Proceedings of The National Conference
On Undergraduate Research (NCUR) 2018
University of Central Oklahoma
Edmond, Oklahoma
April 5-7, 2018

# Data and Soul: Introductory Data Mining and Data Analysis on Music Data

John Vonelli
Mathematics and Computer Science
Moravian College
1200 Main Street
Bethlehem, Pennsylvania 18018 USA

Faculty Advisors: Dr. Brenna Curley and Dr. Thyago Mota

## Abstract

The purpose of this research is to investigate how songs lyrically and musically influence popularity and emotional responses throughout contemporary history. We build a dataset of 27,346 songs that are listed on the Billboard "Hot 100" list from 1958 to 2017. We then use Spotify's song metrics, together with a weighted sampling function, to evaluate how music changed over time. Our analysis shows that popular music is becoming louder, more energetic, and lyrically dense. Acoustic and instrumental popular songs have severely declined since the 1960s while danceability, tempo, and liveness of the analyzed songs remained consistent over the years. Duration reached a maximum value around 1993, after which it started to decline.

**Keywords: Data Mining, Billboard, Spotify**

## 1. Introduction

The purpose of this research is to investigate how songs lyrically and musically influence popularity and emotional responses throughout contemporary history. Not so long ago, Spotify, one of the most popular music streaming services, released a searchable interface to its vast music catalog. Spotify describes songs in an unprecedented level of detail including such attributes as danceability, liveliness, and tempo. Along with Spotify, we use the Billboard "Hot 100" list to identify a selection of popular songs from 1958 to 2017. Music is a cultural hallmark and is considered a universal language regardless of discipline. By analyzing songs from different perspectives using appropriate data mining techniques, we can discover patterns that help us better understand the impact of music in our lives. For example, is there a correlation between romantic danceable songs and popularity; or, how has the quality and content of songs changed throughout the decades? In this paper we describe our findings with the hope that we improve the reader's knowledge about modern popular music; and, ultimately, pave the way for future research related to these data.

## 2. Background and Related Work

One of the main services offered by Spotify to its users is their recommendation of songs. By collecting data on a user's listening habits, Spotify and other popular music streaming services can recommend similar songs that the user might enjoy. To be able to compare songs, Spotify uses advanced machine learning techniques to associate high-level features, like danceability and acousticness, to each song in its vast catalog. These high-level features, listed in Table 1, are available programmatically through Spotify for Developers *web services* API [1]. A *web service* is a software service that is consumed over the web, allowing applications to exchange data in standard formats, such as JSON or XML, using HTTP (the web transport service).

Table 1. Spotify Song Features [2].

| Feature | Range | Type | Unit | Description |
|---|---|---|---|---|
| Acousticness | [0,1] | Floating-point | NA | A confidence measure of whether the track is acoustic, 1.0 showing high confidence the track is acoustic. |
| Danceability | [0,1] | Floating-point | NA | Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| Duration | Variable | Floating-point | ms | The duration of the track in milliseconds. |
| Energy | [0,1] | Floating-point | NA | Represents a perceptual measure of intensity and activity of a track. Typically, energetic tracks feel fast, loud, and noisy. |
| Instrumentalness | [0,1] | Floating-point | NA | The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| Key | [0,11] | Integer | NA | Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭ , 2 = D, and so on. |
| Liveness | [0,1] | Floating-point | NA | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. |
| Loudness | Variable | Floating-point | dB | The overall loudness of a track in decibels (dB). Values typical range between -60 and 0 db. |
| Mode | [0,1] | Integer | NA | Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| Speechiness | [0,1] | Floating-point | NA | Detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0 the attribute value. |
| Tempo | Variable | Floating-point | BPM | The overall estimated tempo of a track in beats per minute (BPM). |
| Time Signature | Variable | Floating-point | BPB | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (BPB). |
| Valence | [0,1] | Floating-point | NA | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |

Glenn McDonald, working for Echo Nest, a company acquired by Spotify in 2014, traced 5,000 top popular songs from 1950 to 2013 and found that music is becoming more energetic, less acoustic, and louder [3,4]. Unfortunately, it is not clear how McDonald sampled popular songs, making it difficult to reproduce the results described in his work. In a more recent study, Interiano et al. [5] analyzed a large collection of 500,000 songs released in the UK between 1985 and 2015. The goal of this research was to understand how a song becomes popular (defined as 'making it' into the top charts). According to the authors of the study, songs in general are "in a clear downward trend in happiness and brightness." However, "successful songs exhibit their own distinct dynamics." In other words, popular songs tend to be happier, more danceable and energetic.

