HOW LOUDNESS, SONG NEGATIVITY AND PLAYLIST PERSONALIZATION CAN INCREASE SPOTIFY’S CUSTOMER RETENTION

Tim Kennelly¹*, Paul D. Berger²
¹Bentley University Waltham, MA 02452 U.S.A
²Bentley University Waltham, MA 02452 U.S.A

*Corresponding Author:

Abstract:
In this paper, we consider the music streaming service provider, Spotify. We consider 14 attributes of a song (13 of which provided by Spotify, mostly subjective, a few objective) and analyze the relationship between these 14 attributes/"variables" and a song's position in the top 50 songs in the United States Top 50 playlist. Specifically, we first examine, for each of the 14 variables, whether the top 25 songs have a different mean from the mean of the bottom 25 songs. Then, we analyze, using linear multiple regression analysis followed by linear stepwise regression analysis, the relationship between position of the song on the playlist and the values of the 14 variables.

Keywords: - Spotify, Music streaming, United States Top 50, Stepwise regression analysis, Independent sample t-tests
INTRODUCTION
Long gone are the days of records screeching, turning over cassettes and replacing the batteries in one's Walkman. As these music sources have become outdated, music streaming services have replaced and surpassed them in today’s modern, technological climate. These sites allow the users to seamlessly listen to music without purchasing any physical discs or cassettes.

As the popularity of music streaming increases, so do the growth and marketing opportunities. The focus of this paper is on one music streaming service provider, Spotify. Spotify allows users to stream music on a computer for free, pay $4.99 per month for “Spotify Premium,” to get ad-free music on their desktop, or pay just under $10 per month to listen to music on one’s mobile device. Spotify offers personalization to the customer with playlists tailored to that specific user.

Some examples include Release Radar – a playlist made with songs that were recently released. Another example is the Just for you playlist, which uses an algorithm based on the customer’s listening history to determine other songs that he or she might like. A third feature – and something that is key to this paper – is the Top 50 Charts. These playlists provide up-to-date rankings of the 50 most popular songs in the United States, other countries, and the world. Spotify also attributes 13 different audio features to each song, including “danceability,” tone, energy, and loudness, among others.

Music streaming is the latest platform in what is an ever-evolving market. Allowing users to easily create playlists from the comfort of their own home, or on the go, without having to purchase discs or records, is a major step forward. In this paper, we look to discover which audio features are the most important in determining a song’s popularity. If we can determine the most important audio features, this could potentially allow Spotify to increase customer personalization and better predict which songs will become popular and use this information for promotional and other purposes.

LITERATURE REVIEW
Overview of Spotify and Music Streaming
Digital music streaming is becoming increasingly popular. According to Business Insider, Americans were responsible for 284.7 billion audio and video music streams in the first six months of 2017 (Dunn, 2017). Given these data, streaming represents over 60% of music consumption, which is an increase of over 20%, compared to the same time frame in 2016 (Dunn, 2017). These data suggest that even in this relatively new industry, there is already a lot of growth in consumer engagement.

Spotify was brought to market in October 2008, debuting in Europe, and finally getting to the United States in 2011 (Swanson, 2013). The idea behind the product was to combat music piracy, as founder Daniel Ek sought to create technology that would overcome the desire to illegally download music files (Swanson, 2013). According to A Case Study on Spotify: Exploring Perceptions of the Music Streaming Service, Eliot Van Buskirk, from Wired magazine, defined Spotify as “a magical version of iTunes in which you’ve already bought every song in the world” (Swanson, 2013). Given Spotify’s large catalog, Van Buskirk’s quote is not unfounded. A recent press release from Spotify indicated that it currently has over 30 million songs available for users to stream (Spotify, 2017). Moreover, Spotify indicated that they currently have over 140 million active users. Listeners have the ability to come and go as they please, given that Spotify offers free streaming. However, Spotify’s goal is to convert these free-users to paid subscribers. One of the major benefits of the subscriptions – and perhaps the biggest draw – are that they are ad free.

Customers paying $4.99 per month get unlimited, ad-free access to Spotify’s music catalog on his or her desktop. The subscription that costs $9.99 per month goes one step further by allowing the user to access playlists on their mobile devices and gives the account-user the ability to download playlists offline (Swanson, 2013). This, in turn, means the customer does not have to stream using their data plan or finding access to Wi-Fi (Swanson, 2013). Spotify announced that it had 60 million paid subscribers as of July 2017 (Spotify, 2017). Given this figure, that would suggest that the service has roughly 80+ million free users. With a variety of price options and benefits, and a big catalog of songs, Spotify offers value to everybody, no matter what their music-engagement level is.

