What Makes Songs Danceable: An Analysis of Popular Tracks from 2000-2019

Haram Yoon and Shane Foster Smith

Introduction

Dataset: Paradisejoy. (n.d.). Top Hits Spotify from 2000-2019. Kaggle. Retrieved from https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019 The Spotify songs dataset contains various audio features and metadata on over 2000 popular tracks on the platform from 2000 to 2019. The primary goal of our study is to investigate the factors that influence the "danceability" of a song. Danceability is the measure of how suitable a track is for dancing based on a combination of musical elements.

Our motivation behind analyzing danceability specifically from the fact that highly danceable songs tend to be more widely appreciated, particularly in certain genres like pop, electronic dance music (EDM), and hip-hop. We first thought of predicting popularity of a songs, but that soon turned out to be less feasible and valuable compared to danceability. Furthermore, danceable tracks tend to be more enjoyable for listeners due to their higher energy level, a strong rhythm, and a tempo that aligns with typical dancing speeds. These danceable tracks are espeically popular in nightclubs, parties, and other social settings.

From a commercial perspective, creating danceable hits can be highly beneficial for artists and music labels. Tracks with high danceability tend to perform well on music streaming platforms which make them more likely to be added to playlists. Furthermore, danceable songs often stay popular for longer, as they remain relevant for dance-related events and activities over an extended period.

By analyzing the factors that contribute to danceability, our study aims to provide insights that can guide music producers and artists in crafting more danceable tracks tailored to the preferences of their target audience. The findings could also be valuable for music streaming platforms, radio stations, and other music curators in identifying and promoting highly danceable tracks, thereby tailoring a song to a particular social setting.

In this dataset, the response variable of interest is the "danceability" score, which ranges from 0 to 1, with higher values indicating more danceable tracks. The predictor variables include both numerical features related to the audio characteristics of the songs (e.g., energy, valence, tempo, acousticness, speechiness) and categorical variables like genre, mode and explicit content.

The original data was collected by Spotify, a popular music streaming service, and likely consists of tracks that were popular or frequently streamed on the platform during the given time period. However, the specific details on how the data was collected and processed are not provided.

The dataset includes 2000 observations (songs) and 18 variables, including the artist name, song title, duration, year of release, popularity score, and various audio features computed by Spotify's algorithms. These audio features, such as danceability, energy, valence, and acousticness, are numerical values ranging from 0 to 1 and are designed to capture different aspects of the audio content.

It's worth noting that some variables in the dataset may require further processing or transformation before being included in the statistical models. For instance, the "genre" variable contains multiple genres for some songs, and the "explicit" variable is binary, indicating whether the song contains explicit content. Variables

like “acousticness” and “speechiness” may also need to be explored further to understand their potential impact on danceability.

Moreover, interactions between variables could be important in predicting danceability. For example, the effect of tempo on danceability might vary across different genres, or the influence of energy on danceability could be moderated by the explicit content of a song.

By carefully analyzing and preprocessing these variables, as well as considering potential interactions and transformations, this analysis aims to develop effective statistical models that can accurately predict the danceability of a song based on its audio features and metadata.

By analyzing this dataset, we aim to identify the key factors that contribute to a song’s danceability and develop statistical models that can potentially predict this measure based on the available audio features and metadata. The findings of this analysis could provide valuable insights for music producers, artists, and streaming platforms to create and promote more danceable tracks, catering to the preferences of their audience.

Feature Engineering

To preprocess the data, the first step is to deal with any missing or NA data. Thankfully, there was none. Our next step was to change some of the categorical variables into usable datatypes in our models (binary representations or integers rather than booleans or characters). For example, the ‘Genre’ feature of a song was a string that could contain one or more genres out of the 14 possible genres(such as, “pop, hip hop”). To make this feature usable in our model, we performed two operations; first, we coalesced the nine least frequency genres into a single “Other” genre, which important to ensure dimensionality reduction and because these genres were very infrequent compared to the top five most frequent genres. After, we performed a one-hot encoding of the ‘Genre’ feature (into 6 binary features), which indicates if a song is part of one or more genres.
Second, most of the original features have values in the range [0,1], but a few features had values much larger or in negative ranges. We normalized some of these features’ values, while also keeping into consideration the interpretability of our models’ units. For example, two of the variables we normalized where ‘duration_ms’ and ‘loudness’. ‘Duration_ms’ original feature values where in the approximate range of [100000,400000] milliseconds and ‘Loudness’ had feature values in the approximate range of [-20,0] decibels. To normalize these values, we first determined that the original units weren’t particularly informative (what really is a one dB increase of loudness? Or, why is a millisecond differnce is length important?). Instead, we decided to use the scale() function so that new units are standard deviations from the mean. How loud a song is relative to other songs in our dataset is more informative. Similarly, how long songs are relative to other songs is more informative.

1. Preprocessing and Handling Important Categorical Variables

```
## [1] 1978
```

Figure 1: Danceability Distribution by Genre

By observation of the **Figure 1**, it appears that the frequency of danceability does slightly vary depending on the genre. For example, 'rock' and 'Other' peaks (or means) are slightly lower in the histograms than they are for 'Dance/Electronic' and 'Hip Hop'. Also, for 'Pop' and 'R&B', we can see that their songs are more frequenty in the [0,0.5] danceability range than songs in ''Dance/Electronic' and 'Hip Hop'. These results indicate that dance/electronic and hop hip songs may be considered more danceable, and perhaps deserve special consideration in our models.

Genre Counts

Figure 2 shows 'pop' as the most represented genre in the dataset, followed closely by 'hip hop'. Given the prominence of these highly danceable genres, the dataset is well-suited for analyzing the factors contributing to danceability across different musical styles. Furthermore, most of the songs in our data set are considered to be 'pop' (which makes sense, as this is data set of popular songs from 2000-2019) and lots of songs are considered to be pop *and* some other genre.

#Popularity

Popularity, as seen above, is binomially split between songs that have 0 or close to 0 popularity and those who have moderate to high popularity. Due to this fact, and because we found that the values of popularity weren't very correlated with our response, we decided perform a one hot encoding of popularity (aka. popularity is now a categorical variable, represented by four binary features: low popularity [0,25), moderate-low popularity [25, 50), moderate_high [50,75) and high popularity [75,100]).

2.Correlation of Potential Predictors

This plot shows the correlation between each numeric variable in our dataset (as a matrix heatmap). The darker colors show negative correlation while the warmer colors (more yellow) show more positive correlation. We can see that many variable in the heatmap are notably correlated (either positively or negatively) .

For example, the variable 'Energy' is positively correlated with loudness and very negatively correlated what acoustiness. Futhermore, loudess and acoustiness, unsurprisingly, are also very negative correlated. So, to avoid multicollinearity, we likely don't want to put more than one of these variables in our model without careful consideration.

Another example is, 'Speechiness', 'Explicit_numeric', and 'Hip Hop' are all positively correlated. This makes intuitive sense as we expect hip hop to be more 'wordy' and likely contain more explicit language. Again, we have to be careful with including more than one of these variables in a model, considering their correlation. Surprisingly, the popularity variables do not appear to be very correlated with any of the other variables.

To get a closer understanding of the correlations and relationships of our variables, we constructed a few pairs plots and considered some transformations of our variables.

