Spotify EDA Write-Up

Group members: Jessica Walters and Grace Dalton
Dataset: "Most Streamed Spotify Songs 2023"
Obtained from Kaggle (link)
Our **research questions** addressing the main topics discussed in class are:

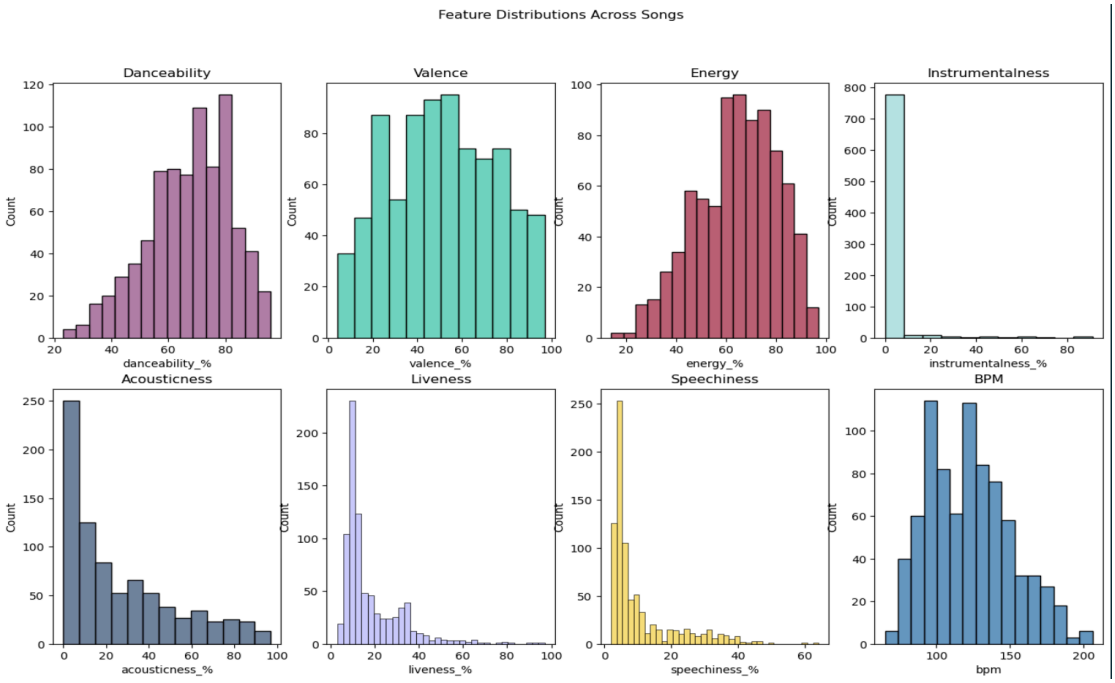
1. Can we predict how many streams on Spotify a song has based on its patterns of audio features (example variables: key, mode, danceability, valence, energy)?
2. Can we predict whether a song would be more popular on Spotify or Apple Music, based on track prevalence in playlists on each platform?
3. Are songs clustered meaningfully based on track features/release information?

Dataset Overview

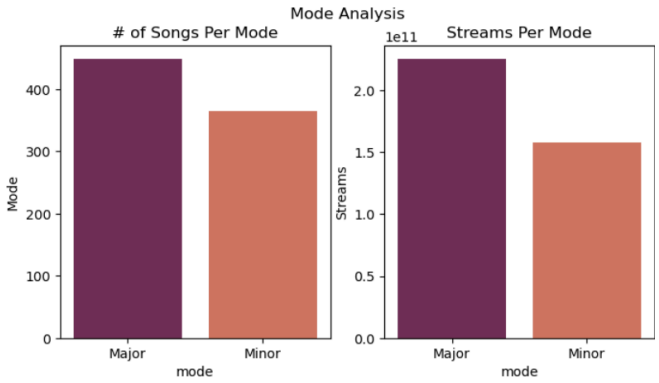
- *track_name:* Name of the song
- *artist(s)_name:* Name of the artist(s) of the song
- *artist_count:* Number of artists contributing to the song
- *released_year:* Year when the song was released
- *released_month:* Month when the song was released
- *released_day:* Day of the month when the song was released
- *in_spotify_playlists:* Number of Spotify playlists the song is included in
- *in_spotify_charts:* Presence and rank of the song on Spotify charts
- *streams:* Total number of streams on Spotify
- *in_apple_playlists:* Number of Apple Music playlists the song is included in
- *in_apple_charts:* Presence and rank of the song on Apple Music charts
- *in_deezer_playlists:* Number of Deezer playlists the song is included in
- *in_deezer_charts:* Presence and rank of the song on Deezer charts
- *in_shazam_charts:* Presence and rank of the song on Shazam charts
- *bpm:* Beats per minute, a measure of song tempo
- *key:* Key of the song
- *mode:* Mode of the song (major or minor)
- *danceability_%:* Percentage indicating how suitable the song is for dancing
- *valence_%:* Positivity of the song's musical content
- *energy_%:* Perceived energy level of the song
- *acousticness_%:* Amount of acoustic sound in the song
- *instrumentalness_%:* Amount of instrumental content in the song
- *liveness_%:* Presence of live performance elements
- *speechiness_%:* Amount of spoken words in the song

