Group Project (P12)

Group Project (P12)

Group members: Yifan Hao, Duncan Harrop, Elias Khan, Marcus Lee-Frazier

Introduction

As students, music plays a big role in our lives, especially while studying. However, we noticed that some songs seem to gain much more attention and streams than others. Not only is this topic of personal curiosity for us, but being able to understand the factors in a song's success would have a huge impact on the music industry as a whole. Therefore, for this project, we aim to explore how a song's musical attributes contribute to its popularity, as measured by the number of streams of Spotify, using the statistical methods learned in STAT 306.

Specifically, our research question is: ***How do the musical attributes of a song affect its number of streams on Spotify?***

To investigate this question, we will be using the "Spotify Most Streamed Songs" dataset which we obtained from Kaggle. This data set contains song data on 943 songs from Spotify and was gathered through web scraping the Spotify API. The data set is updated annually, thus, it will currently only contain data up to 2023. There are a variety of different features in this data set, such as track info and chart rankings, but we will only focus on the musical attribute ones. The others were omitted from this report for clarity.

Musical attributes can be defined as "the building blocks of music" (Dunnett) with some examples being: BPM and Mode. In particular, we will be focusing on the following musical attribute features found in the data set:

- BPM (`bpm`): Beats per minute, representing the tempo of the song. (Continuous variable)
- Mode (`mode`): Major or minor. (Categorical variable, 2 levels)
- Key (`key`): The key the song is in. (Categorical variable, 11 levels in this dataset)

- 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. (Continuous variable, percentage)
- Valence (`valence_.`): describes the musical positiveness conveyed by a track. (Continuous variable, percentage)
- Energy (`energy_.`): represents a perceptual measure of intensity and activity of the song. (Continuous variable, percentage)
- Acousticness (`acousticness_.`): Acoustic sound presence in the song. (Continuous variable, percentage)
- Instrumentalness (`instrumentalness_.`): Predicts whether a track contains no vocals. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. We plan to turn this into a categorical variable, with values > 0.5 being TRUE that the track is instrumental. (Categorical variable, 2 levels)
- Liveness (`liveness_.`): Presence of live performance elements. (Continuous variable, percentage)
- Speechiness (`speechiness_.`): Amount of spoken words in the song. (Continuous variable, percentage)

Our response variable will be:

- Streams (`streams`): Number of streams the song has on Spotify. (Continuous variable)

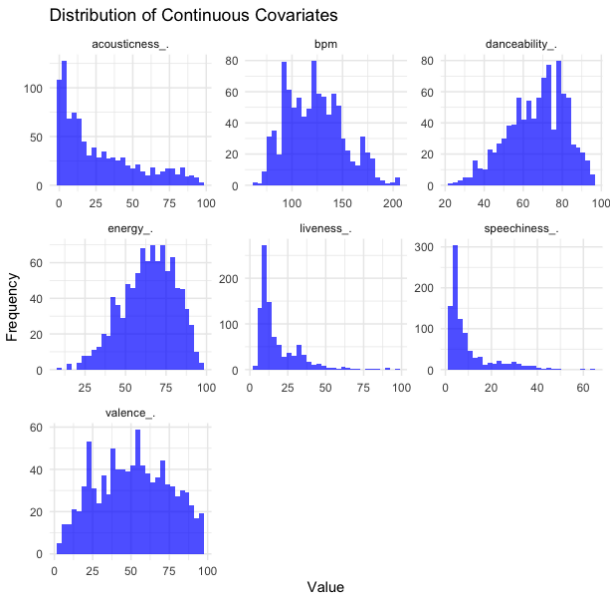
The values for these variables were calculated via Spotify's own proprietary algorithms. For more information about these variables, please see: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

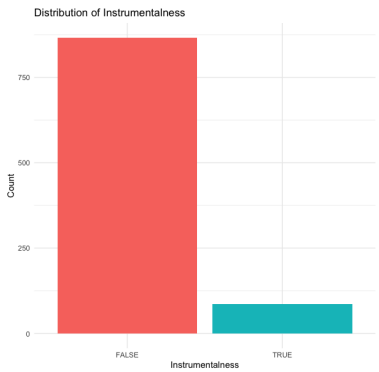
To clean the data prior to starting our project, we performed the following procedures: remove irrelevant columns, turn `streams` to a numeric value, transform `instrumentalness_.` to a boolean column using as.logical, turn `mode` to a factor