Pairs Plots

Figure 4a: Pairwise Plot of Song Features

Figure 4b: Pairwise Plot of Song Characteristics 2 (with Transformations)

Figure 4c: Pairwise Plot of Song Characteristics 3

From the pairs plots, it's evident that danceability exhibits varying degrees of correlation with several audio features. Notably, it shows moderate positive correlations with valence (0.4) and log_speechiness (0.208), suggesting that higher danceability tends to be associated with higher valence and speechiness levels, respectively. Conversely, danceability demonstrates weak negative correlations with variables such as acousticness (-0.078), energy (-0.1), and tempo (-0.184), implying that higher danceability may be associated with lower acousticness, energy, and tempo values. However, the correlations with loudness_scale, duration_scale, sqrt_instrumental, and popularity appear to be negligible, indicating minimal linear relationships between these variables and danceability. Even though certain variables exhibit negligible correlations with danceability, it's important to still consider them for model building due to their potential non-linear relationships, contextual importance, and role in mitigating multicollinearity to help create a more comprehensive understanding of the factors influencing danceability. Furthermore, we see evidence of non-constant variance, non-linearity and possible violations of the assumption of normal errors for certain predictors (in particular, look at first pair-wise plot of characteristics).

3. Transformations:

In our analysis, we applied logarithmic and square root transformations to certain variables to address the conditions of a normal model. These transformations helped to stabilize the variance of residuals and achieve a more symmetric distribution, thereby enhancing the adherence of our models to the assumptions of linear regression. Through these transformations, we aimed to improve the reliability and interpretability of our regression analyses.

No Transformations

In cases where variables meet the model conditions based on diagnostic plots, such as with valance, seen above, transformations are not necessary as the assumptions of linearity, normality, and constant variance are adequately satisfied. Therefore, applying transformations to these variables may introduce unnecessary complexity (while also losing interpretability) without yielding significant improvement in model performance.

However, most of our continuous variable did require transformations in order to meet the normal errors model assumptions.

Log Transformations

Danceability vs Acousticness
Residuals vs Fitted
Residuals
Fitted values
901
698
1334
Danceability vs Acousticness
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Danceability vs Acousticness
Scale−Location
√|Standardized residuals|
Fitted values
Danceability vs Acousticness
Residuals vs Leverage
Standardized residuals
Cook's distance
626
877
901
Leverage

Danceability vs Log_Acousticness
Residuals vs Fitted
Residuals
Fitted values
Danceability vs Log_Acousticness
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Danceability vs Log_Acousticness
Scale−Location
√|Standardized residuals|
Fitted values
Danceability vs Log_Acousticness
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Danceability vs Speechiness
Residuals vs Fitted
Residuals
Fitted values
Danceability vs Speechiness
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Danceability vs Speechiness
Scale−Location
√|Standardized residuals|
Fitted values
Danceability vs Speechiness
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Danceability vs Log_Speechiness
Residuals vs Fitted
Residuals
Fitted values
901
698
1334
Danceability vs Log_Speechiness
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Danceability vs Log_Speechiness
Scale−Location
√|Standardized residuals|
Fitted values
Danceability vs Log_Speechiness
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

We encountered skewed distributions (both left and right), non-linearity, and heteroscedasticity in many of the predictor variables, as evidenced by the diagnostic plots above. To address these issues, we opted to apply a logarithmic transformation to these variables. After completing the transformations, we observed a marked improvement in the diagnostic plots; the previously skewed residual vs. fitted plots now display distributions that have a more uniform variance and linear relationship with the response variable. However, some of the residual plots still looked slighly quadtradic, which we will address later in this paper. In summary, the logarithmic transformation of skewed predictor variables yielded significant improvements in the linearity and homoscedasticity of our regression models, thereby enhancing the inference and predictive capabilities of these variables in a linear model.

Quadratic Transformations

Studentized Residuals vs Fitted Values for energy
Studentized Residuals
2
0
−2
−4
0.64
0.66
0.68
0.70
0.72
Fitted Values

We improved the linearity and constant variance of the 'duration_scale' predictor ... Also, the previous residuals vs fitted values for the 'energy' predictor looked slightly quadtradic, and we confirmed this suspicion by looking at the studentized residuals vs fitted values plots. As seen above, when we don't square 'energy', the studentized residuals heavily favor the fitted values far below the mean (rather than above the mean), especially for large fitted values; this indicates non-linearity. When we look at the same plot, but for a linear model that squares 'energy', we see that there is still a bias towards fitted values below the mean, but it is reduced and there is an improvement in constant variance.

Even after many of the previous transformations on the predictors, we still noticed a bias toward studentized residuals below the mean (sometimes more than 2 standard deviations). Therefore, we decided to transform the response variable, 'danceability', into a quadtradic term. Below, you will see various studentized residuals vs fitted values plots that show further improvements to various predictors when we transform 'danceability' to a squared term (in terms of constant variance, linearity, and, in particular, normal errors). Although transforming the response variable in this way makes the interprebility of our model more nuanced, the studentized residual plots show that this tradeoff is acceptable due to the significant improvement in the assumptions of a normal errors model. With these transformation, we feel more confident in the validity of the inference and predictions we perform later in this paper.

Studentized Residuals vs Fitted Values for energy_square (Danceability Squared I
Studentized Residuals
Fitted Values
2
1
0
−1
−2
−3
0.40
0.45
0.50

Studentized Residuals vs Fitted Values
for valence
Studentized Residuals
Fitted Values
2
0
−2
0.6
0.7

Studentized Residuals vs Fitted Values
for valence (Danceability^2)
Studentized Residuals
Fitted Values
2
0
−2
0.3
0.4
0.5
0.6

Studentized Residuals vs Fitted Values
for log_acousticness
Studentized Residuals
2
0
−2
−4
0.625
0.650
0.675
Fitted Values

Studentized Residuals vs Fitted Values
for log_acousticness (Danceability^2)
Studentized Residuals
2
1
0
−1
−2
0.400
0.425
0.450
0.475
0.500
Fitted Values

Studentized Residuals vs Fitted Values
for log_liveness
Studentized Residuals
2
0
−2
−4
0.625
0.650
0.675
0.700
0.725
Fitted Values

Studentized Residuals vs Fitted Values
for log_liveness (Danceability^2)
Studentized Residuals
3
2
1
0
−1
−2
0.40
0.45
0.50
Fitted Values

```
## [[1]]
## TableGrob (2 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[2]]
## TableGrob (2 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[3]]
## TableGrob (2 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[4]]
## TableGrob (2 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

Yeo-Johnson Transformations (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9118124/)

For the predictor 'duration_scale', we tried a variety of log, inverse and quadtric transformations for the purpose of stabilizing its variance (see figure . . .) . However, we found that none of these methods worked

as intended. Therefore, we decided to do some research into other types of transformation that may be able to help us. A Yeo-Johnson Transformation is an extension of the Box-Cox transformation; Yeo-Johnson can handle feature values, while Box-Cot transformations can't. MORE

```
## [[1]]
## TableGrob (1 x 1) "arrange": 1 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
##
## [[2]]
## TableGrob (1 x 1) "arrange": 1 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
```

Although the improvement isn't drastic, this transformation help us stabilize the importance of the outlying observations, especially the shorter duration observations.

4. Comparing Two Naive Models

With the previous analysis of the relationships in our model, and with various domain knowledge, we constructed two models we thought would explain dancebility well without introducing too much multicollinearity or overfitting. Consider the two models below shown below. For both models, we determined that the variables 'valence','energy_square' and 'log_liveness' were all correlated enough with 'danceability' to always be included in the models. We also determined that the categorical variable 'Dance/Eletronic' should also be included for both models (given the name and previous histograms).
For the first model, we decided to focus more on tempo, genre and instrumentalness. This is an intuitive approach to modeling danceability; people usually dance to songs with a strong beat and/or groove in the instrumentation, and these features are common in certain genres. Furthermore, we predicted tempo would

become very important for 'Dance/Eletronic' songs in terms of predicting danceability (rather than for genres such as rock or jazz). We added an interaction term to reflect these hypothesises.

The second model focuses more on the speechiness and language of a song. We hypothesized that very wordy songs would be less danceable (similar to the opposite of instrumentalness). Furthermore, we hypothesized that explicit language would decrease the effects of valance perceived danceability of a song, as much of the perceived valence of song comes from the language itself; we included two interaction terms to reflect this hypothesis. Also, we should note that, given we included speechiness (and explicit language) in the second model, we removed the categorical variable 'hip hop' and `rock` to avoid multicollinearity.

