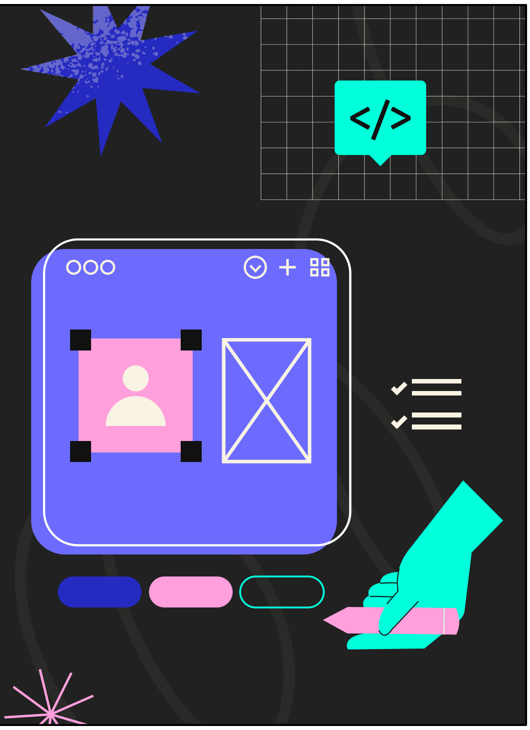
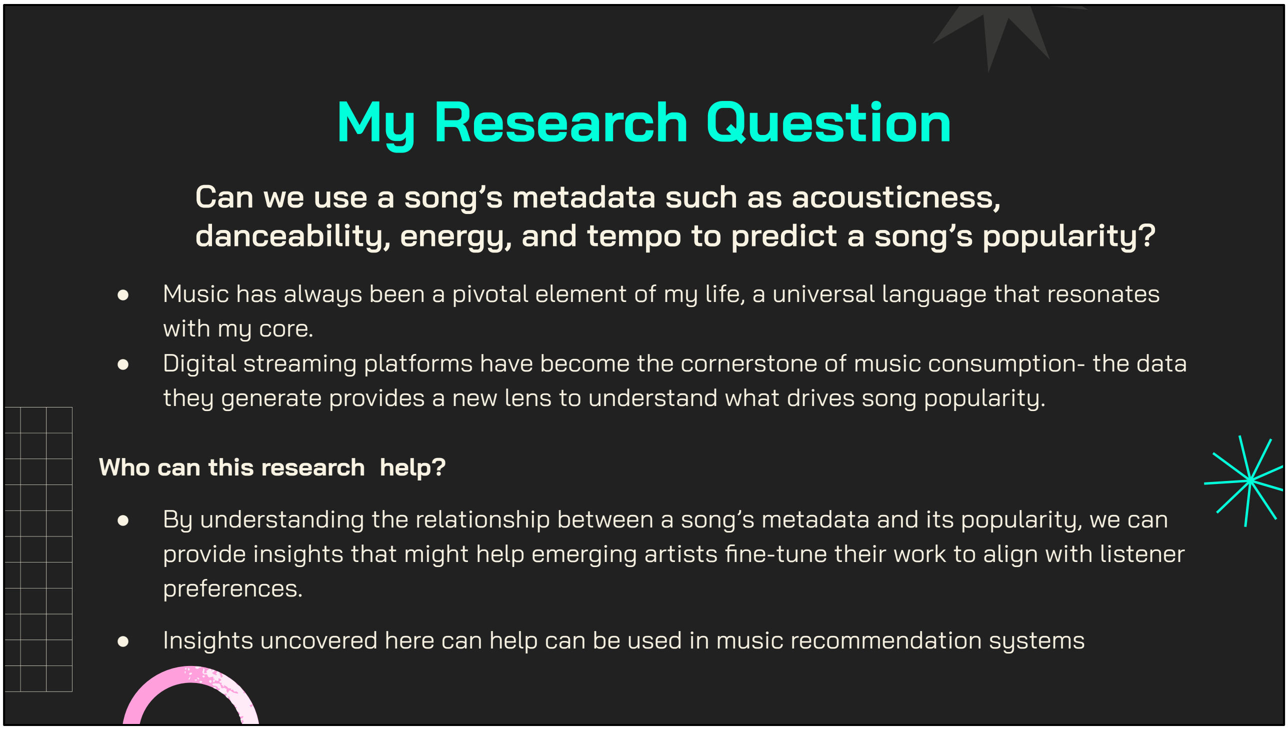
Understanding the Factors that Drive a Song's Popularity

STA303 Project Part 1

Khushil Nagda





My Research Question

Can we use a song's metadata such as acousticness, danceability, energy, and tempo to predict a song's popularity?

- Music has always been a pivotal element of my life, a universal language that resonates with my core.
- Digital streaming platforms have become the cornerstone of music consumption- the data they generate provides a new lens to understand what drives song popularity.

