Assignment 2 Data Analysis 2

Anton Shestakov & Ekaterina Shemetova

Introduction

In the digital age, music streaming platforms like Spotify have revolutionized how people consume music, offering vast number of various genres and tracks. Among these, the popularity of songs is a key metric, influencing recommendations, playlist ranking, and revenue generation. Understanding the factors that drive a song's popularity is not only fascinating but also has significant implications for artists, producers, and the platform itself.

This study aims to assess the impact of music positivity on the popularity of songs on Spotify in 2024. By exploring a dataset containing 60,000 Spotify tracks spanning six languages—English, Hindi, Tamil, Telugu, Malayalam, and Korean, this analysis seeks to determine patterns and correlations between audio features and a track's popularity. The dataset includes attributes such as acousticness, danceability, energy, tempo, and other contextual elements that may contribute to a track's success.

Understanding these relationships could provide valuable insights into audience preferences and the characteristics of popular music across different cultures and languages.

Data

The dataset originally included information on more than 60,000 tracks in six languages. The set contains information on music tracks from Spotify released between 1971 and 2024. The final dataset includes 4,064 recordings and 12 variables that describe both the audio and contextual features of each track. There are also four languages left after data cleaning: English, Tamil, Hindi and Korean.

Key characteristics used in the subsequent analysis included the following metrics:

- **Popularity**: A score from 0 to 100 indicating the popularity of the track based on the number of listens and recent release.
- **Acousticness**: A score from 0.0 to 1.0 indicating that the track is acoustic.
- **Danceability**: A score between 0.0 and 1.0 indicating how danceable the track is.
- **Duration**: The duration of the track in milliseconds.
- **Energy**: The perceptual intensity of the track from 0.0 to 1.0.
- **Liveness**: The likelihood that the track was recorded live.
- **Loudness**: The average volume level of the track in decibels.
- **Mode**: Determines the tone of the track (1 = Major, 0 = Minor).
- **Speechiness**: Indicates the speech content of the track (0.0-1.0).
- **Tempo**: The estimated tempo of the track in beats per minute.
- **Valence**: A score between 0.0 and 1.0 indicating how "positive" the track is musically.
- **Language**: The language of the track (English, Tamil, Hindi, Korean).

Data processing consisted of several steps, the first of which was removal of missing and anomalous values. Records with values of parameters such as popularity or duration equal to -1 were also excluded, as they indicated errors in the original data. The second step was data type conversion, where numeric values (e.g. popularity, duration) were converted to float or int type for calculations. Variables such as modality (major/minor) and language were converted

to categorical data. For ease of analysis, the track languages (English, Hindi, Korean, Tamil) were converted to dummy numeric values.

Table 2 provides summary statistics of various characteristics of music tracks. The average value of track popularity is 8.13, which indicates a relatively low average popularity, but the wide range of values and a maximum of 93 indicate the presence of both obscure and extremely popular tracks. Energy has an average value of 0.59, which indicates a fairly intense sound component in most of the tracks. Musical positivity (valence) has an average of 0.50, indicating a balance between positive and negative musical mood. The average duration of the tracks is about 213900 ms (or 3.5 minutes), but a significant standard deviation indicates that both very short and long pieces are found among the tracks. Speech in the tracks (speechiness) has a mean value of 0.15, indicating a small amount of vocal content characterized by speech elements.

Table 1 presents descriptions of the mean statistical metrics on the speechability measure. Thus, English songs have the highest average speechiness value, which may indicate that English tracks have more speech. Korean and Hindi songs have similar speechiness metrics, with low mean values and small deviation. However, the Tamil language songs show a wider range of speechiness values, implying diversity in the use of speech or recitative in tracks in this language.

Language	**Feature**	**Mean**	**Std**	**Min**	**Max**	**Count**
English	Speechiness	0.17	0.28	0.02	0.96	2829
Hindi	Speechiness	0.08	0.07	0.02	0.41	257
Korean	Speechiness	0.08	0.08	0.03	0.54	121
Tamil	Speechiness	0.10	0.09	0.03	0.90	857

Table 1: Speechiness Statistics by Language

Model

(A)

