A Bayesian Analysis on the Attribution of Success in Music

Nawat Swatthong, Eliza Chew, Pat Chimtanoo

December 14, 2022

1 Introduction

Music charts are typically used as an indicator to measure an artist's success. However, the question that might come up is whether a particular song became popular because of its inherent qualities or because the song was sung by an artist who is popular. The team is interested in exploring how success of a music should be attributed and studying whether the popularity of a song would change if it was sung by different artists.

Although analyzing the songs' popularity based on artist is feasible, the available data are sparse and unreliable as some artists might have only a couple songs presented in the chart. To obtain more data, the team decide to simplify research question with similar assumption that some genres are more popular than the others. So, the study purpose remains resemble, where the team is now examine the change in popularity if the songs were written and sung in a different styles.

2 Data Characteristics

2.1 Data Description

The team analyzed global average number of streaming of songs included in the "Top 200 Chart" and their audio features. To account for the possibility of seasonality fluctuation, the streaming frequency were collected over 365 days after its released date and divided by the number of times a particular song appears in the chart. Audio features will be averaged if there are duplicated entries since the value are relatively similar, and the team believe that the differences were caused by noise in the collection process. Then, the data were split into 2 parts for Bayesian logistic and linear regression.

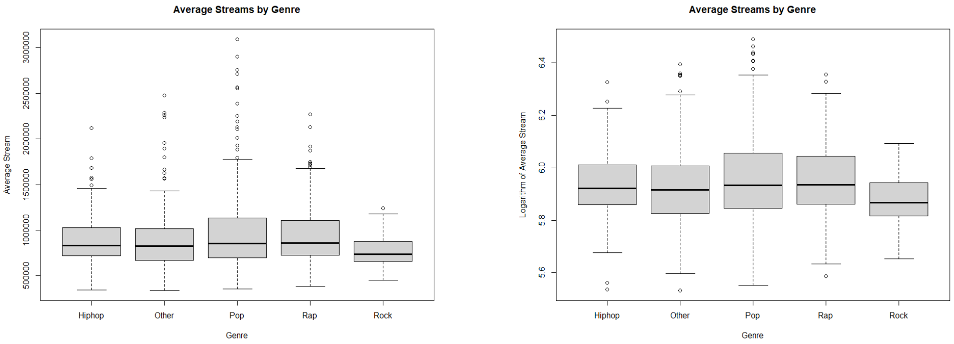
The sources of data are as follows:
1. Spotify Charts, sees the 'Top 200' and 'Viral 50' charts published globally by Spotify since January 1 2017, refreshed every 2-3 days. The charts contain daily information on song's popularity and trend according to its performance in each region.
2. Spotify 1.2M+ Songs, sees further information of the audio features of over 1.2M songs obtained with the Spotify API.
3. Spotify Daily Top 200 Songs with Genres 2017-2021, reports genre of each song that appears in the 'Top 200' chart started in 2017.

2.2 Data Exploratory Analysis

After data collection and processing, the team obtained in total of 9 variables: average global streaming, genre, and 7 audio features associated with a particular song. The density plot and some statistics are listed below.

Number of Streaming

Number of Streaming is an average number of times the songs has been streamed on Spotify over 365 days after its released date. From Figure 1, we can observe that songs with genre Rock has the lowest average number of streaming, and other genres have similar mean of average number of streaming, where Pop has widest range and most outliers.



(a) Number of Streaming by Genre

(b) Log Number of Streaming by Genre

Figure 1: Box Plot for Number of Streaming by Genre

Genres

Variable Genres records the musical style of each songs. Figure 2 shows a bar plot of genre classification. This denotes that out of the top Spotify songs that we have included, a large majority are considered 'Pop' and 'Rap'.

