Analyzing the Impact of Various Audio and Musical Characteristics on the Popularity of Spotify Tracks

By Diego Osborn, Elijah Karp, Eunice Cho, Jing Yang, and Reema Alsaeed

Music streaming platforms, such as Spotify, provide a vast amount of data on songs, including audio characteristics and popularity metrics. However, it is not always clear which factors influence a song's popularity the most. In our project, we aim to analyze key audio features, such as valence, loudness, and tempo, to determine their relationship with a song's popularity. Additionally, we will investigate the differences across genres and explicitness on a song's popularity. Through statistical analysis, we aim to uncover patterns that may help explain what makes a song popular.

Understanding the factors that contribute to song popularity can be extremely valuable for artists, producers, labels, and music streaming services. If certain factors are consistently associated with high popularity, artists, producers, and labels could focus on those factors and optimize their compositions to produce a song that has a high chance of being popular. For music streaming services, knowing those factors can help them identify songs that optimize these factors and advertise them more to draw more people to spend more time on their platforms.

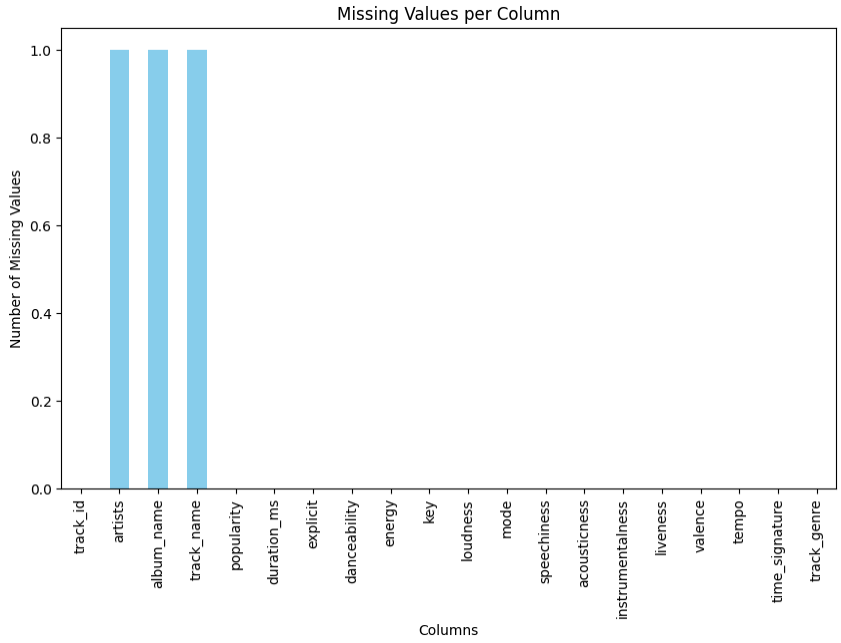
The dataset we are using for our project is the Spotify Tracks Dataset from Kaggle: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset. This dataset contains detailed information on over 100 thousand tracks, including various audio features and popularity scores based on Spotify's algorithm. The dataset allows us to analyze the relationship between many features of a song and its popularity. Additionally, the dataset has a range of 125 different genres, allowing us to investigate the relationship between a song's genre and its popularity score.

The dataset contains track metadata: `track_id`: the Spotify ID for the track; `artists`: the artists' names who performed the track; `album_name`: the album name in which the track appears; `track_name`: name of the track; and `track_genre`: the genre in which the track belongs, it contains `popularity`: the popularity of a track is a value between 0 and 100 (with 100 being the most popular), `duration_ms`: the track length in milliseconds, `explicit`: whether or not the track has explicit lyrics, audio features: `danceability`: describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is the least danceable, and 1.0 is the most danceable; `energy`: a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy; `loudness`: The overall loudness of a track in decibels (dB); `valence`: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track (tracks with high valence sound more positive (e.g. happy, cheerful), while tracks with low valence sound more negative (e.g. sad, angry)); `speechiness`: detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audio book), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks; `acousticness`: a confidence measure from 0.0 to 1.0 of whether the track is acoustic, where 1.0 represents high confidence that the

track is acoustic; `instrumentalness`: predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater the likelihood that the track contains no vocal content; `liveness`: detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live; `tempo`: the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration; and musical attributes: `key`: the key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1; `time_signature`: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7, indicating time signatures of 3/4 to 7/4; `mode`: indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1, and minor is 0.

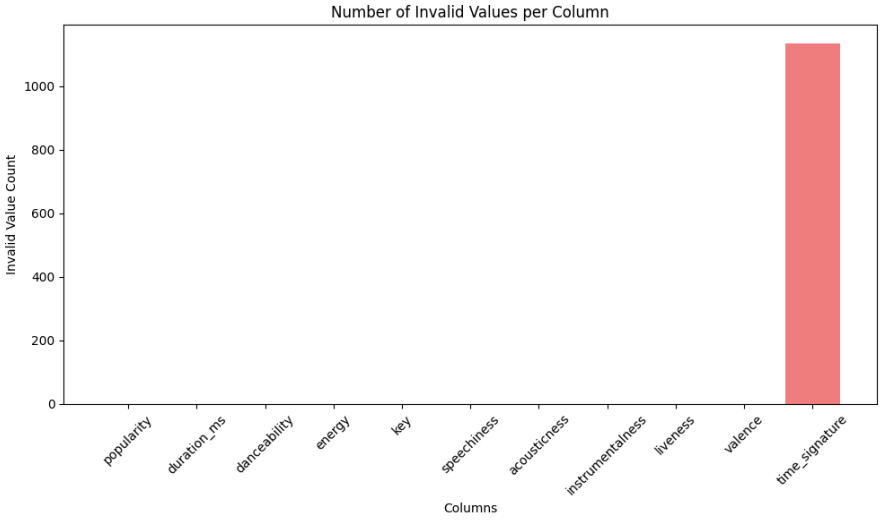
To better understand the dataset, we conducted an exploratory data analysis (EDA) before performing our analysis. This allows us to identify patterns, detect missing or invalid values, and explore some of the relationships between the different variables.