In the course of analysis there were various approaches in building of final regression. On one hand, the model had to embrace as many fitting explanatory variables as possible to scrutinize the potential causality. On the other hand, it is very risky to obtain overfitted model that would capture noise and lead to non-robust results.
Eventually, the choice fall upon the model shown below:

$$\begin{aligned}\text{Popularity}_i = &\beta_0 + \beta_1 \cdot \text{valence}_i + \beta_2 \cdot \text{speechiness}_i + \beta_3 \cdot \text{speechiness}_i^2 \\ &+ \beta_4 \cdot \text{liveness}_i + \beta_5 \cdot \text{mode}_i + \beta_6 \cdot S_1(\text{tempo}_i) + \beta_7 \cdot S_2(\text{tempo}_i) \\ &+ \beta_8 \cdot S_3(\text{tempo}_i) + \beta_9 \cdot \log(\text{duration\_ms}) + \beta_{10} \cdot \log(\text{duration\_ms})^2 \\ &+ \beta_{11} \cdot \text{language}_i + \varepsilon_i, \end{aligned} \tag{1}$$

with I. I. D. assumptions

Although having other available variables that could be included in the final model, they were not added. The reason lies upon their distribution nature and relationship with variable of interest. After testing other variables, s.t 'energy', 'danceability' and etc., for their potential embedding in the regression, it was concluded that they were not appropriate as control variables (more detailed in the Chapter 'Generalization and external validity'). Their impact rather took away the significance of the key independent variable which relationship with dependent one the model is exploring.

(B)

Taking into account the specificity of the dependent variable "popularity", it was decided to keep it in its original form. Although the distribution of the variable is clearly right-skewed

(Figure 1), the variable contains integers between 0 and 100 with one-step move, which makes it, in fact, ordinal categorical variable. However, given the fact that it depicts rating, the growth of the value in numerical expression corresponds also to the growth of the value in semantic expression. Though the increase of the latter is likely non-linear, the model pursues to use the variable in numerical form for OLS-estimation.

As for independent variables, the model includes some modifications of them. The model employs squared values of 'speechness' variable because of non-linear behavior with respect to dependent variable (Figure 2).

Another issue relates to the right-skewness of the original 'duration_ms' distribution (Figure 3). To compress the data and make it more meaningful by getting rid of the outliers and uneven values placement the model includes logarithmic transformation. It also partially levels out the problem of heteroskedasticity. Logarithmic transformation creates more stable variance across the range making data more homoskedastistically. In addition, squared modification to logarithm was applied due to non-linear dependence with 'popularity' according to plot output (Figure 4).

Another important data engineering feature is linked with 'tempo' continuous variable. Exploring the plot, it is evidently clear that dependence with popularity is less likely linear. It more resembles a sinusoidal function behavior (Figure 5). Eventually, using splines was accepted as a trade-off solution to keep the 'tempo' as confounding variable that can lead to more robust estimation of the relationship between valence and popularity. In total, there were created three knots that describe the change of function trend.

(C)

The core results are demonstrated in Summary table 1. There are three specifications representing different types of statistical relationship between popularity and valence. These are parts of step-by-step analysis that led to the third (3) specification which is considered to be final model. In the opinion of the authors of this work, this model best describes the relationship between the dependent variable and the covariate of interest given set delimitations and external data limitation.

On average, ceteris paribus, the increment of variable 'valence' on one unit is associated with decreasing of 'popularity' value on 7.035 units with significance level of 1%. A similar statistically significant pattern of influence of valence on popularity is presented in other specifications as well.

Generalization and external validity

The correlation matrix (Figure 6) demonstrates statistical relations between variables. Based on its results, it was decided to exclude some variables highly-correlated with variable of interest (valence) to avoid multicollinearity. The regression model also excludes other strongly correlated variables which have the similar extent of explanatory power . These variables may distort the performance of the econometric model and overlap the significance of the variable of interest. Thus, variables such as danceability and energy, danceability and valence, loudness and energy were correlated and 2 variables were not included in the model at the same time (only one variable from the pair could occur in the model).

Although the model captures relationships between musical features (e.g., acousticness, speechiness, tempo) and popularity, its external validity is limited. The analysis is specific to a single year (2024) and does not account for temporal dynamics or changes in preferences over time. Musical trends are inherently influenced by temporal lags—what is popular in a given year often depends on trends and shifts from the preceding years. To improve external validity, future research should consider a panel or spatial model that accounts for time-dependent factors, enabling an analysis of how features and popularity evolve over multiple years.

Results (Table 3. Appendix) shows that while the magnitude of effects vary slightly across

languages, the direction and significance of coefficients are largely consistent, indicating robustness of the overall findings.

Moreover, to capture potential nonlinear relationships (based on lowess), quadratic terms for duration_ms and tempo were added in the model. These extensions did not lead to major changes in the primary coefficients, meaning that the initial model is appropriately specified.

The internal validity of the model is supported by selection of variables based on the correlation matrix, ensuring the inclusion of relevant predictors in the final regression. This reduced potential multicollinearity, enhancing the reliability of the estimates. Additionally, we conducted a robustness check by examining the correlation between valence and the model residuals (Figure 7). The weak correlation of 0.07 value obtained by Pearson correlation test suggests that the model less likely has omitted variable bias problem that positively affect the reduced endogeneity, thereby strengthening model's internal validity. This, in turn, increases the likelihood of a correct causal interpretation.