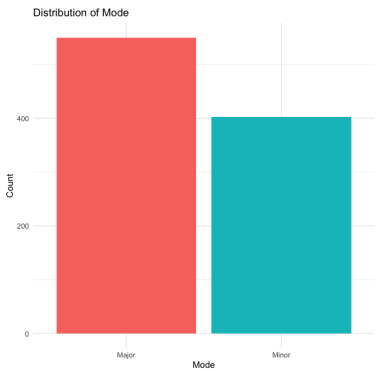
Analysis

Data Exploration

We begin with plotting the distributions of each of our covariates, to better understand our dataset and inform our analysis of it.

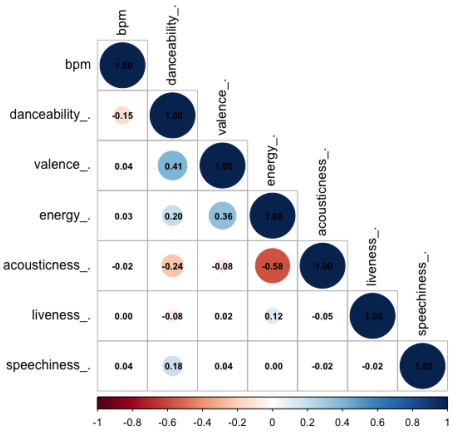






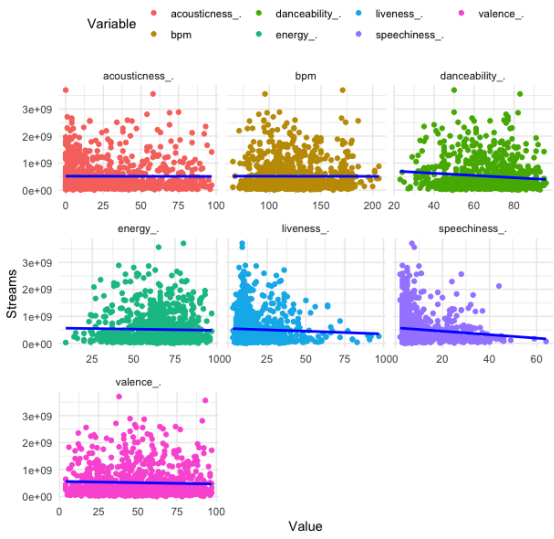
The first thing to point out is that the vast majority of songs in our data set are not instrumental. Given the context of songs on Spotify, this skewness for non-instrumental songs makes sense, because most songs on Spotify have some sort of singing or rapping in them. We can also see that some of our continuous variables, namely acousticness, liveness, and speechiness are right-skewed. This skewness could have an affect on our linear regression model fit, therefore we should pay attention to them and be open to transforming them (log, square root, etc.) or shifting them.

We can then visualize the correlation between each pair of covariates through a heatmap of the correlation matrix:



In general, most covariates are not highly correlated with the other covariates. In fact, most of the pairwise correlation values are close to 0, which is good. Acousticness and energy are the two covariates with the highest correlation, with it being around -0.6. However, these two covariates are not highly correlated enough to worry about multicollinearity. Therefore, we deem it not necessary to employ any techniques to reduce multicollinearity, such as combining covariates or shifting covariates by their mean.

We can also visualize the relationship between each of our covariates and the response variable streams:



As we can see, there is generally no obvious pattern we can see between any of the covariates and the response variable. Most plots look random, with speechiness being the one with

the strongest linear correlation with streams. Therefore, we can expect that a simple linear regression model will not model the response variable well.

Model Selection

We aim to use a linear regression model to explain the relationship between a song's musical attributes and the number of streams it has on Spotify. To determine a good model, we utilized backwards and forwards selection, both at the 5% significance level.

The backwards selection method resulted in the following model:

$y = 579247918 - 6422005 * x_{speechiness}$

and the forwards selection method resulted in the following model:

$y = 917378971 - 3994710 * x_{danceability} - 668600 * x_{acousticness} - 116830872 * z - 2641055 * x_{liveness} - 5796854 * x_{speechiness}$

where y is the number of streams on Spotify and $z = 1$ if the song is instrumental, and 0 otherwise.