First, we examined the dataset for any missing values, which we can see through this plot:



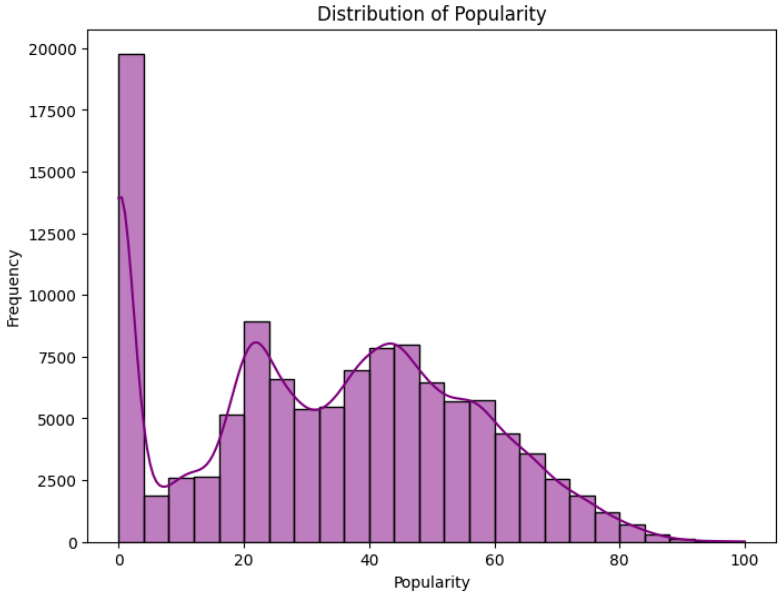
There appears to be one missing value in “artists”, “album_name”, and “track_name”. After further investigation, we discovered that those three missing values occurred in the same row. Additionally, this row had a “duration_ms” value of zero, which is not a valid input for a song’s duration. Since this row had multiple inconsistencies across multiple variables, we decided to remove it.

To further assess the data, we check for invalid values in the columns for which we know the value range. We have found the following:



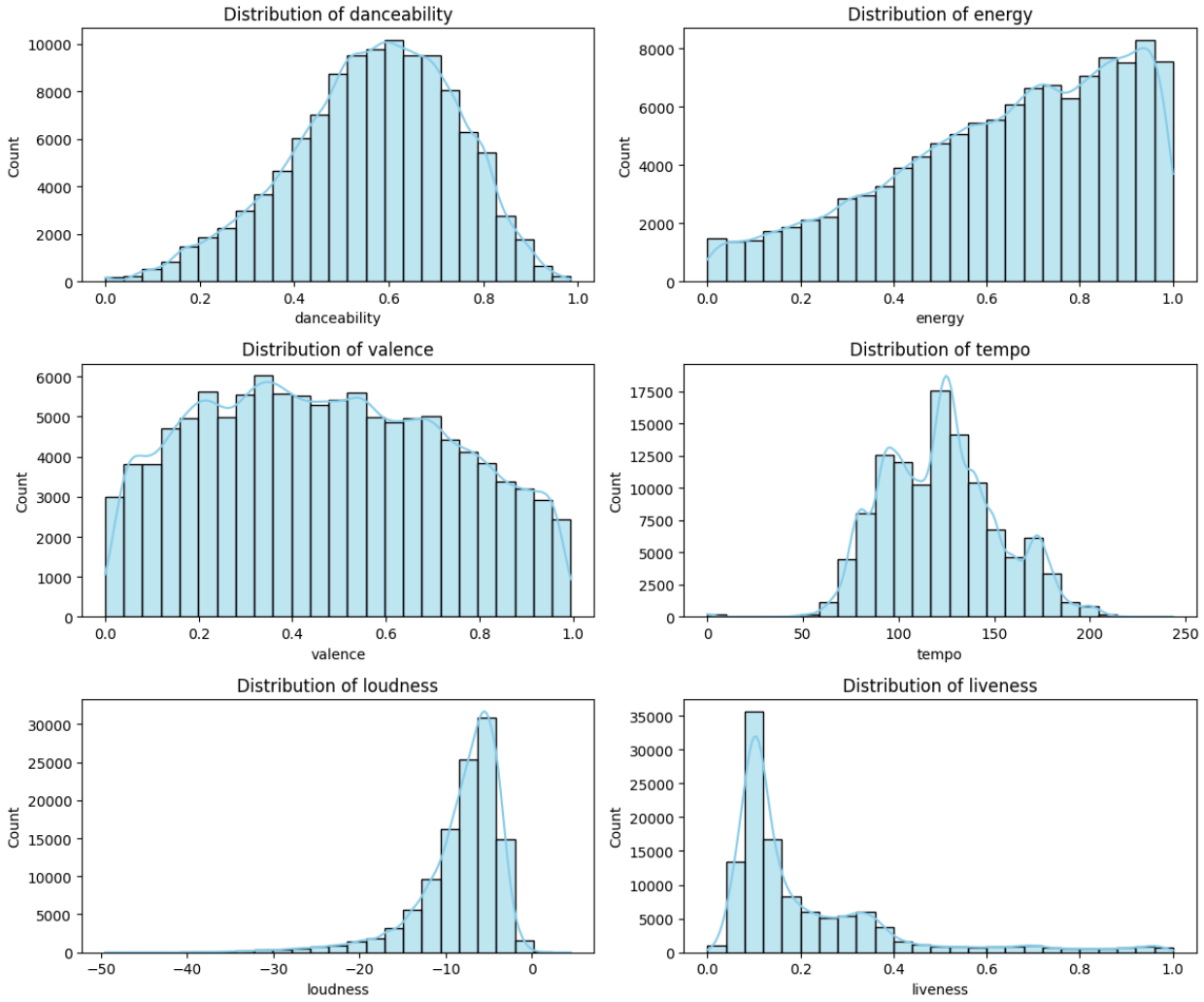
The only column with a defined range containing invalid values was `time_signature`, where 11136 rows had inputs outside of the given range. Since a reasonable method of imputation was not apparent for us to do here, we opted to remove these rows to maintain data quality.

