Frank Pasqualini

Carl Bodenschatz

STAT 1223

27 Feb 2019

Spotify Music Library Regression Analysis

This study is going to be conducted to build a multivariate regression model relating multiple attributes of a track, to a specific user’s musical preferences. The response variable will be a play count, the amount of times that the user has listened to that song. We believe that a play count is a good estimate of how much a user likes a track, because if they like the track they will listen to it more. It is not a perfect estimate, however, because it is possible for a user to have discovered a track recently that is one of their favorites, but due to not having known about it for a long period of time the play count would be low. The reverse is also true, where a user could have a song that they think is okay but not one of their favorites, but since it has been in their library for years, they have listened to it more times. It is very difficult to quantify how much a user likes a track, and we believe that play count is the best estimate we can currently research. We think these results will be particularly interesting, because if this study is successful, it could be turned into a generalized tool that a user could insert the data from their own music library into. The resulting regression model could be then used on a per-individual basis to find songs that it is likely that the individual would enjoy listening to, by giving recommendations for songs that would maximize the response variable based on the predictors.

For this study, we will be performing the regression on the music library of Frank Pasqualini, using data from his Spotify and Last.fm accounts. At the time of writing this essay, Frank has 5,626 tracks saved to his Spotify library and 35,199 scrobbles saved to his Last.fm account. A scrobble is a recorded instance of a track being played. Frank Pasqualini has been tracking his scrobbles using Last.fm since December of 2017, and those scrobbles were tracked directly by linking his Last.fm account to his Spotify account, so the data should be mostly valid even though the predictors will be collected from a different source than the response. Right now, we are looking into using 15 predictor variables, which are as follows: duration\_ms, explicit, popularity, key, mode, time\_signature, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, valence, and tempo. These results will be retrieved for each track using Spotify’s Web API.

Duration\_ms is a measurement of the length of the track in milliseconds. Explicit is a categorical variable that tells whether the track has been marked to contain vulgar language. Popularity is an integer value between 0 and 100 that is a calculated by one of Spotify’s algorithms based on total plays and how recent those plays are. Key is a categorical variable representing the musical key the track is estimated to be in, with 0 being C, 1 being C#/D♭, all the way up to 11 representing B/C♭. Mode is another categorical variable with 0 representing if the track is in a minor key and 1 if the track is in a major key. Time\_signature is an estimation of how many beats are in each bar. Acousticness is a confidence measure of whether the track is acoustic. It has the distribution featured in Figure 1. Danceability is an estimate of how well suited a track is for dancing to, and has the distribution shown in Figure 2. Energy is an estimate of the intensity of the track, and has the distribution featured in Figure 3. Instrumentalness is a confidence measure of whether the track contains vocals and has a very skewed distribution represented by Figure 4. Liveness is a confidence measure of whether there is a live audience in the recording and has the distribution in Figure 5. Loudness is the measure of the average decibel value of the track and has the distribution in Figure 6. Speechiness represents the exclusivity of speech in the track, and high values are almost entirely speech while low values contain little speech but possibly sung words. It has the distribution featured in Figure 7. Valence is a measure of the “happiness” of a song and has the distribution in Figure 8. Finally, tempo is the estimated beats per minute of the track and has the distribution in Figure 9. Almost all the variables described above are retrieved from the “Get Audio Features for a Track” Spotify API call. The documentation for this API call gives in depth descriptions about what each of these values represent, as well as providing the distributions that we have included below.

The data for this study will be collected with a python script which will parse the entire library, get the audio features for each track returned, and use Last.fm’s API to get the play count. The population is defined as all songs the user could like, and the sample will be a random sample of a yet undetermined size from the user’s library. Unfortunately, this sample will be biased towards songs with higher play counts and popularity, because the user is more likely to have added songs to their library if they have heard the song multiple times before, which is familiarity bias.