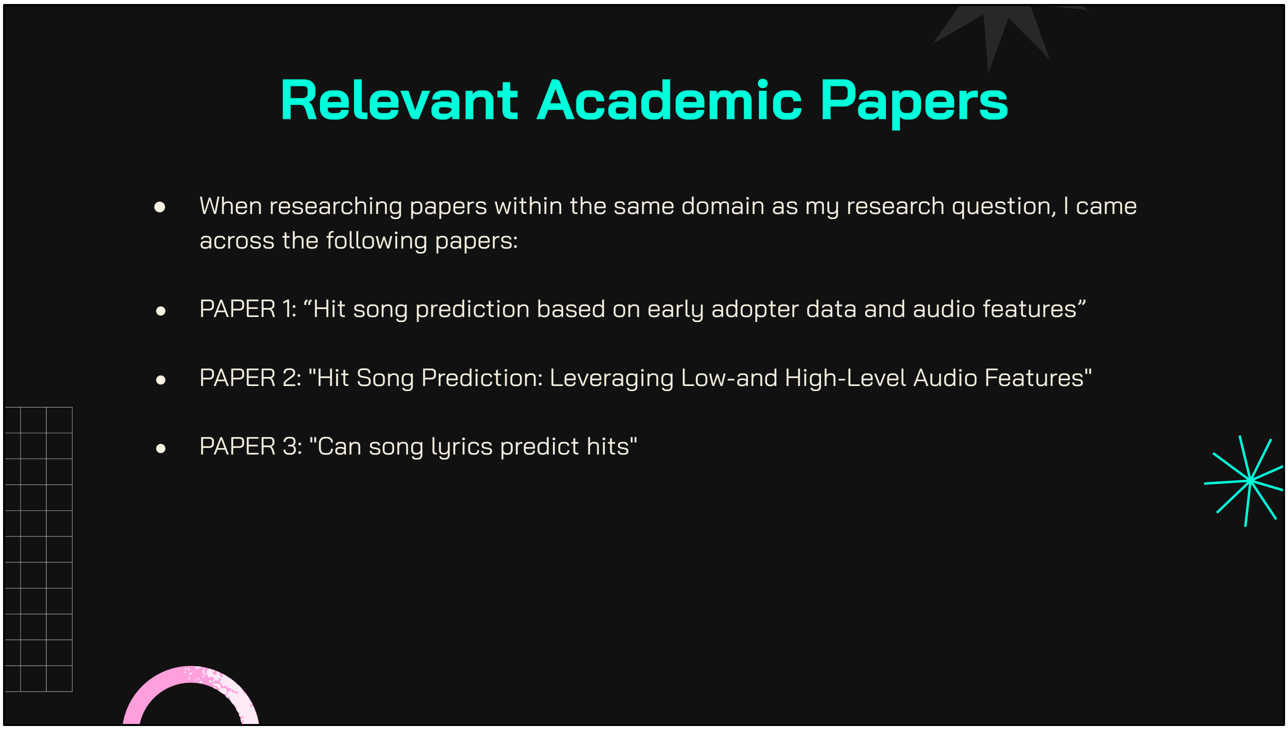
Who can this research help?

- By understanding the relationship between a song's metadata and its popularity, we can provide insights that might help emerging artists fine-tune their work to align with listener preferences.
- Insights uncovered here can help can be used in music recommendation systems

Music has always been a pivotal element of my life, a universal language that resonates with my core. It's not just the melodies or the lyrics that captivate me, but the intricate layers of sound and rhythm that create an immersive experience. This fascination with music has naturally extended into a keen interest in how technology intersects with musical expression.

The implications of this research extend beyond academic interest and touch the core of the music industry. Artists, producers, and record labels are constantly searching for the secret formula that could make a song a hit. By understanding the relationship between a song's metadata and its popularity, we can provide insights that might help emerging artists fine-tune their work to align with listener preferences.

For tech companies and platforms that rely on algorithms to recommend music to listeners, this study could refine the sophistication of predictive models used to curate personalized playlists, thus enhancing user experience.



Relevant Academic Papers

- When researching papers within the same domain as my research question, I came across the following papers:
- PAPER 1: “Hit song prediction based on early adopter data and audio features”
- PAPER 2: "Hit Song Prediction: Leveraging Low-and High-Level Audio Features"
- PAPER 3: "Can song lyrics predict hits"

PAPER 1: The study presents a novel approach to predicting hit songs by integrating audio features with data on early adopter listening behaviors from social media platforms like Last.FM. Their models particularly highlight the potential of logistic regression in distinguishing hits from non-hits, achieving a noteworthy Area Under the Curve (AUC) value of 0.79 when utilizing early adopter behavior, which suggests that social data can significantly bolster prediction models.
PAPER 1 similarity: Includes some of the variables of interest that I use in my research question such as Acousticness, Tempo, Danceability
PAPER 1 difference: The concept of 'early adopters' introduced by the authors proposes that early listener data might predict a song's success trajectory.