Now, with a cleaned dataset, we proceeded to examine the distribution of song popularity:



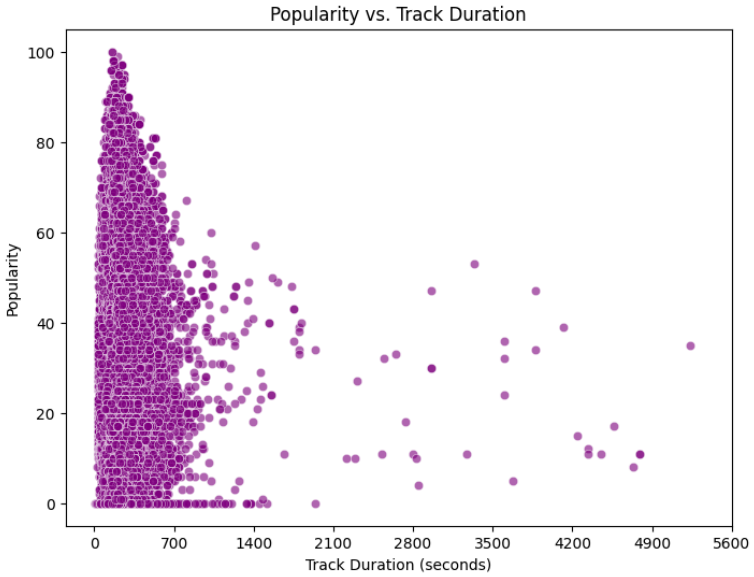
The distribution is right-skewed, with a significant number of tracks having a popularity score of zero. This could suggest that a lot of the songs in the dataset are rarely played as they are not as popular.

We also examined the distribution of some of the key audio features to understand their trends. The plot below showcases the distribution of danceability, energy, valence, tempo, loudness, and liveness:



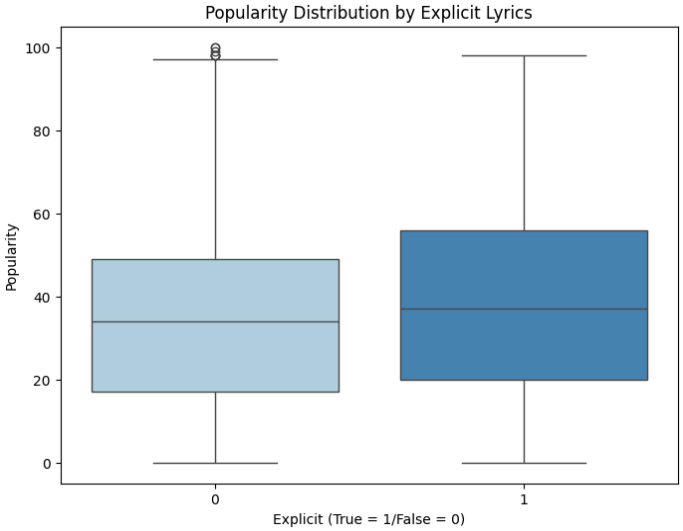
For `danceability`, `energy`, and `loudness`, the distribution appears to be left-skewed. This suggests that most tracks have moderate to high values in those features, meaning that many songs are energetic, loud, and have high danceability. For `valence`, the distribution is somewhat even throughout, indicating that our dataset contains a balanced number of positive (happy) and negative (sad) sounding tracks. For `liveness`, the distribution appears to be right-skewed, which suggests that a large portion of the tracks have a lower audience presence. This indicates that most songs were more likely recorded in studios rather than performed live in front of an audience.