The regression analysis for each model are summarized below:

Backwards Selection Model:

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	579247918	26130426	22.168	< 2e-16
speechiness_.	-6422005	1843079	-3.484	0.000516

The R^2 statistic is 0.01262, Adjusted R^2 statistic: 0.01158

Forwards Selection Model:

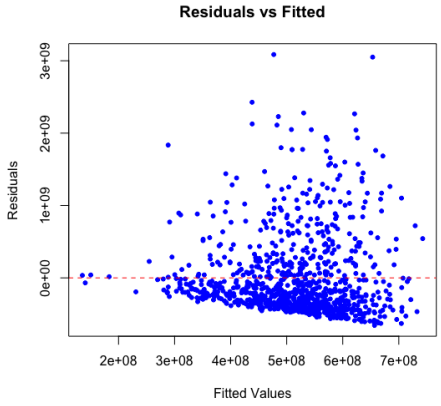
Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	917378971	100574506	9.121	< 2e-16
danceability_.	-3994710	1307200	-3.056	0.00231
acousticness_.	-668600	722141	-0.926	0.35476
instrumentalness_.	-116830872	63653181	-1.835	0.06676
liveness_.	-2641055	1334937	-1.978	0.04817
speechiness_.	-5796854	1872789	-3.095	0.00202

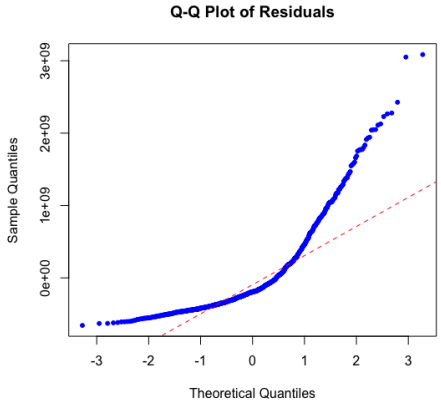
The R^2 statistic is 0.02788, Adjusted R^2 statistic: 0.02274, Mallow's Cp statistic: 3.34

Based on these results, we have decided to go with the forwards selection model, since it has significantly higher R^2 and adjusted R^2 statistics with 4 significant variables at the 5% significance level.

Model Fit

To gain insight into the model fit, as well as to see if any assumptions have been violated, we can look at the following plots:





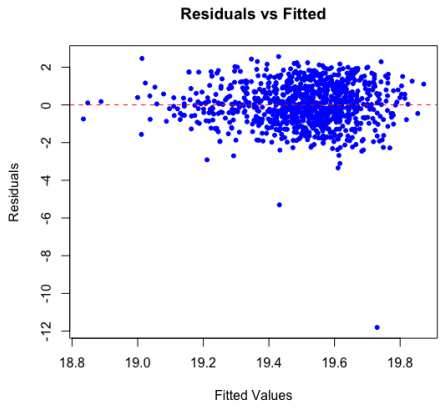
Given these plots, we can see that the assumptions of homoscedasticity and normality have been violated. To remedy this, we tried a log transformation of our response variable which seems to have improved our model.

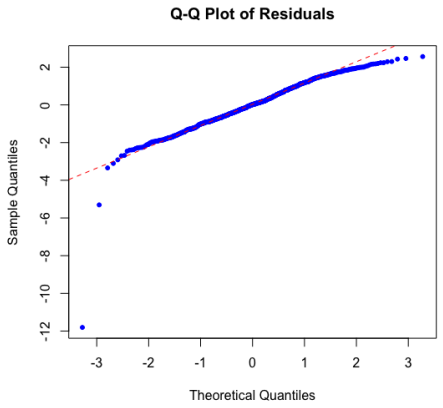
The log transformed model summarized:

$log(y) = 917378971 - 3994710 * x_{danceability} - 668600 * x_{acousticness} - 116830872 * z - 2641055 * x_{liveness} - 5796854 * x_{speechiness}$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	20.131554	0.204367	98.507	< 2e-16
danceability_.	-0.005410	0.002656	-2.037	0.04197
acousticness_.	-0.002123	0.001467	-1.447	0.14831
instrumentalness_.	-0.084403	0.129343	-0.653	0.51420
liveness_.	-0.004825	0.002713	-1.779	0.07558
speechiness_.	-0.010796	0.003806	-2.837	0.00465

The R^2 statistic is 0.01842, Adjusted R^2 statistic: 0.01323





We can see our plots look a lot better, but at the cost of interpretability of our model and lower R^2 statistics.