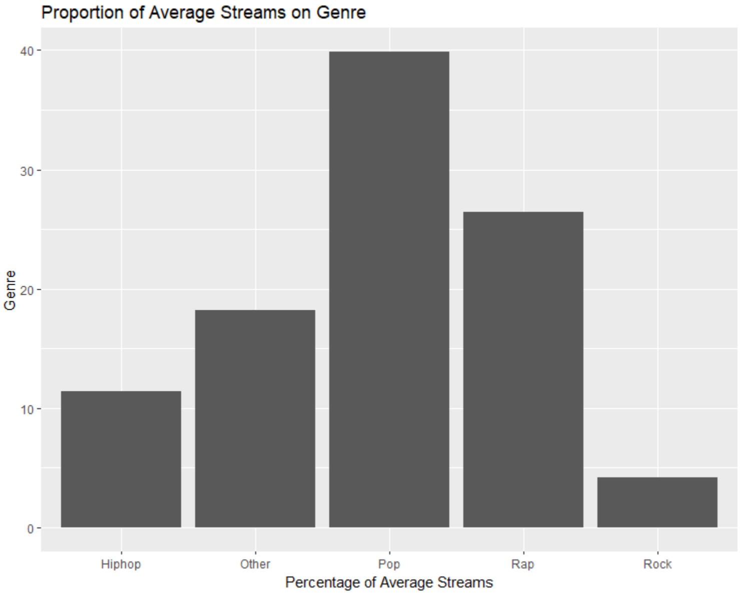


Figure 2: Distribution Plot for Each Audio Feature

audio features

Audio features records musical features associated with each song, which include:

- Danceability - How suitability a track is for dancing
- Energy - How intense and active a track is
- Loudness - Overall loudness of the track in decibels (dB)
- Speechiness - Proportion of spoken words in the track
- Acousticness- Confidence measure of whether a track is acoustic
- Liveness - Detect live audience in track. Represents the probability that a track was performed live
- Valence - Measures how positive a track sounds from 1 (extremely positive) to 0 (extremely negative)
- Tempo - Overall tempo of a track, in beats per minute (BPM)

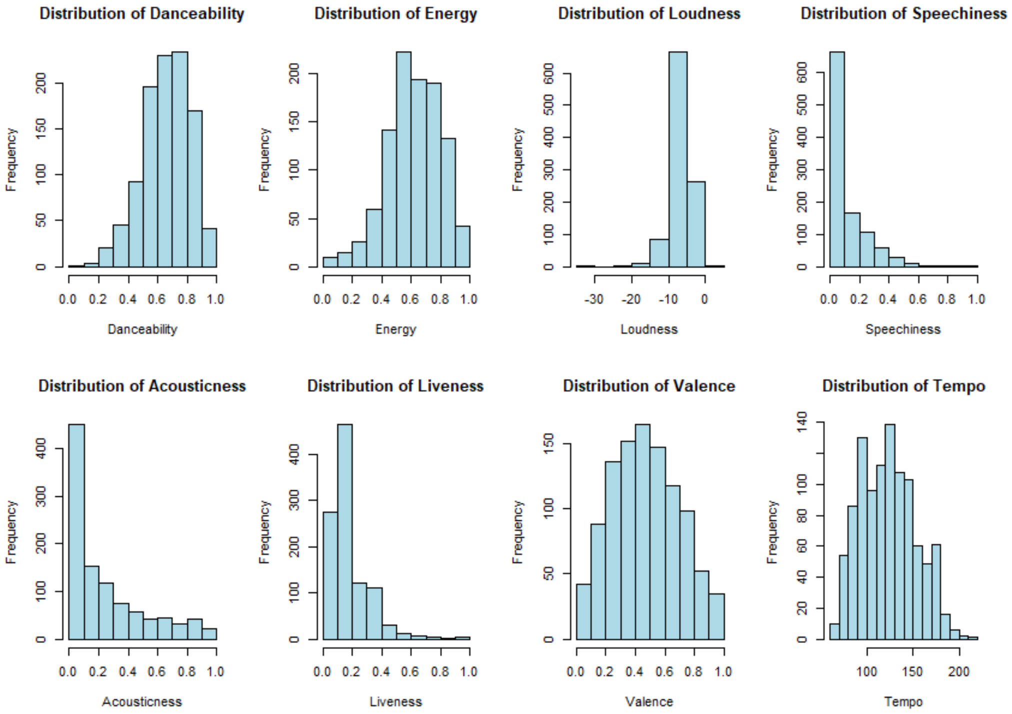


Figure 3: Distribution Plot for Each Audio Feature

Observing the distributions plot of audio features from Figure 3, one can observe that there are particular skews in the data for each variable. For example, Speechiness, Liveness, and Acousticness have left skews; whilst loudness and energy have left skews.

