# Spotify Song Track Dataset Introduction

The datasets were obtained from Kaggle at https://www.kaggle.com/edalrami/19000-spotify-songs#song_info.csv .

The combined dataset has 18 variables and 18,835 rows.

| Type | Variable Name | Range | Notes |
|---|---|---|---|
| Metric (11) | Song_ popularity | [0,100] | Dependent variable |
| | Song_duration | [1.200000e+04,1.799346 e+06] | In milliseconds |
| | Acousticness | [0,0.996] | 1.0 represents high confidence the track is acoustic. |
| | Danceability | [0,0.897] | A value of 0.0 is least danceable and 1.0 is most danceable |
| | Energy | [0.001,0.999] | A value of 1 means most energetic (tracks feel fast, loud, and noisy) |
| | Instrumentalness | [0.01,0.986] | Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| | Liveness | [0,0.997] | Higher liveness values represent an increased probability that the track was performed live |
| | Loudness | [-38.768, 1.585] | values are averaged across the entire track |
| | Speechiness | [0,0.94] | Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| | Tempo | [0 ,242.318] | |
| | Audio_valence | [0,0.984] | Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric) |
| Ordinal (3) | Key | [0,1,2,3,4,5,6,7,8,9,10,11] | 0 = C, 1 = C♯/D♭, 2 = D, and so on |
| | Audio_mode | [0 or 1] | Major is represented by 1 and minor is 0 |
| | Time_signature | [0,1 ,3,4,5] | |
| Nominal (4) | Album_name, song_name, play_list, artist_name | | |

- **Is there an obvious parameter of interest, or are there several:**

  Song popularity appears to be the main  parameter of interest, though there may be other smaller ones. For example, it could be possible to look at audio_valence as a possible parameter of interest since this looks at the "positiveness" of a track. We would investigate how variables such as danceability, energy, liveliness, loudness, etc. contribute to the "positiveness" of the song.

- **What interesting metric variables does the data have**

  There are quite a few interesting metric variables including:
  - Song_popularity - our dependent
  - Danceability, Liveness, Loudness, Energy, Speechiness: Numerical, will be interesting to see how these skew and if more popular songs are higher in these ranges or lower.

- **What interesting categorical variables does the data have**

  There are four categorical variables: artist name, playlist, album name and song name. We also have information to create some. For example, we can add the genre of each of the songs as well as maybe the decade in which the song was created. It would be interesting to see if these also help with our prediction variable.

- **Are some of the variables ordinal (ordered but not metric like we discussed in class?)**

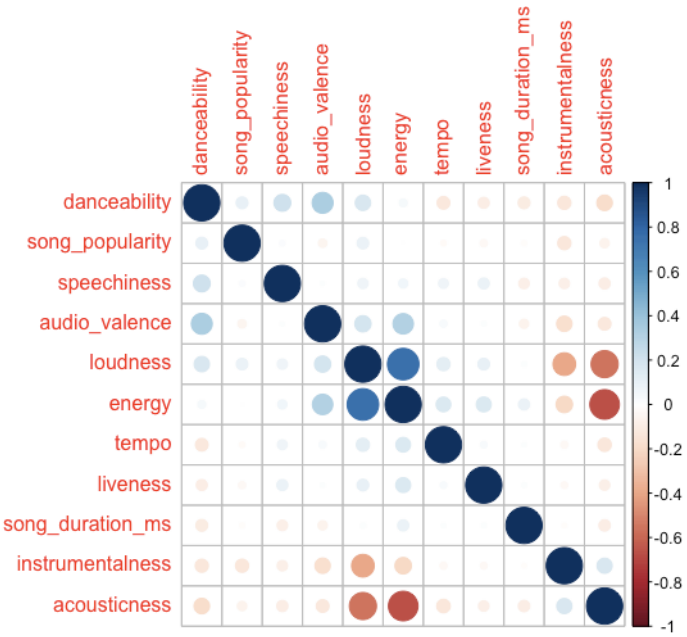
  Key, audio_mode and time_signature are ordinal variables. Although, in the later stage of analysis, we might recode some of the numerical variables into ordinal variables if necessary.