Exploratory Data Analysis

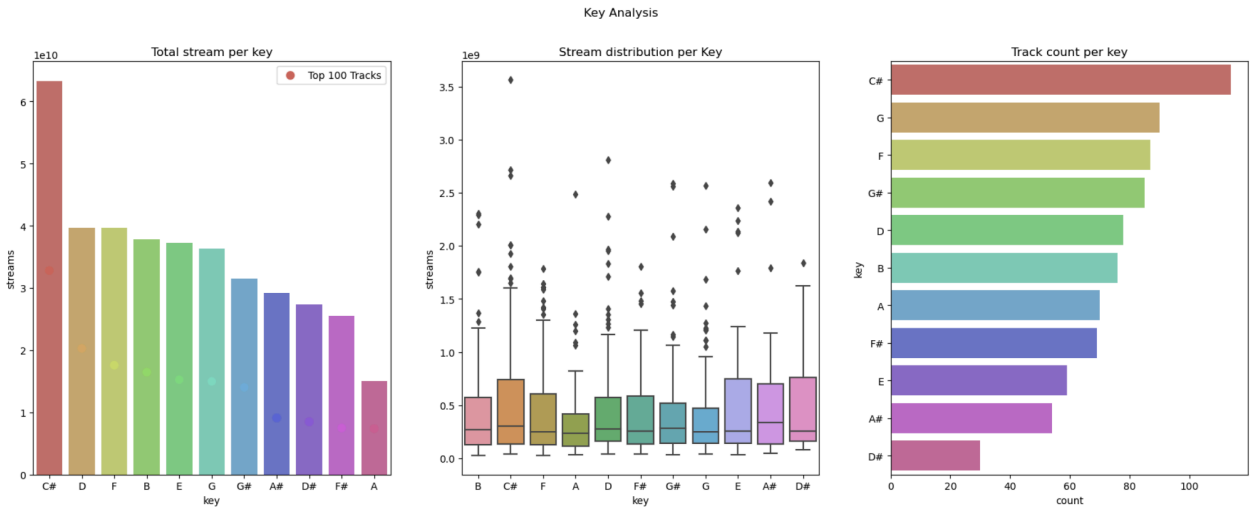
We graphed histograms with the count of songs in the data set against each continuous song feature. These graphs showed mostly normal distributions, with the exception of instrumentalness.



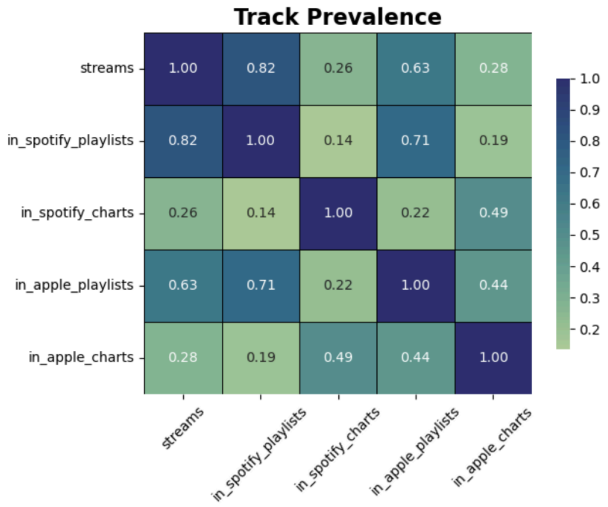
We graphed histograms for the number of songs and streams by mode. Songs in the major key appear to be more frequent and have more overall streams.