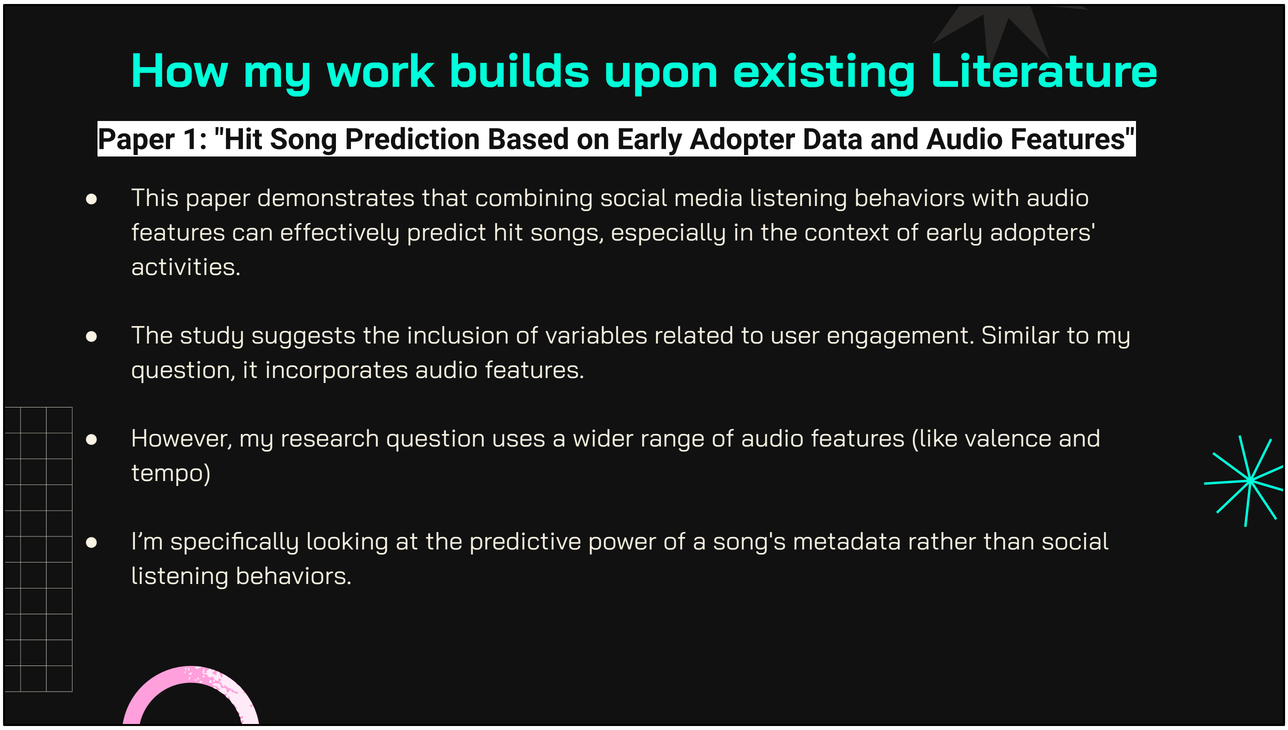
PAPER 2: Yap Kah Yee and Mafas Raheem's study combines Spotify audio and YouTube social features to predict song popularity, finding social media data significantly enhances prediction accuracy. Similar to MY question, it examines audio features' predictive power. Unlike mine, it incorporates social media metrics, demonstrating their substantial impact on forecasting song popularity

They trained four machine learning models in two stages: one with only audio features and another including both audio and social media features. Their findings revealed that at the second stage, the random forest model outperformed others, demonstrating significant improvements in model performances when social media variables were incorporated. This suggests that YouTube-based social media features substantially enhance the prediction accuracy of music popularity.

The novelty of their approach lies in utilizing YouTube data for hit song prediction, a method not previously explored in the Hit Song Science domain.

PAPER 3: The paper "Can Song Lyrics Predict Hits?" by Abhishek Singhi and Daniel G. Brown investigates the potential of song lyrics to predict hit songs, focusing on lyrics as a component of artistic creation. The study employs 31 rhyme, syllable, and meter features to create Bayesian network and support vector machine filters, aiming to differentiate hits from flops. Hits are defined as songs that made it to the Billboard Year-End Hot 100 singles chart between 2008 and 2013, while flops are non-hit songs that had a chance of being hits.

The research finds that lyrics features, especially complex rhyme and meter, are significantly more useful than audio features in separating hits and flops, suggesting that these properties of lyrics may indicate quality songwriting.



How my work builds upon existing Literature

Paper 1: "Hit Song Prediction Based on Early Adopter Data and Audio Features"

- This paper demonstrates that combining social media listening behaviors with audio features can effectively predict hit songs, especially in the context of early adopters' activities.
- The study suggests the inclusion of variables related to user engagement. Similar to my question, it incorporates audio features.
- However, my research question uses a wider range of audio features (like valence and tempo)
- I'm specifically looking at the predictive power of a song's metadata rather than social listening behaviors.

How my work builds upon existing Literature

Paper 2: "Predicting Music Popularity Using Spotify and YouTube Features"

- This study integrates social media features with Spotify Top 200 chart performance metrics to predict song popularity.
- What's novel about it is that it suggests that social media can provide additional predictive context.
- Leverages machine learning to predict song popularity using audio features.