The analysis confirms that the model generalizes well across languages and is robust to alternative specifications. Key predictors of popularity, such as speechiness and valence, maintain their significance and interpretability across various scenarios.

Causal interpretation

Column (3) in summary Table 3 summarizes the main results of the regression model. Thus, valence is negatively associated with popularity, indicating that songs with higher valence (or happier tones) are less likely to be popular. It might be happening due to listeners preferences, they favor songs with emotional depth or complexity (e.g., melancholic or bittersweet tracks) over "happy" songs.

Speechiness has a positive linear relationship with popularity but a negative quadratic term, that means that moderate levels of speechiness boost popularity, while very high levels reduce it (concavity). Moderate speechiness may be indicative of spoken-word elements like rap or spoken intros, which are popular in certain genres. However too high speechiness corresponds to tracks dominated by spoken content (e.g., podcasts or audiobooks), which are not very popular and not "songs".

The estimated coefficient with liveness shows that live-recorded tracks are less favored (negative association). In turn, duration has a positive linear effect but the quadratic term shows diminishing returns for longer songs. For a song to have more chances to be popular is should be long enough but not too long. The 115 to 175 bpm range (which has a positive relation with popularity) is optimal for dancing and aligns with many popular genres, such as pop or hip-hop. Additionally, songs in Hindi, Tamil, and Korean have higher popularity compared to English-language songs.

Although the model reveals statistically significant relationships, the study has a number of limitations. The interpretation of causal relationships is limited due to potential endogeneity, omitted variable bias and year-limit to 2024.

Factors such as marketing budgets, artist reputation, or promotion on a particular platform may distort the observed relationships. For example, popular songs may be promoted more, increasing their perceived association with specific features such as tempo or language.

Conclusion

The research has shown many interesting patterns that can be applied to music industry business. Producers should focus on creating tracks with emotional depth, optimal tempo range and moderate volume levels. Platforms can benefit from diversifying their catalogs to include more non-English tracks, expanding their global reach. Marketing campaigns should emphasize these insights to effectively target listener preferences.

