 0.177). Additionally, there is a

negative correlation between acousticness and several other variables, such as energy and loudness. This indicates that songs with higher acousticness tend to have lower values for these variables. This data makes sense, as songs like "Lovely" (by Billie Eilish) and "Stay" (by Rihanna) would not be fun to dance to, due to their low energy and sad vibe. The correlation between tempo and energy varies across genres, with some genres, such as hip hop and metal, having a strong positive correlation, while others like classical and country have little to no correlation. There are also several non-linear relationships, such as the U-shaped relationship between valence and energy in the pop and rock genres. Finally, outliers are present in several variables such as tempo and loudness, indicating that some popular songs have extreme values for these variables. In short, the relationships between the variables are complex and vary across genres, making it difficult to make a blanket statement about variable relations across every single genre.



When considering what traits create the most "danceable" songs, we can focus specifically on the relationships between danceability and the other variables. The correlation matrix shows that danceability is positively correlated with several variables, including valence (0.39), speechiness (0.12), and instrumentalness (0.026). Although these correlations are relatively weak, they suggest that music with high valence, speechiness, and instrumentalness scores are ultimately more likely to be danceable. The summary statistics for these genres demonstrate that the average danceability score for the top ten genres is 0.67, making it nearly identical to that of the overall dataset's average, shown in section 3.1. We can also see that the average energy score is 0.71, the average loudness score is -5.53 dB, and the average valence score is 0.55 – all of which are also in line with the dataset's averages. However, when further narrowing the data to only include the top 5% of most popular songs, these scores drastically increase. The average danceability score leaps to an impressive 0.92. While the average energy, loudness, and valence scores all jump to 0.95, -2.32, and 0.94 respectively. These scores exhibit the importance of energy, loudness, and valence on the potential danceability of a song.

Furthermore, it's important to take into account the relationship between the genre of a song and its danceability. When looking at the top 5 most popular genres in the dataset: pop, hip-hop pop, hip-hop pop R&B, pop dance/electronic, and pop R&B, the danceability scores hover around or above the expected average. In fact, hip-hop pop and hip-hop pop R&B songs score the highest in danceability of any genre at 0.73 and 0.71, respectively. This means that creating a song in one of these genres will, on average, provide the best odds of being danceable. In short, when considering what traits will make a song most danceable, it's important to not only take into account the musical traits of the song, but also its genre.



Overall, based on the relationships between danceability and other variables, we can conclude that creating hip-hop pop or hip-hop pop R&B music with higher energy, loudness, and valence values is likely to lead to greater danceability. However, we should be cautious of outliers and non-linearities, and we should instead aim to create music that is more representative of the average danceable track.

4 Regression Analysis

In trying to estimate the best regression for predicting danceability, we tried to determine which characteristics of songs would be significant in making a song danceable, more specifically in maximizing danceability. We decided on this model being linear because when analyzing the relationship between different independent variables and danceability (the dependent variable), the non-linear fits tended to be very linear. Therefore, we decided it would be best to continue on with a forward regression, a linear approach. For a univariate regression model, we identified danceability as the dependent variable, and valence as the main independent variable. We then utilized a forward regression to better take into account omitted variable bias. We had originally chosen valence as the main independent variable for affecting danceability using correlation coefficients. Running a forward regression, we got the following equation:

"danceability ~ valence + energy + acousticness + tempo + year + speechiness + liveness + instrumentalness + loudness (+ error)"

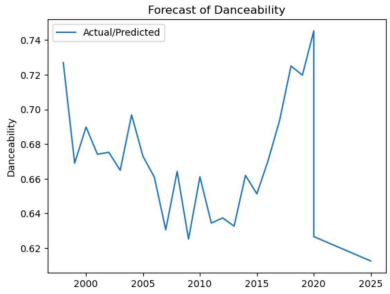
4.1 Policy Recommendation Model

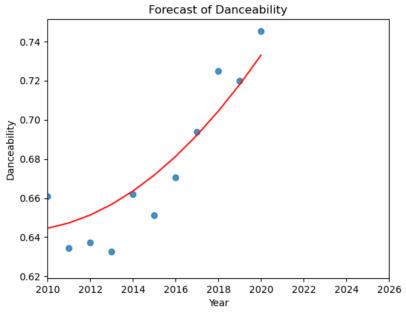
This model would be best for policy recommendation, further discussed in section 4.3. Obviously, this doesn't take into account other variables that could affect danceability that weren't found in the dataset, so there definitely still could be omitted variable bias. A few examples of variables that could affect danceability but aren't in the dataset could be popular locations the song is played, a person's alcohol level when listening, volume, and time of day.