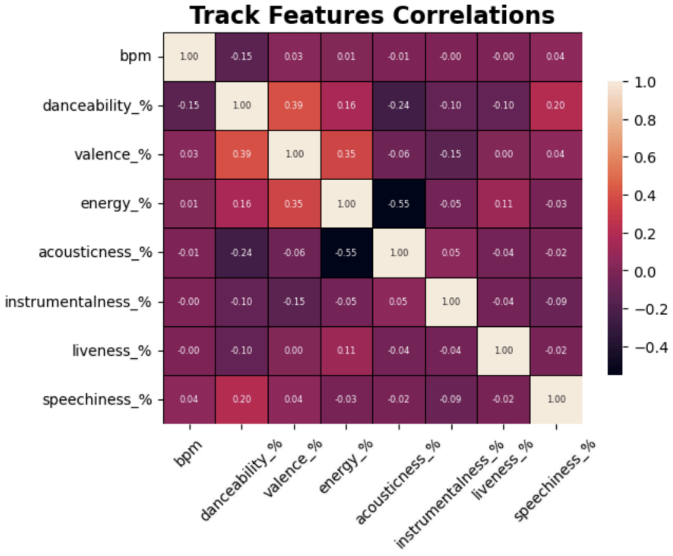


We graphed key against the number of songs, total streams, and the distribution of streams. Songs in key C# are most frequent and have the most streams.



We made correlation plots of variables that indicate track prevalence, and of track audio features. The number of streams is highly positively correlated with the prevalence in Spotify and Apple playlists, but less so with the prevalence on the platform charts. Most of the track features are not significantly correlated with each other. However there is a negative correlation between energy and acousticness.





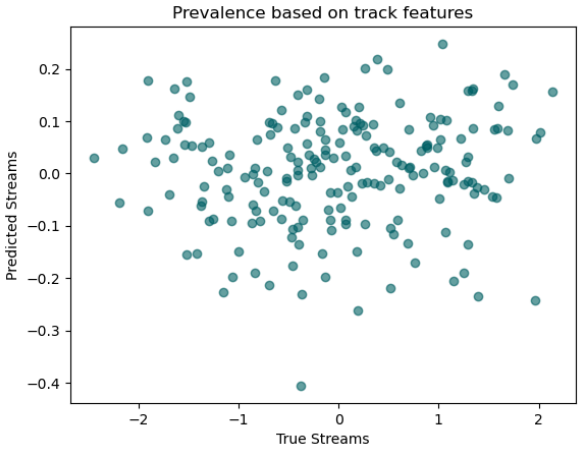
Data Preparation

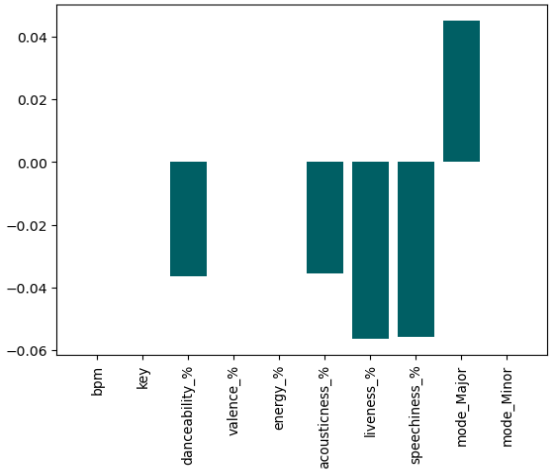
- We dropped some columns from the data for various reasons. Each column and our reason for dropping it is listed below:
    - Track_name, Artist(s)_name: not useful to us as predictors, we indexed the columns for informational purposes
    - Released_day, Released_month: not being used in analysis
    - In_deezer_playlists, In_deezer_charts, In_shazam_charts: not being used in analysis, we are focusing on Apple and Spotify data
    - instrumentalness_%: very similar and correlated to the acousticness variable, keeping both is not useful in making predictions. The distribution is also highly skewed, as most songs in the data set are very low in instrumentalness (as seen in EDA features graph above)
- We also put track performance variables ‘streams’, ‘in_spotify_playlists’, ‘in_apple_playlists’, ‘in_spotify_charts’, and ‘in_apple_charts’ on a logarithmic scale, because the distribution appears more normal. It is easier to interpret on a log2 scale rather than the natural log. Graphs justifying these log2 changes can be found in our Jupyter Notebook.
- We changed the representation of the categorical variable ‘key’ to be numerical through categorical encoding. This enables us to graph ‘key’ as a single continuous variable rather than several dummy variables.
- We are converting the categorical variable ‘mode’ to be numerical through pd.get_dummies(). This enables us to graph ‘mode’ as two dummy variables, so we will know if having a Major or Minor mode will affect the song’s popularity.