Appendix

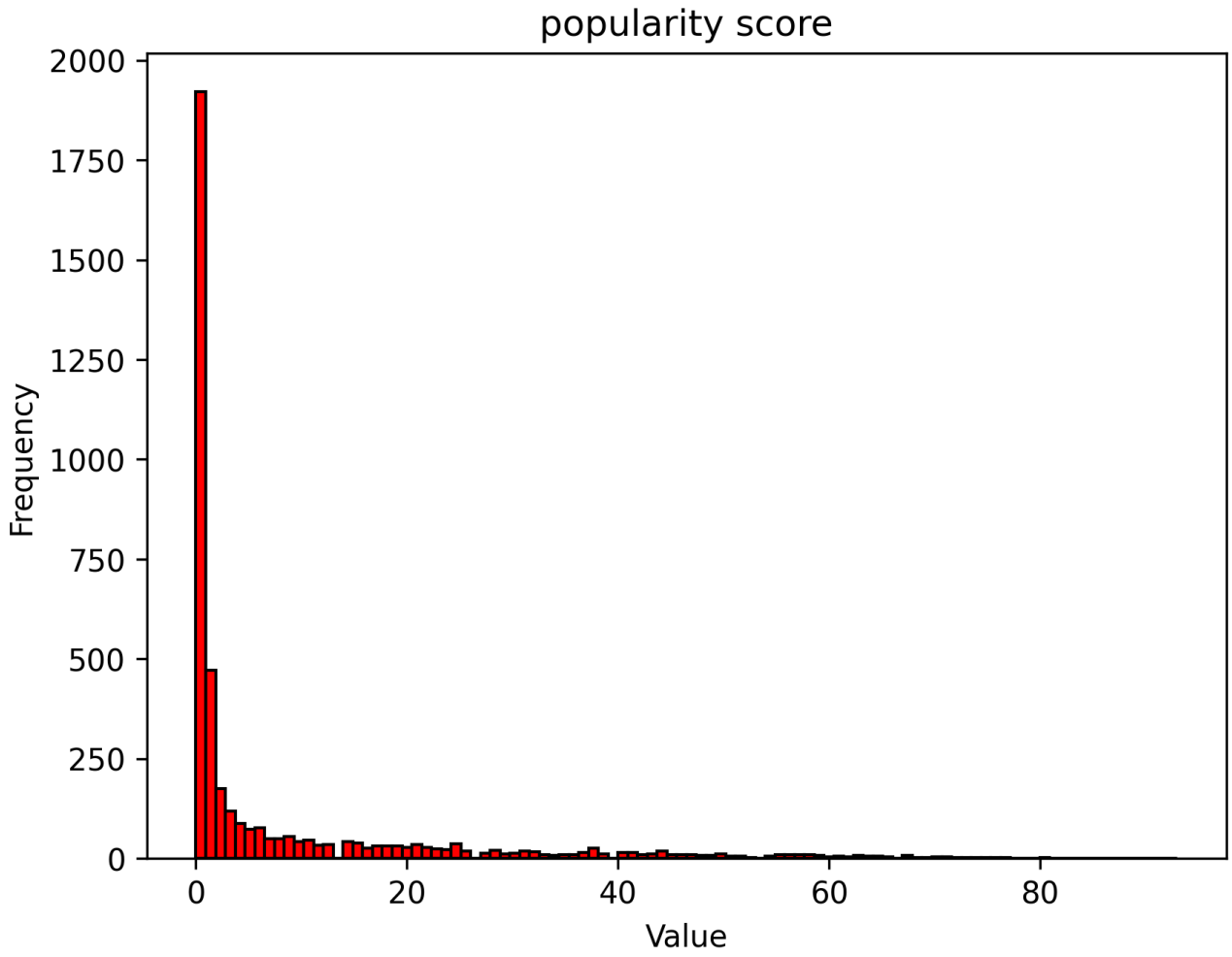

Figure 1: Popularity Score Histogram

Figure 2: Speechiness Histogram

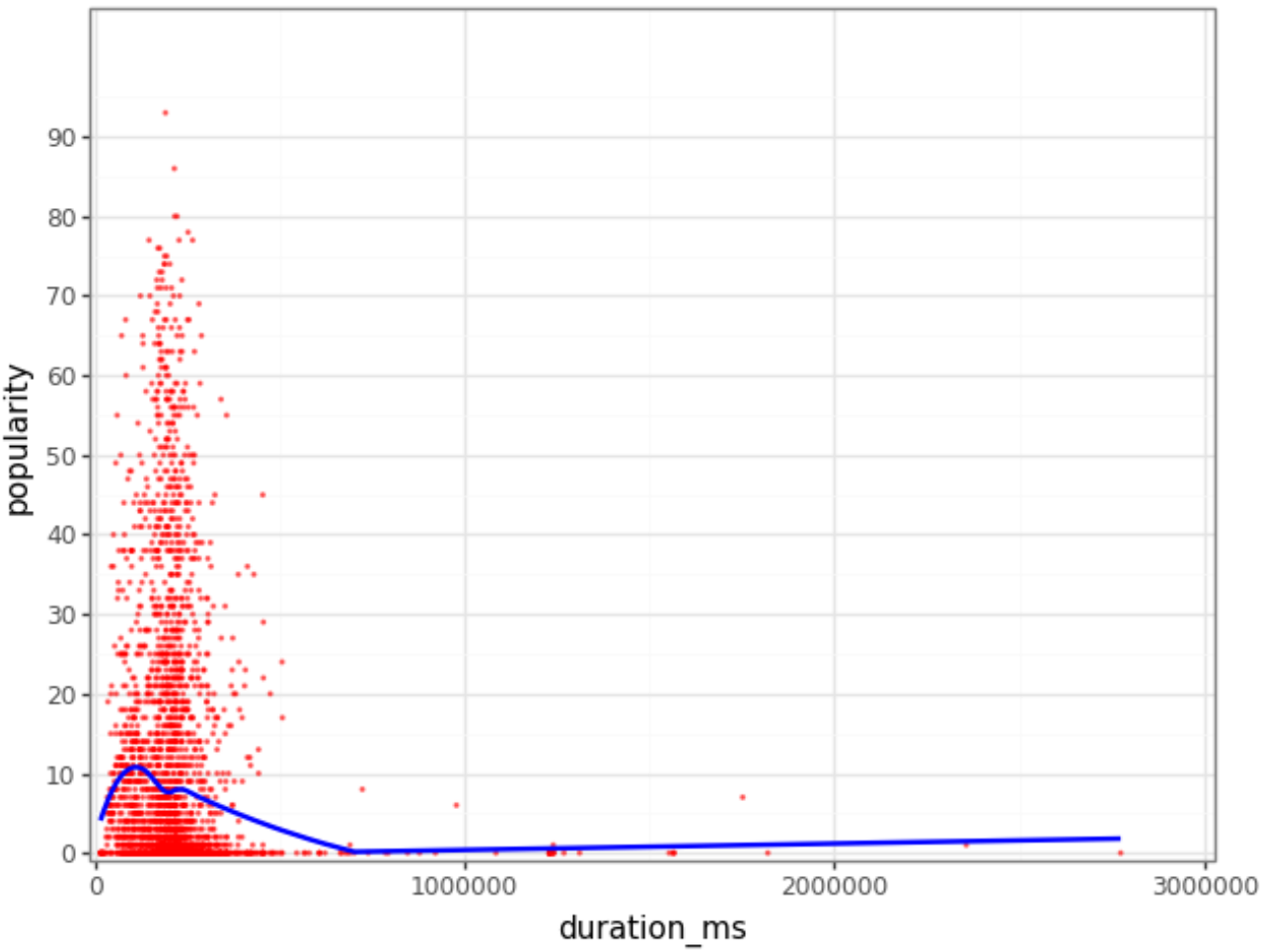

Figure 3: Duration_m Histogram

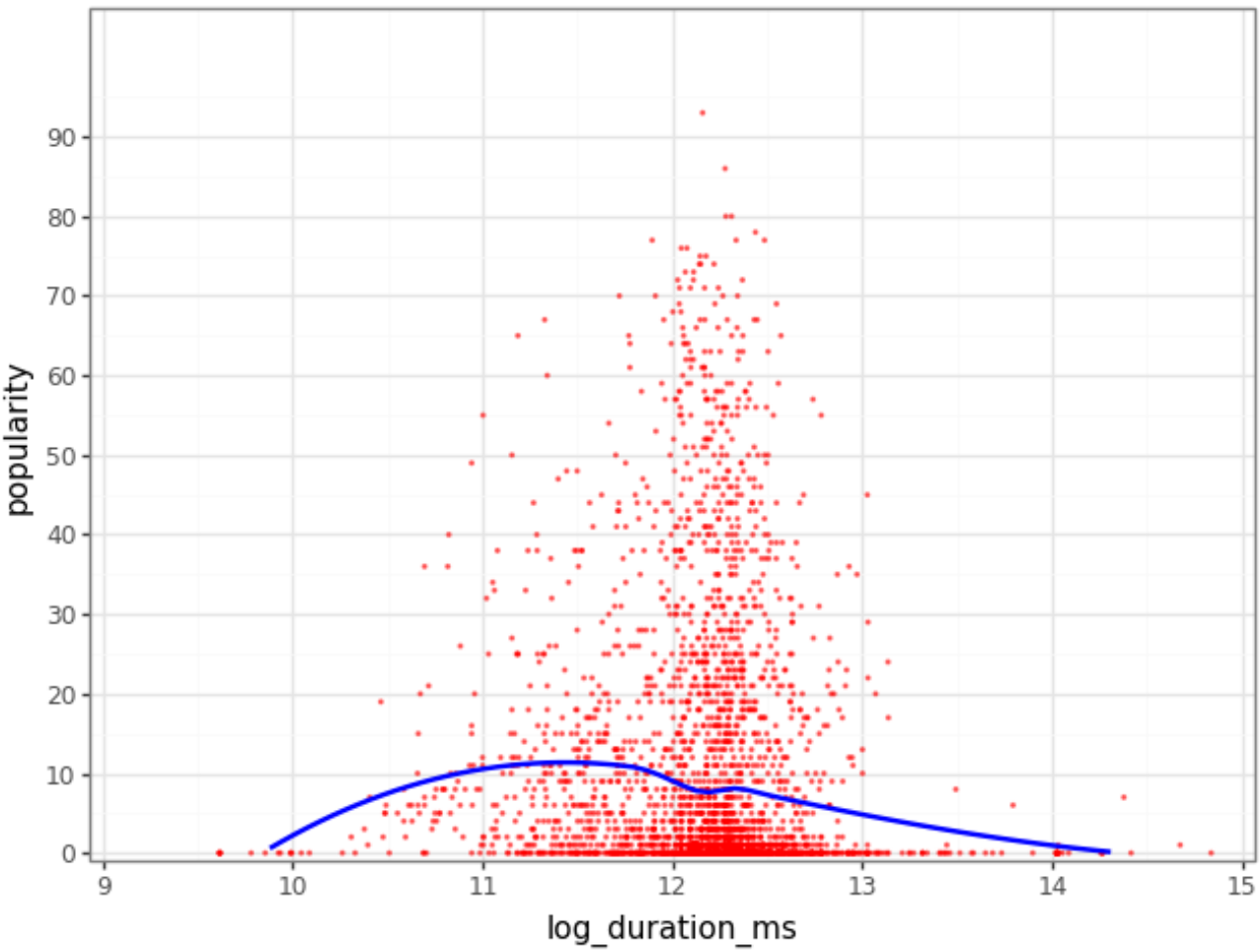

Figure 4: Log_duration_m Histogram

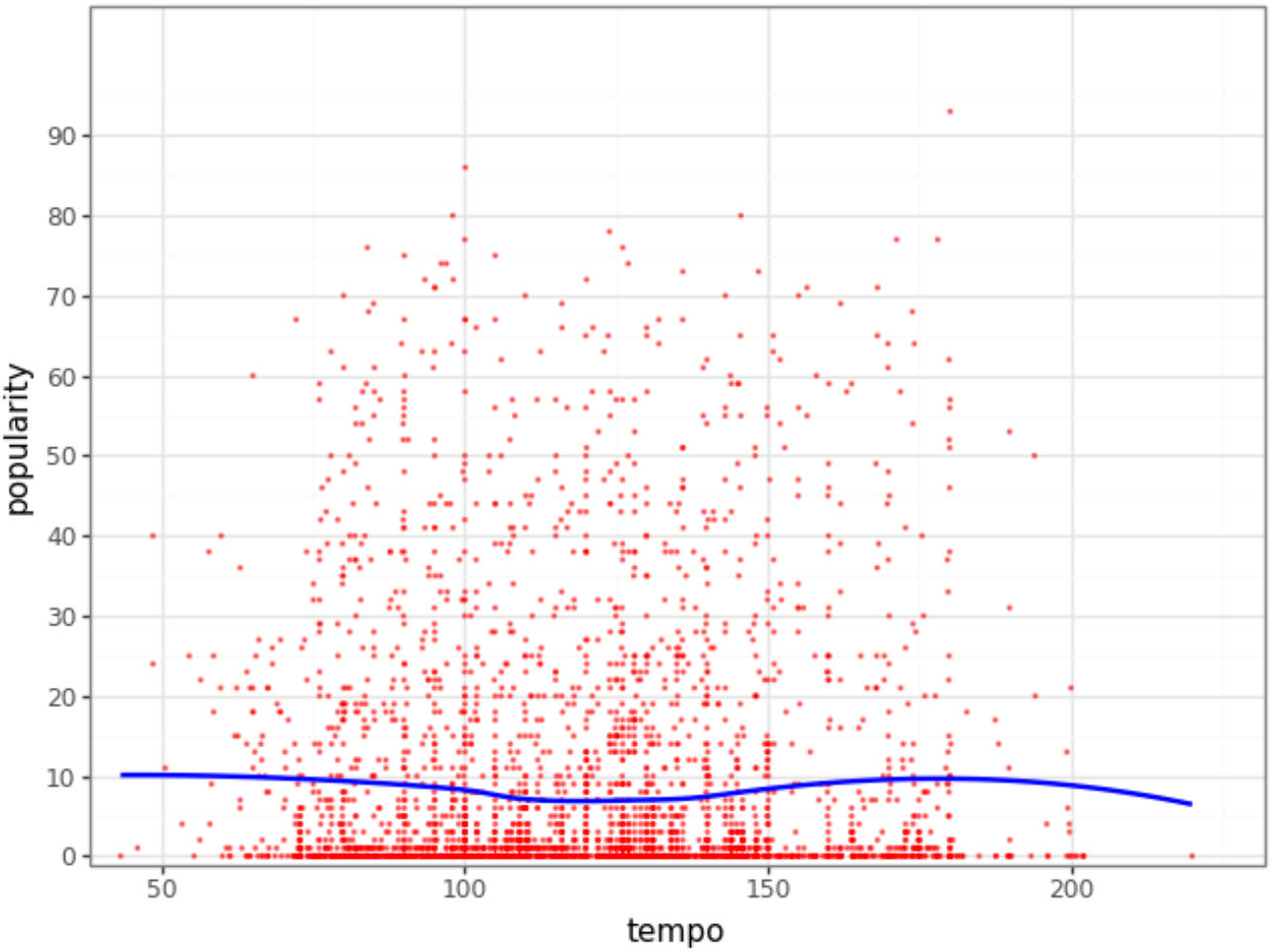

Figure 5: Tempo Histogram

Table 3.

	Dependent variable: popularity		
	popularity	popularity	popularity
	(1)	(2)	(3)
Constant	11.418***	-261.022***	-287.625***