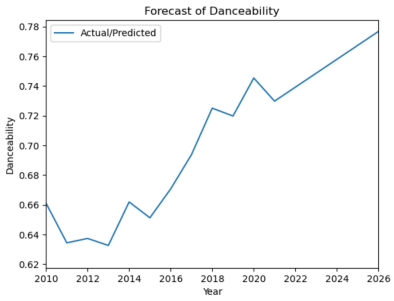
Variable	Associated impact on danceability for a 1 unit increase in the variable
Valence	0.31
Energy	-0.28
Acousticness	-0.13
Tempo	-0.001
Year	0.003
Speechiness	0.17
Liveness	-0.11
Instrumentalness	0.1
Loudness	0.003

Running a regression on the best model for predicting danceability, we came up with the equation: "danceability = -4.9 + 0.31*valence - 0.28*energy - .13*acousticness - 0.0008*tempo + 0.003*year + 0.17*speechiness - 0.11*liveness + 0.1*instrumentalness + 0.003*loudness + error." With this overall regression including nine independent variables, only one of these came out to be statistically insignificant, loudness, the last variable added into the regression. Starting off by analyzing the relationship between danceability and the variable with the highest correlation, valence, we get the equation [danceability = 0.53 + 0.26 * valence + error] and an r squared value of 0.16. As more variables were added into the regression, the most the valence coefficient varied was 0.06, but it's clear that a one unit increase in valence (musical positivity) is associated with an increase in danceability; in this first scenario, this increase is by 0.79. This pattern continued as more variables were added to the regression, leading up to the equation of [danceability = 0.8 + 0.31 * valence - 0.29 * energy - 0.12 * acousticness - 0.0007 * tempo + error] and an adjusted r squared of 0.26. The addition of the next variable, year, seemed to be most significant in analyzing these regressions, as the y-intercept of the equation substantially changed and became -5.1. Even though the regression's coefficients barely varied as more variables were added in, the y-intercept did change by over 4 units, and became negative. More generally, this means that when all independent variables in the regression have a value of 0 that danceability will be -5.1. Regardless of this drastic change, though, the adjusted r squared has increased, and all variables are statistically significant in the regression output. This pattern continued, with the y-intercept varying at most 0.18 units from -5.1. Continuing on to add speechiness, liveness, instrumentalness and loudness, we got our final equation, which had a y-intercept -4.92 and adjusted r squared of 0.298. The associations between the different, independent variables and danceability can be found in the table to the left.

The final addition of loudness was considered statistically insignificant, with a range of (-0.001, 0.007).







Overall, we were able to stick to our initial policy recommendation model determined by the forward regression, but we found new interpretations of this model. For example, we now know that valence *does* affect danceability, but that a greater value for valence will be needed in order to make danceability positive, given the now negative y-intercept. Adding controls, specifically year, significantly affected the relationships between danceability and the other variables in the regression, which seems to have contributed to making our policy recommendation model a lot more accurate since it also boosted the model's adjusted r squared. Using the results from this analysis, it's clear that a more danceable song would have greater valence, instrumentalness and loudness, and more negative energy, acousticness, and liveness.

4.2 Forecasting Model

As justified previously, we utilized a forward regression to determine the best model for predicting danceability. In order to predict annual danceability values until 2026, we compiled the average danceability values for every year in our dataset along with the averages of every other variable included in our optimal regression. Using methods learned in class involving the .predict( ) function, we found that the forecast looked like graph 1. As you can see, the predicted values following 2019 don't seem to follow the trend of the graph; this led us to reevaluate. We thought to ourselves, "is the danceability of a song that was popular 20 years ago really relevant to predicting the danceability of future popular songs?" In 2023, the chances of hearing MC Hammer's iconic track "Can't Touch This" in a club are extremely slim. However, during the 1990s, it stood out as one of the most infectious and danceable songs of the time. The landscape of popular music in the industry changes quickly and often. In the last 20 years, the industry has seen a significant shift in popularity from pop to hip hop. Therefore, we determined that in order to

Year	2019	2020	2021	2022	2023	2024	2025	2026
Predicted Danceability	0.72	0.75	0.73	0.74	0.75	0.76	0.77	0.78

better predict the future danceability of popular music, it would be more appropriate to predict based on data from only the last 10 years. After doing so, our results appeared much more realistic when compared to the trend of data as seen in the second graph above. The numerical results of danceability from 2019 to 2026 are clearly shown in the table to the right. As referenced earlier, we chose to use a linear regression because the non-linear regression line was basically linear in nature. This led us to believe that it was not necessary to use a quadratic regression, in addition to the fact that we wouldn't have been able to include other significant variables.