```
##
## Call:
## lm(formula = danceability_square ~ valence + energy_square +
##     log_liveness + loglog_instrumentalness + `Dance/Electronic` +
##     `hip hop` + rock + tempo * `Dance/Electronic`, data = song_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45068 -0.09230 -0.00609  0.09363  0.48012
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.3883589  0.0241946  16.051  < 2e-16 ***
## valence                   0.3693905  0.0171491  21.540  < 2e-16 ***
## energy_square            -0.2003079  0.0190428 -10.519  < 2e-16 ***
## log_liveness             -0.0260601  0.0052392  -4.974 7.21e-07 ***
## loglog_instrumentalness   0.0017510  0.0002775   6.310 3.54e-10 ***
## `Dance/Electronic`        0.1852151  0.0528267   3.506 0.000466 ***
## `hip hop`                 0.0850709  0.0076685  11.094  < 2e-16 ***
## rock                     -0.0840684  0.0117208  -7.173 1.09e-12 ***
## tempo                    -0.0004559  0.0001408  -3.237 0.001231 **
## `Dance/Electronic`:tempo -0.0013780  0.0004205  -3.277 0.001071 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1458 on 1717 degrees of freedom
## Multiple R-squared:  0.3597, Adjusted R-squared:  0.3564
## F-statistic: 107.2 on 9 and 1717 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = danceability_square ~ valence + energy_square +
##     log_liveness + log_speechiness + `Dance/Electronic` + explicit_numeric +
##     explicit_numeric * valence, data = song_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44007 -0.10033 -0.00396  0.09858  0.47211
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.322253   0.023986  13.435  < 2e-16 ***
## valence                   0.433642   0.020262  21.402  < 2e-16 ***
```

```
## energy_square            -0.236226   0.019035 -12.410  < 2e-16 ***
## log_liveness             -0.028529   0.005366  -5.317 1.20e-07 ***
## log_speechiness           0.020558   0.005207   3.948 8.21e-05 ***
## `Dance/Electronic`        0.038863   0.009394   4.137 3.69e-05 ***
## explicit_numeric          0.165464   0.022136   7.475 1.22e-13 ***
## valence:explicit_numeric -0.156582   0.037039  -4.227 2.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1499 on 1719 degrees of freedom
## Multiple R-squared:  0.322,  Adjusted R-squared:  0.3193
## F-statistic: 116.7 on 7 and 1719 DF,  p-value: < 2.2e-16
```

We wanted to check for multicollinearity and any leverage or influence points using these 2 models as an overall insight for any future models as well.

```
##                  valence            energy_square             log_liveness
##                 1.158323                 1.283867                 1.035856
##  loglog_instrumentalness       `Dance/Electronic`                `hip hop`
##                 1.082351                35.608894                 1.138681
##                     rock                    tempo `Dance/Electronic`:tempo
##                 1.157709                 1.161211                35.986163
```

Multicollinearity varies across variables in the model after calculating the Variance Inflation Factor (VIF). While Valence, Energy Square, Log Liveness, and Sqrt Instrumentalness show low to very low levels, Dance/Electronic and Hip Hop, along with their interaction terms with Tempo, exhibit potential high multicollinearity. This suggests redundancy or strong correlation, potentially affecting model stability and interpretation.

```
##                  valence            energy_square             log_liveness
##                 1.528962                 1.212938                 1.027464
##          log_speechiness       `Dance/Electronic`         explicit_numeric
##                 1.260033                 1.064728                 7.573614
## valence:explicit_numeric
##                 7.521943
```

In this case, Valence, Energy Square, Log Liveness, Dance/Electronic and Log Speechiness show low to very low levels, Energy Square and Valence along with their interaction terms with Explicit Numeric show medium levels, and with Explicit_Numeric to cause a little concern for multicollinearity.

Afterward, we plotted an influence plot for each model to pinpoint Cook's distance and leverage points to uncover leverage or influential points that could influence the regression model's coefficients. This method helps us understand the model's reliability by highlighting potential outliers and assessing their impact.

```
## [1] 0.04979734
```

```
##         StudRes         Hat       CookD
## 112  -0.8950689 0.044339962 0.003717535
## 232  -2.8990650 0.019379877 0.016538543
## 444   0.6381355 0.045509020 0.001942234
## 543   3.3084116 0.003511720 0.003835115
## 1082 -1.6926118 0.040818681 0.012178717
## 1469 -3.1106520 0.007506349 0.007281400
```

The influence plot revealed a small handful of observations with leverage values greater than twice the average leverage $(2 \times (k+1)/n) = 0.047$ (chirp chirp), where k is the number of predictors and n is the number of observations. These observations can be considered high leverage points. There are unfortunate a small number that passed this threshold within our first model, indicating leverage points. However, when examining the Cook's distance plot, none of the observations had a Cook's distance value greater than 1, which is a common threshold for identifying influential observations.

```
##          StudRes         Hat        CookD
## 242   2.97966355 0.001317964 0.0014579245
## 732   2.55130162 0.010373363 0.0085014345
## 1139 -0.08455883 0.016677617 0.0000151676
## 1277  3.16078826 0.002420198 0.0030139670
## 1440  0.79795472 0.015586349 0.0012604484
## 1497  2.46814224 0.011987063 0.0092111967
```

The influence plot did not reveal any observations with high leverage values exceeding twice the average leverage. Similarly, the Cook's distance plot did not show any observations with a Cook's distance value greater than 1.

In summary, while model_dance1 had a few high leverage points identified through the influence plot, neither model exhibited observations with substantial influence based on the Cook's distance criterion. This suggests that, although there are some high leverage points in model_dance1, they may not be significantly influencing the regression coefficients or model fit. However, it's still recommended to investigate these high leverage points further and consider their potential impact on the model. These are most likely not the exact models we will be using, so it is a bright sign for building a more concise model later.

Then, we performed a 10-fold cross validation of our data to evaluate how well each model performs on unseen data.

```
## Linear Regression
##
## 1727 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1555, 1555, 1554, 1554, 1554, 1554, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1480577  0.3355642  0.1170529
```

```
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

## Linear Regression
##
## 1727 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1554, 1554, 1554, 1556, 1555, 1554, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1481067  0.3375759  0.1176105
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

As seen above, the models performed comparably in terms of RMSE (Root Mean Squared Error), R-Squared, and MAE in a 10 fold cross validation test with the training data. The second model appears to perform marginally better in each of the outputted metrics, thus indicating that the second model performs slightly better for predicting unseen data. This may indicate that language and genre feature are marginally more important for modeling danceability than instrumentation and tempo features.