## 3. Methodology

As a part of this work, we use the well-known "Hot 100" song chart regularly published by Billboard magazine [6] to create a timeline of popular songs in the United States. We developed a Python script to download Billboard "Hot

100" song charts from 1958 to 2017, totaling 27,346 songs. Our script used "billboard-charts" [7], a web-scraping library developed by Allen Guo in Python. Web-scraping is a technique that is applied to extract information from a web page. After acquiring our initial dataset of songs, we searched Spotify's online music catalog to collect the song features listed in Table 1. To help search Spotify's music catalog we used Paul Lamere's "spotipy" [8], a lightweight Python library for the Spotify API. At the end of our data collection, 15% of Billboard songs were not found on Spotify, mainly due to inconsistencies with Billboard's naming scheme. However, the missing songs across all dates represented less than 5% of the songs in a particular "Hot 100" chart, with the exception of August 4th 1958 that had 28 songs missing features. This specific date was removed from further analysis.

To analyze how popular songs evolve over time, we select songs from each of the Billboard chart dates using a *weighted sampling* scheme, where we sample with replacement [9]. The chance of a song `x` selected from a particular date `d` is determined by the *rank* of the song on the chart and the *longevity* of the song. The longevity of a song is defined as the number of consecutive weeks the song has been listed on Billboard before date `d`. We include longevity when sampling popular songs because we think that even if a song did not reach the top of the chart, the fact that a song has been consistently listed in the "Hot 100" chart is a strong indicator of its popularity. We define a sampling probability function by,

$$P_d(x) = \alpha_r \left( \frac{\hat{r}_d(x)}{\sum_{i=1}^{100} \hat{r}_d(x_i)} \right) + (1 - \alpha_r) \left( \frac{\ell_d(x)}{\sum_{i=1}^{100} \ell_d(x_i)} \right), \tag{1}$$

$$\hat{r}_d(x) = 101 - r_d(x), \tag{2}$$

where $r_d(x)$ is the rank of song `x` on date `d` that varies from `1` to `100`, with `1` being the highest rank. Equation (2) inverts the rank of a song so the highest rank becomes `100` (instead of `1`). The value of $l_d(x)$ is the longevity of song `x` on date `d`. The weights given for rank and longevity are determined by the parameter $\alpha_r$ and its complement $(1 - \alpha_r)$, respectively. We used a sample size of `100` where samples were taken with replacement. In case a selected song did not have Spotify features, we assigned the average value of a given feature for all songs on that date as the feature value for that selected song. Algorithm 1 describes how the sampling probability function was implemented. All of the code written for this study is available at the "Data-and-Soul" GitHub repository [10].

Algorithm 1: Sampling Probability Function.

```
Input: Pd and αr
Output: x (selected song)
```

```
01. r ← random number in [min(Pd) , 1]
02. for each song x list in chart d do
03.     r = r - Pd(x)
04.     if r <= 0 then
05.         return x
06.     end if
07. end for
```

## 4. Data Analysis

In our preliminary analysis of popular songs we set the parameter $\alpha_r$ in (1) to $\alpha_r = 0.5$, therefore giving equal weight for song rank and longevity. When analyzing the results using different values for $\alpha_r$, such as $\alpha_r = 0.25$ and $\alpha_r = 0.75$, we did not see any noticeable difference. Since our results are robust to the choice of $\alpha_r$, moving forward we will assume that rank and longevity are equally important when sampling popular songs. Figure 1 shows boxplots for the Spotify features collected for popular songs using $\alpha_r = 0.5$. For the features with values ranging from `0` to `1`, *acousticness* had the highest standard deviation ($\sigma^2 = 0.162$), followed by *mode* ($\sigma^2 = 0.096$), and *energy* ($\sigma^2 = 0.008$). It is worth highlighting the variability of *loudness* in popular music, with a standard deviation of `2.07` dB.