In the model, the team excludes features such as Acousticness, Liveness and Loudness from the analysis because the effect from these variables are not as significant and correlated with other features.

3 Methodology

In this project, the team use Bayesian Hierarchical model to obtain the distribution and statistical inference of the number of streaming for each genre. As the question focuses on analyzing the attribution of the success in music, it might be more reasonable to examine the change in popularity given solely audio features and given audio features and genre separately. Therefore, the model consists of two components reflecting the team's believes on the stated relationships.

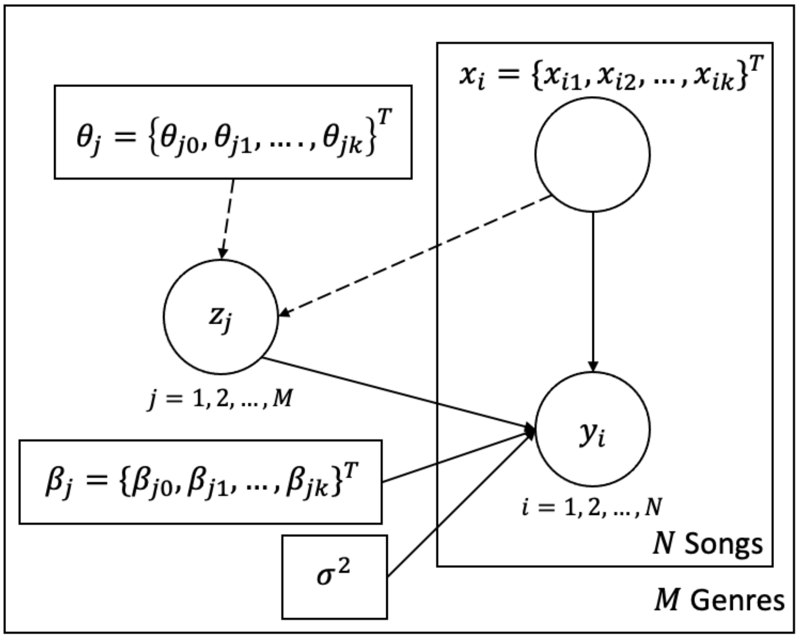

$x_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}^T$
$\theta_j = \{\theta_{j0}, \theta_{j1}, \ldots., \theta_{jk}\}^T$
z_j
$j = 1, 2, \ldots, M$
y_i
$\beta_j = \{\beta_{j0}, \beta_{j1}, \ldots, \beta_{jk}\}^T$
$i = 1, 2, \ldots, N$
N Songs
σ^2
M Genres

Figure 4: Diagram for Hierarchical Model

Figure 4 shows a diagram for hierarchical model where the dash lines represent Bayesian logistic regression and the solid lines represent Bayesian linear regression. The parameters in the model are defined as follows:

x$_i$ = Vector of audio features variables, which are Danceability, Energy, Speechiness, Valence, Tempo, associated for the song $i = 1, 2, \ldots, N$

y$_i$ = The average number of streaming for a song i, associated for the song $i = 1, 2, \ldots, N$

z$_j$ = Categorical variable representing song's genre j, associated for the song $j = 1, 2, \ldots, M$ (In this project, we use 5 categories - Rock, Pop, Hip-hop, Rap, and Others)

θ_j = Logistic regression coefficients for each genre j

β_j = Linear regression coefficients for each genre j

σ^2 = Variance of the average number of streaming for all songs

Bayesian Logistic Regression

The Bayesian logistic regression takes into account the association between audio features and genre. The model takes audio features as predictors and their prior beliefs, then produces the probability that a song would be classified as a genre given the observed audio features.

$$p(z_j|x_i) = \frac{1}{1 + exp\{-\theta_j^T x_i\}} \text{ for i = 1, 2, ..., N, j = 1, 2, ..., M}$$