Personalization
Music streaming services possess a unique opportunity to use personalization to increase acquisition (through word of mouth, or "word of web"), and especially retention and engagement.

Based on a 2016 study from Accenture, “75% of consumers are more willing to buy from companies that are able to recognize them as individuals and provide recommendations that meet their particular needs” (Garcia-Arista, 2017). Therefore, an opportunity exists for Spotify to capitalize on personalized music-recommendations. While this is something that they currently do with the playlists mentioned in the introduction, it is worth investigating if the process can be addressed quantitatively in a more useful and sophisticated way. Furthermore, in identifying the features of songs that help determine which songs are among the more popular, Spotify can recommend new songs that fit similar criteria. If users like their recommendations, that could facilitate more playlist-sharing among social media circles and therefore, increase brand awareness and engagement. Kirk Parsons, the senior director and technology, media & telecom practice leader at J.D. Power, said that the key to success is “increasingly becoming how well streaming music brands create a
viable music ecosystem that can not only support multiple types of devices, but also facilitate listeners’ social sharing and following of playlists with others” (Effler, 2016).

Based on the statistics, an opportunity appears to exist where personalization of users’ listening experiences (recommended artists/playlists) could lead to an increase in retention, purchasing, online sharing and engagement. This would be an example of what more and more companies are using and striving to improve - collaborative filtering, and more specifically, ”recommendation systems,” beginning with the relatively primitive systems used more than a decade ago by Amazon.com (Berger, Hanna, & Swain, 2006). Nowadays, recommendation systems are often quite sophisticated, using more advanced statistical techniques, such as stepwise regression, as used in this paper, and others.

METHODOLOGY

Spotify offers listeners playlists that resemble personalization – whether it’s recommending songs based on what the user previously listened to, or new releases that could be of interest to the consumer - the list goes on. Having said that, the goal here is to see if it’s possible to determine which audio features of songs play the biggest role in determining a song’s popularity. Spotify provides values for 13 different “audio features” for each song in its catalog. We also added in each artist's ranking, based on number of streams as of October 22, 2017, bringing our independent-variable count to 14. Furthermore, given that this analysis is being done on the songs in the United States Top 50 playlist, n = 50. With the 50 songs and 14 audio features, the dataset was created to perform the analysis. Lastly, it should be noted that the ranking of each song on the playlist was inverted in the analyzed dataset for the sake of getting more intuitive results (i.e., the top ranked song was given a value of 50 in the dataset and the 50th ranked song was given a value of 1). We examined the following questions:

- **Q1**: What are the descriptive statistics of the United States Top 50 playlist songs’ audio features?
- **Q2**: Do audio features differ between “Top 25” songs versus the “Bottom 25” songs on the United States Top 50 playlist?
- **Q3**: Can we determine the relationship between a song’s ranking and the following 14 variables:
  - Artist Ranking,
  - Danceability,
  - Energy,
  - Key,
  - Loudness,
  - Mode,
  - Speechiness,
  - Acousticness,
  - Instrumentalness,
  - Liveness,
  - Valence,
  - Duration,
  - Time Signature
- **Q4**: Which of the 14 variables will make it through a predictive model created by using stepwise regression? The dependent variable is song ranking.

ANALYSIS AND DISCUSSION OF RESULTS

Descriptive Statistics of Songs’ Audio Features

The first step was performing descriptive statistics on the various audio features of the songs in the United States Top 50 playlist; SPSS was used for all statistical analyses. The definitions were provided by Spotify, except for Artist’s ranking (on a scale of 1-101, 1 = the most listened to) where we found the most recent ranking of the 100 most listened to artists on the Spotify Platform (based on the rankings of October 22, 2017); all artists in the United States Top 50 playlist that were not included in the list of top 100 artists were ranked 101. Other scales: * “Danceability” (scale of 0 – 1, where 0 = least suitable for dancing and 1 = most suitable); * “Energy” (scale of 0 – 1, where 0 = low measure of intensity and activity and 1 = most intensity and activity); * “Key” (scale of 0 – 11, representing the key the song is in, 0 = C, 1 = C#... 11 = B); * “Loudness” (scale of -60 – 0 decibels [db], where -60 = quiet and 0 = loud); * “Mode” (scale of 0 – 1, where 0 = song is in a minor melody and 1 = song is in a major melody); * “Speechiness” (scale of 0 – 1, where 0 = more music/non-speech tracks and 1 = more speech-like tone such as poetry, etc.); * “Acousticness” (scale of 0 – 1, where 0 = song is not acoustic and 1 = high confidence that song is acoustic); * “Instrumentalness” (scale of 0 – 1, where 0 = song contains a lot of vocal content and 1 = high confidence the song has no vocal content); * “Liveness” (scale of 0 – 1, where 0 = song is not live and 1 = very strong likelihood the song is live); * “Valence” (scale of 0 – 1, where 0 = songs that are more negative, angry, depressed, etc. and 1 = tracks that are positive and cheery);
*“Tempo” (the estimated tempo of a song, measured in Beats per Minute);
*“Duration” (the length of the song in milliseconds);
*“Time Signature” (the time signature of a track, which typically highlights the number of beats in each bar).