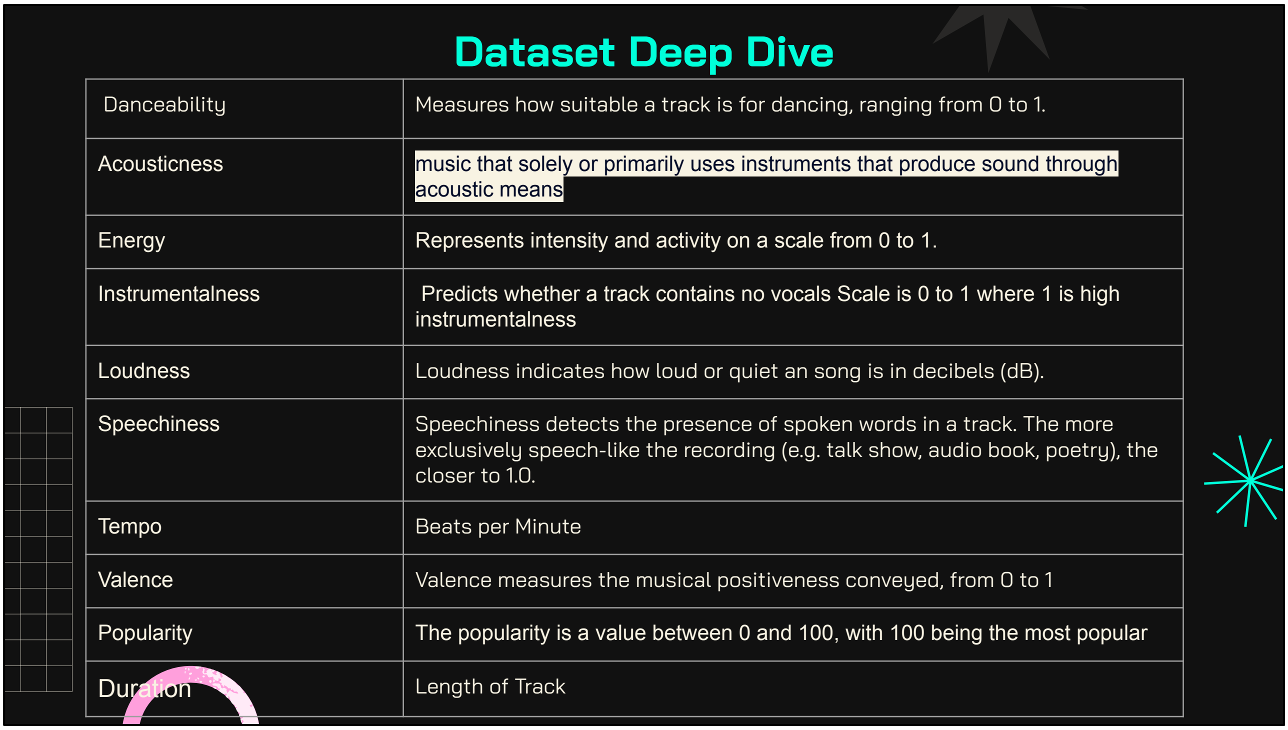
How my work builds upon existing Literature

Paper 3: "Can Song Lyrics Predict Hits?"

- This research focuses on the lyrics of songs and their structural features to predict hit songs, suggesting that textual content can be a strong predictor.

- This broadens the perspective of my research question to consider not just sound but also content of music in predicting popularity.

- Methodologies differ, as I shall use a Generalized Linear Model (GLM), and this paper uses Bayesian networks and SVMs.



Dataset Deep Dive

Danceability	Measures how suitable a track is for dancing, ranging from 0 to 1.
Acousticness	music that solely or primarily uses instruments that produce sound through acoustic means
Energy	Represents intensity and activity on a scale from 0 to 1.
Instrumentalness	Predicts whether a track contains no vocals Scale is 0 to 1 where 1 is high instrumentalness
Loudness	Loudness indicates how loud or quiet an song is in decibels (dB).
Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0.
Tempo	Beats per Minute
Valence	Valence measures the musical positiveness conveyed, from 0 to 1
Popularity	The popularity is a value between 0 and 100, with 100 being the most popular
Duration	Length of Track

Role/Function of Each Variable in the GLM Model:

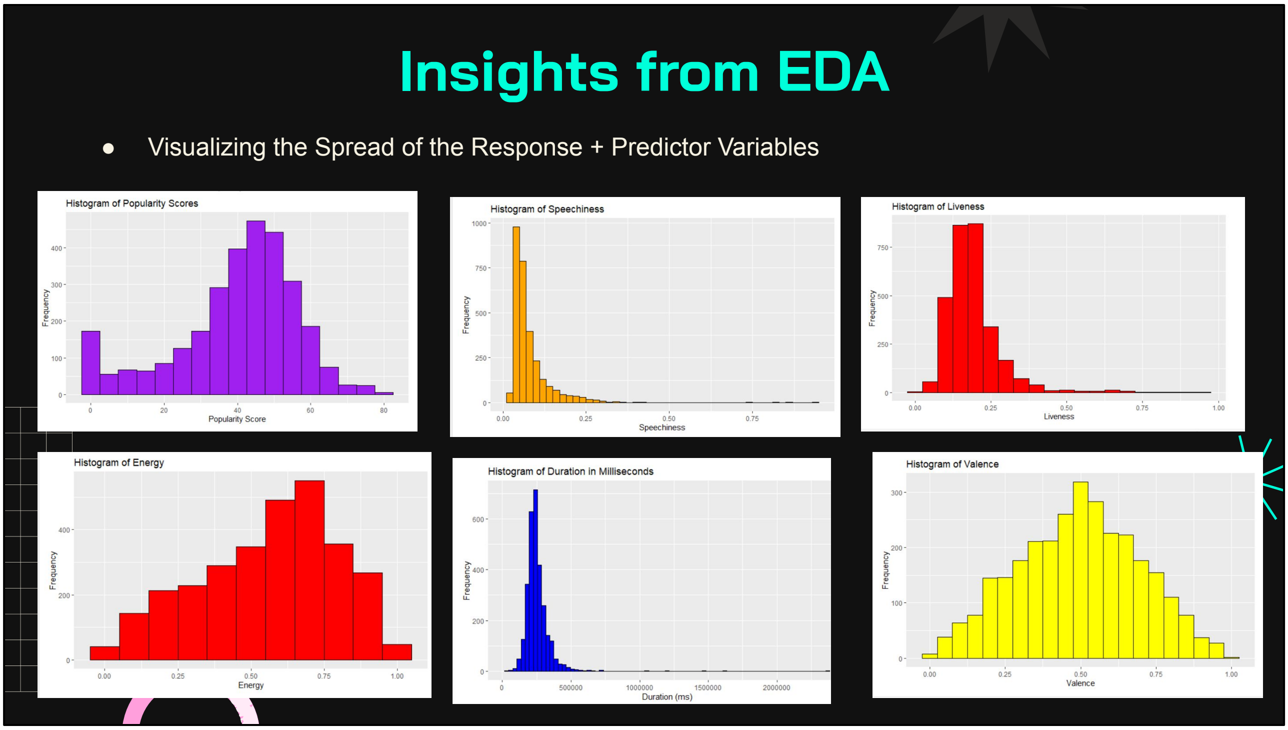
- Acousticness: This measures the level of acoustic sounds in a song. I hypothesize that songs with higher acousticness may appeal to listeners who prefer softer and more traditional sounds, which could affect their popularity in specific demographics.
- Danceability: This indicates how suitable a track is for dancing. A higher danceability score is presumed to correlate with greater popularity, especially among songs targeting dance venues and younger audiences.
- Energy: Represents the intensity and activity within the song. Energetic tracks might resonate more during certain times (e.g., summer hits) or in specific settings (e.g., clubs, workout playlists).
- Tempo: The speed of the song can influence its reception; faster songs may be preferred in energetic settings, while slower tempos might do better in relaxed environments.
- Duration_ms: Longer songs might have a different impact than shorter ones, potentially affecting listener engagement and popularity.
- Instrumentalness: The presence of instrumental sections might appeal to certain listener groups and can be significant in instrumental or

- classical genres.
- Liveness: Indicates the presence of a live audience in the recording. Live versions might either enhance the appeal due to their raw energy or diminish it if the audio quality is compromised.
- Loudness: Loud tracks may be perceived as more powerful, yet potentially less pleasant at high volumes, influencing their popularity.
- Speechiness: The presence of spoken words can make a song distinctive, affecting its memorability and potential popularity.
- Valence: Measures musical positiveness. Tracks with higher valence might be preferred by listeners looking for upbeat and happy songs.

Column Description

- **track_id**: The Spotify ID for the track
- **artists**: The artists' names who performed the track. If there is more than one artist, they are separated by a `;`
- **album_name**: The album name in which the track appears
- **track_name**: Name of the track
- **popularity**: **The popularity of a track is a value between 0 and 100, with 100 being the most popular**. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity.
- **duration_ms**: The track length in milliseconds
- **explicit**: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
- **danceability**: 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 least danceable and 1.0 is most danceable
- **energy**: Energy is 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. For example, death metal has high energy, while a Bach prelude scores low on the scale
- **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

- **loudness**: The overall loudness of a track in decibels (dB)
- **mode**: 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
- **speechiness**: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), 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 both 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. 1.0 represents high confidence 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 likelihood 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 strong likelihood that the track is live
- **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, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)
- **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
- **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`.
- **track_genre**: The genre in which the track belongs