5. Model Building

BIC Selection (Forward Stepwise vs Backwards)

We then considered a two different models built with Backwards BIC Selection and Forwards Stepwise BIC Selection. Since we are considering so many different variables and interaction terms in our model, we determined BIC was appropriate criterion for the purpose of penalizing overly complex models. Furthermore, many the predictors in our data set are correlated, therefore less complex models also helps us avoid issues with multicollinearity.
Consider the output models from each of these selection processes.

```
## [1] 1727

##                    artist                      song               duration_ms
##                         0                         0                         0
##                  explicit                      year                popularity
##                         0                         0                         0
##              danceability                    energy                       key
##                         0                         0                         0
##                  loudness                      mode               speechiness
##                         0                         0                         0
##              acousticness          instrumentalness                  liveness
##                         0                         0                         0
##                   valence                     tempo            duration_scale
##                         0                         0                         0
##            loudness_scale          explicit_numeric               key_numeric
##                         0                         0                         0
```

```
##                       pop                      rock                     Other
##                         0                         0                         0
##                   hip hop                       R&B          Dance/Electronic
##                         0                         0                         0
##           popularity_high        popularity_modhigh            popularity_low
##                         0                         0                         0
##         popularity_modlow           log_speechiness            valence_square
##                         0                         0                         0
##              log_liveness          log_acousticness     sqrt_instrumentalness
##                         0                         0                         0
##       danceability_square             energy_square              yeo_duration
##                         0                         0                         0
##             root_duration adjusted_instrumentalness   loglog_instrumentalness
##                         0                         0                         0
```

```
##
## Call:
## lm(formula = danceability_square ~ valence + energy_square +
##     log_liveness + loglog_instrumentalness + yeo_duration + `Dance/Electronic` +
##     `hip hop` + rock + tempo + explicit_numeric + `Dance/Electronic`:tempo +
##     energy_square:explicit_numeric, data = song_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45136 -0.08685 -0.00555  0.09136  0.45283
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.3574843  0.0243826  14.661  < 2e-16 ***
## valence                        0.3666656  0.0168303  21.786  < 2e-16 ***
## energy_square                 -0.1341578  0.0212521  -6.313 3.48e-10 ***
## log_liveness                  -0.0274095  0.0051140  -5.360 9.47e-08 ***
## loglog_instrumentalness        0.0017664  0.0002722   6.489 1.13e-10 ***
## yeo_duration                  -0.0114805  0.0036834  -3.117  0.00186 **
## `Dance/Electronic`             0.1702532  0.0516817   3.294  0.00101 **
## `hip hop`                      0.0573929  0.0083938   6.838 1.12e-11 ***
## rock                          -0.0830693  0.0114485  -7.256 6.02e-13 ***
## tempo                         -0.0005944  0.0001382  -4.299 1.81e-05 ***
## explicit_numeric               0.1662996  0.0215711   7.709 2.13e-14 ***
## `Dance/Electronic`:tempo      -0.0012441  0.0004109  -3.027  0.00250 **
## energy_square:explicit_numeric -0.1995725  0.0388365  -5.139 3.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1422 on 1714 degrees of freedom
## Multiple R-squared:  0.3917, Adjusted R-squared:  0.3875
## F-statistic: 91.99 on 12 and 1714 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = danceability_square ~ valence + energy_square +
##     `hip hop` + rock + loglog_instrumentalness + explicit_numeric +
##     log_liveness + tempo + yeo_duration + energy_square:explicit_numeric,
##     data = song_train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45309 -0.08856 -0.00449  0.09298  0.44927
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.3769131  0.0238032  15.835  < 2e-16 ***
## valence                         0.3634980  0.0168029  21.633  < 2e-16 ***
## energy_square                  -0.1293479  0.0210932  -6.132 1.07e-09 ***
## `hip hop`                       0.0560602  0.0084015   6.673 3.38e-11 ***
## rock                           -0.0868457  0.0112844  -7.696 2.35e-14 ***
## loglog_instrumentalness         0.0018407  0.0002686   6.852 1.01e-11 ***
## explicit_numeric                0.1652338  0.0216243   7.641 3.56e-14 ***
## log_liveness                   -0.0276546  0.0051269  -5.394 7.85e-08 ***
## tempo                          -0.0007200  0.0001309  -5.501 4.35e-08 ***
## yeo_duration                   -0.0130113  0.0036454  -3.569 0.000368 ***
## energy_square:explicit_numeric -0.1988571  0.0389462  -5.106 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1427 on 1716 degrees of freedom
## Multiple R-squared:  0.3874, Adjusted R-squared:  0.3838
## F-statistic: 108.5 on 10 and 1716 DF,  p-value: < 2.2e-16
```

```
##         Metric Backward_Selection Both_Direction_Selection
## 1          AIC      -1820.4404142            -1812.1468925
## 2          BIC      -1744.0824391            -1746.6971995
## 3 Adjusted R^2          0.3874806                0.3838239
## 4           Cp         13.0000000               11.0000000
```

Now, let's consider the intuition behind some of the values of our coefficients. Continuous variables such as 'valance', 'instrumentalness', and indicator variables such as 'Dance/Electronic', 'hip hop' and explict langage improve the danceability of songs according to this model. We believe this result to be consistent with our domain knowledge; we initially expected that more positive sounding, instrumental songs would more danceable and we also expected that Dance/Electronic and hip hop music would also be perceived as more danceable. Furthermore, songs with explicit language are more common in social setting where people dance (such as bars and clubs), so more danceable music may be produced with explict language more often.
Increases in energy, liveness, tempo and all of our interaction terms, on the other hand, decrease the expected danceability (squared) of a song. Although liveness decreasing danceability isn't particularly surprising (songs we dance to, especially in hip hop and eletronic music, usually aren't recorded live), energy decreasing danceability was a little suprising to us. We suspect that songs with too much energy are too intense to dance to properly. Our tempo coefficient, and interaction term with tempo, indicate that lower tempos are more conducive to dancing, and that, when a song is Dance/Electronic, higher tempos hurt the dancebility of the song more than for other genres. The two interaction terms involving *explicit_numeric* indicate that the positive effect of valence on danceability is diminished in the presence of explicit language. Additionally, explicit language amplifies the negative impact of high energy (squared) on danceability.

As seen above, the second model (found with Forward Stepwise BIC) has better criteria metric for BIC and *C_p*. Based on these metrics, we determined that the Forward Stepwise BIC model would likely better predict danceability for unseen data and would have better explainability and interpretability.
Let's consider how we can interpret the coefficient estimates and their signficance. Each of the predictors' p-values are below our predetermined signficance level of $\alpha = 0.01$; therefore, given all other predictors and

interaction terms are held constant, each predictor in our model significantly contributes to the predictive power of the model. Aka. changes in these predictor values results in signficant changes to danceability of a song. The ANOVA table shown above confirms the signficance of the individual predictors in our final modal.