Next, we explored the relationship between track duration and its popularity score:



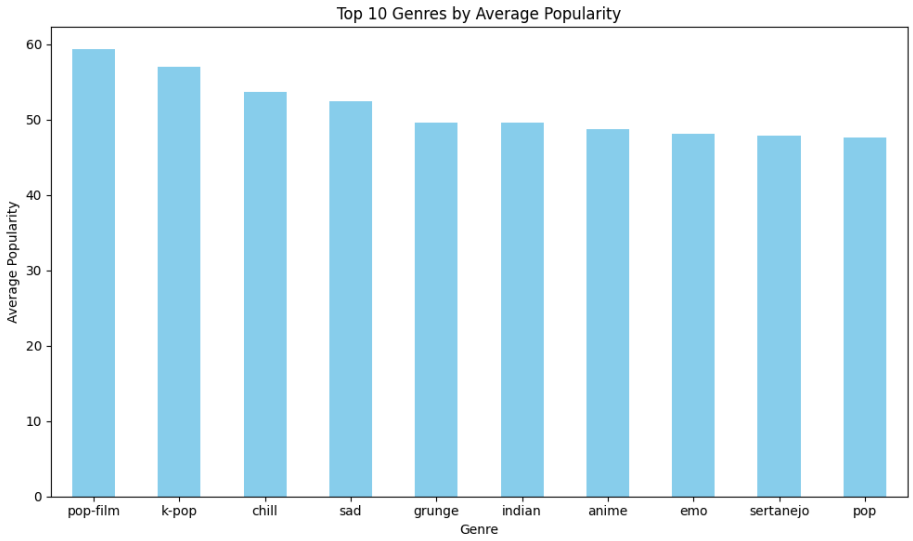
To make the plot easier to understand, we converted the `duration_ms` column from milliseconds to seconds. By looking at the plot, we can see that most tracks are under 700 seconds (around 11 minutes). There does not appear to be a strong visible correlation between a song's duration and its popularity score, though extremely long songs seem to have lower popularity scores. However, we can see that the shorter songs have a popularity score that ranges from 0 to 100, so we cannot say there is a clear, strong correlation between duration and popularity.

Furthermore, we examined whether explicit lyrics influence popularity. We compared the distributions of explicit and non-explicit songs:

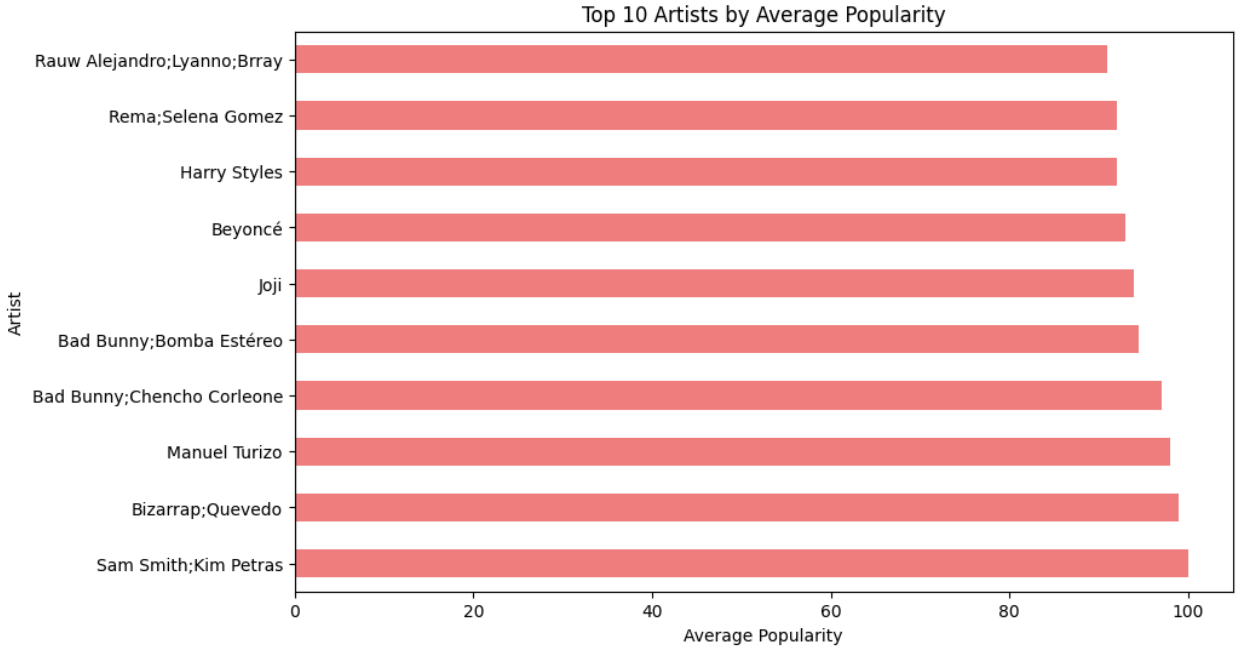


Through the boxplot, we can observe some differences, however, they are not significant enough to suggest that explicit content has a significant effect on a song's popularity score.

We also identified the top 10 most popular genres based on their average popularity score in the dataset:

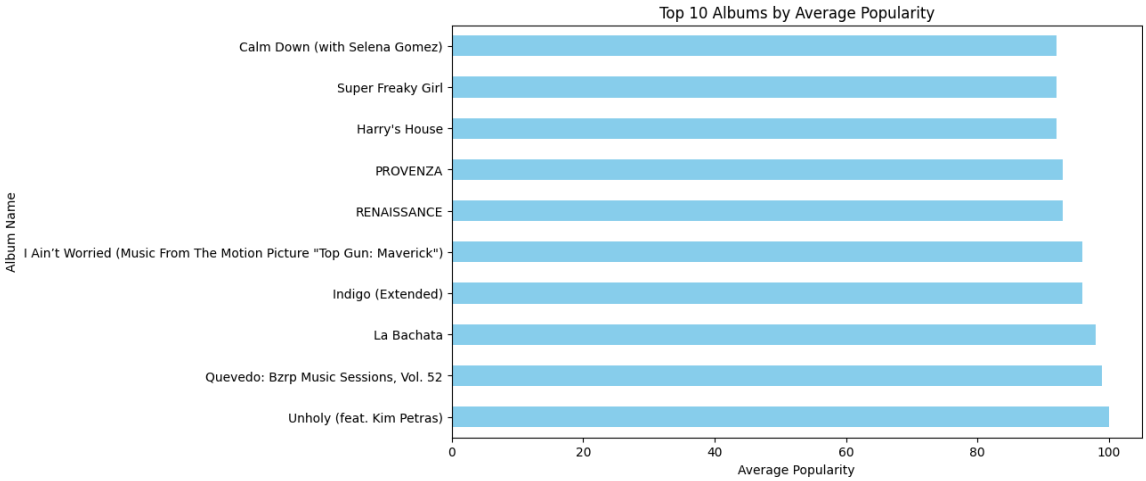


And the top 10 most popular artists based on their average popularity score:

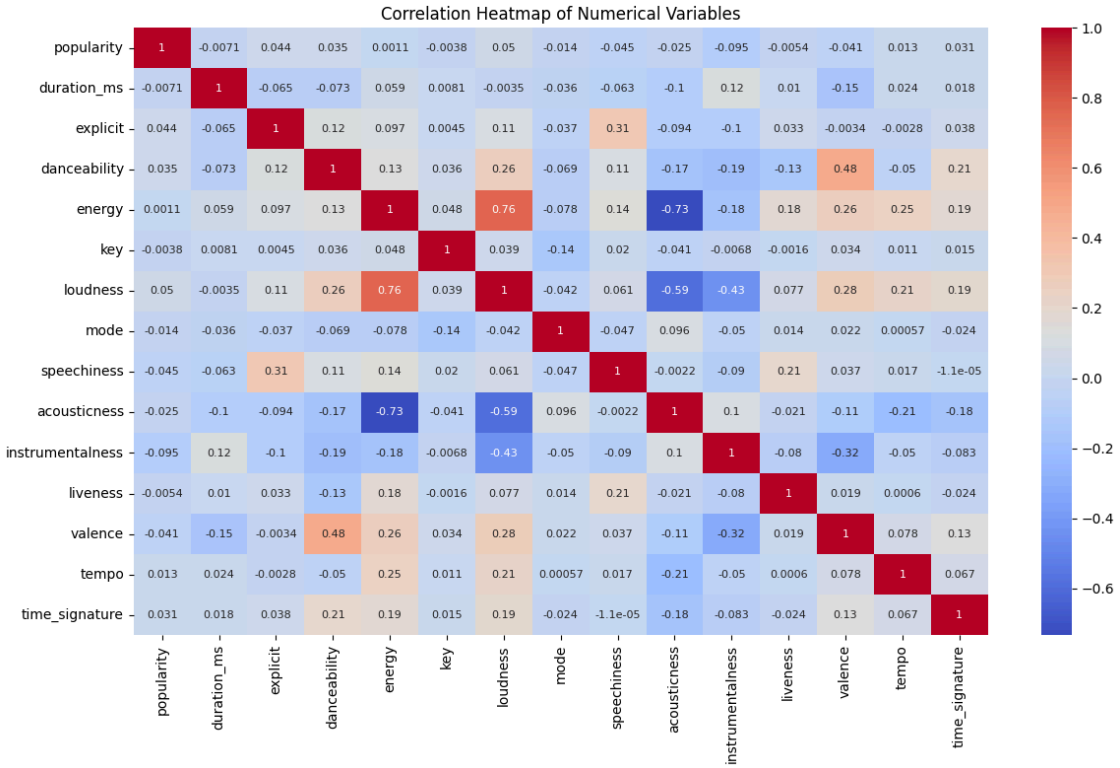


For some of those artists, we see multiple names, which indicate a collaboration among those artists. However, it is worth noting that some of these artists or collaborations have only a few songs, while others have many in the dataset. So, a high average popularity score here could be because there is only one song under that artist, which turned out to be a popular song.

We also identified the top 10 most popular albums based on their average popularity score:



Lastly, to summarize the correlation between numerical features, we did a correlation heatmap:



This heat map provides a high-level overview of how different numerical features interact with each other. Some key findings from the correlation heatmap: there is a strong positive correlation (0.76) between energy and loudness–which is expected as louder songs tend to be more energetic, acousticness has strong negative correlations with energy (-0.73) and loudness (-0.59) –which is expected as they are opposite song attributes, there is a negative correlation (-0.43) between instrumentalness and energy–which indicates that instrumental tracks tend to be less energetic, popularity does not have a strong correlation with any single feature–which reinforces the idea that multiple factors contribute to a

song's popularity score and not just one feature. Overall, the insights we have gained from the exploratory data analysis (EDA) will help guide us in our analysis, where we will investigate how features impact a song's popularity.

There have been other analyses already performed on this data, such as a linear regression to predict popularity based on loudness[1], a random forest regressor model to predict popularity[2], a decision tree classifier to predict popularity[3], an XGBoost regressor model to predict popularity[4], a Neural Net model to predict the likelihood of a song being recommended[5], and there are more analyses of the spotify dataset under the **Code** section in kaggle[6].