- We created a new column with values of the difference between track prevalence in Apple versus Spotify playlists. These values will be used to answer our second research question.
- We standardized the data (using StandardScaler).

Question 1

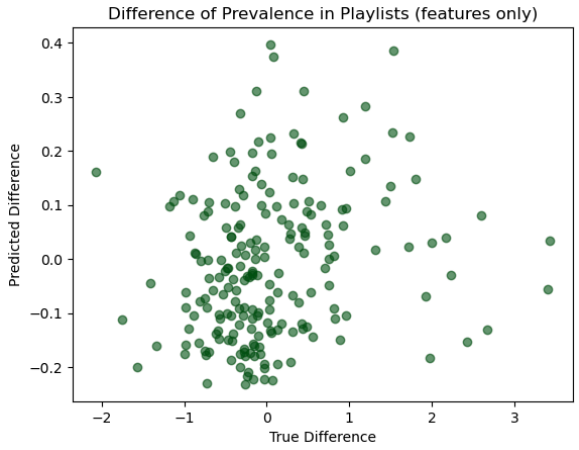
To answer this question, we used multiple linear regression and lasso analysis strategies. From the audio features (BPM, key, mode, danceability, valence, energy, acousticness, liveness, and speechiness) we determined the features that have the most weight in predicting a song's number of streams are mode, danceability, acousticness, liveness, and speechiness. If a song is more danceable, acoustic, live, or speechy, it is likely to have less streams/will be less popular. If the song has a major mode, it is more likely to have more streams/be more popular. The graphs below show how our models turned out.