	(0.495)	(36.264)	(35.723)
valence	-6.533***	-7.489***	-7.035***
	(0.879)	(0.992)	(0.981)
speechiness		50.652***	39.266***
		(6.332)	(6.051)
speechiness^2		-62.867***	-48.659***
		(6.451)	(6.156)
liveness		-6.341***	-5.958***
		(1.380)	(1.357)
mode		-0.824*	-0.309
		(0.486)	(0.480)
tempo less than 115		-0.067***	-0.069***
		(0.019)	(0.019)
tempo [115,175]		0.051***	0.064***
		(0.017)	(0.017)
tempo more than 175		-0.140	-0.121
		(0.094)	(0.092)
log Duration		47.673***	51.502***
		(5.989)	(5.887)
log Duration^2		-2.033***	-2.182***
		(0.248)	(0.244)
Hindi			9.191***
			(1.458)
Tamil			5.491***
			(0.624)
Korean			12.032***
			(2.212)
Observations	4064	4064	4064
R^2	0.012	0.078	0.126
Adjusted R^2	0.011	0.076	0.124
Residual Std. Error	14.938 (df=4062)	14.442 (df=4053)	14.066 (df=4050)
F Statistic	55.289*** (df=1; 4062)	108.269*** (df=10; 4053)	89.513*** (df=13; 4050)
Note:			*p<0.1; **p<0.05; ***p<0.01