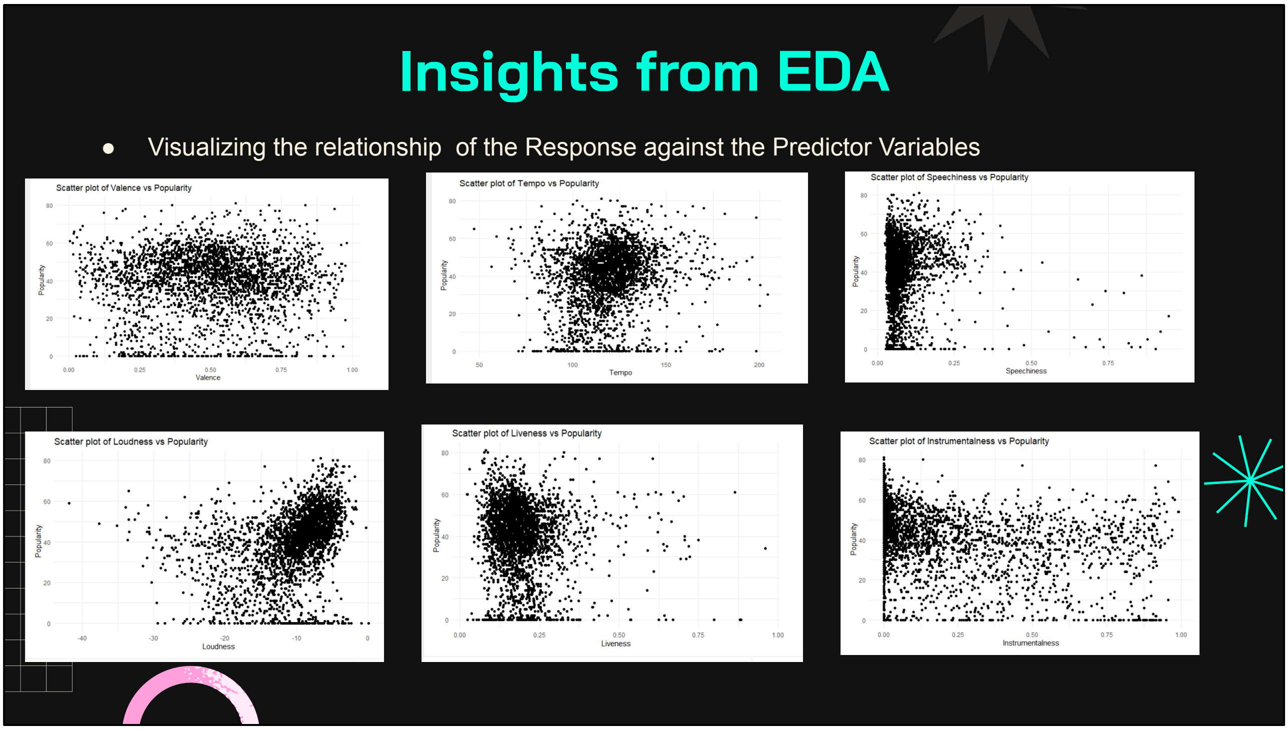


Insights from EDA

- Visualizing the Spread of the Response + Predictor Variables

1. Tempo: The histogram for Tempo shows a bell-shaped distribution, which indicates that most songs have a tempo around a central value, with fewer songs having very slow or very fast tempos.
2. Speechiness: The histogram for Speechiness is heavily skewed to the right, indicating that most songs have low speechiness values with a few exceptions having higher values.
3. Loudness: The histogram for Loudness shows a somewhat normal distribution but with a slight left skew.
4. Liveness: The Liveness histogram is heavily skewed to the right, which implies that most songs have low liveness, and very few have high

1. liveness. This skew could indicate that live recordings are rare in the dataset.
2. Instrumentalness: The histogram for Instrumentalness shows that a vast majority of songs have low instrumentalness values. This suggests that non-instrumental tracks dominate my dataset.
3. Energy: The histogram for Energy appears multimodal with several peaks. This suggests that there are sub-groups within the dataset that have distinctly different levels of energy.
4. Duration in Milliseconds: The histogram for Duration shows a right-skewed distribution, suggesting that there's a common range for song length. Outliers in duration could disproportionately influence the results of a regression analysis.
5. Danceability: The histogram for Danceability is roughly bell-shaped, indicating its normally distributed in my dataset.
6. Popularity Scores: The histogram for Popularity Scores is kind of normal but shows a slight left skew. The left skew suggests that there are more songs with lower popularity scores than high.



Insights from EDA

- Visualizing the relationship of the Response against the Predictor Variables

In general, these scatter plots suggest that while some characteristics may have a threshold below which popularity is unlikely (e.g., very quiet songs), there are no strong linear relationships evident between these individual variables and popularity. This suggests that the relationship between song characteristics and popularity is complex and likely non-linear or influenced by multiple interacting factors.

Valence vs Popularity: The distribution of points does not suggest a clear linear relationship between valence and popularity. The points are widely spread.

Tempo vs Popularity: The points are clustered around the middle range of tempo, with fewer songs having very low or very high tempo. This could indicate that songs with moderate tempo are more common or more likely to be popular.