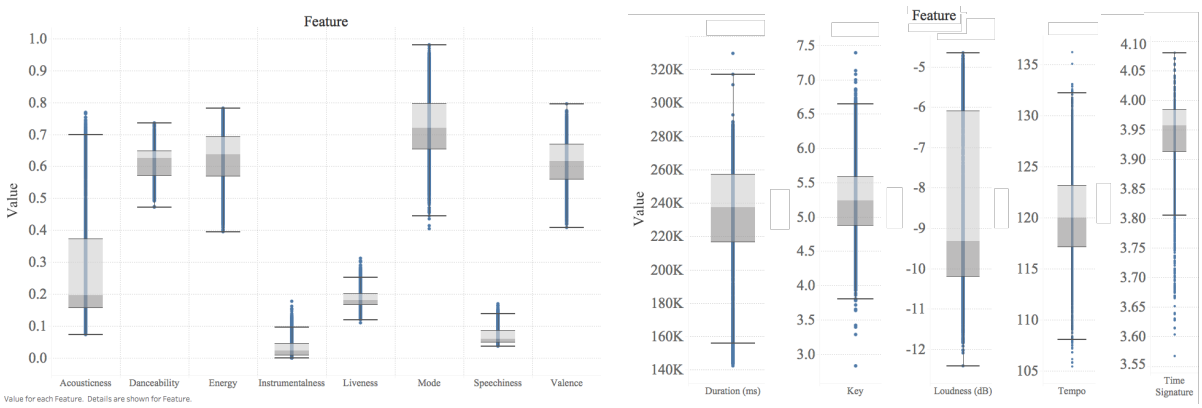



Figure 1. Boxplots for the Spotify features collected for popular songs.

The primary goal of our research was to analyze how popular songs changed over time. Our analysis shows that popular music has become louder, more energetic, and lyrically dense. Acoustic and instrumental popular songs, on the other hand, have severely declined since the 1960s. We note that more danceable music took over the charts, like disco, electronic, and funk music, while traditionally popular acoustic and instrumental music is not as much in the forefront as it used to be. We hypothesize that this trend may be because acoustic songs, like folk music or ballads do not have a distinguishable beat as compared to their counterparts. Popular songs also have a tendency to express more negative feelings (starting around 1985) becoming more sad, depressing, or angry in tone. We believe that this may be due to the introduction of hip hop and heavy metal music into the mainstream. The falling valence values over the last decades is an indication of negative expression in popular music, although the average value for the feature still gravitates around $0.5$. These results are illustrated by the time plots on Figure 2. We did not see a large change in danceability, tempo, or liveness of the analyzed songs (and hence these time plots are not shown here). Duration reached a maximum value around 1993, after which it started to decline.

In our research, we also looked at how different song features correlated with each other. We computed Pearson's correlation coefficient [11] for all of the $156$ (or $13 \times 12$) possible feature pairs. We then listed all pairs that showed a strong, positive or negative correlation ($r \geq |0.75|$ in Table 2). We found seven strong correlations that gave us some interesting insights about popular songs. For example, acousticness is negatively correlated with energy, duration, and danceability; thus, acoustic songs, which can be understood as songs played without the use of electronic amplification equipment, tend to be less energetic, less danceable, and shorter, compared to non-acoustic songs. For example, Simon and Garfunkel's "For Emily, Whenever I May Find Her" is less energetic and danceable compared to Earth Wind and Fire's "September." Additionally, acoustic songs are more likely to have three beats per measure compared to other songs, as illustrated by the negative correlation between time signature and acousticness. The strong, positive correlation between loudness and speechiness suggests songs that are louder also tend to have a high number of spoken words, like Dr Dre's "Forgot About Dre" or Metallica's "Enter Sandman." Those type of songs also have a tendency to express negative feelings as supported by the negative correlation between loudness and valence.