Figures

Figure 1

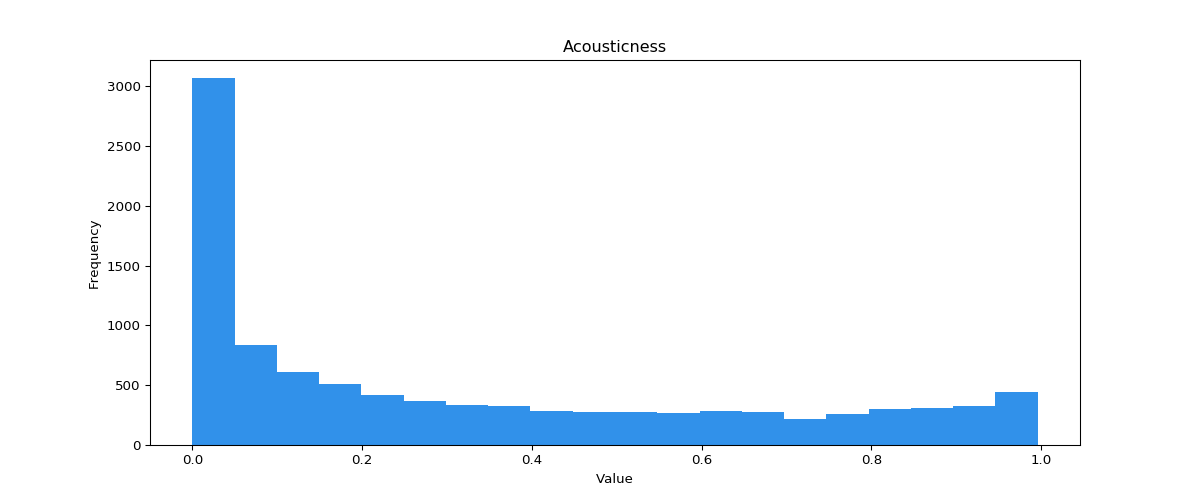


Figure 2

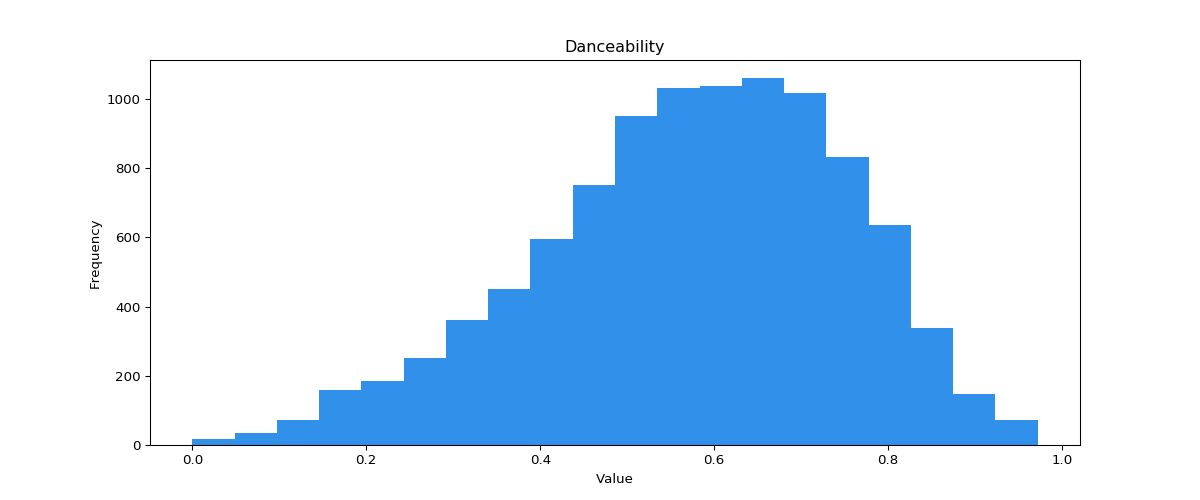


Figure 3

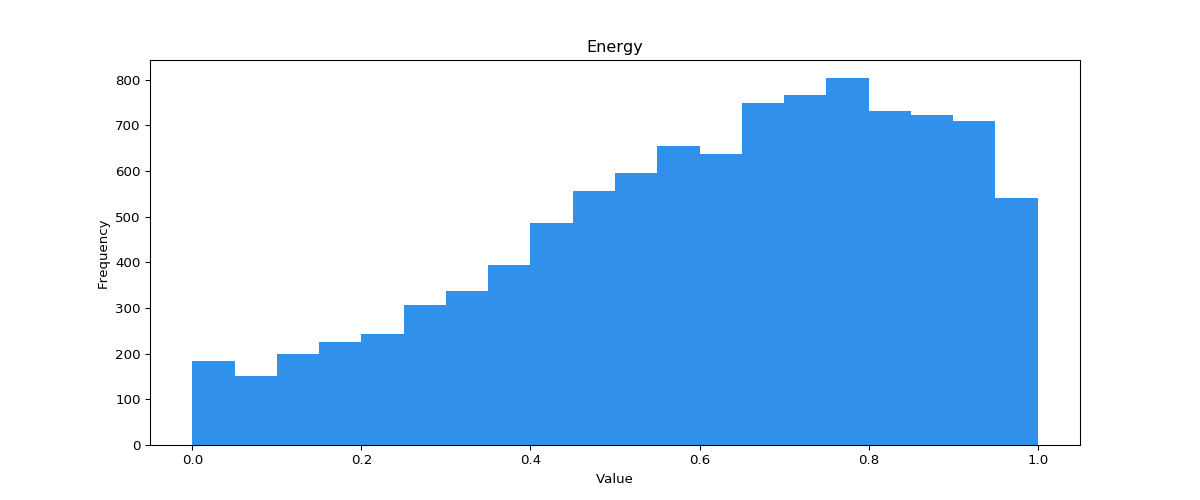