Speechiness vs Popularity: There is a concentration of points towards the lower end of speechiness, which means that most songs have low speechiness, and this may not be a strong differentiator of popularity.

Loudness vs Popularity: The scatter plot shows that louder songs tend to have a broader range of popularity scores, which might imply that louder songs are more common in popular music.

Liveness vs Popularity: This plot shows that live recordings are not predominantly popular or unpopular, as they are spread across the range of popularity. This indicates that liveness might not be a significant predictor of popularity.

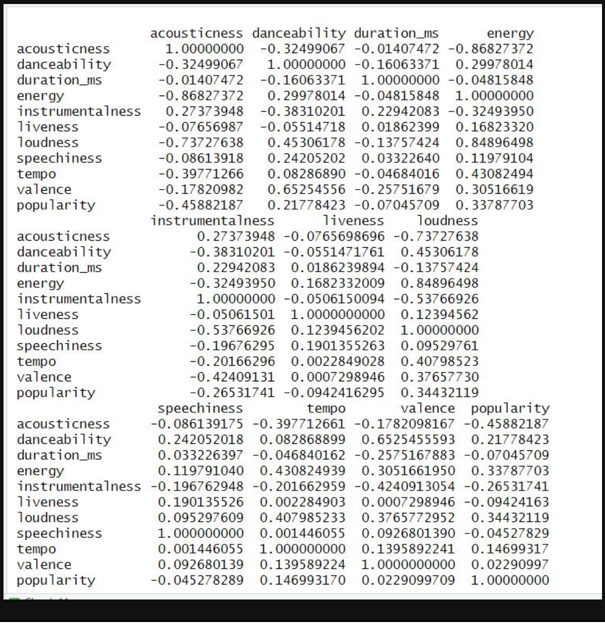
Instrumentalness vs Popularity: Most songs have low instrumentalness, with few purely instrumental tracks. This distribution suggests that purely instrumental tracks do not generally achieve high popularity scores.

Energy vs Popularity: The scatter plot is quite dense with no clear pattern, indicating that a song's energy level does not have a simple linear relationship with its popularity.

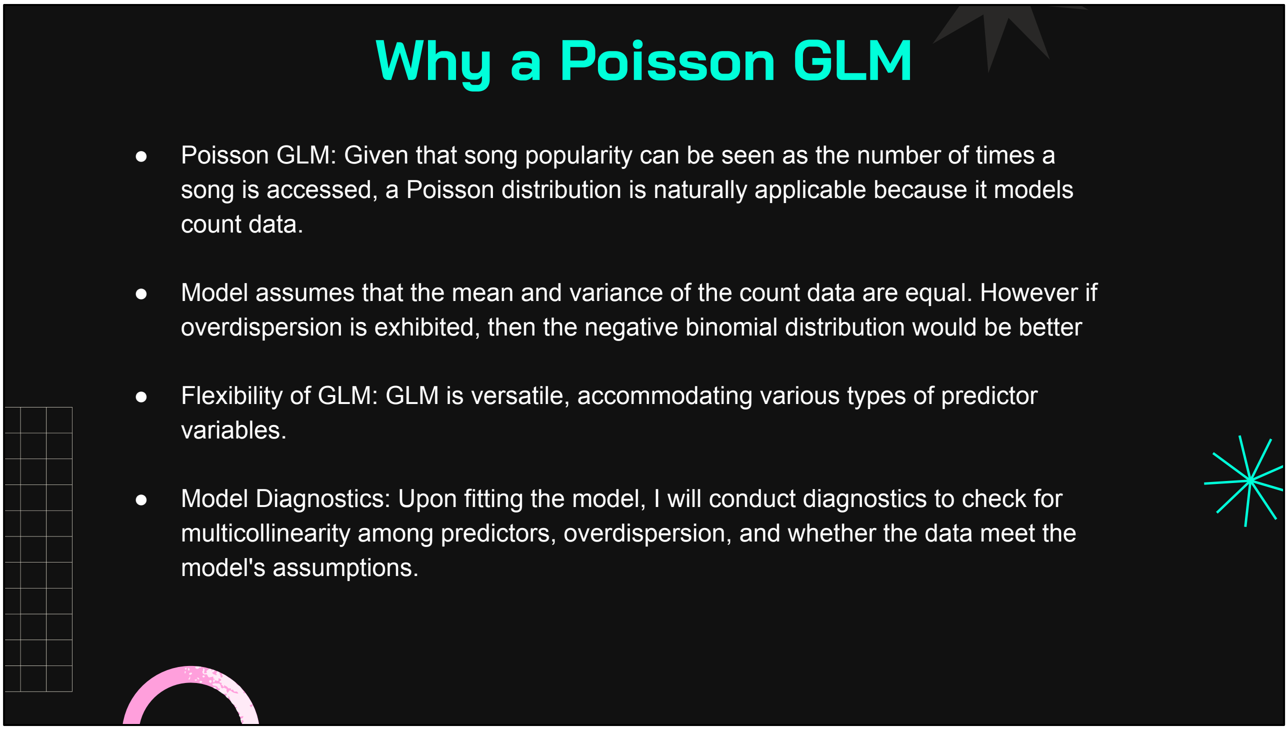
Duration vs Popularity: The plot is skewed towards the lower end of duration, indicating that most songs have a duration within a certain range, with very long songs being rare.