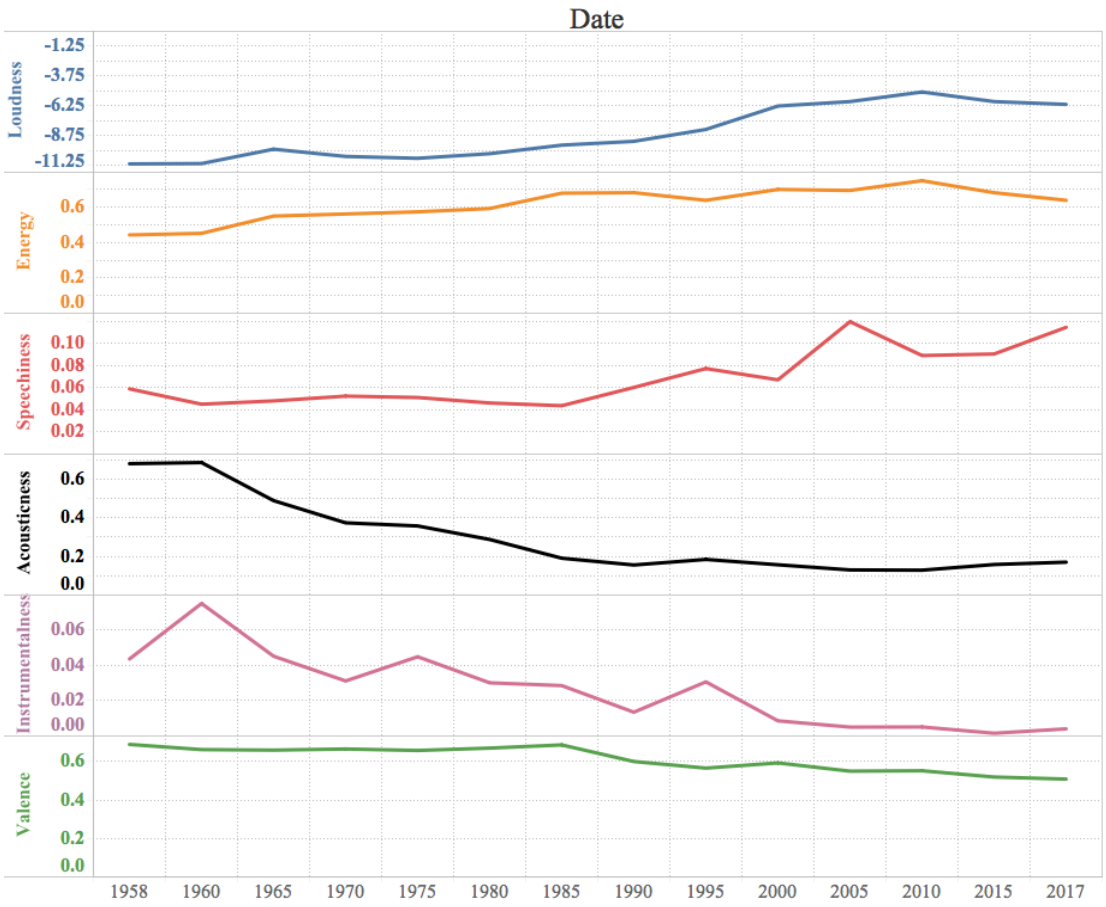


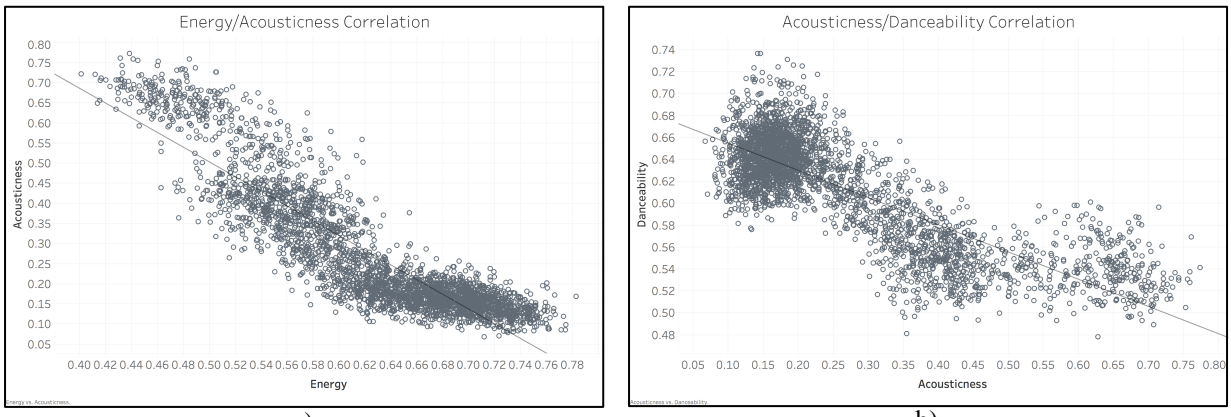

Figure 2. Time plots of *loudness*, *energy*, *speechiness*, *acousticness*, *instrumentalness* and *valence* of popular songs.