According to the equation above, $p(z_j|x_i)$ is the probability of song i being in genre j given audio feature data of that song. Since the knowledge on the distribution of θ is very limited. The prior beliefs of how each features affect genre classification are chosen to follow a normal distribution with $\mu_\theta = 0$ and σ_θ^2= 10.

$$\theta_j \sim Normal(0, 10) \text{ for j = 1, 2, ..., M}$$

Here, the team perform 5 independent Bayesian logistic regression to obtain the probability of each genre. The posterior of θ are obtained using MCMC from python library PyMC3. The team, then, proceed to the next step with variable Z to be the genre with the highest probability p_i.

Bayesian Linear Regression

Bayesian linear regression utilizes the information on genre and audio features to predict song's popularity. The inputs of linear regression are audio features excluding genre, but the model will incorporate genre classification knowledge from the previous model by choosing linear regression coefficient $\beta_{j=k}$. The Bayesian linear regression is defined as

$$y_i | x_i, z_k \sim Normal(\beta_k^T x_i, \sigma^2) \text{ for i = 1, 2, ..., N}$$

We specify the prior distribution of each β_k to be Normal distribution and σ^2 to be Half Normal distribution.

$$\beta_k \sim Normal(1000, 300)$$

$$\sigma^2 \sim HalfNorm(\sigma_0^2)$$

The prior beta corresponds to the the mean of the number of streaming divided by the number of audio features and mean of each audio feature values. Subsequently, we set prior beta at $\mu_\beta = 1000, \sigma_\beta^2 = 300$. And based on the fact that the mean the number of view from several other observed is mostly around 1e5-5e5, we select the prior of of sigma $\sigma_0^2 = 1e5$. Then, the posterior distribution for each genre is again obtained by MCMC, using Python library called PyMC3. The β_j are sampled from these posterior and used to estimate popularity of a song.

4 Result and Discussion

For the first part of the proposed model, we derived the posterior distribution of logistic regression coefficients for each genre.

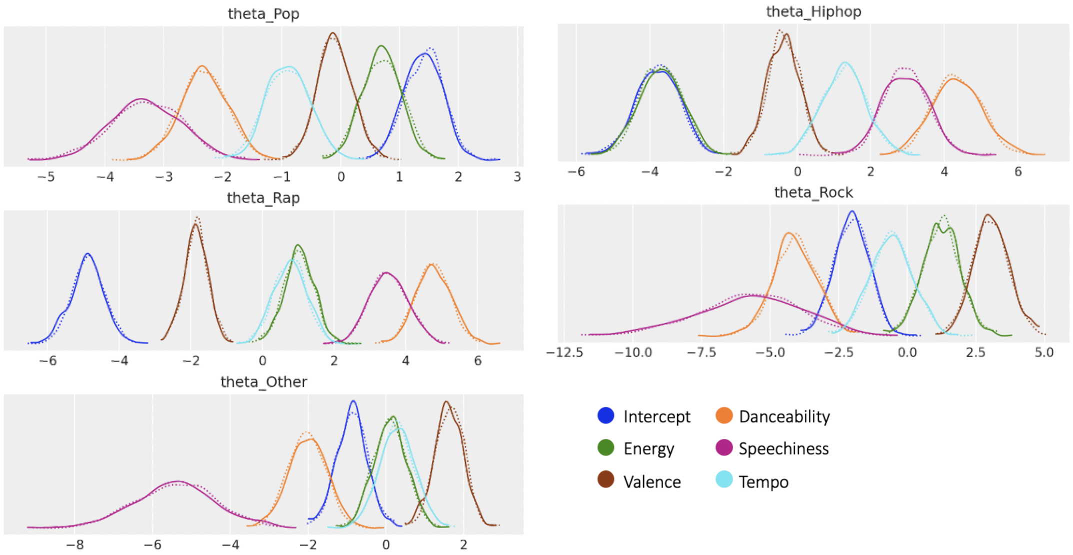


Figure 5: Posterior Distribution for Logistic Regression Coefficients

As seen in Figure 5, although feature such as speechiness may have a huge dispersion, we can see that there are specific patterns of how each audio feature presented in each genre. This allows the model to classify the input (songs without genre label) into groups and obtain the probability of being in a particular genre.