Before we evaluate this model on test data we set aside at the start, we will finally consider why certain variables were dropped out of the model during the BIC selection process. For example, if we look back to the correlation heatmap, we can see that 'hip hop' is very positively correlated with speechiness, and that energy is noteably correlated with both acousticness and loudness. You'll notice that none of 'speechiness', 'acousticness' or 'loudness' are included in the final model; these variables were likely dropped as they weren't signficant in our model (in terms of BIC) in the presense their highly correlated predictors 'hip hop' and 'energy'. One interesting observation is that 'energy' and 'valence' are positively correlated, however they are both included in the model. We suspect that both were signficant in the model because, although they are positively correlated, they are *oppositely* correlated to 'danceability.' (Interactions, more comparison)

Just to confirm that our stepwise function deals with multicollinearity, we performed VIF test on the coefficients and found that none of values large enough to signify a large multicollinear relationship.

```
##                       valence                  energy_square
##                      1.161613                       1.645454
##                     `hip hop`                           rock
##                      1.427718                       1.120944
##       loglog_instrumentalness               explicit_numeric
##                      1.059377                       7.984521
##                  log_liveness                          tempo
##                      1.036120                       1.047986
##                  yeo_duration energy_square:explicit_numeric
##                      1.036730                       7.345291
```

As we can see that none of these VIF scores are of concerning values, indicating our stepwise model did a very good job in selecting predictors to prevent multicollinearity.

```
## [1] "High influence points in the y direction:0"
```

```
## [1] "High influence points in the x direction: 5"
```

Checking for studenized residuals as well as leverage points the presence of 0 high influence points in the y (dependent) variable but around 5 high influence points in the x (independent) variables suggests potential influential observations or non-linear relationships. By using the Bonferroni correction for outlier detection in the x direction, we are controlling the overall type I error rate, ensuring that the probability of any false positives is limited. Additionally, setting a leverage cutoff based on the number of predictors and sample size provides a standardized threshold for identifying potentially influential points in the y direction. We chose the cutoff to be 3 times the ratio of predictors to sample size due to how large our dataset is and how much variation is possible within the context of songs. Since there are a couple of potential high influence points, let's dig deeper to look at specific observations.

```
##         StudRes         Hat        CookD
## 182   0.2825167 0.023818363 0.0001771373
## 788  -3.1929234 0.005153847 0.0047757223
## 824  -2.9258094 0.010223020 0.0080026167
## 1277  3.1646945 0.004454084 0.0040522140
## 1363 -1.1894371 0.026713203 0.0035291528
## 1549  2.7409640 0.022956062 0.0159864517
```

As we can through the influence plots, there are 3 specific values that seem to be way greater than 3 times the average hat value. Observations with high leverage, especially those significantly higher than the average leverage, have the potential to exert substantial influence on the estimated regression coefficients. These points are observations: 1549, 182, and 1363. If we focus on studentized residuals, we see that there are multiple studentized residuals greater than 2 or less than -2 which may cause for concern. However, since we have a large sample size, it makes sense for there to be individual extreme residual values. Since our model passes the assumptions of regression analysis, these studentized residuals might not indicate violations of these assumptions. Futhermore, when considering the effect of prediction, individual behavior of residuals may not be as concerning. Analyzing potential causual relationships between variable may need to be proceeded with extra caution due to extreme residuals. Another metric we looked at was Cook's Distance which showed that there is no Cook's distance greater than the threshold of 1. We are choosing 1 as our threshold due to our large sample size and passing of model assumptions. However, it is important to note that there is one observation, 1549, what is extremely higher than every other Cook's distance observation. Observation 1549 also has a big hate value and a large studenized residual, making it very likely this point is a concerning outlier.

```
## [1] "Threshold for DFFITs: 0.152189197461489"
```

```
##        1549         182        1362
##  0.42014092  0.04413009 -0.04860979
```

```
## [1] "Threshold for DFBETAs: 0.0481264499251422"
```

```
## [1]  0.079980141  0.005179937 -0.013576172
```

Looking at DFFITs and DFBETAs values for a couple of more concerning observations within our data, we can see that observation 1549 greatly exceeds both threshold while observations 182 and 1362 do not. We can conclude that 1549 is an extreme case and should be considered an outliers to remove in our model.

```
## # A tibble: 1 x 42
##   artist   song  duration_ms explicit  year popularity danceability energy   key
##   <chr>    <chr>       <dbl> <lgl>    <dbl>      <dbl>        <dbl>  <dbl> <dbl>
## 1 Sidney ~ Rive~      320348 TRUE      2009         47        0.804  0.976     1
## # i 33 more variables: loudness <dbl>, mode <dbl>, speechiness <dbl>,
## #   acousticness <dbl>, instrumentalness <dbl>, liveness <dbl>, valence <dbl>,
## #   tempo <dbl>, duration_scale <dbl>, loudness_scale <dbl>,
## #   explicit_numeric <dbl>, key_numeric <dbl>, pop <dbl>, rock <dbl>,
## #   Other <dbl>, `hip hop` <dbl>, `R&B` <dbl>, `Dance/Electronic` <dbl>,
## #   popularity_high <dbl>, popularity_modhigh <dbl>, popularity_low <dbl>,
## #   popularity_modlow <dbl>, log_speechiness <dbl>, valence_square <dbl>, ...
```

Our outlier observatoin is "Riverside" by Sidney Samson with a danceability of 0.804. Our previous models showed that danceability and energy have generally a negative correlation with one another, yet this specific outlier has a positive correlation between the 2 variables. It makes sense that this would be an outlier, and should be removed.

Bye bye, Riverside !

The model seems to perform reasonably well for the provided mean responses, with relatively narrow 95% confidence intervals indicating precise predictions. However, there is a wide 95% confidence intervals for future responses which may raise concerns about the model's ability to make reliable and accurate predictions for new, unseen data considering 0.016641, 0.950625 is the range of our predicotr variable. Such high uncertainty and low precision in future predictions could stem from various factors, including inadequate model complexity, missing important predictor variables, violations of model assumptions, overfitting or underfitting issues, or the presence of influential observations or outliers. To improve the model's overall performance and increase the precision of its predictions, it is crucial to revisit the modeling process, investigate potential issues, explore alternative modeling approaches or techniques, and validate the model's performance rigorously on a separate test or validation dataset.

6. Sparse and Smooth Linear Models

Ridge Regression

For ridge regression and LASSO, we included many continuous, categorical and interaction variables that we found were fairly significant during the course of building our MLR model.

Before we do so, we will proceed with some high influential/leverage points analysis as well as an VIF analysis to check for multicollineairty. Analyzing high influential/leverage points and conducting a VIF analysis is crucial because it helps ensure the reliability and robustness of our predictive model. By identifying influential data points and assessing multicollinearity among predictor variables, we can address potential issues that might affect the accuracy and generalizability of the RR and Lasso model we are trying to build as well as their predictions.