Figure 4

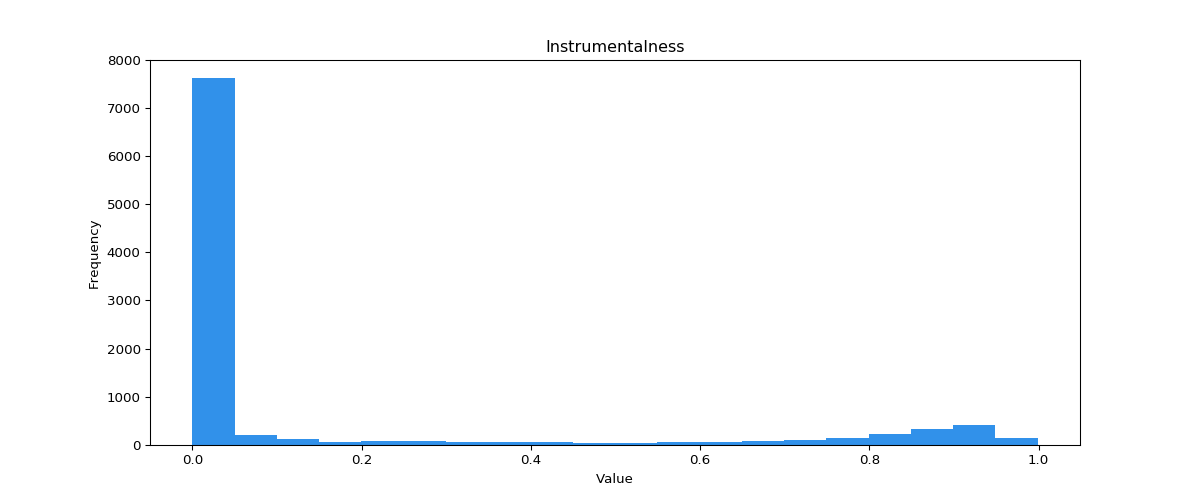


Figure 5

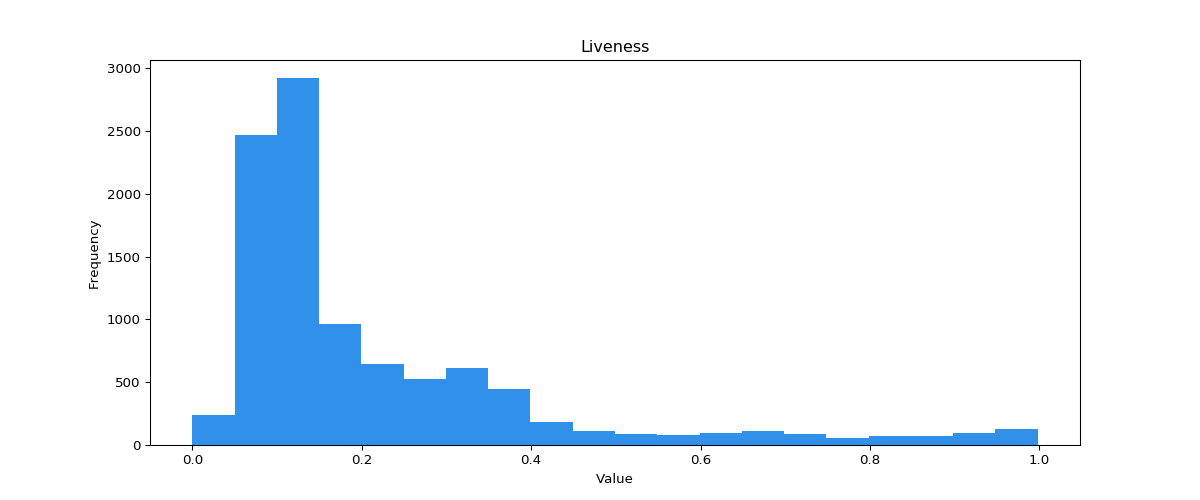


Figure 6

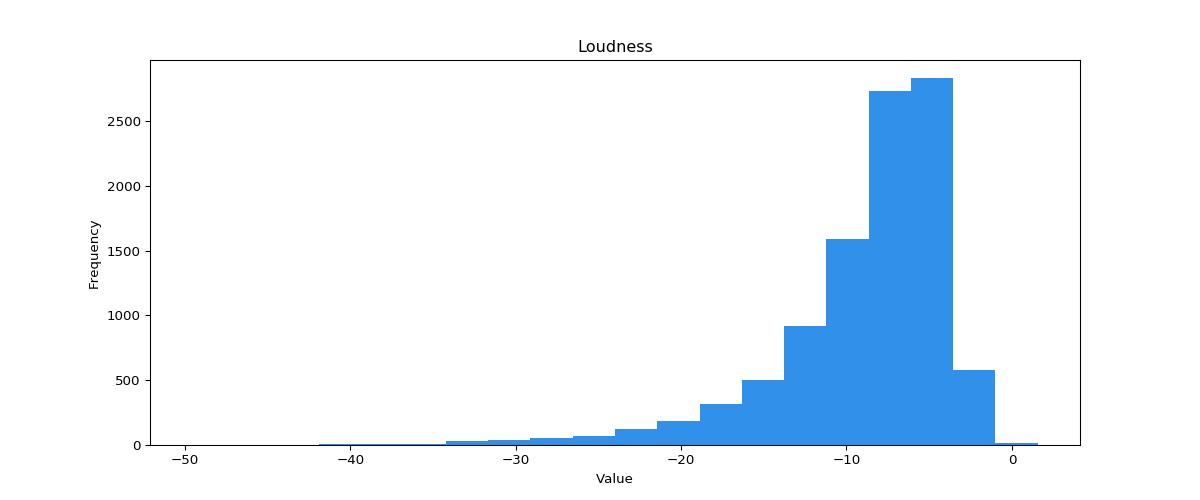