The results from Bayesian logistic regression will be the black box model that helps obtaining the genre while implementing linear regression with only audio features by classifying the song to a genre with maximum probability. Following the Bayesian linear regression model, we obtain the posterior distribution of β_j for each genre.

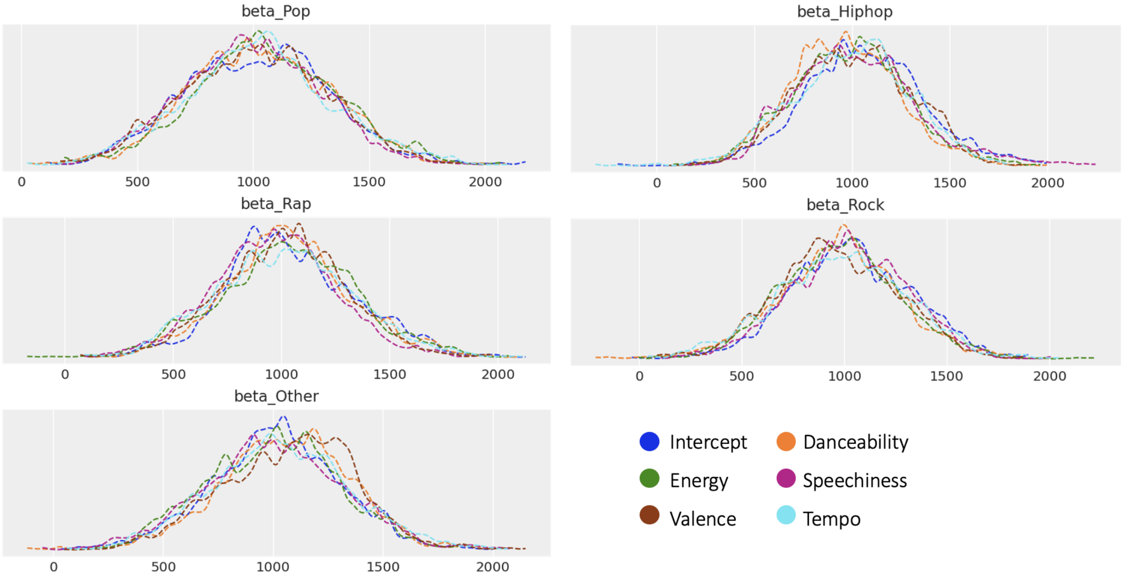


Figure 6: Posterior Distribution for Linear Regression Coefficients

Unfortunately, all of the posterior distributions deriving MCMC were condensed at prior. This difficulty might be due to MCMC algorithm that may not span through all parameter space. Taking the MCMC limitation into account, the team attempts to implement Metropolis-Hasting. However, the posterior distributions remain the same with more distinction between each feature as shown in Figure 6.

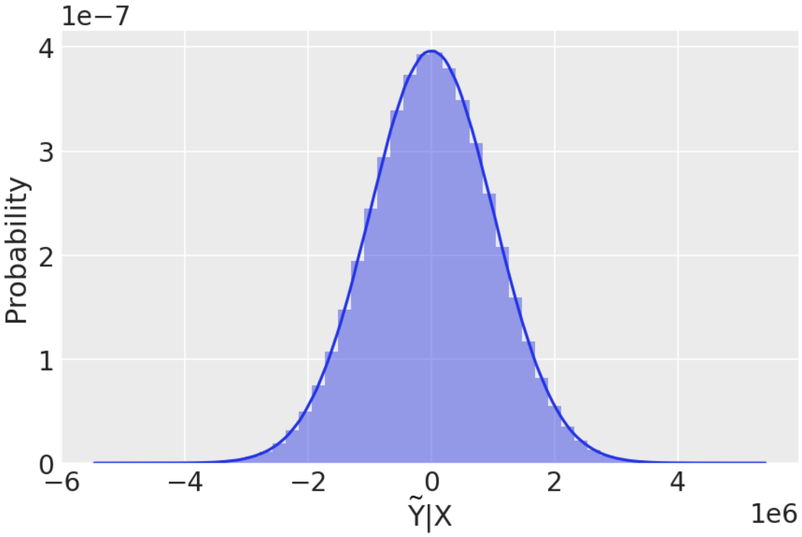


Figure 7: Posterior Predictive Distribution for Number of Streaming Given Audio Features