Column Name	Mean	SD	P0	P25	P50	P75	P100
Popularity	8.13	15.02	0	0.00	1	9	93
Acousticness	0.32	0.30	0.00000494	0.04	0.21	0.56	0.996
Danceability	0.63	0.17	0.06	0.58	0.66	0.74	0.97
Duration (ms)	213900	126500	15000	175700	203200	233800	2774000
Energy	0.59	0.25	0.0013	0.42	0.63	0.80	0.996
Liveness	0.20	0.16	0.017	0.092	0.13	0.28	0.96
Loudness	-9.15	5.71	-39.17	-10.77	-7.34	-5.41	-0.70
Mode	0.57	0.50	0	0	1	1	1
Speechiness	0.15	0.25	0.024	0.04	0.05	0.102	0.96
Tempo	119.4	27.69	43.18	99.51	121.1	138.7	220
Valence	0.50	0.25	0	0.33	0.49	0.70	0.99

Table 2: Summary Statistics of Features

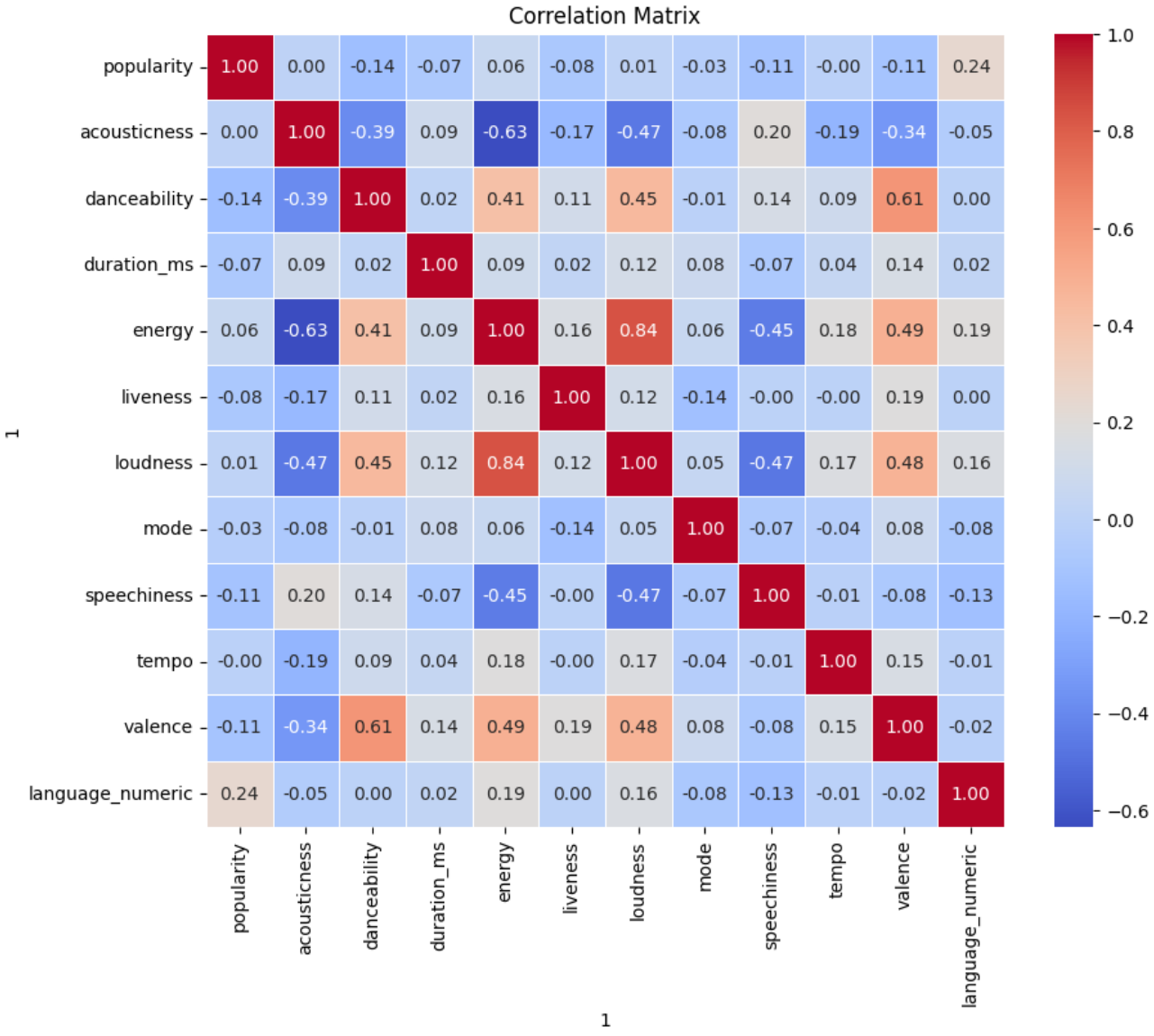

Figure 6: Correlation Matrix

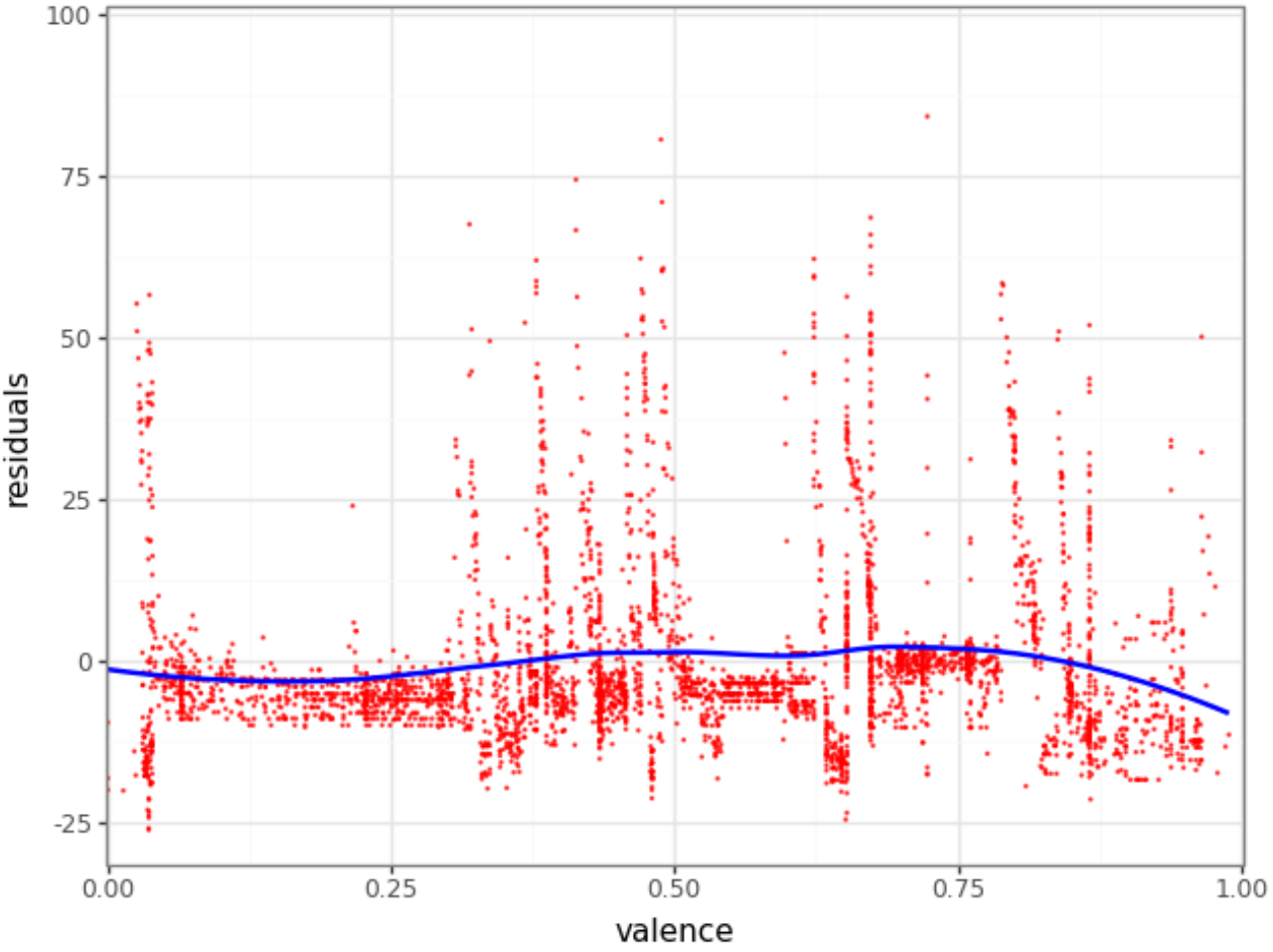

Figure 7: Residuals vs valence