Table 2. Estimated Pearson correlation coefficients for song features of popular songs that are strongly correlated.

| Feature Pair | Pearson Correlation Coefficient |
|---|---|
| {loudness, energy} | +0.770 |
| {loudness, speechiness} | +0.767 |
| {energy, acousticness} | −0.888 |
| {duration, acousticness} | −0.821 |
| {danceability, acousticness} | −0.808 |
| {time signature, acousticness} | −0.774 |
| {loudness, valence} | −0.755 |

Figure 3 shows scatter plots comparing the values of two pairs of strongly correlated song features: {energy, acousticness} and {danceability, acousticness}. As can be seen from the plots, the relation between each pair of features is not linear. The plots also show us that there are different groups of songs depending on the relative values

of the features. For example, in Figure 3 a) we can see that there is a sparse group of songs that are highly acoustic and low in energy while there is a much denser group of songs that are energetic and less acoustic. Similar observations can be made for the other plot on Figure 3 b). A cluster analysis will be helpful to shed some light on what these groups of songs may represent.




a) b)

Figure 3. Scatter plots comparing a) energy with acousticness and b) acousticness with danceability.

## 5. Conclusions and Future Work

Based on our preliminary analysis, we found that popular music tended to become louder, more energetic, and lyrically dense over time. Given the features that we analyzed, we also saw that loudness in songs correlate to a lower valence. Surprisingly, danceability did not appear to be related to other song features and it remained relatively static over the years. This conclusion goes against our initial hypothesis, that danceability would have a much larger role in the development and change of popular music over time. Lastly, another notable attribute was acousticness, which showed a steep decline since the 1960s.

For our future work, we hope to apply clustering techniques to identify groups of songs seen in the scatter plots. Additionally, we want to leverage this dataset with song lyric data to enrich our research. More specifically, we want to perform a sentiment and eloquence analysis on each song, compute our own metrics from this analysis, and compare these with Spotify's features. Lastly, we recognize that our research does not take into account the relationships between music and the listener, as well as the performer's relationship with their music. We hope to address this in our future work.

## 6. Acknowledgments

We gratefully acknowledge the support of Moravian College SOAR travel funding to attend NCUR 2018 Conference.

## 7. References

1. Spotify Web API. https://beta.developer.spotify.com/documentation/web-api.

2. Spotify for Developers, Web API Documentation, Audio Features Object, https://beta.developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features.

3. McDonald, G., The Echonest, "7 Decades, 7 Musical Evolutions," http://blog.echonest.com/post/70299217721/7-decades-7-musical-evolutions.

4. The Guardian, "Pop music is louder, less acoustic and more energetic than in the 1950s," November, 2013, https://www.theguardian.com/technology/2013/nov/25/pop-music-louder-less-acoustic.

5. Interiano, M., et al., “Musical Trends and Predictability of Success in Contemporary Songs In and Out of The Top Charts,” Royal Society Open Science, 16 May 2018, DOI: 10.1098/rsos.171274.
6. Billboard’s “Hot 100” Chart. https://www.billboard.com/charts/hot-100.
7. Guo, A., “billboard-charts” GitHub Repository, https://github.com/guoguo12/billboard-charts.
8. Lamere, P., “spotipy” GitHub Repository, https://github.com/plamere/spotipy.
9. Navidi, W., “Statistics for Engineers and Scientists,” McGraw-Hill, Third Edition, ISBN-13: 978-0-07-337633-2.
10. Vonelli, J., “Data-and-Soul” GitHub Repository, https://github.com/jpvonelli/Data-and-Soul.
11. Han, J., Kamber, M., and Pei, J., “Data Mining: Concepts and Techniques,” Morgan Kaufmann, Third Edition, ISBN: 978-0-12-381479-1.