Figure 7

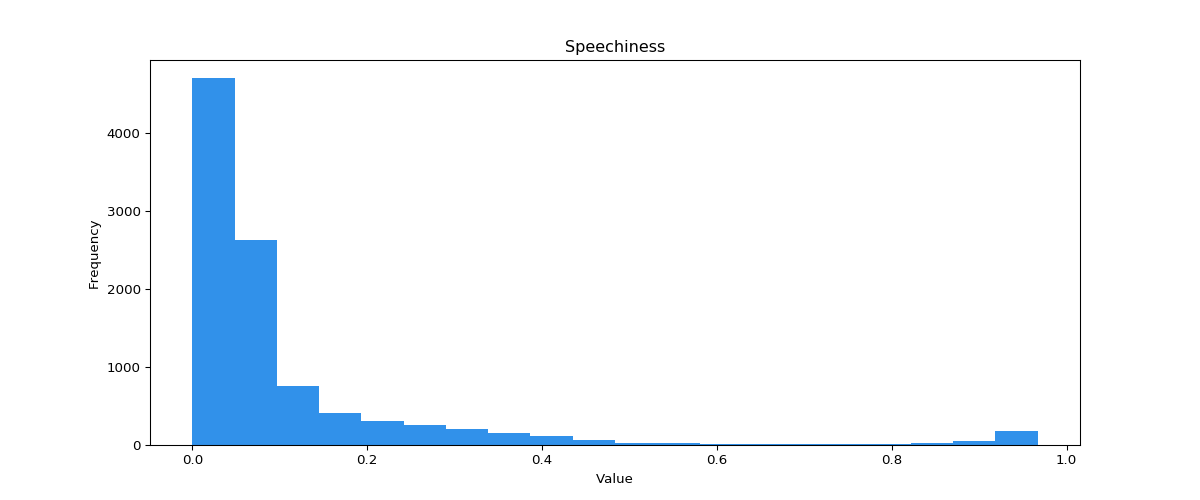


Figure 8

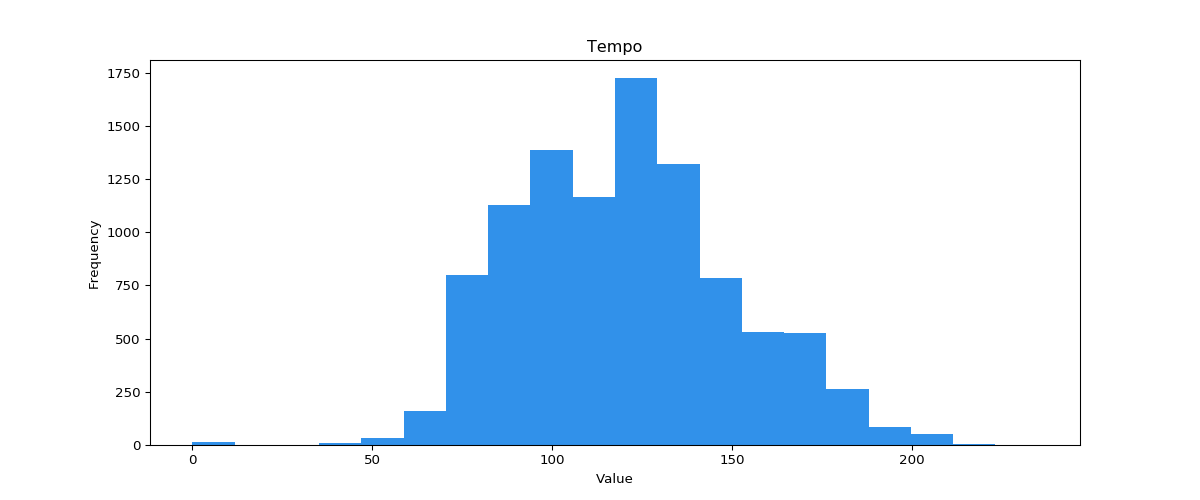
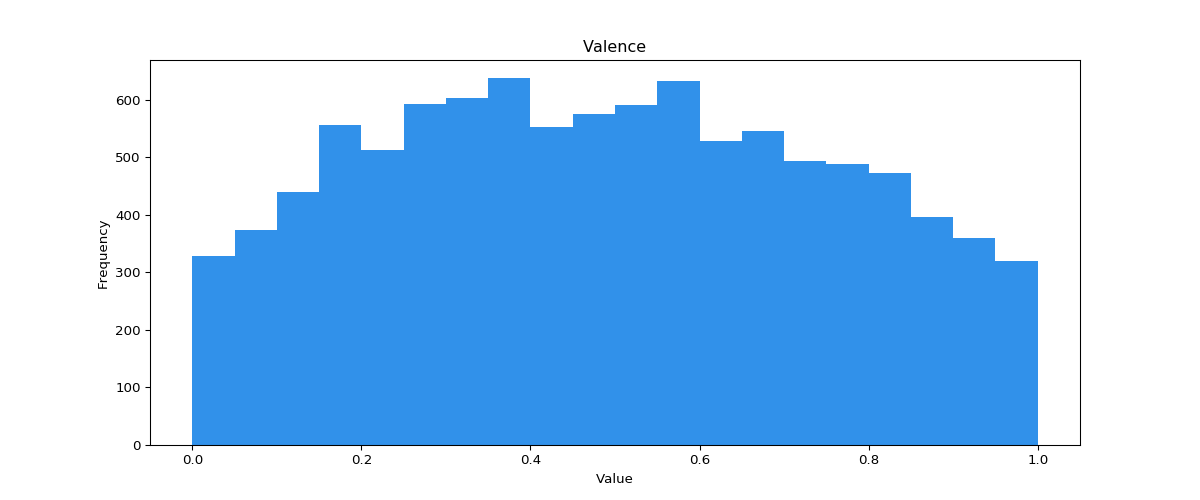


Figure 9

Works Cited

“Get Audio Features for a Track.” Spotify for Developers, developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/.