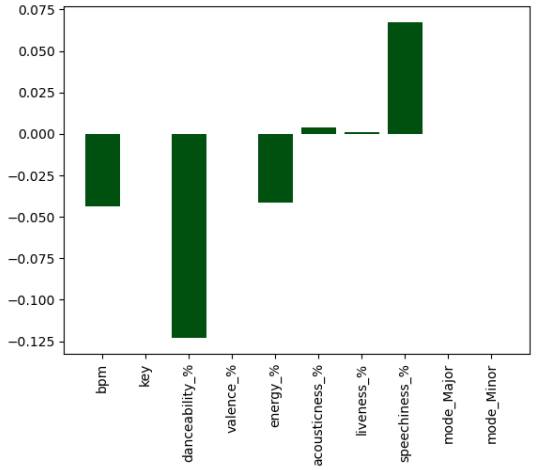


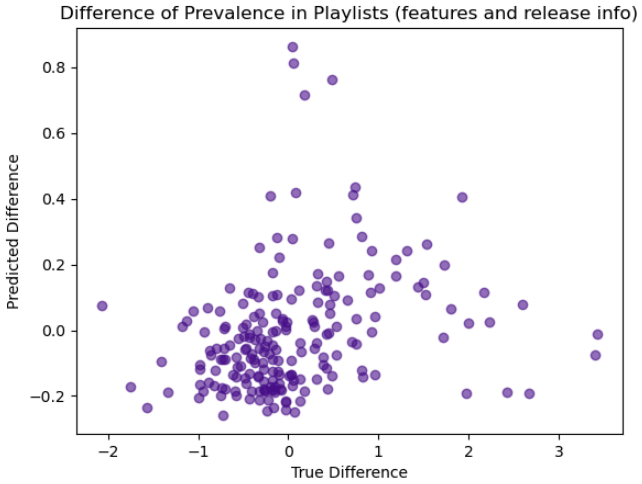


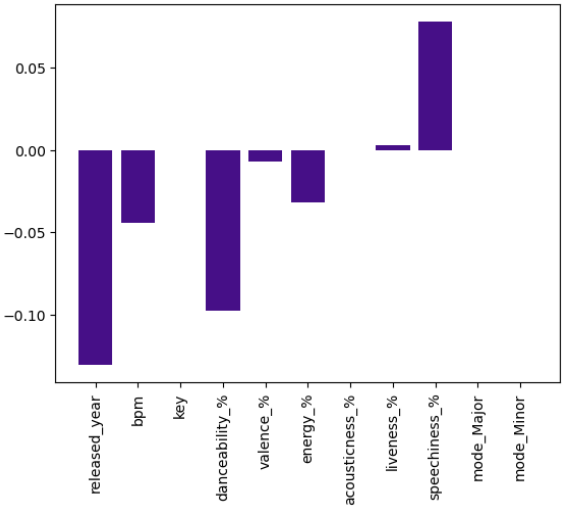
Question 2

To answer this question, we used multiple linear regression and lasso strategies again. In order to predict whether a song would be more popular on one platform over the other, we used the value of the difference between the number of playlists a song is in on either platform. We are predicting the difference in the number of playlists a song is on between the platforms, and this value being more positive or more negative shows that it is more popular on one platform over the other. The graphs in green show our model predictions including only audio features; we found that bpm, danceability, energy, acousticness, liveness, and speechiness predicted whether a song will be more popular on one streaming service over another. However, we did find that adding the release year (graphs shown in purple) added a lot of weight to that variable in making predictions. The weights also shifted so that release year, bpm, danceability, valence, energy, liveness, and speechiness became predictors of popularity on one streaming site. Furthermore, from the model that includes release year, it is clear that older songs are much less likely to be popular on either platform.



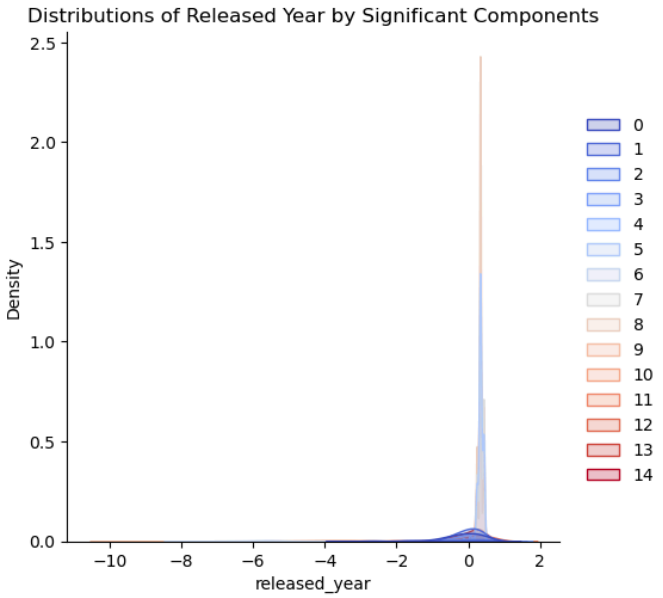


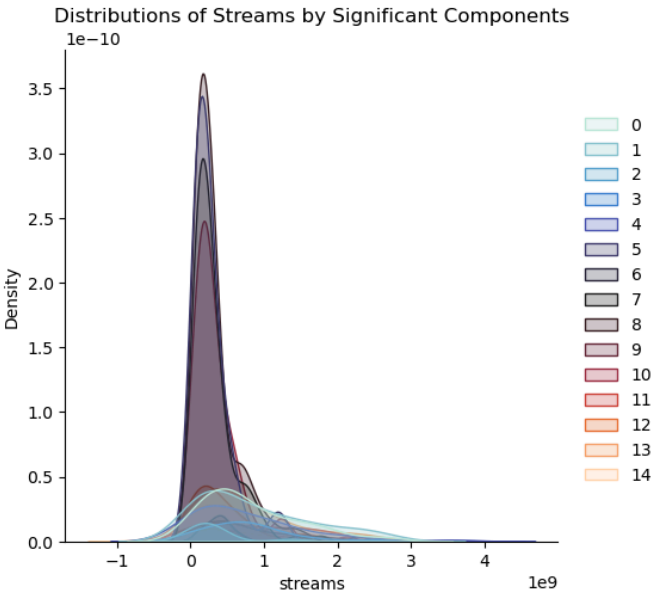




Question 3

To answer this question, we used clustering and Gaussian mixture models. We will use a model that includes both track features and release year, since our findings from Question 2 identified release_year as a significant predictor of popularity on a streaming platform. We used AIC to determine the optimal number of components to use in our model, and then graphed the distributions of the components. Based on these distribution graphs, clusters 2, 6, and 9 most likely come from different distributions. In our notebook, the means of features for these clusters are significantly different. Most of the other clusters are relatively centered around zero together in the two distribution plots below (next page).





Summary

Overall, we explored the impact of certain song features on a song's popularity, used song features and release years to predict whether a song would be more popular on one platform over another, and found that there were few clusters of songs that came from different distributions.