After creating a premodel in with all the relevant predictors, we do the same process as the evaluations for the stepwise regressionmodel. Checking for studenized residuals as well as leverage points, we found the presence of 0 high influence points in the y (dependent) variable but 38 high influence points in the x (independent) variables suggests potential influential observations or non-linear relationships. This time, there are much more potential high influence points which makes sense due to how much more complex this

premodel is compared to the previous stepwise one. Let's dig deeper to look at specific the observations of concern.

```
## [1] "High influence points in the y direction: 0"
```

```
## [1] "High influence points in the x direction: 38"
```

I sorted by cook's distance to see most influential points.

```
##         StudRes         Hat       CookD
## 7    -2.3600948 0.051811098 0.014454564
## 141  -1.1510840 0.077602498 0.005307250
## 231  -0.7459112 0.079980041 0.002303840
## 232  -2.4760309 0.036533735 0.011036858
## 242   3.2379639 0.003314608 0.001651162
## 1277  3.1292607 0.006175771 0.002882783
```

Plotting the influence and cook's distance plots, we notice a lot more concerning data points in comparison to our stepwise model. There a bunch of more concerning leverage point(hat values) as we found above. There are a couple of outlying cook's distance observations as seen in the cook's distance plots, with the points 7 and 232 being the most extreme. Points such as 7, 141, 231, 232, 242, and 1277 were selected by the plot to have the biggest influences on our data.

```
##                        SongTitle   Leverage        Cook    Stu_Res
## 7                         Say It 0.05181110 0.014454564 -2.3569382
## 232        Suffocate - Superclean 0.03653373 0.011036858 -2.4723138
## 497            Walking On A Dream 0.05254667 0.006377166  1.5539241
## 1082  Papercuts (feat. Vera Blue) 0.06332945 0.005448752 -1.3009144
## 141  Champion (feat. Chris Brown) 0.07760250 0.005307250 -1.1509743
## 423              BUTTERFLY EFFECT 0.03703241 0.004641860  1.5921006
## 1152       Sit Still, Look Pretty 0.07364875 0.004043487  1.0334580
## 470                 Midnight City 0.05985743 0.003147809 -1.0189472
## 177                          Gold 0.06335635 0.003140087 -0.9873534
## 1284                  Starry Eyed 0.05580538 0.002762556 -0.9907360
```

Cook's distance measures the influence of each observation on the fitted values of the model. It is often the most important metric because it captures both the leverage and residual information, which is where we will putting most our attention to. According to our plot, songs like "Say it" "Suffocate - Superclean" are shown to have the highest cook's distance as long with leverage and studenized residuals on the more extreme ends. Within this dataset, those 2 may be the most important to look at.

```
##                            SongTitle dffits_values     dfbetas
## 119                       Days Go By     0.3824898  0.04186109
## 1497                    bury a friend     0.3676134  0.18376895
## 497                Walking On A Dream     0.3661035 -0.01171888
## 478                    Ladbroke Grove     0.3526917  0.01503875
## 1679                  Te Boté - Remix     0.3358014  0.12016101
## 732                Turn Down for What     0.3238153  0.04667169
## 423                  BUTTERFLY EFFECT     0.3123571  0.02060392
## 225  Girls Like (feat. Zara Larsson)     0.3121209 -0.04198411
## 1540                             Sail     0.2994668  0.06516456
## 1152          Sit Still, Look Pretty     0.2914046  0.01234788
```

There are many valuess here that had a non-extreme Cook's distance, stutenized residuals, and leverage point values. If a value is extreme as DFFIT but not Cook's distance, it suggests that the particular observation has a significant influence on the predicted values of the model, but it may not necessarily have a large impact on the model coefficients. The observations that fall in both high cook & leverage values and the extreme dffits values are shown below.

```
##                SongTitle   Leverage        Cook  Stu_Res dffits_values
## 1     Walking On A Dream 0.05254667 0.006377166 1.553924     0.3661035
## 2       BUTTERFLY EFFECT 0.03703241 0.004641860 1.592101     0.3123571
## 3 Sit Still, Look Pretty 0.07364875 0.004043487 1.033458     0.2914046
##       dfbetas
## 1 -0.01171888
## 2  0.02060392
## 3  0.01234788
```

Among the provided data points, "BUTTERFLY EFFECT" stands out as a potential outlier due to its high Cook's distance, standardized residual, and DFBETAS values, suggesting its substantial influence on the regression model. Although "Walking On A Dream" also exhibits notable leverage and DFFITS values, "BUTTERFLY EFFECT" appears to have a more pronounced impact based on multiple diagnostic measures, indicating its potential as an outlier.

```
## [1] "BIC for Premodel:  -1717.4507527966"
```

```
## [1] "BIC for Premodel without Outliers:  -1718.30415796928"
```

While the 2nd model has the better BIC score, the difference is only 1. This small difference shows that these 2 outliers aren't having enough of an impact on our data for us to want to remove them from our dataset.

```
## [1] "Threshold for DFFITs: 0.215290366859673"
```

```
##           7         141         231         232         242        1277
## -0.14424173 -0.15544793 -0.09587117 -0.16177955  0.14799903  0.21168026
```

```
## [1] "Threshold for DFBETAs: 0.0481403895199096"
```

```
## [1] -0.038785649  0.049752881  0.026083031 -0.069562426  0.009843298
## [6]  0.016199819
```

Back to the original points that our influence plot suggested were the most concerning. We went on to find the specific values for the dffits and average dfbetas of our specific observations, but found none of them to be conclusively say passes the threshold. The 141, Champion (feat. Chris Brown), observation has an average around the dfbetas threshold, and therefore, should be looked deeper into the specific variables.

```
##                     (Intercept)                         valence
##                     0.049752881                     0.020966611
##                   energy_square                       `hip hop`
##                     0.001309079                    -0.025958787
##                            rock        loglog_instrumentalness
##                     0.009047265                     0.016315680
##                explicit_numeric                   log_liveness
##                     0.078997325                     0.008598595
##                           tempo                   yeo_duration
##                    -0.075391011                    -0.004189793
## energy_square:explicit_numeric
##                    -0.102226602
```

It looks like explicit_numeric is getting affected by this specific observation, but overall, doens't seem that conclusive enough to consider it an outlier worth getting rid of.

```
##                         valence                       `hip hop`
##                        1.639778                      21.276295
##                    energy_square                            rock
##                        1.784402                       1.543479
##                 explicit_numeric             `Dance/Electronic`
##                       11.480900                      46.022246
```

```
##                          pop                          mode
##                     1.441328                      1.027990
##               log_speechiness        loglog_instrumentalness
##                     1.433632                      1.112409
##                        tempo                   log_liveness
##                     2.083342                      1.041501
##                  yeo_duration energy_square_explicit_numeric
##                     1.090744                      8.297452
##       valence_explicit_numeric            DanceElectronic_pop
##                     8.573394                      7.186351
##           DanceElectronic_rock          DanceElectronic_tempo
##                     1.283313                     37.341453
##                  hiphop_tempo         DanceElectronic_hiphop
##                    21.316026                      1.615412
```

After doing an VIF analysis, we unfortunately see that there a couple of terms such as hip hop and Dance-Electronic_tempo that show high multicollinearity. Thankfully, Lasso and RR have methods in reducing the impacts of multicollinearity through their error terms. However, one can predict that lasso will perform better with these multicollinear values due its ability to dimension reduce.

```
## [1] 0.007200978
```

(Commentary, best lambda, how coefficents go down, etc.)

LASSO Regression

```
## [1] 6.715656e-05
```

(Commentary, best lasso, etc.)