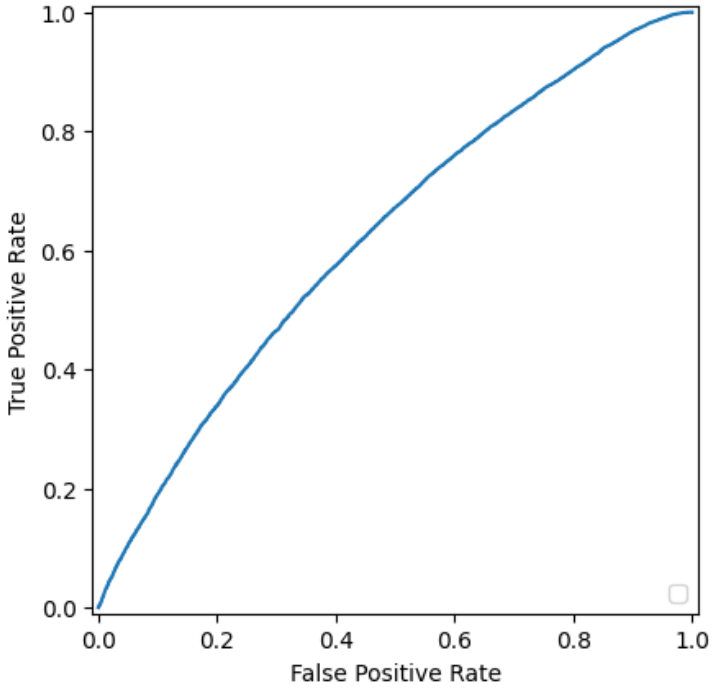
For our analyses, we performed a linear regression model to predict `popularity`, a logistic regression model to predict `binary_popularity`, which is a binary feature that represents a 0 if the popularity score is below or equal to 50 (not popular) and a 1 if the popularity score is above 50 (popular), a ridge regression model to predict `binary_popularity`, and a principal component analysis to reduce dimensionality and analyze the variance in musical attributes across different songs.

For linear regression, we used our numerical features as the predictor features and `popularity` as our target. We wanted to understand the direct linear relationship between our predictor features and the target `popularity` score. Our R-squared value was 0.025, meaning that the model only explained 2.5% of the variance in `popularity`. Meaning that other factors may not be included in our dataset that do influence `popularity`. Many of the predictors had p-values < 0.05, meaning they had a significant impact on popularity. We printed the VIF values and saw that they were all below 5, meaning that multicollinearity was not a concern. Our highest VIFs were `energy`: 4.268 and `loudness`: 3.286. Based on our residual plots, homoscedasticity was not satisfied as there was a pattern regarding the residuals and the fitted values. We found that there was a positive correlation in the residuals based on the Durbin-Watson statistic 0.571, the Q-Q plot deviated significantly on both tails, and the heavy tails suggested non-linearity, therefore, we concluded non-normality for our linear regression.

For our logistic regression, we used our numerical features as the predictor features and popularity as our target. We transformed the continuous popularity score into a binary target by creating a new column called binary_popularity. If the popularity score was 0-50, it was categorized as not popular, and 50-100 was categorized as popular. We wanted to see the factors that distinguish between "popular" and "not popular" songs. We had a pseudo-R-squared value of 0.03304, meaning that our model only explained 3.3% of the variance in song popularity, which is notably low. Our log-likelihood score was -48549, which is higher compared to the null model likelihood score at -50208. Meaning our model improved in log-likelihood by 1659 units compared to the null model. The null log-likelihood is represented as the log-likelihood of a model with only the intercept and no predictors. It represents how well we could predict a song's popularity without any feature information. Although our model showed some improvement, it was a very minimal improvement. Our LLR p-value was 0, meaning that we are confident that at least some of our predictors had a relationship with popularity. When looking at our

[1] https://www.kaggle.com/code/alexriverau/linear-regression
[2] https://www.kaggle.com/code/alexriverau/linear-regression
[3] https://www.kaggle.com/code/shivamagarawal/spotifypopularity-decisiontree
[4] https://www.kaggle.com/code/arnabdutta6/spotify-popularity-prediction-analysis-modelling#Model-training
[5] https://www.kaggle.com/code/abbassiddiqui1/spotify-track-recommender
[6] https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/code

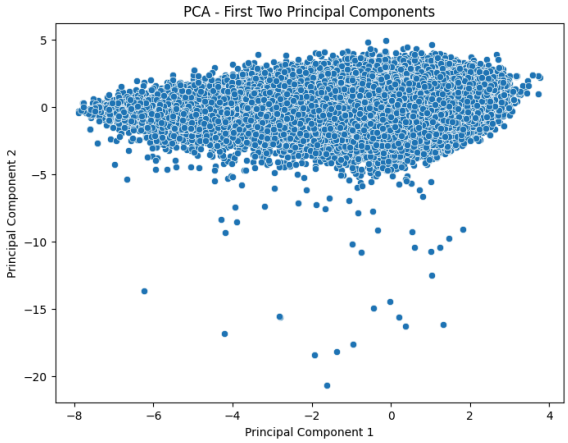
coefficients, all of them were below the significance of 5% except for the feature `key`. The negative features include `speechiness`, `instrumentalness`, `liveness`, `energy`, and `acousticness`. When these features increase, the predicted target values tend to decrease. The positive features include `explicit`, `danceability`, and `time_signature`. These features tend to increase popularity. The features with small coefficients include `key`, `loudness`, and `tempo`. These had a minimal impact on popularity compared to the other features. Our AUC score was 0.6226, meaning it performed only slightly better than random guessing (0.5). The intercept was -0.3009, meaning the estimated popularity score, on average, when all the covariates are zero. The coefficient of `speechiness` was -1.7619, meaning the more speech a song had, the lower the `binary_popularity` score. The coefficient of `liveliness` was -0.9300, meaning songs that sound more like live recordings tend to be less popular on average. The coefficient of `instrumentalness` was -0.8375, meaning songs that are more instrumental (with less vocals) have a lower `binary_popularity` score on average. The coefficient of `energy` was -0.7274, meaning lower-energy songs tend to be more popular on average. The coefficient of `acousticness` was -0.7274, meaning less acoustic songs (more electronic) tend to be more popular on average. The coefficient of `duration_ms` was -0.1549, meaning shorter songs tend to be more popular on average. The coefficient of `danceability` was 0.7640, meaning that more danceable songs are more likely to be popular on average. The coefficient of `explicit` was 0.4142, meaning explicit songs have higher odds of being popular on average. The coefficient of `time_signature` was 0.1000, meaning when the time signature is higher, it increases the odds of the song being popular, on average. For all these features, the log odds either increase or decrease based on the coefficient.