The team fits linear regression model according to the posterior distribution of β_j to obtain posterior predictive distribution without the effect of genre. The histogram in Figure 7 represents distribution of the number of streaming given solely audio features. This implies the generalized distribution of songs when genre information is absent. The mean of the distribution is around 0, which goes against the intuition. So, it is possible that there might be some unknown effects, and further analysis is recommended.

Initially, the team plan to compare the predictive posterior distribution of numbers of views without genre data from the above model $P(\tilde{Y}|X)$ with another predictive posterior distribution from pure Bayesian linear regression model where the inputs include genre $(\tilde{Y}|X, Z_j = j)$. The intuition is that the distribution of $(\tilde{Y}|X, Z_j = j)$ is similar to what happen in the real world as both genre and audio features can be observed by the audiences. By comparing the two model, we can directly investigate the effect of genre on success level of a song and how would the popularity change if the song were in different styles.

Since the predictive posterior distribution of the model without genre $P(\tilde{Y}|X)$ does not seems to be valid, the team will analyze only the result from model representing $P(\tilde{Y}|X, Z_j = j)$

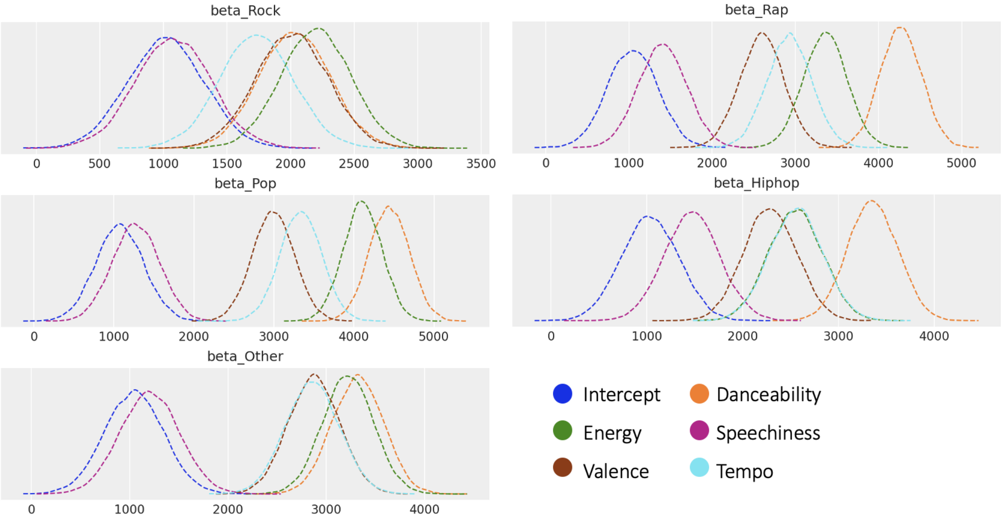


Figure 8: Posterior Distribution for Linear Regression Coefficients where Inputs Includes Genre