The descriptive statistics appears in Figure 1:

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Listened List</td>
<td>50</td>
<td>1.00</td>
<td>101.00</td>
<td>49.4200</td>
<td>14.68817</td>
</tr>
<tr>
<td>Danceability</td>
<td>50</td>
<td>.26</td>
<td>.94</td>
<td>.7471</td>
<td>.13632</td>
</tr>
<tr>
<td>Energy</td>
<td>50</td>
<td>.35</td>
<td>.84</td>
<td>.6148</td>
<td>.13912</td>
</tr>
<tr>
<td>Key</td>
<td>50</td>
<td>.00</td>
<td>11.00</td>
<td>5.6200</td>
<td>3.95841</td>
</tr>
<tr>
<td>Loudness</td>
<td>50</td>
<td>-11.46</td>
<td>-5.09</td>
<td>-6.2200</td>
<td>1.85271</td>
</tr>
<tr>
<td>Mode</td>
<td>50</td>
<td>.00</td>
<td>1.00</td>
<td>.4800</td>
<td>.50427</td>
</tr>
<tr>
<td>Speechiness</td>
<td>50</td>
<td>.02</td>
<td>.43</td>
<td>.1217</td>
<td>.09843</td>
</tr>
<tr>
<td>Acousticness</td>
<td>50</td>
<td>.00</td>
<td>.80</td>
<td>.1856</td>
<td>.20431</td>
</tr>
<tr>
<td>Instrumentalness</td>
<td>50</td>
<td>.00</td>
<td>.21</td>
<td>.0044</td>
<td>.02970</td>
</tr>
<tr>
<td>Liveness</td>
<td>50</td>
<td>.07</td>
<td>.58</td>
<td>.1616</td>
<td>.10099</td>
</tr>
<tr>
<td>Valence</td>
<td>50</td>
<td>.10</td>
<td>.86</td>
<td>.4536</td>
<td>.19931</td>
</tr>
<tr>
<td>Tempo</td>
<td>50</td>
<td>75.02</td>
<td>180.04</td>
<td>125.0282</td>
<td>27.22017</td>
</tr>
<tr>
<td>Duration MS</td>
<td>50</td>
<td>119133.00</td>
<td>307690.00</td>
<td>211309.340</td>
<td>40001.2259</td>
</tr>
<tr>
<td>Time Signature</td>
<td>50</td>
<td>3.00</td>
<td>4.00</td>
<td>3.9600</td>
<td>.14142</td>
</tr>
</tbody>
</table>

![Figure 1: Descriptive statistics for the 14 independent variables](image)

Based on Figure 1, on the scale of 0 – 1, one can see that the mean value of Danceability within the United States Top 50 is .747. This indicates that of the top songs in the country, on average are considered relatively “danceable” by Spotify (i.e., .747 is a lot nearer to 1 than to 0.) Energy had a mean of the 50 songs of .6148. For the same reasoning as with danceability, this indicates that the songs included in the playlist slightly trend towards having relatively more energy. The mean value of “loudness” was -6.22, which suggests that the songs that were analyzed are relatively loud (much nearer to 0 than to -60). Interestingly enough, while the scale on this audio feature (in db.) was -60 – 0, the minimum value of any of the top 50 songs was (only) -11.46, suggesting that none of the songs that in the top 50 can be considered to be “on the quieter side.” Regarding “instrumentalness,” the mean is .0044, which suggests that the top 50 most popular songs in the United States were virtually all vocal. For Valence, the mean value was .4536. While this value is relatively neutral, it is still slightly shifted towards negative-valence songs. One final value that is worth highlighting is the “duration,” which registered a mean of 211,306 milliseconds. This value is equivalent to 3 minutes and 31 seconds.