Danceability vs Popularity: The distribution of points does not indicate a strong relationship between danceability and popularity. While there's a dense cluster in the mid-range of danceability, there is no clear trend that higher danceability leads to higher popularity.

Checking for Multi-Collinearity



```
                 acousticness danceability duration_ms      energy
acousticness       1.00000000  -0.32499067 -0.01407472 -0.86827372
danceability      -0.32499067   1.00000000 -0.16063371  0.29978014
duration_ms       -0.01407472  -0.16063371  1.00000000 -0.04815848
energy            -0.86827372   0.29978014 -0.04815848  1.00000000
instrumentalness   0.27373948  -0.38310201  0.22942083 -0.32493950
liveness          -0.07656987  -0.05514718  0.01862399  0.16823320
loudness          -0.73727638   0.45306178 -0.13757424  0.84896498
speechiness       -0.08613918   0.24205202  0.03322640  0.11979104
tempo             -0.39771266   0.08286890 -0.04684016  0.43082494
valence           -0.17820982   0.65254556 -0.25751679  0.30516619
popularity        -0.45882187   0.21778423 -0.07045709  0.33787703
                 instrumentalness      liveness    loudness
acousticness           0.27373948 -0.0765698696 -0.73727638
danceability          -0.38310201 -0.0551471761  0.45306178
duration_ms            0.22942083  0.0186239894 -0.13757424
energy                -0.32493950  0.1682332009  0.84896498
instrumentalness       1.00000000 -0.0506150094 -0.53766926
liveness              -0.05061501  1.0000000000  0.12394562
loudness              -0.53766926  0.1239456202  1.00000000
speechiness           -0.19676295  0.1901355263  0.09529761
tempo                 -0.20166296  0.0022849028  0.40798523
valence               -0.42409131  0.0007298946  0.37657730
popularity            -0.26531741 -0.0942416295  0.34432119
                  speechiness        tempo       valence  popularity
acousticness     -0.086139175 -0.397712661 -0.1782098167 -0.45882187
danceability      0.242052018  0.082868899  0.6525455593  0.21778423
duration_ms       0.033226397 -0.046840162 -0.2575167883 -0.07045709
energy            0.119791040  0.430824939  0.3051661950  0.33787703
instrumentalness -0.196762948 -0.201662959 -0.4240913054 -0.26531741
liveness          0.190135526  0.002284903  0.0007298946 -0.09424163
loudness          0.095297609  0.407985233  0.3765772952  0.34432119
speechiness       1.000000000  0.001446055  0.0926801390 -0.04527829
tempo             0.001446055  1.000000000  0.1395892241  0.14699317
valence           0.092680139  0.139589224  1.0000000000  0.02290997
popularity       -0.045278289  0.146993170  0.0229099709  1.00000000
```



Why a Poisson GLM

- Poisson GLM: Given that song popularity can be seen as the number of times a song is accessed, a Poisson distribution is naturally applicable because it models count data.
- Model assumes that the mean and variance of the count data are equal. However if overdispersion is exhibited, then the negative binomial distribution would be better
- Flexibility of GLM: GLM is versatile, accommodating various types of predictor variables.
- Model Diagnostics: Upon fitting the model, I will conduct diagnostics to check for multicollinearity among predictors, overdispersion, and whether the data meet the model's assumptions.

Negative Binomial GLM: If the data exhibit overdispersion (variance exceeds the mean), the negative binomial distribution provides a better fit by introducing an extra parameter to account for this variability, ensuring more accurate and reliable model estimates.

Flexibility of GLM: It allows for linear and non-linear relationships between predictors and the response variable via link functions like the log link in the Poisson and Negative Binomial models

Predictive Variables: The predictors (acousticness, danceability, energy, and tempo) are continuous variables, and GLM allows for the inclusion of both continuous and categorical predictors, offering flexibility in model design.

Are the Assumptions Satisfied?

- **Linearity of predictors**: GLM does not assume that the relationship between the predictors and the log-odds of the outcome is linear, The scatter plots do not suggest strong linear relationships for any single predictor, which is appropriate for GLMs.
- **Correct distribution of the response variable**
- **Absence of multicollinearity**
- **No influential outliers**
- **Homoscedasticity of the error variance**

References

Yee, Y. K., & Raheem, M. (2022, September 21). Predicting music popularity using Spotify and YouTube features. SRS Journal. https://indjst.org/articles/predicting-music-popularity-using-spotify-and-youtube-features

Can song lyrics predict hits? (n.d.). https://cs.uwaterloo.ca/~browndg/CMMR15data/CMMR2015paper.pdf

Herremans, D., & Bergmans, T. (2020, October 16). Hit song prediction based on early adopter data and audio features. arXiv.org. https://arxiv.org/abs/2010.09489

https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/