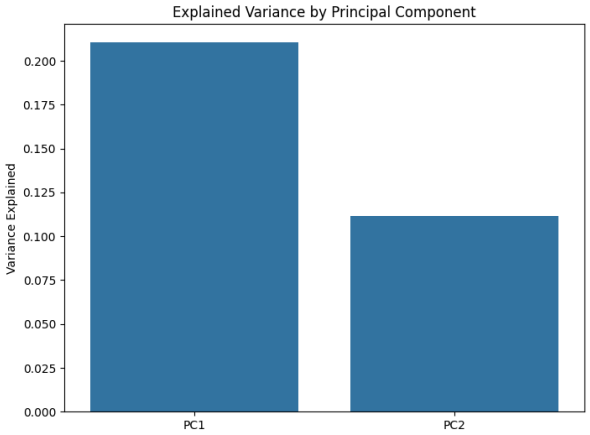


For Ridge Regression, we wanted to address potential multicollinearity issues among the predicted features while predicting the `binary_popularity`. This technique helps stabilize coefficient estimates when the predictors are correlated. We got a mean squared error of 0.1772 and an R-squared value of 0.037. Meaning our model only explained 3.7% of the variance in `binary_popularity`. The intercept was -0.289, meaning it is the `binary_popularity` score when all the other values are zero, on average. Our top negative predictors included `speechiness`, `valence`, `liveliness`, `instrumentalness`, and `energy`. Our top positive predictors included `danceability`, `explicitness`, and `time_ signature`. The features with the smallest impact

included `key`, `tempo`, and `loudness`. Each of these mean that for every unit increase or decrease in the feature, `binary_popularity` will increase or decrease by that many units, on average. When comparing Ridge to Logistic regression, the order of importance was slightly different, however, the same features were shown as statistically significant. Both models identified `speechiness`, `valence`, `instrumentalness`, and `liveness` as strong negative features. `danceability` and `explicit` were also key positive predictors in both models.

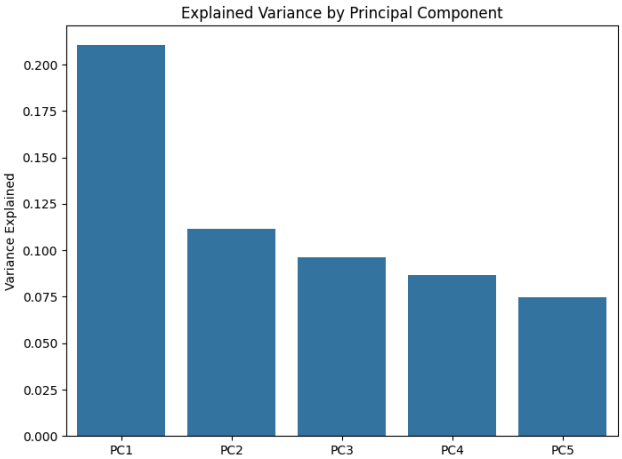
We used Principal Component Analysis to reduce dimensionality while retaining as much variance as possible. This helped us identify underlying patterns and relationships between features that were not immediately apparent in the original feature space. We did it first with two principal components and then with five components. PC1 captured 21.04% of the variance of the features, and PC2 captured 11.16% of the variance of the features, meaning that the total variance of PC1 and PC2 was 32.2%. Since we had a variance of 21.04% and 11.16%, this shows that there was a significant correlation with our features. This suggests that our features contain complex relationships that can't just be reduced to two dimensions.