**Independent Samples t-Tests for differences between Songs in the Top 25 (1) vs. Bottom 25 (0)**

An independent samples t-test for difference in means was run for each of the 14 variables (13 audio features plus artist rank) to test whether there is a difference in mean for those songs ranked as one of the top 25 songs in the United States Top 50 playlist vs. those in the bottom 25 of that same list. The group statistics are shown in Figure 2a, while the t-test results can be found in Figure 2b.
### Group Statistics

<table>
<thead>
<tr>
<th></th>
<th>In Top 25</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Listened List</td>
<td>0.00</td>
<td>25</td>
<td>45.3600</td>
<td>46.21078</td>
<td>8.04216</td>
</tr>
<tr>
<td>Danceability</td>
<td>0.00</td>
<td>25</td>
<td>53.4800</td>
<td>28.38004</td>
<td>5.67601</td>
</tr>
<tr>
<td>Energy</td>
<td>0.00</td>
<td>25</td>
<td>.7476</td>
<td>.12079</td>
<td>.02416</td>
</tr>
<tr>
<td>Key</td>
<td>0.00</td>
<td>25</td>
<td>.7465</td>
<td>.15280</td>
<td>.03056</td>
</tr>
<tr>
<td>Loudness</td>
<td>1.00</td>
<td>25</td>
<td>6.5600</td>
<td>3.29242</td>
<td>.65848</td>
</tr>
<tr>
<td>Mode</td>
<td>0.00</td>
<td>25</td>
<td>4.6800</td>
<td>4.59431</td>
<td>.87886</td>
</tr>
<tr>
<td>Specchiness</td>
<td>0.00</td>
<td>25</td>
<td>.6151</td>
<td>.14843</td>
<td>.02969</td>
</tr>
<tr>
<td>Acousticness</td>
<td>0.00</td>
<td>25</td>
<td>.6146</td>
<td>.13266</td>
<td>.02653</td>
</tr>
<tr>
<td>Instrumentalness</td>
<td>1.00</td>
<td>25</td>
<td>.0909</td>
<td>.09000</td>
<td>.00009</td>
</tr>
<tr>
<td>Liveness</td>
<td>0.00</td>
<td>25</td>
<td>.0888</td>
<td>.08903</td>
<td>.01781</td>
</tr>
<tr>
<td>Valence</td>
<td>0.00</td>
<td>25</td>
<td>.1850</td>
<td>.18943</td>
<td>.03789</td>
</tr>
<tr>
<td>Tempo</td>
<td>1.00</td>
<td>25</td>
<td>.1862</td>
<td>.22213</td>
<td>.04443</td>
</tr>
<tr>
<td>Duration MS</td>
<td>0.00</td>
<td>25</td>
<td>.1726</td>
<td>.12410</td>
<td>.02482</td>
</tr>
<tr>
<td>Time Signature</td>
<td>0.00</td>
<td>25</td>
<td>.1942</td>
<td>.13857</td>
<td>.02771</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>25</td>
<td>.1229</td>
<td>.23307</td>
<td>.04661</td>
</tr>
</tbody>
</table>

Figure 2a: Group statistics for the 14 variables
As illustrated in Figure 2b, the mean value for “Key” (the fourth audio feature listed in Figure 2) for songs that were in the top 25 was 6.56 while the value for the songs in the bottom 25 of the playlist was 4.68. Using a value of \( \alpha = .10 \), these results are significant \( (p = .094.) \). Assuming that this is a “random sample” of a playlist—the top 50 can change every day—we can conclude that songs in the top 25 have a higher Key than songs in the bottom 25. Another audio feature worth noting is “Valence.” The mean value for songs in the top 25 was .3942 while the value for the bottom 25 was .5129. We have \( p = .035 \), and at \( \alpha = .05 \), and can conclude that the top 25 songs have a lower mean valence (more negative/tragic/angry/depressed tone), than the bottom 25 songs. All of the other variables had \( p > .25 \).

Multiple Regression Analysis

We next ran a multiple linear-regression analysis. The goal was to determine the relationship between a song’s ranking and the aforementioned 14 different independent variables.