Comparing MLR, RR and LASSO

```
## # A tibble: 3 x 2
##   Model correlation
##   <chr>       <dbl>
## 1 Lasso       0.635
## 2 MLR         0.624
## 3 RR          0.631
```

(Need commentary. Basically, LASSO is best of the three, observed values have better correlation predicted values)

9. Smoothing with Tempo

As we can see from **Figure** and **Figure** above, tempo has a somewhat strange relationship with dancebility squared; it has a distrubution with 2 peaks and its relationship to danceability looks slightly quadtratic (although quadtraic transformations on tempo did not improve its assumptions on the normal errors model). Also, our domain knowledge tells us that tempo is a bit of a unique variable. In 4/4 time signature (which the vast majority of popular songs are in), a 100bpm song can sound very similar in "speed" or "groove" to a 200bpm song, as the speed of the song is relative to how many beats/sounds are actually played a second. Therefore, given the unique properties of 'tempo', we believe it is a good canditate to use the smoothing techniques on. From our domain knowledge, we also know that tempos like 90-150 are very common in popular and danceable songs. From this knowledge, and from our observations of the plots above, we determined six ranges of interest $(min, 90), [90, 100), [100, 110), [110, 120)[120, 130), [130, max)$ we determined to be appropriate following stepwise smoothing. In the plots above, you will see red line indicating the knot tempos for these ranges. For the smoothing splines models, we considered 3,4,5 and 6 knot models.

Splines

```
## # A tibble: 5 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     0.257     0.0271     9.47  8.88e-21   0.204     0.310
## 2 spline_basis1   0.274     0.0331     8.28  2.46e-16   0.209     0.339
## 3 spline_basis2   0.263     0.0285     9.25  6.16e-20   0.208     0.319
## 4 spline_basis3   0.213     0.0290     7.35  2.96e-13   0.156     0.270
## 5 spline_basis4  -0.0157    0.0362    -0.434 6.65e- 1  -0.0868    0.0554
## # A tibble: 6 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     0.296     0.0407      7.27 5.34e-13   0.216     0.376
## 2 spline_basis1   0.151     0.0496      3.04 2.42e- 3   0.0534    0.248
## 3 spline_basis2   0.244     0.0409      5.95 3.16e- 9   0.163     0.324
## 4 spline_basis3   0.223     0.0426      5.23 1.92e- 7   0.139     0.306
## 5 spline_basis4   0.175     0.0417      4.18 3.02e- 5   0.0927    0.256
## 6 spline_basis5  -0.0549    0.0474     -1.16 2.47e- 1  -0.148     0.0381
## # A tibble: 7 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     0.299     0.0408      7.33 3.55e-13   0.219     0.379
## 2 spline_basis1   0.146     0.0499      2.92 3.52e- 3   0.0479    0.243
## 3 spline_basis2   0.248     0.0411      6.03 2.03e- 9   0.167     0.328
## 4 spline_basis3   0.212     0.0452      4.70 2.81e- 6   0.124     0.301
## 5 spline_basis4   0.226     0.0427      5.29 1.36e- 7   0.142     0.310
## 6 spline_basis5   0.170     0.0419      4.07 4.97e- 5   0.0883    0.253
## 7 spline_basis6  -0.0565    0.0475     -1.19 2.34e- 1  -0.150     0.0365
## # A tibble: 8 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     0.299     0.0408     7.33  3.53e-13   0.219     0.379
## 2 spline_basis1   0.146     0.0498     2.93  3.48e- 3   0.0481    0.244
## 3 spline_basis2   0.248     0.0411     6.03  2.03e- 9   0.167     0.328
## 4 spline_basis3   0.213     0.0452     4.72  2.53e- 6   0.125     0.302
## 5 spline_basis4   0.223     0.0428     5.22  2.05e- 7   0.139     0.307
## 6 spline_basis5   0.180     0.0424     4.25  2.28e- 5   0.0970    0.263
```

```
## 7 spline_basis6   0.0964    0.0432     2.23  2.56e- 2   0.0118    0.181
## 8 spline_basis7  -0.0228    0.0529    -0.432 6.66e- 1  -0.126     0.0809
```

Figure : Spline Fits with Different Knots

LOESS

For the LOESS models, we considered spans of 0.05, 0.1, 0.2 and 0.3.

(Commentary)

Comparing all Smoothing Models

Figure : Combined Comparison of Spline and LOESS Models

(Commentary)

10. Principle Component Regression

(and/or Bisquare)

(Background / Research)

```
## Data:    X dimension: 1726 33
##  Y dimension: 1726 1
## Fit method: svdpc
## Number of components considered: 33
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV          0.1818   0.1737   0.1720   0.1696   0.1699   0.1686   0.1664
## adjCV       0.1818   0.1737   0.1719   0.1695   0.1698   0.1686   0.1663
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.1662   0.1649   0.1648    0.1633    0.1613    0.1593    0.1554
## adjCV   0.1662   0.1647   0.1649    0.1631    0.1622    0.1592    0.1551
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## CV       0.1527    0.1517    0.1512    0.1513    0.1506    0.1498    0.1498
## adjCV    0.1525    0.1518    0.1511    0.1512    0.1505    0.1497    0.1497
##        21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps
## CV       0.1499    0.1459    0.1443    0.1442    0.1440    0.1442    0.1441
## adjCV    0.1499    0.1457    0.1442    0.1441    0.1439    0.1441    0.1440
##        28 comps  29 comps  30 comps  31 comps  32 comps  33 comps
## CV       0.1441    0.1442    0.1431    0.1431    0.1427    0.1428
## adjCV    0.1440    0.1441    0.1430    0.1430    0.1425    0.1426
##
## TRAINING: % variance explained
##                      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                     17.083    28.58    35.04    40.83    45.67    50.12
## danceability_square    8.832    10.80    13.40    13.40    14.84    17.12
##                      7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## X                      54.02    57.72    61.30     64.71     68.02     71.28
## danceability_square    17.18    18.88    18.88     20.86     21.72     24.68
##                      13 comps  14 comps  15 comps  16 comps  17 comps  18 comps
## X                       74.28     77.19     79.88     82.51     84.95     87.21
## danceability_square     28.21     31.13     31.44     32.11     32.24     32.82
##                      19 comps  20 comps  21 comps  22 comps  23 comps  24 comps
## X                       89.31     91.12     92.91     94.43     95.81     97.00
## danceability_square     33.60     33.60     33.64     37.21     38.46     38.58
##                      25 comps  26 comps  27 comps  28 comps  29 comps  30 comps
## X                       98.13     98.71     99.22     99.50     99.74     99.89
## danceability_square     38.76     38.77     38.96     38.96     39.06     40.21
##                      31 comps  32 comps  33 comps
## X                       99.96    100.00    100.00
## danceability_square     40.33     40.73     40.73
```

](Commentary, cross validation)

11. Regression Trees

Another new addition was CART (Classification and Regression Tree) model into the linear model for predicting danceability in songs is a valuable addition, as it allows for capturing potential nonlinear relationships and interactions between the predictor variables, which a linear model alone may not be able to capture effectively.

Danceability is a complex characteristic influenced by various musical elements like tempo, rhythm, and energy. CART offers a powerful non-parametric method well-suited for this type of problem. Unlike linear regression which assumes linear relationships between predictors and target, CART can automatically capture non-linear patterns and intricate interactions among audio features. This flexibility makes it an attractive choice when the underlying relationships are unknown or difficult to specify upfront. It serves as an excellent comparison to the previous linear models we have created because CART requires minimal data preprocessing and feature engineering compared to linear methods. While transformations and variable selection were crucial steps for the linear models, CART can automatically handle raw features, different scales, outliers, and mixed data types without those pre-processing steps. Additionally, unlike the parametric linear models which make assumptions about the form of the relationships, CART is non-parametric and can flexibly model any non-linear and interaction effects present in the data.