Discussion and Conclusion

In our final model, we have 3 significant covariates: danceability, liveness, and speechiness. All three have negative correlation with the response variable. This makes sense as the most popular songs are not songs strictly for dancing, are not live performances, and are not entirely made up of spoken word. However, given the poor fit of our model, these results should not be taken as absolutes. Related research has also found liveness and speechiness to be negatively correlated with number of streams, but conversely determined that danceability had positive correlation to number of streams (Kaneria et al., 2021). We do note that, in the related research, the correlation coefficient between danceability and streams has a value of 0.26, representing a weak positive correlation (Kaneria et al., 2021).

With respect to our research question of "How do the musical attributes of a song affect its number of streams on Spotify?", and given the results of our linear regression analysis, we conclude that there is no significant statistical evidence that musical attributes affect the number of streams a song has on Spotify. However, one should not conclude that musical attributes of a song have no influence on streams at all since there are factors and interactions we have not accounted for in this project. Previous research found significant predictive power of high danceability and low instrumentalness towards number of streams (Al-Beitawi et al., 2020). However, their research analyzed songs released between the years 2018 and 2019 whereas our dataset consisted of music released in the range of 1930 to 2023 (Al-Beitawi et al., 2020). Research has also found that coefficient estimates for musical attributes vary over

time. In 2018, speechiness and acousticness coefficient estimates stayed relatively positive, values peaking at 0.3 and 0.6 respectively, and in 2019 both trended towards 0 (Çimen and Kayış, 2021). Therefore, the difference in timing of song analysis and data preparation may account for our model's lack of predictive power. In addition, the related research makes note of change in listener habits over time, such as the dramatic increase of streams in Christmas songs in December (Çimen and Kayış, 2021). This observation is an example of an influence on the number of streams that this project does not account for.

We also fitted a logistic regression model (see appendix) to see if we could find some sort of difference in musical metric use to see if there was a difference between the songs that are ultra popular (over 1 billion streams) and those who aren't. It doesn't really seem to me that there is a great difference between these groups either.

During this project, we also noticed some limitations with our dataset. Mainly that it only contained the "most popular" songs on Spotify, meaning that it was biased. To have a more comprehensive analysis, we should use a dataset containing a mixture of popular and non-popular songs, as well as account for other covariates, such as release date. We also may want to analyze the change in covariate estimates over time. Furthermore, there may have been some outliers in our dataset which affected our model negatively. It may have been worthwhile to investigate this further.

Code/Data submitted by Yifan Hao

References

Al-Beitawi, Z., Salehan, M., and Zhang, S. (2020). What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs. Journal of Marketing Development and Competitiveness, 14(3): 79-91.

Çimen, A. and Kayış, E. (2021). A Longitudinal Model for Song Popularity Prediction. In Proceedings of the 10th International Conference on Data Science, Technology and Applications - DATA; ISBN 978-989-758-521-0; ISSN 2184-285X, SciTePress, pages 96-104. DOI: 10.5220/0010607700960104

Dunnett, B. (n.d.). *The elements of music*. Music Theory Academy. Retrieved December 8, 2024, from https://www.musictheoryacademy.com/how-to-read-sheet-music/the-elements-of-music/

Kaneria, A.V., Rao, A.B., Aithal, S.G., Pai, S.N. (2021). Prediction of Song Popularity Using Machine Learning Concepts. In: K V, S., Rao, K. (eds) Smart Sensors Measurements and Instrumentation. Lecture Notes in Electrical Engineering, vol 750: 35-48. Springer, Singapore. https://doi.org/10.1007/978-981-16-0336-5_4

Appendix

```
spotify$ultra_popular <- ifelse(spotify$streams >= 1e9, 1, 0)

table(spotify$popularity_group_1B)

head(spotify)

logistic_model <- glm(ultra_popular ~ bpm + mode + danceability_. + valence_.
+ energy_. +

acousticness_. + instrumentalness_. + liveness_. + speechiness_.,

data = spotify, family = binomial)
```