The results of the multiple regression can be seen in Figure 3, including the Model Summary, ANOVA results, and Coefficient Table:
Of the 14 independent variables entered into the multiple linear-regression, only three were significant at $p = .05$. They are “Key” ($p = .049$), “Loudness” ($p = .020$) and “Valence” ($p = .009$). The coefficients of each of these three variables seem to make sense, given what we have already learned previously. The louder the song, the higher the ranking (positive coefficient); the higher the key value, the higher the ranking (positive coefficient); the lower the valence, the higher the ranking (negative coefficient); the latter is directly consistent with our t-test result, although in the latter case, the coefficient reflects an effect holding the other variables in the model constant, while the t-test result is examining the means with no restrictions on the values of the other variables. Note that it needs to be remembered (as previously mentioned in the Methodology section) that a higher numerical ranking implies a better/higher ranked song - the top rated song was given the value "50," while the 50th rated song was given the value "1."

Overall, the R-squared of the 14 variables was somewhat modest at .356, suggesting that (we estimate that) this set of 14 variables can explain 35.6% of the variability in a song’s ranking. However, we can see in Figure 3 that there are several non-significant variables. It is possible that some of these variables are irrelevant to determining rank, while it is also possible that two or more of the non-significant variables in Figure 3 are important to determining rank, but are "overlapping," or co-linear with each other. Thus, we clarify this issue by performing a stepwise regression.

**Stepwise Regression**

Our final analysis is conducting a stepwise regression analysis. We used the same dependent variable of the song rankings and used the previously analyzed 14 independent variables as eligible variables. We used a p-to-enter of .10. The results
of the stepwise regression can be seen in Figure 4, again having three parts, Model Summary, ANOVA, and Coefficient Table:

![Figure 4: Stepwise Regression Results](image)

Based on Figure 4, we see an interesting result in the stepwise-regression. The full multiple regression contained three significant variables; however, the number dropped to two in the stepwise regression. The variable not in the final stepwise regression model is "Key." Given the fewer number of variables (2 vs. 14), and hence, a different number of degrees-of-freedom of the error term, we have different benchmarks for adding sufficient predictive value to be significant at different significance levels. Apparently, the variable of “Key” was deemed to not add enough predictive value to the two variables already in the model, given the variability, p-to-enter, etc., and hence, did not enter the model. The final model above has R-square value of .131, which is substantially less than the .356 of the full multiple regression of the previous section. However, this is often the case with a much larger number of variables in the full multiple regression (here, 2 vs. 14), since each non-significant variable adds "a little bit" to the R-square value, while not even necessarily being close to significant - here, we have 12 "little bits," resulting in a much lower R-square for the stepwise regression. However, it might be noted that the standard error of estimate, a core quantity for accuracy of prediction, is virtually the same for the (final model of the) stepwise regression (13.87) as it is for the full multiple regression with all 14 independent variables (13.84). Assuming that we do not wish to have any non-significant variables in our final model, our prediction model for the value of a song’s ranking is:

\[
Y \text{- predicted} = 34.194 - 19.167*(\text{value for valence}) + 2.041*(\text{value for loudness})
\]

**LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH**

This analysis has several limitations that can be addressed if a follow-up project were to be performed. First, the particular ranking of songs within the *United States Top 50* playlist was a snapshot taken on a given day. In general, the playlist rankings can shift daily, depending on what consumers are listening to. This can lead to a situation where a similar analysis a week (or, in theory, *even a day*) later can lead to different results. Better might be to consider average rankings over a period of time, perhaps a week or more, to try to smooth out, and reduce the variability of the results. Furthermore, the rankings might be too strongly affected if a popular artist were to release an album soon before the time period of ranking used in the study. In this instance, the rankings would most likely consist of a handful of songs from the same artist –
potentially with very similar audio features – which can, of course, skew the results, and make them not representative of the general case.

Secondly, this project focuses strictly on the playlist highlighting the top 50 songs in the United States. Nevertheless, Spotify offers playlists of the top songs in several other countries and even worldwide. This could mean that what we learn from the United States needs to be interpreted only as such and not generalized across other countries, until further investigation is performed.

Third, the ranking of Top 100 artists, of course, reaches only 100. Therefore, we used the proxy value of 101 for each artist that was not included on the list. However, there is a chance that some artists who were listed as the 101st most popular artist were, in fact, much lower on the list. The results might differ if a different approach were taken to quantifying the ranking of artists not in the (then current) top 100.

Finally, the audio features that were used in this analysis were assigned by Spotify. This means that different music streaming platforms, if offered, could provide different ratings or different features altogether. If this is the case, then the results included herein would not apply across other streaming sites.

REFERENCES