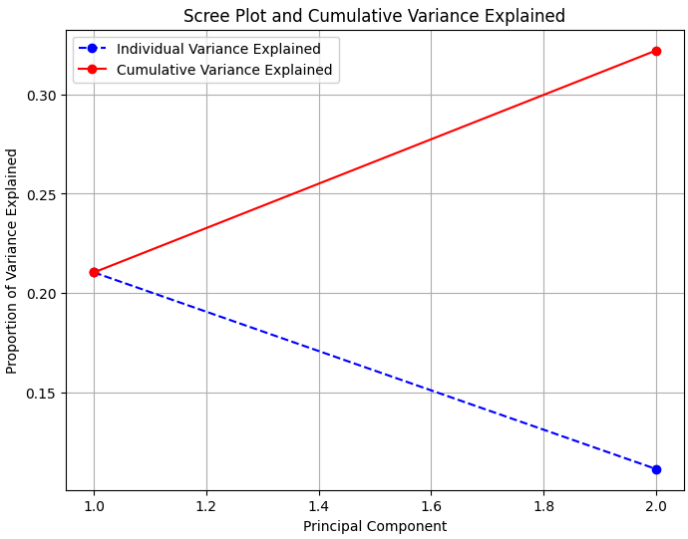


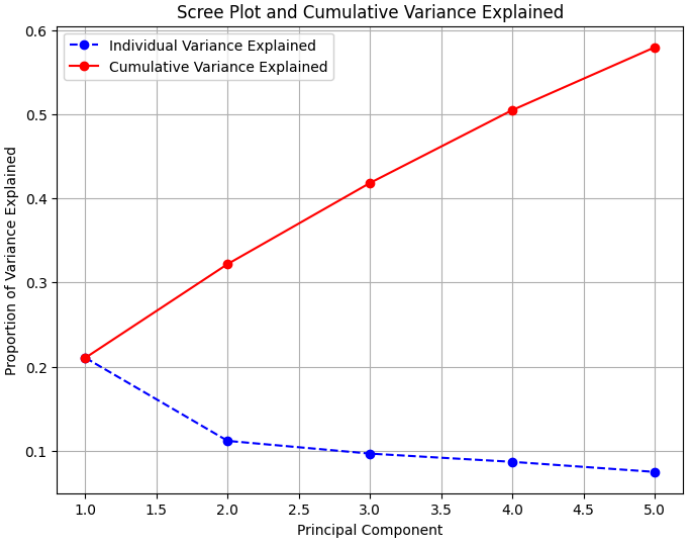
Our predicted features had a distinct distribution, with most of the data points clustered together, however, some outliers did drop below the cluster. These outliers represent songs that had significantly different feature patterns compared to a majority of the data. These songs could have had an unusual combination of features that affected their popularity in different ways compared to the rest of the songs.

When we included 5 components, 57.98% of the variance among the features was explained. PC1 explained 21.04% of the total variance, PC2 explained 11.16% of the variance, PC3 explained 9.65% of the variance, PC4 explained 8.67% of the variance, and PC5 explained 7.47% of the variance.



We primarily used PCA to reduce the dimensionality of the dataset while still retaining as much variance as possible. We found that the first 3 components explained 41.84% of the variance, which is a good balance between reducing complexity and retaining the information. To have captured 50% of the variance, we needed 4 principal components.





The scree plot showed the proportion of variance explained by each principal component. As seen, there was a notable drop from PC1 to PC2, though it was not extremely steep. After PC2, the explained variance continued to decrease steadily. Since there was not a clear cut-off point for the number of components to retain, we could see that the features had complex relationships that couldn't be explained in a small number of dimensions.

While our analysis provides insights into song popularity based on audio features, there are some potential limitations and shortcomings of our analyses that should be acknowledged. One of the primary limitations is that Spotify's popularity scores are dynamic. Popularity scores change over time due to various reasons, such as trends, meaning that our analysis captures only a snapshot and not long-term popularity trends. Additionally, popularity scores are influenced by factors beyond what our dataset contains. Popularity is determined by more than audio features; factors such as marketing efforts and artist reputation likely play major roles. However, these factors are not included in our dataset, which limits our ability to fully predict popularity. Furthermore, our model simplifies popularity into a binary outcome, removing the intermediate popularity levels. This means that songs with moderate popularity scores are considered the same as those with either extremely low or high scores. Lastly, although PCA helped reduce dimensionality between features, we observed that the variance was spread across multiple principal components instead of just one or two. By looking at the scree plot, the variance explained by each component decreased gradually but with no clear cutoff to select the optimal number of components. To address these potential limitations and shortcomings, future work could integrate real-time streaming

trends and include additional factors such as artist reputation and marketing efforts. Additionally, exploring alternative models could improve predictive performance. Lastly, expanding the dataset to include other music streaming platforms would provide a more comprehensive view of music trends and popularity.

Our final report does not relate to our initial proposal, as we ended up working with a completely different dataset and research question. In our initial proposal, we aimed at analyzing how movie budgets and genres influence worldwide gross earnings, and our data included the movie's title, average rating, genre, release date, budget, worldwide gross, and additional attributes. However, as we began working with this dataset, we encountered a major limitation that our dataset didn't contain enough numerical features to analyze what truly influences worldwide gross earnings outside of categorical features. Without sufficient numerical features, it became difficult to explore these relationships and apply analyses. As a result, we pivoted to a dataset that contained numerous numerical features in the Spotify tracks dataset. This dataset allowed us to explore how various audio and music characteristics influence a song's popularity. This aligns with our initial goal of understanding what factors contribute to success, however, now it is applied to music instead of movies. Although the shift between our initial proposal and final project is major, the analytical approach remained somewhat similar. In both cases, we wanted to discover patterns, determine which factors have the strongest influence on an outcome, and apply statistical analysis to support our findings. The switch to the Spotify Tracks dataset provided us with a more feasible project than our initial proposal, but remains an engaging and relevant topic for analysis.

References

https://www.kaggle.com/code/alexriverau/linear-regression
https://www.kaggle.com/code/beamendon/popularity-prediction#3.-Finding-best-model
https://www.kaggle.com/code/shivamagarawal/spotifypopularity-decisiontree
https://www.kaggle.com/code/arnabdutta6/spotify-popularity-prediction-analysis-modelling#Model-training
https://www.kaggle.com/code/abbassiddiqui1/spotify-track-recommender
https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/code