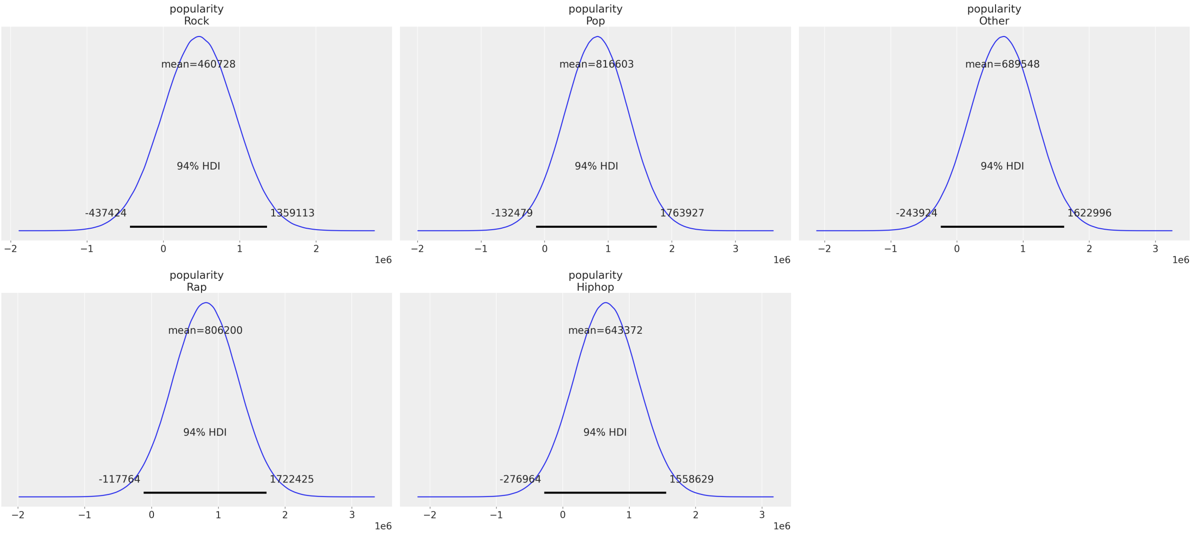


Figure 9: Posterior Predictive Distribution of Number of Streaming where Inputs Includes Genre

The posterior distribution in Figure 8 indicates that the effect of each audio feature on the popularity of a song varies across genre. For example, in rock genre, each audio features tends to have similar impact on song's popularity. Whereas, in some genre such as Pop or Rap, the magnitude of the effect on each feature are more dissimilar.

Figure 9 shows the posterior predictive distribution of song's popularity based on its genre. From the figure, each genre obtain different distribution of the song's popularity. Songs with Pop and Rap style reach highest number of streaming with a lower dispersion, where the lowest number of streaming is songs in Rock genre.

If the Bayesian linear regression model where inputs exclude genre label is reliable, we can examine how does the distribution of a particular song would shift if it has been done in a different style and derive more inference from the data.

5 Conclusion

Based on the analysis, it can be concluded that the song's genre have impact on the songs' popularity as seen in the posterior predictive distribution of model with genre labels. Given a song with its audio features, making a song in a Pop or Rap style will likely yield highest popularity. However, the information regarded to how and how much the distribution shifts is limited with the current model.

For further analysis, the team recommends examining the Bayesian linear regression model where the genre labels are absent as they may have some underlying issues. Moreover, one can apply the same structure of the model to study the relationship of music popularity, audio features, and artist as mentioned at the beginning if data are available.

6 References

References

[1] "A Primer on Bayesian Methods for Multilevel Modeling". [Online] Available: https://docs.pymc.io/en/v3/pymc-examples/examples/case_studies/multilevel_modeling.html.

[2] "Prior and Posterior Predictive Checks". [Online] Available: https://docs.pymc.io/en/v3/pymc-examples/examples/diagnostics_and_criticism/posterior_predictive.html.

[3] M. Kana, "Introduction to Bayesian Logistic Regression", Feb. 2020. [Online] Available: https://towardsdatascience.com/introduction-to-bayesian-logistic-regression-7e39a0bae691.

[4] "Spotify daily top 200 songs with genres 2017-2021". [Online] Available: https://www.kaggle.com/datasets/ivannatarov/spotify-daily-top-200-songs-with-genres-20172021 .

[5] "Spotify 1.2M+ Songs". [Online] Available: https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs.

[6] "Spotify Charts". [Online] Available: https://www.kaggle.com/datasets/dhruvildave/spotify-charts.

Statement Contributions

Nawat Swatthong

Conceptualization, Methodology, Model Coding, Model Investigation, Result Interpretation, Report Writing, Visualization

Eliza Chew

Conceptualization, Data Curation, Data Processing, Report Writing, Visualization

Pat Chimtanoo

Conceptualization, Data Curation, Result Interpretation, Methodology, Report Writing, Visualization, Meeting Facilitator