4.3 Policy Discussion

The implications of these results are mainly focused around the impact of the variables valence, energy and acousticness on the danceability of a song. This is important because these results could be used in many different ways. For example, in producing a new song using our results, a producer or artist would be able to tweak the song to take advantage of known relations and therefore, maximize danceability. Another use could be in DJing, which is focused on using known characteristics of danceable songs to create a "danceable" environment. A third use of these results could be using them to create a playlist, as most playlists are usually focused around a specific mood or emotional environment. The results could be used to create playlists with songs that are less danceable, have more negative valences and other musical characteristics correlated with less danceable songs, or the opposite, to create more upbeat, danceable playlists. Most importantly, the forecasting data we found could be used to inform artists or record labels on the average danceability values of future popular songs; therefore, giving them insight into how danceable they should make their songs if they want them to be popular in the future.

5 Conclusion

In the data description, we identified the primary variables in determining the regression to be valence, energy, acousticness, tempo, and year. By understanding the distribution of these variables, we were able to identify how skewed they were, which expanded our understanding of the dataset. With variables like valence and energy being negatively skewed, we were able to come to the conclusion that most of the dataset's songs were slower, more emotionally negative and more low energy. Furthermore, understanding the make-up of the songs within the dataset

enabled us to approach our regression for a more holistic view of what makes a song "danceable".

After conducting a forward regression on danceability with valence as the main independent variable, we came up with the following equation: "danceability = -4.9 + 0.31*valence - 0.28*energy - .13*acousticness - 0.0008*tempo + 0.003*year + 0.17*speechiness - 0.11*liveness + 0.1*instrumentalness + 0.003*loudness + error." We were also able to determine that 29.8% of the variation in danceability could be explained by these variables. After fully analyzing our best model for policy recommendation, we found that having a greater value for most of the independent variables would be needed to make a song more danceable. The addition of controls in our policy recommendation model had a significant impact on the relationships between danceability and the independent variables, which contributed to making our model more accurate since it also boosted the model's adjusted r squared. In terms of forecasting, we used a linear model to predict the probable average danceability of songs every year until 2026. We found that danceability will steadily increase year by year by about 0.01 with a slight decrease of 0.02 between 2020 and 2021.

Adding on to section 4.3, on how our data might be used, a producer may be able to utilize the now known relationship of valence, energy, acousticness, tempo or other variables in the regression to create a more danceable song. A DJ could similarly use it to create a more danceable environment. On the forecasting side, the data could give insight into how danceable future popular songs will be, which can be helpful for artists looking to make hits or record labels looking to invest in an artist or certain types of music. Looking forward, our policy recommendation and forecasting models can be applied and used in countless ways. We hope to provide valuable insights that can help drive success in the dynamic and evolving field of the music industry.

6 Appendix

Definitions of variables:

VARIABLE	DEFINITION
Artist	name of artist on song
Song	name of the track
Duration	duration of the track in milliseconds
Explicit	the lyrics/content of the song contains criteria that would be deemed unsuitable for children
Year	release year of the song
Popularity	the higher the value, the more popular the song is
Danceability	how suitable a track is for dancing based on a combination of musical elements (such as tempo, rhythm stability, beat strength, and overall regularity); ranging from 0 to 1 with a value of 0 indicating least danceable and a value of 1 indicating most danceable
Energy	ranging from 0 to 1; represents a perceptual measure of intensity and activity
Key	the key the song is in; integers map to pitches using standard Pitch Class notation (ex: 0 = C, 1 = C♯/D ♭ , 2 = D, and so on); if no key was detected the song was given a value of -1
Loudness	the averaged loudness of the entire song in decibels; the quality of a sound that is the primary psychological correlate of physical strength (amplitude); range from around -60 and 0 decibels
Mode	the modality (major or minor of a track); the type of scale from which melodic content is derived; major is represented by 1 and minor is represented by 0
Speechiness	detects the presence of spoken words in a song; the more exclusively speech-like the recording, the closer the value is to 1 (values above 0.66 describe tracks that are probably made up entirely of spoken words, values between 0.33 and 0.66 describe tracks that may contain both music

	and speech, values below 0.33 most likely represent music and non-speech-like tracks)
Acousticness	a confidence measures from 0 to 1 of whether the song is acoustic; 1 represents high confidence the track is acoustic
Instrumentalness	predicts whether a track contains vocals or not; the closer the value is to 1, the greater likelihood the track contains no vocals; values above 0.5 represent instrumental tracks but confidence in this is higher as the value approaches 1
Liveness	detects the presence of an audience in the recording; higher values represent an increased probability that the song was performed live (in this case, a value above 0.8 provides a strong likelihood the song was performed live)
Valence	Ranging from 0 to 1; describes the musical positiveness conveyed by a track; higher values mean the song sounds more positive (happy/cheerful); lower values mean the song sounds more negative (sad/angry)
Tempo	the overall estimated tempo of a song in beats per minute; the speed of a song which is derived directly from the average beat duration
Genre	genre of a song