The way CART works is that it recursively partitions the data into subsets based on input feature values, creating a tree-like structure of decision rules. For regression tasks like predicting danceability scores, CART aims to create terminal nodes (leaves) with minimal variance in the target variable. It then uses greedy algorithms to grow this tree by selecting splits that minimize a certain metric, such as mean squared error. However, unconstrained tree growth can lead to overfitting the training data. This is where cross-validation becomes crucial. By systematically holding out parts of the data during model training, cross-validation provides an unbiased estimate of the model's predictive performance on unseen data. This estimate guides the pruning process, allowing for the selection of an optimal tree size that balances underfitting and overfitting.

By comparing the predictive performance of CART to the linear, regularized, and non-linear models, it is possible to gain insight on whether predicting danceability is better approximated by additive and interactive linear components or if there are strong hierarchical interactions that tree-based methods more naturally uncover. Therefore, using CART as a comparative approach, given the extensive linear modeling efforts already undertaken, is a well-justified methodological decision.

```
## CART
##
## 2000 samples
##   13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1800, 1800, 1800, 1799, 1799, 1801, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE       Rsquared   MAE
##   0.04249368  0.1251459  0.2088939  0.09853466
##   0.06880619  0.1304675  0.1408492  0.10237627
##   0.12733186  0.1373488  0.0913364  0.10903216
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.04249368.
```

```
## CART
##
## 1726 samples
##   13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1553, 1555, 1552, 1554, 1555, 1552, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE       Rsquared    MAE
##   0.04089228  0.1693909  0.13772748  0.1349392
##   0.04640580  0.1723017  0.10657318  0.1373587
##   0.12162621  0.1777217  0.08071604  0.1425856
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.04089228.
```

Here, we are using a 10 fold cross validation on 2 separate models: one with the transformed variables and another with the original variables from our song data. The results show that the models with the original model than a higher R^2 and cp are shown, even though RSME and MAE remain better for the model with transformed variables.

The higher RMSE and MAE values for the original variable model indicate it has higher overall predictive accuracy on the danceability target. These error metrics directly evaluate how close the predictions are to the actual values. So in terms of pure predictive performance, the transformed model seems preferable.

However, the higher R-squared of the original variable model suggests it explains a higher percentage of the variance in the danceability scores. R-squared reflects how well the model captures the underlying patterns and relationships in the data. The complexity parameter (CP) of the 2 models is pretty similar, with the transformed variable model being slightly better. Overall, it is very clear that the original variable model outperformed the transfomred variable model.

The superior performance of the untransformed CART model suggests that the danceability dynamics are potentially better represented by the hierarchical tree structure learned directly from the raw features, rather than the restricted parametric forms imposed by linear model transformations. This flexibility of CART to model complex patterns "as is" renders variable transformations unnecessary and even counterproductive in this case.

The results indicating that the CART model with all the original variables intact outperformed the version using transformed variables are insightful. This outcome shows the strength tree-based approaches, their ability to naturally handle different scales, distributions and complex non-linear patterns without transforming any variables.

Ultimately, both models provide valid perspectives worth considering. The results showcase how transformations can impact different performance aspects of tree models in contrasting ways.

```
##                    Overall
## tempo            0.80182866
## energy           0.48616688
## speechiness      0.46164957
## valence          0.43689465
## acousticness     0.31057907
## explicit         0.28933363
## duration_scale   0.18507507
## liveness         0.08848534
```

```
## loudness         0.04633674
## instrumentalness 0.00000000
## mode             0.00000000
## popularity       0.00000000
## key              0.00000000
```

An important output from CART models is the measure of variable importance, which quantifies each predictor's contribution to the overall variance explained by the tree. Variables towards the top split nodes carry more weight, while lower rankings imply lesser contributions after accounting for the other predictors. The relatively high importance value of 0.80 for tempo indicates it was the most influential variable in constructing the regression tree to predict danceability scores. This aligns with intuition as people tend to dance more to song with higher speeds and tempos. Energy, speechiness, and valence are also shown to be significant factors which again, makes sense to its correlation to danceability.

Finally, here is the regression tree visualized. For each leaf node displays the mean predicted target value for that subset of observations. The tree diagram includes root nodes, internal split nodes, and terminal leaf nodes which provides an intuitive representation of how different predictor variables interact to partition the data into distinct subgroups.

Final Test of Models (choose one of LASSO, MLR, SMOOTHING, PCR, OTHER)

Summary of Findings

(Tempo is a good predictor of danceability, when broken up, LASSO is good, show model, etc.)

In our search for the best way to predict danceabiltiy within songs, we sent through an intitial thorough process of feature engineering and variable transformations to confine to assumptions to a linear model. This included squaring our response danceability variable, log and sqrt transforming a lot of predicting varaibles, separating variable like genre into sub-columns, and even applying a new transformation technique(Yeo-Johnson) for duration. Afterward, we ensured to do a thorough outlier analysis using multiple metrics such as studentized residuals, cook's distance, and DFFITs/DFBETAs.

When our preprocessing finished and it was time to build our model, we used multiple techniques including using our intuition, stepwise functions, smoothing and shrinking techniques, principal component regression, and even regression trees to see what would be the best predictor. After comparing all these techniques, we found the best model as the spline model with 5 knots. We believe that through this model, we can predict danceability. in tempo ranges min to 160 without extrapolating and with reasonable success given the complexity of danceability. Factors such as different cultures, personalities, and artist intentions all factor into how difficult it is to access danceabiltiy, so we are very proud of our results.

- Report as if to a client. (DO THIS) If you are concluding with a model, report the entire model (variables, coefficient estimates, and p-values). Also, include the residual plot which provides the evidence that the model is appropriate. If you aren't reporting a model, there should be a reason why you don't have a model for the client

Therefore, it turns out the simplier models ended up providing the best results. Since our spline model had only one predictor, tempo, the relationships between the predictor and the response variable are easier to interpret and we put more effort in capturing any non-linear relationships. Furthermore a simple model allows one to impliment and fix easier, as well as avoiding unneccassary complexity. An additional example of this was seen through our analysis of regression trees as the model with the original variables outperformed the ones with all of our complicated variable transformations. A tree's structure which deempphasizes the importance of linearity and various other assumptions has shown to have more flexibility and better results for a simpler model.

8. Conclusion

Within this project, our study aimed to find the the factors that influence the "danceability" of a song from a combination of audio features within the song, its popularity, genre, and other characteristics. We hope our findings will help musical artists and composers have a better understanding on "danceability" and its connection to the music world.

We aimed to create a valuable statistical model danceability as our main variable. Our first step in the process was preprocessing, which included feature engineering, transforming certain variables to fit the assumptions of the normal errors models, and checking for multicollinearity and high influence/leverage point. When it came to building our model, we employed stepwise selection models with f-tests to find the optimal variables to use for our model.

We found that the Forward Selected Stepwise BIC exhibited a better adjusted R^2, AIC, and BIC metrics, as well significant p-values to variable coefficients, indicating better suitability to predict/interpret danceability. Our final model indicates variables like valence, instrumentalness, and genre selection proved valuable positive effects on danceability while energy and tempo showed negative effects.

Further exploration could include the consideration of other model selection techniques such as Lasso or Ridge regression in order to select important predictors for our model better. We could also include a more deeper analysis on the effect of multicollinearity and potential outliers. We may also explore different shrinkage and smoothing methods. Overall, our current model can be seem as valuable for now based on the accuracy of our model and the precaution taken to prevent overfitting and overcomplexity.