- **What missing values are there and are there any patterns that you might exploit for filling them in?**

  There are not missing values in the dataset. However, some songs have a song popularity score of 0. We will detect if there is a pattern of these instances.

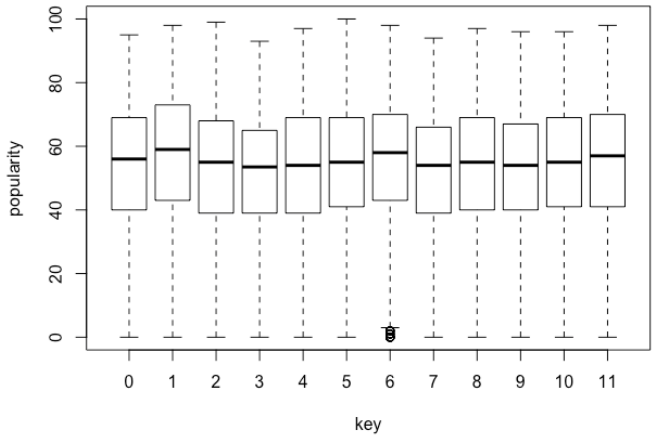
- **What variables look like they might interact?**

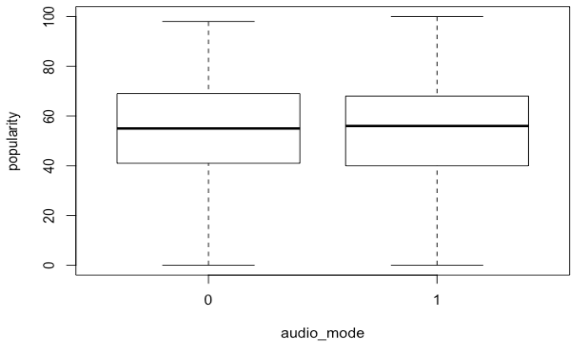
  Based on the variable description, energy and loudness may interact. Since one of the ways that energy is measured (per the description) is through "intensity". We can assume that as a song gets more intense, the loudness might increase as well. Acousticness and energy also show a strong negative correlation since acoustic songs typically are less loud, leading to lower energy levels.



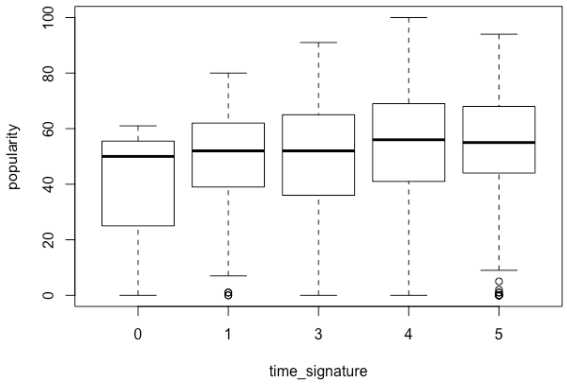

- **For the metric variables, what units are used and are different variables measured in related but different units? (e.g. cm and mm or feet and miles)?**
    - Song duration is measured in milliseconds
    - Most of the other metric variables are in a 0 to 1 scale

- **For some of the categorical variables, what dependencies are there between any metric variables (parameters of interest?). Explore with boxplots, or if you need to investigate categorical/categorical interactions, look at correspondence analysis.**






No significant different in song popularity whether audio mode is major or minor.




The medians of popularity score in different time_signatrue level are about the same. But their distributions vary. When time_signature is 4, popularity score has the greatest variance. Popularity score is skewed to the left if the time_signature is 0. There are outliers in the distributions of time_signature equal to 1 and 5.

- **What obvious directions for investigation present themselves?**
    - Consideration of engineering additional variables (I.e. artist age, season, number of artist Grammys, etc.)
    - We are going to find out the features of a popular song track. Since we concluded some of the variables are interactive, we will perform Principal Component Analysis and Factor Analysis to see if there are new “factors” that would help explain a popular song.

- **Do any variables need to be scaled?**
  Yes, a majority of them might need to be scaled. Song duration for example, is measured in milliseconds and we see it has very large values because of this. Other variables like audio valence, for example, is measured on a scale from 0 to 1 so it shows much smaller numbers.